Learning to Predict Graphs with Fused Gromov-Wasserstein Barycenters

Luc Brogat-Motte [1], Rémi Flamary [2], Céline Brouard [3], Juho Rousu [4], Florence d'Alché-Buc [1]

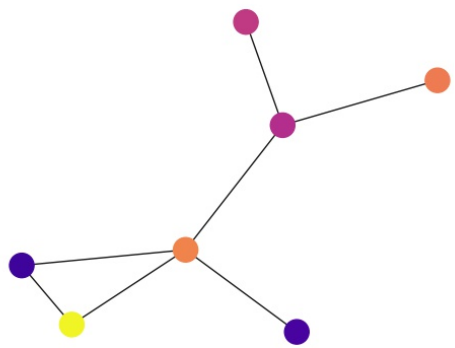
[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2] Ecole Polytechnique, Institut Polytechnique de Paris, CMAP, UMR 7641, Palaiseau, France
[3] Université de Toulouse, INRAE, UR MIAT, France
[4] Department of Computer Science, Aalto University, Finland

Problem setting and contributions

Supervised graph prediction.

Data set: $(x_i, f^*(x_i))_{i=1}^N$

$$x \xrightarrow{f^* = ?}$$



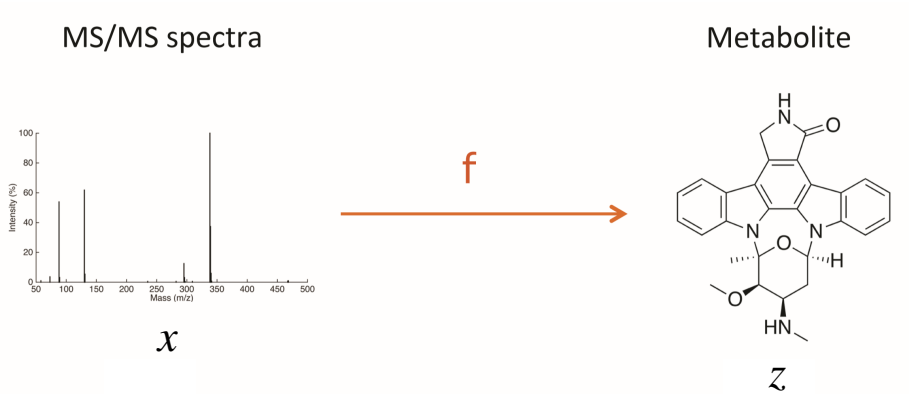
Main contributions in this work.

- We propose a **novel model for graph prediction** with two different training strategies: **kernel method or neural network**.
- We provide **theoretical guarantees** in the nonparametric case.
- We assess the method on two problems: a synthetic and a real-world graph prediction problem.

A real-world graph prediction problem

The metabolite identification problem. The goal is to identify a molecule y from a mass spectrum x.

Dataset. Couples $(x_i, y_i)_{i=1}^N$ are available with typically $N \sim 10^3$.



Molecules are well-represented as graphs where the atoms are nodes and the chemical bonds are edges. Current SOTA methods obtain modest accuracies, and still have difficulty in predicting novel molecules (Stravs et al., 2021).

Outline

I. Supervised graph prediction

II. FGW: a distance between graphs from optimal transport

III. Proposed method

IV. Experiments

I. Supervised graph prediction

Statistical learning problem

Supervised graph prediction. For a given loss $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$. Given a finite sample $(x_i, y_i)_{i=1}^N$ independently drawn from an unknown distribution ρ on $\mathcal{X} \times \mathcal{Y}$, we aim at estimating the target function $f^* : \mathcal{X} \to \mathcal{Y}$ minimizing the expected risk:

$$\mathcal{R}_\Delta(f) = \mathbb{E}_\rho[\Delta(f(X), Y)]. \tag{1}$$

Empirical risk minimization. An estimate $\hat{f}$ of f^* can be obtained by minimizing the empirical risk:

$$\hat{\mathcal{R}}_\Delta(f) = \frac{1}{N} \sum_{i=1}^{N} \Delta(f(x_i), y_i), \tag{2}$$

over a chosen hypothesis space $\mathcal{G}$ of function $f : \mathcal{X} \to \mathcal{Y}$.

Challenges of supervised graph prediction

Challenges.

- **1) Modelling.** How to define a relevant interpolant map from the data sets $(x_i, y_i)_{i=1}^n$? This is tied to finding a proper way of computing weighted averages of outputs.
- **2) Computation.** Because of 1), one ends up with complex models (non-smoothly parameterized models, and costly inference computations), such that solving empirical risk minimization, e.g. via gradient descent, is very difficult.

Overview of the state-of-the-art for graph prediction.

- Using standard structured prediction approaches combined with expert-derived graph representations (Brouard et al., 2016).
- Generative neural networks building graph sequentially (Liao et al., 2020) for **unsupervised** graph prediction.

II. FGW: a distance between graphs from optimal transport

Graphs as metric measure spaces

Feature space. Each node of a graph have a label represented as a vector in $\mathbb{R}^d$.

$$\mathcal{F} \subset \mathbb{R}^d \quad \text{with } |\mathcal{F}| < \infty \tag{3}$$

Output space: discrete graph space.

$$\mathcal{Y} = \left\{ (C, F, h) \,|\, n \leq n_{max}, C \in \{0,1\}^{n \times n}, C^T = C, F = (F_i)_{i=1}^n \in \mathcal{F}^n, h = \frac{1}{n} \mathbb{1}_n \right\} \tag{4}$$

Prediction space: continuous relaxed graph space.

$$\mathcal{Z}_n = \left\{ (C, F, h) \,|\, C \in [0,1]^{n \times n}, C^T = C, F \in \mathrm{Conv}(\mathcal{F})^n, h = \frac{1}{n} \mathbb{1}_n \right\} \tag{5}$$

Fused Gromov-Wasserstein distance (FGW)

The FGW distance (Vayer et al., 2020) is an extension of the GW distance, which can be used to measure the similarity between attributed graphs.

The FGW distance. $\beta \in [0,1]$, $z_1 = (C_1, F_1)$ and $z_2 = (C_2, F_2)$:

$$\mathrm{FGW}_2^2(z_1, z_2) = \min_{\pi \in \mathcal{P}_{n1,n2}} \sum_{i,k,j,l} \Big[(1-\beta) \|F_1(i) - F_2(j)\|_{\mathbb{R}^d}^2 + \beta (C_1(i,k) - C_2(j,l))^2 \Big] \pi_{i,j} \pi_{k,l}.$$

Two graphs are closed if there exists a transport plan matching their nodes and which preserves the labels and the the pairwise similarities between the nodes.

III. Proposed method

Proposed method

I) FGW as a loss function. We propose to estimate a minimizer $f : \mathcal{X} \to \mathcal{Z}_n$ of

$$\mathcal{R}^n_{\text{FGW}}(f) = \mathbb{E}_\rho[\text{FGW}_2^2(f(x), y)] \tag{6}$$

for a given $n \in \mathbb{N}^*$.

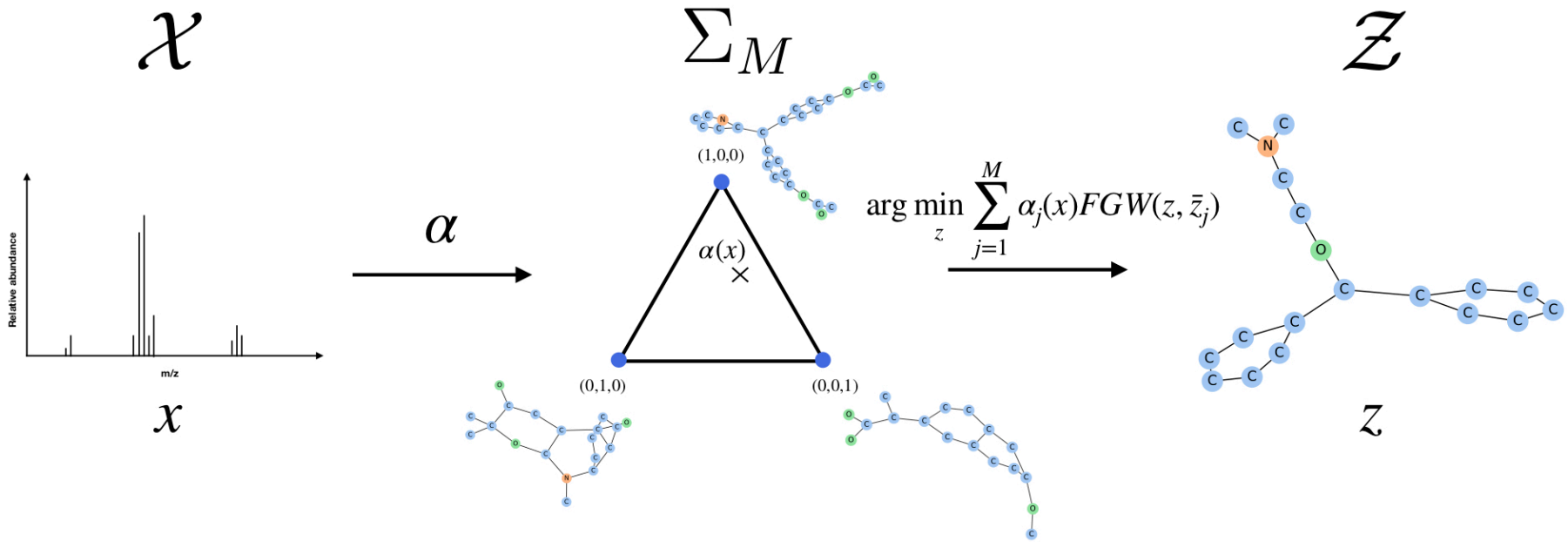
Remark. When n is big enough $\mathcal{Y} \subset \mathcal{Z}_n$[1].

II) Proposed FGW barycentric model. Given M template graphs $\bar{z}_j \in \mathcal{Z}$

$$f_\theta(x) = \underset{z \in \mathcal{Z}_n}{\arg\min} \sum_{j=1}^{M} \alpha_j(x; W)\text{FGW}_2^2(z, \bar{z}_j), \tag{7}$$

where the weights $\alpha_i(x; W) : \mathcal{X} \to \mathbb{R}^+$ are similarity scores between x and x_j.

Properties of the proposed model



Remarks.

- The model's parameters are $\theta = (W, (\bar{z}_j)_{j=1}^M)$.
- The model is invariant under isomorphism.
- In terms of computations, it leverages the advances in computational optimal transport (Peyré et al., 2016, Vayer et al., 2019).

Two fitting strategies

A) Kernel method. Given a p. d. kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $M = N$ and $\bar{z}_j = z_j$

$$\alpha(x) = (K + \lambda I_N)^{-1} k_x \tag{8}$$

with $K = \big(k(x_i, x_j)\big)_{ij} \in \mathbb{R}^{N \times N}$ and $k_x^T = \big(k(x, x_1), \ldots, k(x, x_N)\big)$.

Remark. This **closed-form estimation** is rooted in IOKR (Brouard et al., 2016) and ILE (Ciliberto et al., 2020) frameworks.

B) Neural network.

- $\alpha : \mathcal{X} \to \mathbb{R}^M$ is a neural network.
- α and the template graphs $(\bar{z}_j)_{j=1}^M$ are learned using **stochastic gradient descent**.
- We propose a method to compute a sub-gradient of the loss $\mathrm{FGW}\big(f_\theta(x_i), y_i\big)$.

Theoretical guarantees

Under technical assumptions, the two following guarantees hold for the kernel-based estimator.

Consistency. With probability 1,

$$\lim_{N \to +\infty} \mathcal{R}^n_{\mathrm{FGW}}(\hat{f}) = \mathcal{R}^n_{\mathrm{FGW}}(f^*). \tag{9}$$

Excess-risk bound. With probability $1 - \delta$,

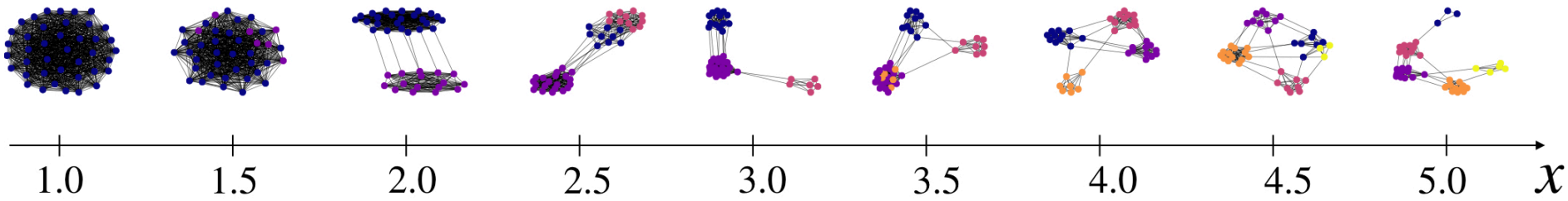
$$\mathcal{R}^n_{\mathrm{FGW}}(\hat{f}) - \mathcal{R}^n_{\mathrm{FGW}}(f^*) \leq c \, \log(4/\delta) \, N^{-1/4}, \tag{10}$$

with c a constant independent of N and δ.

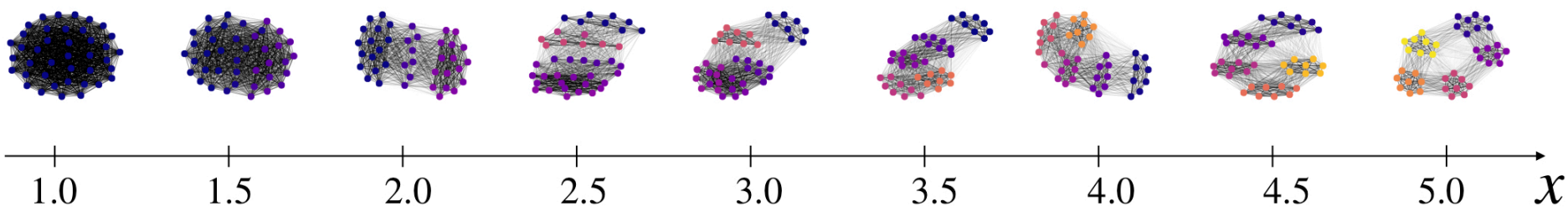
IV. Experiments

Synthetic experiment with neural network

True map. We defined a smooth map $f^* : [0, 6] \to \mathcal{Y}$ which maps x to a stochastic block model with x blocks.



Learned map. We trained the neural network model with 8 templates using 100 i.i.d couples $(x_i, y_i)_{i=1}^{100}$. We obtained the following estimated map $\hat{f} : [0, 6] \to \mathcal{Y}$.



Synthetic experiment with neural network

Learned map visualization.

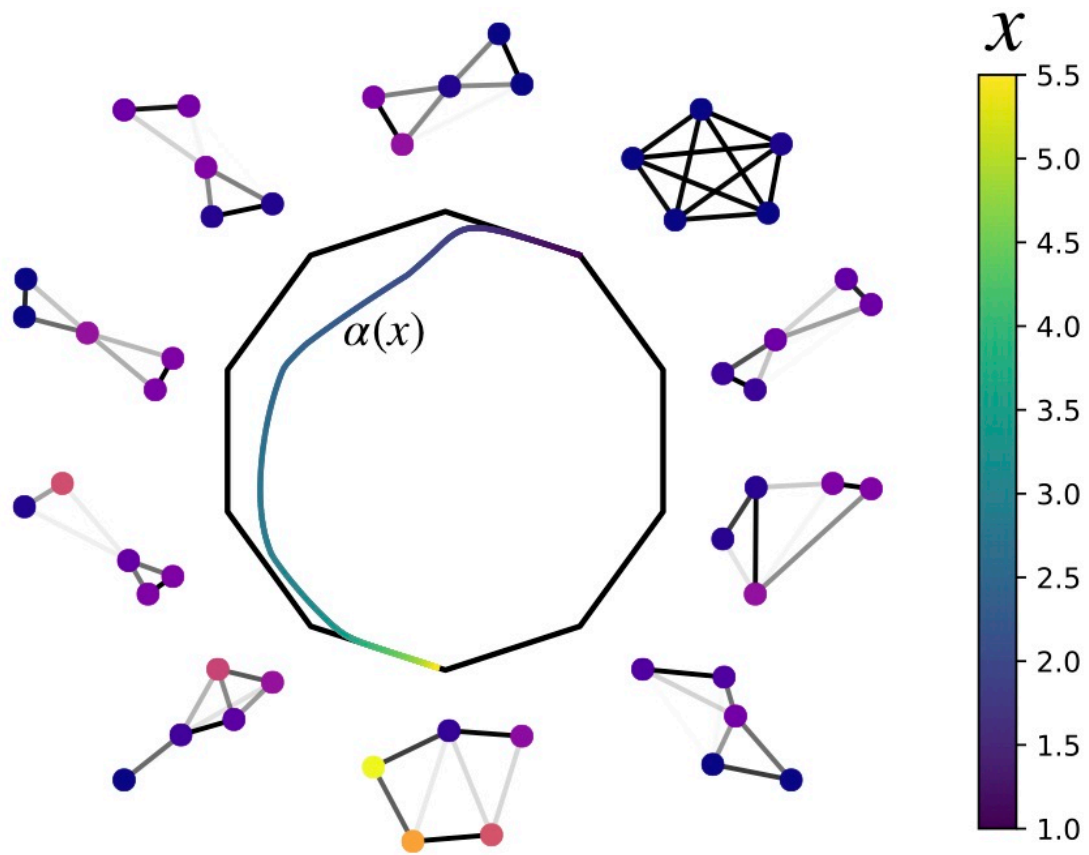


Figure 1: Learned templates $(\bar{z}_j)_{j=1}^m$ on the synthetic dataset and trajectory of the weights $\alpha(x)$ on the simplex as a function of x.

Metabolite identification problem 1/2

Experimental setting. We compare the FGW metric to other graph metrics by changing the metric D in the following model

$$\arg\min_{y \in \mathcal{Y}} \sum_{j=1}^{N} \alpha_j(x; W) D(y, y_j) \tag{11}$$

When D is FGW we recover the proposed method. When D is a kernel we recover the IOKR method for structured prediction.

Obtained results.

- Gaussian fingerprints is state-of-the-arts on this dataset when a candidate set is available.
- FGW greatly benefits from the improved fine and diffuse metrics showing the adaptation potential of the FGW metric to the graph space at hand reaching competitive performance against Fingerprints with linear kernel and beating WL kernels. The method proposed in this work is the first **generic approach** that obtained good Top-k accuracies without using these expert-derived molecular graph representations.

Table of Top-k accuracies

	Top-1	Top-10	Top-20
WL kernel	9.8%	29.1%	37.4%
Linear fingerprint	28.6%	54.5%	59.9%
Gaussian fingeprint	41.0%	62.0%	67.8%
FGW one-hot	12.7%	37.3%	44.2%
FGW fine	18.1%	46.3%	53.7%
FGW diffuse	27.8%	52.8%	59.6%

Table 1: Top-k accuracies for various graph kernels on the metabolite identification dataset.

Comparison with not expert-derived graph representations

Setting 2. In order to define a molecular graph metric, we use the **deep generative graph representations** from MoFlow (Zang et al, 2020) learned from 249.455 molecules and which obtained state-of-the-art results in (unsupervised) molecular graph generation.

Obtained results. FGW diffuse exhibits far better performance than the MoFlow graph distance.

	Top-1	Top-10	Top-20
Gaussian fingerprint	46.2%	77.8%	84.9%
FGW diffuse	40.3%	69.7%	78.3%
MoFlow representat.	20.0%	58.2%	68.4%

Table 2: Top-k accuracies obtained using deep molecular graph representations in comparison to the proposed FGW metric, and expert-derived fingerprint representations.

References

1. Brouard, C., Szafranski, M., & d'Alché-Buc, F. (2016). Input output kernel regression: Supervised and semi-supervised structured output prediction with operator-valued kernels. Journal of Machine Learning Research, 17, np.
2. Ciliberto, C., Rosasco, L., & Rudi, A. (2020). A General Framework for Consistent Structured Prediction with Implicit Loss Embeddings. J. Mach. Learn. Res., 21(98), 1-67.
3. Liao, R., Li, Y., Song, Y., Wang, S., Hamilton, W., Duvenaud, D. K., ... & Zemel, R. (2019). Efficient graph generation with graph recurrent attention networks. Advances in Neural Information Processing Systems, 32.
4. Peyré, G., Cuturi, M., & Solomon, J. (2016, June). Gromov-Wasserstein averaging of kernel and distance matrices. In International Conference on Machine Learning (pp. 2664-2672). PMLR.
5. Stravs, M. A., Dührkop, K., Böcker, S., & Zamboni, N. (2022). MSNovelist: De novo structure generation from mass spectra. Nature Methods, 1-6.
6. Vayer, T., Chapel, L., Flamary, R., Tavenard, R., & Courty, N. (2020). Fused gromov-wasserstein distance for structured objects. Algorithms, 13(9), 212.
7. Titouan, V., Flamary, R., Courty, N., Tavenard, R., & Chapel, L. (2019). Sliced gromov-wasserstein. Advances in Neural Information Processing Systems, 32.
8. Zang, C., & Wang, F. (2020, August). MoFlow: an invertible flow model for generating molecular graphs. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 617-626). ISO 690