Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

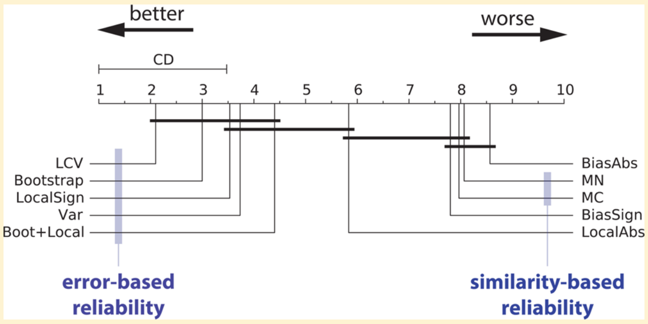
Assessment of Machine Learning Reliability Methods for Quantifying the Applicability Domain of QSAR Regression Models

Marko Toplak,[†] Rok Močnik,[†,‡] Matija Polajnar,[†] Zoran Bosnić,[†] Lars Carlsson,[§] Catrin Hasselgren,[§] Janez Demšar,[†] Scott Boyer,[§] Blaž Zupan,*[,†] and Jonna Stålring*[,§]

[†]Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia

[‡]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Olav Kyrres gate 9, 7489 Trondheim, Norway

[§]Computational Toxicology, Global Safety Assessment, AstraZeneca R&D, Pepparedsleden 1, 43183 Mölndal, Sweden



ABSTRACT: The vastness of chemical space and the relatively small coverage by experimental data recording molecular properties require us to identify subspaces, or domains, for which we can confidently apply QSAR models. The prediction of QSAR models in these domains is reliable, and potential subsequent investigations of such compounds would find that the predictions closely match the experimental values. Standard approaches in QSAR assume that predictions are more reliable for compounds that are “similar” to those in subspaces with denser experimental data. Here, we report on a study of an alternative set of techniques recently proposed in the machine learning community. These methods quantify prediction confidence through estimation of the prediction error at the point of interest. Our study includes 20 public QSAR data sets with continuous response and assesses the quality of 10 reliability scoring methods by observing their correlation with prediction error. We show that these new alternative approaches can outperform standard reliability scores that rely only on similarity to compounds in the training set. The results also indicate that the quality of reliability scoring methods is sensitive to data set characteristics and to the regression method used in QSAR. We demonstrate that at the cost of increased computational complexity these dependencies can be leveraged by integration of scores from various reliability estimation approaches. The reliability estimation techniques described in this paper have been implemented in an open source add-on package (https://bitbucket.org/biolab/orange-reliability) to the Orange data mining suite.

■ INTRODUCTION

At least since the early 1990s, the European Union has actively strived toward decreasing animal testing, marked by the founding of the European Center for Validation of Alternative Methods (ECVAM). However, the European chemical legislation, entitled REACH (Registration Evaluation Authorization and Restriction of Chemicals), increasingly requires chemicals used within the European market to be characterized with respect to their toxicological properties. To refine and replace animal testing, it encourages the use of alternative methods for regulatory purposes, including in vitro as well as in silico methods such as quantitative structure−activity relationships (QSAR). QSARs are widely used to predict properties of chemicals based solely on information about the chemical structure. QSARs model physiochemical properties such as solubility and lipophilicity, absorption, distribution metabolism, and excretion (ADME) and toxicological (T) properties. QSARs are used in the pharmaceutical industry to improve the quality of advancing compounds with respect to ADME and safety as well as in the chemical and cosmetic industries to characterize adverse effects of compounds with respect to humans and ecological systems.

As QSARs are hereby an integral part of major industrial sectors, the Organization for Economic Collaboration and Development (OECD) has developed five main quality criteria for QSARs,[1] one of which is specifically concerned with the applicability domain (AD) of the QSAR. Though there is no scientific consensus on the exact definition of the AD, the Setubal workshop report from 2003 conceptually defined it as “the response and chemical structure space in which the model makes predictions with a given reliability”.[2] The emphasis on the importance of the AD in QSAR modeling[3,4] originates from the notion of the vastness of chemical space and the relatively small coverage by data sets with ADMET annotations. It is generally perceived that chemical compounds “similar” to those used in the development of predictive models (e.g., the training set compounds) fall within the AD, while predictions of dissimilar compounds should not be considered reliable.[1,5] The AD is in practice often considered a binary entity, with compounds falling inside or outside of the AD. However, the OECD guidance document emphasizes the advantage of

Received: November 10, 2013
Published: February 3, 2014

continuous metrics, which may provide a quantitative description of the extent to which a chemical falls within the AD. Continuous descriptions are, for example, used by Sheridan et al.[6] who combine the information from chemical similarity, outcome probabilities as predicted by random forest, and the response value. Alternatively, the prediction confidence in QSAR has been described by Gaussian probability distributions as in predictive distributions[7−9] or recently in a conformal prediction framework.[10]

Most state-of-the-art QSAR models rely on machine learning algorithms such as random forest (RF) and support vector machines (SVM). Interestingly, a concept similar to the AD in QSAR is not well established in machine learning. However, the positioning of a model on the bias-variance spectrum is recognized and a greater variance would, within QSAR notation, correspond to a narrower AD. Within machine learning, prediction confidence is generally addressed by so-called reliability estimation methods, which to some extent overlap with the methods used to define the AD in QSAR. Depending on what information is used for scoring, reliability estimation methods can be classified into three main categories.[11,12] Feature range-based methods place confidence in predictions of examples whose feature values fall within the range of values encountered in the training set. Nearest neighbor methods use the distance to the most similar examples in the training set to infer reliability score from the error of the neighbors or use the information of the training sample density around the point of interest. The third class of methods rely on estimation of prediction error using sensitivity analysis that samples or perturbs the composition of the training set to estimate a distribution of predictions.

Methods based on error estimation and sensitivity analysis are perhaps best related to the OECD's quality criteria and the conceptual definition of AD. Here, we follow Bosnic and Kononenko[12] to suggest that the AD could be quantified by a continuous reliability score that estimates the prediction error directly or associates it with the variability of predictions under permutations in the input data. Compounds whose activity is predicted with a small error would be considered part of the AD, while compounds whose activity is predicted with increased error would deviate from the AD. The reliability is high if the estimated prediction error (or variability) for a point of interest remains small despite local perturbations of the training set. These approaches bypass the notion of "similarity" as directly related to accurate model predictions and allow for a wider range of algorithms, similarity based or not, to describe the prediction confidence. This type of reliability scoring thus relies on accurate estimate of prediction error, a requirement also noted in the predictive distributions framework.[6] Consequently, the correlation between the reliability score and the prediction error is a relevant and general assessment of the quality of a reliability method.

The principal contribution of our work is a finding that standard similarity-based approaches to reliability estimates in QSAR should be complemented with alternative error-based estimation techniques. In particular, we have explored the ML literature, and we examine recently introduced reliability scoring methods by Bosnic and Kononenko[12,13] in the context of QSAR. Their approaches provide continuous estimates of reliability, allow for reliability ranking of predicted compounds, and can be applied with any regression technique employed in QSAR modeling. We compare their reliability scoring techniques to standard similarity-based reliability scores in QSAR[4,14] and evaluate their quality based on the correlation between reliability and the prediction error.

In the paper, we propose techniques to quantify and study the quality of reliability estimation in a QSAR setting and present a new algorithm to use reliability estimates in identification of binary applicability domain. Our study includes 20 public QSAR data sets with continuous response modeled with several state-of-the-art regression techniques. We distinguish between structurally diverse (*global*) data sets and data sets composed of similar (*congeneric*) compounds like those in a lead optimization series in pharmaceutical discovery. Furthermore, we describe the compounds by bulk or structural descriptors or by a combination of the two to examine potential differences in method performance with respect to descriptor type.

■ METHODS

Input to the reliability scoring method is a training data set, regression method, and data instance for which we would like to assess the reliability when predicting a continuous label (response variable) with the model induced from the training data set. This section overviews various reliability estimation methods that include recently proposed approaches from the machine learning community,[12] nearest neighbor-based methods as routinely applied in QSAR and modeling,[4,14] and two ensemble-based methods.

In the following, we assume that the algorithms are given a training data set $\mathcal{T}$ of m data instances described with feature vectors $x \in \mathbb{R}^n$ and a real-valued label $y \in \mathbb{R}$. Given a new (target) data instance $x^* \in \mathbb{R}^n$, we would like to estimate the reliability $r(x^*)$ of its predicted label value $\widehat{y^*}$ by some regression model induced on the training set $\mathcal{T}$.

Reliability Estimation Scores. Below, we provide a brief description of the reliability estimation scoring techniques investigated in this paper. All described methods give reliability scores in which lower (absolute) values mean greater reliability.

Mahalanobis Distance to Nearest Neighbors (MN). Reliability estimation based on Mahalanobis distances assumes that prediction error is lower in *denser* parts of the problem space.[4,14] Space density around x is defined as the sum of Mahalanobis distances to its k nearest neighbors. The Mahalanobis distance between two data instances u and v is a generalization of the Euclidean distance that normalizes the data using the inverse of the covariance matrix S of the training data set

$$\mathrm{d}(u,v) = \sqrt{(u - v)^{\mathrm{T}} S^{-1} (u - v)}$$

Let $\mathcal{N}_k(x^*)$ denote a set of k data instances from $\mathcal{T}$ with lowest Mahalanobis distance to x^*. Then, the reliability estimate is

$$r_{\mathrm{MN}}(x^*) = \sum_{x \in \mathcal{N}_k(x)} \mathrm{d}(x^*,x)$$

Mahalanobis Distance to the Data Set Center (MC). Another possible assumption is that prediction error increases for data instances that are farther from the *center* of the training data set. Let $\bar{x}$ represent a data instance whose feature values are averaged across instances in the training set. The corresponding reliability measure for data instance x^* is its Mahalanobis distance to the training set center

$$r_{\mathrm{MC}}(x^*) = \mathrm{d}(x^*,\bar{x})$$

Sensitivity Analysis Scores (Var, BiasSign, and BiasAbs). Sensitivity analysis estimates changes in prediction due to small perturbations in the training data. To estimate the reliability of the prediction of a continuous label for x^*, we first estimate $\widehat{y^*}$ by inferring the predictive model from the original training set $\mathcal{T}$. Then, we construct a new training set $\mathcal{T}_\varepsilon = \mathcal{T} \cup \{(x^*, y_\varepsilon)\}$, where $y_\varepsilon = \widehat{y^*} + \varepsilon(y_{\max} - y_{\min})$ and where $y_{\min}$ and $y_{\max}$ are the two extreme outcome values from the original training set and ε is taken from the set of possible values, such as $E = \{0.01, 0.1, 0.5, 1.0, 2.0\}$ as originally proposed.[15] Finally, for each ε and related perturbed training set $\mathcal{T}_\varepsilon$, we train a model and use it to estimate the label value $\widehat{y^*}$ of x^*. Three different sensitivity analysis scores were proposed based on this procedure[15]

$$r_{\mathrm{Var}}(x^*) = \frac{1}{|\mathcal{E}|} \sum_{\varepsilon \in \mathcal{E}} \widehat{y}_\varepsilon - \widehat{y_{-\varepsilon}}$$

$$r_{\mathrm{BiasSign}}(x^*) = \frac{2}{|\mathcal{E}|} \sum_{\varepsilon \in \mathcal{E}} (\hat{y}_\varepsilon - \widehat{y^*}) + (\widehat{y_{-\varepsilon}} - \widehat{y^*})$$

$$r_{\mathrm{BiasAbs}}(x^*) = |r_{\mathrm{BiasSign}}(x^*)|$$

The intuition behind sensitivity scores is that reliable prediction will not succumb to local noise and will, regardless of perturbations, predict very similar outcome values for some feature vector x^*.

Bootstrap Variance (Bootstrap). This technique uses b bootstrap samples of the training set $\mathcal{T}$ to construct b regression models, each predicting the outcome $\widehat{y}_i$ of x^*, $i = 1,...,b$. Let $\overline{y} = (1/m)\sum_{i=1}^{m} \widehat{y}_i$ be the mean prediction. The reliability score is then bootstrap-estimated variance, where smaller variance denotes more reliable predictions

$$r_{\mathrm{Bootstrap}} = \frac{1}{m} \sum_{i=1}^{m} (\widehat{y}_i - \overline{y})^2$$

Local Cross-Validation Error (LCV). The expected prediction error at data instance x^* can be estimated from the observed prediction errors for training data instances in the neighborhood of x^*. Let $\mathcal{N} = \mathcal{N}_k(x^*)$ denote the nearest k neighbors of x^*, $\mathcal{N} \subset \mathcal{T}$, and $\hat{y}$ be the label value estimate for $x \in \mathcal{N}$ with the model developed from $\mathcal{N} \backslash x$. The local cross-validation error is then defined as a distance-weighted sum of errors on the nearest neighbors

$$r_{\mathrm{LCV}}(x^*) = \frac{\sum_{x \in \mathcal{N}} \mathrm{d}(x^*, x) \times |\hat{y} - y|}{\sum_{x \in \mathcal{N}} \mathrm{d}(x^*, x)}$$

In our experiments, $\mathrm{d}(x^*, x)$ was the Euclidean distance between the target data instance and its neighbor. According to original proposal of this procedure,[12] the number of neighbors k was set to approximately 5% of the learning data set size.

Local Prediction Error Modeling. (LocalSign, LocalAbs) compares the outcome prediction $\widehat{y^*}$ for the target data instance x^* to the outcome values $y(x)$ of its k nearest neighbors $x \in \mathcal{N}_k(x^*)$ from the training set. This score has a signed and unsigned variant:

$$r_{\mathrm{LocalSign}}(x^*) = \frac{1}{k} \sum_{x \in \mathcal{N}_k(x^*)} (y(x) - \widehat{y^*})$$

$$r_{\mathrm{LocalAbs}}(x^*) = |r_{\mathrm{LocalSign}}(x^*)|$$

Combination of Bootstrap Variance and Local Prediction Error Score. (Boot+Local) has been empirically shown to perform well in some settings[12]

$$r_{\mathrm{Boot+Local}}(x^*) = \frac{1}{2} \times (r_{\mathrm{Bootstrap}}(x^*) + r_{\mathrm{LocalAbs}}(x^*))$$

Notice that r_{BiasSign} and $r_{\mathrm{LocalSign}}$ are signed and are among scoring techniques listed in this section the only ones that can relate to direction and not only to magnitude of prediction error among our set of scores. This property is, however, not explored in our experiments.

Ensemble Approaches. The quality of reliability information scores may depend on the underlying modeling technique and characteristics of the data sets. It would be desirable to detect this dependency and use the optimal scoring method or the right combination of scoring methods. Given the training set, the best reliability scoring method may be determined through internal cross-validation.[13] We further propose an alternative method that uses stacking,[16] a popular ensemble approach from machine learning that is mostly used to fuse predictions coming from a set of predictive models.[17]

Scoring Selection by Internal Cross-Validation (ICV). This method uses cross-validation on the training set (so-called *internal cross-validation*) to determine which reliability estimation method performs best and then uses this method for the target data instance. Internal k-fold cross-validation splits the training data into k subsets. In each iteration, the validation chooses one of the data subsets to estimate the reliability of prediction of a continuous label for its data instances using a regression model developed on all of the $k - 1$ remaining data subsets. Repeating this procedure for each of the k subsets, we obtain a reliability score for all the data instances in the training set. The quality for the given reliability scoring technique is estimated by the Pearson correlation between reliability scores $r(x)$, $x \in \mathcal{T}$ and the predictive error $|y - \hat{y}|$. The process is repeated for all reliability scoring methods, and the best scoring method with the largest correlation to the predictive error is chosen. Given a target data instance x^*, this scoring method is now used to estimate the reliability.

Stacking of Reliability Estimators (Stacking). Similar to ICV, stacking uses internal cross-validation to obtain reliability estimates for each of the data instances in the training set. But instead of using these estimates to select the best method, stacking develops a model that integrates reliability estimates from all available reliability scoring techniques. Notice that with internal cross-validation each data instance $x \in T$ appears in exactly one test set, and thus, we can record its prediction error $|y - \hat{y}|$ and reliability estimates from the various scoring techniques. Let us form a new data set with these reliability estimates as input features and the prediction error as a continuous outcome. This data set constitutes the input to a regression model for the prediction error. Hence, stacking uses a combination of reliability estimates to model the prediction error. The output is a continuous score that can be treated as another (integrated) reliability estimate. For prediction, given a target data instance x^*, its reliability is then estimated by application of this model on a corresponding vector of reliability estimates of x^*. Intuitively, stacking should discard poor reliability scores and emphasize scoring techniques that are highly correlated with prediction error.

Regression Analysis. The reliability methods we investigate in this paper have been designed for regression analysis, that is, for data sets where data instances are labeled with continuous label (outcome, dependent variable). We have chosen three regression approaches that are common in QSAR and can yield models of high accuracy. These are partial least-squares regression (PLS), random forests of regression trees (RF), and support vector machines (SVM) with radial basis function kernel.[18] The regression methods were run as implemented in the Orange data mining suite.[19] Random forests contained 100 unpruned regression trees. Notice that in the experiments we treat the random forest method as a whole and in resampling procedures estimate the accuracies considering the entire forest rather than observing the variance among its trees. To spare computational time, parameters nu, gamma, and C for SVM were estimated from the entire data set by a cross-validation search through a small set of candidate values. This could offer SVM a slight advantage in accuracy, but because our aim was to test reliability scoring, we refrained from separately fitting SVM parameters for each of the many training data sets used in the experiments. The data sets considered by SVM were normalized prior to training.

Reliability Threshold for the AD. The conceptual definition of the AD states that a chemical structure is within the AD if it is predicted with "a given reliability".[2] Notice that this definition of the applicability domain is binary: predictions are either reliable or not. Each of the scoring techniques we use in this paper computes a degree of reliability, hence quantifying the degree to which a compound falls within the AD. Binary AD classification requires a threshold. To compute it, we propose deriving a null distribution of the reliability score (Algorithm 1). Given a data set of activity-labeled compounds, we split this into a training and a test data set. We randomly permute the values of the dependent variable in the training set and use this set to calculate the reliability of predictions for compounds in the test set. We repeat this procedure K times ($K = 100$), each time recording the reliability scores that at the end of the procedure provide for the null distribution. The reliability threshold is the $(1 - p)$-th percentile in the null distribution, with a standard choice for p of 0.05 or 0.01.

Algorithm 1 Inference of reliability threshold for applicability domain.

```
Input:   Training data set $\mathcal{T}$
         Reliability scorer $r(.,.)$
         Percentile parameter $p$ (default: 0.05)
         Number of validation epochs K (default: 100)
Output:  Reliability threshold $\theta$
$D = \emptyset$
for each $k \in \{1, 2, \ldots K\}$ :
    $\mathcal{A} \leftarrow$ random sample of 90% data instances from $\mathcal{T}$
    $\mathcal{B} \leftarrow \mathcal{T} \setminus \mathcal{A}$
    Randomly permutate values of outcome variable in $\mathcal{A}$
    $D \leftarrow D \cup \{r(x, \mathcal{A})$ for $x \in \mathcal{B}\}$
sort $D$ in decreasing order
$\theta \leftarrow p$-th percentile of $D$
```

Notice that the proposed algorithm is suitable for error-based reliability scoring algorithms that rather than the density take into consideration the accuracy of the QSAR procedure in a problem subspace. Outcome label permutation in the training set helps us to obtain reliability scores as they would be observed in the data set with no relation between the independent variables and the outcome. When applying such a threshold on new data, the degree of acceptance of compounds to the AD will depend on the difficulty of the modeling problem and the corresponding quality of the prediction models. For domains that are hard to model (domains with larger prediction errors), we expect that a lower proportion of compounds will be classified as within the AD as compared to domains where the relations between input features and outcome is clearer and easier to infer by some selected QSAR technique.

Prediction with a Reject Option. Reliability scoring approaches considered in this paper provide continuous estimates and allow for reliability ranking of predicted compounds. Such ranking can be explored by prediction with a reject option,[20] where analogous to the usage of the AD in QSAR, the prediction algorithm has the opportunity to decline to predict the response of an example if it is unreliable or if the reliability falls outside of some user-defined threshold. Predictors with reject options have been extensively investigated within classification,[20] but because of the absence of estimates analogous to class probabilities, the reject option is much less studied in regression. In principle, its application would either require a regression method-specific approach to estimate the uncertainty of prediction,[21] or similar to the methods studied here, method-independent reliability estimates. The assumption we explore in the paper is that regressors with reject option would gain in accuracy, and regressors with a higher reliability threshold would have higher accuracy than predictors with a lower reliability threshold.

Evaluation and Quality Scoring. The quality of a reliability estimation method on a given data set was assessed through a 10-fold cross-validation. Cross-validation splits the data into 10 subsets of approximately equal size and across 10 iterations treats one of the subsets as a test set and all the others as a training set. In each iteration, we develop a regression model on the training set and predict the label (outcome, response variable) and reliability for all the instances in the test set. To assess the quality of reliability estimates, we compute the Pearson correlation between the estimated reliabilities and prediction errors for the data instances in the test set. We use the absolute prediction error $|y - \hat{y}|$ for all reliability estimates except r_{BiasSign} and $r_{\text{LocalSign}}$, where the signed error $y - \hat{y}$ is used instead. We also compute the overall predictive accuracy of the regression through the coefficient of determination R^2. We report on average quality of reliability and average predictive accuracy across all 10 iterations of cross-validation.

We compute quality statistics for different combinations of data sets, data features, regression approaches, and reliability estimation methods. The methods examined in the paper are ranked by the quality of the reliability estimates. We report average ranks across the experiments and perform statistical evaluation of the differences between average ranks using the Nemenyi test as proposed by Demsar.[22] The results are presented graphically using critical distance diagrams.[22]

■ DATA SETS

A diverse suite of public QSAR data sets was compiled to empirically assess the utility of various reliability methods and examine their application to global QSARs as well as to more localized data sets.

Data sets originating from high-throughput screening assays usually encompass chemical compounds that may be very diverse in structure and chemical properties. As a representation of this type of QSAR data, we have considered 10 toxicologically relevant assays from the PubChem BioAssay database (http://pubchem.ncbi.nlm.nih.gov/assay). We refer to these data sets as *global*. These PubChem data sets include quantitative data on compound activity, as well as a categorical response, which labels the compounds as either “active” or “inactive”. Categorical labels were used to avoid problems with skewed data sets and response unbalance. Inactive compounds, which were in majority, were randomly excluded from the data to ensure an equal distribution of active and inactive compounds. Also, larger assays were randomly sampled to a maximum size of 5000 compounds. Table 1 shows the selected PubChem bioactivity assays together with the final number of compounds included.

Table 1. Global Data Sets from PubChem[a]

BioAssay ID	target	data set size (m)
1030	ALDH1A1	5000
932	STAT1	5000
2796	AhR	5000
2156	KCNQ2	5000
1239	NF-kB	5000
862	STAT3	3451
1511	hERG	3104
639	ER	2302
1479	thyroid	1635
631	PPARγ	1625

[a] Data sets that included a larger number of data instances were truncated (m = 5000).

Data related to lead optimization that contain structurally similar compounds were obtained from a QSAR benchmark repository.[23] From this collection, we studied the data sets that contained more than 100 compounds (Table 2). We refer to these data sets as *congeneric*. Notice that these data sets were also much smaller than the global data sets as the largest congeneric data set contained 397 compounds.

Table 2. Congeneric Data Sets from Mittal et al.[23] Are Examples of the Data from Tasks Such as Lead Optimization

abbreviation	observed activity	data set size (m)
DHFR	inhibition of dihydrofolate reductase	397
COX2	inhibition of cyclo-oxygenase-2	322
BZR	binding affinity to the benzodiazepine receptor	163
h-PTP	inhibition of human protein tyrosine phosphatase 1B	135
AMPH1	binding affinity to the human amphiphysin-1 SH3 domain	130
EDC	relative binding affinity to the estrogen receptor	123
ACE	inhibition of angiotensin-converting enzyme	114
HIVPR	inhibition of human immunodeficiency virus protease	113
AChE	inhibition of acetylcholinesterase	111
HIVRT	inhibition of HIV-1 reverse transcriptase	101

We characterized the chemical compounds using molecular descriptors as provided in the cheminformatics software toolkit RDKit (http://www.rdkit.org/). To investigate if the quality of the reliability scoring depends on the type of molecular descriptors, we constructed data sets that separately include bulk and structural descriptors and data sets with descriptors of both types (Table 3). These three descriptor sets are commonly used in QSAR modeling, and they provide fundamentally different descriptor vectors; structural descriptors are high dimensional sparse vectors, while bulk descriptors are relatively short and densely populated.

Table 3. Molecular Descriptors Used as Features in Global and Congeneric Data Sets

		n	
descriptor set	included descriptors	congeneric	global
bulk	physiochemical properties, counts, and indices	176	176
structural	structural (circular) fingerprints	≈200	≈2000
combined	bulk and structural descriptors combined	≈300	≈2200

RDkit descriptors for structural properties of chemicals were generated by using the circular fingerprints[24] with default settings and a radius of 1, while the set of 176 RDkit descriptors was used to provide a set of physiochemical properties, counts, and indices. The number of structural descriptors varies between the data sets as they represent counts of all molecular fragments of length up to 2 bonds occurring in each data set.

RESULTS AND DISCUSSION

We begin this section by reporting experiments that study the relation between reliability scores and prediction error. Next, and according to the correlation between these two quantities, we compare various reliability estimation methods and investigate whether their quality depends on the choice of regression methods, type of data sets (congeneric vs global), or type of descriptors. We then investigate whether the proposed procedures have merit when defining fixed reliability thresholds for the applicability domain. Finally, we report on the run time of the reliability scoring procedures.

Reliability Scores and Prediction Error. Our main assumption is that the error of the predicted activity should be lower for the compounds with associated higher reliability of prediction. Figures 1 and 2 show an example study on two data sets that confirms our expectations. We have split the two data sets into training (90%) and test (10%) sets and estimated the reliability of predictions for the test data instances. Both data sets display the trend that data instances for which the prediction was estimated as reliable indeed have lower prediction errors (Figure 1). We have sorted the data instances in the test sets by reliability and report on mean absolute errors when only the k test data instances with the highest reliability are considered (Figure 2). Selecting a subset of compounds with high reliability indeed decreases the error substantially. For instance, for BioAssay 1479, the error is the lowest (MAE ≈ 16) when only 10 data instances with the highest estimated reliability are considered but increases substantially (MAE ≈ 23) when this is assessed on a larger (k > 100) set of data instances that include those with lower reliability. For BioAssay 1239, the error increases from about 1.0 (with k around 10) to 2.5 (k > 200). We observed similar behavior across all the data sets in our study. The effect was, as expected, more pronounced with data sets where the correlation between the reliability estimates and the prediction error was higher.

Evaluation and Comparison of Reliability Estimation Methods. We have assessed the success of the reliability estimation methods for each combination of data sets, feature types, reliability scoring techniques, and QSAR regression methods. The highest correlations between reliability and

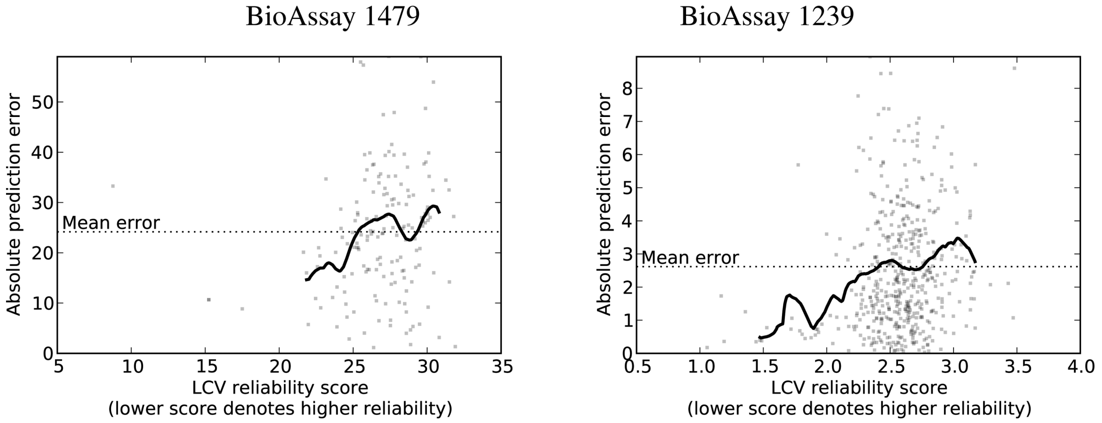


Figure 1. Absolute error and reliability for data instances from the test set for the LCV estimate with random forests on BioAssays 1479 (left) and 1239 (right). Errors and reliabilities for each data instance are presented with dots, while the black line shows these values smoothed with the Epanechnikov kernel.[18] The "mean error" is the average absolute prediction error in a 10-fold cross-validation. Both graphs show that overall the error increases for test data instances of lower reliability.

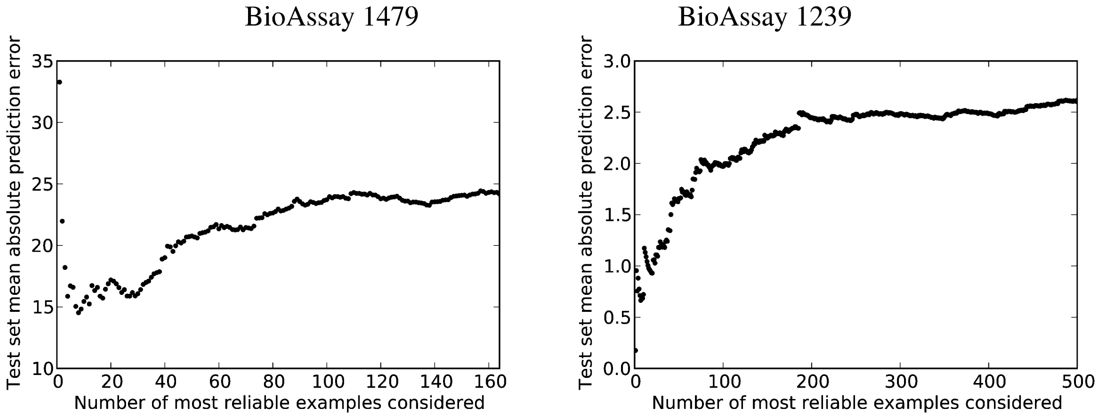


Figure 2. Mean absolute error on subsets of the most reliable data instances from the test set for the LCV estimate with random forests on BioAssays 1479 (left) and 1239 (right). The mean error is lowest if this is estimated from only a few most reliable test data instances (scores close to the origin of the graphs). Error increases when computed from larger set of data instances that include those of lower reliability (scores on the right part of each graph).

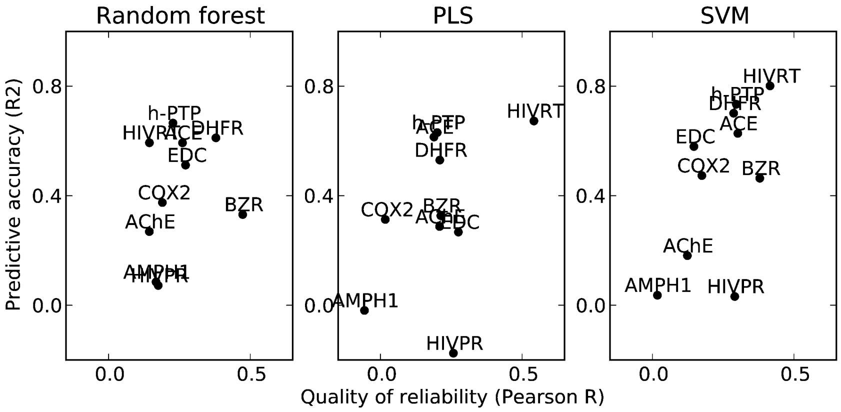


Figure 3. Quality of reliability assessment by stacking and cross-validated model accuracy for congeneric data sets, combined descriptor sets and the three modeling methods.

prediction error were about 0.5, which is consistent with previously reported results on standard machine learning data sets.[12] These correlations and their relation to 10-fold cross-validated predictive accuracy of modeling methods are depicted in Figures 3 and 4. Figure 3 shows these measures for the congeneric data sets and the stacking estimator. The results for the global data sets are presented in Figure 4. The graphs in these figures show that the quality of the estimations substantially varies across different data sets. This is again consistent with previous observations.[12] In all subsequent experiments, we first compared the reliability scoring methods on each particular data set and assigned them ranks based on

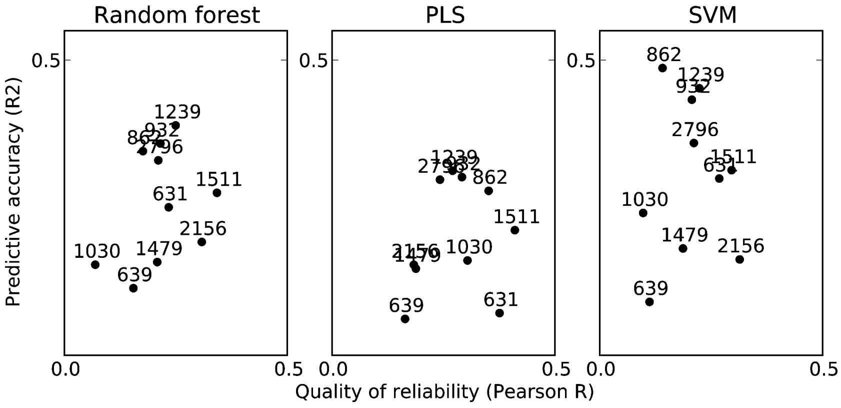


Figure 4. Quality of reliability assessments and cross-validated model accuracy for global data sets, combined descriptor sets, and the three modeling methods. Reliability was estimated by LCV (RF and SVM) and LocalSign (PLS).

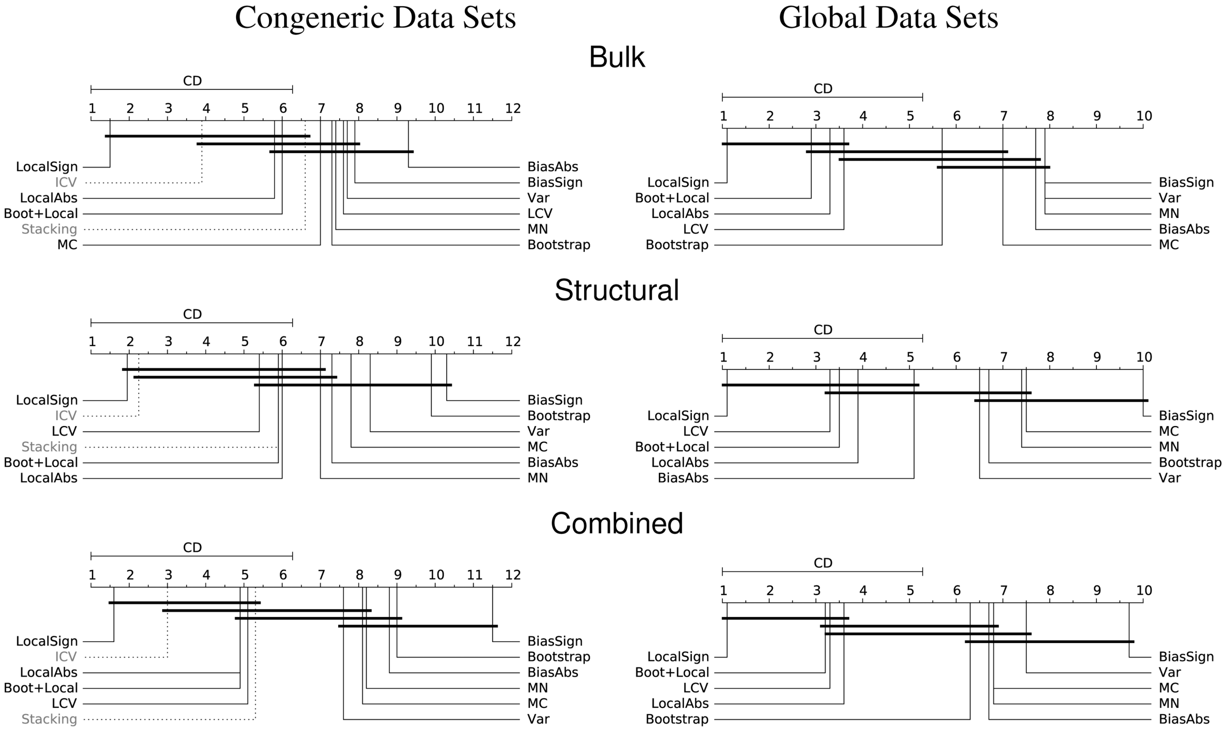


Figure 5. Average ranks from cross-validation testing for quality of reliability scorings and regression by PLS.

the quality of their reliability estimates as computed using the Pearson *R*. We then computed the overall quality of methods as an average rank over all data sets.

We next investigated the differences between reliability scoring methods as applied to congeneric or global data sets and used in combination with a specific regression technique. The results are shown in Figures 5, 6, and 7. Given the regression technique, we ranked the estimation methods according to their quality. Critical distances diagrams show averaged ranks of reliability methods across the data sets. In these graphs, the critical difference (CD, the line segment at the top of each diagram) indicates the required difference in ranks to recognize two approaches as significantly different ($p < 0.05$). The groups of approaches with insignificant differences in quality are connected with bars across the corresponding ranks. For instance, in the diagram at the top right column of Figure 5 (PLS on global data sets with bulk descriptors), the performance of reliability scorers LocalSign, LCV, Boot+Local, and LocalAbs is not significantly different, but they all perform significantly better than Var and MN. In this graph, only LocalSign ranks significantly better than Bootstrap and all other lower-ranked methods.

The rank order of the reliability scoring techniques depends on the regression method. For PLS, the local prediction error modeling method LocalSign is the highest-ranked estimator regardless of the data feature set and data set type. It does not, however, significantly outperform several other estimators, including LocalAbs, Boot+Local, and LCV, the latter with exception of congeneric data sets with bulk description, where LocalSign is significantly better. Importantly, for each data set type and any of the three descriptor types, LocalSign always significantly outranked the two Mahalanobis similarity-based measures MC and MN.

For the other two regression approaches, Bootstrap and LCV perform well for random forests, but the differences with other estimators are less pronounced. LCV is the winner for SVM,

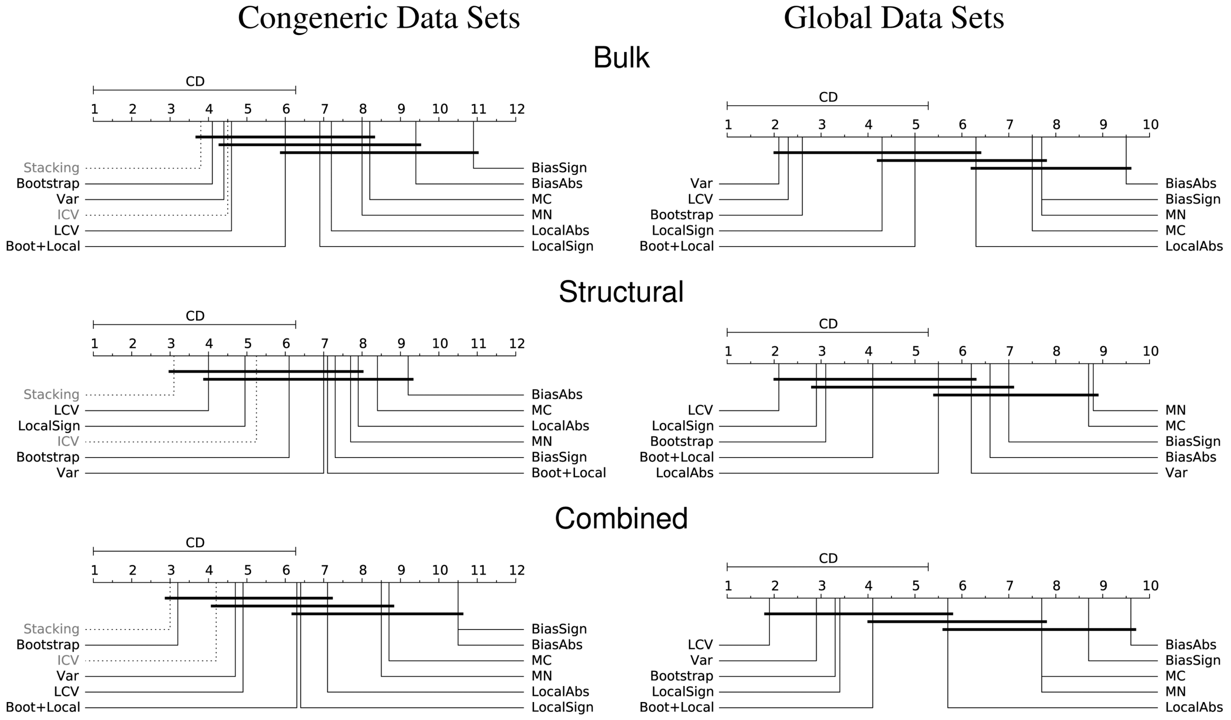


Figure 6. Average ranks from cross-validation testing for quality of reliability scorings and regression by random forests.

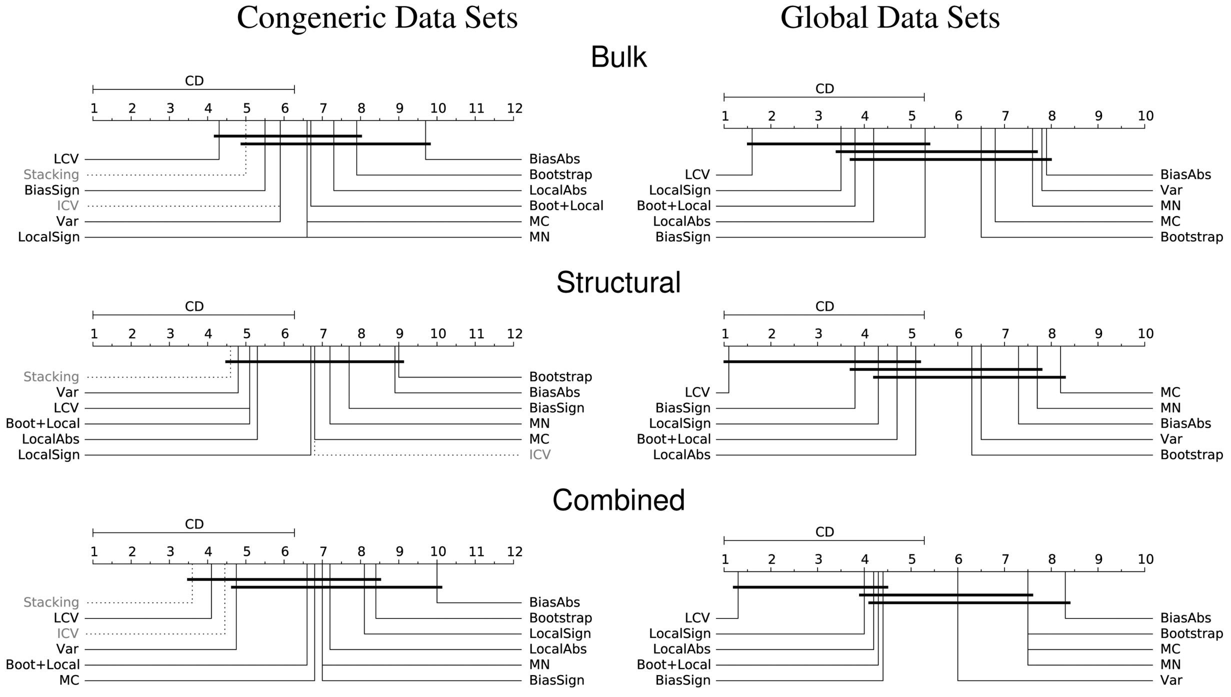


Figure 7. Average ranks from cross-validation testing for quality of reliability scorings and regression by SVM.

but the differences with other top approaches are not statistically significant.

Consistently, the approaches recently introduced in the machine learning community[12] outrank those based on Mahalanobis distance. Across most of experiments, MC and MN ranked worst, with a significant lag behind the best-performing methods.

Reliability Scoring Differences between Data Set Types and between Various Feature Representations. The results displayed in the previous section indicate that the ranking of reliability scoring does not depend on the type of the data set (congeneric vs global) or the descriptor type; the order of the methods on the left and right of the Figure 7 is the same. Similarly, their order changes only slightly across types of type descriptors (rows of graphs). Also, consider the graphs for specific type of data set (say, graphs in the left column of Figure 7) to notice that changing a descriptor set only slightly changes the ordering of reliability scoring techniques.

To further illustrate the inability to give any specific guidance on which descriptor type has the potential to give the most accurate reliability scores, we display the critical distance graphs in Figure 8 for the quality of the ICV scoring techniques. Each critical distance diagram in this figure reports on a rank of reliability quality observed for specific descriptor set as averaged across 10 congeneric data sets. For PLS and random forests, the reliability estimation by ICV performed equally well as the rank differences between the descriptor sets were not significant ($p < 0.05$). The only significant advantage was observed for SVM with the combined descriptor set with respect to the structural set. We conclude that the choice of descriptor type is generally insignificant for the quality of the reliability method.

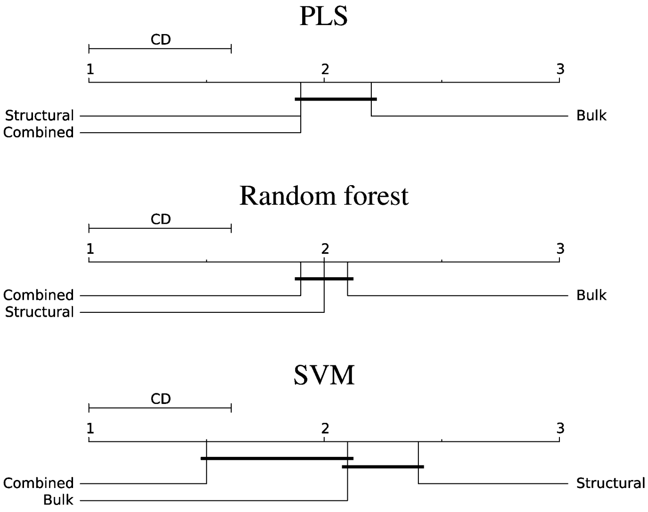


Figure 8. Averaged ranks of quality of ICV reliability estimates on congeneric data sets associated with three different descriptor sets as observed for three different regression techniques.

Quality of Reliability Scoring Methods Regardless of Feature Space and Data Set Type. Ranking the reliability scoring technique does, however, depend on the QSAR regression technique. We can hence investigate aggregated results that provide average reliability method ranks across all different experiments, joining those with different data types and descriptor types but studying them separately for each regression method (Figure 9). Reliability methods in this figure cluster more succinctly, and the differences are more significant. The ranking for PLS is again different than that for random forest and SVM, while the ranking of reliability scoring techniques for the latter two methods is similar. As in the previous rankings, we can expose LCV for random forest and SVM and LocalSign for PLS. Joining the data set and descriptor type confirms the consistently poor performance of classical Mahalanobis-based approaches MN and MC. Absent from our comparison are ensemble-based approaches due to their prohibitively high computational costs in cross-validated procedures on global data sets; the experiments with these were carried out on congeneric data sets only.

Reliability and a Binary Applicability Domain. We have also tested the algorithm for inference of a reliability threshold and classification into an applicability domain. Figure 4 shows results on congeneric data sets with random forests and reliability scoring by Bootstrap. We used cross-validation and evaluated the accuracy only on those compounds from each test set that belonged to the AD ($p < 0.05$). Because of the occasional occurrence of very small sets of selected compounds, we report on cross-validated mean absolute value and compare this to the error observed on all test instances where no selection took place. To avoid overfitting, the reliability thresholds were always inferred on the training set separately for each iteration of cross-validation.

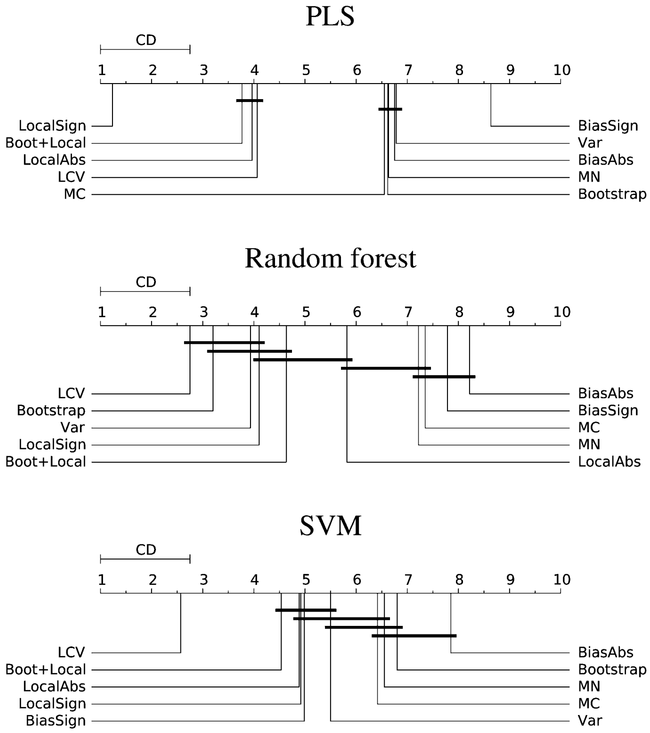


Figure 9. Average ranks from cross-validated quality of reliability scores for three different regression techniques. The reliability score quality was assessed on all available data sets, that is, on all global and congeneric data sets with bulk, structural, and combined descriptors. Average ranks were thus computed on (10 + 10) × 3 = 60 various data sets.

Table 4 reveals that, as expected, the error in AD-compliant compounds is reduced (MAE(AD) < MAE); the only exception occurring with the data sets HIVRT and ACHE.

Table 4. Cross-Validated Random Forest Accuracy on Compounds from Bootstrap-Scored Applicability Domain on Congeneric Data Sets[a]

			AD by Bootstrap	
data set	R^2	MAE	MAE(AD)	P(AD)
HPTP	0.67	0.26	0.21	47%
DHFR	0.61	0.68	0.45	20%
HIVRT	0.59	0.63	0.74	13%
ACE	0.59	1.03	0.86	39%
EDC	0.51	0.93	0.52	18%
COX2	0.38	0.88	0.77	8%
BZR	0.33	0.65	0.33	13%
ACHE	0.27	0.83	0.87	11%
AMPH1	0.09	0.64	0.56	6%
HIVPR	0.07	0.95	0.46	3%

[a] R^2 = cross-validated accuracy reported as average coefficient of determination (estimated from all instances in the test sets). MAE = mean absolute error (estimated from entire test set). MEA(AD) = mean absolute error (estimated only on test compounds from applicability domain), P(AD) = proportion of instances from training set that belong to applicability domain.

Table 5. Running Times (h:min:s or min:s) for one iteration of cross-validation of reliability estimation on PubChem data set 1479[a]

features	learner	MN	MC	BiasAbs, BiasSign, Var	Boot+Local, Bootstrap, LocalAbs, LocalSign	LCV
structural	RF	22:21	0:40	4:19:37	13:03	11:47
structural	PLS	24:44	1:59	34:09:12	1:41:31	7:27
structural	SVM	24:08	2:20	35:05:14	1:38:47	44:45
combined	RF	28:56	0:42	2:55:09	9:24	18:43
combined	PLS	31:49	2:36	43:35:46	2:18:06	9:28
combined	SVM	31:41	2:52	37:54:03	2:01:43	56:53
bulk	RF	1:41	0:05	0:53:50	2:56	7:25
bulk	PLS	1:40	0:04	0:44:47	2:21	0:38
bulk	SVM	1:46	0:09	2:15:32	6:45	3:29

[a] Several methods required similar preprocessing encompassing most of computational time (like BiasAbs, BiasSign and Var); for these, we report on joint execution times.

The number of compounds being classified within the AD ranges from a mere 3% (HIVRP) to 47% (HPTP). To assess if there is any regularity in these numbers, the table also provides cross-validated R^2 estimates that report on the correlation between reliability scores and error of prediction. We can observe that a higher proportion of compounds were classified as inside the AD for the data sets where the reliability estimation is more accurate. For instance, the highest R^2 of 0.67 was observed with the HPTR data where most (47%) of the compounds from the test sets were classified as within the AD. The smallest proportion of AD compounds was observed with HIVPR where the quality of reliability estimates was the lowest.

Our findings support the intuitive perception that data sets that are harder to model or more noisy will have be fewer compounds classified as within the applicability domain. With a less exposed relationship between input features and compound activity, the permutation-inferred null distribution will overlap more with the distribution of scores from the test sets and will hence correctly result in fewer compounds in the AD. The procedure that we proposed seems to correctly adjust the threshold according to the modeling difficulty of the problem domain and ability of QSAR to find true structure−activity dependencies.

Running Times and Choice of Reliability Scoring Technique. Overall, the choice of the reliability estimator is not trivial. It is encouraging, however, that automatic selection or combination of reliability estimates ranks well. We have observed the utility of ensambling by stacking or method selection through internal cross-validation on smaller congeneric data sets. These two methods consistently ranked well, but the difference between the two was never significant. The essential problem with these two methods is their computational complexity. On top of an already expensive reliability estimation, they need to be executed an additional 10 times within an internal cross-validation loop. Most of the investigated reliability methods are computationally expensive (see Figure 5 for execution times). This is especially true for the sensitivity-based methods, which require repeated relearning of the regression model with perturbed data sets for each data instance. When a costly technique like BiasSign is combined with high-quality regression method such as SVM, the reliability estimate for one single data point could require an additional 10 model inferences. Computationally expensive methods perform well and also significantly outperform computationally lighter Mahalanobis distance-based approaches that are currently used in QSAR. The algorithms, however, are embarrassingly parallel and can run concurrently for different test data instances, reliability estimation methods, and regression techniques.

CONCLUSION

Our study investigates the assumption that the smaller the estimated prediction error of a compound is, the more confidently the compound falls into the applicability domain of the QSAR model, aligning with OECD's conceptual interpretation of the applicability domain.[2] Hence, the applicability domain is quantified by reliability methods that estimate the error of prediction. Recently, a set of such approaches was proposed within the machine learning community,[12] and we here study their utility within QSAR. These methods provide a continuous reliability score, which can confidently be transformed into a binary representation if required in the applied setting.

We performed a comprehensive study on over 20 QSAR data sets. This is, to our knowledge, currently the largest and most comprehensive in silico evaluation of reliability estimation techniques in QSAR. Our principal finding is that error-based reliability estimation methods outperform or are at worst as least as accurate as similarity-based methods. Our experimental analysis also shows that the performance of the reliability methods is independent of the compound density structure (congeneric vs global) and descriptor type but depends on the choice of the regression algorithm. Ensambling techniques promote the performance further and leverage the dependence on the regression algorithm.

Overall, our study confirms the usefulness of error-based reliability estimates in QSAR, proposes a scheme for their evaluation and ranking in QSAR, describes an algorithm to relate the estimates to a binary applicability domain, and provides strong evidence that current similarity-based approaches should be complemented with alternative estimation techniques.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: blaz.zupan@fri.uni-lj.si (B.Z.).
*E-mail: jonna.stalring@astrazeneca.com (J.S.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Authors from the University of Ljubljana were supported by Program and Research grants from Slovenian Research Agency (P2-0209, J2-5480) and by a project grant from AstraZeneca.

REFERENCES

(1) *Guidance Document on the Validation of (Quantitative) Structure—Activity Relationship QSAR Models*; OECD Series on Testing and Assessment No.69; OECD Environment Directorate, Environment, Health and Safety Division: Paris, 2007.

(2) Jaworska, J. S.; Comber, M.; Auer, C.; Van Leeuwen, C. *Environ. Health Perspect.* **2003**, *111*, 1358−1360.

(3) Tropsha, A.; Gramatica, P.; Gombar, V. *QSAR Comb. Sci.* **2003**, *22*, 69−77.

(4) Weaver, S.; Gleeson, M. P. *J. Mol. Graphics Modell.* **2008**, *26*, 1315−26.

(5) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−28.

(6) Sheridan, R. P. *J. Chem. Inf. Model.* **2012**, *52*, 814−23.

(7) Sahlin, U.; Filipsson, M.; Öberg, T. *Mol. Inf.* **2011**, *30*, 551−564.

(8) Clark, R. D. *J. Cheminf.* **2009**, *1*, 11.

(9) Wood, D. J.; Carlsson, L.; Eklund, M.; Norinder, U.; Stå lring, J. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 203−19.

(10) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. *Ann. Math. Artif. Intell.* **2013**, DOI: 10.1007/s10472-013-9378-2.

(11) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. *ATLA, Altern. Lab. Anim.* **2005**, *33*, 445−59.

(12) Bosnicć, Z.; Kononenko, I. *Data Knowl. Eng.* **2008**, *67*, 504−516.

(13) Bosnić, Z.; Kononenko, I. *Knowl. Eng. Rev.* **2010**, *25*, 27−47.

(14) Bruneau, P.; Mcelroy, N. R. *J. Chem. Inf. Model.* **2006**, *46*, 1379−1387.

(15) Bosnić, Z.; Kononenko, I. *Appl. Intell.* **2008**, *29*, 187−203.

(16) Wolpert, D. H. *Neural Networks* **1992**, *5*, 241−259.

(17) Dzeroski, S.; Zenko, B. *Mach. Learn.* **2004**, *54*, 255−273.

(18) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer-Verlag: Berlin, 2009.

(19) Demsar, J.; et al. *J. Mach. Learn. Res.* **2013**, *14*, 2349−2353.

(20) Herbei, R.; Wegkamp, M. H. *Can. J. Stat.* **2009**, *34*, 709−721.

(21) Wiener, Y.; El-Yaniv, R. Pointwise Tracking the Optimal Regression Function. In *Advances in Neural Information Processing Systems 25*; 26th Annual Conference on Neural Information Processing Systems, December 3−6, 2012, Lake Tahoe, Nevada,2012, pp 2051−2059.

(22) Demsar, J. *J. Mach. Learn. Res.* **2006**, *7*, 30.

(23) Mittal, R. R.; McKinnon, R. a.; Sorich, M. J. *J. Chem. Inf. Model.* **2009**, *49*, 1810−20.

(24) Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010**, *50*, 742−54.