Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set

Iurii Sushko,[†] Sergii Novotarskyi,[†] Robert Körner,[†] Anil Kumar Pandey,[†] Artem Cherkasov,[‡] Jiazhong Li,[§] Paola Gramatica,[§] Katja Hansen,[||] Timon Schroeter,[||,⊥] Klaus-Robert Müller,[||] Lili Xi,[#] Huanxiang Liu,[▽] Xiaojun Yao,[#] Tomas Öberg,[○] Farhad Hormozdiari,[◆], Phuong Dao[◆] Cenk Sahinalp,[◆] Roberto Todeschini,[¶] Pavel Polishchuk,[+] Anatoliy Artemenko,[+] Victor Kuz'min,[+] Todd M. Martin,[%] Douglas M. Young,[%] Denis Fourches,[□] Eugene Muratov,[+,□] Alexander Tropsha,[□] Igor Baskin,[■] Dragos Horvath,[●] Gilles Marcou,[●] Christophe Muller,[●] Alexander Varnek,[●] Volodymyr V. Prokopenko,[△] and Igor V. Tetko*,[†]

Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum Muenchen—German Research Center for Environmental Health (GmbH), Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany, University of British Columbia, Vancouver Prostate Centre, 2660 Oak str., Vancouver, BC, V6H 3Z6, Canada, QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria, Via Dunant 3, Varese 21100, Italy, Machine Learning Department, Technical University of Berlin, Franklinstr. 28/29, 10587 Berlin, Germany, Bayer Schering Pharma AG, Nonclinical Drug Safety, Müllerstr. 178, 13353 Berlin, Germany, Department of Chemistry, Lanzhou University, Tianshui South Road 222, Lanzhou 730000, China, School of Pharmacy, Lanzhou University, Lanzhou 730000, China, School of Natural Sciences, Linnæus University, 391 82 Kalmar, Sweden, School of Computing Science, Simon Fraser University, Burnaby, British Columbia, Canada, Milano Chemometrics & QSAR Research Group, Dept. Environmental Sciences, University of Milano—Bicocca, 20126 Milan, Italy, A.V. Bogatsky Physico-Chemical Institute of National Academy of Science of Ukraine, Lustdorfskaya doroga 86, Odessa 65080, Ukraine, Clean Processes Branch, National Risk Management Research Laboratory, United States Environmental Protection Agency, Cincinnati, Ohio 45268, United States, Laboratory for Molecular Modeling, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, United States, Department of Chemistry, Moscow State University, 119991, Moscow, Russia, Laboratoire d'Infochimie, UMR 7177 CNRS, Université de Strasbourg, 4 rue B. Pascal, Strasbourg 67000, France, and Institute of Bioorganic & Petrochemistry, Ukrainian Academy of Sciences, Murmanskaya 1, 02660 Kyiv-94, Ukraine

Received July 12, 2010

The estimation of accuracy and applicability of QSAR and QSPR models for biological and physicochemical properties represents a critical problem. The developed parameter of "distance to model" (DM) is defined as a metric of similarity between the training and test set compounds that have been subjected to QSAR/QSPR modeling. In our previous work, we demonstrated the utility and optimal performance of DM metrics that have been based on the standard deviation within an ensemble of QSAR models. The current study applies such analysis to 30 QSAR models for the Ames mutagenicity data set that were previously reported within the 2009 QSAR challenge. We demonstrate that the DMs based on an ensemble (consensus) model provide systematically better performance than other DMs. The presented approach identifies 30−60% of compounds having an accuracy of prediction similar to the interlaboratory accuracy of the Ames test, which is estimated to be 90%. Thus, the *in silico* predictions can be used to halve the cost of experimental measurements by providing a similar prediction accuracy. The developed model has been made publicly available at http://ochem.eu/models/1.

INTRODUCTION

Any QSAR/QSPR prediction of biological and/or physicochemical properties has limited value without an estimated applicability domain of a model. Researchers cannot make much use of a prediction for a particular compound if there is no information available on whether this prediction is reliable or not, in other words, whether the given model is applicable. Currently, this problem is being addressed by ongoing studies of applicability domain (AD) assessment.

The conventional methods for estimating model performance are the root-mean-square error (RMSE) and the Pearson correlation coefficient (R^2) of cross-validation. These measures are easily computable and interpretable. However, in general, some groups of chemical compounds can be predicted well, whereas others allow only low prediction accuracy. Thus, depending on the data composition, one can

* Corresponding author tel./fax: +49-89-3187-3575, e-mail: itetko@vcclab.org.
[†] Helmholtz Zentrum Muenchen—German Research Center for Environmental Health (GmbH).
[‡] University of British Columbia.
[§] University of Insubria.
[||] Technical University of Berlin.
[⊥] Bayer Schering Pharma AG.
[#] Department of Chemistry, Lanzhou University.
[▽] School of Pharmacy, Lanzhou University.
[○] Linnæus University.
[◆] Simon Fraser University.
[¶] University of Milano—Bicocca.
[+] A.V. Bogatsky Physico-Chemical Institute of National Academy of Science of Ukraine.
[%] United States Environmental Protection Agency.
[□] University of North Carolina at Chapel Hill.
[■] Moscow State University.
[●] Université de Strasbourg.
[△] Ukrainian Academy of Sciences.

10.1021/ci100253r © 2010 American Chemical Society
Published on Web 10/29/2010

observe significant differences in the estimated and observed statistical parameters.

In particular, when assessing QSAR model performance, one should not only ensure that the predicted accuracies for the training and testing sets are comparable and high but also that the distribution of descriptors' values is uniform within the sets. Under this assumption, the statistical parameters for the new data should indeed be similar to the estimated average values. However, average values can provide biased results if external data distributed differently compared to the modeling set.

Moreover, the number of experimentally available observation points is usually in the range of hundreds (complex biological properties, such as ADMETox data) to hundreds of thousands of measurements (physicochemical properties or HTS data). These numbers are dramatically smaller than the number of compounds for which estimation of properties is needed, e.g. 2×10^7 commercially available molecules or 10^{20} to 10^{24} synthetically accessible molecules[1] or even 10^{80} to 10^{100} theoretically existing chemical structures. Thus, the scenario when QSAR/QSPR predictive models are intended for chemical structures that are different from the training/testing set molecules is a rule rather than an exception.

Thus, the goal of the AD approaches is to estimate the prediction accuracy for each modeled compound individually. Using this information, one can estimate the accuracy of prediction for an arbitrary data set regardless of its similarity to the set used to validate the model.

QSAR studies can assess the accuracy of predictions in different ways. The simple ones try to distinguish reliable vs. nonreliable predictions. They usually assume that the accuracy of prediction of molecules, which are inside a space of descriptors covered by the training set, is similar to the estimated accuracy of the model. These methods include:

• Descriptor boxes: consider compounds with descriptors, lying in predefined parallelepipeds in multidimensional descriptor space, as being inside of the applicability domain of the model.[2–4]

• Leverage-based: all compounds, whose leverage (known as the Mahalanobis distance) with the training set exceeds some predefined limit, are considered to be outside the applicability domain.[5,6]

Approaches that are more sophisticated directly assess the accuracy prediction of each compound, instead of "inside AD/outside AD" information:

• approaches that evaluate the probability distribution of predictions rather than giving point estimates[3,7]

• empirical approaches based on the "distance to model" concept.

The latter approaches are most commonly used in QSAR modeling[8] and represent the subject of the current study. The "distance to a model" (DM) stands for a numerical measure, which monotonically increases as the accuracy of the model decreases.[8] The AD can be defined on the basis of DM; namely, all compounds that have DM values less than a predefined threshold are considered to be inside the AD. The threshold for the DM is chosen to ensure necessary prediction accuracy for compounds within the AD. For predefined prediction accuracy, DMs covering large numbers of molecules are preferred. Leverage, mentioned before, can be used as a distance to the model. In our analysis, we did not fix a "warning leverage" threshold but, rather, investigated the prediction accuracy for all leverage values.

The accuracy of a model can be specified in terms of RMSE, MAE, classification rate, etc. among others. It is worthwhile to distinguish DMs, based solely on descriptor values, from those that use models' predictions, so-called DMs in the property space. To some extent, this terminology may be confusing, since both types of measures solely rely on the structural information. The DMs in the property space can be, of course, applied to new molecules for which experimental values are not known. Both these measures explore disagreements between models developed with different subsets of the initial training data set. To some extent, the DMs in the property space use descriptors that are normalized according to the target property, while the descriptor space DMs ignore this information (e.g., all descriptors are normalized and contribute equally in the LEVERAGE measure). Thus, if some descriptors are more relevant for a given property, they will have higher impact on the DMs in the property space and *vice versa*. As it has been shown,[2] the DMs in the property space yield higher quality AD assessments compared to the DMs in the descriptor space. We confirmed this observation in our previous study[8] and demonstrated that DMs based on the standard deviation of predictions of the model ensemble outperformed descriptor-based DMs such as leverage.

This and several other studies were used for the analysis of classification models which differ from regression-based modeling by the discrete nature of the target (output) labels, which are commonly selected as "−1" (inactive) and "+1" (active; sometimes "0" and "1" are used instead). Interestingly, most machine learning methods, such as neural networks, KNN, or linear regressions, yield continuous predictions. These quantitative values are frequently used to assess classification accuracy, with values close to "−1" and "+1" considered as more reliable predictions than those that are near 0.[9,10] For example, Manallack et al.[9] showed that the classification accuracy of molecules on soluble and insoluble compounds dramatically increased when only molecules with values close to "−1" and "+1" were considered.

In our previous study we introduced a new DM, STD-PROB, which combined measures used by Manallack et al.[9] with the standard deviation of predictions. The latter measure was the best DM criterion for quantitative models.[8]

In the current study, we extend our benchmarking analysis to 30 classification models developed within the 2009 Ames mutagenicity challenge.[11]

METHODS

Data Sets. *The Data Set of the Ames Test Measurements.* The Ames mutagenicity data set[12] described in our previous article[11] was used in the current benchmarking study. The Ames test relies on the determination of the mutagenic effect of a given compound on histidine-dependent strains of *Salmonella typhimurium*. Thus, the measurable mutagenic ability of a compound may signal its potential carcinogenicity.[13] The Ames test can be used with different bacteria strains and can be performed with or without metabolic activation using liver cells. For this study, all such diverse data were pooled together as described in ref 12. According

Table 1. Summary of the Analyzed QSAR Models[a]

model name	descriptors used	training method	numeric predictions	DM provided
CONS			+	
EPA_2D_FDA	2D		+	
EPA_2D_NN	2D	NN	+	
LNU_Drag_PLS	Dragon	PLS	+	
MSU_FRAG_LR	Fragments	Linear regression	+	
MSU_FRAG_SVM	Fragments	SVM	+	SVM1 AD
OCHEM_ESTATE_ANN	E-State indices	ASNN	+	
PCI_Drag_RF	Dragon	Random forest	+	
PCI_SiRMS.Drag_RF	SiRMS+Dragon	Random forest	+	
PCI_SiRMS_RF	SiRMS	Random forest	+	
TUB_3DDrag_RF	Dragon	Random forest		DA Index
TUB_3DDrag_SVM	Dragon	SVM		DA Index
UBC_ID_IWNN	Inductive descriptors	IWNN		
UBC_ID_NN	Inductive descriptors	NN		
UI_Drag_KNN	Dragon	KNN		
UI_Drag_LDA	Dragon	LDA		
ULP_ISIDA_NB	ISIDA Fragments	Naïve Bayes	+	Trust level
ULP_ISIDA_SQS	ISIDA Fragments	Stochastic QSAR sampler	+	Trust level
ULP_ISIDA_SVM	ISIDA Fragments	SVM	+	Trust level
ULP_ISIDA_VP	ISIDA Fragments	Voted Perceptron	+	Trust level
ULZ_3DDrag_KNN	Dragon	KNN		
ULZ_3DDrag_SVM	Dragon	SVM		
UMB_Drag_DT	Dragon	Decision Tree		
UNC_Drag_KNN	Dragon	KNN		
UNC_Drag_RF	Dragon	Random forest	+	
UNC_Drag_SVM	Dragon	SVM	+	AD Mean
UNC_SiRMS.Drag_RF	SiRMS+Dragon	Random Forest	+	
UNC_SiRMS.Drag_SVM	SiRMS+Dragon	SVM	+	AD Mean
UNC_SiRMS_RF	SiRMS	Random forest	+	
UNC_SiRMS_SVM	SiRMS	SVM	+	AD Mean

[a] There were 30 models including the consensus model. The continuous numeric prediction values were available for 20 models.

to that approach, a molecule can be considered as active if it demonstrates mutagenic activity for at least one strain. Thus, considering that the benchmark set molecules were tested with different strains, there may be a significant variance in results. Moreover, different authors used different thresholds to decide whether a given molecule is active or not. As shown in the Results and Discussion section, we estimated the intra- and interlaboratory accuracies of measurements in the Ames mutagenicity data set to be 94% and 90%, respectively.

The initial data set was randomly divided into training and external test sets. The training set contained 4361 compounds, including 2344 (54%) mutagens and 2017 (46%) nonmutagens. The external test set contained 2181 compounds (1/3 of initial set) including 1172 (54%) mutagens and 1009 (46%) nonmutagens. These data sets were used for the 2009 Ames mutagenicity challenge, where the external test set was given to the participants for "blind predictions".[11]

The Data Sets of Chemical Compounds. To investigate the performance of the QSAR models on the Ames test, we have estimated the prediction accuracy for three external data sets: ENAMINE, EINECS, and HPV. The ENAMINE data set contains over *287 000* drug-like chemicals synthesized in 2009 by the Enamine company (http://www.enamine.net). The HPV (high production volume) data set contains chemicals produced or imported into the United States in quantities over 1 million pounds per year. After filtering out composite substances, stereoisomers, and metals from the HPV data set, 2356 compounds were used for analysis. The EINECS (European Chemical Substances Information System) data set was downloaded from http://ecb.jrc.it/qsar/information-sources and contained 68 779 compounds.

Analyzed Models. Twelve international teams submitted 29 models to the 2009 Ames mutagenicity challenge (the models are summarized in Table 1). All of the models were evaluated according to a 5-fold cross-validation procedure as described in the work by Tetko et al.[8] Additionally, each group developed their models using the whole training set, and these models were "blindly" applied to predict test compounds. The resulting consensus model (CONS) was calculated by averaging the predictions of all 29 individual models. The complete information on descriptors, methods, and specific details about each approach can be found elsewhere,[11] while below we will briefly describe the utilized methodologies.

University of Insubria (UI). Linear discriminant analysis (LDA) was used to develop the UI_Drag_LDA model. The LDA calculates a hyperplane, which subdivides the *n*-dimensional descriptor space into two regions corresponding to analyzed classes of compounds. The model was based on 454 Dragon descriptors, which were selected from a total pool of 2032 descriptors after removing constant and highly correlated ($r > 0.9$) descriptors.

Technical University of Berlin (TUB). The Random Forest model (TUB_3DDrag_RF) was a collection of 50 decision trees where each tree depended on a set of randomly selected descriptors.[14] In comparison to the original work of Breimann,[14] all samples were used to build trees (no bagging). The TUB_3DDrag_SVM model was developed using the libsvm[15] implementation with the radial basis kernel. Both of the models were based on 957 3D Dragon descriptors, which were reduced to 872 by removing the descriptors with constant and missing values.

Lanzhou University (LZU). All of the molecules were converted to 3D structures and optimized using MM+ molecular mechanics with semiempirical PM3 partial charges implemented in the HyperChem program (HyperChem for Windows—Molecular Modeling System, Hypercube, Inc., Gainesville, Florida). The Dragon software[16] was used to calculate 1664 molecular descriptors for each molecule. After deleting the descriptors with constant or highly correlated ($r > 0.95$) values, 716 descriptors remained. Support vector machine—recursive feature elimination (LZU_3DDrag_SVM model)[17] was employed to select calculated descriptors and perform classification of the new molecules as described elsewhere.[11] Another model was calculated using the k nearest neighbors method (LZU_3DDrag_KNN).

Linnæus University (LNU). Partial least-squares discriminant analysis (PLS-DA) was used, which is an extension of PLS regression for classification.[18,19] The initial set of descriptors contained 929 2D Dragon descriptors. After removal of 103 constant variables, 826 remained. Nonsignificant descriptors were further removed using a jack-knife method for significance testing of the PLS procedure.[20] Finally, 82 descriptors were used to develop the LNU_Drag_PLS model.

Helmholtz Zentrum Muenchen, Online CHEmical Modeling Environment (OCHEM). The associative neural network method[21,22] was applied using an ensemble of 50 neural networks. Each neural network had three hidden neurons. Both atom- and bond-type 2D E-state indices[23] (362 descriptors) were used for the structure representation. The filtering of highly correlated $r > 0.95$ indices and singletons (found only in a single molecule) left 233 descriptors, which were used to develop the OCHEM_ESTATE_ANN model.

University of British Columbia (UBC). "Inductive" descriptor IND_I[24–26] and MOE QSAR parameters (http://www.chemcomp.com) were used to quantify the structures of the studied compounds. The *in house* SVL scripts were used to calculate IND_I descriptors from 3D structures of molecules optimized with the MOE MMFF molecular force field. All correlated descriptors ($r > 0.9$) were eliminated, and the most relevant QSAR descriptors (15 and 3 descriptors for the IWNN model and NN models, respectively) were selected according to the Information Gain criteria[27] using Weka software (v. 3.5.8).[28] The weighted nearest neighbor (UBC_ID_NN) and iterative weighted nearest neighbor (UBC_ID_IWNN) models were created as described elsewhere.[11]

Laboratory of Chemoinformatics, Institute of Chemistry, Louis Pasteur University, Strasbourg, France (ULP). Two classes of substructural molecular fragments, "sequences" (I) and "augmented atoms" (II), were used.[29] The ULP_ISIDA_NB model was developed using naive Bayesian approach. The ISIDA/VotedPerceptron ULP_ISIDA_VP model implemented a simple perceptron algorithm re-expressed in terms of the Tanimoto kernel. For nonlinearly separable cases, all perceptrons were combined in a voting pool. The weighted vote was done according to the accuracy of perceptrons for the training set. The ULP_ISIDA_SVM model used libSVM with the Tanimoto similarity coefficient as a kernel. ULP_ISIDA_SQS was created using the stochastic QSAR sampler (SQS) algorithm, which is a genetic algorithm-driven regression tool supporting nonlinear descriptor transformations.[30]

Moscow State University (MSU). The ν-modification support vector machines method[31] and regularized logistic regression implemented in the package LIBLINEAR[32] were used to develop the MSU_FRAG_SVM and MSU_FRAG_LR models, respectively. Optimal values of algorithm parameters were found using the grid search and the cross-validation procedure. Both of the models used the same set of 19 603 fragmental descriptors[33,34] with the size of each fragment up to five non-hydrogen atoms, which were computed using the NASAWIN software.[34] No descriptor selection procedures have been applied.

Physico-Chemical Institute of NAS of Ukraine (PCI). The Simplex representation of molecular structure (SiRMS)[35] was used to calculate 21 378 2D Simplex descriptors (number of tetra-atomic fragments with fixed composition and topological structure).[35,36] In addition, 2D Dragon descriptors (943) were used separately and in combination with Simplex descriptors. The Random Forest (RF)[14] method was employed for obtaining models.[37] The final models have been selected by the highest out-of-bag statistic values. The PCI group contributed three models, based on SiRMS descriptors (PCI_SiRMS_RF), 2D Dragon descriptors (PCI_Drag_RF), and a combination of both (PCI_SiRMS.Drag_RF).

University Milano—Bicocca (UMB). A total of 2489 molecular descriptors[16] were calculated using the Dragon software.[38] Constant and nearly constant descriptors were removed, leading to a final number of 1601 retained descriptors. The CART (Classification and Regression Trees) algorithm, a binary tree classification method,[39] was used to develop the decision trees UMB_Drag_DT model. The final classification tree included 29 descriptors.

University of North Carolina (UNC). The WinSVM program implementing the open-source libSVM package[40] was employed to build and select mutagenicity models. An ensemble of 467 Dragon descriptors calculated for two-dimensional hydrogen-depleted structures, 609 two-dimensional SiRMS descriptors, and a combined set of Dragon/SiRMS descriptors were used as inputs to build the UNC_DRAG_SVM, UNC_SiRMS_SVM, and UNC_SiRMS.DRAG_SVM models respectively.

United States Environmental Protection Agency (EPA). A total of 790 2D descriptors[41] were used. The EPA_2D_FDA model (FDA, Food Drug Administration) was built according to a methodology developed by Contrera et al.[42] For each test chemical, 30−75 of the most similar chemicals from the training set in terms of the cosine similarity coefficient were selected. Then, a local linear regression model was built to predict the test compound. In the EPA_2D_NN model, the three closest chemicals in the training set in terms of the cosine similarity coefficient were selected for the test compound. The predicted mutagenicity was simply the class, which dominated for the three chemicals.

Additional details of models and a detailed description of models and results can be found elsewhere.[11]

Preprocessing of Results. Some models provided prediction as {0,1} while other models provided it as {−1,+1} for mutagenic and nonmutagenic compounds, respectively. In order to be consistent, we converted all predictions to the {−1,+1} values. After this processing, for some models, i.e., neural networks or SVM, there were predictions outside of the [−1,+1] interval. We did not normalize or round these

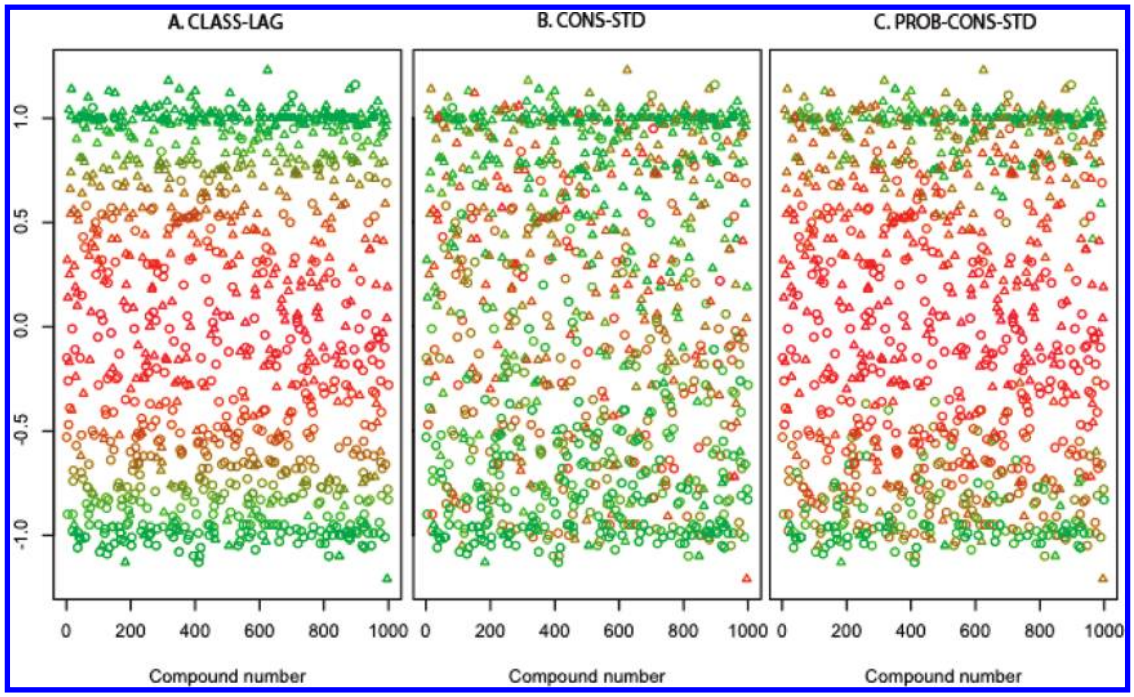

Figure 1. Test set predictions of the OCHEM_ESTATE_ANN model. Three DMs (CLASS-LAG, CONS-STD, and PROB-CONS-STD) are encoded by color. Green represents low values of the corresponding DM; red represents high values. Triangles are mutagens, and circles are nonmutagens according to the Ames mutagenicity test. Values outside the [−1,+1] interval appear due to a specific normalization for neural network training (value −1 corresponded to 0.1, and +1 corresponded to 0.9).

values to [−1,+1]; instead, we used the original values for the calculation of DM as described below.

Numeric prediction values were available only for 20 models (including the consensus model). As some of the investigated DMs require numeric prediction values, only these 20 QSAR models were used in the current study. Several DMs that could be used only with qualitative predictions were applied to all 30 models.

Distance to Model and Applicability Domain. Let us designate any numeric measure calculated solely on the basis of chemical structures or prediction values and which increases with a decrease in the reliability of classification as "distance to model" (DM). Then, on the basis of a model performance, we can identify a threshold for the DM that provides a predefined accuracy of classification. All data set entries with DM values below the threshold form a model's "applicability domain" (AD). Criteria for the performance of distances to the model are suggested in the section below.

Most DMs investigated in this article are developed on the basis of those used previously for regression problems[2,8,43] and were introduced in our preliminary study.

Let us introduce notation to represent predictive modeling entities: J, a compound to be predicted; $y(J)$, a continuous prediction value, calculated by the model; $c(J)$, the predicted class for the given compound J, identified by:

$$c(J) = \begin{Bmatrix} 1, & y(J) > 0 \\ -1, & y(J) \leq 0 \end{Bmatrix} \quad (1)$$

We will designate DM for a compound J as $d(J)$.

CLASS-LAG. For the binary classification problem, labels for the predictive model are discrete and are selected in our study as −1 and +1. However, most machine learning methods give a quantitative number as a result of prediction. The absolute value of the difference between the prediction value and the nearest of the labels can be used as a measure of prediction uncertainty. This measure, referred to as CLASS-LAG, is calculated according to

$$d_{\mathrm{CLASS-LAG}}(J) = \min\{|-1 - y(J)|, |1 - y(J)|\} \quad (2)$$

CLASS-LAG can be interpreted as the amount of rounding to the nearest class label; the more rounding that is required, the less reliable the prediction is expected to be. Thus, the measure punishes deviations from target class values {−1,+1}, both positive and negative deviations (i.e., both 1.2 and 0.8 predicted values have the same DM). Obviously, punishing negative deviations applies only to models that have prediction values outside of the [−1,+1] interval; there were only three models with such predictions: EPA_2D_FDA, LNU_Drag_PLS and OCHEM_ESTATE_ANN.

Figure 1A illustrates the simplicity of this idea: green dots, which are closer to the edge of the class, are predicted to have better prediction accuracy than red dots, located in the "uncertainty area" between the classes, near a value of 0. In this figure, triangles are positive (mutagens) and circles are negative (nonmutagens) predictions. The classes are more mixed together near zero line. The continuous values of predictions may not always be available: some machine learning methods provide only discrete {−1,+1} outputs. In this case, CLASS-LAG is always equal to zero and obviously cannot be used. This DM is the most obvious one, and it was used, e.g., by Mannalack et al.[9]

STD. The standard deviation of the predictions, obtained from an ensemble of models, can be used as an estimator of model uncertainty for a given compound. The general idea is that if different models yield significantly different predictions for a particular compound, then the prediction for this compound is more likely to be unreliable. The sample standard deviation can be used as an estimator of model uncertainty.

Let us assume that $Y(J) = \{y_i(J), i = 1-N\}$ is a set of predictions for a compound J given by a set of N trained models. The corresponding distance to model (STD) is calculated by

$$d_{\mathrm{ASNN-STD}}(J) = \mathrm{stdev}(Y(J)) = \sqrt{\frac{\sum (y_i(J) - \bar{y})^2}{N-1}} \tag{3}$$

This DM has been proven to provide excellent results for the discrimination of highly accurate predictions in the case of regression models.[2,8,9] In the given study, we investigate two variations of the STD measure that differ in the contents of the used models: (i) ASNN-STD, based on predictions of a neural network ensemble of OCHEM_ESTATE_ANN, and (ii) CONS-STD, based on predictions of several models that were built using different machine learning methods and different parameters (and including OCHEM_ESTATE_ANN as one of the models).[2] Although it is possible to calculate STD for virtually any model, i.e., by replicating multiple models of the same method using the bagging technique[44] and computing the standard deviation of predictions, in this study, STD values were available only for OCHEM_ESTATE_ANN.

In our study, we used two variations of this measure: CONS-STD uses quantitative values of predictions to calculate standard deviation, and CONS-STD-QUAL uses qualitative (discretized) values. The rationale for using CONS-STD-QUAL lies in the unavailability of quantitative values for some machine learning methods.

Applicability of the standard deviation to classification tasks follows from the property of the Bernoulli-distribution, which is used for characterizing the distribution of random binary values. The standard deviation of the Bernoulli distribution rises as the probability for each class approaches 0.5, which corresponds to the most uncertain prediction. Hence, both the normal (used in regression tasks) and Bernoulli distributions (used in classification tasks) follow the same law—the prediction uncertainty rises as the standard deviation rises.

In Figure 1B, built on the basis of the OCHEM_ESTATE_ANN model, the green dots denote the highest level of agreement between 20 individual models, used for CONS-STD. These points correspond to low values of standard deviation. The red points, on the contrary, show that the individual models yielded quite a wide range of predictions; so the standard deviation for these points is relatively high. In this figure, we observe that red and green points are mixed, which means the STD measure does not depend on the value of prediction and can provide information that is complementary to CLASS-LAG.

STD-PROB. This DM, suggested in a recent study,[45] combines the two previously mentioned measures into a single value to improve the estimation of prediction accuracy. Having obtained a prediction $p(x)$ for a given compound, we replace this point prediction with a distribution of probabilities. In other words, instead of giving a point prediction, we provide a probabilistic one. We assume the mentioned distribution is Gaussian with a mean $p(x)$ and standard deviation that correspond to its STD value. The suggested distance to model is

$$d_{\mathrm{STD-PROB}}(J) = \min \begin{cases} \mathrm{probability}(c > 0 | N(y(J), d_{\mathrm{STD}}(J))) \\ \mathrm{probability}(c < 0 | N(y(J), d_{\mathrm{STD}}(J))) \end{cases} \tag{4}$$

namely,

$$d_{\mathrm{STD-PROB}}(J) = \min \begin{matrix} \int_0^{+\infty} N(x, y(J), d_{\mathrm{STD}}(J))\, \mathrm{d}x \\ \int_{-\infty}^{0} N(x, y(J), d_{\mathrm{STD}}(J))\, \mathrm{d}x \end{matrix} \tag{5}$$

where $N(x,y(J),d_{\mathrm{STD}}(J))$ is the normal distribution density function with mean $y(J)$ and standard deviation $d_{\mathrm{STD}}(J)$. Here, $y(J)$ is an actual prediction of the analyzed model for a compound J and $d_{\mathrm{STD}}(J)$ is an STD-based distance to model (ASNN-STD or CONS-STD), calculated according to eq 3.

This measure can be graphically illustrated as the square of the area under the curve of the normal distribution density function.

Four examples are given in Figure 2, where the rounded prediction value is always fixed to "+1"; however, the quantitative prediction values and STD values are different. It is obvious that shifting the curve away from the center (decreasing CLASS-LAG) results in a decrease of the filled area. The same effect appears when we make the curve less flat, i.e., decrease the STD value. Thus, STD-PROB combines information about uncertainty from both measures: CLASS-LAG and STD.

STD-PROB has an easy interpretation: values close to 0.5 indicate an equal probability of finding the given compound in either class; i.e., the model cannot provide reliable prediction. On the contrary, values close to 0 indicate a high probability of finding the compound in one of the classes.

We analyze two variations of STD-PROB, ASNN-STD-PROB and CONS-STD-PROB, which correspond to the ASNN-STD and CONS-STD measures, respectively.

Similarly to the depiction of previously introduced measures, green dots in Figure 1C denote compounds whose CONS-STD-PROB value, i.e., minimal area under the probability density chart on the intervals $(-\infty; 0]$ and $[0; +\infty)$, is relatively high. The square of this area is computed using eq 4.

Importantly, the STD-PROB measure is an empirical one. This approach proved to work successfully in our previous study using ensembles of neural networks,[45] and it is applied here to analyze the results produced by other machine learning methods.

CONCORDANCE. This measure shows whether a prediction of an individual model is concordant with predictions of other models within the ensemble. More accurately, CONCORDANCE is the number of models that give the same prediction that the current model does:

$$\mathrm{CONCORDANCE}(J) = \sum_{i=1}^{N} \mathrm{eq}(y(J), y_i(J)) \tag{6}$$

where $y(J)$ and $y_i(J)$ are predictions of compound J, given by the target model and the members of the ensemble, N is the size of the ensemble, and eq is equality indicator (equal to 1 if the arguments are equal and to 0 otherwise).

CORREL. This measure is based on the correlation of vectors of the ensemble's predictions for the target compound and compounds from the training set. More precisely, the

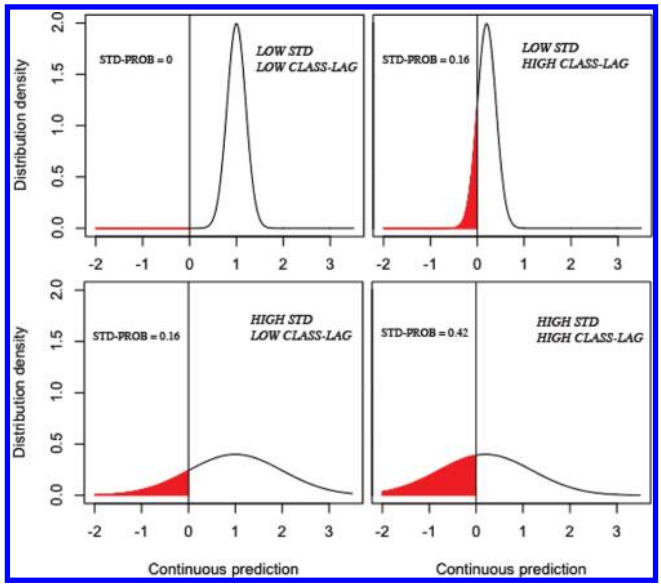

Figure 2. STD-PROB is the square of the filled area on each of the four charts. The charts show how CLASS-LAG and STD affect STD-PROB. STD corresponds to the flatness of the curve, and CLASS-LAG corresponds to the shift of the curve from the center. Larger values of STD correspond to flatter curves and larger STD-PROB values. As CLASS-LAG decreases, the curve shifts more from the center and the STD-PROB value decreases.

CORREL measure for the target compound J is calculated according to the following expression:

$$\mathrm{CORREL}(J) = 1 - \max_{i=1-M} |\mathrm{corr}(\overrightarrow{y(T_i)}, \overrightarrow{y(J)})| \qquad (7)$$

where $\vec{y}(T_i)$ and $\vec{y}(J)$ are vectors of the ensemble's predictions for the training set compound T_i and the target compound J, and corr designates the Spearman rank correlation coefficient between the two vectors, and M is the number of compounds in the training set. The low value of CORREL (i.e., high Spearman correlation coefficient) indicates that for target compound J there is a compound T_k from the training set for which predictions of the ensemble of models are strongly correlated. Indeed, if a compound T_k has the same descriptors as J, then the predictions of the models will be identical for both molecules, and thus CORREL(J) = 0. The performance of this measure for regression models is discussed elsewhere.[8,46]

LEVERAGE. Leverage is a descriptor-based DM; i.e., it is based only on model input but not on output, in contrast to CLASS-LAG, STD, and STD-PROB. LEVERAGE is a special case of Mahalanobis distance, calculated according to expression 8:

$$\mathrm{LEVERAGE}(J) = x(X^T X)^{-1} x^T \qquad (8)$$

where x is a vector of descriptors for compound J and X is the matrix of descriptors for the training set. The LEVERAGE values were available only for the OCHEM_ESTATE_ANN model and were based on E-State indices.

DA-Index. The applicability domain employed by the TUB group is based on the κ, γ, and δ indices introduced by Harmeling et al.[47] The first two indices are heuristics that have been previously used in the chemoinformatics community: κ is the distance (here in this section and below, Euclidian distance calculated using descriptors is assumed) to the k-nearest neighbor, and γ is the mean distance to the k nearest neighbors. The last index, δ, corresponds to the length of the mean vector (i.e., a mean of vectors) to the k nearest neighbors. Since κ and γ are only based on distances, they do not explicitly indicate whether interpolation or extrapolation is expected for prediction. δ allows making this distinction and indicates the degree of extrapolation. Input descriptors for all indexes were weighted following the development of the Gaussian process classification model.[48] The arithmetic mean values of γ and δ indices were used to estimate prediction confidence. A threshold value determined using the training set was used to decide whether a test compound was inside or outside the AD. The output of this decision process was called DA-Index.

AD_MEAN. AD_MEAN values were provided by the UNC group for SVM models that were developed using three sets of descriptors (SiRMS, Dragon, and combined). AD_MEAN corresponds to the average Euclidean distances between a compound and its three nearest neighbors in the training set. All distances are calculated using the entire pool of descriptors. AD_MEAN was available for two models, UNC_SiRMS_SVM and UNC_Drag_RF; therefore, we investigated two respective measures, AD_MEAN1 and AD_MEAN2.

ELLIPS. ELLIPS values were calculated using the EPA_2D_FDA model. A prediction is within the applicability domain of the model 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 ($h_{\max}$) for all of the compounds used in the model.[49] The ratio $h_{00}/h_{\max}$ was used as a distance to the model, referred to as ELLIPS.

SCAvg (Average Similarity Coefficient). The cosine similarity coefficient to the three nearest neighbors used in the EPA_2D_NN method was used as the SCAvg DM.

Two groups classified predicted molecules in several classes with different qualities of prediction as described below.

Trust Level. The applicability domain for the models, provided by the ULP group, is based on a measure, referred to as the *trust score*. This measure has values in the range of {1,2, ...5}, where the "5" corresponds to the highest trust level ("optimal") and 1 is the lowest trust level ("none"). The trust score for a particular compound is based on three factors: (i) the number of models having the compound in their local applicability domain, MINDIFF-OK, as described in ref 50, (ii) the number of dissident predictors in the set (i.e., models that gave predictions, different from the mean prediction), and (iii) the average prediction value, where values close to 0.5 are considered less reliable. Further details on the calculation of the trust score are shown in Figure SF1 (Supporting Information).

SVM1 AD. The applicability domain for the MSU models was computed using the one-class classification approach (novelty detection) based on 1-SVM.[51] The parameters of the 1-SVM method were chosen as follows: the RBF-kernel parameter γ was taken from the same value used for building classification SVM models, while the value of ν was fixed at 0.01.

The SVM1 AD procedure associates the applicability domain of QSAR/QSPR models with the area in the input

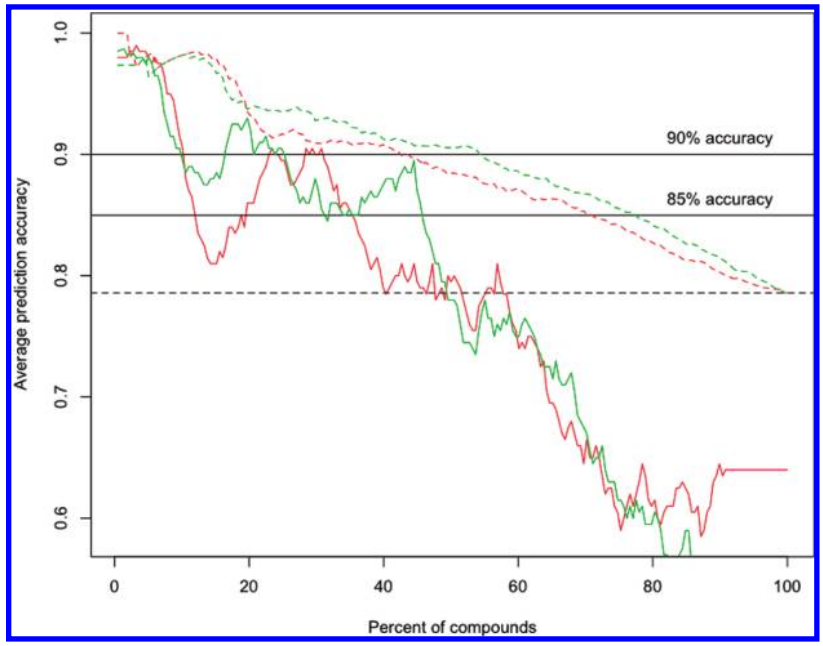

Figure 3. Prediction accuracy of the consensus model as a function of CONS-STD and CONS-STD-PROB. The solid lines (bin-based averaging) show the averaged accuracy on a moving window with a size of 200 compounds. Although there is a trend that the accuracy of prediction decreases with both DMs, the dependency is not smooth, and there are significant fluctuations. The dashed lines (cumulative averaging) indicate the average prediction accuracy for a variable percentage of compounds. Cumulative averaging smooths the variations, which makes it more suitable for the threshold-based comparison of DMs.

descriptor space where the density of training data points exceeds a certain threshold. The main assumption of this procedure is that the predictive performance of the models tends to be higher for the test compounds inside the high density areas than for those that are outside. This could take place since outside the high density area all test objects are located far from training objects, which makes interpolation of the properties from the training to test objects unreliable. Instead of searching a decision surface separating high and low density areas in the input space, the one-class classification 1-SVM approach looks for a hyperplane in the feature space associated with the RBF-kernel.

The ability of novelty detection models to be used as the AD of machine learning models was earlier demonstrated by Bishop.[52] The use of a one-class SVM novelty detection method to assess the applicability domain of models based on structured graph kernels has recently been suggested by Fechner et al.[53]

Benchmarking Criteria. To compare the performances of different DMs, it is necessary to assess their ability to separate predictions with low and high accuracy. Our approach is to determine the percentage of compounds in the training and test sets that are predicted with a DM-defined accuracy. For a particular DM, there are two possible ways to separate compounds, predicted with a given accuracy:

Bin-Based Accuracy Averaging (BBA). BBA groups the compounds, sorted by a particular DM, into bins having an equal number of compounds, averages the accuracy in the bins, and selects a DM threshold, which provides predefined model accuracy for every bin within this threshold. However, this criterion has some drawbacks. First, it does not take into account the actual prediction accuracy as long as it is higher than the threshold. Second, the detection of a DM threshold in practice can be a subjective task and will depend on the size of the bin. For example, when predictions for different models were sorted according to DMs, and their accuracies were averaged using a sliding window of, e.g., 200 molecules, we could observe a significant variation in predictions as a function of the DMs when using one defined threshold (see solid lines in Figure 3).

Integral Accuracy Averaging (IA). Instead of bin-based averaging, one can use the average accuracy of a model for molecules with a DM less than a predefined threshold value. The plots of average predictions of models for a DM less than the predefined threshold are smoother and easier to interpret: i.e., this threshold defines the average (accumulative) accuracy of the model. Moreover, this criterion directly corresponds to, e.g., the average accuracy of inter- or intralaboratory measurements. Therefore, for all further analyses, we used the integral criterion and compared the DMs with respect to their accumulative average accuracy. A threshold of 90% was used. More precisely, we did the following steps to estimate the performance of the investigated DMs:

• For the training and test sets, sort all of the compounds according to DM.

• For each model, identify the largest DM value for which the accumulative accuracy of compounds from the analyzed set (training or test) is ≥ 90% (DM_90%).

• For each model, calculate the percentage of compounds with a DM less than the *respective* DM_90% threshold for the training set (referred to as $C_{\text{TRAIN-90\%}}$, C stands for coverage) and the test set ($C_{\text{TEST-90\%}}$). Notice that thresholds are selected separately for the training and test sets.

Values $C_{\text{TRAIN-90\%}}$ and $C_{\text{TEST-90\%}}$ are used to estimate performances of the DMs for each analyzed model. Indeed, for a given model, the larger $C_{\text{TRAIN-90\%}}$ and $C_{\text{TEST-90\%}}$ values correspond to DMs with larger numbers of reliable predictions. Similar to our previous study,[8] we ranked DMs according to their $C_{\text{TRAIN-90\%}}$ and $C_{\text{TEST-90\%}}$ values (i.e., the DM with the highest $C_{\text{TRAIN-90\%}}$ or $C_{\text{TEST-90}}$ receives a rank

of "1" and so on) and averaged the ranks over all models. These averaged ranks were used to compare different DMs.

Under prediction accuracy, we understand

$$\text{accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{total number of compounds}} \times 100 \quad (9)$$

where true positives, true negatives, and the total number of compounds are within a DM threshold. In addition to the prediction accuracy, the sensitivity and the specificity are frequently used in machine learning methods. These measures are particularly useful for nonbalanced data sets. The Ames data set has a very small imbalance of active and nonactive compounds; therefore, specificity and sensitivity are to a large extent redundant and were not analyzed in this study.

To verify whether there are significant differences between analyzed DMs, we used the Wilcoxon signed-rank test[54] applied to $C_{\text{TRAIN-90\%}}$ and $C_{\text{TEST-90\%}}$ values. This test is used for two-sample designs involving repeated measures, matched pairs, which is the case in our study.

As a graphical illustration of the DM performance, we used cumulative accuracy−coverage plots (see, e.g., dashed lines in Figure 3). On these charts, we plotted prediction accuracy for a group of compounds, having a DM less than some threshold (y axis), against a percentage of this group of compounds in the whole set (x axis). The threshold for DM is not directly present in the chart but is implicitly represented by the x axis.

Additionally, we intended to confirm whether a particular DM can not only separate high and low accuracy predictions but also estimate the external accuracy of prediction. For this purpose, we compare prediction accuracies for compounds within the *same* DM threshold on training and test sets.

There are two drawbacks to the aforementioned accuracy coverage ($C_{\text{TRAIN-90\%}}$ and $C_{\text{TEST-90\%}}$) as an estimator of the DM performance. First, the coverage depends on the accuracy threshold, and different thresholds could possibly result in different rankings of the analyzed DMs. Second, the accuracy coverage depends not only on the ability of DM to separate highly accurate predictions but also on the performance of the analyzed model. Indeed, the models having higher prediction accuracies will probably have higher accuracy coverages.

The AUC (Area under the Curve) Criterion. Another criterion for DM performance that does not have the aforementioned drawbacks is the area under the curve (AUC) parameter, calculated as the area of the square between the bin-based averaging curve and the line of the average model performance. In Figure 3, this is the area of the square between one of the solid lines and the dashed horizontal line. The AUC is higher for the DMs that provide better separation of compounds with higher and lower accuracies compared to the average accuracy of models. Similarly to the accuracy coverage, the weighted accuracy spread can be calculated for both the training set ($\text{AUC}_{\text{TRAIN}}$) and the test set ($\text{AUC}_{\text{TEST}}$).

To rank the investigated DMs, we used both criteria: the accuracy coverage ($C_{\text{TRAIN-90\%}}$ and $C_{\text{TEST-90\%}}$) and the area under the curve ($\text{AUC}_{\text{TRAIN}}$ and AUC_{TEST}).

Table 2. Average Ranks of the DMs Ranked by the Percentages of Compounds with 90% Accuracy[a]

distance to model	average rank (cTRAIN 90%)	average rank (cTEST 90%)
CONS-STD-QUAL-PROB	2.15	1.83
CONCORDANCE	1.65	2.15
CONS-STD-PROB	3.38	2.95
CONS-STD-QUAL	3.7	4.95
ASNN-STD-PROB	6.4	5.48
CONS-STD	4.88	5.75
CLASS-LAG	7.5	6.68
ASNN-STD	8.4	7.78
ELLIPS	9.15	8.98
AD_MEAN1	12.43	10.18
CORREL	10.35	11.65
SCAvg	11.08	11.85
AD_MEAN2	11.3	12.33
LEVERAGE	12.65	12.48

[a] The ranks for both the training and validation sets are shown.

Comparison of Models. The most commonly used measure of model performance is its accuracy on the test set. This measure, however, does not reveal what is the maximum possible performance of a particular model. For this reason, a percentage of compounds that are predicted with a fixed accuracy level (90% in our example) can be identified and used for model ranking.

RESULTS AND DISCUSSION

Comparison of Distances to Model. The calculated $c_{\text{TRAIN-90\%}}$ and $c_{\text{TEST-90\%}}$ values are summarized in Table 2, where DMs are sorted accordingly to their rank on the basis of $c_{\text{TEST-90\%}}$ values (see Table S1 of the Supporting Information for more details). The data demonstrate that the CONS-STD-QUAL-PROB measure appeared to be the best one, considering averaged ranks over all models on the test set. Details for the calculation of averaged ranks can be found in the Supporting Information in Table S1 (part B). According to the Wilcoxon test,[54] the top three models (CONS-STD-QUAL-PROB, CONDCORDANCE, and CONS-STD-PROB) were not significantly different from each other, with $p > 0.05$ for both analyzed sets, but were significantly better ($p < 0.05$) than other investigated measures. The LEVERAGE distance could not separate 90% accuracy predictions for any model ($c_{\text{TEST-90\%}} = c_{\text{TRAIN-90\%}} = 0$); therefore it was not analyzed further.

The rankings based on the accuracy coverage (Table 2) are not significantly different from those based on the AUC (Table 3). Namely, the rankings changed for the four last DMs (LEVERAGE, SCAvg, CORREL, and AD_MEAN2), which were however not significantly different from each other. One difference of the AUC rankings from the accuracy coverage rankings is that, according to the AUC criterion, CLASS-LAG outperformed the ASNN-STD-PROB. For all further analysis, we used the accuracy coverage ($c_{\text{TRAIN-90\%}}$ and $c_{\text{TEST-90\%}}$) because of its simpler and more intuitive interpretation.

According to the PCA plot in Figure 4, some of the models were quite similar, since they were based on the same descriptors and machine-learning methods, e.g., UNC_Drag_RF and PCI_Drag_RF, PCI_SiRMS_RF, and UNC_SiRMS RF. Combining these four models into two did not

Table 3. Averaged Rankings of the DMs Ranked by the AUC Criterion

distance to model	average rank (AUC, training set)	average rank (AUC, test set)
CONS-STD-QUAL-PROB	2.15	1.95
CONCORDANCE	1.4	2.1
CONS-STD-PROB	3.4	2.75
CONS-STD-QUAL	3.8	4.9
CLASS-LAG	6	4.95
ASNN-STD-PROB	6.4	5.65
CONS-STD	5.3	6.1
ASNN-STD	8.05	7.9
ELLIPS	12.1	9.6
AD_MEAN1	10.9	11.25
LEVERAGE	12.85	11.3
SCAvg	11.6	11.7
CORREL	9.95	11.85
AD_MEAN2	11.1	13

affect the sorting of compounds according to the DMs. Therefore, the rankings of the DMs, given in Tables 2 and 3, were not affected.

The dependency of the model performances for the CONS-STD-PROB DM is shown in the cumulative accuracy−coverage plot (Figure 5). The plot indicates that 25−70% of all compounds (depending on the model) are predicted with 90% accuracy. The same kind of plot for the CLASS-LAG DM (Figure 6) reveals poorer performance of the latter measure when it is not used in combination with the STD measure. The difference is visually apparent: for some of the models, CLASS-LAG was not able to separate predictions with 90% accuracy; in Figure 6, these models correspond to curves under the 90% line.

The performance of the CLASS-LAG DM appeared to be very dependent on the model, as can be observed in Figure 6 and in Table S1 (Supporting Information). This can be explained by different distributions of quantitative values of predictions, given by different models. Two histograms in Figure 7 reveal that the prediction values of UNC_SiRMS_SVM are similar to discrete values {−1,+1}; therefore, they contain less information than the predictions by PCI_SiRMS.Drag_RF, which are distributed more uniformly. Indeed, the CLASS-LAG DM failed for the first model, $c_{\text{Test-90\%}} = 0\%$ coverage, and yielded excellent results for the second one, $c_{\text{Test-90\%}} = 62\%$ coverage.

As described in the Methods section, CLASS-LAG punishes both negative and positive deviations from the class labels {−1;+1}. Thus, predictions outside of the [−1,+1] interval (referred to as "outer predictions") are considered less reliable than the exact −1 or +1. There were three models with outer predictions: EPA_2D_FDA, LNU_Drag_PLS, and OCHEM_ESTATE_ANN. When we rounded the outer predictions to {−1;+1} labels, their performance for CLASS-LAG did not change significantly from those for LNU_Drag_PLS and OCHEM_ESTATE_ANN; however, the performance significantly dropped for the EPA_2D_FDA model (see Figure SF2 in the Supporting Information).

The percentage of active (mutagenic) compounds within the range of 90% prediction accuracy is 51−55% and is not significantly different from the percentage of active compounds in the whole test set (53%). Therefore, mutagenic compounds are neither over-represented nor under-represented in the applicability domain of the models. Moreover, the prediction accuracy, sensitivity, and specificity of all of the models were not significantly different within the area of 90% prediction accuracy. Thus, the analysis of specificity and sensitivity is redundant; therefore, we used only prediction accuracy, calculated according to eq 9.

Several distances to the model were investigated in our study. A recently introduced probability-based measure of distance to a binary classification model, CONS-STD-PROB,[45] as well as its qualitative analog, CONS-STD-QUAL-PROB, provided a significantly better separation ($p < 0.05$ using the Wilxocon test) of predictions with low and high accuracy. Therefore, the quality of applicability domain estimation, using these methods, is significantly better than

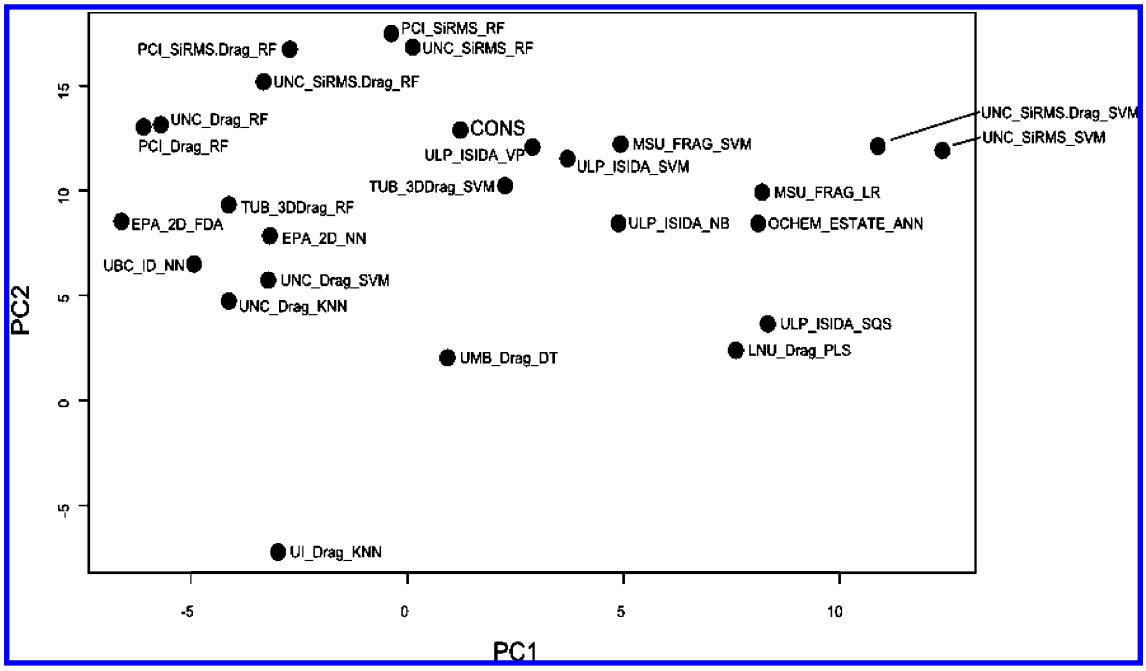

Figure 4. PCA plot of the Ames challenge models, based on the space of predictions for the test set. Four models (UI_Drag_LDA, UBC_ID_IWNN, ULZ_3DDrag_SVM, and ULZ_3DDrag_KNN) are not shown, since they were outliers of this graph.

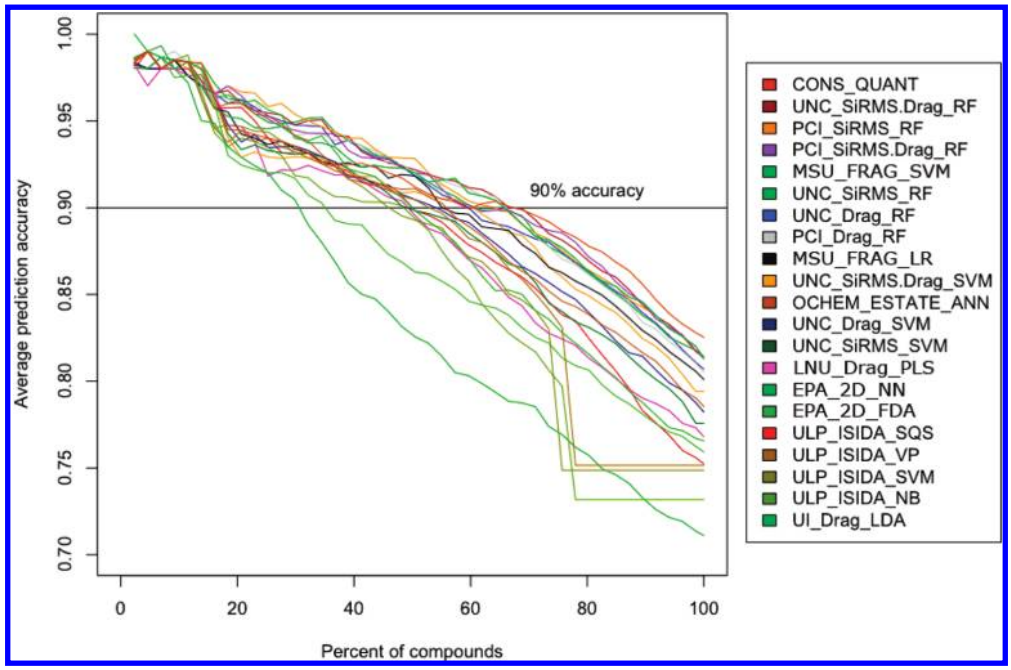

Figure 5. Cumulative accuracy−coverage plot for the CONS-STD-PROB DM based on the test set predictions. Only those 20 models are shown which had numeric prediction values available. The curves show the accumulative accuracy for a particular (variable) percentage of compounds. The curves clearly show that CONS-STD-PROB is highly correlated with the prediction accuracy. The models are ordered according to their overall performance for the test set.

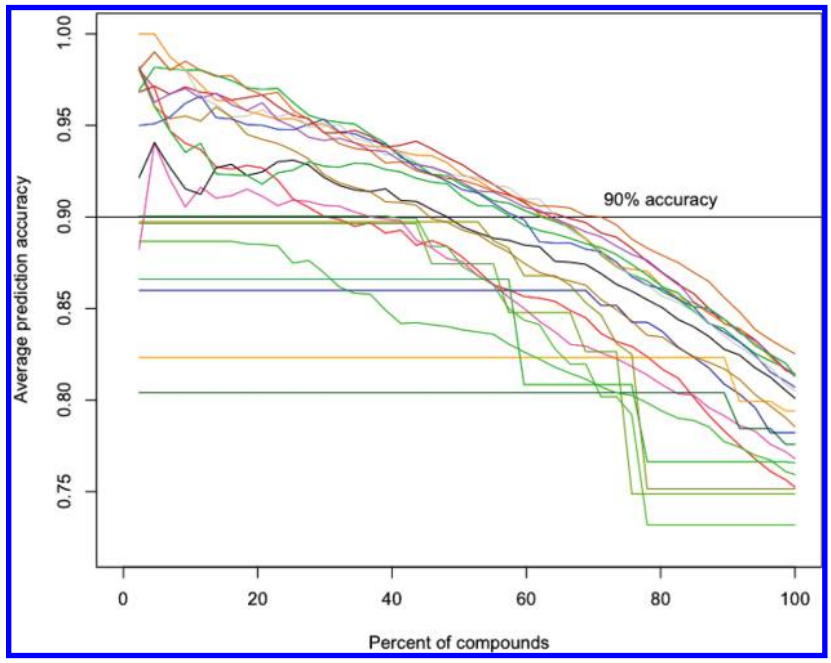

Figure 6. Cumulative accuracy−coverage plot for the CLASS-LAG DM. The plot is based on the test set predictions. The colors of the models are the same as in Figure 5.

that of the traditionally used CLASS-LAG method. It is interesting that CONCORDANCE, i.e., the measure of an agreement of predictions of a considered individual model with other members of the ensemble, was also amid the top three models and provided the best results for the training set. Therefore, it may be reasonable to use this simple measure along with the STD-PROB DMs.

The distances to the model, based on the space of descriptors (LEVERAGE, DA Index, ELLIPS, SCAvg, and AD_MEAN) identified only very small percentages of molecules with >90% accuracy (see Table 2 and Table S1, Supporting Information) and thus performed worse compared to other DMs considered in this study. The measures on which DA Index was based (namely, δ index and γ index) did not outperform DA Index when used as stand-alone DMs; therefore, they were not analyzed. Figure 8 (solid lines) demonstrates AD_MEAN results, which are worse compared to those of CONS-STD-PROB (dashed lines), which identified more than 40% of compounds as having this prediction accuracy for analyzed models.

The PCA plot of the DMs (Figure 9) calculated using the DM-based rankings of Ames challenge compounds reveals high similarity of the five DMs, which are based on the global consensus model. Indeed, these models explore

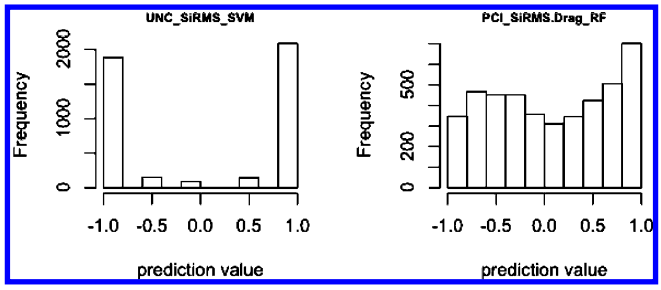

Figure 7. Distribution of prediction values for the two selected models. The prediction values of the model on the left chart resemble rounded discretized "−1" and "+1" values, whereas the values on the right chart have a continuous distribution and therefore provide more information for the estimation of uncertainty. This fact is confirmed in practice: CLASS-LAG of UNC_SiRMS_SVM (left chart) has poor performance (0% coverage of 90% accuracy) in contrast to PCI_SiRMS.Drag_RF (right chart), which separates 63% of compounds with 90% prediction accuracy.

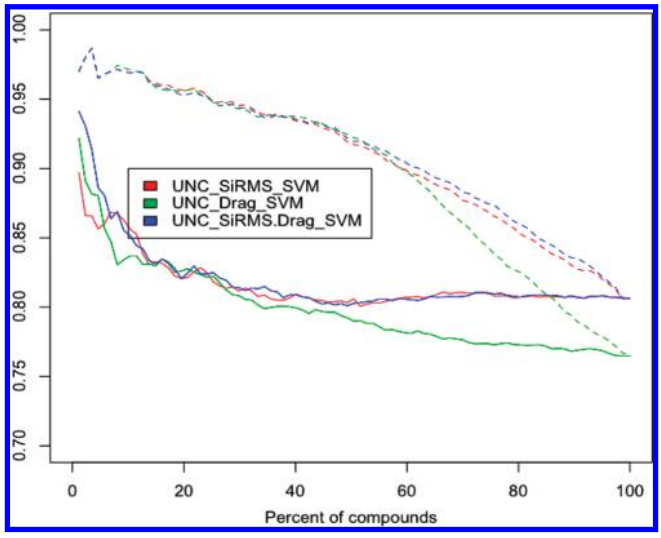

Figure 8. Comparison of AD_MEAN distance to the model (solid line) with CONS-STD-PROB (dashed line).

slightly different aspects of the same data and are strongly intercorrelated (see Table S5 in the Supporting Information). The CONS-STD, CONS-STD-QUAL, and CONCORDANCE DMs form one cluster within which the CONCORDANCE DM provided the best discrimination of the highly accurate predictions (Tables 2 and 3).

Analysis of the Qualitative Distances to Models. As mentioned in the Methods section, several groups provided qualitative AD measures for their respective models. The performance of CONST-STD-PROB for these models binned on several intervals is shown in Table 4 and is compared to the aforementioned models in this section.

Trust Level. This AD-related information, provided by the ULP group, is a generic estimation of the degree of trust for the prediction of a particular compound, ranging from optimal to poor, depending on how concordant individual models were in the prediction of this compound and how many of them had the compound in the applicability domain. We grouped all compounds by trust level and computed defacto prediction accuracy within each group. Results are summarized in Table 4 for the test set.

Prediction accuracy apparently drops with a decrease in trust level, excluding the poor trust level that has only 33 compounds in the corresponding group, which may not be sufficient for an evaluation of prediction accuracy. This measure provides worse results than the CONST-STD-PROB measure, as demonstrated in Table 5.

The 681 molecules with the largest CONS-STD-PROB values have an accuracy of about 52% only (Table 4), i.e., the same as the random guess. Of course, one should not use predicted results for these molecules but rather experimentally measure them. Once measured, such molecules will be important in extending the applicability domain of models and will allow for reliable predictions of new molecules, which are similar to them.

One-Class Classification AD (SVM1 AD). This measure was provided by the MSU group, and it distinguishes compounds inside and outside of AD. Accuracies, grouped by this flag, are summarized in Table 6.

A majority of compounds from the training and test sets were predicted to be inside the applicability domain using SVM1. The prediction accuracy for these compounds was on average 5% higher than those outside of AD. The CONS-STD-PROB method provided a much better separation of molecules; it achieves differences up to 40% for reliable and nonreliable predictions (Table 4).

DA Index. In addition to quantitative values analyzed in the previous section, the TUB group provided qualitative values for their DA_Index, summarized in Table 7.

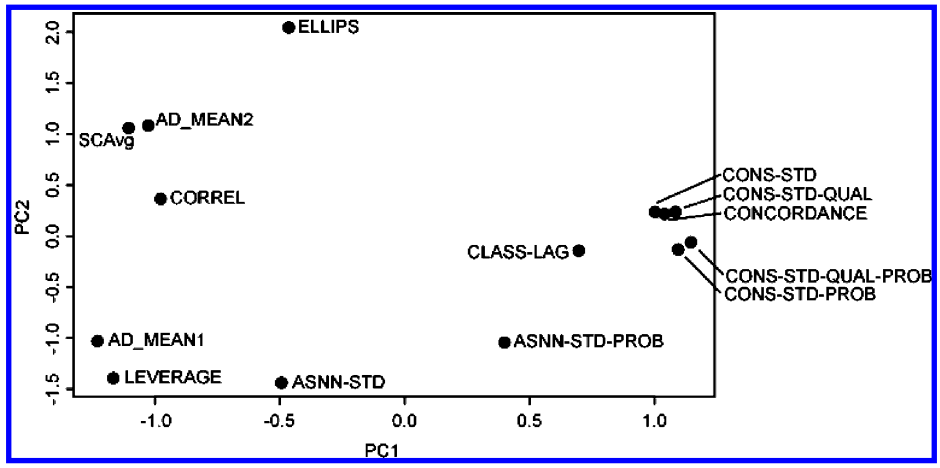

Figure 9. Principal component plot for the analyzed DMs. The PCA was based on the rankings that the DMs gave to the compounds from the training and test sets. Apparently, the five consensus-based DMs form two clusters: CONS-STD, CONS-STD-QUAL, and CONCORDANCE in the first cluster and CONS-STQ-QUAL-PROB and CONS-STD-PROB in the second one.

Table 4. Accuracy of Predictions According to CONS-STD-PROB[a]

	observed prediction accuracy				
number of compounds	ULP_ISIDA_SQS	TUB_3DDrag_SVM	TUB_3DDrag_RF	MSU_FRAG_LR	MSU_FRAG_SVM
500	96%	93%	93%	94%	95%
500	86%	89%	90%	89%	90%
500	76%	79%	81%	80%	83%
500	53%	64%	65%	66%	68%
181	48%	61%	55%	54%	54%
2181	75%	80%	80%	80%	81%

[a] For the first 500 compounds, it achieved an accuracy of 93−96%. This accuracy was higher than other qualitative ADs summarized in Tables 3−5.

Table 5. De Facto Performance of ULP_ISIDA_SQS Model for the Test Set with Regard to Trust Level and CONS-STD-PROB

		observed prediction accuracy	
trust level	number of compounds	trust level	CONS-STD-PROB
optimal	1221	81%	89%
good	512	79%	69%
medium	415	53%	46%
poor (or less)	33	70%	45%
overall test set	2181	75%	

Most compounds (1819, or 83% of the test set) had a DA-Index value of 0, which corresponds to the highest expected accuracy. However, the increase in accuracy of 2−6% was not significant for both TUB models, TUB_3DDrag_SVM and TUB_3DDrag_RF, as shown in Table 7. For the same models, the 500 most accurately predicted compounds identified using CONS-STD-PROB had 93% classification accuracy for both models, as shown in Table 4.

Ability to Estimate Accuracies of Predictions. So far, we investigated the abilities of DMs to separate accurate and inaccurate predictions. The main criterion for such performance was the percentage of compounds that were predicted with 90% accuracy for the training and test sets ($c_{\text{TRAIN-90\%}}$ and $c_{\text{TEST-90\%}}$) with regard to a particular DM. However, it is also important to estimate the expected accuracy of predictions for new molecules. Under the assumption that a model is correctly cross-validated and the investigated DM is consistent, the prediction accuracy for compounds within the same DM threshold should be not significantly different for both 5-CV results and the test set. Thus, the DM selected using 5-CV should cover the same percentage of molecules having about the same accuracy of prediction for the test set. In order to check this assumption, we selected a DM threshold that provides 90% accuracy using 5-CV and calculated accuracies of predictions for compounds within the same threshold on the *test set*.

We have compared accuracies for the training and test sets on compounds, having a DM within the threshold that provides 90% accuracy for the 5-CV result. The comparison was performed for all of the models in combination with all of the investigated DMs. There are 20 models tested against 12 DMs; therefore, there are $20 \times 12 = 240$ comparison cases. We found that the accuracies of predictions for 5-CV and test sets are consistent with significance level $p = 0.01$. With significance level $p = 0.05$, the estimated and observed accuracies were significantly different for two cases (Table S3, part C, Supporting Information), which does not exceed the statistically expected number of failures (for 240 comparison cases, 12 failures at the 0.05 level of signifi-

Table 6. Performance of MSU_FRAG_LR and MSU_FRAG_SVM Models Depending on the SVM1 AD Factor and CONS-STD PROB (For the Same Numbers of Compounds)

		observed prediction accuracy			
		MSU_FRAG_LR		MSU_FRAG_SVM	
SVM1 AD	number of compounds	SVM1 AD	CONS-STD-PROB	SVM1 AD	CONS-STD-PROB
		training set			
inside (= 1)	4194	79%	80%	80%	81%
outside (= −1)	167	75%	59%	79%	53%
overall training set	4361	79%		80%	
		test set			
inside (= 1)	2046	81%	82%	81%	83%
outside (= −1)	135	73%	53%	79%	55%
overall test set	2181	80%		81%	

Table 7. Performance of TUB Models for the Test Set Depending on DA Index and CONS-STD-PROB

		observed prediction accuracy			
		TUB_3DDrag_SVM		TUB_3DDrag_RF	
DA Index	number of compounds	DA Index	CONS-STD-PROB	DA Index	CONS-STD-PROB
0	1819	81%	83%	80%	84%
between 0 and 1	183	75%	62%	78%	61%
1	179	75%	60%	80%	60%
overall test set	2181	80%		80%	

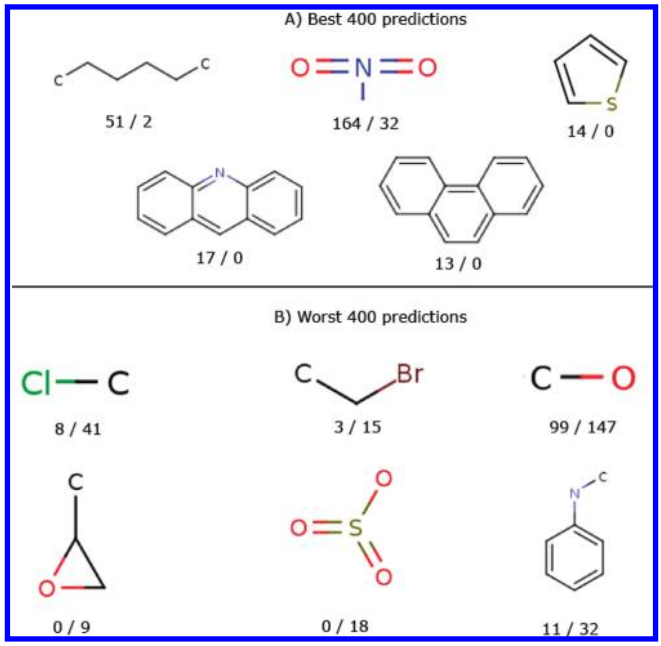

Figure 10. Molecular fragments, presented in the reliably and nonreliably predicted compounds. Shown are the fragments, significantly over-represented in the molecules with the highest accuracy (A) and the lowest accuracy (B) according to CONS-STD-PROB DM. Below the fragments are the numbers of relevant molecules with accurate (left of the slash) and inaccurate predictions (right of the slash).

cance). Thus, the accuracies estimated *a priori* using the training set are in agreement with observed accuracies for the test set.

Substructural Analysis of the Applicability Domain. To determine which types of molecules are predicted accurately and which are not, we have analyzed molecular subfragments for 400 predictions with the highest and lowest accuracies according to the CONS-STD-PROB DM, respectively. We will refer to these sets as "worst-400" and "best-400". We enumerated all of the fragments presented in these molecules and counted the number of molecules containing each fragment in each set.

If a fragment is equally distributed, the number of molecules from the "best-400" (or the "worst-400") containing this fragment should be distributed binomially with p equal to 0.5 and N equal to the total number of the molecules containing this fragment. If this assumption was invalidated with at least a $p < 0.05$ level of significance, we then considered the fragment as over-represented in one of the sets.

An overview of several significant fragments is presented in Figure 10. Apparently, the molecules containing long carbon chains, nitro groups, and thiophene groups were over-represented in the "best-400" predictions. We found out that long carbon chains were mostly presented in nonmutagenic compounds, whereas nitro and thiophene groups are mostly in mutagenic compounds. For the prediction of such compounds, there was a high level of agreement between the models. In contrast, the compounds containing chlorine, bromine, sulfonate, and epoxide groups are not reliably predicted by the models investigated in this study. We plan to provide a more detailed analysis of these fragments to detect "toxicophores", i.e., structural elements responsible for the mutagenicity of analyzed compounds.

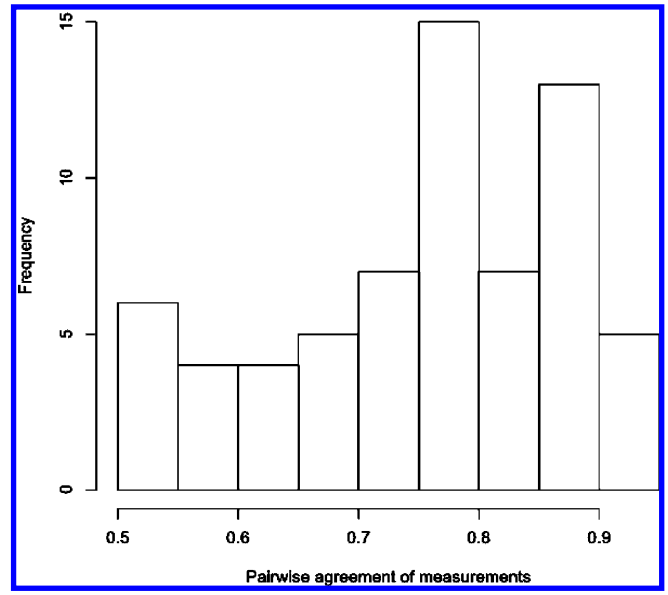

Figure 11. Distribution of the pairwise agreements of the Ames test measurements carried out by 12 laboratories. The data for the plot were taken from a study by Benigni and Giuliani.[55]

Data Variability Analysis. Several studies analyzed the variability of the Ames test experiments. Let us critically review them for a better understanding of the results of our modeling.

The first study by Benigni and Giuliani[55] assessed the Ames tests conducted for 42 compounds by 12 different experimental laboratories. Using the same data, for every pair of laboratories, we calculated the level of agreement as the number of the concordant measurements divided by the total number of measurements. The distribution of agreements of 66 lab pairs is shown in Figure 11. The average pairwise agreement is only 75%. At the same time, Figure 11 reveals that the agreement of results between some laboratories can be sometimes higher than 90%. This result was observed for 4 out of 66 pairs of laboratories (7% of all data). However, it is possible to expect a higher agreement if the data are measured within the same laboratory.

In the study by Piegorsch and Zeiger,[56] the experimental concordance between different laboratories was reported in the range of 70−87%. Each molecule in this set was measured in several experiments either in different laboratories or in the same lab but at different times. The outcomes of experiments were positive (+), weak positive (+W), negative, (−), and questionable (?). Let us consider, similar to how it was done in the analysis by the original authors, positive and weak positive as Ames mutagens and ignore nondecisive experiments, which, of course, are usually expected to be remeasured. Similar to the previous section, let us define the accuracy of one compound as the maximum number of positive or negative tests divided by the total number of decisive experiments. Such accuracy could be expected for our analysis, if we assume that molecules were tested on average just once. The average accuracy of the Ames test was 93% and 90% if we considered molecules with at least two (209 molecules) or three (49 molecules) decisive measurements, respectively.

We further explored this result using the variability of measurements used in our study. For this analysis, we used the Ames test data collected and publicly available at the

OCHEM website (http://ochem.eu). The database contains results for 3205 of the 6542 Ames challenge compounds. We used the same definition of accuracy as above and calculated an average accuracy of 94% for compounds, which had at least three measurements (1680 compounds selected from 189 articles). The variation of the minimal number of measurements from four to seven did not change this number more than ±0.3%. The 94% agreement is conformable with the achievable prediction accuracies of the models investigated in this study.

In this analysis, we mainly considered intralaboratory variations, as compared to the interlaboratory and mixture of the inter- and intralaboratory variations estimated in works of Benigni and Giuliani[55] and Piegorsch and Zeiger,[56] respectively. Unfortunately, it was impossible to carry out interlaboratory analysis in our study as there was an overlap in molecules reported in different articles. Moreover, in some cases, several authors, in particular Errol Zeiger, have contributed to the majority of articles, thus invalidating the goal of the interlaboratory comparison. Therefore, for the comparison of the DMs, we selected the accuracy of 90% obtained in work of Piegorsch and Zeiger[56] as a conservative threshold for interlaboratory comparison.

Confidence of Predictions vs Variability of Experimental Measurements. Different subsets of molecules may behave differently in experiments: some of them may have easily reproducible results (either mutagenic or nonmutagenic), while the other molecules may show higher variability, e.g., because of difficulties in experimental measurements such as metabolic stability, low solubility, etc. It would be interesting to know whether the methods described in the article can differentiate such chemicals.

We have analyzed the variability of measurements for molecules from the Piegorsch and Zeiger data set.[56] The total set contained 239 molecules, but three of them did not have structures defined and were excluded from our analysis. We developed a new ASNN model using all of the Ames challenge molecules with an exception of these 236 molecules, which formed the test set. The confidence of predictions was determined using the ASNN-STD-PROB DM. Amid 50 compounds with the highest and the lowest calculated confidences, we selected molecules that had at least three decisive measurements. There were 14 and 9 such molecules for the top and lower ranges with an agreement of experimental measurements of 96% and 89%, respectively. Moreover, there were also 13 and 21 compounds with questionable measurements within the same intervals.

We applied a similar analysis to the 1680 Ames challenge compounds having at least three measurements. We found that 150 molecules with the highest and the lowest confidence of predictions had an agreement of experimental measurements of 97% and 91%, respectively. Thus, the confidence of predictions determined by the DM correlated with the variability of experimental measurements: the molecules with a higher confidence of predictions have better agreements of experimental measurements and vice versa.

The Prediction Accuracy for EINECS, ENAMINE, and HPV Data Sets. In order to estimate the applicability of the QSAR Ames models to diverse chemical compounds, the OCHEM_ESTATE_ANN model was applied to the ENAMINE, EINECS, and HPV databases, described in the Methods section. The prediction accuracy for these data sets was estimated using bin-based accuracy averaging based on the ASNN-STD-PROB DM.

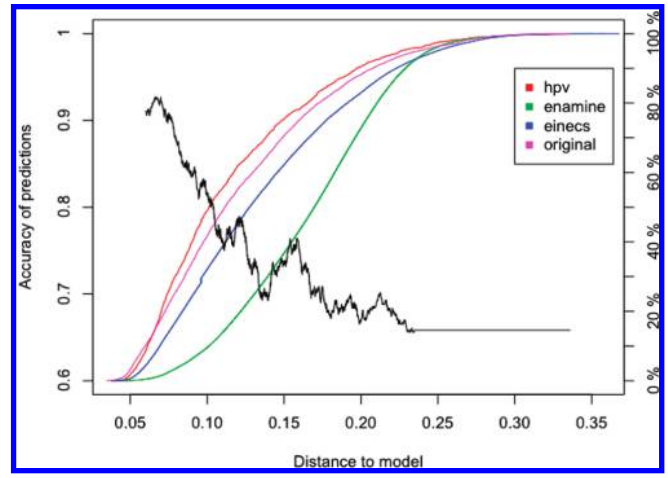

Figure 12. Estimated prediction accuracy for the original Ames challenge data set and the HPV, EINECS, and ENAMINE data sets. The black curve, based on bin-based averaging, plots the prediction accuracy (left *y* axis) against the ASNN-STD-PROB DM. Colored curves show percentages of compounds (right *y* axis) from the four data sets, having ASNN-STD-PROB not more than a particular threshold (*x* axis).

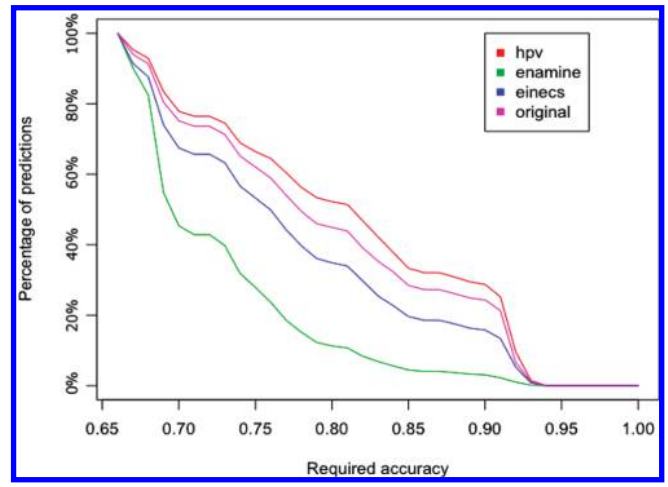

Figure 13. Percentages of compounds (*y* axis) having a required prediction accuracy (*x* axis). This plot is built for four data sets and uses the same data as Figure 12.

In Figure 12, the black curve corresponds to the average prediction accuracy as a function of ASNN-STD-PROB, while four colored curves illustrate the percentages of compounds from the four data sets having DM values less than corresponding thresholds. The plot in Figure 13 shows the percentages of compounds from the four data sets depending on the required prediction accuracy. Apparently, for the HPV and EINECS data sets, the percentages of reliable predictions (with at least 90% estimated prediction accuracy) were 30% and 16%, respectively, which is close to the percentage in the original data set, used for training and validation (25%). However, the percentage of reliable predictions in the ENAMINE data set was only 4%, probably due to a higher chemical diversity of compounds in comparison to the training set.

CONCLUSIONS

In this study, we have analyzed the AD problem for binary classification models. We investigated the relevance of

classical approaches to AD estimation for predictions of quantitative properties. The analysis was based on the Ames mutagenicity data set and involved 30 independent classification models.[11] The model developed by the HMGU group has been made publicly available in OCHEM, Online Chemical Modeling Environment,[57] at http://ochem.eu/models/1.

The analysis in this study was based on abstract measures of prediction uncertainty, referred to as "distances to models" (DMs). While the fact that measures such as CLASS-LAG, which can be used to discriminate accurate and inaccurate predictions, have been known for years, not many researches utilize them to assess the performance of their QSAR methods (frequently, only average model characteristics are reported).

The important message of this study was to demonstrate practical advantages of using DM and AD approaches. The most reliable predictions of the Ames test achieved experimental accuracy (ca. 90%), while unreliable predictions had an accuracy of random guessing (50%). The predictions of the later compounds are useless; one should measure such compounds experimentally rather than rely on predictions.

Several DMs were investigated and benchmarked. The DMs, based on the global consensus model provided significantly better separation ($p < 0.05$ using the Wilxocon test) of low and high accuracy predictions. The top-ranked DMs included a recently introduced probability-based measure of distance to a binary classification model, CONS-STD-PROB,[45] its qualitative analog CONS-STD-QUAL-PROB, as well as another very simple measure, CONCORDANCE, i.e., the agreement of a model's predictions with the global consensus model. Moreover, as shown in Figure 9, these three DMs were strongly correlated. The quality of the AD estimation using these methods was significantly better than tthat of the traditionally used CLASS-LAG method. Nonetheless, while CLASS-LAG did not work for the majority of the analyzed individual models, its performance for the global consensus model was not significantly different from the three aforementioned top-ranked DMs (see Table 2). It is important to mention that all three measures (CONS-STD-QUAL-PROB, CONS-STD-PROB, and CONCORDANCE) implicitly use the predictions given by the consensus model. As the consensus model is the best of all 31 models, these DMs may have performed best because they incorporate information from the best (consensus) model. If we do not consider the consensus-based DMs, the best measures were CLASS-LAG and ASNN-STD-PROB. Importantly, the DMs based on the output of the models outperformed the DMs solely on the basis of molecular structures (e.g., LEVERAGE and AD_MEAN).

Similar to our previous analysis of quantitative QSAR models,[8] we found that the best separation of the reliable and nonreliable predictions was provided by the same DMs. In other words, the compounds having the best prediction accuracy were the same for all of the models, regardless of the descriptors or the machine-learning technique used to develop them. This conclusion is in agreement with the work of Sheridan et al.[58] as well as with our own conclusions that the performance of models is dominated by the size and quality of the training set rather than by the method or the descriptors.[59]

Another important result of this study is the discovery of a correlation between the prediction uncertainty and the variability of experimental measurements of molecules. Namely, we have demonstrated that molecules with more accurate predictions had a higher agreement of experimental measurements and, vice versa, molecules with less accurate predictions showed higher disagreement with experimental measurements. Indeed, molecules from the first group contributed cleaner training sets and thus allowed models to achieve a higher accuracy of predictions for their analogs.

The discrimination of accurate and nonaccurate predictions is important from the practical point of view. If a compound is predicted with the accuracy, which is close to the accuracy of experimental measurements, one can use *in silico* values instead of measuring the activity for this compound. We have shown that the developed models predicted Ames mutagenicity for 35−65% of Ames challenge molecules with an accuracy similar to that of interlaboratory variation. Similar results were also achieved for quantitative models: the octanol/water partition coefficient (log P) was calculated for more than 60% of molecules with experimental accuracy.[60]

An accuracy of 90% was achieved for 35% and 20% of the HPV and EINECS databases of compounds using the ASNN model. However, for a larger and more diverse Enamine data set, only 6% of the compounds were predicted with such accuracy, presumably because of the higher chemical diversity of the Enamine collection. Thus, to increase the accuracy of predictions for such compounds, new experimental measurements are required.

In summary, the differentiation of reliable and nonreliable predictions of *in silico* approaches can decrease experimental costs by delivering accurate predictions for up to 2/3 of the molecules. At the same time, those compounds, which cannot be reliably predicted, should be measured. This additional data will extend the applicability domain of models and will allow reliable prediction of an even larger number of molecules. The combination of *in silico* approaches and experimental measurements can help to avoid redundant measurements, to screen large amounts of molecules even before they are synthesized, and, thereby, to provide significant savings of time and cost for the industry.

ACKNOWLEDGMENT

This study was partially supported with GO-Bio BMBF grant 0313883, FP7 project CADASTER 212668, and Germany−Ukraine collaboration project UKR 08/006, and by the NIH grants R01GM66940 and R21GM076059. We would like to thank all participants of the Ames challenge, who contributed to the development of models used in this study as well as the reviewers for their constructive remarks.

Note Added after ASAP Publication. This paper was published ASAP on October 29, 2010 with an error in the presentation of the names of the authors. The corrected version was published ASAP on November 8, 2010.

Supporting Information Available: Detailed tables with the DM scores and the values of DM comparison criteria. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Ertl, P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–380.

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

(3) Netzeva, T. I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; 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. M.; 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.

(4) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361–1375.

(5) Hemmateenejad, B.; Yazdani, M. QSPR models for half-wave reduction potential of steroids: A comparative study between feature selection and feature extraction from subsets of or entire set of descriptors. *Anal. Chim. Acta* **2009**, *634*, 27–35.

(6) Tropsha, A.; Gramatica, P.; Gombar, V. 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.

(7) Aires, F.; Prigent, C.; Rossow, W. B. Neural Network Uncertainty Assessment Using Bayesian Statistics: A Remote Sensing Application. *Neural Comput.* **2004**, *16*, 2415–2458.

(8) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746.

(9) Manallack, D. T.; Tehan, B. G.; Gancia, E.; Hudson, B. D.; Ford, M. G.; Livingstone, D. J.; Whitley, D. C.; Pitt, W. R. A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 674–679.

(10) Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjogren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von der Saal, W.; Zimmermann, G.; Schneider, G. Development of a Virtual Screening Method for Identification of "Frequent Hitters" in Compound Libraries. *J. Med. Chem.* **2002**, *45*, 137–142.

(11) Muratov, E. *Summer School on Chemoinforamtics*; Obernai: France, 2010; poster no 13.

(12) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K. Benchmark data set for *in silico* prediction of Ames mutagenicity. *J. Chem. Inf. Model.* **2009**, *49*, 2077–2081.

(13) Ames, B. N.; Lee, F. D.; Durston, W. E. An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proc. Natl. Acad. Sci. U.S.A* **1973**, *70*, 782–786.

(14) Breiman, L. Random Forests. *Mach. Learning* **2001**, *45*, 5–32.

(15) Chang, C.; Lin, C. LIBSVM - A Library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed Oct 1, 2010).

(16) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: New York, 2009.

(17) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learning* **2002**, *46*, 389–422.

(18) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab.* **2001**, *58*, 109–130.

(19) Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173.

(20) Martens, H.; Martens, M. Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food Qual. Prefer.* **2000**, *11*, 5–16.

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

(22) Tetko, I. V. Associative neural network. *Methods Mol. Biol.* **2008**, *458*, 185–202.

(23) Hall, L. H.; Kier, L. B.; Brown, B. B. Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1074–1080.

(24) Karakoc, E.; Sahinalp, S. C.; Cherkasov, A. Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. *J. Chem. Inf. Model.* **2006**, *46*, 2167–2182.

(25) Cherkasov, A.; Ban, F.; Li, Y.; Fallahi, M.; Hammond, G. L. Progressive docking: a hybrid QSAR/docking approach for accelerating *in silico* high throughput screening. *J. Med. Chem.* **2006**, *49*, 7466–7478.

(26) Cherkasov, A. Can 'Bacterial-Metabolite-Likeness' Model Improve Odds of '*in silico*' Antibiotic Discovery. *J. Chem. Inf. Model.* **2006**, *46*, 1214–1222.

(27) Cover, T.; Thomas, A. J. *Elements of information theory*; Wiley: New York, 1991; pp 1−543.

(28) Witten, I. H.; Frank, E. *Data mining: practical machine learning tools and techniques with Java implementations*; Morgan Kaufmann: San Francisco, CA, 1999; pp 1−374.

(29) 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 Descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191–198.

(30) Horvath, D.; Bonachera, F.; Solov'ev, V.; Gaudin, C.; Varnek, A. Stochastic versus Stepwise Strategies for Quantitative Structure-Activity Relationship GenerationHow Much Effort May the Mining for Successful QSAR Models Take. *J. Chem. Inf. Model.* **2007**, *47*, 927–939.

(31) Vapnik, V. *Statistical Learning Theory*; Wiley: New York, 1998; pp 1−736.

(32) Fan, R.; Chang, K.; Hsieh, C.; Wang, X.; Lin, C. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* **2008**, *9*, 1871–1874.

(33) Artemenko, N. V.; Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Prediction of Physical Properties of Organic Compounds Using Artificial Neural Networks within the Substructure Approach. *Dokl. Chem.* **2001**, *381*, 317–320.

(34) Baskin, I. I. ; Halberstam, N. M. ; Artemenko, N. V. ; Palyulin, V. A. ; Zefirov, N. S. In *Euroqsar 2002 Designing Drugs and Crop Protectants: Processes, Problems and Solutions*; Ford, M. , Livingstone, D., Dearden, J., Van de Waterbeemd, H., Eds.; Blackwell Science Inc: Bournemouth, 2003; pp 260−263.

(35) Kuz'min, V.; Artemenko, A.; Muratov, E. Hierarchical QSAR technology based on the Simplex representation of molecular structure. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 403–421.

(36) 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. Inhibitors of human rhinovirus 2 replication. *J. Med. Chem.* **2007**, *50*, 4205–4213.

(37) Polishchuk, P. G.; Muratov, E. N.; Artemenko, A. G.; Kolumbin, O. G.; 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.

(38) Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. DRAGON software: an easy approach to molecular descriptor calculations. *MATCH* **2006**, *56*, 237–248.

(39) Breiman, L.; Friedman, H.; Olshen, R. A.; Stone, C. J. *Classification and regression trees*; Wadsworth International Group: Belmont, CA, 1984; pp 1−359.

(40) Fan, R.; Chen, P.; Lin, C. Working Set Selection Using Second Order Information for Training Support Vector Machines. *J. Mach. Learn. Res.* **2005**, *6*, 1889–1918.

(41) 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.

(42) Contrera, J. F.; Matthews, E. J.; Daniel Benz, R. Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Regul. Toxicol. Pharmacol.* **2003**, *38*, 243–259.

(43) Tetko, I. V.; Poda, G.; Ostermann, C.; Mannhold, R. Accurate *in silico* log P Predictions: One Can't Embrace the Unembraceable. *QSAR Comb. Sci.* **2009**, *28*, 845–849.

(44) Breiman, L. Bagging predictors. *Mach. Learning* **1996**, *24*, 123–140.

(45) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Kovalishyn, V. V.; Prokopenko, V. V.; Tetko, I. V. Applicability domain for *in silico* models to achieve accuracy of experimental measurements. *J. Chemom.* **2010**, *24*, 202–208.

(46) 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.* **2009**, *98*, 861–893.

(47) Harmeling, S.; Dornhege, G.; Tax, D.; Meinecke, F.; Müller, K. From outliers to prototypes: Ordering data. *Neurocomputing* **2006**, *69*, 1608–1618.

(48) Schwaighofer, A.; Schroeter, T.; Mika, S.; Hansen, K.; Ter Laak, A.; Lienau, P.; Reichel, A.; Heinrich, N.; Müller, K. A probabilistic

approach to classifying metabolic stability. *J. Chem. Inf. Model.* **2008**, *48*, 785–796.
(49) Montgomery, D.; Peck, E. A.; Vining, G. G. *Introduction to linear regression analysis*; Wiley: New York, 2006; pp 1−639.
(50) Dragos, H.; Gilles, M.; Alexandre, V. Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *J. Chem. Inf. Model.* **2009**, *49*, 1762–1776.
(51) Schölkopf, B.; Smola, A. J. *Learning with kernels*; MIT Press: Cambridge, U.K., 2002; pp 1−644.
(52) Bishop, C. M. Novelty Detection and Neural Network Validation. *IEEE Proc.: Vis. Imag. Sign. Proc.* **1994**, *141*, 217–222.
(53) Fechner, N.; Jahn, A.; Hinselmann, G.; Zell, A. Estimation of the applicability domain of kernel-based machine learning models for virtual screening. *J. Cheminf.* **2010**, *2*, P2.
(54) Wilcoxon, F. Individual Comparisons by Ranking Methods. *Bio. Bull.* **1945**, *1*, 80–83.
(55) Benigni, R.; Giuliani, A. Computer-assisted analysis of interlaboratory Ames test variability. *J. Toxicol. Environ. Health* **1988**, *25*, 135–148.
(56) Piegorsch, W.; Zeiger, E. Measuring intra-assay agreement for the Ames Salmonella assay. *Lect. Notes Med. Inf.* **1991**, *43*, 35–41.
(57) Novotarskyi, S.; Sushko, I.; Körner, R.; Kumar, A.; Rupp, M.; Prokopenko, V.; Tetko, I. OCHEM - on-line CHEmical database & modeling environment. *J. Cheminf.* **2010**, *2*, P5.
(58) 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.
(59) 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.
(60) Tetko, I. V.; Poda, G. I.; Ostermann, C.; Mannhold, R. Large-scale evaluation of log P predictors: local corrections may compensate insufficient accuracy and need of experimentally testing every other compound. *Chem. Biodivers.* **2009**, *6*, 1837–1844.

CI100253R