Gaussian Processes for Classification: QSAR Modeling of ADMET and Target Activity

Olga Obrezanova and Matthew D. Segall*

Optibrium Ltd., 7226 IQ Cambridge, Beach Drive, Cambridge, CB25 9TL, United Kingdom

Received October 21, 2009

In this article, we extend the application of the Gaussian processes technique to classification quantitative structure−activity relationship modeling problems. We explore two approaches, an intrinsic Gaussian processes classification technique and a probit treatment of the Gaussian processes regression method. Here, we describe the basic concepts of the methods and apply these techniques to building category models of absorption, distribution, metabolism, excretion, toxicity and target activity data. We also compare the performance of Gaussian processes for classification to other known computational methods, namely decision trees, random forest, support vector machines, and probit partial least squares. The results indicate that, while no method consistently generates the best model, the Gaussian processes classifier often produces more predictive models than those of the random forest or support vector machines and was rarely significantly outperformed.

INTRODUCTION

Quantitative structure−activity relationship (QSAR) models can be broadly separated into two types: Regression models predict a numerical value of a property based on the structure of a molecule, and classification models predict if a molecule will fall into a property class (e.g., ‘high’ or ‘low’), sometimes with an associated probability of class membership. In general, it would be preferable to predict a numerical value for a property, as this enables selection of molecules based on an arbitrary criterion. However, it often proves to be impossible to build a regression model of acceptably high accuracy. This may be due to variability in the underlying experimental measurements, sparsity of available data, or lack of descriptors with sufficiently high correlation with the observed property. In these cases, a high-quality classification model can often be built which provides discrimination between molecules, at least between broad classes.

Techniques for building regression models based on the Gaussian processes (GP) method have previously been published by the authors[1] and other groups.[2,3] The GP method is a powerful, robust method for nonlinear regression; it does not require subjective determination of model parameters, is able to handle a large pool of descriptors and to select the important ones, is inherently resistant to overtraining, and offers a way of estimating uncertainty in predictions. We have applied this method to build models of various ADMET properties: hERG inhibition, blood-brain barrier (BBB) penetration, solubility at pH 7.4, and intrinsic aqueous solubility.[1,4]

In this paper, we extend the application of GP to the generation of classification models. Two methods will be explored: an intrinsic GP classification technique and an approach using GP regression techniques, combined with a probit analysis. We will describe the underlying theory and compare the performance of these two methods with other classification techniques using seven example data sets, a BBB classification previously published by Zhao et al.,[5] hERG inhibition, human intestinal absorption (HIA),[6] and four activity data sets published by Sutherland et al.[7] A different approach for classification using GP was recently used by Schwaighofer et al. to model metabolic stability.[8]

The models generated with the two GP methods are compared with other, widely used methods for classification that represent the current state of the art, decision trees (DT),[9] support vector machines (SVM),[10] and random forests (RF).[11] We also compare the results with a classification approach based on the linear regression technique partial least squares (PLS).[12]

The underlying theory and resulting equations for the GP classification methods, an outline of the other common methods applied for comparison, the details of the data and the metrics used to assess the quality of the models are described in the Methods and Data Section. In the Results Section, the models generated for the seven data sets are summarized, and the differences between the GP and other methods are examined in more detail for two of the example data sets, BBB and hERG inhibition. Finally, the conclusions of this study are summarized in the Conclusions Section.

METHODS AND DATA

GP Binary Classifier. Here we will briefly describe the underlying method for generating classification models using GPs and give the final formulas to enable the reader to implement this technique. For detailed discussions and derivation of the formulas, we refer the reader to the recent book by Rasmussen and Williams[13] and to the work of Gibbs and Mackay.[14]

The classification problem may be defined as follows. Let $\mathbf{X} = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$ be the matrix of molecular descriptors for the molecules in the training set, where $\mathbf{x}^{(n)} = \{x_i^{(n)}\}_{i=1}^{K}$ is the vector of descriptors associated with molecule n. Let $\mathbf{Y} = \{Y^{(n)}\}_{n=1}^{N}$ be the corresponding vector of class labels (-1 or $+1$) for molecules in the training set. Here, N is the number of compounds in the training set, and K is the number

* Corresponding author e-mail: matt.segall@optibrium.com.

10.1021/ci900406x
Published on Web 04/30/2010

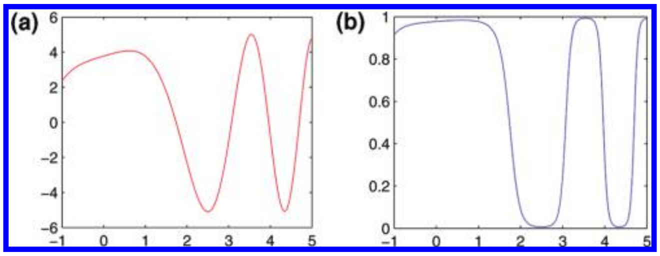

Figure 1. Graphical illustration of GP classification for the case of a one-dimensional descriptor space. Graph (a) shows a latent function drawn from a GP. Graph (b) shows the result of "squashing" this function through the logistic response function (eq 1) to obtain the class membership probability.

of descriptors. We wish to model the probability distribution of the class label y for a molecule given its descriptor vector $\mathbf{x}$, $p(y|\mathbf{x})$.

The GP method proceeds by Bayesian inference directly in the space of functions that link the descriptors of a molecule with the probability of the molecule falling within a class. Initially, a prior probability distribution over functions is assumed, controlled by 'hyperparameters' of the GP. By taking into account the available experimentally observed property values, only those functions from the prior which pass close to or exactly through the training points are considered. The combination of the prior and the data leads to a posterior probability distribution over functions that identifies those most likely to describe the observed data. We can average over all functions in the posterior distribution and take the mean value as the prediction, while the full distribution provides an estimate of the uncertainty for each prediction.

Theoretical Foundation. To apply GPs to a classification problem, we need to identify a variable which can be assigned a GP prior. The class labels are not suitable for this purpose, therefore, we introduce a latent function $f(\mathbf{x})$ to which we can assign a GP prior and use a regression treatment to model $f(\mathbf{x})$. This function then can be 'squashed' by passing it through the logistic response function:

$$g(f) = \frac{1}{1 + \exp(-f)} \tag{1}$$

to obtain the class probability:

$$p(y = +1|\mathbf{x}) = g(f(\mathbf{x})) = \frac{1}{1 + \exp(-f(\mathbf{x}))} \tag{2}$$

This is illustrated in Figure 1 for a one-dimensional input space. It should be noted that the latent function $f(\mathbf{x})$ will never be directly observed and that it will ultimately be integrated out.

The latent function $f(\mathbf{x})$ follows a GP described by covariance function $C(\mathbf{x}^{(n)}, \mathbf{x}^{(m)})$. In this work, we have used a squared exponential covariance function.

The inference steps can be briefly described as follows. First, we obtain the posterior distribution for the latent function value f^* at a new point $\mathbf{x}^*$, given all of the training data:

$$p(f^*|\mathbf{X}, \mathbf{Y}, \mathbf{x}^*) = \int p(f^*|\mathbf{X}, \mathbf{x}^*, \mathbf{f})p(\mathbf{f}|\mathbf{X}, \mathbf{Y})d\mathbf{f} \tag{3}$$

where the posterior over the latent variables can be obtained by Bayes' rule:

$$p(\mathbf{f}|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})}{p(\mathbf{Y}|\mathbf{X})} \tag{4}$$

Next, we use the posterior distribution for f^* (eq 3) and integrate over the logistic response function (eq 2) to obtain the probability of class membership:

$$p(y = +1|\mathbf{x}^*) = \int g(f^*)p(f^*|\mathbf{X}, \mathbf{Y}, \mathbf{x}^*)df^* \tag{5}$$

In the case of a regression problem, all distributions are Gaussian and all integrals can be treated analytically. This is not the case for a classification problem; the likelihood in eq 3 is non-Gaussian and that makes the integral analytically intractable. Therefore, approximation methods for the integrals must be used. Two such methods are described by Rasmussen and Williams[13] and one of these approaches, the method of 'expectation propagation' was used by Schwaighofer et al. in modeling metabolic stability.[8] In this work, we have used the method of variational lower and upper bounds suggested by Gibbs and Mackay.[14] They obtain upper and lower bounds for the unnormalized posterior density $p(\mathbf{Y}|\mathbf{f})p(\mathbf{f})$ (see eq 4). These bounds are parametrized by variational parameters which are optimized to achieve the tightest possible fit. The bounds can then be used to derive approximations for the posterior distribution of f^* (eq 3) and for the class probability (eq 5).

Final Formulas. For the sake of computational efficiency, we have used only the lower bound approximation in our implementation. The final formulas are summarized below. The details of the derivation of the formulas can be found in Gibbs and Mackay.[14]

As a covariance function (to impose a prior on latent variables), we have used a squared exponential covariance function:

$$C(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}) = \theta_1 \exp\left[-\frac{1}{2}\sum_{i=1}^{K}(x_i^{(n)} - x_i^{(m)})^2/r_i^2\right] + \theta_2 + \varepsilon\delta_{nm} \tag{6}$$

where θ_1, θ_2, and $\{r_i\}_{i=1}^{K}$ are hyperparameters. We assume the latent variables to be noise free, but to make the matrix computations well-conditioned, we added the term $\varepsilon\delta_{nm}$, where $\delta_{nm} = 1$ if $n = m$ and $\delta_{nm} = 0$ otherwise, and ε is a small number (for example, $\varepsilon = 0.1$).

The lower bound approximation for the posterior distribution f^* for a new point $\mathbf{x}^*$ is a Gaussian distribution with mean a and variance σ^2:

$$a = \frac{1}{2}\mathbf{k}^T\mathbf{H}^{-1}\mathbf{Y} \tag{7}$$

$$\sigma^2 = C(\mathbf{x}^*, \mathbf{x}^*) - 2\mathbf{k}^T\mathbf{H}^{-1}\mathbf{\Lambda}\mathbf{k}, \tag{8}$$

where the vector $\mathbf{k}$ with components $k_n = C(\mathbf{x}^*, \mathbf{x}^{(n)})$ describes the similarity of the new molecule to the ones in the training set, C is the covariance function (eq 6),

$$\mathbf{H} = \mathbf{I} + 2\mathbf{\Lambda}\mathbf{C} \tag{9}$$

and $\mathbf{C}$ is the covariance matrix. $\mathbf{\Lambda}$ is a diagonal matrix depending on variational parameters ν_n ($n = 1, ..., N$), its elements are defined in the following way:

$$\Lambda_{nn} = \frac{g(\nu_n) - 0.5}{2\nu_n} \tag{10}$$

where $g(\nu)$ is the sigmoid function from eq 1.

This approximation to the posterior distribution for latent variable f^* and eq 5 is then used to derive the probability of a new compound $\mathbf{x}^*$ belonging to class +1:

$$p(y = +1|\mathbf{x}^*) = g(a/\sqrt{1 + \pi\sigma^2/8}) \tag{11}$$

A compound is assigned to the class with the highest probability. As this is a binary classification problem, this is equivalent to a probability threshold of 0.5.

Optimization of Hyperparameters and Variational Parameters. Learning a GP classifier means finding hyperparameters θ_1, θ_2, and r_i ($i = 1, ..., K$) and variational parameters ν_n ($n = 1, ..., N$). This is equivalent to 'fitting' the parameters of a model in other forms of regression. The variational parameters are optimized to ensure that the lower bound on $P(\mathbf{Y}|\mathbf{X}, \Theta)$ is as tight a bound as possible. The hyperparameters of the covariance function should be set to their most probable values given the data. This can be achieved by optimizing a normalizing constant in eq 4.

In summary, the optimal values of hyperparameters and the variational parameters can be found by maximizing the following function:

$$\log Z = \sum_n G(\nu_n) + \frac{1}{8}\mathbf{Y}^T\mathbf{C}\mathbf{H}^{-1}\mathbf{Y} - \frac{1}{2}\log \det(\mathbf{I} + 2\mathbf{\Lambda}\mathbf{C}) \tag{12}$$

where

$$G(\nu_n) = \log(g(\nu_n)) + 0.5\nu_n(g(\nu_n) - 1.5) \tag{13}$$

We have used the conjugate gradient method with the Polak—Ribiere formula[15] to optimize the function log Z (eq 12).

Probit GP. In the previous section, we described an intrinsic binary classification technique. However, we can also use GP regression to directly build a continuous model of the property class variable and then apply a probit transformation to predictions to assign new molecules to a class. This idea is similar to the approach used for GP classifiers, where we applied a logistic transformation to the latent function to obtain class membership probabilities (eqs 1 and 2). We have described GP regression techniques in our previous work.[1,4]

The details are as follows: A GP model is built of the class labels $\mathbf{Y}$, which each take values −1 or +1 (any numerical values could be used in principle, for example 0/1). Let us denote the model prediction for a new point $\mathbf{x}^*$ by a^* and the standard deviation in the prediction by σ^*. For GP models, this uncertainty is calculated by the model and is individual to each molecule.

In the most simple case, ignoring uncertainty in prediction, the class membership can be assigned by applying a threshold to the predictions. In the case of −1, +1 labels it is appropriate to take $t = 0$ as a threshold.

The class probability can be calculated as

$$p(y = +1|\mathbf{x}^*) = 1 - \Phi(t, a^*, \sigma^*) \tag{14}$$

where $\Phi(t, m, \sigma)$ is the cumulative distribution function at t of the normal distribution with mean m and standard deviation σ. In this case, making a prediction is equivalent to using a threshold of 0.5 on probability to assign a class. Here instead of using the logistic transformation function (eq 1), we use the probit transformation, the cumulative distribution function for normal distribution.

Other Classification Techniques. We will compare the performance of the GP classifier and probit models with those generated with DT, RF, SVM and a probit treatment of a PLS model.

Decision Trees (DT). The DT technique applies a recursive partitioning approach to building classification models. Here we used the DT technique implemented within StarDrop's Auto-Modeler software,[16] which is based on the C4.5 algorithm introduced by Quinlan.[9] Models built using the DT method are easy to interpret but often have low-predictive ability. This drawback can be overcome by the tree ensemble techniques, such as RF described below.

The Auto-Modeler generates multiple DTs using different approaches. In this study, the results for the best DT, as measured on the test set, is reported in each case. It should be noted that this represents a small degree of 'fitting' of the results to the test set. Ideally, the robustness of this peformance should be tested on a further, independent validation set. However, for this study only training and test sets were available, so direct comparison of the DT results with other modeling methods should be treated with caution.

The DT method can estimate probabilities of belonging to a class based on the Laplace ratio[9] obtained using the results from the training and test sets. The probabilities are determined for each leaf of the DT, and so all compounds in a leaf will be assigned the same probabilities.

Random Forest (RF). The RF method[11] generates an ensemble of DTs based on different subsets of the training data. To create subsets, molecules from the training set are sampled with replacement, a sampling technique called 'bootstrapping'. Furthermore, a randomly selected subset of the descriptors is chosen to classify the compounds at each node of the tree. Classifications of new molecules are assigned by majority voting across the ensemble. The probability of a molecule belonging to a class is determined by the frequency of prediction across the ensemble (i.e., it is equal to the number of trees predicting the molecule in that class and divided by the total number of trees). In a two-class problem, this is equivalent to using a probability threshold of 0.5 to assign a molecule to a class.

The implementation of RF in the Weka program[17] was used in this study. In each case, an ensemble of 100 trees was generated.

Support Vector Machines (SVM). The SVM method[10] maps the training data points into a high-dimensional space in which points from the different classes are separated as well as possible by a hyperplane in that space. New points are then classified according to which side of the hyperplane they lie.

In this study, we have used the SVM implementation in libSVM[18] accessed throught the Weka interface.[17] The C-SVC method with a radial basis function kernel was used. In each case, the cost and gamma parameters of the kernel were found using a simple grid search to optimize the performace of the SVM using a 10-fold cross validation on the training set and the final model trained on the full training

Table 1. Summary of the Data Sets Used in This Study

		number of compounds			
		training		test	
data set	reference	high	low	high	low
BBB	Zhao et al. 2007[5]	832	260	450	49
hERG	this paper	84	34	33	17
HIA	Zhao et al. 2001[6]	138	20	57	10
COX-2	Sutherland et al. 2003[7]	87	91	61	64
BZR	Sutherland et al. 2003[7]	94	87	63	62
DHFR	Sutherland et al. 2003[7]	84	149	42	118
ER	Sutherland et al. 2003[7]	110	156	70	110

set. The descriptor values were normalized to lie within the range [0,1] before applying the SVM.

Probit Partial Least Squares (PLS) Model. The probit treatment, which we described in relation to the GP probit method, can be applied to a model produced by any regression technique. For comparison, in this study we also use the PLS method[12] to build continuous models and apply a probit treatment to produce classifications.

PLS does not provide an estimate of the uncertainty σ^* in prediction, which we need to be able to calculate class probability $p(y = +1|\mathbf{x}^*)$ (eq 14). Therefore, for a PLS model, the uncertainties are calculated from the actual root-mean-square error (RMSE) on the independent test set, taking into account whether a new compound lies inside or outside the chemical space of the model. The chemical space is determined using a Hotelling's T^2 test in the space of model descriptors. In this case, the uncertainty estimates are not individual for each compound, but one value is assigned to all compounds which lie within the chemical space, and a higher uncertainty to compounds outside of the chemical space.

We use the PLS implementation available in StarDrop's Auto-Modeler.[16] Together with each prediction, models built by the Auto-Modeler produce an uncertainty in prediction, standard deviation σ^*, as described above.

Data Sets. To make a comparison between GPs for classification and other classification techniques we have chosen seven data sets: a BBB data set previously published and modeled by Zhao et al.,[5] a hERG inhibition data set compiled in house and derived from various literature sources, a HIA data set also published by Zhao et al.,[6] and four target activity data sets published and modeled by Sutherland et al.[7] A summary of these data sets is shown in Table 1, and more detail is provided below.

Blood-Brain Barrier (BBB) Data Set. This data set contains 1593 compounds; 1283 compounds that penetrate the blood-brain barrier (class 'high') and 310 compounds with little ability to penetrate the blood-brain barrier (class 'low').[5] The data set is originally based on that published by Adenot and Lahana.[19] Classification of high compounds was assigned on the basis of their central nervous system (CNS) activity. Identifying low compounds is more complicated, and we refer to Adenot and Lahana[19] for a detailed account of the criteria used.

Zhao et al.[5] have modeled this data set using 19 simple molecular descriptors which mostly relate to hydrogen-bonding properties of molecules. These include Abraham descriptors, PSA, log *P*, log *D*, pK_a for acid and base, numbers of rotatable bonds, and hydrogen-bonding donors and acceptors. They also built models using fragmentation schemes. The computational techniques Zhao et al. used to build models included recursive partitioning and binomial PLS methods.

Zhao et al.[5] made the data set available along with the split into training and test sets and the 19 calculated molecular descriptors. To facilitate comparison, we have used the same split and descriptors, although we excluded two duplicate compounds. The final data set we used for modeling contains 1591 compounds (1092 compounds in the training set and 499 compounds in the test set) and 18 descriptors, since one descriptor was excluded as highly correlated during the descriptor-filtering stage (as described below in relation to the hERG data set).

hERG Inhibition Data Set (hERG). Data on hERG (human ether-a-go-go-related gene) potassium channel blockers were derived from various literature sources. A total of 168 compounds with patch-clamp pIC_{50} values for inhibition of the hERG channel expressed in mammalian cells were selected. This data set is an extension of the set we have modeled in the previous work.[1] Here we used a threshold of pIC_{50} = 5 (IC_{50} = 10 μM) to classify compounds into two classes, i.e. compounds with p$IC_{50} \leq 5$ are considered inactive (class 'low') and compounds with p$IC_{50} > 5$ are active (class 'high'). The final data set contains 117 active compounds ('high') and 51 inactive compounds ('low'). The data set and references to literature sources are provided in the Supporting Information.

To generate descriptors and to prepare the hERG set for modeling, we used the automatic modeling procedure from StarDrop's Auto-Modeler which is described in detail in our previous work.[4]

In summary, two-dimensional (2D) SMARTS-based descriptors, which are counts of atom type and functionalities, and whole molecule properties such as log *P*, molecular weight, and polar surface area (a total of 330 descriptors) were calculated.

The initial data set were split into training and test sets containing 70% and 30% of the data, respectively. The split of the initial data set into subsets was performed by cluster analysis based on 2D path-based chemical fingerprints and by the Tanimoto similarity index. Compounds were clustered using an unsupervised nonhierarchical clustering algorithm developed by Butina.[20] Clusters were defined using a Tanimoto similarity index of 0.7. Once the clusters were formed, the cluster centroids and singletons were assigned to the training set. Other cluster members were assigned randomly to the training set until the correct number of compounds were assigned to the training set. The remaining compounds were assigned to the test set.

The calculated descriptors were subjected to a descriptor preselection step that removed descriptors with low variance and low occurrence. Specifically, descriptors with a standard deviation less than 0.0005 and descriptors represented by less than 4% of the compounds in the training set were excluded. Also, highly correlated descriptors are excluded (when the pairwise correlation exceeds 0.95 in the training set), such that just one of the pair remains.

The final hERG set used in modeling contains 157 descriptors.

Human Intestinal Absorption (HIA). The data set derived from Zhao et al.[6] is based on clinically observed HIA and

Table 2. Summary of Model Results for the ADMET Data Sets, Comparing Performance of Models Built with Different Methods[a]

		accuracy training (%)			accuracy test (%)			
data set	method	high	low	overall	high	low	overall	test κ
BBB	GP classifier	99	80	94	100	73	97	0.81
	probit-GP	99	78	94	100	76	97	0.84
	DT	99	86	96	100	78	98	0.85
	RF	100	100	100	100	80	98	0.88
	SVM	99	82	95	100	76	97	0.84
	probit-PLS	99	75	93	100	76	97	0.84
	reference (Binomial PLS)[5]	95	82	92	98	80	97	N/A
hERG	GP classifier	96	82	92	97	65	86	0.66
	probit-GP	99	71	91	100	53	84	0.60
	DT	82	98	93	91	65	82	0.58
	RF	100	100	100	100	47	82	0.54
	SVM	99	47	84	100	47	82	0.54
	probit-PLS	95	44	81	97	41	78	0.43
HIA	GP Classifier	99	65	94	98	50	91	0.58
	probit-GP	100	60	95	100	20	88	0.30
	DT	96	95	96	88	10	85	0.50
	RF	100	100	100	100	30	90	0.42
	SVM	100	100	100	100	40	91	0.53
	probit-PLS	98	30	89	97	19	84	0.09

[a] Where available, the result for the best model in the corresponding reference is also provided for comparison, and the method used is noted.

contains a total of 225 compounds. These have been assigned to two classes: 'high' indicates a HIA greater than or equal to 30% and 'low' indicates a HIA less than 30%.

As neither training/test set split nor descriptors were provided by Zhao et al., the same methods for set splitting and descriptor generation, as described above, for the hERG data set were applied, resulting in 158 compounds in the training set (138 high and 20 low) and 67 compounds in the test set (57 high and 10 low). The final HIA set contains 166 descriptors.

Target Activity Data sets. Target activity data sets for cyclooxygenase-2 (COX-2), benzodiazepine receptor (BZR), dihydrofolate reductase (DHFR), and the estrogen receptor (ER) were obtained from the publication by Sutherland et al.[7] The data set splits for these sets, based on a coverage-based diversity algorithm, 2D structural fingerprints, and Tanimoto similarity, were provided in the supporting information for this reference. Therefore, we have used the same splits for ease of comparison with the models generated in their paper. These are summarized in Table 1.

Sutherland et al. used a set of 1D and 2D descriptors in their models, however, these were not provided by the authors. Therefore, the descriptor generation method described for the hERG data set (see above) has been applied to each of these sets. These resulted in 107, 138, 125, and 119 descriptors for the COX-2, BZR, DHFR, and ER data sets, respectively.

Evaluation of Model Performance. To measure the performance of classification models, we use the κ statistic as well as the overall accuracy and accuracies for individual classes. The κ statistic assesses the model's improvement in prediction over chance and measures the agreement between observed and predicted classification.

Let us assume that the confusion matrix for a binary model has the following form:

$$\begin{array}{lllll} & & \text{Predicted} & \\ & & \text{Class } -1 & \text{Class } +1 \\ \text{Observed} & \text{Class } -1 & \text{TN} & \text{FP} \\ & \text{Class } +1 & \text{FN} & \text{TP} \end{array} \qquad (15)$$

Here, TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives. In this case, the κ statistic is defined as follows:

$$\kappa = \frac{\mathrm{TN} + \mathrm{TP} - \eta}{N - \eta} \qquad (16)$$

where $\eta = [(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP}) + (\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})]/N$ is the agreement expected by chance.

A κ statistic exceeding 0.8 would mean very good agreement and $0.6 \le \kappa < 0.8$ indicates a good agreement between predicted and observed classification.

We will also evaluate model performance using the area under the receiver operating characteristic (ROC) curve, designated as AUC.

RESULTS

An overview of the results obtained for the seven data sets studied is provided below. Subsequently, more detailed results, highlighting the relative strengths of the different modeling methods, are provided for the BBB and hERG data sets.

Summary of Results. The prediction accuracies and test set κ value for each method and data set are summarized in Table 2 and 3.

One immediate observation that may be made is that there is no single method that consistently generates the most accurate model on the independent test set. In three cases, the GP classifier model produces the best model, as measured by κ statistic, in three cases RF, and in a final case the model produced with SFGA in Sutherland et al.[7] (In this last case, the best model generated in this study was that using DT, although as discussed in the DT methods section, this result should be treated with caution).

Table 3. Summary of Model Results for the Activity Data Sets, Comparing Performance of Models Built with Different Methods[a]

		accuracy training (%)			accuracy test (%)			
data set	method	high	low	overall	high	low	overall	test κ
COX-2	GP Classifier	93	86	89	82	72	77	0.54
	probit-GP	95	86	90	77	69	73	0.51
	DT	95	86	90	77	69	73	0.51
	RF	100	100	100	73	74	74	0.47
	SVM	100	100	100	75	73	74	0.49
	probit-PLS	77	85	81	79	72	75	0.51
	ref 7 (SFGA)	83	87	85	75	72	73	0.47
BZR	GP Classifier	81	88	84	70	77	74	0.47
	probit-GP	86	86	86	70	79	74	0.49
	DT	98	78	88	78	73	75	0.50
	RF	100	100	100	65	81	73	0.46
	SVM	86	89	87	67	79	73	0.46
	probit-PLS	85	74	80	78	71	74	0.49
	ref 7 (SFGA)	81	76	79	70	81	75	0.51
DHFR	GP Classifier	79	95	89	71	83	80	0.51
	probit-GP	64	91	82	57	87	79	0.45
	DT	76	95	88	71	81	78	0.48
	RF	100	100	100	59	91	82	0.53
	SVM	64	97	85	60	86	79	0.45
	probit-PLS	64	88	79	62	86	79	0.53
	ref 7 (SIMCA)	85	81	82	74	71	72	0.38
ER	GP Classifier	90	84	86	71	85	79	0.56
	probit-GP	95	86	90	73	84	79	0.57
	DT	94	95	94	69	88	81	0.58
	RF	100	100	100	79	84	82	0.61
	SVM	98	92	95	73	85	80	0.58
	probit-PLS	77	83	80	61	85	76	0.47
	ref 7 (SFGA)	87	76	81	77	80	79	0.56

[a] Where available, the result for the best model in the corresponding reference is also provided for comparison, and the method used is noted.

Table 4. BBB Penetration: Comparison of Classification Models

		accuracy training (%)			accuracy test (%)			test	
method	number desc.	high	low	overall	high	low	overall	κ	AUC
Zhao et al. Models[5]									
binomial PLS	5	95	82	92	98	80	97	na	na
RP	5	93	85	91	98	78	96	na	na
fragments (binPLS)[a]	69	98	94	97	98	88	97	na	na
This Work Classification Techniques, Zhao Descriptors									
GP classifier	18	99	80	94	100	73	97	0.81	0.90
probit-GP	18	99	78	94	100	76	97	0.84	0.92
DT	10	99	86	96	100	78	98	0.85	0.88
RF	18	100	100	100	100	80	98	0.88	0.95
SVM	18	99	82	95	100	76	97	0.84	0.98
probit-PLS	18	99	75	93	100	76	97	0.84	0.89

[a] This model uses different descriptors from the other models.

Where there is a small number of compounds in one class, the accuracy of the GP classifier typically appears to be better than other models for the under-represented class, most notably for the hERG, HIA, and DHFR data sets. Biased data sets arise in many cases; for example, where few active compounds have been identified experimentally or where data is only available for compounds that have progressed downstream and, hence, is heavily biased toward 'successful' compounds. In these situations, it can often be difficult to produce good predictions for the under-represented class.

There does not appear to be a significant difference in overall model accuracy between the ADMET and the activity data sets, with the exception of the BBB model which has a κ value in excess of 0.8, indicating very good agreement. This may be due to the much larger number of compounds in this data set relative to the others or the use of more physicochemically relevant descriptors rather than simple structural descriptors.

Modeling BBB Penetration. The results of modeling the BBB data set are summarized in Table 4. For comparison we have also included the reported performance statistics for the models published by Zhao et al.[5] They have developed a variety of models using different subsets of descriptors and different modeling techniques. The models' accuracy in predicting low compounds ranged from 65 to 88% on the test set, with overall accuracies from 97 to 100%. We have included their best models generated by the binomial PLS method and recursive partitioning (RP) using subsets of 19 simple descriptors and their best reported model built using a fragmentation scheme descriptors.

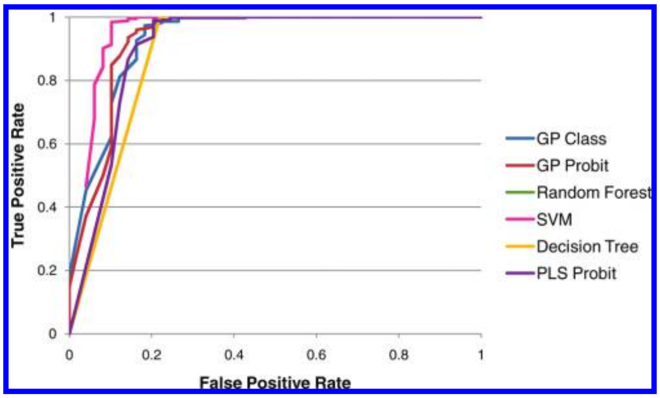

Figure 2. ROC curves for predictions of BBB penetration on the test set for six different models.

The data set contains a larger proportion of high than low compounds. Therefore, we would expect it to be much more difficult to accurately predict low compounds, and the accuracy in predicting low is the most important statistical measure for this set (this is also reflected in the κ statistic).

Overall, all of the models are good and compare well to models produced by Zhao et al. The best models in this study were produced by the RF and DT methods, which achieve 78% accuracy in low class. Although they are not the best models, the GP classifier and the GP probit model are comparable in performance to other models using these descriptors. For comparison, the best Zhao et al. model using the same set of descriptors achieved 80% accuracy in predicting low, and the best model using fragmentation schemes achieved 88% accuracy in low.

Another important performance measure to look at is the area under the ROC curve (AUC) (the last column in Table 4) and also the shape of the ROC curve. Figure 2 shows ROC curves on the test set for five models developed in this work. The SVM model produced the best curve, and the DT model has the worst ROC curve, despite having performance statistics slightly better than that of the SVM model.

All of the techniques used in this work provide fully probabilistic output apart from the DT technique. The DT technique also provides probabilities of belonging to a class based on Laplace ratio,[9] but it is not appropriate in this case to change the probability threshold to assign class membership (as done when constructing ROC curves). Therefore, the DT model is actually represented by a point and not a curve. If for other models we would ignore the probabilistic output and use just the class membership information for ROC curves, the classifiers would be represented by single points, as for the DT model.

In addition to achieving a high accuracy of prediction, a good classifier should provide an estimate of confidence in that prediction, i.e., a probability of belonging to the predicted class. The GP classifier, RF, SVM, and probit models provide an individual probability for each compound. The probabilistic output of DT models is much less sophisticated and accurate than the other techniques used in this work, as reflected in the ROC curve for the DT model (see Figure 2) and smallest AUC (0.88).

For each model, we have considered compounds from the test set which were 'confidently' predicted, that is they have predicted probability for the assigned class of greater than 0.75. We expect that performance of the model on such a subset should be better than on all the data. The results are summarized in Table 5. Comparing the κ statistic and the accuracy for low compounds evaluated on all the test set data (from Table 4) to statistics on 'confident' compounds we can see that only the GP classifier, RF, and SVM models have improved performance. For these compounds, the RF model remains the most predictive.

Table 5. Evaluation of Performance of BBB Models on 'Confident' Compounds From the Test Set

method	number of compounds	accuracy in low (%)	κ	improved?
GP classifier	489	76	0.84	yes
probit-GP	479	74	0.84	no
DT	479	69	0.79	no
RF	493	84	0.90	yes
SVM	492	82	0.89	yes
probit-PLS	476	65	0.77	no

Comparison with Existing BBB Models. In recent years there have been a variety of publications describing predictive models for BBB penetration. They used a variety of different descriptors and computational approaches. Some concentrated on continuous modeling of the logarithm of the brain: blood partition coefficient (log BB), others on classification models predicting high/low. It is beyond of the scope of this paper to provide a comprehensive review of work on classifying BBB penetration, but we refer to a recent review by Clark[21] and the works of Li et al.,[22] Zhao et al.,[5] and Kortagere et al.[23] In general, the accuracy in predicting penetrating BBB compounds exceeds the accuracy in predicting nonpenetrating compounds. Overall, the accuracy in predicting low ranges from 61 to 87% and the accuracy in predicting high ranges from 79 to 99%, and therefore, these models are broadly comparable to those in this study.

Modeling hERG Inhibition. The results of modeling the hERG data set are summarized in Table 6. Again, as in the case of the BBB set, the initial hERG data set contains a larger proportion of high compounds than low. Therefore, the predictions for high compounds are expected to be less accurate than for low compounds. The best model was produced by the GP classifier, which achieved a 65% accuracy in the low class and $\kappa = 0.66$. The κ statistic indicates that this model achieved good agreement between predicted and observed values. The performance of the other models is worse than of GP classifier judging by the κ statistic.

Figure 3 shows ROC curves constructed on the hERG test set compounds for the six models. The GP classifier and probit-GP model have the best curves, and the DT model has the worst ROC curve. As we discussed in relation to the BBB models, this is a reflection of drawbacks of the probabilistic output of DT technique.

To evaluate whether the predicted probabilities provide good estimates of the confidence in prediction for the hERG inhibition models, we performed the same analysis as for the BBB models above. The performance statistics were calculated only for 'confident' compounds from the test set for each model. As before, we have considered a compound as 'confident' if the predicted probability for the assigned class is greater than 0.75. The results are summarized in Table 7. Comparing Tables 6 and 7, one can see that RF, SVM, probit-GP and probit-PLS models have improved

Table 6. Modeling hERG Inhibition: Comparison of Classification Models

		accuracy training (%)			accuracy test (%)			test	
method	number desc.	high	low	overall	high	low	overall	κ	AUC
GP classifier	157	96	82	92	97	65	86	0.66	0.84
probit-GP	157	99	71	91	100	53	84	0.60	0.85
DT	7	98	82	93	91	65	82	0.58	0.78
RF	157	100	100	100	100	47	82	0.54	0.81
SVM	157	99	47	84	100	47	82	0.54	0.84
probit-PLS	157	95	44	81	97	41	78	0.43	0.79

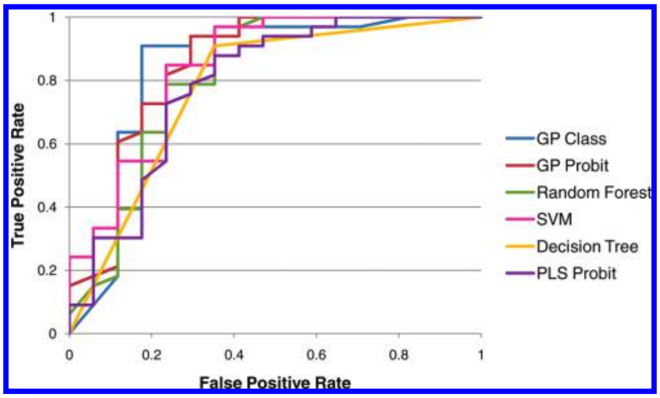

Figure 3. ROC curves for predictions of hERG inhibition on the test set for six different models.

Table 7. Evaluation of Performance of hERG Models on 'Confident' Compounds from the Test Set

method	Number of cpds	accuracy for inactives (%)	κ	improved?
GP classifier	42	64	0.66	no
probit-GP	36	60	0.68	yes
DT	37	14	0.21	no
RF	31	56	0.63	yes
SVM	39	54	0.63	yes
probit-PLS	32	50	0.60	yes

performance for 'confident' predictions. However, despite the improvement, the probit-PLS model is still not a very good model. GP classifier performance remained unchanged when considering only 'confident' predictions. Interestingly, the performance of the DT model on 'confident' predictions is significantly worse than the overall performance on all compounds in the test set, indicating that the assignment of confidence by this method has little meaning in this case.

Comparison with Published hERG Models. A good summary of published classification QSAR models for hERG inhibition is provided by Thai and Ecker.[24] Different values of IC_{50} are proposed in the literature as thresholds to separate compounds into actives and inactives. The most commonly used thresholds are 1 and 10 μM. In our model, we have used $IC_{50} = 10$ μM as a threshold. Thai and Ecker utilized a binary QSAR method to build a classification model with 10 μM threshold, achieving a 75% overall accuracy on a test set of 64 compounds, 86% accuracy in predictive active compounds, and 55% accuracy in predicting inactive compounds.

CONCLUSIONS

This study suggests that there is no single 'best' approach to classification problems. However, Gaussian Process (GP) approaches to classification are comparable in accuracy to those produced by the random forest (RF) and support vector machine (SVM) methods that are widely considered to represent the state of the art in classification modeling. Furthermore, in some cases the GP approaches produce more accurate models and, hence, seem worthy additions to the armory of methods available to tackle classification quantitative structure−activity relationship (QSAR) problems.

In common with GP regression techniques, GP classification methods have a number of advantages; the confidence in prediction is estimated for each individual compound, as Bayesian methods, they are robust to overtraining, and they require no user-determined parameters, meaning that they may be used as part of an automated model-building scheme. One disadvantage of the intrinsic GP classification methods (but not GP-probit) is the computational complexity of the algorithm, which scales as $O(N^3(N + K))$, where N is the number of compounds in the training set, and K is the number of descriptors. For comparison, the computational complexity of GP-probit models can range from $O(N^3)$ to $O(N^4)$ depending on the chosen hyperparameter optimization technique. This means that the GP classification method may be impractical for very large data sets, including large numbers of descriptors.

We have also investigated the correspondence of the estimated confidence in prediction with accuracy. The results have been mixed, showing only a small improvement in accuracy, if any. This suggests that, while the probabilistic methods such as RF and GP capture the uncertainty due to the variation in model fit, they miss a significant source of variability. It is likely that one missing source of variability is the influence of additional descriptors that is not captured by the training set, for example, functionalities that are not well represented in the training set molecules. This illustrates a limitation in currently available descriptors, namely that typically each individual descriptor has a low correlation with the observable being modeled. This means that many descriptors are often required to build a highly predictive model, and the transferability of models to new molecules containing previously unseen structural motifs can be limited. Advanced 'machine learning' techniques such as RFs and GPs may have now reached the limit of predictive power with the currently available descriptor sets.

ACKNOWLEDGMENT

The authors would like to thank Christopher Leeding and Ed Champness for their input into this manuscript.

Supporting Information Available: Structures in SMILES format, observed values and reference sources for 168 compounds of the hERG data set. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Obrezanova, O.; Csányi, G.; Gola, J. M. J.; Segall, M. D. Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties. *J. Chem. Inf. Model.* **2007**, *47*, 1847–1857.

(2) Burden, F. R. Quantitative Structure-Activity Relationship Studies Using Gaussian Processes. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 830–835.

(3) Schwaighofer, A.; Schroeter, T.; Mika, S.; Laub, J.; Laak, A. T.; Sulzle, D.; Ganzer, U.; Heinrich, N.; Muller, K. R. Accurate Solubility Prediction with Error Bars for Electrolytes:A Machine Learning Approach. *J. Chem. Inf. Comput. Sci.* **2007**, *47*, 407–424.

(4) Obrezanova, O.; Gola, J. M. R.; Champness, E. J.; Segall, M. D. Automatic QSAR Modeling of ADME Properties: Blood-Brain Barrier Penetration and Aqueous Solubility. *J.Comput.-Aided Mol. Des.* **2008**, *22*, 431–440.

(5) Zhao, Y. H.; Abraham, M. H.; Ibrahim, A.; Fish, P. V.; Cole, S.; Lewis, M. L.; de Groot, M. J.; Reynolds, D. Predicting Penetration Across the Blood-Brain Barrier from Simple Descriptors and Fragmentation Schemes. *J. Chem. Inf. Model.* **2007**, *47*, 170–175.

(6) Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Butina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. Evaluation of Human Interstinal Absorption Data and Subsequent Derivation of a Qsantitative Structure-Activity Relationship (QSAR) with the Abraham Descriptors. *J. Pharm. Sci.* **2001**, *90*, 749–784.

(7) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Sline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure-Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906–1915.

(8) Schwaighofer, A.; Schroeter, T.; Mika, S.; Hansen, K.; Laak, A. T.; Lienau, P.; Reichel, A.; Heinrich, N.; Muller, K. R. A Probabilistic Approach to Classifying Metabolic Stability. *J. Chem. Inf. Model.* **2008**, *48*, 785–796.

(9) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kauffman Publishers, Inc.: San Mateo, CA, 1993.

(10) Cristianini, N.; Shawe-Taylor, J. *An introduction to Support Vector Machines and other kernel-based learning methods*; Cambridge University Press: Cambridge, U.K., 2000.

(11) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

(12) Wold, S.; Sjstrm, M.; Eriksson, L. Partial Least Squares Projections to Latent Structures (PLS) in Chemistry. In *The Encyclopedia of Computational Chemistry*, Schleyer, P., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P., Schaefer , H. F., III, Schreiner, P. R., Eds.; Wiley: Chichester, U.K., 1998; *Vol. 3*.

(13) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press: Cambridge, Masachusetts, 2006.

(14) Gibbs, M.; MacKay, D. J. C. Variational Gaussian Process Classifiers. *IEEE Trans. Neural Network* **2000**, *11*, 1458–1464.

(15) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numercial Recepies in C: The Art of Scientific Computing*; Cambridge University Press: Cambridge, U.K., 1988.

(16) *StarDrop*, v 4.2.1; Optibrium Ltd.: Cambridge, U.K., 2010.

(17) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutmann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **2009**, *11*, 10–18.

(18) Chang, C.-C.; Lin, C.-J. *LIBSVM: a library for support vector machines*; http://www.csie.ntu.edu.tw/ cjlin/libsvm. Accessed March 19, 2010.

(19) Adenot, M.; Lahana, R. J. Blood-Brain Barrier Permeation Models: Discriminating Between Potential CNS and non-CNS Drugs Including P-Glycoprotein Substrates. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 239–248.

(20) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750.

(21) Clark, D. E. In Silico Prediction of Blood-Brain Barrier Permeation. *Drug Discovery Today* **2003**, *8*, 927–933.

(22) Li, H.; Yap, C. W.; Ung, C. Y.; Xue, Y.; Cao, Z. W.; Chen, Y. Z. Effect of Selection of Molecular Descriptors on the Prediction of Blood-Brain Barrier Penetrating and Nonpenetrating Agents by Statistical Learning Methods. *J. Chem. Inf. Model.* **2005**, *45*, 1376–1384.

(23) Kortagere, S.; Chekmarev, D.; Welsh, W. J.; Ekins, S. New Predictive Models for Blood-Brain Barrier Permeability of Drug-Like Molecules. *Pharm. Res.* **2008**, *25*, 1836–1845.

(24) Thai, K.-M.; Ecker, G. F. A Binary QSAR Model for Classification of hERG Potassium Channel Blockers. *Bioorg. Med. Chem.* **2008**, *16*, 4107–4119.

CI900406X