Journal of Medicinal Chemistry

ARTICLE

pubs.acs.org/jmc

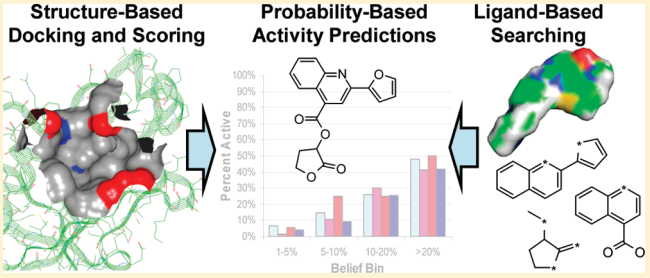
A Unified, Probabilistic Framework for Structure- and Ligand-Based Virtual Screening

Steven L. Swann, Scott P. Brown, Steven W. Muchmore, Hetal Patel, Philip Merta, John Locklear, and Philip J. Hajduk*

Global Pharmaceutical Research and Development, Abbott Laboratories, 100 Abbott Park Road, Abbott Park, Illinois 60064, United States

Supporting Information

ABSTRACT: We present a probabilistic framework for interpreting structure-based virtual screening that returns a quantitative likelihood of observing bioactivity and can be quantitatively combined with ligand-based screening methods to yield a cumulative prediction that consistently outperforms any single screening metric. The approach has been developed and validated on more than 30 different protein targets. Transforming structure-based in silico screening results into robust probabilities of activity enables the general fusion of multiple structure- and ligand-based approaches and returns a quantitative expectation of success that can be used to prioritize (or deprioritize) further discovery activities. This unified probabilistic framework offers a paradigm shift in how docking and scoring results are interpreted, which can enhance early lead-finding efforts by maximizing the value of in silico computational tools.



INTRODUCTION

The use of in silico methods for the identification of new active molecules for a target of interest permeates the early stages of drug discovery research across the pharmaceutical industry. This includes both ligand-based screening approaches (which identify test compounds structurally related to a known active molecule) and structure-based virtual screening methods (which identify molecules that complement a protein active site). Structure-based methods in particular hold special promise in that the search for new molecules is dictated by the binding pocket itself and is not dependent on the existence of known active molecules—increasing the prospects for identifying truly novel chemical matter. However, despite significant efforts to advance the methodology, it is still generally recognized that the performance of most docking and scoring algorithms is inadequate to reliably impact industrial drug discovery productivity.[1]

Given the complexity of molecular recognition and the simplified assumptions used in most scoring functions, it is not surprising that current approaches to docking and scoring often fail to clearly identify leads of sufficient quality or quantity. Numerous attempts have been described in the literature for improving docking and scoring, and in particular pose prediction, including the use of multiple[2] (or adaptive)[3] receptor conformations, normalizing the docking scores,[4] and fusing results from multiple scoring functions.[5] Unfortunately, while performance improvements have been reported retrospectively using these and other schemes, it is difficult to predict a priori what data fusion or other approach will work for a given system. Significantly, it has been analytically demonstrated that data fusion can either increase or decrease performance based on the underlying data distributions for the active and inactive molecules and that none of the most common fusion methods are systematically superior.[6] Thus, in truly prospective cases, it may be required to validate a given approach by generating and testing predictions on small numbers of compounds—at which point the full value of the in silico approach (insofar as occurring long before experimental confirmation) is significantly minimized. Furthermore, in the absence of a framework for understanding whether a given docking and scoring campaign has "succeeded" or "failed," the default is either to test as many of the highest ranked compounds as capacity allows (thereby potentially wasting resources on failed campaigns) or to manually inspect and select a small number of structures for follow-up (potentially missing rich areas of chemical diversity in successful campaigns).

To complement structure-based screening, there has been a longstanding interest in combining these approaches with ligand-based screening methods, which historically have enjoyed more application and success than structure-based screening.[7] Such a unified approach would leverage all available structural and chemical information in the search for new molecules and holds significant potential. Unfortunately, the relatively small number of reports attempting to fuse structure- and ligand-based approaches have yielded mixed results.[8] This is due at least in part

Received: September 10, 2010
Published: February 10, 2011

dx.doi.org/10.1021/jm1013677 | J. Med. Chem. 2011, 54, 1223–1232

to the inability of common fusion strategies to systematically deliver superior performance. Alternative approaches involve either preselecting or rescoring the structure-based docking library using ligand-based methods.[9] In either case, the fusion of a "good" algorithm with those of a "bad" algorithm (in terms of absolute performance against that particular target) tends to yield the "average" result, and there is no current approach to knowing the good from the bad.

We have previously reported on the use of Belief Theory[10] to address these same limitations for ligand-based virtual screening.[11] Belief Theory requires that quantifiable probabilities of an event being true can be obtained. To this end, we created a large database of ligand similarity and potency values to enable the construction of probability assignment curves (PACs), which are empirically derived functions that translate a measure (e.g., a similarity score) into a probability of a compound being active. Moving to a probabilistic framework enables the objective assessment of a screening campaign, in the sense that a virtual screening list can be deprioritized if only low-probability events are retrieved. Quantitative probabilities also enable the straightforward combination of multiple sources of information using well-established principles. In data fusion using Belief Theory, the contribution of any single source of information will be appropriately weighted by the confidence in that value being significant. Thus, this framework naturally and appropriately deemphasizes poor sources of information and maximizes the contribution of highly likely events. Since publication of our ligand-based belief framework,[11] we have subsequently validated these predictions through prospective inquiry for more than 23000 compounds against an additional 20 diverse protein targets (data not shown)—suggesting that appropriately constructed PACs can be universally and prospectively applied.

Given our experience with the broad applicability and utility of a unified probabilistic approach for ligand-based screening, here, we report that the application of Belief Theory to structure-based screening can enable a quantitative estimate of success for an individual screening campaign and allow productive integration with multiple ligand-based methods that systematically outperforms any individual approach. This integration of structure- and ligand-based screening incorporates a truly orthogonal view of chemical similarity and maximizes the value of our computational screening approaches. To accomplish this, we have created a highly curated database of active and decoy molecules from publicly available databases and developed new PACs for both ligand- and structure-based screening that can be quantitatively integrated. The resulting framework consistently outperforms any single screening method and provides a universal platform for virtual screening that leverages all available information (both chemical and structural). In addition, we have made the database available so that PACs for new similarity functions can be developed and integrated into future programs.

RESULTS AND DISCUSSION

Development of PACs for Structure-Based Screening. The generalizability of any PAC is critically dependent on the number and diversity of systems upon which its probabilities have been derived. To address this issue in the context of structure-based docking and scoring, we have assembled a rigorously curated list of known actives for 18 different targets. All actives were taken from the WOMBAT[12] database, and the particular targets were chosen based on the requirement that each have a quality public domain crystal structure available, except for PDGFRB, where we used a previously reported homology model[13] built using c-Kit as the model template. As the ability of a data set to critically assess relative performance between methods is strongly dependent upon the selection of a decoy compound set,[14] we chose a decoy set of nearly 45000 molecules that matched the physicochemical properties (e.g., molecular weight and ClogP) present in the set of active molecules by judiciously sampling from the ZINC[15] database (see Table 1). Using our set of actives and decoys, we evaluated the performance of two different docking programs, GLIDE[16] and FRED,[17] and four scoring functions found within these docking programs: Chemgauss3,[17] ZAPBIND,[18] Chemical Gaussian Overlay (CGO),[17] and GLIDE SP[16] (Table 2). As a performance metric to compare these methods, we use the area under receiver−operator characteristic (AU-ROC), the values for which fall in the range 0−1, with 0.5 corresponding to the performance for a purely random sampling method. As can be appreciated from Table 2, the performance of the docking and scoring algorithms is highly variable both across targets and between different algorithms, with a sustained better-than-random retrieval for all methods.

Table 1. Average Chemical Properties of the 3600 Active Molecules from the WOMBAT Database as Compared to the 45000 Decoys from the ZINC Database

property	WOMBAT	ZINC
MW	395 ± 102	388 ± 68
AlogP	3.4 ± 1.9	3.4 ± 1.4
NRot	6 ± 3	5 ± 2
tPSA	77 ± 41	78 ± 25

As noted earlier, the scores produced by the various scoring functions do not, in and of themselves, provide information as to the strength of the evidence (i.e., probability of a given molecule being active). In addition, while the scores for active molecules do tend to be higher, on average, than inactive molecules (therefore resulting in AU-ROC scores >0.5), the absolute docking score is unrelated to bioactivity and tends to be more a function of the protein target under study. This can be appreciated from Figure 1A, where the range in absolute docking scores for the actives and inactives against all 18 targets is displayed. This target-based influence precludes simply utilizing the raw score for comparisons between and across targets. We therefore transformed the respective docking scores from each program to Z values (see Experimental Section) for both the decoys and the actives across the set of 18 public domain targets. As can be appreciated from Figure 1B (using CGO as the example), the Z values for active molecules form a single distribution that is distinctly shifted relative to the set of inactive molecules (the average Z score for active molecules is 1.1, as compared to an average Z score of 0.0 for the inactive molecules).

Using Z values, we can now construct PACs for each of the scoring function by calculating the fraction of active molecules in each Z bin. The average PAC for each scoring function across all 18 targets is shown in Figure 2, the behavior of which can be described by a simple sigmoidal function (parameters listed in Figure 2B). While there is some variability in the PAC as a function of both target and scoring algorithm (see the Supporting Information), it can be seen that for Chemgauss3, CGO, or GLIDE SP, a Z value of 3 roughly corresponds to a 1% or greater

Table 2. AU-ROC Values for Four Scoring Function Across 18 Protein Targets from the WOMBAT Database[a]

target	PDB	CGO AU-ROC	ChemGuass3 AU-ROC	GLIDE AU-ROC	ZAPBIND AU-ROC
Ache	1EVE	0.64	0.60	0.65	0.60
AR	2AO6	0.87	0.84	0.96	0.73
CDK2	1CKP	0.84	0.56	0.5	0.55
COX-2	1CX2	0.83	0.58	0.8	0.50
c-SRC	2SRC	0.74	0.77	0.79	0.77
DHFR	3DFR	0.61	0.64	0.71	0.65
EGFR	1M17	0.66	0.66	0.68	0.57
ERα	1L2I	0.59	0.59	0.62	0.51
FTASE	3E32	0.83	0.75	0.78	0.52
fXa	1F0R	0.82	0.85	0.85	0.73.
HIV-1 P	1HPX	0.68	0.7	0.67	0.68
HIV-1 RT	3IRX	0.81	0.75	0.76	0.72
MAPK p38	1KV2	0.62	0.77	0.66	0.61
MMP-1	1FBL	0.92	0.88	0.91	0.93
PDE5	1XP0	0.77	0.88	0.8	0.61
PDGFRB	model	0.65	0.56	0.68	0.52
PPARγ	1FM9	0.87	0.88	0.87	0.87
thrombin	1A8I	0.71	0.69	0.73	0.58
average		0.75	0.72	0.75	0.64

[a] A total of 50 actives were chosen for each target (only 49 could be identified for c-SRC) from WOMBAT, along with a set of 45000 decoys from ZINC. All compounds are provided in the Supporting Information. Compounds were docked to their respective target (identified by the PDB code) along with the entire set of decoy molecules as described in the Experimental Section. The expected standard error in these AU-ROC values is ±0.04 units.[22]

probability of being active (10-fold higher than a random background rate of 0.1%), whereas Z values of 4 or higher correspond to a 5% or greater chance of observing activity. It is important to note that no significant enrichment is observed for molecular weight (see Figure 2A), suggesting that these scoring functions are productively capturing some critical aspects of molecular recognition. Using these PACs, we can now assign a probability (or belief, B_i) of observing activity for any given compound against a particular target of interest based upon its docking score and the docking scores of the decoy compounds.

Combining Structure- and Ligand-Based Virtual Screening. While attempts to synthesize information from both ligand-based and structure-based methods have been reported,[9,19] there is no general framework with which to simply and systematically combine the complementary sources of information provided by ligand-centric and structure-centric methods. With a probabilistic framework that now captures the information from structure-based screening, we can combine these results with ligand-based similarity methods using the principles of Belief Theory[10] and provide a single expectation value (or cumulative belief) for observing activity. Probabilities for activity based on the protein structure can be obtained from the PAC in Figure 2, while probabilities based on ligand similarity can be obtained using the PACs derived from two-dimensional graph-based [e.g., extended connectivity fingerprint 6 (ECFP6)[20]] and three-dimensional shape-based [e.g., rapid overlay of chemical structures (ROCS)[21]] metrics in analogy to our previous work.[11] For consistency, we have rederived generalized PACs for ECFP6 and ROCS using the same compounds described above for structure-based screening and augmented with actives from an additional 22 targets for which crystal structures are not available (see the Supporting Information).

Individual beliefs from the various methods can be combined to assign a cumulative belief (see Experimental Section) to each molecule. Tables 3 and 4 list the performance metrics for the combination of beliefs from ECFP6 (ligand-based), ROCS (ligand-based), and CGO (structure-based). In terms of overall retrieval as measured by the AU-ROC curve (Table 3), it can be observed that, in all but two of the protein targets (DHFR and ERα) that we evaluated, the retrieval rates for the cumulative belief result in AU-ROC values that are equal to or better than the highest retrieval rate achieved with any single method. This can also be appreciated from the ROC curves in Figure 3A. On the basis of the retrieval of 50 actives from a library of 50000 compounds, the expected standard error in these AU-ROC values is ±0.04 units.[22] Thus, many of these enhancements are quite significant. It is important to note that similar trends were observed when only the 25 most diverse actives for each target were used (see the Supporting Information, Table S5), suggesting that the methods exhibit good performance across multiple chemotypes. Also listed in Table 3 are the results obtained from alternative fusion methods, including the mean rank, the min rank, and the max rank. The cumulative belief systematically outperforms the mean and max rank fusions, while exhibiting comparable performance to the min rank and mean Z fusions (see Experimental Section). While overall retrieval rates are important, practical utilization of virtual screening results typically involves testing only the top-ranked fraction of the database; therefore, early enrichment is an important measure of success. As shown in Table 4, the enrichment factors at 1% (i.e., for the top-ranked 500 compounds in the database) are also higher, on average, for the cumulative belief as compared to any individual metric. Interestingly, the enrichment observed for the mean rank and max rank fusions are actually *inferior*, on average, to the two ligand-based methods alone. This is not unanticipated[6] and underscores the fact that mean ranking treats all scoring functions equally, even if there is no strong "signal" against certain targets.

A theoretical model[6] for data fusion indicates that enhancement from fusing multiple screening methods can arise from a significant degree of correlation of the true positives from each method when compared to the true negatives (with each method thereby enforcing the signal in the others). As illustrated in Figure 4A,B, this is unlikely to be true in this case for structure-based screening, as essentially no correlation in the actives exists between CGO and either of the ligand-based methods ($R^2 = 0.0$). In our case, the enrichment observed for the cumulative belief is likely due to the orthogonality of the three methods. As shown in the Venn diagram in Figure 4D, the actives recovered using all three methods are only partially overlapped—highlighting the value of using multiple, complementary measures for virtual screening. Similar complementarity of structure- and ligand-based methods has been recently reported for four different targets.[23] While CGO was used for calculation of the cumulative belief in the examples listed, it is important to note that equivalent performance was observed for GLIDE (data not shown).

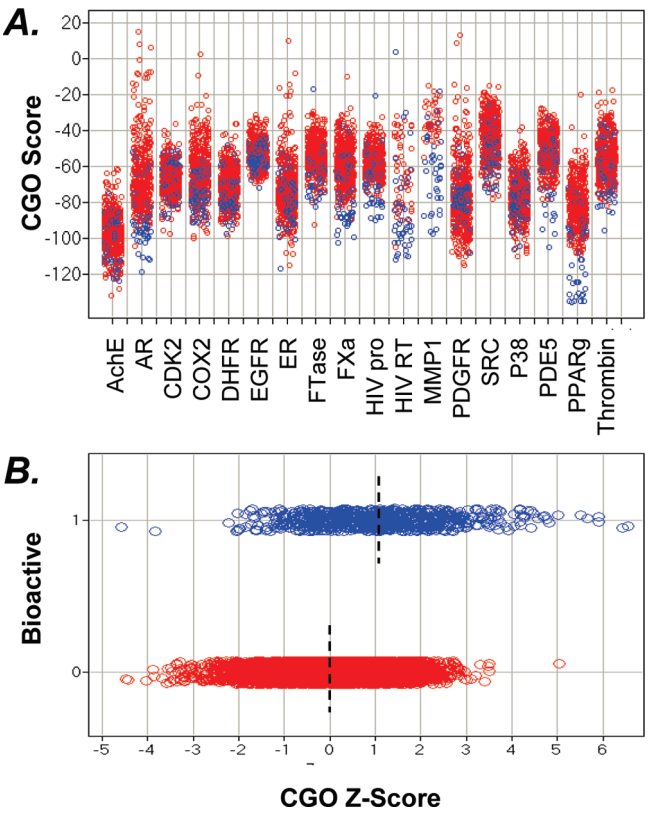
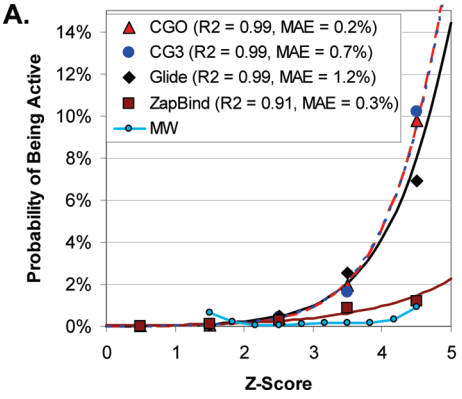


Figure 1. (A) CGO docking scores for active (blue) and inactive (red) compounds as a function of each of the 18 targets listed in Table 2. (B) Distribution of Z scores for active (blue, top) and inactive (red, bottom) across all targets. The average Z score for the active molecules is 1.1 (denoted with a dashed vertical line), while the average Z score for the inactive molecule is 0.0. Only a random 1% sampling of the inactives is shown for clarity.

The min rank and mean Z fusions yield comparable retrospective performance to the cumulative belief, both in terms of overall AU-ROC scores and early enrichment (Tables 3 and 4). In a sense, these fusion methods are closest in kind to the cumulative belief, as exceptional values in one scoring function (either in rank, Z value, or belief) will dominate the final fusion score. Using Belief Theory, scores with unusually strong signal are appropriately weighted, while values "in the noise" are essentially discarded. Similarly, the "best performing" function for each compound will dominate the min rank and mean Z fusion rules. However, only the cumulative belief returns a quantitative expectation for activity that is maintained both for the individual measures and for the combined belief (Figure 3B). This is not possible with any other fusion rule and allows for truly *prospective* assessments of new active molecules. It is this aspect of Belief Theory that enables a paradigm shift in how virtual screening results can be interpreted and utilized in a drug discovery setting. Importantly, the ability of the described PACs to accurately predict the likelihood of activity of new compounds has been extensively validated on external and internal data sets and in prospective biochemical testing, as will be described in the next two sections.

External Validation. To test the validity and generality of our approach, an independent validation set of six additional targets (Table 5) were analyzed using an additional set of compounds from the WOMBAT database (see the Supporting Information, Table S1). The likelihood of activity was assigned to each molecule based on the docking scores and Z scores generated by CGO, using the PACs described above. As is described above



B.

Scoring Function	Fmax	SC_{50}	Slope
CGO	0.40	5.3	0.74
ChemGauss3	0.40	5.4	0.73
GLIDE	0.40	5.6	0.65
ZapBind	0.40	8.5	0.38

C.

$$B_i = \frac{F_{\max}}{1+10^{(SC_{50}-x_i)*Slope}}$$

Figure 2. (A) PACs for the four scoring functions across the set of 18 public domain protein targets from WOMBAT (listed in Table 2). Correlation coefficients (R^2) and mean absolute errors (MAE) from the curve fits are also provided in the legend. (B) Parameters for the four scoring functions, obtained from sigmoidal curve fits to the probability assessment curves in A. (C) The form of the sigmoidal function used to fit the data in A, where B_i is the belief that compound i is active, x_i is the Z value for compound i, F_{max} is the maximum value for the fraction active, SC_{50} is the similarity cutoff at which 50% of the maximum fraction-active is observed, and slope is the is the steepness of the curve.

dx.doi.org/10.1021/jm1013677 |J. Med. Chem. 2011, 54, 1223–1232

Table 3. AU-ROC Values for ECFP6, ROCS, CGO, Their Cumulative Belief, and Various Other Data Fusion Methods against the Training Set of 18 Proteins[a]

target	ECFP6	ROCS	CGO	cumulative belief	mean rank	min rank	max rank	mean Z
AChE	0.64	0.61	0.66	0.76	0.79	0.87	0.60	0.83
AR	0.80	0.67	0.87	0.90	0.89	0.90	0.85	0.90
CDK2	0.50	0.59	0.52	0.59	0.54	0.55	0.57	0.56
COX-2	0.96	0.93	0.81	0.97	0.96	0.97	0.94	0.98
c-Src	0.58	0.63	0.72	0.91	0.87	0.91	0.74	0.91
DHFR	0.83	0.71	0.67	0.77	0.82	0.78	0.75	0.82
EGFR	0.83	0.78	0.65	0.92	0.90	0.91	0.82	0.93
ERα	0.84	0.68	0.63	0.78	0.77	0.92	0.72	0.86
FTase	0.65	0.66	0.88	0.87	0.81	0.89	0.68	0.87
fXa	0.77	0.71	0.84	0.84	0.75	0.87	0.68	0.82
HIV-1 P	0.89	0.53	0.80	0.89	0.74	0.93	0.51	0.76
HIV-1 RT	0.64	0.60	0.80	0.75	0.72	0.89	0.53	0.76
MAPK p38	0.62	0.57	0.63	0.68	0.68	0.69	0.64	0.70
MMP-1	0.88	0.69	0.91	0.94	0.90	0.94	0.81	0.93
PDE5	0.70	0.56	0.73	0.80	0.80	0.78	0.72	0.82
PDGFR	0.90	0.86	0.54	0.91	0.87	0.92	0.72	0.91
PPARγ	0.92	0.85	0.87	0.93	0.90	0.91	0.87	0.91
thrombin	0.87	0.79	0.70	0.96	0.90	0.97	0.78	0.97
average	0.77	0.69	0.73	0.84	0.81	0.87	0.72	0.85

[a] Compounds were docked to their respective target or compared against the reference ligand, along with the entire set of decoy molecules as described in the Experimental Section. The expected standard error in these AU-ROC values is ±0.04 units.[22]

Table 4. Enrichment Values at 1% for ECFP6, ROCS, CGO, Their Cumulative Belief, and Various Other Data Fusion Methods against the Training Set of 18 Proteins[a]

target	ECFP6	ROCS	CGO	cumulative belief	mean rank	min rank	max rank	mean Z
AChE	12	14	6	14	4	14	4	14
AR	32	36	38	36	38	46	32	38
CDK2	28	28	6	28	20	26	16	28
COX-2	38	76	2	76	62	70	56	74
c-Src	50	28	14	54	18	54	12	50
DHFR	14	20	2	20	16	20	16	20
EGFR	52	50	2	60	28	62	18	62
ERα	30	32	8	32	24	36	24	38
FTase	18	26	18	38	20	40	12	42
fXa	8	8	42	14	2	42	2	8
HIV-1 P	20	4	28	30	10	28	8	26
HIV-1 RT	30	12	52	32	28	54	18	38
MAPK p38	12	14	4	18	8	18	6	18
MMP-1	58	18	54	58	54	58	40	60
PDE5	16	16	20	22	14	22	14	16
PDGFR	62	58	2	58	6	62	14	58
PPARγ	56	46	46	56	50	54	48	54
thrombin	52	48	16	52	34	50	32	48
average	33	30	20	39	24	42	21	38

[a] Compounds were assessed against their respective target or reference ligand, along with the entire set of decoy molecules as described in the Experimental Section.

for the training set, the fusion of the docking and ligand-based beliefs for the validation set also results in superior performance, with the cumulative belief resulting in the highest enrichment values and AU-ROC scores for all six targets (see Table 5) and systematically outperforming results obtained with the mean rank. Significantly, the quantitative expectation for bioactivity is again maintained over a wide range of probabilities, as shown in Figure 5. We have also conducted validation studies using internally derived data on an additional nine targets (see the Supporting Information, Table S3) with equally positive results—suggesting that the performance is independent of data source. Thus, while the true applicability domain for the PACs listed here

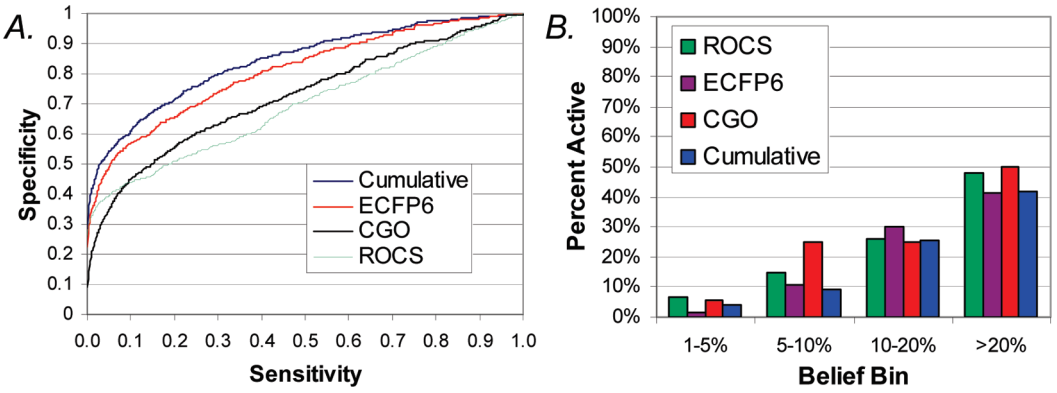


Figure 3. (A) ROC curve comparison against 18 targets from the WOMBAT database using ROCS (green), ECFP6 (red), and CGO (black) alone and for the conjunctive combination of all three functions (cumulative, blue) using Belief Theory. (B) Percent actives in each belief bin across all scoring methods, including combined measures.

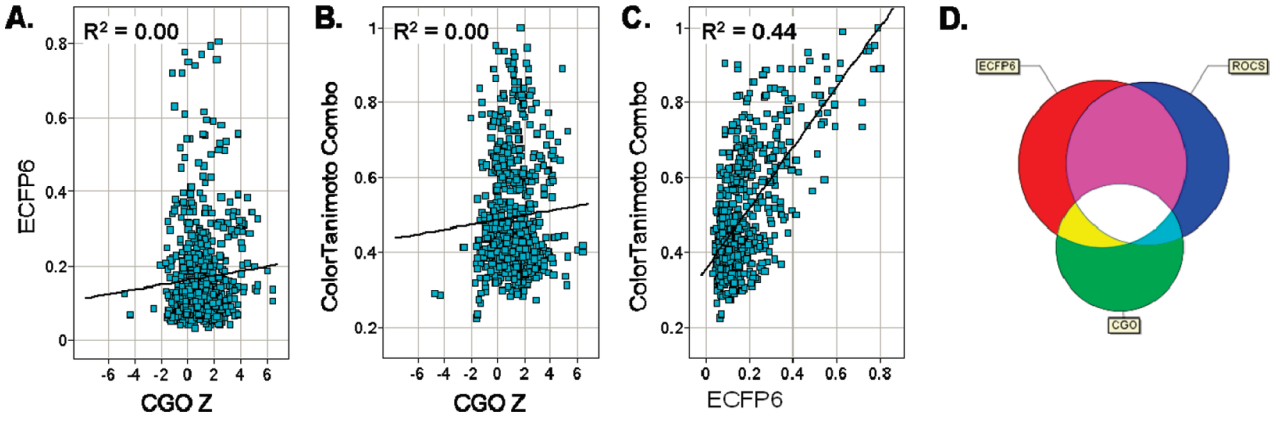


Figure 4. Correlation of (A) CGO *Z* score with ECFP6, (B) CGO *Z* score with Tanimoto Combo, and (C) Tanimoto Combo with ECFP6 for all actives against each of the 18 proteins in the training set. (D) Overlap of recovered actives for ligand- and structure-based virtual screening tools.

Table 5. Enrichment Values at 1% and AU-ROC Scores for ECFP6, ROCS, CGO, Their Cumulative Belief, and Their Mean Rank against a Validation Set of Six Proteins[a]

target	ECFP6	ROCS	CGO	cumulative belief	mean rank	min rank	max rank	mean Z
				enrichment at 1%				
CHK1	24	24	6	24	20	26	18	24
METAP2	54	54	4	54	38	56	14	52
MMP13	22	14	20	28	30	16	28	34
PARP	26	30	28	40	42	34	30	42
PDF	58	38	36	72	58	70	40	70
uPA	32	10	20	32	18	30	16	28
average	36	28	19	42	34	39	24	42
				AU-ROC scores				
CHK1	0.69	0.55	0.67	0.75	0.66	0.73	0.52	0.66
METAP2	0.83	0.72	0.66	0.84	0.80	0.82	0.76	0.81
MMP13	0.84	0.72	0.67	0.87	0.79	0.83	0.69	0.80
PARP	0.89	0.84	0.83	0.95	0.94	0.94	0.86	0.95
PDF	0.85	0.8	0.86	0.98	0.97	0.98	0.89	0.99
uPA	0.76	0.51	0.69	0.81	0.73	0.74	0.64	0.75
average	0.81	0.69	0.72	0.87	0.82	0.84	0.73	0.83

[a] Compounds were assessed against their respective target or reference ligand, along with the entire set of decoy molecules as described in the Experimental Section. The expected standard error in these AU-ROC values is ±0.04 units.[22]

dx.doi.org/10.1021/jm1013677 |*J. Med. Chem.* 2011, 54, 1223–1232

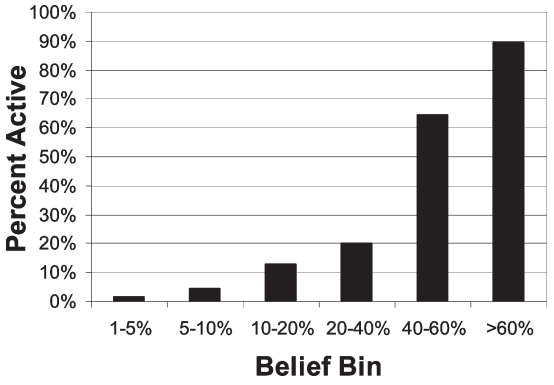


Figure 5. Percent actives in each cumulative belief bin (combining CGO, ECFP6, and ROCS beliefs) across all six targets in the validation set.

will be dependent on the targets in the training set, this result is validation that the PACs reported here for structure- and ligand-based screening will be applicable to the vast majority of protein targets.

Prospective Validation. Both of the validation sets described above are based upon retrospective analyses of reported or known bioactive molecules and a set of presumed inactive decoys. Of course, the true test of this probabilistic virtual screening framework lies in its ability to prospectively identify new active molecules at the expected rate. Our general experience with prospective applications is summarized in the Supporting Information (Table S4), where we have observed excellent agreement between predicted and actual hit rates. Shown in Figure 6 are sets of two newly identified actives for each of four targets—MMP3, PR, PDE4, and Lck (none of which were included in any of the training or validation sets). Compounds **1** and **2** contain an amino-pyrimidine hinge binding element and exhibit activity against Lck. These compounds have known activity against Akt and PI3 kinase,[24] but no activity against Lck has been reported to date. Compounds **3** and **4** inhibit PDE4 and represent novel chemotypes for PDE inhibition as no activity for these compounds has been reported against any member of the phosphodiesterase family. Compounds **5** and **6** were derived from medicinal chemistry programs against GR,[25] and it is therefore not surprising that they exhibit good activity against the highly related PR. The chemotypes represented by compounds **7** and **8** were in fact originally designed for MMP-3 (stromelysin) inhibition[26] and were later evolved into inhibitors of MMP-2/9 (gelatinase-A/B).[27] Overall, the prospective screening campaigns satisfyingly yielded both known and novel chemotypes and at a hit rate consistent with predictions.

CONCLUSIONS

In summary, we have described a unified probabilistic framework that combines both structure- and ligand-based screening methods and provides a quantitative probability that a specific molecule will be active. This framework has been developed on 18 different protein targets and validated on an additional six targets—strongly supporting the general utility of this approach. Moving to a probabilistic interpretation of structure- and ligand-based virtual screening is a significant deviation from the current applications of the available computation tools. This carries significant advantages for drug discovery by leveraging all available information in the search for new molecules, maximizing the diversity of the recovered actives and enabling a risk−benefit analysis of virtual screening results to optimize further discovery activities. The framework described here is immediately extensible to new scoring functions (both ligand- and structure-based), provided that PACs can be derived based on the available data sets (in the Supporting Information). While the development of such PACs can represent a significant up-front investment for new docking and scoring functions, this represents extensive validation and characterization of the true predictive value of any new technique and should therefore be considered a standard requirement for new scoring functions. The method is also amenable to the integration of data from multiple ligands or multiple target structures, thus maximizing not only the use of the computational tools but also the available chemical and structural information for a given target. To preserve the quantitative value of the cumulative belief, care must be taken to ensure that correlation among the various methods is low, and we look forward to new research in this area.

EXPERIMENTAL SECTION

Selection of Decoy and Active Sets from WOMBAT and ZINC. We selected sets of 50 actives (defined as compounds with IC_{50} values less than 1 μM) from WOMBAT for 18 targets under the requirement that each target have an existing crystal structures in the public domain. For each target, we also selected up to an additional 100 compounds that exhibited IC_{50} values between 1 and 10 μM. For the purposes of structure-based screening methods, these additional compounds were considered inactive, whereas they were used to establish differing potency windows for the ligand-based methods. To choose a set of decoys having properties consistent with the actives, we matched mean and standard deviations between the two sets for molecular weight, AlogP, number of rotatable bonds, and tPSA. The exact values of the final matched properties are shown in Table 1. All compounds have been submitted to PubChem, and Substance Identification (SID) values are given in the Supporting Information.

FRED. To prepare the small molecules (both decoys and actives) for docking and scoring with the program FRED (v2.5.5),[17] it was necessary to first generate sets of conformers for each molecule (as FRED treats a single conformer as rigid during docking process). Omega (v2.3.2)[1c,28] was used to generate sets of conformers for each of the decoy and active molecules with default settings except for the following: energy window value of 50, maximum number of conformers of 1000, and root-mean-square conformer clustering threshold of 0.5 Å. We investigate performance of the following scoring functions calculated by FRED: Shapegauss,[29] PLP,[30] Chemgauss3, Chemscore,[31] Screenscore,[32] CGO,[17] and ZapBind.[18] FRED breaks down the docking and scoring procedure into several steps. First, FRED exhaustively docks each conformer for each molecule and ranks each pose based on the user-selected scoring function. The top-ranked poses are then moved on to the optimization stage, in which FRED optimizes the score of each pose using a user-selected scoring function (which can be different from the exhaustive-search selected scoring function). The last stage is final scoring, in which the user can select yet another scoring function (or choose to calculate all of them available within FRED). We perform two separate FRED runs for each target. The first uses Chemgauss3 in the exhaustive phase (selecting top 1000 poses) and also in the optimization phase, followed by calculation of all scoring functions except CGO. The second type of run uses CGO in the exhaustive phase (selecting top 100 poses), which fundamentally changes the ranked list of poses produced, followed by Chemgauss3 in the optimization and final scoring phases. CGO uses similarity to a known bound ligand to bias the conformer ranking in the exhaustive scoring phase, both similarity to the shape as well as placement of protein-interacting hydrogen-bond acceptor and donor groups.

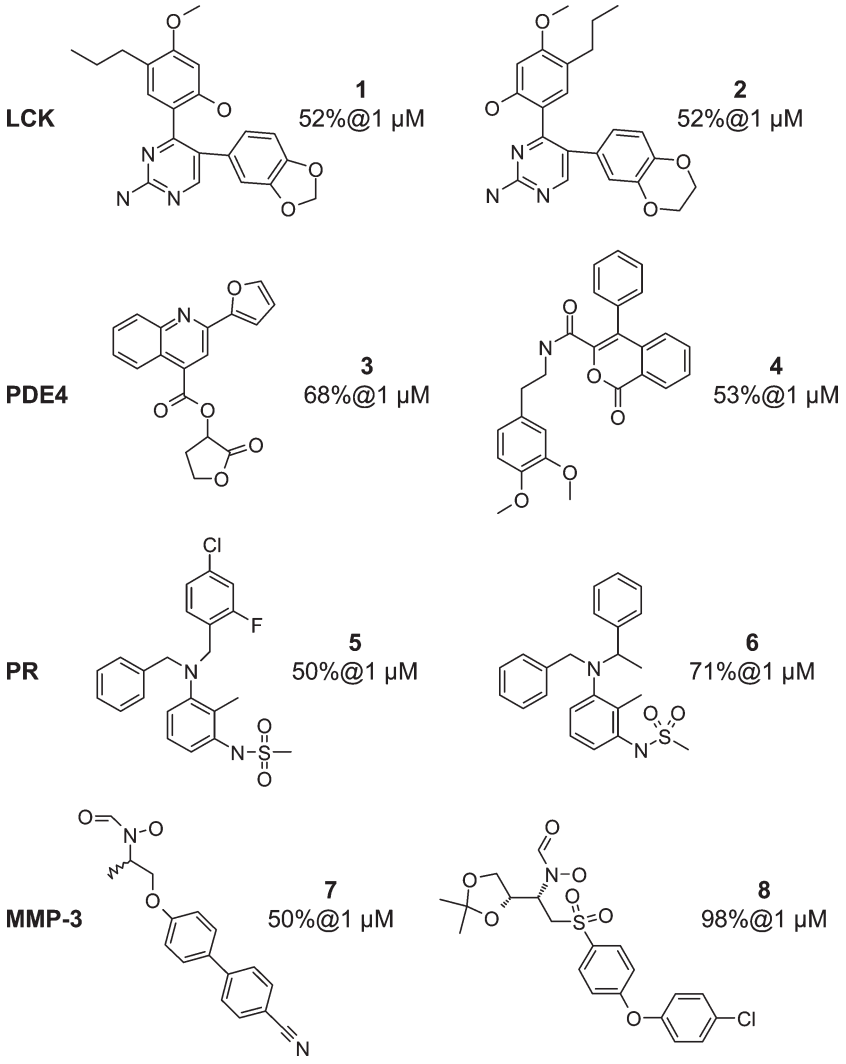


Figure 6. Subset of newly identified actives for four targets (listed at the left), along with the percent inhibition observed at a compound concentration of 1 μM.

In each FRED run, the bound ligand is used as input into FRED, which uses the ligand coordinates to autogenerate a box encompassing the protein active site. We add 4.0 Å to each side of the box in all of the FRED runs, to ensure a large enough box to encompass all of the ligands processed in the docking procedure. To ensure proper scoring of hydrogen-bonding terms, as well as proper atom typing for assigning protein and ligand partial charges (used in the ZapBind calculations), explicit hydrogen atoms are added to each protein−ligand complex (with the ligand in the binding pocket) using a hydrogen atom placement tool in OpenEye's OEChem.[33] After placement, the system is then minimized with rigid heavy atoms using Szybki.[33]

GLIDE. Small molecules were prepared using the version of LigPrep contained within the MAESTRO 9.0 software using all default settings. Protein grids were prepared, with no constraints, using the default setting under the Glide Grid Generation Tool. All molecules were then docked in each protein using Glide in SP Mode using the default settings.

ECFP6. ECFP6 were calculated within Pipeline Pilot v 7.0 (Accelrys) using a bit length of 4096. PACs were calculated as previously described.[11]

ROCS. To perform the three-dimensional overlays we use the program ROCS (v3.0).[21] We use the same sets of conformers generated in the FRED docking runs for both the actives and the decoys (see above). For each bound ligand, we generate five conformations using Omega (v2.3.2),[1c,28] which are then used to perform a multi-conformer overlay using ROCS. We use all default settings to perform the ROCS overlays. ROCS returns shape and color similarity Tanimoto values for the best conformer pair between each bound ligand and every active or decoy molecule. The shape and color Tanimoto values are then used to calculate belief values to be used in constructing the cumulative beliefs.

Generation of Beliefs. Z values are calculated from the docking scores of the active and decoy sets using the following equation

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

where x is the docking score, μ is the mean score for the decoy set, and σ is the standard deviation of the distribution of scores across the decoy set. For each scoring function, the scores for the decoy and active sets are reduced to Z values and binned in integer increments (i.e., $Z = 1, 2, 3, 4$). In each Z bin, we calculate the fraction of active molecules within the bin, which when divided by total number of molecules in the bin gives a probability. Thus, a probability of being active can be assigned, which is a function of Z bin occupied.

To combine the structure-based probabilities, or beliefs, with the ligand-based beliefs produced by ECFP6 and ROCS, we turn to Belief Theory, which provides the framework for the combination of multiple

beliefs (or probabilities) from independent, uncorrelated sources using the conjunctive rule:[34]

$$\text{cumulative belief} = 1 - \prod_{i=1}^{N} (1 - P_i) \quad (2)$$

where N is the number of independent beliefs and P_i is the belief that the i-th measure (on a scale from 0 to 1) is true. An examination of the correlation between the structure- and the ligand-based methods is shown in Figure 4, in which the docking Z scores can be seen to be uncorrelated with the ligand-based measures of ECFP6 and ROCS (correlation coefficients of 0.03 and 0.18, respectively). The distinct lack of correlation allows all beliefs to be conjunctively combined using the above equation, producing a total cumulative belief based on a fusion of the individual beliefs from ECFP6, ROCS, and CGO.

Enrichment Values. Enrichment values at 1% were calculated by counting the number of actives identified in the top-ranked 500 molecules (1% of a 50000 compound database) for each method and then dividing by the expected "random" hit rate of 0.1% (50 actives in a database of 50000 decoys).

Other Fusion Methods. The mean rank was calculated by the simple average of the ranks of the compound across all three scoring functions, while the min and max rank used the lowest or highest rank value, respectively, across the three functions. The mean Z fusion was obtained by calculating the Z value for each compound from the distribution of CGO scores, ECFP6 Tanimoto values, and ROC Color Tanimoto scores and taking the simple average of all Z values.

■ ASSOCIATED CONTENT

Supporting Information. Text file containing all compound structures and activities for the training and validation data sets; information on redevelopment of the ligand-based PACs; and validation results using internal data, prospective data, and only the 25 most diverse actives per target. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
*Tel: 847-937-0368. E-mail: philip.hajduk@abbott.com.

■ ABBREVIATIONS USED

CGO, chemical Gaussian overlay; ROCS, rapid overlay of chemical structures; ECFP, extended connectivity fingerprints; PAC, probability assignment curve

■ REFERENCES

(1) (a) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49* (6), 1455–1474. (b) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, 5912–5931. (c) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56* (2), 235–249.

(2) Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the selection of experimental protein conformations for virtual screening. *J. Chem. Inf. Model.* **2010**, *50* (1), 186–193.

(3) Jain, A. N. Effects of protein conformation in docking: Improved pose prediction through protein pocket adaptation. *J. Comput.-Aided Mol. Des.* **2009**, *23* (6), 355–374.

(4) (a) Larsen, R. J. M.; Morris L. *An Introduction to Mathematical Statistics and Its Applications*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, 2000. (b) Baldi, P.; Benz, R. W. BLASTing small molecules—Statistics and extreme statistics of chemical similarity scores. *Bioinformatics* **2008**, *24* (13), 357–365.

(5) Feher, M. Consensus scoring for protein-ligand interactions. *Drug Discovery Today* **2006**, *11* (9−10), 421–428.

(6) (a) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of data fusion methods in virtual screening: similarity and group fusion. *J. Chem. Inf. Model.* **2006**, *46* (6), 2206–2219. (b) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of data fusion methods in virtual screening: Theoretical model. *J. Chem. Inf. Model.* **2006**, *46* (6), 2193–2205.

(7) McGaughey, G.; Sheridan, R.; Bayly, C.; Culberson, C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47* (4), 1504–1519.

(8) Tan, L.; Geppert, H.; Sisay, M. T.; Gutschow, M.; Bajorath, J. Integrating structure- and ligand-based virtual screening: Comparison of individual, parallel, and fused molecular docking and similarity search calculations on multiple targets. *ChemMedChem* **2008**, *3* (10), 1566–1571.

(9) Lee, H. S.; Choi, J.; Kufareva, I.; Abagyan, R.; Filikov, A.; Yang, Y.; Yoon, S. Optimization of high throughput virtual screening by combining shape-matching and docking methods. *J. Chem. Inf. Model.* **2008**, *48* (3), 489–497.

(10) Shafer, G. *A Mathematical Theory of Evidence*; Princeton University Press: Princeton, NJ, 1976.

(11) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J. Chem. Inf. Model.* **2008**, *48* (5), 941–948.

(12) (a) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: New York, 2004; pp 223−239. (b) Olah, M.; Oprea, T. I. Bioactivity Databases. In *Comprehensive Medicinal Chemistry II*; Taylor, J. B., Triggle, D. J., Eds.; Elsevier: Oxford, 2006; Vol. 3, pp 293−313. (c) Olah, M.; Rad, R.; Ostopovici, L.; Bora, A.; Hadaruga, N.; Hadaruga, D.; Moldovan, R.; Fulias, A.; Mracec, M.; Oprea, T. I. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. In *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*; Schreiber, S. L., Kapoor, T. M., Wess, G., Eds.; Wiley-VCH: New York, 2007; pp 760−786.

(13) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801.

(14) Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: Pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 179–190.

(15) Irwin, J. J.; Shoichet, B. K. ZINC—A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–182.

(16) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749.

(17) McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2011**, in press.

(18) Grant, J. A.; Pickup, B. T.; Nicholls, A.; Smooth, A Permittivity Function for Poisson-Boltzmann Solvation Methods. *J. Comput. Chem.* **2001**, *22* (6), 608–640.

(19) (a) Kraemer, O.; Hazemann, I.; Podjarny, A. D.; Klebe, G. Virtual screening for inhibitors of human aldose reductase. *Proteins* **2004**, *55* (4), 814–823. (b) Ward, R. A.; Perkins, T. D.; Stafford, J. Structure-based virtual screening for low molecular weight chemical starting points for dipeptidyl peptidase IV inhibitors. *J. Med. Chem.* **2005**, *48* (22), 6991–6996. (c) Salam, N. K.; Nuti, R.; Sherman, W. Novel method for generating structure-based pharmacophores using energetic analysis. *J. Chem. Inf. Model.* **2009**, *49* (10), 2356–2368.

(20) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.
(21) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17* (14), 1653–1666.
(22) Hanley, J. A.; McNeil, B. J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **1983**, *148* (3), 839–843.
(23) Kruger, D. M.; Evers, A. Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *ChemMedChem* **2010**, *5* (1), 148–158.
(24) Vennemann, M.; Bär, T.; Maier, T.; Hölder, S.; Beneke, G.; Dehmel, F.; Zülch, A.; Strub, A.; Beckers, T.; Ince, S.; Rehwinkel, H.; Liu, N.; Bömer, U. Fused Pyrimidines as AKT Inhibitors. WO/2010/091824, 2010.
(25) Link, J. T.; Sorensen, B.; Patel, J.; Grynfarb, M.; Goos-Nilsson, A.; Wang, J.; Fung, S.; Wilcox, D.; Zinker, B.; Nguyen, P.; Hickman, B.; Schmidt, J. M.; Swanson, S.; Tian, Z.; Reisch, T. J.; Rotert, G.; Du, J.; Lane, B.; von Geldern, T. W.; Jacobson, P. B. Antidiabetic activity of passive nonsteroidal glucocorticoid receptor modulators. *J. Med. Chem.* **2005**, *48* (16), 5295–5304.
(26) Hajduk, P. J.; Sheppard, G.; Nettesheim, D. G.; Olejniczak, E. T.; Shuker, S. B.; Meadows, R. P.; Steinman, D. H.; Carrera, G. M., Jr.; Marcotte, P. A.; Severin, J.; Walter, K.; Smith, H.; Gubbins, E.; Simmer, R.; Holzman, T. F.; Morgan, D. W.; Davidsen, S. K.; Summers, J. B.; Fesik, S. W. Discovery of Potent Nonpeptide Inhibitors of Stromelysin Using SAR by NMR. *J. Am. Chem. Soc.* **1997**, No. 119, 5818–5827.
(27) Wada, C. K.; Holms, J. H.; Curtin, M. L.; Dai, Y.; Florjancic, A. S.; Garland, R. B.; Guo, Y.; Heyman, H. R.; Stacey, J. R.; Steinman, D. H.; Albert, D. H.; Bouska, J. J.; Elmore, I. N.; Goodfellow, C. L.; Marcotte, P. A.; Tapang, P.; Morgan, D. W.; Michaelides, M. R.; Davidsen, S. K. Phenoxyphenyl sulfone N-formylhydroxylamines (retrohydroxamates) as potent, selective, orally bioavailable matrix metalloproteinase inhibitors. *J. Med. Chem.* **2002**, *45* (1), 219–232.
(28) Boström, J.; Greenwood, J.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graphics Modell.* **2003**, *21* (5), 449–462.
(29) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68* (1), 76–90.
(30) Verkivker, G. M.; Bouzida, D.; Gehlaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 731–751.
(31) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11* (5), 425–445.
(32) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44* (7), 1035–1042.
(33) OpenEye Scientific Software, 9 Bisbee Court, Suite D, Santa Fe, NM 87508; http://www.eyesopen.com/ (accessed August 15, 2010).
(34) Hooper, G. A calculation of the credibility of human testimony. *Phil. Trans. R. Soc.* **1699**, *21*, 359–365.