# Conformal Selection for Efficient and Accurate Compound Screening in Drug Discovery

Tian Bai,[†] Peng Tang,[†] Yuting Xu,[‡] Vladimir Svetnik,[‡] Abbas Khalili,[†] Xiang Yu,[*,¶] and Archer Y. Yang[*,§]

[†]*Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada*

[‡]*MRL, Merck & Co., Inc., Rahway, NJ, USA*

[¶]*Co-corresponding author; MRL, Merck & Co., Inc., Rahway, NJ, USA*

[§]*Corresponding author; Department of Mathematics and Statistics, McGill University; Mila - Quebec AI Institute, Montreal, Quebec, Canada*

E-mail: xiang.yu2@merck.com; archer.yang@mcgill.ca

## Abstract

In drug discovery, the reliability of compound screening based on manual assessments is compromised by potential bias, while existing methods lack robust risk control measures. To address these challenges, we introduced conformal selection as an enhanced approach to optimize the compound screening process with balanced risks and benefits. Leveraging conformal inference, our approach constructs $p$-values for each candidate molecule to quantify statistical evidence for selection. The final selection of molecules is determined by comparing these $p$-values against thresholds derived from multiple testing principles. Our approach offers rigorous control over the false discovery rate, ensuring validity independent of dataset size and requiring minimal assumptions. By avoiding the estimation of prediction errors required in previous approaches, our method achieves higher accuracy (power), thereby improving the ability to identify

https://doi.org/10.26434/chemrxiv-2024-pf3ph ORCID: https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

promising candidates. Furthermore, our method demonstrates superior computational efficiency. We validate these advantages through numerical simulations on real-world datasets.

# Introduction

In drug discovery, the process of selecting a subset of compounds from a diverse molecular pool typically precedes any resource-intensive steps. Compounds selected for further development must demonstrate strong biological activity against their intended targets while remaining inactive against a collection of potentially harmful off-targets. The evaluations for activity on targets and inactivity on off-targets are commonly known as "screening" and "counter-screening" respectively.

Since the relevant biological activities are often unavailable at the time of screening or counter-screening, predictive models for these activities would be invaluable in enabling chemists to make informed decisions during the selection process. Quantitative structure-activity relationship (QSAR) regression models serve this purpose by predicting biological activities based on molecular structure-derived features. Leveraging various machine learning architectures such as the random forest (RF)[1] or deep learning (DL),[2] QSAR models can typically achieve notable prediction accuracy when properly trained. The importance of quantifying uncertainties in QSAR model predictions has also been widely recognized as a key factor for making more reliable decisions. These uncertainty estimates typically provide an assessment of the QSAR model's prediction error or offer a range estimate of molecular activity, rather than a single-point prediction.

Despite significant efforts to enhance QSAR model predictions and uncertainty quantification, there has been limited discussion on the precise decision-making procedures. While current methods offer chemists and analysts a wealth of information for decision-making, the procedures themselves used in practice are often arbitrary, lack rigor, and fail to adequately control the risk of false compound selection. While the eCounterscreen method effectively

https://doi.org/10.26434/chemrxiv-2024-pf3ph ORCID: https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

manages the false selection risk in sufficiently large datasets,[3] it exhibits notable limitations in computational efficiency and its applicability to datasets of arbitrary size.

This paper introduces a unified decision procedure that can efficiently provide guaranteed risk control under minimal assumptions, which are typically already inherent in the use of QSAR models. We adopted conformal selection,[4] a statistical selection methodology. This approach maintains validity regardless of dataset size, performing effectively even when the total number of available training molecules is as low as a few hundred. This method integrates seamlessly with any pre-trained QSAR model, including those based on RF or DL, without requiring adjustments. Furthermore, our approach minimizes computational costs and exhibits superior efficiency compared to eCounterscreen. It achieves this by automatically determining the decision criterion based on a user-specified nominal risk level, eliminating the need for an exhaustive search of appropriate thresholds in the eCounterscreen. Overall, our study demonstrated that the conformal selection method not only ensures valid risk control but also outperforms previous methods in selecting more compounds accurately and efficiently.

# Methods

We will first provide an overview of the entire conformal selection process and then offer detailed explanations for some technical terms.

## Problem Formulation

Before presenting the conformal selection procedure, we first establish some essential notations and concepts. Let $X$ represent the available molecular structure features, and $Y$ denote the molecular activities. We use a dataset $D_{\text{train}} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ to train a QSAR model $\hat{\mu}(X)$ to approximate $Y$. In this dataset, the molecular activities are observed, which may be obtained from prior experiments. The conformal selection procedure also requires

https://doi.org/10.26434/chemrxiv-2024-pf3ph ORCID: https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

another calibration dataset $D_{\text{calib}} = \{(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m}))\}$ where the molecular activities $Y_{n+1}, \ldots, Y_{n+m}$ must also be known.

For the batch of incoming molecules subject to screening, we denote them as $D_{\text{test}} = \{X'_1, \ldots, X'_k\}$, with the corresponding true activity levels $Y'_1, \ldots, Y'_k$ remaining unobserved. For convenience, we assume a screening setting where the objective is to select as many compounds as possible with acceptably high activity levels, characterized by $Y'_j > c$, $j = 1, 2, \ldots, k$ where $c$ is the activity cutoff for a specific target. Our method is also applicable to the counterscreening setting with a straightforward transformation of the data, such as negating $Y'_j$ or transforming $Y'_j$ to binary targets $\mathbf{1}\{Y'_j > c\}$. The final element required is a user-specified risk level $\alpha$. The risk, or the false discovery rate (FDR), is defined as the average fraction of undesirable molecules with low activity levels $Y'_j \leq c$ among all molecules selected for advancement to the next stage, i.e.,

$$\text{FDR} = \mathbf{E}\left[\frac{|R \cap \{j : Y'_j \leq c\}|}{\max(1, |R|)}\right]$$

where $R$ is a subset of $\{1, 2, \ldots, m\}$ representing the indices of the selected molecules and $\mathbf{E}$ is the expectation taken over the joint distribution of the test dataset $D_{\text{test}}$. Intuitively, this value quantifies the percentage of unsuccessful molecular selections. Under risk control, it is advantageous for the selection procedure to identify as many desirable compounds as possible. This selection performance is measured by statistical power, defined as the average proportion of correctly selected molecules among all those suitable for selection, i.e.,

$$\text{Power} = \mathbf{E}\left[\frac{|R \cap \{j : Y'_j > c\}|}{\{j : Y'_j > c\}}\right].$$

## Conformal Selection

Based on the conformal prediction (CP) framework, conformal selection is a model-free selection procedure that provides finite-sample FDR control. The term "finite-sample control"

https://doi.org/10.26434/chemrxiv-2024-pf3ph **ORCID:** https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

signifies that the effectiveness of this procedure in controlling the FDR is not dependent on the size of the dataset. Conformal prediction, initially developed by Vovk et al.,[5] is a model-agnostic framework for uncertainty quantification, offering probabilistic guarantees on range estimates of the target value. Jin and Candès[4] extended this framework to establish the conformal selection procedure, which can be summarized as follows:

1. Train an arbitrary QSAR model $\hat{\mu}$, for example random forest or deep learning, using the training dataset $D_{\text{train}}$.

2. For each pair of molecular structure and activity $(X_{n+i}, Y_{n+i})$, $i = 1, 2, \ldots, m$ in the calibration dataset $D_{\text{calib}}$, compute its corresponding "calibration nonconformity score" $V_i = V(X_{n+i}, Y_{n+i})$ where $V$ is a pre-specified, fixed function called the nonconformity measure. Common choices for $V$ include:

$$V(X, Y) = Y - \hat{\mu}(X) \qquad \text{or} \qquad V(X, Y) = M \cdot \mathbf{1}\{Y > c\} - \hat{\mu}(X)$$

   where $\hat{\mu}(X) = \hat{Y}$ is the QSAR-predicted activity level and $M$ is a sufficiently large number that exceeds the usual activity level by several orders of magnitude. We recommend the second option, the clip method, as it typically yields superior performance in practice.

3. For each incoming test molecule $X'_j$, compute its "conformal $p$-value" defined as follows:

$$p_j = \frac{1}{n+1}\left[\Big|\{i = 1, \ldots, n : V_i < V(X'_j, c)\}\Big| + U \cdot \Big(1 + \Big|\{i = 1, \ldots, n : V_i = V(X'_j, c)\}\Big|\Big)\right]$$

   where $U$ is the realized value from an independent uniformly distributed random variable, $U \sim \text{Unif}(0, 1)$. The conformal $p$-value could be viewed as a smoothed rank of "test nonconformity score" $V(X'_j, c)$ among $V_1, \ldots, V_n$. In the case where the molecular

https://doi.org/10.26434/chemrxiv-2024-pf3ph ORCID: https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

activities are continuous, a simplified version of the conformal $p$-value can be used:

$$p_j = \frac{1}{n+1}\Big[\big|\{i = 1, \dots, n : V_i \leq V(X'_j, c)\}\big| + 1\Big].$$

4. Apply the Benjamini-Hochberg (BH) procedure[6] with a nominal level $\alpha$ to the set of conformal $p$-values obtained in step 3, $p_1, \dots, p_k$ to determine the selection set $R$. The BH procedure is a widely-used method for controlling the false discovery rate (FDR) and can be outlined in the following steps:

   4.1 Order the conformal $p$-values from smallest to largest, and denote the sorted list of $p$-values as $p_{(1)}, \dots, p_{(k)}$.

   4.2 Compare each ordered conformal $p$-values to a series of linearly increasing critical values $\alpha/k, 2\alpha/k, \dots, \alpha$. Specifically, $p_{(1)}$ is compared to $\alpha/k$, $p_{(2)}$ is compared to $2\alpha/k$, and so forth.

   4.3 Determine $r$ as the largest index for which $p$-value is less than its corresponding critical value, i.e., $p_{(r)} < r\alpha/k$. Select every molecule with a $p$-value no larger than $p_{(r)}$, resulting in $R = \{j : \text{The rank of } p_j \text{ is no greater than } r\}$.

Figure 1 summarizes the conformal selection procedure. The only requirement for the validity of this method is data exchangeability, which means that the likelihood of the calibration dataset and test dataset $(X_{n+1}, Y_{n+1}), \dots, (X_{n+m}, Y_{n+m}), (X'_j, Y'_j)$ for $j = 1, 2, \dots, k$, is not affected by the relative order of data points.[4,7] In other words, the dataset is equally likely to be sampled regardless of any permutation applied to the data values (e.g. swapping the first and second data point). This assumption is less stringent compared to identical independent distribution (i.i.d.), and is satisfied when the training set of molecules and the molecules to be predicted are randomly drawn from the same pool. The validity of the procedure does not require any distribution or model assumptions.

https://doi.org/10.26434/chemrxiv-2024-pf3ph ORCID: https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

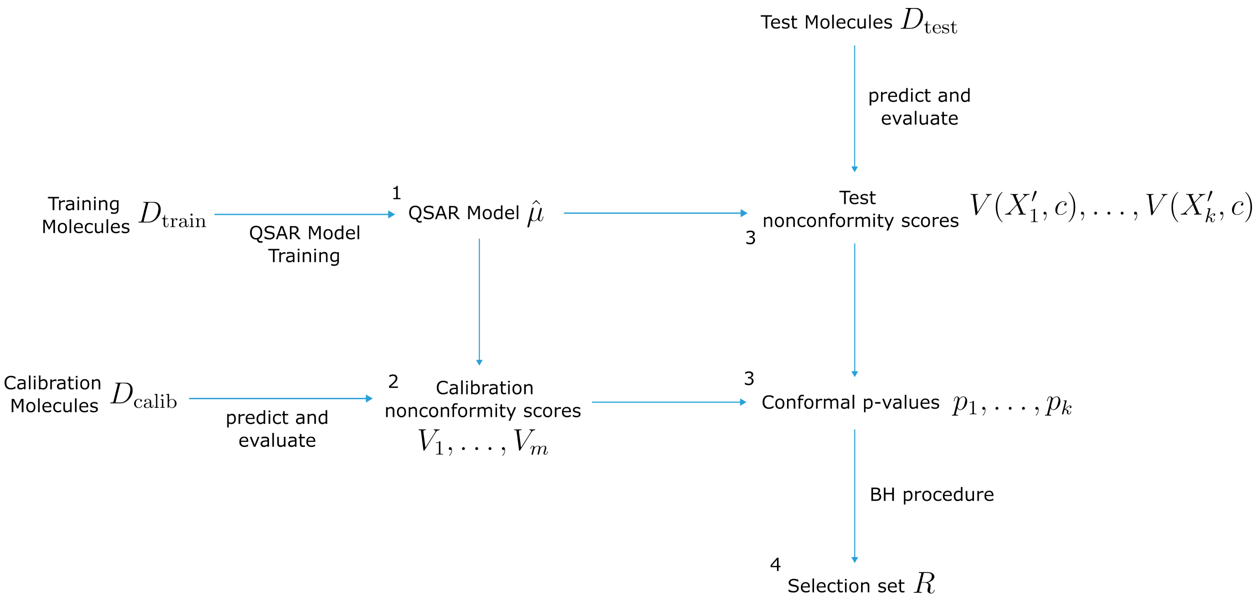
Test Molecules $D_{\text{test}}$
predict and evaluate
1
Training Molecules $D_{\text{train}}$
QSAR Model Training
QSAR Model $\hat{\mu}$
3
Test nonconformity scores $V(X_1', c), \ldots, V(X_k', c)$
Calibration Molecules $D_{\text{calib}}$
predict and evaluate
2
Calibration nonconformity scores $V_1, \ldots, V_m$
3
Conformal p-values $p_1, \ldots, p_k$
BH procedure
4
Selection set $R$

Figure 1: Scheme for the conformal selection method.

## Nonconformity Measure and Nonconformity Scores

The concept of nonconformity measure originates from the conformal prediction (CP) framework, a statistical approach designed to provide reliable predictions and quantifiable measures of uncertainty. This framework extends classical prediction methods by incorporating a formal mechanism to assess the validity of predictions in a data-driven manner. The key principle of conformal prediction is to generate prediction sets or intervals that contain the true outcome with a predefined probability, thereby offering a probabilistic guarantee of correctness.

In the CP framework, a nonconformity measure is a critical component that intuitively quantifies how “atypical” or “nonconforming” an observation is. This framework supports a variety of nonconformity measures, allowing for flexibility in selecting any metric that assesses the alignment of a model with its data. Typically, the nonconformity measure involves information provided by the prediction of QSAR models. While the use of nonconformity measures is conceptually similar in both conformal prediction and conformal selection, there are nuanced differences. We will first outline the role of nonconformity measures in conformal prediction and then relate this to their application in our conformal selection procedure.

https://doi.org/10.26434/chemrxiv-2024-pf3ph **ORCID:** https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

In general, the nonconformity measure is a real-valued function that accepts a pair of feature (structural feature) and target value (activity level) as input. The output value evaluates of how "atypical" this pair is. For example, given a well-trained model $\hat{\mu}$, the pair of feature and target $(X, Y)$ is deemed atypical when the absolute error $|Y - \hat{\mu}(X)|$ is significantly high. The goal of conformal prediction is to provide prediction intervals (PI) for the target at any feature level $X$; for an incoming data point $X$ with an unobserved corresponding target, the conformal PI at $X$ includes all possible target value $Y$ such that the pair $(X, Y)$ is not excessively atypical. The atypicality is assessed based on the nonconformity measures computed from the calibration dataset, called the calibration nonconformity scores $V_1 = V(X_{n+1}, Y_{n+1}), \ldots, V_m = V(X_{n+m}, Y_{n+m})$. For conformal prediction, the choice of the nonconformity measure is crucial to the shape and effectiveness of the resulting PI, and much effort have been devoted to design more efficient nonconformity scores to improve the predictive performance and adaptiveness of the intervals.[8–11] In conformal selection the objective differs, which changes the principle for choosing the nonconformity score. An incoming molecule is selected if its activity level $X_j'$, when paired with the activity level cutoff $c$, is atypical enough, i.e., the nonconformity score $V(X_j', c)$ is sufficiently extreme. The level of atypicality is again accessed using the calibration nonconformity scores. While selecting an appropriate nonconformity measure remains essential for achieving optimal selection performance, the selected nonconformity measure must also satisfy additional requirements, as it is used to construct conformal $p$-values.

## Conformal $p$-values

The problem of deciding whether to select a test molecule can be framed within the hypothesis testing framework. Since the alternative hypothesis typically pertains to a finding or discovery, we formulate it as

$$H_1 : Y_j' > c$$

https://doi.org/10.26434/chemrxiv-2024-pf3ph ORCID: https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

and the acceptance of the alternative hypothesis corresponds to selecting the compound as the activity level $Y_j$ exceeds the cutoff $c$. Naturally, the null hypothesis $H_0 : Y_j' \leq c$ is defined as the complement of $H_1$. Each compound under consideration represents a distinct hypothesis test.

Because our decisions are one-sided (we are only concerned with whether the activity level is exceeds the cutoff, but not how much), it is reasonable to use nonconformity scores that would produce one-sided PIs if applied in conformal prediction, such as the signed error $y - \hat{\mu}(x)$. Hypothesis testing in statistics frequently relies on $p$-values, and we use the conformal $p$-values to decide the hypothesis test described above. As formulated by Vovk et al. and Bates et al.,[5,7] the oracle conformal $p$-value is defined as:

$$p_j^* = \frac{1}{n+1}\left[\left|\{i = 1, \ldots, n : V_i < V(X_j', Y_j')\}\right| + U \cdot \left(1 + \left|\{i = 1, \ldots, n : V_i = V(X_j', Y_j')\}\right|\right)\right].$$

We note the difference between this oracle $p$-value and the conformal $p$-value presented in previous section. As the ground truths $Y_j$ are unobserved, we substitute them with the activity cutoff $c$ to obtain the following $p$-value:

$$p_j = \frac{1}{n+1}\left[\left|\{i = 1, \ldots, n : V_i < V(X_j', c)\}\right| + U \cdot \left(1 + \left|\{i = 1, \ldots, n : V_i = V(X_j', c)\}\right|\right)\right]$$

To preserve the statistical properties of the $p$-values after substitution, an additional condition must be imposed on the nonconformity scores: they must be increasing in their second argument, a property known as monotonicity.[4] Both of the suggested nonconformity scores, the signed error and the clip method, satisfy this condition.

## Controlling the FDR through the BH Procedure

The decision to select a particular molecule, or to accept a single hypothesis test as formulated above, can be made using the conformal $p$-value. For a specified level of risk $\alpha$ (representing the type-I error rate, or the probability of falsely selecting a compound), we accept $H_1$

https://doi.org/10.26434/chemrxiv-2024-pf3ph **ORCID:** https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

and select the corresponding compound only if the $p$-value is less than $\alpha$. However, in the context of multiple simultaneous hypothesis tests, the practice of selecting every compound with a $p$-value below $\alpha$ does not ensure control over the overall false discovery rate (FDR), the proportion of false selections among all actual *selections*.* To achieve FDR control, it is necessary to utilize correction methodologies, such as the Benjamini-Hochberg (BH) procedure,[6] which adjusts for multiple comparisons and regulates the proportion of false discoveries among the selected hypotheses.

The BH procedure is extensively employed in scientific studies involving the simultaneous evaluation of multiple statements or discoveries. Without applying a correction procedure, the likelihood of making a false discovery purely by chance increases as the number of hypotheses tested grows, and asserting discovery without solid evidence is clearly undesirable in formal scientific research. The BH procedure originally relies on the assumption of independence between the input $p$-values. This assumption generally holds for procedures that treat each incoming molecule in isolation. However, because each conformal $p$-value is derived from a shared set of calibration nonconformity scores, the input $p$-values are not independently distributed. Fortunately, Benjamini and Yekutieli[12] generalized the independence assumption to a less restrictive condition one known as positive regression dependency on a subset (PRDS). Under the PRDS property, the Benjamini-Hochberg (BH) procedure remains valid. Jin and Candès[4] proved that the conformal $p$-values are PRDS, thereby ensuring the integration of the CP framework while preserving valid FDR control.

We note that the false discovery rate is only one measure of the risk under the multiple hypothesis test setting. Another widely used metric is the family-wise error rate (FWER),[6] which is defined as the probability of making at least one false rejection among all the hypotheses tested. Numerically, FWER is always higher than FDR, indicating that FWER demands more stringent risk control. FWER is particularly suited to situations where even a single false selection could invalidate the entire result. However, in our setting, such strict

[*] In greater detail, such sequential approach may control the per-comparison error rate (PCER), the expected fraction of false selection among all *decisions* made.

https://doi.org/10.26434/chemrxiv-2024-pf3ph ORCID: https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

control is unnecessary and would result in a considerable loss of statistical power.

## Datasets

In this study, we use a collection of 15 Kaggle QSAR datasets. The Kaggle datasets were originally employed in the 2012 Merck & Co., Inc., Rahway, NJ, USA "Molecular Activity Challenge" Kaggle competition and released in Ma et al.[2] The datasets vary in size and pertain to diverse tasks, including predictions of on-target potency, off-target activity, and absorption, distribution, metabolism, and excretion (ADME) properties. The molecular descriptors used are the combined set of "atom pair" (AP) descriptors from Carhart et al.[13] and "donor-acceptor pair" (DP) descriptors.[14] Each dataset is partitioned into two subsets: a time-split training set and a test set. For this study, we utilize only the training sets, as evaluation of the selection methods requires access to the true activity levels.

To perform selection, each dataset must be assigned a corresponding activity cutoff. In practice, activity cutoffs are not strictly defined and often vary within a reasonable range, as different chemists or analysts may adopt slightly different thresholds. Therefore, we select a range of activity cutoffs for each dataset, ensuring that the proportion of desirable chemicals ranges from 8% to 60%. This allows us to investigate whether the variation in cutoff selection affects the performance of the methods. The number of compounds, number of structural features, activity cutoffs and percentage of desirable chemicals of the 15 datasets are summarized in Table 1. Here, we assume a counterscreening setting, where desirable compounds are defined as those with activity values lower than the specified activity cutoff. This is used to maintain consistency with the eCounterscreen method, which was originally designed for counterscreening.

https://doi.org/10.26434/chemrxiv-2024-pf3ph ORCID: https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

Table 1: Summary of dataset sizes, number of descriptors, activity cutoffs, and percentages of desirable compounds for Kaggle datasets

| **Dataset** | **Number of Compounds** | **Number of Desciptors** | **Activity Cutoff** | **Percentage of Desirable Chemicals** |
|---|---|---|---|---|
| 3A4 | 37,241 | 9,177 | 4.35 | 57.3% |
| CB1 | 8,716 | 5,555 | 6.5 | 31.7% |
| DPP4 | 6,148 | 5,025 | 6 | 34.7% |
| HIVINT | 1,815 | 4,186 | 6 | 27.4% |
| HIVPROT | 3,212 | 5,751 | 4.5 | 5.7% |
| LOGD | 37,388 | 8,623 | 1.5 | 13.3% |
| METAB | 1,569 | 4,372 | 40 | 47.3% |
| NK1 | 9,965 | 5,592 | 6.5 | 9.7% |
| OX1 | 5,351 | 4,601 | 5 | 19.7% |
| OX2 | 11,151 | 5,462 | 6 | 23.2% |
| PGP | 6,399 | 4,731 | -0.3 | 14.7% |
| PPB | 8,651 | 4,991 | 1 | 20.0% |
| RAT_F | 6,105 | 5,525 | 0.3 | 7.8% |
| TDI | 4,165 | 5,712 | 0 | 24.2% |
| THROMBIN | 5,059 | 5,282 | 6 | 36.9% |

# Results

We conducted a series of experiments to compare our purposed method with the approach introduced by Sheridan et al.[3] In each experiment, the dataset is randomly partitioned into three subsets: a training set (50% of the data), a calibration set (35%), and a test set (15%). The training set is used to train a random forest QSAR model, while the test set is employed to evaluate the selection performance of the competing methods. As discussed in previous sections, the calibration set is used to compute nonconformity scores for the conformal selection method, with the clip method serving as the nonconformity measure. The eCounterscreen method, by contrast, requires an additional split for the calibration set. We will refer to them as the calibration-1 set (20% of the whole data) and calibration-2 set (15%). The functionality of these datasets depends on the method used to estimate prediction uncertainty. One common measure of prediction uncertainty is the expected root mean square error (RMSE) of predictions, which can be approximated using different techniques.

In this comparison, we consider two competing approaches: the first, introduced by

https://doi.org/10.26434/chemrxiv-2024-pf3ph ORCID: https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

Sheridan et al. (2004),[15] estimates RMSE through cross-validation and binning; the second, from Sheridan et al. (2013),[16] employs an auxiliary error model. We refer to these two procedures as "eCounterScreen-bin"[15] and "eCounterScreen-pred"[16], respectively. For the eCounterScreen-bin method, we allocate 50% of the overall data for model training and 20% for validation. The data for training and validation are randomly selected from the training set and the calibration-1 dataset. In the eCounterScreen-pred method, the calibration-1 dataset is used exclusively to train the error model. In both approaches, the calibration-2 dataset is employed to determine an appropriate decision threshold for the $z$-score, as outlined in Sheridan et al. (2015).[3] By integrating these two estimation techniques, we establish two distinct decision procedures within the framework of eCounterscreen. The experiments are repeated for 100 times with the average performances reported. For each iteration, a random data split is performed, and all methods are evaluated on the same data partitions. Figure 2 illustrates the data splitting process.

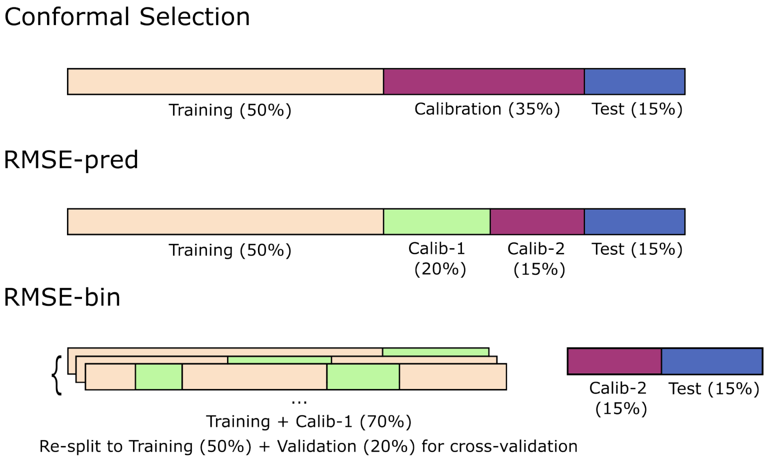

Figure 2: Illustration of the data splitting setups used for the three methods.

Figure 3(a) shows the control of risk (FDR) with varying nominal risk levels, using only a small randomly selected subset (10%) of the entire dataset. This simulates the practical scenarios where only a limited number of compounds are available for model training and calibration. The nominal risk levels vary from 10% to 50% in 5% increments. The actual observed risk levels are assessed on the test sets as the percentage of falsely selected

https://doi.org/10.26434/chemrxiv-2024-pf3ph ORCID: https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

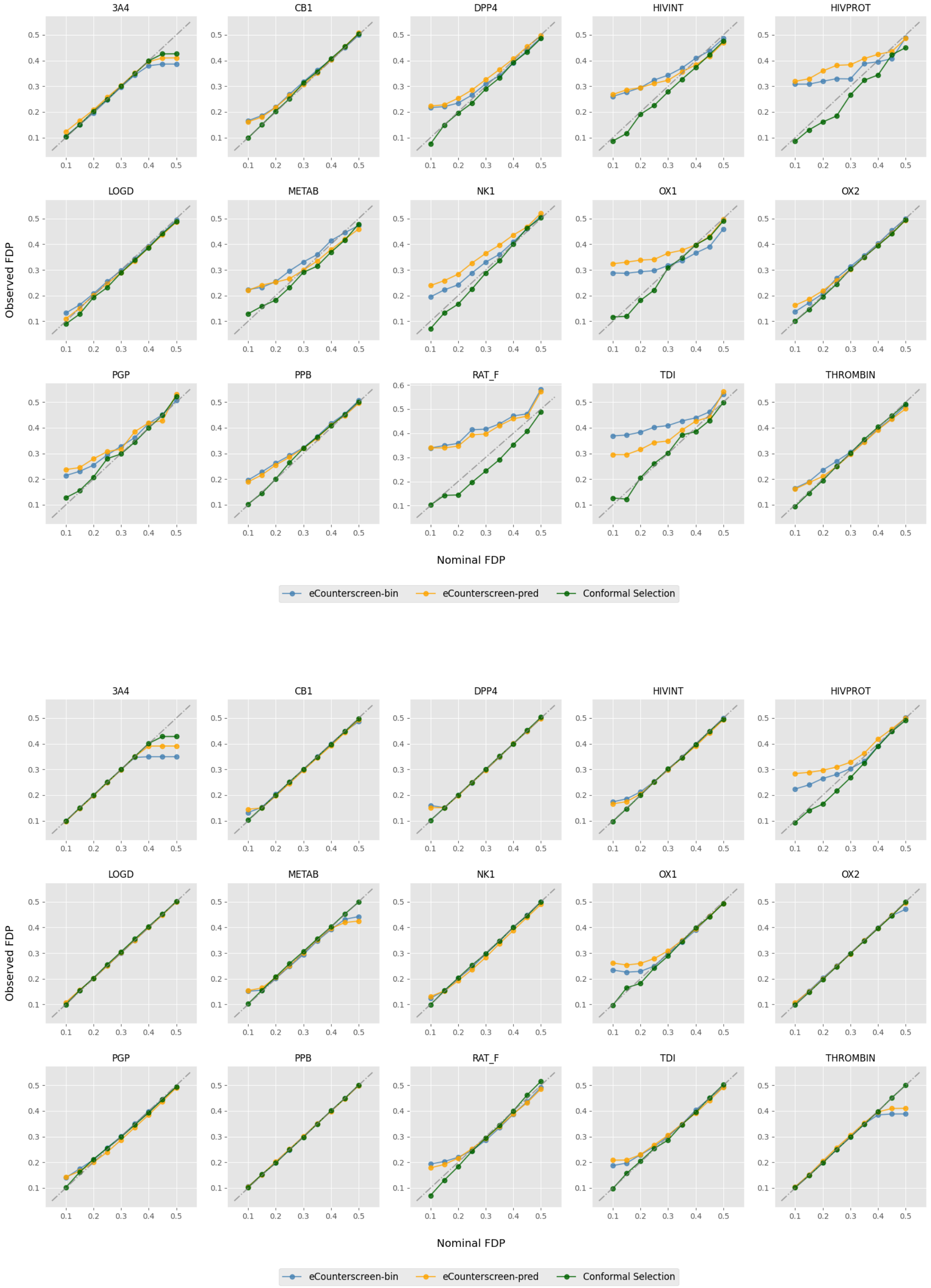

Figure 3: FDP control of conformal selection and eCounterscreen on (a) 10% subsets of the 15 Kaggle datasets, and (b) on the entirety of the datasets, with nominal risk levels varying from 10% to 50%. The grey dashed lines represent perfect risk control, where the observed risk matches the specified risk level exactly.

https://doi.org/10.26434/chemrxiv-2024-pf3ph **ORCID:** https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

molecules. As shown in the plot, the observed risks for the conformal selection method are generally lower than the nominal risk levels (dashed line). The close alignment between the observed and nominal risks demonstrates the *accurate* risk control achieved by the conformal selection method. In contrast, eCounterscreen often exhibit uncontrolled risks, particularly with smaller datasets such as HIVINT or HIVPROT, and when stringent risk thresholds, i.e. low nominal FDP values, are applied. All methods exhibit improved FDP control when the entirety of the datasets is used, as shown in Figure 3(b). Nevertheless, eCounterscreen may still fail to control the FDP under certain settings. Conformal selection provides perfect FDP control for all datasets.

Figure 4 compares the ability of different methods to identify desirable chemicals, i.e. power, using (a) 10% subset of the datasets and (b), the entire dataset. Overall, the power of the conformal selection method is at least comparable to, if not greater than, that of eCounterscreen. For certain datasets, such as OX1 and TDI, the conformal selection method demonstrates a significant performance advantage. This enhanced power may be attributed to the fact that eCounterscreen rely on prediction uncertainty for decision-making, which can be prone to bias or inaccuracies. In contrast, conformal selection inherently avoids this source of potential error, resulting in increased statistical power.

Table 2: Average runtime (in seconds) of different algorithms on 10% subsets and the entirety of the 15 Kaggle datasets

| | 10% Subsets | | | Entirety | | |
|---|---|---|---|---|---|---|
| **Dataset** | **RMSE-bin** | **RMSE-pred** | **Conformal Selection** | **RMSE-bin** | **RMSE-pred** | **Conformal Selection** |
| 3A4 | 405.04 | 6.37 | 3.45 | 69772.69 | 67.67 | 43.27 |
| CB1 | 17.42 | 1.20 | 0.62 | 2089.76 | 16.08 | 7.46 |
| DPP4 | 8.27 | 0.86 | 0.46 | 920.32 | 10.04 | 4.78 |
| HIVINT | 1.50 | 0.42 | 0.27 | 68.30 | 2.48 | 1.11 |
| HIVPROT | 3.59 | 0.60 | 0.27 | 325.55 | 6.25 | 1.91 |
| LOGD | 357.24 | 5.99 | 2.82 | 79632.30 | 70.51 | 38.97 |
| METAB | 1.40 | 0.42 | 0.26 | 66.22 | 1.86 | 0.90 |
| NK1 | 21.41 | 1.36 | 0.59 | 4507.54 | 15.99 | 6.85 |
| OX1 | 5.53 | 0.71 | 0.39 | 1032.14 | 5.76 | 3.07 |
| OX2 | 25.21 | 1.32 | 0.67 | 5518.00 | 14.30 | 8.33 |
| PGP | 5.32 | 0.50 | 0.29 | 454.84 | 5.13 | 2.91 |
| PPB | 8.68 | 0.64 | 0.39 | 889.71 | 7.78 | 4.90 |
| RAT_F | 5.15 | 0.52 | 0.28 | 471.59 | 5.41 | 3.41 |
| TDI | 2.96 | 0.42 | 0.25 | 230.02 | 3.61 | 2.12 |
| THROMBIN | 3.89 | 0.45 | 0.26 | 318.90 | 4.70 | 2.42 |

Table 2 compare the runtime of various algorithms on 10% subsets and the entirety of

https://doi.org/10.26434/chemrxiv-2024-pf3ph **ORCID:** https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

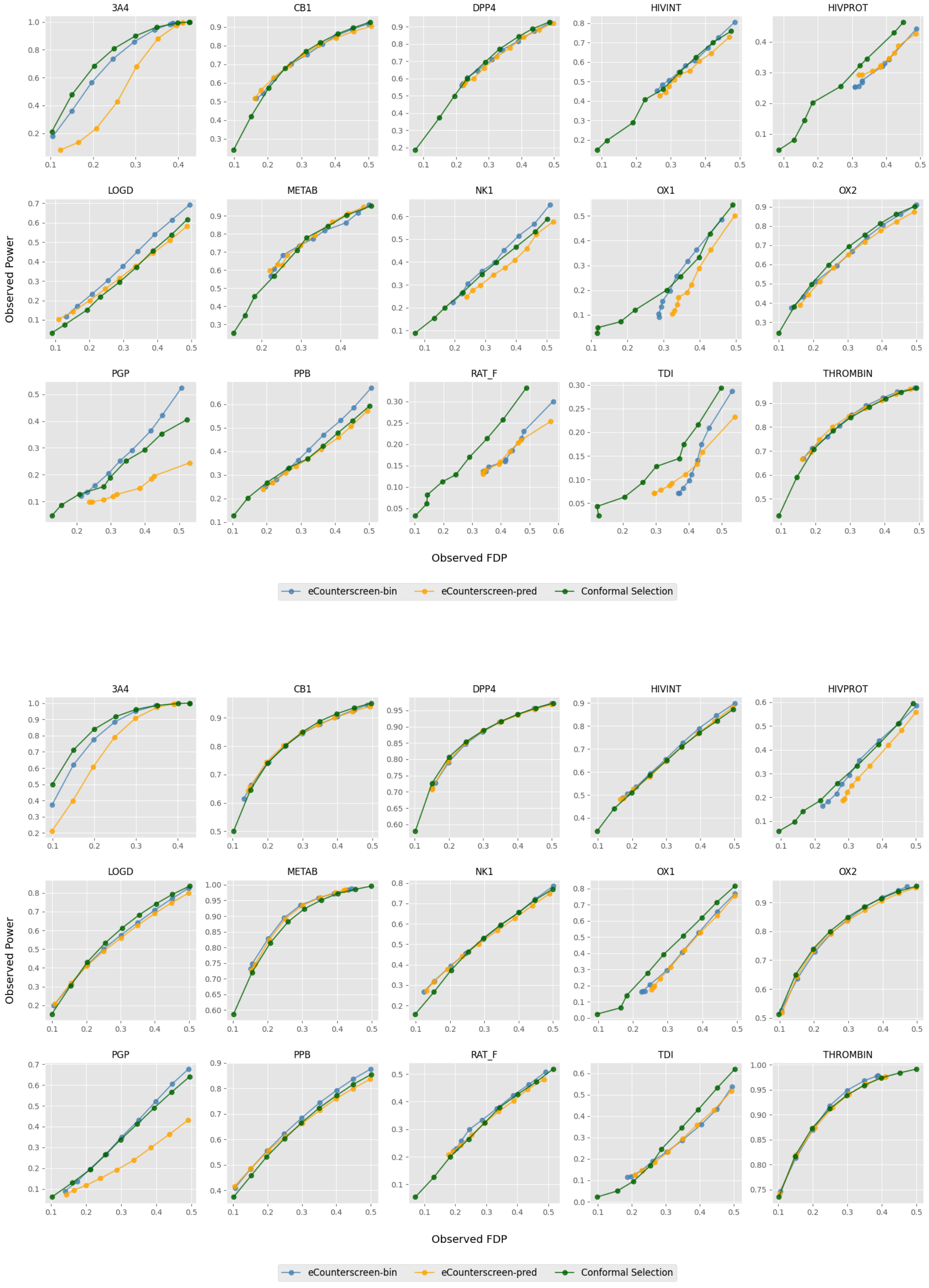

Figure 4: Power of conformal selection and eCounterscreen on (a) 10% subsets of the 15 Kaggle datasets, and (b) on the entirety of the datasets, with nominal risk levels varying from 10% to 50%.

https://doi.org/10.26434/chemrxiv-2024-pf3ph **ORCID:** https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

the 15 Kaggle datasets. For each dataset and across both data proportions, the runtime of the conformal selection method is consistently a fraction of that required by RMSE-pred and is negligible compared to the significantly more computationally intensive RMSE-bin approach. Given its linear computational complexity, conformal selection is well-suited for efficient application to large datasets such as 3A4 or LOGD. In contrast, the high computational demands of RMSE-bin may pose significant limitations or render it infeasible when computational resources are constrained. A more detailed explanation of computational complexity is provided in the Discussion section.

# Discussion

In this paper, we propose the application of conformal selection method to the drug screening and counter-screening processes, which consistently demonstrated valid risk control across all datasets, including those with limited training samples. In contrast, eCounterscreen failed to consistently achieve reliable risk control in such scenarios. This shortcoming arises primarily from the mechanism of eCounterscreen, which estimates "typical threshold level" for the risk, based on historical data. When the historical data is limited in size or of poor quality, the estimation of the threshold can become biased, compromising the method's ability to control risk. This effect is corroborated by our simulated experiments. On the other hand, the risk control provided by the conformal selection method is mathematically guaranteed regardless of the sample size.

The eCounterscreen method bases its decisions on the $z$-score, which rely on an estimate of prediction uncertainty, typically represented by the expected root mean square error (RMSE).[3] The estimation of RMSE has been widely explored by Adaptability Domain (AD) research, a subfield in QSAR.[17,18] One common approach is to use the similarity between the incoming molecule and the training molecules as a predictor.[15,19] However, this similarity-based approach has quadratic computational complexity, making it computation-

https://doi.org/10.26434/chemrxiv-2024-pf3ph ORCID: https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

ally expensive. While later works proposed the use of error models that does not rely on similarity predictors,[16] fitting these error models also incurs significant computational cost. Additionally, the search for an appropriate $z$-score decision cutoff further increases the overall computational burden. In contrast, our conformal selection method requires only three linear iterations to compute the nonconformity score, calculate the conformal $p$-value, and execute the BH procedure. Combined with enhanced statistical power and greater flexibility, these advantages suggest that conformal selection is well-suited for practical implementation in drug discovery screening and counterscreening workflows.

Building upon conformal selection, several potential extensions could be explored. One key area for future work is expanding the method to handle multiple target assays, as the current approach focuses on filtering chemicals based on a single target. While repeating the procedure for each target independently is possible, this may not be optimal in terms of statistical power. Furthermore, sequential application of conformal selection across multiple targets could invalidate the control of the overall false discovery rate (FDR), leading to a statistical issue known as the intersection hypothesis testing (IHT) problem. Addressing this would necessitate additional adjustment methods, which introduce complexity and could further diminish selection power. Thus, developing an integrated procedure capable of selecting candidate chemicals across multiple target assays simultaneously would be a valuable enhancement.

In practice, the testing molecules may not be generated in the same manner as the training and calibration molecules. For instance, chemists might prioritize certain molecular structures when selecting screening compounds,[20] potentially violating the exchangeability assumption. Fortunately, when the disparity between training, calibration and testing molecules can be captured by covariate shift, the weighted conformal selection method[21] offers an efficient solution to this problem. Thus, a natural next step would be to evaluate the performance of this weighted method in drug discovery applications.

Finally, we observed that the predictive accuracy of the QSAR model is critical to the

https://doi.org/10.26434/chemrxiv-2024-pf3ph ORCID: https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

performance of our approach. When the QSAR model demonstrates poor predictive capability, the resulting statistical power of conformal selection is typically low. However, QSAR prediction accuracy is not always perfectly correlated with selection performance. In some cases, once the prediction accuracy, as measured by out-of-sample $R^2$, reaches a certain threshold, further increasing model complexity and prediction accuracy provides minimal improvement in selection power. This insight suggests that blindly increasing QSAR model complexity may not be the optimal approach in practical applications. This observation warrants further investigation in future studies.

https://doi.org/10.26434/chemrxiv-2024-pf3ph **ORCID:** https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

# References

(1) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences* **2003**, *43*, 1947–1958.

(2) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling* **2015**, *55*, 263–274.

(3) Sheridan, R. P.; McMasters, D. R.; Voigt, J. H.; Wildey, M. J. eCounterscreening: using QSAR predictions to prioritize testing for off-target activities and setting the balance between benefit and risk. *Journal of Chemical Information and Modeling* **2015**, *55*, 231–238.

(4) Jin, Y.; Candes, E. J. Selection by Prediction with Conformal p-values. *Journal of Machine Learning Research* **2023**, *24*, 1–41.

(5) Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic learning in a random world*; Springer, 2005; Vol. 29.

(6) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **1995**, *57*, 289–300.

(7) Bates, S.; Candès, E.; Lei, L.; Romano, Y.; Sesia, M. Testing for outliers with conformal p-values. *The Annals of Statistics* **2023**, *51*.

(8) Romano, Y.; Patterson, E.; Candès, E. J. Conformalized Quantile Regression. 2019; https://arxiv.org/abs/1905.03222.

https://doi.org/10.26434/chemrxiv-2024-pf3ph **ORCID:** https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(9) Stutz, D.; Cemgil, A. T.; Doucet, A.; others Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192* **2021**,

(10) Kivaranovic, D.; Johnson, K. D.; Leeb, H. Adaptive, distribution-free prediction intervals for deep networks. International Conference on Artificial Intelligence and Statistics. 2020; pp 4346–4356.

(11) Xie, R.; Barber, R. F.; Candès, E. J. Boosted Conformal Prediction Intervals. *arXiv preprint arXiv:2406.07449* **2024**,

(12) Benjamini, Y.; Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* **2001**, 1165–1188.

(13) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences* **1985**, *25*, 64–73.

(14) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *Journal of Chemical Information and Computer Sciences* **1996**, *36*, 118–127.

(15) 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. *Journal of chemical information and computer sciences* **2004**, *44*, 1912–1928.

(16) Sheridan, R. P. Using random forest to model the domain applicability of another random forest model. *Journal of chemical information and modeling* **2013**, *53*, 2837–2850.

(17) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicability domain estimation by projection of the training set in descriptor space: a review. *Alternatives to laboratory animals* **2005**, *33*, 445–459.

https://doi.org/10.26434/chemrxiv-2024-pf3ph **ORCID:** https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(18) Weaver, S.; Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *Journal of Molecular Graphics and Modelling* **2008**, *26*, 1315–1326.

(19) Sheridan, R. P. Three useful dimensions for domain applicability in QSAR models using random forest. *Journal of chemical information and modeling* **2012**, *52*, 814–823.

(20) Polak, S.; Pugsley, M. K.; Stockbridge, N.; Garnett, C.; Wiśniowska, B. Early drug discovery prediction of proarrhythmia potential and its covariates. 2015.

(21) Jin, Y.; Candès, E. J. Model-free selective inference under covariate shift via weighted conformal p-values. *arXiv preprint arXiv:2307.09291* **2023**,

https://doi.org/10.26434/chemrxiv-2024-pf3ph **ORCID:** https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

# TOC Graphic

Some journals require a graphical entry for the Table of Contents. This should be laid out "print ready" so that the sizing of the text is correct.

Inside the `tocentry` environment, the font used is Helvetica 8 pt, as required by *Journal of the American Chemical Society*.

The surrounding frame is 9 cm by 3.5 cm, which is the maximum permitted for *Journal of the American Chemical Society* graphical table of content entries. The box will not resize if the content is too big: instead it will overflow the edge of the box.

This box and the associated title will always be printed on a separate page at the end of the document.

https://doi.org/10.26434/chemrxiv-2024-pf3ph **ORCID:** https://orcid.org/0009-0008-9112-8670 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0