Assessing the Calibration in Toxicological in Vitro Models with Conformal Prediction

Andrea Morger
Charite Universitatsmedizin Berlin https://orcid.org/0000-0003-4774-6291
Fredrik Svensson
Alzheimer's Research UK UCL Drug Discovery Institute https://orcid.org/0000-0002-5556-8133
Staffan Arvidsson McShane
Uppsala Universitet Uppsala biomedicinska centrum https://orcid.org/0000-0001-6709-7116
Niharika Gauraha
KTH: Kungliga Tekniska Hogskolan https://orcid.org/0000-0002-4446-2800
Ulf Norinder
Biomedicum: Uppsala Universitet Uppsala biomedicinska centrum https://orcid.org/0000-0003-3107-331X
Ola Spjuth
Uppsala Universitet Uppsala biomedicinska centrum https://orcid.org/0000-0002-8083-2864
Andrea Volkamer (andrea.volkamer@charite.de)
Institute of Physiology, Charit´e Universit¨atsmedizin, Berlin, Germany https://orcid.org/0000-0002-3760-580X

Research article

Keywords: toxicity prediction, conformal prediction, data drifts, applicability domain, calibration plots, Tox21 datasets

Posted Date: February 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-220364/v1

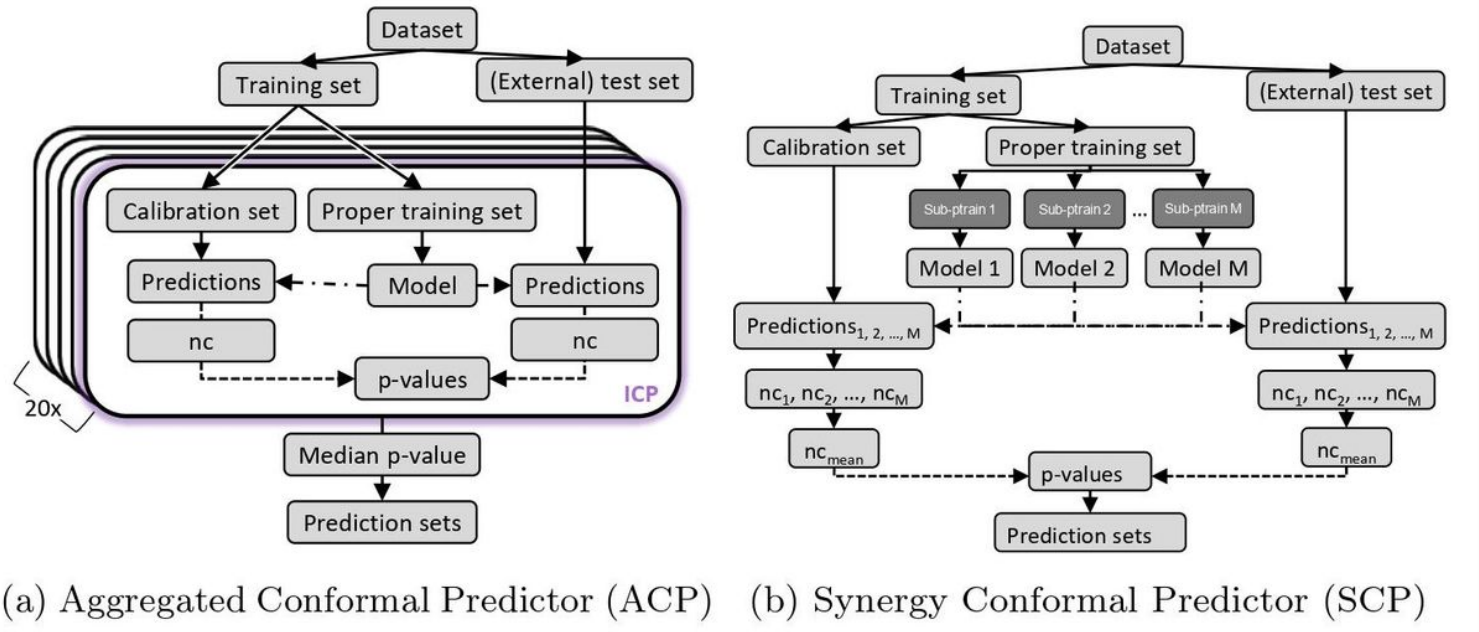
Abstract

Machine learning methods are widely used in drug discovery and toxicity prediction. While showing overall good performance in cross-validation studies, their predictive power (often) drops in cases where the query samples have drifted from the training data's descriptor space. Thus, the assumption for applying machine learning algorithms, that training and test data stem from the same distribution, might not always be fulfilled. In this work, conformal prediction is used to assess the calibration of the models. Deviations from the expected error may indicate that training and test data originate from different distributions. Exemplified on the Tox21 datasets, composed of chronologically released Tox21Train, Tox21Test and Tox21Score subsets, we observed that while internally valid models could be trained using cross-validation on Tox21Train, predictions on the external Tox21Score data resulted in higher error rates than expected. To improve the prediction on the external sets, a strategy exchanging the calibration set with more recent data, such as Tox21Test, has successfully been introduced. We conclude that conformal prediction can be used to diagnose data drifts and other issues relating to model calibration. The proposed improvement strategy — exchanging the calibration data only — is convenient as it does not require retraining of the underlying model.

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the latest manuscript can be downloaded and accessed as a PDF.

Figures



(a) Aggregated Conformal Predictor (ACP) (b) Synergy Conformal Predictor (SCP)

Figure 1

The aggregated conformal prediction methods used in this study. (a) ACP (purple box): The dataset is split into a training set and a test set. The training set is further split into a proper training set to train the model and a calibration set. The predictions made for the test set compounds are used to calculate nonconformity scores (nc) and compared to nonconformity scores in the calibration set to calculate p-values and generate prediction sets. In ACP, multiple models are trained and calibrated with randomly selected proper training and calibration sets, and p-values from these are averaged. (b) SCP: In order to ensure a uniform distribution of p-values, SCP averages the nonconformity scores instead. Multiple models are trained on (subsets of) the proper training set and with each model predictions are made for the test set and for a fixed calibration set.

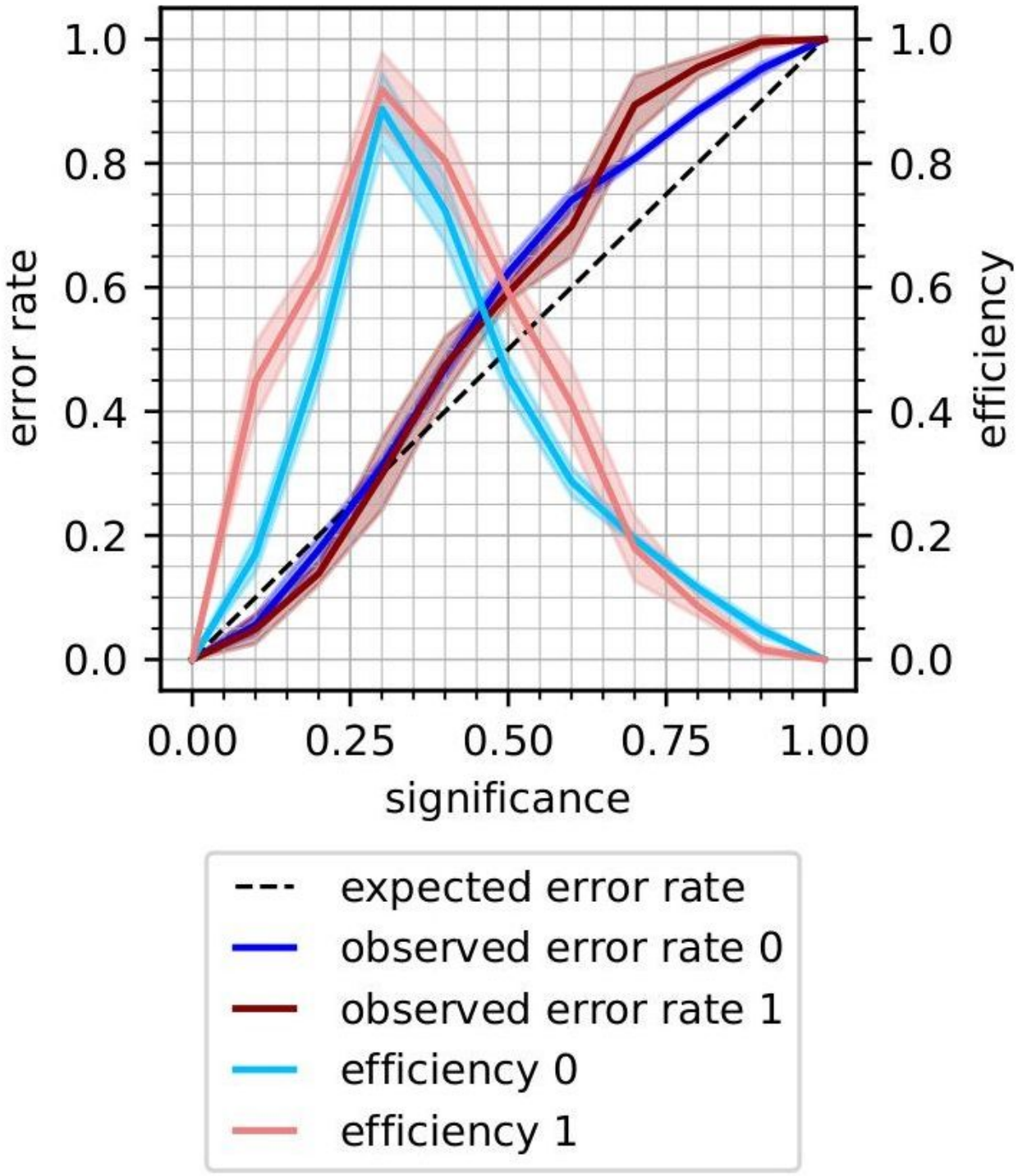


Figure 2

Calibration and efficiency plot. The dark lines show the mean error rate for the active (dark red) and inactive (dark blue) compounds. For a well-calibrated model, the error rate ideally follows the dashed diagonal line. The light coloured lines illustrate the mean efficiency expressed as ratio of single label sets for the active (light red) and inactive (light blue) compounds. The shaded areas indicate the respective

standard deviation within the fivefold cross-validation (CV). Class 0: inactive compounds, class 1: active compounds.

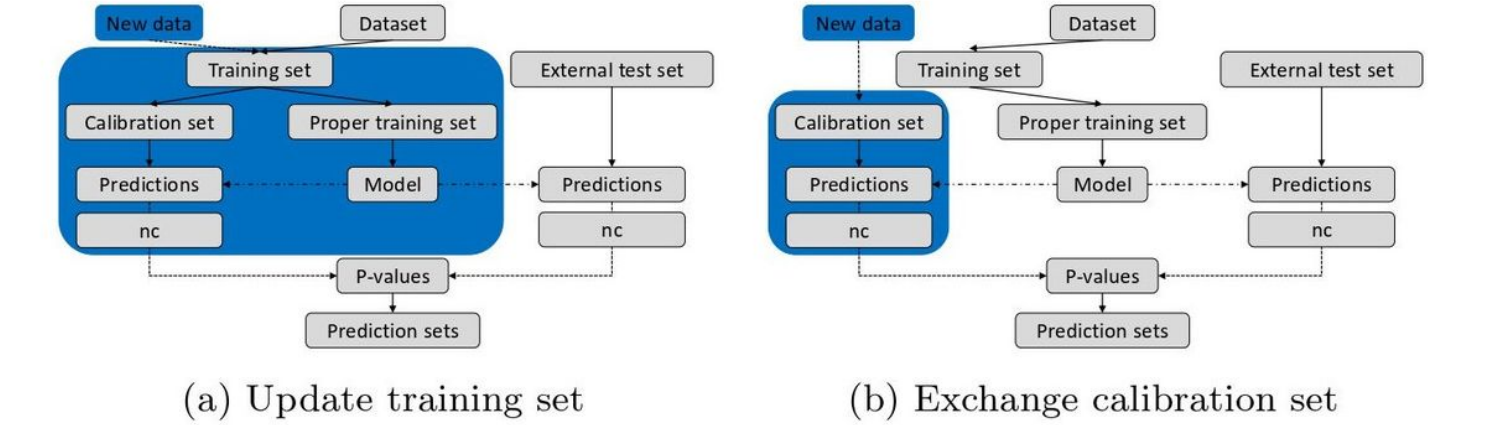
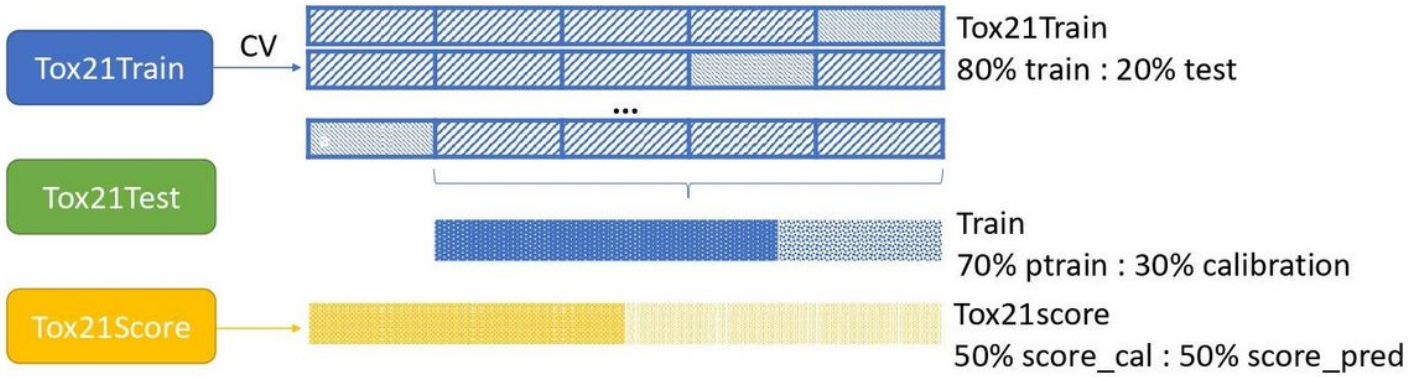


Figure 3

Model update strategies analysed to improve calibration. (a) The whole training set is updated with new data. This involves retraining a new model. (b) Only the calibration set is updated with new data. Models can hereby be re-calibrated without training a new model.



ptrain	calibrate	predict	name	aggregator	special splits
			internal_CV	ACP	
			pred_score	ACP	
			pred_score_SCP	SCP	divide ptrain into 4x 25% sub-ptrain
			train_update	ACP	train = train + Tox21Test
			cal_update	ACP	
			cal_update_2	ACP	

Figure 4

Overview of the experiments discussed in this work. Top: Splitting of Tox21 data into (proper) training, calibration and test set. Bottom: Data for training, calibration, and prediction and aggregator used in the specific experiments.

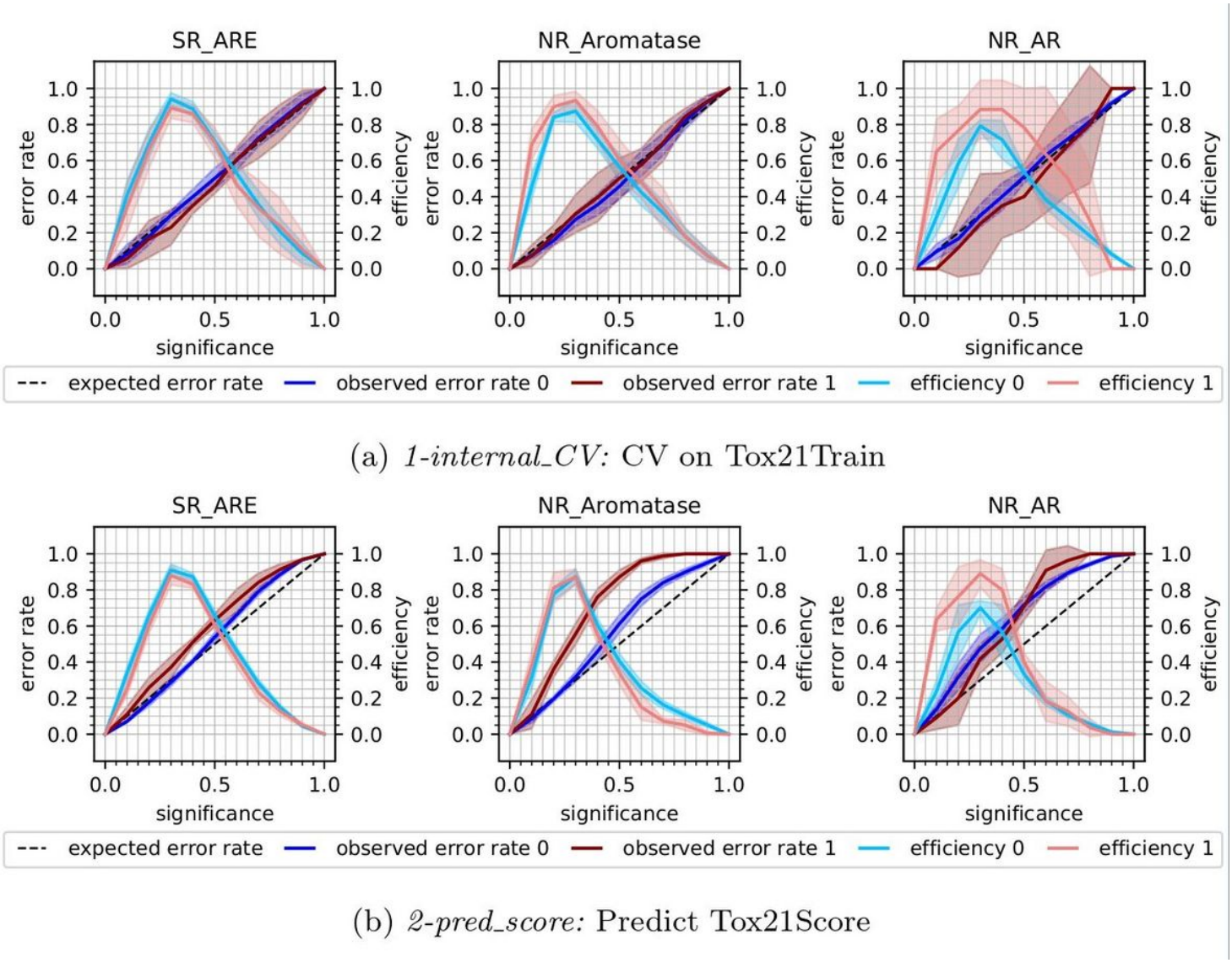


Figure 5

CEPs for models trained on Tox21Train and subsequent (a) internal cross-validation and (b) predictions on Tox21Score. CEPs for a selection of three example endpoints (SR ARE, NR Aromatase, NR AR). Class 0: inactive compounds, class 1: active compounds. For a detailed explanation of all the components in the CEP, see Figure 2.

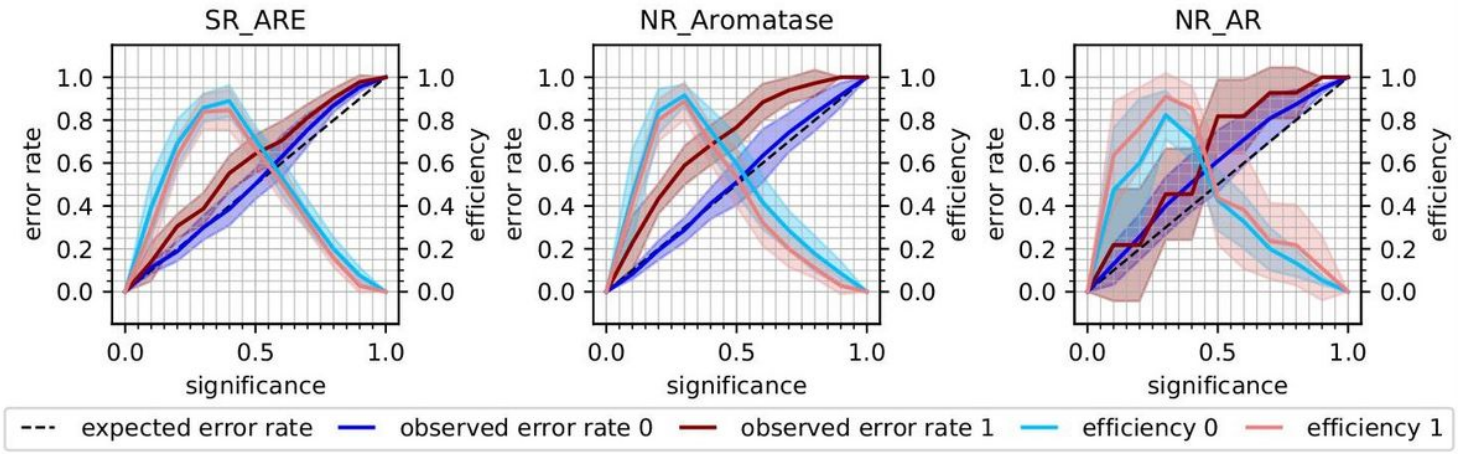
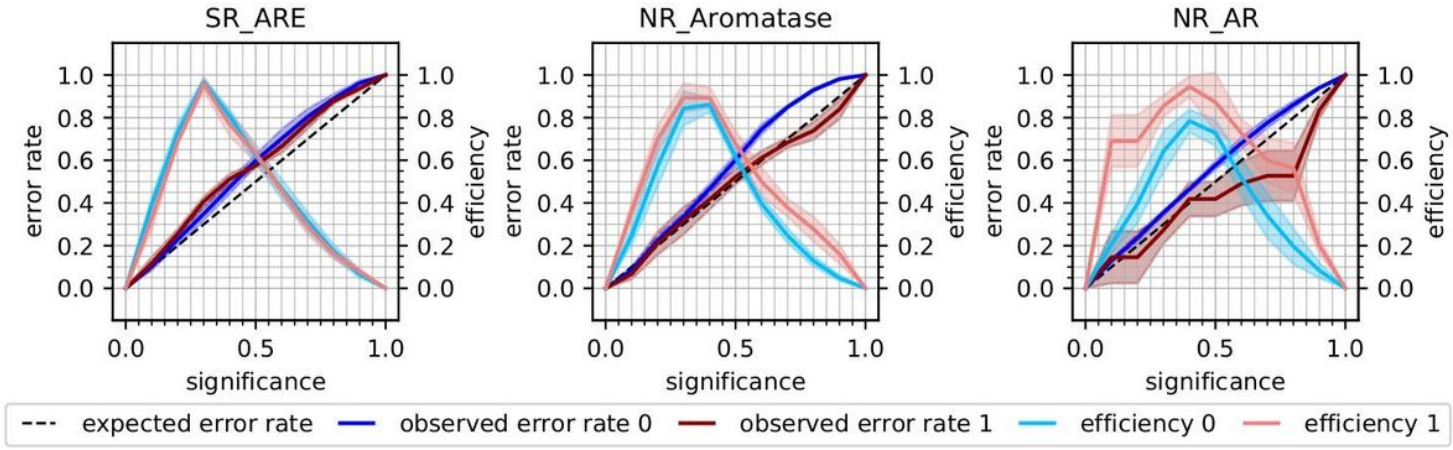
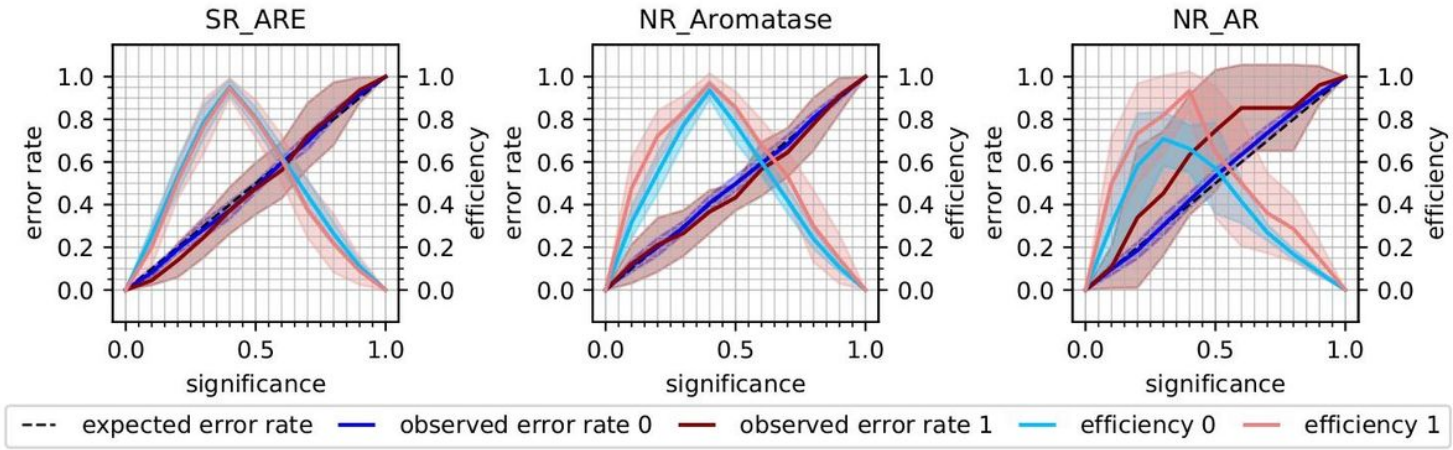


Figure 6

Results from Experiment 3-pred score scp: SCP models were trained on Tox21Train and predictions made for Tox21Score. CEPs are shown for a selection of three example endpoints (SR ARE, NR Aromatase, NR AR). Class 0: inactive compounds, class 1: active compounds. For a detailed explanation of all the components in the CEP, see Figure 2.



(a) *5-cal_update:* Exchange calibration set with Tox21Test



(b) *6-cal_update_2:* Exchange calibration set with half of Tox21Score

Figure 7

Updating the calibration set with more recent data. CEPs for a selection of three example endpoints (SR ARE, NR Aromatase, NR AR). Class 0: inactive compounds, class 1: active compounds. For a detailed explanation of all the components in the CEP, see Figure 2.

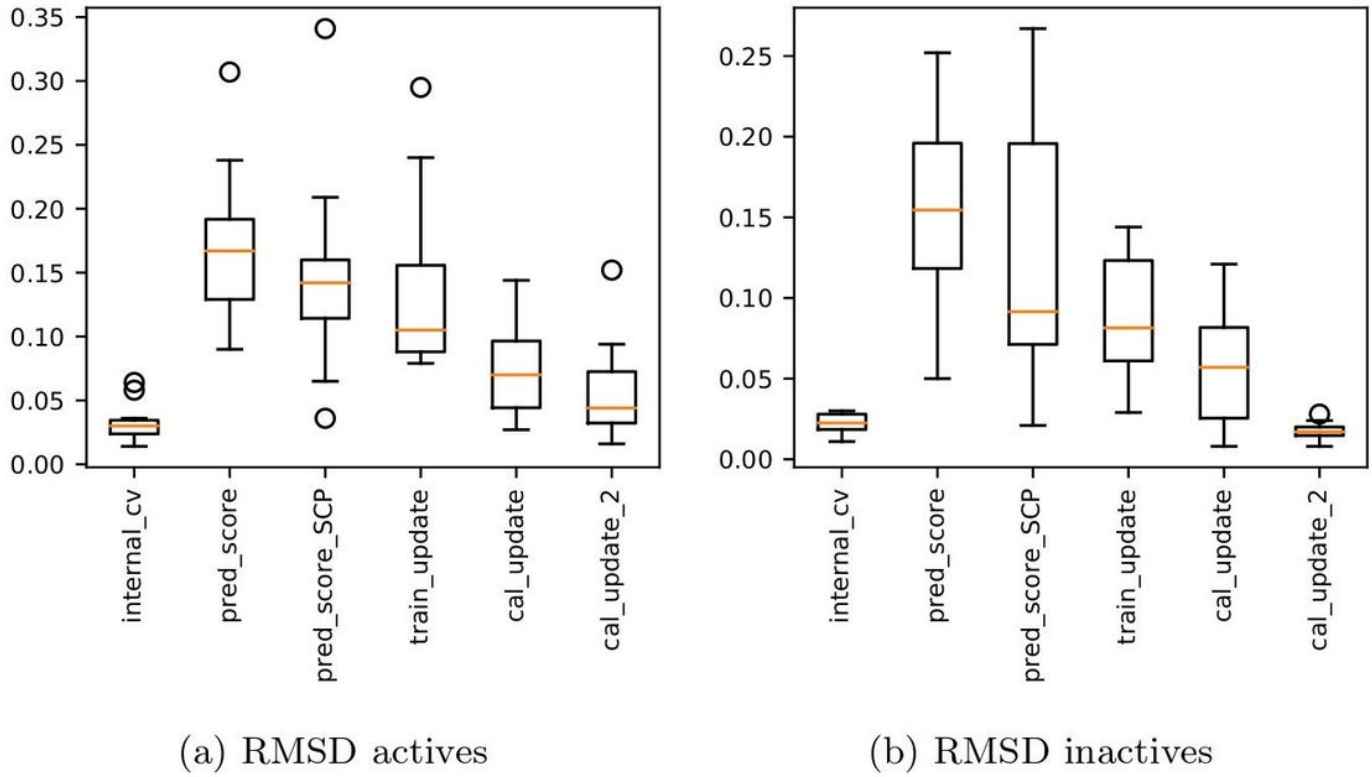


Figure 8

Box plots for the root-mean-square deviation (RMSD) between the expected (diagonal line) and observed error rate for all twelve Tox21 endpoints compared amongst the different experiments are shown. On the left results for the active compounds, on the right for the inactive compounds are plotted. Note that the y-axis ranges differ.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- additionalfiles.tex
- cptox21additionalfile.pdf