Conformal Prediction for Distribution-to-Distribution Regression

Trung-Khang Tran*
School of Computing
National University of Singapore
ttk0106@u.nus.edu

Tuan A. Hoang*
Center for AI Research
VinUniversity
tuan.ha3@vinuni.edu.vn

Viet-Hoang Tran*
Department of Mathematics
National University of Singapore
hoang.tranviet@u.nus.edu

Tan M. Nguyen
Department of Mathematics
National University of Singapore
tanmn@nus.edu.sg

Abstract

Conformal Prediction (CP) provides a model-agnostic framework for constructing finite-sample prediction sets with coverage guarantees and has become a tool for reliable decision-making in high-risk settings. Recent advances have extended Split Conformal Prediction to increasingly complex learning problems, including multilabel, high-dimensional, and functional outputs. In this work, we study CP for Distribution-to-Distribution Regression, a setting of growing interest across many application domains. We show that a naive application of Split CP yields prediction sets that, while marginally valid, are non-interpretable, e.g., lacking a well-defined volume or a tractable sampling procedure, and of limited practical value. To overcome this, we propose a new conformal framework that operates in an orthogonal basis space, producing interpretable prediction sets while preserving coverage guarantees. We further develop an adaptive variant with asymptotic conditional coverage under mild assumptions. Empirical results on synthetic and real-world datasets validate the effectiveness of the proposed approach. The code is publicly available at https://github.com/tuanhacs/conformal_d2d.

1 Introduction

Conformal Prediction (CP) [49, 67, 29, 60, 42, 52] is a model-agnostic and distribution-free framework that quantifies predictive uncertainty by constructing prediction sets with rigorous coverage guarantees, ensuring the true outcome lies within the set at a user-specified confidence level. CP has been widely adopted across many domains, including medical imaging [34, 36, 27], classical machine learning [4, 24], deep learning [51, 69, 73, 71], diffusion models [59, 57, 68, 22], large language models [10, 66], image segmentation [15, 41, 6, 11, 70], and federated learning [35, 30, 37, 53, 23].

Recent work has extended CP to more complex settings, where uncertainty is inherently greater. Notable examples include conformalized quantile regression [54], risk control [3], multi-label classification [48, 46, 62], multivariate regression [26, 60], and functional data analysis [28]. A key challenge in these extensions is preserving the interpretability of CP sets, particularly their ability to quantify uncertainty through volume and support uniform sampling for downstream tasks. While interpretability can often be maintained within the standard *Split Conformal Prediction* framework [47], naive applications may yield poorly interpretable prediction sets in certain settings. One

*These authors contributed equally. Correspondence to: ttk0106@u.nus.edu & tanmn@nus.edu.sg.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

such case is *Distribution-to-Distribution Regression (DDR)* [45], where both inputs and outputs are distributions observed only through finite samples. At test time, the goal is to infer an output distribution from samples drawn from an input distribution. Despite extensive study in prior work, including [20, 25, 58, 9, 56, 32, 16, 44, 8, 7], the intrinsic complexity of distribution-valued data and limited samples per distribution fundamentally constrain predictive accuracy, limiting reliable uncertainty quantification and interpretability of standard Split Conformal Prediction approaches.

Contribution. Motivated by this gap, we propose a CP framework for the DDR setting that ensures coverage guarantees, thereby enhancing confidence in distributional predictions while preserving the interpretability inherent to CP methods. We further introduce an adaptive variant of the framework that provides asymptotic conditional coverage guarantees in the DDR setting, enabling valid uncertainty quantification when conditioning on a given input distribution. The paper is organized as follows:

1. In Section 3, we examine a naive Split Conformal Prediction approach for DDR that follows existing frameworks in the literature, and we rigorously demonstrate why such an approach fails to preserve interpretability.
2. In Section 4, we present our CP framework for DDR. By projecting distributions into a vector space, we show that the resulting framework ensures coverage guarantees while remaining interpretable.
3. In Section 5, we propose an adaptive extension of our framework that provides asymptotic coverage guarantees conditional on the input distribution under mild assumptions.
4. In Section 6, we report empirical results on both synthetic and real datasets across multiple DDR baselines for both the standard and adaptive CP frameworks. We additionally present experiments illustrating the relationship between coverage and calibration set size, as well as a comparison of inference time between the two CP frameworks. We also study the effect of the low-data regime on our framework.

Section 2 provides background on CP and the DDR problem. A table of notation, background on CP, theoretical foundations, and additional experimental details are provided in the Appendix.

2 Preliminaries

This section provides a brief overview of Conformal Prediction (CP) and establishes the problem formulation and notation for Distribution-to-Distribution Regression (DDR). *We refer to Appendix A for more details on CP, and Appendix A.4 for applications of CP in DDR.*

Conformal Prediction (CP). Let $\hat{f} : \mathcal{X} \to \mathcal{Y}$ be a *prediction model*. Given a *target miscoverage level* $\gamma \in (0, 1)$, CP aims to construct a *set-valued predictor* $\mathcal{Q} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ such that, for a test pair $(X_0, Y_0) \in \mathcal{X} \times \mathcal{Y}$, we obtain the following *coverage guarantee*:

$$\mathbb{P}(Y_0 \in \mathcal{Q}(X_0)) \geq 1 - \gamma. \tag{1}$$

A widely used conformal prediction strategy is *Split Conformal Prediction* [47]. In this framework, the available data are partitioned into a training set and an independent *calibration set*. The predictive model $\hat{f}$ is trained exclusively on the training set, while the calibration set, unseen during training, is used to evaluate a *nonconformity score* that quantifies the discrepancy between the model prediction $\hat{f}(X)$ and the observed response Y. In particular, let d_0 denote the size of the calibration set, and let $\{s_i\}_{i=1}^{d_0}$ be the corresponding nonconformity scores, with

$$s_i \coloneqq s\big(\hat{f}(X_i^c), Y_i^c\big), \quad (X_i^c, Y_i^c) \in \text{Calibration Set}.$$

The choice of the score function s is problem-dependent and reflects the data structure and uncertainty relevant to the task [51, 2, 1]. A threshold is then constructed by selecting the empirical quantile

$$\hat{q} \coloneqq \text{the } \frac{\lceil (d_0 + 1)(1 - \gamma) \rceil}{d_0} \text{ quantile of } \{s_i\}_{i=1}^{d_0},$$

equivalently, the $\lceil (d_0 + 1)(1 - \gamma) \rceil$-th order statistic of $\{s_i\}_{i=1}^{d_0}$. The *prediction set* is defined as

$$\mathcal{Q}(X_0) \coloneqq \big\{ Y \in \mathcal{Y} : s(\hat{f}(X_0), Y) \leq \hat{q} \big\},$$

which guarantees the marginal coverage in Equation (1).

Distribution-to-Distribution Regression (DDR). Let $\mathcal{X}$ denote a class of input distributions supported on $\Psi^k \subseteq \mathbb{R}^k$, and let $\mathcal{Y}$ denote a class of output distributions supported on $\Lambda^l \subseteq \mathbb{R}^l$. We assume that each distribution in $\mathcal{X}$ and $\mathcal{Y}$ admits a density with respect to the Lebesgue measure. Importantly, the elements of $\mathcal{X}$ and $\mathcal{Y}$ are not restricted to probability distributions and therefore need not be everywhere nonnegative nor integrate to one. The goal of DDR is to learn a functional mapping $f : \mathcal{X} \to \mathcal{Y}$. To estimate f, we consider a collection of paired distributions $\{(X_1, Y_1), \ldots, (X_d, Y_d)\}$, where $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$. Each input distribution is assumed to be sampled independently and identically from a probability measure $\mathcal{P}$ defined on $\mathcal{X}$, that is, $X_1, \ldots, X_d \overset{\text{i.i.d.}}{\sim} \mathcal{P}$. Due to the nature of DDR in real-world settings, the distributions X_i and Y_i cannot be observed directly for all $i \in \{1, \ldots, d\}$. Instead, we observe independent and identically distributed samples from each distribution. Specifically, the available dataset consists of paired input/output samples $\mathcal{D} = \{(\bar{M}_1, \bar{N}_1), \ldots, (\bar{M}_d, \bar{N}_d)\}$, where $\bar{M}_i = \{M_{i1}, \ldots, M_{im_i}\}$ with $M_{ij} \overset{\text{iid}}{\sim} X_i$, and $\bar{N}_i = \{N_{i1}, \ldots, N_{in_i}\}$ with $N_{ij} \overset{\text{iid}}{\sim} Y_i$. Finally, let $\{x_i\}_{i=1}^d$ and $\{y_i\}_{i=1}^d$ denote the probability density functions corresponding to the distributions $\{X_i\}_{i=1}^d$ and $\{Y_i\}_{i=1}^d$, respectively. Throughout this work, we adopt a general, nonparametric setting with minimal assumptions on the data. In particular, we make only weak assumptions on the structure of the input and output distributions and the underlying measure from which the inputs are sampled.

3 Limitations of Naive Split CP for DDR

In this section, we analyze a naive extension of Split CP to DDR and show that it fails to preserve the interpretability of the resulting prediction sets. We also discuss why interpretability is essential in CP.

Split CP and Interpretability in Classical Settings. In most classical applications, CP can be applied directly via the Split CP framework across diverse learning settings and output spaces. Early work studied classification with discrete outputs, where the prediction set $\mathcal{Q}$ lies in a finite label space [55, 50, 2]. Later extensions considered one-dimensional regression with real-valued outputs, yielding interval-valued sets [24, 54]. Recent work further extends CP to vector-valued inputs and outputs, producing geometrically structured sets in higher dimensions [26, 60]. Despite differences in data modality and output structure across these settings, the resulting CP sets remain *interpretable*: Their volume can be explicitly quantified and uniform sampling from $\mathcal{Q}$ is feasible. This interpretability enables meaningful uncertainty quantification and is essential to CP, as discussed later in this section.

Failure of Naive Split CP in DDR. When inputs and outputs are further generalized to distribution-valued data, a naive application of Split CP fails to preserve the interpretability properties described above. This occurs because the resulting CP sets reside in infinite-dimensional function spaces, where notions such as volume, geometric structure, and uniform sampling are no longer well defined or computationally tractable.

We now illustrate this issue formally. Consider a model $f : \mathcal{X} \to \mathcal{Y}$. Following the standard CP procedure, the dataset $\mathcal{D}$ is split into a training set $\mathcal{D}_t$ of size d_t and a calibration set $\mathcal{C}$ of size d_0. The model is first trained on $\mathcal{D}_t$ to obtain a fitted predictor $\hat{f}$. Then, for each calibration pair $(\bar{M}_i, \bar{N}_i) \in \mathcal{C}$, the underlying input and response distributions are approximated through some approximation function ψ, yielding $\tilde{X}_i := \psi(\bar{M}_i)$, and $\tilde{Y}_i := \psi(\bar{N}_i)$. The calibration nonconformity score is then computed as $s_i := s\Big(\hat{f}(\tilde{X}_i), \tilde{Y}_i\Big)$, where $s : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ measures the discrepancy between the predicted and observed outputs. We denote $\hat{q}$ as the $\lceil (d_0 + 1)(1 - \gamma) \rceil / d_0$ quantile of $\{s_i\}_{i=1}^{d_0}$. Given a test input $\bar{M}_0$ with associated (unobserved) output $\bar{N}_0$, the resulting CP set is

$$\mathcal{Q}_{\text{Naive}}(\bar{M}_0) := \big\{Y \in \mathcal{Y} : s\big(\hat{f}(\tilde{X}_0), Y\big) \leq \hat{q}\big\}. \tag{2}$$

The naive prediction set in Equation (2) still satisfies the marginal coverage guarantee of CP:

$$\mathbb{P}\left(Y_0 \in \mathcal{Q}_{\text{Naive}}(\bar{M}_0)\right) \geq 1 - \gamma.$$

However, the naive CP set $\mathcal{Q}_{\text{Naive}}$ is no longer interpretable in the DDR setting. First, although the supports of $\mathcal{X}$ and $\mathcal{Y}$ lie in Euclidean spaces, the induced spaces of admissible distributions are irregular subsets of infinite-dimensional function spaces. These sets need not be connected or smoothly parametrized, which makes fundamental geometric quantities, such as volume, ill-defined or computationally intractable. Second, the prediction set $\mathcal{Q}_{\text{Naive}}$ consists of distributions rather than

finite- or vector-valued objects. Consequently, sampling an element from $\mathcal{Q}_{\text{Naive}}$ requires sampling an entire distribution, for which no canonical notion of uniformity exists in functional spaces.

Interpretability in CP. Interpretability is a central component of CP, as prediction sets are intended not only to provide coverage guarantees but also to offer meaningful characterizations of predictive uncertainty. In classical settings, the geometry of the conformal set, such as its size, shape, or volume, serves as a direct measure of uncertainty [33], while the ability to sample uniformly from the set supports practical downstream tasks, including visualization [27], simulation [31], and decision-making [65] under uncertainty. When these interpretability properties are lost, conformal prediction sets become difficult to use in practice, even if formal coverage guarantees are retained.

4 An Interpretable CP Framework for DDR

This section provides a CP framework for the DDR problem that satisfies the coverage guarantee in Equation (1). The framework also yields prediction sets with volume characterization and supports uniform sampling. We further analyze the role of the calibration set size and establish finite-sample results on the variability of conditional coverage. We summarize our method in Algorithm 1 in the Appendix.

4.1 Construction of Our Framework

Our construction has four steps: we expand the output distributions in a basis, impose regularity on the resulting coefficients, estimate these coefficients from finite samples, and construct CP sets in coefficient space, then map them back to distributions.

Basis Expansion of Output Distributions. We begin by specifying a functional representation for distributions on Λ^l. To this end, we first fix an orthonormal basis $\{\beta_i\}_{i\in\mathbb{Z}}$ of $L_2(\Lambda)$. For any integer $l \geq 1$, this basis naturally induces a tensor-product basis on the product space Λ^l. Specifically, we construct the collection $\{\beta_\alpha\}_{\alpha\in\mathbb{Z}^l}$ by defining

$$\beta_\alpha(z) \coloneqq \textstyle\prod_{i=1}^{l} \beta_{\alpha_i}(z_i), \text{ where } z = (z_1, \ldots, z_l) \in \Lambda^l.$$

It is well known that the collection $\{\beta_\alpha\}_{\alpha\in\mathbb{Z}^l}$ forms an orthonormal basis of $L_2(\Lambda^l)$, with inner products satisfying $\langle \beta_\alpha, \beta_\gamma \rangle = \mathbb{I}_{\{\alpha=\gamma\}}$. We next describe how an arbitrary distribution in $\mathcal{Y}$ can be expressed in this basis. Given any $Y \in \mathcal{Y}$ that admits a density $y \in L_2(\Lambda^l)$ with respect to the Lebesgue measure, the density y admits the orthogonal expansion $y(z) \coloneqq \sum_{\alpha\in\mathbb{Z}^l} \theta_\alpha(Y)\, \beta_\alpha(z)$, for all $z \in \Lambda^l$, where the expansion coefficients $\theta_\alpha(Y) \in \mathbb{R}$ are given by

$$\theta_\alpha(Y) \coloneqq \int_{\Lambda^l} \beta_\alpha(z)\, y(z)\, dz = \int_{\Lambda^l} \beta_\alpha(z)\, dY(z).$$

Regularity Assumption. We next impose an anisotropic Sobolev ellipsoid assumption on the projection coefficients as stated in Assumption 4.1.

Assumption 4.1. *Assume that the projection coefficients satisfy an anisotropic Sobolev ellipsoid condition. In particular, let $\theta(Y) = \{\theta_\alpha(Y)\}_{\alpha\in\mathbb{Z}^l}$ denote the collection of projection coefficients associated with a distribution Y. We assume $\mathcal{Y} = \{Y : \theta(Y) \in \Theta(\nu, \sigma, \bar{A})\}$, where*

$$\Theta(\nu, \sigma, \bar{A}) = \Big\{ \{\theta_\alpha\}_{\alpha\in\mathbb{Z}^l} : \textstyle\sum_{\alpha\in\mathbb{Z}^l} \theta_\alpha^2 w_\alpha^2(\nu, \sigma) < \bar{A} \Big\},$$

and the anisotropic weight function is given by $w_\alpha^2(\nu, \sigma) = \sum_{i=1}^{l} (\nu_i |\alpha_i|)^{2\sigma_i}$, where $\nu_i, \sigma_i, \bar{A} > 0$.

Empirical Approximation of Output Distributions. In practice, the true distribution Y_i is not directly observed; instead, one has access only to a finite sample drawn from Y_i. Let $\widehat{Y}_i$ denote the empirical distribution constructed from n_i observations by assigning equal mass to each data point. Concretely, $\widehat{Y}_i$ is defined such that $\widehat{Y}_i(N = N_{ij}) = 1/n_i$, for all $j = 1, \ldots, n_i$. We define an estimator of the density y_i using a truncated orthogonal series expansion. The truncation serves to balance the approximation error in accordance with Assumption 4.1. Specifically, we define $\tilde{y}_i(z) \coloneqq \sum_{\alpha\in T_R} \theta_\alpha(\widehat{Y}_i)\, \beta_\alpha(z)$, where the empirical projection coefficients are given by

$$\theta_\alpha(\widehat{Y}_i) = \int_{\Lambda^l} \beta_\alpha(z)\, d\widehat{Y}_i(z) = \frac{1}{n_i} \sum_{j=1}^{n_i} \beta_\alpha(N_{ij}).$$

Here, the index set $T_R = \{\alpha \in \mathbb{Z}^l : 0 < w_\alpha(\nu, \sigma) \leq R\}$ corresponds to a computational truncation that restricts the expansion to finitely many basis functions. Under the anisotropic Sobolev ellipsoid assumption and with an appropriate choice of the truncation parameter R, the resulting estimator achieves an optimal mean-squared error rate. In particular, it can be shown that $\mathbb{E}\left[\|\tilde{y}_i - y_i\|_2^2\right] = \mathcal{O}\Big(n_i^{-2/\left(2+\sum_j 1/\sigma_j\right)}\Big)$, where the rate holds for $n_i \to \infty$ [43].

Construction of Prediction Set. We next describe the conformal calibration step. To this end, we partition the available data and designate a subset as the calibration set, $\mathcal{C} = \{(X_1^c, Y_1^c), \ldots, (X_{d_0}^c, Y_{d_0}^c)\}$. As in the training phase, the latent distributions $\{(X_i^c, Y_i^c)\}_{i=1}^{d_0}$ are not directly observed. Instead, we observe finite samples drawn from these distributions, denoted by $\{(\bar{M}_1^c, \bar{N}_1^c), \ldots, (\bar{M}_{d_0}^c, \bar{N}_{d_0}^c)\}$. Let $\hat{f}$ denote a fitted model obtained from the training data. For each $i \in [d_0]$, we define our score as

$$s_i := s(\hat{f}(X_i^c), Y_i^c) = g\Big(W^{1/2}\big(\theta(\hat{f}(X_i^c)) - \theta(\tilde{y}_i^c)\big)\Big),$$

where $g : \mathbb{R}^{|T_R|} \to \mathbb{R}_{\geq 0}$ is a metric on the coefficient space, and W is a weighting matrix that encodes the relative importance of discrepancies across basis components. For this work, we take $W = \operatorname{diag}\big(w_\alpha^2(\nu, \sigma)\big)$, and $g = \|\cdot\|_2$ for consistency with **Assumption** 4.1.

We next determine the conformal threshold. Let $\hat{q}$ denote the $\lceil (d_0+1)(1-\gamma) \rceil / d_0$ quantile of the empirical distribution of the calibration scores $\{s_i\}_{i=1}^{d_0}$. This threshold determines the admissible level of deviation between predicted and observed coefficients. For a new input distribution X_0, we first construct a prediction set in the coefficient space,

$$\mathcal{E}(X_0) = \Big\{a \in \mathbb{R}^{|T_R|} : g\Big(W^{1/2}\big(\theta(\hat{f}(X_0)) - a\big)\Big) \leq \hat{q}\Big\}.$$

Finally, we map the coefficient-level set back to distribution space and define the CP set as

$$\mathcal{Q}(X_0) = \Big\{y_a(z) = \textstyle\sum_{\alpha \in T_R} a_\alpha \beta_\alpha(z) : a \in \mathcal{E}(X_0)\Big\}.$$

Discussion of Assumption 4.1. This assumption specifies that each output distribution $Y \in \mathcal{Y}$ admits a representation whose projection coefficients $\theta(Y) = \{\theta_\alpha(Y)\}_{\alpha \in \mathbb{Z}^l}$ have finite energy in an anisotropic Sobolev sense. Concretely, this requires the weighted sum $\sum_{\alpha \in \mathbb{Z}^l} \theta_\alpha^2 \, w_\alpha^2(\nu, \sigma)$ to be bounded, which enforces decay of high-frequency components at rates that may differ across coordinates. This condition effectively restricts Y to be sufficiently smooth, while allowing heterogeneous regularity across dimensions. Hence, imposing this assumption amounts to requiring only that the output distributions exhibit reasonable, possibly direction-dependent smoothness, an assumption that is both mild and well aligned with real-world distributional settings.

4.2 Theoretical Properties

Coverage Guarantee. We show that the coverage guarantee in Equation (1) holds for both the coefficient-level prediction set $\mathcal{E}$ and the induced distribution-level prediction set $\mathcal{Q}$. This result confirms that the proposed construction satisfies the fundamental validity requirement of CP.

Theorem 4.2. *For any distribution $X_0 \in \mathcal{X}$, let Y_0 denote the corresponding output distribution, with approximation density $\tilde{y}_0$. Under the exchangeability of the data, the following coverage guarantees hold for any $\gamma \in (0, 1)$:*

$$1-\gamma \;\leq \mathbb{P}(\tilde{y}_0 \in \mathcal{Q}(X_0)) \;\leq\; 1-\gamma+\frac{\bar{d}}{1+d_0}, \quad 1-\gamma \;\leq\; \mathbb{P}(\theta(\tilde{y}_0) \in \mathcal{E}(X_0)) \;\leq\; 1-\gamma+\frac{\bar{d}}{1+d_0},$$

where $\bar{d}$ denotes the maximum number of ties in the set $\{s_1, s_2, \ldots, s_{d_0}, s_0\}$, with $s_0 = s(\hat{f}(X_0), Y_0)$.

The proof of Theorem 4.2 is deferred to Appendix B.1. The coverage statements are equivalent:

$$\tilde{y}_0 \in \mathcal{Q}(X_0) \;\text{ if and only if }\; \theta(\tilde{y}_0) \in \mathcal{E}(X_0), \tag{3}$$

which follows from the definition of the prediction sets and the one-to-one correspondence between coefficient vectors and their associated truncated density representations.

Volume Characterization. We next characterize the volume of the proposed prediction set. Since $\mathcal{E}(X_0)$ is an ellipsoid in coefficient space, its volume has a closed form. By Equation (3), $\mathcal{Q}(X_0)$ is the image of $\mathcal{E}(X_0)$ under the linear coefficient-to-density map, which is an isometry on the truncated basis span. Hence, $\mathcal{Q}(X_0)$ and $\mathcal{E}(X_0)$ have identical induced volume, yielding the following result.

Theorem 4.3 (Volume)**.** *Let $t = |T_R|$, and $g \equiv \|\cdot\|_2$. Recall the Gamma function, $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$. The volume of the prediction set $\mathcal{Q}(X_0)$ is given by*

$$\mathrm{Vol}\left(\mathcal{Q}(X_0)\right) = \frac{\left(\sqrt{\pi}\hat{q}\right)^t}{\left(\Gamma\left(\frac{t}{2}+1\right)\sqrt{\det W}\right)}.$$

The proof of Theorem 4.3 is given in Appendix B.2. Since $w_\alpha(\nu,\sigma)$ admits a uniform lower bound by a constant m depending only on ν and σ, we obtain the following as an immediate consequence.

Corollary 4.4. *Using the same notation as in Theorem 4.3, for $c := \sqrt{\pi}/m$, we have*

$$\mathrm{Vol}\left(\mathcal{Q}(X_0)\right) = \mathcal{O}\left(\left(c\hat{q}\right)^t/\Gamma\left(t/2+1\right)\right).$$

Corollary 4.4 shows that the volume of the CP set depends solely on the quantile $\hat{q}$ and truncation R.

Uniform Sampling. Since $\mathcal{E}(X_0)$ is an ellipsoid in the coefficient space, uniform samples can be generated using standard methods [39]. The one-to-one correspondence between $\mathcal{E}(X_0)$ and $\mathcal{Q}(X_0)$ allows these samples to be mapped to the distribution space, yielding uniform samples from $\mathcal{Q}(X_0)$.

Calibration Set Size. Beyond interpretability and sampling considerations, an additional practical question in CP concerns the size of the calibration set required to achieve a desired coverage level. In practice, the calibration size cannot be chosen arbitrarily large, as the computational cost of conformal prediction scales linearly with the number of calibration samples. The following result addresses this issue by characterizing the conditional coverage behavior induced by a finite calibration set.

Theorem 4.5. *Let $\ell := \lfloor (d_0+1)\gamma \rfloor$. Then the conditional coverage probability satisfies*

$$\mathbb{P}\left(\tilde{y}_0 \in \mathcal{Q}(X_0) \mid \mathcal{C}\right) \;\sim\; \mathrm{Beta}(d_0+1-\ell, \ell).$$

The proof of Theorem 4.5 is provided in Appendix B.3. In practice, one is often interested in empirical coverage behavior evaluated over a finite collection of test instances that share the same calibration set. The following result extends Theorem 4.5 to this scenario by characterizing the distribution of the number of covered test points.

Corollary 4.6. *Let $\{(X_0^{(j)}, Y_0^{(j)})\}_{j=1}^t$ be t exchangeable test pairs, independent of the calibration sample. Define $I := \sum_{j=1}^t \mathbf{1}\left\{\tilde{y}_0^{(j)} \in \mathcal{Q}(X_0^{(j)})\right\}$, and $\ell := \lfloor (d_0+1)\gamma \rfloor$. Then, we have*

$$I \,\Big|\, \mathcal{C} \;\sim\; \mathrm{BetaBinomial}\left(t,\, d_0+1-\ell, \ell\right).$$

The proof of Corollary 4.6 is provided in Appendix B.4.

5 Adaptive CP Framework with Conditional Coverage Guarantees

While the coverage guarantee in Equation (1) provides strong distribution-free uncertainty control on average, it does not ensure reliable coverage for a fixed test input. In particular, marginal validity permits the coverage probability to vary across test instances and may therefore offer limited guarantees at the individual level. In many applications, however, it is desirable to provide uncertainty guarantees that adapt to the specific test input and are thus more informative on an instance-by-instance basis [18, 2, 13, 17, 38]. Based on this motivation, following the framework of [60], we develop an Adaptive CP procedure that targets conditional coverage. Specifically, our goal is to guarantee

$$\mathbb{P}(\tilde{y}_0 \in \mathcal{Q}(X_0) \,|\, X_0) \xrightarrow{\mathbb{P}} 1-\gamma. \tag{4}$$

Here, $\mathbb{P}\left(\tilde{y}_0 \in \mathcal{Q}(X_0) \mid X_0\right)$ is with respect to the randomness of the test output Y_0 given X_0, whereas convergence in probability is with respect to the calibration sample. We first present the construction of the prediction set for Adaptive CP, and then establish its asymptotic conditional coverage guarantee in Equation (4). We further analyze the dependence of the coverage on the size of the calibration set. We summarize our method in Algorithm 2 in the Appendix.

5.1 Construction of Adaptive Prediction Set

Let $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ denote a user-specified distance on the input space $\mathcal{X}$. For a given test distribution X_0, we construct a local calibration subset by selecting the k nearest input distributions from the calibration set with respect to the distance D. Specifically, we define the index set

$$\mathcal{I}_{X_0} \coloneqq \arg\overset{(k)}{\min_{i\in[d_0]}} D(X_0, X_i^c),$$

and the associated local calibration set $\mathcal{C}_{X_0} \coloneqq \{(X_i^c, Y_i^c) : i \in \mathcal{I}_{X_0}\}$. As in the global construction, we take $W = \operatorname{diag}\big(w_\alpha^2(\nu, \sigma)\big)$, and $g = \|\cdot\|_2$. Restricting attention to the calibration pairs in $\mathcal{C}_{X_0}$, we compute localized conformal scores. For the ith element $(X, Y) \in \mathcal{C}_{X_0}$, we define our score as

$$s_i^{X_0}(X, Y) \;:=\; g\Big(W^{1/2}\Big(\theta(\hat{f}(X)) - \theta(\tilde{y})\Big)\Big).$$

We now define the local conformal threshold $\hat{q}_{X_0}$ as the $\lceil (k+1)(1-\gamma)\rceil$-th order statistic of the collection $\{s_i^{X_0}\}_{i=1}^{|\mathcal{C}_{X_0}|}$. This threshold is then used to construct a coefficient-level prediction region,

$$\mathcal{E}(X_0) \coloneqq \Big\{a \in \mathbb{R}^{|T_R|} : g\Big(W^{1/2}\big(\theta(\hat{f}(X_0)) - a\big)\Big) \leq \hat{q}_{X_0}\Big\}.$$

Finally, we obtain the prediction set by mapping $\mathcal{E}(X_0)$ back through the basis expansion, yielding

$$\mathcal{Q}(X_0) \coloneqq \Big\{y_a(x) = \textstyle\sum_{\alpha\in T_R} a_\alpha \beta_\alpha(x) : a \in \mathcal{E}(X_0)\Big\}.$$

5.2 Theoretical Properties

Coverage Guarantee. To establish the conditional coverage property stated in Equation (4), we impose the following regularity assumption on the conditional distribution of the nonconformity scores. A discussion of this assumption is provided at the end of the section.

Assumption 5.1. *Recall that the calibration pairs $(X_1^c, Y_1^c), \ldots, (X_{d_0}^c, Y_{d_0}^c)$ and the test pair (X_0, Y_0) are i.i.d. from a joint distribution $\mathcal{P}$. Assume that for almost every input distribution x, the conditional law $\mathcal{P}_{S|X=x}$ of the score*

$$s_0 = g\Big(W^{1/2}\big(\theta(\hat{f}(X_0)) - \theta(\tilde{y}_0)\big)\Big), \text{ given } \; X_0 = x,$$

is absolutely continuous with respect to the Lebesgue measure on its convex support $\operatorname{Supp}(\mathcal{P}_{S|X=x})$. Moreover, for any $R > 0$, suppose that the conditional density $p_S(\cdot \mid x)$ satisfies, for all

$$s \in \operatorname{Supp}(\mathcal{P}_{S|X=x}) \cap \operatorname{Ball}_{\|\cdot\|}(R),$$

the bounds

$$\lambda_R^x \leq p_S(s \mid x) \leq \Delta_R^x,$$

where λ_R^x and Δ_R^x are local lower and upper bounds, respectively.

Under **Assumptions** 5.1 and 4.1, we show that the prediction set $\mathcal{Q}$ achieves the conditional coverage guarantee in Equation (4). This guarantee is stated formally in the following theorem.

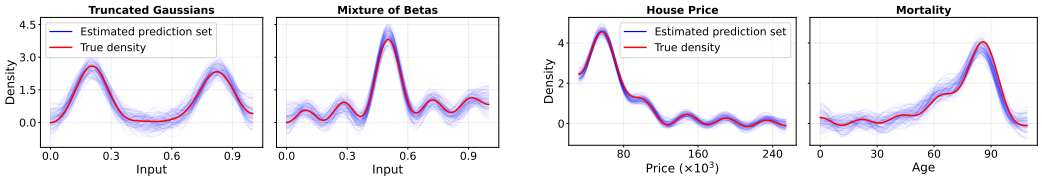
Theorem 5.2 (Asymptotic Conditional Coverage)**.** *Let k denote the number of nearest neighbors used to estimate the local quantile around each test input X_0. Under **Assumptions** 4.1 and 5.1, assume further that $k \to \infty$ and $k/d_0 \to 0$ as $d_0 \to \infty$, and that*

$$\widehat{F}_k(s \mid x) \xrightarrow{\mathbb{P}} F_x(s)$$

for $\mathbb{P}_X$-almost every x and every continuity point s of F_x. Then, for any $\gamma \in (0, 1)$,

$$\mathbb{P}(\tilde{y}_0 \in \mathcal{Q}(X_0) \mid X_0) \xrightarrow{\mathbb{P}} 1 - \gamma, \quad \text{as } d_0 \to \infty.$$

The proof of Theorem 5.2 is provided in Appendix B.5.



(a) Synthetic Datasets (b) Real Datasets

Figure 1: Visualization of distributions uniformly sampled from the proposed conformal prediction sets (blue), compared with the true density (red). Panel (a) presents results on synthetic datasets (Truncated Gaussians and Mixture of Betas), while panel (b) shows results on real-world datasets (House Price and Mortality).

Remark 5.3 (Discussion of the Condition $\widehat{F}_k \to F_x$ in Theorem 5.2)**.** The convergence condition $\widehat{F}_k(s \mid x) \xrightarrow{\mathbb{P}} F_x(s)$ is a standard consistency requirement in nonparametric estimation of conditional distributions, closely related to regularity conditions employed in previous studies of conditional conformal prediction [29, 19]. Intuitively, as $k \to \infty$, the increasing number of local calibration samples reduces estimation variance, while $k/d_0 \to 0$ ensures that the selected neighbors concentrate around the test input. Consequently, the empirical score distribution approaches the true conditional distribution, yielding a consistent estimate of the conditional quantile.

This condition is relatively mild and holds in several common settings. For instance, it is satisfied when the conditional score CDF $F_x(s)$ varies continuously with x under the distance D, and the k-nearest-neighbor radius converges to zero. This includes smooth finite-dimensional parametric families of input distributions, such as Gaussian families, with continuously varying conditional score distributions.

Volume Characterization and Uniform Sampling. As in the marginal setting, the proposed conditional prediction sets inherit favorable geometric properties. In particular, since $\mathcal{E}(X_0)$ is an ellipsoid in the coefficient space, its volume admits a closed-form expression, and uniform sampling can be performed efficiently using standard methods.

Calibration Set Size. We next study the effect of the local calibration set size on conditional coverage. We first introduce the following mild assumption.

Assumption 5.4. *Fix a test input X_0. Assume that the test score $S_0 := s(\hat{f}(X_0), Y_0)$ and the local calibration scores $\{S_i := s(\hat{f}(X_i^c), Y_i^c) : i \in \mathcal{I}_{X_0}\}$ are i.i.d. according to a distribution with continuous cumulative distribution function F_{X_0}.*

Under this assumption, we can explicitly characterize how the achieved conditional coverage depends on the size of the local calibration set through the following result.

Theorem 5.5. *Fix $\gamma \in (0, 1)$ and let $\ell = \lfloor (k+1)\gamma \rfloor$. Under **Assumptions** 4.1 and 5.4, the conditional coverage probability satisfies*

$$\mathbb{P}(\tilde{y}_0 \in \mathcal{Q}(X_0) \mid \mathcal{C}_{X_0}, X_0) \sim \mathrm{Beta}(k + 1 - \ell,\, \ell).$$

The proof of Theorem 5.5 is provided in Appendix B.6.

Discussion of Assumption 5.1. This assumption requires the nonconformity score distribution to have a well-behaved density, so that empirical quantiles from nearby calibration samples remain stable. It rules out pathological cases where small perturbations of the input distribution cause abrupt changes in the score distribution. The bounded-density condition captures local smoothness of the uncertainty structure around each test input, as is standard in conditional and local CP analyses.

Discussion of Assumption 5.4. This assumption requires the test score and the local calibration scores to be i.i.d. with a continuous distribution; it ensures exchangeability within the local neighborhood and rules out ties among scores. This condition is typically met when the score function is continuous, and the calibration points are drawn from a sufficiently small neighborhood around the test input, so the local score distributions are approximately identical.

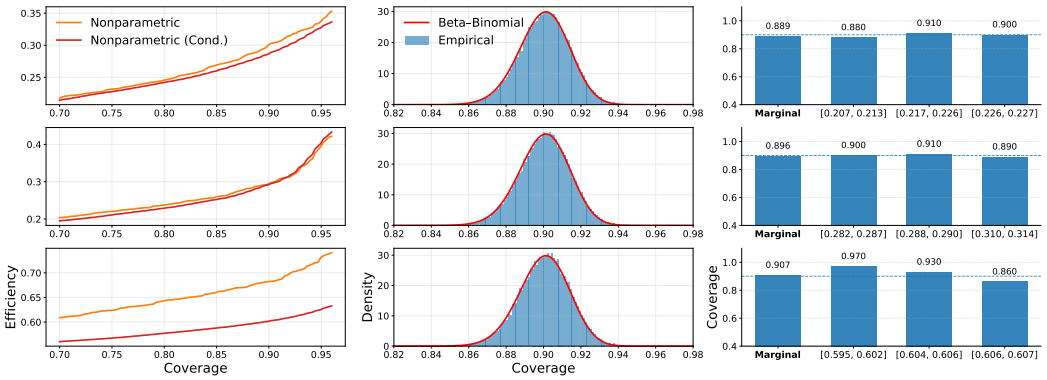


Figure 2: Results on synthetic datasets using the NDR baseline. Rows correspond to Truncated Gaussian, Mixture of Betas, and 3D Dark Matter datasets (top to bottom). Columns correspond to tasks **(I)–(III)** (left to right): efficiency-coverage trade-off between adaptive and non-adaptive CP, comparison of empirical coverage with the theoretical Beta-Binomial distribution, and empirical conditional coverage across different input groups, respectively. Dashed horizontal lines in the rightmost column indicate the nominal coverage level.

6 Experimental Results

This section provides numerical validation showing that (i) the proposed framework satisfies the coverage guarantees established in Theorem 4.2, while achieving a reasonable trade-off between coverage and prediction set volume; (ii) the number of calibration samples needed to achieve a prescribed coverage level for a given number of test instances follows Corollary 4.6; (iii) Adaptive CP conforms to the asymptotic conditional coverage behavior characterized in Theorem 5.2; and (iv) distributions obtained from the resulting prediction sets closely approximate the true distributions.

Overall Setup. We evaluate our proposed method on both synthetic and real-world datasets. For synthetic data, we consider one-dimensional mixtures of two truncated Gaussians [45] and mixtures of three Beta distributions [44], as well as multidimensional dark-matter datasets in 3D and 6D generated following [25]. For real-world evaluation, we use the house-price dataset studied in [8]. As base predictors $\hat{f}$, we consider three models: nonparametric distribution regression (NDR) [45], optimal transport map regression (OTR) [44], and Wasserstein regression (WR) [8].

Ablation Study. To further evaluate the stability and practical behavior of our method, we conduct ablation studies on the truncation parameter R, inference-time efficiency, and hyperparameter sensitivity, and evaluate performance in the low-data regime using the real-world Mortality dataset from [16]. The corresponding results are reported in Appendices C.3, C.4, C.5, and C.6, respectively.

6.1 Simulation Study

We evaluate three tasks: **(I)** examining the trade-off between the target coverage level $1-\gamma$ and prediction set volume; since the volume depends on the calibration threshold $\hat{q}$, we instead report results in terms of average $\hat{q}$ *(Efficiency)* versus $1-\gamma$ to avoid numerical instability; **(II)** validating the relationship between calibration-set size and empirical coverage as characterized in Corollary 4.6; and **(III)** for adaptive CP, empirically verifying the conditional-coverage guarantee in Theorem 5.2 using the evaluation protocol in Appendix A.2. Full experimental details are provided in Appendix C.1.

Truncated Gaussians and Mixture of Betas. Results for NDR tasks **(I)–(III)** under both truncated Gaussian and mixture-of-Betas settings are shown in Figure 2; OTR and WR results are shown in Figure 4 and Figure 5, respectively. For **(I)**, across all models and both settings, the adaptive variant consistently produces prediction sets with comparable or smaller volumes than its non-adaptive counterpart, demonstrating the effectiveness of adaptivity in reducing prediction set size. Both variants exhibit a clear efficiency-coverage trade-off that is consistent with trends reported in the existing literature [71, 72]. For **(II)**, the results closely match the Beta-Binomial relationships derived in Corollary 4.6, demonstrating accurate calibration across varying calibration sample sizes. For **(III)**, adaptive CP closely tracks the target conditional coverage, with deviations bounded by 0.04 (Truncated Gaussians) and 0.06 (Mixtures of Betas).

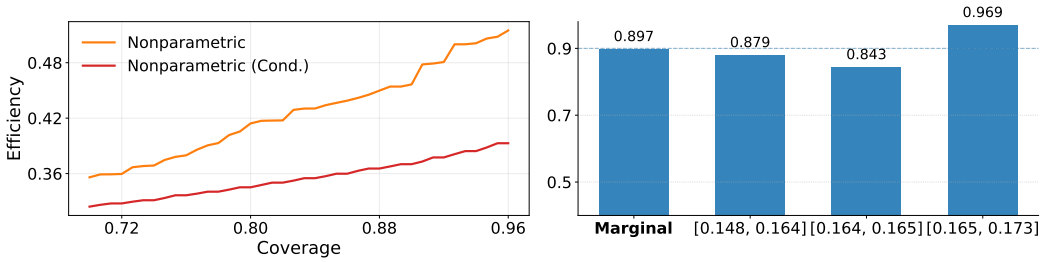


Figure 3: House Price experiments using the NDR base model. The left panel (task **(I)**) compares the efficiency-coverage trade-off between adaptive and non-adaptive CP, while the right panel (task **(III)**) reports empirical conditional coverage across different input groups, with the dashed line indicating the nominal coverage level.

Dark Matter Data. Due to methodological limitations, only NDR is applicable to the multidimensional setting. Results for tasks **(I)**–**(III)** are presented in Figure 2 for 3D data and Figure 6 for 6D data. Consistent with the one-dimensional experiments, task **(I)** demonstrates that adaptive CP achieves substantially better efficiency than its non-adaptive counterpart, while both methods exhibit the expected efficiency-coverage trade-off. For task **(II)**, the empirical results in both dimensions closely match the theoretical Beta-Binomial distribution established in Corollary 4.6. Finally, for task **(III)**, adaptive CP achieves near-nominal conditional coverage, with maximum deviations of 0.04 and 0.02 for the 3D and 6D datasets, respectively. These findings further support the effectiveness of our adaptive framework in higher-dimensional distributional settings.

Visualization. Figure 1(a) demonstrates the effectiveness of the proposed prediction sets via uniform sampling from each set and overlaying with the true density across datasets.

6.2 House Price Dataset

For the real-world evaluation, we consider the House Price dataset introduced in [8], with further details provided in Appendix C.2. We focus on tasks **(I)** and **(III)** defined in Section 6.1, omitting task **(II)** due to the limited sample size. Results for NDR are presented in Figure 3, while the corresponding results for OTR and WR are reported in Figure 7. Consistent with the simulation study, task **(I)** demonstrates that both adaptive and non-adaptive CP exhibit the expected efficiency-coverage trade-off reported in the literature [71], with the adaptive variant consistently achieving comparable or smaller prediction set volumes across all models. For task **(III)**, adaptive CP achieves near-nominal conditional coverage, with deviations bounded by 0.06. These findings further support the effectiveness and reliability of our adaptive framework in real-world distributional regression settings.

Visualization. Figure 1(b) visualizes the prediction sets by uniformly sampling distributions from them and overlaying these samples with the true density.

7 Conclusion

In this paper, we study conformal prediction (CP) for Distribution-to-Distribution Regression (DDR), where directly applying classical Split CP yields prediction sets that lack interpretability despite preserving marginal coverage guarantees. To address this limitation, we propose a CP framework based on projection onto a truncated orthonormal basis, enabling explicit volume characterization and efficient uniform sampling while maintaining coverage validity. We further develop an adaptive variant that incorporates local calibration and achieves asymptotic conditional coverage under mild assumptions. Extensive experiments on synthetic and real-world datasets, spanning both low- and high-dimensional settings and multiple DDR models, validate our theoretical results and demonstrate improved prediction efficiency through adaptivity. Despite these promising results, our framework does not directly accommodate covariate shift [61], and extending it to non-exchangeable settings, together with improving local calibration in high-dimensional distribution spaces, represents an interesting direction for future research.

Acknowledgments and Disclosure of Funding

This paper is dedicated to Professor Richard G. Baraniuk on the occasion of his 60th birthday.

This research is supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-TC-2023-012-SGIL). This research is also supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2023) (A-8002040-00-00, A-8002039-00-00), the NUS Presidential Young Professorship Award (A-0009807-01-00), the NUS Artificial Intelligence Institute–Seed Funding (A-8003062-00-00), and the Cross Faculty Grant 2025, CFG25-012 (A-8004460-00-00).

References

[1] Léo Andéol, Thomas Fel, Florence De Grancey, and Luca Mossina. Confident object detection via conformal prediction and conformal risk control: an application to railway signaling. In *Conformal and Probabilistic Prediction with Applications*, pages 36–55. PMLR, 2023.

[2] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[3] Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

[4] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

[5] Dimitrios Boursinos and Xenofon Koutsoukos. Selective classification of sequential data using inductive conformal prediction. In *2022 IEEE international conference on assured autonomy (ICAA)*, pages 46–55. IEEE, 2022.

[6] Danhui Chen, Ziquan Liu, Chuxi Yang, Dan Wang, Yan Yan, Yi Xu, and Xiangyang Ji. Conformalsam: Unlocking the potential of foundational segmentation models in semi-supervised semantic segmentation with conformal prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24045–24055, 2025.

[7] Xiaoyu Chen, Mengfan Fu, Yujing Huang, and Xinwei Deng. Distribution-in-distribution-out regression. *arXiv preprint arXiv:2405.11626*, 2024.

[8] Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *Journal of the American Statistical Association*, 118(542):869–882, 2023.

[9] Zhicheng Chen, Yuequan Bao, Hui Li, and Billie F Spencer Jr. Lqd-rkhs-based distribution-to-distribution regression methodology for restoring the probability distributions of missing shm data. *Mechanical Systems and Signal Processing*, 121:655–674, 2019.

[10] John Cherian, Isaac Gibbs, and Emmanuel Candes. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842, 2024.

[11] Matt Y Cheung, Ashok Veeraraghavan, and Guha Balakrishnan. Compass: Robust feature conformal prediction for medical segmentation metrics. *arXiv preprint arXiv:2509.22240*, 2025.

[12] Guneet S Dhillon, George Deligiannidis, and Tom Rainforth. On the expected size of conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*, pages 1549–1557. PMLR, 2024.

[13] Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36:64555–64576, 2023.

[14] Yu Feng, Man-Yat Chu, Uroš Seljak, and Patrick McDonald. Fastpm: a new scheme for fast simulations of dark matter and haloes. *Monthly Notices of the Royal Astronomical Society*, 463(3):2273–2286, 2016.

[15] Marius Gade, Kevin Mekhaphan Nguyen, Sol Gedde, and Alvaro Fernandez-Quilez. Impact of uncertainty quantification through conformal prediction on volume assessment from deep learning-based mri prostate segmentation. *Insights into Imaging*, 15(1):286, 2024.

[16] Laya Ghodrati and Victor M Panaretos. Distribution-on-distribution regression via optimal transport maps. *Biometrika*, 109(4):957–974, 2022.

[17] Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf008, 2025.

[18] Leying Guan. Conformal prediction with localization. *arXiv preprint arXiv:1908.08558*, 2019.

[19] Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction, 2022.

[20] Carlos Guestrin, Peter Bodik, Romain Thibaux, Mark Paskin, and Samuel Madden. Distributed regression: an efficient framework for modeling sensor network data. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 1–10, 2004.

[21] Dongjian Hu, Junxi Wu, Shu-Tao Xia, and Changliang Zou. Distribution-informed online conformal prediction. *arXiv preprint arXiv:2512.07770*, 2025.

[22] Peiyan Hu, Xiaowei Qian, Wenhao Deng, Rui Wang, Haodong Feng, Ruiqi Feng, Tao Zhang, Long Wei, Yue Wang, Zhi-Ming Ma, et al. From uncertain to safe: Conformal adaptation of diffusion models for safe pde control. In *Forty-second International Conference on Machine Learning*, 2025.

[23] Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot. One-shot federated conformal prediction. In *International Conference on Machine Learning*, pages 14153–14177. PMLR, 2023.

[24] Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression conformal prediction with random forests. *Machine learning*, 97(1):155–176, 2014.

[25] Andrea Klein. *Scalable Distribution-to-Distribution Regression*. PhD thesis, Department of Physics Jeff Schneider, Robotics Institute, 2014.

[26] Michal Klein, Louis Bethune, Eugene Ndiaye, and Marco Cuturi. Multivariate conformal prediction using optimal transport. *arXiv preprint arXiv:2502.03609*, 2025.

[27] Gilad Kutiel, Regev Cohen, Michael Elad, Daniel Freedman, and Ehud Rivlin. Conformal prediction masks: Visualizing uncertainty in medical imaging. In *International Workshop on Trustworthy Machine Learning for Healthcare*, pages 163–176. Springer, 2023.

[28] Jing Lei, Alessandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1):29–43, 2015.

[29] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

[30] Maud Lemercier, Cristopher Salvi, Theodoros Damoulas, Edwin Bonilla, and Terry Lyons. Distribution regression for sequential data. In *International Conference on Artificial Intelligence and Statistics*, pages 3754–3762. PMLR, 2021.

[31] Benjamin LeRoy and C Shafer. Conformal prediction for simulation models. In *Proceedings of the 2021 ICML Workshop on Distribution-Free Uncertainty Quantification, Online*, volume 24, 2021.

[32] Rui Li, Brian J Reich, and Howard D Bondell. Deep distribution regression. *Computational Statistics & Data Analysis*, 159:107203, 2021.

[33] Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6107–6117, 2025.

[34] Charles Lu, Anastasios N Angelopoulos, and Stuart Pomerantz. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In *International conference on medical image computing and computer-assisted intervention*, pages 545–554. Springer, 2022.

[35] Charles Lu and Jayasheree Kalpathy-Cramer. Distribution-free federated learning with conformal predictions. *arXiv preprint arXiv:2110.07661*, 2021.

[36] Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 12008–12016, 2022.

[37] Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, pages 22942–22964. PMLR, 2023.

[38] Gábor Lugosi and Marcos Matabuena. Conformal and knn predictive uncertainty quantification algorithms in metric spaces. *arXiv preprint arXiv:2507.15741*, 2025.

[39] Pierre L'Ecuyer. Non-uniform random variate generation. In *International Encyclopedia of Statistical Science*, pages 1758–1765. Springer, 2025.

[40] Thomas Mortier, Cas Decancq, and Diego G Miralles. Evaluating conformal prediction in weather forecasting. In *The 37th Benelux Conference on Artificial Intelligence and the 34th Belgian Dutch Conference on Machine Learning*, 2025.

[41] Luca Mossina and Corentin Friedrich. Conformal prediction for image segmentation using morphological prediction sets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 78–88. Springer, 2025.

[42] Sara Narteni, Alberto Carlevaro, Fabrizio Dabbene, Marco Muselli, and Maurizio Mongelli. A novel score function for conformal prediction in rule-based binary classification. *Pattern Recognition*, page 112219, 2025.

[43] Michael Nussbaum. On optimal filtering of a function of many variables in white gaussian noise. *Problemy Peredachi Informatsii*, 19(2):23–29, 1983.

[44] Ryo Okano and Masaaki Imaizumi. Distribution-on-distribution regression with wasserstein metric: Multivariate gaussian case. *arXiv preprint arXiv:2307.06137*, 2023.

[45] Junier Oliva, Barnabás Póczos, and Jeff Schneider. Distribution to distribution regression. In *International Conference on Machine Learning*, pages 1049–1057. PMLR, 2013.

[46] Andreas Paisios, Ladislav Lenc, Jiří Martínek, Pavel Král, and Harris Papadopoulos. A deep neural network conformal predictor for multi-label text classification. In *Conformal and Probabilistic Prediction and Applications*, pages 228–245. PMLR, 2019.

[47] Harris Papadopoulos. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher Rijeka, 2008.

[48] Harris Papadopoulos. A cross-conformal predictor for multi-label classification. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 241–250. Springer, 2014.

[49] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, pages 345–356. Springer, 2002.

[50] Harris Papadopoulos, Vladimir Vovk, and Alex Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

[51] Harris Papadopoulos, Volodya Vovk, and Alex Gammerman. Conformal prediction with neural networks. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, pages 388–395. IEEE, 2007.

[52] Yash Patel, Eduardo Ochoa Rivera, and Ambuj Tewari. Conformal prediction for ensembles: Improving efficiency via score-based aggregation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[53] Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning*, pages 27907–27947. PMLR, 2023.

[54] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[55] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[56] Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *International Conference on Machine Learning*, pages 5897–5906. PMLR, 2019.

[57] Jiankai Sun, Yiqi Jiang, Jianing Qiu, Parth Nobel, Mykel J Kochenderfer, and Mac Schwager. Conformal prediction for uncertainty-aware planning with diffusion dynamics model. *Advances in Neural Information Processing Systems*, 36:80324–80337, 2023.

[58] Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.

[59] Jacopo Teneggi, Matthew Tivnan, Web Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *International Conference on Machine Learning*, pages 33940–33960. PMLR, 2023.

[60] Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. *arXiv preprint arXiv:2501.18991*, 2025.

[61] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

[62] Chhavi Tyagi and Wenge Guo. Multi-label classification under uncertainty: A tree-based conformal prediction approach. In *Conformal and Probabilistic Prediction with Applications*, pages 488–512. PMLR, 2023.

[63] Janette Vazquez and Julio C Facelli. Conformal prediction in clinical medical sciences. *Journal of Healthcare Informatics Research*, 6(3):241–252, 2022.

[64] Guillermo Villate-Castillo, Javier Del Ser, and Borja Sanz. A collaborative content moderation framework for toxicity detection based on conformalized estimates of annotation disagreement. *arXiv preprint arXiv:2411.04090*, 2024.

[65] Harit Vishwakarma, Alan Mishler, Thomas Cook, Niccolo Dalmasso, Natraj Raman, and Sumitra Ganesh. Prune'n predict: Optimizing llm decision-making with conformal prediction. *arXiv preprint arXiv:2501.00555*, 2024.

[66] Harit Vishwakarma, Alan Mishler, Thomas Cook, Niccolo Dalmasso, Natraj Raman, and Sumitra Ganesh. Prune'n predict: Optimizing llm decision-making with conformal prediction. In *Forty-second International Conference on Machine Learning*, 2025.

[67] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[68] Chunjiang Wang, Yunjie Gu, Pengbo Liu, Shang Zhao, and S Kevin Zhou. E-dm: Evaluating diffusion model by conformal prediction. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[69] Håkan Wieslander, Philip J Harrison, Gabriel Skogberg, Sonya Jackson, Markus Fridén, Johan Karlsson, Ola Spjuth, and Carolina Wählby. Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images. *IEEE journal of biomedical and health informatics*, 25(2):371–380, 2020.

[70] Fatima Rabia Yapicioglu, Meltem Aksoy, Tuwe Löfström, Fabio Vitali, and Alberto Rigenti. Conformasegment: A conformal prediction-based, uncertainty-aware, and model-agnostic explainability framework for time-series forecasting. In *World Conference on Explainable Artificial Intelligence*, pages 218–242. Springer, 2025.

[71] Soroush H Zargarbashi, Simone Antonelli, and Aleksandar Bojchevski. Conformal prediction sets for graph neural networks. In *International Conference on Machine Learning*, pages 12292–12318. PMLR, 2023.

[72] Botong Zhang, Shuo Li, and Osbert Bastani. Conformal structured prediction. *arXiv preprint arXiv:2410.06296*, 2024.

[73] Jin Zhang, Ulf Norinder, and Fredrik Svensson. Deep learning-based conformal prediction of toxicity. *Journal of chemical information and modeling*, 61(6):2648–2657, 2021.

Table of Notation

$\mathbb{R}^k$	k-dimensional Euclidean space
$\|\cdot\|_2$	Euclidean norm
$\text{Ball}_{\|\cdot\|}(R)$	a ball of radius R, induced by $\|\cdot\|$
$\Gamma(\cdot)$	Gamma function
$\mathbf{1}\{\cdot\}$	Indicator function
$\text{Beta}(\cdot,\cdot)$	Beta distribution
$\text{BetaBinomial}(\cdot,\cdot,\cdot)$	Beta-Binomial distribution
$\theta(Y)$	collection of projection coefficients associated with distribution Y
$\theta_\alpha(Y)$	expansion coefficients of Y under basis β_α
$L_2(\Omega)$	space of square-integrable functions on Ω with respect to the Lebesgue measure.
$\mathcal{P}(X)$	powerset of a set X
$\text{Vol}(X)$	volume of a set X
$\|X\|$	cardinality of set X
$\mathcal{X}$	class of input distributions
$\mathcal{Y}$	class of output distributions
$\mathcal{P}$	probability measure defined on the class of input distribution
$\mathcal{C}$	calibration set
$\mathcal{E}$	coefficient-level prediction set
$\mathcal{Q}$	distribution-level prediction set
R	truncation parameter
γ	target miscoverage level
d_t	number of Training data points
d_0	number Calibration data points
m_i	number of samples for the ith input distribution
n_i	number of samples for the ith output distribution
$\hat{f}$	prediction model
$\{s_i\}_{i=1}^{d_0}$	nonconformity score
$\hat{q}$	the $\lceil (d_0+1)(1-\gamma) \rceil$-th order statistic of $\{s_i\}_{i=1}^{d_0}$
$\mathcal{I}_X$	set of k nearest neighbor to X in the calibration set.

Appendix of "Conformal Prediction for Distribution-to-Distribution Regression"

Table of Contents

A Background on Conformal Prediction **17**

A.1 Introduction . . . 17
A.2 Evaluation . . . 22
A.3 General Applications . . . 23
A.4 Applications in DDR . . . 24

B Theoretical Proofs **24**

B.1 Proof of the Marginal Coverage Guarantee - Theorem 4.2 . . . 24
B.2 Proof of the Prediction Set Volume - Theorem 4.3 . . . 25
B.3 Proof of the Marginal Coverage Law Given the Calibration Set - Theorem 4.5 . . . 26
B.4 Proof of Distribution of Coverage Given the Calibration Set - Corollary 4.6 . . . 27
B.5 Proof of Asymptotic Conditional Coverage - Theorem 5.2 . . . 27
B.6 Proof of the Conditional Coverage Law under Local Calibration - Theorem 5.5 . . . 28

C Experimental Details **29**

C.1 Simulation Study . . . 29
C.1.1 Truncated Gaussians . . . 29
C.1.2 Mixture of Betas . . . 30
C.1.3 Dark Matter Simulation Data . . . 31
C.2 House Price. . . . 32
C.3 Ablation Study for Truncation Parameter . . . 33
C.4 Ablation Study for Inference Time . . . 34
C.5 Ablation Study for Hyperparameter Sensitivity . . . 35
C.6 Low-Data Regime Analysis . . . 36
C.7 Computational Infrastructure . . . 37
C.8 Broader Societal Impacts . . . 37

A Background on Conformal Prediction

A.1 Introduction

Motivation. Deep learning and modern AI have dramatically improved performance across a wide range of tasks, from natural language processing and computer vision to reinforcement learning and beyond, often producing impressively accurate predictions on complex data. However, these methods are fundamentally probabilistic and data-driven, meaning their outputs are point estimates with associated uncertainties rather than guaranteed truths. In high-risk settings such as medical diagnosis, autonomous driving, or financial risk assessment, relying solely on a single prediction can be dangerous: an overconfident yet incorrect prediction can lead to misdiagnosis, unsafe decisions, or substantial economic loss. Traditional model confidence scores are often poorly calibrated and can be misleading, especially when the model encounters rare or out-of-distribution cases.

Algorithm 1 Conformal Prediction for Distribution-to-Distribution Regression

Require: Calibration set $\mathcal{C} = \{(X_i^c, Y_i^c)\}_{i=1}^{d_0}$ with $Y_i^c = \{N_{ij}^c\}_{j=1}^{n_i^c}$, test input X_0, truncation parameter R, metric g, weight matrix W, miscoverage level $\gamma \in (0,1)$, and parameters $(\nu, \sigma, \bar{A})$ satisfying Assumption 4.1.

Ensure: Prediction set $\mathcal{Q}(X_0)$.

Fix an orthonormal basis $\{\beta_i\}_{i\in\mathbb{Z}}$ of $L_2(\Lambda)$.

Define the product basis $\beta_\alpha(z) \leftarrow \prod_{r=1}^{l} \beta_{\alpha_r}(z_r)$ for $\alpha \in \mathbb{Z}^l$, $z \in \Lambda^l$.

Compute weights $w_\alpha^2(\nu,\sigma) \leftarrow \sum_{r=1}^{l} (\nu_r|\alpha_r|)^{2\sigma_r}$.

Calibration scores:

for $i = 1, \ldots, d_0$ **do**

  Compute empirical coefficients $\theta_\alpha(\widehat{Y}_i^c) \leftarrow \frac{1}{n_i^c} \sum_{j=1}^{n_i^c} \beta_\alpha(N_{ij}^c)$ for all $\alpha \in T_R$.

  Form the truncated density estimate $\tilde{y}_i^c(z) \leftarrow \sum_{\alpha\in T_R} \theta_\alpha(\widehat{Y}_i^c)\, \beta_\alpha(z)$.

  Compute score $s_i \leftarrow g\Big(W^{1/2}\big(\theta(\hat{f}(X_i^c)) - \theta(\tilde{y}_i^c)\big)\Big)$.

end for

Calibrate threshold. Let $\hat{q}$ be the $\lceil (d_0+1)(1-\gamma) \rceil$-th order statistic of $\{s_i\}_{i=1}^{d_0}$.

Prediction set in coefficient space: Define

$$\mathcal{E}(X_0) \leftarrow \{a \in \mathbb{R}^{|T_R|} : g(W^{1/2}(\theta(\hat{f}(X_0)) - a)) \le \hat{q}\}.$$

Map back to distribution space: Define

$$\mathcal{Q}(X_0) \leftarrow \left\{ y_a(z) = \sum_{\alpha\in T_R} a_\alpha \beta_\alpha(z) \,\middle|\, a \in \mathcal{E}(X_0) \right\}.$$

Return $\mathcal{Q}(X_0)$.

Conformal Prediction offers a principled solution to this challenge by providing statistically valid prediction sets or intervals that are guaranteed to contain the true outcome with a user-specified probability level, such as 90 % or 95%. At a high level, conformal methods work by measuring how "non-conforming" a new input looks relative to past data and then using this information to construct uncertainty sets that satisfy finite-sample coverage guarantees under minimal assumptions (essentially just exchangeability of the data). This makes the Conformal Prediction model-agnostic and distribution-free, meaning it can be applied on top of any predictive model, including neural networks, without assuming a specific data distribution, and still deliver reliable uncertainty quantification essential for trustworthy decision-making in safety-critical applications.

Full (Transductive) Conformal Prediction [55, 2]. Historically, full Conformal Prediction is the original form of Conformal Prediction, introduced before its more computationally efficient variants such as Split Conformal Prediction. In the full conformal framework, one considers a collection of $d+1$ exchangeable data points $(X_1, Y_1), \ldots, (X_d, Y_d), (X_{d+1}, Y_{d+1})$, where the learner observes $(X_i, Y_i)_{i=1}^{d}$ and the test input X_{d+1}, and aims to construct a prediction set for the unknown response Y_{d+1}. Unlike Split Conformal Prediction, where the training and calibration stages are separated, full Conformal Prediction allows the predictive model to be trained on all available data, including the candidate test label. The core idea is to loop over candidate outputs $y \in \mathcal{Y}$ and, for each y, fit a model $\hat{f}^{\,y}$ on the augmented sample

$$\{(X_1, Y_1), \ldots, (X_d, Y_d), (X_{d+1}, y)\}. \tag{5}$$

Crucially, the learning algorithm used to produce $\hat{f}^{\,y}$ must be permutation-invariant (symmetric) in its training data.

Given a nonconformity score function $s(\cdot,\cdot;\cdot)$, we compute the scores

$$s_i^y := s\big(X_i, Y_i; \hat{f}^{\,y}\big), \quad i = 1, \ldots, d, \qquad s_{d+1}^y := s\big(X_{d+1}, y; \hat{f}^{\,y}\big). \tag{6}$$

Algorithm 2 Adaptive Conformal Prediction for Distribution-to-Distribution Regression

Require: Calibration set $\mathcal{C} = \{(X_i^c, Y_i^c)\}_{i=1}^{d_0}$, test input X_0, truncation parameter R, distance $\mathcal{D}$, number of neighbors k, metric g, weight matrix W, miscoverage level γ, and parameters $(\nu, \sigma, \bar{A})$ satisfying Assumption 4.1.

Ensure: Adaptive prediction set $\mathcal{Q}(X_0)$.

Fix an orthonormal basis $\{\beta_i\}_{i\in\mathbb{Z}}$ of $L_2(\Lambda)$.

Define the product basis $\beta_\alpha(z) \leftarrow \prod_{r=1}^{l} \beta_{\alpha_r}(z_r)$ for $\alpha \in \mathbb{Z}^l$, $z \in \Lambda^l$.

Compute weights $w_\alpha^2(\nu, \sigma) \leftarrow \sum_{r=1}^{l} (\nu_r |\alpha_r|)^{2\sigma_r}$.

Select local calibration subset:

Define the index set $\mathcal{I}_{X_0} \leftarrow \arg\min_{i\in[d_0]}^{(k)} D(X_0, X_i^c)$,

Define local calibration set $\mathcal{C}_{X_0} \leftarrow \{(X_i^c, Y_i^c) : i \in I_{X_0}\}$.

Compute localized calibration scores:

for $i \in I_{X_0}$ **do**

  Compute empirical coefficients $\theta_\alpha(\widehat{Y}_i^c) \leftarrow \frac{1}{n_i^c} \sum_{j=1}^{n_i^c} \beta_\alpha(N_{ij}^c)$ for all $\alpha \in T_R$.

  Form the truncated density estimate $\tilde{y}_i^c(z) \leftarrow \sum_{\alpha\in T_R} \theta_\alpha(\widehat{Y}_i^c)\, \beta_\alpha(z)$.

  Compute score $s_i^{X_0} \leftarrow g\Big(W^{1/2}\big(\theta(\hat{f}(X_i^c)) - \theta(\tilde{y}_i^c)\big)\Big)$.

end for

Calibrate local threshold: Let $\hat{q}_{X_0}$ be the $\lceil (k+1)(1-\gamma)\rceil$-th order statistic of $\{s_i^{X_0}\}_{i\in I_{X_0}}$.

Prediction set in coefficient space: Define

$$\mathcal{E}(X_0) \leftarrow \{a \in \mathbb{R}^{|T_R|} : g(W^{1/2}(\theta(\hat{f}(X_0)) - a)) \le \hat{q}_{X_0}\}.$$

Map back to distribution space: Define

$$\mathcal{Q}(X_0) \leftarrow \left\{ y_a(z) = \sum_{\alpha\in T_R} a_\alpha \beta_\alpha(z) \,\middle|\, a \in \mathcal{E}(X_0) \right\}.$$

Return $\mathcal{Q}(X_0)$.

Let $\widehat{q}^{\,y}$ be the empirical conformal threshold defined as the $\lceil (d+1)(1-\alpha)\rceil$-th order statistic of $\{s_1^y, \ldots, s_{d+1}^y\}$ (equivalently, the empirical quantile at level $\lceil (d+1)(1-\alpha)\rceil/(d+1)$), i.e.,

$$\widehat{q}^{\,y} := \mathrm{Quantile}\left(s_1^y, \ldots, s_{d+1}^y; \frac{\lceil (d+1)(1-\alpha)\rceil}{d+1} \right). \tag{7}$$

We then include all candidates whose test score is no larger than this threshold:

$$\mathcal{C}(X_{d+1}) := \left\{ y \in \mathcal{Y} \;:\; s_{d+1}^y \le \widehat{q}^{\,y} \right\}. \tag{8}$$

Intuitively, $\mathcal{C}(X_{d+1})$ collects exactly those y for which the augmented point (X_{d+1}, y) looks sufficiently "typical" relative to the d observed samples under the model refit on the augmented dataset.

Theorem A.1 (Full conformal coverage guarantee)**.** *Assume $\{(X_i, Y_i)\}_{i=1}^{d+1}$ are exchangeable (in particular, i.i.d. is sufficient), and the learning algorithm producing $\hat{f}^{\,y}$ is symmetric (permutation-invariant) in its input sample. Then the full Conformal Prediction set $\mathcal{C}(X_{d+1})$ defined above satisfies the finite-sample marginal coverage guarantee*

$$\mathbb{P}(Y_{d+1} \in \mathcal{C}(X_{d+1})) \ge 1 - \alpha. \tag{9}$$

While full Conformal Prediction provides strong theoretical guarantees and serves as the conceptual foundation of conformal inference, its requirement to retrain the model for each candidate y makes it computationally expensive in practice. The split conformal procedure considered in this work can be viewed as a scalable special case of full Conformal Prediction, designed to retain validity while significantly reducing computational cost.

Split (Inductive) Conformal Prediction. In Section 2, we briefly introduced the framework of split Conformal Prediction. We now revisit this method in greater detail, clarifying its connection to full (transductive) Conformal Prediction.

Split Conformal Prediction can be viewed as a computationally efficient *special case* of full Conformal Prediction. Recall that the full Conformal Prediction conceptually evaluates all candidate labels $y \in \mathcal{Y}$ by retraining the model on the augmented dataset that includes (X_{d+1}, y). Split Conformal Prediction avoids this repeated refitting by fixing the predictive model in advance and using a held-out calibration set to assess conformity. This restriction dramatically reduces computational cost while preserving finite-sample validity under exchangeability.

Formally, let $\{(X_i, Y_i)\}_{i=1}^{d}$ be exchangeable samples from an unknown distribution on $\mathcal{X} \times \mathcal{Y}$. Given a target miscoverage level $\gamma \in (0, 1)$, the goal of Conformal Prediction is to construct a set-valued predictor $\mathcal{C} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ such that

$$\mathbb{P}(Y \in \mathcal{C}(X)) \geq 1 - \gamma, \tag{10}$$

where the probability is taken over a fresh test pair (X, Y) drawn from the same distribution.

In the split conformal framework, the dataset is randomly partitioned into two disjoint subsets: a training set and an independent calibration set. Let d_0 denote the size of the calibration set, and write

$$\mathcal{I}_{\text{cal}} = \{1, \ldots, d_0\} \tag{11}$$

for the indices of the calibration points. A prediction model $\hat{f} : \mathcal{X} \to \mathcal{Y}$ is trained *only* on the training set and remains fixed throughout the inference procedure.

Given a user-chosen nonconformity score function

$$s : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}, \tag{12}$$

which quantifies the discrepancy between a prediction and an observed response, we compute the calibration scores

$$s_i := s\big(\hat{f}(X_i), Y_i\big), \qquad i \in \mathcal{I}_{\text{cal}}. \tag{13}$$

The choice of s is problem-dependent and should reflect the relevant notion of uncertainty for the task at hand. An empirical threshold is then defined as the quantile

$$\hat{q} := \text{Quantile}\left(\{s_i\}_{i=1}^{d_0}; \frac{\lceil (d_0 + 1)(1 - \gamma) \rceil}{d_0}\right), \tag{14}$$

equivalently, the $\lceil (d_0 + 1)(1 - \gamma) \rceil$-th order statistic of the calibration scores. The Split Conformal Prediction set is finally given by

$$\mathcal{C}(x) := \big\{y \in \mathcal{Y} : s\big(\hat{f}(x), y\big) \leq \hat{q}\big\}. \tag{15}$$

Theorem A.2 (Split conformal coverage guarantee)**.** *Assume $\{(X_i, Y_i)\}_{i=1}^{d}$ are exchangeable. Then the Split Conformal Prediction set $\mathcal{C}$ defined in (15) satisfies the marginal coverage guarantee*

$$\mathbb{P}(Y \in \mathcal{C}(X)) \geq 1 - \gamma, \tag{16}$$

for any underlying data distribution and any choice of nonconformity score s.

Remark A.3. Split Conformal Prediction arises as a restriction of full Conformal Prediction in which (i) the predictive model is trained once and held fixed, and (ii) conformity is evaluated only on an independent calibration subset rather than on the full augmented sample. This simplification breaks the exact symmetry exploited in full Conformal Prediction but preserves exchangeability between the test point and the calibration scores, which is sufficient to guarantee marginal coverage. The resulting method trades some statistical efficiency for substantial computational savings, making Split Conformal Prediction the method of choice in large-scale and high-dimensional applications.

Remark A.4. Although full Conformal Prediction provides exact finite-sample coverage guarantees, its practical applicability is severely limited for complex or continuous output spaces. In particular, it requires iterating over all candidate outputs $y \in \mathcal{Y}$ and refitting the model on the augmented dataset (X_{d+1}, y) for each candidate. When $\mathcal{Y}$ is continuous, high-dimensional, or structured, as in distribution-to-distribution regression, this procedure becomes computationally intractable due to the uncountable output space and repeated model fitting. While exact computation is possible in certain low-dimensional parametric settings with special score structures, such assumptions do not generally hold here. Consequently, full Conformal Prediction is unsuitable for our setting, motivating the use of split (inductive) Conformal Prediction, which preserves distribution-free validity while remaining computationally feasible.

Extensions of Conformal Prediction. Recent work has extended Conformal Prediction well beyond its classical setting, enabling valid uncertainty quantification in increasingly complex and modern learning problems.

Group-balanced Conformal Prediction [2]. In some applications, it is desirable to guarantee reliable uncertainty quantification not only marginally over the population but also within specific subpopulations. This motivates group-balanced Conformal Prediction, where coverage is enforced separately for each group. Formally, suppose the first component of the input X_i indicates group membership, taking values in $\{1, \dots, G\}$. The goal is to construct a prediction set $\mathcal{C}$ such that, for every group $g \in \{1, \dots, G\}$,

$$\mathbb{P}(Y_{d+1} \in \mathcal{C}(X_{d+1}) \mid X_{d+1} \in g) \geq 1 - \alpha. \tag{17}$$

Standard Conformal Prediction does not, in general, guarantee simultaneous coverage within each group. To address this, group-balanced Conformal Prediction stratifies the calibration scores by group and computes group-specific conformal quantiles. At test time, the prediction set is formed using the quantile corresponding to the group of X_{d+1}, yielding valid group-wise coverage under exchangeability.

Class-conditional Conformal Prediction [2]. In classification problems, it is often desirable to guarantee coverage separately for each ground-truth class rather than only marginally over the entire population. This leads to class-conditional Conformal Prediction, which aims to ensure that prediction sets achieve the target coverage level within every class. Formally, letting $\mathcal{Y} = \{1, \dots, K\}$ denote the set of class labels, the goal is to construct a prediction set $\mathcal{C}$ such that

$$\mathbb{P}(Y_{d+1} \in \mathcal{C}(X_{d+1}) \mid Y_{d+1} = k) \geq 1 - \alpha, \qquad \forall k \in \mathcal{Y}. \tag{18}$$

Unlike group-balanced Conformal Prediction, class membership is not directly observed at test time. To address this, class-conditional Conformal Prediction stratifies the calibration scores by class and computes class-specific quantiles. At inference time, candidate labels are included in the prediction set if their corresponding nonconformity scores fall below the class-specific threshold, yielding valid class-wise coverage under exchangeability.

Conformal Risk Control [3]. Beyond bounding miscoverage, Conformal Prediction has been extended to control more general notions of error through conformal risk control. Instead of enforcing a guarantee on the event $\{Y_{d+1} \notin \mathcal{C}(X_{d+1})\}$, this framework aims to bound the expected value of a user-specified loss function ℓ, yielding guarantees of the form

$$\mathbb{E}[\ell(\mathcal{C}(X_{d+1}), Y_{d+1})] \leq \alpha. \tag{19}$$

This formulation subsumes standard miscoverage as a special case and enables control over richer error metrics, such as the false negative rate or other task-dependent risks. Conformal risk control operates by constructing a family of prediction sets indexed by a tuning parameter and selecting this parameter using calibration data so that the empirical risk satisfies a suitably adjusted threshold. Under exchangeability and mild monotonicity assumptions on the loss, the resulting prediction set enjoys finite-sample, distribution-free guarantees on the target risk, making this approach particularly useful in applications where miscoverage alone does not capture the relevant notion of predictive error.

Conformal Prediction under Covariate Shift [61]. Classical Conformal Prediction relies on the assumption that calibration and test data are drawn from the same distribution. In practice, this assumption is often violated due to covariate shift, where the marginal distribution of the inputs changes from P to P_{test}, while the conditional distribution $Y \mid X$ remains unchanged. To handle this setting, weighted Conformal Prediction modifies the calibration procedure by reweighting calibration points according to their likelihood under the test distribution. Specifically, calibration scores are weighted using importance weights

$$w(x) = \frac{dP_{\text{test}}(x)}{dP(x)}, \tag{20}$$

so that points more representative of the test distribution receive a stronger influence. The conformal threshold is then computed as a weighted quantile of the calibration scores. Under correct specification of the weights and exchangeability after reweighting, this approach yields valid prediction sets for test points drawn from P_{test}, extending Conformal Prediction to settings with covariate shift.

Conformal Prediction under Distribution Drift [21]. Beyond covariate shift, Conformal Prediction has also been extended to settings with distribution drift, where the data distribution evolves gradually over time in a manner that may be unknown or difficult to estimate. In this setting, calibration data and test points are no longer identically distributed, but recent observations are typically more informative about the test distribution. Weighted Conformal Prediction addresses this by assigning larger weights to more recent calibration scores and computing a weighted conformal quantile. The resulting prediction set is constructed as usual by thresholding the nonconformity score. Under mild assumptions, including bounded drift measured via total variation distance, this approach preserves approximate marginal coverage, with explicit bounds that depend on both the magnitude of the drift and the chosen weighting scheme. This framework provides a principled way to maintain valid uncertainty quantification in non-stationary environments such as time-series and streaming data.

A.2 Evaluation

Coverage. Coverage is the central criterion for evaluating the quality of a Conformal Prediction method, as it quantifies the reliability of the resulting prediction sets. Formally, for a target miscoverage level $\gamma \in (0, 1)$, Conformal Prediction aims to ensure that the prediction set $\mathcal{C}(X)$ satisfies

$$\mathbb{P}(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})) \geq 1 - \gamma, \tag{21}$$

independently of the underlying data distribution. In practice, empirical coverage on a test set $\{(X_i^{\text{test}}, Y_i^{\text{test}})\}_{i=1}^{n_{\text{test}}}$ is estimated as

$$\widehat{\text{Cov}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbf{1}\left\{Y_i^{\text{test}} \in \mathcal{C}(X_i^{\text{test}})\right\}, \tag{22}$$

which measures the frequency with which the true outcomes are contained in the predicted sets. Coverage is essential because it provides a rigorous, interpretable guarantee on uncertainty quantification: a method achieving coverage close to the nominal level $1 - \gamma$ can be trusted to capture the true outcome with the prescribed probability, whereas under-coverage indicates overconfident and potentially unsafe predictions.

Prediction Set Size [12]. While coverage ensures reliability, the size of the prediction set determines the practical usefulness and interpretability of Conformal Prediction. Among methods achieving the same coverage level, smaller prediction sets are preferred, as they provide more informative and less conservative uncertainty quantification. The notion of size depends on the output space: for classification, it is typically measured by the cardinality of the label set, while for regression, it may correspond to interval length or volume. Formally, the empirical average size of the prediction sets on a test set $\{X_i^{\text{test}}\}_{i=1}^{n_{\text{test}}}$ is given by

$$\widehat{\text{Size}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |\mathcal{C}(X_i^{\text{test}})|, \tag{23}$$

where $|\mathcal{C}(X)|$ denotes the cardinality or volume of the prediction set. Evaluating prediction set size alongside coverage highlights the fundamental coverage-size trade-off ina Conformal Prediction: overly large sets may trivially achieve coverage but lack informativeness, whereas excessively small sets risk violating the coverage guarantee.

Adaptability. Beyond marginal coverage and average prediction set size, a desirable Conformal Prediction procedure should be adaptive, producing small prediction sets for easy inputs and larger sets for harder or more uncertain inputs in a manner that faithfully reflects the underlying model uncertainty. Adaptability captures whether uncertainty is allocated heterogeneously across the input space rather than uniformly, avoiding overly conservative behavior on simple instances. A common way to evaluate adaptability is to partition the test data into bins based on a proxy for difficulty or uncertainty, such as ranges of the test input, ranges of the predicted output, model confidence scores, or the resulting prediction set size, and then compute empirical coverage separately within each bin. Formally, for a partition $\{\mathcal{B}_k\}_{k=1}^{K}$ of the test set, one evaluates

$$\widehat{\text{Cov}}_k = \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \mathbf{1}\left\{Y_i^{\text{test}} \in \mathcal{C}(X_i^{\text{test}})\right\}, \tag{24}$$

and assesses whether coverage remains close to the nominal level across all bins. Consistent coverage across bins indicates that the method adapts appropriately to varying levels of uncertainty, while systematic deviations reveal a lack of adaptivity or localized miscalibration.

Remark A.5 (Evaluation Procedure for task **(III)**)**.** In our setting, both inputs and outputs are probability distributions, which makes it difficult to define meaningful bins directly in the input or output space. As a result, we evaluate adaptability by partitioning the test samples according to the size of their conformal prediction sets, which serves as an intrinsic, model-agnostic proxy for predictive uncertainty. Specifically, we sort the test instances by prediction set size and divide them into bins such that each bin contains approximately the same number of test samples. For each bin, we compute the empirical coverage separately and examine whether it remains close to the nominal level. This evaluation allows us to assess whether larger prediction sets are indeed allocated to harder instances while maintaining valid coverage across different uncertainty regimes. We visualize this behavior by plotting the empirical coverage across a selected number of bins. Our approach follows that of [60].

A.3 General Applications

Conformal Prediction has been successfully applied across a wide range of learning problems by augmenting point predictions with statistically valid prediction sets. Below, we briefly describe several representative applications, highlighting how Conformal Prediction is used, the form of the prediction sets, and the type of uncertainty control involved.

Multilabel Classification [48]. In multilabel classification, Conformal Prediction is used to output a set of labels for each input rather than a single label vector. Prediction sets are constructed by conformalizing label-wise or joint nonconformity scores derived from probabilistic classifiers, ensuring that the true label set (or each true label) is contained with high probability. These methods often emphasize risk control, such as controlling the false discovery rate or expected set size, and can be adapted to label-conditional settings.

Disease (Sickness) Classification [63]. For disease classification tasks, Conformal Prediction produces a set of plausible diagnoses for a patient. The prediction set consists of disease labels whose scores satisfy a calibrated conformity threshold, guaranteeing that the true diagnosis is included with a prescribed probability. This setting primarily focuses on risk-sensitive decision making, allowing clinicians to identify ambiguous cases and defer decisions when uncertainty is high.

Polyp Segmentation [41]. In polyp segmentation, Conformal Prediction is applied to construct uncertainty-aware segmentation regions rather than a single binary mask. Prediction sets may take the form of pixel-wise confidence regions or spatial bands around predicted boundaries, providing coverage guarantees for the true polyp region. These sets are particularly useful for safety-critical applications, where missing polyp tissue carries a significant risk.

Weather Prediction [40]. In weather forecasting, Conformal Prediction is used to generate prediction intervals or sets for continuous variables such as temperature or precipitation. These sets are typically constructed by conformalizing regression residuals or ensemble forecasts and provide valid coverage without strong distributional assumptions. This application often addresses distribution shift across time or locations through split or adaptive calibration.

Quantile Regression [54]. In quantile regression, Conformal Prediction is used to transform estimated conditional quantiles into valid prediction intervals. By conformalizing lower and upper quantile estimates, the resulting intervals achieve guaranteed marginal coverage even under model misspecification. This approach is particularly effective in heteroskedastic settings and focuses on uncertainty calibration rather than pointwise accuracy.

Toxic Online Comment Identification [64]. For toxic content detection, Conformal Prediction constructs prediction sets over toxicity labels to identify uncertain or borderline cases. When the conformal set contains multiple labels, the system may defer judgment or flag content for human review. This application emphasizes selective risk control, reducing harmful misclassifications in automated moderation systems.

Selective Classification [5]. Selective classification naturally aligns with Conformal Prediction by allowing models to abstain when prediction sets are non-singleton or empty. Predictions are made only when the conformal set contains a single label, while abstention occurs otherwise. This framework provides explicit guarantees on the error rate of accepted predictions and enables principled risk-coverage trade-offs.

A.4 Applications in DDR

While conformal prediction (CP) has been applied broadly across many real-world settings, this work focuses specifically on CP for distribution-to-distribution regression (DDR). For completeness, this section briefly discusses several representative applications of CP within the DDR framework.

Model Reliability Quantification. A practitioner can use the size of the prediction set to decide whether to trust the model, defer to a human expert, or reject the input altogether. This type of decision rule is particularly important in high-stakes domains such as healthcare, genomics, and scientific discovery, where acting on an uncertain prediction can lead to significant consequences. For instance, in medical or gene expression analysis [45], a small prediction set over distributions indicates that the model has identified a well-concentrated and reliable range of plausible biological behaviors, allowing automated conclusions or downstream analysis. In contrast, a large or diffuse prediction set signals substantial uncertainty or heterogeneity in the data, prompting further experimental validation or expert review. In this sense, prediction sets for DDR provide a principled way to assess the reliability of distributional predictions and guide downstream decisions accordingly.

Robust Decision Making. Instead of relying on a single estimated distribution, the practitioner can make decisions that remain reliable across all plausible distributions in the prediction set. Concretely, given $\mathcal{C}(X)$, one can evaluate a decision, such as pricing, allocation, or investment choice, under each candidate distribution and select an option that performs well in the worst-case or conservative scenario. For example, in house price prediction [8], rather than committing to a single estimated price distribution, a buyer or investor can consider all plausible distributions in $\mathcal{C}(X)$ and choose a strategy that avoids overpaying even in the most adverse case. This is particularly important because, in the DDR setting where outputs are complex distributions, model uncertainty and misspecification are more likely; prediction sets therefore provide a principled way to guard against such risks and support more robust decision-making.

Hypothesis Testing. In many applications, the goal is not merely to produce a prediction, but to understand structural properties of the output distribution, such as concentration, multimodality, or tail behavior. However, when data are limited, validating such properties from a single-point estimate of the distribution is inherently unreliable, as small-sample noise or model misspecification can significantly distort these characteristics. In DDR, this challenge is even more pronounced because of the complexity of the distributional outputs. A prediction set $\mathcal{C}(X)$ addresses this by providing a collection of plausible distributions with coverage guarantees, allowing practitioners to assess whether a property holds consistently across the set rather than relying on a single estimate. For example, in age prediction [8], one may wish to determine whether the predicted age distribution is concentrated around a single age group or exhibits multiple plausible age ranges. If this property holds across all distributions in $\mathcal{C}(X)$ over many X, the conclusion is robust; otherwise, it indicates that the data is insufficient to support a definitive claim. In this sense, prediction sets for DDR act as a form of uncertainty-aware data augmentation, enabling more reliable validation of output properties in the complex and low-data setting of DDR.

In summary, conformal prediction in the standard setting has been widely adopted not merely as a predictive tool but as a framework that enables a range of downstream statistical tasks, including reliability-aware decision making, robust optimization, and hypothesis validation. In the DDR setting, these challenges become significantly more pronounced due to the increased complexity of distributional outputs and the higher risk of model misspecification. Our method extends conformal prediction to this setting, providing valid prediction sets for distributions and thereby ensuring that these downstream tasks and statistical objectives can still be carried out in a principled and reliable manner.

B Theoretical Proofs

B.1 Proof of the Marginal Coverage Guarantee - Theorem 4.2

Proof. We have that $\{(X_i, Y_i)\}_{i=1}^{d_0}$ and (X_0, Y_0) are exchangeable. Since $\tilde{y}_i$ (and hence $\theta(\tilde{y}_i)$) is a measurable function of the observed sample from Y_i, applying the same measurable map to each pair yields exchangeable scores $(s_0, s_1, \ldots, s_{d_0})$. Let R denote the rank of s_0 among $(s_0, s_1, \ldots, s_{d_0})$ after random tie-breaking. By exchangeability, R is uniform on $\{1, \ldots, d_0 + 1\}$.

With $k = \lceil (d_0+1)(1-\gamma) \rceil$ and $\hat{q} = s_{(k)}$ computed from $\{s_i\}_{i=1}^{d_0}$, the conformal event is

$$\theta(\tilde{y}_0) \in \mathcal{E}(X_0) \quad \text{if and only if} \quad s_0 \le \hat{q}. \tag{25}$$

Under random tie-breaking, $\{s_0 \le \hat{q}\}$ contains the event $\{R \le k\}$, hence

$$\mathbb{P}\big(s_0 \le \hat{q}\big) \;\ge\; \mathbb{P}(R \le k) \;=\; \frac{k}{d_0+1} \;\ge\; 1-\gamma. \tag{26}$$

If ties occur at the threshold, $\{s_0 \le \hat{q}\}$ may include up to a tie block of size at most $\bar{d}$ beyond rank k. Since R is uniform under random tie-breaking, this adds at most $\bar{d}/(d_0+1)$ probability mass, yielding

$$\mathbb{P}\big(s_0 \le \hat{q}\big) \;\le\; \frac{k}{d_0+1} + \frac{\bar{d}}{d_0+1} \;\le\; 1-\gamma+\frac{\bar{d}}{d_0+1}. \tag{27}$$

Therefore,

$$1-\gamma \;\le\; \mathbb{P}\big(\theta(\tilde{y}_0) \in \mathcal{E}(X_0)\big) \;\le\; 1-\gamma+\frac{\bar{d}}{d_0+1}. \tag{28}$$

By construction of $\mathcal{Q}(X_0)$ via the truncated basis map $a \mapsto y_a$, we have

$$\tilde{y}_0 \in \mathcal{Q}(X_0) \quad \text{if and only if} \quad \theta(\tilde{y}_0) \in \mathcal{E}(X_0), \tag{29}$$

and thus the same bounds apply to $\mathbb{P}(\tilde{y}_0 \in \mathcal{Q}(X_0))$. □

B.2 Proof of the Prediction Set Volume - Theorem 4.3

Proof. Let $t = |T_R|$ and enumerate $T_R = \{\alpha_1, \dots, \alpha_t\}$. Define the finite-dimensional coefficient-to-function map $\Phi : \mathbb{R}^t \to \operatorname{span}\{\beta_\alpha : \alpha \in T_R\}$ by

$$\Phi(a)(z) = \sum_{\alpha \in T_R} a_\alpha \beta_\alpha(z). \tag{30}$$

By construction,

$$\mathcal{Q}(X_0) = \Phi\big(\mathcal{E}(X_0)\big). \tag{31}$$

Since $\{\beta_\alpha\}_{\alpha \in T_R}$ is an orthonormal family in $L_2(\Lambda^l)$, the map Φ is a linear isometry from $\mathbb{R}^t$ equipped with the Euclidean norm to the finite-dimensional subspace $\operatorname{span}\{\beta_\alpha : \alpha \in T_R\} \subset L_2(\Lambda^l)$. Indeed, for any $a, b \in \mathbb{R}^t$,

$$\|\Phi(a) - \Phi(b)\|_{L_2(\Lambda^l)}^2 = \left\| \sum_{\alpha \in T_R} (a_\alpha - b_\alpha)\beta_\alpha \right\|_{L_2(\Lambda^l)}^2 \tag{32}$$

$$= \sum_{\alpha \in T_R} (a_\alpha - b_\alpha)^2 \tag{33}$$

$$= \|a-b\|_2^2. \tag{34}$$

Therefore, the t-dimensional volume of $\mathcal{Q}(X_0)$ in the induced coefficient subspace is equal to the Lebesgue volume of $\mathcal{E}(X_0)$ in $\mathbb{R}^t$:

$$\operatorname{Vol}\big(\mathcal{Q}(X_0)\big) = \operatorname{Vol}\big(\mathcal{E}(X_0)\big). \tag{35}$$

It remains to compute the volume of $\mathcal{E}(X_0)$. Let $m = \theta(\hat{f}(X_0))$. Since $g = \|\cdot\|_2$, the coefficient-space prediction set is

$$\mathcal{E}(X_0) = \left\{ a \in \mathbb{R}^t : \|W^{1/2}(m-a)\|_2 \le \hat{q} \right\}. \tag{36}$$

Thus $\mathcal{E}(X_0)$ is an ellipsoid centered at m. Under the change of variables

$$u = W^{1/2}(a-m), \qquad a = m + W^{-1/2}u, \qquad da = \frac{1}{\sqrt{\det W}}\, du. \tag{37}$$

The defining constraint becomes

$$\|u\|_2 \leq \hat{q}. \tag{38}$$

Hence,

$$\mathrm{Vol}\left(\mathcal{E}(X_0)\right) = \frac{1}{\sqrt{\det W}} \mathrm{Vol}\left(\{u \in \mathbb{R}^t : \|u\|_2 \leq \hat{q}\}\right) \tag{39}$$

$$= \frac{1}{\sqrt{\det W}} \cdot \frac{\pi^{t/2}\hat{q}^t}{\Gamma\left(\frac{t}{2}+1\right)}. \tag{40}$$

Combining this with $\mathrm{Vol}\left(\mathcal{Q}(X_0)\right) = \mathrm{Vol}\left(\mathcal{E}(X_0)\right)$, we obtain

$$\mathrm{Vol}\left(\mathcal{Q}(X_0)\right) = \frac{\pi^{t/2}\hat{q}^t}{\Gamma\left(\frac{t}{2}+1\right)\sqrt{\det W}} = \frac{(\sqrt{\pi}\hat{q})^t}{\Gamma\left(\frac{t}{2}+1\right)\sqrt{\det W}}. \tag{41}$$

This completes the proof. □

B.3 Proof of the Marginal Coverage Law Given the Calibration Set - Theorem 4.5

Proof. For each calibration pair (X_i^c, Y_i^c), $i = 1, \ldots, d_0$, let

$$S_i = s(\hat{f}(X_i^c), Y_i^c) \tag{42}$$

denote its conformal score, and let $S_0 = s(\hat{f}(X_0), Y_0)$ be the score of the test pair (X_0, Y_0). By the exchangeability of the data, the random variables $S_0, S_1, \ldots, S_{d_0}$ are i.i.d. from some distribution with a continuous cumulative distribution function F. Let

$$S_{(1)} \leq S_{(2)} \leq \cdots \leq S_{(d_0)} \tag{43}$$

denote the order statistics of $S_1, \ldots, S_{d_0}$, and define the threshold $T := S_{(d_0+1-\ell)}$. By the definition of the Conformal Prediction set $\mathcal{Q}(X_0)$, we have

$$\tilde{y}_0 \in \mathcal{Q}(X_0) \quad \text{if and only if} \quad S_0 \leq T. \tag{44}$$

Hence, conditional on the calibration sample $\{(X_i^c, Y_i^c)\}_{i=1}^{d_0}$,

$$\mathbb{P}\Big(\tilde{y}_0 \in \mathcal{Q}(X_0) \;\Big|\; \{(X_i^c, Y_i^c)\}_{i=1}^{d_0}\Big) = \mathbb{P}\Big(S_0 \leq T \mid \{S_i\}_{i=1}^{d_0}\Big) \tag{45}$$

$$= F(T). \tag{46}$$

Thus, the conditional coverage probability is the random variable $F(T)$. Next, apply the probability integral transform. Define $U_i := F(S_i), \qquad i = 1, \ldots, d_0$. Since F is continuous and $S_i \sim F$, it follows that $U_i \sim \mathrm{Unif}(0,1)$ i.i.d. Let

$$U_{(1)} \leq \cdots \leq U_{(d_0)} \tag{47}$$

denote the order statistics of $\{U_i\}_{i=1}^{d_0}$. Because F is monotone increasing and $T = S_{(d_0+1-\ell)}$, we have

$$F(T) = F\big(S_{(d_0+1-\ell)}\big) = U_{(d_0+1-\ell)}. \tag{48}$$

A standard result for order statistics of i.i.d. $\mathrm{Unif}(0,1)$ Random variables states that

$$U_{(k)} \sim \mathrm{Beta}(k,\, d_0+1-k), \qquad k = 1, \ldots, d_0. \tag{49}$$

Taking $k = d_0 + 1 - \ell$ yields

$$F(T) = U_{(d_0+1-\ell)} \sim \mathrm{Beta}(d_0+1-\ell,\, \ell). \tag{50}$$

Since $F(T)$ is precisely the conditional coverage probability, we conclude that

$$\mathbb{P}\Big(\tilde{y}_0 \in \mathcal{Q}(X_0) \;\Big|\; \{(X_i^c, Y_i^c)\}_{i=1}^{d_0}\Big) \sim \mathrm{Beta}(d_0+1-\ell,\, \ell), \tag{51}$$

which completes the proof. □

B.4 Proof of Distribution of Coverage Given the Calibration Set - Corollary 4.6

Proof. Let $\ell := \lfloor (d_0+1)\gamma \rfloor$, and let $\{(X_0^{(j)}, Y_0^{(j)})\}_{j=1}^t$ be an exchangeable collection of test pairs, independent of the calibration sample. For each calibration pair (X_i^c, Y_i^c), $i = 1, \ldots, d_0$, define the score $S_i := s(X_i^c, Y_i^c)$, and for each test pair define

$$S_0^{(j)} := s\big(X_0^{(j)}, Y_0^{(j)}\big), \qquad j = 1, \ldots, t. \tag{52}$$

Let $S_{(1)} \le \cdots \le S_{(d_0)}$ be the order statistics of $S_1, \ldots, S_{d_0}$ and set $T := S_{(d_0+1-\ell)}$. By definition of $\mathcal{Q}(\cdot)$, for each j,

$$\tilde{y}_0^{(j)} \in \mathcal{Q}(X_0^{(j)}) \quad \text{if and only if} \quad S_0^{(j)} \le T, \tag{53}$$

and therefore $I = \sum_{j=1}^t \mathbf{1}\Big\{S_0^{(j)} \le T\Big\}$. Conditioning on the calibration data $\{(X_i^c, Y_i^c)\}_{i=1}^{d_0}$, so that T is fixed. By the same exchangeability/continuity argument as in the proof of Theorem 4.5, the calibration scores and each test score are i.i.d. from some continuous CDF F, hence for every j,

$$\mathbb{P}\Big(S_0^{(j)} \le T \,\Big|\, \{(X_i^c, Y_i^c)\}_{i=1}^{d_0}\Big) = F(T). \tag{54}$$

Moreover, since the test pairs are exchangeable and independent of the calibration sample, the indicators $\mathbf{1}\{S_0^{(j)} \le T\}$ are conditionally exchangeable given $\{S_i\}_{i=1}^{d_0}$, and thus

$$I \,\big|\, \{(X_i^c, Y_i^c)\}_{i=1}^{d_0},\ F(T) \ \sim\ \mathrm{Binomial}\big(t,\ F(T)\big). \tag{55}$$

Finally, by Theorem 4.5 we have

$$F(T) = \mathbb{P}\Big(\tilde{y}_0 \in \mathcal{Q}(X_0) \,\Big|\, \{(X_i^c, Y_i^c)\}_{i=1}^{d_0}\Big) \sim \mathrm{Beta}(d_0 + 1 - \ell, \ell). \tag{56}$$

Integrating the conditional binomial law over the random success probability $F(T)$ yields the Beta-Binomial mixture,

$$I \,\big|\, \{(X_i^c, Y_i^c)\}_{i=1}^{d_0} \ \sim\ \mathrm{BetaBinomial}(t,\ d_0 + 1 - \ell,\ \ell)\,, \tag{57}$$

which completes the proof. □

B.5 Proof of Asymptotic Conditional Coverage - Theorem 5.2

Proof. Let a_0 be the coefficient vector associated with $\tilde{y}_0$, and define

$$S_0 := g\Big(W^{1/2}\big(\theta(\hat{f}(X_0)) - a_0\big)\Big)\,. \tag{58}$$

By the definition of $\mathcal{Q}(X_0)$, the event $\{\tilde{y}_0 \in \mathcal{Q}(X_0)\}$ is equivalent to $\{S_0 \le \hat{q}_{X_0}\}$. Hence, the desired conclusion can be rewritten as

$$\mathbb{P}\big(S_0 \le \hat{q}_{X_0} \mid X_0\big) \xrightarrow{\mathbb{P}} 1 - \gamma \qquad \text{as } d_0 \to \infty. \tag{59}$$

To establish (59), we use the regularity assumptions on the conditional distribution of the score S given X, stated above. For x in the support of X, denote by

$$F_x(s) := \mathbb{P}(S \le s \mid X = x) \tag{60}$$

the conditional cumulative distribution function of S given $X = x$, and write $q_{1-\gamma}(x)$ for the $(1-\gamma)$-quantile of this distribution,

$$q_{1-\gamma}(x) := \inf\{s \in \mathbb{R} : F_x(s) \ge 1 - \gamma\}. \tag{61}$$

By the absolute continuity and the local lower and upper bounds on the conditional density $p_S(\cdot \mid x)$ (The above assumption), F_x is strictly increasing and continuous in a neighborhood of $q_{1-\gamma}(x)$, and thus $F_x(q_{1-\gamma}(x)) = 1 - \gamma$.

Let $\hat{F}_k(\cdot \mid X_0)$ denote the k-nearest-neighbor empirical conditional CDF built from the scores $\{s_i^{X_0} : i \in \mathcal{I}_{X_0}\}$,

$$\hat{F}_k(s \mid X_0) := \frac{1}{k} \sum_{i \in \mathcal{I}_{X_0}} \mathbf{1}\{s_i^{X_0} \le s\}, \qquad s \in \mathbb{R}. \tag{62}$$

By construction, letting

$$r_k := \lceil (k+1)(1-\gamma) \rceil, \tag{63}$$

the local conformal threshold $\hat{q}_{X_0}$ is the r_k-th order statistic of the local scores. Since

$$\frac{r_k}{k} \longrightarrow 1-\gamma \qquad \text{as } k \to \infty, \tag{64}$$

the consistency condition in Theorem 5.2, together with the continuity and strict monotonicity of F_{X_0} in a neighborhood of $q_{1-\gamma}(X_0)$, yields

$$\hat{q}_{X_0} \xrightarrow{\mathbb{P}} q_{1-\gamma}(X_0). \tag{65}$$

In particular, by the continuity and strict monotonicity of F_{X_0} in a neighborhood of $q_{1-\gamma}(X_0)$, this pointwise convergence implies

$$\hat{q}_{X_0} \xrightarrow{\mathbb{P}} q_{1-\gamma}(X_0) \qquad \text{as } d_0 \to \infty. \tag{66}$$

Finally, observe that

$$\mathbb{P}(S_0 \le \hat{q}_{X_0} \mid X_0) = F_{X_0}(\hat{q}_{X_0}). \tag{67}$$

By (65) and the continuity of F_{X_0} at $q_{1-\gamma}(X_0)$, we obtain

$$F_{X_0}(\hat{q}_{X_0}) \xrightarrow{\mathbb{P}} F_{X_0}(q_{1-\gamma}(X_0)) = 1-\gamma. \tag{68}$$

This establishes (59), and therefore

$$\mathbb{P}\big(\tilde{y}_0 \in \mathcal{Q}(X_0) \mid X_0\big) = \mathbb{P}\big(S_0 \le \hat{q}_{X_0} \mid X_0\big) \xrightarrow{\mathbb{P}} 1-\gamma, \tag{69}$$

which completes the proof. □

B.6 Proof of the Conditional Coverage Law under Local Calibration - Theorem 5.5

Proof. Let $S_{(1)} \le \cdots \le S_{(k)}$ denote the order statistics of $\{S_i : i \in \mathcal{I}_{X_0}\}$. Fix X_0 and the calibration sample. Conditionally on $\{(X_i^c, Y_i^c)\}_{i=1}^{d_0}$ and X_0, the neighbor index set $\mathcal{I}_{X_0}$ is fixed, and hence the conformal threshold

$$T := S_{(k+1-\ell)} \tag{70}$$

is deterministic. By the definition of $\mathcal{Q}(X_0)$,

$$\tilde{y}_0 \in \mathcal{Q}(X_0) \quad \text{if and only if} \quad S_0 \le T. \tag{71}$$

Therefore, $\mathbb{P}\Big(\tilde{y}_0 \in \mathcal{Q}(X_0) \;\Big|\; \{(X_i^c, Y_i^c)\}_{i=1}^{d_0},\ X_0\Big)$ is equal to $\mathbb{P}\big(S_0 \le T \;\big|\; \{S_i\}_{i\in\mathcal{I}(X_0)},\ X_0\big)$.

Under Assumption 5.4, the conditional distribution $S_0 \mid (X_0, \mathcal{I}_{X_0})$ follows F_{X_0} and is independent of $\{S_i\}_{i\in\mathcal{I}_{X_0}}$. Hence, the right-hand side equals $F_{X_0}(T)$.

Applying the probability integral transform, define

$$U_i := F_{X_0}(S_i), \qquad i \in \mathcal{I}_{X_0}. \tag{72}$$

Since F_{X_0} is continuous, the random variables $\{U_i : i \in \mathcal{I}_{X_0}\}$ are i.i.d. $\mathrm{Unif}(0,1)$. Moreover, their order statistics satisfy

$$F_{X_0}(T) = F_{X_0}\big(S_{(k+1-\ell)}\big) = U_{(k+1-\ell)}. \tag{73}$$

A standard result on order statistics implies that $U_{(j)} \sim \mathrm{Beta}(j,\ k+1-j)$ for $j = 1, \ldots, k$. Taking $j = k+1-\ell$ yields

$$F_{X_0}(T) \sim \mathrm{Beta}(k+1-\ell,\ \ell), \tag{74}$$

which is exactly the claimed distribution for the conditional coverage probability. □

C Experimental Details

C.1 Simulation Study

Experimental setup. Unless otherwise stated, for the Truncated Gaussian and Mixture of Betas datasets, we generate 1000 training, 1000 calibration, and 1000 test distribution pairs. Each input and output distribution is observed through 1000 i.i.d. samples, which are then used to construct empirical distributions and their corresponding coefficient representations.

For a nonparametric distribution regression model that requires bandwidth selection, we further generate 200 distribution pairs that are used exclusively for cross-validation. The bandwidth is selected by minimizing the empirical ℓ_2 error between predicted and observed output coefficients on this validation set.

All output distributions are represented using a Fourier orthonormal basis. For one-dimensional synthetic datasets, the basis is truncated at $R = 10$. For the Dark Matter datasets, the Fourier bases are truncated at $R = 5$ for the 3D setting and $R = 3$ for the 6D setting. The Sobolev weighting parameters are fixed to $\nu = 1$ and $\sigma = 1$ throughout.

We evaluate on three tasks.
(I) Efficiency-coverage trade-off: the target coverage level $1 - \gamma$ is varied from 0.70 to 0.96, and *Efficiency* is computed by averaging over calibration threshold $\hat{q}$ across test inputs (for marginal CP, $\hat{q}$ is identical for all inputs). Both marginal and conditional Conformal Prediction are evaluated. For conditional Conformal Prediction, the neighborhood size is fixed to $K = 500$ nearest neighbors, and the prediction set volumes are averaged over all test points.

(II) Beta-Binomial verification: for the Truncated Gaussian and Mixture of Betas datasets, a fixed pool of 5000 distribution pairs is generated once. For the Dark Matter datasets, a fixed pool of 2300 distribution pairs is used. In each repetition, 1000 samples are randomly selected for calibration and 1000 samples for testing. Empirical coverage is computed on the test set, and this resampling procedure is repeated $50{,}000$ times to obtain the empirical distribution of coverage values.

(III) Conditional coverage: test inputs are sorted by prediction set size and partitioned into groups of equal cardinality from smallest to largest set size, and empirical coverage is computed within each group. Conditional Conformal Prediction again uses $K = 500$ nearest neighbors.

Visualization: 100 coefficient vectors are uniformly sampled from the ℓ_2-ball centered at the predicted output coefficients, with a radius determined by the conformal threshold, and mapped back to the distribution space via the truncated basis expansion.

The data generation procedures for each dataset are described below.

C.1.1 Truncated Gaussians

Data generation. We consider one-dimensional probability distributions supported on $[0, 1]$. For each dataset instance i, we generate an input-output distribution pair (μ_i, ν_i) as follows.

Input distribution. First, we draw the mixture parameters independently:

$$\mu_{i,1}, \mu_{i,2} \sim \mathrm{Unif}[0, 1], \quad \sigma_{i,1}, \sigma_{i,2} \sim \mathrm{Unif}[0.05, 0.1].$$

The input density p_i is defined as an equally weighted mixture of two Truncated Gaussian densities:

$$p_i(x) = \tfrac{1}{2} g(x; \mu_{i,1}, \sigma_{i,1}) + \tfrac{1}{2} g(x; \mu_{i,2}, \sigma_{i,2}), \quad x \in [0, 1].$$

Here, $g(\cdot; \mu, \sigma)$ denotes the Gaussian density with mean μ and standard deviation σ, truncated to $[0, 1]$, defined as

$$g(x; \mu, \sigma) = \tfrac{1}{\sigma} \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{1-\mu}{\sigma}\right) - \Phi\left(\frac{-\mu}{\sigma}\right)}, \quad x \in [0, 1],$$

where ϕ and Φ denote the standard normal probability density function and cumulative distribution function, respectively.

Output distribution. The corresponding output density q_i is defined by reflecting the mixture means around $1/2$, while keeping the same standard deviations:

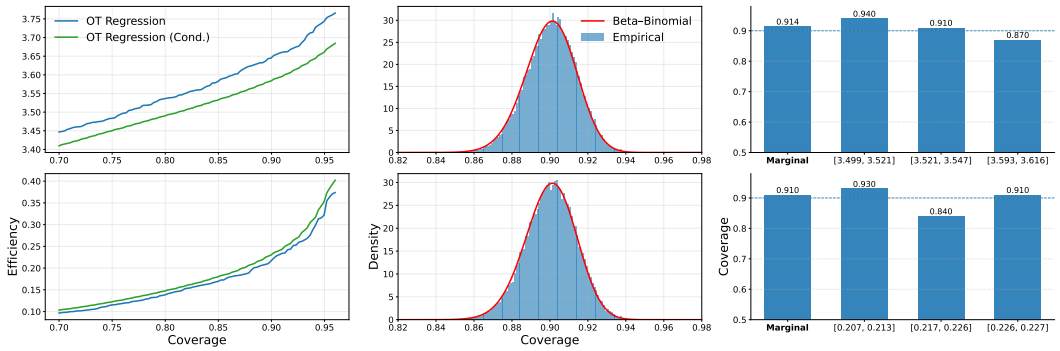


Figure 4: Experiments on Truncated Gaussian (first row) and Mixture of Betas (second row) for task **(I)** (first column), **(II)** (second column), and **(III)** (third column) using the OTR base model.

$$q_i(x) = \tfrac{1}{2} g(x; 1-\mu_{i,1}, \sigma_{i,1}) + \tfrac{1}{2} g(x; 1-\mu_{i,2}, \sigma_{i,2}), \quad x \in [0,1].$$

Sampling. Each input and output distribution is observed only through i.i.d. samples:

$$\{X_{ij}\}_{j=1}^{m_i} \overset{\text{i.i.d.}}{\sim} p_i, \qquad \{Y_{ij}\}_{j=1}^{n_i} \overset{\text{i.i.d.}}{\sim} q_i.$$

Sampling from the mixture is performed by first drawing a Bernoulli variable with probability $1/2$ to select a component, followed by sampling from the corresponding Truncated Gaussian distribution.

C.1.2 Mixture of Betas

Data generation. We consider one-dimensional probability distributions supported on $[0,1]$ constructed as mixtures of Beta distributions combined with random transport-based noise. For each dataset instance i, we generate an input-output distribution pair (μ_i, ν_i) as follows.

Input distribution. First, we generate the input distribution μ_i as a mixture of three Beta components:

$$\mu_i(dx) = \textstyle\sum_{j=1}^{3} \pi_j \, \mathrm{Beta}(x; \alpha_{i,j}, \beta_{i,j}) \, dx, \quad x \in [0,1],$$

where the mixture weights $\{\pi_j\}_{j=1}^{3}$ are fixed and satisfy $\sum_{j=1}^{3} \pi_j = 1$. In all experiments, we set $\pi = (0.3, 0.4, 0.3)$.

The shape parameters of the Beta components are drawn independently as

$$\alpha_{i,j} \sim \mathrm{Unif}[1,10], \qquad \beta_{i,j} \sim \mathrm{Unif}[1,10], \quad j = 1,2,3.$$

Given a fixed sample size m_i, the input distribution is observed through i.i.d. samples

$$\{X_{ij}\}_{j=1}^{m_i} \overset{\text{i.i.d.}}{\sim} \mu_i.$$

Regression operator. The regression operator is defined as a deterministic transport map followed by a random perturbation. Specifically, we define the base regression map $T_0 : [0,1] \to [0,1]$ as

$$T_0(x) = \zeta_4(x), \qquad \zeta_k(x) = x - \tfrac{\sin(\pi k x)}{|k|\pi},$$

which is a strictly increasing and smooth map satisfying $T_0(0) = 0$ and $T_0(1) = 1$.

Random error maps. To model stochastic variability, we introduce random transport maps $T_{\varepsilon,i}$. Let $J \geq 2$ and let

$$0 = U_{(0)} < U_{(1)} < \cdots < U_{(J-1)} < U_{(J)} = 1$$

be the order statistics of $J-1$ i.i.d. random variables drawn from $\mathrm{Unif}[0,1]$. For each interval $[U_{(j)}, U_{(j+1)}]$, we sample an integer-valued random variable K_j from a symmetric distribution around zero. In our experiments, K_j is supported on a finite set of integers.

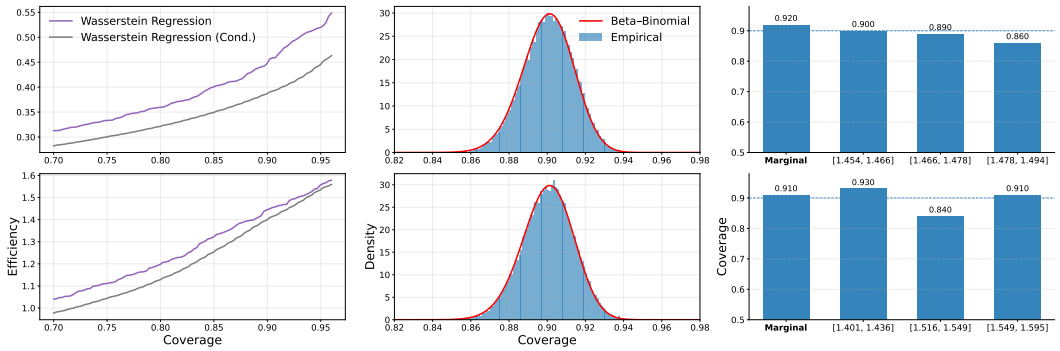


Figure 5: Experiments on Truncated Gaussian (first row) and Mixture of Betas (second row) for task **(I)** (first column), **(II)** (second column), and **(III)** (third column) using the WR base model.

The random error map is then defined as a piecewise transport map:

$$T_{\varepsilon,i}(x) = \sum_{j=0}^{J-1} \mathbf{1}_{[U_{(j)},\, U_{(j+1)})}(x)\, \xi\big(U_{(j)}, U_{(j+1)}, K_j\big)\,(x),$$

where $\xi(\cdot)$ denotes the affine rescaling of ζ_{K_j} to the interval $[U_{(j)}, U_{(j+1)}]$. By symmetry of K_j, the random maps satisfy

$$\mathbb{E}[T_{\varepsilon,i}(x)] = x, \quad \forall x \in [0,1].$$

Output distribution. The output distribution ν_i is defined as the push-forward of μ_i through the composed transport map:

$$\nu_i = T_{\varepsilon,i\#}\, T_{0\#}\mu_i.$$

In practice, given samples $\{X_{ij}\}_{j=1}^{m_i}$, the response samples are generated as

$$Y_{ij} = T_{\varepsilon,i}(T_0(X_{ij}))\,, \quad j = 1, \ldots, m_i,$$

followed by truncation to $[0,1]$ if necessary.

This construction yields nonlinear input-output relationships with distribution-dependent noise, providing a challenging benchmark for distribution-to-distribution regression and Conformal Prediction methods.

C.1.3 Dark Matter Simulation Data

Data generation. We construct synthetic distribution-to-distribution regression datasets based on cosmological Dark Matter simulations generated with the FastPM framework [14], following standard setups used in prior distribution regression studies [25].

Simulation setup. We run particle-mesh N-body simulations with grid resolution $\mathtt{nc} = 64$ under periodic boundary conditions. Two simulations with different numbers of time steps are run, both evolved up to redshift $z = 0.5$:

- a *coarse* simulation with 4 time steps, serving as an approximate evolution;
- a *fine* simulation with 12 time steps, serving as a high-fidelity reference.

Both simulations are initialized from the same initial conditions. The coarse simulation is treated as the input distribution, while the fine simulation is treated as the corresponding output distribution.

Spatial binning and distribution construction. The spatial domain of each simulation is first rescaled to the unit cube $[0,1]^3$. We then partition the space into a regular $16 \times 16 \times 16$ grid of cubic bins. For each bin, we collect all particles whose spatial coordinates fall inside the bin and treat the resulting point cloud as one distribution instance.

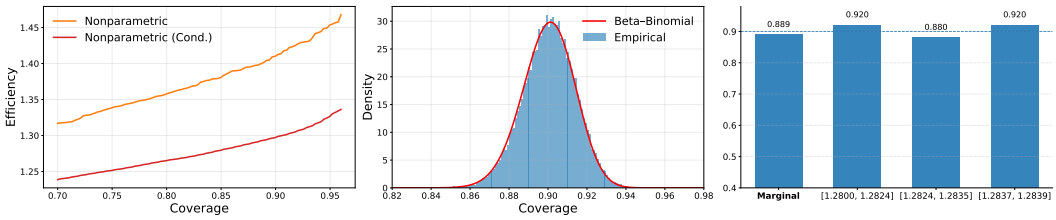


Figure 6: Experiments on Dark Matter 6D for task **(I)** (first column), **(II)** (second column), and **(III)** (third column) using the NDR base model.

Unless otherwise stated, we focus on the bin indexed by $(0, 0, 0)$, yielding paired input-output distributions corresponding to the same spatial region across simulation resolutions. **Feature normalization.** For each distribution instance, the spatial coordinates (x, y, z) of all particles within the bin are rescaled independently to lie in $[0, 1]$. Particle velocity components (v_x, v_y, v_z), when present, are left unnormalized to preserve their physical scale.

Dimensional settings. We consider two variants of the Dark Matter dataset:

- **3D Dark Matter:** distributions consist of spatial coordinates (x, y, z) only;
- **6D Dark Matter:** distributions consist of spatial coordinates and velocities (x, y, z, v_x, v_y, v_z).

In both cases, the input distribution is constructed from the coarse (4-step) simulation, and the output distribution is constructed from the fine (12-step) simulation. This setup induces a challenging distribution-to-distribution regression problem, where the goal is to map approximate particle distributions to their high-fidelity counterparts.

C.2 House Price.

Data. We consider the house price dataset studied in [8]. The data are based on inflation-adjusted median house prices collected from Zillow and consist of bimonthly observations for 306 cities in the United States, covering the period from June 1996 to 2025.

From the raw data, we construct the House Price dataset by extracting columns corresponding to monthly house price measurements over the period from 1996 to 2025 and retaining all rows corresponding to observed house prices across geographic locations. To ensure data integrity, any row containing a non-numeric entry is removed from the analysis. From the resulting cleaned table, we construct time-indexed input-output samples by pairing consecutive monthly columns, such that the empirical distribution of house prices at month t serves as the input distribution and that at month $t+1$ serves as the corresponding output distribution. Each distribution is defined by the collection of house price observations across all retained locations. Aggregating these consecutive month pairs yields a set of distribution-to-distribution regression samples that capture temporal dynamics in housing prices. This dataset is then divided into training, testing, and calibration based on the different tasks.

Experimental setup. For the House Price dataset, we adopt task-specific data splits due to the limited number of available distribution-valued observations.
For task **(I)**, we use 100 distributions for training and 200 distributions for conformal calibration. An additional 20 distributions are reserved for cross-validation of the nonparametric regression model, and all remaining distributions are used for testing. Task-specific settings, including evaluation metrics and target coverage levels, follow the same protocol as in task **(I)** of the simulation study. For conditional Conformal Prediction, we set the neighborhood size to $k = 100$.
For task **(III)**, we use 100 distributions for training and 100 distributions for calibration, with 20 distributions reserved for cross-validation of the nonparametric regression model. The remaining distributions are used for testing. In the adaptive conformal procedure, the neighborhood size is set to $k = 50$.
For visualization task, we use 200 distributions for training, 100 distributions for calibration and 20 distributions reserved for cross-validation.
Task **(II)** is omitted for this dataset due to the limited amount of available data, which makes reliable evaluation infeasible.

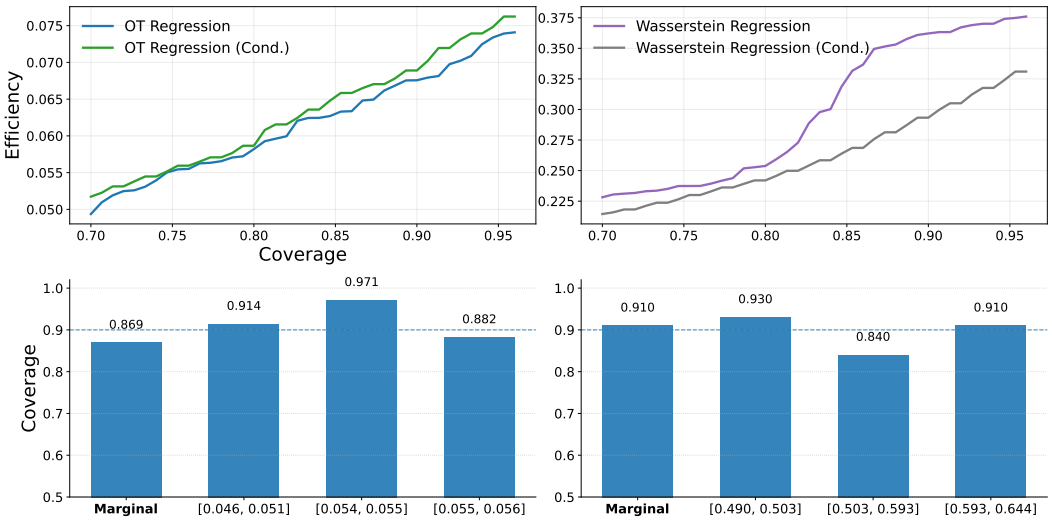


Figure 7: Experiments on House Price Data for task **(I)** (first row), and task **(III)** (second row) using the OTR base model (first column) and the WR base model (second column).

All experiments are conducted using three distribution-to-distribution regression models: Nonparametric Regression, Optimal Transport Regression, and Wasserstein Regression.

C.3 Ablation Study for Truncation Parameter

Motivated by the importance of prediction-set volume in Conformal Prediction (see Appendix A), we conduct an ablation study on the truncation parameter R, which directly controls the index set T_R used in our orthonormal expansion and hence the effective dimension of the conformal set. Recall, the index set is defined as

$$T_R \;:=\; \big\{\alpha \in \mathbb{Z}^l : 0 < w_\alpha(\nu, \sigma) \le R\big\}, \tag{75}$$

and we denote its cardinality by $t := |T_R|$.

By Theorem 4.3, when $g \equiv \|\cdot\|_2$, the volume of the resulting prediction set admits the closed-form expression

$$\mathrm{Vol}\big(\mathcal{Q}(X_0)\big) = \frac{\pi^{t/2}}{\Gamma\big(\frac{t}{2}+1\big)} \cdot \frac{\hat{q}^{\,t}}{\sqrt{\det W}}, \tag{76}$$

which shows that the truncation parameter R influences the prediction-set volume primarily through the induced dimension $t = |T_R|$ (and implicitly through the matrix W). In particular, Corollary 4.4 yields the asymptotic scaling

$$\mathrm{Vol}\big(\mathcal{Q}(X_0)\big) = O\left(\frac{(c\hat{q})^t}{\Gamma\big(\frac{t}{2}+1\big)}\right), \qquad c := \frac{\sqrt{\pi}}{m}, \tag{77}$$

making explicit that the prediction-set volume depends only on the conformal quantile $\hat{q}$ and the truncation-induced dimension t.

Consequently, increasing R, and hence enlarging T_R, can rapidly inflate the prediction-set volume, reflecting the curse-of-dimensionality behavior inherent to volume-based uncertainty sets, while choosing R too small may fail to capture fine-scale structure in the output distributions.

Experiment Setup. We conduct the ablation study on the truncation parameter R using the synthetic datasets considered in the simulation study, namely the Truncated Gaussian and Mixture of Betas settings. For each dataset, we evaluate three distribution-to-distribution regression models. Data splits for training, calibration, cross-validation, and testing follow the same protocol as in the simulation study, and all experimental settings are kept fixed across different values of R.
The truncation parameter is varied over the range $R \in \{2, 3, ..., 20\}$. For each choice of R, we

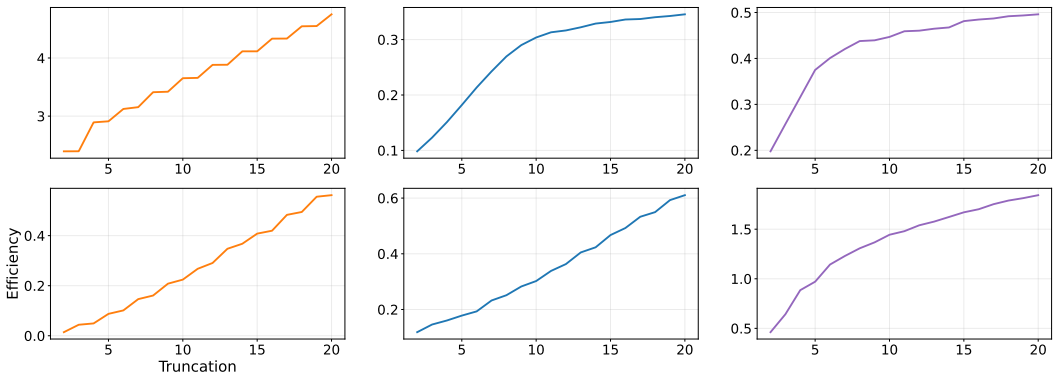


Figure 8: Ablation on the truncation parameter for Truncated Gaussian (first row) and Mixture of Betas (second row) using the NDR (first column), OTR (second column), and WR (third column) base models.

compute the conformal quantile $\hat{q}$ from the calibration scores and use it as the primary measure of efficiency. We report $\hat{q}$ rather than the prediction-set volume, as direct volume computation becomes numerically unstable in higher dimensions, while $\hat{q}$ provides a robust and monotone proxy for prediction-set size.

Results. The results for the Truncated Gaussian and Mixture of Betas settings are reported in Figure 8. Overall, the results demonstrate that computational efficiency is highly sensitive to the choice of the base regression model. Notably, even under identical data and truncation parameters in the Truncated Gaussian setting, the efficiency of Nonparametric Distribution Regression is around 3-4, whereas Optimal Transport Regression attains efficiencies on the order of 0.3-0.4, representing a more than tenfold disparity. This stark contrast empirically confirms our theoretical prediction that efficiency may grow exponentially with the truncation parameter. As a result, careful calibration of the truncation parameter to the specific model-data combination is essential in practice.

C.4 Ablation Study for Inference Time

While adaptive Conformal Prediction substantially reduces prediction-set size, it incurs additional computational overhead at inference time due to the localization step. In particular, for each test input distribution X_0, the adaptive procedure requires identifying a local calibration subset $\mathcal{C}_{X_0}$ by performing a k-nearest-neighbor (KNN) search in the input distribution space $(\mathcal{P}, D)$, followed by the computation of localized conformal scores and quantiles. This stands in contrast to non-adaptive Conformal Prediction, where a single global calibration set and threshold are computed once and reused for all test inputs.

More concretely, as described in Section 5, the additional inference-time cost of adaptive Conformal Prediction arises entirely from the localization step. For a given test input distribution X_0, adaptive Conformal Prediction requires computing the distances. $\{\mathcal{D}(X_0, X_i^c)\}_{i=1}^{d_0}$ between X_0 and all calibration input distributions $\{X_i^c\}_{i=1}^{d_0}$ in the metric space $(\mathcal{P}, \mathcal{D})$. The local index set

$$\mathcal{I}_{X_0} := \arg\min_{\substack{I \subset [n_c] \\ |I|=k}} \sum_{i \in I} \mathcal{D}(X_0, X_i^c) \tag{78}$$

is then obtained by selecting the k nearest neighbors of X_0 among the calibration inputs, which defines the local calibration subset $\mathcal{C}_{X_0}$. This nearest-neighbor search constitutes the dominant inference-time overhead of adaptive Conformal Prediction, with computational complexity scaling linearly with the calibration set size d_0 and the evaluation of the distance function $\mathcal{D}$.

Experiment Setup. We quantify the inference-time overhead of adaptive Conformal Prediction using the nonparametric regression model. This choice is motivated by the fact that inference-time complexity in conformal prediction depends only on the calibration and localization procedures, and not on the underlying regression model.
Experiments are conducted on the Truncated Gaussian and Mixture of Betas datasets. For each

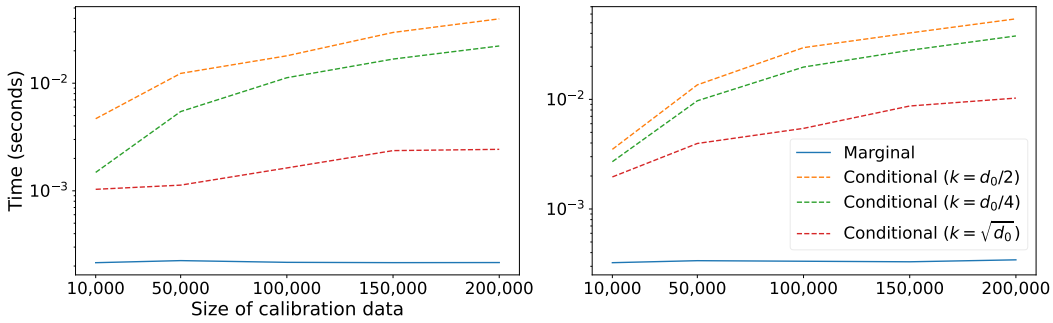


Figure 9: Ablation on Inference Time for Truncated Gaussian (first column) and Mixture of Betas (second column) using the NDR base models.

dataset, we use $1,000$ distributions for training, 200 distributions for cross-validation, and $1,000$ distributions for testing. The size of the calibration set d_0 is varied over

$$d_0 \in \{10,000; 50,000; 100,000; 150,000; 200,000\}$$

To study the effect of localization on inference time, we consider multiple choices of the neighborhood size k. Specifically, we set

$$k \in \left\{ \frac{d_0}{2}, \frac{d_0}{4}, \sqrt{d_0} \right\}$$

For each configuration, inference time is measured by applying conformal prediction to the entire test set and averaging the total runtime to obtain the mean inference time per test distribution.

Results. The results are reported in Figure 9. As expected, marginal Conformal Prediction exhibits nearly constant inference time, whereas adaptive Conformal Prediction incurs increasing computational overhead due to the localization step, which requires computing distances between the test input and all calibration distributions to identify the k nearest neighbors. Across both datasets, inference time grows monotonically with the calibration set size and increases with k, confirming that the KNN search dominates the overall computational cost. While absolute runtimes differ due to dataset-specific distance metrics and data representations, the qualitative scaling behavior is consistent. Notably, adopting a sublinear neighborhood size (e.g., $k = \sqrt{d_0}$) substantially mitigates inference-time overhead, illustrating a practical trade-off between adaptivity and computational efficiency.

C.5 Ablation Study for Hyperparameter Sensitivity

Our framework introduces several hyperparameters, including the orthonormal basis β, the metric g on the coefficient space, the weight matrix W, and the parameters $\nu, \sigma, \bar{A}$ from Assumption 4.1. These choices are introduced both to make the theoretical analysis precise and to follow standard practice in basis-based nonparametric estimation and local conformal calibration [45]. A natural concern, however, is that the proposed method may be sensitive to these choices, since they affect the coefficient representation, the geometry of the prediction set, and the local calibration procedure. In this section, we empirically show that this is not the case: across a range of reasonable hyperparameter choices, our method remains stable in terms of both coverage and prediction-set size.

Experiment setup. To investigate hyperparameter sensitivity, we repeat the same DDR task across the [datasets] while varying one hyperparameter at a time and keeping all other settings fixed, each time we run over 5 seeds to calculate the mean and variance. For each configuration, we measure the empirical coverage and compare it with the nominal target level. For the orthonormal basis β, we replace the default basis with alternative standard choices, including the Legendre basis and a wavelet basis. For the regularity parameters, we vary $\nu \in \{1, 1.5, 2\}$ and $\sigma \in \{0.5, 1, 1.5, 2\}$. For the score function, we consider $g = \|\cdot\|_1$ and $g = \|\cdot\|_3$. For the truncation level, we test $R \in \{3, 5, 8\}$, and for the neighborhood size, we test $k \in \{250, 500, 750\}$. We do not vary W or $\bar{A}$ in this experiment because the weight matrix W is fully determined by the choices of ν and σ, while $\bar{A}$ is a fixed

Table 1: Ablation on Hyperparameter Sensitivity results.

Varied Parameter	Truncated Gaussians	Mixture of Betas
Legendre Basis	0.90 ± 0.013	0.90 ± 0.014
Wavelet Basis	0.90 ± 0.013	0.90 ± 0.013
$\nu = 1$	0.90 ± 0.014	0.90 ± 0.014
$\nu = 1.5$	0.90 ± 0.013	0.90 ± 0.013
$\nu = 2$	0.90 ± 0.013	0.90 ± 0.013
$\sigma = 0.5$	0.90 ± 0.013	0.90 ± 0.013
$\sigma = 1$	0.90 ± 0.014	0.90 ± 0.014
$\sigma = 1.5$	0.90 ± 0.013	0.90 ± 0.013
$\sigma = 2$	0.90 ± 0.014	0.90 ± 0.013
$g = \|\cdot\|_3$	0.90 ± 0.014	0.90 ± 0.013
$g = \|\cdot\|_1$	0.90 ± 0.013	0.90 ± 0.013
$R = 3$	0.90 ± 0.013	0.90 ± 0.014
$R = 5$	0.90 ± 0.013	0.90 ± 0.013
$R = 8$	0.90 ± 0.014	0.90 ± 0.013
$k = 250$	0.87 ± 0.007	0.89 ± 0.016
$k = 500$	0.88 ± 0.009	0.89 ± 0.018
$k = 750$	0.89 ± 0.010	0.90 ± 0.019

regularity constant used in the theoretical assumption and does not enter the implementation of the conformal prediction procedure.

Results. The results are reported in Table 1. Overall, across both the Truncated Gaussians and Mixture of Betas datasets, the empirical coverage remains consistently close to the nominal level, with only minor variation across hyperparameter choices. This indicates that the proposed method is not overly sensitive to hyperparameter selection and remains robust under reasonable perturbations of its main design parameters.

C.6 Low-Data Regime Analysis

Our method is built upon Split Conformal Prediction, which constructs prediction sets using a dedicated calibration split. While this design enables finite-sample coverage guarantees, it also increases data requirements compared to standard training-testing pipelines. This limitation becomes particularly pronounced in low-data settings, where a small number of samples must simultaneously support model training, calibration, and evaluation.

To investigate the behavior of our approach under such constraints, we evaluate performance in a low-data regime using the Mortality dataset of [16], which contains only 36 samples. This dataset represents an extreme yet practically relevant scenario in which both the base regression model and the conformal calibration procedure operate under severe data scarcity. In this low-data setting, we focus our evaluation on task **(I)**, and task **(III)** (recall the definitions in Section 6 and Appendix C.1). These tasks directly probe the reliability and adaptivity of the conformal prediction sets under data scarcity. We do not consider task **(II)**, since accurate estimation of the underlying Beta distribution requires a sufficiently large sample size, which is not available in this regime.

Experimental setup. Experiments use the same three distribution-to-distribution regression models as in previous sections.
For task **(I)**, the data are split so that 10 distributions are used to fit the base regression model, while 20 distributions are reserved for conformal calibration. Cross-validation for the nonparametric regression model is performed using 2 additional distributions, and all remaining distributions are reserved for evaluation. The conditional conformal procedure is applied with a neighborhood size of $k = 10$. Since the available data are extremely limited, we repeat this random partitioning 20 times and report averages over the resulting runs.
For task **(III)**, the same data split is used, with distributions assigned to training, calibration, tuning, and testing as described above. The adaptive conformal procedure is implemented with a neighbor-

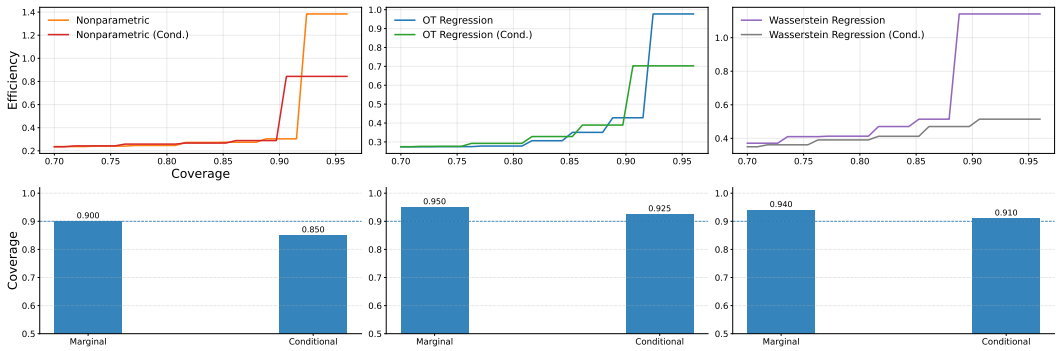


Figure 10: Ablation in the low-data regime on the Mortality dataset for task **(I)** (first row) and task **(III)** (third row) using the NDR (first column), OTR (second column), and WR (third column) base models.

hood size $k = 10$. All results are reported as averages over 5 independent runs.
For the visualization task, the data are split into 20 distributions for training, 2 distributions for cross-validation, and 10 distributions for calibration.

Results. The results are shown in Figure 10. For task **(I)**, the limited sample size leads to a noisier and less smooth efficiency-coverage curve; nevertheless, the overall trend remains consistent with that observed in larger-scale experiments, and the adaptive variant continues to produce prediction sets that are comparable in size or smaller than those of the non-adaptive method. For task **(III)**, adaptive conformal prediction achieves coverage levels close to both the marginal baseline and the nominal target, demonstrating that the proposed method remains effective even in the low-data regime.

Visualization. To further illustrate the effectiveness of our method in the low-data regime, we uniformly sample distributions from the prediction set and overlay them with the true density of the dataset. The resulting visualizations are shown in Figure 1.

C.7 Computational Infrastructure

All experiments are conducted on a CPU-only environment. In the simulation study, generating the Truncated Gaussian and Mixture of Betas datasets takes about 1 minute, while generating the Dark Matter data requires running a numerical simulation and takes on the order of one day.
For the experimental tasks, the end-to-end runtimes (including model training, calibration, and evaluation) are as follows. Task **(I)** takes roughly 2 hours, task **(II)** takes around 3 hours, and task **(III)** takes about 1 hour. The visualization task requires around 10 minutes. The ablation study on the truncation parameter takes about 30 minutes, while the ablation study for inference time takes roughly 5 hours.

C.8 Broader Societal Impacts

This work is primarily methodological and theoretical. Its potential positive societal impact lies in improving uncertainty quantification for distribution-valued prediction tasks, where both inputs and outputs may represent complex empirical distributions rather than finite-dimensional observations. Such settings arise in a range of scientific, medical, environmental, and policy-relevant applications, where reliable uncertainty estimates are important for interpreting predictions and supporting downstream decisions. By developing calibrated prediction sets for distribution-to-distribution regression, the proposed framework may help practitioners assess not only a single predicted distribution but also a statistically valid range of plausible output distributions. This can make uncertainty more transparent in settings where distributional outcomes are central, such as population-level health analysis, demographic forecasting, climate or environmental modeling, economic risk assessment, and scientific simulation. The ability to characterize prediction-set volume and sample from the resulting sets may further support visualization, sensitivity analysis, and decision-making under uncertainty.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction accurately reflect the paper's theoretical contributions and scope. They are supported by the main theorems and are stated under explicit assumptions.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of this work is discussed in Section 7.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We do provide the full set of assumptions and complete proofs for the main theorems.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and conclusions of the paper.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide open access to the data and code in Supplementary Material, with sufficient instructions to faithfully reproduce the main experimental results.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results in Appendix C.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports suitably and correctly defined error bars.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader Societal Impacts are discussed in Appendix C.8.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any data, pre-trained models, image generators, or scraped datasets. Therefore, safeguards for responsible release are not applicable.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper has credited all code and data that have been used.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new assets such as datasets, trained models, or software packages. Therefore, asset documentation is not applicable.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. Therefore, participant instructions, screenshots, and compensation details are not applicable.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve study participants or human-subject research. Therefore, participant risks, risk disclosure, and IRB approval are not applicable.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, a declaration is not required.

Answer: [N/A]

Justification: LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research.