JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination

Ulf Norinder,*[,†] Lars Carlsson,[‡] Scott Boyer,[‡,⊥] and Martin Eklund[‡,§]

[†]H. Lundbeck A/S, Ottiliavej 9, 2500 Valby, Denmark

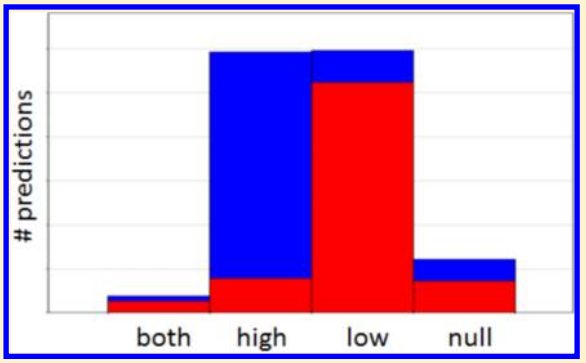
[‡]AstraZeneca Research and Development, SE-431 83 Mölndal, Sweden

[§]Department of Surgery, University of California at San Francisco (UCSF), San Francisco, California 94115, United States

[⊥]Swedish Toxicology Sciences Research Center, SE-151 36 Södertälje, Sweden

Supporting Information

ABSTRACT: Conformal prediction is introduced as an alternative approach to domain applicability estimation. The advantages of using conformal prediction are as follows: First, the approach is based on a consistent and well-defined mathematical framework. Second, the understanding of the confidence level concept in conformal predictions is straightforward, e.g. a confidence level of 0.8 means that the conformal predictor will commit, at most, 20% errors (i.e., true values outside the assigned prediction range). Third, the confidence level can be varied depending on the situation where the model is to be applied and the consequences of such changes are readily understandable, i.e. prediction ranges are increased or decreased, and the changes can immediately be inspected. We demonstrate the usefulness of conformal prediction by applying it to 10 publicly available data sets.



INTRODUCTION

Informed decisions based on predictions from a Quantitative Structure−Activity Relationship (QSAR) model are frequently confounded by a poor understanding of the reliability of the prediction for the example of interest. Substantial efforts have been devoted to research on this topic within the QSAR community over the past decade, and a number of methods have been suggested for estimating the confidence of QSAR predictions. These confidence estimates are typically based on the loosely defined concept of a QSAR model's "applicability domain" (AD).[1−11] In general terms, the assumption is that the farther away a molecule is from a QSAR model's AD, the less reliable the prediction.

At the core of the applicability domain assumption is the concept of a new example being placed in a multidimensional space. This space is sometimes, but not always, defined by a mathematical function that has been used to estimate a probability distribution of the "known" examples relative to an "activity" variable (in QSAR) such that new examples can be located in that multidimensional space based on their properties (as defined by the original model variables and the "landscape") and thereby be assigned an activity (logP, nonmutagen, endocrine disruptor, pIC50, etc.).

The problem with the current approaches to estimating the prediction confidence in QSAR models is that their interpretation is ambiguous. For example, what does it mean that a chemical compound is within or outside the QSAR model's AD by a certain amount according to a certain measure? How do the different metrics used to define the AD relate to each other? Does one AD metric apply throughout the multidimensional space defined by the model despite the fact that we know that both the local density of examples and the local "topology" differs considerably? Consequently, a large number of methods to manage this phenomenon have been developed.[12] The ambiguity in AD estimations arises from the ability of different AD metrics to capture "nearness" and assumes that relative proximity (according to the AD measure) equates to relatively higher accuracy. While this makes intuitive sense, the question (and the genesis of the ambiguity) is 'How close is close enough for an accurate enough prediction?'. To the best of our knowledge, previous methods have only been empirically validated and with little explanation as to why their accuracy or correlation to prediction error of the reliability estimates, in most cases, have been inconsistent although the works of Clark[13] and Sheridan,[11] respectively, have attempted to provide performance measures in a more quantitative way.

What we ideally would like to know is in fact that a particular prediction is derived from an area of property space from which reliable predictions are to be expected. In the following section

Received: February 24, 2014

 © XXXX American Chemical Society dx.doi.org/10.1021/ci5001168 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

we introduce conformal prediction as a solution to the problems with domain applicability estimation (actually rendering the concept of an AD unnecessary) and subsequently apply the method to 10 publicly available data sets.

CONFORMAL PREDICTION

We will in this section introduce the conformal predictor, with a focus on informally explaining the idea behind it. For a more formal description and proofs we refer to Vovk et al.[14] and for some initial, more mathematically oriented, work in the QSAR domain to Eklund et al.[17,15]

We will begin by introducing conformal prediction in the *online transductive* setting, where it is most straightforwardly and naturally applied. We will then describe how conformal prediction can be used in an *off-line inductive* modeling situation, which is the typical case for QSAR applications and which is how we apply it in this paper. In transductive learning, predictions are made for each new example at a time, the true value of the predicted example is then learned and the prediction rule is updated before the next example is predicted. A new prediction is, in online transductive learning therefore, based on all the previous available examples. This contrasts to the off-line inductive framework, where a batch of old examples is used to estimate a prediction rule. This prediction rule is subsequently applied to all new examples.

A *confidence predictor* is a prediction algorithm that outputs a *prediction region*, which contrasts to the *single valued (regression) or single label (classification) predictions* output by standard prediction algorithms, e.g. support vector machines (SVM) and random forest. A prediction region is a set Γ^{ϵ} that contains the true value of an example with probability at least $1-\varepsilon$. In the case of regression, the prediction region is a real valued interval (or set of intervals). For a binary classifier with the two classes represented by A and B, the prediction region could be any of the following sets: $\{A\}$, $\{B\}$, $\{A,B\}$ (both) or $\{\text{null}\}$ (the empty set).

To evaluate the performance of confidence predictors, we introduce the concepts of validity and efficiency. A confidence predictor is said to be *valid* if the frequency of errors (i.e., the fraction of true values outside the prediction region) it commits does not exceed ε at a chosen confidence level $1-\varepsilon$, and *efficient* if the prediction regions it outputs are as small as possible. A prediction region smaller than another is said to be *more efficient*. In classification problems, a measure of efficiency of confidence predictors could be the number of prediction sets containing two or more labels (at a given confidence level). In regression problems, a natural measure of efficiency is the size of the prediction interval.

A *conformal predictor* is a type of confidence predictor. The prediction regions produced by a conformal predictor have the property of always being valid. This means that we only need to worry about their efficiency. This very attractive property holds under the *randomness assumption*, which means that the examples are independently drawn from the same distribution (in fact, the slightly weaker *exchangeability* assumption—meaning that the observations not necessarily are independent but that they do not follow any particular order—is enough to guarantee validity). Note that this assumption is also made for most standard prediction algorithms used in QSAR (e.g., SVM or random forest), i.e. we are not introducing any new assumptions on top of the ones we generally already use.

To construct a conformal predictor's prediction regions, we need to define a *nonconformity measure*. Intuitively, this is a way of measuring how different a new example is from old examples (a nonconformity measure thus serves the same purpose as an AD measure, and most AD measures can be used as nonconformity measures). For example, in a regression problem we may define the nonconformity of a new example (x_n,y_n) as the absolute value of the difference between y_n and the predicted value $\hat{y}_n$ (*i.e.* the unsigned error) calculated from x_n and the old examples. Similarly, in a classification setting, a nonconformity measure may be the predicted probability of an example to belong to a given class. However, a nonconformity measure can also be more complex and take the expected variance or prediction error into consideration (see e.g. Papadopoulos and Haralambous[16]). Given a nonconformity measure, we can compute the *nonconformity score* α of each new example.

To assess how different a new example x_n is from old examples, we need to compare α_n to the nonconformity scores α_j of the previous examples x_j, $j = 1,\ldots,n-1$. To make this comparison we compute the fraction

$$|\{j = 1, \ldots, n: \alpha_j \geq \alpha_n\}|/n \quad (1)$$

i.e. the number of α_j that are as large or larger than α_n divided by the total number of α_j which gives the fraction of examples that have a larger α, e.g. a measure related to the unsigned error, than the new example under investigation. This fraction is called the *p-value* (not to be confused with traditional p-values from statistics) for the new example (x_i, y_i). If the p-value is small (close to $1/n$), then (x_i, y_i) is very nonconforming, i.e. the new example is different from previous examples because of its high α score compared to most of the examples in the training set. On the other hand, if the p-value is large (close to 1), then (x_i, y_i) is very conforming, i.e. very similar to the previous examples.

To obtain the prediction region from a conformal predictor, we assume that the label y_n of a new example x_n will have a value that makes (x_n, y_n) conform with the previous examples. The level of significance ε determines the level of conformity (as measured by the p-value) that we require. Thus, the prediction set from a conformal predictor with a given nonconformity measure and a given significance level is obtained by setting $\Gamma^{\epsilon}(x_1,y_1,\ldots,x_{n-1},y_{n-1},x_n)$ equal to the set $y \in Y$ such that

$$|\{j = 1, \ldots, n: \alpha_j \geq \alpha_n\}|/n > \varepsilon \quad (2)$$

The prediction region is thus defined by the set $y \in Y$ that fulfill criterion (2), i.e. the values in Y that has a p-value greater than ε. To emphasize the validity of the conformal predictor, we remark that the true value of the new example will be within the predicted region with the probability $1 - \varepsilon$.

Conformal prediction as we have discussed it up until now relies on the idea that the prediction rule is updated after each new example (online transductive modeling), which becomes computationally very demanding. Not only because it requires a computational cost similar to that of training a standard machine learning model on the same number of examples, but we also have to evaluate all labels or regions of the example that can lead to different nonconformity scores. *Off-line inductive conformal prediction* (ICP) was introduced as a way to address the computational overhead of conformal prediction.[18,19]

In ICP we start with a training set of examples $(x_1,y_1),\ldots,(x_l,y_l)$. We divide the training set into a *proper training set* $(x_1,y_1,\ldots,x_m,y_m)$ and a *calibration set* $(x_{m+1}, y_{m+1},\ldots, x_l, y_l)$, where m

dx.doi.org/10.1021/ci5001168 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

$< l$. We define the *working set* (external test set) to be the examples we want to predict, i.e. $(x_{l+1},...,x_{l+k})$. This contrasts to the online transductive setting, where, as we have already seen, all training examples are used to calculate nonconformity scores. The ICP prediction region is then obtained by setting $\Gamma^{\epsilon}(x_1,y_1,...,x_l,y_l,x_i)$ equal to the set $y \in Y$ such that

$$|\{j = m + 1, ..., l: \alpha_j \geq \alpha_i\}|/(l - m + 1) > \varepsilon \quad (3)$$

for each working example (x_i,y_i), $i = l+1,..., l+k$. where l is the number of examples in the calibration set. The fraction $|\{j = m+1,...,l: \alpha_j \geq \alpha_i\}|/(l-m+1)$ is the number of α_j in the calibration set that are as large or larger than a new α from the working set (α_i) divided by the total number of observations in the calibration set. The difference to the inductive setting is thus that we define the p-value using *only* the observations in the calibration set.

Equation 3 gives the fraction of examples in the calibration set that have a larger α than the new example to be predicted. The validity of conformal prediction is only guaranteed on average. However, some categories of observations (for example the observations belonging to a certain class in the classification setting) may be more difficult to predict than others, i.e. we may have higher error rate than ε in some categories of observations and lower than ε in others. *Mondrian conformal prediction* was developed to resolve this issue.[13] For Mondrian conformal predictors, we define a set of mutually exclusive categories that examples may belong to and treat each category individually with respect to the comparison of nonconformity scores (see section "Pseudocode for Using the Inductive Conformal Predictions" for a more detailed example of use). For example, if we let the categories be the class labels in a binary classification setting (this is how we use Mondrian conformal prediction in this article), we obtain prediction sets from a label-wise inductive Mondrian conformal predictor by setting $\Gamma^{(\epsilon k:k\in\{0,1\})}(x_1,y_1,...,x_l,y_l,x_i)$ equal to the set $y \in Y$ such that

$$|\{j: K_j = K_i \wedge \alpha_j \geq \alpha_i\}|/|\{j: K_j = K_i\}| > \varepsilon_{K(i,(x_i,y))} \quad (4)$$

where $j = m+1,...,l$, and κ_j is the category observation j belongs to, and κ_i is the assigned category to which observation i belongs. Essentially, Mondrian conformal prediction is conformal prediction applied separately to predefined categories (classes) of observations.

MATERIALS AND METHODS

Data Sets. Ten publicly available data sets were used (see Table 1 for a list of names and end points as well as references and Supporting Information for SMILES and end point activities). The classes of the binary classification data sets are balanced. The structures were described using signature descriptors of signature heights 0−3.[20]

Application of Inductive Conformal Prediction. Random Forest (RF) was used as the underlying learning method in this study.[21]

The data sets were randomly split into a training set (80%) and an external test set (working set) (20%). The training sets were further randomly divided into a "proper" training set (70%) and a calibration set (30%). The proper training set was used for model fitting, and the calibration set for computing the nonconformity scores was used for constructing the prediction regions. This procedure was repeated 100 times.

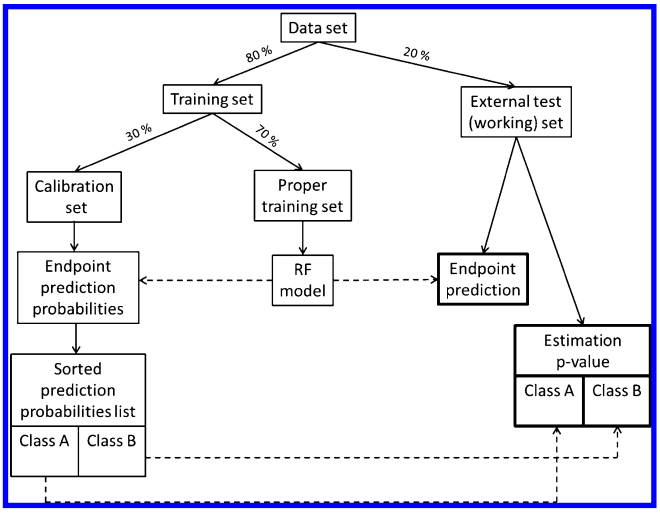
Table 1. Data Set Characteristics

data set	end point	no. of compds	type[a]	ref
Bzr	benzodiazepine receptor	163	binary class	24
cox2	cyclooxygenase-2	322	binary class	24
Dhfr	dihydrofolate reductase	397	binary class	24
Hptp	human protein tyrosine phosphatase 1B	132	binary class	25
f7	factor-7	365	regression	26
il4	interleukin-4	665	regression	26
jak1	Janus kinase-1	921	regression	26
Mgll	monoacylglycerol lipase	1230	regression	26
mmp2	matrix metallopeptidase-2	549	regression	26
prss2	protease, serine, 2 (trypsin-2)	339	regression	26

[a]For classification data sets: balanced classes.

In the classification setting we used a label-wise Mondrian ICP (MICP) to ensure validity for each class label. We defined the nonconformity score to be the probability for the prediction from the decision trees in the forest, i.e. the score of a new compound (example) is equal to the percentage of correct predictions given by the individual decision trees in the forest (see Scheme 1).[22]

Scheme 1. Conformal Prediction in Classification Setting



For regression, we defined a nonconformity measure following eq 16 in ref 16

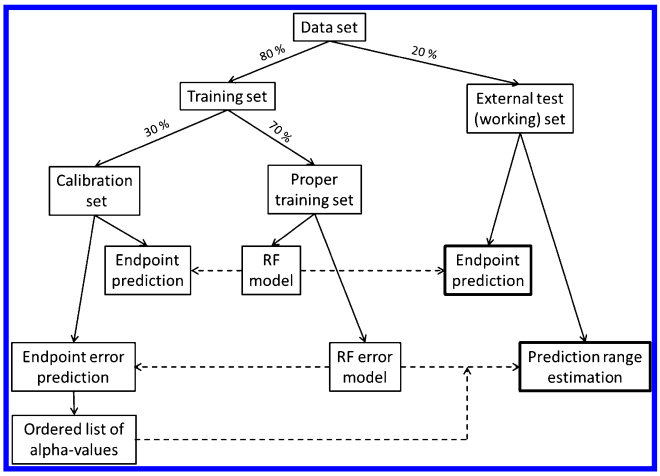
$$\alpha_i = \frac{|y_i - \hat{y}_i|}{\hat{\mu}_i} \quad (5)$$

where $\hat{\mu}_i$ is the error prediction from a derived *error model*. We first trained a *target model* (where the target may be e.g. biological activity) on the proper training set. We then calculated the residuals $|y_i - \hat{y}_i|$ of the target model for all proper training examples and trained the error model using $|y_i - \hat{y}_i|$ as the response variable and the same (signature fingerprints) as descriptors, i.e. the error model is based on self-fit predictions. [Preliminary investigations using cross-validation as well as out-of-bag errors do not result in significant overall differences with respect to prediction ranges for regression as compared to the current scheme.] Equation 5 normalizes the absolute prediction error with the predicted

dx.doi.org/10.1021/ci5001168 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

error on a given example (see Scheme 2). This leads to prediction intervals that are larger for the "difficult" examples

Scheme 2. Conformal Prediction in Regression Setting



and smaller for the "easy" ones. As a result, the ICP can satisfy the required confidence level with intervals that may be tighter on average.

Pseudocode for Using the Inductive Conformal Predictions. 1. Derive RF models: The target model and, for regression, the corresponding error model using the proper training set (Schemes 1 and 2).

2. Create lists of nonconformity scores (α-values) using predictions on the calibration set (two lists for classification (Figure 1) and one list for regression (Figure 2), one for each class, since we are using a label wise Mondrian ICP for classification).

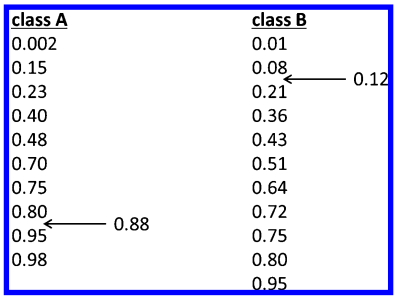


Figure 1. Mondrian class lists for classification.

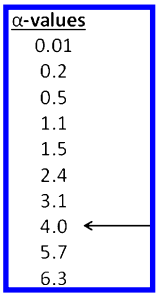


Figure 2. Sorted list of α-values for regression.

3. Sort the list(s) by increasing score values.

4. Set the significance level, e.g. at 0.2 (thus assuming 0.8 confidence level).

5. For each of the compounds in the external test set to be predicted:

1. **Classification case** (Figure 1): obtain the probability of the prediction from the RF target model for each class (in our example the percentage of the trees that give the correct class). For each class assess where this probability is located in the list of predicted probabilities for the corresponding class, e.g. if the probability list for class A has the following sorted list of predicted probabilities from the calibration set: 0.002, 0.15, 0.23, 0.4, 0.48, 0.7, 0.75, 0.8, 0.95, 0.98 the probability for the new compound is predicted to be 0.88, then it will be between 0.8 and 0.95 in the list thus giving a p-value of 0.82 (9/11). A similar lookup is then performed in the sorted probability list of class B where, in this binary classification case, the probability is 0.12 to determine the p-value for that class. If class B has a sorted list of predicted probabilities from the calibration set: 0.01, 0.08, 0.21, 0.36, 0.43, 0.51, 0.64, 0.72, 0.75, 0.8, and 0.95, then the probability for the new compound will be between 0.08 and 0.21 in the list thus giving a p-value of 0.25 (3/12). In this case the exemplified compound with p-values for class A and B of 0.91 and 0.25, respectively, which both are above the significance level set at 0.2, is predicted to belong to both class A and B (both). A compound with a predicted probability from the RF model of 0.10 and 0.90 for class A and B, respectively, with corresponding p-values of 0.18 (2/11) and 0.92 (11/12) would be predicted to only belong to class B.

2. **Regression case** (Figure 2): If the list of sorted α-values is 0.01, 0.2, 0.5, 1.1, 1.5, 2.4, 3.1, 4.0, 5.7, 6.3 with a confidence level set at 0.8 (step 3), we will, after traversing 80% of the list, use the value at that location, e.g. the value 4.0. From the error model the predicted error value of the new compound is computed, and we use eq 5 to calculate the predicted error range for the compound.

Software. Scikit-learn was used for fitting the random forest models (modules ensemble.RandomForestClassifier and ensemble.RandomForestRegressor). The default settings were used, except for the number of trees in the forests (100 trees was used for the target models and 50 trees for the error models).[23]

RESULTS AND DISCUSSION

From Table 2 it appears that conformal predictions using RFs and signature descriptors give conformal predictors that are valid for all cases (the negligible discrepancies observed in

Table 2. Results from Conformal Predictions on Randomly Selected External Test Sets

	CP accuracy (%)				
data set	0.80[a]	0.85[a]	0.90[a]	r^2	rmse
bzr	81.3	85.9	89.5		
cox2	79.5	85.0	90.2		
dhfr	80.3	85.0	89.7		
hptp	79.9	85.4	90.4		
f7	80.0	84.4	90.1	0.655	0.565
il4	80.6	85.3	91.0	0.533	0.389
jak1	80.3	85.0	89.7	0.521	0.497
mgll	80.6	85.1	89.8	0.474	0.712
mmp2	80.4	85.8	89.9	0.519	0.664
prss2	80.8	85.2	89.6	0.544	0.450

[a]Percent accuracy of the conformal predictor at the 0.8, 0.85, and 0.9 confidence levels, respectively.

dx.doi.org/10.1021/ci5001168 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2 from the confidence levels 0.8, 0.85, and 0.9, respectively, result from statistical fluctuations (see ref 14 for a more extensive discussion on this topic)). For regression this means that the number of errors where differences between the predicted and experimental values are outside the prediction region at the set confidence level $(1-\varepsilon)$ in question is less than ε (e.g., a confidence level of 0.8 may not have more that 20% errors for the conformal predictor to be valid). For classification it means that the predicted class is the correct one (including the compounds for which the model predicts both classes). The validity in these experiments is a consequence of that the exchangeability assumption is fulfilled, which can be accomplished by randomly permuting the row order in the data sets.

The consequence of increasing the confidence level, e.g. from 0.8 to 0.9, is that the prediction regions increase. For classification this may mean that the prediction of a compound is suddenly changed from a correct single class prediction, i.e. class low or high, to that the model predicts the compound to belong to both classes, i.e. when the p-values for classes low and high are both above the significance level set for the model, or from being an outlier (null prediction), i.e. none of the p-values for classes low and high are above the significance level set for the model, to the model making either a correct single class prediction, an incorrect single class prediction, or predicting the compound to belong to both classes. Thus, being more confident has its price. This is exemplified in Figures 3a-c for the HTPT data set where the confidence level is increased from 0.80 via 0.85 to 0.90. All four phenomena discussed above, when raising the confidence level, can be observed:

1. The number of "outlier" (null)-class predictions decrease from 245 via 96 to 17 compounds at the 0.85 and 0.9 levels, respectively.
2. The number of both-class predictions is increased from 78 to 248 compounds at the 0.85 level and finally to 578 compounds at the 0.9 level.
3. The percentage of correct predictions increases for the low class.
4. The percentage of correct predictions varies from 87 to 86 back to 87 for the high class.

Conformal prediction permits the user of the model to set the confidence level appropriately depending on the situation at hand. For instance, if it is important to identify compounds above a certain activity threshold from a large pool of possible candidates (thus primarily desiring to keep the prediction regions small by using a low confidence level, e.g. 0.75), then the possible trade-off may be that a fairly large number of compounds will fall outside the predictions range. This would be particularly unfortunate for compounds being overpredicted, i.e. the predicted value is overestimated compared with the true experimental value, but then this should only be the case for maximally 25% of the predicted compounds if the conformal predictor is valid. On the other hand, if it is of crucial importance that the conformal predictor is confident to a high degree (e.g., 0.9), then the prediction ranges will be much wider in order to ensure this. In this case one would perhaps select the 30% compounds with the highest minimum predicted activities considering the prediction range (i.e., predicted activity plus or minus the length of the prediction region) and be aware of the fact that in the worst case one-third of these compounds may possibly have worse activities than minimally estimated given that the conformal predictor is valid at the confidence level 0.9. To clarify this reasoning, we give the following example: Let us assume that the data set in question

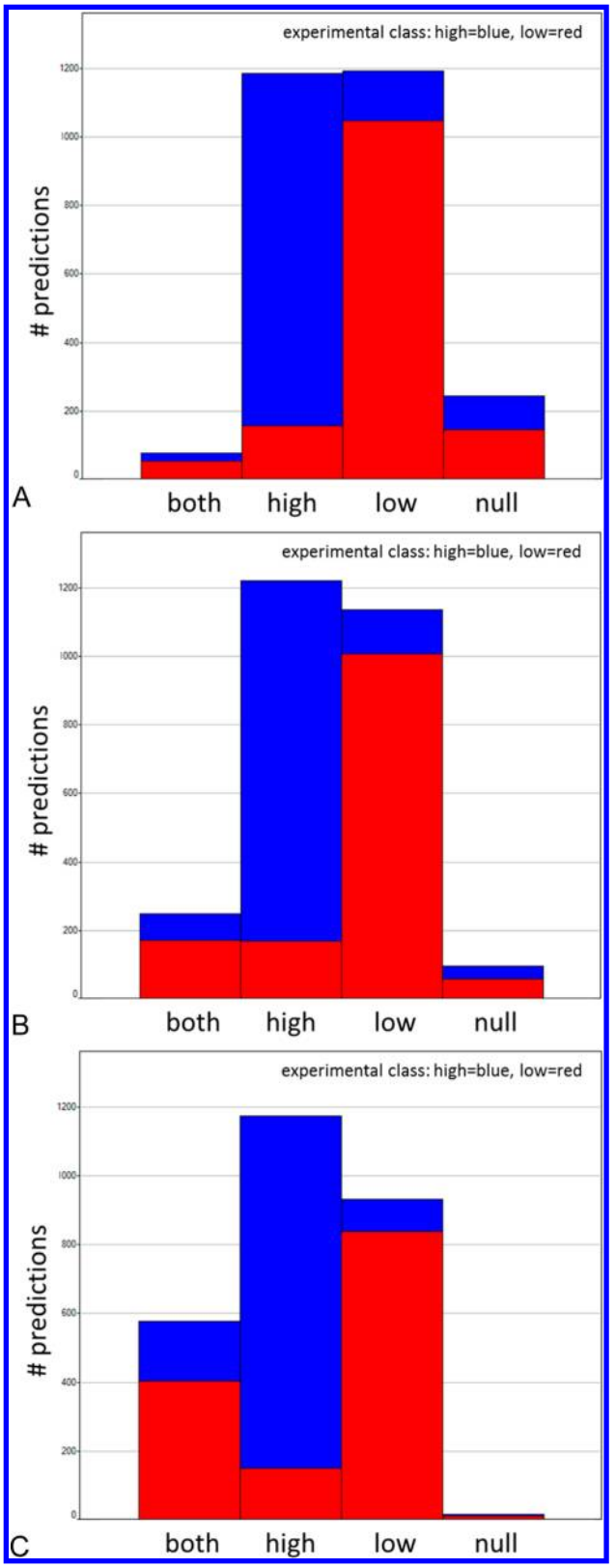


Figure 3. Changes in class memberships (experimental class: high = blue, low = red) for the HTPT data set when the confidence level is increased for 0.80 (a) via 0.85 (b) to 0.90 (c).

contains 100 compounds. If the conformal predictor is valid at the 0.9 confidence level, then a maximum of 10 compounds are allowed to be mispredicted. If we then select the 30% compounds with the highest minimum predicted activities

dx.doi.org/10.1021/ci5001168 | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

considering the prediction range, i.e. 30 compounds, the worst case scenario is where all of these 10 mispredicted compounds are among the 30% selected, i.e. one-third of the selected compounds may, at worst, have activities lower than minimally estimated.

This can nicely be illustrated by the PRSS2 data set. Figure 4a shows the predicted test set with prediction ranges at the 0.8 confidence level. If the desired minimum level of activity should be above 6, i.e. the predicted activity minus the prediction range has a value greater than 6, then Figure 4b shows the compounds remaining. In this case 7 out of 146 compounds are erroneous predicted which is well below what in the worst possible case potentially could be observed (if all erroneous compounds would fall within the activity region of interest). Raising the confidence level to 0.9 results in 108 compounds chosen with minimum predicted activities above 6, i.e. the predicted activity minus the prediction range has a value greater than 6 (Figure 4c). Again only 4 compounds are mispredicted which makes the predictor valid at the 0.9 level. The improvement in precision for this data set is illustrated in Figure 5 where the percentage of erroneous predicted compounds with an activity greater than 6 (false positives) decrease upon introduction of CP prediction ranges from 16% (no intervals) to 5% and 4% for confidence levels 0.8 and 0.9, respectively.

If, on the other hand, screening in a project is conducted with a relatively fast and/or inexpensive assay and the objective is to identify as many high-active compounds as possible without screening the entire available compound collection (library), then the prediction ranges would be used in the following manner:

Select the desired activity for which no compounds should be missed, i.e. predicted activity + prediction range for CP. In this example the desired activity is set to 7.

The result of identifying interesting compounds for screening can be viewed from three different perspectives: The percentage of the compound collection identified for screening, the percentage of erroneous compounds identified with lower activity (false positives) than 7, and the percentage of compounds missed with activities greater than 7 (false negatives). From Figure 6 it can be noted that although there is substantial reduction in identified compounds (only ~10% of the collection is identified) the result also contains a significant proportion of false negative (~59%) when the predicted activities were used without taking prediction ranges into consideration. On the other hand, if the CP confidence level was set to 0.9, then two-thirds (~67%) of the collection is identified and the false negative rate is only 3%. For this example, a useful compromise between screening size and false negative rate may exist at the confidence level of 0.8 where about 58% of the collection is identified and where, at the same time, the false negative rate is only 14%.

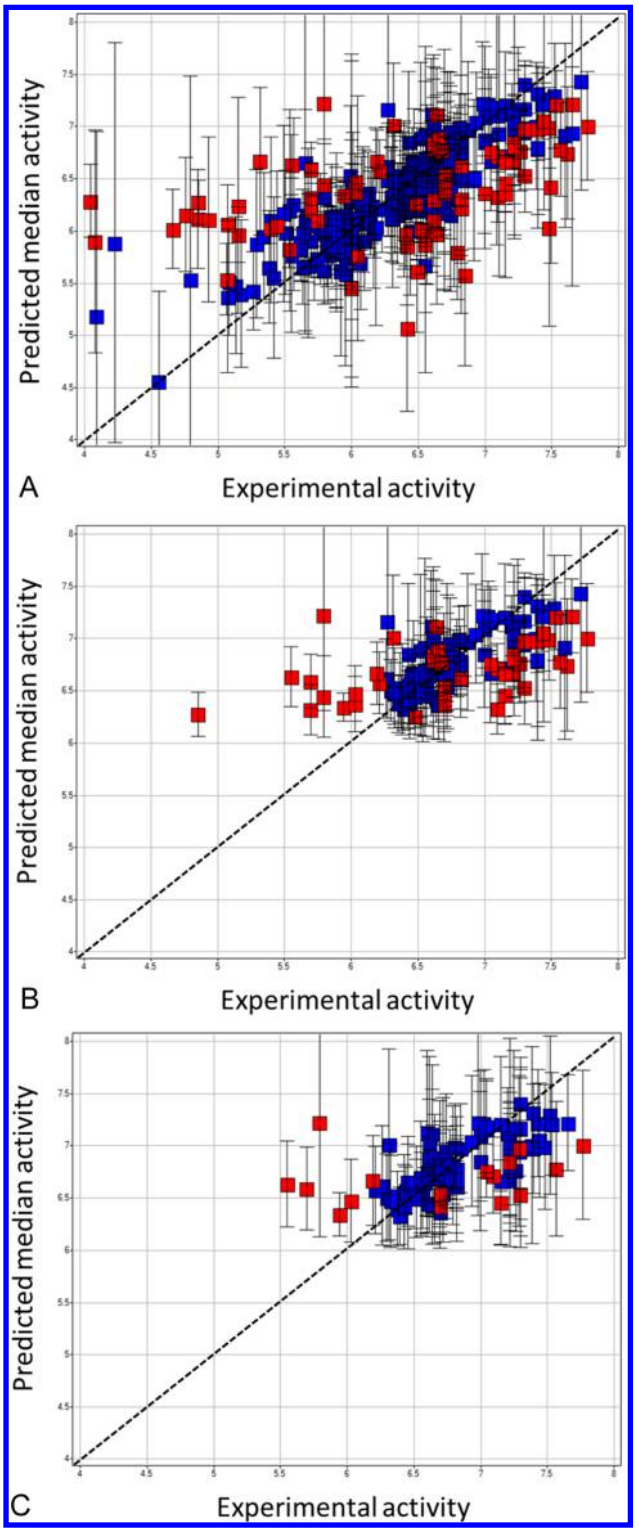


Figure 4. Experimental vs median predicted activities for the PRSS2 data set (a) and changes in median prediction range intervals for compounds with minimum predicted activities, i.e. compound where the predicted activity minus the prediction range, has a value greater than 6 when the confidence level is increased for 0.80 (b) to 0.90 (c). Red squares indicate erroneous predicted compounds, i.e. where the difference between experimental and predicted activity is larger than the prediction range. The prediction ranges are computed from the error model and median activities are the median from all the trees where the compound was assigned to the test set.

CONCLUSION

The use of conformal predictions as part of deriving quantitative structure−activity or structure−property relationships represents several advantages in comparison to the more traditional use of an applicability domain. First, there is a firm definition of a conformal predictor and a well-developed mathematical framework for how to construct them. No ambiguity exists as with the more loosely defined applicability domain concept. Second, the understanding of the confidence level concept in conformal predictions is straightforward, e.g. a

dx.doi.org/10.1021/ci5001168 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

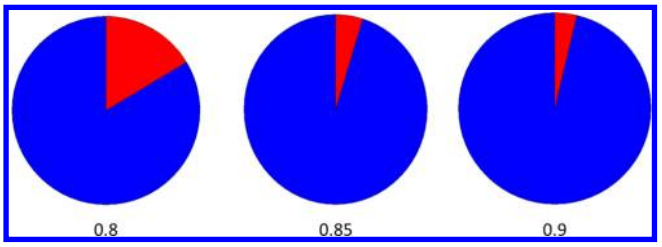


Figure 5. Errors (red) for the retrieved PRSS2 data set compounds with minimum predicted median activities greater than 6 (16.4% error) when the confidence level is introduced and increased from 0.80 (4.8) to 0.90 (3.8).

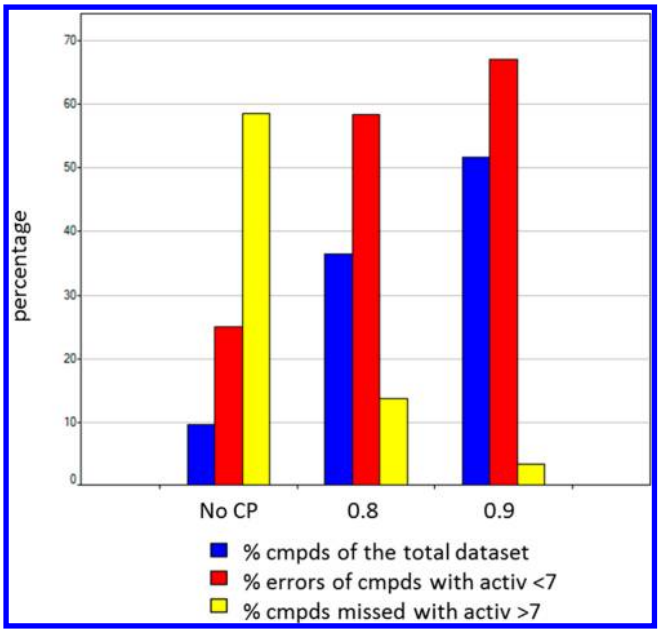


Figure 6. Percentage of retrieved compounds (blue), false positive rate (red), and false negative rate (yellow) for the PRSS2 data set when using only the predicted activities and with inclusion of prediction ranges at the CP confidence levels of 0.8 and 0.9, respectively, when identifying compounds with activities greater than 7.

confidence level of 0.8 means that, at most, there should be 20% errors given the prediction ranges if the predictor is valid. Third, the confidence level can be varied depending on the situation where the model is to be applied and the consequences of such changes are readily understandable, i.e. prediction ranges are increased or decreased, and the changes can immediately be inspected.

■ ASSOCIATED CONTENT

Supporting Information

SMILES and end point activities for the 10 data sets presented in this work. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ulfn@lundbeck.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was supported by the Swedish Council for Working Life and Social Research (FAS), grant number 2012-0073.

■ REFERENCES

(1) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* **2005**, *45*, 839−849.

(2) Schroeter, T. B.; Schwaighofer, A.; Mika, S.; Laak, A. T.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Muller, K.-R. Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 651−664.

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

(4) Weaver, S.; Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graphics Modell.* **2008**, *26*, 1315−1326.

(5) Dragos, H.; Gilles, M.; Varnek, A. Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. *J. Chem. Inf. Model.* **2009**, *49*, 1762−1776.

(6) 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 domain for classification problems: benchmarking of distance to models for Ames mutagenicity set. *J. Chem. Inf. Model.* **2010**, *50*, 2094−2111.

(7) Sheridan, R. P. Three useful dimensions for domain applicability in QSAR models using random forest. *J. Chem. Inf. Model.* **2012**, *52*, 814−823.

(8) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791−4810.

(9) Keefer, C. E.; Kauffman, G. W.; Gupta, R. R. An interpretable, probability-based confidence metric for continuous QSAR models. *J. Chem. Inf. Model.* **2013**, *53*, 368−383.

(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) Sheridan, R. P. Using Random Forest To Model the Domain Applicability of Another Random Forest Model. *J. Chem. Inf. Model.* **2013**, *53*, 2837−2850.

(12) Bosnić, Z.; Kononenko, I. Comparison of approaches for estimating reliability of individual regression predictions. *Data Knowl. Eng.* **2008**, *67*, 504−516.

(13) Clark, R. DPRESS: Localizing estimates of predictive uncertainty. *J. Cheminf.* **2009**, *1*, 11.

(14) Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic learning in a random world*; Springer: New York, 2005.

(15) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. Conformal prediction in QSAR. In *Artificial Intelligence Applications and Innovations: Proceedings, Part II of AIAI 2012 International Workshops: AIAB, AIeIA, CISE, COPA, IIVC, ISQL, MHDW, and WADTMB*; Iliadis, L., Maglogiannis, I., Papadopoulos, H., Karatzas, K., Sioutas, S., Eds.; Springer: Berlin, Heidelberg, 2012; pp 166−175.

(16) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. The application of conformal prediction to the drug discovery process. *Ann. Math. Artif. Intell.* **2013**, [Epub ahead of print], DOI: 10.1007/s10472-013-9378-2.

(17) Papadopoulos, H.; Haralambous, H. Reliable prediction intervals with regression neural networks. *Neural Networks* **2011**, *24*, 842−851.
(18) Papadopoulos, H.; Proedrou, K.; Vovk, V.; Gammerman, A. Inductive confidence machines for regression. *Proceedings of the 13th European Conference on Machine Learning (ECML'02)*; Vol. *2430* of Lecture Notes in Computer Science, 2002; pp 345−356.
(19) Papadopoulos, H.; Vovk, V.; Gammerman, A. Qualified predictions for large data sets in the case of pattern recognition. *Proceedings of the 2002 International Conference on Machine Learning and Applications (ICMLA'02)*; 2002; pp 159−163.
(20) Faulon, J. L.; Collins, M.; Carr, R. D. The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequence. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 427−436.
(21) Brieman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.
(22) Devetyarov, D.; Nouretdinov, I. Prediction with Confidence Based on a Random Forest Classifier. In *Artificial Intelligence and Innovations: Proceedings of 6th IFIP WG 12.5 International Conference*; Papadopoulos, H., Andreou, A. S., Bramer, M., Eds.; Springer: Berlin, Heidelberg, 2010; pp 37−44.
(23) Scikit learn version 0.11. http://scikit-learn.org/0.11 (accessed Dec 14th, 2013).
(24) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR classifiers compared. *J. Chem. Inf. Model.* **2007**, *47*, 219−227.
(25) Mittal, R. R.; McKinnon, R. A.; Sorich, M. J. Comparison data sets for benchmarking QSAR methodologies in lead optimization. *J. Chem. Inf. Model.* **2009**, *49*, 1810−1820.
(26) 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.