Statistical Confidence for Variable Selection in QSAR Models via Monte Carlo Cross-Validation

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

School of Mathematics, Physics and Information Technology, James Cook University, Townsville, Queensland 4811, Australia, Laboratory for Pharmaceutical Technology and Biopharmacy, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium, and Department of Analytical Chemistry and Pharmaceutical Technology, Pharmaceutical Institute, Vrije Universiteit Brussel, B-1050 Brussels, Belgium

Received August 1, 2007

A new variable selection wrapper method named the Monte Carlo variable selection (MCVS) method was developed utilizing the framework of the Monte Carlo cross-validation (MCCV) approach. The MCVS method reports the variable selection results in the most conventional and common measure of statistical hypothesis testing, the *P*-values, thus allowing for a clear and simple statistical interpretation of the results. The MCVS method is equally applicable to the multiple-linear-regression (MLR)-based or non-MLR-based quantitative structure−activity relationship (QSAR) models. The method was applied to blood−brain barrier (BBB) permeation and human intestinal absorption (HIA) QSAR problems using MLR to demonstrate the workings of the new approach. Starting from more than 1600 molecular descriptors, only two (TPSA(NO) and ALOGP) yielded acceptably low *P*-values for the BBB and HIA problems, respectively. The new method has been implemented in the QSAR-BENCH v2 program, which is freely available (including its Java source code) from www.dmitrykonovalov.org for academic use.

INTRODUCTION

Quantitative structure−activity/property relationship (QSAR/QSPR) models[1] are often developed by examining a large number of descriptors based on the molecular structures of sample compounds, where the number of compounds in the sample (n) is smaller or even much smaller than the number of considered descriptors (p), i.e., $n < p$ or $n \ll p$. Diverse automatic variable (i.e., feature) selection[2] techniques exist to arrive at a much smaller subset of descriptors (of size k, $k \ll p$).[3−11] For example, when a QSAR model utilizes the multiple linear regression (MLR) method, Pearson's coefficient of regression (r) or its leave-one-out (LOO) cross-validated equivalent (q) is often used to justify or guide the selection. Both MLR and non-MLR classes of QSAR models normally report how well the models fit the calibration (i.e., training) subset by reporting the mean-squared error (MSE) of calibration,

$$\mathrm{MSE} = \sum_{i=1}^{n_c} (y_i - y_i')^2 / n_c \qquad (1)$$

or its root-mean-squared error (RMSE $= \sqrt{\mathrm{MSE}}$), where y_i' is the activity value estimated for the ith compound and n_c is the size of the calibration subset, and where the ith compound was *included* in the calibration subset. Within the field of machine learning, MSE is also known as empirical error or empirical risk.[12] MSE represents the data fitting ability of a model,[13] where the more flexible the model is in terms of the number of unconstrained parameters or coefficients, the easier it is to minimize MSE.

Even though the MSE values are still commonly reported for QSAR models, the mean-squared error of prediction,

$$\mathrm{MSEP} = \sum_{i=1}^{n_v} (y_i - y_i')^2 / n_v \qquad (2)$$

or its square root (RMSEP $= \sqrt{\mathrm{MSEP}}$) is more suitable for expressing the *predictive* power or accuracy of a QSAR model,[14] where the same notation (y_i') is used but in this case y_i' is the activity value predicted for the ith compound by the QSAR model and n_v is the size of the validation (i.e., test) subset, and where the ith compound was *excluded* from the calibration subset.

It is interesting to note that there exists a wide discrepancy between various theories on how QSAR models should be assessed and how QSAR models are actually assessed in practice. For example, it was pointed out a number of times that the widespread practice of dividing a relatively small QSAR data set (not larger than a few hundred compounds) into calibration and validation subsets was statistically meaningful only for significantly larger data sets.[15] Furthermore, in the case of the MLR-based QSAR models, Shao[16] demonstrated that the leave-one-out cross-validation (LOO-CV) method was inferior to the leave-group-out cross-validation (LGO-CV) method. In particular, the probability of the LOO-CV method to select the MLR model with the

* Corresponding author fax: (617) 4781 5880; e-mail: dmitry.konovalov@jcu.edu.au.
† James Cook University.
‡ Katholieke Universiteit Leuven.
§ Vrije Universiteit Brussel.

10.1021/ci700283s CCC: $40.75
Published on Web 01/31/2008

best predictive ability does not converge to 1 as the total number of observations $n \rightarrow \infty$.[16] Nevertheless, the studies using LOO-CV by far out-number the studies with LGO-CV. If one accepts theoretical arguments[16] in favor of the LGO-CV method, the correct application of the LGO-CV method requires a large number of validation and calibration subsets, e.g., via Monte Carlo cross-validation (MCCV).[17] In practice, however, the available data set is usually divided into a single calibration subset (of size n_c) and a single validation subset (of size $n_v < n_c$ or even $n_v \ll n_c$), i.e., essentially performing a single iteration of the LGO-CV method, which is often referred to as a *holdout*. Validation results obtained from such a single LGO-CV iteration with a relatively small holdout subset ($n_v < 100$) are known to be unreliable,[18] but the problem is often acknowledged just to justify the use of the LOO-CV method in addition to the use of the holdout subset.[15]

There now exist strong indications[16,17] that the MCCV method is a practical approach for measuring (via MSEP) the predictive power of a QSAR model, where the larger the n_v/n ratio, the more accurate MSEP becomes.[16] For example, the MCCV ($n_v/n = 0.5$) method was recently applied to benchmarking of the QSAR models for blood−brain barrier (BBB) permeation, where it was shown that MSEP (cross-validated via MCCV) correctly described the predictive limits of the MLR and *k*-nearest neighbors (*k*NN) methods on the considered data sets.[19] While the MCCV method, arguably, solves the problem of assessing the predictive power of a QSAR model with pre-ordained descriptors,[19] the situation is different for this study, when a much smaller subset of descriptors must be selected from a large pool of available descriptors. The currently existing descriptor selection methods represent significant academic advances in this field; however, there is very little consensus on which method should be preferred (see, for example, a recent review by Dudek et al.[1]). One possible reason for the low uptake of any individual method is that the final results of the methods are difficult to interpret by a person who is not a statistician.

The main objective of this study was to develop a variable selection method that has a clear and simple statistical interpretation for a user (e.g., medicinal chemist) with minimal statistical training. A new method, named the Monte Carlo variable selection (MCVS) method, was developed utilizing the framework of MCCV and reporting the variable selection results in the most conventional and common measure of statistical hypothesis testing, the *P*-values. Using the *filtering* and *wrapper* classification,[2] the MCVS method belongs to a wrapper class of variable selection methods and permits a parallel application of any filtering methods. The MCVS method is equally applicable to the MLR-based or non-MLR QSAR models. The method was applied to BBB permeation and human intestinal absorption (HIA) QSAR problems to demonstrate its workings.

MATERIALS AND METHODS

Monte Carlo Cross-Validation (MCCV). Let a sample of n compounds be described by a $n \times (p + 1)$ matrix $\mathbf{Z}$ with the following structure,

$$\mathbf{Z} = \begin{pmatrix} z_{11} & \cdots & z_{1,p+1} \\ \vdots & \ldots & \vdots \\ z_{n1} & \cdots & z_{n,p+1} \end{pmatrix} = \begin{pmatrix} y_1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ldots & \vdots \\ y_n & x_{n1} & \cdots & x_{np} \end{pmatrix} = (Y, \mathbf{X}) \quad (3)$$

where $Y = (y_1, y_2, ..., y_n)^T$ is the column (i.e., an $n \times 1$ matrix) of activity or response values (the superscript "T" denotes the transpose), $\mathbf{X}$ is an $n \times p$ matrix of descriptor values, and p is the number of structure descriptors or, generally speaking, any predictor variables. The $\mathbf{X}$ matrix could be referenced by its columns (i.e., by descriptors),

$$\mathbf{X} = \left(\begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix} \cdots \begin{bmatrix} x_{1p} \\ \vdots \\ x_{np} \end{bmatrix} \right) = (D_1, ..., D_p) \quad (4)$$

where $D_j = (x_{1j}, x_{2j}, ..., x_{nj})^T$ is the column of the jth descriptor values, $1 \le j \le p$. Another convenient way of referring to the data in the matrix is by rows (i.e., by compounds),

$$\mathbf{X} = \begin{pmatrix} [x_{11}, ..., x_{1p}] \\ \cdots \\ [x_{n1}, ..., x_{np}] \end{pmatrix} = \begin{pmatrix} X_1 \\ \cdots \\ X_n \end{pmatrix} \quad (5)$$

where $X_i = (x_{i1}, x_{i2}, ..., x_{ip})$ is the row of descriptor values for the ith compound, $1 \le i \le n$. The activity response of the ith compound predicted by a QSAR model is denoted by y'_i.

Let $S = \{1, 2, ..., n\}$ denote the complete set of the available compounds numbered from 1 to n. A single iteration of the MCCV method[17] splits S into a validation subset $S_v(I) = S_v(i_1, i_2, ..., i_{n_v})$ (of size $|S_v| = n_v$) and a calibration subset $S_c(I) = S_c(i_{n_v} + 1, i_{n_v} + 2, ..., i_n)$ (of size $|S_c| = n_c = n - n_v$), where $S_v \cap S_c = 0$, $S_v \cup S_c = S$, I is the partition of S into two subsets, and $|\cdot|$ denotes the cardinality operator. The subsets are then encoded by bit set vectors[20,21] of 1's and 0's such that, for example, $S_v = (0_1, 0_2, ..., 0_{i_1-1}, 1_{i_1}, 0_{i_1+1}, ..., 0_{i_{n_v}-1}, 1_{i_{n_v}}, 0_{i_{n_v}+1}, ..., 0_n)$, where without loss of generality $i_1 < i_2 < ... < i_{n_v}$ and the value of "1" in the ith position means that the ith compound is included in the subset.

The case of descriptors is handled similarly to the compounds, where $H = \{1, 2, ..., p\}$ denotes the complete set of descriptors numbered from 1 to p. Then $J = \{j_1, j_2, ..., j_k\}$ denotes a variable selection hypothesis that selects k (i.e., $|J| = k$) descriptors from H; that is, J is a subset of H. Note that the i- and j-based indices are used to distinguish the compound- and descriptor-based entities throughout this study, respectively. A complete set of all available J hypotheses which select exactly k descriptors is denoted by H_k, where $|H_k| = p!/(k!(p - k)!)$.

For a single MCCV iteration, the predictive accuracy of a QSAR model is assessed by the MSEP (denoted as the loss function $L(I,J)$ for brevity),

$$L(I,J) = \mathrm{MSEP}(S_c(I), S_v(I), J) = \sum_{m=1}^{n_v} (y_{i_m} - y'_{i_m}(J))^2/n_v \quad (6)$$

which is reported as its square root (RMSEP), where $y'_{i_m}(J)$ is the activity value predicted by the QSAR model calibrated on the $S_c(I)$ subset of compounds and using the J subset of

descriptors. The MCCV method repeats the above procedure N times, obtaining the average MSEP[16] (denoted as $L(J)$),

$$L(J) = \sum_{\alpha=1}^{N} L(I_\alpha, J)/N \tag{7}$$

where I_α is a random portioning of S into S_v and S_c such that $|S_v| = n_v$ remains constant. The square root of average MSEP was denoted by qms,

$$\mathrm{qms}(J) = \sqrt{L(J)} \tag{8}$$

where the qms_{mc} abbreviation[19] was not used since LOO-CV was not considered.

Monte Carlo Variable Selection (MCVS). The standard approach to the variable selection could be stated as follows: [22] given a data set $\mathbf{Z}$ and a set of hypotheses H_k, choose a hypothesis J that "best" explains the data. In the MCCV context, the approach becomes: for the given sample $\mathbf{Z}$ (and fixed n_v and k), find a descriptor subset J which minimizes $L(J)$. If the number of all possible models h is large, i.e. $h = |H_k| = p!/(k!(p-k)!) \gg 1$, the exhaustive evaluation of all available $L(J)$ becomes computationally impossible, and heuristic optimizations are required to limit the number of considered hypotheses. However, even if all $L(J)$ were calculated and the smallest $L(J_{best})$ was identified, the question of how significant is the superiority of the found descriptor combination J_{best} would remain unanswered.

Historically, for MLR models, the statistical significance of the models is examined via the F-statistic and reported as t-values for individual regression coefficients, partial F-values, and/or overall F-values. While useful in other circumstances, this approach does not assist in achieving the main goal of this study, that is, to quantify how much better (if at all) the best model is, when compared to other models, where most (or a large number) of the considered models are statistically significant in the conventional sense (e.g., in terms of the F-statistic). Moreover, the F-statistic-based approach is essentially equivalent to finding the lowest $L(J_{best})$, since the F-statistic is just an inverse function of MSE.[19] In contrast, the following describes the new approach to variable selection, which focuses on the relative performance of models, some of which could be statistically significant in the conventional sense. This implies that the conventional statistical significance testing (e.g., by reporting F-values) is still required for the found J_{best}.

From the *statistical hypothesis* testing point of view, J denotes an *alternative* hypothesis which postulates that the descriptor subset J best explains the available data and hence should achieve the lowest *test statistic* $L(I,J)$ for a validation sample subset $S_v(I)$. The corresponding *null* hypothesis could then be defined as that there exists a better $J' \neq J$ in terms of achieving the lowest $L(I,J')$ for the given I. The conventional P-value essentially measures the probability of making the false-positive (also known as type-I) error, i.e., the error of rejecting a null hypothesis when it is true. The P-value corresponding to the test statistic $L(I,J)$ could then be calculated exactly (at least in theory) for the considered null hypothesis as $P(I,J) = 0$ if for all $J' \neq J$, $L(I,J) < L(I,J')$, meaning that the null hypothesis is in fact false and hence there is zero probability of false-positive error. If, however, there exists at least one $J' \neq J$ such that $L(I,J') < L(I,J)$, the null hypothesis is in fact true and hence $P(I,J) = 1$. In summary,

$$P(I,J) = \begin{cases} 0, & L(I,J) < L(I,J'),\ \forall\ J' \neq J \\ 1, & L(I,J' \neq J) < L(I,J) \end{cases} \tag{9}$$

$$P(J) = \sum_{\alpha=1}^{N} P(I_\alpha, J)/N \tag{10}$$

For example, if the descriptor subset J achieves the lowest $L(J)$ and has $P(J) < 0.05$, then by using the subset we accept less than 5% chance of the existence of a different descriptor subset (of size k) that has better predictive power for a randomly selected $S_v(I)$. In the majority of the QSAR applications, such a subset should be accepted as the best predictive subset. If, however, the same subset has $P(J) > 0.2$, the fact that it achieves the lowest $L(J)$ is statistically irrelevant, as the false-positive error will be committed in more than 20% of cases, and hence J could not be claimed as the best predictive descriptor subset with sufficient confidence. While the described interpretation of the $P(J)$ value is consistent with the chosen null hypothesis, $P(J)$ could also be interpreted as the *proportion* value to highlight the fact that the utilized null hypothesis is not random. Note that other non-parametric statistical methods such as the bootstrap and jackknife methods derive their P-values in a very similar fashion.

A parsimonious approach to the descriptor selection is supported by keeping the number of descriptors k constant within each MCVS run; i.e., one best descriptor, two best descriptors, etc. are searched for in sequence. The new methodology of selecting k is discussed below in the Blood−Brain Barrier section.

Simulated Annealing (SA-MCVS). This study is concerned with the descriptor selection for data sets with $p > n$, i.e., the number of available descriptors p is greater (or even much greater) than the number of available compounds n. The exhaustive evaluation of all plausible MSEP (i.e., $L(I,J)$), even for relatively small data sets ($n \approx 100$), may not be possible, regardless of available computational power, as the number of distinct combinations $(n!p!)/(n_c!n_v!k!(p-k)!)$ increases exponentially. Moreover, in order to calculate the P-values, it is not sufficient to search for a single optimal subset of descriptors J_{best}. Ideally, for each iteration of MCCV, all possible subsets J should be examined to calculate the P-values exactly. Since such exhaustive evaluation is often impossible for QSAR models, a limited number (m) of best-performing (i.e., with the lowest $L(J)$) subsets could be tracked, offering a computable approximation for the P-values. Obviously, the $P'(J)$ estimations obtained from such a fixed number m of the best-performing J hypotheses will always be less than or equal to the true value,

$$0 \leq P'(J) \leq P(J) \leq 1 \tag{11}$$

The important practical application of the above equation is that the higher the $P'(J)$ value, the more likely that the corresponding null hypothesis is true, i.e., descriptor subset J is not the best available. Or, in other words, only the subsets

which reject the null hypothesis (e.g., $P'(J) < 0.05$) should be verified by re-running the MCVS methods a number of times.

Traditionally, such complex combinatorial optimization/search problems have been studied by the simulated annealing[23−26] (SA) and genetic[27] (GA) algorithms. Both algorithms are heuristics (i.e., not exact), where the SA algorithm is based on the methods of statistical mechanics, while the GA algorithm mimics the genetic evolution of a population of species. The following SA algorithm was used.

Step 1: Start by randomly selecting $J = \{j_1, j_2, ..., j_k\}$.

Step 2: Continue by randomly selecting $S_v(I_\alpha)$ and $S_c(I_\alpha)$ compound subsets. Create new J' by swapping randomly selected $j_\alpha \notin J$ and $j_\beta \notin J$. Calculate the new cost $L_{new} = L(I_\alpha,J')$ and the current cost $L_{curr} = L(I_\alpha,J)$ associated with the current selection of the descriptors J.

Step 3: If $L_{new} \leq L_{curr}$, the new configuration J' is accepted, becoming "current". If $L_{new} > L_{curr}$, calculate the resulting change in the cost value, $\Delta L = (L_{new} - L_{curr})/L_{new} > 0$, where $L_{new} > 0$ is guaranteed by $0 \leq L_{curr} < L_{new}$. For $\Delta L > 0$, the new configuration is accepted with the probability $\Pr(\Delta L) = \exp(-\Delta L/(k_B T_\alpha))$, where T_α is the annealing temperature and k_B is originally the Boltzmann's constant, which becomes just a normalization constant and where the original Boltzmann distribution is used as per Kirkpatrick et al.[23] The role of the annealing temperature is to allow the acceptance of suboptimal configurations at the beginning of the annealing process. This is a key feature of the SA algorithm, allowing the algorithm to escape from local minima of the loss function.[1]

Step 4. Maintain a list[28,29] of up to m (typically $10 \leq m \leq 100$) best-performing descriptor subsets $B_m(\alpha) = \{J_1, J_2, ..., J_m\}$ with the lowest $L(J)$,

$$L_\alpha(J) = \frac{1}{n_J(\alpha)} \sum_{\beta=1}^{n_J(\alpha)} L(I_\beta,J) \qquad (12)$$

where $L_\alpha(J_1) \leq L_\alpha(J_2) \leq ... \leq L_\alpha(J_m)$ and where $n_J(\alpha)$ is the number of times the J descriptor subset has been selected so far (up to the current iteration α), i.e., $n_J \leq \alpha \leq N$. Note that the $O(\log m)$ efficient storage and retrieval of the J configurations could be achieved by Java's TreeMap class (or its equivalent in other computational languages/libraries) since the bit set representation of J maps uniquely to an integer value, where $O(...)$ is the "Big O" algorithm complexity measure describing an asymptotic upper bound and could be interpreted as "order of". The $P'(J)$ values are calculated via

$$P_\alpha(J) = (n_J(\alpha) - b_J(\alpha))/n_J(\alpha) \qquad (13)$$

where $b_J(\alpha)$ is the number of times the J descriptor subset achieved the lowest $L(I,J)$ so far. For each I, it takes $O(m \log m)$ to find J_γ with the lowest $L(I,J_\gamma)$ and then to increment $b_J(\alpha)$. For computational efficiency, the $B_m(\alpha)$ list is allowed to grow to $2m$ elements before it is trimmed back to m elements, since the sorting for the trimming takes $O(m \log m)$ while storing and retrieving J takes only $O(\log m)$.

Step 5: Repeat steps 2−4 with $T_\alpha = (N - \alpha + 1)/N$, where α is the iteration count, obtaining $T_1 = 1$, $T_N = 1/N$, and $0 < T_\alpha \leq 1$. Since $0 < \Delta L \leq 1$, Boltzmann's constant $k_B = -(1/\ln 0.5) = 1.4427$ is selected to achieve $\Pr(\Delta L = 1) = \exp(-(1/k_B)) = 0.5$; i.e., there is at least 50% chance in accepting the new configuration with larger cost value at the beginning of the annealing process ($T_{\alpha \ll N} \approx 1$). The optimality study of the selected sequence of temperatures (annealing schedule) was deemed to be outside the scope of this work. The described algorithm relies on the following conjecture (denoted C1):

$$\text{C1:} \qquad \lim_{N\to\infty} P_N(J) = P(J) \qquad (14)$$

Even though the conjecture is plausible, the exact mathematical proof for an arbitrary QSAR model (including MLR) may not be possible. If found (at least for MLR), such a proof may provide a quantifiable approach to the optimization of the annealing schedule. Note that, since exact evaluation of $P(J)$ is *NP-complete*,[30] the $P_N(J) \rightarrow P(J)$ convergence rate must be exponentially slow; i.e., the opposite would solve the *NP-complete* problem.

The overall algorithm complexity is $O(Nm \log m)$. A reasonable starting minimum for the number is $N = 100n$, i.e., each sample compound will be considered 100 times (on average) for calibration or validation. The user's access to the computing power controls the quality of the solution; i.e., the higher the number N, the higher is the probability of finding the optimal solution. Again, the stability of the solution should always be checked by increasing N until the results converge to the required level of accuracy; e.g., the $P'(J)$ results were stable within 0.01 accuracy in this study. As mentioned earlier, such rigorous validation of the stability of the $P'(J)$ results is only necessary for the low $P'(J)$ values, since strictly speaking the SA algorithm does not guarantee to find the global minimum of $L(J)$.

On a practical note, the algorithm is implemented in the freely available QSAR-BENCH program in such a way that intermediate results are displayed so that the MCVS algorithm could be stopped at any time, if the $P'(J)$ values stop changing within the desired accuracy. [QSAR-BENCH v2 is freely available (including its Java source code) from www.dmitrykonovalov.org for academic use.]

In the case of $k = 1$, the first m descriptors are automatically loaded in the list of best-performing descriptors. If the descriptor columns are sorted by descending order of the absolute values of their correlation to the Y column, then the algorithm starts with the most likely selection of best descriptors. The algorithm is also ideally suited for grid computing and parallel computing.

Genetic Algorithm (GA-MCVS). The main focus of this study was calculating the $P'(J)$ estimates. While the SA algorithm is commonly used, there is no proof that an SA run is guaranteed to find a global minimum of $L(J)$. This was overcome in this study by running SA a number of times with increasing N and verifying that the obtained $P'(J)$ estimates converged within the required accuracy. That is, the estimates were stable to the variation in N and consistently reproducible by different SA runs. The final validation of the obtained $P'(J)$ estimates was obtained by running a genetic algorithm (GA), which is as commonly used as SA

but, at the same time, is a completely different (from SA) optimization algorithm.

The following GA implementation was used:[21]

Step 1: Start by randomly selecting a population of m different descriptor subsets, $B_m(\alpha) = \{J_1, J_2, ..., J_m\}$, where descriptors could be interpreted as alleles (i.e., different variations of a gene) at a locus (i.e., position in a chromosome) and individual J could be viewed as a genotype.[28,29] A subset with a single descriptor ($k = 1$) corresponds to haploid species, a two-descriptor subset ($k = 2$) corresponds to diploid species, and $k > 2$ represents multiploid species.

Step 2: Select $S_v(I_\alpha)$ and $S_c(I_\alpha)$ compound subsets and calculate the average MSEP as per eq 12, which in this case assumes the role of the fitness function.

Step 3: Randomly select two parents from $B_m(\alpha)$ and use them to generate offspring J', where k alleles are randomly drawn from the parental alleles. Assuming the mutation rate to be ϵ ($\epsilon = 0.5$ was used), mutate J' in ϵ of the cases. When the mutation occurs, it replaces one randomly selected allele (i.e., descriptor) with a different allele (also randomly selected) that was not already present in the genotype. Calculate the corresponding $L_\alpha(J')$ and then proceed exactly as per SA. That is, (step 4) maintain the list of between m and $2m$ best descriptor subsets as measured by $L_\alpha(J)$. So, the only difference between the SA and GA is in how the next J is selected. Hence, the algorithm's complexity is exactly the same as for the SA, $O(Nm \log m)$. While SA gradually converges to the best-visited subset, GA continuously samples the genotype space "around" the best m visited subsets.

Step 5: Repeat steps 2 and 3 and report $P'(J)$ as per eq 13.

Method Summary. The MSEP is cross-validated according to the MCCV method. If MSEP is tracked for each MCCV iteration (I), the distribution of MSEP for each descriptor subset (J) could theoretically be obtained (this is the $L(I,J)$ loss function). Ideally, all available $L(I,J)$ should be compared to obtain the true P-value of each descriptor subset, i.e., $P(J)$. Since for any realistic QSAR data set such exhaustive comparison is not possible, the MCVS method uses two heuristic search algorithms (the simulated annealing and genetic algorithms) to obtain a computable approximation of $P(J)$ (denoted as the $P'(J)$ value or proportion estimates). The MCVS method relies on the $\lim_{N\to\infty} P_N(J) = P(J)$ conjecture for the considered heuristics, where N is the number of MCCV iterations.

E-DRAGON Descriptors. The complete set of the so-called DRAGON[31] descriptors is now freely available via the E-DRAGON website (www.vcclab.org/lab/edragon).[32−35] The E-DRAGON website can analyze up to 149 molecules at a time for up to 150 atoms per molecule. When 3D atom coordinates are not available, E-DRAGON accepts a column (up to 149 in length) of SMILES[36,37]-encoded molecules and performs 3D molecular structure optimization via the CORINA[38,39] or OMEGA[40,41] programs. For this study, the E-DRAGON website[34] generated 1666 descriptors for each compound's SMILES. When a descriptor value cannot be calculated, "-999" code is reported to indicate an error. These cases were replaced by zeros to allow for automated processing in this study.

RESULTS AND DISCUSSION

In this study, MLR was used to illustrate the proposed methodology. The MCVS method was employed with $100n \le N \le 100\,000$ and $10 \le m \le 100$. The number of compounds in the validation subsets was set to $n_v = n_c = n/2$ as per discussion in the paper by Konovalov et al.[19] Another argument in favor of $n_v = n_c = n/2$ is that such a choice of n_v maximizes the total number of available unique partitions of the data set into validation and calibration subsets, $n!/(n_c!n_v!)$. Unless stated otherwise, the descriptor columns in the considered data sets were sorted in descending order of the absolute values of their correlation $|r_j|$ to the response values (Y vector), where

$$r_j = \mathrm{cov}(Y,D_j)/\sqrt{\mathrm{var}(Y)\,\mathrm{var}(D_j)} \qquad (15)$$

The descriptor−descriptor correlation is defined by

$$r_{jj'} = \mathrm{cov}(D_j,D_{j'})/\sqrt{\mathrm{var}(D_j)\,\mathrm{var}(D_{j'})} \qquad (16)$$

Blood−Brain Barrier (KS289-logBB). Blood−brain (BB) distribution of a molecule is a key characteristic for assessing the suitability of the molecule as a drug for the central nervous system.[19] The distribution is commonly reported as $\mathrm{logBB} = \log(C_{brain}/C_{blood})$, where C_{brain} and C_{blood} are the equilibrium concentrations of the drug in the brain and the blood, respectively.[42] The KC290-logBB data set was recently reported[19] and was used in this study as the first sample data set, where the first letters of the first two authors' names and the number of utilized compounds were used for labeling the data sets throughout this study. The KC290-logBB data set was further edited by removing a duplicate entry for fluoromisonidazole (i.e., identical compounds 174 and 203),[43] where the compound numbering is from the original Table S1 of Abraham et al.[14] The resulted data set was denoted by KC289. In order to allow for automated generation of molecular descriptors, SMILES[36] encoding of the compounds was used from the KC289 data set. In addition to the SMILES and logBB values, the Iv[14] indicator values from the KC289 were also retained, since they were required to distinguish the origin of the logBB values and to remove the systematic difference of about 0.5 log units between the *in vivo* and *in vitro* distributions;[14] that is, the *in vitro* subset had Iv = 1, and the *in vivo* subset had Iv = 0. When the Iv indicator is used for prediction, it should be set to 1 for the prediction of the *in vitro* logBB values and to 0 for the *in vivo* logBB values. The compound SMILES, names, logBB, and Iv values can be found in Supporting Information Table S1.

Using the SMILES descriptions from the KC289 data set, the E-DRAGON descriptors[34] were calculated. The logBB values, Iv indicators, and E-DRAGON descriptors were combined into the data set denoted as KS289-logBB (Table S1). Table S1 was further processed using version 2 of the QSAR-BENCH program:[19] (1) E-DRAGON error code was replaced with zero; (2) constant and duplicate columns were deleted from the KS289-logBB data set, leaving 1500 (out of the original 1666) E-DRAGON descriptors. After the compounds (in rows) were sorted by their logBB values and the descriptors (in columns) were sorted by their correlation to the logBB column, the resulting KS289-logBB data set

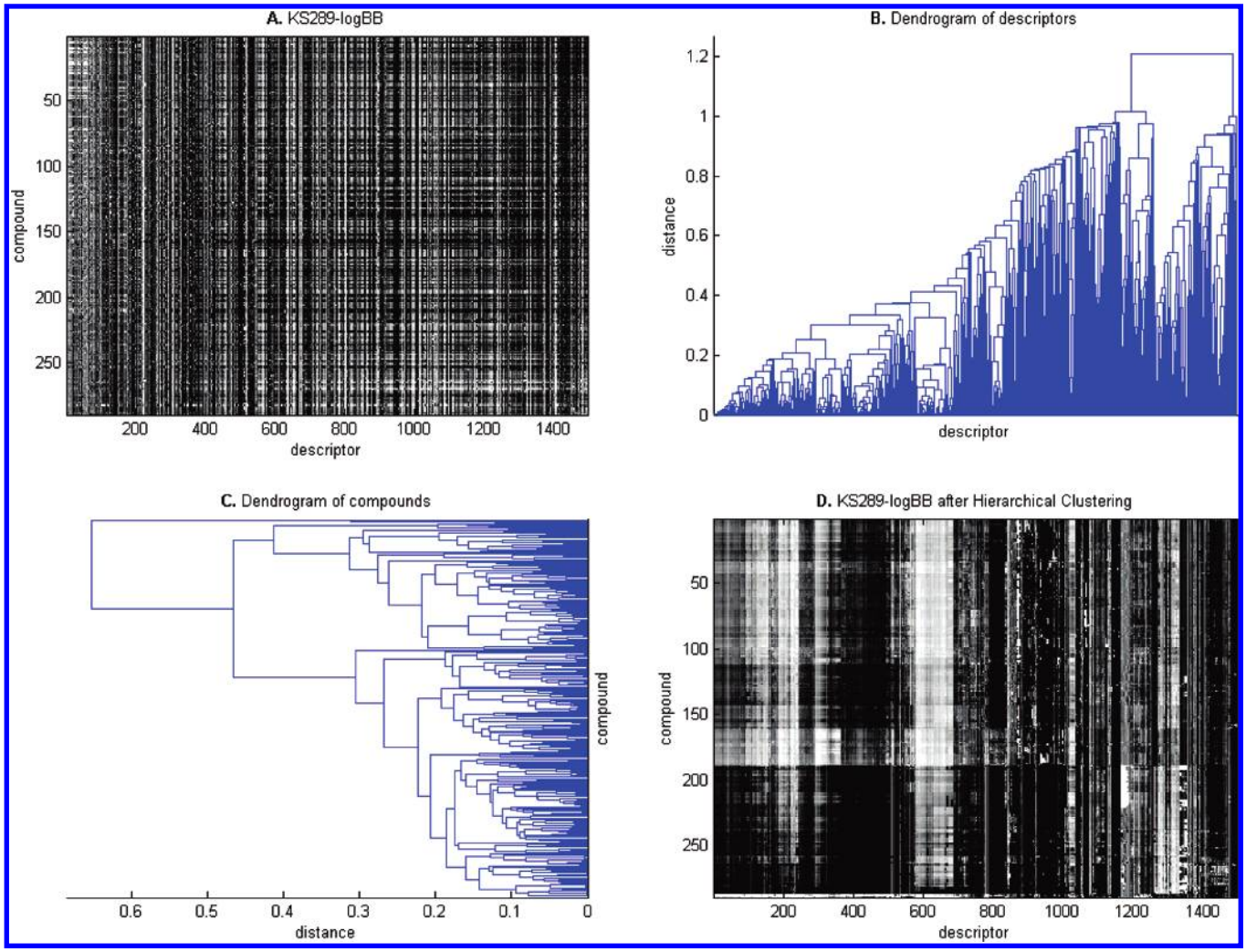

Figure 1. Clustered display of the KS289-logBB data set.

was visualized, as shown in Figure 1A as a gray image. The gray image normalization was achieved by normalizing the logBB and descriptor columns to the [0,1] range, where 0, $0 < x < 1$, and 1 values were represented by black, variable gray, and white pixels, respectively.

When the KS289-logBB data set is viewed as an image (Figure 1A), the data set resembles a standard DNA microarray gene expression experiment,[44] where the descriptors are playing the role of genes. Hierarchical clustering analysis was performed based on the average-linkage method[44,45] as follows (also see the Matlab source code file in the Supporting Information): (1) pairwise distances $d_{ii'}$ were calculated between compounds (in rows), where $d_{ii'} = 1 - \text{corr}(X_i, X_{i'})$ was calculated via Matlab's "pdist" function; (2) a hierarchical cluster tree (Figure 1C) was created from the pairwise distances $d_{ii'}$ using the unweighted pair group method with arithmetic mean[46] (UPGMA) algorithm (the "linkage" function); (3) steps 1 and 2 were repeated for descriptors (in columns), obtaining Figure 1B; (4) Figure 1D displayed the KS289-logBB data set after it was UPGMA sorted by rows and columns. Figure 1D demonstrates that the E-DRAGON descriptors were highly clustered in the KS289-logBB data set, therefore supporting the need for the *P*-value analysis to verify that the descriptor subset with the best qms is in fact the best subset for most cross-validations.

The following variable selection methodology is proposed: from the parsimonious considerations, $k = 1$ should be considered first, and then k should be consecutively increased until a domain-specific "acceptable" *P*-value (normally $P \leq 0.05$) associated with the current k is obtained. Therefore, the proposed methodology asks the question: "What is the smallest number of descriptors that achieve the lowest qms with $P \leq 0.05$?" The question implies that there may not be an answer for some or even all k. This is very different from the commonly used methodology, which searches for the "best" one, two, three, etc. descriptors (e.g., identified by the smallest MSE or MSEP), assuming that there is always an answer.

The following are the first few descriptors most correlated/anti-correlated to logBB values: $r_{\text{TPSA(NO)}} = -0.672$ for the topological polar surface area using N, O polar contributions;[47] $r_{\text{TPSA(Tot)}} = -0.644$ for the topological polar surface area using N, O, S, P polar contributions;[47] $r_{\text{nHDon}} = -0.549$ for the number of H-bond donors;[31] $r_{\text{Hy}} = -0.514$ for the hydrophilic factor;[31,48] $r_{\text{nO}} = -0.499$ for the number of oxygen atoms; $r_{\text{nHAcc}} = -0.488$ for the number of H-bond acceptors.[31]

Note that the Iv indicator should be considered a part of the logBB values rather than as a separate descriptor. For the Iv indicator to be always selected, the original MCVS

Table 1. Predictive Performance of QSAR Models

model	dataset-method	molecular descriptors	qms	F^a	$P'(J)$ value
1	KS289-logBB	[Iv],[b] **TPSA(NO)**	**0.434**	**140**	**0.03**
2	KS289-logBB	[Iv], TPSA(Tot)	0.454	115	0.97
3	KS289-logBB	[Iv], nHDon	0.507	62	1.0
4	KS289-logBB	[Iv], [TPSA(NO)], BEHv5	0.401	126	0.83
5	KS289-logBB	[Iv], [TPSA(NO)], BEHp5	0.401	126	0.89
6	KS289-logBB($\|r_{jj'}\| > 0.9$)[c]	[Iv], [TPSA(NO)], BELe4	0.407	119	0.88
7	KS289-logBB($\|r_{jj'}\| > 0.9$)	[Iv], [TPSA(NO)], RTe+	0.404	123	0.63
8	KS289-logBB	[Iv], [TPSA(NO)], [Ic], SRW09, BELv4, HATS7v	0.35	106	0.66
9	KS289-logBB	[Iv], [TPSA(NO)], [Ic], SRW09, BELv4, HATS8e	0.35	105	0.80
10	KS127-logHIA	ALOGP	0.385	196	0.45
11	KS127-logHIA	Hy	0.401	175	0.71
12	KS127-logHIA($\|r_{jj'}\| > 0.9$)	ALOGP	0.385	196	0.38
13	KS127-logHIA($\|r_{jj'}\| > 0.9$)	Hy	0.401	175	0.69
14	KS127-logHIA($\|r_{jj'}\| > 0.8$)	ALOGP	0.385	196	0.15
15	KS127-logHIA($\|r_{jj'}\| > 0.8$)	TPSA(NO)	0.443	128	0.89
16	KS127-logHIA($\|r_{jj'}\| > 0.7$)	ALOGP	0.385	196	0.05
17	KS127-logHIA	ALOGP, [-Hy][d]	0.385	196	0.33
18	KS127-logHIA	Hy, [-ALOGP]	0.401	175	0.57
19	KS127-logHIA($\|r_{jj'}\| > 0.8$)	[ALOGP], LAI, Neoplastic-80, RDF045m, R5v+, DDI, N-074, IDE	0.3	66	0.85

[a] F statistic was calculated on the whole KS289 data set. [b] Brackets denote preselected or fixed descriptors. [c] The KS289-logBB data set was trimmed by retaining only those descriptors which pairwise correlated with $|r_{jj'}| \leq 0.9$. [d] Denotes deleted descriptor.

algorithm (both the SA and GA versions) was modified to allow for the first few descriptors (k_{fixed}) to be always preselected, i.e., fixed. Then, for example, by placing the Iv indicator values in the first descriptor column and setting $k_{\text{fixed}} = 1$, the Iv indicator is always selected when the MCVS algorithm is run in the QSAR-BENCH program.

For the effective $k_{\text{eff}} = 1$ (QSAR-BENCH settings $k_{\text{fixed}} = 1$ and $k = k_{\text{eff}} + k_{\text{fixed}} = 2$), where the Iv indicator is not counted as a separate descriptor, the SA-MCVS and GA-MCVS algorithms were run on KS289-logBB with $n_v = n/2 = 145$, selecting the TPSA(NO) descriptor as the most accurate with $\text{qms}_{\text{TPSA(NO)}} = 0.434$ and $P' = 0.03$, see Table 1. The TPSA(Tot) and nHDon descriptors were reported as the next most accurate, with comparable qms ($\text{qms}_{\text{TPSA(Tot)}} = 0.454$ and $\text{qms}_{\text{nHDon}} = 0.507$) but unacceptable $P'_{\text{TPSA(Tot)}} = 0.97$ and $P'_{\text{nHDon}} = 1$. That is, for TPSA(Tot), there was about 97% chance that there was a more accurate predictive descriptor, which was most likely to be the TPSA(NO) descriptor in this case. The selection of TPSA(NO) as the best single descriptor is hardly surprising, as the important role of polar surface area is well-known.[49,50] A lesser trivial result was that, even though qms values for the TPSA(NO) and TPSA(Tot) descriptors were quite comparable ($\text{qms}_{\text{TPSA(NO)}} = 0.434$ and $\text{qms}_{\text{TPSA(Tot)}} = 0.454$), $P'_{\text{TPSA(NO)}} = 0.03$ clearly demonstrated the superior prediction power of the TPSA(NO) descriptor over all other considered descriptors (including the TPSA(Tot) descriptor).

The case of a single descriptor is a good illustration of the main purpose of this study. That is, the fact that the TPSA(NO) descriptor was the best possible single descriptor trivially followed from its F-statistic being the largest, and hence there was no need for any searching algorithms. The non-trivial result was the fact that the descriptor was so much better than the other statistically significant descriptors, such as the TPSA(Tot) and nHDon descriptors. This study proposes to compare the statistically significant descriptors by calculating their P-values, which requires re-running the calibration−validation holdout experiments many thousand times.

While running the MCVS algorithms, it was interesting to observe that quite often a particular partition, I, of the compounds in the data set was described best by a newly selected descriptor subset, J. Such subset J would achieve lower qms compared to the currently selected m best descriptor subsets. However, once different holdouts were examined, the corresponding average qms deteriorated (i.e., increased) quickly, and the new J was discarded. The existence of such events highlights the need for extensive cross-validation. That is, a single holdout may be best described by a descriptor subset purely by chance.

The role of the Iv indicator[14] was confirmed by examining the following MLR expressions (Figure 2):

$$\text{logBB} = 0.539 - 0.012 \times \text{TPSA(NO)}, \quad r^2 = 0.451, \quad s = 0.445 \quad (17)$$

$$\text{logBB} = 0.767 - 0.015 \times \text{TPSA(NO)} - 0.343 \times \text{Iv}, \quad r^2 = 0.492, \quad s = 0.430 \quad (18)$$

where r^2 is the proportion of total variation in Y explained by regression and s is the standard error. As expected,[14] the contribution of TPSA(NO) was hardly affected by the inclusion of the Iv indicator.

For the effective $k_{\text{eff}} = 2$ (QSAR-BENCH settings $k = 3$ and $k_{\text{fixed}} = 1$), after extensive trial executions of the MCVS method, the TPSA(NO) descriptor was always selected as one of the descriptors for most of the best m descriptor subsets. Therefore, purely for efficiency reasons, the TPSA(NO) descriptor was fixed in addition to the Iv descriptor (QSAR-BENCH settings $k = 3$ and $k_{\text{fixed}} = 2$), obtaining a range of Burden eigenvalue (BELxx and BEHxx) descriptors[51−55] with $0.4 \leq \text{qms} \leq 0.407$ but $0.83 \leq P' \leq 0.92$. The large P estimates meant that none of the Burden descriptors could be selected as the single preferred addition to the TPSA(NO) descriptor. The results for the best two Burden descriptors are displayed in Table 1 (models 4 and 5) for illustration purposes.

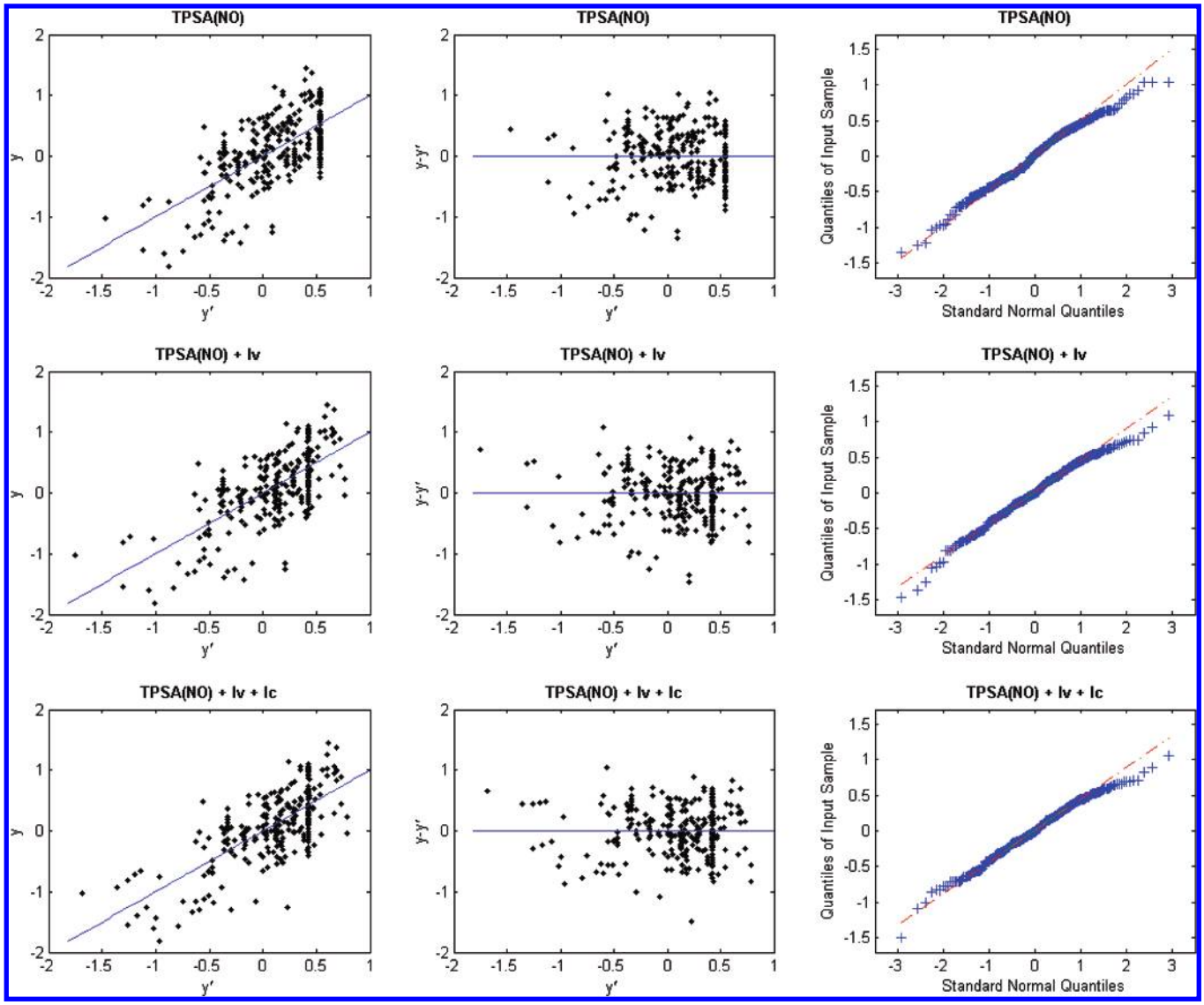

Figure 2. MLR models of the KS289-logBB data set: first column, experimental logBB (y) versus predicted (y'); second column, residual plot; third column, quantile−quantile plot of the residuals.

This situation, concerning the cluster of the Burden descriptors, was anticipated from Figure 1 with the following solution: if there was a combination of best descriptors, such a combination could be identified more easily by working with the remaining representatives of the descriptor "clusters" (Figure 1) after closely correlated descriptors were removed. The question of what should be considered the best representative of a descriptor cluster is likely to be application-specific and hence was not investigated further. In this study, we limited ourselves to one of the simplest rules for dealing with the clusters in a fashion very similar to the procedure of Merkwirth et al.:[56] the descriptors are sorted in descending order of their $|r_j|$ correlation to the response variable, and then, starting from the second descriptor, a descriptor is removed if it is correlated more strongly than some chosen threshold (r_{max}) to the remaining descriptors, $|r_{jj'}| > r_{max}$. The absolute value of correlation was used since, in some cases, the descriptors measure exactly the opposite property; e.g., see the next subsection. In the case of the Burden descriptors, trimming of the KS289-logBB data set by $|r_{jj'}| > 0.90$ was required to arrive at a single Burden descriptor (model 6 in Table 1) but again with an insufficient $P' = 0.88$. Moreover, the GETAWAY RTe+ descriptor[57,58] (R maximal index/weighted by atomic Sanderson electronegativities) was identified with the lowest RMSEP (qms = 0.404) but still insufficient $P' = 0.63$ (model 7 in Table 1).

For $k_{eff} > 1$ (Table 1), we were unable to find a subset of descriptors (in addition to the TPSA(NO) descriptor) with P-values low enough to be considered the *best* predictive descriptors with sufficient statistical confidence. The same situation persisted even after the KS289-logBB data set was trimmed by $|r_{jj'}| > 0.9$, $|r_{jj'}| > 0.8$, $|r_{jj'}| > 0.7$, $|r_{jj'}| > 0.6$, and even $|r_{jj'}| > 0.5$. The fact that the two very different SA and GA search algorithms produce high P-values confirms the absence of such optimal descriptor subsets. The GA results verified that the SA algorithm did not get trapped in local minima since GA included mutation in 50% of the new offspring, which randomly sampled all available combinations.

The absence of the descriptor subsets with low P-values is open for interpretation, with the following being our proposed conjecture (denoted C2):

C2: *A subset of statistically significant structure-based descriptors (predictor variables) describes a causal relation-*

ship between structure and activity (response variable) within the QSAR paradigm if and only if it achieves acceptably low P-value (as measured by MCVS/MCCV or other extensively cross-validated means).

For example, the TPSA(NO) descriptor clearly captures the nature of biochemical processes of the BBB permeation. The same cannot be said about the rest of the descriptors. This could be due to a number of factors: (1) one or more biochemically meaningful descriptors were present but they describe less dominant contributions, which were hidden by the experimental error in the data set; (2) a number of considered descriptors were correlated to some yet unknown descriptor (or descriptors) but the correlation was below the noise level in the data set.

Note that, in the absence of low *P*-value, the lowest qms could still be used for selecting the best predictive subset of statistically significant descriptors.[19] However, the proposed conjecture suggests that, without the low *P*-value, it may be pointless to discuss the "meaning" of the descriptors, as there is no statistical evidence that the subset with the lowest qms is any better than any other subset with comparable qms. For example, working with the KS289-logBB data set, in the majority of the cases the qms values of various descriptor subsets for $k > 1$ were different by less than 0.01, indicating their statistical equality, given the qms of about 0.4 log unit. Therefore, the lowest qms value is a necessary but not sufficient condition for identifying a causal[59] QSAR relationship (i.e., biochemically meaningful descriptors).

Keeping in mind the above discussion, for $k = 3$, the MCVS method often selected the nRCOOH descriptor, which was the number of carboxylic acids (aliphatic). The nRCOOH descriptor values were virtually identical to the Ic indicator[14,60] values from Platts et al.,[60] with the only differences Ic = 1 and nRCOOH = 0 for *p*-phenylbenzoic acid and salicylic acid (2-hydroxybenzoic acid), which had an aromatic carboxylic acid group (nArCOOH = 1). The Ic indicator counts the number of carboxylic acid groups regardless of the nature of their parent carbon (aliphatic or aromatic); that is, in terms of the E-DRAGON descriptors, Ic = nRCOOH + nArCOOH. Using the MCCV method,[19] the nRCOOH and Ic descriptors were included, in turn, with the TPSA(NO) descriptors obtaining $qms_{nRCOOH} = 0.417$ and $qms_{Ic} = 0.403$, respectively. The lower qms value of the Ic descriptor indicates that the nature of the parent carbon is irrelevant;[60] see also the following MLR expression and Figure 2:

$$\log BB = 0.797 - 0.015 \times TPSA(NO) - 0.375 \times Iv - 0.962 \times Ic, \quad r^2 = 0.569, \quad s = 0.396, \quad F = 126 \quad (19)$$

As discussed in the paper by Konovalov et al.,[19] qms = 0.3 is likely to be the theoretical lower bound due to the errors in the data set. After extensive execution of the QSAR-BENCH v2 program within a wide range of *k*-values, the lowest qms achieved was about 0.35, which is consistent with the 0.3 lower bound. Two such examples with $k = 6$ and $k_{fixed} = 3$ are presented in Table 1; see E-DRAGON[34] help pages for the description of the selected descriptors. Note that just the two clearly defined TPSA(NO) and Ic descriptors achieved qms = 0.43, while any three additional descriptors could reduce the error by only 0.08 log unit. This result also highlights the vital importance of the *extensive* LGO-CV (MCCV in this case). The number of possible calibration−validation combinations for non-small n_v, $N_{max} = n!/(n_v!n_c!)$, easily outnumbers any degrees of freedom of a QSAR model, including the number of descriptors used in a MLR model, making overfitting[13] very unlikely.

The *k*NN method[19] was used to check for any clustering/nonlinear effects in the TPSA(NO)-based MLR model of logBB (the use of the same notation *k* for the number of neighbors and the size of descriptor subset is purely coincidental). As expected from Konovalov et al.,[19] the KS289-logBB data set exhibited negligible clustering/nonlinear effects with the Iv and TPSA(NO) descriptors: qms = 0.434 was virtually identical to qms(70NN) = 0.436 obtained with 70 nearest neighbors and the same cross-validation via MCCV with $n_v = n/2 = 145$. The search for an optimal number of nearest neighbors is not considered in this study, as any result of this nature is highly data-specific. A simple rule is used instead, where about half of the available validation subset is used as the nearest neighbors, $k \approx n_v/2$ or $k \approx n/4$. The reported accuracy limit (qms ≈ 0.4) of the non-LFER-based MLR models[19] was reproduced by the Iv, Ic, and TPSA(NO) descriptors, achieving qms = 0.404 and qms(70NN) = 0.406.

Simulated Benchmark (KS289-SB). The main issue with using a real data set such as KS289-logBB is that the correct answer is not known. In order to verify that the proposed approach to the variable selection and the corresponding algorithms do work correctly, simulated data sets were created with exactly known statistical properties. Such simulated benchmark KS289-SB data sets were created from the original KS289-logBB data set by randomly reshuffling all descriptor values in each of the descriptor columns but retaining the original Iv, TPSA(NO), and BEHv5 values. Therefore, each KS289-SB data set contained the original logBB, Iv, TPSA(NO), and BEHv5 values as well as 1498 random descriptors. The main advantage of such an approach is that benchmarking data sets remain statistically identical to the original in every respect except for the correlation of the descriptors to the response variable (i.e., logBB). As expected, both MCVS methods achieved $P' = 0$ for the TPSA(NO) descriptor and $k_{eff} = 1$ (QSAR-BENCH settings $k = 2$ and $k_{fixed} = 1$); $P' = 0.01$ for the TPSA(NO) + BEHv5 combination and $k_{eff} = 2$ ($k = 3$ and $k_{fixed} = 1$); and $P' > 0.4$ for $k_{eff} > 2$.

One interesting result of these benchmarking experiments was that sometimes a combination of three ($k_{eff} = 3$) descriptors, which always included TPSA(NO) and BEHv5, would achieve P' as low as 0.4. Of cause, the specific descriptor would change when a KS289-SB data set was freshly generated. Without knowing that the descriptor was purely random, it would be tempting to look for some meaning of that "correlation" since, after all, the found combination was the best 60% of the time. Therefore, in line with the conjecture, a causal descriptor (or descriptors) must reject the null hypothesis (i.e., that there exist better descriptors) with the level of significance commonly used in most conventional statistical hypothesis testing applications, e.g., $P < 0.05$.

The above benchmarking was repeated for a larger known subset, where logBB, Iv, TPSA(NO), BEHv5, and SRW09 were left nonrandomized in the original KS289-logBB. The

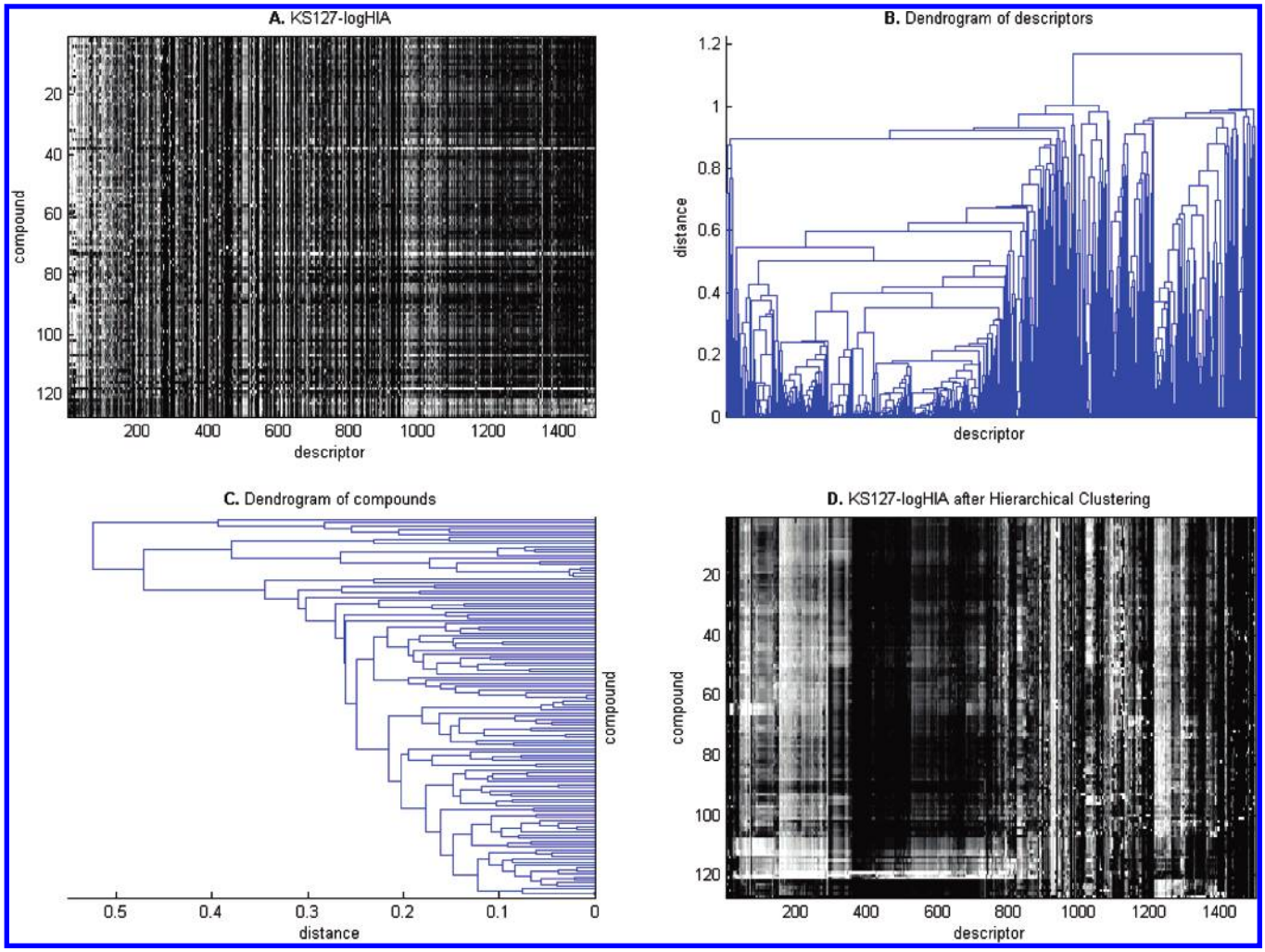

Figure 3. Clustered display of the KS127-logHIA data set.

P' results for $k_{eff} = 1$, $k_{eff} = 2$, and $k_{eff} > 4$ were unchanged, while $k_{eff} = 3$ yielded rather poor $0.08 < P' < 0.16$ for the TPSA(NO) + BEHv5 + SRW09 combination. This result was consistent with the standard MLR analysis, which revealed that the SRW09 values contained a very large random component; e.g., the corresponding F statistic was reduced to F(Iv + TPSA(NO) + BEHv5 + SRW09) = 105 from F(Iv + TPSA(NO) + BEHv5) = 126. In summary, the performed benchmarking confirmed that the MCVS method was not only able to select the statistically significant descriptor subsets but also capable of rejecting the subsets containing random descriptors.

The main underlying philosophy of this study is to report results that could be easily reproduced from the supplied data sets. This is accomplished by implementing the algorithms of this study in the freely available QSAR-BENCH v2 program, which could be used to reproduce the results and/or apply the proposed approach to different data sets. In particular, the results in this section could be verifying by loading the original KS289-logBB data set and randomizing any number of descriptor columns. However, we have found that, if the descriptors were sorted in their correlation order to the response variable and then, for example, the two best-correlated descriptors were left nonrandomized, the two descriptors would not be chosen when looking for the best two-descriptor subset. Even though this may be a trivial MLR result, it is stated to avoid misinterpretation of the described benchmarking procedure.

Note that the selection of any descriptor (BEHv5 in the above example) could be done in QSAR-BENCH by transposing the KS289-logBB data set and saving the result into a text file, which then could be easily edited by searching and moving the descriptor row to the front of the file using any conventional editors (including the Excel program). Once the desired structure of the file is obtained, the data could be loaded and transposed to restore the format of the **Z** matrix.

Human Intestinal Absorption (KS127-HIA). Following the procedure described by Abraham et al.,[61] 127 compounds were selected from those listed by Zhao et al.,[62] which had neither 0 nor 100% HIA values. The corresponding SMILES were obtained using the ChemSketch[63] program and the ChemDB[64,65] and CTB[66] websites. Following the procedure described for the KS289-logBB data set, the E-DRAGON descriptor values were calculated from the SMILES and combined with the %HIA values into the data set denoted by KS127-%HIA (Table S2). In this case, there were 1499 nonconstant and nonduplicate E-DRAGON descriptors, which were visualized as shown in Figure 3.

However, it is known that %HIA as an *activity* response variable has a nonlinear relationship with the *structure* descriptors within the QSAR framework.[61,67,68] Therefore, before the %HIA values could be used with the MLR

method, they must be transformed into a variable which could be modeled linearly using the E-DRAGON descriptors. Abraham et al.[61] used

$$\log \mathrm{HIA} = \log\{\ln[100/(100 - \%\mathrm{HIA})]\} \qquad (20)$$

which is denoted as the logHIA transformation.

The following are the first few descriptors most correlated/anti-correlated to the log HIA values: $r_{\mathrm{ALOGP}} = 0.782$ for Ghose−Crippen octanol−water partition coefficient;[69,70] $r_{\mathrm{Hy}} = -0.764$;[31,48] $r_{\mathrm{MLOGP}} = 0.762$ for Moriguchi octanol−water partition coefficient;[71,72] $r_{\mathrm{nHDon}} = -0.736$;[31] $r_{\mathrm{TPSA(NO)}} = -0.712$;[47] and $r_{\mathrm{TPSA(Tot)}} = -0.688$.[47]

For $k = 1$, the MCVS($n_v = n/2 = 63$) algorithm identified the ALOPG and Hy descriptors having the lowest $\mathrm{qms}_{\mathrm{ALOGP}} = 0.385$ and $\mathrm{qms}_{\mathrm{Hy}} = 0.401$, with the corresponding $P'_{\mathrm{ALOGP}} = 0.45$ and $P'_{\mathrm{Hy}} = 0.71$. Therefore, the logHIA problem appeared to be different from the logBB problem in that there was no single best descriptor, even though ALOGP achieved the lowest qms. Both ALOGP and Hy remained the best by the *P*-values (see Table 1) after the data set was trimmed by $|r_{jj'}| > 0.9$, where the rest of the descriptors had $P' > 0.98$. The corresponding MLR expressions were

$$\log \mathrm{HIA} = -0.119 + 0.2 \times \mathrm{ALOGP}, \quad r^2 = 0.611, \quad s = 0.373, \quad F = 196 \qquad (21)$$

$$\log \mathrm{HIA} = 0.333 - 0.174 \times \mathrm{Hy}, \quad r^2 = 0.584, \quad s = 0.386, \quad F = 175 \qquad (22)$$

The comparable importance of the ALOGP and Hy descriptors is due to their estimating almost exactly opposite properties, molecular lipophilicity and hydrophilicity, respectively. The Hy descriptor could be removed together with other highly correlated descriptors by decreasing the trimming threshold and obtaining P'_{ALOGP} ($|r_{jj'}| > 0.8$) = 0.15 and P'_{ALOGP} ($|r_{jj'}| > 0.7$) = 0.05. The next best *P* estimate was $P'(|r_{jj'}| > 0.7) = 0.99$, verifying that a measure of hydrophobicity/hydrophilicity was the most important single descriptor for the HIA problem (among the considered 1499 descriptors). The ALOGP descriptor estimates the logarithm of the 1-octanol/water partition coefficient (log *P*), which is traditionally used to measure lipophilicity (and hydrophobicity) of a molecule.[69] Interestingly, by removing just the Hy descriptor, the ALOGP's *P*-value improved from $P'_{ALOGP} = 0.45$ to $P'_{ALOGP} = 0.33$.

The above example shows that the proposed MCVS method works extremely well in selecting descriptors which are not highly correlated; e.g., see the P'_{ALOGP} ($|r_{jj'}| > 0.7$) = 0.05 example above. However, if the method is presented with a cluster of similar descriptors, the *P*-values may be unacceptably high for each of the descriptors in the cluster, even if the cluster does represent an existing structure−activity relationship. Then, suitable representatives of each of the clusters must be selected to assess the causal importance of the properties captured by the corresponding clusters. Such unsupervised selection of the cluster representatives could be an important future addition to the MCVS method.

The stability of the found *P*′-values could be verified, but only by increasing n_v/n from its current default of $n_v/n = 0.5$, e.g., obtaining even better P'_{ALOGP} ($|r_{jj'}| > 0.7$) = 0.02 by increasing n_v to $n_v = 80$ from the default value of $n_v = n/2 = 63$. Note that decreasing n_v/n will produce progressively less reliable results, since the probability of the LGO method to select the best *predictive* MLR model converges to 1 only when $n_v/n \to 1$ and $n \to \infty$.[16]

The next best descriptor, which describes affinity to water, was the Hy descriptor. The descriptor is a simple empirical information[48] index given by

$$\mathrm{Hy} = \frac{(1 + N_{\mathrm{Hy}}) \log_2 (1 + N_{\mathrm{Hy}}) + N_{\mathrm{C}}(^1/_A \log_2 {}^1/_A) + {}^1/_A \sqrt{N_{\mathrm{Hy}}}}{\log_2(1 + A)} \qquad (23)$$

where N_{Hy} is the number of hydrophilic groups (−OH, −SH, −NH), N_{C} is the number of carbon atoms, and *A* is the total number of non-hydrogen atoms.[31] There were a number of misconceptions and misquotations regarding the index. First, the original definition by Todeschini et al.[48] used the natural logarithm function, log ≡ ln, as evident from the table of the examples.[73] Subsequently, Todeschini and Consonni[31] changed to $\log_2$ while presenting the original table, without updating it to reflect the new definition. For example, the correct value for H_2O_2 is reported by E-DRAGON as Hy = 3.446,[73] while the example's Hy = 3.64 was obtained using ln. Second, the lower bound of the descriptor was correctly reported as −1,[31,48] for example, for highly hydrophobic molecules such as large ($A \equiv N_{\mathrm{C}} \gg 1$) hydrocarbon compounds. However, in contrast to what was previously stated,[31,48] the Hy descriptor does not have a maximum, as it is a monotonically increasing function of N_{Hy}.[73] When ALOGP was removed from the initial data set, the Hy descriptor did not reproduce the very low *P*-values of ALOGP until a much lower correlation threshold: P'_{Hy} ($|r_{jj'}| > 0.7$) = 0.09, P'_{Hy} ($|r_{jj'}| > 0.6$) = 0.09, and P'_{Hy} ($|r_{jj'}| > 0.5$) = 0.01.

Recalling that the *P* estimates are calculated with constant *k*, it is important to mention that, if this condition is relaxed, it may be difficult to search for small descriptor subsets ($k < 10$) in the presence of more than 1000 descriptors. That is, for any selection of the S_v and S_c subsets, and any small subset of descriptors $J = \{j_1, ..., j_k\}$, there is likely to exist a better subset $J' = \{j'_1, ..., j'_{k'}\}$, $k' > k$, given the large variety of descriptors considered (or even completely random predictors). A version of the SA-MCVS algorithm that allows for such "floating" *k* was implemented and tested to confirm the above observation. The feature was disabled in the current version of QSAR-BENCH to prevent its misuse or misinterpretation, but it could be made available on request.

As in the case of logBB, we could not identify any additional (to ALOGP) descriptors with sufficiently low *P*-values. Moreover, even running the MCVS method with $k = 8$ and fixed ALOGP, the method could not achieve qms better than about qms = 0.3 (Table 1). Comparison to qms = 0.385 obtained with just ALOGP indicated that the rest of the descriptors were essentially random variables in relation to what was left unexplained by the ALOGP descriptor. As in the case of the logBB problem,[19] further discovery of new or existing causal structure−activity relationships in the HIA problem is limited by the lack of better quality experimental data. For example, Wegner et al.[74] did not even attempt MLR of the HIA data set due to the estimated very high experimental error in the data set.

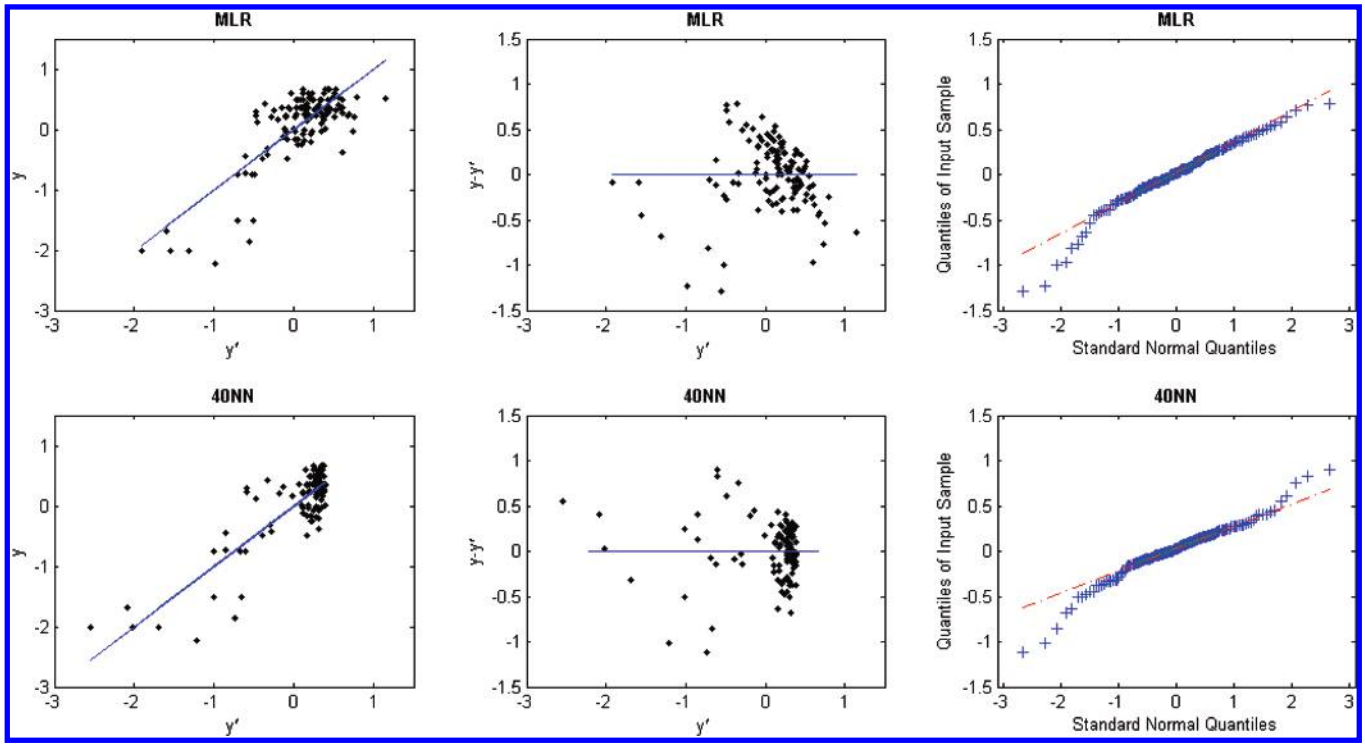

Figure 4. MLR and (k=30)NN models of the KS127-logHIA data set using the ALOGP descriptor: first column, experimental logHIA (y) versus predicted (y'); second column, residual plot; third column, quantile−quantile plot of the residuals.

The kNN method[19] was used to check for clustering/nonlinear effects in the KS127-logHIA-ALOGP data set. Without implying any optimization, 30 (as per the default rule $n/4$) nearest neighbors were utilized (30NN) with the ALOGP descriptor, achieving qms(30NN) = 0.34, which was an improvement from the original qms = 0.385. The qms-(30NN) results indicate that the clustering/nonlinear effects have a comparable or even larger contribution than the inclusion of any additional descriptors; i.e., qms = 0.3 was obtained with eight descriptors. For illustration purposes, the kNN method was also applied to the KS127-logHIA-ALOGP data set without cross-validation, obtaining $r^2 = 0.72$ and $s = 0.317$ (Figure 4), which showed a significant improvement from the MLR's $r^2 = 0.611$ and $s = 0.373$.

Surprisingly, the kNN method with just a single descriptor (ALOGP) achieved cross-validated accuracy (qms(30NN) = 0.34) only 17% worse than the non-cross-validated standard deviation ($s = 0.29$) reported by Abraham et al.,[61] who used five linear free-energy relationship (LFER) descriptors. Moreover, kNN achieved the non-cross-validated $s = 0.317$, which is only 9% worse than the LFER's standard error s. This suggests that the HIA values could be predicted more easily than the logBB values, for which the LFER semiempirical model achieved an accuracy of about $qms_{LFER} = 0.3$, while all non-LFER-based models were at least 33% less accurate ($qms_{non-LFER} \geq 0.4$).[19]

In summary, based on the KS127-logHIA data set and the near normality of the residuals in the last column of Figure 4, the ALOGP descriptor could predict the logHIA activity values via MLR with a predictive accuracy of about ±0.385 and ±0.77 log unit with 68% and 95% confidence, respectively. Using the kNN method, the same descriptor could achieve even higher predictive accuracy of about ±0.34 and ±0.68 log unit with 68% and 95% confidence, respectively.

CONCLUSIONS

A new Monte Carlo variable selection method was proposed and applied to the blood−brain barrier and human intestinal absorption problems using more than 1600 E-DRAGON[34] descriptors. In both considered data sets, there was only a single descriptor which could be interpreted as representing a causal[59] biochemical QSAR relationship: the TPSA(NO) and ALOGP descriptors for the BBB and HIA problems, respectively. To avoid potential misinterpretation of this result, we emphasize that, of cause, the considered problems are very complex and are likely to be explained by more than one descriptor. However, due to the absence of the relevant descriptors among the E-DRAGON descriptors and/or excessive noise in the data sets, we were unable to find more than one descriptor in each data set with sufficiently low P-values.

One of the underlying intentions of this study was to report variable selection results which could be independently reproduced. Therefore, we limited ourselves to the set of freely available E-DRAGON descriptors. The proposed methodology was illustrated on two examples and, it is hoped, could be easily applied to other QSAR problems with the assistance of the freely available QSAR-BENCH program (www.dmitrykonovalov.org). The considered BBB and HIA problems could also be re-examined as additional (or more accurate) data points or descriptors become available.

It was demonstrated that the kNN-MLR method significantly improved the MLR method for the HIA problem, indicating the strong presence of clustering/nonlinearity effects in the data set, which could be investigated in future studies.

By using large validation subsets ($n_v/n \geq 0.5$) and performing the cross-validation many thousand times, MCVS assesses the performance of a subset of variables via

statistically conventional *P*-values, where the null hypothesis is defined as "the rest of all possible descriptor subsets of the same cardinality". The new definition of the null hypothesis (rather than the reference to the random variables) is arguably more applicable to QSAR studies, since it directly compares various "best" descriptor subsets, many of which are statistically significant in the conventional MLR sense.

We appreciate that the conjecture C2 is rather strong. However, the conjecture could be viewed as our attempt to quantify the data-mining approach to the knowledge discovery in the QSAR models.

ACKNOWLEDGMENT

We thank Bruce Litow for useful discussions, as well as Anton Hopfinger and two anonymous reviewers for constructive comments on earlier versions of this manuscript. D.A.K. and D.C. were partly supported by a JCU Internal Research Grant.

Supporting Information Available: KS289-logBB and KS127-logHIA data sets, and Matlab code for Figure 1. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Dudek, A. Z.; Arodz, T.; Galvez, J. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. *Comb. Chem. High Throughput Screening* **2006**, *9*, 213−228.
(2) Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157−1182.
(3) 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.
(4) 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.
(5) Mente, S. R.; Lombardo, F. A recursive-partitioning model for blood-brain barrier permeation *J. Comput.-Aided Mol. Des.* **2005**, *19*, 465−481.
(6) 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.
(7) Rose, K.; Hall, L. H.; Kier, L. B. Modeling Blood-Brain Barrier Partitioning Using the Electrotopological State. *J. Chem. Inf. Model.* **2002**, *42*, 651−666.
(8) 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.
(9) 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.
(10) Pasha, F. A.; Srivastava, H. K.; Srivastava, A.; Singh, P. P. QSTR study of small organic molecules against Tetrahymena pyriformis. *Qsar Comb. Sci.* **2007**, *26*, 69−84.
(11) Deconinck, E.; Ates, H.; Callebaut, N.; Van, Gyseghem, E.; Vander Heyden, Y. Evaluation of chromatographic descriptors for the prediction of gastro-intestinal absorption of drugs. *J. Chromatogr. A* **2007**, *1138*, 190−202.
(12) Schölkopf, B.; Smola, A. J. *Learning with Kernels*; The MIT Press: Cambridge, MA, 2001.
(13) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Persp.* **2003**, *111*, 1361−1375.
(14) 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.
(15) 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.
(16) Shao, J. Linear Model Selection by Cross-Validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486−494.
(17) 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.
(18) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.
(19) Konovalov, D. A.; Coomans, D.; Deconinck, E.; Vander Heyden, Y. Benchmarking of QSAR models for Blood-Brain Barrier Permeation. *J. Chem. Inf. Model.* **2007**, *47*, 1648−1656.
(20) Konovalov, D. A.; Litow, B.; Bajema, N. Partition-distance via the assignment problem. *Bioinformatics* **2005**, *21*, 2463−2468.
(21) Hoffman, B. T.; Kopajtic, T.; Katz, J. L.; Newman, A. H. 2D QSAR Modeling and Preliminary Database Searching for Dopamine Transporter Inhibitors Using Genetic Algorithm Variable Selection of Molconn Z Descriptors. *J. Med. Chem.* **2000**, *43*, 4151−4159.
(22) Wegner, J. K.; Frohlich, H.; Zell, A. Feature selection for Descriptor based classification models. 1. Theory and GA-SEC algorithm. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 921−930.
(23) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671−680.
(24) Guha, R.; Jurs, P. C. Development of linear, ensemble, and nonlinear models for the prediction and interpretation of the biological activity of a set of PDGFR inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2179−2189.
(25) Itskowitz, P.; Tropsha, A. kappa Nearest neighbors QSAR modeling as a variational problem: Theory and applications. *J. Chem. Inf. Model.* **2005**, *45*, 777−785.
(26) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure-Activity-Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.
(27) Wegner, J. K.; Zell, A. Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077−1084.
(28) Konovalov, D. A.; Bajema, N.; Litow, B. Modified SIMPSON $O(n^3)$ algorithm for the full sibship reconstruction problem. *Bioinformatics* **2005**, *21*, 3912−3917.
(29) Konovalov, D. A. Accuracy of four heuristics for the full sibship reconstruction problem in the presence of genotype errors. *Adv. Bioinformatics Comput. Biol.* **2006**, *3*, 7−16.
(30) Davies, S.; Russell, S. NP-Completeness of Searches for Smallest Possible Feature Sets. In *Proceedings of the 1994 AAAI Fall Symposium on Relevance*; AAAI Press: New Orleans, 1994; pp 37−39.
(31) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: New York, 2000.
(32) 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.
(33) Tetko, I. V. Computing chemistry on the web. *Drug Discovery Today* **2005**, *10*, 1497−1500.
(34) E-DRAGON. *Dragon 5.4*; http://www.vcclab.org/lab/edragon/ (accessed June 1, 2007).
(35) VCCLAB. *Virtual Computational Chemistry Laboratory*; www.vcclab.org (accessed May 2, 2007).
(36) Weininger, D.; SMILES, a Chemical Language and Information-System 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(37) SMILES. *Simplified Molecular Input Line Entry System*; www.daylight.com/smiles (accessed May 2, 2007).
(38) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to 3-Dimensional Atomic Coordinates−Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(39) CORINA. *Generation of 3D coordinates*; www.molecular-networks.com/software/corina (accessed May 2, 2007).
(40) Bostrom, J. Reproducing the conformations of protein-bound ligands: A critical evaluation of several popular conformational searching tools. *J. Comput.-Aid. Mol. Des.* **2001**, *15*, 1137−1152.
(41) *OMEGA*. OpenEye Scientific Software, Inc.; www.eyesopen.com/products/applications/omega.html (accessed June 1, 2007).
(42) 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.
(43) Wittekindt, C., personal communication, 2007.
(44) Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 14863−14868.
(45) Sokal, R. R.; Michener, C. D. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* **1958**, *38*, 1409−1438.
(46) Fitch, W. M.; Margoliash, E. Construction of Phylogenetic Trees. *Science* **1967**, *155*, 279−284.

(47) 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.
(48) Todeschini, R.; Vighi, M.; Finizio, A.; Gramatica, P. 3D-modelling and prediction by WHIM descriptors. Part 8. Toxicity and physico-chemical properties of environmental priority chemicals by 2D-TI and 3D-WHIM descriptors. *SAR QSAR Environ. Res.* **1997**, *7*, 173−193.
(49) Liu, X. R.; Tu, M. H.; Kelly, R. S.; Chen, C. P.; Smith, B. J. Development of a computational approach to predict blood-brain barrier permeability. *Drug Metab. Dispos.* **2004**, *32*, 132−139.
(50) Kelder, J.; Grootenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemen, J. P. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* **1999**, *16*, 1514−1519.
(51) Burden, F. R. A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix. *Quant. Struct.-Act. Rel.* **1997**, *16*, 309−314.
(52) Burden, F. R. Molecular-Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.
(53) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discov.* **1998**, *9*−*11*, 339−353.
(54) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(55) Benigni, R.; Passerini, L.; Pino, A.; Giuliani, A. The information content of the eigenvalues from modified adjacency matrices: Large scale and small scale correlations. *Quant. Struct.-Act. Rel.* **1999**, *18*, 449−455.
(56) Merkwirth, C.; Mauser, H. A.; Schulz-Gasch, T.; Roche, O.; Stahl, M.; Lengauer, T. Ensemble methods for classification in cheminformatics. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1971−1978.
(57) Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693−705.
(58) Consonni, V.; Todeschini, R.; Pavan, M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682−692.
(59) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.
(60) 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.
(61) Abraham, M. H.; Zhao, Y. H.; Le, J.; Hersey, A.; Luscombe, C. N.; Reynolds, D. P.; Beck, G.; Sherborne, B.; Cooper, I. On the mechanism of human intestinal absorption. *Eur. J. Med. Chem.* **2002**, *37*, 595−605.
(62) 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.* **2001**, *90*, 749−784.
(63) ACD/ChemSketch; www.acdlabs.com (accessed May 2, 2007).
(64) Chen, J.; Swamidass, S. J.; Bruand, J.; Baldi, P. ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics* **2005**, *21*, 4133−4139.
(65) ChemDB. *ChemicalSearchWeb*; http://cdb.ics.uci.edu/CHEM/Web/cgibin/ChemicalSearchWeb.py (accessed June 26, 2007).
(66) CTD. *The Comparative Toxicogenomics Database*; http://ctd.mdibl.org/ (accessed June 28, 2007).
(67) Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J. Pharm. Sci.* **1999**, *88*, 807−814.
(68) Deconinck, E.; Coomans, D.; Vander Heyden, Y. Exploration of linear modelling techniques and their combination with multivariate adaptive regression splines to predict gastro-intestinal absorption of drugs. *J. Pharmaceut. Biomed.* **2007**, *43*, 119−130.
(69) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.
(70) 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. Model.* **1989**, *29*, 163−172.
(71) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. Simple Method of Calculating Octanol Water Partition-Coefficient. *Chem. Pharm. Bull.* **1992**, *40*, 127−130.
(72) 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*, 976−978.
(73) Mauri, A., personal communication, 2007.
(74) Wegner, J. K.; Frohlich, H.; Zell, A. Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA). *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 931−939.

CI700283S