

# Deep Learning Based Conformal Prediction of Toxicity

*Jin Zhang<sup>1</sup>, Ulf Norinder<sup>2,3,4</sup>, Fredrik Svensson<sup>5\*</sup>*

1. Department of Drug Metabolism and Pharmacokinetics, Janssen Pharmaceutica NV, B-2340

Beerse, Belgium

2. Department of Computer and Systems Sciences, Stockholm University, Box 7003, SE-164 07

Kista, Sweden

3. Department of Pharmaceutical Biosciences, Uppsala University, Box 591, SE-751 24, Uppsala

Sweden

4. MTM Research Centre, School of Science and Technology, Örebro University, SE-701 82

Örebro, Sweden

5. The Alzheimer's Research UK University College London Drug Discovery Institute, The

Cruciform Building, Gower Street, London, WC1E 6BT, UK

\* Corresponding author: [f.svensson@ucl.ac.uk](mailto:f.svensson@ucl.ac.uk)

## **Abstract**

Predictive modelling for toxicity can help reduce risks in a range of applications and potentially serve as the basis for regulatory decisions. However, the utility of these predictions can be limited if the associated uncertainty is not adequately quantified. With recent studies showing great promise for deep learning-based models also for toxicity predictions, we investigate the combination of deep learning-based predictors with the conformal prediction framework to generate highly predictive models with well-defined uncertainties. We use a range of deep feedforward neural networks and graph neural networks in a conformal prediction setting and evaluate their performance on data from the Tox21 challenge. We also compare the results from the conformal predictors to those of the underlying machine learning models. The results indicate that highly predictive models can be obtained that result in very efficient conformal predictors even at high confidence levels. Taken together, our results highlight the utility of conformal predictors as a convenient way to deliver toxicity predictions with confidence, adding both statistical guarantees on the model performance as well as better predictions of the minority class compared to the underlying models.

## Introduction

Deep learning methods have attracted increasing attention in the sphere of drug discovery and design,<sup>1,2</sup> as well as toxicity prediction.<sup>3-5</sup> While in many cases delivering state-of-the-art predictivity, it has proven a challenge to generate accurately calibrated outputs<sup>6</sup> and many implementations struggle with imbalanced data.<sup>7</sup>

Several deep learning methods have been implemented in molecular property prediction.<sup>1</sup> Two of the most popular methods are deep feedforward neural networks and graph neural networks. Deep feedforward neural network uses precalculated molecular representations such as molecular descriptors and fingerprints as the inputs. The weights and bias of the hidden units in multiple fully connected hidden layers are calculated using backpropagation and optimization. The prediction results are given by the activation functions in the output layer. Graph neural networks use molecules represented as nodes that approximate the atoms and edges that approximate the bonds. Graph neural networks are connectionist models that consider the dependence of graphs via message passing between the nodes of graphs. Several variants of graph neural networks have been developed including graph convolutional networks, graph attention networks, message passing neural networks, and r-radius subgraph graph neural networks.<sup>8</sup>

Conformal prediction is a type of confidence predictor that generates predictions with a, by the user, defined error rate.<sup>9</sup> This is achieved by the predictor by outputting prediction ranges rather than single predictions. At a certain confidence level, that fraction of all prediction ranges will include the correct label or value (for classification and regression problems respectively). One of the benefits with conformal prediction is that it can be applied on a class-by-class basis, called a Mondrian conformal predictor, guaranteeing the error rate for each class independently. This is

especially useful for imbalanced classification problems where this can reduce the bias substantially.<sup>10-12</sup>

Conformal prediction is a flexible framework that can be used with any underlying model by calibrating the outputs from the predictor using a calibration set. Any machine learning algorithm can therefore be made into a conformal predictor with only moderate changes to the pipeline.<sup>13</sup> This includes deep learning methods and a few examples have been reported in the literature.<sup>14-16</sup> However, this combination has not previously been explored for toxicity predictions.

With a vision to remove animal testing from regulatory toxicology, there is a drive to develop alternative approaches for safety assessment that also include the use of computational methods.<sup>5</sup> To be useful in such a setting or for any form of risk-assessment, it is crucial that the limitations and reliability of the computational approaches can be accurately assessed. Conformal prediction has previously been suggested to nicely fulfil these requirements with a lot less ambiguity than many other methods to assess prediction reliability.<sup>17,18</sup>

As part of the drive to develop improved methods for toxicity assessment the Tox21 project was launched to provide a wide range of data on toxicologically relevant targets.<sup>19</sup> Data from this project was the basis of the Tox21 challenge<sup>20</sup> and has been used in a range of publications on computational models for toxicity. The successful introduction of more computational methods in supporting regulatory decisions not only require quantification of the prediction confidence but also highly predictive models. Deep learning methods have shown great promise in toxicity predictions,<sup>21-23</sup> notably by achieving first place in the Tox21 Data Challenge.<sup>3</sup> As discussed above, predictions of toxicity is a field where prediction confidence is paramount,<sup>24</sup> but most deep learning methods do not give accurate quantification of the prediction uncertainty. Some methods such as Bayesian neural networks have been shown to work well on toxicity predictions.<sup>25</sup> While

these results are encouraging, they lack the flexibility associated with conformal prediction where any underlying model can be transformed into a conformal predictor. The combination of deep learning and conformal prediction is therefore an attractive choice for toxicity predictions and might help define a robust framework for reliability estimation that is useful also in regulatory contexts.<sup>18</sup> To our knowledge, this has not been demonstrated previously in the literature.

Our aim with this study is to demonstrate how conformal prediction can help construct deep learning-based predictors with associated quantitative uncertainty measures that perform well also for imbalanced data. We demonstrate various approaches to achieve this, using several different deep learning architectures and evaluating the performance of the conformal predictors against the underlying models on data from the Tox21 challenge.

## **Materials and Methods**

### *Data*

We used the training data from the Tox21 challenge (Table 1).<sup>26</sup> The dataset contains toxicities of the tested compounds on 12 nuclear receptors and stress response related targets measured in in vitro test systems. We have previously reported on the use of this data for prediction evaluation.<sup>27</sup>

The compound structures were standardized using the IMI eTOX project standardizer<sup>28</sup> in combination with tautomer standardization using the MolVS<sup>29</sup> standardizer. RDKit molecular descriptors<sup>30</sup> (consisting of 97 structural and physicochemical properties, previously described and listed in ref 12) as well as Morgan fingerprints (FP) (an extended-connectivity FP)<sup>31</sup>, calculated with radius=4 and hashed to 1024 bits, were calculated for all structures. The values of molecular descriptors were scaled to the range between 0 and 1 using the MinMaxScaler function from the scikit-learn package.

**Table 1.** Summary of the Tox21 datasets used in this study.

<b>Dataset</b>	<b>Target</b>	<b>#active</b>	<b>#inactive</b>
nr-ahr	aryl hydrocarbon receptor	942	7103
nr-ar	androgen receptor	376	8843
nr-ar-lbd	androgen receptor ligand binding domain	302	8174
nr-aromatase	aromatase	346	6759
nr-er	estrogen receptor	927	6665
nr-er-lbd	estrogen receptor ligand binding domain	441	8187
nr-ppar-gamma	peroxisome proliferator-activated receptor gamma	219	7848
sr-are	nuclear factor (erythroid-derived 2)-like 2 antioxidant responsive element	1078	6003
sr-atad5	genotoxicity indicated by ATAD5	334	8628
sr-hse	heat shock factor response element	419	7635
sr-mmp	mitochondrial membrane potential	1127	6096
sr-p53	DNA damage p53-pathway	528	7981

### *Modelling*

The datasets are divided into active and inactive and this investigation was accordingly formulated as a binary classification problem. Models were trained using ten-fold cross validation implemented using scikit-learn<sup>32</sup> StratifiedKFold splitting. In each fold, 10% of the training data was set aside for validation (used for early stopping of the deep learning-based methods training and discarded for all other methods for comparability) and 20% of the remaining training data used as a calibration set. The final remaining portion of the training data, called *proper training set* in conformal prediction, was used to train the model.

We applied several deep neural network architectures as well as the decision tree-based algorithms Random Forest<sup>33</sup> (RF) and LightGBM<sup>34</sup>. RFs were implemented using the scikit-learn function RandomForestClassifier and LightGBM using the lightgbm python package. For deep learning we used one four and one eight-layer feed forward deep neural network (DNN4 and DNN8 respectively) as well as several different implementations of graph convolutional neural networks; graph attention network (GAT), graph convolutional network (GCN), and r-radius subgraph graph neural network (GNN)<sup>35</sup>. Details of model setup and hyperparameters are listed in the Supporting Information.

The DNN4 represents the shallow but wide neural network class, whereas DNN8 represents the deep but narrow neural network class. Both classes have been applied for the various classification tasks, but the conclusions on their performance comparisons are contradictory. He et al. concluded that the depth of representations plays a central role in feature recognitions and recommended deeper but narrow networks.<sup>36</sup> However, Zagoruyko et al. reported that a wide but shallow network outperforms in accuracy and efficiency all previous deep but thin networks.<sup>37</sup> Mhaskar et al. also suggested that wide network architectures can be better suited for small datasets whereas deep networks trained on large datasets can represent functions that shallow networks cannot.<sup>38</sup> Here we implemented both classes of deep neural networks. The molecular inputs of graphical neural network methods (GAT, GCN, GNN) are the standardized chemical structures of the compounds represented in SMILES format. The label inputs are the outcomes in different toxicity endpoints of corresponding compounds. The atom feature representations of input molecules in GAT and GCN were calculated using the CanonicalAtomFeaturizer function in the dgl module. The atom type, atom degree, formal charge, hybridization, aromaticity of the atom in the input molecules

were represented in one hot encodings. GNN used similar atom characteristics as well as neighboring bond information for feature representation.

RangerLars (as implemented in Pytorch-tools, based on <https://github.com/mgrankin/over9000>) was used as the optimization algorithm for DNN4, DNN8, GAT, and GCN. RangerLARS is a synergistic optimizer that integrates RAdam<sup>39</sup>, LookAhead<sup>40</sup>, and Layer-wise Adaptive Rate Scaling (LARS)<sup>41</sup> optimizers and inherit their advantages. The RangerLARS optimizer allows large learning rate, weight decay and batch size to be used when training the models, which could result in short training time and better model performance.

The Tox21 datasets are imbalanced with regards to active and inactive compound classes. Class weights for each dataset were calculated using the ‘balanced’ setting LGBM and RF and applied to penalize the ML algorithms for misclassification of the minority class to achieve balanced prediction results. Similar Class weights were calculated and used for the loss function for the DNN, GCN and GAT models. The predicted class probabilities of the compounds were calculated using the softmax function in all deep learning-based models.

### *Conformal prediction*

A conformal predictor will make valid predictions based on a user defined significance level. The significance level is the percentage of, to the user, acceptable errors that the model may commit. A calibration set, e.g. randomly split off from the training set before training, with known labels is used to perform a recalibration of the output from the machine learning model on the test set compounds.

If the prediction outcome for a new compound, after comparison to the calibration set, is higher than the set significance level the new compound is assigned that class label. This comparison is performed for each new compound and each label (class) in the dataset. Thus, for a binary



classification problem four possible outcomes exist. A new compound can be labeled with either of the two classes, assigned both labels (both classification) or none of the labels (empty classification). For discussions related to the validity and efficiency of conformal prediction results, see Results and Discussion section.

To generate conformal predictors, we either employed the package `nonconformist` (<https://github.com/donlnz/nonconformist>), that takes scikit-learn like models and automatically handles the calculations for conformal prediction, or we employed our in-house script for models not compatible with `nonconformist`. These two implementations are designed to be as close as possible to functionally identical and the in-house script was only applied when models were not compatible with the `nonconformist` package. While it is convenient to use `nonconformist`, any model can easily be converted to a conformal predictor. For a practical walkthrough on how this is achieved please see Norinder et al.<sup>13</sup>

For the in-house script we employed the output from the respective models as nonconformity score. For `non-conformist`, all settings were left at default unless otherwise noted. To achieve Mondrian conformal prediction in `nonconformist` the following condition was used in the ICP function. Condition =  $\lambda x: x[1]$

### *Model evaluation*

Conformal prediction generates a prediction interval rather than a single label. For a binary prediction problem four different outcomes are possible: either of the two labels, both labels, or no label. Conformal predictors are proven to be valid as long as the data is exchangeable.<sup>9</sup> Validity is measured as the fraction of predictions that include the correct label, also including cases with both labels in the binary case. At a specified confidence level, the predictor is said to be valid if the fraction of predictions containing the correct label is equal to or higher than the set confidence

level. For example, at the 80% confidence level (sometimes called the 0.2 significance level, with significance being 1 - confidence level) at least 80% of the predictions should include the correct label. Additionally, since we preferably would like the more useful single label output, we calculate the efficiency of the models. Efficiency is defined as the fraction of predictions with a single label (irrespective of it being correct), a model with a higher fraction of single label predictions is more efficient. The desired output from a conformal predictor should be valid with an error rate corresponding to the confidence level (higher validity models are referred to as over conservative) and with as high efficiency as possible. For a more in-depth explanation of conformal prediction including examples we refer the reader to Norinder et al.<sup>13</sup>

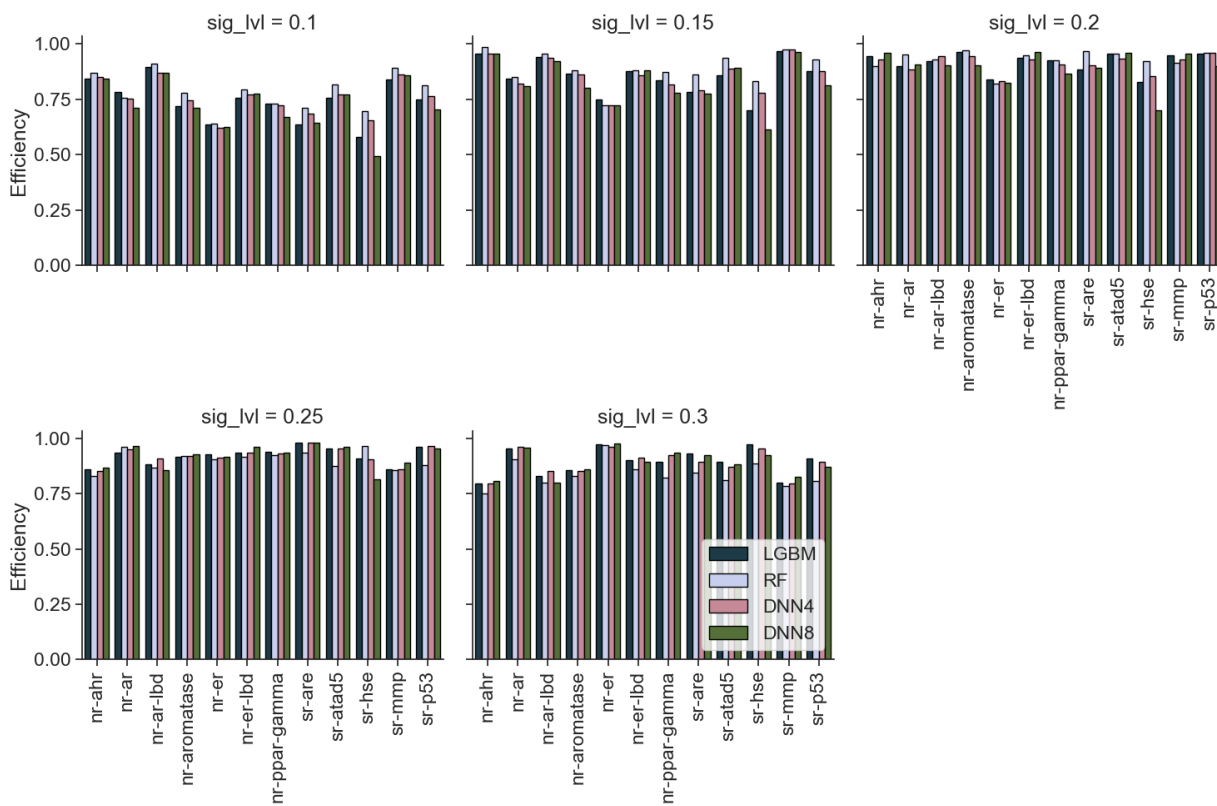
We also evaluated the predictivity of the underlying models using balanced accuracy (BA), sensitivity, specificity, ROC, F1 score, Kappa, Precision, and Matthews correlation coefficient (MCC). These metrics were calculated using the corresponding scikit-learn functions. The calculated metrics for all underlying models were given in the Supporting Information. For comparison we also calculated some of these metrics for both initial predicted class probabilities and the conformal output, where the metrics were calculated only for the single label predictions using the confidence level that generated the highest balanced efficiency (mean value for efficiency of the minority and majority class, respectively).

All model evaluation was calculated using the aggregated results from the cross validation.

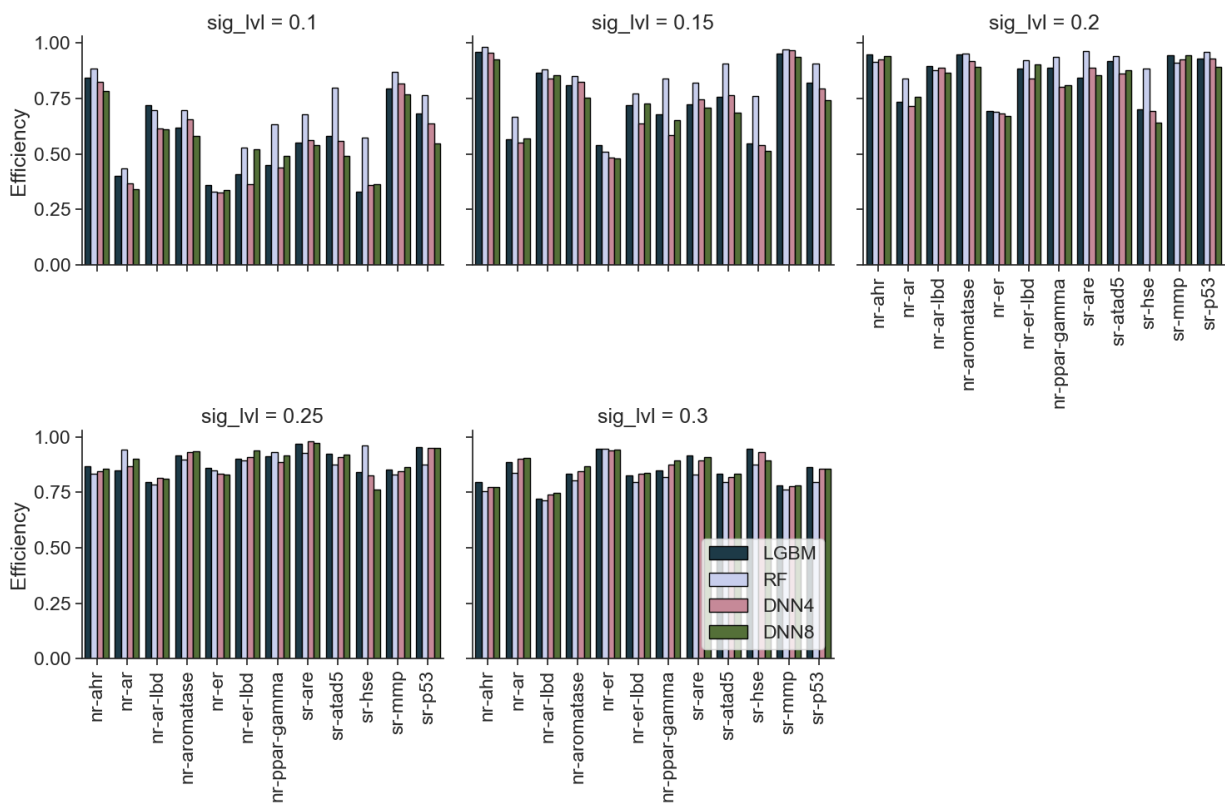
## **Results and Discussion**

For conformal predictors to be useful, the desired output is a high number of single label predictions while also being valid. We therefore evaluate the results primarily using the standard conformal prediction metrics efficiency and validity, for more details please see the Materials and

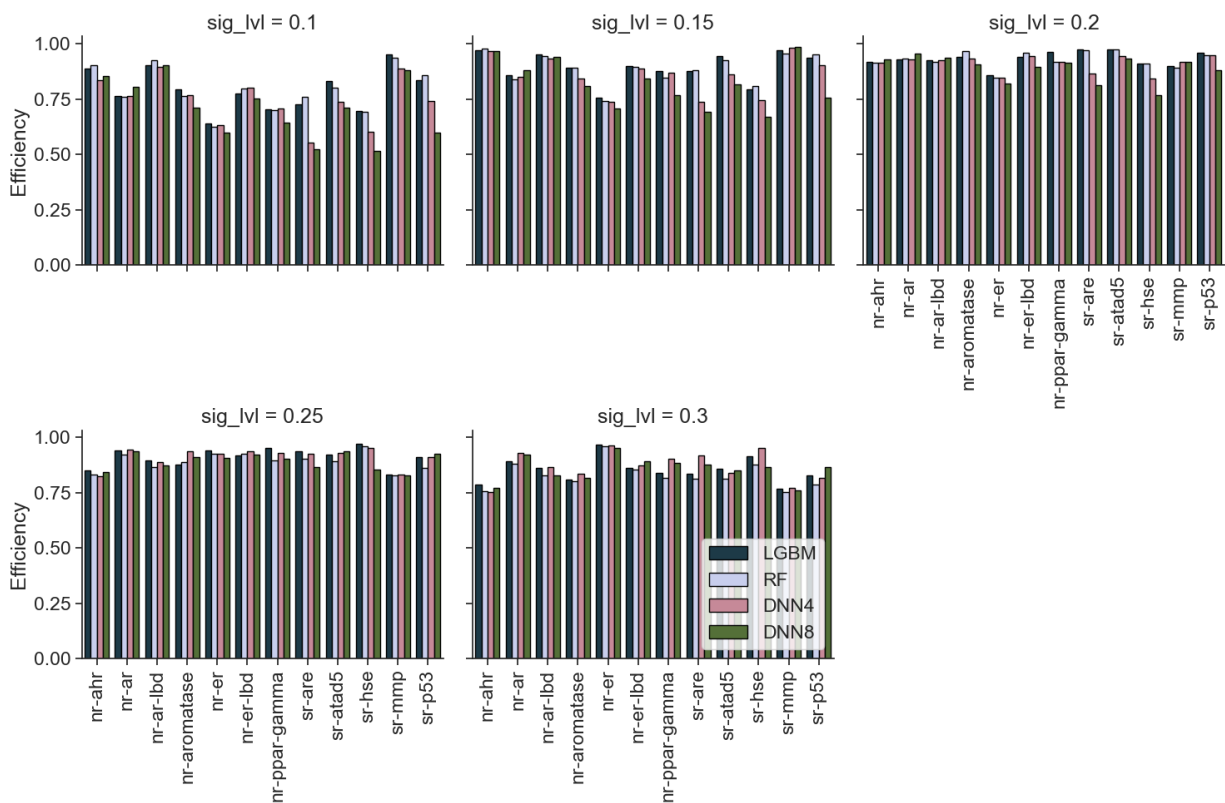
Methods section. All datasets were modelled using ten-fold cross validation using the selected machine learning methods for this study. All methods generated valid (an error rate corresponding to the set confidence level) and efficient conformal predictors (see the Supporting Information for tabled data). Figure 1-6 show the efficiency of the cross-validation test set predictions for the different models based on fingerprints, RDKit descriptors, and graph convolutions for active and inactive compounds.



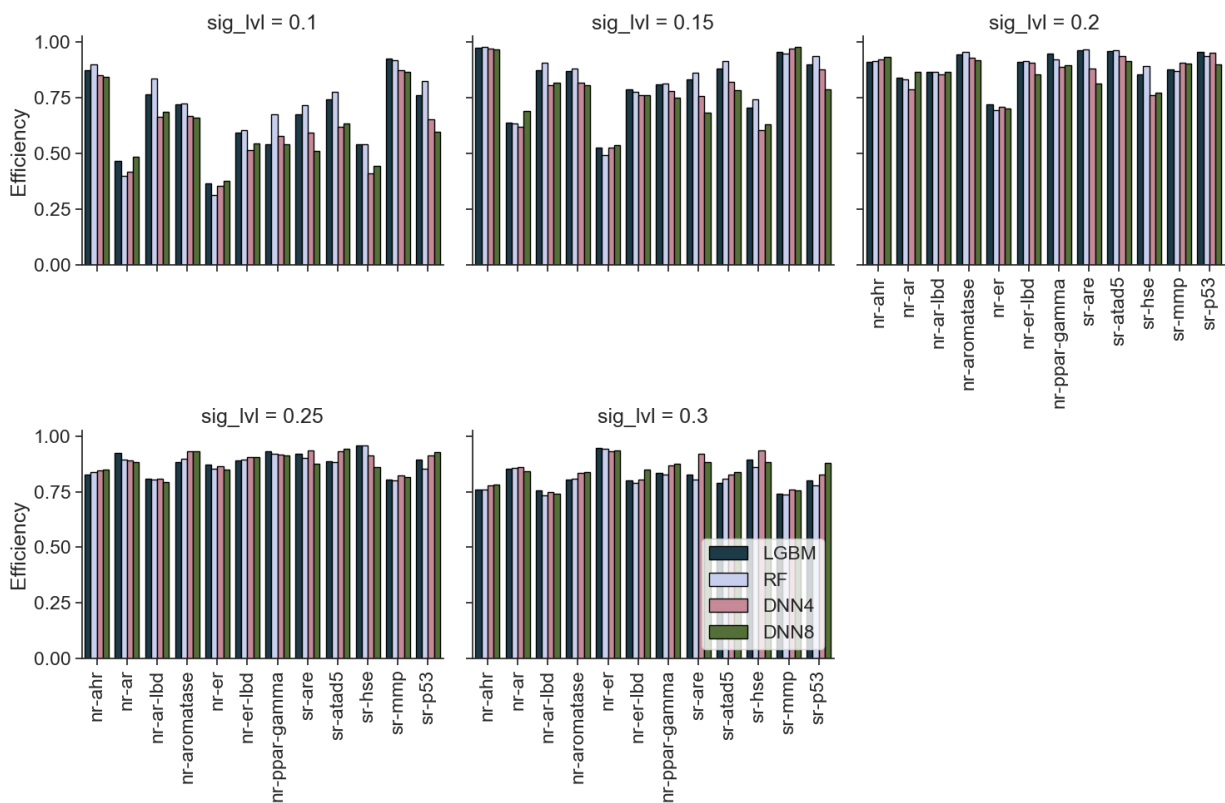
**Figure 1.** Efficiency for the fingerprint-based models for the active class across all datasets at different significance levels (sig\_lvl).



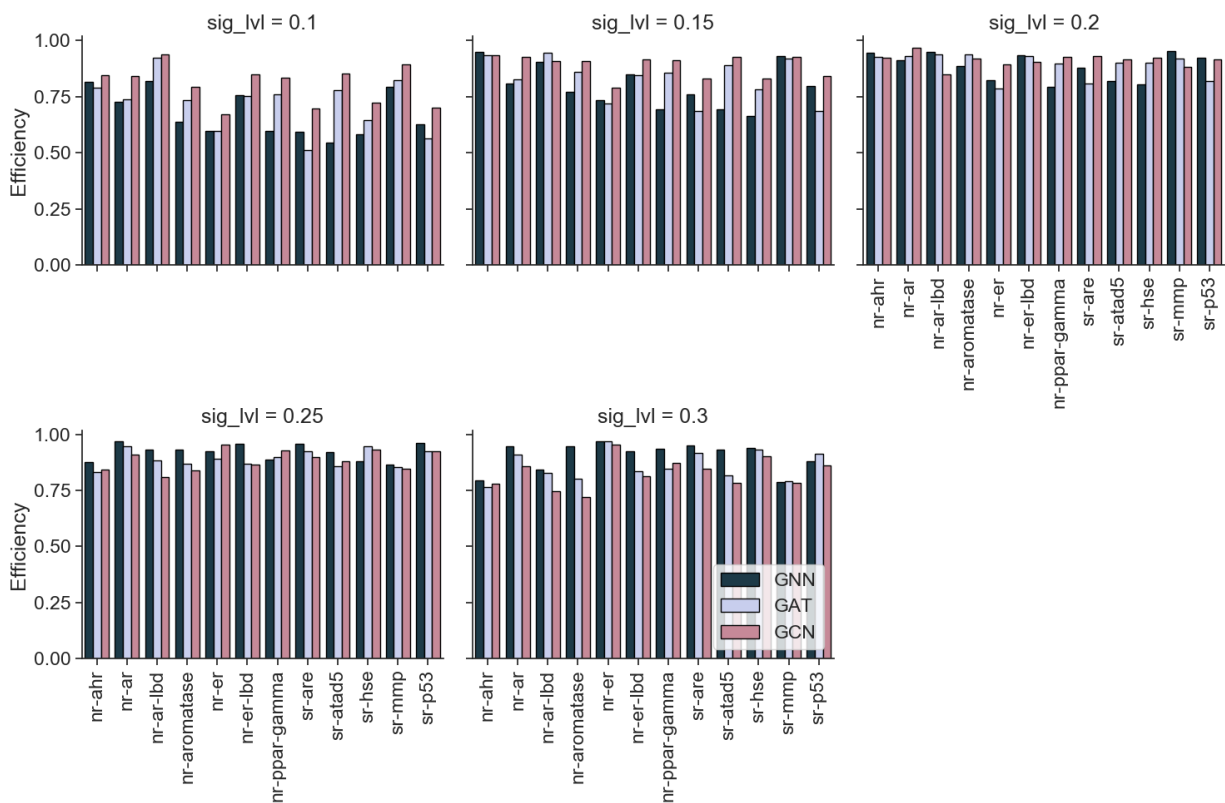
**Figure 2.** Efficiency for the fingerprint-based models for the inactive class across all datasets at different significance levels (sig\_lvl).



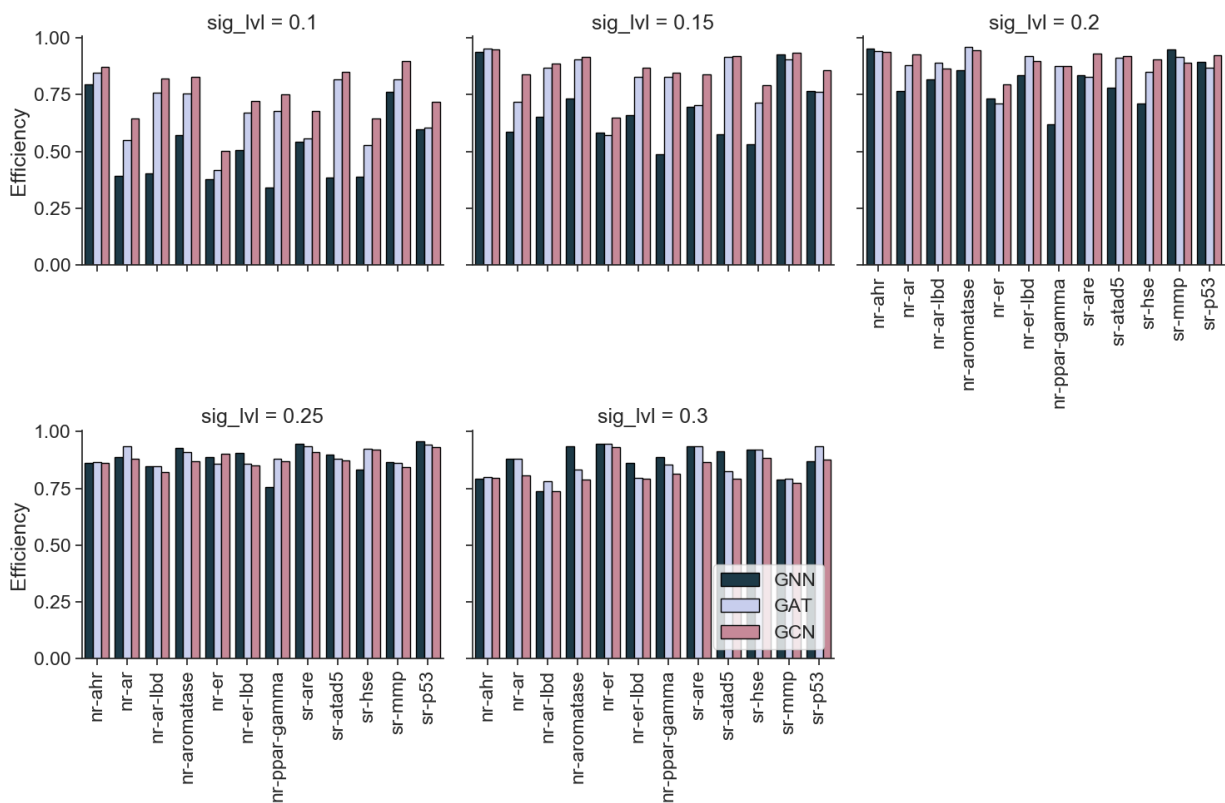
**Figure 3.** Efficiency for the RDKit descriptor-based models for the active class across all datasets at different significance levels (sig\_lvl).



**Figure 4.** Efficiency for the RDKit descriptor-based models for the inactive class across all datasets at different significance levels (sig\_lvl).



**Figure 5.** Efficiency for graph convolution models for the active class across all datasets at different significance levels (sig\_lvl).



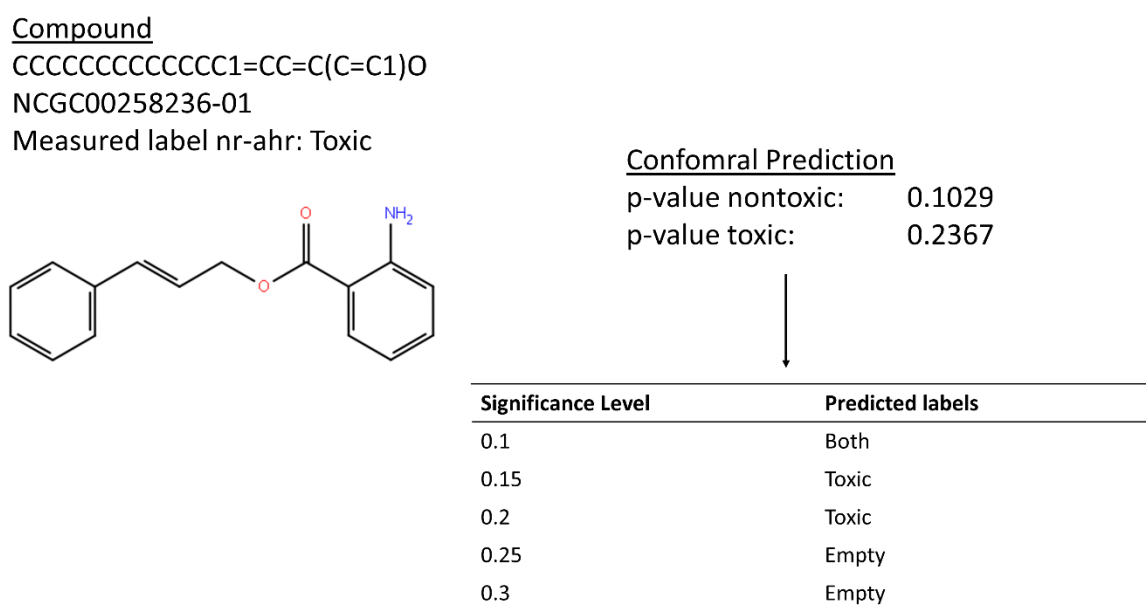
**Figure 6.** Efficiency for graph convolution models for the inactive class across all datasets at different significance levels (sig\_lvl).

Most of the results follow a typical pattern where the average efficiency across datasets peaks at about 75%-80% confidence (0.25-0.2 significance level) while higher confidence levels result in more dual predictions and lower confidence levels in more empty predictions. Not all datasets are equally well modelled with notably lower efficiencies observed for the nr-er dataset. Similar trends have been observed previously.<sup>15</sup> These results demonstrate that conformal predictors can reach levels of efficiency that makes them highly usable in predictive toxicology tasks.

To demonstrate how to interpret the output from a conformal predictor, the output is exemplified in Figure 7 with the results from the fingerprint based nr-ahr DNN8 model (additional examples available in the Supporting Information). A conformal p-value is obtained for each class and in



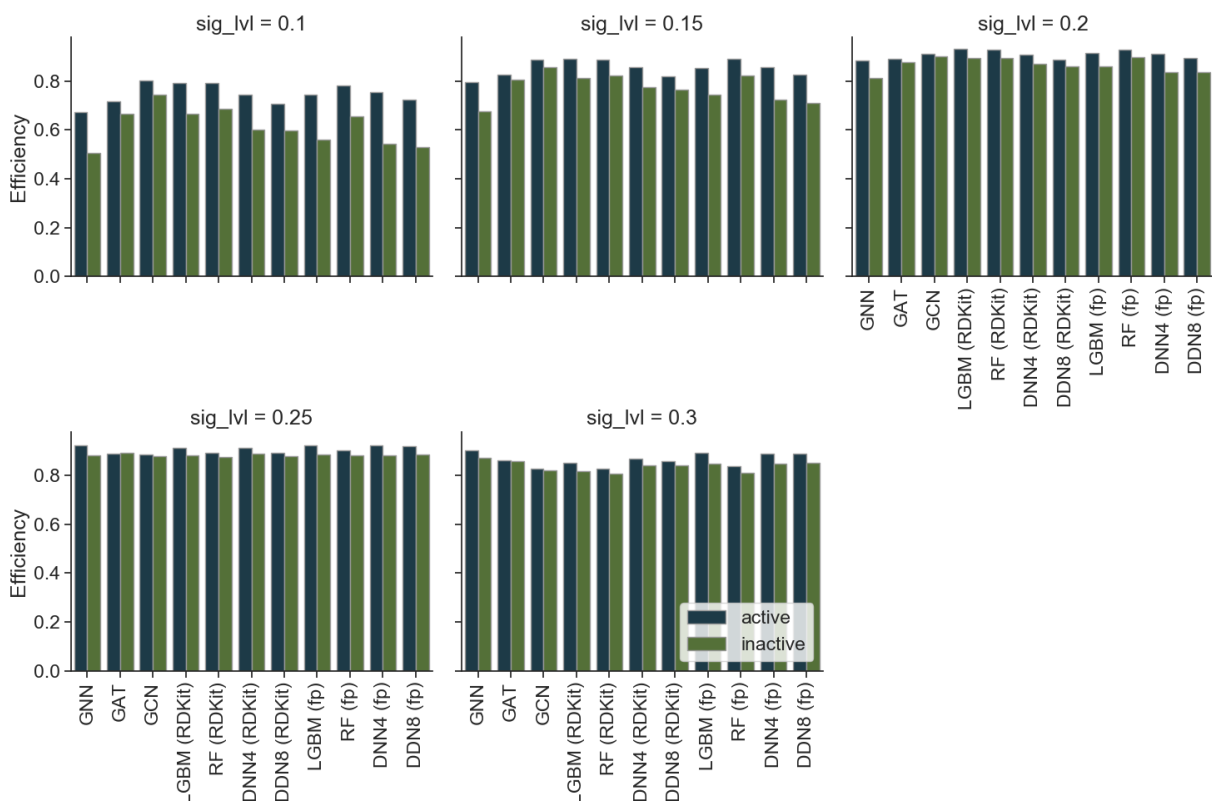
conjunction with the desired significance level these are used to derive the prediction labels (a p-value > the significance level assigns the label). For the example compound the p-values are relatively low for both classes but clearly higher for the toxic class indicating a toxic compound, as also indicated by the label assignment. If a very high confidence is required, no label can be excluded and the compound gets predicted to belong to both classes, but from 85% to 75% confidence the compound received a single label as toxic. The output allows for balanced evaluation of individual compounds and has the potential to enhance chemical risk assessment.



**Figure 7.** Example of the output from the conformal prediction for a selected compound from the nr-ahr dataset. The underlying model is the fingerprint based DNN8. A p-value is generated for each class and compared to the desired confidence level (1-significance level) to derive the labels.

The average efficiencies of the models across all datasets are shown in Figure 8. From a practical point of view, the performance of these models is sufficient to make them highly usable, generating about 90% single label predictions with an overall confidence of 80% or above. Notably is also

that this performance holds for the minority class (toxic compounds) despite the for some datasets strong class imbalance (Table 1). This is a highly desirable feature for toxicity prediction where we are generally mostly interested in the minority class.



**Figure 8.** Average efficiency for the different algorithms across all datasets at different significance levels (sig\_lvl).

Having established that the conformal predictors showed good performance, we also wanted to evaluate the performance of the underlying models and relate this to the output from the conformal predictors. This is a key aspect since the conformal predictors can be constructed from any underlying model. If conformal prediction consistently brings advantages over the underlying models this would thus be highly advantageous to understand. Figure 9 shows the average BA and MCC for the different models derived based on initial predicted class probabilities and conformal

prediction results (full list of metrics available in the Supporting Information). In this study, we implemented two major types of deep learning algorithms i.e. deep feed forward neural networks and graphical neural networks. One of the graph-based models (GCN) outperformed the other models in terms of balanced accuracy and MCC. However, we would like to stress that this might depend on the hyperparameters used (Supporting Information). We also observed that the deep feed forward neural networks (DNN4 and DNN8) gave similar but slightly worse predictivity than LightGBM, which is consistent with what we previously reported.<sup>27</sup>

The Tox21 datasets have been widely used to assess methods for predicting compound toxicity.<sup>42,43</sup> To help put the performance of the models in perspective, we also provide the results for the GCN and GAT models on the Tox21 test sets in the Supporting Information (Figure S4 and S5) along with tabulated performance for previously reported models.<sup>3,44-46</sup> Our best models are comparable to some of the top leaderboard submissions, although not the best performing. Our best model had an average AUC of 0.734 compared to a range of reported AUCs of 0.576-0.874. However, the key aspect we are investigating in this study is the utility of conformal-prediction for toxicity predictions. The comparison of the different properties and performance of the conformal predictors and the respective underlying model is likely to transfer also to other models. Thus, if other models have superior predictive power for the investigated endpoint, those models could be made into conformal predictors and the results would be expected to follow the same trends as reported here.

High efficiency from the conformal predictor is linked to a strong predictive power of the underlying model. Very good performance is observed for the graph-based deep learning model GCN. This method also had a very high efficiency at the 90% confidence level (Figure 5 and 6). Conversely, methods with poorer performance tend to have their peak efficiency at lower

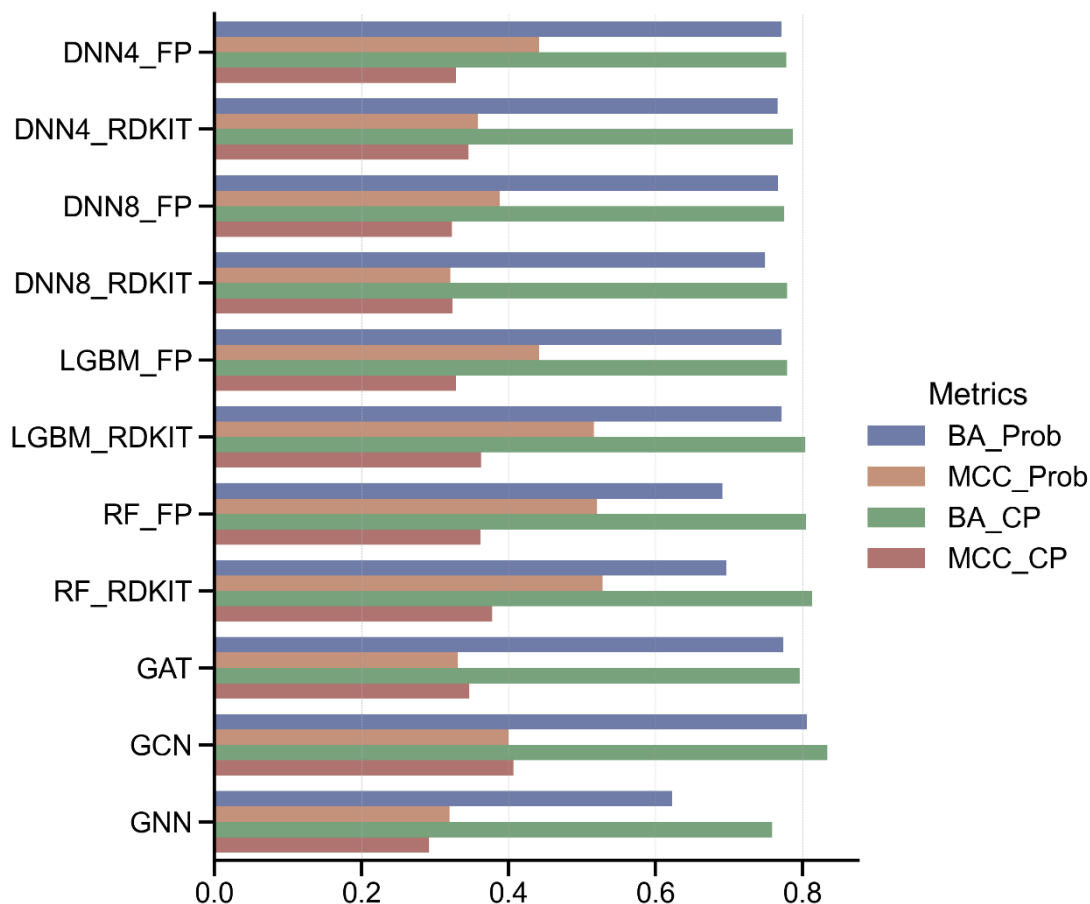
confidence levels. Overall, this shows that the best conformal predictors (highly efficient at high confidence levels) are being derived from the best performing underlying model.

From Figure 9 two main cases emerge with respect to the performance of the underlying model when compared with the conformal prediction framework:

1. High performing underlying models such as GCN, do not gain in predictive performance. This is not that surprising as it is very difficult to achieve better predictions. However, the addition of the conformal prediction framework ensures a stringent handling of the applicability domain (AD) of the model, clearly indicating for which compounds a single label prediction cannot be made.

2. Most of the other models benefit from the CP framework by retrieving more minority class compounds, BA increases, albeit often at the expense of more false positives, MCC decreases. Furthermore, for imbalance data sets the application of the conformal prediction framework resolves the problem of how to adjust the cut-off level for label classification by using a stringent and mathematically proven procedure for reclassification.<sup>10</sup> At the same time, conformal prediction allows the user to set an acceptable confidence level for the predictions.

Overall, conformal prediction tends to retrieve more examples of the minority class at the expense of a slightly higher false positive rate. In a toxicity prediction setting, we believe this is a favorable balance as it reduces the number of potentially toxic instances missed by the model.



**Figure 9.** Selected average performance metrics on the test set for the underlying models and corresponding conformal predictions models at maximum balanced efficiency.

While the best performance was observed for a graph-based deep learning-based model in this study, this was not true for all inputs and datasets. This illustrates that the choice of algorithm is not always straightforward and that in addition to predictivity, other aspects such as training time and interpretability should also be considered. It is therefore important to select the underlying model judiciously. With conformal prediction being agnostic to the type of underlying model, we encourage users to also look beyond the models applied in this study, especially for tasks where other models have been shown to have excellent performance.

Overall, our study highlights two areas where conformal predictors add value compared to the underlying model. Firstly, all our conformal predictors generated valid models, thus delivering the error rate expected at the selected confidence level. This allows the user to tune the predictions to a confidence that fits the task at hand. Secondly, we demonstrated that the conformal predictors generally retrieved more examples of the toxic (minority) class, thus reducing the risk of false negatives, albeit at the cost of more false positives.

The form of conformal prediction used in this study, with one calibration set split at random from the testing data for each cross-validation loop, can in some settings be impacted by what compounds end up in the calibration set. To counter this, and make more efficient use of the available data, an alternative is to train multiple conformal predictors with different calibration set splits and aggregate the results. This can be done either by bootstrapping the different calibration sets or in a cross validation like fashion.<sup>47,48</sup> For deep learning methods the problem can be that the sometimes very long training times make this approach practically difficult. One solution that has been proposed to this is to generate snapshot learners.<sup>14</sup>

The results from this study in our opinion further strengthens the position of conformal prediction for toxicity predictions and regulatory applications. Since the conformal prediction framework associated each instance with a well-defined probability to belong to each class, the result is readily interpretable and more confident predictions are easy to separate from less confident ones. This is especially attractive as any underlying model can be used for conformal prediction. In our opinion, this makes these models highly suitable for computational toxicology.

## **Conclusions**

In this study we have demonstrated that deep learning based conformal predictors can generate highly predictive models with associated confidence for endpoints from the Tox21 challenge. Compared to the underlying models, conformal prediction adds both a controllable error rate as well as better recall of the toxic compounds. In our opinion, the combination of high predictive performance and confidence make deep learning based conformal predictors highly desirable for many tasks in predictive toxicology.

We were able to achieve an efficiency greater than 80% for the toxic class at the 90% confidence level using a graph convolutional network as the underlying model, making it an attractive model for these tasks. However, the benefits from conformal prediction can be leveraged using any underlying machine learning model.

## **Data and Software Availability**

Python code used for the calculations in this study is available on [https://github.com/FredrikSvenssonUK/tox21\\_conformal](https://github.com/FredrikSvenssonUK/tox21_conformal)

All datasets used in this study are available from the Tox21 challenge web portal <https://tripod.nih.gov/tox21/challenge/data.jsp> (Accessed April 21 2018)

## **Supporting Information:**

Tabulated performance data for all models, hyperparameters and details of packages and versions used for modelling.

## **Acknowledgements**

The Alzheimer's Research UK University College London Drug Discovery Institute is core funded by Alzheimer's Research UK (520909).

FS gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan V GPU used for this research.

## Notes

The authors declare no competing financial interest.

## References

- (1) Jing, Y.; Bian, Y.; Hu, Z.; Wang, L.; Xie, X.-Q. S. Deep Learning for Drug Design: An Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. *AAPS J.* **2018**, *20*, 58.
- (2) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250.
- (3) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Frontiers in Environmental Science*. 2016, p 80.
- (4) Tang, W.; Chen, J.; Wang, Z.; Xie, H.; Hong, H. Deep Learning for Predicting Toxicity of Chemicals: A Mini Review. *J. Environ. Sci. Heal. Part C* **2018**, *36*, 252–271.
- (5) Tetko, I. V; Tropsha, A. Joint Virtual Special Issue on Computational Toxicology. *J. Chem. Inf. Model.* **2020**, *60*, 1069–1071.
- (6) Scalia, G.; Grambow, C. A.; Pernici, B.; Li, Y.-P.; Green, W. H. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 2697–2717.



- (7) Johnson, J. M.; Khoshgoftaar, T. M. Survey on Deep Learning with Class Imbalance. *J. Big Data* **2019**, *6*, 27.
- (8) Jie, Z.; Gangu, C.; Zhengyan, Z.; Cheng, Y.; Zhiyuan, L.; Lifeng, W.; Changcheng, L.; Sun, M. Graph Neural Networks: A Review of Methods and Applications. *arXiv* **2018**, 1812.08434.
- (9) Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: New York, 2005; pp 1–324.
- (10) Löfström, T.; Boström, H.; Linusson, H.; Johansson, U. Bias Reduction through Conditional Conformal Prediction. *Intell. Data Anal.* **2015**, *19*, 1355–1375.
- (11) Norinder, U.; Boyer, S. Binary Classification of Imbalanced Datasets Using Conformal Prediction. *J. Mol. Graph. Model.* **2017**, *72*, 256–265.
- (12) Svensson, F.; Norinder, U.; Bender, A. Modelling Compound Cytotoxicity Using Conformal Prediction and PubChem HTS Data. *Toxicol. Res. (Camb)*. **2017**, *6*, 73–80.
- (13) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **2014**, *54*, 1596–1603.
- (14) Cortés-Ciriano, I.; Bender, A. Deep Confidence: A Computationally Efficient Framework for Calculating Reliable Prediction Errors for Deep Neural Networks. *J. Chem. Inf. Model.* **2018**, *59*, 1269–1281.
- (15) Paisios, A.; Lenc, L.; Martinek, J.; Král, P.; Papadopoulos, H. A Deep Neural Network Conformal Predictor for Multi-Label Text Classification. In *Proceedings of the Eighth*

- Symposium on Conformal and Probabilistic Prediction and Applications*; Gammerman, A., Vovk, V., Luo, Z., Smirnov, E., Eds.; Proceedings of Machine Learning Research; PMLR: Golden Sands, Bulgaria, 2019; Vol. 105, pp 228–245.
- (16) Cortés-Ciriano, I.; Bender, A. Reliable Prediction Errors for Deep Neural Networks Using Test-Time Dropout. *J. Chem. Inf. Model.* **2019**, *59*, 3330–3339.
- (17) Svensson, F.; Norinder, U. Conformal Prediction for Ecotoxicology and Implications for Regulatory Decision-Making. In *Methods in Pharmacology and Toxicology*; 2020.
- (18) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling for Regulatory Purposes. A Transparent and Flexible Alternative to Applicability Domain Determination. *Regul. Toxicol. Pharmacol.* **2015**, *71*, 279–284.
- (19) Thomas, R. The US Federal Tox21 Program: A Strategic and Operational Plan for Continued Leadership. *ALTEX* **2018**.
- (20) Huang, R.; Xia, M.; Nguyen, D.-T.; Zhao, T.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Rossoshek, A.; Simeonov, A. Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs. *Frontiers in Environmental Science*. 2016, p 85.
- (21) Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L. Deep Learning for Drug-Induced Liver Injury. *J. Chem. Inf. Model.* **2015**, *55*, 2085–2093.
- (22) Sosnin, S.; Karlov, D.; Tetko, I. V.; Fedorov, M. V. A Comparative Study of Multitask Toxicity Modeling on a Broad Chemical Space. *J. Chem. Inf. Model.* **2018**.
- (23) Kawaguchi, M.; Nukaga, T.; Sekine, S.; Takemura, A.; Susukida, T.; Oeda, S.; Kodama,

- A.; Hirota, M.; Kouzuki, H.; Ito, K. Mechanism-Based Integrated Assay Systems for the Prediction of Drug-Induced Liver Injury. *Toxicol. Appl. Pharmacol.* **2020**, *394*, 114958.
- (24) Wittwehr, C.; Aladjov, H.; Ankley, G.; Byrne, H. J.; de Knecht, J.; Heinzle, E.; Klambauer, G.; Landesmann, B.; Luijten, M.; MacKay, C.; Maxwell, G.; Meek, M. E. (Bette); Pains, A.; Perkins, E.; Sobanski, T.; Villeneuve, D.; Waters, K. M.; Whelan, M. How Adverse Outcome Pathways Can Aid the Development and Use of Computational Prediction Models for Regulatory Toxicology. *Toxicol. Sci.* **2017**, *155*, 326–336.
- (25) Semenova, E.; Williams, D. P.; Afzal, A. M.; Lazic, S. E. A Bayesian Neural Network for Toxicity Prediction. *bioRxiv* **2020**, 2020.04.28.065532.
- (26) Huang, R.; Sakamuru, S.; Martin, M. T.; Reif, D. M.; Judson, R. S.; Houck, K. A.; Casey, W.; Hsieh, J.-H.; Shockley, K. R.; Ceger, P.; Fostel, J.; Witt, K. L.; Tong, W.; Rotroff, D. M.; Zhao, T.; Shinn, P.; Simeonov, A.; Dix, D. J.; Austin, C. P.; Kavlock, R. J.; Tice, R. R.; Xia, M. Profiling of the Tox21 10K Compound Library for Agonists and Antagonists of the Estrogen Receptor Alpha Signaling Pathway. *Sci. Rep.* **2014**, *4*, 5664.
- (27) Zhang, J.; Mucs, D.; Norinder, U.; Svensson, F. LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity—Application to the Tox21 and Mutagenicity Data Sets. *J. Chem. Inf. Model.* **2019**, *59*, 4150–4158.
- (28) IMI ETOX Project Standardizer. *version 0.1.7*. <https://pypi.python.org/pypi/standardiser>.
- (29) MolVS Standardizer. *version 0.0.9*. <https://pypi.python.org/pypi/MolVS>.
- (30) RDKit: Open-Source Cheminformatics (<http://www.rdkit.org>).
- (31) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*,

742–754.

- (32) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (33) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
- (34) Ke, G.; Meng, Q.; Wang, T.; Chen, W.; Ma, W.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst.* **30** **2017**.
- (35) Tsubaki, M.; Tomii, K.; Sese, J. Compound–Protein Interaction Prediction with End-to-End Learning of Neural Networks for Graphs and Sequences. *Bioinformatics* **2018**, *35*, 309–318.
- (36) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *CoRR* **2015**, *abs/1512.03385*.
- (37) Zagoruyko, S.; Komodakis, N. Wide Residual Networks. *CoRR* **2016**, *abs/1605.07146*.
- (38) Mhaskar, H.; Liao, Q.; Poggio, T. When and Why Are Deep Networks Better than Shallow Ones? In *31st AAAI Conference on Artificial Intelligence, AAAI 2017*; 2017.
- (39) Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; Han, J. On the Variance of the Adaptive Learning Rate and Beyond. *arXiv e-prints*. August 1, 2019, p arXiv:1908.03265.
- (40) Zhang, M. R.; Lucas, J.; Hinton, G.; Ba, J. Lookahead Optimizer: K Steps Forward, 1 Step Back. *arXiv e-prints*. July 1, 2019, p arXiv:1907.08610.

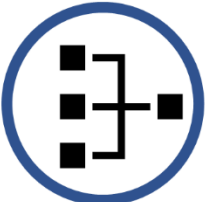
- (41) You, Y.; Gitman, I.; Ginsburg, B. Large Batch Training of Convolutional Networks. *arXiv e-prints*. August 1, 2017, p arXiv:1708.03888.
- (42) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.
- (43) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530.
- (44) Banerjee, P.; Eckert, A. O.; Schrey, A. K.; Preissner, R. ProTox-II: A Webserver for the Prediction of Toxicity of Chemicals. *Nucleic Acids Res.* **2018**, *46*, W257–W263.
- (45) Jiang, J.; Wang, R.; Wei, G.-W. GGL-Tox: Geometric Graph Learning for Toxicity Prediction. *J. Chem. Inf. Model.* **2021**.
- (46) Karim, A.; Mishra, A.; Newton, M. A. H.; Sattar, A. Efficient Toxicity Prediction via Simple Features Using Shallow Neural Networks and Decision Trees. *ACS Omega* **2019**, *4*, 1874–1888.
- (47) Carlsson, L.; Eklund, M.; Norinder, U. Aggregated Conformal Prediction. In *Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings*; Iliadis, L., Maglogiannis, I., Papadopoulos, H., Sioutas, S., Makris, C., Eds.; Springer International Publishing: Berlin, Heidelberg, 2014; pp 231–240.

(48) Vovk, V. Cross-Conformal Predictors. *Ann. Math. Artif. Intell.* **2015**, 74, 9–28.

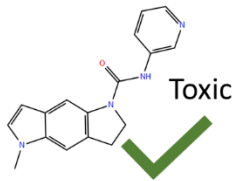
For Table of Contents Only

Deep Learning

- Good Performance
- Statistical Guarantees



With Confidence



Toxic