Article:
Hosni, Z., Gillet, V.J. and Marchese Robinson, R.L. (2026) Explicit applicability domain calculations can help determine when uncertainty estimates are less reliable. ACS Omega, 11 (3). pp. 4722-4743. ISSN: 2470-1343

https://doi.org/10.1021/acsomega.5c11875

ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Explicit Applicability Domain Calculations Can Help Determine When Uncertainty Estimates Are Less Reliable

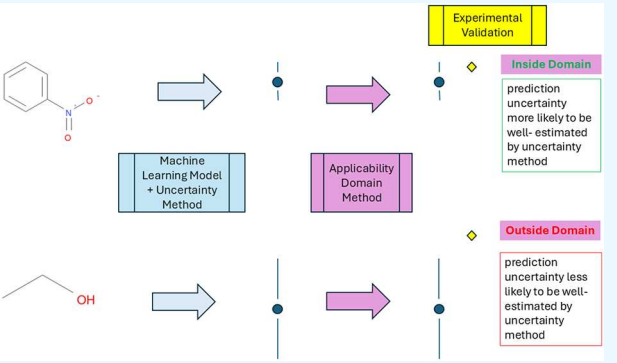
Zied Hosni,[§] Valerie J. Gillet,[*] and Richard L. Marchese Robinson[*,§]

Cite This: *ACS Omega* 2026, 11, 4722−4743

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Quantifying the uncertainty associated with a QSAR prediction is hugely valuable. Conformal regression and Venn-ABERS have emerged as state-of-the-art uncertainty estimation methods for regression and classification QSAR models, respectively. However, their performance is limited when they are applied to compounds sampled from a different distribution to the data used to train the model and/or calibrate their uncertainty estimates. Previous studies have evidenced this when applying these methods to nonrandom train/test splits, e.g., temporal validation, cluster or scaffold splits. Building on these previous studies, we demonstrate that explicit applicability domain calculations, using only structural similarity, can help determine when these uncertainty estimates are less reliable for molecules encountered after model building. By less reliable, we mean the uncertainty estimates for out-of-domain predictions are less likely to reflect the empirically observed model residuals (regression) or probability of observing the predicted class experimentally (classification). After briefly comparing different methods using exemplar data sets, we extensively investigated the implications of computed applicability domain status for uncertainty estimation reliability using a k-nearest neighbors applicability domain approach (nUNC), in combination with Cross-Venn-ABERS Predictors (classification) or Aggregated Conformal Prediction (regression) uncertainty estimation across a wide range of public data sets. Because these are more representative of real-world applications, we focus on the results obtained on nonrandom test sets: temporal and cluster splits defined in previous modeling studies. We also present results for multiple temporal splits (time-splits) of classification and regression industrial data sets. In most cases, we found that nUNC was capable of distinguishing between molecules where the uncertainty estimates were, on average, more (inside the domain) vs less (outside the domain) reliable.



INTRODUCTION

When applying a Quantitative Structure−Activity Relationship (QSAR), it is essential to assess the uncertainty associated with the predictions. In recent years, this topic has received considerable attention in the cheminformatics community.[1−5] Approaches for assessing the uncertainty in the predictions are useful for active learning[6] and multiparameter optimization,[7,8] as well as deciding whether individual predictions are sufficiently reliable to be used for decision making.[5] The main approaches can be distinguished as follows: (i) applicability domain methods which aim to differentiate between predictions that can generally be expected to be reliable (inside the domain) vs those which cannot (outside the domain);[2,5] (ii) continuous metrics which seek to rank compounds according to the expected reliability of their predictions;[9] and (iii) calibrated uncertainty estimation methods which seek to quantify the reliability associated with individual predictions.[3,4]

Here, it should be acknowledged that the terminology around these concepts is not always used entirely consistently in the literature.[5] Indeed, it has been argued that uncertainty estimation approaches can themselves serve to define the applicability domain,[10,11] implying that a distinct definition is redundant. However, this is not a safe assumption. Indeed, Hanser et al. argued that what they term "applicability", "reliability" and "decidability" should be considered in a stepwise fashion.[5] What they term "applicability" and/or "reliability" are addressed by common applicability domain methods. Subsequently, what they term "decidability" would be addressed by uncertainty estimation methods.

The claims of Hanser et al.[5] chime with claims[12] that the uncertainty in predictions arises from three distinct sources: aleatoric (noise in training data), epistemic (uncertainty arising from the modeling process, e.g., due to limited training data), and distributional (new molecules come from a different distribution than data seen during training or calibration of

Received: November 11, 2025
Accepted: December 9, 2025
Published: January 9, 2026

uncertainty estimates). That said, distributional shift can also be seen as a source of epistemic uncertainty.[13]

Two frameworks for quantifying the uncertainty in individual predictions have received particular attention in the cheminformatics community: conformal prediction[1,4,14] and Venn-ABERS.[15−18] They have been proposed for industrial[10,15,19,20] and, in the case of conformal prediction, regulatory[21] applications.

Conformal prediction[1,14,19,22,23] produces prediction intervals which, under the assumption of data "exchangeability", are guaranteed or, for some variations,[14,24] simply expected, on average,[19] to contain the true value with a predefined confidence level. It has also been claimed that this guarantee for the prediction intervals holds for the measured values in practice.[19] These prediction intervals are obtained using a calibration set, which is typically sampled from the same distribution as the training data. In practice, the assumption of "exchangeability" means that new compounds being predicted are expected to come from the same distribution used to build the model and/or calibrate the prediction intervals.[14] When applied to regression tasks, conformal prediction may be referred to as "conformal regression".[4]

Venn-ABERS[15−17] is another calibrated method that is applicable to the modeling of categorical properties (classification) and produces upper and lower bounds on the probability of a particular class label being correct. These bounds can be fused to give a single estimate of the probability that a particular class label is correct. Assuming the raw output of the model used during calibration of the probability estimates is monotonically related to the class labels, these upper and lower bounds of the probability are guaranteed to contain the true class probability,[16,17] indicating that the fused probability estimate should be reliable. By "monotonically related to the class labels", we mean that, for a binary classifier discriminating between actives and inactives, increasing the value of the raw output from the model increases the chance of belonging to the active class.

Given the assumptions of exchangeability and monotonicity, which underpin the reliability of conformal prediction and Venn-ABERS, respectively, it follows that neither can be expected to produce reliable uncertainty estimates for compounds selected from different distributions from those used to build the models and/or calibrate their uncertainty estimates. Indeed, some studies have demonstrated this empirically, by observing that moving to new chemicals not drawn from the same distribution as the training/calibration data can result in a reduction in the ability of conformal prediction[25−28] or Venn-ABERS[15,20,29] to produce reliable uncertainty estimates. In these prior studies, the selection criteria for the test sets were expected to result in a high number of compounds from a different distribution from the training/calibration data. For example, the test sets were reported at a later time point,[20,26,28,29] were based on scaffold-split cross-validation,[15] cluster-split cross-validation[27] or different activity ranges to the training set.[25] We hypothesized that some of these molecules may have been sufficiently similar to the training and/or calibration set for uncertainty estimation to be more reliable than was observed on the test set as a whole.

For practical purposes, it would be hugely valuable to have an applicability domain method that can flag untested compounds as being sufficiently different from those seen during training/calibration that the uncertainty estimates of conformal regression or Venn-ABERS are no longer sufficiently reliable. Importantly, this applicability domain method should be effective for new molecules that may come from novel regions of chemical space without experimental data available to update the calibration of the uncertainty estimates[26,28] or refine the domain.[30]

By 'no longer sufficiently reliable', we mean that the uncertainty estimates for out-of-domain predictions poorly reflect the empirically observed model residuals, for regression, or probability of observing the predicted class experimentally, for classification. We do not simply mean that the predictions become less reliable outside the domain, and the estimated uncertainty becomes higher. Indeed, if the estimated uncertainty merely increased to reflect the reduced reliability in the predictions, the uncertainty estimates would reliably characterize the higher prediction uncertainty. Rather, we expect the uncertainty estimates to be less likely to reflect the reliability of the predictions outside the domain.

In contrast, prior research into applicability domain methods has focused on choosing an appropriate method that ensures that compounds with less reliable predictions tend to lie outside the domain and those with more reliable predictions tend to lie inside the domain.[2,31,32] However, despite the studies mentioned above demonstrating reduced performance of uncertainty estimates for compounds not sampled from the same distribution as the training/calibration set, we were not aware of any prior study that had explicitly investigated the link between applicability domain calculations and the reliability of uncertainty estimates. Knowing whether an applicability domain method could split a given set of newly encountered molecules into those for which uncertainty estimation would be more (inside the domain) vs less (outside the domain) likely to be reliable would be valuable for real-world applications of a model.

Hence, the focus of this study was to systematically explore how the reliability of the uncertainty estimates generated using conformal regression and Venn-ABERS depends on an explicitly computed applicability domain (AD) status.

METHODS

Overview

We built classification and regression models for a variety of public and proprietary data sets, using uncertainty methods to estimate the uncertainty in their predictions. We evaluated the performance of the predictions and uncertainty estimates on random (randomly sampled from the same distribution as the training set) and nonrandom test sets (temporal or cluster splits). For classification tasks, we used Cross-Venn-ABERS Predictors (CVAP),[17] and for regression tasks, we used Aggregated Conformal Prediction (ACP).[33] These uncertainty methods have been shown to typically outperform uncalibrated uncertainty methods and other common variations of Venn-ABERS and Conformal Prediction in prior studies.[1,15,16] (We initially explored some alternative uncertainty methods using the random test sets for a few exemplar public data sets, including common variations of conformal regression[22,34] and Venn-ABERS,[16,17] and concluded ACP and CVAP performed best overall.) We chose Random Forest as the underlying modeling method due to its suitability for a range of QSAR modeling tasks, without the need for extensive tuning.[35,36] For both the modeling and similarity calculations required by the AD method, we represented the compounds using widely employed Morgan fingerprints.[15,37]

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

We then considered how well an AD method we refer to as nUNC, which is loosely adapted from the literature[12,38] (see below), served to differentiate between reliable (inside the domain) and less-reliable (outside the domain) predictions and uncertainty estimates. (We initially explored some alternative AD methods based on the literature,[9,32,39] using a few exemplar public data sets, and concluded nUNC was a reasonable default, even though we do not claim it was optimal for all data sets.) We evaluated the effectiveness of the AD method based on what we term "shift metrics" (described below) and focused on results obtained following the splitting of the nonrandom test sets into inside and outside domain subsets. We focus on the nonrandom test sets as these more closely resemble real-world applications than random test sets.

We note that we make no claim that the uncertainty and AD methods we investigated in our work are optimal for all data sets. Identifying the exact AD method, parameters, and variations of conformal regression and Venn-ABERS, or other uncertainty methods, which are optimal for specific data sets, is a question we leave for future studies. Rather, our work presents a proof-of-concept that explicitly dividing molecules into inside and outside of domain compounds, using an AD method based on structural similarity alone, can identify molecules for which uncertainty methods are less likely to behave reliably.

Selection of Public Data Sets

The public data sets we investigated were previously used to assess the loss of calibration in uncertainty methods using nonrandom train/test splits: the classification data sets by Morger et al. based on data from Tox21[26] and ChEMBL[28] and the ChEMBL regression data sets studied by Wang et al.[27] We refer to the data sets from these publications as the Morger-Tox21, Morger-ChEMBL and Wang-ChEMBL data sets, or "dataset groups", respectively. Each of these data-set groups comprises multiple individual data sets that correspond to *in vitro* assay data against some target. We refer to individual public data sets as targets or endpoints.

Public Classification Data Sets

All the classification tasks were binary, with the active/toxic compounds denoted as members of "class 1" (terminology used when discussing modeling and uncertainty methods).

The Morger-Tox21 data sets[26] were originally assembled for the Tox21 Data Challenge with a training set, called Tox21Train, and two test sets, called Tox21Test and Tox21Score, made available for each target. These two test sets were chronologically released,[26] rather than being randomly sampled from the same distribution as Tox21Train. Hence, we refer to them as nonrandom test sets.

Each of the Morger-ChEMBL[28] data sets was available as a training set and two test sets labeled Update1 and Holdout, which are both derived from chronological splits, based on the date of publication, with Holdout being the most recent. Hence, we refer to them as nonrandom test sets. Morger et al.[28] also describe a third set of test sets, Update2, but we did not use these to limit the study to the same number of nonrandom test sets as the Morger-Tox21 data sets.

For both the Morger-Tox21 and Morger-ChEMBL data sets, the literature training sets were randomly split, using stratified sampling, into subsets of 80% for training and 20% testing, even though we focus on the results obtained for the nonrandom test sets defined in the literature (see above), as these are more representative of real-world applications. The same random 80% subset was used for training models applied to all test sets, to avoid confounding the comparisons of results on different test sets by changes in the training set.

Public Regression Data Sets

In contrast to the classification data sets described above, literature-defined time-splits were not available for the chosen public regression data sets. Instead, Wang et al.[27] defined different splitting strategies, of which "IVIT" and "IVOT" were relevant to our work. We use these terms for their data sets only in order to be clear that we used the same folds. In all cases, they followed a 5-fold cross-validation split on the full data set. Here, we used the same IVIT and IVOT folds provided by Wang et al.,[27] with each fold being used in turn as a test set, with the remaining data being used as the training set. In both cases, we report results obtained over all five folds. Note that for both IVIT and IVOT, in contrast to Wang et al.[27] who treated one of the remaining folds as a validation set, we treated all other folds as the training set.

As described in Wang et al.[27] the IVIT (In-Domain Validation set, In-Domain Test set) folds were obtained via a standard random splitting approach where each fold was sampled from the same distribution. Contrastingly, as previously described,[27] the IVOT (In-Domain Validation set, Out-of-Domain Test set) folds were based on 5-fold cluster cross-validation. For that approach, the compounds were first clustered, such that the minimum distance (1 - Tanimoto similarity, based on a binary fingerprint) between molecules in different clusters was 0.3. The clusters were then combined to form five folds of approximately equal size. We consider each of the IVOT folds to be a nonrandom test set.

Industrial Classification Data Set

This is a classification data set of approximately 5000 compounds from the Syngenta corporate database, where the endpoint is a binary categorical measure of soil persistence based upon the half-life for degradation in soil (DT50) according to a discovery project protocol designed by Syngenta. The assignment to category (DT50 < 100 or >100 days) was based on the geometric mean of multiple values (different soils, soil samples, and repeat measurements). Compounds where the median value and the geometric mean indicated different categories were removed. Here, DT50 < 100 days was treated as "class 1" (terminology used when discussing modeling and uncertainty methods).

Multiple iterative time splits were performed with the strategy for training vs testing configured to represent updating models with new data as new compounds are tested within discovery projects. Thus, for each time-point, all compounds tested up to the previous time-point were used as the training set, and all newly tested compounds available at that time point were used as the test set. In keeping with work by Sheridan and co-workers at Merck,[31] who report employing a similar model update and temporal validation strategy, results were not generated for time points where the number of new compounds in the test set was less than 50, or where there were fewer than 25 compounds per category, due to an assumed lack of robust results.

Industrial Regression Data Set

This comprises logP measurements for approximately 22,000 compounds, retrieved from the Syngenta corporate database. As per the Industrial Classification Data set, multiple iterative time splits for training (on previously assayed compounds),

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

and testing (on newly assayed compounds) were generated based upon regular time points, with time points where the number of compounds in the test set was less than 50 being skipped in keeping with literature precedence.[31]

Data Set Curation Workflow

For all public and industrial data sets, molecular structures were standardized and fingerprints computed (see the section "Fingerprints"). Any molecules that failed any of these steps were removed. Removal of duplicates for the public data sets was performed using InChIs prior to and after standardization. Checks for duplicates in the Industrial data sets were performed using unique corporate identifiers. Full details of the data-set curation workflows, which are similar but not identical to those employed by Morger et al.[26] and Walter et al.,[40] can be found in the Supporting Information.

Fingerprints

Morgan fingerprints of radius 2, which are similar to the ECFP4 fingerprints,[41] and hashed to a fixed-length bit-vector of length 1024, were calculated using RDKit[42] and used as the descriptors for model building. They were also used to compute distances between compounds, for the AD methods and analyses of training set diversity, using the Soergel distance (1−Tanimoto similarity),[43] otherwise known as the Tanimoto distance.[27]

Random Forest Models

Random Forest (RF) models were created for both classification and regression[35,44] using SciKit-Learn[45−47] and used as the basis for all Machine Learning predictions and uncertainty estimates. Given that RF can be expected to perform well with sensible defaults,[35] we did not tune the hyperparameters. The full hyperparameter settings are described in the Supporting Information.

Robustness Assessment

Since RF and the uncertainty estimation methods require random sampling, the robustness of these calculations was assessed by generating all results using five different random seeds. In addition, for the Wang-ChEMBL data sets (see "Public Regression Datasets"), results were also generated over five literature-defined cross-validation folds (IVIT or IVOT).

Meaning of Training Set for Calibrated Uncertainty Methods

In keeping with the literature,[17,22] the "training set" contains both the "proper training sets", used to build the underlying RF models, and the "calibration sets" sampled from the "training set" in order to calibrate the uncertainty estimates for CVAP and ACP. Since these uncertainty methods are based upon multiple proper training/calibration splits of the training set, and the predictions themselves are contingent upon the uncertainty estimates, as explained below, all available training data was used to refine the predictions in practice.

Uncertainty Method for Regression: Aggregated Conformal Prediction (ACP)

Aggregated Conformal Prediction (ACP) builds on top of Inductive Conformal Prediction (ICP),[22] which is therefore described first.

For ICP, the training set was divided into a proper training set and a calibration set, using stratified random splitting, in the ratio 75:25.[4,33] A model was trained on the proper training set using RF. Nonconformity scores, α_i, were calculated for each instance i of the calibration set as the scaled absolute difference between the experimental (y_i) and the predicted ($\hat{y}_i$) value, using eq 1.

$$\alpha_i = \frac{|y_i - \hat{y}_i|}{\lambda_i} \quad (1)$$

In eq 1, λ_i is a scaling factor that is derived from the standard deviation (σ_i) of the predictions of the underlying trees in the RF model, as per eq 2. We chose this approach to nonconformity score calculation based upon prior work by Svensson et al.[4]

$$\lambda_i = e^{\sigma_i} \quad (2)$$

Subsequently, for computing prediction intervals for new compounds, the calibration set nonconformity scores were ranked in descending order and the largest nonconformity score (α_ε) was selected for which the fraction of calibration set compounds with scores equal to or greater than this was approximately the same as the specified significance level (ε), using Algorithm 2 from Papadopoulos et al.[22] The prediction interval (PI) for a new compound (t), e.g., test set compound, at the specified significance level ($PI_{\varepsilon,t}$) was then computed using the selected calibration set nonconformity score (α_ε) and the scaling factor for the new compound (λ_t) using eq 3. In eq 3, $\hat{y}_t$ denotes the prediction of the underlying Random Forest model, and λ_t is computed, as per eq 2, using the standard deviation of the predictions of individual trees (σ_t).

$$PI_{\varepsilon,t} = \hat{y}_t \pm \alpha_\varepsilon \lambda_t \quad (3)$$

Subsequently, ACP applies ICP to the training set N times, i.e., N random partitions into proper training and calibration sets are generated, then combines the resulting prediction intervals to derive an overall prediction and prediction interval. As per Lindh et al.,[33] we computed the final upper and lower limits of the prediction intervals as the median of the upper and lower limits of the N ICP prediction intervals respectively, and computed the corresponding prediction as the midpoint between these final prediction interval limits. (This ensures, as for ICP, that the prediction intervals are symmetric about the prediction, but it does mean that the prediction is contingent upon the specified significance level.) To facilitate performing all calculations within the time limits of the computer cluster used for this work, we set N to 20, as per Morger et al.[26]

Choice of a Bespoke Significance Level for Regression Uncertainty Estimates

In contrast to various prior applications of conformal regression,[1,4] we chose to focus on prediction intervals computed at a significance level of 32%. We explain our rationale for this when describing the metric ENCE below. We only considered other significance levels when computing ECE (see below).

Uncertainty Method for Classification: Cross-Venn-ABERS Predictors (CVAP)

Since CVAP entails combining the outputs of IVAP applied to cross-validated proper training/calibration splits, we first explain how IVAP was applied to a single partition. IVAP[16,17] entails randomly splitting the training set into a proper training and calibration set, using the model trained on the proper training set to compute scores for all calibration compounds and the new compound of interest, and then using Isotonic Regression to obtain an upper and lower estimate for the probability that the new compound belongs to "class 1".

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

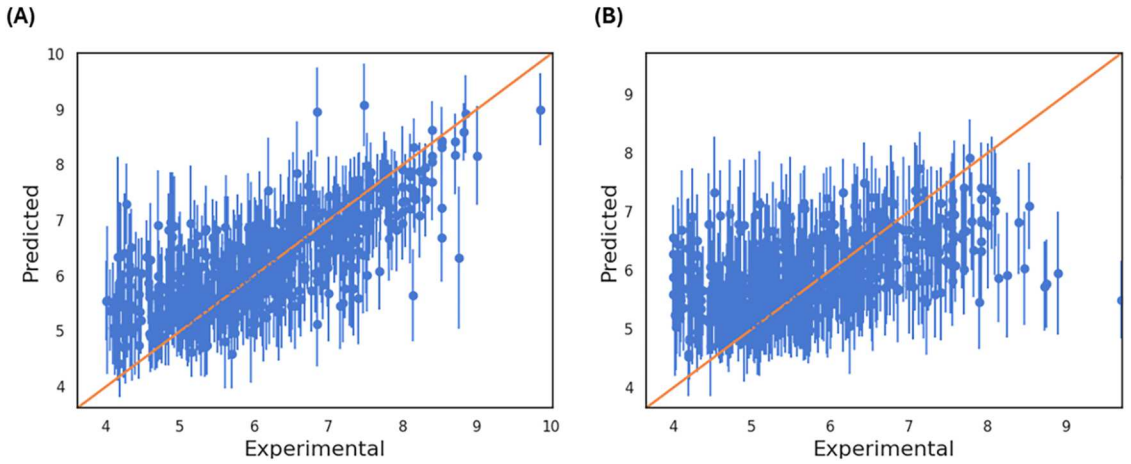


Figure 1. Illustration of regression predictions and uncertainty estimates (prediction intervals) from which prediction and uncertainty estimation performance metrics were computed, with the $y = x$ line shown in orange. Predictions and prediction intervals (ACP, significance level 32%) were generated for the Wang-ChEMBL COX-2 data set on one of the (A) IVIT (random) and (B) IVOT (nonrandom) folds. As explained under Methods, the models were rebuilt using the complement of the fold for which results are shown, which necessarily differs between the IVIT and IVOT scenarios. The corresponding performance metrics were as follows: (A) R^2 = 0.57, RMSE = 0.74, Pearson coefficient = 0.76, Spearman coefficient = 0.74, ENCE = 0.16, SCC = 0.37, efficiency = 1.35, coverage = 69%; (B) R^2 = 0.24, RMSE = 0.93, Pearson coefficient = 0.51, Spearman coefficient = 0.51, ENCE = 0.24, SCC = 0.30, efficiency = 1.49, coverage = 61%.

Subsequently, the upper (p_1) and lower (p_0) probability estimates can be combined into an overall probability estimate (p) of membership of "class 1" using eq 4.

$$p = \frac{p_1}{1 - p_0 + p_1} \tag{4}$$

Here, the scores were the uncalibrated probabilities of "class 1" membership from the RF model.

Subsequently, CVAP[16,17] entails partitioning the training set into K folds of roughly equal size and then using each fold in turn as the calibration set and the remaining training set compounds as the proper training set in order to obtain K sets of p_0 and p_1 as defined above for IVAP. These can be combined into an overall probability estimate (p_{CVAP}) for membership of "class 1", for the new compound of interest, using eq 5, where GM denotes the geometric mean. Here, we set K to 5 and used stratified random splitting.

$$p_{\mathrm{CVAP}} = \frac{\mathrm{GM}(p_1)}{\mathrm{GM}(1 - p_0) + \mathrm{GM}(p_1)} \tag{5}$$

Subsequently, these calibrated probabilities were used to predict whether the compound belonged to "class 1" or "class 0" based on whether the probability for "class 1" exceeded 0.5.

Prediction Performance Metrics

Standard metrics were used to assess the overall prediction performance of the different methods across the various test sets and to check the ability of the AD method to differentiate between, on average, more (inside domain) and less reliable (outside domain) predictions. In the case of regression, R^2 (coefficient of determination), computed based on the method reported by Alexander et al.,[48] RMSE (root-mean-square error), Pearson correlation, and Spearman rank correlation coefficients were calculated. (We prefer the RMSE over the Mean Absolute Deviation (MAD) for analysis, as it also conveys information about the size of model residuals while emphasizing outliers.) For the classification tasks, balanced accuracy, Mathews Correlation Coefficient (MCC), area under the curve (AUC), and Cohen's Kappa were calculated to assess the overall, rather than class-specific, prediction performance.

Uncertainty Estimation Performance Metrics for Regression

For conformal regression, the uncertainty estimates were evaluated using efficiency and validity,[4] as well as a variety of other metrics which have been proposed in the literature for assessing uncertainty estimates for regression models.[27] We assessed validity in terms of the coverage, i.e., the fraction of experimental endpoint values lying inside the prediction interval, which, for a valid conformal predictor, is expected (on average, for sufficiently large sample sizes) to be close to or greater than the complement of the significance level.[4,19,22,27] Efficiency was defined as the average size of the prediction intervals, for which it is desirable to be as small as possible while still maintaining validity.[4] As well as coverage and efficiency, we computed Expected Normalized Calibration Error (ENCE), and Spearman's Rank Correlation Coefficient (SCC) between the size of the prediction intervals and the prediction residuals, as well as Expected Calibration Error (ECE).[27]

We computed ECE[27] as per eq 6, where *coverage* (i%) denotes the observed coverage (%) at the specified significance level of i%, while N denotes the number of significance levels considered. We considered the following significance levels (%): 0, 10, 20, 32, 40, 50, 60, 70, 80, and 90.

$$\mathrm{ECE} = \frac{\sum_i |((100 - \mathrm{coverage}\ (i\%)) - i\%)/100|}{N} \tag{6}$$

We computed ENCE[27] as per eq 7, following binning of the test set compounds according to the prediction interval sizes. In eq 7, N is the number of bins, computed as the number of compounds divided by the minimum number of compounds per bin, rounded down to the nearest integer or set to one if the total number of compounds was less than this minimum. Here, we set the minimum number of compounds per bin to 20 in keeping with literature precedence.[27]

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

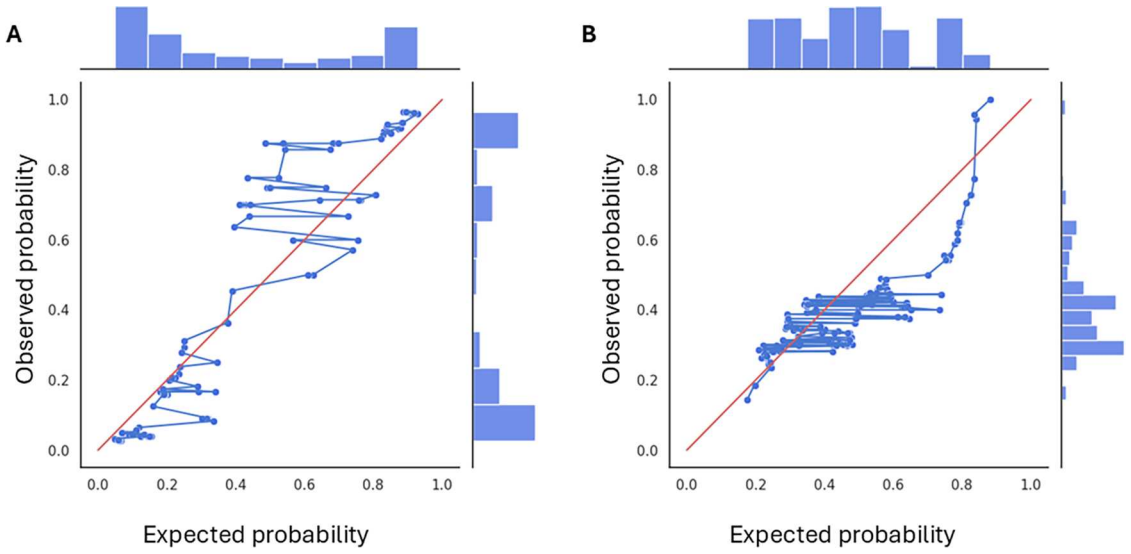


Figure 2. Illustration of classification uncertainty estimates using delta-calibration plots for delta = 0.05 for the CHEMBL206 target from the Morger-ChEMBL data set. (A) CVAP, random test set. (B) CVAP, Holdout test set (temporal validation). The *x*-axis shows the model-estimated probability of class 1 ("Expected Probability"), while the *y*-axis shows the probability of class 1 estimated from the experimental class labels within ± delta of the model probability ("Observed probability"). The density of the points on each axis is shown via histograms. The corresponding performance metrics are as follows: (A) R^2(cal) = 0.90, Pearson coefficient (cal) = 0.95, Spearman coefficient (cal) = 0.96, RMSE (cal) = 0.12; and (B) R^2(cal) = 0.17, Pearson coefficient (cal) = 0.86, Spearman coefficient (cal) = 0.89, RMSE (cal) = 0.12.

Furthermore, MSE(i) denotes the mean-squared error of predictions and mVAR(i) denotes the arithmetic mean of VAR(i,t), across all molecules (t) in the i-th bin In turn, VAR(i,t) denotes the uncertainty method's estimated variance in the distribution of possible values for the predicted property for the t-th molecule. In our context, where we adopt ENCE for the evaluation of conformal regression, it is the squared value of half the prediction interval size. The assumptions underpinning this calculation of VAR(i,t) from the prediction interval sizes are as follows: (i) the predictions are unbiased, meaning the distribution of true values is centered on the prediction, like the prediction intervals; (ii) the prediction intervals correspond closely to the confidence intervals for the same nominal coverage level, i.e., a 32% significance level prediction interval corresponds closely to a 68% confidence interval; (iii) the distribution of possible true values is normally distributed about the prediction, such that a confidence interval of 68% corresponds to twice the standard deviation of this distribution.

If these assumptions are violated, we would expect ENCE to increase. Even if the assumptions are not violated, ENCE will be larger if the uncertainty estimates are poorly calibrated.

$$\mathrm{ENCE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left|\sqrt{\mathrm{mVAR}(i)} - \sqrt{\mathrm{MSE}(i)}\right|}{\sqrt{\mathrm{mVAR}(i)}}\right) \quad (7)$$

Illustration of Regression Predictions, Uncertainty Estimates, and Performance Metrics

Figure 1 illustrates how the performance metrics vary between two sets of predictions, prediction intervals, and experimental endpoint values.

Uncertainty Estimation Performance Metrics for Classification

For the classification tasks, the uncertainty estimates were evaluated using the Stratified Brier Score (SBS)[49] and metrics that can be calculated from the calibration plot.[18] The Brier Score for two classes measures the mean-squared difference between the model-estimated probability of "class 1" and the experimental class labels (1 or 0). The SBS[49] is more appropriate for imbalanced data since it treats each class separately and gives each equal importance by averaging the Brier score for each class. The smaller the SBS is, the closer the estimated probabilities are to the ideal case where all true members of class 1 and 0 would have a model-estimated probability for class 1 of 100 and zero percent, respectively.

However, the SBS still fails to give credit to high uncertainty estimates, i.e., probabilities close to 50% for two classes, in cases where the model predictions are likely to be incorrect. Contrastingly, metrics based on the calibration plot, comparing "observed" probabilities and model-estimated probabilities for "class 1", can give credit to estimated probabilities close to 50% if this is true for the corresponding observed probabilities. The observed probabilities for the calibration plots are obtained by binning the model-estimated probabilities for "class 1" and computing the observed probabilities as the fraction of compounds for which "class 1" is the experimental label within each bin. We used the approach presented in Arvidsson et al.[18] and refer to the calibration plots as "delta-calibration plots". Overlapping bins are defined centered on each test set compound and extend to all test set compounds where their model-estimated probabilities for "class 1" lie within some small value, which we call "delta", of the value associated with the current test set compound. After computing the delta-calibration plots, we calculated a range of performance metrics capturing the relationship between the observed and model-estimated probabilities for "class 1". These were: coefficient of determination (R^2 (cal)), root-mean-square error (*RMSE* (cal)), Pearson coefficient (cal), Spearman coefficient (cal). We use the suffix "(cal)" to emphasize that these were computed from the delta-calibration plot, to avoid confusion with the regression prediction performance metrics.

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

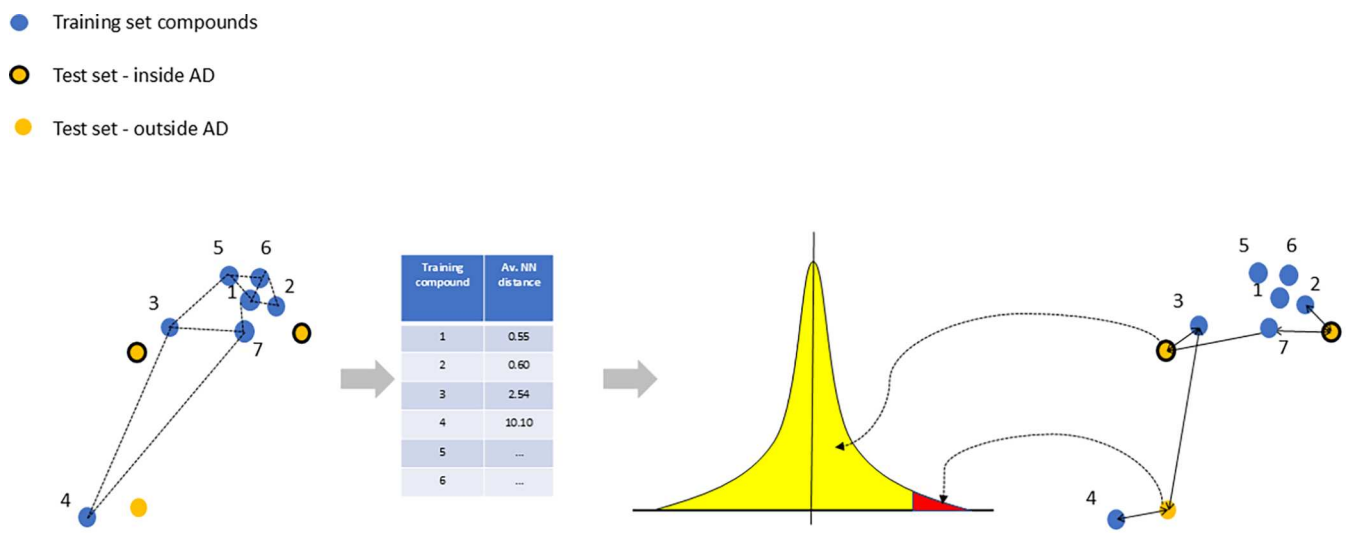


Figure 3. An illustration of the AD method nUNC. For simplicity, we have considered the case of when $k = 2$, i.e., when two nearest-neighbors are used to compute the applicability domain status. The average nearest-neighbor (Av. NN) distances are computed for training and test set compounds, and a statistical test is used to determine whether this value, for any given test set compound, comes from the training set distribution.

We chose delta to be 0.05 based on literature precedence.[18] We briefly considered the sensitivity of the results for a different value of delta and concluded 0.05 was reasonable (details in Supporting Information). In Figure 2, we present example delta-calibration plots for a random and nonrandom test set and the corresponding performance metrics.

Cases Where Metrics Were Not Computed

In some cases, metrics were undefined, e.g., due to zero test set compounds inside or outside the AD method's domain, or due to an endpoint's data distribution resulting in division by zero. We discuss other scenarios that result in metric values being undefined in the Supporting Information.

AD Method

The nUNC approach (Figure 3) was adapted from a k-nearest neighbors (k-NN) approach previously employed by researchers at the University of North Carolina at Chapel Hill in the United States of America (UNC)[38] and, save for our use of Tanimoto distances, is equivalent to the "AD-DD" approach of Fan et al.[12] In brief, this approach assesses whether a new compound comes from the same distribution as the training set compounds by comparing the average distance of that new compound to its nearest neighbors in the training set with the distribution of values for the average distance of training set compounds to their nearest neighbors. Here, the training set comprises both the proper training sets and calibration sets, as described above under "Meaning of Training Set for Calibrated Uncertainty Methods".

In more detail, the AD method works as follows. Initially, for each training set compound, the distance to its k-nearest neighbors is computed and the average distance ($\overline{d}^j$) determined. The distribution of average distances is summarized by the mean ($\frac{1}{J}\sum_j^J \overline{d}^j$) and standard deviation ($sd(\overline{d}^j)$) calculated across all J training set compounds.

For the new compound, e.g., test set compound, the average distance ($\overline{d}$) to its k-nearest neighbors is computed, along with a z-score based upon the training set distribution statistics (eq 8). The corresponding one-tail p-value is computed assuming a normal distribution, where this p-value is the probability of a new compound having an average distance greater than or equal to the observed value ($\overline{d}$) to its nearest neighbors, given the null hypothesis that the average distance to its nearest neighbors comes from the same distribution as the training set. If the one-tail p-value is less than or equal to 5%, the compound is deemed to lie outside the domain.

$$z\text{-score} = \frac{\overline{d} - \left(\frac{1}{J}\sum_j^J \overline{d}^j\right)}{sd(\overline{d}^j)} \quad (8)$$

Here, we used $k = 3$, after briefly exploring different values using the exemplar public data sets and determining that this seemed generally sufficient to ensure that a majority of random test set compounds were deemed to lie inside the domain, in keeping with our intuition. This value of k was also judged likely to leave sufficient numbers of compounds inside and outside the domain for most random and nonrandom test sets, to enable robust performance metrics for both the inside and outside subsets to be computed. This judgment was based on inspecting the number of compounds in these subsets for the exemplar data sets. We did not optimize the value of k based on consideration of these performance metrics.

Finally, we note that, in the work of Fan et al.,[12] the z-score computed as per eq 8, was itself considered a measure of uncertainty. The important difference between this measure of relative uncertainty and the uncertainty methods that we focus on in our work (conformal regression and Venn-ABERS) is that these methods are designed to produce calibrated uncertainties. In other words, these calibrated uncertainties should closely reflect the probability of the predicted class being correct (Venn-ABERS) or the distribution of prediction residuals (conformal regression) if they are reliable. Our focus here was on exploring the extent to which the nUNC AD method could differentiate between compounds where these uncertainty methods were more likely (inside the domain) vs less likely (outside the domain) to be reliable.

Assessing the Impact of the AD Status on Prediction and Uncertainty Estimation Performance Using Shift Metrics

The AD method was used to divide each test set into two subsets: one subset comprised compounds deemed to lie inside and the other comprised compounds deemed to lie outside the

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

domain. The AD method was then evaluated by calculating "shift metrics" that represent the change in the prediction and uncertainty performance metrics when moving from the subset of compounds inside the AD to the subset outside the AD. The shift metrics were calculated as follows: [metric value inside the AD] − [metric value outside the AD]. For metrics where larger (more positive) values are better, e.g., MCC and R^2 (cal), a positive shift metric represents a drop in performance when moving from inside the AD to outside the AD and indicates that the AD method is performing as expected. Conversely, for metrics where smaller values are better, e.g., RMSE (cal) and SBS, a negative shift metric indicates that the AD method is performing as expected. The complete set of shift metrics and their expected signs, when the AD method was working as expected, can be seen in Table 1.

Table 1. All Performance Metrics, Grouped By Their Type, and the Expected Sign of the Shift Metric (Value[Inside Domain] − Value[Outside Domain]) if the AD Method Was Working as Expected to Assign More Reliable Predictions (Or Uncertainty Estimates) to the Inside Domain Subset

type of performance metric	expected sign of shift metric[a]	metric
Regression Prediction	Positive	R^2; Pearson coefficient; Spearman coefficient
	Negative	RMSE
Regression Uncertainty	Positive	Coverage; SCC
	Negative	ENCE; ECE; efficiency
Classification Prediction	Positive	Balanced Accuracy; MCC; AUC; Kappa
	Negative	n/a
Classification Uncertainty	Positive	R^2 (cal); Pearson coefficient (cal); Spearman coefficient (cal)
	Negative	RMSE (cal); SBS

[a]Since the expectation, for a reliable AD method, is that better values would be obtained inside and worse values would be obtained outside the domain, a shift metric with a positive expected sign means that larger (more positive) values of the corresponding metric indicate better performance. Conversely, a shift metric with a negative expected sign means that smaller values of the corresponding metric indicate better performance.

In general, the AD method was evaluated for each data set based on the average (arithmetic mean) shift-metric value, for a given test set type. Here, test set type refers to, for example, Update1 or Holdout for the Morger-ChEMBL targets, or IVOT for the Wang-ChEMBL targets. The shift-metric values that were averaged were computed for different random seeds and, for the Wang-ChEMBL targets, folds. The analyses of the average shift-metric values were complemented by non-parametric statistical significance tests.

Statistical Significance Testing for AD Method Shift Metrics

For each set of averaged shift-metric values, we computed a corresponding one-tail *p*-value, indicating the probability of obtaining results at least as indicative of the AD method working from random splitting of the test set.

Similarly, for each set of averaged shift metrics, we computed corresponding two-tail *p*-values, based on comparing the absolute magnitudes of the observed shift-metric values, with the AD method split of the test set, to those arising from random splitting of the test set. These were computed to allow us to confirm that the minority of cases where the average shift metric had the unexpected sign were chance findings.

Both the one-tail and two-tail *p*-values corresponding to the average shift-metric values were, separately,[50] adjusted for multiple testing,[51] taking into account the results generated using all methods, using a conservative approach.[52]

In order to compute the *p*-values prior to multiple testing adjustment, *precursor* one-tail and two-tail *p*-values were initially computed for each of the underlying shift metrics corresponding to a given average (arithmetic mean) shift metric. This yielded a set of *precursor* one-tail (and two-tail) *p*-values, one for each of the shift-metric values used to compute the corresponding average shift metric. This set of *precursor* *p*-values was combined into a single *p*-value, i.e., each average shift metric was associated with a single one-tail (and two-tail) *p*-value, using a conservative approach: twice the average (arithmetic mean) *p*-value.[53] These combined *p*-values were considered for subsequent analysis and multiple-testing adjustments.

To compute the *precursor* *p*-values, corresponding to a single shift metric prior to averaging over random seeds and, where relevant, folds, the inside and outside AD labels for a given test set were randomly partitioned between test set compounds up to 1000 times and the corresponding shift metrics were computed. (In some cases, slightly fewer partitions yielded splits where the shift metric could be computed, as per the discussion under "Cases where Metrics Were Not Computed". However, at least 941 partitions were used.) Random partitions were generated such that the number of compounds with inside and outside AD labels and the distribution of experimental values were consistent with the split produced by the AD method.

From these random partitions, the *precursor* one-tail *p*-value was computed as the fraction of times that the randomly generated shift-metric values were greater than or equal to the shift metric for the original AD method split, where the shift metric was expected to be positive (see Table 1), or the faction of times that these values were less than or equal to the original value, where the shift metric was expected to be negative. The two-tail *precursor* *p*-value was computed as the fraction of times that the magnitude of the randomly generated shift metric was greater than or equal to the magnitude of the original value. If a shift metric could not be computed for the original split, no *p*-value was computed. The computation of these *precursor* *p*-values is illustrated in Figure 4.

In some cases, the one-tail *p*-values appeared statistically significant even when the average shift-metric did not have the sign expected if the AD method was working (Table 1). These cases were not considered evidence that the AD method was working. (Only cases where the average shiftmetric had the expected sign and the corresponding one-tail *p*-value was statistically significant were considered to be robust evidence that the AD method was working.)[54] These cases arose when, for at least some of the random seed and fold combinations, most or all (>94.5%, as per the significance level described below) of the randomly generated shift-metric values had the unexpected sign and were further from zero than the shift metric corresponding to the AD method splitting of the test set, *but* this also had the unexpected sign, as per the example in Figure 4(C). This could arise because differences in the distributions of experimental values between the subsets (AD method or randomly assigned "inside" vs "outside" the

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

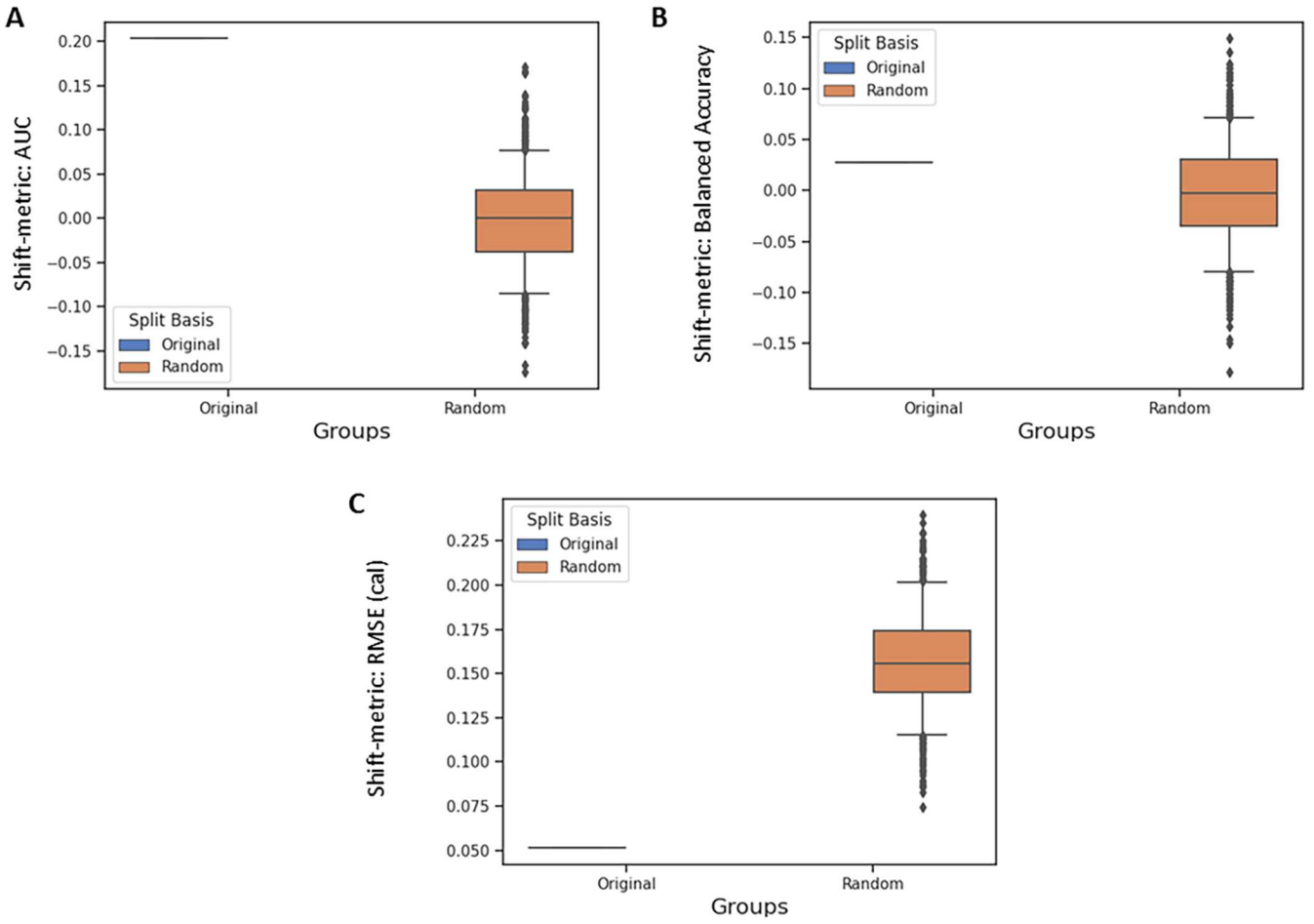


Figure 4. Example box plots illustrating the computation of the precursor one-tail *p*-values, which were computed for each random seed and for the Wang-ChEMBL data sets per fold, prior to combining them into a single one-tail *p*-value for each metric−train/test-split pair for which average shift metrics were computed. Here, all calculations correspond to CVAP and nUNC applied to the following Morger-ChEMBL target−test-set combinations: (A, B) CHEMBL220, Update1; (C) CHEMBL203, Holdout. In cases (A, B), the expected sign for the shift metric, if the AD method was working as expected, was positive; for (C), the expected sign was negative. Hence, the corresponding one-tail *p*-values were as follows: (A) 0.000; (B) 0.278; and (C) 0.000.

Table 2. Public Morger-ChEMBL Classification Datasets: Numbers of Compounds, after Dataset Curation and Fingerprint Calculations, in the Active (A) and Inactive (I) Categories and Identities of the Exemplar Datasets Used for Initial Method Exploration

	literature training set (random 80:20 train:test split)			update1[a]			holdout[a]		
	total	A	I	total	A	I	total	A	I
CHEMBL220	1626	833	793	443	247	196	216	112	104
CHEMBL4078[b]	1996	988	1008	530	275	255	769	270	499
CHEMBL5763	1690	589	1101	376	75	301	248	113	135
CHEMBL203	2066	430	1636	739	213	526	508	167	341
CHEMBL206[b]	758	323	435	180	63	117	249	102	147
CHEMBL279	2597	649	1948	819	307	512	985	299	686
CHEMBL230	1012	540	472	294	78	216	381	168	213
CHEMBL340	1763	495	1268	584	150	434	554	106	448
CHEMBL240[b]	2714	1926	788	714	413	301	735	497	238
CHEMBL2039	1355	645	710	381	192	189	206	72	134
CHEMBL222	901	670	231	286	226	60	127	53	74
CHEMBL228[b]	1097	857	240	468	372	96	274	195	79

[a]Nonrandom (time-split) test sets. [b]Exemplar data sets used for initial exploration of AD parameters (focused on ensuring a majority lay inside the domain for the random test sets), comparison of different uncertainty methods (random 20% test sets) and AD methods (shift metrics for nonrandom test sets), prior to focusing on the default AD (nUNC) and uncertainty method (CVAP) for which we report results.

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

Table 3. Public Morger-Tox21 Classification Datasets: Numbers of Compounds, after Dataset Curation and Fingerprint Calculations, in the Active (A) and Inactive (I) Categories and Identities of the Exemplar Datasets Used for Initial Method Exploration

	Tox21Train (random 80:20 train:test split)			Tox21Test[a]			Tox21Score[a]		
	total	A	I	total	A	I	total	A	I
NR-AhR	6031	690	5341	259	29	230	553	70	483
NR-AR[b]	6690	244	6446	279	3	276	541	11	530
NR-AR-LBD	6271	203	6068	240	4	236	529	8	521
NR-Aromatase[b]	5361	242	5119	204	18	186	484	34	450
NR-ER	5639	606	5033	252	27	225	477	47	430
NR-ER-LBD	6421	261	6160	274	10	264	546	19	527
NR-PPAR-γ	5985	148	5837	254	15	239	552	30	522
SR-ARE[b]	5409	818	4591	223	47	176	503	88	415
SR-ATAD5	6560	224	6336	259	25	234	565	33	532
SR-HSE[b]	6011	282	5729	254	10	244	554	17	537
SR-MMP	5365	819	4546	228	38	190	499	53	446
SR-p53	6283	368	5915	256	28	228	557	37	520

[a]Nonrandom (time-split) test sets. [b]Exemplar data sets used for initial exploration (as described for Table 2).

Table 4. Public Wang-ChEMBL Regression Datasets: Numbers of Compounds along with the Activity Value Distributions, after Dataset Curation and Fingerprint Calculations, and Identities of the Exemplar Datasets Used for Initial Method Exploration

		activity			
	no. compounds[a]	min	max	median	mean (standard deviation)
A2a	199	4.21	9.67	6.33	6.57 (1.25)
ABL1	755	4	9.455	6.3	6.40 (1.26)
Acetylcholinesterase	2964	4	10.7	6.07	6.20 (1.32)
Cannabinoid	1087	4.05	9.7	7.3	7.17 (1.26)
Carbonic	591	4	9.3	7.535	7.16 (1.28)
Caspase	1584	4.01	10.96	5.32	5.85 (1.45)
Coagulation	1588	3.48	10.38	5.99	6.13 (1.25)
COX-1[b]	1306	4	9	5.14	5.27 (0.89)
COX-2[b]	2768	4	10.7	6.03	6.15 (1.18)
Dihydrofolate	573	4	9.41	6.26	6.32 (1.14)
Dopamine[b]	469	4.52	9.54	6.6	6.69 (1.07)
Ephrin	1510	4.05	10.43	6.75	6.76 (1.08)
Estrogen	1618	4	9.7	6.445	6.47 (1.35)
Glucocorticoid	1387	4.05	10.4	7.43	7.29 (1.00)
Glycogen	1724	3.38	10.89	6.52	6.55 (1.16)
HERG	5012	4	9.85	5.3	5.47 (0.89)
JAK2	2388	3.84	10.97	7.29	7.25 (1.19)
LCK	1337	4	11	6.85	6.86 (1.29)
Monoamine	1308	4	10	5.26	5.47 (1.04)
opioid	779	4.18	10.9	6.39	6.61 (1.28)
Vanilloid	1760	4.04	9.8	6.92	6.89 (0.98)

[a]These data sets were split into random and nonrandom (cluster-split) train/test partitions based on the five literature-defined IVIT and IVOT folds, respectively. [b]Exemplar data sets used for initial exploration of AD parameters (focused on ensuring a majority lay inside the domain for the random test sets), comparison of different uncertainty methods (random IVIT folds) and AD methods (shift metrics for nonrandom IVOT folds), prior to focusing on the default AD (nUNC) and uncertainty method (ACP) for which we report results.

domain) were likely to produce shift metrics lying in a particular direction, even when the identities of compounds assigned to those subsets were randomly assigned. However, these cases still do not provide direct evidence that the AD method was working, i.e., assigning molecules with more reliable predictions (or uncertainty estimates) to the inside domain subset, such that the shift metric was observed to have the expected sign.

It should be noted that the standard 5% significance level is an arbitrary convention and larger p-values do not necessarily mean the observed differences arose due to chance.[55] Hence, we considered p-values less than or equal to 5%, when rounded to the nearest percent, to be statistically significant. This was a convenient heuristic for highlighting the results that are least likely to have arisen due to chance. As noted above, only cases where the average shift-metric had the expected sign and the corresponding one-tail p-value was statistically significant were considered to be robust evidence that the AD method was working.Average shift metrics with the expected sign but which were not associated with a statistically significant one-tail p-value were also taken as indicative of the AD method working, but were considered less robust evidence.

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

RESULTS

Model-Ready Data Set Details

The number of compounds and other statistics, following the data-set curation and fingerprint calculations described under METHODS, are presented in Tables 2−6.

Table 5. Syngenta DT50 Classification Dataset: Numbers of Compounds, after Dataset Curation and Fingerprint Calculations, in the Training (Previously Studied Compounds) and Nonrandom Test Sets (Newly Studied Compounds) at Different Time Points

	training			test		
time point	total	<100 days	>100 days	total	<100 days	>100 days
T2	4366	3031	1335	110	66	44
T4	4471	3089	1382	215	163	52
T5	4685	3251	1434	206	148	58
T6	4891	3399	1492	147	98	49

Table 6. Syngenta logP Regression Dataset: Numbers of Compounds, after Dataset Curation and Fingerprint Calculations, in the Training (Previously Studied Compounds) and Nonrandom Test Sets (Newly Studied Compounds) at Different Time Points

time point	training	test
T1	21,149	234
T2	21,383	341
T3	21,724	266
T4	21,990	231

Data Set Distributions

For each data set, we compared the distributions of different data-set subsets using t-SNE.[56,57] The t-SNE plots for all public and Syngenta data sets can be seen in the Supporting Information. These illustrate the differences in shifts in chemical space between random and nonrandom train/test splits, but the degree of data shift observed for the latter, in keeping with reports of real-world time splits,[29] was variable.

Application of the Selected AD and Uncertainty Methods to All Data Sets

We investigated how well the nUNC AD method could differentiate between more (inside the domain) and less (outside the domain) reliable predictions and uncertainty estimates with CVAP (classification) and ACP (regression) across a large range of public and two Syngenta data sets. (For some of the public data sets, e.g., A2a from the Wang-ChEMBL data-set group, it was not possible to compute one or more out-of-domain metrics for some test sets, see "Cases where Metrics Were Not Computed".) We focus on the results for nonrandom test sets that are more relevant to real-world applications than randomly sampled test sets, i.e., the temporal validation test sets for the Morger-ChEMBL (Update1, Holdout), Morger-Tox21 (Tox21Score, Tox21Test), and Syngenta data sets and the IVOT splits for the Wang-ChEMBL data sets. Indeed, for the Syngenta data sets, these represent the gold-standard[29,31,37] of applying models built on previously generated data within discovery projects to compounds subsequently tested within those projects.

We selected the Balanced Accuracy and RMSE metrics to represent trends in prediction performance and Pearson coefficient (cal) and coverage to represent trends in uncertainty estimation performance for classification and regression tasks, respectively. The results for these metrics and corresponding shift metrics are presented in Figures 5, 6, 7, 8, 9. We subsequently discuss the trends by considering all performance metrics for which shift metrics were computed.

Trends Observed for the Nonrandom Test Sets Across All Metrics. We discuss the trends for all metrics and shift metrics, for the nonrandom test sets, in the following subsections. (Graphs showing the values for these metrics and the corresponding shift metrics, including the illustrative examples presented in Figures 5−9, can be seen in the Supporting Information.) We considered the trends in the average shift-metric values, including whether these had the sign expected if the AD method was working to flag less reliable predictions (or uncertainty estimates) as outside the domain, and whether the corresponding one-tail p-values were statistically significant. In addition, for those metrics with an a priori baseline expected for randomly generated predictions or uncertainty estimates, we also considered whether their raw values inside (or outside) the domain were better than expected due to chance for all relevant classification and regression prediction and uncertainty performance metrics: Balanced Accuracy, AUC > 0.5; Kappa, MCC, R^2 (cal), Pearson coefficient (cal), Spearman coefficient (cal), R^2, Pearson coefficient, Spearman coefficient, SCC > 0.

When discussing the trends in average shift-metric performance, we consider the percentage of targets (public data sets) and time-splits (industrial data sets) for which a particular observation was true, for a given combination of metric and nonrandom test set type. We summarize these trends across all combinations being considered, for a given scenario, in terms of the range of these percentages. For example, considering the scenario of uncertainty estimation performance for nonrandom test sets of the Morger-ChEMBL data-set group, 3/12 targets (25%) had average RMSE (cal) shift-metric values with the expected sign for the Holdout test set, the joint smallest percentage. For the same scenario, 11/12 targets (92%) had average SBS shift-metric values with the expected sign for the Update1 test set, the largest percentage. Hence, in this case, we report that 25−92% of targets had uncertainty performance average shift metrics with the expected sign.

Similarly, when discussing whether the relevant raw metric values inside (or outside) the domain tended to be better than expected due to chance, we also considered the percentages for each relevant scenario. For example, for the nonrandom test sets of the Morger-ChEMBL and Morger-Tox21 classification data sets (Holdout, Update1, Tox21Test, Tox21Score), there were 240 raw Balanced Accuracy values computed for inside domain compounds. (These came from 12 Morger-Tox21 targets, 12 Morger-ChEMBL targets, two nonrandom test sets per target-specific data set, and five random seeds.) Of these, 182 were larger than 0.50. Hence, we report that 76% of inside domain Balanced Accuracy values were better than expected due to chance for this scenario.

Public Classification Data Sets: Prediction Performance for Nonrandom Test Sets (Time-Splits). The average prediction performance shift metrics typically had the expected sign, for both the Morger-ChEMBL and Morger-Tox21 data sets. For the Morger-ChEMBL data sets, depending upon the combination of metric and nonrandom test set, 67−83% of targets had average shift metrics with the expected sign. Moreover, for 17−50% of targets, the average prediction

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

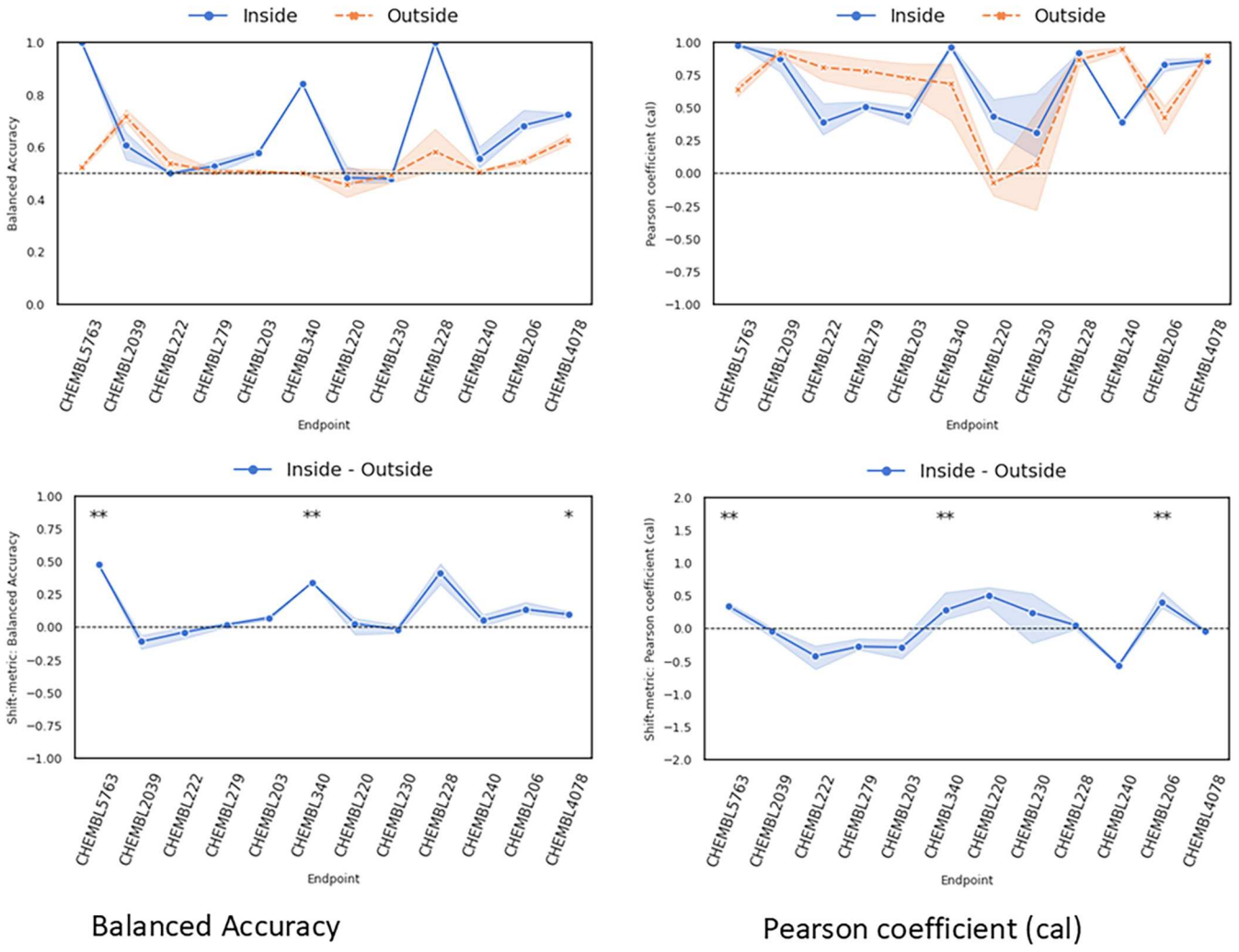


Figure 5. Variation across endpoints of balanced accuracy and Pearson coefficient (cal) values inside vs outside the domain (top panel) and the corresponding shift-metric values (bottom panel), for the Morger-ChEMBL targets and Holdout test set type. The colored dotted lines with markers indicate the arithmetic mean, while the shading indicates the spread of the values (95% percentile interval), across different random seeds. The black dotted lines on the metric figures (top panel) indicate the baseline expected for a random model, while the dotted lines on the shift-metric figures (bottom panel) denote a value of zero, indicating whether the shift metric has the expected sign or not. The annotations are based on the statistical significance testing for AD method shift-metrics: * = average shift-metric had the expected sign and the unadjusted one-tail *p*-value was significant; ** = average shift metric had the expected sign and the multiple testing adjusted one-tail *p*-value was significant; X = average shift metric did not have the expected sign, but the unadjusted two-tail *p*-value was significant; XX = average shift metric did not have the expected sign, but the adjusted two-tail *p*-value was significant.

performance shift metrics had the expected sign and were associated with statistically significant adjusted one-tail *p*-values. An additional 8−25% of targets had average performance shift metrics with the expected sign and one-tail *p*-values, which were statistically significant prior to multiple testing adjustment.

However, while 50−100% (excluding AUC, 25%) of average shift metrics had the expected sign, no statistically significant differences were observed for the Morger-Tox21 data sets. The latter observation probably reflects the smaller numbers of compounds deemed to lie outside the domain, with 4−20 compounds being observed to lie outside the domain for the nonrandom test sets across the Morger-Tox21 data sets, compared to 121−791 for the Morger-ChEMBL data sets.

The relevant inside-domain metric values were typically better than expected by chance: Balanced Accuracy (76%), MCC (76%), AUC (98%), Kappa (76%). This was less frequently observed for the out-of-domain metric values, which could not be computed in some cases due to few compounds in at least one class: Balanced Accuracy (46%), MCC (46%), AUC (82%), Kappa (46%).

Public Classification Data Sets: Uncertainty Estimation Performance for Nonrandom Test Sets (Time-Splits). Again, the average shift metrics typically had the expected sign. However, this was, overall, slightly less frequently the case than for the prediction performance shift metrics.

For the Morger-ChEMBL data sets, depending upon the metric−test set combination, 25−92% of targets had average uncertainty estimation performance shift metrics with the expected sign, with six out of ten combinations (five uncertainty performance metrics paired with two test set types) having a majority with the expected sign. However, only 0−50% of targets had average shift metrics with the expected sign and statistically significant adjusted one-tail *p*-values. Hence, there was less evidence that the AD method was

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

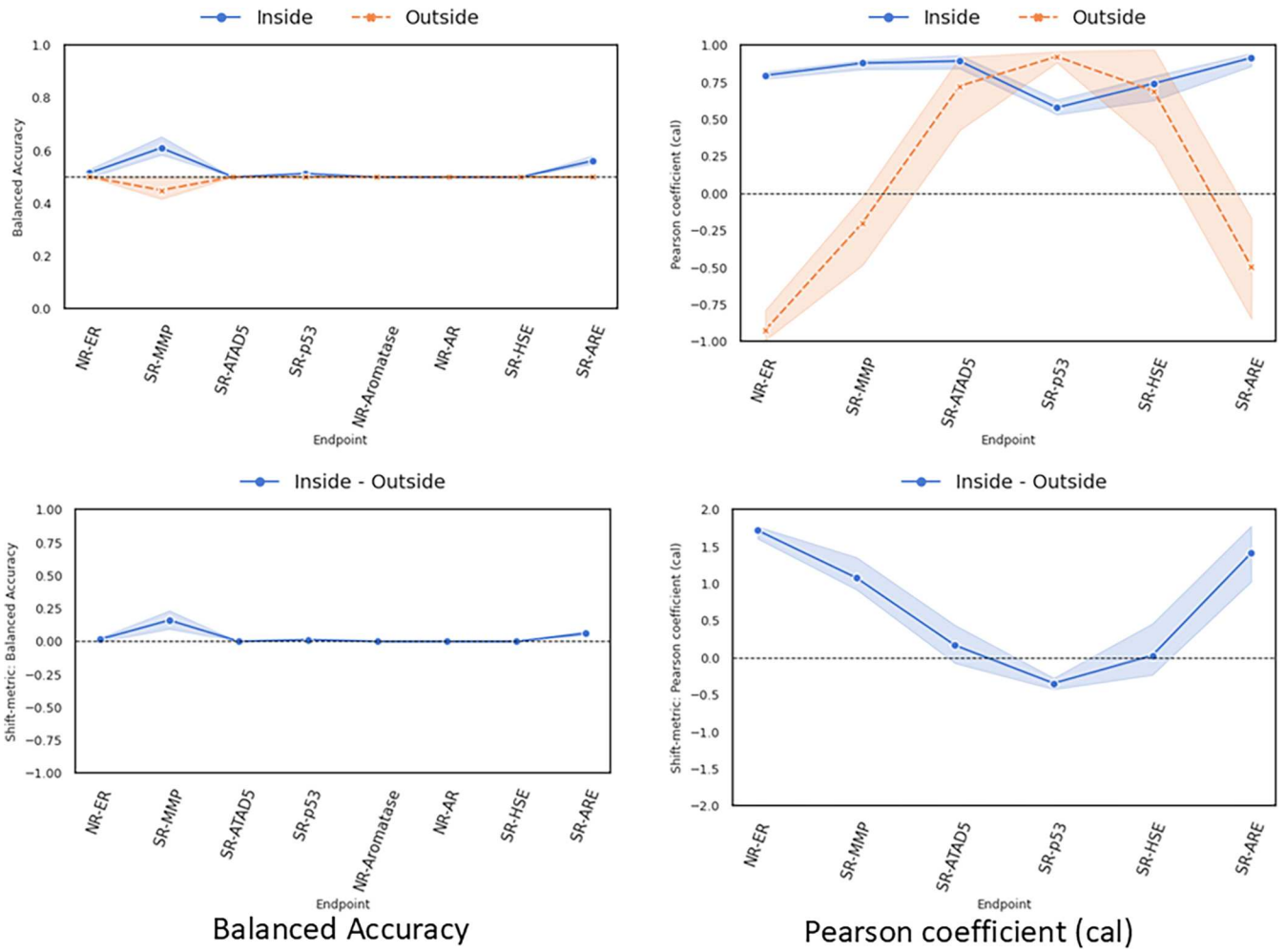


Figure 6. Variation across endpoints of balanced accuracy and Pearson coefficient (cal) values inside vs outside the domain (top panel) and the corresponding shift-metric values (bottom panel), for the Morger-Tox21 targets and Tox21Test test set type. See Figure 5 caption for description of dotted lines, shading, and annotations.

working as expected than for the prediction performance shift metrics. Moreover, for the Morger-Tox21 data sets, none of the *p*-values corresponding to the uncertainty estimation performance shift metrics were statistically significant, albeit 67−100% of the average shift-metric values had the expected sign. Again, this probably reflects the smaller numbers of compounds deemed to lie out-of-domain for the nonrandom test sets for the Morger-Tox21 vs Morger-ChEMBL data sets.

As per the prediction performance metrics, the relevant raw inside-domain metric values were typically better than expected due to chance: R^2 (cal) (74%), Pearson coefficient (cal) (98%), Spearman coefficient (cal) (90%). This was less frequently observed for the corresponding out-of-domain values: R^2 (cal) (39%), Pearson coefficient (cal) (80%), Spearman coefficient (cal) (78%).

Public Regression Data Sets: Prediction Performance for Nonrandom Test Sets (IVOT Folds). The average shift metrics almost always showed the expected sign (85−100% of targets across all prediction performance metrics). However, perhaps in some casesdue to insufficient numbers of test set compounds outside the domain, only one of these (RMSE for Glycogen, see Figure 7) was associated with statistically significant adjusted one-tail *p*-values. Prior to adjusting for multiple comparisons, this was also true for a few other average shift-metric values.

Almost always, the relevant raw in-domain metrics were better than expected by chance: R^2 (92%), Pearson coefficient (100%), Spearman coefficient (100%). This was often, but less frequently, the case for the out-of-domain metrics: R^2 (64%), Pearson coefficient (92%), Spearman coefficient (93%).

Public Regression Data Sets: Uncertainty Estimation Performance for Nonrandom Test Sets (IVOT Folds). The average shift metrics typically showed the expected sign (60−95% of targets across all uncertainty estimation performance metrics), albeit only efficiency (20%) also showed adjusted one-tail *p*-values that were statistically significant. Prior to multiple testing adjustment, some additional average shift metrics for efficiency, coverage, and ECE had the expected sign and were associated with statistically significant one-tail *p*-values.

For the one relevant metric with an a priori random baseline, SCC, the raw in-domain values were almost always (93%) better than expected due to chance. This was less frequently (75%) the case for the corresponding out-of-domain metric values. Looking at the average in-domain coverage values (Figure 7), none of them quite achieved the expected minimum (0.68) for compounds selected from the same distribution as the training/calibration data for the selected significance level (0.32). However, several came very close, including in cases where the out-of-domain values were

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

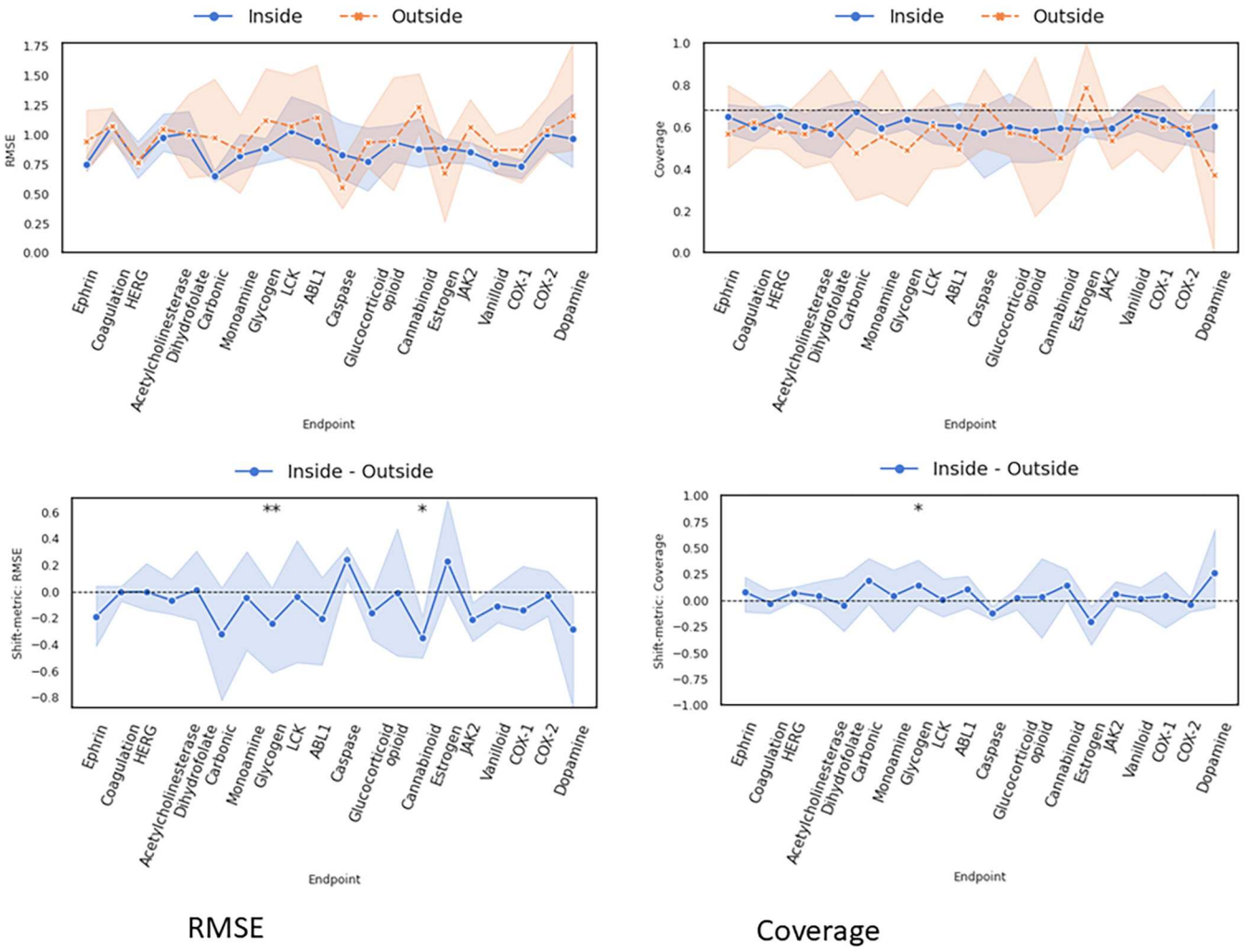


Figure 7. Variation across endpoints of RMSE and coverage values inside vs outside the domain (top panel) and the corresponding shift-metric values (bottom panel), for the Wang-ChEMBL targets and IVOT test set type. The dotted lines, shading, and annotations are as per Figure 5, save for the shading also illustrating the spread of values across different IVOT folds, with the exception of the coverage plot (top RHS), where the black dotted line indicates the minimum expected coverage, on average, for a valid conformal predictor (0.68, i.e., 68%) for the chosen significance level (32%).

considerably lower. This suggests that the nUNC AD method may largely restore the validity of conformal regression in some cases. Moreover, the average coverage outside the domain was typically lower (75% of the time) than the average inside the domain.

Syngenta Classification Data Set: Prediction Performance for Nonrandom Test Sets (Time-Splits). For the prediction performance metrics, the average shift metrics typically (75% of time-splits) showed the expected sign, with the curious exception of AUC. None of the average shift metrics with the expected sign were associated with statistically significant one-tail *p*-values. The in-domain values for these metrics were always better than expected by chance, albeit this was almost always the case for the out-of-domain metric values, notwithstanding the fact that these were typically lower.

Syngenta Classification Data Set: Uncertainty Estimation Performance for Nonrandom Test Sets (Time-Splits). For the uncertainty estimation performance metrics, the corresponding average shift metrics typically showed the expected sign (75−100% of time-splits), albeit these were almost never associated with statistically significant one-tail *p*-values, save for one average SBS shift metric prior to adjustment. For the relevant metrics, both the in-domain and out-of-domain metrics were typically higher than expected due to chance, even though the out-of-domain values were typically lower.

Syngenta Regression Data Set: Prediction Performance for Nonrandom Test Sets (Time-Splits). For the prediction performance metrics, all the average shift metrics had the expected sign. Moreover, just over half of these were associated with statistically significant adjusted one-tail *p*-values and the rest were associated with statistically significant unadjusted one-tail *p*-values. For relevant metrics, the in-domain and out-of-domain metrics were both always better than expected from chance, albeit the in-domain average metrics were consistently larger.

Syngenta Regression Data Set: Uncertainty Estimation Performance for Nonrandom Test Sets (Time-Splits). For the uncertainty estimation performance metrics, all their average shift-metric values had the expected sign, with the exception of SCC for 50% of time-splits. However, for most of the metrics (ECE, ENCE, coverage), only 25% of these results were associated with statistically significant adjusted one-tail *p*-values. (This was the case for all and

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

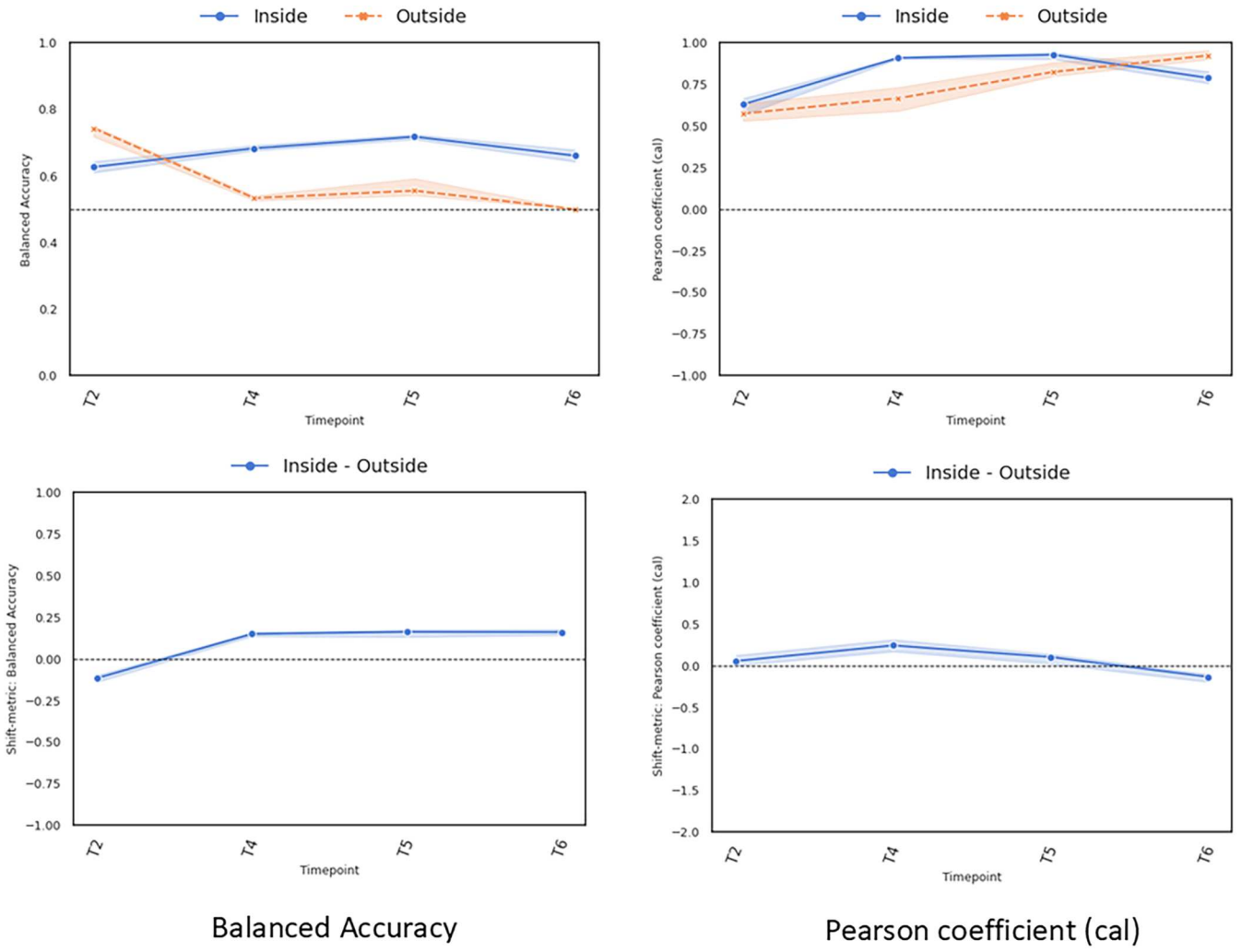


Figure 8. Variation across time-splits of Balanced Accuracy and Pearson coefficient (cal) values inside vs outside the domain (top panel) and the corresponding shift-metric values (bottom panel), for the Syngenta DT50 data set. See Figure 5 caption for description of dotted lines, shading, and annotations.

none of the time-splits in the case of efficiency and SCC, respectively.) For SCC and ENCE, some more of these results were associated with statistically significant unadjusted one-tail *p*-values.

For the one relevant metric with an a priori random baseline, SCC, the in-domain values were always better than expected due to chance, while this was only observed for 50% of time-splits for the out-of-domain values.

Looking at the in-domain coverage values (Figure 9), two of the four time-splits were observed to show average values greater than the theoretical minimum (0.68) expected if the in-domain compounds were sampled from the same distribution as the training set, with another being very close to this. As this was never observed for the out-of-domain values, this indicates the AD method was working to restore validity in these cases.

Summary of Observed Trends with the Selected AD and Uncertainty Methods. We summarize the trends for the average shift metrics, which we discussed in detail above, in Table 7 and Table 8 for the public domain and Syngenta data sets, respectively.

Overall, the results show that the nUNC AD method typically produced the expected sign for the average shift-metric values for both prediction and uncertainty estimation performance metrics for the nonrandom test sets across both the public and Syngenta data sets. (As can be seen in the plots in the Supporting Information, this was even more commonly observed for the random test sets of the public domain data sets. However, as noted above, we focus on the results for the nonrandom test sets, since these are of greater practical relevance.) Some of these results for the nonrandom test sets were statistically significant, according to one-tail *p*-values, even after adjusting for multiple testing. (As might be expected, given that molecules were sampled from the same distribution, the corresponding adjusted one-tail *p*-values for the random test sets were less likely to be statistically significant.) In some cases, i.e., the Morger-ChEMBL data sets and the Syngenta regression (logP) results, both the frequency with which the average shift metrics had the expected sign and the frequency with which they also were associated with statistically significant, adjusted one-tail *p*-values were higher for the prediction performance compared to the uncertainty estimation performance metrics, but this was not consistently observed.

Another important finding was that the prediction and uncertainty metrics inside the domain, according to nUNC, were typically better than expected from chance, for the

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

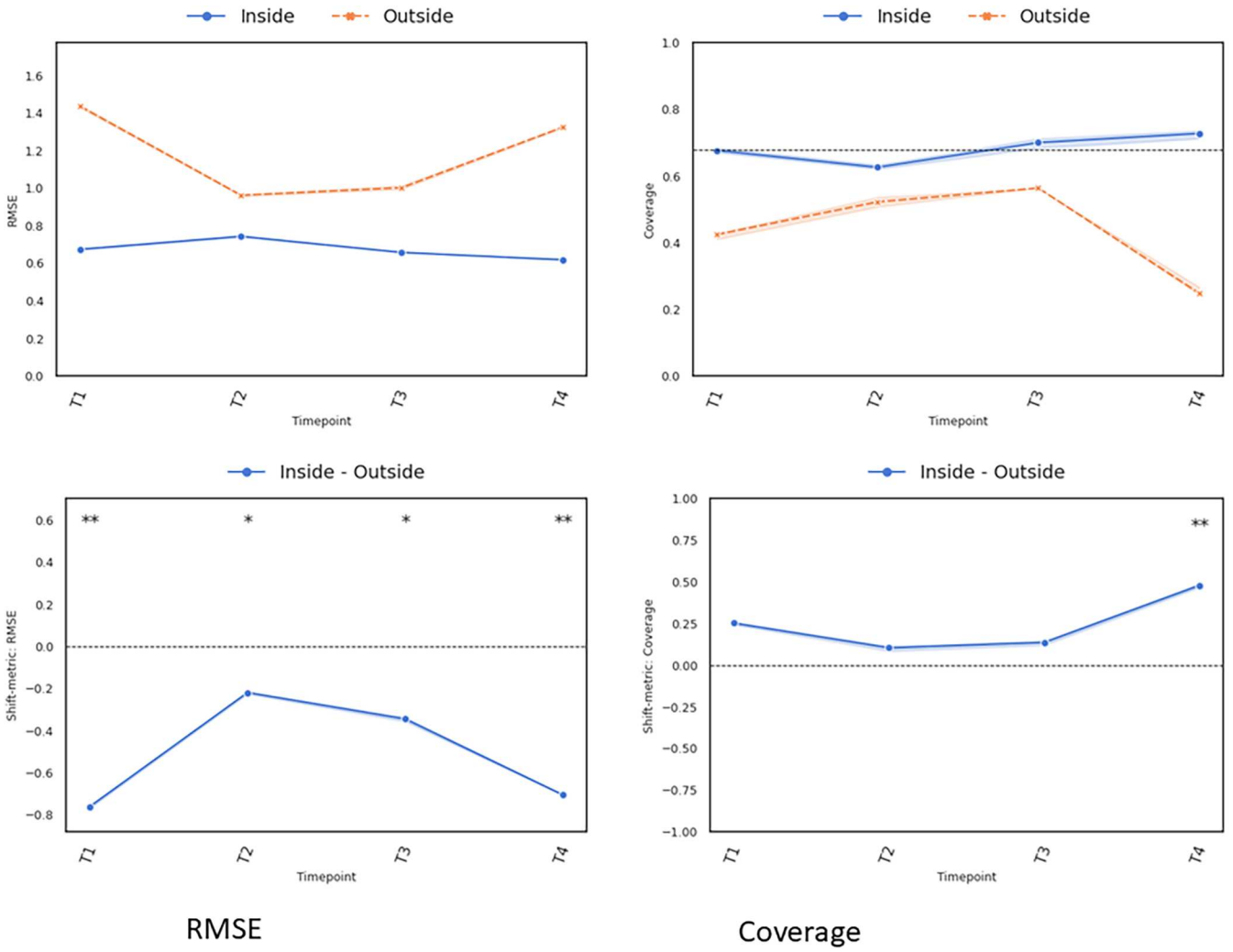


Figure 9. Variation across time-splits of RMSE and coverage values inside vs outside the domain (top panel) and the corresponding shift-metric values (bottom panel), for the Syngenta logP data set. See Figure 7 caption for description of dotted lines, shading, and annotations.

Table 7. Summary of Trends in the Average Shift Metrics for the Non-Random Test Sets from the Public Domain Datasets

prediction type	data-set group	performance metric type	average shift metrics have the expected sign (%)[a]	average shift metrics have the expected sign and are associated with statistically significant adjusted one-tail p-values (%)[a]	average shift metrics do not have the expected sign and are associated with statistically significant adjusted two-tail p-values (%)[a]
Classification	Morger-ChEMBL	Prediction	67−83	17−50	0−8
Classification	Morger-ChEMBL	Uncertainty	25−92	0−50[b]	0−8
Classification	Morger-Tox21	Prediction	25−100	0−0	0−0
Classification	Morger-Tox21	Uncertainty	67−100	0−0	0−0
Regression	Wang-ChEMBL	Prediction	85−100	0−5	0−0
Regression	Wang-ChEMBL	Uncertainty	60−95	0−20	0−0

[a]The range of percentages is presented, out of the total number of targets for which the average shift metric was computed, across the complete set of relevant metric−test set type combinations. [b]In eight out of 10 test set type−metric combinations, at least some of these values were statistically significant.

relevant metrics with an a priori random baseline. In addition, it was notable that validity for conformal regression (ACP) was sometimes restored.

The fact that shift metrics were often not statistically significant, even though the average shift metrics had the expected sign, could reflect insufficient data to ensure that the significance tests had the appropriate level of power to detect genuine differences. In some cases, this could reflect a limited shift in chemical space between the training set and a nonrandom test set, such that few molecules would be deemed to lie outside the domain by even an optimal AD method.

Likewise, the observations that some average shift metrics had the unexpected sign, i.e., the performance metrics were worse inside the domain than outside, were expected to be statistical flukes. The cases where the average shift metrics had the opposite sign to the one expected and the adjusted two-tail p-values were statistically significant represented 6.8% of all cases where the two-tail adjusted p-values were statistically

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

Table 8. Summary of Trends in the Average Shift Metrics for the Time-Splits of the Syngenta Datasets

prediction type	data set	performance metric type	average shift metrics have the expected sign (%)[a]	average shift metrics have the expected sign and are associated with statistically significant adjusted one-tail p-values (%)[a]	average shift metrics do not have the expected sign and are associated with statistically significant adjusted two-tail p-values (%)[a]
Classification	DT50	Prediction	0−75	0−0	0−0
Classification	DT50	Uncertainty	75−100	0−0	0−0
Regression	logP	Prediction	100−100	50−75	0−0
Regression	logP	Uncertainty	50−100	0−100[b]	0−0

[a]The range of percentages is presented, out of all time-splits, across the complete set of relevant metrics. [b]For three out of five relevant metrics, this value was 25%.

significant. That is a bit higher than the expected limit of 5.5% (given that we deemed p-values to be statistically significant when they were less than or equal to 5% when rounded to the nearest percent) of false-discoveries (chance findings)[51] with the multiple testing adjustment procedure we employed.[52] However, the expected limit may be violated in practice for insufficiently large sample sizes.

In summary, we conclude that shift metrics with unexpected signs are statistical flukes and the lack of statistically significant results for some cases where the expected average shift metric was observed may reflect insufficient data to achieve statistical significance. Nonetheless, in keeping with the no-free-lunch theorems,[58,59] and other work suggesting the optimal applicability domain metrics may depend on training set diversity,[60] we do not claim that the nUNC AD method we chose to focus on in this work can be considered optimal for all data sets.

However, while previous work has suggested that structural similarity to one or more nearest neighbors in the training set may be better suited to defining the AD for less diverse training sets,[60] we did not observe this to be the case. This is illustrated by the plots of mean pairwise Tanimoto distance against shift-metric values presented in the Supporting Information for key metrics, for which we anticipated a general trend toward shift-metric values with larger values with the expected sign (Table 1) as training set diversity reduced. This apparent discrepancy might reflect the fact that this earlier work used methods (SIMILARITYNEAREST1, SIMILARITYNEAREST5) which, unlike nUNC, do not scale their measure of distance to training set nearest neighbors to account for training set diversity.[60] Nonetheless, the influence of training set diversity on the behavior of the nUNC AD method is reflected in the smaller proportion of test set compounds deemed to lie outside the domain for the more diverse Morger-Tox21 data sets (2−4% for the nonrandom test sets and 1−3% for the randomly sampled test sets) compared to the other data sets derived from ChEMBL (57−96%, 0−90% for the nonrandom Morger-ChEMBL, Wang-ChEMBL test sets and 5−11%, 0−16% for the randomly sampled test sets) and Syngenta discovery projects (15−39% for the time-split test sets).

Still, in spite of not having optimized this AD method for individual data sets, the fact that we generally observed that this default approach could differentiate between more reliable (inside the domain) and less reliable (outside the domain) uncertainty estimates and predictions, on average, was promising. We consider how this investigation could be extended in future work below.

DISCUSSION

In summary, we typically observed that uncertainty estimates generated using conformal regression (ACP) and Venn-ABERS (CVAP) were less reliable, on average, outside the domain according to the nUNC AD method. This was expected on theoretical grounds and is consistent with previous studies which considered applications to nonrandom test sets, e.g., temporal, scaffold or cluster-based test sets.[15,20,25,27,29] In contrast to earlier studies, we explicitly explored how to identify molecules within sets of nonrandom tests that lie inside the domain, such that the performance of these commonly used uncertainty methods could be (partially) restored. This has practical relevance for pharmaceutical or plant protection product discovery projects, where new compounds for which predictions and uncertainty estimates are needed may lie inside or outside the domain.

While we considered standard conformal prediction approaches which have been explored in various cheminformatics studies,[1,4,14,33] the wider Machine Learning community has developed a variety of novel conformal prediction methodologies which are designed to address the expected loss of validity under data shift, i.e., under "non-exchangeability".[14] These include methodologies designed to work for different kinds of data shifts, which may have various degrees of suitability for recovering validity in uncertainty estimation for predictions made for novel chemistry, i.e., under "covariate shift". However, we are only aware of two publications which report extensions of approaches for handling covariate shift to cheminformatics.[61,62] One of these is only applicable in the special case where the test set distribution is contingent on the training set distribution, e.g., in active learning, and can be defined a priori.[62] In contrast, the work of Laghuvarapu et al.[61] addresses the real-world problem encountered in the present work, where the distribution of molecules, for which predictions are made, is not known in advance. Their approach adjusts the prediction intervals based on estimating the data shift using a set of molecular structures from the chemical space for which predictions are desired. However, they found that the loss of coverage may be only partially restored in practice, because the number of molecular structures, from the new chemical space, which are typically available may be insufficient to fully characterize the data shift. Moreover, this approach would require recalibration for each new distribution of molecules, which would impose greater computational cost than traditional conformal prediction methodologies. Finally, the authors only developed their approach for classification conformal prediction, which was outside the scope of our work, rather than conformal regression.

Here, it should also be noted that some other cheminformatics studies also found that the validity of

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

classification conformal prediction or conformal regression may sometimes be partially or, for some data sets, even wholly restored by updating the calibration set.[26−28] Specifically, these studies considered replacing the standard randomly sampled calibration sets using more recently tested compounds[26,28] or compounds corresponding to different clusters in chemical space.[27] Interestingly, these variations on the standard conformal prediction approach considered in our work partially or wholly restored validity for some data sets even when these compounds were not sampled from the same distribution as the test set compounds.[26−28]

More broadly, other studies, considering various modeling and uncertainty estimation methods, have reported mixed findings regarding whether uncertainty estimation or predictive performance can be expected to be less reliable for chemicals that are structurally distinct from the training and, where applicable, calibration sets.[12,27,37,63−65] In some cases, this could reflect data set dependency of the results,[63] especially where test sets are considered, which may, even where nonrandom, show limited covariate shift.[29,37] Indeed, our own results showed considerable variation in the change in uncertainty estimation performance metrics upon moving from inside to outside the domain across different data sets as well as with different nonrandom test sets, including different folds for the Wang-ChEMBL data sets.

We also observed, in keeping with prior work,[29,66] that other sources of variability in uncertainty estimation performance include the choice of performance metric: not all performance trends, for the same data sets and test sets, were consistent between different metrics. Hence, the fact that not all prior studies concluded that uncertainty estimation is less reliable for novel chemistry could also partially reflect the use of different evaluation metrics and evaluation approaches. (Other ways to evaluate whether uncertainty methods remain robust to novel chemistry include plotting the RMSE against average estimated uncertainty for subsets of molecules with increasingly lower estimated uncertainty.)[64] The fact that different performance metrics could lead to different conclusions is also something which should be carefully considered when evaluating novel variations of conformal regression or other uncertainty estimates which are claimed to (partially) restore coverage under covariate shift. For example, if the (partial) improvement in coverage comes at the expense of overestimating the likely prediction residuals, i.e., making the prediction intervals less efficient in the case of conformal regression, this may not be desirable.

Notwithstanding the variability in results that may be obtained across different data sets and performance metrics, our results across a range of data sets and metrics indicated that prediction performance may sometimes be more strongly affected by the domain status than uncertainty estimation performance. While not consistently observed, this was observed for the Morger-ChEMBL classification data sets and the Syngenta regression (logP) data set. This suggests that the reliability of uncertainty estimation may sometimes be more robust to extrapolation in chemical space than prediction reliability. Similar findings were also reported in some prior studies.[37,63]

Nonetheless, the potential loss of uncertainty estimation reliability due to data shifts was also recognized as a challenge in recent studies, which sought to relate the degree of reduced uncertainty estimation performance to the degree of data shift associated with a given train/test split.[29,63] Indeed, Tossou et al.[63] examined how the reliability of uncertainty estimation using various methods, albeit not conformal regression or Venn-ABERS, varied with distance from the training set within a randomly sampled test set. They subsequently investigated whether the uncertainty calibration "gap" associated with a real-world covariate shift could be, at least partially, corrected. They explored this via partitioning the available data to approximately match the covariate shift expected for a real-world model deployment scenario. However, they did not claim to fully close this calibration "gap". Moreover, their approach supposes that the calibration "gap" is constant irrespective of the direction and starting point from which one is traveling in chemical space.

In contrast to previous work, we propose that compounds for which uncertainty estimation could be less reliable should be flagged using an AD calculation. While different variations of conformal prediction have been proposed to serve the role of classifying compounds as outside the domain, e.g., by generating "null" predictions[1] or abstaining from prediction,[11] those studies did not consider how to use computed AD status to differentiate between reliable and less reliable uncertainty estimates. While finalizing our work for publication, Kim et al.[30] reported an interesting approach that partially addressed this question for conformal regression. However, their domain identification approaches rely upon access to endpoint values for some out-of-domain molecules to construct models underpinning domain status assignment. Since these values would not be available for all real-world applications, this makes their approach less generalizable than the AD method considered in our work.

We also propose that future research should focus on developing new methods for uncertainty estimation, which may be more robust to data shifts, such as extensions of the work of Laghuvarapu and co-workers[61] to enhance conformal regression and Venn-ABERS estimators. Building on other cheminformatics studies of classification conformal prediction and conformal regression[26−28] it would be worth comparing these approaches to replacing the standard randomly sampled calibration sets for conformal regression and Venn-ABERS with calibration sets chosen via cluster splits of the training sets.

These future studies should also investigate the extent to which these and other uncertainty methods[29,37,63,64] would still benefit from flagging molecules deemed to lie outside the domain using an AD method, including consideration of multiple different performance metrics, rather than just the restoration of the desired coverage for conformal regression. It is conceivable that these approaches could be complementary, with the AD method still flagging molecules for which the uncertainty estimates could be less reliable, even when the extent to which that was the case might be reduced for some scenarios. Tuning of the AD method to specific uncertainty estimation techniques and data sets could enhance the complementarity of these approaches.

Hence, future research should also explore other AD methods and the benefits of tuning the AD approach, including AD method parameters, for specific uncertainty methods and data sets. This might build on the recent work of Kaneko, which presented a framework for tuning the AD approach, based on structural similarity to the training set, for specific data sets, albeit this framework simultaneously optimized the trade-off between prediction performance and the percentage inside the domain rather than considering uncertainty

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

estimates for individual predictions.[67] Another interesting possibility would be to benchmark AD methods based on structural similarity, such as the k-nearest neighbors method (nUNC), which was the focus of our work, against thresholds derived from uncertainty estimation approaches. This could include identifying a suitable threshold value for the discordance between the upper (p_1) and lower (p_0) probability estimates obtained from Venn-ABERS, which has been shown to loosely correlate with structural similarity.[15] This could also include, as per a very recent publication, tuning an uncertainty threshold to delimit the AD for specific data sets via a cross-validation procedure.[68] It would also be interesting to see if tuning the AD threshold for specific data sets using a cluster-based cross-validation procedure could yield better results, given recently published work indicating that model selection based on in-domain validation sets may perform poorly when tuning models for out-of-domain performance,[65] albeit other work has suggested limited benefits of model selection using nonrandom splits for uncertainty calibration.[63] Tuning the AD approach for individual data sets could also be performed using an initial temporal validation set prior to evaluating on a further temporal test set and subsequent application to experimentally untested molecules. In isolation or in combination, investigating different kinds of AD methods and task-tuned thresholds could enhance the usefulness of AD methods for specific models beyond the default AD method, which was the focus of our work. However, it should also be noted that tuning the method to particular data sets would also increase the computational cost of model development and might increase the risk of overfitting.

Notwithstanding the potential for future studies to build on and refine our work, we note that our proposed framework has practical applications in its current form. We demonstrated that a structural similarity-based, default AD method can typically be used to identify when uncertainty estimates generated using standard conformal regression and Venn-ABERS methods are, on average, less reliable. While the extent to which differences in uncertainty estimation performance metrics were observed inside vs outside the domain varied between data sets, we note this was also true for the prediction performance metrics. Indeed, this is in keeping with the literature, with reported analyses indicating the extent to which AD methods differentiate between less and more reliable predictions may vary considerably between data sets.[32] Nonetheless, even if they simply flag out-of-domain predictions as potentially "less reliable", rather than completely unreliable, practical applications may reasonably assume that "a QSAR model's predictions are valid only if the subject molecule resides within the model's AD".[69] Indeed, the observation that some observed differences in performance metrics are small inside vs outside the domain may reflect limited extrapolation in chemical space for the out-of-domain compounds for some test sets, i.e., larger differences may be observed for newly encountered molecules for which predictions are desired in a practical setting. By extension, one practical application of the framework investigated in our work would be to only use the predictions and uncertainty estimates for decision-making for compounds inside the domain. A pragmatic simplification would be to assume that predictions outside the domain had maximum uncertainty. The extent to which this would be an appropriate simplification would depend upon the context, i.e., the risks of using potentially less reliable uncertainty estimates to inform decisions.

CONCLUSION

To conclude, we investigated whether explicit applicability domain calculations can differentiate between compounds where uncertainty estimates are more (inside the domain) vs less (outside the domain) reliable. We focused on the use of a k-nearest neighbor method (nUNC) in combination with Cross-Venn-ABERS Predictors (CVAP) and Aggregated Conformal Prediction (ACP), for classification and regression tasks, respectively, across a wide range of public domain data sets and some temporal splits of industrial data sets. These uncertainty estimation methods are widely used and have been extensively studied in cheminformatics over the past decade. Previous studies have identified that these uncertainty estimation methods can lose their validity with nonrandom train/test splits, e.g., temporal, scaffold or cluster splits. However, the key contribution of our study was to investigate whether an explicit applicability domain calculation could differentiate between reliable and less reliable predictions and uncertainty estimates when applying a defined model to a nonrandom test set. This was an important question to address as flagging when these uncertainty estimates become less reliable for novel molecules, including from regions of chemical space without experimental data to refine uncertainty estimates or help define the domain, is of huge value for practical applications.

In some cases, the use of the nUNC applicability domain approach was able to identify compounds where the expected validity of conformal regression was maintained. Moreover, we observed that this applicability domain method was typically able to distinguish between predictions and uncertainty estimates, which were, on average, more (inside the domain) vs less (outside the domain) reliable. In some cases, we concluded that the differences in performance metrics between compounds inside and outside the domain, for predictions and uncertainty estimates, were statistically significant.

ASSOCIATED CONTENT

Data Availability Statement

All Python code used to generate the results reported for this study has been made available under the terms of the GNU Public License (GPL-3.0) on GitHub (https://github.com/syngenta/QSAR_AD_Uncertainty_Paper_Code_release) and a snapshot of the version (https://github.com/syngenta/QSAR_AD_Uncertainty_Paper_Code_release/releases/tag/v4) corresponding to the results reported is also available on Zenodo (10.5281/zenodo.17570872). The code repository includes a README with step-by-step instructions on how to install the necessary dependencies, including the versions of all Open Source Python modules, download the public data sets, and generate the results. However, due to their commercial sensitivity, the Syngenta data sets are not made publicly available.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.5c11875.

File 1: Further details regarding the methods and results, organized under headings matching the corresponding sections in the main text. As indicated in the main text, further details regarding the methods include more detailed explanations of uncertainty estimation performance metrics derived from delta-calibration plots, along

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

with a more detailed explanation of the data set curation workflow. Likewise, further details regarding the results include t-SNE plots of training and test set distributions and the remaining performance metric and shift-metric plots for which the trends are summarized in the main text (PDF)

File 2: Tables of raw results (Tables ES1−ES12), each of which is documented in the README sheet, with Table ES0 documenting how the names used in the raw results correspond to the names used in the manuscript, where these are different (XLSX)

AUTHOR INFORMATION

Corresponding Authors

Valerie J. Gillet − *Information School, University of Sheffield, Sheffield S10 2AH, U.K.;* orcid.org/0000-0002-8403-3111; Email: v.gillet@sheffield.ac.uk

Richard L. Marchese Robinson − *Product Safety Department, Jealott's Hill International Research Centre, Syngenta Crop Protection, Berkshire RG42 6EY, U.K.;* orcid.org/0000-0001-7648-8645; Email: richard.marchese_robinson@syngenta.com

Author

Zied Hosni − *Information School, University of Sheffield, Sheffield S10 2AH, U.K.;* orcid.org/0000-0001-7889-0005

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.5c11875

Author Contributions

[§]Z.H. and R.L.M.R. contributed equally. R.L.M.R. in consultation with colleagues at Syngenta and V.J.G., proposed the original research plan. Z.H. and R.L.M.R. cowrote the Python code used to generate all results reported herein. R.L.M.R., Z.H., and V.J.G. regularly discussed the results and refined the project plan throughout the entirety of the research project. Z.H. wrote the first draft of the manuscript, which was subsequently revised with contributions from R.L.M.R. and V.J.G. All authors were involved in preparing the final draft of the manuscript and have approved the final text.

Funding

The authors are grateful for funding from Syngenta via the Cross-Functional Collaboration Fund (FCC1).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

R.L.M.R. thanks Konstantinos Papachristos (AstraZeneca) for valuable discussions, while he was working at Syngenta, regarding the uncertainty estimation and applicability domain methods investigated in this work, which stimulated the research reported herein. R.L.M.R. further thanks Ben Robinson and Azadi Golbamaki for guidance on the preparation of time-splits for the Syngenta logP dataset and Anne Dalencon for guidance on the preparation of the Syngenta DT50 dataset and proofreading. R.L.M.R. also thanks the following colleagues at Syngenta for their comments during the internal review process: Michael Fernandez Llamosa, Matina Zavitsanou, Nick Mulholland, Emily Scorgie, and Alex Porter. The following researchers are thanked for assistance with clarifying the copyright and/or license terms under which their code could be adapted and redistributed: Ola Spjuth (Uppsala University), Ando Saabas, Greg Landrum, Alex Gammerman, and Paolo Toccaceli (Royal Holloway, University of London). Paolo Toccaceli is also thanked for assistance with understanding how to use the Venn-ABERS implementation employed in this work.

ABBREVIATIONS

RF, Random Forest; ACP, Aggregated Conformal Prediction; CVAP, Cross-Venn-ABERS Predictors; AD, applicability domain; RMSE, root-mean-square error; ECE, expected calibration error; ENCE, expected normalized calibration error

REFERENCES

(1) Cortés-Ciriano, I.; Bender, A.Concepts and Applications of Conformal Prediction in Computational Drug Discovery, 2019, arXiv:1908.03569. arXiv.org e-Print archive https://arxiv.org/abs/1908.03569.

(2) Hanser, T.; Barber, C.; Guesné, S.; Marchaland, J. F.; Werner, S. Applicability Domain: Towards a More Formal Framework to Express the Applicability of a Model and the Confidence in Individual Predictions. In *Advances in Computational Toxicology: Methodologies and Applications in Regulatory Science*; Hong, H., Ed.; Springer International Publishing: Cham, 2019; pp 215−232.

(3) Mervin, L. H.; Johansson, S.; Semenova, E.; Giblin, K. A.; Engkvist, O. Uncertainty Quantification in Drug Design. *Drug Discovery Today* **2021**, *26*, 474.

(4) Svensson, F.; Aniceto, N.; Norinder, U.; Cortes-Ciriano, I.; Spjuth, O.; Carlsson, L.; Bender, A. Conformal Regression for Quantitative Structure−Activity Relationship Modeling—Quantifying Prediction Uncertainty. *J. Chem. Inf. Model.* **2018**, *58* (5), 1132−1140.

(5) Hanser, T.; Barber, C.; Marchaland, J. F.; Werner, S. Applicability Domain: Towards a More Formal Definition. *SAR QSAR Environ. Res.* **2016**, *27* (11), 865−881.

(6) Yin, T.; Panapitiya, G.; Coda, E. D.; Saldanha, E. G. Evaluating Uncertainty-Based Active Learning for Accelerating the Generalization of Molecular Property Prediction. *J. Cheminformatics* **2023**, *15* (1), 105.

(7) Segall, M. D.; Beresford, A. P.; Gola, J. M.; Hawksley, D.; Tarbit, M. H. Focus on Success: Using a Probabilistic Approach to Achieve an Optimal Balance of Compound Properties in Drug Discovery. *Expert Opin. Drug Metab. Toxicol.* **2006**, *2*, 325 DOI: 10.1517/17425255.2.2.325.

(8) Segall, M.D. Multi-Parameter Optimization: Identifying High Quality Compounds with a Balance of Properties. *Curr. Pharm. Des.* **2012**, *18* (9), 1292−1310.

(9) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912.

(10) McShane, S. A.; Norinder, U.; Alvarsson, Jonathan.; Ahlberg, E.; Carlsson, L.; Spjuth, Ola. CPSign: Conformal Prediction for Cheminformatics Modeling. *J. Cheminformatics* **2024**, *16*, 75.

(11) Oršolić, D.; Šmuc, T. Dynamic Applicability Domain (dAD): Compound−Target Binding Affinity Estimates with Local Conformal Prediction. *Bioinformatics* **2023**, *39* (8), No. btad465.

(12) Fan, Y. J.; Allen, J. E.; McLoughlin, K. S.; Shi, D.; Bennion, B. J.; Zhang, X.; Lightstone, F. C. Evaluating Point-Prediction Uncertainties in Neural Networks for Protein-Ligand Binding Prediction. *Artif. Intell. Chem.* **2023**, *1* (1), No. 100004.

(13) Svensson, E.; Friesacher, H. R.; Winiwarter, S.; Mervin, L.; Arany, A.; Engkvist, O. Enhancing Uncertainty Quantification in Drug Discovery with Censored Regression Labels. *Artif. Intell. Life Sci.* **2025**, *7*, No. 100128.

(14) Astigarraga, M.; Sánchez-Ruiz, A.; Colmenarejo, G. Conformal Prediction-Based Machine Learning in Cheminformatics: Current

Applications and New Challenges. *Artif. Intell. Life Sci.* **2025**, *7*, No. 100127.
(15) Mervin, L. H.; Afzal, A. M.; Engkvist, O.; Bender, A. Comparison of Scaling Methods to Obtain Calibrated Probabilities of Activity for Protein−Ligand Predictions. *J. Chem. Inf. Model.* **2020**, *60* (10), 4546−4559.
(16) Manokhin, V. Multi-Class Probabilistic Classification Using Inductive and Cross Venn−Abers Predictors. In *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*; Gammerman, A.; Vovk, V.; Luo, Z.; Papadopoulos, H., Eds.; PMLR, 2017; Vol. *60*, pp 228−240.
(17) Vovk, V.; Petej, I.; Fedorova, V. Large-Scale Probabilistic Predictors with and without Guarantees of Validity. In *Advances in Neural Information Processing Systems*; Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; Garnett, R., Eds.; Curran Associates, Inc, 2015; Vol. *28*.
(18) Arvidsson, S.; Spjuth, O.; Carlsson, L.; Toccaceli PaoloToccaceli, P.; Gammerman, A.; Vovk, V.; Luo, Z.; Papadopoulos, H. Prediction of Metabolic Transformations Using Cross Venn-ABERS Predictors. In *Conformal and Probabilistic Prediction and Applications*, 2017.
(19) Xu, Y.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Development and Evaluation of Conformal Prediction Methods for Quantitative Structure−Activity Relationship. *ACS Omega* **2024**, *9* (27), 29478−29490.
(20) Friesacher, H. R.; Svensson, E.; Arany, A.; Mervin, L.; Engkvist, O. Temporal Evaluation of Probability Calibration with Experimental Errors. In *AI in Drug Discovery*; Clevert, D.-A.; Wand, M.; Malinovská, K.; Schmidhuber, J.; Tetko, I. V., Eds.; Springer Nature Switzerland: Cham, 2025; pp 13−20.
(21) Svensson, F.; Norinder, U. Conformal Prediction for Ecotoxicology and Implications for Regulatory Decision-Making. In *Methods in Pharmacology and Toxicology*; Springer US: New York, NY, 2020.
(22) Papadopoulos, H.; Vovk, V.; Gammerman, A. Regression Conformal Prediction with Nearest Neighbours. *J. Artif. Intell. Res.* **2011**, *40*, 815−840.
(23) Hosni, Z.; Chen, X.; Achour, S.; Saadi, F. Predictive Modeling of Yield Sooting Index Using Machine Learning with Uncertainty Estimation. *ACS Omega* **2025**, *10* (24), 25336−25349.
(24) McShane, S. A.; Ahlberg, E.; Noeske, T.; Spjuth, O. Machine Learning Strategies When Transitioning between Biological Assays. *J. Chem. Inf. Model.* **2021**, *61* (7), 3722−3733.
(25) Cortés-Ciriano, I.; Firth, N. C.; Bender, A.; Watson, O. Discovering Highly Potent Molecules from an Initial Set of Inactives Using Iterative Screening. *J. Chem. Inf. Model.* **2018**, *58* (9), 2000−2014.
(26) Morger, A.; Svensson, F.; Arvidsson McShane, S.; Gauraha, N.; Norinder, U.; Spjuth, O.; Volkamer, A. Assessing the Calibration in Toxicological in Vitro Models with Conformal Prediction. *J. Cheminformatics* **2021**, *13* (1), 35.
(27) Wang, D.; Yu, J.; Chen, L.; Li, X.; Jiang, H.; Chen, K.; Zheng, M.; Luo, X. A Hybrid Framework for Improving Uncertainty Quantification in Deep Learning-Based QSAR Regression Modeling. *J. Cheminformatics* **2021**, *13* (1), 69.
(28) Morger, A.; Garcia de Lomana, M.; Norinder, U.; Svensson, F.; Kirchmair, J.; Mathea, M.; Volkamer, A. Studying and Mitigating the Effects of Data Drifts on ML Model Performance at the Example of Chemical Toxicity Data. *Sci. Rep.* **2022**, *12* (1), No. 7244.
(29) Friesacher, H. R.; Svensson, E.; Winiwarter, S.; Mervin, L.; Arany, A.; Engkvist, O. Temporal Distribution Shift in Real-World Pharmaceutical Data: Implications for Uncertainty Quantification in QSAR Models, 2025, arXiv:2502.03982. arXiv.org e-Print archive https://arxiv.org/abs/2502.03982.
(30) Kim, J. Y.; Vlachos, D. G. Distance-Aware Molecular Property Prediction in Nonlinear Structure−Property Space. *J. Chem. Inf. Model.* **2025**, *65*, 6744.
(31) Sheridan, R. P.; Culberson, J. C.; Joshi, E.; Tudor, M.; Karnachi, P. Prediction Accuracy of Production ADMET Models as a Function of Version: Activity Cliffs Rule. *J. Chem. Inf. Model.* **2022**, *62*, 3275.
(32) Aniceto, N.; Freitas, A. A.; Bender, A.; Ghafourian, T. A Novel Applicability Domain Technique for Mapping Predictive Reliability across the Chemical Space of a QSAR: Reliability-Density Neighbourhood. *J. Cheminformatics* **2016**, *8*, 69 DOI: 10.1186/s13321-016-0182-y.
(33) Lindh, M.; Karlén, A.; Norinder, U. Predicting the Rate of Skin Penetration Using an Aggregated Conformal Prediction Framework. *Mol. Pharmaceutics* **2017**, *14* (5), 1571−1576.
(34) Gauraha, N.; Spjuth, O. Synergy Conformal Prediction for Regression. In *ICPRAM 2021 - Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods*, 2021.
(35) Svetnik, V.; Liaw, A.; Tong, C.; Christopher Culberson, J.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958, DOI: 10.1021/ci034160g.
(36) Sheridan, R. P. Stability of Prediction in Production ADMET Models as a Function of Version: Why and When Predictions Change. *J. Chem. Inf. Model.* **2022**, *62* (15), 3477−3485.
(37) Svensson, E.; Friesacher, H. R.; Arany, A.; Mervin, L.; Engkvist, O. Temporal Evaluation of Uncertainty Quantification Under Distribution Shift. In *AI in Drug Discovery*; Clevert, D.-A.; Wand, M.; Malinovská, K.; Schmidhuber, J.; Tetko, I. V., Eds.; Springer Nature Switzerland: Cham, 2025; pp 132−148.
(38) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. 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*, 1733.
(39) Sahigara, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Defining a Novel K-Nearest Neighbours Approach to Assess the Applicability Domain of a QSAR Model for Reliable Predictions. *J. Cheminformatics* **2013**, *5*, 27 DOI: 10.1186/1758-2946-5-27.
(40) Walter, M.; Allen, L. N.; de la Vega de León, A.; Webb, S. J.; Gillet, V. J. Analysis of the Benefits of Imputation Models over Traditional QSAR Models for Toxicity Prediction. *J. Cheminformatics* **2022**, *14* (1), 32.
(41) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf Model* **2010**, *50* (5), 742−754.
(42) Landrum, G. Fingerprints in the RDKit. https://www.rdkit.org/UGM/2012/Landrum_RDKit_UGM.Fingerprints.Final.pptx.pdf (accessed May 03, 2025).
(43) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983−996.
(44) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.
(45) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(46) SciKit-Learn RandomForestClassifier Documentation (version 1.0.2). scikit-learn. https://scikit-learn.org/1.0/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier (accessed Dec 30, 2024).
(47) SciKit-Learn RandomForestRegressor Documentation (version 1.0.2). scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (accessed Feb 17, 2025).
(48) Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 1316.
(49) Collell, G.; Prelec, D.; Patil, K. R. A Simple Plug-in Bagging Ensemble Based on Threshold-Moving for Classifying Binary and Multiclass Imbalanced Data. *Neurocomputing* **2018**, *275*, 330−340.
(50) Polyak, A.; Rosenfeld, J. A.; Girirajan, S. An Assessment of Sex Bias in Neurodevelopmental Disorders. *Genome Med.* **2015**, *7* (1), 94.

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743

(51) Dudoit, S.; Shaffer, J. P.; Boldrick, J. C. Multiple Hypothesis Testing in Microarray Experiments. *Stat. Sci.* **2003**, *18* (1), 71−103.
(52) Benjamini, Y.; Yekutieli, D. The Control of the False Discovery Rate in Multiple Testing under Dependency. *Ann. Stat.* **2001**, *29* (4), 1165−1188.
(53) Choi, W.; Kim, I. Averaging P-Values under Exchangeability. *Stat. Probab. Lett.* **2023**, *194*, No. 109748.
(54) Baker, M. Statisticians Issue Warning over Misuse of P Values. *Nature* **2016**, *531* (7593), 151.
(55) Amrhein, V.; Greenland, S.; McShane, B. Scientists Rise up against Statistical Significance. *Nature* **2019**, *567*, 305−307.
(56) SciKit-Learn TSNE Documentation. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.manifold.TSNE.html (accessed April 12, 2025).
(57) van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(58) Sterkenburg, T. F.; Grünwald, P. D. The No-Free-Lunch Theorems of Supervised Learning. *Synthese* **2021**, *199* (3), 9979−10015.
(59) Reiss, T.; Cohen, N.; Hoshen, Y. No Free Lunch: The Hazards of Over-Expressive Representations in Anomaly Detection, 2023, arXiv:2306.07284. arXiv.org e-Print archive https://arxiv.org/abs/2306.07284.
(60) Sheridan, R. P. The Relative Importance of Domain Applicability Metrics for Estimating Prediction Errors in QSAR Varies with Training Set Diversity. *J. Chem. Inf. Model.* **2015**, *55*, 1098.
(61) Laghuvarapu, S.; Lin, Z.; Sun, J. CoDrug: Conformal Drug Property Prediction with Density Estimation under Covariate Shift. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 37728−37747.
(62) Fannjiang, C.; Bates, S.; Angelopoulos, A. N.; Listgarten, J.; Jordan, M. I. Conformal Prediction under Feedback Covariate Shift for Biomolecular Design. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119* (43), No. e2204569119.
(63) Tossou, P.; Wognum, C.; Craig, M.; Mary, H.; Noutahi, E. Real-World Molecular Out-Of-Distribution: Specification and Investigation. *J. Chem. Inf. Model.* **2024**, *64* (3), 697−711.
(64) Whitehead, T. M.; Irwin, B. W. J.; Hunt, P.; Segall, M. D.; Conduit, G. J. Imputation of Assay Bioactivity Data Using Deep Learning. *J. Chem. Inf. Model.* **2019**, *59* (3), 1197−1204.
(65) Fooladi, H.; Vu, T. N. L.; Mathea, M.; Kirchmair, J. Evaluating Machine Learning Models for Molecular Property Prediction: Performance and Robustness on Out-of-Distribution Data. *J. Chem. Inf. Model.* **2025**, *65*, 9871.
(66) Rasmussen, M. H.; Duan, C.; Kulik, H. J.; Jensen, J. H. Uncertain of Uncertainties? A Comparison of Uncertainty Quantification Metrics for Chemical Data Sets. *J. Cheminformatics* **2023**, *15* (1), 121.
(67) Kaneko, H. Evaluation and Optimization Methods for Applicability Domain Methods and Their Hyperparameters, Considering the Prediction Performance of Machine Learning Models. *ACS Omega* **2024**, *9* (10), 11453−11458.
(68) Liu, J.-W.; Liu, K.-Y.; Deng, Y.-C.; Shi, S.-H.; Fu, X.-Z.; Jiang, Y.-P.; Fang, J.; Zhang, Q.; Jiang, D.-J.; Liu, S.; Cao, D.-S. Uncertainty-Aware Deep Learning and Structural Feature Analysis for Reliable Nephrotoxicity Prediction. *J. Chem. Inf. Model.* **2025**, *65*, 9082.
(69) Yang, S.; Kar, S. Applicability Domain for Trustable Predictions. In *Computational Toxicology: Methods and Protocols*; Nicolotti, O., Ed.; Springer US: New York, NY, 2025; pp 131−149.

https://doi.org/10.1021/acsomega.5c11875
ACS Omega 2026, 11, 4722−4743