An Empirical Process for the Design of High-Throughput Screening Deck Filters

Bradley C. Pearce,*,† Michael J. Sofia,† Andrew C. Good,‡ Dieter M. Drexler,§ and
David A. Stock||

Bristol-Myers Squibb Pharmaceutical Research Institute, 5 Research Parkway,
Wallingford, Connecticut 06492

Received November 17, 2005

A process for objective identification and filtering of undesirable compounds that contribute to high-throughput screening (HTS) deck promiscuity is described. Two methods of mapping hit promiscuity have been developed linking SMARTS-based structural queries with historical primary HTS data. The first compares an expected assay hit rate to actual hit rates. The second examines the propensity of an individual compound to hit multiple assays. Statistical evaluation of the data indicates a correlation between the resultant functional group filters and compound promiscuity. These data corroborate a number of commonly applied filters as well as producing some unexpected results. Application of these models to HTS collection triage reduced the number of in-house compounds considered for screening by 12%. The implications of these findings are further discussed in the context of the HTS screening set and combinatorial library design as well as compound acquisition.

INTRODUCTION

In recent years, high-throughput screening (HTS) has become a dominant strategy for lead identification within the pharmaceutical industry. A significant advantage of HTS is its ability to produce unanticipated discoveries, the likelihood of which improves with having a diverse, high quality compound screening deck. Serendipitous hits are particularly important with less explored targets where existing knowledge bases have gaps, limiting the rational approaches available for lead identification. Finding these discoveries from the milieu of HTS data is made more difficult when HTS assays contain high levels of random noise. While many factors contribute to such noise in screening, compound quality is undoubtedly a major component.[1−3] Methods that enhance screening compound quality are thus of importance as they facilitate the reduction of false positives (and potentially negatives) that consume valuable resources and lengthen development times.

The suitability of HTS hits for transition into a medicinal chemistry program depends on multiple criteria including synthetic tractability, patentability, and the potential for ADMET (absorption, distribution, metabolism, excretion, toxicity) liabilites. A popular technique for compound triage is to employ property filters such as the “rule of 5” guidelines outlined by Lipinski.[4] In addition to property-based constraints, a number of researchers have developed filters that flag or remove unwanted compounds prior to screening.[5−7] Filters that remove compounds with reactive functional groups based on Daylight SMARTS were first published several years ago by Glaxo Wellcome.[8] Others have published more general rules eliminating problematic functional groups.[7] Traditionally our own approach toward the design of functional group (FG) filters has relied, in part, on literature guidance and, to a significant extent, on the input of experienced medicinal chemists. These chemists participate in HTS compound hit triage teams. However, as a recent publication from Pharmacia pointed out, agreement about which compounds should be allowed is elusive.[9] This study found that, rather than achieving consensus, any two medicinal chemists will agree only 28% of the time on the tractability of a given compound. Further, the same medicinal chemist will reject the same compounds 50% of the time when it appears in different points on a list. Thus, decisions about compound suitability depend on the biases of the individual medicinal chemist and even that is inconsistent. As a consequence it is clear that a less subjective method for determining compound suitability is needed.

In order reduce the subjectivity of compound property or FG filters, we have applied empirical methods to help validate filter utility through analysis of in-house HTS data. Initially a high level data view was applied to determine trends relating to “frequent hitter” or “promiscuous” molecules flagged by either a property or FG filter. Next a more detailed analysis relating FG filters to HTS compound promiscuity was undertaken. This investigation comprised two quantitative views of compound promiscuity, both statistically linked to a particular FG filter. The linkage of HTS data results to specific types of molecules depends on the structural integrity of the samples, and this was also addressed. Examples are discussed in the context of applying filters to library design, acquiring compounds from external sources, and building a screening deck.[10]

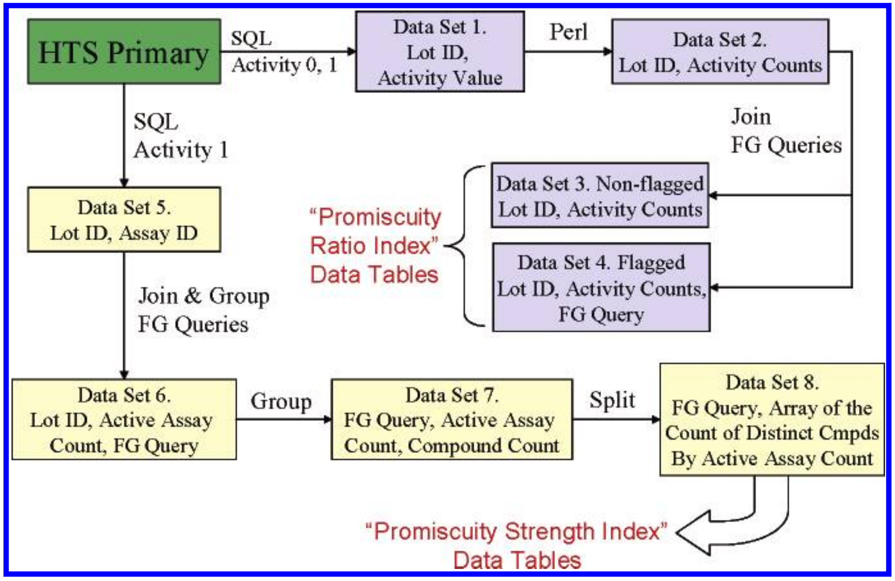
METHODS

In the section that follows the techniques applied to HTS data extraction, filter constraints and promiscuity index definitions are described in detail.

* Corresponding author phone: (203)677-6904; fax: (203)677-7702; e-mail: Bradley.pearce@bms.com.
† New Leads Cheminformatics; Applied Biotechnologies.
‡ Computer Assisted Drug Design, Applied Biotechnologies.
§ Discovery Analytical Sciences; Pharmaceutical Candidate Optimization.
|| Non-Clinical Biostatistics.

10.1021/ci050504m CCC: $33.50 © 2006 American Chemical Society
Published on Web 04/05/2006

Scheme 1. Process Map for Primary HTS Data Mining



Property Filters. We currently use a modification of the phase I clinical development guidelines described by Wenlock wherein, a compound is flagged if it has two or more violations where the molecular weight is >639, cLogP is < -3.0 or >5.5, the number of hydrogen bond donors >5, the number of hydrogen bond acceptors >9, and the number of rotatable bonds >14.[11] A hard molecular weight cutoff of <130 or >900 is also applied.

Functional Group Compound Filters (FG Filters). The FG filters consist of a series of molecular query strings written using the SMARTS coding language described by Daylight.[12] The FG filters that comprise this study are a combination of exclusion and informational filters. Exclusion FG filters are those intended for compound removal from screening decks. Informational filters are useful for compound annotation. For example, trifluoromethyl ketones may not be explicitly excluded from screening, but annotation could be helpful in the context of the assay under evaluation (e.g. serine protease). These filters, like many others currently being applied within the industry for triage, are based on chemical intuition and experience (bias). A set of 180 of these SMARTS strings, including a full set of exclusion and informational filters, can be found in the Supporting Information (functional group SMARTS). A protocol was written using SciTegic's Pipeline Pilot to identify compounds that were flagged by any of the FG filters. The protocol can be downloaded from Scitegic's user group Web site.[13]

High-Throughput Screening (HTS) Data Extraction. The BMS corporate HTS data are stored in Oracle tables in three stages as primary, retest, and concentration response curves. A map for the processing of primary HTS data for all aspects of this analysis is shown in Scheme 1.

The compound HTS primary activity classification data was obtained from Oracle using an sql query, resulting in data set 1 (all data set data not shown), which is a large table of approximately 61 million records. Data set 1 in the form of rows of column lot IDs (Oracle primary key) and activity values ($0 =$ inactive, $1 =$ active) were processed using a Perl script,[14] resulting in data set 2 containing 1.4 million records. Data set 2 is a table of columns having unique lot IDs and counts of inactive and active assays. For each given compound in data set 2, the table lists the total number of times the compound was designated active or inactive in HTS. A join by lot ID of data set 2 and all FG flagged compounds provided nonflagged and flagged data sets 3 and 4.[15] From data sets 3 and 4 the percent actives for HTS nonflagged and flagged classifications (by FG filter) were obtained. These form the basis of the "promiscuity ratio index" described below.

Following the alternate pathway in Scheme 1, another sql query was written to extract distinct lot ID and assay ID records from the Oracle tables containing primary HTS screening data. Only compounds that were active as defined by the screener's threshold for a particular assay were included. This resulted in data set 5 containing approximately 1.8 million records from 362 different assays run over the last 12 years.

From the entire set of compounds active in primary HTS assays, 17% of compounds are flagged using the FG filters used in this analysis. Data set 6 contains compounds by lot ID linked to a particular FG query along with a count of assays in which the compound was active (active assay count). Grouping the compounds in data set 6 by FG query and active assay count provided data set 7. Finally, a split of data set 7 produced data set 8. Data set 8 contains 168 rows with an array of 51 columns. Rows of the first column contain the name of each FG query, with subsequent columns housing the number of active compounds associated with each active assay count. Column 2 contains the number of compounds active in exactly one assay, going across to column 51, which contains the number of compounds active in exactly 56 assays. Data set 8 forms the basis of the "promiscuity strength index" described below.

Determining Compound Promiscuity. One approach to defining promiscuity is to examine active compounds from the primary (initial) HTS data. For this high level view, compounds have been classified into one of three categories: inactive, active and promiscuous. We consider a compound inactive if it has been tested in at least 15 assays and was inactive in all of them. This ensures that a compound has seen a reasonable number of assays [>4% of total] to be considered inactive. More recently registered compounds

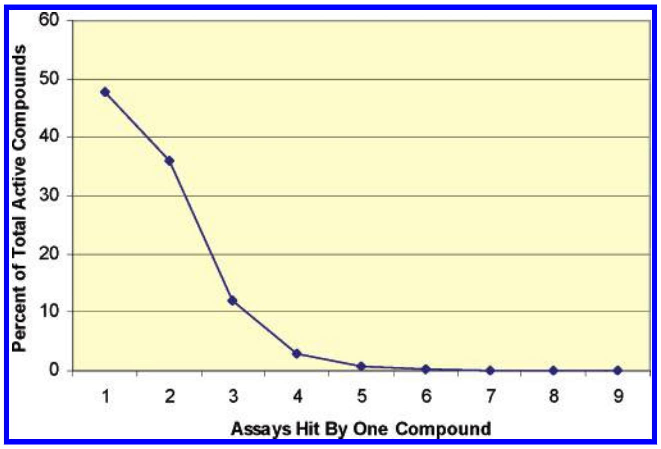


Figure 1. Compound promiscuity as measured by assay hit rate from primary HTS.

have been excluded from the inactive list since these were tested in less than 15 assays. The results of these criteria applied to all active compounds are shown in Figure 1. It is clear from these data that the number of compounds active across multiple assays falls rapidly as the number of assays increases. For all active compounds in primary HTS screens, approximately 48% of compounds are active in exactly one assay but only 0.8% are active in exactly five assays.

For these studies, a compound was classified as active if it were active in at least one but less than 7 assays. Based on the data shown in Figure 1 we empirically chose to define compounds active in 7 or more assays (0.1% of total active compounds) as promiscuous.

Promiscuity Ratio Index (PRI). The percentage of nonflagged HTS primary actives for each lot ID was calculated from data set 3 (Scheme 1). The resulting mean value of 1.925% was used to compare the activity of compounds flagged by the FG filters to those not flagged. For each FG filter the mean percent actives was calculated along with the standard error and variance from data set 4. A "promiscuity ratio index" or PRI was calculated for each FG filter by dividing the mean percent actives by the nonflagged HTS mean, which in this case equals 1.925. For each FG filter the PRI value relates the found actives to expected activity based on the benchmark HTS data and thus provides a measure of promiscuity. A PRI < 1 indicates fewer actives than expected, and a PRI > 1 indicates more actives than expected. The statistical significance of each PRI value was evaluated by constructing 95% confidence intervals for each ratio. A confidence interval containing one indicates there is no difference from chance in comparison to the nonflagged HTS primary actives. Confidence bounds that lie entirely below one indicate a hit rate significantly less than the nonflagged group, while confidence intervals entirely above one indicate a hit rate significantly greater than the nonflagged group. The confidence intervals were constructed using percentiles of the normal distribution and a Taylor approximation to the variance of a function of random variables.[16] These statistical correlations provided three rule classes: significantly less than HTS (less), no different from HTS (same), and significantly greater than HTS (greater). The full datasheet is included as Supporting Information (Table 2).

Promiscuity Strength Index (PSI). An additional ranking of promiscuity of compounds flagged by a FG filter was obtained from the number of distinct compounds exhibiting activity across an increasing assay count. The PSI data were tabulated for each FG filter. This is included as Supporting Information (Table 3). Listed in Table 3 by FG filter are the associated number of active compounds, the PSI value, standard error, *t* value, probability function, and rule class. The PSI value is calculated from data set 8 (Scheme 1) and is the mean of the number of hits for each compound flagged by the particular FG filter. The PSI value is 3.29 for HTS benchmark. A FG filter that flags many compounds across a high number of assay counts will have a higher mean probability than HTS benchmark and vice versa.

The *t*-value is the standard, independent samples' *t* statistic, as described by Urdan,[17] and the *p*-value is the associated probability level. The *t* statistic can be thought of as a signal-to-noise ratio. The signal is the difference between FG filter and HTS benchmark means. The estimate of noise is derived from the variation of the data around the means. The *p*-value provides a measure of how likely the associated *t*-value is produced under pure chance conditions. That is, the *p*-value is the probability of observing a *t*-value as extreme as, or more extreme than, the one obtained under the assumption that there is no difference between the FG filter and HTS data. When the FG filter mean is less than the HTS mean, and the *p*-value is less than 0.05, the rule classification is designated as less. When the FG filter mean is greater than the HTS benchmark mean, and the *p*-value is less than 0.05, the rule classification is designated as greater. When the *p*-value is greater than 0.05, the rule classification is designated as same. This provides a complementary set of three rule classes for the PSI index.

Structural Integrity (SI) Data. The criteria for a compound passing SI analysis are (1) having a major peak greater than 75% of total integrated UV peak area by LC and (2) having the expected molecular weight confirmed by MS.[18,19] Structural integrity data have been accumulated for a relatively small portion of the screening deck, and, as such, SI data were sparse for a number of compounds flagged by the FG filters. A set of FG filters that flagged potentially unstable compounds in a DMSO storage environment were selected and used to query SI data. Additional compound sets, flagged by the unstable FG queries, were selected at random to help fill gaps in the SI data. In some cases there were still fewer than 10 compounds available for SI analysis. The previously analyzed data set and a set of approximately 900 new compounds submitted for SI analysis were combined and the results are included in Table 4 as Supporting Information. As a benchmark, nonflagged FG SI data have an overall pass rate of approximately 73%.

RESULTS

In the methods section, it was shown that the number of individual compounds active across multiple assays falls rapidly as the number of assays increases (Figure 1). Inactive, active, and promiscuous classifications were defined. Compound promiscuity was defined as a single compound active in 7 or more HTS assays. Researchers at Roche have defined an activity cutoff of 8 assays as their definition for a promiscuous or a "frequent hitter" compound.[20] The high level trends remain the same within a range of assay cutoffs defining promiscuous hits. However defined, the percentage

Table 1. (a) Assay Data and Filter Flagging Statistics for Molecules in Each HTS Promiscuity Classification[a] and (b) Breakdown of Calculated Property Filter Data of Molecules in Each Promiscuity Classification[a]

(a)

HTS classification	mean assays per compound	% of assays active in	% property flagged	% functional group flagged
inactive	33 (14)	0	3.3	4.3
active	80 (42)	3.6 (1.9)	4.8	5.5
promiscuous	161 (62)	7.3 (3.5)	4.4	10.9

(b)

HTS classification	mean aLogP	mean MW	mean HBA	mean HBD	mean solubility	mean surface area
inactive	3.3 (2.0)	398 (131)	4.6 (2.1)	1.4 (1.2)	−5.4 (2.1)	377 (121)
active	3.2 (2.1)	412 (140)	4.8 (2.3)	1.6 (1.3)	−5.5 (2.3)	391 (128)
promiscuous	3.0 (2.3)	347 (131)	4.3 (2.6)	1.4 (1.4)	−4.9 (2.5)	329 (120)

[a] Values in parentheses represent the standard deviation.

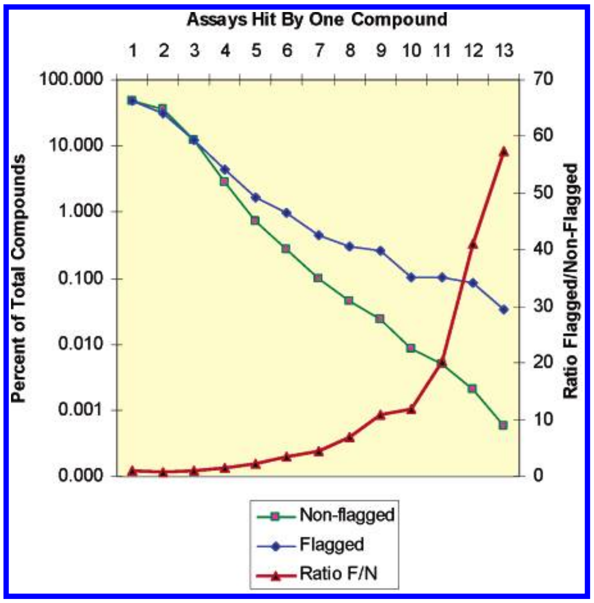


Figure 2. Assay hit rates of FG filter flagged vs non-flagged compounds and their ratio.

of compounds classified as promiscuous is small, but their representation within the group of active compounds, as will be shown, is significant. Information was gathered by making comparisons across activity classifications as shown in Table 1a .

The resultant compound clusters defined in Table 1a were then analyzed using our property and FG filters. An overall breakdown by calculated property filter data is shown in Table 1b.

For the second phase of this analysis we examined in more detail our current FG filters as they relate to HTS primary hit rate promiscuity. A log plot of the percentage of FG flagged and nonflagged compounds by number of assays hit is shown on the left *y*-axis of Figure 2. The ratio of flagged to nonflagged compounds as a function of the number of assays hit by the same compound is shown on the right *y*-axis.

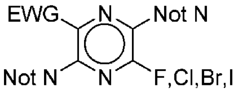
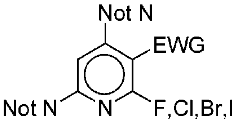
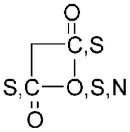
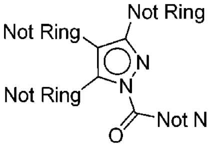
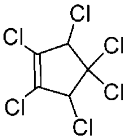
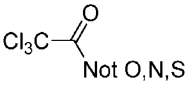
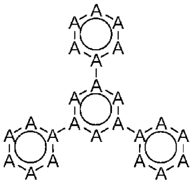
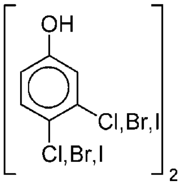
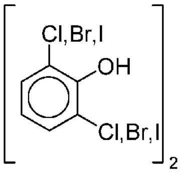
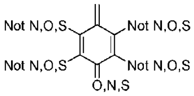
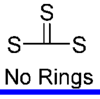
FG filters have also been analyzed systematically to determine “promiscuity ratio index” (PRI) and “promiscuity strength index” (PSI) behavior (Tables 2 and 3 respectively, Supporting Information). A summary of the FG filter linkages to overall promiscuity is given in Table 4 (Supporting Information), which highlights the particular FG filter, the PRI statistical classification, the PSI statistical classification, the SI pass rate, and the promiscuity linkage rating. Compound structural integrity (SI) data are included in this analysis and were enhanced with additional data where possible. Summarized in Table 5 are six examples each of extreme ranges of the PRI and PSI rule classifications along with the query structure to better visualize the FG filter. A rigorous translation of the SMARTS query into a visual representation may not be possible in terms that are universally accepted. The visual representations are not strictly Markush or ISIS queries. The SMARTS query code has been included in the Supporting Information to provide an unambiguous definition source.

A control study for the PRI and PSI indices was also undertaken to test these methods abilities to separate problematic substructures from those well represented in known clinical agents. PRI and PSI results for these control examples are shown in Table 6.

DISCUSSION

Compound functional group and property filters are routinely employed by drug discovery screening operations to remove compounds from screening collections. This process helps reduce false positives, improve the tractability of screening hits, and reduce screening costs. While numerous papers have discussed the merits of this approach, there has been a paucity of hard data correlating promiscuous screening hits to a compound’s characteristics. Shoichet and co-workers have described an interesting correlation of assay hit promiscuity with compound aggregation,[21,22] and some recent success has been obtained using recursive partitioning models to predict which compounds are likely to form aggregates.[23,24] However, one is limited by the breadth and quality of the training sets using this approach, and aggregation is but one possible mechanism for promiscuity. A number of groups have used in-house HTS data in an effort to improve this situation. Schneider et al. evaluated frequent hitters from HTS using a set of 345 different descriptors including properties.[20] Their most predictive frequent hitter models were neural nets based on Ghose and Crippen descriptors, which are atom classifications developed for measuring contributions of hydrophobicity and molar refractivity. The Roche researchers were unable to demonstrate that a meaningful substructure was common to HTS promiscuity. However, using the software Leadscope[25] they did

Table 5. Weak and Strong Linkage Examples to FG Filters[a]

Functional Group Filter	Distinct Cmpds	Query Structure	PRI	PRI Class	PSI	PSI Class	% Pass SI	Linkage
2halo pyrazine 5EWG	78		0.11	Less	1.00	Same	0	LOW
2halo pyridine 3EWG	890		0.82	Less	2.72	Less	56	LOW
Activated 4mem ring	28		0.15	Less	5.00	Same	0	LOW
Acyl pyrazole	249		1.08	Same	2.76	Less	14	LOW
Perchloro cp	22		1.3	Same	2.24	Same		MED
Trichloromethyl ketone	67		0.9	Same	3.14	Same		MED
Branched polycyclic aromatic	729		3.73	Greater	9.49	Greater	50	HIGH
Gte 5 phenolic OH	66	≥ 5 phenolic OH anywhere in molecule	3.77	Greater	9.02	Greater	17	HIGH
Polyhalo phenol c	14		5.95	Greater	11.92	Greater	100	HIGH
Polyhalo phenol d	40		9.84	Greater	23.30	Greater	25	HIGH
Quinone methide	160		6.91	Greater	9.10	Greater	29	HIGH
Thio xanthate	52		4.68	Greater	14.78	Greater	33	HIGH

[a] EWG = sulfonyl, trifluoromethyl, cyano, nitro, carbonyl.

find fragments that correlated with higher hit rates. A recent publication by Diller and Hobbs discusses the linkage of HTS results to several property descriptors and to 17 basic substructures such as cyano, ether, and tertiary aniline.[26] Due to inherent compound redundancy with the screening of approximately 200 million ECLiPS combinatorial library compounds, a statistical model was built to identify and remove false positives. The authors were able to determine

Table 6. Promiscuity Measures for Control Functional Group Filters

functional group filter	active compds	sum of total assays	PRI	PRI class	PSI	PSI class	% pass SI	compds tested SI	linkage
all HTS data	NA	NA	1.00	NA	3.29	NA	73	NA	NA
diphenylmethane	61 959	2 827 462	1.23	greater	3.75	greater	74	2547	high
ethylcarbamate	9750	402 123	0.83	less	3.81	greater	79	271	med
hydantoin	2456	145 135	0.91	less	4.18	greater	83	425	med
indole	37 936	1 195 077	0.83	less	3.88	greater	80	1447	med
isoquinoline	1698	62 692	0.74	less	3.52	same	92	702	low
phenethylamine	31 391	1 415 404	1.43	greater	3.20	less	86	2897	med
SI > 75% purity	NA	NA	1.47	greater	4.50	greater	100	NA	high
SI < 50% purity	NA	NA	2.05	greater	5.10	greater	0	NA	high

the probability of a flagged set of compounds as being active in HTS. Their results indicate larger and more lipophilic molecules as more likely to be strongly active, which makes sense from a potency standpoint as mentioned by the authors. One needs to distinguish between activity (potency) and promiscuity (selectivity), which was not evaluated in their study. Given the nature of the compound sets, they did not address offending functional groups, nor were structural integrity issues discussed. Compound insolubility could contribute to an inactive HTS status but was not assessed in their study or this current analysis.

High-Level View. We have attempted to further address these issues though the studies described above. Table 1a shows that, as expected, the mean number of assays seen per compound increases from inactive to promiscuous classifications, weighting promiscuity toward the more venerable compounds in the data set. The screening data associated with older compounds escaped filters that are currently employed, and one would thus envisage that they would exhibit greater promiscuity. As expected, the percentage of assays a given compound was found to be active in is greatest for the promiscuous classification. The percentage of compounds flagged by the property and FG filters both show a slight increase when comparing the inactive and active classification. The two are found to diverge, however, within the promiscuous category. In this high level view, property-based filters show no association with promiscuous compounds, while the FG filters show strong associations.

Table 1b provides a more detailed examination of the property filter results. An analysis of the data contained within the table shows a weak trend for promiscuous compounds to be, on average, smaller and more soluble. This might be considered consistent with the work of Hann et al., which showed that the probability of a successful binding event decreases exponentially with increasing molecular complexity.[27] The trend is only slight, however, and its relevance is open to question. Contrasting with the property group filters, FG filter hit frequency did trend higher with increasing compound promiscuity. This is further illustrated by looking again at HTS active hits, which were annotated as either flagged (hit a FG filter) or nonflagged of Figure 2. A fairly dramatic trend emerges wherein compounds flagged by the FG filters exhibit an increasing proportion of the promiscuity pool.

FG Filter PSI and PRI Indices. To develop a better sense of how well the FG filters were performing in terms of flagging promiscuous HTS hits, primary HTS data were examined in some detail in the context of our PRI and PSI indices. The PRI provides a measurement of relative HTS assay hit frequency. The PRI data results were statistically analyzed, and where confidence bounds could be calculated, a determination of statistical significance from mean HTS hit rates could be made (Supporting Information, Table 2). As expected, FG filters that did not flag many compounds have wide confidence intervals, and those that flagged many have narrow confidence intervals. For example the pentafluorophenylester filter, having only 9 examples, has a low PRI (0.37) but is not statistically different from the HTS benchmark (1.0). In contrast the 2halo_pyridine_3EWG, having 890 examples, exhibits a higher PRI (0.82) which is significantly less than the HTS benchmark. The PSI measures the tendency of a particular flagged compound to remain active across an increasing assay count (Supporting Information, Table 3). The PSI is reported as a value ranging from 1 to approximately 23. A value of one indicates a mean probability of only one active assay per distinct compound and is the lowest measure of activity. For the nonflagged HTS compounds the PSI value is 3.29 and establishes a benchmark for comparison. Nitrosamine is a FG filter example having a low PSI of 2.12, where of the 133 compounds flagged that are active in at least one assay, only 9 remain active across 5 or more assays and one compound is active in 11 or more assays. On the other hand, of the 45 compounds flagged by the thio_xanthate FG filter that are active in at least one assay, fully 32 remain active across 11 or more assays. The thio_xanthate FG filter has a PSI of 14.8 and is considerably greater than the HTS benchmark. A PSI value this high indicates a strong level of promiscuity.

Classification Scheme. Based on these data, a simple classification scheme linking the FG filter to the PRI and PSI has been developed using a sliding scale of high, medium, and low. The ratings depend on the PRI and PSI rule classifications, which are both statistically determined. If both indicators have greater rule classes or if one rule class is greater and one is the same as the HTS benchmark (as defined by its associated confidence intervals), the FG filter is given a high linkage to promiscuity rating. A medium linkage rating is given if both rule classes are the same or if one is greater and the other less. A low linkage rating is given if both rule classes are less or if one is the same and the other less. For the FG filters used in this analysis, 64% were designated as having a high linkage, 18% medium, and 9% having a low linkage to promiscuity. Fourteen filters (8%) have not been rated since either no compounds or too few compounds were flagged in our HTS screening set. However rare, these unrated filters are still potentially applicable to filtering out acquisition compounds, combinatorial library compounds, and other new additions to compound collections.

Structural Integrity (SI). An analysis of compound structural integrity (SI) highlights that the connectivity between a FG filter flag and screening data is highly dependent on the integrity of the sample. We are unaware of a publication relating structural integrity to molecular reactivity patterns. Although a comprehensive study is currently underway as part of the COMDECOM consortium,[28,29] it is anticipated that from the COMDECOM analysis compound stability models will be developed to help anticipate problem compounds. Several researchers have investigated general factors contributing to compound stability in DMSO solution mimicking HTS storage conditions.[30−33] The scope of this manuscript can only be considered a survey of the set of compounds defined by the filters. More importantly, it does not take into account differences between actual SI at the time the assay was run versus when the SI analysis was performed. This is particularly true for older screening data. As such, the SI data should only be considered a rough confirmation of anticipated compound stability patterns. For example, as expected none of the 11 carbonyl_halide (e.g. acid chloride) compounds tested passed SI.

Promiscuity Classification Examples. Starting with examples of weakly linked FG filters shown in Table 5, 2halo_pyrazine_5EWG had 78 examples of which 77 have the same core structure, a 3-substituted-2-chloro-5,6-dicyanopyrazine. This is an activated heterocycle which, along with the limited SI data, suggests that it may be unstable in DMSO solution. In contrast, the related 2halo_pyrazine_3EWG has 15 diverse examples with a high SI pass rate and an overall medium linkage to promiscuity. These examples illustrate that it is important to view the results in the context of physical nature of the compounds flagged. A different set of 2halo_pyrazine_5EWG examples may provide differing linkage results. Overall, the eight instances of halo substituted π-deficient heterocycles do not exhibit a high level of promiscuity. Another case in point is the activated acetylene filter (Supporting Information). This FG filter showed a low linkage to promiscuity. Experience has taught synthetic organic chemists of its electrophilic nature. In fact, 90% of these examples are acetylene carboxamides, which would have a lower reactivity toward biological nucleophiles. If these are excluded, the PRI doubles, and the PSI goes to a greater classification, indicating a high linkage to promiscuity, highlighting the need for a coding change to the activated_acetylene SMARTS. Compounds containing the perchloro_cp resemble the insecticide chlordane and, while unattractive from a tractability standpoint, are not particularly promiscuous. The reversible electrophiles, boronate ester, α-keto-π-deficient heterocycle, and trichloromethyl ketone (Table 5), all exhibit low to medium promiscuity linkages. Trifluoromethyl ketone, aldehyde, and alpha_dicarbonyl FG filters show higher levels of promiscuity but are at the lower end of the high ranking. Reversible electrophiles are known to be time-dependent inhibitors of certain enzymes such as serine proteases, and they can be specific in their binding with biological substrates.[34−38] The acyl_pyrazole example is one of several where an acyl moiety is bonded to a moderately good leaving group and is therefore a potential acylating agent. Related acylating agents such as acyl_123_triazole, acyl_134_triazole, and acyl_imidazole do not show high promiscuity. A broader survey of all acylating agents by classification shows mixed levels of promiscuity. The mean SI pass rate of all acylating agents identified by the FG filters is less than half the mean of the total, indicating instability of this class of compounds. Acylating agents that are stable enough to survive wet DMSO may nevertheless be capable of derivatizing biological nucleophiles. Six examples of FG filters that are strongly linked to promiscuity (Table 5) all have PRI and PSI values significantly above the HTS benchmark. In part, these filters have been designed based on years of medicinal chemistry experience to flag compounds with tractability issues. Many of the highly promiscuous compounds flagged by the FG filters are undesirable drug leads. Some of these compounds may exert their promiscuity by aggregation or other mechanisms. All of the polyhalophenol templates show high promiscuity. Experiences at BMS have demonstrated that polyhalophenols show consistently high assay hit rates but are typically devoid of meaningful structure−activity relationships. The related branched polycyclic aromatics and compounds containing greater than 5 phenolic groups behave similarly. Quinone methide is an example of a reasonably stable Michael acceptor that is quite promiscuous in many assays suggesting a nonspecific binding event to a biological nucleophile.

Controls. The PRI/PSI control examples shown in Table 6, based on their FG queries, do not show promiscuous linkages to the HTS data used in this analysis. These results provide validation of the methods used to define compound promiscuity in this study. Several substructures that are well represented in known clinical agents were chosen. These include diphenylmethane, ethyl carbamate, hydantoin, indole, isoquinoline, and phenethylamine. The diphenylmethane moiety is prevalent in many drugs on the market and as such may be considered a promiscuous scaffold.[39−41] Ethyl carbamate, hydantoin, and isoquinoline scaffolds would not be expected to be promiscuous in screening. Some scaffolds, such as indole and phenethylamine, might be considered GPCR-like privileged substructures, and compounds containing such privileged substructures may be pharmacologically promiscuous across multiple but related assays.[42−44] Promiscuity could be better understood by looking at compound promiscuity within target classes or between target classes and by looking at assay type such as biochemical versus whole cell. However, this is beyond the scope of this study. In addition to molecular feature controls, we looked at compounds grouped within the structural integrity classifications of highly pure (>75% purity) and impure (<50% purity). Since most of the SI data were generated at the request of program chemists, it is not surprising that PRI and PSI are at the low end of a high linkage classification. Impure compounds show increased levels of promiscuity. However, this does not fully account for the strongly observed promiscuity associated with some FG filters. It makes sense that less pure compounds would show a higher level of promiscuity since degradation products such as fluorescent materials could interfere with assay readouts.

The FG filters applied within this study only flag approximately 15% of the promiscuous compounds (active ≥ 7 assays) from HTS screens. If one looks at nonflagged compounds showing activity in 11 or more assays, about 16K compounds remain. An inspection of the most active 100 of these show questionable compounds that appear to have fluorescent chromophores or that may aggregate in a fashion described by Shoichet.[21] Within the entire set of 16K compounds there are a number of prominent functional group and scaffold arrangements that were missed by the filters. Rebuilding an extensive list of SMARTS based on these missed compounds is an ongoing effort.

CONCLUSIONS

An evaluation of primary HTS data carried out at BMS over the last 12 years was analyzed for promiscuity correlations to a set of compound functional group and property filters. Lipinski-type property descriptors did not show significant correlations to screening deck compound promiscuity. While widely used in drug discovery organizations, property filters seem better used as part of an HTS triage carried out at a later stage rather than a strict application to primary screening data. However, it is still useful to implement property filters as part of library design strategies and compound acquisitions in order to maintain higher compound tractability. The exact nature of such compound property cutoffs remain a matter of debate within the scientific community.[6,45,46]

In contrast to property filters, much stronger promiscuity correlations were found with a number of SMARTS-based FG filters. To aid in the application of such filters, two assessments of promiscuity have been developed. The PRI measures an observed/expected FG filter hit rate, and the PSI measures the strength of that hit rate across multiple assays. These were combined with a statistical analysis of in-house assay data to provide confidence intervals that link the filters to a promiscuity category (high, medium, and low). These designed FG filters form the basis for compound annotation and exclusion lists. SMARTS-based FG filter controls representing common molecular features were not associated with promiscuity and provided validation to the methods employed. Functional group filter impact depends on the number of compounds flagged and their relative promiscuity in HTS. The removal of all compounds flagged by the high linkage FG filters from our screening data resulted in a 12% reduction in compounds considered for triage, a significant time saving impact for hit-to-lead chemistry teams. Since filter design is an iterative process, it is anticipated that more compound characteristics leading to false positives will be identified. It has been shown that molecules having structural similarity tend to have similar biological properties.[47] Similarly, the data we have presented supports the hypothesis that molecules having a common functional group, mediating nondiscriminate binding to biological substrates, will tend to behave promiscuously as a group in HTS screening campaigns.

ACKNOWLEDGMENT

The authors wish to acknowledge Brett Beno for helpful discussions and translation of an initial set of filters based on SLNs to SMARTS. The authors also express acknowledgment to Martyn Banks, Mark Hermsmeier, Debbie Loughney, Nelly Masias, Nick Meanwell, Ramesh Padmanabha, Julie Pearce, and Dora Schnur for helpful discussions. Members of the Scitegic Help team for their excellent support.

Supporting Information Available: Statistical analysis of the promiscuity ratio index (PRI) (Table 2), statistical analysis of the promiscuity strength index (PSI) (Table 3), summary table of functional group query promiscuity linkages (Table 4), and functional group SMARTS. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Walters, W. P.; Namchuk, M. A guide to drug discovery: Designing screens: how to make your hits a hit. *Nat. Rev. Drug Discovery* **2003**, *2* (4), 259−266.

(2) Rishton, G. M. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discovery Today* **2002**, *8* (2), 86−96.

(3) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of extremely noisy high-throughput screening data using a naive Bayes classifier. *J. Biomol. Screening* **2004**, *9* (1), 32−36.

(4) 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* (1−3), 3−25.

(5) Blower, P. E., Jr.; Cross, K. P.; Fligner, M. A.; Myatt, G. J.; Verducci, J. S.; Yang, C. Systematic analysis of large screening sets in drug discovery. *Curr. Drug Discovery Technol.* **2004**, *1* (1), 37−47.

(6) Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E. Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totaling 2.7 Million Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (2), 643−651.

(7) Walters, W. P.; Murcko, M. A. Prediction of ‘drug-likeness’. *Adv. Drug Delivery Rev.* **2002**, *54* (3), 255−271.

(8) 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* (5), 897−902.

(9) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds. *J. Med. Chem.* **2004**, *47* (20), 4891−4896.

(10) Presented in part, at the 229th American Chemical Society National Meeting, San Diego, CA, March 2005; CINF-085.

(11) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A Comparison of Physiochemical Property Profiles of Development and Marketed Oral Drugs. *J. Med. Chem.* **2003**, *46* (7), 1250−1256.

(12) *Daylight Chemical Information Systems*; Mission Viejo Inc.: CA. http://www.daylight.com.

(13) *Pipeline Pilot Version 4.5.1.0*; Scitegic: San Diego, CA. http://www.scitegic.com.

(14) *Perl Version 5.8.1*; http://www.perl.com.

(15) *JMP Version 5.0.1a*; SAS: Cary, NC. http://www.jmp.com.

(16) Meyer, P. *Introductory Probability and Statistical Applications*; Addison-Wesley Publishing Company: Reading, MA, 1970.

(17) Urdan, T. C. *Statistics in Plain English*; Lawrence Erlbaum Association: London, 2001.

(18) Compounds are individually submitted to the LC−UV/MS based structural integrity (SI) assay as samples of 10 μL, 3 mM solutions in DMSO formatted in shallow 96-well-plates. The samples are diluted and mixed 1:14 with 130 μL of a solvent cocktail of 2-propanol/water/acetonitrile (33%:33%:33%, v:v:v) utilizing a TECAN liquid handler. The solvent mixture is used to ensure solubility of the compounds in LC and MS compatible solvents. The samples are then analyzed by rapid gradient in-line LC−UV/MS utilizing UV detection at 220 nm, and ion-trap mass spectrometers employing positive-negative-switching with data-dependent full scan MS/MS methods in electrospray ionization mode (ESI), and if necessary, also in atmospheric pressure chemical ionization mode (APCI). The acquired SI data are automatically processed by “intelligent” in-house software developed in Visual Basic and ThermoElectron Finnigan Xcalibur Developers Toolkit (XDK).

(19) Josephs, J. L.; Sanders, M.; Langish, R. A.; Hnatyshyn, S. Y.; Salyan, M. E.; Shipkova, P.; Drexler, D.; Flynn, M. J.; Burdette, H. L.; Balimane, P.; Zvyaga, T. A.; Chong, S. In *High Quality High Throughput LC/MS Strategy for Profiling of Drug Candidates with Applications to Structural Integrity, Permeability, Stability, and Related Assays*; Proceedings of the 50th ASMS Conference, Orlando, Florida, June 02−06, 2002, A021460.pdf, 2002; 2002.

(20) Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E.-M.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjoegren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von Saal, W.; Zimmermann, G.; Schneider, G. Development of a Virtual Screening Method for Identification of “Frequent Hitters” in Compound Libraries. *J. Med. Chem.* **2002**, *45* (1), 137−142.

(21) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. *J. Med. Chem.* **2002**, *45* (8), 1712−1722.

(22) McGovern, S. L.; Helfand, B. T.; Feng, B.; Shoichet, B. K. A Specific Mechanism of Nonspecific Inhibition. *J. Med. Chem.* **2003**, *46* (20), 4265−4272.

(23) Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and Prediction of Promiscuous Aggregating Inhibitors among Known Drugs. *J. Med. Chem.* **2003**, *46* (21), 4477−4486.

(24) Doman, T. N.; Shoichet, B. K. In *Identification and Prediction of Self-Aggregating Compounds in Drug Discovery Settings*; Abstracts, 36th Central Regional Meeting of the American Chemical Society, Indianapolis, IN, United States, June 2−4, 2004; 2004; pp INV-025.

(25) Leadscope Inc.: Columbus, OH. http://www.leadscope.com.

(26) Diller, D. J.; Hobbs, D. W. Deriving Knowledge through Data Mining High-Throughput Screening Data. *J. Med. Chem.* **2004**, *47* (25), 6373−6383.

(27) 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* (3), 856−864.

(28) Herklots, H. Outsourcing the search for leads. *Modern Drug Discovery* March, 2000; pp 48−50.

(29) COMDECOM information: http://www.specs.net.

(30) Kozikowski, B. A.; Burt, T. M.; Tirey, D. A.; Williams, L. E.; Kuzmak, B. R.; Stanton, D. T.; Morand, K. L.; Nelson, S. L. The effect of freeze/thaw cycles on the stability of compounds in DMSO. *J. Biomol. Screening* **2003**, *8* (2), 210−215.

(31) Kozikowski, B. A.; Burt, T. M.; Tirey, D. A.; Williams, L. E.; Kuzmak, B. R.; Stanton, D. T.; Morand, K. L.; Nelson, S. L. The effect of room-temperature storage on the stability of compounds in DMSO. *J. Biomol. Screening* **2003**, *8* (2), 205−209.

(32) Cheng, X.; Hochlowski, J.; Tang, H.; Hepp, D.; Beckner, C.; Kantor, S.; Schmitt, R. Studies on repository compound stability in DMSO under various conditions. *J. Biomol. Screening* **2003**, *8* (3), 292−304.

(33) Buko, A. M.; Beckner, C.; Hepp, D.; Helias, N.; Zhu, F.; Nemcek, T.; Schmitt, R. J.; Hochlowski, J. In *Stability testing of chemical diversity in liquid DMSO storage*; Abstracts of Papers, 222nd ACS National Meeting, Chicago, IL, United States, August 26−30, 2001, 2001; pp ANYL-210.

(34) Zhong, J.; Groutas, W. C. Recent developments in the design of mechanism-based and alternate substrate inhibitors of serine proteases. *Curr. Top. Med. Chem. (Sharjah, United Arab Emirates)* **2004**, *4* (12), 1203−1216.

(35) Sanderson, P. E. J. Small, noncovalent serine protease inhibitors. *Med. Res. Rev.* **1999**, *19* (2), 179−197.

(36) Ilies, M. A.; Scozzafava, A.; Supuran, C. T. Therapeutic applications of serine protease inhibitors. *Expert Opin. Ther. Pat.* **2002**, *12* (8), 1181−1214.

(37) Mehdi, S. Synthetic and naturally occurring protease inhibitors containing an electrophilic carbonyl group. *Bioorg. Chem.* **1993**, *21* (3), 349−59.

(38) Lin, C.; Lin, K.; Luong, Y.-P.; Rao, B. G.; Wei, Y.-Y.; Brennan, D. L.; Fulghum, J. R.; Hsiao, H.-M.; Ma, S.; Maxwell, J. P.; Cottrell, K. M.; Perni, R. B.; Gates, C. A.; Kwong, A. D. In Vitro Resistance Studies of Hepatitis C Virus Serine Protease Inhibitors, VX-950 and BILN 2061: structural analysis indicates different resistance mechanisms. *J. Biol. Chem.* **2004**, *279* (17), 17508−17514.

(39) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887−2893.

(40) Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipskind, P. A. Characteristic Physical Properties and Structural Fragments of Marketed Oral Drugs. *J. Med. Chem.* **2004**, *47* (1), 224−232.

(41) Hajduk, P. J.; Bures, M.; Praestgaard, J.; Fesik, S. W. Privileged molecules for protein binding identified from NMR-based screening. *J. Med. Chem.* **2000**, *43* (18), 3443−3447.

(42) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification Of Biological Activity Profiles Using Substructural Analysis And Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (2), 165−179.

(43) Schnur, D.; Hermsmeier, M. A. In *Classpharmer and the quest for privileged substructures*; Abstracts of Papers, 228th ACS National Meeting, Philadelphia, PA, United States, August 22−26, 2004, 2004; pp CINF-080.

(44) Hermsmeier, M. A.; Schnur, D.; Pearce, B. C. In *Systematic bioactivity classification of ligands onto a protein target ontology: Application for library design and virtual profiling of a compound collection*; Abstracts of Papers, 227th ACS National Meeting, Anaheim, CA, United States, March 28−April 1, 2004, 2004; pp CINF-014.

(45) Charifson, P. S.; Walters, W. P. Filtering databases and chemical libraries. *Mol. Diversity* **2002**, *5* (4), 185−197.

(46) Lajiness, M. S.; Shanmugasundaram, V. Strategies for the identification and generation of informative compound sets. *Methods Mol. Biol. (Totowa, NJ, United States)* **2004**, *275* (Chemoinformatics), 111−129.

(47) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45* (19), 4350−4358.

CI050504M