KMM-CP: Practical Conformal Prediction under Covariate Shift via Selective Kernel Mean Matching

Siddhartha Laghuvarapu[1] **Rohan Deb**[1] **Jimeng Sun**[1]

[1]Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, IL, USA.,

Abstract

Uncertainty quantification is essential for deploying machine learning models in high-stakes domains such as scientific discovery and healthcare. Conformal Prediction (CP) provides finite-sample coverage guarantees under exchangeability, an assumption often violated in practice due to distribution shift. Under covariate shift, restoring validity requires importance weighting, yet accurate density-ratio estimation becomes unstable when training and test distributions exhibit limited support overlap. We propose KMM-CP, a conformal prediction framework based on Kernel Mean Matching (KMM) for covariate-shift correction. We show that KMM directly controls the bias–variance components governing conformal coverage error by minimizing RKHS moment discrepancy under explicit weight constraints, and establish asymptotic coverage guarantees under mild conditions. We then introduce a selective extension that identifies regions of reliable support overlap and restricts conformal correction to this subset, further improving stability in low-overlap regimes. Experiments on molecular property prediction benchmarks with realistic distribution shifts show that KMM-CP reduces coverage gap by over 50% compared to existing approaches. The code is available at `https://github.com/siddharthal/KMM-CP`

1 INTRODUCTION

Reliable uncertainty quantification is essential for deploying machine learning models in high-stakes domains such as scientific discovery and healthcare. Conformal Prediction (CP) Vovk et al. [2005] provides a principled framework for constructing prediction sets with finite-sample marginal coverage guarantees under minimal distributional assumptions. In standard split conformal prediction, a trained model is evaluated on a held-out *calibration set*, whose prediction residuals are used to determine a quantile threshold for forming prediction sets. These guarantees rely on exchangeability between calibration and test data—an assumption that is frequently violated in real-world settings due to distribution shift Tibshirani et al. [2019], Barber et al. [2023]. Extensions of conformal prediction to non-exchangeable or distribution-shift settings have been studied in recent work.

A common and practically relevant form of distribution shift is *covariate shift*, where the marginal distribution of inputs changes while the conditional distribution $Y \mid X$ remains stable Sugiyama et al. [2007]. This setting is especially relevant in scientific discovery, where the data-generating process is consistent under comparable experimental conditions Laghuvarapu et al. [2023], Fannjiang et al. [2022]. Under covariate shift, weighted conformal prediction restores validity by reweighting calibration samples using the density ratio between test and calibration covariate distributions Tibshirani et al. [2019]. When the ratio is known or accurately estimated, coverage guarantees are recovered.

In practice, density ratio estimation can be unstable, particularly in moderate or high-dimensional settings and under limited support overlap. Ratios may be estimated via probabilistic classification Bickel et al. [2009], direct density-ratio fitting methods such as KLIEP or LSIF Sugiyama et al. [2008], Kanamori et al. [2009], or kernel density estimation Silverman [1986b]. Classifier-based methods depend on well-calibrated probability estimates and can yield extreme weights when the test distribution assigns mass to regions poorly represented in calibration data, while kernel density approaches suffer from the curse of dimensionality. In both cases, heavy-tailed or highly concentrated importance weights reduce the effective sample size (ESS), leading to unstable quantile estimation and degraded coverage.

In this work, we propose *KMM-CP*, a framework for conformal prediction under covariate shift based on Kernel Mean

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3237–3261.

Matching (KMM) Huang et al. [2007b]. Rather than explicitly estimating density ratios, KMM aligns the weighted calibration distribution with the test covariate distribution in a reproducing kernel Hilbert space (RKHS) by minimizing the Maximum Mean Discrepancy (MMD) Gretton et al. [2012]. The resulting weights are obtained via a constrained optimization that enforces boundedness, thereby controlling variance and improving stability in high-dimensional settings. Although KMM-based reweighting has appeared in recent work Almeida et al. [2025], its connection to conformal coverage has not been systematically analyzed. In particular, prior work does not explicitly relate moment-matching quality and effective sample size to coverage behavior, nor compare KMM with alternative density-ratio estimators from a stability perspective.

We further introduce a selective extension of KMM for low-overlap or disjoint-support settings. When test regions lack calibration support, enforcing global moment matching can induce extreme weights and collapse the ESS. Our formulation jointly optimizes calibration weights and target selection variables, restricting correction to regions of shared support. This improves the bias–variance tradeoff and limits uncertainty quantification to regimes where shift correction is reliable.

Empirically, we evaluate KMM-CP on molecular property prediction tasks with substantial covariate shift. In high-dimensional settings, our method consistently reduces coverage gap and improves efficiency relative to classifier- and kernel density–based baselines, demonstrating robustness in low-overlap regimes. Our main contributions are as follows.

Main Contributions

1. We propose KMM-CP, a practical framework for conformal prediction under covariate shift that uses Kernel Mean Matching (KMM) to align calibration and test distributions via kernel mean embeddings, improving stability without explicit density estimation.
2. We introduce a selective extension for low-overlap regimes that restricts correction to regions of shared support, improving effective sample size and calibration reliability.
3. We provide analysis connecting moment-matching quality and effective sample size to conformal coverage under distributional shift, explaining the improved stability of KMM-based reweighting.
4. We evaluate KMM-CP on molecular property prediction benchmarks with substantial covariate shift, achieving over 50% reduction in coverage gap compared to existing baselines.

2 PRELIMINARIES

2.1 CONFORMAL PREDICTION FRAMEWORK

Conformal Prediction (CP) is a framework for constructing prediction sets with finite-sample coverage guarantees. We consider a classification setup where each data point is $Z = (X, Y)$ with $X \in \mathbb{R}^d$ and $Y \in [K] = \{1, \dots, K\}$. Beyond a point estimate from a base classifier f, we want a confidence level encoded as a prediction set $\hat{C}(X) \subseteq [K]$. The main goal is *valid coverage*: for a target level $\alpha \in (0, 1)$ (e.g. 0.1), the set should contain the true label with at least $1 - \alpha$ probability. Formally, for a new test point (X_{N+1}, Y_{N+1}), $\hat{C}$ is $(1 - \alpha)$-valid if

$$\mathbb{P}\big\{Y_{N+1} \in \hat{C}(X_{N+1})\big\} \geq 1 - \alpha. \tag{1}$$

Split conformal prediction. In practice, the coverage guarantee in (1) is typically attained using split (inductive) conformal prediction, which partitions the data into a training set and a calibration set D_{cal}. A base predictor is fit on the training set, and conformity scores on the calibration set are used to choose a threshold so that the resulting set-valued predictor attains the target coverage. Under exchangeability of the calibration and test points, the procedure is distribution-free and achieves the validity condition (1); see Vovk et al. [2005, Theorem 2.1].

2.2 CONFORMAL PREDICTION UNDER COVARIATE SHIFT

Standard conformal prediction assumes calibration and test data are i.i.d. (or exchangeable), an assumption that is rarely realistic in practice. In many scientific settings, however, while the marginal distribution of covariates X may shift between calibration and test time, the conditional distribution $Y \mid X$ remains stable. For example, molecular properties such as target activity or physicochemical characteristics are governed by underlying physical laws and typically remain consistent under comparable experimental conditions. This motivates the *covariate shift* setting, where the marginal law of X changes but the conditional law of $Y \mid X$ is invariant.

Formally, let $D_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^N$ denote the calibration sample and (X_{N+1}, Y_{N+1}) the test point. Under covariate shift, $P_{Y|X}^{\text{cal}} = P_{Y|X}^{\text{test}}$, so the joint laws decompose as

$$P^{\text{cal}}(dx, dy) = P_X^{\text{cal}}(dx)\, P_{Y|X}^{\text{cal}}(dy \mid x), \tag{2}$$

$$P^{\text{test}}(dx, dy) = P_X^{\text{test}}(dx)\, P_{Y|X}^{\text{cal}}(dy \mid x). \tag{3}$$

The calibration data satisfy $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P^{\text{cal}}$, while $(X_{N+1}, Y_{N+1}) \sim P^{\text{test}}$.

Importance weighting under covariate shift. Under covariate shift, one typically assumes $P_X^{\text{test}} \ll P_X^{\text{cal}}$, so the

likelihood ratio

$$w(x) := \frac{dP_X^{\text{test}}}{dP_X^{\text{cal}}}(x)$$

is well-defined. Tibshirani et al. [Tibshirani et al., 2019] view this as a special case of *weighted exchangeability*: independent (but not identically distributed) samples are weighted exchangeable with weights given by appropriate Radon–Nikodym derivatives [Tibshirani et al., 2019, Lemma 2]. In particular, if $Z_1, \ldots, Z_N \sim P^{\text{cal}}$ and $Z_{N+1} \sim P^{\text{test}}$, then one may take $w_i \equiv 1$ for $i \leq N$ and $w_{N+1}(x, y) = w(x)$. Weighted conformal prediction then forms a cutoff as a *weighted quantile* of the calibration scores, where the weight $w(x)$ of the candidate test point enters through the weighted conformal rank/quantile construction [Tibshirani et al., 2019, Lemma 3 and Theorem 2]. Using this weighted cutoff (instead of the usual unweighted cutoff) yields prediction sets with the desired marginal coverage under P^{test}.

Theorem 1 (Coverage under covariate shift [Tibshirani et al., 2019, Theorem 2])**.** *Assume the covariate shift model* (2)–(3) *and* $P_X^{test} \ll P_X^{cal}$*. Let* $w(x) = \frac{dP_X^{test}}{dP_X^{cal}}(x)$*, and construct the weighted conformal prediction set* $\hat{C}$ *using calibration nonconformity scores and the weighted-quantile rule of Tibshirani et al. [2019, Theorem 2] with weights* $w_i \equiv 1$ *for* $i \leq N$ *and* $w_{N+1}(x, y) = w(x)$*. Then*

$$\mathbb{P}_{test}\{Y_{N+1} \in \hat{C}(X_{N+1})\} \geq 1 - \alpha,$$

where the probability is over $(X_{N+1}, Y_{N+1}) \sim P^{test}$ *(and the randomness in the calibration sample).*

Remark 1. *The guarantee above is exact when* w *is known. If* w *is replaced by an estimate* $\hat{w}$ *(e.g., learned from unlabeled test covariates), finite-sample distribution-free coverage is no longer guaranteed and instead depends on the accuracy of* $\hat{w}$ *[Tibshirani et al., 2019].*

2.3 DENSITY RATIO ESTIMATION

In practice, estimating the density ratio under covariate shift is non-trivial. Let p_s and p_t denote the source (calibration) and target (test) marginal densities of X, respectively, and define the density ratio

$$w(x) = \frac{p_t(x)}{p_s(x)}.$$

A common strategy is *classifier-based density-ratio estimation*, where a probabilistic classifier distinguishes source ($D = 0$) from target ($D = 1$) samples. Let $\eta(x) = \mathbb{P}(D = 1 \mid X = x)$ and π_s, π_t denote class priors. Then

$$w(x) = \frac{\pi_s}{\pi_t} \frac{\eta(x)}{1 - \eta(x)}.$$

Another approach is *kernel density estimation* (KDE), which separately estimates p_s and p_t via

$$\hat{p}_s(x) = \frac{1}{n_s} \sum_{i=1}^{n_s} K_h(x - x_i^{(s)}), \quad \hat{p}_t(x) = \frac{1}{n_t} \sum_{i=1}^{n_t} K_h(x - x_i^{(t)}),$$

and forms $\hat{w}(x) = \hat{p}_t(x)/\hat{p}_s(x)$.

Classifier-based methods can be sensitive to probability calibration and may produce high-variance importance weights under limited support overlap [Shimodaira, 2000, Sugiyama and Kawanabe, 2012], while KDE-based approaches suffer from the curse of dimensionality in moderate or high dimensions [Wand and Jones, 1995, Silverman, 1986a]. We instead employ *Kernel Mean Matching* (KMM), which estimates importance weights by directly matching kernel mean embeddings of the weighted source and target distributions, avoiding explicit density estimation [Huang et al., 2007a].

3 KERNEL MEAN MATCHING

3.1 KERNEL MEAN MATCHING UNDER COVARIATE SHIFT

Let $\{x_i\}_{i=1}^n \sim P_S$ and $\{z_j\}_{j=1}^m \sim P_T$ denote source and target samples, related by a covariate shift. Kernel Mean Matching (KMM) [Huang et al., 2006] seeks to correct this shift by constructing nonnegative weights $w_1, \ldots, w_n$ such that the reweighted empirical source measure

$$P_S^w := \frac{1}{n} \sum_{i=1}^{n} w_i \, \delta_{x_i} \tag{4}$$

approximates (after normalization) P_T. Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space (RKHS) with a characteristic kernel k and feature map φ. For probability measures Q and R on $\mathbb{R}^d$, define the associated MMD by

$$\text{MMD}(Q, R) := \sup_{\|h\|_{\mathcal{H}} \leq 1} \left| \mathbb{E}_{X \sim Q}[h(X)] - \mathbb{E}_{X \sim R}[h(X)] \right|.$$

KMM solves the empirical Maximum Mean Discrepancy (MMD) minimization problem:

$$\min_w \left\| \frac{1}{n} \sum_{i=1}^{n} w_i \varphi(x_i) - \frac{1}{m} \sum_{j=1}^{m} \varphi(z_j) \right\|_{\mathcal{H}}^2 \tag{5}$$

subject to the following for some $\epsilon < 1$

$$0 \leq w_i \leq B, \qquad \left| \frac{1}{n} \sum_{i=1}^{n} w_i - 1 \right| \leq \epsilon. \tag{6}$$

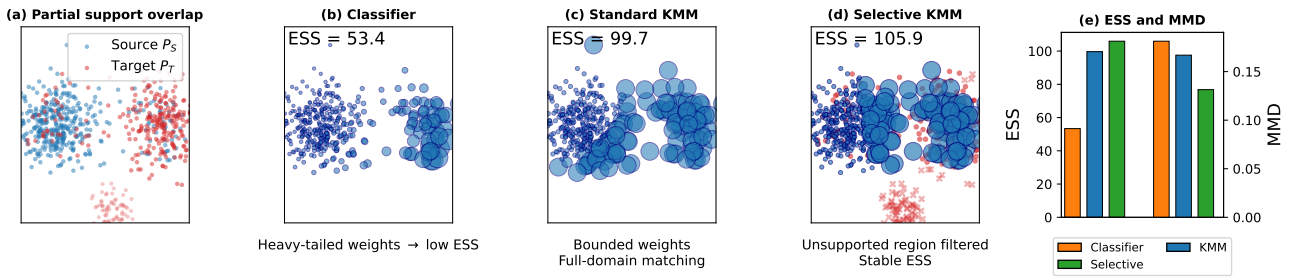


Figure 1: Synthetic experiment: (a) Covariate shift with partial support overlap, the target includes regions poorly supported by the source. (b) Classifier-based density-ratio assigns heavy-tailed weights to low-support regions, reducing effective sample size (ESS). (c) Standard KMM enforces bounded weights and matches moments over the full domain, improving stability but still correcting unsupported regions. (d) Selective KMM jointly optimizes source weights and target selection variables, filtering unsupported regions and stabilizing weights. (e) Classifier weighting reduces ESS, KMM lowers moment discrepancy (MMD), and selective KMM achieves both stable ESS and improved moment matching.

3.2 COVERAGE GUARANTEES UNDER KMM REWEIGHTING

Let $s : \mathbb{R}^d \times [K] \to \mathbb{R}$ be a conformity score. For $t \in \mathbb{R}$ define the conditional score CDF

$$g_t(x) := \mathbb{P}\big(s(x, Y) \le t \,\big|\, X = x\big),$$

where $(X, Y) \sim P_S$ (equivalently $(X, Y) \sim P_T$) since $P^S_{Y|X} = P^T_{Y|X}$. The target (population) CDF of the score is

$$F_T(t) := \mathbb{P}_{(X,Y)\sim P_T}\big(s(X, Y) \le t\big) = \mathbb{E}_{X\sim P^T_X}\big[g_t(X)\big].$$

Given weights w_i from KMM, define $\tilde{w}_i := \frac{w_i}{\sum_{j=1}^n w_j}$, $\sum_{i=1}^n \tilde{w}_i = 1$, and the corresponding weighted empirical covariate measure $\widehat{P}^w_{S,X} := \sum_{i=1}^n \tilde{w}_i\, \delta_{x_i}$. This weighted empirical covariate measure induces a reweighted (population) score CDF

$$F_w(t) := \mathbb{E}_{X\sim \widehat{P}^w_{S,X}}\big[g_t(X)\big] = \sum_{i=1}^n \tilde{w}_i\, g_t(x_i).$$

The first step toward coverage analysis is to control the population discrepancy $\sup_t |F_w(t) - F_T(t)|$ by the MMD discrepancy between the covariate measures.

Theorem 2 (**CDF stability under MMD control**). *Assume that for each $t \in \mathbb{R}$ there exists $\tilde{g}_t \in \mathcal{H}$ such that*

$$\sup_{t\in\mathbb{R}} \|\tilde{g}_t\|_{\mathcal{H}} \le B_{\mathcal{H}}, \qquad \sup_{t\in\mathbb{R}} \|g_t - \tilde{g}_t\|_\infty \le \varepsilon_\infty.$$

Then, for any probability measure Q on $\mathbb{R}^d$,

$$\sup_{t\in\mathbb{R}} \Big|\mathbb{E}_{X\sim Q}[g_t(X)] - \mathbb{E}_{X\sim P^T_X}[g_t(X)]\Big| \le B_{\mathcal{H}}\, \mathrm{MMD}(Q, P^T_X) + 2\,\varepsilon_\infty.$$

In particular, with $Q = \widehat{P}^w_{S,X}$ we obtain

$$\sup_{t\in\mathbb{R}} |F_w(t) - F_T(t)| \le B_{\mathcal{H}}\, \mathrm{MMD}(\widehat{P}^w_{S,X}, P^T_X) + 2\,\varepsilon_\infty.$$

Using this we can show the following.

Theorem 3 (**Coverage error for KMM-weighted split conformal**). *Let $(X_1, Y_1), \dots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P_S$ be a calibration sample, let $Z_1, \dots, Z_m \overset{\text{i.i.d.}}{\sim} P^T_X$ denote unlabeled target covariates used to compute the KMM weights, and let $(X_{n+1}, Y_{n+1}) \sim P_T$ be an independent test point. Define $\widehat{F}_w(t) := \sum_{i=1}^n \tilde{w}_i\, \mathbf{1}\{s(X_i, Y_i) \le t\}$, and $\mathrm{ESS} := 1/\sum_{i=1}^n \tilde{w}_i^2$. Let the weighted $(1-\alpha)$-quantile be $\widehat{q}_{1-\alpha} := \inf\big\{t \in \mathbb{R} : \widehat{F}_w(t) \ge 1-\alpha\big\}$, and define the set-valued predictor*

$$\widehat{C}(x) := \big\{y \in [K] : s(x, y) \le \widehat{q}_{1-\alpha}\big\}.$$

Assume that with probability 1 the calibration scores $s(X_1, Y_1), \dots, s(X_n, Y_n)$ are all distinct and that the conditions of Theorem 2 hold with constants $B_{\mathcal{H}}$ and ε_∞. Then for any $\delta \in (0, 1)$, conditional on $X_1, \dots, X_n, Z_1, \dots, Z_m$, with probability at least $1 - \delta$ over $Y_1, \dots, Y_n$ we have

$$\Big|\mathbb{P}_{(X,Y)\sim P_T}\big(Y \in \widehat{C}(X)\big) - (1-\alpha)\Big| \le B_{\mathcal{H}}\, \mathrm{MMD}_{\mathcal{H}}(\widehat{P}^w_{S,X}, P^T_X) + 2\,\varepsilon_\infty + \left(5 + \sqrt{\tfrac{1}{2}\log\tfrac{1}{\delta}}\right)\sqrt{\tfrac{1}{\mathrm{ESS}}}.$$

Corollary 1 (Asymptotic validity of KMM-weighted split conformal). *Assume the conditions of Theorem 3. Allow the approximation error in Theorem 2 to depend on (n, m), and write it as $\varepsilon_{\infty,n,m}$. If, as $n, m \to \infty$,*

$$\mathrm{MMD}_{\mathcal{H}}\big(\widehat{P}^w_{S,X}, P^T_X\big) \to 0, \mathrm{ESS} \to \infty, \varepsilon_{\infty,n,m} \to 0,$$

in probability, then the following holds in probability

$$\mathbb{P}_{(X,Y)\sim P_T}\big(Y \in \widehat{C}(X)\big) \to 1 - \alpha$$

Theorem 3 reveals a clear bias–variance decomposition: the MMD term controls systematic distributional mismatch between the reweighted calibration distribution and the target distribution, while the $1/\sqrt{\mathrm{ESS}}$ term captures variance

induced by weight concentration. Accurate covariate-shift correction therefore requires both small moment discrepancy and stable (non-degenerate) weights.

Kernel Mean Matching is well-suited to this objective. Unlike classifier- or KDE-based density-ratio estimation, which can yield heavy-tailed or highly concentrated weights under limited support overlap, KMM minimizes an RKHS discrepancy under explicit box and mass constraints. These constraints control weight variance, preserve effective sample size, and stabilize quantile estimation. As a result, KMM directly targets the two terms governing coverage in our bound—moment alignment and ESS—without explicit density estimation. We further analyze these stability considerations in relation to other density estimation procedures in Appendix C and empirically validate them.

3.3 SELECTIVE KERNEL MEAN MATCHING VIA JOINT TARGET SELECTION

When the target distribution P_T is not well-supported by the source distribution P_S, the standard density ratio diverges. Forcing the KMM optimizer to match moments over the entire target domain requires extreme source weights, which collapses the Effective Sample Size (ESS) and inflates the variance of our coverage bounds (Theorem 3).

To address this, we incorporate a selective mechanism that restricts conformal correction to regions of shared support. Specifically, we introduce test-instance selection weights $\alpha_j \in [0,1]$ applied to the target samples z_j, alongside the standard source weights $w_i \in [0,B]$ applied to x_i. We jointly optimize the source weights and the target selection variables to match the selected target moments. This formulation is related to double-weighting strategies for covariate shift adaptation Segovia-Martín et al. [2023], though our objective and constraints differ and are tailored to conformal calibration rather than risk minimization.

$$\min_{w,\alpha} \left\| \frac{1}{n}\sum_{i=1}^{n} w_i \varphi(x_i) - \frac{1}{m}\sum_{j=1}^{m} \alpha_j \varphi(z_j) \right\|_{\mathcal{H}}^2 \tag{7}$$

subject to the constraints:

$$\begin{aligned} &0 \le w_i \le B, \qquad \left| \frac{1}{n}\sum_{i=1}^{n} w_i - \frac{1}{m}\sum_{j=1}^{m} \alpha_j \right| \le \epsilon, \\ &0 \le \alpha_j \le 1, \qquad \frac{1}{m}\sum_{j=1}^{m} \alpha_j \ge \tau. \end{aligned} \tag{8}$$

Here, $\tau \in (0,1)$ enforces a acceptance yield (the fraction of target points we wish to make predictions for). The KKT conditions suggest that when gradients clearly favor inclusion or exclusion, many selection variables α_j may approach the boundaries 0 or 1 (see Appendix B). Empirically, we observe such sparsity: the learned α_j often concentrate near the extremes, particularly for target points poorly supported by the source distribution. This induced sparsity directly controls the two terms in the coverage error bound:

1. **Bias Reduction (MMD Control):** By restricting inference to the selected subset, moment matching is concentrated on regions of shared support. In these regions, bounded source weights suffice to approximate the selected target moments, reducing the effective MMD between P_S^w and induced target distribution $\tilde{P}_T$.
2. **Variance Reduction (**$\mathrm{ESS} \uparrow$**):** Excluding unsupported target points prevents excessive weight concentration on a few samples, yielding a more uniform weight distribution and larger Effective Sample Size (ESS), thereby stabilizing the $C_2\sqrt{1/\mathrm{ESS}}$ term.

4 KERNEL MEAN MATCHING BASED CONFORMAL PREDICTION

We describe the practical implementation of split conformal prediction under covariate shift using KMM and its selective extension, focusing on classification. We also discuss class-conditional calibration; details are provided in Appendix D.

KMM-Based Split Conformal Prediction. Given labeled data $\mathcal{D}$ and unlabeled test covariates $z_j{}_{j=1}^{m}$, we split $\mathcal{D}$ into training and calibration sets, train a predictor on the training split, and compute conformity scores on the calibration data. KMM is then applied between calibration and test covariates to obtain importance weights, which are used to compute a weighted empirical quantile of the calibration scores.

Selective KMM. The double-weighted formulation in 3.3 induces sparsity in the selection variables α_j via complementary slackness, concentrating mass near the boundaries. In finite samples, the solutions cluster near 0 and 1 but are not exactly discrete. We therefore adopt a two-stage procedure: first obtain α_j from the double-weighted objective, then threshold the learned α values to retain sufficiently large-α test points, and finally rerun standard KMM on the retained subset. Coverage applies conditionally to the selected test instances under the assumptions of Section 3.1. The unified procedure is summarized in Algorithm 1.

Connection to Selective Conformal Inference. Our formulation is conceptually related to selective conformal inference Angelopoulos et al. [2021], which guarantees coverage on a data-dependent subset of test instances. Classical selective methods typically assume exchangeability and use predictive confidence to define this subset. In contrast, our selection mechanism is driven by support mismatch under covariate shift: guarantees are restricted to regions with stable reweighting and reliable density-ratio estimation.

Selective guarantees are particularly relevant in high-stakes domains. In molecular discovery, conformal prediction is

Algorithm 1 Split Conformal Prediction with (Selective) KMM

Require: Labeled data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, test covariates $\{z_j\}_{j=1}^{m}$, miscoverage level α_{conf}, kernel k, weight bound B, threshold θ (optional)

Ensure: Prediction sets $\hat{C}(z_j)$

Split $\mathcal{D}$ into training and calibration sets $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{n}$

Train predictor f and define conformity score $S(x, y)$

Compute calibration scores $S_i = S(x_i, y_i)$

if Selective KMM **then**

  Solve double-weighted KMM to obtain α_j

  $\mathcal{T}_{\text{sel}} \leftarrow \{z_j : \alpha_j \geq \theta\}$ (or top τm)

else

  $\mathcal{T}_{\text{sel}} \leftarrow \{z_j\}_{j=1}^{m}$

end if

Solve KMM between $\mathcal{D}_{\text{cal}}$ and $\mathcal{T}_{\text{sel}}$ to obtain w_i

Normalize $\tilde{w}_i = w_i / \sum_{k=1}^{n} w_k$

Compute weighted quantile $\hat{q}_{1-\alpha_{\text{conf}}}$ satisfying

$$\sum_{i=1}^{n} \tilde{w}_i \mathbf{1}\{S_i \leq t\} \geq 1 - \alpha_{\text{conf}}$$

for $z_j \in \mathcal{T}_{\text{sel}}$ **do**

  $\hat{C}(z_j) = \{y : S(z_j, y) \leq \hat{q}_{1-\alpha_{\text{conf}}}\}$

end for

used to prioritize active compounds for experimental validation, where missing truly active molecules is costly; abstaining in low-support regions is therefore practical. Similarly, in healthcare, models may defer out-of-distribution cases to clinicians while maintaining calibrated guarantees on well-supported patient groups.

5 EXPERIMENTAL DETAILS

In this section, we evaluate whether the proposed methods correct covariate shift and restore conformal coverage on small-molecule property prediction tasks, where conformal prediction is widely used to prioritize compounds for wet-lab validation Astigarraga et al. [2025]. These benchmarks derive from real-world drug discovery pipelines Huang et al. [2021] and are standard testbeds for molecular uncertainty quantification Laghuvarapu et al. [2023].

5.1 DATASETS

We conduct experiments on widely evaluate molecular property prediction benchmarks Fooladi et al. [2025] from Therapeutics Data Commons (TDC) Huang et al. [2021]. We evaluate on five classification tasks: **Tox21** (multi-target toxicity from high-throughput assays), **AMES** (mutagenicity prediction), **hERG** (cardiac toxicity via potassium channel inhibition), **BBB-Martins** (blood–brain barrier penetration), and **HIV** (bioactivity via inhibition of HIV replication). These tasks span toxicity and bioactivity prediction and exhibit substantial structural diversity across molecular scaffolds. In such regimes, reliable uncertainty quantification is critical: overconfident predictions on novel compounds can lead to unsafe or inefficient decisions, while overly conservative predictions hinder prioritization. Conformal prediction offers a principled mechanism for controlling predictive uncertainty, making these benchmarks well-suited for evaluating robustness under covariate shift.

5.2 BASELINES AND METHODS

We compare against the following approaches:

1. **Vanilla Conformal Prediction.** Standard split conformal prediction assuming exchangeability.
2. **KDE Density Ratio Estimation (64D).** Importance weights via kernel density estimation in 64-dimensional learned feature space.
3. **KDE Density Ratio Estimation (8D).** KDE using an 8-dimensional linear projection to assess sensitivity to dimensionality, as suggest by CoDrugLaghuvarapu et al. [2023]. We refer to this as *CoDrugKDE*.
4. **Logistic Classifier-Based Density Ratio Estimation.** Logistic regression domain classifier distinguishing calibration and test samples.
5. **KMM**: Kernel moment matching with bounded importance weights but no selective filtering.
6. **SKMM (KMM + α Filtering)** Joint optimization of calibration weights (β) and target selection variables (α), corresponding to our selective KMM framework.

5.3 IMPLEMENTATION DETAILS

Covariate Shift Construction. To simulate realistic structural shift, we consider Fingerprint split Wu et al. [2018], following prior work Laghuvarapu et al. [2023]. In this split, molecules are represented using binary embeddings (fingerprints) and partitioned based on distance from the dataset centroid in fingerprint space. The 15% farthest molecules form the test set, while the remainder constitutes the calibration pool. This induces reduced scaffold overlap and support mismatch, reflecting practical discovery settings where new chemical series differ from historical data.

Model Training. All predictive models use the AttentiveFP graph neural network architecture Xiong et al. [2019]. Models are trained with five independent random seeds per dataset. Conformal calibration and evaluation are performed separately for each run, and results are averaged.

Weight Construction. For weighting-based methods, covariate representations are extracted from the final hidden layer of AttentiveFP (64-dimensional embeddings).

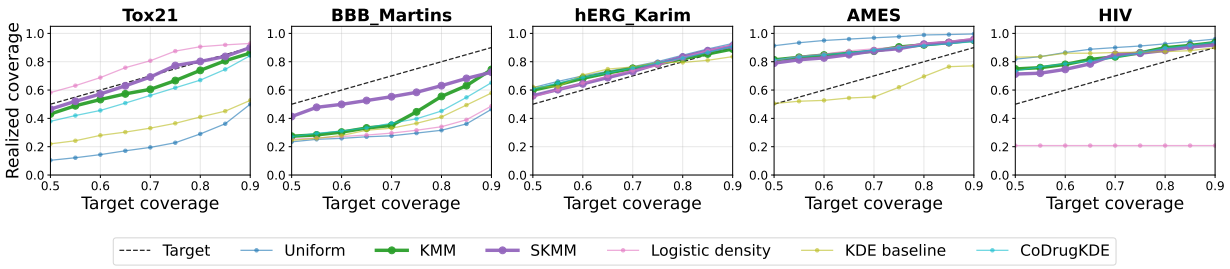


Figure 2: Calibration curves under covariate shift (global calibration). The dashed line denotes nominal coverage. Uniform CP undercovers systematically, density-ratio baselines partially correct this, KMM improves alignment, and SKMM most closely tracks the target across datasets.

These embeddings serve as inputs for density ratio estimation and kernel moment matching, reflecting realistic high-dimensional representation learning regimes. For the *CoDrugKDE* baseline, we evaluate an 8-dimensional linear projection to assess sensitivity to dimensionality.

Conformal Setup. Equation 1 describes the standard marginal coverage guarantee of conformal prediction, where coverage holds over the joint distribution of (X, Y). In addition to global calibration, we also evaluate *class-conditional (Mondrian) conformal prediction*, which computes quantile thresholds separately within predefined groups—in our case, class labels. Mondrian conformal prediction guarantees coverage conditional on class membership, i.e., $\mathbb{P}\Big(Y_{N+1} \in \hat{C}(X_{N+1}) \mid Y_{N+1} = c\Big) \geq 1-\alpha$ for each class c. For molecular activity tasks, we calibrate on the active class (C1), while for toxicity tasks we calibrate on the non-toxic class (C0), prioritizing reliable coverage for outcomes most critical to downstream drug discovery decisions. Additional details of this are provided in Appendix D.

KMM Hyperparameters. For kernel mean matching (KMM), the kernel bandwidth is selected using the maximum test power criterion following prior work Gretton et al. [2012]. The weight upper bound is set to $B = 30$, and the target selection fraction is fixed at $\alpha = 0.5$, corresponding to selecting 50% of the test molecules. These hyperparameters were not tuned per dataset.

Additional implementation details and hyperparameters are provided in Appendix E. All reported results are averaged over five independent random runs and detailed error bars are included in Appendix F.

6 RESULTS AND DISCUSSION

6.1 COVERAGE UNDER COVARIATE SHIFT

We first evaluate vanilla (uniform) conformal prediction under both no-shift and fingerprint-based covariate shift. Results are summarized in Table 1. To quantify calibration error, we report the mean absolute deviation (MAD) from

Table 1: Mean absolute deviation (MAD) of uniform conformal prediction under random (no-shift) and shifted splits.

Dataset	Random	Shifted
AMES	**0.0121**	0.0844
BBB-Martins	**0.0346**	0.4579
HIV	**0.0044**	0.1830
Tox21	**0.0249**	0.4366
hERG-Karim	**0.0254**	0.0384

nominal coverage. Specifically, for target coverage levels $1-\alpha \in \{0.5, 0.55, \ldots, 0.9\}$,

$$\text{MAD} = \frac{1}{K}\sum_{k=1}^{K} \left|\widehat{\text{cov}}(1-\alpha_k) - (1-\alpha_k)\right|,$$

where $\widehat{\text{cov}}(1-\alpha_k)$ denotes empirical coverage at level $1-\alpha_k$, and K is the number of evaluated levels. This metric aggregates deviation across coverage levels and provides a measure of miscalibration.

Under random (no-shift) splits, uniform conformal prediction achieves low MAD (e.g., 0.0044 on HIV and 0.0121 on AMES), confirming nominal coverage under exchangeability. Under fingerprint-based covariate shift, calibration deteriorates sharply: MAD exceeds 0.43 on Tox21 and 0.45 on BBB-Martins, and increases by more than an order of magnitude even on milder shifts such as HIV. These results demonstrate that exchangeability violations substantially degrade conformal validity in realistic molecular settings, motivating explicit covariate-shift correction.

6.2 GLOBAL AND MONDRIAN CALIBRATION

Table 2 reports mean absolute deviation (MAD) from nominal coverage under both global and Mondrian conformal calibration. Uniform conformal prediction performs poorly under covariate shift, exhibiting the highest deviation across nearly all datasets, with mean ranks of 5.4 (global) and 5.0 (Mondrian). In particular, uniform MAD exceeds 0.40 on Tox21 and BBB-Martins, confirming that exchangeability violations severely degrade calibration.

Density-ratio baselines partially mitigate this effect but

Table 2: Mean absolute deviation (MAD) from nominal coverage under covariate shift for global (G) and Mondrian (M) calibration. Lower is better. Best per column is **bold**; second-best is **gray bold**. SKMM attains the lowest MAD under global calibration and the best overall mean rank across both settings.

	Uniform		KDE		CoDrug KDE		Logistic		KMM		SKMM	
Dataset	G	M	G	M	G	M	G	M	G	M	G	M
AMES	0.166	0.084	0.137	**0.046**	0.093	0.090	0.102	0.096	**0.078**	0.079	**0.065**	**0.072**
Tox21	0.409	0.437	0.323	0.331	0.103	0.110	0.073	0.089	**0.053**	**0.057**	**0.015**	**0.020**
BBB-Martins	0.424	0.458	0.364	0.393	0.329	0.381	0.396	0.443	**0.319**	**0.353**	**0.213**	**0.270**
hERG	0.034	0.038	0.096	0.094	0.031	0.030	**0.020**	**0.022**	0.028	0.036	**0.009**	**0.019**
HIV	0.185	0.183	0.291	0.283	0.145	0.151	0.297	0.269	**0.087**	**0.098**	**0.013**	**0.022**
Mean Rank	5.4	5.0	5.0	4.4	3.4	3.6	4.0	4.2	**2.2**	**2.6**	**1.0**	**1.2**

remain inconsistent. Logistic weighting substantially improves Tox21 (0.073 vs. 0.409 under uniform), yet remains unstable on BBB-Martins (0.396) and HIV (0.297). KDE-based approaches are particularly sensitive to the dimensionality of the representation, with MAD exceeding 0.30 on multiple datasets under global calibration. Even the CoDrug KDE variant remains above 0.30 in BBB-Martins and fails to consistently match nominal coverage.

In contrast, KMM (no filtering) consistently reduces deviation across datasets. On Tox21 (global), MAD decreases from 0.073 (logistic) to 0.053 under KMM, with similar improvements in BBB-Martins (0.319) and HIV (0.087), indicating more stable moment alignment under covariate shift. The selective extension (SKMM) further improves calibration, achieving the lowest MAD on all five benchmarks under global calibration—reducing MAD to 0.015 in Tox21, 0.009 in hERG, and 0.013 in HIV. Under Mondrian calibration, SKMM remains competitive and attains the best mean rank (1.2), showing that its gains extend to class-conditional settings.

Furthermore, figure 2 corroborates these observations: SKMM consistently tracks the diagonal target line across coverage levels, while uniform and KDE baselines exhibit systematic undercoverage under shift. KMM narrows this gap, and SKMM further stabilizes calibration across both toxicity and activity prediction tasks. We observe similar trends global calibration, depicted in Appendix E.

Overall, the results demonstrate that moment-matching-based reweighting improves calibration stability under covariate shift, and that selective filtering further enhances robustness in low-overlap regimes. On average across datasets under global calibration, SKMM achieves a 55% reduction in MAD relative to CoDrugKDE, highlighting substantial improvements in coverage stability.

6.3 BIAS–VARIANCE AND SELECTIVE FILTERING

To understand the source of improvement under covariate shift, we examine the dominant terms in our coverage bound,

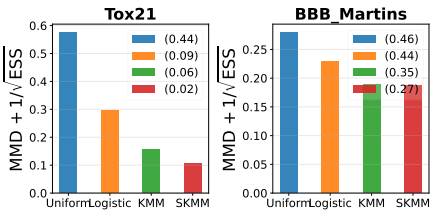


Figure 3: Bias–variance proxy ($\mathrm{MMD} + \sqrt{1/\mathrm{ESS}}$) under covariate shift. MAD in brackets.

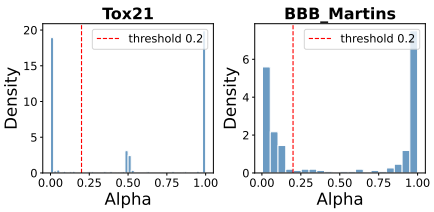


Figure 4: Selection weights under SKMM.

namely $\mathrm{MMD} + \sqrt{1/\mathrm{ESS}}$, which capture bias (distributional mismatch) and variance (weight concentration), respectively. Figure 3 reports this bias–variance proxy across methods. Uniform conformal prediction exhibits large combined error due to both high MMD and lack of shift correction. The ordering of $\mathrm{MMD} + \sqrt{1/\mathrm{ESS}}$ across methods closely aligns with empirical calibration performance, with KMM and SKMM achieving the smallest values

Selective KMM (SKMM) further reduces this quantity by restricting correction to regions of shared support, achieving the smallest combined error across datasets. To better understand this behavior, Figure 4 visualizes the learned selection weights α. The distribution of α concentrates near the boundaries, with approximately half of the test instances retained (as controlled by $\tau = 0.5$). The learned sparsity is consistent across datasets and remains stable under shift. and exhibit the same qualitative trends.

7 CONCLUSION

We presented KMM-CP, a framework for conformal prediction under covariate shift that leverages Kernel Mean Matching with a selective extension for limited-support regimes. By combining bounded moment matching with target-aware filtering, the proposed method improves coverage stability relative to standard density-ratio approaches, particularly in high-dimensional molecular representation spaces. Our analysis clarifies how moment alignment and effective sample size govern coverage under shift, and experiments across multiple molecular property benchmarks demonstrate more reliable coverage under structural distribution shift.

We acknowledge several limitations. The selective filtering step reduces the number of evaluated samples, and performance depends on the quality of learned representations and kernel hyperparameters. Although our synthetic shift construction reflects realistic structural mismatch, real-world deployment may involve more complex distribution shifts. Finally, while we demonstrate effectiveness on small-molecule property prediction tasks, generalization to other domains and data modalities remains to be validated.

References

Duarte C. Almeida, João Bravo, Jacopo Bono, Pedro Bizarro, and Mário A. T. Figueiredo. High-probability risk control under covariate shift. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266 of *Proceedings of Machine Learning Research*, pages 133–152. PMLR, 2025.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Aaditya Ramdas. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Mario Astigarraga, Andrés Sánchez-Ruiz, and Gonzalo Colmenarejo. Conformal prediction-based machine learning in cheminformatics: Current applications and new challenges. *Artificial Intelligence in the Life Sciences*, 7: 100127, 2025.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *Annual Review of Statistics and Its Application*, 2023.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10:2137–2155, 2009.

Clara Fannjiang, Stephen Bates, Anastasios N. Angelopoulos, Jennifer Listgarten, and Michael I. Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022. doi: 10.1073/pnas.2204569119.

Hosein Fooladi, Thi Ngoc Lan Vu, Miriam Mathea, and Johannes Kirchmair. Evaluating machine learning models for molecular property prediction: performance and robustness on out-of-distribution data. *Journal of Chemical Information and Modeling*, 65(19):9871–9891, 2025.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL `https://proceedings.neurips.cc/paper_files/paper/2006/file/a2186aa7c086b46ad4e8bf81e2a3a19b-Paper.pdf`.

Jiayuan Huang, Arthur Gretton, Karsten M. Borgwardt, Bernhard Schölkopf, and Alex J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, volume 19, 2007a.

Jiayuan Huang, Alex Smola, Arthur Gretton, Karsten Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, volume 19, 2007b.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

Siddhartha Laghuvarapu, Zhen Lin, and Jimeng Sun. Codrug: Conformal drug property prediction with density estimation under covariate shift. *Advances in Neural Information Processing Systems*, 36:37728–37747, 2023.

Mufei Li, Jinjing Zhou, Jiajing Hu, Wenxuan Fan, Yangkang Zhang, Yaxin Gu, and George Karypis. Dgl-lifesci: An open-source toolkit for deep learning on graphs in life science. *ACS Omega*, 2021.

José I. Segovia-Martín, Santiago Mazuelas, and Anqi Liu. Double-weighting for covariate shift adaptation. *arXiv preprint arXiv:2305.08637*, 2023.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986a.

Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986b.

Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. The MIT Press, 2012.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 20, 2008.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall, 1995.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.

A PROOFS OF THEORETICAL RESULTS

A.1 PROOF OF THEOREM 2

Fix a probability measure Q on $\mathbb{R}^d$ and define

$$F_Q(t) := \mathbb{E}_{X\sim Q}[g_t(X)].$$

Then, for any fixed $t \in \mathbb{R}$,

$$F_Q(t) - F_T(t) = \mathbb{E}_{X\sim Q}[g_t(X)] - \mathbb{E}_{X\sim P_X^T}[g_t(X)].$$

Add and subtract $\tilde{g}_t$ to obtain

$$\begin{aligned}\left|F_Q(t) - F_T(t)\right| \le &\left|\mathbb{E}_{X\sim Q}[\tilde{g}_t(X)] - \mathbb{E}_{X\sim P_X^T}[\tilde{g}_t(X)]\right| \\ &+ |\mathbb{E}_{X\sim Q}[g_t(X) - \tilde{g}_t(X)]| + \left|\mathbb{E}_{X\sim P_X^T}[g_t(X) - \tilde{g}_t(X)]\right|.\end{aligned}$$

Since $|g_t(x) - \tilde{g}_t(x)| \le \|g_t - \tilde{g}_t\|_\infty$ for all x, and both Q and P_X^T are probability measures, we have

$$|\mathbb{E}_{X\sim Q}[g_t(X) - \tilde{g}_t(X)]| \le \|g_t - \tilde{g}_t\|_\infty, \qquad \left|\mathbb{E}_{X\sim P_X^T}[g_t(X) - \tilde{g}_t(X)]\right| \le \|g_t - \tilde{g}_t\|_\infty.$$

For the first term, if $\|\tilde{g}_t\|_{\mathcal{H}} = 0$, then $\tilde{g}_t = 0$ in $\mathcal{H}$ and hence

$$\left|\mathbb{E}_{X\sim Q}[\tilde{g}_t(X)] - \mathbb{E}_{X\sim P_X^T}[\tilde{g}_t(X)]\right| = 0.$$

Otherwise, define

$$h_t := \frac{\tilde{g}_t}{\|\tilde{g}_t\|_{\mathcal{H}}},$$

so that $\|h_t\|_{\mathcal{H}} = 1$. By the definition of $\mathrm{MMD}_{\mathcal{H}}$,

$$\begin{aligned}\left|\mathbb{E}_{X\sim Q}[\tilde{g}_t(X)] - \mathbb{E}_{X\sim P_X^T}[\tilde{g}_t(X)]\right| &= \|\tilde{g}_t\|_{\mathcal{H}} \left|\mathbb{E}_{X\sim Q}[h_t(X)] - \mathbb{E}_{X\sim P_X^T}[h_t(X)]\right| \\ &\le \|\tilde{g}_t\|_{\mathcal{H}}\, \mathrm{MMD}_{\mathcal{H}}(Q, P_X^T).\end{aligned}$$

Combining the bounds gives

$$|F_Q(t) - F_T(t)| \le \|\tilde{g}_t\|_{\mathcal{H}}\, \mathrm{MMD}_{\mathcal{H}}(Q, P_X^T) + 2\|g_t - \tilde{g}_t\|_\infty.$$

Taking the supremum over $t \in \mathbb{R}$ and using the uniform bounds

$$\sup_{t\in\mathbb{R}} \|\tilde{g}_t\|_{\mathcal{H}} \le B_{\mathcal{H}}, \qquad \sup_{t\in\mathbb{R}} \|g_t - \tilde{g}_t\|_\infty \le \varepsilon_\infty,$$

we obtain

$$\sup_{t\in\mathbb{R}} |F_Q(t) - F_T(t)| \le B_{\mathcal{H}}\, \mathrm{MMD}_{\mathcal{H}}(Q, P_X^T) + 2\varepsilon_\infty.$$

Finally, applying this bound with $Q = \widehat{P}_{S,X}^{\,w}$ and recalling that

$$F_w(t) = \mathbb{E}_{X\sim \widehat{P}_{S,X}^{\,w}}[g_t(X)],$$

yields

$$\sup_{t\in\mathbb{R}} |F_w(t) - F_T(t)| \le B_{\mathcal{H}}\, \mathrm{MMD}_{\mathcal{H}}(\widehat{P}_{S,X}^{\,w}, P_X^T) + 2\varepsilon_\infty.$$

This proves the theorem.

A.2 PROOF OF THEOREM 3

Proof. Throughout the proof, condition on $X_1, \dots, X_n, Z_1, \dots, Z_m$ so the weights $\tilde{w}_1, \dots, \tilde{w}_n$ are deterministic and the labels $Y_1, \dots, Y_n$ remain independent.

1. **Reduce coverage error to a CDF error and an overshoot term.** By definition of $\widehat{C}$,

$$\mathbb{P}_{(X,Y)\sim P_T}\big(Y \in \widehat{C}(X)\big) = \mathbb{P}_{(X,Y)\sim P_T}\big(s(X,Y) \le \widehat{q}_{1-\alpha}\big) = F_T(\widehat{q}_{1-\alpha}).$$

Since $\widehat{F}_w(\widehat{q}_{1-\alpha}) \ge 1-\alpha$ by definition of $\widehat{q}_{1-\alpha}$,

$$F_T(\widehat{q}_{1-\alpha}) - (1-\alpha) = \big(F_T(\widehat{q}_{1-\alpha}) - \widehat{F}_w(\widehat{q}_{1-\alpha})\big) + \big(\widehat{F}_w(\widehat{q}_{1-\alpha}) - (1-\alpha)\big),$$

and therefore

$$\Big|F_T(\widehat{q}_{1-\alpha}) - (1-\alpha)\Big| \le \sup_{t\in\mathbb{R}} |F_T(t) - \widehat{F}_w(t)| + \big(\widehat{F}_w(\widehat{q}_{1-\alpha}) - (1-\alpha)\big).$$

Because $\widehat{F}_w$ is right-continuous and has jumps of size exactly $\tilde{w}_i$ at the distinct points $s(X_i, Y_i)$, the distinct-scores assumption implies that every jump has size at most $\max_i \tilde{w}_i$. Moreover, by minimality of $\widehat{q}_{1-\alpha}$ we have $\widehat{F}_w(t) < 1-\alpha$ for all $t < \widehat{q}_{1-\alpha}$, hence the overshoot satisfies

$$0 \le \widehat{F}_w(\widehat{q}_{1-\alpha}) - (1-\alpha) \le \max_{1\le i\le n} \tilde{w}_i.$$

Therefore

$$\Big|\mathbb{P}_{(X,Y)\sim P_T}\big(Y \in \widehat{C}(X)\big) - (1-\alpha)\Big| \le \sup_{t\in\mathbb{R}} |F_T(t) - \widehat{F}_w(t)| + \max_{1\le i\le n} \tilde{w}_i.$$

2. **Decompose $\sup_t |F_T(t) - \widehat{F}_w(t)|$ into bias and variance.** By the triangle inequality,

$$\sup_{t\in\mathbb{R}} |F_T(t) - \widehat{F}_w(t)| \le \sup_{t\in\mathbb{R}} |F_T(t) - F_w(t)| + \sup_{t\in\mathbb{R}} |F_w(t) - \widehat{F}_w(t)|.$$

The first term is controlled by Theorem 2:

$$\sup_{t\in\mathbb{R}} |F_T(t) - F_w(t)| \le B_{\mathcal{H}}\, \mathrm{MMD}_{\mathcal{H}}(\widehat{P}^w_{S,X}, P^T_X) + 2\,\varepsilon_\infty.$$

It remains to control $\sup_t |F_w(t) - \widehat{F}_w(t)|$.

3. **An expectation bound for the weighted empirical process.** Define

$$\Delta := \sup_{t\in\mathbb{R}} \Big|\widehat{F}_w(t) - F_w(t)\Big| = \sup_{t\in\mathbb{R}} \left|\sum_{i=1}^n \tilde{w}_i \Big(\mathbf{1}\{s(X_i,Y_i) \le t\} - g_t(X_i)\Big)\right|.$$

Let $Y'_1, \dots, Y'_n$ be an independent copy of $Y_1, \dots, Y_n$ conditional on $X_1, \dots, X_n$ (and thus also on the weights). Define

$$\widehat{F}'_w(t) := \sum_{i=1}^n \tilde{w}_i\, \mathbf{1}\{s(X_i, Y'_i) \le t\}.$$

By Jensen's inequality and $F_w(t) = \mathbb{E}[\widehat{F}_w(t) \mid X_1, \dots, X_n, Z_1, \dots, Z_m]$,

$$\mathbb{E}\big[\Delta \,\big|\, X_1, \dots, X_n, Z_1, \dots, Z_m\big] \le \mathbb{E}\left[\sup_t \Big|\widehat{F}_w(t) - \widehat{F}'_w(t)\Big| \,\Big|\, X_1, \dots, X_n, Z_1, \dots, Z_m\right].$$

Introduce i.i.d. Rademacher signs $\varepsilon_1, \dots, \varepsilon_n$ independent of everything else. By standard symmetrization,

$$\mathbb{E}\big[\Delta \,\big|\, X_1, \dots, X_n, Z_1, \dots, Z_m\big] \le 2\,\mathbb{E}\left[\sup_{t\in\mathbb{R}} \left|\sum_{i=1}^n \tilde{w}_i\, \varepsilon_i\, \mathbf{1}\{s(X_i,Y_i) \le t\}\right| \,\Big|\, X_1, \dots, X_n, Z_1, \dots, Z_m\right].$$

Condition further on $X_1, \dots, X_n, Z_1, \dots, Z_m, Y_1, \dots, Y_n$ and write $s_i := s(X_i, Y_i)$. Let π be a permutation such that $s_{\pi(1)} \le \cdots \le s_{\pi(n)}$. Then

$$\sup_{t \in \mathbb{R}} \left| \sum_{i=1}^{n} \tilde{w}_i \, \varepsilon_i \, \mathbf{1}\{s_i \le t\} \right| = \max_{0 \le k \le n} \left| \sum_{j=1}^{k} \tilde{w}_{\pi(j)} \, \varepsilon_{\pi(j)} \right|.$$

Define $M_k := \sum_{j=1}^{k} \tilde{w}_{\pi(j)} \, \varepsilon_{\pi(j)}$ with $M_0 := 0$. By Doob's L^2 maximal inequality,

$$\mathbb{E}\left[\max_{0 \le k \le n} M_k^2 \,\middle|\, X_1, \dots, X_n, Z_1, \dots, Z_m, Y \right] \le 4 \, \mathbb{E}\left[M_n^2 \,\middle|\, X_1, \dots, X_n, Z_1, \dots, Z_m, Y \right] = 4 \sum_{i=1}^{n} \tilde{w}_i^2.$$

Taking square roots and using Jensen's inequality yields

$$\mathbb{E}\left[\sup_{t \in \mathbb{R}} \left| \sum_{i=1}^{n} \tilde{w}_i \, \varepsilon_i \, \mathbf{1}\{s_i \le t\} \right| \,\middle|\, X_1, \dots, X_n, Z_1, \dots, Z_m, Y \right] \le 2 \sqrt{\sum_{i=1}^{n} \tilde{w}_i^2}.$$

Combining the last two displays and unconditioning on Y gives

$$\mathbb{E}\big[\Delta \,\big|\, X_1, \dots, X_n, Z_1, \dots, Z_m\big] \le 4 \sqrt{\sum_{i=1}^{n} \tilde{w}_i^2} = 4 \sqrt{\tfrac{1}{\mathrm{ESS}}}.$$

4. **A high-probability bound via bounded differences.** View Δ as a function of the independent variables $Y_1, \dots, Y_n$ conditional on $X_1, \dots, X_n, Z_1, \dots, Z_m$. If one replaces Y_i by another value $Y_i^\sharp$ (leaving all other Y_j fixed), then for every t the quantity $\widehat{F}_w(t)$ changes by at most $\tilde{w}_i$, and $F_w(t)$ is unchanged. Hence Δ changes by at most $\tilde{w}_i$. McDiarmid's inequality implies that for any $u > 0$,

$$\mathbb{P}\big(\Delta - \mathbb{E}[\Delta \mid X, Z] \ge u \,\big|\, X, Z\big) \le \exp\left(-\frac{2u^2}{\sum_{i=1}^{n} \tilde{w}_i^2} \right) = \exp\big(-2u^2 \, \mathrm{ESS}\big),$$

where X denotes $(X_1, \dots, X_n)$ and Z denotes $(Z_1, \dots, Z_m)$. Taking $u := \sqrt{\frac{1}{2\mathrm{ESS}} \log \frac{1}{\delta}}$ yields that with probability at least $1 - \delta$,

$$\Delta \le 4 \sqrt{\tfrac{1}{\mathrm{ESS}}} + \sqrt{\tfrac{1}{2\mathrm{ESS}} \log \tfrac{1}{\delta}}.$$

On the event from item 4, we combine items 1 and 2 to obtain

$$\begin{aligned} \left| \mathbb{P}_{(X,Y) \sim P_T}\big(Y \in \widehat{C}(X)\big) - (1 - \alpha) \right| &\le \sup_t |F_T(t) - F_w(t)| + \sup_t |F_w(t) - \widehat{F}_w(t)| + \max_i \tilde{w}_i \\ &\le B_{\mathcal{H}} \, \mathrm{MMD}_{\mathcal{H}}(\widehat{P}_{S,X}^{w}, P_X^T) + 2\,\varepsilon_\infty + \Delta + \max_i \tilde{w}_i. \end{aligned}$$

Finally, $\max_i \tilde{w}_i \le \sqrt{\sum_{i=1}^{n} \tilde{w}_i^2} = \sqrt{1/\mathrm{ESS}}$, so item 4 gives

$$\Delta + \max_i \tilde{w}_i \le 5 \sqrt{\tfrac{1}{\mathrm{ESS}}} + \sqrt{\tfrac{1}{2\mathrm{ESS}} \log \tfrac{1}{\delta}}.$$

Substituting this completes the proof.

□

A.3 PROOF OF COROLLARY 1

Proof. Fix $\eta > 0$ and $\delta \in (0,1)$. By Theorem 3, conditional on $X_1, \ldots, X_n, Z_1, \ldots, Z_m$, with probability at least $1-\delta$ over $Y_1, \ldots, Y_n$,

$$\left|\mathbb{P}_{(X,Y)\sim P_T}\big(Y \in \widehat{C}(X)\big) - (1-\alpha)\right| \le B_{\mathcal{H}}\,\mathrm{MMD}_{\mathcal{H}}\big(\widehat{P}^{\,w}_{S,X}, P^T_X\big) + 2\,\varepsilon_{\infty,n,m} + \left(5 + \sqrt{\tfrac{1}{2}\log\tfrac{1}{\delta}}\right)\sqrt{\tfrac{1}{\mathrm{ESS}}}.$$

Define the random bound

$$R_{n,m}(\delta) := B_{\mathcal{H}}\,\mathrm{MMD}_{\mathcal{H}}\big(\widehat{P}^{\,w}_{S,X}, P^T_X\big) + 2\,\varepsilon_{\infty,n,m} + \left(5 + \sqrt{\tfrac{1}{2}\log\tfrac{1}{\delta}}\right)\sqrt{\tfrac{1}{\mathrm{ESS}}}.$$

Then,

$$\begin{aligned}
&\mathbb{P}\Big(\left|\mathbb{P}_{(X,Y)\sim P_T}\big(Y \in \widehat{C}(X)\big) - (1-\alpha)\right| > \eta\Big) \\
&\qquad = \mathbb{E}\Big[\mathbb{P}\Big(\left|\mathbb{P}_{(X,Y)\sim P_T}\big(Y \in \widehat{C}(X)\big) - (1-\alpha)\right| > \eta \,\Big|\, X_1, \ldots, X_n, Z_1, \ldots, Z_m\Big)\Big] \\
&\qquad \le \mathbb{E}\Big[\mathbf{1}\{R_{n,m}(\delta) > \eta\} + \mathbb{P}\big(\text{the bound above fails} \,\big|\, X_1, \ldots, X_n, Z_1, \ldots, Z_m\big)\Big] \\
&\qquad \le \mathbb{P}\big(R_{n,m}(\delta) > \eta\big) + \delta,
\end{aligned}$$

where the probability is over $X_1, \ldots, X_n, Z_1, \ldots, Z_m, Y_1, \ldots, Y_n$. By assumption, $R_{n,m}(\delta) \to 0$ in probability for each fixed δ, hence $\mathbb{P}(R_{n,m}(\delta) > \eta) \to 0$. Taking $\limsup$ and then letting $\delta \downarrow 0$ yields

$$\lim_{n,m\to\infty} \mathbb{P}\Big(\left|\mathbb{P}_{(X,Y)\sim P_T}\big(Y \in \widehat{C}(X)\big) - (1-\alpha)\right| > \eta\Big) = 0.$$

□

B JUSTIFICATION FOR SELECTIVE KMM

Theorem 4 (Conditional boundary rule under a yield constraint)**.** *Fix w and consider the optimization over $\alpha \in \mathbb{R}^m$*

$$\begin{aligned}
\min_{\alpha} \quad & J(\alpha) \\
\textit{s.t.} \quad & 0 \le \alpha_j \le 1 \quad \textit{for all } j \in [m], \\
& \frac{1}{m}\sum_{j=1}^{m} \alpha_j \ge \tau,
\end{aligned}$$

where J is convex and differentiable. Assume $\tau < 1$, so Slater's condition holds. Let $\alpha^\star$ be an optimal solution, and let $(\lambda^\star, \mu^\star, \nu^\star)$ be KKT multipliers, where $\lambda^\star_j \ge 0$ and $\mu^\star_j \ge 0$ correspond to the constraints $-\alpha_j \le 0$ and $\alpha_j - 1 \le 0$, and $\nu^\star \ge 0$ corresponds to the yield constraint $\tau - \frac{1}{m}\sum_{j=1}^m \alpha_j \le 0$. Then for each $j \in [m]$ the following statements hold:

1. *If $\frac{\partial J}{\partial \alpha_j}(\alpha^\star) - \frac{\nu^\star}{m} > 0$, then $\alpha^\star_j = 0$.*
2. *If $\frac{\partial J}{\partial \alpha_j}(\alpha^\star) - \frac{\nu^\star}{m} < 0$, then $\alpha^\star_j = 1$.*
3. *If $\frac{\partial J}{\partial \alpha_j}(\alpha^\star) - \frac{\nu^\star}{m} = 0$, then $\alpha^\star_j$ may be interior or boundary; in particular, any interior coordinate $\alpha^\star_j \in (0,1)$ must satisfy the equality.*

Proof. Since $\tau < 1$, there exists a strictly feasible point, for example $\alpha^0_j = (\tau+1)/2$ for all j, which satisfies $0 < \alpha^0_j < 1$ and $\frac{1}{m}\sum_j \alpha^0_j = (\tau+1)/2 > \tau$. Hence Slater's condition holds, and KKT conditions are necessary and sufficient for optimality.

The KKT conditions assert the existence of multipliers $(\lambda^\star, \mu^\star, \nu^\star)$ satisfying:

1. (Primal feasibility) $0 \le \alpha^\star_j \le 1$ for all j and $\frac{1}{m}\sum_{j=1}^m \alpha^\star_j \ge \tau$.

2. (Dual feasibility) $\lambda_j^\star \geq 0$, $\mu_j^\star \geq 0$ for all j, and $\nu^\star \geq 0$.
3. (Complementary slackness)

$$\lambda_j^\star \alpha_j^\star = 0, \qquad \mu_j^\star(\alpha_j^\star - 1) = 0, \qquad \nu^\star \left(\tau - \frac{1}{m} \sum_{j=1}^{m} \alpha_j^\star \right) = 0.$$

4. (Stationarity) For each $j \in [m]$,

$$\frac{\partial J}{\partial \alpha_j}(\alpha^\star) - \frac{\nu^\star}{m} - \lambda_j^\star + \mu_j^\star = 0.$$

Fix an index $j \in [m]$ and define the shifted gradient

$$g_j^\star := \frac{\partial J}{\partial \alpha_j}(\alpha^\star) - \frac{\nu^\star}{m}.$$

By stationarity,

$$g_j^\star = \lambda_j^\star - \mu_j^\star.$$

We now analyze cases using complementary slackness:

1. If $\alpha_j^\star = 0$, then $\mu_j^\star(\alpha_j^\star - 1) = \mu_j^\star(-1) = 0$ implies $\mu_j^\star = 0$, hence $g_j^\star = \lambda_j^\star \geq 0$.
2. If $\alpha_j^\star = 1$, then $\lambda_j^\star \alpha_j^\star = \lambda_j^\star = 0$, hence $g_j^\star = -\mu_j^\star \leq 0$.
3. If $\alpha_j^\star \in (0, 1)$, then $\lambda_j^\star \alpha_j^\star = 0$ and $\alpha_j^\star > 0$ imply $\lambda_j^\star = 0$, and $\mu_j^\star(\alpha_j^\star - 1) = 0$ and $\alpha_j^\star - 1 \neq 0$ imply $\mu_j^\star = 0$. Therefore $g_j^\star = 0$.

These implications yield the desired conditional boundary rule:

1. If $g_j^\star > 0$, then $\alpha_j^\star$ cannot equal 1 (which would force $g_j^\star \leq 0$) and cannot be interior (which would force $g_j^\star = 0$). Hence $\alpha_j^\star = 0$.
2. If $g_j^\star < 0$, then $\alpha_j^\star$ cannot equal 0 (which would force $g_j^\star \geq 0$) and cannot be interior (which would force $g_j^\star = 0$). Hence $\alpha_j^\star = 1$.
3. If $g_j^\star = 0$, then the KKT conditions allow $\alpha_j^\star$ to be interior (in which case necessarily $\lambda_j^\star = \mu_j^\star = 0$) or boundary (for example $\alpha_j^\star = 0$ with $\lambda_j^\star = 0$ or $\alpha_j^\star = 1$ with $\mu_j^\star = 0$). This proves the final claim.

□

Corollary 2 (Thresholding structure for SKMM target selection)**.** *Fix w and consider the SKMM subproblem in $\alpha \in \mathbb{R}^m$ with yield constraint*

$$\begin{aligned} \min_{\alpha} \quad & J(\alpha) := \left\| \frac{1}{n} \sum_{i=1}^{n} w_i \varphi(x_i) - \frac{1}{m} \sum_{j=1}^{m} \alpha_j \varphi(z_j) \right\|_{\mathcal{H}}^2 \\ \text{s.t.} \quad & 0 \leq \alpha_j \leq 1 \quad \textit{for all } j \in [m], \\ & \frac{1}{m} \sum_{j=1}^{m} \alpha_j \geq \tau, \end{aligned}$$

where $\tau < 1$. Let $\alpha^\star$ be an optimal solution, and let $\nu^\star \geq 0$ be a KKT multiplier for the yield constraint. Define the shifted partial derivatives

$$g_j^\star := \frac{\partial J}{\partial \alpha_j}(\alpha^\star) - \frac{\nu^\star}{m}, \qquad j \in [m].$$

Then:

1. *If* $g_j^\star > 0$*, then* $\alpha_j^\star = 0$*.*
2. *If* $g_j^\star < 0$*, then* $\alpha_j^\star = 1$*.*
3. *If* $\alpha_j^\star \in (0,1)$*, then* $g_j^\star = 0$*.*

Equivalently, there exists a scalar threshold $t^\star := \nu^\star/m$ *such that for every* j*,*

$$\alpha_j^\star = 0 \;\Rightarrow\; \frac{\partial J}{\partial \alpha_j}(\alpha^\star) \geq t^\star, \qquad \alpha_j^\star = 1 \;\Rightarrow\; \frac{\partial J}{\partial \alpha_j}(\alpha^\star) \leq t^\star, \qquad 0 < \alpha_j^\star < 1 \;\Rightarrow\; \frac{\partial J}{\partial \alpha_j}(\alpha^\star) = t^\star.$$

Proof. First note that for the SKMM objective, $J(\alpha)$ is a convex quadratic function of α because

$$J(\alpha) = \|\mu_w - \mu_\alpha\|_{\mathcal{H}}^2, \qquad \mu_w := \frac{1}{n}\sum_{i=1}^{n} w_i \varphi(x_i), \qquad \mu_\alpha := \frac{1}{m}\sum_{j=1}^{m} \alpha_j \varphi(z_j),$$

and expanding yields

$$J(\alpha) = \|\mu_w\|_{\mathcal{H}}^2 - \frac{2}{m}\sum_{j=1}^{m} \alpha_j \langle \mu_w, \varphi(z_j)\rangle_{\mathcal{H}} + \frac{1}{m^2}\sum_{j=1}^{m}\sum_{l=1}^{m} \alpha_j \alpha_l \, k(z_j, z_l),$$

whose Hessian with respect to α is $(2/m^2)\, K_{TT}$, where $K_{TT} = (k(z_j, z_l))_{j,l}$ is positive semidefinite. Hence J is convex and differentiable.

Since $\tau < 1$, the point $\alpha_j^0 = (\tau + 1)/2$ is strictly feasible, so Slater's condition holds and KKT conditions are necessary and sufficient for optimality. Therefore Theorem 4 applies directly to this choice of J, yielding the three implications stated in the corollary. The equivalent threshold formulation follows by defining $t^\star := \nu^\star/m$. □

C INSTABILITY OF CLASSIFIER-BASED DENSITY RATIOS

A common alternative to KMM for estimating the density ratio $w(x) = P_T(x)/P_S(x)$ is probabilistic classification. By training a classifier $\hat{p}(x)$ to predict the probability that a sample x belongs to the target domain, the estimated weight is given by $\hat{w}(x) = \hat{p}(x)/(1 - \hat{p}(x))$. The following proposition demonstrates why this approach can fail under covariate shift with limited support.

Proposition 1 (Unbounded MMD under Bounded Classifier Error)**.** *Let the true density ratio be* $w^*(x) = P_T(x)/P_S(x)$ *and let* $\hat{w}(x)$ *be the density ratio estimated via a classifier* $\hat{p}(x)$*. For any* $\epsilon > 0$ *and* $M > 0$*, there exists a sufficiently small* δ *and a sample containing such points such that* $\mathrm{MMD}(P_S^{\hat{w}}, P_T) > M$*.*

Proof. Assuming equal prior class probabilities for the source and target domains, the true conditional probability of a sample belonging to the target domain is $p^*(x) = P_T(x)/(P_T(x) + P_S(x))$. The true density ratio is exactly $w^*(x) = p^*(x)/(1 - p^*(x))$.

Consider a region in the covariate space where the source distribution has very low mass but the target distribution is well-supported. In this region, $P_S(x) \to 0$, which implies $p^*(x) \to 1$. Let $p^*(x) = 1 - \delta$ for some sufficiently small $\delta > 0$. The true weight is:

$$w^*(x) = \frac{1 - \delta}{\delta} \approx \frac{1}{\delta}.$$

Assume our classifier slightly overestimates the probability of the target class such that $\hat{p}(x) = 1 - \delta^2$. The absolute error of our classifier is:

$$|\hat{p}(x) - p^*(x)| = |(1 - \delta^2) - (1 - \delta)| = \delta - \delta^2 < \delta.$$

By choosing $\delta \leq \epsilon$, the classifier satisfies the bounded error condition. However, the estimated weight in this region becomes:

$$\hat{w}(x) = \frac{1 - \delta^2}{\delta^2} \approx \frac{1}{\delta^2}.$$

The deviation of the estimated weight from the true weight is:

$$|\hat{w}(x) - w^*(x)| = \left| \frac{1-\delta^2}{\delta^2} - \frac{1-\delta}{\delta} \right| = \frac{1-\delta}{\delta^2}.$$

As $\delta \to 0$, the difference between the estimated weight and the true weight grows as $\mathcal{O}(1/\delta^2)$.

Recall the empirical MMD objective in the RKHS $\mathcal{H}$ with feature map φ:

$$\mathrm{MMD}(P_S^{\hat{w}}, P_T) = \left\| \frac{1}{n} \sum_{i=1}^{n} \hat{w}(x_i)\varphi(x_i) - \frac{1}{m} \sum_{j=1}^{m} \varphi(z_j) \right\|_{\mathcal{H}}.$$

Because $\hat{w}(x_i)$ grows unboundedly as $\delta \to 0$ for points in this low-density region, the term $\hat{w}(x_i)\varphi(x_i)$ will dominate the empirical source mean. Thus, for any arbitrary bound M, there exists a sufficiently small δ and a sample containing such points such that $\mathrm{MMD}(P_S^{\hat{w}}, P_T) > M$.

This demonstrates that minimizing a classifier's probability estimation error does not safely bound the MMD or the variance of the weights, leading to degraded conformal prediction coverage. KMM explicitly resolves this by enforcing the bound $w_i \leq B$ and directly matching the moments. □ □

D MONDRIAN CONFORMAL PREDICTION

D.1 MONDRIAN CONFORMAL PREDICTION

Mondrian Conformal Prediction (MCP) extends standard conformal prediction by conditioning calibration on a predefined partition of the data Vovk [2012]. Instead of enforcing marginal coverage over the entire population, MCP guarantees coverage within each subgroup defined by a class label or other stratification variable.

Let $g(X) \in \mathcal{G}$ denote a grouping function. In classification tasks, a common choice is $g(X, Y) = Y$, leading to class-conditional calibration. For each group $c \in \mathcal{G}$, define the group-specific conformity scores

$$\{S_i : g(X_i, Y_i) = c\}.$$

Let n_c denote the number of calibration points in group c. The group-specific empirical CDF is

$$F_c(t) = \frac{1}{n_c} \sum_{i:g(X_i,Y_i)=c} \mathbf{1}\{S_i \leq t\}.$$

The Mondrian conformal threshold for group c is then

$$\hat{q}_{1-\alpha}^{(c)} = \inf \{t : F_c(t) \geq 1 - \alpha\}.$$

The resulting prediction set for a new input (X_{N+1}, Y_{N+1}) with group $c = g(X_{N+1})$ satisfies

$$\mathbb{P}\big(Y_{N+1} \in \hat{C}(X_{N+1}) \mid g(X_{N+1}) = c\big) \geq 1 - \alpha,$$

under exchangeability within each group.

In imbalanced molecular datasets, class-conditional calibration prevents dominant classes from overwhelming rare but important categories (e.g., active or toxic compounds).

D.2 WEIGHTED MONDRIAN CONFORMAL PREDICTION

Under covariate shift, we extend MCP by incorporating importance weights within each group. Let w_i denote the calibration weights (e.g., obtained via KMM or SKMM). Define the normalized group-specific weights

$$\tilde{w}_i^{(c)} = \frac{w_i}{\sum_{j:g(X_j,Y_j)=c} w_j}, \quad \text{for } g(X_i, Y_i) = c.$$

The weighted group CDF becomes

$$F_c^w(t) = \sum_{i:g(X_i,Y_i)=c} \tilde{w}_i^{(c)} \mathbf{1}\{S_i \leq t\}.$$

The Mondrian weighted quantile is

$$\hat{q}_{1-\alpha}^{(c),w} = \inf \{t : F_c^w(t) \geq 1 - \alpha\}.$$

The prediction set is constructed exactly as in the global weighted case, except that quantiles are computed separately within each group.

Practical Choice of Groups. In our experiments, we use class-conditional Mondrian calibration: $g(X, Y) = Y$. For toxicity datasets, we calibrate within the non-toxic class to avoid undercoverage of rare toxic compounds; for activity datasets, we calibrate within the active class. This choice reflects domain-specific asymmetry in decision importance. In Algorithm 2, we summarize the application of KMM-CP for class conditional guarentees under covariate shift.

Algorithm 2 Mondrian Split Conformal Prediction with KMM / Selective KMM

Require: Labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, unlabeled test covariates $\{z_j\}_{j=1}^{m}$, miscoverage level α_{conf}, kernel $k(\cdot, \cdot)$, weight bound B, selection fraction τ (optional)

Ensure: Prediction sets $\hat{C}(z_j)$

Split $\mathcal{D}$ into training set $\mathcal{D}_{\text{train}}$ and calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{n}$

Train predictor f on $\mathcal{D}_{\text{train}}$

Define conformity score $S(x, y)$ based on f

if Selective KMM is enabled **then**

  Solve double-weighted KMM to obtain selection weights α_j

  Define selected subset $\mathcal{T}_{\text{sel}}$

else

  Set $\mathcal{T}_{\text{sel}} = \{z_j\}_{j=1}^{m}$

end if

Solve standard KMM between $\mathcal{D}_{\text{cal}}$ and $\mathcal{T}_{\text{sel}}$ to obtain source weights w_i

for each class c **do**

  Let $\mathcal{I}_c = \{i : y_i = c\}$

  Normalize weights within class:

$$\tilde{w}_i^{(c)} = \frac{w_i}{\sum_{k \in \mathcal{I}_c} w_k} \quad \text{for } i \in \mathcal{I}_c$$

  Compute class-conditional quantile $\hat{q}_{1-\alpha}^{(c)}$:

$$\sum_{i \in \mathcal{I}_c} \tilde{w}_i^{(c)} \mathbf{1}\{S_i \le t\} \ge 1 - \alpha_{\text{conf}}$$

end for

for each $z_j \in \mathcal{T}_{\text{sel}}$ **do**

  Let $\hat{c} =$ predicted class of z_j

  Output

$$\hat{C}(z_j) = \{y : S(z_j, y) \le \hat{q}_{1-\alpha}^{(\hat{c})}\}$$

end for

E ADDITIONAL EXPERIMENTAL DETAILS

E.1 CODE AVAILABILITY

All code used to reproduce the experiments is publicly available at:

`https://github.com/siddharthal/KMM-CP`

The repository contains complete training scripts, data preprocessing pipelines, and exact random seeds for all five runs per experiment. All results reported in the paper are fully reproducible.

E.2 DATASET STATISTICS

Table 3 summarizes dataset sizes and splits.

Table 3: Dataset statistics and data splits.

Dataset	Total N	Train	Calib	Test
AMES	7,280	5,096	1,092	1,092
BBB-Martins	2,033	1,423	305	305
HIV	41,132	28,793	6,169	6,170
Tox21	7,267	5,087	1,090	1,090
hERG	13,445	9,411	2,017	2,017

E.3 TRAINING DETAILS AND HYPERPARAMETERS

All models use the AttentiveFP Xiong et al. [2019] graph neural network architecture implemented via the `dgllife` library Li et al. [2021], with training performed in PyTorch.

Optimization.

- Optimizer: Adam
- Learning rate: 1×10^{-3}
- Learning rate scheduler: StepLR (step_size=1, $\gamma = 0.8$)

Training Loop.

- Maximum epochs: 50
- Batch size: 128
- Early stopping: patience = 5 (based on training loss)

Model Architecture.

- Graph embedding dimension: 512
- Hidden dimension (classifier): 64
- GNN layers (AttentiveFP steps): 3
- Output dimension: 2 (binary classification)
- Total parameters: 10M

Training time Model training takes approximately 5–25 minutes per dataset on a single NVIDIA A100 GPU, depending on dataset size.

Loss Function.

- Cross-entropy loss
- Class weights enabled (inverse class frequency)

E.4 CONFORMAL PREDICTION PARAMETERS

Non-conformity Score. Non-conformity scores are defined using the predicted class probabilities from the trained AttentiveFP model.

Feature Representation. Penultimate-layer embeddings from the trained model are used as feature representations for density-ratio estimation and KMM.

E.5 IMPLEMENTATION OF KMM

Kernel Mean Matching (KMM) optimization is solved using `cvxopt` library.

Weight Constraints.

- Weight upper bound: $B = 30$
- Target selection fraction (SKMM): $\tau = 0.5$
- Second-stage filtering threshold on α: 0.2

These hyperparameters are fixed across all datasets and are not tuned per dataset.

Kernel Bandwidth Selection. We use an RBF kernel with bandwidth selected via a median-distance heuristic Gretton et al. [2012]. Specifically, we compute the median pairwise distance between calibration and test embeddings and evaluate bandwidths from the set:

$$\sigma \in \{0.01,\ 0.1,\ 0.5,\ 1,\ 2\} \times \text{median\_distance}.$$

The bandwidth maximizing standardized MMD (z-scored across permutations) is selected.

This strategy balances sensitivity to local discrepancies and stability under high-dimensional embeddings, and is commonly used in kernel two-sample testing.

Computation Time. KMM optimization is performed on CPUs and requires approximately:

- 6 minutes for the smallest dataset,
- up to 50 minutes for the largest dataset.

E.6 COMPUTE RESOURCES

All experiments were conducted on a cluster with:

- 4 NVIDIA A100 GPUs,
- 64 CPU cores.

E.7 IMPLEMENTATION OF DENSITY-RATIO BASELINES

KDE Baseline. For the KDE-based density-ratio baseline, we estimate source and target marginal densities using a Gaussian kernel. The bandwidth is selected via a held-out validation split by maximizing the log-likelihood over candidate values $\{0.01, 0.1, 1.0, 10\}$. The same bandwidth is used for both source and target density estimation. The density ratio is then formed as $\hat{w}(x) = \hat{p}_t(x)/\hat{p}_s(x)$.

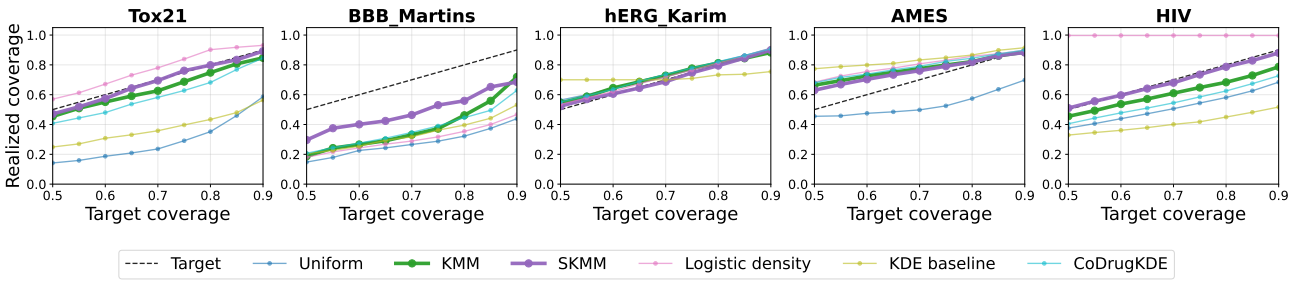


Figure 5: Calibration coverage under global conformal prediction across datasets.

Logistic Regression Baseline. For the classifier-based density-ratio baseline, we train a logistic regression classifier to distinguish calibration and test samples using 64-dimensional embeddings extracted from the final layer of the trained predictor. The classifier is implemented using the `scikit-learn` library with default regularization settings. Density ratios are obtained from the predicted probabilities via the standard odds transformation.

F ADDITIONAL RESULTS

F.1 CALIBRATION COVERAGE PLOTS FOR GLOBAL CP

Figure 5 presents calibration curves under covariate shift for global conformal prediction across all datasets. Each curve plots realized coverage against target coverage levels from 0.5 to 0.95, averaged over five random runs. The dashed line indicates perfect calibration.

Under covariate shift, uniform conformal prediction systematically undercovers, with deviations increasing at higher target coverage levels. Density-ratio baselines partially correct this behavior but exhibit residual miscalibration and variability across datasets. In particular, classifier- and KDE-based weighting methods often fail to fully align with the nominal coverage line at moderate-to-high coverage levels.

In contrast, KMM produces curves that more closely track the nominal diagonal, indicating improved distributional alignment. The proposed selective KMM (SKMM) further tightens this alignment, yielding the smallest deviations from nominal coverage across datasets. The improvement is consistent across both toxicity and activity prediction tasks, and persists across the full range of coverage levels.

These plots visually corroborate the mondran CP trends and ranking results reported in the main text.

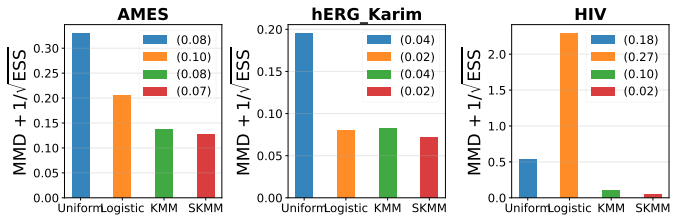
F.2 ADDITIONAL BIAS-VARIANCE AND SELECTIVE FILTERING PLOTS

In this section, we provide supplemental plots to support the analysis detailed in Section 6.3 for the remaining datasets: AMES, hERG, and HIV.Figure 6a presents the bias-variance decomposition across these benchmarks, while Figure 6b illustrates the corresponding α distribution plots. Overall, the bias-variance trends observed here closely mimic the results shown in Figure 3; specifically, the KMM-based methods continue to exhibit higher stability across varying shift intensities. Similarly, the α distributions demonstrate emergent sparsity across nearly all experimental settings. However, the hERG_Karim dataset exhibits a denser α distribution; this is potentially attributable to a higher intrinsic overlap in the covariate support.

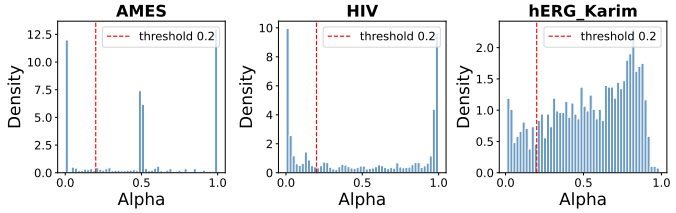
F.3 FILTERED FRACTION

Table 4 reports the average filtration ratio under global and Mondrian calibration, defined as the percentage of test instances retained after the selective KMM filtering stage (mean $\pm$ standard deviation across five runs). Recall that the target selection fraction is set to $\tau = 0.5$, corresponding to retaining approximately 50% of the test instances.

Across datasets, the observed retention rates range from roughly 58% to 77%, indicating that the learned selection variables α_j concentrate mass on a substantial but restricted subset of the test distribution. Also note that SKMM does not collapse to extreme filtering: a majority of test instances are retained across all benchmarks. This confirms that the method focuses



(a) Bias-Variance (AMES)



(b) α Distribution (AMES)

Figure 6: Figure. (a) illustrates the bias-variance proxy($\mathrm{MMD} + \sqrt{1/\mathrm{ESS}}$) , while (b) displays the emergent sparsity in the weight distribution.

Table 4: Average filtration ratio (%) under two-stage Selective KMM (mean $\pm$ std).

Dataset	Fraction Filtered
AMES	71.2 ± 5.9
Tox21	58.3 ± 0.5
BBB-Martins	60.1 ± 10.1
hERG-Karim	76.7 ± 5.9
HIV	62.2 ± 1.4

correction on regions where density-ratio estimation is stable, rather than aggressively discarding data.

F.4 FULL CALIBRATION RESULTS ACROSS DATASETS

The following tables report full calibration curves (mean $\pm$ std over five runs) for all datasets under both global and Mondrian conformal prediction across target coverages from 0.5 to 0.95. Best results per column are shown in **bold**, and second best in **gray bold**.

Overall trends. Across all datasets and coverage levels, SKMM consistently achieves the lowest or near-lowest deviation from nominal coverage. This pattern holds for both global and class-conditional (Mondrian) calibration. KMM (Beta-only) typically ranks second, while classifier- and KDE-based baselines exhibit larger systematic deviations, particularly at higher target coverage levels.

Global calibration. Under global conformal prediction, uniform weighting frequently exhibits pronounced undercoverage as the target coverage increases. Density-ratio baselines partially mitigate this effect but remain unstable across datasets. In contrast, KMM substantially improves calibration alignment, and SKMM further reduces deviation across nearly all coverage levels. The average rank columns confirm this consistency: SKMM achieves the best mean rank on every dataset under global calibration.

Mondrian calibration results show analogous results for Mondrian conformal prediction. The same qualitative behavior persists: SKMM maintains the strongest alignment with nominal coverage across coverage levels, followed by KMM. Importantly, improvements are not restricted to a specific coverage region; gains remain stable from moderate (0.6–0.7) to high (0.9–0.95) target coverage.

Table 5: AMES — Global calibration (mean ± std). Best is **bold**; second best is **gray bold**.

Method	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg. Rank
Uniform	**0.455±0.000**	**0.458±0.004**	**0.475±0.006**	0.485±0.004	0.498±0.005	0.525±0.014	0.574±0.034	0.636±0.040	0.698±0.034	0.824±0.048	4.60
KDE	0.774±0.127	0.788±0.113	0.799±0.104	0.810±0.092	0.833±0.081	0.848±0.073	0.866±0.051	0.898±0.046	0.915±0.036	**0.933±0.022**	4.80
CoDrugKDE	0.680±0.041	0.716±0.040	0.738±0.042	0.767±0.035	0.790±0.034	0.823±0.042	0.845±0.033	0.874±0.027	**0.899±0.033**	**0.933±0.027**	3.00
Logistic	0.684±0.058	0.725±0.068	0.755±0.051	0.778±0.037	0.807±0.030	0.831±0.036	0.858±0.044	0.875±0.028	**0.898±0.027**	0.925±0.030	4.10
KMM	0.664±0.026	0.694±0.025	0.726±0.021	**0.750±0.012**	**0.775±0.017**	**0.799±0.021**	**0.822±0.027**	**0.861±0.017**	0.885±0.024	0.915±0.019	2.60
SKMM	**0.630±0.028**	**0.669±0.024**	**0.703±0.019**	**0.734±0.009**	**0.761±0.010**	**0.791±0.020**	**0.819±0.029**	**0.862±0.022**	0.883±0.025	0.927±0.020	1.90

Table 6: AMES — Mondrian calibration (mean ± std). Best is **bold**; second best is **gray bold**.

Method	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg. Rank
Uniform	**0.470±0.006**	**0.499±0.014**	**0.530±0.025**	**0.569±0.035**	0.606±0.045	0.642±0.033	0.683±0.035	0.732±0.038	0.809±0.048	0.916±0.055	4.00
KDE	**0.585±0.080**	**0.605±0.083**	**0.613±0.094**	**0.642±0.096**	**0.658±0.112**	**0.704±0.090**	0.758±0.085	0.808±0.053	0.817±0.064	0.847±0.086	2.80
CoDrugKDE	0.685±0.031	0.710±0.034	0.735±0.039	0.764±0.038	0.792±0.032	0.820±0.028	0.838±0.033	0.865±0.031	**0.901±0.026**	0.926±0.030	4.10
Logistic	0.691±0.052	0.722±0.033	0.746±0.033	0.777±0.038	0.795±0.041	0.819±0.043	0.847±0.029	0.865±0.031	**0.894±0.020**	0.926±0.030	4.70
KMM	0.669±0.020	0.698±0.024	0.723±0.029	0.744±0.030	0.767±0.030	0.803±0.026	**0.834±0.029**	**0.855±0.027**	0.881±0.024	**0.926±0.028**	3.00
SKMM	0.655±0.013	0.680±0.020	0.706±0.022	0.740±0.026	**0.765±0.032**	**0.801±0.029**	**0.837±0.032**	**0.857±0.028**	0.889±0.028	**0.936±0.026**	2.40

Table 7: BBB_Martins — Global calibration (mean ± std). Best is **bold**; second best is **gray bold**.

Method	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg. Rank
Uniform	0.148±0.053	0.179±0.072	0.226±0.069	0.243±0.078	0.266±0.081	0.288±0.083	0.322±0.077	0.374±0.064	0.439±0.087	0.659±0.110	6.00
KDE	0.205±0.094	0.226±0.112	0.254±0.116	0.294±0.112	0.317±0.114	0.362±0.125	0.396±0.142	0.442±0.173	0.532±0.195	0.698±0.230	3.80
CoDrugKDE	**0.207±0.098**	**0.244±0.100**	**0.275±0.100**	**0.308±0.102**	**0.347±0.089**	**0.390±0.084**	0.446±0.093	0.495±0.108	0.629±0.080	0.770±0.055	2.40
Logistic	0.180±0.068	0.216±0.066	0.243±0.077	0.268±0.078	0.289±0.079	0.317±0.073	0.354±0.067	0.400±0.059	0.467±0.072	0.693±0.084	5.00
KMM	0.187±0.061	0.242±0.108	0.266±0.119	0.291±0.111	0.331±0.096	0.372±0.084	**0.462±0.125**	**0.559±0.110**	**0.719±0.112**	**0.838±0.038**	2.60
SKMM	**0.296±0.121**	**0.375±0.171**	**0.401±0.166**	**0.424±0.157**	**0.464±0.142**	**0.530±0.135**	**0.559±0.138**	**0.652±0.094**	**0.685±0.091**	**0.812±0.086**	1.20

Table 8: BBB_Martins — Mondrian calibration (mean ± std). Best is **bold**; second best is **gray bold**.

Method	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg. Rank
Uniform	0.189±0.005	0.202±0.002	0.208±0.003	0.216±0.004	0.222±0.004	0.236±0.010	0.252±0.010	0.287±0.022	0.367±0.069	0.683±0.088	5.70
KDE	0.214±0.015	0.226±0.022	0.241±0.020	**0.269±0.020**	0.282±0.025	0.308±0.040	0.343±0.072	0.408±0.086	0.473±0.121	0.632±0.173	3.90
CoDrugKDE	0.218±0.014	**0.235±0.020**	**0.248±0.026**	0.265±0.027	**0.291±0.041**	0.315±0.044	0.357±0.067	0.432±0.037	0.511±0.062	**0.687±0.075**	2.50
Logistic	0.202±0.003	0.209±0.006	0.218±0.005	0.226±0.011	0.237±0.013	0.252±0.019	0.272±0.018	0.311±0.047	0.386±0.065	0.658±0.090	5.00
KMM	**0.219±0.013**	0.227±0.018	0.242±0.015	0.264±0.023	0.279±0.013	**0.353±0.078**	**0.437±0.027**	**0.496±0.070**	**0.604±0.089**	**0.686±0.085**	2.50
SKMM	**0.272±0.037**	**0.312±0.031**	**0.329±0.028**	**0.383±0.091**	**0.433±0.094**	**0.453±0.112**	**0.510±0.112**	**0.575±0.082**	**0.604±0.092**	0.672±0.095	1.40

Table 9: HIV — Global calibration (mean ± std). Best is **bold**; second best is **gray bold**.

Method	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg. Rank
Uniform	0.376±0.063	0.406±0.063	0.439±0.064	0.472±0.064	0.506±0.064	0.544±0.063	0.581±0.065	0.626±0.065	0.684±0.061	0.779±0.049	4.40
KDE	0.329±0.124	0.346±0.126	0.361±0.132	0.379±0.137	0.401±0.144	0.418±0.160	0.450±0.168	0.482±0.188	0.517±0.203	0.577±0.255	5.60
CoDrugKDE	0.405±0.101	0.443±0.105	0.478±0.110	0.510±0.108	0.545±0.112	0.585±0.121	0.625±0.116	0.674±0.110	0.727±0.089	0.803±0.091	3.30
Logistic	0.997±0.002	0.997±0.002	0.997±0.002	0.997±0.002	0.997±0.002	0.997±0.002	0.997±0.002	0.997±0.002	**0.997±0.002**	**0.997±0.002**	4.50
KMM	**0.455±0.058**	**0.493±0.056**	**0.537±0.048**	**0.571±0.042**	**0.610±0.042**	**0.648±0.033**	**0.683±0.031**	**0.729±0.035**	0.787±0.031	0.863±0.034	2.20
SKMM	**0.510±0.053**	**0.556±0.056**	**0.596±0.050**	**0.641±0.052**	**0.680±0.051**	**0.735±0.043**	**0.787±0.040**	**0.829±0.034**	**0.880±0.031**	**0.945±0.028**	1.00

Table 10: HIV — Mondrian calibration (mean ± std). Best is **bold**; second best is **gray bold**.

Method	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg. Rank
Uniform	0.385±0.063	0.415±0.064	0.448±0.063	0.480±0.066	0.512±0.067	0.546±0.068	0.581±0.068	0.619±0.067	0.667±0.064	0.732±0.060	4.40
KDE	0.341±0.124	0.358±0.123	0.373±0.130	0.392±0.132	0.411±0.139	0.427±0.150	0.457±0.159	0.479±0.171	0.515±0.177	0.552±0.201	5.60
CoDrugKDE	0.409±0.096	0.445±0.103	0.481±0.106	0.509±0.109	0.544±0.106	0.574±0.107	0.615±0.119	0.659±0.115	0.705±0.112	0.769±0.097	3.40
Logistic	0.969±0.001	0.969±0.001	0.969±0.001	0.969±0.001	0.969±0.001	0.969±0.001	0.969±0.001	**0.969±0.001**	**0.969±0.001**	**0.972±0.009**	4.20
KMM	**0.452±0.059**	**0.488±0.060**	**0.527±0.057**	**0.565±0.048**	**0.600±0.044**	**0.640±0.039**	**0.674±0.035**	0.714±0.038	0.759±0.032	0.826±0.030	2.30
SKMM	**0.505±0.059**	**0.548±0.054**	**0.593±0.053**	**0.631±0.048**	**0.676±0.053**	**0.713±0.048**	**0.770±0.040**	**0.816±0.043**	**0.861±0.034**	**0.910±0.031**	1.10

Table 11: Tox21 — Global calibration (mean ± std). Best is **bold**; second best is **gray bold**.

Method	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg. Rank
Uniform	0.142±0.014	0.159±0.012	0.187±0.018	0.209±0.018	0.236±0.019	0.291±0.020	0.352±0.009	0.460±0.015	0.586±0.010	0.760±0.029	5.80
KDE	0.249±0.021	0.270±0.023	0.308±0.023	0.331±0.023	0.358±0.017	0.397±0.023	0.434±0.032	0.480±0.037	0.564±0.028	0.709±0.070	5.20
CoDrugKDE	0.408±0.117	0.444±0.116	0.480±0.125	0.537±0.129	0.582±0.145	0.628±0.132	0.682±0.137	0.769±0.066	0.846±0.038	0.906±0.027	3.90
Logistic	0.570±0.063	0.613±0.076	0.671±0.094	0.731±0.117	0.780±0.101	0.839±0.071	0.902±0.036	0.918±0.034	**0.931±0.021**	**0.940±0.009**	2.80
KMM	**0.454±0.027**	**0.509±0.021**	**0.551±0.005**	**0.591±0.008**	**0.626±0.015**	**0.687±0.006**	**0.748±0.011**	**0.807±0.013**	0.847±0.015	0.905±0.013	2.30
SKMM	**0.471±0.024**	**0.519±0.014**	**0.574±0.033**	**0.643±0.069**	**0.695±0.056**	**0.761±0.028**	**0.798±0.018**	**0.834±0.024**	**0.892±0.014**	**0.956±0.004**	1.00

Table 12: Tox21 — Mondrian calibration (mean ± std). Best is **bold**; second best is **gray bold**.

Method	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg. Rank
Uniform	0.137±0.017	0.154±0.013	0.175±0.010	0.202±0.017	0.225±0.019	0.257±0.019	0.317±0.013	0.386±0.009	0.518±0.014	0.623±0.011	6.00
KDE	0.247±0.019	0.268±0.021	0.305±0.023	0.326±0.029	0.353±0.016	0.386±0.018	0.429±0.032	0.469±0.030	0.542±0.039	0.669±0.078	5.00
CoDrugKDE	0.399±0.121	0.438±0.115	0.473±0.126	0.523±0.130	0.575±0.149	0.627±0.145	0.678±0.139	0.751±0.104	0.845±0.038	0.906±0.027	4.00
Logistic	0.590±0.062	0.636±0.086	0.692±0.100	0.760±0.108	0.807±0.087	0.873±0.041	0.903±0.034	0.916±0.033	**0.925±0.020**	**0.934±0.008**	2.80
KMM	**0.448±0.022**	**0.504±0.033**	**0.546±0.005**	**0.585±0.017**	**0.616±0.010**	**0.675±0.025**	**0.743±0.012**	**0.809±0.012**	0.860±0.016	0.907±0.014	2.20
SKMM	**0.466±0.025**	**0.521±0.013**	**0.570±0.029**	**0.627±0.051**	**0.688±0.048**	**0.767±0.017**	**0.796±0.018**	**0.830±0.016**	**0.891±0.015**	**0.951±0.003**	1.00

Table 13: hERG_Karim — Global calibration (mean ± std). Best is **bold**; second best is **gray bold**.

Method	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg. Rank
Uniform	0.564±0.026	0.600±0.024	0.643±0.034	0.692±0.032	0.734±0.037	0.776±0.045	0.820±0.044	0.862±0.037	0.912±0.024	0.958±0.014	4.30
KDE	0.700±0.180	0.700±0.180	0.700±0.180	0.700±0.180	**0.700±0.180**	0.709±0.170	0.733±0.151	0.737±0.149	0.755±0.157	0.807±0.177	5.50
CoDrugKDE	0.551±0.037	0.592±0.036	0.636±0.045	0.690±0.042	0.737±0.057	0.779±0.060	0.822±0.053	0.863±0.046	**0.907±0.032**	**0.949±0.021**	3.90
Logistic	**0.535±0.043**	**0.578±0.039**	**0.622±0.043**	**0.674±0.037**	0.719±0.028	**0.765±0.030**	0.813±0.035	0.861±0.027	0.910±0.027	0.957±0.015	2.50
KMM	0.545±0.027	0.587±0.038	0.645±0.045	0.686±0.057	0.728±0.055	0.776±0.055	**0.812±0.060**	**0.844±0.074**	0.883±0.066	0.930±0.065	3.60
SKMM	**0.523±0.046**	**0.570±0.053**	**0.606±0.067**	**0.645±0.067**	**0.687±0.058**	**0.747±0.053**	**0.795±0.046**	**0.846±0.028**	**0.899±0.018**	**0.944±0.016**	1.20

Table 14: hERG_Karim — Mondrian calibration (mean ± std). Best is **bold**; second best is **gray bold**.

Method	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg. Rank
Uniform	0.594±0.044	0.629±0.040	0.664±0.037	0.696±0.036	0.729±0.037	0.768±0.038	0.812±0.032	**0.853±0.019**	**0.899±0.015**	**0.941±0.009**	4.30
KDE	**0.545±0.129**	**0.545±0.129**	**0.591±0.115**	0.612±0.115	0.620±0.120	0.642±0.125	0.642±0.125	0.662±0.140	0.684±0.137	0.716±0.071	4.30
CoDrugKDE	0.571±0.043	0.612±0.045	0.649±0.042	0.681±0.044	0.720±0.045	0.759±0.041	**0.801±0.035**	**0.840±0.026**	0.887±0.014	0.929±0.010	3.20
Logistic	**0.546±0.046**	0.590±0.043	0.631±0.045	**0.675±0.036**	**0.716±0.033**	**0.758±0.030**	0.811±0.024	0.861±0.018	**0.909±0.018**	**0.950±0.011**	2.50
KMM	0.572±0.058	0.609±0.057	0.656±0.062	0.694±0.072	0.726±0.075	0.764±0.085	**0.802±0.090**	0.831±0.093	0.868±0.089	0.914±0.069	4.40
SKMM	0.548±0.078	**0.589±0.074**	**0.625±0.069**	**0.666±0.067**	**0.706±0.063**	**0.753±0.062**	0.797±0.055	0.839±0.040	0.884±0.029	0.931±0.023	2.30