ARTICLES

Analysis of Data Fusion Methods in Virtual Screening: Theoretical Model[†]

Martin Whittle,* Valerie J. Gillet, and Peter Willett

Department of Information Studies, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, U.K.

Jens Loesel

Pfizer Global Research and Development, Pfizer Limited, Ramsgate Road, Sandwich, Kent, CT13 9NJ, U.K.

Received December 20, 2004

This paper presents a theoretical model of how data fusion can be used to combine the results of multiple similarity searches of chemical databases. The model is based on frequency distributions of similarity values that are fused using a multiple integration over regions defined by the particular fusion rule that is being applied. For pairwise fusion, the resulting double integrals are straightforward to evaluate for simple model distributions. Similarity values for recovered-active and recovered-nonactive frequency distributions are independently modeled using a constant background, linearly biased terms, and a first-order correlated term. The model shows that two standard fusion rules can give performance enhancements in some cases but that the results of fusion are dependent on many factors that, taken together, can lead to seemingly inconsistent levels of enhancement.

INTRODUCTION

Data fusion was first developed for applications in signal processing and involves combining data from different sensors with the goal of increasing the value of that data in comparison with any of the individual sensors.[1] In the context of virtual screening, the different sensors correspond to different ways of ranking a database of previously untested molecules in order of decreasing probability of activity, with the aim of enabling better decisions to be made as to which few molecules should go forward for biological testing. There is extensive, and growing, literature describing its use for both ligand-based and structure-based virtual screening.[2−14]

In this and the companion paper,[15] we focus on the use of similarity searching[16] for ligand-based virtual screening. Similarity searching is based on the similar property principle[17] and involves finding those database molecules that are structurally most similar to an input reference structure. Data fusion can be used to enhance the effectiveness of similarity searching in two ways. If multiple similarity measures are available to quantify the degree of resemblance between the reference structure and each of the database structures, then a ranking of the database can be generated using each such measure, e.g., using different types of 2D fingerprint or of similarity coefficient. We refer to the combination of these various rankings to give a single output ranking, the normal practice in virtual screening applications, as *similarity fusion*. Alternatively, *group fusion* involves combining the multiple rankings obtained from a fixed similarity measure and multiple reference structures.[11,18]

Studies in many application domains have demonstrated that while data fusion can enhance system performance, however defined, the degree of enhancement is very variable, or even negative in some cases. This inconsistency has inspired several empirical studies that seek to determine why, and under what circumstances, fusion is effective.[6,12,19−26] In this paper, we present a novel theoretical model of both similarity fusion and group fusion. Our starting point was the belief that if one could describe the operation of data fusion in theoretical, rather than empirical terms, then it ought to be possible to develop new types of fusion rule that would ensure the maximization of search performance. However, as will be seen, our analysis has revealed that the operation of a data fusion system is far more complex than previously realized, involving subtle interactions between a range of factors that, taken together, severely complicate attempts to predict the effect of fusion on search performance. The principal contribution of our work is hence explanatory in character; that said, it has been possible to identify some circumstances in which fusion has at least the potential to increase the effectiveness of screening and also circumstances in which fusion is unlikely to provide any benefits.

METHODS

Data Fusion. In chemoinformatics, similarity fusion has usually involved the combination of nearest neighbor lists by using the ranks as a score. This provides a simple way of standardizing the similarity scores, but does mean that some information is lost. Here, we have used linear range scaling to define the score

$$S^*(i,j) = \frac{S(i,j) - S_{\min}(i)}{S_{\max}(i) - S_{\min}(i)} \qquad (1)$$

where $S(i,j)$ is the similarity between reference molecule i

[†] Dedicated to Professor Johann Gasteiger.

* Corresponding author e-mail: m.whittle@sheffield.ac.uk.

10.1021/ci049615w CCC: $33.50 © 2006 American Chemical Society
Published on Web 08/29/2006

and comparison molecule j, $S_{\max}(i)$ and $S_{\min}(i)$ are the maximum and minimum values in the ranked list for reference i, and $S^*(i,j)$ is the scaled similarity. This transformation maps the original values onto the range 0−1, but since it is a simple linear scaling there can be no difference between the rank positions of comparison structures scored using the scaled or unscaled version of the result. For group fusion we consistently found that better results were obtained by fusing the scaled similarities rather than the ranks.[11] The combination of lists then proceeds using one of several fusion rules. For similarity fusion,[8] we obtain m lists of N scaled similarity values $S_k^*(i,j)$ relating reference i and comparison structures j using similarity measures k and compute a fused score $S_{FUS}(i,j)$ for each recovered structure from

$$S_{FUS}(i,j) = F_{k=1}^{m}[S_k^{\ *}(i,j)] \tag{2}$$

If structure j is not found in one of the lists (as occurs when the ranked lists are truncated at some point, e.g., the top 5% of the ranking), then the scaled similarity is assumed to be zero.

The fusion rule, F, indicates the method used to combine the similarities. The SUM-rule has been frequently used for rank-based work, and in this case the expression becomes

$$S_{FUS}(i,j) = \sum_{k=1}^{m}[S_k^{\ *}(i,j)] \tag{3}$$

The MAX-rule is another basic scheme that is straightforward to implement. In this case F operates on the set of similarities to choose the maximum value for each j and eq 3 becomes

$$S_{FUS}(i,j) = \max\,[S_1^{\ *}(i,j), S_2^{\ *}(i,j), S_3^{\ *}(i,j), ..., S_m^{\ *}(i,j)] \tag{4}$$

The MAX-rule has proved particularly successful in our own studies of group fusion.[11] For this technique the combination is performed over lists obtained from a number of different reference structures rather than different measures, and the details are therefore slightly different.

Analytical Comparison of Data Fusion Rules Using Scaled Similarities. Efforts to understand the process of data fusion for results that are presented in terms of ranks involve a discrete description that leads to some difficult analysis.[26] Conversely, when scaled similarities are used, the process can be described in continuous space. By counting the number of compounds recovered with similarity values between sets of limits we can, after averaging over a given target class, establish a discrete frequency distribution $F_T(s)$ for the occurrence of similarity values s_{ij}. Since, in our case, these refer to scaled similarity values, each reference structure generates its own frequency distribution over range 0−1, and the distributions can in practice be averaged for a whole activity class. Division of these frequencies by the total number of compounds retrieved, N, gives the probability that compounds will be recovered over a given range of scaled similarity. If a large number of bins are used this can effectively be treated as a continuous function, and, for a given metric, we can thus define a probability density $\Phi_T(s)$ for the occurrence of each value, s_{ij}, of the similarity coefficient over a range between s and $s + \Delta s$

$$\Phi_T(s) = \operatorname*{Lim}_{\Delta s\to 0}\left\{\frac{\Pr(s < s_{ij} \le s + \Delta s)}{\Delta s}\right\} \tag{5}$$

where Pr() refers to the probability of retrieving a compound over the given range of similarity values. This is a true probability density function (pdf) since it obeys the criterion $\int_0^1\Phi_T(s)\mathrm{d}s = 1$, and it is related to the original discrete frequency distribution over a range between s and $s + \Delta s$ by

$$F_T(s + \Delta s/2) = N\int_s^{s+\Delta s}\Phi_T(s)\mathrm{d}s \tag{6}$$

where N is the total number of compounds retrieved. We can similarly define a probability distribution for the n active compounds in a given class

$$\phi(s) = \operatorname*{Lim}_{\Delta s\to 0}\left\{\frac{\Pr(s < s_{ij} \le s + \Delta s)_A}{\Delta s}\right\} \tag{7}$$

where $\Pr()_A$ refers to the probability of retrieving an active compound over the given range of similarity values. We will call this a *partial* probability density since $\int_0^1\phi(s)\mathrm{d}s = n/N$ is less than one, but it represents the total probability of finding an active compound in those retrieved. The similarity frequency distribution $f_A(s)$ for active compounds at the midpoint between s and $s + \Delta s$ is then recovered from

$$f_A(s + \Delta s/2) = N\int_s^{s+\Delta s}\phi(s)\mathrm{d}s \tag{8}$$

The partial probability density and associated frequency distribution for nonactives is

$$\Phi(s) = \Phi_T(s) - \phi(s)$$

$$f_{NA}(s + \Delta s/2) = N\int_s^{s+\Delta s}\Phi(s)\mathrm{d}s \tag{9}$$

Clearly, $\int_0^1\Phi(s)\mathrm{d}s = (N - n)/N$ is the probability that any compound chosen at random from those retrieved does not belong to the active class of interest. A sketch of representative functions, modeled by normal distributions, is shown in Figure 1, where the reference structure is located at maximum similarity. Virtual screening by similarity search in-

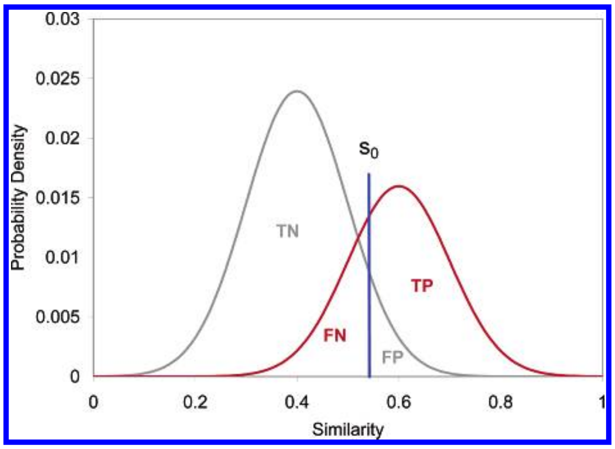

Figure 1. A sketch of the probability densities, ϕ, for recovered actives (red) and Φ, for recovered nonactives (gray), over the full range of available similarity values. To the right of the cutoff similarity value of S_0 the area under the red curve labeled *TP* represents the true positives, that under the gray curve *FP* the false positives, and to the left of the cutoff *TN* represents the true negatives and *FN* the false negatives.

volves ranking the contents of a database by similarity with respect to a known biologically active reference structure and retaining only the top few percent for physical screening. Here, these are represented by values greater than the cutoff at similarity value s_0, shown as a blue line. Of the structures with similarity values above this cutoff those under the red curve represent the *recovered-actives*, also known as true positives (*TP*). Those under the gray curve are *recovered*-nonactives or false positives (*FP*). To the left of the cutoff, those active structures that have not been recovered are called false negatives (*FN*), while those under the gray curve are correctly categorized as nonactive and labeled true negatives (*TN*).

In continuous space the available tools for manipulating such distributions are familiar. Thus the cumulative recall *R* and precision[27] *P* can be expressed in terms of the frequencies f_A and f_{NA} or equivalently in terms of the partial probability densities as defined:

$$R[N(s_0)] = TP/(TP + FN) = \int_{s_0}^{1}\phi(s)\mathrm{d}s/\int_{0}^{1}\phi(s)\mathrm{d}s \qquad (10)$$

$$P[N(s_0)] = TP/(TP + FP) = \int_{s_0}^{1}\phi(s)\mathrm{d}s/\int_{s_0}^{1}\Phi_T(s)\mathrm{d}s \qquad (11)$$

Here each measure is expressed for the total number of compounds collected at similarity s_0, which is equivalent to the rank $N(s_0)$

$$N(s_0) = (TP + FP) = N\int_{s_0}^{1}\Phi_T(x)\mathrm{d}x \qquad (12)$$

where, in this case, *N* would be the total number of compounds in the database. Our objective is now to extend these methods to analyze the effect of simple fusion rules such as SUM and MAX.

When two or more similarity measures are combined by fusion, the same formalism can be applied by extension to higher dimensions. Thus, for the combination of two similarity measures, we must consider two-dimensional distributions, representing the joint probability that a compound is found with similarity *x* according to one measure and *y* according to another[28]

$$\Phi_T(x,y) = \underset{\Delta x \to 0}{\mathrm{Lim}}\,\underset{\Delta y \to 0}{\mathrm{Lim}} \left\{\frac{\Pr(x < x_{ij} \le x + \Delta x)\ \mathrm{AND}\ \Pr(y < y_{ij} \le y + \Delta y)}{\Delta x \Delta y}\right\} \qquad (13)$$

Here, $\Phi_T(x,y)$ represents the bivariate probability density for all recovered compounds. By analogy with the one-dimensional case eqs 5−9, this can be split into contributions from the nonactive and active components to give $\Phi_T(x,y) = \Phi(x,y) + \phi(x,y)$ where $\Phi(x,y)$ and $\phi(x,y)$ are joint partial probability densities. Such two-dimensional similarity distributions can be displayed by plotting the similarity values from two measures one against the other. An example is shown in Figure 2 where the points represent scaled similarity values obtained from the comparison of bit-string representations using the Forbes coefficient plotted against those obtained using the Russell-Rao coefficient. For bit-strings of length *n*, with *a* bits set in the target string, *b* bits set in the comparison string, and *c* bits common to both, the Russell-Rao coefficient is defined by $S_R = c/n$ and the Forbes by $S_F = cn/ab$. In practice, these can be replaced by modified forms,[8] $S_{MR} = c/a$ and $S_{MF} = c/b$, but scaling makes the distinction superfluous. The integration used to obtain rank and recall must now be performed over a general region, Q, which determines how the measures are combined. For our purposes, the only restriction on this region is that it should contain the point (1,1) corresponding to the position of the reference compound, and an example of such a region is sketched in the figure. Figure 2 represents the superimposed results from many reference molecules and therefore represents an average distribution for the activity class. The overlaid illustration of a permissible region Q is the analogue of the precise similarity cutoff in Figure 1, and different values of rank will be reached for each reference structure at this point. Conversely, the extent of this region will be different for each reference structure if the same number of compounds are collected by each, but these differences are minimized for scaled similarities.

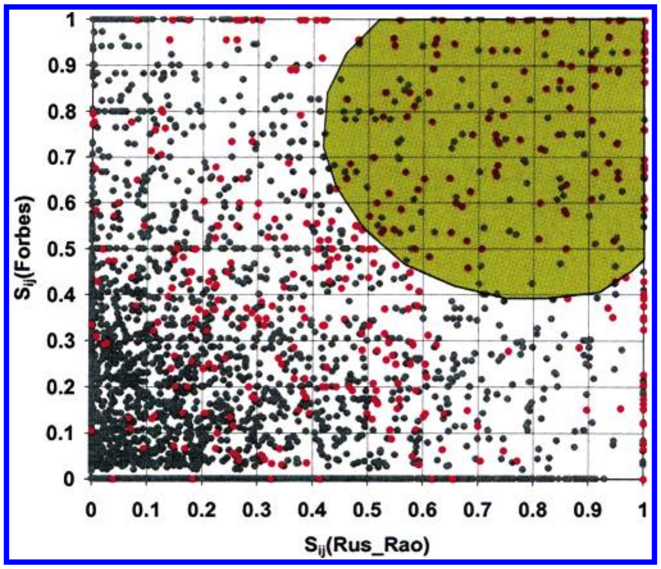

Figure 2. Paired similarity values S_{ij} for the scaled Forbes and Russell-Rao coefficients obtained using 83 wound healing agents from the MDDR as the active set. Results were taken up to rank 2000 and the whole of the MDDR (less the 83 known agents) was used as the nonactive set. (red circle) Recovered-active values; (black circle) recovered-nonactive values; only 10% of these points are shown for clarity. The colored area shows an example integration region Q containing the point (1,1).

The data shown in Figure 2 represents scaled similarity measurements obtained using BCI fingerprints[29] for each of the 83 reference structures taken from the wound healing agents activity class identified in the *MDL Drug Data Report* (MDDR) database.[30] The version of the database used here contained 102 443 compounds, and the top 2000 compounds retrieved were recorded in each case. The plot represents the recovered-active similarity values as measured by both coefficients in red and the recovered-nonactive values in gray. Because the latter significantly outnumber the former they were randomly sampled for display, and only about 1 in 10 are shown. The number of points in each of the grid squares shown is then proportional to the average density over each square, and so the plot is essentially a graphical representation of the bivariate densities $\phi(x,y)$ in red and $\Phi(x,y)$ in gray. The points within the colored region Q represent those retrieved using similarity fusion according to some unspecified rule.

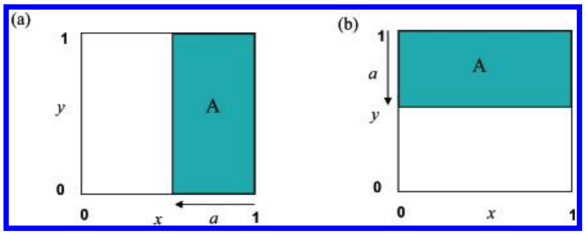

Figure 3. Sketch illustrating example integration regions for simple recall without fusion. The axes *x* and *y* represent similarity values obtained by using different measures. The colored areas show the integration regions, A, appropriate for collecting compounds with similarity values (a) for $x > 1-a$ and (b) $y > 1-a$.

Extending our notation to the higher dimensionality, the total number of compounds recovered for a given region Q, i.e., the rank, becomes

$$N(\mathrm{Q}) = N \int\int_{\mathrm{Q}} \Phi_T(x, y)\mathrm{d}x\mathrm{d}y \tag{14}$$

The number of actives recovered is obtained by integration of the recovered-active density over the same region:

$$n(\mathrm{Q}) = N \int\int_{\mathrm{Q}} \phi(x, y)\mathrm{d}x\mathrm{d}y \tag{15}$$

The recall and precision follow directly from these quantities using expressions equivalent to eqs 10 and 11

$$R(\mathrm{Q}) = \frac{\int\int_{\mathrm{Q}} \phi(x, y)\mathrm{d}x\mathrm{d}y}{\int\int_{\mathrm{T}} \phi(x, y)\mathrm{d}x\mathrm{d}y};\ P(\mathrm{Q}) = \frac{\int\int_{\mathrm{Q}} \phi(x, y)\mathrm{d}x\mathrm{d}y}{\int\int_{\mathrm{Q}} \Phi_T(x, y)\mathrm{d}x\mathrm{d}y} \tag{16}$$

where T signifies the total available region, and the recall is stated in terms of those compounds retrieved. Seen as a function of Q, the precision is a measure of the effectiveness of a given region for accumulating actives. The denominator is the cost of using the region in terms of rank, while the numerator gives the return in terms of actives retrieved. We now proceed by determining the shape of the regions, Q, appropriate for SUM and MAX fusion with the objective of comparing their relative effectiveness (for our application to similarity fusion, as in Figure 2, measured in terms of the number of actives recovered at a given rank) against single measure recovery and each other. For comparative purposes it will be convenient to express the single measure recall as a double integral, and, as the simplest, we start with the regions relevant to this calculation.

Single-Measure Recall. For distributions of a single variable *x*, the number of compounds recovered at a similarity value $x = s_0$ can be obtained using eq 12, and a parallel expression can be written for the number of recovered actives. When the density is expressed as a bivariate distribution, these quantities can be formally obtained using eqs 14 and 15, but the alternative variable *y* is redundant. Thus the integration for *y* is performed over all available values 0−1, while integrating over all values of *x* from 1 down to the required limit. We express this as $x = 1 - a$ since the rank and recall then conveniently increase with increasing values of the parameter *a*. The appropriate integration region is therefore bounded by the line $x = 1 - a$ as shown in Figure 3(a). We label this region A (equivalent to Q for this specific case) and will adopt the convention that the value of the associated integral takes the italicized symbol. Thus, for an arbitrary function $f(x,y)$ the resulting expression for the integral is

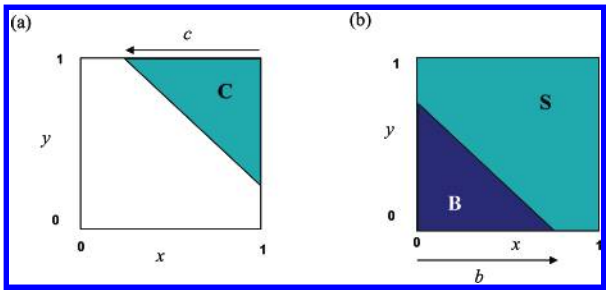

Figure 4. Sketch of regions for the integration of SUM-rule fusion, which is carried out in two stages. (a) For $c < 1$ the integral is evaluated for region C using the parameter *c* to define the limits. For $c > 1$ the integral is evaluated for region B using the parameter *b* to define the limits for $b < 1$. The integral for the required region, S, is obtained by subtraction from the total integral.

$$A(a) = \int_0^1 \mathrm{d}y \int_{1-a}^1 f(x,y)\mathrm{d}x \tag{17}$$

Figure 3(b) shows the equivalent region for the single similarity measure *y*; in this case the variable *x* is redundant, and the region is bounded by $y = 1-a$. The integral hence becomes

$$A(a) = \int_0^1 \mathrm{d}x \int_{1-a}^1 f(x,y)\mathrm{d}y \tag{18}$$

In general these routes are not equivalent, since either the *x*- or *y*-measure may give the better recall for real data. However, for our studies we have elected to always compare with the *x*-measure, and thus we have chosen examples to ensure that eq 17 always represents the larger values when there is nonequivalence.

SUM-Rule Fusion. Fusion using the SUM-rule takes the similarity values *x* and *y* for each compound, adds them together, and ranks the results. Compounds with a combined similarity higher than some given value are retained for consideration. Thus the corresponding region is such that $x + y \geq z$ where *z* is a combined similarity value. This region is shown in Figure 4(a) labeled C and extends to fill the whole box as the rank increases (i.e. as the similarity threshold decreases).

The area increments associated with changing values of *z* are thus far from regular. The variable *c*, shown in the diagram, increases with the rank and is therefore a "natural" choice for a single variable that describes the region. As the rank increases and lower values of similarity are considered, the region extends beyond the halfway mark as in Figure 4(b).

The integration is best performed in two sections, and since compounds with high similarity are normally recovered first, we start with the region above the diagonal as shown in Figure 4(a) and for which $0 \leq c \leq 1$. In terms of the variable *c* the equation of the diagonal line limiting the region C yields $x = 2 - c - y$, and thus for a general function the required integral for region C can then be written as

$$C = \int_{1-c}^1 \mathrm{d}y \int_{2-c-y}^1 f(x,y)\mathrm{d}x; \quad c \leq 1 \tag{19}$$

At $c = 1$, half of the box is covered, and we obtain a total contribution that we label T_1. For regions that extend below the diagonal it is convenient to use the variable *b* shown in

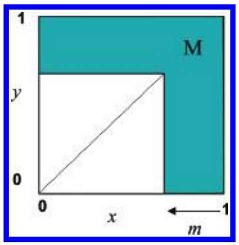

Figure 5. Sketch of an example integration region for MAX-rule fusion. Integration limits for the region M are defined using the parameter m, which represents the range indicated in the figure.

Figure 4(b) and to evaluate the integral B corresponding to the region B. For this variable, the equation of the limiting line is expressed as $x + y = b$, and we obtain

$$B = \int_0^b \mathrm{d}y \int_0^{b-y} f(x,y)\mathrm{d}x; \quad b \le 1 \tag{20}$$

The maximum extent of this region is reached when half of the box is covered and $b = 1$. At this point we obtain a total contribution for this half of the box that we label T_2. For this region of the box the variable c has the range $1 \le c \le 2$ corresponding to an extension beyond the limits of the box. Once this part of the integration has been evaluated for a given function the result can be expressed in terms of c by substituting with $b = 2 - c$. The value of integral over the whole box can now be obtained from $T = T_1 + T_2$, and the required integral for this range of c is then obtained by subtraction from this total. For any value of c, the total integral, S, (equivalent to Q for the case of SUM-rule fusion) can thus be expressed as

$$\begin{aligned} S &= C: \ c \le 1 \\ S &= T - B: \ c \ge 1 \end{aligned} \tag{21}$$

MAX-Rule Fusion. The MAX-rule ranks each compound using the highest value of x or y. Points with $x > y$ appear below the diagonal while points with $y > x$ appear above the diagonal, but when these values are ranked the largest individual values appear first whether they are x's or y's. The corresponding region, M, (equivalent to Q for the case of MAX-rule fusion) is therefore the right-angled shape shown in Figure 5.

As increasing values of rank are accessed, the intersection point of this region moves down the diagonal. Parallel to the definition of a for single measures, we define the parameter m describing the thickness of the "arms" shown in the diagram. Using this variable the required integral can be expressed as

$$\begin{aligned} M &= \int_0^1 \mathrm{d}y \int_{1-m}^1 f(x,y)\mathrm{d}x + \int_{1-m}^1 \mathrm{d}y \int_0^{1-m} f(x,y)\mathrm{d}x \\ &= \int_0^1 \int_0^1 f(x,y)\mathrm{d}x\mathrm{d}y - \int_0^{1-m} \mathrm{d}y \int_0^{1-m} f(x,y)\mathrm{d}x \end{aligned} \tag{22}$$

Other Fusion Rules. The MIN-rule[5] works such as the MAX-rule except that it scores each retrieved structure using the minimum value of the similarity. The collection of values less than a given magnitude thus corresponds to integration over the white L-shaped region in Figure 6. However, because the scored values are then ranked with the highest value top, the required integration region is the complement of this shape and is represented by the cyan square. It thus corresponds to the Boolean AND operator and is of interest because it selects all matched values (associated with those compounds appearing in both similarity lists) before any unmatched values (associated with those compounds appearing in only one of the lists) are chosen, and these are frequently the regions with the highest precision or concentration of actives.

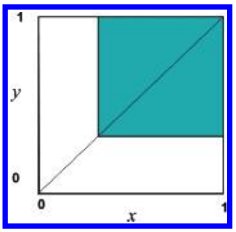

Figure 6. Sketch of the integration region appropriate for fusion of two similarity lists represented by the values x and y using the MIN-rule in two-dimensions.

Other simple fusion rules can also be considered; for example, weighted versions of the SUM-rule correspond to a triangular region bounded by a line with different slope. The PRODUCT rule[5] multiplies values of x and y before ranking the result, and the corresponding region is thus $xy \ge z$, which is bounded by a hyperbola. The CombMNZ-rule[20] is a weighted average scheme that multiplies the SUM-rule score for each retrieved structure by the number of nonzero values attributed to that structure. This gives results that are close to the SUM-rule but that cannot be expressed simply in terms of an integration region.

Fusion Enhancement. For similarity fusion to be effective, the number of actives recovered within a given region, that is up to a given rank under a specified fusion rule, must exceed the number retrieved using a single measure; that is either of the regions shown in part (a) or (b) of Figure 3. Specifically, suppose that the best single measure retrieves n actives at rank N and that n_F actives are recovered at rank N_F by fusing with data from some other measure. The ranks, N and N_F, are just the total number of compounds retrieved in each case. For these methods the retrieval precision is thus $P = n/N$ for the single measure and $P_F = n_F/N_F$ for the data fusion technique. These can be computed from the distributions using eq 16 with the appropriate regions. To make a comparison between methods we must compare the numbers of actives retrieved by each at the same rank $N_F = N$. Thus, a simple measure of enhancement, the ratio of actives retrieved by fusion to those recovered by the single measure, is directly related to the ratio of precision for those measures: $n_F/n = P_F/P$. For the SUM-rule, for example, this might suggest that a net gain would be achieved if the recovered-active values tend to cluster toward the top right-hand triangle of such a plot *more so* than the recovered-nonactive values. To examine more closely where advantages for fusion might lie we need to investigate what kind of density distributions lead to successful fusion for a given rule and, conversely, which if any of the fusion rules is best to use given some prior knowledge about the density distributions. There are innumerable different forms that a bivariate function can take, but we can begin to answer these questions by probing the results for a few idealized cases. In the following section we use a simple, analytically tractable model distribution designed to embrace the most basic underlying trends in such a function. We then apply the methods that we have described above to explore the consequences of these contributions for the retrieval enhancement that may be obtained using simple data fusion rules. Although the structure of this model is informed by real data, the objective of this exercise is not to fit results but rather to study the influence of basic functional variations on the effectiveness of data fusion. Applications of our methods to real experimental data will be described in a companion paper.[15]

MATHEMATICAL MODEL

Model Bivariate Distributions. The model functions introduced here lead to tractable expressions and are intended to broadly represent features that are clearly present in many real situations. Moreover, they have also made a qualitative appearance in several previous rationalizations of results of data fusion.[12,20,22] The terms chosen are insufficient to model similarity densities such as those shown in Figure 1, which represent values taken from a complete database, or distributions containing very dense regions,[31] but they may be more appropriate as approximations to the densities of the scaled and truncated distributions that we have studied,[11] since these correspond to the tails of the complete distributions at high similarity. However, we stress again that these are "toy" distributions whose primary purpose is to aid the study of data fusion itself. First, for measures that lead to recall values above the random expectation the active similarities must in some way be biased toward the target compound. This is a basic corollary of the similar property principle. We model this using a simple linear term. Second, two measures may give comparable rank ordering of similarity values leading to a nonzero Pearson's Correlation Coefficient when the values are compared. We model this with a cross term, which has the same underlying mathematical form as Pearson's Correlation Coefficient. The active similarity joint densities $\phi(x,y)$ can thus be expressed as the sum of a uniform density ϕ_0, a pair of linear bias terms with coefficients ϕ_x, ϕ_y describing the degree of asymmetry, and a term with the coefficient ϕ_{xy} describing the degree of correlation

$$\phi(x,y) = \phi_0 + \phi_x\left(x - \frac{1}{2}\right) + \phi_y\left(y - \frac{1}{2}\right) + \phi_{xy}\left(x - \frac{1}{2}\right)\left(y - \frac{1}{2}\right) \tag{23}$$

for the range $0 \leq x \leq 1$, $0 \leq y \leq 1$. These functional forms are centered on (1/2,1/2) so that changes in the coefficients do not affect the total integrated density. All parameters used in this paper satisfy the criterion that the density is everywhere non-negative; in particular we have ensured that $|\phi_x + \phi_y| \leq \phi_0$ and $|\phi_{xy}| \leq 2\phi_0$. The densities describing the frequency distribution of recovered active to recovered nonactive similarities can be expanded in a parallel manner about the same center with an independent set of coefficients

$$\Phi(x,y) = \Phi_0 + \Phi_x\left(x - \frac{1}{2}\right) + \Phi_y\left(y - \frac{1}{2}\right) + \Phi_{xy}\left(x - \frac{1}{2}\right)\left(y - \frac{1}{2}\right) \tag{24}$$

with similar restrictions to avoid negative densities. This symmetrical choice of function origin leads to integrals that, apart from the leading term, vanish when taken over the full range of values ($0 \leq x \leq 1$; $0 \leq y \leq 1$) thus avoiding awkward normalization issues. Equations 23 and 24 then represent a dissection of the general functions into basic components with symmetries that are expected to couple significantly with the integration regions of interest to data fusion. Specifically, bias of the recovered-active distributions relative to the recovered-nonactive distributions correspond to good recovery methods and, with the addition of correlation, provide the tools for investigating the prediction that the combination of comparably good but different methods is needed for successful fusion.[12,22] More complex distributions are best tackled by simulation techniques, as described later.

Substituting eq 23 for the recovered-active density into eq 15, each term generates its own contribution: the first corresponding to the uniform density, second and third to the linear bias terms, and the fourth to a contribution from the correlated term

$$n(\mathrm{Q}) = N[\phi_0 Q_0 + \phi_x Q_x + \phi_y Q_y + \phi_{xy} Q_{xy}] \tag{25}$$

where

$$Q_0 = \int\int_{\mathrm{Q}} \mathrm{dQ};\ Q_x = ;\ Q_x = \int\int_{\mathrm{Q}} \left(x - \frac{1}{2}\right)\mathrm{dQ}$$

$$Q_y = \int\int_{\mathrm{Q}} \left(y - \frac{1}{2}\right)\mathrm{dQ}\ ;\ Q_{xy} = \int\int_{\mathrm{Q}} \left(x - \frac{1}{2}\right)\left(y - \frac{1}{2}\right)\mathrm{dQ} \tag{26}$$

The number of nonactives recovered in a region can similarly be obtained by substituting eq 24 into eq 14 giving the total number of compounds retrieved in the same form:

$$N(\mathrm{Q}) = N[(\phi_0 + \Phi_0)Q_0 + (\phi_x + \Phi_x)Q_x + (\phi_y + \Phi_y)Q_y + (\phi_{xy} + \Phi_{xy})Q_{xy}] \tag{27}$$

For the model functions used here, the integrals Q_x, Q_y, and Q_{xy} can make no overall contribution (i.e. when the region Q represents the whole box) to the number of compounds retrieved. In this case the fraction of active compounds, λ, in the sample can be expressed as

$$\lambda = \frac{n}{N} = \frac{\phi_0}{\phi_0 + \Phi_0} \tag{28}$$

Since integration of the total probability density $\Phi_T(s) = \phi(s) + \Phi(s)$ over the whole region, T, must give 1, the uniform density parameters are constrained by $\phi_0 + \Phi_0 = 1$ and $\lambda = \phi_0$. Integration of eq 27 over the whole region gives the maximum rank or total number of compounds available, $N(\mathrm{T}) = N$. Integration of eq 25 over the whole region yields $n(\mathrm{T}) = n$, the total number of actives available in the whole region. For our purposes it is convenient to report results as the normalized quantities $N^*(\mathrm{Q}) = N(\mathrm{Q})/N(\mathrm{T})$ and $n^*(\mathrm{Q}) = n(\mathrm{Q})/n(\mathrm{T})$, which is effectively the recall since $n(T)$ represents the total number of available actives. Since, for the model functions used here, the only integrals that make a contribution to these totals are from the constant background term, we can identify $N(\mathrm{T}) = N(\phi_0 + \Phi_0) = N$ and $n(\mathrm{T}) = N\phi_0 = n$.

For distributions of this form, we can thus find the total number of compounds collected by a given region, i.e., the rank, and also the number of actives recovered for that region in terms of the integrals given in eq 26. In the following sections we evaluate these integrals for the specific regions

Table 1. Expressions for the Integrals Discussed in the Text with Subscripts Given in the Second Row as a Function of the Appropriate Parameter *a*, *c*, *b*, *or m* for Each of the Integrands Given in the Top Row[a]

integrand	1	$(x - 1/2)$	$(y - 1/2)$	$(x - 1/2)(y - 1/2)$
subscript	*0*	*x*	*y*	*xy*
A	a	$- 1/2a^2 + 1/2a$	0	0
C	$1/2c^2$	$- 1/6c^3 + 1/4c^2$	$- 1/6c^3 + 1/4c^2$	$1/24c^4 - 1/6c^3 + 1/8c^2$
B	$1/2(2 - c)^2$	$1/6(2 - c)^3 - 1/4(2 - c)^2$	$1/6(2 - c)^3 - 1/4(2 - c)^2$	$1/24(2 - c)^4 - 1/6(2 - c)^3 + 1/8(2 - c)^2$
T	1	0	0	0
M	$- m^2 + 2m$	$(1/2)m^3 - m^2 + (1/2)m$	$(1/2)m^3 - m^2 + (1/2)m$	$- (1/4)m^4 + (1/2)m^3 - (1/4)m^2$

[a] The integrals for single measure recall are labeled *A*. Integrals for the SUM-rule region labeled *S* in the text are obtained by combining expressions labeled *C* and *B* and the totals *T* (eq 21). Integrals labeled *M* refer to integrals for the MAX region.

of interest. The results for single-measure recall are needed for comparison, and we start with these. We then evaluate the integrals for the SUM-rule region. By comparing these methods of integrating the same functions we will then investigate fusion enhancement for the correlated and biased terms separately. Finally, in this section, we give results for the MAX-rule, which can be obtained by a parallel procedure.

Single-Measure Recall. To find the recall obtained using a single measure we need to perform the integration using the appropriate region shown in part (a) or (b) of Figure. Substituting for $f(x,y)$ in eq 17 with the model distribution for recovered actives $\phi(x,y)$ and scaling with the total number of compounds recovered, N, we obtain

$$n(a) = NA(a) = N\int_0^1 dy \int_{1-a}^1 \phi(x,y)dx \qquad (29)$$

from which the single-measure recall using the *x*-measure is obtained directly. Alternatively, we could use eq 18 to obtain the recall using the *y*-measure, but all of our comparisons will be made against the *x*-measure.

Equation 29 can now be split into a sum of terms corresponding to each of the terms from the expression eq 23. Thus we obtain a special case of eq 25

$$n(a) = N[\phi_0 A_0 + \phi_x A_x + \phi_y A_y + \phi_{xy} A_{xy}] \qquad (30)$$

where

$$A_0 = \int_0^1 \mathrm{d}y \int_{1-a}^1 \mathrm{d}x; \quad A_x = \int_0^1 \mathrm{d}y \int_{1\text{-}a}^1 \left(x - \frac{1}{2}\right)\mathrm{d}x;$$

$$A_y = \int_0^1 \mathrm{d}y \int_{1\text{-}a}^1 \left(y - \frac{1}{2}\right)\mathrm{d}x;$$

$$A_{xy} = \int_0^1 \mathrm{d}y \int_{1-a}^1 \left(x - \left(\frac{1}{2}\right)\right)\left(y - \frac{1}{2}\right)\mathrm{d}x \qquad (31)$$

These integrals are just a special case of eq 26 for the single-measure integration region. Similarly, an expression for the single measure rank can be obtained from eq 27 by substituting for Q_0, Q_x, Q_y, and Q_{xy} with A_0, A_x, A_y, and A_{xy} from eq 31. Expressions for these and all the other analogous integrals discussed later are given in Table 1. Consulting this it will be seen that A_y and A_{xy} vanish for all *a* when integrated using the region given in Figure 3a. Clearly, A_y vanishes by symmetry, and the zero value of A_{xy} reflects the fact that the recall obtained by single measure *x* is independent of single measure *y* even if these values are correlated.

Fusion Using the SUM-Rule Region. Substituting for $f(x,y)$ in eqs 19 and 20 with each of the integrands used in eq 24 we can thus obtain expressions for *B* and *C* as functions of the variable *c*. These are collected in Table 1. Using eq 18 we can then evaluate integrals for the SUM-rule region, *S*, over the full range of *c*. These are again a special case of eq 26 and are labeled S_0, S_x, S_y, and S_{xy}. For example

$$S_x(c) = -\frac{1}{6}c^3 + \frac{1}{4}c^2; \quad 0 \le c \le 1$$

$$S_x(c) = -\frac{1}{6}(2 - c)^3 + \frac{1}{4}(2 - c)^2; \quad 1 \le c \le 2 \qquad (32)$$

since, in this case, $T_x = 0$ (see eq 21). Using these relations we can write down analogous expressions to eqs 25 and 27 to obtain the recall and rank using the SUM-rule.

Fusion Enhancement. As discussed previously, the analysis of fusion enhancement requires that we compare the recall obtained using a given fusion rule with that obtained using a single measure at the same rank. Crucially, it can be seen by comparing the expressions for *C* or *B* for the SUM-rule and *A* for single-measure recall in Table 1 that the functions obtained for the integration region appropriate to a single measure are quite different in form to those obtained using the region appropriate for SUM-rule fusion; they depend on different powers of the appropriate parameter. In particular, the contribution arising from the correlated term, A_{xy}, vanishes for all values of *x* and *y*. Because of this property, it becomes possible to obtain an analytical result for data fusion enhancement when correlated distributions are used. At this point it is therefore convenient to discuss the results for the correlated term separately from those of the linearly biased terms, which require numerical solution.

Correlated Contribution. In this section we consider SUM-rule fusion using only the correlated term in association with the constant background. Thus, in effect, we set $\phi_x = \phi_y = \Phi_x = \Phi_y = 0$. Using the integrals from Table 1 for a single measure region we can evaluate eqs 25 and 27 to write the rank and the number of actives recovered as

$$N(a) = N[A_0 + (\Phi_{xy} + \phi_{xy})A_{xy}] \qquad (33)$$

$$n(a) = N[\phi_0 A_0 + \phi_{xy} A_{xy}] \qquad (34)$$

where we have also used $\phi_0 + \Phi_0 = 1$. Similarly, for the case of SUM fusion we obtain

$$N_S(c) = N[S_0 + (\Phi_{xy} + \phi_{xy})S_{xy}] \qquad (35)$$

$$n_S(c) = N[\phi_0 S_0 + \phi_{xy} S_{xy}] \qquad (36)$$

To make a comparison between SUM fusion and single measure recall, we must now compare the values of $n_S(c)$ and $n(a)$ at equivalent rank $N_S(c) = N(a)$. Setting $\Psi_{xy} = (\Phi_{xy}$

+ ϕ_{xy}) and equating (33) and (35) gives

$$A_0 + \Psi_{xy}A_{xy} = S_0 + \Psi_{xy}S_{xy} \tag{37}$$

However, because they are obtained for different regions, integrals from the A and S series are expressed in terms of different variables: a and c, respectively. In this case, as we have seen, $A_{xy} = 0$ over its whole range, and it is therefore possible to obtain a simple polynomial relation between variables a and c. Thus, eq 37 becomes

$$A_0 = S_0 + \Psi_{xy}S_{xy} \tag{38}$$

From eq 34, with $A_{xy} = 0$, the number of actives recovered using a single measure is $n(a) = NA_0\phi_0$. But, under the constraint of equivalent rank, we can substitute for A_0 in this expression using eq 38 to obtain an expression for the number of recovered-actives using a single measure in terms of the variable c rather than a:

$$n(c) = N\phi_0(S_0 + \Psi_{xy}S_{xy}) \tag{39}$$

This may now be compared directly with the number of actives recovered using the integration region appropriate to the SUM-rule, eq 36—already expressed in terms of c. The algebraically most convenient way to do this is by difference, when the uniform contributions cancel to give

$$\Delta n = n_S(c) - n(c) = N(\phi_{xy} - \phi_0\Psi_{xy})S_{xy} \tag{40}$$

or in terms of the original coefficients:

$$\Delta n = N(\phi_{xy}\Phi_0 - \Phi_{xy}\phi_0)S_{xy} \tag{41}$$

We have thus derived a result for the change in the number of actives retrieved when SUM-fusion is employed rather than a single measure, which is expressed in terms of differences in the degree of correlation between the recovered active similarities and the recovered-nonactive similarities. If the degree of correlation is expressed relative to the appropriate uniform densities, $\phi^0_{xy} = \phi_{xy}/\phi_0$ and $\Phi^0_{xy} = \Phi_{xy}/\Phi_0$, and we substitute for the fraction of active compounds, λ in eq 28, we obtain

$$\Delta n^* = (1 - \lambda)[\phi^0_{xy} - \Phi^0_{xy}]S_{xy} \tag{42}$$

This expression shows how the recall enhancement, $\Delta n^* = \Delta n/n$, depends on the active fraction, λ, the difference in degree of correlation for the actives ϕ^0_{xy} and the nonactives Φ^0_{xy}, and the integral S_{xy}, a geometric factor arising from SUM-rule characteristics. It is an analytical result obtained for our simple model, which shows that correlation between the positions of recovered compounds using different measures can lead to enhanced retrieval performance at high rank. It is however of wider relevance since the degree of correlation defined here is directly related to the Product-Moment Correlation Coefficient (also known as Pearson's Correlation Coefficient), which can be extracted for any bivariate distribution. The recall enhancement is shown in Figure 7 for $\phi_{xy} = 0.2$ and an uncorrelated recovered nonactive distribution, $\Phi_{xy} = 0$, and shows that positive enhancement at high rank is balanced by negative enhancement at low rank. The sign change seen in this result is a direct consequence of the integral S_{xy}. Results for the linearly biased component, which will be discussed later, are also shown in this figure for comparison.

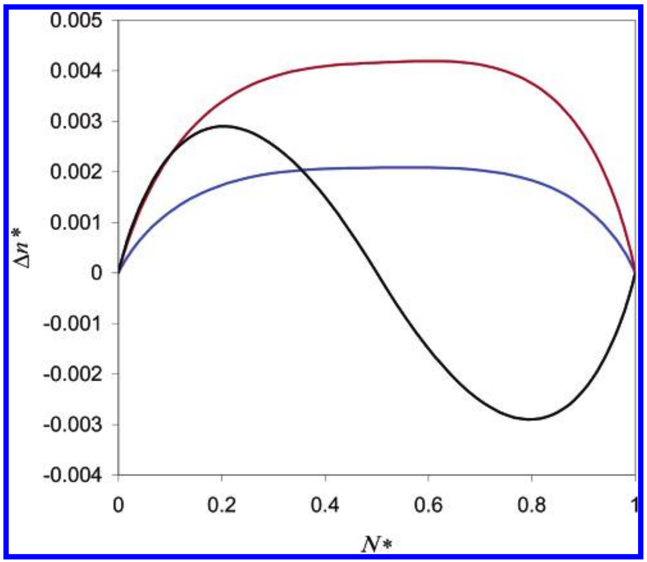

Figure 7. SUM-rule recall enhancement Δn^* plotted against the normalized rank N^*, with $\Phi_0 = \phi_0 = 0.5$, $\Phi_{xy} = \Phi_x = \Phi_y = 0.0$; (blue line) $\phi_x = \phi_y = 0.05$, $\phi_{xy} = 0.0$; (red line) $\phi_x = \phi_y = 0.1$, $\phi_{xy} = 0.0$; (black line) $\phi_x = \phi_y = 0.0$, $\phi_{xy} = 0.2$.

Notably, when the same degree of correlation is present in both recovered actives and nonactives it can be seen that Δn must vanish over the whole range. Moreover, if the nonactive values are correlated more than the active values, then the sign of the enhancement changes over the whole range. When $\Phi_{xy} = 0$, i.e., the nonactives are uncorrelated (as in Figure 7), the enhancement is governed by a simple linear dependence on the degree of correlation of the active similarities obtained by the two measures that are being fused.

To summarize, the above analysis shows that it is possible for data fusion to enhance the performance of a similarity search at high rank by combining the results of two retrieval methods. In this case, the driving force is correlation: enhancement is improved if the order in which the active compounds are recovered by the two measures is more related than it is for nonactive compounds.

An alternative way to express the enhancement is through the relative improvement $\epsilon = \Delta n/n(c)$ shown in Figure 8 (this also includes results discussed in the next section). Using eqs 39 and 42, and with $\phi_0 + \Phi_0 = 1$, we obtain

$$\epsilon = \frac{(1 - \lambda)[\phi^0_{xy} - \Phi^0_{xy}]S_{xy}}{S_0 + [\lambda\phi^0_{xy} + (1 - \lambda)\Phi^0_{xy}]S_{xy}} \tag{43}$$

The maximum values of 4−5% in relative enhancement are comparable with those seen in real experiments.[8] If the degree of correlation for the actives is less than that for the nonactives, then the sign of the enhancement is clearly reversed and positive results are obtained at low rank, $N^* > 0.5$. However, we can see from Figure 8 that this will lead to much smaller values of positive relative enhancement.

For low active fractions, $\lambda \rightarrow 0$, which is the norm in virtual screening applications, the last expression hence becomes

$$\epsilon = \frac{[\phi^0_{xy} - \Phi^0_{xy}]S_{xy}}{S_0 + \Phi^0_{xy}S_{xy}} \tag{44}$$

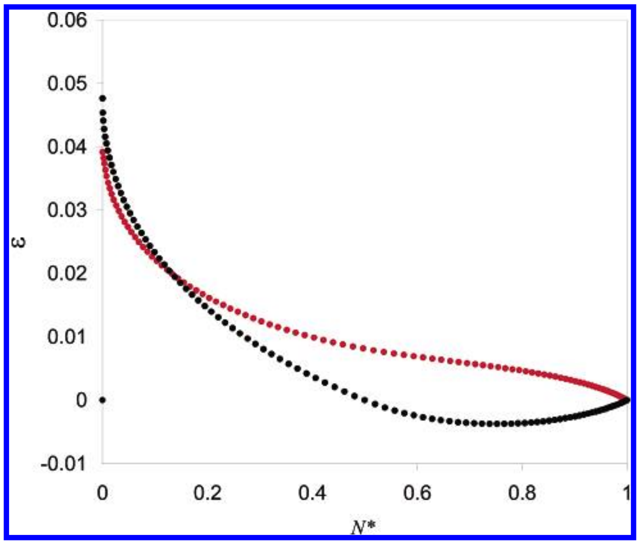

Figure 8. Relative fusion enhancement, ϵ, plotted against the normalized rank N^* using the SUM-rule with $\Phi_0 = \phi_0 = 0.5$, $\Phi_{xy} = \Phi_x = \Phi_y = 0.0$: (black circle) $\phi_x = \phi_y = 0.0$, $\phi_{xy} = 0.2$; (red circle) $\phi_x = \phi_y = 0.1$, $\phi_{xy} = 0.0$.

Notice that higher values of the recovered nonactive correlation $\Phi_{xy}{}^0$ reduce the size of the relative improvement if the difference between recovered active and recovered nonactive correlations in the numerator is held constant. For uncorrelated recovered nonactives values, $\Phi_{xy} = 0$, eq 44 further simplifies to show that the relative improvement is, in this case, linearly related to S_{xy}/S_0 and the relative degree of correlation of the active similarities.

LINEARLY BIASED CONTRIBUTION

In this section we consider the linearly biased contributions in isolation and thus choose $\phi_{xy} = \Phi_{xy} = 0$. In place of eqs 33−36 we now obtain

$$N(a) = N[A_0 + (\Phi_x + \phi_x)A_x + (\Phi_y + \phi_y)A_y] \quad (45)$$

$$n(a) = N[\phi_0 A_0 + \phi_x A_x + \phi_y A_y] \quad (46)$$

$$N_S(c) = N[S_0 + (\Phi_x + \phi_x)S_x + (\Phi_y + \phi_y)S_y] \quad (47)$$

$$n_S(c) = N[\phi_0 S_0 + \phi_x S_x + \phi_y S_y] \quad (48)$$

It is then convenient to define

$$\Psi_x = \Phi_x + \phi_x; \quad \Psi_y = \Phi_y + \phi_y \quad (49)$$

so that the condition of equivalent rank $N = N_S$ can be written concisely as

$$A_0 + \Psi_x A_x + \Psi_y A_y = S_0 + \Psi_x S_x + \Psi_y S_y \quad (50)$$

As before, we use this relation to express the variable a in terms of c in order to make a comparison of the actives retrieved: $n(a)$ and $n_S(c)$. However, since $A_y = 0$ but $A_x \neq 0$, only partial simplification of the expression is possible. In this case, substituting for the A's from Table 1 and collecting terms leads to a quadratic equation in a

$$-\frac{1}{2}\Psi_x a^2 + \left[1 + \frac{1}{2}\Psi_x\right]a = S_0 + \Psi_x S_x + \Psi_y S_y \quad (51)$$

where the terms on the right depend on c only. In principle the solution gives values of a in terms of c, at which an equivalent rank is reached by single measure recall. However, in this case the analytic expressions obtained are unpromising, and we have resorted to numerical solutions. Using the integrals given in Table 1, values of S_0, S_x, and S_y and thus n_S and N_S were computed at 100 equally spaced values of c over the range 0−2. Equation 51 was then solved at each value of c to obtain the equivalent values of a. This value of a was then substituted back into eq 46 to obtain the value of $n(a)$ for comparison with the equivalent value of $n_S(c)$ at the appropriate value of c. The process was repeated for various values of the coefficients in eqs 23 and 24. The degree of enhancement, $\Delta n^* = [n_S(c) - n(c)]/n(c)$, so obtained is plotted for a variety of values ϕ_x,ϕ_y in Figure 9, chosen specifically to investigate the effect of superposition of similarity densities by maintaining a constant overall value of $\phi_x + \phi_y$. In all of the cases shown, the recovered-nonactive values are assumed to be evenly distributed with $\Phi_x = \Phi_y = 0$.

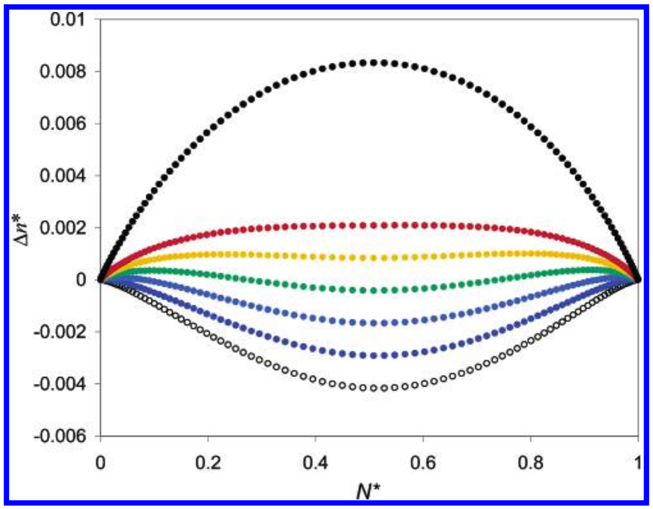

Figure 9. Fusion enhancement for the SUM-rule Δn^* plotted against the normalized rank N^*. From the bottom with $\Phi_0 = \phi_0 = 0.5$, $\Phi_{xy} = \Phi_x = \Phi_y = 0.0$, $\phi_{xy} = 0.0$: (white circle) $\phi_x = 0.1$, $\phi_y = 0.0$; (light blue circle) $\phi_x = 0.09$, $\phi_y = 0.01$; (dark blue circle) $\phi_x = 0.08$, $\phi_y = 0.02$; (green circle) $\phi_x = 0.07$, $\phi_y = 0.03$; (yellow circle) $\phi_x = 0.06$, $\phi_y = 0.04$; (red circle) $\phi_x = 0.05$, $\phi_y = 0.05$; (black circle) $\phi_x = 0.0$, $\phi_y = 0.1$.

With $\phi_x = 0.1$, $\phi_y = 0.0$ there is strong negative enhancement $n_S < n$; since the density for these parameters has the same symmetry as the region for single-measure recall, this is the most effective method and fusion cannot improve on the result. As the proportion of ϕ_y is increased the density more closely matches the SUM-rule region, and for $\phi_x = \phi_y = 0.05$ there is significant positive enhancement using the SUM-rule. The upper curve for $\phi_x = 0.0$, $\phi_y = 0.1$ is included for completeness, but this only shows that the SUM-rule is an improvement over performing single-measure recall along the x-axis (i.e. using Figure 3(a)), which in this case has a symmetry orthogonal to the density. By symmetry, integration along the y-axis would give the best result in this case, and the SUM-rule would give a negative enhancement in comparison.

For the linearly biased terms the best enhancement is thus obtained when, for the recovered-active distribution $\phi_x = \phi_y > 0$, against an evenly distributed background of recovered-nonactive values. For similarity fusion, this scenario models a situation in which two measures offer equally

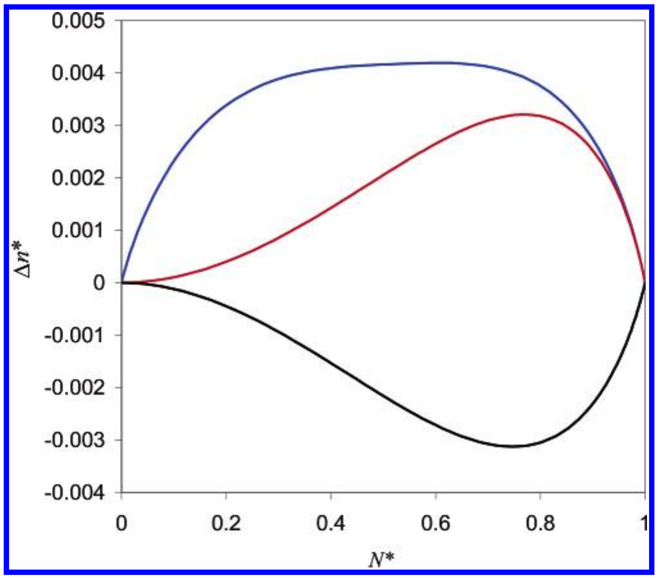

Figure 10. Fusion enhancement Δn^* plotted against the normalized rank N^* with $\Phi_0 = \phi_0 = 0.5$, $\Phi_{xy} = \Phi_x = \Phi_y = 0.0$: SUM-rule: (blue line) $\phi_x = \phi_y = 0.1$, $\phi_{xy} = 0.0$; MAX-rule: (red line) $\phi_x = \phi_y = 0.1$, $\phi_{xy} = 0.0$; MAX-rule: (black line) $\phi_x = \phi_y = 0.0$, $\phi_{xy} = 0.2$

good retrieval of actives relative to nonactive structures. However, they also recover compounds in a significantly different order since the correlation term is zero. To summarize, in this case, the driving force for improved retrieval performance using data fusion is bias: the enhancement is positive if both of the independent retrieval methods are equally good at retrieving more active compounds than nonactive compounds at high rank.

Fusion Using the MAX-Rule Region. The region appropriate to the MAX-rule is shown in Figure 5 where we defined the parameter *m*, describing the thickness of the "arms". Expressions for the relevant integrals in terms of this parameter can be found in Table 1, and, following a parallel development to that given above for the SUM-rule, we can thus obtain the analogue of eq 41 for the MAX-rule

$$\Delta n = n_M(m) - n(m) = N(\phi_{xy}\Phi_0 - \Phi_{xy}\phi_0)M_{xy} \quad (52)$$

where in this case the integral M_{xy} is a geometric factor arising from MAX-rule characteristics. For the linearly biased terms we again obtain a quadratic, which can be solved numerically as before. The resulting MAX-rule fusion enhancement is compared with the SUM-rule enhancement for linear bias in Figure 10.

The enhancement for the correlated term using the MAX-rule is strongly negative, but, from eq 52, negatively correlated recovered-active values relative to recovered-nonactive values lead to an equivalent positive result. The combined linearly biased terms with $\phi_x = \phi_y$ again lead to a positive result. This shows that the SUM-rule is superior over the whole of the range for equivalent values of ϕ_x and ϕ_y. The enhancement obtained for nonequivalent bias is plotted in Figure 11, with the same values of ϕ_x and ϕ_y as used for the SUM-rule results of Figure 9.

With $\phi_x = 0.1$, $\phi_y = 0.0$ there is strong negative enhancement; since the density for these parameters has the same symmetry as the region for single measure recall, this is the most efficient method. As the proportion of ϕ_y is increased the density overlaps more strongly with the MAX-rule region, and for $\phi_x = \phi_y = 0.05$ the result is just positive over the whole range. The SUM-rule results in Figure 9 are not entirely symmetric, and stronger asymmetry does appear for higher values of ϕ_x and ϕ_y but it is already clearly evident in the MAX-rule results of Figure 11. Notably, if there is any imbalance in the densities the enhancement rapidly becomes negative for low values of N^*, equivalent to high rank. In Figure 12 we compare the relative enhancement for the SUM and MAX rules for equally biased distributions: the SUM-rule performance is clearly much superior at low rank for linearly biased distributions.

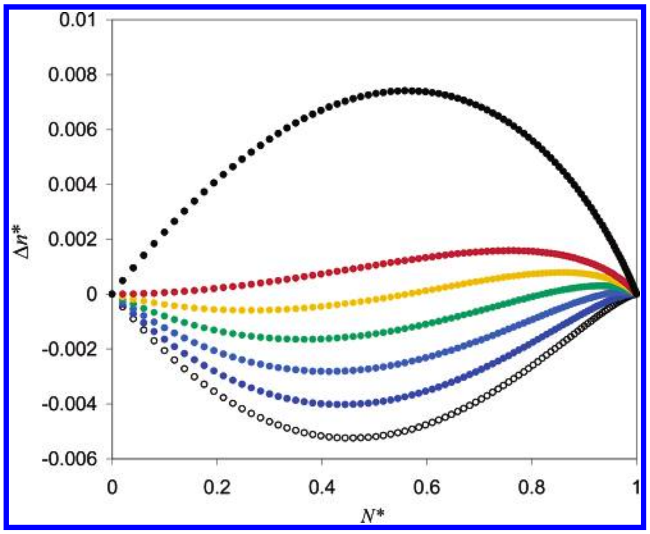

Figure 11. Fusion enhancement $\Delta n^* = n_m^* - n^*$ for the MAX-rule plotted against the normalized rank N^*. From the bottom with $\Phi_0 = \phi_0 = 0.5$, $\Phi_{xy} = \Phi_x = \Phi_y = 0.0$, $\phi_{xy} = 0.0$: (white circle) $\phi_x = 0.1$, $\phi_y = 0.0$; (light blue circle) $\phi_x = 0.09$, $\phi_y = 0.01$; (dark blue circle) $\phi_x = 0.08$, $\phi_y = 0.02$; (green circle) $\phi_x = 0.07$, $\phi_y = 0.03$; (yellow circle) $\phi_x = 0.06$, $\phi_y = 0.04$; (red circle) $\phi_x = 0.05$, $\phi_y = 0.05$; (black circle) $\phi_x = 0.0$, $\phi_y = 0.1$.

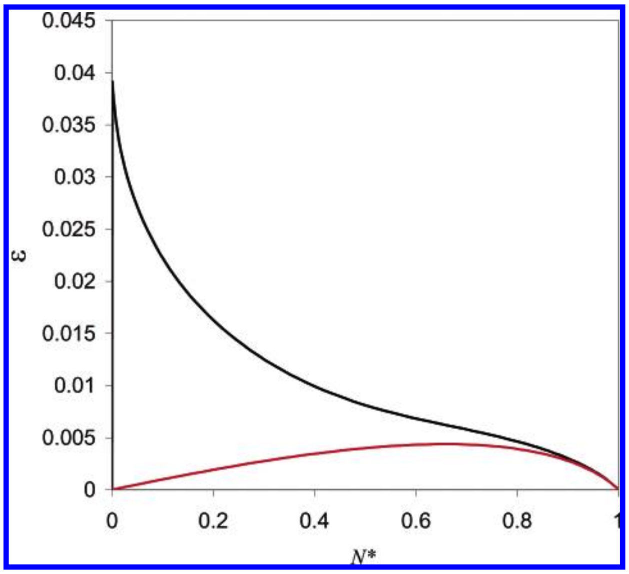

Figure 12. Relative fusion enhancement, ϵ, plotted against the normalized rank N^*. SUM-rule with $\Phi_0 = \phi_0 = 0.5$, $\Phi_{xy} = \Phi_x = \Phi_y = 0.0$, $\phi_{xy} = 0.0$: (black line) $\phi_x = \phi_y = 0.1$; MAX-rule (red line) $\phi_x = \phi_y = 0.1$.

A SIMULATION APPROACH

Simulating Data Fusion. The previous section has demonstrated the explanatory power of our theoretical approach, but the model is limited in that it can be applied only to simple distributions and fusion rules that describe the combination of two data sets. However, for more complex forms, simulation is a very useful tool with a more general range of application, and one that can be extended to the combination of several sets. The simulation of data fusion is straightforward and comprises three elements: the generation of suitable data with known distributions; a fusion algorithm; and a means of analyzing the results. We have used random variables x_i and y_i chosen on the range 0−1 to represent, for example, the similarity values obtained by two experiments. Random variables can be chosen from any given distribution using a rejection method[32] that we have adapted for bivariate distributions as follows. A pair of random numbers, x_i and y_i, is first chosen on the range 0−1 from a flat distribution. A third random number, s ($0 < s < 1$), is then chosen and compared with the desired distribution $f(x,y)$. If $s \leq f(x_i,y_i)$, then the values x_i and y_i are accepted, but if $s > f(x_i,y_i)$, then they are discarded and the process is repeated. For our simulations, $f(x,y)$ is just a scaled version of the desired probability density $\phi(x,y)$ with a peak value of 1 to minimize the number of rejections needed. Simulations have been performed in which 100 000 values are taken in total with some proportion, λ, of these randomly assigned as active, the rest being labeled nonactive. In general the distribution parameters for the values assigned as active and nonactive are different.

Data fusion is performed using an algorithm that is essentially identical to that used in practice. The SUM and MAX rules are implemented by calculating values of a new quantity, z_i, from

$$z_i = \frac{1}{2}(x_i + y_i);\ z_i = \max\,[x_i,y_i] \tag{53}$$

respectively. Values x_i, y_i, and z_i are then independently ranked with the highest values top, and the number of actives—i.e., those values with index i that have been labeled as such—are counted in each case up to a given rank. Knowing the total number of pairs flagged as “active” these results are then converted to recall values. Comparison of the recall values obtained using fused results (z_i) and single measures (x_i, y_i) then leads directly to values of the enhancement factor ϵ. A simulation study based on these ideas yielded results in very close agreement with those predicted by the theory above, thus supporting the model that we have presented.

Inclusion of Unmatched Values. The mathematical model considers a pair of lists in which each numerical value is matched to a value in the other list. These matched values represent similarities that would be assigned to the same chemical structure in a real data fusion trial. This will only be the case if all of the available similarity data is fused. However, if a large database is being searched, lists are normally truncated to include only the most similar compounds to a query; for example, an operational virtual screening system might consider just those molecules in the top 1% of the ranking resulting from a similarity search. Under these circumstances some of the compounds in one list will not appear in the second list and vice versa: we call these *unmatched* values.

These unmatched similarity distributions are rather awkward to represent mathematically but can readily be accommodated in a simulation. Suppose all the compounds in a database are ranked by similarity using methods X and Y and suppose that each list is now truncated to the first 1000 compounds and the similarities scaled to the range 0−1. Compounds with similarities below the truncation point are given a scaled similarity of zero. Some of the first 1000 compounds in list X will also appear in the first 1000 compounds of list Y, but some will appear below the truncation point in list Y and be assigned a value zero. If the measures X and Y are uncorrelated there is no way of predicting which of the chosen compounds in X will be above or below the truncation line in Y and vice versa. Hence, to simulate this scenario, proportions of the recovered-active and recovered-nonactive values are randomly set to zero in one list or the other when the values are generated. When generated in this way, the distribution of nonzero unmatched values in each list is determined by the same distribution parameters (excepting correlation) entered for the matched values. Thus if the x_i values are biased toward high similarity by a positive value of ϕ_x, then the unmatched values in this list will retain this bias.

An implementation of these ideas shows that unmatched values can make a significant contribution to data fusion but that it is also rank dependent. At high rank, for the SUM-rule, our results support the suggestion (by Lee in the context of textual information retrieval[20]) that higher recovered-active overlap than recovered-nonactive overlap is a factor that leads to positive fusion enhancement. Beitzel et al. have argued that the combination of lists containing a significant number of high-ranking unmatched documents is likely to lead to fusion enhancement;[25] however, we have shown previously that completely matched values can also lead to positive fusion, and we thus cannot agree with these authors’ stronger statement that unmatched values are a necessary condition for effective retrieval.[33]

DISCUSSION

In this paper, we have developed a theoretical approach to the problem of combining continuous-valued similarity data from several different measures. This shows that the origin of data fusion enhancement for simple fusion rules can be traced to a combination of differences between the recovered-active and recovered-nonactive multivariate distributions and the geometrical difference between the regions of the multivariate distributions that the fusion rules access. This is succinctly summarized by eq 16 and its extension to higher dimensions, which effectively links similarity distributions and fusion rules directly to the precision. The approach has been applied to bivariate distributions of matched values through a mathematical model using idealized similarity distributions aimed at capturing the most basic functional variations. These have focused on three contributions: the effects of differential correlation and bias and commonality between recovered-active (relevant, true-positive) and recovered-nonactive (nonrelevant, false-positive) values as obtained by two measurement schemes.

The three different contributions that we have studied have been shown to lead to very different behavior when operated on by the SUM-fusion or MAX-fusion rule. We have shown that if the values representing two different similarity measures between active compounds are correlated with more than those of the recovered-nonactive values, then data fusion can lead to improved recall at high rank (relative to list length, i.e., low N^*) using the SUM-rule. However, this improvement becomes negative as lower ranks are accessed. The same type of data leads to reduced values of the recall over the whole range when fusion is performed using the MAX-rule. Reduced correlation of the recovered-active similarities relative to the recovered-nonactive values (anticorrelation of the recovered-active similarities if the recovered-nonactive values are uncorrelated) reverses these results so that the MAX-rule becomes advantageous in these cases.

Our analysis has made clear the overall complexity of the fusion process. Thus, the combination of just two lists of similarity values actually depends on eight distinct distributions: the recovered-active and recovered-nonactive distributions of both lists for both matched and unmatched compounds. Both the recovered-active and recovered-nonactive similarity values are treated in the same way by a data fusion algorithm, and because of the complex interactions between the aspects that we have discussed (differential correlation, bias and commonality, and rank above which molecules are retrieved) the outcomes of data fusion are not open to intuitive prediction. It is hence not surprising that the behavior of data fusion is frequently found to be inconsistent for different data sets.

For matched values, our work would suggest that positive fusion results for two measures should be obtained, at least for some values of rank, if the bivariate distribution of recovered-active values is (a) different from that of the recovered-nonactive distribution and (b) matches the appropriate integration region for the fusion rule under consideration more closely than either of the single measure regions. Both points can be conflated and restated concisely by saying that the precision, P(Q) (eq 16), for the integration region Q associated with the fusion rule under consideration should be higher than that of either of the single measure regions at the same rank. For this to happen at all using the SUM and MAX fusion rules it is clear that a broad spread of values throughout the integration region is preferable to tightly correlated values, which tend to cluster along the $y = x$ diagonal of the region. Thus we would agree that Kendall's tau for rank-based fusion[22] (or equivalently, Pearson's Correlation Coefficient for the fusion of similarity values) can be useful to determine whether the outputs of two measurement schemes are sufficiently different to enable the possibility of effective data fusion.

There are limitations in the model that we have presented. First, it is not clear how the integration regions described in this work might be extended to some of the more sophisticated fusion rules that are available,[24] and for these, simulation is likely to provide a better route to improved understanding. Second, we have not investigated to what extent our models are relevant to rank-based fusion, although there seems no reason such data cannot be comparably represented as distributions. Third, it is complex to include the important effects of unmatched contributions, which can be better analyzed using a simulation approach. Even so, the theory does provide a firm basis for rationalizing the results of data fusion when applied to similarity searching; in the companion paper[15] we examine some of our experimental data fusion results in light of the model presented here.

ACKNOWLEDGMENT

We thank the following: Pfizer Global Research and Development for funding; Alexander Alex for helpful discussions and guidance; the referees for their comments on this manuscript; Digital Chemistry Ltd., MDL Information Systems Inc., and Tripos Inc. for software and data; and the Royal Society and the Wolfson Foundation for hardware and laboratory support.

REFERENCES AND NOTES

(1) Klien, L. A. *Sensor and Data Fusion Concepts and Applications*, 2nd ed.; SPIE: Bellingham, WA, 1999.

(2) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Moseley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

(3) 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.

(4) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: a Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(5) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of Molecular Similarity Measures Using Data Fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1−16.

(6) Wang, R.; Wang, S. How Does Consensus Scoring Work For Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *43*, 449−457.

(7) Salim, N.; Holliday, J.; Willett, P. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−440.

(8) Whittle, M.; Willett, P.; Klaffke, W.; van Noort, P. Evaluation of Similarity Measures for Searching the *Dictionary of Natural Products* Database. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 449−457.

(9) Raymond, J. W.; Jalaie, M.; Bradley, M. P. Conditional Probability: A New Fusion Method for Merging Disparate Virtual Screening Results. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 601−609.

(10) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein−Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.

(11) Whittle, M.; Gillet, V. J.; Willet, P.; Loesel, J.; Alexander, A. Enhancing the Effectiveness of Virtual Screening by Fusing Nearest-Neighbour Lists: A Comparison of Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840−1848.

(12) 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.

(13) Barber, 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.

(14) Oda, A.; Tsuchida, K.; Takakura, T.; Yamaotsu, N.; Hirono, S. Comparison of Consensus Scoring Strategies for Evaluating Computational Models of Protein-Ligand Complexes. *J. Chem. Inf. Model*. **2006**, *46*, 380−391.

(15) Whittle, M.; Gillet, V. J.; Willet, P.; Loesel, J. Analysis of Data Fusion Methods in Virtual Screening: Applications to Similarity and Group Fusion. *J. Chem. Inf. Model.* **2006**, *46*, 2206−2219.

(16) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(17) Johnson, M.; Maggoria, G. M. *Concepts and Applications of Molecular Similarity;* Wiley: New York, 1990.

(18) 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.

(19) Belkin, N. J.; Cool, C.; Croft, W. B.; Callan, R. K. Effect of Multiple Query Representations on Information System Performance. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993; pp 339−346.

(20) Lee, J. H. Analyses of Multiple Evidence Combination. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1997; pp 267−276.

(21) Turner, K.; Ghosh, J. Linear and Order Statistics Combiners for Pattern Classification. In *Combining Artificial Neural Nets*; Sharkey, A. J. C., Ed.; Springer-Verlag: London, 1999; pp 127−161.

(22) Ng, K. B.; Kantor, P. B. Predicting the Effectiveness of Naïve Data Fusion on the Basis of System Characteristics. *J. Am. Soc. Inf. Sci.* **2000**, *51*, 1177−1189.

(23) Rao, N. S. V. On Fusers that Perform Better than Best Sensor *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 904−909.

(24) Wu, S.; Crestani, F.; Gibb, F. New Methods of Results Merging for Distributed Information Retrieval. *Lect. Notes Comput. Sci.* **2003**, *2924*, 84−100.

(25) Beitzel, S. M.; Jensen, E. C.; Chowdhury, A.; Grossman, D.; Goharian, N.; Frieder, O. Fusion of Effective Retrieval Strategies in the Same Information Retrieval System. *J. Am. Soc. Inf. Sci. Tech*. **2004**, *55*, 859−868.

(26) Hsu, D. F.; Taksa, I. Comparing Rank and Score Combination Methods for Data Fusion in Information Retrieval. *Inf. Retriev.* **2005**, *8*, 449−480.

(27) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures. *J. Mol. Graphics Modell.* **2000**, *18*, 343−357.

(28) Bendat, J. S.; Piersol, A. G. *Random Data: Analysis and Measurement Procedures,* 2nd ed.; Wiley-Interscience: New York, 1986.

(29) The BCI software is available from Digital Chemistry Ltd. at URL http://www.digitalchemistry.co.uk.

(30) The *MDL Drug Data Report* database is available from MDL Information Systems at URL http://www.mdl.com/.

(31) Cheng, C.; Maggiora, G.; Lajiness, M.; Johnson, M. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 909−915.

(32) Press: W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipies in C*, 2nd ed.; Cambridge University Press: Cambridge, 1994.

(33) Beitzel, S. M.; Jensen, E. C.; Chowdhury, A.; Frieder, O.; Grossman, D.; Goharian, N. Disproving the Fusion Hypothesis: An Analysis of Data Fusion via Effective Information Retrieval Strategies. *Proceedings of the 2003 ACM Symposium on Applied Computing (SAC),* 2003; pp 823−827.

CI049615W