Geometry of the Space of Partitioned Networks: A Unified Theoretical and Computational Framework

Stephen Y Zhang[1], Fangfei Lan[2], Youjia Zhou[2], Agnese Barbensi[3], Michael P H Stumpf[1], Bei Wang[2], and Tom Needham[4]

[1]University of Melbourne
[2]University of Utah
[3]University of Queensland
[4]Florida State University

Abstract

Network data are ubiquitous in the real world to capture pairwise or high-order relations among objects. We introduce a class of measure-theoretic network objects called *partitioned measure networks* that generalize a number of objects used to model network data in the literature, such as graphs, hypergraphs, and augmented graphs (i.e., graphs whose nodes are assigned categorical classes). We then propose a metric called a *partitioned network distance* between partitioned measure networks that extends the Gromov-Wasserstein distance between graphs and the co-optimal transport distance between hypergraphs. We characterize the geometry of the space of partitioned measure networks, thereby providing a unified theoretical treatment of generalized network structures that encompass both pairwise and higher-order relations. In particular, we show that our metric defines an Alexandrov space of non-negative curvature, and leverage this structure to define gradients for certain functionals commonly arising in geometric data analysis tasks. We extend our framework to the setting where nodes have additional label information, and derive efficient computational schemes to utilize the partitioned network distance in practice. Equipped with these theoretical and computational tools, we demonstrate the utility of our framework in a suite of applications, including hypergraph alignment, clustering and dictionary learning from ensemble data, multi-omics alignment, as well as multiscale network alignment.

Contents

1 Introduction

Modelling relations among objects or concepts is a central task across the natural sciences, engineering, as well as arts and humanities. Graphs are a conventional approach to model pairwise relations between objects. Many real-world systems, however, involve higher-order interactions among three or more objects. In biochemical reaction networks, reactions typically involve multiple chemical species simultaneously [33]. In coauthorship networks, papers are written jointly by any number of authors [83]. And in a theatre play, each scene can be viewed as a (higher-order) interaction between a set of characters [23]. These systems cannot be modelled as graphs without information loss. Instead, we need to introduce more general structures such as hypergraphs, simplicial complexes, and cell complexes [6].

A natural question arising from the study of graphs is how to compare them: specifically, how to characterize the distance or degree of similarity between two graphs. This is not straightforward, since two graphs may vary in the number of nodes and comparisons must be invariant under permutation [66] and other symmetries. The complexity involving graph data has driven the development of an extensive tool set for graph comparison and matching [75], including spectral methods [29, 30, 46, 66] and graph kernels [61, 9]. Among these methods, Gromov-Wasserstein couplings of *metric measure spaces* (i.e., metric spaces equipped with measures) have proven fruitful from theoretical and computational perspectives [44]. A *coupling* between measures is a relaxed notion of correspondence between objects [44]; and it creates a joint probability space with the desired measures as its marginals. By modelling graphs as metric measure spaces and considering couplings between them, a notion of Gromov-Wasserstein distance between graphs emerges in terms of a *least distortion* principle. This distance is in fact a pseudometric, and the space of graphs (considered up to a natural notion of equivalence) can thus be formalized as a metric space endowed with the Gromov-Wasserstein metric [19]. The geometry of the space of metric measure spaces, endowed with this metric, was studied in detail by Sturm [65] and was shown to be an Alexandrov space with curvature bounded below. This is a powerful characterization that allows well-defined notions of geodesics, tangent spaces, and gradient flows.

Computationally, the Gromov-Wasserstein framework formulates the distance between two graphs as an optimal value of a non-convex quadratic program over a set of feasible couplings. Efficient computational schemes exist to find local minima of this problem, which have given rise to an algorithmic tool box for dealing with graphs that has gained popularity in the statistics and machine learning communities [52].

While the Gromov-Wasserstein approach to graph comparison has been a foundational tool for understanding the space of graphs from a geometric viewpoint, in its basic form, it is insufficient to model higher-order systems such as hypergraphs. A *hypergraph* consists of a set of nodes and a set of hyperedges (i.e., subsets of nodes); if each hyperedge contains exactly two nodes, then a hypergraph reduces to a graph. A hypergraph can be used to encode multi-way relations among nodes. Recently, Chowdhury et al. [22] introduced Gromov-Wasserstein type distances between hypergraphs, based on the co-optimal transport framework of Redko et al. [55]. Framing hypergraphs as metric measure spaces, Chowdhury et al. showed that the space of hypergraphs can be characterized as a metric space; however, an in-depth geometric description of this metric space remains to be fully developed.

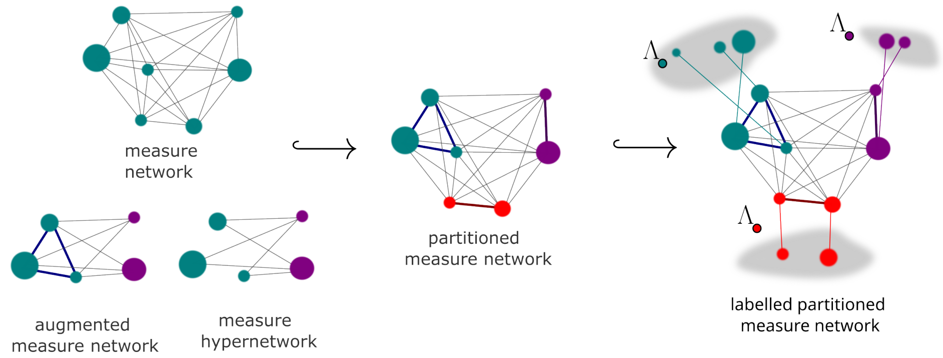
measure network
augmented measure network
measure hypernetwork
partitioned measure network
Λ
Λ
Λ
labelled partitioned measure network

Figure 1: A schematic representation of the types of generalized networks which can be embedded into the space of partitioned measure networks. From left to right: (i) Examples of a measure network (or a weighted graph) whose nodes are endowed with a probability measure; a measure hypernetwork (or a weighted bipartite graph) with a node probability measure; and an augmented measure network (or a measure network) whose node set has been partitioned into two classes. (ii) Illustration of a partitioned measure network with $k = 3$ partitions, that is, a measure network whose nodes have been partitioned into 3 classes. (iii) Illustration of a labelled partitioned measure network, or a partitioned measure network whose nodes are endowed with features in some auxiliary metric spaces.

In this paper, we introduce a general class of network objects called *partitioned measure networks*, which generalize a number of measure-theoretic objects recently introduced in the literature: measure networks [19], measure hypernetworks [22], and augmented measure networks [26]; see Figure 1, and Remark 2.3 for further discussions. A *k-partitioned measure network* is a graph structure whose nodes are separated into k

classes, and this partitioning should be taken into account when comparing these structures. For example, a hypergraph can be encoded as a bipartite graph between two partitions (that correspond to sets of nodes and hyperedges respectively), so that any hypergraph can be modelled as a 2-partitioned measure network; further examples are provided in Example 2.9. We also consider k-partitioned networks whose nodes come with attributes in some auxiliary metric space, referred to as *labelled k-partitioned measure networks.*

Contributions. We highlight our contributions below. First, we equip the space of partitioned measure networks with a family of transport-based distances formulated as the minimum of a quadratic distortion functional over *k-partitioned couplings* (i.e., measure couplings that respect the class structures of the partitioned networks). This choice of distance is shown to be a *bona fide* metric (up to a natural equivalence relation), and we explicitly construct isometric embeddings of measure networks, hypernetworks, and augmented measure networks into this space; see Definition 2.8 and Theorem 1.

Second, we characterize geodesics in the space of partitioned measure networks and show that it is an Alexandrov space with curvature bounded below; see Theorem 3. As an extension to our analysis, we consider the addition of *labels* to partitioned measure networks, which provides a generalization of the so-called *Fused Gromov-Wasserstein* problem [71]. When labels reside in a Hilbert space, we show that our geometric characterization of geodesics and curvature also applies; see Theorem 5.

Third, our theoretical contributions provide a unified treatment of a family of generalized networks, which encompass multiple network objects recently introduced in the literature [19, 55, 22, 26]. This allows us to provide a common geometric description of these spaces. To the best of our knowledge, in the settings of labelled measure networks, measure hypernetworks, and augmented measure networks, these characterizations are new. We conclude our theoretical contributions in Section 4, with a brief discussion of the Riemannian geometric concepts of tangent spaces, exponential and logarithmic maps, and gradients in the space of k-partitioned measure networks, which are crucial for practical applications in learning algorithms. We remark here that a recent paper [5] also provides a general framework for studying several variants of the Gromov-Wasserstein distance, but that the results therein are disjoint from the ones presented here: the framework of the present paper captures different variants of Gromov-Wasserstein distance than that of [5], and the curvature bounds and Riemannian structures established here are not treated in [5].

Fourth, we demonstrate the utility of our framework on computational case studies, which bring together the theoretical ideas. We provide a connection between the Gromov-Wasserstein network matching problem and a family of spectral network alignment algorithms [29, 46]. To our knowledge, this connection has not been explicitly pointed out in prior work. We show that partitioned measure networks provide a more natural and flexible extension to the hypergraph alignment problem, and we demonstrate in numerical experiments that transport-based approaches are more accurate and efficient than spectral methods.

We illustrate the practicality of our methods with applications to metabolic network alignment, simultaneous sample and feature alignment in multi-omics data, and multi-scale network matching. We formulate each of these problems in terms of partitioned network matching. In addition, we investigate some more complex tasks on the space of partitioned networks that exploit their geometric properties. For example, we provide computational characterizations of geodesics and barycenters in the space of measure hypernetworks. We introduce geodesic dictionary learning as a bi-level problem on the space of (partitioned) measure networks, from which we motivate linearized dictionary learning [74] as a fast approximate algorithm. We conclude by demonstrating the utility of geodesic dictionary learning in synthetic hypergraph block models as well as networks derived from atomic and topological representations of small molecules.

Overview. The plan for the paper is as follows. In Section 2, we give precise definitions of various spaces of generalized networks and define our new variant of Gromov-Wasserstein distance—the partitioned network distance—on them. We establish metric properties of the partitioned network distance. We generalize these ideas in Section 3 by introducing node features to the partitioned networks. We define a new metric on these objects and establish its metric properties, including curvature bounds. The deferred proofs from previous sections follow as corollaries to these more general results. We study the Riemannian structures following from our curvature bounds in Section 4, with a focus on computational examples. In Section 5, we give an extended collection of computational examples and applications of our framework. This is supplemented by an open source code repository, available at `https://github.com/zsteve/partitioned_networks`. We conclude the main paper with a discussion of future directions in Section 6, followed by technical aspects of our numerical methods in the Appendix A.

2 Metric geometry of spaces of generalized networks

2.1 Spaces of generalized networks

We first review some notions of generalized networks, and Gromov-Wasserstein type distances between them, which have appeared previously in the literature.

2.1.1 Generalized networks

Let us first recall some definitions of generalized network structures in the literature. These use the following notational conventions: given measure spaces (X, μ) and (Y, ν), we use $L^p(\mu)$ to denote the space of p-integrable functions on X (for $p \in [1, \infty]$) and we use $\mu \otimes \nu$ to denote the product measure on $X \times Y$.

Definition 2.1 (Various notions of generalized networks)**.** *Let $p \in [1, \infty]$.*

1. *A **measure p-network** [19] is a triple $N = (X, \mu, \omega)$, where X is a Polish space, μ is a Borel probability measure on X, and $\omega : X \times X \to [0, \infty)$ is an element of $L^p(\mu \otimes \mu)$. Let $\mathcal{N}^p$ denote the collection of all measure p-networks.*

2. *A **measure p-hypernetwork** [55, 22] is a five-tuple $H = (X, \mu, Y, \nu, \omega)$ consisting of Polish spaces X and Y endowed with Borel probability measures μ and ν, respectively, and $\omega : X \times Y \to [0, \infty)$ is an element of $L^p(\mu \otimes \nu)$. Let $\mathcal{H}^p$ denote the collection of all measure p-hypernetworks.*

3. *An **augmented measure p-network** [26] is a six-tuple $A = (X, \mu, Y, \nu, \omega_X, \omega_{XY})$, consisting of Polish spaces X and Y endowed with Borel probability measures μ and ν, respectively, and $\omega_X : X \times X \to [0, \infty)$ and $\omega_{XY} : X \times Y \to [0, \infty)$ are elements of $L^p(\mu \otimes \mu)$ and $L^p(\mu \otimes \nu)$, respectively. Let $\mathcal{A}^p$ denote the collection of all augmented measure p-networks.*

We will frequently suppress explicit mention of the parameter p, as the appropriate value will typically be either unimportant or clear from context; e.g., when referring to certain L^p-type metrics. The functions ω, ω_X, ω_{XY}, etc., will loosely be referred to as ***network kernels***.

Example 2.2. We now provide prototypical examples of the structures defined above.

1. A **metric measure space** is a measure network (X, μ, ω) such that ω satisfies the axioms of a metric, and the topology of X is induced by the metric. This was the original setting where the Gromov-Wasserstein distances were formulated; see [43, 44] and the related work [64]. More general examples of measure networks frequently come from the setting of graph theory, where X is a finite set of nodes, ω encodes node affinities, and μ is some choice of weights on the nodes (e.g., uniform). For example, weighted adjacency functions are used as network kernels in this framework to represent protein-protein interaction networks in [78], and heat kernels are used in [21] for the purpose of uncovering community structures in graph datasets.

2. A **hypergraph** is a set V of nodes and a set E of subsets of V, each of which is referred to as a **hyperedge**. The containment or incidence relation can be encoded as a function $\omega : V \times E \to \{0, 1\}$, so that picking probability measures μ and ν (say, uniform) yields a measure hypernetwork (V, μ, E, ν, ω). This representation was used in [22], where it was applied, for example, to simplify hypergraphs representing complicated social interactions. The notion of a measure hypernetwork therefore generalizes the notion of a hypergraph. One also obtains a measure hypernetwork from a **data matrix**, where X is an indexing set for samples, Y is an indexing set for features , $\omega(x, y)$ is the value of the matrix for sample x and feature y and μ and ν are some choices of weights. Applications to analysis of ensembles of data matrices was a main motivation for the introduction of this formalism in [55].

3. One can obtain an **augmented measure network** by taking (X, μ, Y, ν, ω) to be a measure hypernetwork representation of a data matrix, setting $\omega_{XY} = \omega$ and taking ω_X to be some relational function on the rows, such as distance between samples in the data space. This approach was taken to model multi-omics data in [26], with a view toward integrating several single-cell multi-omics datasets.

Remark 2.3 (Representations as graph structures)**.** Figure 1 provides schematic representations of the generalized networks defined above. To interpret this, one should conceptualize a measure network as a fully connected, weighted graph with a node set X and edge weights encoded by ω. From this perspective, it is natural to consider a measure hypernetwork as a bipartite weighted graph, where X and Y define the two classes of nodes. Then ω defines edge weights for the complete bipartite graph, while there are no edges joining pairs of nodes in X or pairs of hyperedges in Y. Finally, an augmented measure network can be viewed similarly, with the difference being that edges between nodes in X are permitted with weights encoded by ω_X (and bipartite edge weights being encoded by ω_{XY}). These interpretations lead to the k-partitioned measure network formalism introduced in Section 2.2. See Definition 2.8 and Remark 2.10 for a more formal justification of these representations.

Remark 2.4 (Finite and continuous spaces)**.** The concept of measure hypernetwork was introduced in [55], somewhat less formally, and primarily as a model for data matrices (so X and Y were assumed to be finite). This definition was formalized and extended to the setting of infinite spaces in [22] (in a slightly less general form than what is presented here, as network kernels were assumed therein to be bounded and the underlying Polish spaces were assumed to be compact for many results). The computational examples of practical interest are, of course, always defined over finite spaces; the main motivation for extending the definition to infinite spaces is to allow us to consider the collection of all measure hypernetworks as a complete metric space with respect to the distance defined below. Similarly, the notion of an augmented measure network was introduced for finite spaces in [26], and the more formal definition provided above is novel.

2.1.2 Generalized network distances

For each flavour of generalized network described in Definition 2.1, there has been an associated notion of distance introduced in the literature. They all have a similar structure, defined in terms of optimizing over measure couplings, as in the Kantorovich formulation of classical optimal transport (see [73, 51] as general references on optimal transport). Below, we use $\mathrm{proj}_X : X \times X' \to X$ and $\mathrm{proj}_{X'} : X \times X' \to X'$ to denote coordinate projections on some product sets, and for a Borel measurable map $p : X \to Y$ of topological spaces, we use $p_\# \mu$ to denote the pushforward to Y of a Borel measure μ on X.

Definition 2.5 (Measure coupling)**.** *For probability spaces (X, μ) and (X', μ'), we say that a measure π on $X \times X'$ is a **coupling** of μ and μ' if its left and right marginals are equal to μ and μ', respectively; that is, $(\mathrm{proj}_X)_\# \pi = \mu$ and $(\mathrm{proj}_{X'})_\# \pi = \mu'$. Let $\Pi(\mu, \mu')$ denote the set of all couplings of μ and μ'.*

Let us now establish some convenient notational conventions that will be used throughout the rest of the paper. We always use N and N' to stand for measure networks, with the underlying data always being implicitly denoted $N = (X, \mu, \omega)$ and $N' = (X', \mu', \omega')$. We similarly use H, H' for measure hypernetworks and A, A' for augmented hypernetworks. That is, when referring to H, the underlying data is implicitly given by $H = (X, \mu, Y, \nu, \omega)$, unless explicitly stated otherwise. For functions $\tau : A \times B \to \mathbb{R}$ and $\tau' : A' \times B' \to \mathbb{R}$, the difference $\tau - \tau'$ is understood to be the function defined on $A \times A' \times B \times B'$ as follows:

$$\begin{aligned} \tau - \tau' : A \times A' \times B \times B' &\to \mathbb{R} \\ (a, a', b, b') &\mapsto \tau(a, b) - \tau'(a', b'). \end{aligned}$$

Given a measure space (X, μ), we use $\|\cdot\|_{L^p(\mu)}$ to denote the standard norm on $L^p(\mu)$. With these conventions in mind, we recall some metrics which have been introduced on the generalized network spaces of Definition 2.1.

Definition 2.6 (Generalized network distances)**.** *Let $p \in [1, \infty]$.*

1. *The **network p-distance** or **Gromov-Wasserstein p-distance** [19] between measure networks $N, N' \in \mathcal{N}^p$ is*

$$d_{\mathcal{N}^p}(N, N') := \inf_{\pi \in \Pi(\mu, \mu')} \frac{1}{2} \|\omega - \omega'\|_{L^p(\pi \otimes \pi)}.$$

2. *The **hypernetwork p-distance** [55, 22] between measure hypernetworks $H, H' \in \mathcal{H}^p$ is*

$$d_{\mathcal{H}^p}(H, H') := \inf_{\substack{\pi \in \Pi(\mu, \mu') \\ \xi \in \Pi(\nu, \nu')}} \frac{1}{2} \|\omega - \omega'\|_{L^p(\pi \otimes \xi)}.$$

3. *The **augmented network** p-**distance** [26] between augmented measure networks $A, A' \in \mathcal{A}^p$ is*

$$d_{\mathcal{A}^p}(A, A') := \inf_{\substack{\pi \in \Pi(\mu,\mu') \\ \xi \in \Pi(\nu,\nu')}} \frac{1}{2} \left(\|\omega_X - \omega'_{X'}\|^p_{L^p(\pi\otimes\pi)} + \|\omega_{XY} - \omega'_{X'Y'}\|^p_{L^p(\pi\otimes\xi)} \right)^{1/p}.$$

In the following section, we unify these generalized network concepts and distances (as well as others) under a common framework. We will use this common framework to derive various metric properties of these distances simultaneously.

2.2 Partitioned measure networks and generalized networks

One should observe the similarities between the various notions of generalized network in Definition 2.1, and the optimal transport-inspired distances between them described in Definition 2.6. In this section, we describe a new structure that simultaneously generalizes these ideas.

2.2.1 Partitioned measure networks

Let us now introduce a new generalized network structure.

Definition 2.7 (Partitioned measure network)**.** *Let k be a positive integer and $p \in [1,\infty]$. A k-**partitioned measure** p-**network** is a structure of the form $P = \big((X_i,\mu_i)_{i=1}^k, \omega\big)$, where*

- *each (X_i,μ_i) is a Polish probability space such that $X_i \cap X_j = \varnothing$ for $i \neq j$, and*
- *using the notation $X := \sqcup_i X_i$ and $\mu := \sum_i \mu_i$, $\omega : X \times X \to [0,\infty)$ is an element of $L^p(\mu\otimes\mu)$.*

To simplify notation, we sometimes write (X_i,μ_i) instead of $(X_i,\mu_i)_{i=1}^k$, (X_i) instead of $(X_i)_{i=1}^k$ and (μ_i) instead of $(\mu_i)_{i=1}^k$. We use $\mathcal{P}_k^p$ to denote the collection of all k-partitioned measure p-networks.

It is often the case that the particular values of k and p are not important, in which case we abuse terminology and refer to the objects defined above as **partitioned measure networks**. In line with those established above, we follow the notational convention that P and P' are implicitly assumed to stand for partitioned measure networks $P = ((X_i,\mu_i),\omega)$ and $P' = ((X'_i,\mu'_i),\omega')$. Observe that a 1-partitioned measure network is just a measure network; that is, $\mathcal{N}^p = \mathcal{P}_1^p$. Next, we observe below that we could embed the various notions of generalized networks (from Definition 2.1) into the space of partitioned measure networks $\mathcal{P}_2^p$.

Definition 2.8 (Generalized network embeddings)**.** *We have the following families of embeddings.*

1. *For each k, let $\varepsilon_{k,k+1} : \mathcal{P}_k^p \to \mathcal{P}_{k+1}^p$ be the map taking $P \in \mathcal{P}_k^p$ to*

$$\varepsilon_{k,k+1}(P) := \left((X_i,\mu_i)_{i=1}^{k+1}, \varepsilon_{k,k+1}(\omega)\right),$$

where X_{k+1} consists of a single abstract point, μ_{k+1} is the associated Dirac measure, and the network kernel is defined by

$$\varepsilon_{k,k+1}(\omega) : (\sqcup_{i=1}^{k+1} X_i)\times(\sqcup_{i=1}^{k+1} X_i) \to \mathbb{R}, \qquad \varepsilon_{k,k+1}(\omega)(u,v) = \begin{cases} \omega(u,v) & (u,v) \in (\sqcup_{i=1}^k X_i)\times(\sqcup_{i=1}^k X_i); \\ 0 & \textit{otherwise.} \end{cases}$$

In particular, $\varepsilon_{1,2}$ gives an embedding $\mathcal{N}^p \hookrightarrow \mathcal{P}_2^p$. For $k < \ell$, we define $\varepsilon_{k,\ell} : \mathcal{P}_k^p \to \mathcal{P}_\ell^p$ by

$$\varepsilon_{k,\ell} := \varepsilon_{\ell-1,\ell} \circ \cdots \circ \varepsilon_{k,k+1}.$$

2. *Let $\varepsilon_{\mathcal{H}} : \mathcal{H}^p \to \mathcal{P}_2^p$ be the map taking $H \in \mathcal{H}^p$ to*

$$\varepsilon_{\mathcal{H}}(H) := \left(((X,\mu),(Y,\nu)), \varepsilon_{\mathcal{H}}(\omega)\right),$$

where

$$\varepsilon_{\mathcal{H}}(\omega)(u,v) := \begin{cases} \omega(u,v) & u \in X \textit{ and } v \in Y; \\ 0 & \textit{otherwise.} \end{cases}$$

3. Let $\varepsilon_{\mathcal{A}} : \mathcal{A}^p \to \mathcal{P}_2^p$ be the map taking $A \in \mathcal{A}^p$ to

$$\varepsilon_{\mathcal{A}}(A) := (((X,\mu),(Y,\nu)), \varepsilon_{\mathcal{A}}(\omega_X, \omega_{XY}))\,,$$

where

$$\varepsilon_{\mathcal{A}}(\omega_X, \omega_{XY})(u,v) := \begin{cases} \omega_X(u,v) & u, v \in X; \\ \omega_{XY}(u,v) & u \in X \text{ and } v \in Y; \\ 0 & \text{otherwise.} \end{cases}$$

Example 2.9 (Partitioned networks)**.** The generalized network embeddings defined above show that measure networks, hypernetworks and augmented hypernetworks can be considered as 2-partitioned networks, so that this structure encompasses those described in Example 2.2. Another source of examples of k-partitioned networks is the notion of a dataset with a categorical class structure. That is, consider a (say, finite) dataset of points X such that each $x \in X$ belongs to one of k different classes—for example, X could consist of a set of images, and the images could be assigned classes based on subject matter (e.g., cats, dogs, etc.). This class structure can be encoded as probability measures by taking μ_i to be a uniform measure supported on those points belonging to class i. The supports then define the required partition of $X = \sqcup_i X_i$ and any choice of network kernel ω on X gives a representation of the multiclass dataset as a k-partitioned measure network.

Remark 2.10 (Intuition for the embeddings)**.** Remark 2.3 gives interpretations of generalized networks in terms of graph structures, illustrated schematically in Figure 1. The embeddings defined above formalize these intuitive descriptions mathematically. For example, if $H \in \mathcal{H}^p$ is a representation of a hypergraph, then $\varepsilon_{\mathcal{H}}(H)$ gives a representation of H as a bipartite graph.

2.2.2 Partitioned network distance

We now define a distance between partitioned measure networks, using the concept of a partitioned coupling.

Definition 2.11 (Partitioned coupling)**.** *Given k-tuples of probability spaces $(X_i, \mu_i)_{i=1}^k$ and $(X_i', \mu_i')_{i=1}^k$, let*

$$\Pi_k\left((\mu_i), (\mu_i')\right) := \Pi(\mu_1, \mu_1') \times \cdots \times \Pi(\mu_k, \mu_k').$$

*An element $(\pi_i)_{i=1}^k$ of $\Pi_k\left((\mu_i), (\mu_i')\right)$ is a k-**partitioned coupling**. To simplify notation, we sometimes denote k-partitioned couplings as (π_i) instead of $(\pi_i)_{i=1}^k$.*

Definition 2.12 (Partitioned network distance)**.** *For $p \in [0,\infty)$, the **partitioned network p-distance** between partitioned measure networks $P, P' \in \mathcal{P}_k^p$ is*

$$d_{\mathcal{P}_k^p}(P,P') := \inf_{(\pi_i) \in \Pi_k((\mu_i),(\mu_i'))} \frac{1}{2}\left(\sum_{i,j=1}^k \|\omega - \omega'\|_{L^p(\pi_i \otimes \pi_j)}^p\right)^{1/p}. \tag{1}$$

For $p = \infty$, we define

$$d_{\mathcal{P}_k^\infty}(P,P') := \inf_{(\pi_i) \in \Pi_k((\mu_i),(\mu_i'))} \frac{1}{2} \max_{i,j} \|\omega - \omega'\|_{L^\infty(\pi_i \otimes \pi_j)}. \tag{2}$$

Remark 2.13. In the above definition and throughout the rest of the paper, we slightly abuse terminology and consider each π_i as a probability measure on $X \times X' = (\sqcup_j X_j) \times (\sqcup_j X_j')$ (which is supported on the subset $X_i \times X_i'$).

Remark 2.14 (Connection to labelled Gromov-Wasserstein distance [57])**.** A notion of Gromov-Wasserstein distance is introduced which is essentially equivalent to our partitioned network distance [57]. However, the treatment in that paper is very much from a computational perspective, and a formal, a general definition of the distance is not provided. The usefulness of such a metric is demonstrated by applications to cross-modality matching for biological data. We remark that [57] uses the terminology **labelled Gromov-Wasserstein distance**; we use the term **labelled** differently below, in Section 3.

When $k = 1$, $d_{\mathcal{P}_1^p}$ is simply the Gromov-Wasserstein distance $d_{\mathcal{N}^p}$. We will show below that, for arbitrary k, $d_{\mathcal{P}_k^p}$ induces a metric on $\mathcal{P}_k^p$ modulo a natural notion of equivalence, which generalizes the known result in the Gromov-Wasserstein case. Before doing so, we show that the embeddings of generalized networks of Definition 2.8 preserve the notions of distance that we have defined so far.

Theorem 1. *The maps from Definition 2.8 preserve generalized network distances:*

1. *for $k < \ell$, $d_{\mathcal{P}_\ell^p}(\varepsilon_{k,\ell}(P), \varepsilon_{k,\ell}(P')) = d_{\mathcal{P}_k^p}(P, P')$; in particular, $d_{\mathcal{P}_2^p}(\varepsilon_{1,2}(N), \varepsilon_{1,2}(N')) = d_{\mathcal{N}^p}(N, N')$;*
2. *$d_{\mathcal{P}_2^p}(\varepsilon_{\mathcal{H}}(H), \varepsilon_{\mathcal{H}}(H')) = d_{\mathcal{H}^p}(H, H')$;*
3. *and $d_{\mathcal{P}_2^p}(\varepsilon_{\mathcal{A}}(A), \varepsilon_{\mathcal{A}}(A')) = d_{\mathcal{A}^p}(A, A')$.*

Proof. We provide details for the $p < \infty$ case, with the proof for $p = \infty$ following by similar arguments. For the first claim, it suffices to consider the case where $\ell = k + 1$. Let $P, P' \in \mathcal{P}_k^p$. Any $(\pi_i) = (\pi_i)_{i=1}^k \in \Pi_k((\mu_i), (\mu_i'))$ extends uniquely to a $(k+1)$-partitioned coupling of $(\mu_1, \ldots, \mu_k, \mu_{k+1})$ and $(\mu_1', \ldots, \mu_k', \mu_{k+1}')$, namely, $(\pi_1, \ldots, \pi_k, \pi_{k+1})$, where π_{k+1} the Dirac mass on the singleton set $X_{k+1} \times X_{k+1}'$. We have that

$$\begin{aligned}\sum_{i,j=1}^{k+1} \|\varepsilon_{k,k+1}(\omega) - \varepsilon_{k,k+1}(\omega')\|_{L^p(\pi_i \otimes \pi_j)}^p &= \sum_{i=1}^{k+1} \|\varepsilon_{k,k+1}(\omega) - \varepsilon_{k,k+1}(\omega')\|_{L^p(\pi_i \otimes \pi_{k+1})}^p \\ &\quad + \sum_{j=1}^{k+1} \|\varepsilon_{k,k+1}(\omega) - \varepsilon_{k,k+1}(\omega')\|_{L^p(\pi_{k+1} \otimes \pi_j)}^p \\ &\quad + \sum_{i,j=1}^{k} \|\varepsilon_{k,k+1}(\omega) - \varepsilon_{k,k+1}(\omega')\|_{L^p(\pi_i \otimes \pi_j)}^p \\ &= \sum_{i,j=1}^{k} \|\omega - \omega'\|_{L^p(\pi_i \otimes \pi_j)}^p,\end{aligned}$$

where the last line follows because $\varepsilon_{k,k+1}(\omega) = \varepsilon_{k,k+1}(\omega') = 0$ on the supports of $\pi_i \otimes \pi_{k+1}$ and $\pi_{k+1} \otimes \pi_j$. And moreover, $\varepsilon_{k,k+1}(\omega) = \omega$ and $\varepsilon_{k,k+1}(\omega') = \omega'$ on the support of each $\pi_i \otimes \pi_j$ with $i, j < k$. Since the k-partitioned coupling (π_i) was arbitrary, the first claim follows.

Let us now prove the third claim; the case for hypernetworks is proved using similar arguments, so we omit the details here. Let $A, A' \in \mathcal{A}^p$ and let $\pi \in \Pi(\mu, \mu')$ and $\xi \in \Pi(\nu, \nu')$. Then $(\pi_1, \pi_2) = (\pi, \xi)$ is a 2-partitioned coupling of (μ, ν) and (μ', ν'), and we have

$$\sum_{i,j=1}^{2} \|\varepsilon_{\mathcal{A}}(\omega) - \varepsilon_{\mathcal{A}}(\omega')\|_{L^p(\pi_i \otimes \pi_j)}^p = \|\omega_X - \omega_{X'}'\|_{L^p(\pi \otimes \pi)}^p + \|\omega_{XY} - \omega_{X'Y'}'\|_{L^p(\pi \otimes \xi)}^p,$$

by reasoning similar to the above. Since π and ξ were arbitrary, this completes the proof. □

2.2.3 Metric properties of the partitioned network distance

To describe the exact sense in which $d_{\mathcal{P}_k^p}$ is a distance, we need to introduce some equivalence relations on partitioned measure networks.

Definition 2.15 (Strong and weak isomorphism)**.** *We say that k-partitioned measure networks P and P' are **strongly isomorphic** if, for each $i = 1, \ldots, k$, there is a Borel measurable bijection $\phi_i : X_i \to X_i'$ (with Borel measurable inverse) such that $(\phi_i)_\# \mu_i = \mu_i'$, and $\omega(x, y) = \omega'(\phi_i(x), \phi_j(y))$ for every pair $(x, y) \in X_i \times X_j$. The tuple $(\phi_i) = (\phi_i)_{i=1}^k$ is called a **strong isomorphism**.*

*A tuple (ϕ_i) of maps $\phi_i : X_i \to X_i'$ is called a **weak isomorphism** from P to P' if $(\phi_i)_\# \mu_i = \mu_i'$, and $\omega(x, y) = \omega'(\phi_i(x), \phi_j(y))$ for $\mu_i \otimes \mu_j$-almost every pair $(x, y) \in X_i \times X_j$. We do not require each ϕ_i to be a bijection in this definition.*

*We say that P and P' are **weakly isomorphic** if there exists a third partitioned measure network $\overline{P} = ((\overline{X}_i, \overline{\mu}_i), \overline{\omega})$ and weak isomorphisms (ϕ_i) from $\overline{P}$ to P and (ϕ_i') from $\overline{P}$ to P'.*

It is straightforward to show that weak isomorphism defines an equivalence relation on $\mathcal{P}$; we use $P \sim P'$ to denote that P is weakly isomorphic to P'. For $P \in \mathcal{P}_k^p$, let $[P]$ denote its equivalence class under this relation and let $[\mathcal{P}_k^p]$ denote the collection of all equivalence classes.

Using the embeddings from Definition 2.8, there is an induced notion of weak isomorphism on the spaces of generalized networks. Let $[\mathcal{N}^p]$, $[\mathcal{H}^p]$ and $[\mathcal{A}^p]$ denote the collections of weak isomorphism equivalence classes of measure networks, measure hypernetworks and augmented measure networks, respectively.

Weak isomorphisms of measure networks and of measure hypernetworks are introduced in [19] and [22], respectively. It is straightforward to show that the induced notions from Definition 2.15 agree with those already established in the literature. The aforementioned papers show that $d_{\mathcal{N}^p}$ and $d_{\mathcal{H}^p}$ descend to well-defined metrics on $[\mathcal{N}^p]$ and $[\mathcal{H}^p]$, respectively. The following theorem generalizes these results, in light of Theorem 1.

Theorem 2. *The k-partitioned network p-distance $d_{\mathcal{P}_k^p}$ induces a well-defined metric on $[\mathcal{P}_k^p]$.*

Putting Theorems 1 and 2 together, we obtain the following corollary.

Corollary 2.16. *The generalized network distances $d_{\mathcal{N}^p}$, $d_{\mathcal{H}^p}$ and $d_{\mathcal{A}^p}$ induce well-defined metrics on $[\mathcal{N}^p]$, $[\mathcal{H}^p]$ and $[\mathcal{A}^p]$, respectively. The embeddings from Definition 2.8 induce isometric embeddings of each of these spaces into $[\mathcal{P}_2^p]$.*

We will abuse notation and continue to denote the induced metric on $[\mathcal{P}_k^p]$ as $d_{\mathcal{P}_k^p}$, and take a similar convention for the other induced metrics.

Remark 2.17. The case of $\mathcal{N}^p$ in the corollary was proved in [19] and the case of $\mathcal{H}^p$ was proved in [22] (those papers assumed boundedness of the ω-functions, but this restriction is easily lifted in those proofs). In [26], a relaxed version of triangle inequality was proved for the augmented network distance in the finite setting. Corollary 2.16 strengthens [26, Proposition 1] to show that $d_{\mathcal{A}^p}$ satisfies the true (non-relaxed) triangle inequality.

Theorem 2 follows as an easy corollary of a more general result in the following section, so we defer its proof to Section 3.1.2.

2.2.4 Geodesics and curvature

We can say more about the metric properties of the partitioned network distance. To state our next result, we first recall some standard concepts from metric geometry; see [10, 13] as general references.

Definition 2.18 (Geodesics and curvature)**.** *Let (X, d) be a metric space.*

1. *A* ***geodesic*** *between points $x, y \in X$ is a path $\gamma : [0,1] \to X$ with $\gamma(0) = x$, $\gamma(1) = y$ and such that, for all $0 \leqslant s \leqslant t \leqslant 1$, we have*
$$d(\gamma(s), \gamma(t)) = (t - s)d(x, y).$$
If there is a geodesic joining any two points in X, we say that (X, d) is a ***geodesic space****.*

2. *Suppose that (X, d) is a geodesic space. We say that (X, d) has* ***curvature bounded below by zero*** *if for every geodesic $\gamma : [0,1] \to X$ and every point $x \in X$, we have*
$$d(\gamma(t), x)^2 \geqslant (1-t)d(\gamma(0), x)^2 + td(\gamma(1), x)^2 - t(1-t)d(\gamma(0), \gamma(1))^2$$
for all $t \in [0,1]$.

3. *We say that (X, d) is an* ***Alexandrov space of non-negative curvature*** *if it is a complete geodesic space with curvature bounded below by zero.*

The next main result follows from a more general result below. We defer its proof to Section 3.2.3.

Theorem 3. *For any $k \geqslant 1$, $([\mathcal{P}_k^2], d_{\mathcal{P}_k^2})$ is an Alexandrov space of non-negative curvature.*

Combining Theorem 3 with Corollary 2.16 immediately yields the following result.

Corollary 2.19. *Each of the spaces* $([\mathcal{N}^2], d_{\mathcal{N}^2})$*,* $([\mathcal{H}^2], d_{\mathcal{H}^2})$ *and* $([\mathcal{A}^2], d_{\mathcal{A}^2})$ *is an Alexandrov space of non-negative curvature.*

Remark 2.20 (Prior curvature results)**.** The fact that $([\mathcal{N}^2], d_{\mathcal{N}^2})$ is an Alexandrov space of non-negative curvature was essentially proved by Sturm in [65, Theorem 5.8] (there the space of *symmetric measure networks* was considered, i.e., where the function $\omega : X \times X \to \mathbb{R}$ is assumed to be symmetric; the proof still works if this assumption is dropped, as was observed in [20]). This result is new for the other spaces of generalized networks in Corollary 2.19.

3 Extension to labelled networks

We consider, as an extension to the discussion so far, the setting of k-partitioned measure networks where each element of X_i $(1 \leqslant i \leqslant k)$ is associated with a label element that lives in a metric space.

3.1 Labelled partitioned measure networks

Let us begin by defining a labelled notion of a k-partitioned measure network, and the distances between these objects.

Definition 3.1 (Labelled k-partitioned measure networks)**.** *Let* $(\Lambda_i, d_{\Lambda_i})$ *$(1 \leqslant i \leqslant k)$ be fixed metric spaces, which we consider as spaces of* ***labels****. A* ***labelled** k**-partitioned measure** p**-network*** *is a tuple* $\big(P, (\iota_i)_{i=1}^k\big)$*, where*

- $P = \big((X_i, \mu_i)_{i=1}^k, \omega\big) \in \mathcal{P}_k^p$;
- *each of the* $\iota_i : X_i \to \Lambda_i$ *is a measurable function, which we refer to as a* ***labelling function****;*
- *each function* $X_i \times X_i \to \mathbb{R}$ *defined by* $(x, y) \mapsto d_{\Lambda_i}(\iota_i(x), \iota_i(y))$ *belongs to* $L^p(\mu_i \otimes \mu_i)$.

We frequently simplify notation and write $L = (P, (\iota_i))$ *for a labelled k-partitioned network. We denote by* $\mathcal{LP}_k^p$ *the space of labelled k-partitioned measure p-networks, where it is understood that the label spaces* $(\Lambda_i, d_{\Lambda_i})$ *are fixed.*

Example 3.2 (Node-attributed networks)**.** The main examples of labelled partitioned networks come from **node-attributed network** structures. For example, consider a measure network $N = (X, \mu, \omega) \in \mathcal{N}^p$ representing a graph via some graph kernel ω. In applications, the node set X may be attributed with auxiliary data—for example, if the graph encodes user interactions on a social network, then each node may be attributed with additional user-level statistics. This situation can be modelled as a function $\iota : X \to \Lambda$, where Λ is the attribute space (e.g., $\Lambda = \mathbb{R}^n$). The structure (N, ι) defines an element of $\mathcal{LP}_1^p$.

The partitioned network distance (2.12) can be naturally generalized to $\mathcal{LP}_k^p$.

Definition 3.3 (Labelled partitioned network distance)**.** *Let* $1 \leqslant p < \infty$ *and let* $L = (P, (\iota_i))\,, L' = (P', (\iota_i')) \in \mathcal{LP}_k^p$ *be two labelled partitioned measure networks. Then we define the* ***labelled** k**-partitioned network distance*** *to be*

$$d_{\mathcal{LP}_k^p}(L, L') = \inf_{(\pi_i) \in \Pi_k\big((\mu_i), (\mu_i')\big)} \frac{1}{2} \left(\sum_{i,j=1}^k \|\omega - \omega'\|_{L^p(\pi_i \otimes \pi_j)}^p + \sum_{i=1}^k \|d_{\Lambda_i} \circ (\iota_i, \iota_i')\|_{L^p(\pi_i)}^p \right)^{1/p}. \tag{3}$$

This extends to the $p = \infty$ *case as*

$$d_{\mathcal{LP}_k^\infty}(L, L') = \inf_{(\pi_i) \in \Pi_k\big((\mu_i), (\mu_i')\big)} \frac{1}{2} \max\left(\max_{i,j} \|\omega - \omega'\|_{L^\infty(\pi_i \otimes \pi_j)}, \max_i \|d_{\Lambda_i} \circ (\iota_i, \iota_i')\|_{L^\infty(\pi_i)} \right). \tag{4}$$

Remark 3.4. We could include a balance parameter to weight the contributions of the network kernel term (i.e., the first summation) versus the labelling function term (the second summation) in (3). Such a parameter is included in the definition of *Fused Gromov-Wasserstein (FGW) distance* [71], which has a similar structure. The connection between $d_{\mathcal{LP}_k^p}$ and FGW distance is explained precisely in Section 3.1.2. We avoid the inclusion of the balance parameter in our formulation, as it is unimportant from a theoretical standpoint and can been absorbed into the definitions of network kernels and label functions in practical applications.

A simple (but useful) observation is that the distances can be written as nested ℓ^p-norms[1]. For the rest of the paper, let $\|\cdot\|_p$ denote the ℓ^p-norm on $\mathbb{R}^n$ for $p \in [1,\infty]$. We abuse notation and use the same symbol $\|\cdot\|_p$ for the norm on spaces of various dimensions, with the specific meaning always being clear from context. Then the labelled k-partitioned network distance can be expressed as

$$d_{\mathcal{LP}_k^p}(L,L') = \inf_{(\pi_i)\in\Pi_k((\mu_i),(\mu_i'))} \frac{1}{2}\left\|\left(\left\|\left(\|\omega-\omega'\|_{L^p(\pi_i\otimes\pi_j)}\right)_{i,j}\right\|_p, \left\|\left(\|d_{\Lambda_i}\circ(\iota_i,\iota_i')\|_{L^p(\pi_i)}\right)_i\right\|_p\right)\right\|_p, \tag{5}$$

for all $p \in [1,\infty]$. We have made one more abuse of notation by considering the collection $\left(\|\omega-\omega'\|_{L^p(\pi_i\otimes\pi_j)}\right)_{i,j}$, which is most naturally indexed as a $k \times k$ matrix, as an element of $\mathbb{R}^{k^2}$ in order to apply the ℓ^p-norm to it.

3.1.1 Metric properties of the labelled distance

We now show that $d_{\mathcal{LP}_k^p}$ defines a metric, up to a natural notion of equivalence. Strong and weak isomorphisms of partitioned networks (Definition 2.15) extend to the case of labelled partitioned measure networks in a straightforward way.

Definition 3.5 (Weak isomorphism of labelled partitioned measure networks)**.** *We say that labelled k-partitioned measure networks $L = (P,(\iota_i))$ and $L' = (P',(\iota_i'))$ are **strongly isomorphic** if the underlying partitioned measure networks P and P' are strongly isomorphic (see Definition 2.15) via bijections $\phi_i : X_i \to X_i'$ such that $\iota_i(x) = \iota_i'(\phi(x))$ for μ_i-almost every $x \in X_i$.*

*We say that L and L' are **weakly isomorphic** if there exists $\overline{L} = (\overline{P},(\overline{\iota}_i)) \in \mathcal{LP}_k$, with $\overline{P} = ((\overline{X}_i,\overline{\mu}_i),\overline{\omega})$, such that*

- *there exist weak isomorphisms (ϕ_i) and (ϕ_i') from $\overline{P}$ to P and P', respectively; that is, $\phi_i : \overline{X}_i \to X_i$ and $\phi_i' : \overline{X}_i \to X_i'$ satisfy the conditions given in Definition 2.15;*
- *and the maps ϕ_i and ϕ_i' additionally satisfy*

$$\overline{\iota}_i(x) = \iota_i(\phi(x)) = \iota_i'(\phi'(x)),$$

for $\overline{\mu}_i$-almost every $x \in \overline{X}_i$.

One can easily verify that weak isomorphism again defines an equivalence relation on $\mathcal{LP}_k^p$, and we write $[P,(\iota_i)]$ for equivalence classes and $[\mathcal{LP}_k^p]$ for the quotient space.

The next theorem is analogous to Theorem 2, which establishes the metric properties of $d_{\mathcal{P}_k^p}$. In fact, the deferred proof of Theorem 2 will follow easily from this result (see Section 3.1.2).

Theorem 4. *The labelled k-partitioned network p-distance $d_{\mathcal{LP}_k^p}$ induces a well-defined metric on $[\mathcal{LP}_k^p]$.*

The proof will use some important technical lemmas.

Lemma 3.6. *The infima in (3) and (4) are always realized by partitioned couplings.*

Proof. Let $(P,(\iota_i)),(P',(\iota_i')) \in \mathcal{LP}_k^p$. We have $\Pi_k((\mu_i),(\mu_i')) = \Pi(\mu_1,\mu_1') \times \cdots \times \Pi(\mu_k,\mu_k')$. By [65, Lemma 1.2], each $\Pi(\mu_i,\mu_i')$ is compact (as a subspace of the space of probability measures on $X_i \times X_i'$, with the weak topology), so it follows that $\Pi_k((\mu_i),(\mu_i'))$ is compact as well. By the proof of [22, Lemma 24], for each (i,j), the function

$$\Pi(\mu_i,\mu_i') \times \Pi(\mu_j,\mu_j') \to \mathbb{R} : (\pi_i,\pi_j) \mapsto \|\omega-\omega'\|_{L^p(\pi_i\otimes\pi_j)}$$

[1] In this paper, L^p is the norm defined in terms of a measure, whereas ℓ^p is the standard norm on $\mathbb{R}^n$, which does not depend on any measure.

is continuous in the $p < \infty$ case and lower semicontinuous in the $p = \infty$ case. Similarly, the function $\Pi(\mu_i, \mu_i') \to \mathbb{R} : \pi_i \mapsto \|d_{\Lambda_i} \circ (\iota_i, \iota_i')\|_{L^p(\pi_i)}$ is continuous (respectively, lower semicontinuous) if $p < \infty$ (respectively, $p = \infty$). It is then straightforward to see that the objectives of (3) and (4) inherit these properties as functions on $\Pi_k((\mu_i), (\mu_i'))$. In either case, it follows from compactness that the infima are achieved. □

The proofs of Theorem 4 and results later in the paper will use the following standard result from optimal transport theory. In the statement, and throughout the paper, we use $\mathrm{proj}_i : Y^0 \times Y^1 \times \cdots \times Y^n \to Y^i$ to denote the coordinate projection map from a product of sets to its ith factor.

Lemma 3.7 (Gluing Lemma; see, e.g., Lemma 1.4 of [65])**.** *Let (Y_i, ν_i) be Polish probability spaces (for $i = 0, \ldots, n$). For a collection of measure couplings $\xi_i \in \Pi(\nu_{i-1}, \nu_i)$, $i = 1, \ldots, n$, there is a unique probability measure $\tilde{\xi}$ on $Y_0 \times Y_1 \times \cdots \times Y_n$ with the property that*

$$\left(\mathrm{proj}_{i-1} \times \mathrm{proj}_i\right)_{\#} \tilde{\xi} = \xi_i,$$

for all $i = 1, \ldots, n$.

The measure $\tilde{\xi}$ from the lemma is called the **gluing** of the measures ξ_i. We denote it as $\tilde{\xi} = \xi_1 \boxtimes \xi_2 \boxtimes \cdots \boxtimes \xi_n$.

Proof of Theorem 4. The function $d_{\mathcal{LP}_k^p}$ is clearly symmetric. We now establish the triangle inequality. Let $L = (P, (\iota_i)), L' = (P', (\iota_i')), L'' = (P'', (\iota_i'')) \in \mathcal{LP}_k^p$. By Lemma 3.6 there exist partitioned couplings $(\pi_i) \in \Pi_k((\mu_i), (\mu_i'))$ and $(\pi_i') \in \Pi_k((\mu_i'), (\mu_i''))$ that realize the infima in the distances, respectively, between P, P' and P', P''. Let $\widetilde{\xi_i} = \pi_i \boxtimes \pi_i' \boxtimes \pi_i''$ denote the probability measure on $X_i \times X_i' \times X_i''$, for $i = 1, \ldots, k$, obtained from the Gluing Lemma (Lemma 3.7). Letting ξ_i denote the pushforward to $\widetilde{\xi_i}$ to $X_i \times X_i''$, we have that $(\xi_i) \in \Pi_k((\mu_i), (\mu_i''))$. Using the expression (5), we get the desired triangle inequality from the triangle inequality for the L^p- and ℓ^p-norms and for d_{Λ_i}:

$$\begin{aligned}
&2 \cdot d_{\mathcal{LP}_k^p}(L, L'') \\
&\leqslant \left\| \left(\left\| \left(\|\omega - \omega''\|_{L^p(\xi_i \otimes \xi_j)} \right)_{i,j} \right\|_p, \left\| \left(\|d_{\Lambda_i} \circ (\iota_i, \iota_i'')\|_{L^p(\xi_i)} \right)_i \right\|_p \right) \right\|_p \\
&= \left\| \left(\left\| \left(\|\omega - \omega''\|_{L^p(\tilde{\xi}_i \otimes \tilde{\xi}_j)} \right)_{i,j} \right\|_p, \left\| \left(\|d_{\Lambda_i} \circ (\iota_i, \iota_i'')\|_{L^p(\tilde{\xi}_i)} \right)_i \right\|_p \right) \right\|_p \\
&\leqslant \left\| \left(\left\| \left(\|\omega - \omega' + \omega' - \omega''\|_{L^p(\tilde{\xi}_i \otimes \tilde{\xi}_j)} \right)_{i,j} \right\|_p, \left\| \left(\|d_{\Lambda_i} \circ (\iota_i, \iota_i') + d_{\Lambda_i} \circ (\iota_i', \iota_i'')\|_{L^p(\tilde{\xi}_i)} \right)_i \right\|_p \right) \right\|_p \\
&\leqslant \left\| \left(\left\| \left(\|\omega - \omega'\|_{L^p(\tilde{\xi}_i \otimes \tilde{\xi}_j)} + \|\omega' - \omega''\|_{L^p(\tilde{\xi}_i \otimes \tilde{\xi}_j)} \right)_{i,j} \right\|_p, \left\| \left(\|d_{\Lambda_i} \circ (\iota_i, \iota_i')\|_{L^p(\tilde{\xi}_i)} + \|d_{\Lambda_i} \circ (\iota_i', \iota_i'')\|_{L^p(\tilde{\xi}_i)} \right)_i \right\|_p \right) \right\|_p \\
&\leqslant \left\| \left(\left\| \left(\|\omega - \omega'\|_{L^p(\tilde{\xi}_i \otimes \tilde{\xi}_j)} \right)_{i,j} \right\|_p + \left\| \left(\|\omega' - \omega''\|_{L^p(\tilde{\xi}_i \otimes \tilde{\xi}_j)} \right)_{i,j} \right\|_p, \right. \right. \\
&\qquad\qquad\qquad\qquad \left. \left. \left\| \left(\|d_{\Lambda_i} \circ (\iota_i, \iota_i')\|_{L^p(\tilde{\xi}_i)} \right)_i \right\|_p \right\| + \left(\|d_{\Lambda_i} \circ (\iota_i', \iota_i'')\|_{L^p(\tilde{\xi}_i)} \right)_i \right\|_p \right) \right\|_p \\
&\leqslant \left\| \left(\left\| \left(\|\omega - \omega'\|_{L^p(\tilde{\xi}_i \otimes \tilde{\xi}_j)} \right)_{i,j} \right\|_p, \left\| \left(\|d_{\Lambda_i} \circ (\iota_i, \iota_i')\|_{L^p(\tilde{\xi}_i)} \right)_i \right\|_p \right) \right\|_p \\
&\qquad\qquad\qquad\qquad + \left\| \left(\left\| \left(\|\omega' - \omega''\|_{L^p(\tilde{\xi}_i \otimes \tilde{\xi}_j)} \right)_{i,j} \right\|_p, \left\| \left(\|d_{\Lambda_i} \circ (\iota_i', \iota_i'')\|_{L^p(\tilde{\xi}_i)} \right)_i \right\|_p \right) \right\|_p \\
&= \left\| \left(\left\| \left(\|\omega - \omega'\|_{L^p(\tilde{\pi}_i \otimes \tilde{\pi}_j)} \right)_{i,j} \right\|_p, \left\| \left(\|d_{\Lambda_i} \circ (\iota_i, \iota_i')\|_{L^p(\tilde{\pi}_i)} \right)_i \right\|_p \right) \right\|_p
\end{aligned}$$

$$+ \left\| \left(\left\| \left(\|\omega' - \omega''\|_{L^p(\tilde{\pi'}_i \otimes \tilde{\pi'}_j)} \right)_{i,j} \right\|_p , \left\| \left(\|d_{\Lambda_i} \circ (\iota'_i, \iota''_i)\|_{L^p(\tilde{\pi'}_i)} \right)_i \right\|_p \right) \right\|_p$$

$$= 2 \cdot d_{\mathcal{LP}_k^p}(L, L') + 2 \cdot d_{\mathcal{LP}_k^p}(L', L'').$$

We note that lines where measures are changed in the L^p-norms follow by marginalization (for example, the first equality which exchanges ξ_i for $\tilde{\xi}_i$ and ξ_j for $\tilde{\xi}_j$). This proves that the triangle inequality holds.

Finally, let us show that $d_{\mathcal{LP}_k^p}(L, L') = 0$ if and only if L and L' are weakly isomorphic. Suppose that L and L' are weakly isomorphic. Let $\overline{L} \in \mathcal{LP}_k$ denote the auxiliary space from the definition of weak isomorphism. It is easy to show that $d_{\mathcal{LP}_k^p}(\overline{L}, L) = d_{\mathcal{LP}_k^p}(\overline{L}, L') = 0$, so $d_{\mathcal{LP}_k^p}(L, L') = 0$ follows by symmetry and the triangle inequality. Conversely, suppose that $d_{\mathcal{LP}_k^p}(L, L') = 0$. By Lemma 3.6, there is a partitioned coupling (π_i) such that $\|\omega - \omega'\|_{L^p(\pi_i \otimes \pi_j)} = \|d_{\Lambda_i} \circ (\iota_i, \iota'_i)\|_{L^p(\pi_i)} = 0$ for all $i, j = 1, \ldots, k$. Define $\overline{X}_i = X_i \times X'_i$, $\overline{\mu}_i = \pi_i$ and $\overline{\omega}((x, x'), (y, y')) = \omega(x, y)$. The maps $\phi_i : \overline{X}_i \to X_i$ and $\phi'_i : \overline{X}_i \to X'_i$ from the definition of weak isomorphism are coordinate projection maps. One can then show that this gives a weak isomorphism of P and P'. Finally, define a new labelling function $\overline{\iota}_i : \overline{X}_i \to \Lambda_i$ by $\overline{\iota}_i(x, x') = \iota_i(x)$. Since d_{Λ_i} is a metric, it must be that $\iota_i(x) = \iota'_i(x')$ for π_i-almost every $(x, x') \in X_i \times X'_i$, so this labelling function satisfies the condition in the definition of weak isomorphism. □

3.1.2 Consequences and comparisons to other results

We now give a proof of Theorem 2, which says that the (unlabelled) partitioned network distance is a metric, and which then implies that various other generalized network distances in the literature are metrics as well (Corollary 2.16).

Proof of Theorem 2. Consider the map which takes a k-partitioned measure network P to the labelled k-partitioned measure network $(P, (\iota_i))$, where $(\Lambda_i, d_{\Lambda_i})$ is the one-point metric space for all i (hence ι_i is the constant map for all i). Clearly, we have

$$d_{\mathcal{P}_k^p}(P, P') = d_{\mathcal{LP}_k^p}\big((P, (\iota_i)), (P', (\iota'_i))\big),$$

since the labelling term in the definition of $d_{\mathcal{LP}_k^p}$ vanishes. Thus the map $P \mapsto (P, (\iota_i))$ induces a bijection from $[\mathcal{P}_k^p]$ to $[\mathcal{LP}_k^p]$ which takes $d_{\mathcal{P}_k^p}$ to $d_{\mathcal{LP}_k^p}$, and it follows that $d_{\mathcal{P}_k^p}$ is a metric. □

Next, we give a more precise comparison between the distance $d_{\mathcal{LP}_k^p}$ and the Fused Gromov-Wasserstein (FGW) distance of Vayer et al. [71]. The FGW distance is defined in the context of labelled measure networks; that is, in the $k = 1$ setting, where we write elements as (N, ι), with $N = (X, \mu, \omega) \in \mathcal{N}$ and $\iota : X \to \Lambda$, and in which case the distance $d_{\mathcal{LP}_1^p}$ reduces to

$$d_{\mathcal{LP}_1^p}((N, \iota), (N', \iota')) = \inf_{\pi \in \Pi(\mu, \mu')} \frac{1}{2} \left(\|\omega - \omega'\|^p_{L^p(\pi \otimes \pi)} + \|d_\Lambda \circ (\iota, \iota')\|^p_{L^p(\pi)} \right)^{1/p}, \tag{6}$$

for $p < \infty$. In contrast, the FGW distance depends on several more parameters, but the version of it which is closest to (6) would read as

$$\begin{aligned} & d_{\mathrm{FGW},p}((N, \iota), (N', \iota')) \\ & \quad = \inf_{\pi \in \Pi(\mu, \mu')} \frac{1}{2} \left(\int_{X \times X' \times X \times X'} \big(|\omega(x, y) - \omega'(x', y')| + d_\Lambda(\iota(x), \iota'(x')) \big)^p \, d\pi(x, x') d\pi(y, y') \right)^{1/p} \\ & \quad = \inf_{\pi \in \Pi(\mu, \mu')} \frac{1}{2} \| |\omega - \omega'| + d_\Lambda \circ (\iota, \iota') \|_{L^p(\pi \otimes \pi)}. \end{aligned}$$

Although the distances $d_{\mathcal{LP}_1^p}$ and $d_{\mathrm{FGW},p}$ treat the same type of object, the above shows that their formulations are subtly but legitimately distinct.

Remark 3.8 (Triangle Inequality for Fused Gromov-Wasserstein)**.** The situation described above is slightly murky, as several articles following [71] have formulated the FGW distance more in line with (6); see e.g., [68, 11, 31]. However, as far as we are aware, the primary references for FGW distance have not established a triangle inequality for any of its formulations. In particular, [71, 68] both give relaxed variants of the triangle inequality for different versions of FGW, where the larger side of the inequality involves an extra scale factor. Theorem 4 therefore gives a novel proof of the triangle inequality for FGW, when expressed in the form (6). We note that the triangle inequality for FGW was also recently established via an independent argument in [5, Corollary 4.3].

3.2 Alexandrov geometry of labelled partitioned networks

Next, we characterize geodesics and curvature in the space of labelled partitioned measure networks. For the rest of the section, we will assume the following conventions:

- We assume that the label spaces $(\Lambda_i, d_{\Lambda_i})$ are geodesic spaces. This is sometimes specialized further to assume that the label spaces are Hilbert spaces, but this specialization will always be pointed out explicitly in the statements of our results.
- We will restrict our attention to the case $p = 2$, and simply write $\mathcal{LP}_k$ in place of $\mathcal{LP}_k^2$.

The following is a generalization of Theorem 3; recall that the proof of that theorem was deferred—we will prove it in Section 3.2.3 as a corollary.

Theorem 5. *Let each Λ_i $(1 \leqslant i \leqslant k)$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\Lambda_i}$. Then, for any $k \geqslant 1$, $([\mathcal{LP}_k], d_{\mathcal{LP}_k})$ is an Alexandrov space of non-negative curvature.*

We will prove this by establishing the necessary properties as propositions. The proof techniques used in this section are largely adapted from the seminal work of Sturm [65].

3.2.1 Geodesic structure

We first prove two results on the geodesic structure of $[\mathcal{LP}_k]$.

Proposition 3.9. *For any $k \geqslant 1$, $([\mathcal{LP}_k], d_{\mathcal{LP}_k})$ is a geodesic space. For labelled k-partitioned measure networks $L = (P, (\iota_i)), L' = (P', (\iota'_i)) \in \mathcal{LP}_k$, a geodesic from $[P, (\iota_i)]$ to $[P', (\iota'_i)]$ is given by $[L^t] = [P^t, (\iota_i^t)]$, $t \in [0,1]$, defined as follows. The underlying k-partitioned measure network P^t is*

$$P^t = \big((X_i \times X'_i, \pi_i), \omega^t\big),$$

where (π_i) is a k-partitioned coupling which realizes $d_{\mathcal{LP}_k}(L, L')$, and $\omega^t : (\sqcup_i X_i) \times (\sqcup_i X'_i) \to \mathbb{R}$ is defined by

$$\omega^t((x, x'), (y, y')) = (1-t)\omega(x, y) + t\omega'(x', y'). \tag{7}$$

The labelling function (ι_i^t) is given by

$$\iota_i^t : X_i \times X'_i \longrightarrow \Lambda_i, \quad \iota_i^t(x, x') = \gamma_i^{x,x'}(t), \quad x, x' \in X_i \times X'_i,$$

where $\gamma_i^{x,x'} : [0,1] \to \Lambda_i$ is a geodesic between $\iota_i(x)$ and $\iota'_i(x')$ for each $1 \leqslant i \leqslant k$.

Proof. It is straightforward to show that L^0 is weakly isomorphic to L and L^1 is weakly isomorphic to L'. To show that $[L^t]$ defines a geodesic, it suffices to show that

$$d_{\mathcal{LP}_k}(L^s, L^t) \leqslant (t-s) d_{\mathcal{LP}_k}(L, L'), \tag{8}$$

for all $s, t \in [0,1]$ with $s < t$ (see, e.g., [18, Lemma 1.3]). Let $(\pi_i) \in \Pi_k((\mu_i), (\mu'_i))$ be optimal (Lemma 3.6) and set

$$1_{\pi_i} := (\mathrm{id}_{X_i \times X'_i} \times \mathrm{id}_{X_i \times X'_i})_\# \pi_i \in \Pi(\pi_i, \pi_i)$$

for each $i = 1, \ldots, k$. Then

$$4 \cdot d_{\mathcal{LP}_k}(L^s, L^t)^2 \leqslant \sum_{i,j=1}^k \|\omega^s - \omega^t\|^2_{L^2(1_{\pi_i} \otimes 1_{\pi_j})} + \sum_{i=1}^k \|d_{\Lambda_i} \circ (\iota_i^s, \iota_i^t)\|^2_{L^2(1_{\pi_i})}. \tag{9}$$

Applying the various definitions, it is straightforward to show that, for each pair (i,j),

$$\|\omega^s - \omega^t\|^2_{L^2(1_{\pi_i}\otimes 1_{\pi_j})} = (t-s)^2 \|\omega - \omega'\|^2_{L^2(\pi_i \otimes \pi_j)}.$$

To bound the last term in (9), observe that the term $\|d_{\Lambda_i} \circ (\iota_i^s, \iota_i^t)\|^2_{L^2(1_{\pi_i})}$ is equal to

$$\iint\limits_{(X_i \times X_i')^2} d_\Lambda(\iota_i^s(x,x'), \iota_i^t(y,y'))^2 \, \mathrm{d}1_\pi((x,x'),(y,y')) = \int_{X_i \times X_i'} d_{\Lambda_i}(\iota_i^s(x,x'), \iota_i^t(x,x'))^2 \, \mathrm{d}\pi_i(x,x').$$

We have that, for all $0 \leqslant s \leqslant t \leqslant 1$,

$$d_{\Lambda_i}(\iota_i^s(x,x'), \iota_i^t(x,x')) = d_{\Lambda_i}(\gamma_i^{x,x'}(s), \gamma_i^{x,x'}(t)) = (t-s) d_{\Lambda_i}(\iota_i(x), \iota_i(x')),$$

where the second equality follows by geodesity of $\gamma^{x,x'}$. This implies

$$\|d_{\Lambda_i} \circ (\iota_i^s, \iota_i^t)\|^2_{L^2(1_{\pi_i})} = (t-s)^2 \int_{X_i \times X_i'} d_\Lambda\big(\iota_i(x), \iota_i(x')\big)^2 \, \mathrm{d}\pi_i(x,x').$$

Putting all of this together yields the desired inequality (8). □

Proposition 3.10. *Let us now assume that each of the $\Lambda_i, 1 \leqslant i \leqslant k$ are inner product spaces with inner products $\langle \cdot, \cdot \rangle_{\Lambda_i}$, associated norms $\| \cdot \|_{\Lambda_i}$ and metrics d_{Λ_i} induced by their norms. Then any geodesic in $[\mathcal{LP}_k]$ can be written in the form given in Proposition 3.9: for any geodesic $[P^t, (\iota_i^t)], t \in [0,1]$ between $[L]$ and $[L']$, there exists an optimal coupling $(\pi_i) \in \Pi_k((\mu_i),(\mu_i'))$ such that $[P^t, (\iota_i^t)]$ is weakly isomorphic to $((X_i \times X_i', \pi_i), \omega^t, \gamma_i^t)$ where ω^t is given by (7) and where*

$$\gamma_i^{x,x'}(t) := (1-t)\iota_i(x) + t\iota_i'(x').$$

We will use some additional notation and terminology in subsequent proofs.

Definition 3.11. *Let Λ be an inner product space with inner product $\langle \cdot, \cdot \rangle_\Lambda$ and induced norm $\| \cdot \|_\Lambda$. For a probability space (Z, π), consider the space of functions $\iota : Z \to \Lambda$ such that*

$$\int_Z \|\iota(z)\|^2_\Lambda \mathrm{d}\pi(z) < \infty.$$

We denote the space of such functions, considered up to almost-everywhere equality, as $L^2(\pi, \Lambda)$. This is an inner product space with inner product defined by

$$\langle \iota, \iota' \rangle_{L^2(\pi,\Lambda)} := \int_Z \langle \iota(z), \iota'(z) \rangle_\Lambda \mathrm{d}\pi(z).$$

We let $\| \cdot \|_{L^2(\pi,\Lambda)}$ denote the associated norm.

Proof of Proposition 3.10. Let $(P,(\iota_i)),(P',(\iota_i')) \in \mathcal{LP}_k$ and let $[P^t, (\iota_i^t)]$ be an arbitrary geodesic from $[P^0, (\iota_i^0)] = [P, (\iota_i)]$ to $[P^1, (\iota_1^i)] = [P', (\iota_i')]$ with $P^t = ((X_i^t, \mu_i^t), \omega^t) \in \mathcal{P}_k$. We will show that $(P^t, (\iota_i^t))$ is (pointwise, in time) weakly isomorphic to a geodesic in the form described in Proposition 3.9.

For each $t \in [0,1]$, let $X^t = \sqcup_i X_i^t$. Fix an integer n and consider a dyadic decomposition of the time domain, $t_0 = 0, t_1 = \frac{1}{2^n}, \ldots, t_i = \frac{i}{2^n}, \ldots, t_{2^n} = 1$. For each $j = 1, \ldots, 2^n$, choose an optimal k-partitioned coupling $(\pi_i^j)_{i=1}^k \in \Pi_k\big((\mu_i^{t_{j-1}}), (\mu_i^{t_j})\big)$ (via Lemma 3.6). Consider the gluings (Lemma 3.7)

$$\widetilde{\pi}_i = \pi_i^1 \boxtimes \cdots \boxtimes \pi_i^{2^n} \in \mathbb{P}\big(X^0 \times X^{2^{-n}} \times X^{2 \cdot 2^{-n}} \times \cdots \times X^1\big), \qquad i \in \{1, \ldots, k\}.$$

Let $\mathrm{proj}_t : X^0 \times X^{2^{-n}} \times \cdots \times X^1 \to X^t$ denote coordinate projection for each $t \in \{0, 2^{-n}, \ldots, 1 - 2^{-n}, 1\}$ and define

$$\pi_i = (\mathrm{proj}_0 \times \mathrm{proj}_1)_\# \widetilde{\pi}_i \in \Pi_k(\mu^0, \mu^1)$$

for each $i = 1, \dots, k$. Then, by suboptimality,

$$4 \cdot d_{\mathcal{LP}_k}((P^0, (\iota_i^0), (P^1, (\iota_i^1))^2 \leqslant \sum_{i,j=1}^{k} \|\omega^0 - \omega^1\|^2_{L^2(\pi_i \otimes \pi_j)} + \sum_{i=1}^{k} \|d_{\Lambda_i} \circ (\iota_i^0, \iota_i^1)\|^2_{L^2(\pi_i)}. \tag{10}$$

First, consider the term $\|\omega^0 - \omega^1\|^2_{L^2(\pi_i \otimes \pi_j)}$ on the right hand side of (10). For any choice of $t \in \{0, 2^{-n}, \dots, 1 - 2^{-n}, 1\}$, let

$$\xi_i^t := (\mathrm{proj}_0 \times \mathrm{proj}_1 \times \mathrm{proj}_t)_\# \widetilde{\pi}_i \in \mathbb{P}(X^0 \times X^1 \times X^t).$$

We have

$$\begin{aligned}
&\|\omega^0 - \omega^1\|^2_{L^2(\pi_i \otimes \pi_j)} \\
&= \left\| t \left(\frac{1}{t} (\omega^0 - \omega^t) \right) + (1-t) \left(\frac{1}{1-t} (\omega^t - \omega^1) \right) \right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)} && (11) \\
&= \frac{1}{t} \left\| \omega^0 - \omega^t \right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)} + \frac{1}{1-t} \left\| \omega^t - \omega^1 \right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)} - \frac{1}{t(1-t)} \left\| (1-t)(\omega^0 - \omega^t) - t(\omega^t - \omega^1) \right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)}, && (12)
\end{aligned}$$

where (11) uses marginalization to replace $\pi_i \otimes \pi_j$ with $\xi_i^t \otimes \xi_j^t$, and where (12) is derived by applying the following identity, which holds in an arbitrary inner product space with associated norm $\| \cdot \|$:

$$\|ta + (1-t)b\|^2 = t\|a\|^2 + (1-t)\|b\|^2 - t(1-t)\|a - b\|^2. \tag{13}$$

Bearing in mind that $t = k2^{-n}$ for some k, the first term in (12) satisfies

$$\begin{aligned}
\frac{1}{t} \left\| \omega^0 - \omega^t \right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)} &= 2^n \cdot \frac{1}{k} \left\| \omega^0 - \omega^{k2^{-n}} \right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)} \\
&= 2^n \cdot \frac{1}{k} \left\| \sum_{\ell=1}^{k} (\omega^{(\ell-1)2^{-n}} - \omega^{\ell 2^{-n}}) \right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)} \\
&\leqslant 2^n \cdot \frac{1}{k} \left(\sum_{\ell=1}^{k} \left\| \omega^{(\ell-1)2^{-n}} - \omega^{\ell 2^{-n}} \right\|_{L^2(\xi_i^t \otimes \xi_j^t)} \right)^2 && (14) \\
&\leqslant 2^n \sum_{\ell=1}^{k} \left\| \omega^{(\ell-1)2^{-n}} - \omega^{\ell 2^{-n}} \right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)}, && (15)
\end{aligned}$$

where (14) follows by the triangle inequality for the L^2-norm and (15) is Jensen's inequality. Similarly, the second term in (12) satisfies

$$\frac{1}{1-t} \left\| \omega^t - \omega^1 \right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)} \leqslant 2^n \sum_{\ell=k+1}^{2^n} \left\| \omega^{(\ell-1)2^{-n}} - \omega^{\ell 2^{-n}} \right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)},$$

so that, after marginalizing, we have

$$\frac{1}{t} \left\| \omega^0 - \omega^t \right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)} + \frac{1}{1-t} \left\| \omega^t - \omega^1 \right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)} \leqslant 2^n \sum_{\ell=1}^{2^n} \left\| \omega^{(\ell-1)2^{-n}} - \omega^{\ell 2^{-n}} \right\|^2_{L^2\left(\pi_i^{(\ell-1)2^{-n}} \otimes \pi_j^{\ell 2^{-n}}\right)}. \tag{16}$$

Next, consider the term $\|d_{\Lambda_i} \circ (\iota_i^0, \iota_i^1)\|^2_{L^2(\pi_i)}$. The following uses the notation of Definition 3.11. We have, similar to the above,

$$\begin{aligned}
&\|d_{\Lambda_i} \circ (\iota_i^0, \iota_i^1)\|^2_{L^2(\pi_i)} \\
&= \left\| t \left(\frac{1}{t} (\iota_i^0 - \iota_i^t) \right) + (1-t) \left(\frac{1}{1-t} (\iota_i^t - \iota_i^1) \right) \right\|^2_{L^2(\xi_i^t, \Lambda_i)}
\end{aligned}$$

$$= \frac{1}{t}\|\iota_i^0 - \iota_i^t\|^2_{L^2(\xi_i^t,\Lambda)} + \frac{1}{1-t}\|\iota_i^t - \iota_i^1\|^2_{L^2(\xi_i^t,\Lambda)} - \frac{1}{t(1-t)}\|(1-t)(\iota_i^0 - \iota_i^t) - t(\iota_i^t - \iota_i^1)\|^2_{L^2(\xi_i^t,\Lambda)},$$

where we have used the definition of d_{Λ_i}, as well as marginalization and the general identity (13). Repeating the arguments above, we obtain

$$\frac{1}{t}\left\|\iota_i^0 - \iota_i^t\right\|^2_{L^2(\xi_i^t,\Lambda_i)} + \frac{1}{1-t}\left\|\iota_i^t - \iota_i^1\right\|^2_{L^2(\xi_i^t,\Lambda_i)} \leqslant 2^n \sum_{\ell=1}^{2^n} \left\| d_{\Lambda_i} \circ \left(\iota_i^{(\ell-1)2^{-n}}, \iota_i^{\ell 2^{-n}}\right)\right\|^2_{L^2\left(\pi_i^{(\ell-1)2^{-n}}\right)}. \tag{17}$$

Summing the right hand sides of (16) and (17) over all $i, j = 1, \ldots, k$ gives

$$\begin{aligned}
&\sum_{i,j=1}^{k} 2^n \sum_{\ell=1}^{2^n} \left\|\omega^{(\ell-1)2^{-n}} - \omega^{\ell 2^{-n}}\right\|^2_{L^2\left(\pi_i^{(\ell-1)2^{-n}} \otimes \pi_j^{\ell 2^{-n}}\right)} + \sum_{i=1}^{k}\sum_{\ell=1}^{2^n} \left\| d_{\Lambda_i} \circ \left(\iota_i^{(\ell-1)2^{-n}}, \iota_i^{\ell 2^{-n}}\right)\right\|^2_{L^2\left(\pi_i^{(\ell-1)2^{-n}}\right)} \\
&= 2^n \sum_{\ell=1}^{2^n} \left(\sum_{i,j=1}^{k} \left\|\omega^{(\ell-1)2^{-n}} - \omega^{\ell 2^{-n}}\right\|^2_{L^2\left(\pi_i^{(\ell-1)2^{-n}} \otimes \pi_j^{\ell 2^{-n}}\right)} + \sum_{i=1}^{k} \left\| d_{\Lambda_i} \circ \left(\iota_i^{(\ell-1)2^{-n}}, \iota_i^{\ell 2^{-n}}\right)\right\|^2_{L^2\left(\pi_i^{(\ell-1)2^{-n}}\right)} \right) \\
&= 2^n \sum_{\ell=1}^{2^n} 4 \cdot d_{\mathcal{LP}_k}\left((P^{(\ell-1)2^{-n}}, (\iota_i^{(\ell-1)2^{-n}})), (P^{\ell 2^{-n}}, (\iota_i^{(\ell)2^{-n}}))\right)^2 \qquad (18) \\
&= 4 \cdot d_{\mathcal{LP}_k}\left((P^0, (\iota_i^0)), (P^1, (\iota_i^1))\right)^2, \qquad (19)
\end{aligned}$$

where (18) follows by the optimality of the π_i^j's and (19) follows because $[P^t]$ is assumed to be a geodesic. Combining this with (10), we have

$$\begin{aligned}
&d_{\mathcal{LP}_k}\left((P^0, (\iota_i^0)), (P^1, (\iota_i^1))\right)^2 \\
&\quad \leqslant d_{\mathcal{LP}_k}\left((P^0, (\iota_i^0)), (P^1, (\iota_i^1))\right)^2 \\
&\qquad - \frac{1}{4(t(1-t))}\left(\sum_{i,j=1}^{k} \left\|(1-t)(\omega^0 - \omega^t) - t(\omega^t - \omega^1)\right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)} \right. \\
&\qquad\qquad \left. + \sum_{i=1}^{k} \|d_{\Lambda_i} \circ \left((1-t)(\iota_i^0 - \iota_i^t), t(\iota_i^t - \iota_i^1)\right)\|^2_{L^2(\xi_i^t)} \right),
\end{aligned}$$

so that the term in parentheses on the right hand side must vanish. This shows that the partitioned coupling (π_i) which we have constructed is, in fact, optimal. We also have that, for all t in the dyadic decomposition,

$$0 = \sum_{i,j=1}^{k} \left\|(1-t)(\omega^0 - \omega^t) - t(\omega^t - \omega^1)\right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)} = \sum_{i,j=1}^{k} \left\|\left((1-t)\omega^0 + t\omega^1\right) - \omega^t\right\|^2_{L^2(\xi_i^t \otimes \xi_j^t)}.$$

Similarly,

$$\begin{aligned}
0 &= \sum_{i=1}^{k} \|d_{\Lambda_i} \circ \left((1-t)(\iota_i^0 - \iota_i^t), t(\iota_i^t - \iota_i^1)\right)\|^2_{L^2(\xi_i^t)} \\
&= \sum_{i=1}^{k} \|\left((1-t)\iota_i^0 + t\iota_i^1\right) - \iota_i^t\|^2_{L^2(\xi_i^t,\Lambda_i)} = \sum_{i=1}^{k} \|d_{\Lambda_i} \circ \left(\left((1-t)\iota_i^0 + t\iota_i^1\right), \iota_i^t\right)\|^2_{L^2(\xi_i^t)}.
\end{aligned}$$

Observe that, by the properties described in the Gluing Lemma (Lemma 3.7), we have $\xi_i^t \in \Pi(\pi_i, \mu_i^t)$, so that the above calculation shows $d_{\mathcal{LP}_k}\left((P^t, (\iota_i^t)), (\overline{P}^t, (\overline{\iota}_i^t))\right) = 0$, where $(\overline{P}^t, (\overline{\iota}_i^t))$ is a geodesic as in the specific construction from Proposition 3.9.

So far, we have shown that $d_{\mathcal{LP}_k}\left((P^t, (\iota_i^t)), (\overline{P}^t, (\overline{\iota}_i^t))\right) = 0$ for any t in the form of a dyadic number, i.e., $t = j2^{-n}$ for some j and n. By the density of the dyadic numbers in $[0,1]$ and by continuity of the maps $t \mapsto [\overline{P}^t, (\overline{\iota}_i^t)]$ and $t \mapsto [P^t, (\iota_i^t)]$, it follows that $d_{\mathcal{LP}_k}\left((P^t, (\iota_i^t)), (\overline{P}^t, (\overline{\iota}_i^t))\right) = 0$ holds for any $t \in [0,1]$. This completes the proof. □

3.2.2 Completeness and curvature

We now complete the proof of Theorem 5 by establishing the remaining required properties. Throughout this section, we suppose that the label spaces Λ_i are Hilbert spaces with the same notation as Proposition 3.9 used for inner products and norms.

We first show that the space of labelled networks is complete. The proof will use the following result.

Lemma 3.12 (See, e.g., [37]). *If Λ is a Hilbert space, then so is $L^2(\pi, \Lambda)$.*

Proposition 3.13. *Let each $\Lambda_i, 1 \leqslant i \leqslant k$ be a Hilbert space. Then the space $\left([\mathcal{LP}_k], d_{\mathcal{LP}_k}\right)$ is complete.*

Proof. The proof follows the strategy of the proofs of [65, Theorem 5.8] or [22, Theorem 1], so we treat it somewhat tersely. Let $[P^n, (\iota_i^n)]$, $n \geqslant 1$, be a Cauchy sequence of labelled partitioned networks in $[\mathcal{LP}_k]$, with $P^n = ((X_i^n, \mu_i^n), \omega^n)$. Assume, without loss of generality (via a subsequence argument), that $d_{\mathcal{LP}_k}\left((P_n, (\iota_i^n)), (P_{n+1}, (\iota_i^{n+1}))\right) \leqslant 2^{-n}$. Invoking Lemma 3.6, we may choose partitioned couplings (π_i^n) for each n which achieve $d_{\mathcal{LP}_k}\left((P_n, (\iota_i^n)), (P_{n+1}, (\iota_i^{n+1}))\right)$. Gluing the first N of these measures yields a probability measure $\pi_i^1 \boxtimes \pi_i^2 \boxtimes \cdots \boxtimes \pi_i^N$ on $X_i^1 \times X_i^2 \times \cdots \times X_i^N$ for each $i = 1, \ldots, k$. Let π_i denote the projective limit measure on $\Pi_{\ell=1}^\infty X_i^\ell$.

For each N, define maps

$$\Omega^N : \left(\bigsqcup_{i=1}^k \Pi_{\ell=1}^\infty X_i^\ell\right) \times \left(\bigsqcup_{i=1}^k \Pi_{\ell=1}^\infty X_i^\ell\right) \to \mathbb{R}$$
$$\left((x^\ell)_\ell, (y^\ell)_\ell\right) \mapsto \omega^N(x^N, y^N)$$

and

$$I_i^N : \Pi_{\ell=1}^\infty X_i^\ell \to \Lambda_i$$
$$(x^\ell)_\ell \mapsto \iota_i^N(x^N)$$

Since $d_{\mathcal{LP}_k}\left((P_n, (\iota_i^n)), (P_{n+1}, (\iota_i^{n+1}))\right) \leqslant 2^{-n}$, it must be that

$$\frac{1}{4}\|\omega_n - \omega_{n+1}\|^2_{L^2(\pi_i \otimes \pi_j)} \leqslant 2^{-2n} \quad \text{and} \quad \frac{1}{4}\|\iota_{i,n} - \iota_{i,n+1}\|^2_{L^2(\pi_i, \Lambda_i)} \leqslant 2^{-2n},$$

where we use the notation of Definition 3.11 in the second term. It follows that the sequence (Ω^N) is Cauchy in the Hilbert space $L^2(\pi_i \otimes \pi_j)$ and that (I_i^N) is Cauchy in the Hilbert space $L^2(\pi_i, \Lambda_i)$ (this is Hilbert because we assumed that Λ_i is Hilbert; see Lemma 3.12). Let $\Omega := \lim_{N \to \infty} \Omega^N$ and $I_i := \lim_{N \to \infty} I_i^N$.

Putting these constructions together, we have constructed a labelled k-partitioned network

$$\left(\left(\left(\Pi_{\ell=1}^\infty X_i^\ell, \pi_i\right), \Omega\right), (I_i)\right).$$

One can then show that its weak isomorphism class is the limit of the original Cauchy sequence. □

Finally, we establish a curvature bound for the space of labelled partitioned networks.

Proposition 3.14. *Assume that all label spaces Λ_i are Hilbert spaces. Then the space $([\mathcal{LP}_k], d_{\mathcal{LP}_k})$ has curvature bounded below by zero.*

Proof. We need to establish the triangle comparison inequality from Definition 2.18. Let $[L], [L'] \in [\mathcal{LP}_k]$ be two labelled partitioned networks and let $[L^t], 0 \leqslant t \leqslant 1$ be a geodesic connecting them. Let $[L''] \in [\mathcal{LP}_k]$ be given. We seek to show that

$$4d_{\mathcal{LP}_k}(L^t, L'')^2 + 4t(1-t)d_{\mathcal{LP}_k}(L, L')^2 \geqslant 4(1-t)d_{\mathcal{LP}_k}(L, L'')^2 + 4td_{\mathcal{LP}_k}(L', L'')^2. \tag{20}$$

Using the characterization of geodesics in $[\mathcal{LP}_k]$ from Proposition 3.10, we may assume without loss of generality that $L^t = (P^t, (\iota_i^t))$ has the form described in Proposition 3.9. Let (ξ_i) be an optimal k-partitioned coupling of L^t to L''; then ξ_i is supported on $X_i \times X_i' \times X_i''$. Expanding the left hand side of (20), we have

$$\sum_{i,j=1}^k \left(\|\omega'' - \omega^t\|^2_{L^2(\xi_i \otimes \xi_j)} + t(1-t)\|\omega - \omega'\|^2_{L^2(\pi_i \otimes \pi_j)}\right) + \sum_{i=1}^k \left(\|\iota'' - \iota^t\|^2_{L^2(\xi_i, \Lambda_i)} + t(1-t)\|\iota - \iota'\|^2_{L^2(\pi_i, \Lambda_i)}\right).$$

Marginalizing ξ_i and using the structures of ω^t and ι^t, this can be rewritten as

$$
\begin{aligned}
&\sum_{i,j=1}^{k} \Big(\|\omega'' - \omega^t\|^2_{L^2(\xi_i\otimes\xi_j)} + t(1-t)\|\omega - \omega'\|^2_{L^2(\xi_i\otimes\xi_j)}\Big) + \sum_{i=1}^{k} \Big(\|\iota'' - \iota^t\|^2_{L^2(\xi_i,\Lambda_i)} + t(1-t)\|\iota - \iota'\|^2_{L^2(\xi_i,\Lambda_i)}\Big) \\
&= \sum_{i,j=1}^{k} \Big((1-t)\|\omega'' - \omega\|^2_{L^2(\xi_i\otimes\xi_j)} + t\|\omega'' - \omega'\|^2_{L^2(\xi_i\otimes\xi_j)}\Big) + \sum_{i=1}^{k} \Big((1-t)\|\iota'' - \iota\|^2_{L^2(\xi_i,\Lambda_i)} + t\|\iota'' - \iota'\|^2_{L^2(\xi_i,\Lambda_i)}\Big) \\
&\geqslant 4(1-t)d_{\mathcal{LP}_k}(P,P'')^2 + 4t d_{\mathcal{LP}_k}(P',P'')^2,
\end{aligned}
$$

where the second line follows from a computation which holds in an arbitrary inner product space, explained below, and the inequality in the last line follows from sub-optimality. To expand on the inner product space calculation, we use the identity (13) to deduce that (for an arbitrary inner product $\langle\cdot,\cdot\rangle$ with norm $\|\cdot\|$),

$$
\begin{aligned}
\|c - ((1-t)a + tb)\|^2 + t(1-t)\|a-b\|^2 &= \|c\|^2 + \|(1-t)a + tb\|^2 - 2\langle c, (1-t)a + tb\rangle + t(1-t)\|a-b\|^2 \\
&= \|c\|^2 + t\|b\|^2 + (1-t)\|a\|^2 - 2\langle c, (1-t)a + tb\rangle \\
&= (1-t)\big(\|c\|^2 + \|a\|^2 - 2\langle c, a\rangle\big) + t\big(\|c\|^2 + \|b\|^2 - 2\langle c, b\rangle\big) \\
&= (1-t)\|c-a\|^2 + t\|c-b\|^2.
\end{aligned}
$$

□

3.2.3 The case of unlabelled networks

We now proceed with the deferred proof of Theorem 3, which says that the space of (unlabelled) k-partitioned networks is an Alexandrov space of non-negative curvature.

Proof of Theorem 3. Following the proof of Theorem 2 (Section 3.1.2), we can consider $d_{\mathcal{P}_k^2}$ as an instance of $d_{\mathcal{LP}_k^2}$, where the target metric spaces Λ_i are all taken to be the one-point space, which can be considered as a trivial Hilbert space. The proof then follows immediately from Theorem 5. □

3.3 Interpretation of partitioned distance as a labelled distance

We end this section by proving that the k-partitioned network distance $d_{\mathcal{P}_k^p}$ can itself be realized as a sort of labelled distance, where labels are allowed to take the value ∞. To keep exposition clean, we recapitulate and rework some of our notation before precisely stating our result.

3.3.1 Notation for networks labelled in extended metric spaces

In this section, we write $\mathcal{LN}^p_{\mathrm{ext}}$ for the space of **networks with labels in a (fixed) extended metric space** Λ. This is essentially the same as $\mathcal{LP}_1^p$, except that the label space is allowed to be an **extended metric space**, or a metric space whose distances are allowed to take the value ∞. Elements of $\mathcal{LN}^p_{\mathrm{ext}}$ will be denoted (N,ι), where $N = (X,\mu,\omega) \in \mathcal{N}^p$ and $\iota : X \to \Lambda$ is the label function. The labelled network distance is then defined by a formula identical to (6):

$$
d_{\mathcal{LN}^p_{\mathrm{ext}}}((N,\iota),(N',\iota')) = \inf_{\pi\in\Pi(\mu,\mu')} \frac{1}{2}\Big(\|\omega-\omega'\|^p_{L^p(\pi\otimes\pi)} + \|d_\Lambda\circ(\iota,\iota')\|^p_{L^p(\pi)}\Big)^{1/p}.
$$

The notion of weak isomorphism extends to $\mathcal{LN}^p_{\mathrm{ext}}$; we denote the quotient space as $[\mathcal{LN}^p_{\mathrm{ext}}]$. The proofs above also extend to show that $d_{\mathcal{LN}^p_{\mathrm{ext}}}$ induces a well-defined extended metric on $[\mathcal{LN}^p_{\mathrm{ext}}]$.

We now consider the particular extended metric space $\Lambda^{(k)} := \{1,\ldots,k\}$, with extended metric satisfying $d_{\Lambda^{(k)}}(i,j) = \infty$ for all $i \neq j$. To a k-partitioned measure network P, we associate an element of $\mathcal{LN}^p_{\mathrm{ext}}$, with labels in $\Lambda^{(k)}$, via the map

$$
\mathcal{P}_k^p \ni P = ((X_i,\mu_i),\omega) \mapsto ((X,(1/k)\mu,\omega),\iota) \in \mathcal{LN}^p_{\mathrm{ext}}, \qquad \iota(x) = i \Leftrightarrow x \in X_i. \tag{21}
$$

Here, as in Definition 2.7, $\mu = \sum_i \mu_i$ is considered as a measure on $X = \sqcup_i X_i$, so that $\frac{1}{k}\mu$ is a probability measure.

3.3.2 Partitioned distance as a labelled network distance

We now prove the main result of Section 3.3.

Theorem 6. *The map* (21) *is an isometric embedding with respect to* $d_{\mathcal{P}^p_k}$ *and* $d_{\mathcal{LN}^p_{\mathrm{ext}}}$.

Observe that

$$\|d_{\Lambda^{(k)}} \circ (\iota, \iota')\|_{L^p(\pi)} = \begin{cases} \infty & \text{if there exists } (x,x') \in \mathrm{supp}(\pi) \text{ with } \iota(x) \neq \iota'(x'); \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

The proof of the theorem is based on this observation and the following lemma.

Lemma 3.15. *Let* $\pi \in \Pi(\sqcup_i \mu_i, \sqcup_i \mu_i')$ *such that* $\|d_{\Lambda^{(k)}} \circ (\iota, \iota')\|_{L^p(\pi)} = 0$. *Then there exists a unique* k*-partitioned coupling* $(\pi_i) \in \Pi_k\big((\mu_i), (\mu_i')\big)$ *such that*

$$\pi = \mathrm{inc}_{\#}\left(\sqcup_i \pi_i\right),$$

where $\mathrm{inc} : \sqcup_i (X_i \times X_i') \to (\sqcup_i X_i) \times (\sqcup_i X_i')$ *is the inclusion map. Moreover, any coupling of this form yields* $\|d_{\Lambda^{(k)}} \circ (\iota, \iota')\|_{L^p(\pi)} = 0$.

Proof. Suppose that π is a coupling with $\|d_{\Lambda^{(k)}} \circ (\iota, \iota')\|_{L^p(\pi)} < \infty$. Then $(x,x') \in \mathrm{supp}(\pi)$ implies $\iota(x) = \iota'(x')$, i.e., $x \in X_i$ and $x' \in X_i'$ for some common index i. Therefore the support of π is contained in $\sqcup_i (X_i \times X_i')$. We also see that the mass of each block $X_i \times X_i'$ must be equal to $\frac{1}{k}$, since, for all i,

$$\pi(X_i \times X_i') = \pi((\sqcup_j X_j) \times X_i') = \sqcup_j \mu_j'(X_i') = \frac{1}{k}\sum_j \mu_j'(X_i') = \frac{1}{k}\mu_i'(X_i') = \frac{1}{k}.$$

For each i, define π_i by $\pi_i(A) = k \cdot \pi(A)$ for each Borel set $A \subset X_i \times X_i'$. We claim that $\pi_i \in \Pi(\mu_i, \mu_i')$. Indeed, for any Borel set $B \subset X_i$,

$$\pi_i(B \times X_i') = k \cdot \pi(B \times X_i') = k \cdot \pi(B \times (\sqcup_i X_i')) = k \cdot \sqcup_j \mu_j(B \cap X_j) = \mu_i(B),$$

and the other marginal condition follows similarly.

We will now show that

$$\pi(C) = \mathrm{inc}_{\#}\left(\sqcup_i \pi_i\right)(C)$$

holds for any Borel subset C of $(\sqcup_i X_i) \times (\sqcup_i X_i')$. In light of the discussion above, we may assume without loss of generality that $C \subset \sqcup_i (X_i \times X_i')$. Then we have

$$\begin{aligned} \mathrm{inc}_{\#}(\sqcup_i \pi_i)(C) = \sqcup_i \pi_i(\iota^{-1}(C)) &= \frac{1}{k}\sum_i \pi_i(\mathrm{inc}^{-1}(C) \cap (X_i \times X_i')) \\ &= \frac{1}{k}\sum_i \pi_i(C \cap (X_i \times X_i')) = \sum_i \pi(C \cap (X_i \times X_i')) = \pi(C). \end{aligned}$$

This shows the existence part of the statement.

To prove uniqueness, suppose that (π_i) satisfies $\pi = \mathrm{inc}_{\#}(\sqcup_i \pi_i)$. For a Borel set $A \subset X_i \times X_i'$

$$\pi_i(A) = \sum_j \pi_j(A) = k \cdot \sqcup_j \pi_j(A) = k \cdot \mathrm{inc}_{\#}(\sqcup_j \pi_j)(A) = k \cdot \pi(A),$$

so the formula for π_i is unique.

Finally, the last statement follows because any coupling of this form is supported on $\sqcup_i (X_i \times X_i')$. □

Proof of Theorem 6. Let $P = ((X_i, \mu_i), \omega)$ and $P' = ((X_i', \mu_i'), \omega')$ be elements of $\mathcal{P}^p_k$ with images under the map (21) denoted (N, ι) and (N', ι'), respectively. A k-partitioned coupling (π_i) of (μ_i) and (μ_i') yields a coupling $\pi = \mathrm{inc}_{\#}((1/k)\sum_i \pi_i) \in \Pi((1/k)\sum_i \mu_i, (1/k)\sum_i \mu_i')$, as in Lemma 3.15. Then

$$\frac{1}{2}\|\omega - \omega'\|_{L^p(\pi \otimes \pi)} + \|d_{\Lambda^{(k)}} \circ (\iota, \iota')\|_{L^p(\pi)} = \frac{1}{2}\left(\sum_{i,j=1}^{k} \|\omega - \omega'\|^p_{L^p(\pi_i \otimes \pi_j)}\right)^{1/p},$$

so that $d_{\mathcal{LN}^p_{\text{ext}}}((N,\iota),(N',\iota')) \leqslant d_{\mathcal{P}^p_k}(P,P')$. Similarly, for any $\pi \in \Pi((1/k)\sum_i \mu_i, (1/k)\sum_i \mu'_i)$ with $\|d_{\Lambda^{(k)}} \circ (\iota,\iota')\|_{L^p(\pi)} = 0$, we can find a k-partitioned coupling $(\pi_i) \in \Pi_k((\mu_i),(\mu'_i))$ via Lemma 3.15 and use it to show that $d_{\mathcal{LN}^p_{\text{ext}}}((N,\iota),(N',\iota')) \geqslant d_{\mathcal{P}^p_k}(P,P')$. □

The following corollary shows that the various generalized network distances which have appeared in the recent literature can all essentially be considered as special cases of the labelled extended network distance. The result follows as a direct consequence of Corollary 2.16 and Theorem 6.

Corollary 3.16. *The embeddings from Definition 2.8 induce isometric embeddings of the space of measure networks $([\mathcal{N}], d_{\mathcal{N}^p})$, the space of measure hypernetworks $([\mathcal{H}], d_{\mathcal{H}^p})$ and the space of augmented measure networks $([\mathcal{A}], d_{\mathcal{A}^p})$, respectively, into the space of labelled networks $([\mathcal{LN}^p_{\text{ext}}], d_{\mathcal{LN}^p_{\text{ext}}})$.*

3.3.3 Labelled partitioned distance as a labelled network distance

The work above can be directly adapted to show that the space of labelled k-partitioned networks also embeds into the space of labelled networks. Consider the space $\mathcal{LP}^p_k$ of labelled k-partitioned p-networks with labels in some arbitrary metric spaces $(\Lambda_i, d_{\Lambda_i})$. Now consider the extended metric space $\Lambda = (\sqcup_i \Lambda_i) \times \Lambda^{(k)}$ with extended metric d_Λ defined by

$$d_\Lambda((a,i),(b,j)) = \begin{cases} d_{\Lambda_i}(a,b) & \text{if } i = j \text{ and } a, b \in \Lambda_i; \\ \infty & \text{otherwise.} \end{cases}$$

Given an element $L = (P,(\iota_i))$ of $\mathcal{LP}^p_k$, with $P = ((X_i,\mu_i),\omega)$, we associate an element of $\mathcal{LN}^p_{\text{ext}}$ with labels in Λ via the map

$$\mathcal{LP}^p_k \ni L \mapsto ((X,(1/k)\mu,\omega),\iota), \qquad \iota(x) = (\iota_i(x), i) \Leftrightarrow x \in X_i. \tag{23}$$

The techniques used above likewise yield a proof of the following.

Theorem 7. *The map (23) is an isometric embedding with respect to $d_{\mathcal{LP}^p_k}$ and $d_{\mathcal{LN}^p_{\text{ext}}}$.*

4 Riemannian structure of partitioned networks

We now focus again on the space $\mathcal{LP}_k = \mathcal{LP}^2_k$, endowed with the metric $d_{\mathcal{LP}_k} = d_{\mathcal{LP}^2_k}$. We consider the scenario where the label spaces $(\Lambda_i, d_{\Lambda_i})$ are Hilbert spaces endowed with their associated distances. We have showed in Theorem 5 that $([\mathcal{LP}_k], d_{\mathcal{LP}_k})$ is a non-negatively curved Alexandrov space. This property endows $[\mathcal{LP}_k]$ with synthetic versions of various structures seen in Riemannian geometry, such as tangent spaces and exponential maps [53]. Rather than following the general constructions of these structures, we follow the approach of Sturm in [65, Chapter 6] and develop equivalent versions which are more specific to the metric at hand. In this section, we describe these structures and present some example applications to geometric data analysis.

4.1 Tangent spaces

We develop notions of tangent spaces and exponential maps for $[\mathcal{P}_k]$ and $[\mathcal{LP}_k]$. These concepts are introduced in detail for $[\mathcal{LP}_k]$, and the case of $[\mathcal{P}_k]$ then follows by considering partitioned networks as special cases of labelled partitioned networks, as in the proof of Theorem 2 (see Section 3.1.2).

4.1.1 The labelled case

For clarity, we remind the reader of some notational conventions, while introducing some new ones. Let $L = (P,(\iota_i)) = \big((X_i,\mu_i),\omega,(\iota_i)\big) \in \mathcal{LP}_k$, where we continue to assume that the label spaces Λ_i are Hilbert spaces. As above, we write $X = \sqcup_i X_i$ and endow it with the measure $\mu = \sum_i \mu_i$. Given another element $P' \in \mathcal{LP}_k$ and a k-partitioned coupling $\pi \in \Pi_k((\mu_i),(\mu'_i))$, we write $\pi = \sum_i \pi_i$, and consider this as a measure on $X \times X'$ (where $X' = \sqcup_i X'_i$). Finally, we define

$$\Lambda_X := \bigoplus_{i=1}^k L^2(\mu_i, \Lambda_i).$$

We denote an element of Λ_X as g, where we can canonically write $g = (g_1, \ldots, g_k)$ with $g_i \in L^2(\mu_i, \Lambda_i)$. In this way, we consider the label function data as an element $\iota = (\iota_1, \ldots, \iota_k) \in \Lambda_X$.

Definition 4.1 (Synthetic tangent space)**.** *We define the* ***synthetic tangent space*** *of $[\mathcal{LP}_k]$ at a point $[L]$ to be*

$$T_{[L]}[\mathcal{LP}_k] := \left(\bigcup_{((X_i,\mu_i),\omega,(\iota_i))\in[L]} (L^2(\mu \otimes \mu) \oplus \Lambda_X) \right) / \sim .$$

In the above, the union is taken over all labelled k-partitioned measure networks $((X_i, \mu_i), \omega, (\iota_i))$ in the weak isomorphism equivalence class $[L]$. The equivalence relation is defined as follows. For two representatives

$$((X_i, \mu_i), \omega, (\iota_i)), \quad ((X'_i, \mu'_i), \omega', (\iota'_i)) \in [L]$$

and functions

$$(f, g) \in L^2(\mu \otimes \mu) \oplus \Lambda_X, \quad (f', g') \in L^2(\mu' \otimes \mu') \oplus \Lambda_{X'},$$

we write $(f, g) \sim (f', g')$ if and only if there exists a k-partitioned coupling $(\pi_i) \in \Pi_k(\mu, \mu')$ such that

$$f(x, y) = f'(x', y') \quad \text{for} \quad \pi \otimes \pi - \text{a.e. } (x, x', y, y') \in X \times X' \times X \times X'$$

and, writing $g = (g_1, \ldots, g_k)$ and $g' = (g'_1, \ldots, g'_k)$,

$$g_i(x_i) = g'_i(x'_i) \quad \text{for} \quad \pi_i - \text{a.e. } (x_i, x'_i) \in X \times X'.$$

The equivalence class of (f, g) is denoted $[f, g]$.

The space $[\mathcal{LP}_k]$ has a natural notion of an exponential map, defined as follows.

Definition 4.2 (Exponential map)**.** *For a labelled k-partitioned measure network $[L] \in [\mathcal{LP}_k]$, let $[f, g] \in T_{[L]}[\mathcal{LP}_k]$ be a tangent vector with $(f, g) \in L^2(\mu \otimes \mu) \oplus \Lambda_X$. We define the* ***exponential map*** *by*

$$\exp_{[L]} : T_{[L]}[\mathcal{LP}_k] \mapsto [\mathcal{LP}_k], \quad \exp_{[L]}([f, g]) := [((X_i, \mu_i), \omega + f), \iota + g].$$

We can now provide a geodesic characterization of the exponential map on $([\mathcal{LP}_k], d_{\mathcal{LP}_k})$, analogous to the one given for measure networks in [20], at least for labelled partitioned measure networks which are "inherently finite". That is, we say that an element $[L] \in [\mathcal{LP}_k]$ is **finite** if the equivalence class $[L]$ contains a representative $L' \in [L]$ such that all sets X'_i are finite. In the following, take the following terminology convention: if we refer to $[L] \in [\mathcal{LP}_k]$ as finite, we implicitly assume without loss of generality that L is finite. Observe that, even if L is finite, the equivalence class $[L]$ contains elements which are not finite, hence the need for care in the terminology here.

Proposition 4.3. *Let $[L] \in [\mathcal{LP}_k]$ be a finite labelled k-partitioned measure network. There exists $\epsilon_{[L]} > 0$ and $\eta_{[L]} > 0$ such that for any tangent vector represented by $(f, g) \in L^2(\mu \otimes \mu) \oplus \Lambda_X$ satisfying $|f(x, y)| < \epsilon_{[L]}$ and $|g(x)| < \eta_{[L]}$ for all $(x, y) \in X \times X$, the path defined by*

$$[\gamma_t] = [((X_i, \mu_i), \omega + tf), \iota + tg], \quad t \in [0, 1]$$

is a geodesic from $[L]$ to $\exp_{[L]}([f, g])$.

Proof. Up to weak isomorphism, we may assume that $[\gamma_t]$ takes the form

$$[\gamma_t] = [((X_i \times X_i, \Delta_{\mu_i,\mu_i}), \omega_t), \iota_t],$$

where

$$\begin{aligned} \omega_t &: (x, x', y, y') \mapsto (1 - t)\omega(x, y) + t(\omega(x', y') + f(x', y')) \\ \iota_t &: (x, x') \mapsto (1 - t)\iota(x) + t(\iota(x') + g(x')) \end{aligned}$$

and Δ_{μ_i,μ_i} denotes the **diagonal coupling** of (μ_i, μ_i); that is, $\Delta_{\mu_i,\mu_i} = (\mathrm{id}_{X_i} \times \mathrm{id}_{X_i})_{\#}\mu_i$. By Proposition 3.9, verifying that $[\gamma_t]$ is a geodesic amounts to deducing the condition on (f, g) for $(\Delta_{\mu_i,\mu_i})_{i=1}^k$ to be the optimal couplings between L and $\gamma(1)$. Let $(\pi_i) \in \Pi_{\mathcal{P}_k}(\mu, \mu)$ be any competitor coupling. The corresponding matching cost is then

$$\begin{aligned}
&\frac{1}{2}\sum_{X^4} \pi(x,x')\pi(y,y')|\omega(x,y) - \omega(x',y') - f(x',y')|^2 + \frac{1}{2}\sum_{X^2} \pi(x,x')\|\iota(x) - \iota(x') - g(x')\|^2 \\
&= \frac{1}{2}\sum_{X^4} \pi(x,x')\pi(y,y')f(x',y')^2 + \sum_{X^4} \pi(x,x')\pi(y,y')\left[\frac{1}{2}|\omega(x,y) - \omega(x',y')|^2 - f(x',y')(\omega(x,y) - \omega(x',y'))\right] \\
&\qquad + \frac{1}{2}\sum_{X^2} \pi(x,x')\|g(x')\|^2 + \sum_{X^2} \pi(x,x')\left[\frac{1}{2}\|\iota(x) - \iota(x')\|^2 - \langle g(x'), \iota(x) - \iota(x')\rangle\right].
\end{aligned}$$

In the above, the inner product and norms are the induced structures on $\oplus_i \Lambda_i$. The first and third terms amount to the matching cost between L and $\gamma(1)$ under the diagonal couplings $(\Delta_{\mu_i,\mu_i})_i$ and so it is sufficient to deduce conditions on (f, g) so that the second and fourth terms are non-negative.

Consider first the sum

$$\sum_{X^4} \pi(x,x')\pi(y,y')\left[\frac{1}{2}|\omega(x,y) - \omega(x',y')|^2 - f(x',y')(\omega(x,y) - \omega(x',y'))\right].$$

Clearly, if $|\omega(x,y) - \omega(x',y')|^2 = 0$ $\pi \otimes \pi$-a.e. then this term vanishes. Otherwise there exists at least one $(x, x', y, y') \in X^4$ such that $|\omega(x,y) - \omega(x',y')| > 0$, since we consider finite networks. Among such values, pick

$$\epsilon_{[L]} = \frac{1}{2}\min\left\{|\omega(x,y) - \omega(x',y')| : |\omega(x,y) - \omega(x',y')| > 0\right\}$$

Then, for f satisfying $|f(x',y')| \leqslant \epsilon_{[L]}$ for all $(x', y') \in X \times X$, we have that

$$\begin{aligned}
|f(x',y')(\omega(x,y) - \omega(x',y'))| &= |f(x',y')||\omega(x,y) - \omega(x',y')| \\
&\leqslant \epsilon_{[L]}|\omega(x,y) - \omega(x',y')| \\
&\leqslant \frac{1}{2}|\omega(x,y) - \omega(x',y')|^2,
\end{aligned}$$

and so the sum is non-negative. Next consider the labels

$$\sum_{X^2} \pi(x,x')\left[\frac{1}{2}\|\iota(x) - \iota(x')\|^2 - \langle g(x'), \iota(x) - \iota(x')\rangle\right].$$

Applying the same reasoning as previously, we note that if $\|\iota(x) - \iota(x')\|^2 = 0$ π-a.e. then the sum vanishes. Otherwise, there exists at least one $(x, x') \in X^2$ for which $\|\iota(x) - \iota(x')\|^2 > 0$, and pick

$$\eta_{[L]} = \frac{1}{2}\min\left\{\|\iota(x) - \iota(x')\| : \|\iota(x) - \iota(x')\| > 0\right\}.$$

Then for g such that $\|g(x')\| \leqslant \eta_{[L]}$ for $x \in X$, we have that

$$\langle g(x'), \iota(x) - \iota(x')\rangle \leqslant \|g(x')\|\|\iota(x) - \iota(x')\| \leqslant\leqslant \eta_{[L]}\|\iota(x) - \iota(x')\| \leqslant \frac{1}{2}\|\iota(x) - \iota(x')\|^2,$$

and this sum is also non-negative. □

Definition 4.4 (Logarithm map)**.** *Let $[L], [L'] \in [\mathcal{LP}_k]$ and let $(\pi_i) \in \Pi_k((\mu_i), (\mu_i'))$ be an optimal k-partitioned coupling of L to L'. Consider the representative $\hat{L} = ((\hat{X}_i, \hat{\mu}_i), \hat{\omega}, (\hat{\iota}_i)) \in [L]$ with*

$$\hat{X}_i = X_i \times X_i', \quad \hat{\mu}_i = \pi_i, \quad \hat{\omega}(x, x', y, y') = \omega(x, y), \quad \hat{\iota}_i(x, x') = \iota_i(x).$$

Similarly, define the representative $\hat{L}' = ((\hat{X}_i, \hat{\mu}_i), \hat{\omega}', (\hat{\iota}_i')) \in [L']$, where

$$\hat{\omega}'(x, x', y, y') = \omega'(x', y'), \quad \hat{\iota}_i'(x, x') = \iota_i'(x').$$

We define

$$\log^{\pi}_{[L]}([L']) = [\omega' - \omega, \iota' - \iota] \in T_{[L]}[\mathcal{LP}_k]. \tag{24}$$

It follows by definition of $\exp_{[L]}$ *that*

$$\exp_{[L]}(\log^{\pi}_{[L]}([L'])) = [L'].$$

Proposition 4.5. *Let $[L] \in [\mathcal{LP}_k]$ be a finite labelled k-partitioned measure network and let $\epsilon_{[L]}$ and $\eta_{[L]}$ be as in Proposition 4.3. The exponential map $\exp_{[L]}$ is injective on the set of tangent vectors admitting representations $(f, g) \in L^2(\mu \otimes \mu) \oplus \Lambda_X$ satisfying $|f(x, y)| < \epsilon_{[L]}/2$ and $|g(x)| < \eta_{[L]}/2$.*

Proof. Recall that $\exp_{[L]}([f, g]) = \exp_{[L]}([f', g'])$ if and only if $d_{\mathcal{LP}_k}(\exp_{[L]}([f, g]), \exp_{[L]}([f', g'])) = 0$. Let π be an optimal coupling between $\exp_{[L]}([f, g])$ and $\exp_{[L]}([f', g'])$. Then the corresponding distortion is

$$\begin{aligned}
&\frac{1}{2}\sum_{X^4} \pi(x, x')\pi(y, y')|\omega(x, y) + f(x, y) - \omega(x', y') - f'(x', y')|^2 + \frac{1}{2}\sum_{X^2} \pi(x, x')\|\iota(x) + g(x) - \iota(x') - g'(x')\|^2 \\
&= \sum_{X^4} \pi(x, x')\pi(y, y')\left[\frac{1}{2}|f(x, y) - f'(x', y')|^2 + \frac{1}{2}|\omega(x, y) - \omega(x', y')|^2\right. \\
&\qquad\qquad\qquad\qquad \left. + (\omega(x, y) - \omega(x', y'))(f(x, y) - f'(x', y'))\right] \\
&\qquad + \sum_{X^2} \pi(x, x')\left[\frac{1}{2}\|g(x) - g'(x')\|^2 + \frac{1}{2}\|\iota(x) - \iota(x')\|^2 + \langle \iota(x) - \iota(x'), g(x) - g'(x')\rangle\right]
\end{aligned}$$

Assuming we have that $|f(x, y)|, |f'(x, y)| < \epsilon_{[L]}/2$ and $|g(x)|, |g'(x)| < \eta_{[L]}/2$, then for all $(x, x', y, y') \in X^4$:

$$|f(x, y) - f'(x', y')| < \epsilon_{[L]}, \quad \|g(x) - g'(x')\| < \eta_{[L]}.$$

Then

$$\begin{aligned}
(\omega(x, y) - \omega(x', y'))(f(x, y) - f'(x', y')) &\leqslant |\omega(x, y) - \omega(x', y')||f(x, y) - f'(x', y')| \\
&\leqslant \epsilon_{[L]}|\omega(x, y) - \omega(x', y')| \\
&\leqslant \frac{1}{2}|\omega(x, y) - \omega(x', y')|^2,
\end{aligned}$$

and by the same reasoning we have that

$$\langle \iota(x) - \iota(x'), g(x) - g(x')\rangle \leqslant \frac{1}{2}\|\iota(x) - \iota(x')\|^2.$$

It follows that the second two terms in each of the sums are non-negative: $\frac{1}{2}|\omega(x, y) - \omega(x', y')|^2 + (\omega(x, y) - \omega(x', y'))(f(x, y) - f'(x', y')) \geqslant 0$ and $\frac{1}{2}\|\iota(x) - \iota(x')\|^2 + \langle \iota(x) - \iota(x'), g(x) - g'(x')\rangle \geqslant 0$.

As a result, $d_{\mathcal{LP}_k}(\exp_{[L]}([(f, g)]), \exp_{[L]}([(f', g')])) = 0$ implies that $|f(x, y) - f'(x', y')| = 0$ $\pi \otimes \pi$-a.e. and $\|g(x) - g'(x')\|^2 = 0$ π-a.e. Therefore we conclude that $[f, g] = [f', g']$. □

4.1.2 The unlabelled case

Recall from Section 3.1.2 that $[\mathcal{P}_k]$ can be considered as a subspace of $[\mathcal{LP}_k]$, where the attribute spaces Λ_i are 0-dimensional Hilbert spaces. Under this identification, the concepts and results from Section 4.1.1 can be specialized to $[\mathcal{P}_k]$. It is more convenient to express the specialized concepts directly in the notation of $[\mathcal{P}_k]$, rather than in the notation coming from the embedding $[\mathcal{P}_k] \hookrightarrow [\mathcal{LP}_k]$. For the sake of convenience, we summarize these expressions in the language of partitioned networks below.

Definition 4.6 (Tangent space for partitioned networks)**.** *We define the* ***synthetic tangent space*** *of $[\mathcal{P}_k]$ at a point $[P]$ to be*

$$T_{[P]}[\mathcal{P}_k] := \left(\bigcup_{((X_i, \mu_i), \omega) \in [P]} L^2(\mu \otimes \mu)\right) / \sim .$$

In the above, the union is taken over all k-partitioned measure networks $((X_i,\mu_i),\omega)$ in the weak isomorphism equivalence class $[P]$. The equivalence relation is defined as follows. For two representatives

$$((X_i,\mu_i),\omega)\,, \quad \big((X'_i,\mu'_i),\omega'\big) \in [P]$$

and functions

$$f \in L^2(\mu \otimes \mu), \quad f' \in L^2(\mu' \otimes \mu'),$$

we write $f \sim f'$ if and only if there exists a k-partitioned coupling $(\pi_i) \in \Pi_k(\mu,\mu')$ such that

$$f(x,y) = f'(x',y') \quad \textit{for} \quad \pi \otimes \pi - a.e.\ (x,x',y,y') \in X \times X' \times X \times X'.$$

The equivalence class of f is denoted $[f]$.

Definition 4.7 (Exponential map for partitioned networks)**.** *For a k-partitioned measure network $[P] \in [\mathcal{P}_k]$, let $[f] \in T_{[P]}[\mathcal{P}_k]$ be a tangent vector with $f \in L^2(\mu\otimes\mu)$. We define the* ***exponential map*** *by*

$$\exp_{[P]} : T_{[P]}[\mathcal{P}_k] \mapsto [\mathcal{P}_k], \quad \exp_{[P]}([f]) := [(X_i,\mu_i),\omega + f].$$

Definition 4.8 (Logarithm map for partitioned networks)**.** *Let $[P],[P'] \in [\mathcal{P}_k]$ and let $(\pi_i) \in \Pi_k((\mu_i),(\mu'_i))$ be an optimal k-partitioned coupling of P to P'. Consider the representative $\hat{P} = ((\hat{X}_i,\hat{\mu}_i),\hat{\omega}) \in [P]$ with*

$$\hat{X}_i = X_i \times X'_i, \quad \hat{\mu}_i = \pi_i, \quad \hat{\omega}(x,x',y,y') = \omega(x,y).$$

Similarly, define the representative $\hat{L}' = ((\hat{X}_i,\hat{\mu}_i),\hat{\omega}') \in [L']$, where

$$\hat{\omega}'(x,x',y,y') = \omega'(x',y').$$

We define

$$\log^{\pi}_{[P]}([P']) = [\omega' - \omega] \in T_{[P]}[\mathcal{P}_k]. \tag{25}$$

It follows by definition of $\exp_{[P]}$ that

$$\exp_{[P]}(\log^{\pi}_{[P]}([P'])) = [P'].$$

4.2 Gradients

Tasks in geometric statistics such as Fréchet means are often formulated in terms of minimization of functionals over a manifold [50]. To make sense of gradient flows, we need a notion of gradients. For simplicity, we will discuss the case of k-partitioned measure networks $\mathcal{P}_k$ (i.e., without labels). However, we remark that analogous results can be obtained for labelled graphs where labels are valued in Hilbert spaces.

Definition 4.9 (Gradients of functionals)**.** *Let $F : [\mathcal{P}_k] \to \mathbb{R}$ be a functional on the space of k-partitioned measure networks. For a network $[P] \in [\mathcal{P}_k]$ and a tangent vector $[f] \in T_{[P]}[\mathcal{P}_k]$, we define the* ***directional derivative*** *of F, if it exists, to be*

$$D_{[f]}F([P]) := \lim_{t\downarrow 0} \frac{F(\exp_{[P]}([tf])) - F([P])}{t}.$$

A functional F is said to be ***strongly differentiable*** *(following [65, Definition 6.23]) at a point $[P] \in [\mathcal{P}_k]$ if all of its directional derivatives exist, and if there exists a tangent vector $[g] \in T_{[P]}[\mathcal{P}_k]$ such that for any $[f] \in T_{[P]}[\mathcal{P}_k]$ and for every $(\pi_i) \in \Pi_k(\mu,\mu')$ such that $\|\omega - \omega'\|_{L^2((X\times X')^2,\pi\otimes\pi)} = 0$, it holds that*

$$D_{[f]}F([P]) = \langle f, g\rangle_{L^2(\pi\otimes\pi)}.$$

Here, $((X_i,\mu_i),\omega)$ and $((X'_i,\mu'_i),\omega')$ are two representatives of $[P]$, and $f \in L^2(X^2,\mu\otimes\mu)$ and $g \in L^2(X'^2,\mu'\otimes\mu')$ are representatives of $[f]$ and $[g]$ respectively. We then write $[\nabla F(P)] := [g]$ and refer to $[\nabla F(P)]$ as the ***gradient*** *of F at $[P]$.*

The following proposition characterizes some basic properties of the gradient. It uses the concept of the **norm** of a tangent vector. For $[f] \in T_{[P]}[\mathcal{P}_k]$, with $P = ((X_i,\mu_i),\omega)$ and $f \in L^2(\mu \otimes \mu)$, define

$$\|[f]\|_{T_{[P]}[\mathcal{P}_k]} := \|f\|_{L^2(\mu\otimes\mu)}.$$

The fact that this is well-defined follows easily from the nature of the equivalence relation used to construct the tangent space.

Proposition 4.10. *If $F : [\mathcal{P}_k] \to \mathbb{R}$ is strongly differentiable at $[P] \in [\mathcal{P}_k]$, then the gradient $[\nabla F(P)] \in T_{[P]}[\mathcal{P}_k]$ is unique and*

$$\|[\nabla F(P)]\|_{T_{[P]}[\mathcal{P}_k]} = \sup\{D_{[f]}F([P]) \;:\; [f] \in T_{[P]}[\mathcal{P}_k], \|[f]\|_{T_{[P]}[\mathcal{P}_k]} = 1\}.$$

Proof. This follows directly from [65, Lemma 6.24] for the case of partitioned measure networks by requiring couplings to respect partitions where necessary. □

4.3 Calculating gradients

Motivated by some practical applications, in this section, we compute expressions for gradients of two functionals defined over $[\mathcal{P}_k]$. Namely, we consider the Fréchet functional and its generalization to the problem of geodesic dictionary learning. We will put these expressions to use in Section 5.4, where we conduct some numerical computations with partitioned measure networks. Since in this section we focus on practical utility for numerical applications, some computations are done formally. A rigorous theoretical treatment of gradient flows has been addressed in the context of measure networks by Sturm [65]. In particular, a rigorous analysis of the dictionary learning problem may be a useful area for future study. More generally, the barycenter computation problem remains an active area of research even in the case of measures in $\mathbb{R}^d$ (e.g. [70, 2]).

Our gradient computations will make repeated use of the following result.

Proposition 4.11. *Let $[P], [P'] \in [\mathcal{P}_k]$ and let (π_i) be an optimal k-partitioned coupling. There exist representatives $\overline{P} \in [P]$ and $\overline{P}' \in [P']$ whose underlying sets and measures are the same and the diagonal couplings give an optimal k-partitioned coupling. If $[P]$ and $[P']$ are finite, then the representatives can also be taken to be finite.*

In particular, if (π_i) is an optimal k-partitioned coupling of P and P', then such representatives are given by

$$\overline{P} = \big((X_i \times X_i', \pi_i), ((x,x'),(y,y')) \mapsto \omega(x,y)\big) \quad \text{and} \quad \overline{P}' = \big((X_i \times X_i', \pi_i), ((x,x'),(y,y')) \mapsto \omega'(x',y')\big).$$

The proof is a straightforward verification that the proposed $\overline{P}$ and $\overline{P}'$ satisfy the conditions; see [22, Lemma 12] for details in the case of hypernetworks. When k-partitioned measure networks satisfy the conditions in the first paragraph, we say that the networks are **aligned**; the conclusion of the proposition is that, when considered up to weak isomorphism, we can assume without loss of generality that any pair of partitioned measure networks is aligned. Moreover, given a finite collection of partitioned measure networks $\{[P_i], 1 \leqslant i \leqslant N\}$, repeated application of the proposition shows that we can assume without loss of generality that each P_i, $i \geqslant 2$, is aligned to P_1.

4.3.1 Fréchet functional

Define the **Fréchet functional** for a finite collection of partitioned measure networks $[P_1], \ldots, [P_N] \in [\mathcal{P}_k]$ to be the maps $F : [\mathcal{P}_k] \to \mathbb{R}$ given by

$$F([R]) = \frac{1}{N}\sum_{i=1}^{N} d_{\mathcal{P}_k}([R],[P_i])^2. \tag{26}$$

Based on the discussion following Proposition 4.11, we can assume without loss of generality that R and the P_i are aligned, so that they are of the form $((X_i,\mu_i),\omega_R)$ and $((X_i,\mu_i),\omega_{P_i})$, respectively, and an optimal k-partitioned coupling of R to each P_i is given by diagonal couplings (Δ_{μ_i,μ_i}). We claim that, under the assumption that all networks $R, P_1, \ldots, P_N$ are finite, the gradient of F is represented by $\nabla F(R) \in L^2(X^2, \mu^{\otimes 2})$, with

$$\nabla F(R) = \omega_R - \frac{1}{N}\sum_{i=1}^{N} \omega_{P_i}. \tag{27}$$

The Fréchet functional was studied in [20] in the setting of measure networks. Although the case of partitioned measure networks can be treated by the same approach, we include a derivation of (27) for completeness. We first consider the setting where $N = 1$, in which case the Fréchet functional simplifies to

$$F([R]) = d_{\mathcal{P}_k}([P],[R])^2.$$

We fix two representatives $R = ((X_i, \mu_i), \omega_R)$ and $P = ((X_i, \mu_i), \omega_P)$ such that an optimal k-partitioned coupling between R and P is given by diagonal couplings $(\Delta_i) = (\Delta_{\mu_i,\mu_i})$. Following our usual convention, we let $\Delta = \sum_i \Delta_i$, considered as a measure on $\sqcup_i X_i$. Let $f \in L^2(X^2, \mu \times \mu)$ be a representative of a tangent vector in $T_{[R]}[\mathcal{P}_k]$. Then by definition of the directional derivative,

$$D_{[f]}F([R]) = \lim_{t \downarrow 0} \frac{F(\exp_{[R]}(t[f])) - F([R])}{t}.$$

By the proof of Proposition 4.3 (specialized to the case of partitioned networks), we may assume without loss of generality that (Δ_i) is an optimal k-partitioned coupling between $\exp_R(tf)$ and P, for t sufficiently small. Then, for small enough t,

$$\begin{aligned}
&\frac{1}{t}(F(\exp_R(tf)) - F(R)) \\
&\qquad = \frac{1}{t}\left(d_{\mathcal{P}_k}(\exp_R(tf), P)^2 - d_{\mathcal{P}_k}(R,P)^2\right) \\
&\qquad = \sum_{i=1}^{k} \frac{1}{t}\left(\frac{1}{2}\|\omega_R + tf - \omega_P\|^2_{L^2(\Delta_i^{\otimes 2})} - \frac{1}{2}\|\omega_R - \omega_P\|^2_{L^2(\Delta_i^{\otimes 2})}\right) \\
&\qquad = \sum_{i=1}^{k} \frac{t}{2}\|f\|^2_{L^2(\mu_i^{\otimes 2})} + \langle f, \omega_R - \omega_P\rangle_{L^2(\Delta_i^{\otimes 2})} + \frac{1}{2t}\left(\|\omega_R - \omega_P\|^2_{L^2(\Delta_i^{\otimes 2})} - \|\omega_R - \omega_P\|^2_{L^2(\Delta_i^{\otimes 2})}\right).
\end{aligned}$$

Taking $t \to 0$ yields

$$D_{[f]}F([R]) = \sum_{i=1}^{k} \langle f, \omega_R - \omega_P\rangle_{L^2(\Delta_i^{\otimes 2})} = \langle f, \omega_R - \omega_P\rangle_{L^2(\Delta^{\otimes 2})},$$

hence a representative of $[\nabla F(R)]$ is

$$\nabla F(R) = \omega_R - \omega_P \in L^2(\mu \otimes \mu).$$

The formula (27) follows by linearity.

4.3.2 Geodesic dictionary learning

Let $[P_1], \ldots, [P_N] \in [\mathcal{P}_k]$ be a collection of finite k-partitioned measure networks. We consider a generalization of the Fréchet mean, which seeks to find a **dictionary** of m **atoms** (i.e. representatives or archetypes) $[D_1], \ldots, [D_m] \in [\mathcal{P}_k]$ (each of which we assume to be finite) and a collection of N vectors in the $(m-1)$-dimensional **probability simplex** $\alpha_1, \ldots, \alpha_N \in \Delta^m$ (that is, each $\alpha_j \in \mathbb{R}^m$ has nonnegative entries which sum to one), that provide a useful set of reference points for summarizing the original dataset. In what follows, we give a brief heuristic derivation of gradient expressions which can be used for approximating such a dictionary (see Section 5.5 for applications involving such a derivation). A rigorous treatment of this difficult bi-level optimization problem is out of the scope for our present paper. We point out that, compared to the better-understood analogous problem of learning Wasserstein barycenters or dictionaries in Euclidean space [58, 70, 2], in our setting, issues such as uniqueness of barycenters have not been rigorously addressed.

We first informally define a **barycenter operator** to be any assignment taking a proposed dictionary $\{D_j\}_{j=1}^m$ together with an vector $\alpha \in \Delta^m$, whose entries are denoted $\alpha(1), \ldots, \alpha(m)$, to

$$B\left(\{D_j\}_{j=1}^m, \alpha\right) \in \arg\min_{R \in \mathcal{P}_k} \sum_{j=1}^m \alpha(j) d_{\mathcal{P}_k}(D_j, R)^2. \tag{28}$$

That is, there is not necessarily a unique minimizer, so a barycenter operator must involve a choice. In practice, barycenters are approximated by some algorithm, so the necessity of making a choice models a realistic situation (although the approximators are likely to only return a local minimizer).

Next, the loss functional for the **geodesic dictionary learning problem** is the bi-level optimization problem:

$$F\left(\{D_j\}_{j=1}^m, \{\alpha_i\}_{j=1}^N\right) = \frac{1}{N}\sum_{j=1}^N d_{\mathcal{P}_k}\left(B(\{D_j\}, \alpha_j), P_j\right)^2. \tag{29}$$

For practical purposes, we replace this with an easier problem by taking the following heuristic approximation of the barycenter operator. Given a dictionary $\{D_j\}$, a basepoint P and a weight vector α, we define the **local barycenter operator** to return $B_P(\{D_j\}, \alpha) = ((X_i, \mu_i), \omega_B)$, where it is assumed that all atoms have been aligned to $P = ((X_i, \mu_i), \omega)$, so that $D_j = ((X_i, \mu_i), \omega_{D_j})$, and ω_B is defined by

$$\omega_B = \sum_{j=1}^m \alpha(j)\omega_{D_j},$$

We then consider

$$F\left(\{D_j\}_{j=1}^m, \{\alpha_i\}_{j=1}^N\right) = \frac{1}{N}\sum_{j=1}^N d_{\mathcal{P}_k}\left(B_{P_j}(\{D_j\}, \alpha_j), P_j\right)^2.$$

To approximate this via gradient descent, we hold all arguments constant besides one and run a gradient step on the induced functional, and iterate this process through arguments. We claim that, for $\{\alpha_i\}_{i=1}^N$ held constant, the gradient of F in each of the D_i (that is, holding other atoms constant) is given by

$$\nabla_{D_i} F(D_i) = \frac{1}{N}\sum_{j=1}^m \omega_{D_j}\left(\sum_{\ell=1}^N \alpha_\ell(i)\alpha_\ell(j)\right) - \frac{1}{N}\sum_{\ell=1}^N \alpha_\ell(i)\omega_{P_\ell}. \tag{30}$$

where we have assume that all spaces are aligned to a common k-partitioned measures space, as in Proposition 4.11. Similarly, the Fréchet gradient in each of the α_ℓ is given by

$$\nabla_{\alpha_\ell} F(\alpha_\ell) = \frac{1}{N}\left[\langle \omega_{D_i}, \omega_B - \omega_{P_\ell}\rangle\right]_{i=1}^m. \tag{31}$$

Heuristic derivations of these expressions are provided below.

By linearity (as in the case of Fréchet means), let us take $N = 1$. Then the functional simplifies to

$$F\left(\{D_j\}_{j=1}^m, \alpha\right) = d_{\mathcal{P}_k}(B_P(D, \alpha), P)^2.$$

We wish to consider the functional induced by holding all but one of the atoms fixed; without loss of generality, suppose that only D_1 varies and $D_2, \ldots, D_m$ and α are fixed. Then we consider the functional

$$\tilde{F}(D_1) = F(\{D_i\}_{i=1}^m, \alpha).$$

Let $f \in L^2(X^2, \mu^{\otimes 2})$ be a tangent vector at D_1 and let $\omega_{B(t)}$ denote the network kernel of $\exp_{D_1}(tf)$—explicitly, $\omega_{B(t)} = \omega_B + \alpha_1 tf$. By a similar computation to the one used in the derivation of the gradient to the Fréchet functional, we have, for sufficiently small t (so that the ideas of Proposition 4.3 apply),

$$\frac{1}{t}(\tilde{F}(\exp_{D_1}(tf)) - \tilde{F}(D_1))$$

$$
\begin{aligned}
&= \frac{1}{t}\left(d_{\mathcal{P}_k}(B_P(\{\exp_{D_1}(tf), D_2, \dots, D_m\}, \alpha), P)^2 - d_{\mathcal{P}_k}(B_P(\{D_j\}, \alpha), P)^2\right) \\
&= \sum_{i=1}^{k} \frac{1}{t}\left(\frac{1}{2}\|\omega_{B(t)} - \omega_P\|^2_{L^2(\Delta_i^{\otimes 2})} - \frac{1}{2}\|\omega_B - \omega_P\|^2_{L^2(\Delta_i^{\otimes 2})}\right) \\
&= \sum_{i=1}^{k} \frac{1}{t}\left(\frac{1}{2}\|\omega_B + \alpha_1 tf - \omega_P\|^2_{L^2(\Delta_i^{\otimes 2})} - \frac{1}{2}\|\omega_B - \omega_P\|^2_{L^2(\Delta_i^{\otimes 2})}\right) \\
&= \sum_{i=1}^{k} \frac{\alpha_1^2 t}{2}\|f\|^2_{L^2(\mu_i^{\otimes 2})} + \alpha_1\langle f, \omega_B - \omega_P\rangle_{L^2(\Delta_i^{\otimes 2})} + \frac{1}{2t}\left(\|\omega_B - \omega_P\|^2_{L^2(\Delta_i^{\otimes 2})} - \|\omega_B - \omega_P\|^2_{L^2(\Delta_i^{\otimes 2})}\right).
\end{aligned}
$$

The claimed formula (30) then follows by a straightforward calculation.

When deriving the formula (31), we observe that this amounts to computing the derivative of a function defined on $\mathbb{R}^m$. The derivation then follows by elementary methods.

5 Applications and algorithms

In this section, we discuss a large number of applications using our formulation of the partitioned measure networks. In Section 5.1, we first give an overview of numerical algorithms involved in these applications. In Section 5.2, we discuss network matching and comparison using partitioned measure networks. We comment on the connection between measure network matching using Gromov-Wasserstein type distances and spectral network alignment such as EigenAlign (Section 5.2.1). We support our analysis using experiments involving synthetically generated graphs and hypergraphs (Section 5.2.2), real-world metabolic networks (Section 5.2.3), and multi-omics data (Section 5.2.4). In Section 5.3, we further expand the applications to study multiscale network matching using partitioned measure networks, by studying networks that arise from 3D mesh objects (Section 5.3.1) and protein-protein intersections (Section 5.3.2). In Section 5.4, we demonstrate via simple examples in computing geodesic interpolations between three hypergraphs and their barycenter. Finally in Section 5.5, we further extend the study of barycenter to the problem of dictionary learning using partitioned measure networks, that is, given an ensemble of partitioned measure networks, learn a basis such that each ensemble member could be described as a convex combination of the basis elements. We study nonlinear and linear dictionary learning in Section 5.5.1 and Section 5.5.2 respectively, with examples that arise from hypergraph stochastic block model (Section 5.5.3) and mutagenicity data (Section 5.5.4). For implementation details, see Appendix A.

5.1 Numerical algorithms

Practical applications of our framework in machine learning and statistics hinges upon numerical solution of quadratic programs that arise from the matching problem introduced in Definition 2.12 and its extensions. While significant progress has been made developing and analyzing numerical approaches for the case of Gromov-Wasserstein matchings (for measure networks) [52, 19, 67], co-optimal transport (for measure hypernetworks) [55, 69], and augmented Gromov-Wasserstein (for augmented measure networks) [26], our framework allows us to consider in generality (labelled) k-partitioned measure networks, from which each of these algorithms emerges as a special case. We provide a brief overview in what follows, and we defer technical details of specific algorithms to Appendix A due to space considerations.

As we are interested in numerical calculations related to generalized networks, we will assume all networks to be finite in this section and where appropriate use matrix notation to represent functions defined on finite spaces. That is, for a k-partitioned measure network $P = ((X_i, \mu_i), \omega)$ (see (2.7)) we take each X_i to be a finite set and denote its cardinality by $|X_i|$. Thus μ_i belonging to the probability simplex $\Delta^{|X_i|}$ (see Section 4.3.2) is a discrete probability distribution which we often write as a column vector, and ω a matrix of dimensions $| \sqcup_i X_i| \times | \sqcup_i X_i|$. Furthermore, we write

$$
\omega = \begin{bmatrix} \omega_{11} & \cdots & \omega_{1k} \\ \vdots & \ddots & \vdots \\ \omega_{k1} & \cdots & \omega_{kk} \end{bmatrix} = [\omega_{ij}]_{1 \leqslant i,j \leqslant k},
$$

where ω_{ij} corresponds to the restriction of ω onto $X_i \times X_j$, i.e. a submatrix of dimensions $|X_i| \times |X_j|$. In what follows, we use angle brackets to denote the Frobenius inner product between vectors or matrices, i.e. $\langle A, B \rangle = \mathrm{tr}(A^\top B)$.

We take $p = 2$ in the definition of the (labelled) partitioned network distance (see (3.3) and Definitions 2.12), as this allows for a efficient scheme for evaluating computationally the value of $\|\omega - \omega'\|^2_{L^2(\pi \otimes \xi)}$. For ease of notation, we will re-write the (squared) objective from (3):

$$\min_{\pi_i \in \Pi(\mu_i, \mu'_i), 1 \leqslant i \leqslant k} \frac{1}{2} \sum_{i,j=1}^{k} \|\omega - \omega'\|^2_{L^2(\pi_i \otimes \pi_j)} + \sum_{i=1}^{k} \|d_{\Lambda_i} \circ (\iota_i, \iota'_i)\|^2_{L^2(\pi_i)}. \tag{32}$$

Up to scaling (ω, ω') and d_{Λ_i} by constant factors, this is equivalent to the problem as written in (3) with $p = 2$. In particular, we note that the factor of $1/2$ is associated to terms quadratic in the π_i, which will simplify expressions later. Adopting the notation of [52], we define $L(\omega, \omega')$ to be the 4-way **distortion tensor**

$$L(\omega, \omega')_{ijkl} = \frac{1}{2} |\omega_{ik} - \omega'_{jl}|^2. \tag{33}$$

Introduce also for the labelled setting

$$C_i(x, x') = \frac{1}{2} d_{\Lambda_i}(\iota_i(x), \iota'_i(x'))^2, 1 \leqslant i \leqslant k \tag{34}$$

as cost matrices for matching labels in each label metric space Λ_i. Although our theoretical setup in Definition 3.1 assumes the existence of labelling functions ι_i and label metric spaces Λ_i, in practice our computations depend only on the matrices C_i and so the labellings are not made explicit. For instance, C_i may be constructed from kernels and thus understood to correspond to squared pairwise distances in a reproducing kernel Hilbert space.

Using the quantities we have now introduced, the problem (32) can be written as

$$\min_{\pi_i \in \Pi(\mu_i, \mu'_i), 1 \leqslant i \leqslant k} \frac{1}{2} \sum_{i,j=1}^{k} \langle L(\omega_{ij}, \omega'_{ij}), \pi_i \otimes \pi_j \rangle + \sum_{i=1}^{k} \langle C_i, \pi_i \rangle. \tag{35}$$

By setting appropriate terms to zero (following the lines of Definition 2.8), we can recover the optimal transport matching problems on generalized measure networks such as measure hypernetworks and measure networks (both labelled and unlabelled), as well as the standard optimal transport. Additionally, we can consider regularized variants of this problem which may yield favourable results in practice [24, 52, 7], both in terms of numerical solution schemes as well as properties of the solution:

$$\min_{\pi_i \in \Pi(\mu_i, \mu'_i), 1 \leqslant i \leqslant k} \frac{1}{2} \sum_{i,j=1}^{k} \langle L(\omega_{ij}, \omega'_{ij}), \pi_i \otimes \pi_j \rangle + \sum_{i=1}^{k} \langle C_i, \pi_i \rangle + \sum_{i=1}^{k} \varepsilon_i \Omega_i(\pi_i). \tag{36}$$

A common choice of Ω is the relative entropy $\Omega(\pi) = \mathrm{KL}(\pi | \mu \otimes \mu')$, which is consistent with the existing formulations of regularized co-optimal transport [55] and Gromov-Wasserstein transport [52].

Solving (35) (or (36)) amounts to finding minimizers of a (regularized) constrained, non-convex quadratic program in the couplings $(\pi_i)_{i=1}^k$. Naïve solutions of these problems using general purpose solvers is not scalable [32]. We develop several iterative algorithms to this end: since the problem (35) is non-convex, different choices of algorithm may converge to different minima. In summary, algorithmic approaches to solving (35) or (36) can be divided into (a) approaches relying on iterative solution of the standard linear optimal transport or Gromov-Wasserstein transport as algorithmic primitives, and (b) approaches based on gradient descent. We refer the interested reader to Appendix A for details. As our examples illustrate, the algorithm of choice depends heavily on the application at hand.

5.2 Network matching and comparison

In this section, we illustrate the utility of our theoretical and algorithmic framework via network matching and comparison. We first discuss a connection between Gromov-Wasserstein measure network alignment and a spectral network alignment method, as well as their respective generalizations to hypergraphs. Together with numerical results, we show that the optimal transport framework has a better behaviour, in terms of both accuracy and scalability.

We next consider an application to metabolic network alignment. We model this problem as one of labelled hypergraph matching (i.e. $k = 2$ for our partitioned setup), and solve an *unbalanced* transport problem due to the lack of a one-to-one matching between network elements. We find that, while incorporating label information alongside the hypergraph structure is essential to obtaining meaningful alignments, the hypergraph relational structure provides information that is crucial for refining the alignment. That is, incorporating the hypergraph structure improves significantly upon using labels alone.

Last, we turn to a problem of simultaneous sampling and feature alignment in multi-omics data, wherein networks are derived from general data matrices (see Example 2.2). This is a problem for which co-optimal transport and augmented Gromov-Wasserstein have been previously developed [55, 69, 26], viewing data matrices as hypergraphs where nodes are samples and hyperedges are features. These algorithms fall under our partitioned framework with $k = 2$. We show that partitioned networks are a flexible and more general tool for modelling multi-omics data, and results in improved alignment accuracy.

5.2.1 Relation to spectral network alignment

We first comment on the connection of Gromov-Wasserstein measure network matching to a (perhaps widely known) family of *spectral alignment* approaches. As introduced in Definition 2.6, for $p = 2$ and measure networks $(X, \mu, \omega), (X', \mu', \omega')$, the Gromov-Wasserstein (measure network) alignment problem is to solve

$$\min_{\pi \in \Pi(\mu, \mu')} \frac{1}{2} \langle L(\omega, \omega'), \pi \otimes \pi \rangle, \quad L(\omega, \omega')_{ijkl} = \frac{1}{2} |\omega_{ik} - \omega'_{jl}|^2, \tag{37}$$

which corresponds to partitioned measure network matching of Definition 2.12 with $k = 1$. This approach was studied in depth by [78] for network alignment. Spectral alignment methods are a family of approaches that have gained attention for graph alignments [29, 30, 46, 47] and also for hypergraphs [60]. Briefly, for two input graphs $G = (V, E), G' = (V', E')$, spectral network alignment seeks a node matching between V and V' that optimally preserves graph structure in a way similar to the Gromov-Wasserstein problem. This leads to a quadratic assignment problem (QAP), which upon being relaxed amounts to solving for the Perron-Frobenius eigenvector of a square matrix with dimensions $|V \times V'| \times |V \times V'|$ with all positive entries. We now make this problem description concrete. We abuse notation and also write G_{ij}, G'_{ij} to denote the (binary) adjacency matrices of the graphs G, G' respectively.

Feizi et al. [29] defined a matching score, for $(i, j), (k, l) \in (V \times V')^2$:

$$A_{ijkl} = \begin{cases} s_1 & (i,k) \in E \wedge (j,l) \in E'; \\ s_2 & (i,k) \notin E \wedge (j,l) \notin E'; \\ s_3 & \text{otherwise}. \end{cases} \tag{38}$$

Here, $s_1, s_2, s_3 > 0$. The first case corresponds to matching edges to edges (referred to as "matches" in [29]) with score s_1, the second case corresponds to matching non-edges with non-edges ("neutrals") with score s_2, and the final case corresponds to matching non-edges to edges, or vice versa ("mismatches") with score s_3.

It is immediately clear that A_{ijkl} plays the same role (but with opposite sign, since in [29] the aim is to *maximize* the matching score), as the tensor $L(G_{ik}, G'_{jl})$ in the Gromov-Wasserstein network alignment setting. While A_{ijkl} is a *matching score* (larger is better), L_{ijkl} is a *distortion* (smaller is better). The authors further derived an identity for A (Equation 3.3 of [29]):

$$A = (s_1 + s_2 - 2s_3)(G \otimes G') + (s_3 - s_2)(G \otimes \mathbf{1} + \mathbf{1} \otimes G') + s_2(\mathbf{1} \otimes \mathbf{1}),$$

which is also a convenient formula for the Gromov-Wasserstein setup, for computing $\langle A, x \otimes x \rangle$.

The graph alignment problem is then formulated as a QAP in terms of an unknown alignment matrix $y \in \mathbb{R}^{|V| \times |V'|}$ (which by an abuse of notation, we will also write as a vector of length $|V \times V'|$):

$$\begin{aligned} & \max_y y^\top A y \\ \text{s.t.} \qquad & y\mathbf{1} \leqslant 1, y^\top \mathbf{1} \leqslant 1, y \in \{0, 1\}. \end{aligned} \tag{39}$$

Since direct solution of this problem is intractable, Feizi et al. proposed an algorithm called *EigenAlign* which first solves a relaxation of (39) where the integer and row/column-sum constraints are replaced with non-negativity and unit-ball constraints:

$$\begin{aligned} & \max_y y^\top A y \\ \text{s.t.} \qquad & y \geqslant 0, \|y\|_2 \leqslant 1. \end{aligned} \tag{40}$$

The solution to this problem is shown to be v, the Perron-Frobenius eigenvector of the positive alignment score matrix A. In the second step of the algorithm, the solution v to the relaxed problem (40) is projected back onto the constraint set by solving a linear assignment problem:

$$\begin{aligned} & \max_y v^\top y \\ \text{s. t.} \qquad & y \geqslant 0 : y\mathbf{1} \leqslant 1, y^\top \mathbf{1} \leqslant 1, y \in \{0,1\}. \end{aligned} \tag{41}$$

This can be understood as solving for y that maximizes its similarity to the relaxed solution v that also satisfies the bijectivity constraints. The objective function for the Gromov-Wasserstein alignment problem has the exact same form as (39), since $\langle L(G,G'), \pi \otimes \pi \rangle = \operatorname{vec}(\pi)^\top \operatorname{mat}(L(G,G')) \operatorname{vec}(\pi)$.

The correspondence between the Gromov-Wasserstein (37) and spectral alignment problems (39, 40, 41) has some subtlety. Under the simplex constraint $\pi \in \Pi(\mu,\mu')$ the problem (37) is invariant to constant shifts in the distortion tensor L, since for any $c \in \mathbb{R}$,

$$\langle L(\omega,\omega') + c, \pi \otimes \pi \rangle = \langle L(\omega,\omega'), \pi \otimes \pi \rangle + c.$$

For binary incidence matrices $\omega_{ij}, \omega'_{ij} \in \{0,1\}$, we can introduce a *shifted* version of the distortion tensor, where $\eta > 0$ is a small constant:

$$\overline{L}(\omega,\omega')_{ijkl} = L(\omega,\omega)_{ijkl} - \frac{1}{2} - \eta.$$

Then $\overline{L}$ is strictly negative. We can therefore re-write the problem (37) equivalently as

$$\min_{\pi \in \Pi(\mu,\mu')} \langle \overline{L}, \pi \otimes \pi \rangle = \operatorname{vec}(\pi)^\top \operatorname{mat}(\overline{L}) \operatorname{vec}(\pi) \iff \max_{\pi \in \Pi(\mu,\mu')} \operatorname{vec}(\pi)^\top \operatorname{mat}(-\overline{L}) \operatorname{vec}(\pi).$$

Therefore, we may choose $A = \operatorname{mat}(-\overline{L})$ in (39, 40, 41) since it is a positive matrix. This corresponds to $s_1 = s_2 = 1/2 + \eta$ and $s_3 = \eta$. On the other hand, crucially the objective (40) is *not* invariant under additive shifts to the matrix A, since

$$y^\top (A + c\mathbf{1}\mathbf{1}^\top) y = y^\top A y + c|\mathbf{1}^\top y|^2$$

and $|\mathbf{1}^\top y|$ is not constant on the 2-norm ball. Therefore, while additive shifts of the distortion tensor leave the Gromov-Wasserstein problem (37) unchanged, different choices of the shift lead to different relaxed spectral problems (40). We remark that, since the Perron-Frobenius theorem restricts EigenAlign to positive alignment matrices, one cannot straightforwardly take $A = -\operatorname{mat}(L)$.

The main remaining difference between Gromov-Wasserstein and EigenAlign lies in the constraints: inequality constraints on the row and column sums of π are replaced instead with equality constraints. When $|V| = |V'|$ and node weights are chosen to be uniform, this amounts to the set of bi-stochastic matrices. In a sense, the relaxed problem solved by Gromov-Wasserstein departs less from (39) than EigenAlign. Noting that $\Pi(\mu,\mu') \subseteq \operatorname{Prob}(X \times X')$ and that $\{x \in \mathbb{R}^k : |x| = 1, x \geqslant 0\} \subseteq \{x \in \mathbb{R}^k : \|x\| \leqslant 1, x \geqslant 0\}$, the spectral problem solved by EigenAlign is in fact itself a relaxation of the corresponding Gromov-Wasserstein problem. Together with the observation that Gromov-Wasserstein finds a solution in a single step while EigenAlign requires two consecutive steps, this suggests that Gromov-Wasserstein network alignment may behave more favorably since the matching constraints are retained throughout the algorithm and can better inform the alignment.

This spectral alignment framework can be extended to the problem of hypergraph alignment [60, 38, 45], although hypergraphs introduce the additional complication that in general, hyperedges of a hypergraph may have edges of differing degree. For the simpler case of K-uniform hypergraphs (hypergraphs in which each hyperedge spans exactly K nodes), the matching score matrix A_{ijkl} can be extended to a matching score *tensor* $A_{(i_1,j_1),\dots,(i_K,j_K)}$ which has dimensions $|V \times V'|^K$. Writing y as a $|V \times V'|$ matching vector, a generalized matching objective is

$$\begin{aligned} &\max_y \left\langle A, \otimes_{i=1}^K y \right\rangle \\ \text{s. t.} \quad &y\mathbf{1} \leqslant 1, y^\top \mathbf{1} \leqslant 1, y \in \{0, 1\}. \end{aligned} \tag{42}$$

In [60], this problem is tackled in an analogous way to the EigenAlign algorithm (39, 40, 41), that is, a relaxation of (42) onto the unit norm ball is derived which amounts to a generalized tensor eigenproblem which can be approximately solved using higher-order power iterations [36]. This is then projected back onto the constraint set by solving a linear assignment problem. Non-uniform hypergraphs are converted to uniform hypergraphs by introducing a dummy node repeatedly to hyperedges as needed until all hyperedges have the same degree. In contrast, co-optimal transport-based matchings of hypergraphs still boils down to a quadratic problem (as opposed to higher-order) in the coupling π, regardless of hypergraph degree. Furthermore, optimal transport handles non-uniform hypergraphs naturally.

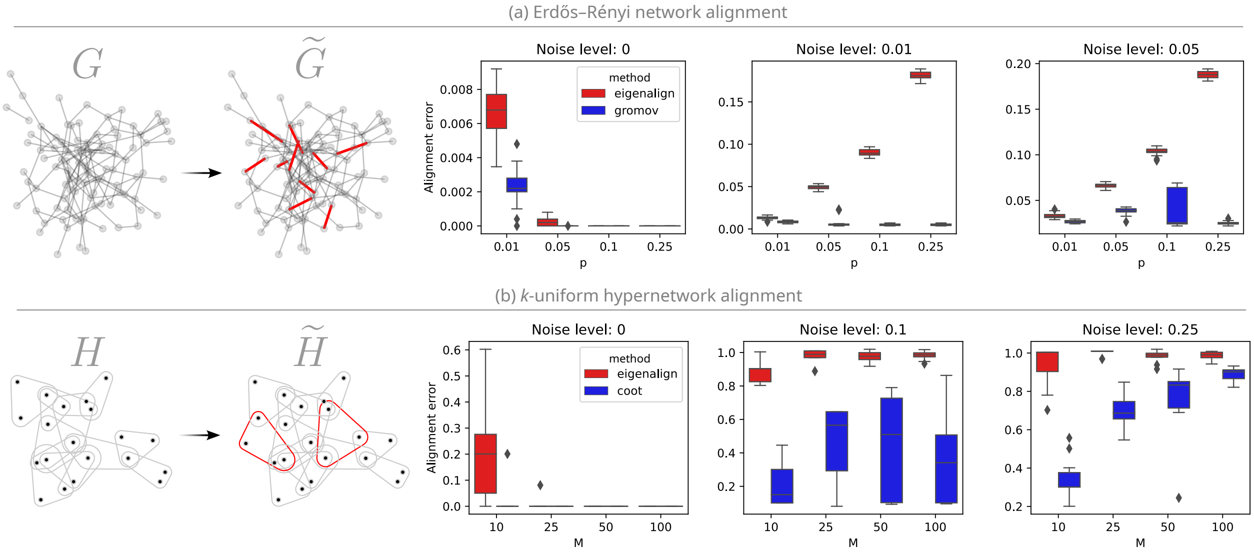

Figure 2: (a) Illustration of Erdös-Rényi random graph alignment problem with noise (edges due to noise shown in red); alignment errors (measured in terms of the distortion functional (37)) achieved by EigenAlign and Gromov-Wasserstein under permutation and noise. (b) Illustration of random 3-uniform hypergraph alignment problem with noise (hyperedges due to noise shown in red); alignment errors (measured in terms of the objective (42)) achieved by higher-order EigenAlign and COOT under permutations and noise.

5.2.2 Comparison to spectral network alignment for random graphs and hypergraphs

For this first set of experiments, we use synthetic datasets of graphs and hypergraphs. In Figure 2(a) we investigate the relative performance of spectral alignment and Gromov-Wasserstein alignment, considering Erdös-Rényi (ER) graphs of size $N = 100$ with parameter $p \in \{0.01, 0.05, 0.1, 0.25\}$. For a randomly sampled ER graph G, we form a copy $\widetilde{G}$ in which nodes have been relabelled via a random permutation. Optionally, we also add noise in the form of random addition or deletion of edges independently with probability $q \in \{0, 0.01, 0.05\}$. We align G to $\widetilde{G}$ using both the implementation of EigenAlign from [29] and Gromov-Wasserstein using a proximal gradient algorithm (see Algorithm 3), similar to the approach taken by [79]. Since the proximal gradient algorithm yields a coupling that is dense but potentially vanishingly small for most entries (i.e. strictly on the interior of the constraint set), we apply a "rounding" of the result onto an extreme point of the coupling polytope to yield a sparse permutation matrix. For each alignment, we measure the *alignment error* by calculating the corresponding distortion functional (37) to measure the alignment quality. In the absence of noise, $\widetilde{G}$ and G are isomorphic since they are represented by adjacency matrices that are identical up to permutation, and a distortion of zero corresponds to a perfect matching. Non-zero noise breaks this isomorphism (so that the ground truth node matching may no longer be the "right" one after adding noise), so the lower the distortion the better the alignment. In this sense, the distortion is an objective measure of alignment quality rather than the coupling itself. In all cases we consider, we find that Gromov-Wasserstein finds an alignment that yields a lower distortion than EigenAlign, shown in Figure 2(a). At a conceptual level, this can be understood since the Gromov-Wasserstein problem arises as a relaxation of the quadratic assignment problem (39) that accounts for the quadratic objective and the assignment constraints jointly, whereas the EigenAlign approach adopts a two step approach, first relaxing the assignment constraint to a norm ball constraint (40) and then projecting back onto the assignment polytope (41). Because of this, the assignment constraints in the second step cannot inform the quadratic program in the first step.

In Figure 2(b) we turn to hypergraph alignments. For hypergraphs, the scope of the higher-order spectral alignment approach is limited to dealing with uniform hypergraphs, and furthermore the time and space complexity scale exponentially in the order of the hypergraph. We therefore consider random 3-uniform graphs for $N = 25$ nodes and $M = \{10, 25, 50, 100\}$ hyperedges. Each hyperedge is obtained by sampling 3 nodes uniformly without replacement from the node set. Given a hypergraph H, we form a copy $\widetilde{H}$ by randomly relabelling nodes and hyperedges, and then replacing a fraction $q \in \{0, 0.1, 0.25\}$ of hyperedges with independently sampled hyperedges. The spectral alignment approach only aligns nodes (since for uniform hypergraphs a node alignments also induces hyperedge alignments), so we quantify the quality of alignments in terms of the objective of (42) rather than the co-optimal transport distortion which depends on both node and hyperedge couplings. As in the graph alignment case, we find that co-optimal transport alignments (using again the proximal method of Algorithm 3) perform as well or better compared to spectral alignments in all cases.

Spectral hypergraph alignments are restricted to uniform hypergraphs and are computationally expensive, while co-optimal transport does not have these limitations. Measuring the computation time for spectral alignment and co-optimal transport alignment for $N \in \{5, 10, 25\}$, we find that spectral alignment is several orders of magnitude more expensive in terms of runtime. Runs for $N = 50$ with spectral alignment have failed due to memory usage exceeding the available 32 GB.

5.2.3 Metabolic network alignment

Metabolic networks (and chemical reaction networks more generally) are an example of systems in which higher-order relations are essential to retain information: chemical species may be modelled as nodes and reactions as hyperedges, which may involve any number of reactants simultaneously [33]. We consider the metabolic networks of *E. coli* and halophilic archaeon DL31, retrieved from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [34] with accession numbers `eco01100` and `hah01100` respectively. We model each metabolic network as a labelled measure hypernetwork, where nodes are identified with metabolite compounds and hyperedges are identified with enzymes which catalyze reactions involving multiple compounds (multiple reactants and products). For simplicity, we discard directionality information and model the metabolic networks as undirected hypergraphs (i.e. we do not distinguish between reactants and products within each hyperedge). For `eco01100` (the source network) we construct a measure hypernetwork with 984 metabolites and 1005 reaction terms, and for `hah01100` (the target network) a measure hypernetwork with 679 metabolites and 558 reaction terms. We find that the minimum and maximum hyperedge sizes are 2 and 9, respectively, in both the source and target hypergraphs. This verifies the heterogeneous, non-uniform nature of these hypernetworks. We visualize each network in Figure 3(a), showing the associations between compounds (nodes, red) and reactions (hyperedges, blue). In contrast to the previous synthetic example, we now must align two hypergraphs that are non-uniform and different in size. Within our framework, the unbalanced, fused hypergraph alignment scheme is the most suitable approach and we demonstrate the effectiveness of this method.

As we mentioned previously in Section 5.2.1, this hypergraph alignment problem was addressed using a spectral approach by [60]. As the metabolic hypernetworks are non-uniform, dummy nodes are added to produce a d-uniform hypergraph, where d is the maximum hyperedge degree in the original non-uniform hypergraph. This uniform hypergraph is then represented as a $|V|^d$ adjacency tensor. Due to the size of the alignment tensor, in [60] a fairly involved computational scheme is described that exploits its symmetry

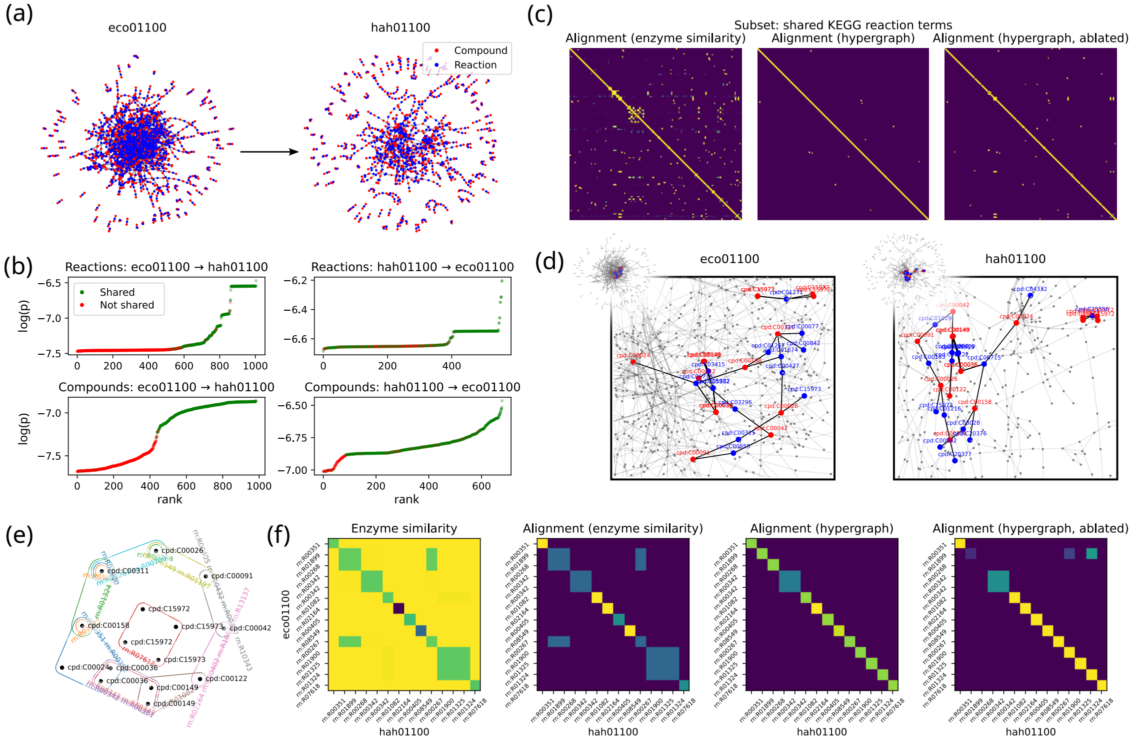

Figure 3: (a) Genome-scale metabolic networks of *E. coli* (`eco01100`) and halophilic archaeon DL31 (`hah01100`). (b) Nodes (compounds) and hyperedges (enzymes) ranked by total out-going probability mass as found by unbalanced alignment, coloured by whether its true match is shared in the target network. (c) Hyperedge couplings for the subset of reaction terms (enzymes) common to both organisms, found using (left) enzyme similarity only, (middle) labelled hypergraph alignment (with both metabolite and enzyme similarities provided), and (right) ablated, labelled hypergraph alignment (with metabolite similarities provided, but not enzyme similarities). (d) Zoom-in on conserved tricarboxylic acid (TCA) cycle subnetwork as shown in genome-scale metabolic network layout from (a). (e) Hypergraph layout of TCA cycle subnetwork, shown as a rubber band diagram. (f) Alignment of reaction terms in TCA cycle subnetwork, from left to right: enzyme similarity matrix for TCA cycle reaction terms; and alignments found using enzyme similarity, labelled hypergraph alignment (with both metabolite and enzyme similarities provided), and ablated labelled hypergraph alignment (with metabolite similarities provided, but not enzyme similarities), respectively.

properties. Even so, distributed computing is necessary to speed up the alignment, which was reported to take over two hours to match the two networks (559 metabolites and 537 reactions for `hah01100`, 794 metabolites and 923 reactions for `eco01100`) [60, Supplementary Materials].

For metabolite compounds and reaction terms, we construct pairwise cost matrices between the source and target using a similar approach to [60]. For metabolites, similarity scores are calculated using the cheminformatics package ChemmineR [14]. For any two enzymes (e_1, e_2), the similarity score is taken to be $1/N(e_1, e_2)$ where $N(e_1, e_2)$ is the number of enzyme entries in the lowest common level in the Enzyme Commission (EC) classification [3]. Since there is not a one-to-one correspondence between the two metabolic networks, we solve an unbalanced variant (see Section A.5) of the labelled hypergraph alignment problem between the two hypernetworks using a proximal gradient variant of Algorithm 5, using $\alpha = 0.9, \varepsilon = 10^{-3}, \lambda = 0.1$ and 250 and 1000 inner and outer iterations respectively. This takes less than a minute utilizing 4 cores of an Intel Xeon Gold 6242 system, several orders of magnitude faster than the higher-order spectral method employed by [60].

To assess the quality of the matching, we use the fact that compounds and enzymes conserved between these two organisms are known. For a subset of nodes and hyperedges therefore, we have a biological "ground truth" correspondence to compare against. Since only parts of the two metabolic networks are shared, we expect the unbalanced matching to reflect this and assign more mass to shared components. In Figure 3(b) we show reactions and compounds in each organism, ranked by the (log) total mass assigned to it by the unbalanced matching algorithm. Components which have a true match in the target network are shown in green: it is clear that more mass is assigned to components with a true match, and non-overlapping components tend to be down-weighted. These weights can be thought of as a measure of confidence in the alignment. We remark that the `eco01100` network is larger, so a significant fraction of its components do not have true matches in the `hah01100` network. Despite this, we still observe a separation in `hah01100` between components with and without true matches.

To understand how our alignment method depends on the node and hyperedge labels and relational information encoded in the hypergraph structure, we perform two additional alignments where some of this information is hidden (i.e., ablation study). To study the performance of alignment using hyperedge label information alone without the hypergraph structure, we directly align enzymes (hyperedges) using the enzyme similarity matrix by solving a standard optimal transport problem using the proximal point method [76]. We also consider hyperedge alignment using node labels and the hypergraph structure, when hyperedge labels are hidden. To do this, we set the hyperedge-hyperedge cost matrix to zero and recompute the alignment with the same parameters using only the node-hyperedge incidence matrix and the node-node cost matrix. In Figure 3(c), for the subset of enzymes or reaction terms shared between both organisms, we show alignments obtained using only hyperedge labels *(enzyme similarity)*, the full labelled hypergraphs *(hypergraph)*, and only node labels *(hypergraph, ablated)*. Compared to the full labelled hypergraph alignment result, we find that using only enzyme similarity leads to a much more noisy alignment, with large amounts of mass assigned away from the diagonal. Suppressing hyperedge labels leads to a slightly worse alignment of hyperedges compared to the full labelled alignment, but still significantly cleaner than using hyperedge labels alone.

Finally, we focus on the tricarboxylic acid (TCA) cycle, a fundamental metabolic process that is conserved between both organisms. In Figure 3(d) we highlight this subnetwork, and in Figure 3(e) we show its rubber band visualization. In Figure 3(f), we find that a full labelled hypergraph alignment near-perfectly matches the components, while the ablated hypergraph alignment without hyperedge labels again does slightly worse. In contrast, the enzyme similarity score does not provide full information about the matching, and hence alignment based on enzyme similarity alone performs much worse. Together, these results indicate that utilizing the hypergraph structure in combination with label information is crucial for achieving a good alignment between the two metabolic networks, outperforming alignments where either label or relational information are suppressed.

5.2.4 Multi-omics sample and feature alignment

Co-optimal transport has previously been employed for simultaneously matching samples and features between heterogeneous datasets [55]. One particularly popular example is that of multi-omics datasets, where two or more sets of features (e.g. gene expression and protein marker expression) are observed in samples (cells) [69]. This problem can be cast in the setting of hypernetwork alignment by interpreting samples and features as nodes and hyperedges respectively, and the sample-by-feature data matrix as the membership function. In [69] the application of unbalanced co-optimal transport was demonstrated to improve alignment quality, and in [26] the augmented Gromov-Wasserstein matching (see Definition 2.6) is introduced: this corresponds to partitioned network matching when only node-node information is provided in addition to node-hyperedge relations, but not hyperedge-hyperedge information. Here we consider $1,000$ cells sampled from the same CITE-seq dataset as in [26], in which 15 genes and their corresponding marker proteins were measured.

Partitioned measure networks allow pairwise relations *within* as well as *between* partitions to be modelled, so we incorporate pairwise similarities in each domain. This is in addition to the sample-feature information contained directly in the data matrix. We choose to capture this by calculating sample-sample and feature-feature correlations. Under the partitioned measure network alignment framework, we expect that pairwise structure between samples (respectively features) should be preserved by the alignment. To get an initial understanding of the data, we show the PCA embeddings and the partitioned network data constructed from each modality in Figure 4(a): $\omega_{00}^{(\mathrm{RNA},\mathrm{ADT})}, \omega_{11}^{(\mathrm{RNA},\mathrm{ADT})}$ are correlation matrices between samples and features respectively, and $\omega_{01}^{(\mathrm{RNA},\mathrm{ADT})}$ are the data matrices capturing cell-wise gene or marker expression. Put simply, we expect that correlated (anti-correlated) pairs of genes should be matched to correlated (anti-correlated) pairs of proteins, and similarly for cells between modalities.

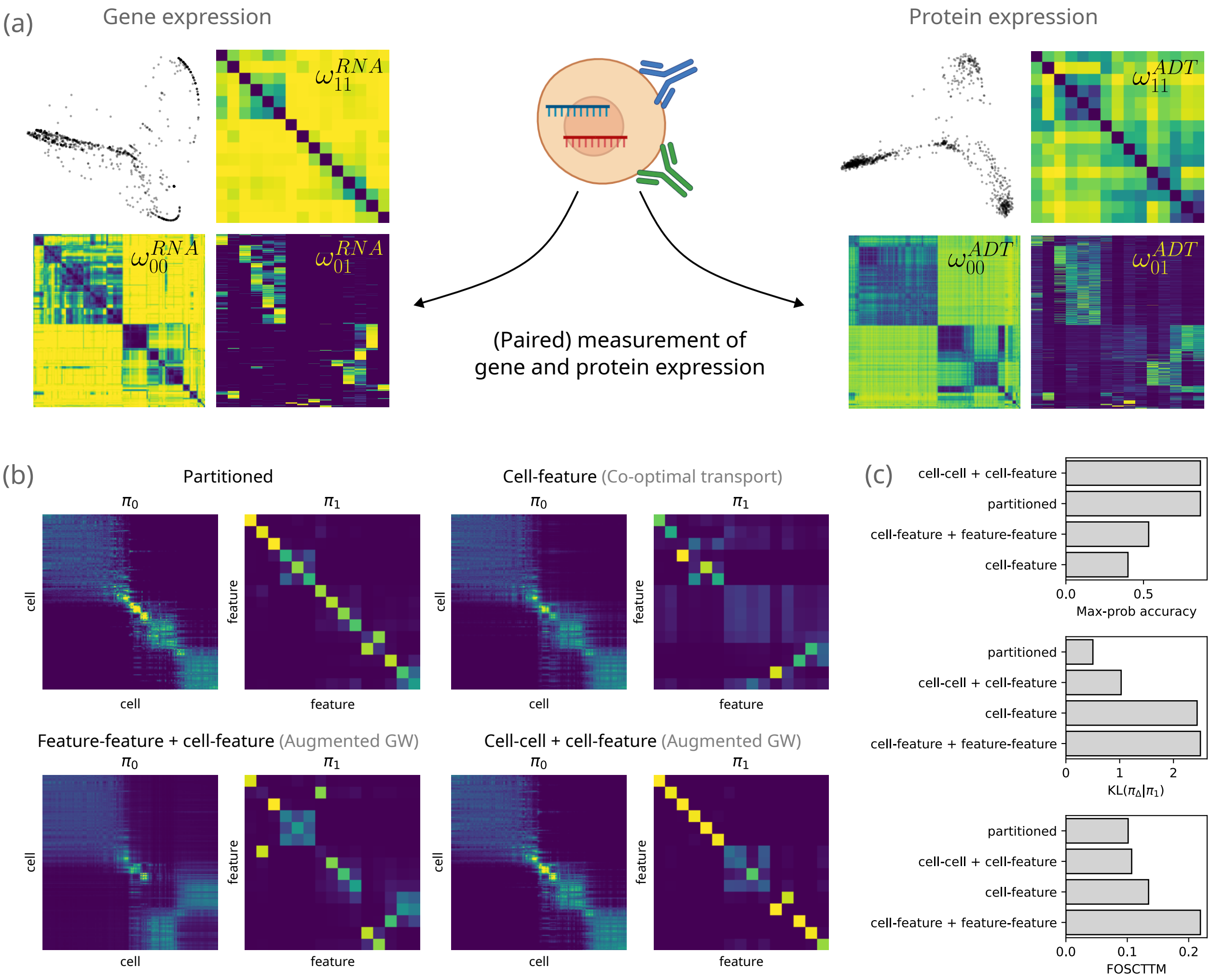

Figure 4: (a) Illustration of CITE-seq dataset shown in transcriptomic (gene expression or RNA) and surface marker (protein expression or ADT) modalities. For modality m, $\omega_{01}^m, \omega_{00}^m, \omega_{11}^m$ denote the data matrix, sample-sample similarities, and feature-feature similarities, respectively. (b) Sample and feature alignments obtained using partitioned network matching, co-optimal transport, and augmented Gromov-Wasserstein with sample-sample information and feature-feature information respectively. The ground truth matching for samples and features corresponds to the diagonal. (c) Quality of feature alignments in terms of (top) maximum probability assignment (higher is better); (middle) reverse KL-divergence to the diagonal coupling (lower is better); and (bottom) quality of sample alignments in terms of FOSCTTM (lower is better).

We then compute an entropy-regularized alignment of the RNA and protein partitioned measure networks using Algorithm 4. We introduce a parameter $\alpha \in [0, 1]$ to control the trade-off between the contribution of Gromov-Wasserstein type terms $(\omega_{00}, \omega_{11})$ and co-optimal transport type terms (ω_{01}), scaling these inputs by $\sqrt{\alpha}$ and $\sqrt{1-\alpha}$ respectively. We choose values $\alpha \in \{0, 0.1, \ldots, 0.9, 1\}$. For each partition, different levels of entropic regularization $\varepsilon_{0,1} \in \{\texttt{5e-4}, \texttt{1e-3}, \texttt{5e-3}, \texttt{1e-2}, \texttt{5e-2}, \texttt{0.1}, \texttt{0.5}\}$ are used, as it is well known that regularization level may play a role in the alignment quality [26]. Finally, we consider the special cases where pairwise information (i.e. the Gromov-Wasserstein term) on samples, features, or both, are suppressed. We implement this by setting $\omega_{00}^{\mathrm{RNA,ADT}}$, $\omega_{11}^{\mathrm{RNA,ADT}}$ to zero as needed.

For each set of parameter values we compute the alignment, and then calculate the fraction of gene transcripts which are correctly matched to their corresponding protein in terms of maximum assigned probability. We show in Figure 4(b) the best matchings obtained for the settings of partitioned matching, co-optimal transport, and augmented Gromov-Wasserstein (AGW) on samples and features respectively. In terms of identification of features (Figure 4(c)), we found that the partitioned alignments and AGW with sample-sample information were both able to correctly assign 13/15 (87%) features, in terms of maximum probability. On the other hand, AGW with feature-feature information and co-optimal transport correctly assigned 8/15 and 6/15 features respectively. While the fraction of correctly matched features by maximum probability gives an indication of the alignment accuracy, it does not account for the level of uncertainty in the matching. To account for this, we also calculate the KL divergence of the diagonal (ground truth) coupling relative to the alignment, reasoning that alignments that produce the correct matching with a higher certainty should have a lower divergence (i.e., lower divergence is better). We find that the partitioned matching produces a more informative alignment (KL = 0.502) compared to AGW (KL = 1.027), which can also be assessed visually from the couplings. Finally, in Figure 4(c) bottom, we assess the quality of sample matchings in terms of the fraction of samples closer than true match (FOSCTTM) which is a standard performance metric in the single cell alignment literature [26], for which a lower value indicates a better alignment. Both the partitioned alignment and AGW with sample-sample information produce sample alignments of similar quality, whereas AGW with feature-feature information and co-optimal transport have worse performance.

5.3 Partitioned networks for multiscale network matching

Whereas the previous examples focus on hypergraphs (i.e., partitioned networks with $k = 2$), our framework can be used to model multiscale data by setting $k \geqslant 2$. This insight was obtained in [22]: a multi-scale graph with k simplification levels can be modelled as k coupled hypergraphs. In Section A.6, we show how the work of [22] can be framed and extended in terms of k-partitioned measure networks. Specifically, we can model relations between nodes in the same simplification level, as well as between simplification levels using partitioned measure networks. We demonstrate the application of our framework for matching geometric networks (obtained from 3D objects), as well as non-geometric protein-protein interaction networks.

5.3.1 Multiscale point cloud matching

We apply multiscale matching to networks derived from 3D models of a wolf and a centaur from the TOSCA object database [12]. In [22], co-optimal transport was employed to find semantic matchings between two poses of the centaur graph across multiple scales. In their experiments, the two poses of the centaur graph have the same number of nodes at each level and are nearly identical in structure. The co-optimal transport framework of [22] is also applicable to finding semantic matchings between a wolf and a centaur, where the two graphs are significantly different in their size, connectivity, and semantic components. Indeed, the "true" semantic correspondence between the objects is not one-to-one, since the wolf has four limbs and the centaur has six. In this section, we solve the multiscale object matching problem using a partitioned measure network formulation and compare it against previous approaches.

For each input graph, a multi-scale topological simplification was produced using the heat kernel multiscale reduction of [22, Section 5.3] with $k = 3$ simplification levels. We take each of the $\omega_{i,i+1}, \omega'_{i,i+1}$, for $0 \leqslant i < 2$ to be binary incidence matrices of node-hyperedge relations between successive reductions. Pairwise relations $\omega_{ii}, \omega'_{ii}, 0 \leqslant i < 3$ are constructed from the graph shortest path distances on each simplification level.

In Figure 5 we visualize the alignments obtained using multiple methods: (1st row) multiscale COOT (using the algorithm of [22]); (2nd row) Gromov-Wasserstein measure network matching obtained independently on each simplification level; (3rd row) the multiscale alignment using k-partitioned networks (Algorithm 4) and (4th row) its unbalanced variant (Algorithm 6). At each simplification level, each node in the centaur graph is connected to a node in the wolf graph with the maximum matching probability. While multiscale COOT finds a matching that is largely consistent across scales, the matching is very noisy especially at fine scales as evidenced by discontinuities in the colour gradient in the centaur. On the other hand, independent Gromov-Wasserstein matchings at each simplification level leads to more locally consistent matchings (continuous colour gradients), but the matching fails to be consistent across scales; this is apparent, for instance, by looking at the matchings of the head/neck regions across simplification levels (blue arrows), as well as the arms of the centaur. The multiscale partitioned network alignment produces a semantically reasonable, consistent matching at each scale, correctly matching the head, neck, hind legs, as well as tail regions. Due to the difference in the number of limbs across each model, we observe mismatches among the arms of the centaur, which is not entirely surprising. Interestingly, a small portion of the front legs of the centaur model

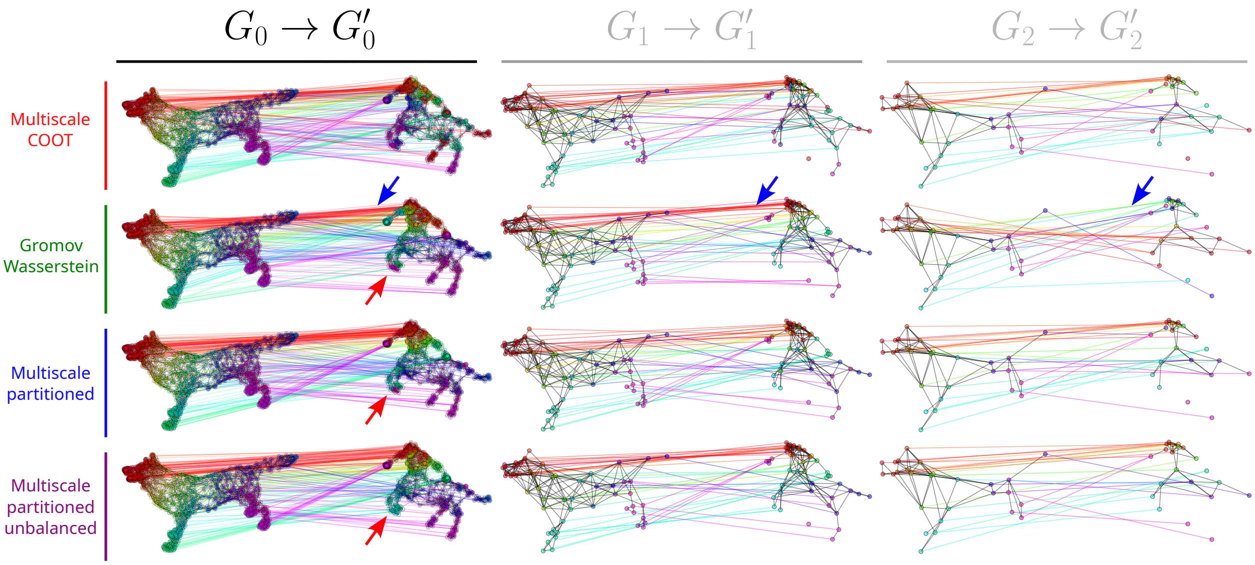

Figure 5: Multiscale network matchings: TOSCA wolf and centaur. Each row corresponds to a different algorithm, and each column corresponds to a graph simplification level. 1st row: multiscale COOT [22]. 2nd row: Gromov-Wasserstein measure network coupling obtained independently on each simplification level. 3rd and 4th row: the multiscale alignment using k-partitioned networks (3rd row) and its unbalanced variant (4th row). Blue arrows: matching of the wolf's head is inconsistent across scales for Gromov-Wasserstein alignment. Red arrows: discontinuity in centaur's front leg disappears with an unbalanced alignment.

is matched to the hind legs of the wolf (red arrows), seen as a discontinuity in the colour gradient. However, these mismatches disappear in the unbalanced partitioned measure network alignment. In other words, there is a lack of one-to-one correspondence between the limbs of a centaur and those of a wolf causing a number of mismatches, and unbalanced matching may alleviate these issues.

In addition to the visual assessment of semantic matchings, we show in Table 1(a) different components of the objective (82) for the matchings found by multiscale COOT, Gromov-Wasserstein, and partitioned measure network alignment, respectively. This table provides us with an unbiased quantification of alignment quality directly in terms of the distortion. These results confirm our visual observations from Figure 5: Gromov-Wasserstein alignment at each scale produces the minimal Gromov-Wasserstein loss reflecting preservation of pairwise relations at each individual scale, but a very high COOT loss indicates a lack of consistency across scales. Conversely, multiscale COOT minimizes the COOT loss while producing the highest Gromov-Wasserstein loss, which suggests the reverse. The partitioned alignment on the other hand yields a much lower Gromov-Wasserstein loss, while achieving a COOT loss only marginally worse than that found by multiscale COOT. These results demonstrate that the partitioned multiscale alignment is able to incorporate both pairwise and multiscale information effectively to simultaneously align networks at multiple scales.

5.3.2 Multi-scale biological network matching

Our multi-scale network matching approach is not limited to geometric graphs, e.g., those constructed from a point cloud sampled from 3D objects. We now consider a dataset of protein-protein interaction (PPI) networks [72], in which nodes and edges correspond to protein species and biochemical interactions respectively. We take G_0 to be the PPI network of high-confidence interactions among $1,004$ proteins, and G'_0 to be the PPI network with 20% more low-confidence interactions. For each of G_0, G'_0 we construct a progressive topological simplification using the heat kernel reduction described in [22], yielding multi-scale reductions $\{G_i\}_{i=0}^2$ and $\{G'_i\}_{i=0}^2$. We visualize in Figure 6(a) each multiscale reduction, in which nodes are coloured by the leading non-trivial eigenvector of L_{G_0}, the graph Laplacian of G_0. Nodes in the low-confidence networks G'_0 are also coloured using the ground truth node correspondence. We then calculate matchings at each scale by employing partitioned measure network alignment, multiscale co-optimal transport, as well as independent

	COOT_loss	**GW_loss**
Gromov-Wasserstein	0.036033	0.010600
Multiscale COOT	0.018261	0.019294
Partitioned alignment (projected gradient)	0.019534	0.012787

(a) TOSCA wolf and centaur: COOT and GW distortion losses.

Method	**COOT_loss**	**GW_loss**	**Node correctness**	**Edge correctness**
Gromov-Wasserstein	0.058684	0.016181	0.613546	0.967680
Multiscale COOT	0.028927	0.021701	0.036853	0.565782
Partitioned (proximal)	0.030732	0.016125	0.585657	0.964916
Partitioned (block)	0.032357	0.014105	0.597610	0.951580

(b) Protein-protein interaction network: COOT and GW distortion losses, as well as node and edge correctness.

Table 1: Gromov-Wasserstein (GW) and co-optimal transport (COOT) loss for multiscale network matching. (a) TOSCA object matching. (b) Protein-protein interaction network matching.

Gromov-Wasserstein matchings at each simplification level.

In this example, we care about the *exact* node matchings and so we opt to solve the exact, unregularized network matching problem: for the partitioned measure network alignment as well as Gromov-Wasserstein measure network alignment, we use a proximal gradient algorithm (Algorithm 3). In Figure 6(b), the matchings from each of these algorithms are shown across all three simplification levels. We also show a set of "ground truth" couplings: between G_0 and G'_0, this is the identity coupling, whereas between G_i and G'_i (for $i = 1, 2$) an approximate ground truth coupling is found by calculating a matrix of pairwise correlations between coarse-grained nodes and then solving a linear assignment problem. From a visual assessment of each of the matchings, we observe that both the partitioned and Gromov-Wasserstein alignment find matchings between G_0 and G'_0 that resemble the ground truth, while multiscale COOT performs quite poorly: this reflects the fact that multiscale COOT is unaware of the pairwise adjacency information at each scale. Between G_1 and G'_1, we find that the partitioned alignment continues to resemble the ground truth, but both Gromov-Wasserstein and multiscale COOT matchings begin to deviate significantly. Finally, between G_2 and G'_2, we find that Gromov-Wasserstein continues to appear differently from the ground truth.

Due to the non-geometric nature of the input graphs, an effective direct visualization of the matchings is very difficult. To demonstrate the difference in matching results in a clearer way, we employ the Louvain community detection algorithm [8] which finds a partitioning of G_0 (and hence G'_0) into $m = 12$ communities. Together with the hypergraph coupling between simplification levels i and $i + 1$, the matching π_i at simplification level i induces a matching of communities. In Figure 6(c), we show the induced community matchings for each alignment method, as well as for the ground truth. At level 0, the partitioned alignment and Gromov-Wasserstein both produce nearly perfect alignments, while multiscale COOT performs poorly. At level 1, however, both Gromov-Wasserstein and multiscale COOT perform poorly, while partitioned alignment continues to perform well. Finally, at level 2 from the ground truth matching, it is apparent that too much information is lost by applying coarsening to the graphs to correctly identify communities. However, the matching induced by the partitioned alignment is still closer to the ground truth, for instance in L_1 norm ($L_1 = 0.265$) compared to Gromov-Wasserstein ($L_1 = 0.319$) and multiscale COOT ($L_1 = 0.508$).

In Table 1(b), we show the loss terms similar to the previous TOSCA example: we observe that Gromov-Wasserstein fails to find consistent matchings across scales as evidenced by a high COOT loss. On the other hand, multiscale COOT leads to poor preservation of pairwise relations within each simplification level, indicated by a high Gromov-Wasserstein loss. In contrast, partitioned alignments are able to find multiscale matchings that are consistent within each scale as well as across scales. Furthermore, partitioned alignment methods yield node and edge correctness scores for level 0 that are comparable to Gromov-Wasserstein, which was found to outperform most other competing alignment methods in [78].

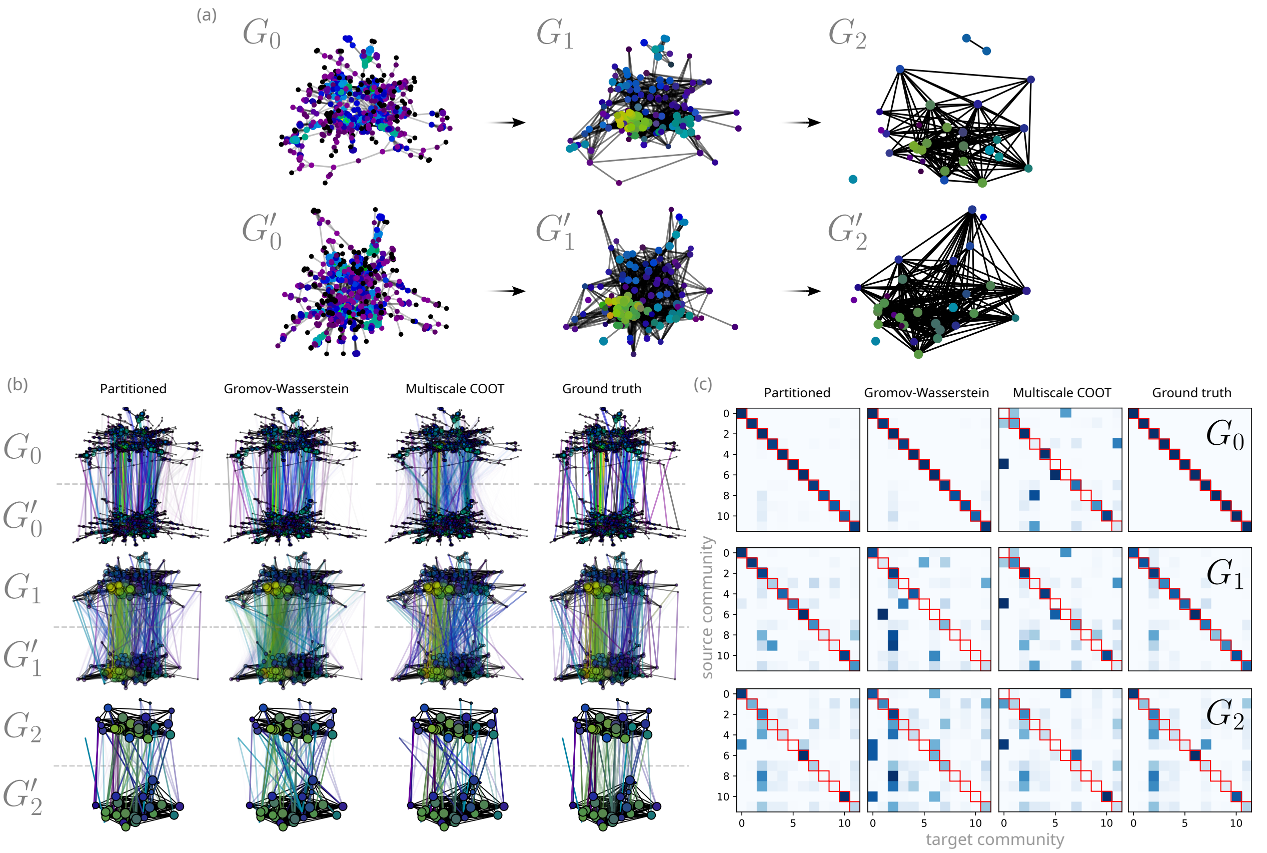

Figure 6: Multiscale matchings for protein-protein interaction networks: (a) Network layouts of successive simplifications $\{G_i\}$ and $\{G'_i\}$ for $i = 0, 1, 2$ with nodes coloured by the leading nontrivial eigenvector of L_G. (b) Matchings found at each scale using (from left to right): partitioned measure network matching; multiscale COOT matching; Gromov-Wasserstein measure network matching at each level; and the ground truth matching. (c) Matchings of Leiden communities induced by the node-level matching between G_i and G'_i at each simplification level i. "Ground truth" shows the best possible matching of Leiden communities using couplings at this level of granularity.

5.4 Geodesics and Fréchet means

In Section 4.3, we introduce the Fréchet functional (26) on the space of partitioned measure networks and calculate its gradient. Recall from Theorem 1 that the spaces of measure networks and measure hypernetworks isometrically embed into the space of partitioned measure networks, we recover from (26) the Fréchet functional on measure networks [20, 52] and measure hypernetworks [22] as special cases. For simplicity we consider the unlabelled case here, although in general our results can be straightforwardly extended to measure networks with labels valued in an inner product space (see e.g., [74]).

In practice, a stationary point of the functional (26) can be found via gradient descent on the space of partitioned measure networks using the "blow-up" scheme of [20] which progressively carries out alignment of network representatives as per Proposition 4.11. This approach exploits the empirical observation that optimal couplings of measure networks tend to be sparse (see e.g., [20, Appendix C] and [21, Theorem 2]), allowing the geodesics of Proposition 3.9 to be explicitly constructed. While the approach of [20] handles only measure networks, we implement an extension of [20, Algorithms 1-3] in order to handle measure hypernetworks. We remark here that, although the question of sparsity of optimal couplings is open in the general quadratic case, in the setting of co-optimal transport between measure hypernetworks, the alternating scheme of Algorithm 1 is guaranteed to yield a sparse coupling. This is because each iterate is the solution of a *linear* program and is therefore sparse. In what follows, we refer to this approach as the *free support* method, since the size of the networks is determined as part of the optimization procedure.

As an alternative to the more involved free support method, we also consider fixing the networks to those described by matrices of a fixed size, as done in [52]. This makes the optimization much easier: as detailed in Section A.7, this problem can be solved by alternating between solving independent alignment problems and a closed form update for the barycenter. We refer to this approach as the *fixed support* method.

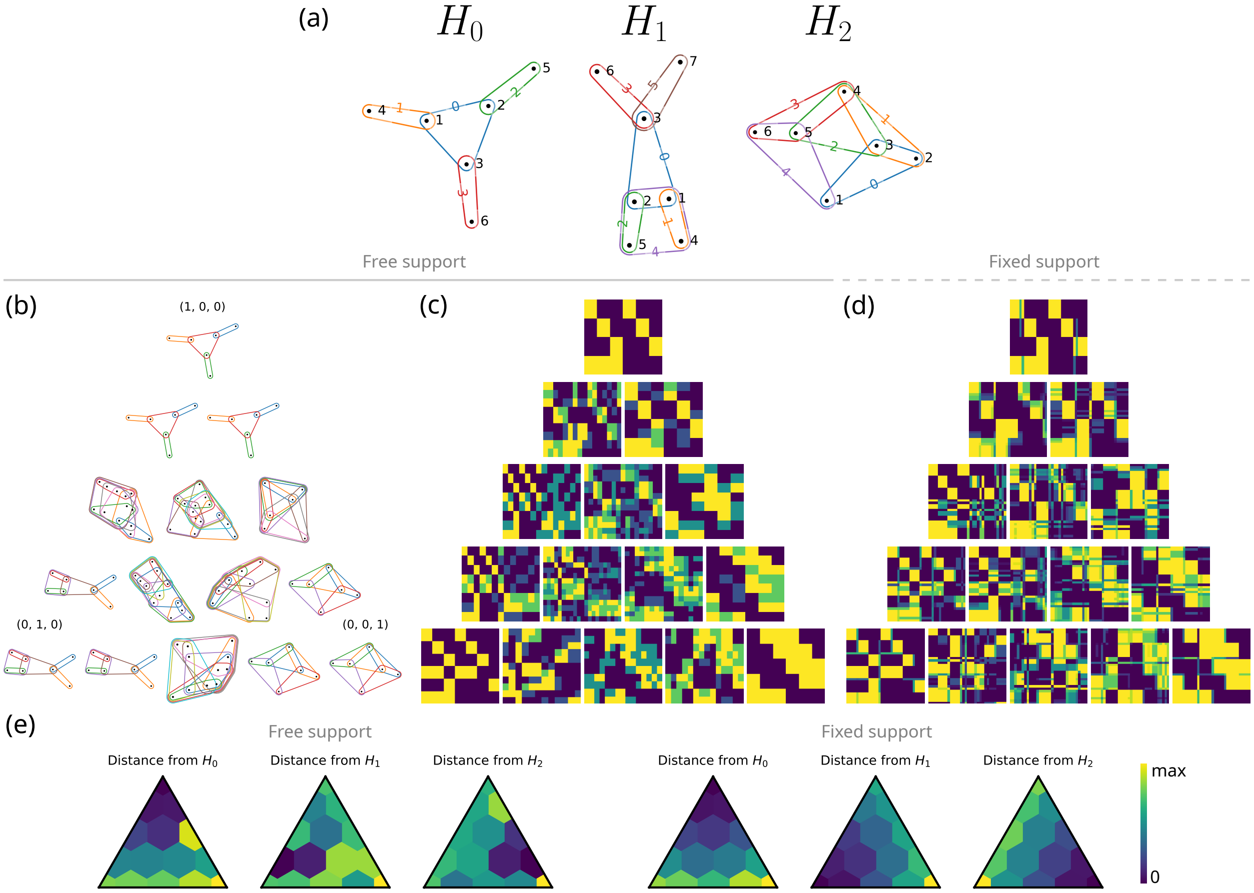

Figure 7: Geodesic interpolations between three hypergraphs H_0, H_1, H_2, shown in (a) as “rubber-band” diagrams. Interpolations found using the free-support method are visualized in (b) as rubber band diagrams and also directly as their membership functions ω in (c). (d) Fixed support (using 32 × 32 matrices) interpolations visualized as membership functions. (e) Hypergraph distance $d_{\mathcal{H},2}(H_i, H_{\mathrm{interp}})$ between each of the three input hypergraphs and geodesic interpolations, for the free support and fixed support barycenters, respectively.

In Figure 7(a), we take three hypergraphs H_0, H_1, and H_2 shown as “rubber-band” diagrams in which nodes are represented as points, and hyperedges are shown as “rubber bands” around the convex hull of contained nodes. We interpolate between these hypergraphs by computing the Fréchet means of $\{H_0, H_1, H_2\}$ with various weights (w_0, w_1, w_2), using both the free support and fixed support methods. Since the rubber band visualization is unable to represent weighted node-hyperedge relationships, a threshold (in this case we chose $\omega(x, y) > 1/4$) is applied to the membership function $\omega(x, y)$ before visualization. As a result, Figure 7(b) cannot reflect the true nature of the interpolation in the measure hypernetwork space. In Figure 7(c), we show the interpolated function ω, making the interpolation readily apparent. We remark that the size (number of nodes and hyperedges) of each interpolating measure hypernetwork is determined by the blowup scheme (in general, larger than each of the inputs), and that the ordering of rows and columns in the visualization is arbitrary. In Figure 7(d), we show the membership functions of barycenters computed using the fixed support method. We note a close resemblance between the results of the free and fixed support approaches.

Finally, in Figure 7(e), we compute the measure hypernetwork distance (per Definition 2.6) from each interpolated hypernetwork to each input hypernetwork. For true interpolations, we would expect the distance to vary affinely across the simplex. However, in the case of free support barycenters, we observe that is not always the case. This suggests that local minima may have played a role in the calculation of the weighted Fréchet means and calculating the distances to each endpoint, reflecting the non-convex nature of the distance and alignment computation. On the other hand, for fixed support barycenters, we recover the expected trend. This reflects the computationally simpler nature of the fixed support method, and may indicate that the fixed support barycenters are more accurate representations of the true hypergraph barycenter.

5.5 Linear and non-linear dictionary learning

We may extend the barycenter question and ask for *several* characteristic partitioned measure networks that best describe an ensemble of partitioned measure networks, rather than a single barycenter or Fréchet mean. One common method is to learn a small basis of representatives (or archetypes), such that each ensemble member can be approximated by a convex combination of these basis elements. Also known as dictionary learning, this has become a classical analysis approach for vector-valued data [39] and has recently been extended to graphs [74, 77] and topological descriptors such as merge trees [40] using the Gromov-Wasserstein framework. We extend dictionary learning to the setting of partitioned measure networks, which also covers the settings of measure networks and measure hypernetworks.

5.5.1 Nonlinear (geodesic) dictionary learning

We first recall the geodesic dictionary learning problem first stated in Section 4.3.2. Given an ensemble of N k-partitioned measure networks $\{P_1, \dots, P_N\} = \{P_i \in \mathcal{P}_k\}_{i=1}^N$, we aim to find a dictionary $\mathcal{D} = \{D_1, \dots, D_m\} = \{D_j \in \mathcal{P}_k\}_{j=1}^m$ (where $m \ll N$) such that each P_i could be described by elements in $\mathcal{D}$. Formally, we denote each $P_i = \big(X_i, \mu_i, \omega_i\big)$, and the geodesic dictionary learning problem is

$$\min_{\{D_j\}_{j=1}^m \in \mathcal{P}_k, \{\alpha_i\}_{i=1}^N \in \Delta^m} \frac{1}{N} \sum_{i=1}^N d_{\mathcal{P}_k} \left(B(\mathcal{D}, \alpha_i), P_i\right)^2 , \tag{43}$$

where $B(\mathcal{D}, \alpha)$ is the barycenter operator (28) for $\mathcal{D}$ and each $\alpha \in \Delta^m$ encodes the corresponding coefficients. We derive formal expressions for the gradients of this function in Section 4.3.2. Like related methods [58, 80], solving problem (43) is a non-convex, bi-level minimization problem which is not straightforward even to find a local optimum. While problems of this nature can be solved using more involved schemes such as [77], we propose to simplify the problem by taking $B(\mathcal{D}, \alpha)$ to be the *fixed support* barycenter operator, where we fix the size (i.e. number of nodes) of the barycenter *a priori* and approximate it iteratively (see Section A.7, and [52]). In practice, we also fix the sizes of the dictionary atoms $\{D_j\}_{j=1}^m$ *a priori* and seek a local minimum solution for dictionary networks of fixed size by a simple gradient descent.

5.5.2 Linear dictionary learning

Even after fixing the support size of barycenters and dictionary atoms, solution of the bi-level problem (43) is computationally demanding due to the need for inner-loop computations of the barycenter operator. As an alternative, *linear* dictionary learning approaches have been proposed [74, 56], in which the (Fréchet) barycenter operator $B(\mathcal{D}, \alpha)$ is replaced with its Euclidean equivalent, a weighted superposition of the atoms $\mathcal{D}[\alpha] = (X, \mu, \sum_{j=1}^m \alpha_j \omega_{\mathcal{D}_j})$. Since reconstructions from the dictionary are carried out in the Euclidean space, this eliminates the nested optimization arising from the barycenter computation, at the cost of departing from the natural geodesic structure of the space of (partitioned) measure networks. We remark that in the setting where the data $P_i \in \mathcal{P}_k, 1 \leqslant i \leqslant N$ are sufficiently close together (in the sense of the injectivity radius of the exponential map), that linear dictionary learning is equivalent to the nonlinear case.

The linearized equivalent of (43) is

$$\min_{\{\omega_{D_j}\}_{j=1}^m \in L^2(X^2, \mu^{\otimes 2}),\ \{\alpha_i\}_{i=1}^N \in \Delta^m} \frac{1}{N} \sum_{i=1}^N d_{\mathcal{P}_k} \left(\mathcal{D}[\alpha_i], P_i\right)^2 , \tag{44}$$

where, for brevity, we denote the linear combination of atoms in $\mathcal{D}$ with coefficients in α to be

$$\mathcal{D}[\alpha] = \left(X, \mu, \sum_{j=1}^{m} \alpha_j \omega_{D_j} \right) \in \mathcal{P}_k.$$

In this formulation, we ask for $m \ll N$ atoms, set $\{\omega_{D_j}\}_{j=1}^m \in L^2(X^2, \mu^{\otimes 2})$, and for each input network P_i, we work with a corresponding set of coefficients $\{\alpha_{ij}\}_{j=1}^m$ such that the reconstructed network described by incidence matrix $\mathcal{D}[\alpha_i]$ is close to P_i in the sense of the optimal transport metric. In the above, we fix the partitioned measure space to (X, μ) respectively. We note that the computation of $d_{\mathcal{P}_k}$ involves solution of a nonlinear program for the coupling π. We can expand the $d_{\mathcal{P}_k}$ terms within the objective (44):

$$\min_{\{\omega_{D_j}\}_{j=1}^m \in L^2(X^2,\mu^{\otimes 2}),\, \{\alpha_i\}_{i=1}^N \in \Delta^m, \{\pi_i \in \Pi_k(\mu,\mu_i)\}_{i=1}^N} \frac{1}{N} \sum_{i=1}^{N} \left\langle L\left(\mathcal{D}[\alpha_i], \omega_{P_i}\right), \pi_i \otimes \pi_i \right\rangle. \tag{45}$$

Minimizing in π_i can be done independently and in parallel for each of the $1 \leqslant i \leqslant N$ inputs. Fixing π, we have a non-convex quadratic program in ω_{D_j} and the coefficients α_i. We solve the problem (44) using a stochastic projected gradient descent.

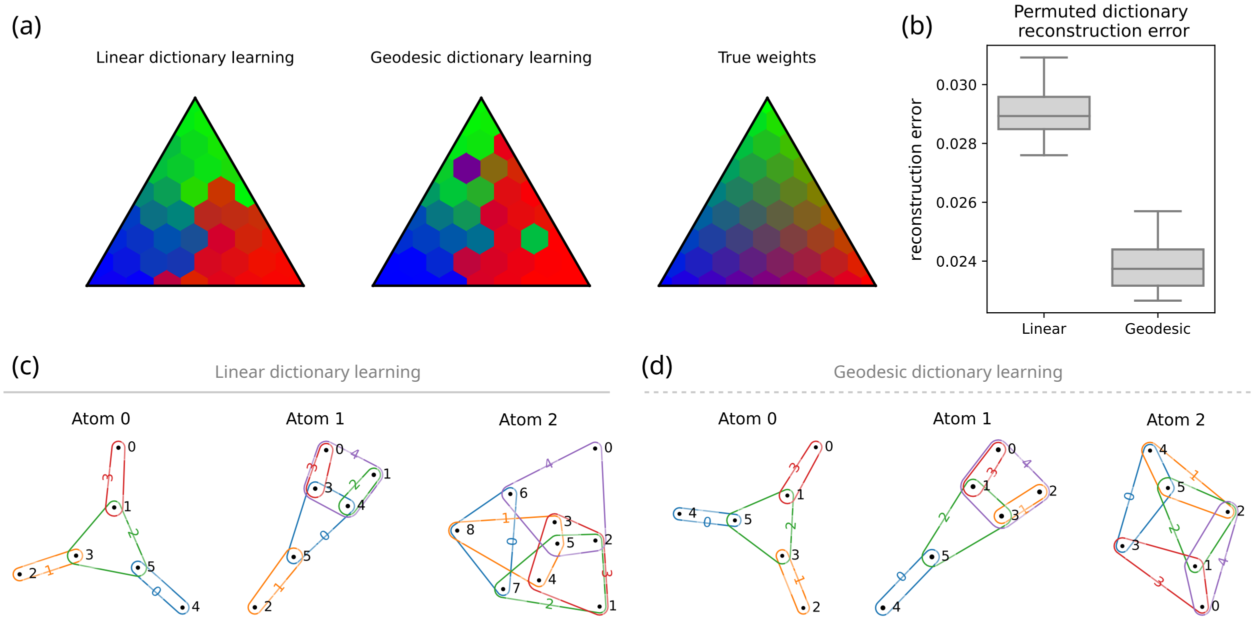

Figure 8: (a) Learned dictionary weights for each interpolation instance using linear dictionary learning (correlation = 0.894) and geodesic dictionary learning (correlation = 0.809), compared to the true weights. (b) Measure hypernetwork distances between true networks and dictionary learning reconstructions, under random permutation of dictionary elements. (c) Atoms learned by linear dictionary learning. (d) Atoms learned by geodesic dictionary learning.

For an experiment, we consider again the example from Section 5.4 involving three hypergraphs. We generate barycenters of these three hypergraphs with mixture weights uniformly spaced across a barycentric grid inside a simplex (i.e., a triangle in this example) using the "blow-up" algorithm. This procedure produces an ensemble of 45 hypergraphs across the grid, in which the number of nodes ranged from 6 to 23, and the number of hyperedges between 4 and 17. By construction, the input hypergraphs H_0, H_1, H_2 are the ground truth atoms, and weights $\{w_{ij}\}_{1 \leqslant i \leqslant N, 0 \leqslant j \leqslant 2}$ serve as the ground truth mixture coefficients. In Figure 8(a) we show the dictionary weights learned by the linear and geodesic dictionary learning respectively, compared to the true weights. Although geodesic dictionary learning (at least in theory) accounts correctly for the underlying geometry of hypergraphs, we find that linearized dictionary learning yields better results in practice. A potential explanation for this result is that linearization avoids the need to solve a non-convex inner loop problem, replacing this with Euclidean averaging.

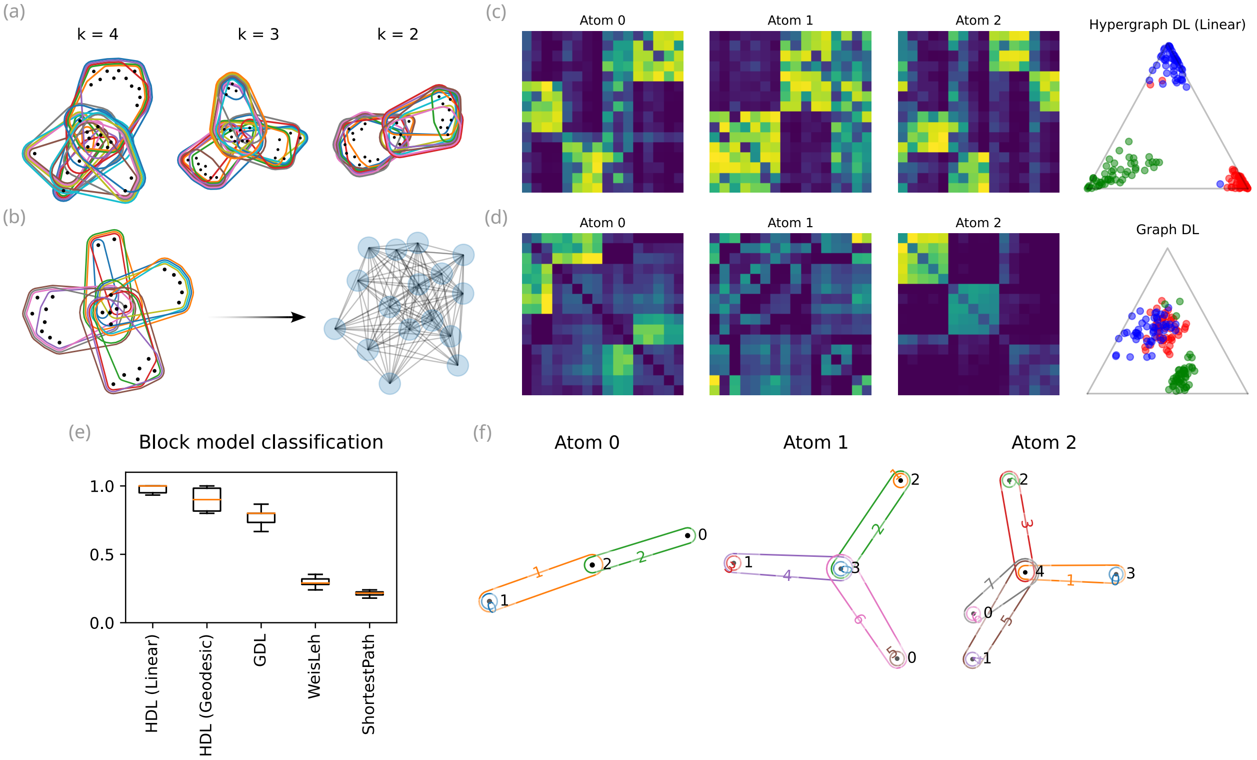

Figure 9: (a) Sample instances from the hypergraph stochastic block model for $k = 4, 3, 2$ blocks respectively. (b) Illustration of hypergraph flattening. (c) left: learned hypergraph atoms capture the three underlying block models; note the presence of hyperedges that span the node set between blocks; right: Barycentric projection of learned weights, coloured by the ground truth label. (d) Learned graph atoms (left) and weights (right) from Gromov-Wasserstein dictionary learning with flattened hypergraphs as input. (e) 10-fold cross-validation results for SVM classification of hypergraph samples. (f) Coarse-grained hypergraph representations of dictionary atoms learned by linear HDL.

To illustrate what we gain by using geodesic over linearized dictionary learning, for each method we randomly permute the learned dictionary atoms and re-calculate the reconstruction error. We show in Figure 8(b) the reconstruction error across 10 independent permutations. In the space of measure hypernetworks, this amounts to changing the choice of representative for the dictionary atoms. This leads to an increased reconstruction error for linearized dictionary learning, relative to geodesic dictionary learning: this tallies with geodesic dictionary learning objective being defined independent of choices of representative. Overall, we find that linearized dictionary learning has an advantage over its geodesic variant due to computational simplicity, and this is the algorithm we use in the following sections.

Finally, we show in Figure 8(c) and (d) the atoms learned by linear and geodesic dictionary learning respectively. Both methods could reliably reconstruct the atoms shown in Figure 7(a), which are used to generate the input ensemble.

5.5.3 Example: stochastic block model for hypergraphs

To further validate our numerical algorithm using synthetic data, we devise a simple generative model for random non-uniform hypergraphs that is analogous to the stochastic block model in the case of simple graphs, similar to the model introduced in [27]. For a (possibly random) number of nodes $|V|$ and a fixed number of blocks k, we partition the node set V into k blocks of size $\lfloor |V|/k \rceil$, with one block containing any remainders. A random number, $|E|$, of hyperedges are then sampled. With probability p, a sampled hyperedge will span $m < \lfloor N/k \rceil$ nodes, all residing in one of the k blocks chosen uniformly at random. Otherwise the hyperedge will consist of m nodes sampled uniformly from all nodes in V. The key parameters at play are k and p, controlling respectively the number of blocks and the level of noise.

We set up three block models with $k = 2, 3, 4$ blocks and $p = 2/3$; 50 hypergraphs are sampled from each model. For each hypergraph, the number of nodes and hyperedges are chosen uniformly from $[16, 32]$ and $[24, 32]$ respectively. We show examples of sampled hypergraphs for each k in Figure 9(a). In total, we generate a dataset of 150 hypergraphs. We apply both linearized and geodesic hypergraph dictionary learning (HDL) algorithms to this dataset to learn three atoms.

To compare against graph-based methods, we apply the graph dictionary learning (GDL) method of [74] using flattened hypergraphs. These flattened hypergraphs are obtained by taking the sum of hyperedges, i.e., $\frac{1}{|E|}\sum_{i=1}^{|E|} \mathbf{1}_{e_i}\mathbf{1}_{e_i}^\top$, that is, putting a weighted connected component in place of each hyperedge; see Figure 9(b). As can be seen from the learned atoms and coefficients in Figure 9(c) and (d), HDL is able to accurately learn distinct atoms corresponding to different values of k. By comparison, GDL finds atoms that appear to be mixed.

We use the learned coefficients from HDL and GDL as features for support vector machine (SVM) classification. In Figure 9(e) we show the accuracy in 10-fold cross validation for both (linear and geodesic) HDL and GDL. For comparison, we also show accuracy achieved by two commonly used graph kernels, the Weisfeiler-Lehmann isomorphism test and the shortest path kernel [62], applied to the flattened hypergraphs. We find that both variants of HDL achieve high accuracy, outperforming GDL. We also observe that linearized HDL outperforms geodesic HDL. Strikingly, the two graph kernel approaches perform much worse than either HDL or GDL. Finally, in Figure 9(f), we show rubber-band diagrams of the atoms found by linearized HDL.

5.5.4 Example: mutagenicity dataset

A real-world example is the well known MUTAG dataset [25] containing 188 small molecule structures, labelled by their mutagenicity (mutagenic or non-mutagenic). This is a standard benchmarking dataset that has been widely adopted for testing graph learning algorithms [62]. Starting from a molecular graph where nodes are atoms and edges are chemical bonds, we propose to lift these chemical structures to partitioned measure networks as follows. Using a graph heat kernel [21], we construct an atom-atom connectivity network encoding proximity of atoms in the molecule. We also construct a persistent homology (PH) hypergraph following the procedure described in [4] to encode information about geometric cycles: nodes and hyperedges correspond to atoms and PH generators, respectively. We show the ensemble of persistence diagrams from this dataset in Figure 10(a), where points are coloured by the ground truth mutagenicity label; see [48] for an user-friendly introduction to persistent homology and [28] for the seminal work on the topic. A separation between mutagenic and non-mutagenic compounds can be visually discerned from these persistence diagrams, suggesting that persistent homology may be sensitive to molecular features that play a role in mutagenicity. In Figure 10(b) we illustrate the computation of the connectivity kernel and the PH-hypergraph from an example molecular graph.

Taking either the PH-hypergraphs or (heat kernel, PH)-partitioned networks as input data, we run linearized partitioned network dictionary learning for $k = 8$ atoms and extract the learned atoms and coefficients. We reason that the learned dictionaries should capture aspects of the molecular structure that are predictive of molecular properties such as mutagenicity. Using the learned coefficients, we train a RBF-kernel SVM for binary classification of mutagenicity. We also apply the linear Gromov-Wasserstein dictionary learning algorithm of [74] using the clique expansion of the PH-hypergraph as input. For reference, we consider several other popular approaches for encoding geometric and topological information in machine learning tasks. Specifically, we compare to persistence images [1] (PersImg), as well as the Weisfeiler-Lehmann graph kernel [61] using either the molecular connectivity graphs (WL-conn) or clique-expansion of the persistent homology hypergraph (WL-ph).

We find that both hypergraph dictionary learning (HDL) and partitioned network dictionary learning (PDL) achieved comparable performance, suggesting that information on PH generators alone is sufficient to distinguish between mutagenic and non-mutagenic compounds in the majority of cases; see Figure 10(c). On the other hand, Gromov-Wasserstein dictionary learning exhibits worse performance despite being provided with PH information in the form of the clique-expansion of the PH hypergraph. This reflects the loss of higher-order information incurred by "flattening" of a hypergraph to a graph [63]. Similarly, we find that classification using persistence images (encoding information about the persistence diagram, but not node-generator membership information) as well as Weisfeiler-Lehman kernels (encoding only pairwise

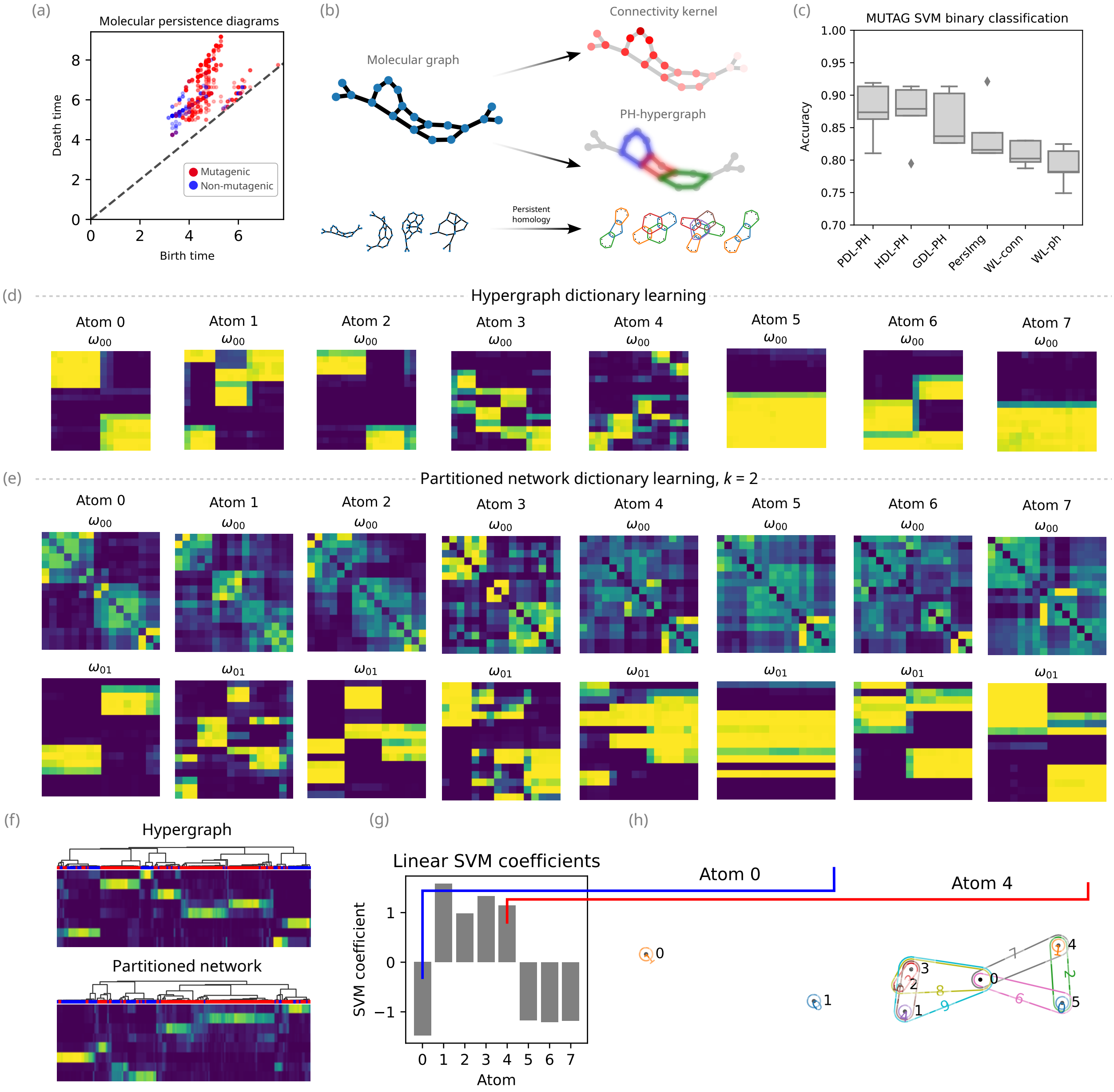

Figure 10: (a) Superimposed persistence diagrams (1-dimensional persistent homology or H_1) for molecular structures in the MUTAG dataset, coloured by mutagenicity. (b) Illustration of connectivity kernel and persistent homology (PH) hypergraph construction from molecular graphs. (c) SVM binary classification accuracy for mutagenicity, using different representations or kernels, shown over 5-fold cross validation. (d) Atoms learned by hypergraph dictionary learning (HDL). (e) Atoms learned by partitioned network dictionary learning (PDL). (f) Dictionary weights learned from HDL and PDL, respectively. (g) Coefficients learned from HDL weights by linear SVM. (h) Atom 0 (predicted to be associated with non-mutagenic compounds) and Atom 4 (predicted to be associated with mutagenic compounds).

relationships) perform relatively poorly. In order to ensure a fair comparison across the different methods we consider, we do not use node labels in this analysis. We remark that our dictionary learning algorithm can be straightforwardly extended to incorporate vector-valued node labels [74] and we expect that including that additional information would further improve classification performance.

Figures 10(d) and (e) illustrate the atoms found by HDL and PDL, respectively. The learned atoms of the partitioned network dictionary consist of pairwise node-node similarities, as well as node-hyperedge memberships. The partitioned network atoms capture both local detail as well as topological information in a coupled fashion. By comparison, with only hypergraph information, nodes with membership in the same homology generators may be indistinguishable. The dictionary coefficients in Figure 10(f) illustrate how each molecule in the dataset is decomposed in terms of soft membership to each of the $k = 8$ archetypes (atoms or representatives). From visual inspection, it is already clear that the learned topics are largely able to disentangle mutagenic compounds from non-mutagenic ones. To quantify this, we train a linear SVM on the hypergraph dictionary coefficients, and extract the contribution of each atom towards the "mutagenic" class; see Figure 10(g). We find that atoms are clearly separated into mutagenicity and non-mutagenicity contributing factors. In Figure 10(h), we show example PH-hypergraph archetypes that are indicative of mutagenicity or non-mutagenicity. For instance, Atom 0, which contributes towards non-mutagenicity, displays a simple PH-hypergraph structure with two generators that do not share nodes. On the other hand, Atom 4 contributes towards mutagenicity and displays a complex PH-hypergraph structure with many interlinked generators. These findings suggest that compounds with more cycles are more likely to be mutagenic, and this is consistent with the chemical literature [25].

6 Discussion

We develop a theoretical footing for the analysis of generalized network objects, modelled by the space of k-partitioned measure networks. We equip this space with a family of metrics $d_{\mathcal{P}_k^p}$ $(p \geqslant 1)$ that extends the well-known p-Gromov-Wasserstein metric originally developed for measure metric spaces [65, 44] and recently applied to measure networks [19] and measure hypernetworks [22]. When $p = 2$, we further provide a geometric characterization of the space of partitioned measure networks in terms of geodesics, curvature bounds, as well as its tangent space. We additionally consider the case where additional *labels* (valued in a metric space) are available. We prove metric properties in the labelled setting, and for $p = 2$, we show that our geometric characterizations also apply (when labels are valued in an inner product space). Based on these ideas, we provide a range of numerical examples illustrating the applicability of our framework across multiple domains in network analysis and data science. We believe our work will be of broad interest to the network science, geometry, and statistics communities.

Our work leaves several open avenues for future research. Among these, providing a geometric characterization of labelled networks when labels are valued in more general spaces, such as Riemannian manifolds (for instance, conditions for uniqueness of geodesics and curvature). Additionally, a rigorous analysis of functionals on the space of partitioned measure networks and their gradient flows remains to be constructed. Whereas we have introduced notions of tangent vectors and gradients and have formally shown their calculation and application, we have not addressed issues such as the existence and uniqueness of minimizer in the general case.

On the application front, some of the example applications we have presented could be fruitful problems to study separately. Among these, we think that multiscale network alignment and partitioned network dictionary learning are particularly interesting. As an example of a specific application requiring extensions of our theory, in [82], we develop a *partial transport* variant for applications in topological data analysis. Finally, we remark that there are some alternative formulations of optimal transport applied to tensors [35] that our framework does not cover. It remains an open question whether similar theoretical results to ours can be established for the tensors.

References

[1] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8):1–35, 2017.

[2] Jason M Altschuler and Enric Boix-Adsera. Wasserstein barycenters are NP-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022.

[3] Amos Bairoch. The enzyme database in 2000. *Nucleic acids research*, 28(1):304–305, 2000.

[4] Agnese Barbensi, Iris HR Yoon, Christian Degnbol Madsen, Deborah O Ajayi, Michael PH Stumpf, and Heather A Harrington. Hypergraphs for multiscale cycles in structured data. *arXiv preprint arXiv:2210.07545*, 2022.

[5] Martin Bauer, Facundo Mémoli, Tom Needham, and Mao Nishino. The Z-Gromov-Wasserstein distance. *arXiv preprint arXiv:2408.08233*, 2024.

[6] Christian Bick, Elizabeth Gross, Heather A Harrington, and Michael T Schaub. What are higher-order networks? *SIAM Review*, 65(3):686–731, 2023.

[7] Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *International conference on artificial intelligence and statistics*, pages 880–889. PMLR, 2018.

[8] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[9] Karsten M Borgwardt and Hans-Peter Kriegel. Shortest-path kernels on graphs. In *Fifth IEEE international conference on data mining (ICDM'05)*, pages 8–pp. IEEE, 2005.

[10] Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

[11] 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.

[12] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008.

[13] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Society, 2022.

[14] Yiqun Cao, Anna Charisi, Li-Chang Cheng, Tao Jiang, and Thomas Girke. ChemmineR: a compound mining framework for R. *Bioinformatics*, 24(15):1733–1734, 2008.

[15] Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913, 2020.

[16] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

[17] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.

[18] Samir Chowdhury and Facundo Mémoli. Explicit geodesics in Gromov-Hausdorff space. *Electronic Research Announcements*, 25(0):48–59, 2018.

[19] Samir Chowdhury and Facundo Mémoli. The Gromov–Wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019.

[20] Samir Chowdhury and Tom Needham. Gromov-Wasserstein averaging in a Riemannian framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 842–843, 2020.

[21] Samir Chowdhury and Tom Needham. Generalized spectral clustering via Gromov-Wasserstein learning. In *International Conference on Artificial Intelligence and Statistics*, pages 712–720. PMLR, 2021.

[22] Samir Chowdhury, Tom Needham, Ethan Semrad, Bei Wang, and Youjia Zhou. Hypergraph co-optimal transport: Metric and categorical properties. *Journal of Applied and Computational Topology*, pages 1–60, 2023.

[23] Corinna Coupette, Jilles Vreeken, and Bastian Rieck. All the world'sa (hyper) graph: A data drama. *Digital Scholarship in the Humanities*, 39(1):74–96, 2024.

[24] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[25] Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

[26] Pinar Demetci, Quang Huy Tran, Ievgen Redko, and Ritambhara Singh. Breaking isometric ties and introducing priors in gromov-wasserstein distances. In *International Conference on Artificial Intelligence and Statistics*, pages 298–306. PMLR, 2024.

[27] Ioana Dumitriu, Hai-Xiao Wang, and Yizhe Zhu. Partial recovery and weak consistency in the non-uniform hypergraph stochastic block model. *Combinatorics, Probability and Computing*, pages 1–51, 2021.

[28] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28:511–533, 2002.

[29] Soheil Feizi, Gerald Quon, Mariana Recamonde-Mendoza, Muriel Médard, Manolis Kellis, and Ali Jadbabaie. Spectral alignment of networks. *arXiv preprint arXiv:1602.04181*, 2016.

[30] Soheil Feizi, Gerald Quon, Mariana Recamonde-Mendoza, Muriel Medard, Manolis Kellis, and Ali Jadbabaie. Spectral alignment of graphs. *IEEE Transactions on Network Science and Engineering*, 7(3):1182–1197, 2019.

[31] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[32] Yanjun Han, Philippe Rigollet, and George Stepaniants. Covariance alignment: from maximum likelihood estimation to Gromov-Wasserstein. *arXiv preprint arXiv:2311.13595*, 2023.

[33] Jürgen Jost and Raffaella Mulas. Hypergraph Laplace operators for chemical reaction networks. *Advances in mathematics*, 351:870–896, 2019.

[34] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[35] Tanguy Kerdoncuff, Rémi Emonet, Michaël Perrot, and Marc Sebban. Optimal tensor transport. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7124–7132, 2022.

[36] Tamara G Kolda and Jackson R Mayo. Shifted power method for computing tensor eigenpairs. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1095–1124, 2011.

[37] Nicholas J Korevaar and Richard M Schoen. Sobolev spaces and harmonic maps for metric space targets. *Communications in Analysis and Geometry*, 1(4):561–659, 1993.

[38] Andrew Lauziere, Ryan Christensen, Hari Shroff, and Radu Balan. An exact hypergraph matching algorithm for posture identification in embryonic C. elegans. *PLOS One*, 17(11):e0277343, 2022.

[39] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[40] Mingzhe Li, Sourabh Palande, Lin Yan, and Bei Wang. Sketching merge trees for scientific visualization. In *IEEE Workshop on Topological Data Analysis and Visualization (TopoInVis) at IEEE VIS*, pages 61–71, 2023.

[41] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

[42] Dirk A Lorenz, Paul Manns, and Christian Meyer. Quadratically regularized optimal transport. *Applied Mathematics & Optimization*, 83(3):1919–1949, 2021.

[43] Facundo Mémoli. On the use of Gromov-Hausdorff Distances for Shape Comparison. In M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, editors, *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association, 2007.

[44] Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

[45] Tom Michoel and Bruno Nachtergaele. Alignment and integration of complex networks by hypergraph-based spectral clustering. *Physical Review E*, 86(5):056111, 2012.

[46] Huda Nassar, Nate Veldt, Shahin Mohammadi, Ananth Grama, and David F Gleich. Low rank spectral network alignment. In *Proceedings of the 2018 World Wide Web Conference*, pages 619–628, 2018.

[47] Efe Onaran and Soledad Villar. Projected power iteration for network alignment. In *Wavelets and Sparsity XVII*, volume 10394, pages 298–305. SPIE, 2017.

[48] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1):17, 2017.

[49] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

[50] Xavier Pennec. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25:127–154, 2006.

[51] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[52] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pages 2664–2672. PMLR, 2016.

[53] Conrad Plaut. Metric spaces of curvature $\geqslant k$. In *Handbook of geometric topology*, pages 819–898. Elsevier, 2001.

[54] Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

[55] Ievgen Redko, Titouan Vayer, Rémi Flamary, and Nicolas Courty. Co-optimal transport. *Advances in neural information processing systems*, 33:17559–17570, 2020.

[56] Antoine Rolet, Marco Cuturi, and Gabriel Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In *Artificial Intelligence and Statistics*, pages 630–638. PMLR, 2016.

[57] Jayoung Ryu, Romain Lopez, Charlotte Bunne, and Aviv Regev. Cross-modality matching and prediction of perturbation responses with labeled Gromov-Wasserstein optimal transport. *arXiv preprint arXiv:2405.00838*, 2024.

[58] Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

[59] Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced Gromov Wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021.

[60] Tie Shen, Zhengdong Zhang, Zhen Chen, Dagang Gu, Shen Liang, Yang Xu, Ruiyuan Li, Yimin Wei, Zhijie Liu, Yin Yi, et al. A genome-scale metabolic network alignment method within a hypergraph-based framework using a rotational tensor-vector product. *Scientific reports*, 8(1):16376, 2018.

[61] Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.

[62] Giannis Siglidis, Giannis Nikolentzos, Stratis Limnios, Christos Giatsidis, Konstantinos Skianis, and Michalis Vazirgiannis. Grakel: A graph kernel library in Python. *Journal of Machine Learning Research*, 21(54):1–5, 2020.

[63] Balasubramaniam Srinivasan, Da Zheng, and George Karypis. Learning over families of sets-hypergraph representation learning for higher order tasks. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 756–764. SIAM, 2021.

[64] Karl-Theodor Sturm. On the geometry of metric measure spaces. *Acta Mathematica*, 196(1):65–131, 2006.

[65] Karl-Theodor Sturm. *The Space of Spaces: Curvature Bounds and Gradient Flows on the Space of Metric Measure Spaces*. Memoirs of the American Mathematical Society. American Mathematical Society, 2023.

[66] Thomas W Thorne and Michael P H Stumpf. Graph spectral analysis of protein interaction network evolution. *Journal of the Royal Society, Interface / the Royal Society*, 9(75):2653–2666, 2012.

[67] Alexis Thual, Quang Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced Gromov Wasserstein. *Advances in Neural Information Processing Systems*, 35:21792–21804, 2022.

[68] Vayer Titouan, 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.

[69] Quang Huy Tran, Hicham Janati, Nicolas Courty, Rémi Flamary, Ievgen Redko, Pinar Demetci, and Ritambhara Singh. Unbalanced co-optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10006–10016, 2023.

[70] Tomas Vaskevicius and Lénaïc Chizat. Computational guarantees for doubly entropic wasserstein barycenters. *Advances in Neural Information Processing Systems*, 36, 2024.

[71] Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused Gromov-Wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

[72] Vipin Vijayan, Vikram Saraph, and Tijana Milenković. Magna++: maximizing accuracy in global network alignment via both node and edge conservation. *Bioinformatics*, 31(14):2409–2411, 2015.

[73] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[74] Cédric Vincent-Cuaz, Titouan Vayer, Rémi Flamary, Marco Corneli, and Nicolas Courty. Online graph dictionary learning. In *International Conference on Machine Learning*, pages 10564–10574. PMLR, 2021.

[75] Peter Wills and François G Meyer. Metrics for graph comparison: a practitioner's guide. *PLOS One*, 15(2):e0228728, 2020.

[76] Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact Wasserstein distance. In *Uncertainty in artificial intelligence*, pages 433–453. PMLR, 2020.

[77] Hongteng Xu, Jiachang Liu, Dixin Luo, and Lawrence Carin. Representing graphs via Gromov-Wasserstein factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):999–1016, 2022.

[78] Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable Gromov-Wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019.

[79] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-Wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pages 6932–6941. PMLR, 2019.

[80] Hongtengl Xu. Gromov-Wasserstein factorization models for graph clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6478–6485, 2020.

[81] Stephen Zhang, Gilles Mordant, Tetsuya Matsumoto, and Geoffrey Schiebinger. Manifold learning with sparse regularised optimal transport. *arXiv preprint arXiv:2307.09816*, 2023.

[82] Stephen Y Zhang, Michael PH Stumpf, Tom Needham, and Agnese Barbensi. Topological optimal transport for geometric cycle matching. *arXiv preprint arXiv:2403.19097*, 2024.

[83] Youjia Zhou, Archit Rathore, Emilie Purvine, and Bei Wang. Topological simplifications of hypergraphs. *IEEE Transactions on Visualization and Computer Graphics*, 29(7):3209–3225, 2022.

A Details on numerical algorithms

In this section, we follow the notations introduced in Section 5.1.

A.1 Co-optimal transport

By taking $k = 2$, $\omega_{ii} = 0$, $\omega_{ij} = \omega_{ji}$, and L defined in (33), it is easy to verify the following identity:

$$\langle L(\omega_{12}, \omega'_{12}), \pi_1 \otimes \pi_2 \rangle = \langle L(\omega_{12}, \omega'_{12}) \otimes \pi_2, \pi_1 \rangle = \langle L(\omega_{12}^\top, (\omega'_{12})^\top) \otimes \pi_1, \pi_2 \rangle = \langle L(\omega_{12}^\top, (\omega'_{12})^\top), \pi_2 \otimes \pi_1 \rangle.$$

We obtain from (35) the case of matching between labelled measure hypernetworks $H = (X_1, \mu_1, X_2, \mu_2, \omega_{12})$ and $H' = (X'_1, \mu'_1, X'_2, \mu'_2, \omega'_{12})$:

$$\min_{\pi_1 \in \Pi(\mu_1, \mu'_1), \pi_2 \in \Pi(\mu_2, \mu'_2)} \langle L(\omega_{12}, \omega'_{12}), \pi_1 \otimes \pi_2 \rangle + \langle C_1, \pi_1 \rangle + \langle C_2, \pi_2 \rangle + \varepsilon_1 \Omega_1(\pi_1) + \varepsilon_2 \Omega_2(\pi_2), \tag{46}$$

where for full generality we include the possibility of a regularization of each of the π_i when $\varepsilon_i > 0$, as per (36).

The unregularized problem is a bilinear program in (π_1, π_2) since the objective can be rewritten as $\langle L(\omega_{12}, \omega'_{12}) + C_1 \oplus C_2, \pi_1 \otimes \pi_2 \rangle$ where $(A \oplus B)_{ijkl} = A_{ij} + B_{kl}$. In the case of unlabelled measure hypernetworks where $C_1 = C_2 = 0$, this has been studied in detail by [55, 22], among others. The alternating scheme of [55] for finding a stationary point presents itself:

$$\begin{aligned} \pi_1 &\leftarrow \min_{\pi_1 \in \Pi(\mu_1, \mu'_1)} \langle L(\omega_{12}, \omega'_{12}) \otimes \pi_2 + C_1, \pi_1 \rangle + \varepsilon_1 \Omega_1(\pi_1), \\ \pi_2 &\leftarrow \min_{\pi_2 \in \Pi(\mu_2, \mu'_2)} \langle L(\omega_{12}^\top, \omega'^\top_{12}) \otimes \pi_1 + C_2, \pi_2 \rangle + \varepsilon_2 \Omega_2(\pi_2). \end{aligned} \tag{47}$$

Employing the identity [52, Proposition 1] we have for $\pi \in \Pi(\mu, \mu')$ that

$$L(\omega, \omega') \otimes \pi = \sum_{kl} L(\omega, \omega')_{\cdot\cdot kl} \otimes \pi_{kl} = \eta(\omega, \omega') - \omega \pi (\omega')^\top, \tag{48}$$

where

$$\eta(\omega, \omega') = \frac{1}{2} \left(\omega^{\wedge 2}\right) \mu \mathbf{1}^\top + \frac{1}{2} \mathbf{1} \left(\omega'^{\wedge 2} \mu'\right)^\top, \tag{49}$$

and $\omega^{\wedge 2}, \omega'^{\wedge 2}$ are understood entrywise. $\eta(\omega, \omega')$ depends only upon (μ, μ'), the marginals of π.

Algorithm 1 Alternating minimization: labelled hypergraphs (co-optimal transport)

```
Input: Incidence matrices $\omega_{12}, \omega'_{12}$, probability measures $\mu_i, \mu'_i$, $i = 1, 2$, label cost matrices $C_{1,2}$ (optional).
Parameters: entropic regularization levels $\varepsilon_1, \varepsilon_2 \geqslant 0$ (optional)
Initialize couplings: $\pi_i \leftarrow \mu_i \otimes \mu'_i$, $i = 1, 2$.
for $t = 1, 2, \ldots,$ max_iter do
    $\pi_1 \leftarrow \arg\min_{\pi_1 \in \Pi(\mu_1, \mu'_1)} \langle L(\omega_{12}, \omega'_{12}) \otimes \pi_2 + C_1, \pi_1 \rangle + \varepsilon_1 \Omega_1(\pi_1),$
    $\pi_2 \leftarrow \arg\min_{\pi_2 \in \Pi(\mu_2, \mu'_2)} \langle L(\omega_{12}^\top, \omega'^\top_{12}) \otimes \pi_1 + C_2, \pi_2 \rangle + \varepsilon_2 \Omega_2(\pi_2),$
end for
Output: couplings $(\pi_1, \pi_2)$
```

A.2 General matchings of partitioned measure networks

For general k-partitioned measure networks, the problem (35) is quadratic in $(\pi_i)_{i=1}^k$. We propose to obtain an approximate solution to this problem using block coordinate descent separately in each of the π_i, while holding $(\pi_j)_{j\neq i}$ fixed. Each block update amounts to solution of a problem closely resembling Fused Gromov-Wasserstein matching [71]:

$$\min_{\pi_i \in \Pi(\mu_i, \mu_i')} \frac{1}{2}\langle L(\omega_{ii}, \omega_{ii}'), \pi_i \otimes \pi_i \rangle + \langle M[\pi_{-i}] + C_i, \pi_i \rangle + \varepsilon_i \Omega_i(\pi_i), \quad 1 \leqslant i \leqslant k, \tag{50}$$

where $\pi_{-i} = (\pi_j)_{j\neq i}$ and

$$\begin{aligned} M[\pi_{-i}] &= \sum_{j\neq i} \left(\frac{1}{2} L(\omega_{ij}, \omega_{ij}') + \frac{1}{2} L(\omega_{ji}^\top, \omega_{ji}'^\top) \right) \otimes \pi_j \\ &= \frac{1}{2} \sum_{j\neq i} \left[\eta(\omega_{ij}, \omega_{ij}') - \omega_{ij} \pi_j ({\omega'}_{ij})^\top + \eta(\omega_{ji}^\top, {\omega'}_{ji}^\top) - \omega_{ji}^\top \pi_j \omega_{ji}' \right], \end{aligned} \tag{51}$$

Each subproblem (50) in π_i amounts to the minimization of a non-convex quadratic objective on a closed convex domain, and so a stationary point can be found using the conditional gradient algorithm of [71, Algorithm 1]. We remark that a similar algorithm for the case of augmented measure networks was introduced in [26].

Algorithm 2 Alternating minimization: labelled k-partitioned networks

```
Input: Matrices $\{\omega_{ij}\}_{i,j=1}^k, \{\omega_{ij}'\}_{i,j=1}^k$, probability measures $\mu_i, \mu_i'$, $1 \leqslant i \leqslant k$, label cost matrices $(C_i)_{i=1}^k$ (optional)
Parameters: entropic regularization levels $\varepsilon_i \geqslant 0, 1 \leqslant i \leqslant k$ (optional)
Initialize couplings: $\pi_i \leftarrow \mu_i \otimes \mu_i'$, $1 \leqslant i \leqslant k$.
for $t = 1, 2, \dots, \texttt{max\_iter}$ do
    for $1 \leqslant i \leqslant k$ do
        $\pi_i \leftarrow \operatorname*{arg\,min}_{\pi_i \in \Pi(\mu_i, \mu_i')} \frac{1}{2}\langle L(\omega_{ii}, \omega_{ii}'), \pi_i \otimes \pi_i \rangle + \langle M[\pi_{-i}] + C_i, \pi_i \rangle + \varepsilon_i \Omega_i(\pi_i)$,    with $M[\pi_{-i}]$ as per (51)
    end for
end for
Output: couplings $\{\pi_i\}_{i=1}^k$
```

A.3 Proximal gradient methods

As an alternative to relying on exact solvers for the unregularized problem, an entropic proximal gradient algorithm can also be used to solve the problem (35). These algorithms have been shown to perform favorably in terms of computational complexity as well as empirical results [79, 76]. Writing L to be the objective function of (35),

$$\mathsf{L}(\pi_1, \dots, \pi_k) = \frac{1}{2} \sum_{i,j=1}^k \langle L(\omega_{ij}, \omega_{ij}'), \pi_i \otimes \pi_j \rangle + \sum_{i=1}^k \langle C_i, \pi_i \rangle, \tag{52}$$

and choosing a regularization level (inverse step size schedule) $\lambda^t > 0$, a proximal gradient descent on the objective L starting from an initialization $(\pi_i^0)_{i=1}^k$ generates the iterates for $t \geqslant 0$:

$$(\pi_i^{t+1})_{i=1}^k \leftarrow \operatorname*{arg\,min}_{\pi_i \in \Pi(\mu_i, \mu_i'),\, 1 \leqslant i \leqslant k} \mathsf{L}(\pi_1, \dots, \pi_k) + \lambda^t \operatorname{KL}(\otimes_i \pi_i | \otimes_i \pi_i^t), \tag{53}$$

where $\operatorname{KL}(\alpha|\beta)$ denotes the (generalized) Kullback-Leibler divergence between probability distributions (positive measures)

$$\operatorname{KL}(\alpha|\beta) = \langle \alpha, \log(\mathrm{d}\alpha/\mathrm{d}\beta) \rangle - |\alpha| + |\beta|. \tag{54}$$

Replacing L with its linearization about $(\pi_1^t, \ldots, \pi_k^t)$ yields the proximal gradient method [49, 78],

$$(\pi_i^{t+1})_{i=1}^k \leftarrow \operatorname*{arg\,min}_{\pi_i \in \Pi(\mu_i, \mu_i'), 1 \leqslant i \leqslant k} \sum_{i=1}^k \langle \nabla_i \mathsf{L}(\pi_1^t, \ldots, \pi_k^t), \pi_i \rangle + \lambda^t \operatorname{KL}(\otimes_i \pi_i | \otimes_i \pi_i^t), \tag{55}$$

where ∇_i denotes the gradient of L in its ith argument. Since (for all probability measure inputs) $\operatorname{KL}(\otimes_i \pi_i | \otimes_i \pi_i^t) = \sum_i \operatorname{KL}(\pi_i | \pi_i^t)$, the proximal gradient update decouples in each of the π_i:

$$\pi_i^{t+1} \leftarrow \operatorname*{arg\,min}_{\pi_i \in \Pi(\mu_i, \mu_i')} \langle \nabla_i \mathsf{L}(\pi_1^t, \ldots, \pi_k^t), \pi_i \rangle + \lambda^t \operatorname{KL}(\pi_i | \pi_i^t), \quad 1 \leqslant i \leqslant k. \tag{56}$$

Rewriting each problem leads to an entropic optimal transport problem which can be solved via Sinkhorn iterations [24]:

$$\pi_i^{t+1} \leftarrow \operatorname*{arg\,min}_{\pi_i \in \Pi(\mu_i, \mu_i')} \lambda^t \operatorname{KL}(\pi_i | \pi_i^t \odot e^{-M_i/\lambda^t}), \quad M_i = \nabla_i \mathsf{L}(\pi_1^t, \ldots, \pi_i^t), \quad 1 \leqslant i \leqslant k. \tag{57}$$

Noting that

$$\frac{\partial}{\partial \pi} \frac{1}{2} \langle L(\omega, \omega'), \pi \otimes \pi \rangle = \left(\frac{1}{2} L(\omega, \omega') + \frac{1}{2} L(\omega^\top, {\omega'}^\top) \right) \otimes \pi,$$

we have the following formula for $\nabla_i \mathsf{L}$:

$$\nabla_i \mathsf{L}(\pi_1, \ldots, \pi_k) = \frac{1}{2} \left(L(\omega_{ii}, \omega_{ii}') + L(\omega_{ii}^\top, {\omega'}_{ii}^\top) \right) \otimes \pi_i + M[\pi_{-i}] + C_i$$

where $M[\pi_{-i}]$ is defined in (51).

Algorithm 3 Proximal gradient: labelled k-partitioned networks

```
Input: Matrices $\{\omega_{ij}\}_{i,j=1}^k, \{\omega_{ij}'\}_{i,j=1}^k$, probability measures $\mu_i, \mu_i'$, $1 \leqslant i \leqslant k$, label cost matrices $(C_i)_{i=1}^k$ (optional).
Parameters: inverse step size schedule $\lambda^t, t \geqslant 0$.
Initialize couplings: $\pi_i^1 \leftarrow \mu_i \otimes \mu_i'$, $1 \leqslant i \leqslant k$.
for $t = 1, 2, \ldots,$ max_iter do
    for $1 \leqslant i \leqslant k$ do
        $M_i \leftarrow \frac{1}{2}\left(L(\omega_{ii}, \omega_{ii}') + L(\omega_{ii}^\top, {\omega'}_{ii}^\top)\right) \otimes \pi_i^t + M[\pi_{-i}^t] + C_i$     (see (51))
        $\pi_i^{t+1} \leftarrow \arg\min_{\pi_i \in \Pi(\mu_i, \mu_i')} \lambda^t \operatorname{KL}(\pi_i | \pi_i^t \odot e^{-M_i/\lambda^t})$     (solve using Sinkhorn algorithm).
    end for
end for
Output: couplings $\{\pi_i\}_{i=1}^k$
```

We remark that when all the $\omega_{ii} = 0$, then L coincides with its linearization and Algorithm 3 is in fact a *proximal point* method.

Proposition A.1 (Convergence of proximal point method)**.** *The limiting iterates as $t \to \infty$ of the problem* (53) *converge to a stationary point of* (35)*.*

Proof. Writing $\pi = (\pi_1, \ldots, \pi_k)$, (53) can be written as

$$\min_{\pi \in U} u(\pi, \pi^t),$$

where

$$u(\pi, \pi') = \mathsf{L}(\pi_1, \ldots, \pi_k) + \lambda \operatorname{KL}(\otimes_i \pi_i | \otimes_i \pi_i')$$

and the set $U = \times_{i=1}^k \Pi(\mu_i, \mu_i')$ is closed and convex as a Cartesian product of closed, convex sets. Now we note that

- $u(\pi, \pi) = \mathsf{L}(\pi)$ for all $\pi \in U$.
- $u(\pi, \pi') \geqslant \mathsf{L}(\pi)$ by non-negativity of the KL-divergence.
- $u(\pi, \pi')$ is smooth in both of its arguments.

Taken together and applying [54, Proposition 1], we satisfy the conditions for [54, Theorem 1].

We remark that the unbalanced case (where the hard marginal constraints are replaced by soft constraints) can be handled in the same way, if the marginal penalties are smooth. The constraint set is then $U = \times_{i=1}^{k} \mathcal{M}_+$ which is also closed and convex. □

A.4 Projected gradient descent

When solving the regularized problem (36) and setting $\Omega_0 = \Omega_1 = \mathrm{KL}$, a projected gradient descent approach [52] can be used. Then the minimization problem has the form

$$\min_{\pi_i \in \Pi(\mu_i, \mu_i'), 1 \leqslant i \leqslant k} \mathsf{L}(\pi_1, \ldots, \pi_k) + \sum_{i=1}^{k} \varepsilon_i \, \mathrm{KL}(\pi_i | \mu_i \otimes \mu_i'). \tag{58}$$

For a gradient step size $\eta_i > 0$, a projected mirror descent step in each π_i reads

$$\pi_i^{t+1} \leftarrow \mathrm{Proj}^{\mathrm{KL}}_{\Pi(\mu_i, \mu_i')} \left[\pi_i^t \odot \exp\left(-\eta_i (\nabla_i \mathsf{L}(\pi_1^t, \ldots, \pi_k^t) + \varepsilon_i \log(\pi_i^t / \mu_i \otimes \mu_i')) \right) \right], \quad 1 \leqslant i \leqslant k. \tag{59}$$

Choosing $\eta_i = 1/\varepsilon_i$, we get

$$\pi_i^{t+1} \leftarrow \mathrm{Proj}^{\mathrm{KL}}_{\Pi(\mu_i, \mu_i')} \left[e^{-\varepsilon_i^{-1} \nabla_i \mathsf{L}(\pi_1, \ldots, \pi_k)} \mu_i \otimes \mu_i' \right], \quad 1 \leqslant i \leqslant k. \tag{60}$$

Each of these projections can be computed through the Sinkhorn algorithm [52, 24]. Similarly, one may consider $\Omega_i = \frac{1}{2} \| \cdot \|^2_{L^2(\mu_i \otimes \mu_i')}$, in which case the projected L^2-gradient descent steps are

$$\pi_i^{t+1} \leftarrow \mathrm{Proj}^{F}_{\Pi(\mu_i, \mu_i')} \left[-\frac{1}{\varepsilon_i} \nabla_i \mathsf{L}(\pi_1^t, \ldots, \pi_k^t)(\mu_i \otimes \mu_i') \right], \quad 1 \leqslant i \leqslant k. \tag{61}$$

Here, both the gradient and projection steps are calculated in $L^2(\mu \otimes \mu')$. The L^2-projection $\mathrm{Proj}^{F}_{\Pi(\mu, \mu')}(A)$ can be carried out by solving a quadratically regularized optimal transport problem. This formulation has the notable advantage of producing couplings which are sparse, i.e. identically zero outside of a support set [81, 42].

Algorithm 4 Projected gradient: labelled k-partitioned networks, regularized matchings

```
Input: Matrices $\{\omega_{ij}\}_{i,j=1}^{k}, \{\omega'_{ij}\}_{i,j=1}^{k}$, probability measures $\mu_i, \mu_i'$, $1 \leqslant i \leqslant k$, label cost matrices $(C_i)_{i=1}^{k}$ (optional).
Parameters: entropic regularization levels $\varepsilon_i \geqslant 0, 1 \leqslant i \leqslant k$ (optional)
Initialize couplings: $\pi_i^1 \leftarrow \mu_i \otimes \mu_i'$, $1 \leqslant i \leqslant k$.
for $t = 1, 2, \ldots, \texttt{max\_iter}$ do
    for $1 \leqslant i \leqslant k$ do
        $M_i \leftarrow \frac{1}{2} \left( L(\omega_{ii}, \omega'_{ii}) + L(\omega_{ii}^\top, {\omega'}_{ii}^\top) \right) \otimes \pi_i^t + M[\pi_{-i}^t] + C_i$    (see (51))
        $\pi_i^{t+1} \leftarrow \arg\min_{\pi_i \in \Pi(\mu_i, \mu_i')} \varepsilon_i \, \mathrm{KL}(\pi_i | e^{-\varepsilon_i^{-1} M_i} \mu_i \otimes \mu_i')$
    end for
end for
Output: couplings $\{\pi_i\}_{i=1}^{k}$
```

A.5 Unbalanced matchings

We now consider the setting of unbalanced transport, in which marginal constraints are relaxed and replaced with penalty functions that enforce a "soft" marginal constraint. Unbalanced transport has been well studied from both a theoretical and practical viewpoint for the transportation of measures [17, 16, 41], and has since been extended to the setting of (Fused) Gromov-Wasserstein matchings between metric measure spaces [59, 67] and co-optimal transport [69]. An unbalanced formulation of the partitioned network alignment problem is valuable in practical settings when there may only be partial correspondences between networks, such as in the metabolic network alignment example of Figure 3.

The unbalanced transport problem for labelled partitioned measure networks includes unbalanced (Fused) Gromov-Wasserstein [59, 67] and co-optimal transport [69] as sub-cases. We let $\lambda_{1,2} > 0$ enforce the soft marginal constraints for the source and target respectively, and we denote by $\mathcal{M}_+(X)$ the space of positive measures supported on X. For generality, and because this makes efficient computational schemes possible, we optionally allow an entropic regularization with a coefficient $\varepsilon \geqslant 0$. Then we pose the problem of (entropically regularized) *unbalanced* matching as

$$\begin{aligned}\min_{\substack{\pi_i \in \mathcal{M}_+(X_i \times X_i'),\, 1 \leqslant i \leqslant k, \\ m(\pi_i) = m(\pi_j),\, \forall 1 \leqslant i,j \leqslant k}} \quad & \widetilde{\mathsf{L}}(\pi_1, \dots, \pi_k) + \varepsilon \sum_{i,j=1}^{k} \mathrm{KL}(\pi_i \otimes \pi_j | \mu_i \otimes \mu_i' \otimes \mu_j \otimes \mu_j') \\ & + \lambda_1 \sum_{i,j=1}^{k} \mathrm{KL}(\pi_i \mathbf{1} \otimes \pi_j \mathbf{1} | \mu_i \otimes \mu_j) + \lambda_2 \sum_{i,j=1}^{k} \mathrm{KL}(\pi_i^\top \mathbf{1} \otimes \pi_j^\top \mathbf{1} | \mu_i' \otimes \mu_j').\end{aligned} \tag{62}$$

In the above, $\widetilde{\mathsf{L}}$ denotes a variant of the objective function (52), modified to ensure that the function $\widetilde{\mathsf{L}}$ and overall objective to be minimized remains homogeneous in $(\pi_1, \dots, \pi_k)$:

$$\widetilde{\mathsf{L}}(\pi_1, \dots, \pi_k) = \frac{1}{2} \sum_{i,j=1}^{k} \langle L(\omega_{ij}, \omega_{ij}'), \pi_i \otimes \pi_j \rangle + \sum_{i=1}^{k} m(\pi_i) \langle C_i, \pi_i \rangle, \tag{63}$$

where $m(\pi) = \int \mathrm{d}\pi$ is the total mass of π. This is different to the setup in [67], in which the quadratic nature of the Gromov-Wasserstein term in the coupling π conflicts with the linearity of the fused term. Importantly, $\widetilde{\mathsf{L}}$ coincides with L when its inputs are restricted to be probability measures, i.e. $m(\pi_i) = 1$.

Extending the definition of partitioned measure networks (Definition 2.7), we will allow μ_i to be positive measures in $\mathcal{M}_+(X_i)$ for each $1 \leqslant i \leqslant k$, but require that $m(\mu_i) = m(\mu_j), i \neq j$. The motivation is to eliminate the non-uniqueness under scaling (e.g., $(\mu_i, \mu_j) \to (\lambda \mu_i, \lambda^{-1} \mu_j)$) that becomes a particular issue in the special case of co-optimal transport [69].

First, we state the following identity for the KL-divergence (see (54)),

$$\begin{aligned}\mathrm{KL}(\alpha \otimes \beta | \alpha' \otimes \beta') &= m(\beta)\, \mathrm{H}(\alpha|\alpha') + m(\alpha)\, \mathrm{H}(\beta|\beta') - m(\alpha) m(\beta) + m(\alpha') m(\beta') \\ &= m(\beta)\, \mathrm{KL}(\alpha|\alpha') + m(\alpha)\, \mathrm{KL}(\beta|\beta') + (m(\alpha) - m(\alpha'))(m(\beta) - m(\beta')),\end{aligned} \tag{64}$$

where we have defined the *relative entropy* term

$$\mathrm{H}(\alpha|\alpha') = \int \log\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\alpha'}\right) \mathrm{d}\alpha. \tag{65}$$

Thus, $(\alpha, \beta) \mapsto \mathrm{KL}(\alpha \otimes \beta | \alpha' \otimes \beta')$ is 2-homogeneous up to additive constants [59]. Since $\widetilde{\mathsf{L}}$ is 2-homogeneous, the objective of (62) is also 2-homogeneous. We remark that this is important, since if we used L instead of $\widetilde{\mathsf{L}}$, under the scaling $\pi_i \mapsto \lambda \pi_i$, quadratic terms would dominate when $\lambda \to +\infty$ and linear terms when $\lambda \to 0$.
Special case: measure hypernetworks. A special case is when $k = 2$ and $\omega_{ii} = \omega_{ii}' = 0$: this amounts to an *unbalanced, fused* co-optimal transport problem. This was the focus of [69] which considered the unlabelled setting, and we discuss it for completeness. In this case, $\widetilde{\mathsf{L}}$ can be written in a bilinear form in (π_1, π_2):

$$\widetilde{\mathsf{L}}(\pi_1, \pi_2) = \left\langle \frac{1}{2} L(\omega_{12}, \omega_{12}') + \frac{1}{2} L(\omega_{21}^\top, \omega_{21}'^\top) + C_1 \otimes \mathbf{1} + \mathbf{1} \otimes C_2, \pi_1 \otimes \pi_2 \right\rangle. \tag{66}$$

Up to additive constants, the problem (62) for $k = 2$ and $\omega_{ii} = \omega'_{ii} = 0$ can be re-written as

$$\begin{aligned}\min_{\pi_1,\pi_2} \widetilde{\mathsf{L}}(\pi_1,\pi_2) &+ 2\lambda_1 \left((m(\pi_1)+m(\pi_2))\,\mathrm{H}(\pi_1\mathbf{1}|\mu_1) + (m(\pi_1)+m(\pi_2))\,\mathrm{H}(\pi_2\mathbf{1}|\mu_2)\right) \\ &+ 2\lambda_2 \big((m(\pi_1)+m(\pi_2))\,\mathrm{H}(\pi_1^\top\mathbf{1}|\mu_1') + (m(\pi_1)+m(\pi_2))\,\mathrm{H}(\pi_2^\top\mathbf{1}|\mu_2')\big) \\ &+ 2\varepsilon \big((m(\pi_1)+m(\pi_2))\,\mathrm{H}(\pi_1|\mu_1\otimes\mu_1') + (m(\pi_1)+m(\pi_2))\,\mathrm{H}(\pi_2|\mu_2\otimes\mu_2'))\big) \\ &- (\lambda_1+\lambda_2+\varepsilon)\big(m(\pi_1)^2 + m(\pi_2)^2 + 2m(\pi_1)m(\pi_2)\big)\,,\end{aligned} \tag{67}$$

where the minimum is taken over $\pi_1 \in \mathcal{M}_+(X_1 \times X_1')$ and $\pi_2 \in \mathcal{M}_+(X_2 \times X_2')$ such that $m(\pi_1) = m(\pi_2)$. Notice that $\widetilde{\mathsf{L}}$ is bilinear, but the terms corresponding to the soft marginal constraints contain quadratic terms in π_1 and π_2. However, since we have the constraint $m(\pi_1) = m(\pi_2)$, we could judiciously swap $m(\pi_1)$ and $m(\pi_2)$ to derive an objective that remains equivalent under the mass equality constraint:

$$\begin{aligned}\min_{\pi_1,\pi_2} \widetilde{\mathsf{L}}(\pi_1,\pi_2) &+ 4\lambda_1 \left(m(\pi_2)\,\mathrm{H}(\pi_1\mathbf{1}|\mu_1) + m(\pi_1)\,\mathrm{H}(\pi_2\mathbf{1}|\mu_2)\right) \\ &+ 4\lambda_2 \big(m(\pi_2)\,\mathrm{H}(\pi_1^\top\mathbf{1}|\mu_1') + m(\pi_1)\,\mathrm{H}(\pi_2^\top\mathbf{1}|\mu_2')\big) \\ &+ 4\varepsilon \big(m(\pi_2)\,\mathrm{H}(\pi_1|\mu_1\otimes\mu_1') + m(\pi_1)\,\mathrm{H}(\pi_2|\mu_2\otimes\mu_2')\big) \\ &- 4(\lambda_1+\lambda_2+\varepsilon)m(\pi_1)m(\pi_2).\end{aligned} \tag{68}$$

In what follows, we drop the factor of 4 as it can be absorbed into the coefficients:

$$\begin{aligned}\min_{\pi_1,\pi_2} \widetilde{\mathsf{L}}(\pi_1,\pi_2) &+ \lambda_1 \left(m(\pi_2)\,\mathrm{H}(\pi_1\mathbf{1}|\mu_1) + m(\pi_1)\,\mathrm{H}(\pi_2\mathbf{1}|\mu_2)\right) \\ &+ \lambda_2 \big(m(\pi_2)\,\mathrm{H}(\pi_1^\top\mathbf{1}|\mu_1') + m(\pi_1)\,\mathrm{H}(\pi_2^\top\mathbf{1}|\mu_2')\big) \\ &+ \varepsilon \big(m(\pi_2)\,\mathrm{H}(\pi_1|\mu_1\otimes\mu_1') + m(\pi_1)\,\mathrm{H}(\pi_2|\mu_2\otimes\mu_2')\big) \\ &- (\lambda_1+\lambda_2+\varepsilon)m(\pi_1)m(\pi_2).\end{aligned} \tag{69}$$

We will for now remove the constraint $m(\pi_1) = m(\pi_2)$. In this setting, we can tackle the problem by alternating block minimization in (π_1, π_2). Fixing π_2 and rearranging the objective function, the update in π_1 amounts to

$$\min_{\pi_1\in\mathcal{M}_+(X_1\times X_1')} \langle M[\pi_2], \pi_1\rangle + \lambda_1 m(\pi_2)\,\mathrm{KL}(\pi_1\mathbf{1}|\mu_1) + \lambda_2 m(\pi_2)\,\mathrm{KL}(\pi_1^\top\mathbf{1}|\mu_1') + \varepsilon m(\pi_2)\,\mathrm{KL}(\pi_1|\mu_1\otimes\mu_1'), \tag{70}$$

where

$$\begin{aligned}M[\pi_2] = &\left(\frac{1}{2}L(\omega_{12},\omega_{12}') + \frac{1}{2}L(\omega_{21}^\top,\omega'^{\top}_{21})\right)\otimes\pi_2 \\ &+ m(\pi_2)C_1 + \big(\langle C_2,\pi_2\rangle + \lambda_1\,\mathrm{H}(\pi_2\mathbf{1}|\mu_2) + \lambda_2\,\mathrm{H}(\pi_2^\top\mathbf{1}|\mu_2') + \varepsilon\,\mathrm{H}(\pi_2|\mu_2\otimes\mu_2')\big)\,\mathbf{1}.\end{aligned} \tag{71}$$

Fixing π_1, the update in π_2 is

$$\min_{\pi_2\in\mathcal{M}_+(X_2\times X_2')} \langle M[\pi_1], \pi_2\rangle + \lambda_1 m(\pi_1)\,\mathrm{KL}(\pi_2\mathbf{1}|\mu_2) + \lambda_2 m(\pi_1)\,\mathrm{KL}(\pi_2^\top\mathbf{1}|\mu_2') + \varepsilon m(\pi_1)\,\mathrm{KL}(\pi_2|\mu_2\otimes\mu_2'), \tag{72}$$

where

$$\begin{aligned}M[\pi_1] = &\left(\frac{1}{2}L(\omega_{12}^\top,\omega'^{\top}_{12}) + \frac{1}{2}L(\omega_{21},\omega_{21}')\right)\otimes\pi_1 \\ &+ m(\pi_1)C_2 + \big(\langle C_1,\pi_1\rangle + \lambda_1\,\mathrm{H}(\pi_1\mathbf{1}|\mu_1) + \lambda_2\,\mathrm{H}(\pi_1^\top\mathbf{1}|\mu_1') + \varepsilon\,\mathrm{H}(\pi_1|\mu_1\otimes\mu_1')\big)\,\mathbf{1}.\end{aligned} \tag{73}$$

To enforce the equal mass constraint $m(\pi_1) = m(\pi_2)$, given (π_1, π_2) with $m(\pi_1) \neq m(\pi_2)$, we use the following to project onto the constraint set:

$$(\pi_1,\pi_2) \mapsto \left(\sqrt{\frac{m(\pi_2)}{m(\pi_1)}}\pi_1, \sqrt{\frac{m(\pi_1)}{m(\pi_2)}}\pi_2\right). \tag{74}$$

Algorithm 5 Unbalanced matchings: labelled measure hypernetworks

```
**Input:** Matrices $\omega_{12}, \omega'_{12}$, positive measures $\mu_i, \mu'_i$, $i = 1, 2$, label cost matrices $C_{1,2}$ (optional).
**Parameters:** Entropic regularisation parameter $\varepsilon \geqslant 0$, unbalanced parameters $\lambda_1, \lambda_2 > 0$.
Initialize couplings: $\pi_i \leftarrow \mu_i \otimes \mu'_i / \sqrt{m(\mu_i) m(\mu'_i)}$, $i = 1, 2$.
**for** $t = 1, 2, \ldots,$ max_iter **do**
    $\pi_1 \leftarrow \underset{\pi_1 \in \mathcal{M}_+(X_1 \times X'_1)}{\arg\min} \langle M[\pi_2], \pi_1 \rangle + \lambda_1 m(\pi_2) \operatorname{KL}(\pi_1 \mathbf{1} | \mu_1) + \lambda_2 m(\pi_2) \operatorname{KL}(\pi_1^\top \mathbf{1} | \mu'_1) + \varepsilon m(\pi_2) \operatorname{KL}(\pi_1 | \mu_1 \otimes \mu'_1)$
    $\pi_1 \leftarrow \sqrt{\frac{m(\pi_2)}{m(\pi_1)}} \pi_1$
    $\pi_2 \leftarrow \underset{\pi_2 \in \mathcal{M}_+(X_2 \times X'_2)}{\arg\min} \langle M[\pi_1], \pi_2 \rangle + \lambda_1 m(\pi_1) \operatorname{KL}(\pi_2 \mathbf{1} | \mu_2) + \lambda_2 m(\pi_1) \operatorname{KL}(\pi_2^\top \mathbf{1} | \mu'_2) + \varepsilon m(\pi_1) \operatorname{KL}(\pi_2 | \mu_2 \otimes \mu'_2)$
    $\pi_2 \leftarrow \sqrt{\frac{m(\pi_1)}{m(\pi_2)}} \pi_2$
**end for**
**Output:** couplings $\{\pi_i\}_{i=1}^k$
```

This projection, also used in [69], can be shown to be equivalent to the KL-projection of (π_1, π_2) onto the set $m(\pi_1) = m(\pi_2)$.

General case. For general partitioned measure networks, the objective function $\widetilde{\mathsf{L}}$ introduces terms that are non-trivially quadratic in π_i and is therefore less straightforward to solve. While in the balanced case these kinds of problems are typically tackled using a Frank-Wolfe algorithm [44], such an approach is not feasible for problems with soft constraints. As done in related works [67, 59], we propose to solve the problem instead via a biconvex relaxation. Consider *two* partitioned couplings $(\pi_1, \ldots, \pi_k)$ and $(\xi_1, \ldots, \xi_k)$. We us the relaxation of (62):

$$\begin{aligned} \min_{\substack{\pi_i, \xi_i \in \mathcal{M}_+(X_i \times X'_i), 1 \leqslant i \leqslant k \\ m(\pi_i) = m(\xi_j), 1 \leqslant i,j \leqslant k}} & \widetilde{\mathsf{L}}(\pi_1, \ldots, \pi_k; \xi_1, \ldots, \xi_k) + \varepsilon \sum_{i,j=1}^k \operatorname{KL}(\pi_i \otimes \xi_j | \mu_i \otimes \mu'_i \otimes \mu_j \otimes \mu'_j) \\ & + \lambda_1 \sum_{i,j=1}^k \operatorname{KL}(\pi_i \mathbf{1} \otimes \xi_j \mathbf{1} | \mu_i \otimes \mu_j) + \lambda_2 \sum_{i,j=1}^k \operatorname{KL}(\pi_i^\top \mathbf{1} \otimes \xi_j^\top \mathbf{1} | \mu'_i \otimes \mu'_j), \end{aligned} \tag{75}$$

where we define the relaxed version of (63):

$$\widetilde{\mathsf{L}}(\pi_1, \ldots, \pi_k; \xi_1, \ldots, \xi_k) = \frac{1}{2} \sum_{i,j=1}^k \langle L(\omega_{ij}, \omega'_{ij}), \pi_i \otimes \xi_j \rangle + \sum_{i=1}^k \left\langle \frac{1}{2}(C_i \otimes \mathbf{1} + \mathbf{1} \otimes C_i), \pi_i \otimes \xi_i \right\rangle.$$

The form of the second term ensures symmetry under the exchange of (π, ξ) and that $\widetilde{\mathsf{L}}(\pi_1, \ldots, \pi_k; \pi_1, \ldots, \pi_k) = \mathsf{L}(\pi_1, \ldots, \pi_k)$. The problem (75) is now convex separately in $(\pi_i)_{i=1}^k$ and $(\xi_i)_{i=1}^k$ respectively. Fixing $(\xi_i)_{i=1}^k$

and minimizing in $(\pi_i)_{i=1}^k$, we find that the problem decouples across partitions in each of the π_i:

$$\begin{aligned}\min_{\substack{\pi_i \in \mathcal{M}_+(X_i \times X_i'), 1 \leqslant i \leqslant k \\ m(\pi_i) = m(\pi_j), 1 \leqslant i,j \leqslant k}} & \widetilde{\mathsf{L}}(\pi_1, \dots, \pi_k; \xi_1, \dots, \xi_k) \\ & + \left(\sum_i m(\pi_i)\right)\left(\varepsilon \sum_j \mathrm{H}(\xi_j | \mu_j \otimes \mu_j') + \lambda_1 \sum_j \mathrm{H}(\xi_j \mathbf{1} | \mu_j) + \lambda_2 \sum_j \mathrm{H}(\xi_j^\top \mathbf{1} | \mu_j')\right) \\ & + \lambda_1 \left(\sum_j m(\xi_j)\right) \sum_i \mathrm{KL}(\pi_i \mathbf{1} | \mu_i) + \lambda_2 \left(\sum_j m(\xi_j)\right) \sum_i \mathrm{KL}(\pi_i^\top \mathbf{1} | \mu_i') \\ & + \varepsilon \left(\sum_j m(\xi_j)\right) \sum_i \mathrm{KL}(\pi_i | \mu_i \otimes \mu_i'). \end{aligned} \tag{76}$$

Relaxing the mass equality constraint $m(\pi_i) = m(\pi_j), 1 \leqslant i, j \leqslant k$, the above problem amounts to k regularized unbalanced optimal transport problems that can be solved independently and in parallel. The resulting couplings can be projected onto the set $\{m(\pi_i) = m(\pi_j), 1 \leqslant i, j \leqslant k\}$:

$$(\pi_i)_{i=1}^k \mapsto \left(\frac{(m(\pi_1) \dots m(\pi_k))^{1/k}}{m(\pi_i)} \pi_i\right)_{i=1}^k. \tag{77}$$

Similarly, fixing $(\pi_i)_{i=1}^k$, the problem in $(\xi_i)_{i=1}^k$ is

$$\begin{aligned}\min_{\substack{\xi_i \in \mathcal{M}_+(X_i \times X_i'), 1 \leqslant i \leqslant k \\ m(\xi_i) = m(\xi_j), 1 \leqslant i,j \leqslant k}} & \widetilde{\mathsf{L}}(\pi_1, \dots, \pi_k; \xi_1, \dots, \xi_k) \\ & + \left(\sum_j m(\xi_j)\right)\left(\varepsilon \sum_i \mathrm{H}(\pi_i | \mu_i \otimes \mu_i') + \lambda_1 \sum_i \mathrm{H}(\pi_i \mathbf{1} | \mu_i) + \lambda_2 \sum_i \mathrm{H}(\pi_i^\top \mathbf{1} | \mu_i')\right) \\ & + \lambda_1 \left(\sum_i m(\pi_i)\right) \sum_j \mathrm{KL}(\xi_j \mathbf{1} | \mu_j) + \lambda_2 \left(\sum_i m(\pi_i)\right) \sum_j \mathrm{KL}(\xi_j^\top \mathbf{1} | \mu_j') \\ & + \varepsilon \left(\sum_i m(\pi_i)\right) \sum_j \mathrm{KL}(\xi_j | \mu_j \otimes \mu_j'), \end{aligned} \tag{78}$$

and the same projection (77) can be used to enforce the mass equality constraint in $(\xi_i)_{i=1}^k$. It is important to note that this scheme aims to solve the biconvex relaxation (75) which is in general only a lower bound for (62). In particular, at convergence, we may have $\pi_i \neq \xi_i$ in general. While this biconvex relaxation scheme was studied for the Gromov-Wasserstein setting by [59], they were unable to prove tightness or that the two sets of couplings $(\pi_i)_i, (\xi_i)_i$ coincide.

A remark on partial transport. On the other hand we may consider *partial* transport, where some fraction $0 \leqslant s \leqslant 1$ of mass is required to be transported with the remainder being discarded and thus incurring zero cost. This problem was considered by [15] in the case of Gromov-Wasserstein transport. For two probability measures μ, μ', define the set of partial couplings of mass s to be

$$\Pi(\mu, \mu'; s) = \{\pi \geqslant 0 : \pi \mathbf{1} \leqslant \mu, \pi^\top \mathbf{1} \leqslant \mu', m(\pi) = s\}.$$

Then, the partial matching problem amounts to solving

$$\min_{\pi_i \in \Pi(\mu_i, \mu_i'; s)} \mathsf{L}(\pi_1, \dots, \pi_k). \tag{79}$$

This amounts to the minimization of a non-convex objective on a convex and compact constraint set, and similar to [15], we can tackle it via conditional gradient method. In particular, to compute the descent directions in each of the π_i, we need to solve:

$$\min_{\pi_i \in \Pi(\mu_i, \mu_i'; s)} \langle \nabla_i \mathsf{L}(\pi_1^t, \dots, \pi_k^t), \pi_i \rangle.$$

Each of these is a partial optimal transport problem which can be solved using the virtual point approach of [15].

Algorithm 6 Unbalanced matchings: labelled partitioned measure networks via biconvex relaxation

```
Input: Matrices $\{\omega_{ij}\}_{i,j=1}^k, \{\omega'_{ij}\}_{i,j=1}^k$, positive measures $\mu_i, \mu'_i$, $1 \leqslant i \leqslant k$, label cost matrices $(C_i)_{i=1}^k$ (optional)
Parameters: Marginal penalties $\lambda_1, \lambda_2 > 0$, entropic regularization $\varepsilon \geqslant 0$ (optional).
Initialize couplings: $\pi_i \leftarrow \mu_i \otimes \mu'_i / \sqrt{m(\mu_i)m(\mu'_i)}$, $1 \leqslant i \leqslant k$.
Initialize additional couplings: $\xi_i \leftarrow \pi_i$, $1 \leqslant i \leqslant k$.
for $t = 1, 2, \ldots,$ max_iter do
    Update $(\pi_1, \ldots, \pi_k)$ by solving (76) independently for each $1 \leqslant i \leqslant k$.
    Rescale $(\pi_i)_{i=1}^k$ following (77)
    Update $(\xi_1, \ldots, \xi_k)$ by solving (78) independently for each $1 \leqslant i \leqslant k$.
    Rescale $(\xi_i)_{i=1}^k$ following (77)
    $(\pi_i)_{i=1}^k, (\xi_i)_{i=1}^k \leftarrow \left(\sqrt{\frac{m(\xi)}{m(\pi)}}\pi_i\right)_{i=1}^k, \left(\sqrt{\frac{m(\pi)}{m(\xi)}}\xi_i\right)_{i=1}^k$
end for
Output: couplings $\{\pi_i\}_{i=1}^k, \{\xi_i\}_{i=1}^k$
```

A.6 Partitioned networks for multiscale network matching

Chowdhury et al. [22] introduced a generalized co-optimal transport problem for multiscale network matching. Given an input graph G, they produced successive topological simplifications $\mathcal{G} = \{G = G_1, \ldots, G_k\}$. At each level $1 \leqslant i \leqslant k-1$, the nodes of G_i are partitioned among the nodes of G_{i+1}. In this way, the coupling of G_i to G_{i+1} can be modelled as a hypergraph in which nodes and hyperedges are identified with nodes in G_i and G_{i+1} respectively. We now show that $\mathcal{G}$ can be formulated as a partitioned network with k partitions. Let X_i be the node set of the ith simplification level G_i. Let $\omega_{i,i+1}$ (for $1 \leqslant i \leqslant k-1$) be the function encoding relations between nodes in the ith and nodes the $(i+1)$th simplification:

$$\omega(x,y) = \begin{cases} \omega_{i,i+1}(x,y), & x \in X_i, y \in X_{i+1} \text{ for } i = 1, \ldots, k-1; \\ \omega_{i+1,i}(x,y), & x \in X_{i+1}, y \in X_i \text{ for } i = 1, \ldots, k-1; \\ 0, & \text{otherwise.} \end{cases}$$

Together with a choice of weights $(\mu_i)_{i=1}^k$, $((X_i, \mu_i)_{i=1}^k, \omega)$ is a partitioned measure network encoding the multiscale network $\mathcal{G}$.

Given two graphs G and G' and their respective simplifications $\mathcal{G}$ and $\mathcal{G}'$, we can then construct two partitioned measure networks: $((X_i, \mu_i)_{i=1}^k, \omega)$ and $((X'_i, \mu'_i)_{i=1}^k, \omega')$. For a candidate coupling $(\pi_i)_{i=1}^k$, the corresponding distortion functional is

$$\sum_{i=1}^{k-1} \|\omega_{i,i+1} - \omega'_{i,i+1}\|^p_{L^p(\pi_i \otimes \pi_{i+1})}. \tag{80}$$

The partitioned measure network alignment problem induced by this distortion is equivalent to the one proposed in [22, Algorithm 1], i.e.

$$\min_{\pi_i \in \Pi(\mu_i, \mu'_i), 1 \leqslant i \leqslant k} \sum_{i=1}^{k-1} \langle L(\omega_{i,i+1}, \omega'_{i,i+1}), \pi_i \otimes \pi_{i+1} \rangle. \tag{81}$$

This multiscale graph matching problem therefore fits into the problem of matchings of partitioned measure networks. Encouragingly, while the formulation of [22] was in terms of pairs of couplings $(\pi_i, \xi_i), 1 \leqslant i \leqslant k-1$ under the constraint $\xi_i = \pi_{i+1}$, the derivation of the problem from the viewpoint of partitioned measure networks allows us to directly and naturally formulate the problem in terms of a single set of couplings $(\pi_1, \ldots, \pi_k)$.

Furthermore, by modifying the function ω, we can incorporate pairwise information on each of the graphs

G_i:

$$\omega(x,y) = \begin{cases} \omega_{i,i+1}(x,y), & x \in X_i, y \in X_{i+1} \text{ for } i = 1,\dots,k-1; \\ \omega_{i+1,i}(x,y), & x \in X_{i+1}, y \in X_i \text{ for } i = 1,\dots,k-1; \\ \omega_{ii}(x,y), & x, y \in X_i \times X_i \text{ for } i = 1,\dots,k; \\ 0, & \text{otherwise.} \end{cases}$$

This choice of ω leads to the problem

$$\min_{\pi_i \in \Pi(\mu_i,\mu'_i)} \sum_{i=1}^{k-1} \langle L(\omega_{i,i+1},\omega'_{i,i+1}), \pi_i \otimes \pi_{i+1}\rangle + \frac{1}{2}\sum_{i=1}^{k} \langle L(\omega_{ii},\omega'_{ii}), \pi_i \otimes \pi_i\rangle + \sum_{i=1}^{k} \varepsilon_i \operatorname{KL}(\pi_i|\mu_i \otimes \mu'_i), \tag{82}$$

which incorporates Gromov-Wasserstein like (i.e. quadratic in π) terms. In the above we allow optionally for entropy regularization, $\varepsilon_i \geqslant 0$. For $\varepsilon_i > 0$, applying the projected gradient descent approach of Section A.4 leads to the update rule

$$\begin{aligned} &\pi_i^{t+1} \leftarrow \operatorname{Proj}^{\mathrm{KL}}_{\Pi(\mu_i,\mu'_i)} \left(e^{-\varepsilon_i^{-1}\nabla_i \mathsf{L}(\pi_1^t,\dots,\pi_k^t)} \mu_i \otimes \mu'_i \right), \\ &\nabla_i \mathsf{L}(\pi_1,\dots,\pi_k) = \\ &\begin{cases} L(\omega_{12},\omega'_{12}) \otimes \pi_2 + \frac{1}{2}\left(L(\omega_{11},\omega'_{11}) + L(\omega_{11}^\top,\omega'^\top_{11})\right) \otimes \pi_1, & i = 1; \\ L(\omega_{i-1,i}^\top,\omega'_{i-1,i}) \otimes \pi_{i-1} + L(\omega_{i,i+1},\omega'_{i,i+1}) \otimes \pi_{i+1} + \frac{1}{2}\left(L(\omega_{ii},\omega'_{ii}) + L(\omega_{ii}^\top,\omega'^\top_{ii})\right) \otimes \pi_i, & 2 \leqslant i \leqslant k-1; \\ L(\omega_{k-1,k}^\top,\omega'^\top_{k-1,k}) \otimes \pi_{k-1} + \frac{1}{2}\left(L(\omega_{kk},\omega'_{kk}) + L(\omega_{kk}^\top,\omega'^\top_{kk})\right) \otimes \pi_k, & i = k. \end{cases} \end{aligned} \tag{83}$$

When we look for an unregularized solution and $\varepsilon_i = 0$, a block coordinate descent scheme similar to the one proposed in [22] can be employed. The block update in each of the π_i works out to be a Fused Gromov-Wasserstein problem which can be tackled for instance using the Frank-Wolfe scheme of [71]. Alternatively, a proximal gradient approach similar to the one described in Section A.3 can be employed, in which case the gradient steps are the same as in (57).

An unbalanced formulation of this problem can also be solved by using the same biconvex relaxation approach laid out in Section A.5. While this problem falls into the scope of Algorithm 6, it is in fact a sub-case since each partition i is only coupled to its "adjoining" partitions (rather than all partitions in the general case). We detail below the specific updates for the biconvex relaxation, in terms of two sets of couplings, $(\pi_i)_{i=1}^k, (\xi_i)_{i=1}^k$. For each of the $\pi_i, 1 \leqslant i \leqslant k$, solve

$$\begin{aligned} \min_{\pi_i \in \mathcal{M}_+(X_i \times X'_i)} \langle M_i, \pi_i \rangle &+ \left(\varepsilon \sum_j m(\xi_j)\right) \operatorname{KL}(\pi_i|\mu_i \otimes \mu'_i) \\ &+ \left(\lambda_1 \sum_j m(\xi_j)\right) \operatorname{KL}(\pi_i \mathbf{1}|\mu_i) + \left(\lambda_2 \sum_j m(\xi_j)\right) \operatorname{KL}(\pi_i^\top \mathbf{1}|\mu'_i) \end{aligned} \tag{84}$$

where

$$\begin{aligned} M_i &= L_i + \sum_j \left(\varepsilon \operatorname{H}(\xi_j|\mu_j \otimes \mu'_j) + \lambda_1 \operatorname{H}(\xi_j \mathbf{1}|\mu_j) + \lambda_2 \operatorname{H}(\xi_j^\top \mathbf{1}|\mu'_j)\right), \\ L_i &= \begin{cases} \frac{1}{2} L(\omega_{12},\omega'_{12}) \otimes \xi_2 + \frac{1}{2} L(\omega_{11},\omega'_{11}) \otimes \xi_1, & i = 1; \\ \frac{1}{2}\left(L(\omega_{i,i+1},\omega'_{i,i+1}) \otimes \xi_{i+1} + L(\omega_{i-1,i}^\top,\omega'^\top_{i-1,i}) \otimes \xi_{i-1}\right) + \frac{1}{2} L(\omega_{ii},\omega'_{ii}) \otimes \xi_i, & 2 \leqslant i \leqslant k-1; \\ \frac{1}{2} L(\omega_{k-1,k}^\top,\omega'^\top_{k-1,k}) \otimes \xi_{k-1} + \frac{1}{2} L(\omega_{kk},\omega'_{kk}) \otimes \xi_k & i = k. \end{cases} \end{aligned} \tag{85}$$

Similarly, in each of the $\xi_i, 1 \leqslant i \leqslant k$, we solve

$$\begin{aligned} \min_{\xi_i \in \mathcal{M}_+(X_i \times X'_i)} \langle M'_i, \xi_i \rangle &+ \left(\varepsilon \sum_j m(\pi_j)\right) \operatorname{KL}(\xi_i|\mu_i \otimes \mu'_i) \\ &+ \left(\lambda_1 \sum_j m(\pi_j)\right) \operatorname{KL}(\xi_i \mathbf{1}|\mu_i) + \left(\lambda_2 \sum_j m(\pi_j)\right) \operatorname{KL}(\xi_i^\top \mathbf{1}|\mu'_i), \end{aligned} \tag{86}$$

where

$$\begin{aligned} M_i' &= L_i' + \sum_j \left(\varepsilon \, \mathrm{H}(\pi_j | \mu_j \otimes \mu_j') + \lambda_1 \, \mathrm{H}(\pi_j \mathbf{1} | \mu_j) + \lambda_2 \, \mathrm{H}(\pi_j^\top \mathbf{1} | \mu_j') \right) , \\ L_i' &= \begin{cases} \frac{1}{2} L(\omega_{12}, \omega_{12}') \otimes \pi_2 + \frac{1}{2} L(\omega_{11}^\top, \omega_{11}'^\top) \otimes \pi_1, & i = 1; \\ \frac{1}{2} \left(L(\omega_{i,i+1}, \omega_{i,i+1}') \otimes \pi_{i+1} + L(\omega_{i-1,i}^\top, \omega_{i-1,i}'^\top) \otimes \pi_{i-1} \right) + \frac{1}{2} L(\omega_{ii}^\top, \omega_{ii}'^\top) \otimes \pi_i, & 2 \leqslant i \leqslant k-1; \\ \frac{1}{2} L(\omega_{k-1,k}^\top, \omega_{k-1,k}'^\top) \otimes \pi_{k-1} + \frac{1}{2} L(\omega_{kk}^\top, \omega_{kk}'^\top) \otimes \pi_k & i = k. \end{cases} \end{aligned} \tag{87}$$

These updates, together with the projections of Algorithm 6 onto the mass equality constraint sets, give a numerical approach to approximating a solution of the general unbalanced multiscale alignment problem.

A.7 Barycenters with fixed support

As an alternative to the blow-up scheme of [20], we can consider an approximation of the barycenter problem where we restrict our approach to seeking a minimizer over network representatives of a fixed size. This is the same as the approach of [52], which was developed in the setting of the Gromov-Wasserstein distance. That is, for an input ensemble of partitioned measure networks $\{P^{(i)}, 1 \leqslant i \leqslant N\}$, we consider a barycenter $\overline{P} = ((\overline{X}_i, \overline{\mu}_i)_{i=1}^k, \overline{\omega})$ in which we have fixed the cardinalities of $\overline{X}_i$ to $|\overline{X}_i| = n_i$. We also prescribe the probability measures $\overline{\mu}_i$ for $\overline{P}$, so that it remains to find the optimal function $\overline{\omega}$. Expanding the definition of the partitioned network distance between hypernetworks, we have

$$\min_{\overline{\omega}} \sum_{i=1}^N w_i d_{\mathcal{P}_k}(\overline{P}, P^{(i)})^2 = \min_{\overline{\omega}} \sum_{i=1}^N \left[w_i \min_{\pi^{(i)} \in \Pi_k(\overline{\mu}, \mu^{(i)})} \|\overline{\omega} - \omega^{(i)}\|_{L^2(\pi^{(i)} \otimes \pi^{(i)})}^2 \right] . \tag{88}$$

From this, it is apparent that an alternating scheme can be developed by minimizing separately in the couplings $\pi^{(i)} \in \Pi_k(\overline{\mu}, \mu^{(i)})$ and in the function $\overline{\omega}$. Fixing $\overline{\omega}$, the objective (88) can be minimized in each of the $\pi^{(i)}, 1 \leqslant i \leqslant N$ by solving N independent partitioned network matching problems. Fixing the couplings $\{\pi^{(i)}\}_{i=1}^N$, the minimization problem in $\overline{\omega}$ becomes

$$\min_{\overline{\omega}} \sum_i w_i \, \|\overline{\omega} - \omega^{(i)}\|_{L^2(\pi^{(i)} \otimes \pi^{(i)})}^2,$$

this amounts to minimizing a quadratic objective and therefore has a closed form solution.

Proposition A.2 (Barycenter update for fixed couplings)**.** *For fixed couplings $\pi^{(i)}, 1 \leqslant i \leqslant N$, the objective* (88) *is quadratic and minimized in $\overline{\omega}$ at*

$$\overline{\omega}_{jl}^\star = \frac{1}{\overline{\mu}_j \otimes \overline{\mu}_l} \sum_{i=1}^N w_i \pi_j^{(i)} \omega_{jl}^{(i)} {\pi_l^{(i)}}^\top, \quad 1 \leqslant j, l \leqslant k. \tag{89}$$

We note that the derivation of the form of this update is identical to that of [52], except for the presence of two possibly distinct couplings in the summand.