The Application of Neural Networks in Conformational Analysis. 1. Prediction of Minimum and Maximum Interatomic Distances

Shaun N. Jordan and Andrew R. Leach*,†

Department of Chemistry, University of Southampton, Southampton, Hampshire SO17 1BJ, U.K.

John Bradshaw†

Glaxo Research and Development Ltd., Park Road, Ware, Hertfordshire SG12 0DP, U.K.

Received February 13, 1995®

Feed-forward neural networks have been used to predict the maximum and minimum distances between pairs of heteroatoms in a wide variety of molecules. The distances are predicted solely from information derived from connection tables. The data to train and test the networks were obtained by extracting from a database all compounds with a defined number of atoms in the shortest path between a specified pair of heteroatoms and by subjecting such molecules to a conformational search in order to determine the range of distances possible. This information was then used to train the neural networks to predict the minimum and maximum distances between the pairs of heteroatoms. The performance of each network was evaluated using a set of test molecules. Paths containing six and nine atoms were considered, with the molecules being extracted from the Cambridge Structural Database and the Available Chemicals Database, respectively. For the six-atom problem the cumulative effect of incorporating the various network descriptors was also investigated. For the six-atom path the networks were able to predict 75% of the maximum distances and 50% of the minimum distances within a tolerance of 0.5 Å. The predictive ability decreased for the nine-atom problem to 74% for the maximum distances and 47% for the minimum distances. The performance of the networks was compared with a random simulation using the appropriate distribution of distances; the neural networks were able to predict the appropriate distance with an accuracy between 25% and 36% better than the random method.

INTRODUCTION

The properties of a molecule are intimately linked to the conformations that it adopts and so an understanding of a molecule's conformational space is important in rationalizing and predicting its behavior. This is particularly so in current approaches to drug design, where the structure of a pharmacologically important macromolecule or alternatively a 3D pharmacophore of the binding site is often used to discover new inhibitors or other strongly binding molecules. The availability of such structural information means that it is important to consider not only the chemical properties of the proposed molecules (i.e., whether they possess the required functional groups) but also their structural properties (i.e., whether can they position these groups in the appropriate relative orientations to satisfy the three-dimensional requirements of the binding site).

There are two basic approaches to structure-based design.[1] Methods in the first category aim to identify previously synthesised molecules that show the correct properties to interact with the receptor. This is usually achieved by searching a "three dimensional" (3D) database.[2] Methods in the second class use the model of the binding site to design completely new molecules, an approach often referred to as "de novo" design. In this paper we shall be concerned with the database-searching approaches to ligand design and discovery.

The size of most corporate or commercially available chemical databases means that a huge number of potential ligands must often be considered. The requirement is thus to identify as efficiently as possible those compounds that best match the chemical and structural requirements of the target receptor site. This is usually achieved by gradually filtering out molecules according to their chemical and conformational properties. In general, it is found that the elimination of molecules based upon their "two-dimensional" properties (e.g., presence of functional groups) can be performed rather effectively, and many database systems have offered such a capability for several years. It thus remains to consider the structural properties.

The first 3D databases contained only a single conformation for each molecule, and it was relatively straightforward to check whether or not that conformation could satisfy the geometric constraints. It has always been recognized that to store just a single conformation (or even several representative conformations) may be a significant limitation, as many molecules are very flexible and there is no guarantee that the conformation(s) stored in the database would be the one adopted at the binding site. Nevertheless, some notable successes have been reported using this limited approach.[3] Several solutions to the problem of including "full" conformational flexibility into a 3D database search have been proposed, and some of these algorithms have been implemented in commercially available modeling systems. The ChemDBS-3D system from Chemical Design[4] performs a conformational search on each molecule as it is entered into the database. The conformational search generates a series of binary keys which code the distances between pairs of

† Current address: Glaxo Research & Development Ltd., Glaxo Medicines Research Centre, Chemistry Building, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, U.K.

® Abstract published in *Advance ACS Abstracts*, April 1, 1995.

0095-2338/95/1635-0640$09.00/0

"pharmacophoric groups". The default set of pharmacophoric groups comprises heteroatoms, hydrogen bonding acceptors, hydrogen bond donors, and ring centroids. At search time, the molecular keys are compared with a key appropriate to the target pharmacophore. Matching molecules are then subjected to the conformational search once more in order to regenerate the appropriate conformation. A key feature of the conformational search algorithm is the use of "rules"[5] which enable the space to be explored rapidly. The Catalyst system[6] from Molecular Simulations Inc. generates a set of conformations that are intended to span the conformational space of each molecule in the database. An alternative approach is to perform the conformational analysis at search time. The UNITY system[7] of Tripos Associates stores in the database a single conformation of each molecule. This conformation is then "adjusted", by rotating single bonds, to try and force it to satisfy the requirements of the pharmacophore. The Tripos system uses a torsional minimizer[8] that is derived from the "tweak" algorithm of Fine, Shenkin, and colleagues[9,10] to perform the adjustment. The tweak algorithm gives the changes in dihedral angle that are required to enable a pair of atoms to adopt a specified distance. The MACCS-3D search system from MDL Information Systems[11] uses a similar approach, but here a target function that relates the variation in an interatomic distance to the changes in torsion angles of the intervening bonds is minimized. Clark, Willett, and Kenny have investigated the use of distance geometry algorithms in 3D database searching.[12] Their system has two components, the first of which is a screen based upon triangle smoothing that eliminates molecules that could not possibly achieve the required distance. The second component of their system is an explicit conformational search using a standard distance geometry algorithm. The incorporation of conformational flexibility using these methods was found to greatly increase the computational time required in comparison with the rigid search.

A key feature of most approaches to 3D database searching is the use of distance screens, which aim to eliminate molecules that cannot satisfy the constraints. Two types of screens can be identified in current 3D systems: screens that calculate the range of distances between any pair of atoms and screens that implicitly represent the actual distances possible. Triangle smoothing[13] is a well-known approach for generating the limits on the minimum and maximum distances between atoms. Triangle smoothing has the advantage that the distance ranges it provides are self-consistent at the triangle level, dependent only upon the quality of the bond length, bond angle, and van der Waals radius information present. By contrast, distance screens derived from conformational search algorithms are limited by any inherent biases and inaccuracies of the conformational search method. The main drawback of triangle smoothing when used to generate upper and lower distances is that the bounds produced do not necessarily correspond to distances that can be achieved in a realistic, low-energy conformation of the molecule. In part this is because such a conformation must satisfy not only triangle inequalities but also quadrangle, pentangle, and hexangle inequalities.[14] A simple example suffices to illustrate this point: the lower bound distance obtained when triangle smoothing is used to calculate the lower bound distance between the amide nitrogen and the carbonyl oxygen of the carboxylic acid group in 4-acetamidobenzoic acid (Figure 1) is equal to the sum of the vdw radii (ca. 3.3 Å depending upon the van der Waals radii used), compared to 6.4 Å in the crystal structure of this molecule.

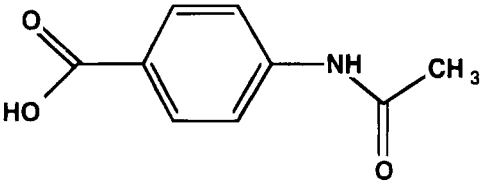

Figure 1. 4-Acetoamidobenzoic acid.

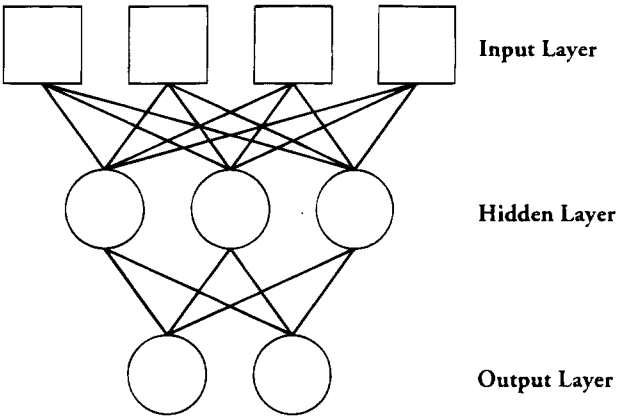

Figure 2. A schematic illustration of a 4 × 3 × 2 feed-forward neural network.

This example nicely illustrates the rationale behind the work reported in this paper: as chemists, we would be able to explain why the distance between the amide nitrogen and the carbonyl oxygen in 4-acetamidobenzoic acid must be significantly greater than the sum of the vdw radii and moreover should also be able to calculate an approximate value for the expected upper and lower bounds. This is done by recognizing the patterns inherent in the chemical diagram and the way in which molecules are constructed from smaller building blocks. It was this notion that led us to investigate the applications of neural networks in predicting the conformational properties of molecules.

NEURAL NETWORKS

Neural nets are considered to be one of the most exciting developments in computational science in recent years.[15] They have been applied to a wide variety of problems that involve the recognition of patterns and nonlinear relationships. Chemical applications[16] include the prediction of protein secondary structure from primary sequence,[17] the interpretation of NMR spectra,[18] and the rationalization of quantitative structure−activity relationships.[19] In this work we have used feed-forward neural networks which contain layers of elementary computational units (which we shall refer to as nodes) with connections between all pairs of nodes in adjacent layers, Figure 2. The input to each node is obtained by adding together the weighted outputs from the nodes to which it is connected in the previous layer. Thus

$$I_{pi} = \sum_j w_{ij} o_{pj}$$

I_{pi} is the input to the node i (in layer n, say) for pattern p, o_{pj} is the output from node j in the previous layer ($n - 1$), and w_{ij} is the strength (or weight) of the connection between i

and j. The output value for the node is obtained by passing the input value through a transfer function. A common transfer function is the logistic function

$$O_{pj} = \frac{1}{1 + e^{-I_{pj}}}$$

The logistic transfer function can be differentiated (which is important for the learning mechanism that is used). The transfer function is usually biased so that it steps around zero. This is achieved by adding a bias to each noninput node in the network

$$I_{pi} = \sum_j w_{ij} o_{pj} + \theta_i$$

A neural net must first be trained to perform the desired task. We use supervised learning in which the network is presented with a set of sample inputs and outputs, and the strengths of the connections (weights and biases) are then adjusted until the network finds the values giving the best agreement between the input and output. A variety of methods can be used to adjust the weights and biases, a particularly popular method being the generalized delta rule.[20] Using the current set of weights, each input pattern is presented to the network, and the corresponding outputs are determined. The error for a pattern is calculated as the square of the difference between the expected and the actual outputs for the output nodes

$$E_p = \frac{1}{2}\sum_i (t_{pi} - o_{pi})^2$$

E_p is the sum of squares error for pattern p, determined by summing the squares of the differences between the target outputs t_{pi} for the pattern on output nodes i and the actual output o_{pi} for the nodes i. The factor 1/2 is introduced for normalization reasons.

Defining s_{pi} as the net input to the output node i for the pattern p

$$s_{pi} = \sum_j w_{ij} o_{pj}$$

and defining δ_{pi} as

$$\delta_{pi} = -\frac{\partial E_p}{\partial s_{pi}}$$

It can be shown that for an output node

$$\delta_{pi} = f'_i(s_{pi})(t_{pi} - o_{pi})$$

f'_i is the first derivative of the transfer function (hence the requirement that the transfer function is continuous and differentiable). For a hidden node no specific target value is available, but these can be computed from the errors in the nodes in the next layer forward (the method is thus often described as "backpropagation"). For a hidden node j we have

$$\delta_{pi} = f'_i(s_{pi}) \sum_j \delta_{pj} w_{ij}$$

The weight between the units i and j is changed using the following expression

$$\Delta_p w_{ij}(t) = \eta \delta_{pi} o_{pj}$$

η is the learning rate. If η is infinitesimally small, then the generalized delta rule is equivalent to a true steepest descents algorithm. However, η is typically assigned a value between 0.1 and 1.0. In common with other minimization procedures, the generalized delta rule can become stuck in local minima; to attempt to overcome this problem, a momentum term is included which incorporates previous weight changes into the calculation

$$\Delta_p w_{ij}(t) = \eta \delta_{pi} o_{pj} + \alpha \Delta_p w_{ij}(t - 1)$$

Many variants and enhancements have been made to the generalized delta rule; for example, Monte Carlo methods can be used in conjunction with simulated annealing to explore the space.[21] One modification that is employed in the work reported here is the use of "fuzzy data"; random noise (in the form of a small random perturbation of a noninput node's activation) is added to the network. This enhances the ability of the network to deal with noisy data and to generalize.[22]

As we have seen, the adjustment of weights in a feed-forward neural network is an iterative process. If this process is performed indefinitely, there is a possibility that the network will become too specific to the training data set and so reduce its ability to correctly predict unseen data. This is called overtraining. Therefore, the training process should be stopped when the predictive ability of the network starts to fall. This is achieved by using three data sets—a training set, a testing set, and a cross-validating set. The training set is used to adjust the weights using the generalized delta rule; the cross-validating set is used to detect overtraining by monitoring the predictive ability of the network; and the testing set is used to test the performance of the trained network. Training thus proceeds until the total error falls below a given value, or until overtraining is indicated by reference to the cross-validating data set. Overtraining is indicated when the ability of the network to predict the output for the cross-validating data set starts to systematically fall. Once trained, the net can then be used in a predictive fashion: when presented with an input the network will generate an output. In this work we evaluate the performance of the network by presenting it with a test set of molecules that were not used during the training phase.

Neural networks can only accept numerical inputs, and so it is necessary to devise an appropriate means of representing the input and output patterns. We shall use the term descriptor to refer to one or more features of the input pattern. Descriptors can be either binary or continuous. Binary descriptors take values of 0 or 1, whereas continuous descriptors may take any value. To illustrate the difference between binary and continuous descriptors, let us consider how a set of traffic lights, where only one light is on at any time, might be characterized:

	Binary	Continuous
red	{1,0,0}	{1}
amber	{0,1,0}	{2}
green	{0,0,1}	{3}

There can be problems using continuous descriptors to describe non-numeric, discrete features as they assign significance to relationships which do not exist in the binary data. In the above example, if we use continuous descriptors, then we are stating that the green light has three times the significance of the red light when input to a neural network.

This relationship is clearly not present when binary descriptors are used. The inputs to a neural network should ideally fall within the range 0 to 1 when using the backpropagation learning algorithm. It is therefore common to normalize continuous descriptors between 0.1 and 0.9, to cope with the possibility of a new pattern (e.g., from a molecule in the testing set) requiring a value beyond the range of values of the molecules in the training set. Outputs were similarly normalized to lie between 0.1 and 0.9.

The number of variables in a neural network is equal to the number of weights (i.e., connections) added to the number of biases (which equals the number of nodes not in the input layer). For a three-layer network with I input nodes, H hidden nodes, and O output nodes the total number of adjustable parameters is thus $H(I + O) + H + O$. If there are many fewer input patterns than parameters, then the neural network may simply "memorize" the data and be unable to make any generalizations. Similarly, if there are too many pieces of data, then the network may be unable to generalize at all. The ratio of the number of patterns in the training set to the number of variables has been used in an ad hoc way to determine whether these problems of memorization and inability to generalize are likely to arise. A value between 1.8 and 2.2 has been suggested.[23]

PREDICTING INTERATOMIC DISTANCES FROM A 2-D REPRESENTATION

To try and predict the range of distances that can be achieved between a pair of atoms in a molecule it is clearly necessary to have some means of representing the constitution of the molecule to the neural network, and in particular those aspects of the molecule's constitution that are considered to be important in determining the minimum and maximum distances between the pair of atoms. Our current approach is based on the assumption that those parts of the molecule that lie between the two atoms are largely responsible for the range of distances that can be adopted.

When there are rings present in a molecule, then there may be more than one path between a pair of atoms. We have assumed that the shortest path determines the conformational properties of one atom relative to the other. Bradshaw and Maliski have previously described how the maximum possible distance between a pair of atoms may be predicted using a simple bond counting scheme along the "most restrictive path".[24] In most cases the most restricted path is the same as the shortest path. However, due to some ambiguity in the definition of the most restricted path we decided to use the shortest path as the basis for our work. In those cases where there was more than one path of minimum length, the first one found was arbitrarily selected as the one to use. We anticipated that some or all of the following features of the shortest path might be important in determining the conformations available to it: the atoms on the path and their hybridization states, the bond orders and any geometric isomerism, the degree of substitution at each atom, and stereochemical effects. It was thus necessary to have a means of representing these features to the network. For some properties a binary input is possible; for others, continuous values were used. The descriptors used at various stages of this work are as follows.

Bond order: The bonds along the path are classified as single, double, triple, or aromatic and described using four binary numbers as follows:

single:	{1,0,0,0}
double:	{0,1,0,0}
triple:	{0,0,1,0}
aromatic:	{0,0,0,1}

Atom types: To represent all possible atom types using binary descriptors would require too many descriptors (at least 10 bits per atom). Consequently, we decided to encode each atom according to its hybridization state and its row in the periodic table. Three binary descriptors were used to characterize hybridization:

sp^3	{1,0,0}
sp^2	{0,1,0}
sp	{0,0,1}

Two binary descriptors were used to describe the period of each atom:

1st row	{1,0}
Other rows	{0,1}

Substitution at atoms on the path: Bulky substituent groups can play an important role in determining the accessible conformations of a molecule, a classic example being the influence of bulky groups on the conformations of cyclohexane rings. We required a scheme that can represent substitution in a concise format. Our method is based on Kier and Hall's algorithm for calculating the topological states of an atom within a molecule.[25] The topological state implicitly contains information about the atom types, bonding, and substitution patterns within the molecule. To determine the topological state, a "delta" value δ_i is determined for each non-hydrogen atom within the molecule:

$$\delta_i = \frac{Z_i^v - h_i}{Z_i - Z_i^v - 1}$$

Z_i is the number of electrons in atom i, Z_i^v is the number of valence electrons in atom i, and h_i is the number of hydrogens bonded to atom i.

All the paths between atoms i and j in the molecule are found and the elements of the topological matrix **T** are calculated

$$t_{ij} = \frac{(\prod_{k=1}^{n_{ij}} \delta_k)^{1/n_{ij}}}{n_{ij}}$$

n_{ij} is the number of atoms in the path between i and j.

Where more than one path exists between atoms i and j, the t_{ij} value of each path is calculated separately and added together to give the matrix element. The topological state of the atom is then determined by adding the rows of the topological state matrix **T**

$$S_i = \sum_{j=1}^{N} t_{ij}$$

N is the number of non-hydrogen atoms in the molecule. The topological state implicitly includes information about the bonding and atom type by virtue of the electron counts and number of bonded hydrogens in the delta values;

information about substitution comes from the use of intramolecular paths when calculating the elements of the topological matrix. As we explicitly represent bond order and atom type information using our other descriptors we developed a simplified topological state to represent just substitution patterns. This was done by assigning a δ value of 1 to each atom. The elements of the topological state matrix then become

$$t_{ij} = \frac{1}{n_{ij}}$$

The topological state of each atom is calculated by adding the rows of the topological state matrix and normalizing by dividing by $\sqrt{N}$.

One intuitively expects "local" substitution to be more important than substitution far removed from the site of interest. We have therefore extended our substitution-based topological states indicators to describe local substitution. This is done by setting t_{ij} to zero for paths longer than some cutoff. Thus

$$t_{ij} = \frac{1}{n_{ij}} \text{ if } n_{ij} < \text{cutoff}$$

$$t_{ij} = 0 \text{ if } n_{ij} \geq \text{cutoff}$$

In this work, topological state descriptors with cutoffs of three atoms and six atoms were employed. It was anticipated that the S3 descriptor would characterize the local substitution at the atoms actually on the path and S6 would indicate the presence of substitution and also ring systems just off the path. Figure 3 and Table 1 show how these descriptors reflect these types of substitution. In the unsubstituted chain in molecule A (Figure 3) the S3 values increase as one moves toward the center of the path. However, in molecule B, S3 peaks at atom 2 (where the local substitution is the greatest) and also shows a small increase at atom 5, the cutoff removing any effect due to the distant naphthalene group. In molecule C, the S3 descriptor does not differentiate between the isopropyl substituent and the phenyl ring. However, these are distinguished by the S6 value due to its larger cutoff value.

Stereochemistry: Geometric isomerism (the presence of cis or trans double bonds in the path) was coded using a continuous descriptor d

$$d = \frac{\text{number of cis double bonds}}{\text{maximum number of double bonds in path}}$$

The denominator equals the total number of double bonds wholly contained within the path (two for a path of six atoms and three for a path of nine atoms). The presence of rings and the stereochemical features associated with them can affect the interatomic distance between atoms, Figure 4. Two binary descriptors were used to encode whether a path travels cis or trans through a ring. The presence of one of these two descriptors also serves to indicate to the network that the path passes through a ring. The stereochemical descriptors were chosen to be global (i.e., functions of the entire path rather than for each bond or atom) to reduce the number of descriptors required.

A separate neural network was used for each problem, characterized according to the path length and whether the minimum or maximum distance was to be predicted. The data used to train and test each of the networks was generated as follows. Molecules containing one or more pairs of heteroatoms separated by a shortest path of a specified length were extracted from either the Cambridge Structural Database (CSD[26]) or from the Available Chemicals Database (ACD[27]). The conformational search program COBRA[28] was then used to explore the conformational space of each molecule in order to determine the range of distances for each pair of heteroatoms separated by the appropriate path length. The following heteroatoms were considered: O, N, P, S and halogen, giving a total of 15 possible heteroatom pairs. In the case of molecules from the CSD, the stereochemistry of the molecule was taken to be that of the X-ray structure.

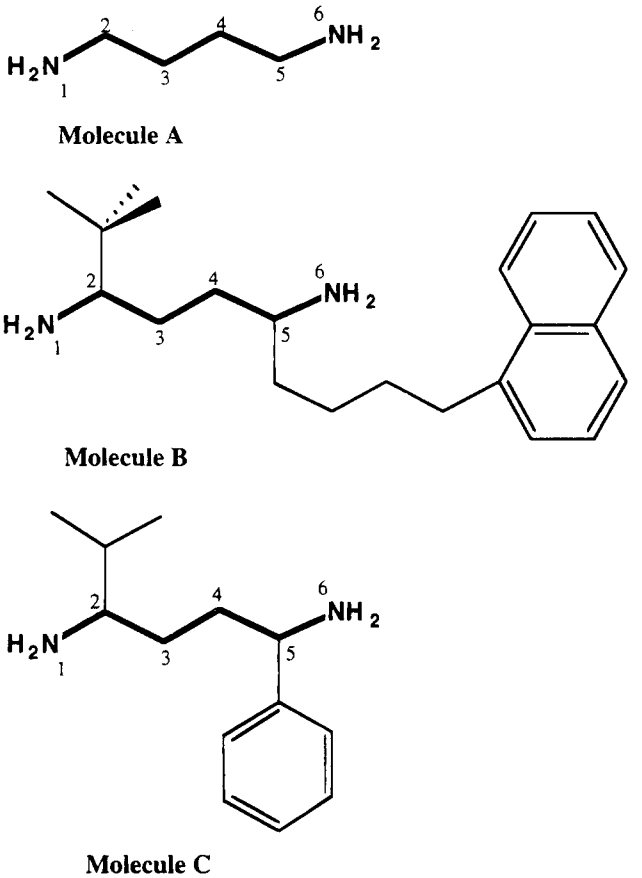

Figure 3. Use of cutoffs in the calculation of the topological state that indicate varying degrees of local substitution. For molecule B the local substitution is most pronounced at atom 2 and so the S3 descriptor is largest for that atom. The S6 descriptor takes into account substitution further removed from the path, and so the value at atom 5 is larger than for atom 2.

Table 1. Effect of the Topological State Cutoff in Characterizing Substitution Types

	atom number					
	1	2	3	4	5	6
			molecule A			
S3	1.83	2.33	2.66	2.66	2.33	1.83
S6	2.45	2.78	2.91	2.91	2.78	2.45
			molecule B			
S3	2.17	3.83	3	3	3.17	2.17
			molecule C			
S3	2.17	3.5	3	3	3.5	2.17
S6	3.45	4.32	4.43	5.14	5.73	4.18

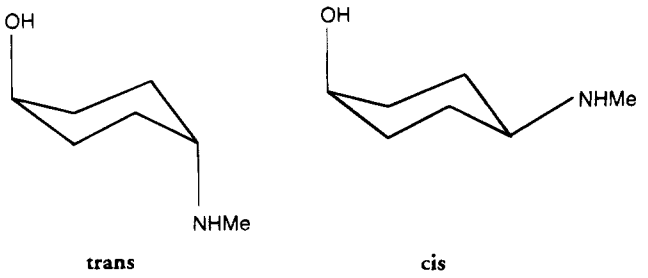

Figure 4. Effects of ring stereochemistry on interatomic distances; the distance between the hydroxyl oxygen and the amine nitrogen is different for the two isomers shown.

The molecules in the ACD were supplied in the form of three-dimensional structures generated by the CONCORD program.[29] As CONCORD assigns an arbitrary configuration to the molecule, all unique configurations were generated by COBRA and then subjected to conformational analysis. COBRA uses a model-building approach to explore the conformational space of molecules.[5] The program contains a database of molecular fragments, each of which has associated with it one or more three-dimensional conformations (templates). The templates correspond to the conformations that the fragment can adopt when present in larger molecules. For example, the cyclohexane ring has the chair, twist boat, and boat templates. The program determines which of the fragments are present in the molecule, and it then constructs three-dimensional conformations by joining together appropriate templates. Different combinations of templates give rise to different conformations of the molecule. The conformational space explored by the program corresponds to all possible combinations of templates and can be searched using a variety of algorithms such as the depth-first search. We have also developed more sophisticated ways of exploring the search tree, and in this work we have used a new method based on the A* algorithm that is able to directly identify the two conformations in which a specified pair of atoms achieve their minimum and maximum separations.[30] For those molecules that contain more than one pair of heteroatoms separated by the appropriate path length, all such pairs of atoms were considered.

As with other methods for exploring conformational space,[31,32] various user-selectable parameters determine how a conformational search is performed with COBRA, and the criteria used to evaluate whether a conformation is "acceptable" or not. Principal among these is the pairwise interatomic close-contact ratio: should the ratio of the sum of the vdw radii for any pair of atoms in a 1,*n* relationship ($n > 3$) to the interatomic separation of the atoms in a partial or full conformation exceed the threshold then that conformation (or partial conformation) is rejected. This parameter was set to 2.0 in the current study. While the structures produced by COBRA are not minimized, a close-contact ratio of 2.0 means that all of the conformations produced will be near energy minima. If the template energy threshold is activated then higher energy templates (e.g., boat cyclohexane) are initially not used in the conformational search. Such higher energy forms are only considered if no acceptable conformations can be generated using their lower energy counterparts. In this study, however, no threshold was applied, and all templates were available. The strain relief algorithms,[33] which are employed if no acceptable conformations can be produced using the default templates, are not at present applicable to the search methods based on the A* algorithm, and so molecules for which only strained conformations could be generated were ignored. We have also described the use of a distance geometry algorithm to generate conformations for parts of a molecule (e.g., large rings) that are not present in the fragment database and to construct conformations for highly strained ring systems which cannot be constructed by joining together rigid templates.[34] These algorithms were also not employed in the studies reported here. Finally, an upper time limit of 5 min was placed on the analysis of any one molecule.

Having explored the conformational space and derived the minimum and maximum distances between appropriate pairs of heteroatoms, the data were divided into training and testing sets for each path length, the networks were trained, and their performance was then evaluated using the testing set. The training and testing data were selected so as to ensure that they contained a representative sample of distances. The testing data sets contained approximately 25% of the total data. Each pattern input to the network consisted of a path, together with the associated minimum or maximum distance obtained from the conformational search. The distances were scaled to floating numbers between 0.1 and 0.9. For each problem, various network architectures were evaluated, though due to the size of the networks and the amount of training data it was not possible to perform an exhaustive exploration. For the six-atom path problem we investigated how the predictive performance of the networks changed as the various descriptors were incorporated. In all cases the performance of the networks were compared with the results that could be expected by simply selecting random numbers from the observed distribution of distances.

RESULTS

The total numbers of molecules (i.e., unique configurations) subjected to the conformational search were 6842 and 15 371 for the six- and nine-atom path problems, respectively. After the conformational search, the minimum and maximum distances between a total of 5716 unique six atom paths and 8004 unique nine atom paths were obtained. Many of the nine-atom path molecules contained stereocenters off the path, resulting in many configurations that had the same shortest path between the end atoms. This proliferation of configurations explains why a significant proportion of the nine-atom molecules exceeded the time limit (3149 versus 276 of the six-atom path molecules). The distributions of the minimum and maximum distances for the six- and nine-atom paths are shown in Figure 5.

First we considered the effects of including the various descriptors on a subset of molecules from the six-atom path molecules. This subset comprised the 2381 molecules in which the two atoms at the ends of the path were both nitrogen. With the reduced descriptor sets that only used a few of the descriptors to code for a path it was frequently the case that a given pattern would occur more than once. For this reason, no duplicate patterns (i.e., the same set of descriptors and the same minimum and/or maximum distance) were permitted in the data sets used to train and test the networks. The first two networks used as descriptors just the bond order and atom hybridization and period information. A total of 44 descriptors were used for each path ($\equiv 5 \times 4 + 6 \times 4$).

Two $44 \times 10 \times 1$ networks were used to predict the maximum and minimum separations. The 1604 unique

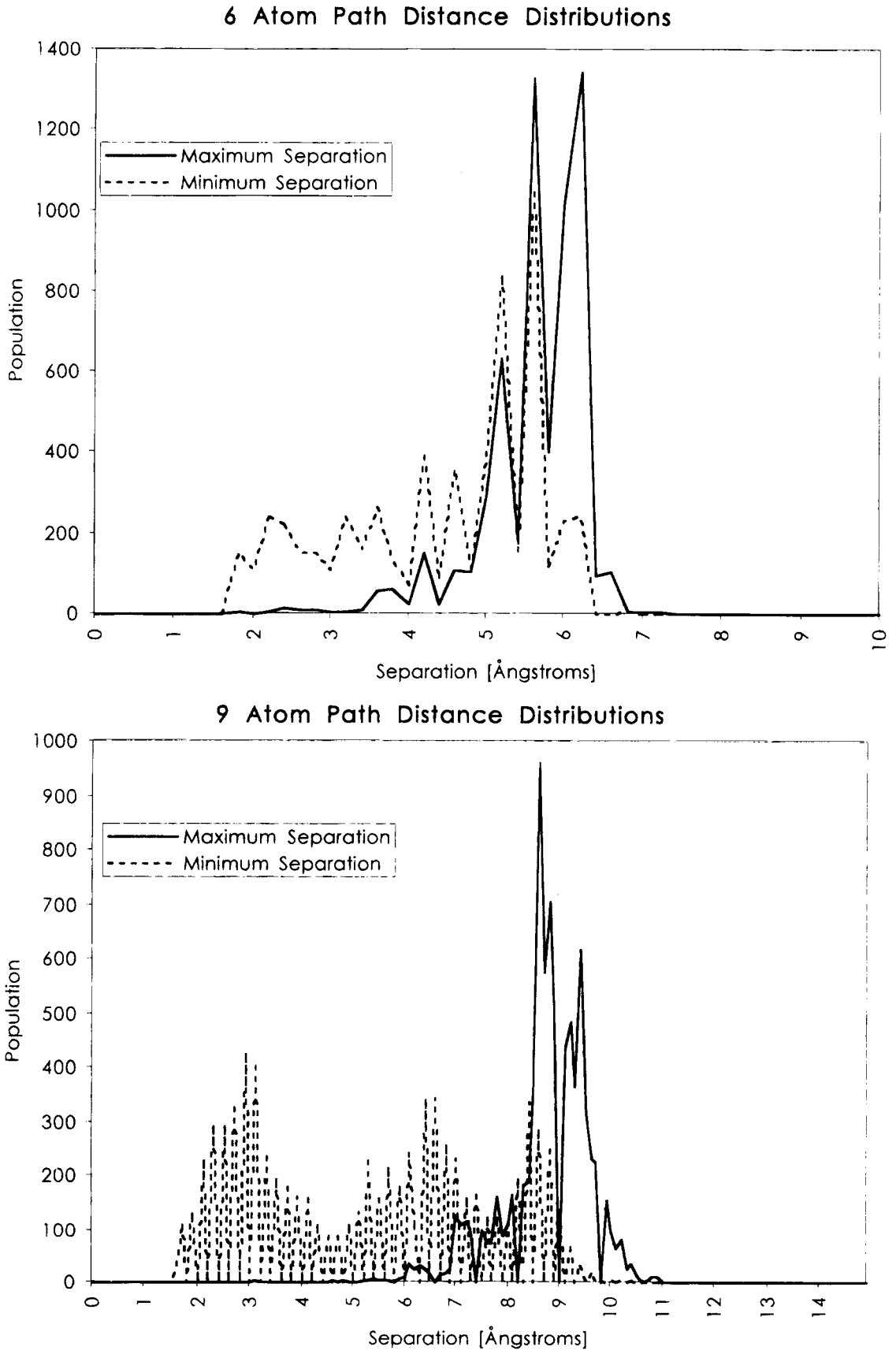

Figure 5. Distribution of distances for the upper and lower path separations for the six- (top) and nine- (bottom) atom paths.

patterns were divided into a training set of 1185 patterns and 419 testing patterns. The upper bound network was trained for 900 epochs (where an epoch corresponds to the presentation of each pattern in the training set to the network once); the lower bound network required 1000 epochs to reach its optimal solution. The ability of the networks to predict the upper and lower distances in the testing set are reported in Table 2. The networks are able to correctly predict the upper bound separation in 63% of the molecules in the testing set and 43% of the molecules for the lower

Table 2. Results for Building the Model

error tolerance (Å)	upper bound				lower bound			
	random no. simulation (%)	basic descriptors (%)	substit descriptors (%)	sterochem descriptors (%)	random no. simulation (%)	basic descriptors (%)	substit descriptors (%)	sterochem descriptors (%)
0.00	0	0	0	0	0	0	0	0
0.10	11	16	15	19	5	10	23	27
0.20	21	27	31	41	10	18	31	40
0.30	30	41	51	61	14	27	40	49
0.40	41	52	63	71	19	35	47	59
0.50	49	63	73	80	23	43	55	66
0.60	57	71	78	85	26	48	60	70
0.70	63	79	83	88	30	56	64	75
0.80	69	84	87	91	35	61	70	79
0.90	74	88	90	92	40	65	74	82

Table 3. Results for the Full Atom Paths

error tolerance (Å)	6 atom paths				9 atom paths			
	upper neural net (%)	upper random no. (%)	lower neural net (%)	lower random no. (%)	upper neural net (%)	upper random no. (%)	lower neural net (%)	lower random no. (%)
0.00	0	0	0	0	0	0	0	0
0.10	19	11	11	5	20	11	11	2
0.20	36	21	22	10	37	22	20	4
0.30	50	30	32	14	51	31	30	7
0.40	62	41	41	19	65	40	38	9
0.50	75	49	50	23	74	48	47	11
0.60	81	57	56	26	80	56	54	13
0.70	86	63	61	30	84	62	61	15
0.80	89	69	66	35	88	68	66	17
0.90	93	74	70	40	90	73	71	20

bound separation, within a tolerance of 0.5 Å. This compares with prediction accuracies of 49% and 23%, respectively, by randomly selecting distances from the observed distribution in the training set. There were no systematic biases in the prediction of either upper or lower bound distances. However, it is still possible for the neural network to predict values that are outside the triangle smoothed limits for that particular case. However, none of the predicted upper distance values fall outside the maximum value in the entire set nor are any of the predicted lower distances less than the minimum value in the entire set.

As would be expected, a description based solely upon bond order and atom type cannot account for all the variation in the data; simple inspection reveals many cases where the atom types and bond orders are identical but where the minimum and maximum distances are significantly different. The substitution descriptors were thus incorporated. The simplified topological states values (S) were employed, together with the S3 and S6 cutoff descriptors. This gave nine descriptors per atom and four descriptors per bond in the path, a total of 74 input nodes. At this level of description, a total of 2615 unique patterns were obtained from the data; these were divided into a training set of 1952 and a testing set of 663. Two 74 × 10 × 1 networks were used; to train the upper bound network required 400 epochs whilst to train the lower bound network required 4700 epochs. The performance of these networks was approximately 10% better than the previous network (Table 2).

Finally, the binary descriptor to describe the presence of cis or trans double bonds and two binary descriptors to characterize cis or trans ring systems were included. The number of descriptors was now 78. A set of 1925 patterns was chosen to train the network (1900 epochs for the upper bound and 2100 epochs for the lower bound) with 690 patterns being used for testing. Using networks with an 78 × 8 × 1 architecture, it was found that the performance had improved by approximately 7% for the upper bound problem and 10% for the lower bound problem (Table 2).

Having demonstrated that all the descriptors made a significant contribution to the performance of the networks, we next considered the full datasets for the six and nine atom paths. For the six atom path, the 5716 paths were divided into a training set of 3812, a testing set of 1474 and a cross validating set of 430. A 78 × 25 × 1 network was employed. Training required 2000 epochs for the upper bound and 1300 epochs for the lower bound. For the nine atom paths, the 8004 data patterns were divided into a training set of 5470, a testing set of 2062 and a cross validating set of 572. A 113 × 25 × 1 architecture was employed. The results are given in Table 3 and illustrated graphically in Figure 6. Our networks provided a prediction accuracy (at the 0.5 Å tolerance level) of 75% and 50% for the minimum and maximum six-atom path problem, respectively, and 74% and 47% for the maximum and minimum nine-atom path problem, respectively. By contrast, the corresponding figures for the random simulations, where distances were predicted using a random number generator which was weighted according to the distribution of the distances in the actual data sets, were 49%, 23%, 48%, and 11%. The neural networks thus achieved a prediction accuracy between 25% and 36% better than the random method.

DISCUSSION AND CONCLUSIONS

Over 1000 papers have now been published describing the application of neural networks to chemical problems. Our neural networks provide a link between the two-dimensional representation of the molecule and the three-dimensional

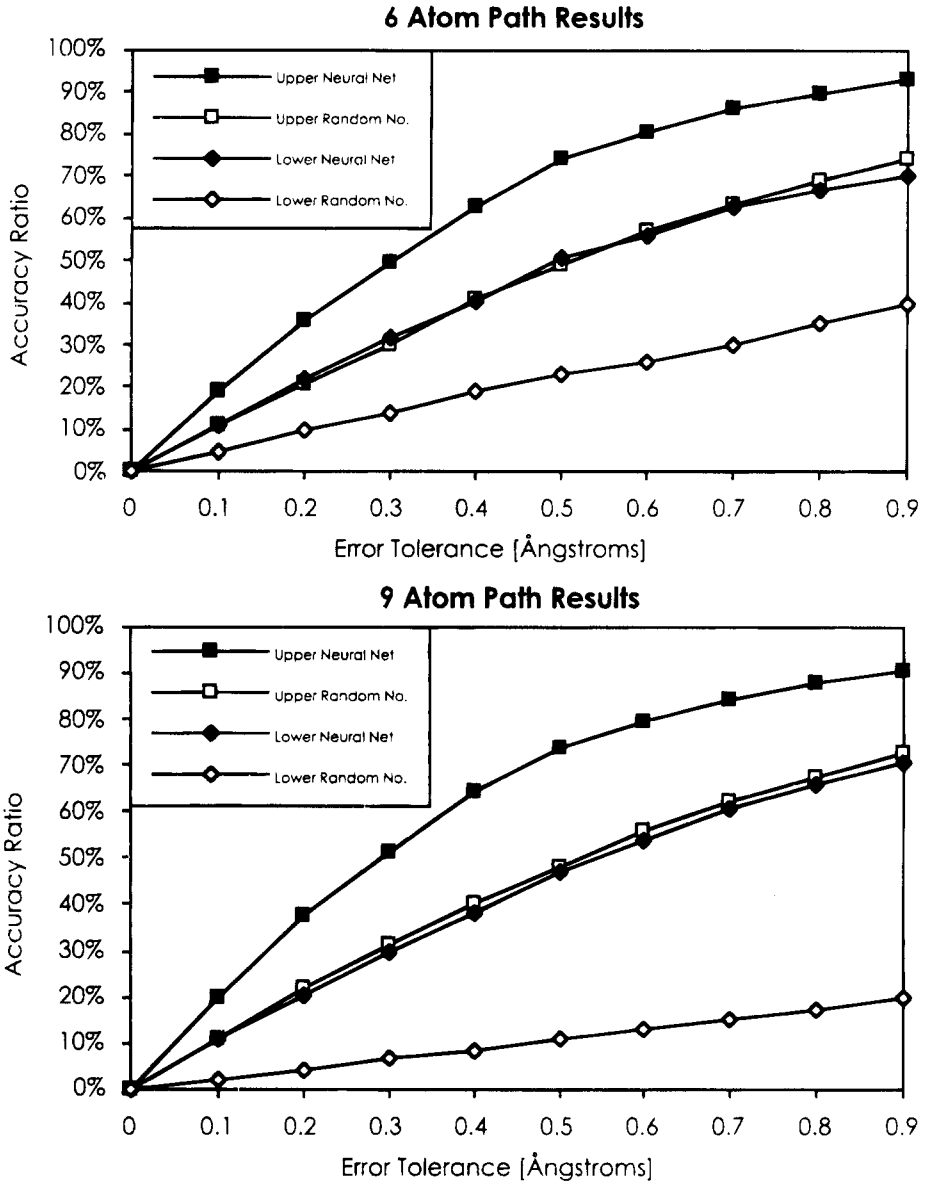

Figure 6. Results for the neural network simulations. The y axis gives the proportion of the distances that were correct to within a given tolerance (x axis).

conformations that it adopts. It is important to recognize the limitations of our work to date. The most obvious shortcoming is that an accuracy of 74%, let alone 47%, is not sufficiently reliable for use in most database screening applications, where it is usually desired to identify "all" possible hits. Nevertheless, in comparison to the random simulations the neural networks do show significantly enhanced performance, comparable to that achieved in other neural network applications such as the prediction of protein secondary structure. Moreover, it should be also recognized that the elements of other 3D database searching systems are subject to potential inaccuracies; for example, a random conformational search procedure such as distance geometry is not guaranteed to locate any particular conformation. A systematic search may miss conformations due to the selection of too large a torsional increment.

Clearly, the performance of a network can only be so good as the quality of the data used to train and test it and the possible presence of inaccurate data cannot be discounted. Various checks were performed to ensure that the data were as accurate as possible. For example, molecules that had unusual minimum or maximum distances were visually checked. The minimum and maximum distances were also compared with the values obtained from triangle smoothing and any significant deviations (such as occur for the 4-acetamidobenzoic acid) noted and investigated as we have described previously.[30]

Our work has shown that a considerable amount of information can be inferred about the conformational space of a molecule from the connection table alone. Unfortunately, no systematic study has been performed to address this question, and it is almost certainly molecule-dependent. Thus, the conformational properties of completely rigid systems can in general be predicted well, but this accuracy decreases as the molecule becomes more flexible. We have made a number of assumptions that may be the source of further errors. We have assumed that the shortest path is critical in determining the range of distances between a pair of atoms. The descriptors used to characterize the shortest path may be incomplete. We have demonstrated above how the incorporation of additional information about the path improves the network performance. It is of course entirely feasible that an alternative way to represent such information might produce an improvement in performance, or that

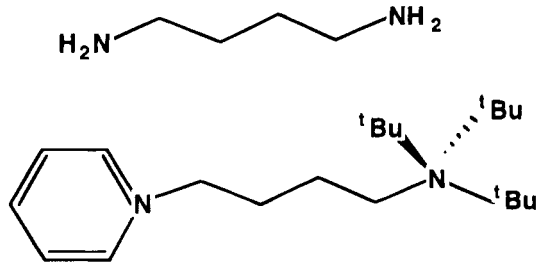

Figure 7. Effect of terminal substitution on the minimum separation; it might be anticipated that the presence of a bulky substituent would prevent an atom from approaching as close to another atom when compared to an unsubstituted analogue. Thus the minimum N−N distance in the top molecule would be expected to be smaller than for the bottom molecule.

information not present in our descriptors might also be important in characterizing the distance ranges. The problem of memorization is an ever-present one when developing neural network models; this limits the size of network that can be satisfactorily represented. Different descriptors may be required for the minimum and maximum distance problems. For example, we considered that the consistently poorer performance for predicting the minimum distance might be due to steric interactions between parts of the molecule that lie off the shortest path, as illustrated in Figure 7. To investigate this possibility, we trained and tested a network to predict the minimum distance that included descriptors that indicated the degree of steric bulk on the two end groups of the path. We added two binary descriptors to flag the presence of a ring at the head or tail of the path, and two continuous descriptors to indicate the number of substituents at the head or tail atom. These changes resulted in an increase in accuracy of between 1 and 3% for the prediction of the minimum separation. Our observation that the minimum distance is more difficult to predict than the upper distance has also been noted by other workers, such as Bradshaw and Maliski;[24] we have also noticed that the lower bounds produced by triangle smoothing deviate more from the actual distances obtained from a conformational search than the upper bounds.[30] Our networks required significantly more training for the minimum distance problems than for the corresponding maximum distances, indicating difficulties in finding a consistent model.

The performance of the networks was, in general, significantly better for acyclic paths than for paths that contain one or more ring. Thus, for the six-atom acyclic paths the network performances were 73% for the minimum distances and 95% for the maximum distances. The corresponding figures were 65% and 78% for paths that contained at least one ring. Rings impose restrictions on the conformational space that can be accessed by a molecule, and whilst we anticipate that the descriptors that we have employed do provide implicit information to the network about the rings present, it is possible that this does not encode for more subtle conformational effects. It is also possible that arbitrarily choosing the first "shortest path" between two atoms influences the performance for cyclic compounds. Again, we emphasize that, in theory at least, it should be possible to enhance the performance of the network by increasing the number of descriptors, but one must be wary of the dangers of the network simply memorizing the data rather than "learning".

ACKNOWLEDGMENT

A.R.L. thanks the SERC for an Advanced Fellowship and for providing computing equipment. S.N.J. thanks the University of Southampton and Glaxo Research and Development for a research studentship and financial support. We would like to thank Dr. D.V.S. Green for assistance with the conformational search and Dr. M. M. Hann for his continued interest in the project.

REFERENCES AND NOTES

(1) Lewis, R. A.; Leach, A. R. Current Methods for Site-Directed Structure Generation. *J. Comp.-Aided Mol. Des.* **1994**, *8*, 467−475.
(2) Martin, Y. C. 3D Database Searching in Drug Design. *J. Med. Chem.* **1992**, *35*, 2145−2154.
(3) Kuntz, I. D. Structure-Based Strategies For Drug Design And Discovery. Science **1992**, *257*, 1078−1082.
(4) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3-D Databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312−316.
(5) Dolata, D. P.; Leach, A. R.; Prout, K. WIZARD: AI in Conformational Analysis. *J. Computer-Aided Mol. Des.* **1987**, *1*, 73−85.
(6) Catalyst. BioCAD/Molecular Simulations Inc.
(7) Tripos Associates, St. Louis, MO.
(8) Hurst, T. Flexible 3D Searching: the Directed Tweak Algorithm. *J. Chem. Inf. Comp. Sci.* **1994**, *34*, 190−195.
(9) Fine, R. M.; Wang, H.; Shenkin, P. S.; Yarmush, D. L.; Levinthal, C. Predicting Antibody Hypervariable Loop Conformations II: Minimization and Molecular Dynamics Studies of MCPC603 From Many Randomly Generated Loop Conformations Proteins. *Structure, Function Genetics* **1986**, *1*, 342−362.
(10) Shenkin, P. S.; Yarmush, D. L.; Fine, R. M.; Wang, H.; Levinthal, C. Predicting Antibody Hypervariable Loop Conformation. I. Ensembles of Random Conformations for Ringlike Structures. *Biopolymers* **1987**, *26*, 2053−2085.
(11) MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA 94577.
(12) Clark, D. E.; Willet, P.; Kenny, P. W. Pharmacophoric Pattern-Matching In Files Of 3-Dimensional Chemical Structures−Use Of Bounded Distance Matrices For The Representation And Searching Of Conformationally Flexible Molecules. *J. Mol. Graph.* **1992**, *10*, 194−204. Clark, D. E.; Willet, P.; Kenny, P. W. Pharmacophoric Pattern-Matching In Files Of 3-Dimensional Chemical Structures−Implementation Of Flexible Searching. *J. Mol. Graph.* **1993**, *11*, 146−156.
(13) Crippen, G. M. Distance Geometry and Conformational Calculations. Chemometrics Research Studies Series 1; Wiley: New York, 1981.
(14) Crippen, G. M.; Havel, T. F. Distance Geometry and Molecular Conformation. Chemometrics Research Studies Series 15; Wiley: New York, 1988.
(15) Bearle, R.; Jackson, T. Neural Computing; Institute of Physics Publishing: Bristol, 1990.
(16) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503 −527.
(17) Qian, N.; Sejnowski, T. J. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **1990**, *202*, 865−884.
(18) Mayer, B.; Hansen, T.; Nute, D.; Albersheim, P.; Darvill, A.; York, W.; Sellers, J. Identification of the 1H-NMR spectra of complex oligosaccharides with artificial neural networks. *Science* **1991**, *251*, 542−544.
(19) Aoyama, T.; Suzuki Y.; Ichikawa, H. Neural networks applied to quantitative structure-activity relationship analysis. *J. Med Chem.* **1990**, *33*, 2583−2590.
(20) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning internal representations by error propagation. In *Parallel Distributed Processing;* Rumelhart, D. E., McClelland J. L., Eds.; MIT Press: Cambridge, 1986; Vol. 1, pp 322−328.
(21) Maggiora, G. M.; Elrod, D. W. Computational neural networks as model-free mapping devices. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 732−741.
(22) Rouvray, D. H. Making the right connection. *Chem. Brit.* June **1993**, 495−498.
(23) So, S.; Richards, W. G. Applications of neural networks: QSAR of derivatives of 2,4-diamino-5-(substituted-benzyl)pyramidines as DHFR inhibitors. *J. Med. Chem.* **1992**, *35* 3201−3207.
(24) Bradshaw, J.; Maliski, E. G. Use of the most restrictive paths in 3D search strategy. Paper presented at the 4th Chemical Congress of North America, New York, 25th−30th August, 1991.

(25) Hall, L. H.; Kier, L. B. Determination of topological equivalence in molecular graphs from the topological state. *Quant. Struct.-Act. Relat.* **1990**, *9*, 115−131.

(26) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; MacRae, C. F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. Integrated 3D search facilities for the Cambridge Structural Database (CSD). *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187−204.

(27) The ACD is available from MDL Information Systems Inc., 14600 Catalina St., San Leandro, CA 94577.

(28) Leach, A. R.; Prout, K. Automated conformational analysis: directed conformational search using the A* algorithm. *J. Chem. Inf. Comput. Sci.* **1990**, *11*, 1193−1205.

(29) Rusinko, III, A.; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. CONCORD: a program for the rapid generation of high quality 3D molecular structures. The University of Texas at Austin and Tripos Associates: St. Louis, MO, 1988.

(30) Leach, A. R. An algorithm to identify a molecule's "most different" conformations. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 661−670.

(31) Howard, A. E.; Kollman, P. A. An Analysis of Current Methodologies for Conformational Search of Complex Molecules. *J. Med. Chem.* **1988**, *31*, 1669−1675.

(32) Leach, A. R. A Survey of Methods for exploring the conformational space of small and medium sized molecules. *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: **1991**, 1−54.

(33) Leach, A. R.; Prout, C. K.; Dolata, D. P. The Application of Artificial Intelligence to the Conformational Analysis of Strained Molecules. *J. Comput. Chem.* **1990**, *11*, 680−694.

(34) Leach, A. R.; Smellie, A. S. A Combined Model-Building and Distance Geometry approach to Automated Conformational Analysis. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 379−385.

CI940240W