Conformal Regression for QSAR Modelling – Quantifying Prediction Uncertainty

Fredrik Svensson[1,2], Natalia Aniceto[1], Ulf Norinder[3,4], Isidro Cortes-Ciriano[1], Ola Spjuth[5], Lars Carlsson[6], Andreas Bender[1]*

[1] Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

[2] IOTA Pharmaceuticals, St Johns Innovation Centre, Cowley Road, Cambridge CB4 0WS, UK

[3] Swetox, Karolinska Institutet, Unit of Toxicology Sciences, Forskargatan 20, SE-151 36 Södertälje, Sweden

[4] Department of Computer and Systems Sciences, Stockholm University, Box 7003, SE-164 07 Kista, Sweden

[5] Department of Pharmaceutical Biosciences, Uppsala University, Box 591, SE-75124, Uppsala Sweden

[6] Quantitative Biology, Discovery Sciences, IMED Biotech Unit, AstraZeneca, SE-43183, Mölndal, Sweden

*fs447@cam.ac.uk

Abstract

Making predictions with an associated confidence is highly desirable as it facilitates decision making and resource prioritization. Conformal regression is a machine learning framework that allows the user to define the required confidence and delivers predictions that are guaranteed to be correct to the selected extent. In this study, we apply conformal regression to model molecular properties and bioactivity values and investigate different ways to scale the outputted prediction intervals to create as efficient (i.e. narrow) regressors as possible. Different algorithms to estimate the prediction uncertainty were used to normalize the prediction ranges and the different approaches were evaluated on 29 publicly available datasets. Our results show that the most efficient conformal regressors are obtained when using the natural exponential of the ensemble standard deviation from the underlying random forest to scale the prediction intervals, but other approaches were almost as efficient. This approach afforded an average prediction range of 1.65 pIC50 units at the 80 % confidence level when applied to bioactivity modeling. The choice of nonconformity function has a pronounced impact on the average prediction range with a difference of close to one log unit in bioactivity between the tightest and widest prediction range. Overall, conformal regression is a robust approach to generate bioactivity predictions with associated confidence.

Introduction

Quantitative Structure-Activity Relationship (QSAR) studies explore the relationship between molecular structure and bioactivity or molecular structure and properties (QSPR).[1,2] QSAR methods form the backbone of early stage predictions in drug discovery, and are routinely applied throughout the pharmaceutical industry.[3] However, the confidence or reliability in the predictions generated by QSAR models are sometimes difficult to accurately assess and quantify, something that can hamper the usefulness of predictive modelling.[4,5] Depending on the uncertainty in the data used to train the models, as well as the machine learning algorithms used, varying levels of uncertainty will be associated with the predictions generated by the model.[6,7] Methods that can be used to estimate the confidence associated with a prediction have therefore received increasing attention.

Model applicability domain has been a key concept for prediction reliability during many years and is still an active research field.[8] Also other methods have been applied to estimate the errors or control the confidence in predictors,[9] these methods include the use of experimental and predictive probability distributions,[10] Bayesian methods[11] including Gaussian processes,[12] reliability-density neighborhoods,[13] and ensemble model variance,[14] as well as the use of confidence predictors.[15]

Clearly, knowing the uncertainties associated with a prediction can be very helpful. Different propensity for error exists in any context where predictions are made, quantifying the uncertainty associated with a prediction allow for informed decisions based on the outcome. In a QSAR setting, it can allow prioritizing between different compounds based on the uncertainty associated with their predicted values of interest or to help guide the number of compounds that should be selected for experimental determination. For example, the user can choose to trust in predictions

with a low uncertainty while experimentally determine the values for instances with higher uncertainty.

In a regression context, a conformal predictor is a type of confidence predictor that outputs prediction ranges with a guaranteed maximum error rate corresponding to a user-defined confidence level.[15] For a conformal regression model at the 80 % confidence level, at least 80 % of all generated prediction ranges will include the correct value. This is achieved by comparing new instances to previous examples of known outcome through a *nonconformity* function. Conformal predictors can be applied both to classification and regression tasks and can be used with any underlying machine learning algorithm, thus making conformal prediction a flexible framework.

It is possible to define many different nonconformity measures, each defining a different conformal predictor. Conformal predictors can be evaluated using the concepts of *validity* and *efficiency*. A prediction from a conformal predictor is considered to be correct if the corresponding prediction range includes the correct value. The predictor is said to be *valid* if the overall frequency of errors does not exceed that of the chosen significance level (defined as 1 – the confidence level). The efficiency of a predictor measures the size of the prediction ranges. A predictor with smaller prediction ranges is said to be more efficient. Since a conformal predictor by design is always valid, the variability between different conformal predictors is mainly how efficient the predictor is.

An inductive conformal predictor utilizes a calibration set with known labels to infer new predictions, and this study will only consider this type of conformal predictor. In the simplest case, the prediction range from such a conformal regressor can be derived by evaluating the size of the residuals ($|y_i - \hat{y}_i|$) on the calibration set. The prediction range that includes the correct value for

a proportion of the calibration set corresponding to the set confidence level is then applied to all new predictions. The simplicity of this approach allows it to be combined with any regression algorithm. However, the problem with this approach is that different instances are generally associated with different levels of uncertainty. Rather than generating the same prediction ranges for every instance, it is desirable to scale the ranges to reflect the uncertainty of the individual prediction.

Since a conformal regressor relies on the known prediction errors from previous examples, a larger library of known instances will help making the prediction ranges more accurate and more efficient. A conformal predictor requires these calibration examples in addition to the instances used to train the underlying predictor, thus a larger number of instances are ideally required compared to using only the underlying predictor. In some cases, where data is scarce, this might limit the usefulness of the approach.

Earlier studies have investigated different methods for improving the tightness of the prediction regions and to normalize these to the individual uncertainties of new instances.[16–18] Papadopoulos, Vovk, and Gammerman[16] explored nonconformity functions based on *k* nearest neighbors, using either the distance from an example to the neighbors, the standard deviation of the neighbor labels or a combination of both. Johansson *et al.*[17] described an application of random forest based conformal regression where the out of bag examples are used to derive the prediction ranges either though a separate error prediction model or by using the out of bag error for the nearest neighbors, eliminating the need of a separate calibration set. However, none of these studies applied the methods to bioactivity prediction.

Conformal prediction has attracted increasing attention in recent years and several studies have investigated both theoretical aspects and applications.[19–32] These studies have highlighted several

strengths associated with conformal prediction, such as its excellent handling of imbalanced data,[26,27,33] and built-in definition of the applicability domain.[20,34] Previous studies have also applied conformal prediction to QSAR modelling.[24,29] Although these studies have shown the usefulness of conformal prediction as a basis for QSAR, a more systematic approach evaluating different conformal prediction methods in the context of QSAR is still missing in the literature, and further understanding on how different nonconformity functions can impact the prediction ranges afforded in bioactivity and property modelling is still required.

In this study, we investigated the performance of different conformal regression approaches for QSAR. We applied six distinct nonconformity measures based on different ways to estimate the uncertainty of predictions for individual instances. The performance of the different measures is evaluated on 29 datasets with bioactivity or molecular properties as output values. Special emphasis was put on investigating the effects of the different nonconformity measures on model efficiency, i.e. the size of the prediction ranges.

Methods

Data Sets

Table 1 describes the 29 datasets used in this study. The datasets were extracted from ChEMBL[35,36] (version 19) as described by Cortes-Ciriano and Bender[37]. Some of these datasets have also been used in previous studies.[20,38–42]

The structures where standardized using the IMI eTOX project standardizer[43] in combination with tautomer standardization using the MolVS standardizer[44]. All activity values where converted to pIC50 values (i.e. $-\log_{10}$ IC50). Chemical structures were encoded using 97 RDKit[45] descriptors

(see Supporting Information for full list of descriptors). These descriptors have shown good performance in previous studies.[26,46]

Table 1. Summary of the datasets used in this study.

Dataset	Number of compounds	Outcome/Target	Data Range	ChEMBL ID/reference
Properties				
AQUAX	1,277	Solubility in water (log S)	-11.6 – 1.6	Huuskonen[38]
AZ solubility	1,763	Solubility in buffer pH 7.4	-4 – 0.2	CHEMBL3301364
AZ LogD	4,197	Octanol/water distribution at pH 7.4	-1.5 – 4.5	CHEMBL3301363
Biological activity		(pIC50)		
F7	353	Factor VII	4.0 – 8.2	Chen *et al.*[39]
IL4	632	Interleukin 4	4.7 – 8.3	Chen *et al.*[39]
MMP2	533	Matrix metalloproteinase-2	5.1 – 10.3	CHEMBL333
hERG	4,325	hERG potassium ion channel	2.4 – 9.9	Czodrowski[47]
JAK1	804	Tyrosine-protein kinase JAK1	6.3 – 9.9	Chen *et al.*[39]
JAK2	608	Tyrosine-protein kinase JAK2	3.8 – 9.8	CHEMBL2971
GCR	748	Glucocorticoid receptor	4.0 – 10.4	CHEMBL2034
AR	688	Androgen Receptor	4.2 – 9.7	CHEMBL1871
Estrogen β	593	Estrogen receptor β	4.0 – 9.5	CHEMBL242
Estrogen α	649	Estrogen receptor α	4.1 – 9.7	CHEMBL206
DAT	910	Dopamin transporter	4.1 – 10.2	CHEMBL238
VEGF2	3,501	Vascular endothelial growth factor receptor 2	4.0 – 9.8	CHEMBL279
SERT	1,682	Serotonin transporter	4.0 – 10.6	CHEMBL228
SRC	1,882	Tyrosine-protein kinase SRC	2.3 – 9.9	CHEMBL267

PKCα	323	Protein kinase C alpha	4.0 – 9.4	CHEMBL299
PTK2	295	Focal adhesion kinase 1	4.0 – 9.3	CHEMBL2695
PR	908	Progesterone receptor	4.1 – 10.2	CHEMBL208
mTOR	1,014	Serine/threonine-protein kinase mTOR	4.0 – 10.1	CHEMBL2842
MAP ERK2	118	MAP kinase ERK2	4.1 – 9.7	CHEMBL4040
CDK2	808	Cyclin-dependent kinase 2	3.5 – 10	CHEMBL301
Aurora-A	745	Serine/threonine-protein kinase Aurora-A	4.0 – 9.8	CHEMBL4722
Vanilloid	962	Vanilloid receptor	4.0 – 9.8	CHEMBL4794
SELE	155	E-selectin	4.0 – 9.2	CHEMBL3890
MGLL	1,112	Monoacylglycerol lipase	4.8 – 8.4	Chen *et al.*[39]
PRSS2	315	Protease, serine, 2	4.0 – 7.8	Chen *et al.*[39]
Toxicity				
Tox pyriformis	1,083	toxicity against *T. pyriformis*	0.3 – 6.3	Sushko *et al.*[48]

Conformal regression and nonconformity measures

For an in-depth definition of the conformal prediction/regression methodology we refer the reader to Vovk, Gammerman, Shafer[15] and for a practical explanation of the implementation to Norinder *et al*.[20]

In conformal regression, we defined the nonconformity measure α as

$$\alpha_i = \frac{|y_i - \hat{y}_i|}{\lambda_i} \quad \text{(eq. 1)}$$

where y_i is the observed value, $\hat{y}_i$ the predicted value, and λ_i is a factor for scaling the prediction range for instance *i*.

In this study, we choose to apply scaling methods based on measures that in some aspect quantify the underlying uncertainty of the prediction for individual instances. The intention is to generate predictions where instances associated with a greater uncertainty have a wider prediction range. A number of different approaches to assess the uncertainty of a prediction were used to derive the denominator λ, these are discussed below and a summary of these methods are presented in Table 2.

Table 2. Descriptions of the different nonconformity functions used in this study.

Method	**Description**
Ensemble Standard Deviation (ESD)	The standard deviation of the predictions from the underlying ensemble of trees was used to assess uncertainty.
Ensemble interpercentile range (EIR)	The interpercentile range, 10^{th} to 90^{th} percentile to give the central 80 % range, from the ensemble predictions.
Error model (EM)	A separate model was used to predict the error of each instance.
Distance to training center (DTC)	The distance in two component PCA space to the center of all training data.
IPCA NN	Uncertainty was derived based on the average distance to the five NN in two dimensional IPCA.
t-SNE NN	Uncertainty was derived based on the average distance to the five NN in two dimensional t-SNE.

1. Ensemble based methods

The first approach to estimate the uncertainty was to use the distribution of the predictions from the ensemble. The ensemble standard deviation has previously been reported in the literature as being an effective indicator of the relative confidence in the prediction.[14,40,49–51] If there is a high

variance in the output from the underlying predictors, that indicates a greater uncertainty in the prediction. Two ways of integrating this information in the nonconformity score were applied, the standard deviation of the predictions from the individual predictors and the interpercentile range (the range of prediction values between two percentiles) of the ensemble predictions.

2. Error prediction model

Another way to measure the uncertainty that has been applied as a basis for nonconformity scores in previous studies is to apply a separate model to predict the size of the residual for each instance.[20,32] We applied a random forest (RF) based error model using the predicted error associated with instance *i* as λ_i.

3. Data distribution

Uncertainty can also be estimated based on how a new instance relates to the instances in the training set. If there are many instances with similar features in the training data, the expectation would be that the predictor performs better than when there are fewer instances with similar features. We therefore applied methods to define λ_i by taking into account the distance to and density of training examples in relation to the instance being predicted.

Our first approach was to create a two component PCA of the training data and project new instances in this space. The distance to the average object (PCA origin) was then calculated for the new instance and used as λ_i.

Predictive uncertainty based on density was derived from t-student distributed stochastic neighborhood embedding (t-SNE) and IPCA. The scikit-learn implementations were used for this calculation, with default settings. t-SNE has been demonstrated to conserve the multidimensional structure of the data (in terms of relative vicinity of instances) during its projection into a low-dimensional map.[52] The dimensionality reductions were employed to the same descriptors used as

input for the modelling, and the average distance to the five nearest neighbors in a two component reduced space were used as λ_i.

Weighting of the nonconformity score

For the nonconformity measures based on λ_i derived from the ensemble standard deviation, we also implemented a weighted expression defined as

$$\lambda_i = e^{w \cdot SD} \quad \text{(eq. 2)}$$

where w is a weighing factor. The exponent term of this non-conformity score was adapted from a non-conformity score reported by Papadopoulus *et al*.[16]

Machine Learning

All machine learning models were constructed using the scikit-learn[53] Python package. Default values for the parameters were used unless otherwise specified. For ensemble based methods 500 estimators were used.

Initial regression models were developed using RF, Lasso, Gradient Boosting, Ridge Regression, Bayesian Ridge, Adaptive Boosting (AdaBoost), Automatic Relevance Determination (ARD), Elastic Net, and partial least squares (PLS).

Inductive conformal predictors were based on the nonconformist[54] Python package and the code was modified according to the different nonconformity functions.

We applied the aggregated conformal predictor approach described by Carlsson *et al*.[22] using 200 iterations. In each iteration, the data was randomly split into 20 % test set, 20 % calibration set, and 60 % training set. This allows the split for test and calibration set to be carried out multiple times, thereby reducing the variability from different splits.

Evaluation of the predictors

The individual regression techniques were evaluated using R2 and RMSD. Although these metrics are useful to summarize the performance of models in the form of single numbers, they might not be useful as measures of the predictive ability in all settings, as has been demonstrated before.[55] Importantly, when evaluating models multiple parameters should always be considered.

We chose to evaluate the validity of the predictors by comparing the expected error rate, ε, to the observed error rate for the confidence levels 70-100 % at every 1 % step expressed as the 2-norm of the resulting error vector. This way, the validity over the most typical confidence levels are considered. A smaller value indicated that the error rates generated by the models more closely correspond to the set confidence level.

The efficiency was calculated as the average prediction range at the 80 % confidence level. Prediction ranges were calculated as the full span on both sides of the point prediction (i.e. plus and minus the uncertainty).

Visualization

All plots were made using matplotlib.[56] For all the boxplots, the box extends from the lower to upper quartile values of the data and the whiskers extends a further 1.5 times the interquartile range with all data points outside that range represented as individual dots.

Results and Discussion

We explored an array of different regression techniques on our datasets and the results indicate a higher performance (R^2 and RMSD) of ensemble-based models compared with single models (Figure 1; see Supporting Information for detailed values).[57,58]

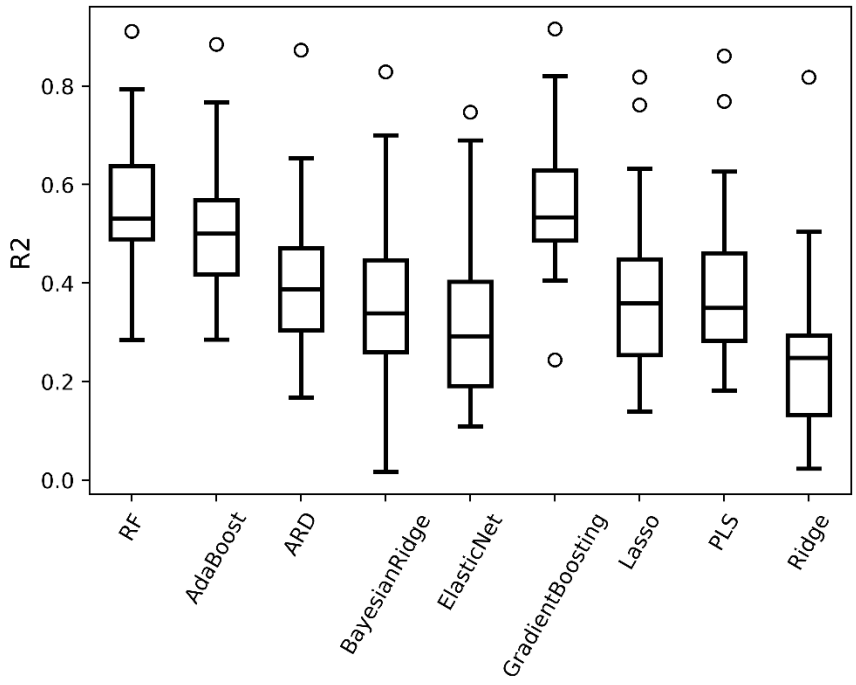

Figure 1. Distribution of R^2 values for different regression methods across all the datasets when evaluated on test data.

In the context of conformal prediction, ensemble models also offer the opportunity to use values generated from the underlying estimators to derive λ. Since RF had among the best overall performance in our study, and is a widely used algorithm with efficient implementations in most machine learning software, we chose to continue the analysis using RF as the underlying machine learning algorithm for our conformal predictors.

We evaluated different nonconformity measures (Table 2) with respect to model validity and average prediction range at the 80 % confidence level. Table 3 shows the validity of the different algorithms across all datasets. Although some differences in validity can be seen across the different nonconformity functions, all models produced an error rate that was very close to the defined significance level. This is gratifying, as it shows that the conformal predictors produce

valid models regardless of the choice of nonconformity function, in agreement with the underlying theory.

Table 3. Validity of the different conformal regression models (lower value indicates better correspondence between expected and observed error, see methods for details). All the tested nonconformity functions produced valid models with only small deviations from the expected error.

Dataset	**ESD**	**EIR**	**EM**	**DTC**	**IPCA NN**	**t-SNE NN**
Properties						
AQUAX	0.093	0.107	0.210	0.069	0.151	0.104
AZ solubility	0.075	0.087	0.142	0.042	0.135	0.103
AZ LogD	0.056	0.068	0.148	0.030	0.115	0.086
Biological activity						
F7	0.155	0.146	0.168	0.097	0.177	0.144
IL4	0.136	0.152	0.225	0.082	0.174	0.132
MMP2	0.113	0.115	0.204	0.041	0.148	0.144
hERG	0.056	0.054	0.046	0.030	0.124	0.084
JAK1	0.120	0.115	0.101	0.056	0.151	0.112
JAK2	0.093	0.100	0.210	0.049	0.123	0.122
GCR	0.107	0.117	0.185	0.056	0.134	0.102
AR	0.118	0.141	0.158	0.075	0.138	0.105
Estrogen β	0.124	0.153	0.180	0.075	0.085	0.126
Estrogen α	0.117	0.133	0.210	0.091	0.086	0.136
DAT	0.086	0.120	0.074	0.062	0.140	0.136
VEGF2	0.057	0.072	0.163	0.026	0.121	0.102

SERT	0.080	0.108	0.197	0.040	0.129	0.102
SRC	0.071	0.083	0.073	0.039	0.103	0.143
PKCα	0.129	0.157	0.147	0.105	0.119	0.101
PTK2	0.125	0.134	0.115	0.122	0.121	0.094
PR	0.093	0.108	0.093	0.068	0.113	0.124
mTOR	0.115	0.123	0.209	0.042	0.154	0.133
MAP ERK2	0.188	0.233	0.314	0.321	0.167	0.089
CDK2	0.115	0.106	0.187	0.086	0.140	0.135
Aurora-A	0.095	0.085	0.109	0.060	0.150	0.137
Vanilloid	0.132	0.130	0.089	0.059	0.128	0.128
SELE	0.221	0.282	0.318	0.183	0.274	0.234
MGLL	0.088	0.124	0.123	0.060	0.167	0.139
PRSS2	0.094	0.188	0.245	0.086	0.154	0.064
Toxicity						
Tox pyriformis	0.092	0.127	0.151	0.077	0.160	0.111

Figure 2 shows how the prediction ranges are affected by the confidence level. At high confidence levels the prediction range is wide in order to include the correct value in most cases, while for low confidence levels more errors are accepted and the prediction ranges can become narrower. We chose to evaluate the prediction ranges at the 80 % confidence level (i.e. expected error rate of 0.2) as in the authors' experience this represents a reasonable balance between predictor efficiency and expected error rate that would translate into actionable models for QSAR applications.

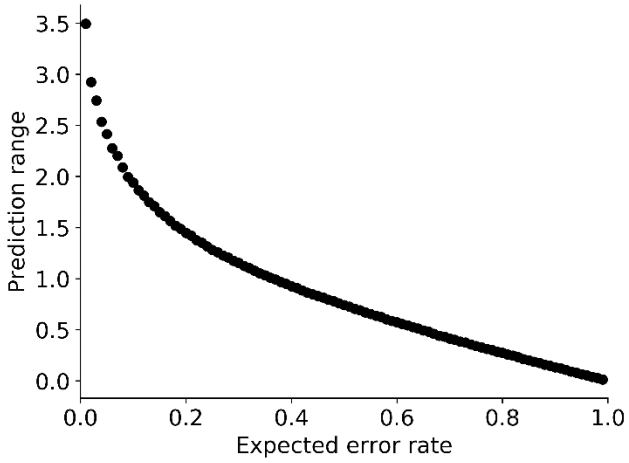

Figure 2. Average prediction ranges for different expected errors (based on the confidence level) on the dataset 'AQUAX' using the ESD based nonconformity function. This illustrates how the prediction range increases as the expected error (1 – confidence level) decreases.

Despite producing similar results in terms of validity, the efficiency of the applied nonconformity functions differed greatly with almost one log unit difference in average prediction range between the best performing nonconformity function and the worst (Figure 3, Figure 4, and Table 4). The performance of many of the individual models is somewhat poor; for many datasets even the best models produce an average prediction range at the 80 % confidence level spanning around 30 % of the full range of the dependent variable in the dataset. However, the much tighter prediction ranges achieved for some datasets, most notably AQUAX where the average prediction range spans about 10 % of the data range, indicate that the poor predictions associated with some datasets is not inherent to the method. However, it is important to stress the fact that the tightness of the prediction ranges is closely related to the quality of the underlying models and that many of the datasets that generate very wide prediction ranges also have poor RMSD (see Supporting Information).

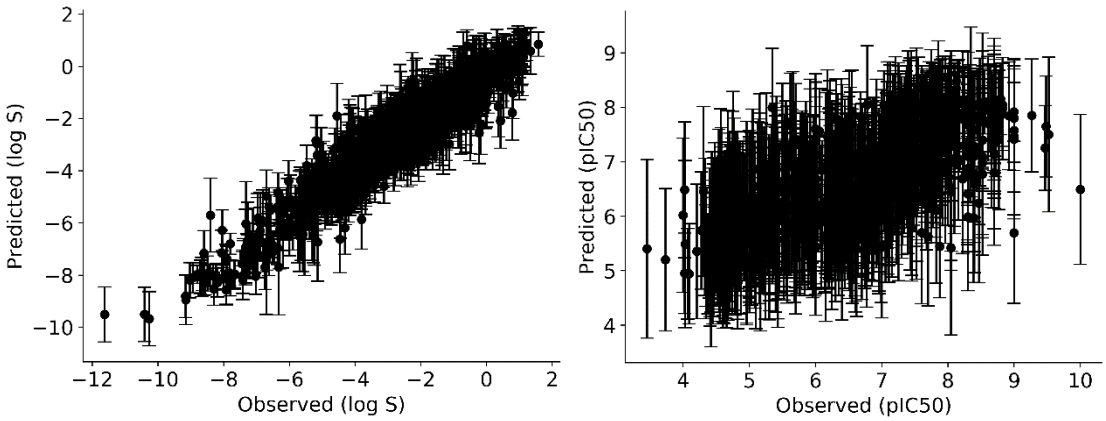

Figure 3. The prediction ranges using the ESD for the best performing (most efficient in relation to data range) dataset 'AQUAX' (left) and the worst performing 'CDK2' (right). There is a clear difference in terms of the usefulness of the generated predictions.

Based on the results in Table 4, clearly the nonconformity function has a pronounced impact on the prediction range. The difference between the tightest prediction range and the widest is often close to one log unit, enough to have substantial implications for the usability of the model. Overall the conformal predictors deriving the uncertainty from the underlying ensemble model had the highest efficiency with a slight preference for using the ensemble standard deviation (see Supporting Information for detailed comparison of the performance). Inspection of the datasets with data on biological activity reveals that most of the average prediction ranges fall between one and two pIC50 units at the 80 % confidence level. This can be put in context by comparing to the experimental uncertainty in ChEMBL IC50 data that has been estimated to have a SD of 0.68 and a mean unsigned error of 0.55 pIC50 units.[59] The underlying experimental uncertainty is important to keep in mind as it limits the prediction accuracy that can be achieved.[37,42,60]

Table 4. The mean prediction ranges at the 80 % confidence level. The most efficient method (generating the tightest prediction rages) was ESD.

Dataset	ESD	EIR	EM	DTC	IPCA NN	t-SNE NN
Properties						
AQUAX	1.45	1.48	1.57	1.77	2.29	1.73
AZ solubility	2.07	2.11	2.15	2.56	2.81	3.20
AZ LogD	1.82	1.87	1.91	2.37	2.70	2.54
Biological activity						
F7	1.38	1.38	1.44	1.64	1.88	1.47
IL4	0.91	0.91	0.97	1.15	1.34	1.04
MMP2	1.48	1.51	1.58	2.02	2.10	1.72
hERG	1.44	1.45	1.50	1.92	2.25	3.26
JAK1	1.46	1.48	1.44	1.74	1.86	1.62
JAK2	2.03	2.11	2.09	2.51	2.46	2.20
GCR	1.54	1.61	1.66	1.71	3.23	1.87
AR	1.70	1.75	1.77	1.97	2.63	1.93
Estrogen β	1.88	1.99	1.95	2.17	3.39	2.14
Estrogen α	1.86	1.91	1.96	2.05	3.07	2.08
DAT	1.89	1.95	1.93	2.46	3.28	2.20
VEGF2	2.08	2.16	2.19	2.60	2.90	3.81
SERT	1.83	1.91	1.97	2.40	2.77	2.36
SRC	1.89	1.95	1.91	2.32	3.55	2.53
PKCα	1.95	2.01	2.10	2.46	2.81	2.08
PTK2	1.78	1.76	1.79	2.27	2.84	1.91
PR	1.51	1.61	1.59	2.09	2.20	1.81
mTOR	1.92	1.97	1.99	2.64	2.76	2.26
MAP ERK2	1.94	2.01	2.09	2.72	2.36	1.91

CDK2	2.14	2.19	2.25	2.57	2.83	2.43
Aurora-A	1.87	1.90	1.88	2.47	2.35	2.20
Vanilloid	1.91	1.94	1.87	2.29	2.42	2.18
SELE	2.10	2.16	2.09	2.56	7.18	2.17
MGLL	1.98	2.01	1.94	2.50	2.88	2.23
PRSS2	0.97	0.98	1.12	1.43	1.30	1.05
Toxicity						
Tox pyriformis	1.01	1.04	1.07	1.26	1.45	1.21

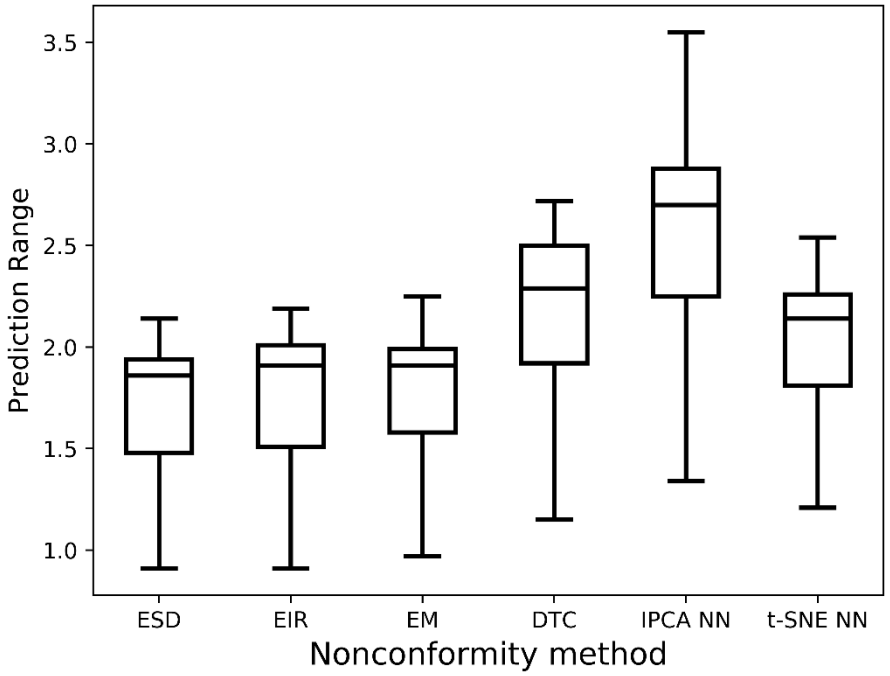

Figure 4. Boxplots reporting the average prediction ranges generated by the different nonconformity methods across all datasets considered (data in Table 4). (Outliers not shown)

To fine tune the resulting prediction ranges we also applied different scaling factors to the uncertainty measure (eq. 2). This approach has been shown to be beneficial in previous studies.[16] This was done using the standard deviation of the ensemble since this was the best performing nonconformity function. The results for applying different scaling factors is shown in Figure 5 and Table 5.

Table 5. Prediction ranges using different scaling factors based on the ensemble standard deviation (eq. 2). Best average performance was obtained using w=1.

Dataset w =	**0.05**	**0.25**	**0.5**	**0.75**	**1**	**1.25**
Properties						
AQUAX	1.49	1.46	1.46	1.41	1.40	1.40
AZ solubility	2.07	2.07	2.05	2.05	2.03	2.02
AZ LogD	1.90	1.87	1.84	1.81	1.77	1.74
Biological activity						
F7	1.40	1.39	1.37	1.34	1.32	1.34
IL4	0.91	0.90	0.90	0.89	0.89	0.90
MMP2	1.60	1.54	1.51	1.45	1.40	1.39
hERG	1.50	1.46	1.44	1.41	1.39	1.38
JAK1	1.43	1.43	1.42	1.42	1.41	1.43
JAK2	2.03	2.00	1.97	1.96	1.92	1.93
GCR	1.62	1.58	1.57	1.52	1.49	1.50
AR	1.73	1.69	1.67	1.65	1.62	1.64
Estrogen β	1.88	1.89	1.87	1.82	1.79	1.79
Estrogen α	1.91	1.89	1.86	1.80	1.77	1.78
DAT	1.92	1.90	1.86	1.84	1.83	1.84
VEGF2	2.10	2.07	2.05	2.03	2.02	2.01

SERT	1.86	1.83	1.81	1.77	1.76	1.75
SRC	1.92	1.91	1.89	1.84	1.83	1.82
PKCα	2.04	2.02	1.94	1.90	1.85	1.91
PTK2	1.78	1.72	1.70	1.66	1.66	1.67
PR	1.58	1.55	1.52	1.49	1.47	1.47
mTOR	1.87	1.85	1.82	1.78	1.80	1.80
MAP ERK2	1.73	1.72	1.71	1.68	1.64	1.90
CDK2	2.12	2.12	2.05	2.04	2.01	2.03
Aurora-A	1.90	1.91	1.83	1.81	1.80	1.82
Vanilloid	1.85	1.86	1.84	1.82	1.82	1.84
SELE	1.90	1.93	1.89	1.91	1.86	2.06
MGLL	1.91	1.91	1.90	1.89	1.89	1.92
PRSS2	1.00	0.99	0.98	0.97	0.96	1.00
Toxicity						
Tox pyriformis	1.08	1.05	1.04	1.03	1.00	0.99

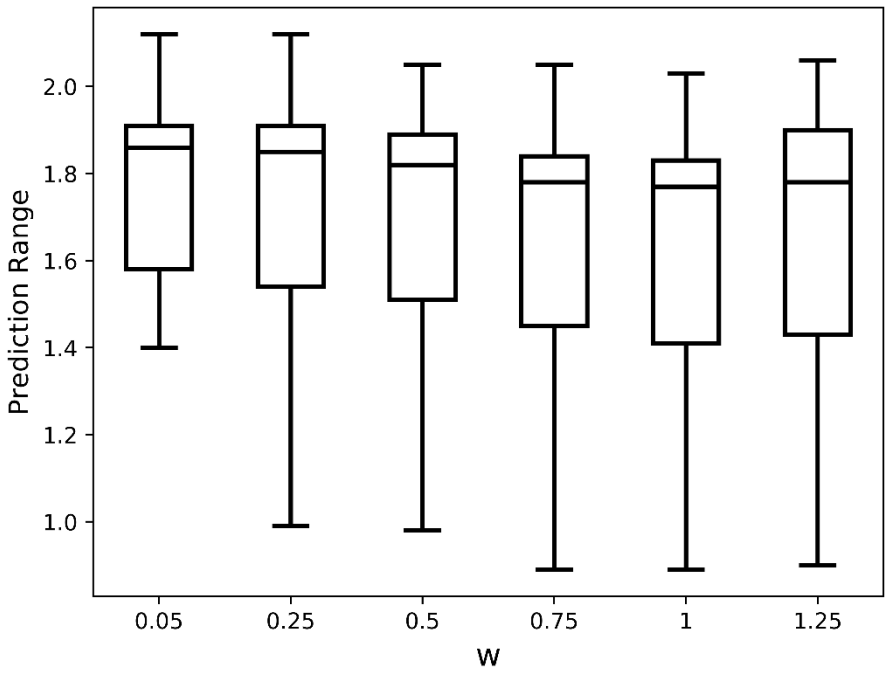

Figure 5. Distribution of average prediction ranges for different weight factors across all the datasets (data in Table 5). (Outliers not shown)

The results from the weighted functions showed a clear trend, where reducing the weighting factor, *w*, produced a larger spread of the mean prediction ranges. This means that, as *w* increases, some of the datasets become more efficiently predicted, which can be seen by the marked decrease of the lower limit of the distribution. This, in turn, may represent important improvements in the ability to produce useful predictions ranges in the context of experimental variability. The inverse relationship between *w* and efficiency might be explained by the fact that larger *w* values produce larger absolute differences across the scale of SD values (e.g. SD values of 0.1 and 0.9 show fold differences of 2.2 and 1.1 for w=1 and w=0.1, respectively). This means that increasing the value of *w* dilutes the differences in disagreement rate within any given regression ensemble. The best performance was obtained with *w* set to 1, this was also slightly better than using the standard deviation without the exponential scaling, producing average prediction ranges of 1.63 and 1.66, respectively, across all datasets (Table 4 and Table 5).

Overall, our results show how different conformal regressors can be used to generate QSAR models with an associated level of confidence. The best results were obtained using conformal regressors based on ensemble models and scaling the nonconformity score using the standard deviation of the individual predictors in the ensemble, but using the ensemble interpercentile range or a separate error model also produced similar results. Aside from generating the tightest prediction ranges, the ensemble standard deviation is an effective metric to use in the sense that the data is intrinsically generated by the underlying model. However, since the performance of the conformal models is dependent on how well the uncertainty of an individual prediction can be estimated it is likely that the best performance can be obtained by specifically tailoring the

nonconformity function for the problem and dataset at hand, likewise can tailoring of the underlying predictive model likely lead to improvements in the efficiency. Still, the above-mentioned approach delivered high performing and robust results across all the analyzed datasets.

Conclusions

Conformal regression offers a number of benefits when used for QSAR modeling, the most apparent being that the generated prediction regions come with a statistical guarantee that allows for confidence in the predictions.

In this study, we applied conformal regression to model 29 different datasets and evaluated several ways to calculate nonconformity scores. All methods consistently generated valid predictions but varied greatly in terms of efficiency, illustrating the importance of choosing a suitable nonconformity score. The most efficient models (delivering the tightest prediction ranges) were obtained using the natural exponential of the random forest ensemble standard deviation in the nonconformity function but other approaches were almost as efficient. The most efficient predictions generated an average prediction range of 1.65 pIC50 units when predicting the bioactivity of ChEMBL datasets at the 80 % confidence level.

Supporting Information

List of descriptors used for model training. Performance of different ML algorithms.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Corresponding Author

*fs447@cam.ac.uk

Acknowledgements

The research at Swetox (UN) was supported by Knut & Alice Wallenberg Foundation and Swedish Research Council FORMAS.

ICC has received funding from the European Union's Framework Programme For Research and Innovation Horizon 2020 (2014-2020) under the Marie Curie Sklodowska-Curie Grant Agreement No. 703543 (I.C.C.).

This project was financially supported by the Swedish Foundation for Strategic Research.

Conflict of Interest Statement

ICC holds equity interest in Evariste technologies.

Abbreviations

DTC, Distance to Training Center; EIR, Ensemble Interpercentile Range; EM, Error Model (EM); ESD, Ensemble Standard Deviation; RF, Random Forest; SD, Standard Deviation

References

(1) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010.

(2) Dearden, J. C. The History and Development of Quantitative Structure-Activity Relationships (QSARs). *IJQSPR* **2016**, *1*, 1–44.

(3) Cumming, J. G.; Davis, A. M.; Muresan, S.; Haeberlein, M.; Chen, H. Chemical Predictive Modelling to Improve Compound Quality. *Nat. Rev. Drug Discovery* **2013**, *12*, 948–962.

(4) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20*, 269–276.

(5) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **2010**, *29*, 476–488.

(6) Cortes-Ciriano, I.; Bender, A. Improved Chemical Structure–Activity Modeling Through Data Augmentation. *J. Chem. Inf. Model.* **2015**, *55*, 2682–2692.

(7) Roy, K.; Ambure, P.; Aher, R. B. How Important Is to Detect Systematic Error in Predictions and Understand Statistical Applicability Domain of QSAR Models? *Chemom. Intell. Lab. Syst.* **2017**, *162*, 44–54.

(8) Hanser, T.; Barber, C.; Marchaland, J. F.; Werner, S. Applicability Domain: Towards a More Formal Definition. *SAR QSAR Environ. Res.* **2016**, *27*, 865–881.

(9) Bosnić, Z.; Kononenko, I. Comparison of Approaches for Estimating Reliability of Individual Regression Predictions. *Data Knowl. Eng.* **2008**, *67*, 504–516.

(10) Wood, D. J.; Carlsson, L.; Eklund, M.; Norinder, U.; Stålring, J. QSAR with Experimental and Predictive Distributions: An Information Theoretic Approach for Assessing Model Quality. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 203–219.

(11) Lazic, S.; Edmunds, N.; Pollard, C. Predicting Drug Safety and Communicating Risk: Benefits of a Bayesian Approach. *Toxicol. Sci.* **2017**, *162*, 89–98.

(12) Cortes-Ciriano, I.; van Westen, G. J. P.; Lenselink, E. B.; Murrell, D. S.; Bender, A.; Malliavin, T. Proteochemometric Modeling in a Bayesian Framework. *J. Cheminform.* **2014**, *6*, 35.

(13) Aniceto, N.; Freitas, A. A.; Bender, A.; Ghafourian, T. A Novel Applicability Domain Technique for Mapping Predictive Reliability across the Chemical Space of a QSAR: Reliability-Density Neighbourhood. *J. Cheminform.* **2016**, *8*, 69.

(14) 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.

(15) Vovk, V.; Gammerman, A.; Shafer, G. Algorithmic Learning in a Random World; Springer: New York, 2005; pp 1–324.

(16) Papadopoulos, H.; Vovk, V.; Gammerman, A. Regression Conformal Prediction with Nearest Neighbours. *J. Artif. Intell. Res.* **2011**, *40*, 815–840.

(17) Johansson, U.; Boström, H.; Löfström, T.; Linusson, H. Regression Conformal Prediction with Random Forests. *Mach. Learn.* **2014**, *97*, 155–176.

(18) Papadopoulos, H.; Haralambous, H. Reliable Prediction Intervals with Regression Neural Networks. *Neural Networks* **2011**, *24*, 842–851.

(19) Carlsson, L.; Ahlberg, E.; Boström, H.; Johansson, U.; Linusson, H. Modifications to P-Values

of Conformal Predictors. In *Statistical Learning and Data Sciences: Third International Symposium, SLDS 2015, Egham, UK, April 20-23, 2015, Proceedings*; Gammerman, A., Vovk, V., Papadopoulos, H., Eds.; Springer International Publishing: Cham, 2015; pp 251–259.

(20) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **2014**, *54*, 1596–1603.

(21) Vovk, V. Conditional Validity of Inductive Conformal Predictors. *Mach. Learn.* **2013**, *92*, 349–376.

(22) Carlsson, L.; Eklund, M.; Norinder, U. Aggregated Conformal Prediction. In *Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings*; Iliadis, L., Maglogiannis, I., Papadopoulos, H., Sioutas, S., Makris, C., Eds.; Springer International Publishing: Berlin, Heidelberg, 2014; pp 231–240.

(23) Norinder, U.; Boyer, S. Conformal Prediction Classification of a Large Data Set of Environmental Chemicals from ToxCast and Tox21 Estrogen Receptor Assays. *Chem. Res. Toxicol.* **2016**, *29*, 1003–1010.

(24) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. The Application of Conformal Prediction to the Drug Discovery Process. *Ann. Math. Artif. Intell.* **2013**, *74*, 117–132.

(25) Linusson, H.; Johansson, U.; Boström, H.; Löfström, T. *Efficiency Comparison of Unstable Transductive and Inductive Conformal Classifiers*; 2014; Vol. 437.

(26) Svensson, F.; Norinder, U.; Bender, A. Modelling Compound Cytotoxicity Using Conformal Prediction and PubChem HTS Data. *Toxicol. Res. (Camb).* **2017**, *6*, 73–80.

(27) Löfström, T.; Boström, H.; Linusson, H.; Johansson, U. Bias Reduction through Conditional Conformal Prediction. *Intell. Data Anal.* **2015**, *19*, 1355–1375.

(28) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling for Regulatory Purposes. A Transparent and Flexible Alternative to Applicability Domain Determination. *Regul. Toxicol. Pharmacol.* **2015**, *71*, 279–284.

(29) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. Application of Conformal Prediction in QSAR. In *IFIP Advances in Information and Communication Technology*; 2012; Vol. 382 AICT, pp 166–175.

(30) Johansson, U.; Ahlberg, E.; Boström, H.; Carlsson, L.; Linusson, H.; Sönströd, C. Handling Small Calibration Sets in Mondrian Inductive Conformal Regressors. In *Statistical Learning and Data Sciences: Third International Symposium, SLDS 2015, Egham, UK, April 20-23, 2015, Proceedings*; Gammerman, A., Vovk, V., Papadopoulos, H., Eds.; Springer International Publishing: Cham, 2015; pp 271–280.

(31) Cortés-Ciriano, I.; Bender, A.; Malliavin, T. Prediction of PARP Inhibition with Proteochemometric Modelling and Conformal Prediction. *Mol. Inform.* **2015**, *34*, 357–366.

(32) Cortés-Ciriano, I.; van Westen, G. J. P.; Bouvier, G.; Nilges, M.; Overington, J. P.; Bender, A.; Malliavin, T. E. Improved Large-Scale Prediction of Growth Inhibition Patterns Using the NCI60 Cancer Cell Line Panel. *Bioinformatics* **2016**, *32*, 85–95.

(33) Norinder, U.; Boyer, S. Binary Classification of Imbalanced Datasets Using Conformal Prediction. *J. Mol. Graph. Model.* **2017**, *72*, 256–265.

(34) Norinder, U.; Rybacka, A.; Andersson, P. L. Conformal Prediction to Define Applicability Domain – A Case Study on Predicting ER and AR Binding. *SAR QSAR Environ. Res.* **2016**, *27*, 303–316.

(35) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

(36) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2016**, *45*, D945–D954.

(37) Cortés-Ciriano, I.; Bender, A. How Consistent Are Publicly Reported Cytotoxicity Data? Large-Scale Statistical Analysis of the Concordance of Public Independent Cytotoxicity Measurements. *ChemMedChem* **2016**, *11*, 57–71.

(38) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773–777.

(39) Chen, H.; Carlsson, L.; Eriksson, M.; Varkonyi, P.; Norinder, U.; Nilsson, I. Beyond the Scope of Free-Wilson Analysis: Building Interpretable QSAR Models with Machine Learning Algorithms. *J. Chem. Inf. Model.* **2013**, *53*, 1324–1336.

(40) Tetko, I. V; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, 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.

(41) Cortes-Ciriano, I. Benchmarking the Predictive Power of Ligand Efficiency Indices in QSAR. *J. Chem. Inf. Model.* **2016**, *56*, 1576–1587.

(42) Cortes-Ciriano, I.; Bender, A.; Malliavin, T. E. Comparing the Influence of Simulated Experimental Errors on 12 Machine Learning Algorithms in Bioactivity Modeling Using 12 Diverse Data Sets. *J. Chem. Inf. Model.* **2015**, *55*, 1413–1425.

(43) IMI eTOX project standardizer version 0.1.7, https://pypi.python.org/pypi/standardiser

(44) MolVS standardizer version 0.0.9, https://pypi.python.org/pypi/MolVS

(45) RDKit version 2016_03_01: Open-source cheminformatics, http://www.rdkit.org

(46) Svensson, F.; Norinder, U.; Bender, A. Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 439–444.

(47) Czodrowski, P. hERG Me Out. *J. Chem. Inf. Model.* **2013**, *53*, 2240–2251.

(48) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.;

Radchenko, E. V; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 533–554.

(49) Sushko, Y.; Novotarskyi, S.; Körner, R.; Vogt, J.; Abdelaziz, A.; Tetko, I. V. Prediction-Driven Matched Molecular Pairs to Interpret QSARs and Aid the Molecular Optimization Process. *J. Cheminform.* **2014**, *6*, 48.

(50) Tetko, I. V; Novotarskyi, S.; Sushko, I.; Ivanov, V.; Petrenko, A. E.; Dieden, R.; Lebon, F.; Mathieu, B. Development of Dimethyl Sulfoxide Solubility Models Using 163 000 Molecules: Using a Domain Applicability Metric to Select More Reliable Predictions. *J. Chem. Inf. Model.* **2013**, *53*, 1990–2000.

(51) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K.-R.; Xi, L.; Liu, H.; Yao, X.; Öberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V; Tetko, I. V. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **2010**, *50*, 2094–2111.

(52) Van Der Maaten, L.; Hinton, G.; van der Maaten, G. H. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

(53) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(54) nonconformist package version 1.2.5, https://github.com/donlnz/nonconformist

(55) Roy, K.; Das, R. N.; Ambure, P.; Aher, R. B. Be Aware of Error Measures. Further Studies on Validation of Predictive QSAR Models. *Chemom. Intell. Lab. Syst.* **2016**, *152*, 18–33.

(56) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 99–104.

(57) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.

(58) Merkwirth, C.; Mauser, H.; Schulz-Gasch, T.; Roche, O.; Stahl, M.; Lengauer, T. Ensemble Methods for Classification in Cheminformatics. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1971–1978.

(59) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC50 Data – A Statistical Analysis. *PLoS One* **2013**, *8*, e61007.

(60) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public Ki Data. *J. Med. Chem.* **2012**, *55*, 5165–5173.

For Table of Contents Use Only