# Quantitative Structure−Activity Relationship Studies Using Gaussian Processes

Frank R. Burden†

School of Chemistry, Monash University, Victoria 3800, Australia

Received August 11, 2000

A Gaussian process method (GPM) is described and applied to the production of some QSAR models. These models have the potential to solve a number of problems which arise in QSAR modeling in that no parameters have to be supplied and only one hyperparameter is used in finding the optimal solution. The application of the method to QSAR is illustrated using data sets of compounds active at the benzodiazepine and muscarinic receptors as well as the data set of the toxicity of substituted benzenes to the ciliate, *Tetrahymena Pyriformis*.

## 1. INTRODUCTION

Quantitative structure−activity relationship (QSAR) methods were developed by Hansch and Fujita,[1] and they have been successfully applied to many drug and agrochemical design and optimization 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. Hence, the method is often applicable to the analysis of in vivo data. As useful as "traditional" QSAR methods have been, they still exhibit a number of difficulties and shortcomings which relate to the molecular representations used, and the methods by which SAR models are developed and validated.

It must be emphasized, at the outset, that the purpose of the Gaussian processes method (GPM) is purely predictive and *not* explanatory. Given a data set of compounds of known bioactivity and a set of molecular descriptors, the GPM can be used to predict the bioactivity of new compounds which are not included in the training data set, but it does not explain why such compounds have the activity that they do.

Finding structure−activity relationships is essentially a regression or pattern recognition process. Historically, linear regression methods such as MLR (multiple linear regression) and PLS (partial least squares) have been used to develop QSAR models though more recently nonlinear modeling utilizing artificial neural networks have been favored. 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. nonlinear) relationships between structure and activity. It is important that QSAR methods should produce unambiguous models, not rely on any subjective decisions about the functional relationships between structure and activity, and be easy to validate. Recently, regression methods based on artificial neural networks (ANNs) have been shown to overcome some of these problems as they can account for nonlinear SARs and can deal with linear dependencies which sometimes appear in real SAR problems.

Lucic et al.[2,3] have shown that in some cases nonlinear multiregression methods, which include cross-terms of higher order in the variables, may perform as well as or better than neural networks. However, for highly nonlinear problems neural networks provide a versatile modeling method and their 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.[4] Feed forward, back-propagation 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 $O(N^2P^2)$,[5] 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 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. These problems of overtraining and overfitting, which were carefully addressed by Tetko et al.,[6,7] have been incorporated in a Bayesian statistical regularisation procdure (BRANNs) which in turn has been further modified by the addition of automatic relevance determination (BRANN-ARD) which can be used to prune and analyze the network by means of weightings ascribed to the input molecular indices.[8,9]

The primary purpose of this paper is to show how to produce a robust QSAR model using the new technique of

† Corresponding author phone: (03) 9905 4593; e-mail: frank.burden@sci.monash.edu.au.

10.1021/ci000459c CCC: $20.00 
Published on Web 04/26/2001

Gaussian processes[10−12] which effectively bypass the need to define a model by means of a set of coefficients or weights and go directly from the input data to the predictions. The method is efficient for medium sized data sets (<2000) though modeling times can increase dramatically as the sets get larger since the time determining step is a matrix inversion which scales as $N^3$ for direct inversion or as $N^2$ for approximate inversions. The latter can be used with little reduction in accuracy in most cases.

## 2. METHODS

**2.1. Gaussian Processes.** In the context of this paper, the Gaussian processes method (GPM) constitutes a method of solving regression problems. The usual coefficients or weights associated with other regression methods are absent and an exact Bayesian analysis is accomplished using matrix manipulations.

Only a brief overview of GPs is given here since the problem is well explained in a paper by MacKay,[12] from which the terminology and some of the equations are reproduced. In the context of a QSAR analysis there is typically some noisy bioactivity data, the scalar target $t_n$ ($n$ = 1.. number of compounds), and some indices as vectors $\mathbf{x}_n$, which may or may not be noisy; the noise will be zero when the indices are constructed purely from structural formulas. Given that complete data set $D$ is denoted by $\{\mathbf{x}^{(n)}, t_n\}_{n=1}^N$, then we wish to predict the output $t_{N+1}$ given a novel input $x_{N+1} \notin D$ by using a nonlinear function $y(\mathbf{x})$ which underlies the data and has a set of parameters $\mathbf{w}$.

If the set of input vectors is denoted by $\mathbf{X}_N \equiv \{\mathbf{x}_n\}_{n=1}^N$ and the corresponding target values by the vector $\mathbf{t}_N \equiv \{t_n\}_{n=1}^N$, then the inference of $y(\mathbf{x})$ is described by the posterior probability distribution

$$P(y(\mathbf{x})|\mathbf{t}_N,\mathbf{X}_N) = \frac{P(\mathbf{t}_N|y(\mathbf{x}),\mathbf{X}_N)P(y(\mathbf{x}))}{P(\mathbf{t}_N|\mathbf{X}_N)} \quad (1)$$

the first term on the right-hand side $P(\mathbf{t}_N|y(\mathbf{x}),\mathbf{X}_N)$ is the probability of the data, given the function y($\mathbf{x}$), which in the case of regression problems is often implicitly assumed to be a separable Gaussian distribution. The term $P(y(\mathbf{x}))$ is the prior distribution on functions assumed by the model and is implicit in the choice of the parameters and regularizers used during the model adaptation to the data.

The idea of a Gaussian Process is, without parametrizing $y(\mathbf{x})$, to place a prior $P(y(\mathbf{x}))$ on the space of functions. The simplest type of prior over functions is called a Gaussian process. It can be thought of as a generalization of a Gaussian distribution over a finite vector space to a function space of infinite dimensions. It is specified by a mean and a covariance function rather than a mean a covariance matrix. The covariance function $C(\mathbf{x}, \mathbf{x}')$, at the point $\mathbf{x}$, $\mathbf{x}'$, expresses the expected covariance of the possible functions, $y$. Any particular $y(\mathbf{x})$ is assumed to be a single sample of this Gaussian distribution.

Given $N$ data points $\mathbf{X}_N$, $\mathbf{t}_N = \{t_n\}_{n=1}^N$, where the $t$ can be real numbers (regression models) or categorical variables, e.g. $t \in \{0,1\}$, (a classification problem). Assuming that a function $y(\mathbf{x})$ underlies the data, the problem is to predict a value for $t_{N+1}$ for a new input $\mathbf{x}^{(N+1)}$.

In this paper a linear model is assumed

$$y(\mathbf{x};\mathbf{w}) = \sum_{h=1}^{H} w_h \phi_h(\mathbf{x}) \quad (2)$$

where the $\phi_h$ are radial basis functions centered at points $\{\mathbf{c}_h\}_{h=1}^H$ with

$$\phi_h(\mathbf{x}) = \exp\left[\frac{(\mathbf{x}-\mathbf{c}_h)^2}{2r^2}\right] \quad (3)$$

where $r$ is a scaling factor. Hence the model is linear in the parameters $\mathbf{w}$ but nonlinear in $\mathbf{x}$.

The functions $y(\mathbf{x})$ are obtained by inferring the parameters $\mathbf{w}$. The posterior probability of these parameters in Bayesian terms is given by

$$P(\mathbf{w}|\mathbf{t}_N, \mathbf{X}_N) = \frac{P(\mathbf{t}_N|\mathbf{w},\mathbf{X}_N)P(\mathbf{w})}{P(\mathbf{t}_N|\mathbf{X}_N)} \quad (4)$$

where the term $P(\mathbf{t}_N|\mathbf{w},\mathbf{X}_N)$ states the probability of the data points when the $\mathbf{w}$ have been ascertained. This probability distribution is taken to be Gaussian since $t_n$ will differ from $y(\mathbf{x}^{(n)};\mathbf{w})$ by measurement error.

The distribution of weights $P(\mathbf{w})$ is also taken to be a Gaussian since the chance of occurrence of very large or very small weights is taken to be negligible. In this parametric approach, the implementation of the inference of $P(\mathbf{w}|\mathbf{t}_N,\mathbf{X}_N)$ is commonly effected by minimizing the objective function

$$M(\mathbf{w}) = \log[P(\mathbf{t}_N|\mathbf{w},\mathbf{X}_N)P(\mathbf{w})] \quad (5)$$

with respect to $\mathbf{w}$.

However it is possible to use some nonparametric methods whereby the predictions are obtained **without** representing the unknown function $y(\mathbf{x})$ as an explicit parametrized function of the $\mathbf{w}$.

Assuming $H$ basis functions and $N$ input points $\{\mathbf{x}^{(n)}\}$, the matrix $\mathbf{R}$ is defined as a matrix of the basis functions $\{\phi_h(\mathbf{x}^{(n)})\}$ at the points $\{\mathbf{x}^{(n)}\}$.

$$R_{nh} \equiv \{\phi_h(\mathbf{x}^{(n)})\} \quad (6)$$

The vector $\mathbf{y}_N$ is defined as the vector of values $y(\mathbf{x})$ at the N points.

$$y_n \equiv \sum_h R_{nh}\, w_h \quad (7)$$

If the prior distribution $P(\mathbf{w})$ is given by the Gaussian

$$P(\mathbf{w}) = \text{Normal}(\mathbf{0},\sigma_w^2\mathbf{I}) \quad (8)$$

then $\mathbf{y}$, from eq 7 must also be a Gaussian with a mean of zero.

The covariance matrix, $\mathbf{Q}$, of $\mathbf{y}$ is given by

$$\mathbf{Q} = \langle \mathbf{y}\mathbf{y}^{\mathrm{T}} \rangle = \langle \mathbf{R}\mathbf{w}\mathbf{w}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}} \rangle = \mathbf{R}\langle \mathbf{w}\mathbf{w}^{\mathrm{T}} \rangle \mathbf{R}^{\mathrm{T}} = \sigma_W^2 \mathbf{R}\mathbf{R}^{\mathrm{T}} \quad (9)$$

so that

$$P(\mathbf{y}) = \text{Normal}(\mathbf{0},\mathbf{Q}) = \text{Normal}(\mathbf{0},\sigma_w^2\mathbf{R}\mathbf{R}^{\mathrm{T}}) \quad (10)$$

This effectively removes the need to determine the vectors

**w** leaving only the parameter $\sigma_w^2$ and this is the defining property of a Gaussian process. "*The probability distribution of a function* $y(\mathbf{x})$ *is a Gaussian process if for any finite selection of points* $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, ..., $\mathbf{x}^{(N)}$, *the marginal density* $P(y(\mathbf{x}^{(1)}), y(\mathbf{x}^{(2)}), ..., y(\mathbf{x}^{(N)}))$ *is a Gaussian.*[12]"

If the target $t_n$ differs from the corresponding $y_n$ by Gaussian noise of variance $\sigma_\upsilon^2$, then **t** has the prior Gaussian distribution

$$P(\mathbf{t}) = \text{Normal}(\mathbf{0}, \mathbf{Q}+\sigma_\upsilon^2\mathbf{I}) \tag{11}$$

If the covariance of **t** is denoted by **C** where

$$\mathbf{C} = \mathbf{Q} + \sigma_\upsilon^2\mathbf{I} = \sigma_w^2\mathbf{R}\mathbf{R}^{\mathrm{T}} + \sigma_\upsilon^2\mathbf{I} \tag{12}$$

even if $H \ll N$ where **Q** does not have full rank, **C** will have full rank since $\sigma_\upsilon^2\mathbf{I}$ is of full rank.

The elements of **Q** look like

$$Q_{pq} = [\sigma_w^2\mathbf{R}\mathbf{R}^{\mathrm{T}}]_{pq} = \sigma_w^2\sum_h \phi_h(\mathbf{x}^{(p)})\phi_h(\mathbf{x}^{(q)}) \tag{13}$$

from which

$$C_{pq} = \sigma_w^2\sum_h \phi_h(\mathbf{x}^{(p)})\phi_h(\mathbf{x}^{(q)}) + \sigma_\upsilon^2\delta_{pq} \tag{14}$$

where $\delta_{pq} = 1$ if $p = q$ and 0 otherwise.

When $H \rightarrow \infty$ the sum in eq 14 can be replaced by an integral which reduces to the elements of **C** to

$$C_{pq} = C(\mathbf{x}^{(p)}, \mathbf{x}^{(q)}) + \sigma_\upsilon^2\delta_{pq} \equiv \theta_0 + \theta_1 \exp\left[-\frac{(x^{(p)} - x^{(q)})^2}{4r^2}\right] \tag{15}$$

**2.2. Implementation of Gaussian Processes for Regression.** Having formed a covariance matrix **C**, the task is to infer $t_{N+1}$ given the observed vector, $\boldsymbol{t}_N$. Since the joint density $P(t_{N+1}|\mathbf{t}_N)$ is a Gaussian distribution, then

$$P(t_{N+1}|\mathbf{t}_N) = \frac{P(t_{N+1}, \mathbf{t}_N)}{P(\mathbf{t}_N)} \tag{16}$$

is also a Gaussian.

The $\mathbf{C}_{N+1}$ matrix can be split up as follows

$$\mathbf{C}_{N+1} = \begin{bmatrix} [\mathbf{C}_N] & [\mathbf{k}] \\ [\mathbf{k}^{\mathrm{T}}] & \kappa \end{bmatrix} \tag{17}$$

from which, after some algebra

$$P(t_{N+1}|\mathbf{t}_N) = \frac{1}{Z}\exp\left[\frac{(t_{N+1} - \hat{t}_{N+1})^2}{2\sigma_{\hat{t}_{N+1}}^2}\right] \tag{18}$$

where the predictive mean is

$$\hat{t}_{N+1} = \mathbf{k}^{\mathrm{T}}\,\mathbf{C}_N^{-1}\,\mathbf{t}_N \tag{19}$$

and the error bars are

$$\sigma_{\hat{t}_{N+1}}^2 = \kappa - \mathbf{k}^{\mathrm{T}}\,\mathbf{C}_N^{-1}\,\mathbf{k} \tag{20}$$

It is important to note that $\mathbf{C}_{N+1}$ does not need to be inverted, only $\mathbf{C}_N$. The prediction produced by the Gaussian process depends only on **C** and even if $H \gg N$ the computational requirement (for exact inversion of $\mathbf{C}_N$) scale as $N^3$ or less for approximate inversion.

The only constraint on the choice of covariance function is that it must generate a nonnegative definite covariance matrix for any set of points $\{\mathbf{x}_N\}_{n=1}^N$. If the hyperparameters of the covariance function are given by $\Theta$, then the elements of the covariance matrix of **t** are given by

$$C_{pq} = C(\mathbf{x}^{(p)}, \mathbf{x}^{(q)}; \Theta) + \delta_{pq}\,N(\mathbf{x}^{(p)}; \Theta) \tag{21}$$

and $N(\mathbf{x}^{(p)};\Theta)$ can be set as a constant, $\theta_0$, for the input independent noise usually encountered in QSAR studies. A useful form of the covariance function,[13] with an additional linear term is

$$C(\mathbf{x}^{(p)}, \mathbf{x}^{(q)}) = \theta_0 + \sum_{i=1}^{M}\theta_{1,i}x_i^{(p)}\,x_i^{(q)} + \theta^2 \exp\left[-\frac{1}{2}\sum_{i=1}^{M}\frac{(x_i^{(p)} - x_i^{(q)})^2}{r_i^2}\right] \tag{22}$$

where $x_i$ is the $i$th component of **x**, an **M** (here the number of molecular indices) dimensional vector and $\theta_0$, $\theta_1$, $\theta_2$, $r_i \in \Theta$. $r_i$ is the length scale associated with each input and characterizes the distance in the $i$th direction over which $y$ is expected to vary significantly.

The problem can now be reduced to finding the most probable values for the hyperparameters $\Theta$ so that

$$P(t_{N+1}|\mathbf{x}_{N+1}, D) \cong P(t_{N+1}|\mathbf{x}_{N+1}, D, \boldsymbol{\Theta}_{MP}) \tag{23}$$

The posterior probability of $\Theta$ is

$$P(\boldsymbol{\Theta}|D) \propto P(\mathbf{t}_N|\{\mathbf{x}_n\}, \boldsymbol{\Theta})P(\boldsymbol{\Theta}) \tag{24}$$

The log of $P(\boldsymbol{\Theta}|D)$ (the evidence for the hyperparameters), $L$ is

$$L = -\log\det\mathbf{C}_N - \frac{1}{2}\,\mathbf{t}_N\mathbf{C}_N^{-1}\mathbf{t}_N - \frac{N}{2}\log 2\pi \tag{25}$$

with a derivative with respect to the hyperparameters

$$\frac{\partial L}{\partial\theta} = -\frac{1}{2}Trace\left(\mathbf{C}_N^{-1}\frac{\partial\mathbf{C}_N}{\partial\theta}\right) + \frac{1}{2}\,\mathbf{t}_N^{\mathrm{T}}\,\mathbf{C}_N^{-1}\frac{\partial\mathbf{C}_N}{\partial\theta}\mathbf{C}_N^{-1}\mathbf{t}_N \tag{26}$$

The evaluation of $\dfrac{\partial\mathbf{C}_N}{\partial\theta}$ is straightforward so that the problem reduces to maximizing $L$ by searching for the most probable value of $\boldsymbol{\Theta}$, that is $\boldsymbol{\Theta}_{MP}$. This can be done by a Monte Carlo method, a gradient method, a simplex method, or a genetic algorithm. The main difficulty occurs when $L$ is multimodal so that the search for $\boldsymbol{\Theta}_{MP}$ can lead to false optimizations of $L$. However this can be ameliorated if sensible initial values of the hyperparameters are chosen.

**2.3. Generalized Linear Model.** For many problems the GPM can be reduced to a much more tractable form known as the generalized linear model (GLM). The GLM used here is just a Gaussian process as outlined above, with the linear term removed (i.e. $\theta_{1,i} = 0$ for $i \in N$) and with all of the

length vectors scaled equally so that $r_i^2 \rightarrow \theta_3$ for $i \in N$ in eq 22. Here $\Theta$ is a three-dimensional vector of hyperparameters and its components can be named as $\theta_0$ the data noise, $\theta_2$ the standard deviation of the weights, and $\theta_3$ the width of the Gaussian basis functions, hence the search space is much smaller than for the full GPM. The covariance function is now written as

$$C(\mathbf{x}^{(p)}, \mathbf{x}^{(q)}) = \theta_0 + \theta_2 \exp\left[-\frac{1}{2}\sum_{i=1}^{M}\frac{(\mathbf{x}_i^{(p)} - \mathbf{x}_i^{(q)})^2}{\theta_3}\right] \quad (27)$$

and the evidence for the hyperparameters is given by[13]

$$L = -\frac{\theta_0}{2}\mathbf{t}_N^2 + \frac{\theta_0^2}{2}\mathbf{t}_N^T\Phi^T\mathbf{C}_N^{-1}\Phi\mathbf{t}_N - \frac{1}{2}\log\det\mathbf{C}_N - \frac{H}{2}\log\theta_2 - \frac{N}{2}\log\frac{2\pi}{\theta_0} \quad (28)$$

The hyperparameter $\theta_3$ is contained in $\Phi$ and is the width of the Gaussian functions. It is also necessary to decide on the number and placement of the basis functions. They need to be placed so that they have a significant value at the sample points and be sufficient in number. Preliminary calculations showed that good results were obtained if a basis function was placed at each sample point and an equal number was spread throughout the sample space. Similarly, it was found that the noise parameter could be set at 0.1, which corresponds to the level of 10% error noise expected in the standardized activity data, and the width of the Gaussians could be set at 0.5 for all of the data sets. This leaves only one hyperparameter, $\theta_2$ the weight distribution, to be determined. Since eq 28 is multimodal a preliminary scan over a wide range of $\theta_2$ values was carried out initially to find the region containing the maximum in the evidence (eq 28) followed by a simplex search.

**2.4. Practical Considerations.** The actual implementation of a Gaussian process is relatively straightforward. There are a few matrix multiplications to be evaluated and the covariance matrix $\mathbf{C}_N$ must be constructed and inverted. This inversion, which is the time-consuming step, can be done directly for small $N$ (i.e. $N \sim< 1000$ for current desktop computers) or by approximate means for larger $N$.

The Gaussian process equations were encoded in MATLAB[14] and added to a suite of MATLAB programs, written by the author, which includes an implementation of a Bayesian regularized neural network (BRANN) and a Bayesian regularized neural network with automatic relevance determination (BRANN-ARD).[8] The extra GPM and GLM code is based on work by Barber[15] and modified for the present purpose. It incorporates the search for the most probable hyperparameters, $\mathbf{\Theta}_{MP}$, by the methods mentioned above.

**2.5. Molecular Indices.** A combination of five sets of easily computed molecular indices was employed in this work: the well studied Randic[16] index (R); the valence modification to the Randic index by Kier and Hall (K);[17] and an atomistic (A)[18] index developed by Burden which has now been enhanced by the recognition of aromatic atoms and hydrogen atom donors and acceptors (B). The R and K indices are produced from the path length and valence electron counts in the molecule. The enhanced atomistic indices (A,B) count the numbers of each type of these atoms present in the molecule. Two further indices have been added the first counts the number of rings of various sizes (G) and the second counts some common functional groups (F). The indices and their associations are enumerated in Table 1.

**Table 1.** Molecular Indices Used in the QSAR Analyses

| index atomistic | element | no. of connections | atom type | rings | no. in ring |
|---|---|---|---|---|---|
| A1 | mol. mass | | | G3 | 3 |
| A2 | H | 1 | H1 | G4 | 4 |
| A3 | C | 2 | C2 (sp) | G5 | 5 |
| A4 | C | 3 | C3 ($sp_2$) | G6 | 6 |
| A5 | C | 4 | C4 ($sp_3$) | G7 | 7 |
| A6 | N | 1 | N1 | G8 | 8 |
| A7 | N | 2 | N2 | | |
| A8 | N | 3 | N3 | fragments | |
| A9 | N | 4 | N4 | F1 | H−O−C |
| A10 | O | 1 | O1 | F2 | H−O−N |
| A11 | O | 2 | O2 | F3 | C−O−C |
| A12 | F | 1 | F1 | F4 | N−O−C |
| A13 | Si | 2 | Si2 | F5 | N−O−N |
| A14 | Si | 3 | Si3 | F6 | C═O |
| A15 | Si | 4 | Si4 | F7 | O═C−N |
| A16 | P | 2 | P2 | F8 | O═C−O |
| A17 | P | 3 | P3 | F9 | N═O |
| A18 | P | 4 | P4 | F10 | O═N═O |
| A19 | P | 5 | P5 | | |
| A20 | S | 1 | S1 | Randic[16] | |
| A21 | S | 2 | S2 | R1 | $^0\chi$ |
| A22 | S | 3 | S3 | R2 | $^1\chi$ |
| A23 | S | 4 | S4 | R3 | $^2\chi$ |
| A24 | Cl | 1 | Cl1 | R4 | $^3\chi$ |
| A25 | Br | 1 | Br1 | R5 | $^4\chi$ |
| A26 | I | 1 | I1 | | |

| extended | element | no. of connections | atom type | Kier and Hall[17] | |
|---|---|---|---|---|---|
| B1 | C(Ar) | | c | K1 | $^0\chi^v$ |
| B2 | N(Ar) | | n | K2 | $^1\chi^v$ |
| B3 | O(Ar) | | o | K3 | $^2\chi^v$ |
| B4 | S(Ar) | | s | K4 | $^3\chi^v$ |
| B9 | H donor | | (N)H | K5 | $^4\chi^v$ |
| B10 | H donor | | (O)H | | |
| B11 | H | | N═(O) | | |

The four types of index, R, K, A, and B, are complementary, and we have been shown in previous studies[19] that their combination yields better QSAR models than the individual indices alone.

**2.6. Large Benzodiazepine Data Set.** This is a set of 245 compounds that act on the Bz receptor and was culled from the literature.[20−27] They do not have a common substructure so that the nature of the molecular indices used in forming the model become more important.

**2.7. Muscarinic Data Set.** This is a set of 162 compounds that act on the $M_1$ muscarinic receptor and was culled from the literature.[28−33] Muscarinic compounds are used in the treatment of memory related problems such as Alzheimer's disease. The compounds in this data set do not have a common substructure but do fall into small subsets with common structures. $IC_{50}$ values were measured as the concentration necessary to displace 50% of [$^3$H]quinuclidinyl benzilate (QNB) from the $M_1$ muscarinic receptor. A test set of 15% of the compounds was used.

**2.8. Toxicity Data Set.** The toxicological data set (TOX) has been modeled recently[34] by the BRANN method. It consists of 277 substituted benzenes and their toxicity to the ciliate *Tetrahymena Pyriformis* as measured by Schultz and

**Table 2.** Statistics of QSAR Models Derived from the Three Data Sets Using the RKABGF Indices

| data set | method | variables[a] | SEF[b] | $R^2$ | SEP[b] | $Q^2$ |
|---|---|---|---|---|---|---|
| BZD[c] | MLR | 39 | 0.18 | 0.47 | 0.20 | 0.32 |
| BZD | ANN[d] | 22[e] | 0.13 | 0.73 | 0.14 | 0.66 |
| BZD | BRANN | 39 | 0.12 | 0.75 | 0.12 | 0.71 |
| BZD | GPM | 39 | 0.12 | 0.76 | 0.14 | 0.66 |
| BZD | GLM | 39 | 0.12 | 0.78 | 0.13 | 0.71 |
| MUS[c] | MLR | 29 | 0.12 | 0.56 | 0.19 | 0.34 |
| MUS | ANN | 20[e] | 0.10 | 0.72 | 0.12 | 0.47 |
| MUS | BRANN | 29 | 0.13 | 0.57 | 0.13 | 0.39 |
| MUS | GPM | 29 | 0.12 | 0.61 | 0.14 | 0.35 |
| MUS | GLM | 29 | 0.12 | 0.62 | 0.16 | 0.39 |
| TOX[c] | MLR | 32 | 0.09 | 0.69 | 0.11 | 0.71 |
| TOX | ANN | 14[e] | 0.07 | 0.85 | 0.10 | 0.67 |
| TOX | BRANN | 32 | 0.06 | 0.86 | 0.08 | 0.81 |
| TOX | GPM | 32 | 0.07 | 0.81 | 0.08 | 0.82 |
| TOX | GLM | 32 | 0.08 | 0.80 | 0.09 | 0.81 |

[a] Chosen indices were as listed in Table 1 minus those represented by less than four compounds. BZD variables: R01−K05, A01, A02, A04 − A08 A10- A12,A21,A24,A25,A26,B01,B02,B04,B10,G04 − G07,F01,F03,F04,F06,F07,F08,F10. MUS variables: R01−K05, A01,A02, A04,A05,A08 A10,A11,A21,A24,A25,B01,B02,B03,B04,G05,G06,F03, F04,F08. TOX variables: R01−K05, A01,A02,A04,A05,A06,A08 A10-A12,A24 − A26,B01,B02,B10,G06,F01 − F03,F07,F08,F10 [b] SEF = standard error of fit; SEP = standard error of prediction. [c] BZD set, 245 samples; MUS set, 162 samples; TOX set, 277 samples. [d] Details of the ANN and BRANN paprameters can be found in refs 8, 9, and 34. [e] Number of principal components used in forming the model.

modeled in parts by Cronin et al.,[35−37] using MLR techniques. A test set of 15% of compounds was used in the analysis of all three data sets.

**2.9. Procedure Used in Forming the Model.** For each data set the following steps were taken. The data set was divided into a training set, and a test set was chosen by a K-means clustering algorithm clustering on X and Y values taken together. Clustering on X and Y data is the preferred method in that it clusters the compounds according to all of the given information in a manner akin to PLS; it generally gives slightly poorer training statistics than Y clustering but gives superior predictive statistics. The training set data was mean centered and the means obtained were subtracted from the test set data, which consisted of 15% of the total data set. Indices which were represented by less than four samples in any data set were discarded.

## 3. RESULTS

Table 1 shows the molecular indices (descriptors) used, and Table 2 shows the particular selection for each data set. Table 2 presents the results for calculations on the BZD, MUS, and the TOX data sets using the GPM and the GLM method and the MLR method as well as the ANN and BRANN method. The statistics of the models, the standard error of fit (SEF), coefficient of determination ($R^2$), and the standard error of prediction on the test set (SEP) with its coefficient of determination ($Q^2$) are also shown.

The models produced by the GPM and GLM are shown alongside those generated by multiple linear regression (MLR), back-propagation neural networks (ANN), and back-propagation neural networks with Bayesian regularizers (BRANN).

The MLR results are poorer than the other methods though in the MUS case this is marginal since the data set is not fitted well by any method except by the ANN, probably due to overtraining. However the final ANN model in each case has been chosen out of a series of training runs by reference to the test set which can produce spurious results since the test set is not now independent of the training set. The BRANN method is essentially an ANN calculation with a Bayesian regularizer which is used to stop the training at the evidence maximum.[8] The best model was chosen from a set of training runs using the maximum in the evidence as the objective criterion.

The GPM and GLM calculations for each data set give consistent results and are all similar to the BRANN calculations. However the GPM takes considerably longer than the GLM method. The reason for this can be seen from eq 22 where the number of hyperparameters that need to be optimized amount to $2 \times M + 2$, where $M$ is the number of indices used in forming the model; in the case of the BZD set this amounts to 80 hyperparameters. The complexity of a surface with 80 variables renders the task of finding the true maximum lengthy and prone to the finding of false maxima. In the present case, 10 runs were performed, and the run giving the maximum evidence has been reported. The 10 runs usually found several different maxima which all gave similar fit statistics but varied in their predictive capacity; the one giving the maximum evidence gave the best, or very close to the best, predictions.

It can be seen from Table 2 that the GLM method gave very similar results to the GPMs. As discussed in section 2.3, the three hyperparameters of the GPM are the width of the Gaussian basis functions, the noise in the data, and the spread of the weights. It was found that the width could be set to 0.5 and the noise to 0.1 for the three data sets, leaving only the spread of the weights to be determined. GLM calculations on the BZD set took approximately 3 min on a 400 MHz PentiumII, while the GPM calculations took approximately 15 min.

## 4. CONCLUSIONS

The Gaussian process method is based on clearly defined statistical principles and is easily programmed. Its main advantage is the lack of fitting parameters and that the final solution is obtained by maximizing the evidence with respect to some hyperparameters. However, for many cases, as illustrated here, the number of hyperparameters can be reduced to a single one, the spread of the weights (even though the weights are not explicitly evaluated). For more complex models the three GLM hyperparameters may need to be optinized. The time-consuming step if the inversion of a covariance matrix and this inversion scales as $N^3$ for exact inversion or $N^2$ for approximate inversion, which may be necessary for larger data sets. For most QSAR studies in the literature the size of the data sets are below 300 samples which is ideal for using a GPM or GLM. The two methods are also useful when quick and accurate predictions are desired, though they are not easily analyzed for the contributions of each of the independent variables. When these contributions are desired the GLM method can be useful in checking for data set integrity and model predictivity prior to using other methods such as Bayesian regularized neural networks with or automatic relevance determination.[8]

The reader is encouraged to use a web search engine to find out more about Gaussian processes where much of the new work and conference papers are published.[11]

## REFERENCES AND NOTES


(1) Hansch, C.; Fujita, T. $\rho-\sigma$-p Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc*. **1964**, *86*, 1616

(2) Lucic, B.; Trinajstic, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 121−132.

(3) Lucic, B.; Trinajstic, N. Nonlinear Multivariate Regression Outperforms Several Concisely Designed Neural Networks on Three QSPR Data Sets. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 403−413.

(4) (a) Hadamard, J. Sur les problemes aux derivees parielies et leur signification physique. Bull. Univ. of Princeton, **1902**, 49−52. (b) Tikhonov, A.; Arsenin, V. *Solution of Ill-posed Problems*; Winston: Washington, DC, 1977.

(5) Goutte, C. Statistical Learning and Regularization for Regression. Ph.D. Thesis, University of Paris, 1997.

(6) Tetko, I. V.; Villa, A. E. P.; Livingstone, D. J. Neural Network Studies. 2. Variable Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794−803.

(7) Kovalishyn, V. V.; Tetko, I. V.; Luik, A. I.; Kholodovych, V. V.; Villa, A. E. P.; Livingstone, D. J. Neural Network Studies. 3. Variable Selection in the Cascade-Correlation Learning Architecture**.** *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 651−659.

(8) Burden, F. R.; Winkler, D. A. Robust QSAR Models using Bayesian Regularized Neural Networks. *J. Med. Chem.* **1999**, *42*, 3183−3187.

(9) Burden, F. R.; Ford, M.; Whitley, D.; Winkler, D. A. Use of Automatic Relevance Determination in QSAR Studies using Bayesian Neural Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1423−1430.

(10) Williams, C. K. I.; Rasmussen, C. E. *Gaussian Processes for Regression in "Advances in Neural Information Processing Systems 8"*; Touretzky, D. S., Mozer, M. C., Hasselmo, M. E., Eds.; MIT Press: 1996.

(11) Gaussian Process Homepage http://bayes.imm.dtu.dk/gp/.

(12) Gibbs, M.; MacKay, D. J. Efficient Implementation of Gaussian Processes available from http://wol.ra.phy.cam.ac.uk/mackay/.

(13) Rasmussen, C. E. Evaluation of Gaussian Processes and other Methods for Non-Linear Regression abstract. 1996 Ph.D. Thesis, Graduate Department of Computer Science, University of Toronto.

(14) *MATLAB*; The MathWorks, Inc.: Natick, U.S.A., 1999.

(15) Barber, D. Bayesian Methods − Computer Practical, http://www.mbfys.kun.nl/∼davidb/.

(16) Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(17) Kier, L. B.; Hall, L. H. The Molecular Connectivity Chi Indexes and kappa Shape Indexes in Structure−Property Modelling. In *Reviews in Computational Chemistry*; Lipkowitz K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1995; Vol. 2, pp 367−422, and references therein.

(18) Burden, F. R. Using Artificial Neural Networks to Predict Biological Activity from Simple Molecular Strctural considerations. *Quant. Struct.-Activ. Relat*. **1996**, *15*, 7−11

(19) Winkler, D. A.; Burden, F. R.; Watkins, A. J. R. Atomistic Topological Indices Applied to Benzodiazepines using Various Regression Methods. *Quant. Struct.-Activ. Relat*. **1998**, *17*, 14−19.

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

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

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

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

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

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

(26) Allen, M. S.; Hagen, T. J.; Trudell, M. L.; Codding, P. W.; Skolnick, P.; Cook, J. M. Synthesis of novel 3-substituted b-carbolines as benzodiazepine receptor ligands: Probing the benzodiazepine pharmacophore. *J. Med. Chem.* **1988**, *31*, 1854−1861.

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

(28) Orlek, B. S.; Blaney, F. E.; Brown, F.; Clark, M. S. G.; Hadley, M. S.; Hatcher, J.; Riley, G. J.; Rosenberg, H. E.; Wadsworth, H. J.; Wyman, P. Comparison of Azabicyclic Esters and Oxadiazoles as Ligands for the Muscarinic Receptor. *J. Med. Chem.* **1991**, *34*, 2726−2735.

(29) Wadsworth, H. J., Jenkins, S. M.; Orlek, B. S.; Cassidy, F.; Clark, M. S. G.; Brown, F.; Riley, G. J.; Graves, D.; Hawkins, J.; Naylor, C. Synthesis and Muscarinic Activity of Quinuclidin-3-yltriazole and -tetrazole Derivatives. *J. Med. Chem*. **1992**, *35*, 1280−1290.

(30) Ward, J. S.; Merritt, L.; Klimkowski, V. J.; Lamb, M. L.; Mitch, C. H.; Bymaster, F. P.; Sawyer, B.; Shannon, H. E.; Olesen, P. H.; Honoré, T.; Sheardown, M. J.; Sauerberg, P. Novel functional M1 selective muscarinic agonists. 2. Synthesis and structure−activity relationships of 3-pyrazinyl-1,2,5,6-tetrahydro-1-methylpyridines. Construction of a molecular model for the M1 pharmacophore. *J. Med. Chem.* **1992**, *35*, 4011−4019.

(31) Sauerberg, P.; Olesen, P. H.; Nielsen, S.; Treppendahl, S.; Sheardown, M. J.; Honoré, T.; Mitch, C. H.; Ward, J. S.; Pike, A. J.; Bymaster, F. P.; Sawyer, B. D.; Shannon, H. E. Novel functional M1 selective muscarinic agonists. Synthesis and structure−activity relationships of 3-(1,2,5-thiadiazolyl)-1,2,5,6-tetrahydro-1-methylpyridines. *J. Med. Chem.* **1992**, *35*, 2274−2283.

(32) Jenkins, S. M.; Wadsworth, H. J.; Bromidge, S.; Orlek, B. S.; Wyman, P. A.; Wiley: G. J.; Hawkins, J. Substituent Variation in Azabicyclic Triazole and Tetrazole-Based Muscarinic Receptor Ligands. *J. Med. Chem.* **1991**, *35*, 2392−2406.

(33) Sauerberg, P.; Kindtlet, J. W.; Nielsen, L.; Sheardown, M. J.; Honoré, T. Muscarinic Cholinergic Agonists and Antagonists of the 3-(3-Alkyl-1,2,4-oxadiazol-5-yl_-1,2,5,6-Tetrahydropyridine Type. Synthesis and Structure−Activity Relationships. *J. Med. Chem.* **1991**, *34*, 687−692.

(34) Burden, F. R.; Winkler, D. A. A Quantitative Structure−Activity Relationships Model for the Acute Toxicity of Substituted Benzenes to *Tetrahymena pyriformis* using Bayesian Regularized Neural Networks. *Chem. Res. Toxicol.* **2000**, *13*, 436−440.

(35) Dearden, J. C.; Cronin, M. T. D.; Schultz, T. W.; Lin, D. T. QSAR study of the toxicity of Nitrobenzenes to Tetrahymena pyriformis. *Quantum Struct.-Act. Relat.* **1995**, *14*, 427−432.

(36) Cronin, M. T. D. Bryant, S. E. Dearden, J. C. Schultz, T. W. Quantitative Structure−Activity study of the Toxicity of Benzonitriles to the ciliate Tetrahymena Pyriformis. *SAR QSAR Environ. Res.* **1995**, *3*, 1−13.

(37) Cronin, M. T. D.; Gregory, B. W.; Schultz, T. W. Quantitative Structure−Activity Analyses of Nitrobenzene to Tetrahymena Pyriformis. *Chem. Res. Toxicol*. **1998**, *11*, 902−908.


CI000459C