



**CHALMERS**  
UNIVERSITY OF TECHNOLOGY

## Temporal Evaluation of Probability Calibration with Experimental Errors

Downloaded from: <https://research.chalmers.se>, 2024-11-06 01:20 UTC






Citation for the original published paper (version of record):

Friesacher, H., Svensson, E., Arany, Á. et al (2025). Temporal Evaluation of Probability Calibration with Experimental Errors. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 14894 LNCS: 13-20.  
[http://dx.doi.org/10.1007/978-3-031-72381-0\\_2](http://dx.doi.org/10.1007/978-3-031-72381-0_2)

N.B. When citing this work, cite the original published paper.



# Temporal Evaluation of Probability Calibration with Experimental Errors

Hannah Rosa Friesacher<sup>1,3</sup> , Emma Svensson<sup>2,3</sup> , Adam Arany<sup>1</sup> ,  
Lewis Mervin<sup>4</sup> , and Ola Engkvist<sup>3,5</sup> 

<sup>1</sup> ESAT-STADIUS, KU Leuven, Leuven 3000, Belgium

<sup>2</sup> ELLIS Unit Linz, Institute for Machine Learning, Johannes Kepler University Linz, Linz 4040, Austria

<sup>3</sup> Molecular AI, Discovery Sciences, R&D, AstraZeneca Gothenburg, Mölndal 431 83, Sweden

`rosafriesacher@live.at`

<sup>4</sup> Molecular AI, Discovery Sciences, R&D, AstraZeneca Cambridge, Cambridge CB2 0AA, UK

<sup>5</sup> Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg 412 96, Sweden

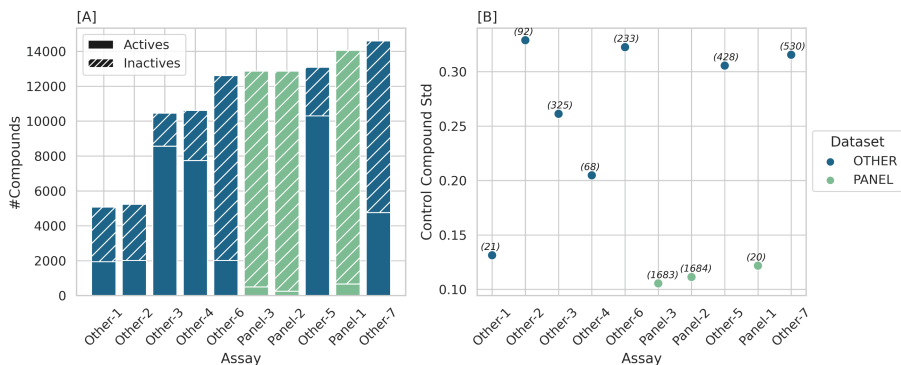
## 1 Introduction

The quantification of uncertainties associated with neural network predictions can facilitate optimal decision-making and accelerate workflows where time and resource efficiency are essential. In drug discovery, computational tools exist that estimate predictive uncertainties to enable the assessment of costs and risk in the discovery and development pipeline [11]. There are various sources of uncertainty in machine learning. A common classification found in literature is the distinction between aleatoric uncertainty, which originates from uncertainty in the data, and epistemic sources, which quantifies uncertainty inherent in the choice of model. We refer to Hüllermeier & Waegemann [6] and Gruber et al. [3] for a deeper discussion of uncertainty sources. It is important to point out that modern neural networks often fail to give realistic estimates of the uncertainty associated with a prediction in classification tasks, resulting in poorly calibrated models [4, 11]. There are various calibration methods for classification models, that aim to obtain better uncertainty estimates by fitting a calibrating model to a separate dataset in a post-hoc manner. Another strategy to achieve more reliable predictions is the incorporation of model uncertainty, by taking into account model variance, which increases when the model is overfitting or the test instance lies outside the domain of the training data. This work compares the performance of single-task classification models trained on industry-scale assay data in a temporal analysis. In contrast to random or cluster-based strategies to split the data, temporal splits simulate most accurately the drug discovery pipeline in pharmaceutical companies [16]. A temporal splitting strategy enables model training on older data and prediction on subsequent folds. We use temporal splits to compare the performance and calibration of Random Forest (RF)

models for classification tasks with and without post-hoc calibration using two different calibration approaches. Furthermore, we investigate whether the inclusion of data uncertainty in the form of probabilistic labels improves uncertainty estimation. Finally, we use the temporal setting to investigate how the temporal evolution of the test set affects model calibration.

## 2 Methods

We evaluate single-task classification models on data from ten assays and two assay categories, including 'Panel' and 'Other' assays [5]. The assays are labeled using the assay category combined with a number from 1 to 5, e.g. 'Panel-1'. The 'Panel' category comprises cross-project assays such as undesired off-target effects, whereas 'Other' includes project-specific assays from on-target activity screens. The data solely includes affinity data with pIC50 or pEC50 as endpoints. The assays were chosen to be representative, exhibiting various assay sizes and active ratios. Figure 1[A] summarizes the number of measurements and the ratio of actives for all assays used in our study. Standardized SMILES were obtained using the method described in the MELLODDY-TUNER [1] package and extended connectivity fingerprints (ECFPs) of size 1024 and radius 2 were generated with RDKit [8]. Given that the date of each measurement is available, a real temporal split was performed. After ordering the data according to the measurement date, the data was split into five folds of equal size, so that each fold represented a specific period in the assay history. For generating single-task classification models, two label types were used to assess if the incorporation of aleatoric uncertainty improves model performance. First, hard labels were generated using a pIC50/pEC50 threshold of 6 for assigning active or inactive labels based on the result. This specific threshold was chosen because the models will be deployed in the early stages of the drug discovery pipeline, in which the desired binding affinity of drug candidates is in the micromolar range ( $10^{-6}$  molar concentration) corresponding to a pIC50/pEC50 of 6. Second, the same threshold was applied and the assay-specific measurement error, corresponding to the standard deviation of the control compound measurements, was used to obtain probabilistic labels. In detail, a normal distribution  $X \sim \mathcal{N}(\mu, \sigma^2)$  was generated, where  $\mu$  corresponded to the chosen threshold and  $\sigma^2$  to the standard deviation of the control compound of the respective assay. In this step, the control compound corresponded to the compound with the most measurements in the respective assay. Subsequently, the CDF of these assay-specific distributions was used to obtain the probabilistic label [9]. Figure 1[B] shows the standard deviation (Std) of the control compound as well as the available number of measurements to calculate the Std for every assay.



**Fig. 1. Overview over Assay data.** Assays from two categories, 'Other' and 'Panel', were used. [A] The number of measurements (#Compounds) of each assay as the sum of active and inactive compounds (pIC50/pEC50 threshold = 6) is shown. [B] shows the standard deviation (Std) and the number of measurements (in brackets) of the control compound for every assay.

## 2.1 Model Generation

Random Forest (RF) models were generated using scikit-learn. The maximum depth of the trees and the required number of estimators were tuned using a validation dataset. Probability-like outputs were generated by taking the ratio of decision trees in an RF that voted for a specific test instance to be active. Furthermore, Probabilistic Random Forests (PRF) [15] were generated using probabilistic labels as ground truth. A detailed description of the PRF training procedure can be found in Mervin et al. [9]. Post-hoc probability calibration techniques fit a calibration model to the raw scores of a classifier using a separate calibration dataset. In our work, we use the validation dataset for this step. Two uncertainty calibration approaches were used, namely Platt scaling [14] and Venn-ABERS (VA) predictors [18]. Platt scaling [14] involves fitting a logistic regression to the classification scores to counteract over- or underfitted uncertainty estimations. For calibration with VA predictors [18] two isotonic regression functions are trained on the validation data and the test instance, representing the two possible hypotheses that the test instance is active versus inactive. As such, two different probabilities are obtained from the isotonic regression models, corresponding to a lower and an upper bound on the probability, which are subsequently condensed to a point estimate as proposed by Tocatelli et al. [17]. For more detail on VA predictors we refer to Mervin et al. [10].

## 3 Results

### 3.1 Incorporation of Aleatoric Uncertainty Using Measurement Errors

Table 1 summarizes the Binary Cross Entropy (BCE $\downarrow$ ) loss and the Adaptive Calibration Error (ACE $\downarrow$ ) [12] for five model repeats of all model types trained on two example datasets, namely the Panel-1 and Other-3 assays. The first

**Table 1. Overview over RF model performance based on two example assays.** Averages over five model repeats are shown. The best results for each metric are marked in bold, while not significantly worse scores are indicated in italics.

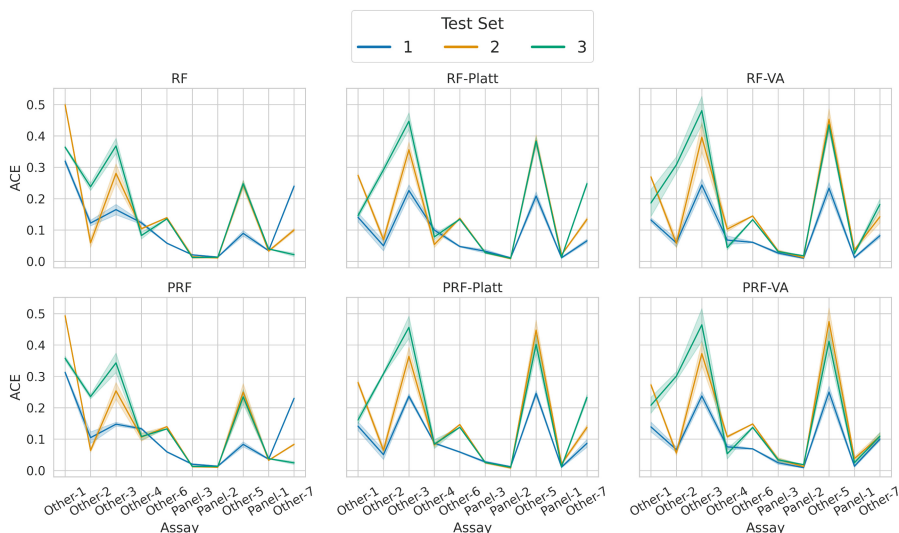
Method	Panel-1		Other-3	
	BCE ↓	ACE ↓	BCE ↓	ACE ↓
<b>Hard Labels</b>				
RF	<i>0.187 ± 0.005</i>	0.032 ± 0.001	0.312 ± 0.009	0.192 ± 0.008
RF-Platt	<b>0.182 ± 0.004</b>	<b>0.029 ± 0.002</b>	0.235 ± 0.007	0.117 ± 0.006
RF-VA	<i>0.183 ± 0.002</i>	0.037 ± 0.001	<b>0.211 ± 0.007</b>	<b>0.089 ± 0.006</b>
<b>Probabilistic Labels</b>				
PRF	<i>0.181 ± 0.002</i>	<i>0.032 ± 0.002</i>	0.307 ± 0.01	0.187 ± 0.008
PRF-Platt	<b>0.181 ± 0.001</b>	<b>0.03 ± 0.001</b>	0.229 ± 0.005	0.112 ± 0.004
PRF-VA	0.185 ± 0.002	0.038 ± 0.002	<b>0.212 ± 0.006</b>	<b>0.092 ± 0.006</b>

three folds were used for model training, while the last fold was used for testing. Using probabilistic labels instead of hard labels improves the calibration error and the BCE loss of the RF and RF-Platt models trained on Other-3 assay data. Models for the Panel-1 assay do not show any improvements when incorporating aleatoric error. This result could be explained by the difference in standard deviations shown in Fig. 1[B], which are used for generating the probabilistic labels. Given that the measurement error of the Panel-1 assay is smaller compared to the Other-3 assay the normal distribution used for generating the probabilistic labels is narrower, resulting in probabilistic labels that are more similar to the hard labels, thus leading to similar results of RF and PRF models. The post-hoc calibration methods improve the BCE loss and ACE scores of Other-3 models, with RF-VA performing best in terms of both metrics, with a BCE and ACE of  $0.211 \pm 0.007$  and  $0.089 \pm 0.006$ , respectively. The results for the Panel-1 assay show that in terms of ACE the RF-Platt model performs slightly better than the uncalibrated RF model, while the PRF models did not improve after calibration. In general, the control compounds of the Panel assays exhibit smaller standard deviations than those of the Other assays, as illustrated in Fig. 1[B]. The results of the assays omitted from Table 1 reveal that using probabilistic labels generally leads to better BCE scores for Other assays. In contrast, such clear improvements can not be observed for Panel assays. This could be a result of the differences mentioned above in standard deviations of the control compounds between the assay categories. However, there are also exceptions from this trend, such as the Other-1 assay, which does not show improvements when including probabilistic labels, despite the large standard deviation of its control compound. Hence, we conclude that it is required to look at the model performance on the individual assay to find the best calibration method for that specific dataset. For all assays, the same model performs best in terms of BCE scores when comparing models trained with hard labels versus probabilistic labels. This is also true in terms of

the ACE results, except for the Other-1 assay, for which the RF model performs best for hard labels and the VA-calibrated model is best for probabilistic labels. However, the difference between PRF and PRF-VA is not significant. A more elaborate study is required to understand the effect of probabilistic labels on probability calibration in detail, which will be the object of our future research but is outside the scope of this abstract.

### 3.2 Probability Calibration Across Evolving Test Sets

Figure 2 shows the performance of five model repeats of different RF models across all ten assays and test sets in terms of ACE. The models were trained on one fold and then used for separately predicting three test folds representing subsequent time spans in the assay history. Test set 1 corresponds to the fold closest in time to the training fold, while test set 3 represents the fold furthest away. The ACE for test set 1 is the smallest across all models for the majority of assays as shown in Fig. 2, indicating that the models are better calibrated for compounds measured closer in time to the training fold. This pattern can also be observed in some assays when comparing test sets 2 and 3, however, the tendency is not as clearly visible as for test set 1. One of the reasons for the observed behavior could be a distribution shift in training and test data that increases as we progress in time, which is supported by a paper by Ovadia et al. [13], in which an increasing distribution shift was reported to impair probability calibration.



**Fig. 2. Model calibration over time.** The Adaptive Calibration Error (ACE) is shown for five model repeats across all assays. The models were trained on one training fold. Test Set 1 is closest in time to the training set, whereas Test Set 3 is furthest away.

## 4 Conclusion and Outlook

In this study, we showed that using probabilistic labels in combination with probability calibration approaches can improve uncertainty estimation in RF models. In addition, we present a comprehensive analysis of how model calibration changes over time using temporal splits of internal data from a pharmaceutical company. Based on these preliminary results, we will take further steps to understand model calibration in a temporal setting. Furthermore, we will extend our study to other model architectures, including multi-layer perceptrons (MLP), to investigate if the same conclusions can be drawn for other model types. Finally, we will explore uncertainty estimation methods to account for model uncertainty, including deep ensembles [7] and Monte-Carlo Dropout [2], to analyze if these approaches improve probability calibration.

**Acknowledgements.** Many thanks to Susanne Winiwarter of AstraZeneca in Gothenburg for her valuable advice during the data preparation. This study was partially funded by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions grant agreement “Advanced machine learning for Innovative Drug Discovery (AIDD)” No. 956832.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. MELLODDY-TUNER. <https://github.com/melloddy/MELLODDY-TUNER>
2. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (20–22 Jun 2016). <https://proceedings.mlr.press/v48/gal16.html>
3. Gruber, C., Schenk, P.O., Schierholz, M., Kreuter, F., Kauermann, G.: Sources of uncertainty in machine learning – a statisticians’ view (2023). <https://doi.org/10.48550/arXiv.2305.16703>
4. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, 06–11 Aug 2017, vol. 70, pp. 1321–1330. PMLR (2017). <https://proceedings.mlr.press/v70/guo17a.html>
5. Heyndrickx, W., et al.: Melloddy: Cross-pharma federated learning at unprecedented scale unlocks benefits in qsar without compromising proprietary information. *J. Chem. Inf. Model.* (2023). <https://doi.org/10.1021/acs.jcim.3c00799>
6. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.* **110**, 457–506 (2019). <https://doi.org/10.1007/s10994-021-05946-3>

7. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., et al. (eds.) *Advances in Neural Information Processing Systems*. vol. 30. Curran Associates, Inc. (2017), [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf)
8. Landrum, G.: RDKit: Open-source cheminformatics (2006). <https://doi.org/10.5281/zenodo.6961488>
9. Mervin, L., Trapotsi, M.A., Afzal, A., Barrett, I., Bender, A., Engkvist, O.: Probabilistic random forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty (2021). <https://doi.org/10.26434/chemrxiv.14544291>
10. Mervin, L.H., Afzal, A.M., Engkvist, O., Bender, A.: Comparison of scaling methods to obtain calibrated probabilities of activity for protein-ligand predictions. *J. Chem. Inf. Model.* **60**(10), 4546–4559 (2020). <https://doi.org/10.1021/acs.jcim.0c00476>, pMID: 32865408
11. Mervin, L.H., Johansson, S., Semenova, E., Giblin, K.A., Engkvist, O.: Uncertainty quantification in drug design. *Drug Discovery Today* **26**(2), 474–489 (2021). <https://doi.org/10.1016/j.drudis.2020.11.027>
12. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)*. <https://doi.org/10.48550/arXiv.1904.01685>
13. Ovadia, Y., et al.: Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift. In: Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc. (2019). [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf)
14. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **10**(3), 61–74 (1999)
15. Reis, I., Baron, D., Shahaf, S.: Probabilistic random forest: a machine learning algorithm for noisy data sets. *Astron. J.* **157**(1), 16 (2018). <https://doi.org/10.3847/1538-3881/aaf101>
16. Sheridan, R.P.: Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. Chem. Inf. Model.* **53**(4), 783–790 (2013). <https://doi.org/10.1021/ci400084k>
17. Toccaceli, P., Nouretdinov, I., Luo, Z., Vovk, V., Carlsson, L., Gammerman, A.: *Escape wp1-probabilistic prediction* (2016)
18. Vovk, V., Petej, I.: Venn-abers predictors (2014). <https://doi.org/10.48550/arXiv.1211.0025>



**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

