Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion

Claire M. R. Ginn, David B. Turner, and Peter Willett*

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK

Allan M. Ferguson and Trevor W. Heritage

Tripos Inc., 1699 South Hanley Road, St. Louis, Missouri 63144

Received July 8, 1996[⊗]

EVA is a new molecular descriptor that provides a concise summary of the fundamental frequency components of a molecule's infrared range vibrational spectrum in a vector format. Target structures from the Starlist database are used to demonstrate the effectiveness of the descriptor for similarity searching and its difference from a conventional similarity measure based on the matching of two-dimensional (2D) fingerprints. The use of data fusion on the rankings resulting from the EVA-based and the 2D-based similarity measures results in a combined ranking that can be more effective in simulated property prediction experiments than either of the individual rankings.

INTRODUCTION

The *similar property principle*[1,2] states that molecules that are structurally similar to each other are expected to exhibit similar properties (or activities). Thus, if an active molecule, such as an initial weak lead, is used as the target structure in a similarity search of a chemical database,[3,4] then the nearest neighbors are expected to have a greater probability of exhibiting the same activity than would be the case if they were selected at random. Experiments with a wide range of types of descriptor have demonstrated the general effectiveness of this approach to database searching.[5−9]

Most current systems for similarity searching are based on two-dimensional (2D) fragment substructures, with two molecules being regarded as similar if they have many fragments in common. These fragments can either be descriptors that have been developed specifically for activity correlation purposes, such as the atom pair[10] or the topological torsion,[11] or those that are encoded in a bit-string for use in the screening stage of a 2D substructure search.[12] There is currently much interest in the development of similarity measures that take account of the three-dimensional (3D) structures of molecules.[2] A range of types of information has been used for the calculation of such measures, including interatomic distances,[8,13−16] valence and torsion angles,[17,18] calculated physicochemical properties,[7,19] and molecular fields.[14,20−22]

The availability of different similarity measures has led to several comparative studies that seek to identify the best measure(s) for similarity searching, *i.e.*, those that are most successful in retrieving additional actives given an active target structure. Examples of such comparative studies have been reported by Kearsley *et al.*,[7] Sheridan *et al.*,[8] Brown and Martin,[9] Pepperrell and Willett,[13] and Johnson *et al.*,[23] *inter alia*. While such comparisons can identify measures that are of general applicability, quantitative structure−activity relationship (QSAR) studies have demonstrated that many different types of molecular feature may be of importance for a specific class of activity. For example, the classical Hansch approach to QSAR involves multivariate regression on parameters such as the partition coefficient and the molar refractivity,[24] while the more recent 3D QSAR approaches involve the analysis of steric, electrostatic, and hydrophobic fields.[25] A plethora of descriptors can be calculated using molecular orbital techniques, and these have also been found to provide effective parameters in QSAR studies (see, *e.g.*, ref 26). Thus, while a particular similarity measure may well provide an appropriate ranking of a database for a particular type of feature that is important for activity, the ranking will take no account of features that are described by alternative similarity measures. Accordingly, it should be possible to increase the performance of a similarity search by basing the ranking on several different types of similarity measure, rather than on just a single one as in current approaches.

In this paper, we report the use of a new descriptor for similarity searching, called EVA, in files of 3D chemical structures. The next section introduces the EVA descriptor, and this is followed by a comparison of the effectiveness of similarity searches using EVA with analogous searches using a conventional, 2D fingerprint similarity measure. We then introduce the concept of *data fusion*,[27] which allows different similarity measures to be combined, and describe simulated property prediction experiments that explore the effectiveness of this approach to the combination of measures of structural similarity. The paper concludes with a summary of our principal findings.

THE EVA DESCRIPTOR

Many popular methods for 3D QSAR seek to correlate biological activity with the values of various types of molecular field at points in a 3D grid surrounding each of the molecules in a dataset.[25] These methods require that the molecules be aligned, so that equivalent structural features can be matched. Molecular alignment can be relatively

[*] To whom all correspondence should be addressed. E-mail P.WILLETT@SHEFFIELD.AC.UK.

[⊗] Abstract published in *Advance ACS Abstracts,* December 15, 1996.

S0095-2338(96)00466-0 CCC: $14.00

straightforward if a set of rigid structural analogues is to be analyzed but becomes increasingly problematic as the diversity of the dataset increases, even if no account is taken of conformational flexibility. Recent work at Shell Research Limited has resulted in the development of an alternative molecular descriptor for QSAR, called EVA (for EigenVAlue), that is based on calculated spectral properties of molecules and that does not require the alignment of a dataset to generate a statistical model.[28-32]

The main premise of the work at Shell is that a molecular structure may be uniquely defined by its fundamental vibrational fingerprint (as reflected in its infrared spectrum), which is readily calculated from the molecular electronic potential energy function. Typically, an approximate 3D structure for a molecule is created using a structure generation program; the geometry is then optimized using a molecular mechanics or semiempirical/*ab initio* quantum mechanics method. The fundamental molecular vibrational frequencies and directions (corresponding to the normal coordinate eigenvalues and eigenvectors) are then calculated by means of a classical normal coordinate analysis. The theoretical descriptor (EVA) is derived from the vibrational frequencies alone; vibrational intensity and atomic displacement information are not included in the descriptor.

The number of fundamental vibrations for a structure of N atoms is equal to the number of vibrational degrees of freedom, which is $3N-6$ for a nonlinear molecule and $3N-5$ for a linear molecule. Therefore, one of the major difficulties of utilizing the calculated vibrational frequencies in QSAR modeling or similarity calculations is that, unless the structures all contain the same number of atoms, different molecules in a dataset will have different numbers of fundamental frequencies, *i.e.*, the data are usually in nonstandard form. As a result, the direct comparison of the frequency sets for each structure is not possible. Furthermore, even where the number of vibrations is the same for two molecules, it is still difficult to establish which vibrations should be compared between the two structures, owing to the inherent and effectively indeterminate contributions from individual atoms. The EVA approach overcomes these problems in a three-stage process. In the first step, the data are projected onto a bounded frequency scale (BFS) selected to cover the full range of vibrational frequencies that may be obtained—in this case, infrared wave numbers in the range 0−4000 cm^{-1}. In the second step, a Gaussian kernel of fixed standard deviation (σ) is placed over each and every vibrational frequency value for a given structure. The result of this procedure is a series of $3N-6$ (or $3N-5$) overlapping kernels for the structure. The value for the descriptor, EVA_x, at point x on the BFS can then be determined by summing the contributions from each and every one of the $3N-6$ ($3N-5$) overlaid Gaussian curves:

$$\text{EVA}_x = \sum_{i=1}^{3N-6} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-f_i)^2/2\sigma^2}$$

where f_i is the ith vibrational frequency. In practice an heuristic, based on the chosen σ term, can be used to exclude from consideration the vast majority of kernels that make a negligible contribution to a given EVA variable. In the final step the BFS is sampled at fixed increments of L cm^{-1}, which results in the $4000/L$ values that comprise an EVA descriptor. This procedure is repeated for each and every member of a dataset. In QSAR studies a typical effective starting point for an EVA analysis has been found to be a σ of 10 cm^{-1} and an L of 5 cm^{-1}.[31] This results in a descriptor vector consisting of 800 variables, and, therefore, partial least squares (PLS) regression techniques are generally used to generate robust QSAR models.

The precise form of the descriptor depends on the σ term that is chosen; values of L need only be sufficiently small in relation to the σ term such that there is no information loss.[30] The purpose of the σ term is to “smear out” the frequency values so as to enable a frequency in one structure to be related to that in another structure (interstructural kernel overlap). Low values of σ will result in a very “spiky” EVA profile (Figure 1a) which has characteristics analogous to that of the 0’s and 1’s of a binary bit-string such as the UNITY 2D descriptor described below. The effect of gradually increasing the value of σ is to smooth out and merge these spikes (Figure 1b) which means that an EVA variable for a given structure may consist of contributions from more than one vibrational frequency (intrastructural kernel overlap). Previous studies with QSAR datasets using PLS regression[31] have indicated that there tends to be a smooth and gradual change in descriptor information content as σ is increased. Even at very large σ values (100 cm^{-1} or more) it was found that good QSAR models could be obtained, albeit generally at the cost of greater model complexity (*i.e.*, an increased number of PLS latent variables). However, PLS is a sophisticated, iterative regression procedure designed to weight variables with a view to maximizing the biological endpoint explained. In contrast similarity calculations used for searching purposes are usually done in the absence of a biological endpoint and typically are a crude, unweighted comparison of descriptor vectors. The effect of choice of σ on similarity searching performance is thus not clear, and we have, therefore, selected a range of σ terms for the experiments described below.

The use of a fixed Gaussian σ term means that each part of the spectrum is equally weighted prior to analysis. It is important to reiterate that the purpose of the EVA data standardization procedure is not to simulate an experimental IR-range vibrational spectrum (intensity information has been discarded) but, rather, to apply a probability density kernel to the calculated vibrational frequencies so as to derive a multivariate descriptor. In addition, the nature of the EVA smoothing technique is such that the descriptor can in certain circumstances be relatively insensitive to small conformational differences between a pair of structures under comparison.[32] For example, if the shift in frequency of a particular normal mode vibration of a pair of conformers is sufficiently small relative to the chosen Gaussian standard deviation then the derived EVA descriptor sets will also differ little, as judged by, for example, predictions made using a QSAR model. However, if σ is small relative to the frequency shift, then the descriptor sets will be different; *i.e.*, the descriptor exhibits increased conformational sensitivity in such circumstances. Therefore, through the appropriate choice of σ, the EVA descriptor can be considerably less sensitive to small changes or differences in molecular conformation than many 3D QSAR methods, such as CoMFA. The main advantage of EVA is, of course, that the technique is totally invariant to the relative orientation of structures, thus removing the computer-intensive align-

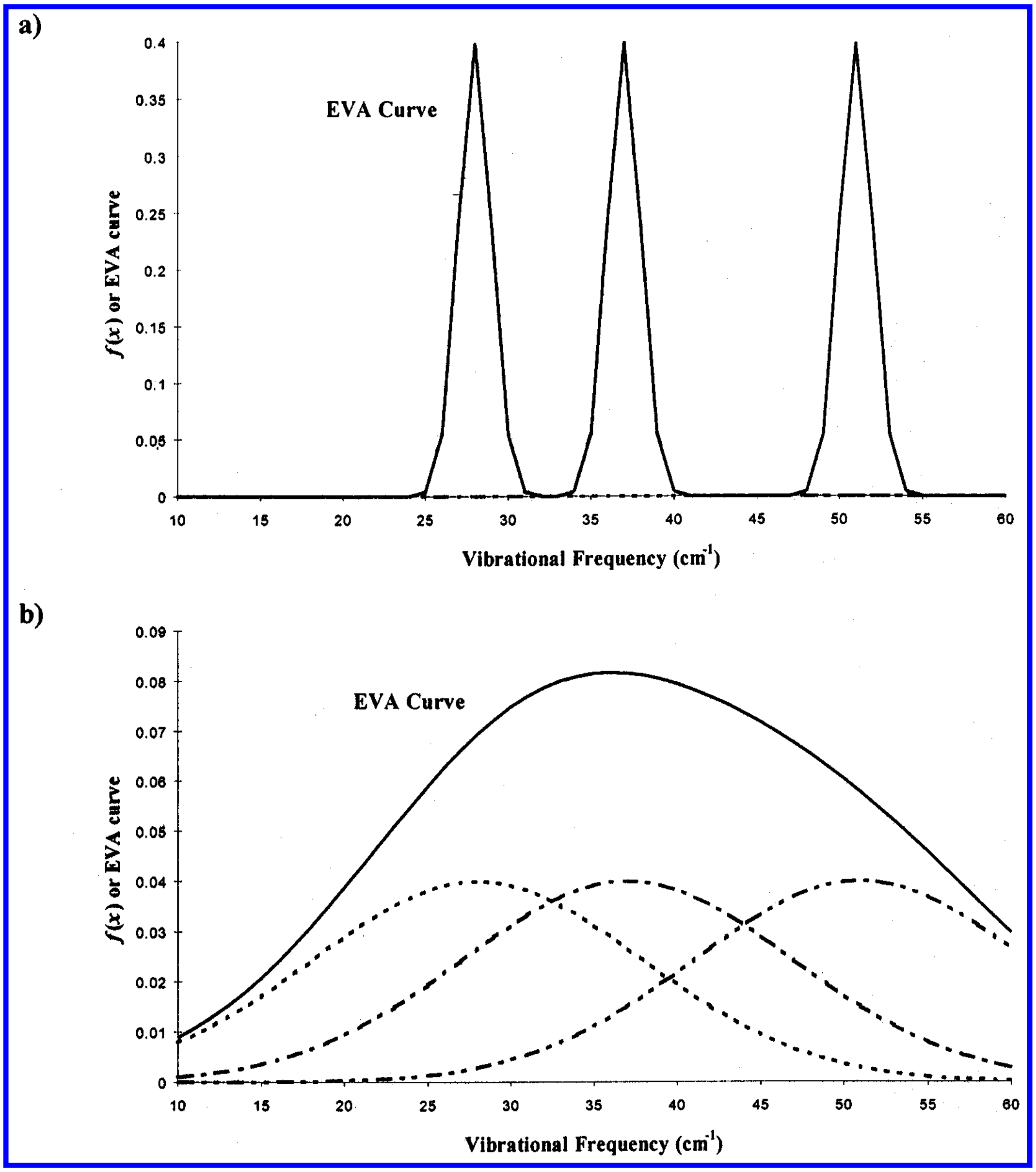


Figure 1. EVA descriptor generation based upon three hypothetical vibrational frequencies at 28, 37, and 51 cm^{-1} and using (a) a spiky Gaussian kernel ($\sigma = 1$ cm^{-1}) and (b) a more smoothed Gaussian kernel ($\sigma = 10$ cm^{-1}). The "EVA" curves are determined by summing the overlaid Gaussian kernels, and the frequency scale is then sampled at fixed intervals of L cm^{-1}. Note that in (a) the "EVA" curve obscures the underlying Gaussian kernels since there is no overlap of the kernels at non-negligible values.

ment procedures that are required for many of the other approaches that have been suggested for 3D similarity searching, such as molecular fields.[14,20-22]

EXPERIMENTAL DETAILS

The EVA approach was developed to provide a vibrational parameter for QSAR analyses of small datasets, either by means of discriminant models (active/inactive classification) or of robust regression analyses (typically using PLS with cross-validation).[28,29] In this paper, we consider its use for similarity searching with the similarity between a target molecule and a database molecule being calculated from a comparison of their EVA descriptors. The effectiveness of several EVA-based similarity measures is compared with that of a 2D, fragment-based similarity measure. Firstly, simulated property-prediction experiments are reported in which search effectiveness is evaluated through predictions of target log P values from the log P values of selected numbers of nearest neighbors. Secondly, comparisons are made between the nearest neighbor lists for the two types of similarity measure to find the nearest neighbors in common. Full details of these experiments are presented by Turner.[30]

The dataset used here was generated from the SMILES strings in the Starlist file,[33] which contains a large number of molecules for which high quality measured log P data are available. The experimental uncertainty of these measurements are quoted as average log P deviations of ±0.05 for solutes where $-3.0 < \log P < 6.0$ and ±0.10 for the remaining solutes. The SMILES strings were converted to 3D structures using CONCORD;[34] these structures were minimized using the MAXIMIN2 minimizer in SYBYL 6.0[35] with the default parameter settings, and the minimized structures then submitted to MOPAC 6.0[36] for full geometry optimization using the PM3 Hamiltonian. This step is necessary since the normal coordinate analysis (NCA) required to calculate the fundamental vibrational frequencies is only meaningful at stationary points on the energy surface. Once this optimization had been completed a MOPAC NCA is performed (using the FORCE keyword) so as to determine the molecular vibrational frequencies. The final dataset that was used in our experiments consisted of 8178 structures.

Studies of the use of EVA for QSAR have shown that the use of various Gaussian spread (σ) values in the generation of the descriptor can, in some instances, have a significant effect on the statistical quality of the derived QSAR models.[30] The same study also indicated that the choice of L is not important provided that L is not too large in relation to the chosen σ. Once these critical L values ($L^{\sigma}_{\mathrm{crit}}$) are exceeded then there is information loss (omission of peaks) with unpredictable effects on the information content of the EVA descriptor. However, in the interests of computational efficiency it is useful to choose as high an L as possible so as to minimize the size of the descriptor vector. Several runs were hence undertaken to evaluate the effect of choice of Gaussian spread on the search effectiveness of the EVA-based similarity measures. These tests were performed using the Dice and Cosine association coefficients and the Manhattan distance coefficient.[5] The Gaussian spread (σ) values chosen were 4, 10, 20, 40, and 50 cm^{-1}, with sampling intervals (L) of 2, 5, 10, 20, and 25 cm^{-1}, respectively.

For comparison, a UNITY (version 2.3)[35] database was created from the selected 8178 Starlist structures to enable searches to be carried out using a conventional, 2D fragment-based similarity measure. A 2D fingerprint was generated for each of the molecules using the default, general-purpose, UNITY screen definition file, in which various types of structural feature are mapped, generally as a pattern of a few bits, to a 988-member bit-string. Note that, unlike the EVA descriptor, hydrogen atoms are generally ignored when generating 2D bit-string descriptors. The similarity between the fingerprint describing a target structure and the fingerprint describing a database structure is calculated in the UNITY system by means of the Tanimoto association coefficient (which is monotonic with the Dice coefficient).[5]

COMPARISON OF EVA AND 2D SIMILARITY MEASURES

Simulated Property Prediction. The effectiveness of the EVA descriptor was evaluated using a "leave-one-out" approach based on the similar property principle.[1,5] A set of 100 target structures was chosen at random from our Starlist dataset, and the similarity of each of these to each of the database structures then determined using the Dice, Cosine, or Manhattan similarity coefficients in their forms generalized to deal with nonbinary data.[21] The structures were ranked in order of decreasing similarity (or increasing distance when using the Manhattan coefficient) to each of the target structures in turn. The log P value for each target was then estimated by the arithmetic mean of the log P values of the n nearest neighbors to the target. Log P estimates were made for all integral values of n in the range 1−100, resulting in 100 different log P estimates for each target and for each similarity measure. The estimated log P values were compared to the known target log P values, and the mean absolute (*i.e.*, unsigned) error of estimate of log P over the 100 target structures was then determined for each of the n nearest neighbors.

A matched pairs t test was used to test the significance of the differences between the mean error scores resulting from a set of EVA searches and from the corresponding set of 2D searches. Here

$$t = \frac{\bar{D}}{SE(\bar{D})}$$

where $\bar{D}$ is the mean of the differences between each pair of scores and $SE(\bar{D})$ is the standard error of the difference scores. Once t has been calculated, it is possible to obtain a value for the probability, p, that a difference between the mean errors as large as that observed could have arisen from two samples drawn from populations with the same mean value.

Figure 2 plots the mean unsigned error in the estimation of the 100 target log P values for various values of n using the 2D and a selected number of EVA similarity measures; plots for many other types of EVA measure are presented by Turner.[30] It is possible to compare the 2D and EVA results with three benchmark levels of performance. The first of these involves the calculation of the mean absolute error in the estimates of log P using the mean of all the structures in the Starlist as the best estimate of the log P value of each target. The value of this benchmark for the Starlist dataset is 1.23, and it will be seen from Figure 2 that all of the selected similarity measures perform better (provide closer target log P estimates) than this benchmark for all of the values of n considered here. None of the measures, however, approach the upperbound performance denoted by the "Best Possible" benchmark indicated in Figure 2. This upperbound is obtained by estimating the log P for each target structure from its n nearest neighbors as determined by the known log P values (rather than its n nearest neighbors as determined by the similarity measures). All of the measures, conversely, are very much better than the performance denoted by "Randomized" in Figure 2, which involves replacing each of the nearest neighbors for each target structure with randomly selected structures prior to the estimation of the target log P values.

Table 1 reports the results of paired t test comparisons between the 2D measure and the various EVA measures. With the Manhattan EVA measures, the tests indicate that there are no significantly different means (SDMs) to those of the 2D measure with the exception of a limited range of n values when $\sigma = 50$ cm^{-1}. Plots of the mean absolute error of log P prediction indicate that, for this limited range of n values, the EVA measure gives better average log P predictions than does the 2D measure. However, these predictions are better by about 0.10−0.12 log units, which is only just outside the original uncertainty in the measured log P values. With the Dice EVA measures, the tests indicate that there are many significantly different predictions to those of the 2D measure for all values of σ. However, except for the 4 cm^{-1} measure, these differences only arise if $n > 12$, the value of n depending upon the σ term. The 20 cm^{-1} measure provides the least number of significantly different results to those of the 2D measure, whilst σ terms both lower and higher than 20 cm^{-1} give predictions that are increasingly significantly different to those of the 2D measure. Plots of the mean absolute error of log P prediction indicate that the Dice measures based on σ terms of 10 cm^{-1} or more perform better than the 2D measure; however, the mean predictive differences, in the range 0.10−0.14 log units, are again not large when compared with the uncertainty in the measured log P values. With the Cosine EVA measure, nearly all of

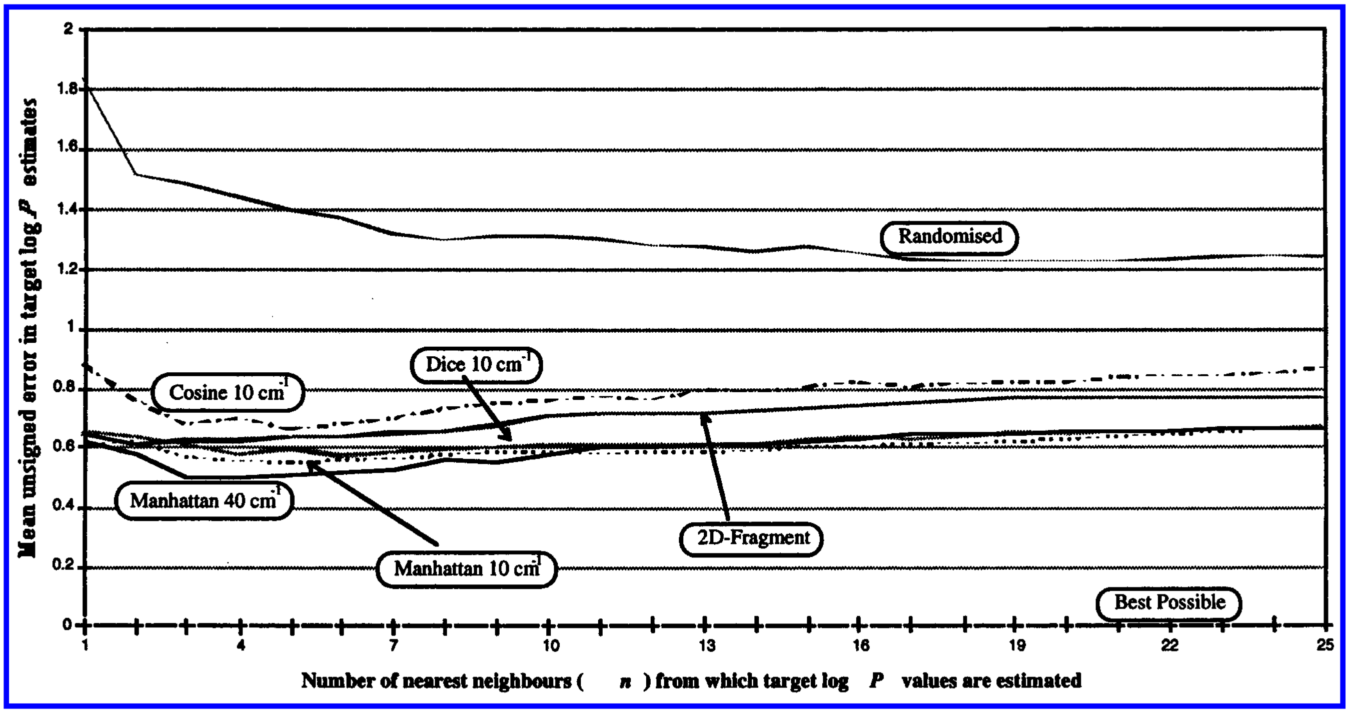


Figure 2. Mean unsigned error in log *P* estimated from 1 to 25 nearest neighbors using the 2D and three EVA similarity measures. The randomized and "Best Possible" benchmarks have also been indicated; the latter benchmark is indistinguishable from 0 on this scale. As expected, as *n* is increased the randomized estimates tend toward the mean of all the database log *P* values (1.23).

Table 1. Matched Pairs *t* Test Results for Comparing Log *P* Predictions Made Using 2D and EVA Similarity Measures[a]

	EVA similarity coefficient		
Gaussian term σ (cm^{-1})	Manhattan	Dice	Cosine
4	NONE ($\forall n\ p > 0.085$)	$\forall n\ p < 0.02$ $n > 1 \Rightarrow p < 0.005$	$\forall n$: $p \le 0.008$
10	NONE ($\forall n\ p > 0.18$)	$n > 12 \Rightarrow 0.005 < p < 0.05$	$\forall n$: $p < 0.006$
20	NONE ($\forall n\ p > 0.10$)	$n > 59 \Rightarrow 0.03 < p < 0.05$	$n > 2 \Rightarrow p < 0.003$
40	NONE ($\forall n\ p > 0.07$)	$n > 45 \Rightarrow 0.02 < p < 0.05$	$n > 1 \Rightarrow p < 0.05$ $n > 4 \Rightarrow p < 0.01$
50	$24 < n < 38 \Rightarrow 0.04 < p < 0.05$	$n > 23 \Rightarrow 0.02 < p < 0.05$	$n > 2 \Rightarrow p < 0.05$ $n > 5 \Rightarrow p \le 0.0008$

[a] The listed values of *n* (the number of nearest neighbors from which log *P* estimates are made) and *p* indicate *t* tests that show significantly different means ($p < 0.05$); "NONE" indicates that the log *P* predictions are not significantly different for any of the 100 values of *n*.

the tests indicate significant differences in predictive performance between the EVA and 2D measures. The absolute error plots show that the average log *P* predictions are little different (<0.05 log units) to those of the 2D measure; the $\sigma = 4$ cm^{-1} results provide, again, an exception to this general conclusion, with the Cosine log *P* prediction errors being 0.14−0.27 log units poorer than those of the 2D measure.

Overall it can be said that the performance of the EVA measures is in many cases significantly different to that of the 2D measure, particularly where the Dice or Cosine coefficient is used. However, these significant differences are not observed in the region of the nearest neighbor lists where a browsing chemist might be expected to pay most attention or where one might carry out an actual prediction (say, up to ten structures).

Table 2 reports pairwise comparisons of the significance of the differences between the sets of predictions made using different similarity coefficients and equivalent σ terms. It appears that for equivalent values of σ, the differences between the Manhattan−Cosine and Dice−Cosine pairs of measures are significant for most values of *n*, including very low values of *n*. The differences between the Manhattan and Dice are, for most values of *n*, highly significant only for $\sigma = 4$ cm^{-1} and $\sigma = 10$ cm^{-1}, while for σ terms of 20 cm^{-1} or more there are generally SDMs only where *n* is larger than about 45. Absolute error plots indicate that the Dice and Manhattan coefficients perform equally well except for $\sigma = 4$ cm^{-1}, where the Dice coefficient performs better than the Manhattan coefficient. This performance, with a few exceptions where *n* is less than about 10, is in all cases better than that of the Cosine-based measures. Turner[30] also reports a detailed analysis of the effect of the σ parameter on predictive performance and demonstrates, as is suggested by some of the results in Table 1, that the better EVA results tend to be obtained with the larger values of σ.

Comparison of Nearest Neighbor Lists. The second set of experiments studied the extent to which the EVA and 2D similarity measures tend to return alternative orderings of essentially the same sets of nearest neighbors or whether they tend to return sets of nearest neighbors that are quite different from one another. Counts were made, for each and every target, of how many of the top *x* nearest neighbors obtained using one similarity measure were present in the top *y* nearest neighbors obtained using one of the other measures, with the results being given as mean values over the same set of

Table 2. Matched Pairs *t* Test Results for Comparing Log *P* Predictions Made Using Selected Pairs of EVA−Based Similarity Measures[a]

	similarity coefficients under comparison		
gaussian term σ (cm^{-1})	Manhattan and Cosine	Manhattan and Dice	Dice and Cosine
4	$\forall n$: $p < 0.02$ $n > 2 \Rightarrow p < 0.00001$	$\forall n$: $p < 0.02$ $n > 2 \Rightarrow p < 0.0002$	$n > 4 \Rightarrow p < 0.05$ $n > 5 \Rightarrow p < 0.01$
10	$\forall n$: $p < 0.01$	$n > 5 \Rightarrow p < 0.01$	$\forall n$: $p < 0.05$ $n > 2 \Rightarrow p < 0.01$
20	$\forall n$: $p < 0.05$ $n > 3 \Rightarrow p < 0.00004$	$46 < n < 53$ and $n > 59 \Rightarrow p < 0.05$ $n > 68 \Rightarrow p < 0.01$	$n > 2 \Rightarrow p < 0.01$ $n > 4 \Rightarrow p < 0.0002$
40	$n > 3 \Rightarrow p < 0.01$	$n > 44 \Rightarrow 0.01 < p < 0.05$	$n > 3 \Rightarrow p < 0.01$
50	$n > 3 \Rightarrow p < 0.05$ $n > 5 \Rightarrow p < 0.0002$	$n = 9 \Rightarrow p \approx 0.03$ $n > 52 \Rightarrow 0.001 < p < 0.05$	$n > 2 \Rightarrow p < 0.04$ $n > 4 \Rightarrow p < 0.003$

[a] See the caption to Table 1 for further information.

Table 3. Counts of the Mean Number of EVA Nearest Neighbors (When Averaged over 100 Target Structures) in Common with 2D Nearest Neighbors[a]

EVA similarity measure	nearest neighbors under comparison (*x*/*y*)					
	5/5	5/100	30/30	30/100	100/30	100/100
			Manhattan			
4	1.2	2.8	5.6	7.7	8.5	13.5
10	1.6	3.7	8.8	12.7	12.5	22.0
20	1.6	3.7	8.8	12.9	12.6	23.0
40	1.7	3.6	8.4	12.5	12.1	22.1
50	1.6	3.6	8.3	12.4	12.1	22.0
			Dice			
4	1.4	3.2	7.8	11.6	11.4	20.4
10	1.5	3.5	8.0	12.1	11.9	22.0
20	1.5	3.5	7.7	11.6	11.4	21.4
40	1.5	3.4	7.4	11.2	11.0	20.7
50	1.5	3.3	7.4	11.3	11.0	20.7
			Cosine			
4	1.3	3.1	6.6	9.6	9.6	16.7
10	1.4	3.4	7.4	10.7	10.8	19.3
20	1.5	3.5	7.9	11.6	11.4	21.2
40	1.5	3.5	7.8	11.6	11.6	21.7
50	1.4	3.4	7.8	11.6	11.6	21.5
mean	1.5	3.4	7.7	11.4	11.3	20.5

[a] The table lists the number of nearest neighbors in the top *x* (5, 30, or 100) nearest neighbors of each EVA measure found in the top *y* (5, 30, or 100) nearest neighbors of the 2D measure for various combinations of *x* and *y*.

100 target structures as were used in the predictive experiments. However, these mean values clearly do not indicate the differences between the specific hit lists of individual targets, and a closer look was hence taken at the types of structure retrieved by the various similarity measures for a (necessarily) limited number of selected targets.

The results of the pairwise-count experiments are presented in summary form in Table 3, which lists the number of nearest neighbors in the top *x* (5, 30, or 100) nearest neighbors for each EVA measure that were found in the top *y* (5, 30, or 100) nearest neighbors for the 2D measure. The results indicate that, overall, the hits returned by the 2D measure are quite different to those returned by the EVA measures; for example, there are on average only 20.5 EVA top 100 hits appearing in the top 100 of the 2D hits. On the whole, each of the EVA measures shares a very similar average number of hits with the 2D measure, with the 4 cm^{-1}-based measures generally having the least numbers of such shared nearest neighbors. The target structures for which there appear to be the least difference between the EVA measures as a whole (that is, when averaged over all the EVA measures), and the 2D measure are those for which there are sufficient numbers of close analogues present in the Starlist. In these cases both the EVA-based and 2D-based measures tend to return the analogues as nearest neighbors. Some examples of these structures are given in Table 4. For example, there are a large number of mono- and dihalogenated ethylacetoacetatebenzal structures present such that, in a 30/30 comparison, the use of ethylacetoacetate,3-chlorobenzal as a target gives an average of 21.8 hits shared by the EVA and 2D nearest neighbor lists.

If those targets are considered for which there appears to be the greatest difference between the EVA measures as a whole and the 2D measure, then for 16 of the target structures there were found to be no shared top five nearest neighbors for any one of the EVA-based similarity measures when compared to those of the 2D-based measure. Table 5 details the numbers of shared nearest neighbors for some of these target structures where there was a noticeable lack of overlap in the search outputs, with the top three nearest neighbors of the 2D and Manhattan $\sigma = 40$ cm^{-1} searches for some of these target molecules being shown in Figures 3−6. In these figures, (a) is the target structure, (b) the top three EVA nearest neighbors, and (c) the top three 2D UNITY nearest neighbors. Part (d) of these figures is discussed later in the paper.

Indomethacin is a target structure in Table 5 that had no structures common to the top 30 hits of the 2D measure and *any* of the EVA measures, and an average of only 2.1 of the top five EVA nearest neighbors present in the top 100 of the 2D nearest neighbors. The top three indomethacin nearest neighbors for the EVA and 2D searches are illustrated in Figure 3. As indicated in the figure, the log *P* values for the EVA nearest neighbors are, in this case, very much closer to that of the target than are those of the 2D nearest neighbors. The EVA nearest neighbors, whilst having roughly the same number of heavy atoms (25) as the target, are structurally quite diverse. In contrast both the first two nearest neighbors in the 2D search contain the indole substructure found in the target itself and have 10 or so fewer heavy atoms; this observation is discussed further below.

1,4-Benzodioxan provides another example of a molecule retrieving radically different sets of EVA and 2D nearest neighbors. Figure 4 shows that the 2D top three nearest neighbors are very different to and have very much higher molecular weights than the target structure, despite quite high similarity scores of 0.86 to 0.89. The EVA measure here provides nearest neighbors that are much more obviously comparable to the target.

Generally, the larger molecular weight target structures, such as podophyllotoxin and fluxofenim shown in Figures

Table 4. Counts of the Mean Numbers of EVA Nearest Neighbors (When Averaged over All 15 EVA Measures) in Common with 2D Nearest Neighbors for a Number of Selected Targets Where There Is a Noticeable Degree of Commonality in the Search Outputs[a]

	nearest neighbors under comparison (*x*/*y*)					
target	5/5	5/100	30/30	30/100	100/30	100/100
ethylacetoacetate,3-chlorobenzal	2.4	5.0	21.8	25.4	25.9	38.3
cyanamphenicol	0.0	5.0	21.3	24.2	25.3	32.7
AC−Ala−Tyr−Phe−N[b] tripeptide	0.8	5.0	12.7	29.7	24.1	68.3
4-fluorotryptophan	3.9	4.7	12.4	16.5	16.9	28.2

[a] See the caption to Table 3 for further information. [b] The Starlist peptide-nomenclature uses AC to identify the C-terminal amino acid residue and N to identify the N-terminal residue.

Table 5. Counts of the Mean Numbers of EVA Nearest Neighbors (When Averaged over All 15 EVA Measures) in Common with 2D Nearest Neighbors for a Number of Selected Targets Where There Is a Noticeable Difference in the Search Outputs[a]

	nearest neighbors under comparison (*x*/*y*)					
target	5/5	5/100	30/30	30/100	100/30	100/100
indomethacin	0.0	2.1	0.0	4.8	1.0	8.7
fluxofenim	0.0	1.1	1.1	3.5	2.1	7.6
podophyllotoxin	0.0	1.3	1.1	4.1	2.1	9.7
chlorothalonil	0.0	0.3	1.9	4.9	7.1	20.9
1,4-benzodioxan	0.0	1.5	1.8	4.5	3.5	9.5
2-phenoxypyridine	0.0	1.4	1.6	4.7	2.7	10.0

[a] See the caption to Table 3 for further information.

5 and 6, seem to give quite structurally diverse sets of hits with both the EVA and 2D measures. However, the EVA nearest neighbors tend to have a much closer number of atoms to the target than the 2D nearest neighbors. This behavior is exemplified by the figures in Table 6, which detail the differences in size, expressed in terms of counts of either all atoms or heavy atoms alone (*i.e.*, excluding hydrogens), between the target structures and the nearest neighbor structures retrieved by the 2D and Manhattan $\sigma = 40$ cm^{-1} searches, when averaged over all 100 target structures. Matched pairs *t* tests indicate that the 2D and EVA hit atom-counts are statistically significantly different ($p \leq 0.000031$) for all *n*.

Overall, the qualitative analyses described in this section indicate that the different (or similar) log *P* predictive performance of the EVA-based and 2D-based measures is not simply the result of a better ordering of essentially the same set of nearest neighbors but is often due to the retrieval of quite different sets of nearest neighbors. We hence suggest that the EVA-based and 2D-based measures provide valid, alternative sets of nearest neighbors. It thus seems at least feasible that it might be possible to identify a single, combined measure that would retrieve a set of nearest neighbors characterizing both types of measure and offering a better level of predictive performance than either of them. This suggestion is considered further in the next section.

COMBINATION OF SIMILARITY MEASURES BY DATA FUSION

Introduction to Data Fusion. Data fusion is the generic name given to a range of techniques for combining different sources of information into a single decision function that can provide better estimates, in this case of bioactivity, than can the individual sources.[27,37,38] These techniques have been developed for use in situations where several different sensors evaluate a situation and then assign values to attributes describing that situation. Data fusion takes place when these individual evaluations are combined in some way to arrive at a final, overall decision that is expected to provide a more reliable judgment of the current situation than can the individual systems that have been combined. Other characteristics of data fusion include the following: the ability to draw conclusions even if one sensor is damaged since information is still available from the other sensors (which is of importance in the many military applications that have been reported); the coverage of disparate areas, times, and quantities (as occurs when several different types of spectroscopic evidence are used to identify an unknown chemical sample in structure-elucidation studies) and an increased level of confidence in the results since different sensors are more likely to agree when they are issuing correct, as against incorrect, decisions (as when several doctors are consulted to obtain agreement on a difficult diagnosis).

Examples of the application of data fusion techniques that have been studied thus far include signal processing (where several different sensors try to discriminate between signals carrying useful information and those that represent merely noise of some sort) and optical character recognition (where several different recognition algorithms try to identify a particular character). The work described in this section originates from studies of the use of data fusion in a further application domain, that of ranked-output searching in databases of textual documents. Here, a query, containing keywords that describe a user's information need, is compared with the natural-language texts of documents in a database to find those documents that are most similar to the query.[39,40] The similarity measures that are used for text retrieval are all based on the number of keywords common to a document and a query, but there are many ways in which this number can be weighted, *e.g.*, by taking account of the numbers of times that query words occur in the document and/or in the database as a whole or by using natural-language processing techniques that allow for the matching of phrases, rather than of just the individual words. It is thus possible to generate many different rankings of a given set of documents in response to a given query, and there has been much recent interest in the use of data fusion to combine such variant rankings so as to increase the numbers of relevant documents that are retrieved.[41−44] In the following we investigate the effectiveness of this approach in combining the rankings resulting from the use of different measures of chemical similarity.

Implementation of Data Fusion for Similarity Searching. Assume that a target molecule is searched against a database using two different similarity measures, giving two different rankings of the database in order of decreasing similarity with the target structure. Data fusion methods can then be applied either to the raw similarity coefficient values

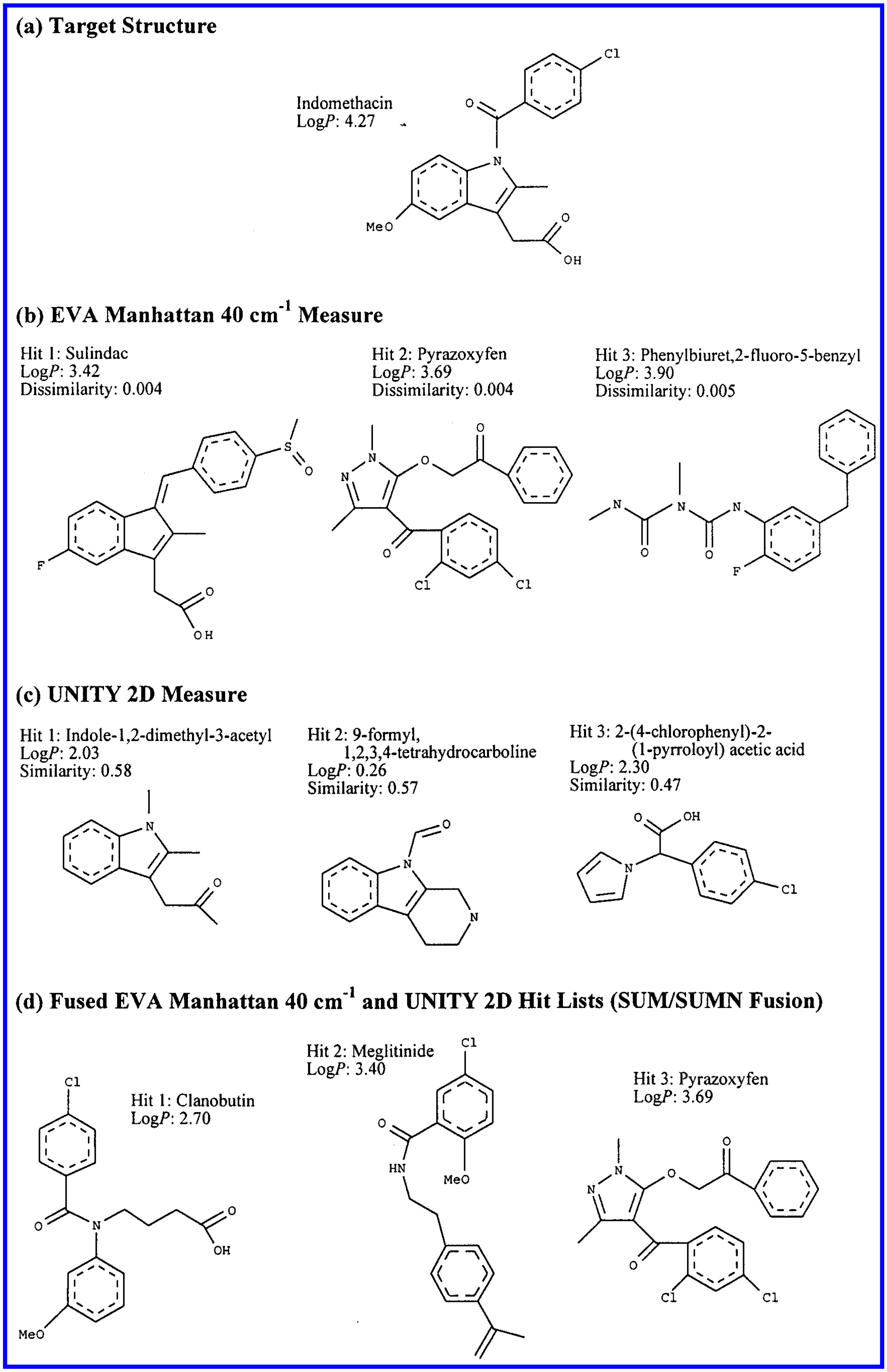
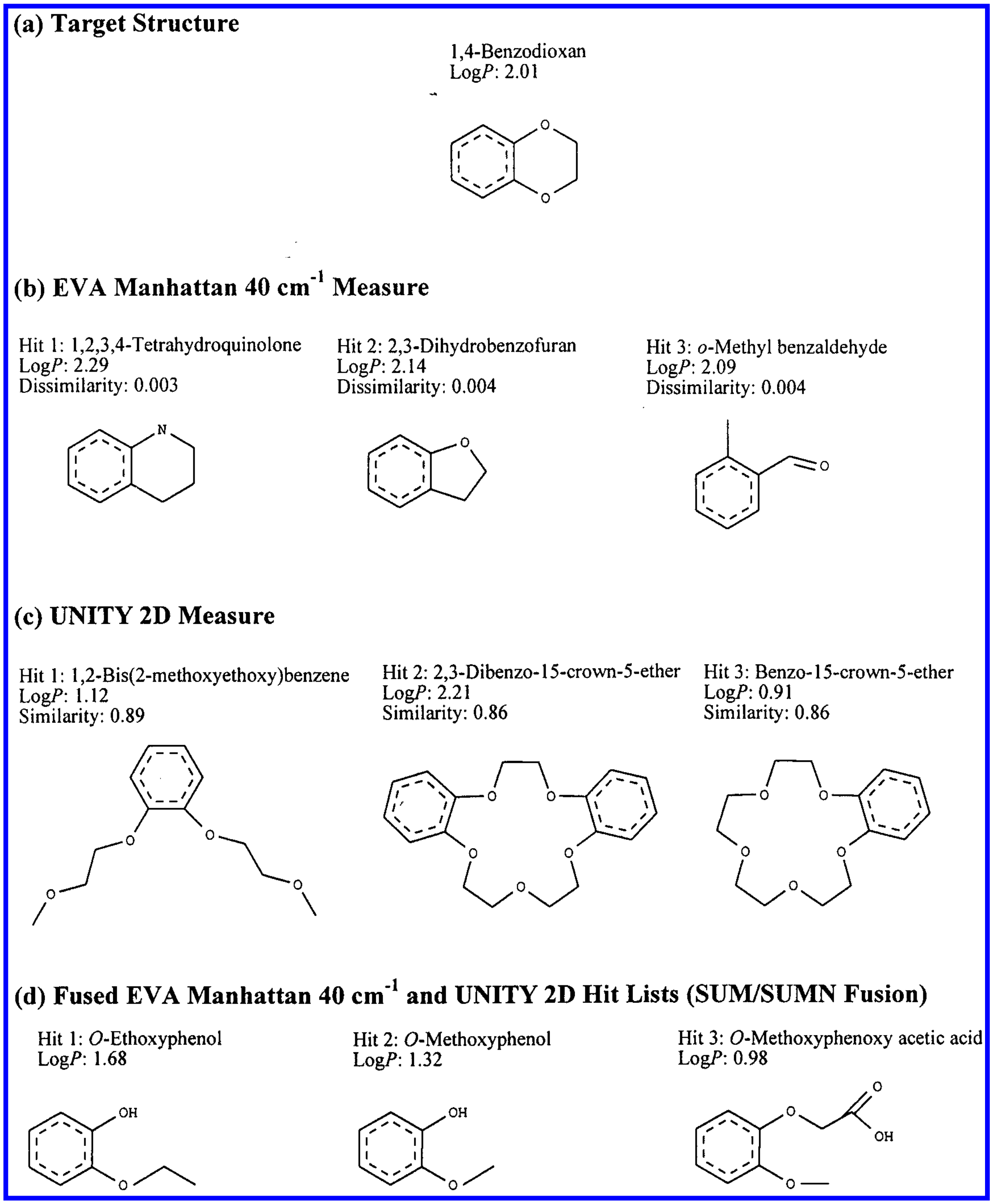


Figure 3. (a) Indomethacin target structure; top three nearest neighbors for (b) EVA Manhattan $\sigma = 40$ cm^{-1}, (c) 2D UNITY, and (d) fused EVA and 2D UNITY measures. Hydrogens suppressed and stereochemistry omitted.

output by each measure or to the rank positions that are obtained when the coefficient values are sorted into decreasing order (or increasing order if a distance coefficient, rather than an association coefficient, is used). We have chosen to combine the rankings in the work reported here for two reasons. Firstly, because it avoids the need to identify an



(a) Target Structure

1,4-Benzodioxan
Log*P*: 2.01

(b) EVA Manhattan 40 cm^{-1} Measure

Hit 1: 1,2,3,4-Tetrahydroquinolone
Log*P*: 2.29
Dissimilarity: 0.003

Hit 2: 2,3-Dihydrobenzofuran
Log*P*: 2.14
Dissimilarity: 0.004

Hit 3: *o*-Methyl benzaldehyde
Log*P*: 2.09
Dissimilarity: 0.004

(c) UNITY 2D Measure

Hit 1: 1,2-Bis(2-methoxyethoxy)benzene
Log*P*: 1.12
Similarity: 0.89

Hit 2: 2,3-Dibenzo-15-crown-5-ether
Log*P*: 2.21
Similarity: 0.86

Hit 3: Benzo-15-crown-5-ether
Log*P*: 0.91
Similarity: 0.86

(d) Fused EVA Manhattan 40 cm^{-1} and UNITY 2D Hit Lists (SUM/SUMN Fusion)

Hit 1: *O*-Ethoxyphenol
Log*P*: 1.68

Hit 2: *O*-Methoxyphenol
Log*P*: 1.32

Hit 3: *O*-Methoxyphenoxy acetic acid
Log*P*: 0.98

Figure 4. (a) 1,4-Benzodioxan target structure; top three nearest neighbors for (b) EVA Manhattan $\sigma = 40$ cm^{-1}, (c) 2D UNITY, and (d) fused EVA and 2D UNITY measures. Hydrogens suppressed and stereochemistry omitted.

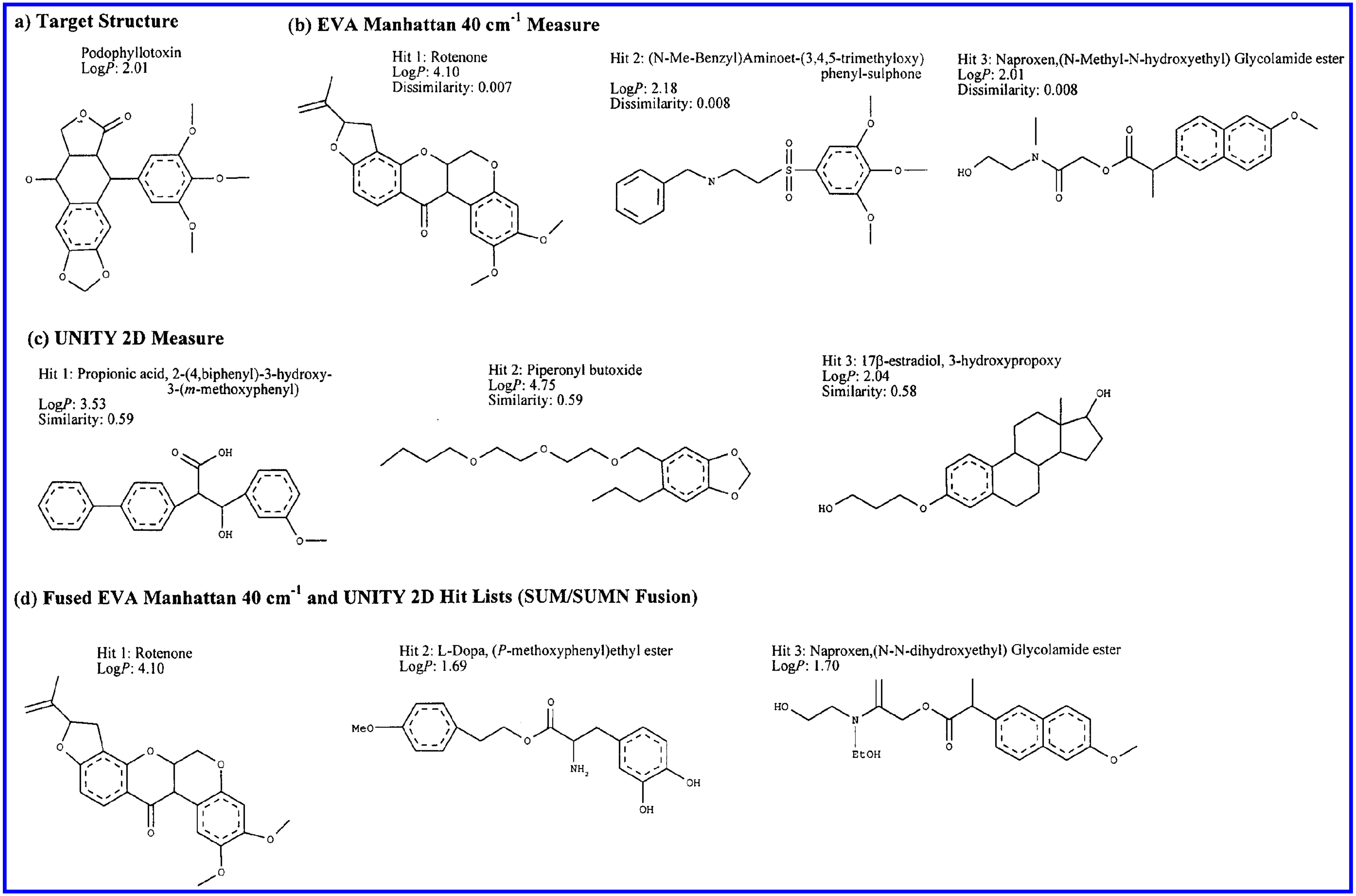
appropriate standardization method, and, secondly, since searchers are normally interested in some number of nearest neighbors, rather than those molecules that have similarities greater than some threshold value (this latter search constraint implying expert knowledge of the characteristics of the particular similarity measure that is being used to generate the rankings).

Formally, assume that a target molecule is matched against a database using a similarity measure s_1 and that the *i*th molecule ($1 \leq i \leq I$, where I is the number of molecules in the database) receives a similarity score such that it is ranked at position rs_1 ($1 \leq rs_1 \leq n$, where n is the number of nearest neighbors that the user wishes to see as the output from the similarity search). Similarly, rs_2 is the rank position of that same *i*th molecule when a second similarity measure, s_2, is used. Data fusion then involves the calculation of a new score of the form

$$f(rs_1, rs_2)$$

for each of the up to 2*n* molecules retrieved by the individual searches, and their reranking in order of these new scores. The fusion procedure thus takes two rankings as its input and returns a single, fused ranking that can then be evaluated in just the same way as a ranking resulting from one of the individual measures. Note that while the preceding discussion has considered only pairs of rankings, the same basic technique is applicable to any number of rankings.

Following the studies of Belkin *et al.* on combining search outputs in text retrieval,[41,42] we have used the following fusion criteria for the creation of combined rankings: SUM, SUMN, and MIN. SUM takes the sum of the individual



a) Target Structure

(b) EVA Manhattan 40 cm^{-1} Measure

(c) UNITY 2D Measure

(d) Fused EVA Manhattan 40 cm^{-1} and UNITY 2D Hit Lists (SUM/SUMN Fusion)

Podophyllotoxin
LogP: 2.01
Hit 1: Rotenone
LogP: 4.10
Dissimilarity: 0.007
Hit 2: (N-Me-Benzyl)Aminoet-(3,4,5-trimethyloxy) phenyl-sulphone
LogP: 2.18
Dissimilarity: 0.008
Hit 3: Naproxen,(N-Methyl-N-hydroxyethyl) Glycolamide ester
LogP: 2.01
Dissimilarity: 0.008
Hit 1: Propionic acid, 2-(4,biphenyl)-3-hydroxy-3-(m-methoxyphenyl)
LogP: 3.53
Similarity: 0.59
Hit 2: Piperonyl butoxide
LogP: 4.75
Similarity: 0.59
Hit 3: 17β-estradiol, 3-hydroxypropoxy
LogP: 2.04
Similarity: 0.58
Hit 1: Rotenone
LogP: 4.10
Hit 2: L-Dopa, (P-methoxyphenyl)ethyl ester
LogP: 1.69
Hit 3: Naproxen,(N-N-dihydroxyethyl) Glycolamide ester
LogP: 1.70
HO
N
OH
MeO
NH2
EtOH

Figure 5. (a) Podophyllotoxin target structure; top three nearest neighbors for (b) EVA Manhattan $\sigma = 40$ cm^{-1}, (c) 2D UNITY, and (d) fused EVA and 2D UNITY measures. Hydrogens suppressed and stereochemistry omitted.

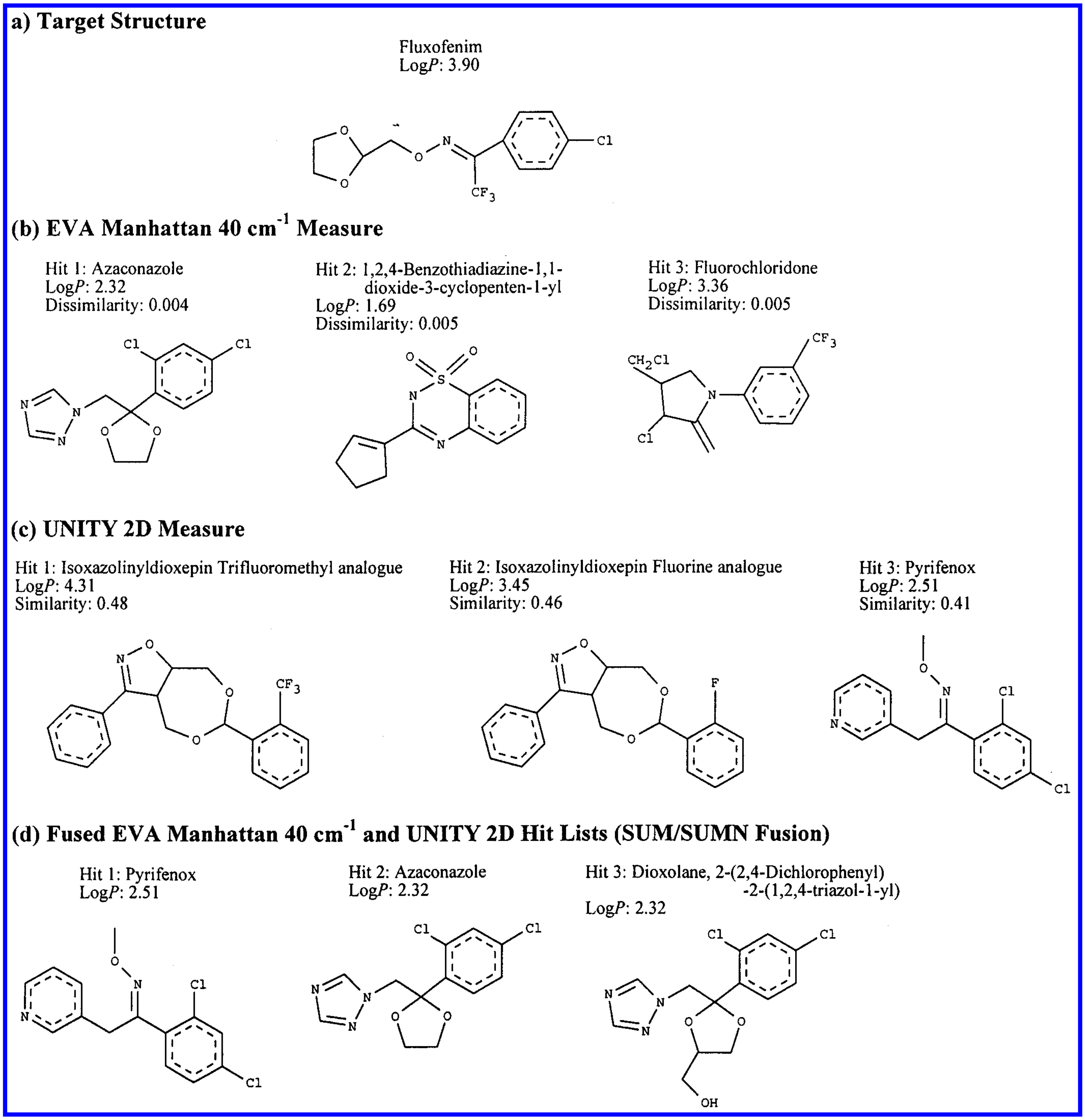


Figure 6. (a) Fluxofenim target structure; top three nearest neighbors for (b) EVA Manhattan $\sigma = 40$ cm^{-1}, (c) 2D UNITY, and (d) fused EVA and 2D UNITY measures. Hydrogens suppressed and stereochemistry omitted.

ranks, rs_1 and rs_2, for each molecule. A default value of $n + 1$ is used for rs_1 (or rs_2) if a molecule does not appear in the set of nearest neighbors resulting from the corresponding similarity measure. SUMN sums the ranks for each molecule and then divides the result by the number of rankings for which the molecule was in the set of n nearest neighbors, so that a molecule that was ranked moderately by both methods would fare better than one that was ranked highly by one method and not at all by the other. MIN takes the minimum of the two ranks for each molecule, with a default value of $n + 1$ for rs_1 (or rs_2) if a molecule does not appear in the set of nearest neighbors resulting from the corresponding similarity measure. We have also used a further criterion, MAX, that is additional to those tested by Belkin *et al.* MAX takes the maximum of the two ranks for each molecule, with a default value of 0 for rs_1 (or rs_2) if a molecule does not appear in the set of nearest neighbors resulting from the corresponding similarity measure. MAX is a counter-intuitive criterion for data fusion, focusing as it does on the worst possible result given by an individual measure and thus provides a check on the utility of the other, more intuitive criteria.

Evaluation of Fused Rankings. The initial experiments used the four criteria above to fuse the 2D UNITY rankings with the EVA rankings based on the Manhattan 40 cm^{-1} EVA measure. The four types of fused rankings are compared with the input EVA and 2D rankings using the simulated property prediction approach that has been described previously, with 100 target structures being searched against the Starlist file of 8178 molecules. A summary of

Table 6. Mean (μ), Standard Deviation (st dev), and Range of Absolute Differences between the Target and Nearest Neighbor Atom Counts Taken over All 100 Target Structures for Various Values of n^a

		2D			EVA Manhattan 40 cm^{-1}			
n	atoms	μ	st dev	range	μ	st dev	range	matched pairs t test scores p
1	all	3.61	4.29	0−23	1.43	1.56	0−7	8.6×10^{-7}
	heavy only	1.92	2.17	0−12	0.82	0.91	0−4	2.6×10^{-6}
2	all	3.88	4.15	0−25	1.50	1.47	0−6	6.3×10^{-8}
	heavy only	2.01	2.31	0−13	0.95	0.94	0−4	1.6×10^{-6}
5	all	5.16	5.65	0−32	2.06	1.86	0−9	2.1×10^{-7}
	heavy only	2.61	2.89	0−16	1.30	1.18	0−4	3.1×10^{-5}
10	all	7.60	7.75	0−40	2.06	1.48	0−6	2.6×10^{-10}
	heavy only	3.62	3.48	0−16	1.38	1.28	0−6	2.2×10^{-8}
25	all	7.49	7.57	0−39	2.37	1.68	0−7	1.3×10^{-9}
	heavy only	3.48	3.39	0−17	1.66	1.23	0−6	7.0×10^{-6}
100	all	10.83	9.01	0−40	3.06	2.29	0−13	4.2×10^{-13}
	heavy only	4.92	4.16	0−18	2.05	1.70	0−9	1.8×10^{-9}

[a] This has been done for the 2D UNITY and EVA Manhattan 40 cm^{-1} searches. Counts were done both including (all) and excluding (heavy only) hydrogen atoms from consideration. The matchedpairs t test was applied to assess the statistical significance of the differences between the atom counts.

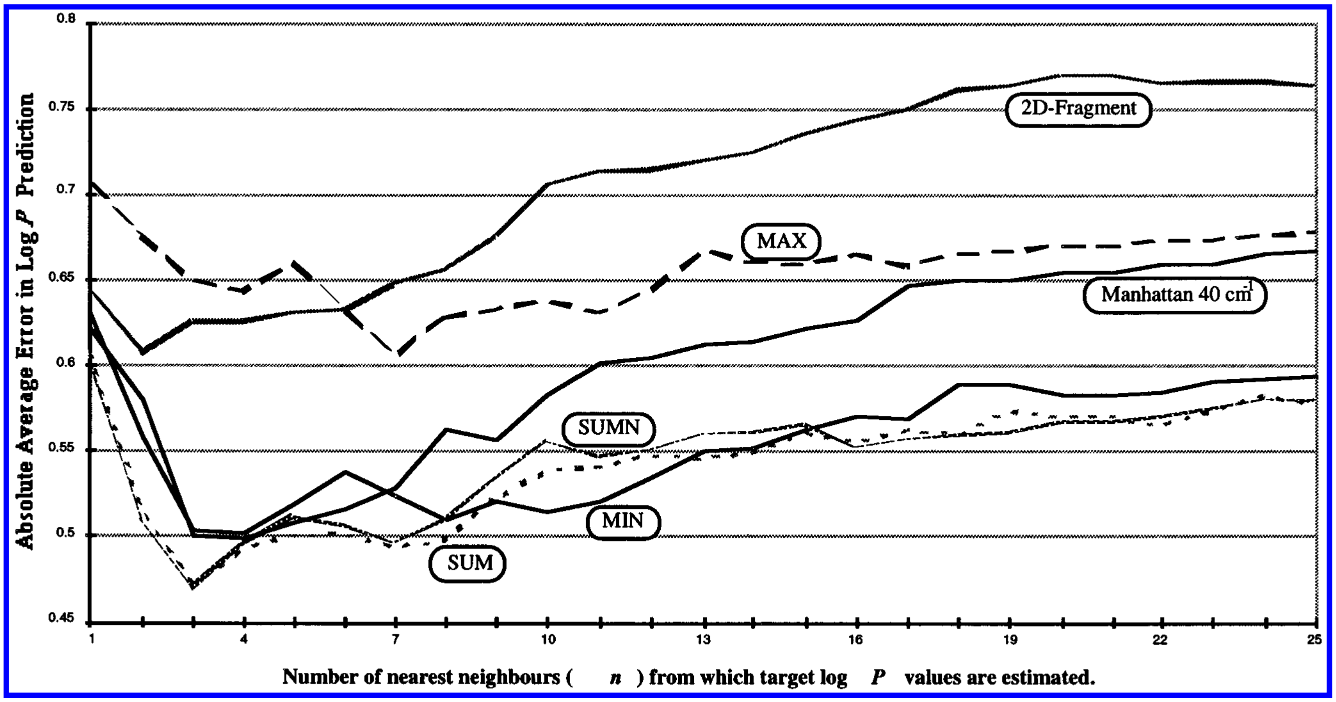


Figure 7. Mean unsigned error in log P estimated from 1 to 25 nearest neighbors using the EVA Manhattan 40 cm-1 and 2D measures and fused rankings derived from the hitlists of these two measures. Matched pairs t tests (Table 7) are used to establish the statistical significance of the differences indicated in this figure.

the results is presented in Figure 7, which shows mean unsigned error plots for the estimation of log P using between 1 and 25 nearest neighbors. With the obvious exception of the MAX fusion criterion, all of the fused rankings appear for the most part to be better than the original 2D and EVA rankings. Similar results to these are obtained when the 2D measure is fused with any one of the EVA measures described previously; *e.g.*, Figure 8 provides plots analogous to those given in Figure 7 but using the Dice 4 cm^{-1} EVA measure.

The statistical significance of these results is again tested by means of the matched pairs t test described above; in this case a restricted range of n values are tested, *viz.*, n = 1, 2, 5, 10, 20, 50, and 100. This was done for the Manhattan-, Dice-, and Cosine-based measures where σ = 4 and 40 cm^{-1}. The results of these tests indicate that only where $n \geq 5$ is there a tendency for a fused measure to give statistically significantly different log P predictions to those of one or other of the original unfused measures. Furthermore, it is found that only in a very few cases are the fused predictions statistically significantly different to *both* sets of original unfused rankings. Thus, while inspection of plots such as those given in Figures 7 and 8 strongly suggest that fusion tends to improve property-prediction performance, the t tests indicate that these differences are statistically significant for only about 50% of the cases tested. In addition, there appeared to be no obvious relationship between the presence or absence of SDMs between the original (unfused) EVA and 2D UNITY measures and the subsequent presence or absence of SDMs between these measures and the derived fused measures.

Comparison of Nearest Neighbor Lists. It seems reasonable to expect that little is to be gained from data fusion for those target structures where the hit lists that are being combined have many nearest neighbors in common; instead a minor reordering of the hits is likely to result. This

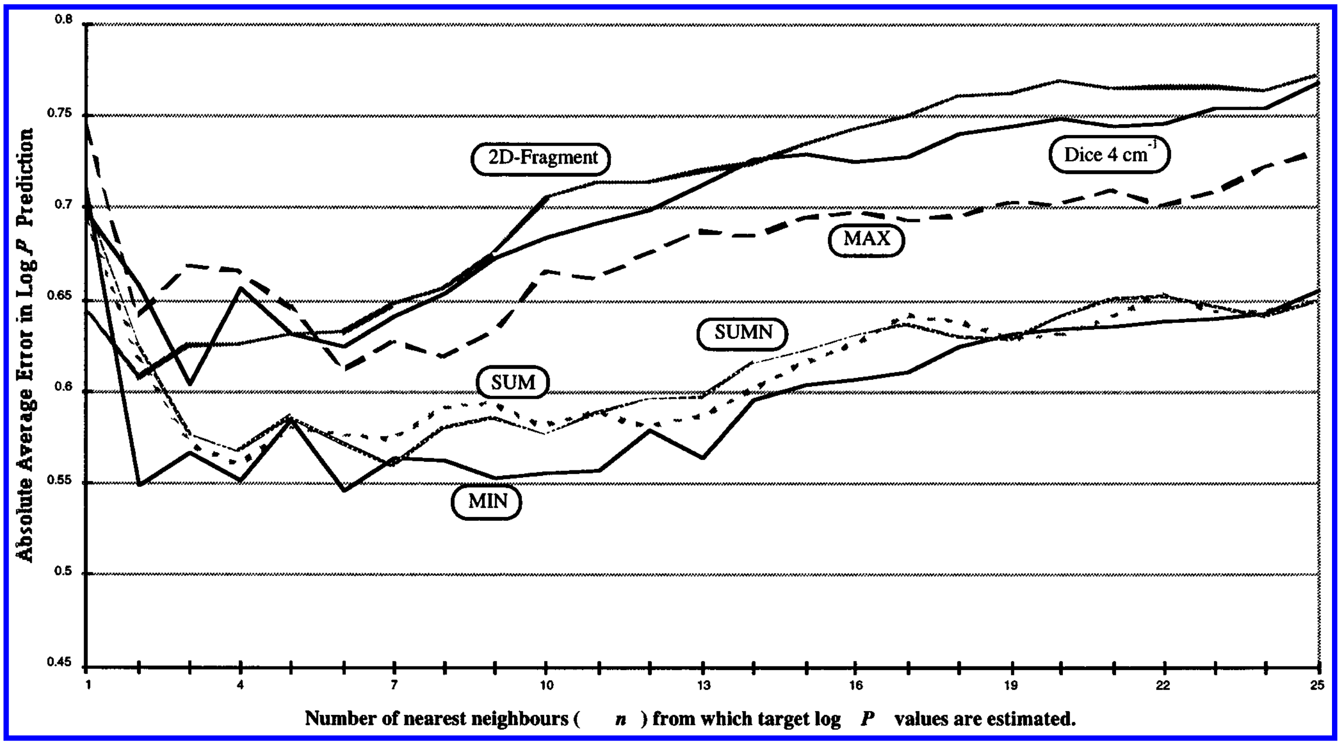


Figure 8. Mean unsigned error in log *P* estimated from 1 to 25 nearest neighbors using the EVA Dice 4 cm^{-1} and 2D measures and fused rankings derived from the hitlists of these two measures. Matched pairs *t* tests (Table 7) are used to establish the statistical significance of the differences indicated in this figure.

Table 7. Mean (μ), Standard Deviation (st dev) and Range of Absolute Differences between the Target and Nearest Neighbor Atom Counts Taken over All 100 Target Structures for Various Values of *n*[a]

		fused			matched pairs *t* test scores	
n	atoms	μ	st dev	range	Manhattan 40 cm^{-1} *p*	Unity 2D *p*
1	all	2.09	2.18	0−10	5.15×10^{-4}	6.60×10^{-5}
	heavy only	1.16	1.16	0−5	7.44×10^{-4}	1.05×10^{-4}
2	all	1.86	1.92	0−7	6.38×10^{-2}	2.78×10^{-6}
	heavy only	1.26	1.29	0−6	3.67×10^{-3}	4.79×10^{-4}
5	all	2.40	2.70	0−18	2.12×10^{-1}	1.07×10^{-5}
	heavy only	1.45	1.51	0−6	2.80×10^{-1}	2.14×10^{-4}
10	all	3.05	2.99	0−17	1.86×10^{-3}	7.28×10^{-8}
	heavy only	1.76	1.73	0−8	4.04×10^{-2}	2.76×10^{-6}
25	all	4.09	4.30	0−22	2.65×10^{-4}	6.82×10^{-5}
	heavy only	2.34	2.35	0−12	8.95×10^{-3}	2.77×10^{-3}
100	all	6.97	8.68	0−47	2.41×10^{-5}	4.16×10^{-4}
	heavy only	3.58	3.91	0−21	2.15×10^{-4}	1.03×10^{-2}

[a] This has been done for fused (SUM) 2D UNITY and EVA Manhattan $\sigma = 40$ cm^{-1} searches. See the caption to Table 6 for further information.

is confirmed when looking at, for example, the hits obtained using the target structures listed in Table 4. The targets of greatest interest are, therefore, those for which there appear to be substantial differences between the nearest neighbor lists that are to be fused. A number of such target structures and their respective hits have been previously identified (Table 5 and Figures 3−6), and these target structures are discussed further below, with particular reference to the fusion of the Manhattan 40 cm^{-1} and 2D rankings.

For indomethacin (Figure 3), the EVA log *P* predictions remain the best, and the 2D predictions remain the worst. The MIN hits are better than 2D UNITY but worse than all others, including those of SUM and SUMN, which in these tests frequently but not always gave the same fused rankings. The fused (SUM/SUMN) nearest neighbors shown in Figure 3d (and the EVA only hits) are much more comparable in size (based upon heavy and all atom counts) to the target than are those of the 2D UNITY measure. This is a general finding when fusion is applied using the SUM/SUMN methods but, due to the nature of the MIN criterion, is often less marked when that fusion criterion is used. In this particular case, the MAX criterion provides better predictions than 2D alone; generally, however, this fusion criterion has a log *P* prediction performance that is worse than that of any of the other measures, whether fused or not.

In the case of 1,4-benzodioxan, the EVA and MIN log *P* predictions are the best, with the SUM/SUMN fused hits shown in Figure 4d again being much more similar in size and structure to the target structure than are the 2D hits. The SUM/SUMN hit list provides the best log *P* predictions for podophyllotoxin (Figure 5), with the MIN hits here being poorer than EVA alone but better than the 2D measure alone. Finally, for fluxofenim, the 2D measure log *P* predictions remain substantially the best. The SUM/SUMN fused hits shown in Figure 6d provide log *P* predictions that are better than those of the EVA measure but are, nonetheless, much poorer than those of the 2D measure alone. In this case, the best fused search was that based on the MIN criterion.

We have commented previously on the disparity in molecular size between the target structure and either the EVA or the 2D nearest neighbor structures. The left-hand portion of Table 7 details the difference in size (mean, standard deviation, and range) between the target structure and selected nearest neighbors in the fused (2D and Manhattan 40 cm^{-1}) rankings and thus complements the results presented previously in Table 6. The right-hand portion of Table 7 compares these differences (using matched pairs *t* tests) with those for the individual rankings that were merged to form the fused ranking using the SUM fusion criterion. It will be seen that fusion has lessened the substantial differences in molecular size identified in Table 6, with an output that, as might be expected, is intermediate between those resulting from the individual 2D and EVA rankings.

Thus far, we have considered the fusion of an EVA ranking with the 2D ranking. Experiments were also carried out in which the Cosine, Dice, and Manhattan 4 cm^{-1} rankings were fused, and the resultant fused ranking compared with the rankings for the individual measures. While there were small increases in predictive performance, these were far less than when an individual ranking was fused with the 2D ranking, and entirely comparable results were obtained when the Cosine, Dice, and Manhattan 40 cm^{-1} rankings were fused. This suggests, hardly surprisingly, that data fusion is most appropriate when the individual rankings that are being combined are very different in character from each other.

CONCLUSIONS

In this paper, we have discussed the use of measures based on the EVA descriptor for similarity searching and compared its performance with that of a standard similarity searching procedure based on the matching of 2D fingerprints. The simulated property-prediction experiments have established that a limited number of the EVA measures significantly outperform the 2D measure, although this behavior is restricted to certain ranges of nearest neighbors and of EVA Gaussian parameters.

Overall, given the computationally intensive, albeit "once-only", preprocessing required to calculate normal mode frequencies using semiempirical methods, further testing of the effectiveness of EVA-based similarity measures is required before its use could be generally recommended. There is also a clear requirement for a comparison to be made of the effectiveness of EVA in relation to other efficient, alignment-free 3D measures, such as atom-mapping[13] or atom-triplets.[15] However, EVA-based searching would be a more attractive proposition if the computational requirements of this preprocessing stage could be substantially reduced, *e.g.*, by using molecular mechanics rather than semiempirical techniques.

The 2D measure is constrained by the requirement that its nearest neighbors must share identical, albeit small, specific fragments with the target although the interconnectivity and multiplicity of these fragments is not specified. There is no such requirement in the case of the EVA measure, where similarity is based on a descriptor in which structural features are implicit rather than explicit. It is thus, perhaps, not very surprising that the EVA measures retrieve sets of nearest neighbors that are often very different to the sets retrieved by the 2D measure; moreover, these tend to contain structures with more similar numbers of atoms to the target than those of the 2D measure.

The different sets of structures retrieved by the EVA and 2D measures suggest that their outputs might usefully be combined, and experiments with several different data fusion criteria show that such combined rankings can in some cases yield superior results, in terms of simulated property prediction. While the idea of combining different similarity measures is not new,[13] there have been only two systematic studies, of which we are aware, of this approach to similarity searching, both of which appeared while the present paper was being prepared for publication. Masui and Yoshida have used the SUMN criterion to combine the similarity scores (rather than the ranks) obtained in searches of a database containing mass, IR, and ^{1}H and ^{13}C NMR spectral data when one or more of the spectra are missing for a particular sample molecule.[45] In work more analogous to that reported here, a group at Merck[7,8] used the MIN fusion criterion and a similarity-based, rather than rank-based, version of SUM to combine pairs of 2D and 3D rankings in searches of the Standard Drug File database and found that significant improvements in performance could be achieved in simulated property prediction experiments. Taken with the experiments reported here, we believe that data fusion provides a simple, and potentially valuable, way of improving the effectiveness of similarity searching in chemical databases.

ACKNOWLEDGMENT

We thank the following: the Biotechnology and Biological Sciences Research Council, the Engineering and Physical Sciences Research Council and Shell Research Centre for funding; BioByte Corp. for provision of the Starlist database; and Tripos Inc. for provision of the CONCORD, SYBYL, and UNITY software.

REFERENCES AND NOTES

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

(2) *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Chapman and Hall: Glasgow, 1994.

(3) Ash, J. E.; Warr, W. A.; Willett, P. *Chemical Information Systems*; Ellis Horwood: Chichester, 1991.

(4) Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. *Rev. Comput. Chem.* **1995**, *7*, 1−66.

(5) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.

(6) Hagadone, T. R. Molecular Substructure Similarity Searching: Efficient Retrieval in Two-Dimensional Structure Databases. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 515−521.

(7) 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.

(8) 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.

(9) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(10) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(11) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: a New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.

(12) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nearest-Neighbor Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36−41.

(13) Pepperrell, C. A.; Willett, P. Techniques for the Calculation of Three-Dimensional Structural Similarity Using Inter-Atomic Distances. *J. Comput.-Aid. Mol. Design* **1991**, 5, 455−474.

(14) Perry, N. C.; van Geerestein, V. J. Database Searching on the Basis of Three-Dimensional Molecular Similarity Using the SPERM Program. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 607−616.

(15) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. A New Method for Rapid Characterisation of Molecular Shape: Applications in Drug Design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79−85.

(16) Good, A. C.; Ewing, T. J. A.; Gschwend, D. A.; Kuntz, I. D. New Molecular Shape Descriptors: Application in Database Screening. *J. Comput.-Aid. Mol. Design* **1995**, *9*, 1−12.

(17) Fisanick, W.; Cross, K. P.; Rusinko, A. Similarity Searching on CAS Registry Substances. 1. Global Molecular property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664−674.

(18) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. Similarity Searching in Files of Three-Dimensional Chemical Structures: Comparison of Fragment-Based Measures of Shape Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141−147.

(19) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.

(20) Good, A. C.; Hodgkin, E. E.; Richards, W. G. Similarity Screening of Molecular Data Sets. *J. Comput.-Aid. Mol. Design* **1992**, *6*, 513−520.

(21) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of Similarity Coefficients and Standardisation Methods for Field-Based Similarity Searching. *SAR QSAR Environ. Res.* **1995**, *3*, 101−130.

(22) Wild, D. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Alignment of Molecular Electrostatic Potentials with a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 159−167.

(23) Johnson, M. A.; Maggiora, G. M.; Lajiness, M. S.; Moon, J. B.; Petke, J. D.; Rohrer, D. C. Molecular Similarity Analysis: Applications in Drug Discovery. In *Advanced Computer-Assisted Techniques in Drug Discovery*; VCH: New York, 1994; pp 89−110.

(24) Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington DC, 1995.

(25) *3D QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993.

(26) Cocchi, M.; Menziani, M. C.; Fanelli, F.; De Bendetti, P. G. Theoretical Quantitative Structure-Activity Relationship Analysis of Congeneric and Non-Congeneric α_1-Adrenoceptor Antagonists: A Chemometric Study. *J. Mol. Struct. (Theochem)* **1995**, *331*, 79−93.

(27) Hall, D. L. *Mathematical Techniques in Multisensor Data Fusion*; Artech House: Boston, 1992.

(28) Jonathan, P.; McCarthy, W. V.; Roberts, A. M. I. Discriminant Analysis with Singular Covariance Matrices. A Method Incorporating Cross-Validation and Efficient Randomised Permutation Tests. *J. Chemomet.* **1996**, *10*, 189−213.

(29) Ferguson, A. M.; Heritage, T.; Jonathan, P.; Pack, S. E.; Phillips, L.; Rogan, J.; Snaith, P. J. EVA: a New Theoretically-Based Molecular Descriptor For Use in QSAR/QSPR Analysis. Submitted for publication.

(30) Turner, D. B. *An Evaluation of a Novel Molecular Descriptor (EVA) for QSAR Studies and the Similarity Searching of Chemical Structure Databases*; Ph.D. Thesis, University of Sheffield, 1996.

(31) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Evaluation of a Novel Infra-red Range Vibration-Based Descriptor (EVA) for QSAR Studies. 1. General Application. Manuscript in preparation.

(32) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Evaluation of a Novel Infra-red Range Vibration-Based Descriptor (EVA) for QSAR Studies. 2. Validation. Manuscript in preparation.

(33) Leo, A. J. Calculating log P_{oct} from Structures. *Chem. Rev.* **1993**, *93*, 1281−1306.32.

(34) Pearlman, R. S. Rapid Generation of High Quality Approximate 3D Molecular Structures. *Chem. Des. Automat. News* **1987**, *2*, 1−6.

(35) SYBYL and UNITY are produced by Tripos Inc., St Louis, MO 63144 U.S.A.

(36) Stewart, J. J. P. MOPAC: A Semi-Empirical Molecular Orbital Program. *J. Comput.-Aid. Mol. Design* **1990**, *4*, 1−105.

(37) Crowley, J. L.; Demazeau, Y. Principles and Techniques for Sensor Data Fusion. *Signal Proc.* **1993**, *32*, 5−27.

(38) Kokar, M.; Kim, K. Preface to the Special Section on Data Fusion: Architectures and Issues. *Control. Eng. Pract.* **1994**, *2*, 803−809.

(39) Salton, G. *Automatic Text Processing*; Addison-Wesley: Reading, MA, 1989.

(40) *Information Retrieval. Data Structures and Algorithms*; Frakes, W. B., Baeza-Yates, R., Eds.; Prentice Hall: Englewood Cliffs, 1992.

(41) Belkin, N. J.; Kantor, P.; Cool, C.; Quatrain, R. Combining Evidence for Information Retrieval. In *TREC-2, Proceedings of the Second Text Retrieval Conference*; Harman, D., Ed.; National Institute of Standards and Technology: Gaitherburg, 1994, pp. 35−44.

(42) Belkin, N. J.; Kantor, P.; Fox, E. A.; Shaw, J. A. Combining the Evidence of Multiple Query Representations for Information Retrieval. *Inf. Proc. Manag.* **1995**, *31*, 431−448.

(43) Voorhees, E. M.; Gupta, N. K.; Johnson-Laird, B. Learning Collection Fusion Strategies. *Proceedings of the 18th International Conference on Research and Development in Information Retrieval*. **1995**, 172−179.

(44) Lee, J. H. Combining Multiple Evidence from Different Properties of Weighting Schemes. *Proceedings of the 18th International Conference on Research and Development in Information Retrieval* **1995**, 180−188.

(45) Masui, H.; Yoshida, M. SPECTRA: a Spectral Information Management System Featuring a Novel Combined Search Function. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 294−298.

CI960466U