Journal of Pharmaceutical Sciences 110 (2021) 42-49

Contents lists available at ScienceDirect

Journal of Pharmaceutical Sciences

journal homepage: www.jpharmsci.org

Minireview

Predicting With Confidence: Using Conformal Prediction in Drug Discovery

Jonathan Alvarsson [a], Staffan Arvidsson McShane [a], Ulf Norinder [a, b, c], Ola Spjuth [a, *]

[a] Department of Pharmaceutical Biosciences and Science for Life Laboratory, Uppsala University, Box 591, SE-75124, Uppsala, Sweden
[b] Department of Computer and Systems Sciences, Stockholm University, Box 7003, SE-16407, Kista, Sweden
[c] MTM Research Centre, School of Science and Technology, Örebro University, SE-70182 Örebro, Sweden

ARTICLE INFO

Article history:
Received 5 August 2020
Revised 28 September 2020
Accepted 29 September 2020
Available online 17 October 2020

Keywords:
QSAR
Conformal prediction
Predictive modeling
Confidence
Applicability domain

ABSTRACT

One of the challenges with predictive modeling is how to quantify the reliability of the models' predictions on new objects. In this work we give an introduction to conformal prediction, a framework that sits on top of traditional machine learning algorithms and which outputs valid confidence estimates to predictions from QSAR models in the form of prediction intervals that are specific to each predicted object. For regression, a prediction interval consists of an upper and a lower bound. For classification, a prediction interval is a set that contains none, one, or many of the potential classes. The size of the prediction interval is affected by a user-specified confidence/significance level, and by the nonconformity of the predicted object; *i.e.*, the strangeness as defined by a nonconformity function. Conformal prediction provides a rigorous and mathematically proven framework for *in silico* modeling with guarantees on error rates as well as a consistent handling of the models' applicability domain intrinsically linked to the underlying machine learning model. Apart from introducing the concepts and types of conformal prediction, we also provide an example application for modeling ABC transporters using conformal prediction, as well as a discussion on general implications for drug discovery.

Introduction

Prediction of different endpoints based on chemical structure constitutes an important problem in drug discovery projects. A common approach is to use predictive modeling with QSAR (Quantitative Structure-Activity Relationships) with the aim to correlate chemical structures to a response (target) value, such as the biological activity towards a certain target, physiological properties such as solubility, or other measurable effects such as cytotoxicity. QSAR predictions can then support decision making in drug discovery, such as prioritizing between compounds and experiments.[1] QSAR is a ligand-based method which often relies on machine learning algorithms for making predictions, and a key challenge when constructing and using these types of models is the concept of confidence in predictions; *i.e.*, how much can you trust the predictions made by this approach on a novel compound that has never been tested or sometimes not even synthesized?

In this article we give an introduction to conformal prediction, a framework operating on top of a regression or classification algorithm in predictive modeling. Conformal prediction adds several benefits to predictive modeling, mainly by assigning a valid measure of the confidence in predictions that is specific to the predicted object. In QSAR, the predictions made by the conformal predictor thus already take into account the *strangeness* of a new compound compared to training data, delivering an alternative to the concept of applicability domain that is commonly used within this field.[2] The rest of this article is organised as follows: first in (2) we give some background on conformal prediction in QSAR and introduce the concepts of validity and efficiency, then in (3) we describe general applications of conformal prediction in drug discovery, in (4) we present a case study on ABC transporters, in (5) we describe different approaches to conformal prediction and conclude in (6) with discussing the implications of using conformal prediction in drug discovery.

Methods

QSAR modeling constructs *in silico* prediction models from a set of collected training compounds described by a set of chemical

Conflicts of interest: OS declare shares in Aros Bio AB, a company developing software that utilizes conformal prediction methodology.

* Corresponding author.
E-mail address: ola.spjuth@farmbio.uu.se (O. Spjuth).

https://doi.org/10.1016/j.xphs.2020.09.055

structural descriptors,[3] *e.g.*, molecular fingerprints or physiochemical parameters, using an appropriate algorithm of some sort, *e.g.*, Random Forest, Support Vector Machines or Neural Networks. The derived model is then used to predict new test compounds for the model property in question such as biological activity or toxicity. An important aspect during model building is the applicability domain (AD) of the derived model, *i.e.*, for which compounds the model can make reliable predictions, that needs to be determined by some metric.[4] There exist a plethora of suggested methods and metrics for creating the AD that are more or less closely linked to the underlying model,[5] and with varying degrees of mathematical rigor.

Conformal Prediction

Predictions made by standard learning algorithms are typically point predictions, *e.g.*, a real value such as 2.4 for a regression problem or "active" (predicting active vs inactive) for a classification problem. These point predictions have no measure of uncertainty assigned to them and either an external validation set or cross validation is needed in order to estimate the average performance of the model. However, using the average performance of a model as a level of confidence for predictions from a model gives the same confidence region to all predictions made by the model. With conformal prediction we can improve the handling of prediction confidence. If the new compound being predicted is similar to the compounds used in the training data we would typically trust the prediction to a higher degree than if it was chemically dissimilar compared to training compounds of a small chemical space (the background of applicability domain). Conformal prediction addresses this by returning *prediction regions*, *i.e.*, intervals for regression problems and sets of labels for classification problems. For a given compound and a given confidence level the conformal predictor provides a prediction interval that the true value should lay within with a probability of the given confidence, *e.g.*, in the interval (2.29, 2.53), or that the true class is in the prediction set {active}. These prediction regions are very similar to confidence intervals used in statistics but they are not based only on overall statistics but on the individual predictions. The region size is dependent on multiple factors, such as strangeness of the test compound compared to training compounds, the desired confidence of the prediction and the overall *efficiency* of the predictor, further discussed in Section Efficiency.

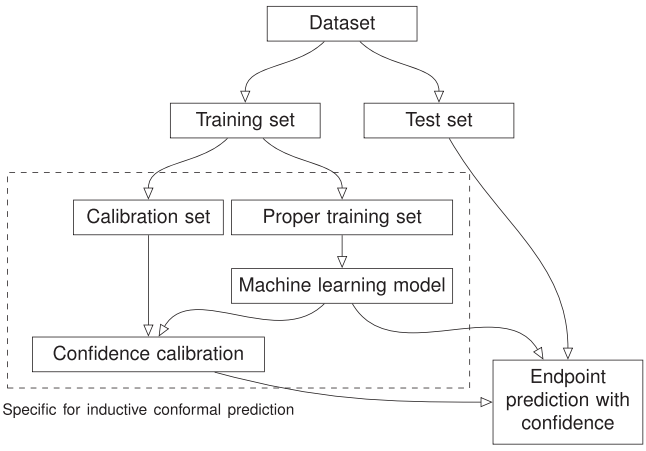

Fig. 1. In the typical workflow of machine learning, all available data is initially split up into two disjoint sets; the training set used for generating the model and an external test used for evaluating final model performance. In inductive conformal prediction, the section surrounded by the dashed line, the training set is further split into a calibration set and a proper training set. The proper training set is used for training a machine learning model and the calibration set is used for calibrating the predictions made by the model to yield the conformal prediction. How the calibration is performed differs between, *e.g.*, regression and classification.

Inductive conformal prediction (ICP) is the most widely used approach to conformal prediction. Just as any other conformal prediction method it acts as a layer on top of an underlying machine learning algorithm, it then adds calibration by splitting the training set into a calibration set and a proper training set (Fig. 1). A model is then constructed based on the proper training set, and the prediction region (confidence estimate of confidence) is obtained from the calibration set.

The most commonly used ICPs for classification are Mondrian ICPs, addressing a common problem in machine learning which is in dealing with classification problems having imbalanced data (many more examples in the training data having one of the class labels). There are many techniques used for balancing such data sets, *e.g.*, under/oversampling[6–8] and boosting.[9] In Mondrian conformal prediction, the prediction for each class is estimated separately using an individual calibration set per class. This has been shown to work well even for severely imbalanced datasets,[10,11] without the need for additional balancing techniques.

In binary classification (active/inactive classes) there exists four outcomes for conformal prediction:

1. active
2. inactive
3. both classes (active and inactive)
4. no class assignment (empty class)

A conformal prediction is deemed correct if it includes the correct class which means that "both" predictions are always correct and "empty" predictions are always erroneous.

What are the implications of "both" and "empty" predictions? For "both" predictions it means that the predicted compound is similar to both sets of known compounds at the set significance level (significance level = error rate or 1 − confidence level) and *vice versa* for "empty" predictions, *i.e.*, the predicted compound is too dissimilar to both classes of known compounds for the model to give a reliable prediction. The implication for the first case ("both") is lack of information to distinguish between the 2 classes and that new information through additional descriptors must be provided in order to result in a single class prediction. The implication for the second case ("empty") is that the predicted compound is out-of the applicability domain and that the class for the compound should be determined (*e.g.*, experimentally) and the compound later incorporated into the updated model in order to expand the model's applicability domain.

The Mondrian conformal classification predictor outputs p-values for each class, which are used slightly differently than in standard hypothesis testing in statistics. In essence, these p-values are the ranking of a test object compared to known instances of each class. The prediction sets are calculated from these p-values together with the desired confidence by finding the p-values that are equal to or larger than the *significance* threshold $\epsilon = 1-$ confidence level (*i.e.*, the percentage of accepted errors). By only including p-values over the desired significance level, the true label is excluded with a probability of ϵ. So for a class with p-value of, *e.g*, 0.85 that class will be part of the prediction set at a confidence of 0.15 or higher (or, equivalently, at an ϵ of 0.85 or lower).

Conformity and Nonconformity

Central to conformal prediction is the use of a *nonconformity measure* to assess how dissimilar or 'strange' a new object is

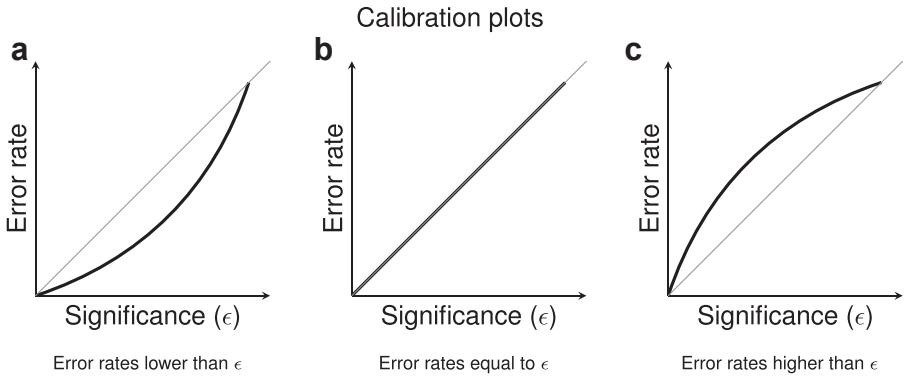

Fig. 2. Calibration plots plot the error rate as a function of the significance ϵ. In the best of worlds the error rate is equal to the significance (b) and we obtain the error rate that we ask for. However, something to look out for is when the error rate is higher than the significance as in sub figure (c). Sometimes the error rates are lower than the significance (a), although this is in some sense wrong it is not a problem in practice since making fewer errors than anticipated is normally perceived as an advantage.

compared to the data that the model has been built upon. An alternative is to present this as a *conformity measure* which measures similarity instead of dissimilarity (going from one to the other is trivial, *e.g.*, by defining conformity: $=-$nonconformity or conformity: $=1-$ nonconformity). Although it is possible to use either a conformity or nonconformity measure, nonconformity measures are used more frequently as it follows the original work by Vovk et al. were they prefer nonconformity due to it typically being easier to construct a measure of an object's distance to another set of objects in some space than its corresponding closeness and to follow conventions in mathematical statistics.[12] However, some people find it more natural to think in terms of similarity and prefer conformity measures. The nonconformity measure is calculated via a nonconformity function, which in most cases comes from the prediction of a machine learning algorithm. Examples of nonconformity measures include distance to the decision hyperplane when using a Support Vector Machine, the out-of-bag measure when using a Random Forest, or the SoftMax value from a neural network. It is important to note that it is proven that any function can be used as a nonconformity function,[12] but that the choice of this function is the key to obtaining a predictor with high efficiency (see Section Efficiency and[13]).

Validity

Conformal prediction has been mathematically proven to produce valid predictions, given that data is exchangeable, meaning that examples are drawn from the same distribution of data and that there is no particular ordering of the examples.[12] Validity implies that for a given confidence level of, *e.g.*, 0.9 the predictor will include the correct value in its prediction intervals or prediction sets in at least 90% of all predictions (*i.e.*, the predictor is correct at least 90% of the time).

It is important to note that all machine learning methods in general assume exchangability, or, in fact, the slightly stricter assumption of the data being *Independent and Identically Distributed* (IID) which means that there are no new requirements introduced by conformal prediction since the same assumptions on the underlying data are made when using any machine learning methodology.

When the exchangability assumption does not hold this might be discovered in a calibration plot (see Fig. 2), the error rate is plotted versus the significance and the model is considered valid (well-calibrated) if the result is a straight diagonal line, *i.e.*, we obtain the error-rate we ask for when making predictions. Deviations from the expected error rates can mainly be attributed to either: lack of exchangeability for the set of predicted compounds, or statistical fluctuations due to small sets of predicted compounds[14,15] since conformal prediction will provide valid predictions "over time", *i.e.*, given enough predictions (law-of-large-numbers).[12]

Efficiency

Efficiency is another important concept in conformal prediction, especially given the guaranteed validity for the derived model. There are many different definitions of efficiency in conformal prediction,[12] where some are dependent on the confidence level and others are not. A predictor can always be correct in 100% of its predictions by always predicting all possible values, but such predictions are not informative, *i.e.*, a predictor that when asked if a compound is toxic always answers with both yes and no is always correct, but not particularly useful. The efficiency of a predictor is a measure of how specific the predictions can be, while still remaining valid. For regression, efficiency is typically defined as the mean or median prediction interval width and for classification as the ratio of single label predictions (a higher ratio is preferable) or ratio of prediction sets with two or more labels (a smaller value is preferable).

In order to construct a predictor that is as efficient as possible, several aspects need to be taken into consideration and investigated such as descriptors, learning algorithms, validation techniques as well as parameters to optimize, *e.g.*, various algorithmic settings and descriptors selection procedures. Additionally when using conformal predictions, a good *nonconformity* or *conformity measure* is needed (see Section Conformity and Nonconformity).

Conformal Prediction Applied to Drug Discovery

Regression

In this section we will investigate a few examples to illustrate how conformal predictors operate. We will start with a regression problem of predicting LogD for omeprazol using our online service found at https://cplogd.service.pharmb.io/. The service contains a QSAR model built on approximately 1.6 million calculated LogD values at pH 7.4 found in the ChEMBL database.[16] In Fig. 3 we show how the confidence affects the prediction region where a higher confidence forces the predictor to assign a larger prediction interval (*i.e.*, being less specific); requiring a high confidence in the prediction result in larger intervals, and the opposite, allowing for more errors lets the predictor be more specific. For the extreme value of confidence set to 1.0 (in Fig. 3), *i.e.*, we want to be

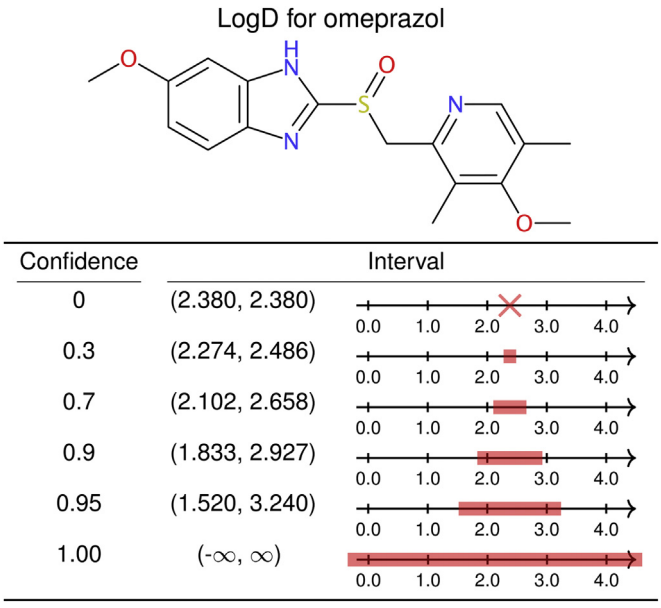

Fig. 3. Example of LogD prediction for omeprazol at different confidence levels. At confidence 0 the algorithm gives a single point and then as the confidence is increased, so is the interval. The only way to safely guarantee a correct prediction, i.e., confidence 1, is to provide an infinitely large interval.

absolutely sure that the predicted interval covers the true value, we see that in order to be that certain an infinitely wide interval is needed which is not useful at all. At the other extreme, if we set the confidence to 0, we obtain a point value without any useful measure of certainty.

Another of the advantages of conformal prediction is that predictions are object related, an easier to predict compound will produce smaller prediction intervals compared to a more difficult one. This is visualized in Fig. 4 where we predict LogD for the well known drug paracetamol and for 4-acetamidothiophenol. The prediction interval for paracetamol is roughly half the size of that of 4-acetamidothiophenol, indicating that paracetamol is more similar to examples used for training than 4-acetamidothiophenol and the predictor can thus provide a more specific prediction for the well known drug.

Classification

For a binary classification problem with the classes A and N (*e.g.*, active and non-active compounds) there are, as already mentioned, four possible prediction sets: {A}, {N}, {A, N} and {Ø}. The sets {A} and {N} are single label sets and are the best result for an end user; the predictor will provide a single label (class) for the test compound indicating one or the other of the two possible classes. For the {A, N} prediction the predictor cannot, for the given confidence, distinguish between the two classes. The empty prediction set, {Ø}, means that the test compound is difficult to predict and no reliable predictions can be made by the predictor.

We will now look at an example of how the classes A = active and N = nonactive are derived from the p-values by using the prediction of off-target binding from our online service http://ptp.service.pharmb.io/ [17] by use of gene LCK(Fig. 5). The service contains 31 QSAR models modeling known adverse targets. More specifically we will look at a model modeling binding to the target connected to the LCK gene. This model is based on 2662 known actives, 283 known non-actives and 4963 assumed non-actives. For more details about the model see Refs.[17] Predicting binding for omeprazol results in p-values 0.137 for the active class and 0.488 for non-active class, both relatively low p-values. In Fig. 5 we investigate how the desired confidence affects the prediction set; at a low confidence the prediction set is as small as possible (even empty when having a confidence of 0.512 or lower) and then the predictor set grows larger as confidence increases. At a confidence of 0.87 the predictor will start to include both classes.

Case Study: Prediction of ABC Transporters Using Conformal Prediction

Data Sets and Descriptors

The Bcrp, BSEP and Pgp inhibition data sets were obtained from Montanari et al.[18] The transporters and the number of inhibitors and non-inhibitors are listed in Table 1. The compounds were characterized using 97 different physiochemical RDKit[19] descriptors previously successfully used in *in silico* model building.[20]

ATP-Binding Cassette (ABC) transporters are membrane proteins that mediate translocation of substrates across cellular membranes.[21] P-glycoprotein (P-gp/ABCB1), Breast Cancer Resistance Protein (BCRP/ABCG2) and Bile salt export pump (BSEP/ABCB11) inhibitors exhibit a wide variation with respect to chemical structures.[22] P-gp is an efflux transporters found in tissues such as the intestine, brain and kidney.[23] Blockade or absence of intestinal P-gp results in decreased extrusion and increased availability of drugs that are P-gp substrates, which may lead to toxicity.[24] BCRP plays an important role in drug disposition[25] and BSEP is an

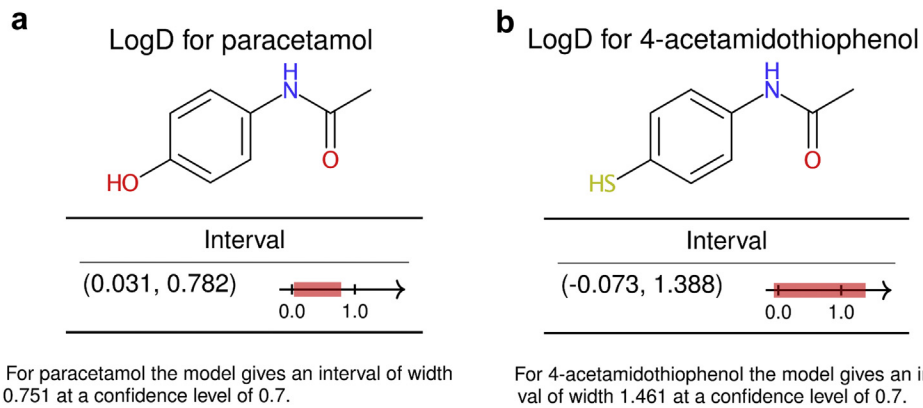

Fig. 4. The interval given by the model represents how certain the model is about the specific compound. For paracetamol (a) the interval is smaller than for 4-acetamidothiophenol (b) because the model is more certain about the prediction of LogD for that compound.

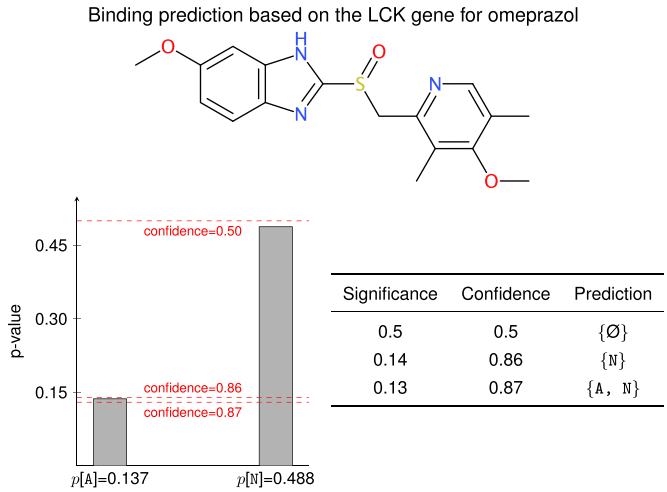

Fig. 5. Prediction of binding for omeprazol to the LCK target using a classification model. The prediction is given as a set of p-values, one for each class, forming the final prediction sets when deciding on a desired confidence. Formation of the prediction set can be visualized with the bar plot on the left, where a p-value above the significance level (defined as 1 − confidence) places it into the prediction set.

important transporter where bile secretory failure causes cholestasis and drug-induced liver injury (DILI).[26]

Study Design

The case study consists of classification modeling for the three different transporters. The training and test sets for each data set from the original publication were combined and duplicates removed. An external test set, 20%, was selected for each data set using stratified random sampling. The remaining part of each data set was used as training set for model building.

Model Building

Mondrian ICP models[11,27] were trained based on an underlying random forest (RF) model with 100 trees (Scikit-learn version 0.20.4[28]). The nonconformist package (version 1.2.5)[29] was used for building the Mondrian conformal predictors. Default values were used for all parameters unless explicitly stated and a calibration set of 30% was randomly selected from the training set and the remaining part (*proper* training set) used for model building. The conformity measure used was the probabilities from the RF ensemble.

Results

The results of the case study is presented in Table 2 and Figs. 6 and 7, where the table presents validity and efficiency for a set of pre-determined significance levels.

Table 1
The Targets in the Datasets, Number of Active Inhibitors (A) and Non-Inhibitors (N), and Percentage Inhibitors (A (%)) for the Calibration, Proper Train and Test Sets.

Transporter	Training						Test		
	Calibration			ProperTrain					
	A	N	A (%)	A	N	A (%)	A	N	A (%)
BcrpINH	130	150	≈46%	300	352	≈46%	108	125	≈46%
BSEPINH	38	124	≈23%	87	292	≈23%	31	105	≈23%
PgpINH	166	145	≈53%	391	333	≈54%	104	119	≈47%

Fig. 6a and b contains calibration plots both for each class separately and for both of them together ('Overall'). For BSEPINH we see that all curves are below the diagonal for small significance values. In fact the 'Overall' and 'non-inhibitor' class line stays below or close to the diagonal most of the time, meaning that they give an error rate lower (over-conservative) or well corresponding to the required significance level for all significance levels. However, at significance around 0.6 the error rates for the 'Inhibitor' class is close to 0.7. For the PgpINH dataset there is a clear trend that the 'Inhibitor' class is predicted with an error rate lower than the required significance and the 'noninhibitor' class with an error rate that is higher, but the 'Overall' line is below or near the diagonal meaning that when looking at both classes at the same time the error rate is the required one. However, the deviations observed in Fig. 6a and b are minor considering the small data set sizes and can be considered as statistical fluctuations.[12]

Fig. 6c and d shows the p-values for the two classes plotted against each other. This gives a visualisation of the data used for making predictions. We see that a few compounds are strongly indicated as being in the wrong class, *e.g.*, for BSEPINH there is a cyan dot almost all the way up in the left corner, however there are no inhibitors in the test set for BSEPINH with a p-value higher than 0.4 for non-inhibitor so the bottom right corner of that plot only contains the correct class.

Fig. 6e and f shows how the distribution of prediction labels varies as we vary the significance level. The significance level printed in the grey box (0.22 for BSEPINH and 0.12 for PgpINH) corresponds to the significance level giving rise to the highest number of single label predictions. As before we see that at significance level 0 we obtain only multiset predictions and at 1.0 we obtain only empty set predictions.

When using a conformal prediction classifier one has to set a significance ϵ. Once ϵ has been set it is possible to create a confusion matrix for the test set. A classic confusion matrix consists of 2 rows and 2 columns but in the case of conformal prediction there are two more rows, one for empty predictions and one for multiclass predictions, (in the case of binary classification 'Both' classes). Fig. 7 shows a visualisation of confusion matrixes for the BSEPINH dataset and the PgpINH dataset at two different significance levels. Once again we see that at lower ϵ we obtain more multiclass predictions and at higher ϵ we obtain more empty prediction sets.

Approaches to Conformal Prediction

The two most widely used approaches for conformal prediction are inductive conformal prediction (ICP) and transductive conformal prediction (TCP). In ICP, as previously described in Section Conformal Prediction, all training examples are split up into two sets and a single machine learning model is trained using the proper training set (see Fig. 1) and this model is used for all future predictions until enough new data has been collected for the model to be worth updating. ICP is possible to use in both classification and regression problems, TCP on the other hand is only available for classification. In TCP, the online approach which was the one first described, all training examples are used in both training the underlying learning model and calibrating the predictions. A new model is trained for each class and each predicted object (*e.g.*, for a binary classification two models are learnt for *every* prediction). This makes TCP much more computationally demanding, which can be practically infeasible for larger datasets, but in general it provides better results than the ICP approach since it does not require to set aside training examples for calibration.

The ICP approach has been further developed into aggregated conformal prediction (ACP)[30] in which many ICP's are produced by randomly sampling the same training dataset multiple times, and

Table 2
Results From the Classification.

Transporter	Significance Level	Validity Inhibitor Class	Validity Non-Inhibitor Class	Efficiency	Efficiency Inhibitors	Efficiency Non-Inhibitor
BcrpINH	0.1	0.954	0.896	0.632	0.759	0.520
	0.15	0.926	0.840	0.790	0.870	0.720
	0.2	0.917	0.800	0.893	0.954	0.840
	0.25	0.880	0.760	0.974	0.991	0.960
BSEPINH	0.1	0.968	0.971	0.765	0.613	0.810
	0.15	0.935	0.895	0.882	0.806	0.905
	0.2	0.935	0.838	0.978	0.968	0.981
	0.25	0.935	0.810	0.971	1.000	0.962
PgpINH	0.1	0.929	0.891	0.942	0.943	0.941
	0.15	0.893	0.840	0.961	0.964	0.958
	0.2	0.871	0.790	0.907	0.929	0.882
	0.25	0.793	0.748	0.834	0.843	0.824

Validity is defined as percentage of correct predictions where the label set includes the correct label. The efficiency measure used is fraction of single label predictions. For many practical applications an efficiency of around 0.8 or higher is desirable. See also plots in Fig. 6 for a graphical illustration.

cross conformal prediction (CCP)[31] in which the random sampling used in ACP is replaced by the same sampling approach as is used in cross validation. Fig. 1 shows an outline of the construction of an ICP, where the ICP specific section is performed repeatedly for ACP and CCP and the final prediction is an aggregation of the individual predictions using the mean or median prediction.

For further reading about conformal prediction, if you found the application in this work interesting, we recommend Norinder et al.[2]

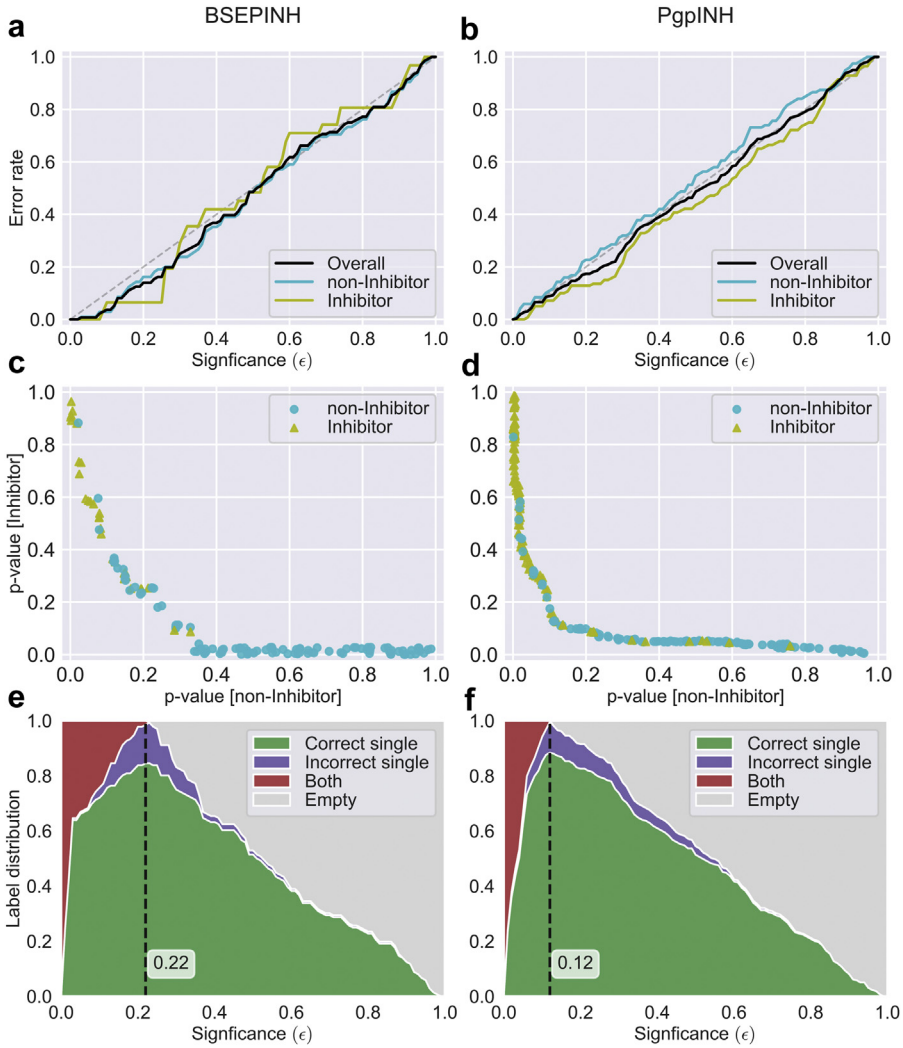

Fig. 6. Plots for two classification datasets from the case study. (a and b) Calibration curves for the two datasets show error rates plotted for different significance values. We want the curves to be near the diagonal. (c and d) P0/P1 plot, plotting the two p-values against each other while coloring the dots according to observed class. Preferably we want to see one class in the bottom right corner and the other class in the top left corner. (e and f) Label distribution plots for different significance values, for lower significance values the model incorporate more and more labels in the prediction and at significance 0 all predictions are predicted as both classes.

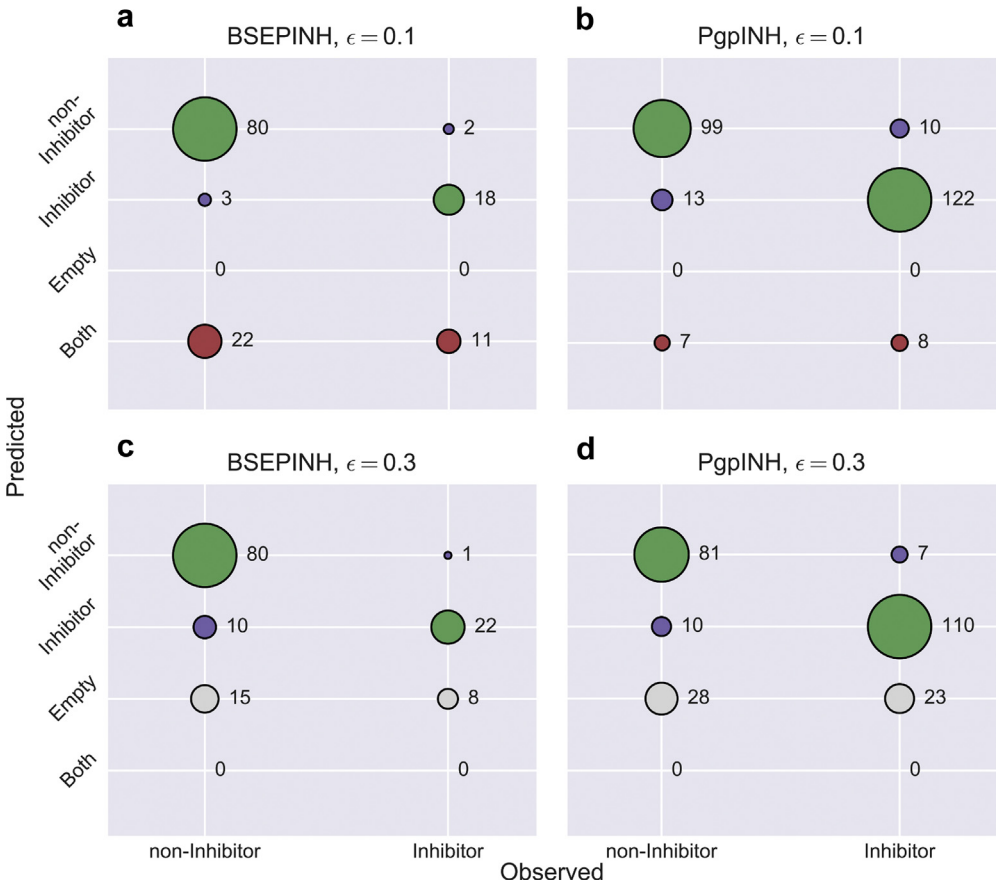

Fig. 7. Plot of confusion matrix for the BSEPINH dataset and the PgpINH dataset at significance 0.1 (a, b) and 0.3 (c, d). We see that for both targets going from $\epsilon = 0.3$ to 0.1 means a decrease of empty class predictions and an increase of both class predictions. i.e., at a lower significance level, the predictor will have to create larger prediction sets to not violate the required significance.

and Eklund et al.,[32] both describing conformal prediction in the drug discovery setting but from a more mathematical viewpoint. If you are specifically interested in papers dealing with imbalanced QSAR datasets in conformal prediction we recommend Norinder et al.[10] and Sun et al.[11] For someone interested in implementation details and mathematical theorems a good place to start is http://alrw.net/, or even the original book introducing conformal prediction by Vovk et al.[12]

Worth noting is that recently there has been a shift in naming conventions, where the inductive conformal predictor in some literature is now referred to as Split Conformal Predictor (SCP) as in Colombo and Vovk[33] written by one of the original authors of conformal prediction.

Discussion

Predictive modeling in drug discovery is still relatively new, and there has been (and still is) resistance among some scientists to use predictions to aid decision making. It is hard to know how much to trust a point estimate made by a computer software and in this situation it may be considered reasonable to chose not to trust them at all. In the authors experience it is a convincing argument in this discussion that the conformal prediction provides confidence measures and that they are mathematically proven to be valid, and also that the results depend on the rather natural question: How certain do you want the prediction to be?

As we have seen in both the regression and classification examples presented herein, the general outcome of conformal prediction is that the higher the desired level of confidence — the larger the prediction interval or prediction set (*i.e.*, less specific predictions). This is easy to comprehend — if you need to be very certain about an estimate you will then need to provide a large prediction interval. Prediction intervals also depend on the test objects themselves, where a larger prediction interval is assigned to more difficult objects and *vice versa.* One observation the authors have made in drug discovery projects is that when users are faced with prediction intervals, it is not uncommon that they at first sight are disappointed that the interval is larger than they expected. However when faced with the decision to simply remove the prediction interval and go back to point predictions as they have done before, they tend to prefer to keep the prediction interval and continue to use the prediction leading to, in our opinion, more informed and balanced decisions.

Conformal prediction is a relatively new framework, and unknown to many readers and practitioners. For conformal prediction classification an additional challenge arises in that it is difficult to compare the results with more familiar methods of statistical model evaluation. Mondrian conformal prediction, as used in this work, results in two p-values per predicted object (one for each class label) that do not have to sum to 1. An object-specific confidence interval is derived from these p-values, and while this has the benefits of being a valid measure of confidence for the prediction, it is nevertheless difficult to compare to more traditional

used measures like area under the receiver operator characteristics curve, balanced accuracy, Matthews correlation coefficient, or just specificity and sensitivity, due to the presence of multi-label and/or empty-label predictions.

In order to use conformal prediction, a significance level needs to be set. As earlier eluded to, for conformal prediction to be as effective as possible the efficiency (often defined as the percentage of single label predictions) must be high at significance levels that are supportive to the user of the models when attempting to make the decision at hand, *e.g.*, whether to send a set of compounds to synthesis and/or biological screening or determine whether the compounds are toxic or not. Therefore, different significance levels need to be investigated in order to understand the performance of the conformal prediction model. It may also be so that a significance level set at the beginning of a drug discovery project, where more errors are acceptable and less expensive tests are used, may not be acceptable later on when more expensive investigations have to be performed. This, in turn, may warrant a lower significance level and the consequences of such a change must be understood for future predictions.

Conformal prediction has many advantages over traditional machine learning models yielding point estimates, and a key objective of this manuscript is to introduce and explain these advantages. Despite not being able to directly evaluate conformal predictors using traditional accuracy metrics, and the educational aspect of having to decide on a significance level when making predictions, in our experience scientists who have learnt to interpret prediction intervals from conformal prediction are unwilling to go back to point estimates or use *ad-hoc* approaches to applicability domain estimation without mathematical rigor. Given the increase over the last years in the number of published studies presenting results from conformal prediction, it is clear that conformal prediction is a methodology worth grasping.

Acknowledgements

This project was financially supported by FORMAS (grant 2018-00924) and the Swedish Foundation for Strategic Research (grant BD15-0008SB16-0046).

References

1. Spjuth O, Eklund M, Ahlberg Helgee E, Boyer S, Carlsson L. Integrated decision support for assessing chemical liabilities. *J Chem Inf Model*. 2011;51(8):1840-1847.
2. 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(6):1596-1603.
3. Kubinyi H. *QSAR: Hansch Analysis and Related Approaches*. vol. 1. VcH; 1993.
4. O. OECD, Principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models. 2004. Available at: https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf.
5. Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of qsar models. *Chemometr Intell Lab Syst*. 2015;145:22-29.
6. Chawla NV. Data mining for imbalanced datasets: an overview. In: *Data Mining and Knowledge Discovery Handbook*. Springer; 2009:875-886.
7. Sales AP, Tomaras GD, Kepler TB. Improving peptide-mhc class i binding prediction for unbalanced datasets. *BMC Bioinf*. 2008;9(1):385.
8. Lee PH. Resampling methods improve the predictive power of modeling in class-imbalanced datasets. *Int J Environ Res Publ Health*. 2014;11(9):9776-9789.
9. Parvin H, Minaei-Bidgoli B, Alinejad-Rokny H. A new imbalanced learning and dictions tree method for breast cancer diagnosis. *J Bionanoscience*. 2013;7(6):673-678.
10. Norinder U, Boyer S. Binary classification of imbalanced datasets using conformal prediction. *J Mol Graph Model*. 2017;72:256-265.
11. Sun J, Carlsson L, Ahlberg E, Norinder U, Engkvist O, Chen H. Applying mondrian cross-conformal prediction to estimate prediction confidence on large imbalanced bioactivity data sets. *J Chem Inf Model*. 2017;57(7):1591-1598.
12. Vovk V, Gammerman A, Shafer G. *Conformal Prediction, Algorithmic Learning in a Random World*. 2005:17-51.
13. Svensson F, Aniceto N, Norinder U, et al. Conformal regression for quantitative structure activity relationship modeling-quantifying prediction uncertainty. *J Chem Inf Model*. 2018;58(5):1132-1140.
14. Linusson H, Johansson U, Boström H, Löfström T. Reliable confidence predictions using conformal prediction. In: *Pacific-asia Conference on Knowledge Discovery and Data Mining*. Springer; 2016:77-88.
15. Balasubramanian V, Ho S-S, Vovk V. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*. Newnes; 2014.
16. Lapins M, Arvidsson S, Lampa S, et al. A confidence predictor for logd using conformal regression and a support-vector machine. *J Cheminf*. 2018;10(1):17.
17. Lampa S, Alvarsson J, Arvidsson Mc Shane S, Berg A, Ahlberg E, Spjuth O. Predicting off-target binding profiles with confidence using conformal prediction. *Front Pharmacol*. 2018;9:1256.
18. Montanari F, Knasmüller B, Kohlbacher S, et al. Vienna livertox workspace—a set of machine learning models for prediction of interactions profiles of small molecules with transporters relevant for regulatory agencies. *Front Chem*. 2020;7:899.
19. Landrum G, et al. Rdkit: open-source cheminformatics. Available at: http://www.rdkit.org.
20. Svensson F, Norinder U, Bender A. Modelling compound cytotoxicity using conformal prediction and pubchem hts data. *Toxicol Res*. 2017;6(1):73-80.
21. Licht A, Schneider E. Atp binding cassette systems: structures, mechanisms, and functions. *Cent Eur J Biol*. 2011;6(5):785.
22. Glavinas H, Krajcsi P, Cserepes J, Sarkadi B. The role of abc transporters in drug resistance, metabolism and toxicity. *Curr Drug Deliv*. 2004;1(1):27-42.
23. Sharom FJ. The p-glycoprotein multidrug transporter. *Essays Biochem*. 2011;50:161-178.
24. Huang S-M, Lertora JJ, Markey SP, Atkinson Jr AJ. *Principles of Clinical Pharmacology*. Academic Press; 2012.
25. Mao Q, Unadkat JD. Role of the breast cancer resistance protein (bcrp/abcg2) in drug transport—an update. *AAPS J*. 2015;17(1):65-82.
26. Stieger B. Role of the bile salt export pump, bsep, in acquired forms of cholestasis. *Drug Metab Rev*. 2010;42(3):437-445.
27. Shafer G, Vovk V. A tutorial on conformal prediction. *J Mach Learn Res*. 2008;9:371-421.
28. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
29. S. Henrik Linusson Boŕas, nonconformist package 1.2.5. Available at: https://github.com/donlnz/nonconformist.
30. Carlsson L, Eklund M, Norinder U. Aggregated conformal prediction. In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer; 2014:231-240.
31. Vovk V. Cross-conformal predictors. *Ann Math Artif Intell*. 2015;74(1-2):9-28.
32. Eklund M, Norinder U, Boyer S, Carlsson L. The application of conformal prediction to the drug discovery process. *Ann Math Artif Intell*. 2015;74(1-2):117-132.
33. Colombo N, Vovk V. Training conformal predictors. *arXiv*. 2020;2005:7037.