Atomistic Topological Indices Applied to Benzodiazepines using Various Regression Methods

David A. Winkler¹, Frank R. Burden²* and Andrew J. R. Watkins²

¹CSIRO Division of Molecular Science, Private Bag 10, Clayton South MDC, Clayton, Victoria 3169, Australia.
²Chemistry Department, Monash University, Clayton, Victoria 3168, Australia.

Abstract

The use of atomic positionally independent molecular representations of a set of benzodiazepines is shown to yield a comparable structure-activity model than the traditional physicochemical representations. The best model was produced using an artificial neural network; the results are compared with models produced by MLR, PCR and PLS.

The atomistic representation compares favourably with a functional group representation and the idea of implicit information in such representations is discussed. It appears that the complex interrelationships between elements of a simple molecular representation may be capable of encoding subtle molecular structural elements.

Key words: Benzodiazepines, neural network, principal component regression, partial least squares, topological indices

1 Introduction

The field of Quantitative Structure Activity Relationships (QSAR) is now well established [1] and this work has demonstrated that the biological activity of a set of molecules acting via the same mode of action can be mathematically related to some simple physicochemical properties or molecular structure parameters [2]. These methods have been widely adopted in the pharmaceutical and agrochemical industries. The relationships between biological activity and physicochemical parameters have normally been expressed by power series expansion with the physicochemical parameters as the independent variables [3]. In practice, linear and quadratic terms are mainly used without the use of cross terms.

There is a need to be able to predict the biological activity by more ab initio means so that the need to synthesize new candidate molecules is reduced or removed. Such ab initio methods range from the prediction of molecular properties by use of large basis set molecular orbital methods to simply using molecular formulae and valency rules. Since the interactions of small molecules with biological systems is extremely complex it is clear that even a detailed understanding of the electronic structure of a molecule is not sufficient, and additional knowledge of the molecular site where the interaction is usually desirable. Historically, one of the most successful ways of discovering more active lead molecules is to begin with an active molecule then chemically modify it in a carefully controlled manner.

At the most basic level the biological activity of a molecule depends on its steric, electronic and lipophilic properties. As the molecular sites of action are usually receptors, or enzyme active sites with often complex three-dimensional shapes, it is clear why these properties are so important for recognition, binding and effect. The size of a molecule is determined by the number of atoms or, at a slightly more refined level, the empirical formula, which also gives the elemental composition. Similarly the polarity of the molecule is, in part, determined by the nature and polarity of its functional groups. The number of heteroatoms in a molecule give some indication of molecular polarity. Rules of valency often allow such simple representations to imply higher level representations. For example, in a molecule containing several carbonyl groups it may only be necessary to specify the number of oxygen atoms which are attached to the substituent by a single connection, that is a double-bonded oxygen. Brown and Martin [4] used concepts drawn from database fingerprinting theory to generalize the molecular fragment descriptors concept. They compared structural keys generated by the MACCS database, Daylight and Unity 2D hashed fingerprints and several other fragment generation methods for their ability to separate active from inactive compounds. The Unity hashed fingerprint concept [5] was recently modified to create molecular holograms [6] as efficient representations. In a previous paper [7] an atomistic representation was developed and applied to a set of dihydrofolate reductase inhibitors. Here we apply atomistic and functional group descriptors to a set of benzodiazepines.

Benzodiazepines have been used therapeutically as anxiolytics, as tranquillisers, and as anticonvulsants in epilepsy. They act via the benzodiazepine site (BzR) on the γ-aminobutyric acid receptor (GABA<sub>A</sub>) family, and have been subject to extensive research, with over twenty QSAR studies having been carried out e.g. [8–10]. Many types of compound have been shown to bind at the BzR e.g., benzodiazepines, arylpyrazolo-quinolines, β-carbolines, imidazo-pyridazines, cyclo-pyrrrolones. Structure-activity relationships in Bz receptor ligands have been reviewed recently [11–12].

We have taken a well characterized set of 57 benzodiazepines which bind to the GABA<sub>A</sub> receptor and used these to investigate three features:

Abbreviations

ANN, Artificial Neural Network; QSAR; MLR, Multiple Linear Regression; PCR, Principal Component Regression; PLS, Partial Least Squares.

* To receive all correspondence
the relative importance of atomistic versus functional group-based representations of structures
the influence of neural network architecture on the “predictiveness” of the QSAR relationship as assessed by cross-validation procedures.
the efficacy of artificial neural networks compared with other QSAR paradigms such as multiple linear regression, partial least squares etc., in deriving quantitative structure-activity relationships.

2 Materials and Methods

2.1 Biological Data

The data set used for studying the application of the simple atomistic and functional group-based representations was a set of fifty-seven 1,4-benzodiazepin-2-ones (I) [13] used in a study by Maddelena and Johnston [14].

The molecular representation used in that work involved traditional physicochemical substituent constants, and used a neural network to map structure representations to biological activity. The results in this paper will be compared with the results from Maddelena and Johnston. The structures, biological data and representations are summarized in Table 1.

Table 1. Structures and biological data [11].

<table>
<thead>
<tr>
<th>Name</th>
<th>R7</th>
<th>R1</th>
<th>R2'</th>
<th>R6'</th>
<th>R3'</th>
<th>R8</th>
<th>log IC₅₀a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ro 05-3061</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.602</td>
</tr>
<tr>
<td>Ro 05-4865</td>
<td>F</td>
<td>Me</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.230</td>
</tr>
<tr>
<td>Ro 05-6820</td>
<td>F</td>
<td>H</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>0.869</td>
</tr>
<tr>
<td>Ro 05-6822</td>
<td>F</td>
<td>Me</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>0.708</td>
</tr>
<tr>
<td>nitrazepam</td>
<td>NO₂</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.000</td>
</tr>
<tr>
<td>Ro 05-4435</td>
<td>NO₂</td>
<td>H</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>0.176</td>
</tr>
<tr>
<td>fentanyl</td>
<td>NO₂</td>
<td>Me</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>0.580</td>
</tr>
<tr>
<td>Ro 05-4082</td>
<td>NO₂</td>
<td>Me</td>
<td>Cl</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>0.342</td>
</tr>
<tr>
<td>Ro 05-3590</td>
<td>NO₂</td>
<td>H</td>
<td>CF₁</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>0.544</td>
</tr>
<tr>
<td>Ro 20-7736</td>
<td>NHOH</td>
<td>Me</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td></td>
<td>1.982</td>
</tr>
<tr>
<td>Ro 05-3072</td>
<td>NH₃</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td></td>
<td>2.587</td>
</tr>
<tr>
<td>Ro 05-3418</td>
<td>NH₃</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td></td>
<td>2.663</td>
</tr>
<tr>
<td>Ro 20-1815</td>
<td>NH₂</td>
<td>Me</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.813</td>
</tr>
<tr>
<td>Ro 05-4619</td>
<td>NH₂</td>
<td>H</td>
<td>Cl</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.875</td>
</tr>
<tr>
<td>Ro 05-4528</td>
<td>CN</td>
<td>Me</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>2.580</td>
</tr>
<tr>
<td>Ro 20-2541</td>
<td>CN</td>
<td>Me</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.477</td>
</tr>
<tr>
<td>Ro 20-2533</td>
<td>Et</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.556</td>
</tr>
<tr>
<td>Ro 20-5747</td>
<td>CH₂=CH₂</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.380</td>
</tr>
<tr>
<td>Ro 20-5397</td>
<td>CHO</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.633</td>
</tr>
<tr>
<td>Ro 20-3053</td>
<td>COMe</td>
<td>H</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.255</td>
</tr>
<tr>
<td>Ro 05-2921</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>2.544</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>R7</th>
<th>R1</th>
<th>R2'</th>
<th>R6'</th>
<th>R3'</th>
<th>R8</th>
<th>log IC₅₀a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ro 05-4336</td>
<td>H</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.322</td>
</tr>
<tr>
<td>Ro 05-4520</td>
<td>H</td>
<td>Me</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.146</td>
</tr>
<tr>
<td>Ro 05-4608</td>
<td>H</td>
<td>Me</td>
<td>Cl</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>0.580</td>
</tr>
<tr>
<td>halazepam</td>
<td>Cl</td>
<td>CH₂CF₃</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.964</td>
</tr>
<tr>
<td>Ro 06-9098</td>
<td>NO₂</td>
<td>CH₂OCH₃</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>2.633</td>
</tr>
<tr>
<td>Ro 20-1310</td>
<td>Cl</td>
<td>C(CH₃)₃</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>2.792</td>
</tr>
<tr>
<td>Ro 07-2750</td>
<td>Cl</td>
<td>(CH₃)₂OH</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.389</td>
</tr>
<tr>
<td>Ro 22-4683</td>
<td>NO₂</td>
<td>C(CH₃)₃</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>2.477</td>
</tr>
<tr>
<td>Ro 07-4419</td>
<td>H</td>
<td>F</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td></td>
<td>1.279</td>
</tr>
<tr>
<td>Ro 07-3953</td>
<td>Cl</td>
<td>H</td>
<td>F</td>
<td>F</td>
<td>H</td>
<td></td>
<td>0.204</td>
</tr>
<tr>
<td>Ro 07-4065</td>
<td>Cl</td>
<td>Me</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td></td>
<td>0.613</td>
</tr>
<tr>
<td>Ro 07-5193</td>
<td>Cl</td>
<td>H</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td></td>
<td>0.477</td>
</tr>
<tr>
<td>Ro 22-3294</td>
<td>Cl</td>
<td>H</td>
<td>Cl</td>
<td>H</td>
<td>H</td>
<td></td>
<td>0.845</td>
</tr>
<tr>
<td>Ro 07-5220</td>
<td>Cl</td>
<td>Me</td>
<td>Cl</td>
<td>H</td>
<td>H</td>
<td></td>
<td>0.740</td>
</tr>
<tr>
<td>Ro 13-3780</td>
<td>Br</td>
<td>Me</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td></td>
<td>0.380</td>
</tr>
<tr>
<td>Ro 11-4878</td>
<td>Cl</td>
<td>H</td>
<td>F</td>
<td>H</td>
<td>Me</td>
<td></td>
<td>0.544</td>
</tr>
<tr>
<td>meclonazepam</td>
<td>NO₂</td>
<td>H</td>
<td>Cl</td>
<td>H</td>
<td>Me</td>
<td>H</td>
<td>0.079</td>
</tr>
<tr>
<td>Ro 11-6896</td>
<td>NO₂</td>
<td>Me</td>
<td>F</td>
<td>H</td>
<td>Me</td>
<td>H</td>
<td>0.845</td>
</tr>
<tr>
<td>Ro 06-7263</td>
<td>Cl</td>
<td>H</td>
<td>H</td>
<td>Me</td>
<td>H</td>
<td></td>
<td>1.690</td>
</tr>
<tr>
<td>Ro 06-7263</td>
<td>Cl</td>
<td>H</td>
<td>H</td>
<td>OH</td>
<td>H</td>
<td></td>
<td>1.255</td>
</tr>
<tr>
<td>Ro 05-4619</td>
<td>NH₂</td>
<td>H</td>
<td>Cl</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.875</td>
</tr>
<tr>
<td>Ro 05-4528</td>
<td>CN</td>
<td>Me</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>2.580</td>
</tr>
<tr>
<td>Ro 20-2541</td>
<td>CN</td>
<td>Me</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.477</td>
</tr>
<tr>
<td>Ro 20-2533</td>
<td>Et</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.556</td>
</tr>
<tr>
<td>Ro 20-5747</td>
<td>CH₂=CH₂</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.380</td>
</tr>
<tr>
<td>Ro 20-5397</td>
<td>CHO</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.633</td>
</tr>
<tr>
<td>Ro 20-3053</td>
<td>COMe</td>
<td>H</td>
<td>F</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>1.255</td>
</tr>
<tr>
<td>Ro 20-2921</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>2.544</td>
</tr>
</tbody>
</table>

*IC₅₀ in nmol L⁻¹.
2.2 Molecular Descriptors

In order to carry out a QSAR study, it is necessary to convert the molecular structure and other properties of molecules into a mathematical representation. Representations can range from empirical formulae, chemical graphs, traditional π, MR parameters, to molecular orbital descriptions using large basis sets and electron correlation. However, for most chemical purposes the structural formula, sometimes including stereochemistry, is sufficient for chemists to predict a wide range of properties and to use it as a basis for planning the synthesis of other molecules. From structural formulae can be derived a variety of molecular descriptors for using in prediction of chemical reactivity, physical properties etc. [15].

We adopted a very simple atomistic representation of molecules [7] which superficially appears to disregard much information such as topology and stereochemistry. Each atom is classified by its element and number of connections such as H1, for hydrogen (since all hydrogens have only one connection), or with carbon: C4 (sp³), four connections; C3 (sp²), three connections; C2 (sp), two connections. The number of each type of atom in the molecule become the numerical values of each independent variable [7].

The other simple representation comes out of this work done by Andrews et al. [16] on deducing the relative contributions of functional groups in molecules to receptor binding. There is some overlap with the atomistic representation, but the focus of the work is the decomposition of molecules into functional groups capable of interacting with receptors. As these functional groups were derived from 200 drug molecules of diverse structure binding to a diverse range of receptors, they are likely to be capable of general application. The functional groups found to be statistically significant in Andrews’ analyses were: CO₂-, PO₄³⁻, N⁺, N, OH, C=O, O/S ethers, halogens, C(sp³), C(sp²), and an entropic term related to the number of freely rotatable bonds in the molecule.

2.3 Regression Methods

The relationship between the molecular structure and the biological activity was modelled using Multiple Linear Regression, MLR, Partial Least Squares, PLS[r] [17], and an artificial neural network, ANN [18]. The use of ANN’s in this context has proved to be better than either of MLR and PLS[r].

All calculations were strictly cross-validated so that in each regression the training set data was scaled and the scaling of this set applied to the validation set. This ensures that the validation set is truly independent of the training set. In the results reported for the partial least squares, PLS[r], are for the number of latent variables that gave the minimum standard error of prediction.

The ANN’s used were three layer fully connected, feed forward networks which were trained by the use of back propagation. Other larger or multi-layered networks did not show any improvement over the three-layered networks. The work was carried out on a Pentium® based desktop computer using MATLAB [18] programs written by the authors. The following neural network parameters were used:

- Momentum rate = 0.5
- Learning rate = 1.0
- Network connectivity = full
- Transfer functions: input layer: linear, other layers: sigmoidal.

The data was mean centered and then scaled between 0 and 1.

\[ y_i(\text{scaled}) = (y_i - y_{\text{min}})/(y_{\text{max}} - y_{\text{min}}) \]

The data order was randomised thus ensuring that when the data set was split into validation and training sets, the validation set would be truly representative of the training set and not feature a disproportionate number of a particular compound-type. The data set was automatically split into a series of training and validation sets, for use in cross-validation. The validation sets consisted of 5 compounds taken in turn from the 50 to 55 compounds in the data set being considered. This gave 10 or 11 neural networks to be trained and the SEP results, reported in Table 2, represent the merging of the SEPs calculated for each network. The Standard Error of Prediction is given, for a single training run, \( k \), as

\[ \text{SEP}_k = \frac{1}{N} \left( \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \right) \]

and these are combined for the \( M = 10 \) or 11 runs by

\[ \text{SEP}_{\text{total}} = \sqrt{\frac{1}{M} \sum_{j=1}^{M} (\text{SEP}_j)^2} \]

The data presented in Table 2 has been scaled between 0 and 1 and has a mean of 0.4208 and a standard deviation of 0.2426. (The unscaled log IC₅₀ data has a mean of 1.2120 and a standard deviation of 0.7498, a maximum of 2.792 and a minimum of 0.079.)

It was necessary to remove some compounds from the full set of 57 compounds for each representation. This comes about because it happens that a particular atomistic or functional group is only represented by a single compound. Under these circumstances it is not possible for the neural network to learn a rule with regard to that group. For instance, in the case of the positionally representation AT1, of Table 2, halazepam is the only compound with a–CH₂CF₃ substituent at position R1.

Care was taken to keep the network sufficiently small, in terms of the number of weights to be computed, that overfitting was unlikely to occur. We computed the ratio,

\[ \rho = \frac{\text{number of input samples}}{\text{number of ANN weights}} \]

for all networks, and observed the relationship between \( \rho \) and the optimum cross-validated architectures. Manallack and Livingstone [19] suggest that \( \rho \) should lie in the range 1.8–2.2 for a set of dihydrololate reductase inhibitors. In the present case we have used networks with \( \rho \) values in the range 0.88 to 2.12. When \( \rho < 1 \) there is the strong possibility of overfitting, though this is not necessarily so when there is dependence between weights, as may occur with neural networks. Since all of the methods were strictly cross-validated, the SEPs may be taken as a reliable measure of the goodness of fit.
3 Results

The standard error of predictions (SEPs) and correlation coefficients, using the various representations, are shown in Table 2. Numerous ANN architectures were tested; the network with the lowest cross-validated SEP was deemed to be the optimal architecture. It is the SEPs of these optimal architectures that are shown in Table 2. A typical plot of calculated vs observed activities is shown in Figure 1. As the Table 2 shows SEPs obtained using ANNs are generally significantly lower than those obtained using the linear techniques.

As a test of the possible occurrence of chance effects, the activity data from representation γ AT4 was randomised, giving a SEP for the neural network of 0.466 as compared with 0.117 for the unscrambled data. This shows that chance effects are unlikely to be important for these data sets and calculations.

Rather surprisingly, the model does not suffer when positional information is removed from the representation (i.e. the position of substituents is ignored). Indeed, the best model using the atomistic approach (representation AT4) was positionally-independent. An indication of the quality of the QSAR derived using another positionally independent representation (AT3) is given in Table 2.

Two other notable features emerge from these results—the model obtained using the atomistic representation provides a SEP comparable to the model using the functional group representation. Removal of variables from AT4 and FG1 found not to be statistically significant in multiple linear regression does not greatly increase the resulting SEP.

4 Discussion

4.1 Positional Dependence vs Positional Independence

As Table 2 shows, the model does not suffer when positional information is removed from the representation. This result is counter-intuitive, as the position of various substituents is often critical to biological activity. We can rationalize by assuming that position is often implied in the representation. For example, the only two positions to feature alkyl substituents were R1 and R7.

Table 2. Cross-validated Standard Errors of Prediction, SEP and Correlation Coefficients, R, of the various regressions: Data scaled 0 to 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>AT1</th>
<th>AT2</th>
<th>AT3</th>
<th>AT4</th>
<th>AT5</th>
<th>FG1</th>
<th>FG2</th>
<th>PCh</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEP</td>
<td>0.219</td>
<td>0.231</td>
<td>0.228</td>
<td>0.233</td>
<td>0.191</td>
<td>0.240</td>
<td>0.203</td>
<td>0.195</td>
</tr>
<tr>
<td>R</td>
<td>0.721</td>
<td>0.695</td>
<td>0.582</td>
<td>0.603</td>
<td>0.712</td>
<td>0.454</td>
<td>0.640</td>
<td>0.705</td>
</tr>
<tr>
<td><strong>PLSr:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LV</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>SEP</td>
<td>0.188</td>
<td>0.207</td>
<td>0.224</td>
<td>0.214</td>
<td>0.191</td>
<td>0.206</td>
<td>0.203</td>
<td>0.189</td>
</tr>
<tr>
<td>R</td>
<td>0.754</td>
<td>0.696</td>
<td>0.575</td>
<td>0.627</td>
<td>0.712</td>
<td>0.621</td>
<td>0.640</td>
<td>0.708</td>
</tr>
<tr>
<td><strong>ANN:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEP</td>
<td>0.062</td>
<td>0.088</td>
<td>0.116</td>
<td>0.117</td>
<td>0.127</td>
<td>0.143</td>
<td>0.126</td>
<td>0.124</td>
</tr>
<tr>
<td>R</td>
<td>0.975</td>
<td>0.945</td>
<td>0.908</td>
<td>0.903</td>
<td>0.883</td>
<td>0.840</td>
<td>0.877</td>
<td>0.881</td>
</tr>
<tr>
<td>ρ</td>
<td>0.61</td>
<td>0.52</td>
<td>0.61</td>
<td>0.88</td>
<td>1.16</td>
<td>1.62</td>
<td>1.61</td>
<td>1.02</td>
</tr>
</tbody>
</table>

h Positionally independent. Input parameters: π(R7), F(R7), MR(R1), Σ(R1), μ(R1), MR(R2’), F(R2’), MR(R6’), s_(p)(R3), s_(R8), N = 50. Compounds removed: Ro 07-9957, Ro 05-3590, Halazepam, Ro 13-3780, Ro 06-7263, Ro 20-7078, Ro 20-8895.

Figure 1. Sample output from an 11 : 6 : 1 neural network using the AT3 (positionally independent) representation as input.
Thus, input variables such as C4 must be atoms attached to either of these positions.

### 4.2 Explicit vs Implicit Information

The results raise the issue of implicit versus explicit information in a representation of molecular structure. The model obtained using the atomistic representation provides a SEP comparable to the model using the functional group representation. Superficially, it would appear that the atomistic representation doesn’t contain enough information to model biological activity. Intuitively it would seem to provide less information than the functional group representation. However, there is much information implied in the representation. Take, for example, a compound which contains a single C3 atom and an single O1 atom; this must imply a carbonyl group. Thus, the presence of the carbonyl group is implied in the atomistic representation.

Further evidence for this arises when the AT5 and FG2 representations are compared. These are the representations which contained only input variables which were found to be statistically significant using a multiple linear regression analysis. In each representation, four variables were found to be significant, and each four-variable set is very similar to the other. Both features C4, Hal, and N (N3 in the case of AT5), FG1 also features NO2, whilst AT5 featured N3 and O1 which is effectively a nitro group. That is, all the information in the functional group representation would appear to be implied in the atomistic representation.

### 4.3 Non-linearity of Response Surface

Even when the atomistic representation was reduced to four atom types C4, N3, O1 and halogen, from an MLR study, the SEP was found comparable to the values from the physicochemical representation. These four atom types contribute to the model of the activity in a non-linear and interactive fashion when artificial neural networks are used in the modelling process. A similar conclusion is reached when the atomistic representation is compared to a functional group representation. In the case of the functional group representation, the input parameters could be reduced to four; namely, C4, N, NO2 and halogen. The atomistic representation is the preferred method since it is simpler to encode for larger data sets and is potentially transferable to other systems.

This raises the question — how non-linear is the benzodiazepine data set? It is not possible to conclude from this analysis that the data set is largely linear. Indeed, the lower performance of linear techniques such as least squares regression compared with a non-linear technique such as an ANN suggests that significant non-linearity exists in the data set. It is possible that much of the non-linearity in the data set exists in the four variables that were not excluded. The observation that linear techniques are less successful in modelling the data set when integer input variables are used as opposed to continuous input variables suggests that there is considerable non-linearity in the response surface for these representations. These non-linearities may relate to higher order representations implicit in the data.

### 4.4 Complexity of Representations

While the representation we used is simple, it is still an intrinsically more “information rich” representation than other indices, such as that of Randic [23], where all molecular properties are compressed into a single value. It appears that the complex inter-relationships between elements of a simple molecular representation may be capable of encoding subtle molecular structural elements. While this paper was being refereed, a related molecular representation came to our notice. The holographic QSAR (HQSAR) method uses a modified version of the bit string fingerprint approach used in database searches. This generates molecular fragments within certain sizes (usually 4–7 atoms) and hashes them into bins to generate fingerprints of given length. Molecular holograms are generated by counting the number of times each bin is set rather than using a bit string containing either 0 or 1 in each bin as in database searching fingerprinting. In molecular holograms, as in our case with smaller fragments (typically 1–3 atoms), positional information is implied rather than explicit. This concept has been discussed recently by Kier and Hall [20–21]. Brown and Martin [4] also found considerable inherent or latent information relevant to the forces of ligand-receptor binding in their fragment-based structural descriptors. The compression of molecular information in molecular representations spans a very wide range, from the representation of a molecule by a single number (for example the Weiner index) to a full high level ab initio wavefunction. Our approach, and that of related methods such as molecular holograms, falls in the middle. Our intent is to find the minimal representation which accounts sufficiently for molecular properties to give useful and statistically valid QSARs.

### 4.5 Significance of $\rho$ Parameter

The ANN architectures had values of $\rho$ less than recommended by Andrea and Kalayeh [22] in the study of dihydrofolate reductase inhibitors showing that a general rule for $\rho$ is hard to define. The $\rho$ value of 2 is simply a “rule of thumb” which cannot replace a more rigorous cross-validation method in choosing an neural net architecture. Indeed, our work has shown that the best cross-validated architectures can occur at $\rho$ values considerably less than 2. A careful monitoring of the SEP’s for various validated methods is necessary when studying any new data set.

We have shown that the use of a positionally independent atomistic representation can provide a better SAR model for a set of fifty-seven benzodiazepines than using a set of traditional positionally dependent physicochemical parameters [14]. If this result can be generalized from the two examples studied so far, it has far-reaching implications. Any new candidate molecule can be simply screened by the use of the previously fitted model and a simple atomistic representation. As with the positionally dependent physicochemical parameter study, artificial neural networks were found to provide a superior fit than linear techniques such as MLR and PLS.

Finally it seems unlikely that the absence of positional information can be expected to give such good results as a general rule. Further work is being undertaken using topological indices [23–25], molecular multipole moments [26], as well as the chemically...
intuitive molecular index, CIMI, of Burden [27], used in conjunction with the atomistic representation.

5 References