JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Chemical and Biological Properties of Frequent Screening Hits

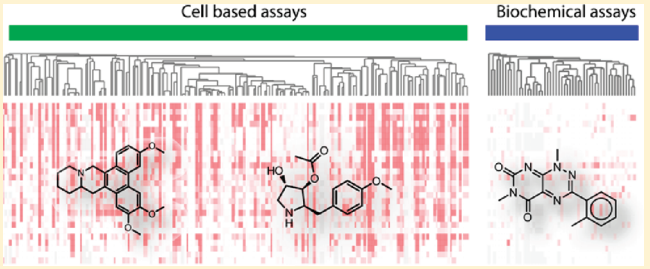
Jianwei Che,*[,†] Frederick J. King,[†,‡] Bin Zhou,[†] and Yingyao Zhou*[,†]

[†]Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121, United States

[‡]Developmental and Molecular Pathways, Novartis Institutes for BioMedical Research, 250 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

Supporting Information

ABSTRACT: High-throughput screening (HTS) has become an important technology for the drug discovery process. It has been noted that certain compounds frequently appear as hits in many screening campaigns. By mining an HTS database covering large chemical space and diverse biological functions, we identified many novel chemical features, as well as several biological processes that were associated with a significant portion of frequent hits. However, we also noted that several marketed drugs also contained characteristics that commonly were associated with frequent hits. This observation suggested that current generally employed strategies for triaging compounds may result in the removal of compounds with desirable properties. Therefore, we developed a novel strategy that overlaid chemical scaffolds and biological processes, along with empirical hit frequency data, in order to provide a more functional frequent hit triage strategy; the risk of removing biologically relevant frequent hits was reduced compared to the typical empirical hit frequency-based filtering strategy.



INTRODUCTION

High-throughput screening (HTS) has been used in the pharmaceutical industry as a key technology for the identification of lead molecules for novel therapeutics.[1] Its versatility also has impacted the academic community by providing a means to identify chemical probes for basic biomedical research as outlined in the NIH roadmap.[2] A phenomenon often associated with HTS is the repetitive identification of "familiar" active molecules, which are often known as "frequent hits" or "promiscuous hits".[3] Frequent HTS hits are defined as active molecules across a large number of statistically independent assays and tend not to be considered as promising leads for drug development. In fact, it is considered desirable to remove frequent hits as early as possible in the lead discovery workflow.

Polypharmacology is associated with a number of successful drug molecules;[4] therefore, high selectivity may not necessarily be a prerequisite or even desirable for an HTS lead.[5] Hu et al. studied the chemical scaffold relationship of ~35 000 bioactive compounds using pooled data set from ChEMBL[6] and BindingDB.[7] They showed that many known drugs demonstrated multitargeting behavior.[8] In our experience, there is often a trade-off between potency and selectivity for early lead candidates, as their suboptimal properties can be improved by medicinal chemistry efforts. Therefore, hit frequency by itself is often insufficient to enable scientists to make a decision on whether it is appropriate to filter out certain frequent hits without compromising the opportunities for discovering good leads.

Characteristics of nonspecific screening hits have been reported previously.[3a,9] Shoichet et al. reported aggregation of particular organic molecules into colloid-like particles as a potential mechanism for nonspecifically inhibiting enzyme activity.[9a,b,d] The rationale for the promiscuity observed was attributed to partial unfolding of the enzyme. Recently, Jadhav et al. identified three modalities of artifacts from a screen for thiol protease inhibitors: aggregation, autofluorescence, and reactivity.[9c] Baell et al. illustrated certain substructure features that contributed to pan assay activity.[9e] Each of these factors has the potential to contribute to a compound's broad spectrum activities. However, all these studies were performed on very specific assay systems, primarily enzymatic assays, and on limited number of compounds. Therefore, it would be extremely valuable to expand these studies toward cell-based assay systems and include a much larger-scale HTS database.

The focus of this study is to gain insights into the fundamental biological driving forces behind frequent hits in more complex cellular systems, which will hopefully provide a more critical approach to protect drug-like frequent hits from being discriminated against by their high hit rates. At the Genomics Institute of the Novartis Research Foundation (GNF), we have performed a large number of cellular and biochemical HTS over the past decade. The screens covered a diverse range of disease areas, biological processes, screening formats, and popular drug target families.[10] Therefore, the accumulated data, offered us a unique opportunity to study frequent hits at not only a large scale, but also a rich biological context that had not been described in previous studies.

The insights gained in this study regarding the characteristics associated with frequent hits offer practical guidance for an improved HTS hit triage process. As illustrated in Figure 1, we

Received: January 3, 2012
Published: March 21, 2012

   dx.doi.org/10.1021/ci300005y | J. Chem. Inf. Model. 2012, 52, 913−926

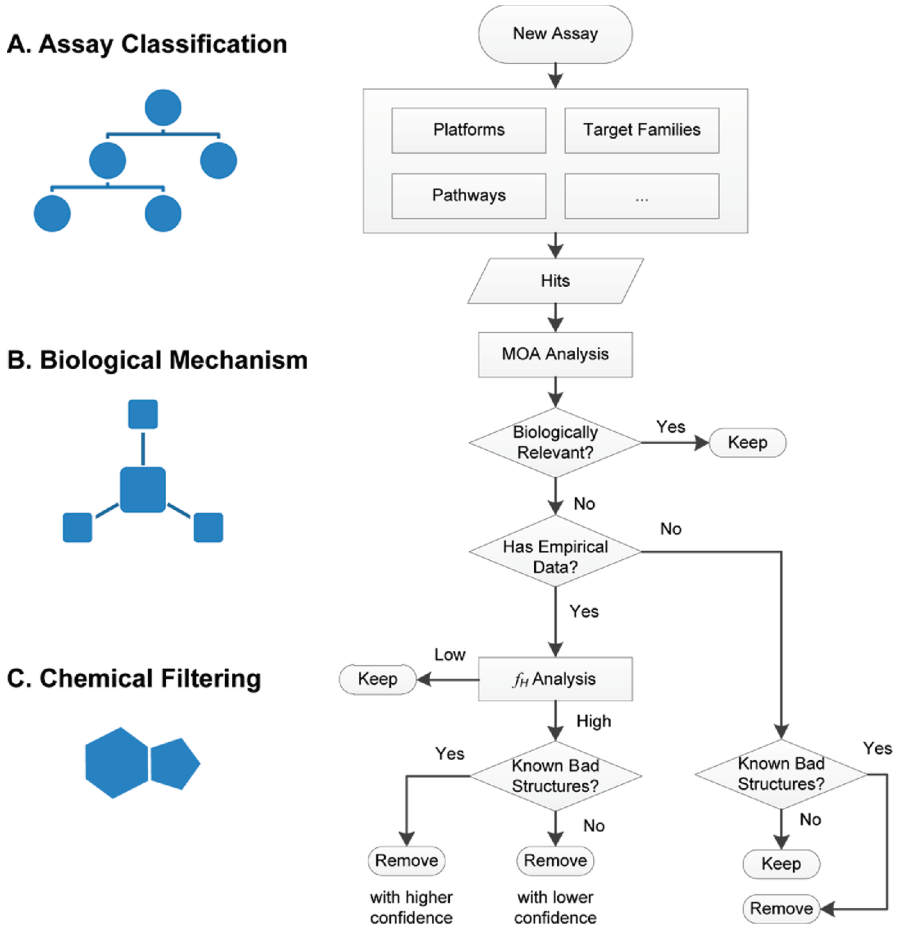


Figure 1. Workflow of the "ABC" HTS frequent hits triaging strategy.

propose a process called "ABC", where "A" is for *a*ssay classification, "B" is for *b*iological mechanisms, and "C" is for *c*hemical filtering. Assay classification refers to whether the screen was cell-based or biochemical, along with further categorization in terms of target class and readouts. As one can imagine, frequent hits for kinases most likely are not frequent hits for proteases or G protein-coupled receptors (GPCRs). Similarly, screens with fluorescent readouts can have very different frequent hits with respect to luciferase-based screens. Therefore, the HTS data used to identify frequent hits need to be relevant to one another. Step B addresses the potential biological mechanism of actions (MOA) that may impart the broad activity observed. This step is intended to provide crucial information on the biological properties of the hits, so that frequent hits are retained as long as their MOA directly connects to the biological function of interest. The chemical filtering (step C) further identifies frequent hits based on prevalent "promiscuous" chemical structural features derived from in-house or publicly available biological data sets. Traditionally, only step C has been applied as the criterion for hit triage, i.e., the percentage of total assays where a compound is active. Our study shows step A and B play as an important role as C, and all three should be combined in any successful HTS hit triaging approach.

In addition to the new ABC hit triage strategy proposed here, it is to our knowledge that similar comprehensive large-scale study on HTS frequent hits has not been reported before, and the informatics analysis methods described here could also provide examples of mining large compound-assay matrices.

MATERIALS AND METHODS

GNF Databases. The HTS database consists of a matrix of 2 942 760 unique small molecule structures across 277 assays that have been accumulated in the GNF Lead Discovery Database (LDDB) over the past ten years. The screens contain 55 biochemical assays and 222 cellular assays. The matrix is 29% dense, due to changes of the composition of the screening libraries and varying QC filters applied over the period. However, each compound structure used in this study had been screened in at least ten different assays. Activity data from each assay were normalized based on activity rank, and the resultant *Z*-scores were used in the activity matrix. Among the approximately three million compounds, there were 2783 known drugs that were used in the comparative analyses.

We also extensively collected publically available biological annotations to help interpret the biological properties of the compounds. The integrated annotation database[11] covers data sources, including MeSH,[12] GeneGo,[13] and PubMed.[14] The compound–protein interaction data were mainly derived from GeneGo.

Hit Frequency f_H. Here, we defined hit frequency of a compound to be

$$f_H = \begin{cases} \dfrac{N_{hit}}{N_{tot}}, & \text{if } N_{tot} \geq 10 \text{ and } N_{hit} \geq n_0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

dx.doi.org/10.1021/ci300005y | *J. Chem. Inf. Model.* 2012, 52, 913–926

where N_{hit} is the number of assays in which the molecule is considered as a hit, and N_{tot} is the total number of assays, in which the molecule has been tested. We set $f_H = 0$, if the compound is active in less than n_0 assays or tested in less than 10 assays. When all assays or all cellular assays were analyzed, we set n_0 to be 5. When all biochemical assays were analyzed, we chose n_0 to be 3.

Usually, the determination of whether a molecule is designated as a hit is a reflection of the compound's percentile rank and not a particular activity. The typical HTS hit rate that is used ranges from 0.1 to 1% of the total compounds screened. For this reason when a single activity cutoff was required, we defined the most active 0.1% of the total compounds screened as hits (i.e., $Z > 3.09$). To make sure that our conclusions were insensitive to this subjective cutoff under such circumstance, we have also repeated all the calculations using a hit rate of 0.5%.

At GNF, although the median compound activity for a screen is always normalized to 1.0, the activity data still heavily depends on the varying assay sensitivities. This means wide signal ranges for the HTS hits can exist. For example, in inhibition assays, a hit rate of 0.1% could be equivalent to an activity as low as 50% in some assays, but as high as 95% in others. Therefore, only quantile-normalized Z scores were used for hit determination in our study.

Compound Clustering. To characterize molecular structural relationships, we first constructed an all-by-all distance matrix using the Tanimoto similarities of their ChemAxon 2D topological fingerprints.[15] Then a standard hierarchical clustering algorithm was applied. The resultant hierarchical tree was used to construct disjointed clusters by cutting the tree at a 0.85 similarity threshold. Among the entire compound set of three million molecules, most were never active in any assays. Therefore, it was unnecessary to cluster and analyze the complete collection of compounds. We first ranked all compounds based on their f_H values and kept the top 20 000 molecules. Toward the bottom of the list of these 20 000 molecules, the f_H values were zero or near zero, which ensured that there was enough background nonfrequent hits to enable the identification of molecules with abnormally high f_H. This process was repeated for all the 277 assays, separately for the subset of biochemical assays, and for the subset of cellular assays.

Ontology-Based Pattern Identification (OPI) Algorithm. Instead of arbitrarily setting an f_H threshold (e.g., 0.1), we applied an OPI algorithm[16] to identify frequent hits from various HTS assay sets, either by a scaffold-driven or an MOA-driven approach. The detailed steps of the algorithm were described in a review,[17] and it has prior applications to large HTS data sets.[10,18] Ontology-based pattern identification (OPI) algorithm provides a nonheuristic alternative to identify frequent hits, which minimizes the false positives and only identifies compounds whose chemotypes or MOAs were robustly associated with high-frequency hit patterns. We outline the key steps of the algorithm that were employed in this study below.

In a scaffold-driven OPI analysis, compounds first were clustered into chemical scaffold families whose members shared a structure similarity greater than a Tanimoto value of 0.85. Second, compounds were ranked according to their f_H values in descending order, so frequent hits were populated toward the top of the list and nonhits were at the bottom. Third, for each given compound family, the OPI algorithm determined a family specific f_H cutoff, so that the subset of molecules with f_H values above the cutoff were most statistically distant from the remaining members in terms of their locations on the sorted list. The cutoff values are compound family specific, because the f_H distributions are family specific. The process of determining an f_H cutoff value is conceptually similar to clustering each compound family member into two groups, such as k-means, based on their f_H values, except OPI made use of an iterative statistical optimization process to make sure the probability (p-value) of the grouping due to random noise was minimized. Within each family, the molecules with f_H above the optimal cutoff were then labeled as frequent hits.

The application of OPI enabled dynamically identifying molecules for each cluster, whose f_H were abnormally high within the family and across other families measured by a statistically significant p-value. The main reason the OPI algorithm has worked well in similar cheminformatics applications is that it is able to boost the signal-to-noise ratio by taking advantage of the principle of structure−activity relationship,[19] i.e., a chemical scaffold assigned a lower p-value is much more likely to be a true frequent hit when multiple structurally similar compounds also share high f_H values. In contrast, if only a small fraction of members in a scaffold family have high f_H values, less significant p-values would be assigned because the isolated incidences of high f_H could be caused by artifacts such as compound impurities.

Similar to the chemical scaffold-driven OPI analysis described above, we attempted to determine the identity of the common MOAs associated with frequent hits using a MOA-driven OPI analysis. 4353 structures in the original data matrix have available annotations covering 294 MOA categories defined by the MeSH[12] ontology database. We applied the same OPI algorithm to identify MOA groups associated with high f_H values. The only difference here was that the 294 MOA groups substituted the structural information used in the chemical analysis described above. All three million compounds were sorted by their descending f_H values first; then the OPI algorithm was iteratively applied to each MOA group to identify the statistically optimal subset of MOA members as frequent hits. For the result, p-values were assigned to those MOA groups, where multiple members shared significantly higher f_H values compared to the remaining compounds. As a molecule could have multiple MeSH annotations and each MOA group was analyzed independently, a compound could result in as a frequent hit for multiple mechanisms.

PubChem Data Set. Screening data for a total of 427 primary assays, which included 54 cell-free and 373 cell-based assays based on the BioAssay Ontology study[20] and text searches in PubChem, were retrieved from the public domain. The data matrix was 17% dense, compared to 29% in the GNF data set. The largest PubChem screen retrieved assayed ~360 000 compounds, compared to a typical size of over 1 million compounds in GNF screens. To exclude screens based on focused compound libraries, only the PubChem assays tested for more than 10 000 compounds with a hit rate lower than 0.3% were retained. The definition of a "hit" was taken from its PubChem designation of being "active". A higher hit rate threshold was adopted here in order to retain more assays. As the result, the final PubChem data set consisted of 551 200 compounds and 90 assays, including 78 cellular assays and 12 biochemical assays. Based on eq 1, 349 compounds with $f_H > 0$ for cellular assays and 5 compounds with $f_H > 0$ for biochemical assays were found.

RESULTS AND DISCUSSIONS

No Single Hit Frequency Is a Good Threshold. In our HTS data set, 4301 compounds were marked as actives in at

dx.doi.org/10.1021/ci300005y | J. Chem. Inf. Model. 2012, 52, 913−926

least five separate assays, using a 0.1% hit-calling threshold. Their f_H values were calculated based on the proportion of active assays (see Materials and Methods, eq 1). For all hits from the 277 assays, 10% of those with the highest f_H values only occupied ~1.7% of the chemical space covered by all the unique hit structures (Figure 2a). Without a hit triaging strategy

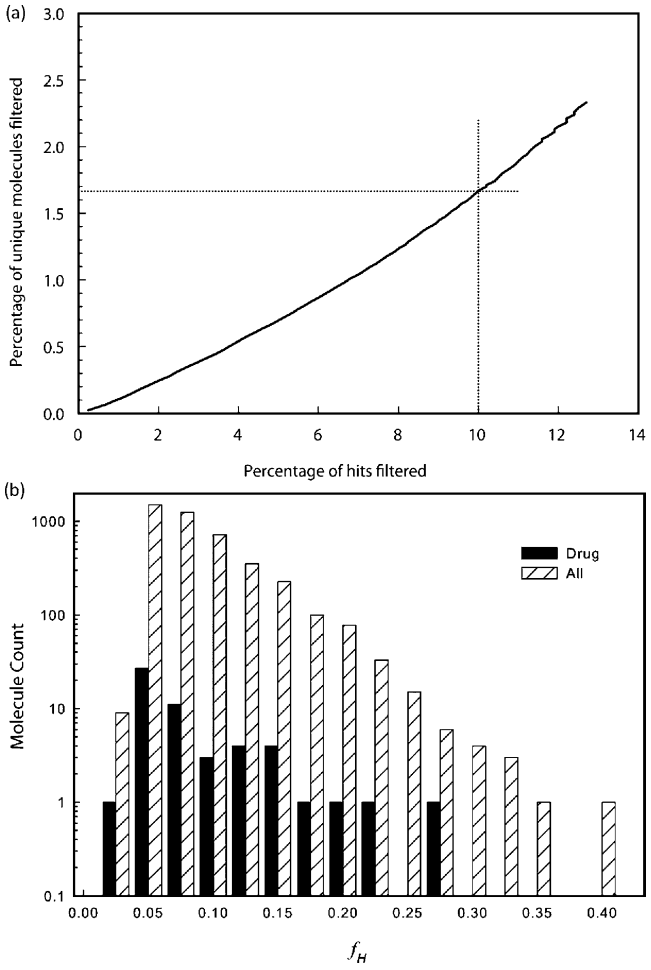


Figure 2. (a) Percentage of most frequent hits in all assays as a function of unique structures. The 10% most frequent hits only correspond to about 1.7% unique molecular structures. (b) f_H distributions for frequent hits in all assays and for known drugs.

to account for frequent hits, the same set of nonspecific molecules would be repetitively selected for reconfirmation by multiple projects. The common wisdom has been to either filter them out at the beginning of the HTS hit selection process or even permanently exclude them from the screening library.[21]

However, it is practically difficult to flag and deprioritize frequent hits solely by their f_H values. As shown in Figure 2b, f_H values of screening hits formed a continuous exponential distribution. Although the majority of hits had relatively low f_H, there was a long tail of high f_H hits that could account for as many as 25% of the hits within some screens. This unimodal distribution implied there was not an obvious f_H cutoff value to segregate frequent hits from nonfrequent hits. As the definition of frequent hit itself was subjective, the f_H-based triaging process could be difficult to implement on a consistent basis. The situation was further complicated by the observation that some compounds with significant f_H scores were marketed drugs. For example, the f_H distribution of 54 known drugs (with $f_H > 0$) contained in the database was overlaid with all the compounds in Figure 2b. In contrast to the common belief that marketed drugs frequently demonstrate activity toward specific cellular targets, the overall trend of f_H distribution was comparable between drugs and screening hits. For example, we found drugs such as Sutent scored as a hit in 20% of the kinase assays screened, which is a frequency consistent with literature reports.[22] Also, 8-iso-13,14-dihydro-15-keto Prostaglandin F2α was a hit in 14% of all assays; Aurantimycin-A hit in as many as 27% of the cellular assays profiled. Our analysis on f_H distributions argues that frequent hits should not be triaged simply based on their f_H values. Doing so imposes a concrete risk of eliminating leads that potentially have drug-like characteristics.

(Step A) Assay Classification: Dependency between Assay Types and Frequent Hits. As we mentioned previously, f_H can be tightly coupled with particular assay platforms. For example, Staurosporine has a high f_H among kinase assays and, consequently, tends to be considered as an uninteresting promiscuous hit in this context. However, it likely would not be designated as a frequent hit in screening program focusing on GPCR assays. The f_H score assigned to a compound would depend upon the protein target family that a screening center focuses on. Biochemical assays and cellular assays were the two largest assay groups in our HTS portfolio; therefore, we studied potential dependencies between f_H and these two assay types.

We initially applied a heuristic f_H cutoff of 0.1 in order to assign promiscuous hit designations to 1214 hits across the 55 biochemical assays and 1076 hits spanning the 222 cellular assays. There were only about 6% (61 compounds) in common between the two frequent hit lists. Although different f_H cutoff values were applied, the observation that frequent hits derived from the two assay types were dramatically distinct remained unchanged. Because the exact same compound collection had been screened in both assay systems, the large difference observed implied the biological mechanisms behind the frequent hits must be fundamentally different. The strong dependence of f_H values on assay type highlighted our concern for the "one-size-fits-all" approach that is generally used in triaging frequent hits.

Next, we applied the OPI[17] algorithm to all of the cell-based and biochemical assays, respectively (see Materials and Methods). The frequent hits were identified with p-values less than 0.05. This method found 990 compounds as frequent hits for the 55 biochemical assays, 2263 compounds for the 222 cellular assays, and 2142 compounds for both assays combined. Again, there is an about 6% (104 compounds) overlap between cellular and biochemical frequent hits. This degree of overlap was consistent with the earlier naïve approach of using a 0.1f_H cutoff. The exact molecules identified as frequent hits were not identical but nevertheless largely overlapping. A more detailed depiction of the overlap among the frequent hits is shown in a Venn diagram in Figure 3. There were only 96 molecules shared by all three frequent hit lists. Since the cellular assays made up the majority of the total assays, a large overlap existed between the frequent hits identified from the cellular and all assays. The marginal overlap observed between the hits of the biochemical assays and the cellular assays was not due to compounds that were differentially screened in one of the assay classifications. Cellular frequent hits were tested in 22 of the biochemical assays on average, and 88% were tested in over 10 biochemical assays. Similarly, biochemical frequent hits were tested in 78 cellular assays on average, and more than 99% were tested in over 24 cellular assays. Because all the frequent hits were tested sufficiently across both assay types, the overlap

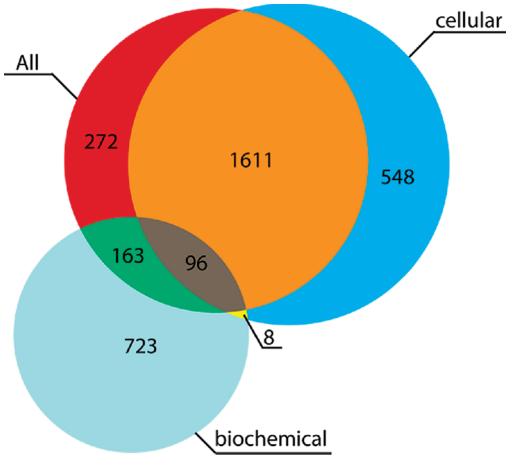


Figure 3. Venn diagram of frequent hits from all assays, biochemical assays and cellular assays derived based on the OPI algorithm.

observed above was quite robust. These OPI-derived frequent hit lists were used hereafter, as they were not derived based on any subjective f_H cutoff value.

Despite the general lack of common promiscuous scaffolds between the two assay types, they nevertheless shared a few chemotypes that were notable. Figure 4 outlines a few such examples, where their f_H values in both the cellular (i.e., f_H^c) and biochemical (i.e., f_H^e) assays are listed. As one might expect, some of these molecules are believed to be cytotoxic to cultured cells. For example, Staurosporine analogs (GNF-301, GNF-70) were active in many cellular and biochemical assays. Some other lesser known kinase inhibitors were also active in many cellular and biochemical assays, such as molecules GNF-689, GNF-695, and GNF-586 that are depicted in Figure 4.[23] In addition, fluorescent molecules GNF-276 and GNF-343 likely were identified as frequent hits due to the nature of the assay detection methods. Interestingly, the antibiotic spiroxin (GNF-721) also appeared to be active in many cellular and biochemical assays. As the examples in Figure 4 had been tested in a sufficiently large number of assays in both assay types, they were unlikely found simply due to statistical noise. Notably, we observed subtle structural modifications could result in dramatic changes in the selectivity profiles. For example, although compounds GNF-695, GNF-586, GNF-301, and GNF-70 are structurally closely related, their f_H values are a few fold different for both the cellular and biochemical assays. In both cases, a highly promiscuous molecule is turned into a less promiscuous one with simple peripheral modifications. This will be discussed in further detail below.

Our HTS assays were classified into 12 different target families, such as kinases, GPCRs, nuclear hormone receptors, transcription factors, etc. Due to the biology, technology, and the disease focus of the company, assays corresponding to a target family may tend to be coupled with either biochemical or cellular types, which might account for the observed difference in frequent hits between two assay classes. To examine this possibility, Fisher's exact test was performed and three associations were found to be statistically significance: kinases ($p < 10^{-4}$) and proteases ($p = 0.003$) for biochemical assay and GPCR ($p = 0.0007$) for cellular assays. The biochemical assays consisted of many kinase targets, namely 32 out of 55. Also, 20 out of 222 cellular assays were performed for identifying inhibitors of a particular kinase. Similarly, the proportion of protease enzymatic assay (5 out of 55) was also much higher in biochemical assays than that in cellular assays (7 out of 222). On the other hand, GPCR assays were only performed using a cellular setting (33 out of 222). We then applied Fisher's exact test to evaluate the association of each frequent hit to these target families. There were no frequent hit found to be significantly associated with protease screens. Among the 2263 cellular frequent hits, only 75 compounds showed statistically significant association ($p \leq 0.05$) with the GPCR family of screens. Likewise, only 45 compounds out of 990 biochemical frequent hits showed statistical significant dependence with kinase screens.

To further examine if the frequent hits could be associated with on-target activities, we identified a total of six assay pairs in our database, where both cell-based and biochemical assays were performed to identify inhibitors of the same protein target. The number of frequent hits tested in these assay pairs ranged from 34 to 1394. The subset that was considered hits in these assays had low overlaps with an average of 4%, in agreement with the 6% observed across the complete assay panel. If only biochemical frequent hits were considered, the overlap is 1%. If most biochemical frequent hits found had been on-target, the overlap would have been much higher. In our previous analysis with the complete assay panel, the majority of biochemical frequent hits were not more active in kinase than other protein families, despite kinases overrepresented there. Combined, this suggested that the activities of most biochemical frequent hits might be due to certain artifacts instead of their target-binding characteristics.

(Step B) Biological Mechanism Analysis of Frequent Hits. *Common MOA for Frequent Hits.* Our early analysis showed that the removal of the top 10% of the most frequent hits only reduced the chemical space covered by all of the hits by 1.7% (Figure 2a). However, we have shown that a triage strategy based on f_H values alone may filter out potentially valuable lead candidates. In fact, our experience suggests that researchers rarely rely on f_H alone for triaging molecules in practice. We have seen some promiscuous hits repeatedly chosen for post HTS confirmations, despite their relatively high f_H values. Ascribing a biological context to frequent hits often was a more enabling and effective filter than f_H alone. For example, the annotation of Staurosporine as a "nonspecific kinase inhibitor," provided scientists with key information to eliminate the compound from further development.

An MOA-driven OPI analysis (see Materials and Methods) was applied to the f_H vector of all the HTS compounds in order to identify MOAs that were enriched among frequent hits. The results were rather informative. For example, 92 structures were defined as "protein synthesis inhibitors" based on MeSH annotation (MeSH ID 82011500). Without relying on an arbitrary f_H cutoff, OPI determined that the top 3241 most frequent hits of the f_H -sorted compound list contained 19 out of 92 annotated protein synthesis inhibitors. If the MOA, "protein synthesis inhibitors", had no connection with a compound's hit frequency, one would only expect 0.1 compounds from this MOA family to appear in the top 3241 list ($92 \times 3241/3\,000\,000 = 0.1$). The chance of observing 19 protein synthesis inhibitors in this set corresponds to an enrichment factor of 190 and a strikingly low p-value of 10^{-37}. Repeating the same analysis using frequent hits from the cellular assays alone also confirmed the same discovery of a p-value of 10^{-39}. On the other hand, this MOA was only marginally significant for biochemical assays, giving a p-value of 0.03. This example illustrates how our analysis appears to correctly determine that protein synthesis inhibition is an MOA that scores as a hit in many cellular assays, likely due to general cytotoxicity. In contrast, compounds with this same biological mechanism did

dx.doi.org/10.1021/ci300005y | J. Chem. Inf. Model. 2012, 52, 913−926

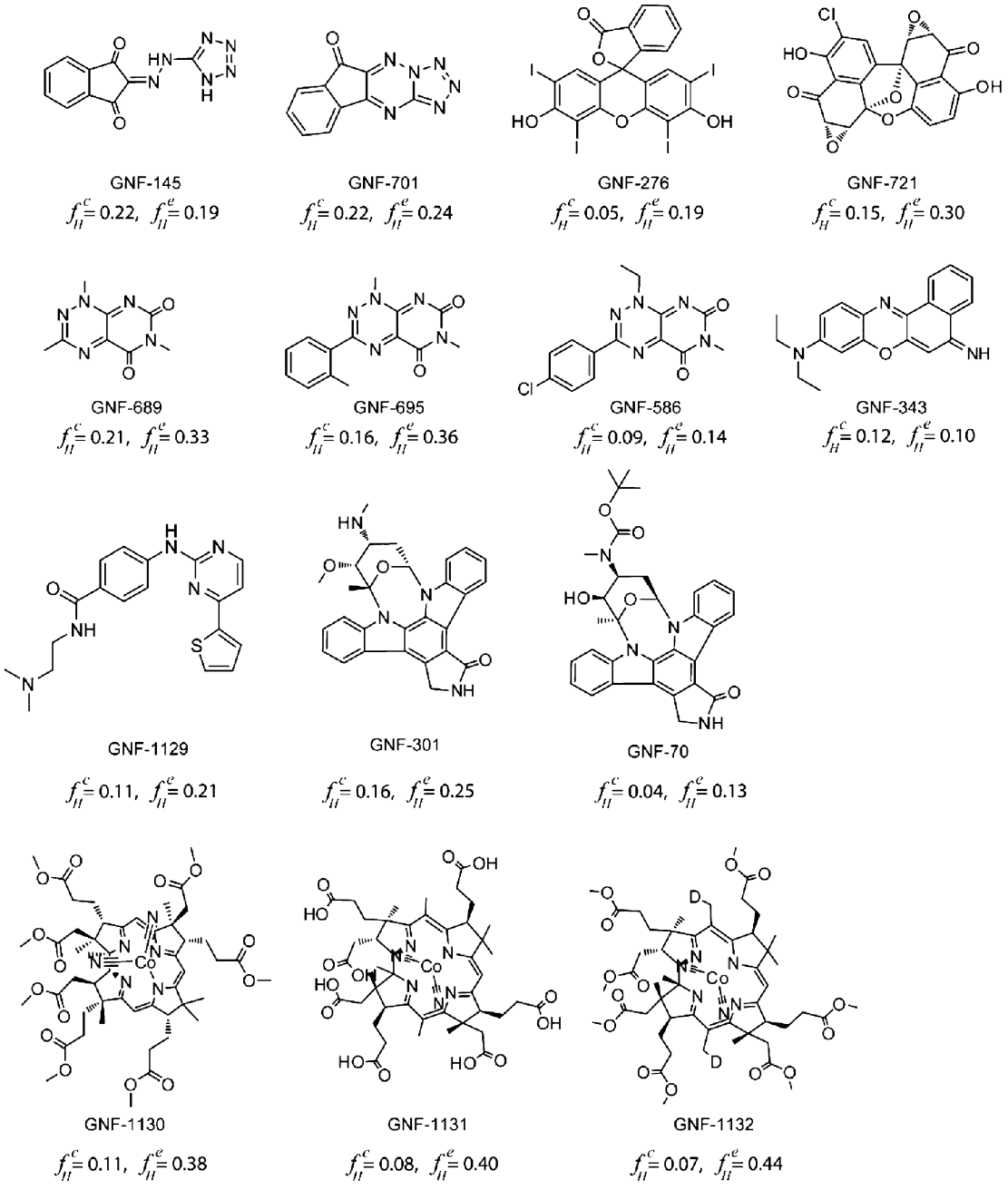


Figure 4. Selected common frequent hits in both cellular and biochemical assays. f_H^c denotes hit frequency in cellular assays, and f_H^e denotes the hit frequency in biochemical assays.

not affect most biochemical assays, since most proteins in biochemical assays were not related to protein translation.

Without relying on a subjective f_H cutoff value, the OPI algorithm provided a rather unique statistical framework that effectively distinguished true frequent hits from the remaining members for a given MOA group. This is particularly important for analyses driven by the MeSH MOA classification, because MeSH ontology terms are often very vague and broad, such as "antimetabolites, antineoplastic" and "protein kinase inhibitors." Taking the protein kinase inhibitors classification for instance, there were 22 molecules under this category in our screening data matrix. They included compounds with a wide range of f_H values, from very nonselective compounds such as Staurosporine, to more selective ones such as Sorafenib and Imatinib, to very specific ones such as Laptinib. Although "protein kinase inhibitors" was a statistically significant mechanism for frequent hits, represented by a p-value of 10^{-6}, obviously not all kinase inhibitors had equivalent statistical significance. The OPI algorithm determined that three molecules, Staurosporine (f_H = 8%), Staurosporine racemates (f_H = 6%), and Flavopiridol (f_H = 6%) were the only true frequent hits within this group, and the remaining molecules could be considered nonpromiscuous. Indeed, Staurosporine and Flavopiridol are known for their broad spectrum kinase activities.[22] In contrast, 19 nonpromiscuous kinase inhibitors identified by OPI included drugs like Gefitinib and Imatinib. Despite the lack of desirable granularity in MeSH MOA ontology categories, it is encouraging that the OPI algorithm was quite resistant to such ambiguity and was able to segregate promiscuous compounds from the rest.

dx.doi.org/10.1021/ci300005y | J. Chem. Inf. Model. 2012, 52, 913−926

Previously, Crisman et al. proposed inhibition of kinases, topoisomerases, and protein phosphatases as the most dominant mechanism for off target activity in cell-based assays.[24] Equipped with a large-scale HTS data matrix of 4353 MeSH structures across 294 MOA categories, we were able to study the biological properties of frequent hits with much greater resolution. Table 1 summarizes the 14 MOA groups that not only had *p*-values less than 0.001, but also included over 10% of their members within our compound deck identified as frequent hits. Our study identified several new MOAs not widely described previously in the literature as frequent hits, such as ionophores. Ionophores disrupt ion concentration gradients across cell membrane, which potentially lead to reduced viability of certain cultured cells. Some ionophores are also used as antibiotics, which is consistent with antibiotics being one of the MOA's listed in Table 1. Consistent with other reports, cardiotonic agents were also seen to have significant promiscuous activity.[25]

Table 1. Key MOAs of HTS Frequent Hits

MeSH	description	$\log_{10} P$	hit rate (%)
82004639	emetics	−14.0	83
82014344	trypanocidal agents	−11.2	31
82007476	ionophores	−15.0	28
82000903	antibiotics, antineoplastic	−27.2	27
82011500	protein synthesis inhibitors	−34.5	20
82000563	amebicides	−10.0	19
82002400	cathartics	−9.8	18
82002316	cardiotonic agents	−24.1	16
82003049	coccidiostats	−7.7	16
82000969	antinematodal agents	−9.4	15
82019384	nucleic acid synthesis inhibitors	−11.0	14
82000964	antimetabolites, antineoplastic	−12.7	14
82047428	protein kinase inhibitors	−5.7	14
82005456	fluorescent dyes	−9.0	13

These findings suggest a frequent hit triage strategy based upon the MOA of the compound hits. When that MOA is unrelated to the primary purpose of the biological assay, scientists can confidently remove the frequent hits from further validations. Otherwise, if the MOA is of biological relevance, such as Midostaurin for an oncology indication, the hit should be retained. When hits have an ascribed MOA that is relevant to a specific screening campaign, they nevertheless should be retained as lead candidates despite their high f_H scores. This strategy would ensure that many frequent hits that are cytotoxic to cells be retained for screens where the focus is to identify agents cytotoxic to tumor cell lines. Frequent hits with MOAs irrelevant to the biological focus of the screen could be removed with confidence. Promiscuous hits that lacked any mechanistic annotations would then need to be addressed based on either logistic considerations or structural filters described below.

MOA Distributions Across HTS Campaigns. In a conventional HTS, the activity of a small molecule is often represented by a single value such as percent inhibition. As a result, the response due to different MOAs can be convoluted and difficult to differentiate. The large HTS data matrix used here represented a diverse collection of independent biological systems, which presented us with a unique opportunity to identify biological mechanisms accountable for the activity pattern across each individual assay. In combination, they could help to answer questions such as, "what MOAs contribute to biochemical or cellular specific activities, and what MOAs contribute to generic frequent hits for both"? The analysis in the previous section effectively compressed this large matrix into a single f_H vector for common MOA identification, where arguably the signal of some MOA groups might have been too weak to detect. To address this issue, we reanalyzed MOAs on each individual assays first and then identified reoccurring MOAs afterward.

For each of the 277 assays, we applied the same OPI algorithm to each MeSH category in order to identify MOA members that had statistically unusually potent activities compared to the large background of the remaining compounds (see Materials and Methods). The calculation yielded a *p*-value for each MOA category that quantified the association strength between the MOA and its activity in a particular assay. Figure 5 is a heat map of the resulting log *P* values of 278 MOA categories across 219 assays (only 118 MOAs are shown for clarity). Assays are further segregated into cellular and biochemical assays and then hierarchically clustered independently into two assay trees, columnwise. MOA classes of similar enrichment profiles across assays were clustered into an MOA tree, rowwise. The lower the *p*-value (red), the stronger the biological association between the MOA and the particular assay.

The majority of the MOAs in the heat map generally displayed rather selective patterns. For example, "Anti-HIV agents" (MeSH 82019380) and "Reverse Transcriptase Inhibitor" (MeSH 82018894) showed activity only in HIV viral infection assay and lacked notable activity across all other cellular and biochemical assays. Another example is "Anti-infective agents" (MeSH 82000890), which only showed activity in antibacterial screens such as "*Mycobacterium smegmatis*" and "tuberculosis". The lack of activity of this MOA in screens using human cell lines implies that the hits from these screens were high quality lead candidates that were not only potent but also had limited liabilities within mammalian host systems. Another interesting example is "Antimalarial Inhibitors" (MeSH 82000962). This class of compounds was extremely enriched in the malaria cellular screens and lacked significant activity in most other screens. The exceptions were two cellular screens, BCR-Abl ($p < 10^{-5.2}$) and the TLR9 antagonist screens ($p < 10^{-3.9}$). The antimalaria compounds' activities in BCR-Abl may be related to the observation that DHFR inhibitor-like compounds also target this kinase.[18] Interestingly, it also has been reported that antimalarial compounds, such as chloroquine, hydroxychloroquine, and quinacrine are antagonists of TLR9,[26] and they have been used for immune-mediated inflammatory disorders such as rheumatoid arthritis, systemic lupus erythematosus, and Sjögren's syndrome.[27]

This suggests that an MOA-based analysis across a large assay panel can also help identify drug repositioning opportunities. This approach is quite different from an individual compound-based reposition exercise. First, the initial observation is rather insensitive to the assay typically found in HTS experiments, because each MOA row in the heat map is supported by multiple underlying compound members. Second, the structural−activity relationship is not assumed in this analysis. In fact, the molecules within the same MOA class can be structurally diverse, and the linkage of an MOA with another assay could theoretically enable new indications of not only one but a class of diverse molecules. Third, it provides clues to common biological processes under different diseases for further investigation.

For the purpose of this study, we were most interested in the MOAs of compounds that repeatedly were identified across a diverse range of screens in order to potentially identify a biological explanation to why certain compounds were frequent hits. To this end, there was a group of MOAs active across a wide range of the cellular assays in our data set, depicted in the

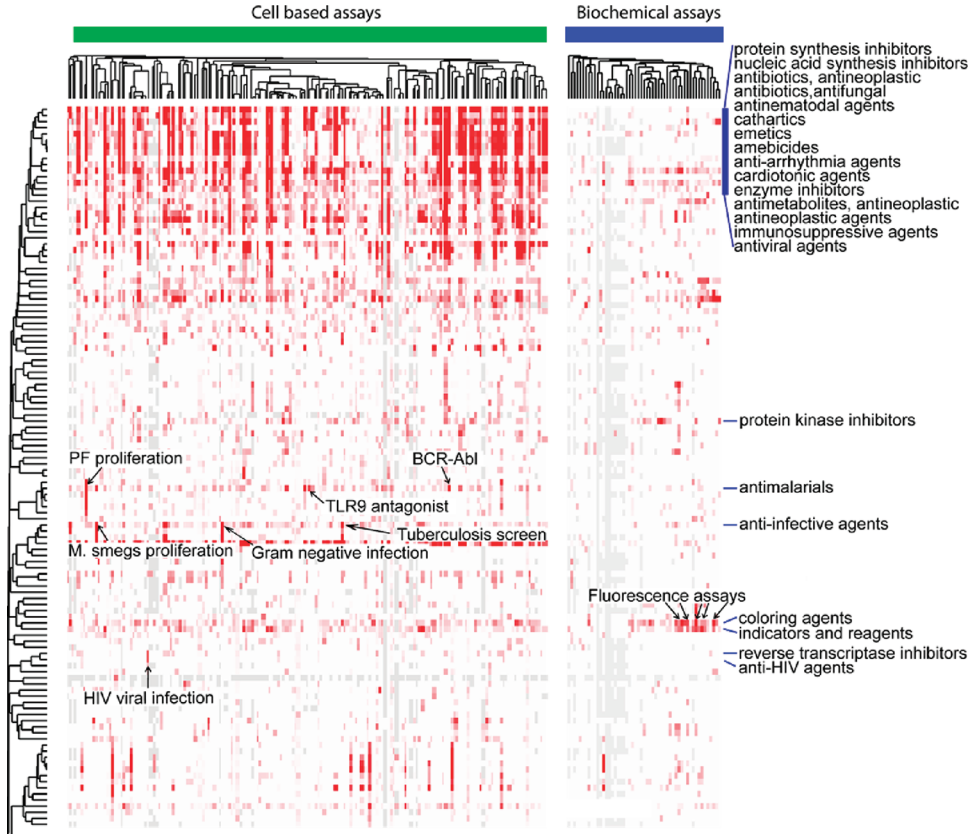


Figure 5. Enrichment heat map of MeSH MOA categories across all HTS assays. The color scale represents the log-scale *p*-values for the enrichment of the particular MOA in a given assay. The darker red represents better *p*-values (lower). Some representative MeSH groups and assays discussed in the main text are labeled for convenience.

top portion of Figure 5. Reconfirming our previous observations (Table 1), the MOA "protein synthesis inhibitors" clearly showed a high hit rate in the majority of cellular assays. Expectedly, this MOA was not prominent in the frequent hits originating from the biochemical assays, as its promiscuity likely was due to modulating a few vital cellular targets rather than nonspecific binding to many different targets. Since many protein synthesis inhibitors are used as antibiotics, this provides a rationale for the observation that MOA "antibiotics", "antineoplastic" and "protein synthesis inhibitors" were close neighbors in the MOA tree. Another example of nonselective MOAs was "nucleic acid synthesis inhibitors", where these molecules would be predicted to disrupt a fundamental cellular process that is required for survival. This supports a model where MOAs that negatively impact cellular viability prominently contribute to the frequent hits identified from cell-based but not biochemical-based screens. This would imply that certain MOAs, such as "anti-arrhythmia agents" could show broad cytotoxicity across cellular assays.

In contrast, most MOAs were quite specific on particular biochemical systems. There was rarely an MOA class with activity across the entire range of biochemical assays. Our biochemical assay targets contained not only a diversity of protein families, but also targets originating from multiple organisms, including human, bacterial, and viral proteins. "Protein kinase inhibitors" (MeSH 82047428) showed activities only in a few kinase enzymatic assays. Further examination identified three classes that showed relatively broad activities in biochemical assays: "Indicators and Reagents" (MeSH 82007202), "Coloring Agents" (MeSH 82004396), and "Fluorescent Dyes" (MeSH 82005456). Their broad profiles were largely associated with the compounds' intrinsic physiochemical properties that likely interfered with the assay detection mechanism in our data set.[9c]

The MOA heat map provided an interesting summarization of biological activities embedded in our large HTS database. This concept could be applied to other screening data set of interest, which potentially would allow one to detect nontrivial new disease indications of certain MOA classes. Such MOA-based hit analysis methods can provide an effective knowledge-driven screening validation tool,[11] as well as the critical biological context for a better MOA evidence-based hit triage strategy. Bear in mind that our MOA analysis is limited due to the low percentage of compounds carrying MOA annotations, as well as many known MOA members that may not be included in our screening library. As the result, additional MOAs also accountable for cellular frequent hits may not be detected in this study.

Distinct Mechanisms between Cellular and Biochemical Frequent Hits. We have observed that there was little overlap in the chemotypes between frequent hits in cellular assays and biochemical assays, and their MOA enrichment patterns were also strikingly different (Figure 5). This phenomenon suggested that frequent hits could employ very different molecular activities. For example, a frequent hit could bind either nonspecifically to multiple gene/protein targets or selectively to a few targets that are critical to cellular viability. The previous MOA analyses showed frequent hits from the biochemical screens tended to have predicted poor target selectivity, such as the pan kinase inhibitor Staurosporine, or produced detection artifacts, such as dye molecules that interfere with fluorescence-based

dx.doi.org/10.1021/ci300005y | J. Chem. Inf. Model. 2012, 52, 913−926

biochemical signal detection. Although cellular frequent hits could also result from the interaction of multiple targets, such as "dirty" kinase inhibitors, we hypothesize that most of them more likely impact a small number of key biological nodes that result in cytotoxicity.

We examined this hypothesis by studying the molecular networks impacted by frequent hits. First, we used the GeneGo[13] database to extract protein target annotation for the 2171 out of 2783 known drugs and the 262 out of 2263 promiscuous molecules in our cellular assays. Figure 6 illustrates their distributions in terms of the number of predicted targets. A large number of known drugs (~40%) and cellular frequent hits (~48%) had five or less known targets, which supported our assumption that a significant portion of the frequent hits were as selective as some known drugs in terms of their protein targets. The last bin in Figure 6

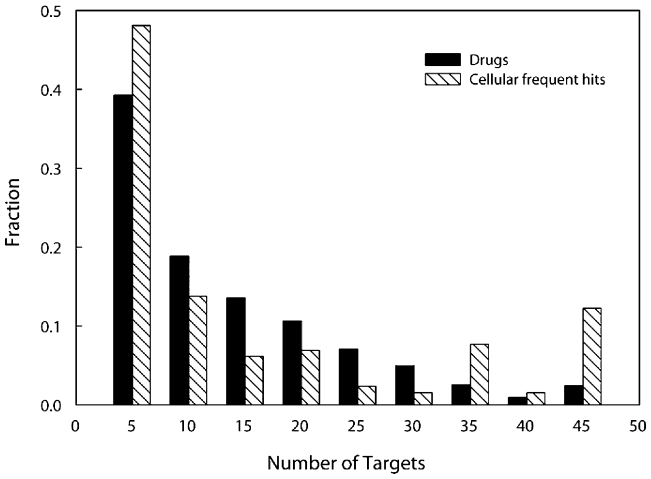


Figure 6. Distribution of number of targets for known drugs and for cellular frequent hits.

represented all molecules with 45 or more known targets. For cellular frequent hits, this bin almost entirely consists of Staurosporine analogs with the exception of Vincristine, a mitotic inhibitor. For known drugs, this bin contained a diverse set of molecules, including Vincristine, Midostaurin, a Staurosporine analog, and several steroids such as Etiocholanolone and Allopregnanolone. In addition, non-specific molecules such as Linoleic acid and Resveratrol also had many annotated protein targets.

The compound−protein network extracted above provided a list of targets for promiscuous molecules. One interesting aspect of the list was the ability to identify proteins frequently targeted by cellular-only frequent hits but not by the known drugs. Among them, many are vital to cell survival or cell growth. The top three targets were WEE1, HDAC3, and Tank-binding kinase 1 (TBK1). WEE1, a Ser/Thr kinase, is a component for cellular cell size checkpoint, which controls the time point of mitosis initiation. Without WEE1, cell division occurs prematurely. Interestingly, a WEE1 inhibitor (MK-1775) is currently in clinical trials for solid tumors.[28] HDAC3 is critical for RNA transcription regulation and cell growth, and TBK1 closely interacts with and mediates NF-κB activation in response to certain growth factors.

Among the popular targets shared between both frequent hits and known drugs, ATP binding cassette members, ABCB1 and ABCG2, are on top of the list. They are involved in transporting various molecules, including drugs, across cellular membranes. Another example was EGFR inhibitors, where multiple drugs are clinically approved, including Gefitinib, Erlotinib, and some earlier

Table 2. Key GO Groups and Pathways for Cellular Frequent Hits

description	$\log_{10} P$
cell cycle G1-S growth factor regulation	−44.7
signal transduction TGF-beta, GDF, and Activin signaling	−42.8
inflammation MIF signaling	−41.5
development VEGF signaling and activation	−39.1
CYPs-exo	−38.0
GO:0004674: protein serine/threonine kinase activity	−36.7
development EGFR signaling pathway	−35.0
development VEGF signaling via VEGFR2-generic cascades	−32.7
development hemopoiesis, erythropoietin pathway	−29.8
cell cycle G1-S interleukin regulation	−29.2
development A2A receptor signaling	−27.8
signal transduction Erk interactions: Inhibition of Erk	−26.7
CYPs C/EBP-mediated regulation	−26.2
immune response gastrin in inflammatory response	−26.2
signal transduction ERBB-family signaling	−26.1
inflammation IL-10 anti-inflammatory response	−25.7
development regulation of angiogenesis	−25.2

tyrphostins.[29] Also, many nonspecific kinase inhibitors in our frequent hit lists also inhibited EGFR as seen by the high enrichment of protein kinase inhibitors in enzymatic EGFR screens. Therefore, if one were looking for novel EGFR inhibitors, one should not remove all frequent hits based on any ad hoc f_H cutoff value.

To understand the biological pathways that were targeted by frequent hits, we selected 194 targets that were included in the MOA annotation and had at least 10 promiscuous compounds. We then applied a gene ontology enrichment analysis. Table 2 lists the top 17 biological pathways (p-value $<10^{-25}$) that are targeted by these compounds. Protein kinase pathways and cell cycle regulators were significant represented within the list, as expected. For instance, "cell cycle G1/S growth factor regulation" was the most significant process. The core components of this pathway (i.e., Cdc25A, cyclins, CDK2, CDK4, and CDK6) regulate the cellular G1/S checkpoint and hence are required for DNA replication and the initiation of the S phase. This process has implications in diseases, such as cancer. "Signal transduction of TGF-β, GDF, and activin signaling" was another process heavily targeted by frequent hits. This pathway controls a diverse set of cellular processes, including cell proliferation, recognition, differentiation, apoptosis, and developmental fate.[30] Clearly, the pathways identified here are essential to cell survival, which supports our initial hypothesis. Multiple targets involved in these processes such as CDK2/4/6 are also experimental drug targets with compounds in various stages of clinical trial for cancers.

(Step C) Chemical Features of Frequent Hits. After biological MOA-based triage at step B, one might further apply structure-based methods. As atomic structures ultimately are the underlying factors that dictate a molecule's activities, we explored the chemical space of frequent hits and identified reliable chemotypes that correlated with their designation as frequent hits in biochemical and cellular assays, respectively. Simple empirical rules such as "the rule of five"[31] are widely adopted by modern drug discovery efforts in order to provide predictions as to the "drug-likeliness" of small molecules based on their physicochemical properties. As one might expect, properties such as molecular weight and log P of HTS hits and drugs are quite different (data not shown), because lead molecules from an HTS

dx.doi.org/10.1021/ci300005y | J. Chem. Inf. Model. 2012, 52, 913−926

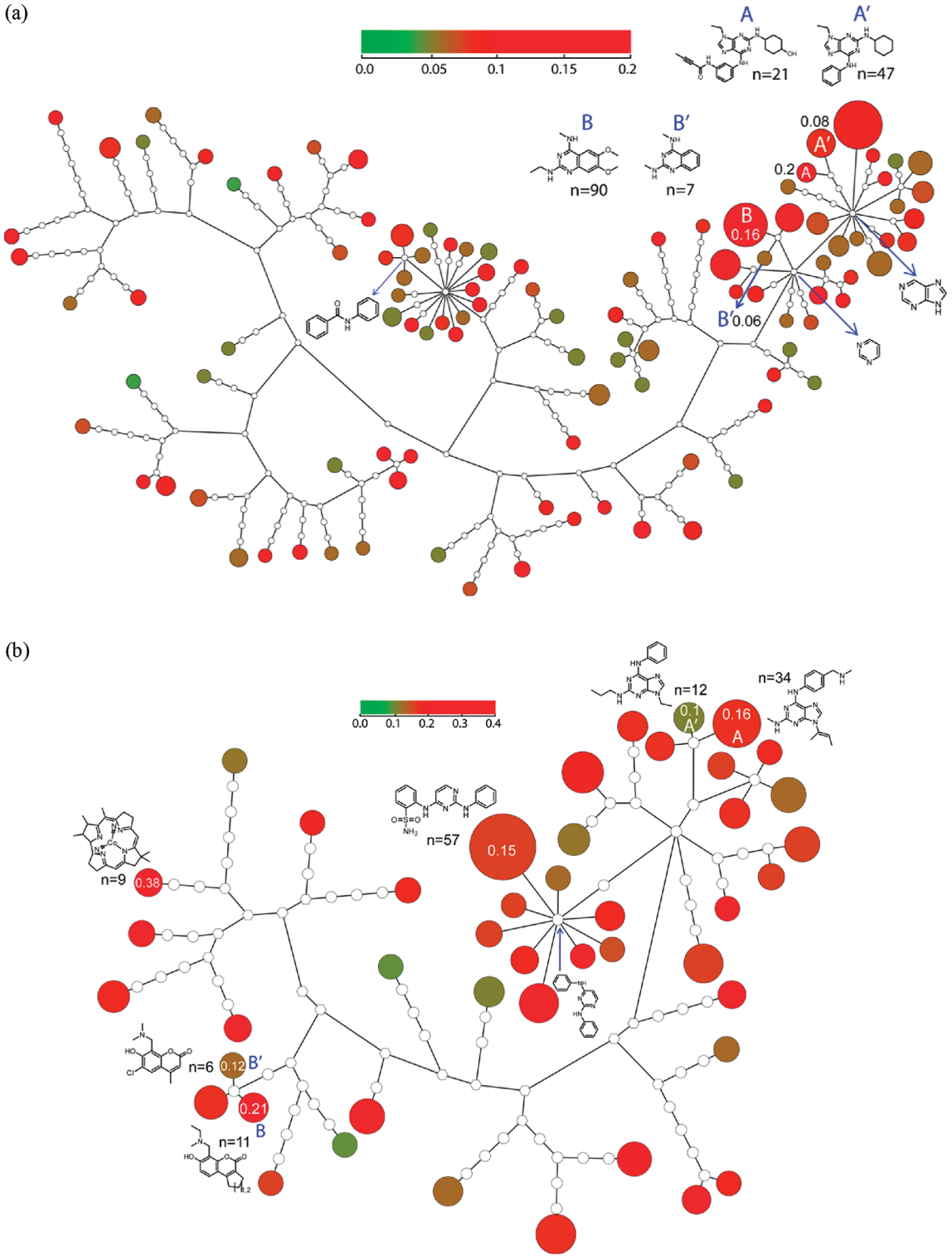


Figure 7. Tree diagrams of frequent hit scaffolds with five or more members per scaffold from cellular (a) and biochemical (b) assays. The leaf nodes representing scaffolds are colored based on their median f_H and sized based on the number of molecules within. Example chemical scaffolds are shown with the cluster size and median f_H values.

tend to undergo extensive modifications in order to improve their potency, bioavailability, and pharmacokinetics properties before becoming drugs. However, the frequent hits shared nearly the same physicochemical properties as the rest of the HTS hits, i.e., there was no obvious correlation between such properties and the compounds' selectivity profiles, regardless of the assay types used for screening. Although these properties might be partially accountable for some of the differential activities of a compound between biochemical and cellular assays via their relationships to cell permeability, they cannot explain why some compounds bind to multiple enzymes or regulate multiple cellular systems, while others demonstrated clear selectivity.

dx.doi.org/10.1021/ci300005y | J. Chem. Inf. Model. 2012, 52, 913−926

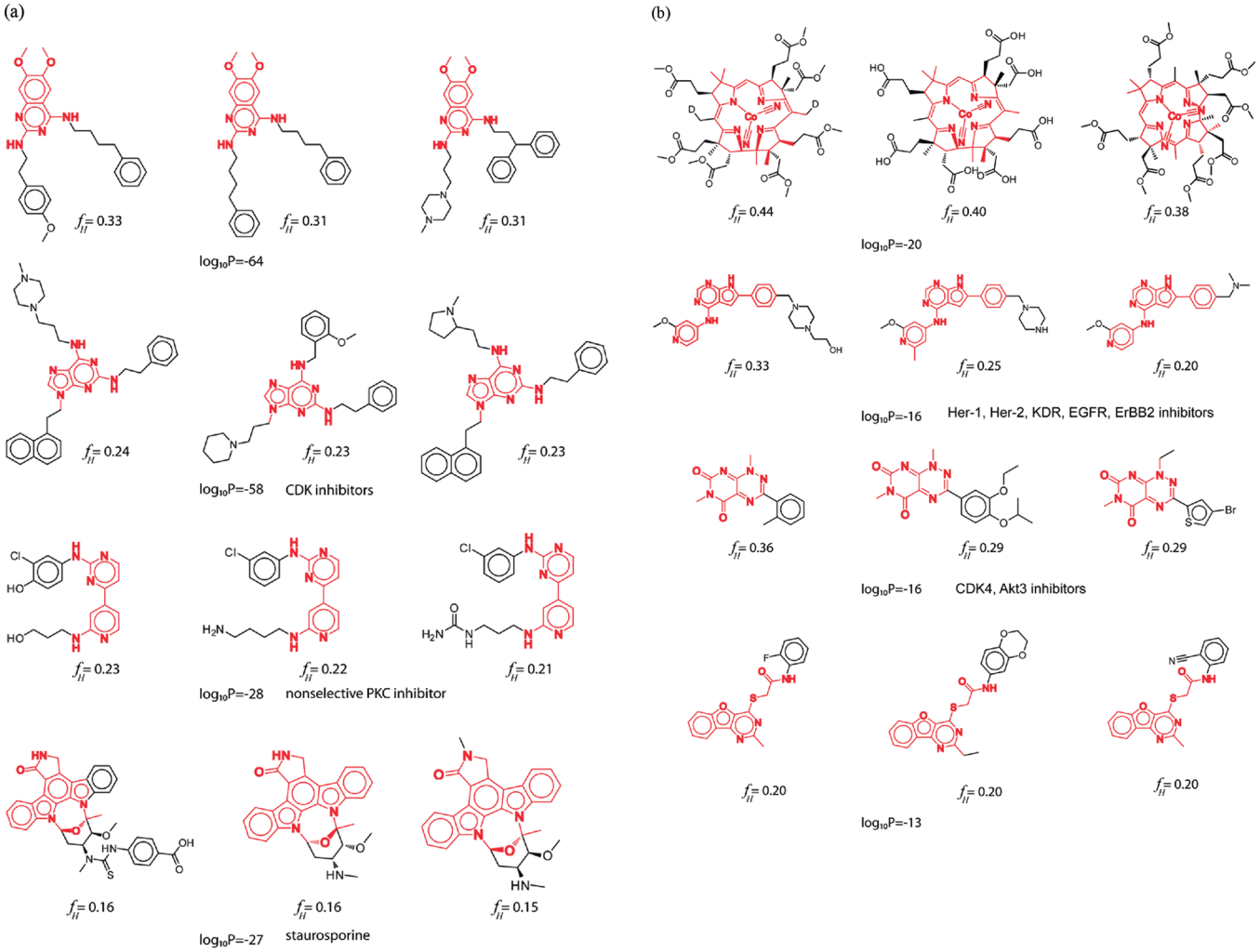


Figure 8. Selected prominent scaffolds of frequent hits in cellular (a) and biochemical (b) assays. The statistical *p*-values and f_H values are shown for each scaffold.

There were a total of 479 OPI structure families with *p*-value ≤ 0.05 that had at least two promiscuous members identified from both assay types. Among them, 387 biochemical (Supporting Information Table S1) and 778 cellular frequent hits (Supporting Information Table S2) have structures that were publically available. For the convenience of visualization and discussion, a selected subset of the nonproprietary structural families that had five or more members derived from biochemical and cellular assays were displayed in Figure 7. The leaf nodes of the tree represented the maximum common substructures (MCSSs) of each cluster identified above. By further trimming peripheral functional groups, higher level nodes represent simpler and smaller scaffolds shared by the lower level nodes.[32] The highest level scaffolds were then clustered by their chemical fingerprints. The color of each leaf node represented the median f_H of all molecules belonging to that scaffold, and its size was proportional to the number of molecules sharing the scaffold. This hierarchical tree rendered a global view of the distribution of frequent hits in chemical space. In addition, it outlined how the promiscuous propensity of certain scaffolds varied upon chemical modification. For example, a region of abundant frequent hits shown in Figure 7a largely consists of purine and pyrimidine analogs. When two pairs of scaffolds labeled as (A and A′) and (B and B′) were compared, although both pairs shared large portions of common structure, the addition of small functional groups could readily transform a relatively less promiscuous cluster of compounds into highly promiscuous molecules. This observation highlights the inherited risk associated with a structure-based triage strategy, i.e. a low frequency hit might mistakenly be predicted as a high frequency hit due to its structural similarity to a previously identified promiscuous chemotype. Therefore, whenever a large historical assay database is available, it is recommended that the chemotype evidence be combined with experimentally derived f_H phenotypic filters (Figure 1).

Figure 7b shows the results for frequent hits in biochemical assays. Several key scaffolds were depicted in the graph. As previously noted, there was little overlap between frequent hits derived from the two assay types. This is reinforced here at the scaffold level by visually comparing the examples between Figure 7a and those in Figure 7b. The ones shared among the two platforms were dominated by the purine and diaminopyrimidine scaffolds and staurosporine analogs.

Figure 8a shows the top four clusters that had the highest statistical significance, i.e., the lowest OPI *p*-values, identified in the cellular assay panel. Keep in mind that statistical significance is not equivalent to high f_H values. Therefore, these are not the four clusters of the highest f_H values. In the context of cellular assays, physical effects such as compound aggregation

dx.doi.org/10.1021/ci300005y | J. Chem. Inf. Model. 2012, 52, 913−926

are not expected to play a significant role in the observed promiscuity. Instead, one would expect true nonspecific target binding to be one of the driving forces here. Indeed, some of the compounds in Figure 8a are well-known "dirty" kinase inhibitors, such as Staurosporine and nonselective PKC inhibitors. These compound families inhibit a wide range of kinases and consequently alter a wide range of cellular activities. While Staurosporine itself may not be suitable for drug development, its close analog Midostaurin (PKC412) is currently in phase III clinical trial for acute myeloid leukemia (AML) patients with FLT-3 mutations[33] and phase II trial in aggressive systemic mastocytosis (ASM).[34] This is despite the observation that Midostaurin inhibits over 40% of kinases with $k_d < 3$ μM.[22] In Figure 8b, we show the most significant scaffolds from biochemical assays that were not covered by top cellular frequent hit scaffolds. The first scaffold is a cobalt-containing porphyrin-like molecule. The exact reason for their high f_H values is not clear. The second cluster is the deazapurine scaffold, which inhibited a wide range of kinases. Likewise, the third cluster of pyrimidotriazinedione analogs are also known to inhibit various kinases. However, we were also not entirely clear about the reason for the promiscuity of the fourth class of molecules, abenzofuranpyrimidine scaffold.

Frequent Hit Analysis in the PubChem Data Set. An attempt was made to apply the similar analyses described above to the PubChem data set (see Materials and Methods). Due to the much smaller data matrix, there were only 349 compounds with $f_H > 0$ for 78 cellular assays and 5 compounds with $f_H > 0$ for 12 biochemical assays. They were identified without applying a nontrivial f_H cutoff (see Supporting Information Tables S3 and S4). Among the five biochemical hits, PubChem CID2347892 ($f_H = 0.27$) was also identified in our frequent biochemical hit list ($f_H = 0.19$); two molecules were not in our screening collection. The other two did not show broad enough activity in the GNF assays. Among the 349 compounds with nontrivial f_H from the PubChem cellular assays, 179 were not in the GNF collection. The 23 out of the remaining 170 PubChem cellular hits were also similarly identified as GNF cellular frequent hits, a strong overlap of 10.5% ($p = 10^{-50}$). This validates the frequent hits originating from GNF's proprietary data matrix. The overlap of frequent hits increased if close analogs were considered. This analysis provided confidence that the promiscuous hits listed in Supporting Information Table S1 were largely transferable and provided indirect public-domain support to the related conclusions presented in this study. Among the MOAs of promiscuous hits identified in both data sets, the antibiotic toxoflavin had an f_H of 0.33, and this high f_H was consistent with one of the common MOA for promiscuity identified previous by in-house data. Other examples included antiparasitic drug nitazoxanide ($f_H = 0.27$), and several cardiotonic agents like digoxin ($f_H = 0.27$), and caradrin ($f_H = 0.26$). The cardiotonic agents also were one of the top MOAs from in-house cellular frequent hits (Figure 5). The Pubchem biochemical data panel was too small to allow us to robustly check whether the lack of overlap in frequent hits between cellular and biochemical assays also holds for a nonproprietary data set. We hope the observations and conclusions presented in this study can be reexamined as the proprietary and public-domain data set grows significantly in the future.

CONCLUSIONS

Through mining a large HTS data matrix consisting of three million structures screened across 277 assays, popular chemical structural features and biological mechanisms of action of frequent screening hits were identified in this study. In the chemical space covered by the compound library, the overlap between frequent hits in biochemical and cellular assays was merely about 6%. Further MOA and target data analyses indicated that frequent hits in biochemical assays typically either targeted a large number of proteins or tended to induce assay readout artifacts, while frequent hits in cellular assays were more likely to target only a few proteins in biological pathways essential for cellular viability. The number of cellular targets associated with cellular frequent hits was often comparable to that of known drugs. Because of the distinct mechanism for frequent hits, step A of our proposed ABC frequent hit triage strategy was to classify the assays into the proper assay groups, so that a helpful f_H could be calculated and applied.

The large scale HTS database utilized in this study covered a variety of assay technologies and disease areas. Therefore, even though the exact numbers presented here were subjected to the limitations of our particular data source, the general conclusions derived from our analyses likely are valid and applicable to other screening facilities. In particular, we have cross-validated a significant number of our cellular frequent hits by an available PubChem data set.

Based on the GNF database, we concluded that the chemical and biological properties of frequent hits were rather distinct between biochemical and cellular assays. Therefore, one should not transfer empirical hit rate data across the two platforms. Biochemical frequent hits caused by physicochemical artifacts and lacking biological mechanisms of interest, could be filtered out with minimum risk. In contrast, caution should be given for filtering cellular frequent hits. One could employ MOA biological analysis in step B to ensure that compounds targeting biologically relevant pathways are carefully examined. Lastly, one can apply step C, where known promiscuous chemical structural features (Supporting Information Table S1−S4), if available, can be utilized to remove additional molecules on top of empirical f_H data (Figure 1).

The main contribution of this study was to suggest an improved mechanism-based triage strategy to help remove frequent hits, while still minimizing the risk of removing biologically relevant compound candidates. In addition, several computational approaches were described here for identifying a comprehensive collection of structural features that illustrate "what" constitutes frequent hits, as well as shed some new lights on "why" they become frequent hits biologically. Answers to both questions are critical aspects that can actively contribute to drug discovery pipelines with higher quality hit molecules.

ASSOCIATED CONTENT

Supporting Information

Publicly available compounds that were identified as frequent hits within cellular and biochemical assays are listed in the excel spreadsheets. Compound structures, external database identifiers, and f_H values are included. This information is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jche@gnf.org (J.C.) and yzhou@gnf.org (Y.Z.).

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of high-throughput screening in biomedical research. *Nat Rev. Drug Discov.* **2011**, *10* (3), 188−195.

(2) Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH molecular libraries initiative. *Science* **2004**, *306* (5699), 1138−1139.

(3) (a) 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*, 1712−1722. (b) Ryan, A. J.; Gray, N. M.; Lowe, P. N.; Chung, C. W. Effect of detergent on 'promiscuous' inhibitors. *J. Med. Chem.* **2003**, *46*, 3448−3451. (c) Seidler, J.; McGovern, S. L.; Doman, T.; Shoichet, B. K. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J. Med. Chem.* **2003**, *46*, 4477−4486.

(4) Scheiber, J.; Chen, B.; Milik, M.; Sukuru, S. C. K.; Bender, A.; Mikhailov, D.; Whitebread, S.; Hamon, J.; Azzaoui, K.; Urban, L.; Glick, M.; Davies, J. W.; Jenkins, J. L. Gaining Insight into Off-Target Mediated Effects of Drug Candidates with a Comprehensive Systems Chemical Biology Analysis. *J. Chem. Inf. Model.* **2009**, *49* (2), 308−317.

(5) (a) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25* (2), 197−206. (b) Keiser, M. J.; Setola, V.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462* (7270), 175. (c) Xie, L.; Xie, L.; Kinnings, S. L.; Bourne, P. E. Novel Computational Approaches to Polypharmacology as a Means to Define Responses to Individual Drugs. *Annu. Rev. Pharmacol. Toxicol.* **2011**, *52*, 361−379.

(6) ChEMBL. https://www.ebi.ac.uk/chembldb/ (accessed August, 2011).

(7) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35* (suppl 1), D198−D201.

(8) Hu, Y.; Bajorath, J. r. Polypharmacology Directed Compound Data Mining: Identification of Promiscuous Chemotypes with Different Activity Profiles and Comparison to Approved Drugs. *J. Chem. Inf. Model.* **2010**, *50* (12), 2112−2118.

(9) (a) Coan, K. E. D.; Maltby, D. A.; Burlingame, A. L.; Shoichet, B. K. Promiscuous Aggregate-Based Inhibitors Promote Enzyme Unfolding. *J. Med. Chem.* **2009**, *52* (7), 2067−2075. (b) Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K. High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.* **2005**, *1* (3), 146−148. (c) Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; Austin, C. P.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A. Quantitative Analyses of Aggregation, Autofluorescence, and Reactivity Artifacts in a Screen for Inhibitors of a Thiol Protease. *J. Med. Chem.* **2009**, *53* (1), 37−51. (d) McGovern, S. L.; Helfand, B. T.; Feng, B. Y.; Shoichet, B. K. A specific mechanism of nonspecific inhibition. *J. Med. Chem.* **2003**, *46*, 4265−4272. (e) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53* (7), 2719−2740.

(10) Yan, S. F.; King, F. J.; He, Y.; Caldwell, J. S.; Zhou, Y. Learning from the Data: Mining of Large High-Throughput Screening Databases. *J. Chem. Inf. Model.* **2006**, *46* (6), 2381−2395.

(11) Zhou, Y.; Zhou, B.; Chen, K.; Yan, S. F.; King, F. J.; Jiang, S.; Winzeler, E. A. Large-Scale Annotation of Small-Molecule Libraries Using Public Databases. *J. Chem. Inf. Model.* **2007**, *47* (4), 1386−1394.

(12) MeSH. http://www.nlm.nih.gov/mesh/ (accessed August, 2010).

(13) GeneGO. http://www.genego.com/ (accessed August, 2011).

(14) PubMed. http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed (accessed August, 2011).

(15) Csizmadia, F. JChem: Java Applets and Modules Supporting Chemical Database Handling from Web Browsers. *J. Chem. Inf. Comp. Sci.* **2000**, *40* (2), 323−324.

(16) Zhou, Y.; Young, J. A.; Santrosyan, A.; Chen, K.; Yan, S. F.; Winzeler, E. A. In silico gene function prediction using ontology-based pattern identification. *Bioinformatics* **2005**, *21* (7), 1237−1245.

(17) Yan, S. F.; King, F. J.; Chanda, S. K.; Caldwell, J. S.; Winzeler, E. A.; Zhou, Y., Mining High-Throughput Screening Data by Novel Knowledge-Based Optimization Analysis. In *Pharmaceutical Data Mining*; John Wiley & Sons, Inc.: New York, 2009; pp 205−233.

(18) Plouffe, D.; Brinker, A.; McNamara, C.; Henson, K.; Kato, N.; Kuhen, K.; Nagle, A.; Adrián, F.; Matzen, J. T.; Anderson, P.; Nam, T.-g.; Gray, N. S.; Chatterjee, A.; Janes, J.; Yan, S. F.; Trager, R.; Caldwell, J. S.; Schultz, P. G.; Zhou, Y.; Winzeler, E. A. In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (26), 9059−9064.

(19) Yan, S. F.; Asatryan, H.; Li, J.; Zhou, Y. Novel Statistical Approach for Primary High-Throughput Screening Hit Selection. *J. Chem. Inf. Model.* **2005**, *45* (6), 1784−1790.

(20) (a) Visser, U.; Abeyruwan, S.; Vempati, U.; Smith, R.; Lemmon, V.; Schurer, S. BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. *BMC Bioinf.* **2011**, *12* (1), 257. (b) Schürer, S. C.; Vempati, U.; Smith, R.; Southern, M.; Lemmon, V. BioAssay Ontology Annotations Facilitate Cross-Analysis of Diverse High-Throughput Screening Data Sets. *J. Biomol. Screening* **2011**, *16* (4), 415−426.

(21) (a) Muegge, I.; Heald, S. L.; Brittelli, D. Simple Selection Criteria for Drug-like Chemical Matter. *J. Med. Chem.* **2001**, *44* (12), 1841−1846. (b) Lipinski, C. A.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25. (c) Walters, W. P.; Ajay, A. A.; Murcko, M. A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* **1999**, *3*, 384−387.

(22) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Patel, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2008**, *26* (1), 127−132.

(23) Lacrampe, J.; Fernand, A.; Connors, W. R.; Ho, C. Y.; Richardson, A. G.; Freyne, E.; Edgard, J.; Buijnsters, P.; Jacobus, J.; Bakker, C. A. *3-Phenyl analogs of toxoflavine as kinase inhibitors*. WO/2004/007498, Jan. 22, 2004.

(24) Crisman, T. J.; Parker, C. N.; Jenkins, J. L.; Scheiber, J.; Thoma, M.; Kang, Z. B.; Kim, R.; Bender, A.; Nettles, J. H.; Davies, J. W.; Glick, M. Understanding False Positives in Reporter Gene Assays: in Silico Chemogenomics Approaches To Prioritize Cell-Based HTS Data. *J. Chem. Inf. Model.* **2007**, *47* (4), 1319−1327.

(25) Burniston, J.; Ellison, G.; Clark, W.; Goldspink, D.; Tan, L.-B. Relative toxicity of cardiotonic agents. *Cardio. Toxicol.* **2005**, *5* (4), 355−364.

(26) Karlsson, L. S.; Siquan; Rao, N. L.; Venable, J.; Thurmond, R. TLR7/9 Antagonists as Therapeutics for Immune-Mediated Inflammatory Disorders. *Inflamm. Allergy−Drug Targets* **2007**, *6*, 223−235.

(27) Van Beek, M. J.; Piette, W. W. Antimalarials. *Dermatol Clin* **2001**, *19* (1), 147−160.

(28) Stathis, A.; Oza, A. Targeting WEE1-like protein kinase to treat cancer. *Drug News Perspec.* **2010**, *23* (7), 425−429.

(29) Gazit, A.; Yaish, P.; Gilon, C.; Levitzki, A.; Tyrphostins, I synthesis and biological activity of protein tyrosine kinase inhibitors. *J. Med. Chem.* **1989**, *32* (10), 2344−2352.
(30) (a) Attisano, L.; Wrana, J. L. Signal Transduction by the TGF-β Superfamily. *Science* **2002**, *296* (5573), 1646−1647. (b) Massagué, J. TGF-β Signal Transduction. *Annu. Rev. Biochem.* **1998**, *67* (1), 753−791. (c) Roberts, A.; Mishra, L. Role of TGF-[beta] in stem cells and cancer. *Oncogene* **2005**, *24* (37), 5667−5667.
(31) 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.* **2001**, *46* (1−3), 3−26.
(32) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2006**, *47* (1), 47−58.
(33) Daunorubicin, Cytarabine, and Midostaurin in Treating Patients With Newly Diagnosed Acute Myeloid Leukemia. http://www.clinicaltrials.gov/show/NCT00651261 (accessed March 8, 2012).
(34) Phase II PKC412 in aggressive systemic mastocytosis and mast cell leukemia. http://www.clinicaltrials.gov/ct2/show/NCT00233454?term=Midostaurin&rank=10 (accessed March 8, 2012).