Similarity network aggregation for the analysis of glacier ecosystems

Similarity network aggregation for the analysis of glacier ecosystems

Roberto Ambrosini[1], Federica Baccini[2,*], Lucio Barabesi[3]

[1]Dipartimento di Scienze e Politiche Ambientali, Università degli Studi di Milano, Via Celoria, 26 - Corpo C, Milano, 20133, Italy
{roberto.ambrosini@unimi.it}

[2]Dipartimento di Ingegneria informatica, automatica e gestionale "Antonio Ruberti", Università degli Studi di Roma "La Sapienza", Via Ariosto, 25, Roma, 00185, Italy
{baccini@diag.uniroma1.it}

[3]Dipartimento di Economia Politica e Statistica, Università degli Studi di Siena, Piazza San Francesco, 7, Siena, 53100, Italy
{lucio.barabesi@unisi.it}

Abstract

The synthesis of information deriving from complex networks is a topic receiving increasing relevance in ecology and environmental sciences. In particular, the aggregation of multilayer networks, i.e. network structures formed by multiple interacting networks (the layers), constitutes a fast-growing field. In several environmental applications, the layers of a multilayer network are modelled as a collection of similarity matrices describing how similar pairs of biological entities are, based on different types of features (e.g. biological traits). The present paper first discusses two main techniques for combining the multi-layered information into a single network (the so-called monoplex), i.e. Similarity Network Fusion (SNF) and Similarity Matrix Average (SMA). Then, the effectiveness of the two methods is tested on a real-world dataset of the relative abundance of microbial species in the ecosystems of nine glaciers (four glaciers in the Alps and five in the Andes). A preliminary clustering analysis on the monoplexes obtained with different methods shows the emergence of a tightly connected community formed by species that are typical of cryoconite holes worldwide. Moreover, the weights assigned to different layers by the

*Corresponding author

SMA algorithm suggest that two large South American glaciers (Exploradores and Perito Moreno) are structurally different from the smaller glaciers in both Europe and South America. Overall, these results highlight the importance of integration methods in the discovery of the underlying organizational structure of biological entities in multilayer ecological networks.

Keywords: Multilayer ecological networks; Similarity network fusion; Similarity matrix averaging; Generalized distance correlation; Communities in networks.

Declarations

- Conflict of interest/Competing interests: The authors have no competing interests to declare that are relevant to the content of this article.
- Availability of data and materials: After acceptance, the data used in this study are available from the corresponding author upon reasonable request.
- Authors' contributions: Federica Baccini and Lucio Barabesi contributed to the analysis and developements of the methodology. Roberto Ambrosini collected the data. Federica Baccini performed the data analysis and the computation of the matrix integration methods. Roberto Ambrosini provided the data and interpreted them from an ecological perspective. All the authors contributed to the writing of the manuscript.

1 Introduction

The analysis of complex networks is a rapidly expanding area in data science, since many real-world structures can be explored by assessing the connections and the interplay between their parts (Barabási, 2013; Estrada, 2011; Newman, 2018). In the field of ecology, Mittelbach and McGill (2019) is an authoritative source for appraising the growing use and application of complex networks. Environmental and ecological networks aim at modelling the interactions among a set of biological and abiotic entities, which are represented as nodes of a graph. Interactions characterize living systems, because organisms necessarily interact with each other and with the environment. The reciprocal actions between ecological populations are the basis of community ecology and are central to shape biodiversity. As an example of ecological network, the different plant species in a certain ecosystem could be interpreted as the graph nodes, while pollination by the same pollinator species represents the link between them.

Recent advances in network science have pointed out that standard networks, i.e. networks constituted by pairwise interactions among a set of nodes, are often not sufficient to describe the full complexity of ecological interactions (Bianconi, 2018; De Domenico, 2022). In order to implement a more realistic depiction of an ecological system, a key ingredient consists in considering different types of interactions between organisms (Hutchinson et al., 2019; Pilosof et al., 2017). Indeed, by continuing the previously considered example, different plant species in an ecosystem are likely to be pollinated by multiple pollinator species, instead of a single one. In this setting, multilayer networks (Bianconi, 2018; De Domenico, 2022; Kivelä et al., 2014) are an efficient way to represent systems where multiple types of interactions (i.e. the layers) occur between the nodes. More explicitly, each layer of a multilayer network constitutes a graph, where the edges are representative of a given type of interaction. The nodes may be partially different in different layers, and even interlayer edges may be present. Hutchinson et al. (2019) provide many real examples of multilayer networks in ecology, with special emphasis on the interactions among species.

Different families of multilayer networks exist, each aimed at modelling different settings. A frequently-adopted type of multilayer network is the so-called 'multiplex' (Bianconi, 2018; De Domenico, 2022). A multiplex network has the same set of nodes on each layer (the node 'replicas'), and each layer represents a different relation type. The present work focuses on multiplex networks and, in particular, on techniques for aggregating the multi-layered information into a single graph, i.e. the so-called monoplex. The resulting aggregated network may be central to evaluate the strong connections between the different nodes and – eventually – to implement a cluster analysis for searching cohesive groups (densely connected groups of nodes, according to a chosen metric) in the original multiplex network. In addition, an appropriate analysis of the correlation between the monoplex and the single layers may be helpful to detect the global community structure.

We emphasize that a multiplex is often used to model similarities between pairs of entities (the nodes), where the similarity is computed based on several feature types. This defines a multiplex similarity network consisting of a collection of adjacency similarity matrices, whose dimension is given by the number of nodes (see e.g. F. Baccini, Barabesi, and Petrovich, 2023). In the framework of multiplex similarity networks, we discuss two main techniques for integrating the similarity networks of the multiplex in order to obtain a single monoplex with suitable properties. The first technique is Similarity Network Fusion, the widely-used approach initially suggested by Wang, Jiang, et al. (2012) and Wang, Mezlini, et al. (2014). Then, the more recent approach introduced by F. Baccini, Barabesi, and Petrovich (2023), namely Similarity Matrix Average, is revised and discussed. The latter approach is designed to "average" a collection of similarity matrices by exploiting advances in the theory of barycenters in appropriate matrix spaces. We finally emphasize that, even if a simple strategy for aggregating a multiplex network may be based on a weighted average of the original adjacency matrices corresponding to the layers, such a method could be unsuitable, since some layers may be over-represented and the choice of weights is subjective (see e.g. Kivelä et al., 2014, and references therein). Thus, an approach based on the integration of similarity matrices may be preferable.

The Similarity Network Fusion and the Similarity Network Average are actually descriptive methodologies. However, a body of literature exists that deals with model-based techniques for analyzing and detecting groups in multilayer networks. For example, we emphasize the following recent proposals. Barbillon et al. (2017) and Bar-Hen, Barbillon, and Donnet (2022) respectively consider a stochastic block model for multiplex networks based on a probabilistic structure of latent random variables for the nodes. MacDonald, Levina, and Zhu (2022) also suggest a latent stochastic block model for multiplex networks, which especially takes into account the part of network architecture shared by the layers. In addition, Krivitsky, Koehly, and Marcum (2020) introduce a generalized exponential family for modeling random multiplex networks. Finally, Kuncheva and Montana (2015) enhance community detection in multiplex networks by using a method based on locally adaptive random walks. In general, these model-based techniques assume rather stringent assumptions, such as the independence of the latent random variables, and involve a considerable burden in computing the maximum-likelihood estimates of the parameters, especially when the number of nodes is elevate, since the likelihood cannot be evaluated in a closed form. Moreover, when the model-based approach is finalized to grouping, the number of groups is often prefixed. For such reasons, we solely aim to introduce exploratory methods for aggregating multiplex networks, which do not require assumptions, and are not necessarily tailored for grouping. Thus, the two considered descriptive techniques could be suitable in the environmental framework.

The effectiveness of the Similarity Network Fusion and the Similarity Network Average for the study of ecological networks is tested on a real-world dataset describing the abundance of microbial species in glacier ecosystems. Glaciers host active ecological communities and have been recognized as ecosystems in their own right, where the dominant life forms are microbes, especially bacteria (Anesio and Laybourn-Parry, 2012; Hodson et al., 2008). Most ecological studies on glaciers have focused on cryoconite holes, i.e. small ponds that form on the glacier surface with a tiny layer of debris at the bottom, and are considered as the most biodiverse ecosystems in these harsh environments (Ambrosini et al., 2017; Cook et al., 2016). The scope of this analysis is twofold. First, we aim at giving insights into the similarities between microbial species based on their abundance in different glaciers. Subsequently, on the basis of the similarity between the analysed microbial species, we assess how much distinct glaciers are similar in terms of the relative abundance of microbial species.

The paper is organized as follows. Section 2 formalizes the problem of integrating the information on a multiplex similarity network; the same section also contains the description and the analysis of the techniques adopted for the similarity network aggregation procedure. Section 3 presents the application of the considered methods to the dataset registering the relative abundance of microbial species in cryoconiye holes of glaciers worldwide. Finally, some conclusions are drawn in Section 4.

2 Integration of multiplex similarity networks

In this section the problem of integrating the multiplex similarity network into the monoplex is described, by assuming two main possible geneses for the multiplex. In both geneses, the multiplex is modelled as a set of similarity matrices.

A first framework where a multiplex similarity network can originate is the case in which m measurements referring to a fixed set of entities are available. For each measurement, it is possible to define a graph (the "similarity network") where the vertices are the items, and the edges are weighted with a suitable similarity score between items. The similarity can be computed via suitable similarity measures, many choices of which are available (see De Bello et al., 2017; Eck and Waltman, 2009; Newman, 2018). The intuitive idea behind the use of similarity measures is that a high similarity score is associated to pairs of objects having a similar profile based on some measurement. Then, in the resulting similarity network, such objects will be connected by a tighter edge. If m measurements are available, m similarity networks (the "layers"), each one formed by the same set of nodes, are originated, and the resulting structure is referred to as "multiplex similarity network". Each layer of the multiplex is characterized by a similarity network whose weighted adjacency matrix is the corresponding similarity matrix. As a real-world example in the ecological setting, the items could represent n species in an ecosystem, and the measurements could denote m distinct functional traits. Similarities between species can then be computed based on each trait, thus producing a multiplex similarity network where each layer represents a specific trait. Properly, a multiplex similarity network is represented by means of m similarity matrices of order $(n \times n)$, namely $\mathbf{S}_1, \ldots, \mathbf{S}_m$.

A formalization of the genesis of such multiplex similarity network is now necessary. A generic entity, or object, is described, for the l-th measurement, as a vector of n observations, say $(x_{1l}, \ldots, x_{nl})$, for $l = 1, \ldots, m$. In general, the n observations could be also multivariate and can assume values in a given Euclidean space. If the l-th measurement is discrete or continuous, one option to compute similarities between observations is to consider a transform of the Euclidean distance as a measure of similarity. Thus, a possible choice for the elements of the similarity matrix $\mathbf{S}_l = (s_{ij}^{(l)})$ is given by

$$s_{ij}^{(l)} = \exp\left(-\frac{d_{\mathrm{E}}(x_{il}, x_{jl})^2}{\sigma}\right), \tag{1}$$

where $\sigma > 0$ is a scaling parameter to be fixed, and d_{E} is the appropriate Euclidean distance (see e.g. Wang, Jiang, et al., 2012). Alternatively, if the l-th measurement x_{il} assumes values in $\{0, 1\}$ for $i = 1, \ldots, n$, the similarity matrix $\mathbf{S}_l = (s_{ij}^{(l)})$ could be chosen as follows

$$s_{ij}^{(l)} = \mathbf{1}_{\{x_{il} = x_{jl}\}}, \tag{2}$$

where $\mathbf{1}_A$ represents the indicator function of a set A (see e.g. De Bello et al., 2017). When (2) is adopted, the similarity matrix $\mathbf{S}_l$ is a joint presence-absence matrix. When (1) or (2) are selected, $s_{ij}^{(l)}$ turns out to be a normalized similarity measure in $[0, 1]$ and $\mathbf{S}_l$ is symmetric.

A second framework where a multiplex structure arises concerns the modelling of relations among a fixed set of n groups described by the membership to such groups of m different sets of elements. This situation can be modelled by using a multilayer network characterized by the following two properties: (i) each layer is a bipartite network, and (ii) one vertex set of the bipartition is common to all the layers (the set of n groups), while the other one can change between layers (each of the m different sets of elements). As a real-world example in the ecological field, one can consider n groups of herbivore species that interact with some plant species, in m distinct ecosystems. A task of interest could be the analysis of the relations among the n herbivore species, taking into account all the distinct ecosystems. This situation can be modelled as a multilayer network structured as above, where each layer represents an ecosystem, and contains a bipartite network formed by the set of herbivore species (common to all the layers), and the set of plant species they graze in each ecosystem.

If the above multilayer structure is assumed, a multiplex network could be useful to explore the relations within the set of nodes that are common to all the layers. The multiplex can be obtained by considering the "one-mode projection" of each layer onto the common subset of nodes (the groups), which can in turn be used to compute suitable similarities among the set of groups. The one-mode projection is a particular network that can be computed starting from a bipartite network (see e.g. Newman, 2018). More specifically, let $\mathbf{B}_l = (b_{ij}^{(l)})$ denote the incidence matrix of order $(p_l \times n)$ corresponding to the bipartite network on the l-th layer, whose generic elements are defined as follows

$$b_{ij}^{(l)} = \begin{cases} 1 & \text{if item } i \text{ in the } l\text{-th layer belongs to group } j \\ 0 & \text{otherwise.} \end{cases}$$

Coherently with the previously-used terminology, the l-th bipartite network can be interpreted as describing the membership of p_l items to n groups.
The one-mode projection on the set of l items can then be defined as $\mathbf{G}_l = (g_{ij}^{(l)}) = \mathbf{B}_l^{\mathrm{T}} \mathbf{B}_l$. In particular, $\mathbf{G}_l$ is a symmetric matrix of order $(n \times n)$, whose generic element g_{ij} represents the number of items of the l-th layer which belong to both the i-th and j-th group.

Interestingly, as stated above, the one-mode projection $\mathbf{G}_l$ may be used to compute a similarity matrix $\mathbf{S}_l$. The cosine and the Jaccard similarity are two popular examples of similarity measures that can be computed starting from the one-mode projection $\mathbf{G}_l$ (see e.g. Newman, 2018, Sec. 7.6). Indeed, the Jaccard coefficient is formally given by

$$s_{ij}^{(l)} := J_{ij}^{(l)} = \frac{g_{ij}^{(l)}}{g_{ii}^{(l)} + g_{jj}^{(l)} - g_{ij}^{(l)}}, \tag{3}$$

while the cosine similarity turns out to be

$$s_{ij}^{(l)} := C_{ij}^{(l)} = \frac{g_{ij}^{(l)}}{(g_{ii}^{(l)} g_{jj}^{(l)})^{1/2}} \tag{4}$$

(for more details, see F. Baccini, Barabesi, and Petrovich, 2023). Finally, by assuming (3) and (4), $s_{ij}^{(l)}$ is a normalized similarity measure in $[0, 1]$ and $\mathbf{S}_l$ is symmetric, since $\mathbf{G}_l$ is symmetric.

2.1 Similarity network fusion

Similarity Network Fusion (SNF) is an unsupervised technique first introduced by Wang, Mezlini, et al. (2014) with the purpose of integrating multilayered data. Specifically, given a multiplex similarity network, SNF iteratively updates the similarity matrix corresponding to each layer in order to obtain a unique similarity network, which is an "averaged" representation of the single layers. The technique has been successfully used in many different frameworks. For example, in genetics and epigenetics, Wang, Mezlini, et al. (2014) and F. Baccini, Bianchini, and Geraci (2022) finalize the methodology to the aggregation of different types of genome data; F. Baccini, Barabesi, A. Baccini, et al. (2022) consider instead an application to scientometrics by integrating a multiplex similarity network built on different scholar interactions between scientific journals.

SNF is built upon an iterative procedure named Cross Diffusion Process (CDP) suggested by Wang, Jiang, et al. (2012). CDP aims at enhancing two types of edges in the desired monoplex, i.e. strong links (with a large weight compared to the other links of the network) which are present in some layers, and links which are shared by all the layers. Let us assume that $\mathbf{P}_l = (p_{ij}^{(l)})$ denotes the normalized matrix obtained by $\mathbf{S}_l$ as follows

$$p_{ij}^{(l)} = \frac{s_{ij}^{(l)}}{\sum_{k=1}^{n}\sum_{h=1}^{n} s_{kh}^{(l)}}.$$

In addition, for the l-th layer, a further "locally" normalized matrix is introduced, namely $\mathbf{Q}_l = (q_{ij}^{(l)})$, where

$$q_{ij}^{(l)} = \begin{cases} \frac{s_{ij}^{(l)}}{\sum_{h\in N_i} s_{ih}^{(l)}} & j \in N_i \\ 0 & \text{otherwise,} \end{cases}$$

and N_i represents the index set of the k-th nearest neighbours for the i-th node, computed via the k-Nearest Neighbours (k-NN) algorithm. We recall that the parameter k must be specified in advance, and it is fixed throughout the CDP procedure. For more details about the k-NN algorithm, see e.g. Hastie et al. (2009). The introduction of $\mathbf{Q}_l$ aims at eliminating the mutual influence between distant nodes in the l-th layer. Indeed, the edge weight between non-neighbouring nodes in the l-th layer is null in $\mathbf{Q}_l$.

The matrices $\mathbf{P}_1, \dots, \mathbf{P}_m$ are called "initial status" matrices, since they constitute the first step of the iterative procedure CDP. Let $\mathbf{P}_{l,t}$ denote the l-th matrix computed at the t-th iteration of the CDP. For the l-th layer, the updating rule is given by

$$\mathbf{P}_{l,t+1} = \mathbf{Q}_l \left(\frac{1}{m-1} \sum_{h=1,h\neq l}^{m} \mathbf{P}_{h,t} \right) \mathbf{Q}_l^{\mathrm{T}},$$

for $t = 0, 1, \dots,$ and where $\mathbf{P}_{l,0} = \mathbf{P}_l$ for $l = 1, \dots, m$.

Wang, Jiang, et al. (2012) show that CDP converges, in the sense that $\|\mathbf{P}_{l,t} - \mathbf{P}_{l,\infty}\|_{\mathrm{F}} \to 0$

as $t \to \infty$, where $\mathbf{P}_{l,\infty}$ is a suitable limit matrix, while $\|\mathbf{X}\|_{\mathrm{F}} = \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{1/2}$ represents the usual Frobenius norm of a matrix $\mathbf{X}$. Moreover, let us assume that there exists $t = T$ for which convergence is reached according to the accuracy criterion $\|\mathbf{P}_{l,T+1} - \mathbf{P}_{l,T}\|_{\mathrm{F}} < \varepsilon$ for each $l = 1, \ldots, m$ and for a pre-fixed $\varepsilon > 0$. Thus, the last step of SNF consists in averaging the matrices $\mathbf{P}_{1,T}, \ldots, \mathbf{P}_{m,T}$, i.e.

$$\mathbf{P}_{+T} = \frac{1}{m}\sum_{l=1}^{m} \mathbf{P}_{l,T}. \tag{5}$$

Finally, the matrix (5) is taken as the adjacency matrix of the desired monoplex. A similarity matrix, say $\mathbf{S}_{\mathrm{SNF}}$, is then suitably obtained by re-weighting $\mathbf{P}_{+T}$.

Besides the enforcement of strong links, SNF also boosts weaker links that are shared by many layers, in the sense that densely connected groups of nodes are maintained and strengthened with tighter links in the monoplex, since they potentially represent strong relationships among groups of elements. However, some drawbacks in the use of SNF should be highlighted. A first observation is that the choice of the parameter k in the k-NN algorithm is rather arbitrary, and different values of k might result in significantly different monoplex structures. Second, each layer equally contributes to the structure of the monoplex defined by (5), i.e. SNF implicitly assumes that all the layers have the same relevance in determining the overall structure of the relations.

2.2 Similarity matrix average

In the present section, we analyse the Similarity Matrix Average (SMA) method suggested by F. Baccini, Barabesi, and Petrovich (2023) for computing a monoplex network starting from a multiplex similarity network. SMA relies on some mathematical properties of similarity matrices, and exploits some recent advances in the theory of barycenters of objects lying in abstract spaces, with the ultimate goal of computing a matrix average (Álvarez-Esteban et al., 2016; Bhatia and Congedo, 2019).

Although results concerning matrix barycenters are usually considered in the space of positive definite matrices (Bhatia and Congedo, 2019), similarity matrices are more naturally handled in the space of completely positive matrices. Formally, the space $\mathcal{CP}_n$ of completely positive matrices is defined as

$$\mathcal{CP}_n = \{\mathbf{X} \in \mathbb{R}^{n\times n} : \mathbf{X} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y}, \exists \mathbf{Y} \in \mathbb{R}^{p\times n}, \mathbf{Y} \geq \mathbf{0}\},$$

where the notation $\mathbf{X} \geq \mathbf{0}$ indicates that the matrix $\mathbf{X}$ has nonnegative elements. For details on the properties of this class of matrices, we refer the reader to the monographs by Berman and Shaked-Monderer (2003) and Shaked-Monderer and Berman (2021). If $\mathcal{P}_n$ represents the space of symmetric positive semidefinite matrices, a necessary condition for a matrix $\mathbf{X}$ to be completely positive is that $\mathbf{X} \in \mathcal{P}_n$ and $\mathbf{X} \geq \mathbf{0}$, even if – contrary to intuition – this condition is not generally sufficient. Hence, $\mathcal{CP}_n$ is a proper subset of $\mathcal{P}_n$ and, in addition, $\mathcal{CP}_n$ is a closed convex cone in $\mathcal{P}_n$ (Berman and Shaked-Monderer, 2003).

F. Baccini, Barabesi, and Petrovich (2023) show that several commonly-adopted similarity matrices are completely positive. Noticeably, the similarity matrices obtained by using the Jaccard index (3) and the cosine similarity (4) belong to the space $\mathcal{CP}_n$. The fact that many similarity matrices belong to $\mathcal{CP}_n$ is of central importance for SMA. Indeed, in SMA the problem of obtaining a monoplex starting from a multiplex similarity network is formulated as a minimization problem carried out in the space of completely positive matrices. Formally, the adjacency matrix of the monoplex can be computed as the barycenter $\mathbf{S}_+$ of the m similarity matrices $\mathbf{S}_1, \ldots, \mathbf{S}_m$ representing the layers, by solving the following minimization problem

$$\mathbf{S}_+ = \arg \min_{\mathbf{X} \in \mathcal{CP}_n} \sum_{l=1}^{m} w_l d(\mathbf{X}, \mathbf{S}_l)^2, \tag{6}$$

where $d : \mathcal{CP}_n \times \mathcal{CP}_n \to \mathbb{R}^+$ is a metric on $\mathcal{CP}_n$, and the weights $\mathbf{w} = (w_1, \ldots, w_m)^{\mathrm{T}}$ are such that $w_l \geq 0$ and $\sum_{l=1}^{m} w_l = 1$. In a general framework, $\mathbf{S}_+$ provides the so-called Fréchet mean (see e.g. Bacák, 2014). The Fréchet mean might not be suitable in an arbitrary metric space, though being highly appropriate in a geodesic metric space of non-positive curvature, i.e. a Hadamard space such as $\mathcal{CP}_n$. The existence and uniqueness of the minimizer $\mathbf{S}_+$ in a Hadamard space are assured by Theorem 2.4 by Bacák (2014). Obviously, the optimization problem in (6) also depends on the choice of the metric d and of the weights $\mathbf{w}$.

Concerning the metric choice, the cases of the Frobenius, Riemannian and Wasserstein metrics are discussed in F. Baccini, Barabesi, A. Baccini, et al. (2022). If the classical Frobenius metric is adopted in (6), i.e.

$$d_{\mathrm{F}}(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}} = \mathrm{tr}((\mathbf{X} - \mathbf{Y})^{\mathrm{T}}(\mathbf{X} - \mathbf{Y}))^{1/2},$$

we obtain the weighted matrix average suggested by Abdi et al. (2005), i.e.

$$\mathbf{S}_{+\mathrm{F}} = \sum_{l=1}^{m} w_l \mathbf{S}_l. \tag{7}$$

F. Baccini, Barabesi, and Petrovich (2023) emphasize that $\mathbf{S}_{+\mathrm{F}} \in \mathcal{CP}_n$, since (7) is a weighted sum of completely positive matrices with positive weights. Therefore, such a procedure is basic and intuitive, as well as the simplest to implement in practice. However, this proposal may involve a “swelling effect” as a drawback, meaning that $\mathbf{S}_{+\mathrm{F}}$ may show an increase in the determinant with respect to the single addends of the matrix average (7). As a result, the final monoplex might not be highly representative of the information contained in the multiplex.

An alternative choice for d in (6) is the Riemannian metric, i.e.

$$\begin{aligned} d_{\mathrm{R}}(\mathbf{X}, \mathbf{Y}) &= \|\log(\mathbf{X}^{-1/2}\mathbf{Y}\mathbf{X}^{-1/2})\|_{\mathrm{F}} \\ &= \mathrm{tr}(\log(\mathbf{X}^{-1/2}\mathbf{Y}\mathbf{X}^{-1/2})^{\mathrm{T}} \log(\mathbf{X}^{-1/2}\mathbf{Y}\mathbf{X}^{-1/2}))^{1/2}. \end{aligned}$$

Bhatia (2009) (Chapter 6, Theorem 6.1.6) remarks that d_{R} naturally arises in the framework of Riemannian geometry (for further details, see also Bhatia and Congedo, 2019, and references therein). Moreover, d_{R} may be considered as the matrix version of the Fisher-Rao metric for probability laws (see in turn Chapter 6 in Bhatia, 2009). For this metric choice, the matrix given by (6), say $\mathbf{S}_{+\mathrm{R}}$, exists and is the unique solution of the nonlinear matrix equation in $\mathbf{X}$ given by

$$\sum_{l=1}^{m} w_l \log(\mathbf{X}^{1/2}\mathbf{S}_l^{-1}\mathbf{X}^{1/2}) = \mathbf{0}, \tag{8}$$

even if it is expressible in closed form solely for $m = 2$ (see e.g. Lim and Pálfia, 2014a,b). For $m > 2$, Lim and Pálfia (2014a) introduce an iterative procedure which produces the solution $\mathbf{S}_{+\mathrm{R}}$ of (8). In addition, $\mathbf{S}_{+\mathrm{R}}$ may be less prone than $\mathbf{S}_{+\mathrm{F}}$ to the swelling effect (Lim and Pálfia, 2014a,b).

A final proposal for d in (6) is given by the Wasserstein metric, i.e.

$$d_{\mathrm{W}}(\mathbf{X},\mathbf{Y}) = \mathrm{tr}(\mathbf{X} + \mathbf{Y} - 2(\mathbf{X}^{1/2}\mathbf{Y}\mathbf{X}^{1/2})^{1/2})^{1/2},$$

which is of great importance in the theory of optimal transport and quantum information (for more details, see e.g. Bhatia, Jain, and Lim, 2019). Bhatia, Jain, and Lim (2019) emphasize that d_{W} displays many interesting features and, among others, it corresponds to a metric in Riemannian geometry. For this metric selection, Álvarez-Esteban et al. (2016) show that the matrix given by (6), say $\mathbf{S}_{+\mathrm{W}}$, is the unique solution of the following nonlinear matrix equation in $\mathbf{X}$

$$\mathbf{X} = \sum_{l=1}^{m} w_l (\mathbf{X}^{1/2}\mathbf{S}_l\mathbf{X}^{1/2})^{1/2}. \tag{9}$$

F. Baccini, Barabesi, and Petrovich (2023) also remark that $\mathbf{S}_{+\mathrm{W}} \in \mathcal{CP}_n$ (see also Bhatia, Jain, and Lim, 2019). The solution of the nonlinear matrix equation (9) is known in a closed form only in the case $m = 2$ (see e.g. F. Baccini, Barabesi, and Petrovich, 2023). For $m > 2$, Álvarez-Esteban et al. (2016) introduce a fixed-point algorithm which generates as a limiting matrix the solution $\mathbf{S}_{+\mathrm{W}}$ of (9). It is worth noticing that $(\mathbf{S}_{+\mathrm{F}} - \mathbf{S}_{+\mathrm{W}})$ is positive semidefinite based on Theorem 9 by Bhatia (2009). This property implies that the swelling effect is mitigated, unlike in the case of the Frobenius metric.

In order to introduce a possible choice of the elements of the weight vector $\mathbf{w}$, a measure of "closeness" between each pair of the m similarity matrices $\mathbf{S}_1, \ldots, \mathbf{S}_m$ is needed. A suitable measure is given by the RV coefficient proposed by Robert and Escoufier (1976). The matrix $\mathbf{R} = (r_{ij})$ of order $(m \times m)$ is then considered, where r_{ij} represents the RV coefficient between $\mathbf{S}_i$ and $\mathbf{S}_j$, i.e.

$$r_{ij} = \frac{\langle \mathbf{S}_i, \mathbf{S}_j \rangle_{\mathrm{F}}}{\|\mathbf{S}_i\|_{\mathrm{F}} \|\mathbf{S}_j\|_{\mathrm{F}}},$$

where $\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathrm{F}} = \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{Y})$ represents the Frobenius inner product for two matrices $\mathbf{X}$ and $\mathbf{Y}$ of the same order. It holds that $r_{ij} \in [0, 1]$, and the closeness between $\mathbf{S}_i$ and $\mathbf{S}_j$ increases as r_{ij} approaches one.

When $\mathbf{S}_{+\mathrm{F}}$ is adopted, Abdi et al. (2005) suggest to consider the eigendecomposition of $\mathbf{R}$, i.e. $\mathbf{R} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\mathrm{T}}$, where $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_m)$ is the orthogonal matrix whose columns are the eigenvectors of $\mathbf{R}$, and $\boldsymbol{\Lambda}$ is the diagonal matrix whose entries are the positive eigenvalues of $\mathbf{R}$ considered in nonincreasing order. Abdi et al. (2005) propose the choice

$$\mathbf{w}_{\mathrm{F}} = \frac{1}{\mathbf{1}^{\mathrm{T}}\mathbf{q}_1}\,\mathbf{q}_1. \tag{10}$$

In practice, a principal component analysis is considered on $\mathbf{R}$ and the first eigenvector is used for implementing $\mathbf{w}$. This choice reflects the idea that layers with larger projections on $\mathbf{q}_1$ are more similar to the other layers than the layers with smaller projections. Therefore, the elements of this eigenvector should provide suitable weights for $\mathbf{S}_{+\mathrm{F}}$, which is a linear combination of $\mathbf{S}_1, \ldots, \mathbf{S}_m$.

In the cases of $\mathbf{S}_{+\mathrm{R}}$ and $\mathbf{S}_{+\mathrm{W}}$, it is not obvious if the previous choice of $\mathbf{w}$ is suitable, since these averages are not linear functions of $\mathbf{S}_1, \ldots, \mathbf{S}_m$. Alternatively, F. Baccini, Barabesi, and Petrovich (2023) suggest the choice

$$\mathbf{w}_{\mathrm{R}} = \mathbf{w}_{\mathrm{W}} = \frac{1}{\mathbf{1}^{\mathrm{T}}(\mathbf{R} - \mathbf{I}_m)\mathbf{1}}\,(\mathbf{R} - \mathbf{I}_m)\mathbf{1}. \tag{11}$$

The rationale behind the weights in Equations 10 and 11 is that the more a similarity matrix is correlated with the other layers, the more it is representative of the whole set of similarity matrices – and hence it should receive a larger weight with respect to the others. As a result, groups of layers that are assigned similar weights collect similarity matrices that are likely to be structurally similar.

We conclude by remarking that the matrices $\mathbf{S}_{+\mathrm{F}}$, $\mathbf{S}_{+\mathrm{R}}$ and $\mathbf{S}_{+\mathrm{W}}$ represent the adjacency matrices of the desired monoplex in the practical implementation of the method. We also remark that SMA could be generally adopted with similarity matrices defined in $\mathcal{P}_n$, since the above results also hold in this space (F. Baccini, Barabesi, and Petrovich, 2023). However, even if the method could be adopted for similarity matrices belonging to different spaces, the described properties of SMA are not necessarily guaranteed.

3 Application to the analysis of glacier data

In cryoconite holes, bacteria form complex ecological networks (Gokul et al., 2016) whereby different populations interact, sometimes using peculiar metabolic ways (Franzetti, Tagliaferri, et al., 2016; Pittino, Zawierucha, et al., 2023). However, it is still unknown whether the same ecological relationships between the bacterial populations of cryoconite holes occur worldwide, or whether a biogeography of glacier bacterial communities exist (Darcy et al., 2011; Franzetti, Tatangelo, et al., 2013; Pittino, Ambrosini, et al., 2023).

In the following, we performed the similarity network aggregation analysis described in Section 2 on a set of ecological networks constructed from the relative abundances of the bacterial populations sampled on nine glaciers.

3.1 Dataset description and processing

The dataset consists of the relative abundances obtained from 16S rDNA metatranscriptomics of the bacterial communities sampled between 2013 and 2018 on 4 glaciers in the Alps (Forni, Cedec, and East Zebrù in the Ortles-Cevedale group in Italy, Morteratsch in the Bernina group in Switzerland) and five glaciers in the Andes (Iver in the Plomo Group, El Morado in the Maipo Valley, Exploradores in the San Valentin group in Chile, Perito Moreno in the Southern Patagonian Ice Field in Argentina). Details on glaciers, sampling methods, laboratory and bioinformatics analyses are provided by Pittino, Maglio, et al. (2018) and Pittino, Zordan, et al. (2021). Briefly, 16S rDNA was extracted from cryoconite collected in 140 holes (11-23 holes per glacier), Illumina sequenced and clustered in Amplicon Sequences Variants (ASVs) with DADA2 (Callahan et al., 2016). ASVs, which can roughly be considered to correspond to bacterial strains, were then taxonomically classified using the SILVA classifier (Quast et al., 2012). Overall, the dataset consists of 135 ASVs, where the number of cryoconite holes considered varies between glaciers. Table 1 reports the exact number of sites (cryoconite holes) considered for each glacier.

Table 1: Number of sites where the measurement of relative abundance of the considered ASVs was performed.

Glacier	# Cryoconite holes
Cedec	11
Exploradores	15
Forni	20
Iver	15
Iver Est	15
El Morado	15
Morteratsch	23
Perito Moreno	15
Zebrú	11

For the present analysis, the ASVs that are not present on any glacier are filtered out. This excludes 9 of the 135 ASVs available. In view of computing similarities between pairs of ASVs for each glacier, the ASVs that are totally absent in at least one glacier are filtered out. Indeed, the complete absence of an ASV is represented as a null vector of dimension equal to the number of different sites of the glacier. This constitutes an issue when computing the similarity between such ASV and each of the remaining ones, since a null vector has similarity equal to 0 with all the other vectors, and produces a similarity matrix which is not completely positive. As a consequence, the integration methods described in Section 2 could not be applied. By excluding the ASVs that are absent from at least one glacier, the number of ASVs drastically reduces to 16, represented by $1,626,370$ sequences. The taxonomy of the considered ASVs is reported in Table 2. We highlight that this filtering is acceptable if we focus the attention on the most common ASVs that can be found on all the analysed glaciers. A more detailed analysis of the excluded ASVs goes beyond the scope of the present work, and is left for future investigations.

Overall, the filtering process yields 9 relative abundance matrices, one for each glacier, where one dimension is fixed and equal to 16 (the number of ASVs) and the other varies according to the number of cryoconite holes sampled on each glacier (see Table 1).

3.2 Integration of the multiplex similarity network of ASVs

In order to assess the existence of differences between glaciers in terms of the relative abundance of ASVs, a multiplex similarity network of ASVs is computed. In particular, the aim is to select the ASVs that are present in every glacier, and then compute the similarities between pairs of ASVs based on their relative abundance within every glacier. The result is a multiplex similarity network where each layer represents a glacier, and the nodes are the considered ASVs. All the ASVs included in the analysis belong to orders and genera typical of cryoconite holes worldwide and, in particular, of those situated on mountain glaciers (Fillinger et al., 2021). Formally, the 9 matrices can be interpreted as a multilayer network formed by 9 bipartite networks with the structure discussed at the beginning of Section 2. Indeed, each matrix is the adjacency matrix of a bipartite network, where the nodes in one set of the bipartition represent the ASVs, and the other set varies between layers and represents the set of sites. We remark that the set of nodes representing the ASVs is the same on each layer. Moreover, a node corresponding to an ASV is connected to all the nodes corresponding to sites where the ASV is present. These links are weighted with the relative abundance value of that ASV in the connected site. Subsequently, the similarity between pairs of ASVs is computed as in (1), and a multiplex similarity network is obtained.

In order to analyse the overall organization of the considered ASVs, both SNF and SMA are applied to the multiplex similarity network of ASVs. To perform the integration we used the R implementation of SNF, available at `http://compbio.cs.toronto.edu/SNF/SNF/Software.html`, and the Python implementation of SMA available at `https://github.com/DedeBac/SimilarityMatrixAggregation`. Four different monoplex networks are obtained: one resulting from SNF, i.e. $\mathbf{S}_{\mathrm{SNF}}$, and one for each of the three different metric choices available in SMA and discussed in Section 2.2, i.e $\mathbf{S}_{+\mathrm{F}}$, $\mathbf{S}_{+\mathrm{R}}$ and $\mathbf{S}_{+\mathrm{W}}$. For each resulting network, the Louvain algorithm for modularity optimization (Blondel et al., 2008) is then applied to individuate strongly connected groups of nodes. We recall that modularity is a scalar value that measures the density of links within a community of nodes. In particular, the strength of a network partition is measured by comparing the density of the links falling inside a community with links between different communities. The Louvain algorithm efficiently implements the search for an optimal quality partition based on modularity. The networks are depicted in Figures 1-4. In the figures, the nodes are coloured according to the cluster they have been assigned to. The cluster and

Table 2: Taxonomy of the 16 ASVs included in the analysis. The ASV number is assigned in the data collection phase and it is unique for each sample.

ASV	Domain	Phylum	Class	Order	Family	Genus
ASV_166579	Bacteria	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Chitinophagaceae	*Ferruginibacter*
ASV_179530	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	*Rhizobacter*
ASV_184088	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Acidovorax*
ASV_184128	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Variovorax*
ASV_184156	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Unclassified_Comamonadaceae*
ASV_185564	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Polaromonas*
ASV_185576	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Polaromonas*
ASV_185593	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	*Polaromonas*
ASV_190905	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	*Unclassified_Oxalobacteraceae*
ASV_218814	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	*Bradyrhizobium*
ASV_310446	Bacteria	Bacteroidetes	Cytophagia	Cytophagales	Cytophagaceae	*Hymenobacter*
ASV_310875	Bacteria	Bacteroidetes	Cytophagia	Cytophagales	Cytophagaceae	*Hymenobacter*
ASV_439437	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	*Mesorhizobium*
ASV_489786	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Microbacteriaceae	*Cryobacterium*
ASV_88033	Bacteria	Cyanobacteria/Chloroplast	Cyanobacteria	Unclassified_Cyanobacteria	Unclassified_Cyanobacteria	*Unclassified_Cyanobacteria*
ASV_92579	Bacteria	Deinococcus-Thermus	Deinococci	Deinococcales	Deinococcaceae	*Deinococcus*

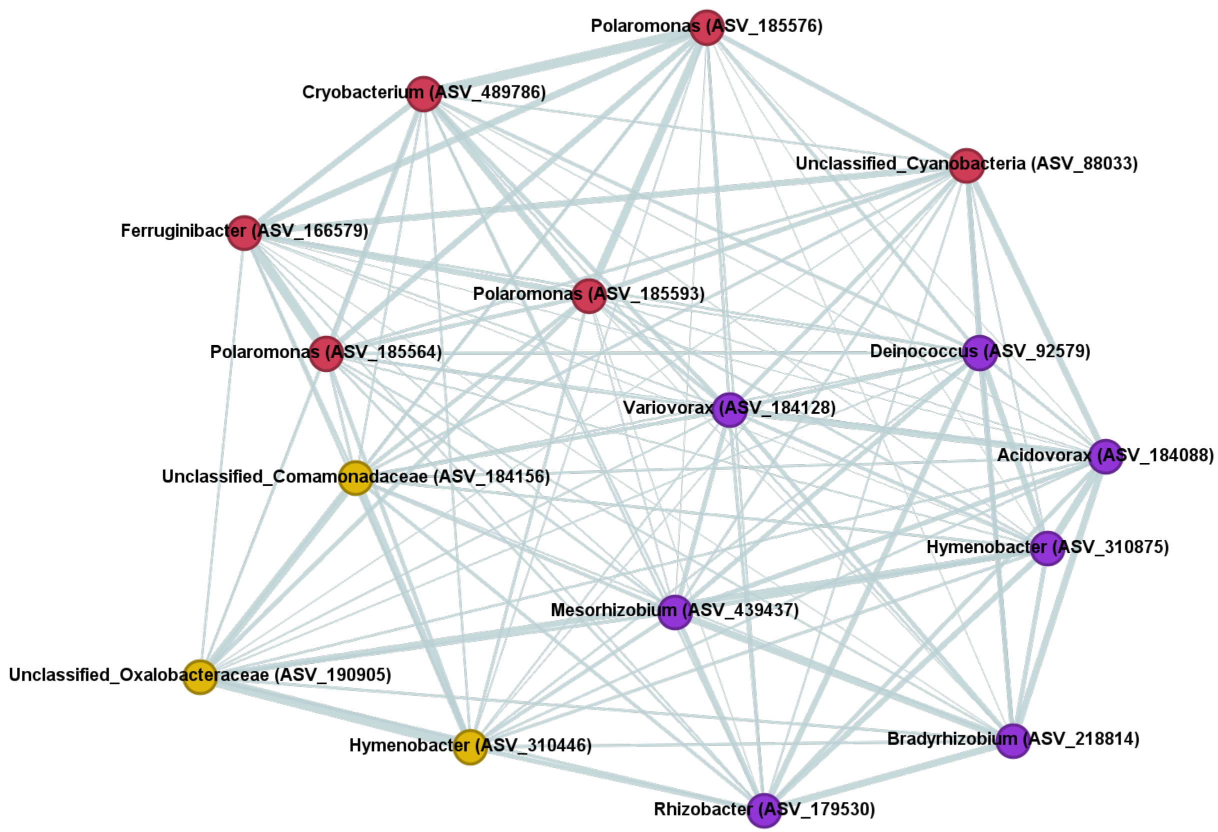

Figure 1: Monoplex obtained with SNF ($\mathbf{S}_{\mathrm{SNF}}$). Different colours correspond to distinct communities individuated by the Louvain algorithm for modularity optimization. Node labels report the genus and the identification number of the corresponding ASVs.

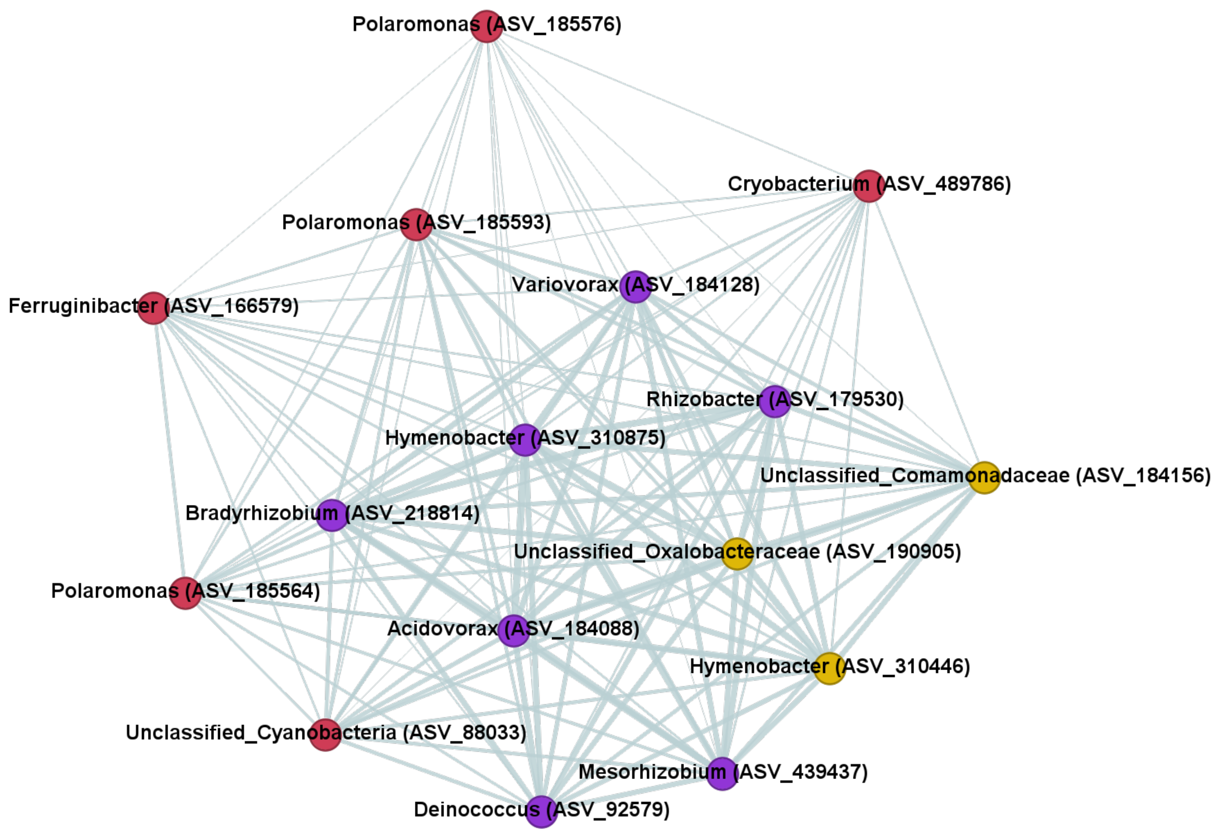

Figure 2: Monoplex obtained with SMA and the Frobenius metric ($\mathbf{S}_{+\mathrm{F}}$). Different colours correspond to distinct communities individuated by the Louvain algorithm for modularity optimization. Node labels report the genus and the identification number of the corresponding ASVs.

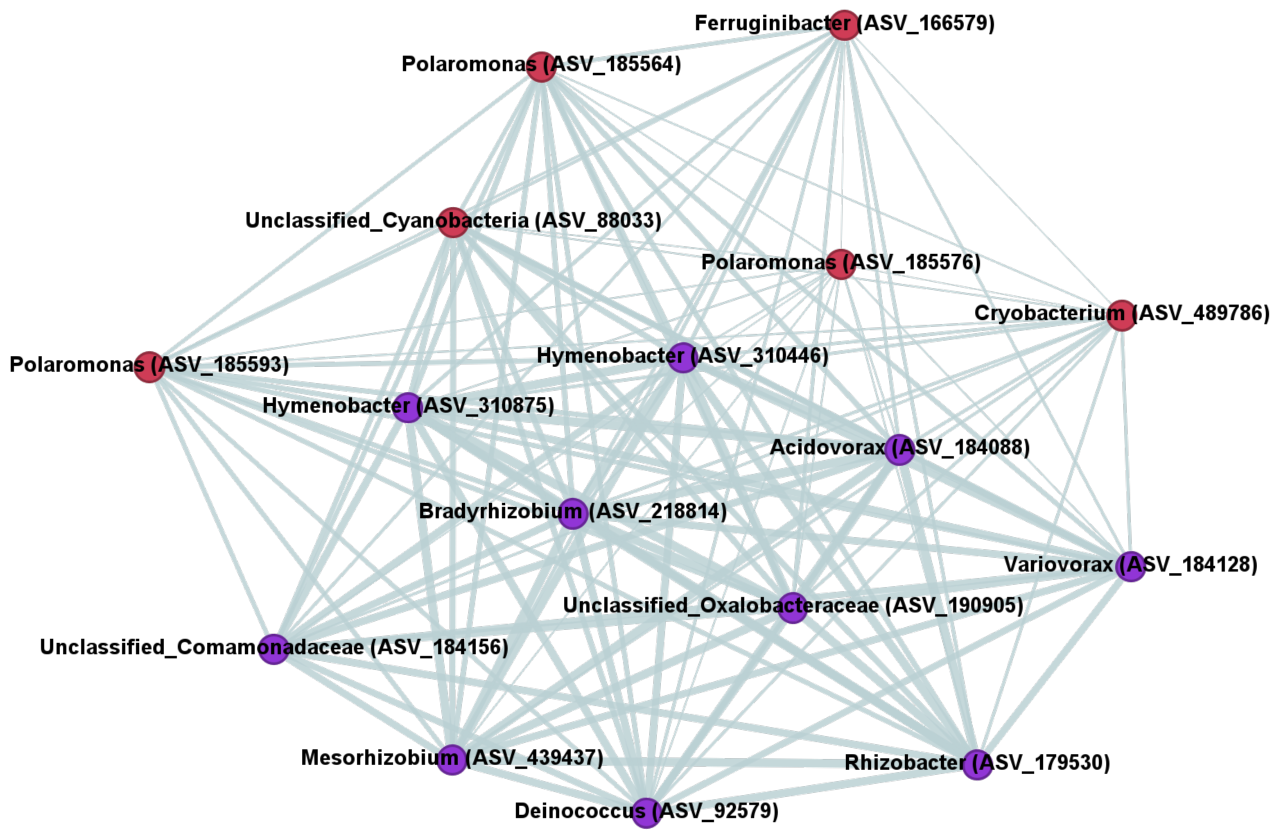

Figure 3: Monoplex obtained with SMA and the Riemannian metric ($\mathbf{S}_{+\mathrm{R}}$). Different colours correspond to distinct communities individuated by the Louvain algorithm for modularity optimization. Node labels report the genus and the identification number of the corresponding ASVs.

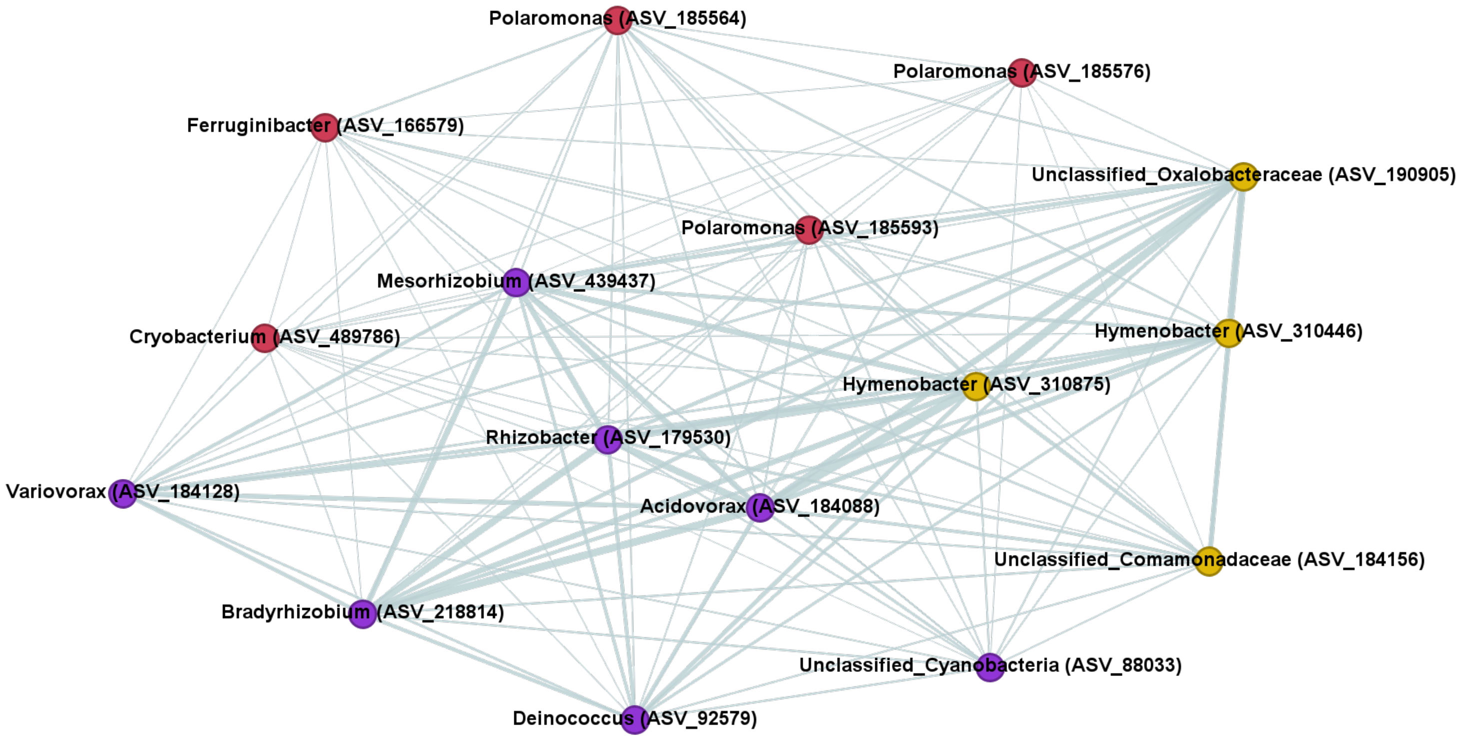

Figure 4: Monoplex obtained with SMA and the Wasserstein metric ($\mathbf{S}_{+\mathrm{W}}$). Different colours correspond to distinct communities individuated by the Louvain algorithm for modularity optimization. Node labels report the genus and the identification number of the corresponding ASVs.

visual analysis was performed using the Gephi software (Bastian, Heymann, and Jacomy, 2009) and the visualization algorithm ForceAtlas2 (Jacomy et al., 2014).

All the ASVs included in the analysis belong to orders and genera typical of cryoconite holes worldwide, particularly on mountain glaciers (Pittino, Ambrosini, et al., 2023). Except for $\mathbf{S}_{+\mathrm{R}}$, for which only two clusters are individuated, the Louvain algorithm partitions the networks into three clusters. In all the networks, the obtained modules consistently identified a group (red nodes) of ASVs composed by the genera *Polaromonas* (3 ASVs), *Cryobacterium*, *Ferruginibacter* and *Cyanobacteria* (in $\mathbf{S}_{+\mathrm{W}}$ *Cyanobacteria* are instead assigned to the purple cluster). Cyanobacteria usually dominate cryoconite hole bacterial communities, where they are typically primary producers and ecosystem engineers (Rozwalak et al., 2022). The other taxa belonging to the red cluster show a broad variety of metabolisms, which allow them to survive on a broad variety of energy sources in the harsh supraglacial environment (Franzetti, Tatangelo, et al., 2013). Overall, the red cluster is composed of bacterial strains that are abundant and typical of cryoconite holes worldwide (Boetius et al., 2015; Fillinger et al., 2021; Franzetti, Tatangelo, et al., 2013).

In the network obtained with SNF, in $\mathbf{S}_{+\mathrm{F}}$ and in $\mathbf{S}_{+\mathrm{W}}$ (Figures 1, 2 and 4) the Louvain algorithm identifies a further module (yellow nodes) composed of one ASV classified as *Himenobacter*, and two ASVs classified only at the family level as *Comamonadaceae* and *Oxalobacteraceae* (see Table 2). In $\mathbf{S}_{+\mathrm{W}}$ a second Himenobacter ASV is also added to the yellow cluster. However, the lack of classification at the genus level and the inconsistency in the classification of the Himenobacter ASVs prevent from further interpretations of the ecological role of the yellow cluster in the cryoconite hole communities.

3.3 Results and discussion

The first issue to be investigated is the coherence between the various monoplex networks obtained with different techniques. To this aim, a measure of correlation between the similarity matrices (i.e. the similarity networks) is adopted. An appropriate correlation measure is the so-called "generalized distance correlation" introduced by Székely, Rizzo, and Bakirov (2007). In the present setting, the generalized distance correlation can be interpreted as an extension to similarity matrices of the Pearson correlation coefficient, since it assumes values in the interval $[0, 1]$, with values close to 0 indicating no or very weak association, and values close to 1 suggesting a stronger association. Table 3 displays the values of generalized distance correlation between pairs of monoplex networks obtained with different methods. It is apparent that all the networks are strongly associated, thus indicating a significant structural similarity between them. Therefore, this argument suggests a coherence between the different methods.

The second issue of interest is the relative impact of the single layers (i.e. the glaciers) in determining the final monoplex. As highlighted in Section 2.2, unlike SNF, SMA automatically assigns weights to the contribution of each layer to the final network by using the RV-coefficients matrix. The resulting weights may be interpreted as the relative importance of the single layers

Table 3: Values of the generalized distance correlation between monoplex networks obtained with different methods.

	$\mathbf{S}_{\text{SNF}}$	$\mathbf{S}_{+\text{F}}$	$\mathbf{S}_{+\text{R}}$	$\mathbf{S}_{+\text{W}}$
$\mathbf{S}_{\text{SNF}}$	1	0.82659	0.78627	0.99790
$\mathbf{S}_{+\text{F}}$	–	1	0.99000	0.83454
$\mathbf{S}_{+\text{R}}$	–	–	1	0.79449
$\mathbf{S}_{+\text{W}}$	–	–	–	1

for the determination of the final monoplex. Moreover, networks (in this case, glaciers) that are assigned similar weights are likely to have a similar structure. Table 4 reports the weights assigned to each layer (i.e. the glacier) in the SMA procedure, in case of the adoption of the different choices given by (10) and (11).

Table 4: Weights assigned to the similarity networks of the 9 different layers.

Glacier	$\mathbf{w}_{\text{F}}$	$\mathbf{w}_{\text{R}}$
Cedec	0.11333	0.11355
Exploradores	0.10664	0.10616
Forni	0.11026	0.11009
Iver	0.11003	0.10990
Iver Est	0.11418	0.11459
El Morado	0.11424	0.11463
Morteratsch	0.11043	0.11037
Perito Moreno	0.10787	0.10749
Zebrú	0.11301	0.11321

From Table 4 it can be observed that Exploradores and Perito Moreno are assigned very similar weights, both slightly lower than those assigned to the other glaciers. Interestingly, Exploradores and Perito Moreno are large glaciers (> 85 km^2) reaching low altitudes (the sampling areas were ≤ 200 m a.s.l.), while the other glaciers (both in the Andes and the Alps) are medium to small (≤ 15 km^2) mountain glaciers at ≥ 2500 m a.s.l. Thus, Exploradores and Perito Moreno can show very different ecological conditions and communities with different structures with respect to the other glaciers, which are reflected by their similar weights.

As previously observed, SNF does not assign specific weights to the contribution of the single layers. However, it is possible, a posteriori, to adopt suitable network correlation measures to evaluate the structural similarity between the final monoplex and the single layers. An appropriate choice of such measure is the generalized distance correlation previously discussed. Table 5 reports the values of distance correlation between the monoplex obtained with SNF and the single similarity networks corresponding to different glaciers. Coherently with the weights assigned by the SMA algorithm, Exploradores and Perito Moreno show a weaker association with the final monoplex, when compared with the other glaciers.

Table 5: Values of the generalized distance correlation between the monoplex obtained with SNF and the single layers.

Glacier	
Cedec	0.74192
Exploradores	0.67823
Forni	0.84335
Iver	0.74968
Iver Est	0.71703
El Morado	0.82913
Morteratsch	0.87491
Perito Moreno	0.68240
Zebrú	0.70132

4 Conclusion

The present paper discusses the use of two network integration methods, namely Similarity Network Fusion and Similarity Matrix Average, to combine multi-source information coming from ecological networks into a single network. The two methods are valid alternatives to integrate a multiplex similarity network, and are shown to give coherent results. A major difference between SNF and SMA is that the second method includes an automatic way of assigning weights to the contribution of the layers based on the structure of the multiplex. In contrast, when SNF is used, the impact of the layers in the final monoplex can only be evaluated a posteriori by computing, for instance, the generalized distance correlation. The effectiveness of the two methods in the analysis of ecological networks is tested on a real-world dataset recording the relative abundance of microbial species in the ecosystems of nine glaciers. Despite being based on a very small subset of all the ASVs observed on the considered glaciers (19 out of 126), the investigation was able to identify an ecological network between cosmopolitan glacier bacteria strains, which may form the "basic" ecological network of glacier ecosystems. Further investigation on a larger set of glaciers will be certainly necessary to assess the generality of this ecological network. Nevertheless, the present analysis can provide glacier ecologists with a background hypothesis for future investigations. Moreover, the weights assigned to the similarity networks and the generalized distance correlations between them consistently suggest that the two large Patagonian glaciers (Exploradores and Perito Moreno) host ecological networks different from those of the other smaller mountain glaciers. This pattern supports the hypothesis that the general features of a glacier, as its altitude, size, and whether or not its tongue crosses the local tree line, can affect their ecological networks (Pittino, Zordan, et al., 2021). Overall, this work opens to the use and development of similarity network integration techniques to combine different characteristics describing biological entities with the aim of discovering organization patterns in ecological communities.

Acknowledgements

The authors are grateful to Pablo Alviz-Gazitua, Roberto Sergio Azzoni, Biagio Di Mauro, Guglielmina Diolaiuti, and Michael Seeger for their help during field work. Andrea Franzetti and Francesca Pittino kindly provided sequencing data. We thank the administration of Parco Nazionale dello Stelvio for the permission of collecting samples on Cedec, Forni and Zebrú glaciers through their collaboration agreement with the University of Milan; the CONAF (Corporación Nacional Forestal of Chile) for the permission and support in collecting supraglacial samples on the Exploradores Glacier, and the Administración de Parques Nacionales of Argentina (APN) for the permission to collect samples on the Perito Moreno Glacier with project N. 110-CPA-2016. La Venta team provided logistic support during the expeditions.

Bibliography

Abdi, Hervé et al. (2005). "DISTATIS: The analysis of multiple distance matrices". In: *Proceedings of the IEEE Computer Society: International Conference on Computer Vision and Pattern Recognition, San Diego, CA*, pp. 42–47.

Álvarez-Esteban, Pedro C et al. (2016). "A fixed-point approach to barycenters in Wasserstein space". In: *Journal of Mathematical Analysis and Applications* 441.2, pp. 744–762.

Ambrosini, Roberto et al. (2017). "Diversity and assembling processes of bacterial communities in cryoconite holes of a Karakoram glacier". In: *Microbial Ecology* 73, pp. 827–837.

Anesio, Alexandre M and Johanna Laybourn-Parry (2012). "Glaciers and ice sheets as a biome". In: *Trends in Ecology & Evolution* 27.4, pp. 219–225.

Bacák, Miroslav (2014). "Computing medians and means in Hadamard spaces". In: *SIAM Journal on Optimization* 24.3, pp. 1542–1566.

Baccini, Federica, Lucio Barabesi, Alberto Baccini, et al. (2022). "Similarity network fusion for scholarly journals". In: *Journal of Informetrics* 16.1, p. 101226.

Baccini, Federica, Lucio Barabesi, and Eugenio Petrovich (2023). "Similarity matrix average for aggregating multiplex networks". In: *Journal of Physics: Complexity.*

Baccini, Federica, Monica Bianchini, and Filippo Geraci (2022). "Graph-based integration of histone modification profiles". In: *Mathematics* 10.11, p. 1842.

Bar-Hen, Avner, Pierre Barbillon, and Sophie Donnet (2022). "Block models for generalized multipartite networks: applications in ecology and ethnobiology". In: *Statistical Modelling* 22.4, pp. 273–296.

Barabási, Albert-László (2013). “Network science”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371.1987, p. 20120375.

Barbillon, Pierre et al. (2017). “Stochastic block models for multiplex networks: an application to a multilevel network of researchers”. In: *Journal of the Royal Statistical Society Series A* 180.1, pp. 295–314.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy (2009). “Gephi: an open source software for exploring and manipulating networks”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 3. 1, pp. 361–362.

Berman, Abraham and Naomi Shaked-Monderer (2003). *Completely Positive Matrices.* World Scientific, Singapore.

Bhatia, Rajendra (2009). *Positive Definite Matrices.* Princeton University Press, Princeton, NJ.

Bhatia, Rajendra and Marco Congedo (2019). “Procrustes problems in Riemannian manifolds of positive definite matrices”. In: *Linear Algebra and its Applications* 563, pp. 440–445.

Bhatia, Rajendra, Tanvi Jain, and Yongdo Lim (2019). “On the Bures–Wasserstein distance between positive definite matrices”. In: *Expositiones Mathematicae* 37.2, pp. 165–191.

Bianconi, Ginestra (June 2018). *Multilayer Networks: Structure and Function.* Oxford University Press. Oxford, United Kingdom.

Blondel, Vincent D. et al. (2008). “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008.

Boetius, Antje et al. (2015). “Microbial ecology of the cryosphere: sea ice and glacial habitats”. In: *Nature Reviews Microbiology* 13.11, pp. 677–690.

Callahan, Benjamin J et al. (2016). “DADA2: High-resolution sample inference from Illumina amplicon data”. In: *Nature methods* 13.7, pp. 581–583.

Cook, Joseph et al. (2016). “Cryoconite: the dark biological secret of the cryosphere”. In: *Progress in Physical Geography* 40.1, pp. 66–111.

Darcy, John L et al. (2011). “Global distribution of Polaromonas phylotypes-evidence for a highly successful dispersal capacity”. In: *PLoS One* 6.8, e23742.

De Bello, Francesco et al. (2017). “Towards a more balanced combination of multiple traits when computing functional differences between species”. In: *Methods in Ecology and Evolution* 12, pp. 443–448.

De Domenico, Manlio (2022). *Multilayer Networks: Analysis and Visualization.* Springer Cham, Switzerland.

Eck, Nees Jan van and Ludo Waltman (2009). “How to normalize cooccurrence data? An analysis of some well-known similarity measures”. In: *Journal of the American Society for Information Science and Technology* 60.8, pp. 1635–1651.

Estrada, Ernesto (Oct. 2011). *The Structure of Complex Networks: Theory and Applications.* Oxford University Press, Oxford, United Kingdom.

Fillinger, Lucas et al. (2021). “Spatial and Annual Variation in Microbial Abundance, Community Composition, and Diversity Associated With Alpine Surface Snow”. In: *Frontiers in Microbiology* 12, p. 781904.

Franzetti, Andrea, Ilario Tagliaferri, et al. (2016). “Light-dependent microbial metabolisms drive carbon fluxes on glacier surfaces”. In: *The ISME Journal* 10.12, pp. 2984–2988.

Franzetti, Andrea, Valeria Tatangelo, et al. (2013). “Bacterial community structure on two alpine debris-covered glaciers and biogeography of Polaromonas phylotypes”. In: *The ISME journal* 7.8, pp. 1483–1492.

Gokul, Jarishma K et al. (2016). “Taxon interactions control the distributions of cryoconite bacteria colonizing a High Arctic ice cap”. In: *Molecular Ecology* 25.15, pp. 3752–3767.

Hastie, Trevor et al. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Vol. 2. Springer, New York, NY.

Hodson, Andy et al. (2008). “Glacial ecosystems”. In: *Ecological Monographs* 78.1, pp. 41–67.

Hutchinson, Matthew C. et al. (2019). “Seeing the forest for the trees: Putting multilayer networks to work for community ecology”. In: *Functional Ecology* 33.2, pp. 206–217.

Jacomy, Mathieu et al. (2014). “ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software”. In: *PLOS ONE* 9.6, e98679.

Kivelä, Mikko et al. (2014). “Multilayer networks”. In: *Journal of Complex Networks* 2.3, pp. 203–271.

Krivitsky, Pavel N., Laura M. Koehly, and Christopher Steven Marcum (2020). “Exponential-family random graph models for multi-layer networks”. In: *Psychometrika* 85.3, pp. 630–659.

Kuncheva, Zhana and Giovanni Montana (2015). “Community detection in multiplex network using locally adaptive random walks”. In: *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 1308–1315.

Lim, Yongdo and Miklós Pálfia (2014a). "Weighted deterministic walks for the least squares mean on Hadamard spaces". In: *Bulletin of the London Mathematical Society* 46.3, pp. 561–570.

— (2014b). "Weighted inductive means". In: *Linear Algebra and its Applications* 453, pp. 59–83.

MacDonald, Peter W., Elizaveta Levina, and Ji Zhu (2022). "Latent space models for multiplex networks with shared structure". In: *Biometrika* 109.3, pp. 683–706.

Mittelbach, Gary G and Brian J McGill (2019). *Community Ecology, 2nd ed.* Oxford University Press.

Newman, Mark (2018). *Networks.* Oxford University Press, Oxford, United Kingdom.

Pilosof, Shai et al. (2017). "The multilayer nature of ecological networks". In: *Nature Ecology & Evolution* 1.4, p. 0101.

Pittino, Francesca, Roberto Ambrosini, et al. (2023). "Geographical variability of bacterial communities of cryoconite holes of Andean glaciers". In: *Scientific Reports* 13.1, p. 2633.

Pittino, Francesca, Maurizio Maglio, et al. (2018). "Bacterial communities of cryoconite holes of a temperate alpine glacier show both seasonal trends and year-to-year variability". In: *Annals of Glaciology* 59.77, pp. 1–9. DOI: 10.1017/aog.2018.16.

Pittino, Francesca, Krzysztof Zawierucha, et al. (2023). "Functional and Taxonomic Diversity of Anaerobes in Supraglacial Microbial Communities". In: *Microbiology Spectrum* 11.2, e01004–22.

Pittino, Francesca, Simone Zordan, et al. (2021). "Diel transcriptional pattern contributes to functional and taxonomic diversity in supraglacial microbial communities". In: *bioRxiv*, pp. 2021–01.

Quast, Christian et al. (2012). "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools". In: *Nucleic Acids Research* 41.D1, pp. D590–D596.

Robert, Paul and Yves Escoufier (1976). "A unifying tool for linear multivariate statistical methods: the *RV*-coefficient". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 25.3, pp. 257–265.

Rozwalak, Piotr et al. (2022). "Cryoconite–From minerals and organic matter to bioengineered sediments on glacier's surfaces". In: *Science of The Total Environment* 807, p. 150874.

Shaked-Monderer, Naomi and Abraham Berman (2021). *Copositive and Completely Positive Matrices.* World Scientific, Singapore.

Székely, Gábor J., Maria L. Rizzo, and Nail K. Bakirov (2007). "Measuring and testing dependence by correlation of distances". In: *The Annals of Statistics* 35.6, pp. 2769–2794.

Wang, Bo, Jiayan Jiang, et al. (2012). "Unsupervised metric fusion by cross diffusion". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2997–3004.

Wang, Bo, Aziz M Mezlini, et al. (2014). "Similarity network fusion for aggregating data types on a genomic scale". In: *Nature Methods* 11.3, pp. 333–337.