



CHALMERS
UNIVERSITY OF TECHNOLOGY

Temporal distribution shift in real-world pharmaceutical data: Implications for uncertainty quantification in QSAR models

Downloaded from: <https://research.chalmers.se>, 2026-03-17 01:18 UTC

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

Friesacher, H., Svensson, E., Winiwarter, S. et al (2025). Temporal distribution shift in real-world pharmaceutical data: Implications for uncertainty quantification in QSAR models. *Artificial Intelligence in the Life Sciences*, 8. <http://dx.doi.org/10.1016/j.aillsci.2025.100132>

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



Research article

Temporal distribution shift in real-world pharmaceutical data: Implications for uncertainty quantification in QSAR models

Hannah Rosa Friesacher^{a,b},^{*} Emma Svensson^{a,c}, Susanne Winiwarter^d, Lewis Mervin^e, Adam Arany^b, Ola Engkvist^{a,f}

^a Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, 431 83, Sweden

^b ESAT-STADIUS, KU Leuven, Leuven, 3000, Belgium

^c ELLIS Unit Linz & Institute for Machine Learning, Johannes Kepler University Linz, Linz, 4040, Austria

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

^e Molecular AI, Discovery Sciences, R&D, AstraZeneca, Cambridge, CB2 0AA, UK

^f Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, 412 96, Sweden

ARTICLE INFO

Keywords:

Uncertainty quantification
Probability calibration
Temporal evaluation
Distribution shift
Drug discovery

ABSTRACT

The estimation of uncertainties associated with predictions from quantitative structure–activity relationship (QSAR) models can accelerate the drug discovery process by identifying promising experiments and allowing an efficient allocation of resources. Several computational tools exist that estimate the predictive uncertainty in machine learning models. However, deviations from the i.i.d. setting have been shown to impair the performance of these uncertainty quantification methods. We use a real-world pharmaceutical dataset to address the pressing need for a comprehensive, large-scale evaluation of uncertainty quantification approaches in the context of realistic distribution shifts over time. We investigate the performance of several popular uncertainty estimation methods for classification models, including ensemble-based and Bayesian approaches. Furthermore, we use this real-world setting to systematically assess the distribution shifts in label and descriptor space and their impact on the capability of the uncertainty quantification methods. Our study reveals significant shifts over time in both label and descriptor space and a clear connection between the magnitude of the shift and the nature of the assay. Moreover, we show that pronounced distribution shifts impair the performance of popular uncertainty quantification methods used in QSAR models. This work highlights the challenges of identifying uncertainty quantification techniques that remain reliable under distribution shifts introduced by real-world data.

1. Introduction

The development of new therapeutic agents is a time- and resource-consuming process, characterized by high failure rates and development spans of over a decade until a compound can be put on the market [1,2]. The use of artificial intelligence (AI), or more precisely, machine learning (ML) approaches, can contribute to easing these problems by using the extensive amount of data produced in the drug discovery pipeline to train computational models that can effectively support future projects with their expert knowledge [3]. During early-stage drug discovery, a part of the vast chemical space is screened to identify promising molecular compounds, which are subsequently optimized to achieve the desired properties [4]. The large scale and complexity of this early-stage screening make it an ideal application

for ML models with their high computational power and predictive abilities [3,5].

Quantitative structure–activity relationship (QSAR) models are well-established in computer-aided drug discovery for identifying compounds with desired features. They enable the prediction of biological activities or properties of chemical compounds based on their molecular structure. However, the reliability of these approaches is crucial to optimally support an informed decision-making process, which ultimately saves money and time in the lengthy and costly drug discovery pipeline.

Uncertainty quantification is a powerful tool to increase the reliability of ML models and the confidence in deploying them to real-world applications [6]. Various sources can lead to uncertainty in the predictions obtained from neural networks. A common classification found

* Corresponding author at: Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, 431 83, Sweden.
E-mail address: rosa.friesacher@kuleuven.be (H.R. Friesacher).

in literature is the distinction between aleatoric uncertainty, which originates from uncertainty in the data, and epistemic sources, which quantifies uncertainty inherent in the choice of model [7,8]. Optimally, estimates of the predictive uncertainty should represent the total uncertainty originating from these different sources. Uncertainty quantification methods can be classified into Bayesian approaches [9–12], ensemble-based models [13,14], conformal predictors [15,16], evidential learning [17–21] and distance-based approaches [22]. Furthermore, multiple techniques exist that can improve the uncertainty estimates post hoc by calibrating them using a simple function trained on a separate calibration dataset [23–25]. Many of these computational tools have been explored for drug discovery applications to enable the estimation of predictive uncertainties in molecular property prediction tasks [26,27]. However, available uncertainty quantification methods vary in their ability to capture all sources of uncertainty correctly, and there is no clear agreement in previous studies on which approach estimates these uncertainties most reliably [19,28–36].

Furthermore, the available uncertainty quantification methods have primarily been evaluated on public data lacking temporal information about the measurements, which is needed to perform data splits that cohere with the history of the assay of interest. Due to this lack of temporal information, the use of temporal splitting techniques for cross-validation is not possible, which is needed to realistically evaluate model performance over time, as reported by Sheridan [37] for classification and Landrum et al. [38] for regression tasks. Alternative splitting strategies that do not require temporal input include random splits or approaches that are based on the chemical structure of the chemical compounds. However, these methods are usually too optimistic or pessimistic compared to the true prospective prediction as they do not reflect the evolution of data in real-world pharmaceutical drug discovery projects [37,38].

The first part of this work investigates the evolution of real-world pharmaceutical data and the resulting distribution shift. Dunder et al. [39] reported an intrinsic assumption of many training algorithms that the data is independent and identically distributed (i.i.d.). While this assumption is foundational for traditional ML models, it imposes significant constraints and oversimplifies the complexities of realistic scenarios [40]. Consequently, the simplified problems may fail to accurately represent or address the challenges inherent in real-world datasets, such as the pharmaceutical data included in this study. In the context of probability calibration, deviations from the i.i.d. setting have been shown to impair the performance of common uncertainty estimation methods previously reported to improve model calibration under i.i.d. conditions [41,42].

Therefore, the second part of this work compares common uncertainty quantification approaches that employ real-world temporal splits to evaluate model performance in a more realistic setting. In binary classification problems, neural networks typically give probability-like predictions that can be directly interpreted as an estimate of the confidence in the prediction. Previous work has concluded that modern neural networks often fail to give realistic estimates of the uncertainty associated with a prediction in classification tasks, resulting in poorly calibrated models [26,27,43]. Several approaches exist in the literature that use more sophisticated techniques to improve the reliability of these uncertainty estimates. For a more straightforward comparison between the uncertainty quantification methods used in this work, we classify them into two categories, namely train-time uncertainty estimation approaches and post hoc probability calibration methods.

Train-time uncertainty quantification approaches refer to Bayesian methods or ensemble-based techniques inspired by the Bayesian framework to estimate the posterior distribution of predictions from a set of models [13,14,44]. The idea of these approaches is to construct the posterior distribution, which correctly reflects the epistemic uncertainty. Assuming that the model cannot generalize well to inputs outside the domain of the training data, model accuracy is expected to decrease for these inputs. If the model is well-calibrated, it should also

produce less confident predictions in these cases. Thus, the predictions are expected to be consistent if the input is similar to the training data, and to become increasingly diverse as the inputs move away from the training data. It is important to acknowledge that the model behavior described above reflects a best-case outcome and that modeling approaches might not behave as expected, as reported by Abe et al. [45] in the context of deep ensembles.

We consider three methods for train-time uncertainty estimation in this work. Deep ensembles and Monte Carlo (MC) dropout aim to improve the performance by obtaining numerous base estimators to determine the model variance [7,13,14]. Furthermore, we compare the ensemble-based strategies with a full Bayesian neural network trained with the Bayes-by-Backprop approach [10]. Bayes-by-Backprop allows to quickly obtain samples from the posterior distribution of the neural network weights using a variational approximation scheme.

While these train-time uncertainty quantification approaches aim to achieve better uncertainty estimation by accounting for the epistemic uncertainty, post hoc calibration approaches improve model calibration by applying an additional post-processing step to the scores retrieved from a separately trained classifier. These methods require a separate dataset, called a calibration set, to train the calibrating function used in the post-processing step. For this study, we tested two post hoc calibration techniques, including the commonly used Platt scaling approach [23] and Venn-ABERS predictors [24], which were previously shown to enhance the probability calibration of classifier predictions [35,46]. Platt scaling fits a logistic regression to the classification scores of the calibration set to counteract over- or underfitted uncertainty estimations, while Venn-ABERS predictors use the more flexible isotonic regression functions to calibrate the probability point estimates.

To our knowledge, only a few studies have addressed the performance of uncertainty estimation approaches under temporal shifts. In some of these works, temporal splitting approaches are applied to ChEMBL [47] data, using the publication date as a Ref. [19]. As the date of publication does not correspond to the date when the experiment was conducted, it remains questionable how accurately this information can reflect the timeline in a pharmaceutical company. Other studies [48,49] used internal data from pharmaceutical companies with the necessary information to perform proper temporal splits. Rodríguez-Pérez et al. [48] studied the performance of multitask graph neural networks for uncertainty estimation, focusing on intrinsic clearance data. Another recent work compares the uncertainty estimation of various regression models for pharmacokinetic property prediction of potential drug molecules [49]. However, both of these studies use training data from a fixed time span for all experiments and, therefore, do not address model performance over time. Furthermore, they do not address shifts in the data caused by the temporal splitting strategy. Svensson et al. [50] recently published an extensive temporal study comparing uncertainty estimation methods trained on drug-target interaction data with and without censored data. While this study provides a comprehensive guide on handling uncertainty estimation methods for regression tasks, a comparable large-scale study that applies temporal splitting strategies to different biological assays has yet to be published for classification approaches.

In this work, we aim to address these gaps by assessing the performance of different uncertainty estimation approaches using single-task classification models over time and in the context of assay-specific distribution shifts in the data. The models were trained on an internal dataset, which has already been studied in previous works [50–52]. This dataset includes drug-target interaction data from different biochemical assays, providing the additional information required to perform temporal splits. First, we analyze the history of the individual assays and how the data distribution shifts in label and descriptor space over time. Next, we compare available probability calibration and train-time uncertainty quantification methods and explore possible connections between their performance and distribution shifts.

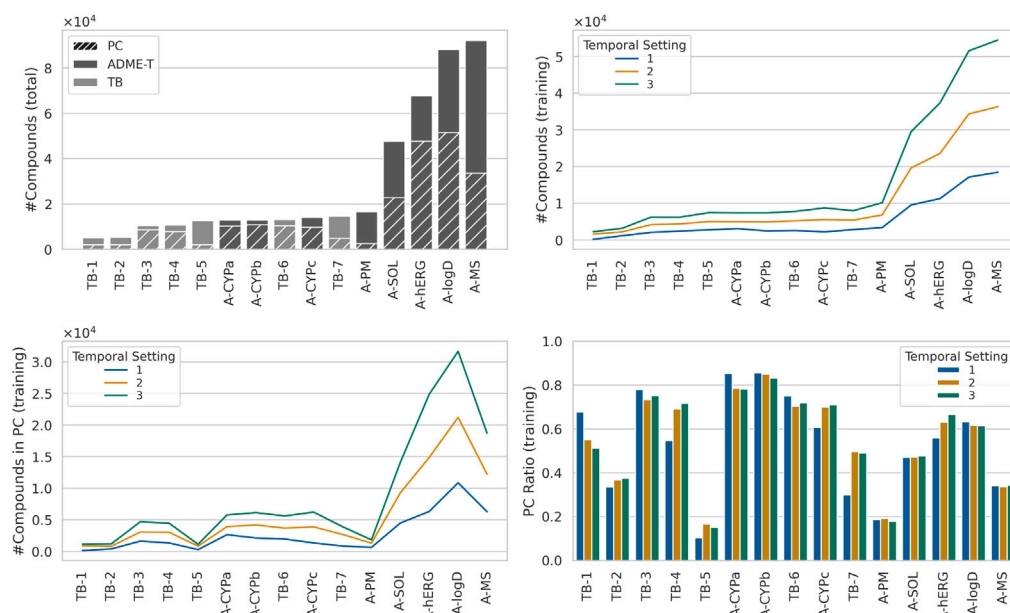


Fig. 1. Overview of datasets. The upper left panel plots the total size of the individual assays ordered according to assay size. The striped areas in the bar indicate the number of compounds belonging to the preferred class (PC) in each assay. The upper right panel shows the amount of training data in each temporal setting across all assays, with 1, 2, or 3 time spans used for training. In the bottom panels, the total number (left) and ratio (right) of training set compounds belonging to the PC are displayed for each temporal setting and assay. Compound counts are shown in units of 10^4 .

2. Methods

The following section is structured into three parts to examine the material and methods used in this study. The first section describes the assay data and the splitting strategy used to generate splits representing different time spans in the assay history. The second part addresses the modeling approaches, including the baseline estimators and the more sophisticated uncertainty estimation approaches. Finally, the last section provides insight into the experiments and metrics used to compare the uncertainty estimation approaches comprehensively.

2.1. Data

This study uses internal data from 15 biological assays to gain insight into the properties of real-world pharmaceutical data and subsequently trains binary classifiers for each assay separately. Parts of the dataset [51,52] or the whole dataset [50] have already been used in previous studies. The included assays are diverse and represent different optimization problems typically addressed during the drug discovery workflow to enhance the pharmacokinetic and pharmacodynamic properties of a drug candidate. Furthermore, the assays exhibited different sizes, modeled endpoints, and ratios of compounds belonging to the preferred class (PC) in the individual datasets. The total size of the assays is illustrated in the upper left panel of Fig. 1. In addition, Table 1 provides detailed information on the assays used in this study. The assays were assigned to two categories, Target-Based (TB) and ADME-T, and subsequently labeled based on size and category affiliation. The labels of the ADME-T assays were created from the class name and the assay description, together with the prefix “A” to indicate the affiliation to the ADME-T category (e.g., A-logD for the lipophilicity assay). In contrast, the TB assays were ordered according to size and numbered consecutively (TB-1 for the smallest to TB-7 for the largest TB assay).

TB assays. The TB category includes project-specific assays from activity screens to identify active substances on a specific target of interest. Active substances are compounds that modulate the function of a protein, for example, by inhibiting or activating the target. This work includes seven TB assays, with assay sizes ranging from 5082 to 14,605 measured compounds. As opposed to the ADME-T assays,

further specifics regarding these biological assays cannot be disclosed due to proprietary constraints.

ADME-T assays. Assays in the ADME-T category typically assess the pharmacokinetic properties and toxicity profile of a drug candidate. These properties are connected to the absorption, distribution, metabolism, and excretion (ADME) of a compound, while the toxicity screens identify compounds that hit unintended targets. The ADME-T category comprises assays that assess the general features of a compound, which are typically relevant for its success in the drug discovery and development pipeline [53,54]. The assays, which include data from various projects, are usually comparatively large. In our study, eight ADME-T assays were used, including five large assays with measurement numbers between 16,511 and 92,161 and three smaller assays comprising 12,875 and 14,062 data points, which measure interactions with Cytochrome P450 (CYP).

As opposed to the TB assays, the ADME-T assays included in this study are widely used in the drug discovery process, which allows the disclosure of more detailed descriptions of the assays. The CYP assays measure the inhibition of one of the two CYP isoforms, CYP3A4 (A-CYPa) and CYP2C9 (A-CYPb and A-CYPc). These isoforms play an essential role in drug metabolism and the detection of drug-drug interactions [55–57]. Two distinct assay types are available, exploring different interactions with the CYP isoforms. The CYP2C9 (I) and CYP3A4 assays measure drug molecule disappearance using liquid chromatography-mass spectrometry, while the CYP2C9 (II) assay measures CYP inhibition using a fluorescent substrate. In both assays, weaker interactions with the CYP protein are usually favorable to avoid rapid decomposition of the drug molecule and drug-drug interactions. The permeability assay (A-PM) evaluates the flux of a compound across a Caco-2 cell, reflecting its potential in vivo absorption, which is measured in $1e-6$ cm/s [58]. High velocities are favorable, indicating a compound’s ability to cross biological membranes. The solubility assay (A-SOL) assesses the maximum concentration of a compound in an aqueous solution at pH 7.4. A Dimethyl sulfoxide (DMSO) stock solution is used, and the organic solvent is evaporated to obtain a solid sample. Compounds with high solubility are preferred to allow sufficient dissolution in biological fluids [59]. The hERG assay (A-hERG) provides vital insight into a compound’s toxicity profile by measuring

Table 1

Overview of the assay data. Details of the assays used in this study, including assay size, assay unit, and modeled endpoint. Descriptions of the ADME-T assays are shown. The ratio of compounds belonging to the preferred class (PC) is reported for each assay. The last column indicates the corresponding threshold T used to assign compounds to classes and if the preferred class (PC) lies above or below T ($PC < / > T$).

Abbreviation	Assay description	Assay size	PC ratio	Assay unit	Modeled end-point	Threshold T (PC < / > T)
Target-Based						
TB-1	NA	5,082	0.38	μM	pIC50	> 6
TB-2	NA	5,237	0.39	μM	pIC50	> 6
TB-3	NA	10,465	0.82	μM	pIC50	> 6
TB-4	NA	10,624	0.73	μM	pIC50	> 6
TB-5	NA	12,612	0.16	μM	pEC50	> 6
TB-6	NA	13,093	0.79	μM	pIC50	> 6
TB-7	NA	14,605	0.33	μM	pIC50	> 6
ADME-T						
A-CYPa	CYP3A4	12,875	0.79	μM	pIC50	< 5
A-CYPb	CYP2C9 (I)	12,876	0.84	μM	pIC50	< 5
A-CYPc	CYP2C9 (II)	14,062	0.30	μM	pIC50	< 5
A-PM	Permeability	16,511	0.15	1e-6cm/s	logP	> 1
A-SOL	Solubility	47,607	0.48	μM	logS	> 2
A-hERG	Toxicity	67,687	0.70	μM	pIC50	< 5
A-logD	Lipophilicity	88,114	0.58	-	logD	> 3
A-MS	Metabolic Stability	92,161	0.36	$\mu\text{l}/\text{min}/1\text{e}6$	logMS	< 1

its inhibiting effects on the human Ether-a-go-go Related Gene (hERG) potassium channel. Inhibition of hERG is correlated to severe cardiac side effects by prolonging the QT interval [60]. Therefore, inhibiting interactions with hERG is usually undesirable. The lipophilicity of a compound is obtained in the A-logD assay by measuring the logarithm of the distribution coefficient between octanol and aqueous phase at pH 7.4. Lipophilicity is crucial since it significantly affects drug absorption, metabolism, and safety. A logD greater than 3 has previously been identified as a trigger for safety concerns [61,62]. Finally, the metabolic stability assay (A-MS) measures how fast a compound is metabolized in rat hepatocytes. The in vivo hepatic clearance is measured in $\mu\text{l}/\text{min}/\text{million}$ cells. In general, low values for hepatic clearance are desirable, as they imply slower decomposition and, therefore, higher bioavailability of the drug molecule [63,64].

Binary classification of compounds. The measured values were converted to a logarithmic scale, and a suitable threshold was determined for each assay individually. Assay-specific thresholds were defined to determine if a compound belongs to the PC. All TB assays obtain a compound's inhibiting potency by measuring the IC50, except TB-5, in which the EC50 was used. The IC50 value measures the compound concentration needed to inhibit a protein's activity by half, while the EC50 value indicates the compound's concentration that triggers half of the maximum possible effect. Subsequently, these values were converted to pIC50/pEC50 by taking the logarithm of the measurements converted from micromolar (μM) to molar. The PCs in the TB assays comprise compounds with a pIC50/pEC50 value above 6, indicating that a substance achieves the desired effect when its IC50/EC50 value is below 1 μM . In addition, four ADME-T assays, including the three CYP and the hERG assay, contain pIC50 values. Since these assays aim to detect interaction with off-targets, lower pIC50 thresholds of 5 were selected to decrease the risk of false negatives. The PC includes compounds with pIC50 values below this threshold. For A-logD, A-SOL, A-PM, and A-MS, individual thresholds were chosen as shown in Table 1 to assign compounds with desirable properties to the PC.

Temporal split. For each assay, we split the data into five roughly equally sized folds using the date of each measurement. Each fold represents a specific time span in the history of the assay. These folds were then used to set up three experimental settings, using one, two, or three folds for training the QSAR models. In each case, the first subsequent fold was used for validation, including model selection and calibration where applicable. We only evaluated each setting on the first fold following the validation set for consistency between test sets. However, all remaining folds could, in principle, be used. Fig. 2 illustrates the temporal splitting strategy. Considering all assays and

settings, 45 separate training datasets were used throughout this work. For experiments in which the results of all three settings are used, the assays are labeled as *Assay Abbreviation [Temporal Setting]*. Naturally, the training dataset sizes and number of training compounds belonging to the PC vary across the temporal settings as shown in the upper right and lower left panel of Fig. 1. Furthermore, the bottom right plot in Fig. 1 demonstrates that the ratios of PC compounds in the training sets differ between temporal settings.

2.2. Models

Fig. 3 provides an overview of the models compared in this study. All architectures used in this work stem from a Random Forest (RF) or a multilayer perceptron (MLP). Both approaches are commonly used in research addressing uncertainty estimation in QSAR modeling [30,35,36].

Extended connectivity fingerprints (ECFP) [65] were generated and used as model input. ECFPs are one of the most widely used fingerprint types for molecular property prediction. While more advanced molecular representations and model architectures exist, such as graph neural networks for molecular graphs or language models for SMILES representations, ECFPs have been demonstrated to perform comparably to these more sophisticated approaches [66,67].

Furthermore, since our study aims to gain insight into uncertainty estimation in QSAR models rather than finding the best approach or comparing molecular representations, we opted for the simple and popular ECFP representation. The RDKit package [68] was used to generate ECFPs of length 4096 from the SMILES [69] of the compound structures. Due to additional computational constraints, we concentrated on RF and MLP models as suitable choices for examining uncertainty quantification in a temporal context.

Model generation. A Python package is publicly available at <https://github.com/MolecularAI/uq4dd>, which contains the code used for model generation and evaluation inspired by the design pattern proposed by Hartog et al. [70]. The hyperparameter tuning for the two baseline estimators, RF and MLP, was performed using an exhaustive grid search. The exact parameter space search is described in Table S1 of the supplementary material. The binary cross-entropy (BCE) loss was calculated to compare the model performance on a validation set. Since assay data often exhibits distributional changes over time, the temporal splitting approach described in this paper was also used to obtain validation sets that reflect this progression, as shown in Fig. 2. This allowed the selection of hyperparameters that perform best under these realistic conditions.

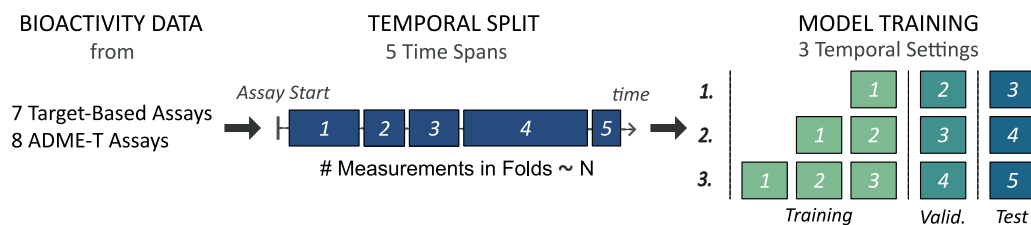


Fig. 2. Overview of the temporal split and model training. The data in each assay was assigned to 5 time spans to create three temporal settings, each with increasing amounts of training (Training) data. The subsequent two folds were used for validation (Valid.) and testing (Test). The validation data also served as a calibration set used in post hoc calibration approaches.

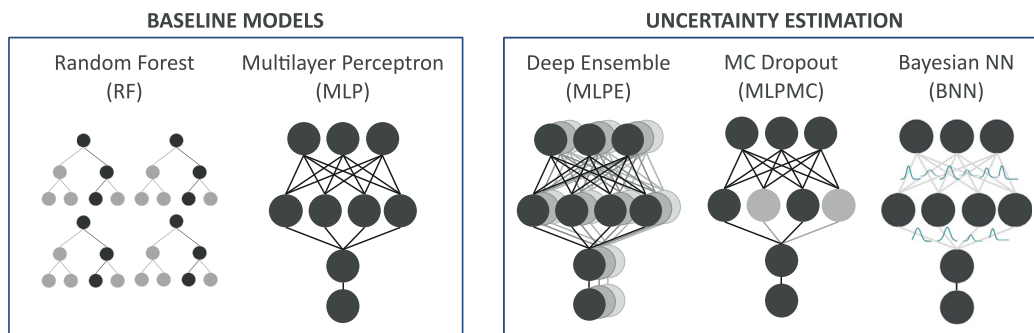


Fig. 3. Overview of the classification models. The architectures of the baseline models and train-time uncertainty quantification methods compared in this study are shown. All models were trained in a single-task manner. The hyperparameters of the baselines, RF and MLP, were tuned in an extensive grid search. The baseline MLP was used as the basis for the three uncertainty quantification methods, deep ensembles, MC dropout, and a Bayesian neural network.

The RF models were generated using scikit-learn [71]. During hyperparameter tuning, the maximum depth of the trees and the required number of estimators of each assay and temporal setting were individually tuned using the validation BCE loss. Probability-like outputs were generated from the ratio of decision trees in the RF that classified a test instance as active.

The MLP models were trained using PyTorch [72] with the BCE loss function. Similarly, the model selection, including early stopping, was optimized using the validation loss for every assay and temporal setting. The network architecture was optimized for the number of hidden units, the number of hidden layers, and the dropout rate. Additionally, the learning rate and scaling factor of a ReduceOnPlateau learning rate scheduler were also optimized. Adam was used as an optimization algorithm [73] to train the neural networks. Probability-like scores were obtained by applying a sigmoid function to the output of the MLP.

The baseline models were further modulated to generate more sophisticated uncertainty estimation methods. Train-time uncertainty quantification methods were trained using the MLP as the foundation. Furthermore, post hoc probability calibration methods were applied to selected models. A detailed description of the train-time uncertainty estimation and post hoc calibration methods can be found below.

Train-time uncertainty quantification. Train-time uncertainty quantification approaches aim to estimate uncertainty during model training by accounting for model variance. In contrast to post hoc calibration methods, they do not apply a post-processing step to the scores of the classifier. In this work, we compare two ensemble-based techniques inspired by the Bayesian theorem: deep ensembles (MLPE) and Monte Carlo (MC) dropout (MLPMC). We also include one full Bayesian neural network trained with the Bayes-by-Backprop approach (BNN). These methods aim to estimate the posterior distribution over the parameters of the neural network [13,14,44]. Theoretically, the posterior distribution $P(\theta|D)$ over model parameters θ , given data D , can be computed using the Bayesian theorem

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta) d\theta}. \quad (1)$$

When working with high-dimensional posteriors, the calculation of the closed-form solution of the Bayesian equation is usually infeasible due to the intractability of the evidence term in the denominator, which requires solving a highly complex integral. To circumvent this problem, sampling-based methods are often used that retrieve samples $\theta := \{\theta_1, \theta_2, \dots, \theta_N\}$ from the posterior distribution, so that $\theta_n \sim P(\theta|D)$. During inference, the predictions of the sampled models are averaged to obtain a mean estimate of the target label y given the descriptor x :

$$P(y|x, D) \approx \frac{1}{N} \sum_{n=1}^N P(y|x, \theta_n). \quad (2)$$

Both ensemble-based approaches, deep ensembles (MLPE) and MC dropout (MLPMC), approximate the Bayesian treatment by estimating the predictive uncertainty using numerous base estimators. Deep ensembles use multiple randomly initialized models as base estimators, corresponding to different local minima in the loss landscape [13]. In this work, 25 base estimators were trained, and their predictions were averaged to obtain a point estimate. MC dropout applies dropout during inference by setting a number of randomly selected neurons to zero to introduce stochasticity [14]. To generate MC dropout (MLPMC) models, 400 forward passes using dropout were aggregated, with the average being the final prediction of the models.

To compare the ensemble-based methods with a full Bayesian approach, we include Bayesian neural networks (BNN) trained with the Bayes-by-Backprop method in the comparison study. We used a previously published repository for the Bayes-by-Backprop method accessible at <https://github.com/ThirstyScholar/bayes-by-backprop> as a template for our implementation of the BNN approach. In the Bayesian setting, neural network weights are treated as random variables rather than point estimates, which allows model variance to be accounted for in the posterior distribution of the weights. Since the parameter space is usually high-dimensional, the closed-form solution of the posterior distribution cannot be solved.

Bayes-by-Backprop provides a quick solution for obtaining samples from the approximate posterior distribution of neural network weights W using a variational approximation scheme. The underlying idea of Bayes-by-Backprop is to learn the optimal parameters θ^* of

a surrogate distribution that minimizes the Kullback–Leibler (KL) divergence [74] between the simpler surrogate $q(W|\Theta)$ and the complex posterior distribution $P(W|D)$

$$\Theta^* = \operatorname{argmin}_{\Theta} KL[q(W|\Theta)|P(W|D)]. \quad (3)$$

The computation of the resulting KL divergence requires the intractable closed-form solution of the posterior. To avoid the calculation of the posterior, the evidence lower bound (ELBO) can be computed, which provides a lower bound on the log likelihood of the observed data.

$$P(D) \geq E_{q(W|\Theta)}(\log \frac{P(D|W)P(W)}{q(W|\Theta)}) = ELBO \quad (4)$$

Using the ELBO, a computable loss function can be derived that is used in the backpropagation framework of the Bayes-by-Backprop approach:

$$\mathcal{L}(\Theta, D) = \operatorname{argmin}_{\Theta} KL[q(W|\Theta)|P(W)] - \mathbb{E}_{q(W|\Theta)}(\log P(D|W)). \quad (5)$$

When sampling from the surrogate distribution, the introduced stochasticity prevents using a backpropagation scheme. To allow the computation of gradients, the local reparametrization trick is applied. Instead of sampling directly from the proposal function, a deterministic transformation function with learnable parameters is used to convert a sample of parameter-free noise into a sample of the proposal function. We refer to Blundell et al. [10] for more technical details of the Bayes-by-Backprop approach.

Post hoc probability calibration. Two post hoc probability calibration techniques were fitted to each model using the validation set. These approaches included Platt scaling [23] and Venn-ABERS (VA) predictors [24]. Platt scaling fits a logistic regression to the classification scores to counteract over- or underfitted uncertainty estimations [23]. Two isotonic regression functions were trained on the validation set and a given test instance [24] for calibration with VA predictors. The functions represent the hypothesis that the test instance is active versus inactive. As such, the probabilities obtained from the isotonic regression functions correspond to a lower and an upper bound on the estimated probability. Finally, these bounds were condensed to a point estimate, as proposed by Toccaceli et al. [75]. The suffixes -P and -VA indicate models calibrated with Platt scaling or Venn-ABERS predictors, respectively. For instance, the calibrated MLPE model was labeled MLPE-P or MLPE-VA.

2.3. Experiments

The first part of this study focuses on the data characteristics resulting from the temporal splitting strategy. We studied the shift in label and descriptor space over time, mainly concentrating on the differences between TB and ADME-T assays. The shift in label space was assessed by comparing the PC ratios in each time span. Shifts in the descriptor space were quantified using the maximum mean discrepancy (MMD) [76], which provides a kernel-based estimate for the distance between the distributions of two datasets. MMD has been successfully applied to different tasks, including the detection of distribution shift [77,78], adversarial sample detection [79], and unsupervised domain adaptation [80,81]. Furthermore, MMD was shown to be applicable to a wide range of problems in molecular biology [82]. Lee et al. [83] used MMD on graph representations of chemical structures to score molecules obtained from a generative model. In this work, we compute the MMD to detect distribution shifts between the ECFP spaces of the training datasets $X := \{x_1, \dots, x_M\}$ and the test datasets $Z := \{z_1, \dots, z_N\}$ which are distributed according to $P(X)$ and $Q(Z)$.

$$\begin{aligned} MMD(P, Q) &= \frac{1}{M^2} \sum_{i=1}^M \sum_{j=1}^M k(x_i, x_j) + \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N k(z_i, z_j) \\ &\quad - \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N k(x_i, z_j). \end{aligned} \quad (6)$$

When using the Tanimoto coefficient [84] as kernel, the MMD lies between 0 and 1, with 0 indicating no differences and 1 indicating no shared features between the compounds in the datasets. To investigate the overlap of chemical scaffolds in the training and test datasets, Murcko scaffolds of the chemical structures were retrieved using RD-Kit [68]. The proportion of test set scaffolds that overlapped with those in the training set was determined.

The second part of the study aims to determine how well different uncertainty quantification methods estimate the probability that compounds have a certain desirable feature, such as being active on a TB assay or inactive on an ADME-T toxicity assay. To improve reproducibility, ten repetitions were generated for each method using random initialization. The reliability of the models was assessed by calculating various performance metrics on the test set predictions, as described below. The metrics were calculated for each repetition and subsequently aggregated to obtain the mean and standard deviation for each score and model across all datasets. For each assay and temporal setting, a two-sided, independent t-test was used to assess whether the difference between the best model and any other model was statistically significant. To compare model performance, the AUC under the receiver operating characteristic (ROC) curve [85] (AUC [↑]), the binary cross-entropy (BCE [↓]) and the adaptive calibration error [86] (ACE [↓]) of the predictions were calculated using the test set predictions. The calibration error was obtained by ordering the predictions and assigning them to ten bins. Subsequently, the difference between the mean probabilities and the ratio of the instances belonging to the PC was computed for each bin. The ACE was calculated by taking the mean of the differences in the bins. Another commonly used calibration error is the expected calibration error, which is similar to the ACE but uses equally spaced instead of equally sized bins [86]. However, this binning strategy can overestimate the calibration error when handling imbalanced datasets due to the high variance of the predictions in the sparsely populated bins [86]. In this context, the ACE provides a more robust estimate and is, therefore, the preferred estimator for the probability calibration error in this study. Note that the ACE is an improper score [87,88], so a perfect calibration error of 0 does not automatically correspond to the best model. Thus, the ACE was always evaluated in combination with the BCE for a more comprehensive analysis.

3. Results and discussion

We divide the results of this study into two consecutive parts to investigate the efficacy of uncertainty quantification and probability calibration methods using real-world temporal data. The first part addresses the properties of the data in the context of a distribution shift in the label and descriptor space. In the second section, we compare the probability calibration of various uncertainty quantification methods and set the results in the context of the underlying distribution shift resulting from the temporal split.

3.1. Distribution shifts over time

Shift in label space. The ratios of the PC in different time spans of an assay were compared to evaluate shifts in the label space. As previously reported in Section 2.1, the ratio of compounds belonging to the PC varies across temporal settings. The magnitude of this difference is dependent on the assay, as shown in the bottom right panel of Fig. 1. Some assays exhibit larger differences, like TB-7, while for others, such as A-PM, the ratio is more stable over time. To evaluate the distribution shift in label space between training and test data, the difference in the ratios of compounds in the PC was calculated. The left panel in Fig. 4 illustrates these differences between the training and the test set for each assay. The ratio of compounds belonging to the PC in each time span is listed for all assays in Table S2 of the supplementary material. We assessed all three temporal settings, using one, two, or

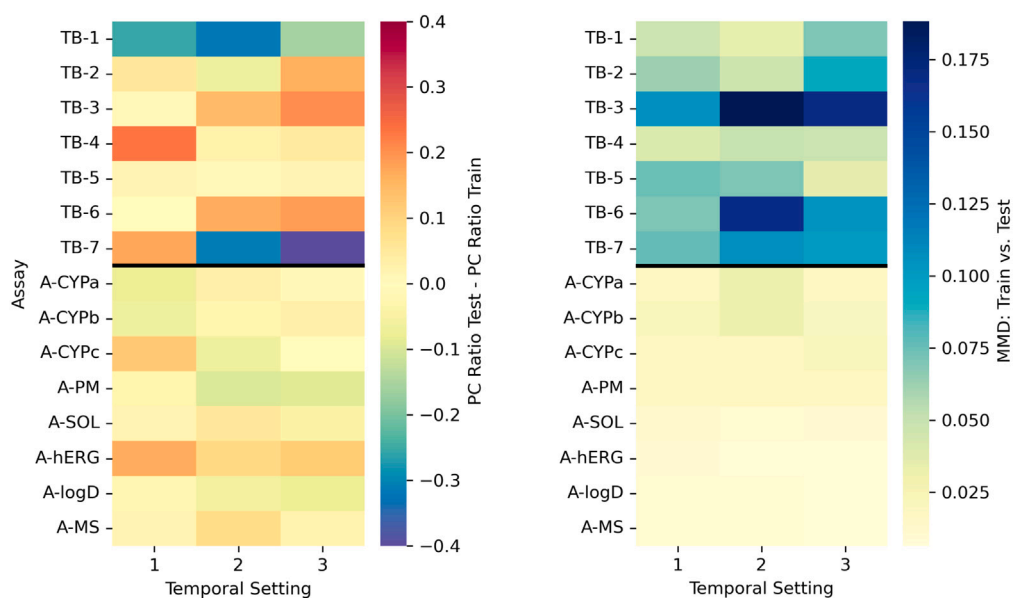


Fig. 4. Quantification of the distribution shifts between the training and test datasets over time. The shift in label space and in the descriptor space is illustrated for each temporal setting, using the data of 1, 2, or 3 time spans for training. Results are shown for each assay. The left panel shows the shift in label space in terms of the difference in ratios of the preferred class (PC) between the training and test datasets. The right panel shows the MMD in the descriptor space between the training and test datasets for each temporal setting and assay, quantifying shifts in descriptor space.

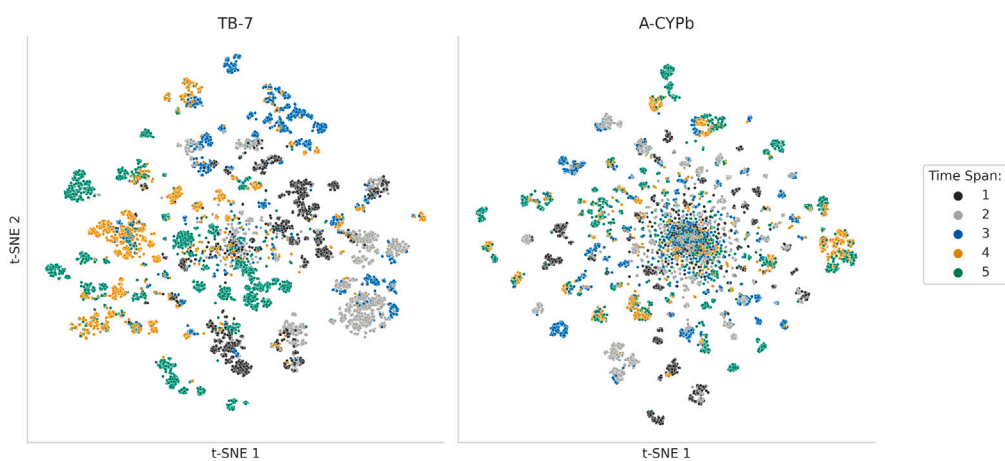


Fig. 5. T-SNE plots of the ECFP space. T-SNE plots of the ECFP space are shown for one example of each assay category to illustrate how the explored chemical space changes over time. Compounds are colored according to the time span that they were assigned to. The t-SNE plots of the remaining TB and ADME-T assays are shown in Figure S1 and Figure S2 in the supplementary material.

three time spans as training data. The second consecutive span after the training folds was considered the test set. We refer to Fig. 2 for a more comprehensive explanation of the different training settings. The left heatmap of Fig. 4 shows that TB assays evolve differently in label space over time than ADME-T assays. Generally, the differences in PC ratios between the training and test sets are smaller in ADME-T assays, while the more extreme values in TB assays indicate larger shifts in label space. Recall that the TB category includes project-based assays, which aim to find modulators for a specific target of interest. Therefore, a plausible explanation for the larger shifts in label space could be that various chemical series are tested in search of promising compounds. These series may differ in their abilities to modulate a target, leading to changing ratios of preferred compounds over time. Some assays show an enrichment of the PC over time, as observed in TB-3 and TB-6. However, this pattern was not observed in all assays, and the TB-1 and TB-7 assays even show opposite tendencies. Two datasets exhibited high shifts in label and small shifts in descriptor space, which might indicate the presence of activity cliffs. These two assays included the

second temporal setting of TB-1 and the first temporal setting of TB-4. Note that label shifts might also result from chemical series with activity values close to the classification threshold. This could lead to similar chemical compounds labeled differently, despite the absence of an actual activity cliff. A detailed discussion of the shifts in descriptor space can be found in the next paragraph. Similarly to the TB assays, the more stable ratios in the ADME-T assays can also be attributed to the nature of this assay category. These assays are not specific to individual projects and are used to evaluate the pharmacokinetic and toxicity profiles of compounds. Based on the results above, it is highly questionable whether the i.i.d. assumption for these assays remains valid over time, particularly in the TB category. This category includes some challenging assays, such as TB-3 and TB-7, which exhibit significant label shifts.

Shift in descriptor space. The shifts in ECFP space are visualized using two-dimensional t-SNE plots to reveal patterns and clusters in the dataset. The t-SNE plots for TB-7 and A-CYPb are shown in Fig. 5, while the plots for the remaining assays can be found in Figure S1 and

Figure S2 in the supplementary material. Fig. 5 reveals a clear pattern in the TB-7 assay, which indicates a shift in the chemical space over the assay history. Furthermore, chemically similar compounds tend to be assigned to the same time span, as indicated by the color purity in various clusters. In contrast, the t-SNE plot of the ADME-T assay does not show a clear pattern in clusters and color gradients. To quantify the shift in descriptor space, the MMD was calculated between the training and test sets. The MMD of the three temporal settings is shown in the right panel of Fig. 4. The observed tendencies are similar to the patterns reported for the label shifts. In general, the TB assays exhibit larger shifts than the ADME-T assays, which is also supported by the patterns seen in the t-SNE plots. These results can again be explained by the different characteristics of the two assay categories, resulting in distinct developments through the descriptor space over time. To find promising compounds in the TB assays, various chemical series are usually screened, containing chemically similar compounds. As a result, large shifts are observed when comparing the descriptor space of compounds assigned to different time spans. Interestingly, when comparing the ratio of scaffolds in the test set shared with the training set, no apparent differences in the scaffold overlap between the TB and ADME-T assays could be detected (Table S3 in the supplementary material). This means that the difference in distribution shifts between the ADME-T and the TB assays is not a result of the scaffold overlap between the training and test sets. In conclusion, the shifts in descriptor space are more pronounced in TB assays, while those in ADME-T assays are comparatively small. Similar to the shifts in label space, the i.i.d. assumption might not be appropriate, particularly in the TB assays.

3.2. Probability calibration study

We assessed the performance of various uncertainty estimation methods in three experiments, focusing on the probability calibration of the models. Throughout this part of the study, the model performance is assessed separately for TB and ADME-T assays. The first experiment compares the baseline methods and the uncertainty quantification approaches limited to the third temporal setting, in which three time spans were used as training datasets. The second section investigates the change in model performance over time by comparing the three temporal settings. Moreover, the results will be linked to the assay-specific conclusions about the label and data shift drawn in Section 3.1. The third experiment concentrates on the potential of post hoc probability calibration methods in the context of distribution shifts between the calibration and test sets. The results of the majority of models are presented. The numerical results of all methods are listed in Section 4 in the supplementary material.

Comparison of uncertainty estimation methods. We compared the predictive performance of RF, MLP, MLPE, MLPMC, and BNN in terms of AUC, BCE, and ACE. For a straightforward comparison, we limit the reported results to the third temporal setting, in which three time spans were used for model training. The data assigned to the last time span was used as a test set. The AUC values of the models are listed in Table 2. The AUC results for the TB assays show that the MLPs, as well as the non-Bayesian uncertainty estimation methods, outperform the RFs and BNNs on most datasets. In more detail, the MLPE model is always among the best approaches. The MLP and the MLPMC models retrieve results that are not statistically different from the best model in 6 and 5 out of 7 datasets. The BNN is the best model for assay TB-4, while the RF approach is consistently outperformed. In contrast, the results for the ADME-T assays show that either the MLPEs, or the BNNs, or both outperform the other approaches in 7 out of 8 assays. The remaining assay is A-CYPa, for which the RF approach achieves the best result.

The model calibration is analyzed using the ACE and the BCE scores. The results are illustrated in Fig. 6. The analysis of the BCE and ACE values reveals trends similar to those observed in the AUC scores. The results of the models trained on TB assays show that the MLPE approach

is among the best-performing approaches for all datasets, except for TB-1, for which MLPMC performs best in terms of ACE. The baseline MLP matches the performance of MLPE in 4 out of 7 times in terms of BCE and in 5 out of 7 assays in terms of ACE. Furthermore, the models perform overall worse on TB-1 and TB-2 in terms of BCE. A reason for this result could be the small dataset size of these two assays, which might lead to overfitting and, therefore, poorer calibration of the models. Furthermore, both assays exhibit comparatively large shifts in label and descriptor space, as shown in Fig. 4, which introduces additional difficulties in generalizing well over time. The results of the models trained on the ADME-T data demonstrate the superiority of the MLPE and BNN methods. More specifically, in all assays, either MLPE or the BNN performs best with regard to BCE and ACE. A-CYPa is the only exception for which RF and MLPMC perform best. The MLPE and BNN models achieve the best results across both metrics in the same number of assays, namely in 4 out of 8 cases.

It is surprising that the more sophisticated uncertainty estimation methods improve the probability calibration of the baselines trained on ADME-T assays but fail to do so for models trained on TB category datasets. The MLPE and the BNN approaches account for epistemic uncertainty by including model variance in their predictions. The deep ensemble approach retrieves model samples representing different local minima of the loss surface, while the BNN treats the neural network weights as probability distributions. Koh et al. [42] showed that the performance of a classifier can degrade significantly when there is a distribution shift between the source and target domains. Moreover, Garg et al. [89] reported specifically for the presence of label shift that a classifier that is ideal for the source domain might no longer be ideal for the target domain. In general, deep ensemble approaches have been shown to outperform other uncertainty estimation approaches, including approximate Bayesian neural networks, in terms of predictive accuracy and calibration without [90] and under distribution shift in the descriptor space [41,91]. These conclusions are supported by the results in this study, which show that MLPE performs better than other uncertainty estimation methods under distribution shift.

However, despite being the best uncertainty estimation approach, the MLPE models rarely outperform the baseline MLP in the presence of distribution shift. An additional reason for the failure of the uncertainty estimation methods to generate better uncertainty estimates is that the baseline could be their inability to handle the shifts in label space well. This conclusion might be transferrable to model calibration, thus explaining the difference between architectures trained on TB and ADME-T assays to produce better-calibrated probabilities. A large study that compared uncertainty estimation approaches used for regression models trained on the same dataset also reported that the deep ensemble and Bayesian neural network approaches outperform other common uncertainty estimation approaches for regression [50]. In contrast to the classification setting, the uncertainty estimates of baselines trained with TB assays could also be improved in the regression study. This observation could indicate that uncertainty estimation models for regression tasks are less sensitive to shifts in the label space than approaches for classification. A reason for this observation could be that the model can access the actual values of the measurements in the regression setting, which might attenuate the shift in the target space. For example, strongly inactive and weakly inactive compounds exhibit different target values for regression models, while this information is lost in classification tasks due to the application of binary classification thresholds.

Uncertainty estimation over time. We assessed the quality of the uncertainty estimates over time by comparing the model performance in all three temporal settings. The results are displayed in Fig. 7. The plots show that the model performance in one time span rarely allows conclusions about the performance of the same approach at another point in time. In 3 out of 7 TB assays, a single model is always among the models with the best BCE score in all three temporal settings. These

Table 2

Summary of AUC results for the third temporal setting. AUC results for the baselines (RF and MLP) and train-time uncertainty quantification models (MLPE, MLPMC, and BNN) are reported. The models were trained with compounds from three time spans. The mean and standard deviation of 10 model repetitions are shown. The best-performing approach and those not significantly different from it, as determined by a two-sided t-test, are highlighted in bold.

	RF	MLP	MLPE	MLPMC	BNN
Target-Based					
TB-1	0.437 ± 0.025	0.586 ± 0.065	0.614 ± 0.006	0.592 ± 0.061	0.502 ± 0.024
TB-2	0.778 ± 0.018	0.793 ± 0.01	0.795 ± 0.002	0.791 ± 0.01	0.289 ± 0.026
TB-3	0.708 ± 0.023	0.765 ± 0.009	0.768 ± 0.001	0.761 ± 0.009	0.73 ± 0.002
TB-4	0.909 ± 0.027	0.95 ± 0.007	0.956 ± 0.0	0.952 ± 0.003	0.957 ± 0.001
TB-5	0.676 ± 0.028	0.896 ± 0.008	0.9 ± 0.001	0.897 ± 0.008	0.78 ± 0.171
TB-6	0.641 ± 0.06	0.768 ± 0.007	0.771 ± 0.001	0.768 ± 0.007	0.701 ± 0.007
TB-7	0.533 ± 0.086	0.71 ± 0.026	0.718 ± 0.004	0.718 ± 0.03	0.479 ± 0.014
ADME-T					
A-CYPa	0.675 ± 0.015	0.625 ± 0.013	0.627 ± 0.002	0.625 ± 0.013	0.622 ± 0.016
A-CYPb	0.581 ± 0.012	0.644 ± 0.005	0.648 ± 0.001	0.643 ± 0.006	0.661 ± 0.001
A-CYPc	0.631 ± 0.024	0.714 ± 0.003	0.714 ± 0.0	0.715 ± 0.003	0.734 ± 0.003
A-PM	0.613 ± 0.031	0.769 ± 0.004	0.77 ± 0.001	0.765 ± 0.004	0.784 ± 0.023
A-SOL	0.692 ± 0.008	0.78 ± 0.011	0.792 ± 0.002	0.779 ± 0.013	0.511 ± 0.011
A-hERG	0.632 ± 0.009	0.729 ± 0.005	0.735 ± 0.001	0.729 ± 0.005	0.652 ± 0.003
A-logD	0.641 ± 0.015	0.833 ± 0.003	0.839 ± 0.002	0.834 ± 0.003	0.828 ± 0.004
A-MS	0.694 ± 0.006	0.71 ± 0.01	0.716 ± 0.003	0.713 ± 0.006	0.745 ± 0.007

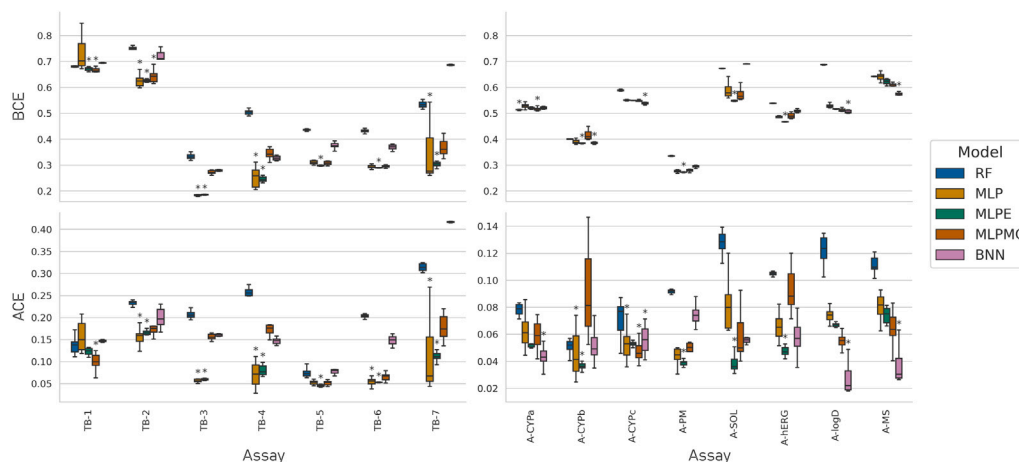


Fig. 6. Summary of BCE and ACE scores for the third temporal setting. The first column shows the results for TB assays, while the second one reports the performance of models trained on ADME-T assays. BCE scores are plotted in the first row, and ACE scores in the second row. Results for the baselines (RF and MLP) and train-time uncertainty quantification models (MLPE, MLPMC, and BNN) are reported. The models were trained with compounds from three time spans. The results of 10 model repetitions were aggregated. For each assay, the best-performing approach and those not significantly different from it, as determined by a two-sided t-test, are marked with an asterisk.

approaches include MLPMC for TB-1, MLP for TB-6, and MLPE for TB-7. Regarding the ACE scores, the MLPE is consistently among the best methods for TB-2, and the MLP is among the best for TB-3. The BCE results of the ADME-T assays show consistent results over time for 4 assays, including the RFs for A-CYPa and the MLPE approach for A-PM, A-SOL, and A-hERG. The ACE results of the ADME-T assays reveal a consistent model performance in 2 out of 8 assays, namely the MLPs for A-PM and the MLPE for A-hERG. In general, model performance in terms of ACE is slightly less stable over time than in terms of BCE. Interestingly, Svensson et al. [50] applied regression modeling techniques to the same data and reported more consistent model performance for the ADME-T assays.

Plotting the model performance on all temporal settings confirms the conclusions drawn in Section 3.2. The MLP and MLPE methods obtain the best results in terms of both BCE and ACE for the TB assays. Both are among the best-performing approaches in at least one temporal setting in all assays, except TB-4, where MLP is not among the best models in any setting. Interestingly, regarding the BCE results, the MLP is as often among the best models as the MLPEs, namely for 13 out of the 21 settings. The ACE scores reveal a similar result. MLP is among the best methods for 11, and MLPE for 13 out of 21 settings. All other approaches trained on the TB assays are much less often among the best-performing models.

As described in Section 3.1, TB-1[2] and TB-4[1] exhibited large shifts in label space and small shifts in descriptor space, which might indicate the presence of activity cliffs. Interestingly, the performance of models trained on these datasets did not differ considerably from the results of other TB assays. Thus, if the label shift was large, model performance was not affected significantly by the presence of shifts in descriptor space. A reason for this result could be the overall large shifts observed in TB assays, which makes the prediction tasks generally harder, diminishing the impact of activity cliffs on model performance. Furthermore, the combination of high label and low descriptor space shifts could also be an artifact of chemical series exhibiting activity values close to the classification threshold rather than actual activity cliffs, as described in Section 3.1.

The results for the ADME-T assays also support the results from the previous experiment. Concerning the ACE scores, the MLPE and BNNs outperform all other models, with MLPE being among the best methods for 13 and BNN for 11 out of 24 settings. The MLPs and the MLPMCs are among the best-calibrated methods in 5 and 7 settings, respectively. The MLPEs outperform the other approaches in terms of BCE, obtaining significant results in 13 settings. The MLPMCs are among the best methods in 8, and the BNNs in 7 out of 24 settings.

In conclusion, the MLPE approach generates the best-performing models for most ADME-T assays. Considering the high computational

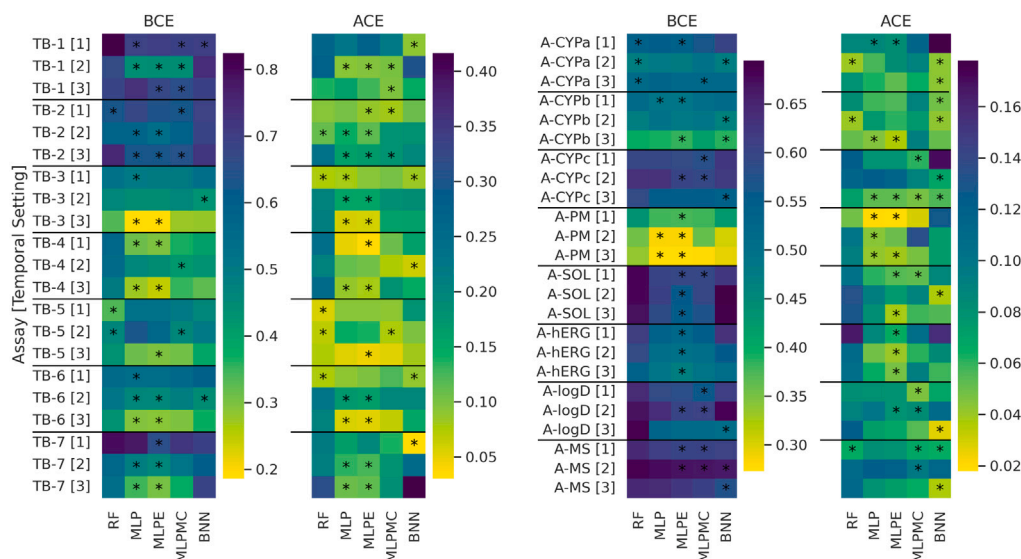


Fig. 7. Summary of BCE and ACE scores across all temporal settings. The first two columns show the ACE and BCE scores for TB assays, while the last two report the performance of models trained on ADME-T assays. The temporal setting is indicated in brackets after the assay abbreviation. Results for the baselines (RF and MLP) and train-time uncertainty quantification models (MLPE, MLPMC, and BNN) are reported. The models were trained with compounds from three time spans. Averages over 10 model repetitions are shown. For each assay, the best-performing approach and those not significantly different from it, as determined by a two-sided t-test, are marked with an asterisk.

resources required to train these MLPEs, another good choice are BNNs. This approach is much more time- and resource-effective and generates well-calibrated models for many ADME-T datasets. Note that a fixed variance was chosen for the Gaussian prior and that tuning this hyperparameter could even improve the performance of the BNNs. Given that deep ensemble models always lead to more underconfident predictions [92], the good performance of these methods indicates overfitting of the baselines. The MLPEs and BNNs counteract this behavior for the ADME-T assays but not for datasets from the TB category. The ACE values for TB assays are higher than for the ADME-T assays, as illustrated in Fig. 6, indicating worse model calibration. This result indicates room for improvement in probability calibration and eliminates good baseline model calibration as a reason for the inability of MLPE and BNN to improve the ACE of the MLP. An alternative reason for the failure of MLPEs and BNNs could be the shift in label space, which is considerably larger for TB assays.

Based on these results, it is questionable whether the costly generation of MLPEs for improved uncertainty estimation is justified for datasets with large distribution shifts, such as the TB data. As discussed in Section 3.2, this shift might be difficult to handle for the uncertainty quantification methods, resulting in the inability of uncertainty quantification methods to improve the probability calibration.

Post hoc probability calibration. The effects of two post hoc calibration methods, Platt scaling and Venn-ABERS predictors, were assessed. Furthermore, the distribution shift between calibration and test sets was taken into account by obtaining the MMD between the two datasets. The MMD between the calibration and test datasets is reported in Table S4 in the supplementary material. For the sake of clarity, we only include the results of the third temporal setting for the MLPE approach, which was reported to be one of the best-performing models in previous sections of this study. Only the ACE is reported since the post hoc calibration step includes the application of monotonously increasing functions, which cannot correct non-monotonous distortions and, therefore, does not change the ranking of the predictions. Hence, these approaches only affect the probability calibration while the AUC scores of the models remain constant. The results of all calibrated models across all temporal settings and assays can be found in Section 4 in the supplementary material.

Fig. 8 shows the performance of the Platt-scaled MLPEs (MLPE-P) and the MLPEs calibrated with Venn-ABERS predictors (MLPE-VA). The

results are shown separately for TB and ADME-T assays, and the assays are ordered according to increasing MMD values. In general, post hoc scaling leads to better results in 4 out of 7 TB assays and in 4 out of 8 ADME-T assays. In 4 out of these 8 cases, both MLPE-P and MLPE-VA are the best models, while in 4 cases, either MLPE-P or MLPE-VA performs best. Both panels in Fig. 8 show that with increasing MMD between calibration and test set, the calibrating abilities of the post hoc calibration methods decrease. A stronger trend can be detected for the TB assays, for which the shifts are larger than for the ADME-T assays. Nevertheless, the pattern is also visible in the right panel plotting the results for ADME-T, which show no improvements after post hoc calibration for assays with large MMD, like A-CYPc and A-CYPa. The reported trends can also be seen for other approaches, albeit less clearly for some models.

An intuitive explanation for this pattern is that post hoc probability approaches perform better if the calibration set is similar to the test set, and worse if the calibration and test sets are different. Ovadia et al. [41] showed that post hoc calibration approaches improve model calibration in the i.i.d. setting. However, their calibrating abilities degrade as the data shift increases, so that they are ultimately outperformed by train-time uncertainty quantification methods in the presence of large shifts. These findings are supported by the post hoc calibration results in this study, which report good calibrating properties of the post hoc calibration methods in the case of small shifts, while for larger shifts, they are outperformed by the train-time versions of the methods.

4. Conclusions

Uncertainty estimation emerges as a critical tool in the cost- and resource-intensive drug discovery process, facilitating the evaluation of experimental risks and costs. In this framework, the quality of the uncertainty estimates is crucial to ensure the reliability of the models. This study evaluates uncertainty estimation approaches for classification tasks in a practical, real-world context.

A temporal splitting strategy was applied to internal data from a pharmaceutical company to simulate the evolution of drug-target interaction data. This splitting approach allowed us to determine distribution shifts in pharmaceutical data over time and to identify uncertainty estimation methods that remain reliable as assay data evolves and distributional shifts occur. In pharmaceutical research, datasets are

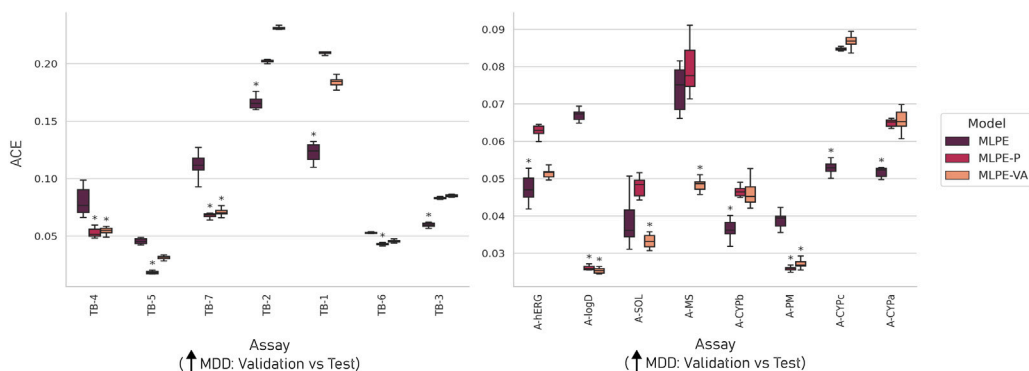


Fig. 8. Summary of ACE scores of post-hoc probability calibration approaches using the third temporal setting. The left panel shows the ACE scores of models trained on TB assays, while the right panel reports the ACE performance of ADME-T models. The assays in each panel are ordered according to increasing distribution shifts in descriptor space between the calibration and the training set. The distribution shift is determined by the maximum mean discrepancy (MMD) between the two datasets using the Tanimoto coefficient on the ECFP space. Results for the deep ensembles (MLPE), the Platt-scaled deep ensembles (MLPE-P), and the ensembles calibrated with a Venn-ABERS predictor (MLPE-VA) are reported. The models were trained with compounds from three time spans. The results of 10 model repetitions were aggregated. For each assay, the best-performing approach and those not significantly different from it, as determined by a two-sided t-test, are marked with an asterisk.

continually updated, requiring models to be retrained. It is therefore essential to assess whether a chosen uncertainty estimation strategy continues to deliver accurate uncertainty estimates as the data evolves or whether adjustments are necessary. Our findings offer valuable insights into uncertainty estimation and highlight the challenges posed by real-world applications.

The analysis of the pharmaceutical data showed that the distribution shifts in label and descriptor space over time strongly depended on the nature of the individual assays. The project-specific TB assays exhibited more pronounced distribution shifts, while the more abundantly used toxicity screens and assays assessing the pharmacodynamic properties (ADME-T assays) showed moderate and more stable shifts in descriptor space and little shift in label space over time. These results suggest that the i.i.d. assumption might not be accurate, especially for target-specific assays with larger shifts in label- and descriptor space. In general, real-world data is likely to exhibit distribution shifts over time, particularly in assay data, where chemical series are naturally present. Hence, we also expect the ADME-T data not to be completely i.i.d.. However, in the case of the ADME-T assays, the distributions between training and test are more similar than for TB assays, as shown in Fig. 4.

A comparison of common uncertainty quantification methods revealed that the deep ensembles and the Bayesian neural network achieved the best-calibrated results for ADME-T assays that show small distribution shifts in the data. Recently, these two approaches have also been shown to outperform other common uncertainty quantification methods in regression tasks using the same dataset [50]. Given that training deep ensembles demands a lot of computational resources, Bayesian neural networks might be the best choice when striving for a fast method that produces well-calibrated estimates. For TB assays exhibiting more pronounced distribution shifts in both label and descriptor space, the deep ensemble method performed best. However, the baseline MLP matched the performance of the uncertainty quantification methods for many datasets. Due to the high computational effort required to train deep ensembles, choosing a simple MLP for assays with large distribution shifts might be the best and most efficient solution.

Surprisingly, most ensemble-based and full Bayesian uncertainty estimation methods did not outperform the baseline MLP for TB assays. Recall that these approaches infer probability distributions over model parameters, thus allowing the generation of a predictive distribution. Assuming the model is well-calibrated, the variance of its predictive distributions should reflect its confidence. More specifically, the variance of these predictive distributions is expected to increase when the model cannot make accurate predictions, such as when the input data differs from the training distribution. Therefore, well-calibrated models that account for model variance should be able to handle a lack

of generalizability due to shifts in the descriptor space. This leads to the assumption that their inability to produce better-calibrated results might stem from the shifts in label space rather than the descriptor space. Furthermore, Svensson et al. [50] used the TB assays in regression models and reported an improved model calibration using ensemble-based models and Bayesian neural networks. Thus, these approaches might be able to generate better-calibrated results in a regression setting, when the actual bioactivity values can be accessed rather than less-informative binary labels.

The analysis of the performance of uncertainty estimation approaches over time showed that it is difficult to draw conclusions from results from one point in time in the assay history to another. In general, model performance was unstable over time. Only for a few assays could one method be identified that was among the best-performing approaches at all considered time points in the assay history. In conclusion, a reevaluation of classification model performance is required for all assays as soon as more recent data is added. Interestingly, the performance of regression models is more stable over time for ADME-T assays, as shown recently by Svensson et al. [50].

Lastly, two post hoc calibration approaches, including Platt scaling and Venn-ABERS predictors, were tested on their ability to improve the probability calibration of deep ensembles. Overall, the calibration methods were able to achieve better-calibrated results for some assays, while for others, the calibration stayed the same or even deteriorated after the post hoc calibration step. The calibrating capabilities were dependent on the distribution shift between the calibration and test sets, with declining performance when the shifts in the descriptor space increased.

The uncertainty estimation approaches discussed in this study have been previously demonstrated to improve model calibration on toy data or datasets that do not account for the distribution shift caused by the evolution of the data over time. However, previous studies show that improved performance on i.i.d. data often fails to translate into better outcomes under distribution shift [41,42], which is also supported by the findings in this study. This study highlights the challenges introduced by real-world data, emphasizing the complexity of identifying effective strategies for uncertainty estimation in QSAR models.

CRediT authorship contribution statement

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

Adam Arany: Writing – review & editing, Supervision, Methodology, Funding acquisition. **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.

Funding sources

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 with Leuven.AI and received funding from the Flemish Government (AI Research Program).

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

The authors sincerely thank their colleagues for the useful discussions. Special appreciation goes to András Formanek and Antoine Passemiers at KU Leuven for their valuable insights. For this work, the editors have granted a waiver on data release requirements for publication of original research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.aills.2025.100132>.

Data availability

The data that has been used is confidential.

References

- Singh N, Vayer P, Tanwar S, Poyet JL, Tsaïoun K, Villoutreix BO. Drug discovery and development: introduction to the general public and patient groups. *Front Drug Discov* 2023;3:1201419.
- Laermann-Nguyen U, Backfisch M. Innovation crisis in the pharmaceutical industry? A survey. *SN Bus Econ* 2021;1(12):164.
- Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. *Nature* 2023;616(7958):673–85.
- Hertzberg RP, Pope AJ. High-throughput screening: New technology for the 21st century. *Curr Opin Chem Biol* 2000;4(4):445–51.
- Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 2009;25(18):2397–403.
- Apostolakis G. The concept of probability in safety assessments of technological systems. *Science* 1990;250(4986):1359–64.
- Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn* 2019;110:457–506. <http://dx.doi.org/10.1007/s10994-021-05946-3>.
- Gruber C, Schenk PO, Schierholz M, Kreuter F, Kauermann G. Sources of uncertainty in machine learning – A statisticians' view. 2023. <http://dx.doi.org/10.48550/arXiv.2305.16703>, arXiv:2305.16703.
- Neal RM. In: *Bayesian learning for neural networks*, vol. 118, Springer Science & Business Media; 2012.
- Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural network. In: *International conference on machine learning*. PMLR; 2015, p. 1613–22.
- Izmailov P, Vikram S, Hoffman MD, Wilson AGG. What are Bayesian neural network posteriors really like? In: *International conference on machine learning*. PMLR; 2021, p. 4629–40.
- Kim Q, Ko JH, Kim S, Park N, Jhe W. Bayesian neural network with pretrained protein embedding enhances prediction accuracy of drug-protein interaction. *Bioinformatics* 2021;37(20):3428–35.
- Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. *Advances in neural information processing systems*, vol. 30, Curran Associates, Inc.; 2017, URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.
- Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Balcan MF, Weinberger KQ, editors. *Proceedings of the 33rd international conference on machine learning*. Proceedings of machine learning research, vol. 48, New York, New York, USA: PMLR; 2016, p. 1050–9, URL: <https://proceedings.mlr.press/v48/gal16.html>.
- Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. 2021, arXiv preprint arXiv:2107.07511.
- Taquet V, Blot V, Morzadec T, Lacombe L, Brunel N. MAPIE: an open-source library for distribution-free uncertainty quantification. 2022, arXiv preprint arXiv:2207.12274.
- Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. *Adv Neural Inf Process Syst* 2018;31.
- 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.
- Wang D, Wu Z, Shen C, Bao L, Luo H, Wang Z, et al. Learning with uncertainty to accelerate the discovery of histone lysine-specific demethylase 1A (KDM1A/LSD1) inhibitors. *Brief Bioinform* 2023;24(1):bbac592.
- Oh D, Shin B. Improving evidential deep learning via multi-task learning. In: *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, 2022, p. 7895–903.
- Soleimany AP, Amini A, Goldman S, Rus D, Bhatia SN, Coley CW. Evidential deep learning for guided molecular property prediction and discovery. *ACS Central Sci* 2021;7(8):1356–67.
- Liu J, Lin Z, Padhy S, Tran D, Bedrax Weiss T, Lakshminarayanan B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Adv Neural Inf Process Syst* 2020;33:7498–512.
- Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classif* 1999;10.
- Vovk V, Petej I. Venn-ubers predictors. 2014, <http://dx.doi.org/10.48550/arXiv.1211.0025>, arXiv:1211.0025.
- Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining*. 2002, p. 694–9.
- 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>.
- Yu J, Wang D, Zheng M. Uncertainty quantification: Can we trust artificial intelligence in drug discovery? *Iscience* 2022;25(8).
- Dheur V, Taieb SB. A large-scale study of probabilistic calibration in neural network regression. In: *International conference on machine learning*. PMLR; 2023, p. 7813–36.
- Schweighofer K, Aichberger L, Ielanskyi M, Klambauer G, Hochreiter S. Quantification of uncertainty with adversarial models. *Adv Neural Inf Process Syst* 2023;36:19446–84.
- Mervin LH, Trapotsi MA, Afzal AM, Barrett IP, Bender A, Engkvist O. Probabilistic random forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty. *J Cheminform* 2021;13:1–17.
- Rayka M, Mirzaei M, Mohammad Latifi A. An ensemble-based approach to estimate confidence of predicted protein–ligand binding affinity values. *Mol Inform* 2024;43(4):e202300292.
- Fan Z, Yu J, Zhang X, Chen Y, Sun S, Zhang Y, et al. Reducing overconfident errors in molecular property classification using posterior network. *Patterns* 2024.
- Friesacher HR, Engkvist O, Mervin L, Moreau Y, Arany A. Achieving well-informed decision-making in drug discovery: a comprehensive calibration study using neural network-based structure-activity models. *Journal of Cheminformatics* 2025;17(1):29.
- 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.
- Mervin LH, Afzal AM, Engkvist O, Bender A. Comparison of scaling methods to obtain calibrated probabilities of activity for protein–ligand predictions. *J Chem Inf Model* 2020;60(10):4546–59. <http://dx.doi.org/10.1021/acs.jcim.0c00476>, PMID: 32865408.

- [36] Dutschmann TM, Kinzel L, Ter Laak A, Baumann K. Large-scale evaluation of k-fold cross-validation ensembles for uncertainty estimation. *J Chem Inf Model* 2023;15(1):49.
- [37] 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>.
- [38] Landrum GA, Beckers M, Lanini J, Schneider N, Stiefl N, Riniker S. SIMPD: an algorithm for generating simulated time splits for validating machine learning approaches. *J Chem Inf Model* 2023;15(1):119.
- [39] Dunder M, Krishnapuram B, Bi J, Rao RB. Learning classifiers when the training data is not IID. In: *IJCAI*, vol. 2007, Citeseer; 2007, p. 756–61.
- [40] Cao L. Beyond iid: Non-iid thinking, informatics, and learning. *IEEE Intell Syst* 2022;37(4):5–17.
- [41] 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: Wallach H, Larochelle H, Beygelzimer A, d'Alché Buc F, Fox E, Garnett R, editors. *Advances in neural information processing systems*. 32, Curran Associates, Inc.; 2019, URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf.
- [42] Koh PW, Sagawa S, Marklund H, Xie SM, Zhang M, Balsubramani A, et al. Wilds: A benchmark of in-the-wild distribution shifts. In: *International conference on machine learning*. PMLR; 2021, p. 5637–64.
- [43] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Precup D, Teh YW, editors. *Proceedings of the 34th international conference on machine learning*. Proceedings of machine learning research, vol. 70, PMLR; 2017, p. 1321–30, URL: <https://proceedings.mlr.press/v70/guo17a.html>.
- [44] Sheridan RP. Three useful dimensions for domain applicability in QSAR models using random forest. *J Chem Inf Model* 2012;52(3):814–23.
- [45] Abe T, Buchanan EK, Pleiss G, Zemel R, Cunningham JP. Deep ensembles work, but are they necessary? *Adv Neural Inf Process Syst* 2022;35:3364–60.
- [46] Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd international conference on machine learning*. 2005, p. 625–32.
- [47] Zdrzil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, et al. The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res* 2024;52(D1):D1180–92.
- [48] Rodríguez-Pérez R, Trunzer M, Schneider N, Faller B, Gerebtzoff G. Multispecies machine learning predictions of in vitro intrinsic clearance with uncertainty quantification analyses. *Mol Pharm* 2022;20(1):383–94.
- [49] Stoyanova R, Katzberger PM, Komissarov L, Khadraoui A, Sach-Peltason L, Groebke Zbinden K, et al. Computational predictions of nonclinical pharmacokinetics at the drug design stage. *J Chem Inf Model* 2023;63(2):442–58.
- [50] Svensson E, Friesacher HR, Winiwarter S, Mervin L, Arany A, Engkvist O. Enhancing uncertainty quantification in drug discovery with censored regression labels. *Artif Intell Life Sci* 2025;100128.
- [51] Friesacher HR, Svensson E, Arany A, Mervin L, Engkvist O. Temporal evaluation of probability calibration with experimental errors. In: *International workshop on AI in drug discovery*. Springer; 2024, p. 13–20.
- [52] 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.
- [53] DiMasi JA. Risks in new drug development: approval success rates for investigational drugs. *Clin Pharmacol Ther* 2001;69(5):297–307.
- [54] Van De Waterbeemd H, Gifford E. ADMET in silico modelling: towards prediction paradise? *Nat Rev Drug Discov* 2003;2(3):192–204.
- [55] Deodhar M, Al Rihani SB, Arwood MJ, Darakjian L, Dow P, Turgeon J, et al. Mechanisms of CYP450 inhibition: understanding drug-drug interactions due to mechanism-based inhibition in clinical practice. *Pharmaceutics* 2020;12(9):846.
- [56] Ioannides C. *Cytochromes P450: metabolic and toxicological aspects*. CRC Press; 1996.
- [57] Furge LL, Guengerich FP. Cytochrome P450 enzymes in drug metabolism and chemical toxicology: An introduction. *Biochem Mol Biol Educ* 2006;34(2):66–74.
- [58] Shah P, Jogani V, Bagchi T, Misra A. Role of Caco-2 cell monolayers in prediction of intestinal drug absorption. *Biotechnol Prog* 2006;22(1):186–98.
- [59] Di L, Fish PV, Mano T. Bridging solubility between drug discovery and development. *Drug Discov Today* 2012;17(9–10):486–95.
- [60] Keating MT, Sanguinetti MC. Molecular genetic insights into cardiovascular disease. *Science* 1996;272(5262):681–5.
- [61] Waring MJ. Lipophilicity in drug discovery. *Expert Opin Drug Discov* 2010;5(3):235–48.
- [62] Chen M, Borlak J, Tong W. High lipophilicity and high daily dose of oral medications are associated with significant risk for drug-induced liver injury. *Hepatology* 2013;58(1):388–96.
- [63] Masimirembwa CM, Thompson R, Andersson TB. In vitro high throughput screening of compounds for favorable metabolic properties in drug discovery. *Comb Chem High Throughput Screen* 2001;4(3):245–63.
- [64] Di L, Kerns EH, Hong Y, Kleintop TA, Mc Connell OJ, Hury DM. Optimization of a higher throughput microsomal stability screening assay for profiling drug discovery candidates. *SLAS Discov* 2003;8(4):453–62.
- [65] Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;50(5):742–54.
- [66] Jiang D, Wu Z, Hsieh C-Y, Chen G, Liao B, Wang Z, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J Chem Inf Model* 2021;13:1–23.
- [67] 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.
- [68] Landrum G. RDKit: Open-source chemin. 2006, <http://dx.doi.org/10.5281/zenodo.6961488>.
- [69] 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.
- [70] 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.
- [71] 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.
- [72] Paszke A, et al. PyTorch: An imperative style, high-performance deep learning library. In: *Advances in neural information processing systems*, vol. 32, Curran Associates, Inc.; 2019.
- [73] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: *International conference on neural representations*. 2015.
- [74] Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat* 1951;22(1):79–86.
- [75] Toccaceli P, Nouredinov I, Luo Z, Vovk V, Carlsson L, Gammerman A. ExCAPE WPI-probabilistic prediction. Royal Holloway; 2016.
- [76] Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. *J Mach Learn Res* 2012;13(1):723–73.
- [77] Rabanser S, Günnemann S, Lipton Z. Failing loudly: An empirical study of methods for detecting dataset shift. *Adv Neural Inf Process Syst* 2019;32.
- [78] Ouyang L, Key A. Maximum mean discrepancy for generalization in the presence of distribution and missingness shift. 2021, arXiv preprint arXiv:2111.10344.
- [79] Gao R, Liu F, Zhang J, Han B, Liu T, Niu G, et al. Maximum mean discrepancy test is aware of adversarial attacks. In: *International conference on machine learning*. PMLR; 2021, p. 3564–75.
- [80] Yan H, Ding Y, Li P, Wang Q, Xu Y, Zuo W. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, p. 2272–81.
- [81] Zhu Y, Zhuang F, Wang J, Ke G, Chen J, Bian J, et al. Deep subdomain adaptation network for image classification. *IEEE Trans Neural Netw Learn Syst* 2020;32(4):1713–22.
- [82] Borgwardt KM, Gretton A, Rasch MJ, Kriegel HP, Schölkopf B, Smola AJ. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 2006;22(14):e49–57.
- [83] Lee S, Jo J, Hwang SJ. Exploring chemical space with score-based out-of-distribution generation. In: *International conference on machine learning*. PMLR; 2023, p. 18872–92.
- [84] Holliday JD, Hu C, Willett P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb Chem High Throughput Screen* 2002;5(2):155–66.
- [85] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29–36.
- [86] Nixon J, Dusenberry MW, Zhang L, Jerfel G, Tran D. Measuring calibration in deep learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops*. 2019, <http://dx.doi.org/10.48550/arXiv.1904.01685>.
- [87] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Amer Statist Assoc* 2007;102(477):359–78. <http://dx.doi.org/10.1198/01621450600001437>, arXiv:https://doi.org/10.1198/01621450600001437.
- [88] Bröcker J. Reliability, sufficiency, and the decomposition of proper scores. *Q J R Meteorol Soc: A J Atmos Sci Appl Meteorol Phys Ocean* 2009;135(643):1512–9.
- [89] Garg S, Wu Y, Balakrishnan S, Lipton Z. A unified view of label shift estimation. *Adv Neural Inf Process Syst* 2020;33:3290–300.
- [90] Gustafsson FK, Danelljan M, Schon TB. Evaluating scalable bayesian deep learning methods for robust computer vision. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, p. 318–9.
- [91] Mehrtens HA, Kurz A, Bucher TC, Brinker TJ. Benchmarking common uncertainty estimation methods with histopathological images under domain shift and label noise. *Med Image Anal* 2023;89:102914.
- [92] Rahaman R, Thiery A. Uncertainty quantification and deep ensembles. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang P, Vaughan JW, editors. *Advances in neural information processing systems*, vol. 34, Curran Associates, Inc.; 2021, p. 20063–75, URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/a70dc40477bc2adceef4d2c90f47eb82-Paper.pdf.