JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Perspective

pubs.acs.org/jcim

Combination of Similarity Rankings Using Data Fusion

Peter Willett*

Information School, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, United Kingdom

■ INTRODUCTION

Similarity searching is one of the most common techniques for ligand-based virtual screening and involves scanning a chemical database to identify those molecules that are most similar to a user-defined reference structure using some quantitative measure of intermolecular structural similarity.[1−5] The similar property principle states that molecules that are structurally similar are also likely to exhibit similar properties,[2,6,7] and hence ranking a database in order of decreasing similarity with a bioactive reference structure is expected to highlight database structures that have high a priori probabilities of exhibiting the same activity.

Any similarity measure has three principal components: the representation that is used to describe each of the structures that are to be considered; the weighting scheme that is used to assign weights to different parts of the structure representation that reflect their relative degrees of importance; and the similarity coefficient that is used to quantify the degree of resemblance between two suitably weighted representations. Multiple approaches have been described for each of these three components, resulting in a potentially vast number of similarity measures that could be used for virtual screening. This multiplicity has resulted in many studies that seek to determine which measures are most effective using some quantitative measure of screening performance. While certain types of measures are known to provide a reasonable level of retrieval effectiveness (e.g., those making use of Pipeline Pilot ECFP_4 fingerprints,[8] of occurrence-based frequency weighting,[9] or of the Tanimoto similarity coefficient[10]), there is a general recognition that there is no single similarity measure that will provide optimal screening in all circumstances.[11−14] This situation has been well summarized by Sheridan and Kearsley, when they note that "we have come to regard looking for 'the best' way of searching chemical databases as a futile exercise. In both retrospective and prospective studies, different methods select different subsets of actives for the same biological activity and the same method might work better on some activities than others".[15]

Given that there is no single, consistently effective, similarity searching method that can be used to rank a database in decreasing similarity order, it has been suggested that multiple searches should be carried out. The results of these individual searches are then combined (or merged or fused) into a single ranking that is the final output presented to the user for subsequent compound selection and biological testing. Such combination approaches have been used not only in similarity searching and other types of ligand-based virtual screening, where they are normally referred to as *data fusion*,[16] but also in structure-based virtual screening, where they are normally referred to as *consensus scoring*.[17] There is hence much interest in combining these two approaches to virtual screening.[18−21]

This perspective discusses the use of data fusion in similarity-based virtual screening; other similarity-related applications of data fusion include the analysis of molecular diversity and of structure−activity landscapes inter alia.[22−25] A previous review provided an overview of data fusion methods in ligand-based virtual screening up to 2005.[16] However, the technique has now been so widely adopted that it is difficult to provide a comprehensive review, with a Google Scholar search in late 2012 for "Data Fusion" AND "Virtual Screening" identifying over 350 post-2005 items. Accordingly, after a description of the basic approach and the various ways in which it can be implemented in the next section, the review focuses on two specific aspects of data fusion: the various fusion rules that have been described in the literature for combining rankings; and work at Sheffield that seeks to provide a rationale for why data fusion methods work in practice. The focus here is the combination of similarity rankings but many of the methods described here are equally applicable to the combination of the rankings that result from the use of, e.g., machine learning techniques for screening chemical databases.

■ DATA FUSION: THE BASICS

Data fusion is the name given to a body of techniques that combine multiple sources of data into a single source, with the expectation that the resulting fused source will be more informative than will the individual input sources.[26−28] The idea of combining different information sources is hardly a novel one since people need to use more than just a single sense (i.e., the ability to see, hear, feel, smell, or taste), for many, if not most, situations in daily life, and since the decisions of groups of people, such as a jury, are usually considered to be superior to those of an individual. However, it was only in the 1980s, principally in response to the United States' Department of Defense's need to identify and to track military targets, that interest developed in computational methods for combining information that had been obtained in digital form from different types of sensor.[29] A formal definition of data fusion used by the Department of Defense is as follows: "data fusion is a multilevel, multifaceted process dealing with the automatic detection, association, correlation, estimation, and combination of data and information from multiple sources".[30] It will be realized that this definition is very broad, and methods are now being used for many, more specialized, purposes; for example, Dasarathy listed 40 review articles summarizing work on data fusion for applications in astronomy, database systems, neurocomputing, remote sensing, and speech processing inter alia.[31]

Received: November 14, 2012
Published: January 8, 2013

```
For $i$ := 1 to $n$
    For $j$ := 1 to $N$
        Compute the similarity, $S_i(d_j)$, for the $j$-th database-structure using the $i$-th similarity scoring function
For $j$ := 1 to $N$
    Use a fusion rule, $F$, to combine the set of $n$ scores $\{S_i(d_j)\}$ for the $j$-th database-structure to give its fused score, $FS_j$
Rank the database in decreasing order of the fused scores, $FS_j$
```

Figure 1. Basic procedure for fusion of ranked similarity lists.

Data fusion can take place at three different levels. In data-level fusion, the raw data generated by multiple sensors are combined directly, or after the use of appropriate normalization to ensure that the data are commensurate; in feature-level fusion, feature extraction methods are used to generate derived representations of the raw data that are then combined; and decision-level fusion involves combining decisions that have been arrived at independently by the available sensors.[29] In the virtual screening context, a sensor is a computational procedure that computes a score for each of the structures in a database and then ranks the structures in decreasing order of these scores. The sensor outputs are hence sorted lists of similarity scores (or the rankings corresponding to those sorted scores) that are assumed to reflect the structures' relative probabilities of exhibiting the biological activity of interest. If multiple sensors are available, then data-level fusion can be used to combine the scores or rankings resulting from the individual screening methods, with the expectation that the fusion will result in a greater, or at least a more consistent, degree of clustering of the true actives toward the top of the ranking than will the use of a single screening method.

The basic procedure that we shall consider here is hence as shown in Figure 1. In this algorithm, it is assumed that there are n different ways available for computing the similarity score for each of the N structures in the database that is being searched. The fusion rule, F, is a procedure that takes as input the n similarity scores for each database structure and next fuses these to produce a single fused score for each such structure. The resulting list of N fused scores is then ranked in decreasing order to yield the final output of the procedure.

The procedure shown in Figure 1 is far simpler than those required in many other applications of data fusion, but it still provides a wide range of ways in which fusion can be implemented. Three factors are of particular importance: the nature of the n different database searches that comprise the first part of the procedure; the nature of the scores that are merged by the fusion rule; and the nature of the fusion rule that is used to combine the n scores for each database structure. The first two factors are described in the remainder of this section, with the third factor being described in the next section.

Whittle et al. note that two basic approaches are in use for similarity-based virtual screening:[32] *similarity fusion*, where n different similarity measures are used with a single reference structure; and *group fusion*, where n different reference structures are used with a single similarity measure. These two approaches could, of course, be combined, although this does not appear to have been studied in detail thus far.

Similarity fusion was first reported in the late 1990s in papers from research groups at Merck and at Sheffield. Sheridan and co-workers described facilities available in the Merck in-house similarity searching system,[33,34] which offered a range of different types of fingerprint (both 2D and 3D), with intermolecular similarities based on the Dice coefficient.[35] When a search was carried out, the user had the option of identifying two specific fingerprint-types, and data fusion took place using either similarity or rank data. In the former case, the system computed the arithmetic mean of the Dice values for the two chosen fingerprints and sorted the database structures into decreasing order of the mean similarities. Alternatively, the database was sorted into decreasing similarity order for each of the two fingerprints, and then, a note made of the minimum rank (i.e., nearer the top of the ranking) for each database structure; the structures were finally sorted into decreasing order of these minimum ranks. Thus, the fusion in this case involved fusing similarity rankings produced using different types of structure representation. Shortly afterward, Ginn et al. at Sheffield reported the use of 2D, 3D, and spectral descriptors with different types of similarity coefficients, so that fusion here involved fusing similarity rankings produced by varying both the structure representation and the similarity coefficient.[36,37] They found that the best results were generally obtained using ranked data and the SUM fusion rule (vide infra). Both the Merck and Sheffield groups concluded that fused searches performed on average as well as, or slightly better than, the best individual searches; since the latter often varied from one search to another, it was concluded that the use of a fusion rule would generally provide a more consistent level of search performance than would a single similarity measure.

The alternative, group-fusion approach was first described in detail by Willett and co-workers at Sheffield,[32,38] drawing on several earlier studies that had considered the use of multiple molecules for searching.[39−42] Group fusion involves combining the rankings (or similarities) that are obtained when multiple reference structures are used to search a database with a single similarity measure, e.g., using Daylight fingerprints and the Tanimoto coefficient. Whittle et al.[32] and Hert et al.[38] studied the effectiveness of group fusion, in comparison with both conventional similarity searching and similarity fusion. They found that better results were obtained from using similarity scores, rather than rank positions, and that the best results were obtained using the MAX fusion rule (vide infra). This combination (similarity data and the MAX rule) far out-performed conventional similarity searching and has since been widely adopted as a standard approach to similarity searching with multiple reference structures.[10] Analyses using a range of bioactivity classes demonstrated that group fusion was particularly effective with structurally diverse sets of actives, which present severe problems to conventional similarity searching and similarity fusion.[32,43] Group fusion is often implemented using the hits obtained in an initial HTS experiment, where the activity data can be erroneous. However, studies of turbo similarity searching,[43] where group fusion is carried out with molecules of presumed, but untested, activity suggest that the presence of false hits need not invalidate the use of the procedure (although there will clearly be some point at which the volume of erroneous data will outweigh the benefits of fusion).

The scores that are merged by the fusion rule can be of two types: either the structure's actual similarity, as computed using some particular similarity measure; or the rank of the structure when all of the N computed similarities are ranked in decreasing order of the scores for the chosen similarity measure. There has been some discussion as to the relative merits of these two types of score.[32,37,44,45] As discussed in the next section, most of the fusion rules that are available are applicable to both types of score. Since ranks are derived from similarities, the former involve a loss of information, but the loss is often ignored since it is of less consequence than in other applications. This is because the principal requirement of a fusion method is to enable the medicinal chemist to decide whether a specific database structure should be considered for further analysis, and the structure's position in the ranking is normally sufficient to enable the chemist to make this decision. Moreover, fusing similarity scores when similarity fusion is being used can introduce some degree of bias since the distribution of scores from different measures may not be the same, even if the measures yield the same range of scores (e.g., values between zero and unity for many of the association coefficients that are used for similarity searching) or if range-scaling is applied to the raw similarity scores to ensure that this is the case. Rank-based fusion rules are hence often used in practical applications, at least when similarity fusion is adopted.

FUSION RULES

A fusion rule takes as input n ($n \geq 2$) sets of N similarities or ranks and produces as output a single such set, normally as a ranked list from which the top-ranked database structures can be selected for further analysis. Fusion rules can either be unsupervised, meaning that the fusion rule operates directly on the similarity or rank information, or supervised, meaning that an additional training procedure is required. To date, unsupervised methods have been more widely used, not least because similarity searching is normally carried out at an early stage in a lead-discovery program, when only limited structural and activity data are available; supervised methods, conversely, become more appropriate as a research program progresses and starts to generate additional structure−activity data. However, the availability of such data means that a wide range of alternative approaches to virtual screening are available, e.g., machine learning methods based on naïve Bayesian classifiers, random forests, or support vector machines inter alia.[46]

Unsupervised Fusion Rules. Many of the unsupervised fusion rules that have been used in virtual screening derive from work in the discipline of information retrieval (IR).[47−50] The central task in IR is the retrieval of documents from a text database that are relevant to a user's query, a situation that is clearly analogous to the virtual screening task of retrieving molecules from a chemical database that have the same biological activity as a reference structure.[2] An early study of the use of data fusion methods was undertaken by Belkin et al., who described a range of fusion rules that are based on simple arithmetic operations and that have since been widely adopted in IR.[51] Examples of such arithmetic fusion rules are shown in Figure 2 where, as before, d_j denotes the jth database structure and where there are n sets of similarity scores or ranks to be fused (when considering these rules, it should be remembered that a large similarity score, e.g., 0.95 for a Tanimoto coefficient, will correspond to a small rank, i.e., at or near to the top of a ranked list).

Fusion rule	Formula
MAX	$\max\{S_1(d_j), S_2(d_j)\ldots S(d_j)\ldots S_n(d_j)\}$
MIN	$\min\{S_1(d_j), S_2(d_j)\ldots S(d_j)\ldots S_n(d_j)\}$
SUM	$\frac{1}{n}\sum_{i=1}^{n} S_i(d_j)$
MED	$\mathrm{median}\{S_1(d_j), S_2(d_j)\ldots S(d_j)\ldots S_n(d_j)\}$
ANZ	$\frac{1}{p}\sum_{i=1}^{n} S_i(d_j)$
MNZ	$p\sum_{i=1}^{n} S_i(d_j)$
EUC	$\sqrt{S_1(d_j)^2 + S_2(d_j)^2 + \cdots S_i(d_j)^2 \cdots + S_n(d_j)^2}$
RRF	$\sum_{i=1}^{p} \frac{1}{R_i(d_j)}$

Figure 2. Fusion rules.

The first two rules, MAX and MIN, involve assigning a database structure d_j a score that is the maximum, or the minimum, similarity that it has achieved in the complete set of n similarity searches. The next two rules compute average scores: SUM is the arithmetic mean of the individual scores (or equivalently the arithmetic sum of them), while MED eschews the mean in favor of the median; the geometric or harmonic means might also be used for this purpose. Virtual screening often uses some cutoff similarity or rank position, with only those molecules above this threshold (e.g., the top-1% of the ranking) being passed on for further processing, this essentially meaning that all database structures below the threshold are assigned a score of zero. Let p ($p \leq n$) be the number of such nonzero similarity scores for a database structure; then the rules ANZ and MNZ are obtained by multiplying SUM by either $1/p$ or p, respectively, so that these rules focus on database structures that occur frequently above the threshold. The EUC rule views the set of similarity scores for each database structure as an n-dimensional vector and, then, computes the Euclidean norm for the vector.

The fusion rules described thus far can be used for group fusion (where the same similarity measure is employed in all the searches) with either the original similarity scores, $S_i(d_j)$ (the form used in Figure 2) or the ranks, $r_i(d_j)$, obtained from sorting the similarities into decreasing order. Their use for similarity fusion (where different similarity measures are used) depends on the extent to which the different measures that are to be combined have the same range and distribution (vide supra); if this is not the case, then the rules should only be used with rank data. The final rule in Figure 2, RRF (reciprocal rank fusion), was reported as being effective for IR by Cormack et al.[52] and involves summing over the p rankings in which a database structure occurs to yield a score analogous to a harmonic mean. This rule is more appropriate for use with ranked data, since its use with similarity data requires all of the similarity scores to be nonzero (or the inclusion of an arbitrary constant in the denominator of the expression) to avoid an undefined value for the fusion score

In addition to the rules in Figure 2, Chen et al. briefly describe two fusion rules—Borda fusion and Condorcet fusion—that have been used in IR and that are based on voting behavior in elections.[53] Analysis of Borda fusion demonstrated that it was equivalent to the SUM rule when rank data were used, while Condorcet fusion was found to be computationally expensive and to be generally inferior to the simple rules discussed above. Zhang and Muegge describe another voting method that considers just the top 20% of a similarity ranking and then allocates votes in a manner analogous to the RRF rule.[54] Tan et al. describe a very simple rule, called parallel selection, in which molecules are selected from the top of each ranked list in turn until the desired output size is required.[19] This "round-robin" approach was found to be more effective than the other fusion rules tested in a study by Svensson et al. that combined ligand-based and structure-based screening methods.[21]

The last unsupervised fusion rule to be described here involves the use of Pareto ranking.[55] Pareto ranking has previously found application in chemoinformatics as a fitness function in studies of genetic algorithms[56−58] but can also be used to combined similarity search outputs, with the Pareto rank of a database structure being defined as the number of structures that are ranked above it (or, equivalently, have a higher similarity score) in all of the n ranked lists. This definition inevitably results in large numbers of tied ranks. These are broken either by secondary application of the SUM fusion rule or by a second (or subsequent) round of Pareto ranking using the number of molecules ranked above it in all but one (all but two, all but three, etc.) ranked lists. Cross et al. suggested that the latter was the better of the two Pareto rules and that it was also superior to the SUM ranks rule.[55]

Given the range of rules that have been described, it is reasonable to ask whether there is any difference between them. Ginn et al. summarized early work on data fusion at Sheffield and suggested that using the SUM rule with rank data was generally the most effective for similarity fusion,[37] a conclusion that seems to have been adopted as a benchmark in some subsequent studies of supervised fusion methods (vide infra). Later work at Sheffield, and subsequently elsewhere, has focused on group fusion, with the initial studies by Hert et al.[38,59] suggesting the general effectiveness of the MAX rule with similarity data. This has been extensively tested and is now probably the fusion rule of choice.[10,60−64] Support for this approach comes from a study by Nasr et al. of multiple-molecule query methods that used no less than 41 publicly available data sets taken from the MUV, NCI AIDS, and WOMBAT databases and from literature QSAR analyses.[65] Of the group-fusion rules that were studied, the MAX rule worked best with similarity data, and unsurprisingly, the MIN rule worked best with rank data. All the experiments used the Tanimoto coefficient for computing intermolecular similarities (as do most of the others considered in this review), but Nasr et al. note that entirely comparable results were obtained when the overlap coefficient, the cosine coefficient, or the Euclidean distance[35] were used in its place.

Chen et al. have recently reported a detailed comparison of the fusion rules in Figure 2 when used in group fusion with MDDR and WOMBAT data.[53] Their experiments involved comparing not just the rules themselves but also the number of reference structures involved in a search, and the fraction of each ranked similarity list involved in the fusion stage (i.e., whether all of the ranked database structures should be used or just some small percentage of the top-ranked structures). Hardly surprisingly, fusion was found to be most effective when the largest numbers of reference structures were employed, but their other conclusions were more unexpected. First, fusion was most effective if just the top-ranked database structures (e.g., the top 1−5%) in each ranked list were used. Chen et al. ascribe this to the similar property principle, under which these molecules are the ones that have the greatest probability of having the same activity as the reference structure(s) whereas the overwhelming majority of the database structures occurring lower down the rankings will be inactive and contribute little information to the fusion procedure. Second, while the MAX similarity and MIN rank rules performed well (as expected), the most effective searches were achieved when the RRF rule was used to fuse the ranked lists.

Chen et al. suggest that the observed effectiveness of the RRF rule arises from the relationship that exists between the probability that a database structure is active and its reciprocal rank in a similarity search. Figure 3, which is based on results presented by Hert et al.,[66] shows the relationship for similarity searches in the MDDR database. The probabilities in this figure were obtained by averaging the number of times that a database structure at a specific rank position had the same activity as the reference structure, when averaged over 8294 individual similarity searches for 11 different MDDR activity classes. It will be seen that the probabilities follow a hyperbolalike distribution, falling rapidly away as the ranked list is traversed. By definition, plotting the reciprocal of rank against rank results in a hyperbola, and thus, fusion-rule scores based on summing the reciprocal ranks of multiple similarity searches, such as the $1/R_i(d_j)$ term in RRF, may be expected to mimic the probability of activity in a similarity search. Chen et al. show that the scores $FS(d_j)$ resulting from RRF do indeed mirror probability curves (such as that shown in Figure 3) more closely than do the scores resulting from other unsupervised rules, hence providing a clear rationale for the observed effectiveness of the RRF fusion rule.[53]

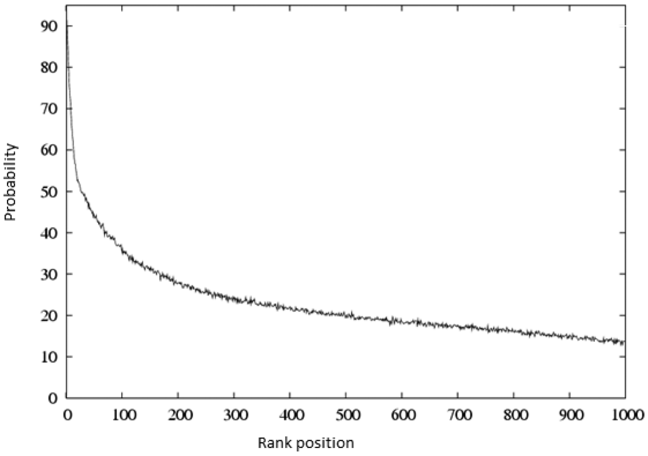

Figure 3. Plot of probability of activity at a specific rank position against rank position, with results averaged over 8294 similarity searches of the MDDR database.

The principal focus of the study by Chen et al. was the use of group fusion, but the effectiveness of the RRF rule was also demonstrated in similarity fusion studies using the MDDR, WOMBAT, and MUV data sets.[53] This suggests that RRF, which was first described (in the IR context[52]) several years after the detailed comparisons of Whittle et al.[32] and Hert et al.,[38] is of quite general applicability and that it should henceforth be considered as an alternative to SUM or MAX, which have been the rules of choice in most previous applications of data fusion.

Supervised Fusion Rules. The unsupervised fusion rules considered thus far require as input just the n sets of N similarity scores; this section considers supervised fusion rules that are based on sets of structures for which bioactivity data are available. Specifically, these approaches seek to identify a quantitative relationship between the structural similarity of a pair of molecules, as determined using some particular similarity measure, and their corresponding biological similarities. Then, given a reference structure of known activity and the similarity of a database structure to it in multiple similarity searches, one can calculate the probability that the database structure has the same activity, with fusion being effected by combining the probabilities for the individual similarity searches.

The first such rule to be described in the literature was the conditional probability approach of Raymond et al.[67] Assume that a search has been carried out using an active reference structure. The similarity scores are converted to probabilities of activity using Bayes rule

$$P(A|S_x) = \frac{P(S_x|A)P(A)}{P(S_x)}$$

where $P(A|S_x)$ is the conditional probability of a database structure being active given that it has a similarity S_x to the reference structure using some particular similarity measure (or, more generally, some particular virtual screening method). $P(A|S_x)$ is calculated from the following: $P(S_x)$, the unconditional probability of obtaining similarity score S_x; $P(A)$, the unconditional probability of a database structure being active; and $P(S_x|A)$, the conditional probability that a similarity score of S_x will occur given that the reference structure and a database structure have the same activity. $P(S_x)$ is estimated by fitting a probability distribution to the similarity scores obtained using a particular similarity measure and a data set containing known actives and inactives. $P(S_x|A)$ is estimated by clustering the database into activity classes and then performing all pairwise intracluster similarity comparisons. $P(A)$ is independent of the similarity measure that is being used and can hence be treated as an arbitrary constant, meaning that the Bayes rule simplifies to

$$P(A|S_x) = \frac{P(S_x|A)}{P(S_x)}$$

It is then assumed that the n sets of N similarity scores that are to be fused are statistically independent, so that the probability of activity of each structure can be obtained by multiplying together its n $P(A|S_x)$ scores (which were fitted to five-parameter exponential curves for ease of use). The database structures are then ranked in decreasing order of the resulting probabilities to give the final fused output. Two points should be noted. First, since $P(A)$ has been ignored, $P(A|S_x)$ is not a true probability and thus takes values greater than unity. Second, the independence assumption is clearly seriously flawed, but the authors argue that its use results in an effective fused ranking; specifically, fusion of three different similarity measures (both 2D and 3D descriptors were considered) resulted in more effective screening than did fusion using the simple SUM rule, although there was no significant difference when just two similarity measures were fused.

Baber et al. describe an approach to data fusion that is based on logistic regression.[68] Given a set of training data, the approach uses a sigmoidal function of the form

$$P_{ij} = \frac{1}{1 + e^{-z}}$$

to compute the probability P_{ij} that the jth database structure has the same activity as the reference structure given that the two structures have a similarity of x using some particular similarity measure. In this equation, $z = a + bx$, and the constants a and b are computed using the available training data. If n measures are to be fused, z takes the form $z = a + \sum b_i x_i$ and the logistic regression is refitted; coefficients are not transferred between models. For this reason, a large amount of training data is required. The fused score is then given by

$$FS_j = \frac{1}{1 + e^{-(a + \sum b_i x_i)}}$$

SUM fusion and logistic regression consistently outperformed any single measure, with the latter being generally the better of the two fusion rules.

Muchmore et al.[69] use belief theory, which provides a systematic approach to evidence combination that enables the calculation of a degree of belief given the evidence available from different sources (i.e., from different similarity searches in the present context). The approach is based on a large file of in-house IC_{50} data, in which all the intermolecular similarities have been calculated using a range of different similarity measures and in which a pair of molecules is defined as being active if one has nanomolar potency and if their potencies differ by less than one log unit. Plots of the percentage of such active pairs having a particular similarity against similarity were found to closely resemble standard dose−response curves, and the plots were hence fitted to sigmoidal curves (as in the work of Baber et al.,[68] whereas Raymond et al.[67] used an exponential curve to fit the similarity and activity data (vide supra)). These curves were of the form

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

Here, B_i is the belief that a pair of molecules are equally active using the ith similarity measure, x_i is the similarity value for the ith similarity measure, $F_{\max}$ is the maximal probability that any two molecules will be active (and is hence analogous to the maximal observed dose in a dose−response curve), SC_{50} is the similarity value for which half of the observed maximal probability is observed, and slope is the steepness of the curve. Data fusion is effected using Hooper's rule. Given multiple beliefs B_1, B_2, etc. (representing the belief functions for a particular database structure given different similarity searches), the cumulative belief is computed to be

$$\text{cumulative belief} = 1 - \prod_{i=1}^{n} (1 - B_i)$$

Muchmore et al. found that this fusion rule yielded rankings that were comparable to the SUM rule but noted the greater interpretability of the belief approach.[69] It has been used subsequently in a lead-hopping application that combines the results of ROCS, Daylight, and ECFP_6 similarity searches.[70] It has also been extended to allow for the fusion of rankings resulting from both ligand-based and structure-based virtual screening, with comparative experiments demonstrating the general superiority of this combined approach to the unsupervised MIN, MAX, and SUM fusion rules.[20]

Tiikkainen et al. present a data fusion method based on bioactivity data from the NCI-60 data set.[63,71] This data set contains the results of screening small molecules against 60 cancer cell lines, and was used to compute a measure of biological similarity for each pair of molecules by matching the corresponding cytotoxicity profiles. An empirical conditional probability distribution was then developed as follows. Define

$$N(c, a, r, b) = N(c \geq a; r \geq b)$$

as the number of molecule-pairs that have a chemical similarity $\geq a$ using chemical similarity measure c and that have a biological similarity $\geq b$ using biological similarity measure r. Then let

$$F(c = a|r = b) = \frac{N(c, a, r, b)}{N(c, a, r, -1.0)}$$

so that F represents the fraction of all the molecule-pairs which have a chemical similarity of at least a using similarity measure c and which also have a biological similarity of at least b using similarity measure r. This can be extended to two similarity measures by considering the fraction of all the molecule-pairs that have the following: chemical similarity at least a_1 using measure c_1; chemical similarity at least a_2 using measure c_2; and biological similarity at least b using measure r (in fact, their work used only a single measure of biological similarity so that r can be ignored). Tabulating F over a grid of the possible c_1 and c_2 scores then provides a probability distribution for biological similarity given a pair of chemical similarity scores, and hence a way of predicting biological similarity given two different similarity measures. Rather than using biological profiles, there seems no reason why this approach could not be used with activity data for a single biological target, hence providing a simple fusion procedure given sufficient training data.

Mention should also be made here of two probabilistic combination methods that have been described for the identification of polypharmacology but that may also be applicable to virtual screening. Keiser et al. report a detailed analysis of similarities in MDDR activity classes, comparing the distributions of interclass and intraclass similarities to enable the computation of the significance of an observed similarity between two sets of molecules.[72] These ligand-based data are then used to identify significant relationships between the biological receptors corresponding to each of the activity classes. Yera et al. report an approach that again uses distributions of similarity scores to compute the probability of observing a specific similarity value, but with these distributions, and the corresponding probabilities, being specific to an individual molecule rather than being averaged over a data set.[73] The observed similarities for a molecule of unknown activity with the known actives for some activity class are then used to compute the probability that the unknown molecule will exhibit the current activity.

Yera et al. report some limited experiments comparing their approach with those described by Keiser et al. and by Muchmore et al., but there have been no extended comparisons to date such as those by Nasr et al. and Chen et al. for unsupervised fusion rules (vide supra). The increasing availability of large volumes of linked chemical and biological data means that detailed comparisons of supervised rules are likely to become more common in the future, and such comparisons will hopefully include studies of the extent to which the success of the rules depends on the precise nature of the training data that is available; for the present, the unsupervised rules such as SUM, MAX, or RRF provide extremely simple ways of combining the results from multiple similarity searches.

■ WHY DOES DATA FUSION WORK?

Many experimental evaluations have been carried out (as summarized in the earlier review by Willett[16] and in the more recent studies reported here) of the effectiveness of data fusion. These studies have demonstrated that the technique works, in that screening effectiveness is normally at least as good as, and often superior to, that resulting from use of a single similarity search method, especially with group fusion; moreover, the effectiveness varies less from search to search than when a single search is carried out, meaning that data fusion will result in a more consistent level of performance. It would clearly be of interest to be able to elucidate the reason(s) for the enhanced performance, and this section summarizes recent work at Sheffield that has sought to determine how and why fusion works in practice.[74−76]

Experimental Rationale. A basic premise of applications of data fusion in virtual screening is that multiple similarity searches are better able to retrieve bioactive molecules than can single searches. The many previous studies demonstrate that this is normally the case in practice, but without any systematic test having been made of the validity (or otherwise) of the underlying premise. As noted previously, data fusion has been much studied in IR, and Spoerri described experiments to test the premise in the context of matching a query against a text database using multiple search engines.[77] He found that a given document had a greater probability of being relevant to a query the more search engines that retrieved that document, with this probability increasing rapidly in line with the number of search engines retrieving it, a phenomenon that he referred to as the Authority Effect. Given the many similarities between IR and chemoinformatics,[78] Holliday et al. sought to assess the applicability of the Authority Effect since its existence would then provide a firm basis for the use of fusion methods.[76]

The experiments of Holliday et al. involved searches for 11 different MDDR activity classes and for 14 different WOMBAT activity classes, with average results being computed using 10 different searches for each such activity class. Five different fingerprints (BCI bit-strings, Daylight fingerprints, ECFP_4 fingerprints, MDL keys, and Unity fingerprints) and five different similarity coefficients (Cosine, Forbes, Russell-Rao, Simple Match, and Tanimoto) were available, giving a total of 25 different similarity measures. A reference structure was searched using each of these different similarity measures, and a threshold applied to retrieve just the top 1% or 5% of each the resulting rankings. A note was then taken of how many database structures occurred in a single ranking, in two rankings, in three rankings, etc. (note that while this procedure is clearly reminiscent of similarity fusion, it does not involve combining the multiple rankings with an actual fusion rule to generate a final output ranking).

The same basic behavior was observed in all the searches: very many structures were retrieved in just a single search, and then the numbers dropped away rapidly as the number of searches increases. The skewed nature of the data suggests a power-law relationship[79,80] between the numbers of structures and numbers of searches, i.e., a relationship of the form

$$O = \frac{a}{n^b}$$

where n is the number of similarity searches, O is the overlap (i.e., the number of structures appearing in n of the 25 sets of nearest neighbors), and a and b are constants. If a power-law holds then a plot of $\log(O)$ against $\log(n)$ will give a straight line with a slope of $-b$. An example of such a plot (using logarithms to base 10) is shown in Figure 4, which is based on averaging the top-1% results over similarity fusion searches of all the MDDR activity classes. While the plot shows some deviations from an exact straight line (as is often the case in power-law studies[79]), the general form of the relationship is clear, with an r^2 value of 0.959 and with a b value of −2.17 (i.e., close to the inverse square relationship that is often referred to as Lotka's Law[81]).

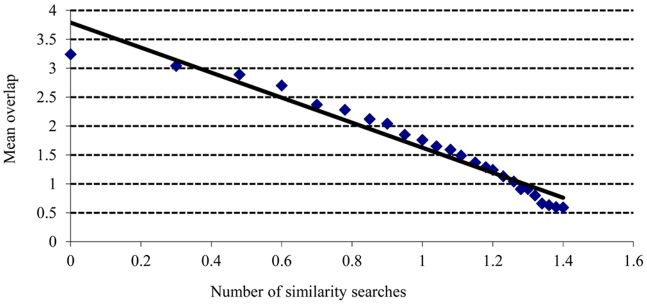

Figure 4. Log–log plot of the overlap (i.e., the mean numbers of database structures) retrieved in a given number of similarity searches using the top 1% of the search rankings for MDDR activity classes.

If data fusion is to be effective as a screening strategy then the increasingly small sets of retrieved structures retrieved by increasingly large numbers of searches should contain a greater fraction of actives and this is indeed found to be the case: while the search precision is initially very low for molecules retrieved by just a few searches, it then increases very rapidly as the number of searches moves toward the maximum. This behavior is shown in the log–log plot of Figure 5 where the precision follows an approximate cubic relationship with the number of searches. Holliday et al. hence concluded that the probability of activity of a database structure increases in line with its frequency of retrieval in multiple similarity searches, hence providing a rationale for the use of fusion methods in similarity-based virtual screening. They also reported analogous experiments using a group fusion-like approach in which searches for multiple reference structures were carried out using the Tanimoto coefficient and ECFP_4 fingerprints. Drawing on previous work by Mitzenmacher,[82] Holliday et al. presented a mathematical model that generated a power law distribution in fair agreement with the overlaps observed in their virtual screening experiments (such as those shown in Figure 4). However they were unable to provide a comparable approach for modeling the precision results (such as those shown in Figure 5), and suggested this as a topic warranting further research.

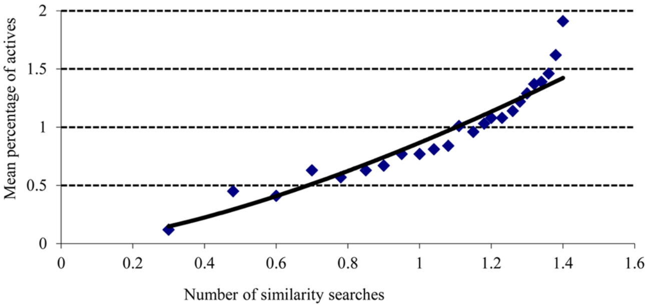

Figure 5. Log–log plot of the mean percentage of actives database structures retrieved in a given number of similarity searches using the top 1% of the search rankings for MDDR activity classes. The curve shown is the best-fitting cubic relationship.

Theoretical Rationale. Whittle et al. reported an extended attempt to develop an analytical model of data fusion when used for similarity searching[74] and then assessed the extent to which the model agreed with experimental data obtained from searches for MDDR activity classes.[75] The study assumed that normalized similarity values were being fused, and concentrated on SUM or MAX similarity fusion of ranked lists based on two different similarity coefficients; however, they also reported the extension of their methods to similarity fusion using multiple representations, and to group fusion.

The model is based on detailed consideration of the frequency distributions for the similarities between a reference structure and the database structures. Thus if two different similarity coefficients are being used (e.g., the Tanimoto coefficient and the Hamming distance) the model considers the numbers of similarities at least s in magnitude using the first coefficient and at least t in magnitude using the second coefficient (in a manner reminiscent of Tiikkainen et al. (vide supra)) and the extent to which the resulting area of similarity space is accessible to the particular fusion rule that is being used. For example, the shaded section of Figure 6 represents database structures that would be retrieved with a SUM fusion score of at least $s + t$ for the similarity measures X and Y. Whittle et al. showed that increases in screening effectiveness could be expected to result from data fusion when this area was more strongly populated by actives and/or more weakly populated by inactives than was the case when a conventional similarity search was carried out using just a single similarity coefficient. The analysis showed that this will generally be the case for the SUM and MAX rules if sufficient training data are available.

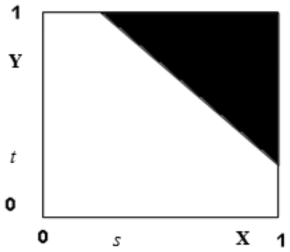

Figure 6. Data fusion using the SUM rule and similarity measures X and Y. The shaded area in the upper-right-hand part of the figures includes all the database structures that would be retrieved with a SUM fusion score of at least $s + t$.

The model is complex and requires a large volume of training data, with the fusion of just two similarity lists involving no less than eight different similarity distributions. Assume a similarity threshold is applied to the ranked lists resulting from each of the coefficients. Then, the following similarity distributions need to be considered: the distributions for the top-ranked actives and for the top-ranked inactives for each of the two similarity coefficients for both the matched and the unmatched database structures (where we define a matched (or unmatched) database structure to be one that is retrieved in the top part of both (or just one) of the ranked lists). Even slight variations in the relative contribution of each of these distributions can have a substantial effect on the performance of a fusion rule, and it is hence hardly surprising that the performance of data fusion can vary. That said, it must be emphasized that such variations are normally on a smaller scale than those observed when a single similarity search is executed, which is why one of data fusion's principal merits is that it provides a more consistent level of screening than does conventional similarity searching. The analytical model is not applicable if there are many unmatched database structures in the two lists of top-ranked database structures that are to be fused. Whittle et al. hence developed a simulation model of similarity fusion that gave comparable results to the analytical model when the latter can be applied but that could also encompass unmatched lists.[83]

Despite the inherent complexities of the model, three general conclusions were drawn from the study: that the SUM rule is likely to out-perform the MAX rule in similarity fusion; that the converse applies in group fusion; and that group fusion is generally far superior to similarity fusion. These conclusions are all fully supported by previous studies of data fusion, but the large body of training data needed here means that the practical application of the model is likely to be very limited. Indeed, if training data are available then alternatives to data fusion may be preferable. For example, 2D or 3D substructure searching would normally be the methods of choice if a specific scaffold or pharmacophoric pattern had been implicated as being responsible for, or at least involved in, activity. Conventional similarity searching, and similarity fusion, are appropriate when just a single active is available, such as a literature or competitor compound. Group fusion is appropriate when multiple actives are available and can be surprisingly effective in operation as noted above. However, if multiple actives are available, as well as multiple inactives (which are, of course, generally available in abundance in a practical context), then machine learning tools can be used to exploit this information in arguably more appropriate ways for ligand-based virtual screening.[46,84] Thus, using MAX-based group fusion as a benchmark, comparative experiments have demonstrated the generally superior effectiveness of PLS-based discriminant analysis procedure in experiments using the WOMBAT database;[85] a support vector machine and binary kernel discrimination (BKD) in experiments using pesticide data from the Syngenta corporate database;[86] and BKD using a range of data from the MIV, NCI AIDS, and WOMBAT databases and from literature QSAR analyses.[65] An exception to this general trend is provided by a study of Chen et al.,[87] who used sets of actives from the MDDR database that were of high, medium, or low structural diversity. While a naïve Bayesian classifier[88] was generally superior to MAX-based group fusion, the latter was preferable when just a few, highly diverse active structures were available for database searching.

■ CONCLUSIONS

Data fusion has been widely used as way of enhancing the effectiveness of similarity-based virtual screening. It will generally result in a level of screening effectiveness that is at least comparable to, and often exceeds, that obtained in a conventional similarity search, and that is more consistent from one search to another. It first found application in similarity fusion, where a single reference structure is searched using different similarity measures; it has since been extended to encompass multiple reference structures, this group fusion approach normally being noticeably superior to similarity fusion. Both approaches can benefit from the availability of training data linking similarity scores and probabilities of activity, but unsupervised fusion rules are available that enable effective searches to be carried out even in the absence of such data. Given the general success of data fusion, it is natural to seek a satisfactory theoretical description of the approach that could yield further improvements in screening effectiveness. However, the attempts that have been made to date have been only partially successful, with Whittle et al. noting that that "whilst we believe that our theoretical and experimental work provide a useful interpretive tool for the analysis of data fusion results, the complexities that we have identified in the fusion process mean that it will be difficult to develop fusion methods that can be expected consistently to out-perform individual similarity searches".[75] Data fusion is hence an effective tool for similarity-based virtual screening, but one that still lacks a firm basis: the provision of same is hence perhaps the most obvious area for further study.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: p.willett@sheffield.ac.uk. Telephone: 0044-114-2222633. Fax: 0044-114-2780300.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

My thanks are due to Eleanor Gardiner for helpful comments and to Val Gillet, Claire Ginn, Jerome Hert, John Holliday, Evangelos Kanoulas, Nurul Malim, Christoph Mueller, Naomie Salim, and Martin Whittle for their contributions to the work on data fusion at Sheffield.

■ REFERENCES

(1) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.

(2) Willett, P. Similarity methods in chemoinformatics. *Ann. Rev. Inf. Sci. Technol.* **2009**, *43*, 3−71.

(3) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205−216.

(4) Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-art in ligand-based virtual screening. *Drug Discovery Today* **2011**, *16*, 372−376.

(5) Willett, P. Similarity-based data mining in files of two-dimensional chemical structures using fingerprint-based measures of molecular resemblance. *WIRES Data Mining Knowledge Discovery* **2011**, *1*, 241−251.

(6) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley: New York, 1990.

(7) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activities? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(8) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(9) Arif, S. M.; Holliday, J. D.; Willett, P. Analysis and use of fragment occurrence data in similarity-based virtual screening. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 655−668.

(10) Stumpfe, D.; Bajorath, J. Similarity searching. *WIRES Comput. Mol. Sci.* **2011**, *1*, 260−282.

(11) 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.

(12) Sheridan, R. P. Chemical similarity searches: when is complexity justified? *Expert Opin. Drug Discovery* **2007**, *2*, 423−430.

(13) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How similar are similarity searching methods? A principal components analysis of molecular descriptor space. *J. Chem. Inf. Model.* **2009**, *49*, 108−119.

(14) Bender, A. How similar are those molecules after all? Use two descriptors and you will have three different answers. *Expert Opin. Drug Discovery* **2010**, *5*, 1141−1151.

(15) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.

(16) Willett, P. Data fusion in ligand-based virtual screening. *QSAR Comb. Sci.* **2006**, *25*, 1143−1152.

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

(18) 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*, 489−497.

(19) Tan, L.; Geppert, H.; Sisay, M. T.; Gütschow, 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*, 1566−1571.

(20) Swann, S. L.; Brown, S. P.; Muchmore, S. W.; Patel, H.; Mert, P.; Locklear, J.; Hajduk, P. J. A unified, probabilistic framework for structure- and ligand-based virtual screening. *J. Med. Chem.* **2011**, *54*, 1223−1232.

(21) Svensson, F.; Karlen, A.; Skold, C. Virtual screening data fusion using both structure- and ligand-based methods. *J. Chem. Inf. Model.* **2012**, *52*, 225−232.

(22) Ruiz, I. L.; Urbano-Cuadrado, M.; Gómez-Nieto, M. A. Data fusion of similarity and dissimilarity measurements using Wiener-based indices for the prediction of the NPY Y5 receptor antagonist capacity of benzoxazinones. *J. Chem. Inf. Model.* **2007**, *47*, 2235−2241.

(23) Medina-Franco, J. L.; Maggiora, G. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. *Chem. Biol. Drug Des.* **2007**, *70*, 393−412.

(24) Maggiora, G. M.; Shanmugasundaram, V. Molecular similarity measures. *Method. Mol. Biol.* **2010**, *672*, 39−100.

(25) Medina-Franco, J. L.; Yongye, A. B.; Perez-Villanueva, J.; Houghten, R. A.; Martinez-Mayorga, K. Multitarget structure−activity relationships characterized by activity-difference maps and consensus similarity measure. *J. Chem. Inf. Model.* **2011**, *51*, 2427−2439.

(26) Hall, D. L.; McMullen, S. A. H. *Mathematical Techniques in Multisensor Data Fusion*; Artech House: Norwood MA, 2004.

(27) Liggins, M. E.; Hall, D. L.; Llinas, J. *Handbook of Multisensor Data Fusion: Theory and Practice*; CRC Press: Boca Raton FL, 2008.

(28) Mitchell, H. B. *Multi-Sensor Data Fusion: An Introduction*; Springer: Berlin, 2007.

(29) Hall, D. L.; Llinas, J. Introduction to Multisensor Data Fusion. *Proc. IEEE* **1997**, *85*, 6−23.

(30) Klein, L. A. *Sensor and Data Fusion Concepts and Applications*, 2nd ed.; SPIE Optical Engineering Press: Bellingham, 1999.

(31) Dasarathy, B. V. A representative bibliography of surveys in the information fusion domain. *Inf. Fusion* **2010**, *11*, 299−300.

(32) Whittle, M.; Gillet, V. J.; Willett, P.; Alex, A.; Loesel, J. Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: A comparison of similarity coefficients. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840−1848.

(33) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physicochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

(34) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128−136.

(35) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(36) Ginn, C. M. R.; Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Similarity searching in files of three-dimensional chemical structures: evaluation of the EVA descriptor and combination of rankings using data fusion. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23−37.

(37) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1−16.

(38) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.

(39) Shemetulskis, N. E.; Weininger, D.; Blankey, C. J.; Yang, J. J.; Humblet, C. Stigmata: an algorithm to determine structural commonalities in diverse datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862−871.

(40) Singh, S. B.; Sheridan, R. P.; Fluder, E. M.; Hull, R. D. Mining the chemical quarry with joint chemical probes: an application of latent semantic structure indexing (LASSI) and TOPOSIM (Dice) to chemical database mining. *J. Med. Chem.* **2001**, *44*, 1564−1575.

(41) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 746−753.

(42) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.

(43) 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 techniques to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model.* **2006**, *46*, 462−470.

(44) Yang, J.-M.; Chen, Y.-F.; Shen, T.-W.; Kristal, B. S.; Hsu, D. F. Consensus scoring criteria for improving enrichment in virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134−1146.

(45) Hsu, D. F.; Taksa, I. Comparing rank and score combination methods for data fusion in information retrieval. *Inf. Retriev* **2005**, *8*, 449−480.

(46) Goldman, B. B.; Walters, W. P. Machine learning in computational chemistry. *Ann. Reports Comput. Chem.* **2006**, *2*, 127−140.

(47) Spärck Jones, K.; Willett, P. *Readings in Information Retrieval*; Morgan Kaufmann: San Francisco, CA, 1997.

(48) Singhal, A. Modern information retrieval: A brief overview. *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.* **2001**, *24* (4), 35−43.

(49) Manning, C. D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, 2008.

(50) Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*, 2nd ed.; Addison-Wesley: Harlow, 2011.

(51) Belkin, N. J.; Kantor, P.; Fox, E. A.; Shaw, J. B. Combining the evidence of multiple query representations for information retrieval. *Inf. Proc. Manag.* **1995**, *31*, 431−448.

(52) Cormack, G. V.; Clarke, C. L. A.; Buettcher, S. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Boston, MA, July 19−23, 2009; pp 758−759.

(53) Chen, B.; Mueller, C.; Willett, P. Combination rules for group fusion in similarity-based virtual screening. *Mol. Inf.* **2010**, *29*, 533−541.

(54) 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.

(55) Cross, S.; Baroni, M.; Carosati, E.; Benedetti, P.; Clementi, S. FLAP: GRID molecular interaction fields in virtual screening. Validation using the DUD data set. *J. Chem. Inf. Model.* **2010**, *50*, 1442−1450.

(56) Gillet, V. J.; Khatib, W.; Willett, P.; Fleming, P. J.; Green, D. V. S. Combinatorial library design using a multiobjective genetic algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 375−385.

(57) Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1079−1087.

(58) Nicolaou, C. A.; Apostolakis, J.; Pattichis, C. S. De novo drug design using multiobjective evolutionary graphs. *J. Chem. Inf. Model.* **2009**, *49*, 295−307.

(59) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.

(60) Williams, C. Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Mol. Diversity* **2006**, *10*, 311−332.

(61) Kogej, T.; Engkvist, O.; Blomberg, N.; Muresan, S. Multi-fingerprint based similarity searches for targeted class compound selection. *J. Chem. Inf. Model.* **2006**, *46*, 1201−1213.

(62) Hristozov, D. P.; Oprea, T. I.; Gasteiger, J. Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 617−640.

(63) Tiikkainen, P.; Markt, P.; Wolber, G.; Kirchmair, J.; Distinto, S.; Poso, A.; Kallioniemi, O. Critical comparison of virtual screening methods against the MUV data set. *J. Chem. Inf. Model.* **2009**, *49*, 2168−2178.

(64) Abdo, A.; Salim, N.; Ahmed, A. Implementing relevance feedback in ligand-based virtual screening using Bayesian inference network. *J. Biomol. Screening* **2011**, *16*, 1081−1088.

(65) Nasr, R. J.; Swamidass, S. J.; Baldi, P. F. Large scale study of multiple-molecule queries. *J. Cheminf.* **2009**, DOI: 10.1186/1758-2946-1-7.

(66) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbour information. *J. Med. Chem.* **2005**, *48*, 7049−7054.

(67) Raymond, J. W.; Jalaie, M.; Bradley, P. P. Conditional probability: a new fusion method for merging disparate virtual screening results. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 601−609.

(68) 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.

(69) 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*, 941−8.

(70) Martin, Y. C.; Muchmore, S. Beyond QSAR: lead hopping to different structures. *QSAR Comb. Sci.* **2009**, *28*, 797−801.

(71) Tiikkainen, P.; Poso, A.; Kallioniemi, O. Comparison of structure fingerprint and molecular interaction field based methods in explaining biological similarity of small molecules in cell-based screens. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 227−239.

(72) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197−206.

(73) Yera, E. R.; Cleves, A. E.; Jain, A. N. Chemical structural novelty: on-targets and off-targets. *J. Med. Chem.* **2011**, *54*, 6771−6785.

(74) 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.

(75) 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.

(76) Holliday, J. D.; Kanoulas, E.; Malin, N.; Willett, P. Multiple search methods for similarity-based virtual screening: analysis of search overlap and precision. *J. Cheminf.* **2011**, DOI: 10.1186/1758-2946-3-29.

(77) Spoerri, A. Authority and ranking effects in data fusion. *J. Amer. Soc. Inf. Sci. Technol* **2008**, *59*, 450−460.

(78) Willett, P. Textual and chemical information retrieval: different applications but similar algorithms. *Inf. Res.* **2000**, *5* (2); available at http://InformationR.net/ir/5-2/infres52.html.

(79) Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **2005**, *46*, 323−351.

(80) Benz, R. W.; Swamidass, S. J.; Baldi, P. Discovery of power-laws in chemical space. *J. Chem. Inf. Model.* **2008**, *48*, 1138−1151.

(81) Pao, M. L. An empirical examination of Lotka's Law. *J. Amer. Soc. Inf. Sci.* **1986**, *37*, 26−33.

(82) Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. *Internet Math.* **2004**, *1*, 226−251.

(83) Whittle, M.; Gillet, V. J.; Willett, P. A simulation study of the use of similarity fusion for ligand-based virtual screening. In *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*; Lodhi, H., Yamanishi, Y., Eds.; IGI Global: Hershey PA, 2010; pp 46−59.

(84) Melville, J. L.; Burke, E. K.; Hirst, J. D. Machine learning in virtual screening. *Comb. Chem. High Throughput Screening* **2009**, *12*, 332−343.

(85) Askjaer, S.; Langgard, M. Combining pharmacophore fingerprints and PLS-discriminant analysis for virtual screening and SAR elucidation. *J. Chem. Inf. Model.* **2008**, *48*, 476−488.

(86) Wilton, D. J.; Harrison, R. F.; Willett, P.; Delaney, J.; Lawson, K.; Mullier, G. Virtual screening using binary kernel discrimination: analysis of pesticide data. *J. Chem. Inf. Model.* **2006**, *46*, 471−477.

(87) Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N. Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 53−62.

(88) Bender, A. Bayesian methods in virtual screening and chemical biology. *Methods Mol. Biol.* **2011**, *672*, 175−196.