Understanding False Positives in Reporter Gene Assays: in Silico Chemogenomics Approaches To Prioritize Cell-Based HTS Data

Thomas J. Crisman,[†] Christian N. Parker,[‡] Jeremy L. Jenkins,[†] Josef Scheiber,[†] Mathis Thoma,[†] Zhao Bin Kang,[†] Richard Kim,[†] Andreas Bender,[†] James H. Nettles,[†] John W. Davies,[†] and Meir Glick*[,†]

Lead Discovery Center, Novartis Institutes for Biomedical Research Inc., 250 Massachusetts Avenue, Cambridge, Massachusetts 02139, and Lead Discovery Center, Novartis Pharma AG, Postfach, CH-4002 Basel, Switzerland

Received December 11, 2006

High throughput screening (HTS) data is often noisy, containing both false positives and negatives. Thus, careful triaging and prioritization of the primary hit list can save time and money by identifying potential false positives before incurring the expense of followup. Of particular concern are cell-based reporter gene assays (RGAs) where the number of hits may be prohibitively high to be scrutinized manually for weeding out erroneous data. Based on statistical models built from chemical structures of 650 000 compounds tested in RGAs, we created "frequent hitter" models that make it possible to prioritize potential false positives. Furthermore, we followed up the frequent hitter evaluation with chemical structure based in silico target predictions to hypothesize a mechanism for the observed "off target" response. It was observed that the predicted cellular targets for the frequent hitters were known to be associated with undesirable effects such as cytotoxicity. More specifically, the most frequently predicted targets relate to apoptosis and cell differentiation, including kinases, topoisomerases, and protein phosphatases. The mechanism-based frequent hitter hypothesis was tested using 160 additional druglike compounds predicted by the model to be nonspecific actives in RGAs. This validation was successful (showing a 50% hit rate compared to a normal hit rate as low as 2%), and it demonstrates the power of computational models toward understanding complex relations between chemical structure and biological function.

INTRODUCTION

High throughput screening is a common way for modern drug companies to discover active compounds for lead optimization projects.[1] Screens are divided into biochemical (e.g., enzymatic or ligand-binding) and cell-based assays. In biochemical assays the percent change of activity (either inhibition or activation) caused by a certain compound is measured in vitro using a purified protein target in a defined buffer system. Cell-based assays measure the effect of a compound on a particular readout within the context of the entire cell. There are various cell-based screening formats and assays.[2] This work will focus on reporter gene assays (RGA) where the activity of a compound eventually depends on how much it affects the expression of certain reporter genes.[3] Although RGA screens do not identify the specific target being modulated, they do reveal the effect of the compound on gene expression as a downstream effect along the pathway they have been designed to monitor. All HTS data can be noisy in terms of false positives and negatives,[4] and cell-based assays are thought to be noisier than enzymatic screens. This assumption is made because cell-based screens tend to have higher hit rates than enzymatic ones *but with lower confirmation rates*. Given the time, expense, and depletion of compound material involved in validation, further attention to triaging hitlists from cell-based screens is warranted.

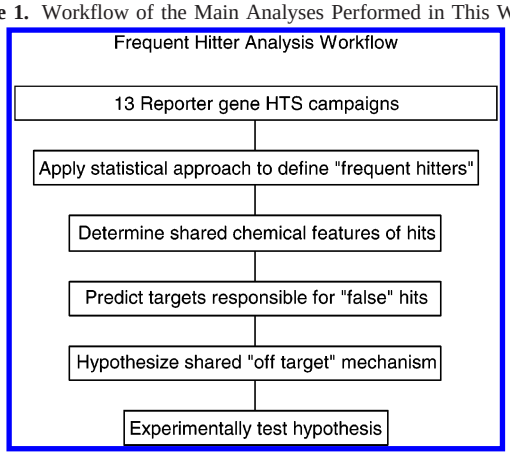
Assay artifacts are inherent to the assay format and the chemical structure being tested for both the case of biochemical and cell-based assays; compounds that are reactive, fluorescent, quenchers, aggregators, chelators, and reducing tend to read as false positives or negatives due to their ability to affect cells, the target protein, or the detection method.[5] Any assay has error sources, such as reagent and temperature errors or screening technology errors in liquid handling, reading, and time. Errors related to storage conditions such as compound precipitation from DMSO stock solutions are also possible.[6] The above mechanisms of errors and artifacts will not be discussed here. While it is possible to exclude, for example, reactive or unstable compounds, by filtering out undesirable substructures or with more sophisticated statistical models, the current study focuses on other types of frequent hitters which, in practice, are more difficult to triage, namely those based on off-target activities.

It should be noted that the exact definition of a frequent hitter is a subject of debate, although their existence is not disputed. One study defines frequent hitters as compounds occurring in 7 of 15 assays,[7] while another defines them as compounds which appear as hits in 8 of 161 assays.[8] Numerous experiments have been done to discover why certain compounds are promiscuous. For example, aggregators can be flagged with biochemical assays as suggested by Shoichet's group.[9−12] These studies focus on a general,

* Corresponding author phone: (617)871-7130; e-mail: meir.glick@novartis.com.
[†] Novartis Institutes for Biomedical Research Inc.
[‡] Novartis Pharma AG.

10.1021/ci6005504 CCC: $37.00
Published on Web 07/04/2007

Scheme 1. Workflow of the Main Analyses Performed in This Work[a]



[a] This is the collection of data from 13 reporter gene assay HTS campaigns: the validation of a valid "frequent hitter" definition; the analysis of their shared chemical features; a prediction of targets enriched among reporter gene assay frequent hitters; and an experimental validation of the target hypotheses.

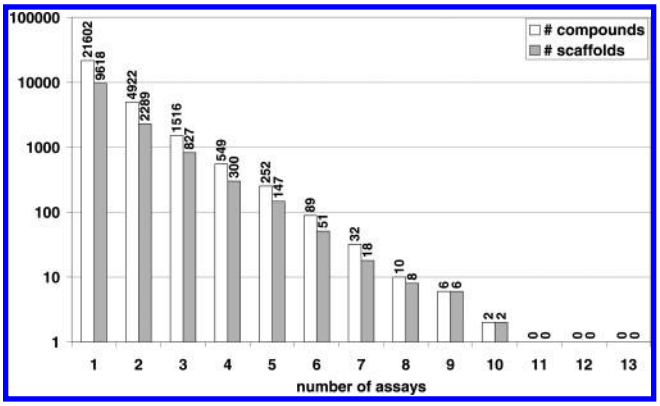


Figure 1. Distribution of frequent hitters (compounds and scaffolds) in 13 RGA data sets. The number of unique molecules or scaffolds which were determined to be active in an assay are plotted versus the number of assays in which they were discovered to be active. While most of the compounds are specific actives in single or few assays, also more promiscuous frequent hitters can clearly be observed.

nontarget-specific mechanism of promiscuity, relevant only to biochemical assays. In contrast, the current study focuses on a novel classification of frequent hitters based upon their disruptive off-target cellular activities that influence reporter gene read out. A number of promiscuous mechanisms by which compounds could interact with cells influencing RGAs can be envisioned: general cytotoxicity,[13] modulation of gene expression efficiency,[14] modulation of the reporter enzyme activity (e.g., for the commonly used luciferase reporter gene[15]), or color shifting.

In silico methods of triaging compounds are numerous and diverse, from simple quantitative properties such as the Lipinski rule of 5[16] to more sophisticated statistical models built on molecular fingerprints.[17] Recently, Laplacian-modified Naïve Bayesian classifiers have proven to be useful in creating models which can work well even with noisy data.[18] Further, multiple-Category Bayesian Models can be trained on annotated commercial data sources such as WOMBAT,[19] stARLITe,[20] Jubilant BioSys database,[21] and GVK Biosciences[22] databases. Nidhi and co-workers[23] have shown that such models are effective for associating specific targets with orphan ligands. In this study, we apply a similar in silico target fishing approach to flag primary hits that cannot otherwise be visually inspected by a medicinal chemist in a timely fashion.

The outline of this work is schematically represented in Scheme 1. The first step is the collection of data from 13 reporter gene assay HTS campaigns, followed by the validation of a frequent hitter definition relevant in practice. Next, we analyze shared chemical features of the frequent hitters and predict targets enriched among this group of compounds—targets we propose to be among the causes for encountering frequent hitters in reporter gene assays. The last step is an experimental validation of the target hypotheses on a new data set.

METHODS

Reporter Gene Assay Data Sets. The data set used in the present study was a 650 000 compound subset of our screening collection and their associated activity from 13 different reporter gene HTS campaigns. The screening format in each case was a RGA which tested for antagonists of specifically stimulated molecular pathways. Various promoters were used to select the pathway, but firefly luciferase was always the reporter gene and always in mammalian cell lines (e.g., Jurkat, CHO, CCL39, C2C12, MC3T3, BL2, UMR106). The 13 diverse RGAs assayed inhibition of transcription, cell stimulation, GPCR activity, pathway signaling, protein secretion, mRNA stability, enzyme inhibition, and both nuclear and membrane receptor binding. The primary screens consisted of one compound per well at a concentration of 2.5−20 μM. Activity was defined as >30−60% inhibition at the screening concentration. The 13 training sets contained a total of 41 006 (28 933 unique) molecular hits. These compounds were either purchased from external vendors based on drug likeness or were synthesized in-house through lead optimization projects and exploratory chemistry.[24] Figure 1 depicts the number of screens in which each compound or scaffold were active out of the 13 training data sets. Scaffolds in the HTS sets were determined using the "Murcko Assemblies" component in Pipeline Pilot based on the work of Bemis and Murcko.[25] Although this type of frequent hitter analysis shows that there are compounds which frequently occur in different screening assays, it provides little prospective value since their number is very small and, further, does not explain why any compound is a frequent hitter.

Databases. The World of Molecular Bioactivity Database (WOMBAT)[19] was used to train our Multiple-Category Bayesian Models. WOMBAT 2006.1 contains 154 236 entries (136 091 unique SMILES) with biological activity on 1320 unique targets. It was built on 6801 different series from 6791 different papers published in medicinal chemistry journals between 1975 and 2005. The MDL Drug Data Report (MDDR)[26] is a database which contains over 132 000 molecules which have biological relevance, and each yearly update adds approximately 10 000 molecules.

Training and Prediction of Targets Using Multiple-Category Bayesian Models. Laplacian-Modified Naïve Bayesian Classification Models were built for each individual target and target family using components from Pipeline Pilot (SciTegic, Inc.). These models assume all variables are independent and use a Laplacian correction to reduce the

bias caused through descriptors less prevalent in the data set. The derivation of the Laplacian-modified Bayesian models has been described previously.[23] All models which were built contain multiple probabilities for multiple features. The overall probability ($P_{combined}$) for a compound having activity for a particular target is determined by summing the logs of the probabilities of activity for each feature.[27] Extended Connectivity Fingerprints (ECFPs) were used as chemical descriptors,[28,29] since it has been established that Bayesian models built using circular fingerprints work well in virtual screening tasks[30] and also with primary screening results.[18] Other recent work on target prediction for small molecules employed 3D descriptors[31] which were shown to perform better if structures with high 2D similarity are not known. Current target prediction methods were also subject of a recent review.[32]

Prospective Use of the Frequent Hitter Model on an External Data Set. The 132 000 MDDR compound were ranked according to the best frequent hitter Bayesian model defined earlier (compounds that hit in 3 or more RGAs) and annotated with targets using our target fishing model. From this list, 160 compounds (selected as a manageable number for an experimental procedure) that were ranked with a score of over 20 for at least one of the top 25 overenriched targets and which did not appear in the WOMBAT training set were selected to test prospectively. (A Bayes score of above 20 for a target prediction has, based on our experience, been shown to possess a certain confidence.) Our primary hypothesis was that molecules selected by chemical features would have a tendency to appear active in any RGA with luciferase activity inhibition as the readout. Our secondary hypothesis was that this was due to cytotoxicity.

Experimental Evaluation of the Predicted RGA Activity. Compound activity was measured in two additional cell based assays: an RGA not used in the original modeling exercise and a non-RGA bacterial growth inhibition assay. This second assay was included as a means of determining if the compounds were acting against generic biological mechanisms such as glycolysis or membrane disruption.

Experimental Evaluation of Predicted Cell Viability. Inhibition of cell growth (in vitro) was measured using the reduction of the tetrazolim dye resazurin by functional mitochondria as a means to assess cell viability.[33] These experiments were conducted using both bacterial and mammalian cell lines under conditions, such as compound concentration and incubation time, similar to the reporter gene assays. The growth inhibition assays used the same or similar cell lines as those used in the RGA assays.

RESULTS AND DISCUSSION

The outline of this work is represented in Scheme 1. The first step is the collection of data from 13 reporter gene assay HTS campaigns, followed by the validation of a frequent hitter definition relevant in practice which we discuss in the following paragraph.

Generation of Models To Define Reporter Gene Assay Frequent Hitters. Figure 1 illustrates the distribution of individual compounds or chemical scaffolds having overlapping activity in multiple assays. It can be seen that the number of individual compounds with activity in a large number of assays (e.g., more than four or five) is rather small.

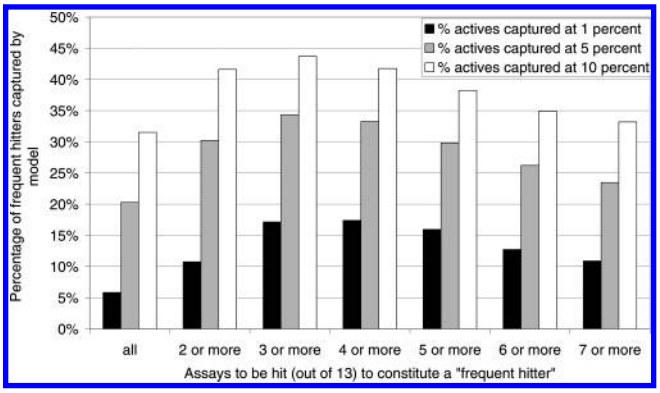


Figure 2. Recall rates of RGA active compounds seeded into the entire Novartis collection by the RGA frequent hitter model at varying cutoffs of assays used to determine frequent hitter status for model training. For example, "7 or more" indicates that the compounds used to train the model were active in 7 or more historical RGA assays.

Simply removing these would thus have little impact on biological followup. We were interested in exploring a shared mechanism of this nonspecific behavior that could be generalized to new compounds. However, when separating the data set into frequent hitters and nonfrequent hitters, the question of where to set the cutoff between the classes needs to be answered. Therefore, a number of statistical models were generated to determine which cutoff should be used to create *the most predictive* model. The models varied in the number of assays in which the compounds were found to be active: For example, the definition of frequent hitter for the first model was "compounds which hit in *one or more* assays", the next model assigned this property to compounds "which hit in *two or more* assays", etc., up to the most stringent definition of frequent hitters as those "which hit in *seven or more* assays". The entire Novartis Screening Collection was used as a background compound distribution for model training.

Figure 2 depicts the performance of each model for recalling all RGA actives from the entire Novartis Screening Collection as a background. The percentage of actives found in the top 1%, 5%, and 10% of the total ranked data set is shown. The most striking result from this figure was that the models built on frequent hitters from "three or more" or "four or more" assays captured 17−18% of the test set actives in the first 1% suggested by the model. Increasing the number of active assays did not improve this metric of model performance. Figure 3 shows the area below the receiver operating characteristic (ROC) curve for each model. ROC curves provide a quantitative measure of enrichment which allows for easier comparison across models. ROC scores of 0.5 indicate random recall, and a ROC score of 1 indicates perfect recall. In this case, the ROC score indicates the ability of each frequent hitter model to rank RGA actives more highly than random compounds. Again, the models based on a frequent hitter definition as compounds hitting in three or more assays or in four or more assays performed similarly well, with ROC scores in the 0.75 range. The ability to generate models for reporter gene assays frequent hitters corroborates our hypothesis that frequent hitters share features in chemical space. Because the model built on compounds active in three or more assays performed best, compounds active in three or more assays are heretofore

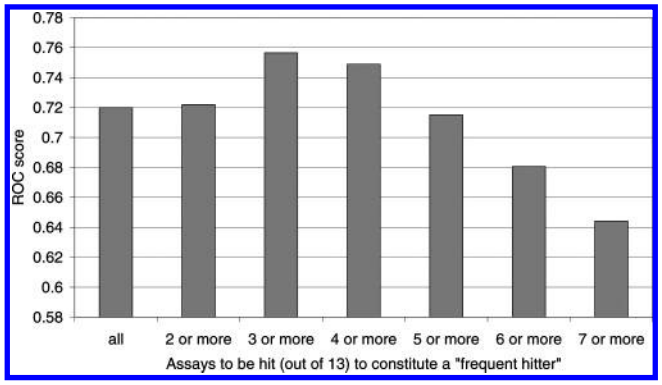


Figure 3. Areas below the ROC curves of RGA active compounds seeded into the entire Novartis collection by the RGA frequent hitter model at varying cutoffs.

considered to be frequent hitters for the evaluation purposes of the study.

In order to gain insight into the nature of what actually makes the frequent hitters similar as a group we used a complimentary in silico method to annotate potential off-target cellular targets for these RGA hits.

Prediction of Targets for Reporter Gene Assay Frequent Hitters. It was postulated that the source of bias in RGA frequent hitters may be particular off-target proteins that, when modulated, lead to disruption of the RGA readout, creating a false positive signal. Therefore an attempt to predict the most likely target class for each RGA compound hit was made using a WOMBAT-trained Multiple-Category Bayesian Model described previously.[23] Since the model was trained for 1320 different potential targets, 1320 different scores were calculated for each compound. Bayes scores are relative and vary from model to model, where a higher score means that the compound is more likely to be correctly classified by the model. While the target prediction model employed here probably does not provide correct structure-target associations in each single case, overall validation of the method[23] gave the correct target among the top 3 guesses in about 77% of all cases. The top five target classifications were kept for each compound in both the RGA data set and the entire Novartis screening collection. Targets ranking in the top five for a compound with a Bayes score higher than 20 were kept. The criteria of a Bayes score above 20 and a target number cutoff of 5 was selected based on previous work with the multiple class models.[23] While each individual prediction might contain uncertainty, the large number of predictions employed in the current work should nonetheless be able to point out tendencies regarding the nature of more frequently predicted targets, as opposed to others.

To separate the frequency of target prediction that is specific to these RGA hits from that inherent in our corporate collection's limited chemical space, we used the same target prediction model for our entire library. Table 1 describes the resulting target enrichment for RGA frequent hitters, where the definition of a frequent hitter is a compound that was a hit in 3 or more of 13 hitlists in this study. Target enrichment is the ratio between the target's predicted prevalence for RGA frequent hitters and the target's predicted prevalence in the entire Novartis screening collection. For the RGA data set, a particular target class must be predicted a minimum of 20 times in order to be considered significant, and all targets which were not represented by at least 20 compounds in the RGA screens were removed. Only targets enriched by more than 4-fold in the RGA set versus the background set are shown in Table 1. The most frequent targets enriched for the RGA frequent hitters are those which relate to apoptosis, cell differentiation, or other cell-wide functions. It is noteworthy that an estimated 20−30% of pharmaceutical discovery programs currently focus on kinases,[34] leading to a natural abundance of compounds in the collection favoring kinase modulation. This underscores the importance of knowing the predicted targets for the entire Novartis compound collection and employing them as a baseline to which the targets of the RGA frequent hitters are compared. Even by employing this kinase-heavy baseline, the most "RGA-biased" targets are kinases (Table 1). Mixed lineage kinases (MLK) are responsible for signaling events that have been reported to lead to apoptotic cell death,[35,36] and MLK overexpression has been documented in tumor cells.[37] Also highly predicted are cyclin dependent kinases (CDK). Cyclin dependent kinases regulate cell cycle, and inhibiting CDKs causes arrests in cell growth and mitotic division.[38,39] Topoisomerase II inhibitors were also heavily predicted for the frequent hitters. Topoisomerase II function is clearly linked to DNA replication and transcription by cleaving strands of DNA to relieve strain and allow for easier transcription.[40] Protein phosphatase 1 (PP1) and protein

Table 1. Enrichment of Targets Predicted for RGA Frequent Hitters versus the Entire Novartis Screening Collection

WOMBAT06 target name	% of frequent hitters	% of Novartis screening collection	target enrichment
TrkA	2.275	0.019	119.10
MLK-3	0.846	0.016	54.24
PKA	2.187	0.043	51.20
PKC	2.479	0.054	45.69
MLK-1	0.671	0.015	43.79
CaMPKase II	0.233	0.007	32.84
PP2A	0.233	0.008	29.75
PP1	0.233	0.008	27.79
PKC-beta-II	0.350	0.014	24.79
PSA	0.204	0.010	20.30
Na(+)/(K+)-ATPase	0.262	0.013	19.62
POTA	0.146	0.008	18.38
CDK4	1.429	0.095	15.01
CDK2	1.487	0.108	13.73
MTL-R	0.262	0.022	11.71
FPR1	0.321	0.028	11.59
H(+)/K(+) ATPase	1.779	0.207	8.60
CDK1	2.712	0.316	8.59
KinA	0.962	0.119	8.10
v-Abl	2.391	0.322	7.42
NOXase	0.175	0.024	7.18
FcERI	0.146	0.020	7.12
CDK4/cyclin D1	0.146	0.021	6.78
ERG2	0.175	0.030	5.82
topoisomerase II	0.292	0.053	5.53
FGFR-2	0.437	0.080	5.47
Na(+)/K(+) ATPase	0.787	0.153	5.15
CDK4/cyclin D	1.896	0.396	4.79
AdoMetDC	0.204	0.045	4.52
mAb 1B3	0.846	0.188	4.50
L-CA	0.146	0.033	4.40
CXCR	0.496	0.113	4.37
Tie-2	0.379	0.088	4.29
EGFR	2.479	0.589	4.21
GTPase	0.262	0.063	4.16
VEGFR-2	3.295	0.800	4.12
NET	0.350	0.086	4.08
V-ATPase	0.496	0.123	4.02

Table 2. Enrichment of Target Families Predicted for RGA Frequent Hitters versus the Entire Novartis Screening Collection

WOMBAT06 family name	% of frequent hitters	% of Novartis screening collection	family overenrichment
peptidase M14, hydrolase; peptidase M10A; peptidase M10A, hydrolase; metalloprotease	9.54	0.04	233.2
serine/threonine-protein kinase	9.89	0.05	201.1
phosphatase	0.99	0.01	87.9
ATPase alpha/beta chains	4.25	0.06	72.9
heat shock protein 90	2.03	0.03	68.0
serine/threonine-protein phosphatase, hydrolase; peptidase M10A; peptidase M10A, hydrolase	1.07	0.04	29.9
carnitine/choline acetyltransferase	1.39	0.06	23.0
serine/threonine-protein kinase, transferase	30.23	1.36	22.2
LDLR	0.83	0.04	21.2
P-type ATPase	4.61	0.22	20.9
type II topoisomerase	2.58	0.14	18.7
androgen carrier	1.51	0.08	18.2
choline/ethanolamine kinase	1.23	0.07	17.2
enzymeserine/threonine-protein kinase, transferase	1.83	0.11	17.2
uracil-DNA glycosylase, glycosidase; hydrolase; peptidase M10A; peptidase M10A, hydrolase	3.10	0.19	16.6
NUDIX hydrolase	1.63	0.10	16.3
amino acid-polyamine-organocation (APC) transporter	1.19	0.08	15.8
zinc-containing alcohol dehydrogenase	1.11	0.07	15.7
rRNA adenine N-6-methyltransferase	5.36	0.38	14.0
tubulin	4.73	0.34	14.0
eukaryotic AdoMetDC, lyase	1.19	0.09	14.0
peptidase M24, hydrolase; peptidase M10A; peptidase M10A, hydrolase	4.65	0.33	13.9
transcription factor	1.07	0.08	13.6
equilibrative nucleoside transporter	1.07	0.08	13.5
tyrosine-protein kinase, transferase	27.93	2.15	13.0
polypeptide deformylase, hydrolase; peptidase M10A; peptidase M10A, hydrolase	0.99	0.08	12.2

phosphatase 2A (PP2A) are major protein phosphatases in number and importance; inhibiting them alters their ability to dephosphorylate serine-threonine residues that occur in approximately 30% of proteins in the cell.[41,42] Thus, it is reasonable that chemotypes modulating these targets are highly represented among frequent hitters in cell-based screens. We also acknowledge that several of the kinase-inhibiting compounds shown here, particularly the staurosporine chemotypes, are known pan-kinase inhibitors. Further, the podophyllotoxin chemotypes predicted to inhibit topoisomerase II are known inhibitors of both topoisomerase II and tubulin polymerization. Therefore, despite the enrichment of specific predicted targets, it is conceivable that multiple off-target binding events contribute to the cellular effects that yield promiscuous RGA activity.

In addition to predicted target enrichment, we compared predicted target *family* enrichment (Table 2). The rationale for this analysis is that while a given target may not have been significantly enriched in Table 1, its respective family may be enriched when the target is taken as a whole with other members of its family. Only the highest ranked family prediction was used for each compound. Enrichment ratios lower than 12 are not shown. The results of the target family comparison were largely consistent with the individual target data with one large exception, the addition of protease inhibitors. These family types consist largely of MMP, carboxypeptidase, and amino peptidase inhibitors. Although the exact method of inhibition is not known, proteasome inhibitors have been shown to reduce luciferase activity generically.[43]

Table 3. Enrichment of Targets Predicted for RGA Frequent Hitters versus All RGA Hits

WOMBAT06 target name	times more expressed in frequent hitters	WOMBAT06 target name	times more expressed in frequent hitters
TrkA	6.31	PSA	2.64
MLK-3	2.73	Na(+)/(K+)-ATPase	3.74
PKA	5.41	CDK4	2.72
PKC	5.42	CDK2	2.20
MLK-1	2.20	MTL-R	2.21
CaMPKase II	3.01	FPR1	2.88
PP2A	4.19	H(+)/K(+) ATPase	2.77
PP1	3.86	CDK1	2.11
PKC-beta-II	4.52		

Table 4. Physicochemical Property Distributions of RGA Hits versus the Novartis Screening Collection[a]

	RGA		Novartis	
	mean	SD	mean	SD
ALogP	3.9	2.1	3.3	1.9
molecular weight	396.4	135.3	375.6	114.2
H-bond donor	1.6	1.4	1.4	1.3
H-bond acceptor	4.5	2.4	4.3	1.9
solubility	−5.9	2.6	−5.3	2.3
polar surface area	75.2	43.6	75.5	38.2

[a] Solubility is 'aqueous solubility' and is expressed as Log*S*, where *S* is equal to the solubility in mol/L. This calculation is based on the 'electrotopological state indices' published by Tetko et al.[44] The polar surface area was calculated utilizing methods created by Ertl, Rohde, and Selzer.[45]

The significant target enrichments for frequent hitters strongly suggests that off-target effects may be responsible in some cases for false-positive activity in the cell-based screen. Table 3 compares the target prediction prevalence of the RGA frequent hitters versus target prediction performed on *all* RGA hits. Again, proteins regulating cell cycle, signal transduction cascades, and other critical cell events are enriched. This indicated that the overenrichment of these particular cytotoxicity-related targets versus the entire collection was a result of features found more commonly among frequent hitters than general RGA actives.

Physicochemical Property Distributions Are Not Responsible for Frequent Hitters in Reporter Gene Assays. Next, we looked at whether biases in RGA hits could be attributed to simple differences in physiochemical properties (Table 4). In contrast to the target prediction results, we did not find statistically significant difference between physicochemical properties of RGA hits and the Novartis screening collection, suggesting that off-target effects were a more realistic source of false positive signals.

Prospective Validation of the Frequent Hitter Model on an External Data Set. In order to test the predictive power of this methodology, 132 000 compounds from the MDDR were initially ranked according to the frequent hitter Bayesian model. Here, we used the optimal model (Figures

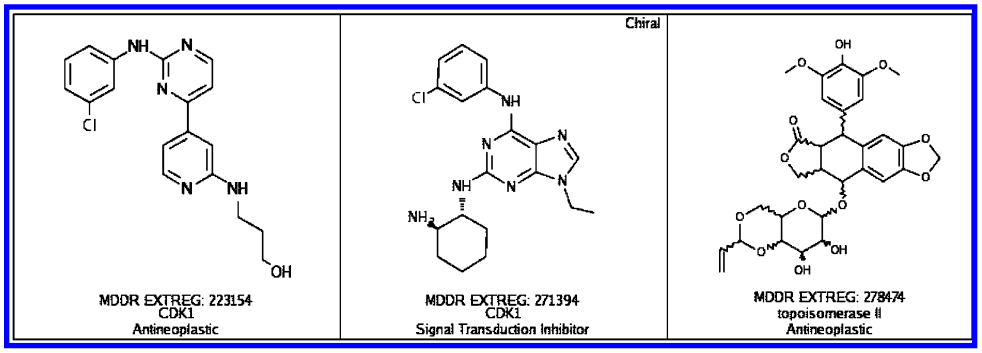


Figure 4. MDDR compounds predicted and then confirmed experimentally to be RGA frequent hitters. The predicted target name based on the WOMBAT models and the known MDDR activity class are depicted.

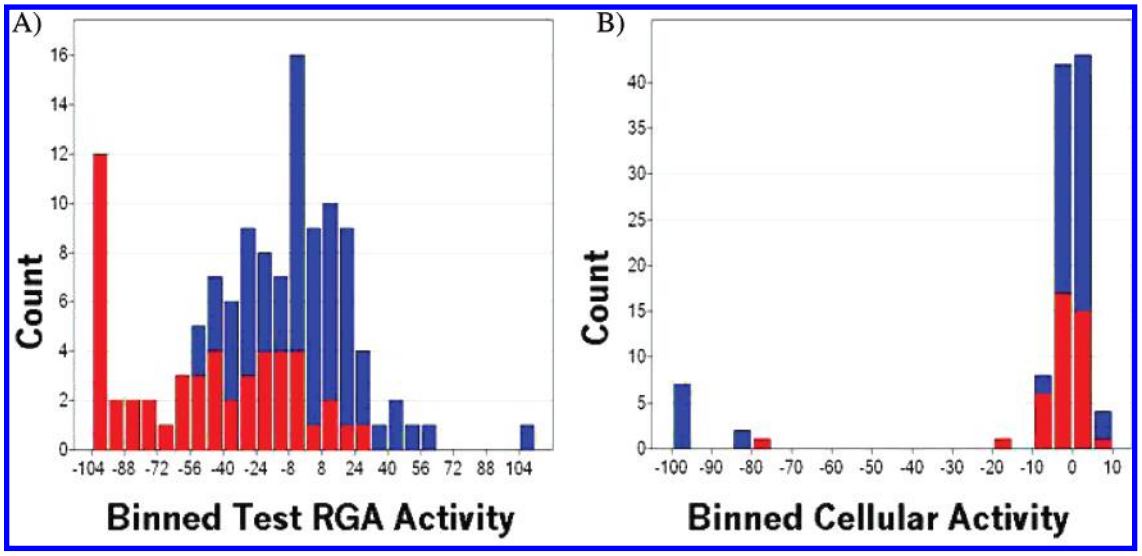


Figure 5. Activity frequency distributions for compounds predicted to have targets associated with RGA frequent hitters. The activity is described as % inhibition where more a negative value means a stronger inhibition. Compounds colored red are cytotoxic in mammalian cells and compounds colored blue are nontoxic (see Figure 7). A) The results of the selected compounds in an additional RGA using a luciferase read-out but with a target unrelated to any of the 13 RGAs used to train the models. B) The results for the selected compounds in an antibacterial growth inhibition assay where the majority of compounds were found to be inactive.

2 and 3), which was trained on compounds that hit in 3 or more RGAs. From this list, 160 compounds (selected as a manageable number for an experimental procedure) that were ranked with a score of over 20 for at least one of the top 25 overenriched targets and which did not appear in the WOMBAT training set were selected to test prospectively. Figure 4 depicts sample RGA MDDR frequent hitter compounds identified by the model. The MDDR Biology field describes compound EXTREG 223154 as an "antineoplastic agent that potently and selectively inhibits the cyclin-dependent protein kinases cdk1 and cdk2 (IC50 = 20 and 50 nM, respectively)". Likewise, EXTREG 271394 is an "antineoplastic agent, a potent cyclin-dependent type 1 kinase (CDK1) inhibitor (IC50 = 24 nM) with high selectivity over other kinases such as protein kinase C-alpha (PKC-alpha), protein kinase A (PKA) and epidermal growth factor (EGF) protein kinase (IC50 = 6.1, 125 and > 10 μM, respectively)". EXTREG 278474 is being described as "more potent than the reference in inhibiting topoisomerase II (IC50 = 2.0 mcg/ml vs 10.0 mcg/ml for etoposide)".

Test compound activities were initially measured in two cell-based assays: an RGA not used in the original modeling exercise (Figure 5A) and a non-RGA bacterial growth inhibition assay (Figure 5B). This second assay was included as a means of determining if the compounds were acting against generic biological mechanisms such as glycolysis or membrane disruption. Like the previous 13 RGAs used for model building, the test RGA consisted of a mammalian cell line and relied on a decrease in luciferase expression and signal to measure inhibition; however, the reporter gene was under a different promoter system from any used in the

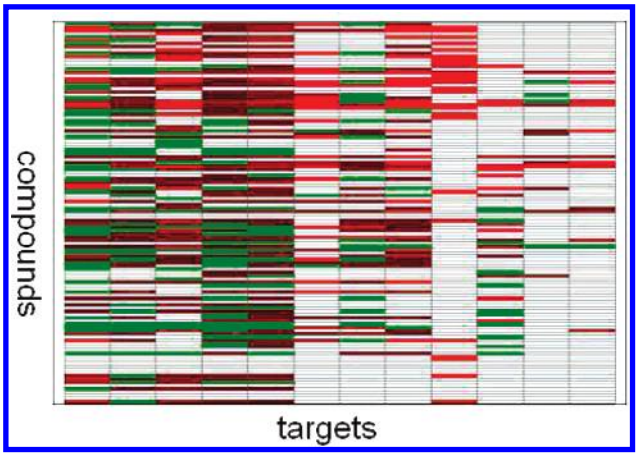


Figure 6. Activity heat map for compounds predicted to have targets associated with RGA frequent hitters at the same RGA assays that were used for model building. The activity is described as % inhibition where compounds colored red are strong inhibitors. Compounds colored green are inactive, and compounds colored white were not tested.

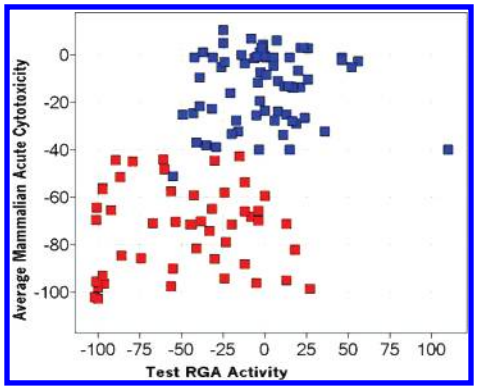


Figure 7. Average cytotoxicity versus RGA activity. The activity (X-axis) and cytotoxicity (Y-axis) are described as % inhibition where a more negative value means a stronger inhibition. In this figure the activity of the 160 test compounds was predicted to show RGA frequent hitter activity, their activity in an additional RGA assay, and their average activity in two assays of acute mammalian cytotoxicity.

previous assays. Thus, the assay served to test the generalizability of the frequent hitter and target prediction models. In this RGA the selected compounds yielded a disproportionate number of actives: approximately 50% of the compounds were active as compared to the overall hit rate of ~2% for the complete compound collection. These results indicate that compounds which frequently hit in RGA with a luciferase inhibition readout are highly predictable in mammalian cells irrespective of the promoter being studied and that RGA hits are prioritizable by this approach. Intriguingly, this shifted activity distribution was only seen for the compounds in the RGA and not in the bacterial growth inhibition assay (Figure 5B), suggesting that the source of the promiscuous inhibition in RGAs is not general cellular effects of the compounds applicable across eukaryotic and prokaryotic cell lines but rather specific effects on mammalian cells. The 160 compounds' inhibitory effect was further tested in the same RGA assays that were used for model building as shown in Figure 6. Due to reagents and compound availability constraints the figure contains 586 data points from 12 out of the 13 assays. In 120 out of the 586 data points (20.4%) the compounds show 50% or higher inhibition. This ratio is significantly higher than the normal hit rate in a HTS campaign. This result further supports the findings from Figure 6 that RGA frequent hitters are to some extent predictable.

We have hypothesized that off-target binding produces promiscuous activity in RGAs; in practice however, it is stress on the cell from off-target binding that most likely translates into frequent hitting across various assay systems. In other words, antineoplastic compounds would generally appear as actives in assays where a decrease in luminescence is the measure of activity, since their cytotoxicity would decrease cell count and the luciferase signal would be lowered artificially. We examined the possibility that RGA frequent hitters were affecting the readout of the luciferase signal due specifically to cytotoxicity. Cytotoxicity is dependent on a number of different parameters, including the cell line used, the time the compound was incubated with the cells, the number of cells used, the assay format used to measure cell death (or growth) as well as the concentration at which the compound was tested. Therefore we used an average value of growth inhibition for two different cell lines as an indication of cytotoxicity. Figure 7 shows the relationship between the activity of the selected test compounds in our additional RGA assay and the average activity of these compounds in the measure of cytotoxicity. The results suggest that a significant portion of the RGA frequent hitters

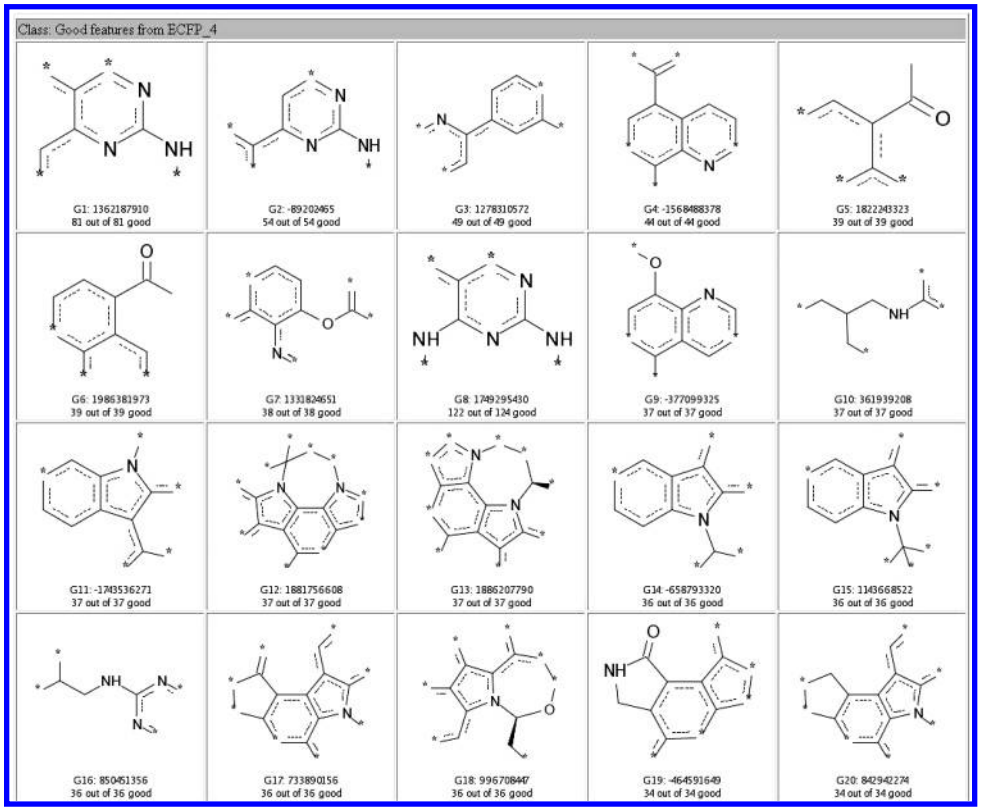


Figure 8. Analysis of the structural features associated with frequent hitter compounds in reporter gene assays. Various classes of nitrogen heterocycles as well as polycyclic compounds are over-represented among frequent hitters.

are acting to inhibit cell viability. In fact almost all of the compounds found active in the additional RGA appeared to be active due to various forms of cytotoxicity. The fact that these compounds were perturbing the cellular physiology and therefore causing false positives is as expected. Given the correlation between the documented cytotoxic effects and overenriched targets, the target prediction method appears to be identifying the cytotoxic frequent hitters as well as a putative mechanism of activity.

Interpretation of Features Causing False-Positives in Reporter Gene Assays. Figure 8 gives an analysis of the structural features associated with frequent hitter compounds in reporter gene assays. Various classes of nitrogen heterocycles as well as polycyclic compounds are over-represented among frequent hitters. In particular, aminopyrimidines are highly correlated with frequent hitting; these are well-known substructures found in kinase inhibitors.

CONCLUSIONS

Cell-based screens can generate false positive hits in a number of ways. This work provides a possibility for generating models and for detecting frequent hitters in reporter gene assays, a method than can be applied directly in practice, e.g., for HTS hit list triaging. We did not find a statistically significant difference in physicochemical properties of RGA hits and the Novartis collection, suggesting that off-target effects were a more realistic source of false positive signals. These off-target effects were then the subject of further research.

Given that the *reduced* luciferase signal is the readout for compound activity in antagonist reporter gene assays, it was hypothesized that compounds affecting cell cycle and cell growth could be a consistent and predictable source of false positives in RGAs (since also cell death decreases luciferase readout). Based on a chemogenomics database, we also predicted likely targets of these false-positive compounds and showed that the predicted cellular targets were associated with undesirable cellular effects such as cytotoxicity and luciferase inhibition (e.g., protease inhibitors). More specifically, the most frequently predicted targets related to apoptosis and cell differentiation. They include mixed lineage kinases, cyclin-dependent kinases, topoisomerase II, protein phosphatase 1, and protein phosphatase 2. Inhibition of all of these targets can be rationalized to give false positive results. Experimental validation was performed on an external validation set of 160 molecules and a 50% hit rate of predicted frequent hitters, compared to 2% usually encountered in reporter gene assays, corroborated our model. Our mechanistic hypothesis was that frequent hitters in RGAs which share the luciferase readout act to disrupt normal cell cycle progression or possibly even induce apoptosis. This is supported by the observation that for the 160 compounds predicted to be RGA frequent hitters, RGA activity is closely correlated with inhibition of cell viability (Figure 7).

Tools like the one developed here are particularly useful in case of high hit rates where further analysis is needed to prioritize compounds. Although many cytotoxic compounds can be manually identified by experienced medicinal chemists, it is impractical to perform for the tens of thousands of hits that can result from cell-based screening of very large chemical collections. Having systematic methods to separate fact from artifact is important for all aspects of pharmaceutical discovery. We show here that deconvoluting frequent hitter properties into predicted target activities translates directly into improved hit list triaging in practice.

Note Added after ASAP Publication. This article was released ASAP on May 25, 2007, with no Supporting Information paragraph. The correct version was posted on July 11, 2007.

Supporting Information Available: MDDR compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Macarron, R. Critical review of the role of HTS in drug discovery. *Drug Discovery Today* **2006**, *11*, 277−279.
(2) Johnston, P. A.; Johnston, P. A. Cellular platforms for HTS: three case studies. *Drug Discovery Today* **2002**, *7*, 353−363.
(3) Bronstein, I.; Fortin, J.; Stanley, P. E.; Stewart, G. S.; Kricka, L. J. Chemiluminescent and bioluminescent reporter gene assays. *Anal. Biochem.* **1994**, *219*, 169−181.
(4) Diller, D. J.; Hobbs, D. W. Deriving knowledge through data mining high-throughput screening data. *J. Med. Chem.* **2004**, *47*, 6373−6383.
(5) Rishton, G. M. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discovery Today* **2003**, *8*, 86−96.
(6) 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. Screen.* **2003**, *8*, 292−304.
(7) Pearce, B. C.; Sofia, M. J.; Good, A. C.; Drexler, D. M.; Stock, D. A. An Empirical Process for the Design of High-Throughput Screening Deck Filters. *J. Chem. Inf. Model.* **2006**, *46*, 1060−1068.
(8) 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.; Sjogren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von der 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*, 137−142.
(9) 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*, 4477−4486.
(10) 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.
(11) McGovern, S. L.; Shoichet, B. K. Kinase Inhibitors: Not Just for Kinases Anymore. *J. Med. Chem.* **2003**, *46*, 1478−1483.
(12) Nowlin, D.; Bingham, P.; Berridge, A.; Gribbon, P.; Laflin, P.; Sewing, A. Analysing the output from primary screening. *Comb. Chem. High Throughput Screening* **2006**, *9*, 331−337.
(13) Walum, E.; Hedander, J.; Garberg, P. Research perspectives for prescreening alternatives to animal experimentation On the relevance of cytotoxicity measurements, barrier passage determinations and high throughput screening in vitro to select potentially hazardous compounds in large sets of chemicals. *Toxicol. Appl. Pharmacol.* **2005**, *207*, 393−397.
(14) Andrews, J. M.; Newbound, G. C.; Lairmore, M. D. Transcriptional modulation of viral reporter gene constructs following induction of the cellular stress response. *Nucleic Acids Res.* **1997**, *25*, 1082−1084.
(15) Lindenmeyer, M. T.; Garcia-Pineres, A. J.; Castro, V.; Merfort, I. Sesquiterpene lactones inhibit luciferase but not beta-galactosidase activity in vitro and ex vivo. *Anal. Biochem.* **2004**, *328*, 147−154.
(16) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimentation and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(17) Davies, J. W.; Glick, M.; Jenkins, J. L. Streamlining lead discovery by aligning in silico and high-throughput screening. *Curr. Opin. Chem. Biol.* **2006**, *10*, 343−351.
(18) Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. *J. Chem. Inf. Model.* **2006**, *46*, 193−200.
(19) WOMBAT 2006 Sunset Molecular Systems. http://sunsetmolecular.com/products/?id=4 (accessed June 2006).
(20) Inpharmatica Index. http://www.inpharmatica.co.uk/StARLITe/Index.htm (accessed June 2006).
(21) Accelrys. http://www.accelrys.com/products/catalyst/catalystproducts/jubilant.html (accessed June 2006).
(22) GVK Biosciences. http://www.gvkbio.com/informatics/dbprod.htm (accessed June 2006).
(23) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124−1133.
(24) Schuffenhauer, A.; Popov, M.; Schopfer, U.; Acklin, P.; Stanek, J.; Jacoby, E. Molecular diversity management strategies for building and

enhancement of diverse and focused lead discovery compound screening collections. *Comb. Chem. High Throughput Screening* **2004**, *7*, 771−781.
(25) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(26) MDL Drug Data Report (MDDR). http://www.mdl.com/products/knowledge/drug_data_report/ (accessed July 2006).
(27) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* **2004**, *47*, 4463−4470.
(28) Klon, A. E.; Glick, M.; Davies, J. W. Combination of a naïve Bayes classifier with consensus scoring improves enrichment of high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 4356−4359.
(29) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for structures. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.
(30) Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, *9*, 199−204.
(31) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging Chemical and Biological Space: "Target Fishing" Using 2D and 3D Molecular Descriptors. *J. Med. Chem.* **2006**, *49*, 6802−6810.
(32) Jenkins, J. L.; Bender, A.; Davies, J. W. In silico target fishing: Predicting biological targets from chemical structure. *Drug Discovery Today: Technol.* **2007**, *3*, 413−421.
(33) O'Brian, J.; Wilson, I.; Orton, T.; Pognan, F. Investigation of the Alamar Blue (resazurin) fluorescent dye for the assessment of mammalian cell cytotoxicity. *Eur. J. Biochem.* **2000**, *267*, 5421−5426.
(34) Cohen, P. Protein kinases-the major drug targets of the twenty-first century? *Nat. Rev. Drug Discovery* **2002**, *1*, 309−315.
(35) Kim, K. Y.; Kim, B. C.; Xu, Z.; Kim, S. J. Mixed lineage kinase 3 (MLK3)-activated p38 MAP kinase mediates transforming growth factor-beta-induced apoptosis in hepatoma cells. *J. Biol. Chem.* **2004**, *279*, 29478−29484.
(36) Cha, H.; Dangi, S.; Machamer, C. E.; Shapiro, P. Inhibition of mixed-lineage kinase (MLK) activity during G2-phase disrupts microtubule formation and mitotic progression in HeLa cells. *Cell. Signalling* **2006**, *18*, 93−104.
(37) Xu, Z.; Maroney, A. C.; Dobrzanski, P.; Kukekov, N. V.; Greene, L. A. The MLK family mediates c-Jun N-terminal kinase activation in neuronal apoptosis. *Mol. Cell Biol.* **2001**, *21*, 4713−4724.
(38) Sherr, C. J.; Roberts, J. M. Inhibitors of mammalian G1 cyclin-dependent kinases. *Genes Dev.* **1995**, *9*, 1149−1163.
(39) Polyak, K.; Kato, J. Y.; Solomon, M. J.; Sherr, C. J.; Massague, J.; Roberts, J. M.; Koff, A. p27Kip1, a cyclin-Cdk inhibitor, links transforming growth factor-beta and contact inhibition to cell cycle arrest. *Genes Dev.* **1994**, *8*, 9−22.
(40) Baldwin, E. L.; Osheroff, N. Etoposide, topoisomerase II and cancer. *Curr. Med. Chem. Anticancer Agents* **2005**, *5*, 363−372.
(41) Hunter, T. Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell* **1995**, *80*, 225−236.
(42) Cohen, P. T. Novel protein serine/threonine phosphatases: variety is the spice of life. *Trends Biochem. Sci.* **1997**, *22*, 245−251.
(43) Deroo, B.; Archer, T. K. Proteasome Inhibitors Reduce Luciferase and b-Galactosidase Activity in Tissue Culture Cells. *J. Biol. Chem.* **2002**, *277*, 20120−20123.
(44) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.
(45) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.

CI6005504