Ann Math Artif Intell (2017) 81:145–154
DOI 10.1007/s10472-017-9550-1

Current application of conformal prediction in drug discovery

Two useful applications

Ernst Ahlberg[1] · **Oscar Hammar**[2] · **Claus Bendtsen**[3] · **Lars Carlsson**[2]

Published online: 28 April 2017

Abstract We present two applications of conformal prediction relevant to drug discovery. The first application is around interpretation of predictions and the second one around the selection of compounds to progress in a drug discovery project setting.

Keywords Drug discovery · Conformal prediction · Interpretation

1 Introduction

Pharmaceutical research and development is a process that typically takes at least 10 years and where costs exceed $1bn [1, 2]. With high failure rates and low over all throughput the decision making process becomes very important. However, due to both ethical and cost considerations it is difficult to obtain data and information that links the effect in model systems to the human outcome. In reality this means that the data fully reflecting human

✉ Ernst Ahlberg
ernst.ahlberg@gmail.com

Oscar Hammar
oscar.hammar@astrazeneca.com

Claus Bendtsen
claus.bendtsen@astrazeneca.com

Lars Carlsson
lars.a.carlsson@astrazeneca.com

[1] Predictive Compound ADME & Safety, Drug Safety & Metabolism, AstraZeneca, Innovative Medicines & Early Development, Mölndal, Sweden

[2] Quantitative Biology, Discovery Sciences, AstraZeneca, Innovative Medicines & Early Development, Mölndal, Sweden

[3] Quantitative Biology, Discovery Sciences, AstraZeneca, Innovative Medicines & Early Development, Cambridge, UK

disease and therapeutic benefits is very limited and oftentimes skewed, in the sense that only marketed drugs and late stage failures become public knowledge. In the case where one wants to study a side effect, it is usually possible to find examples where the side effect has not been seen, but more difficult to find examples where the drug was actually responsible for causing the side effect. One example is Torsades de Pointes, an acute form of arrhythmia that leads to cardiac arrest. In light of the data available, hERG, the human Ether-a-go-go related gene, is believed to be a possible cause of Torsade de Pointes [3]. hERG is an ionchannel which when blocked can cause a prolonged heart beat, long QT syndrome, which in turn can cause irregular heartbeats and failure of the cardiac function. Using the available data, cell based experiments were devised to screen for hERG activity on a massive scale, validated using a limited number of drugs but applied to thousands of compounds as a marker for hERG activity. Based on the massive datasets generated in the cell based screens computational models have been created to predict hERG activity. In essence, the gold standard data is the limited set of compounds that reached the clinic and where an effect was seen. A surrogate model, hERG inhibition, was created to help project design away from hERG activity. In early stages computational models are used as yet another surrogate model to screen out compounds long before synthesis, thus creating a pseudo model of a pseudo model. Each such step is associated with error and there is a risk that compounds become optimized towards the pseudo models rather than the actual issue. Therefore it is very important that decisions that are taken are based on a solid scientific foundation and that one can assess the confidence of an individual prediction. The example with hERG is just one, but in practice compounds are tested against batteries of cell based screens in an effort find a suitable drug candidate.

1.1 Drug discovery process

The path to bring a chemical compound into a launched drug is long and expensive. It can be divided into two parts, first the part where one decides on which compound to choose, commonly denoted discovery and the second part, commonly denoted development, where clinical studies are performed and one also optimizes the tablet formulation and synthesis for large scale production.

1.1.1 Discovery

The earliest phase of drug discovery is target selection and validation. The focus of this effort is to locate a drugable biological target on which it is possible to intervene with treatment through delivery of a drug. To locate the target one studies disease mechanisms and biological pathways. Such studies depend on the availability of information and data from previous studies but equally from publications in the public domain. When a target has been located, animal studies can be used to evaluate if disturbing a pathway or knocking out a specific gene can result in a desired response. When the target activity has been confirmed and validated the process continues to design a drug or chemical compound that can intervene on the target and regulate the signaling process such that the cascade is either amplified or reduced. To easily evaluate the efficiency of the interaction with the target an *in vitro* assay is constructed. Such an assay usually measures inhibition or activation of the target and is used to identify compounds that have a desired effect. Compounds with high activity are selected and analyzed for common chemical features and properties and the compounds are grouped into chemical series that are optimized later. There is a drive to keep early selection as diverse as possible to avoid failure further on due to restricted chemistry which cannot be sufficiently optimised. When a limited set of chemical series have been selected the process

focuses on optimization. The desire is to have a compound that is selective and only interacts with the desired target. In this phase *in vitro* assays are in place to remove compounds that hit too many other targets and that may cause safety issues, like Torsade de Pointes, mutagenicity, cyto-toxicity renal failure etc. In parallel possible routes of administration are analyzed and the uptake of the drug into the body is investigated.

The whole process is very iterative and there is an experimental cycle within each of the phases. In short the experimental cycle goes through four phases: design, make, test and analysis. Optimizing flows and decisions in this cycle has the potential to dramatically reduce cost through smarter testing and shorter cycle times. One way to influence the cycle is to use available data and algorithmic learning to predict the outcome of standard tests. In the end of discovery, compounds are shortlisted based on the most promising candidates and a single compound is put forward to development.

1.1.2 Development

The development phase is heavily dominated by clinical and safety studies. The pre-clinical and clinical safety package is designed to test both acute and long term effects of the drug in the body and spans from 14 day maximum tolerated dose studies to year long carcinogenicity studies. Once the compound is selected the focus shifts from developing the compound to optimizing synthesis and particle properties of the formulation. The formulation is the part of the tablet or suspension that is used to deliver the drug to the body. In most cases an oral formulation is preferred since a tablet can be self administrated by the patient. With a working formulation, clinical trials are initiated, first in small groups of healthy volunteers and then in a diseased population. Initially the clinical development has a focus on understanding the safe dose range (phase I) and then to assess the therapeutic dose requirements and study efficacy (phase II). This information is then used to test the therapeutic benefit and value as a novel medicine (phase III). If the drug is proven effective and without severe side effects it will be presented to the authorities for approval.

In all phases of discovery and development the decision process is key to the success of the drug. Where data analysis in the clinical phase is very strict, its counterpart in discovery is oftentimes less stringent due to the experimental and research focus of the effort. Even though the overarching goal is clear, the path to get there is usually not. Historically this has led to difficulties in obtaining go/no go decisions, i.e. to decide whether or not to progress the next phase of discovery or into development [4].

To support the process, models are built to predict the outcome of tests and decisions at different stages [5]. Effective prioritization builds on known and tangible confidence in the available information and the ability to rank and weigh the different options. This is where conformal prediction can be applied to give more solid decision support directly in the project setting. Furthermore, conformal prediction can also be applied to guide the testing strategy and for example be applied to focus testing in areas of low confidence or reduce testing for a particular target.

The remainder of the paper will focus on application and interpretation of conformal prediction in a drug discovery setting and how to utilize existing data to aid informed decisions.

1.2 Conformal prediction

In the late 20th century, machine-learning algorithms grew popular as a tool to predict various properties of compounds. Many companies provided various tools and pharmaceutical

companies applied these so called Quantitative Activity-Structure Relationships (QSAR) [6] where a the model is an approximation of a relationship between an object (compound) and its label (biological activity). For machine learning in general, methods like support vector machines [7] and random forests [8] gained a lot of interest. Predominately, due to the fact that they produced highly accurate predictions without requiring any deeper understanding of the underlying problem. Accuracy alone may be sufficient in many applications. However, in drug discovery there is also a need to understand how a molecule could be altered to improve certain properties and this could be accomplished by allowing for interpretable models by applying for example sensitivity analysis.

Another desire is to understand how confident one could be in a prediction and a lot of efforts were made to define criteria as to when a model could be applied or not [9]. During the 1990s, theories were developed that were summarized in Algorithmic Learning in a Random World [10]. One of the methods developed was Conformal Prediction (CP), where prediction sets or ranges are formed corresponding to a prescribed confidence. To derive a prediction for a new object a nonconformity score is assigned to it by applying some function, in most cases this is a prediction from a machine-learning model. This nonconformity score is then compared to nonconformity scores of other objects, where the label is known, by applying the same function. A randomness test is then performed that determines whether a particular label or region should be part of the prediction.

Validity and efficiency The nice property with CPs is that if the data that is being modelled is generated randomly from some unknown probability distribution, then a CP will produce correct predictions for a fraction of the predictions it makes corresponding to at least the confidence. This property is called *Validity* [10]. Comparing two different CPs can not be done in the traditional sense of studying for example accuracy but rather by looking at the size of the prediction sets or ranges, the *Efficiency* of the CP. Given two valid methods, the one that produces smaller prediction sets or regions is more efficient [10].

Confidence and credibility From the outset, conformal prediction uses a preset significance level and reports the label set for that level of confidence, all labels that have a p-value larger than the significance level. At a given level of confidence it may not be possible to obtain any label. In drug discovery it is desirable to always obtain a label, thus one can report the most credible label and its associated confidence and credibility. For binary classification the credibility can be defined as the largest p-value for an object and the confidence is one minus the p-value for the other label. A formal definition is given in [10].

1.3 Teaching schedules

Teaching schedules [10] are used in retrospective evaluation of models and modeling strategies. The application of teaching schedules within drug discovery builds on the fact that the chemistry in the projects changes over time. Machine-learning based models are taught in a semi-offline setting where a set of compounds are measured for various properties in different assays. New compounds are being designed and synthesized based on the experimental results and these new compounds are subsequently measured in the same assays. This is an iterative procedure with a lag since the model does not immediately learn new labels. The goal for the chemists is to make novel compounds, using known chemical reactions and building blocks. This means that the chemistry shifts over time, and thus if one wants to assess the performance of a model the time aspect of data acquisition cannot be ignored. That does not mean that if a compound had been tested earlier or later that it would obtain

a different result, it only means that the chemical design of new compounds is altered over time. Teaching schedules is a way to make use of the information of when a sample was obtained and use that date in the testing process [11]. Figure 1 illustrates the concept and how it is used. Basically, when starting a retrospective analysis, a starting date, t_0, is set together with a time to next model update, t_δ. In essence the initial training data, defined as all data generated before t_0 is used to predict the compounds generated in the time window between t_0 and $t_0 + t_\delta$. The process is then repeated for each t_δ until present day. This allows the modeler not only to see how well a model will perform over time but also how the model would be affected by a change in the testing strategy. In this way, teaching schedules are used to mimic the data generating process in drug discovery and give a relevant picture of model performance over time. By applying teaching schedules it is also possible to investigate the impact of changes in updating frequency, t_δ. For some very large data sets, the impact on the model made by a comparatively small number of new examples can lead to that t_δ can be extended, lowering the maintenance burden for the modelers.

2 Application of conformal prediction

Application of any computer-based modeling technique for decision support in drug discovery requires that the predictions can be easily interpreted and that a measure of confidence is readily available for each prediction. To replace an existing method, it is required that the new method is either cheaper or more reliable. The limiting factor is often speed and interpretability. With conformal prediction however, a new possibility arises that gives us the possibility to affect the testing strategies in drug discovery. Below we describe how conformal predictions can be used to aid the drug discovery process, saving resource by both suggesting changes and suggest which compounds to test.

2.1 Interpretation of predictions

In the QSAR modeling community, non-linear machine-learning methods like support vector machines [7] and random forests [8] are by many still believed to be black box methods [12], where it is difficult or impossible to interpret the results in an understandable way. In 2009 we suggested a compound specific approach for interpretation of non-linear machine-learning models [13] with an extra added benefit of providing easy visualisations of the

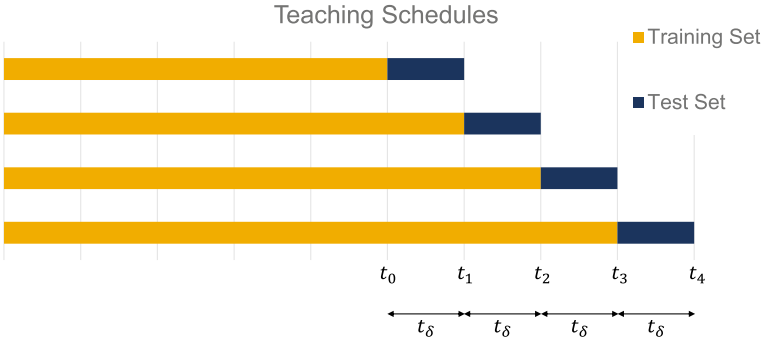


Fig. 1 Visualization of the concept of teaching schedules

interpretation. Given a prediction for a new compound, it is valuable to understand if slight modifications to the object will alter the prediction. By viewing the model as a function that maps the objects to the labels one can calculate a gradient with respect to the components of the object. Figure 2 is an illustration of such a model function where the dot indicates the local point of interest and the vertical is the response. Computing the gradient for the point and identifying the component responsible for the largest change will show how the object can be changed to alter the activity of the object. In identifying the component that gives the largest contribution to the prediction a recommendation can be made on what to change to move into a more favourable region. The approach was adapted for CP in 2015 [14]. In ordinary classification the decision function determining the predicted label is of interest. For conformal classification it is more interesting to investigate the change in p-value space since the p-values determine the prediction set. In the classification case, a conformal predictor results in a p-value for each label, thus calculating the gradient for the most credible p-value will identify the component that have the highest influence on the prediction and the resulting prediction set, directly comparable to the method presented in [13].

In the drug discovery setting this method gives the chemist an easy interpretation of the prediction and an idea on how to alter the compound in order to obtain a more favourable prediction and in the end a safer and better compound. When chemical features are used to describe the object the component with the largest influence on the prediction can be mapped back to the original structure, as can be seen in Fig. 3. By visual inspection it is possible to detect the part of the compound that if changed has the highest potential to alter the prediction. In Fig. 3 the compound is altered from (a) to (b) and the components that drive the prediction are highlighted in red and blue respectively. When compound is no longer predicted to have the unfavourable activity, highlighted in red, the drivers of the new prediction are highlighted in blue.

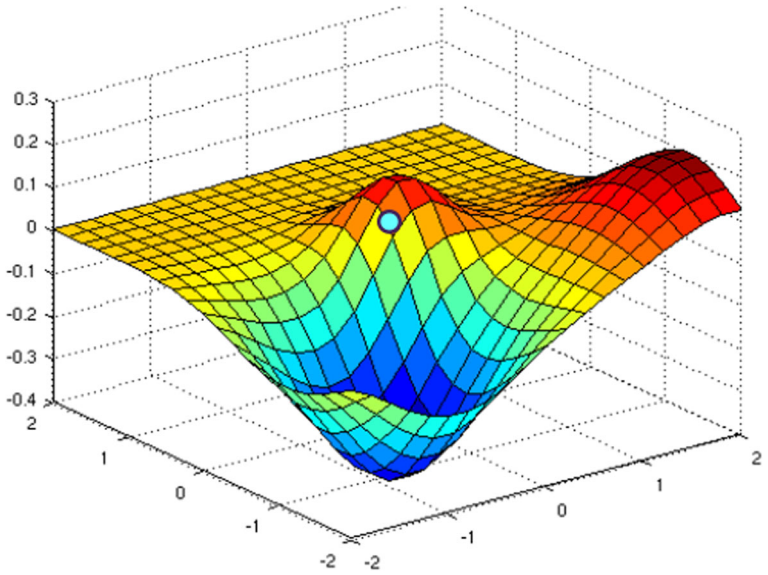
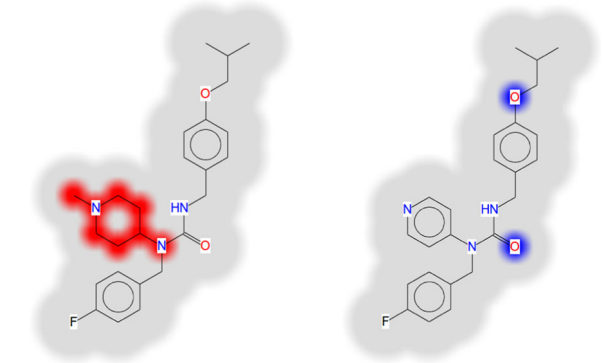


Fig. 2 An illustration of a smooth model function. The *horizontal plane* represents the object space and the vertical axis represents the label as a real value. At any point in feature space it is possible to get an understanding of what the model function looks like in a local neighborhood to this point



(a) Compound with unfavourable prediction (b) Altered compound

Fig. 3 Visualization of a compound with an undesired property, highlighted red, and how to alter the compound to obtain a more favourable prediction. The aliphatic ring, piperidine, in (**a**) has been replaced with an aromatic ring, pyridine, in (**b**)

2.2 Reduction of experimental testing

In drug discovery, cell based experimental testing, *in vitro* testing is the bread and butter of early testing both for main target effects and off target effects like safety liabilities. A compound that gets selected as a drug candidate has been tested in at least 100 different *in vitro* tests and most drug discovery projects creates hundreds of compounds. Even if testing can come as cheap as a few dollars per test this still sums up to a substantial investment. A method to reduce testing and still keep the predictive ability going forward has the ability to have a big impact on the drug discovery process. In the drug discovery setting one wants to progress compounds with favourable properties and stop compounds with unfavourable properties. Conformal prediction gives that possibility through progressing compounds predicted to have favourable properties with high confidence and credibility. At the same time, one wants to stop compounds with unfavourable properties. Thus testing can be performed primarily on compounds with low confidence and credibility regardless of whether they are predicted to have favourable or unfavourable properties as compounds with sufficient confidence and credibility would either pass or be stopped due to undesired properties. In this work mondrian conformal prediction [10] has been been used to obtain p-values for the respective labels, where the mondrian addition here is that the p-value calibrations are label specific, thus the calibration set for a specific label only includes compounds experimentally tested with that label. Furthermore the teaching schedule approach described has also been used, with a monthly updating frequency. Instead of using a specific user defined significance level, the confidence and credibility notations by Vovk et al have been used to rank predictions. In this simple case, a mondrian conformal prediction classification model is used. From the model, p-values are obtained for each class that can be converted to confidence and credibility. Compounds predicted with high confidence and credibility have been not been included in future training as to simulate that those compounds were never tested.

Early tests on internal data shows that application of mondrian conformal prediction allows for significant reductions in the number of compounds sent for experimental testing,

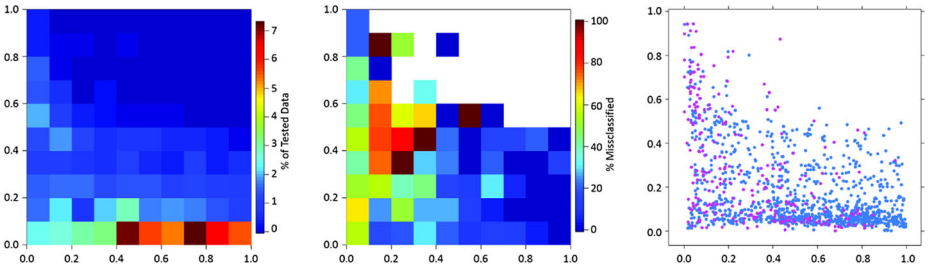
Table 1 Different thresholds for p-values and the associated fraction of molecules withheld from experimental testing

Cutoff	% reduction	% error	% error desired class	% error undesired class
0.9	11.1	1.6	0.0	4.0
0.8	22.5	1.4	0.3	3.6
0.7	35.1	2.2	0.3	6.7
0.6	47.1	3.0	0.4	8.3
0.5	58.1	4.1	1.2	10.3

It is also shown what the fraction of errors are overall and for the two types of outcomes

whilst still retaining the predictive ability of the model over time when predicting compounds novel to the model. If such a strategy is implemented, early studies indicate that testing could be reduced by 50% without severely affecting the predictive performance of the model. For more details see Table 1, where the cutoff describes the lower limit of confidence and credibility to exclude compounds from testing and the % reduction is the percentage of compounds excluded from testing. For the compounds excluded from testing the table shows the overall error rate and the error rates for the two classes, the desired and the undesired class. As the cutoff is lowered the errors increase, but the errors are not evenly distributed between the classes, thus it is less likely that an error is made in the desirable region compared to the undesired region. The model will continue to learn over time as this approach refocuses testing in areas where the confidence and credibility is low, thus potentially this could further reduce testing.

Over time there is a testing bias for the desired class as chemists adopt and progress compounds and series with favourable properties. This means that a testing strategy where all new compounds are measured is focused on a chemical environment that is well covered by the model. Figure 4 describes this in detail, all three figures shows a p-value grid where where the x-axis corresponds to the desired class and the y-axis corresponds to the undesired class respectively. Part (a) of the figure shows the density of tested compounds on the p-value grid and highlights the fact that most compounds are predicted to fall into the desirable category. To a large extent this is also correct, as can be seen in subfigure (b) where the fraction of misclassified compounds are shown. Specifically note that the error rate in the



(a) Density of tested examples (b) Fraction of missclassified examples (c) Examples colored by experimental activity

Fig. 4 The data in these plots are represented by their predicted p-values. The x axis represents a desired property and the y-axis an undesired property

bottom right corner is very low. Finally subfigure (c) shows the distribution of the test set, where the desired activity is colored blue.

3 Discussion and Conclusions

Conformal prediction is gaining interest in the cheminformatics community. There has been a desire to complement predictions with some concept of confidence, however no strong theories apart from theories describing conformal prediction have been developed.

Although functions represented by machine-learning models often are non smooth, we have seen practical utility in the gradient based method used to interpret prediction when applied to compound property predictions. The extension of this method to models based on conformal predictors has not been extensively studied but early results look promising. The value added is in this case that a design chemist could get a better understanding on how to modify a compound to improve a property.

The second application mentioned here, where we try to reduce the number of compounds tested in experiments, is relatively new. Initial results look very promising, where strategies based on static thresholds on various outputs of conformal predictors are used to select compounds for testing. This approach has a huge potential to speed up the discovery process and reduce costs. If coupled with a virtual testing strategy where all compounds are predicted before synthesis it could reduce cost and effort even more. This strategy would also encourage chemists to probe new chemistry as compounds that can not be predicted with high confidence and credibility will be tested.

Future directions of research on conformal prediction applied within drug discovery is of extreme interest. In particular, developments that would combine conformal prediction with reinforcement or active learning would allow us to implement experimental testing strategies that potentially could improve upon the results we have shown here.

References

1. Paul, S.M., Mytelka, D.S., Dunwiddie, C.T., Persinger, C.C., Munos, B.H., Lindborg, S.R., Schacht, A.L.: How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery, 1–12 (2010)
2. DiMasi, J.A.: Cost of Developing a New Drug. Tech. Rep. R&D Cost Study Briefing, Tufts Center for the Study of Drug Development, Boston, MA (2014)
3. Curran, M.E., Splawski, I., Timothy, K.W., Vincen, G., Green, E.D., Keating, M.T.: A molecular basis for cardiac arrhythmia: HERG mutations cause long QT syndrome. Cell **80**(5), 795–803 (1995). doi:10.1016/0092-8674(95)90358-5, http://www.sciencedirect.com/science/article/pii/0092867495903585
4. Scannell, J.W., Bosley, J.: When quality beats quantity: decision theory, drug discovery, and the reproducibility crisis. PLoS ONE **11**(2), 1–21 (2016). doi:10.1371/journal.pone.0147215
5. Spjuth, O., Eklund, M., Helgee, E.A., Boyer, S., Carlsson, L.: Integrated Decision Support for Assessing Chemical Liabilities. J. Chem. Inf. Model. **51**(8), 1840 (2011)
6. Gramatica, P.: [Online accessed january 26, 2012]. Available from: http://qsarworld.com/tempfileupload/shorthistoryofqsar.pdf (2008)
7. Christianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, 1st edn. Cambridge University Press, Cambridge, UK (2004)
8. Breimann, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
9. Mathea, M., Klingspohn, W., Baumann, K.: Chemoinformatic classification methods and their applicability domain. Mol. Inf. **35**(5), 160–180 (2016). doi:10.1002/minf.201501019

10. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer New York, Inc., Secaucus, NJ, USA (2005)
11. 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. **27**(3), 203–219 (2013). doi:10.1007/s10822-013-9639-5
12. Guha, R.: On the interpretation and interpretability of quantitative structure–activity relationship models. J. Comput. Aided Mol. Des. **22**(12), 857–871 (2008). doi:10.1007/s10822-008-9240-5
13. Carlsson, L., Ahlberg, E., Boyer, S.: Interpretation of nonlinear QSAR models applied to Ames mutagenicity data. J. Chem. Info. Model. **49**(11), 2551–2558 (2009)
14. Ahlberg, E., Spjuth, O., Hasselgren, C., Carlsson, L.: Interpretation of Conformal Prediction Classification Models, pp. 323–334. Springer International Publishing, Cham (2015)