Towards Reliable Uncertainty Estimates for Drug Discovery: A Large-scale Temporal Study of Probability Calibration

Hannah Rosa Friesacher [1 2] **Emma Svensson** [1 3] **Adam Arany** [2] **Lewis Mervin** [4] **Ola Engkvist** [1 5]

Abstract

Quantifying the uncertainties associated with predictive models can facilitate optimal decision-making and accelerate workflows where time and resource efficiency are essential. Computational tools exist that estimate the predictive uncertainty, which is useful for assessing the costs and risks involved with deploying machine learning models. In drug discovery, these tools can provide valuable insights into the efficient allocation of resources by identifying promising experiments, thereby reducing the overall costs associated with the development of therapeutic agents. We address the pressing need for a comprehensive, large-scale temporal evaluation of probability calibration methods, specifically focusing on drug-target interactions. We investigate the performance of several calibration-free uncertainty estimation and post-hoc probability calibration methods. Furthermore, we systematically compare the effect of different training set sizes and shifts in active ratios on the capability of the uncertainty estimation methods.

1. Introduction

Uncertainty quantification is a powerful tool to increase the reliability of machine learning models and the confidence in deploying them to real-world applications (Apostolakis, 1990). Various sources can lead to uncertainty in the predictions obtained from machine learning models. 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 (Hüllermeier & Waegeman, 2019; Gruber et al., 2023). Available uncertainty quantification methods vary in their ability to capture all sources of uncertainty correctly.

When modeling classification problems, models typically give probability-like predictions that can be directly interpreted as an estimate of the confidence in the prediction. This is considered a calibration-free approach to uncertainty quantification. Similarly, ensemble-based approaches inspired by the Bayesian theorem to estimate the posterior distribution of predictions from a set of models are also calibration-free (Sheridan, 2012; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017). Ensembles enable the estimation of model uncertainty, by accounting for model variance, which increases when the model is overfitting or the test instance lies outside the domain of the training data. Previous work has identified 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 (Guo et al., 2017; Mervin et al., 2021a). Instead, various calibration methods have been developed for classification models which aim to obtain better uncertainty estimates by fitting a calibrating model to a separate dataset in a post-hoc manner (Platt, 1999; Vovk & Petej, 2014).

During early-stage drug discovery, a part of the vast chemical space is screened to identify promising molecular compounds as potential drugs (Hertzberg & Pope, 2000). The large scale and complexity of the screening, make it an ideal application for machine learning models with their high computational power and predictive abilities (Bleakley & Yamanishi, 2009). However, the long and costly process of the entire drug discovery pipeline, including years-long clinical trials, etc., means that reliability and confidence in the models are crucial for their deployment. Computational tools that estimate predictive uncertainties facilitate the assessment of costs and risk in the discovery and development pipeline (Mervin et al., 2021a). So far, the available uncertainty quantification methods have mostly been evaluated

[1] Molecular AI, Discovery Sciences, R&D, AstraZeneca Gothenburg, 431 83 Sweden [2] ESAT-STADIUS, KU Leuven, 3000 Belgium [3] ELLIS Unit Linz, Institute for Machine Learning, Johannes Kepler University Linz, 4040 Austria [4] Molecular AI, Discovery Sciences, R&D, AstraZeneca Cambridge, CB2 0AA UK [5] Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, 412 96 Sweden. Correspondence to: Rosa Friesacher <rosa.friesacher@kuleuven.be>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

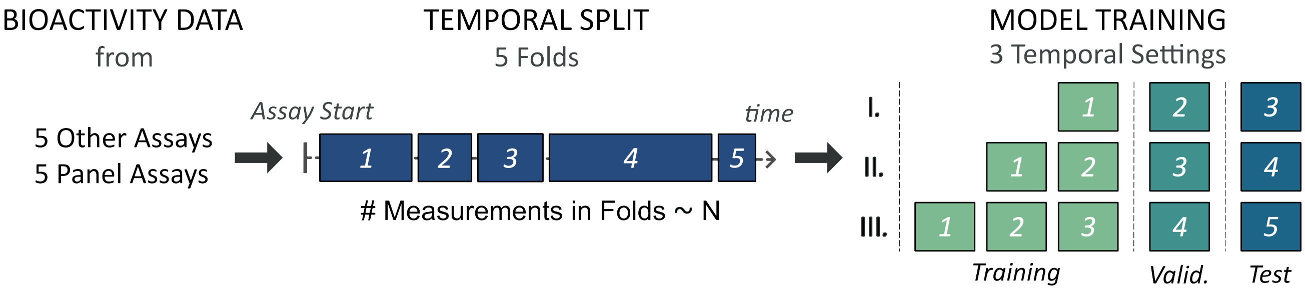

Figure 1. **Overview of the temporal split and model generation.** Five folds per assay were generated to create three temporal settings, each with increasing amounts of training (Training) data. The subsequent two folds were used for validation (Valid.) and testing (Test). The validation data also served as a calibration set used in post-hoc calibration approaches.

on public data that lack the information needed to draw realistic conclusions about how the methods perform over time in real-world pharmaceutical drug discovery projects (Sheridan, 2013).

In this work, we evaluate the performance of uncertainty quantification for single-task classification models trained on industry-scale assay data in a temporal analysis. A temporal splitting strategy enables model training and calibration on older data and evaluation of test predictions on data from subsequent experiments. Similarly to a cluster-based splitting strategy, a temporal split is more challenging than a standard random split. However, in contrast to a cluster-based splitting strategy, another advantage of the temporal split is that it more accurately simulates the realistic drug discovery pipeline in pharmaceutical companies (Sheridan, 2013). Our analysis compares the predictive performance and calibration of estimated uncertainties by four calibration-free approaches, with and without two post-hoc calibration methods.

2. Methods

We extracted internal data from a pharmaceutical company belonging to ten assays, individually used as single-task classification datasets. The assays were categorized as proposed by Heyndrickx et al. (2023) into "Panel" and "Other", and labelled according to size. The "Panel" category comprises cross-project assays to detect undesired effects of compounds that hit unintended targets. The "Other" category includes project-specific assays from activity screens to identify substances that are active on a target of interest. The selected assays were chosen to be representative, exhibiting various assay sizes and active ratios.

Apart from absolute observed measurements, the original data included censored labels indicating if a measurement was greater or smaller than a given value. However, all measurements were transformed to binary labels active or inactive. To do so, the negative logarithm of the compound's concentration needed for inhibiting half of a target's activity (pIC50) or for triggering half of the maximum response (pEC50) was used. We applied a fixed threshold of 6 pIC50 or pEC50, corresponding to $1\mu M$ concentration. During this processing step, censored labels indicating a measurement that could not be classified according to the threshold were removed. Duplicated measurements of a compound were aggregated using the median, prioritizing observed measurements over censored ones. Finally, all molecular compounds were encoded by first standardizing their SMILES strings (Weininger, 1988) using the MELLODDY-TUNER (Mel) package and then generating extended connectivity fingerprints (ECFPs) of hashing length 1024 and radius 2 with RDKit (Landrum, 2006).

Temporal split. For each assay, we split the data into five, roughly equally sized, folds using the date of each measurement. These folds were then used to set up three experimental settings, using one, two, or three folds for training the machine learning models. In each case, the first subsequent fold was used for validation, including model selection, and calibration where applicable. We only evaluated each setting on the first fold following the validation set for consistency between test sets. However, all remaining folds could in principle be used. Fig. 1 illustrates the temporal splitting strategy.

Considering all assays and settings, 30 individual training datasets were used throughout this work. They are labeled Category-Assay Number [#Training Folds]. Naturally, the size of the training sets varies among all experiments as shown in the top panel of Fig. 2. Additionally, we quantify the ratio of active compounds in each training dataset in the middle plot and the shifts in label distribution in the lower panel. The latter was evaluated by comparing the active ratio in the combined training and validation set with the active ratio in the test set.

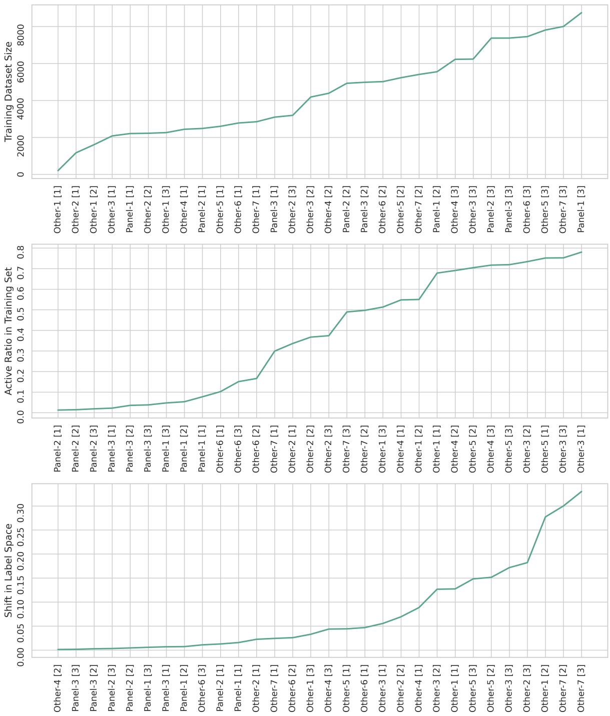

Figure 2. **Overview of assay data.** For all assays and temporal settings. (Upper) shows the size of the training sets. (Middle) illustrates how the ratio of active compounds changes across training sets. (Lower) shows the difference in active ratios between the combined training and validation folds and the test fold, i.e. the shift in label space.

2.1. Model Generation

All models used in this work stem from either a Random Forest (RF) or a fully connected neural network (MLP). Both approaches are commonly used in research addressing uncertainty estimation in machine learning (Dutschmann et al., 2023; Mervin et al., 2020; 2021b). Furthermore, uncertainty estimation comes quite naturally with the ensemble-like nature of RFs, and MLPs can be easily combined with the ECFP fingerprint representation. Note, that more sophisticated options of molecular representations and model architectures exist, like graph neural networks for molecular graph representations or language models for SMILES representations. However, since our study aims to gain insight into uncertainty estimation in bioactivity predicting models rather than finding the best model, or comparing molecular representations, we opted for the simple ECFP representation. Because of these reasons and the fact that we were required to restrict the methods included in this paper due to computational limitations, we considered RF and MLP models a good selection to understand uncertainty quantification in a temporal setting.

The RF models were generated using scikit-learn (Pedregosa et al., 2011), optimizing the maximum depth of the trees and the required number of estimators of each assay and temporal setting individually based on the validation loss. Probability-like outputs were generated from the ratio of decision trees in the RF that classified a test instance as active. The MLP models were trained using PyTorch (Paszke et al., 2019) with the binary cross-entropy (BCE) loss function. Due to the low active ratio in 9 of the Panel assay datasets, a weighted BCE loss was used for all datasets from the Panel category. Similarly, the model selection including early stopping was optimized using the validation loss for every assay and temporal setting. The network architecture was optimized for the number of hidden units, number of hidden layers, and dropout rate. Additionally, the learning rate and scaling factor of a ReduceOnPlateu learning rate scheduler were also optimized. Probability-like scores were obtained by applying a sigmoid function to the output of the MLP. The model selection for the two base estimators was performed using an exhaustive grid search. The exact parameter space search is detailed in Appendix Table 2. [1]

Calibration-free uncertainty quantification. We explored two calibration-free approaches to improve the uncertainty estimate of the base MLP model. Both methods utilize model variance by creating an ensemble of predictions. First, we generated 25 individually trained MLPs from randomly initialized weight distributions as proposed by (Lakshminarayanan et al., 2017) as a Deep Ensemble (MLPE). Second, we applied dropout during inference of a single trained MLP and drew 400 instances of predictions proposed by (Gal & Ghahramani, 2016) as MC-dropout (MLPMC). In both cases, the individual probability-like scores of each prediction were aggregated by taking the average as the final prediction of the improved models.

Post-hoc probability calibration. Two post-hoc probability calibration techniques were fitted to each model using the validation set, namely Platt scaling (Platt, 1999) and Venn-ABERS (VA) predictors (Vovk & Petej, 2014). Platt scaling fits a logistic regression to the classification scores to counteract over- or underfitted uncertainty estimations (Platt, 1999). For calibration with VA predictors, two isotonic regression functions were trained on the validation set and a given test instance (Vovk & Petej, 2014). The two isotonic regression functions represent the hypothesis that the test instance is active versus inactive. As such, the probabilities obtained from the isotonic regression functions correspond to a lower and an upper bound on the estimated probability. Finally, these bounds were condensed to a point estimate, as proposed by Toccaceli et al. (2016).

[1] A Python package containing our proposed method will be made available open-source upon acceptance.

3. Results

To thoroughly evaluate how the different approaches perform over time, we make ten repetitions of each experiment and provide averaged results paired with their standard deviation. The standard deviations are also used to perform unpaired two-sided t-tests, to check the statistical significance of the top-performing models. We test the best performance against all other scores, presenting all scores without significant ($p > 0.05$) differences as the top-performing methods.

First, the predictive accuracy of each method is evaluated and compared in terms of the area under the receiver operating characteristic curve (AUC). As the post-hoc calibration methods do not significantly impact the ordering of predictions, these methods have been omitted from the first experiment. Next, the calibration of the predicted probabilities is evaluated using the BCE and the adaptive calibration error (ACE) (Nixon et al., 2019). The ACE estimates the true calibration error by discretizing the predicted probabilities into equally sized bins based on the number of predictions in each bin. It then takes the weighted average of the errors over all bins. The equally sized bins distinguish the ACE from the otherwise more commonly used expected calibration error (ECE). Here, the ACE is preferred over ECE as it is more robust towards skewed distributions of the predictions. We independently compare the BCE and ACE performances of the methods on all assays and temporal settings in terms of training set size and distribution shift.

3.1. Comparison of Model Accuracy

Table 1 shows the AUC scores across all assay datasets ordered by ascending size of the training set. Only the third temporal setting is shown for readability, but the full results for all settings can be seen in Appendix Table 3. We see no clear pattern in AUC performance regarding increasing training data size. For 5 out of the 10 assays (18 out of 30 all datasets), the RF model is best at ranking the test instances and significantly outperforms the more flexible and complex MLP approaches. In the remaining cases, the MLPE model is always among the best-performing methods in terms of AUC.

However, a single significantly best method can often not be identified due to the lack of significant differences between the methods. In 10 out of all 30 datasets, at least 3 models have comparable AUC. This means that despite substantial differences in the calibrating properties of a model, as seen in the following section, all models can have comparable ranking abilities. Roth & Bajorath (2024) found similar results, reporting that various model architectures can produce accurate predictions, despite showing large differences in probability calibration.

While the increase in training dataset size does not lead to an evident trend in AUC, comparing the active ratio in the training datasets and the AUC performance of the respective models reveals a clear pattern. Fig. 2 shows that there is a noticeable difference in terms of active ratio between the training datasets. Models trained on datasets with small active ratios exhibit lower AUC scores. Table 3 in the Appendix illustrates that in particular, the models trained on Other-1 [3], Panel-1 [2], and all three training sets of the Panel-2 assay fail to achieve accurate predictions, yielding an AUC score lower than 0.65. Four of these training sets are among the five datasets with the smallest positive ratio. Thus, the poor model performance is likely due to an insufficient number of active compounds available in the training sets, resulting in models that primarily predict compounds as inactive. Because of the poor performance, we exclude the models trained in these five mentioned settings, from the model calibration analysis addressed in the following sections.

3.2. Effect of Training Set Size on Model Calibration

Next, we compare the calibrating properties of the methods by evaluating the BCE and ACE scores for all datasets. The results are shown as heatmaps in Fig. 3 for BCE and Fig. 4 for ACE. All numerical results of this experiment can be found in Section B.2 of the Appendix. The rows are sorted by ascending training set size and white rectangles indicate the best-performing methods per row.

The heatmaps show that both scores tend to decrease with increasing training dataset size, indicated by the color gradient from lighter green in the first rows to dark blue for the larger datasets. The pattern is much more apparent in the BCE scores, but can also be detected in the ACE heatmap. This general trend might stem from the model's overfitting on the smaller training sets, which manifests as probabilistic error rather than impaired ranking which would have been detectable in the AUC scores (Guo et al., 2017).

Some outliers exist, such as the Panel-3 [1] and Panel-1 [1] datasets, for which all methods achieve better scores than they do for other datasets of comparable sizes. The AUC scores in Table 3 in the Appendix indicate that for both these datasets the RF model performs much better than the MLP methods. The same trend can be observed in terms of BCE (Fig. 3) and ACE (Fig. 4), and only post-hoc calibration of the MLP models yields results comparable to the RF performance. In general, the ACE scores of the individual methods reveal that RF models tend to be better calibrated than MLP approaches for smaller training set sizes, and vice versa for larger training sets 4.

Importantly, the post-hoc calibrated versions of the MLP, MLPE, and MLPMC models exhibit better ACE performance than their calibration-free counterparts for large

Table 1. **Overview of AUC scores across datasets and methods.** Averages of 10 model repeats are shown. The best-performing method as well as the methods that are statistically indistinguishable from the best one are marked in bold.

Dataset	RF	MLP	MLPE	MLPMC
Other-1 [3]	**0.6443±0.0048**	**0.6371±0.0656**	**0.6457±0.0038**	**0.6487±0.0619**
Other-2 [3]	**0.8235±0.0078**	0.7345±0.0373	0.7758±0.0062	0.7344±0.0415
Other-4 [3]	**0.9584±0.0006**	0.9562±0.0029	0.957±0.0005	0.9563±0.0029
Other-3 [3]	0.7451±0.0096	**0.7835±0.0055**	**0.7872±0.0008**	**0.7841±0.0054**
Panel-2 [3]	0.4725±0.0105	**0.5706±0.0135**	**0.5688±0.0016**	**0.57±0.0137**
Panel-3 [3]	**0.7443±0.0051**	0.6069±0.0134	0.6183±0.0051	0.607±0.0133
Other-6 [3]	**0.8995±0.0022**	0.8527±0.0167	0.8682±0.003	0.8527±0.0167
Other-5 [3]	0.6739±0.0057	**0.7639±0.017**	**0.7657±0.0015**	**0.7619±0.0171**
Other-7 [3]	**0.8379±0.0027**	0.7289±0.0313	0.7584±0.0067	0.729±0.029
Panel-1 [3]	**0.6633±0.0098**	0.6454±0.0175	**0.661±0.0012**	0.6454±0.0175

datasets. This observation is likely an effect of the calibrating properties of model ensembling, which tends to make the models more underconfident. Hence, calibration-free ensemble-based models have been reported to not always lead to better-calibrated models (Rahaman & Thiery, 2021). Considering the large amount of training data used for these models, the baseline MLP model might not suffer from overconfidence as much as, for example, models trained on middle-sized datasets. Hence, the ensemble-based models MLPE and MLPMC might to a greater extent lead to underconfident and poorly calibrated predictions that are corrected in the post-hoc calibration step.

Note, that ACE decouples predictive performance from probability calibration making it an improper scoring rule (Gneiting & Raftery, 2007). This means that the best-calibrated model according to the ACE might not correspond to the most accurate predictive model. If a proper scoring rule is preferred, one should focus on the BCE score for the comparison given that both calibration and accuracy of a model affect BCE. As a result, the model performance in terms of BCE across the datasets lies somewhat between the scores AUC and ACE.

3.3. Effects of Label Shifts on Model Calibration

In the final experiment, we analyze the effect of distribution shifts in label space on model performance and probability calibration within the framework of the temporal split. Fig. 5 and Fig. 6 illustrate the same results as in Fig. 3 and Fig. 4 but sorted differently. In these heatmaps, the rows correspond to the datasets ordered by ascending label distribution shift between the combined training and validation folds versus the test fold.

What stands out in this version of the analysis is a steady increase in BCE with increasing label shifts, indicated by the color gradients of the heatmap. In more detail, the models for datasets with a distribution shift smaller than the one observed in Other-2 [1] perform comparably well, as indicated by the dark blue colored regions. There is a noticeable change in color in the row of the Other-2 [1] dataset, which might result from the combination of increasing label shift and a small training set size of this specific dataset as shown in Fig. 2. However, most of the rows below Other-2 [1] show similarly high BCE, with some exceptions including Other- [3], Other-3 [3], and Other-7 [3]. The low BCE for these assays might be a result of their comparably large training sets (Fig. 2). More training data help to counteract model overfitting, resulting in the higher AUC score seen for Other-4 [3] (Table 1) which in turn contributes positively to the BCE.

A similar trend as the one found in the BCE analysis can be observed in the ACE scores, albeit less prominent. A potential explanation for the decline in model performance with increasing label shift could be overconfidence due to model overfitting manifesting in poorly calibrated predictions with higher BCE and ACE scores. Interestingly, similar observations were reported for shifts in the descriptor space (Ovadia et al., 2019). The authors showed that with an increasing distribution shift, neural networks fail to produce reliable uncertainty estimates due to increasing model overconfidence.

Importantly, the tendency described above can be found across all methods indicating that a shift in label space generally impairs model performance. Recall, that the white rectangles in Fig. 5 and Fig. 6 indicate the best-performing models on each dataset. The heatmaps show a slight trend that RF models outperform MLP approaches on datasets with smaller label shifts, while the post-hoc calibrated MLPs seem to perform better on datasets with large shifts. However, this tendency does not apply to all datasets.

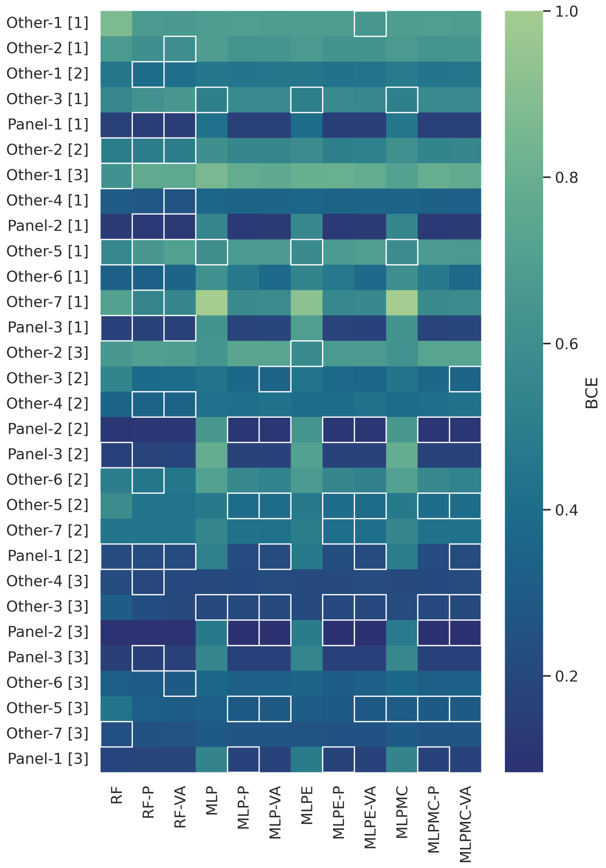

Figure 3. **BCE across methods and datasets.** The datasets are sorted by ascending size of training data. Averages of 10 model repeats are shown. The white rectangles mark the significantly best score and those results that are statistically indifferent from the best one.

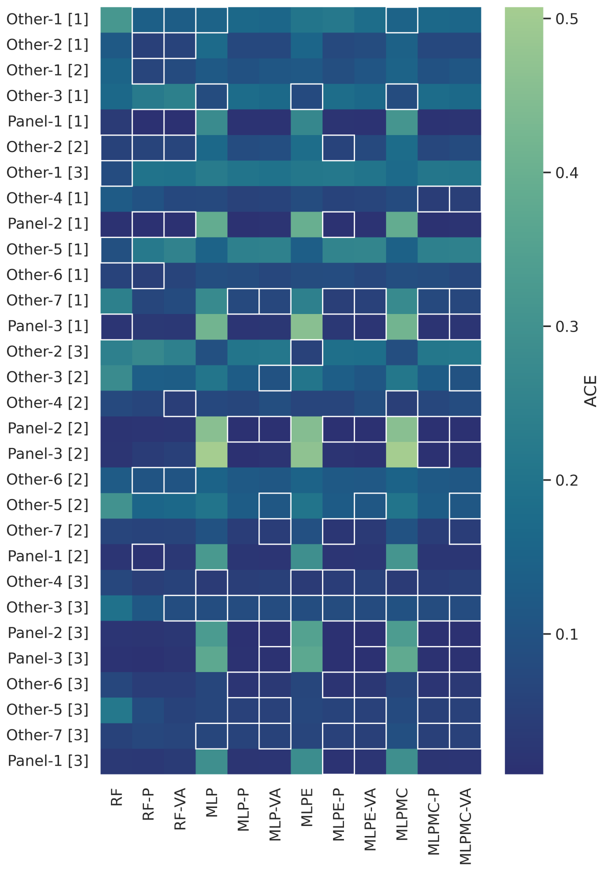

Figure 4. **ACE across methods and datasets.** The datasets are sorted by ascending size of training data. Averages of 10 model repeats are shown. The white rectangles mark the significantly best score and those results that are statistically indifferent from the best one.

4. Conclusion and Outlook

In this study, we assessed the effect of training set sizes and shifts in label distribution on the probability calibration of uncertainty quantification methods using a temporal split. We found that model calibration improves with increasing training data, a potential result from decreased model overfitting when training on larger datasets. Interestingly, RF models could often match or even outperform the more complex MLP models, especially for smaller datasets. For larger datasets post-hoc calibrated versions of the MLP, MLPE, and MLPMC models exhibited lower calibration errors.

In addition, we examined the impact of distribution shifts in label space on model calibration. Note, that a large distribution shift was observed for some assays as a result of the temporal split, indicating a violation of the i.i.d. assumption. Overall, increasingly large distribution shifts in label space impaired model calibration, but a counteracted effect could be seen by larger training set sizes.

Based on these preliminary results we aim to further explore the changes in model calibration throughout time in the proposed temporal setting. We will extend our study to other assay categories, to investigate how label shifts affect the probability calibration of models trained on even larger amounts of data. These categories will include ADME assays, which measure the absorption, distribution, metabolism, and excretion of a substance. ADME assays typically comprise large amounts of bioactivity data as they evaluate particularly important properties of a drug candidate.

Furthermore, we will explore if incorporating aleatoric uncertainties in the form of probabilistic labels (Reis et al., 2018; Mervin et al., 2021b) can enhance the quality of uncertainty estimates. In summary, our work provides impor-

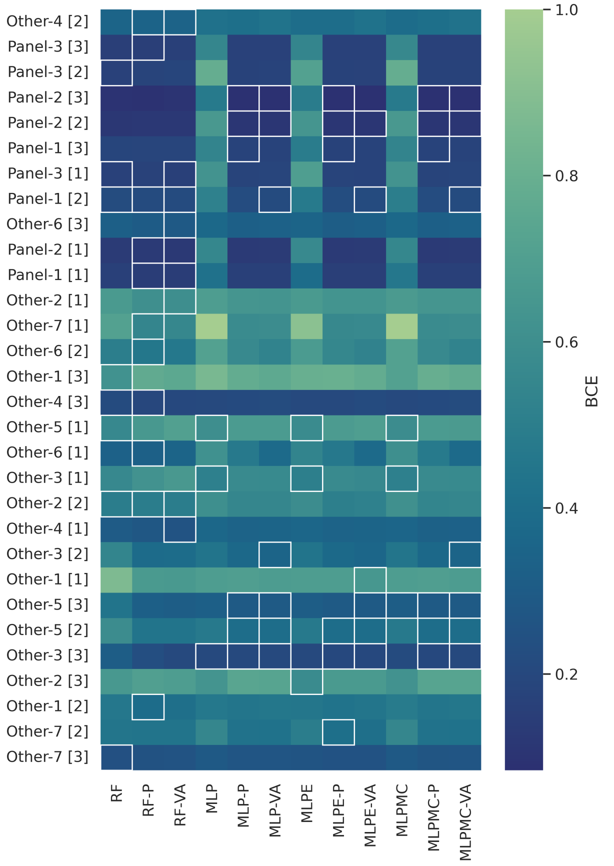

Figure 5. **BCE across methods and datasets.** The datasets are sorted according to ascending label distribution shifts. Averages of 10 model repeats are shown. The white rectangles mark the significantly best score and those results that are statistically indifferent from the best one.

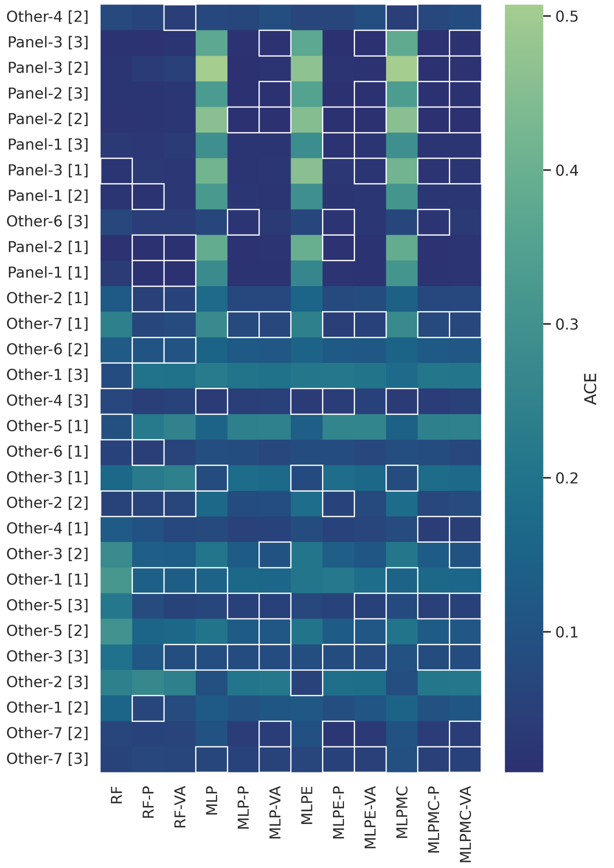

Figure 6. **ACE across methods and datasets.** The datasets are sorted according to ascending label distribution shifts. Averages of 10 model repeats are shown. The white rectangles mark the significantly best score and those results that are statistically indifferent from the best one.

tant insight into the calibrating abilities of machine learning models on real pharmaceutical data, which is essential to achieving reliable uncertainty estimates for an efficient drug discovery process.

Acknowledgements

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.

References

MELLODDY-TUNER. URL `https://github.com/melloddy/MELLODDY-TUNER`.

Apostolakis, G. The Concept of Probability in Safety Assessments of Technological Systems. *Science*, 250(4986): 1359–1364, 1990.

Bleakley, K. and Yamanishi, Y. Supervised Prediction of Drug–Target Interactions Using Bipartite Local Models. *Bioinform.*, 25(18):2397–2403, 2009.

Dutschmann, T.-M., Kinzel, L., Ter Laak, A., and Baumann, K. Large-scale evaluation of k-fold cross-validation ensembles for uncertainty estimation. *Journal of Cheminformatics*, 15(1):49, 2023.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New

York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Gruber, C., Schenk, P. O., Schierholz, M., Kreuter, F., and Kauermann, G. Sources of uncertainty in machine learning – a statisticians' view, 2023.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Hertzberg, R. P. and Pope, A. J. High-Throughput Screening: New Technology for the 21st Century. *Curr. Opin. Chem. Biol.*, 4(4):445–451, 2000.

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. doi: 10.1021/acs.jcim.3c00799.

Hüllermeier, E. and Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110:457 – 506, 2019. doi: 10.1007/s10994-021-05946-3.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Landrum, G. Rdkit: Open-source cheminformatics, 2006.

Mervin, L. H., Afzal, A. M., Engkvist, O., and Bender, A. Comparison of scaling methods to obtain calibrated probabilities of activity for protein–ligand predictions. *Journal of Chemical Information and Modeling*, 60(10): 4546–4559, 2020. doi: 10.1021/acs.jcim.0c00476. PMID: 32865408.

Mervin, L. H., Johansson, S., Semenova, E., Giblin, K. A., and Engkvist, O. Uncertainty quantification in drug design. *Drug Discovery Today*, 26(2):474–489, 2021a. ISSN 1359-6446. doi: 10.1016/j.drudis.2020.11.027.

Mervin, L. H., Trapotsi, M.-A., Afzal, A. M., Barrett, I. P., Bender, A., and Engkvist, O. Probabilistic random forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty. *Journal of Cheminformatics*, 13:1–17, 2021b.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. doi: 10.48550/arXiv.1904.01685.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. 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., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf.

Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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., and Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10, 06 1999.

Rahaman, R. and Thiery, A. Uncertainty quantification and deep ensembles. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 20063–20075. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a70dc40477bc2adceef4d2c90f47eb82-Paper.pdf.

Reis, I., Baron, D., and Shahaf, S. Probabilistic random forest: A machine learning algorithm for noisy data sets. *The Astronomical Journal*, 157(1):16, dec 2018. doi: 10.3847/1538-3881/aaf101.

Roth, J. P. and Bajorath, J. Relationship between prediction accuracy and uncertainty in compound potency prediction using deep neural networks and control models. *Scientific Reports*, 14(1):6536, 2024.

Sheridan, R. P. Three Useful Dimensions for Domain Applicability in QSAR Models Using Random Forest. *J. Chem. Inf. Model.*, 52(3):814–823, 2012.

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. doi: 10.1021/ci400084k.

Toccaceli, P., Nouretdinov, I., Luo, Z., Vovk, V., Carlsson, L., and Gammerman, A. Excape wp1-probabilistic prediction, 2016.

Vovk, V. and Petej, I. Venn-abers predictors, 2014.

Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, 1988.

A. Hyperparameter Search

In Table 2 the hyperparameters explored in the model selection for the RF model and the base MLP are presented. The architecture of the MLP comprises a given number of hidden layers with either a fixed number of units or a steadily decreasing number of units determined by the decreasing parameter. Furthermore, dropout and the ReLU activation function are applied to each hidden layer. An exhaustive grid search was used to find the optimal hyperparameters for every assay and temporal setting based on the validation BCE loss.

Table 2. **Model selection.** Considered hyperparameter space in the model selection of the RF and MLP models.

Base Model	Hyperparameter	Explored space
RF	n_estimators	{50, 100, 250, 500, 1000, 1500}
	max_deapth	{5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 1000000}
MLP	Optimizer	{Adam}
	Learning rate	{0.00005, 0.0001, 0.0005, 0.001}
	Weight decay	{0.0005}
	Scheduler	{ReduceOnPlateu}
	Scheduler Factor	{0.1, 0.5}
	Scheduler Patience	{50}
	Batch size	{64}
	Number of hidden layers	{2, 3, 4}
	Hidden dimension	{64, 128, 256, 512}
	Decreasing dimension	{False, True}
	Dropout	{0, 0.25, 0.5, 0.75}

B. Numerical Results

The following section gives the full numerical results for all experiments, complementary to the ones provided in Section 3.

B.1. Additional Model Accuracy Results

Table 3 presents the model accuracy in terms of AUC for all available datasets and temporal settings.

Table 3. **Overview of AUC scores across datasets and methods.** Averages of 10 model repeats are shown. The best-performing method as well as the methods that are statistically indistinguishable from the best one are marked in bold.

Dataset	RF	MLP	MLPE	MLPMC
Other-1 [1]	0.6869±0.018	**0.614±0.1249**	**0.7266±0.0053**	**0.614±0.1249**
Other-2 [1]	**0.7262±0.0084**	0.6821±0.0208	0.6894±0.002	0.6616±0.0617
Other-1 [2]	**0.8411±0.0015**	0.7886±0.0116	0.7962±0.0007	0.7902±0.0111
Other-3 [1]	0.6253±0.0208	**0.6576±0.0198**	**0.6634±0.0054**	**0.6576±0.0198**
Panel-1 [1]	**0.7407±0.0121**	0.6141±0.0702	0.6583±0.0013	0.6086±0.0775
Other-2 [2]	**0.8147±0.0068**	0.7266±0.1034	0.7906±0.0012	0.7291±0.0959
Other-1 [3]	**0.6443±0.0048**	**0.6371±0.0656**	**0.6457±0.0038**	**0.6487±0.0619**
Other-4 [1]	**0.9505±0.002**	0.8971±0.0118	0.8979±0.0006	0.9001±0.0089
Panel-2 [1]	**0.6083±0.0084**	0.4781±0.0086	0.4811±0.0047	0.4781±0.0086
Other-5 [1]	**0.6839±0.0101**	**0.6831±0.0192**	**0.6935±0.0025**	**0.6859±0.0175**
Other-6 [1]	**0.7237±0.0106**	0.6679±0.0104	0.6813±0.0058	0.6685±0.0104
Other-7 [1]	**0.8129±0.0041**	0.779±0.0109	0.7922±0.0023	0.779±0.0109
Panel-3 [1]	**0.7807±0.0198**	0.4622±0.033	0.46±0.0049	0.475±0.0422
Other-2 [3]	**0.8235±0.0078**	0.7345±0.0373	0.7758±0.0062	0.7344±0.0415
Other-3 [2]	0.7503±0.0049	**0.7682±0.0084**	**0.7739±0.0033**	**0.7673±0.0084**
Other-4 [2]	**0.9029±0.0012**	0.8763±0.0096	0.8812±0.0018	0.8773±0.0094
Panel-2 [2]	0.5646±0.0177	**0.5775±0.0315**	**0.6006±0.0128**	**0.5775±0.0315**
Panel-3 [2]	**0.6504±0.008**	0.4752±0.0669	0.4981±0.0067	0.4801±0.0327
Other-6 [2]	**0.7466±0.0128**	0.6123±0.044	0.6508±0.0066	0.6123±0.044
Other-5 [2]	**0.6924±0.0149**	**0.6977±0.0134**	**0.705±0.003**	**0.6977±0.0134**
Other-7 [2]	0.7271±0.0079	0.7458±0.0155	**0.7686±0.001**	0.7458±0.0155
Panel-1 [2]	**0.6317±0.0171**	0.5637±0.0134	0.5663±0.001	0.5619±0.0148
Other-4 [3]	**0.9584±0.0006**	0.9562±0.0029	0.957±0.0005	0.9563±0.0029
Other-3 [3]	0.7451±0.0096	**0.7835±0.0055**	**0.7872±0.0008**	**0.7841±0.0054**
Panel-2 [3]	0.4725±0.0105	**0.5706±0.0135**	**0.5688±0.0016**	**0.57±0.0137**
Panel-3 [3]	**0.7443±0.0051**	0.6069±0.0134	0.6183±0.0051	0.607±0.0133
Other-6 [3]	**0.8995±0.0022**	0.8527±0.0167	0.8682±0.003	0.8527±0.0167
Other-5 [3]	0.6739±0.0057	**0.7639±0.017**	**0.7657±0.0015**	**0.7619±0.0171**
Other-7 [3]	**0.8379±0.0027**	0.7289±0.0313	0.7584±0.0067	0.729±0.029
Panel-1 [3]	**0.6633±0.0098**	0.6454±0.0175	**0.661±0.0012**	0.6454±0.0175

B.2. Additional Model Calibration Results

In the following tables, the full numerical results are provided, as a complement to the heatmaps in Fig. 3,4,5,6. Table B.2 presents the results in terms of BCE for the RF and MLP models with and without post-hoc calibration, while Table B.2 presents the BCE results for the MLPE and MLPMC models. The best-performing models from both Tables B.2 and B.2 are marked in bold for each dataset and temporal setting.

Table 4. **Overview of BCE scores across datasets and RF and MLP methods.** Averages of 10 model repeats are shown. The best-performing method as well as the methods that are statistically indistinguishable from the best one are marked in bold, including also the models in Table B.2.

Dataset	RF	RF-P	RF-VA	MLP	MLP-P	MLP-VA
Other-1 [1]	0.8722±0.0077	0.6654±0.0069	0.6607±0.0066	0.6868±0.0174	0.6925±0.0189	0.6807±0.024
Other-2 [1]	0.6645±0.0156	0.6056±0.0084	**0.5968±0.0066**	0.6842±0.028	0.636±0.0266	0.6324±0.011
Other-1 [2]	0.4586±0.0009	**0.3945±0.0014**	0.4116±0.0028	0.46±0.0306	0.4446±0.0098	0.4533±0.0125
Other-3 [1]	0.561±0.0149	0.6225±0.0297	0.6591±0.0325	0.5105±0.0128	**0.5728±0.0097**	0.5676±0.0125
Panel-1 [1]	0.1556±0.0011	**0.1439±0.0013**	**0.1445±0.0021**	0.4228±0.1345	0.1586±0.0008	0.1565±0.0021
Other-2 [2]	**0.4939±0.0051**	**0.495±0.0041**	**0.4924±0.006**	0.6003±0.06	0.5483±0.0435	0.5459±0.0395
Other-1 [3]	**0.62±0.0031**	0.7706±0.0063	0.7515±0.0092	0.8591±0.0716	0.7799±0.0251	0.7584±0.0245
Other-4 [1]	0.2964±0.0039	0.2798±0.0072	**0.2561±0.0056**	0.3665±0.0391	0.3394±0.018	0.3486±0.0216
Panel-2 [1]	0.1328±0.0049	**0.1269±0.0025**	**0.1261±0.0011**	0.5514±0.0932	0.1293±0.0006	0.1335±0.0011
Other-5 [1]	**0.5611±0.024**	0.656±0.0278	0.7051±0.0233	**0.5991±0.0624**	0.6719±0.0217	0.6717±0.02
Other-6 [1]	0.3315±0.0028	**0.3246±0.0038**	0.3459±0.0036	0.6119±0.0616	0.4684±0.0118	0.3781±0.0035
Other-7 [1]	0.7094±0.0083	**0.5428±0.0056**	0.5529±0.0071	1.1537±0.253	0.5804±0.0176	0.5851±0.0154
Panel-3 [1]	**0.1654±0.0117**	0.1665±0.0057	**0.162±0.0021**	0.6274±0.086	0.1779±0.0011	0.1808±0.0041
Other-2 [3]	0.6548±0.0098	0.7023±0.0116	0.6817±0.0181	0.6346±0.0347	0.7355±0.0558	0.7309±0.071
Other-3 [2]	0.5389±0.0036	0.3875±0.0025	0.3895±0.003	0.4401±0.0213	0.3692±0.0052	**0.3475±0.0065**
Other-4 [2]	0.3492±0.0014	**0.343±0.0018**	**0.3416±0.0017**	0.423±0.0186	0.4077±0.0239	0.4325±0.0262
Panel-2 [2]	0.1155±0.0018	0.1173±0.0021	0.1188±0.0033	0.6587±0.0653	**0.1108±0.0005**	**0.1111±0.0009**
Panel-3 [2]	**0.1643±0.0012**	0.1781±0.0022	0.1863±0.0045	0.7883±0.0578	0.1703±0.0	0.1735±0.0035
Other-6 [2]	0.4967±0.0056	**0.4551±0.0074**	0.4623±0.0062	0.7114±0.0437	0.5646±0.0135	0.5306±0.0196
Other-5 [2]	0.587±0.011	0.4361±0.0077	0.4412±0.0102	0.4674±0.0086	**0.4008±0.006**	**0.3989±0.0084**
Other-7 [2]	0.4486±0.0044	0.4404±0.0036	0.4417±0.0063	0.5484±0.0894	0.4281±0.0135	0.4266±0.0156
Panel-1 [2]	**0.2195±0.0112**	**0.213±0.0077**	**0.2161±0.0037**	0.5173±0.0783	0.2228±0.005	**0.2155±0.0022**
Other-4 [3]	0.2136±0.001	**0.1966±0.0013**	0.2003±0.0018	0.2065±0.0116	0.2122±0.0065	0.2166±0.0061
Other-3 [3]	0.3116±0.0082	0.2331±0.0057	0.2065±0.007	**0.2084±0.0099**	**0.2024±0.0045**	**0.2035±0.0043**
Panel-2 [3]	0.0955±0.0012	0.0946±0.0015	0.0985±0.003	0.4718±0.1242	**0.0842±0.0006**	**0.0844±0.0009**
Panel-3 [3]	0.1543±0.0008	**0.1531±0.0007**	0.158±0.0024	0.5478±0.0285	0.1659±0.0013	0.1652±0.0013
Other-6 [3]	0.3199±0.0033	0.2996±0.0024	**0.2887±0.0024**	0.3686±0.0373	0.3218±0.0129	0.331±0.0106
Other-5 [3]	0.4381±0.003	0.3213±0.0015	0.3086±0.0026	0.3171±0.0232	**0.2948±0.0127**	**0.293±0.0124**
Other-7 [3]	**0.242±0.0012**	0.2514±0.0011	0.257±0.0018	0.2848±0.0203	0.2668±0.0095	0.2687±0.0133
Panel-1 [3]	0.1852±0.0039	0.1812±0.003	0.1842±0.0017	0.5353±0.173	**0.1747±0.0022**	0.1765±0.0024

Table 5. **Overview of BCE scores across datasets and MLPE and MLPMC models.** Averages of 10 model repeats are shown. The best-performing method as well as the methods that are statistically indistinguishable from the best one are marked in bold, including also the models in Table B.2.

Dataset	MLPE	MLPE-P	MLPE-VA	MLPMC	MLPMC-P	MLPMC-VA
Other-1 [1]	0.6847±0.0021	0.6882±0.0048	**0.6505±0.0059**	0.6868±0.0174	0.6925±0.0189	0.6807±0.024
Other-2 [1]	0.6661±0.0074	0.6282±0.0051	0.6294±0.0016	0.6692±0.0322	0.6385±0.0219	0.6383±0.0183
Other-1 [2]	**0.4523±0.0018**	0.4335±0.0009	0.4472±0.0017	**0.4808±0.0339**	0.4456±0.0095	0.4548±0.0131
Other-3 [1]	0.5038±0.0037	0.5707±0.002	0.5632±0.0037	0.5105±0.0128	0.5728±0.0097	0.5676±0.0125
Panel-1 [1]	0.3927±0.0132	0.1553±0.0002	0.1537±0.0003	0.4592±0.1107	0.157±0.0009	0.1565±0.0022
Other-2 [2]	0.5861±0.0032	0.5052±0.0008	0.5187±0.0024	0.6092±0.053	0.5416±0.0456	0.5476±0.0417
Other-1 [3]	0.7993±0.009	0.8024±0.005	0.7783±0.005	0.7119±0.0294	0.7926±0.032	0.7688±0.0253
Other-4 [1]	0.3587±0.0014	0.34±0.0004	0.3479±0.0017	0.3529±0.0312	0.328±0.0131	0.3319±0.0137
Panel-2 [1]	0.5639±0.012	0.1293±0.0001	0.1332±0.0003	0.5514±0.0932	0.1293±0.0006	0.1335±0.0011
Other-5 [1]	**0.5743±0.0078**	0.6779±0.0021	0.6895±0.0029	**0.588±0.0565**	0.6724±0.0227	0.6705±0.0205
Other-6 [1]	0.5337±0.0195	0.4663±0.005	0.3777±0.0016	0.5995±0.0589	0.4706±0.0112	0.3782±0.0038
Other-7 [1]	0.9157±0.0338	0.5564±0.0037	0.5669±0.0035	1.1537±0.253	0.5804±0.0176	0.5851±0.0154
Panel-3 [1]	0.6922±0.0064	0.1776±0.0003	0.1763±0.0013	0.6259±0.0881	0.1779±0.0012	0.1826±0.005
Other-2 [3]	**0.5744±0.0059**	0.6669±0.0099	0.664±0.008	0.6242±0.029	0.7289±0.0612	0.731±0.07
Other-3 [2]	0.441±0.0046	0.3769±0.0063	0.3579±0.0083	0.4445±0.019	0.3674±0.0053	**0.3478±0.0064**
Other-4 [2]	0.4019±0.004	0.3988±0.0039	0.4254±0.0054	0.3891±0.019	0.4066±0.0252	0.4282±0.0282
Panel-2 [2]	0.642±0.0089	**0.1106±0.0002**	**0.1102±0.0004**	0.6587±0.0653	**0.1108±0.0005**	**0.1111±0.0009**
Panel-3 [2]	0.7069±0.0177	0.1703±0.0	0.1686±0.0008	0.7868±0.0578	0.1703±0.0	0.1717±0.0015
Other-6 [2]	0.673±0.0103	0.5515±0.0018	0.527±0.0056	0.7114±0.0437	0.5646±0.0135	0.5306±0.0196
Other-5 [2]	0.4682±0.0022	**0.3986±0.0016**	**0.3994±0.0018**	0.4674±0.0086	**0.4008±0.006**	**0.3989±0.0084**
Other-7 [2]	0.4944±0.0111	**0.4086±0.0008**	0.411±0.0009	0.5484±0.0894	0.4281±0.0135	0.4266±0.0156
Panel-1 [2]	0.4724±0.0041	0.2249±0.0002	**0.215±0.0003**	0.499±0.0952	0.2223±0.0048	**0.2161±0.0026**
Other-4 [3]	0.2035±0.0023	0.208±0.0013	0.2147±0.0018	0.2063±0.0116	0.2118±0.0066	0.2167±0.0061
Other-3 [3]	0.2043±0.0025	**0.199±0.0009**	**0.2001±0.0015**	0.2147±0.0093	**0.2006±0.0046**	**0.2036±0.0047**
Panel-2 [3]	0.4866±0.0147	**0.084±0.0001**	0.0846±0.0002	0.4758±0.1226	**0.0843±0.0006**	**0.0843±0.0008**
Panel-3 [3]	0.5445±0.0035	0.1648±0.0004	0.1629±0.0003	0.5566±0.0281	0.1656±0.0013	0.1652±0.0011
Other-6 [3]	0.3477±0.0074	0.3077±0.0025	0.3172±0.0023	0.3686±0.0373	0.3218±0.0129	0.331±0.0106
Other-5 [3]	0.3094±0.0035	0.295±0.0011	**0.2919±0.0009**	**0.3064±0.0174**	**0.2934±0.0124**	**0.295±0.0121**
Other-7 [3]	0.2624±0.0041	0.2549±0.0017	0.2543±0.002	0.2904±0.0233	0.2632±0.0102	0.2675±0.0144
Panel-1 [3]	0.4881±0.0093	**0.1728±0.0003**	0.1746±0.0005	0.5353±0.173	**0.1747±0.0022**	0.1765±0.0024

Similarly, Tables B.2 present the model calibration results in terms of ACE for the RF and MLP models and B.2 presents the results for the MLPE and MLPMC models. Also in this case, the best-performing models across all options from both tables are marked in bold for each dataset and temporal setting.

Table 6. **Overview of ACE scores across datasets and RF and MLP methods.** Averages of 10 model repeats are shown. The best-performing method as well as the methods that are statistically indistinguishable from the best one are marked in bold, including also the models in Table B.2.

Dataset	RF	RF-P	RF-VA	MLP	MLP-P	MLP-VA
Other-1 [1]	0.3179±0.0048	**0.1384±0.0121**	**0.1306±0.0101**	**0.1482±0.0507**	0.1644±0.0353	0.1599±0.019
Other-2 [1]	0.1196±0.0094	**0.0496±0.0132**	**0.0591±0.0119**	0.1696±0.0302	0.0728±0.0302	0.0722±0.0112
Other-1 [2]	0.1532±0.0014	**0.0627±0.0024**	0.0793±0.004	0.1192±0.0345	0.0965±0.008	0.1103±0.0122
Other-3 [1]	0.1656±0.0166	0.2225±0.0218	0.2404±0.021	**0.0817±0.0166**	0.1767±0.0044	0.167±0.0065
Panel-1 [1]	0.0345±0.0014	**0.012±0.002**	**0.0117±0.0022**	0.2772±0.0978	0.0155±0.005	0.016±0.0017
Other-2 [2]	**0.057±0.0097**	**0.0594±0.0067**	**0.0624±0.0096**	0.1646±0.0436	0.0841±0.0202	0.0894±0.0213
Other-1 [3]	**0.0845±0.0031**	0.1995±0.0013	0.1944±0.0023	0.2232±0.0279	0.2027±0.006	0.1943±0.0076
Other-4 [1]	0.1241±0.0046	0.1021±0.0085	0.0717±0.0102	0.0762±0.0319	0.0543±0.0111	0.0581±0.0098
Panel-2 [1]	0.0135±0.0006	**0.0117±0.0012**	**0.0105±0.0023**	0.3868±0.0611	0.0129±0.0011	0.0167±0.0016
Other-5 [1]	**0.0935±0.0115**	0.2177±0.0195	0.2487±0.0202	0.1482±0.0574	0.2416±0.0138	0.243±0.0121
Other-6 [1]	0.0581±0.0009	**0.0475±0.0016**	0.0612±0.0017	0.0869±0.004	0.0823±0.004	0.0692±0.0035
Other-7 [1]	0.2388±0.0032	0.0657±0.0065	0.0789±0.0075	0.2723±0.0419	**0.0739±0.0287**	**0.0705±0.0234**
Panel-3 [1]	**0.021±0.0008**	0.0311±0.0059	0.0265±0.0043	0.4174±0.0467	0.0223±0.0022	0.026±0.0066
Other-2 [3]	0.2412±0.0063	0.2665±0.0075	0.242±0.0078	0.0939±0.0135	0.2054±0.0238	0.2128±0.039
Other-3 [2]	0.276±0.0026	0.1339±0.0044	0.1328±0.0027	0.2051±0.0171	0.1285±0.0039	**0.0993±0.0057**
Other-4 [2]	0.0754±0.0025	**0.0632±0.0033**	0.0441±0.0023	0.0721±0.0162	0.0657±0.0145	0.0868±0.0138
Panel-2 [2]	0.0168±0.0013	0.0216±0.0023	0.0235±0.0033	0.457±0.035	**0.0094±0.0017**	**0.0107±0.0012**
Panel-3 [2]	0.0193±0.0016	0.0394±0.0022	0.0482±0.0032	0.5073±0.0283	0.0126±0.0016	0.0166±0.0065
Other-6 [2]	0.1242±0.0009	**0.1049±0.0017**	**0.1058±0.0019**	0.1496±0.0046	0.1181±0.0048	0.1156±0.0055
Other-5 [2]	0.3011±0.0081	0.1565±0.0075	0.1627±0.0096	0.2029±0.0091	0.1293±0.0034	**0.1134±0.0065**
Other-7 [2]	0.0627±0.0072	0.0588±0.013	0.0594±0.0153	0.0996±0.0198	0.0434±0.0132	**0.0424±0.0185**
Panel-1 [2]	0.0206±0.0032	**0.0152±0.0035**	0.0296±0.0028	0.323±0.0531	0.0231±0.0021	0.0208±0.0033
Other-4 [3]	0.0684±0.0013	0.0467±0.001	0.0571±0.0018	**0.0344±0.011**	0.0439±0.0044	0.0482±0.0073
Other-3 [3]	0.1909±0.0071	0.1146±0.0054	**0.0853±0.0059**	**0.0856±0.0108**	**0.0835±0.0043**	**0.083±0.0049**
Panel-2 [3]	0.0217±0.0009	0.02±0.0012	0.0248±0.0022	0.3282±0.0856	0.0105±0.0016	**0.0087±0.0016**
Panel-3 [3]	0.0162±0.0015	0.0142±0.0023	0.0202±0.0043	0.3733±0.0185	0.013±0.002	**0.0126±0.0033**
Other-6 [3]	0.0675±0.0016	0.0423±0.0022	0.0429±0.0029	0.0666±0.0185	**0.0209±0.006**	**0.0307±0.0073**
Other-5 [3]	0.2149±0.002	0.0803±0.0021	0.0571±0.0024	0.0652±0.0142	**0.0527±0.0086**	**0.0525±0.0093**
Other-7 [3]	0.0571±0.0015	0.071±0.0015	0.062±0.0062	**0.067±0.0193**	0.058±0.0065	**0.056±0.0131**
Panel-1 [3]	0.0319±0.0016	0.0286±0.0017	0.0366±0.0014	0.2918±0.1007	0.0185±0.0021	0.0223±0.0018

Table 7. **Overview of ACE scores across datasets and MLPE and MLPMC models.** Averages of 10 model repeats are shown. The best-performing method as well as the methods that are statistically indistinguishable from the best one are marked in bold, including also the models in Table B.2.

Dataset	MLPE	MLPE-P	MLPE-VA	MLPMC	MLPMC-P	MLPMC-VA
Other-1 [1]	0.2059±0.0067	0.2171±0.0046	0.1809±0.0152	**0.1482±0.0507**	0.1644±0.0353	0.1599±0.019
Other-2 [1]	0.1551±0.0109	0.0749±0.0112	0.0823±0.0045	0.1427±0.0263	0.0714±0.0228	0.0729±0.0077
Other-1 [2]	0.1166±0.0025	0.0874±0.0012	0.1098±0.0037	0.1476±0.0344	0.0965±0.0074	0.11±0.0127
Other-3 [1]	**0.0775±0.0043**	0.1788±0.0009	0.1646±0.0009	**0.0817±0.0166**	0.1767±0.0044	0.167±0.0065
Panel-1 [1]	0.2651±0.0109	0.0181±0.0005	0.0154±0.0009	0.3105±0.0762	0.0156±0.0046	0.0165±0.0024
Other-2 [2]	0.1804±0.0032	**0.0579±0.0023**	0.0732±0.0059	0.1771±0.0316	0.0695±0.0122	0.0805±0.0249
Other-1 [3]	0.2098±0.0039	0.2134±0.0018	0.2001±0.0028	0.1686±0.0219	0.2082±0.0081	0.2063±0.0067
Other-4 [1]	0.0814±0.0023	0.0584±0.0016	0.0615±0.0011	0.0806±0.0392	**0.0417±0.011**	**0.0473±0.0082**
Panel-2 [1]	0.398±0.0076	**0.0127±0.0005**	0.0155±0.0006	0.3868±0.0611	0.0129±0.0011	0.0167±0.0016
Other-5 [1]	0.1347±0.0105	0.2527±0.0011	0.2567±0.0028	0.1422±0.0545	0.2422±0.0146	0.2428±0.0122
Other-6 [1]	0.0811±0.0012	0.0839±0.0002	0.0682±0.0014	0.0866±0.0039	0.0826±0.0042	0.0693±0.003
Other-7 [1]	0.2418±0.008	**0.0463±0.0093**	**0.0517±0.0092**	0.2723±0.0419	**0.0739±0.0287**	**0.0705±0.0234**
Panel-3 [1]	0.4569±0.0035	0.0273±0.0006	**0.022±0.0017**	0.4164±0.0481	**0.0184±0.0044**	**0.0223±0.0059**
Other-2 [3]	**0.0553±0.0062**	0.1864±0.0023	0.1839±0.0027	0.0876±0.021	0.2101±0.0294	0.2139±0.0359
Other-3 [2]	0.2036±0.0033	0.1347±0.0064	0.1113±0.0068	0.2096±0.0155	0.1266±0.004	**0.0992±0.0058**
Other-4 [2]	0.0614±0.0031	0.0632±0.0022	0.0866±0.002	**0.0464±0.013**	0.068±0.0178	0.0829±0.0188
Panel-2 [2]	0.4491±0.0049	**0.0102±0.0019**	**0.0102±0.001**	0.457±0.035	**0.0094±0.0017**	**0.0107±0.0012**
Panel-3 [2]	0.4666±0.0095	0.0178±0.0018	0.0128±0.0013	0.5066±0.0283	**0.009±0.0022**	0.0133±0.0024
Other-6 [2]	0.1488±0.001	0.1181±0.0015	0.1163±0.002	0.1496±0.0046	0.1181±0.0048	0.1156±0.0055
Other-5 [2]	0.2031±0.0038	0.1265±0.0032	**0.115±0.0018**	0.2029±0.0091	0.1293±0.0034	**0.1134±0.0065**
Other-7 [2]	0.095±0.0036	**0.0238±0.0038**	0.0326±0.0024	0.0996±0.0198	0.0434±0.0132	**0.0424±0.0185**
Panel-1 [2]	0.2913±0.0036	0.0222±0.0004	0.0228±0.0011	0.3101±0.0656	0.0225±0.0016	0.0231±0.003
Other-4 [3]	**0.0352±0.0022**	**0.042±0.0015**	0.0528±0.0021	**0.0349±0.0113**	0.0436±0.0041	0.0481±0.0059
Other-3 [3]	0.0848±0.0024	**0.0808±0.001**	**0.0806±0.0014**	0.0972±0.0099	**0.0813±0.0047**	**0.083±0.0051**
Panel-2 [3]	0.3521±0.0102	0.0113±0.0005	**0.0096±0.001**	0.3326±0.0835	**0.0102±0.0016**	**0.0089±0.0013**
Panel-3 [3]	0.3717±0.0023	0.0121±0.0011	**0.0103±0.0009**	0.3791±0.0181	0.0128±0.0022	**0.0129±0.0032**
Other-6 [3]	0.0667±0.0053	**0.0165±0.0026**	0.0254±0.0022	0.0666±0.0185	**0.0209±0.006**	**0.0307±0.0073**
Other-5 [3]	0.0683±0.0036	0.0581±0.0027	**0.0529±0.0021**	0.0785±0.0167	**0.0494±0.0075**	**0.0527±0.0088**
Other-7 [3]	0.0614±0.0048	**0.0518±0.0025**	**0.048±0.004**	0.0916±0.0228	**0.0503±0.0073**	**0.052±0.0123**
Panel-1 [3]	0.2852±0.0057	**0.0146±0.0015**	0.0214±0.0009	0.2918±0.1007	0.0185±0.0021	0.0223±0.0018