Comparing Performance of Different Inductive and Transductive Conformal Predictors Relevant to Drug Discovery

Lars Carlsson LARS.A.CARLSSON@ASTRAZENECA.COM
Quantitative Biology, Discovery Sciences, AstraZeneca IMED Biotech Unit, Mölndal, Sweden

Claus Bendtsen CLAUS.BENDTSEN@ASTRAZENECA.COM
Quantitative Biology, Discovery Sciences, AstraZeneca IMED Biotech Unit, Cambridge, UK

Ernst Ahlberg ERNST.AHLBERG@ASTRAZENECA.COM
Predictive Compound ADME & Safety, Drug Safety & Metabolism, AstraZeneca IMED Biotech Unit, Mölndal, Sweden

Abstract

We present an evaluation of the impact of transductive, inductive, aggregated and cross inductive mondrian conformal prediction on the validity and efficiency of predictions. The aim of the study is to give guidance to which methods perform best where there is limited data. The evaluation has been made on a large public dataset of Ames mutagenicity data, relevant for drug discovery, a spam dataset and a diverse set of drug discovery datasets. When considering predictions only, the transductive conformal predictor performs the best in terms of validity. If however more information is required, for example interpretation of a prediction, then any of the methods that calculate an averaged p-value should be considered.

Keywords: Drug discovery, Conformal Prediction, Validity, Efficiency

1. Introduction

In drug discovery a huge amount of virtual compounds, not yet synthesized, are generated. Some of the virtual compounds are selected for synthesis and some of those compounds are later tested in cell based assays to understand biological activity and alleviate safety concerns. The biological endpoint for these assays varies from ion-channel and GPCR inhibition to mutagenicity assessed through the Ames reverse mutation test. The bottom line is that these tests are used to screen out compounds that have undesirable properties for novel medicines.

The cost of synthesizing and testing compounds is high and in an effort to reduce those costs machine-learning models are used to assess the compounds before synthesis. In drug discovery this type of models is referred to as Quantitative Structure-Activity Relationship (QSAR) models. The typical use case is that a project chemist has a set of compounds in a graphical user interface (GUI) and submits the compounds to be predicted by a set of QSAR models. When the computation is finished the GUI is updated with the results. Since this is an interactive process there is a limitation to the time a user can be expected to wait for the results. Traditionally, just giving a prediction from any

machine learning model is fast but without any estimate of the associated confidence in the particular prediction. One way to get an associated confidence is to use conformal prediction (Vovk et al., 2005; Shafer and Vovk, 2008). The strongest theoretical guarantees exist for transductive conformal prediction which is assumed to be as close to exact validity as possible but at a high computational cost in terms of prediction time. Between the simple machine-learning prediction case and the transductive conformal prediction there is inductive conformal prediction of different flavours, including aggregated conformal and cross conformal prediction. There is a need to analyze the efficiency of these methods both with respect to the quality of the predictions, defined by validity and efficiency, and with respect to the time to prediction to better understand the trade-offs between these two aspects.

Conformal prediction (Vovk et al., 2005; Shafer and Vovk, 2008) is a method that use existing data to determine valid prediction regions for new examples. Thus instead of giving a point estimate, a conformal prediction model gives a prediction region that contains the true value with probability equal to or higher than a predefined level of confidence. Such a prediction region can be obtained under the assumption that the observed data is exchangeable (Vovk et al., 2005). Conformal prediction relies on a nonconformity measure that describes how different each example looks compared to the other examples. By comparing the nonconformity measure of the predicted example with the ordered nonconformity measures of the known examples it is possible to obtain a p-value for each possible label. A conformal predictor can then report the results in two ways, either predicting the label of the largest p-value where the largest p-value is the credibility of the prediction and one minus the second largest p-value is the confidence, describing how unlikely it is that the prediction should contain another label. The other option is to give a prediction given a user defined level of significance. In that case the prediction will be the set of labels with a p-value larger than the significance. For clarity it should be noted that p-values can be more difficult to interpret than probabilities, and one should consider using Venn predictors (Vovk et al., 2005) when the latter are desired.

To build an ICP the training data is split into a proper training set, used to build a model, and a calibration set. The calibration set is then used to compute the p-value for each label by using the nonconformity measure. For this purpose, consider the prediction setting where training examples $(z_1, \ldots, z_l)$, where z_i is composed of an object x_i of arbitrary description and a label y_i that could be either 0 or 1. The problem is to predict the set of labels, Γ^ϵ, of a new object x_n when a confidence, $1 - \epsilon$ is predefined. The training data has been randomly split in to a proper training set $(z_1, \ldots, z_m)$ and a calibration set $(z_{m+1}, \ldots, z_l)$ of size n. The predicted set of labels of an ICP is,

$$\Gamma^\epsilon(z_1, \ldots, z_l, x) = \{Y \in [0, 1] : p_Y > \epsilon\},$$

where

$$p_Y := \frac{|\{i = m + 1, \ldots, l | \alpha_i \leq \alpha^y\}|}{l - m + 1}.$$

The purpose of this study is to evaluate Inductive Conformal Prediction (ICP), Simplified Aggregated Conformal Prediction (SACP), Aggregated Conformal Prediction (ACP), Cross Conformal Prediction (CCP) and Transductive Conformal Prediction (TCP) with respect to computational effort, validity and efficiency. The study focuses on the boundary

of predictive ability and computational effort for inductive methods versus a transductive method where available data is limited. The setup of the study is made so that the only difference between the various conformal predictors will be in how they use the training data. In every other aspect they will use the same algorithms and datasets when evaluated. Both the ACP and CCP are based on the ICP, but instead of only making one selection for proper training and calibration set, sampling of the available data is applied. For CCP the training set is split into k folds and then a model is built for each fold using one fold at a time as a calibration set and the reminder as proper training set. In the ACP, a subset of the training data is randomly sampled, without replacement, from the training set to obtain a calibration set. The rest of the training data is used for proper training. The procedure is repeated k times. Both CCP and ACP then averages the outcomes from the k p-values for each label. The SACP is simplified in the sense that it uses a lower number of models to average p-values than the ACP.

The remainder of this paper is organized as follows. In the methods section we present the study procedure and the parameters that have been investigated. Next, we show results when the procedure is applied to two different datasets. In the last section we conclude the paper.

2. Method

We propose to compare different flavors of conformal prediction. From a prediction service point of view the wall clock time to prediction is of high importance, but it is equally important that the predictions obtained are as accurate as possible. The best model will be the one that deviates the least from exact validity and has high efficiency. The deviation from exact validity is measured as the Euclidean norm of the difference of the observed error and the expected error for a given set of predefined significance levels. To measure efficiency, the *fuzziness* or *observed fuzziness* (Vovk et al., 2014) of the predictions will be calculated. Fuzziness is defined as the sum of all p-values except the largest p-value for the predicted objects. In this case the true label of a predicted object does not need to be known. When the true label is known it is possible to calculate the observed fuzziness which is the sum of all p-values for the incorrect class labels. These measures will also be compared to a qualitative estimate of the computational effort for training and testing.

The evaluation of conformal predictors have been limited to Mondrian conformal predictors using support vector machines with a radial basis function kernel as described in Eklund et al. (2013). The Mondrian taxonomy was based on the individual labels, thus ensuring class specific validity. Furthermore, scikit-learn (Pedregosa et al., 2011) was used to implement the Mondrian conformal predictor where the nonconformity measures were defined as the decision function values of the C-SVC algorithm. Based on this setup five different flavors of Mondrian conformal predictors were evaluated namely TCP, ICP, SACP, ACP and CCP. The SACP, ACP and CCP are built on ICP but use averaging of results from multiple ICPs for each prediction, in this case 10 models were used, for ACP and CCP, and the average p-values were reported for both ACP and CCP. The difference between the two is that the ACP uses stratified random sampling to obtain the calibration set whereas CCP uses a 10 fold cross validation approach. Running the ACP in this fashion leads to a computational cost equivalent to the cost of running the CCP. The simplified version of

ACP, SACP, averages p-values based on two models in the same fashion as ACP by using the same calibration set and proper training set sizes for each model as both the ACP and the CCP. Thus, the SACP has a training and prediction computational cost which is five times smaller than for the ACP or the CCP.

Given a dataset D, stratified random sampling without replacement was used to draw a subset of 3000 examples from D which in turn was split into a training and a test set, stratified sampling was used here as well and the test set equated to 500 of the examples in the subset. Hyper-parameters for the SVM were preset, $C = 10.0$ and $\gamma = 10^{-3}$.

To evaluate the selected conformal predictors based on validity and efficiency eight datasets were selected, one that describes activities (labels) of compounds (objects) in the Ames mutagenicity test (Ames et al., 1973), the Spambase dataset (Bache and Lichman.) and an additional six datasets taken from the ExCAPE DB (Sun et al., 2017). All datasets were trained as outlined above. The procedure of training was repeated a large number of times and both validity and efficiency, for all the different methods, were evaluated each time using the alternative one-tail test of Wilcoxon signed-rank test in R (R Core Team, 2017) to test whether one method would produce either larger or smaller values than another. The tests were carried out pairwise with respect to the methods for validity and efficiency, respectively. Significance was called at $\alpha = 0.01$.

3. Results

The numerical experiments were all conducted on a standard Linux machine with an Intel Core I7 processor running at 2.5 GHz for the results in the following two subsections. The results in the last subsection were generated on Amazon EC2 using a `c4.2xlarge`. The execution times for the TCP were much longer than for all the other ICP based predictors. For all datasets a prediction with the TCP would take around six seconds, on the standard Linux machine, whereas for all other methods it would take less than a tenth of a second, disregarding the training time for the ICP based predictions.

3.1. Experimental results on Ames data

For Ames data there is a well established link between subgraphs of compounds and mutagenicity. For that reason, the object used when applying the proposed method to Ames data is the signature (Faulon and Churchwell, 2003; Faulon et al., 2003) descriptor. The signature descriptor describes a compound by a set of strings and corresponding counts where the strings represent subgraphs of the compound, centered at an atom (vertex) and expanded to include all neighbor vertexes h bonds (edges) away from the centre vertex. Each string is a canonical representation of a subgraph written as a directed acyclic graph. Signature descriptors are calculated for all atoms in a compound. The signature descriptors used in this work were generated by the Chemistry Development Kit (Steinbeck et al., 2003) and the signatures were limited to $h \in [1, 3]$ thus one to three bonds away from the centre atom.

The Ames data used here is described in Hansen et al. (2009). The set consist of 6512 objects with a total number of 31831 features. The dataset has two labels, mutagen and non-mutagen, describing the ability of the compound to induce reverse mutation in the E-coli based cell lines.

The results of the Wilcoxon signed-rank test are shown in Tables 1 and 2, based on 100 repeated runs.

Method	ICP	SACP	ACP	CCP	TCP
ICP	-	**6.4e-03**	**2.8e-05**	**5.6e-07**	**3.8e-11**
SACP	9.9e-01	-	3.2e-02	**8.6e-04**	**1.6e-05**
ACP	1.0	9.7e-01	-	3.4e-02	**2.2e-05**
CCP	1.0	1.0	9.7e-01	-	**3.7e-04**
TCP	1.0	1.0	1.0	1.0	-

Table 1: Wilcoxon signed-rank test p-values for two alternative hypotheses concerning validity for the different methods applied to the Ames data. The p-values are shown for the methods in the left column having greater validity values than the methods in the first row. Bold face indicates significant results.

Method	ICP	SACP	ACP	CCP	TCP
ICP	-	3.6e-01	4.2e-01	4.9e-01	1.0
SACP	6.4e-01	-	6.1e-01	6.9e-01	1.0
ACP	5.8e-01	3.9e-01	-	6.3e-01	1.0
CCP	5.1e-01	3.1e-01	3.7e-01	-	1.0
TCP	**9.2e-14**	**8.5e-14**	**8.5e-14**	**8.5e-14**	-

Table 2: Wilcoxon signed-rank test p-values for two alternative hypotheses concerning fuzziness applied to the Ames dataset. The p-values are shown for the methods in the left column having greater fuzziness values than the methods in the first row. Bold face indicates significant results.

3.2. Experimental results on Spambase data

This dataset was used with the objects and labels provided at the UCI website. The classification of spam e-mails is diverse, with examples from advertisements and chain letters. The collection of data comes from personal e-mails and postmasters reporting individual e-mails as being spam. The non-spam e-mails were retrieved from work and personal e-mails. The results for comparing the different methods on Spambase data, by applying the Wilcoxon signed-rank test, are shown in Tables 3 and 4, based on 100 repeated runs.

Method	ICP	SACP	ACP	CCP	TCP
ICP	-	**4.6e-06**	**2.8e-09**	**6.0e-09**	**5.2e-14**
SACP	1.0	-	**1.6e-03**	**1.8e-02**	**3.7e-08**
ACP	1.0	1.0	-	8.0e-01	**8.3e-06**
CCP	1.0	9.8e-01	2.1e-01	-	**4.0e-07**
TCP	1.0	1.0	1.0	1.0	-

Table 3: Wilcoxon signed-rank test p-values for two alternative hypotheses concerning validity for the different methods applied to the Spambase dataset. The p-values are shown for the methods in the left column having greater validity values than the methods in the first row. Bold face indicates significant results.

Method	ICP	SACP	ACP	CCP	TCP
ICP	-	9.9e-01	1.0	1.0	1.0
SACP	1.5e-02	-	9.4e-01	9.6e-01	1.0
ACP	**4.8e-04**	5.7e-02	-	5.8e-01	1.0
CCP	**1.8e-04**	3.6e-02	4.2e-01	-	1.0
TCP	$<$ **2.2e** $-$ **16**	$<$ **2.2e** $-$ **16**	$<$ **2.2e** $-$ **16**	$<$ **2.2e** $-$ **16**	-

Table 4: Wilcoxon signed-rank test p-values for two alternative hypotheses concerning fuzziness applied to the Spambase dataset. The p-values are shown for the methods in the left column having greater fuzziness values than the methods in the first row. Bold face indicates significant results.

3.3. Experimental results on a diverse set of data

These datasets were collected from the ExCAPE DB (Sun et al., 2017). All of them are representing binary classification with at least more than 3000 examples. Most of these datasets are highly imbalanced with respect to the number of examples of each class. We have chosen a subset of datasets where the imbalance is more moderate. A short description of the datasets is shown in Table 5. The results for comparing the different methods on

Entrez ID	Nr of examples	Approx. nr of features
154	344186	6.8e+05
367	6739	5.0e+05
1576	19324	2.6e+05
1588	5767	2.1e+05
2908	6766	6.8e+05
3091	10303	6.8e+05

Table 5: A summary of some of the characteristics of the ExCAPE DB data.

these datasets by applying the Wilcoxon signed-rank test are shown in Tables 6, 7 and 8, based on 40 repeated runs.

Method	ICP	SACP	ACP	CCP	TCP
Spambase					
ICP	-	8.41e-02	1.45e-01	8.41e-02	**2.39e-04**
SACP	9.17e-01	-	6.89e-01	5.44e-01	3.19e-02
ACP	8.57e-01	3.14e-01	-	4.34e-01	**9.52e-03**
CCP	9.17e-01	4.60e-01	5.70e-01	-	2.38e-02
TCP	1.00	9.69e-01	9.91e-01	9.77e-01	-
Ames					
ICP	-	**7.27e-03**	**2.30e-04**	**5.44e-04**	**2.68e-06**
SACP	9.93e-01	-	1.40e-01	2.55e-01	3.26e-02
ACP	1.00	8.62e-01	-	6.37e-01	4.07e-01
CCP	9.99e-01	7.48e-01	3.66e-01	-	1.25e-01
TCP	1.00	9.68e-01	5.96e-01	8.77e-01	-
154					
ICP	-	2.92e-02	**9.03e-03**	1.17e-02	**3.75e-07**
SACP	9.71e-01	-	2.32e-01	3.60e-01	**1.06e-03**
ACP	9.91e-01	7.71e-01	-	6.07e-01	**2.87e-03**
CCP	9.89e-01	6.44e-01	3.96e-01	-	**5.43e-04**
TCP	1.00	9.99e-01	9.97e-01	9.99e-01	-
367					
ICP	-	1.75e-02	**6.18e-03**	**9.53e-04**	8.46e-01
SACP	9.83e-01	-	5.08e-01	4.26e-01	9.98e-01
ACP	9.94e-01	4.96e-01	-	3.07e-01	1.00
CCP	9.99e-01	5.78e-01	6.97e-01	-	1.00
TCP	1.56e-01	**1.74e-03**	**8.58e-05**	**1.09e-06**	-
1576					
ICP	-	2.12e-01	1.51e-02	2.17e-02	3.28e-01
SACP	7.91e-01	-	5.47e-02	5.69e-02	5.89e-01
ACP	9.85e-01	9.46e-01	-	6.11e-01	9.88e-01
CCP	9.79e-01	9.44e-01	3.93e-01	-	9.72e-01
TCP	6.75e-01	4.15e-01	1.27e-02	2.85e-02	-
1588					
ICP	-	**3.63e-04**	**2.90e-04**	**9.55e-05**	**5.91e-07**
SACP	1.00	-	4.75e-01	4.26e-01	3.49e-01
ACP	1.00	5.29e-01	-	5.13e-01	3.74e-01
CCP	1.00	5.78e-01	4.90e-01	-	4.30e-01
TCP	1.00	6.54e-01	6.29e-01	5.74e-01	-
2908					
ICP	-	**2.38e-03**	**2.54e-03**	**5.84e-04**	2.72e-01
SACP	9.98e-01	-	5.17e-01	1.88e-01	1.00
ACP	9.98e-01	4.87e-01	-	2.38e-01	9.99e-01
CCP	9.99e-01	8.15e-01	7.65e-01	-	1.00
TCP	7.32e-01	**3.37e-04**	**1.02e-03**	**4.08e-05**	-

3091					
ICP	-	2.02e-02	2.10e-03	3.40e-02	1.00
SACP	9.80e-01	-	2.17e-01	5.74e-01	1.00
ACP	9.98e-01	7.85e-01	-	8.84e-01	1.00
CCP	9.67e-01	4.30e-01	1.18e-01	-	1.00
TCP	**1.58e-06**	**1.31e-13**	**2.36e-14**	**1.05e-13**	-

Table 6: Wilcoxon signed-rank test p-values for two alternative hypotheses concerning validity for the different methods applied to all datasets. The p-values are shown for the methods in the left column having greater validity values than the methods in the first row. Bold face indicates significant results.

Method	ICP	SACP	ACP	CCP	TCP
Spambase					
ICP	-	4.68e-01	8.44e-01	8.36e-01	1.00
SACP	5.36e-01	-	9.37e-01	9.36e-01	1.00
ACP	1.58e-01	6.39e-02	-	3.59e-01	1.00
CCP	1.67e-01	6.48e-02	6.44e-01	-	1.00
TCP	**9.30e-24**	**9.30e-24**	**9.30e-24**	**7.17e-15**	-
Ames					
ICP	-	8.97e-01	9.88e-01	9.73e-01	1.00
SACP	1.05e-01	-	9.66e-01	8.10e-01	1.00
ACP	1.24e-02	3.52e-02	-	8.56e-02	1.00
CCP	2.79e-02	1.93e-01	9.16e-01	-	1.00
TCP	**1.79e-16**	**4.20e-18**	**1.49e-20**	**1.77e-22**	-
154					
ICP	-	9.43e-01	9.98e-01	1.00	1.00
SACP	5.80e-02	-	7.39e-01	1.00	1.00
ACP	**2.14e-03**	2.64e-01	-	1.00	1.00
CCP	**7.78e-05**	**8.48e-08**	**9.41e-08**	-	1.00
TCP	**7.69e-15**	**7.17e-15**	**7.17e-15**	**9.30e-24**	-
367					
ICP	-	7.01e-02	4.56e-01	3.18e-01	1.00
SACP	9.31e-01	-	9.74e-01	9.61e-01	1.00
ACP	5.48e-01	2.67e-02	-	3.74e-01	1.00
CCP	6.86e-01	3.95e-02	6.29e-01	-	1.00
TCP	**3.16e-11**	**4.19e-20**	**1.94e-20**	**3.72e-23**	-
1576					
ICP	-	9.64e-01	9.45e-01	9.90e-01	1.00
SACP	3.71e-02	-	3.39e-01	5.63e-01	1.00
ACP	5.58e-02	6.65e-01	-	8.02e-01	1.00
CCP	1.06e-02	4.41e-01	2.01e-01	-	1.00

TCP	**1.69e-17**	**3.26e-20**	**4.73e-21**	**8.51e-21**	-
1588					
ICP	-	9.67e-01	9.00e-01	9.70e-01	1.00
SACP	3.33e-02	-	7.45e-02	1.16e-01	1.00
ACP	1.02e-01	9.27e-01	-	7.80e-01	1.00
CCP	3.05e-02	8.86e-01	2.23e-01	-	1.00
TCP	**2.38e-12**	**3.26e-20**	**9.30e-24**	**9.30e-24**	-
2908					
ICP	-	9.57e-01	9.71e-01	9.69e-01	1.00
SACP	4.38e-02	-	4.34e-01	3.89e-01	1.00
ACP	2.98e-02	5.70e-01	-	5.06e-01	1.00
CCP	3.16e-02	6.15e-01	4.98e-01	-	1.00
TCP	**1.13e-15**	**2.35e-17**	**9.30e-24**	**9.30e-24**	-
3091					
ICP	-	9.92e-01	9.86e-01	9.97e-01	1.00
SACP	**8.33e-03**	-	2.88e-01	4.64e-01	1.00
ACP	1.40e-02	7.16e-01	-	6.68e-01	1.00
CCP	**3.05e-03**	5.40e-01	3.35e-01	-	1.00
TCP	**4.08e-19**	**1.77e-22**	**9.30e-24**	**9.30e-24**	-

Table 7: Wilcoxon signed-rank test p-values for two alternative hypotheses concerning fuzziness for the different methods applied to all datasets. The p-values are shown for the methods in the left column having greater validity values than the methods in the first row. Bold face indicates significant results.

Method	ICP	SACP	ACP	CCP	TCP
Spambase					
ICP	-	8.12e-02	2.72e-01	1.73e-01	1.00
SACP	9.20e-01	-	8.07e-01	6.99e-01	1.00
ACP	7.32e-01	1.96e-01	-	2.17e-01	1.00
CCP	8.30e-01	3.04e-01	7.85e-01	-	1.00
TCP	**2.77e-17**	**1.13e-20**	**4.19e-22**	**9.02e-22**	-
Ames					
ICP	-	6.47e-01	8.65e-01	7.56e-01	1.00
SACP	3.56e-01	-	7.92e-01	6.18e-01	1.00
ACP	1.37e-01	2.11e-01	-	2.58e-01	1.00
CCP	2.47e-01	3.85e-01	7.45e-01	-	1.00
TCP	**2.22e-11**	**5.12e-12**	**2.25e-09**	**7.92e-13**	-
154					
ICP	-	9.56e-01	9.98e-01	1.00	1.00
SACP	4.47e-02	-	7.53e-01	1.00	1.00
ACP	**2.04e-03**	2.50e-01	-	1.00	1.00

CCP	**7.63e-05**	**1.12e-08**	**2.05e-08**	-	1.00
TCP	**8.28e-15**	**9.30e-24**	**9.30e-24**	**9.30e-24**	-
367					
ICP	-	4.57e-02	2.53e-01	1.96e-01	1.00
SACP	9.55e-01	-	9.49e-01	9.08e-01	1.00
ACP	7.50e-01	5.16e-02	-	3.35e-01	1.00
CCP	8.07e-01	9.34e-02	6.68e-01	-	1.00
TCP	**2.75e-09**	**2.00e-18**	**1.09e-19**	**2.53e-21**	-
1576					
ICP	-	5.71e-01	5.32e-01	5.93e-01	1.00
SACP	4.33e-01	-	5.59e-01	6.86e-01	1.00
ACP	4.71e-01	4.45e-01	-	5.89e-01	1.00
CCP	4.11e-01	3.17e-01	4.15e-01	-	1.00
TCP	**8.97e-12**	**9.15e-11**	**3.35e-14**	**4.23e-14**	-
1588					
ICP	-	9.67e-01	8.30e-01	9.35e-01	1.00
SACP	3.33e-02	-	8.56e-02	1.47e-01	1.00
ACP	1.73e-01	9.16e-01	-	7.80e-01	1.00
CCP	6.63e-02	8.55e-01	2.23e-01	-	1.00
TCP	**3.85e-12**	**7.55e-19**	**4.19e-22**	**1.86e-23**	-
2908					
ICP	-	9.26e-01	8.78e-01	8.42e-01	1.00
SACP	7.55e-02	-	2.94e-01	2.35e-01	1.00
ACP	1.24e-01	7.09e-01	-	3.82e-01	1.00
CCP	1.61e-01	7.68e-01	6.22e-01	-	1.00
TCP	**1.60e-12**	**3.83e-17**	**2.84e-14**	**1.13e-20**	-
3091					
ICP	-	8.17e-01	4.68e-01	5.44e-01	1.00
SACP	1.85e-01	-	8.26e-02	1.02e-01	1.00
ACP	5.36e-01	9.19e-01	-	5.78e-01	1.00
CCP	4.60e-01	9.00e-01	4.26e-01	-	1.00
TCP	**1.31e-12**	**2.78e-16**	**1.13e-20**	**3.47e-21**	-

Table 8: Wilcoxon signed-rank test p-values for two alternative hypotheses concerning observed fuzziness for the different methods applied to all datasets. The p-values are shown for the methods in the left column having greater validity values than the methods in the first row. Bold face indicates significant results.

4. Discussion

In this study five flavors of conformal prediction have been studied with respect to validity, fuzziness and observed fuzziness. From the evaluation based on both datasets it is clear that the validity of the TCP is better than all the ICP based methods. However, there

is a significant difference in validity between all ICP based methods and the ICP method itself. One reason may be that the datasets are relatively small and the information used for calibration and nonconformity measures is more sensitive to the random partition for the ICP which is reduced for the other ICP based methods. In contrast, all ICP based methods, as well as ICP alone, are more efficient than the TCP. The possible explanation for this is that when validity is compromised this results in smaller prediction sets. However, sacrificing validity in terms of improving efficiency is not of interest when making predictions within drug discovery. It is important to have a predictor that behaves as outlined by the theory.

Given the size of the datasets studied and prediction times, TCP is the preferred method if we would only consider predictions. But, when interpreting predictions, by for example calculating a gradient (Ahlberg et al., 2015), the number of actual predictions increase with at least the number of non-zero descriptors which makes it prohibitive to use a TCP. Recall, that for the dataset sizes studied here, a single prediction using TCP will take around six seconds. In this case all the methods averaging p-values of an ICP become viable options.

Furthermore, the use of different types of ICPs that combine p-values of a number of individual ICPs is of great interest. These methods could be of potential interest in both a distributed data environment as well as in making predictions on large datasets in a parallel computing environment. We remark that the Spambase dataset is a not a dataset of relevance for drug discovery. It was added as a reference so that this study could be extended to other datasets by others of interest to this type of methods.

Acknowledgments

The authors would like to thank the reviewers for invaluable feedback and constructive comments that helped progressing and clarifying the work reported.

References

Ernst Ahlberg, Ola Spjuth, Catrin Hasselgren, and Lars Carlsson. Interpretation of conformal prediction classification models. In *International Symposium on Statistical Learning and Data Sciences*, pages 323–334. Springer International Publishing, 2015.

Bruce N. Ames, Frank D. Lee, and William E. Durston. An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proceedings of the National Academy of Sciences*, 70(3):782–786, 1973. doi: 10.1073/pnas.70.3.782. URL http://www.pnas.org/content/70/3/782.abstract.

K. Bache and M. Lichman. *UCI machine learning repository.* http://archive.ics.uci.edu/ml, 2013. (accessed Apr 10, 2014).

Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. The application of conformal prediction to the drug discovery process. *Annals of Mathematics and Artificial Intelligence*, pages 1–16, 2013. ISSN 1012-2443. doi: 10.1007/s10472-013-9378-2. URL http://dx.doi.org/10.1007/s10472-013-9378-2.

Jean-Loup Faulon and Carla J Churchwell. Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences. 43:721–734, 2003.

Jean-Loup Faulon, Donald P Jr Visco, and Ramdas S Pophale. Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. 43:707–720, 2003.

Katja Hansen, Sebastian Mika, Timon Schroeter, Andreas Sutter, Antonius ter Laak, Thomas Steger-Hartmann, Nikolaus Heinrich, and Klaus-Robert Müller. Benchmark data set for in silico prediction of ames mutagenicity. *Journal of Chemical Information and Modeling*, 49(9):2077–2081, 2009. doi: 10.1021/ci900161g. URL http://dx.doi.org/10.1021/ci900161g. PMID: 19702240.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, March 2008. URL http://www.jmlr.org/papers/volume9/shafer08a/shafer08a.pdf.

Christoph Steinbeck, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen. The chemistry development kit (cdk) an open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.*, 43(2):493–500, 2003. doi: 10.1021/ci025584y. PMID: 12653513.

Jiangming Sun, Nina Jeliazkova, Vladimir Chupakin, Jose-Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jörg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, Nikolay Kochev, Thomas J. Ashby, and Hongming Chen. Excape-db: an integrated large scale dataset facilitating big data analysis in chemogenomics. *Journal of Cheminformatics*, 9 (1):17, 2017. ISSN 1758-2946. doi: 10.1186/s13321-017-0203-5. URL http://dx.doi.org/10.1186/s13321-017-0203-5.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 0387001522.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alex Gammerman. Criteria of efficiency for conformal prediction. Technical report, Royal Holloway, 2014.