Development and Validation of *k*-Nearest-Neighbor QSPR Models of Metabolic Stability of Drug Candidates

Min Shen,[†] Yunde Xiao,[†] Alexander Golbraikh,[†] Vijay K. Gombar,*[,‡] and Alexander Tropsha*[,†]

Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, CB# 7360, University of North Carolina, Chapel Hill, North Carolina 27599-7360, and Drug Metabolism and Pharmacokinetics, Mechanism and Extrapolation Technologies (MAI.A-3101), GlaxoSmithKline, 3030 Cornwallis Road, Research Triangle Park, North Carolina 27709

Received October 30, 2002

Computational ADME (absorption, distribution, metabolism, and excretion) models may be used early in the drug discovery process in order to flag drug candidates with potentially problematic ADME profiles. We report the development, validation, and application of quantitative structure−property relationship (QSPR) models of metabolic turnover rate for compounds in human S9 homogenate. Biological data were obtained from uniform bioassays of 631 diverse chemicals proprietary to GlaxoSmithKline (GSK). The models were built with topological molecular descriptors such as molecular connectivity indices or atom pairs using the *k*-nearest neighbor variable selection optimization method developed at the University of North Carolina (Zheng, W.; Tropsha, A. A novel variable selection QSAR approach based on the *k*-nearest neighbor principle. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 185−194.). For the purpose of validation, the whole data set was divided into training and test sets. The training set QSPR models were characterized by high internal accuracy with leave-one-out cross-validated R^2 (q^2) values ranging between 0.5 and 0.6. The test set compounds were correctly classified as stable or unstable in S9 assay with an accuracy above 85%. These models were additionally validated by in silico metabolic stability screening of 107 new chemicals under development in several drug discovery programs at GSK. One representative model generated with MolConnZ descriptors predicted 40 compounds to be metabolically stable (turnover rate less than 25%), and 33 of them were indeed found to be stable experimentally. This success (83% concordance) in correctly picking chemicals that are metabolically stable in the human S9 homogenate spells a rapid, computational screen for generating components of the ADME profile in a drug discovery process.

Introduction

It has been estimated that nearly 50% of drugs fail because of unacceptable efficacy, which includes poor bioavailability resulting from ineffective intestinal absorption and limited metabolic stability.[1] Clearly, in addition to pharmacological potency and toxicity, the absorption, distribution, metabolism, and excretion (ADME) properties are crucial determinants of the ultimate clinical success of a drug candidate.[1] To reduce the cost and improve the efficiency of experimental drug discovery, the pharmaceutical industry welcomed the "fail fast, fail cheap" [2] concept. One of the recent trends in the pharmaceutical industry has been the integration of what has traditionally been considered the "drug development" stage into the early phases of drug discovery. The aim of this paradigm shift is the prompt identification, and possibly elimination, of candidate molecules that are unlikely to survive later stages of drug development. To this end, in vitro ADME screens have been implemented in most pharmaceutical companies.[3]

As valuable as these experimental filters are, they do have some limitations. For example, they require physical samples of compounds for testing, and despite significant technical advances, they remain time-consuming and resource-intensive. Thus, there is currently much interest in the development and application of computational methods for predicting "druglikeness".[4] Such methods could be applied to virtual compounds or existing libraries, permitting rapid and cost-effective elimination of poor candidates prior to synthesis.

The main advantage of quantitative structure−property relationship (QSPR) methods lies in the fact that once such a relationship is ascertained, it becomes of valuable assistance in the prognosis of the property of interest of new molecules before they are actually synthesized and tested. An automated variable selection QSPR method based on the *k*-nearest-neighbor (*k*NN) classification principle was introduced recently in one of our laboratories.[5] This *k*NN QSPR method is formally built upon the active analogue principle, which implies that chemically similar compounds display similar profiles of their physical and biological properties. All compounds are characterized by multiple chemical descriptors, and chemical similarity is evaluated by Euclidean distance between points representing compounds in multidimensional descriptor space. The selection of the optimal subset of descriptors, which affords

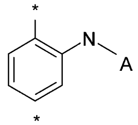
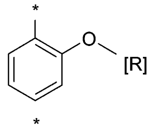
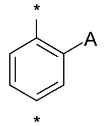
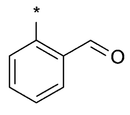
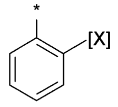
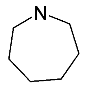
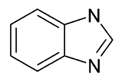
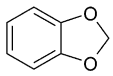
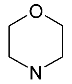
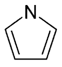
* To whom correspondence should be addressed. For V.K.G.: phone, 919-483-9504; fax, 919-315-6073; e-mail, vkg88567@gsk.com. For A.T.: phone, 919-966-2955; fax, 919-966-0204; e-mail, alex_tropsha@unc.edu.

[†] University of North Carolina.

[‡] GlaxoSmithKline.

10.1021/jm020491t CCC: $25.00 © 2003 American Chemical Society
Published on Web 05/30/2003

Table 1. Representative Structural Moieties in the Training Set[a]

Structural Moiety	Occurrence	Structural Moiety	Occurrence
Amino Acids	37	Alcohols	123
Bases, nucleosides	11	Carbamates	62
Carboxylic acid	35	Nitrile	39
Nitro	17	Sulfone	66
Urea	17	Ketone	82
	61		159
	20		44
	103		11
	27		17
	47		35

[a] $*$ = any atom. A = aromatic atom. [R] = saturated C atom. [X] = any halogen atom.

the highest correlation between the actual and predicted values of the target property for all compounds, is achieved by using simulated annealing as a stochastic optimization algorithm.

In this paper, we have applied the *k*NN QSPR approach to a data set of 631 compounds that were tested in the GSK Metabolic and Viral (MV) Diseases Center of Excellence for Drug Discovery (CEDD) for metabolic turnover in human liver S9 homogenate. Our objective was to establish robust and validated in silico QSPR models for the screening of new compounds. Frequently, the high value of q^2 for the training set is considered as a sufficient criterion of a QSPR model's accuracy. However, as we showed recently,[6] this value alone does not guarantee the acceptable predictive ability of a QSPR model. All QSPR models developed herein have been extensively validated using several criteria of robustness and accuracy.[6,7] These models were applied to an external set of 107 chemicals that were under in vitro assay at the GSK MV CEDD as these models were being developed. An average concordance of about 85% was observed between in vitro measurements and in silico prediction. The QSPR models developed and validated in this study can be used to evaluate metabolic turnover rates of large chemical databases or virtual libraries.

Data Set

A data set of 631 diverse compounds was used for the model generation. The complete structures of the compounds in the training set cannot be disclosed at this time because they are still in the discovery stage at GSK. However, their chemical diversity can be described to some extent by representative structural moieties that occur in the training set molecules. These fragments are shown in Table 1.

Methods

1. Metabolic Stability. In vitro metabolic stability of test compounds was assessed using pooled human liver S9 homogenate. Test compounds (final concentration of 1 μM) were incubated at 37 °C in the presence of the protein (5 mg/mL) and required cofactors. The incubation mixture was sampled at 0, 15, 30, and 60 min. Following precipitation with acetonitrile, samples were analyzed for test compound by LC/MS/MS methods. The percent turnover at 30 min was calculated by the following formula:

$$\text{turnover } \% = 1 - \frac{\text{peak area at 30 min}}{\text{peak area at 0 min}}$$

Thereby, each compound is assigned a value of turnover rate ranging from 0% to 100%. The experimental protocol is controlled by GSK.[8] Considering the practical significance of metabolic transformation rates in the context of drug design, we divided compounds into four subclasses: stable class (<25% turnover), moderately stable class (25−50%), moderately unstable class (50−75%), and unstable class (>75%).

2. Molecular Descriptors. The generation of molecular descriptors involves translation of chemical structures into numerical values. We applied two different types of descriptors in this study: MolConnZ descriptors[9] and atom pair (AP) descriptors.[10]

2.1. MolConnZ Descriptors. The MolConnZ program[8] is designed to carry out the computation of a wide range of topological indices of molecular structure. These indices include (but are not limited to) the following descriptors:[11−23] differential molecular connectivity indices, κ molecular shape indices, electrotopological state indices, graph's radius and diameter, Wiener and Platt indices, Shannon information indices, counts of different vertices, counts of paths and edges between different kinds of vertices, and many other topological indices.

For the present work, a version of the MolConnZ package[9] available at GSK servers was used. Initially, 310 different descriptors were generated for the available set of 631 compounds. However, many of these descriptors had zero variance and were eliminated, leaving 190 descriptors that were used eventually.

2.2. AP Descriptors. The AP descriptors were generated using an approach initiated by Carhart et al.[10] The key components for defining a set of atom pair descriptors include the definition of atom types and the classification of distance bins. An atom pair is a type of substructure defined in terms of the atom types and the shortest path (or graph distance) between two atoms. The graph distance is defined as the smallest number of atoms along the path connecting two atoms in a molecular structure. The general form of AP descriptors is as follows:

$$\text{atom type } i\text{—(distance)—atom type } j$$

Here, distance is the molecular graph distance (i.e., shortest path) between atom types *i* and *j*. In this study, the following 15 types of atoms were used: (1) negative charge center, NCC; (2) positive charge center, PCC; (3) hydrogen bond acceptor, HA; (4) hydrogen bond donor, HD; (5) aromatic ring center, ARC; (6) nitrogen atoms, N; (7) oxygen atoms, O; (8) sulfur atoms, S; (9) phosphorus atoms, P; (10) fluorine atoms, FL; (11) chlorine, bromine, iodine atoms, HAL; (12) carbon atoms, C; (13) all other elements, OE; (14) triple bond center, TBC; (15) double bond center, DBC. Further, distance bins were defined in the interval between graph distance 1 (i.e., zero atoms separating an atom pair) to 15 or greater. For this human S9 metabolic turnover data set of 631 compounds, 625 AP descriptors with nonzero value and nonzero variance were generated using the GenAP program developed in the Laboratory for Molecular Modeling, University of North Carolina.

3. Pearson Correlation Analysis. Many MolConnZ descriptors are highly correlated with each other. If two descriptors are strongly correlated, typically one of them is discarded.[24] To eliminate highly correlated descriptors, the Pearson correlation analysis was used. The Pearson correlation between two variables reflects the degree to which the variables are related. The correlation coefficient *r* between two variables *X* and *Y* for *N* compounds is calculated as follows:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)}\sqrt{\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}} \quad (1)$$

The value of *r* can take values between −1 and +1. A correlation of +1 means that there is a perfect positive linear relationship between variables, while −1 means a perfect negative linear relationship. Following pairwise Pearson correlation analysis, one of the highly correlated variables ($r > 0.9$) was removed and the number of MolConnZ descriptors was reduced from 190 to 106. The SAS[25] package was used for the analysis.

4. *k*NN QSPR Method. The *k*NN QSPR method[5] used here employs the *k*NN classification principle[26] combined with simulated annealing algorithm for variable selection. The activity of each compound in a dataset is predicted in a leave-one-out cross-validation process as a weighted average activity of its *k* nearest neighbors (i.e., most similar compounds) in the dataset. The procedure seeks to optimize simultaneously (i) the selection of variables (nvar) from the original pool of all molecular descriptors that are used to calculate similarities between compounds (i.e., distances in the nvar-dimensional descriptor space), (ii) number of nearest neighbors (*k*) used to estimate the activity of each compound, and (iii) value of q^2. Specifically, the *k*NN QSPR procedure involves the following steps.

(1) Select a subset of nvar descriptors randomly (nvar is a number between 1 and the total number of available descriptors) as a hypothetical descriptor pharmacophore[27] (HDP). nvar is usually set to different values in a range between 10 and 50 in several different runs, and for each fixed value of nvar, we generate at least 10 models.

(2) For each HDP, compute q^2 by a leave-one-out cross-validation procedure as described below.

(3) Repeat steps 1 and 2 until the maximum q^2 for a given number of nvar is achieved. This optimization process is driven by generalized simulated annealing (see below) using q^2 as the objective function.

The standard leave-one-out procedure has been implemented as follows.

(1) Eliminate a compound in the training set and predict its target property on the basis of the *k*NN principle, i.e., as the weighted average property of the *k* most similar molecules (*k* is set to 1 initially). The similarity between any two compounds *i* and *j* is evaluated as the Euclidean distance between their representations in the descriptor space,

$$d_{i,j} = \sqrt{\sum_{k=1}^{\text{nvar}} (X_{ik} - X_{jk})^2} \quad (2)$$

using only the subset of descriptors that corresponds to the current trial HDP.

The *X* descriptors generated with MolConnZ were range-scaled prior to distance calculations, and scaling was not necessary for AP descriptors. The reason for scaling the MolConnZ descriptors was that their absolute ranges differ quite significantly, sometimes by orders of magnitude, unlike AP descriptors, which are integers ranging between zero to no more than a couple of dozens of AP counts. Thus, the scaling was used to avoid giving descriptors with significantly higher ranges a greater weight upon distance calculations in multidimensional MolConnZ descriptor space.

The original *k*NN method was enhanced in this work by using the weighted molecular similarity as opposed to algebraic averaging as follows. In the original method,[5] the activity of each compound was predicted as the algebraic average activity of its *k* nearest compounds in the training set. However, since the Euclidean distances, in the selected descriptor space, between a compound and each of its *k* nearest neighbors may not be the same, the neighbor with the smaller distance from a compound was given a higher weight in calculating the predicted activity as shown in

$$w_i = \frac{\exp(-d_i)}{\sum_{k \text{ nearest neighbors}} \exp(-d_i)} \quad (3)$$

and

$$\hat{y} = \sum w_i y_i \quad (4)$$

Here, d_i is the Euclidean distance between a compound and its neighbor *i*, w_i is the weight for the nearest neighbor *i*, y_i is the actual experimental activity value for the nearest neighbor *i*, and $\hat{y}$ is the predicted activity value of the compound.

(2) Repeat step 1 until every compound in the training set has been excluded and its activity predicted once.

(3) Calculate the leave-one-out cross-validated R^2 (q^2) value using

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (5)$$

Here, y_i and $\hat{y}_i$ are the actual and predicted properties of the *i*th compound, respectively, and $\bar{y}$ is the average activity of all compounds in the training set. Both summations are over all compounds in the training set.

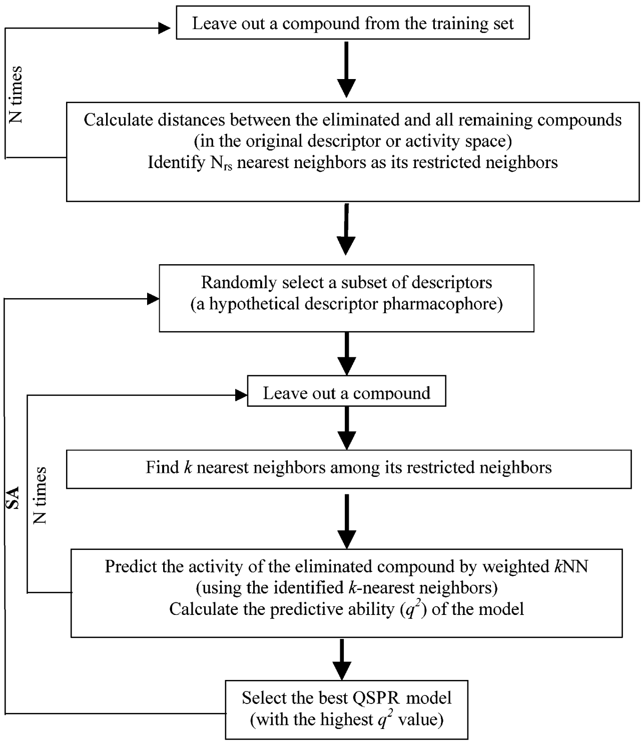

Leave out a compound from the training set
N times
Calculate distances between the eliminated and all remaining compounds
(in the original descriptor or activity space)
Identify N$_{rs}$ nearest neighbors as its restricted neighbors
Randomly select a subset of descriptors
(a hypothetical descriptor pharmacophore)
Leave out a compound
SA
N times
Find k nearest neighbors among its restricted neighbors
Predict the activity of the eliminated compound by weighted kNN
(using the identified k-nearest neighbors)
Calculate the predictive ability (q^2) of the model
Select the best QSPR model
(with the highest q^2 value)

Figure 1. Flowchart of the modified *k*NN method.

(4) Repeat steps 1−3 for $k = 2, 3, 4$, etc. Formally, the upper limit of k is the total number of compounds in the data set; however, the best value has been found empirically to lie between 1 and 5. The k value that leads to the highest q^2 value is chosen for the current *k*NN QSPR model.

Further details of the *k*NN method implementation including the description of the simulated annealing procedure used for stochastic sampling of the descriptor space are given elsewhere.[5] In summary, the *k*NN QSPR algorithm generates both an optimum k value and an optimal subset of nvar descriptors, which together afford a QSPR model with the highest value of q^2. Figure 1 shows the overall flowchart of the current implementation of the *k*NN method.

5. Applicability Domain of QSAR Models. Formally, a QSPR model can predict the target property for any compound for which its chemical descriptors can be calculated. However, since the training set models are developed by interpolating activities of nearest-neighbor compounds, a special similarity threshold should be introduced to avoid making predictions for compounds that differ substantially from the training set molecules. This threshold D_T is calculated as follows.

$$D_T = \bar{y} + Z\sigma \qquad (6)$$

Here, $\bar{y}$ is the average Euclidean distance of k nearest neighbors used to predict the target property of each compound in the training set, σ is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level. We set the default value of this parameter at 0.5, which formally places the allowed distance threshold at one-half of the standard deviation assuming Boltzman distribution of distances between k nearest neighbor compounds in the training set. Thus, if the distance of the external compound from at least one of its nearest neighbors in the training set exceeds this threshold, the prediction is deemed unreliable.

Table 2. Frequently Used α Values and the Corresponding Critical Values of Z_c for One-Tail Test[a]

α	Z_c
0.10	1.28
0.05	1.64
0.01	2.33
0.001	3.10

[a] For $z \geq 4$, $\alpha = \{1/[\sigma(2\pi)^{1/2}]\}\, e^{-z^2/2}$.

6. Robustness of QSPR Models. To evaluate the statistical significance of QSPR models, we have employed a standard hypothesis testing approach.[28] The robustness of training set models was examined by comparing these models to those derived for random data sets. The latter data sets were generated by assigning target property values to the training set compounds randomly but restricting these values to fall within the same range as the actual values of the original training set. In practice, this is achieved by random shuffling of compound properties prior to QSPR analysis.

According to the standard hypothesis testing approach, two alternative hypotheses are formulated:

$$H_0\colon\quad h = \mu$$

$$H_1\colon\quad h > \mu$$

where μ is the average value of q^2 for random data sets and h is the q^2 value for the actual data set. Thus, the null hypothesis H_0 states that the QSPR model for the actual data set is not significantly better than random models whereas the alternative hypothesis H_1 assumes the opposite, i.e., that the actual model is significantly better than random models. The decision-making is based on the standard one-tail test, which involves the following procedure.

(1) Determine the average value of q^2 (μ) and its standard deviation (σ) for random datasets.

(2) Calculate the Z score that corresponds to the q^2 value for the actual dataset:

$$Z = (h - \mu)/\sigma \qquad (7)$$

(3) Compare this Z score with the tabular critical values of Z_c at different levels of significance (σ)[28] to determine the level at which H_0 should be rejected. If the Z score is higher than the tabular values of Z_c (cf. Table 2), one concludes that at the level of significance that corresponds to that Z_c, H_0 should be rejected and, therefore, H_1 should be accepted. In this case, it is concluded that the result obtained for the actual dataset is statistically much more significant than the results obtained for random data sets at a given level of significance.

7. Model Validation: Training and Test Set Selection. Often, a value of q^2 greater than 0.5 is regarded as a proof of the high predictive ability of the model (for instance, these criteria are used as an ultimate indicator of a model's predictive power by the popular software CoMFA marketed by Tripos[29]). Although a low value of q^2 can indeed serve as an indicator of the low predictive ability of the model, the opposite is not necessarily true; i.e., as we have demonstrated recently,[6] high q^2 values do not imply that the model has any significant predictive power. The predictive power of a QSPR model can be evaluated by its ability to predict accurately the target property of compounds not used in model development. Thus, to establish the predictive power of a model, one needs to split the available data set into the training and test sets. A model is then developed for the training set compounds and is evaluated by its accuracy in predicting the target property of the test set compounds. The following algorithm was used in this work to divide a set of N compounds into the training and test sets.[30]

(1) The total volume V occupied by N points representing compounds in the descriptor space is estimated as in ref 6; the volume corresponding to one representative point is then equal to V/N.

(2) Select a compound with the highest activity.

(3) Include this compound in the training set.

(4) Construct a sphere with the center on the representative point of this compound and with radius $R = c(V/N)1/K$. Here, K denotes the number of descriptors (dimensionality of the descriptor space), and c is the dissimilarity level. (Dissimilarity level was varied to construct more examples with different number of compounds in the training and test sets.)

(5) Include compounds corresponding to representative points within this sphere (except the compound at the center) in the test set.

(6) Exclude all points within this sphere from the initial set of compounds.

(7) Let n be the number of remaining compounds. If $n = 0$, go to step 11; otherwise, go to step 8.

(8) Let m be the number of spheres already constructed. Calculate distances d_{ij}, with $i = 1, ..., n$ and $j = 1, ..., m$, from the representative points of the remaining compounds to the sphere surfaces.

(9) Select a compound with the smallest d_{ij}.

(10) Go to step 3.

(11) Stop.

This algorithm discussed in more detail in the original publication[30] allows construction of training sets that cover all descriptor space areas occupied by representative points. The higher is the dissimilarity level c, the smaller is the training set and the larger is the test set. It is expected that the predictive ability of a QSPR model generally decreases when the dissimilarity level increases. Obviously, the selection of training and test set compounds is sensitive to the types of descriptors used in calculations.

Results and Discussion

1. *k*NN QSPR Modeling Using MolConnZ Descriptors. 1.1. Model Robustness. As discussed above, the robustness of a QSPR model should be established by comparing results for the actual data set with those for the data sets with randomized activity values. Thus, 10-, 20-, 30-, 40-, 50-, and 60-descriptor (nvar) models were generated for the data sets of 631 compounds. Figure 2 shows the plots of q^2 vs nvar for the actual and random data sets. For each nvar, the result is the average of 10 independent models. Overall, the actual QSPR models give consistently higher q^2 values than those for the random data sets.

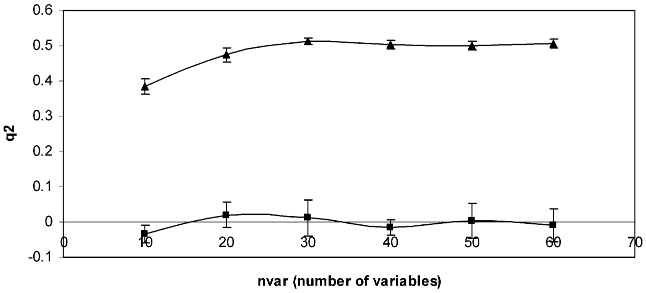


Figure 2. Plot of q^2 vs the number of variables selected for kNN QSPR models using MolConnZ descriptors. The results for both actual (triangles) and random (rectangles) data sets are shown. Every q^2 value is the average of 10 independent calculations.

Table 3. Standard One-Tail Hypothesis Testing for a 20-Descriptor kNN QSPR Model Using MolConnZ Descriptors

data sets	q^2	Z score
random 1	−0.031	0.047
random 2	0.029	0.959
random 3	−0.143	−1.655
random 4	0.066	1.521
random 5	−0.018	0.245
random 6	−0.048	−0.211
random 7	−0.074	−0.606
random 8	0.039	1.111
random 9	−0.054	−0.302
random 10	−0.107	−1.107
average of random 1−10	−0.034	
standard deviation	0.066	
actual	0.511	8.284

The statistical examination of the results was performed with one-tail hypothesis testing as described in the Methods section. The q^2 values for the 20-descriptor kNN QSPR models obtained from 10 different random data sets are shown in Table 3. This table also lists the average q^2 value, the standard deviation of the q^2 values, and the Z score for the 20-descriptor model for the actual data set. A Z score of 8.28 indicates that there is a probability of only about 10^{-15} that the model constructed for the actual data set is spurious.

In the kNN QSPR method, nvar can be set to any value that is less than the total number of descriptors. Figure 2 allows one to examine the relationship between the statistical robustness of the actual model (characterized by the difference between q^2 values for actual and random models) and nvar. Initially, as the value of nvar increases, the model performance also improves dramatically until it reaches a plateau at nvar of 30. These results suggest that the optimal values of nvar for models built with MolConnZ descriptors are between 30 and 50.

1.2. Model Generation and Evaluation Using Training and Test Sets. Having established that several robust kNN QSPR models for the entire original data set can be obtained, we have concentrated on the development of predictive (i.e., validated) models. Using the training and test set selection procedure discussed in Methods, the entire data set was divided into a training set of 572 compounds and a test set of the remaining 59 compounds. The q^2 value for a 30-descriptor model for the training set was 0.516. The accuracy of prediction for the training set is shown in Table 4a. The model categorized 256 compounds as stable, and 216 of them were actually stable as deter-

Table 4. Accuracy Analysis of *k*NN QSPR Model Using MolConnZ Descriptors

(a) Cross-Validated Training Set, 572 Compounds

		predicated class			
actual class		stable	mod. stable	mod. unstable	unstable
stable	281	216	55	10	0
mod. stable	97	30	45	18	4
mod. unstable	73	6	20	32	15
unstable	121	4	17	44	56
total	572	256	137	104	75

(b) Test Set, 59 Compounds

		predicated class			
actual class		stable	mod. stable	mod. unstable	unstable
stable	41	33	6	2	0
mod. stable	9	5	3	1	0
mod. unstable	4	2	0	1	1
unstable	5	0	0	2	3
total	59	40	9	6	4

(c) External Validation Set, 96 Compounds

		predicated class			
actual class		stable	mod. stable	mod. unstable	unstable
stable	66	42	19	5	0
mod. stable	9	2	6	1	0
mod. unstable	12	1	6	2	3
unstable	9	5	1	1	2
total	96	50	32	9	5

mined experimentally. Thus, the accuracy of this training set model was 84% when applying the screen for “acceptable” (slowly metabolizing) chemicals. Thirty additional compounds were predicted as stable but actually were found to belong to the moderately stable class, which is adjacent to the stable class. On the other side of the stability spectrum, the model assigned 75 compounds to the unstable class, and 56 of them were actually assigned to this class experimentally. Thus, the accuracy of this model was 75% when applied to select “unacceptable” (rapidly metabolizing) chemicals. Fifteen additional compounds were predicted as members of the unstable class but actually belonged to the adjacent moderately unstable class.

Table 4b shows the accuracy of prediction for the test set. The model assigned 40 compounds to the stable class, and 33 of them were indeed defined as stable experimentally. This implies that the accuracy of this model with respect to screening for “true positive” stable compounds was 83%. Five additional compounds were predicted as stable but actually belonged to the adjacent moderately stable class, which can be also considered as an acceptable prediction. This model also categorized four compounds as unstable, and three of them were actually assigned to this class experimentally. Only one compound was erroneously predicted as unstable but was experimentally assigned to the moderately unstable class, which is adjacent to the unstable class. It should be noted that the model produced no false positive or false negative predictions.

1.3. External Validation. In the course of developing QSPR models described above, the experimental data for 107 additional compounds had become available. These compounds constituted an external validation data set. Eleven of these compounds were found to be outside the model applicability domain based on the threshold criteria (cf. eq 6). The results of prediction for the remaining 96 compounds are shown in Table 4c. The model predicted 50 compounds as stable, and 42 among them actually belonged to the stable class, implying 84% accuracy of prediction. Concurrently, the same model predicted five compounds as unstable, and two of them were experimentally unstable. The remaining three compounds actually belonged to the moderately unstable class.

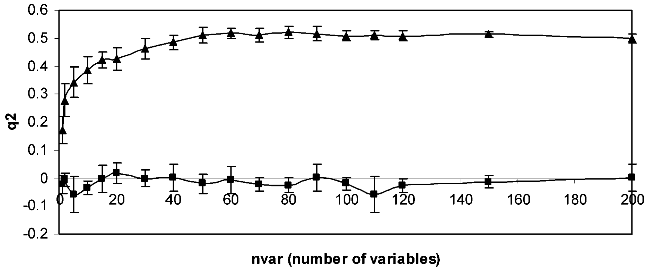


Figure 3. Plot of q^2 vs the number of variables selected for the *k*NN QSPR models using AP descriptors. The results for both actual (triangle) and random (rectangle) data sets are shown. Every q^2 value is the average of 10 independent calculations.

It is obvious from these studies that the disagreements between predictions from our model and the actual in vitro assay results are minimal for both stable and unstable categories. This result confirms that kNN QSPR models using MolConnZ descriptors are reliable for the prediction of metabolic stability and can aid in selecting “acceptable” and eliminating “unacceptable” compounds.

1.4. Importance of Pearson Correlation Analysis. We have also developed QSPR models using all 190 descriptors, but the detailed results are not listed here. We found that the best models generated after deletion of highly correlated descriptors using Pearson correlation analysis had practically the same quality as models built using all 190 descriptors; both had the average accuracy about 85%. Thus, in this case, the exclusion of highly correlated descriptors was reasonable. It did not influence the accuracy of QSPR models but made the variable selection procedure faster and more efficient. If QSPR models built with a fewer number of descriptors have high predictive ability, no additional analysis based on all descriptors is required. However, in some cases a small variance between two or more correlated descriptors may be very important for structure−property relationships,[24] and including all of them in QSPR studies may dramatically improve the predictive power of models. Thus, if the predictive power of QSPR models built using low-correlated descriptors is poor, it may be necessary to perform analysis with all descriptors.

2. *k*NN QSPR Modeling Using Atom Pair Descriptors. 2.1. Model Robustness. To examine the robustness of a QSPR model, 1-, 2-, 5-, 10-, 15-, 20-, 30-, 40-, 50-, 60-, 70-, 80-, 90-, 100-, 110-, 150-, and 200-descriptor models were established for both the actual data set and data sets with randomized activity values. Figure 3 shows the plots of q^2 vs nvar for the actual and random data sets. For each nvar, the result is the average of 10 independent models. Overall, the q^2 vs

Table 5. Standard One-Tail Hypothesis Testing for a 20-Descriptor *k*NN QSPR Model Using AP Descriptors

data sets	q^2	*Z* score
random 1	0.023	0.789
random 2	−0.035	−0.295
random 3	−0.046	−0.501
random 4	−0.102	−1.547
random 5	0.037	1.050
random 6	0.069	1.648
random 7	−0.044	−0.463
random 8	−0.027	−0.146
random 9	−0.079	−1.118
random 10	0.012	0.583
average of random 1−10	−0.019	
standard deviation	0.054	
actual	0.503	9.759

Table 6. Accuracy Analysis of *k*NN QSPR Model Using AP Descriptors

(a) Training Set, 560 Compounds

		predicated class			
actual class		stable	mod. stable	mod. unstable	unstable
stable	276	218	44	13	1
mod. stable	94	42	34	16	2
mod. unstable	70	13	24	27	6
unstable	120	8	22	38	52
total	560	281	124	94	61

(b) Test Set, 71 Compounds

		predicated class			
actual class		stable	mod. stable	mod. unstable	unstable
stable	46	40	5	1	0
mod. stable	12	7	2	3	0
mod. unstable	7	1	2	1	3
unstable	6	0	1	3	2
total	71	48	10	8	5

(c) External Validation Set, 78 Compounds

		predicated class			
actual class		stable	mod. stable	mod. unstable	unstable
stable	51	43	6	2	0
mod. stable	9	3	4	2	0
mod. unstable	10	5	2	2	1
unstable	8	5	2	1	0
total	78	56	14	7	1

nvar relationships indicate that the actual QSPR models give consistently higher q^2 values than those for random data sets.

Initially, as the value of nvar increased, the model performance also improved dramatically until it reached a plateau at nvar of 50. The results presented in Figure 3 suggest that the optimal nvar of AP descriptors is between 50 and 70.

The statistical examination of the results was performed with one-tail hypothesis testing as described in Methods. The q^2 values for the 20-descriptor *k*NN QSPR models obtained from 10 different random data sets are shown in Table 5. This table also lists the average q^2 value, the standard deviation of the q^2 values, and the *Z* score for the 20-descriptor model for the actual data set. A *Z* score of 9.76 indicates that there is a probability of only about 10^{-20} that the model constructed for the actual data set is spurious.

2.2. Model Generation and Evaluation Using Training and Test Sets. To validate the QSPR model, we selected 560 compounds into the training set and left 71 compounds in the test set based on the training and test set selection procedure discussed in Methods. The q^2 value for a 50-descriptor model for the training set was 0.511, and the accuracy of prediction for the training set is shown in Table 6a. The model categorized 281 compounds as stable, and 218 among them were actually stable as determined experimentally. This means that the accuracy of this model is 78% when applying the screen for “acceptable” chemicals. Forty-two additional compounds were predicted as stable but actually belonged to the moderate stable class, which is adjacent to the stable class. Concurrently, this model classified 61 compounds as unstable, and 52 among them actually belonged to the unstable class. That is, the accuracy of this model is 85% when applying the model to screen for “unacceptable” chemicals. Six compounds are predicted as unstable but actually belong to the adjacent moderately unstable class.

Table 6b shows the accuracy analysis for the test set. The model assigned 48 compounds to the stable class, and 40 of them were indeed defined as stable experimentally. This implies that the accuracy of this model with respect to screening for “true positive” stable compounds was 83%. Seven additional compounds were predicted as stable but actually belonged to the adjacent moderately stable class, which can be also considered as the acceptable prediction. This model also categorized five compounds as unstable, and two of them were actually assigned to this class experimentally. The remaining three compounds actually belonged to the moderately unstable class, which means that one will not be acceptable chemicals when applying the screen for “unacceptable” compound selection. It should be noted that the model produced no false positive or false negative predictions.

2.3. External Validation. When applying this model for predicting the external validation data set including 107 compounds, 29 of these compounds were regarded as too dissimilar from the training set based on the threshold criteria (cf. eq 6). The results of prediction for the remaining 78 compounds are shown in Table 6c. The model predicted 56 compounds as stable, and 43 among them actually belonged to the stable class. This means that the accuracy of this model is 77% as applied to screening for metabolically stable compounds. Concurrently, this model predicted one compound as unstable, but none actually belonged to the unstable class. This incorrectly predicted that the compound was experimentally assigned to the moderately unstable class, which is adjacent to the unstable class.

Similar to our studies with MolConnZ descriptors, the disagreement between predictions from the model and the results from the actual in vitro assay is minimal in the stable and unstable categories, confirming that the screen can be very effective in picking “acceptable” and eliminating “unacceptable” compounds.

Conclusions

We have developed several thoroughly validated QSPR models of metabolic stability measured in human S9 in vitro metabolic stability assay for a series of GSK compounds. Compounds were divided into four classes according to their stability. These models were shown to be very accurate in assigning compounds to extreme classes (i.e., stable and unstable). Comparable results were obtained using MolConnZ and AP descriptors. These models can be used as an in silico screen to predict the metabolic stability of diverse chemicals undergoing ADME testing. They can be applied to analyze existing chemical databases and combinatorial library design as one of the evaluation indices for ADME properties in ongoing drug discovery programs. Finally, the approaches applied in this paper to develop the in silico screen for human S9 metabolic turnover can be used in ADME QSPR studies of other series of compounds.

References

(1) Kennedy, T. Managing the discovery/development interface. *Drug Discovery Today* **1997**, *2*, 436−444.

(2) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of “drug-likeness”. *Drug Discovery Today* **2000**, *5*, 49−58.

(3) Smith, D. A.; van de Waterbeemd, H. Pharmacokinetics and metabolism in early drug discovery. *Curr. Opin. Chem. Biol.* **1999**, *3*, 373−378.

(4) Walters, W. P. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* **1999**, *3*, 384−387.

(5) Zheng, W.; Tropsha, A. A novel variable selection QSAR approach based on the *k*-nearest neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.

(6) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(7) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *Quant. Struct.−Act. Relat.*, in press.

(8) Department of Drug Metabolism and Pharmacokinetics, Center of Excellence for Drug Discovery for Metabolic and Viral Diseases, GlaxoSmithKline, Research Triangle Park, North Carolina.

(9) *MolConnZ*, version 3.5; Hall Associates Consulting: Quincy, MA, 1998.

(10) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure−activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.*, **1985**, *25*, 64−73.

(11) Randić, M. On characterization on molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(12) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. The E-state fields. Applications to 3D QSAR. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 513−520.

(13) Kier, L. B. Inclusion of symmetry as a shape attribute in kappa-index analysis. *Quant. Struct.−Act. Relat.* **1987**, *6*, 8−12.

(14) Hall, L. H.; Kier, L. B. Determination of topological equivalence in molecular graphs from the topological state. *Quant. Struct.−Act. Relat.* **1990**, *9*, 115−131.

(15) Hall, L. H.; Mohney, B. K.; Kier, L. B. The electrotopological state: An atom index for QSAR. *Quant. Struct.−Act. Relat.* **1991**, *10*, 43−51.

(16) Hall, L. H.; Mohney, B. K.; Kier, L. B. The electrotopological state: Structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.

(17) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: San Diego, CA, 1999.

(18) Kier, L. B.; Hall, L. H. A differential molecular connectivity index. *Quant. Struct.−Act. Relat.* **1991**, *10*, 134−140.

(19) Petitjean, M. Applications of the radius−diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331−337.

(20) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17.

(21) Platt, J. R. Prediction of isomeric differences in paraffin properties. *J. Phys. Chem.* **1952**, *56*, 328.

(22) Shannon, C.; Weaver, W., Eds. Mathematical theory of communication, University of Illinois Press: Urbana, IL, 1963.

(23) Bonchev, D.; Mekenyan, O.; Trinajstic, N. Isomer discrimination by topological information approach. *J. Comput. Chem.* **1981**, *2*, 127−148.

(24) Randić, M. On characterization of chemical structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672−687.

(25) Web site: http://www.sas.com/.

(26) Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. *Chemometrics*; John Wiley & Sons: New York, 1986.

(27) Tropsha, A.; Zheng, W. Identification of the descriptor pharmacophores using variable selection QSAR: Applications to database mining. *Curr. Pharm. Des.* **2001**, *7*, 599−612.

(28) Gilbert, N. *Statistics*; W. B. Sounders, Co.: Philadelphia, PA, 1976.

(29) *SYBYL Theory Manual*, version 6.7; Tripos Associates Inc.: St. Louis, MO, 2000.

(30) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on rational division of experimental datasets into diverse training and test sets. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 357−369.

JM020491T