Using Predicted Bioactivity Profiles to Improve Predictive Modelling

Ulf Norinder[1,2,3], Ola Spjuth[2,4], Fredrik Svensson[5]*

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

2. Department of Pharmaceutical Biosciences, Uppsala University, Box 591, SE-75124, Uppsala Sweden

3. MTM Research Centre, School of Science and Technology, Örebro University, SE-70182 Örebro, Sweden

4. Science for Life Laboratory, Uppsala University, Box 591, SE-75124, Uppsala 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

Abstract

Predictive modelling is a cornerstone in early drug development. Using information for multiple domains or across prediction tasks has the potential to improve the performance of predictive modelling. However, aggregating data often leads to incomplete data matrices that might be limiting for modelling. In line with previous studies, we show that by generating predicted bioactivity profiles, and using these as additional features, prediction accuracy of biological endpoints can be improved. Using conformal prediction, a type of confidence predictor, we present a robust framework for the calculation of these profiles and the evaluation of their impact. We report on the outcomes from several approaches to generate the predicted profiles on 16 datasets in cytotoxicity and bioactivity and show that efficiency is improved the most when including the p-values from conformal prediction as bioactivity profiles.

Introduction

Machine learning has established itself as a key technology in the drug discovery process.[1,2] Predictive modelling can be used to improve compounds properties and ADME-PK,[3] improve on target activity, identify off-targets, and reduce the risk of toxicity.

Whereas standard QSAR modelling typically relies on chemical features, biological features such as HTS fingerprints[4] and affinity fingerpints[5] have been shown to improve predictions when combined with chemical features.[6,7] Studies have also demonstrated that data from similar targets in other species can improve the predictions[8,9] and even entirely predicted bioactivity features have been used.[10–13] Essentially, these methods use information on the compounds bioactivity profile to help infer a new unknown activity.[14]

Other developments in the field include new machine learning algorithms. Deep neural networks have recently attracted increasing attention also for applications in drug discovery.[15] One of the attractive

features of these methods is the ability to transfer learning, where information from one endpoint can be used to improve the prediction of another.[16] However, this is not prominent in all settings.[17]

Together, these developments have shown a lot of promise for methods that combine and learn from multiple data sources. However, a common limitation integrating data from multiple domains is that the resulting data matrix is often sparse.[18] The usage of imputed bioactivity data can circumvent this limitation and has been shown to be able to greatly improve the predictive performance of bioactivity models.[14,19–21] There are currently multiple machine learning based methods for predicting bioactivity profiles.[22–24]

Useful models do not only require accurate predictions, but also some measure of confidence in how likely the prediction is to be correct, ideally on an instance by instance basis.[25,26] Conformal predicton,[27] is a type of confidence prediction, generating predictions with a fixed error rate determined by the user set confidence level. As such, conformal prediction forms a well-defined framework for assigning predicted class probabilities. Conformal prediction is also a good choice for bioactivity modelling as it handles data imbalance very well.[28,29]

We propose that data from multiple endpoints can be integrated using predicted bioactivity profiles generated from conformal predictors. Multiple single endpoints are modelled and these models are used to predict the activity of all compounds in the training matrix. Using these predicted bioactivities as additional features a new model is trained to predict the outcome of interest. We report on different approaches to generate the predicted bioactivity features and show that this approach can improve the predictions compared to standard approaches.

Methods

Datasets

We used the 16 datasets from PubChem Bioassay[30,31] previously reported by Svensson, Norinder, and Bender[32] measuring compound cytotoxicity. 11 of these were used in the development of the methods

and five as validation sets. These were complemented by ten bioactivity datasets,[18,33] also originating from PubChem. These datasets have also previously been used in predictive modelling. All the datasets used in this study are detailed in Table 1. Compounds were labelled as negative (inactive) or positive (active).

Table 1. The datasets used in this study.

Cytotoxicity data sets	**Number of compounds**	
	Negative Class	**Positive Class**
AID 2275	29745	193
AID 847	40958	194
AID 903	52445	338
AID 624418	385836	524
AID 719	83904	937
AID 648	85197	924
AID 602141	357738	1302
AID 1486	215443	2408
AID 1825	288346	2259
AID 588856	400999	3018
AID 2717	296777	3181
AID 430 (Validation)	61506	1121
AID 620 (Validation)	55759	706
AID 598 (Validation)	32946	33
AID 50464 (Validation)	80022	5139
AID 463 (Validation)	86336	364
Bioactivity data sets		
AID 687014	52572	4320
AID 463190	52443	4449

AID 588726	51858	5034
AID 652054	51857	5035
AID 485346	51461	5431
AID 2796	51322	5570
AID 504652	50420	6472
AID 743279	47459	9433
AID 1814	40780	16112
AID 2314	30586	26306

Compounds were represented using 97 different physiochemical RDKit[34] descriptors (cytotoxicity data sets[32]) or by the fingerprints used by de la Vega de Leon et al.[18] (bioactivity data sets); these features are referred to as Structural Features. See the respective studies for details on compound preparation.

The cytotoxicity data sets were divided into two groups consisting of 11 and 5 data sets (Table 1). The 5 Validation data sets were used to evaluate the performance when adding a fixed array of bioactivity profiles to new datasets.

The experimental assay matrix completeness, i.e. percentage of compounds tested in all assays, was 21.6 % and 100 % for the cytotoxicity and bioactivity data sets, respectively. The experimental values of the assays were never used in the added bioactivity profiles.

Modelling

Prior to modelling, a 20 % random test set was split from the data for model evaluation. For the bioactivity data sets the training and test set were setup to be identical for each endpoint since they all contain the same compounds. The result of the split with respect to overlap of compounds between the test set of the investigated end point and the corresponding training sets for the added bioactivity profiles were, on average, 4.6 % for the cytotoxicity datasets while all training and test sets for the bioactivity datasets were completely disjoint, i.e. the overlap was 0%. QSAR models were trained for all

dataset using the Structural Features. Separate models were trained for all endpoints using Random Forests (RF)[35] consisting of 100 trees. In addition to standard RF models, models were also trained using Aggregated Mondrian Conformal Prediction (CP)[36] with 20 aggregated models per outcome. These models were based on underlying RF models with 100 trees, we refer to these models as CP-RF. CP is a well-calibrated framework for class probability assignments where predictions are a set of p-values; one p-value (related to probabilities) for each class.[27,37] In this case, since we are looking at binary predictions (active or inactive) we obtain two p-values per predicted endpoint.

Both standard RF and CP predictions were used to generate predictions for all outcomes for each compound in all datasets. These values constitute additional features (Bioactivity Features). A new model (Augmented Model) was constructed for each endpoint based on Structural Features with the addition of Bioactivity Features (referred to as Bioactivity Profile) for all other endpoints than the one being considered; see Figure 1. We added the bioactivity features either as the predicted label (0 for inactive and 1 for active, referred to as Binary Bioactivity Profile, for conformal models the whichever label had the largest p-value was added), or using the number of trees from the RF (RF probability) or the conformal p-value for the active class (Continuous Bioactivity Profile). Predicted bioactivity profiles were only added from other data sets from the same source (Cytotoxicity or Bioactivity, Table 1).

Scikit-learn[38] (version 0.20.4) was used for building the underlying random forest models and nonconformist (version 1.2.5) for building the Mondrian conformal predictors. All parameters were kept at default unless explicitly stated: A calibration set of 30 % was randomly selected from the training set and the remaining part used for building the model. In order to perform class-wise calibration (Mondrian conformal prediction) the python code was augmented with the following statement:

```
icp = IcpClassifier(nc, condition=lambda x: x[1])
```

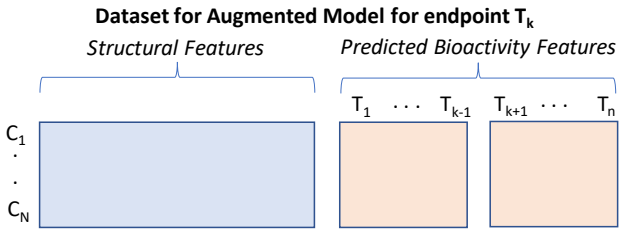

Figure 1. Datasets for Augmented Models are constructed for all compounds C_i from both Structural (chemical) Features and predicted Bioactivity Features, where the latter are the predicted values using a model trained on only Structural (chemical) Features for each endpoint T_k.

Similarity analyses were performed whereby the similarity to the nearest neighbour in the training set was calculated for all corresponding test as well as calibration set compounds, respectively, and averaged over each set for each endpoint (see Supporting Information). All features were scaled to range 0-1 for the features to have the same influence on the distance calculation. This only affected the cytotoxicity datasets since the features were physiochemical RDKit[34] descriptors of various scales while the bioactivity datasets used fingerprints (0 or 1) as features. The Tanimoto similarity for the cytotoxicity datasets were based on the corresponding Tanimoto similarity formula for continuous variables while the traditional formula was used for the bioactivity datasets.[39]

Model evaluation

Conformal predictors are proven to always be valid (the error rate matches that defined by the operator) if the data is exchangeable.[27] This is achieved by outputting a set of labels for each prediction. For a binary prediction, this set can be either empty, contain either of the two labels on their own, or both labels. While the validity defined as the fraction of sets containing the correct label is guaranteed by the method, the models will have varying efficiencies (number of single label predictions) depending on how well the target is modelled. At a given confidence level, a more predictive model should afford better efficiency, i.e. a higher fraction of single label predictions. We therefore choose to evaluate model performance by looking at the efficiency (defined as the fraction of single label predictions). More

traditional measures for imbalanced datasets such as balanced accuracy (BA) and Matthews correlation coefficient (MCC) have also been included in the Supporting Information for comparison. These measures were all calculated using only single label predictions.

For a more in-depth explanation of the concepts behind the practical application of conformal prediction we refer the reader to Norinder et al.[37]

Results and Discussion

The first approach we investigated for adding Bioactivity Features was to add binary predicted labels, active or inactive. The results from these experiments are shown in Figures 2-5 and Table 2 and 3, displaying the change in model efficacy with and without the added binary bioactivity profiles. A clear improvement, with higher efficiencies for the higher confidence levels, could be observed across all the datasets when adding the additional labels. This was also mirrored by better model performance when evaluated using MCC and BA (Supporting Information). These results hold true for both the negative (inactive) and the positive (active) class. Importantly, the models remain valid for all the experiments (Table 2 and 3).

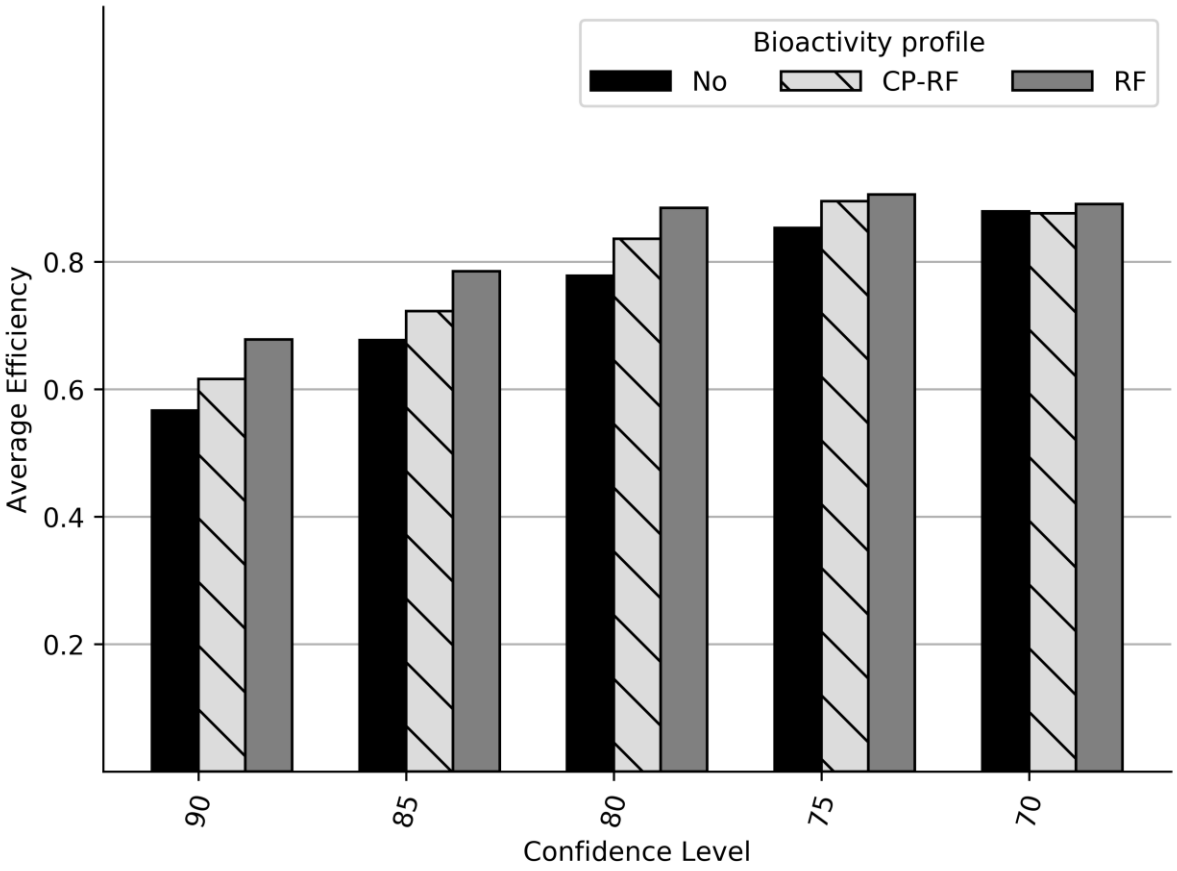

Figure 2. Average efficiency of the positive class for the cytotoxicity datasets at different confidence levels with and without the binary bioactivity profile features. An increased efficiency is observed when adding the bioactivity profiles.

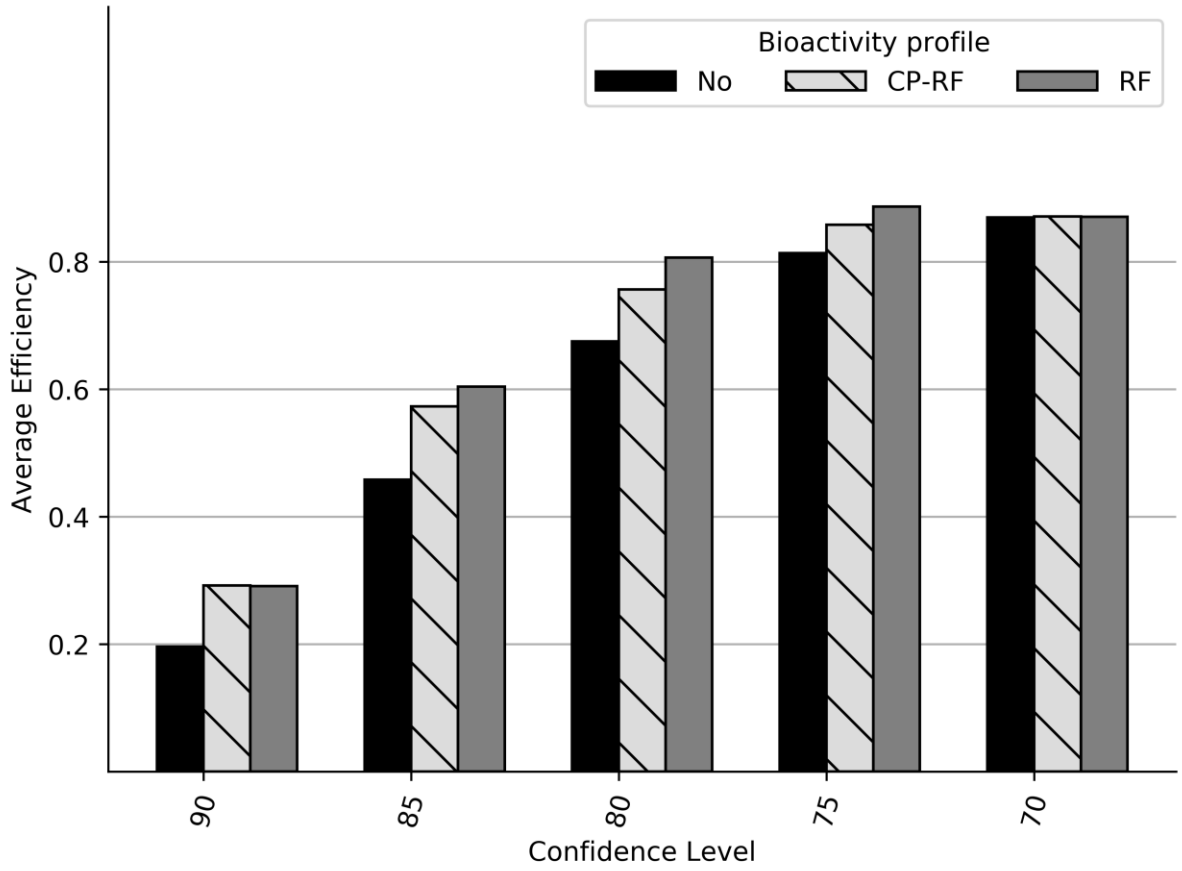

Figure 3. Average efficiency of the negative class for the cytotoxicity datasets at different confidence levels with and without the binary bioactivity profile features. An increased efficiency is observed when adding the bioactivity profiles.

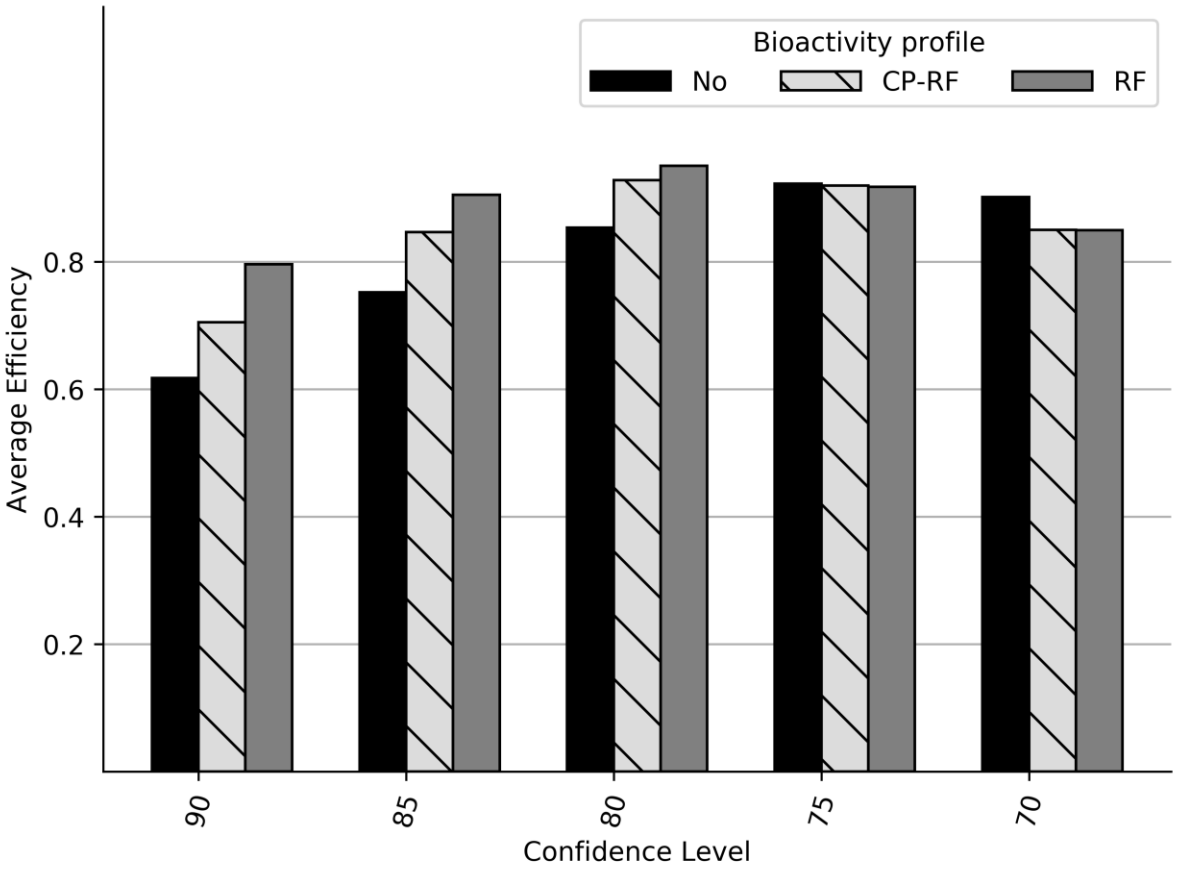

Figure 4. Average efficiency of the positive class for the bioactivity datasets at different confidence levels with and without the binary bioactivity profile features. An increased efficiency is observed when adding the bioactivity profiles.

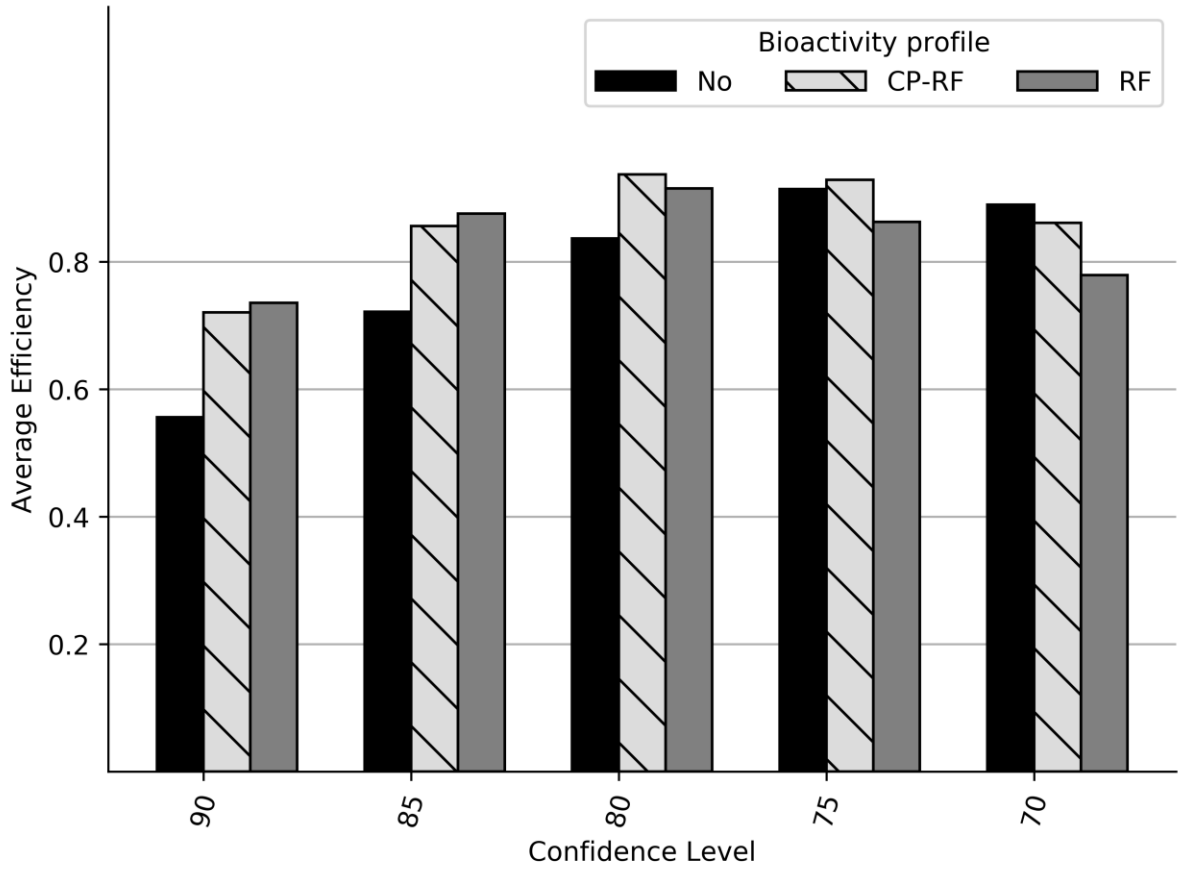

Figure 5. Average efficiency of the negative class for the bioactivity datasets at different confidence levels with and without the binary bioactivity profile features. An increased efficiency is observed when adding the bioactivity profiles.

Although adding the predicted binary class labels as bioactivity profiles improved the predictive performance of the models, the method introduces a boundary problem as a cut-off needs to be defined to assign the labels. This may be especially challenging when using imbalanced data where the classifiers, most likely, will be expected to prefer the majority class. Alternatively, the predicted class label can be forced by assigning it with the largest conformal prediction p-value of the two classes. This however, completely overrides setting significance levels in conformal prediction, making it a less desirable approach. By instead adding the raw output (number of trees for the random forest or p-values for the classes from the conformal models) these issues are circumvented.

To investigate this, the datasets were modelled with and without the addition of the predicted continuous bioactivity profiles and the results evaluated. Addition of the predicted continuous bioactivity profiles increase the efficiencies of the final models (Figure 6-9), both when adding the conformal p-values for each label or when using the number of trees predicting the outcome from the random forest model. Compared to using the predicted labels the continuous profiles generated slightly more efficient models (Table 2 and 3).

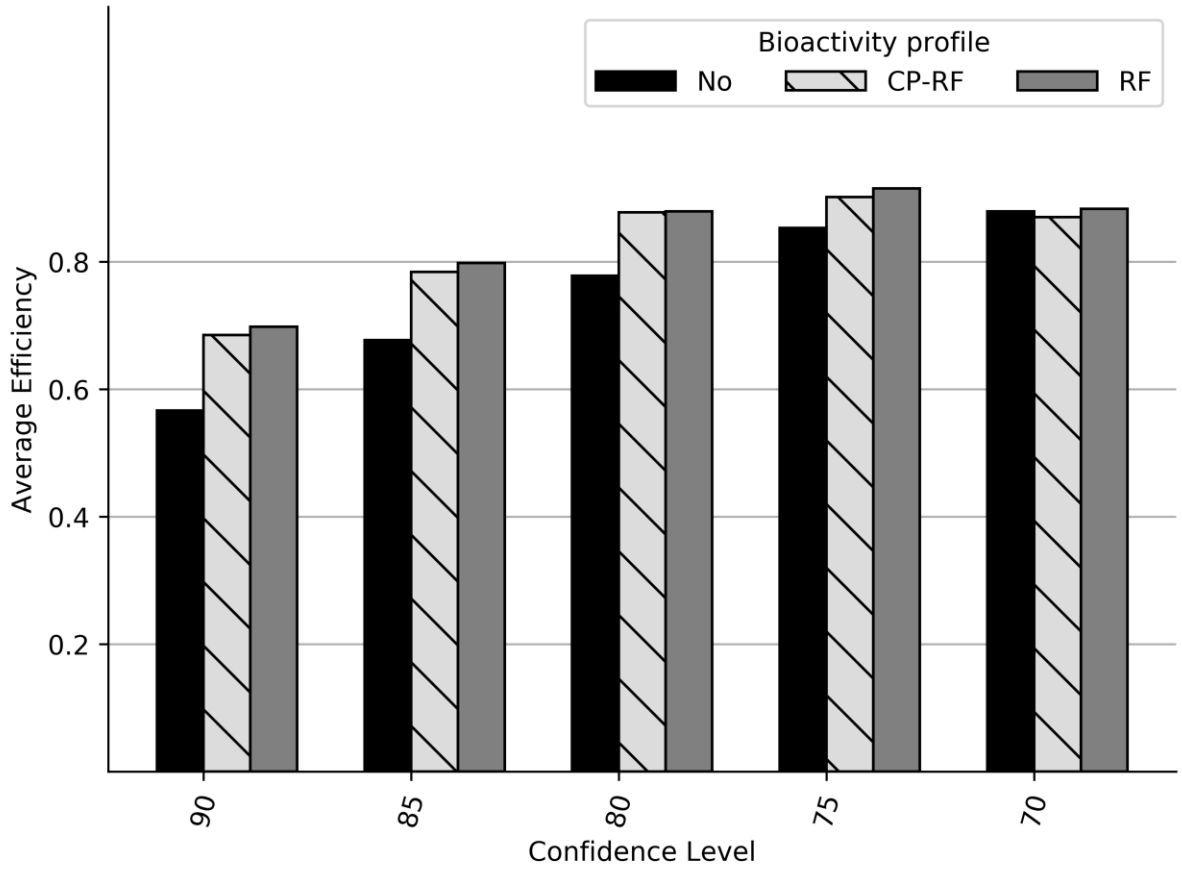

Figure 6. Average efficiency of the positive class for the cytotoxicity datasets at different confidence levels with and without the continuous bioactivity profile features. An increased efficiency is observed when adding the bioactivity profiles.

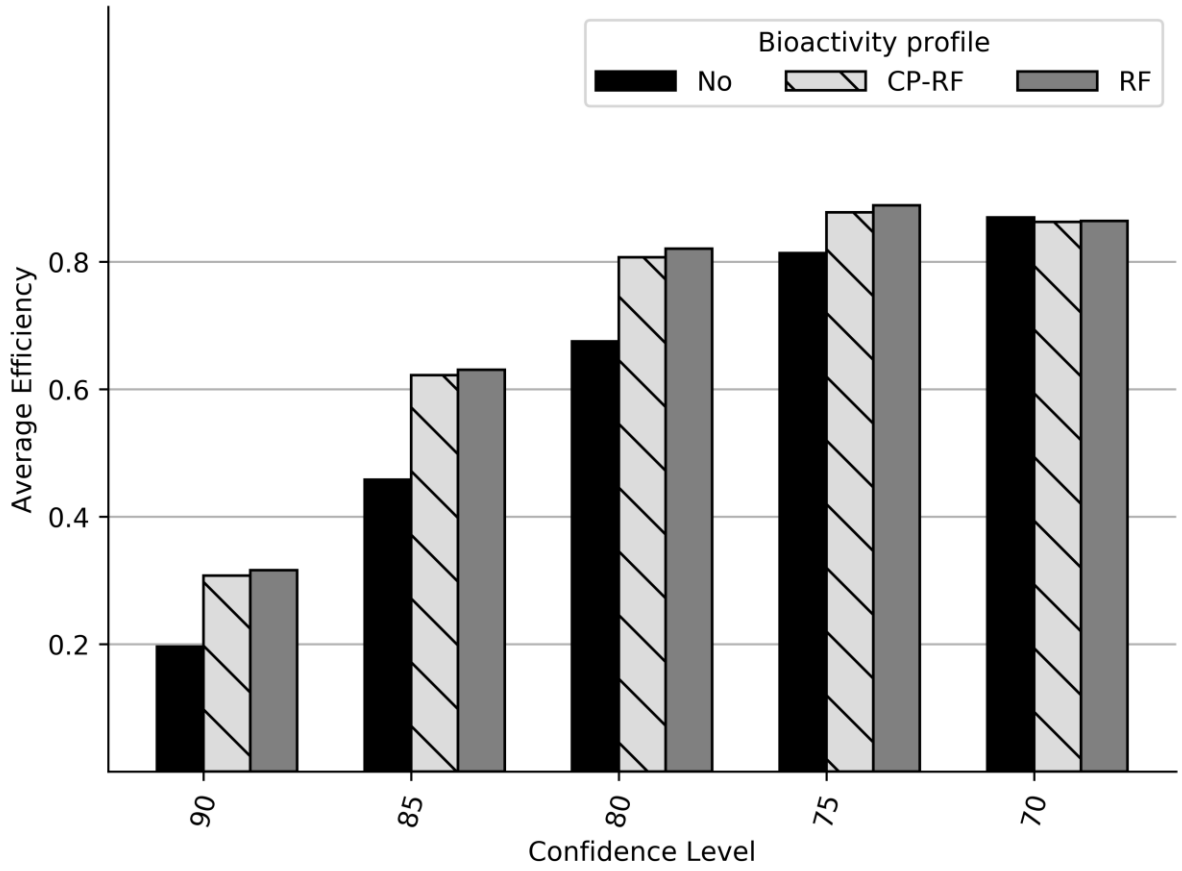

Figure 7. Average efficiency of the negative class for the cytotoxicity datasets at different confidence levels with and without the continuous bioactivity profile features. An increased efficiency is observed when adding the bioactivity profiles.

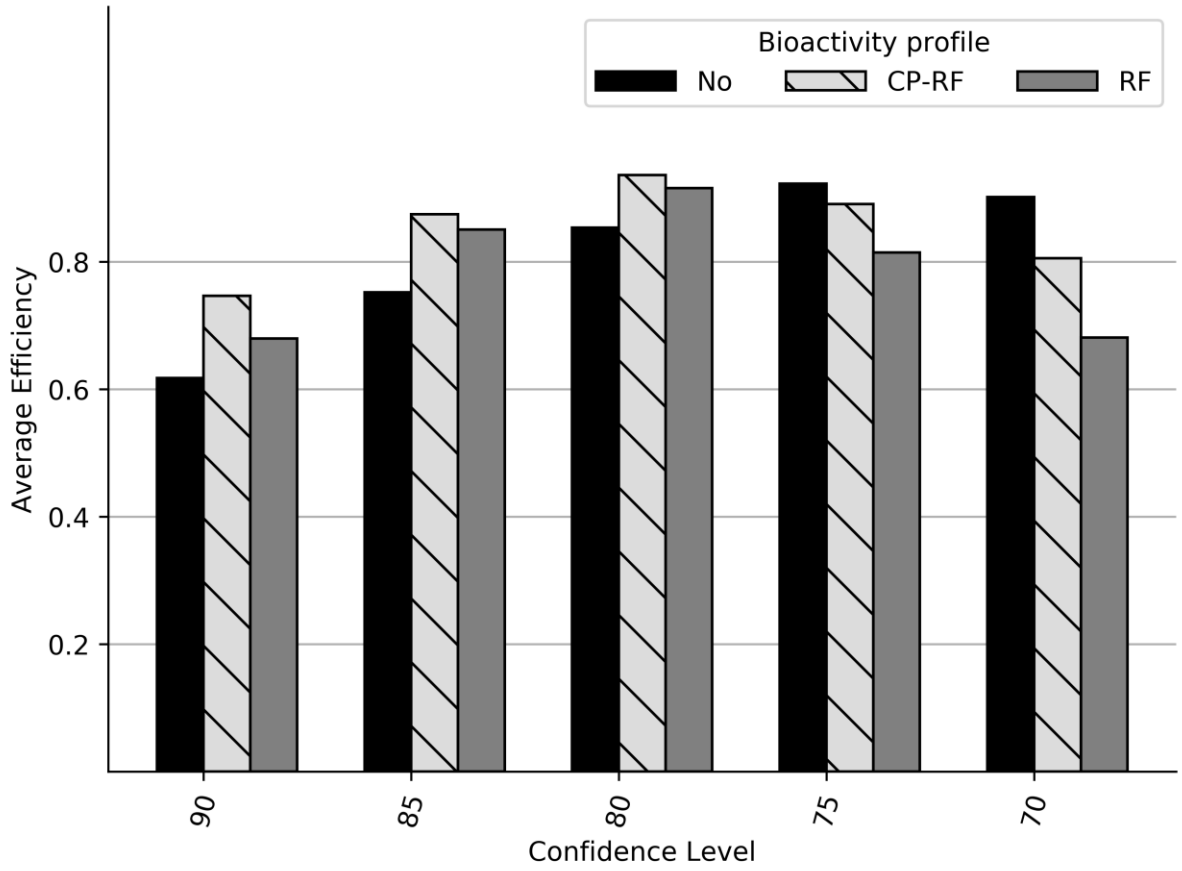

Figure 8. Average efficiency of the positive class for the bioactivity datasets at different confidence levels with and without the continuous bioactivity profile features. An increased efficiency is observed when adding the bioactivity profiles.

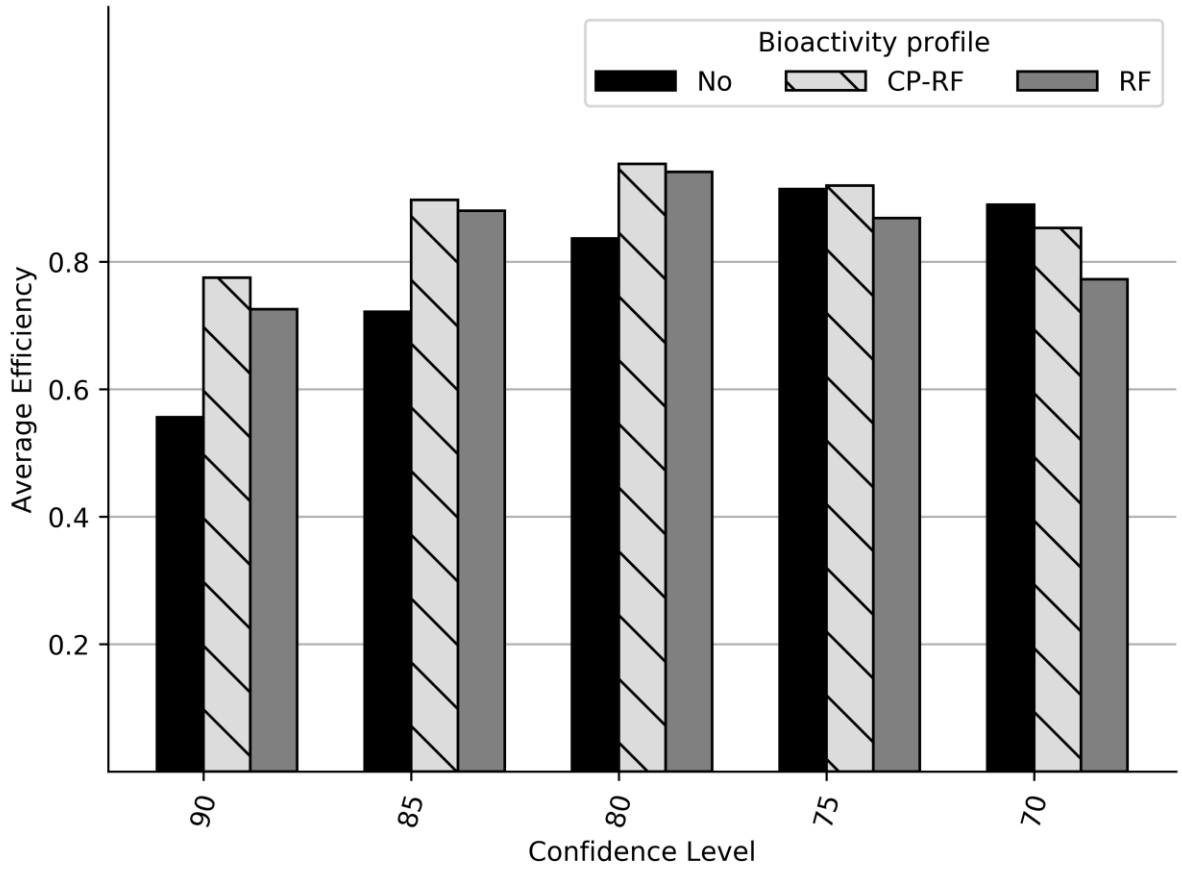

Figure 9. Average efficiency of the negative class for the bioactivity datasets at different confidence levels with and without the continuous bioactivity profile features. An increased efficiency is observed when adding the bioactivity profiles.

Interestingly, for the cytotoxicity dataset the two different continuous profiles (RF number of trees or conformal p-values) generated very similar results but with a slight preference for the RF derived profiles. In contrast, for the bioactivity datasets the conformal derived continuous profiles had a clear edge in performance.

Maximum efficiency for the datasets was observed for a confidence level of about 75-80%, in line with what has been observed previously for this type of data.[32,33]

To check that the increased performance observed when using the bioactivity profiles was not due to the additional number of features contributing spurious correlations, we also included 20 columns of

random variables as features for the cytotoxicity data sets. The random features did not improve model performance (see Supporting Information for details).

Table 2. Average validity and efficiency of the different methods for the bioactivity data sets

Confidence Level (%)	Bioactivity Descriptors	Mean Validity Positive Class	Mean Validity Negative Class	Mean Efficiency Positive Class	Mean Efficiency Negative Class
90	None	0.92	0.91	0.62	0.56
85	None	0.87	0.87	0.75	0.72
80	None	0.83	0.82	0.85	0.84
75	None	0.77	0.77	0.92	0.91
70	None	0.72	0.72	0.90	0.89
90	Conformal Class Label	0.85	0.90	0.70	0.72
85	Conformal Class Label	0.79	0.85	0.85	0.86
80	Conformal Class Label	0.73	0.81	0.93	0.94
75	Conformal Class Label	0.67	0.76	0.92	0.93
70	Conformal Class Label	0.62	0.72	0.85	0.86
90	Conformal p-value	0.90	0.92	0.75	0.77
85	Conformal p-value	0.83	0.88	0.87	0.90
80	Conformal p-value	0.77	0.84	0.94	0.95
75	Conformal p-value	0.71	0.80	0.89	0.92

70	Conformal p-value	0.66	0.76	0.81	0.85
90	RF Class Label	0.89	0.79	0.80	0.74
85	RF Class Label	0.83	0.72	0.91	0.88
80	RF Class Label	0.78	0.66	0.95	0.92
75	RF Class Label	0.74	0.61	0.92	0.86
70	RF Class Label	0.70	0.57	0.85	0.78
90	RF Number of Trees	0.82	0.91	0.68	0.73
85	RF Number of Trees	0.72	0.86	0.85	0.88
80	RF Number of Trees	0.63	0.80	0.92	0.94
75	RF Number of Trees	0.55	0.74	0.81	0.87
70	RF Number of Trees	0.47	0.69	0.68	0.77

Table 3. Average validity and efficiency of the different methods for the cytotoxicity datasets.

Confidence Level (%)	Bioactivity Descriptors	Mean Validity Positive Class	Mean Validity Negative Class	Mean Efficiency Positive Class	Mean Efficiency Negative Class
90	None	0.98	0.94	0.57	0.20
85	None	0.94	0.90	0.68	0.46
80	None	0.89	0.87	0.78	0.67
75	None	0.84	0.83	0.85	0.81
70	None	0.76	0.79	0.88	0.87
90	Conformal Class Label	0.97	0.93	0.62	0.29
85	Conformal	0.92	0.89	0.72	0.57

	Class Label				
80	Conformal Class Label	0.86	0.85	0.84	0.76
75	Conformal Class Label	0.81	0.82	0.90	0.86
70	Conformal Class Label	0.75	0.78	0.88	0.87
90	Conformal p-value	0.97	0.94	0.68	0.31
85	Conformal p-value	0.92	0.91	0.78	0.62
80	Conformal p-value	0.85	0.88	0.88	0.81
75	Conformal p-value	0.78	0.85	0.90	0.88
70	Conformal p-value	0.72	0.81	0.87	0.86
90	RF Class Label	0.97	0.94	0.68	0.29
85	RF Class Label	0.91	0.91	0.79	0.60
80	RF Class Label	0.84	0.87	0.88	0.81
75	RF Class Label	0.78	0.84	0.91	0.89
70	RF Class Label	0.72	0.80	0.89	0.87
90	RF Number of Trees	0.97	0.94	0.70	0.32
85	RF Number of Trees	0.91	0.90	0.80	0.63
80	RF Number of Trees	0.84	0.87	0.88	0.82
75	RF Number of Trees	0.78	0.84	0.92	0.89
70	RF Number of Trees	0.71	0.80	0.88	0.86

Although generating the models for the continuous bioactivity profiles is relatively straight forward for the size of data sets used in this study, we also wanted to show their utility when applied as a fixed set to new datasets without adding additional models. This allows for more rapid training of models for new targets.

We applied the 11 cytotoxicity models from the first part to generate features for the remaining five validation sets (Table 1). The results from the resulting models are shown in Figure 10 and 11. The results closely mimics those seen for the previous datasets.

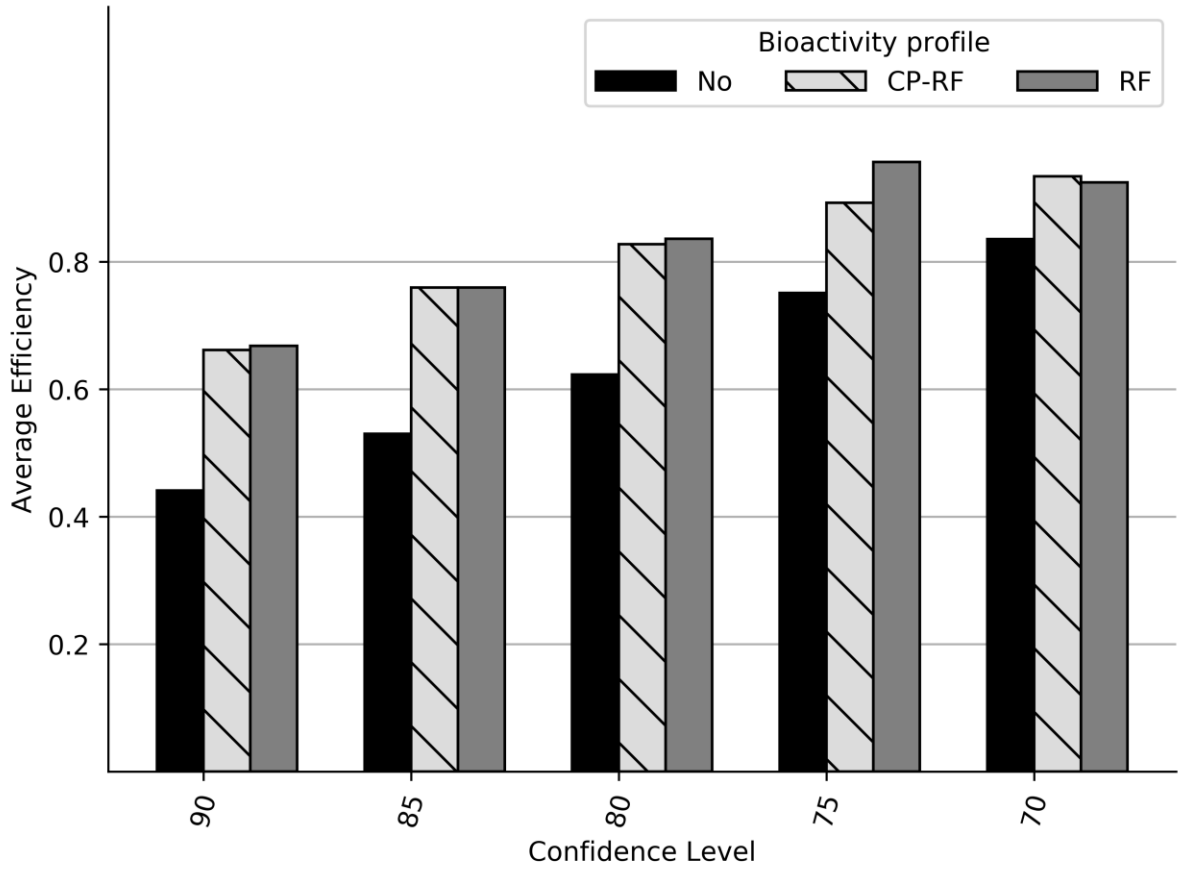

Figure 10. Efficiency for the positive class of the five external cytotoxicity data sets.

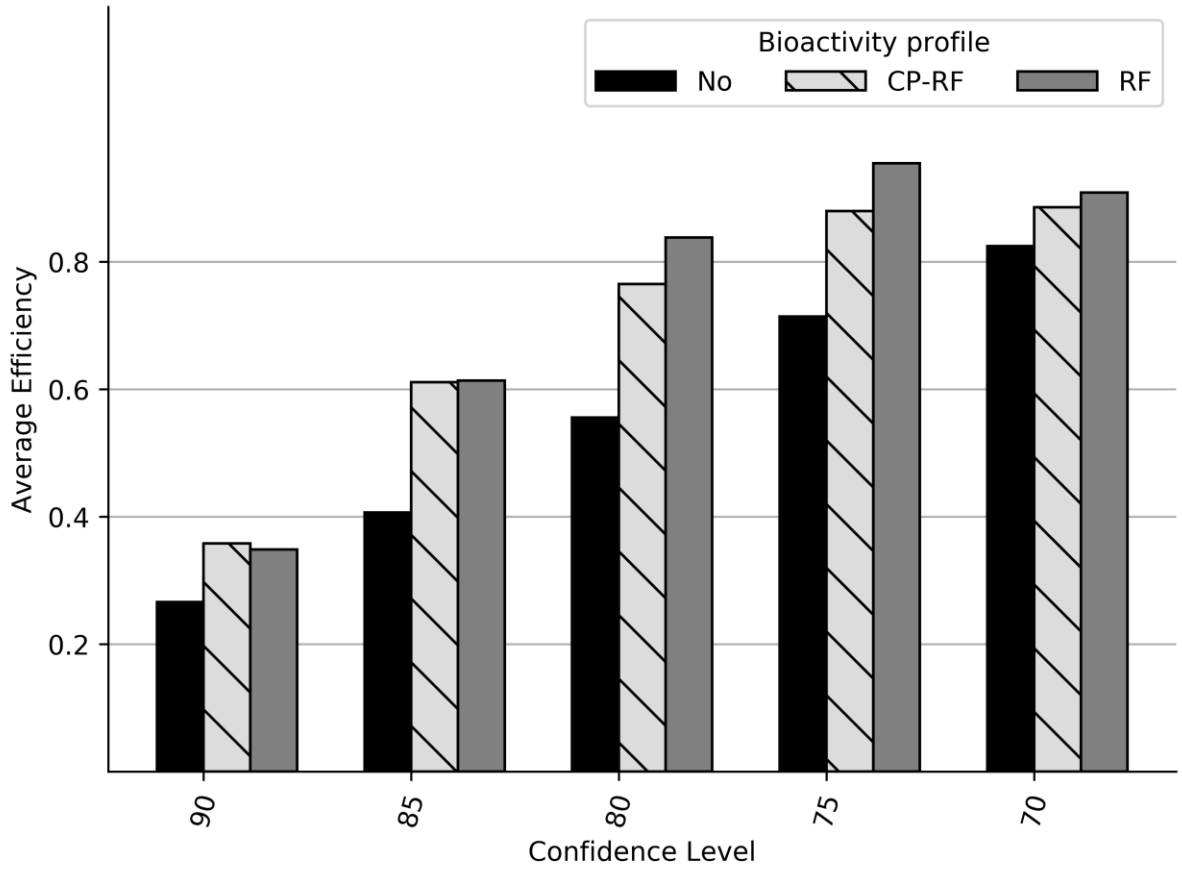

Figure 11. Efficiency for the negative class of the five external cytotoxicity data sets.

It might be expected that adding information from targets with a high similarity to the one being considered would be more beneficial to the model than targets that are less similar. Using a slightly modified form of the similarity measure described by Helal et al.[4] (defined as the Pearson correlation coefficient of binary compound activities in the assays using the measured activities for compounds measured in both assays) we compared the change in model performance with the average similarity of the five most similar assays from the bioactivity profile to the target being considered. The changes in model performance before and after the addition of the predicted bioactivity profiles alongside this similarity is shown in Figure 12. However, we did not observe a clear relationship between the target similarities and the gain in performance. Notably, it was clear that the models with a very high performance also before the added predicted bioactivity profiles did not on average benefit from the

addition for most cases, which is understandable. The lack of a clear relationship may reflect the balance between adding new information to the model, requiring a level of dissimilarity, without reaching such low levels of similarity that no valuable inferences can be drawn from the data.

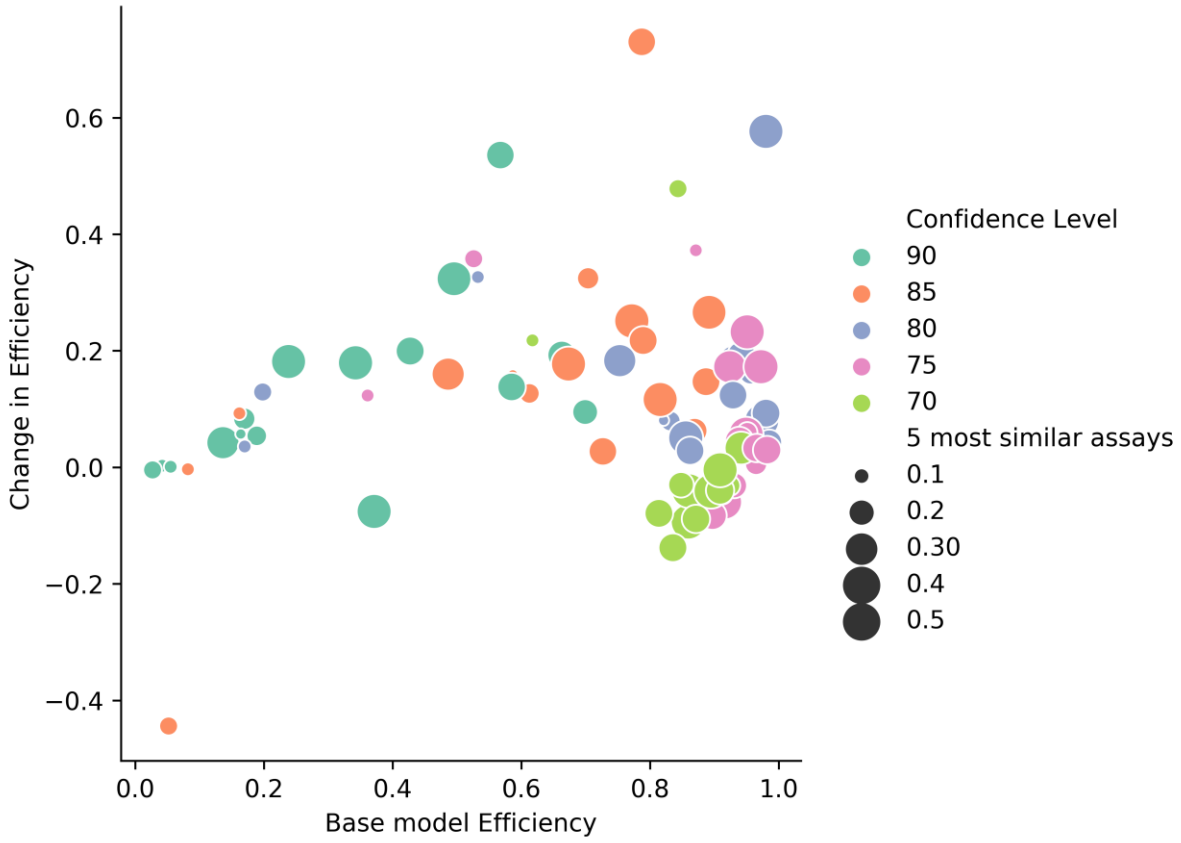

Figure 12. Change in efficiency after adding the predicted continuous bioactivity profiles generated by conformal prediction to the Cytotoxicity datasets models. Marker size is based on the average similarity to the five most similar targets in the predicted bioactivity profile (bigger is more similar).

Overall, our results show that using predicted bioactivity features can serve to increase model performance. This is in line with previous findings in the literature. Compared to previous studies we use the more rigorous conformal prediction framework for method evaluation, and we investigate the differences between using predicted labels and probabilities, showing that adding continuous bioactivity features gives a larger increase in model performance.

Conclusions

Similarly to what has been observed with bioactivity fingerprints, the addition of predicted bioactivity profiles increase the performance of QSAR modelling. The best performance was obtained when using the raw output, rather than the predicted labels, from the predictive models as additional features, either the number of trees or p-value for RF and conformal models, respectively.

The addition of the conformal framework provides both well-defined probabilities to generate the predicted bioactivity profiles as well as a robust way to measure the influence of the profiles on the models.

Supporting Information Available:

Additional performance metrics for the modelled datasets and results from using randomly generated features in place of the bioactivity profiles.

Acknowledgements

This work was supported by Alzheimer's Research UK (ARUK). The ARUK UCL Drug Discovery Institute is core funded by Alzheimer's Research UK (520909).

References

(1) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* **2018**, *23*, 1538–1546.

(2) Cumming, J. G.; Davis, A. M.; Muresan, S.; Haeberlein, M.; Chen, H. Chemical Predictive Modelling to Improve Compound Quality. *Nat. Rev. Drug Discov.* **2013**, *12*, 948–962.

(3) Lombardo, F.; Desai, P. V; Arimoto, R.; Desino, K. E.; Fischer, H.; Keefer, C. E.; Petersson, C.;

Winiwarter, S.; Broccatelli, F. In Silico Absorption, Distribution, Metabolism, Excretion, and Pharmacokinetics (ADME-PK): Utility and Best Practices. An Industry Perspective from the International Consortium for Innovation through Quality in Pharmaceutical Development. *J. Med. Chem.* **2017**, *60*, 9097–9113.

(4) Helal, K. Y.; Maciejewski, M.; Gregori-Puigjané, E.; Glick, M.; Wassermann, A. M. Public Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from PubChem's Bioassay Repository. *J. Chem. Inf. Model.* **2016**, *56*, 390–398.

(5) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, Å.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting Ligand Binding to Proteins by Affinity Fingerprinting. *Chem. Biol.* **1995**, *2*, 107–118.

(6) Riniker, S.; Wang, Y.; Jenkins, J. L.; Landrum, G. A. Using Information from Historical High-Throughput Screens to Predict Active Compounds. *J. Chem. Inf. Model.* **2014**, *54*, 1880–1891.

(7) Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.; Glick, M. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. *ACS Chem. Biol.* **2012**, *7*, 1399–1409.

(8) Mervin, L. H.; Bulusu, K. C.; Kalash, L.; Afzal, A. M.; Svensson, F.; Firth, M. A.; Barrett, I.; Engkvist, O.; Bender, A. Orthologue Chemical Space and Its Influence on Target Prediction. *Bioinformatics* **2018**, *34*, 72–79.

(9) Kar, S.; Das, R. N.; Roy, K.; Leszczynski, J. Can Toxicity for Different Species Be Correlated?: The Concept and Emerging Applications of Interspecies Quantitative Structure-Toxicity Relationship (i-QSTR) Modeling. *Int. J. Quant. Struct. Relationships* **2016**, *1*, 23–51.

(10) Mason, D. J.; Eastman, R. T.; Lewis, R. P. I.; Stott, I. P.; Guha, R.; Bender, A. Using Machine

Learning to Predict Synergistic Antimalarial Compound Combinations With Novel Structures. *Front. Pharmacol.* **2018**, *9*, 1096.

(11) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. "Bayes Affinity Fingerprints" Improve Retrieval Rates in Virtual Screening and Define Orthogonal Bioactivity Space: When Are Multitarget Drugs a Feasible Concept? *J. Chem. Inf. Model.* **2006**, *46*, 2445–2456.

(12) Li, X.; Kleinstreuer, N. C.; Fourches, D. Hierarchical Quantitative Structure–Activity Relationship Modeling Approach for Integrating Binary, Multiclass, and Regression Models of Acute Oral Systemic Toxicity. *Chem. Res. Toxicol.* **2020**, *33*, 353–366.

(13) Martin, E. J.; Polyakov, V. R.; Zhu, X.-W.; Tian, L.; Mukherjee, P.; Liu, X. All-Assay-Max2 PQSAR: Activity Predictions as Accurate as Four-Concentration IC50s for 8558 Novartis Assays. *J. Chem. Inf. Model.* **2019**, *59*, 4450–4459.

(14) Whitehead, T. M.; Irwin, B. W. J.; Hunt, P.; Segall, M. D.; Conduit, G. J. Imputation of Assay Bioactivity Data Using Deep Learning. *J. Chem. Inf. Model.* **2019**, *59*, 1197–1204.

(15) 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.

(16) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57*, 2490–2504.

(17) Rodríguez-Pérez, R.; Bajorath, J. Prediction of Compound Profiling Matrices, Part II: Relative Performance of Multitask Deep Learning and Random Forest Classification on the Basis of Varying Amounts of Training Data. *ACS Omega* **2018**, *3*, 12033–12040.

(18) de la Vega de León, A.; Chen, B.; Gillet, V. J. Effect of Missing Data on Multitask Prediction

Methods. *J. Cheminform.* **2018**, *10*, 26.

(19) Martin, E.; Mukherjee, P.; Sullivan, D.; Jansen, J. Profile-QSAR: A Novel Meta-QSAR Method That Combines Activities across the Kinase Family To Accurately Predict Affinity, Selectivity, and Cellular Activity. *J. Chem. Inf. Model.* **2011**, *51*, 1942–1956.

(20) Martin, E. J.; Polyakov, V. R.; Tian, L.; Perez, R. C. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50s for Realistically Novel Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 2077–2088.

(21) Irwin, B. W. J.; Mahmoud, S.; Whitehead, T. M.; Conduit, G. J.; Segall, M. D. Imputation versus Prediction: Applications in Machine Learning for Drug Discovery. *Futur. Drug Discov.* **2020**, *2*, FDD38.

(22) Mervin, L. H.; Afzal, A. M.; Drakakis, G.; Lewis, R.; Engkvist, O.; Bender, A. Target Prediction Utilising Negative Bioactivity Data Covering Large Chemical Space. *J. Cheminform.* **2015**, *7*, 51.

(23) Lampa, S.; Alvarsson, J.; Arvidsson Mc Shane, S.; Berg, A.; Ahlberg, E.; Spjuth, O. Predicting Off-Target Binding Profiles With Confidence Using Conformal Prediction. *Front. Pharmacol.* **2018**, *9*, 1256.

(24) Lee, K.; Lee, M.; Kim, D. Utilizing Random Forest QSAR Models with Optimized Parameters for Target Identification and Its Application to Target-Fishing Server. *BMC Bioinformatics* **2017**, *18*, 567.

(25) Roy, K.; Ambure, P.; Aher, R. B. How Important Is to Detect Systematic Error in Predictions and Understand Statistical Applicability Domain of QSAR Models? *Chemom. Intell. Lab. Syst.* **2017**, *162*, 44–54.

(26) Roy, K.; Ambure, P.; Kar, S. How Precise Are Our Quantitative Structure–Activity Relationship

Derived Predictions for New Query Chemicals? *ACS Omega* **2018**, *3*, 11392–11406.

(27) Vovk, V.; Gammerman, A.; Shafer, G. Algorithmic Learning in a Random World; Springer: New York, 2005; pp 1–324.

(28) Löfström, T.; Boström, H.; Linusson, H.; Johansson, U. Bias Reduction through Conditional Conformal Prediction. *Intell. Data Anal.* **2015**, *19*, 1355–1375.

(29) Norinder, U.; Boyer, S. Binary Classification of Imbalanced Datasets Using Conformal Prediction. *J. Mol. Graph. Model.* **2017**, *72*, 256–265.

(30) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Res.* **2012**, *40*, D400–D412.

(31) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955–D963.

(32) Svensson, F.; Norinder, U.; Bender, A. Modelling Compound Cytotoxicity Using Conformal Prediction and PubChem HTS Data. *Toxicol. Res. (Camb).* **2017**, *6*, 73–80.

(33) Norinder, U.; Svensson, F. Multitask Modeling with Confidence Using Matrix Factorization and Conformal Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 1598–1604.

(34) RDKit: Open-Source Cheminformatics (http://www.rdkit.org).

(35) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

(36) 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.

(37) 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.

(38) 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.

(39) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform.* **2015**, *7*, 20.

for Table of Contents use only

Using Predicted Bioactivity Profiles to Improve Predictive Modelling

Ulf Norinder, Ola Spjuth, Fredrik Svensson