



**CHALMERS**  
UNIVERSITY OF TECHNOLOGY

## **Enhancing uncertainty quantification in drug discovery with censored regression labels**

Downloaded from: <https://research.chalmers.se>, 2025-05-23 11:52 UTC

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


Svensson, E., Friesacher, H., Winiwarter, S. et al (2025). Enhancing uncertainty quantification in drug discovery with censored regression labels. *Artificial Intelligence in the Life Sciences*, 7. <http://dx.doi.org/10.1016/j.aillsi.2025.100128>

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



## Research article

# Enhancing uncertainty quantification in drug discovery with censored regression labels

Emma Svensson <sup>a,b</sup> <sup>\*</sup>, Hannah Rosa Friesacher <sup>a,c</sup>, Susanne Winiwarter <sup>d</sup>, Lewis Mervin <sup>e</sup>, Adam Arany <sup>c</sup>, Ola Engkvist <sup>a,f</sup>

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

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

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

<sup>d</sup> Drug Metabolism and Pharmacokinetics, Research and Early Development Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, 431 83, Sweden

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

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

## ARTICLE INFO

Dataset link: <https://github.com/MolecularAI/uq4dd>

## Keywords:

Uncertainty quantification  
Censored regression  
Temporal evaluation  
Distribution shift  
Deep learning  
Drug discovery  
Molecular property prediction

## ABSTRACT

In the early stages of drug discovery, decisions regarding which experiments to pursue can be influenced by computational models for quantitative structure–activity relationships (QSAR). These decisions are critical due to the time-consuming and expensive nature of the experiments. Therefore, it is becoming essential to accurately quantify the uncertainty in machine learning predictions, such that resources can be used optimally and trust in the models improves. While computational methods for QSAR modeling often suffer from limited data and sparse experimental observations, additional information can exist in the form of censored labels that provide thresholds rather than precise values of observations. However, the standard approaches that quantify uncertainty in machine learning cannot fully utilize censored labels. In this work, we adapt ensemble-based, Bayesian, and Gaussian models with tools to learn from censored labels by using the Tobit model from survival analysis. Our results demonstrate that despite the partial information available in censored labels, they are essential to reliably estimate uncertainties in real pharmaceutical settings where approximately one-third or more of experimental labels are censored.

## 1. Introduction

Drug discovery is a challenging field of research where experiments are time-consuming and expensive. In addition, the development of therapeutic agents bears a high risk of failure resulting in the abandonment of a drug candidate during the later stages of the drug discovery pipeline, which leads to an extensive waste of money and time. To optimize the use of resources and accelerate the drug development workflow, machine learning models are often applied to support a smart allocation of these resources [1]. In the context of machine learning-assisted drug discovery, uncertainty quantification enables safer and more reliable deployment of computational models by increasing human confidence in the models [2]. The effects are highly relevant for the drug discovery pipeline as they allow users to judge results based on the predicted uncertainty quantification before deciding how to progress in the experimental workflow [3]. Applying machine learning models to early-stage drug discovery requires modeling the complex chemical space where data availability is typically

limited. This low-data problem is another effect of the time-consuming and costly experiments needed to generate data. As such, there is a continuously increasing need for application-specific uncertainty quantification methods in molecular property prediction, particularly in the modeling of affinity scores and drug side effects through quantitative structure–activity relationships (QSAR) [4].

Uncertainty quantification in machine learning is typically disentangled into its underlying sources, which provide a deeper understanding of the factors contributing to overall predictive uncertainty [2]. The two primary sources of uncertainty are aleatoric and epistemic uncertainty [5,6]. Aleatoric uncertainty refers to the inherent stochastic variability within experiments. It is often considered irreducible because it cannot be mitigated through additional data or model improvements. In drug discovery specifically, aleatoric uncertainty can reflect the inherent unpredictability of interactions between certain molecular compounds as a result of either biological stochasticity or human intervention. As such, proper quantification of aleatoric uncertainty can

\* Corresponding author at: Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, 431 83, Sweden.  
E-mail address: [svensson@ml.jku.at](mailto:svensson@ml.jku.at) (E. Svensson).

lead to better risk management in drug discovery by highlighting areas where outcomes are inherently uncertain [7].

On the other hand, epistemic uncertainty encompasses uncertainties related to the model's lack of knowledge, which can stem from insufficient training data or model limitations [5,6]. Unlike aleatoric uncertainty, epistemic uncertainty can be reduced by acquiring additional data or through improvements to the model. Understanding and quantifying epistemic uncertainty in drug discovery allows researchers to strategically guide data collection efforts, focusing on areas of the chemical space where the model's predictions are less certain [8]. By further separating these individual aspects of epistemic uncertainty, targeted enhancements to the machine learning model itself can be achieved. Recent work by Gustafsson et al. [9] emphasizes the importance of uncertainty quantification in out-of-distribution scenarios, highlighting how epistemic uncertainty can inform about when a model will likely fail due to a lack of relevant training data. In real-world drug discovery, it is especially advantageous to understand the various sources of epistemic uncertainty, as navigating the vast chemical space efficiently and effectively can significantly impact the success of a project.

Approaches that quantify uncertainty in machine learning regression tasks can be broadly categorized into several types: Bayesian learning [10], ensemble-based methods [11–14], distance-based approaches [15,16], mean–variance estimation [17–19], evidential learning [20], conformal prediction [21], and quantile regression [22], among others. Recent studies have compared and benchmarked these methods on publicly available datasets for modeling of molecular properties [8,23–27]. Despite these efforts, no single method has emerged as consistently superior across all evaluation metrics and tasks [3,28].

The prior research mentioned above has primarily been performed on publicly available datasets using random or scaffold-based ways to split the data for evaluation. These means have been shown to either overestimate machine learning models, in the case of the random split, or underestimate the performance with the scaffold-based split [23,27,29]. Hirschfeld et al. [23] 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. [27] point out that public benchmarks do not allow proper temporal evaluation as they lack relevant information and sufficient replications for reliable statistics. Temporal evaluation based on information available in public data, such as ChEMBL, for molecular property prediction can be misleading [30]. The reason is that the time stamp of data points in public data relates to when the compound was added to the public domain rather than when the experiment was performed. Instead, modeling the real evolution of experiments in a pharmaceutical company is what makes a temporal evaluation truly useful [31]. Earlier work on internal pharmaceutical assay-based data from Merck compares a temporal splitting strategy with random and structure-based splitting strategies [32]. Sheridan [32] concludes that the temporal option best approximates the true predictive performance, but they do not explore uncertainty quantification.

In addition to the limited amount of precise data in drug discovery, partial information in the form of censored labels is often generated during experiments. Censored labels arise when the experiment's measurement range is exceeded, such that the exact value cannot be recorded. For instance, a fixed range of compound concentrations is typically used to test the compound's effect on biochemical processes in an assay. If no response is observed within this range, the experiment may only indicate that the response lies above or below the tested concentrations rather than providing a specific value. This results in a censored label, where the true value is known only to exceed or fall below a certain threshold. While censored labels can be easily included in classification tasks by categorizing observations as active or inactive [33], integrating them into regression models that predict continuous values is far less trivial. Due to this challenge, censored data has not yet been properly utilized in regression tasks within

drug discovery, despite its potential to enhance model accuracy and uncertainty quantification.

The problem of censored regression has been widely studied in other fields, particularly in survival analysis, where Hollander et al. [34] demonstrated that the information contained in data decreases after censorship. Despite the reduced information in censored labels, Hüttel et al. [35] further showed that these labels are crucial in Bayesian active learning due to their contribution to increased entropy and mutual information between the predicted and true data distributions. Additionally, censored labels have enhanced quantile regression in survival analysis, as demonstrated by Pearce et al. [36]. In the context of drug discovery, limited work has explored ways to incorporate censored labels for regression tasks. Arany et al. [37] proposed an adaptation of the mean squared error (MSE) to account for censored labels by using a one-sided squared loss. However, their approach did not address uncertainty quantification. This approach was first explored for uncertainty quantification in an earlier version of our work on a smaller subset of the datasets provided here [38]. To the best of our knowledge, no prior work has yet adapted both the MSE and the more general Gaussian negative log-likelihood (NLL) for uncertainty quantification in predictive modeling for drug discovery. Therefore, further exploration of how censored labels can be effectively integrated into regression models to improve prediction accuracy and uncertainty estimation in drug discovery is needed.

In this work, we address the challenge of effectively utilizing censored regression labels during QSAR modeling in drug discovery, a domain where data is often scarce, and uncertainty quantification is critical for decision-making. While previous approaches have primarily focused on high-quality data, our work expands the ability of machine learning models to incorporate partial, yet valuable information provided in censored labels, thereby enhancing the reliability of the uncertainty estimates. To achieve this, we develop and extend existing methods, adapting them to the unique demands of drug discovery tasks. We also refine evaluation techniques to rigorously assess the impact of incorporating censored labels, ensuring that the benefits are demonstrated across relevant metrics. Our contributions are summarized as follows:

- We derive extended versions of ensemble-based models, Bayesian models, and Gaussian mean–variance estimators capable of learning from additional partial information available in censored regression labels.
- Similarly, we adapt available evaluation methods to compare models trained with and without the additional censored labels.
- Furthermore, we provide a large-scale comparison between the resulting censored regression models in a comprehensive temporal evaluation using internal pharmaceutical assay-based data.
- Finally, we showcase how the resulting model predictions can be used in practical applications to streamline and aid the drug discovery process in a case study.

## 2. Methods

The methodologies employed in our study are divided into two parts detailing the data and the modeling approaches. First, we describe the data used in our analysis, including the sources and types of biological assays explored. Following this, we outline the models and techniques implemented to leverage this data, focusing on how they are adapted to handle the unique challenges presented by censored regression labels.

The analysis in this work was performed on data from 15 internal biological assays, categorized into two distinct groups: project-specific target-based assays and cross-project assays related to side effects, such as Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADME-T) properties [39]. This division allows us to explore the effects of censored labels on uncertainty estimation using different assay types. Similar datasets were utilized in the recent study by Friesacher et al.

**Table 1**

**Data overview.** A detailed overview of the assays used in this analysis, including the short name used throughout the results. For each assay, key properties are shown, such as the proportion of the censored data, the end-point, the assay size, and the experimental standard deviation derived from a control compound. Due to proprietary constraints, detailed descriptions of target-based assays cannot be disclosed.

Short name	Assay description	End-point	Goal	Assay size	Censoring		Control std
					<	>	
<b>Target-based</b>							
Target 1	N/A	pIC50	↑	5,082	32%	1%	0.13
Target 2	N/A	pIC50	↑	5,237	43%	0%	0.24
Target 3	N/A	pIC50	↑	10,465	0%	0%	0.36
Target 4	N/A	pIC50	↑	10,624	12%	2%	0.20
Target 5	N/A	pEC50	↑	12,612	35%	0%	0.32
Target 6	N/A	pIC50	↑	13,093	0%	0%	0.31
Target 7	N/A	pIC50	↑	14,605	25%	0%	0.32
<b>ADME-T</b>							
ADME-T CYP 1	CYP3A4	pIC50	↓	12,875	61%	0%	0.11
ADME-T CYP 2	CYP2C9 (a)	pIC50	↓	12,876	63%	0%	0.11
ADME-T CYP 3	CYP2C9 (b)	pIC50	↓	14,062	58%	0%	0.12
ADME-T Perm.	Permeability	Log Perm.	↑	16,511	8%	0%	0.17
ADME-T Solub.	Solubility	Log Solub.	↑	47,607	5%	6%	0.21
ADME-T hERG	Toxicity	pIC50	↓	67,687	42%	0%	0.19
ADME-T LogD	Lipophilicity	LogD	≈ 2	88,114	0%	8%	0.10
ADME-T CLint	Metabolic Stability	Log $C_{int}^I$	↓	92,161	8%	6%	0.14

The “Goal” column in the table indicates the desired performance of a drug candidate on this assay (↑ for high, ↓ for low values). While LogD targets can be project-specific, a target value of 2 is generally assumed, as excessively high or low values are typically undesirable.

[33], providing a relevant benchmark for our work. An overview of the assays used in this study is presented in Table 1.

The target-based assays are crucial for ongoing drug development projects and, therefore, cannot be disclosed fully. They all model either the half-maximal inhibitory concentration (IC50) or the half-maximal effective concentration (EC50), as seen in Table 1. ADME-T assays are important in the pharmaceutical industry to test the pharmacokinetic profile and safety of drug candidates. Four of the included ADME-T assays also measure IC50. Three of the included ADME-T assays measure the inhibiting effects on two isoforms of Cytochrome P450 to check for potential drug–drug interactions. Two CYP2C9 assays are included, which differ in the methods used for measuring CYP inhibition. In the CYP2C9 (b) assay, a fluorescent substrate is used to determine CYP inhibition. The two other CYP assays, CYP2C9 (a) and CYP3A4 apply an updated protocol, where the degradation of a drug is measured using liquid chromatography-mass spectrometry (LC-MS). The solubility assay determines aqueous solubility in a high throughput fashion, starting from a dimethyl sulfoxide (DMSO) stock solution where the organic solvent is evaporated to have a solid sample. The result gives the maximum concentration of a compound in an aqueous solution at pH 7.4. The included toxicity assay determines the compound’s potential to inhibit the human Ether-à-go-go-Related Gene (hERG) potassium channel. This hERG inhibition is correlated to severe cardiac side effects by prolonging the time it takes the heart to contract and relax, *i.e.* the QT interval.

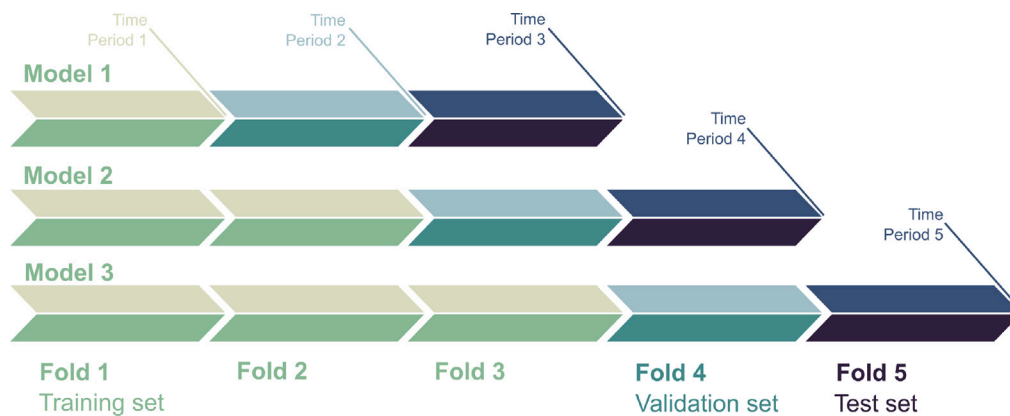
The remaining ADME-T assays model other important properties that influence the pharmacokinetics of a drug candidate. The permeability assay measures the flux of a compound across the Caco-2 cell layer. The permeability is measured in 1E-6 cm/s and is related to in vivo absorption. To obtain a compound’s lipophilicity, the logarithm of the distribution coefficient between octanol and aqueous phase at pH 7.4 is obtained (LogD). Finally, the Clint assay assesses the metabolic stability of a compound in  $\mu\text{L}/\text{min}/\text{million}$  cells. It measures how fast the compound is metabolized in rat hepatocytes and enables the predictions of in vivo hepatic clearance.

During the data preparation, the measurements of all target-based assays and the ADME-T assays measuring CYP and hERG inhibition were transformed to pIC50/pEC50 by taking the negative log base 10 of the measurements after transforming them to molar units. In addition, the measurements of the solubility, permeability, and Clint assays were transformed to log10-scale. The assay measuring lipophilicity was not modified, as it already comprised results on the log scale. Duplicated

measurements for molecular compounds in the data were then aggregated using the median of the result. The standard deviation for this aggregation was also used in the study to indicate the experimental error. For each assay, a control compound was available which had been tested many times more than any other compound. The standard deviation of the control compound was used as an indication of the general experimental error of the assay, also presented in Table 1. The control compounds were removed before modeling each assay.

All remaining molecular compounds were then encoded with RD-Kit (version 2023.03.3) [40] from SMILES strings [41] to Morgan Fingerprints (ECFP) [42] of size 1024 and radius 2. Other, more advanced ways to encode molecular compounds exist, such as the graph-based ChemProp model [7] and the pre-trained language-based CDDD model [43]. Models based on the resulting embeddings from these neural network encoders have been compared and shown improvements in prior work [23,26,30]. However, Dutschmann et al. [26] showed that fingerprints perform best in combination with random forest and are close second to CDDD in combination with neural networks. Additional prior work has similarly shown that neural networks perform better in combination with fingerprints compared to other machine learning methods [44,45]. While we employ the fingerprint representations in this study for simplicity and computational reasons, we strongly encourage considering state-of-the-art, learned representations before deploying our proposed methods in practical applications.

**Data censoring.** Given that the determination of exact experimental results is often connected with additional experimental efforts, a significant proportion of the data is provided with censoring. We follow the definition of censored regression proposed by Hüttel et al. [35]. Formally, if we define the true experimental label  $y_n^* \in \mathbb{R}$  of a molecular compound with features  $\mathbf{x}_n \in \mathbb{R}^d$  of length  $d$  from an assay, a censored label provides only an observation of a threshold  $z_n$  below or above which the true result lies. In the case of right-censoring this means that instead of observing  $y_n^*$  we observe  $y_n = \min(y_n^*, z_n)$  and for left-censoring we instead observe  $y_n = \max(y_n^*, z_n)$ . As such, we introduce a mask  $m_n$  which is 1 if  $y_n^* > z_n$  and  $-1$  if  $y_n^* < z_n$ , otherwise  $m_n = 0$ . As a result, we get the dataset  $\mathcal{D}_{\text{censored}} = \{\mathbf{x}_n, y_n, m_n\}_{n=1}^N$  for each assay with total number of labels  $N$ . In particular, note that the censoring threshold can vary for different labels which is different from fixed-value censoring where the threshold is fixed [46]. Furthermore, we assume that the true distribution and the censoring distribution are conditionally independent given the molecular compound’s features, *i.e.* that  $y_n^* \perp z_n | \mathbf{x}_n$ . Labels that are not censored are henceforth



**Fig. 1. Five-fold temporal split.** Illustrating the five folds created based on the date of the experiments 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 the second subsequent fold is used for testing. Note that the folds were created to have roughly equal size, not based on fixed time intervals.

referred to as observed labels. An observed dataset was also prepared for each assay to evaluate if the performance improved by including the censored labels. These datasets contain only the observed labels defined by  $\mathcal{D}_{\text{observed}} = \{\mathbf{x}_n, y_n, m_n\}_{\forall n: m_n=0}$ .

While prior work mentions the possibility of extending the definition of censored regression to include both right- and left-censoring at once [35,37], the majority of previous applications focus only on right-censoring. During the aggregation of duplicated measurements in the data preparation for this work, observed labels were always prioritized over censored labels. If no observed labels were available, the compounds were assigned the most common censored threshold among the potential duplicated measurements. Table 1 lists the percentage of censored labels for each assay used in this work divided between left- and right-censoring, according to the log-scale of the end-points. Note that while many assays include only left-censored labels, *i.e.*,  $y_n^* < z_n$ , two assays have a more balanced amount of censored labels, namely the ADME-T assays for solubility and metabolic stability. There is also one of the included assays that only have right-censored labels, namely the ADME-T assay for lipophilicity. Finally, two of the target-based assays, Target 3 and Target 6, do not have any censored labels. These two assays are useful in the model comparison to determine how all of the considered models compare when no censored labels are available.

**Temporal split.** Apart from the modeling of censored data, a key contribution of our work relates to evaluating the uncertainty quantification of molecular property prediction in a temporal setting. As such, we simulated 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. Fig. 1 illustrates the folds and resulting three settings used to evaluate and compare trained models as time evolves in this work, similar to how it was previously done by Svensson et al. [38]. The time intervals were chosen to create roughly equally sized folds regarding the number of observed labels.

By comparing the feature-space and label distributions between the three training folds and their respective validation and test sets, we identified various scenarios of in- versus out-of-distribution cases among the datasets. For the feature-space analysis, we created t-SNE projections of the molecular compounds in each dataset shown in Fig. S2 of the supplementary material. For the label-space analysis, we compared the distributions of all observed labels of each fold as shown in Fig. S1 of the supplementary material.

Generally, for the ADME-T assays no clear distribution shift is present in either the feature-space or the label-space over time. On the contrary, for the target-based assays both the feature-space and the label-space shift over time. These observed trends are naturally expected due to the fundamental difference between the two kinds of

assays. ADME-T assays include final drug candidates from all kinds of drug discovery projects that stem from a diverse chemical space. Target-based assays, on the other hand, are more focused on a given chemical space, which can shift over time as the focus of the project changes. This important distinction would not be possible without the temporal split proposed in this work, and the effects of it on model performance, as well as calibration of predicted uncertainty, is explored during the evaluation of the models.

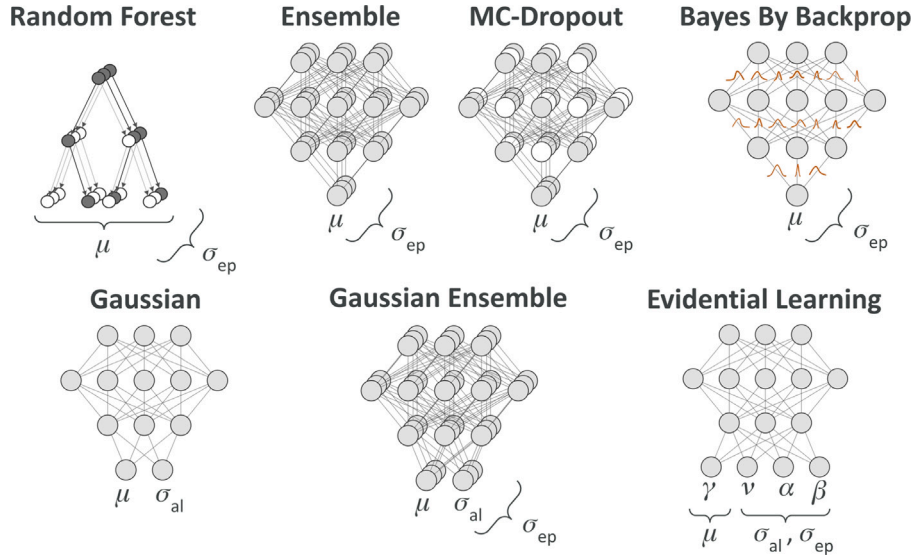
### 2.1. Models for uncertainty quantification

In this work, we propose five models that extend existing techniques to handle censored labels alongside two baseline models that only learn from observed labels without accounting for censorship. Fig. 2 gives an overview of all models and how they can estimate epistemic and/or aleatoric uncertainty. Ensemble-based and Bayesian approaches, such as Random Forests [11], ensembles of neural networks [13], Monte Carlo (MC) Dropout [12] and the approximation of a Bayesian neural network known as Bayes by Backprop [10] are commonly used to estimate epistemic uncertainty. For aleatoric uncertainty, two primary methods are available: the Gaussian mean-variance estimator [18] and evidential deep learning [20]. By creating an ensemble of the Gaussian model, we can also derive estimates of epistemic uncertainty. Evidential deep learning, on the other hand, simultaneously models four parameters that can be used to directly estimate aleatoric and epistemic uncertainty. We extend all of these models, except the Random Forest and Evidential model, to learn from both observed and censored data, while the two remaining approaches are used as baselines. A detailed description of how the model selection is performed for each model is provided in Table S1 of the supplementary material. While the data used in this work is proprietary and cannot be disclosed, the full methodology is available on GitHub at <https://github.com/MolecularAI/uq4dd>. Our framework is inspired by the design pattern proposed by Hartog et al. [47].

**Ensemble-based approaches.** In general, ensemble-based approaches are useful to model epistemic uncertainty by estimating the model variance. An ensemble is defined as a set of  $K$  base estimators  $f(\mathbf{x}_n)$ , for which the average of the individual base estimators' predictions is taken as the final prediction by the ensemble and the variance of the predictions as an estimate of the predictive uncertainty as follows,

$$\mu_n = \frac{1}{K} \sum_{k=1}^K f_k(\mathbf{x}_n), \quad \sigma_{\text{ep},n}^2 = \frac{1}{K} \sum_{k=1}^K (f_k(\mathbf{x}_n) - \mu_n)^2. \quad (1)$$

In a Bayesian framework, the uncertainty in model parameters  $\theta$  results in the predictive uncertainty of the model  $p(y_n|\mathbf{x}_n, \theta)$ . The true posterior distribution of the model parameters for a given dataset can be



**Fig. 2. Overview of models.** Illustrations of all models used in this study. The top row shows ensemble-based and Bayesian methods for which epistemic uncertainty can be obtained from the standard deviation in sampled predictions. All models in the bottom row produce aleatoric estimates of uncertainty. Additionally, epistemic uncertainty can be derived from the Gaussian Ensemble and the Evidential model.

described as  $p(\theta|D)$ , such that the predictive uncertainty of the Bayesian model average is defined by  $p(y_n|x_n, D) = \int_{\theta} p(y_n|x_n, \theta)p(\theta|D)d\theta$  [6,48]. As shown by Lakshminarayanan et al. [13] and Gal [48], the variance of ensemble predictions approximates the epistemic part of this true posterior distribution.

One of the simplest and most widely used traditional machine learning methods for ensemble-based uncertainty quantification takes the decision trees in a Random Forest as an ensemble [11]. We use the implementation of decision tree regressors from Scikit-learn [49] where each decision tree is trained independently using different sub-samples of the training data. The final prediction is obtained by averaging the predictions of all the trees and the variance constitutes the epistemic uncertainty estimate. Similarly, ensembles can be created by training multiple instances of the same neural network architecture with different random weight initialization [13]. We first consider the neural network architecture with a single output, which means that each base estimator makes a prediction such that the final prediction by the ensemble again is the average of the individual predictions and the variance estimates the epistemic uncertainty [50]. We use 50 randomly initialized and independently trained neural networks to create this ensemble and hence refer to it as Ensemble. All neural networks in this work are implemented and trained with PyTorch [51].

Furthermore, MC-Dropout is another technique that uses dropout in neural networks to approximate Bayesian inference [12]. Dropout is typically used when training neural networks to prevent overfitting by randomly setting a fraction of the neurons to zero. During MC-Dropout, multiple samples are drawn from the model using dropout during the inference phase. Similar to the ensemble-based approaches, the variance between these predictions provides an estimate of the epistemic uncertainty. We use the same neural network architecture from the Ensemble in our MC-Dropout model but train only one model and draw 500 samples from it during the inference step.

In a standard regression framework, the mentioned base estimators of these models are trained using the MSE loss. We propose to use the adapted CensoredMSE loss from Arany et al. [37] to allow the models to learn also from additional censored labels where available. As such, we define a one-sided error of a given prediction  $\mu_n$  and the true label  $y_n$  as follows,

$$\epsilon_n = \begin{cases} \min(z_n - \mu_n, 0), & \text{if } m_n = -1, \\ y_n - \mu_n, & \text{if } m_n = 0, \\ \max(z_n - \mu_n, 0), & \text{if } m_n = 1, \end{cases} \quad (2)$$

and get the CensoredMSE as  $\mathcal{L}^{\text{MSE}} = \frac{1}{N} \sum_{n=1}^N \epsilon_n^2$ . The interpretation of the proposed loss is that when the model correctly predicts a value above the threshold of a right-censored label or below the threshold of a left-censored label the error is set to zero. This reflects a perfect prediction and does not result in any weight updates due to the zero gradient of a constant.

*Bayes by backprop.* Blundell et al. [10] proposed a Bayesian version of neural networks, known as Bayes by Backprop, that approximates the posterior distribution of the network weights. In traditional neural networks, the weights are deterministic, meaning that a single set of values is learned during training. However, in Bayes by Backprop, the weights are modeled as random variables with associated probability distributions, typically Gaussian. This allows the network to maintain a distribution over the weights, reflecting the uncertainty about the true values of these parameters given the observed data.

To train the Bayes by Backprop model, the goal is to approximate the posterior distribution  $p(\mathbf{w}|D)$  of the weights  $\mathbf{w}$  given the training data  $D$ . Since exact Bayesian inference is intractable for neural networks, a variational approximation is used. The approach involves defining a variational distribution  $q(\mathbf{w}|\theta)$ , parameterized by  $\theta$ , which approximates the true posterior. The optimal variational parameters  $\theta$  are learned by minimizing the variational free energy, also known as the evidence lower bound (ELBO), which balances the trade-off between fitting the data well and staying close to the prior distribution  $p(\mathbf{w})$ . The loss function used to train the Bayes by Backprop model is therefore

$$\mathcal{L}^{\text{BNN}} = \text{KL}[q(\mathbf{w}|\theta) \parallel p(\mathbf{w})] - \mathbb{E}_{q(\mathbf{w}|\theta)}[\log p(D|\mathbf{w})], \quad (3)$$

where the first term is the Kullback–Leibler (KL) divergence between the variational distribution  $q(\mathbf{w}|\theta)$  and the prior distribution  $p(\mathbf{w})$  and the second term is the expected log-likelihood of the data under the variational distribution. The KL divergence penalizes deviations from the prior distribution while the expected log-likelihood encourages the model to find weight distributions that explain the observed data well.

In our implementation, we convert the same neural network used in the ensemble-based approaches into a Bayesian neural network and use the CensoredMSE as the log-likelihood to allow the Bayes by Backprop model to learn from censored labels. The KL divergence is unaffected by censored labels as it relates solely to the learned weight distribution with respect to the prior. During inference, we sample 500 sets of weights from the learned distribution, and the resulting ensemble of

models generates a distribution of predictions. As for the ensemble-based approaches, the spread of these predictions provides an estimate of the epistemic uncertainty, which reflects the model's uncertainty due to limited data or model capacity.

**Gaussian models.** So far, the base estimators of all described models produce only a prediction of the regression label. Next, we introduce the mean–variance–estimator originally proposed by Nix and Weigend [18]. More recent work by Choi et al. [19] suggests that the uncertainty estimates by this model are related to the heteroscedastic aleatoric uncertainty as it captures the noise inherent in data. As such, we define the mean–variance–estimator as a neural network with two outputs: one representing the predicted value  $\mu_n$  and the other representing the model's estimate of the aleatoric uncertainty  $\sigma_{\text{al},n}^2$ . The predicted variance is processed through a Softplus function to ensure non-negativity and stability. This model is trained using the NLL, which in the case of a presumed Gaussian distribution becomes the following,

$$\begin{aligned} \mathcal{L}^{\text{NLL}} &= \frac{1}{N} \sum_{n=1}^N -\log \varphi(y_n | \mathbf{x}_n, \theta) \\ &= \frac{1}{N} \sum_{n=1}^N \frac{1}{2} \log(\sigma_{\text{al},n}^2) + \frac{(y_n - \mu_n)^2}{2\sigma_{\text{al},n}^2} + \text{constant}, \end{aligned} \quad (4)$$

where  $\varphi(y_n | \mathbf{x}_n, \theta)$  is the probability density function for the Gaussian distribution with parameters  $\mu_n$  and  $\sigma_{\text{al},n}^2$  estimated by the model with parameters  $\theta$ . Note that the constant term  $\frac{1}{2} \log(2\pi)$  is omitted from the loss. Additionally, note that the variance is lower bound by a small number ( $\xi = 1e^{-6}$ ) for numerical stability as done in the PyTorch implementation of the Gaussian NLL [51].

We refer to this model as Gaussian going forward and additionally make a Gaussian Ensemble from 5 independently trained Gaussian models. A similar approach was used by Busk et al. [52], who suggests we get the aleatoric estimate from the expected value of all predicted variances and the epistemic estimate from the variance of the prediction. Furthermore, this decomposition of the sources of uncertainty is similar to how it is done for Gaussian Mixture Models [53,54]. Hub-schneider et al. [53] argues that this decomposition results in epistemic uncertainty estimates that vanish with more data, thus being reducible, in contrast to the aleatoric part which does not, thus being irreducible. As such, these results also correspond to the definition of epistemic and aleatoric sources of uncertainty [2,5,6].

To adapt this learning objective to be able to handle additional censored labels where available, we take inspiration from the Tobit model [55] used more recently in Bayesian active learning for survival analysis by Hüttel et al. [35]. The intuition behind this method is that for censored labels we use the integral of the probability density function that is within the correct censored region. Thus, for a left-censored label with threshold  $z_n$  the objective is to maximize the integral of  $\varphi(y_n | \mathbf{x}_n, \theta)$  below  $z_n$ , in other words to minimize the negative logarithm of the cumulative distribution function  $\Phi(z_n | \mathbf{x}_n, \theta)$ . On the contrary, for right-censored labels with threshold  $z_n$  the objective is to minimize the negative logarithm of the complement to the cumulative distribution function  $1 - \Phi(z_n | \mathbf{x}_n, \theta)$ . In summary, our proposed CensoredNLL is defined as,

$$\begin{aligned} \mathcal{L}^{\text{NLL}} &= -\frac{1}{N} \sum_{n=1}^N (1 - |m_n|) \log \varphi(y_n | \mathbf{x}_n, \theta) \\ &\quad + |m_n| \log \begin{cases} \Phi(z_n | \mathbf{x}_n, \theta), & \text{if } m_n = -1, \\ 1 - \Phi(z_n | \mathbf{x}_n, \theta), & \text{if } m_n = 1. \end{cases} \end{aligned} \quad (5)$$

**Evidential deep learning.** Another probabilistic approach to uncertainty estimation is through evidential learning using a neural network as proposed by Amini et al. [20]. This approach leverages the principles of evidential reasoning, where the network outputs parameters of a distribution that characterizes both aleatoric and epistemic uncertainty. In evidential deep learning, the neural network outputs four parameters

$\gamma_n \in \mathbb{R}, \nu_n > 0, \alpha_n > 1, \beta_n > 0$ , which together model the prediction  $\mu_n \sim \mathcal{N}(\gamma_n, \sigma_n^2 \nu_n^{-1})$  and variance  $\sigma_n^2 \sim \Gamma^{-1}(\alpha_n, \beta_n)$ . In the implementation, the outputs for  $\nu_n, \alpha_n$ , and  $\beta_n$  are all passed through a Softplus function, and then 1 is added to  $\alpha_n$ . Through this hierarchical model, the parameters express the epistemic uncertainty and aleatoric uncertainty estimates as,

$$\sigma_{\text{ep},n}^2 = \frac{\beta_n}{\nu_n(\alpha_n - 1)}, \quad \sigma_{\text{al},n}^2 = \frac{\beta_n}{\alpha_n - 1}. \quad (6)$$

The Evidential model is trained using the evidential loss, which optimizes the predicted parameter  $\gamma_n, \nu_n, \alpha_n, \beta_n$  to balance their fit to the data and simultaneous calibration of uncertainty estimates. The loss is defined as,

$$\mathcal{L}^{\text{EL}} = \frac{1}{N} \sum_{n=1}^N \frac{1}{2} \log \left( \frac{\pi}{\nu_n} \right) + \log \left( \frac{\Gamma(\alpha_n)}{\Gamma(\alpha_n + \frac{1}{2})} \right) - \alpha_n \log(2\beta_n(1 + \nu_n)) \quad (7)$$

$$+ \left( \alpha_n + \frac{1}{2} \right) \log((y_n - \gamma_n)^2 \nu_n + 2\beta_n(1 + \nu_n)) \quad (8)$$

$$+ \lambda |y_n - \gamma_n| \cdot (2\nu_n + \alpha_n), \quad (9)$$

where  $\Gamma$  is the Gamma distribution and  $\lambda$  is a hyperparameter determining the level of regularization. The first part of the evidential loss (Eq. (7)) corresponds to the log-likelihood of the predicted mean and variance of the normal-inverse Gamma distribution. The second part (Eq. (8)) optimizes the predictions toward the true target values  $y_n$  while also accounting for the uncertainty in the predictions. Finally, the last part (Eq. (9)) provides the regularization that penalizes the model for being overly confident, we use  $\lambda = 1$ .

While evidential learning has been used previously to estimate uncertainty in drug discovery-related tasks [27,56] its effectiveness at disentangling aleatoric and epistemic uncertainty has recently been questioned Juergens et al. [57], therefore we decided not to provide an extension of this model for cases with censored labels. Instead, the Evidential model is used solely as a baseline to benchmark the performances of the other approaches with that of this commonly used model.

## 2.2. Evaluation

The focus of this work is to determine the performance of models on realistic pharmaceutical data. As such all models are primarily evaluated on the full test sets containing observed and, if available, censored labels. The MSE loss is used to evaluate the performance of the predictions made by the models, in cases of censored labels we use the one-sided squared error from Eq. (2). Other metrics are required to evaluate the accuracy and calibration of the predicted uncertainties. We consider two ways to evaluate predicted uncertainty: one that evaluates only the accuracy or calibration of the uncertainty and another that evaluates 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 [8,25,53,54]. 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. Again, we adopt this evaluation technique for censored labels by using the one-sided squared error from Eq. (2).

Furthermore, the Gaussian NLL [58] and the Expected Normalized Calibration Error (ENCE) [59] are two global metrics that evaluate the intertwined predictive performance and calibration of uncertainties. We use the extended version of the Gaussian NLL from the Tobit model that can handle censored labels as defined in Eq. (5), and does not omit any constant terms. The ENCE metric is derived from the error-based calibration plot proposed by Levi et al. [59], which is made from a

binned representation of the root MSE (RMSE) and the root mean variance (RMV), *i.e.*, predicted uncertainty. In the case of censored labels, we again employ the one-sided squared error before taking the root of the mean. 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,

$$\text{RMSE}_b = \sqrt{\frac{1}{|b|} \sum_{i \in b} (y_i - \mu_i)^2}, \quad \text{RMV}_b = \sqrt{\frac{1}{|b|} \sum_{i \in b} \sigma_i^2}. \quad (10)$$

The bins are then summarized to give the ENCE metric as follows,

$$\text{ENCE} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{|\text{RMSE}_b - \text{RMV}_b|}{\text{RMV}_b}. \quad (11)$$

Other 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 [23,25–27]. However, this score has been criticized due to the stochasticity and unreliability of the result Rasmussen et al. [60]. Statistically, a data point with high predicted uncertainty can still result in a prediction with low error and vice versa. Additionally, it is non-trivial how to adjust this score for censored labels. Therefore, we discard the metric from our analysis.

### 3. Experiments

We start our experiments by evaluating the effect of training models with the additional partial information available in the real pharmaceutical assay-based setup in the form of censored regression labels. This comparison is conducted as an ablation study, comparing the difference in performance between models trained with and without censored data for each assay. Next, we provide an in-depth comparison between the models on each temporal setting of every assay. The model comparison evaluates the predictive performance and calibration separately for aleatoric and epistemic uncertainty estimates. Finally, we provide practical illustrations of the resulting uncertainty estimates for the best-performing model in terms of the aleatoric and epistemic parts respectively.

#### 3.1. Ablation study

We compare the performance of each model trained with and without the censored data to determine the impact of training uncertainty-aware machine learning models with additional partial information in the form of censored regression labels for molecular property prediction. Each test set is kept the same, containing all available data including the censored labels. However, the baseline models are trained only on the observed labels from the training datasets, *i.e.*  $\mathcal{D}_{\text{observed}}$ . Similarly, the validation set, used for hyper-parameter optimization and early stopping, only contains observed labels for the baseline models. We first evaluate the differences in NLL as a measurement of the intertwined predictive accuracy and calibration of uncertainty estimates. We define the test score from the model trained without censored data as  $\text{NLL}_{\text{observed}}$  and the test score from the model trained with censored data as  $\text{NLL}_{\text{censored}}$ . As NLL is minimized for more accurate predictions and better-calibrated uncertainty estimates, we define the evaluation metric as  $\Delta\text{NLL} = \text{NLL}_{\text{observed}} - \text{NLL}_{\text{censored}}$  such that a positive difference indicates that the performance was enhanced for the model trained with censored data. Next, we do the same analysis in terms of the MSE to evaluate the effect of including the censored labels on the predictive accuracy independent from the uncertainty estimation.

We evaluate the difference in MSE and NLL for each assay and temporal setting and test the statistical significance using a two-sided Mann–Whitney–Wilcoxon test [61]. Due to the vast number of statistical tests required, we apply Bonferroni correction to counteract the multiple comparisons problem [62]. As such, we use a significance level

$0.05/m$  where  $m$  is the number of datasets times the number of models in the comparison. If the difference is positive and significant ( $p < 0.05/m$ ) the model trained with censored data is deemed best. Similarly, if the difference is negative and significant ( $p < 0.05/m$ ) the model trained without censored data is deemed best. The results of the ablation study are summarized in Fig. 3. In addition, the results on test sets containing only observed labels are provided in Fig. S3 of the supplementary material. The magnitude of all underlying scores for Fig. 3 and Fig. S3 are also supplied in Fig. S4 and S5 of the supplementary material.

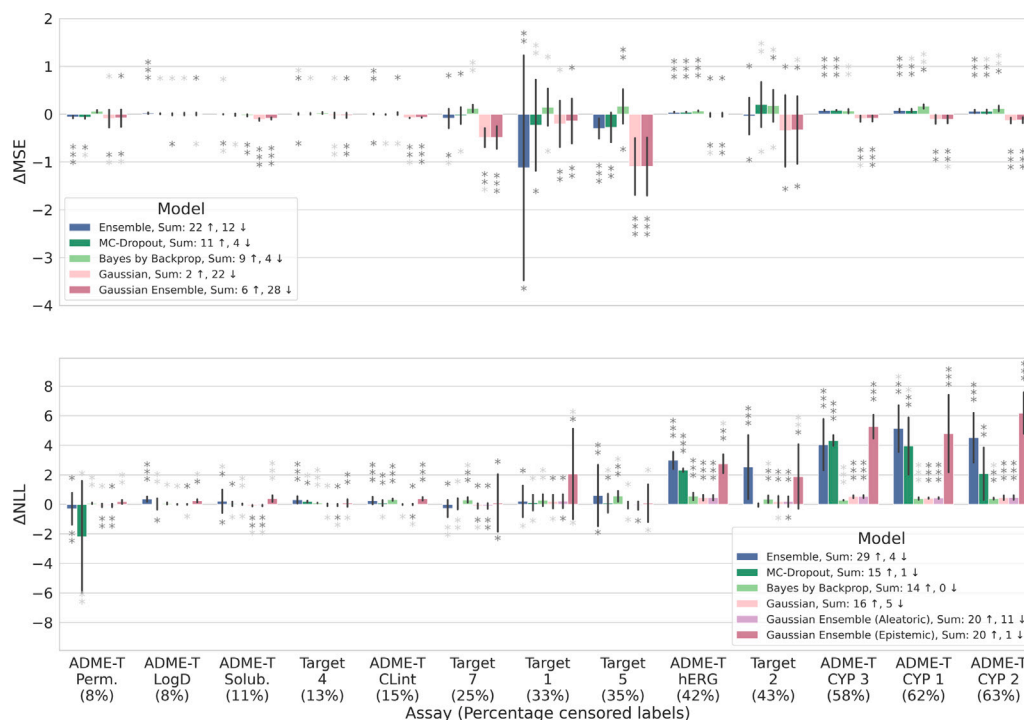
Throughout the ablation study, the assays are ordered according to the percentage of censored labels available in the full dataset. Note that the two assays without any available censored labels, Target 3 and Target 6, have been excluded. Each star in Fig. 3 and Fig. S3 indicates that the model trained with or without censored data was significantly better in one of the three temporal settings. As such, each bar can have a maximum of three stars. If a star is located above the bar, the model trained with censored data is the best, and vice versa if it is located below the bar. The dark gray stars mark the significance under Bonferroni correction, but the uncorrected significances are also shown as the light gray stars. The total number of significant cases under Bonferroni correction is summarized in the legend for each model. Note that the total number of datasets in this ablation study is 39, including the 13 assays and 3 temporal settings. The error bars illustrate the standard deviation between the temporal settings.

The  $\Delta\text{NLL}$  results in the bottom panel of Fig. 3 show predominately enhanced performances when the models are trained with the censored regression labels. Particularly, the Ensemble and Gaussian Ensemble models show significant enhancements in a majority of the datasets when trained with censored labels. All other models are enhanced or comparable in all but a few datasets. By instead considering each assay at a time, we see that for all assays apart from ADME-T Perm. and ADME-T Solub. more models are improved by the censored labels than hurt by them. The two exceptions are among the assays with the smallest percentages of censored labels available. In this case, the Bonferroni correction could be considered too restrictive given the much smaller number of comparisons in each analysis. In terms of the uncorrected significance, 8 out of the 13 assays see a strong majority of enhanced performances with the censored labels.

Furthermore, the magnitudes of the  $\Delta\text{NLL}$  scores in Fig. 3 generally become larger for assays with higher proportions of censored labels, with a more substantial shift occurring around 35%. A more detailed analysis of this trend for each model respectively is seen in Fig. S5 of the supporting material. There, we see that for Ensemble and MC-Dropout the improvements grow for assays with  $> 35\%$  censored labels. For these assays, the NLL is on average enhanced by 52% for the Ensemble model and 48% for the MC-Dropout. Similarly, the Gaussian model and both uncertainty sources of the Gaussian Ensemble model also see substantial enhancements for assays with  $> 35\%$  censored labels. For the Gaussian model and the aleatoric part of the Gaussian Ensemble, the NLL is enhanced by 34% on average for these assays. The NLL for the epistemic part of the Gaussian Ensemble for these assays on the other hand is enhanced by 56% on average. Finally, the Bayes by Backprop model is consistently improved already for assays with  $> 13\%$  censored labels although by a slightly lower average enhancement of 24%.

While the NLL provides an overview of the overall performance of these uncertainty estimation methods, it is also interesting to evaluate the effect of the censored labels on the predictive accuracy on its own. The ablation study in terms of  $\Delta\text{MSE}$  is presented in the top panel of Fig. 3. Note that the Gaussian Ensemble only occurs once in this result, as the different uncertainty estimates do not impact the MSE. There appears to be a greater difference between the effects of the censored labels on the two loss functions in terms of predictive accuracy compared to the NLL. Recall that the Ensemble, MC-Dropout, and Bayes by Backprop models are adapted using the one-sided squared loss, whereas the two Gaussian models use the Tobit loss with similarities to the NLL. Regarding the significant differences, the results for Ensemble,





**Fig. 3. Ablation study.** Difference between MSE (top) and NLL (bottom) for models trained with and without censored labels. The error bars aggregate the three temporal settings. For each temporal setting, a star indicates whether one of the models trained with or without censored labels was significantly better. As such, each bar can have a maximum of three stars. Stars above a bar indicate that the censored model was significantly better, with Bonferroni correction ( $p < 0.05/m$ ) in dark gray and without correction ( $p < 0.05$ ) in light gray. In contrast, the stars below a bar indicate the opposite. The legend summarizes the number of Bonferroni corrected significant differences per model.

MC-Dropout, and Bayes by Backprop are similar in terms of  $\Delta$ MSE as they were in terms of  $\Delta$ NLL. The Ensemble model is enhanced for a majority of datasets, and the other two are enhanced or comparable in all but a few datasets. The improvements are however smaller in magnitude, as seen especially in Fig. S4 of the supplementary material. Proportionally, the enhancements of the MSE are still, on average, around 20% for these models on the assays with  $> 35\%$  censored labels. The combined results of these models in terms of MSE and NLL, indicate that the inclusion of censored labels during the training procedure mostly impacts the uncertainty estimation abilities of the models rather than predictive accuracy.

On the contrary, the predictive accuracy of the Gaussian models trained with the Tobit loss is mostly not improved by including the censored labels. The magnitude of the loss in predictive accuracy for the Gaussian models is largest for Target assays with  $\geq 25\%$  censored labels included for the adapted version of the models, as seen in Fig. S4 of the supplementary material. The proportional losses in MSE are comparable to the enhancements in NLL, consisting of around  $-30\%$  on average for all assays. This is an unfortunate drawback of our proposed adaption of the Gaussian models using the Tobit loss and should be considered when using the models in practice. However, as these machine learning approaches are most often used in applications for high-throughput screening in the early stages of the drug discovery pipeline, the ranking of molecular compounds is usually considered more critical than the exact prediction of a given property. As such, we believe the enhancement of the uncertainty estimates can outweigh the loss in predictive accuracy. Furthermore, it is possible that a weighting scheme included in the Tobit loss functions could make the Gaussian models trained with censored labels favorable also in terms of predictive accuracy.

As a final step in this ablation study, we analyze the impact of training the models with and without the censored labels on solely the observed test labels. Similar to the analysis of the  $\Delta$ MSE, this is useful to determine which impact the adapted training procedure has on censored versus observed test labels. While this analysis is

interesting, we believe it is secondary to the analysis performed on the full set of test labels. The reason is that the full test set is the only one that simulates the real scenario in the drug discovery process. In reality, there would be no way of excluding the compounds that in hindsight would result in censored labels when tested in the lab. Fig. S3 in the supplementary material shows the  $\Delta$ MSE and  $\Delta$ NLL on only observed test labels (i.e. excluding the censored labels). These scores are mostly not improved by including censored labels during training. The performance of the Bayes by Backprop model is mostly not significantly affected by the inclusion of censored labels during its training. The predictive accuracy of the Ensemble and MC-Dropout models is typically worse when censored labels are included while the impact on the NLL is more varying. The performance of the Gaussian models is almost always negatively impacted by the inclusion of censored labels during training.

By considering the full datasets, including the censored labels, as the true underlying data distributions, we can conclude that the results on the observed labels only illustrate a crucial point. Namely, that the censored labels make up a critical part of the distributions, such that machine learning models trained with or without them do not transfer to the test set corresponding to the opposite case. Models trained without censored labels do not perform well on test sets with censored labels, and models trained on full datasets do not perform well on test sets without censored labels. We believe this is a natural consequence of the vast imbalance in the censored labels. As seen in Table 1 most assays only include censoring in one direction. Only two of the assays explored in this study have a more balanced occurrence of left- and right-censored labels, namely ADME-T Solub. and ADME-T CLint. However, the total proportion of censored labels for these assays is not large enough to significantly impact the difference in performance between the models trained with and without the censored labels as discussed above. While the reduced performance of the models trained with censored labels on observed test labels is unfortunate, we argue that it is so far an unavoidable side-effect of the optimal performance achieved on true test sets overall. In the future, other



**Fig. 4. Predictive accuracy.** Comparing the predictive accuracy of all models in terms of MSE, aggregated over 10 experiments. For each dataset, the best model in terms of mean MSE is marked with a star together with any other models not statistically worse based on a one-sided Mann–Whitney–Wilcoxon test with Bonferroni correction. Apart from the Random Forest and Evidential models, all other models are trained with censored labels where available.

ways of incorporating censored training labels could be explored, such as using the absolute values of censored labels without the additional censored information.

All in all, this ablation study has shown that the inclusion of censored labels during training of uncertainty-aware methods can in some cases negatively impact the predictive accuracy as well as the performance on observed labels separately. However, it has also shown that including censored labels during machine learning is important to accurately estimate well-calibrated uncertainties in real pharmaceutical scenarios, where censoring occurs and cannot be avoided. The enhancements are especially prominent when the proportion of censored labels makes up roughly one-third or more of all available data for a given assay. Given the overall majority of improvements in the NLL seen with censored labels included during training on the full test sets, we recommend the community utilize this data in the future during QSAR modeling for regression tasks. Furthermore, we continue to compare these models to each other and the additional baselines in the following section.

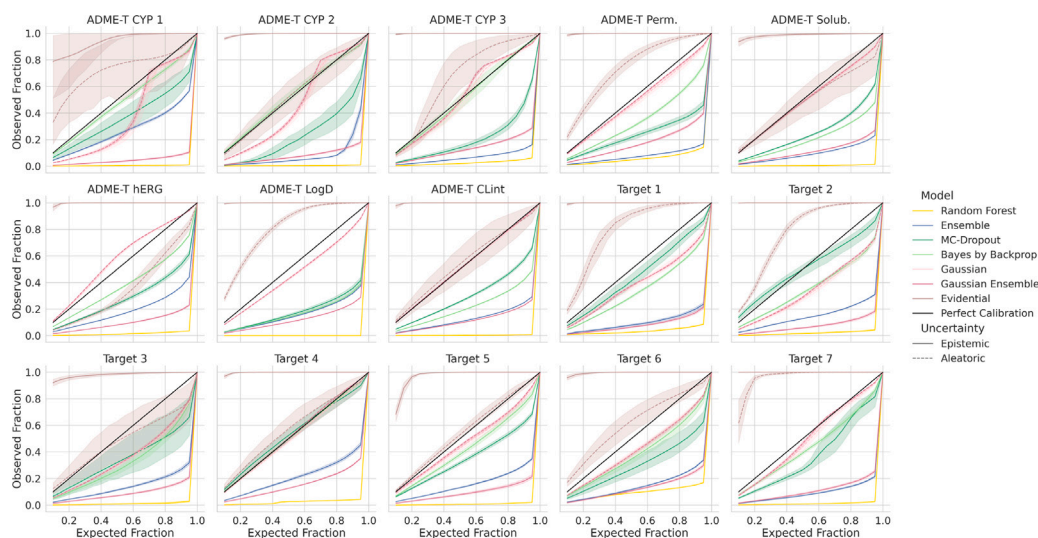
### 3.2. Model comparison

We compare the proposed models trained with all available data, including censored labels, on each dataset by obtaining their predictive accuracy in terms of MSE, their local calibration in terms of the confidence-based calibration curves, and their overall performance in terms of NLL and ENCE. Where uncertainty estimates are evaluated, we separate the aleatoric and epistemic parts. For each dataset, the three temporal settings are evaluated separately such that the results can be compared between different sizes of training sets, and as time evolves. For instance, it is relevant to know if conclusions drawn from the first temporal setting also hold throughout time, *i.e.* on the following temporal settings. In this analysis, the assays are first ordered by category and then according to the overall size of the dataset. For the scores, MSE, NLL, and ENCE the model with the best average score over 10 repeated experiments is marked with a star. Additionally, any other performances not significantly different from the best model according to a one-sided Mann–Whitney–Wilcoxon test are also marked with stars [61]. Similar to the ablation study, we use Bonferroni correction to counteract the multiple comparisons problem [62]. Here the significance level is  $p < 0.05/m$ , where  $m$  is the number of comparisons per dataset. The two target-based assays without censored labels are included. Apart from the proposed extended models, which are trained

with censored labels, two additional baselines are included: a Random Forest model and a model trained with evidential deep learning. These models are trained without censored labels as explained in the Methods section.

**Predictive accuracy.** The predictive performances of the models are presented in Fig. 4 in terms of MSE. An initial observation is that the MSE is generally lower for the models trained on the ADME-T assays. This result is expected given the diversity and size of the ADME-T assays, compared to the target-based assays. More data means that the models can learn to generalize better and the lack of shifts in both the feature space and the label space provides an easier prediction task. For seven out of the eight ADME-T assays, the Ensemble model is the best or among the best for all temporal settings. In a couple of instances, the MC-Dropout model is comparable, and for the ADME-T Perm. assay the Bayes by Backprop model significantly outperforms all other models. In the target-based assays, there is a greater variation in the best-performing models. For three target-based assays, the Bayes by Backprop model is consistently among the best across all three temporal settings, and in one target-based assay, the Ensemble model is consistently among the best. However, for the remaining three assays Target 1, Target 2, and Target 7 the best-performing model changes drastically over time. This observation is of great importance, as it illustrates that, typically, the conclusions about the best-performing model drawn early in the development of an ADME-T assay are reliable and still hold throughout the later development of the assay. On the other hand, as target-based assays might significantly shift in chemical space over time the same is not always true for the model comparison on these kinds of assays. For target-based assays, a new model comparison might be needed later on in the development of the assays to keep the best-performing model up-to-date.

For the most part, the Random Forest and Evidential models, which are trained without censored labels, perform poorly compared to the best-performing models. This is also true for the Evidential model on the two assays that do not have any censored labels, meaning that neither of the other models has the advantage of being trained on more data. We interpret this as an indication that the poor performance of the Evidential model compared to the other models is not primarily related to the fact that the other models are trained with censored labels. The Random Forest model, however, does reach comparable results to the Bayes by Backprop model for some of the temporal settings on the two assays, but it is never significantly better than all other models on any datasets. As such, it might be interesting to try to extend the



**Fig. 5. Calibration curves.** Full calibration curves for all uncertainty estimates on the third temporal setting containing three folds in the training set, aggregated over 10 experiments. The black line in each panel illustrates what a perfectly calibrated model would look like. Apart from the Random Forest and Evidential models, all other models are trained with censored labels.

Random Forest model to also handle censored labels in the future. Note, that this should be possible similar to how the one-sided square loss was used for models originally trained using MSE. However, it would require lower-level changes to the implementations in Scikit-learn [49] which we deem out of the scope of this study given that it does not significantly outperform any of the other proposed methods herein. We do not believe it would be possible to simply include the censored labels as they are during the training of the Random Forest, given the non-fixed window of censoring per assay.

Considering all datasets, *i.e.* assays and their respective temporal settings, the Ensemble model performs best overall, by being among the best-performing models for 28 out of the 45 dataset instances. The only approach that matches the performance of the Ensemble model is the Bayes by Backprop model which is among the best for 19 dataset instances. From the remaining models, the MC-Dropout model is only among the best-performing models for 9 dataset instances, usually together with the Ensemble model. The Evidential model and the Random Forest model are among the best together with the Bayes by Backprop model in 7 versus 5 cases each. When training with two folds of the Target 1 assay, the two Gaussian models perform best together with the Ensemble model. However, this is the only instance where these models are among the best.

It is reasonable and compatible with previous conclusions, that the Ensemble model achieves the highest predictive accuracy, *i.e.* lowest MSE [63]. The reason is that this model contains the highest number of individually trained models. As such, the consensus in the predictions by the base models can reach a higher accuracy in cases where the problem is solvable by the available data. In comparison to the Gaussian Ensemble, the regular Ensemble has a slight advantage in terms of solely predictive accuracy because it is only trained to optimize the MSE. The Gaussian model simultaneously needs to optimize the calibration of the predicted uncertainty and thus might need to sacrifice some predictive accuracy if they are miss-aligned. For the Random Forest, the number of decision trees was optimized for each dataset in the model selection, as such in some cases the ensemble contains more than 50 trees. However, as the Random Forest is less complex than a neural network it is still reasonable that the neural network ensemble performs better. Out of the remaining models, the Evidential model relies on a single trained base model, but the Bayesian models instead learn an infinite set of model parameters.

Overall, our results show the Ensemble as the best-performing alternative considering only the predictive accuracy of the considered

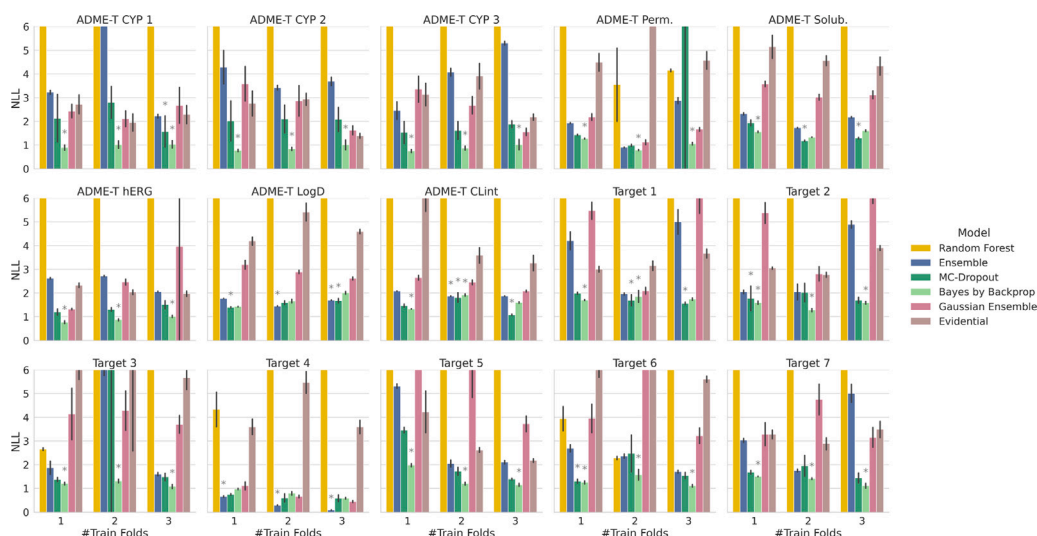
models. However, for practical applications, the computational cost of this model also needs to be taken into consideration. It is substantially more demanding to train 50 individual models for the Ensemble compared to the single model architecture used in the Bayesian approaches. As such, it is worth keeping in mind, that when Bayes by Backprop performs comparable to the Ensemble, it will also be significantly faster to train.

**Calibration.** Next, we evaluate the local calibration of the uncertainty estimates of each model in terms of the confidence-based calibration curves described in the Methods section shown in Fig. 5. We focus this analysis on the third and biggest temporal setting. The results from the two earlier settings are provided in Fig. S8 of the supplementary material. For the most part, we see that all aleatoric estimates are much better calibrated than many of the epistemic estimates. This is an interesting observation as it holds for models that produce both epistemic and aleatoric uncertainty estimates including the Gaussian Ensemble and the Evidential model, as well as the simple Gaussian model which only produces aleatoric estimates. The observation is true for ADME-T and target-based assays alike and thus irrespective of distribution shifts or dataset size.

Among the epistemic uncertainty estimates, a clear trend shows that the Evidential model is grossly under-confident, whereas many of the ensemble-based models are over-confident, especially the Random Forest model. This means that the Evidential model predicts CIs that are far too wide and, on the contrary, that the ensemble-based models predict very narrow CIs. For the three ADME-T CYP assays, the Bayes by Backprop estimates of the epistemic uncertainty are well-calibrated. Similarly, for many of the target-based assays, both the Bayes by Backprop model and the MC-Dropout model produce well-calibrated estimates of the epistemic uncertainty. These observations corroborate previously reported findings that ensembles tend to be under-calibrated in terms of the epistemic uncertainty compared to Bayesian approaches [53].

As pointed out by Rasmussen et al. [60], these confidence-based calibration curves can in some cases be misleading in determining whether predicted uncertainties accurately depict small versus large errors. Perfect calibration could in principle be achieved by a model that always predicts the same level of uncertainty. As such, Levi et al. [59] propose to additionally measure the dispersion of the predicted uncertainties using the coefficient of variation,  $c_v$ , as follows

$$c_v = \frac{\sqrt{\frac{\sum_{n=1}^N (\sigma_n - \mu_\sigma)^2}{N-1}}}{\mu_\sigma}, \quad \text{where } \mu_\sigma = \frac{1}{N} \sum_{n=1}^N \sigma_n. \quad (12)$$



**Fig. 6. Combined accuracy of uncertainty estimation and predictive performance.** Comparing the NLL of all epistemic uncertainty estimating models, aggregated over 10 experiments. For each dataset, the best model in terms of average NLL is marked with a star together with any other models not statistically worse based on a one-sided Mann-Whitney-Wilcoxon test with Bonferroni correction. Apart from the Random Forest and Evidential models, all other models are trained with censored labels where available.

We provide these values for all models and datasets in Fig. S9 of the supplementary material. According to these results, we deem that all of the best-performing models exhibit sufficient dispersion in their uncertainty estimates. There are some cases where the dispersion is very low for the Random Forest model, but this is only the case on some of the very smallest training sets where it is understandable that the model struggled to learn regardless. Additionally, many previous studies on uncertainty estimation have shown that the calibration can typically be improved by recalibrating the uncertainties [25,60,64]. It is, however, not straightforward how recalibration could be applied in the context of censored labels. Thus, we leave this to be explored in future work.

**Overall performance.** In the final step of our model comparison, we compare the models using the global scores NLL and ENCE which take both the predictive accuracy and the calibration of uncertainties into account. The results for NLL of the epistemic estimates are shown in Fig. 6 while the results for NLL of the aleatoric estimates are provided in Fig. S6 of the supplementary material. Additionally, the ENCE scores are provided in Fig. S7 of the supplementary material, where similar conclusions can be observed as for the NLL scores. The NLL scores are highly varying for some of the poorly performing models, and therefore the plots have been cut to show the NLL below 2.5 in the case of the aleatoric models and 6 for the epistemic models. This was done for readability and does not hide any information about the best-performing models.

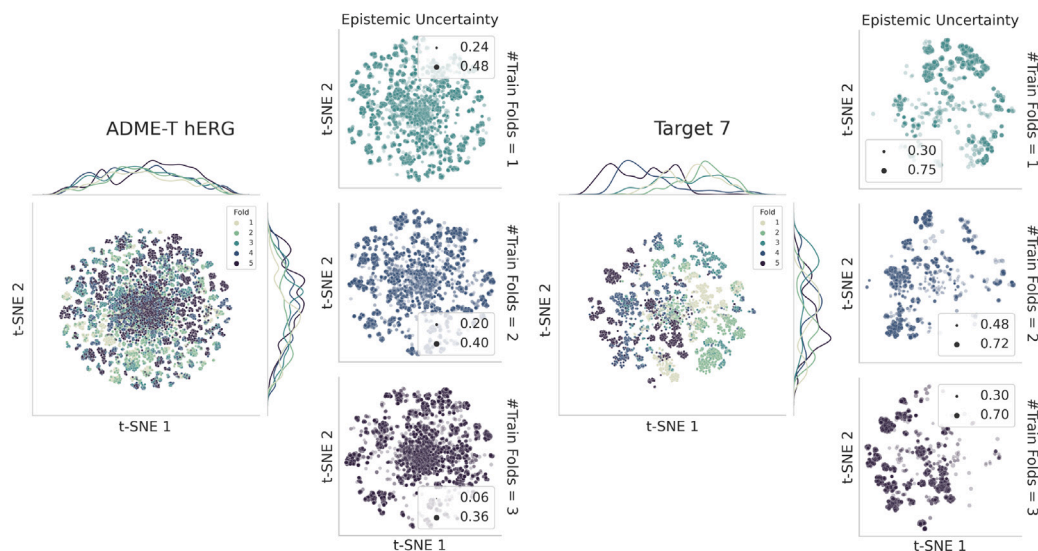
In terms of epistemic uncertainty, the Bayes by Backprop model is overwhelmingly best in terms of NLL, with 35 out of 45 best-performing scores. Only on the Target 4 assays, does the Ensemble model outperform Bayes by Backprop consistently throughout time. On 7 out of all 15 assays, the Bayes by Backprop model outperforms all other models entirely in all temporal settings. For an additional 4 assays, Bayes by Backprop is consistently among the best-performing models throughout time. This observation means that in terms of the NLL, the conclusions drawn early for an assay from both categories can oftentimes be trusted to hold throughout time. We believe that this robustness for the target-based assays, which exhibit distribution shifts, speaks to the reliability of the epistemic uncertainty estimates in accounting for distribution shifts where present. As such, the best-performing model in terms of predictive accuracy might change over time, only due to differences in how well the models handle the distribution shifts. Still, the overall best-performing model is reliable for all assay types irrespective of distribution shifts.

Similarly to what was observed in the MSE, Fig. 6 shows that the NLL scores are generally slightly lower for the ADMET-T assays compared to the target-based assays. Target 4 is again an outlier with quite low NLL scores for the best-performing models compared to the other target-based assays. Recall from Eq. (5) that the NLL includes a term with the squared error. Thus, it is likely that this trend arises from that source. In the ENCE scores presented in Fig. S7 of the supplementary material, the same trend can be partially observed but it is not as prominent. Apart from this, the Bayes by Backprop model also outperforms all other models in all temporal settings for 8 assays in terms of ENCE. Contrary to the NLL, the Ensemble is never among the best-performing models in terms of ENCE, but the Evidential model is consistently among the best for 3 assays. Regardless of which model is the best, the one that is best in the first temporal setting is in most cases also best in the subsequent temporal settings, also in terms of ENCE.

For the aleatoric uncertainty estimates, shown in Fig. S6 of the supplementary material, the Evidential model is significantly outperformed by either or both of the Gaussian models in all but 3 of the 24 ADMET-T dataset instances. Among the target-based assays, there are more times that the Evidential model is comparable to or outperforms the Gaussian models. However, the Target 5 assay is the only one where the Evidential model is consistently among the best through time. On the other hand, the Gaussian Ensemble is the single best or among the best for the vast majority of datasets, namely in 39 out of 45. In the case of the NLL for the aleatoric uncertainty estimates, like for the MSE, the best-performing model is almost exclusively robust throughout time for the ADMET-T assays but not always robust for the target-based assays. The same conclusions can be seen in terms of the ENCE score. Note also that the NLL and ENCE scores for the aleatoric uncertainty estimates are mostly lower than for the epistemic uncertainty estimates, similar to the results seen in the calibration curves.

### 3.3. Case study

Finally, we perform a case study on one example assay from each category to explore the practical implications and usefulness of the epistemic and aleatoric estimates respectively. In the case study of aleatoric uncertainty, we compare the predicted uncertainty estimates with ground truth experimental error in the form of the standard deviation between duplicated experiments. For this analysis, we cannot include censored labels due to the aggregation strategy described in the Methods section. Additionally, we exclude experiments repeated less



**Fig. 7. Epistemic estimates by Bayes by Backprop.** A practical illustration of the epistemic uncertainty estimates by the Bayes by Backprop model on two assays: ADME-T hERG without clear distribution shifts over time, and Target 7 with prominent distribution shifts over time.

than three times and those with no variation reported in the data. We chose the assays with the most available compounds with duplicated experiments, e.g. ADME-T hERG and Target 6. For the epistemic case study, we instead pick Target 7 as a target-based assay due to its particularly challenging distribution shift. Similarly to the aleatoric case study, we compare the results with the ADME-T hERG assay, which has no visible distribution shift in either the label or feature space as seen in Fig. S1 and Fig. S2 of the supplementary material respectively.

This case study is solely possible due to the temporal split and the additional information available through the internal pharmaceutical data, which are generally not accessible or trustworthy in public datasets. Similar remarks were made by Sheridan [32] and point to the necessity of openness from the pharmaceutical industry to learn from each other and push the field forward.

**Epistemic uncertainty quantification.** For the practical case study of the epistemic uncertainty estimates, we illustrate the epistemic uncertainty estimates averaged over the 10 repeated experiments of the Bayes by Backprop model on the ADME-T hERG and the Target 7 assays. The Bayes by Backprop model is chosen as it is significantly the best model in terms of NLL across all temporal settings of both of these example assays, as shown in Fig. 6. The case study compares the epistemic uncertainty between different regions of the t-SNE projections of each test set, as illustrated in the rightmost part of each panel in Fig. 7. It also relates this to the full t-SNE projection including the folds used for training and validation, as seen in the leftmost parts of the same figure. Note, that this only accounts for distributional shifts in a linear approximation of the feature space. As the fundamental nature of the feature space is non-linear, this analysis cannot cover these non-linearities. Similarly, we cannot evaluate any parts of the epistemic uncertainty related to other factors than the distribution shift in this analysis. The other factors might include epistemic uncertainty that relates to the choice of model architecture or training procedures. We have to rely on the metrics used in the model comparison section above to account for these factors of epistemic uncertainty.

In the full t-SNE projections of each dataset containing all five folds, we see the trend described in the Methods section. The chemical space of the ADME-T hERG assay is highly diverse, already in the first fold, and does not shift substantially over time. However, some clusters are formed in the outer edges of the chemical space, where the different folds are well separated. For the Target 7 assay on the other hand, there is a clear shift in the chemical space from the right side of the plot representing the initial measurements to the left side showing the

later folds. Similar trends are also present in the label distributions as seen in Fig. S1 of the supplementary material. Furthermore, the label distribution for the Target 7 assays does not shift continuously over time. Instead, it shifts greatly toward higher pIC50 values in the second fold and then back toward lower values in the last two folds. The remaining three plots to the right in each panel of Fig. 7 show the t-SNE projections of each test set, i.e. folds 3, 4, and 5, separately. Here, the size of the data points is determined by the averaged epistemic uncertainty predicted by the Bayes by Backprop model trained on the three temporal settings respectively, i.e. with an increasing number of training folds. Note that the legends of these plots detail the respective minimum and maximum predicted uncertainties on the given test set.

For the ADME-T hERG assay, only minor differences can be observed between compounds positioned in various regions of the feature space. This is expected given the lack of overall distribution shifts between the folds. Additionally, we evaluate the span of uncertainty estimates, according to the minimum and maximum values listed in the legend of each plot. The third temporal setting has the largest span of 0.3 compared to 0.24 and 0.2 for the first two settings. This indicates that the model is better at separating the compounds in the final temporal setting compared to the previous one, likely due to the increased size of the training data. Another effect of the increasing training data can be seen in the difference in the maximum predicted uncertainty between the folds. For the model trained on three folds, the maximum is 0.36, which is smaller than the values seen for the models trained on less data, which are 0.4 for the model trained on two folds and 0.48 for the model trained on one fold. This is a direct and expected behavior of the epistemic uncertainty, that more data results in more certain models i.e. lower predicted uncertainties. In the case of the target-based assay, we note that the maximum estimated uncertainty on each test set is very high, between 0.7–0.75. This can be a result of the general distribution shift present between every fold or the overall small dataset. In both cases overall high uncertainty is to be expected.

Based on the distribution shift present in the feature space, and considering that epistemic uncertainty should account for distributional shifts, we expect that distribution shifts are reflected in the predicted epistemic uncertainties. As such, our analysis provides empirical evidence to support this claim, and it illustrates that the uncertainty estimates cover additional sources of uncertainty related to the model itself, such as poor generalization in cases when training data is limited. It is important to understand all sources of uncertainty when basing future critical decisions on model predictions, such as in drug discovery.

Considering the cases addressed in this section of our study, we can deduce practical suggestions on how the identified sources can impact the continued drug discovery process. If poor generalization is determined, indicated by overall high uncertainty estimates and low predictive performance of a model, the model needs to be retrained with more data before deployment. Another alternative is to reconsider the choice of model, but our temporal study shows that the Bayes by Backprop model remains the best choice in the future, despite the addition of more data. When distribution shifts are identified, such as in the target-based assay, further data exploration in chemical spaces with high uncertainty estimates is needed before deployment.

Further research is necessary to disentangle the sources of epistemic uncertainty, including distribution shifts and other model-related sources. One alternative approach would be to quantify the distribution shift using other means, either with distance-based approaches, such as the average Tanimoto or cosine similarity [15] between an inference compound and compounds in the training set, or the interpretable method proposed by Kulinski and Inouye [65]. Additionally, more advanced pre-training procedures can be used, that are trained to incorporate distribution shifts more effectively [66]. 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.

**Aleatoric uncertainty quantification.** To examine the aleatoric estimates by the best-performing model from the model comparison above, we need a way of quantifying the true experimental error. As such, we consider the variation between duplicated measurements of a single compound on the assay to provide this information. However, note that this information was not provided to the models during training because duplicated measurements were aggregated as part of the data preprocessing described in the Methods section. This procedure of aggregation is common when applying machine learning to molecular property prediction, but an important question is if the available methods to estimate aleatoric uncertainty can learn the experimental error from such preprocessed data.

One potential issue with the described estimates of the true experimental error is that the number of duplicated measurements varies greatly between compounds. Some compounds are tested many times whereas many are tested only a handful of times or only once. The compounds for which there is only one measurement do not give any indication about the experimental error and ones where only a few duplicates are available might not be as reliable as those where the experiment was repeated many times. For this reason, we consider only compounds from the test sets that have more than 2 measurements and ones where the variation was greater than exactly zero.

After filtering for duplicated experiments in the available data, the ADME-T hERG assay contains 712 compounds with duplicated measurements from the test sets of all three temporal settings combined. The Target 6 assay similarly contains 1712 duplicated experiments in total between the three test sets. In this analysis, we focus on the aleatoric estimates averaged over the 10 repeated experiments of the Gaussian Ensemble. In the previous experiments of this study, this model emerged as the significantly best or top-performing model across all temporal settings of the ADME-T hERG and Target 6 assay. Fig. 8 illustrates the relationship between the predicted aleatoric estimates and the available experimental errors, colored by the three test sets.

In general, we cannot necessarily expect the model to learn the correct scale of the experimental errors since the experimental error is not included in the training data. Instead, we are looking to evaluate if any upward trends can be observed, such that compounds with larger experimental error also prompt greater predicted aleatoric uncertainty. However, no such trend can be observed for any of the models or test sets on the two assays. Given that the aleatoric estimates are still well-calibrated according to the observed errors, as seen in the calibration curves of Fig. 5 and the overall scores shown in Fig. S6

and Fig. S7 of the supplementary material, this could indicate that the learned aleatoric uncertainties reflect something other than the experimental error examined here. The model may be able to learn trends in the label space given similar compounds in the feature space. For instance, activity cliffs can exist where very similar compounds prompt significantly different responses in an assay [67]. Such relationships could be explored in future research, but from our analysis, it would also be interesting to try modeling these molecular properties without aggregating duplicated measurements.

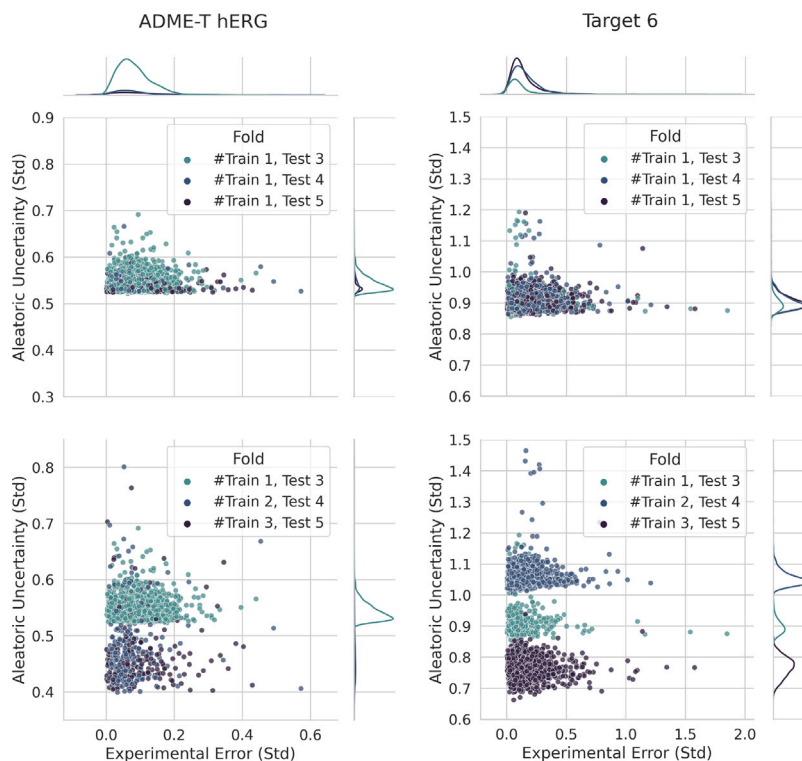
Additionally, we compare the scale of the predicted aleatoric estimates across different test sets and assay types. First, we see that the estimates are significantly lower on the ADME-T hERG assay than the Target 6 assay. This might relate to several factors that make the two assays distinctly different. It could simply be a result of the overall size of the datasets: recall from Table 1 that the Target 6 assay has only 13,093 labels compared to the 67,687 labels available for ADME-T hERG. However, this trend also corresponds well with the observed standard deviation of the control compounds on the two assays, which was 0.19 for the ADME-T hERG assay and 0.31 for the Target 6 assay. These errors from the control compounds should be more reliable and thus give a better general estimate of the overall homoscedastic noise present in each assay.

Furthermore, an important discrepancy can be observed by comparing the distributions of the aleatoric estimates on each test set between the first row and the second row of Fig. 8. In the first row, all predictions are made by the same model, the one trained on only the first fold. Here, the distributions of the aleatoric estimates are roughly the same for each test set of the two example assays respectively. In the second row, the illustrated predictions of each test set are made by different models, trained on an increasing amount of folds. Thus, these are the models used throughout all previous results of this study. Apart from these models being trained on different amounts of data, they are also optimized individually as explained in the model selection of the supplementary material. Therefore, these models can have vastly different model architectures and be trained using different training procedures such as learning rate.

Crucially, none of these factors should have any effect on the aleatoric uncertainty as it is often categorized in literature as irreducible [2,5,6]. Despite this, we do see distinct differences between most of the distributions of aleatoric estimates by each of these different models in the bottom row of Fig. 8. Specifically, we see some instances where increasing amounts of training data result in generally lower aleatoric estimates. However, there is also one case where the opposite is true, between test folds 3 and 4 of the Target 6 assay. We can only assume this relation to be because of the differences in model architectures. Therefore, our empirical results raise questions about whether the explored models' supposed estimates of aleatoric uncertainty are really fully disentangled from the epistemic uncertainty. Similarly, recent theoretical work on deriving suitable measures for aleatoric, epistemic, and total uncertainty has found that the aleatoric and epistemic parts do not necessarily have to add up to the total uncertainty [68]. As such, the disentanglement of the sources of uncertainty should be considered an ongoing field of research that needs more work to fully determine how these estimates should be categorized and how they relate to the underlying noise in the data.

#### 4. Conclusions

The low-data challenge in QSAR modeling during drug discovery is typically accompanied by additional, often overlooked, partial information from censored labels. Despite their potential value, censored labels have not yet been fully utilized due to the lack of suitable methods to incorporate them during machine learning. Recognizing this gap, we have developed extensions to several established machine learning models, enabling them to effectively learn from censored labels while providing robust and reliable uncertainty quantification.



**Fig. 8. Aleatoric estimates by Gaussian ensemble.** A practical illustration of the aleatoric uncertainty estimates by the Gaussian Ensemble on two example assays. Comparing the predicted aleatoric uncertainty to the experimental error. (Top row) Predictions by the model trained only on the first fold. (Bottom row) Predictions by the respective models trained on all folds up to the given test set.

Our results showed a particular advantage of including the censored labels during training in cases where a high enough proportion of the experimental labels are censored. For the Bayes by Backprop model, the enhancements by including censored labels grew when they made up  $> 13\%$  of the data, whereas  $> 35\%$  was needed for prominent improvements with the other models. Despite these enhancements, drawbacks such as potentially decreased predictive performance and overall performance on specifically observed labels should be considered before adopting our suggested approach. These side effects are likely a result of the commonly unbalanced nature of the censored labels and should be addressed in future work.

Through a comprehensive temporal study using internal pharmaceutical datasets, we demonstrated the importance of these extended models in reliably estimating uncertainties in predicted key affinity scores, and side effects, of potential drug compounds. Our model comparison included approaches that estimate both aleatoric and epistemic uncertainties, providing a more complete understanding of prediction confidence. Specifically, we found that a straightforward ensemble of individually trained neural networks achieves generally high predictive accuracy. However, when accounting for the calibration of uncertainty and computational cost, we recommend using the Bayesian Bayes by Backprop model instead.

Thanks to the temporal evaluation, we were able to detect key differences between the distributions of target-based assays versus ADME-T assays. The ADME-T assays are more diverse in terms of the chemical space and thus exhibit less of a shift throughout time compared to the target-based assays where distinct differences could be observed for different time points. In light of these trends, we showed that the best-performing model in terms of predictive accuracy and calibration of aleatoric uncertainty are typically robust throughout time for ADME-T assays but not always for target-based assays. As such, an evaluation of the best model for ADME-T assays can be trusted to hold without reevaluation whereas target-based assays may need to be reassessed occasionally.

Finally, our case study showed how the uncertainty estimates from the most effective models can be practically applied to inform and guide ongoing drug development efforts, offering valuable insights for risk management and decision-making. Herein, we found that the epistemic uncertainty estimates correlate well with theoretical assumptions, but that the aleatoric uncertainty estimates require further analysis to understand their relation to the underlying, inherent noise in the data.

#### CRediT authorship contribution statement

**Emma Svensson:** Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Hannah Rosa Friesacher:** Writing – review & editing, Software, Methodology, Data curation, Conceptualization. **Susanne Winiwarter:** Writing – review & editing, Data curation. **Lewis Mervin:** Writing – review & editing, Supervision, Conceptualization. **Adam Arany:** Writing – review & editing, Supervision, Methodology. **Ola Engkvist:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

#### Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to proof-read parts of the manuscripts and correct grammatical errors. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

#### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

We thank our colleagues for many insightful discussions. We are also sincerely grateful for the valuable feedback that we have received from our reviewers. 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. HRF, and AA are affiliated to Leuven.AI and received funding from the Flemish Government (AI Research Program).

## Appendix A. Supplementary materials

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.ailsci.2025.100128>.

## Data availability

This work was conducted on internal data that cannot be disclosed beyond the information provided in Table 1. Despite this, we believe the work brings great value to the community by showcasing crucial aspects of the data that could not otherwise be analyzed. First, the temporal evaluation would not be possible to the same extent on public data, which led to conclusions about how the models compare over time. Second, while censoring in experimental labels are naturally occurring in internal data, it is less common in public data. Therefore, the key contribution of adapting and evaluating current uncertainty quantification approaches to censored labels, would not be possible without the proprietary data. All methodology is available in our code on GitHub at <https://github.com/MolecularAI/uq4dd>. Instructions to prepare the programming environment, as well as how to run the training, inference, and evaluation procedures on similar public data from Therapeutics Data Commons [69] can also be found there. Throughout this work, all experiments were run on a cluster of servers with diverse Nvidia GPUs using Python 3.11 with PyTorch 2.0.1 [46].

## References

- [1] Mervin LH, Johansson S, Semenova E, Giblin KA, Engkvist O. Uncertainty Quantification in Drug Design. *Drug Discov Today* 2021;26(2):474–89. <http://dx.doi.org/10.1016/j.drudis.2020.11.027>.
- [2] Apostolakis G. The Concept of Probability in Safety Assessments of Technological Systems. *Science* 1990;250(4986):1359–64, URL <http://www.jstor.org/stable/2878382>.
- [3] Yu J, Wang D, Zheng M. Uncertainty Quantification: Can We Trust Artificial Intelligence in Drug Discovery? *IScience* 2022;25(8):104814. <http://dx.doi.org/10.1016/j.isci.2022.104814>.
- [4] Hansch C, Fujita T.  $p$ - $\sigma$ - $\pi$  Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J Am Chem Soc* 1964;86(8):1616–26. <http://dx.doi.org/10.1021/ja01062a035>.
- [5] 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.
- [6] Hüllermeier E, Waegeman W. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Mach Learn* 2021;110:457–506. <http://dx.doi.org/10.1007/s10994-021-05946-3>.
- [7] Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing Learned Molecular Representations for Property Prediction. *J Chem Inf Model* 2019;59(8):3370–88. <http://dx.doi.org/10.1021/acs.jcim.9b00237>.
- [8] Heid E, McGill CJ, Vermeire FH, Green WH. Characterizing Uncertainty in Machine Learning for Chemistry. *J Chem Inf Model* 2023;63(13):4012–29. <http://dx.doi.org/10.1021/acs.jcim.3c00373>.
- [9] Gustafsson FK, Danelljan M, Schön TB. How Reliable is Your Regression Model's Uncertainty Under Real-World Distribution Shifts? In: *Transactions on machine learning research*. 2023, URL <https://openreview.net/forum?id=WJt2Pc3qtI>.
- [10] Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight Uncertainty in Neural Network. In: *International conference on machine learning*. PMLR; 2015, p. 1613–22.
- [11] Sheridan RP. Three Useful Dimensions for Domain Applicability in QSAR Models Using Random Forest. *J Chem Inf Model* 2012;52(3):814–23. <http://dx.doi.org/10.1021/ci300004n>.
- [12] Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: *International conference on machine learning*. PMLR; 2016, p. 1050–9.
- [13] 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.
- [14] Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction. *J Chem Inf Model* 2020;60(6):2697–717. <http://dx.doi.org/10.1021/acs.jcim.9b00975>.
- [15] Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK. Similarity to Molecules in the Training Set is a Good Discriminator for Prediction Accuracy in QSAR. *J Chem Inf Comput Sci* 2004;44(6):1912–28. <http://dx.doi.org/10.1021/ci049782w>.
- [16] Berenger F, Yamanishi Y. A Distance-Based Boolean Applicability Domain for Classification of High Throughput Screening Data. *J Chem Inf Model* 2018;59(1):463–76. <http://dx.doi.org/10.1021/acs.jcim.8b00499>.
- [17] Bishop CM. *Mixture Density Networks*. Technical Report, Birmingham: Aston University; 1994.
- [18] Nix DA, Weigend AS. Estimating the Mean and Variance of the Target Probability Distribution. In: *Proceedings of 1994 IEEE international conference on neural networks*. ICNN'94, vol. 1, IEEE; 1994, p. 55–60.
- [19] 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, IEEE; 2018, p. 6915–22.
- [20] Amini A, Schwarting W, Soleimany A, Rus D. Deep Evidential Regression. In: *Advances in neural information processing systems*, vol. 33, Curran Associates, Inc.; 2020, p. 14927–37.
- [21] Vovk V, Gammerman A, Shafer G. *Algorithmic learning in a random world*, vol. 29, Springer; 2005.
- [22] Koenker R, Hallock KF. Quantile Regression. *J Econ Perspect* 2001;15(4):143–56, URL <https://www.aeaweb.org/articles?id=10.1257/jep.15.4.143>.
- [23] Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *J Chem Inf Model* 2020;60(8):3770–80. <http://dx.doi.org/10.1021/acs.jcim.0c00502>.
- [24] Kim Q, Ko JH, Kim S, Park N, Jhe W. Bayesian Neural Network with Pretrained Protein Embedding Enhances Prediction Accuracy of Drug-Protein Interaction. *Bioinform* 2021;37(20):3428–35. <http://dx.doi.org/10.1093/bioinformatics/btab346>.
- [25] Wang D, Yu J, Chen L, Li X, Jiang H, Chen K, et al. A Hybrid Framework for Improving Uncertainty Quantification in Deep Learning-Based QSAR Regression Modeling. *J Cheminf* 2021;13(1):1–17. <http://dx.doi.org/10.1186/s13321-021-00551-x>.
- [26] Dutschmann TM, Kinzel L, Ter Laak A, Baumann K. Large-Scale Evaluation of k-Fold Cross-Validation Ensembles for Uncertainty Estimation. *J Cheminf* 2023;15(1):49. <http://dx.doi.org/10.1186/s13321-023-00709-9>.
- [27] Yin T, Panapitiya G, Coda ED, Saldanha EG. Evaluating Uncertainty-Based Active Learning for Accelerating the Generalization of Molecular Property Prediction. *J Cheminf* 2023;15(1):105. <http://dx.doi.org/10.1186/s13321-023-00753-5>.
- [28] Lanini J, Huynh MTD, Scebbi G, Schneider N, Rodríguez-Pérez R. Unique: A Framework for Uncertainty Quantification Benchmarking. *J Chem Inf Model* 2024;64(22):8379–86. <http://dx.doi.org/10.1021/acs.jcim.4c01578>.
- [29] Svensson E, Hoedt PJ, Hochreiter S, Klambauer G. HyperPCM: Robust Task-Conditioned Modeling of Drug-Target Interactions. *J Chem Inf Model* 2024;64(7):2539–53. <http://dx.doi.org/10.1021/acs.jcim.3c01417>.
- [30] Lenselink EB, Ten Dijke N, Bongers B, Papadatos G, Van Vlijmen HW, Kowalczyk W, et al. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *J Cheminf* 2017;9(1):1–14. <http://dx.doi.org/10.1186/s13321-017-0232-0>.
- [31] Fluetsch A, Di Lascio E, Gerebtzoff G, Rodríguez-Pérez R. Adapting Deep Learning QSPR Models to Specific Drug Discovery Projects. *Mol Pharm* 2024;21(4):1817–26. <http://dx.doi.org/10.1021/acs.molpharmaceut.3c01124>.
- [32] Sheridan RP. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J Chem Inf Model* 2013;53(4):783–90. <http://dx.doi.org/10.1021/ci400084k>.
- [33] Friesacher HR, Svensson E, Arany A, Mervin L, Engkvist O. Towards Reliable Uncertainty Estimates for Drug Discovery: A Large-scale Temporal Study of Probability Calibration. In: *ICML 2024 AI for science workshop*. 2024, URL <https://openreview.net/forum?id=5B8tsit5s>.
- [34] Hollander M, Proschan F, Sconing J. Measuring Information in Right-Censored Models. *Naval Res Logist* 1987;34(5):669–81. [http://dx.doi.org/10.1002/1520-6750\(198710\)34:5<669::AID-NAV3220340507>3.0.CO;2-L](http://dx.doi.org/10.1002/1520-6750(198710)34:5<669::AID-NAV3220340507>3.0.CO;2-L).
- [35] Hüttel FB, Riis C, Rodrigues F, Pereira FC. Bayesian Active Learning for Censored Regression. 2024, arXiv preprint [arXiv:2402.11973](https://arxiv.org/abs/2402.11973).
- [36] Pearce T, Jeong JH, Jia Y, Zhu J. Censored Quantile Regression Neural Networks for Distribution-Free Survival Analysis. In: *Advances in neural information processing systems*, vol. 35, Curran Associates, Inc.; 2022, p. 7450–61.
- [37] Arany A, Simm J, Oldenhof M, Moreau Y. SparseChem: Fast and Accurate Machine Learning Model for Small Molecules. 2022, arXiv preprint [arXiv:2203.04676](https://arxiv.org/abs/2203.04676).



- [38] Svensson E, Friesacher HR, Arany A, Mervin L, Engkvist O. Temporal Evaluation of Uncertainty Quantification Under Distribution Shift. In: International workshop on AI in drug discovery. Springer; 2024, p. 132–48. [http://dx.doi.org/10.1007/978-3-031-72381-0\\_11](http://dx.doi.org/10.1007/978-3-031-72381-0_11).
- [39] Heyndrickx W, Mervin L, Morawietz T, Sturm N, Friedrich L, Zalewski A, et al. MELLODDY: Cross-pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information. *J Chem Inf Model* 2023;64(7):2331–44. <http://dx.doi.org/10.1021/acs.jcim.3c00799>.
- [40] Landrum G. RDKit: Open-Source Cheminformatics. 2006, <http://dx.doi.org/10.5281/zenodo.6961488>, URL <http://www.rdkit.org>.
- [41] Weininger D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J Chem Inf Comput Sci* 1988;28(1):31–6. <http://dx.doi.org/10.1021/ci00057a005>.
- [42] Morgan HL. The Generation of a Unique Machine Description for Chemical Structures – A Technique Developed at Chemical Abstracts Service. *J Chem Doc* 1965;5(2):107–13. <http://dx.doi.org/10.1021/c160017a018>.
- [43] Winter R, Montanari F, Noé F, Clevert DA. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. *Chem Sci* 2019;10(6):1692–701. <http://dx.doi.org/10.1039/C8SC04175J>.
- [44] Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem Sci* 2018;9(24):5441–51. <http://dx.doi.org/10.1039/C8SC00148K>.
- [45] Van Tilborg D, Alenicheva A, Grisoni F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J Chem Inf Model* 2022;62(23):5938–51. <http://dx.doi.org/10.1021/acs.jcim.2c01073>.
- [46] Powell JL. Censored Regression Quantiles. *J Econ* 1986;32(1):143–55. [http://dx.doi.org/10.1016/0304-4076\(86\)90016-3](http://dx.doi.org/10.1016/0304-4076(86)90016-3).
- [47] Hartog PBR, Svensson E, Mervin L, Genheden S, Engkvist O, Tetko IV. Registries in Machine Learning-Based Drug Discovery: A Shortcut to Code Reuse. In: International workshop on AI in drug discovery. Springer; 2024, p. 98–115. [http://dx.doi.org/10.1007/978-3-031-72381-0\\_9](http://dx.doi.org/10.1007/978-3-031-72381-0_9).
- [48] Gal Y. Uncertainty in deep learning [Ph.D. thesis], Department of Engineering, University of Cambridge; 2016.
- [49] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011;12:2825–30, URL <http://jmlr.org/papers/v12/pedregosa11a.html>.
- [50] Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, 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* 2008;48(9):1733–46. <http://dx.doi.org/10.1021/ci800151m>.
- [51] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in neural information processing systems, vol. 32, Curran Associates, Inc.; 2019.
- [52] Busk J, Jørgensen PB, Bhowmik A, Schmidt MN, Winther O, Vegge T. Calibrated Uncertainty for Molecular Property Prediction using Ensembles of Message Passing Neural Networks. *Mach Learn: Sci Technol* 2021;3(1):015012.
- [53] Hubschneider C, Huttmacher R, Zöllner JM. Calibrating Uncertainty Models for Steering Angle Estimation. In: 2019 IEEE intelligent transportation systems conference. ITSC, IEEE; 2019, p. 1511–8.
- [54] Yang CI, Li YP. Explainable Uncertainty Quantifications for Deep Learning-Based Molecular Property Prediction. *J Cheminf* 2023;15(1):13. <http://dx.doi.org/10.1186/s13321-023-00682-3>.
- [55] Tobin J. Estimation of Relationships for Limited Dependent Variables. *Econometrica: J Econ Soc* 1958;26(1):24–36. <http://dx.doi.org/10.2307/1907382>.
- [56] Wang R, Liu Z, Gong J, Zhou Q, Guan X, Ge G. An Uncertainty-Guided Deep Learning Method Facilitates Rapid Screening of CYP3A4 Inhibitors. *J Chem Inf Model* 2023;63(24):7699–710. <http://dx.doi.org/10.1021/acs.jcim.3c01241>.
- [57] Juergens M, Meinert N, Bengs V, Hüllermeier E, Waegeman W. Is Epistemic Uncertainty Faithfully Represented by Evidential Deep Learning Methods? In: International conference on machine learning, vol. 235, PMLR; 2024, p. 22624–42.
- [58] Zadrozny B, Elkan C. Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers. In: International conference on machine learning. PMLR; 2001, p. 609–16.
- [59] Levi D, Gispan L, Giladi N, Fetaya E. Evaluating and Calibrating Uncertainty Prediction in Regression Tasks. *Sensors* 2022;22(15):5540. <http://dx.doi.org/10.3390/s22155540>.
- [60] Rasmussen MH, Duan C, Kulik HJ, Jensen JH. Uncertain of Uncertainties? A comparison of Uncertainty Quantification Metrics for Chemical Data Sets. *J Cheminf* 2023;15(1):121. <http://dx.doi.org/10.1186/s13321-023-00790-0>.
- [61] Mann HB, Whitney DR. On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other. *Ann Math Stat* 1947;50–60.
- [62] Dunn OJ. Multiple Comparisons Among Means. *J Amer Statist Assoc* 1961;56(293):52–64.
- [63] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In: Advances in neural information processing systems, vol. 32, Curran Associates, Inc.; 2019.
- [64] Janet JP, Duan C, Yang T, Nandy A, Kulik HJ. A Quantitative Uncertainty Metric Controls Error in Neural Network-Driven Chemical Discovery. *Chem Sci* 2019;10(34):7913–22.
- [65] Kulinski S, Inouye DI. Towards Explaining Distribution Shifts. In: International conference on machine learning. PMLR; 2023, p. 17931–52.
- [66] Bertolini M, Clevert DA, Montanari F. Explaining, Evaluating and Enhancing Neural Networks' Learned Representations. In: International conference on artificial neural networks. Springer; 2023, p. 269–87.
- [67] Maggiora GM. On Outliers and Activity Cliffs Why QSAR Often Disappoints. *J Chem Inf Model* 2006;46(4). <http://dx.doi.org/10.1021/ci060117s>, 1535–1535.
- [68] Sale Y, Bengs V, Caprio M, Hüllermeier E. Second-Order Uncertainty Quantification: A Distance-Based Approach. In: International conference on machine learning. PMLR; 2024.
- [69] Huang K, Fu T, Gao W, Zhao Y, Roohani YH, Leskovec J, et al. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. In: Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 1). 2021, URL <https://openreview.net/forum?id=8nvgNORnoWr>.