Gaussian Processes for Classification: Modeling of Blood-Brain Barrier Penetration and hERG Inhibition

Olga Obrezanova and Matthew D. Segall*

Optibrium Ltd., 7226 IQ Cambridge, Beach Drive, Cambridge, CB25 9TL, U.K.

Abstract

In this article we extend the application of Gaussian Processes technique to classification 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 blood-brain barrier penetration and hERG inhibition. We also compare performance of Gaussian Processes for classification to other known computational methods, namely decision trees, bagging and probit PLS.

1 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; classification models predict if a molecule will fall into a property class (e.g. whether 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] We have applied this method to build models of various ADME 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 Gaussian Processes 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

*Corresponding author phone: +44(0)1223 815900; e-mail: info@optibrium.com the underlying theory and compare the performance of these two methods with other classification techniques using two example data sets, a blood-brain barrier classification previously published by Zhao *et al.*[5] and hERG inhibition. A different approach for classification using Gaussian Processes was recently used by Schwaighofer *et al.* to model metabolic stability.[6]

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 Section 2. Section 3 details results for the two example data sets and compares the GP methods with other methods. Finally, Section 4 summarizes the results of this study and draws some conclusions.

2 METHODS AND DATA

2.1 Gaussian Processes Binary Classifier

Here we will briefly describe the underlying method for generating classification models using GPs and give the final formulae to enable the reader to implement this technique. For detailed discussions and derivation of the formulae we refer the reader to the recent book by Rasmussen and Williams[7] and the work of Gibbs and Mackay.[8]

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

2.1.1 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, it will ultimately be integrated out.

The latent function $f(x)$ 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 $f^\star$ at a new point $\mathbf{x}^\star$, given all of the training data

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

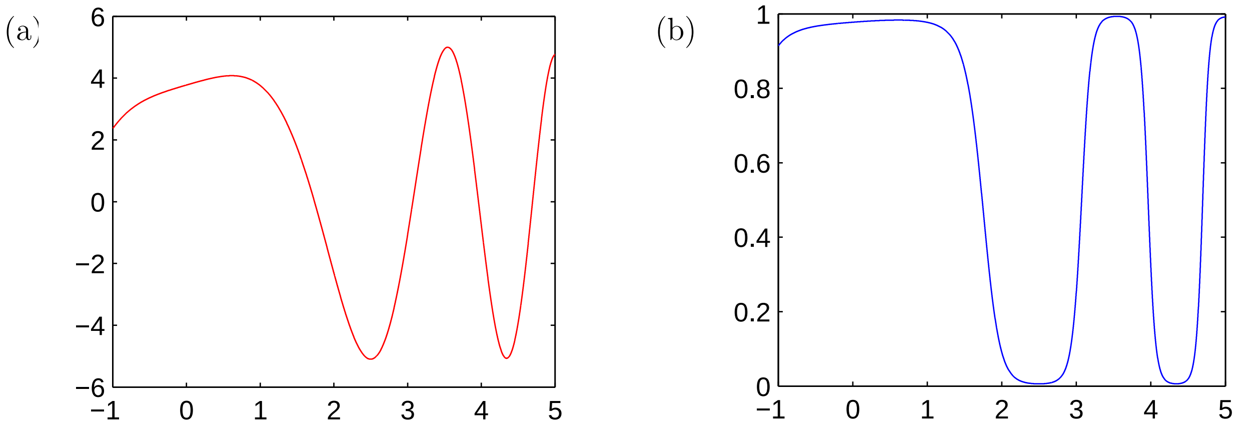


Figure 1: Graphical illustration to Gaussian Processes classification for the case of a one-dimensional descriptor space. Graph (a) shows a latent function drawn from a Gaussian Process. Graph (b) shows the result of "squashing" this function through the logistic response function (eq 1) to obtain the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

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 $f^\star$ (eq 3) and integrate over the logistic response function (eq 2) to obtain the probability of class membership:

$$p(y=+1|\mathbf{x}^\star) = \int g(f^\star) p(f^\star|\mathbf{X},\mathbf{Y},\mathbf{x}^\star)\,\mathrm{d}f^\star. \tag{5}$$

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 integrals must be used. Two such methods are described by Rasmussen and Williams[7] and one of these approaches, the method of 'expectation propagation' was used by Schwaighofer *et al.* in modeling metabolic stability.[6] In this work we have used the method of variational lower and upper bounds suggested by Gibbs and Mackay.[8] 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 parameterized 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 $f^\star$ (eq 3) and for the class probability (eq 5).

2.1.2 Final Formulae

For the sake of computational efficiency, we have used only the lower bound approximation in our implementation. The final formulae are summarized below. The details of the derivation of the formulae can be found in Gibbs and Mackay.[8]

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

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

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 $f^\star$ for a new point $\mathbf{x}^\star$ is a Gaussian distribution with mean a and variance σ^2;

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

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

where the vector $\mathbf{k}$ with components $k_n = C(\mathbf{x}^\star, \mathbf{x}^{(n)})$ 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 $(n = 1 \ldots N)$, 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 $f^\star$ and eq 5 are then used to derive the probability of a new compound $\mathbf{x}^\star$ belonging to class +1:

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

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.

2.1.3 Optimization of Hyperparameters and Variational Parameters

Learning a GP classifier means finding hyperparameters θ_1, θ_2 and r_i $(i = 1 \ldots K)$ and variational parameters ν_n $(n = 1 \ldots N)$. The variational parameters are optimized to ensure that the lower bound on $P(\mathbf{Y}|\mathbf{X}, \Theta)$ is as tight 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 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[9] to optimize the function $\log Z$ (eq 12).

2.2 Probit Gaussian Processes

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 in Section 2.1 where we applied a logistic transformation to the latent function to obtain class membership probabilities (eqs 1, 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 $\mathbf{x}^\star$ by $a^\star$ and the standard deviation in the prediction by $\sigma^\star$. 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

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

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.

2.3 Other Classification Techniques

We will compare performance of the GP classifier and probit models with other modeling methods, namely decision trees (DT), bagging and a probit treatment of a PLS model.

2.3.1 Decision Trees

The DT technique applies a recursive partitioning approach to building classification models. Here we used the DT technique implemented within StarDrop's Auto-Modeller software,[10] which is based on the C4.5 algorithm introduced by Quinlan.[11] 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, one example of which is given in Section 2.3.2.

The DT method can estimate probabilities of belonging to a class based on the Laplace ratio[11] obtained using the results from the training and test sets. The probabilities are determined for each leaf of the decision tree and so all compounds in a leaf will be assigned the same probabilities.

2.3.2 Bagging

In this work we used the simplest tree ensemble method, bagging,[12,13] where an ensemble is created by training multiple trees on different subsets of original data. To create subsets molecules from the training set are sampled with replacement, a sampling technique called 'bootstrapping'. 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 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. We have used an ensemble of 100 trees in the models generated in this study.

Bagging is a special case of the random forest technique in which each tree is built using a random subset of descriptors.

2.3.3 Probit PLS Model

The probit treatment which we described in Section 2.2 can be applied to a model produced by any regression technique. For comparison, in this study we also use the Partial Least Square (PLS) method[14] to build continuous models and apply a probit treatment to produce classifications.

PLS does not provide an estimate of the uncertainty $\sigma^\star$ in prediction, which we need to be able to calculate class probability $p(y = +1|\mathbf{x}^\star)$ (eq 14). Therefore, for a PLS model, the uncertainties are calculated from the actual RMSE on the independent test sets, taking into account whether a new compound lies in the chemical space of the model or outside. 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-Modeller.[10] Together with each prediction, models built by the Auto-Modeller produce an uncertainty in prediction, standard deviation $\sigma^\star$, as described above.

2.4 Data Sets

To make a comparison between GPs for classification and other classification techniques we have chosen two data sets; a BBB dataset previously published and modeled by Zhao *et al.*[5] and a hERG inhibition data set compiled in house and derived from various literature sources.

2.4.1 Blood–Brain Barrier Data Set (BBB)

This data set contains 1593 compounds; 1283 BBB+ compounds (penetrating blood-brain barrier) and 310 BBB- compounds (little ability to penetrate blood-brain barrier).[5] The data set is originally based on that published by Adenot and Lahana.[15] Classification of BBB+ compounds was assigned on the basis of their central nervous system (CNS) activity. Identifying BBB- compounds is more complicated and we refer to Adenot and Lahana[15] 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, logP, logD, 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 split 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 (see Section 2.4.2).

2.4.2 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 $pIC_{50} = 5$ ($IC_{50} = 10\ \mu$M) to classify compounds into two classes, i.e. compounds with $pIC_{50} \leq 5$ are considered inactive (class '-') and compounds with $pIC_{50} > 5$ are active (class '+'). The final data set contains 117 active compounds ('+') and 51 inactive compounds ('-'). The data set and references to literature sources are provided in the Supporting Information.

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

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

The initial data set was 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 the Tanimoto similarity index. Compounds were clustered using an unsupervised non-hierarchical clustering algorithm developed by Butina.[16] 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 pre-selection 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.

2.5 Evaluation of Model Performance

To measure the performance of classification models we use the kappa statistic as well as the overall accuracy and accuracies for individual classes. The kappa 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:

$$\begin{array}{cccc} & & \text{Predicted} & \\ & & \textit{Class -1} & \textit{Class 1} \\ \text{Observed} & \textit{Class -1} & TN & FP \\ & \textit{Class 1} & FN & TP \end{array} \tag{15}$$

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

$$\kappa = \frac{TN + TP - \eta}{N - \eta}, \tag{16}$$

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

A kappa statistic exceeding 0.8 would mean very good agreement and $0.6 \leq \kappa < 0.8$ 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.

3 RESULTS

3.1 Blood-Brain Barrier Penetration

The results of modeling the BBB data set are summarized in Table 1. 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 BBB- 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.

Table 1: Blood-Brain Barrier Penetration: Comparison of Classification Models

		accuracy (%) train[a]			accuracy (%) test[b]			test	test
method	desc.	+	-	overall	+	-	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)[c]	69	98	94	97	98	88	97	na	na
	This work classification techniques, Zhao descriptors								
DT	10	99	86	96	100	78	98	0.85	0.88
Bagging	18	99	88	97	100	78	98	0.85	0.96
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
probit-PLS	18	99	75	93	100	76	97	0.84	0.89

[a] Training set 1092 compounds.
[b] Test set 499 compounds.
[c] This model is using different descriptors from the other models.

The data set contains a larger proportion of BBB+ compounds than BBB- compounds. Therefore, we would expect it to be much more difficult to accurately predict BBB- compounds and therefore the accuracy in predicting BBB- 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 Bagging and DT methods, which achieve 78% accuracy in BBB- 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 BBB-, and the best model using fragmentation schemes achieved 88% accuracy in BBB-.

Another important performance measure to look at is the area under the ROC curve (AUC) (the last column in Table 1) 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 Bagging model produced the best curve and the DT model has the worst ROC curve, despite having performance statistics equal to the Bagging model. All the techniques used in this work provide fully probabilistic output apart from the decision trees technique. The DT technique also provides probabilities of belonging to a class based on Laplace ratio,[11] 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.

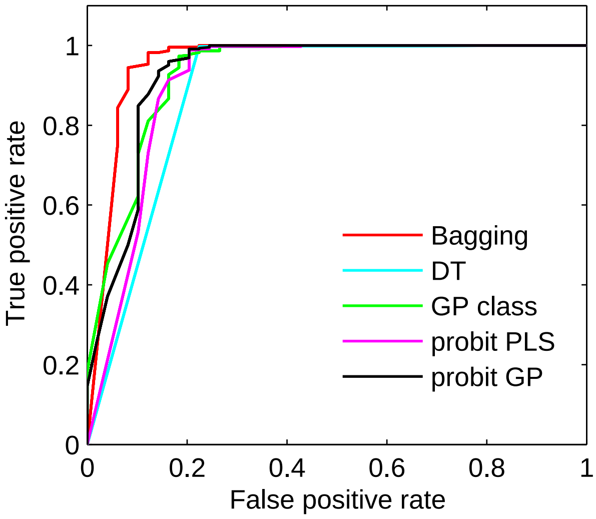


Figure 2: ROC curves for predictions of BBB+/- on the test set for five different models.

Table 2: Evaluation of Performance of BBB Models on 'Confident' Compounds From the Test Set[a]

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

[a] Test set contains 499 compounds.

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, Bagging 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 in the interval $[0, 0.25]$ or $[0.75, 1]$. 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 2. Comparing the κ statistic and accuracy for BBB- compounds evaluated on all the test set data (from Table 1) to statistics on 'confident' compounds we can see that only the Bagging and GP classifier models have improved performance.

3.1.1 Comparison with Existing BBB Models

In recent years there have been a variety of publications describing predictive models for blood-brain barrier 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 (logBB), others on classification models predicting BBB+/-. 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,[17] and the works of Li *et al.*,[18] Zhao *et al.*[5] and Kortagere *et al.*[19] In general, the accuracy in predicting penetrating BBB compounds exceeds the accuracy in predicting non-penetrating compounds. Overall, the accuracy in predicting BBB- ranges from 61% to 87% and the accuracy in predicting BBB+ ranges from 79% to 99% and therefore these models are broadly comparable to those in this study.

3.2 Modeling hERG Inhibition

The results of modeling the hERG data set are summarized in Table 3. Again, as in the case of the BBB set, the initial hERG data set contains a larger proportion of active compounds than inactive. Therefore, the predictions for active compounds are expected to be less accurate than for active compounds. The best model was produced by the GP classifier, which achieved a 65% accuracy in the inactive class and $\kappa = 0.66$. The κ-statistic indicates that this model achieved good agreement between predicted and observed values. The performance of other models is worse than of GP-classifier judging by the κ-statistic.

Table 3: Modeling hERG Inhibition: Comparison of Classification Models

		training set[a]				test set[b]				
		accuracy (%)				accuracy (%)				
method	desc.	inactive	active	all	κ	inactive	active	all	κ	AUC
DT	7	82	98	93	0.82	65	91	82	0.58	0.78
Bagging	157	100	100	100	1.00	59	97	84	0.61	0.85
GP classifier	157	82	96	92	0.81	65	97	86	0.66	0.84
probit-GP	157	71	99	91	0.75	53	100	84	0.60	0.85
probit-PLS	157	44	95	81	0.45	41	97	78	0.43	0.79

[a] Training set 118 compounds.
[b] Test set 50 compounds.

Figure 3 shows ROC curves constructed on the hERG test set compounds for the five models. The Bagging model and probit-GP model have the best curves and the DT model has the worst ROC curve. As we discussed in Section 3.1, 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 in Section 3.1. 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 lies in interval $[0, 0.25]$ or $[0.75, 1]$. The results are summarized in Table 4. Comparing Tables 3 and 4 one can see that Bagging, probit-GP and probit-PLS models have improved 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.

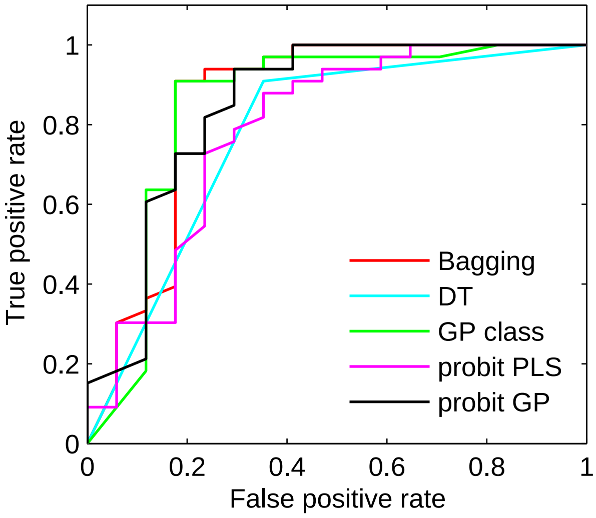


Figure 3: ROC curves for predictions of hERG inhibition on the test set for five different models.

Table 4: Evaluation of Performance of hERG Models on 'Confident' Compounds from the Test Set[a]

method	number of cpds	accuracy for inactives (%)	κ	improved?
DT	37	14	0.21	no
Bagging	36	63	0.72	yes
GP classifier	42	64	0.66	no
probit-GP	36	60	0.68	yes
probit-PLS	32	50	0.60	yes

[a] Test set contains 50 compounds.

3.2.1 Comparison with Published hERG Models

A good summary of published classification QSAR models for hERG inhibition is provided by Thai and Ecker.[20] 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 1μM and 10μM. In our model we have used $IC_{50} = 10\mu$M as a threshold. Thai and Ecker utilized a binary QSAR method to build a classification model with $10\mu M$ 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.

4 CONCLUSIONS

This study suggests that GP approaches to classification are comparable in accuracy to those produced by the DT ensemble methods that are widely considered to represent the state of the art in classification modeling. 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 bagging and GPs 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 DT ensembles and GPs may have now reached the limit of predictive power with the currently available descriptor sets.

ACKNOWLEDGEMENT

We would like to thank Christopher Leeding for proof-reading the paper.

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

(1) Obrezanova, O.; Csányi, G.; Gola, J.M.R.; 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. Model.* **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.P. Predicting Penetration Across the Blood-Brain Barrier from Simple Descriptors and Fragmentation Schemes. *J. Chem. Inf. Model.* **2007**, *47*, 170–175.

(6) 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.

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

(8) Gibbs, M.; Mackay, D. J. C. Variational Gaussian Process Classifiers. *IEEE Transactions on Neural Networks.*, **2000**, *11*, 1458–1464.

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

(10) *StarDrop, Version 4.2.1*; Optibrium Ltd.: Cambridge, U.K.

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

(12) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140.

(13) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.

(14) Wold, S.; Sjöström, M.; Eriksson, L. Partial Least Squares Projections to Latent Structures (PLS) in Chemistry. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. von R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P.; Schaefer III, H. F.; Schreiner, P. R., Eds.; Wiley: Chichester, 1998; Vol. 3, pp. 2006–2022.

(15) 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.

(16) 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.

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

(18) 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.

(19) 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.

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

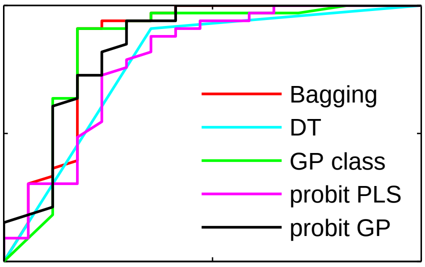
For Table of Contents Use Only

Gaussian Processes for Classification: Modeling of Blood-Brain Barrier Penetration and hERG Inhibition

Olga Obrezanova and Matthew D. Segall*

BBB +/−
1591 cpds

hERG active/ inactive
168 cpds


Bagging
DT
GP class
probit PLS
probit GP