Conformal Selection for Efficient and Accurate Compound Screening in Drug Discovery

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

[†]*Department of Mathematics and Statistics, McGill University, 845 Sherbrooke Street West, Montreal, Quebec, H3A 0G4, Canada*

[‡]*Pharmacokinetics, Dynamics, Metabolism, and Bioanalytical, Merck & Co., Inc., South San Francisco, California 94080, United States*

[¶]*Merck Research Laboratories, Merck & Co., Inc., 126 East Lincoln Avenue, Rahway, New Jersey 07065, United States*

[§]*Mila - Quebec AI Institute, 6666 Saint-Urbain Street, Montreal, Quebec, H2S 3H1, Canada*

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

Abstract

Reliable compound screening is fundamental to drug discovery, yet the process remains undermined by lack of robust risk controls of false compound selection or omission in current methods. 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/omission 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 power, thereby improving the ability to identify promising candidates. 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.[1,2] 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.

For instance, consider the development of a new cancer drug targeting the B-Raf V600E mutation, a common driver of melanoma. The primary "screening" goal is to identify compounds from a virtual library that strongly inhibit this mutant protein (e.g., have a predicted activity level, or IC50, below 100 nM). Similarly, a critical "counter-screening" step is to ensure these compounds do not block the hERG potassium channel, as hERG inhibition can lead to severe cardiac side effects. Given a large library of candidate molecules, a robust decision-making framework is essential to select only the most promising candidates for expensive synthesis and laboratory testing while controlling the risk of pursuing false leads.

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) models[3,4] serve this purpose by predicting biological activities based on molecular structure-derived features. Leveraging various machine learning architectures such as random forest[5] or deep neural networks,[6] QSAR models can typically achieve notable prediction accuracy when properly trained. In addition, to guide robust decision-making, single-point predictions from QSAR models are often accompanied by uncertainty

quantification techniques that offer assessment of prediction errors or a range estimate of molecular activity.

Despite significant efforts to enhance QSAR model predictions and uncertainty quantification,[7–11] there has been limited discussion how to systematically account for prediction uncertainties in 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 only approximately valid and may fail to adequately control the risk of false compound selection or omission. The state-of-the-art screening approach, eCounterscreen, developed by Sheridan et al.,[12] uses QSAR model predictions to prioritize compounds for counterscreen assays. While this method effectively manages the false selection risk in sufficiently large datasets,[12] it exhibits notable limitations in computational efficiency and its applicability to datasets of arbitrary size.

This paper introduces a unified and efficient decision procedure that provides guaranteed risk control under the minimal assumptions inherent in QSAR modeling. We adopt conformal selection,[13] a statistical methodology that offers several key advantages: first, our framework provides rigorous control of the false discovery rate (FDR) when selecting desirable molecules and can be adapted to control the false omission rate (FOR) when removing undesirable ones, while maintaining strong overall selection performance. This statistical guarantee holds regardless of dataset size, ensuring effective performance even with only a few hundred available reference molecules for model training and calibration. For example, in our numerical experiments on the 3A4 dataset,[6] when the nominal FDR and FOR levels are both set to 0.1, the realized values are 0.099 and 0.098, respectively. Simultaneously, our method achieves a 34% improvement in selective performance on the dataset, as measured by statistical power. Second, the method integrates seamlessly with any pre-trained QSAR model without requiring modification. Finally, it is highly efficient, automatically determining decision criteria from a user-specified risk tolerance level. This eliminates the computationally expensive search for thresholds required by methods like 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 that have been previously measured. We train a QSAR model $\hat{\mu}(X)$ on a training dataset $D_{\text{train}} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ to approximate Y. In addition, suppose that we also have access to a hold-out calibration dataset $D_{\text{calib}} = \{(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})\}$. In both datasets, the molecular activities are observed, which may be obtained from prior experiments.

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 molecules with high activity levels, characterized by $Y'_j > c$, where c is the activity cutoff for a specific target. In a counterscreening scenario where we want to select molecules with low activities, one can equivalently work with the transformed responses, such as negated activities $-Y'_j$ or binarized activities $\mathbf{1}\{Y'_j < c\}$.

Our goal is to obtain a subset of selected molecules $S \subseteq \{1, \ldots, k\}$ from the test dataset D_{test}, such that S contains as many molecules as possible that satisfies $Y'_j > c$, while controlling the overall selection error. The selection error is measured by the false discovery

rate (FDR), which is the expected fraction of selected molecules whose true activity levels are low, i.e. $Y'_j \leq c$. Formally, it is defined as follows:[13,14]

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

where $\mathbf{E}$ is the expectation taken over the joint distribution of the test dataset D_{test}. Intuitively, this value quantifies the percentage of unsuccessful molecular selections, and thus also quantifies the selection risk. 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 (or sensitivity), defined as the average proportion of correctly selected molecules among all those suitable for selection,[13,14] i.e.,

$$\text{Power} = \mathbf{E}\left[\frac{|S \cap \{j : Y'_j > c\}|}{|\{j : Y'_j > c\}|}\right]. \tag{2}$$

Conformal Selection for False Discovery Rate Control

Conformal selection is based on the conformal prediction (CP) method,[15–17] which is initially developed by Vovk et al.[18] It is a model-free framework for uncertainty quantification, offering probabilistic guarantees on range estimates of the target value. Jin and Candès[13] extended this framework to establish the conformal selection procedure, which is a model-free selection procedure that provides finite-sample FDR control. The terms "model-free" and "finite-sample control" signify that the effectiveness of this procedure in controlling the FDR is not dependent on the specific choice of the underlying QSAR model and the size of the dataset.

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}), $\ldots$, (X_{n+m}, Y_{n+m}), (X'_j, Y'_j) for $j = 1, 2, \ldots, k$, is not affected by the relative order of data points.[13,16,17] 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 and independently drawn from the same pool. The validity of the procedure does not require any distribution or model assumptions.

The conformal selection procedure can be summarized as follows:

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

2. We predict on D_{calib} to compare the predictions against the actual observed activities. Specifically, for each pair of molecular structure and activity (X_{n+i}, Y_{n+i}), $i = 1, 2, \ldots, m$ in the calibration dataset D_{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. For each test molecule X'_j for $j = 1, \ldots, k$ in D_{test}, compute its "test nonconformity score" $V(X'_j, c)$, where the unknown Y'_j is replaced by the activity cutoff c. Common choices for V include:

$$V(X, Y) = Y - \hat{\mu}(X) \qquad \text{or} \qquad V(X, Y) = M \cdot \mathbf{1}\{Y > c\} - \hat{\mu}(X) \tag{3}$$

   where $\hat{\mu}(X) = \widehat{Y}$ 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, called the *clip method*, as it typically yields superior selection power in practice.

3. For each incoming test molecule X'_j for $j = 1, \ldots, k$ in D_{test}, compute its "conformal p-value" defined as follows:

$$p_j = \frac{1}{m+1}\Big[\big|\{i = 1, \ldots, m : V_i \leq V(X'_j, c)\}\big| + 1\Big]. \tag{4}$$

   The p-value could be viewed as the rank of test nonconformity score $V(X'_j, c)$ among the

calibration nonconformity scores $V_1, \ldots, V_m$. In the case where the molecular activities are discrete, a randomized tie-breaking version of the conformal p-value can be used:

$$p_j = \frac{1}{m+1}\left[\left|\{i = 1, \ldots, m : V_i < V(X'_j, c)\}\right| + U \cdot \left(1 + \left|\{i = 1, \ldots, m : V_i = V(X'_j, c)\}\right|\right)\right] \tag{5}$$

where U is the realized value from an independent uniformly distributed random variable, $U \sim \text{Unif}(0, 1)$.

4. Apply the Benjamini-Hochberg (BH) procedure[14] with a *nominal level* α to the set of conformal p-values $p_1, \ldots, p_k$ obtained in step 3 to determine the selection set S. The BH procedure is a widely-acknowledged method for controlling the 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 $p_{(1)}, \ldots, p_{(k)}$.

   4.2 Compare each ordered conformal p-values to a series of linearly increasing critical values $\alpha/k, 2\alpha/k, \ldots, \alpha$. Specifically, $p_{(1)}$ is compared to α/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 in the test dataset with a p-value no larger than $p_{(r)}$, resulting in $S = \{j : p_j \leq p_{(r)}\}$.

We refer practitioners to the step-by-step procedure outlined above for implementing conformal selection as either a screening or counter-screening strategy. Figure 1 offers a visualization to the general conformal selection workflow and Figure 2 summarizes the conformal selection procedure. In the following sections, we briefly discuss the key sub-steps involved in the conformal selection process.

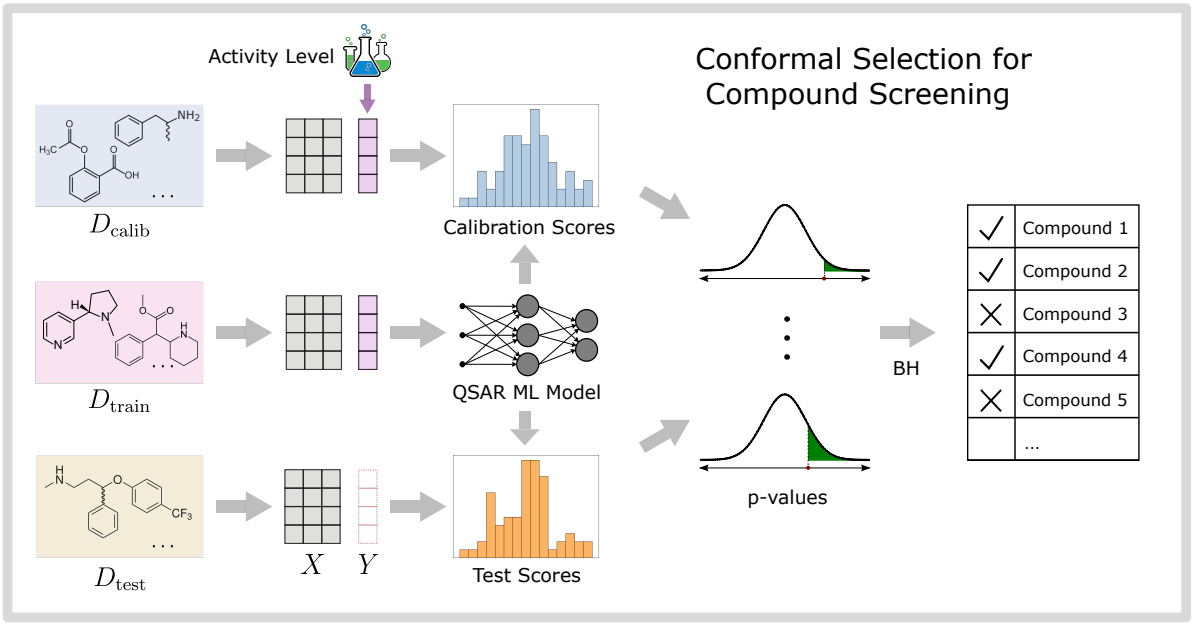


Figure 1: The conformal selection procedure. For the training data (D_{train}) and calibration data (D_{calib}), the assay activity (Y) are observed, while they remain unobserved for the testing data (D_{test}). Conformal selection uses a *pre-trained* QSAR model to compute conformal p-values by comparing calibration and test nonconformity scores. Subsequently, the Benjamini-Hochberg (BH) procedure is applied to control the selection's False Discovery Rate (FDR).

Nonconformity Measure and Nonconformity Scores

The concept of nonconformity measure originates from the conformal prediction framework. In the CP framework, a nonconformity measure is a critical component that intuitively quantifies how "atypical" or "nonconforming" an observation is.[16] 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 important 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.

In general, the nonconformity measure is a real-valued function that accepts a pair of

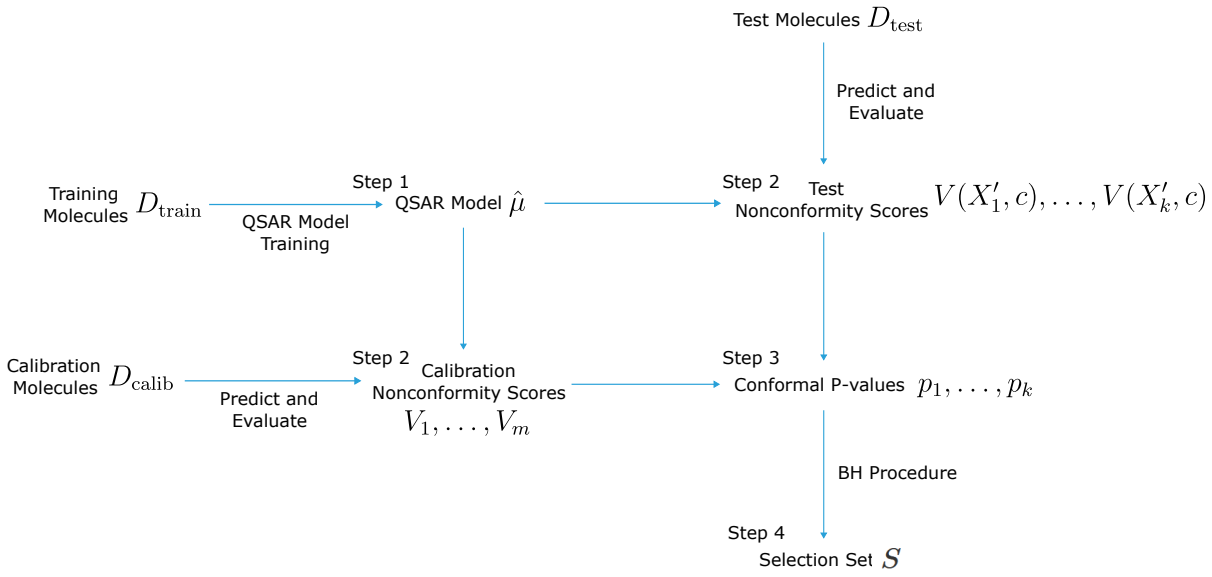


Figure 2: Scheme for the conformal selection method. Step 1: train a QSAR model using the training dataset. Step 2: compute nonconformity scores for both the calibration and test compounds based on the QSAR model's predictions. Step 3: calculate the conformal p-values for the test compounds using the nonconformity scores. Step 4: apply the BH procedure to the resulting p-values to obtain the final selection set.

feature (structural feature) and target value (activity level) as input. The output value evaluates the atypicality of this pair. 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.[19–22] In conformal selection the objective differs, which changes the principle for choosing the nonconformity score. An in-

coming molecule is selected if its structural feature 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, namely monotonicity, 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 the (random) testing problem as

$$H_0 : Y'_j \leq c \qquad \text{and} \qquad H_1 : Y'_j > c \tag{6}$$

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 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 exceeds the cutoff, but not whether it lies in an interval), it is reasonable to use nonconformity scores that would produce one-sided PIs if applied in conformal prediction, such as the signed error $y - \hat{\mu}(x)$. In this way, the generated PIs are naturally connected to hypothesis tests via conformal p-values, and we could use the conformal p-values to decide the hypothesis test described above. As formulated by Vovk et al. and Bates et al.,[18,23] the *oracle* conformal

p-value is defined as:

$$p_j^* = \frac{1}{m+1}\left[\left|\{i = 1, \ldots, m : V_i < V(X_j', Y_j')\}\right| + U \cdot \left(1 + \left|\{i = 1, \ldots, m : V_i = V(X_j', Y_j')\}\right|\right)\right]. \tag{7}$$

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:

$$p_j = \frac{1}{m+1}\left[\left|\{i = 1, \ldots, m : V_i < V(X_j', c)\}\right| + U \cdot \left(1 + \left|\{i = 1, \ldots, m : V_i = V(X_j', c)\}\right|\right)\right] \tag{8}$$

To preserve the statistical properties of the p-values after substitution, an additional condition must be imposed on the nonconformity scores $V(\cdot, \cdot)$: they must be increasing in their second argument, a property known as monotonicity.[13] 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 α (representing the type-I error rate, or the probability of falsely selecting a compound), we accept H_1 and select the corresponding compound only if the p-value is less than α. However, in the context of multiple simultaneous hypothesis tests, the practice of selecting every compound with a p-value below α does not ensure control over the overall FDR, the proportion of false selections among all actual *selections*.* To achieve FDR control, it is necessary to utilize correction methodologies, such as the BH procedure,[14] which adjusts for multiple comparisons and

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

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, the independence assumption was generalized to a less restrictive condition known as positive regression dependency on a subset (PRDS), by Benjamini and Yekutieli.[24] Under the PRDS property, the BH procedure remains valid. Jin and Candès[13] 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 FDR is only one measure of the risk under the multiple hypothesis test setting. Another widely used metric is the family-wise error rate (FWER),[14] 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 control is unnecessary and would result in a considerable loss of selection power.

In the following two sections, we explore two methodological extensions of the conformal selection framework: (1) adapting conformal selection to control the False Omission Rate (FOR), and (2) integrating FDR- and FOR-controlling sets to support dynamic decision-making processes.

Conformal Removal for False Omission Rate Control

Our proposed method offers dual error control. In addition to managing the selection error via the false discovery rate (FDR) for the selected subset S, it also provides control over the omission error for the compounds that are removed.

The problem of compound removal can be framed as a hypothesis test for each compound Y'_j in the test data, using a user-specified threshold c':

$$H_0 : Y'_j \geq c' \quad \text{and} \quad H_1 : Y'_j < c'. \tag{9}$$

The goal is to maximize the number of removed compounds that meet the criterion $Y'_j < c'$, while keeping the overall removal error in check. To achieve this, we define the false omission rate (FOR) for the set of removed compounds, R, as the expected fraction of desirable compounds among all those removed:

$$\text{FOR} = \mathbf{E}\left[\frac{|R \cap \{j : Y'_j > c'\}|}{\max(1, |R|)}\right]. \tag{10}$$

Analogous to the statistical power for selection (i.e. sensitivity), we define specificity, the statistical power for removal, as:

$$\text{Specificity} = \mathbf{E}\left[\frac{|R \cap \{j : Y'_j \leq c'\}|}{|\{j : Y'_j \leq c'\}|}\right]. \tag{11}$$

Under FOR control below a nominal level β, it is therefore desirable to maximize specificity, leading to the correct removal of more truly undesirable compounds.

We note that if the selection set S is set to be FDR-controlling, it is impossible to simultaneously control the FOR with $R = S^c$. To ensure both FDR control for the selection set S and FOR control for the removal set R, we allow S and R to be determined separately using different procedures in the following discussions.

To determine R, the conformal selection framework can be adapted to control the FOR

by using different nonconformity measures – a variant we name as conformal removal. We apply the procedure on the negated calibration data $\{(X_{n+j}, -Y_{n+j})\}_{j=1}^{m}$ and test data $\{(X'_\ell, -Y'_\ell)\}_{\ell=1}^{k}$ and the negated threshold $-c'$. Specifically, given a QSAR model $\hat{\mu}$ that predicts the activity levels Y, the nonconformity measure V can be set as

$$\begin{aligned} V(X,Y) &= -Y - (-\hat{\mu}(X)) \qquad \text{or} \qquad V(X,Y) = M \cdot \mathbf{1}\{-Y > -c'\} - (-\hat{\mu}(X)) \\ &= \hat{\mu}(X) - Y, \qquad\qquad\qquad\qquad = \hat{\mu}(X) - M \cdot \mathbf{1}\{Y < c'\}. \end{aligned} \tag{12}$$

where M is a large constant. Both nonconformity measures satisfy a modified monotonicity condition[13] under our setup. We then set R as the final selection outcome of this adapted procedure, whose FOR is controlled in finite sample.

Combining FDR- and FOR-controlling Sets

In practical drug discovery pipelines, analysts may wish to construct, in parallel, a selection subset and a removal subset to guide decision-making. Specifically, using conformal selection procedures, we can obtain both a FDR-controlling selection set S at nominal level α and a FOR-controlling removal set R at nominal level β. This leads to a set of molecules selected under the FDR criterion but not removed under the FOR criterion (defined as the green zone $Z_{\text{green}} = S \setminus R$), and a set of molecules removed under the FOR criterion but not selected under the FDR criterion (defined as the red zone $Z_{\text{red}} = R \setminus S$). These two zones therefore contain only molecules for which the two procedures are in agreement. The remaining molecules – those for which the FDR and FOR decisions conflict – constitute the grey zone $Z_{\text{grey}} = (S \cap R) \cup (S^c \cap R^c)$. For different choices of nominal levels α and β, we may observe one of the following two scenarios:

(1) $Z_{\text{green}} = S,\ Z_{\text{red}} = R,\ Z_{\text{grey}} = S^c \cap R^c,$ if $S \cap R = \emptyset$,

(2) $Z_{\text{green}} = S \setminus R,\ Z_{\text{red}} = R \setminus S,\ Z_{\text{grey}} = S \cap R,$ if $S \cap R \neq \emptyset$.

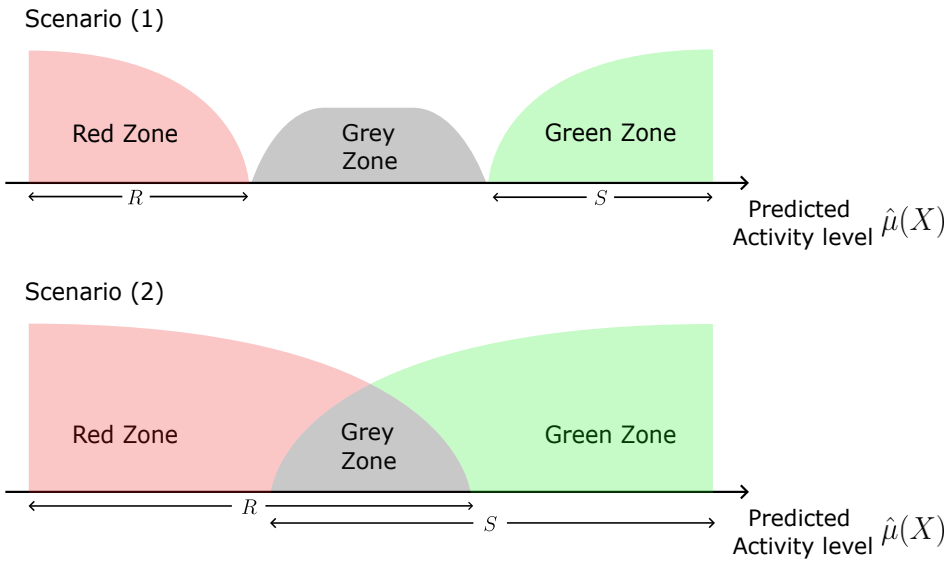


Figure 3: Illustration of the two scenarios on how the selection/removal sets interact.

Figure 3 illustrates the interaction between the selection set S and the removal set R. The horizontal axis denotes the predicted activity level $\hat{\mu}(X)$ for each molecule. Since we assume a screening setting here, S (resp. R) comprises all molecules with predicted activity above (resp. below) a certain cutoff. In a counterscreening scenario, the horizontal axis is simply reversed so that higher values of $\hat{\mu}(X)$ correspond to R. One can see that under scenario (1), molecules in Z_{green} and molecules in Z_{red} achieve exact finite-sample FDR and FOR control respectively, as they coincide with the outputs of the conformal selection procedures. Under scenario (2), although exact theoretical guarantees no longer hold, our numerical experiments, presented in the later sections, show that the two set of molecules Z_{green} and Z_{red} nonetheless maintain highly accurate, approximate control of FDR and FOR. Intuitively, this follows from the fact that Z_{green} and Z_{red} correspond to strictly more stringent selection criteria than the original sets S and R. Consequently, practitioners can customize the nominal levels α and β to control the FDR and FOR for the green and red zones.

Building on the three-zone framework, we propose a flexible screening strategy suitable for practical applications. Specifically, all compounds in the green zone are selected, those in

the red zone are discarded, and compounds in the grey zone are deferred for further, potentially more targeted, decision-making processes (e.g. wet-lab experiments to verify biological activity). This structure allows practitioners to adjust the level of conservativeness in the grey zone selection step based on specific objectives or resource constraints. Notably, the two extremes – fully conservative (selecting only the green zone) and fully liberal (including the entire grey zone) – correspond to controlling the FDR and the FOR, respectively. Intermediate strategies provide a principled trade-off, maintaining power within a satisfactory and interpretable range.

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.[6] 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.[25] and "donor-acceptor pair" (DP) descriptors.[26] Each dataset is originally provided as two subsets: one with the true activity levels and one without. For this study, we utilize only the first subset and treat it as the whole dataset, 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 in the numerical experiments, 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. Accordingly, we binarize the response variables to ensure compatibility with the conformal selection procedure.

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

FDR-controlled Selection

We conducted a series of experiments to compare our purposed method with the approach introduced by Sheridan et al.[12] 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 (performance evaluation of other types of QSAR models are included in the Supporting Materials), 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 comparison with other types of measures is included in the Supporting Materials. 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. This value is a function of unknown variables and thus cannot be analytically computed in general. Consequently, for eCounterscreen, we consider two different approaches for RMSE estimation: the first, introduced by Sheridan et al. (2004),[27] estimates RMSE through cross-validation and binning; the second, from Sheridan et al. (2013),[28] employs an auxiliary error model. We refer to these two procedures as "eCounterscreen-bin" [27] and "eCounterscreen-pred" [28], 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 combined 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).[12] By integrating these two estimation techniques, we establish two distinct decision procedures within the framework of eCounterscreen.

These procedures differ solely in the uncertainty estimation method applied to the eCounterscreen process. For each dataset, both procedures and conformal selection are executed over 100 iterations, with the average performance metrics reported. For each iteration, a random data split is performed, and all methods are evaluated on the same data partitions. Figure 4 illustrates the data splitting process.

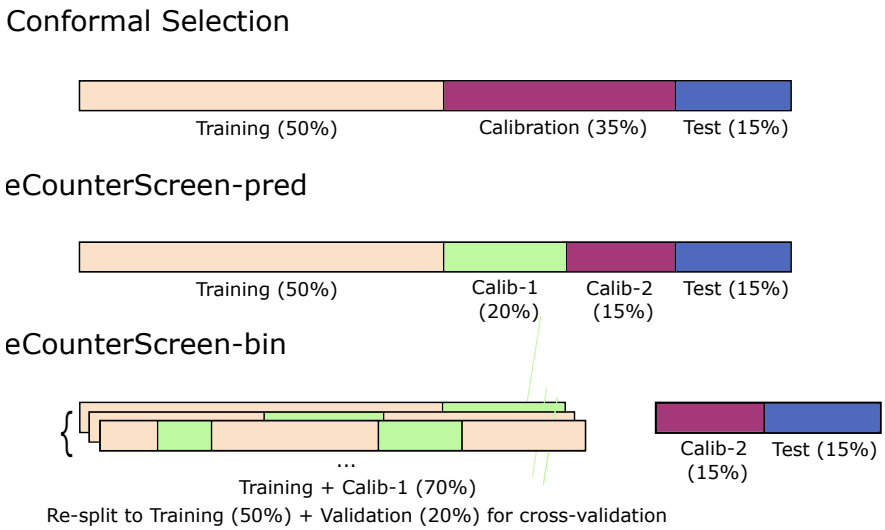
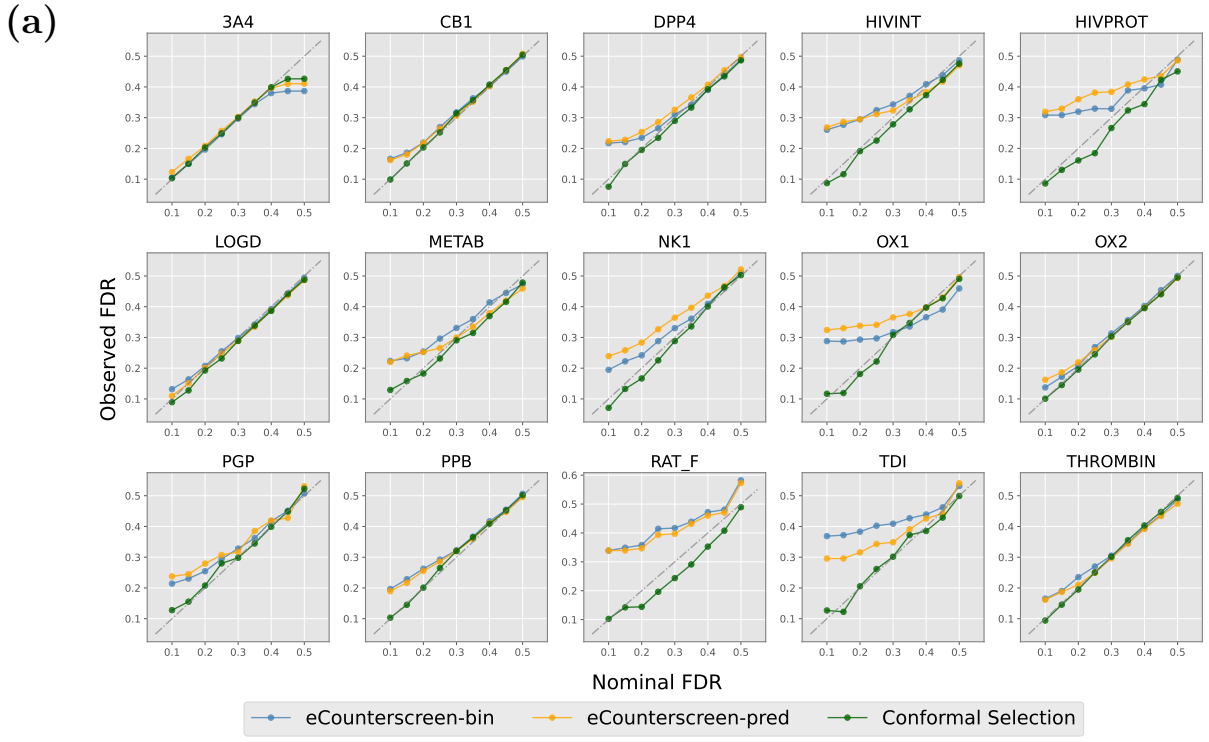


Figure 4: Illustration of the data splitting setups used for the three methods.

In Figure 5, we compare the risk control performance of various methods across datasets of differing sizes. Figure 5(a) shows the control of risk (FDR) with varying nominal FDR 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 FDR levels vary from 0.1 to 0.5 in 0.05 increments. The actual observed risk levels are assessed on the test sets as the percentage of falsely selected molecules. As shown in the plot, the observed risks for the conformal selection method are generally lower than the nominal FDR levels (dashed line). The close alignment between the observed and the nominal FDR 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 FDR values, are applied. All methods exhibit improved FDR control when the entirety of the datasets is used, as shown in Figure 5(b). Nevertheless, eCounterscreen may still fail to control the FDR under certain settings. Conformal selection provides perfect FDR control for all datasets.

We specifically note that the FDR curves for the 3A4 and THROMBIN datasets level



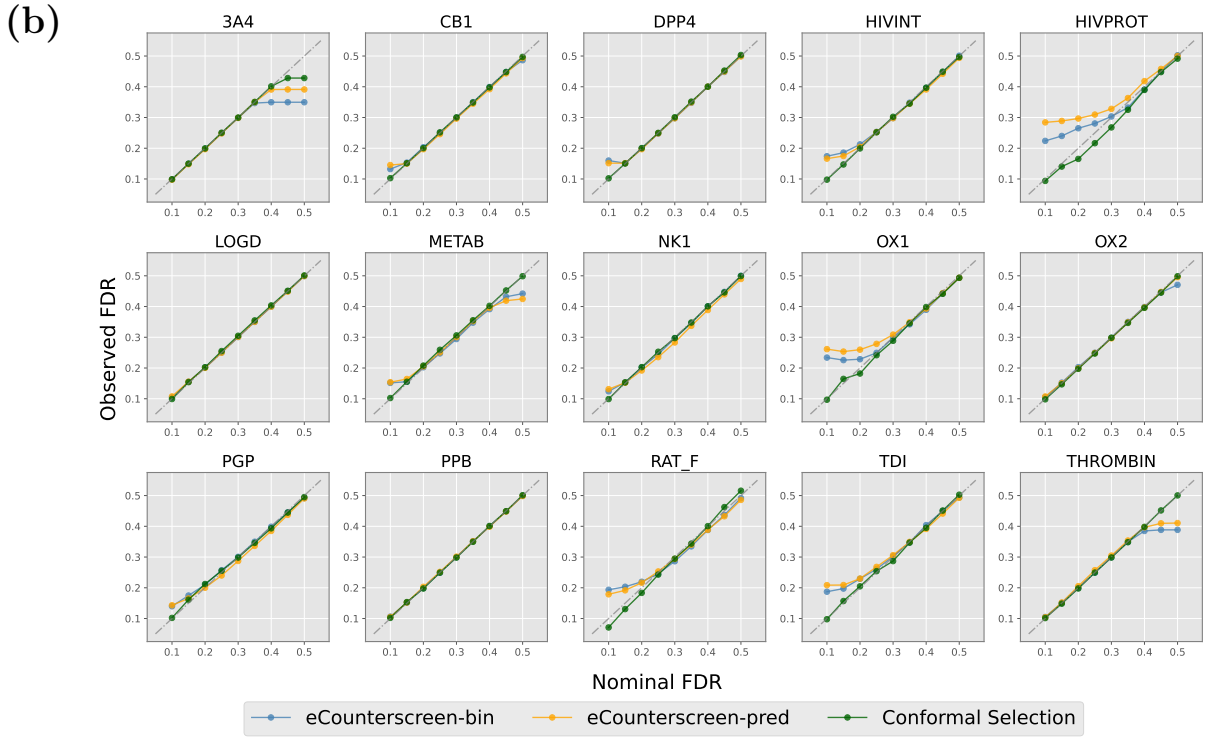


Figure 5: FDR control of conformal selection and eCounterscreen on (a) 10% subsets of the 15 Kaggle datasets, and (b) on the entirety of the datasets, with nominal FDR levels varying from 0.1 to 0.5. The grey dashed lines represent perfect risk control, where the observed FDR (y-axis) matches the specified FDR level (x-axis) exactly.

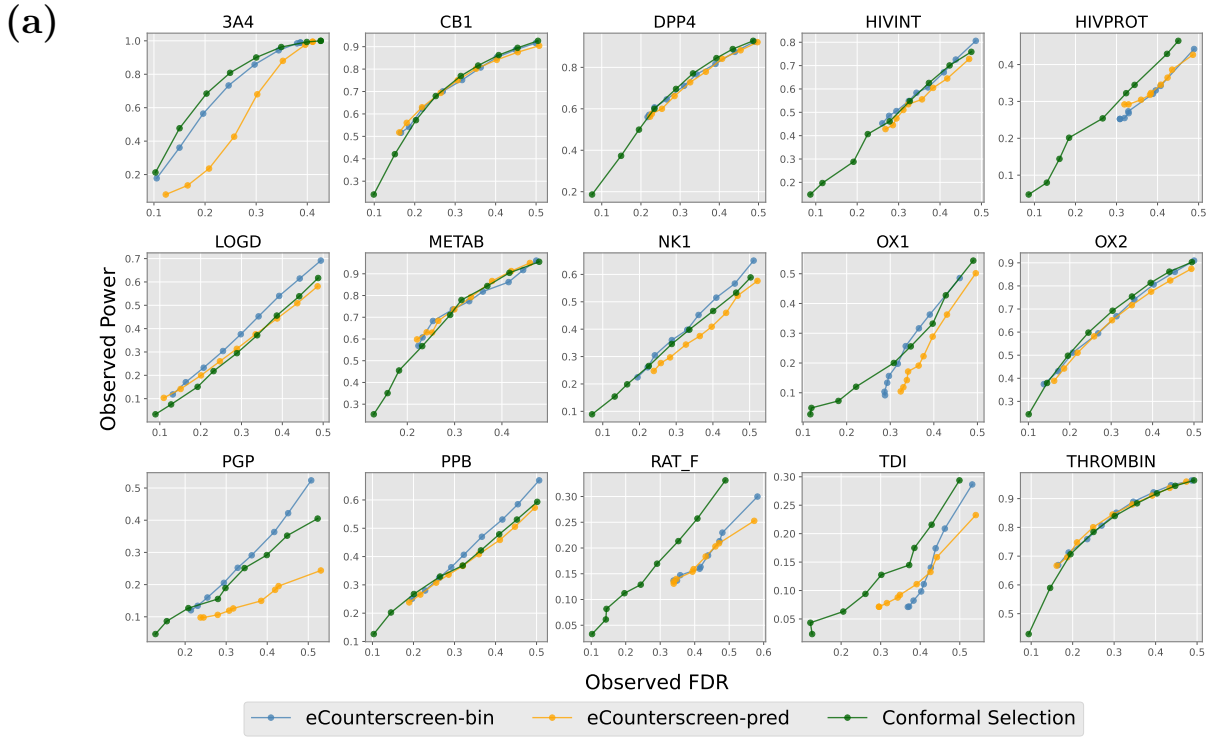
off as the nominal FDR threshold increases beyond a certain point. This occurs because, at sufficiently high nominal FDR levels, all compounds in the dataset are selected, preventing any further increase in the observed FDR.

Figure 6 compares the ability of different methods to identify desirable chemicals, (i.e. power in our context), 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. The # column denotes the total number of compounds in each dataset.

	10% Subsets				**Entirety**			
Dataset	#	**eCS-bin**	**eCS-pred**	**ConfSel**	#	**eCS-bin**	**eCS-pred**	**ConfSel**
3A4	3,724	405.04	6.37	3.45	37,241	69772.69	67.67	43.27
CB1	871	17.42	1.20	0.62	8,716	2089.76	16.08	7.46
DPP4	614	8.27	0.86	0.46	6,148	920.32	10.04	4.78
HIVINT	181	1.50	0.42	0.27	1,815	68.30	2.48	1.11
HIVPROT	321	3.59	0.60	0.27	3,212	325.55	6.25	1.91
LOGD	3,738	357.24	5.99	2.82	37,388	79632.30	70.51	38.97
METAB	156	1.40	0.42	0.26	1,569	66.22	1.86	0.90
NK1	996	21.41	1.36	0.59	9,965	4507.54	15.99	6.85
OX1	535	5.53	0.71	0.39	5,351	1032.14	5.76	3.07
OX2	1,115	25.21	1.32	0.67	11,151	5518.00	14.30	8.33
PGP	639	5.32	0.50	0.29	6,399	454.84	5.13	2.91
PPB	865	8.68	0.64	0.39	8,651	889.71	7.78	4.90
RAT_F	610	5.15	0.52	0.28	6,105	471.59	5.41	3.41
TDI	416	2.96	0.42	0.25	4,165	230.02	3.61	2.12
THROMBIN	505	3.89	0.45	0.26	5,059	318.90	4.70	2.42

Table 2 compares the runtime of various algorithms on 10% subsets and the entirety of the 15 Kaggle datasets, while Figure 7 visualizes the relative runtime trends as a function of the number of compounds on a logarithmic scale. For each dataset and across both data proportions, the runtime of the conformal selection method is consistently a fraction of that required by eCounterscreen-pred and is negligible compared to the significantly more compu-



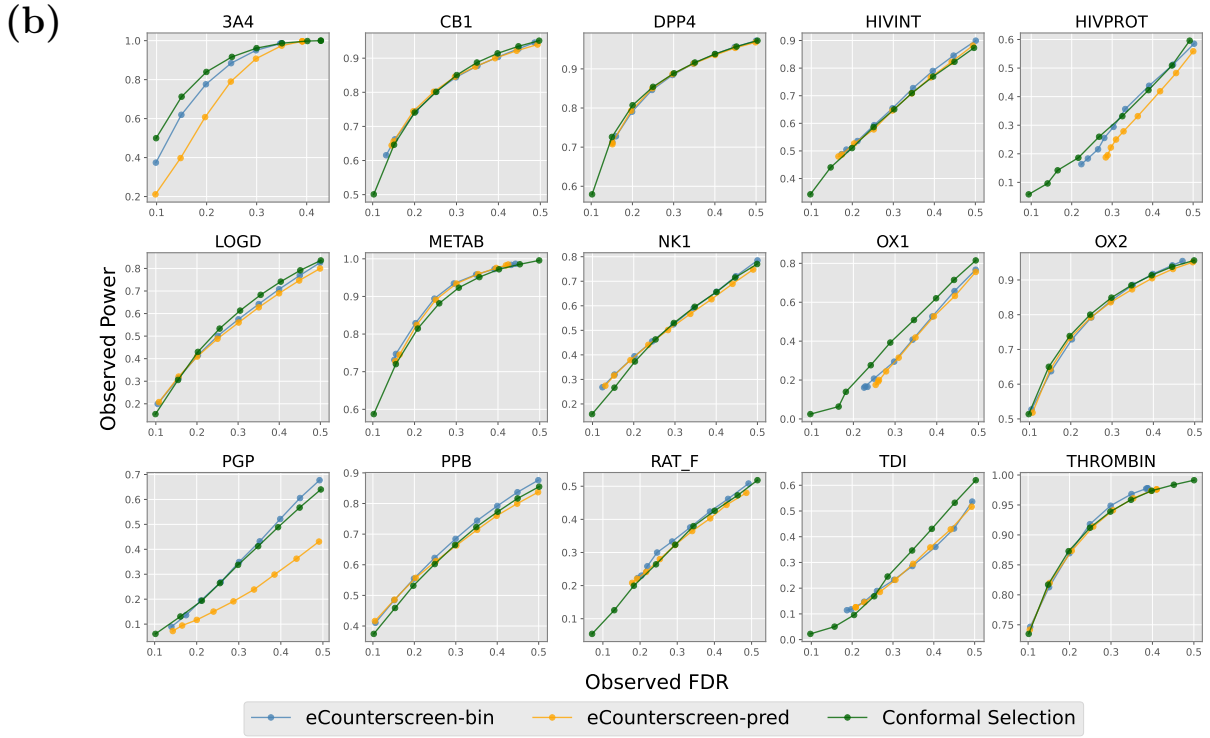


Figure 6: Power of conformal selection and eCounterscreen against the observed FDR on (a) 10% subsets of the 15 Kaggle datasets, and (b) on the entirety of the datasets, with nominal FDR levels varying from 0.1 to 0.5. The x-axis represents the observed FDR level, and the y-axis represents the observed power.

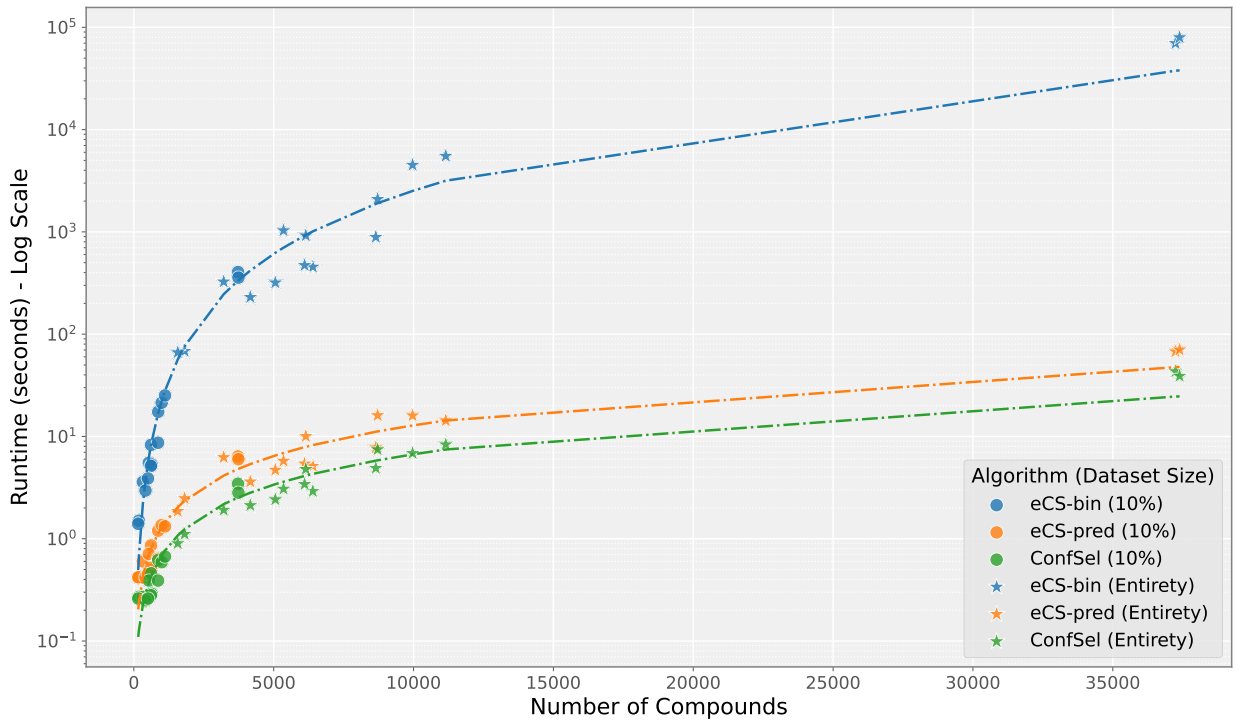


Figure 7: Scatter plot comparing the runtime and scalability of three algorithms against the number of compounds. The data points show performance on various datasets as in Table 2, and the dashed lines represent the trend for each algorithm, derived from a linear regression on the log-transformed data.

tationally intensive eCounterscreen-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 eCounterscreen-bin may pose significant limitations or render it infeasible when computational resources are constrained. This efficiency is reflected in the consistently low values and gentle slope of the approximate trend curve for conformal selection in Figure 7, underscoring the algorithm's strong scalability for large-scale applications. A more detailed explanation of computational complexity is provided in the Discussion section.

FOR-controlled Removal

Using the same datasets and experimental configurations as in the previous study, we evaluate the performance of conformal selection for FOR control in the full data setting (i.e., without down-sampling). The nominal FOR level varies from 0.02 to 0.20 in increments of 0.02. Figure 8 depicts the empirical FOR (a) and the corresponding specificity (b). The results show that conformal selection provides accurate finite-sample control of FOR while achieving satisfactory specificity. The FOR curves may level off due to an analogous reason as in the selection case.

Selection/Removal via Three Zones

Finally, we verify the claimed FDR/FOR control guarantees of the purposed three-zone scheme. With again the same datasets and configurations, we consider 18 different pairs of nominal levels where the FDR level ranges from 0.1 to 0.5 in 0.05 increments, and the FOR level is set at either 0.05 or 0.1. Figure 9 shows the observed FDR for the green zone and FOR for the red zone. We see that for every combination of nominal levels, the green zone and the red zone still maintain accurate FDR and FOR control respectively. In Figure 9, we similarly observe the FDR and FOR curves leveling off, since no further selection or omission errors can occur. In addition, the FOR curve may decline due to specific interactions between the selection and omission procedures.

Figure 10 illustrates the potential power range of the three-zone method, when all compounds in the green zone and a subset of those in the grey zone are selected. The lower and upper solid curves correspond to scenarios in which none or all grey zone compounds are selected, respectively. For any partial selection of the grey zone, the resulting power lies along a curve contained within the shaded region.

It is worth noting that the power of the standard conformal selection procedure also corresponds to a curve within this shaded region (cf. Figure 6), since the conformal selection output always includes the green zone and is a subset of the union of the green and

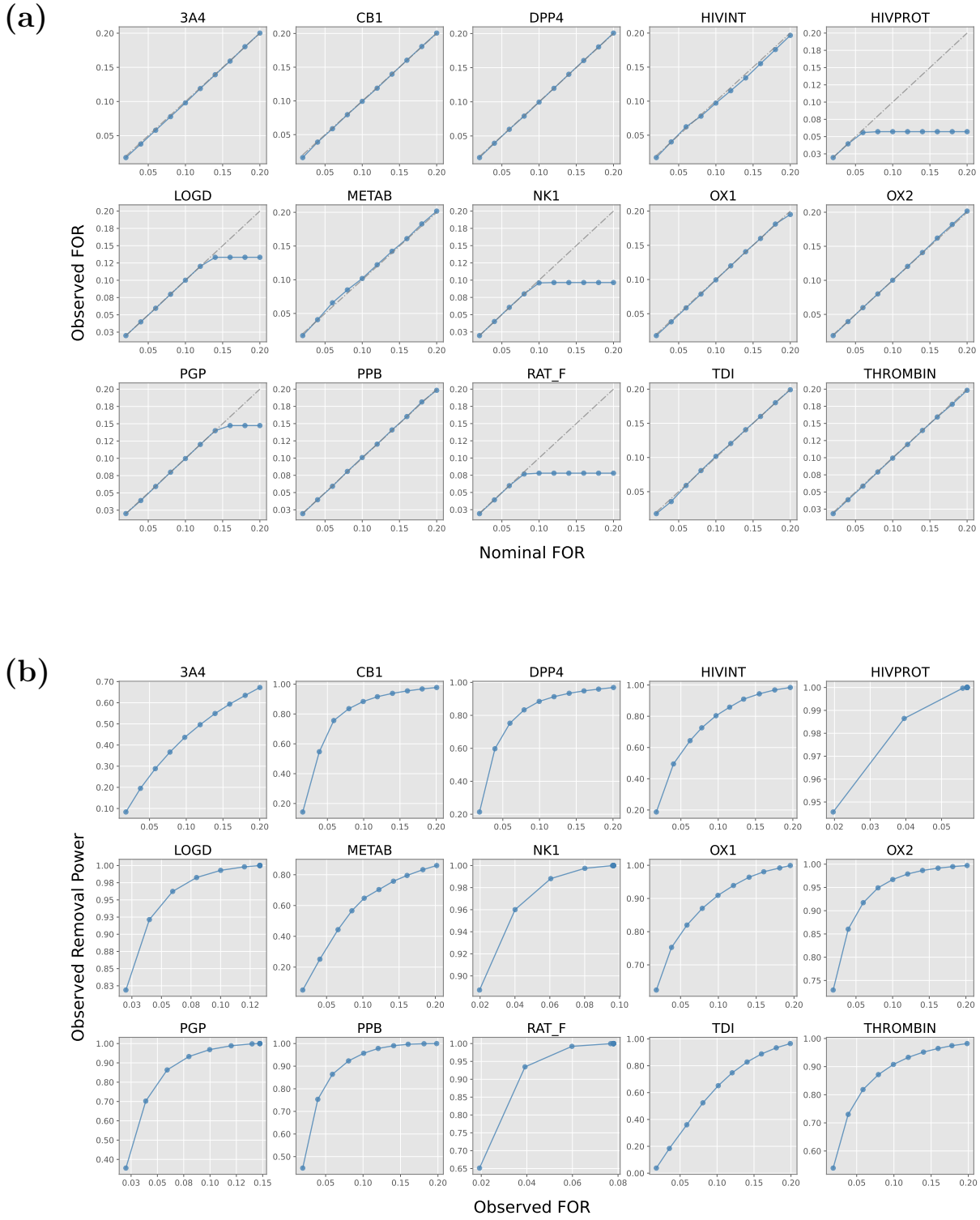


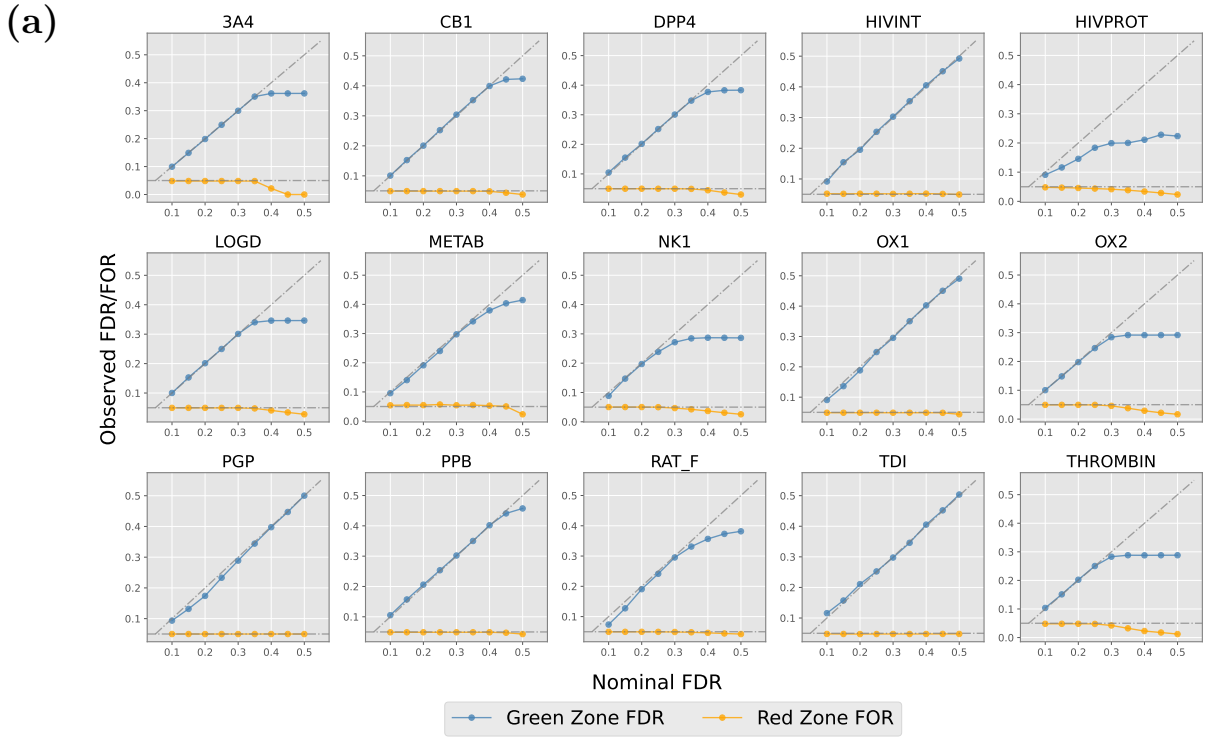
Figure 8: Observed FOR (a) and specificity (b) of conformal selection on the 15 Kaggle datasets, with nominal FOR level ranging from 0.02 to 0.2. For (a), the grey dashed line represents perfect FOR control, where the observed FOR (y-axis) matches the specified FOR level (x-axis) exactly. For (b), the x-axis represents the observed FOR, and the y-axis represents the observed specificity.

grey zones. The three-zone framework offers greater flexibility in accommodating complex practical constraints, while still achieving an overall satisfactory power profile.

Discussions

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-scores computed for each compound, which quantify the confidence associated with each selection. It relies on an estimate of prediction uncertainty, typically represented by the expected root mean square error (RMSE).[12] The estimation of RMSE has been widely explored by Adaptability Domain (AD) research, a subfield in QSAR.[29,30] One common approach is to use the similarity between the incoming molecule and the training molecules as a predictor.[27,31] However, this similarity-based approach has quadratic computational complexity, making it computationally expensive. While later works focusing on RMSE estimation proposed the use of error models that does not rely on similarity predictors,[28] 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,



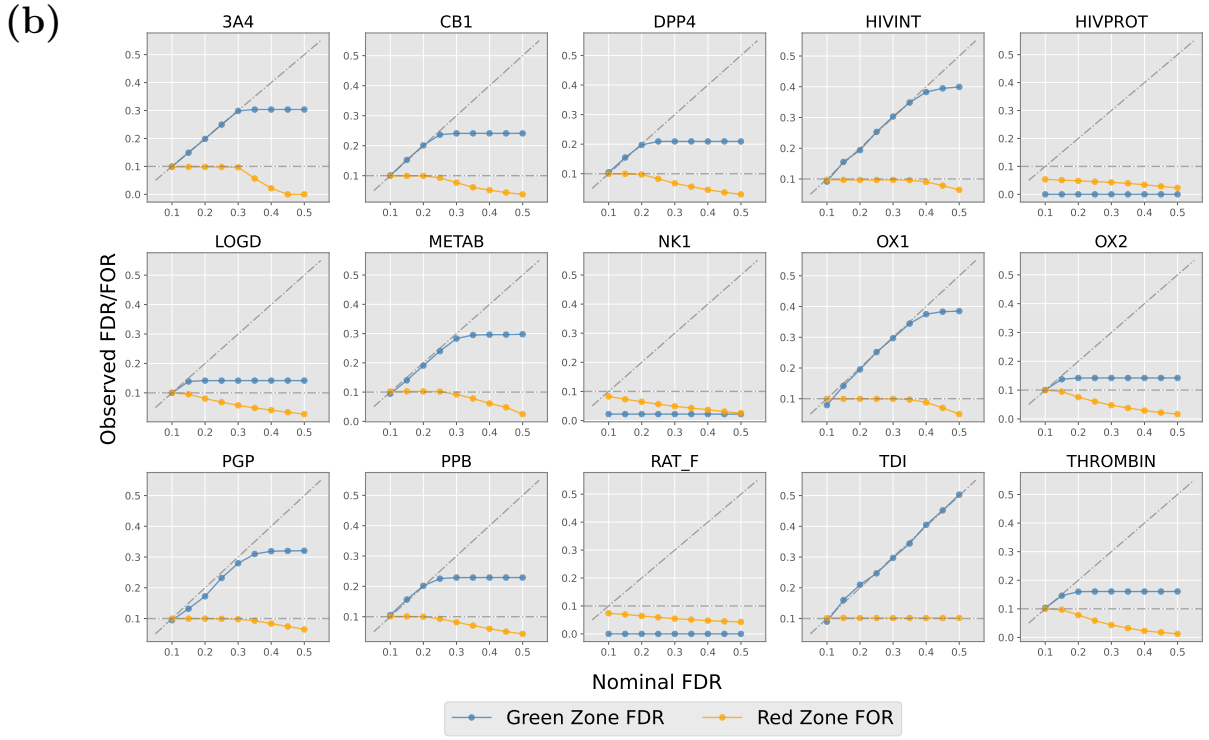
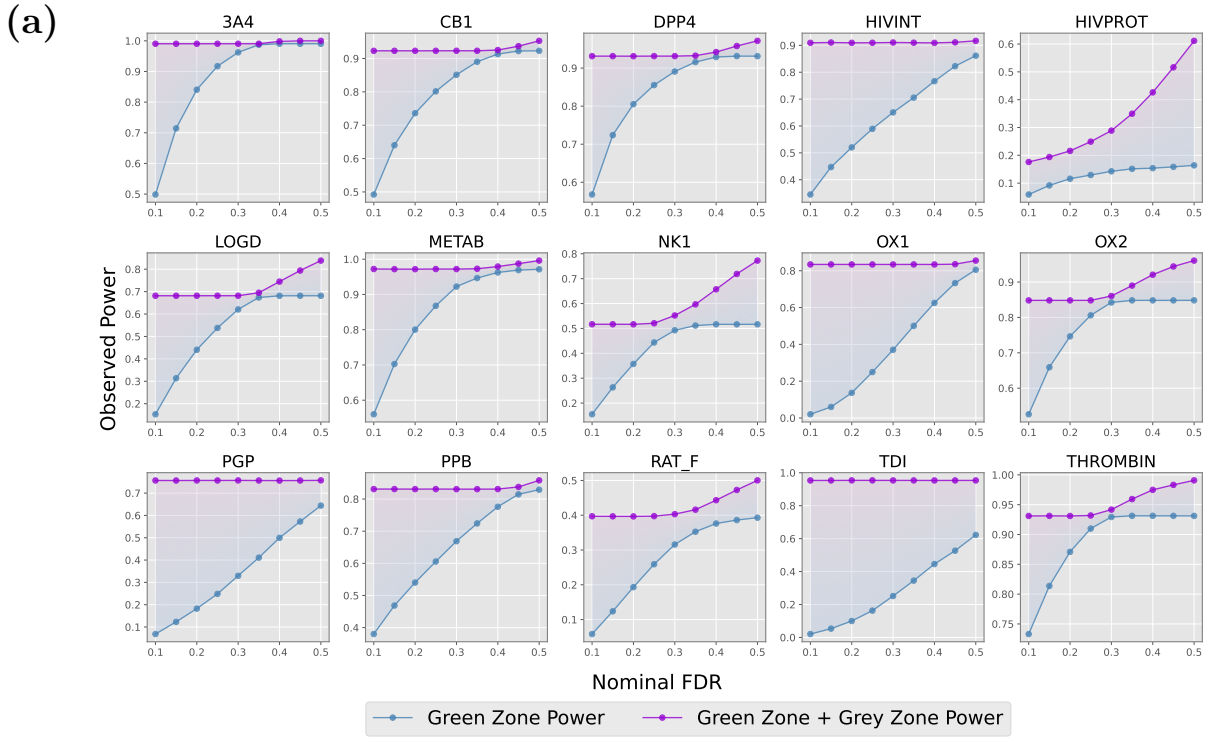


Figure 9: Observed green zone FDR and red zone FOR on the 15 Kaggle datasets, with nominal FDR level ranging from 0.1 to 0.5, and nominal FOR level set as 0.05 (a) or 0.1 (b). The grey dashed lines represent perfect FDR/FOR control.



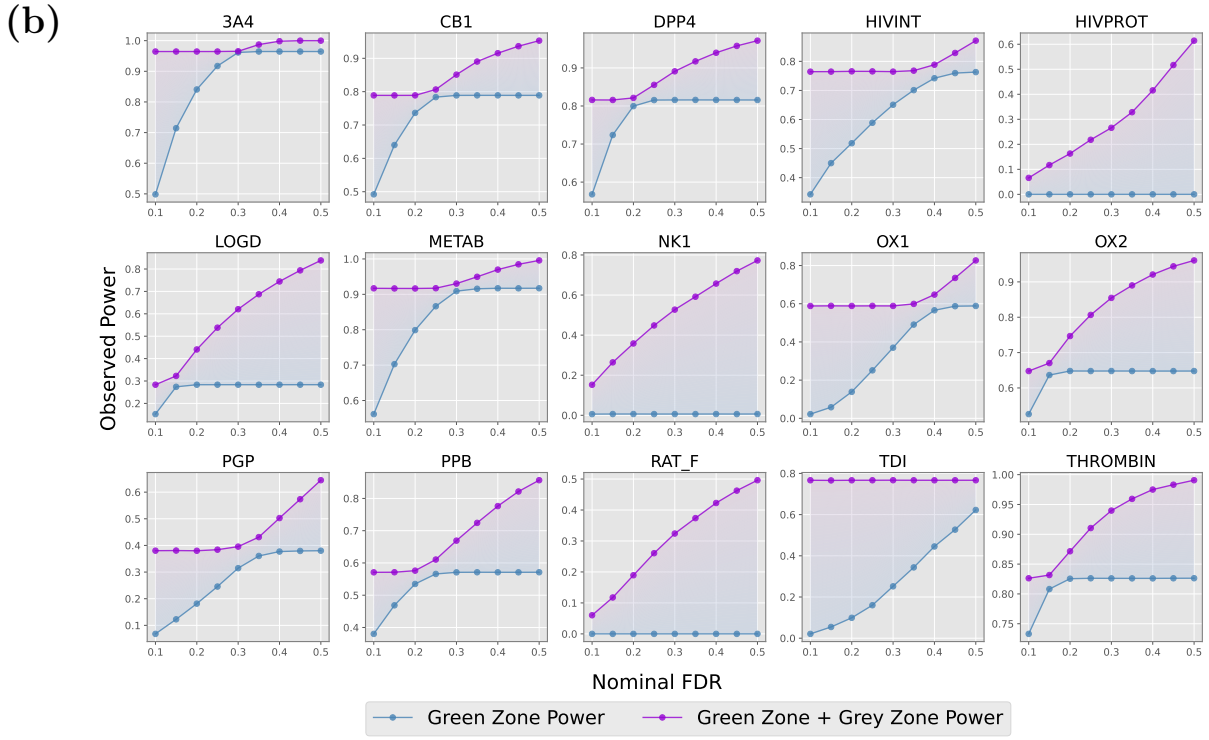


Figure 10: Observed green zone power and green-grey zone power on the 15 Kaggle datasets, with nominal FDR level ranging from 0.1 to 0.5, and nominal FOR level set as 0.05 (a) or 0.1 (b). The shaded region indicates potential power values achieved when the entire green zone and a portion of the grey zone are selected.

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.

While the prediction uncertainty estimates are not required for the conformal selection procedure, they can be incorporated by utilizing nonconformity measures that depend on prediction uncertainties. One example would be the "uncertainty-aware" clipped score $V(X, Y) = M \cdot \mathbf{1}\{Y > c\} - \hat{\mu}(x)/\hat{\sigma}(x)$, where $\hat{\sigma}$ is an estimate of the prediction uncertainty of $\hat{\mu}$. Nevertheless, our experiments (with results provided in the Supporting Materials) indicate that incorporating this term does not improve selection performance. Essentially, conformal selection without $\hat{\sigma}$ estimates already accounts for prediction uncertainty through its comparison of test scores and calibration scores, which inherently captures potential prediction errors. Including the $\hat{\sigma}$ term is therefore unnecessary and may undermine selection performance by disrupting the original ordering of the scores.

Compared to the work of Jin et al.,[13] our study extends the conformal selection framework in several important directions for practical applications in drug discovery. On the methodological front, we extend its application to FOR control and introduce a three-zone paradigm to provide greater flexibility for real-world decision making. From an applied perspective, we demonstrate that conformal selection outperforms traditional methods in identifying promising compounds. We also perform comprehensive benchmarking to assess the robustness of the method across a range of factors, including different predictive models, data preprocessing techniques, and formulations of the nonconformity score (see Supporting Materials for more details).

Limitations and Future Works

Since the current approach focuses on filtering chemicals based on a single target, one key area for future work is expanding the method to handle multiple target assays. While repeating the procedure for each target independently is possible, this may not be optimal in terms of statistical power.[32] Furthermore, sequential application of conformal selection across multiple targets could invalidate the control of the overall FDR, leading to a statistical issue known as the intersection hypothesis testing (IHT) problem.[33,34] Addressing this would necessitate additional adjustment methods,[35–38] 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.

Another limitation of the current work is that its theoretical guarantees hinge on the exchangeability assumption—an assumption that may not hold in practical molecular screening scenarios. In particular, the testing molecules may not be generated under the same distribution as the training and calibration molecules. For example, chemists often prioritize certain structural motifs when selecting compounds for screening,[39] which can introduce deviations from exchangeability. Such disparities between training, calibration, and testing data can be characterized by a covariate shift model, where the marginal distribution of structured features X changes while the conditional distribution $Y \mid X$ remains invariant, or more broadly by a distribution shift, where even the conditional relationship $Y \mid X$ differs. As a direction for future research, it would be valuable to quantify the reliability of our method under non-exchangeable conditions, building upon recent theoretical advances in conformal inference under dependence,[40] and to extend our framework to accommodate covariate shift[41,42] and general distributional shift settings.[43]

Finally, we observed that the predictive accuracy of the QSAR model plays a crucial role in determining the performance of our approach. In section B.2 of the Supporting Materials, we further examined how model prediction accuracy influences selection power. When the QSAR model exhibits poor predictive capability, the resulting statistical power of conformal

selection tends to be low. However, prediction accuracy is not always perfectly aligned with selection performance. In particular, once the predictive accuracy, as measured by the out-of-sample R^2, exceeds a certain threshold, further increases in model complexity and predictive precision yield diminishing returns in selection power. This observation suggests that, while ensuring a well-calibrated predictive model is essential, indiscriminately increasing QSAR model complexity may not be optimal in practice. The nuanced tradeoff between prediction accuracy and selection performance warrants further theoretical and empirical investigation.

Acknowledgement

This work was supported by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA; the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN2024-06780); and the FRQNT Team Research Project Grant (FRQNT 327788).

We also thank Yixuan Li for valuable assistance with the ablation study simulations presented in the Supporting Materials.

Data and Software Availability

Related Python scripts for reproducing this work are available at `https://github.com/Tian-Bai/confsel-drug`. The Kaggle QSAR datasets underlying this study are available at `https://zenodo.org/records/16388887`.

Conflict of Interest Statement

The authors declare no competing financial interests.

Author Contributions

XY, AY, and TB conceptualized the study and developed the core idea. Together with YX and VS, they designed the overall methodology and experimental framework. TB, BY, and PT conducted the literature review, implemented the code for numerical experiments, collected data, and carried out the experiments. Technical support and troubleshooting during the experimental process were provided by XY, AY, YX, VS, AK, and TB. Data collection and preprocessing was performed by PT, TB and BY. TB, AY, XY, YX, and VS interpreted the experimental results. TB was responsible for generating all figures, tables, and visualizations, and also drafted the initial version of the manuscript. All authors contributed to the manuscript's editing and revision. TB and AY prepared the final version for submission. AY, XY, and AK supervised the project. AY and XY additionally coordinated collaboration among team members, guided the overall project direction, and secured funding and other necessary resources.

Supporting Information Available

Additional details regarding the numerical experiments, and extra ablation experiments investigating different aspects of the algorithm.

References

(1) Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L. Principles of early drug discovery. *British Journal of Pharmacology* **2011**, *162*, 1239–1249.

(2) Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nature Reviews Drug Discovery* **2003**, *2*, 369–378.

(3) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of biological activity

of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* **1962**, *194*, 178–180.

(4) Cherkasov, A. et al. QSAR modeling: where have you been? Where are you going to? *Journal of Medicinal Chemistry* **2014**, *57*, 4977–5010.

(5) 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.

(6) 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.

(7) Alvarsson, J.; McShane, S. A.; Norinder, U.; Spjuth, O. Predicting with confidence: using conformal prediction in drug discovery. *Journal of Pharmaceutical Sciences* **2021**, *110*, 42–49.

(8) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. *Journal of Chemical Information and Modeling* **2014**, *54*, 1596–1603.

(9) Bosc, N.; Atkinson, F.; Felix, E.; Gaulton, A.; Hersey, A.; Leach, A. R. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *Journal of Cheminformatics* **2019**, *11*, 1–16.

(10) Arvidsson McShane, S.; Norinder, U.; Alvarsson, J.; Ahlberg, E.; Carlsson, L.; Spjuth, O. CPSign: conformal prediction for cheminformatics modeling. *Journal of Cheminformatics* **2024**, *16*, 75.

(11) 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*, 29478–29490.

(12) 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.

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

(14) 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.

(15) Shafer, G.; Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research* **2008**, *9*.

(16) Angelopoulos, A. N.; Bates, S. Conformal prediction: a gentle introduction. *Foundations and Trends® in Machine Learning* **2023**, *16*, 494–591.

(17) Angelopoulos, A. N.; Barber, R. F.; Bates, S. Theoretical foundations of conformal prediction. arXiv preprint arXiv:2411.11824, 2024.

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

(19) Romano, Y.; Patterson, E.; Candes, E. Conformalized quantile regression. Advances in Neural Information Processing Systems. 2019.

(20) Stutz, D.; Dvijotham, K. D.; Cemgil, A. T.; Doucet, A. Learning optimal conformal classifiers. International Conference on Learning Representations. 2022.

(21) 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.

(22) Xie, R.; Barber, R.; Candès, E. J. Boosted conformal prediction intervals. *Advances in Neural Information Processing Systems* **2024**, *37*, 71868–71899.

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

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

(25) 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.

(26) 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.

(27) 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.

(28) 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.

(29) 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.

(30) 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.

(31) 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.

(32) Bai, T.; Zhao, Y.; Yu, X.; Yang, A. Y. Multivariate conformal selection. International Conference on Machine Learning. 2025.

(33) Roy, S. N. On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics* **1953**, *24*, 220–238.

(34) Berger, R. L.; Hsu, J. C. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* **1996**, *11*, 283–319.

(35) Šidák, Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* **1967**, *62*, 626–633.

(36) Simes, R. J. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **1986**, *73*, 751–754.

(37) Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **1979**, 65–70.

(38) Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **1988**, *75*, 800–802.

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

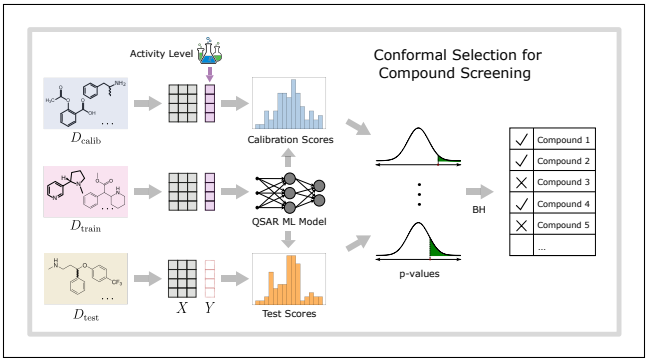
(40) Barber, R. F.; Candes, E. J.; Ramdas, A.; Tibshirani, R. J. Conformal prediction beyond exchangeability. *The Annals of Statistics* **2023**, *51*, 816–845.

(41) Tibshirani, R. J.; Foygel Barber, R.; Candes, E.; Ramdas, A. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems* **2019**, *32*.

(42) Jin, Y.; Candès, E. J. Model-free selective inference under covariate shift via weighted conformal p-values. *Biometrika* **2025**, asaf066.

(43) Xu, R.; Chen, C.; Sun, Y.; Venkitasubramaniam, P.; Xie, S. Wasserstein-regularized conformal prediction under general distribution shift. *International Conference on Learning Representations* **2025**,

TOC Graphic


Activity Level
Conformal Selection for
Compound Screening
D_{calib}
Calibration Scores
D_{train}
QSAR ML Model
BH
D_{test}
X
Y
Test Scores
p-values
Compound 1
Compound 2
Compound 3
Compound 4
Compound 5
...