Improved Naïve Bayesian Modeling of Numerical Data for Absorption, Distribution, Metabolism and Excretion (ADME) Property Prediction

Anthony E. Klon,* Jeffrey F. Lowrie, and David J. Diller

Department of Molecular Modeling, Pharmacopeia Drug Discovery, Inc., P.O. Box 5350, Princeton, New Jersey 08543-5350

Received April 10, 2006

We have implemented a naïve Bayesian classifier which models continuous numerical data using a Gaussian distribution. Several cases of interest in the area of absorption, distribution, metabolism, and excretion prediction are presented which demonstrate that this approach is superior to the implementation of naïve Bayesian classifiers in which continuous chemical descriptors are modeled as binary data. We demonstrate that this enhanced performance, upon comparison with other implementations, is independent of the descriptor sets chosen. We also compare the performance of three implementations of naïve Bayesian classifiers with other previously described models.

INTRODUCTION

Bayes's rule of conditional probability[1] is a widely used method of statistical inference which has been applied to many real-world problems, including, but not limited to, pattern recognition, machine learning, medicine, public health,[2,3] text classification,[4] traffic safety,[5] bioinformatics,[6,7] and cheminformatics.[8,9] Specific cheminformatics applications of naïve Bayesian classifiers have included the creation of quantitative structure−activity relationship (QSAR) models;[8] the prediction of protein activity inhibition by small molecules;[9] absorption, distribution, metabolism, and excretion (ADME) prediction;[10] prioritizing the hits from high-throughput screening campaigns;[11,12] and enrichment of the results from high-throughput docking.[13,14]

Bayes' theorem relates the probability that a hypothesis is true given some evidence to the probability of the evidence when the hypothesis is true. In the case of cheminformatics problems, the evidence for a given molecular structure may be a set of descriptors such as molecular fingerprints, the molecular weight, the log of the octanol/water partition coefficient (logP), polar surface area (PSA), the number of rotatable bonds, or the number of hydrogen-bond donors and acceptors, whereas the hypotheses are whether (or not) a compound will have a particular biological activity such as protein kinase inhibition.[9] In the specific cases presented in this paper, the biological activities we wish to predict are the probabilities of passive intestinal absorption (PIA), blood−brain barrier penetration (BBB), and serum protein binding (SPB) for a given compound. On the basis of the distributions of active and inactive compounds in a given descriptor space, the probability a compound will appear in a particular area of the descriptor space given that its activity can be estimated from a training set is derived. Bayes' theorem allows this to be converted to the probability a compound will be active or inactive given its location in the descriptor space. Because of the curse of dimensionality, it is not practical to estimate the probability densities in the full descriptor space. As a result, we make the "naïve" assumption that the descriptors impact the activity independently from one another. This simplified approach has been shown consistently to be comparable or even superior to other, more complicated machine-learning techniques such as support vector machines, decision-tree induction, or neural networks.[15,16] Bayesian classifiers have also been shown to be superior to logistic regression modeling. It has been previously shown that, although discriminative learning approaches such as logistic regression have a lower asymptotic error, Bayesian classifiers tend to approach their asymptotic error at a much higher rate.[17] As a result, logistic regression models are prone to overfitting to the training data for a small number of observations. The implication is that, as the number of observations in the training set increases, a Bayesian classifier will initially be superior to logistic regression, but logistic regression methods would be predicted to match and eventually even surpass the performance of Bayesian classifiers with a large number of observations.[17] Because it is not possible to know a priori how many observations are required in the training set to yield superior performance when using logistic regression in a particular case, naïve Bayesian classifiers are a reasonable choice to model training data sets consisting of a few hundred compounds.

Previous studies have shown that naïve Bayesian classifiers are tolerant of noisy data sets.[11] Naïve Bayesian categorization thus provides an attractive solution to the problem of predicting biological activity because of the algorithm's robustness when dealing with the extremely noisy data sets faced by researchers in the pharmaceutical industry. In particular, the prediction of ADME properties may be quite problematic. The data sets used to train machine-learning algorithms in the prediction of human passive intestinal absorption are fraught with pitfalls. Even a cursory review of the literature reveals confounding problems such as oral bioavailability and intestinal absorption being used interchangeably. Many published reports do not distinguish between actively and passively transported drugs. Furthermore, compounds with measured intestinal absorption reported in the literature often come from different laboratories,

* Corresponding author tel.: 609-452-3676; fax: 609-655-4187; e-mail: aklon@pcop.com.

10.1021/ci0601315 CCC: $33.50
Published on Web 07/21/2006

raising the concern about reproducibility of the reported values. Indeed, several and sometimes widely differing values for human intestinal absorption may be reported by different groups for a single compound. Another test case for the use of naïve Bayes is the prediction of serum protein binding. The prediction of whether a given compound is likely to bind to human serum albumin is determined on the basis of its similarity to compounds which are known binders. The predominant descriptors in this case are the chemical fingerprints of the compounds in the data set. This provides another rationalization for the selection of naïve Bayes because the classifier has been previously shown to perform well with chemical fingerprints such as the extended connectivity fingerprints and functional connectivity fingerprints (FCFP) available in the software package Pileline Pilot.[18] Bayesian classifiers' tolerance of noisy data provides another advantage over logistic regression models. For a small set of training compounds, logistic regression models are more prone to overfitting to noise in the data set than Bayesian models.

We use the prediction of several ADME properties as specific cases to construct and test our approach to building naïve Bayesian models. PIA contributes greatly to the bioavailability of a drug. Once absorbed into the bloodstream, other factors such as SPB become important in determining whether a particular drug will reach its intended target under physiological conditions. BBB is also an important quantity impacting the effectiveness or toxicity of the given compound. Certain drugs, such as those intended to bind to targets in the central nervous system, are obviously only effective if they are capable of crossing the BBB. Furthermore, it may be desirable to ensure that certain drugs remain in the bloodstream and not cross the BBB where they may have toxic side effects. Knowledge of such properties is obviously crucial to the early success of a drug discovery program. A computational tool which can accurately predict the ADME properties of a given compound or chemical series is obviously of extreme interest in the lead optimization stage of drug discovery prior to in vivo testing. Being able to identify compounds with potentially undesirable characteristics allows medicinal chemists to concentrate their expertise on compounds which are less likely to fail at the later stages in the drug discovery process. Furthermore, understanding why a compound has undesirable ADME characteristics is just as important as knowing that it does. By using descriptors which are intuitively understandable, such as the PSA, logP, number of hydrogen-bond donors and acceptors, or the presence or absence of a particular chemical substructure, compounds with a better ADME profile can be proposed by medicinal chemists and tested prior to synthesis.

We present here an in-house implementation of a naïve Bayesian classifier which models binary as well as continuous numerical data.[15,16] This implementation makes the assumption that numerical values have a Gaussian distribution, though we acknowledge that this assumption, although reasonable, may not be entirely accurate in all cases. Indeed, outside the field of cheminformatics, there has been considerable interest in developing other Bayesian models of continuous numerical data, such as kernel density estimation.[15] We compare our implementation to the Laplacian-modified naïve Bayesian classifier in Pipeline Pilot[18] and the binary QSAR model in the Molecular Operating Environment[19] (MOE). We also compare the performance of other previously described models.[20]

METHODS

Laplacian-Modified Naïve Bayes with Continuous Numerical Data. For a given hypothesis H with one or more pieces of evidence E, Bayes' rule for conditional probability[1] is

$$\Pr[H|E] = \frac{\Pr[E|H] \times \Pr[H]}{\Pr[E]}$$

where $\Pr[H]$ is the prior probability of H being true without any knowledge of E. $\Pr[H/E]$ is the conditional probability of H being true given E. For example, H might refer to whether a given compound will show a desired biological activity such as kinase inhibition, intestinal absorption, or toxicity, while E might refer to molecular descriptors such as logP, PSA, or molecular fingerprints. If there are multiple pieces of evidence associated with a given hypothesis, then E is the combination of all pieces of evidence. The conditional probability of E given H is

$$\Pr[E|H] = \frac{A}{B}$$

where B is the total number of samples for which H is true and A is the total number of times E is observed when H is true.

If we "naively" assume that each piece of evidence E_i affects the hypothesis independently from the others, the combined probability is obtained by multiplying the individual probabilities for E_i. In the cases presented here, the hypothesis may be either BBB penetration, PIA, or SPB. For a number of pieces of evidence N, each piece of evidence E_i for a given hypothesis will be either chemical descriptors or fingerprint bits. The combined probability of an event, given the "naïve" assumption, occurring if the hypothesis is true is

$$\Pr[\text{true}|E] = \frac{\Pr[\text{true}]}{\Pr[E]} \times \prod_{i=1}^{N} \Pr[E_i|\text{true}]$$

Similarly, the combined probability of the events occurring if the hypothesis is false is

$$\Pr[\text{false}|E] = \frac{\Pr[\text{false}]}{\Pr[E]} \times \prod_{i=1}^{N} \Pr[E_i|\text{false}]$$

To calculate these Bayesian probabilities, we first estimate the numerators in the previous equations by determining the likelihood of a given hypothesis being true or false within a training set. Because we assume that the events are not correlated, the likelihood that the hypothesis is true is obtained by multiplying the probabilities of each event E_i:

$$\text{likelihood of true: } \prod_i \Pr[E_i|\text{true}] \times \Pr[\text{true}]$$

Similarly, the likelihood that the hypothesis is false is obtained by

$$\text{likelihood of false: } \prod_i \Pr[E_i|\text{false}] \times \Pr[\text{false}]$$

The probabilities can then be obtained by normalizing the likelihoods so that their sums are equal to 1:

$$\Pr[\text{true}] = \frac{\text{likelihood[true]}}{\text{likelihood[true]} + \text{likelihood[false]}}$$

$$\Pr[\text{false}] = \frac{\text{likelihood[false]}}{\text{likelihood[false]} + \text{likelihood[true]}}$$

If, however, a given E does not occur in the training set for a particular H, then likelihood(true) = 0 and likelihood(false) = 0. The probability then becomes

$$\Pr[E_i|H] = 0$$

When the individual probabilities for E_i are multiplied together, this leads to unstable behavior because the total probability becomes 0, regardless of the probabilities for other E_i:

$$\text{likelihood}[H|E]: \prod_{i=1}^{N} \Pr[E_i|H] \times \Pr[H] = 0$$

Such a situation might arise, for example, in the case of under-represented fingerprint bits. Suppose that a given feature (e.g., a carboxylic acid) occurs only once in a given data set and for a compound in the training set for which the hypothesis is false (e.g., likely to be absorbed in the intestine). The resulting probability that the hypothesis would be true for any test compound having this feature would be 0. (In our trivial example, this would lead to the rather absurd conclusion that no compounds containing a carboxylic acid will be absorbed in the intestine.) A Laplacian estimator is therefore applied by adding a value of 1 to each $\Pr[E_i|H]$ in the numerator and a value of N to the denominator, where N is the total number of pieces of evidence. This gives each E which occurs with a frequency of 0 a small, nonzero value:

$$\text{likelihood}[H|E]: \prod_{i=1}^{N} \left(\frac{A_i + 1}{B + N}\right) \times \Pr[H]$$

Although it is customary to use a value of 1 for the Laplacian estimator, there is no requirement to do so. An arbitrary weight μ can be used instead to determine how influential each of the conditional probabilities for E_i are:

$$\text{likelihood}[H|E]: \prod_{i=1}^{N} \left(\frac{A_i + \frac{\mu}{N}}{B + \mu}\right) \times \Pr[H]$$

There is no reason that μ must be divided into N equal parts in the numerator. μ may instead be multiplied by p_i:

$$\text{likelihood}[H|E]: \prod_{i=1}^{N} \left(\frac{A_i + \mu\rho_i}{B + \mu}\right) \times \Pr[H]$$

where p_i is the a priori probability of E_i.

Numerical values are handled by the assumption that they have a normal probability distribution. The probability distribution function for a normal distribution with mean μ and standard deviation σ is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

The resulting expression for the probability distribution function when H is true where E_i is equal to a particular value is therefore

$$f_i(E_i = x|\text{true}) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\mu_i)^2/2\sigma_i^2}$$

Combining the probability distribution function with our previous expression for the likelihoods gives

$$\text{likelihood of true: } \prod_b \Pr[E_b|\text{true}] \prod_n f_n(E_n = x|\text{true}) \times \Pr[\text{true}]$$

$$\text{likelihood of false: } \prod_b \Pr[E_b|\text{false}] \prod_n f_n(E_n = x|\text{false}) \times \Pr[\text{false}]$$

where b denotes binary evidence and n denotes numerical evidence.

Finally, we normalize the denominator as described previously, resulting in the probabilities for a given hypothesis being true when the evidence is a collection of binary and numerical data.

BBB-Penetration Data Set. The BBB-penetration data set of 178 compounds was taken directly from the supplementary data provided by Garg and Verma.[21] The classification of the molecules into the predefined training and test sets was retained, with 129 and 49 compounds in the training and test sets, respectively. A cutoff of observed log BB $\geq$ 0.477 was used to define the compounds most likely to cross the blood−brain barrier. This number corresponds to three times the concentration of the molecule in the brain as that in the bloodstream.

Human Passive Intestinal Absorption Data Set. A set of 500 initial compounds was taken from the medicinal chemistry literature. Care was taken to remove all compounds which were known P-glycoprotein substrates. Of the compounds which did not have a reported intestinal absorption, the PIA was inferred for compounds with a high (>90%) reported bioavailability. Low values of oral bioavailability were not used because the low value could be attributed to either poor PIA or other factors such as low metabolic stability. Compounds for which human intestinal absorption was not reported or for which there was no reasonable way to infer an accurate passive intestinal absorption were evicted from this data set. The final data set after filtering contained a total of 264 compounds from various sources.[22−29] Approximately 75% of the compounds were randomly assigned to the training set using Pipeline Pilot, with the remaining compounds assigned to the test set. The total numbers of compounds in the training and test sets were 205 and 59, respectively.

Because of the differences in the reported values of PIA in the literature, all of the naïve Bayesian classifiers studied

in this paper were tested using three cutoffs for the PIA training set (90%, 80%, and 70%) in order to define what constitutes a well-absorbed compound. This resulted in each Bayesian classifier generating three separate models for PIA. It is important to note that the resulting Bayesian probabilities are not the predicted fraction of the compound passively absorbed but are rather the probability that a particular compound will be absorbed at a level greater than the specified cutoff used in the training set.

Serum Protein Binding Data Set. The 260 compounds used for the serum protein binding data set were taken from the set used to train ADME Profiler's SPB model.[20] The compounds were randomly divided into the training and test sets using Pipeline Pilot by assigning ~75% of the compounds to the training set, resulting in 207 and 53 compounds in the training and test sets, respectively. Cutoffs of 90% and 95% in the training set were used to generate Bayesian models classifying compounds likely to bind to carrier proteins in the blood, resulting in the creation of two models for each naïve Bayes model.

Third-Party Software. Two commercial chemically aware software packages were used which implement slightly different versions of the naïve Bayesian classifier. Both MOE[19] and Pipeline Pilot[18] use a Laplacian-modified naïve Bayesian classifier, and both implementations use a binning strategy to model numerical data. However, MOE also applies a Gaussian smoothing factor to correct for the case where a compound may be close to the boundary between two bins.[8]

ADME Profiler Models. The models available in ADME Profiler are described by Cheng et al.[30] The BBB-penetration and PIA models were created using a multivariate statistics approach. The sets of training compounds for BBB penetration and PIA were used to generate ellipses delineating the 95% and 99% confidence limits when ALogP is plotted against fast PSA[30] (FPSA). The training sets for these two models are described elsewhere.[30] Compounds in the PIA test set were ranked using the FAbsT2 value calculated by ADME Profiler. FAbsT2 is the Mahalanobis distance from the center of the region in the ALogP versus FPSA plane bounded by compounds with >90% PIA. Compounds in the BBB test set were ranked using the log BbR value. Log BbR is the predicted value for the logarithm of brain concentration/blood concentration calculated by a least-median-of-squares regression to a set of training compounds. The compounds used for the BBB-penetration and PIA test sets are those described previously in this paper for testing the naïve Bayesian classifiers.

The SPB[31] model is generated by calculating the one-dimensional molecular similarity[32] of a molecule to a set of 126 training compounds from the Physician's Desk Reference[33] at the 90% and 95% levels. Each compound in the training set has a characteristic similarity threshold for both the 90% and 95% levels. If a test compound has a one-dimensional similarity greater than the predefined thresholds at either level, it is considered likely to bind to serum protein. This calculation is further modulated by the compound's calculated ALogP value. A compound is considered to be likely to bind at the 90% level for ALogP values ≥ 4.0 and is considered likely to bind at the 95% level for ALogP values ≥ 5.0.

Table 1. Mean and Standard Deviations for Descriptors from the BBB-Penetration Training Set

BBB penetrant	nHDon	nHAcc	TPSA(NO)	MLogP	MW	NRB
yes	0.66 (0.78)	2.72 (1.99)	24.34 (19.54)	3.37 (0.44)	278.10 (109.79)	3.55 (2.52)
no	1.35 (1.44)	4.05 (2.51)	52.87 (40.96)	1.98 (1.35)	247.22 (119.22)	3.13 (3.24)

Table 2. Mean and Standard Deviations for Descriptors from the PIA Training Set

cutoff	PIA	Hy	TPSA(NO)	MLogP
90%	yes	0.07 (0.9)	61.10 (35.41)	2.40 (1.39)
	no	1.90 (3.25)	107.42 (77.38)	0.76 (2.34)
80%	yes	0.21 (1.59)	64.35 (44.91)	2.29 (1.55)
	no	2.05 (3.14)	111.63 (75.80)	0.59 (2.32)
70%	yes	0.28 (1.59)	65.40 (44.92)	2.22 (1.57)
	no	2.21 (3.32)	117.17 (78.48)	0.46 (2.40)

Descriptor Selection for the BBB-Penetration Data Set. The descriptors used in a recent study to create a model of BBB penetration[21] were used to construct one model for each of the three implementations of the naïve Bayesian classifier. DRAGON[34] was used to calculate the topological polar surface area[35] for nitrogen and oxygen atoms [TPSA(NO)], the Ghose−Crippen−Viswanadhan octanol−water partition coefficient[36,37] (ALogP), the molecular weight (MW), the number of rotatable bonds (nRB), the number of hydrogen-bond acceptors (nHAcc), and the number of of hydrogen-bond donors (nHDon). Table 1 shows the mean and standard deviation of the values of the six descriptors from the training set.

In addition, we calculated sets of descriptors which were native to Pipeline Pilot, MOE, and ADME Profiler. For Pipeline Pilot, the descriptors used were SciTegic's implementations of the ALogP, MW, nHDon, nHAcc, nRB, and the polar surface area for nitrogen and oxygen atoms [PSA-(NO)]. For MOE, the additional set of descriptors used consisted of the sum of the atomic van der Waals surface area (VSA) contributions for three descriptors implemented in scientific vector language: the octanol/water partition coefficient (SlogP-VSA), molar refractivity (SMR-VSA), and Gasteiger charges (PEOE-VSA). ADME Profiler's BBB-penetration model[20,38] uses the calculated values for FPSA and ALogP for each compound. The values for the native descriptors calculated by MOE and ADME Profiler were subsequently exported and used to train our implementation of the naïve Bayesian classifier, generating two new Bayesian models.

Descriptor Selection for the PIA Data Set. A total of 1664 descriptors were calculated for the compounds comprising the PIA data set using DRAGON. Only those descriptors with a correlation of ±0.70 or more/less were retained, and this set was further reduced by removing those descriptors highly correlated with one another. The remaining descriptors were the hydrophilic factor[39] (Hy), TPSA(NO), and the Moriguchi log P[40,41] (MLogP). The Hydrophilic factor is essentially a surrogate for the number of hydrogen-bond acceptors and donors. In particular, a high correlation (0.99) was observed between Hy and nHDon. These descriptors were then used to train the three implementations of the naïve Bayesian classifier. Table 2 shows the mean and standard deviation for the values of the three descriptors calculated for the compounds in the PIA training set.

Table 3. Mean and Standard Deviations for ALogP Values from Compounds in the SPB Training Set

cutoff	SPB	ALogP
95%	yes	3.25 (1.87)
	no	1.35 (2.10)
90%	yes	3.21 (1.79)
	no	0.75 (1.89)

ADME Profiler uses FPSA and ALogP to predict PIA, and these values were used to build a Bayesian model using our implementation for a comparison of the two models' performance. Pipeline Pilot was also used to build a Bayesian model using the following descriptors: FCFPs with a neighborhood size of six (FCFP_6), ALogP, MW, nHDon, nHAcc, nRB, and PSA(NO). MOE was also used to calculate the following descriptors to generate a Bayesian model using the binary QSAR approach: nHDon, nHAcc, TPSA, SlogP, nRB, and MW.

Descriptor Selection for the SPB Data Set. ADME Profiler uses 1D similarity to a set of reference compounds and ALogP to predict SPB. The calculation of 1664 descriptors using DRAGON showed a high correlation between serum protein binding and ALogP. No other descriptors were found which were highly correlated to SPB and uncorrelated to ALogP. Table 3 displays the mean and standard deviation for the ALogP values calculated for the compounds in the SPB training set. Therefore, we proceeded to build the Bayesian models of SPB using fingerprints and logP. We calculated the MACCS structural keys and SLogP-VSA to build a binary QSAR model using MOE and FCFP_6 and ALogP to build a Bayesian model using Pipeline Pilot. Our implementation of the naïve Bayesian classifier was used to create two models. One model was based upon an in-house implementation of FCFP_6 and ALogP as calculated by DRAGON. The MACCS structural keys and SLogP-VSA calculated by MOE were used to generate the second model.

Receiver Operating Characteristic (ROC) Curves. The predictive ability of all of the models was evaluated by calculating the area under the ROC curves[42] using Pipeline Pilot. The ROC curve demonstrates the model's sensitivity, the ability to identify true positives, and specificity, the ability to avoid false negatives. The area under the ROC curve is a quantitative measure of the model's performance. A value of 1.0 represents the ability to perfectly discriminate between true positives and true negatives, while a value of 0.5 indicates that the model has no predictive ability.

RESULTS AND DISCUSSION

Prediction of BBB Penetration. Table 4 shows the calculated area under the ROC curves for all models used to predict BBB penetration. For all three Bayesian models constructed to predict BBB penetration, the set of six descriptors used by Garg and Verma[21] to train the classifiers resulted in the best performance. Modeling numerical data with a Gaussian distribution resulted in ~5% improvement compared to the models constructed using Pipeline Pilot and MOE with the same descriptors. Pipeline Pilot was used to calculate ALogP, PSA(NO), MW, nRB, nHAcc, and nHDon prior to training a naïve Bayesian classifier. The model resulted in ~5% poorer performance relative to the Pipeline Pilot model built with the set of six descriptors calculated with DRAGON and was ~10% worse than the Gaussian Bayesian model with the same set of descriptors. Molecular fingerprints (FCFP_6) were used to build a Bayesian model in Pipeline Pilot but resulted in substantially reduced performance with this particular data set.

Table 4. Prediction of BBB Penetration

model	area under ROC curve	descriptors
Gaussian naïve Bayes	0.9464	nHDon, nHAcc, TPSA(NO), MLogP, MW, nRB
	0.8438	SlogP-VSA, SMR-VSA, PEOE-VSA
	0.8578	FPSA, ALogP98
Pipeline Pilot	0.9068	nHDon, nHAcc, TPSA(NO), MLogP, MW, nRB
	0.7949	FCFP_6, ALogP, MW, nHDon, nHAcc, nRB, PSA(NO)
	0.8578	ALogP, MW, nHDon, nHAcc, nRB, PSA(NO)
binary QSAR	0.8904	nHDon, nHAcc, TPSA(NO), MLogP, MW, nRB
	0.8462	SlogP-VSA, SMR-VSA, PEOE-VSA
ADME Profiler	0.8788	FPSA, ALogP98

When training both our implementation and MOE's binary QSAR with the set of universal descriptors available in MOE (SLogP-VSA, SMR-VSA, and PEOE-VSA), the performance of both models was comparable but ~10% worse relative to the set of six descriptors used by Garg and Verma. The comparable performance of the Gaussian naïve Bayes and binary QSAR models suggests that in this case treating numerical values did not substantially improve the data. This is likely because the set of universal descriptors available in MOE are preoptimized for individual bins. SLogP-VSA is divided into 10 bins, SMR-VSA into eight bins, and PEOE-VSA into 14 bins. The authors state that these bins are created from the set of training compounds such that they are equally populated, "resulting in non-uniform boundaries". Hence, this is distinct from a situation in which a set of numerical values for other descriptors will be divided into bins with equal intervals, which will not necessarily be equally populated. Indeed, if the numerical values truly fit a Gaussian distribution, then the bins would necessarily not be equally populated.

The Gaussian Bayesian model outperformed ADME Profiler by ~5% when using the six descriptors reported by Garg and Verma calculated with DRAGON. ADME Profiler slightly (~2%) outperformed the Gaussian Bayesian model trained using the values for FPSA and ALogP calculated by Profiler. This difference in performance is because of Profiler's generous classification of compounds in the FPSA/ALogP plane relative to the Bayesian model. Using the set of six descriptors results in a model with higher dimensionality than the Gaussian Bayesian classifier uses to generate a more precise prediction of BBB penetration for a particular compound.

Figure 1 compares the results of the predictions from the ADME Profiler model and the Gaussian naïve Bayesian classifier trained using the set of six descriptors described by Garg and Verma for BBB penetration. The 49 compounds from the test set are plotted in FPSA/ALogP space and are labeled by their calculated Bayesian probabilities of BBB penetration. Both models correctly predict that the 19 compounds outside the 95% confidence ellipse are not likely

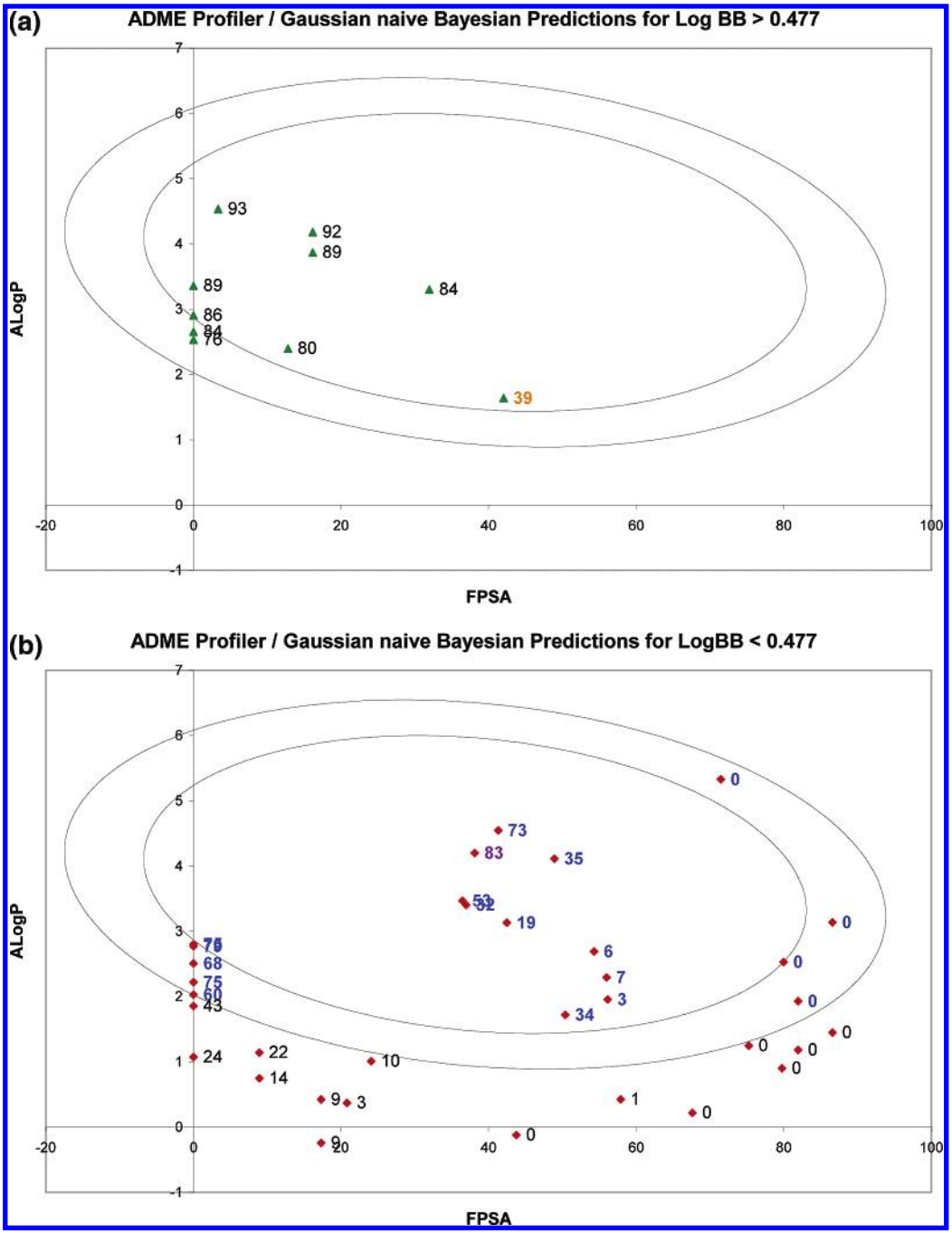

Figure 1. Plots of 49 compounds from the BBB penetration test set consisting of 11 true actives (a) and 38 true negatives (b) in FPSA/ALogP space. The two ellipses correspond to the 99% (inner) and 95% (outer) confidence regions for compounds predicted to be BBB-penetrant (observed log BB ≥ 0.477) in the ADME Profiler model. BBB-penetrant compounds are displayed as green triangles, and nonpenetrant compounds are shown as red diamonds. The numbers adjacent to each data point represent the calculated Bayesian probabilities as computed by a Gaussian naïve Bayesian model trained using nRB, nHAcc, nHDon, MLogP, TPSA(NO), and MW. Data points with blue numbers represent instances where ADME Profiler predicted that the compound was likely to cross the BBB while the Bayesian classifier correctly identified compounds as unlikely to cross the BBB. The data point with the violet number represents the case where both models incorrectly identified the compound as being likely to cross the BBB. The data point with the orange number represents the lone instance in the test set where the Bayesian classifier mistakenly classified the compound as unlikely to cross the BBB but the ADME Profiler model correctly identified it as likely to do so.

to cross the BBB (some data not shown). When a cutoff of 80% is used for the probabilities to discriminate whether compounds are likely to cross the BBB, the Bayesian classifier correctly identifies 10 of the 11 true positives in the test set (green triangles). ADME Profiler correctly predicts all 11 compounds but has much a higher rate of false positives. There are a total of 18 cases where the ADME Profiler model mistakenly identified compounds as likely to cross the BBB, but the Bayesian classifier correctly determined them to have a low probability of doing so. Figure 2 shows several examples of true negatives which were misclassified by the ADME Profiler model, along with a number of true positives correctly identified by each model that lie in the same region of FPSA/AlogP space. One example is seen where both models incorrectly identify a compound as likely to cross the BBB (tibolone). Also

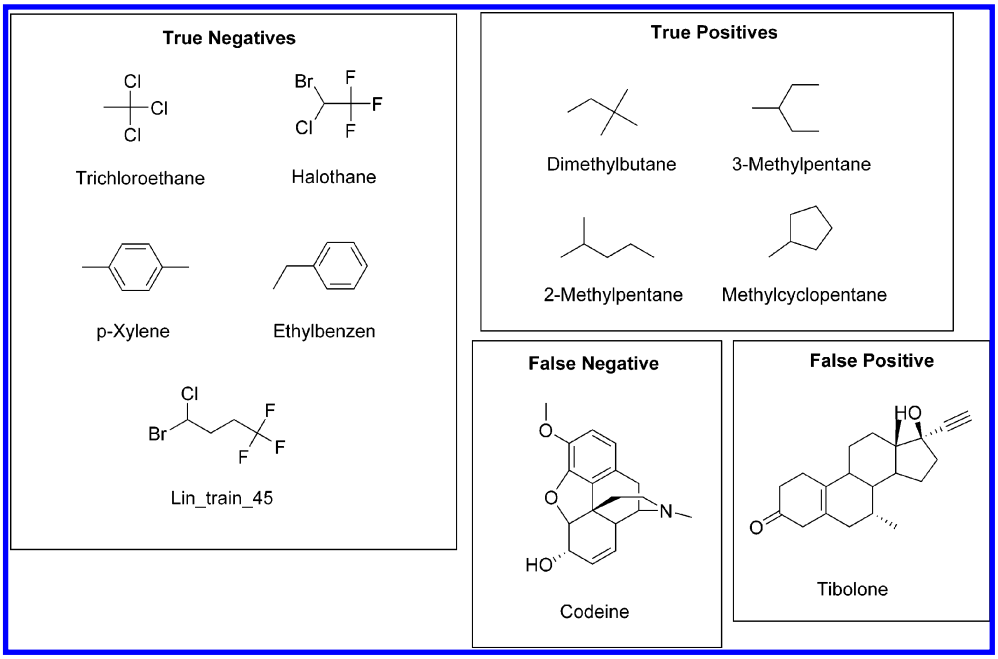

Figure 2. Five true negatives and four true positives correctly classified by naïve Bayes as BBB-penetrant but misclassified by the ADME Profiler model. Also shown are one false positive, which both models misclassified, and one false negative, which ADME Profiler correctly predicted but the Bayesian classifier misclassified as not crossing the BBB.

Table 5. Prediction of Human Intestinal Absorption

	area under ROC curve, cutoff for good passive intestinal absorption			
model	90%	80%	70%	descriptors
Gaussian naïve Bayes	0.7657	0.8030	0.9063	Hy, TPSA(NO), MLogP
	0.7040	0.8044	0.8906	FPSA, AlogP98
Pipeline Pilot	0.7189	0.8030	0.8854	Hy, TPSA(NO), MlogP
	0.7063	0.7956	0.8872	FCFP_6, AlogP, MW, nHDon, nHAcc, nRB, PSA(NO)
binary QSAR	0.6777	0.7481	0.8247	Hy, TPSA(NO), MlogP
	0.7051	0.7837	0.8299	nHDon, nHAcc, TPSA, SlogP, nRB, MW
ADME Profiler	0.6389	0.7437	0.8524	FPSA, AlogP98

observed is a single instance of a compound, codeine, which was misclassified by the Bayesian classifier as unlikely to cross the BBB but was properly identified by ADME Profiler.

Prediction of Human Intestinal Absorption. The calculated area under the ROC curves for all models used to predict PIA are shown in Table 5. For all three cutoffs used to define the "good" versus "bad" compounds in the PIA training set, the Gaussian naïve Bayesian classifier outperformed the other two implementations of the Bayesian classifier as well as the ADME Profiler model. The differences in performance between the Gaussian naïve Bayesian classifier and the other models were most significant at the 90% cutoff for the training data. As the criteria for well-absorbed compounds in the PIA training set was decreased, the performance of all models improved, with the differences between the Gaussian Bayesian models and the others becoming smaller.

The best performance of the Gaussian Bayesian model was obtained using three descriptors calculated by DRAGON [Hy, TPSA(NO), and MLogP]. Using these descriptors at the 90% cutoff for the training data, the Gaussian Bayesian model outperformed a naïve Bayesian model trained using Pipeline Pilot by ~5%. At the 80% cutoff, the performance between the Gaussian Bayes and Pipeline Pilot implementations was identical, and at the 70% cutoff, the Gaussian Bayesian model was slightly (~2%) superior. The Gaussian Bayesian model outperformed the binary QSAR model in MOE by ~9%, ~6%, and ~8% at the 90%, 80%, and 70% cutoffs, respectively.

Bayesian models were calculated with Pipeline Pilot and MOE at each cutoff for the set of training compounds using natively calculated descriptors. When Pipeline Pilot was used to create a naïve Bayesian model using FCFP_6, ALogP, MW, nHDon, nHAcc, nRB, and PSA(NO), the performance was essentially unchanged from that of the models trained using the set of three descriptors calculated using DRAGON at all cutoffs used for the training data. When MOE was used to calculate a binary QSAR model using nHDon, nHAcc, TPSA, SLogP, nRB, and MW, the performance was comparable to that of the Pipeline Pilot model trained with native descriptors at the 90% cutoff, ~2% worse at the 80% cutoff, and ~6% worse at the 70% cutoff for the training data.

When the Gaussian Bayesian classifier was trained with the ADME Profiler descriptors FPSA and ALogP, the Bayesian model outperformed the Profiler model by ~6% using the 90% and 80% cutoff values for the training set compounds and ~4% at the 70% cutoff. Figure 3 shows the comparison of the predicted PIAs in the ADME Profiler model and Gaussian naïve Bayesian model using data trained using a 90% cutoff for well-absorbed compounds. Com-

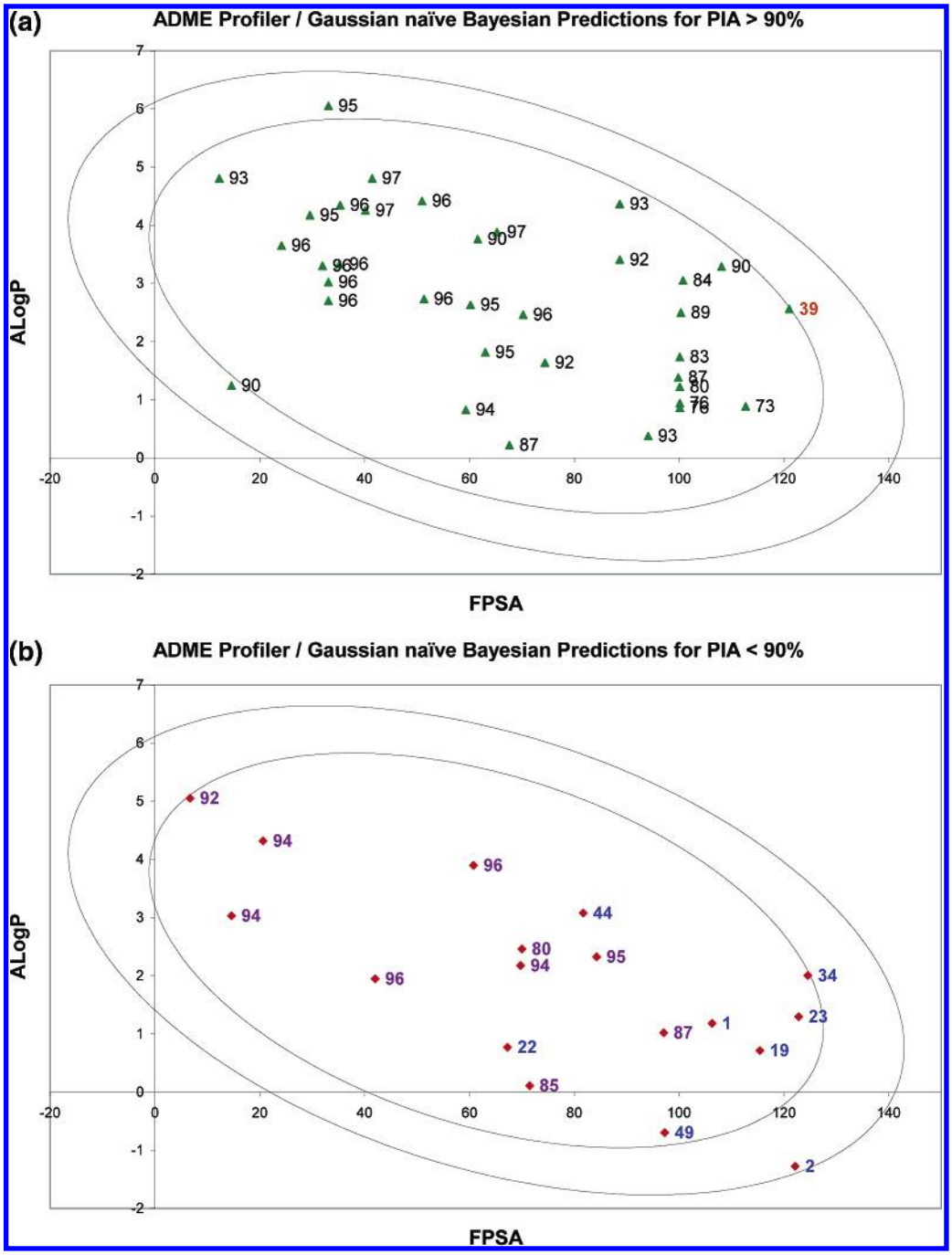

Figure 3. Plots of 59 compounds from the PIA test set consisting of 35 true actives (a) and 24 true negatives (b) in FPSA/ALogP space. The two ellipses correspond to the 99% (inner) and 95% (outer) confidence regions for well-absorbed compounds in the ADME Profiler model. Well-absorbed compounds (PIA ≥ 90%) are shown as green triangles, while poorly absorbed compounds (PIA < 90%) are depicted as red diamonds. The numbers adjacent to each point correspond to the probability that the compound will be well-absorbed, as calculated by a Gaussian naïve Bayesian classifier trained using the Hy, TPSA(NO), and MLogP descriptors. The eight data points with blue numbers correspond to the cases where the Bayesian model was able to correctly classify the compound as poorly absorbed and where the ADME Profiler model predicted the same compound as being well-absorbed. The data points with violet numbers correspond to the 10 cases where both models falsely identified the compound as being well-absorbed. The data point with the orange number corresponds to the single case in the test set where ADME Profiler correctly identified the compound as well-absorbed but where it was misclassified by the Bayesian classifier.

pounds are plotted in FPSA/ALogP space and are labeled by the Bayesian probability of having a PIA ≥ 90% as predicted by the model trained using Hy, TPSA(NO), and MLogP. The region of FPSA/ALogP space predicted by ADME Profiler to be populated by well-absorbed compounds is delineated by inner and outer ellipses showing the 99% and 95% confidence levels, respectively. Compounds are deemed to be well-absorbed in the ADME Profiler model if their PIA is ≥ 90%. Both models were able to correctly identify six compounds from the test set that are known to be poorly absorbed and are outside of the 95% confidence ellipse (data not shown). Within the area defined by the 99% confidence ellipse, there are 10 compounds which both ADME Profiler and the Bayesian models predict to be well-absorbed but which are not. Within the area bounded by the 95% confidence ellipse, there are eight compounds (Figure

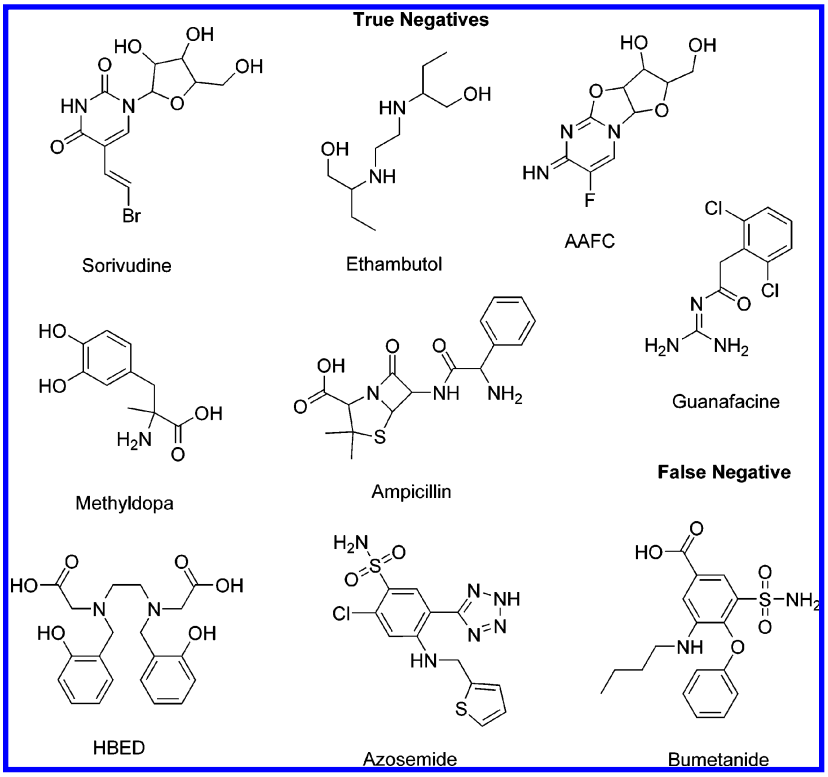

Figure 4. Eight compounds which the model in ADME Profiler incorrectly predicted as likely to have a high (≥90%) PIA but which the Gaussian naïve Bayesian classifier correctly identified as being poorly absorbed. Also shown is the one case from the test set where the Bayesian model incorrectly classified the compound as being poorly absorbed but where the model in ADME Profiler correctly identified it as being well-absorbed.

Table 6. Prediction of Serum Protein Binding

	area under ROC curve, cutoff for SPB		
model	95%	90%	descriptors
Gaussian naïve Bayes	0.7624	0.8268	FCFP_6, ALogP
	0.7487	0.6941	SLogP-VSA, MACCS structural keys
Pipeline Pilot	0.7607	0.8153	FCFP_6, ALogP
binary QSAR	0.5949	0.5426	SLogP-VSA, MACCS structural keys
ADME Profiler	0.7282	0.7648	ALogP, 1D similarity

4), which were falsely identified by ADME Profiler as being well-absorbed compounds but which the Bayesian model correctly classified as having a low probability of being well-absorbed. Also observed is one case where the Bayesian model misclassifies bumetanide, a well-absorbed compound which ADME Profiler is able to correctly identify (Figure 4).

Prediction of Serum Protein Binding. Table 6 shows the calculated area under the ROC curves for each SPB model. Unlike the BBB prediction models, where the inclusion of chemical fingerprints in the model was detrimental, and the PIA models, where their use had no effect on the predictive ability, they are essential to the construction of the SPB models because of the fact that a crucial component of this activity is an interaction with a specific protein—human serum albumin.

Two types of models were created using the Gaussian naïve Bayesian classifier. One of these types of models used FCFP_6 as calculated in Pipeline Pilot and ALogP as calculated in DRAGON, with the Laplacian weight increased to 100, to predict those compounds in the test set which bind at the 95% and 90% levels. The performances of these models were only slightly better than those of the Bayesian models created in Pipeline Pilot using FCFP_6 and ALogP, a result which is not likely to be statistically significant. A total of 7405 unique bits were used to completely describe all 260 compounds. Given that substantially more information is encoded in fingerprint space, which is quite large compared to the distribution of ALogP values, it seems reasonable to conclude that FCFP_6 would dominate the calculation of the Bayesian probabilities and that the two types of models should show similar performance. With both the Gaussian naïve Bayesian and Pipeline Pilot models, decreasing the cutoff from 95% to 90% increased the accuracy of the predictions by ∼5−6%.

A second type of Gaussian Bayesian model was constructed using SLogP-VSA and MACCS structural keys calculated in MOE and was compared to the models constructed using the binary QSAR method. In this case, the Gaussian Bayesian model showed substantially improved performance (∼15%) when compared to the binary QSAR model at both the 95% and 90% cutoff values used for compounds in the training set. This difference in performance is likely because the MACCS structural keys contain a total of 166 types, which is far smaller than the FCFP_6 calculated in Pipeline Pilot. This results in a significantly reduced ability to discriminate subtle differences between chemical structures. In this case, the calculated values for the SLogP-VSA descriptors become more important in calculating the probability of a compound to bind to serum protein at the 95% and 90% levels than the ALogP values in the Pipeline Pilot and Gaussian Bayesian models.

The dramatically improved performance of the Gaussian Bayesian classifier over the binary QSAR model is particu-

larly noteworthy. From our previous experience with the prediction of BBB penetration, it was clear that the performance of the two methods using the set of universal descriptors was comparable. Furthermore, the MACCS structural keys are expected to be binary fingerprints, and because both implementations model binary data in the same way, the results would have been expected to be comparable. The striking improvement in performance of the Gaussian Bayesian models stems from the fact that CCG's implementation of the MACCS structural keys does not generate a bit for the presence or absence of a given feature but rather a count for the number of times that feature is observed in a given compound. What is obtained is therefore not a bitstring but rather a set of 166 descriptors which are functional group counts. The Gaussian Bayesian model thus treats these descriptors not as binary data but as numerical values. The information content encoded in these 166 descriptors is therefore higher than would be expected from a simple bit string, resulting in the superior performance observed in the models.

The performance of the Gaussian naïve Bayesian classifier was compared to the results obtained by ADME Profiler. The Bayesian model trained with FCFP_6 and ALogP outperformed ADME Profiler by ~3% at the cutoff corresponding to 95% of a given compound bound to protein and ~6% at the 90% cutoff. ADME Profiler's performance was comparable to that of the Gaussian Bayesian model trained using the MACCS structural keys and SLogP-VSA at the 95% cutoff but was better than the model by ~7% at the 90% cutoff. It is important to note that the SPB data set used for these comparisons was the set of training compounds used for ADME Profiler's model, so good performance is expected. Application of the ADME Profiler SPB prediction model to a different set of compounds would be expected to result in less predictive ability.

There was a significant decrease in the performance of both the binary QSAR and Gaussian naïve Bayesian models trained using the MACCS structural keys and SLogP-VSA descriptors when using a cutoff of 90% of the compounds bound to serum protein as opposed to 95% bound. This decrease in performance was ~5% for both Bayesian models, suggesting that this may result from the specific compounds randomly selected for the training set relative to those of the test set. Both the Gaussian naïve Bayesian algorithm and the algorithm implemented in Pipeline Pilot performed well using models trained with FCFP_6 and ALogP at both the 95% and 90% cutoff values, suggesting that these descriptors are more appropriate choices for modeling the SPB data set. In particular, the larger number of fingerprint bits in FCFP_6 relative to the MACCS structural keys allows these models to extract more structural information from the set of training compounds. This implies that more structural information is present in the training set than is being utilized by the MACCS structural keys, leading to improved performance of the Bayesian classifiers overall.

The Improved Performance of the Bayesian Models Is Due to the Use of Higher Dimensional Data. The performance of ADME Profiler in the prediction of BBB penetration and PIA is quite impressive given the simplicity of the model. Indeed, using only two descriptors, the ADME Profiler model in general compares quite favorably to several of the Bayesian models (Tables 4 and 5). As can be seen from the comparison of ADME Profiler with the Gaussian naïve Bayesian models in Figures 1 and 2, the ADME Profiler model tends to be overly generous in classifying compounds within the confidence ellipses as likely to be well-absorbed or able to cross the BBB. This is to be expected because the ADME Profiler models were built using only positive data. The Bayesian classifier is better able to avoid these false positives because of the higher dimensionality of the models (six descriptors used for the BBB-penetration model and three descriptors used for the PIA model). This allows the Bayesian model to deconvolute situations such as that observed in Figure 1, where there are a number of compounds with a calculated FPSA of 0 and calculated ALogP values between 2.0 and 3.0. Several of these compounds are known to cross the BBB but are superimposed in the FPSA/ALogP plane with several which are not. The Bayesian model is able to discriminate between the two classes in this case by using the four additional descriptors (nHDon, nHAcc, NRB, and MW). Figure 2 shows five of these compounds which are correctly classified by the Bayesian model as true negatives, along with four compounds which are true positives.

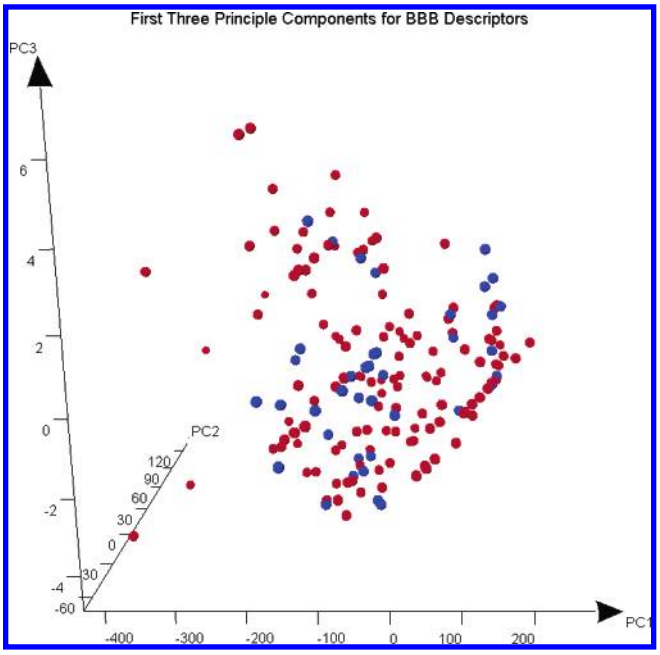

Figure 5. Plot of the first three principle components resulting from a principle components analysis of the six descriptors used to model BBB penetration. Compounds with an observed log BB $\geq$ 0.477 are shown in blue, while compounds with an observed log BB $<$ 0.477 are shown in red.

Figure 5 shows a plot of the first three principle components resulting from a principle components analysis of the six descriptors used to create the Bayesian BBB-penetration model. Together, the first three principle components account for 90% of the variability in the observed data (Table 7). The figure clearly shows that, even in higher dimensions, compounds known to cross the BBB occupy the same regions of descriptor space as those compounds which do not cross the BBB. The PCA results illustrate that the Bayesian algorithm is able to discriminate between classes by assigning probabilities to each piece of evidence individually, rather than by simply applying a nonlinear cutoff in the higher dimensions. This result is especially important given that there is a substantial overlap in the numerical values between

Table 7. Importance of Components in BBB-Penetration Model

	proportion of variance	cumulative proportion
PC 1	0.5645	0.5645
PC 2	0.2549	0.8195
PC 3	0.0836	0.9031
PC 4	0.0688	0.9719
PC 5	0.0145	0.9865
PC 6	0.0135	1.0000

the "good" and "bad" compounds used to train the Bayesian classifier (Tables 1−3).

Another example of the Bayesian classifier's ability to distinguish between classes of compounds with a desired activity can be seen in the plot of the PIA test set compounds shown in Figure 3. Ethambutol (Figure 4) lies in the center of the ellipse with calculated FPSA and ALogP values of 67.25 and 0.77 and is surrounded by a number of well-absorbed compounds. Yet the Bayesian classifier calculates only a 22% probability of having a PIA ≥ 90 (Figure 3). This compound has a reported PIA of 80%, so the model has correctly classified this compound as being unlikely to have a PIA ≥ 90%.

Improved Performance of Gaussian Modeling versus Binning of Numerical Data. Binning the values of numerical descriptors has the disadvantage that compounds with values close to the edges of the bins are treated equally with compounds whose values lie in the center of the bins. The binary QSAR model in MOE attempts to overcome this problem by applying a Gaussian smoothing function to the binned values, but problems in modeling the data still exist for bins which are nearly empty. The solution which was chosen in this case was to then add the Laplacian modifier to the probabilities obtained for each bin after the Gaussian smoothing. Binning procedures such as those implemented in Pipeline Pilot and MOE then add a small constant, the Laplacian correction, to the probabilities of each individual descriptor using the prior probability that the hypothesis is true. By contrast, Gaussian modeling of the numerical values essentially calculates a weight which is specific for each numerical descriptor, independently of the other descriptors in the training set. Furthermore, the Gaussian weight is neither an arbitrarily assigned value nor is it assigned by the user. This takes into account differences between distributions of different descriptors while avoiding user bias. Binning carries the implicit assumption that each bin for a given numerical descriptor is independent of the values in the other bins for that same descriptor. This assumption is not entirely accurate, and by modeling numerical data as a Gaussian distribution, the relationship between numerical values for a given descriptor is taken into consideration, leading to a more robust model.

Although we assume here that numerical data fit a Gaussian distribution, this is not necessarily the case in all instances. A Gaussian distribution, however, is a reasonable assumption if the number of compounds in the training set is not too small. The descriptors calculated for the data sets in this paper (molecular weight, logP, etc.) are likely to fit such a distribution if the set of training compounds is truly representative of the larger set of test and training compounds being modeled. Cases where the distribution of descriptor values are truly non-Gaussian could be determined by calculating the relevant descriptors and plotting their resulting values as a histogram. Non-Gaussian distributions should become apparent with such a priori assessment of the data, and different distributions could be considered in place of the Gaussian.

Validation of Statistical ADME Models. A significant potential source of error in statistical ADME models is due to the noisy nature of biological data. In particular, there may be considerable variation not only from different observations in the same lab but between different laboratories as well. The selection of a single cutoff for biological activity will likely introduce false positives and false negatives, as the reported value for a given biological activity may be a statistical outlier, or the cutoff may lie close to the mean value for several observations of a single compound. Unlike logistic regression models, naïve Bayesian classifiers tend not to overfit to the training data and are therefore more tolerant of experimental noise. If the standard deviation of the observed values for a given biological activity is large, two cutoffs could be used in principle instead of one. For example, when developing a passive absorption model, compounds in the training set classified as likely to be absorbed could be those with a measured HIA ≥ 90%, while those which are classified as not likely to be absorbed could be those with a measured HIA < 70%.

The construction of a statistical model for ADME property prediction using experimentally measured values is problematic because of the variability inherent in any such data set. Compounds are usually culled from the literature after being reported by a variety of research groups using different experimental protocols. Terms such as bioavailability and absorption may be used interchangeably and incorrectly. No distinction may be made between actively transported or passively transported molecules. For some compounds, the mode of transport may not even be known. Reported values may come from a single measurement or multiple measurements. Significant information may be absent from the original report such as the standard deviation of the reported values. Methods used to measure ADME properties may include human, rat, mouse, dog, and monkey models, or some combination of the above. Preferably, the validation of statistical models for ADME prediction would use a set of training and test compounds obtained from one laboratory using a single experimental protocol. Ideally, there would also be multiple measurements obtained for each compound, so the true values of the mean and standard deviation could be determined and used to assist in model construction in the manner described above. This would allow the construction of better statistical models, as well as the ability to do a better comparative evaluation of the performance of different models.

CONCLUSION

We have implemented a Gaussian naïve Bayesian classifier capable of modeling both binary and numerical data. The use of such classifiers is commonplace outside the field of cheminformatics. We have shown that modeling numerical data as a Gaussian distribution yields generally superior results when compared to those of implementations which model numerical data by creating a series of bins to convert values into binary data. Implementation of the Gaussian

model is straightforward and could be implemented in-house or integrated into existing software packages. Further improvements could include the implementation of other methods such as kernel density estimation or other probability distributions to extend the modeling of numerical values to the cases where the values do not fit a Gaussian distribution. One example might be an attempt to use the scores resulting from high-throughput docking as descriptors where the values demonstrate a clear bimodal distribution (unpublished results). An additional advantage of using naïve Bayesian classifiers is their inherent extensibility. Compounds can be easily added to the training set from the literature or from in-house studies, resulting in more robust predictions as more training data become available.

REFERENCES AND NOTES

(1) Bayes, T. Essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. London* **1763**, *53*, 370−418.

(2) Muscatello, D. J.; Churches, T.; Kaldor, J.; Zheng, W.; Chiu, C.; Correll, P.; Jorm, L. An automated, broad-based, near real-time public health surveillance system using presentations to hospital Emergency Departments in New South Wales, Australia. *BMC Public Health* **2005**, *5*, 141.

(3) Ivanov, O.; Wagner, M. M.; Chapman, W. W.; Olszewski, R. T. Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance. *Proc. AMIA Symp.* **2002**, 345−349.

(4) Zorkadis, V.; Karras, D. A.; Panayotou, M. Efficient information theoretic strategies for classifier combination, feature extraction and performance evaluation in improving false positives and false negatives for spam e-mail filtering. *Neural Networks* **2005**, *18* (5−6), 799−807.

(5) Miaou, S. P.; Song, J. J. Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accid. Anal. Prev.* **2005**, *37* (4), 699−720.

(6) Liu, H.; Wong, L. Data mining tools for biological sequences. *J. Bioinf. Comput. Biol.* **2003**, *1* (1), 139−167.

(7) Chinnasamy, A.; Sung, W. K.; Mittal, A. Protein structure and fold prediction using Tree-Augmented naive Bayesian classifier. *J. Bioinf. Comput. Biol.* **2005**, *3* (4), 803−819.

(8) Labute, P. Binary QSAR: A new method for the determination of quantitative structure activity relationships. *Pac. Symp. Biocomput. '99* **1999**, 444−455.

(9) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* **2004**, *47* (18), 4463−4470.

(10) Sun, H. A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (2), 748−757.

(11) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enriching extremely noisy high-throughput screening data using a naive Bayes classifier. *J. Biomol. Screening* **2004**, *9* (1), 32−36.

(12) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening* **2005**, *10* (7), 682−686.

(13) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 2743−2749.

(14) Yoon, S.; Smellie, A.; Hartsough, D.; Filikov, A. Surrogate docking: Structure-based virtual screening at high throughput speed. *J. Comput.-Aided Mol. Des.* **2005**, *19* (7), 483−497.

(15) John, G. H.; Langley, P. In *Estimating Continuous Distributions in Bayesian Classifiers*, Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, Canada, 1995; Besnard, P., Hanks, S., Eds.; Morgan Kaufmann Publishers: Montreal, Canada, 1995; pp 338−345.

(16) Witten, I. H.; Frank, E. Statistical modeling. In *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*; Cerra, D. D., Ed.; Morgan Kaufmann Publishers: New York, 2000; pp 82−88.

(17) Ng, A. Y.; Jordan, M. I. *On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes*; MIT Press: Cambridge, MA, 2002; Vol. 14.

(18) *Pipeline Pilot*, 5.1; SciTegic, Inc.: San Diego, CA, 2005.

(19) *Molecular Operating Environment*, 2005.06; Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2005.

(20) Delisle, R. K.; Lowrie, J. F.; Hobbs, D. W.; Diller, D. J. Computational ADME/Tox modeling: Aiding understanding and enhancing decision making in drug design. *Curr. Comput.-Aided Drug Des.* **2005**, *1* (4), 325−345.

(21) Garg, P.; Verma, J. In silico prediction of blood brain barrier permeability: An artificial neural network model. *J. Chem. Inf. Model.* **2006**, *46* (1), 289−297.

(22) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.

(23) Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure−activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci.* **2002**, *90* (6), 749−784.

(24) Wohnsland, F.; Faller, B. High-throughput permeability pH profile and high-throughput alkane/water log P with artificial membranes. *J. Med. Chem.* **2001**, *44*, 923−930.

(25) Chiou, W. L.; Barve, A. Linear correlation of the fraction of oral dose absorbed of 64 drugs between humans and rats. *Pharm. Res.* **1998**, *15* (11), 1792−1795.

(26) Chiou, W. L.; Jeong, H. Y.; Chung, S. M.; Wu, T. C. Evaluation of using dog as an animal model to study the fraction of oral dose absorbed of 43 drugs in humans. *Pharm. Res.* **2000**, *17* (2), 135−140.

(27) Chiou, W. L.; Buehler, P. W. Comparison of oral absorption and bioavailability of drugs between monkey and human. *Pharm. Res.* **2002**, *19* (6), 868−874.

(28) Matsson, P.; Bergstrom, C. A. S.; Nagahara, N.; Tavelin, S.; Norinder, U.; Artursson, P. Exploring the role of different drug transport routes in permeability screening. *J. Med. Chem.* **2005**, *48*, 604−613.

(29) Varma, M. V. S.; Sateesh, K.; Panchagnula, R. Functional role of P-glycoprotein in limiting intestinal absorption of drugs: Contribution of passive permeability to P-glycoprotein mediated efflux transport. *Mol. Pharm.* **2005**, *2* (1), 12−21.

(30) Cheng, A.; Diller, D. J.; Dixon, S. L.; Egan, W. J.; Lauri, G.; Mertz, K. M., Jr. Computation of the physio-chemical properties and data mining of large molecular collections. *J. Comput. Chem.* **2002**, *23*, 172−183.

(31) Egan, W. J.; Dixon, S. L. Activity prediction models. U.S. Patent Publication number 2003/0073128 A1, April 17, 2003.

(32) Dixon, S. L.; Merz, K. M., Jr. One-dimensional molecular representations and similarity calculations: Methodology and validation. *J. Med. Chem.* **2001**, *44*, 3795−3809.

(33) *The PDR Electronic Library*; Medical Economics Co. Inc.: Montvale New Jersey, 1999; Vol. 1999.2.

(34) Todeschini, R.; Consonni, V. A.; Mauri, M. P. *DRAGON*, 5; Talete srl: Milano, Italy, 2005.

(35) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of a fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.

(36) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure−activity relationships I. Partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1985**, *7* (4), 565−577.

(37) Viswanadhan, N. N.; Reddy, M. R.; Bacquet, R. J.; Erion, M. D. Assessment of methods used for predicting lipophilicity: Application to nucleosides and nucleoside bases. *J. Comput. Chem.* **1993**, *14* (9), 1019−1026.

(38) Egan, W. J.; Mertz, K. M., Jr.; Baldwin, J. J. Prediction of drug absorption using multivariate statistics. *J. Med. Chem.* **2000**, *43* (21), 3867−3877.

(39) Todeschini, R.; Vighi, M.; Finizio, A.; Gramatica, P. 3D-modelling and prediction by WHIM descriptors. Part 8. Toxicity and physicochemical properties of environmental priority chemicals by 2D-TI and 3D-WHIM descriptors. *SAR QSAR Environ. Res.* **1997**, *7* (1−4), 173−193.

(40) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. Simple method of calculating octanol/water partition coefficient. *Chem. Pharm. Bull.* **1992**, *40* (1), 127−130.

(41) Moriguchi, I.; Hirono, S.; Nakagome, I.; Hirano, H. Comparison of reliability of log P values for drugs calculated by several methods. *Chem. Pharm. Bull.* **1994**, *42* (4), 976−978.

(42) Witten, I. H.; Frank, E. ROC curves. In *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*; Cerra, D. D., Ed.; Morgan Kaufmann Publishers: New York, 2000; pp 141−147.

CI0601315