JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling?

Todd M. Martin,*[,†] Paul Harten,[†] Douglas M. Young,[†] Eugene N. Muratov,[‡,§] Alexander Golbraikh,[‡] Hao Zhu,[∥,⊥] and Alexander Tropsha[‡]

[†]Sustainable Technology Division, National Risk Management Research Laboratory, Office of Research and Development, United States Environmental Protection Agency, 26 West Martin Luther King Drive, Cincinnati, Ohio 45268, United States

[‡]Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, Eshelman School of Pharmacy, University of North Carolina, Beard Hall 301, CB#7563, Chapel Hill, North Carolina 27599, United States

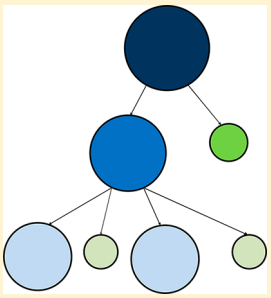
[§]Laboratory of Theoretical Chemistry, Department of Molecular Structure, A.V. Bogatsky Physical Chemical Institute National Academy of Sciences of Ukraine, Lustdorfskaya Doroga 86, Odessa 65080, Ukraine

[∥]Department of Chemistry, Rutgers University, Camden, New Jersey 08102, United States

[⊥]The Rutgers Center for Computational and Integrative Biology, Rutgers University, Camden, New Jersey 08102, United States

Supporting Information

ABSTRACT: Prior to using a quantitative structure activity relationship (QSAR) model for external predictions, its predictive power should be established and validated. In the absence of a true external data set, the best way to validate the predictive ability of a model is to perform its statistical external validation. In statistical external validation, the overall data set is divided into training and test sets. Commonly, this splitting is performed using random division. Rational splitting methods can divide data sets into training and test sets in an intelligent fashion. The purpose of this study was to determine whether rational division methods lead to more predictive models compared to random division. A special data splitting procedure was used to facilitate the comparison between random and rational division methods. For each toxicity end point, the overall data set was divided into a modeling set (80% of the overall set) and an external evaluation set (20% of the overall set) using random division. The modeling set was then subdivided into a training set (80% of the modeling set) and a test set (20% of the modeling set) using rational division methods and by using random division. The Kennard-Stone, minimal test set dissimilarity, and sphere exclusion algorithms were used as the rational division methods. The hierarchical clustering, random forest, and *k*-nearest neighbor (*k*NN) methods were used to develop QSAR models based on the training sets. For *k*NN QSAR, multiple training and test sets were generated, and multiple QSAR models were built. The results of this study indicate that models based on rational division methods generate better statistical results for the test sets than models based on random division, but the predictive power of both types of models are comparable.



1. INTRODUCTION

The main goal of QSAR studies is the development of predictive models which can be used in computer-aided drug discovery and design for prediction of activities or properties of new compounds (i.e., those included in chemical databases or combinatorial libraries). Prior to using the model for external predictions, its predictive power should be established and validated. Thus, model validation has become a standard (and in some laboratories such as ours, mandatory) part of QSAR modeling. In the past, model validation was performed using leave one out (LOO) cross-validation. LOO cross-validation has been determined to yield an overoptimistic estimate of predictive ability (if model feature selection is not redone for each hold out).[1]

External validation is now a "must have" tool for evaluating the reliability of QSAR models.[2] In this procedure, typically the overall set is randomly divided into a training set and a test set. QSAR models are developed based on the training set and then are used to make predictions for the test set. The advantage of this approach is that the test set compounds are "unknown" to the models since they are excluded from the model development procedure, especially the variable selection.

The disadvantage of dividing the overall set randomly is that it does not provide any rationale for selecting test set chemicals. Rational division algorithms have been developed which attempt to "intelligently" divide data sets into training and test sets with the goal of producing more predictive models. Rational design algorithms include the Kohonen Self-Organizing Map method,[3,4] the Kennard-Stone method,[5−8] D-optimal design,[9−11] D-optimal onion design,[12] and sphere exclusion based methods.[8,13,14] Previous studies have indicated the rational division algorithms are superior to the simple random splitting and activity sorting methods[13,15,16] although it is unclear which rational division method is the best because of conflicting results.[12,16]

Received: July 18, 2012

   dx.doi.org/10.1021/ci300338w | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

Wu and co-workers[15] used several different techniques to divide training and test sets for NIR spectral data including the Kennard-Stone method, D-optimal design, Kohonen self-organizing maps, and random selection. The D-optimal and Kennard-Stone methods yielded high classification rates for the three test sets. The Kohonen SOM and random selection methods yielded slightly lower classification rates for the test sets. According to Wu and co-workers, the Kennard-Stone method is more practical than the D-optimal design method. Olsson and co-workers reported that D-optimal design, D-optimal onion design,[12] and the space filling design of Marengo and Todeschini[17] achieved similar prediction errors for an octanol−water partition coefficient data set consisting of 551 diverse chemicals. Gramatica and co-workers used D-optimal experimental design to develop training and test sets for a NO_3 radical tropospheric degradability data set.[18] They noted that since the structure variability of the test set was smaller than that of the training set, the D-optimal design method might have overestimated the prediction power of their QSAR models.

Golbraikh and co-workers[13,14] utilized several variants of the sphere exclusion algorithm to develop training and test sets for several different end points including the ED_{50} for amino acid anticonvulsant agents, antitumor activity of epidophyllotoxin derivatives, binding affinities to the dopamine D_1 receptor, and binding affinities to the histamine H_1 receptor. They reported the sphere exclusion algorithms yielded models with higher predictive power than those obtained from training sets selected from activity ranking and random selection. In addition, the sphere exclusion algorithms were able to produce training and test sets which met the splitting criteria outlined above.

In general, rational design algorithms are designed to meet the following criteria:

1. The training set chemicals must be structurally diverse enough to cover the whole descriptor space of the overall data set.[13,15,19,20]

2. The compounds in the training and test sets should be close to each other.

The goal of this study is to verify whether rational design methods do yield more predictive QSAR models than random splitting approaches. In this study, a special data splitting methodology was employed to assess the benefits of rational design. For each toxicity end point, the overall data set was randomly divided into a modeling set and an evaluation set. The modeling set was then subdivided into a training set and a test set using a rational design method and also by using random selection. QSAR models were developed based on the training sets, and then predictions were made for the test and external evaluation sets.

In this study calculations were performed separately by researchers at the US Environmental Protection Agency (USEPA) and at the University of North Carolina (UNC) at Chapel Hill. The two research groups used different rational design methods and QSAR methodologies in order to fairly evaluate if rational design methods truly improve the external predictive ability of QSAR models. The USEPA group used the Kennard-Stone rational design method and the hierarchical clustering QSAR methodology.[21] The UNC group used the sphere exclusion rational design method[14] and the random forest (Breiman 2001) and *k* Nearest Neighbors (*k*NN) (Zheng and Tropsha, 2000) QSAR methods. Additionally, the minimal test set dissimilarity[22] and random forest approach[23] were used by the UNC team in collaboration with A.V. Bogatsky Physical Chemical Institute NASU (UNC2). Each group performed calculations for four different toxicity end points.

2. MATERIALS AND METHODS

2.1. Experimental Data Sets. A *P. promelas* LC_{50} data set consisting of 809 chemicals was compiled from the ECOTOX database.[24,25] The toxicity is expressed as the concentration (in mol/L)) which kills 50% of fathead minnow after 96 h. The *P. promelas* LC_{50} data set (or subsets of it) has been modeled by several researchers.[26−29]

A *T. Pyriformis* IGC_{50} data set consisting of 1093 chemicals was compiled by Zhu and co-workers[30] from the work of Schultz and co-workers.[31−33] The toxicity is expressed as the 50% growth inhibitory concentration (in mmol/L) of the *T. pyriformis* organism (a protozoan ciliate) after 40 h. This data set represents the largest number of toxicity data tested in a single laboratory by a single, reliable, and robust method.[34] Our final IGC_{50} data set consisted of 1085 chemicals (8 salts were omitted). This data set has been modeled using various sophisticated QSAR techniques,[21,30] but the training and test sets were not generated by rational division techniques in these studies.

An oral rat LD_{50} data set of 7286 chemicals was compiled from the ChemIDplus database.[25,35] The oral rat LD_{50} end point represents the amount of the chemical (mass of the chemical per body weight of the rat) which when orally ingested kills half of rats. This data set was recently modeled by a variety of QSAR techniques,[36] but the training set was selected to match the training set used in the Toxicity Prediction by Komputer Assisted Technology (TOPKAT) software in this study.

A bioaccumulation factor (BCF) data set of 643 chemicals was compiled from several different databases.[37−40] The bioconcentration factor BCF is defined as the ratio of the chemical concentration in biota as a result of absorption via the respiratory surface to that in water at steady state.[41] The final data set consists of 610 chemicals (after removing salts, mixtures, and ambiguous compounds). The modeled end point was the Log (BCF).

The structures for all of the chemicals in these data sets were validated according to the procedure detailed by Young et al.[42]

2.2. Test and External Evaluation Sets. A special splitting procedure was used to facilitate the comparison between the random division and rational division methods into the training and test sets. For each toxicity end point, the overall data set was randomly divided into a modeling set (80% of the overall set) and external evaluation set (20% of the overall set) (Figure 1). The modeling set was then subdivided into Training Set I (80% of the modeling set) and Test Set I (20% of the modeling set) using the rational division algorithms (Kennard-Stone for USEPA, sphere exclusion and minimal test set dissimilarity for UNC). Multiple training and test sets were generated for the *k*NN method by using the sphere exclusion method. Next, the modeling set was also subdivided into Training Set II and Test Set II using random division. The training sets were used for QSAR model development. Predictions were then made for the associated test set (i.e., the models based on Training Set I were used to make predictions for Test Set I) and then for the evaluation set.

2.3. Kennard-Stone (KS) Algorithm (USEPA). The Kennard-Stone algorithm[5−8] was developed to produce a

dx.doi.org/10.1021/ci300338w | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

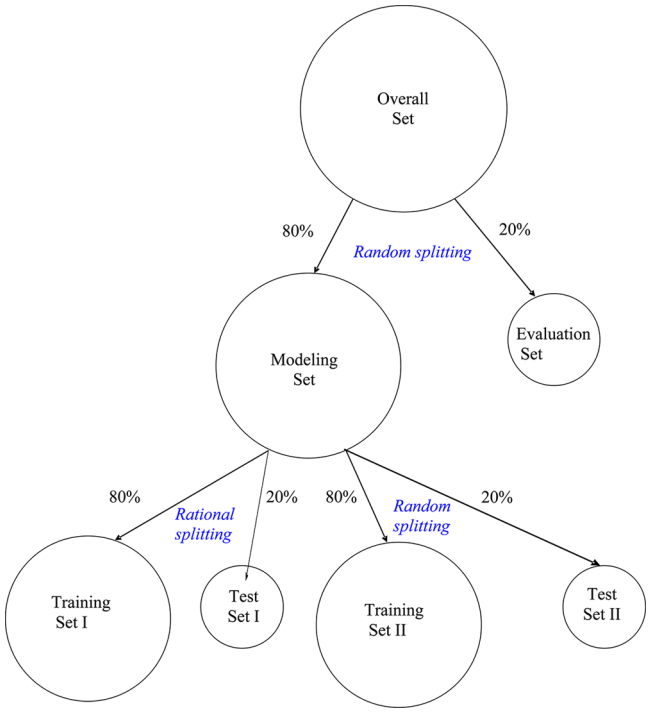


Figure 1. Splitting methodology.

division when no standard experimental design can be applied. The Kennard-Stone algorithm selects the objects so they are divided evenly throughout the descriptor space of the original data set. The Kennard-Stone algorithm as carried out in this study is as follows:

1. The first two members of the training set are selected by choosing the two chemicals that are the farthest apart in terms of Euclidean distance.[5] The Euclidean distance between chemical i and chemical j is given by

$$distance_{ij} = \sqrt{\sum_{k=1}^{\#descriptors} (X_{ik}^n - X_{jk}^n)^2} \quad (1)$$

where X_{ik}^n and X_{jk}^n are the normalized descriptor values for descriptor k for chemicals i and j. The normalized descriptor values are given by

$$X_{ik}^n = \frac{X_{ik} - \bar{X}_k}{\sigma_k} \quad (2)$$

where X_{ik}^n and X_{ik} are the normalized and non-normalized descriptor values for descriptor k for compound i, and $\bar{X}_k$ and σ_k are the mean and standard deviation for descriptor k for the overall set. Descriptors with standard deviations of zero were removed from the list of descriptors.

2. Find the chemical which has the maximum dissimilarity (maximum minimum distance) from each of the previously selected chemicals and place this chemical in the training set.

3. Repeat step 2 until the desired number of chemicals have been added to the training set. In this study 80% of the compounds in the modeling set were placed in the training set.

4. The remaining chemicals (20% of the modeling set) were placed in the test set.

2.4. Sphere Exclusion Algorithm (UNC). The sphere exclusion algorithm[14] was used to generate training and test sets for random forest and kNN model development. In the latter case, multiple pairs of training and test sets were generated. In these calculations, one descriptor from a pair of highly correlated descriptors was excluded (threshold of $|R| = 0.95$ was used); constant and near-constant descriptors were also excluded. Then the remaining descriptors were normalized by range-scaling so as each descriptor was distributed within the interval $[0,1]$

$$X_{ij}^n = \frac{X_{ij} - X_{j,\min}}{X_{j,\max} - X_{j,\min}} \quad (3)$$

where X_{ij}^n and X_{ij} are the normalized and non-normalized values of descriptor j for compound i, and $X_{j,\min}$ and $X_{j,\max}$ are the minimum and maximum values of descriptor j. Let $D_{\min}$ and $D_{\max}$ be the minimum and maximum elements of distance matrix **D**, respectively. N probe sphere radii are defined by the following formulas: $R_{\min} = R_1 = D_{\min}$, $R_{\max} = R_N = D_{\max}/4$ and $R_i = R_1 + (i - 1) \times (R_N - R_1)/(N - 1)$, where $i = 2, ..., N - 1$. The algorithm begins with the calculation of the distance matrix D between points that represent compounds in the descriptor space. Each probe sphere radius corresponds to one division of the modeling set into a pair of the training and test sets.

A sphere-exclusion algorithm used in the present study consisted of the following steps:

1. Select a compound or several compounds. For the random forest calculations, one initial compound was selected randomly. For kNN studies, the initial compounds were the most active and most inactive compounds in the set.

2. Include it (or them) in the training set.

3. Construct a probe sphere around this (these) compound(s).

4. Select other compounds from this (these) sphere(s) and include all of them alternately into the test and the training set.

5. Exclude all compounds from within this (these) sphere(s) from further consideration.

6. If no more compounds are left, stop. Otherwise (as in random forest calculations), select the next compound as follows. Let m be the number of probe spheres constructed and n be the number of remaining compounds. Let d_{ij} ($i = 1,...,m$; $j = 1,...,n$) be the distances between the remaining compounds and the probe sphere centers. Select a compound corresponding to the lowest d_{ij} value and go to step 2. In kNN calculations, the next compound was selected randomly from the remaining compounds.

In case of the random forest QSAR method, just one division into training and test sets was performed (in which the size of the test set was 20% of that of the modeling set). For kNN QSAR, the sphere exclusion algorithm was run ten times, and multiple splits into the training and test sets of the same size were selected.

2.5. Minimal Test Set Dissimilarity (UNC). The test set was selected as follows: a dissimilarity matrix for all initial modeling set molecules was developed using structural descriptors obtained using different variable selection procedures.[22] In our opinion, subset of preselected parameters is preferable because a lot of descriptors in the initial pool are weakly correlated with the desired toxicity end point. The dissimilarity matrix is based on the estimation of structural dissimilarity between all investigated molecules. A measure of structural dissimilarity for molecules M, M' can be calculated using the Euclidean distance in the multidimensional space of structural parameters S

dx.doi.org/10.1021/ci300338w | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

$$SD(M, M') = \sqrt{\sum_{i=1}^{n} (S_i - S'_i)^2} \quad (4)$$

where n is the number of descriptors in the data set. Thus, the total structural dissimilarity toward the rest of the initial modeling set compounds can be calculated for every molecule as the sum of the corresponding Euclidean distances. In the beginning, all the compounds were sorted by activity and then placed in bins of N molecules (N = 5 in the case that 20% of compounds were selected for test set) according to their rank. Thus, bin 1 contained five most active compounds; bin 2 contained compounds with ranks 6−10 and so on. Finally, the compound with minimal dissimilarity value was selected into the test set from every bin.

2.6. Molecular Descriptors. The descriptors used by USEPA were described in detail in a previous publication.[21] The descriptors consisted of 790 descriptors in the following classes: E-state values and E-state counts, constitutional descriptors, topological descriptors, walk and path counts, connectivity, information content, 2D autocorrelation, Burden eigenvalue, molecular property (such as the octanol−water partition coefficient), Kappa, hydrogen bond acceptor/donor counts, molecular distance edge, and molecular fragment counts.

The descriptors used by UNC were calculated using Dragon version 5.4.[43] Initial use of Dragon yielded thousands of chemical descriptors for each training set. Redundant descriptors were identified by analyzing correlation coefficients between all pairs of descriptors. If the correlation coefficient between two descriptor types for all training set compounds was higher than 0.95, then one of them was removed. As a result, the number of Dragon descriptors used for the training set ranged from 485 to 1,064 for each toxicity end point. There are overlaps between Dragon descriptors and EPA descriptors, but both included unique types of descriptors as well.

2D Simplex descriptors[44,45] generated by HiT QSAR Software[46] developed in A.V. Bogatsky Physical-Chemical Institute NASU were also used by the UNC team. A detailed description can be found elsewhere.[46,47] 4000−8000 simplex descriptors were generated (depending on the data set) during the initial stage of work. The following kinds of differentiation of atoms in simplexes were used: 1) nature of atom, 2) partial charge, 3) lipophilicity, 4) refraction, and 5) a mark that characterizes atoms as possible hydrogen bond donors or acceptors. Atom and H-bond characteristics were used directly as an atom mark in simplex, but for other differentiations, atom marks were generated corresponding to their (i) partial charge ($A \le -0.05 < B \le 0 < C \le 0.05 < D$), (ii) lipophilicity ($A \le -0.5 < B \le 0 < C \le 0.5 < D$), and (iii) refraction ($A \le 1.5 < B \le 3 < C \le 8 < D$). Constant, low-variance, and correlated ($|R| \ge 0.5$) descriptors were excluded prior to modeling. Thus, descriptor pools of 150−400 simplex descriptors (depending on the data set) were selected for the statistical processing.

2.7. QSAR Methods. *2.7.1. Hierarchical Clustering (USEPA).* The hierarchical clustering method was used to develop several QSAR models for each training set. The hierarchical clustering method is described in greater detail in a previous publication.[21] For completeness, this method is described briefly below. The hierarchical clustering method uses a variation of Ward's Minimum Variance Clustering Method[48] to produce a series of clusters from the initial training set. For a training set of n chemicals, initially there will be n clusters (each cluster contains one chemical and has a variance of zero). Each following step of the method adds two of the clusters together into one cluster so that the increase in variance over all clusters in the system is minimized. The process of combining clusters while minimizing variance continues until all the chemicals are merged into a single cluster. After the clustering is complete, each cluster is analyzed to determine if an acceptable QSAR model can be developed. A genetic algorithm technique is used to determine an ideal descriptor set for characterizing the toxicity values of the chemicals within that cluster.[21] The use of a genetic algorithm based technique insures the models are developed based solely on the knowledge of the chemicals in the training set.

Before any cluster model can be used to make a prediction for a test chemical, it must be determined whether the test chemical falls within the domain of applicability for the model. The first constraint, the model ellipsoid constraint, checks if the test chemical is within the multidimensional ellipsoid defined by the ranges of descriptor values for the chemicals in the cluster (for the descriptors appearing in the cluster model). The model ellipsoid constraint is satisfied, if the leverage of the test compound (h_{00}) is less than the maximum leverage value for all the compounds used in the model.[49] The second constraint, the Rmax constraint, checks if the distance from the test chemical to the centroid of the cluster is less than the maximum distance for any chemical in the cluster to the cluster centroid. The final constraint is the cluster must contain at least one example of each of the molecular fragments that are present in the test chemical.[21] The predictions are made using the weighted average of the predictions from the closest cluster from each step in the hierarchical clustering.[21] That is each predicted value from this method represents the average of the predictions from one or more multilinear regression models.

2.7.2. Random Forest (UNC). In machine learning, a random forest is a predictor that consists of many decision trees and outputs the prediction that combines outputs from the individual trees. The random forest algorithm was developed by Breiman and Cutler.[50] In this study, the implementation of the random forest algorithm available in the R Project (version 2.7.1)[51] was used. In the random forest modeling procedure, N samples were randomly drawn with replacement from the original data set. These samples were used to construct n training sets and to build n trees. For each node of the tree, m variables were randomly chosen from all the available descriptors. The best data split was obtained using these m variables for each training set. In this study, only the defined parameters (n = 500 and m = 13) were used for the model development. Predictions were made by averaging predicted activities over all trees in the final forest.

Another implementation of random forest approach[23] was also used by the UNC team for processing of simplex descriptors. The trial version of the random forest software is available at http://qsar.110mb.com by request. Each tree was grown as follows:

1. A bootstrap sample, which will be a training set for the current tree, is produced from the whole modeling set of N compounds. Compounds which are not in the current tree training set are placed in the out-of-bag (OOB) set (OOB set size ≈ $N/3$).

2. The best split by the CART algorithm is chosen from among the m randomly selected descriptors from the initial pool. m was the only tuning parameter and was selected according to maximal value of R^2_{oob}.

dx.doi.org/10.1021/ci300338w | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

3. Each tree is grown to the largest possible extent (without pruning). The number of trees in the forest, n, was set at 500 (250 for the oral rat LD50 data set). R^2_{oob} is the major criterion for model selection.

Predictions were made by averaging predicted activities over all trees in the final forest.

2.7.3. k Nearest Neighbors QSAR (UNC). The *k*NN QSAR method employs the *k*NN pattern recognition principle and a stochastic variable selection procedure. First, a subset of *nvar* descriptors is selected randomly. The model based on this set of *nvar* descriptors is validated by LOO cross-validation, where each compound is eliminated from the training set and its biological activity is predicted as the weighted average activity of k most similar molecules (k = 1 to 5). The similarity between compounds was characterized by the Euclidean distance between compounds in multidimensional descriptor space. In general, the Euclidean distances in the descriptor space between a compound and each of its k nearest neighbors are not the same. Thus, the neighbor with the smaller distance from a compound is given a higher weight in calculating the predicted activity as follows

$$\hat{y}_i = \frac{\sum_{j=1}^{k} y_j w_{ij}}{\sum_{j=1}^{k} w_{ij}} \tag{5}$$

where y_j is the observed activity value for nearest neighbor j, and $\hat{y}_i$ is the predicted activity value for compound i; weights w_{ij} are defined as

$$w_{ij} = \left(1 + \frac{d_{ij}^2}{\sum_{j'=1}^{k} d_{ij'}^2}\right)^{-1} \tag{6}$$

d_{ij} are the Euclidean distances between compound i and its k nearest neighbors. A method of simulated annealing (SA) with the Metropolis-like acceptance criteria is used to sample the entire descriptor space in order to maximize the value of LOO R^2 (q^2)

$$q^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2} \tag{7}$$

The value k is optimized during the model building process to give the best prediction of the training set. For prediction, the same equations and k value are used. Additional details on the implementation of the SA procedure for variable selection are given elsewhere.[26]

The applicability domain is defined as the Euclidean distance threshold D_T between a compound under prediction and its closest nearest neighbor of the training set. It is calculated as follows:

$$D_T = \bar{y} + Z\sigma \tag{8}$$

Here $\bar{y}$ is the average Euclidean distance between each compound and its k nearest neighbors in the training set (where k is the parameter optimized in the course of QSAR modeling, and the distances are calculated using descriptors selected by the optimized model only), σ is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level [2,3,7,24]. The default value of Z was set at 0.5, which formally places the allowed distance threshold at the mean plus one-half of the standard deviation. The AD was also defined in the entire descriptor space. In this case, eq 8 was used, k = 1, Z = 0.5, and the Euclidean distances were calculated using all descriptors. Thus, if the distance of the external compound from its nearest neighbor in the training set within either the entire descriptor space or the selected descriptor space exceeds these thresholds, the prediction is not made. In general, predictive power decreases while Z-cutoff value increases [25].

The *k*NN QSAR models were considered acceptable if the following conditions were satisfied for the test sets[13]

$$q^2 > 0.5 \tag{9}$$

$$R^2 > 0.6 \tag{10}$$

$$\frac{(R^2 - R_0^2)}{R^2} < 0.1 \text{ and } 0.9 \le k \le 1.1 \tag{11}$$

or

$$\frac{(R^2 - R'^2_0)}{R^2} < 0.1 \text{ and } 0.9 \le k' \le 1.1 \tag{11'}$$

and

$$|R_0^2 - R'^2_0| < 0.3 \tag{12}$$

where q^2 is the leave one out correlation coefficient for the training set, R^2 is the square of the Pearson's correlation coefficient (coefficient of determination) between the observed and predicted toxicities for the test set, R_0^2 and R'^2_0 are coefficients of determination for trendlines through the origin between the observed and predicted and predicted and observed toxicities for the test sets. The acceptable models were used for consensus prediction of the external evaluation sets: each compound of the external evaluation set was predicted by all acceptable models, for which it was within the applicability domain, and the mean predicted value over all these models was obtained and used as the consensus prediction value.

*k*NN QSAR models were built for three data sets: LC_{50}, IGC_{50}, and BCF. For LD_{50} data set, due to its large size only similarity search models (*k*NN QSAR without variable selection, i.e. using all descriptors) were used in this study.

3. RESULTS AND DISCUSSION

The statistical results for the USEPA (Kennard-Stone and hierarchical clustering methods) group are presented in Tables 1 and 2. The results for the UNC group (sphere exclusion and random forest methods) are presented in Tables 3 and 4. The results for the UNC group (sphere exclusion and the *k*NN

Table 1. Splitting Results in Terms of the R^2 Correlation Coefficient (USEPA − Kennard-Stone and Hierarchical Clustering)

	R^2 test set		R^2 external evaluation set		R^2 external evaluation set (no applicability domain, i.e. 100% coverage)	
end point	rational	random	rational	random	rational	random
LC_{50}	0.81	0.66	0.67	0.60	0.67	0.58
IGC_{50}	0.86	0.84	0.85	0.85	0.80	0.83
LD_{50}	0.76	0.53	0.55	0.53	0.47	0.47
BCF	0.89	0.80	0.73	0.81	0.71	0.68

dx.doi.org/10.1021/ci300338w | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Splitting Results in Terms of the Prediction Coverage (USEPA − Kennard-Stone and Hierarchical Clustering)

	coverage test set		coverage external evaluation set	
end point	rational	random	rational	random
LC_{50}	100%	95%	99%	99%
IGC_{50}	100%	99%	98%	98%
LD_{50}	98%	82%	84%	83%
BCF	100%	96%	95%	94%

Table 3. Splitting Results in Terms of the R^2 Correlation Coefficient (UNC - Sphere Exclusion and Random Forest)

	R^2 test set		R^2 external evaluation set		R^2 external evaluation set (no applicability domain, i.e. 100% coverage)	
end point	rational	random	rational	random	rational	random
LC_{50}	0.72	0.66	0.69	0.67	0.68	0.65
IGC_{50}	0.82	0.81	0.81	0.80	0.81	0.81
LD_{50}	0.74	0.65	0.66	0.65	0.63	0.62
BCF	0.88	0.83	0.87	0.89	0.81	0.80

Table 4. Splitting Results in Terms of the Prediction Coverage (UNC - Sphere Exclusion and Random forest)

	coverage test set		coverage external evaluation set	
end point	rational	random	rational	random
LC_{50}	100%	80%	85%	82%
IGC_{50}	100%	83%	86%	79%
LD_{50}	100%	79%	85%	79%
BCF	100%	82%	81%	77%

methods are presented in Tables 5 and 6). The results for the UNC2 group (minimal test set dissimilarity and random forest)

Table 5. Splitting Results in Terms of the R^2 Correlation Coefficient (UNC - Sphere Exclusion and *k*NN)

	R^2(av) test set		R^2 external evaluation set		R^2 external evaluation set (no applicability domain, i.e. 100% coverage)	
end point	rational	random	rational	random	rational	random
LC_{50}	0.65	0.63	0.64	0.62	0.62	0.61
IGC_{50}	0.80	0.78	0.82	0.80	0.81	0.79
LD_{50}			0.61	0.59	0.56	0.54
BCF	0.81	0.75	0.81	0.8	0.81	0.79

Table 6. Splitting Results in Terms of the Prediction Coverage (UNC - Sphere Exclusion and *k*NN)

	coverage test set		coverage external evaluation set	
end point	rational	random	rational	random
LC_{50}	97%	83%	90%	84%
IGC_{50}	94%	80%	86%	82%
LD_{50}			81%	71%
BCF	49%	39%	83%	85%

are presented in Tables 7 and 8. The results in Tables 1-8 are given in bar chart form in the Supporting Information.

The statistical difference in the correlation coefficients using random division and rational selection was determined using a two sided comparison[52]

$$\hat{z} = \frac{|\dot{z}_1 - \dot{z}_2|}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \tag{13}$$

where n_i is the sample size and

$$\dot{z}_i = \frac{1}{2}\ln\frac{1 + r_i}{1 - r_i} \tag{14}$$

where r_i is the square root of the R^2 correlation coefficient for method i. In order to reject the null hypothesis that the correlation coefficients are different, $\hat{z}$ must be greater than 1.96 ($\alpha = 0.05$).

3.1. Results for the Test Sets. Results from the USEPA group indicated that for all four data sets, the Kennard-Stone rational division method yielded better results than those for random division ($\Delta R^2 = 0.02-0.23$) for the *test* set (see Table 1). The R^2 values were statistically different ($\hat{z} > 1.96$) for all of the end points except for the IGC_{50} end point ($\hat{z} = 0.67$). The difference in prediction coverage (4−5%) was small for three out of four data sets (see Table 2). For the oral rat LD_{50} data set, however, there was a 16% increase in prediction coverage.

The test set results for UNC (sphere exclusion and random forest methods) are similar to those of the USEPA group; however, the difference in the prediction statistics between the rational (sphere exclusion) and random divisions was lower ($\Delta R^2 = 0.01-0.09$) (see Table 3). The R^2 values were statistically different only for the LD_{50} data set. Finally, the prediction coverages were about 20% higher (see Table 4).

The test set results for UNC (sphere exclusion and *k*NN methods) are presented in Tables 5 and 6. It should be noted that multiple models have been built for only three out of four data sets (the oral rat LD_{50} contains too many compounds for the *k*NN approach to be practical). Since multiple divisions into training and test sets were used in these calculations, the average R^2 over all test sets are given in Table 5. The correlation coefficients for rational division were slightly higher than those for random division ($\Delta R^2 = 0.02-0.06$). Coverage values, however, were 10−14% higher than those for random division. The fact that rational selection yielded 1.6−2.1 times as many models as random division could explain the increase in prediction coverage.

The results of UNC2 team for the test set were slightly different from the results for the USEPA and UNC groups (see Tables 7 and 8). Rational division achieved better results ($\Delta R^2 = 0.03-0.19$) for three of the end points but performed slightly worse for the LD_{50} data set. The difference in correlation coefficients was only statistically significant for the BCF data set (although $\hat{z}$ ranged from 1.55 to 1.79 for the other end points). Again the difference in coverage was small (except for the LC_{50} end point where the difference in coverage was 18%).

In summary rational division methods (Kennard-Stone, sphere exclusion, and minimal test set dissimilarity) usually yielded better results than random division for the *test* sets. It was expected that the prediction results obtained using rational division would be better (higher predictivity and coverage) than for random selection because in rational division test set molecules would be more similar to training set compounds.

dx.doi.org/10.1021/ci300338w | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 7. Splitting Results in Terms of the R^2 Correlation Coefficient (Minimal Test Set Dissimilarity and Random Forest Approach)

	R^2 test set		R^2 external evaluation set			R^2 external evaluation set (no applicability domain, i.e. 100% coverage)		
end point	rational	random	rational	random	OOB	rational	random	OOB
LC_{50}	0.72	0.60	0.66	0.69	0.66	0.63	0.67	0.65
IGC_{50}	0.91	0.88	0.84	0.85	0.85	0.84	0.85	0.85
LD_{50}	0.53	0.58	0.59	0.58	0.65	0.58	0.58	0.64
BCF	0.88	0.69	0.84	0.83	0.84	0.83	0.83	0.83

Table 8. Splitting Results in Terms of the Prediction Coverage (Minimal Test Set Dissimilarity and Random Forest Approach)

	coverage test set		coverage external evaluation set		
end point	rational	random	rational	random	OOB
LC_{50}	89%	71%	68%	71%	62%
IGC_{50}	99%	97%	98%	96%	97%
LD_{50}	98%	95%	96%	96%	95%
BCF	99%	94%	94%	95%	92%

Five times out of fifteen the difference in predictivity was statistically significant, and six times the difference in coverage exceeded 15%.

3.2. Results for the External Evaluation Sets. For the USEPA group, there was no statistical difference between Kennard-Stone rational division method and random division (ΔR^2 = −0.08−0.07) for the *external evaluation* set (when the applicability domain was applied). There was also no statistical difference between the splitting methods when no applicability domain was applied. In addition there was no significant difference in coverage (ΔCOV ∼ 0.01). Again the UNC (sphere exclusion and random forest methods) results agreed with those of the USEPA group: there was no difference between sphere exclusion and random divisions (ΔR^2 = −0.02−0.02). This result did not change when the applicability domain was removed. Coverage of models obtained for the sphere exclusion algorithm was a little bit higher than for random selection (ΔCOV = 0.03−0.07); however, this difference was much smaller than for the test set (ΔCOV ∼ 0.20).

For UNC (sphere exclusion and *k*NN methods) prediction results for external evaluation sets are given in Tables 5 and 6. The prediction statistics (R^2 and coverage) are for prediction by training sets by at least 50% of models. We can see that in all cases sphere-exclusion slightly outperforms random division (ΔR^2 = 0.01−0.02). However these differences were not statistically significant (the maximum $\hat{z}$ was 0.8). Removing the applicability domain did not change this result. For three of the end points, rational division yielded a slight improvement in prediction coverage (ΔCOV = 4−10%). For the BCF end point, the prediction coverage was slightly lower.

The results for the UNC2 group were similar to the other two groups- there was virtually no difference in the prediction performance between rational and random division for the external evaluation set. In addition to random and rational divisions, the model with no test set was also developed by UNC2 team. It was created for two reasons. (i) One will not lose any part of structural information contained in the modeling set compounds during test set formation (test set molecules are not used for models development and selection). (ii) RF possesses its own out-of-bag (OOB) set for model development and selection. Every tree in the forest has it own out-of-bag set created by compounds that were not selected in the training set for the given tree. As was described by Breiman,[50] OOB set contains approximately 1/3 of the total number of molecules in the data set. It was shown[53] that prediction accuracy of OOB set and 5-fold external cross-validation procedure is nearly the same. Thus, it was expected that the model with no test set (based on OOB statistics) will be the most accurate and possess the highest coverage for the external evaluation set. However, this was not observed in this study. There was no difference in predictivity and coverage for three out of the four data sets for the external evaluation set. Moreover, for the *P. promelas* data set, the "no test set model" coverage was 9% lower than that for the random division. Only for the oral rat LD_{50} data set was the difference in predictivity between these two models statistically significant. However, the advantage of the model without the test set over the "minimal dissimilarity test set" model was not statistically significant.

3.3. Discussion. In this paper it was shown that rational design yields better results for the internal test set but does not yield better results for the external test set. This can be explained by the fact that rational design ensures that each member of the test set will have structural analogs in the training set (and thus should be more accurately predicted than a randomly selected set). However, the compounds in the external set are not guaranteed to have analogs in the training set. In addition methods such as the Kennard-Stone method place structural outliers in the training set (which may cause problems if the training set is very heterogeneous). Thus while the statistical results are better for the test set, rational design does not guarantee better results for a brand new data set.

In Tables 1, 3, 5, and 7, two sets of results are given for each external set. In the first set applicability domains were applied, and in the second set applicability domains were ignored. In both sets there is no statistical difference between the random selection and rational design methods. This indicates that rational selection will not yield significantly better results for high-confidence predictions (in terms of the applicability domains).

4. CONCLUSIONS

In these studies, we tried to address the important problem of the QSAR model validation: which method of division of a data set into training and test sets is better: rational or random. For this purpose, four data sets were selected. To approach the problem, each data set was randomly divided into a modeling and external evaluation set that included 80% and 20% of the entire data set, respectively. The modeling sets were divided into training and test sets rationally and randomly. Training and test sets included 80% and 20% of the modeling sets, respectively. Training sets were used to build QSAR models, and test sets were used for their validation. External evaluation

dx.doi.org/10.1021/ci300338w | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

sets were used to find the "true" predictive power of QSAR models. Test and external evaluation sets were not used in the process of model building. The following rational division methods were used: Kennard-Stone algorithm, sphere exclusion algorithm, and minimal test set dissimilarity method. The following QSAR methods were used: hierarchical clustering, two implementations of random forest, and *k* nearest neighbors (*k*NN) QSAR.

In hierarchical clustering, the Kennard-Stone algorithm was used to rationally divide a modeling set into training and test sets. For one of the random forest implementations, the sphere exclusion algorithm was used to generate one pair of training and test sets. For another random forest implementation, the minimal test set dissimilarity method was used. These three methods of division of the modeling set into training and test sets were found to have no observed influence on model predictivity or coverage for the external evaluation sets for the QSAR methods that were used. Since both test and external evaluation set were performed in "blind prediction" mode, one can compare their prediction performance. It was observed that the rational selection methods yielded better prediction statistics for the test set but not for the external evaluation set. Thus, they yield an overly optimistic estimate of prediction ability. For these rational division methods and their corresponding QSAR methodologies, random selection provides a more accurate estimate of prediction ability since the statistics for the test set and external evaluation set are essentially equivalent.

For *k*NN QSAR studies, division of a data set into training, test, and external evaluation sets is a common practice. In this study we found that roughly twice as many predictive models were generated using the sphere exclusion algorithm than by using random division. The additional models may account for the 5% (on average) increase in the prediction coverage for the external set. Since the R^2 values for the test and external sets are comparable, the use of the sphere exclusion algorithm does not yield an unrealistic estimate of model performance. Taking into account these results, it is our opinion that in *k*NN QSAR studies sphere exclusion rather than random division should be used. However, since the difference between the results is relatively small, random division is also acceptable.

■ ASSOCIATED CONTENT

Supporting Information

The results in Tables 1-8 are given in bar chart form. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: (513)-569-7682. E-mail: martin.todd@epa.gov.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.

(2) Eriksson, L.; Jaworska, J. S.; 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 Perspect.* **2003**, *111*, 1361−1375.

(3) Zupan, J.; Novic, M.; Ruisanchez, I. Kohonen and Counterpropagation Artificial Neural Networks in Analytical Chemistry. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 1−23.

(4) Gasteiger, J.; Zupan, J. Neuronale Netze in der Chemie. *Angew. Chem.* **1993**, *32*, 503.

(5) Bourguignon, B.; de Aguiar, P. F.; Khots, M. S.; Massart, D. L. Optimization in Irregularly Shaped Regions: pH and Solvent Strength in Reversed-Phase High-Performance Liquid Chromatography Separations. *Anal. Chem.* **1994**, *66*, 893−904.

(6) Bourguignon, B.; de Aguiar, P. F.; Thorre, K.; Massart, D. L. Application of Nonlinear-Regression Functions for the Modeling of Retention in Reversed-Phase LC. *J. Chromatogr. Sci.* **1994**, *32*, 144−152.

(7) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137−148.

(8) Snarey, M.; Terrett, N. K.; Willet, P.; Wilton, D. J. Comparison of Algorithms for Dissimilarity-Based Compound Selection. *J. Mol. Graphics Modell.* **1997**, *15*, 372−385.

(9) Federov, V. V. *Theory of Optimal Experiments, Moscow University*; Academic Press: New York, 1972.

(10) de Aguiar, P. F.; Bourguignon, B.; Khots, M. S.; Massart, D. L.; Phan-Than-Luu, R. D-Optimal Designs. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 199−210.

(11) Carlson, R. *Design and Optimization in Organic Synthesis*; Elsevier: Amsterdam, 1992.

(12) Olsson, I.-M.; Gottfries, J.; Wold, S. D-Optimal Onion Designs in Statistical Molecular Design. *Chemom. Intell. Lab. Syst.* **2004**, *73*, 37−46.

(13) Golbraikh, A.; Tropsha, A. Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Training and Test Set Selection. *J. Comput.-Aided. Mol. Des.* **2002**, *16*, 357−369.

(14) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. Rational Selection of Training and Test sets for the Development of Validated QSAR Models. *J. Comput.-Aided. Mol. Des.* **2003**, *17*, 241−253.

(15) Wu, W.; Walczak, B.; Massart, D. L.; Heuerding, S.; Erni, F.; Last, I. R.; Prebble, K. A. Artificial Neural Networks in Classification of NIR Spectral Data: Design of the Training Set. *Chemom. Intell. Lab. Syst.* **1996**, *33*, 35−46.

(16) Gramatica, P.; Pilutti, P. *Evaluation of different statistical approaches for the validation of quantitative structure-activity relationships*; The European Commission - Joint Research Centre, Institute for Health & Consumer Protection - ECVAM: Ispra, Italy, 2004.

(17) Marengo, E.; Todeschini, R. A New Algorithm for Optimal, Distance-Based Experimental Design. *Chemom. Intell. Lab. Syst.* **1992**, *16*, 37−44.

(18) Gramatica, P.; Pilutti, P.; Papa, E. Predicting the NO_3 Radical Tropospheric Degradability of Organic Pollutants by Theoretical Molecular Descriptors. *Atmos. Environ.* **2003**, *37*, 3115−3124.

(19) Sjostrom, M.; Eriksson, L. Applications of Statistical Experimental Design. In *Chemometric Methods in Molecular Design*; Van de Warerbeemd, H., Ed.; VCH: New York, 1995.

(20) Gramatica, P.; Pilutti, P.; Papa, E. Validated QSAR Prediction of OH Tropospheric Degradability: Splitting into Training-Test Set and Consensus Modeling. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1794−1802.

(21) Martin, T. M.; Harten, P.; Venkatapathy, R.; Das, S.; Young, D. M. A Hierarchical Clustering Methodology for the Estimation of Toxicity. *Toxicol. Mech. Methods* **2008**, *18*, 251−266.

(22) Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N.; Volineckaya, I. L.; Makarov, V. A.; Riabova, O. B.; Wutzler, P.; Schmidtke, M. Quantitative Structure-Activity Relationship Studies of [(Biphenyloxy)Propyl]Isoxazole Derivatives - Human Rhinovirus 2 Replication Inhibitors. *J. Med. Chem.* **2007**, *50*, 4205−4213.

(23) Polishchuk, P. G.; Muratov, E. N.; Artemenko, A. G.; O.G., K.; Muratov, N. N.; Kuz'min, V. E. Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity. *J. Chem. Inf. Model.* **2009**, *49*, 2481−2488.

(24) USEPA. ECOTOX Database. http://cfpub.epa.gov/ecotox/ (accessed 10/23/08).

(25) USEPA. User's Guide for T.E.S.T. (Toxicity Estimation Software Tool) A Program to Estimate Toxicity from Molecular Structure, Version 3.0. http://www.epa.gov/ORD/NRMRL/std/cppb/qsar/testuserguide.pdf (accessed 10/27/09).
(26) Eldred, D. V.; Weikel, C. L.; Jurs, P. C.; Kaiser, K. L. E. Prediction of Fathead Minnow Acute Toxicity of Organic Compounds from Molecular Structure. *Chem. Res. Toxicol.* **1999**, *12*, 670−678.
(27) He, L.; Jurs, P. C. Assessing the Reliability of QSAR Model'S Predictions. *J. Mol. Graphics Modell.* **2005**, *23*, 503−523.
(28) Kaiser, K. L. E.; Niculescu, S. P. Using Probabilistic Neural Networks To Model the Toxicity of Chemicals to the Fathead Minnow (Pimephales promelas): A Study Based on 865 Chemicals. *Chemosphere* **1999**, *38*, 3237−3245.
(29) Martin, T. M.; Young, D. M. Prediction of the Acute Toxicity (96-h LC_{50}) of Organic Compounds to the Fathead Minnow (*Pimephales promelas*) Using a Group Contribution Method. *Chem. Res. Toxicol.* **2001**, *14*, 1378−1385.
(30) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Öberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinational QSAR Model of Chemical Toxicants Tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **2008**, *48*, 766−784.
(31) Schultz, T. W. TETRATOX: *Tetrahymena pyriformis* Population Growth Impairment Endpoint - A Surrogate for Fish Lethality. *Toxicol. Methods* **1997**, *7*, 289−309.
(32) Schultz, T. W. TETRATOX. http://www.vet.utk.edu/TETRATOX/index.php (accessed 2/24/11).
(33) Schultz, T. W.; Netzeva, T. I. Development and Evaluation of QSARs for Ecotoxic Endpoints: The Benzene Response-Surface Model for Tetrahymena Toxicity. In *Modeling Environmental Fate and Toxicity*; Cronin, M. T. D., Livingstone, D. J., Eds.; CRC Press: Boca Raton, FL, 2004.
(34) Cronin, M. T. D.; Aptula, A. O.; Duffy, J. C.; Netzeva, T. I.; Rowe, P. H.; Valkova, I. V.; Schultz, T. W. Comparative Assessment of Methods To Develop QSARs for the Prediction of the Toxicity of Phenols to *Tetrahymena pyriformis*. *Chemosphere* **2002**, *49*, 1201−1221.
(35) NLM. ChemIDplus Advanced Database. http://chem.sis.nlm.nih.gov/chemidplus/ (accessed 2/24/11).
(36) Zhu, H.; Martin, T. M.; Ye, L.; Sedykh, A.; Young, D. M.; Tropsha, A. Quantitative Structure-Activity Relationship Modeling of Rat Acute Toxicity by Oral Exposure. *Chem. Res. Toxicol.* **2009**, *22*, 1913−1921.
(37) Dimitrov, S.; Dimitrova, N.; Parkerton, T.; Combers, M.; Bonnell, M.; Mekenyan, O. Base-Line Model for Identifying the Bioaccumulation Potential of Chemicals. *SAR QSAR Environ. Res.* **2005**, *16*, 531−554.
(38) Arnot, J. A.; Gobas, F. A. P. C. A Review of Bioconcentration Factor (BCF) and Bioaccumulation Factor (BAF) Assessments for Organic Chemicals in Aquatic Organisms. *Environ. Rev.* **2006**, *14*, 257−297.
(39) EURAS. Establishing a bioconcentration factor (BCF) Gold Standard Database. http://www.euras.be/eng/project.asp?ProjectId=92 (accessed 5/20/09).
(40) Zhao, C. B., E.; Chana, A.; Roncaglioni, A.; Benfenati, E. A New Hybrid System of QSAR Models for Predicting Bioconcentration Factors (BCF). *Chemosphere* **2008**, *73*, 1701−1707.
(41) Hamelink, J. L. Current Bioconcentration Test Methods and Theory. In *Aquatic Toxicology and Hazard Evaluation*; Mayer, F. L., Hamelink, J. L., Eds.; ASTM STP: West Conshohocken, PA, 1977; Vol. *634*, pp 149−161.
(42) Young, D. M.; Martin, T. M. Are the Chemical Structures in Your QSAR Correct? *QSAR Comb. Sci.* **2008**, *27*, 1337−1345.
(43) Talete. Dragon Version 5.4. http://www.talete.mi.it/ (accessed 5/26/09).
(44) Artemenko, A. G.; Muratov, E. N.; Kuz'min, V. E.; Kovdienko, N. A.; Hromov, A. I.; Makarov, V. A.; Riabova, O. B.; Wutzler, P.; Schmidtke, M. Identification of Individual Structural Fragments of N,N′-(Bis-5-nitropyrimidyl)dispirotripiperazine Derivatives for Cytotoxicity and Antiherpetic Activity Allows the Prediction of New Highly Active Compounds. *J. Antimicrob. Chemother.* **2007**, *60*, 68−77.
(45) Muratov, E. N.; Artemenko, A. G.; Kuz'min, V. E.; Lozitsky, V. P.; Fedchuk, A. S.; Lozitska, R. N.; Boschenko, Y. A.; Gridina., T. L. Investigation of Anti-Influenza Activity Using Hierarchic QSAR Technology on the Base of Simplex Representation of Molecular Structure. *Antiviral Res.* **2005**, *65*, A62−A63.
(46) Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N. Hierarchical QSAR Technology on the Base of Simplex Representation of Molecular Structure. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 403−421.
(47) Kuz'min, V. E.; Artemenko, A. G.; Lozitsky, V. P.; Muratov, E. N.; Fedtchouk, A. S.; Dyachenko, N. S.; Nosach, L. N.; Gridina, T. L.; Shitikova, L. I.; Mudrik, L. M.; Mescheriakov, A. K.; Chelombitko, V. A.; Zheltvay, A. I.; Eynde, J.-J. V. The Analysis of Structure- Anticancer and Antiviral Activity Relationships for Macrocyclic Pyridinophanes and Their Analogues on the Basis of 4D QSAR Models (Simplex Representation of Molecular Structure). *Acta Biochim. Pol.* **2002**, *49*, 157−168.
(48) Romesburg, H. C. *Cluster Analysis for Researchers*; Lifetime Learning Publications: Belmont, CA, 1984.
(49) Montgomery, D. C. *Introduction to linear regression analysis*; John Wiley and Sons: New York, 1982.
(50) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.
(51) Dalgaard, P. *Introductory Statistics with R*; Springer: New York, 2008.
(52) Sachs, L. *Applied Statistics. A Handbook of Techniques*; Springer-Verlag: New York, 1984.
(53) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Model.* **2003**, *43*, 1947−1958.