

Conformal Graph Prediction with Z-Gromov-Wasserstein Distances

Gabriel Melo^{1,*}

Thibaut de Saivre^{1,*}

Anna Calissano²

Florence d’Alché-Buc¹

¹LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France

²Department of Statistical Science, University College London, London, UK

*Equal Contribution

Abstract

Supervised graph prediction addresses regression problems where the outputs are structured graphs. Although several approaches exist for graph-valued prediction, principled uncertainty quantification remains limited. We propose a conformal prediction framework for graph-valued outputs, providing distribution-free coverage guarantees in structured output spaces. Our method defines nonconformity via the Z-Gromov-Wasserstein distance, instantiated in practice through Fused Gromov-Wasserstein (FGW), enabling permutation invariant comparison between predicted and candidate graphs. To obtain adaptive prediction sets, we introduce Score Conformalized Quantile Regression (SCQR), an extension of Conformalized Quantile Regression (CQR) to handle complex output spaces such as graph-valued outputs. We evaluate the proposed approach on a synthetic task and a real problem of molecule identification.

1 INTRODUCTION

Motivated by various applications such as molecular identification [Nguyen et al., 2019] in chemistry or scene understanding in computer vision [Shit et al., 2022], Supervised Graph Prediction (SGP) has recently attracted a growing interest in Machine Learning. This task consists in learning a predictive model that maps an input variable of any modality (text, image, tabular or distributional) to a target graph of arbitrary size. Approaches in the literature leverage surrogate regression in graph representation spaces [Brouard et al., 2016] or end-to-end learning [Shit et al., 2022, Yang et al., 2024, Krzakala et al., 2024] that boils down to graph-valued regression [Calissano et al., 2022]. However, none of these provides confidence sets, posing risks for discovery or identification when experimental validation is costly.

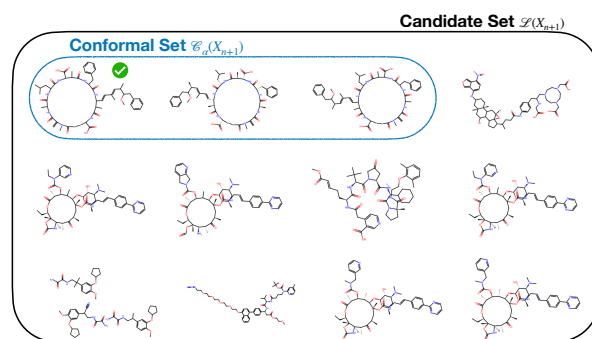


Figure 1: Conformal metabolite prediction set at 90% marginal coverage for molecule identification. The green check-mark denotes the ground-truth.

Uncertainty quantification for graph-valued data aims to produce a set of plausible graphs rather than a single prediction. Although random graph models are well studied [Frieze and Karoński, 2015], parametric approaches are often unrealistic in practice, especially for attributed graphs. Non-parametric methods therefore provide a more flexible alternative.

In this paper, we adopt Conformal Prediction (CP), an agnostic, distribution free, and post-training framework¹. CP [Shafer and Vovk, 2008, Fontana et al., 2023] provides finite-sample coverage guarantees under minimal assumptions, requiring only exchangeability of the data, and can be applied on top of any pre-trained predictor without modifying training. Extending CP to graph-valued outputs raises several fundamental challenges. Even in the multivariate Euclidean setting, constructing informative conformal sets is known to be difficult due to the curse of dimensionality and the lack of natural orderings [Dheur et al., 2025, Kondratyev et al., 2025, Thurin et al., 2025]. These challenges are amplified in the graph setting, where outputs live in a highly structured, non-Euclidean, and combinatorial space.

¹The code is available at <https://github.com/gabrielmelo00/GraphConf>.

For a graph with attributes on nodes and edges, uncertainty can revolve around the uncertainty at the global level, providing a conformal set made of a set of graphs with more or less different structure, or at a local level, measuring uncertainty at the level of the nodes and edges attribute. In this paper, we opt for a global framework. This choice is also supported by a specific focus on molecular identification tasks, where only a specific amount of configurations are possible.

A central challenge is therefore the choice of a suitable non-conformity score, which determines the structure and informativeness of the resulting conformal sets. Since graphs are defined up to node permutation, one must either work in the quotient space or use a permutation-invariant discrepancy. We adopt Optimal Transport distances, namely Gromov–Wasserstein and its variants, to define permutation-invariant scores and extend Conformal Prediction to graph-valued outputs with validity in the quotiented graph space.

We focus on Supervised Graph Prediction settings where each input is associated with a finite, input-dependent candidate set of graphs, as commonly assumed in structured output prediction and molecular identification.

Standard conformal prediction relies on a single global threshold, implicitly assuming homogeneous uncertainty across inputs. To account for input-dependent variability, we introduce *Score Conformalized Quantile Regression* (SCQR), which calibrates conditional quantiles of the non-conformity score instead of a global cutoff. SCQR yields locally adaptive conformal sets while preserving marginal coverage guarantees.

Finally we empirically test our novel framework on a synthetic image-to-graph task and metabolite identification from mass spectra, assessed on Spectraverse, a recent benchmark in metabolomics [Gupta et al., 2026].

In short, our contributions include:

- a framework for Conformal Graph Prediction based on Z -Gromov-Wasserstein non-conformity scores and the proof of its validity in quotiented graph spaces;
- a locally adaptive variant, Score Conformalized Quantile Regression (SCQR), with proven marginal coverage;
- a set of numerical experiments showing the effectiveness and the versatility of the framework on a synthetic image-to-graph task and real molecule prediction problem.

2 RELATED WORK

Prediction Models for Graphs. We consider supervised models whose outputs are graphs, including barycenter-based methods [Brogat-Motte et al., 2022, Yang et al., 2024],

graph regression [Calissano et al., 2022], deep end-to-end predictors [Krzakala et al., 2024, Shit et al., 2022], and graph-level autoencoders [Krzakala et al., 2025]. In molecular applications, graphs may also be predicted indirectly via SMILES representations [Zhang et al., 2025]. Our framework is model-agnostic and applies to both direct graph predictors and SMILES-based pipelines. Unlike works that use graphs as input structures for node-, edge-, or graph-level prediction, our focus is on prediction problems whose outputs are themselves graphs.

Conformal Prediction for Graphs. Recent efforts have extended CP to non-Euclidean domains. For node-level tasks, Huang et al. [2023] introduced conformalized GNNs, establishing the importance of permutation invariance in base predictors, while [Lunde et al., 2025] defined a node based CP sets. Related approaches construct conformal prediction sets for graph neural networks, including node classification and link prediction on fixed input graphs [Zargarbashi et al., 2023, H Zargarbashi and Bojchevski, 2024]. Zhang et al. [2024] proposed a general framework for conformal structured prediction, constructing structured prediction sets via implicit representations, e.g., directed acyclic graphs for hierarchical labels. CP has also been studied for knowledge graph (KG) embeddings, turning its raw plausibility scores into entity or predicate answer sets that provably cover the true answer at a user-specified confidence level [Zhu et al., 2025b,a,c].

Closer to our setting, Calissano et al. [2024] developed CP sets for populations of unlabeled graphs, utilizing quotient spaces to handle the lack of node correspondence. Our work extends these concepts to graphs with categorical *attributes* on nodes. In addition, by adopting the Z -Gromov-Wasserstein distance, we work in a permutation invariant setting. Thus, whereas prior graph-related CP methods mainly use graphs as inputs or define sets over discrete entity/predicate completions of a KG, our setting requires calibrated sets in an unordered, attributed graph-valued output space.

3 BACKGROUND

We begin by reviewing the two central components of our framework: conformal prediction and the Z -Gromov-Wasserstein (Z -GW) distance, a metric on metric measure spaces, under which graphs naturally appear as an example of discrete metric measure spaces.

Notation Let $\Sigma_n := \{a \in \mathbb{R}_+^n : \sum_{i=1}^n a_i = 1\}$ denote the probability simplex, and let $\sigma_n := \{P \in \{0, 1\}^{n \times n} : P\mathbf{1}_n = \mathbf{1}_n, P^\top \mathbf{1}_n = \mathbf{1}_n\}$ be the set of permutation matrices, where $\mathbf{1}_n \in \mathbb{R}^n$ denotes the all-ones vector. For finite sets $\mathcal{X} = \{x_1, \dots, x_n\}$ and $\mathcal{Y} = \{y_1, \dots, y_m\}$, any discrete probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ can

be written as $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$ with $a \in \Sigma_n$ and $b \in \Sigma_m$. We identify admissible couplings $\pi \in \Pi(\mu, \nu)$ with nonnegative matrices $\pi \in \mathbb{R}_+^{n \times m}$ satisfying $\pi \mathbf{1}_m = a$, $\pi^\top \mathbf{1}_n = b$. When both measures are uniform, i.e. $\mu(\{x_i\}) = \nu(\{y_j\}) = \frac{1}{n}$ for all i, j and $n = m$, the set $\Pi(\mu, \mu)$ coincides with the Birkhoff polytope.

3.1 CONFORMAL PREDICTION

Definition 3.1 (Exchangeability). *A sequence of random variables $(Z_1, \dots, Z_{n+1})$ taking values in a measurable space $\mathcal{Z}$ is said to be exchangeable if, for any permutation σ of $\{1, \dots, n+1\}$ and for any $(z_1, \dots, z_{n+1}) \in \mathcal{Z}^{n+1}$, $\mathbb{P}(Z_1 = z_1, \dots, Z_{n+1} = z_{n+1}) = \mathbb{P}(Z_{\sigma(1)} = z_1, \dots, Z_{\sigma(n+1)} = z_{n+1})$.*

Conformal Prediction (CP) provides finite-sample, distribution-free guarantees on the coverage of prediction sets. In the regression setting, we observe exchangeable pairs (X_i, Y_i) and have a pre-trained base predictor $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$. The framework relies on a *nonconformity score* function $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ that measures the discrepancy between the target y and the prediction $f_\theta(x)$. For standard real-valued regression, a typical choice is the absolute residual:

$$s(x, y) = |y - f_\theta(x)|. \quad (1)$$

Given a held-out calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$, we compute the scores $R = \{R_1, \dots, R_n\}$ where $R_i = s(X_i, Y_i)$. We then compute the adjusted empirical quantile:

$$\hat{q}_{1-\alpha} = \text{Quantile} \left(\{R_i\}_{i=1}^n, \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \right). \quad (2)$$

For a new input X_{n+1} , the conformal prediction set is defined as:

$$\mathcal{C}(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{q}_{1-\alpha}\}. \quad (3)$$

By exchangeability of the calibration and test samples, the nonconformity scores $\{R_i\}_{i=1}^{n+1}$ with $R_i = s(X_i, Y_i)$ form an exchangeable sequence. Consequently, the rank of R_{n+1} among $\{R_i\}_{i=1}^{n+1}$ is uniformly distributed, which implies the marginal coverage guarantee [Vovk et al., 2005]

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha,$$

without any assumptions on the data distribution or the accuracy of the predictor. We refer to Angelopoulos et al. [2023], Shafer and Vovk [2008] for more details on conformal prediction.

Conformalized Quantile Regression To extend standard conformal prediction to better handle heteroscedasticity, Conformalized Quantile Regression (CQR) [Romano et al.,

2019] improves the conformal framework by leveraging quantile regression [Koenker and Bassett, 1978] to construct input-adaptive prediction intervals. Instead of relying on a fixed, globally calibrated residual, CQR employs two base models, $\hat{\psi}_{\alpha/2}(x)$ and $\hat{\psi}_{1-\alpha/2}(x)$, trained to estimate the lower and upper $(\alpha/2, 1 - \alpha/2)$ quantiles of the conditional distribution $Y|X = x$. The nonconformity score is then defined as $s(x, y) = \max\{\hat{\psi}_{\alpha/2}(x) - y, y - \hat{\psi}_{1-\alpha/2}(x)\}$, which measures the signed distance of the target to the boundaries of the predicted interval. By computing the $(1 - \alpha)$ -th quantile of these scores, $\hat{q}_{1-\alpha}$, on a calibration set, the resulting prediction intervals $\mathcal{C}(x) = [\hat{\psi}_{\alpha/2}(x) - \hat{q}_{1-\alpha}, \hat{\psi}_{1-\alpha/2}(x) + \hat{q}_{1-\alpha}]$ achieve valid marginal coverage while dynamically adjusting their width to account for local uncertainty in the data.

3.2 Z-GROMOV WASSERSTEIN DISTANCE

To measure discrepancies between structured outputs such as graphs, which will later serve as the basis for conformal scoring, we work in the space of Z -networks equipped with the Z -Gromov-Wasserstein distance.

Definition 3.2 (Metric Measure Spaces). *A metric measure space (mm-space) is defined as triple $(\mathcal{X}, d_{\mathcal{X}}, \mu_{\mathcal{X}})$, where $(\mathcal{X}, d_{\mathcal{X}})$ is a compact metric space and $\mu_{\mathcal{X}}$ is a Borel probability measure on $\mathcal{X}$ with $\mu_{\mathcal{X}}(\mathcal{X}) = 1$ and full support, $\text{supp}(\mu_{\mathcal{X}}) = \mathcal{X}$.*

Gromov-Wasserstein (GW) distances were introduced to compare mm-spaces via optimal transport [Mémoli, 2011]. Subsequent work in the machine learning literature proposed several variants to handle structured and attributed data, most notably Fused GW (FGW) [Vayer et al., 2019]. The Z -Gromov-Wasserstein (Z -GW) distance [Bauer et al., 2025] unifies these approaches by replacing the metric-valued distance function in classical GW with a general measurable pairwise relation taking values in a metric space $(\mathcal{Z}, d_{\mathcal{Z}})$, so that classical GW, FGW and other formulations arise as particular instances of this construction.

Definition 3.3 (Z -networks). *Let $(\mathcal{Z}, d_{\mathcal{Z}})$ be a metric space. A Z -network is a triple $(\mathcal{X}, \omega_{\mathcal{X}}, \mu_{\mathcal{X}})$, where $\mathcal{X}$ is a measurable space, $\mu_{\mathcal{X}}$ is a probability measure on $\mathcal{X}$, and $\omega_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{Z}$ is a measurable function encoding pairwise relational information. Unlike mm-spaces, $\omega_{\mathcal{X}}$ is not required to be a metric.*

Z-Gromov-Wasserstein distance. Given two Z -networks $(\mathcal{X}, \omega_{\mathcal{X}}, \mu_{\mathcal{X}})$ and $(\mathcal{Y}, \omega_{\mathcal{Y}}, \mu_{\mathcal{Y}})$, the Z -Gromov-Wasserstein p -distance ($p \geq 1$) is defined as

$$\text{GW}_p^Z(\mathcal{X}, \mathcal{Y}) = \frac{1}{2} \inf_{\pi \in \Pi(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}})} \left(\iint_{(\mathcal{X} \times \mathcal{Y})^2} d_{\mathcal{Z}}(\omega_{\mathcal{X}}(x, x'), \omega_{\mathcal{Y}}(y, y'))^p d\pi(x, y) d\pi(x', y') \right)^{1/p}. \quad (4)$$

When $\mathcal{Z} = \mathbb{R}$ and $\omega_{\mathcal{X}}$ is the usual metric distance ($\omega_{\mathcal{X}} = d_{\mathcal{X}}$), $\text{GW}_p^{\mathcal{Z}}$ reduces to the classical GW distance.

Definition 3.4 (Weak isomorphism [Bauer et al., 2025]). Two $\mathcal{Z}$ -networks $(\mathcal{X}, \omega_{\mathcal{X}}, \mu_{\mathcal{X}})$ and $(\mathcal{Y}, \omega_{\mathcal{Y}}, \mu_{\mathcal{Y}})$ are said to be weakly isomorphic, written $\mathcal{X} \sim \mathcal{Y}$, if there exists a $\mathcal{Z}$ -network $(\mathcal{W}, \omega_{\mathcal{W}}, \mu_{\mathcal{W}})$ and measure-preserving maps $\phi_{\mathcal{X}} : \mathcal{W} \rightarrow \mathcal{X}$ and $\phi_{\mathcal{Y}} : \mathcal{W} \rightarrow \mathcal{Y}$ such that

$$\omega_{\mathcal{W}}(w, w') = \omega_{\mathcal{X}}(\phi_{\mathcal{X}}(w), \phi_{\mathcal{X}}(w')) = \omega_{\mathcal{Y}}(\phi_{\mathcal{Y}}(w), \phi_{\mathcal{Y}}(w'))$$

for $\mu_{\mathcal{W}} \otimes \mu_{\mathcal{W}}$ for almost every pair (w, w') .

Weak isomorphism and metricity. The distance $\text{GW}_p^{\mathcal{Z}}$ defines a pseudometric on the space of $\mathcal{Z}$ -networks. To obtain a genuine metric, one considers equivalence classes under *weak isomorphism*. We denote by $\mathfrak{M}$ the collection of all $\mathcal{Z}$ -networks $(\mathcal{X}, \omega_{\mathcal{X}}, \mu_{\mathcal{X}})$, and by $\mathcal{M} := \mathfrak{M}/\sim$ the corresponding quotient space under weak isomorphism.

Theorem 3.5 ([Bauer et al., 2025, Thm. 29]). For any separable metric space $(\mathcal{Z}, d_{\mathcal{Z}})$ and $p \geq 1$, the distance $\text{GW}_p^{\mathcal{Z}}$ induces a genuine metric on $\mathcal{M}$.

Remark 3.6 (Working with representatives in practice). Although the quotient space $\mathcal{M} := \mathfrak{M}/\sim$ is the natural mathematical domain on which $\text{GW}_p^{\mathcal{Z}}$ is well defined, it is not directly accessible in practice. Each element of $\mathcal{M}$ corresponds to an equivalence class of isomorphic realizations, and any concrete representative in $\mathfrak{M}$ is inherently arbitrary. In practical settings, observed data are therefore provided as explicit $\mathcal{Z}$ -networks in $\mathfrak{M}$ (e.g., graphs, attributed graphs, or meshes), and learning algorithms operate on these representations. Consequently, both empirical datasets and prediction tasks are formulated at the level of $\mathfrak{M}$, while permutation-invariant losses such as $\text{GW}_p^{\mathcal{Z}}$ ensure consistency with the underlying quotient structure.

4 CONFORMAL GRAPH PREDICTION

4.1 PROBLEM SETUP

Graphs as $\mathcal{Z}$ -networks. Graphs can be naturally viewed as finite $\mathcal{Z}$ -networks. Let $G = (V, E, F)$ be a graph, where V is the node set with $|V| = n \in \mathbb{N}^*$, $E \subseteq V \times V$ is the edge set, and $\Psi : V \rightarrow \mathbb{R}^d$ is a node feature map, with $\Psi(i)$ denoting the feature associated to node $i \in V$. We associate to G a $\mathcal{Z}$ -network (V, ω_G, μ_G) , where $\mu_G \in \mathcal{P}(V)$ is a probability measure on V (assumed uniform in this case), and $\omega_G : V \times V \rightarrow \mathcal{Z}$ encodes pairwise relational information between nodes. This representation provides a unified way to encode both graph structure and node attributes, enabling the comparison of graphs within the general $\mathcal{Z}$ -GW distance.

Example 4.1 (Fused Network Gromov-Wasserstein Distance). Choosing the pairwise structure ω_G such that it

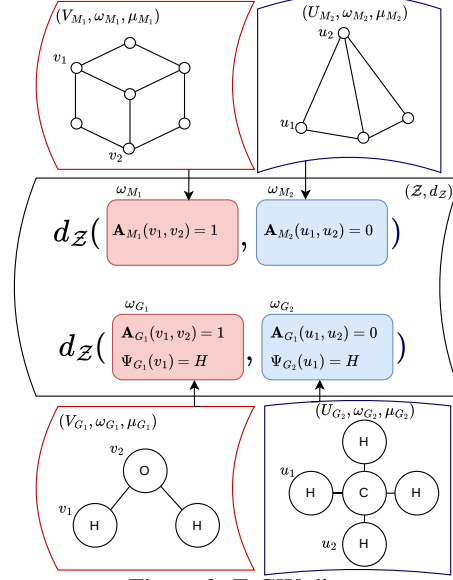


Figure 2: $\mathcal{Z}$ -GW distance

combines structural and node-level information allows to retrieve the so-called Fused Network Gromov-Wasserstein Distance introduced by Yang et al. [2024]:

$$\omega_G(i, k) = (\mathbf{A}(i, k), \mathbf{X}(i, k), \Psi(i)) \in \mathcal{Z} := \Omega \times \mathbb{R}^m \times \mathbb{R}^d,$$

where the matrix $\mathbf{A} : V \times V \rightarrow \Omega$ defined from the set of edges E encodes graph connectivity (e.g. Adjacency, Shortest Path, Laplacian, etc), $\mathbf{X} : V \times V \rightarrow \mathbb{R}^m$ represents edge features, and $\Psi(i) \in \mathbb{R}^d$ denotes the feature associated to node i . For instance, when $\mathbf{A}$ is the adjacency matrix, $\Omega = \{0, 1\}$. The product space $\mathcal{Z}$ is equipped with a weighted ℓ_q metric

$$d_{\mathcal{Z}}((a, x, \psi), (a', x', \psi')) = \left(\beta d_{\Omega}(a, a')^q + \gamma d_{\mathbb{R}^m}(x, x')^q + (1 - \gamma - \beta) d_{\mathbb{R}^d}(\psi, \psi')^q \right)^{1/q}, \quad (5)$$

for $\gamma, \beta \in [0, 1]$, $q \geq 1$.

Example 4.2 (Fused Gromov-Wasserstein Distance [Vayer et al., 2020]). Setting $\gamma = 0$ and $\beta \in [0, 1]$, we recover the Fused Gromov-Wasserstein (no more edge features).

Example 4.3 (Gromov-Wasserstein Distance [Mémoli, 2011]). To retrieve Gromov-Wasserstein distance between unlabeled graphs (no more feature nodes), we set $\gamma = 0$ and $\beta = 1$.

Discrete $\mathcal{Z}$ -Gromov-Wasserstein for graphs. Let $G_1 = (V_1, \omega_1, \mu_1)$ and $G_2 = (V_2, \omega_2, \mu_2)$ be finite graphs with $|V_1| = n$ and $|V_2| = m$ and uniform measures. The discrete $\mathcal{Z}$ -GW distance between G_1 and G_2 is defined as

$$\min_{\pi \in \Pi(\mu_1, \mu_2)} \left(\sum_{i, k \in V_1} \sum_{j, l \in V_2} d_{\mathcal{Z}}(\omega_1(i, k), \omega_2(j, l))^p \pi_{ij} \pi_{kl} \right)^{1/p}. \quad (6)$$

In the finite graph setting with uniform measures, weak isomorphisms admit a simple characterization (see Proposition 4.4 below). As a consequence, GW_p^Z defines a permutation-invariant distance on graphs.

Proposition 4.4 (Weak isomorphism and permutation invariance). *Let (V_1, ω_1, μ_1) and (V_2, ω_2, μ_2) be finite Z -networks associated to G_1 and G_2 , respectively, with $|V_1| = |V_2| = n$ and uniform measures $\mu_1 = \mu_2$. Then $G_1 \sim_n G_2$ if and only if there exists a permutation matrix $P \in \sigma_n$ such that $\omega_1 = P^\top \omega_2 P$.*

From now on, we specialize the spaces $\mathfrak{M}$ and $\mathcal{M}$ to graphs. We denote by $\mathfrak{M}$ the collection of all graphs with at most N nodes, i.e. $\mathfrak{M} = \bigsqcup_{n=1}^N \mathcal{G}_n$, in which $\mathcal{G}_n$ are the space of all graphs with n nodes represented in a labeled form, and by $\mathcal{M} := \bigsqcup_{n=1}^N (\mathcal{G}_n / \sim_n)$, the union of the quotient graph spaces of size n , where graphs are identified up to node permutation. For simplicity, we note $\mathcal{M} = \mathfrak{M} / \sim$.

Graph-valued prediction. We consider a graph-valued prediction model

$$f_\theta : \mathcal{X} \rightarrow \mathfrak{M}$$

that maps inputs $x \in \mathcal{X}$ (e.g., spectra, images, or other descriptors) to graph-valued outputs $y \in \mathfrak{M}$. In the remainder, we assume that f_θ has been trained on a separate dataset $\mathcal{D}_{\text{train}}$ by minimizing a loss measuring discrepancy between predicted and ground-truth graphs. This training set is independent of the calibration data $\mathcal{D}_{\text{val}} = \{(X_i, Y_i)\}_{i=1}^n$. Crucially, such losses must be invariant to node relabeling; the GW_p^Z distance, for instance, satisfies this requirement by construction. In the experimental part, we consider three off-the-shelves graph prediction models f_θ : ANY2GRAPH[Krzakala et al., 2024], a general-purpose end-to-end graph predictor restricted to small-size graphs, and EMBCOS[De Waele et al., 2026] and MSALIGN[Krzakala et al., 2026], models specialized on molecule identification from mass spectra.

4.2 GENERAL FRAMEWORK FOR CONFORMAL GRAPH PREDICTION

Given the previous set up, for a given input x , we get a concrete graph $\hat{y} = f_\theta(x) \in \mathfrak{M}$. Uncertainty guarantees, however, must be invariant to node relabeling and are therefore stated on the quotient space $\mathcal{M}$.

Assumption 4.5 (Exchangeable data). *The observed data $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable and take values in $\mathcal{X} \times \mathfrak{M}$.*

Definition 4.6 (Canonical projection). *The canonical projection $h : \mathfrak{M} \rightarrow \mathcal{M} := \mathfrak{M} / \sim$ maps a graph $y \in \mathfrak{M}$ to its equivalence class*

$$h(y) := [y] = \{y' \in \mathfrak{M} \mid y' \sim y\},$$

corresponding to all node relabelings of the same underlying graph.

Given a graph predictor f_θ and representative-valued data in $\mathfrak{M}$, we define the nonconformity score

$$s : \mathcal{X} \times \mathfrak{M} \rightarrow \mathbb{R}_+, \quad s(x, y) := \text{GW}_p^Z(f_\theta(x), y).$$

Since GW_p^Z is invariant under node permutation, the score s is invariant under the equivalence relation $\sim$. As a consequence, s factors through the canonical projection h and induces a well-defined score

$$\tilde{s} : \mathcal{X} \times \mathcal{M} \rightarrow \mathbb{R}_+, \quad \tilde{s}(x, \tilde{y}) = s(x, y), \quad \tilde{y} = h(y),$$

which depends only $\tilde{y} \in \mathcal{M}$.

Lemma 4.7 (Exchangeability is preserved under quotient maps). *Let $(X_i, Y_i)_{i=1}^n$ be an exchangeable sequence of random variables taking values in $\mathcal{X} \times \mathfrak{M}$. Let $h : \mathfrak{M} \rightarrow \mathcal{M} := \mathfrak{M} / \sim$ be the canonical projection onto equivalence classes, and define $\tilde{Y}_i := h(Y_i)$. Then the induced sequence $(X_i, \tilde{Y}_i)_{i=1}^n$ taking values in $\mathcal{X} \times \mathcal{M}$ is exchangeable.*

Lemma 4.7 ensures that the exchangeability of the observed data is preserved when passing from representative-valued outputs in $\mathfrak{M}$ to their equivalence classes in the quotient space $\mathcal{M}$.

We now establish conformal validity for Z -GW-based conformal sets. Let $R_i = s(X_i, Y_i)$ be the calibration scores and let $\hat{q}_{1-\alpha}$ denote their empirical $(1 - \alpha)$ -quantile. For a new input X_{n+1} , define

$$\mathcal{C}_\alpha^{\mathfrak{M}}(X_{n+1}) = \left\{ y \in \mathfrak{M} : \text{GW}_p^Z(f_\theta(X_{n+1}), y) \leq \hat{q}_{1-\alpha} \right\}. \quad (7)$$

Proposition 4.8 (Conformal validity on $\mathfrak{M}$ and $\mathcal{M}$). *Let $(X_i, Y_i)_{i=1}^{n+1}$ be exchangeable random variables taking values in $\mathcal{X} \times \mathfrak{M}$, and let $f_\theta : \mathcal{X} \rightarrow \mathfrak{M}$ be a predictor. Define the nonconformity score*

$$s(x, y) := \text{GW}_p^Z(f_\theta(x), y),$$

and construct the conformal prediction set $\mathcal{C}_\alpha^{\mathfrak{M}}(x) \subseteq \mathfrak{M}$ at level $\alpha \in (0, 1)$ using s . Then the following hold:

1. (Marginal coverage)

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_\alpha^{\mathfrak{M}}(X_{n+1})) \geq 1 - \alpha.$$

2. (Well-defined on Quotient) $\mathcal{C}_\alpha^{\mathfrak{M}}(x)$ is a union of equivalence classes and therefore induces a well-defined prediction set on the quotient space, $\mathcal{C}_\alpha^{\mathcal{M}}(x) := h(\mathcal{C}_\alpha^{\mathfrak{M}}(x)) \subseteq \mathcal{M}$, which satisfies

$$\mathbb{P}(h(Y_{n+1}) \in \mathcal{C}_\alpha^{\mathcal{M}}(X_{n+1})) \geq 1 - \alpha.$$

All proofs are deferred to Appendix A.

4.3 PRACTICAL RESTRICTIONS

The conformal set defined in Eq. 7 is an implicit subset of the graph space $\mathfrak{M}$, specified through a membership predicate rather than explicit enumeration. For any candidate graph $y \in \mathfrak{M}$, membership is determined by evaluating the non-conformity score $s(x, y)$ and comparing it to the calibrated threshold. In many graph domains such as molecular data, however, $\mathfrak{M}$ is combinatorially large, making the explicit materialization of $\mathcal{C}(x)$ computationally infeasible.

To obtain a finite and tractable prediction set, we intersect the implicit conformal set with an input-dependent candidate library $\mathcal{L}(x) \subset \mathfrak{M}$ (e.g., a metabolite database determined by a mass spectrum):

$$\mathcal{C}_{\mathcal{L}}(x) = \mathcal{C}(x) \cap \mathcal{L}(x) = \{y \in \mathcal{L}(x) : s(x, y) \leq \hat{q}_{1-\alpha}\}. \quad (8)$$

In retrieval tasks, the candidate library is typically constructed to be complete, in the sense that the ground-truth output is contained almost surely, i.e., $\mathbb{P}(Y_{n+1} \in \mathcal{L}(X_{n+1})) = 1$. Under this assumption, $\mathcal{C}_{\mathcal{L}}(x)$ coincides with the exact conformal prediction set restricted to the reduced output space $\mathcal{L}(x)$ and therefore inherits the marginal coverage guarantee.

Remark 4.9. *If the ground-truth output Y_{n+1} may lie outside $\mathcal{L}(X)$, the coverage degrades according to the probability of library incompleteness: $\mathbb{P}(Y_{n+1} \in \mathcal{C}_{\mathcal{L}}(X_{n+1})) \geq (1 - \alpha) - \mathbb{P}(Y_{n+1} \notin \mathcal{L}(X_{n+1}))$.*

5 LOCALLY ADAPTIVE GRAPH CONFORMAL PREDICTION

Previously, we extended conformal prediction to graph-valued outputs using a single global threshold on a nonconformity score. However, a global cutoff implicitly assumes homogeneous uncertainty across inputs. In practice, this assumption is rarely satisfied: some instances are intrinsically easy, while others are highly ambiguous. A single threshold therefore tends to produce overly conservative sets for easy inputs and risks undercoverage for difficult ones.

In complex, high-dimensional, or structured output spaces (e.g. graphs, manifolds, or functional data), applying standard CQR directly in the output space $\mathfrak{M}$ is often computationally or conceptually intractable, since conditional quantiles are not naturally defined for structured objects. To provide distribution-free *locally adaptive* guarantees in these settings, we introduce *Score Conformalized Quantile Regression* (SCQR).

Score Conformalized Quantile Regression Given a point predictor $f_{\theta} : \mathcal{X} \rightarrow \mathfrak{M}$ and a nonconformity score $s(x, y) \in \mathbb{R}$, with $x \in \mathcal{X}$ and $y \in \mathfrak{M}$. SCQR relaxes the Graph conformal prediction one-size-fits-all assumption by

allowing the threshold to depend on input-dependent attributes $\omega(x) \in \Omega$, which capture heteroscedasticity. These attributes may be the input itself (identity map) or derived quantities reflecting its complexity, e.g. size of candidate set or some model embeddings.

Concretely, we first train a quantile regression model $\psi : \Omega \rightarrow \mathbb{R}$ using the pinball loss to estimate $(1 - \alpha)$ -conditional quantile of $s(x, y)$ on a training set. We again compute the adaptive residuals $E_i = s(X_i, Y_i) - \psi(\omega(X_i))$. We then compute the empirical quantile of the residuals: $\hat{q}_{1-\alpha} = \text{Quantile}\left(\{E_i\}_{i=1}^n, \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right)$. For a new input X_{n+1} , the conformal prediction set is:

$$\mathcal{C}(X_{n+1}) = \{y \in \mathfrak{M} : s(X_{n+1}, y) \leq \psi(\omega(X_{n+1})) + \hat{q}_{1-\alpha}\}. \quad (9)$$

The residuals E_i may take both negative and positive values. Since $s(x, y) \geq 0$ and $s(x, y) = 0$ indicates perfect conformity, negative residuals correspond to examples that are easier than predicted given $\omega(x)$, while positive residuals indicate an underestimation of difficulty. Retaining both is essential for adaptive and non-conservative conformal calibration.

Proposition 5.1 (Marginal Coverage Guarantee of SCQR). *Let $\{(X_i, Y_i)\}_{i=1}^{n+1}$ be exchangeable random variables, and let $\omega : \mathcal{X} \rightarrow \Omega$ be a deterministic feature map. Assume that the nonconformity score $s : \mathcal{X} \times \mathfrak{M} \rightarrow \mathbb{R}$ and the quantile regression function $\psi : \Omega \rightarrow \mathbb{R}$ are fixed measurable functions, independent of the calibration sample. Then the SCQR prediction set expressed in Eq. 9 satisfies*

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha. \quad (10)$$

Remark 5.2 (One-Sided Nature of SCQR). *Unlike the original CQR for real-valued regression which often produces two-sided intervals $[q_{\text{low}}, q_{\text{high}}]$, SCQR is inherently one-sided. This is because nonconformity scores are typically designed to be non-negative, where smaller values represent better fits.*

6 NUMERICAL EXPERIMENTS

To illustrate the general applicability of our approach, we consider both a synthetic graph prediction benchmark and a real-world metabolite retrieval task. We refer to Appendix C for more details on implementations and extra results.

6.1 SYNTHETIC DATASET

We use the synthetic Coloring dataset [Krzakala et al., 2024], consisting of pairs (X_i, Y_i) where X_i is an image encoding a graph-coloring instance and Y_i is its ground-truth graph. Nodes take one of four discrete colors (blue, green, yellow, red) and edges indicate spatial neighboring. The task is to recover the underlying graph from the image, providing a controlled benchmark for image-to-graph prediction.

Candidate sets. For each test instance (X_i, Y_i) , we construct the candidate set $\mathcal{L}(x)$ by selecting all graphs in the test split that share the same node-type configuration as Y_i , i.e., the same number of blue, green, yellow, and red nodes. This yields a controlled identification problem where structurally distinct graphs share identical node statistics.

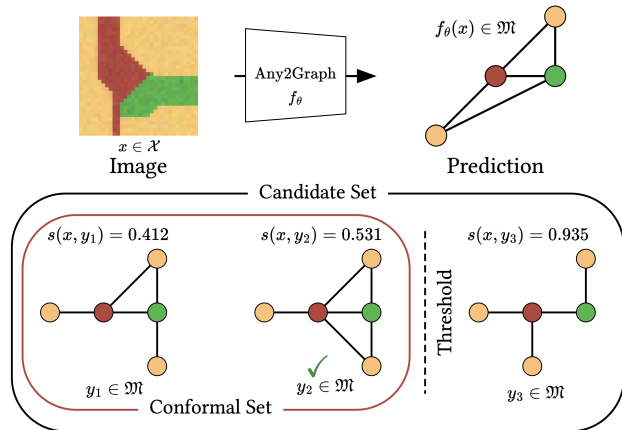


Figure 3: Example of the retrieval on the Coloring dataset. The green check-mark denotes the ground-truth.

Figure 3 illustrates the conformal prediction mechanism on a single Coloring instance. ANY2GRAPH predicts a graph that does not exactly match the ground truth, yet the conformal set, constructed by thresholding FGW distances to all candidates, correctly contains it. The threshold, calibrated on held-out data, automatically adjusts to the predictor’s error level. Candidates with FGW distance above the threshold are excluded, while the ground truth, being structurally close enough, remains inside the conformal set.

Model and protocol. We use ANY2GRAPH [Krzakala et al., 2024] as graph predictor f_θ (82% test accuracy), trained with the PMFGW loss. We use 100k samples for training and 10k each for calibration and testing.

6.2 MOLECULE RETRIEVAL

In many scientific applications, researchers aim to identify which molecules (also referred as metabolites) are present in a biological or chemical sample. A standard experimental tool for this task is mass spectrometry, which measures ionized molecules according to their mass-to-charge ratio. In tandem mass spectrometry (MS/MS), a molecule is further fragmented, producing a spectrum whose peaks provide information about its underlying structure. The computational goal is then to retrieve, from a set of candidate molecules, the molecular structure that most likely generated the observed spectrum.

We evaluate our framework on MS/MS-based molecule identification using the benchmark *Spectraverse* [Gupta et al., 2026]. Each input $X_i \in \mathcal{X}$ is an MS/MS spectrum and

$Y_i \in \mathfrak{M}$ is the corresponding molecular graph. *Spectraverse* contains spectra acquired under different *adduct types*, corresponding to different ionized forms of a molecule, for example after gaining or losing a small charged group. Since adducts can change the observed spectrum, each adduct type defines a distinct dataset for evaluating our method. For this paper, each adduct type can simply be understood as a separate dataset of paired mass spectra and molecules, together with the associated candidate molecules used for retrieval.

From spectra to graphs. We consider two existing methods for mapping spectra to molecular structures. EMB-COS [De Waele et al., 2026] learns a shared embedding space for spectra and molecules from scratch, while MSALIGN [Krzakala et al., 2026] leverages frozen pre-trained encoders DREAMS [Bushuiev et al., 2025] and CHEMBERTA [Ahmad et al., 2022, Chithrananda et al., 2020] for greater expressivity. Both output predicted SMILES strings (a standard text-based encoding of molecular structure), which are deterministically converted to molecular graphs via RDKit [Landrum et al., 2013], where nodes correspond to atom types. This yields the two-step pipeline MS/MS $\rightarrow$ SMILES $\rightarrow$ Graph, which we treat as a graph-valued predictor $f_\theta : \mathcal{X} \rightarrow \mathfrak{M}$.

Remark 6.1. *Our focus is uncertainty quantification in a graph-valued output space, not improving MS/MS-to-graph prediction. We rely on the two-step pipeline above because direct MS/MS-to-graph approaches were not sufficiently accurate on this task.*

Candidate sets. For each spectrum x , we follow the same candidate library pipeline as *MassSpecGym* [Bushuiev et al., 2024], which retrieves all molecules whose mass matches the observed spectrum x and caps the set at $|\mathcal{L}(x)| \leq 256$. All candidates are then converted to graphs via the same deterministic mapping, yielding a finite set $\mathcal{L}(x) \subset \mathfrak{M}$.

Remark 6.2. *By construction of the candidates, the ground-truth molecule is always contained in $\mathcal{L}(x)$, so the completeness assumption $\mathbb{P}(Y_{n+1} \in \mathcal{L}(X_{n+1})) = 1$ is satisfied.*

6.3 EXPERIMENTAL PROTOCOL

Distance and nonconformity score. Our nonconformity score is based on the Z -GW, instantiated in practice as FGW with task-specific cost matrices. For Coloring, we use FGW with adjacency cost (A) and Feature Diffusion (FD) initialization of the solver. For Molecule Retrieval, we use FGW with Laplacian cost (L) and identity initialization (Id).

Calibration. Standard conformal prediction (CP) estimates a single global $(1 - \alpha)$ quantile from the calibration set, regardless of the input. Score Conformalized Quantile Regression (SCQR) instead conditions the threshold on an input-dependent attribute $\omega(x)$, allowing the prediction set

Table 1: Comparison of conformal prediction (CP) and score conformalized quantile regression (SCQR) on graph-valued outputs with finite candidate sets. Acc denotes the base predictor retrieval accuracy. We report the metrics on non-empty predicted conformal sets, hence the changes in candidate set sizes.

Dataset	Model f_θ	Method	Emp. Cov. (%)	Conformal Set Size ↓		Candidate Set Size		Reduction % ↑		Empty rate (%) ↓
				Mean	Median	Mean	Median	Mean	Median	
Coverage $1 - \alpha = 90\%$										
Colors	ANY2GRAPH Acc: 82%	CP	90.2	4	1	205	223	95.8%	98.9%	8.9%
		SCQR $ \mathcal{L}(x) $	90.3	4	1	201	223	95.6%	98.9%	8.8%
		SCQR ResNet(x)	93.0	4	1	205	218	96.2%	98.9%	5.7%
Spectraverse Adduct: [M+CH ₃ COOH-H] [−]	MSALIGN Acc: 71.8%	CP	90.0	59	56	152	159	60.4%	61.0%	0.0%
		SCQR $ \mathcal{L}(x) $	90.0	47	48	147	159	64.5%	64.5%	0.0%
		SCQR DREAMS(x)	89.7	29	19	151	159	79.4%	86.7%	0.7%
Coverage $1 - \alpha = 80\%$										
Spectraverse Adduct: [M+CH ₃ COOH-H] [−]	EMBCos Acc: 59.9%	CP	84.7	51.9	53	144.3	133	64.1%	63.5%	0.0%
		SCQR $ \mathcal{L}(x) $	84.7	52.1	54	144.3	133	63.9%	63.5%	0.0%
		SCQR DREAMS(x)	79.7	27.8	18	144.1	159	80.0%	84.3%	0.0%

size to adapt to properties of x . For the synthetic Coloring task, we consider two choices of $\omega(x)$: (i) the candidate set cardinality $\omega(x) = |\mathcal{L}(x)|$, and (ii) an image embedding obtained from a ResNet(x) variant [He et al., 2016]. For the molecule retrieval task, we similarly consider two variants: (i) the candidate set cardinality $\omega(x) = |\mathcal{L}(x)|$, and (ii) the learned spectral embedding $\omega(x) = \text{DREAMS}(x)$ of the input spectrum x , projected via Random Fourier Features [Rahimi and Recht, 2007]. Because the DREAMS(x) embedding space is high-dimensional and only on the order of a thousand adduct-specific calibration points were available, directly regressing the conformal score in this space led to overfitting. We therefore apply random Fourier features (RFF), which regularizes the SCQR fit and enabled stable training despite the limited calibration data.

Evaluation metrics. All results are reported at nominal level $1 - \alpha = 0.9$. We measure empirical coverage, conformal set size (mean and median), reduction relative to the full candidate library, and empty-set rate.

6.4 RESULTS

Coverage Validity and Efficiency. Across both tasks, empirical coverage remains close to the nominal levels, confirming validity under Z -GW scoring, as reported in Table 4 and Fig. 5. We evaluate efficiency through conformal set size and reduction relative to the candidate library. On *Coloring*, SCQR conditioned on $|\mathcal{L}(x)|$ performs similarly to CP: the median set size is 1 for both, with over 95% average reduction, and conditioning on the ResNet embedding brings only marginal further gains in mean reduction. On *Molecule Retrieval*, conditioning matters more. SCQR based on $|\mathcal{L}(x)|$ offers only a modest improvement over CP, while conditioning on the spectral embedding DREAMS(x) yields a much sharper reduction in conformal set size, roughly halving it relative to CP while keeping coverage close to nominal. The

same pattern holds at the lower 80% target with the weaker EMB COS base predictor: $|\mathcal{L}(x)|$ -conditioned SCQR tracks CP closely, whereas DREAMS(x)-conditioned SCQR again produces markedly smaller sets.

Choice of conditioning attribute. Table 2 compares conditioning choices for SCQR on the Coloring dataset. While marginal coverage is achieved by all variants, the choice of $\omega(x)$ has a significant impact on conditional coverage. SCQR with image embeddings yields the most robust worst-slab coverage, outperforming candidate set size conditioning, suggesting that richer input representations better capture heteroscedasticity across instances.

Table 2: Marginal and worst-slab coverage for different SCQR conditioning attributes on the Coloring dataset. Slabs group inputs by number of candidates.

Method	Marginal Coverage	Worst-Slab Coverage
CP	90.2%	64.7%
SCQR $ \mathcal{L}(x) $	90.3%	67.6%
SCQR ResNet(x)	90.2%	73.5%

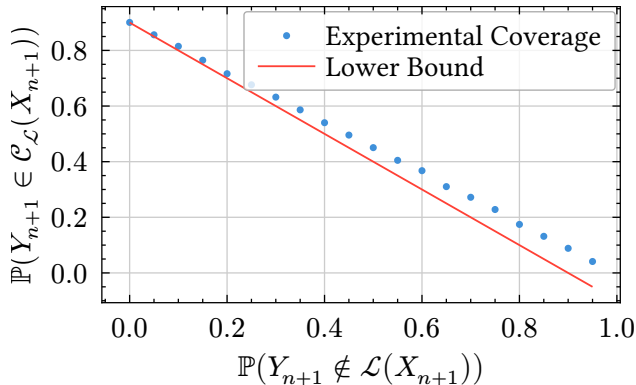


Figure 4: Empirical coverage as a function of the probability of ground-truth absence from the candidate library $\mathbb{P}(Y_{n+1} \notin \mathcal{L}(X_{n+1}))$, evaluated on the Coloring dataset.

Candidate set incompleteness. When the assumption that the ground-truth output is always contained in the candidate library fails, Remark 4.9 provides a theoretical lower bound on coverage as a function of the incompleteness probability $\mathbb{P}(Y_{n+1} \notin \mathcal{L}(X_{n+1}))$. Figure 4 empirically validates this bound on the Coloring dataset by artificially removing the ground truth from the candidate library with increasing probability. Empirical coverage degrades gracefully and remains above the theoretical lower bound throughout.

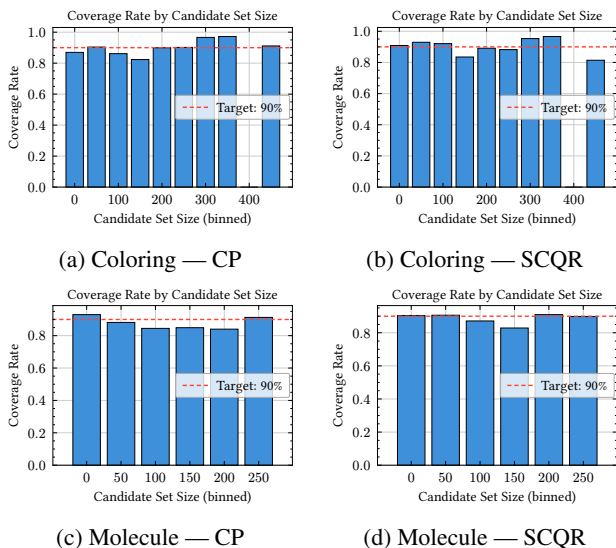


Figure 5: Empirical coverage versus candidate set size, for CP (left) and SCQR (right). For SCQR, the attribute function $\omega(x)$ is task-dependent: $\omega(x) = |\mathcal{L}(x)|$ (candidate set size) for Coloring, and $\omega(x) = \text{DREAMS}(x)$ (mass-spectrum embedding) for Molecule Retrieval.

7 CONCLUSION

We introduced a conformal prediction framework for graph-valued outputs grounded in the Z -Gromov–Wasserstein geometry. By defining nonconformity through Z -GW, the method provides distribution-free coverage guarantees while respecting the permutation-invariant structure of graphs. We further proposed Score Conformalized Quantile Regression (SCQR), which yields adaptive prediction sets by conditioning on input-dependent signals, improving efficiency without compromising validity. More broadly, the framework extends naturally to other metric for graph data and to other structured output spaces representable as Z -networks, such as meshes, point clouds, and distributions, suggesting a general geometry-aware approach to uncertainty quantification in structured prediction.

Acknowledgements

The authors thank Aymeric Dieuleveut for helpful discussions and feedback and Paul Krzakala for his insights on

the experimental part of the work. The first author is funded through the PEPR IA FOUNDRY (ANR-23-PEIA-0003) and the second author through the PRCE Far-see (ANR-24-CE23-0921).

It also received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement 101120237 (ELIAS). The author(s) would also like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme Representing, calibrating & leveraging prediction uncertainty from statistics to machine learning, where work on this paper was undertaken. This work was supported by EPSRC grant EP/Z000580/1.

References

- Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.
- Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and trends® in machine learning*, 16(4):494–591, 2023.
- Martin Bauer, Facundo Mémoli, Tom Needham, and Mao Nishino. The z-gromov-wasserstein distance. *Journal of Machine Learning Research*, 26(291):1–57, 2025.
- Luc Brogat-Motte, Rémi Flamary, Céline Brouard, Juho Rousu, and Florence d’Alché Buc. Learning to predict graphs with fused gromov-wasserstein barycenters. In *International Conference on Machine Learning*, pages 2321–2335. PMLR, 2022.
- Céline Brouard, Huibin Shen, Kai Dührkop, Florence d’Alché Buc, Sebastian Böcker, and Juho Rousu. Fast metabolite identification with input output kernel regression. *Bioinformatics*, 32(12):i28–i36, 2016.
- Roman Bushuiev, Anton Bushuiev, Niek F de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, et al. Massspecgym: A benchmark for the discovery and identification of molecules. *Advances in Neural Information Processing Systems*, 37:110010–110027, 2024.
- Roman Bushuiev, Anton Bushuiev, Raman Samusevich, Corinna Brungs, Josef Sivic, and Tomáš Pluskal. Self-supervised learning of molecular representations from millions of tandem mass spectra using dreams. *Nature Biotechnology*, pages 1–11, 2025.
- Anna Calissano, Aasa Feragen, and Simone Vantini. Graph-valued regression: Prediction of unlabelled networks in a non-euclidean graph space. *Journal of Multivariate Analysis*, 190:104950, 2022.

- Anna Calissano, Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction sets for populations of graphs. *arXiv preprint arXiv:2404.18862*, 2024.
- Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.
- Gaetan De Waele, Marek Wydmuch, Willem Waegeman, et al. Small molecule retrieval from tandem mass spectrometry: what are we optimizing for? *arXiv preprint arXiv:2602.16507*, 2026.
- Victor Dheur, Matteo Fontana, Yorick Estievenart, Naomi Desobry, and Souhaib Ben Taieb. A unified comparative study with generalized conformity scores for multi-output conformal regression. in *Int. Conf. on Mach. Learning (ICML)*, 2025.
- Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.
- Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2015.
- Vishu Gupta, Hantao Qiang, Hsin-Hsiang Chung, Ehud Herbst, and Michael A Skinnider. Comprehensive curation and harmonization of small-molecule ms/ms libraries in spectraverse. *Analytical Chemistry*, 98(5):3934–3943, 2026.
- Soroush H Zargarbashi and Aleksandar Bojchevski. Conformal inductive graph neural networks. In *International Conference on Learning Representations*, volume 2024, pages 55175–55197, 2024.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Kexin Huang, Ying Jin, Emmanuel Candes, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems*, 36:26699–26721, 2023.
- Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica: Journal of the Econometric Society*, 46(1): 33–50, 1978.
- Vladimir Kondratyev, Alexander Fishkov, Nikita Kotelevskii, Mahmoud Hegazy, Remi Flamary, Maxim Panov, and Eric Moulines. Neural optimal transport meets multivariate conformal prediction. *arXiv preprint arXiv:2509.25444*, 2025.
- Paul Krzakala, Junjie Yang, Rémi Flamary, Florence d’Alché Buc, Charlotte Laclau, and Matthieu Labeau. Any2graph: Deep end-to-end supervised graph prediction with an optimal transport loss. *Advances in Neural Information Processing Systems*, 37:101552–101588, 2024.
- Paul Krzakala, Gabriel Melo, Charlotte Laclau, Florence d’Alché Buc, and Rémi Flamary. The quest for the graph level autoencoder (grale). *arXiv preprint arXiv:2505.22109*, 2025.
- Paul Krzakala, Gabriel Melo, Camille Lançon, Charlotte Laclau, Rémi Flamary, Etienne Thévenot, and Florence d’Alché Buc. Msalign: Aligning molecule and mass spectra foundation models for metabolite identification. *arXiv preprint arXiv:2605.19752*, 2026.
- Greg Landrum et al. Rdkit documentation. *Release*, 1(1-79): 4, 2013.
- Robert Lunde, Elizaveta Levina, and Ji Zhu. Conformal prediction for network-assisted regression. *Journal of the American Statistical Association*, (just-accepted):1–22, 2025.
- Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.
- Dai Hai Nguyen, Canh Hao Nguyen, and Hiroshi Mamitsuka. Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Briefings in bioinformatics*, 20(6):2028–2043, 2019.
- Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.
- Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.
- Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.
- Suprosanna Shit, Rajat Koner, Bastian Wittmann, Johannes Paetzold, Ivan Ezhov, Hongwei Li, Jiazhen Pan, Sahand Sharifzadeh, Georgios Kaissis, Volker Tresp, et al. Relationformer: A unified framework for image-to-graph generation. In *European conference on computer vision*, pages 422–439. Springer, 2022.
- Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. *arXiv preprint arXiv:2501.18991*, 2025.

- Titouan Vayer, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pages 6275–6284. PMLR, 2019.
- Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.
- Junjie Yang, Matthieu Labeau, and Florence d’Alché Buc. Exploiting edge features in graph-based learning with fused network gromov-wasserstein distance. *Transactions on Machine Learning Research*, 2024.
- 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.
- Botong Zhang, Shuo Li, and Osbert Bastani. Conformal structured prediction. *arXiv preprint arXiv:2410.06296*, 2024.
- Yiwen Zhang, Keyan Ding, Yihang Wu, Xiang Zhuang, Yi Yang, Qiang Zhang, and Huajun Chen. Breaking the modality barrier: Generative modeling for accurate molecule retrieval from mass spectra. *arXiv preprint arXiv:2511.06259*, 2025.
- Yuqicheng Zhu, Daniel Hernández, Yuan He, Zifeng Ding, Bo Xiong, Evgeny Kharlamov, and Steffen Staab. Predicate-conditional conformalized answer sets for knowledge graph embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4145–4167, 2025a.
- Yuqicheng Zhu, Nico Potyka, Jiarong Pan, Bo Xiong, Yunjie He, Evgeny Kharlamov, and Steffen Staab. Conformalized answer set prediction for knowledge graph embedding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 731–750, 2025b.
- Yuqicheng Zhu, Jingcheng Wu, Yizhen Wang, Hongkuan Zhou, Jiaoyan Chen, Evgeny Kharlamov, and Steffen Staab. Certainty in uncertainty: Reasoning over uncertain knowledge graphs with statistical guarantees. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8741–8763, 2025c.

A MORE ON THEORETICAL RESULTS

A.1 PROOF OF PROPOSITION 4.4

Proof. Fix orderings $V_1 = \{v_1, \dots, v_n\}$ and $V_2 = \{w_1, \dots, w_n\}$ and identify ω_1, ω_2 with their matrix representations.

($\Rightarrow$) Assume $G_1 \sim G_2$. By Definition 3.4, there exist a finite Z -network $W = (V_W, \omega_W, \mu_W)$ and measure-preserving maps $\phi_1 : V_W \rightarrow V_1$ and $\phi_2 : V_W \rightarrow V_2$ such that

$$\omega_W = \phi_1^* \omega_1 = \phi_2^* \omega_2 \quad \mu_W \otimes \mu_W\text{-a.e.}$$

Since μ_1 is uniform and measure preserving,

$$\mu_W(\phi_1^{-1}(\{v\})) = \mu_1(\{v\}) = \frac{1}{n} \quad \forall v \in V_1,$$

hence ϕ_1 is surjective; similarly for ϕ_2 . Writing $\mu_W = \sum_{k=1}^N c_k \delta_{u_k}$ with $c_k > 0$, the above implies

$$\sum_{u_k : \phi_1(u_k) = v} c_k = \frac{1}{n} \quad \forall v \in V_1,$$

and analogously for ϕ_2 .

Since V_W is finite with full-support measure, the pullback identity holds everywhere:

$$\omega_W(u, u') = \omega_1(\phi_1(u), \phi_1(u')) = \omega_2(\phi_2(u), \phi_2(u')) \quad \forall u, u' \in V_W.$$

Choose a section $s : V_1 \rightarrow V_W$ with $\phi_1 \circ s = \text{id}_{V_1}$ and define $\varphi := \phi_2 \circ s : V_1 \rightarrow V_2$. Then

$$\omega_1(v_i, v_j) = \omega_2(\varphi(v_i), \varphi(v_j)) \quad \forall i, j,$$

so $\omega_1 = \varphi^* \omega_2$. Since μ_1 and μ_2 are uniform atomic measures and φ is measure-preserving, each fiber $\varphi^{-1}(\{w\})$ has mass $1/n$. As $|V_1| = |V_2| = n$, this implies that all fibers are singletons and hence φ is bijective. Let $P = (P_{ij}) \in \sigma_n$ be the permutation matrix defined by

$$P_{ij} := \mathbb{I}_{\{w_j = \varphi(v_i)\}},$$

so that $\omega_1 = P^\top \omega_2 P$.

($\Leftarrow$) If $\omega_1 = P^\top \omega_2 P$ for some $P \in \sigma_n$, let φ be the associated bijection and set $W := G_1$, $\phi_1 := \text{id}$, $\phi_2 := \varphi$. Both maps are measure-preserving and the pullback identity holds everywhere, hence $G_1 \sim G_2$. $\square$

A.2 EXTRA RESULTS FOR SECTION 4

Lemma A.1 (Measurable maps preserve equality in distribution). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let Z and Z' be random variables taking values in a measurable space $(\mathcal{Z}, \Sigma_{\mathcal{Z}})$, and let $g : \mathcal{Z} \rightarrow \mathcal{Y}$ be a measurable map into another measurable space $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$. If*

$$Z \stackrel{d}{=} Z',$$

then

$$g(Z) \stackrel{d}{=} g(Z').$$

Proof. Let $B \in \Sigma_{\mathcal{Y}}$ be arbitrary. Since g is measurable, $g^{-1}(B) \in \Sigma_{\mathcal{Z}}$, and therefore

$$\mathbb{P}(g(Z) \in B) = \mathbb{P}(Z \in g^{-1}(B)) = \mathbb{P}(Z' \in g^{-1}(B)) = \mathbb{P}(g(Z') \in B).$$

Since this holds for all measurable sets $B \subseteq \mathcal{Y}$, we conclude that $g(Z)$ and $g(Z')$ have the same distribution. $\square$

A.3 PROOF OF LEMMA 4.7

Proof. Exchangeability of $(X_i, Y_i)_{i=1}^n$ means that for any permutation σ of $\{1, \dots, n\}$,

$$(X_1, Y_1, \dots, X_n, Y_n) \stackrel{d}{=} (X_{\sigma(1)}, Y_{\sigma(1)}, \dots, X_{\sigma(n)}, Y_{\sigma(n)}).$$

Define the measurable map $g : \mathcal{X} \times \mathfrak{M} \rightarrow \mathcal{X} \times \mathcal{M}$, $g(x, y) := (x, h(y))$. Since h is measurable, so is g . Applying g componentwise and invoking Lemma A.1, we obtain

$$(X_1, \tilde{Y}_1, \dots, X_n, \tilde{Y}_n) \stackrel{d}{=} (X_{\sigma(1)}, \tilde{Y}_{\sigma(1)}, \dots, X_{\sigma(n)}, \tilde{Y}_{\sigma(n)}),$$

where $\tilde{Y}_i = h(Y_i)$. Since this holds for all permutations σ , the sequence $(X_i, \tilde{Y}_i)_{i=1}^n$ is exchangeable. $\square$

A.4 PROOF OF PROPOSITION 4.8

Proof. By invariance of s under $\sim$, it factors through the canonical projection h and induces a score on $\mathcal{M}$. By Lemma A.1, the score sequence $\{s(X_i, Y_i)\}_{i=1}^{n+1}$ is exchangeable. Standard conformal prediction theory Angelopoulos et al. [2023] then yields

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_\alpha^{\mathfrak{M}}(X_{n+1})) \geq 1 - \alpha.$$

By construction, membership in $\mathcal{C}_\alpha^{\mathfrak{M}}(x)$ is characterized by the inequality $s(x, y) \leq \hat{q}_{1-\alpha}$. Since $s(x, y) = \tilde{s}(x, h(y))$, this criterion depends on y only through its equivalence class $h(y)$: for any $y, y' \in \mathfrak{M}$ with $h(y) = h(y')$,

$$y \in \mathcal{C}_\alpha^{\mathfrak{M}}(x) \iff y' \in \mathcal{C}_\alpha^{\mathfrak{M}}(x).$$

In other words, $\mathcal{C}_\alpha^{\mathfrak{M}}(x)$ is a union of equivalence classes, containing either all node relabelings of a given graph or none of them. In particular, denoting by $h^{-1}(B) := \{y \in \mathfrak{M} : h(y) \in B\}$ the preimage of a set $B \subseteq \mathcal{M}$,

$$h^{-1}(\mathcal{C}_\alpha^{\mathcal{M}}(x)) = \mathcal{C}_\alpha^{\mathfrak{M}}(x), \quad \text{where } \mathcal{C}_\alpha^{\mathcal{M}}(x) := h(\mathcal{C}_\alpha^{\mathfrak{M}}(x)).$$

Let $\tilde{Y}_{n+1} := h(Y_{n+1})$; its law is the pushforward measure $P_{\tilde{Y}_{n+1}} = h_\# P_{Y_{n+1}}$, which is well defined since $\mathcal{M}$ is equipped with the quotient σ -algebra, under which h is measurable by construction. Then

$$\begin{aligned} \mathbb{P}(\tilde{Y}_{n+1} \in \mathcal{C}_\alpha^{\mathcal{M}}(X_{n+1})) &= \mathbb{P}(Y_{n+1} \in h^{-1}(\mathcal{C}_\alpha^{\mathcal{M}}(X_{n+1}))) \\ &= \mathbb{P}(Y_{n+1} \in \mathcal{C}_\alpha^{\mathfrak{M}}(X_{n+1})), \end{aligned}$$

Combining with the coverage guarantee in $\mathfrak{M}$ yields $\mathbb{P}(\tilde{Y}_{n+1} \in \mathcal{C}_\alpha^{\mathcal{M}}(X_{n+1})) \geq 1 - \alpha$. $\square$

A.5 PROOF OF PROPOSITION 5.1

Proof. Define the residuals $E_i = s(X_i, Y_i) - \psi(\omega(X_i))$ for $i = 1, \dots, n+1$. Since $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable and the functions s , ω , and ψ are measurable and independent of the calibration data, the sequence $(E_1, \dots, E_{n+1})$ is exchangeable.

For $\beta \in (0, 1)$, let $q_\beta(V_1, \dots, V_k)$ denote the $\lceil \beta k \rceil$ -th smallest value of $(V_1, \dots, V_k)$, so that the threshold reads $\hat{q}_{1-\alpha} = q_{1-\alpha}(E_1, \dots, E_n, +\infty)$. First note that

$$E_{n+1} \leq q_{1-\alpha}(E_1, \dots, E_n, +\infty) \iff E_{n+1} \leq q_{1-\alpha}(E_1, \dots, E_n, E_{n+1}),$$

since replacing $+\infty$ by E_{n+1} leaves the $\lceil (n+1)(1-\alpha) \rceil$ -th smallest value unchanged whenever E_{n+1} lies above it, and both events hold whenever E_{n+1} lies below it. By exchangeability of $(E_1, \dots, E_{n+1})$, the event that E_{n+1} is among the $\lceil (n+1)(1-\alpha) \rceil$ smallest of the $n+1$ residuals has probability at least $\lceil (n+1)(1-\alpha) \rceil / (n+1)$, hence

$$\mathbb{P}(E_{n+1} \leq \hat{q}_{1-\alpha}) \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \geq 1 - \alpha.$$

Since $Y_{n+1} \in \mathcal{C}(X_{n+1}) \iff E_{n+1} \leq \hat{q}_{1-\alpha}$, the marginal coverage guarantee follows. $\square$

B EVALUATION METRICS

This appendix details the metrics reported in Table 4 for comparing conformal prediction (CP) and score-conditional quantile regression (SCQR) on graph-valued outputs with finite candidate sets.

Setup. For each test input x , let $\mathcal{L}(x)$ denote the associated candidate set of graphs, and let $\mathcal{C}_\alpha(x) \subseteq \mathcal{L}(x)$ be the conformal prediction set constructed at nominal coverage level $1 - \alpha$. Let y denote the ground-truth graph.

Empirical Coverage. Empirical coverage is computed as

$$\widehat{\text{Cov}}(1 - \alpha) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(x,y) \in \mathcal{D}_{\text{test}}} \mathbb{I}\{y \in \mathcal{C}_\alpha(x)\}.$$

Both CP and SCQR are evaluated at the same nominal level $1 - \alpha$.

Prediction Set Size. The prediction set size for an input x is given by $|\mathcal{C}_\alpha(x)|$. We report both the mean and the median of this quantity across the test set to account for potential skewness in the distribution of set sizes.

Relative Reduction. To quantify efficiency relative to the original candidate set, we define the relative reduction for each input as

$$\text{Reduction}(x) = \frac{|\mathcal{L}(x)| - |\mathcal{C}_\alpha(x)|}{|\mathcal{L}(x)|} \times 100\%.$$

We report the mean and median reduction over the test set.

Empty Set Rate. The empty set rate is defined as

$$\frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(x,y) \in \mathcal{D}_{\text{test}}} \mathbb{I}\{\mathcal{C}_\alpha(x) = \emptyset\}.$$

This quantity reflects how often the conformal procedure returns no candidate graph and is reported as a sanity check in finite-candidate settings.

Remark B.1. Reporting both mean and median statistics is crucial due to the typically heavy-tailed distribution of candidate set sizes. While both CP and SCQR are guaranteed to satisfy marginal coverage, differences in set size and reduction reflect their relative efficiency, with SCQR expected to yield more adaptive prediction sets.

Remark B.2. We distinguish between validity, defined as marginal coverage, and efficiency, defined as the size of the conformal prediction set. Since prediction set size is meaningful only when the ground truth is contained in the conformal set, we report mean and median **prediction set size** and **relative reduction** conditioned on coverage, i.e., over the subset

$$\{x \mid (x, y) \in \mathcal{D}_{\text{test}}, y \in \mathcal{C}_\alpha(x)\}.$$

As a consequence, the reported candidate set statistics in Table 4 may differ across methods, since they are computed only over the subset of inputs for which the conformal set contains the ground truth. Because each method achieves coverage on slightly different instances, the corresponding candidate set averages are evaluated on different subsets. Coverage itself is always reported over the full test set.

C EXPERIMENTAL DETAILS

C.1 DATASETS

We describe here the datasets and evaluation protocols used for the Coloring.

Coloring (Synthetic). For the Coloring task, we use the synthetic dataset introduced in Krzakala et al. [2024]. The dataset already provides predefined test and validation splits, in which we used 10k for each. We use the test split for conformal calibration and report all evaluation metrics on the validation split.

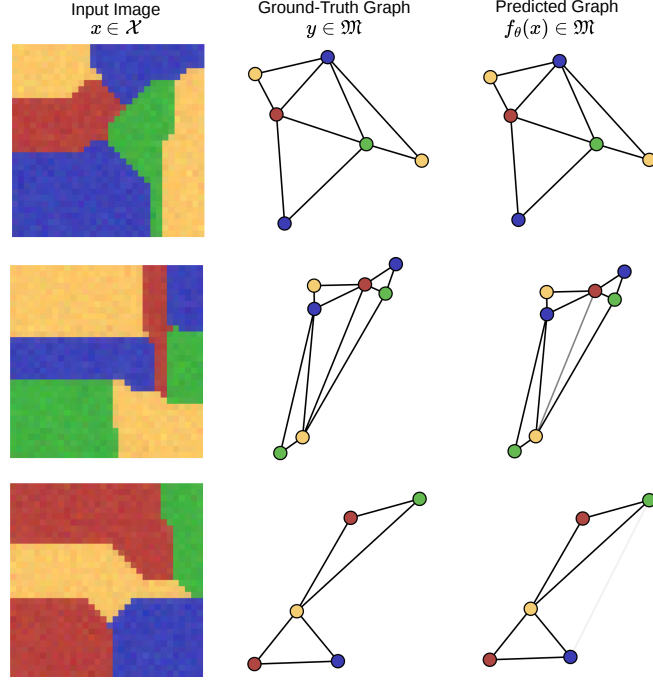


Figure 6: Coloring dataset examples using ANY2GRAPH as graph predictor.

C.2 TRAINING

For the Coloring task, we use ANY2GRAPH [Krzakala et al., 2024]. Both models were trained using the default configurations provided in their respective official repositories, without additional hyperparameter tuning.

Training was performed on NVIDIA A100 GPU. Conformal calibration and evaluation were conducted on CPU.

SCQR calibration based on candidate set size $|\mathcal{L}(x)|$ was done by fitting a linear quantile regressor to predict calibration set non-conformity scores from $|\mathcal{L}(x)|$.

SCQR calibration based on DREAMS embeddings was done by training a 2-layer neural network to predict non-conformity scores from them.

The architecture is as follows:

- Linear(768, 384)
- ReLU
- Linear(384, 1)

The network was trained with the following parameters:

Table 3: embedding-based SCQR neural network training parameters

Parameter	Value
Optimizer	Adam
Learning rate	0.001
Batch size	32
Early stopping patience	3
Early stopping minimum loss improvement	0.0001

Complete training on the 10 000 samples of the calibration set takes a few seconds.

C.3 EXECUTION TIME

Execution time is a concern when working with graphs, especially in conformal prediction where a single prediction may be compared to thousands of candidates. We report here the effects of various FGW hyperparameters on computation time on the Coloring dataset.

The following results were obtained using a Lenovo P16s Gen 4 with an intel ultra 7 chip.

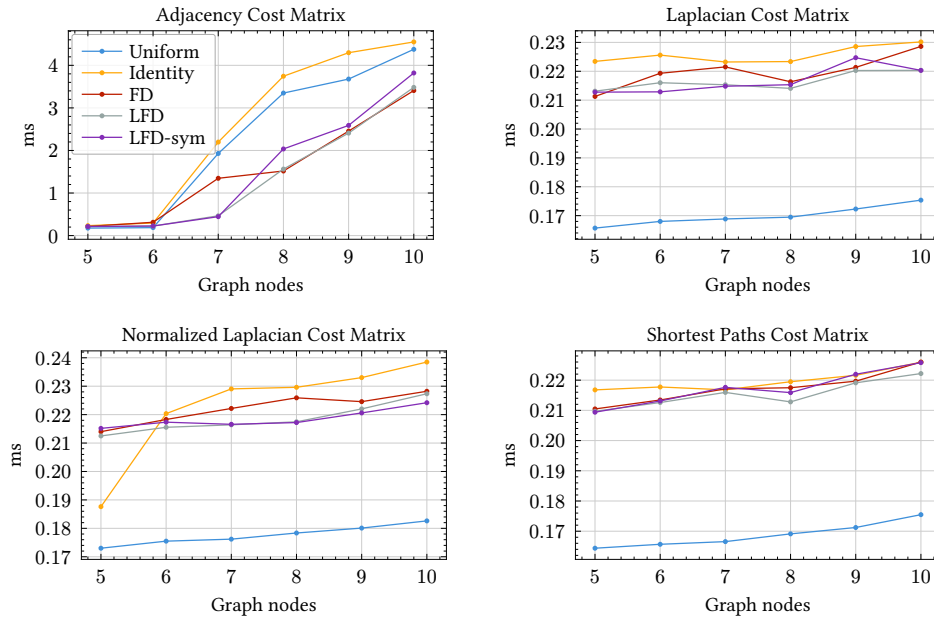


Figure 7: Time taken to compute a FGW distance between a pair of Coloring dataset graphs with various amounts of nodes. Notice how the performance of uniform initialization is offset from the others, given that it is the default initialization of the FGW algorithm, not requiring prior computations. The adjacency cost matrix scales very badly with the amount of nodes. At 10 nodes per graph, it is already 20 times slower than the other options.

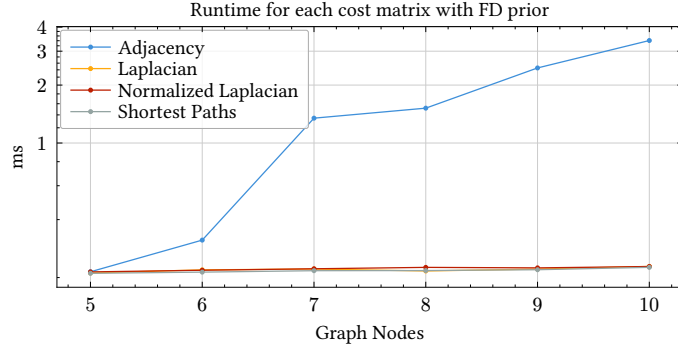


Figure 8: Computation time of the FGW distance between a pair of coloring dataset graphs with the FD prior. Although the adjacency cost matrix yields good conformal performance, it is not recommended when working with large graphs.

C.4 EXTRA RESULTS

Table 4: Comparison of conformal prediction (CP) and score conformalized quantile regression (SCQR) on graph-valued outputs with finite candidate sets. Acc denotes the base predictor retrieval accuracy.

Dataset	Model f_θ	Method	Emp. Cov. (%)	Conformal Set Size ↓		Candidate Set Size		Reduction % ↑		Empty rate (%) ↓
				Mean	Median	Mean	Median	Mean	Median	
Coverage $1 - \alpha = 90\%$										
Colors	ANY2GRAPH Acc: 82%	CP	90.2	4	1	205	223	95.8%	98.9%	8.9%
		SCQR $ \mathcal{L}(x) $	90.3	4	1	201	223	95.6%	98.9%	8.8%
		SCQR ResNet(x)	93.0	4	1	205	218	96.2%	98.9%	5.7%
Spectraverse Adduct: [M+CH ₃ COOH-H] [−]	MSALIGN Acc: 71.8%	CP	90.0	59	56	152	159	60.4%	61.0%	0.0%
		SCQR $ \mathcal{L}(x) $	90.0	47	48	147	159	64.5%	64.5%	0.0%
		SCQR DREAMS(x)	89.7	29	19	151	159	79.4%	86.7%	0.7%
Spectraverse Adduct: [M+HCOOH-H] [−]	MSALIGN Acc: 31.1%	CP	90.0	166	184	250	256	33.4%	25.8%	0.0%
		SCQR $ \mathcal{L}(x) $	90.9	165	184	252	256	34.4%	25.8%	0.9%
		SCQR DREAMS(x)	90.2	143	154	250	256	42.6%	36.3%	0.2%
Coverage $1 - \alpha = 80\%$										
Spectraverse Adduct: [M+CH ₃ COOH-H] [−]	EMBCOS Acc: 59.9%	CP	84.7	51.9	53	144.3	133	64.1%	63.5%	0.0%
		SCQR $ \mathcal{L}(x) $	84.7	52.1	54	144.3	133	63.9%	63.5%	0.0%
		SCQR DREAMS(x)	79.7	27.8	18	144.1	159	80.0%	84.3%	0.0%
Spectraverse Adduct: [M+K] ⁺	EMBCOS Acc: 25.0%	CP	87.2	120.0	95	255.8	256	53.1%	62.9%	0.0%
		SCQR $ \mathcal{L}(x) $	86.4	119.0	93	255.8	256	53.5%	63.7%	0.0%
		SCQR DREAMS(x)	83.4	112.8	85	255.8	256	55.9%	66.8%	0.0%
Spectraverse Adduct: [M] ⁺	EMBCOS Acc: 37.9%	CP	80.2	124.8	146	250.0	256	49.7%	43.0%	0.0%
		SCQR $ \mathcal{L}(x) $	88.8	165.0	151	250.6	256	34.4%	40.6%	0.0%
		SCQR DREAMS(x)	79.3	108.0	144	250.0	256	56.2%	43.8%	2.6%

C.5 RESULTS BY HYPERPARAMETER

The choices of the G_0 FGW prior transport and of the cost matrix both influence the speed of convergence and the performance of the produced distances for conformal set prediction.

Table 5: Summary of acronyms used in the experiments.

Category	Acronym	Description
Cost Matrix C	A	Adjacency
	L	Laplacian
	L^{sym}	Laplacian (symmetrically normalized)
	SP	Shortest Paths
Initial Transport Plan G_0	FD	Feature Diffusion: Solution to the <i>Earth Mover's Distance</i> problem for given nodes of feature augmented vector (F, AF)
	LFD	Laplacian Feature Diffusion: Same as Feature Diffusion but with features (F, LF)
	LFD-sym	Sym-Norm Laplacian Feature Diffusion: Same as Feature Diffusion but with features $(F, L^{sym}F)$
	Id	Identity matrix: when the two graphs have the same amount of nodes, else use the solver's default initialization.
	Uniform	Uniform: FGW solver default initialization (uniform transport).

Table 6: Cost matrix impact on different tasks. The colors task uses the feature diffusion G_0 prior.

Task	Cost Matrix	Empirical Coverage	Mean set size ↓	Median set size ↓	Mean reduction (%) ↑	Median reduction (%) ↑	Empty rate (%) ↓
Colors	A	90.2%	4	1	95.8%	98.9%	9.0%
	L^{sym}	90.4%	4	1	95.8%	98.9%	8.5%
	L	90.8%	4	1	95.5%	98.7%	8.3%
	SP	90.1%	205	223	2.2%	0.0%	2.6%

Table 7: G_0 prior impact on different tasks. The colors task uses the adjacency cost matrix. Note that uniform G_0 prior often converges very slowly.

Task	G_0	Empirical Coverage	Mean set size ↓	Median set size ↓	Mean reduction (%) ↑	Median reduction (%) ↑	Empty rate (%) ↓
Colors	LFD	90.4%	4	1	95.8%	98.9%	8.7%
	FD	90.2%	4	1	95.8%	98.9%	9.0%
	Uniform	90.1%	4	1	95.8%	98.9%	8.7%
	LFD-sym	90.0%	4	1	95.8%	98.9%	8.8%
	Id	90.4%	6	2	94.4%	97.9%	5.1%