ARTICLES

Application of Belief Theory to Similarity Data Fusion for Use in Analog Searching and Lead Hopping

Steven W. Muchmore, Derek A. Debe, James T. Metz, Scott P. Brown, Yvonne C. Martin, and Philip J. Hajduk*

Pharmaceutical Discovery Division, GPRD, Abbott Laboratories, Abbott Park, Illinois 60064-6098

Received December 3, 2007

A wide variety of computational algorithms have been developed that strive to capture the chemical similarity between two compounds for use in virtual screening and lead discovery. One limitation of such approaches is that, while a returned similarity value reflects the perceived degree of relatedness between any two compounds, there is no direct correlation between this value and the expectation or confidence that any two molecules will in fact be equally active. A lack of a common framework for interpretation of similarity measures also confounds the reliable fusion of information from different algorithms. Here, we present a probabilistic framework for interpreting similarity measures that directly correlates the similarity value to a quantitative expectation that two molecules will in fact be equipotent. The approach is based on extensive benchmarking of 10 different similarity methods (MACCS keys, Daylight fingerprints, maximum common subgraphs, rapid overlay of chemical structures (ROCS) shape similarity, and six connectivity-based fingerprints) against a database of more than 150 000 compounds with activity data against 23 protein targets. Given this unified and probabilistic framework for interpreting chemical similarity, principles derived from decision theory can then be applied to combine the evidence from different similarity measures in such a way that both capitalizes on the strengths of the individual approaches and maintains a quantitative estimate of the likelihood that any two molecules will exhibit similar biological activity.

INTRODUCTION

Ligand-based virtual screening (LBVS) is a powerful tool for identifying new biologically active compounds.[1,2] Most computational approaches for LBVS rely on some measure of chemical similarity between the compounds being compared, with the assumption that more similar compounds will have a higher chance of being active[1,3] A large number of diverse methods have been developed that perceive different aspects of chemical similarity, including substructure descriptor counts,[4–6] atom connectivities,[7] molecular features,[8–10] and molecular shape.[11,12] In a typical ligand-based virtual screen, one or more reference active molecules are analyzed and compared to a test database, and the top-ranking list of compounds (those most similar to the query) are returned, evaluated, and ultimately tested for biological activity. Such approaches have been successfully applied to the identification of novel bioactive leads for a number of targets, including recent examples on K(ATP) channels,[13] 5-lipoxygenase,[14] and the MCH receptor.[15]

Despite the potential power of ligand-based approaches, there are several issues that limit their broader applicability and utility in the field. The first is that the calculation will always return a rank-ordered list of the most similar molecules, and most approaches do not estimate the probability that any test molecule will in fact be active. It may very well be that no molecules in the test database are of sufficient similarity to the references molecule(s) to warrant bioassay follow up. While one method for estimating overall retrieval rates from LBVS has been proposed (referred to as the FP-KL-BDACCS score),[16] it only applies to fingerprinting techniques that generate uncorrelated bit settings (and therefore cannot be applied to connectivity or shape descriptions). A second limitation is that the different methods for evaluating similarity often retrieve very different lists of compounds. While it has been recognized that one can potentially exploit the complementarity between different similarity measures (often referred to as *data fusion* or *consensus scoring*),[17–23] there is no universally accepted approach for balancing either the absolute similarity values or the resulting rank order of compounds from the different virtual screening campaigns. Thus, there exists a need for a quantitative estimate of the probability of identifying bioactive hits through similarity searching, along with a rigorous framework for combining the results of multiple similarity metrics while retaining the estimates of success.

Here, we describe a probabilistic interpretation of chemical similarity based on extensive benchmarking against a large data set of compounds with known bioactivity. The resulting relationships between chemical similarity and bioactivity allow derivation of probabilities that any compound pair with a given level of similarity will in fact be equipotent. The resulting probabilities always map on a scale from 0.0 to

* To whom correspondence should be addressed. Mailing address: Abbott Laboratories, R46Y, AP-10, 100 Abbott Park Road, Abbott Park, IL 60064-3500. Phone: (847) 937-0368. Fax: (847) 938-2478. E-mail: Philip.hajduk@abbott.com.

10.1021/ci7004498 CCC: $40.75
Published on Web 04/17/2008

1.0 regardless of the similarity metric being considered, and thus provide a common framework for interpreting the various measures. In addition, this probabilistic framework allows fusion of results from different similarity measures using Dempster−Schafer Theory (DST; see http://www.glennshafer.com for a useful listing of books and literature references).[24] DST (also known as belief theory) is a mathematical theory of evidence that has been developed to combine separate pieces of information that can arise from different sources.[25−29] This is precisely the challenge that we face in combining results from multiple similarity searches. As an analogy, DST, and in particular Schafer's rule of combination, has been used extensively in engineering as a way to combine information from multiple sensors that have differing degrees of reliability.

In this study, a total of 10 different similarity measures have been evaluated using this probabilistic approach, and their utility both alone and in combination in retrieving active compounds and reporting quantitative estimates of success is described. By evaluating the performance of similarity metrics in this way, we are able to assess both the performance of an individual measure of chemical similarity, as well as provide a combined estimate of the probability that two compounds will be equipotent.

RESULTS AND DISCUSSION

Generation of the Probability Assignment Curve. Developing a *probability assignment* is the basic function in DST, and is an expression of the level of confidence that can be ascribed to a particular measurement. In this work, we develop the probability assignment for each of the similarity measurements by using an in-house database of activity measurements. Detailed descriptions of the data set and subsequent analyses are given in the Methods section, but will be briefly reviewed here. A data set of more than 66 000 compounds with associated activity data against 23 different protein targets was compiled, containing nearly 2400 compounds that exhibited an IC_{50} value between 1 and 10 nM against one of these targets. Each of the active compounds (IC_{50} value between 1 and 10 nM) was then compared to a random subset of the total data set using 10 different measures of similarity: Daylight fingerprints,[4] MACCS keys, extended connectivity fingerprints (ECFPs) and functional class fingerprints (FCFPs) with lengths of 2, 4, and 6,[7] a Tanimoto derived from the maximum common subgraph (MCS), and rapid overlay of chemical structures (ROCS) shape overlap[11] combined with chemical group matching. In this work, the ROCS similarity was the "comboscore" (shape overlap + atom-based overlap) divided by 2 to put it on scale with the other metrics. In the end, similarity metrics were calculated for 8082 active pairs (defined as two compounds with IC_{50} values within 1 log unit of each other), 58 605 inactive pairs (defined as two compounds with IC_{50} values that differ by more than 1 log unit), and an additional 104 429 random pairs from our corporate database (also treated as inactive pairs). The similarity as well as the change in potency was tabulated for all of these pairwise comparisons. An example of the resulting data is given in Figure 1 for Daylight fingerprints. This plot is similar to the neighborhood plot method previously reported;[30] however, the analysis of the information is significantly different (see below).

Once the distribution between the measure of similarity and the activity difference of the pairs has been established (as shown in Figure 1), it is possible to generate a *probability assignment* curve by simply calculating the ratio of those molecular pairs that have potency values within a defined variance versus their similarity values. Here, we define a compound pair as being active if the difference in potency between the two compounds is less than 1 log unit (red symbols in Figure 1). The resulting ratios are shown in Figure 2 for six similarity measures. It was observed that there are significant differences between similarity thresholds for the different techniques. For example, using a Daylight similarity cutoff of 0.85, 27% of the compound pairs are active (Figure 2A). This is in excellent agreement with the results reported previously.[3] In contrast, to achieve this same level of active pairs using ECFP_6 fingerprints (Figure 2B), a cutoff of only 0.42 is required. Interestingly, all of the plots shown in Figure 2 resemble simple dose–response curves, and can in fact be fit to a sigmoidal equation (see Methods) to yield a continuous function that can be used to translate any value for these similarity measures into a quantitative probability that any compound pair will in fact be active. The results of these curve-fitting exercises are shown in Table 2. F_{max} is the maximal probability that any two compounds with similarity approaching 1.0 will in fact be active, and is analogous to the maximal observed response in a dose–response curve. All of these values fall in the range of 0.3–0.5, reflecting the fact that even small changes to a molecule (including stereoisomers) can significantly affect activity. The similarity value for which half of the maximal response is observed is termed the SC_{50} (for similarity cutoff at 50%), which is analogous to an IC_{50}. This is a critical parameter, as it qualitatively represents the equivalence point for all of the different measures.

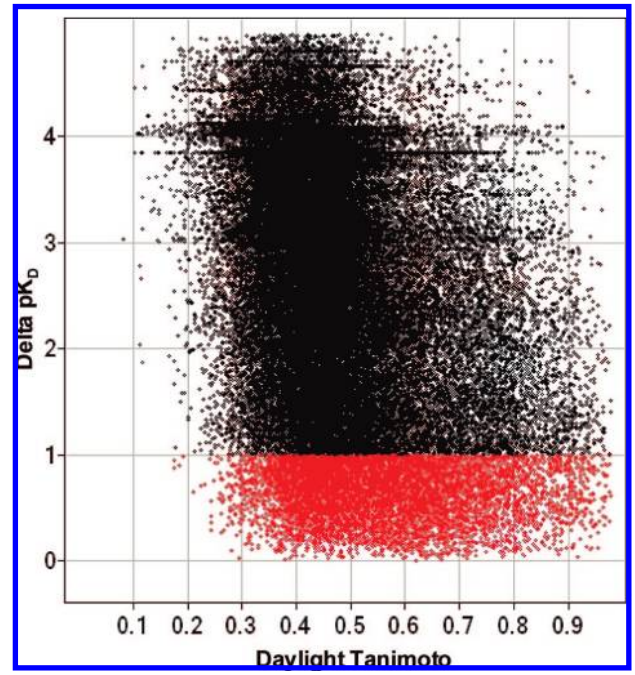

Figure 1. Plot of Daylight Tanimoto similarity values vs potency change for 66 687 pairs of compounds from our corporate database. The red circles denote those pairs of compounds that differ in potency against the same target by less than 1 log unit and were considered active as described in the text.

Evaluation of Individual and Combined Performance. Each similarity metric was analyzed alone and in combination for the ability to retrieve active pairs from the database. In order to evaluate the utility of the various metric combi-

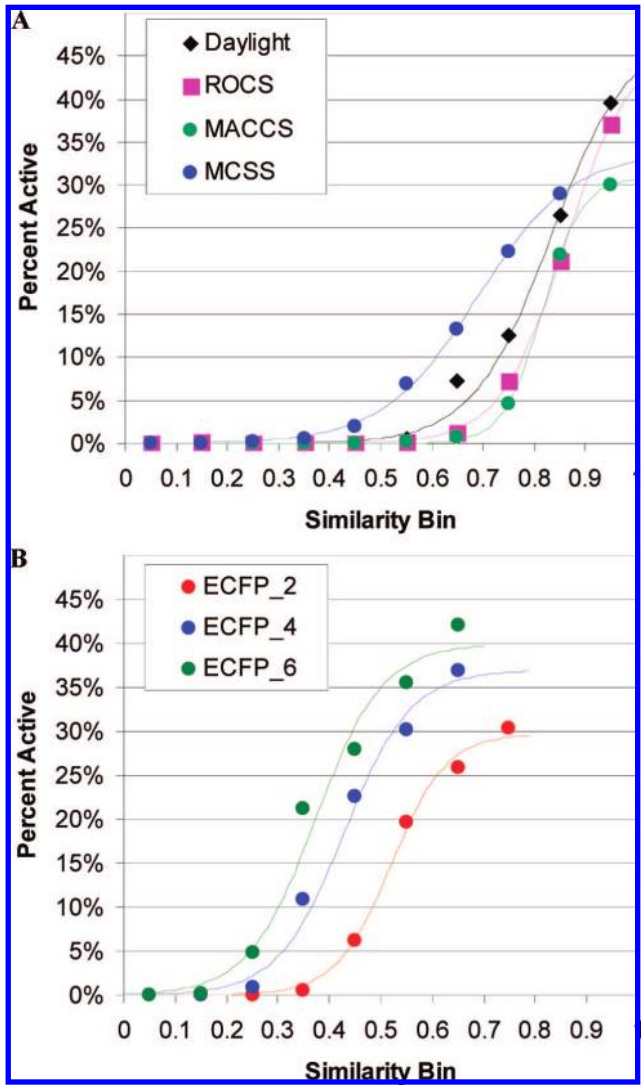

Figure 2. Probability assignment curves as described in the text for Tanimoto similarity values using (A) Daylight fingerprints (black diamonds), ROCS (average of shape and color, magenta squares), MACCS keys (green circles), maximum common substructure, and (B) Scitegic's extended connectivity fingerprints (ECFPs) using lengths of 2 (red circles), 4 (blue circles), and 6 (green circles). Plotted is the fraction of compound pairs that are equipotent (defined as a difference in potency of less than 1 log unit) out of all compound pairs that fall into a given similarity bin. Each data set was fit to a sigmoidal curve (eq 1) with the fitted values shown in Table 1. Similarity bins with less than 200 compound pairs were omitted from the analysis.

Table 1. Sigmoidal Curve Parameters for the Probability Assignment Curves Given in Figure 1

	F_{max}[a]	SC_{50}[b]	slope[c]
Daylight	0.48	0.83	5.7
ROCS	0.46	0.86	6.8
MCSS	0.34	0.69	4.7
MACCS	0.31	0.82	11.3
FCFP_2	0.33	0.75	5.6
FCFP_4	0.46	0.60	5.4
FCFP_6	0.40	0.45	6.4
ECFP_2	0.30	0.52	8.1
ECFP_4	0.37	0.43	6.7
ECFP_6	0.40	0.37	6.7

[a] Maximal fraction of compounds that are equipotent. [b] Similarity cutoff (analogous to an IC_{50}) at which the fraction of active pairs is half-maximal. [c] Slope (analogous to a Hill slope) for the sigmoidal curve given in eq 1.

Table 2. AU-ROC and BEDROC Values for the Individual and Combined Beliefs for Active Pair Retrieval Using the Training Set Comprised of 171 116 Compound Pairs

Tanimoto	AU-ROC[a]	SD[b]	BEDROC[c]	SD[d]	recovery at 1%[e]	SD[f]
Daylight	0.84	0.01	0.42	0.02	0.18	0.02
ROCS	0.77	0.01	0.36	0.02	0.19	0.02
MCSS	0.80	0.01	0.39	0.02	0.19	0.02
MACCS	0.84	0.01	0.39	0.02	0.19	0.02
FCFP_2	0.82	0.01	0.42	0.02	0.21	0.02
FCFP_4	0.82	0.01	0.45	0.02	0.22	0.02
FCFP_6	0.82	0.01	0.45	0.02	0.23	0.02
ECFP_2	0.82	0.01	0.45	0.02	0.23	0.02
ECFP_4	0.82	0.01	0.46	0.02	0.23	0.02
ECFP_6	0.82	0.01	0.46	0.02	0.24	0.02
E6R[g]	0.85[i]	0.01	0.46	0.02	0.26[i]	0.02
E6RD[h]	0.86[i]	0.01	0.47[i]	0.02	0.26[i]	0.02

[a] Area under the receiver operating characteristic curve. [b] Standard deviation in the AU-ROC value from 100 trials of retrieving 500 randomly selected active pairs in a total data set of 100 000 comparisons as described in the Methods section. [c] Boltzmann-enhanced discrimination of receiver operating characteristic curve using $\alpha = 20$.[31] [d] Standard deviation in the BEDROC value from 100 trials of retrieving 500 randomly selected active pairs in a total data set of 100 000 comparisons as described in the Methods section. [e] Fraction of actives recovered in the top-ranked 1% of the library. [f] Standard deviation in the fraction of actives recovered at 1% of the library from 100 trials of retrieving 500 randomly selected active pairs in a total data set of 100 000 comparisons as described in the Methods section. [g] Combination (using belief theory) of beliefs from ECFP_6 fingerprints and ROCS. [h] Combination (using belief theory) of beliefs from ECFP_6, Daylight, and ROCS. [i] $p < 0.01$ as compared to ECFP_6 alone.

nations in identifying these pairs, three parameters were calculated: the area under the receiver operating characteristic curve (AU-ROC), the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) curve, and the fraction of activities recovered in the top-ranked 1% of the database.[31] The AU-ROC value measures the overall performance in retrieving actives, but can be significantly biased by gains at the bottom of the rank order. The BEDROC value, on the other hand, was specifically designed to measure the initial enhancement by weighting the initial build-up of the ROC curve using an exponential function controlled by the "early recognition" parameter α (where a higher value of α corresponds to greater emphasis on early recognition).[31] As described above, similarity metrics were calculated for 8082 active pairs (defined as two compounds with IC_{50} values within 1 log unit of each other) and 163 034 inactive pairs (defined as two compounds with IC_{50} values that differ by more than 1 log unit). BEDROC estimates for a total of 8100 active pairs in a data set of 171 116 comparisons would suffer from severe saturation error, with a maximum relative error in the BEDROC value (Δ_{max}) of more than 45% using a value of 20 for α.[31] Therefore, a bootstrap approach (see Methods) was implemented in which multiple sets of 500 randomly chosen active pairs in a background of 100 000 total comparisons were created. This ratio of actives to inactives ensures that the *maximum* error in the BEDROC value is less than 5% using a value of 20 for α. AU-ROC and BEDROC values ($\alpha = 20$) were then calculated for each of these subsets and averaged. The final values (with standard deviations) are given in Table 2.

All of the individual similarity metrics performed reasonably well in retrieving active pairs from our data set, with

AU-ROC values ranging from 0.77 to 0.84 and BEDROC values ranging from 0.36 to 0.46 (see Table 2). The ECFP_6 metric was the best overall individual performer, with the highest BEDROC value (0.44) and an AU-ROC value (0.82) second only to Daylight. The ECFP_6 fingerprints were then combined (in a pairwise fashion) with each of the other metrics using the individual beliefs derived from the probability assignment curves (Figure 2) using Hooper's rule for combining concurrent evidence:[29,32]

$$\text{joint belief} = 1 - (1 - B_1)(1 - B_2) \quad (1)$$

where B_1 is the first belief (in this case the belief in the ECFP_6 value) and B_2 is the second belief (taken from the list of alternative similarity metrics). The conjunctive combination of beliefs for ECFP_6 and ROCS was the highest scoring pair, yielding small but statistically meaningful gains in both the AU-ROC values and recovery rates of 18–24% (see Table 2). This combination of 2D fingerprints and 3D shape overlap makes intuitive sense, as it can be expected that the two methods would capture different aspects of compound similarity. In fact, of all the metrics, the ROCS similarity values are the least correlated with ECFP_6, with a Pearson correlation (R) of only 0.562 (a full correlation matrix between all parameters is given in the Supporting Information, Table S1). This substantiates the expectation that combining the most orthogonal descriptors yields the greatest increases in retrieval rates.

The beliefs for ECFP_6 and ROCS were then combined with each of the remaining 8 metrics, in which the combination with Daylight yielded a slight gain in both the AU-ROC and BEDROC values (see Table 2). Given the high correlation between ECFP_6 and the remaining metrics (R values ranging from 0.716 to 0.985; see Table S1), it is not surprising that more substantial gains are not realized. In fact, combination of the ECFP_6/ROCS/Daylight triplet of beliefs with additional metrics yielded no further gains in performance. A listing of all combinations used in this study and the corresponding performance measures is given in the Supporting Information (Table S2).

Combining the individual beliefs derived from ECFP_6, ROCS, and Daylight produces a joint probability that any given pair of molecules will be active. As an example, consider a pair of molecules with ECFP_6, ROCS, and Daylight Tanimoto similarities of 0.27, 0.50, and 0.44, respectively. The individual beliefs derived from the probability assignment curves shown in Figure 2 are 6.9%, 0.2%, and 0.3% for ECFP_6, ROCS, and Daylight, respectively. The joint belief for this triplet of individual beliefs is 7.3%. This value can be used as a quantitative estimate of the confidence that a test compound will be active to the query. However, joining beliefs in this way requires that the individual sources of belief are independent. This is clearly not the case for the 10 metrics described here. In fact, the Pearson correlation (R) between the Tanimoto values derived from Daylight and ECFP6 fingerprints is 0.808 ($R^2 = 0.653$) (a full correlation matrix between all parameters is given in the Supporting Information, Table S1). The effect of descriptor correlation is illustrated in Figure 3, where the joint belief consistently overestimates the probability that compound pairs will be active (blue circles). However, this overestimate is systematic and, as shown in Figure 3, perfectly linear. Use of the fitted linear function to correct the resulting belief

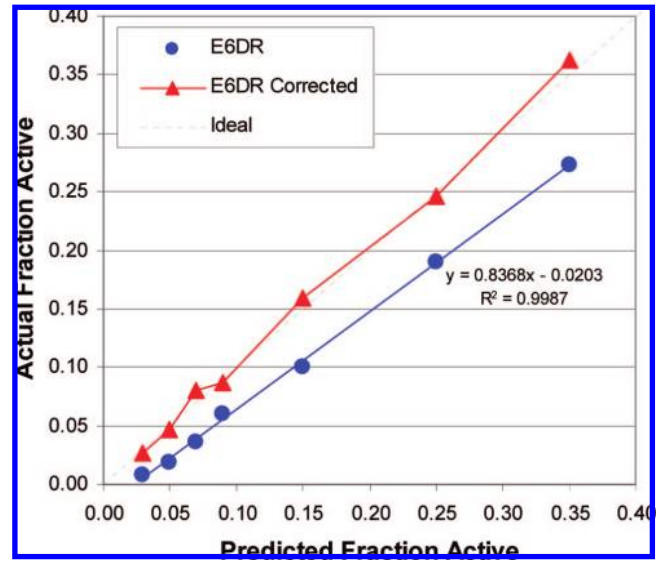

Figure 3. Effects of combining confidences on the predicted vs actual fraction of active pairs. Shown with blue circles is the combined joint belief for of ECFP_6, Daylight, and ROCS, which exhibits a systematic overestimation of the actual fraction of active pairs. This overestimation can be corrected by scaling the combined beliefs by the linear fit to produce the corrected combined belief shown with red triangles.

values from the ECFP_6/Daylight/ROCS combined belief yields a corrected joint belief (red triangles in Figure 3) that faithfully reproduces the actual fraction of active pairs.

Application to an External Test Set. In order to evaluate the applicability of the corrected joint belief to the retrieval of actives from naïve data sets, a set of publicly reported lead hops derived from 28 protein targets was assembled.[33,34] An average of 4.8 active compounds were identified for each target, resulting in a total of 134 compounds (a full listing of these compounds is given in the Supporting Information, Table S3). Similar to the analysis describe above, exhaustive pairwise comparisons were performed using 10 similarity metrics, yielding 310 active pairs in a total data set of 8911 pairwise comparisons (a full listing of these comparisons is given in the Supporting Information, Table S4). In this case, compounds were treated as active if they were reported to have activity against the same protein target, and as inactive if they were derived from different targets. As above, BEDROC estimates for a total of 310 active pairs in a data set of 8911 comparisons would suffer from severe saturation error ($\Delta_{max} > 30\%$).[31] Therefore, a bootstrap approach was again implemented in which multiple test sets of 50 randomly chosen active pairs in a background of 8651 total comparisons were created. This ratio of actives to inactives is very similar to that used for the training set and results in a maximum error for the BEDROC value of approximately 6%. AU-ROC, BEDROC and recovery rate values were then calculated for each of these subsets and averaged. The final values (with standard deviations) are given in Table 3.

As with the internal training set of more than 171 000 compound pairs, all of the individual similarity metrics performed reasonably well in retrieving active pairs from the validation set, with AU-ROC values ranging from 0.77 to 0.86 and BEDROC values ranging from 0.36 to 0.52 (see Table 3). While the ECFP_6 fingerprints perform well on this set (BEDROC value of 0.50), the functional class fingerprints (FCFP) perform slightly better (see Table 3).

Table 3. AU-ROC and BEDROC Values for the Individual and Combined Beliefs for Active Pair Retrieval Using the Validation Set of 134 Compounds[a]

Tanimoto	AU-ROC[b]	SD[c]	BEDROC[d]	SD[e]	recovery at 1%[f]	SD[g]
Daylight	0.77	0.04	0.43	0.06	0.31	0.03
ROCS	0.75	0.04	0.38	0.06	0.21	0.04
MCSS	0.78	0.04	0.38	0.07	0.25	0.06
MACCS	0.75	0.04	0.45	0.06	0.30	0.06
FCFP_2	0.85	0.03	0.49	0.06	0.28	0.07
FCFP_4	0.87	0.03	0.52	0.07	0.36	0.10
FCFP_6	0.86	0.03	0.52	0.06	0.36	0.08
ECFP_2	0.82	0.04	0.49	0.07	0.32	0.08
ECFP_4	0.82	0.04	0.50	0.06	0.35	0.07
ECFP6	0.82	0.03	0.50	0.07	0.36	0.09
E6RD[h]	0.85[l]	0.03	0.56[l]	0.06	0.43[l]	0.06
MAX(E6RD*)[i]	0.79[l]	0.04	0.41[l]	0.07	0.31	0.06
MIN(E6RD*)[j]	0.82	0.04	0.49	0.07	0.32	0.08
SUM(E6RD*)[k]	0.84[l]	0.03	0.54[l]	0.07	0.44[l]	0.06

[a] Compounds were derived from the work of Martin[33,34] and are listed in Supporting Information S3. [b] Area under the receiver operating characteristic curve. [c] Standard deviation in the AU-ROC value from 100 trials of retrieving 50 randomly selected active pairs in a total data set of 8651 comparisons as described in the Methods section. [d] Boltzmann-enhanced discrimination of receiver operating characteristic curve using $\alpha = 20$.[31] [e] Standard deviation in the BEDROC value from 100 trials of retrieving 50 randomly selected active pairs in a total data set of 8651 comparisons as described in the Methods section. [f] Fraction of actives recovered in the top-ranked 1% of the library. [g] Standard deviation in the fraction of actives recovered at 1% of the library from 100 trials of retrieving 500 randomly selected active pairs in a total data set of 100 000 comparisons as described in the Methods section. [h] Combination (using belief theory) of beliefs from ECFP_6, Daylight, and ROCS. [i] Data fusion using the maximum value of the self-scaled ranked similarities from ECFP_6, Daylight, and ROCS. [j] Data fusion using the minimum value of the self-scaled ranked similarities from ECFP_6, Daylight, and ROCS. [k] Data fusion using the sum of the self-scaled ranked similarities from ECFP_6, Daylight, and ROCS. [l] $p < 0.01$ as compared to ECFP_6 alone.

Nonetheless, the BEDROC value using the joint belief calculated from ECFP_6, ROCS, and Daylight outperforms any individual metric. Representative ROC curves are shown in Figure 4 for the ECFP_6/ROCS/Daylight combination and each of the three metrics alone, where the gains in overall retrieval and early recognition can be clearly visualized. Thus, the joint belief calculated from ECFP_6, ROCS, and Daylight appears to be a generally superior metric for identifying active leads from large databases.

Many other attempts as fusing similarity metrics to improve retrieval rates have been reported. A recent analysis has evaluated the use of the MIN-, MAX-, and SUM-fusion rules, in which the minimum, maximum, and summed values for the combination of self-scaled metrics are used, respectively.[17,18] We have calculated the AU-ROC and BEDROC values and recovery rates using three fusion rules for the combination of the normalized rank-order scores (not belief values) for ECFP_6, ROCS and Daylight combination, and the results are shown in Table 3. On this data set, the MAX- and MIN-fusion rules are inferior to the use of ECFP_6 alone, and significantly inferior to the joint belief for the ECFP_6/ROCS/Daylight combination. However, the SUM-fusion rule performs comparably to the ECFP_6/ROCS/Daylight joint belief combination. This is again satisfying, as it evidence that the three individual metrics are productively capturing different aspects of chemical similarity that can be merged in a number of ways. However, while the sum of the normalized ranks performs comparably on this data set, the quantitative estimate for the probability that any pair of compounds will in fact be equally active is lost using this metric. In contrast, the belief theory combination gives superior performance and retains this quantitative assessment.

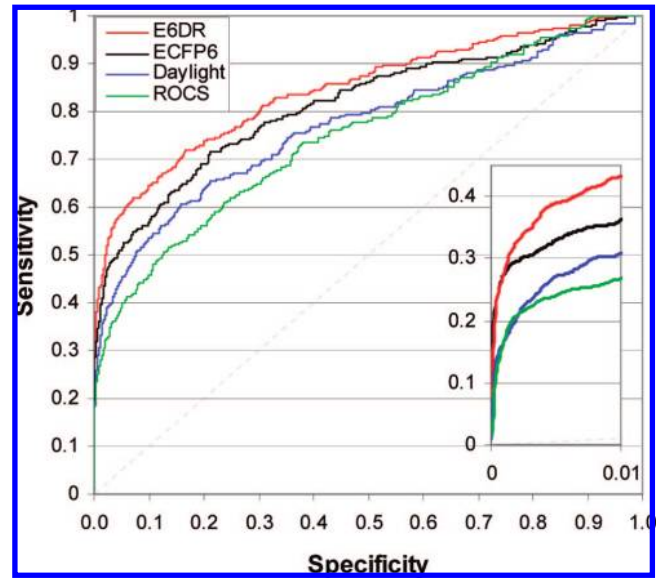

Figure 4. Representative receiver operating characteristic (ROC) curves for retrieval of 50 known lead hops from a background of 8651 compound pairs using ROCS (green), Daylight fingerprints (blue), ECFP_6 fingerprints (black), and the joint belief for ECFP_6, Daylight, and ROCS (red) as described in the text. The inset shows an expanded region of the initial build-up.

Use of Belief Theory in Lead Identification. As mentioned above, an important aspect of the use of joint belief is that, along with high performance, an estimate of the *confidence* for which a given compound pair will be active can be quantitated. For example, active compounds for four different targets from the external set are shown in Figure 5. Using the first compound as the query, it is observed that close analogs (second column of compounds in Figure 5) have joint beliefs in the range of 20–50% (suggesting that 20–50% of the compounds tested at this similarity level will *in fact* be equipotent to the query). These types of changes typically involve addition of substituents (e.g., targets 5 and 17), heteroatom switches (e.g., target 32), homologations (e.g., target 8), or other conservative changes. This level of joint belief should be used for close analog searches around high-throughput screening (HTS) hits or other active leads for which information on structure–activity relationships (SAR) is desired. However, true lead-hopping (where substantially different chemotypes are desired) occurs at a joint belief level of ~1–5%. As shown in Figure 5, hits at this level of belief exhibit ring deletions coupled with atom switches (e.g., target 17), ring opening coupled with substituent additions (e.g., targets 5 and 32), ring replacements (e.g., target 8), and other major changes to the query. Importantly, this level of belief suggests that only 1–5 compounds out of 100 that meet this belief threshold will in fact be active. Thus, expectations can be set as to how many compounds to screen and how many actives to expect. This is distinctly different from simply testing the top-ranked compounds from a virtual screen—regardless of the absolute scores (or similarity values). For example, in searching for alternative chemotypes in an advanced lead optimization

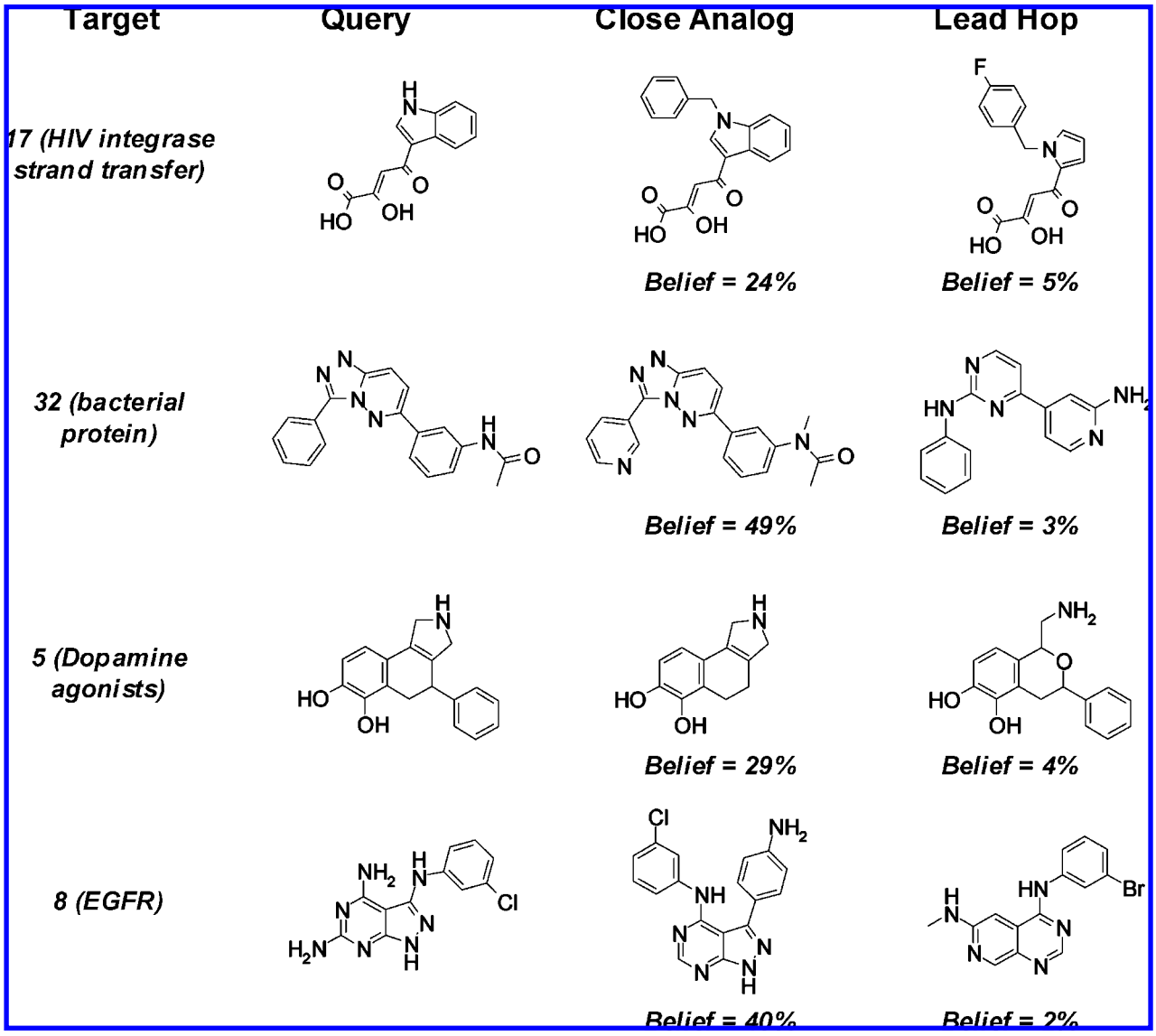

Figure 5. Examples of close analogs and lead hops, as well as their calculated joint beliefs using ECFP_6, Daylight, and ROCS as described in the text.

program (for which many hundreds or even thousands of analogs may exist in the compound collection), a simple virtual screen of the library will yield the closest analogs as the top hits. This trivial result must be corrected for by removing similar chemotypes, clustering, defining allowed similarity ranges, or other means. In contrast, analog searching around a compound for which few analogs exist will certainly return a ranked list, but all of the members may be so distantly related that there is little to no chance of finding active compounds. Calculating the joint belief allows an objective estimation of the confidence that a compound will be active, such that collections rich in analogs can be partitioned based on belief and collections lacking true analogs will not yield any hits with high belief.

CONCLUSIONS

The method of data fusion described in this work has a number of implications in the combination of results from a variety of computational methods for describing molecular similarity. The method relies on the availability of a large data set of molecular pairs for which both of the members of the pair have had accurate activity data measured. Applying principles from Dempster−Shafer belief theory to this large data set, we have developed a computational framework which allows a common scale to be developed for fundamentally distinct approaches to defining molecular similarity. This not only allows the identification of the most preferred method for combining the results of different measures, but also provides a quantitative expectation of the percentage of active compounds that will be contained in a combined library. This important aspect of the current work enables a researcher to select the level of belief required for their particular application.

Although the current work is focused on combination of results of similarity calculations, applications of belief theory to the combination of results can be found in many areas of cheminformatics and computational chemistry. A belief theory approach could be used in a variety of contexts, including the search for both very high and very low similarity compound analogs, bioisostere identification, and virtual screening. The potential to rationally combine very disparate descriptions of chemical similarity and generate a quantitative estimate for

success should extend the utility and interpretability of these tools and enhance drug discovery efforts.

METHODS

Data Collection and Curation. Activity data were derived from our corporate electronic database for the following 23 protein targets: 5-HT1A, Akt-1, CB2, Chk-1, Cot kinase, D4, DPP-4, farnesyl transferase, ghrelin, glucocorticoid receptor, HCV polymerase, HCV protease, HDAC, Jnk-1, KDR, Lck kinase, MCH, MetAP2, NNRs, P2x7, PARP-1, V1b, and VR1. All data were obtained from in vitro enzymatic or ligand-competition assays, and IC_{50} and K_I values were used consistently within a given target data set, but no distinction was made between data sets. Compounds with less than 10 or greater than 60 heavy atoms were excluded from the analysis. An average of 2871 compounds were obtained per target, resulting in a total of 66 026 compounds with potencies ranges from 0.1 nM to 100 μM. An average of 600 compounds per target exhibited IC_{50} or K_I values below 100 nM, while an average of 195 compounds per target exhibited IC_{50} or K_I values below 10 nM. The resulting physicochemical properties of these 66,026 compounds were the following: MW = 395 ± 104, ClogP = 3.5 ± 1.6, and polar surface area (PSA) = 83 ± 34.

Similarity Comparisons. Pairwise similarity measurements were performed using Tanimoto coefficients with Daylight fingerprints (version 4.83),[4] MACCS keys, extended connectivity fingerprints (ECFPs), and functional class fingerprints (FCFPs) with lengths 2, 4, and 6.[7] MACCS, ECFP, and FCFP similarities were calculated within Pipeline Pilot v5.0 (Scitegic) using 1024 bits. Maximum common substructures (MCS) were identified using the MCS capabilities in OEChem (OpenEye Scientific), and Tanimoto similarities were calculated using heavy atom counts in the MCS and each comparator molecule.[35] Finally, Tanimoto similarities derived from three-dimensional atomic volume overlaps were obtained using the ROCS software[11] from OpenEye. A maximum of 50 conformations of each comparator molecule was generated using OMEGA, and ROCS was then used to identify the conformers that yielded the highest volume overlap between the two comparator molecules. For the purposes of this analysis, the average of the ROCS "shape" and "color" Tanimoto was used. A correlation matrix relating the degree of correlation between the 10 similarity methods used in this work in given in the Supporting Information (Table S1).

As it was computationally impractical to calculate exhaustive pairwise similarity measures for all 66 026 compounds in the data set (representing more than 4 billion comparisons), the number of comparisons was performed on a subset of the data. First, "active" query molecules were defined as those compounds that exhibited an IC_{50} or K_I value between 1 and 10 nM. An average of 104 compounds from each target met this criterion. Next, each of these active query molecules was compared to a random 1% of compounds tested against the same target. This resulted in an average of 28 comparisons for each active molecule, generating a total of 66 687 pairwise comparisons over all 23 targets in the test set (i.e., 104 actives × 28 comparisons × 23 targets ~ 66 687 pairs). In order to simulate the ability of the similarity measures to discriminate known actives from a large data set of inactives, the similarity measurements described above were also performed on a set of 104 429 pairs of compounds that were randomly selected from our corporate repository. It is important to note that only 3 of the 104 429 "random pairs" (less than 0.003%) exhibited Daylight similarity values in excess of 0.8, indicating that these random compounds are not biased by coincidental selection of highly similar compounds. In addition, these molecules shared very similar overall physicochemical properties to the 66 026 compounds derived from activity data against 23 targets (see above), with MW = 390 ± 99, ClogP = 3.5 ± 1.6, and PSA = 82 ± 36. Thus, these random selections appear to be reasonably matched compounds that approximate true noise.

Probability Assignment Curves. The probability assignment curves shown in Figure 2 were constructed in the following manner. As described above, "active" query molecules are defined as those compounds that exhibited an IC_{50} or K_I value between 1 and 10 nM. A test compound was then considered "active" if the IC_{50} or K_I value was within 1 log unit of that for the query molecule. Thus, for a query molecule with an IC_{50} of 5 nM, all test compounds with IC_{50} values between 0.5 and 50 nM were considered active. Analysis of the 66 687 target-specific compound pairs using this criterion for equipotency yielded 8082 active pairs (the IC_{50} values for the remaining 58 605 pairs differed by more than 1 log unit and were considered inactive). As a result, the probability of randomly selecting an active compound pair from this data set is approximately 12%. This is substantially higher than the expected frequency from a large and diverse compound collection, as hit rates from high-throughput screening of million-member compound collections tend to be substantially less than 1%. To simulate the background noise similarity for inactive compound pairs, the similarity values derived for the 104 429 compound pairs randomly selected from our corporate repository were added as inactive pairs to each of the 23 targets to result in a final background hit rate of 0.1%.

Similarity values were then binned in 0.1 unit intervals and the number of total pairs (target-specific data set plus noise) and active pairs were counted for each similarity bin. The resulting plots of similarity versus fraction active closely resemble standard dose–response curves (see Figure 2), and the data were therefore fit to sigmoidal curves of the following form:

$$B_i = \frac{F_{\max}}{1 + 10^{(SC_{50} - x_i)\cdot \text{slope}}}$$

where B_i is the belief that a pair of compounds are equally active, x_i is the similarity value for the ith similarity measure, $F_{\max}$ is the maximum value for the fraction active, SC_{50} (by analogy to the IC_{50}) is the similarity cutoff at which 50% of the maximum fractional active is observed, and the slope (by analogy to the Hill slope) is the steepness of the curve. The fitted values for all of the similarity metrics used in this study are listed in Table 1. In order to assess the uncertainty in the sigmoidal curve parameters, probability assignment curves were recalculated after omitting data from each of the 23 targets in the set. It was observed that the values for $F_{\max}$, SC_{50}, and the slope exhibited standard deviations of less than 5% in all cases.

Performance Evaluation. The utility of each metric (both alone and in combination) to retrieve active pairs from the database was evaluated by calculating the area under the re-

ceiver operating characteristic curve (AU-ROC) and the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC)[31] curve. To guard against saturation errors in calculation of the BEDROC value, 100 test sets of 500 randomly chosen active pairs (from the total of 8082 active pairs) and 99 500 inactive pairs (from the total of 163 034 inactive pairs) were created. This ratio of active to total molecules ($R_a = 0.005$) ensures that the *maximum* error in the BEDROC value (Δ_{max}) is less than 5% using a value of 20 for α. AU-ROC and BEDROC values ($\alpha = 20$) were then calculated for each of these subsets and averaged. The final values (with standard deviations) are given in Table 2.

External Test Set. An external test set of 134 reported lead hops against 28 protein targets was constructed (see Supporting Information Table S2).[33,34] As with the training set above, exhaustive pairwise comparisons were performed with all 134 compounds using the 10 different similarity metrics (Daylight, ROCS, MCSS, MACCS, FCFP_2, FCFP_4, FCFP_6, ECFP_2, ECFP_4, and ECFP_6; see above), yielding 310 active pairs (defined as compound pairs derived from the same target) in a total data set of 8911 pairs (a listing of all pairwise comparisons along with calculated similarities is given in Supporting Information Table S3). To minimize saturation error in calculation of the AU-ROC and BEDROC values, a bootstrap approach was implemented in which 100 test sets of 50 randomly chosen active pairs in a background of 8651 total comparisons were created. AU-ROC and BEDROC values ($\alpha = 20$) were then calculated for each of these subsets and averaged. The final values (with standard deviations) are given in Table 3.

Supporting Information Available: Tables of similarity measure correlations, the AU-ROC and BEDROC results for all comparisons, the list of 134 compounds used in the validation study, and the full listing of all pairwise similarity comparisons for the validation study. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* **2007**, *12*, 225–233.
(2) Renner, S.; Schneider, G. Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem* **2006**, *1*, 181–185.
(3) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity. *J. Med. Chem.* **2002**, *45*, 4350–4358.
(4) Leo, A.; Weininger, A. *Daylight Chemical Information Systems*, Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA, 1995.
(5) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **2002**, *1*, 882–894.
(6) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
(7) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* **2005**, *10*, 682–686.
(8) Nettles, J. H.; Jenkins, J. L.; Williams, C.; Clark, A. M.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Flexible 3D pharmacophores as descriptors of dynamic biological space. *J. Mol. Graph. Modell.* **2007**, *26*, 622–633.
(9) Fischer, J. R.; Rarey, M. SwiFT: an index structure for reduced graph descriptors in virtual screening and clustering. *J. Chem. Inf. Model.* **2007**, *47*, 1341–1353.
(10) Hessler, G.; Zimmermann, M.; Matter, H.; Evers, A.; Naumann, T.; Lengauer, T.; Rarey, M. Multiple-ligand-based virtual screening: methods and applications of the MTree approach. *J. Med. Chem.* **2005**, *48*, 6575–6584.
(11) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *48*, 1489–1495.
(12) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74–82.
(13) Carosati, E.; Mannhold, R.; Wahl, P.; Hansen, J. B.; Fremming, T.; Zamora, I.; Cianchetta, G.; Baroni, M. Virtual screening for novel openers of pancreatic K(ATP) channels. *J. Med. Chem.* **2007**, *50*, 2117–2126.
(14) Franke, L.; Schwarz, O.; Muller-Kuhrt, L.; Hoernig, C.; Fischer, L.; George, S.; Tanrikulu, Y.; Schneider, P.; Werz, O.; Steinhilber, D.; Schneider, G. Identification of natural-product-derived inhibitors of 5-lipoxygenase activity by ligand-based virtual screening. *J. Med. Chem.* **2007**, *50*, 2640–2646.
(15) Muchmore, S. W.; Souers, A. J.; Akritopoulou-Zanze, I. The use of three-dimensional shape and electrostatic similarity searching in the identification of a melanin-concentrating hormone receptor 1 antagonist. *Chem. Biol. Drug. Des.* **2006**, *67*, 174–176.
(16) Vogt, M.; Bajorath, J. Introduction of a Generally Applicable Method to Estimate Retrieval of Active Molecules for Similarity Searching using Fingerprints. *ChemMedChem* **2007**, *2*, 1311–1320.
(17) 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*, 2206–2219.
(18) 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*, 2193–2205.
(19) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model.* **2006**, *46*, 462–470.
(20) Baber, J. C.; Shirley, W. A.; Gao, Y.; Feher, M. The use of consensus scoring in ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 277–288.
(21) Zhang, Q.; Muegge, I. Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. *J. Med. Chem.* **2006**, *49*, 1536–1548.
(22) Feher, M. Consensus scoring for protein-ligand interactions. *Drug Discov. Today* **2006**, *11*, 421–428.
(23) Glen, R. C.; Adams, S. E. Similarity Metrics and Descriptor Spaces - Which Combinations to Choose. *QSAR Comb. Sci.* **2006**, *25*, 1133–1142.
(24) Dempster, A. P. Upper and Lower Probabilities Induced by a Multivalued Mapping. *Ann. Stat.* **1967**, *28*, 325–339.
(25) Hughes, C. The representation of uncertainty in medical expert systems. *Med. Inform. (London)* **1989**, *14*, 269–279.
(26) van Ginneken, A. M.; Smeulders, A. W. Reasoning in uncertainties. An analysis of five strategies and their suitability in pathology. *Anal. Quant. Cytol. Histol.* **1991**, *13*, 93–109.
(27) Shafer, G. R. Perspectives on the Theory and Practice of Belief Functions. *Int. J. Approx. Reas.* **1990**, *3*, 1–40.
(28) Srivastava, R. P.; Shenoy, R. P.; Shafer, G. R. Propagating beliefs in AND-trees. *Int. J. Intel. Syst.* **1995**, *10*, 647–664.
(29) Shafer, G. R. The combination of evidence. *Int. J. Intel. Syst.* **2007**, *1*, 155–179.
(30) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.
(31) Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
(32) Hooper, G. A calculation of the credibility of human testimony. *Phil. Trans. R. Soc.* **1699**, *21*, 359–365.
(33) Martin, Y. C. Pharmacophore Modeling: 1 - Methods. In *Comprehensive Medicinal Chemistry II*; Mason, J. S., Ed.; Elsevier: Oxford, 2006; pp 119–147.
(34) Martin, Y. C. Pharmacophore Modeling: 2 - Applications. In *Comprehensive Medicinal Chemistry II*; Mason, J. S., Ed.; Elsevier: Oxford, 2006; pp 515–536.
(35) Hagadone, T. R. Molecular Substructure Similarity Searching: Efficient Retrieval in Two-Dimensional Structure Databases. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 515–521.

CI7004498