Benchmarking of QSAR Models for Blood-Brain Barrier Permeation

Dmitry A. Konovalov,*,† Danny Coomans,† Eric Deconinck,‡ and Yvan Vander Heyden‡

School of Mathematics, Physics and Information Technology, James Cook University, Townsville, Australia, and Deptartment of Analytical Chemistry and Pharmaceutical Technology, Pharmaceutical Institute, Vrije Universiteit Brussel, Brussels, Belgium

Received March 20, 2007

Using the largest available database of 328 blood−brain distribution (logBB) values, a quantitative benchmark was proposed to allow for a consistent comparison of the predictive accuracy of current and future logBB/quantitative structure−activity relationship (-QSAR) models. The usefulness of the benchmark was illustrated by comparing the global and *k*-nearest neighbors (*k*NN) multiple-linear regression (MLR) models based on the linear free-energy relationship (LFER) descriptors, and one non-LFER-based MLR model. The leave-one-out (LOO) and leave-group-out Monte Carlo (MC) cross-validation results ($q^2 = 0.766$, qms = 0.290, and $qms_{mc} = 0.311$) indicated that the LFER-based *k*NN-MLR model was currently one of the most accurate predictive logBB-QSAR models. The LOO, MC, and *k*NN-MLR methods have been implemented in the QSAR-BENCH program, which is freely available from www.dmitrykonovalov.org for academic use.

INTRODUCTION

The blood−brain (BB) distribution of a molecule (reported as logBB) is a key characteristic for assessing the suitability of a molecule as a drug for the central nervous system (CNS).[1] When the new paradigm of drug discovery is worked within,[2] high-throughput chemoinformatics/bioinformatics or combinatorial chemistry methods may generate a vast number of candidate CNS compounds with desired therapeutic properties. However, the candidate compounds must also be screened for suitable absorption, distribution, metabolization, excretion, and toxicity properties including acceptable logBB values for the CNS compounds. Quantitative structure−activity/property relationship (QSAR/QSPR) models offer such in silico screening by predicting logBB from the molecular structure of the compounds, hence supporting the "fail fast, fail cheap" business model[3] of drug development.

A number of QSAR−logBB models have been developed over the years,[4−15] and new models are being continuously proposed.[16] Two classes of the logBB models stand out. The first is the multiple linear regression (MLR) model based on the linear free-energy relationship (LFER) descriptors.[1] A recent LFER-based MLR (LFER-MLR) model was trained on what is currently the largest logBB data set (328 data points) and achieved $r^2 = 0.75$ and $s = 0.3$ log units.[1] It was suggested that the LFER-MLR model had reached the predictive limit obtainable from the data set since the experimental errors of the logBB measurements were estimated to be around 0.3.[1] It was also suggested[1] that more complex methods in either descriptors or computation did not result in any significantly better prediction of logBB. Therefore, the development of logBB models with better predictive power may now be limited by the lack of more accurate logBB measurements rather than by the development of better QSAR methods. The only potential problem of the LFER-MLR model was identified in the difficulty to calculate the LFER descriptors for structurally diverse drug candidates.[6] This critique may no longer be applicable since the descriptors could be calculated by the ADME Boxes program, which is available from Pharma Algorithms.[17]

The second class of the logBB models consists of MLR models where other (non-LFER) descriptors were selected[4,6,8,9,12,16] from more than 3000 descriptors. Unfortunately, the published models are difficult to compare since they used a variety of training and test subsets as well as different validation methodologies.[1] As it stands, it is not clear if non-LFER-based MLR models or other more complex logBB models,[10,11] for example, artificial neural networks (ANN), have any practical advantage over the LFER-MLR model.[1] This situation motivated the main objective of this study: to propose a quantitative benchmark for consistently comparing the *predictive* accuracy of the logBB QSAR models. The usefulness of the benchmark was illustrated by performing the benchmark calculations for the LFER-MLR[1] model and the non-LFER-based model of Narayanan and Gunturi[4] (NG-MLR). One immediate implication of such a benchmark was to reconsider a number of QSAR models which were developed by the "data mining" of more than 3000 currently available descriptors. Since many descriptors are highly correlated, there are potentially a vast number of descriptor combinations which would produce a "reasonable" fit of training as well as (typically small) test sets.

The second objective was to examine whether the LFER-MLR model had reached the limit of predictive accuracy on the proposed benchmark. This was done by using *k*-nearest neighbors (*k*NN) regression,[11] which was similar (if not identical) to the local lazy regression (LLR).[18]

MATERIALS AND METHODS

Benchmark Data Set. The largest available database[1] of logBB values contained 328 data points and was denoted as AI328 (first letters of the first two authors were used for

* Corresponding author. E-mail: dmitry.konovalov@jcu.edu.au.
† James Cook University.
‡ Vrije Universiteit Brussel.

10.1021/ci700100f CCC: $37.00
Published on Web 06/30/2007

labeling throughout this study). Abraham et al.[1] demonstrated that the common AI328 data set could be assembled from in vivo and in vitro logBB values for blood, plasma, and serum to the rat brain. Out of the 328 records, 302 were identified as unique compounds.[1] Further examination revealed that the SMILES were identical for the following pairs of data points: 225/126, 255/246, 323/260, 310/261, 318/269, and 319/270 (numbering is from Table S1 of Abraham et al.[1]). Even though multiple entries per compound are acceptable for MLR, other models may require a unique data point per compound; for example, *k*-nearest neighbors[11,18] models would produce skewed results for multiple entries. Hence, for benchmarking purposes, multiple logBB and LFER values corresponding to the same compound were averaged, if they were different, arriving at the final data set of 291 compounds (denoted as KC291-LFER, see Supporting Information Table S1): the same LFER values but different a logBB in 225/126 and 255/246 and the same logBB but a different LFER in 323/260, 310/261, 318/269, and 319/270. Another requirement on the compounds was the availability of corresponding structural molecular descriptors; for example, the Parameter Client (PCLIENT)[19] Web site could not calculate the DRAGON[20] descriptors for the first five compounds from the original data set, that is, [Ar], [Kr], [Ne], [Rn], and [Xe], and hence they were omitted from the benchmark data set.

Benchmark of Predictive Accuracy. Let a sample of n compounds be described by a $n \times (p + 1)$ matrix $\mathbf{Z} = (Y,D_1,D_2,...,D_p)$, where p is the number of descriptors or, generally speaking, any predictor variables; $Y = (y_1,y_2,...,y_n)^T$ is the column of observed logBB values (the superscript "T" denotes the transpose) or responses; and $D_j = (x_{1j},x_{2j},...,x_{nj})^T$ is the column vector of the jth molecular descriptor or LFER indicator. Other useful notations are $\mathbf{Z}^T = (Z_1,Z_2,...,Z_n)$, $\mathbf{X}^T = (X_1,X_2,...,X_n)$, and $X_i^T = (x_{i1},x_{i2},...,x_{ip})$, where $Z_i^T = (y_i,x_{i1},x_{i2},...,x_{ip}) = (y_i,X_i^T)$ is the complete data record for the ith compound; $Y' = (y'_1,y'_2,...,y'_n)^T$, with y'_i being the logBB value of the ith compound predicted by a QSAR model. The standard set of MLR statistics was considered: the sum of squares due to error, $SSE = \sum_{i=1}^{n} (y_i - y'_i)^2$; the sum of squares due to regression, $SSR = \sum_{i=1}^{n} (y'_i - \bar{y})^2$; the total sum of squares, $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$; the corresponding mean values $MSE = SSE/(n - p - 1)$, $s = \sqrt{MSE}$, $MSR = SSR/p$, and $MST = SST/(n - 1)$; the proportion of total variation in Y explained by regression, $r^2 = 1 - SSE/SST$; r^2 is adjusted for available degrees of freedom, $r_a^2 = 1 - MSE/MST$; $F = MSR/MSE$, which rejects the null hypothesis, H_0:$\beta_1 = \beta_2 = ... = \beta_p = 0$, at level α if $F > F_{p,n-p}(1 - \alpha)$, where $(\beta_1,\beta_2,...,\beta_p)$ are the MLR coefficients and $F_{p,n-p}$ is an F distribution with p and $n - p$ degrees of freedom. If the given data set satisfies $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i$, where error ϵ_i is independently drawn from a normal distribution $N(0,\sigma^2)$, then the expected value of MSE becomes $E(MSE) = \sigma^2$. Even though the MLR statistics are well-known and are useful for comparing MLR models, they (strictly speaking) do not measure the predictive power of the MLR models, which is required for the benchmarking between MLR- as well as non-MLR-based logBB models.

The predictive power of a QSAR model is typically studied by cross-validation, when a given data set is split into two subsets: a calibration (i.e., training) subset S_c (of size n_c) and a validation (i.e., test) subset S_v (of size $n_v = n - n_c$). The predictive accuracy of the model could be assessed by the averaged mean squared error of prediction, $MSEP(n_v) = (1/Nn_v) \sum_{i=1}^{N} ||Y_{S_v(i)} - Y'_{S_v(i)}||^2$ (also reported as its square root, $RMSEP = \sqrt{MSEP}$), where: $||\cdot||^2$stands for the Euclidean norm of a vector; $Y_{S_v(i)}$ is the $S_v(i)$ subvector of Y; $Y'_{S_v(i)}$ is the vector of the $S_v(i)$ predicted Y values, where the model training is done using *only* the $S_c(i)$ calibration subset; and N is the number of different cross validations or different splits. For the leave-one-out (LOO) cross-validation (LOOCV), $n_v = 1$ and $N = n$,[4,8,12,16] the predictive accuracy of a logBB-QSAR model is also assessed by $r^2 = cov^2(Y,Y')/[var(Y)\ var(Y')] = 1 - SSE/SST$ (known as q^2, r_{cv}^2 or r_{press}^2) and RMSEP (known as qms),[16] where SSE, in this case, is also known as the predictive residual sum of squares. Since the logBB-MLR models have a varied number of descriptors, r_a^2 (denoted as q_a^2) should also be considered/reported to prevent the selection of a MLR model which achieves higher q^2 by simply increasing the number of descriptors. Note that we used a common statistics textbook notation MSE for SSE/$(n - p - 1)$, while MSE also quite commonly denotes SSE/n. To distinguish the two variants, we denoted SSE/n by the mean squared error of prediction (MSEP), reflecting its equal applicability to non-MLR models.

The leave-group-out[8,12] (LGO) cross-validation is another method for assessing the predictive accuracy of the QSAR models. A single iteration of the LGO method was commonly used when the available data set was split into a training subset and a test subset (not used to obtain the training equation or algorithm). When RMSEP was calculated for such a single LGO split, it was denoted as rms_{test}, throughout this study. For the logBB problem, the training and test subsets were normally on the order of 100 compounds or smaller. Even though such single iteration of the cross-validation was commonly reported, the statistical validity of such a practice is questionable; for example, it was suggested that the practice was only statistically meaningful for substantially larger data sets.[21] The statistically consistent (even for such relatively small data sets) LGO cross-validation is known as the Monte Carlo (MC) cross-validation (MCCV) method.[22−24] The ith iteration of the MCCV method randomly splits the sample data set into two subsets: $S_c(i)$ and $S_v(i)$.[23] By repeating the procedure N times, the averaged MSEP is calculated with the corresponding RMSEP, $qms_{mc}(n_v) = \sqrt{(1/Nn_v)\sum_{i=1}^{N}||Y_{S_{v(i)}} - Y'_{S_v(i)}||^2}$. For $n_v = 1$ and $N \gg n$, the MCCV method converges to the LOO method, and hence the standard LOO method is computationally more efficient (for $n_v = 1$); that is, the LOO procedure needs to be run only n times. However, when $n_v \gg 1$ values are considered, the exhaustive enumeration of all different $S_v(i)$ sets becomes numerically impossible; that is, $n!/(n_v!n_c!)$ iterations are required, while the MCCV method requires at most only n^2 iterations.[23] Note that the MCCV method was developed to assist in MLR model selection, which is exactly what is required for the benchmarking purposes since the majority of the logBB models are MLRs. The MCCV method is a variation of the LGO method which was shown to be more accurate than the LOO method in describing the predictive ability of the MLR models.[24] In particular, the probability of the LOO cross-validation method to select the MLR model with the best predictive

ability does not converge to one as the total number of observations $n \rightarrow \infty$.[24] However, the probability of the LGO method to select the best predictive MLR model converges to one only when $n_v/n \rightarrow 1$ and $n \rightarrow \infty$.[24] Shao[24] suggested using $n_c = n^{3/4}$, which in the case of the KC291 data set became $n_c = 70$ and $n_v = 221$, while the majority of the published logBB models selected $n_c \geq n_v$. We therefore adopted the choice of Abraham et al.,[1] $n_c \approx n_v \approx n/2$, as potentially more acceptable for the current logBB community. Note that Xu et al.[23] also developed a corrected MCCV which was specific to MLR models and therefore could not be used for generic benchmarking.

The q^2 statistic was commonly reported to indicate the predictive power of the logBB models, that is, according to the standard interpretation of r^2 as the proportion of response variability explained by the regression. There is however a potential problem in using the statistic for the descriptor-mining models when the LOO cross-validation is employed (by maximizing q^2) to select normally a small subset of descriptors from sometimes thousands of available descriptors since *all* information in a particular data set is utilized for the selection. For such cases, q^2 could be viewed as a measure of a model's internal self-consistency rather than its predictive power. For example, Narayanan and Gunturi[4] reported $q^2 = 0.68$ for their V1 selection of descriptors and calibration set of 88 compounds.[12] As shown, later in this study, q^2 obtained that way could be overestimated. Moreover, if the LOO or MCCV methods were strictly adhered to, the descriptors selected for a particular calibration subset set would (generally) not be the same as the descriptors selected for any other calibration subset or even the total set. Even if some descriptor set is selected (say, five descriptors), the descriptors are drawn from a large pool of descriptors, and the resulting equation cannot be retrained because the descriptors drawn from the large pool to fit the new data set may not be the same five descriptors as originally used.

The proposed benchmark focuses on the *predictive* power of the models. Therefore, the commonly reported r^2 statistic must be excluded from the benchmark/comparison. This is because it has been shown theoretically[25−27] and is already quite commonly acknowledged[1] that a higher r^2 does not guarantee higher predictive accuracy of non-MLR models.

***k*-Nearest Neighbors.** The presence or absence of nonlinearity and/or clustering could be checked by the LLR[18] where MLR is performed on *k*NNs[11] of a data point. The *k*NN model is essentially just a local MLR model, but the *k*NN idea is more generic and could be used with other more complex (non-MLR) models, especially for clustered data sets. The standard Euclidian distance (or metric) was used to measure the pairwise distances, $d_{ii'}^2 = ||X_i - X_{i'}||^2 = \sum_{j=1}^{k} (x_{ij} - x_{i'j})^2$, in this study. Since the distance measure is required for the *k*NN model to work and given the possibly diverse ranges of descriptors, the descriptor values could be standardized to a mean of zero and a variance of one: $x'_{ij} = (x_{ij} - \mu_j)/s_j$, where $\mu_j = {}^1/_n \sum_{i'=1}^{n} x_{i'j}$ and $s_j^2 = 1/(n-1) \sum_{i'=1}^{n} (x_{i'j} - \mu_j)^2$. When the standardized descriptor values, x'_{ij}, were used, the pairwise distances were calculated from x'_{ij} rather than from x_{ij}. The corresponding r^2 statistic is denoted as q_k^2 to track the k number of neighbors. Since the q^2 statistic is not available for the *k*NN model, the q_k^2 statistic is used instead for the benchmarking. This substitution by q_k^2 is clearly justified because it is a stricter test having fewer data points to work with, compared to q^2 for $k < n - 1$.

Formally, in vector-matrix notation, the *k*NN model predicts the logBB value of the ith compound via $y'_i = W_i^T B_i(k)$, where $W_i^T = (1, x_{i1}, x_{i2}, ..., x_{ip})$ is the $1 \times (p+1)$ row vector containing p molecular descriptors for the ith compound and $B(i,k) = (b_0, b_1, ..., b_p)^T$ is the $(p+1) \times 1$ column vector of regression coefficients which depends on the k-nearest neighbors of the ith compound. The $B(i,k)$ vector is calculated via $B(i,k) = (W_{ik}^T W_{ik})^{-1} W_{ik}^T Y_{ik}$ (or via more numerically stable QR decomposition), where $W_{ik}^T = (W_{i_1}, W_{i_2}, ..., W_{i_k})$ is the matrix consisting of the k-nearest neighbors of the ith compound; $\{i_1, i_2, ..., i_k\} = NN_{ik}$ is the k-nearest neighbor subset of sample indices $\{1,2,...,n\}$, with $i \notin \mathrm{NN}_{ik}$; $Y_{ik} = (y_{i_1}, y_{i_2}, ..., y_{i_k})^T$.

RESULTS AND DISCUSSION

The LFER Descriptors. Our KC291-MLR results reproduced very closely the original results of Abraham et al.[1] The MLR model (first row of Figure 1 and Table 1) of the KC291 data set became

$$\mathrm{logBB} = 0.513 + 0.189E - 0.604S - 0.635A - 0.621B + 0.638V - 1.171\mathrm{Ic} - 0.418\mathrm{Iv} \quad (1)$$

with $r_a^2 = 0.746$, $r^2 = 0.753$, $s = 0.301$, and $F = 122$ and where E, S, A, B, and V were the LFER[1] parameters; Ic[1,13] and Iv[1] were indicator variables. The quantile−quantile plot (third column in Figure 1) of the residuals confirmed that the residuals were in fact distributed very closely to normal. The predictive power of the model was assessed according to the proposed benchmark obtaining $q^2 = 0.732$, qms = 0.314, and $\mathrm{qms_{mc}}(n_v = 145) = 0.316$ (Model 2 in Table 1). Hence, the KC291-MLR model could predict the in vitro and in vivo logBB values from the LFER descriptors within an accuracy of about ± 0.3[1] and ± 0.6 log units with 68% and 95% confidence (due to the near normality of the residuals), respectively. We found that $N = 3 \times 10^4$ in the MCCV method was sufficient to obtain $\mathrm{qms_{mc}}$ within 0.001 accuracy.

It was generally assumed that[1] (1) the predictive accuracy of a QSAR model could only be guaranteed within the chemical space (e.g., visualized by the first two principal components)[1] of the original training data set and (2) a larger homogeneous training set would yield a global MLR model with better predictive power. Figure 2 clearly validated the first assumption on the KC291 data set, where $q^2(k\mathrm{NN}) > q^2$ for most considered k values and for $k > 50$ with nonstandardized and standardized descriptor values, respectively (k was varied from 30 to 230). For example, the ($k = 70$)NN model achieved $q^2(70\mathrm{NN}) = 0.753$ (Model 3 in Table 1), while $k = 70$ furthest neighbors (70FN) exhibited a much lower predictive power, $q^2(70\mathrm{FN}) = 0.143$ (Model 4 in Table 1). The second assumption was validated by comparing MLR and 70NN results (Table 1) for KC291, KC90 (in vitro subset of KC291, Iv = 1), and KC201 (in vivo subset of KC291, Iv = 0). The global MLR model had in fact better predictive accuracy on the whole KC291 set ($q^2 = 0.732$ and $F = 114$, Model 2) than on each of the KC90 ($q^2 = 0.672$ and $F = 36$, Model 8) and KC201 ($q^2 = 0.719$ and $F = 86$, Model 10) subsets. The same held for the 70NN model, where the

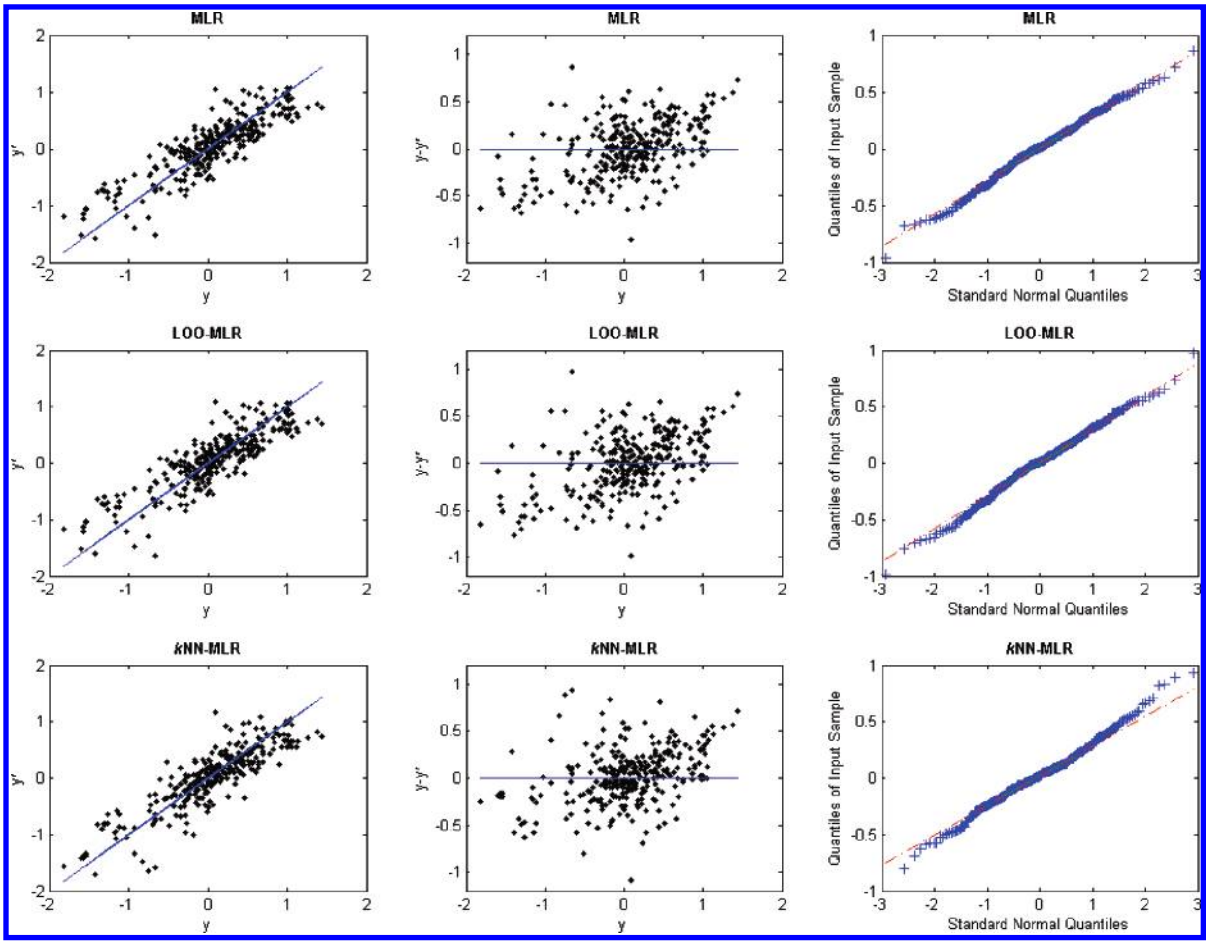

Figure 1. MLR models of the KC291 data set: first column, experimental LogBB (y) vs predicted (y'); second column, residual plot; third column, quantile−quantile plot of the residuals. The kNN model used $k = 70$ and standardized (mean zero, variance one) LFER, Ic, and Iv values.

cross-validated predictive accuracy of kNN was better on the whole set than on each of the subsets ($q^2 = 0.753/F = 140$ vs $q^2 = 0.668/F = 41$ and $q^2 = 0.730/F = 101$; Models 3, 9, and 11, respectively, in Table 1).

One of the objectives of this study was to examine whether the global MLR based on the LFER descriptors had reached the limit of predictive accuracy. On the KC291 data set, the 70NN model achieved (Model 3 in Table 1) cross-validated $q^2(70NN) = 0.753$ and $qms(70NN) = 0.297$, which were better than $q^2 = 0.732$ and $qms = 0.314$ of the global MLR (Model 2). It appeared that the main reason for that was due to the KC201 set being not completely homogeneous as seen from $q^2(70NN) = 0.730 > q^2 = 0.719$ (comparing Models 11 and 10), while KC90 exhibited better homogeneity than KC201 with $q^2(70NN) = 0.668 < q^2 = 0.672$ (Models 9 and 8). For a completely homogeneous data set, any of its subsets would exhibit lower predictive accuracy (on average), that is, $q^2(k\text{NN}) < q^2$. Abraham et al.[1] showed that data on in vivo and in vitro distribution from blood, plasma, and serum to the rat brain could be effectively combined into the single AI328 data set. Our $q^2(k\text{NN})$ results demonstrated that the in vivo subset, 207 compounds from AI328 or the corresponding 201 compounds in KC291, exhibited very minor clustering effects due to groups of data points of a different nature and/or experimental quality. Note that the MCCV results ($qms_{mc}(70NN) = 0.314 < qms_{mc} = 0.316$, Models 3 and 2 in Table 1) also support the existence of minor clustering, but the detected effect is even smaller than when using the q^2 statistic. On the practical level, the presented kNN results supported (1) the proposed merger[1] of logBB data sets of different origins (into a single AI328/KC291 data set) by identifying only very minor clustering effects ($qms_{mc} - qms_{mc}(70NN) = 0.002$), which were negligible when comparing them to the estimated experimental error of about 0.3, and (2) the proposition that the LFER-based logBB MLR model (and hence kNN) achieved the theoretical limit of predictive accuracy of about 0.3 log units obtainable from the AI328 (and hence KC291) data set.[1]

When the MLR, LOO, and kNN results were analyzed, terbinafine (from the KC201 set) was identified as a possible outlier. The compound's logBB in the AI328 data set was 0.095, while 1.059, 1.084, and 1.178 were predicted from the LFER ($S = 1.89$, $E = 1.38$, $A = 0$, $B = 1.03$, and $V = 2.606$) descriptors by the global MLR, LOO, and ($k = 70$)-NN models, respectively. Moreover, ADME Boxes (version 3.5.20) produced $S = 1.74$, $E = 1.34$, $A = 0$, $B = 0.86$, and $V = 2.606$, which resulted in $logBB(70NN) = 1.275$ making the discrepancy between experimental and theoretical results even greater. When terbinafine was removed from KC291, the 70NN model improved $qms_{mc} = 0.314 \rightarrow 0.311$ on the remaining KC290 data set (Model 6 in Table 1). Note that the removal of terbinafine did not improve the homogeneity of the KC200 set (obtained from KC201 by removing terbinafine), $q^2(70NN) = 0.751 > q^2 = 0.732$, Models 13 and 12 in Table 1.

Table 1. Predictive Performance of LogBB−QSAR Models

model	data set method	molecular descriptors (MDs)	q^2 (q_a^2, F)	qms (qms$_{mc}$)	rms$_{test}$
1	AI328[1]-MLR	LFER,[1] Ic,[1,13] Iv[1]	0.736, (0.730, 130)	0.308 (0.313)[a]	
2	KC291-MLR	LFER, Ic, Iv	0.732, (0.725, 114)	0.314 (0.316)[a]	
3	KC291-70NN	LFER, Ic, Iv	0.753, (0.747, 140)	0.297 (0.314)[a]	
4	KC291-**70FN**	LFER, Ic, Iv	0.143, (0.122, 54)	0.553 (0.62)[a]	
5	KC290-MLR	LFER, Ic, Iv	0.741 (0.735, 119)	0.304 (0.307)[a]	
6	**KC290-70NN**	**LFER, Ic, Iv**	**0.766 (0.760, 150)**	**0.290 (0.311)[a]**	
7	KC81-MLR	LFER, Ic, Iv	0.768 (0.746, 39)	0.314 (0.350)[a]	
8	KC90-MLR	LFER	0.672 (0.652, 36)	0.211 (0.219)[a]	
9	KC90-70NN	LFER	0.668 (0.649, 41)	0.212	
10	KC201-MLR	LFER, Ic	0.719 (0.710, 86)	0.342 (0.352)[a]	
11	KC201-70NN	LFER, Ic	0.730 (0.722, 101)	0.335 (0.347)[a]	
12	KC200-MLR	LFER, Ic	0.732 (0.723, 91)	0.335 (0.345)[a]	
13	KC200-70NN	LFER, Ic	0.751 (0.743, 112)	0.323 (0.340)[a]	
14	KC291-MLR	V1,[b] Ic, Iv	0.581 (0.574, 25)	0.387 (0.393)[a]	
15	KC291-MLR	V2,[b] Ic, Iv	0.527 (0.519, 28)	0.411 (0.416)[a]	
16	KC291-MLR	V3,[b] Ic, Iv	0.530 (0.522, 29)	0.409 (0.415)[a]	
17	KC81-MLR	V1, Ic, Iv	0.601 (0.574, 25)	0.412 (0.436)[a]	
18	NG88[4]-MLR	V1	0.679 (66)	0.416	0.539[c]
19	KC81-MLR	V2, Ic, Iv	0.622 (0.597, 28)	0.401 (0.425)[a]	
20	NG88[4]-MLR	V2	0.676 (67)	0.418	0.548[c]
21	KC81-MLR	V3, Ic, Iv	0.637 (0.613, 29)	0.393 (0.421)[a]	
22	NG88[4]-MLR	V3	0.664	0.428	0.566[c]
23	RH106[12]-MLR	3 MDs	0.62		0.48
24	RH106[43]-MLR	5 MDs	0.70		0.43
25	WD103[16]-MLR	3 MDs	0.68	0.42	
26	OW79[14]-PLS	31 MDs	0.65		
27	CB119[8]-MLR	3 MDs	0.648		0.428
28	KK113[9]-CODESSA	5 MDs	0.752		0.179
29	KK113[9]-ISIDA	2800 MDs	0.803		0.05
30	ML191[7]-MLR	30+ MDs	0.53	(0.57, N=10, n_v = 47)	
31	S57[5]-PLS	3 MDs	0.552		0.326
32	HX115[6]-MLR	3 MDs	0.736		0.490[d]
33	HX96[15]-MLR	4 MDs	0.701		
34	FS61[41]-MLR	3 MDs	0.688		0.628−0.789
35	PI150[42]-MLR	2−3 MDs	0.60−0.83		
36	SK55[44]-MLR/PLS	2−7 MDs	0.786−0.811		0.490−0.708
37	GV132[10]-ANN	8 MDs			0.319[e]

[a] MCCV with N = 30 000 and $n_v = n/2$. [b] V1: PSA,[32] SsssN,[35] and ALOGP.[36] V2: S1K,[33] SsssN, and ALOGP. V3: Xu,[34] SsssN, and ALOGP. [c] Calculated from Table 3 and supplementary tables of Narayanan and Gunturi.[4] [d] Calculated from Table 6 of Hou and Xu.[6] Average of 0.576 and 0.616 from Table 1 of Feher et al.[41] [e] Calculated from test data set in Table 3 of Garg and Verma.[10]

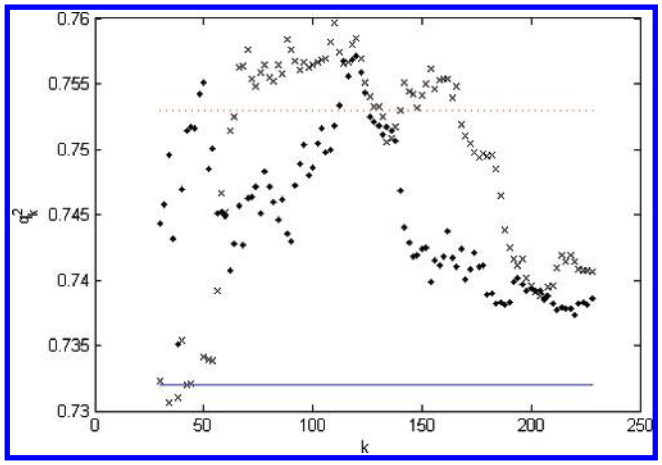

Figure 2. The *k*NN-MLR model of the KC291-LFER data set. Solid and dotted horizontal lines are the $q^2 = 0.732$ and $r^2 = 0.753$ levels obtained by LOOCV and MLR of the whole KC291-LFER data set, respectively. Crosses and dots are q_k^2 obtained with and without standardization of the LFER descriptor values, respectively.

Given the difficulty in obtaining the experimental logBB values, it is generally assumed that structural descriptors could be calculated with high accuracy and consistency. However, even in the case of well-studied[28] LFER descriptors, this is not the case; for example, the ADME Boxes program produced LFER values for terbinafine somewhat different from the LFER values in the AI238 database.[1] We examined a few other components, and most *S*, *E*, *A*, and *B* descriptor values were reported by ADME Boxes as different from those of the AI238 database. For example, acetylsalicylic acid (2-acetoxybenzoic acid, aspirin, logBB = −0.5,[12] SMILES = "CC(=O)Oc1ccccc1C(=O)O") yielded LFER values of $A = 0.57$, $B = 0.77$, $L = 6.113$, $S = 1.42$, $E = 0.84$, and $V = 1.2879$ when calculated by ADME Boxes in comparison to those of the experimental LFER ($A = 0.49$, $B = 1$, $L = 5.458$, $S = 0.8$, $E = 0.781$, and $V = 1.2879$).[13] Given that the LFER-based models achieved one of the highest predictive accuracy values, for consistency, the LFER values in the KC291/AI328 data set should be recalculated by the ADME Boxes since the program is currently the preferred source of the LFER values.[29]

Narayanan and Gunturi (2005).[4] When SMILES from Table S1[1] are used, more than 2000 molecular descriptors could be calculated for the compounds in the KC291 data set via the PCLIENT[19,30] Web site. The **Iv**[1] indicator was retained from the AI328/KC201-LFER data set since it was required to distinguish the origin of the data points (in vitro

or in vivo). The **Ic** indicator variable[1,13] for carboxylic acids (**Ic** = 1) was also retained to assist with a more consistent comparison with the A-V LFER descriptors. When required, the conversion between SMILES and the 2D structure was performed using ACD/ChemSketch version 10.0, which is freely available from www.acdlabs.com.[31]

Narayanan and Gunturi[4] identified six combinations of descriptors (V1−V6 models) for the logBB prediction. We limited our analysis to the first three models, which were based on only three descriptors. The following are the names of the five PCLIENT descriptors with the corresponding descriptor numbers from Narayanan and Gunturi:[4] TPSA[32] (from *molecular properties*) for 320 (2D van der Waals surface area); S1K[33] (from *topological descriptors*) for 144 (κ shape index), Xu[34] (from *topological descriptors*) for 30 (topological Xu index), SsssN[35] (from *ET-state indices*) for 254 (atomic type E-state index), and ALOGP[36,37] (from *molecular properties*) for 311 (AlogP98). The considered descriptors were found to be highly correlated. For example, S1K values could be regressed from the remaining four descriptors with $r^2 = 0.952$, $r_a^2 = 0.951$, $s = 1.228$, and $F = 1418$, where MLR was performed on the KC291-NG data set containing the five descriptors for the compounds from the KC291 data set (see Table S1).

Narayanan and Gunturi[4] used 88 compounds with the corresponding logBB values from Rose et al.[12] in their training set (denoted NG88). The NG88 set contained the following seven compounds which were excluded as outliers from the AI328[1] database of logBB data points and hence from the current KC291 data set: SK&F93319 (SKB4) [−1.3],[13,38] icotidine (SKB3) [−2.0],[38] BBCPD11 (SKB13) [−2.15],[38] cimetidine (SKB1) [−1.42],[13] ranitidine (SKB9) [−1.23],[13] phenserine [1.0],[13] and acetylsalicylic acid [−0.5],[13] where the numbers in square brackets are the corresponding logBB values from Rose et al.[12] and where the compound codes cross-referenced to Table S1[1] are in round brackets.

The KC81 set of compounds was created by retaining only the compounds contained in both KC291 and NG88 data sets, and retaining the logBB, Ic, and Iv values from KC291. The selected PCLIENT descriptors were verified by reproducing the original results of Narayanan and Gunturi[4] with sufficient accuracy, see the KC81 and NG88 rows in Table 1 (Models 17 and 18, 19 and 20, and 21 and 22). In particular, our qms values were especially close to the reported standard error (SE), 0.412 versus 0.416, 0.401 versus 0.418, and 0.393 versus 0.428. For the benchmark case of the whole KC291 set, the predictive power ($0.519 \le q_a^2 \le 0.574$, $0.387 \le \text{qms} \le 0.411$) of the V1−V3 descriptors (Models 14, 15, and 16) was significantly lower than the power of the LFER descriptors ($q_a^2 = 0.725$ and qms = 0.314; Model 2), where q_a^2 was used rather than q^2 to account for the difference in the number of descriptors (seven in KC291-LFER and five in each of the KC291−V1/V2/V2 sets).

In the case of KC291-LFER data set, the LOOCV and MCCV results were in qualitative agreement with each other. In the case of KC291-NG, the situation was somewhat different. From the LOOCV point of view, all three (V1−V3) models exhibited lower predictive power (if measured by q^2) when presented with a larger set (KC291) compared to their original NG88/KC81 data set. The predictive accuracy of the V1, V2, and V3 models dropped from $q^2 = 0.601$ to $q^2 = 0.581$ (Models 17 and 14), $q^2 = 0.622$ to $q^2 = 0.527$ (Models 19 and 15), and $q^2 = 0.637$ to $q^2 = 0.530$ (Models 21 and 16). However, the MCCV results indicated a slight improvement in the predictive power: $\text{qms}_{mc} = 0.436$ (Model 17) improved to $\text{qms}_{mc} = 0.393$ (Model 14), $\text{qms}_{mc} = 0.425$ (Model 19) to $\text{qms}_{mc} = 0.416$ (Model 15), and $\text{qms}_{mc} = 0.421$ (Model 21) to $\text{qms}_{mc} = 0.415$ (Model 16). The mixed LOOCV results were obtained for qms mixed with only the V1 model improved when the model was applied to the complete KC291 data set.

This discrepancy between the LOO and LGO (via MCCV) cross-validation results was very interesting because it forced the issue of which cross-validation method was better. Shao[24] showed that the LGO cross-validation method had a higher probability of selecting a MLR model with higher predictive power than the LOOCV method, and therefore the qms_{mc} statistic (or MSEP(n_v)) should be preferred over the q^2 and qms statistics. Also, the q^2 values reported by the descriptor-mining logBB models might be overestimated and therefore should not be used for the assessment of the models' predictive power; that is, smaller q^2 values obtained on a larger data set might not mean the reduction of predictive accuracy when measured by qms_{mc}. The considered case study also showed that the qms_{mc} statistic could be quite resilient to the size of a data set, further supporting its suitability as a benchmarking quantity: the $\text{qms}_{mc}(n_v = {}^1/_2 n = 40)$ values obtained from the KC81−V1/V2/V3 models were comparable with the corresponding $\text{qms}_{mc}(n_v = {}^1/_2 n = 145)$ values obtained from the KC291−V1/V2/V3 models: qms_{mc}(KC81−V1) = 0.436 versus qms_{mc}(KC291−V1) = 0.393; qms_{mc}(KC81−V2) = 0.425 versus qms_{mc}(KC291−V2) = 0.416; and qms_{mc}(KC81−V3) = 0.421 versus qms_{mc}(KC291−V3)=0.415.

The MCCV results also highlighted the problem of using a single cross-validation split of data into calibration and validation subsets. In the case of the V1−V3 models, RMSEP (denoted rms_{test} in Table 1) calculated ($N = 1$, $n_c = 88$) from the predicted logBB values for $n_v = 28$ validation compounds was around 0.539−0.566, which was actually much higher than the correct value of around 0.393−0.415 obtained from $N = 30\,000$ Monte Carlo cross-validations with $n_v = 140$ and $n_c = 141$.

In summary, in the benchmarked case study for the logBB models of Narayanan and Gunturi,[4] the originally reported q^2 was overestimated by at least 0.1, while the actual predictive accuracy of the models (as measured by qms_{mc}) was underestimated by at least 0.1 log unit.

Other LogBB-QSAR Models. Exact benchmarking of all or even the most recent logBB-QSAR models[5−9,12,14−16] was outside the scope of this paper since the utilized descriptors were calculated by mostly proprietary software programs which were in some cases available for purchase. Currently, an academically positive trend is emerging where a large number of descriptors are becoming freely available for computation, for example, via the Virtual Computational Chemistry Laboratory,[19] the Chemistry Development Kit (CDK),[39] and MODEL[40] Web sites/programs. However, the freely calculated descriptors may not match the previously reported descriptors exactly, for example, due to differences in descriptor and/or three-dimensional molecular structure optimization algorithms. For example, in the case of Narayanan and Gunturi,[4] we attempted to match the required

descriptors to the freely available descriptors and obtained results similar but not identical to the reported results. Hence, the following comparison of predictive accuracy of the logBB-QSAR models had only an indicative nature. Most of the considered models reported their results for various test sets (not used for training). However, because of the different testing sets, it was not possible to compare the models consistently, leaving the question of which model should be preferred in practice unanswered.

As discussed, ideally, the qms_{mc} statistic should be used for benchmarking, which was rarely reported.[7] In the absence of the qms_{mc} statistic, qms is the next most suitable predictive statistic. In the situations when even qms was not reported, there was a common tendency to present rms_{test} obtained from a single LGO split of the available data sets. Such a practice is not statistically meaningful[21] when applied to small validation subsets; for example, $n_v = 13$ yielded either $rms_{test} = 0.326$ or 0.671 depending on the composition of the validation subset.[5] As shown for the models of Narayanan and Gunturi,[4] a single split with a relatively small number of validation compounds could yield an rms_{test} significantly different from the "correct" (e.g., MCCV) value. In the case of Narayanan and Gunturi,[4] rms_{test} was 0.1 higher than the much more accurate $qms_{mc} = 0.4$, meaning that it would be feasible to obtain $rms_{test} = 0.3$ (i.e., 0.1 lower than qms_{mc}) just by a fortunate selection of the validation subset.

The models of Wichmann et al.[16] (qms = 0.42; Model 25 in Table 1), Cabrera et al.[8] ($s_{press} = 0.428$; Model 27), Sun[5] ($rms_{test} = 0.326$; Model 31), Feher et al.[41] ($rms_{test} = 0.628-0.789$; Model 34), Pan et al.[42] (Model 35), and Hou and Xu[6] ($rms_{test} = 0.490$; Model 32) used a preordained set of descriptors, while the models of Narayanan and Gunturi[4] (Models 18, 20, and 22), Rose et al.[12,43] ($s_{press} = 0.43-0.48$; Models 23 and 24), Ooms et al.[14] (Model 26), Mente and Lombardo[7] ($N = 10$, $n_c = 142$, $n_v = 47$, and $qms_{cv} = 0.57$; Model 30), Katritzky et al.[9] ($n_v = 19$, s_{test}(CODESSA) = 0.179, and s_{test}(ISIDA) = 0.05; Models 28 and 29), Hou and Xu[15] (Model 33), Pan et al. (Model 35), Subramanian and Kitchen[44] ($n_c = 142$, $n_v = 39$, and $0.490 \leq rms_{test} \leq 0.708$; Model 36), and Garg and Verma[10] ($n_v = 50$ and $s_{test} = 0.319$; Model 37) were developed by data mining (to various extents) of the available descriptors to achieve the best fit of their respective training sets. The case study of the Narayanan and Gunturi[4] models indicated that it was likely that the models which reported qms would retain their reported predictive qms statistics when applied to a larger data set such as KC291. On the other hand, rms_{test} values do not reflect the actual predictive accuracy of the models since the values were obtained from a single LGO split with a relatively low n_v. The exact verification would require recalculation of the corresponding descriptors, which is outside the scope of this study.

The predictive statistics of the LFER-MLR/*k*NN models ($0.290 \leq qms \leq 0.314$ and $0.311 \leq qms_{mc} \leq 0.316$; Models 2, 3, 5, and 6 in Table 1) indicated that the LFER-based models had the best cross-validated predictive accuracy of all considered logBB models. Out of the considered models, only those of Sun[5] (Model 31), Katritzky et al.[9] (Models 28 and 29), and Garg and Verma[10] (Model 37) achieved a rms_{test} accuracy comparable to the accuracy of LFER-based models ($0.290 \leq qms \leq 0.314$). However, the LFER models had their qms values verified by qms_{mc} (qms values were underestimated by only at most 0.02). The model of Garg and Verma[10] was an ANN which provided little transparency in how each of the eight used descriptors contributed to logBB. This makes the model less helpful in assisting the design of a compound with desirable logBB properties. The same applies to the ISIDA model of Katritzky et al.,[9] which used more than 2800 fragment descriptors. The CODESSA model of Katritzky et al.[9] used five descriptors which were selected from a pool of 475 descriptors. The CODESSA results are promising, but the proposed MLR model was trained on 113 compounds and validated on an external test set of only 19 compounds; hence, it remains to be seen how well the model performs on the KC291/AI328 data sets in terms of MCCV.

Our intension was to highlight the academic need for current and future logBB models to be reported in a way that could be independently reproduced (with minimum cost) and applied to a standard benchmark (e.g., qms_{mc} statistic on the KC291 data set) for objective and consistent comparison. However, the benchmark may not resolve the issue of data-mined descriptors, where all available data may be used for descriptor selection. In some areas of bioinformatics, this problem is addressed by objective testing of the models via the process of blind prediction. For example, the critical assessment of methods of protein structure prediction[45,46] has been successfully run for more than 10 years, where experimental results of the three-dimensional structure of proteins were withheld while the sequences of experimentally solved proteins were released to the public.

CONCLUSIONS

The *k*NN method, with the *k*NN-MLR model being a MLR instance of it, works in line with the central premise of medicinal chemistry in that structurally similar molecules have similar biological activities.[47] The *k*NN-MLR method does not make any assumptions about the presence or the number of clusters in a sample and works equally well even in the presence of nonlinear effects and should at least be considered whenever the MLR method is used.[18]

In the case of the considered KC291/AI328 logBB benchmark, a variety of logBB data sources were combined, which were different in quality, method (in vivo and in vitro), and even medium (blood, plasma, and serum). On such a compositionally diverse data set, the *k*NN-MLR method achieved better predictive results comparing to the global MLR method. However, the improvement was quite minor and well within the experimental error of the data set, confirming the statistical validity of the data amalgamation.

Using the *k*NN-MLR method, we could not invalidate the implied suggestion of Abraham et al.[1] in that it might not be possible to develop a logBB model with RMSEP (cross-validated or not) significantly better than about 0.3 due to the experimental errors in the KC291/AI328 data set. Therefore, any logBB model that is trained on KC291/AI328 (or its subset) and yields a RMSEP (on calibration or validation subsets) of significantly less than 0.3 is likely to be overfitting.

A benchmarking procedure based on the KC291 data set and LOO/MC cross-validation was proposed to start addressing the question of how to consistently compare the predictive accuracy of the available logBB models. The question is of practical importance for the pharmaceutical

industry since most reported models are either (1) not readily available and therefore must be reproduced "in-house", incurring potentially considerable developmental expense, or (2) available for purchase and hence the expenditure must be justified. The first benchmarking results indicated that the LFER[1] descriptors remained to be one of the most accurate for predicting logBB. The logBB models of Narayanan and Gunturi[4] were examined as a case study of the data-mining approach to descriptor selection. The models could not reproduce their reported q^2 values on a larger data set such as the KC291 data set, while their qms and qms_{mc} statistics remained much more stable. This indicated that the q^2 statistic might not be applicable as a measure of the predictive power of the models and was merely a measure of internal self-consistency.

The reported MLR, LOOCV, MCCV, and *k*NN results in Table 1 were obtained (and hence could be easily reproduced) via the Java-based QSAR-BENCH program, which is freely available from www.dmitrykonovalov.org and could be run on all Java-enabled computational platforms (Java 1.5+ is required), for example, Windows, Mac OSX, and Unix/Linux. The program can import comma- or tab-delimited text files containing calibration/validation (a $n \times (p + 1)$ **Z** matrix is expected) and prediction (a $n \times p$ **X** matrix is expected) QSAR data sets.

It has been known for quite sometime[24] that the leave-group-out cross validation (LGOCV) method should be preferred over the LOOCV method for assessing the predictive power of MLR models. However, among the considered logBB models only the study of Mente and Lombardo[7] reported LGO cross-validated statistics (with $N = 10$). The low rate of LGO acceptance may be due to the lack of such functionality in commonly used statistical packages. By releasing the freely available QSAR-BENCH program (which could perform LGOCV via the MCCV method for MLR), we hope to assist in the acceptance of the LGOCV method, which is known to be statistically more valid than LOOCV (at least for the MLR-based QSAR models).

And finally, the usefulness of any predictive model for logBB is not simply due to the lowest qms (or qms_{mc}) obtained by the model. The model should consist of methods/algorithms which could be easily retrained as new data become available. The models with a preordained set of descriptors clearly satisfy this requirement, while the descriptor-mining models may not as new data may yield a new set of "best" descriptors. Another practically important requirement is for the equations or algorithms to have a clear interpretation so that a drug designer would know what needs to be done in order to increase or decrease logBB. While the LFER-MLR model is probably the most accurate predictive logBB model currently available, the LFER descriptors are based on empirically determined LFER values for simple molecules or "fragments". The LFER descriptors for more complex molecules are calculated by a "black box" via a set of quite complicated fragment-based rules,[48] which are difficult to interpret and even reproduce.[49] Therefore, it is still desirable to develop logBB models, for example, similar to the ones developed by Narayanan and Gunturi,[4] which are based on clearly defined structural descriptors and comparable in accuracy to the LFER-based logBB models.

ACKNOWLEDGMENT

We thank Michael Abraham for supplying the AI328 data set in electronic format, Pharma Algorithms for supplying an evaluation version of the ADME Boxes program, the CODESSA-PRO team for supplying a demo version of the CODESSA-PRO program, and Anton Hopfinger and two anonymous reviewers for helpful comments on earlier versions of this manuscript.

Supporting Information Available: The KC291 data set and *k*NN-MLR results. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Abraham, M. H.; Ibrahim, A.; Zhao, Y.; Acree, W. E. A data base for partition of volatile organic compounds and drugs from blood/plasma/serum to brain, and an LFER analysis of the data. *J. Pharm. Sci.* **2006**, *95*, 2091−2100.
(2) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: Towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.
(3) Clark, D. E. In silico prediction of blood-brain barrier permeation. *Drug Discovery Today* **2003**, *8*, 927−933.
(4) Narayanan, R.; Gunturi, S. B. In silico ADME modelling: prediction models for blood-brain barrier permeation using a systematic variable selection method. *Bioorg. Med. Chem.* **2005**, *13*, 3017−3028.
(5) Sun, H. A Universal Molecular Descriptor System for Prediction of LogP, LogS, LogBB, and Absorption. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 748−757.
(6) Hou, T. J.; Xu, X. J. ADME Evaluation in Drug Discovery. 3. Modeling Blood-Brain Barrier Partitioning Using Simple Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2137−2152.
(7) Mente, S. R.; Lombardo, F. A recursive-partitioning model for blood-brain barrier permeation *J. Comput.-Aided Mol. Des.* **2005**, *19*, 465−481.
(8) Cabrera, M. A.; Bermejo, M.; Perez, M.; Ramos, R. TOPS-MODE approach for the prediction of blood-brain barrier permeation. *J. Pharm. Sci.* **2004**, *93*, 1701−1717.
(9) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Dobchev, D. A.; Fara, D. C.; Karelson, M.; Acree, J. W. E.; Solov'ev, V. P.; Varnek, A. Correlation of blood-brain penetration using structural descriptors. *Bioorg. Med. Chem.* **2006**, *14*, 4888−4917.
(10) Garg, P.; Verma, J. In Silico Prediction of Blood Brain Barrier Permeability: An Artificial Neural Network Model. *J. Chem. Inf. Model.* **2006**, *46*, 289−297.
(11) 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.
(12) Rose, K.; Hall, L. H.; Kier, L. B. Modeling Blood-Brain Barrier Partitioning Using the Electrotopological State. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 651−666.
(13) Platts, J. A.; Abraham, M. H.; Zhao, Y. H.; Hersey, A.; Ijaz, L.; Butina, D. Correlation and prediction of a large blood-brain distribution data set−an LFER study. *Eur. J. Med. Chem.* **2001**, *36*, 719−730.
(14) Ooms, F.; Weber, P.; Carrupt, P. A.; Testa, B. A simple model to predict blood-brain barrier permeation from 3D molecular fields. *Biochim. Biophys. Acta* **2002**, *1587*, 118−125.
(15) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery - 1. Applications of genetic algorithms to the prediction of blood-brain partitioning of a large set of drugs. *J. Mol. Model.* **2002**, *8*, 337−349.
(16) Wichmann, K.; Diedenhofen, M.; Klamt, A. Prediction of Blood-Brain Partitioning and Human Serum Albumin Binding Based on COSMO-RS σ-Moments. *J. Chem. Inf. Model.* **2007**, *47*, 228−233.
(17) *ADME Boxes*; PharmaAlgorithms Inc: Toronto, Canada. http://www.ap-algorithms.com/adme_boxes.htm (accessed April 29, 2007).
(18) Guha, R.; Dutta, D.; Jurs, P. C.; Chen, T. Local Lazy Regression: Making Use of the Neighborhood to Improve QSAR Predictions. *J. Chem. Inf. Model.* **2006**, *46*, 1836−1847.
(19) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V.; Radchenko, E.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory - design and description. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453−463.
(20) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; WILEY-VCH: New York, 2000.

(21) Duffy, E. M.; Jorgensen, W. L. Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water. *J. Am. Chem. Soc.* **2000**, *122*, 2878−2888.
(22) Xu, Q. S.; Liang, Y. Z. Monte Carlo cross validation. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 1−11.
(23) Xu, Q. S.; Liang, Y. Z.; Du, Y. P. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *J. Chemom.* **2004**, *18*, 112−120.
(24) Shao, J. Linear Model Selection by Cross-Validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486−494.
(25) Schölkopf, B.; Smola, A. J. *Learning with Kernels*; The MIT Press: Cambridge, MA, 2001.
(26) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.
(27) Vapnik, V. *Statistical Learning Theory*; John Wiley and Sons: New York, 1998.
(28) Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835−845.
(29) Abraham, M. H. Personal communication, 2007.
(30) VCCLAB Virtual Computational Chemistry Laboratory. www.vcclab.org (accessed May 2, 2007), 2005.
(31) ACD/ChemSketch. www.acdlabs.com (accessed May 2, 2007).
(32) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(33) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109−116.
(34) Ren, B. A New Topological Index for QSPR of Alkanes. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 139−143.
(35) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.
(36) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure−Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(37) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for 3-Dimensional Structure-Directed Quantitative Structure-Activity-Relationships. 1. Partition-Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565−577.
(38) Kaznessis, Y. N.; Snow, M. E.; Blankley, C. J. Prediction of blood-brain partitioning using Monte Carlo simulations of molecules in water. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 697−708.
(39) Steinbeck, C.; Han, Y. Q.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.
(40) Li, Z. R.; Han, L. Y.; Xue, Y.; Yap, C. W.; Li, H.; Jiang, L.; Chen, Y. Z. MODEL-Molecular Descriptor Lab: A web-based server for computing structural and physicochemical features of compounds. *Biotechnol. Bioeng.* **2007**, in press.
(41) Feher, M.; Sourial, E.; Schmidt, J. M. A simple model for the prediction of blood-brain partitioning. *Int. J. Pharm.* **2000**, *201*, 239−247.
(42) Pan, D.; Iyer, M.; Liu, J.; Li, Y.; Hopfinger, A. J. Constructing Optimum Blood Brain Barrier QSAR Models Using a Combination of 4D-Molecular Similarity Measures and Cluster Analysis. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2083−2098.
(43) Rose, K.; Hall, L. H.; Hall, L. M.; Kier, L. B. Modeling Blood-Brain Barrier Partitioning Using Topological Structure Descriptors; Elsevier: New York, 2003. Elsevier MDL Web Site. http://www.mdl.com/solutions/white_papers/MDLQSARreprint.jsp (accessed Mar 1, 2007).
(44) Subramanian, G.; Kitchen, D. B. Computational models to predict blood−brain barrier permeation and CNS activity. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 643−664.
(45) CASP Protein Structure Prediction Center. http://predictioncenter.gc.ucdavis.edu/ (accessed Mar 1, 2007).
(46) Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 334−339.
(47) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(48) Japertas, P.; Didziapetris, R.; Petrauskas, A. Fragmental Methods in the Design of New Compounds. Applications of The Advanced Algorithm Builder. *Quant. Struct.-Act. Relat.* **2002**, *21*, 23−37.
(49) Leo, A. J. Calculating log P_{oct} from Structures. *Chem. Rev.* **1993**, *93*, 1281−1306.

CI700100F