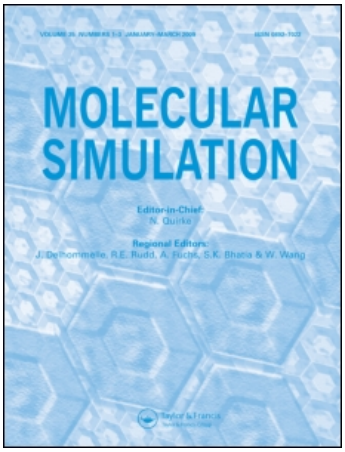


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Robust QSAR Models from Novel Descriptors and Bayesian Regularised Neural Networks

David A. Winkler[a]; Frank R. Burden[b]
[a] CSIRO Division of Molecular Science Private Bag 10, Clayton South MDC, Clayton, Victoria, Australia [b] Department of Chemistry, Monash University, Clayton, Australia

To cite this Article Winkler, David A. and Burden, Frank R.(2000) 'Robust QSAR Models from Novel Descriptors and Bayesian Regularised Neural Networks', Molecular Simulation, 24: 4, 243 — 258
To link to this Article: DOI: 10.1080/08927020008022374
URL: http://dx.doi.org/10.1080/08927020008022374

Molecular Simulation, 2000, Vol. 24, pp. 243–258
Reprints available directly from the publisher
Photocopying permitted by license only

ROBUST QSAR MODELS FROM NOVEL DESCRIPTORS AND BAYESIAN REGULARISED NEURAL NETWORKS

DAVID A. WINKLER[a,*] and FRANK R. BURDEN[b]

[a]*CSIRO Division of Molecular Science Private Bag 10, Clayton South MDC, Clayton, Victoria 3169, Australia;* [b]*Department of Chemistry, Monash University, Clayton 3168, Australia*

(Received April 1999; accepted May 1999)

The QSAR method, using multivariate statistics, was developed by Hansch and Fujita, and it has been successfully applied to many drug and agrochemical design problems. As well as speed and simplicity QSAR has advantages of being capable of accounting for some transport and metabolic processes which occur once the compound is administered.

Until recently QSAR analyses have used relatively simple molecular descriptors based on substituent constants (*e.g.*, Hammett constants, π, or molar refractivities), physicochemical properties (*e.g.*, partition coefficients), topological indices (*e.g.*, Randic and Weiner indices). Recently several new representations have been devised: atomistic; molecular eigenvalues and BCUT indices derived therefrom; E-state fields; topological autocorrelation vectors; various molecular fragment-based hash codes. These representations have advantages in speed of computation, in more accurately representing molecular properties most relevant to activity, or in being more generally applicable to diverse chemical classes acting at a common receptor, than traditional representations.

Historically, linear regression methods such as MLR (multiple linear regression) and PLS (partial least squares) have been used to develop QSAR models. Regression is an "ill-posed" problem in statistics, which sometimes results in QSAR models exhibiting instability when trained with noisy data. In addition traditional regression techniques often require subjective decisions to be made on the part of the investigator as to the likely non-linear relationship between structure and activity, and whether there are cross-terms. Regression methods based on neural networks offer some advantages over MLR methods as they can account for non-linear SARs, and can deal with linear dependencies which sometimes appear in real SAR problems. However, some problems still exist in the development of SAR models using conventional backpropagation neural networks.

We have used a specific type of neural network, the Bayesian Regularized Artificial Neural Network (BRANN), in the development of SAR models. The advantage of BRANN is that the models are robust and the validation process, which scales as $O(N^2)$ in normal regression methods, is unnecessary. These networks have the potential to solve a number of problems

*Corresponding author.

which arise in QSAR modelling such as: choice of model; robustness of model; choice of validation set; size of validation effort; and optimization of network architecture. The application of the methods to QSAR of compounds active at the benzodiazepine and muscarinic receptors will be illustrated.

Keywords: QSAR; neural nets; Bayes' Theorem; drug design; molecular descriptors

INTRODUCTION

The phenomenal growth of computer technologies in the last decade [1] has been paralleled by the development of new methods for the design of bioactive agents. Concurrently, the development of important new technologies such as molecular biology and combinatorial chemistry has promised enormous increases in the amount of biological target information and structure-activity data which is available for drug design [2, 3].

Within the last decade, computer-aided molecular design methods have matured [4] and they are now responsible for an increasing number of new drugs reaching clinical trial. Recent developments include: the use of neural networks and genetic algorithms for variable selection and structure-activity mapping; novel 2D and 3D molecular representations; the use of similarity and diversity measures [5]; the simulation of combinatorial synthesis and high throughput screening *via* 'virtual receptor' design and intelligent database searching; new algorithms for ligand docking and ligand design; protein homology and motif modelling [6]. These have proven very useful for understanding relationships between biological activity and molecular structure and for facilitating the design of new drugs, pesticides and biological macromolecules. A recent, seminal example is the work of Colman and von Itzstein and coworkers in the rational design of neuriminidase inhibitors as anti-influenza drugs [7]. Molecular modelling approaches are described in other symposium papers and the intention in this work is to focus on new QSAR methods.

This paper reviews our recent work on three of these developments: novel molecular representations; Bayesian regularized neural networks; and the simulation of combinatorial chemistry, and their application to bioactive compound discovery.

DRIVERS FOR CHANGE

Historically, new bioactive molecules have been discovered by one of several methods [8]: exploratory chemical synthesis; screening of natural products;

chemical analoguing of lead compounds; observation of side effects; serendipity.

Although these methods have been very successful, accounting for the large majority of drugs and agrochemicals discovered to date, they are becoming increasingly expensive and time-consuming. As the activity of newly introduced bioactive products increases, it becomes more difficult to discover new chemical entities with substantial advantages. Clearly it has become very important to find new methods for extracting useful molecular design information from the large amounts of structure-activity data.

QSAR Methods

QSAR attempts to quantify the mapping of molecular structural or physicochemical parameters to biological activity. The method using multivariate statistics was developed by Hansch and Fujita in the 1960s [9] and it has been successfully applied to many drug and agrochemical design problems.

QSAR has advantages of speed and simplicity and it can, in some cases, account for transport and metabolic processes which occur once the compound is administered. Hence, the method is often applicable to the analysis of *in vivo* or toxicological data. However classical QSAR has limitations in that it cannot handle stereoisomers, cannot find models for compounds where the base structure varies widely, it relies on non-linear dependencies and interaction terms between the parameters being provided and QSAR analyses can be difficult to interpret in terms of mechanism at the molecular level [10]. The new QSAR methods, 3D QSAR and neural nets, have been developed in an attempt to overcome some of these shortcomings.

NEW MOLECULAR REPRESENTATIONS

Many types of molecular representation have been proposed, from Hansch parameters (σ, π, MR, log P *etc.*) to chemical graph-based methods [11–13]. Recently several new representations have been devised: molecular eigenvalues [14]; E-state fields [11]; topological autocorrelation vectors [15]; various molecular fragment-based hash codes [16–18]; and molecular holograms [19,20]. These representations may have advantages over traditional representations in their speed of computation, in more accurately representing molecular properties most relevant to receptor activity, or in being more generally applicable to diverse chemical classes acting at a common

receptor. One of our major research interests is the discovery of new, computationally efficient, information rich molecular descriptors and we review our recent work in this area.

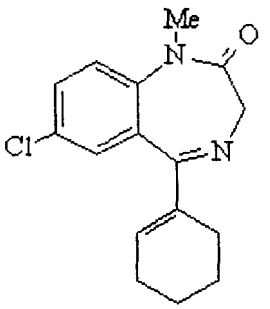
Atomistic Representations

In this deceptively simple approach [16], molecules are represented simply by counting the numbers of atoms of specific elemental type, with specific numbers of connections (a measure of the hybridization). For instance a carbon atom with four connections is denoted C4, those with three connections C3 *etc.*, and the numbers of atoms of each type can be totalled. Figure 1 illustrates an encoding example.

Although simple this representation is adequate to encode not only physicochemical parameters, such as lipophilicity and molar refractivity, but also biological activity (DHFR inhibition [16]). It has been shown to be a surprisingly good index for such a simple representation and it must be inferred that it encapsulates inherent information about the molecular structure. Its main failing is that isomers are represented by the identical indices and that the full molecular geometry is under represented.

Molecular Eigenvalues

A previous version of this eigenvalue index [14] has been developed further by Pearlman to become the BCUT (Burden, CAS, University of Texas)



Car	C4	C3	C2	Nar	N3	N2	N1	O2	O1	S	P
6	6	4	0	0	1	1	0	0	1	0	0

FIGURE 1 Example representation of tetrazepam.

index [21]. The eigenvalue indices can be thought of as quantifying the most electronegative and electropositive atoms in the molecules. This comes about because the diagonal elements of the modified adjacency matrix have been ascribed atom specific values while the off-diagonal elements have values proportional to the bond orders. The diagonalization process in effect ascribes the bond electrons back to the atoms (the trace of the matrix remaining invariant). We are studying new eigenvalue indices formed from molecular matrices representing steric, and lipophilic properties of molecules and their topologies.

Functional Group Representations

Another simple representation comes out of the work of Andrews and his coworkers [22] on deducing the relative contributions of functional groups in molecules to the binding of ligands to receptors. There is some overlap with the atomistic representation of Burden, but the focus of the work was the decomposition of molecules into functional group motifs capable of interacting with receptors. These functional groups were derived from 200 drug molecules of diverse structure binding to a diverse range of receptors and are likely to be of general applicability. The functional group representations found to be statistically significant in Andrews' analyses were: CO_2^-, PO_4^{2-}, N^+, N, OH, C=O, O/S ethers, halogens, Csp^3, Csp^2 and an entropic term related to the number of freely rotatable bonds in the molecule. As in Burden's atomistic approach, the molecule is simply broken down into it's constituent numbers of functional groups, which are then used as a representation. We are extending this work by extending the range of atom types used, including rings and hydrogen bond donors and acceptors. We are also exploring new topological indices based on paths between functional groups rather than atoms.

Molecular Multipole Moments

The above representations are simple to implement for very large numbers of compounds with diverse structures. However, a recent paper by Platt and Silverman [23] introduced a another representation which is intuitively appealing. They generated the zero-, first- and second-order molecular multipole moments with respect to atomic mass, and charge. A previous paper [24] provided a method of defining a rotationally invariant frame of reference for the second order mass moment (the principle moments of inertia with respect to the centre of mass) and second order charge moment

(the electric quadrupole moment with respect to the "centre of dipole"). They utilized this representation to carry out 3D QSAR (the CoMMA method). This method has the advantage over simple representations of accounting for 3D properties of molecules, such as shape, isomerism and conformation albeit at greater computational expense. We have extended this work by replacing the mass moments with pseudo mass moments correlating more strongly with molecular volumes, and by generating analogous 'lipophilic' molecular multipole moments (hydropoles), utilizing the hydrophobic atom constant approach of Abraham and Leo [25]. Such hydropole moments may find application to a wide range of QSAR problems, by analogy with the application of the HINT! lipophilic fields in CoMFA [26].

Molecular Holograms

The molecular hologram descriptors were derived from a common strategy to increase the efficiency of database searching by translation of chemical structure representations into bit strings (molecular fingerprints). Several approaches to fingerprinting have been implemented within commercial software [18].

Fingerprints are generated using a combination of the keyed and hashed fingerprinting approaches. The hashed fragments encode all unique linear, branched, and cyclic fragments, including overlapping fragments. Typically fragments of size 4 to 7 atoms, ignoring hydrogen atoms, are generated and hashed into bins of the fingerprint. Each corresponding fragment is then mapped to a pseudo-random integer in the range 0 to 231 using the CRC (cyclic redundancy check) algorithm. The integer generated by the CRC algorithm is unique and reproducible for each and every unique molecule. The hashing then occurs by folding the pseudo-random integer for a particular molecule into the bin range defined. Since the length of the bit string is considerably smaller than the integer to which the molecule is mapped, it is possible for different molecules to hash to the same bin (a 'collision'). Unique fragments are always hashed into the same bin. Hologram lengths are usually prime numbers so as to minimize collisions. Rather than just setting bits in each bin, a molecular hologram retains a count of the number of times each bin is set. In the case of the hashed fingerprint, the representation is purely binary, whereas in the case of the hologram, the bins contain information about the number of fragments hashed into each bin. The selection of the hologram length leading to the "best" HQSAR is based on that PLS analysis that gives either the highest

cross validated r^2 or the lowest standard error associated with the cross validation analysis.

Complexity of Representations

It appears that the complex interrelationships between elements of a simple molecular representation may be capable of encoding subtle molecular structural elements. There is much information implied in the representation. For example, a compound which contains a single sp^2 carbon atom and an single sp^2 oxygen atom must contain a carbonyl group. Thus, the presence of the carbonyl group is implied in an atomistic representation. The previous analyses of these data also suggest that significant non-linearity exists in the data set. These non-linearities may relate to higher order representations implicit in the data. These concepts have been discussed by Kier and Hall [11]. Brown and Martin [27] also found considerable inherent or latent information relevant to the forces of ligand-receptor binding in their fragment-based structural descriptors. Our results suggest, not surprisingly, that larger, more complex fragment descriptors are better able to encode information relevant to ligand-protein binding, than smaller, atom- or functional group-based fragments.

Molecular holograms, eigenvalue descriptors, molecular multipole moments, chemical graph theory, and several other developments have significantly improved the mathematical description of molecules for use in SAR studies and rational design.

STRUCTURE-ACTIVITY MAPPING

Finding structure-activity relationships is essentially a regression or pattern recognition process. Historically, linear regression methods such as MLR (multiple linear regression), PCR (principal component regression), and PLS (partial least squares) [28] have been used to develop QSAR models. Regression is an "ill-posed" problem in statistics, which sometimes results in QSAR models exhibiting instability when trained with noisy data. In addition traditional regression techniques often require subjective decisions to be made on the part of the investigator as to the likely functional (*e.g.*, non-linear) relationships between structure and activity. It is important that QSAR methods be quick, give unambiguous models, not rely on any subjective decisions about the functional relationships between structure and activity, and be easy to validate.

Neural Networks

Structure-activity relationships may be modelled by other pattern recognition methods such as artificial neural networks (ANN) [29]. In most instances ANNs have proven to be better than MLR, PCR or PLS because of their ability to handle nonlinearities in the activity surfaces. The parallel processing nature of the ANNs give them the characteristics of speed, reliability and generalisation. Networks can produce reasonable results even when some input data are missing or inaccurate, and they can generalize well when faced with new data outside its normal range of training. They do not require any assumptions to be made about the nature of the functional relationships between molecular descriptors and the mechanical, chemical or biological property being mapped. They produce better SAR models, since they are able to model complicated activity surfaces better than conventional (linear) methods. The backpropagation neural net will be discussed here, as it is the type most widely applied to QSAR problems [30].

The dispersal of knowledge over the whole of the entire network is one of the main reasons why the networks are excellent at pattern recognition. Analysis of the distribution of connection weights has shown that, after training, various parts of the network specialize in detecting specific parts of a pattern to be recognized [31].

The non-linear output from each neuron is another important feature of ANNs. If many neurons each with non-linear inputs and outputs are interconnected then the network essentially becomes a universal mathematical function approximator or model free mapping device [32]. Hence, ANNs are useful in predicting results from complex data interactions involving multiple variables where the inter-relationships among them are complex, non-linear, poorly understood or fuzzy.

Neural network training can be regularized, a mathematical process which converts the regression into a well-behaved, "well-posed" problem. The mathematics of 'well-posedness' and regularization can be found in the papers by Hadamard and Tikhonov [33]. Feed forward, backpropagation neural nets still present some problems, principal of which are overtraining, overfitting, network architecture optimization, and selection of the best QSAR model. Overtraining results from running the neural network training for too long and results in a loss of ability of the trained net to generalize. Overtraining can be avoided by used of a validation set. However, the effort to cross validate QSAR models scales as $O(N^2P^2)$ [34], where N is the number of data points and P is the number of input parameters. Where the training data set is large and diverse, as may occur in

combinatorial discovery, this can result in a prohibitively large validation times. Validation procedures also produce a family of similar QSAR models and it is not clear which of these models is preferred, or how they may be combined to give the 'best' model. It is also not obvious which neural net architecture (*e.g.*, number of hidden layers; number of nodes per hidden layer; fully connected or not) gives the best QSAR model necessitating an additional architecture optimization step. Overfitting results from the use of too many adjustable parameters to fit the training data and is avoided by use of test sets of data, not used in the training and validation steps.

Bayesian Regularized Neural Nets

Recently, we showed [35–37] how to produce robust QSAR models using a special type of neural network, a Bayesian regularized artificial neural network (BRANN). These types of neural network have found applicability in numerous other areas of data modelling but, with a very recent exception [38], have not been used for drug design. Bayesian regularized neural networks offer substantial advantages in QSAR analysis compared with other methods such as conventional backpropagation neural networks and regression techniques.

The Bayesian method is summarized in the papers by Mackay [39] and Buntine and Weigend [40] and only a brief summary will be provided here. Bayesian methods are optimal methods for solving learning problems. Any other method not approximating them should not perform as well on average [40]. They are very useful for comparison of data models (something orthodox statistics cannot do well) as they automatically and quantitatively embody "Occam's Razor". Complex models are automatically self-penalizing under Bayes' Rule [40]. Bayesian methods are complementary to neural networks as they overcome the tendency of an over-flexible network to discover non-existent, or overly complex data models.

Unlike a standard backpropagation neural network training method where a single set of parameters (weights, biases, *etc.*) are used, the Bayesian approach to neural network modelling considers all possible values of network parameters weighted by the probability of each set of weights. Bayesian inference is used to determine the posterior probability distribution of weights, and related properties from a prior probability distribution according to updates provided by the training set using the BRANN model. Where orthodox statistics provide several models with several different criteria for deciding which model is best, Bayesian statistics only offers one answer to a well-posed problem.

Bayesian methods can simultaneously optimize the regularization constants in neural nets, a process which is very laborious using cross validation. There is no better method for reliably finding and identifying better models using only the training set [39].

SIMULATION OF COMBINATORIAL DISCOVERY

The development of combinatorial chemistry, and the resultant large increases in the numbers of chemical entities screened for drug activity, has resulted in a paradigm shift in the way new drug leads are discovered. It is now possible to generate millions of chemical analogues in a relatively short time with greatly reduced effort. Rapid screening techniques have necessarily emerged to keep pace with the generation of new combinatorial libraries.

As powerful as these new methods of combinatorial and mass screening are, they are still only capable of accessing a very tiny part of the 'universe' of possible chemistries estimated to lie in the range from $10^{60}-10^{400}$. This recognition is driving the quest for methods of simulation of combinatorial synthesis and high throughput screening '*in silico*'. Methods to allow exploration of much larger region of combinatorial space would be of considerable interest in allowing a focusing of the combinatorial chemistry effort into chemical species with inherent novelty and receptor efficacy.

We have developed a QSAR-based 'virtual receptor' which allows rapid screening of very large (potentially up to 10^{12} compounds) virtual combinatorial libraries [36]. We employ computationally efficient, information rich molecular representations, coupled with neural networks, to generate our virtual receptors. Related work [29] has given indirect support to these concepts in showing that ANNs could provide a pharmacophore model. A similar flexible pharmacophore model of another receptor type has recently been described using a genetic algorithm approach [41].

APPLICATIONS OF CONCEPTS

Holographic QSAR 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 ($GABA_A$) family, at which many other classes of compounds also bind *e.g.*, benzodiazepines,

arylpyrazolo-quinolines, β-carbolines, imidazo-pyridazines, cyclopyrrolones. We have used molecular holograms to generate QSAR models for a set of benzodiazepines active at the $GABA_A$ receptor [19]. We compared the relative performance of molecular hologram representations with other traditional representations of structures, and assessed the influence of hologram fragment size on the validity of the model.

The data set used was a set of fifty-seven 1,4-benzodiazepin-2-ones (Scheme 2) [42]. This data set has been analyzed by several groups [43, 29, 17] using a number of molecular representations and SAR mapping techniques.

We found the quality of the model increases as the size of fragments and fragment complexity increases. The cross validated squared correlation coefficient q^2 was 0.784 and a training squared correlation coefficient r^2 was 0.982 with standard error = 0.119 for the best holographic QSAR model. We also found that a unit length molecular hologram and the atomistic representation used by Winkler *et al.* [17] are encoding similar information. Neural net models gave much better cross validated statistics than did PLS studies suggesting that linear regression methods are not capable of accounting for interactions between parameters and non linearity in the structure-activity response surface.

Benzodiazepine QSAR Models Using Bayesian Neural Nets

We recently reported [37] the use of Bayesian regularized neural nets to develop robust QSAR models for a series of benzodiazepines active at the

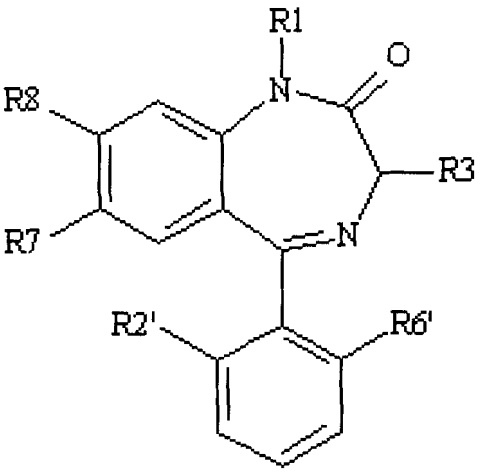


FIGURE 2 General benzodiazepine structure.

$GABA_A$ binding site of the benzodiazepine receptor (BzR). We used two benzodiazepine data sets, the 55 member set used in the holographic QSAR study above, and a considerably larger and more diverse set containing 245 compounds. The latter compounds do not have a common sub-structure so that the nature of the molecular indices used in forming the model was more important. The study used Randic/Kier and Hall topological indices and atomistic descriptors to produce a robust QSAR model. The smaller data set, which had a low experimental error, gave a training r^2 of 0.91 and a test set q^2 value of 0.82, which compared favourably with the quality of the models derived by others using a less stringent leave-one-out (LOO) cross validation procedure. Our models were validated using a test set not employed in training. We found that the QSAR model was independent of the number of hidden layer nodes beyond 4 in the Bayesian neural.

Bayesian neural net analysis of the larger, more diverse, but lower quality data set also produce a good QSAR model, although the test set q^2 value of 0.69 was lower than for the smaller set.

M_1 Muscarinic QSAR Model Using Bayesian Neural Nets

We recently reported [37] the use of Bayesian regularized neural nets to develop robust QSAR models for a diverse series of compounds active at the postsynaptic M_1 muscarinic receptor site. Muscarinic compounds are used in the treatment of memory related problems such as Alzheimer's disease. The 162 compounds in this data set did not have a common sub-structure but fell into small subsets with common structures. The study used Randic/Kier and Hall topological indices and atomistic descriptors to produce a robust QSAR model. The test set for the best QSAR model produced a q^2 of 0.63 and a training r^2 of 0.70. The Bayesian neural net produced a much better model than a PLS analysis, indicating that the response surface contains significant nonlinearity. Again, the model was robust, with successive training sessions producing virtually identical QSAR models.

Virtual Receptors

We applied some of the new fragment-based molecular representations, and neural networks to the concept of virtual receptors. We used two simple representations: the atomistic representation [16]; the functional group-based representation [22].

A data set was compiled from the literature [42, 44 – 51]. It consisted of 321 compounds of diverse structure: benzodiazepines, arylpyrazolo-quinolines,

β-carbolines, imidazo-pyridazines, cyclopyrrolones. These were broken up into two sets: 21 compounds served as the test set, whilst the other 300 compounds formed the training/validation sets.

A 5-layer 21:8:5:3:1 network proved to be most successful at modelling the data as it had the lowest predictive error. A sample output from a 21:8:5:3:1 neural network is shown in Figure 3.

The resulting virtual receptor was used to screen a virtual combinatorial library simulated by 40,000 compounds from the Maybridge chemical database. Several compounds were predicted to have high affinity for the benzodiazepine receptor and their activity is being investigated. We are continuing to develop our methods of simulating combinatory discovery by developing better representations and screening larger databases representing virtual chemical libraries. These tools will enable very large databases (those containing > 1012 compounds currently exist) to be screened and a much smaller library with a greater population of potential leads to be obtained. These enriched libraries can then be assessed by more accurate, but computationally demanding methods such as docking, or synthesized directly by combinatorial chemistry.

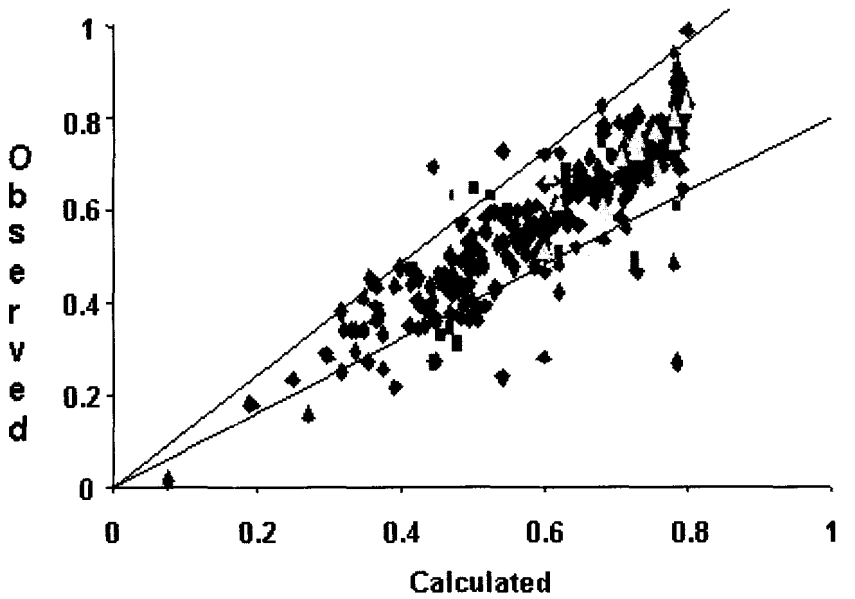


FIGURE 3 Output from a 21:8:5:3:1 network representing a Bz virtual receptor. Black = training set; dark grey = validation set; light grey = test set. (See Color Plate IV).

References

[1] Baskett, F. and Hennessy, J. L. (1993). "Microprocessors: from Desktops to Supercomputers". *Science*, **261**, 864–871.

[2] Pool, R. (1993). "Beyond Databases and E-mail". *Science*, **261**, 841–843.

[3] Wulf, W. A. (1993). "A Collaboratory Opportunity". *Science*, **261**, 854–855.

[4] Anderson, P. S., Kenyon, G. L. and Marshall, G. R., Eds., Perspectives in Drug Discovery, *ESCOM*, Leiden, 1993.

[5] Richards, W. G. (1995). "Molecular Similarity and Dissimilarity". In: *Modelling of Biomolecular Structures and Mechanisms*, Pullman, A., Ed., Kluwer Academic Publishers, Amsterdam, pp. 365–69.

[6] Blundell, T. L., Carney, D. P., Gardner, S., Hayes, F. R. F., Howlin, B., Hubbard, T. J. P., Overington, J. P., Singh, D. A., Sibanda, B. L. and Sutcliffe, M. (1988). "Knowldge-based Protein Modelling and Design". *Eur. J. Biochem.*, **172**, 513–520.

[7] von Itzstein, M., Wu, W.-Y., Kok, G. B., Pegg, M. S., Dyason, J. C., Jin, B., Phan, T. V., Smythe, M. L., White, H. F., Oliver, S. W., Colman, P. M., Varghese, J. N., Ryan, D. M., Woods, J. M., Bethell, R. C., Hotham, V. J., Cameron, J. M. and Penn, C. R. (1993). "Rational Design of Potent Sialidase-based Inhibitors of Influenza Virus Replication". *Nature*, **363**, 418–423.

[8] Rouvray, D. H. (1993). "Making the Right Connection". *Chem. Brit.*, **29**, 495–498.

[9] Hansch, C. and Fujita, T. (1964). "ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure". *J. Am. Chem. Soc.*, **86**, 1616.

[10] Holan, G. and Winkler, D. A. (1992). "QSAR and Rational Design of Insecticides", In: *Rational Approaches to Structure, Activity and Ecotoxicology of Agrochemicals*, Draber, W. and Fujita, T., Eds., CRC Press, Boca Raton, pp. 123–146.

[11] Kier, L. B. and Hall, L. H., *Molecular Connectivity in Structure-Activity Analysis*, J. Wiley and Sons, New York, 1986; Kier, L. B. and Hall, L. H., "The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modelling". In: *Reviews in Computational Chemistry*, Lipkowitz, K. B. and Boyd, D. B., Eds., VCH Publishers, NY, 1995, **2**, 367–422.

[12] Randic, M. (1975). "On Characterization of Molecular Branching". *J. Amer. Chem. Soc.*, **97**, 6609–6615.

[13] Randic, M. and Trinajstic, N. (1993). "In Search of Graph Invariants of Chemical Interest". *J. Molec. Struct.*, **300**, 551–571.

[14] Burden, F. R. (1997). "A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix". *Quant. Struct.-Activ. Relat.*, **16**, 309–14.

[15] Bauknecht, H., Zell, A., Bayer, H., Levi, P., Wagener, M., Sadowski, J. and Gasteiger, J. (1996). "Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists". *J. Chem. Inf. Comput. Sci.*, **36**, 1205–1213.

[16] Burden, F. R. (1996). "Using Artificial Neural Networks to Predict Biological Activity from Simple Molecular Structural Considerations". *Quant. Struct.-Activ. Relat.*, **15**, 7–11.

[17] Winkler, D. A., Burden, F. R. and Watkins, A. J. R. (1998). "Atomistic Topological Indices Applied to Benzodiazepines Using Various Regression Methods". *Quant. Struct.-Activ. Relat.*, **17**, 14–19.

[18] Brown, R. D. and Martin, Y. C. (1996). "Use of Structure-Activity Data to Compare Structure-based Clustering Methods and Descriptors for Use in Compound Selection". *J. Chem. Inf. Comput. Sci.*, **36**, 572–584.

[19] Winkler, D. A. and Burden, F. R., "Holographic QSAR of Benzodiazepines". *Quant. Struct.-Activ. Relat.*, **17**, 224.

[20] Tripos Associates, 1699 South Hanley Road, Suite 303, St. Louis, MO 63144.HQSAR Software v 1.0. Tripos Associates: (http://www.tripos.com/products/hqsar.html)

[21] Pearlman, R. S. and Smith, K. M., "Novel Software Tools for Chemical Diversity", In: *3D QSAR and Drug Design*, Kubinyi, H., Folkers, G. and Martin, Y. C., Eds., Kluwer/Escom, London, 1998, Volume 2, pp. 339–353; Stanton, D. T. (1999). "Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies". *J. Chem. Inf. Comput. Sci.*, **39**, 11–20.

[22] Andrews, P. R., Craik, D. J. and Martin, J. L. (1984). "Functional Group Contributions to Drug-Receptor Interactions". *J. Med. Chem.*, **27**, 1648–57.

[23] Silverman, B. D. and Platt, D. R. (1996). "Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superimposition". *J. Med. Chem.*, **39**, 2129–40.

[24] Platt, D. E. and Silverman, B. D. (1996). "Registration, Orientation, and Similarity of Molecular Electrostatic Potentials through Multiple Matching". *J. Comput. Chem.*, **17**, 358–66.

[25] Abraham, D. J. and Leo, A. J. (1987). "Extension of the Fragment Method to Calculate Amino Acid Zwitterion and Side Chain Partition Coefficients". *Proteins: Struct. Funct. Genetics*, **2**, 130–152.

[26] Cramer, R. D., Patterson, D. E. and Bunce, J. D. (1988). "Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins". *J. Am. Chem. Soc.*, **110**, 5959–5967.

[27] Brown, R. D. and Martin, Y. C. (1997). "The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding". *J. Chem. Inf. Comput. Sci.*, **37**, 1–9.

[28] Brereton, R. G., *Chemometrics: Applications of Mathematics and Statistics to Laboratory Systems*, Ellis Horwood, New York, 1990.

[29] Maddalena, D. and Johnston, G. A. R. (1995). "Prediction of Receptor Properties and Binding Affinities of Ligands to Benzodiazepine/$GABA_A$ Receptors Using Neural Networks". *J. Med. Chem.*, **38**, 715–724.

[30] Salt, D. W., Yildiz, N., Livingstone, D. J. and Tinsley, C. (1992). "The Use of Artificial Neural Networks in QSAR", *Pestic. Sci.*, **36**, 161–170.

[31] Rumelhart, D. E. and McClelland, J. L., Parallel distributed processing: Explorations in the microstructure of cognition. MIT Press, Cambridge, 1986, Vols. I and II.

[32] Maggiora, G. M. and Elrod, D. W. (1992). "Computational Neural Networks as Model-free Mapping Devices". *J. Chem. Inf. Comput. Sci.*, **32**, 732–741. Hornik, K., Stinchcombe, M. and White, H. (1988). "Multilayer Feed Forward Networks are Universal Approximators". *Neural Networks*, **2**, 359–366.

[33] (a) Hadamard, J., "Sur les problemes aux derivees parielies et leur signification physique". Bull. Univ. of Princeton, 1902, pp. 49–52. (b) Tikhonov, A. and Arsenin, V., "Solution of Ill-posed Problems". Winston, Washington, DC, 1977.

[34] Goutte, C. (1997). "*Statistical Learning and Regularization for Regression*". *Ph.D. Thesis*, University of Paris.

[35] Burden, F. R. and Winkler, D. A. (1999). "A QSAR Model for the Acute Toxicity of Substituted Benzenes Towards Tetrahymena Pyriformis using Bayesian Regularized Neural Networks". *Chem. Res. Toxicol.*, Submitted.

[36] Burden, F. R. and Winkler, D. A. (1999). "New QSAR Methods Applied to Structure-Activity Mapping and Combinatorial Chemistry". *J. Chem. Inf. Comput. Sci.*, **39**, 236–242.

[37] Winkler, D. A. and Burden, F. R. (1999). "Robust QSAR Models using Bayesian Regularized Neural Networks". *J. Med. Chem.*, Submitted.

[38] Ajay, Walters, W. P. and Murcko, M. (1998). "Can we Learn to Distinguish between "Drug-like" and "Nondrug-like" Molecules?". *J. Med. Chem.*, **41**, 3314–3324.

[39] (a) MacKay, D. J. C. (1992). "A Practical Bayesian Framework for Backprop Networks", *Neural Computation*, **4**, 415–447. (b) Mackay, D. J. C. (1995). "Probable Networks and Plausible Predictions – a Review of Practical Bayesian Methods for Supervised Neural Networks", *Comput. Neural Sys.*, **6**, 469–505. (c) Mackay, D. J. C. (1992). "Bayesian Interpolation". *Neural Computat.*, **4**, 415–447.

[40] Buntine, W. L. and Weigend, A. S. (1991). "Bayesian Back-Propagation". *Complex Sys.*, **5**, 603–643.

[41] Walters, D. E. and Hinds, R. M. (1994). "Genetically Evolved Receptor Models: A Computational Approach to Construction of Receptor Models". *J. Med. Chem.*, **37**, 2527–36.

[42] Haefely, W., Kyburz, E., Gerecke, M. and Möhler, H. (1985). "Recent Advances in the Molecular Pharmacology of Benzodiazepine Receptors and the Structure-Activity Relationships of their Agonists and Antagonists". *Adv. Drug Res.*, **14**, 165–322.

[43] So, S.-S. and Karplus, M. (1996). "Genetic Neural Networks for Quantitative Structure-Activity Relation-ships: Improvements and Application of Benzodiazepine Affinity for Benzodiazepine/GABAA Receptors". *J. Med. Chem.*, **39**, 5246–5256.

[44] Harrison, P. W., Barlin, G. B., Davies, L. P., Ireland, S. J., Matyus, P. and Wong, M. G. (1996). "Syntheses, pharmacological evaluation and molecular modelling of substituted 6-alkoxyimidazo[1,2-β]pyridazines as new ligands for the benzodiazepine receptor". *Eur. J. Med. Chem.*, **31**, 651–662.

[45] Davies, L. P., Barlin, G. B., Ireland, S. J. and Ngu, M. M. L. (1992). "Substituted imidazo[1,2-β]pyridazines. New compounds with activity at central and peripheral benzodiazepine receptors". *Biochem. Pharmacol.*, **44**, 1555–1561.

[46] Barlin, G. B., Davies, L. P., Davis, R. A. and Harrison, P. W. (1994). "Imidazo[1,2-β] pyridazines. XVII* Synthesis and central nervous system activity of some 6-(alkylthio and chloro)-3-(methoxy, unsubstituted and benzamidomethyl)-2-aryl-imidazo[1,2-β]pyridazines containing methoxy, methylenedioxy and methyl substituents". *Aust. J. Chem.*, **47**, 2001–2012.

[47] Fryer, R. I., Zhang, P., Rios, R., Gu, Z.-Q., Basile, A. S. and Skolnick, P. (1993). "Structure-activity relationship studies at the benzodiazepine receptor (BzR): A comparison of the substituent effects of pyrazoloquinoline analogues". *J. Med. Chem.*, **36**, 1669–1673.

[48] Wang, C.-G., Langer, T., Kamath, P. G., Gu, Z.-Q., Skolnick, P. and Fryer, R. I. (1995). "Computer-aided molecular modelling, synthesis and biological evaluation of 8-(benzyloxy)-2-phenylpyrazolo[4,3-χ]quinoline as a novel benzodiazepine receptor agonist ligand". *J. Med. Chem.*, **38**, 950–957.

[49] Hollinshead, S. P., Trudell, M. L., Skolnick, P. and Cook, J. M. (1990). "Structural requirements for agonist actions at the benzodiazepine receptor: studies with analogues of 6-(benzyloxy)-4-(methoxymethyl)-β-carboline-3-carboxylic acid ethyl ester". *J. Med. Chem.*, **33**, 1062–1069.

[50] Allen, M. S., Hagen, T. J., Trudell, M. L., Codding, P. W., Skolnick, P. and Cook, J. M. (1988). "Synthesis of novel 3-substituted β-carbolines as benzodiazepine receptor ligands: Probing the benzodiazepine pharmacophore". *J. Med. Chem.*, **31**, 1854–1861.

[51] Yokoyama, N., Ritter, B. and Neubert, A. D. (1982). "2-Arylpyrazolo[4,3-χ]quinolin-3-ones: Novel agonist, partial agonist and antagonist benzodiazepines". *J. Med. Chem.*, **25**, 337–339.