Multi-task Modeling with Confidence using Matrix Factorization and Conformal Prediction

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

1. Swetox, Unit of Toxicology Sciences, Karolinska Institutet, Forskargatan 20, SE-151 36 Södertälje, Sweden

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

3. Alzheimer's Research UK UCL Drug Discovery Institute, University College London, Cruciform Building, Gower Street, London, WC1E 6BT, UK

4. The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, UK

Macau, multi-task learning, confidence, conformal prediction

ABSTRACT

Multi-task prediction of bioactivities is often faced with challenges relating to the sparsity of data and imbalance between different labels. We propose class conditional (Mondrian) conformal predictors using underlying Macau models as a novel approach for large scale bioactivity prediction. This approach handles both high degrees of missing data and label imbalances while still producing high quality predictive models. When applied to ten assay endpoints from PubChem, the models generated valid models with an efficiency of 74.0 - 80.1 % at the 80 % confidence level with similar performance both for the minority and majority class. Also when deleting progressively larger portions of the available data (0 - 80 %) the performance of the models remained robust with only minor deterioration (reduction in efficiency between 5-10 %). Compared to using Macau without conformal prediction the method presented here significantly improves the performance on imbalanced datasets.

Introduction

Multi-task learning methods have recently attracted considerable attention for modelling large scale bioactivity data.[1,2] Probabilistic Matrix Factorization is one approach that can be used in this setting since it allows for the modelling of sparse data with multiple possible outcomes.[3] A recently published development in this area is Macau, a Scalable Bayesian Multi-relational Factorization with Side Information using MCMC.[4] This new method incorporates several highly desirable features in a unified Bayesian framework and has been shown to give good performance on bioactivity data[5] while also handling large portions of missing data.[6]In contrast to most other methods with similar capabilities, like deep neural networks, matrix factorization provides models that are readily interpretable while providing similar predictive performance.[6]

Multi-task prediction of large scale bioactivity data is usually associated with two significant challenges. The available data is often incomplete and highly imbalanced, something that is challenging for many machine learning approaches.

Conformal predictors are examples of confidence predictors operating at a user-defined confidence level.[7] Data imbalance has been shown to be gracefully handled by class conditional (often called Mondrian) conformal predictors.[8–11] This is achieved by generating calibration sets for each of the classes separately, tuning the predictor to the observed distributions for the specific class. Similarly, a Mondrian conformal predictor that is calibrated on each class separately will not only handle imbalance between active and inactive compounds but in a multi-target prediction setting also handle different distributions between target labels.[12] These features have recently been shown to be valuable additions to bioactivity prediction.[13]

Previous studies applying Macau to large scale bioactivity data presented by de la Vega de León and co-workers[6] have shown less satisfactory performance when modeling imbalanced data sets. This is particularly notable for the minority class with respect to the retrieval of these compounds.

Conformal prediction (CP) is designed as a flexible framework allowing the use of any underlying predictor. Combining Macau with conformal prediction should yield a robust multi-task predictor capable of handling sparse data and highly imbalanced datasets.

In this study we report on the combination of Macau with conformal prediction to generate a multi-task confidence predictor. The predictor was evaluated on ten high-throughput screening datasets (endpoints) originating from PubChem[14] compiled by de la Vega de León and co-workers[6] and the amount of information provided in order to generate the model was varied to evaluate how well the proposed procedure handles missing data.

MATERIALS AND METHODS

Dataset

We used the same HTSFP10 dataset as reported by de la Vega de León, Chen and Gillet in the form submitted under "Availability of data and materials" including all descriptors (Morgan fingerprints of radius two hashed to 1,024 bits).[6]

The dataset consist of ten different PubChem HTS assays[15] and 56,892 compounds with complete coverage of the assay data matrix, i.e. all compounds have been measured in all assays. The activity is binary (active, inactive). A summary of the dataset is shown in Table 1.

Table 1. PubChem ID and percentage of active compounds in the ten HTSFP10 assays. The sets range from balanced to a fairly high imbalance of 1:12.

AID assay id	# active compounds	% active
687014	4,321	7.6
463190	4,450	7.8
588726	5,034	8.8
652054	5,035	8.9
485346	5,431	9.5
2796	5,570	9.8
504652	6,472	11.4
743279	9,433	16.6
1814	16,113	28.3
2314	26,307	46.2

Experiments

The dataset was randomly divided into a training set (75 %) and a test set (25 %). The training set was subsequently randomly subdivided into a calibration set (20 %) and a proper training set (80 %).

Twenty pairs of calibration set and proper training set were generated for aggregated conformal prediction,[16] each used to train and calibrate a model for prediction of the test set. Final prediction was made using the median of the predicted p-values. Each calibration set and test set contained 8,533 and 14,223 compounds respectively. The Macau models were trained using the 20 proper training sets and the mean value used for prediction.

Seven different experiments were performed, six of them with varying degrees and distributions of missing values in the proper training set (Table 2). The data was deleted either at random across the whole data frame (overall) or by randomly deleting the same fraction of active and inactive compounds for each outcome (individual). The introduction of missing values generated empty cells in the multi-target matrix, i.e. the matrix always contained the same number of rows and columns in all experiments, but maintained the size of the matrix making it sparser.

Table 2. Experiments with various amounts of proper training set missing values.

Experiment	% proper training set missing values	distribution[a]
1	0	
2	20	individual
3	50	individual
4	80	individual
5	20	overall
6	50	overall
7	80	overall

a. individual = same percentage of missing value in each assay; overall = percentage of missing values distributed across the entire multi-target matrix.

Software

The Macau Python package was installed from https://github.com/jaak-s/macau. For the classification, we used the code for the Macau classification analysis as deposited by de la Vega de León *et al*. (https://doi.org/10.5281/zenod o.1230488) with minor modifications to allow for the above mentioned split of train and test set and conformal post processing.

All parameters (keywords) were kept at default values except for the number of iterations to drop during training (burnin) that was set to 300 (default: 400).

Application of Conformal Prediction

The framework of Mondrian Conformal Prediction[17] was applied to each pair of calibration set and test set using an in-house Python script. This means that test set compounds were ranked against each of the two calibration class sets (one sorted list in descending order of prediction value for each class) and accordingly assigned a class label with respect to the chosen significance level (percentage of acceptable errors). Macau "classification" is a regression analysis with two types of class labels (1 = active and 0 = inactive), and the classification cut-off (decision boundary) is mid-way between the two classes at 0.5. The Conformal Prediction conformity value used in this analysis is the difference between the predicted value and the decision boundary keeping the class label in mind. If, for example, a test compound has a predicted value of 0.7 then the conformity value for the active class is 0.2 (0.7 – 0.5) and -0.2 for the inactive class (0.5 – 0.7). Comparing these two values against the sorted lists of corresponding values for the calibration compounds belonging to each class, respectively, determine the compounds' corresponding CP p-value for each class.

The median p-value for each compound in the test set over all 20 predictions (one prediction for each generated calibration set) for each class were used for final class assignment (aggregated conformal prediction[16]) by comparing the p-value to the selected *confidence level*. A confidence level of 80 % (0.8) would correspond to a *significance level* of 0.2 (1 - confidence level) and any predictions having a p-value exceeding this significance level will be assigned the label under consideration.

Evaluation of results

Conformal predictors are usually evaluated based on their *Validity* and *Efficiency*. A consequence of the conformal approach is that instances can be assigned between zero and two labels for a binary classification problem. Validity measures how many instances have been assigned the correct label (including both predictions), and should correspond to the user determined confidence label of the predictor, while efficiency measures the number of instances assigned only one label irrespective of the predicted class (correct label or not). The desired outcome is a predictor that is valid while being as efficient as possible.

Validity and efficiency are defined as:

Validity (active class) = (NB + TPS) / (NB + TPS + FNS + NE)

Validity (inactive class) = (NB + TNS) / (NB + TNS + FPS + NE)

Efficiency = (TPS + FNS + TNS + FPS) / (TPS + FNS + TNS + FPS + NB + NE)

where TPS, FNS, TNS, FPS are single-labeled true positives, false negative, true negative and false positives, respectively, and NB, NE are the number of compounds classified as both and empty, respectively. For a more in-depth introduction to these metrics and the principles behind conformal prediction we refer readers to Norinder *et al.*[18]

When evaluating if a predictor is valid, we consider a validity greater than one percent below the set validity level to be acceptable. For example, a predictor with a set confidence level of 80 % is considered valid if the achieved validity is 79 %.

For single label predictions we also calculate Matthews correlation coefficient (MCC), F1 score as well as recall, precision and sensitivity for the minority class and specificity for the majority class (see Supporting Information). Details on how the both or empty label predictions were handled for these methods is also given in the Supporting Information.

RESULTS AND DISCUSSION

Overall, all of the considered assay endpoints are well predicted. For convenience, we have chosen to display and discuss mainly around the results of the conformal predictors at an 80 % confidence level but additional results for other confidence levels are available in the Supporting Information. At this confidence level (80 %), the conformal predictors for the full dataset for all but one of the endpoints, inactive class of 743279, achieve the set validity of 80 % and an efficiency between 80.1 - 74.0 %, as depicted in Figures 1 – 3. The two methods used for generating sparser training data yielded comparable results.

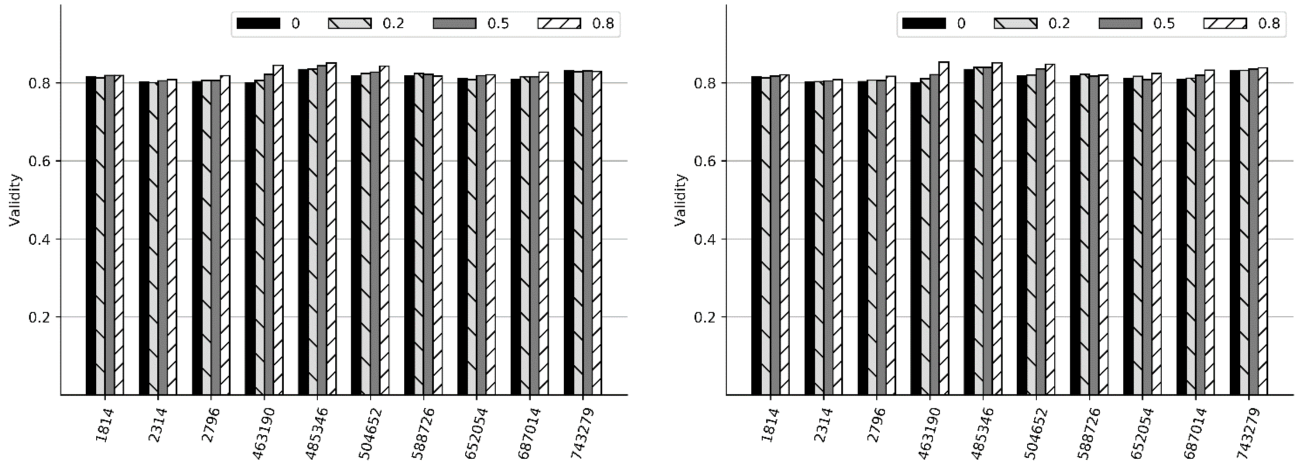


Figure 1. Test set validities for the active class across the different datasets at confidence level 80 % (0.8). Results for individual class deletions left and overall deletions right. Size of the deletion is indicated by the different bar styles; 0, 0.2, 0.5 and 0.8 from left to right for the respective datasets. The models achieve the expected validity of around 80 %.

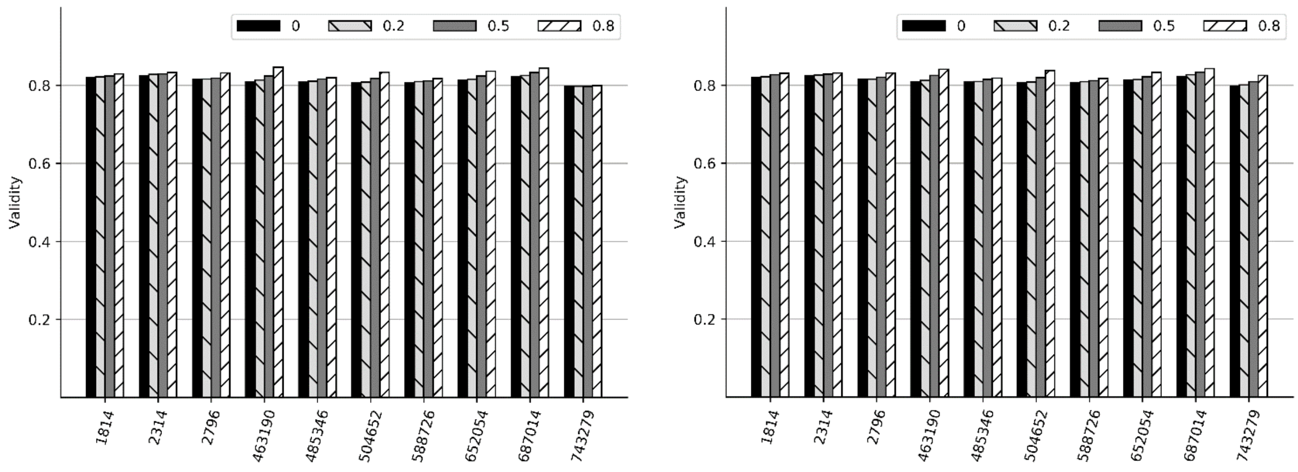


Figure 2. Test set validities for the inactive class across the different datasets at confidence level 80 % (0.8). Results for individual class deletions left and overall deletions right. Size of the deletion is indicated by the different bar styles; 0, 0.2, 0.5 and 0.8 from left to right for the respective datasets. The models achieve the expected validity of around 80 %.

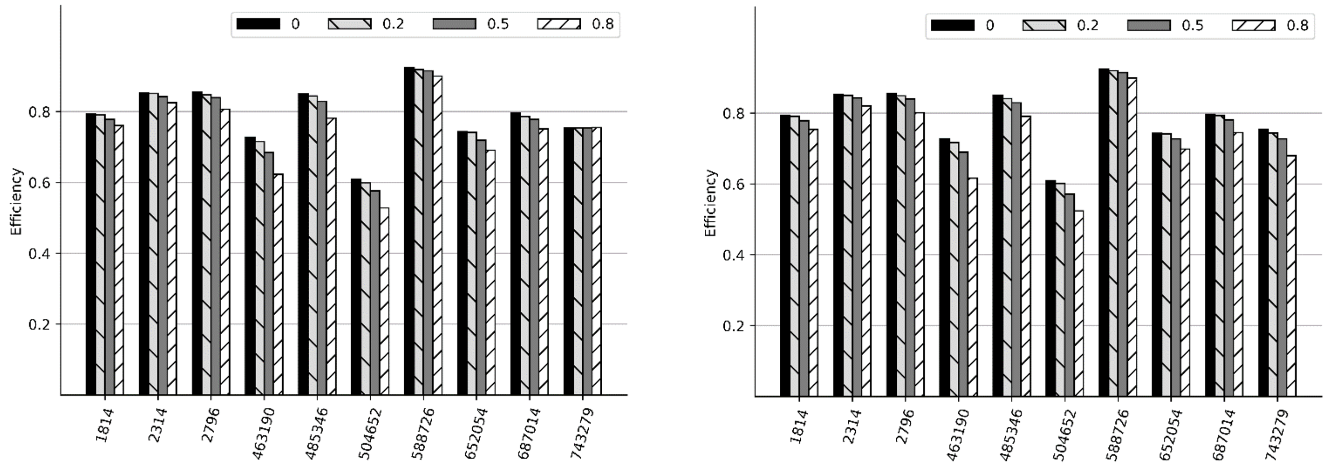


Figure 3. Test set efficiencies across the different datasets at confidence level 80 % (0.8). Results for individual class deletions left and overall deletions right. Size of the deletion is indicated by the different bar styles; 0, 0.2, 0.5 and 0.8 from left to right for the respective datasets.

Figure 4 shows how the predicted labels distribute for outcome 2796 using conformal prediction at the 80 % confidence level. Only a small fraction of the compounds do not receive single label predictions.

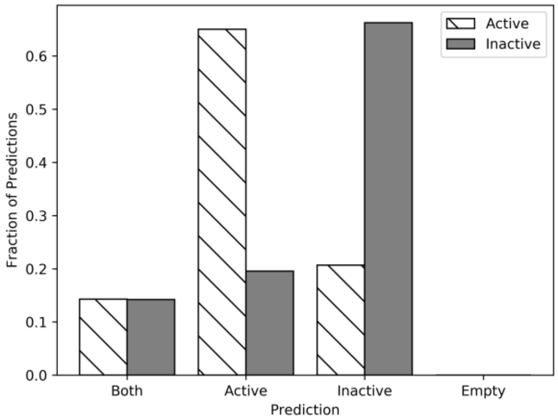


Figure 4. Outcomes of the conformal predictor at the 80 % confidence level for AID 2796. Most compounds receive a single label prediction (Active or Inactive) and a small portion is predicted as Both.

Using more traditional classification metrics, at the 80 % confidence level the conformal predictors achieve an average sensitivity for the test set single label predictions of 0.76 while generating single label predictions for between 80.1 - 74.0 % of the active compounds (corresponds to the efficiency).

In previous applications of Macau on bioactivity data,[6] one of the main limitations of Macau when applied to bioactivity data was the poor recall of the minority class. Gratifyingly, when combined with a conformal predictor the performance is excellent also on the minority class. A comparison of the model recall using Macau with and without conformal prediction is shown in Figure 5. On average, using the full data matrix, the recall of the active class was 0.14 for the Macau models and 0.62 with the addition of conformal prediction (80 % confidence level). It should here be noted that the recall calculated for the CP models is defined as:

TPS / P

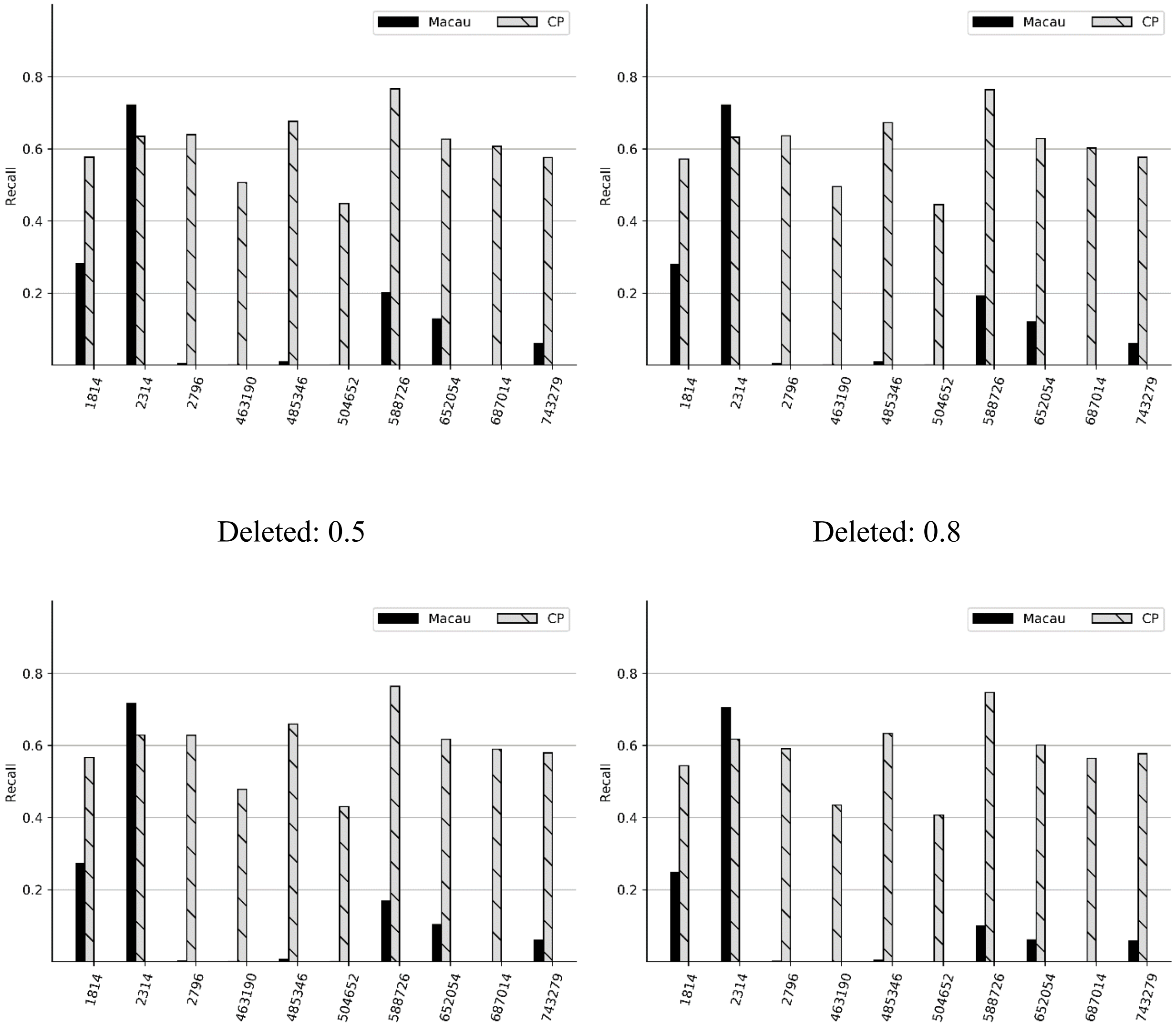
where TPS are single-labeled true positives, P all active compounds with TPS $\leq$ TP (true positives). Thus, the recall reported and used for the CP comparison with the corresponding traditional recall value for Macau models can be looked upon as a lower estimate of recall for CP models since, in a traditional setting, some of the *both* class predicted active compounds would, in most cases, be classified as true positives.

Looking at the results from a practical perspective, using the conformal Macau models at the 80 % confidence level to predict the bioactivity and then screening the compounds that receive a single label active prediction would result in an average of 3,733 compounds being screened while locating an average of 1,347 active compounds.

A closer look at the retrieval of the minority class for non-conformal Macau shows that it is closely related to the ratio of active and inactive compounds of the specific assay endpoint. For endpoint 2314, where the classes are fairly well balanced, a similar performance with or without the CP framework is observed while for endpoints such as 687014 and 463190 with more pronounced imbalances (Table 1) the retrieval of the minority class is only a few compounds. For endpoints 687014 and 463190 very few minority class compounds are retrieved out of in total 1,090 and 1,143 such compounds, respectively, in the test set.

Although the CP based models handle the imbalance much better, there is a tradeoff with respect to the increased number of false positive predictions for the minority class (Supporting Information). While this might be a significant drawback in some applications, for bioactivity predictions it is likely a favorable tradeoff to identify more active compounds despite a corresponding increase in false positives. This is especially true if the predictions will subsequently be validated experimentally.

Deleted: 0 Deleted: 0.2



Deleted: 0.5

Deleted: 0.8

Figure 5. The performance (recall) of Macau with and without conformal prediction for the test set at confidence level 80 % (0.8). When combined with conformal prediction the retrieval of the minority class is greatly enhanced.

One of the main benefits with Macau is the ability to handle sparse data. To test how the method handles different amounts of missing data within the framework of CP, we deleted different proportions of proper training data and retrained the predictor. Results when deleting 20 %, 50 %, and 80 % of all data in each endpoint is shown in Figures 1-3. Not only do the models overall handle large amounts of missing data, but looking at the individual class predictions the Mondrian conformal predictors still recover the minority class at the set level of confidence, i.e. the predictors are valid, in all but for a few endpoints. In six cases for the majority class of endpoint 743279 (individual assay endpoint deletion of values) at the 75 % (Supporting Information) and 80 % confidence levels, respectively, the models are not valid. A few other endpoint cases (7) deviate from the set validity at the 75 % and 85 % confidence levels for the minority class out of a total of 90 models built altogether.

Looking at the efficiency of the models there is a slight drop in the number of single label predictions when the number of missing data points increase. However, this reduction, between 5 % and 10 %, is relative minor considering the large amount of missing data.

We also performed experiments by deleting 20 %, 50 %, and 80 % of the data but randomly across the entire bioactivity matrix (overall deletion). This will generate different distributions of missing data for the investigated endpoints. Also, in this case the models handled significant portions of missing data without any significant drop in performance (Figures 1-3). In four cases (out in total 90 models) the predictions for the minority class were not valid at the investigated confidence levels.

It should here be noted that the total error in validities for the 9 and 8 models associated with the minority and majority class respectively, are on average only 1.30 % for the former and 1.24 % for the latter class. The breach in validity is minor for most practical applications and consequently, these models are most likely useful for predicting new compounds.

Furthermore, a slight increase in the validities for the minority as well as the majority class is observed as the amount of missing data increase in the proper training sets (Figures 6 and 7). This is expected in CP since the efficiencies at the same time slightly decreased as a consequence of more and more compounds being labels as "both" (Figure 8). However, the decrease in efficiency is quite modest considering the fact that the amount of missing data has increased from 0 % to 80 % in the proper training sets that constitute the basis for building the models.

Overall, the ability of Macau to handle missing data remains prominent also within the conformal framework.

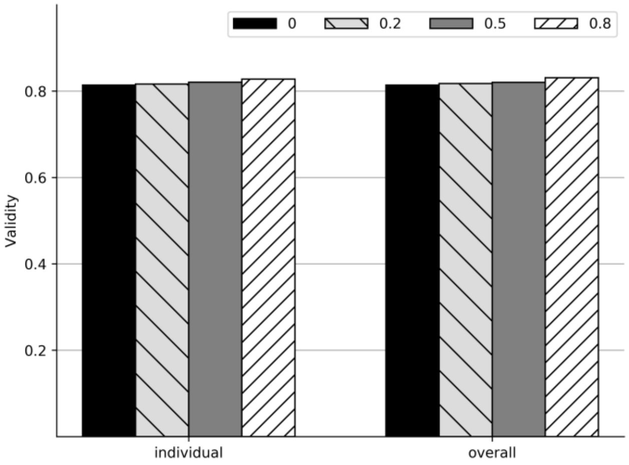


Figure 6. Average test set validities for the active class across all datasets at the 80 % (0.8) confidence level. Individual class deletions left and overall deletions right. Size of the deletion is indicated by the different bar styles; 0, 0.2, 0.5 and 0.8 from left to right for the respective deletions.

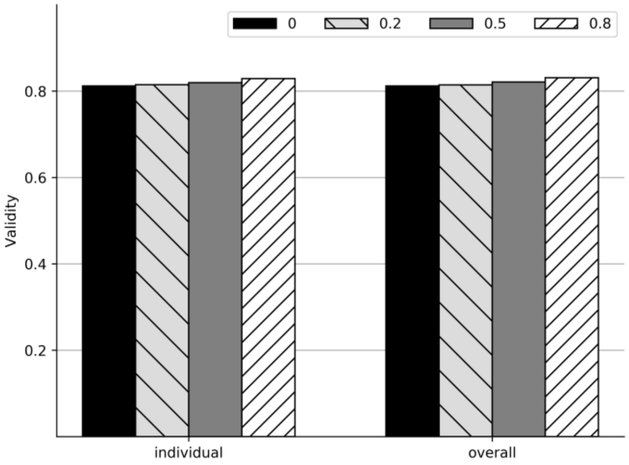


Figure 7. Average test set validities for the inactive class across all datasets at the 80 % (0.8) confidence level. Individual class deletions left and overall deletions right. Size of the deletion is indicated by the different bar styles; 0, 0.2, 0.5 and 0.8 from left to right for the respective deletions.

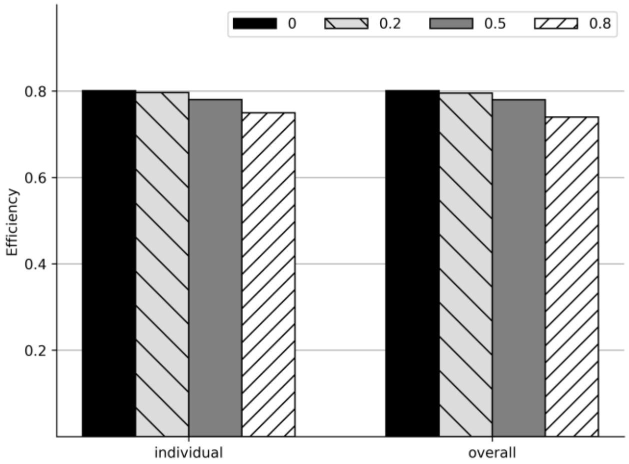


Figure 8. Overall test set efficiencies across all datasets at the 80 % (0.8) confidence level. Individual class deletions left and overall deletions right. Size of the deletion is indicated by the different bar styles; 0, 0.2, 0.5 and 0.8 from left to right for the respective deletions.

In this study, we have demonstrated that Mondrian conformal Macau is an excellent approach to multi-target bioactivity prediction that can handle both label imbalance and sparsity of data. However, a potential drawback with this approach is that the models are relatively computationally expensive to train, making the method unsuitable for models that need to be retrained at a high frequency. Still, for models that are expected to stay in production for some time, the presented method is a robust and high performing choice. It should also be noted here that the purpose of this investigation is not to primarily generate the best possible model, for which parameter (keyword) optimization is most likely needed, but to highlight the advantages of combining multi-target matrix factorization with a confidence predictor such as conformal prediction.

CONCLUSIONS

We show that confidence prediction using Macau coupled with conformal prediction is an excellent approach for predicting large scale multi-target bioactivity data. This approach can handle highly imbalanced data as well as missing data while delivering high quality predictions with associated confidence.

When applied to ten assay endpoints the average performance of the conformal Macau predictors was very strong with validities closely corresponding to the set confidence levels and an efficiency (fraction of single label predictions) of 80.1 - 74.0 % at the 80 % confidence level, with similar performance for both the majority and minority class. This despite the fact that the most imbalanced target label had more than twelve times of the majority label. This is in stark contrast to the performance of the Macau models without CP where the minority class in many cases was poorly predicted.

ASSOCIATED CONTENT

Supporting Information.

Tabulated values for the method performance on the 10 different assay endpoints for the 7 experiments.

AUTHOR INFORMATION

Corresponding Author

* f.svensson@ucl.ac.uk

Notes

The Authors declare no competing financial interest.

ACKNOWLEDGMENT

The ARUK UCL Drug Discovery Institute is core funded by Alzheimer’s Research UK (registered charity No. 1077089 and SC042474). The Francis Crick Institute receives its core funding from Cancer Research UK (FC001002), the UK Medical Research Council (FC001002), and the Wellcome Trust (FC001002).

REFERENCES

(1) Sosnin, S.; Vashurina, M.; Withnall, M.; Karpov, P.; Fedorov, M.; Tetko, I. V. A Survey of Multi-Task Learning Methods in Chemoinformatics. *Mol. Inform.* **2018**, doi:10.1002/minf.201800108.

(2) 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**, doi:10.1021/acs.jcim.8b00685.

(3) Cobanoglu, M. C.; Liu, C.; Hu, F.; Oltvai, Z. N.; Bahar, I. Predicting Drug–Target Interactions Using Probabilistic Matrix Factorization. *J. Chem. Inf. Model.* **2013**, *53*, 3399–3409.

(4) Simm, J.; Arany, A.; Zakeri, P.; Haber, T.; Wegner, J. K.; Chupakhin, V.; Ceulemans, H.; Moreau, Y. Macau: Scalable Bayesian Factorization with High-Dimensional Side Information Using MCMC. In *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*; 2017.

(5) Yang, M.; Simm, J.; Lam, C. C.; Zakeri, P.; van Westen, G. J. P.; Moreau, Y.; Saez-Rodriguez, J. Linking Drug Target and Pathway Activation for Effective Therapy Using

Multi-Task Learning. *Sci. Rep.* **2018**, *8*, 8322.

(6) de la Vega de León, A.; Chen, B.; Gillet, V. J. Effect of Missing Data on Multitask Prediction Methods. *J. Cheminform.* **2018**, *10*, 26.

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

(8) Sun, J.; Carlsson, L.; Ahlberg, E.; Norinder, U.; Engkvist, O.; Chen, H. Applying Mondrian Cross-Conformal Prediction To Estimate Prediction Confidence on Large Imbalanced Bioactivity Data Sets. *J. Chem. Inf. Model.* **2017**, *57*, 1591–1598.

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

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

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

(12) Shi, F.; Ong, C. S.; Leckie, C. Applications of Class-Conditional Conformal Predictor in Multi-Class Classification. In *2013 12th International Conference on Machine Learning and Applications*; 2013; Vol. 1, pp 235–239.

(13) Bosc, N.; Atkinson, F.; Felix, E.; Gaulton, A.; Hersey, A.; Leach, A. R. Large Scale Comparison of QSAR and Conformal Prediction Methods and Their Applications in Drug Discovery. *J. Cheminform.* **2019**, *11*, 4.

(14) Chen, B.; Wild, D. J. PubChem BioAssays as a Data Source for Predictive Models. *J. Mol. Graph. Model.* **2010**, *28*, 420–426.

(15) 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.

(16) 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.

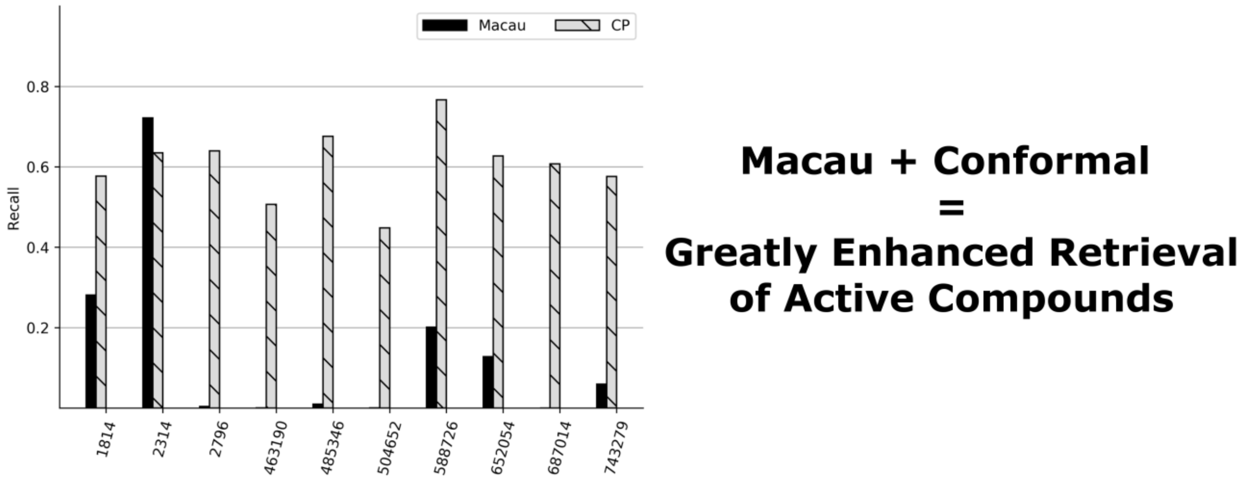
(17) Vovk, V. Conditional Validity of Inductive Conformal Predictors. *Mach. Learn.* **2013**, *92*, 349–376.

(18) 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.

for Table of Content use only

Multi-task Modeling with Confidence using Matrix Factorization and Conformal Prediction

Ulf Norinder, Fredrik Svensson


Macau
CP
Recall
Macau + Conformal
=
Greatly Enhanced Retrieval
of Active Compounds