Multivariate Conformal Prediction using Optimal Transport

Anonymous authors
Paper under double-blind review

Abstract

Conformal prediction (CP) quantifies the uncertainty of machine learning models by constructing sets of plausible outputs. These sets are constructed by leveraging a so-called conformity score, a quantity computed using the input point of interest, a prediction model, and past observations. CP sets are then obtained by evaluating the conformity score of all possible outputs, and selecting them according to the rank of their scores. Due to this ranking step, most CP approaches rely on a score functions that are univariate. The challenge in extending these scores to multivariate spaces lies in the fact that no canonical order for vectors exists. To address this, we leverage a natural extension of multivariate score ranking based on optimal transport (OT). Our method, **OT-CP** , offers a principled framework for constructing conformal prediction sets in multidimensional settings, preserving distribution-free coverage guarantees with finite data samples. We demonstrate tangible gains in a benchmark dataset of multivariate regression problems and address computational & statistical trade-offs that arise when estimating conformity scores through OT maps.

1 Introduction

Conformal prediction (CP) Gammerman et al. (1998); Vovk et al. (2005); Shafer & Vovk (2008) has emerged as a simple yet powerful framework for quantifying the prediction uncertainty of machine learning algorithms without imposing distributional assumptions on the data. Given a sequence of observed data

$$D_n = \{(x_1, y_1), \dots, (x_n, y_n)\}, \quad x_{n+1},$$

the objective is to construct a set that contains the unobserved response y_{n+1} with a prescribed coverage probability $1-\alpha$. This is typically achieved by computing a conformity score $S(x, y, \hat{y}) \in \mathbb{R}$, for example, a prediction error of $\hat{y}$, for each observation (x, y) in D_n, and then ranking these scores. The conformal prediction set for the new input x_{n+1} consists of all candidate responses y such that their score $S(x_{n+1}, y, \hat{y})$ ranks sufficiently low relative to the empirical distribution of scores $\{S(x_i, y_i, \hat{y})\}_{i=1}^n$ to meet the target confidence level.

In recent years, CP has witnessed rapid methodological and theoretical development Barber et al. (2023); Park et al. (2024); Tibshirani et al. (2019); Guha et al. (2024), reflecting its growing applicability to challenging learning scenarios Straitouri et al. (2023); Lu et al. (2022). Applications span a wide range of domains, including uncertainty quantification for active learning Ho & Wechsler (2008), anomaly detection Laxhammar & Falkman (2015); Bates et al. (2021), few-shot learning Fisch et al. (2021), time series forecasting Chernozhukov et al. (2018); Xu & Xie (2021); Chernozhukov et al. (2021b); Lin et al. (2022); Zaffran et al. (2022), performance guarantees for learning algorithms Holland (2020); Cella & Ryan (2020), and, more recently, uncertainty calibration for large language models Kumar et al. (2023); Quach et al. (2023). We refer the reader to the surveys Balasubramanian et al. (2014); Angelopoulos et al. (2024) for comprehensive reviews.

A key ingredient of CP is the notion of order: the inclusion of a candidate response depends on how its score ranks relative to past observations. Consequently, classical CP methods are primarily designed for univariate scores $S(x, y, \hat{y}) \in \mathbb{R}$. This poses a challenge for problems involving multivariate responses or vector-valued scores $S(x, y, \hat{y}) \in \mathbb{R}^d$ with $d \geq 2$, where ranking is not as straightforward as in the univariate case.
While conformal prediction for multivariate responses $y \in \mathbb{R}^d$ has been widely studied, much less attention has been paid to the setting where the conformity score itself is vector-valued $S(x, y, \hat{y}) \in \mathbb{R}^d$, which introduces distinct methodological and theoretical challenges.

Ordering Vector Distributions using Optimal Transport. In parallel to these developments, and starting with the seminal reference of Chernozhukov et al. (2017) and more generally the pioneering work of Hallin et al. (2021; 2022; 2023), multiple references have explored the possibilities offered by the optimal transport theory to define a meaningful ranking or ordering in a multidimensional space. Simply put, the analog of a rank function computed on the data can be found in the optimal Brenier map that transports the data measure to a uniform, symmetric, centered measure of reference in $\mathbb{R}^d$. As a result, a simple notion of a univariate rank for a vector $z \in \mathbb{R}^d$ can be found by evaluating the distance of the image of z (according to that optimal map) to the origin. This approach ensures that the ordering respects both the geometry, i.e., the spatial arrangement of the data and its distribution: points closer to the center get lower ranks.

Contributions We propose to leverage recent advances in computational optimal transport Peyré & Cuturi (2019), using notably differentiable transport map estimators Pooladian & Niles-Weed (2021); Cuturi et al. (2019), and apply such map estimators in the definition of multivariate score functions. More precisely:

- **OT-CP** : We extend conformal prediction techniques to multivariate score functions by leveraging optimal transport ordering, which offers a principled way to define and compute a higher-dimensional quantile and cumulative distribution function. As a result, we obtain distribution-free uncertainty sets that capture the joint behavior of multivariate predictions that enhance the flexibility and scope of conformal predictions.
- We propose a computational approach to this theoretical ansatz using the entropic map Pooladian & Niles-Weed (2021) computed from solutions to the Sinkhorn (1964) problem Cuturi (2013). We prove that our approach preserves the coverage guarantee while being tractable.
- We show the application of **OT-CP** using a recently released benchmark of regression tasks Dheur et al. (2025).

We acknowledge the concurrent proposal of Thurin et al. (2025), who adopt a similar approach to ours, with, however, a few important practical differences, discussed in more detail in Section 5

2 Background

2.1 Univariate Conformal Prediction

Conformal prediction is a flexible, model-agnostic framework that provides valid uncertainty quantification for machine learning models. It enables the construction of prediction sets (or intervals) that guarantee a desired coverage level such as 90% under the assumption that the data points are exchangeable which holds when for example the data are independently and identically distributed, or i.i.d.

To apply conformal prediction, begin by splitting your dataset into two similar but independent parts: a training set D_{tr} used to fit a predictive model $\hat{y}$, and a calibration set D_n, used to calibrate uncertainty estimates. Once the model is trained on the training data, compute nonconformity scores $s_i = s(x_i, y_i) \in \mathbb{R}$ on the calibration set D_n, where $s(x, y)$ measures how atypical a label y is for a new input x_{n+1}, given the model's prediction. From the calibration scores, determine the threshold $q_{1-\alpha}$ as the $(1-\alpha)(n+1)$-th empirical quantile of the scores $\{s_i\}_{i=1}^n$, where n is the number of calibration examples. At test time, given a new input x_{n+1}, the conformal prediction set $\mathcal{R}(x_{n+1})$ is defined as the set of labels y whose score does not exceed the quantile i.e.

$$\mathcal{R}(x_{n+1}) = \{y \in \mathcal{Y} : s(x_{n+1}, y) \leq q_{1-\alpha}\} \tag{1}$$

We now state the main coverage property of conformal prediction (Vovk et al., 2005).

Proposition 2.1. *Let $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ be exchangeable random variables taking values in $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{R}(x_{n+1})$ be the prediction set constructed as above using a scalar valued nonconformity scores computed on a calibration set of size n. Then:*

$$\mathbb{P}(Y_{n+1} \in \mathcal{R}(X_{n+1})) \geq 1 - \alpha, \quad \forall \alpha \in (0, 1).$$

That is, the prediction set contains the true label with probability at least $1-\alpha$, regardless of the underlying distribution, provided that the data are exchangeable. More details and proofs of this proposition can be found in the recent comprehensive review (Angelopoulos et al., 2024).

2.2 Multivariate Conformal Prediction

While many conformal methods exist for univariate prediction, we focus here on those applicable to *multivariate* outputs. As recalled in (Dheur et al., 2025), several alternative conformal prediction approaches have been proposed to tackle multivariate prediction problems. Some of these methods can directly operate using a simple predictor (e.g., a conditional mean) of the response y, while some may require stronger assumptions, such as requiring an estimator of the joint probability density function between x and y, or access to a generative model that mimics the conditional distribution of y given x) Izbicki et al. (2022); Wang et al. (2022). For simplicity, we restrict our attention to approaches that make no such assumption, reflecting our modeling choices for **OT-CP** .

M-CP . We will consider the template approach of Zhou et al. (2024) to use classical CP by aggregating a score function computed on each of the d outputs of the multivariate response. Given a conformity score s_i (to be defined next) for the i-th dimension, Zhou et al. (2024) define the following aggregation rule:

$$s_{\text{M-CP}}(x,y) = \max_{i\in[d]} s_i(x,y_i). \tag{2}$$

As Dheur et al. (2025), we will use *conformalized quantile regression* Romano et al. (2019) to define the score functions above, for each output $i \in [d]$, where the conformity score is given by:

$$s_i(x,y_i) = \max\left\{\hat{l}_i(x) - y_i, y_i - \hat{u}_i(x)\right\},$$

with $\hat{l}_i(x)$ and $\hat{u}_i(x)$ representing the lower and upper conditional quantiles of $Y_i|X=x$ at levels α_l and α_u, respectively. In our experiments, we consider equal-tailed prediction intervals, where $\alpha_l = \frac{\alpha}{2}$, $\alpha_u = 1-\frac{\alpha}{2}$, and α denotes the miscoverage level.

Merge-CP . An alternative approach is simply to use a squared Euclidean aggregation,

$$s(x,y) := \|\hat{y}(x) - y\|_2,$$

where the choice of the norm (e.g., ℓ_1, ℓ_2, or ℓ_∞) depends on the desired sensitivity to errors across tasks. This approach reduces the multidimensional residual to a scalar conformity score, leveraging the natural ordering of real numbers. This simplification not only makes it straightforward to apply univariate conformal prediction methods, but also avoids the complexities of directly managing vector-valued scores in conformal prediction. A variant consists of applying a Mahalanobis norm Johnstone & Cox (2021) in lieu of the squared Euclidean norm, using the covariance matrix Σ estimated from the training data Johnstone & Cox (2021); Katsios & Papadopulos (2024); Henderson et al. (2024),

$$s(x,y) := \|\Sigma^{-1/2}(\hat{y}(x) - y)\|_2,$$

2.3 Kantorovich Ranks

A naive way to define ranks in multiple dimensions might be to measure how far each point is from the origin and then rank them by that distance. This breaks down if the distribution of the data is stretched or skewed in certain directions. To correct for this, Hallin et al. (2021) developed a formal framework of center-outward distributions and quantiles, also called Kantorovich ranks Chernozhukov et al. (2017), extending the familiar univariate concepts of ranks and quantiles into higher dimensions by building on optimal transport theory.

Optimal Transport Map and Multivariate Quantile Region Let μ and ν be probability measures on $\Omega \subset \mathbb{R}^d$, representing the source and target distributions, respectively. One seeks a map $T : \Omega \to \Omega$ that pushes forward μ to ν while minimizing the average transportation cost:

$$T^\star \in \operatorname*{arg\,min}_{T_\#\mu=\nu} \int_\Omega \|x - T(x)\|^2 \, d\mu(x). \tag{3}$$

Brenier (1991) ensures that if the source measure μ admits a density, there exists a solution to equation 3 that is the gradient of a convex function $\phi : \Omega \to \mathbb{R}$, i.e., $T^\star = \nabla\phi$. In the one-dimensional case, the cumulative distribution function (CDF) of a distribution $\mathbb{P}$ is the unique increasing map that transports it to the uniform distribution. This monotonicity property extends to higher dimensions through the gradient of a convex function $\nabla\phi$. Thus, the optimal transport map in multiple dimensions can be seen as a natural analog of the univariate CDF: both provide a unique, monotone transformation from one distribution to another. Following Hallin et al. (2021); Chernozhukov et al. (2017), the center-outward distribution of a random variable $Z \sim \mathbb{P}$ is defined as the optimal transport map $T = \nabla\phi$ that pushes $\mathbb{P}$ forward to the uniform distribution $\mathbb{U}$ on the unit ball $B(0, 1)$: $T : \Omega \to B(0, 1)$ with $T_\#\mathbb{P} = \mathbb{U}$. This allows one to define the multivariate rank (and associated quantile) of a point Z as $\text{Rank}(Z) = \|T(Z)\|$, i.e., its radial distance from the origin. The inverse map, $(T^\star)^{-1} = \nabla\phi^*$, where ϕ^* is the convex conjugate of ϕ is referred to as the center-outward quantile map. The multivariate quantile region is a generalization of univariate quantiles to higher dimensions, representing a region in the sample space that contains a specified proportion of probability mass. The quantile region is then defined as: $\mathcal{R}_\alpha = \{z \in \mathbb{R}^d : \|T(z)\| \leq 1 - \alpha\}$. By construction of the spherical uniform distribution, $\|T(Z)\| \sim \text{Uniform}(0, 1)$, which implies the following coverage property: $\mathbb{P}(Z \in \mathcal{R}_\alpha) = 1 - \alpha$.

2.4 Entropic Map.

A convenient estimator of the Brenier map $T^\star$ from samples $(z_1, \ldots, z_n)$ and $(u_1, \ldots, u_m)$ is the entropic map Pooladian & Niles-Weed (2021): Let $\varepsilon > 0$ and write $K_{ij} = [\exp(-\|z_i - u_j\|^2/\varepsilon)]_{ij}$, the kernel matrix. Define,

$$\mathbf{f}^\star, \mathbf{g}^\star = \operatorname*{arg\,max}_{\mathbf{f}\in\mathbb{R}^n, \mathbf{g}\in\mathbb{R}^m} \langle \mathbf{f}, \tfrac{\mathbf{1}_n}{n}\rangle + \langle \mathbf{g}, \tfrac{\mathbf{1}_m}{m}\rangle - \varepsilon\langle e^{\frac{\mathbf{f}}{\varepsilon}}, K e^{\frac{\mathbf{g}}{\varepsilon}}\rangle \,. \tag{4}$$

The Equation (4) is an unconstrained concave optimization problem known as the regularized OT problem in dual form (Peyré & Cuturi, 2019, Prop. 4.4) and can be solved numerically with the Sinkhorn algorithm Cuturi (2013). Equipped with these optimal vectors, one can define the maps, valid out of sample:

$$f_\varepsilon(z) = \min\nolimits_\varepsilon([\|z - u_j\|^2 - \mathbf{g}^\star_j]_j)\,, \quad g_\varepsilon(u) = \min\nolimits_\varepsilon([\|z_i - u\|^2 - \mathbf{f}^\star_i]_i)\,, \tag{5}$$

where for a vector $\mathbf{u}$ or arbitrary size s we define the log-sum-exp operator as $\min_\varepsilon(\mathbf{u}) := -\varepsilon \log(\frac{1}{s}\mathbf{1}_s^T e^{-\mathbf{u}/\varepsilon})$. Using the Brenier (1991) theorem, linking potential values to optimal map estimation, one obtains an estimator for $T^\star$ and it inverse given weights $p_j(z) := \frac{\exp\left(-\left(\|z-u_j\|^2-\mathbf{g}^\star_j\right)/\varepsilon\right)}{\sum_{k=1}^m \exp\left(-\left(\|z-u_k\|^2-\mathbf{g}^\star_k\right)/\varepsilon\right)}$,

$$T_\varepsilon(z) := z - \nabla f_\varepsilon(z) = \sum_{j=1}^m p_j(z) u_j \quad T_\varepsilon^{\text{inv}}(u) := \sum_{i=1}^n q_j(u) z_j \tag{6}$$

where the weights and $q_j(u)$ arising for a vector u from the Gibbs distribution of $[\|z_i - u\|^2 - \mathbf{f}^\star_i]_i$.

3 Kantorovich Conformal Prediction

3.1 Optimal Transport Merging

We are interested in setting where the score function $S(x, y) \in \mathbb{R}^d$ is multidimensional $d \geq 2$. In this case, we can no longer benefit from the ordering of real values and the inequality $S(x_{n+1}, y) \leq q_{1-\alpha}$ in Equation (1) no longer make sense. This motivates us to introduce *optimal transport merging*, a procedure that maps vector-valued scores $S(x, y) \in \mathbb{R}^d$ to scalar scores, allowing direct application of one-dimensional conformal prediction. Specifically, we define:

$$S_{\text{OT}-\text{CP}}(x, y) = \|T^\star \circ S(x, y)\|\,,$$

where $T^\star$ is the optimal transport map pushing the distribution of scores onto the uniform distribution on the unit ball defined in Equation (3). This transformation effectively merges the multivariate geometry into a scalar quantity while preserving ordering, enabling us to return to the well-understood one-dimensional setting. The resulting scalar scores $Z_i = \|S_{\mathrm{OT-CP}}(X_i, Y_i)\|$ can be conformalized using the standard empirical CDF $F_{n_{cal}}$ over the calibration set. At test time, we define the prediction region:

$$\mathcal{R}_\alpha(x) = \left\{y \in \mathcal{Y} : \|T_\epsilon \circ S(x, y)\| \leq q_{1-\alpha}^{n_{cal}}\right\},$$

where T_ϵ is an empirical approximation of $T^\star$, learned on an independent sample. More precisely, we split the dataset $D_n = D_{hold} \cup D_{cal}$ in two part where D_{hold} is used to estimate the entropic map T_ϵ defined in Equation (6) and D_{cal} to obtain $q_{1-\alpha}^{n_{cal}}$ the $(1-\alpha)$-quantile of the transported scores $\{\|S_{\mathrm{OT-CP}}(x, y)\| : \text{ for } (x, y) \in D_{cal}\}$. This reduction to the scalar case ensures finite-sample marginal coverage; Equation (1) directly implies $\mathbb{P}(Y_{n+1} \in \mathcal{R}_\alpha(X_{n+1})) \geq 1 - \alpha$, while retaining the flexibility of multivariate scoring functions. As $n \to \infty$, $\hat{T}$ and the empirical distribution converge to their population counterparts Yet, the core strength of this method lies in its finite-sample guarantee: by reducing to the one-dimensional case, OTCP inherits the calibration and distribution-free validity of classical conformal prediction, without requiring asymptotic assumptions.

Our proposed conformal prediction framework **OT-CP** with optimal transport merging score function generalizes the **Merge-CP** approaches. More specifically, under the additional assumption that we are transporting a source Gaussian (resp. uniform) distribution to a target Gaussian (resp. uniform) distribution, the transport map is affine Gelbrich (1990); Muzellec & Cuturi (2018) with a positive definite linear map term.

3.2 Localized Kantorovich Conformal Prediction

We generalize the OTCP score by introducing a unified framework for localizing the transport map using kernel-weighted distributions. These variants aim to better capture local structure in the residuals and improve approximation to conditional coverage by adapting the conformal score to the geometry of the input space. Consider a kernel function $H : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ that measures similarity between points. Following del Barrio et al. (2024), we define a kernel-weighted empirical distribution over residuals obtained in in the hold-out set D_{hold} in each region A_k as

$$\mathbb{P}_{A_k} = \sum_{j=1}^{n} w_j(A_k)\, \delta_{Z_j}, \text{ where } w_j(A_k) = \frac{H(X_j, c_k)}{\sum_{\ell=1}^{n} H(X_\ell, c_k)},$$

and c_k denotes a representative point of the region (e.g., its centroid). The choice of kernel H controls the weighting behavior. For example, setting $H(x, c_k) = \mathbb{1}_{x \in A_k}$ recovers uniform weights over the region (i.e., $w_j(A_k) = 1/n_k$ for $X_j \in A_k$), while a Gaussian kernel $H(x, c_k) = \exp\left(-\|x - c_k\|^2/2\sigma^2\right)$ yields soft, distance-based weighting. More generally, H may incorporate adaptive metrics or local density estimates. The kernel formulation thus unifies hard and soft partitioning schemes: indicator kernels recover classical OTCP with uniform weights, while smooth kernels allow finer, geometry-aware localization within each region. We define a single map for each region

$$S_{\mathrm{OTCP}}^{A_k}(x, y) = \|T_{A_k} \circ S(x, y)\|, \quad \forall x \in A_k.$$

For finite-sample conditional coverage guarantees require the partitions A_k to have positive diameter: $\mathrm{diam}(A_k) > 0$. This ensures that each region contains enough samples to meaningfully estimate the local transport map. Under regularity conditions, the map T_{A_k} pushing $\mathbb{P}(\cdot \mid A_k)$ to the reference measure converges to the conditional transport map $T_{Y|X=x}$ as $A_k \to \{x\}$ and $n \to \infty$, see (del Barrio et al., 2024, Theorem 3.2 and Corollary 3.4) Similar assumptions are required in methods like Distributional Conformal Prediction Chernozhukov et al. (2021a), where conditional coverage relies on accurate estimation of $\mathbb{P}_{Y|X=x}$. Theoretically, consistency requires the number of partitions K to grow with sample size, typically under the conditions $K \to \infty$ and $K/n \to 0$. This balances two competing requirements: small regions to approximate the true conditional law, and large regions to ensure enough local samples for stability. In practice, even modest choices (e.g., $K = 5$ or 10) already improve conditional coverage.

Approximate Conditional Coverage The partition splits $\mathcal{X}$ into K disjoint cells $A_1, \ldots, A_K$ (each with $\mathrm{diam}(A_k) > 0$), and a transport map T_{A_k} is computed for each. Conditional coverage is approximated by applying OTCP separately in each cell, leveraging conditional exchangeability. Indeed, if $(X_1, \ldots, X_n, \ldots) \in$

$\bigcup_{i=1}^{K} A_k$ is exchangeable, then for each k, the conditional sequence $(X_i \mid A_k)$ is also exchangeable. Indeed, by definition, for any permutation π of $\{1, \dots, n\}$, we have

$$(X_1, \dots, X_n) \overset{d}{=} (X_{\pi(1)}, \dots, X_{\pi(n)}).$$

Now, fix k, and any subsequence $(X_i : X_i \in A_k)$ and let σ be any permutation of the indices i such that $X_i \in A_k$. Since the full joint law is invariant under *all* permutations, in particular it's invariant under permutations that only rearrange the values inside A_k and leave the others unchanged. Hence

$$(X_i : X_i \in A_k) \overset{d}{=} (X_{\sigma(i)} : X_i \in A_k).$$

Similar arguments used for local validity can be dated at least to Lei & Wasserman (2012); Vovk (2012). Now, it suffices to apply Proposition 2.1 to empirical conditional distribution $\mathbb{P}_{A_k}$ on each partition A_k to obtain:

$$\mathbb{P}(Y_{n+1} \in \mathcal{R}(X_{n+1}) \mid X_{n+1} \in A_k) \geq 1 - \alpha, \quad \forall k \in [K].$$

Computational Trade-offs Localizing the OTCP score either through clustering or kernel-based weights introduces a trade-off between conditional accuracy and computational cost. Hard clustering (e.g., K-means) is particularly efficient: given N residuals and a reference distribution of size M (e.g., $M = 8192$), a global Sinkhorn map costs $O(NM)$. Partitioning the data into k clusters of sizes $N_1, \dots, N_k$ leads to a total cost $\sum_{i=1}^{k} O(N_i M) = O(NM)$, matching the global cost while producing localized maps better suited to capture input-space heterogeneity. In contrast, soft localization solves k OT problems over the full dataset, each reweighted by a kernel centered at a representative point. Each problem still incurs $O(NM)$ cost, yielding an overall complexity of $O(kNM)$ scaling linearly with k. Although more expensive, soft methods are trivially parallelizable: both the kernel weights and Sinkhorn maps for different centers can be computed independently. A third option is pointwise localization at test time, where a new transport map is computed from the K nearest training points for each test input. This highly adaptive method, similar to that proposed in del Barrio et al. (2024), incurs $O(KM)$ per test query. However, in our experiments, it brought no notable gains over hard clustering. In summary, hard clustering strikes a practical balance: it improves conditional adaptation without exceeding the global cost, and remains especially effective in high dimensions where fine-grained structure is hard to estimate globally.

3.3 Implementation with the Entropic Map

We assume access to residuals samples $(z_1, \dots, z_n)$, and a discretization of the uniform grid on the sphere, $(u_1, \dots, u_m)$, with sizes n, m that will be usually different, $n \neq m$. Learning the entropic map estimator as in Section 3.3 requires running the Sinkhorn algorithm for a given regularization ε on a $n \times m$ cost matrix. At test time, for each evaluation, computing the Gibbs weights requires computing the distances of a new score z to the uniform grid. The complexity is therefore $O(nm)$ when training the map and conformalizing its norms, and $O(m)$ to transport a conformity score for a given y.

Sampling on the sphere. Following Hallin et al. (2021), we begin by constructing the target distribution $\mathbb{U}_{n+1}$ as a discretized version of a spherical uniform distribution. It is defined such that the total number of points $n + 1 = n_R n_S + n_o$, where n_o points are at the origin: n_S unit vectors $\mathbf{u}_1, \dots, \mathbf{u}_{n_S}$ are uniform on the sphere and n_R radius are regularly spaced as $\left\{\frac{1}{n_R}, \frac{2}{n_R}, \dots, 1\right\}$. The grid discretizes the sphere into layers of concentric shells, with each shell containing n_S equally spaced points along the directions determined by the unit vectors. The discrete spherical uniform distribution places equal mass over each points of the grid, with $n_o/(n + 1)$ mass on the origin and $1/(n + 1)$ on the remaining points. This ensures isotropic sampling at fixed radius onto $[0, 1]$. Additionally, we borrow inspiration from the review provided in (Nguyen et al., 2024) and pick their *Gaussian based* mapping approach (Basu, 2016). This consists of mapping a low-discrepancy sequence $w_1, \dots, w_L$ on $[0, 1]^d$ to a potentially low-discrepancy sequence $\theta_1, \dots, \theta_L$ on $\mathbb{S}^{d-1}$ through the mapping $\theta = \Phi^{-1}(w)/\|\Phi^{-1}(w)\|_2$, where Φ^{-1} is the inverse CDF of $\mathcal{N}(0, 1)$ applied entry-wise.

3.4 Differences with Vector Quantile Regression (VQR)

VQR aims to estimate the full conditional quantile map using optimal transport, offering a rich, model-based representation of multivariate uncertainty to construct such maps under a mean-independence constraint

Carlier et al. (2016); Rosenberg et al. (2022); Pegoraro et al. (2023), extending earlier L_1-based multivariate quantile approaches Chaudhuri (1996). In contrast, our framework does not assume access to a predictive model and generalizes conformal prediction to vector-valued non-conformity scores. VQR formulates quantile regression as solving a large-scale linear program over a transport plan $\Pi \in \mathbb{R}^{T^d \times N}$ between a T^d-point quantization grid in $[0,1]^d$ and N data points, subject to: $\Pi \mathbf{1}_N = \frac{1}{T^d}\mathbf{1}_{T^d}$, and $\Pi^\top \mathbf{u} = \bar{X}$, where $\mathbf{u} \in \mathbb{R}^{T^d \times d}$ is the grid of quantile targets and $\bar{X}$ the empirical mean of covariates. These constraints ensure uniform marginals and mean-independence. However, the number of variables scales as $T^d N$, which becomes computationally prohibitive beyond $d = 2$. In our experiments, the (NL)VQR implementation hardcodes $T^d \approx 8000$, limiting scalability. While dual relaxations mitigate some of the computational burden, they do not yield explicit quantile maps. Moreover, enforcing monotonicity requires a second OT step via Vector Monotone Rearrangement (VMR) Rosenberg et al. (2022), adding cost at test time. VQR also lacks built-in coverage guarantees and typically relies on scalar conformalization over post-estimated regions Feldman et al. (2023), e.g., using scores such as $S(x, y) = \max(\text{dist}(y, R(x)), \text{dist}(y, R^c(x)))$. By contrast, OTCP estimates a transport map from residuals to a continuous reference (e.g., uniform on the sphere) via Sinkhorn regularization with complexity $O(NM)$, avoiding gridding and scaling well with dimension. In our evaluation, we matched the number of OTCP target points to the VQR grid ($T^d \approx 8000$) and found OTCP outperformed VQR in coverage tasks. This should not be seen as a critique of VQR's modeling power, but as evidence that its generality does not necessarily yield practical benefits for calibrated prediction in higher dimensions.

4 Experiments

4.1 Setup and Metrics

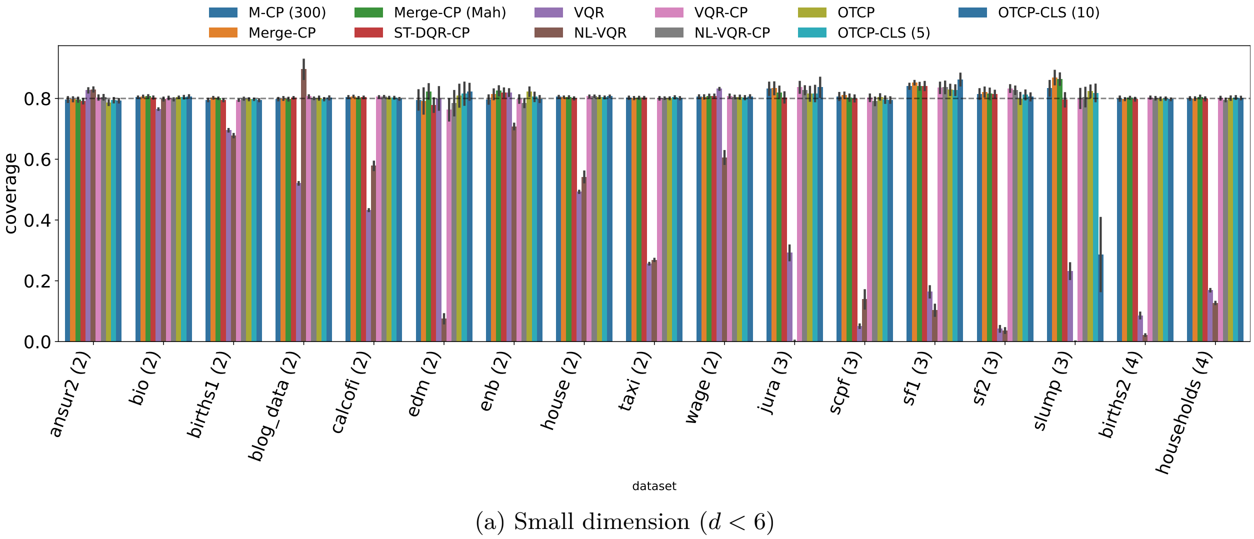
We borrow the experimental setting provided by Dheur et al. (2025) and benchmark multivariate conformal methods on a total of 24 tabular datasets. Total data size n in these datasets ranges from 103 to 50,000, with input dimension p ranging from 1 to 348, and output dimension d ranging from 2 to 16. We adopt their approach, which is to rely on a multivariate quantile function forecaster (MQF^2, Kan et al., 2022), a normalizing flow that is able to quantify output uncertainty conditioned on input x. However, in accordance with our stance mentioned in the background section, we will only assume access to the conditional mean (point-wise) estimator for **OT-CP** . If one has access to a smooth joint or conditional distribution $P_{Y|X}$, one may define vector-valued scores using samples from $P_{Y|X}$ (as in PCP), or extract a conditional map $T_x \# P_{Y|X=x} = \mathbb{U}$ to a reference distribution. Our framework remains compatible with such maps: if T pushes Z to a reference, then any invertible function f induces a new map $T_f = f \circ T \circ f^{-1}$, allowing transport on conformity scores $s(x, y)$ via composition. While promising, estimating such conditional maps can be computationally intensive and is left for future work.

We incorporated several vector quantile regression (VQR) baselines into the evaluation: `ST-DQR-CP`, `VQR`, `NL-VQR`, `VQR-CP`, and `NL-VQR-CP`. We also implemented a simple localization strategy for OTCP based on k-means clustering, resulting in two variants: `OTCP-CLS (5)` and `OTCP-CLS (10)`, where the number in parentheses indicates the number of clusters used. Furthermore, we report two conditional coverage metrics to assess the quality of local calibration. As is common in the field, we evaluate the methods using several metrics, including marginal coverage (MC), and mean region size (Size). The latter is using importance sampling, leveraging (when computing test time metrics only), the generative flexibility provided by the MQF^2 as an invertible flow. See (Dheur et al., 2025) and their code for more details.

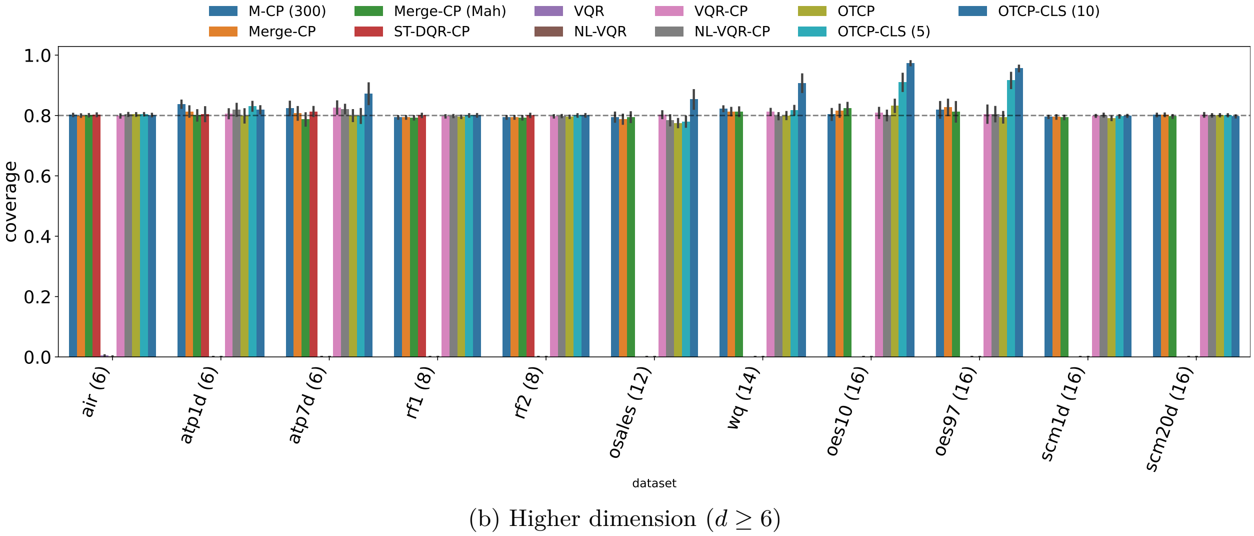
4.2 Hyperparameter Choices

We apply default parameters for all three competing methods, **M-CP** and **Merge-CP** , using (or not) the Mahalanobis correction and set a target coverage $1 - \alpha = 0.8$. For **M-CP** using conformalized quantile regression boxes, we follow (Dheur et al., 2025) and leverage the empirical quantiles return by MQF^2 to compute boxes (Zhou et al., 2024).

OT-CP : our implementation requires tuning two important hyperparameters: the entropic regularization ε and the total number of points used to discretize the sphere m, not necessarily equal to the input data sample size n. These two parameters describe a fundamental statistical and computational trade-off. On the

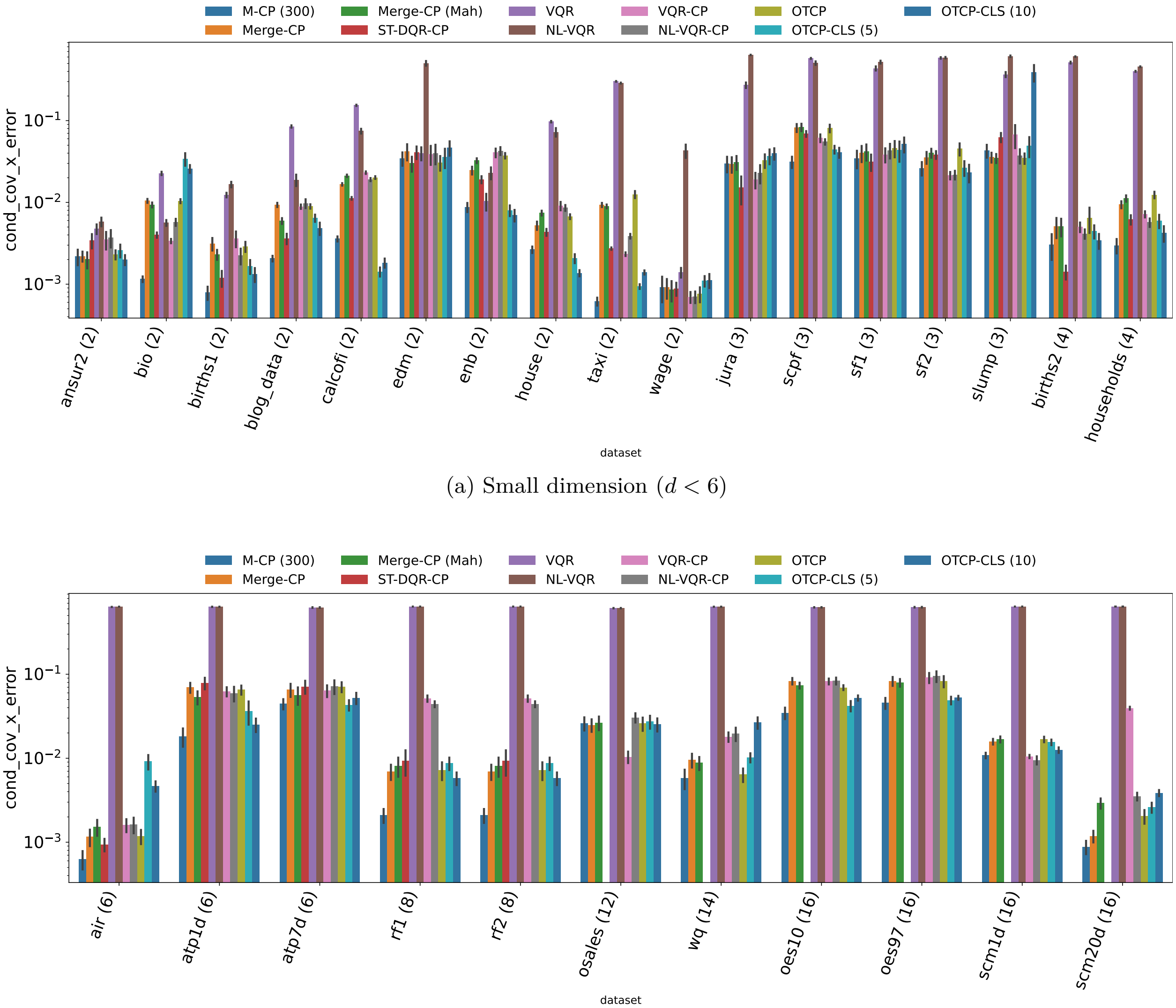


(a) Small dimension ($d < 6$)



(b) Higher dimension ($d \geq 6$)

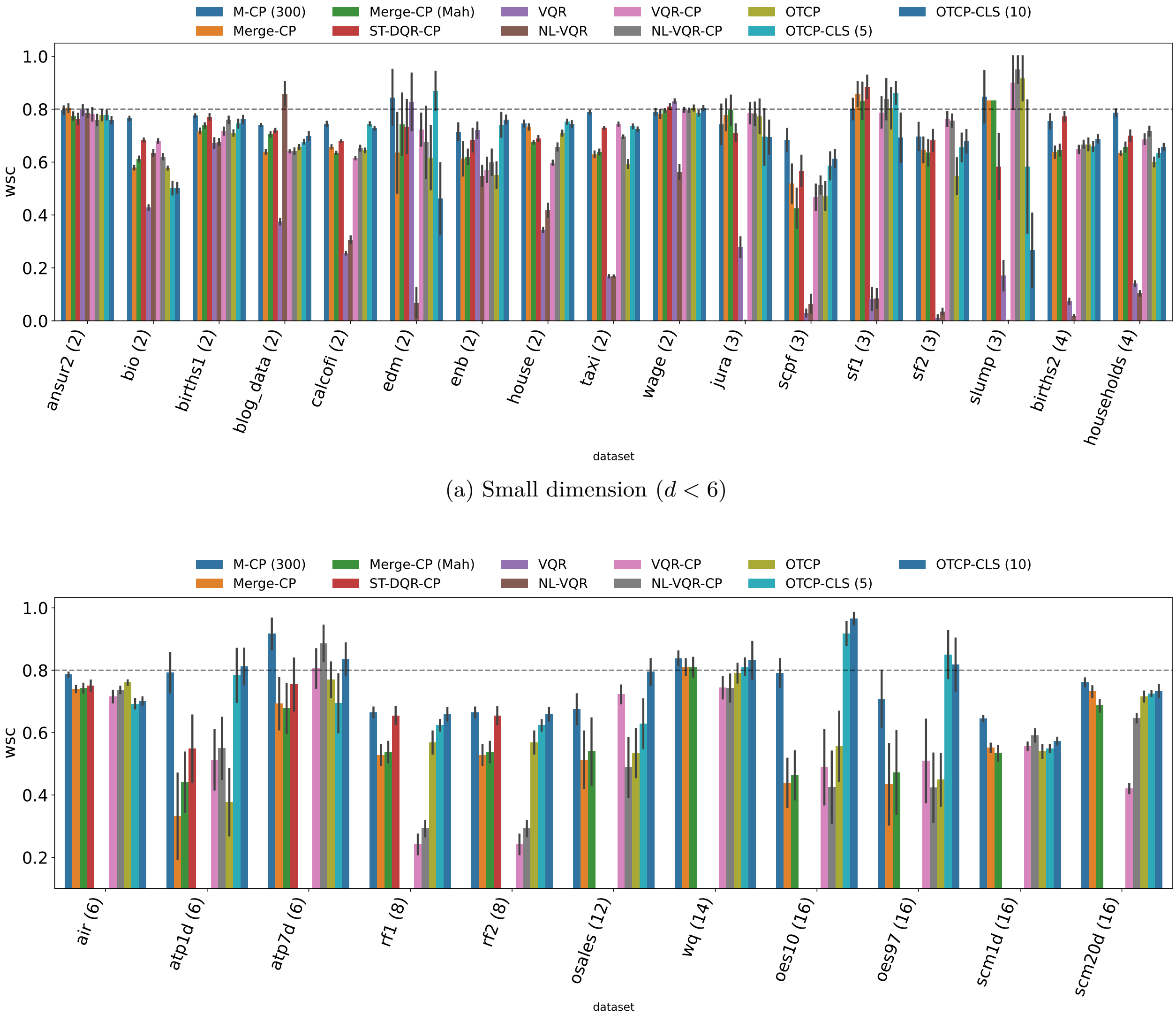
Figure 1: Marginal Coverage obtained wrt dimension of the score function



(a) Small dimension ($d < 6$)

(b) Higher dimension ($d \geq 6$)

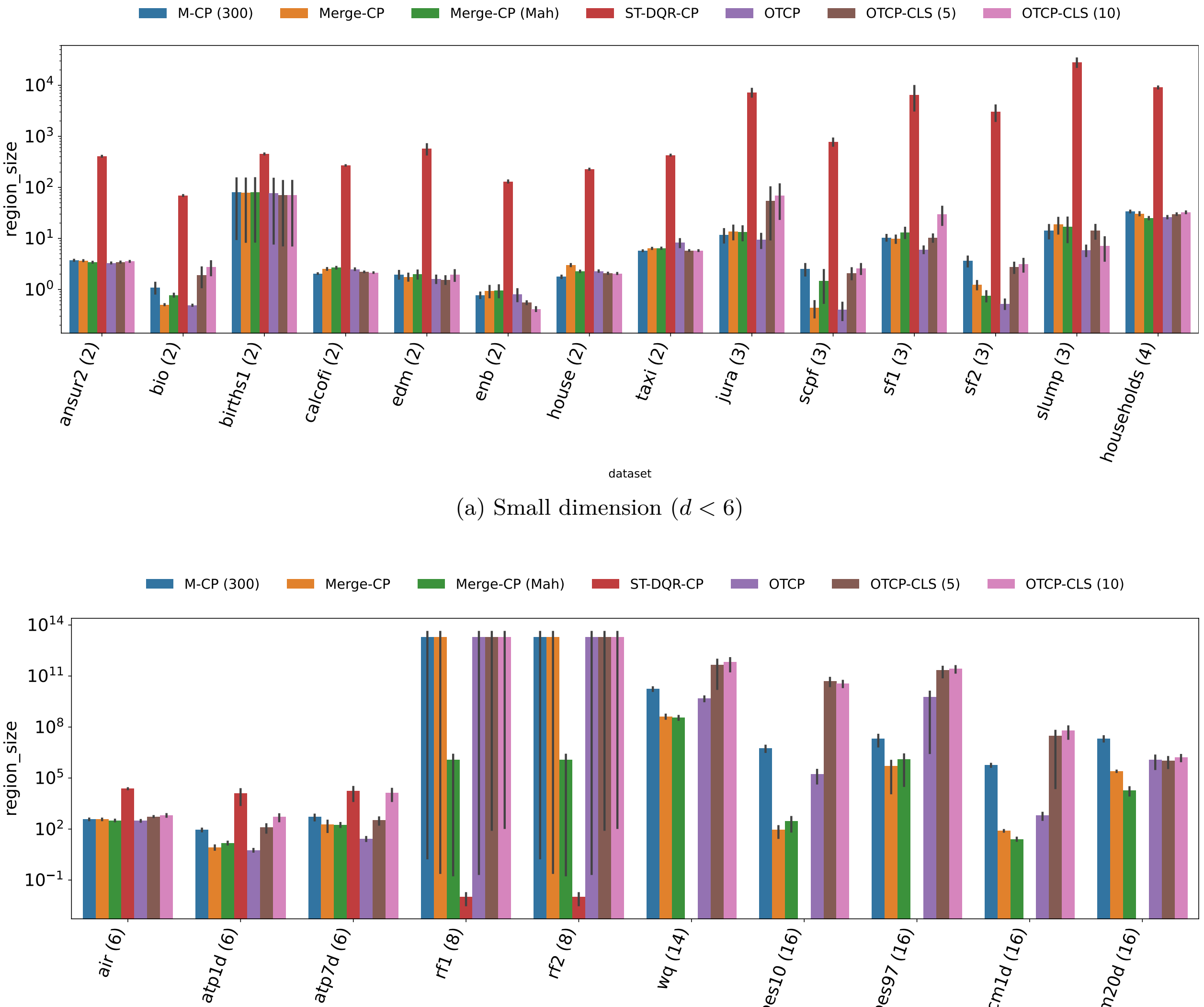
Figure 2: Conditional Coverage Error (CEC-X) wrt dimension of the score function



(a) Small dimension ($d < 6$)

(b) Higher dimension ($d \geq 6$)

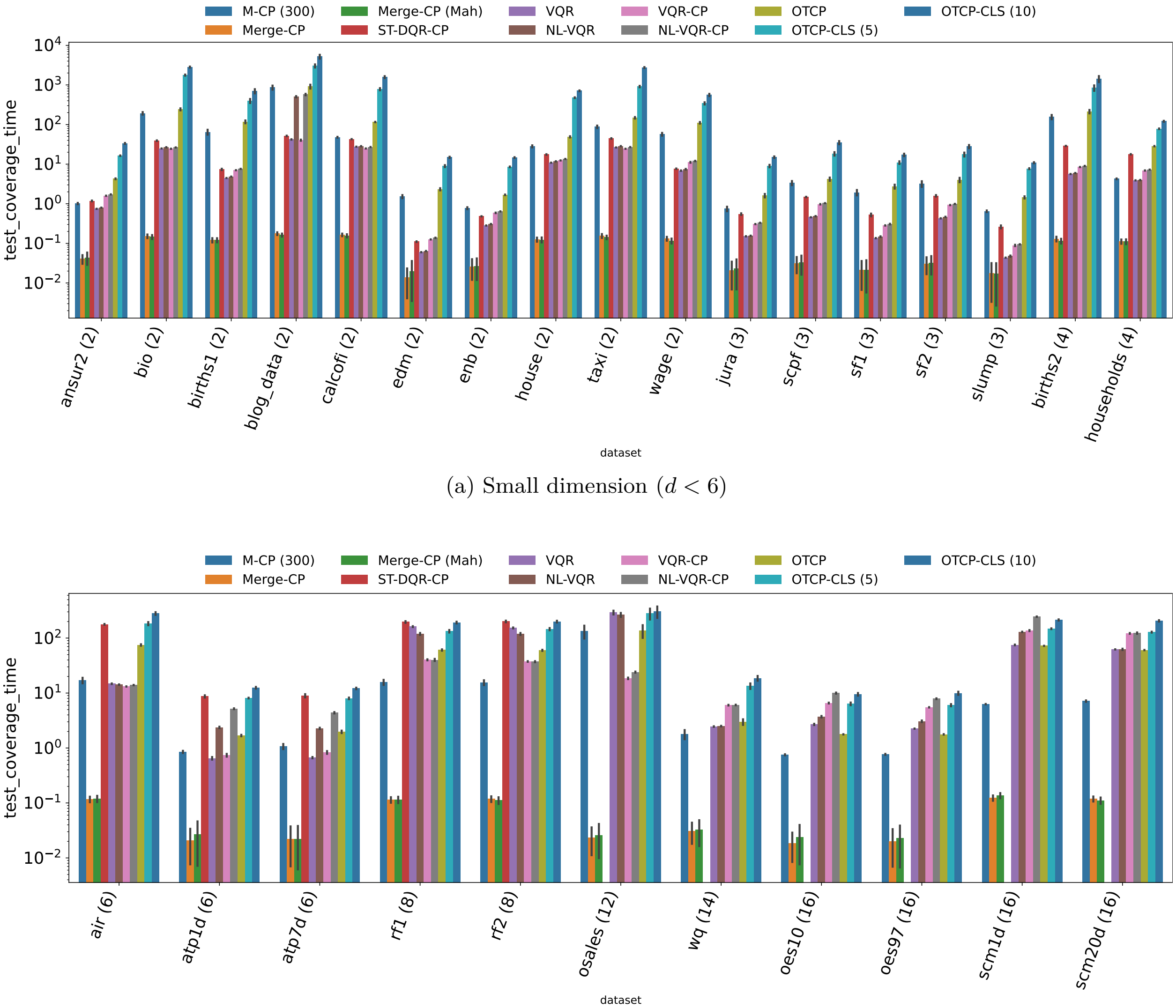
Figure 3: Worst case Conditional Coverage (WSC) wrt dimension of the score function



(a) Small dimension ($d < 6$)

(b) Higher dimension ($d \geq 6$)

Figure 4: Region size as function of dimension of the score function.



(a) Small dimension ($d < 6$)

(b) Higher dimension ($d \geq 6$)

Figure 5: Computational time as function of dimension of the score function.

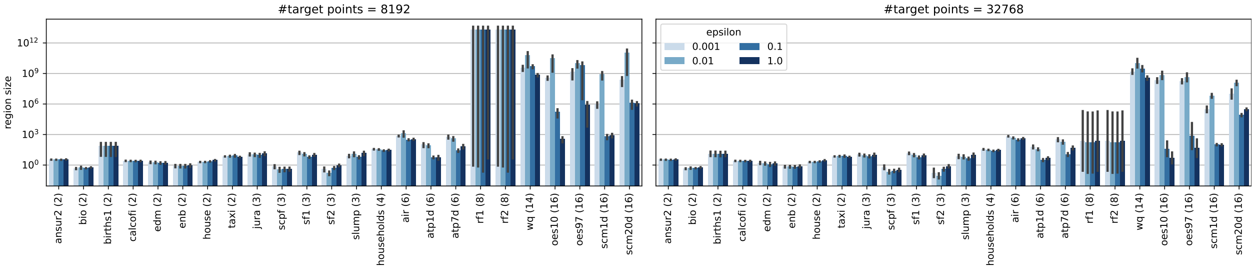


Figure 6: This plot details the impact of the two important hyperparameters one needs to set in **OT-CP** : number of target points m sampled from the uniform ball and the ε regularization level. As can be seen, larger sample size m improves region size (smaller the better) for roughly all datasets and regularization strengths. On the other hand, one must tune ε to operate at a suitable regime: not too low, which results in the well-documented poor statistical performance of unregularized / linear program OT, nor too high, which would lead to a collapse of the entropic map to the sphere. Using OTT-JAX and its automatic normalizations, we see that $\varepsilon = 0.1$ works best overall.

one hand, it is known that increasing m will mechanically improve the ability of T_ε to recover in the limit $T^\star$ (or at least solve the semi-discrete (Peyré & Cuturi, 2019) problem of mapping n data points to the sphere). However, large m incurs a heavier computational price when running the Sinkhorn algorithm. On the other hand, increasing ε improves on *both* computational and statistical aspects, but deviates further the estimated map from the ground truth $T^\star$ to target instead a blurred map. We have experimented with these aspects and derive from our experiments that both m and ε should be increased to track increase in dimension. As a sidenote, we do observe that debiasing the outputs of the Sinkhorn (Sinkhorn, 1964) algorithm does not result in improved results, which agrees with the findings in (Pooladian et al., 2022). We use the OTT-JAX toolbox (Cuturi et al., 2022) to compute these maps.

4.3 Results

We present results by differentiating datasets with small dimension $d \leq 6$ from datasets with higher dimensionality $14 \leq d \leq 16$, that we expect to be more challenging to handle with OT approaches, owing to the curse of dimensionality that might degrade the quality of multivariate quantiles. Results in Figure 1a and Figure 2a indicate an improvement (smaller region for similar coverage) on 15 out of 18 datasets in lower dimensions, this edge vanishing in the higher-dimensional regime. Compared to Vector Quantile Regression approaches, we observed that VQR-based methods tend to underperform on our benchmark tasks, likely due to scalability issues and the absence of inherent coverage guarantees. In contrast, the localized versions of OTCP (`OTCP-CLS`) demonstrate improved conditional coverage, consistent with our expectations. These results confirm the benefit of incorporating even simple localization techniques into OTCP to better adapt to heterogeneous regions of the input space. Ablations provided in Appendix highlight the role of ε and m, the entropic regularization strength and the sphere size respectively. These results show that results for high m tend to be better but more costly, while the tuning of the regularization strength ε needs to be tuned according to dimension (Vacher & Vialard, 2022). Finally, Figure 7 provides an illustration of the non-elliptic CP regions outputted by **OT-CP** , by pulling back the rescaled uniform sphere using the inverse entropic mapping.

5 Conclusion

We have proposed **OT-CP** , a new approach that can leverage a recently proposed formulation for multivariate quantiles that uses optimal transport theory and optimal transport map estimators. We show the theoretical soundness of this approach, but, most importantly, demonstrate its applicability throughout a broad range of tasks compiled by (Dheur et al., 2025). Compared to similar baselines that either use a conditional mean regression estimator (**Merge-CP**), or more involved quantile regression estimators (**M-CP**), **OT-CP** shows overall superior performance, while incurring, predictably, a higher train / calibration time cost.

Our approach relies on the estimation of an optimal transport map from a finite set of score samples, a task with known statistical challenges that become acute in high dimensions (Chewi et al., 2024). As the dimension d of the score vector grows, the number of samples (n_{hold}) needed to faithfully represent the geometry of the score distribution increases exponentially. This curse of dimensionality directly impacts **OT-CP** 's performance. Our results for datasets with $d \geq 6$ clearly illustrate this limitation. The diminishing advantage of OT-CP over simpler baselines in these settings is not a failure of the concept, but rather a reflection of the fundamental statistical cost of non-parametric map estimation. While we demonstrate that a pragmatic choice of hyperparameters can still yield reasonable results, this highlights a clear boundary for the current applicability of our method. Overcoming this will likely require moving beyond unstructured point clouds, for instance by exploring OT methods tailored for structured distributions or factorized assumptions, which remains a promising avenue for future research.

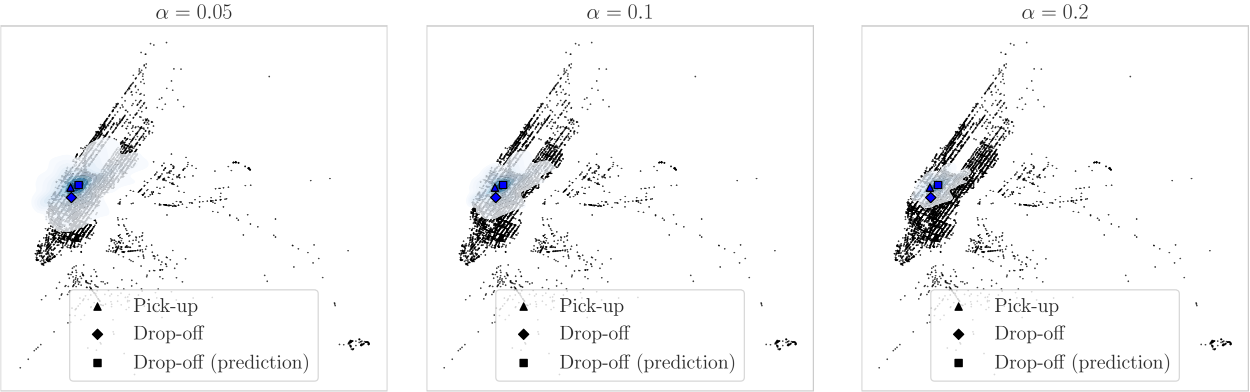


Figure 7: K-means localized Conformal sets, recovered by mapping back the reduced sphere on the Manhattan map, on a prediction for the `taxi` dataset. We use the inverse entropic map mentioned in Section 3.3, mapping back the gridded sphere of size $m = 2^{15}$ for each level, plotting its outer contour.

6 Concurrent Work.

In work developed independently and appearing on arXiv at nearly the same time, Thurin et al. (2025) proposed to leverage OT in CP with a similar approach, deriving a similar CP set and analyzing a variant with asymptotic conditional coverage under additional regularity assumptions. However, our methods differ in several key aspects. On the computational side, our implementation leverages general entropic maps (Section 3.3) without compromising finite-sample coverage guarantees, and we prove that validity holds even when using approximate maps. This yields a procedure that scales as $O(nm)$, decouples the number of calibration points n from the target grid size m, and allows the map to be pre-trained once and reused at test time, which is crucial for large-scale applications. In contrast, their approach requires solving a linear assignment problem, using for instance the Hungarian algorithm, which has cubic complexity $O(n^3)$ in the number of target points, and which also requires having a target set on the sphere that is of the same size as the number of input points. With our notations in Section 3.3, they require $n = m$, whereas we set m to anywhere between 2^{12} and 2^{15}, independently of n, providing smoother approximations in high dimension. Beyond computation, we also introduce localized and cluster-based variants of entropic maps to improve conditional coverage efficiently, offering a different route to adaptivity than their kNN-based OT-CP+. Finally, we provide extensive empirical benchmarks and sensitivity analyses (grid size m, regularization ε) across dozens of multivariate regression tasks, giving practical guidance for using OT-based CP in real-world settings.

References

Angelopoulos, A. N., Barber, R. F., and Bates, S. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

Balasubramanian, V., Ho, S.-S., and Vovk, V. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Basu, K. *Quasi-Monte Carlo Methods in Non-Cubical Spaces*. Stanford University, 2016.

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. Testing for outliers with conformal p-values. *arXiv preprint arXiv:2104.08279*, 2021.

Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4), 1991. doi: 10.1002/cpa.3160440402.

Carlier, G., Chernozhukov, V., and Galichon, A. Vector quantile regression: an optimal transport approach. 2016.

Cella, L. and Ryan, R. Valid distribution-free inferential models for prediction. *arXiv preprint arXiv:2001.09225*, 2020.

Chaudhuri, P. On a geometric notion of quantiles for multivariate data. *Journal of the American statistical association*, 91(434):862–872, 1996.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223 – 256, 2017. doi: 10.1214/16-AOS1450. URL `https://doi.org/10.1214/16-AOS1450`.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. Exact and robust conformal inference methods for predictive machine learning with dependent data. *Conference On Learning Theory*, 2018.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021a.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864, 2021b.

Chewi, S., Niles-Weed, J., and Rigollet, P. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 2024.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013.

Cuturi, M., Teboul, O., and Vert, J.-P. Differentiable ranking and sorting using optimal transport. *Advances in neural information processing systems*, 32, 2019.

Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., and Teboul, O. Optimal transport tools (ott): A jax toolbox for all things wasserstein, 2022. URL `https://arxiv.org/abs/2201.12324`.

del Barrio, E., Sanz, A. G., and Hallin, M. Nonparametric multiple-output center-outward quantile regression. *Journal of the American Statistical Association*, pp. 1–15, 2024.

Dheur, V., Fontana, M., Estievenart, Y., Desobry, N., and Taieb, S. B. Multi-output conformal regression: A unified comparative study with new conformity scores, 2025. URL `https://arxiv.org/abs/2501.10533`.

Feldman, S., Bates, S., and Romano, Y. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023.

Fisch, A., Schuster, T., Jaakkola, T., and Barzilay, R. Few-shot conformal prediction with auxiliary tasks. *ICML*, 2021.

Gammerman, A., Vovk, V., and Vapnik, V. Learning by transduction, 1998.

Gelbrich, M. On a formula for the l^2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1), 1990.

Guha, E., Natarajan, S., Möllenhoff, T., Khan, M. E., and Ndiaye, E. Conformal prediction via regression-as-classification. *arXiv preprint arXiv:2404.08168*, 2024.

Hallin, M., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139 – 1165, 2021. doi: 10.1214/20-AOS1996. URL https://doi.org/10.1214/20-AOS1996.

Hallin, M., La Vecchia, D., and Liu, H. Center-outward r-estimation for semiparametric varma models. *Journal of the American Statistical Association*, 117(538):925–938, 2022.

Hallin, M., Hlubinka, D., and Hudecová, Š. Efficient fully distribution-free center-outward rank tests for multiple-output regression and manova. *Journal of the American Statistical Association*, 118(543): 1923–1939, 2023.

Henderson, I., Mazoyer, A., and Gamboa, F. Adaptive inference with random ellipsoids through conformal conditional linear expectation. *arXiv preprint arXiv:2409.18508*, 2024.

Ho, S.-S. and Wechsler, H. Query by transduction. *IEEE transactions on pattern analysis and machine intelligence*, 2008.

Holland, M. J. Making learning more transparent using conformalized performance prediction. *arXiv preprint arXiv:2007.04486*, 2020.

Izbicki, R., Shimizu, G., and Stern, R. B. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32, 2022.

Johnstone, C. and Cox, B. Conformal uncertainty sets for robust optimization. In Carlsson, L., Luo, Z., Cherubin, G., and An Nguyen, K. (eds.), *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pp. 72–90. PMLR, 08–10 Sep 2021. URL https://proceedings.mlr.press/v152/johnstone21a.html.

Kan, K., Aubet, F.-X., Januschowski, T., Park, Y., Benidis, K., Ruthotto, L., and Gasthaus, J. Multivariate quantile function forecaster. In *International Conference on Artificial Intelligence and Statistics*, pp. 10603–10621. PMLR, 2022.

Katsios, K. and Papadopulos, H. Multi-label conformal prediction with a mahalanobis distance nonconformity measure. In Vantini, S., Fontana, M., Solari, A., Boström, H., and Carlsson, L. (eds.), *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 230 of *Proceedings of Machine Learning Research*, pp. 522–535. PMLR, 09–11 Sep 2024. URL https://proceedings.mlr.press/v230/katsios24a.html.

Kumar, B., Lu, C., Gupta, G., Palepu, A., Bellamy, D., Raskar, R., and Beam, A. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Laxhammar, R. and Falkman, G. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Annals of Mathematics and Artificial Intelligence*, 2015.

Lei, J. and Wasserman, L. Distribution free prediction bands. *arXiv preprint arXiv:1203.5422*, 2012.

Lin, Z., Trivedi, S., and Sun, J. Conformal prediction intervals with temporal dependence. *Transactions of Machine Learning Research*, 2022.

Lu, C., Lemay, A., Chang, K., Höbel, K., and Kalpathy-Cramer, J. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12008–12016, 2022.

Muzellec, B. and Cuturi, M. Generalizing point embeddings using the wasserstein space of elliptical distributions. *Advances in Neural Information Processing Systems*, 31, 2018.

Nguyen, K., Bariletto, N., and Ho, N. Quasi-monte carlo for 3d sliced wasserstein. In *The Twelfth International Conference on Learning Representations*, 2024.

Park, J. W., Tibshirani, R., and Cho, K. Semiparametric conformal prediction. *arXiv preprint arXiv:2411.02114*, 2024.

Pegoraro, M., Vedula, S., Rosenberg, A. A., Tallini, I., Rodolà, E., and Bronstein, A. M. Vector quantile regression on manifolds. *arXiv preprint arXiv:2307.01037*, 2023.

Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11, 2019.

Pooladian, A.-A. and Niles-Weed, J. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

Pooladian, A.-A., Cuturi, M., and Niles-Weed, J. Debiaser beware: Pitfalls of centering regularized transport maps. In *International Conference on Machine Learning*, pp. 17830–17847. PMLR, 2022.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2023.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Rosenberg, A. A., Vedula, S., Romano, Y., and Bronstein, A. M. Fast nonlinear vector quantile regression. *arXiv preprint arXiv:2205.14977*, 2022.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 2008.

Sinkhorn, R. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879, 1964.

Straitouri, E., Wang, L., Okati, N., and Rodriguez, M. G. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*, pp. 32633–32653. PMLR, 2023.

Thurin, G., Nadjahi, K., and Boyer, C. Optimal transport-based conformal prediction, 2025. URL `https://arxiv.org/abs/2501.18991`.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vacher, A. and Vialard, F.-X. Parameter tuning and model selection in optimal transport with semi-dual brenier formulation. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

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

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world.* Springer, 2005.

Wang, Z., Gao, R., Yin, M., Zhou, M., and Blei, D. M. Probabilistic conformal prediction using conditional random samples. *arXiv preprint arXiv:2206.06584*, 2022.

Xu, C. and Xie, Y. Conformal prediction interval for dynamic time-series. *ICML*, 2021.

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pp. 25834–25866. PMLR, 2022.

Zhou, Y., Lindemann, L., and Sesia, M. Conformalized adaptive forecasting of heterogeneous trajectories. *arXiv preprint arXiv:2402.09623*, 2024.

A Appendix

We conduct several additional ablation experiments to further analyze and support the main findings presented in Section 4. These experiments are designed to systematically examine the influence of key hyperparameters and experimental settings on the coverage, runtime metrics and region size across different datasets and dimensionalities. By isolating and varying specific components of our approach and baselines, we aim to provide a clearer insights into the robustness and generalizability of our methods. The results of these ablation studies are summarized in the following figures and tables.

A.1 Benchmark Datasets

Table 1: Summary of the 24 multivariate regression datasets used in our experiments. The datasets are grouped and sorted by dimension to align with the analysis of low-dimensional ($d < 6$) versus high-dimensional ($d \geq 6$) performance.

(a) Small Dimension Datasets ($d < 6$)

Dataset	**Samples (n)**	**Input dim. (p)**	**Output dim. (d)**
ansur2	6068	98	2
bio	45730	8	2
births1	1577	11	2
calcofi	50000	1	2
edm	154	16	2
enb	768	8	2
house	21613	14	2
taxi	50000	4	2
jura	359	15	3
scpf	1137	23	3
sf1	1066	10	3
sf2	1066	10	3
slump	103	7	3
households	7207	4	4

(b) Higher Dimension Datasets ($d \geq 6$)

Dataset	**Samples (n)**	**Input dim. (p)**	**Output dim. (d)**
air	500	11	6
atp1d	337	411	6
atp7d	296	411	6
rf1	9005	64	8
rf2	9005	64	8
wq	1060	16	14
oes10	403	298	16
oes97	263	262	16
scm1d	9803	280	16
scm20d	8966	60	16

A.2 Benchmark Metrics

We evaluate the conformal methods using several classical metrics that assess both coverage properties and region efficiency. We follow and refer to Dheur et al. (2025) for more details

Region size. Smaller regions are preferred for sharper uncertainty quantification, provided coverage guarantees are maintained. The size of a prediction region $\hat{R}(x)$ is defined as

$$|\hat{R}(x)| = \int_{\mathcal{Y}} \mathbf{1}\{y \in \hat{R}(x)\}\, dy. \tag{7}$$

Since this integral is intractable in high dimensions, we approximate it using importance sampling with the predictive density $\hat{f}(y \mid x)$:

$$|\hat{R}(x)| \approx \frac{1}{K} \sum_{k=1}^{K} \frac{\mathbf{1}\{\hat{Y}^{(k)} \in \hat{R}(x)\}}{\hat{f}(\hat{Y}^{(k)} \mid x)}, \quad \hat{Y}^{(k)} \sim \hat{f}(\cdot \mid x). \tag{8}$$

Worst Slab Coverage (WSC). WSC quantifies how well coverage is preserved across all directions in the input space, capturing conditional validity. For a direction $v \in \mathbb{R}^d$, the slab coverage is defined as

$$WSC_v = \inf_{a<b} \Big\{ \hat{P}_{D_{\text{test}}}\big(y_i \in \hat{R}(x_i) \,\big|\, a \le v^\top x_i \le b\big) \quad : \quad \hat{P}_{D_{\text{test}}}(a \le v^\top x_i \le b) \ge \delta \Big\}, \tag{9}$$

where $\delta \in (0, 1]$ is a minimal mass threshold. The worst-slab coverage is then

$$WSC = \min_{v_j \in S^{d-1}} WSC_{v_j}, \tag{10}$$

with S^{d-1} the unit sphere, approximated by sampling random vectors v_j.

Coverage Error Conditional on X (CEC-X). Partition the input space into clusters $A_1, \dots, A_J$ (e.g., using k-means++). Then

$$CEC\text{-}X = \frac{1}{|D_{\text{test}}|} \sum_{i=1}^{|D_{\text{test}}|} \sum_{j=1}^{J} \Big(\hat{P}_{D_{\text{test}}}(y^{(i)} \in \hat{R}(x^{(i)}) \mid x^{(i)} \in A_j) - (1-\alpha) \Big)^2. \tag{11}$$

Rationale: Measures deviations from the target coverage $1-\alpha$ within regions of the input space.

Coverage Error Conditional on V (CEC-V). More robust to high-dimensional inputs, since conditioning is done on predictive density rather than raw features. Define $V = \hat{f}(\hat{Y} \mid X)$ with $\hat{Y} \sim \hat{f}(\cdot \mid X)$. Construct a feature vector v_x from order statistics of $\log V$, and cluster in this density space. Then compute

$$CEC\text{-}V = \frac{1}{|D_{\text{test}}|} \sum_{i=1}^{|D_{\text{test}}|} \sum_{j=1}^{J} \Big(\hat{P}_{D_{\text{test}}}(y^{(i)} \in \hat{R}(x^{(i)}) \mid v_{x^{(i)}} \in A_j) - (1-\alpha) \Big)^2. \tag{12}$$

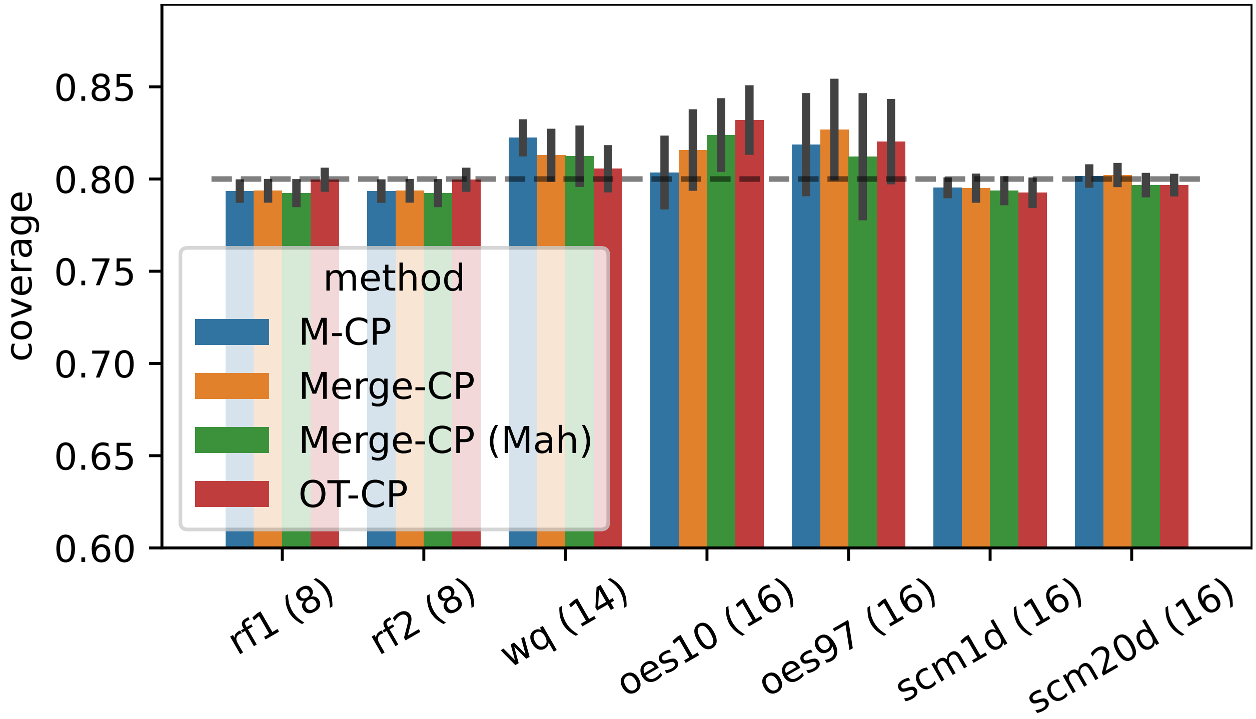


Figure 8: Coverage for higher dimensional datasets, corresponding to the setting displayed in Figure 1b.

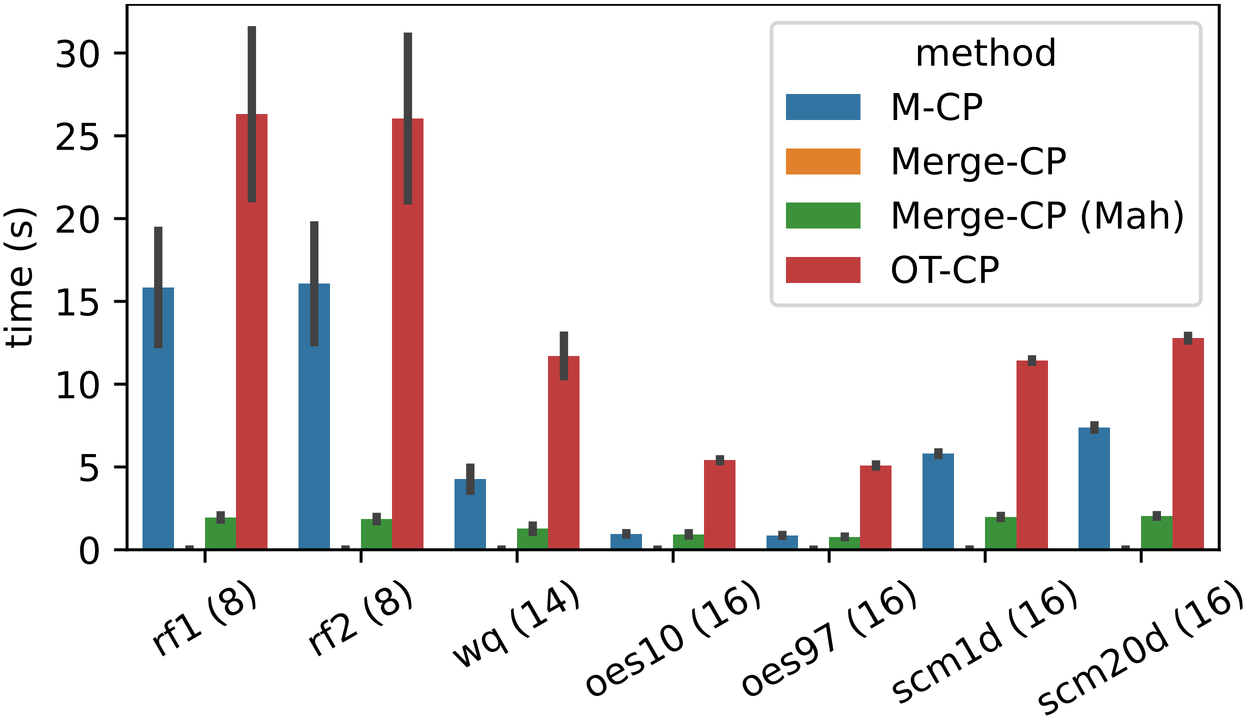


Figure 9: Runtimes for higher dimensional datasets, corresponding to the setting displayed in Figure 5b.

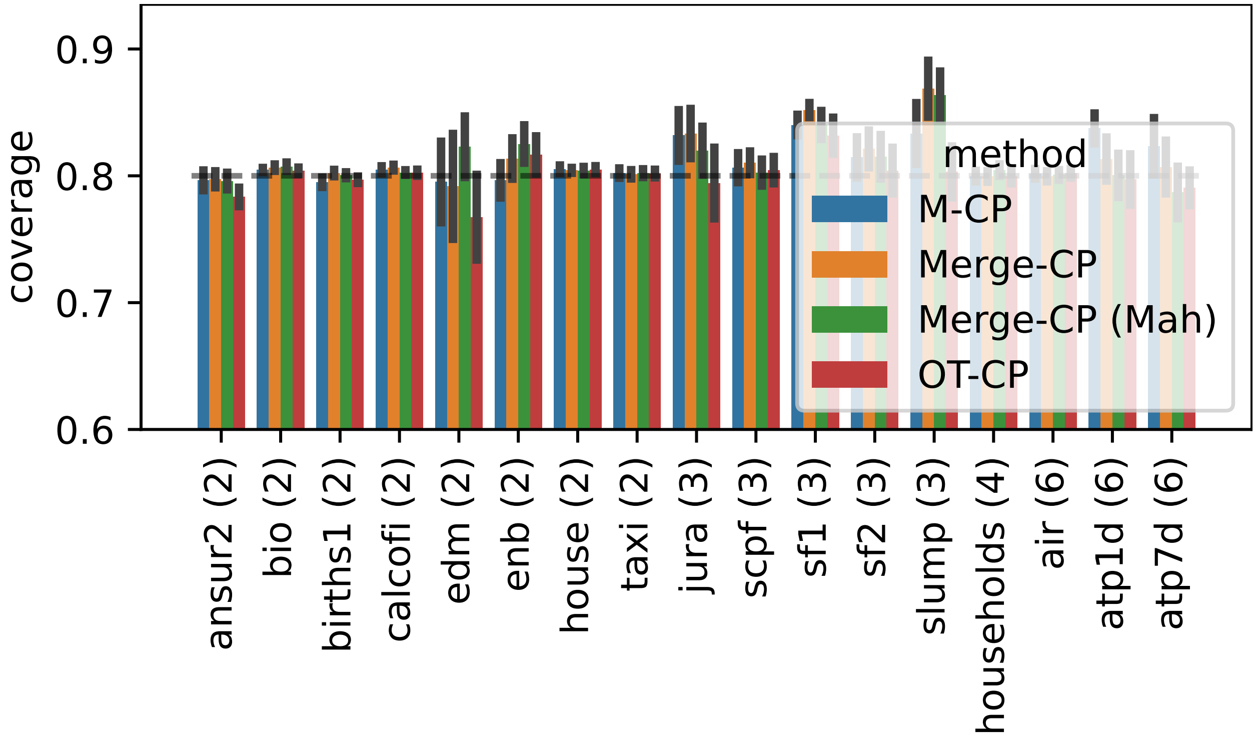


Figure 10: Coverage of all baselines on small dimensional datasets, corresponding to the region sizes given in Section 5.

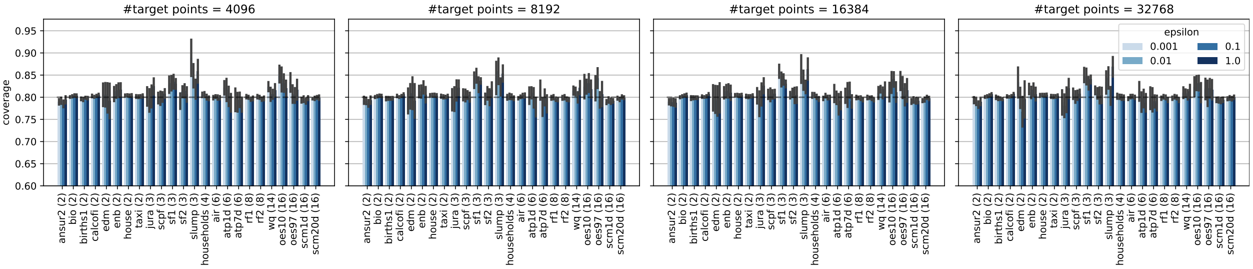


Figure 11: Ablation: coverage quality as a function of hyperparameters, with the setting corresponding to Figure 6.

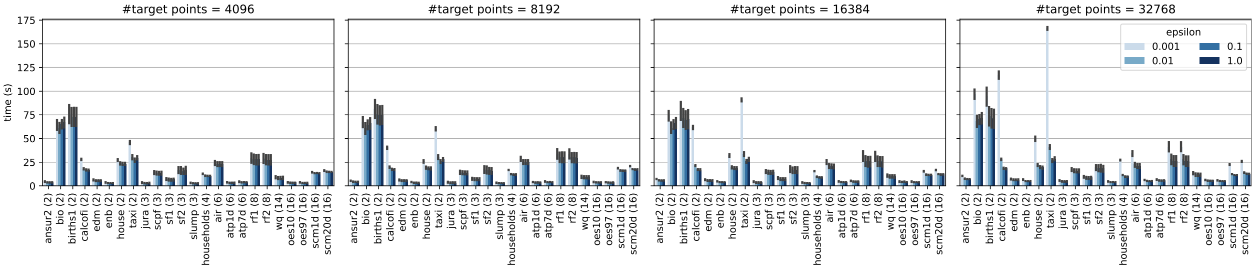


Figure 12: Ablation: running time as a function of hyperparameters, with the setting corresponding to Figure 6 .

		ansur2 (2)	bio (2)	births1 (2)	calcofi (2)	edm (2)	enb (2)	house (2)	taxi (2)	jura (3)	scpf (3)	sf1 (3)	sf2 (3)
ϵ	#target												
0.001	4096	3.3 ± 0.064	0.46 ± 0.057	78 ± 70	2.6 ± 0.089	1.9 ± 0.3	0.81 ± 0.21	2 ± 0.051	7 ± 0.12	13 ± 2.6	0.78 ± 0.4	14 ± 2.6	0.82 ± 0.32
	8192	3.4 ± 0.059	0.45 ± 0.057	78 ± 70	2.6 ± 0.089	1.9 ± 0.29	0.81 ± 0.2	2 ± 0.05	7 ± 0.13	11 ± 2.6	0.73 ± 0.23	16 ± 3.9	0.4 ± 0.16
	16384	3.4 ± 0.059	0.46 ± 0.058	78 ± 70	2.6 ± 0.093	1.8 ± 0.28	0.83 ± 0.21	2 ± 0.048	7 ± 0.13	12 ± 2.3	0.87 ± 0.34	21 ± 4.8	0.44 ± 0.2
	32768	3.4 ± 0.063	0.46 ± 0.058	78 ± 70	2.6 ± 0.092	1.9 ± 0.3	0.81 ± 0.2	2 ± 0.05	7 ± 0.13	12 ± 2.6	1.2 ± 0.47	16 ± 2.9	0.57 ± 0.18
0.01	4096	3.3 ± 0.055	0.55 ± 0.12	78 ± 70	2.5 ± 0.084	1.9 ± 0.3	0.81 ± 0.21	2 ± 0.05	7.5 ± 0.63	11 ± 2.8	0.43 ± 0.15	12 ± 2.1	0.2 ± 0.086
	8192	3.3 ± 0.054	0.56 ± 0.13	78 ± 70	2.5 ± 0.082	1.8 ± 0.3	0.8 ± 0.21	2 ± 0.049	7.5 ± 0.69	10 ± 2.6	0.37 ± 0.15	12 ± 2.8	0.17 ± 0.063
	16384	3.3 ± 0.045	0.56 ± 0.12	78 ± 70	2.5 ± 0.082	1.7 ± 0.24	0.8 ± 0.21	2 ± 0.05	7.5 ± 0.71	13 ± 4.3	0.4 ± 0.18	11 ± 2.9	0.19 ± 0.076
	32768	3.3 ± 0.064	0.56 ± 0.12	78 ± 70	2.5 ± 0.085	1.7 ± 0.26	0.82 ± 0.22	2 ± 0.049	7.5 ± 0.69	10 ± 2.7	0.41 ± 0.17	12 ± 2.6	0.18 ± 0.071
0.1	4096	3.3 ± 0.058	0.49 ± 0.011	78 ± 70	2.5 ± 0.084	1.6 ± 0.25	0.81 ± 0.21	2.3 ± 0.065	8.3 ± 1.4	9.2 ± 2.8	0.37 ± 0.15	6.6 ± 0.96	0.48 ± 0.1
	8192	3.3 ± 0.059	0.49 ± 0.011	78 ± 70	2.5 ± 0.084	1.6 ± 0.26	0.8 ± 0.21	2.3 ± 0.065	8.2 ± 1.5	9.4 ± 2.9	0.4 ± 0.15	6.1 ± 0.89	0.53 ± 0.11
	16384	3.3 ± 0.054	0.49 ± 0.012	78 ± 70	2.5 ± 0.081	1.6 ± 0.26	0.8 ± 0.21	2.3 ± 0.058	8.2 ± 1.4	9.4 ± 2.9	0.37 ± 0.12	6.4 ± 0.83	0.45 ± 0.092
	32768	3.3 ± 0.051	0.49 ± 0.011	77 ± 70	2.5 ± 0.083	1.5 ± 0.25	0.79 ± 0.2	2.3 ± 0.057	8.2 ± 1.4	8.9 ± 2.9	0.36 ± 0.12	6.5 ± 1.2	0.5 ± 0.1
1	4096	3.6 ± 0.055	0.65 ± 0.019	78 ± 70	2.5 ± 0.1	1.7 ± 0.27	0.92 ± 0.24	3 ± 0.13	6.4 ± 0.14	13 ± 4	0.45 ± 0.16	9.5 ± 1.9	0.84 ± 0.13
	8192	3.6 ± 0.067	0.59 ± 0.013	78 ± 70	2.5 ± 0.099	1.7 ± 0.26	0.91 ± 0.24	3 ± 0.14	6.3 ± 0.14	13 ± 4	0.42 ± 0.14	10 ± 1.8	0.93 ± 0.16
	16384	3.5 ± 0.072	0.57 ± 0.016	78 ± 70	2.5 ± 0.099	1.7 ± 0.27	0.91 ± 0.24	3 ± 0.13	6.4 ± 0.14	14 ± 4	0.48 ± 0.17	9.8 ± 1.7	0.91 ± 0.17
	32768	3.5 ± 0.061	0.6 ± 0.028	78 ± 71	2.5 ± 0.1	1.7 ± 0.27	0.91 ± 0.24	2.9 ± 0.13	6.4 ± 0.15	13 ± 4	0.47 ± 0.17	10 ± 1.7	0.9 ± 0.17

Table 2: Mean region size for varying ε and the number of target points in the ball.

		slump (3)	households (4)	air (6)	atp1d (6)	atp7d (6)
ϵ	#target					
0.001	4096	15 ± 7.6	37 ± 1.4	$2.6 \times 10^3 \pm 1.9 \times 10^3$	81 ± 19	$8.5 \times 10^2 \pm 4.5 \times 10^2$
	8192	7.9 ± 2	36 ± 1.9	$7.1 \times 10^2 \pm 56$	99 ± 41	$5.9 \times 10^2 \pm 1.8 \times 10^2$
	16384	11 ± 3.7	34 ± 1.3	$6.9 \times 10^2 \pm 52$	65 ± 19	$9.4 \times 10^2 \pm 3 \times 10^2$
	32768	12 ± 4.3	36 ± 2.6	$6.8 \times 10^2 \pm 36$	87 ± 28	$5.1 \times 10^2 \pm 2 \times 10^2$
0.01	4096	20 ± 6.8	37 ± 1.6	$8.5 \times 10^2 \pm 1 \times 10^2$	85 ± 24	$7.9 \times 10^2 \pm 4.1 \times 10^2$
	8192	12 ± 4.9	34 ± 1.7	$1.3 \times 10^3 \pm 7 \times 10^2$	82 ± 24	$4 \times 10^2 \pm 1.5 \times 10^2$
	16384	7.1 ± 2.2	33 ± 0.81	$5.5 \times 10^2 \pm 47$	$1.1 \times 10^2 \pm 26$	$3.7 \times 10^2 \pm 68$
	32768	10 ± 4	31 ± 0.97	$4.8 \times 10^2 \pm 51$	42 ± 9.1	$2.8 \times 10^2 \pm 98$
0.1	4096	5.8 ± 1.3	27 ± 1.3	$3.2 \times 10^2 \pm 32$	8.1 ± 1.7	33 ± 9.2
	8192	5.9 ± 1.3	26 ± 1.3	$3.1 \times 10^2 \pm 33$	5.7 ± 1	27 ± 6.9
	16384	5.9 ± 1.4	25 ± 1	$3.1 \times 10^2 \pm 34$	4 ± 1.4	26 ± 7.7
	32768	5.1 ± 1.1	25 ± 1	$3.1 \times 10^2 \pm 34$	3.8 ± 0.88	16 ± 5.1
1	4096	14 ± 5.3	29 ± 1.3	$4.3 \times 10^2 \pm 31$	6.2 ± 1.7	69 ± 25
	8192	15 ± 5.3	30 ± 2.1	$3.4 \times 10^2 \pm 38$	5.6 ± 2.2	69 ± 25
	16384	16 ± 5.6	28 ± 1.1	$4.1 \times 10^2 \pm 36$	6.1 ± 2	76 ± 27
	32768	15 ± 5.5	29 ± 1.9	$4.3 \times 10^2 \pm 38$	5.6 ± 1.5	73 ± 24

Table 3: Mean region size for varying ε and the number of target points in the ball.

		rf1 (8)	rf2 (8)	wq (14)	oes10 (16)	oes97 (16)	scm1d (16)	scm20d (16)
ϵ	#target							
0.001	4096	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$7.1 \times 10^9 \pm 3 \times 10^9$	$2.9 \times 10^8 \pm 8.3 \times 10^7$	$8.7 \times 10^8 \pm 4 \times 10^8$	$4 \times 10^7 \pm 3.6 \times 10^7$	$1.7 \times 10^7 \pm 1.1 \times 10^7$
	8192	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$3.7 \times 10^9 \pm 1.9 \times 10^9$	$3.7 \times 10^8 \pm 1.3 \times 10^8$	$1.4 \times 10^9 \pm 1.2 \times 10^9$	$9.3 \times 10^5 \pm 5 \times 10^5$	$2.5 \times 10^8 \pm 1.9 \times 10^8$
	16384	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$6.6 \times 10^9 \pm 3.2 \times 10^9$	$5.6 \times 10^8 \pm 4.3 \times 10^8$	$2.5 \times 10^8 \pm 1.3 \times 10^8$	$3.5 \times 10^5 \pm 1.3 \times 10^5$	$8.9 \times 10^7 \pm 5.7 \times 10^7$
	32768	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$3.1 \times 10^9 \pm 1.2 \times 10^9$	$5.5 \times 10^8 \pm 3 \times 10^8$	$3.1 \times 10^8 \pm 9.5 \times 10^7$	$9.7 \times 10^5 \pm 4.5 \times 10^5$	$1.3 \times 10^9 \pm 1.3 \times 10^9$
0.01	4096	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$1.1 \times 10^{10} \pm 7.3 \times 10^9$	$4.3 \times 10^9 \pm 3.8 \times 10^9$	$3.5 \times 10^9 \pm 2.5 \times 10^9$	$4.1 \times 10^8 \pm 3.8 \times 10^8$	$1.3 \times 10^{11} \pm 1.1 \times 10^{11}$
	8192	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$6.4 \times 10^{10} \pm 6 \times 10^{10}$	$3 \times 10^{10} \pm 2.8 \times 10^{10}$	$1 \times 10^{10} \pm 6.1 \times 10^9$	$8.1 \times 10^8 \pm 5.5 \times 10^8$	$1.1 \times 10^{11} \pm 1.1 \times 10^{11}$
	16384	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$3.3 \times 10^9 \pm 7.9 \times 10^8$	$1.1 \times 10^9 \pm 4.3 \times 10^8$	$1 \times 10^{10} \pm 5.7 \times 10^9$	$4.8 \times 10^7 \pm 3.7 \times 10^7$	$1.3 \times 10^9 \pm 8.3 \times 10^8$
	32768	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$5.1 \times 10^{11} \pm 4.9 \times 10^{11}$	$6.5 \times 10^9 \pm 5 \times 10^9$	$4 \times 10^9 \pm 3.2 \times 10^9$	$1.6 \times 10^7 \pm 9.5 \times 10^6$	$2.7 \times 10^8 \pm 1.3 \times 10^8$
0.1	4096	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$8.7 \times 10^9 \pm 3.7 \times 10^9$	$4.8 \times 10^4 \pm 3.2 \times 10^4$	$6 \times 10^9 \pm 6 \times 10^9$	$1.5 \times 10^3 \pm 6.7 \times 10^2$	$1.3 \times 10^6 \pm 6.4 \times 10^5$
	8192	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$4.8 \times 10^9 \pm 1.5 \times 10^9$	$1.7 \times 10^5 \pm 1.3 \times 10^5$	$6 \times 10^9 \pm 6 \times 10^9$	$6.2 \times 10^2 \pm 2.8 \times 10^2$	$1.2 \times 10^6 \pm 8.7 \times 10^5$
	16384	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$1.3 \times 10^{10} \pm 6.8 \times 10^9$	$5.2 \times 10^4 \pm 4.7 \times 10^4$	$5.6 \times 10^9 \pm 5.6 \times 10^9$	$2.2 \times 10^2 \pm 46$	$2.9 \times 10^5 \pm 1 \times 10^5$
	32768	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$7.4 \times 10^9 \pm 2.9 \times 10^9$	$7.6 \times 10^3 \pm 5.1 \times 10^3$	$9.2 \times 10^7 \pm 8.1 \times 10^7$	$1.1 \times 10^2 \pm 17$	$1.1 \times 10^5 \pm 3.1 \times 10^4$
1	4096	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$8 \times 10^8 \pm 2 \times 10^8$	$6.6 \times 10^2 \pm 3.4 \times 10^2$	$8.3 \times 10^5 \pm 8.1 \times 10^5$	$4.1 \times 10^2 \pm 76$	$5.2 \times 10^5 \pm 6.5 \times 10^4$
	8192	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$6.9 \times 10^8 \pm 1.7 \times 10^8$	$3.5 \times 10^2 \pm 1.8 \times 10^2$	$7.7 \times 10^5 \pm 7.6 \times 10^5$	$8.5 \times 10^2 \pm 3.1 \times 10^2$	$1.1 \times 10^6 \pm 3.9 \times 10^5$
	16384	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$5.3 \times 10^8 \pm 1.2 \times 10^8$	$2.2 \times 10^2 \pm 1.5 \times 10^2$	$4 \times 10^5 \pm 4 \times 10^5$	$1.3 \times 10^2 \pm 14$	$4.7 \times 10^5 \pm 1.8 \times 10^5$
	32768	$2 \times 10^{13} \pm 2 \times 10^{13}$	$2 \times 10^{13} \pm 2 \times 10^{13}$	$5.5 \times 10^8 \pm 1.5 \times 10^8$	$1.9 \times 10^2 \pm 1.6 \times 10^2$	$3.1 \times 10^5 \pm 3.1 \times 10^5$	$1 \times 10^2 \pm 11$	$3.4 \times 10^5 \pm 6.4 \times 10^4$

Table 4: Mean region size for varying ε and the number of target points in the ball.