Temporal Evaluation of Uncertainty Quantification Under Distribution Shift

Citation for the original published paper (version of record):
Svensson, E., Friesacher, H., Arany, Á. et al (2025). Temporal Evaluation of Uncertainty Quantification Under Distribution Shift. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 14894 LNCS: 132-148.
http://dx.doi.org/10.1007/978-3-031-72381-0_11

(article starts on next page)

Temporal Evaluation of Uncertainty Quantification Under Distribution Shift

Emma Svensson[1,3](✉), Hannah Rosa Friesacher[2,3], Adam Arany[2], Lewis Mervin[4], and Ola Engkvist[3,5]

[1] ELLIS Unit Linz, Institute for Machine Learning, Johannes Kepler University Linz, 4040 Linz, Austria
svensson@ml.jku.at
[2] ESAT-STADIUS, KU Leuven, 3000 Leuven, Belgium
[3] Molecular AI, Discovery Sciences, R&D, AstraZeneca Gothenburg, 431 83 Mölndal, Sweden
[4] Molecular AI, Discovery Sciences, R&D, AstraZeneca Cambridge, Cambridge CB2 0AA, UK
[5] Department of Computer Science and Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden

Abstract. Uncertainty quantification is emerging as a critical tool in high-stakes decision-making processes, where trust in automated predictions that lack accuracy and precision can be time-consuming and costly. In drug discovery, such high-stakes decisions are based on modeling the properties of potential drug compounds on biological assays. So far, existing uncertainty quantification methods have primarily been evaluated using public datasets that lack the temporal context necessary to understand their performance over time. In this work, we address the pressing need for a comprehensive, large-scale temporal evaluation of uncertainty quantification methodologies in the context of assay-based molecular property prediction. Our novel framework benchmarks three ensemble-based approaches to uncertainty quantification and explores the effect of adding lower-quality data during training in the form of censored labels. We investigate the robustness of the predictive performance and the calibration and reliability of predictive uncertainty by the models as time evolves. Moreover, we explore how the predictive uncertainty behaves in response to varying degrees of distribution shift. By doing so, our analysis not only advances the field but also provides practical implications for real-world pharmaceutical applications.

Keywords: uncertainty quantification · temporal evaluation · distribution shift · deep learning · drug discovery · molecular property prediction

1 Introduction

Uncertainty quantification enables safer and more reliable deployment of machine learning models in real-world applications by increasing the confidence

D.-A. Clevert et al. (Eds.): AIDD 2024, LNCS 14894, pp. 132–148, 2025.
https://doi.org/10.1007/978-3-031-72381-0_11

of humans in the models [2]. The effects are particularly important in high-stakes decision-making processes that rely on machine learning as they allow users to judge results based on the predicted uncertainty quantification before basing critical decisions on the results [11]. Drug discovery is a complex field of research where experiments are time-consuming, expensive, and high-risk, therefore wrong decisions regarding which experiments to make can be highly wasteful [29]. Additionally, the early stages of drug discovery rely on modeling the complex chemical space where data availability is typically limited, another effect of the time-consuming and costly experiments needed to generate data. As such, there is a continuously increasing need to develop application-specific uncertainty quantification methods in molecular property prediction and modeling of quantitative structure-activity relationships (QSAR) [15].

Approaches that quantify uncertainty in machine learning for regression tasks can be classified into Bayesian learning [7], ensemble-based [12,25,36,38], distance-based [4,40], mean-variance-estimation [6,8,31], evidential learning [1], etc. Several recent efforts have been made to compare and benchmark the available methods on publicly available datasets related to molecular property prediction or QSAR modeling [10,16,18,23,42,47]. However, no consensus has been reached regarding a single method that consistently outperforms the other methods across evaluation metrics and tasks [48]. Hirschfeld et al. [18] stress the need for a more realistic evaluation, such as a temporal data split, to gain insights into the real implications and nuances between the approaches. Additionally, Yin et al. [47] point out that public benchmarks do not allow proper temporal evaluation as they lack relevant information and sufficient replications for reliable statistics.

Prior work that uses temporal evaluation on public data for molecular property prediction can be misleading [27]. The reason is that the available information regarding the time of data points in public data does not relate to the real evolution of experiments in a pharmaceutical company, which is what makes a temporal evaluation truly useful in real drug discovery. Earlier work on internal pharmaceutical assay-based data from Merck compares a temporal splitting strategy with random and structure-based splitting strategies [39]. Sheridan [39] concludes that the temporal option best approximates the true predictive performance, but they do not explore uncertainty quantification.

Uncertainty quantification can be disentangled to detail the underlying sources behind the uncertainty, which gives a more comprehensive understanding of the factors that contribute to the total predictive uncertainty. In machine learning, the two main sources of uncertainty can be derived from are the aleatoric and the epistemic parts [2,20,22]. Aleatoric uncertainty is the inherent stochastic variability in experiments, also considered irreducible as it cannot be reduced with additional data or changes to the model. Epistemic uncertainty includes all remaining sources, such as lack of knowledge and model limitations. The epistemic uncertainty can be reduced with additional data or changes to the model, but understanding which adjustments are needed requires further dissection of the predicted uncertainty [13]. Awareness of the aleatoric uncertainty in

molecular property prediction can lead to better risk management by recognizing and quantifying the unpredictable nature of certain properties or parts of the chemical space [46]. Quantified epistemic uncertainty, on the other hand, can be used during drug discovery to guide the search through the chemical space by redirecting data collection [16]. If the parts of the epistemic uncertainty that relate to missing data or distribution shift can be effectively separated from the remaining model uncertainty, it can also aid in developing the machine learning model.

In this work, we provide a sought-after comparison of available methods for uncertainty quantification in a temporal evaluation of assay-based QSAR modeling for real pharmaceutical data. We focus the analysis on ensemble-based approaches that quantify predictive uncertainty and attempt to further disentangle the uncertainty between distributional uncertainty and model uncertainty such that the results are most useful in guiding the real-world search for new drugs. Additionally, we explore the effects of including lower-quality data through censored labels during training.

2 Methods

Our analysis has been performed on data from ten internal biological assays differentiated by the categories proposed by Heyndrickx et al. [17], namely Panel, Other, and ADME assays. The Panel category includes cross-project assays related to undesired off-target effects. The Other category includes on-target activity from project-specific assays. The ten assays presented in this work belong to these two categories. Larger assays of the ADME type, related to Absorption, Distribution, Metabolism, and Excretion, are left for future work. The respective distributions of observed experimental labels for each assay are shown in the bottom half of Fig. 1.

All but one of the assays model pIC50 values, while the Other 3 assay models pEC50. Due to the infeasibility of performing an unlimited number of experiments to find exact experimental results, such as pIC50 values, significant proportions of the data are provided as censored labels. Censored labels define a threshold below or above which the true results lie, e.g. the censored label < 3 pIC50 means that the true pIC50 value is below three. In some cases, the censored labels have been included in model training, as explained further in the following section. However, note that the available censored labels are highly imbalanced, as for all but two assays less than 1% of the censored labels are lower bound, i.e. $>$. The Other 2 assay has just above 1% of $>$ censored labels and the Other 4 assay has 2%, while the $<$ labels typically make up between 30–60% of each assay's total number of results. There are two assays without any censored labels, namely Other 5 and 7. Data points that are not censored are called observed labels in the remainder of this work.

Duplicated measurements for molecular compounds in the data are aggregated using the median of the result and the standard deviation is stored for later reference. Each molecular compound is then encoded with RDKit [26]

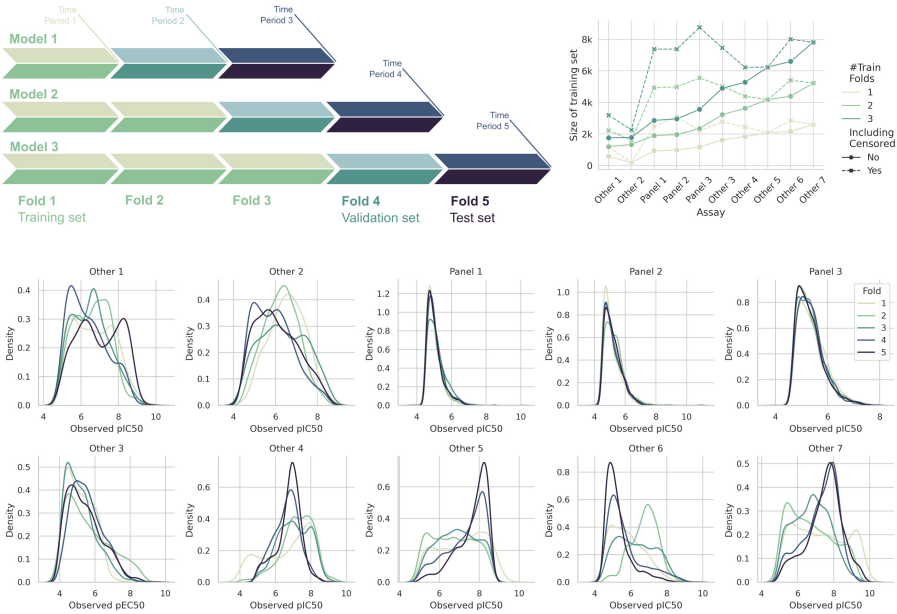

Fig. 1. Five-fold temporal split. (Upper left) Five folds and how they are used to create three temporal settings, each with more training data. For each setting, the first subsequent fold is used for validation and calibration, and the second subsequent fold is used for testing. (Upper right) Training data size for each assay and temporal setting, with and without including available censored labels. (Lower) Distribution of observed labels across the temporal folds for two example assays, one from each category.

from SMILES strings [43] to Morgan Fingerprints [30] of size 1024 and radius 2. Other. More advanced ways to encode molecular compounds exist, such as the graph-based ChemProp model [46] and the pre-trained language-based CDDD model [44]. Models based on the resulting embeddings from these neural network encoders have been compared and shown improvements in prior work [10,18,27]. Specifically, Dutschmanm et al. [10] showed that fingerprints perform best in combination with RF and are close second to CDDD in combination with neural networks. While the fingerprint representations are used in our study for simplicity, we encourage considering state-of-the-art, learned representations before deploying the proposed methods in practical applications.

Temporal Split. The main contribution of our work relates to evaluating the uncertainty quantification of molecular property prediction in a temporal evolving setting. As such, we simulate realistic assay-based modeling of pharmaceutical projects by splitting the data of each assay into five folds based on the date of the experiment. Where duplicated measurements were aggregated, the first experiment date of all measurements was used. The upper left panel in Fig. 1

illustrates the folds and resulting three settings that can be used to evaluate trained models as time evolves. The time intervals are chosen to create roughly equally sized folds regarding the number of observed labels. The resulting sizes of training sets for each assay are shown in the top right panel of Fig. 1. The solid lines show only observed results while the dashed lines include the censored labels. Note that the size of the setting with one train fold also corresponds to the size of the validation and test sets respectively, as individual, subsequent folds are used for these.

As previously mentioned, the lower part of Fig. 1 illustrates the distribution of observed labels in each fold of every assay. Note particularly, the shift in distributions between folds in the Other assays compared to the highly similar label distributions over time in the Panel assays. The assays are ordered according to the overall dataset size throughout this work.

2.1 Ensemble-Based Modeling

We compare three ensemble-based approaches for regression QSAR modeling of several internal biological assays. As such, we consider each assay t as an individual single-task dataset $\mathcal{D}_t := \{(\mathbf{x}^n, y_t^n)\}_{n=1}^N$ of molecular compounds represented by a one-dimensional numerical embeddings $\mathbf{x}^n \in \mathbb{R}^e$ and continuous activity labels $y_t^n \in \mathbb{R}$. An ensemble is defined as a set of K base estimators $\hat{y}_t^n = f(\mathbf{x}^n)$. We consider two base estimators, a decision tree regressor and a multi-layer perceptron (MLP), i.e. fully connected deep neural network. We take the average of the individual base estimators' predictions as the final prediction by the ensemble μ_t and define the variance of the predictions as an estimate of the predictive uncertainty σ_t^2, as follows

$$\mu_t(\mathbf{x}^n) = \frac{1}{K}\sum_{k=1}^{K} f_k(\mathbf{x}^n), \quad \sigma_t^2(\mathbf{x}^n) = \frac{1}{K}\sum_{k=1}^{K}(f_k(\mathbf{x}^n))^2 - (\mu_t(\mathbf{x}^n))^2. \tag{1}$$

The ensemble of decision tree regressors results in a Random Forest (RF) model [38] while we use the MLP base estimator to create a Deep Ensemble (DE) as proposed by Lakshminarayanan et al. [25] and an MC-Dropout model as proposed by Gal & Ghahramani [12]. Prior work has compared similar methods for variability in QSAR modeling [41]. The DE combines base 50 MLPs trained from different weight initialization whereas the MC model generates 500 samples from a single trained base MLP with dropout turned on during inference.

In a Bayesian framework, the uncertainty in model parameters ω results in the predictive uncertainty of the model $p(y_t^n|\mathbf{x}^n, \omega)$. The true posterior distribution of model parameters for a given dataset can be described as $p(\omega|\mathcal{D}_t)$, such that the predictive uncertainty of the Bayesian model average is defined by $p(y_t^n|\mathbf{x}^n, \mathcal{D}_t) = \int_\Omega p(y_t^n|\mathbf{x}^n, \tilde{\omega})p(\tilde{\omega}|\mathcal{D}_t)\mathrm{d}\tilde{\omega}$ [11,20]. As shown by both Lakshminarayanan et al. [25] and Gal & Ghahramani [11], the variance in ensemble predictions provides an approximation of the epistemic part of this true posterior distribution. Figure 2 gives an overview of the three ensemble-based methods

considered in our work. The remainder of this section gives details about the training procedures used in the evaluation of the three methods.

Training Details. The Random Forest is implemented using scikit-learn [34] and the two neural network-based models are trained with PyTorch [32]. All models are initially trained with a Mean Squared Error (MSE) loss only on data points with observed labels. However, in addition, we include versions of the neural network-based models for which censored labels are also included in the training data. We denote these models as DE+ and MC+ in the result. Note that these extended models are not provided for the Other 5 and 7 assays, which do not include censored labels. Training these extended models requires adjustments to the loss function, as censored labels only give a one-sided view of the true result. We adopt the CensoredMSE defined by Arany et al. [3] with a one-sided squared error applied for the censored labels as follows

$$\mathcal{L}(\mathbf{x}^n, y_t^n) = \frac{1}{N}\sum_{n=1}^{N} \begin{cases} \min\left(y_t^n - \mu_t(\mathbf{x}^n), 0\right)^2, & \text{if censored label } < y_t^n, \\ (y_t^n - \mu_t(\mathbf{x}^n))^2, & \text{if observed label } y_t^n, \\ \max\left(y_t^n - \mu_t(\mathbf{x}^n), 0\right)^2, & \text{if censored label } > y_t^n. \end{cases} \tag{2}$$

To compare the models trained on censored labels fairly against the ones trained only on observed labels we only include the censored labels in the training sets. Thus, the validation and test sets are identical between the models. We believe this could hinder the censored models somewhat, especially due to the imbalance between lower and upper-bound labels.

We optimize the hyperparameters for each base estimator detailed in Table 1 of the Appendix for each assay and each temporal setting individually using a grid search based on the validation MSE loss. It would not be computationally feasible to optimize the DE model in terms of any score that incorporates the calibration of uncertainty estimates due to the large number of models that would need to be trained. Therefore, we do not consider this option for any of the models to ensure a fair comparison. However, such optimization schemes should be considered for practical applications.

Evaluation. While the MSE loss is used to evaluate the performance of the predictions made by the models, other metrics are required to evaluate the accuracy and calibration of the predicted uncertainties. We consider two types of ways to evaluate predicted uncertainty, ones that evaluate only the accuracy or calibration of the uncertainty and ones that evaluate predictive performance intertwined with how well-calibrated the predicted uncertainty is. A detailed way to evaluate the predicted uncertainties by themselves is by comparing the confidence-based calibration curve to the identity function which corresponds to perfect calibration [16,19,42,45]. The confidence-based calibration curve is obtained by computing the z% confidence interval (CI) for every predicted uncertainty in the test set. Next, the observed fraction of errors within each CI is calculated for several expected fractions between 0 and 1.

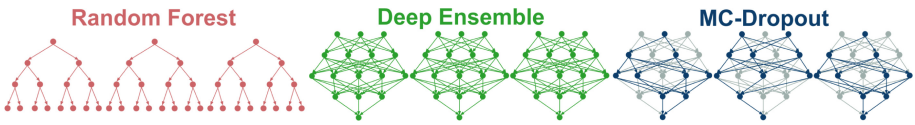

Fig. 2. Ensemble-based models. Three approaches to ensemble-based modeling including uncertainty quantification.

Furthermore, the Gaussian Negative Log Likelihood (NLL) [49] and the Expected Normalized Calibration Error (ENCE) [28] are two global metrics that evaluate the intertwined predictive performance and calibration of uncertainties. The Gaussian NLL is defined as,

$$\text{NLL} = \frac{1}{2N} \sum_{n=1}^{N} \left(\ln(2\pi) + \ln(\sigma_t^2(\mathbf{x}^n)) + \frac{(y_t^n - \mu_t(\mathbf{x}^n))^2}{\sigma_t^2(\mathbf{x}^n)} \right). \tag{3}$$

The ENCE metric is derived from the error-based calibration plot proposed by Levi et al. [28] which is made from a binned representation of the Root MSE and the Root Mean Variance (RMV), i.e. predicted uncertainty. Computationally, the errors and corresponding predicted uncertainties are ordered based on increasing predicted uncertainty and split into a set $\mathcal{B}$ of bins. For each bin b of size $|b|$ the RMSE and RMV are calculated as,

$$\text{RMSE}_b = \sqrt{\frac{1}{|b|} \sum_{i \in b} (y_t^i - \mu_t(\mathbf{x}^i))^2}, \quad \text{RMV}_b = \sqrt{\frac{1}{|b|} \sum_{i \in b} \sigma_t^2(\mathbf{x}^i)}. \tag{4}$$

Finally, the bins are summarized to give the ENCE metric as follows,

$$\text{ENCE} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{|\text{RMSE}_b - \text{RMV}_b|}{\text{RMV}_b}. \tag{5}$$

Several additional metrics have been proposed and used to evaluate uncertainty estimates in drug discovery applications, such as Spearman's Rank Correlation Coefficient between predicted uncertainties and corresponding errors [10,18,42,47]. However, this score has been criticized due to the stochasticity and unreliability of the result [35]. Statistically, a data point with high predicted uncertainty can still result in a prediction with low error and vice versa. Therefore, we discard the metric from our analysis.

Recalibration. Several post hoc alternatives have been proposed to recalibrate predicted uncertainties by ensemble-based models [21,28,35], as the original estimates have been found to underestimate the epistemic uncertainty [9,37]. Janet et al. [21] recalibrate the uncertainty estimates based on a maximum-likelihood estimation strategy on the NLL, while Levi et al. [28] propose a re-scaling of the predicted uncertainty based on the NLL similar to temperature scaling [14].

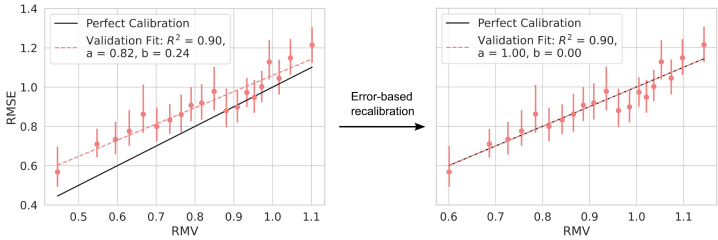

Fig. 3. Error-based recalibration. Linear recalibration of uncertainty estimates based on the validation set.

Most recently, Rasmussen et al. [35] instead proposed to recalibrate the predicted uncertainty using the fit of the RMSE versus RMV curve described above as the error-based calibration plot. The latter is the strategy that we adopt in this work and Fig. 3 illustrates an example of a recalibration on the validation set of one of our datasets. A linear regression is fitted to the binned RMSE versus RMV results on the validation set, resulting in parameters a_{val} for the slope and b_{val} for the intercept. During inference the predicted standard deviation is then shifted according to $\sigma_{\text{cal}} = a_{\text{val}} \cdot \sigma + b_{\text{val}}$.

3 Experiments

In the experimental setup, we first analyze and compare the performance of the models averaged over ten repeated experiments on all assays and temporal settings. The global evaluation scores are shown in Fig. 4 and the confidence-based calibration curves are shown in Fig. 5. We then provide a more in-depth case study of the predictions by one of the best-performing models on the Other 6 assay, which exhibits a particularly challenging distribution shift in terms of both the feature and label space. Here, we illustrate how the predicted uncertainties relate to the distribution shift in the feature space and suggest how the model's predictions could have practical implications for future decisions in the given drug discovery project.

Model Comparison. Figure 4 presents an overview of the MSE and recalibrated NLL and ENCE scores. Note that the recalibration step only affects the predicted uncertainties and therefore does not affect the MSE. In the figure, the models can be compared in several ways: 1) as the training set size increases over time for each assay with increasing #Train folds, 2) as the overall size of the assay increases, going from smallest assays in the left-most columns to larger assays in the right-most columns, 3) in terms of the varying amounts of label shifts between the Panel and Other assays, or 4) in terms of the different metrics.

The first observable trend is that predictive performance is higher for the Panel assays compared to the Other assays. This is not surprising given the

Fig. 4. Benchmarking overview. Results for each assay and temporal setting averaged over ten repeated experiments. DE+ and MC+ are trained with censored labels as supplementary lower-quality data. However, these models do not apply to Other 5 and 7 as they do not include censored labels.

constant distribution over time as illustrated in Fig. 1. A similar trend can be observed in the NLL but not in terms of ENCE. As the Gaussian NLL includes the squared error term, a likely conclusion is that distribution shifts do not generally hurt the calibration of uncertainty estimates. This conclusion is also reasonable as the predictive uncertainty from ensemble-based approaches model specifically the epistemic uncertainty which should cover distribution shifts. In general, the ranking of the methods from the MSE scores are often the same in the NLL while they can vary in terms of the ENCE. For example, for the Other 1 assay the DE is always among the best models for all three temporal settings in terms of the MSE and NLL scores, while in terms of the ENCE score, it is outperformed by the RF model in the first two temporal settings.

For the most part, the performances of the two MLP-based models are usually indistinguishable from each other for the cases trained with and without censored data respectively. On the contrary, there are no general trends regarding whether the RF model or the MLP-based models are best. This changes depending on the assay, metric, and even temporal setting. However, the versions of the neural network-based models trained with supplementary censored labels, DE+ and MC+, do not generally improve the predictive performance or the calibration of uncertainty estimates over their respective base versions. Only one instance occurs where the DE+ is better than the DE model and all other models for all three scores, namely the Other 2 model trained on 2 folds. However, as this result is not consistent across the other two temporal settings of the assay, it is more likely the result of statistical variability. The non-competitive results with

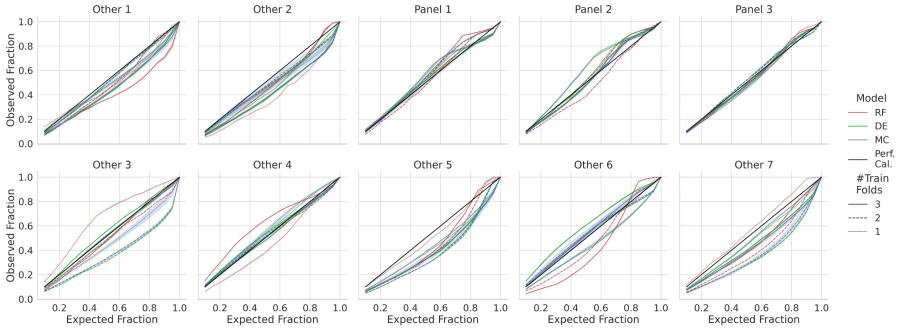

Fig. 5. Confidence-based calibration over time. Discrete visualization of the observed fraction of results for each expected confidence interval based on the predicted uncertainty. The black, solid lines illustrate perfect calibration.

the censored models require further analysis, but we believe it could be due to the uneven nature of the censored labels available. As described in Sect. 2, the vast majority of the censored labels are upper bound (<). Given that all models are evaluated only on observed labels for a fair comparison, the imbalance in the censored labels may shift the models' understanding of the label distributions.

In light of the overall poor performance of the models trained with censored data, we have omitted these models from the confidence-based calibration curves presented in Fig. 5. The curves are shown for each assay and temporal setting with error bands illustrating the confidence from the ten repeated experiments. A majority of calibration curves are not far away from being perfectly calibrated. This indicates that most models produce useful uncertainty estimates. The possibility of such intuitive interpretations of the calibration curves is not as easily derived from the scores presented in Fig. 4. The reason for this is that the ENCE score is unbounded, such that it can be hard to determine whether achieved scores are useful or not. For the calibration curves in Fig. 4 it is clear that they are significantly closer to being perfectly calibrated than to the extreme cases of completely over- or under-calibration. On the other hand, it is harder to compare the models and temporal settings in terms of the calibration curves, as many of the curves are indistinguishable. However, in practical applications where perhaps a particular confidence is of interest, a closer evaluation of the calibration curves can be crucial to distinguish between the models.

Case Study. Finally, we provide a practical case study of one of the Other assays, Other 6, which exhibits a particularly challenging evolution of the data throughout time. Our case study aims to test the top-performing model from the model comparison above in this demanding setting to determine in detail how well the predictive uncertainties perform, and how individual predictions can be used in practice to impact future decisions of the drug discovery project.

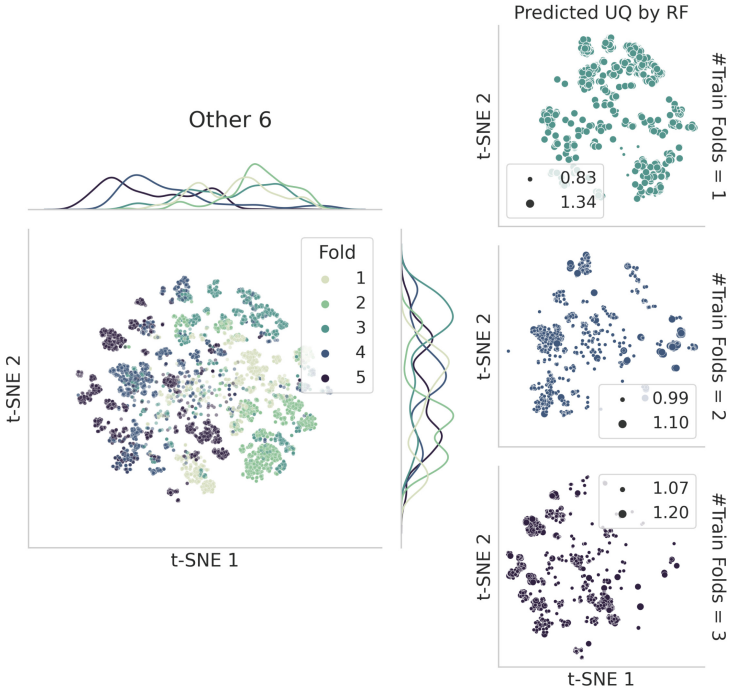

Fig. 6. Practical temporal evaluation. A t-SNE projection of the Other 6 assay, colored by temporal fold. The left panel illustrates the full dataset, where a distribution shift can be seen throughout time. The remaining panels in the right column, show individual test sets with predicted uncertainty by the RF model presented as the size of data points.

The leftmost part of Fig. 6 illustrates the feature space of the compounds tested on the assay decomposed to two t-SNE projections and colored by the five temporal folds. A clear distribution shift in the feature space can be observed in the t-SNE projection where the second fold tends more toward the bottom right corner of the feature space and the last two folds shift drastically to the left side of the plot. Similarly, highly varying label distributions were seen between the same folds in the lower part of Fig. 1 in Sect. 2. Also, the label distribution does not shift continuously over time, but instead first shifts greatly toward higher pIC50 values in the second fold and then back toward more extreme lower values by the last two folds. In the remaining three plots to the right in Fig. 6, the t-SNE projections of each test set, i.e. folds 3, 4, and 5, are repeated separately. Here, the size of the data points is determined by the recalibrated predicted uncertainty of the RF model trained on the three temporal settings respectively, i.e. with an increasing number of training folds. The RF model is chosen for this analysis due to being best-performing on the Other 6 assay in terms of the

ENCE score. Note that the legends of these plots detail the respective minimum and maximum predicted uncertainties on the given test set.

We observe that the model trained on the least amount of data namely on only the first fold and tested on fold 3, seen in the top panel of the right column in Fig. 6, indicates overall high uncertainty for most test compounds. A likely explanation is that the amount of training data was insufficient for the model to learn from, meaning that it overfitted and could not generalize well to the test compounds. The described scenario is also corroborated by the relatively poor MSE score seen for RF trained on one fold of assay Other 6 in Fig. 4 compared to the same model trained on two and three folds respectively. For the models trained on two and three folds, the span of predicted uncertainties is notably much smaller, 0.11 and 0.13, compared to the first model, 0.51. As a result, we can observe more distinct patterns in the predicted uncertainty between different regions of the feature space. The regions with high uncertainty predicted by the model trained on two folds seem uncorrelated with proximity to training data. However, when the third and final training fold is included, it is clear that the clusters with the highest predicted uncertainty are also located furthest away from the training data. The same trend is reflected in the ENCE scores presented in Fig. 4 where the calibration error of the model trained on two folds is significantly worse than the one achieved by the model trained on three folds.

Given the distribution shift present in the feature space, and that the ensemble-based model's predicted uncertainty accounts for epistemic uncertainty, it follows our expectation that the distribution shift should be reflected in the estimated uncertainty. As such, our analysis provides empirical evidence to support this claim, but it also illustrates that the uncertainty estimates cover additional sources of uncertainty related to the model itself such as overfitting. It is important to understand all sources of uncertainty when basing future high-stakes decisions on them, such as in drug discovery. Considering the identified cases in this case study, we provide practical suggestions on how the identified sources can impact the continued drug discovery process. If overfitting is determined, such as through overall high uncertainty estimates and low performance seen in the model train on one fold, the modeling requires overall more data before deployment. Another alternative would be to reconsider the choice of model, but our temporal split shows that the RF continues to be the best choice in the future when more data is included. When distribution shifts are instead identified, such as seen later in the given project for the model trained on three folds, more data exploration is needed in the chemical spaces where the uncertainty estimates are high before deployment.

Further research is necessary to disentangle the sources of epistemic uncertainty between distribution shifts and other model-related sources, such that more reliable measurements of these situations can be obtained. One alternative approach would be to quantify the distribution shift using other means, either with distance-based approaches, such as the average Tanimoto similarity [40] between an inference compound and compounds in the training set, or the

interpretable method proposed by Kulinski and Inouye [24]. Additionally, more advanced pre-training procedures can be used, that are trained to incorporate distribution shift more effectively [5]. After the distribution shift has been independently quantified, the predicted epistemic uncertainty could be re-evaluated such that the remaining model uncertainty is disentangled from this information.

4 Conclusions

In this comparison between three ensemble-based uncertainty quantification approaches evaluated temporally on data from multiple biological assays, we have shown varying results between the assays emphasizing the impact of individual assay characteristics on predictive outcomes. No single model was consistently best across evaluation metrics or assays, but some conclusions could be drawn for particular assays. Specifically, we analyzed the results in light of the varying presence of shifts in label distributions and feature space distributions in the assays over time. While doing so we found that predictive performance and calibration of uncertainty can be robust and reliable for assays without distribution shifts and that the method can be used to identify data points outside of the training distribution when distribution shifts are present. As such, we give insights and provide practical advice on how uncertainty estimates by ensemble-based models can be used to impact future decision-making in high-stakes situations such as drug discovery. Incorporating lower-quality data in the form of censored labels did not yield improvements in the predictive performance of the models. Suggestions were given as to why this could be the case, such as the uneven nature of the censored labels and the evaluation strategy. Future work can explore other ways to include the censored labels or extend the analysis to other modeling approaches that allow censored labels, such as Censored Quantile Regression [33]. Overall, this study has gained valuable insights into how distribution shift affects uncertainty quantification in assay-based QSAR modeling, which can impact real-world pharmaceutical drug discovery.

Acknowledgements. We thank our colleagues and reviewers for their valuable feedback, especially Susanne Winiwarter at AstraZeneca in Gothenburg for her advice and guidance 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.

Appendix

Table 1 presents the hyperparameters explored in the model selection for the RF model and the base MLP used for both the DE and MC models. A grid search

was used to find the optimal hyperparameters for every assay and temporal setting based on the validation MSE loss. Additionally, the MLPs were trained using the Adam optimizer with a weight decay of 0.0005, the learning rate was reduced when plateauing with a patience of 50 epochs, and a batch size of 64 was used.

Table 1. Model selection. Considered hyperparameter space for model selection of RF and base MLP during grid search based on validation MSE loss.

Base Model	Hyperparameter	Explored space
RF	n_estimators	{50, 100, 250, 500, 1000}
	min_samples_leaf	{2, 10, 0.25, 0.5, 0.75}
	min_samples_split	{1, 25, 50, 100, 250, 500}
MLP	Learning rate	{0.00005, 0.0001, 0.0005, 0.001}
	Scheduler Factor	{0.1, 0.5}
	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}

References

1. Amini, A., Schwarting, W., Soleimany, A., Rus, D.: Deep evidential regression. In: Advances in Neural Information Processing Systems, vol. 33, pp. 14927–14937. Curran Associates, Inc. (2020)
2. Apostolakis, G.: The concept of probability in safety assessments of technological systems. Science **250**(4986), 1359–1364 (1990)
3. Arany, A., Simm, J., Oldenhof, M., Moreau, Y.: SparseChem: fast and accurate machine learning model for small molecules. arXiv preprint arXiv:2203.04676 (2022)
4. Berenger, F., Yamanishi, Y.: A distance-based boolean applicability domain for classification of high throughput screening data. J. Chem. Inf. Model. **59**(1), 463–476 (2018)
5. Bertolini, M., Clevert, D.A., Montanari, F.: Explaining, evaluating and enhancing neural networks' learned representations. In: International Conference on Artificial Neural Networks, pp. 269–287. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-44192-9_22
6. Bishop, C.M.: Mixture Density Networks. Technical report. Aston University, Birmingham (1994)
7. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: International Conference on Machine Learning, pp. 1613–1622. PMLR (2015)

8. Choi, S., Lee, K., Lim, S., Oh, S.: Uncertainty-aware learning from demonstration using mixture density networks with sampling-free variance modeling. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 6915–6922. IEEE (2018)
9. D'Angelo, F., Fortuin, V.: Repulsive deep ensembles are bayesian. In: Advances in Neural Information Processing Systems, vol. 34, pp. 3451–3465. Curran Associates, Inc. (2021)
10. Dutschmann, T.M., Kinzel, L., Ter Laak, A., Baumann, K.: Large-scale evaluation of k-fold cross-validation ensembles for uncertainty estimation. J. Cheminf. **15**(1), 49 (2023)
11. Gal, Y.: Uncertainty in Deep Learning. Ph.D. thesis, Department of Engineering, University of Cambridge (2016)
12. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059. PMLR (2016)
13. Gruber, C., Schenk, P.O., Schierholz, M., Kreuter, F., Kauermann, G.: Sources of Uncertainty in Machine Learning–A Statisticians' View. arXiv preprint arXiv:2305.16703 (2023)
14. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330. PMLR (2017)
15. Hansch, C., Fujita, T.: p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J. Am. Chem. Soc. **86**(8), 1616–1626 (1964)
16. Heid, E., McGill, C.J., Vermeire, F.H., Green, W.H.: Characterizing uncertainty in machine learning for chemistry. J. Chem. Inf. Model. **63**(13), 4012–4029 (2023)
17. 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)
18. Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R., Coley, C.W.: Uncertainty quantification using neural networks for molecular property prediction. J. Chem. Inf. Model. **60**(8), 3770–3780 (2020)
19. Hubschneider, C., Hutmacher, R., Zöllner, J.M.: Calibrating uncertainty models for steering angle estimation. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 1511–1518. IEEE (2019)
20. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach. Learn. **110**, 457–506 (2021)
21. Janet, J.P., Duan, C., Yang, T., Nandy, A., Kulik, H.J.: A quantitative uncertainty metric controls error in neural network-driven chemical discovery. Chem. Sci. **10**(34), 7913–7922 (2019)
22. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
23. Kim, Q., Ko, J.H., Kim, S., Park, N., Jhe, W.: Bayesian neural network with pretrained protein embedding enhances prediction accuracy of drug-protein interaction. Bioinformatics **37**(20), 3428–3435 (2021)
24. Kulinski, S., Inouye, D.I.: Towards explaining distribution shifts. In: International Conference on Machine Learning, pp. 17931–17952. PMLR (2023)
25. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)

26. Landrum, G.: RDKit: Open-Source Cheminformatics (2006). https://doi.org/10.5281/zenodo.6961488, http://www.rdkit.org
27. Lenselink, E.B., et al.: Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. J. Cheminf. **9**(1), 1–14 (2017)
28. Levi, D., Gispan, L., Giladi, N., Fetaya, E.: Evaluating and calibrating uncertainty prediction in regression tasks. Sensors **22**(15), 5540 (2022)
29. 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)
30. Morgan, H.L.: The generation of a unique machine description for chemical structures - a technique developed at chemical abstracts service. J. Chem. Doc. **5**(2), 107–113 (1965)
31. Nix, D.A., Weigend, A.S.: Estimating the mean and variance of the target probability distribution. In: Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94), vol. 1, pp. 55–60. IEEE (1994)
32. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. (2019)
33. Pearce, T., Jeong, J.H., Jia, Y., Zhu, J.: Censored quantile regression neural networks for distribution-free survival analysis. In: Advances in Neural Information Processing Systems, vol. 35, pp. 7450–7461. Curran Associates, Inc. (2022)
34. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
35. Rasmussen, M.H., Duan, C., Kulik, H.J., Jensen, J.H.: Uncertain of uncertainties? a comparison of uncertainty quantification metrics for chemical data sets. J. Cheminf. **15**(1), 121 (2023)
36. Scalia, G., Grambow, C.A., Pernici, B., Li, Y.P., Green, W.H.: Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. J. Chem. Inf. Model. **60**(6), 2697–2717 (2020)
37. Schweighofer, K., Aichberger, L., Ielanskyi, M., Klambauer, G., Hochreiter, S.: Quantification of Uncertainty with Adversarial Models. In: Advances in Neural Information Processing Systems, vol. 36. Curran Associates, Inc. (2023)
38. Sheridan, R.P.: Three useful dimensions for domain applicability in QSAR models using random forest. J. Chem. Inf. Model. **52**(3), 814–823 (2012)
39. 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)
40. Sheridan, R.P., Feuston, B.P., Maiorov, V.N., Kearsley, S.K.: Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. J. Chem. Inf. Comput. Sci. **44**(6), 1912–1928 (2004)
41. Tetko, I.V., et al.: Critical Assessment of QSAR Models of Environmental Toxicity Against Tetrahymena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. J. Chem. Inf. Model. **48**(9), 1733–1746 (2008)
42. Wang, D., et al.: A hybrid framework for improving uncertainty quantification in deep learning-based QSAR regression modeling. J. Cheminf. **13**(1), 1–17 (2021)
43. 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)
44. Winter, R., Montanari, F., Noé, F., Clevert, D.A.: Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem. Sci. **10**(6), 1692–1701 (2019)

45. Yang, C.I., Li, Y.P.: Explainable uncertainty quantifications for deep learning-based molecular property prediction. J. Cheminf. **15**(1), 13 (2023)
46. Yang, K., et al.: Analyzing learned molecular representations for property prediction. J. Chem. Inf. Model. **59**(8), 3370–3388 (2019)
47. Yin, T., Panapitiya, G., Coda, E.D., Saldanha, E.G.: Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction. J. Cheminf. **15**(1), 105 (2023)
48. Yu, J., Wang, D., Zheng, M.: Uncertainty quantification: can we trust artificial intelligence in drug discovery? iScience **25**(8), 104814 (2022)
49. Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: International Conference on Machine Learning, pp. 609–616. PMLR (2001)