ARTICLES

Critical Assessment of QSAR Models of Environmental Toxicity against *Tetrahymena pyriformis:* Focusing on Applicability Domain and Overfitting by Variable Selection

Igor V. Tetko,*[,†,‡] Iurii Sushko,[†] Anil Kumar Pandey,[†] Hao Zhu,[§] Alexander Tropsha,[§] Ester Papa,[||] Tomas Öberg,[⊥] Roberto Todeschini,[#] Denis Fourches,[▽] and Alexandre Varnek[▽]

Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute of Bioinformatics and Systems Biology, Neuherberg D-85764, Germany, Institute of Bioorganic & Petrochemistry, National Ukrainian Academy of Sciences, Kyiv-94 02660, Ukraine, Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products and Carolina Exploratory Center for Cheminformatics Research, School of Pharmacy, CB 7360, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria, Varese, Italy, School of Pure and Applied Natural Sciences, University of Kalmar, SE-391 82 Kalmar, Sweden, Department of Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza, 1, 20126 Milano, Italy, and Laboratoire d'Infochimie, Institut de Chimie, Louis Pasteur University, Strasbourg, France

Received May 2, 2008

The estimation of the accuracy of predictions is a critical problem in QSAR modeling. The "distance to model" can be defined as a metric that defines the similarity between the training set molecules and the test set compound for the given property in the context of a specific model. It could be expressed in many different ways, e.g., using Tanimoto coefficient, leverage, correlation in space of models, etc. In this paper we have used mixtures of Gaussian distributions as well as statistical tests to evaluate six types of distances to models with respect to their ability to discriminate compounds with small and large prediction errors. The analysis was performed for twelve QSAR models of aqueous toxicity against *T. pyriformis* obtained with different machine-learning methods and various types of descriptors. The distances to model based on standard deviation of predicted toxicity calculated from the ensemble of models afforded the best results. This distance also successfully discriminated molecules with low and large prediction errors for a mechanism-based model developed using *log P* and the Maximum Acceptor Superdelocalizability descriptors. Thus, the distance to model metric could also be used to augment mechanistic QSAR models by estimating their prediction errors. Moreover, the accuracy of prediction is mainly determined by the training set data distribution in the chemistry and activity spaces but not by QSAR approaches used to develop the models. We have shown that incorrect validation of a model may result in the wrong estimation of its performance and suggested how this problem could be circumvented. The toxicity of 3182 and 48774 molecules from the EPA High Production Volume (HPV) Challenge Program and EINECS (European chemical Substances Information System), respectively, was predicted, and the accuracy of prediction was estimated. The developed models are available online at http://www.qspr.org site.

INTRODUCTION

Toxic environmental chemicals may be damaging to the environment and human health, and therefore they represent a considerable danger to society. Unfortunately, there is a great gap in the number of chemical compounds for which experimental physicochemical properties and toxicity data are available and those for which such information is needed.[1] In the 1990s the EPA Office of Toxic Substances (OTS) listed approximately 70000 industrial chemicals, and about 1000 chemicals have been added each year. However, even simple experimental properties have been measured only for a small fraction of these compounds.[2] Moreover, for some important environmental pollutants, such as DDT, there are still no reliable data on their basic physicochemical properties, e.g. water solubility and lipophilicity,[3] despite more than 60 years of studies and more than 7500 articles dealing with the toxicity of this pesticide.[3,4]

The European Union has recently approved a new regulation, called REACH - Registration, Evaluation and Authorization of Chemicals, that will create a database of chemicals

* Corresponding author phone: +49-89-3187-3575; fax: +49-89-3187-3585; e-mail: itetko@vcclab.org. Corresponding author address: Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute for Bioinformatics and Systems Biology, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany.
[†] Institute of Bioinformatics and Systems Biology.
[‡] National Ukrainian Academy of Sciences.
[§] University of North Carolina at Chapel Hill.
[||] University of Insubria.
[⊥] University of Kalmar.
[#] University of Milano-Bicocca.
[▽] Louis Pasteur University.

10.1021/ci800151m CCC: $40.75
Published on Web 08/26/2008

used in the EU (see http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm). This law requires assessment of physicochemical properties and adverse effects (e.g., carcinogenic and mutagenic properties) of all compounds, which are produced in excess of 1 ton/year, which will lead to the registration of more than 30000 compounds. The implementation of REACH requires demonstration, by means of experimental tests, of the safe manufacturing of chemicals and their safe use throughout the supply chain. The total cost of tests required for the registration of compounds is estimated to be 5 billion euros during the next 11 years (http://news.bbc.co.uk/2/hi/europe/4444550.stm).

The REACH advocates the use of nonanimal testing methods and, in particular, QSAR/QSPR approaches in order to decrease the number and costs of animal tests. For example, the REACH system requires that nonanimal methods should be used for the majority of tests in the 1−10 ton band of chemicals produced in large volumes. In November 2004, the OECD member countries agreed on the principles for validating (Q)SAR models to enable their use in regulatory assessment of chemical safety. An OECD Expert Group on (Q)SARs was established for this purpose, and the first version of the OECD (Q)SAR Application Toolbox has been released (http://toolbox.oasis-lmc.org). In February 2007, the OECD published a "Guidance Document on the Validation of (Q)SAR Models" that summarized the (Q)SAR model validation principles accepted by the OECD: http://www.oecd.org/document/23/0,2340,en_2649_34365_33957015_1_1_1_1,00.html.

When using *in silico* predictions, one should have a clear understanding and knowledge of the applicability domain of the developed models, which is defined according to the OECD guidelines as "*the response and chemical structure space in which the model makes predictions with a given reliability*".[5,6] Thus, methods to estimate the accuracy of prediction and development of practical protocols for the implementation of REACH are very important. Considering the volume of measurements and lack of experience with the implementation of such global test policies, the EU has launched a number of projects, e.g. OSIRIS (Optimized Strategies for Risk assessment of Industrial Chemicals through integration of nontest and test information), CAESAR (Computer Assisted Evaluation of Substances According to Regulations), NOMIRACLE (NOvel Methods for Integrated Risk Assessment of CumuLative stressors in Europe), MODELKEY (MODELs for assessing and forecasting the impact of environmental KEY pollutants on marine and freshwater ecosystems and biodiversity), and CADASTER (CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment), to meet these goals.

In our previous study[7] we have developed several models using different QSAR modeling approaches to predict aqueous toxicity of molecules against *Tetrahymena pyriformis* and demonstrated that consensus models yielded both the highest chemical space coverage and prediction accuracy. The growth inhibition of the ciliated protozoan *T. pyriformis* is a commonly accepted toxicity screening tool that has been under development and implementation by Schultz and coworkers for many years.[8−12] In the past ten years, this group has published the results from the standard *T. pyriformis* toxicity test protocol for more than 1000 different compounds providing a unique data set for modeling aquatic toxicity. All these data were measured in one laboratory under strict experimental control and thus represent a valuable source for benchmarking QSAR models to access toxicity of chemical compounds.

In the previous study, we have analyzed the performance of both individual and consensus models taking into consideration their applicability domains (ADs).[7] It was shown that there was an increase of the accuracy of predictions for compounds that were within ADs of all models where the AD was defined. The ADs were identified using threshold values of similarities between test and training set molecules using different metrics of distances to models (DM) functions. The threshold values were selected according to our previous individual experiences. However, due to the size limitation of that article and its primary focus on model development and validation we did not discuss the quantitative aspects of the AD nor investigate the respective DM functions in details. The goal of this study is to expand upon the previous one in terms of most suitable quantitative estimates of model prediction accuracy and to provide a quantitative analysis of different types of DM. In particular, we posed the following questions. Are DM defined for one method/set of descriptors be used with other approaches? How can we benchmark different DM approaches? Is there a best definition for the DM? Can we predict errors of property prediction for molecules in the external sets? Following our analysis we have established best practices for model development and validation. Furthermore, we have developed an online server, which predicts aquatic toxicity of chemical compounds from their structure using both individual and consensus models as well as provides robust estimates of the accuracy of predictions.

METHODS

Data Set. The data for our analysis were compiled from several publications of the Schultz group[8−12] and from the Web site of the Tetratox database (http://www.vet.utk.edu/TETRATOX/) as described in our previous study.[7] *T. pyriformis* toxicity of each compound was used as the logarithm of 50% growth inhibitory concentration (pIGC50) values. For the modeling purpose, the data set was divided into three parts: 1) the training set that consisted of 644 compounds; 2) the validation set that consisted of 339 compounds; and 3) the second validation set, which included 110 unique compounds from the most recent publication of the Schultz group.[13] None of the compounds was included in more than one set.

Experimental Accuracy of Data. The experimental analysis of reproducibility of toxicity against *T. pyriformis* was performed by Schultz et al.[11] for 51 molecules. The authors divided all molecules into two groups according to the expected mechanism of their action: chemicals considered as reactive and those thought to have a narcosis mode of action. The authors reported higher variability of measurements for the molecules from the former group. Using their data we estimated $\sigma = 0.38$ (Mean Absolute Error, MAE = 0.24, $N = 27$) and $\sigma = 0.21$ (MAE = 0.13, $N = 24$) for molecules with reactive and narcosis modes, respectively.

QSAR Approaches. Table 1 summarizes QSAR approaches used, and Table 2 summarizes the statistical

Table 1. Overview of Contributing QSAR Modeling Approaches and Distances to Models

					distance to models	
nn	group	modeling techniques	descriptors	abbreviation	descriptor space	property-based space
1	UNC	ensemble of 192 kNN models	MolconnZ	kNN-MZ	EUCLID	STD
2	UNC	ensemble of 542 kNN models	Dragon	kNN-DR	EUCLID	STD
3	VCCLAB	ensemble of 100 neural networks	E-state indices	ASNN-ESTATE		CORREL, STD
4	ULP	kNN	ISIDA fragments	kNN-FR	EUCLID, TANIMOTO	
5	ULP	MLR	ISIDA fragments	MLR-FR	EUCLID, TANIMOTO	
6	UI	OLS	Dragon	OLS-DR	LEVERAGE	
7	UK	PLS	Dragon	PLS-DR	LEVERAGE	PLSEU
8	UNC	SVM	MolconnZ	SVM-MZ		
9	UNC	SVM	Dragon	SVM-DR		
10	ULP	SVM	ISIDA fragments	SVM-FR		
11	ULP	MLR	molecular properties (CODESSA-Pro)	MLR-COD		
12		average of all models	-	CONS		STD

Table 2. Statistical Parameters of the Calculated Models

	training set				validation			
	internal LOO		5-CV		set 1		set 2	
nn	R^{2a}	$RMSE^a$	R^{2a}	$RMSE^a$	R^{2a}	$RMSE^a$	R^{2a}	$RMSE^a$
ASNN-ESTATE	0.84	0.42	0.82	0.44	0.85	0.41	0.66	0.52
kNN-DR	0.92	0.30[c]	0.80	0.50	0.84	0.41	0.59	0.57
kNN-FR	0.77	0.51	0.73	0.55	0.71	0.56	0.37	0.71
kNN-MZ	0.91	0.32[c]	0.76	0.53	0.83	0.43	0.49	0.64
MLR-COD	0.72	0.55	0.69	0.59	0.71	0.57	0.58	0.58
MLR-FR	0.94	0.26[d]	0.74	0.55	0.49	0.56	0.43	0.67
OLS-DR	0.75	0.53	0.77	0.51	0.77	0.50	0.58	0.58
PLS-DR	0.88	0.36[b]	0.79	0.48	0.81	0.46	0.59	0.57
SVM-DR	0.93	0.28[d]	0.81	0.46	0.70	0.57	0.53	0.61
SVM-FR	0.95	0.24[d]	0.80	0.48	0.76	0.51	0.38	0.70
SVM-MZ	0.89	0.35[b]	0.77	0.51	0.77	0.50	0.58	0.58
CONS	0.92	0.31[b]	0.83	0.44	0.85	0.40	0.67	0.51

[a] R^2 is coefficient of determination, and *RMSE* is the Root Mean Squared Error. [b] Corresponds to significant differences in *RMSE* for the training and validation set 1 at the significance level $p < 0.05$. [c] Corresponds to significant differences in *RMSE* for the training and validation set 1 at the significance level $p < 0.01$. [d] Corresponds to significant differences in *RMSE* for the training and validation set 1 at the significance level $p < 0.001$.

parameters for all models. In total, eleven models differing in the types of descriptors and modeling techniques have been applied. Full computational details of each approach as well as analysis of results can be found elsewhere.[7] All QSAR toxicity models were first developed based on the training set only, and their accuracy was estimated using the Leave-One-Out (LOO) cross-validation. Following this analysis we performed "blind prediction" of molecules from both validation sets. The performances of the individual models for the validation sets were used to compare the prediction ability of the models.

Consensus Model. The predicted toxicity for test set molecules using the consensus ensemble model was calculated as a simple nonweighted average of individual predictions with all eleven models listed in Table 1. The statistical parameters of both individual and consensus models are summarized in Table 2. The consensus model had similar prediction ability to that of the Associative Neural Network (ASNN) model for all three sets for 100% coverage.[7]

Distances to Models. Model AD is an active area of modern QSAR research. Generally, there is no universal technique of defining the AD.[6,14,15] Each AD definition is usually based on some arbitrarily defined distance (or similarity) of the analyzed molecule to the training set compounds and/or model for the given property. In our previous study[7] each participating group adopted its own definition of the distance of a molecule to the model/training set in the context of the respective QSAR methods. Below, we described these definitions in details.

University of North Carolina at Chapel Hill in the United States (UNC): This group used the ensemble of variable selection k Nearest Neighbors (kNN) and Support Vector Machine (SVM) methods,[16] which were applied to descriptors calculated with Dragon[17] and MolconnZ[18] software packages.

The AD for models derived using the kNN approach was calculated from the distribution of similarities between each compound and its k nearest neighbors in the training sets. The similarities were defined as distances between a molecule i and a training set Ω. They were computed as the average Euclidean distance to the k nearest neighbors of this molecule in the training set using only a subset of variables identified by the modeling procedure as optimal[15,19]

$$EU_m = \frac{\sum_{j=1}^{k} d_j}{k} \quad (1)$$

where d_j is the distance of a query compound to its k^{th} nearest neighbor, and m is index of the model.

The distribution of distances (pairwise similarities) between each compound and its k nearest neighbors in the training set

is computed to produce an applicability domain threshold, D_T, calculated for each kNN model as follows:

$$D_T = \bar{y} + Z\sigma \quad (2)$$

Here, $\bar{y}$ is the average Euclidean distance of the k nearest neighbors of each compound within the training set, σ is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level. Typically, the default value of this parameter is set at 0.5, which formally places the boundary for which compounds will be predicted at one-half of the standard deviation (assuming a Boltzmann distribution of distances between each compound and its k nearest neighbors in the training set). Thus, if the distance of the external compound from all of its nearest neighbors in the training set exceeds this threshold, the prediction is considered unreliable.

In total $M = 192$ and $M = 542$ individual models were calculated using MolconnZ and Dragon descriptors, respectively. The average values of the distances to each individual model $m = 1,...,M$

$$EUCLID = \overline{EU}_m \quad (3)$$

was used to estimate a distance of a molecule to the final ensemble of models. Notice, that the minimal value of EUCLID is observed when the training set model was built with $k = 1$. The same definition of DM was also used for models built with the SVM method.

University of Louis Pasteur in France (ULP): This group used kNN, SVM, and Multiple Linear Regression (MLR) methods and fragmental descriptors calculated with ISIDA software.[20−22]

Applicability domains in ISIDA-MLR and ISIDA-kNN models were estimated with an approach similar to that of the UNC with an exception that only one ISIDA-MLR and one ISIDA-kNN model were calculated. Thus there was no averaging over models. For both approaches the distances were calculated using $k = 3$, which was an optimal number of nearest neighbors for the kNN model. The minimal and maximal occurrences of fragments (which were selected by the regression) within compounds in the training set were also retrieved for the ISIDA-MLR model. These values defined an allowed range for each fragment. For a validation compound, the distance to the training set was considered as infinite if one of its fragment descriptors was outside the corresponding range defined for the training set.

University of Insubria in Italy (UI): This group used Ordinary Least Squares regression (OLS) and genetic algorithm to calculate the best linear model using descriptors available within the Dragon[17] software.

Hat values from the leverage matrix representing the "distance" of the molecule to the model structural space were calculated as

$$LEVERAGE = \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x} \quad (4)$$

where $\mathbf{x}$ is the vector of descriptors of a query compound, and $\mathbf{X}$ is the matrix formed with rows corresponding to the descriptors of molecules from the training set. In linear modeling, the leverage, which is frequently notated as h, ranges between $1/N$ and 1, and averages $(K+1)/N$ for the N compounds in the learning data set, where K is the number of model variables. The residual of a compound in the data set has a variance $\sigma^2(1 - h)$, but if an external compound has a leverage h, its prediction error has variance $\sigma^2(1 + h)$. Once the leverage of an external compound gets big, one starts extrapolating outside the range of the learning data and can no longer have much faith that the model itself is valid. The molecules with

$$LEVERAGE = h > 3(K + 1)/N \quad (5)$$

were identified as structurally outlying in the original model as proposed by us earlier.[15]

University of Kalmar in Sweden (UK): This group used Partial Least Squares (PLS) method and Dragon[17] descriptors. Two distances in space of descriptors were used. The first one was LEVERAGE, which was also employed by the UI group. However, since different descriptors were selected in OLS and in PLS models, the nominal DM values in both models were different. The second DM was a distance to the PLS model, PLSEU, which is calculated using the UNSCRAMBLER program as described in its manual or in the book.[23] This distance corresponds to the error in calculation (back-projection) of the vector of input variables from the latent variables and PLS weights.

Virtual Computational Chemistry Laboratory in Germany (VCCLAB): The Associative Neural Networks[24−26] were applied to analyze E-state indices. The ASNN model was based on an ensemble of 100 neural networks. Thus, for any molecule each model calculated one predicted value, i.e. we had 100 predicted values calculated with the ensemble. These 100 values were used to form a vector $\mathbf{Y}_{calc}$. This vector corresponded to a new representation of a molecule in the property-based model space for both training and test set molecules. For each analyzed molecule, i, in the validation set we determined a molecule, z, in the training set with a maximum correlation coefficient

$$CORREL(i,z) = \max R^2(\mathbf{Y}^i\mathbf{Y}^j) \quad (6)$$

between them.[14] R^2 corresponds to 1-Euclidian distance, if $\mathbf{Y}_{cal}$ are normalized to zero mean and unit variance (the normalization does not influence R^2). Thus, this similarity measure corresponded to the minimal Euclidian distance between the validation and training set molecules in the space of activities predicted from models. In the previous study[7] a cut-of value $CORREL > 0.7$ was used to define the applicability domain (AD) of the ASNN model.

Thus, in total four types of distances to models were used in our previous publication.[7] In addition, we have also applied the range of descriptors and that of predicted experimental values to estimate the AD. However, for the current study we have considered only the distance-based approaches and excluded the range-based approaches, which can be difficult to quantify due to the "empty space" problems.[14]

Other DMs. The target property predictions for external set molecules in the UNC and VCCLAB approaches were calculated as an average resulting from the application of the ensembles of models. Several studies[27−30] have indicated that standard deviation of predictions of models correlate with the accuracy of predictions. Thus, we also considered the standard deviation of model predictions

$$STD = \frac{1}{N-1}\sum (Y_{calc} - \bar{Y}_{calc})^2 \quad (7)$$

as an additional metric characterizing the distance of molecules from the ensemble of models.

The Tanimoto index is frequently used in chemoinformatics to measure similarity of molecules. In our study we used the Jaccard/Tanimoto correlation between molecules a and b defined as

$$TANIMOTO(a,b) = \frac{\sum x_{a,i}x_{b,i}}{\sum x_{a,i}x_{a,i} + \sum x_{b,i}x_{b,i} - \sum x_{a,i}x_{b,i}} \quad (8)$$

where $x_{a,i}$ and $x_{b,i}$ are fragment counts, $i=1,...,F$, in each molecule. A maximal value of this index can be considered as a similarity of a validation molecule to the training set. We calculated this index with the ISIDA fragments. The different sets of descriptors were used in the kNN and MLR models, thus contributing two Tanimoto similarities. Both CORREL and TANIMOTO serve to measure similarity between molecules. Their complements to 1, i.e. 1-CORREL and 1-TANIMOTO, were used as DM in our study.

Thus, our study included 14 DMs of 6 different types (EUCLID, LEVERAGE, PLSEU, CORREL, STD and TANIMOTO). The DM was named by combining its type (STD, EUCLID, etc.) and abbreviation of the method (see Table 1) in which the DM was calculated.

Comparison of DMs. An objective analysis requires some statistical tests, which can be used to rank different DMs. Let us assume that the calculated errors, $e_i = Y^i_{exp} - Y^i_{calc}$, $i = 1,...,N$ follow the Gaussian distribution

$$N(0,\sigma^2(e)) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e^2}{2\sigma^2}\right) \quad (9)$$

where σ is a standard deviation of the errors. The simplest hypothesis is to consider that all errors are generated with only "One Gauss" distribution with some σ_0, which can be estimated as the standard deviation of errors in the data set. Under this assumption there is no dependency of the accuracy of prediction of molecules upon any DM. It is also possible that the calculated errors are generated from more than one Gaussian distributions σ_q, $q = 1,...,Q$. In general, we do not know which Gaussian distributions were used to generate each particular error. However, it may happen that some DMs correlate with the σ_q, used to generate the data. In this case molecules with a smaller DM value will have smaller errors and *vice versa* (Figure 1). We can, e.g., bin all errors in several intervals according to the DM values and try to estimate within each bin the parameters of the original Gaussian distributions used to generate the errors. As a result of this analysis we calculate a Mixture of Gaussian Distributions (MGD) σ_g, $g = 1,...,G$, which in the ideal case will help restoring the initial Gaussian distributions, σ_q, $q = 1,...,Q$, used to generate the data (see Figure 1). Since the original Gaussian distributions σ_q, $q = 1,...,Q$ that were used to generate the data are generally not known (except for simulated data), we need some statistical criteria to measure the success of our analysis, i.e. whether the use of a given DM does allow for the discrimination of molecules with small and larger errors. As an alternative hypothesis we can assume that DM does not discriminate against such molecules. Thus the use of the MGD calculated using this DM does not provide any significant advantage in comparison to the assumption that all errors are generated with only "One Gauss" distribution. The statistical tests described in the next sections discriminate between both these situations.

Likelihood Score. The $N(0,\sigma^2(e_i))$ corresponds to a probability that a given error e_i is generated according to the given Gaussian distribution. A probability to observe k errors, $e_1,e_2,...,e_k$, is given by a product

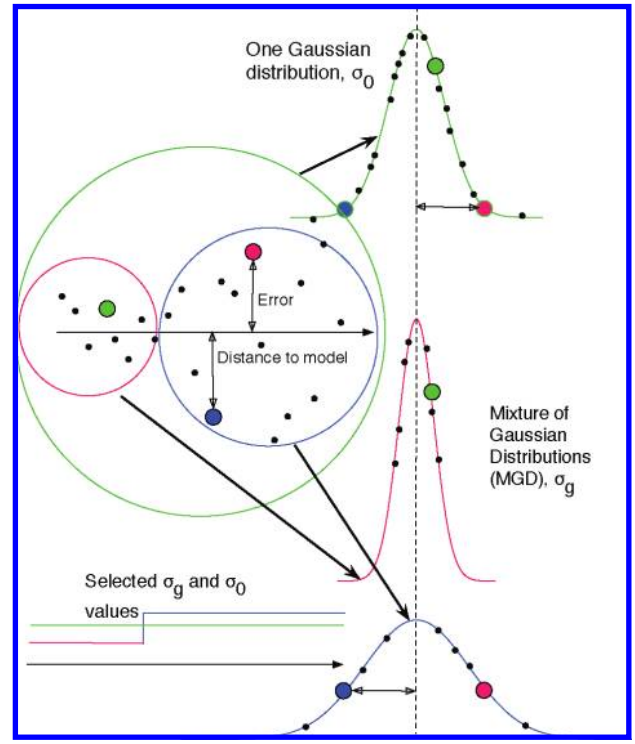

Figure 1. A cartoon illustration of the Mixture of Gaussian distributions. The errors of prediction are ordered according to the distance to models (DM). The molecules with larger values of DM have larger errors on average compared to molecules with smaller DM values. The probability of an error is proportional to the heights of the Gaussian distribution curves shown on the right side of the figure. The Mixture of two Gaussian distributions calculates higher probabilities for three molecules represented by colored circles. A wider distribution implies a higher probability of large errors (blue and red points), while a narrower Gaussian distribution corresponds to higher probabilities of small errors (green points).

$$\Pi N(0,\sigma^2(e_i)) \quad (10)$$

of individual probabilities that is known in statistics as a likelihood function. The maximization of this function corresponds to the calculation of the most probable distribution, which describes the data. From a computational point of view, it is more convenient to work with the log transform of the probabilities. Let us define log score functions

$$S(G_g) = \Sigma \log N(0,\sigma_g^2(e_i))$$
$$S(G_0) = \Sigma \log N(0,\sigma_0^2(e_i)) \quad (11)$$

where G_g and G_0 correspond to the MGD and "One Gauss" distribution, respectively. Here σ_g is selected for each analyzed molecule depending on its DM, while σ_0 is the same for all molecules. The increase of these functions corresponds to the increase of probability that the observed errors are produced with MGD or "One Gauss" distribution, respectively. In case the difference in log scores

$$D(G_g,G_0) = \Sigma \log N(0,\sigma_g^2(e_i)) - \log N(0,\sigma_0^2(e_i))) = S(G_g) - S(G_0) \quad (12)$$

is significantly higher than 0, we can conclude that a use of the MGD provides a better description of calculated errors (and thus the corresponding DM significantly correlates with the errors of molecules) compared to the assumption that all errors are produced with only "One Gauss" distribution. Let us refer to score $S(G_g)$ calculated for the MGD as the

"MGD score" and to score $S(G_0)$ calculated for "One Gauss" distribution as the "One Gauss" score.

Estimation of MGD. To calculate the MGD for a given DM we first ordered all molecules according to the distance. Then we subdivided the data into groups with the same number of molecules, L, which varied in the range $L = 30$, 40, 50,..., $N/2$. For each group we calculated the standard deviation, which was used as σ_g in the MGD. We also required that σ_g were monotonically increasing with the DM. The number L, which minimized the $S(G_g)$ score, was selected as the optimal one.

Estimation of Significance. The bootstrap test with $k = 10000$ replicas was used to estimate whether the score $S(G_g)$ of the MGD distribution was significantly higher than the score $S(G_0)$ of the "One Gauss" distribution. To do it we calculated for each analyzed molecule, $i = 1,..., N$, a difference in scores $dif(e_i) = \log N(0,\sigma_g^2(e_i)) - \log N(0,\sigma_0^2(e_i))$, where σ_g was selected for the molecule according to its DM value. Then we selected with replacement N values from the distribution of all values $dif(e_i)$ by chance and summed them together. The selection was repeated $k = 10000$ times, and the number of runs, C, when the sum was negative, i.e., when the MGD score $S(G_g)$ was smaller or equal to the "One Gauss" score $S(G_0)$, was counted. The p-values reported in Tables S1 and S2 (see the Supporting Information) are the ratios of the counts C to the total number (10000) of replicas.

Cumulative Fraction Plot. Assuming that the errors are generated according to the Gaussian distribution, one can easily estimate the theoretical number of molecules which should have their errors within the given prediction interval.[31,32] For example, 68%, 95%, and 99% of errors should be within σ, 2σ, and 3σ intervals, respectively. Notice that this estimation does not change in case if not one but several Gaussian distributions are used to generate the errors. At least three different plots will be provided for each data set in our article. The first plot, "Optimal", will be a line with identical estimated and theoretical numbers. The second plot, "One Gauss", will be calculated under the assumption that all errors are generated with "One Gauss" distribution. A difference in "One Gauss" and "Optimal" plots may indicate that the errors are generated with several Gaussian distributions (e.g., see Figure 2D,F). The third MGD plot will show whether the use of a mixture of Gaussian distributions and given DM helps in detecting Gaussian distributions used to generate the errors. In the ideal case, when the use of a MGD detects the underlying distributions, both the MGD and the "Optimal" plot will coincide (Figure 2D), while in the worst case, when the analyzed DM is not correlated with errors, the MGD and "One Gauss" plots will be very similar (Figure 2F). Finally, all three plots will coincide for a trivial case, when the data are generated with only one Gaussian distribution (Figure 2B).

EPA High Production Volume (HPV) and EINECS (European Chemical Substances Information System) Data Sets. The HPV Challenge database[33] was downloaded from http://www.epa.gov/ncct/dsstox/sdf_hpvcsi.html. The EINECS data set was downloaded from http://ecb.jrc.it/qsar/information-sources. Composite and metal-containing molecules were filtered out leaving 3182 and 48774 molecules in the HPV and EINECS data sets, respectively. These data were used to demonstrate how our models could be employed for the toxicity predictions against *T. pyriformis* for a diverse set of molecules.

RESULTS

Describing the Errors. The first analysis evaluated whether analyzed DMs correlated with the accuracy of predictions for training and validation sets. To better understand the results of this section let us consider three simulated examples.

Theoretical Examples. In all these examples the position of a data case e_i in the set, i, is used as its DM.

1) The data cases e_i, $i = 1,..., 100$ were generated according to one Gaussian distribution, $N(0,\sigma^2)$, and $\sigma = 1$ (Figure 2A).

2) The data cases were generated with three Gaussian distributions with $\sigma_1 = 0.3$, $i = 1,..., 33$, $\sigma_2 = 1$, $i = 34,..., 66$ and $\sigma_3 = 3$, $i = 67,..., 100$ (Figure 2C).

3) The same data from example 2 were used. However, we randomly shuffled the position of the data cases in the data set (Figure 2E).

No significant MGD (i.e., MGD that has score $S(G_g)$ significantly higher ($p < 0.05$) than "One Gauss" score $S(G_0)$ according to the bootstrap test) could be expected for the first example, since all data cases were generated with one Gaussian distribution. The other two examples had the data generated with three different Gaussian distributions. However, the DM, i.e., position of the sample in the set, was correlated with the Gaussian distributions only in the second example. Thus, we can expect to find significant MGD only for this data set.

The analysis of data for the first example detected a mixture of 2 Gaussian distributions (Figure 2A) with $\sigma_1 = 0.71$ ($i = 1,... 26$) and $\sigma_2 = 1.1$ ($i = 27,... 101$). The difference in σ-values was very small and was a chance effect of data sampling. The MGD score $S(G_g) = 134$ was not significantly different ($p > 0.05$) from the "One Gauss" score $S(G_0) = 135$. Thus, as expected, no significant correlation of DM with errors was detected for these data. The "Optimal", "One Gauss", and MGD plots were all very similar and practically coincided (Figure 2B). Indeed, these plots were originated from the same Gaussian distribution.

The analysis of the data from the second example (Figure 2C) revealed a different picture. First, there was a significant deviation of the "One Gauss" plot compared to the "Optimal" plot (Figure 2D). The "One Gauss" predicted a larger fraction of small errors and a smaller fraction of large errors compared to the theoretical numbers. Thus, estimation of errors with "one Gauss" distribution will predict smaller than observed number of cases with large errors, i.e. will result in a number of outliers. The estimation of a MGD calculated 3 Gaussian distributions (Figure 2C). The MGD score, 141, and the "one Gauss" score, 189, were significantly different at $p < 0.0001$. The fraction plot using MGD coincided with that of the "Optimal" plot. Thus, the use of the MGD correctly predicted the experimental errors, and the position of the molecules in the data set allowed for the discrimination of errors generated with different Gaussian distributions.

The "one Gauss" plot for the third example was exactly the same as in the second example. Indeed, exactly the same data were used to produce it. Only a mixture of two MGD was found (Figure 2E). The MGD score, 186, was nonsig-

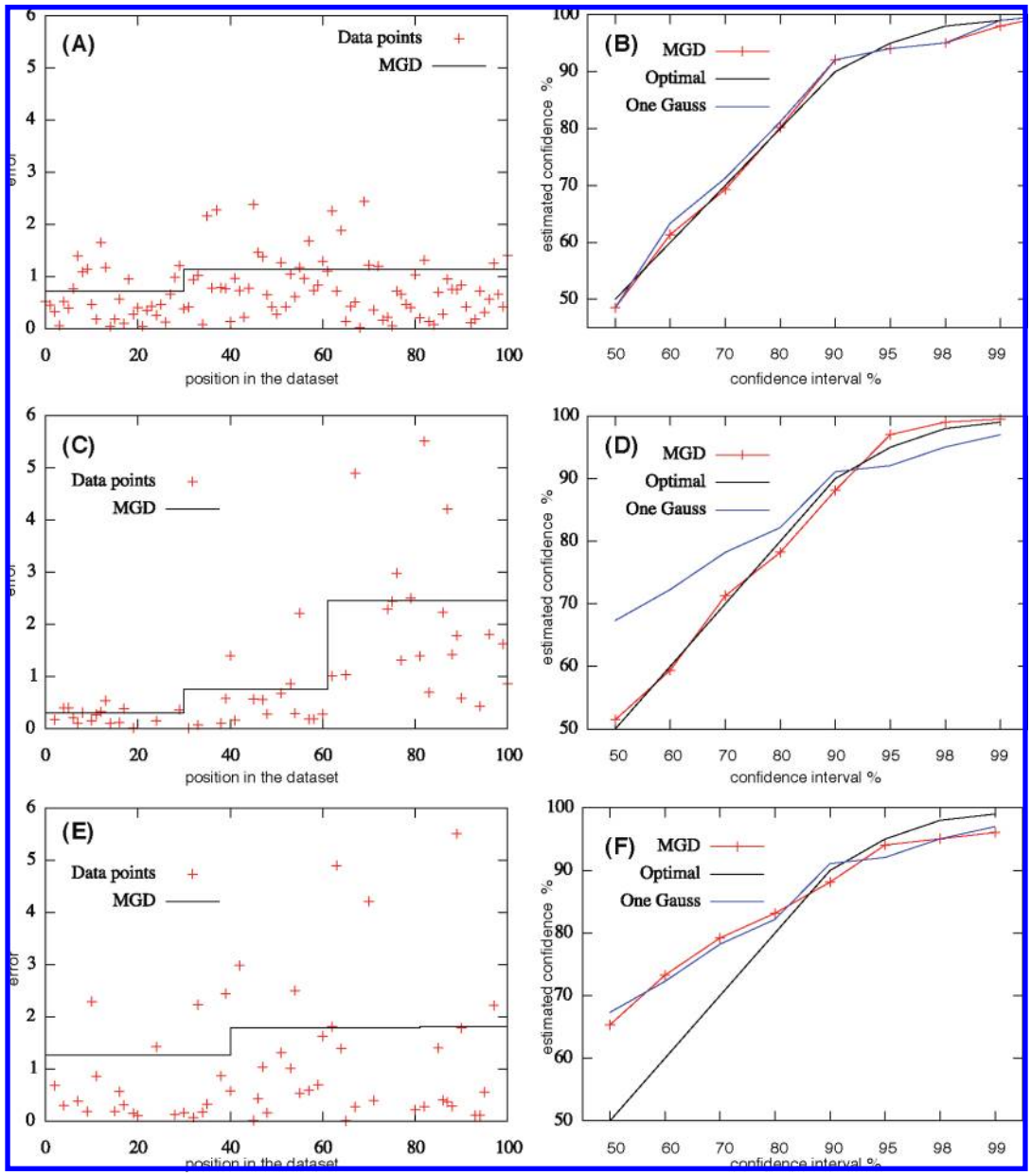

Figure 2. Distribution of data points and calculated MGD for three simulated examples are shown on the left panel. The corresponding fraction plots are shown on the right panel. (A) Data points were generated with one Gaussian distribution. (C,E) Data points were generated with three Gaussian distributions. (E) The data points were shuffled.

nificantly different ($p > 0.05$) from the "one Gauss" score, and fraction plots for both of these distributions practically coincided. Indeed, the shuffling of the positions of the cases in the data set made impossible the discrimination of data points generated with the different Gaussian distributions.

Thus, as it was expected, MGD detected significant dependency between DM, position of the error in the sample, and the errors for the second example only. The use of MGD was of no advantage for errors generated with one Gaussian distribution in the first example (Figure 2A). Indeed, if all observed errors are generated with just one Gaussian distribution, then no DM can discriminate molecules with small and large errors. The "One Gauss" and the "Optimal" plot for such data are practically the same (Figure 2B). Contrary to that, large differences between the "Optimal" and the "One Gauss" plot indicate examples where the use of MGD could be advantageous. However, this is only a necessary condition, since one should also have appropriate DM, which is able to differentiate molecules with low and larger errors. Therefore, although both parts D and F of Figure 2 demonstrated the same deviation, only in the second example there was a correlation of DM and errors. The bootstrap test was important to distinguish significant MGD from those that could be calculated by chance as an effect of data sampling. Notice, that instead of the bootstrap test one can also employ statistical tests to directly compare the fraction plots, using e.g. the Kolmogorov−Smirnov test,[32] and to draw similar conclusions on significant differences. However, we preferred to use the bootstrap test, which in our opinion affords a more simple interpretation.

Analysis of Experimental Data. For this analysis we joined both validation sets in order to improve statistical results. The DM defined with one model can be applied to estimate the accuracy of predictions for any other model. Thus we applied 14 DM to all 12 models, and the results are

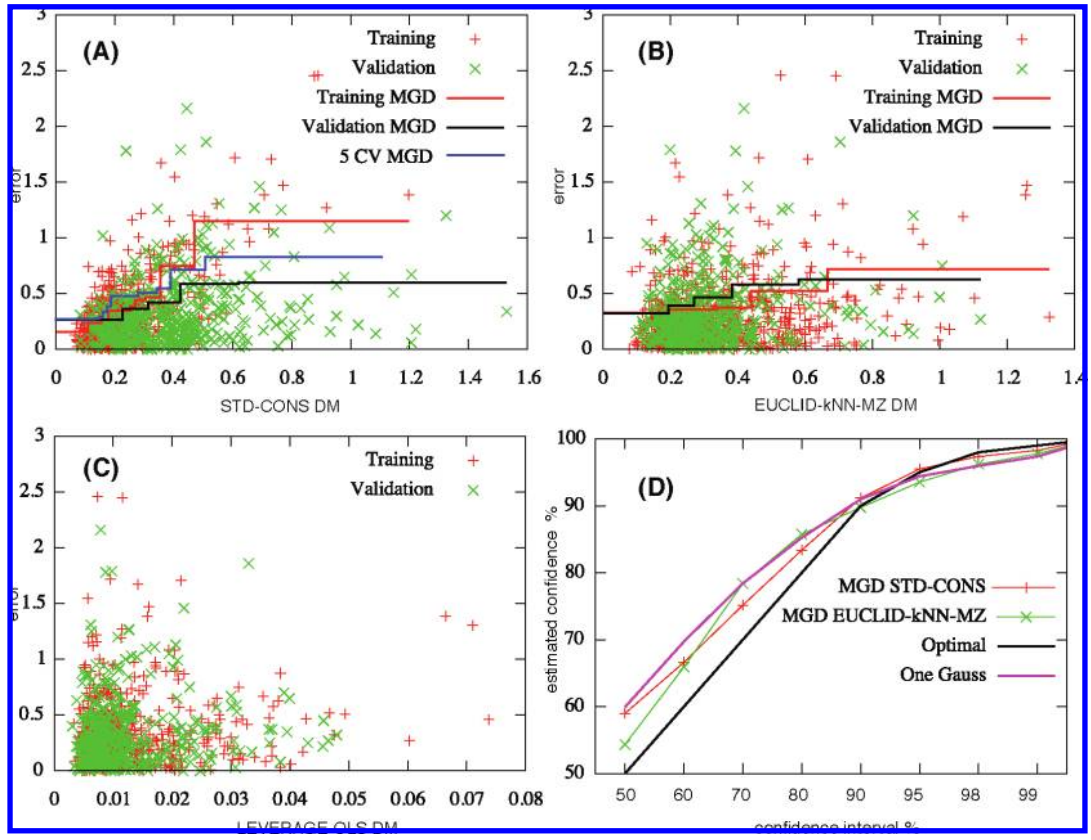

Figure 3. Analysis of the ASSN-ESTATE model. The MGD for the training and joint validation sets are shown for STD-CONS (A) and EUCLID-kNN-MZ (B) DMs. Each horizontal span covers L molecules with close distances. The numbers L were optimized for each DM. The Y-values of the span correspond to the standard deviation of errors of molecules and are used as σ_g for the MGD. The distribution of errors for the LEVERAGE-OLS-DR, which did not calculate significant MGD, is shown in (C). (D) Plots estimated vs theoretical confidence intervals.

summarized in Table S1. The results for all methods and DM were ranked according to their $S(G_g)$ scores.

Examples of Analyses. *ASNN Model.* An example of analysis of results in the ASNN model (Figure 3) demonstrates that "one Gauss" plot was significantly different from the "Optimal" plot. Thus the observed errors were, presumably, generated with several Gaussian distributions. The use of MGD confirmed this result.

The STD-CONS and STD-ASNN calculated the lowest $S(G_g)$ scores for the training and validations sets of the ASNN method. The MGD calculated using STD-CONS DM allowed the best separation of molecules with small and large errors for the training set. For example, molecules from the training set with STD-CONS < 0.19 and STD-CONS > 0.73 had average errors of 0.19 and 0.78 log units, respectively. Thus, the most and least reliably predicted molecules had errors, which differed by a factor of 4. The EUCLID-kNN-MZ distance had a smaller $S(G_g)$ score and provided the worse discrimination of molecules with small and large errors for the same set. The most reliable predictions according to this measure had an average error of 0.31 log units, while the least reliable predictions had an average error of 0.57 log units for EUCLID-kNN-MZ distances <0.23 and >0.75, respectively. Figure 3A,B demonstrates that the ASNN model errors correlated better with the STD-CONS distance and not with the EUCLID-kNN-MZ for the training set (red line). This difference, however, is not so obvious for the validation set (black line on Figure 2A,B), for which both DMs demonstrated similar performances. The fraction plots for the STD-CONS were closer to the "Optimal" plot compared to the EUCLID-kNN-MZ (Figure 3D) thus indicating the higher discrimination of the former DM. The LEVERAGE OLS as well as several other DM (Table S1) did not calculate MGD with a significant score and thus did not discriminate molecules with small and large errors for the training set. This result was also apparent from an absence of apparent correlations between this DM and errors (Figure 3C).

For the joint validation set the minimum score was calculated with the STD-ASNN, which corresponded to the standard deviation of models in the ensemble of neural networks. The scores $S(G_g)$ provided a correct ordering of bootstrap probabilities (Table S1): indeed, by minimizing score $S(G_g)$ we minimized the probabilities implicitly.

OLS Model. The OLS model included six Dragon descriptors

$$\log(\mathrm{IGC}_{50}{}^{-1}) = -18(\pm 0.7) + 0.065(\pm 0.002)\mathrm{AMR} - 0.50(0.04)\mathrm{O\text{-}056} - 0.30(0.03)\mathrm{O\text{-}058} - 0.29(0.02)\mathrm{nHAcc} + 0.046(0.005)\mathrm{H\text{-}046} + 16(0.7)\mathrm{Me}\quad N = 664, R^2 = 0.75, RMSE = 0.53 \quad (13)$$

In our previous work[7] we used LEVERAGE-DM to determine AD of the model. Figure 4 shows that MGD calculated using this DM discriminated molecules with low and large errors. However, the use of CONS-DM provided significantly better results. Indeed, the former DM calculated MGD with $\sigma = 0.5$ and $\sigma = 0.66$ for molecules with the lowest and largest errors from the joint validation set. The MGD calculated with CONS-DM for the same set had minimum $\sigma = 0.36$ and maximum $\sigma = 1.2$, respectively.

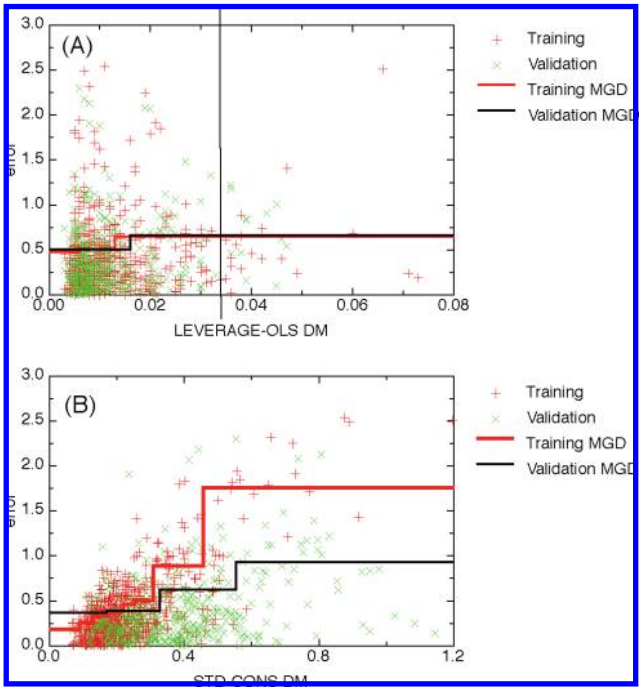

Figure 4. Analysis of the OLS-DR model given by eq 13. The STD-CONS DM provides better discrimination of molecules with low and large errors compared to that of LEVERAGE-OLS DM. The vertical line at panel (A) corresponds to the leverage threshold $3(K+1)/N = 3*7/664 = 0.33$ used to identify outliers in our previous study.[7]

Thus, the second DM better discriminated molecules with reliable and nonreliable predictions. Depending on the purpose of the analysis, the latter metric could be used to identify molecules that are predicted either accurately (e.g., registration within REACH) or inaccurately (e.g., selection of new molecules to extend the model AD). The small discrimination power of the LEVERAGE-DM does not allow performing such a selection efficiently.

Mechanism Based Model. Schultz et al.[13] analyzed a simple model

$$\log(\mathrm{IGC}_{50}{}^{-1}) = 0.545\ \log \mathrm{P} + 16.2\mathrm{A}_{\max} - 5.91$$
$$N = 392,\ R^2 = 0.83,\ RMSE = 0.31 \qquad (14)$$

which was developed using $N = 384$ molecules (8 outlying molecules were excluded). This model is based only on two descriptors, namely the octanol−water partition coefficient (log P) and the Maximum Acceptor Superdelocalizability ($A_{\max}$). This equation predicted molecules from the test set (the second validation set in our study) with $RMSE = 0.54$ log units. The authors pointed out that the distance to the descriptor centroid did not allow them to differentiate molecules with low and large errors.[13] However, the MGD calculated using, e.g. STD-ASNN DM (Figure 5), successfully accomplished this goal for the molecules from both training and validation data sets. Interestingly, five out of eight outlying molecules (benzoyl isothiocyanate, pentafluoronitrobenzene, pentafluorobenzyl alcohol, α,α,α-4-tetrafluoro-o-toluidine, 4-chloro-3,5-dinitrobenzonitrile, 1,5-difluoro-2,4-dinitrobenzene), which were excluded from the original equation, in fact had large STD-ASNN deviations (>0.27) and contributed to the Gaussian distribution with the largest $\sigma = 0.49$. Thus, the low prediction ability of eq 14 for these five outlying molecules could be due to their structural diversity as compared to other molecules in the training set.

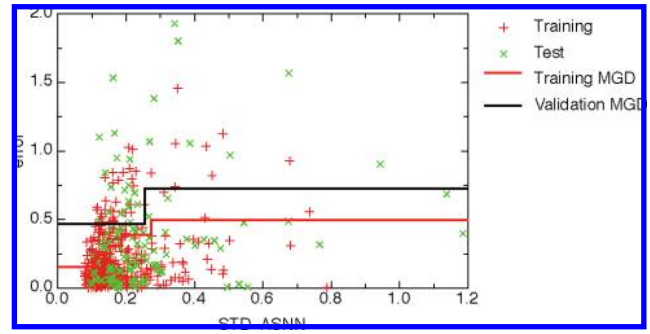

Figure 5. MGD for the model given by eq 14. The use of STD-ASNN DM allowed for the discrimination of molecules with low and large errors in both training and validation sets.

The "Universal" DMs. As mentioned above, all twelve analyzed DMs were applied to estimate MGD and calculate their scores for the training and the joint validation sets. The STD-CONS provided the lowest scores for all twelve and for eleven models (Table S1, Table 3) for the training and the joint validation set, respectively. The STD-ASNN calculated the minimal score for the ASNN-ESTATE model in the joint validation set. When the STD-CONS was excluded from consideration, the STD-ASNN became the best DM for all models from the training set and for ten models from the joint validation set.

To better compare the DMs, we ranked their scores (Table S1). The DM contributing the MGD with the lowest score received rank 1 and so on. The DM with scores nonsignificantly different compared to the $S(G_0)$ score, i.e. DM which failed to differentiate molecules with small and large errors, were not used in the scoring. The averaged ranks of the DM as well as a number of times the DM failed to find significant MGD are summarized in Table 3. The STD-CONS and STD-ASNN, as expected, topped the list. Only these two DMs found significant Gaussian distributions for all analyses. The STD-kNN-DR, which corresponded to the standard deviations in the ensembles of the kNN models, was the third best approach.

Thus, three STD-based models provided the best differentiation of molecules with small and large prediction errors for all analyzed models. This result is surprising since the 14 models considered in this study were developed with different sets of descriptors and machine learning approaches. One may expect that DMs directly developed for the analyzed model should provide better results for it. However, this was not the case, and these DMs appeared as "the universal DMs" for this data set. Since, e.g., STD-CONS always ranked the molecules in the same order, we can conclude that the error of predicting a new molecule did not depend on the descriptors or machine learning method used but on their similarity to the model, i.e., to the training set molecules. Notice that of course different models developed with different descriptors and methods had different performances. However, all these models provided the highest, intermediate, and lowest prediction accuracies for the same groups of molecules on average. Similar results were also reported by Sheridan et al.[34] who observed that accuracy of predictions in their analysis of several QSAR sets practically depended neither on the used set of descriptors nor on the used QSAR method. These results can be related to the "neighborhood behavior" principle asserting that similar molecules in general have similar activities.[35]

Table 3. Performances of MGDs on the Training and on the Joint Validation Sets

	average rank			highest rank[b]			nonsignificant MGD[c]		
DM[a]	LOO	5-CV	valid.	LOO	5-CV	valid.	LOO	5-CV	valid.
STD-CONS	1	1.8	1.1	12	2	11			
STD-ASNN	2	1.2	2.5		10	1			
STD-kNN-DR	6.6	4.3	4.1				2		
STD-kNN-MZ	9.2	8.3	5.3				3		
EUCLID-kNN-DR	7.1	4.9	5.4				3		
LEVERAGE-PLS	8.4	5	6.3				4		
EUCLID-kNN-MZ	7.5	7.1	6.4				3		
TANIMOTO-kNN-FR	7	6.1	6.8				2		
TANIMOTO-MLR-FR	8.3	8.3	9				2		1
CORREL-ASNN	10.7	10.8	9.4				4		1
LEVERAGE-OLS-DR	12.3	12.6	11.1				6		2
EUCLID-MLR-FR	7	9.3	11.5				2		7
PLSEU-PLS	11.1	11.8	11.5				6		7
EUCLID-kNN-FR	12.1	13.3	12.1				10	3	11

[a] The ranks of DM were calculated as follows: each DM (14 in total) was used with each model (12) to calculate MGD. For each model the DM with the lowest MGD $S(G_g)$ score (eq 11) received rank 1. The DM with the second lowest score received rank 2, etc. The average ranks of each DM (over all 12 models) are show in the first three columns. [b] The number of times when the DM provided the MGD with the lowest score. [c] The number of times when no significant MGD was calculated. See also Table S1 (Supporting Information).

Predicting Errors. Our analysis was so far only descriptive. We estimated Gaussian distributions for the training and validation set data separately and then analyzed whether they provided a better description of errors compared to one Gaussian distribution. However, the DM calculated with the training set data could be used to predict errors for molecules from the validation set. To perform this prospective study, we estimated the MGD (i.e., calibrated them) using the training set and then applied the obtained MGD to predict errors of the molecules in the validation sets.

Overfitting of LOO Results by Variable Selection. Before doing this study let us critically examine the results of Table 2. As it was discussed in our previous paper the training and first validation sets were sampled from the same initial data set. Thus, one would expect that LOO results calculated for the training set should provide a reliable estimation of the performance of the analyzed models for the validation set.

However, validation set errors of eight models (including the CONS model) were significantly higher at $p < 0.05$ compared to the results of the same models for the training set. Thus, for these models the LOO results for the training set provided a biased estimation of the performance of the method. However, other methods, e.g. ASNN-ESTATE, kNN-FR, MLR-COD, and OLS-DR, calculated similar errors for both the first validation and the training set as expected.

These significant differences in performance of methods are, in fact, a consequence of implicit differences in the LOO calculation procedures employed by different groups in our first joint publication. Indeed, the main focus of the previous study was to compare all methods according to their blind predictions of both validation sets. However, despite all groups formally using the same LOO procedure, the majority of groups applied the LOO *after the variable selection*, and some groups did not use variable selection at all. Moreover, in some cases the LOO q^2 (UI) was used in the Genetic Algorithm for variable selection. The variable selection procedures resulted in a different degree of overfitting (from no overfitting to a strong one) and provided significant differences in results for the training and first validation sets for some models.

For example, the SVM-FR method calculated the lowest LOO *RMSE* of 0.24 log units for the training set. The *RMSE* of this method for the first validation set was 0.51 log units, i.e. two times bigger. As was indicated in our previous article, the SVM-FR LOO results for this method were calculated *after the variable selection*. The initial set of variables included more than 1000 descriptors, many of which were correlated, and only 109 descriptors were selected for the final model. The use of variable selection for this method resulted in a significant overfitting of the LOO procedure, and its results did not reflect the predictive power of the model. On the other hand, the ASNN approach did not use any variable selection. The LOO *RMSE* = 0.42 log units for this method corresponded to the *RMSE* = 0.44 calculated for the first validation set.

5-Fold Cross-Validation of Data. To overcome the problem of overfitting due to variable selection we performed 5-fold cross-validation with variable selection in each step of the analysis. For each fold, we first selected variables using the corresponding training set, developed the model, and then applied it to predict molecules, which were excluded from the training set. The *RMSE* calculated using the 5-fold CV were nonsignificantly different between the training and the first validation set for all analyzed methods (Table 2). Thus, for all these methods the 5-fold CV *RMSE* provided a correct estimation of the performance of the methods for the validation set, which was generated from the same distribution of molecules. The MGD fitted for the 5-fold cross-validated data again identified STD-ASNN and STD-CONS as the two best DMs (Table 3). However, in this study the STD-ASNN provided the lowest scores in 10 out of 12 analyses and thus became the top-ranked approach. The STD-kNN-DR was the third best DM again.

It is interesting that the analysis of the 5-CV data resulted in just three nonsignificant MGDs, and all of them were for EUCLID-kNN-FR. This number was much smaller compared to 47 nonsignificant MGD calculated with the overfitted LOO data. Thus the overfitting made the errors more unpredictable, i.e. the errors did not correspond to the property of data but rather reflected noise in the data set due to selected variables and the method.

Table 4. Analysis of Validation Set Errors Predicted with the MGDs

		calibrated on 5-CV set				on validation set 1	
		validation set 1		validation set 2			
	rank	*RMSE*[a]	Δerr[b]	*RMSE*	Δerr	*RMSE*	Δerr
STD-ASNN	5	0.53	0.06	0.62	0.05	0.58	0.07
LEVERAGE-PLS	5.7	0.50	0.04	0.54	0.07	0.52	0.09
EUCLID-kNN-MZ	7.9	0.45	0.05	0.51	0.10	0.57	0.06
EUCLID-kNN-DR	8.4	0.45	0.05	0.52	0.09	0.57	0.07
LEVERAGE-OLS-DR	10	0.50	0.04	0.52	0.09	0.51	0.09
TANIMOTO-kNN-FR	10.4	0.50	0.04	0.54	0.07	0.52	0.08
STD-kNN-DR	11.2	0.46	0.05	0.52	0.09	0.56	0.07
TANIMOTO-MLR-FR	11.2	0.51	0.04	0.53	0.07	0.52	0.09
CORREL-ASNN	11.4	0.49	0.04	0.54	0.07	0.53	0.08
STD-CONS	12	0.65	0.16	0.72	0.12	0.54	0.07
EUCLID-MLR-FR	12	0.49	0.04	0.52	0.09	0.52	0.09
PLSEU-PLS	12	0.49	0.04	0.5	0.10	0.5	0.11
STD-kNN-MZ	12	0.48	0.04	0.56	0.05	0.59	0.06
EUCLID-kNN-FR	12	0.50	0.04	0.52	0.09	0.5	0.10
average error[c]		0.49		0.6		0.6	

[a] Average predicted *RMSE* (e.g., using STD-ASNN DM we predicted *RMSE* for all 12 analyzed models and averaged them). [b] Average absolute differences between predicted and actual *RMSE* for all methods (e.g., using STD-ASNN DM we predicted *RMSE* for all 12 models and calculated average absolute difference between predicted and *RMSE* errors for all models). [c] Average *RMSE* of all methods for the given set. See also Table S2.

Use of MGD To Predict Errors. The MGD fitted to the distances and 5-fold cross-validation (5-CV) errors was used to predict the *RMSE* errors for the molecules from the validation sets. An example of MGD calculated using the 5-CV procedure is shown in Figure 3A as a blue line. This MGD mapped the STD-CONS distances to values σ_g. For example, minimal STD-CONS distances in the range of [0,0.15] corresponded to $\sigma_g = 0.25$, while distances larger than 1.1 corresponded to $\sigma_g = 0.80$. These ranges and values σ_g were used to predict errors for molecules from the validation sets. To do this we, first, calculated STD-CONS for each new molecule and, second, estimated its error using the ranges and MGD values σ_g calculated by the 5-CV procedure. Thus for molecule with STD-CONS = 0.1, belonging to the interval [0,0.15], we predicted its average square of the error as $(\sigma_{g(d=0.1)})^2 = 0.25*0.25 = 0.0625$. We made such predictions for all molecules from the validation set, $i = 1,..., M$, and estimated the *RMSE* error for the validation set as

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1,...,M}\sigma^2_{g(d_i)}} \qquad (15)$$

where $\sigma_{g(d_i)}$ is the MGD value corresponding to the STD-CONS distance d_i. The predicted *RMSE* were compared with the calculated values.

Table S2 shows performance of analyzed DMs for all models, while Table 4 summarizes all results similar to Table 3. First, all DMs correctly recognized the higher complexity of validation set 2 and predicted higher errors for this set compared to validation set 1. Thus, all DMs were useful in discriminating data sets of different complexity on the qualitative basis. Second, the STD-ASNN DM calculated the highest number, 6, of significant MGD for validation set 1. The LEVERAGE-PLS and EUCLID-kNN-MZ also calculated significant MGDs for 3 and 1 data set, respectively.

It is interesting that using the STD-CONS MGD we predicted higher errors of models for both validation sets than the calculated ones (compare "*RMSE*" vs "*RMSE* pred" in Table S2). This decreased its $S(G_g)$ scores. The same effect was also observed for the STD-ANN. This DM predicted larger than observed average errors for 11 and 9 out of 12 models for validation sets 1 and 2, respectively. However, the errors estimated with the latter approach were more close to the observed errors. On average the absolute differences between predicted and calculated errors were 0.06−0.05 (set 1−2) and 0.16−0.12 (set 1−2) log units for the STD-ASNN and STD-CONS DM, respectively. Other DMs also tended to predict higher than actually calculated errors of models for validation set 1 but not for validation set 2. Because of the tendency to overestimate errors some STD-ASNN MGDs were not significant.

It was also possible to calibrate the MGD on one of the validation sets. We performed such an analysis and fitted the MGD using results calculated for the first validation set (Table S2). The errors predicted with these MGDs were similar to those calculated for MGDs fitted on 5-fold cross-validation results.

Prediction of Toxicity of Molecules in the HPV and EINECS Databases. The ASNN-ESTATE model and STD-ASNN DM provided one of the most accurate predictions of the toxicity values[7] and estimation of the errors. Therefore we decided to evaluate a performance of this method for prediction of molecules from both industrial databases. For this analysis we redeveloped the ASNN model using all 1093 molecules and calculated $R^2 = 0.86$, $RMSE = 0.39$ using LOO. The 5-fold cross-validation analysis yielded similar results, i.e., $R^2 = 0.85$ and $RMSE = 0.41$.

The MGD calibrated 5-CV data were used to estimate prediction errors for molecules from the training data set and the HPV and the EINECS databases. The distributions of molecules according to the expected errors for both sets are shown in Figure 6. There is a dramatic difference in distributions of the molecules in the training set compared to the both industrial data sets, which have remarkable similarity. While about 24% (67%) of molecules from the training set had predicted errors on the order of experimental

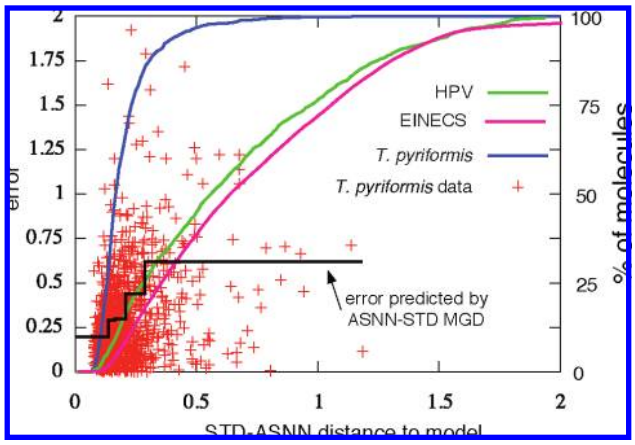

Figure 6. Analysis of molecules from *T. pyriformis* ($N = 1093$ molecules), EINCES ($N = 48774$), and HPV Challenge data sets ($N = 3182$ molecules). The ASNN model errors for the training data set ordered vs STD-ASNN DM are shown as red crosses. MGDs calculated using these data are shown as a black line. The cumulative percentages of molecules as a function of the STD-ASNN distance to model for the HPV (green), EINECS (magenta), and *T. pyriformis* (blue) data sets are also shown. The HPV and EINECS data sets have a much smaller percentage of molecules with short distances (reliably predicted) to the model.

measurements, i.e. <0.21 (<0.38) log units, only 5% (17%) and 2% (10%) of molecules had similar predicted errors in the HPV and EINECS data sets, respectively. Moreover, above 25% of molecules in each industrial data set had STD > 1 log unit, while there were only 2 molecules with such large STD-ASNN distances in our training set. Thus it is possible that the accuracy of prediction for these molecules in the HPV and EINECS data sets could be even lower than predicted by the MGD.

Online Implementation. The models developed by all groups are available for online calculations at the following Web site: http://www.qspr.org. The users can upload data or draw molecules using the JME editor of Peter Ertl. In case molecules are submitted in a 2D format a conversion of 2D=>3D structure is performed using the CORINA program[36] provided by the Molecular Network GmbH. The calculation of descriptors is done with the DRAGON software,[17] the Fragmentor program developed by the ULP,[22] and the E-state descriptor program implemented in the PCLIENT.[26] The server estimates accuracy of prediction using the STD DM.

DISCUSSION

Our results indicate that the standard deviation of models in the ensemble provided the best estimation of the accuracy of predictions of models for the calculation of compound toxicity against *T. pyriformis*. The standard deviation measures the degree of disagreement or divergence of models for a new molecule. The larger is the disagreement, the lower accuracy of predictions is expected for this molecule. The disagreement appears due to structural features that are underrepresented in the training set or do not cover the same range of values in the training and validation sets. The more different is the analyzed molecule to the training set molecules, the higher variation of predictions of models is expected for it.

We have also shown that DM developed with one method and one set of descriptors could be also used to estimate the accuracy of models developed with a different set of descriptors or/and machine learning methods. For example, DM developed with neural networks, STD-ASNN, or k-Nearest Neighbors (STD-kNN-DR), or a consensus model (STD-CONS), in most cases provided better discrimination of molecules with low and large errors for all analyzed models, even if these models were developed with different sets of descriptors and different machine learning methods. Moreover, we have also demonstrated that STD-ASNN DM successfully discriminated molecules with low and large errors for a model based on log P and the Maximum Acceptor Superdelocalizability descriptors.[13] Considering that the distance to the descriptor centroid did not allow the authors to differentiate molecules with low and large errors,[13] our approach can significantly complement methods based on the mechanism of action of molecules by estimating the prediction errors of molecules in such models. This could be particularly useful for the prediction of new scaffolds of molecules, for which determination of the mechanism can be difficult. It is also important to mention that experimental log P values required in eq 14 may not be known for some chemicals. The use of predicted log P values can introduce additional errors. Indeed *RMSE* > 1 were calculated for each of the 18 public and commercial programs benchmarked on more than 96000 molecules.[37] Such large errors could invalidate the model of Schultz[13] when it will be applied to a diverse set of molecules.

Thus, the diversity and distribution of data in the training set but not the computational approaches and descriptors of molecules are the limiting factors determining the accuracy of predictions and applicability domain of the models. This conclusion is in tune with similar results for the prediction of physicochemical parameters, namely lipophilicity[38] and aqueous solubility[4] of molecules, as well as it further confirms similar conclusions of Sheridan and co-workers.[34]

The limited diversity of molecules in the training set naturally limits the applicability of models developed in our study (see Figure 6). Indeed, our estimations suggest that the developed models are able to predict only about 17% of HPV and 10% of EINECS molecules with accuracy comparable to the experimental one. The efforts of the EU REACH program to register about 30000 compounds during the next ten years will challenge models built using data sets of similar chemical diversity. An application of our models to this set is unlikely to cover greater fractions of molecules compared to those we reported for the HPV and EINECS data sets. New experimental measurements of some compounds will be still required. The toxicity of molecules against of *T. pyriformis* has been studied for over twenty years.[8−12] Despite this fact the amount of data, but what is even more important, the chemical diversity of the data set, remains critically low. Moreover, considering results of Sheridan et al.[34] and our studies there is absolutely no reason why any other models developed with this training set and other descriptors will have significantly better accuracy of prediction. Considering that this property is one of the most extensively studied the situation with availability of data for other endpoints can be even worse, and consequently even less predictable models can be expected.

From a practical point of view, the experimental efforts should be focused not on the very detailed analysis of a congeneric set of molecules but on screening as many

different scaffolds as possible. This could help to develop better models with larger applicability domains. The compounds, which would provide the highest improvement in the accuracy of models, could be selected following D-optimal design or similar space-filling design algorithms to cover underrepresented scaffolds of molecules. Multivariate characterization to select compounds, before chemical and biological testing, which differ substantially in the chemical descriptor space, provides an example of such a coherent strategy to ensure diversity in the training set.[39] In other words, a more detailed analysis of a congeneric series is much less valuable than wider screening of different scaffolds. This is particularly true for methods, which are facing the prediction of large and diverse series of molecules, such as HPV, EINECS, or REACH data set.

Another important conclusion of our study concerns overfitting of models by variable selection procedure. Indeed, the statistical QSAR/QSPR models have received a lot of criticism during the last few years, in particular because of the incorrect validation of methods and, as a consequence, the misinterpretation of the results. A very similar problem of overfitting has been addressed by us for the neural network method.[40] For example, some LOO results (Table 2) were overfitted by variable selection. Thus they provided an incorrect estimation of the accuracy of models even for the validation set that was sampled from the same distribution of data, as was discussed in the Results section.

To avoid this problem some authors suggested leaving a part of the data as an external set, which can be used to estimate the performance of the model.[15,41] We used this approach in our previous study. This procedure is well justified for large data sets, like the one we used in our studies. However, Hawkins[42] correctly pointed out that it does not use all available data and thus may result in a lower prediction ability of models compared to those developed with all data. This problem is critically important for small data sets, as it was demonstrated, e.g. for the multivariate modeling by Martens and Dardenne.[43]

The cross-validation with variable selection on each validation fold used in this study (see also other papers[21,42,44]) actually reused the whole data set for the external validation. Indeed, since all optimizations were performed inside of the validation step, this procedure does perform ***a blind external prediction*** of the validation subset on each validation fold. Thus, it does not at all contradict the idea of using the external set[15,41] but extends it to the whole training set. In the end one can develop a model using the whole data, and this model (since it is developed with more data than used in the cross-validation procedure) will perform at least as accurate as estimated using the *n*-fold cross-validation procedure.

Of course, the cross-validation will result in the selection of different variables for each validation set and in different models. This is not a drawback but a considerable advantage of this procedure, since the cross-validation does ***estimate the impact of the variable selection on the final model***. Neglecting the variable selection could result in a serious overfitting as shown in Table 2 and discussed above. Considering the importance of accurate toxicity prediction of environmental chemicals, or in fact any biologically significant property in general, we summarize the data modeling procedure advocated in this paper as follows.

1) Model development. Develop your model using your favorite method(s) and all available data. When this study is completed, estimate the accuracy of your final model as follows.

2) Model validation.

a. Divide your initial set on *n*-subsets (e.g., $n = 5$ was used in this study, larger *n* or LOO can be recommended for small data sets).

b. Select one subset as the validation set.

c. Use the remaining $n-1$ sets to develop a model using ***exactly the same approach*** as in step 1.

d. Apply the model to the validation set and store the predictions.

e. Go to step b) and repeat the analysis until all subsets are used as the validation sets.

f. Estimate the performance of your model using values calculated from step d).

This procedure estimates the expected accuracy of the model, which was built using all data. The validation should be performed only once. Multiple runs of this procedure with an attempt to improve the cross-validation results will again lead to the overfitting. Nonsatisfactory results will mean that there are not enough data or/and the data are not accurate enough and/or used descriptors are not adequate to model the analyzed property. Instead of *n*-fold cross-validation one can also use bootstrap aggregation (bagging) procedures.[45] Like with *n*-fold cross-validation the variable selection should be applied on each bootstrap run. The prediction of data not used for the variable selection (out-of-bag samples) provides an unbiased estimation of the method performance.[46]

Abbreviations: DM - distance to a model; LOO - Leave-One-Out; *RMSE* - Root Mean Squared Error; STD - DM calculated as standard deviation of models in the ensemble; *MAE* - Mean Absolute Error; "One Gauss" - Gaussian distribution parametrized with σ equal to the standard deviation of all errors in the analyzed data set; "One Gauss" plot - fraction plot calculated under the assumption that all errors are generated with "One Gauss"; "One Gauss" score - "One Gauss" score $S(G_0)$ calculated using eq 11; "Optimal" plot - diagonal line corresponding to equal theoretical and estimated confidence intervals; MGD - Mixture of Gaussian Distributions; MGD score - MGD score $S(G_g)$ calculated using eq 11; significant MGD - MGD that has an MGD score significantly higher ($p < 0.05$) than a "One Gauss" score according to the bootstrap test (see the Methods section); "MGD" plot - fraction plot estimated using MGD.

ACKNOWLEDGMENT

This study was supported in part by the Go-Bio BMBF grant 0313883 (VCCLAB) and the NIH RoadMap grant GM076059 and the EPA STAR grant RD832720 (UNC). The authors would like to thank Prof. J. Gasteiger and Molecular Networks GmbH for providing access to their CORINA program at the Web site http://www.qspr.org and Prof. T. W. Schultz for many years of research on chemical toxicity in *T. pyriformis* that made this study possible.

Supporting Information Available: MGD scores and probabilities for training and validation sets as well as predicted *RMSE* for the validation sets (Tables S1 and S2).

This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Tetko, I. V. Prediction of physicochemical properties. In *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*; Ekins, S., Ed.; John Wiley & Sons, Inc.: NJ, 2007; Vol. 1, pp 241−275.

(2) Karickhoff, S. W.; Carreira, L. A.; Melton, C.; McDaniel, V. K.; Vellino, A. N.; Nute, D. E. In *Computer Prediction of Chemical Reactivity - The Ultimate SAR*; C. f. E. R. I., Ed.; Environmental Research Brief EPA/600/M-89/017; U.S. Environmental Protection Agency: Cincinnati, OH, 1989.

(3) Pontolillo, J.; Eganhouse, R. P. The search for reliable aqueous solubility (Sw) and octanol-water partition coefficient (Kow) data for hydrophobic organic compounds: DDT and DDE as a case study. In *Investigations*; U. S. G. S. W.-R., Ed.; Reston, VA, 2001; p 55.

(4) Balakin, K. V.; Savchuk, N. P.; Tetko, I. V. In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: trends, problems and solutions. *Curr. Med. Chem.* **2006**, *13*, 223–241.

(5) Bassan, A.; Worth, A. P., Computational Tools for Regulatory Needs. In *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals*; Ekins, S., Ed.; John Wiley & Sons, Inc.: NJ, 2007; Vol. 1, pp 751−775..

(6) Netzeva, T. I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M. T.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D.; Schultz, T.; Stanton, D. W.; van de Sandt, J. J.; Tong, W.; Veith, G.; Yang, C. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* **2005**, *33*, 155–173.

(7) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR Modeling of Chemical Toxicants Tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **2008**, *48*, 766–784.

(8) Schultz, T. W.; Netzeva, T. I.; Cronin, M. T. Evaluation of QSARs for ecotoxicity: a method for assigning quality and confidence. *SAR QSAR Environ. Res.* **2004**, *15*, 385–397.

(9) Cronin, M. T. D.; Schultz, T. W. Development of quantitative structure-activity relationships for the toxicity of aromatic compounds to *Tetrahymena pyriformis*: Comparative assessment of the methodologies. *Chem. Res. Toxicol.* **2001**, *14*, 1284–1295.

(10) Dimitrov, S. D.; Mekenyan, O. G.; Sinks, G. D.; Schultz, T. W. Global modeling of narcotic chemicals: ciliate and fish toxicity. *J. Mol. Struct. (THEOCHEM)* **2003**, *622*, 63–70.

(11) Seward, J. R.; Sinks, G. D.; Schultz, T. W. Reproducibility of toxicity across mode of toxic action in the Tetrahymena population growth impairment assay. *Aquat. Toxicol.* **2001**, *53*, 33–47.

(12) Aptula, A. O.; Roberts, D. W.; Cronin, M. T.; Schultz, T. W. Chemistry-toxicity relationships for the effects of di- and trihydroxybenzenes to *Tetrahymena pyriformis*. *Chem. Res. Toxicol.* **2005**, *18*, 844–854.

(13) Schultz, T. W.; Hewitt, M.; Netzeva, T. I.; Cronin, M. T. D. Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. *QSAR Comb. Sci.* **2007**, *26*, 238–254.

(14) Tetko, I. V.; Bruneau, P.; Mewes, H. W.; Rohrer, D. C.; Poda, G. I. Can we estimate the accuracy of ADME-Tox predictions. *Drug Discovery Today* **2006**, *11*, 700–707.

(15) 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. *QSAR Comb. Sci.* **2003**, *22*, 69–77.

(16) Shen, M.; Xiao, Y.; Golbraikh, A.; Gombar, V. K.; Tropsha, A. Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J. Med. Chem.* **2003**, *46*, 3013–3020.

(17) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; WILEY-VCH: Weinheim, 2000; p 667.

(18) Molconn-Z. http://www.edusoft-lc.com/molconn (accessed Jun 10, 2008).

(19) 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*, 241–53.

(20) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693–703.

(21) Tetko, I. V.; Solov'ev, V. P.; Antonov, A. V.; Yao, X.; Doucet, J. P.; Fan, B.; Hoonakker, F.; Fourches, D.; Jost, P.; Lachiche, N.; Varnek, A. Benchmarking of linear and nonlinear approaches for quantitative structure-property relationship studies of metal complexation with ionophores. *J. Chem. Inf. Model.* **2006**, *46*, 808–819.

(22) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptor. Curr. Comput.-Aided Drug Des. 2008, in press.

(23) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and megavariate data analysis: Principles and applications*. Umetrics: Umeå, 2001; p 425.

(24) Tetko, I. V. Associative neural network. *Neural Process. Lett.* **2002**, *16*, 187–199.

(25) Tetko, I. V. Neural network studies. 4. Introduction to associative neural networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717–728.

(26) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory - design and description. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453–463.

(27) Geman, S.; Bienenstock, E.; Doursat, R. Neural networks and the bias-variance dilemma. *Neural. Comput.* **1992**, *4*, 1–58.

(28) Tetko, I. V.; Luik, A. I.; Poda, G. I. Applications of neural networks in structure-activity relationships of a small number of molecules. *J. Med. Chem.* **1993**, *36*, 811–814.

(29) Beck, B.; Breindl, A.; Clark, T. QM/NN QSPR models with error estimation: Vapor pressure and *log P*. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1046–1051.

(30) Chalk, A. J.; Beck, B.; Clark, T. A quantum mechanical/neural net model for boiling points with error estimation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 457–462.

(31) Schwaighofer, A.; Schroeter, T.; Mika, S.; Laub, J.; ter Laak, A.; Sulzle, D.; Ganzer, U.; Heinrich, N.; Muller, K. R. Accurate solubility prediction with error bars for electrolytes: a machine learning approach. *J. Chem. Inf. Model.* **2007**, *47*, 407–424.

(32) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. Numerical Recipes in C++. The Art of Scientific Computing; 2nd ed.; Cambridge, 2002; p 1002.

(33) Wolf, M. A.; Burch, J.; Martin, M.; Richard, A. M. DSSTox EPA High Production Volume Challenge Program Structure-Index Locator File: SDF File and Documentation. Updated version: HPVCSI_v2c_3548_15Feb2008. http://www.epa.gov/ncct/dsstox (accessed Jun 10, 2008).

(34) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.

(35) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.

(36) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.

(37) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of Molecular Lipophilicity: State of the Art and Comparison of Log P Methods on More Than 96,000 Compounds. J. Pharm. Sci. 2008, in press.

(38) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407–1421.

(39) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; 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.

(40) Tetko, I. V.; Livingstone, D. J.; Luik, A. I. Neural network studies. 1. Comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826–833.

(41) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694–701.

(42) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.

(43) Martens, H. A.; Dardenne, P. Validation and verification of regression in small data sets. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 99–121.

(44) Wegner, J. K.; Zell, A. Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077–1084.

(45) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140.

(46) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.

CI800151M