Computational and Structural Biotechnology Journal 21 (2023) 4988–5002

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

Delving into gene-set multiplex networks facilitated by a k-nearest neighbor-based measure of similarity

Cheng Zheng [a], Man Wang [b], Ryo Yamada [a], Daigo Okada [a,*]

[a] Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, South Research Bldg. No.1(5F), 53 Shogoinkawahara-cho, Sakyo-ku, Kyoto, 6068507, Kyoto, Japan

[b] Department of Signal Transduction, Research Institute for Microbial Diseases, Osaka University, 3-1 Yamadaoka, Suita, 5650871, Osaka, Japan

ARTICLE INFO

Keywords:
Gene set
Multiplex network
k-nearest neighbor-based similarity
Multiplex clustering coefficient
Multiplex PageRank centrality
Gene set enrichment analysis
Gene set co-expression

ABSTRACT

Gene sets are functional units for living cells. Previously, limited studies investigated the complex relations among gene sets, but documents about their altering patterns across biological conditions still need to be prepared. In this study, we adopted and modified a classical k-nearest neighbor-based association function to detect inter-gene-set similarities. Based on this method, we built multiplex networks of gene sets for the first time; these networks contain layers of gene sets corresponding to different populations of cells. The context-based multiplex networks can capture meaningful biological variation and have considerable differences from knowledge-based networks of gene sets built on Jaccard similarity, as demonstrated in this study. Furthermore, at the scale of individual gene sets, the structural coefficients of gene sets (multiplex PageRank centrality, clustering coefficient, and participation coefficient) disclose the diversity of gene sets from the perspective of structural properties and make it easier to identify unique gene sets. In gene set enrichment analysis (GSEA), each gene set is treated independently, and its contextual and relational attributes are ignored. The structural coefficients of gene sets can supplement GSEA with information about the overall picture of gene sets, promoting the constructive reorganization of the enriched terms and helping researchers better prioritize and select gene sets.

1. Introduction

A gene set, by definition, is a set of gene symbols or other equivalent strings with a specific identifier that usually summarizes its essential information [1]. For example, a gene set called "GSE9006_HEALTHY_VS_TYPE_1_DIABETES_PBMC_AT_DX_DN" from the IMMUNESIGDB database includes 198 gene symbols (e.g., CYP4F3, BASP1, and VNN3). Its naming came from case-control microarray research that found these 198 genes down-regulated in peripheral blood mononuclear cells (PBMCs) of healthy subjects compared to newly diagnosed type I diabetes patients [2,3]. The genes in a gene set may have a mixture of discovery sources, such as biological discovery or computational inference [1], and creation of the set may rely on curators who browse online publications about genes and manually select and annotate genes to form a gene set based on their domain knowledge [4]. Nowadays, gene-set databases are expanding rapidly, and it becomes more and more important to understand the relationship among gene sets that could empower us to identify vital gene sets/pathways or mitigate their redundancy [5–7].

Many existing models measure the similarity of two gene sets either via memberships or annotation knowledge of their constituent genes (e.g., Jaccard coefficients [8] and kappa statistics [9]) or through gene expression profiles by computing a multivariate association (e.g., analysis of gene co-expression/correlation structures [10] and multivariate associations of gene expression vectors [11]). However, the knowledge-based similarity may differ substantially from the multivariate association inferred from the gene expression because the knowledge is static and ignores the biological context of studies and the complex interactions among genes. While multivariate associations are context-based methods [12], they usually have specific presumptions and significantly higher computational costs for large-scale experiments. Thus, a thorough investigation of multivariate association/similarity measures regarding their practical performance and capability of being applied to gene-set studies is warranted.

* Corresponding author.
E-mail addresses: zheng.cheng.68e@st.kyoto-u.ac.jp (C. Zheng), dokada@genome.med.kyoto-u.ac.jp (D. Okada).

https://doi.org/10.1016/j.csbj.2023.09.042
Received 18 April 2023; Received in revised form 22 September 2023; Accepted 28 September 2023
Available online 11 October 2023

In addition, we are interested in the intrinsic features of gene sets in their interacting communities under specific biological conditions. A multiplex network is a mathematical framework that aligns with our purpose [13,14]. Gene sets can be nodes in the layers of the multiplex network, and the weights of their incident edges can be scalar similarities defined as above. The structural coefficients of a multiplex network integrate the data from all the layers, offering clues to identify characteristic gene sets in the network [15–17]. Additionally, the findings of network properties of gene sets may help improve the interpretability of gene set enrichment analysis (GSEA) [8,18]. GSEA and its variants often emphasize the gene-level statistics or effect size/significance of enrichment [19,20]. The contextual and relational attributes of gene sets are often omitted in GSEA [8,21]. By delving into the gene set communities, we may have more opportunities to understand the overall picture of gene sets and have extra criteria to select enriched terms in GSEA.

2. Methods

2.1. General workflow

A multiplex network consists of a list of regular (monoplex) networks called layers. The layers share the same group of constituent nodes, and the inter-layer relation is simple because a node only connects with its identical node in other layers (for formal definitions, see section 2.6). A real-life example of a multiplex network can be a multiplex social network for a group of people. The same node represents the same person in networks corresponding to different life scenarios, including friendship, hobbies, colleagues, and social media followers. Fig. 1 shows the essential steps in building a multiplex network of gene sets. The analysts can choose and fetch the target collection of gene sets from a database, for example, gene ontology gene sets from org.Hs.eg.db. It also depends on the analysts to determine the data source for computing similarities in each layer of the multiplex network that establishes the biological context for measurement in that layer, which in our experiment is a cell population of a scRNA-seq dataset (section 2.2). Thus, the number of different cell populations equals the number of layers (Fig. 1.A). Other types of samples may work as well.

The Jaccard coefficient is a popular semantic or knowledge-based similarity measure for gene sets. We can modify it to be context-based by adding weights to genes according to their average expression levels in a cell population. Two gene sets are likely dissimilar if their common genes have no expression whereas their private genes dominate, and vice versa. However, the modified Jaccard coefficient cannot capture the total effect of gene expressions on similarities among gene sets because the interactions among genes are crucial and complicated. This is where multivariate association functions may come into play. We designed simulation and running time comparison studies to compare the performance of classical multivariate association functions (for more details, see sections 2.3-2.5). In particular, we focused on a k-nearest-neighbor (KNN)-based association function [12,22]. Readers can also check the comprehensive review by Josse and Holmes [12].

Once we are using a similarity measure ($\mathcal{S}_{\mathbf{KNN}}$ in section 2.4) to compute a list of similarity matrices corresponding to multiple biological conditions, we can filter edges with lower-than-threshold weights and instantiate a multiplex network of gene sets and study its structural properties (section 2.6, Fig. 1.B, C). At a high level, the layer-wise relations signify the connections between biological contexts. For illustration purposes, we select cell populations (subtypes of immune cells) that are well-known to researchers. Other than layers, the distributions of subcollections of gene sets, for example, multiplex communities or families of gene ontology gene sets, can also give insights into the behaviors of a group of gene sets of interest. For example, GO:0070161 (anchoring junction) and GO:0005681 (spliceosomal complex) are two parent gene sets in the level-3 gene ontology-cellular component (GO-CC) gene sets. By observing the distribution of their families, including descendant gene sets, we can get acquainted with all the versions of connections between these two families in both the knowledge-based and context-based networks (section 2.7). To quantify the difference in inter-family connections between the Jaccard network and multiplex networks, we introduce two metrics called within-family and between-family fold changes (section 2.7).

At the scale of individual gene sets, the structural coefficients are multifaceted relational attributes of gene sets about their connectivity across layers, which include a multiplex PageRank centrality, a multiplex clustering coefficient, and a multiplex participation coefficient (Fig. 1.C, section 2.7). The multiplex PageRank centrality and clustering coefficients are generalized from their counterparts in the monoplex network by rewarding “consistent” players with high centrality or the clustering tendency in their neighborhoods across layers. The participation coefficient evaluates the uniformity or bias of connections of a gene set over the layers in a multiplex network. Those heterogeneous features reveal the diversity of gene sets from the structural perspective and are helpful in GSEA applications. Gene sets are independently analyzed in GSEA to estimate their member genes’ deviation from a random distribution on a target gene list. Researchers frequently use a singular metric of the effect size, for example, a normalized enrichment score (NES), to determine the priority of gene sets for interpretation. We explored the potential of using the structural coefficients to expand the current scope of GSEA by reorganizing the enriched terms and providing more clues on prioritizing gene sets for better exposing terms deserving attention (section 2.8).

2.2. Single-cell RNA-seq datasets and gene sets

The pre-processed single-cell RNA-seq datasets used in this study include pbmc3k.final fetched from SeuratData, which consists of 2638 immune cells collected from the peripheral blood of a healthy human donor [23], and E-MTAB-11536 (EMBL-EBI), which involves 329,762 immune-related human cells from 16 tissues of 12 deceased donors [24]. The pre-processing included quality control, gene count normalization, log transformation, variable gene detection, feature scaling, and data integration when applicable.

The collections of GO-CC and GO biological process (GO-BP) gene sets were accessed through org.Hs.eg.db (version 3.14.0) in gseGO [25–28]. Their IDs, descriptions, and symbols of constituent genes were formatted and stored in JSON files for further use [29]. ImmuneSigDB (v2022.1/2023.1) is a collection of gene sets stemming from 389 published immunology studies. Researchers manually design comparison experiments for differential gene expression analyses. Upregulated or downregulated genes are packaged as gene sets [3]. They can be downloaded from the Molecular Signatures Database published by the Broad Institute [30].

2.3. Notation: similarity matrix of gene sets

Constructing a distance matrix (metric or non-metric) or a similarity matrix can provide insights into the relationship among observations not embedded in a common vector space [31]. For example, let $\mathcal{T} = \{U_1, \cdots, U_n\}$ be a collection of n gene sets, and given a similarity function $\rho : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, the $n \times n$ matrix with its ij-th entry as $\rho(U_i, U_j)$ is called a similarity matrix. We note that the similarity functions do not have a common definition, and association functions or the inverse of distance/dissimilar functions can work as similarity measures. The Jaccard coefficient is a popular similarity function for gene sets based on their semantic relationships. In contrast, given gene expression profiles of gene sets (e.g., gene count matrices restricted to the genes in each gene set), we can deploy numerical models to calculate the similarity.

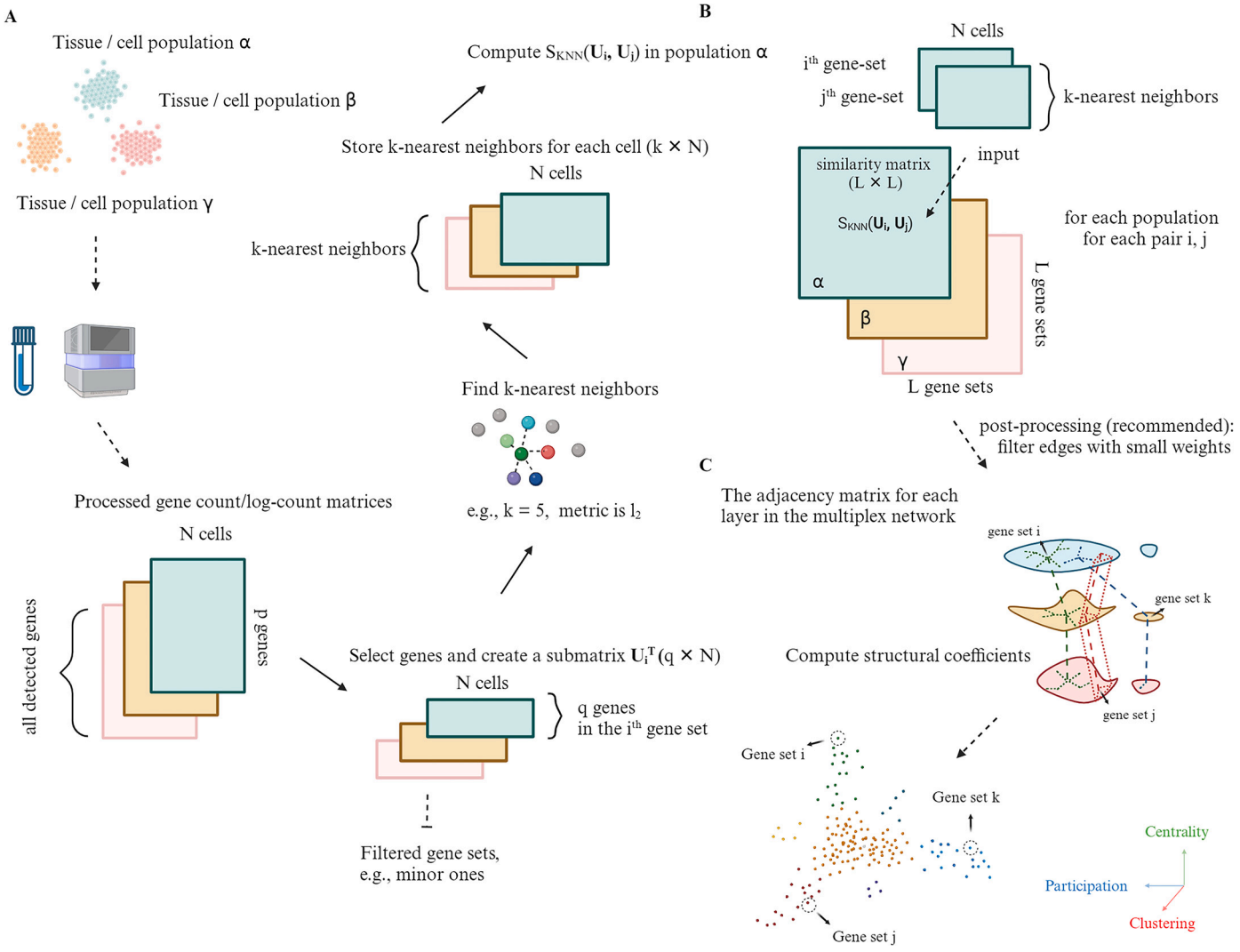

Fig. 1. Essential steps in building a multiplex network of gene sets based on $\mathcal{S}_{\mathbf{KNN}}$. (A) Cells from three populations α, β, γ are collected and sequenced for their transcripts subjected to further preprocessing. For the simplicity of illustration, we assume that the number of samples in the gene expression matrix of each layer is N, and all matrices have the same p detected genes. These constraints are not mandatory in practice. For any gene set i in a cell population, selecting q genes (rows) in the gene set among the p genes of the processed gene expression matrix produces a submatrix that determines the distribution of cells in that feature space. Gene sets with too few or too many overlapped genes or that display excessively sparse expression values in the submatrix are excluded. Different layers may have slightly different collections of gene sets that pass the filtration. We take their intersections to make the filtering collections the same. An implemented KNN algorithm with a chosen k and a metric will take the $N \times q$ submatrix as the input and return a $N \times k$ index matrix that stores the indices of the k-nearest neighbors for each cell. (B) $\mathcal{S}_{\mathbf{KNN}}$ uses a pair of index matrices corresponding to two gene sets to compute the inter-gene-set similarity (section 2.4). By iterating the computation for all pairs of gene sets in all cell populations, three $L \times L$ gene-set similarity matrices are obtained. (C) It is optional but recommended to set a threshold to truncate entries with low values in the similarity matrices to reduce the effects of random noise. The processed similarity matrices can be treated as weighted adjacency matrices for layers in a multiplex network. An example of the downstream analysis of the multiplex network is computing the structural coefficients of gene sets (section 2.7).

2.4. Classical methods for measuring similarity/multivariate association

Jaccard coefficient The Jaccard coefficient is defined as

$$\mathbf{JC}(U,V) = \frac{|U \cap V|}{|U \cup V|},$$

where U, V are two gene sets and $|\cdot|$ refers to the number of elements in a finite set [32]. To account for the effects of gene expression, we propose a modified Jaccard coefficient by weighing genes according to their average expression levels in a scRNA-seq dataset. If the mean expression level of gene k is ω_k, then the modified Jaccard coefficient is

$$\mathbf{JC_{mod}}(U,V) = \begin{cases} \dfrac{\sum_{k \in U \cap V} \omega_k}{\sum_{k' \in U \cup V} \omega_{k'}} & \text{if } \sum_{k' \in U \cup V} \omega_{k'} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

The modified RV coefficient, distance correlation coefficient, and Mantel coefficient have closed-form definitions and related tools in Python (Table 1). For the KNN-based association, we gave its definition and used sklearn.neighbors to implement it in Python.

Modified RV coefficient Suppose that gene sets U and V include p and q detected genes, and gene expression submatrices $\mathbf{U} \in \mathbb{R}^{N \times p}$ and $\mathbf{V} \in \mathbb{R}^{N \times q}$ are column-centered, where N is the number of cells in the dataset. Inputs $\mathbf{U}$ and $\mathbf{V}$ to the **RV** function return an empirical coefficient for their multivariate linear association. The modified RV coefficient ($\mathbf{RV_{mod}}$ in Table 1) was devised to eliminate the dependency on sample size by the original RV coefficient [12,33].

Distance correlation coefficient With similar notation, the empirical distance covariance (**dCov** in Table 1) for $\mathbf{U} \in \mathbb{R}^{N \times p}$ and $\mathbf{V} \in \mathbb{R}^{N \times q}$ is defined as

Table 1
Functions to measure similarity/multivariate association.

Name	Empirical Formula: $\mathbf{f(U,V)}$	Input → Output	Python Package
JC	$\frac{\lvert U\cap V\rvert}{\lvert U\cup V\rvert}$	$\mathcal{T}\times\mathcal{T}\to[0,1]$	–
$\mathbf{JC_{mod}}$	$\frac{\sum_{k\in U\cap V}\omega_k}{\sum_{k'\in U\cup V}\omega_{k'}}$ if $\sum_{k'\in U\cup V}\omega_{k'}\neq 0$, otherwise 0	$\mathcal{T}\times\mathcal{T}\to[0,1]$	–
$\mathbf{RV_{mod}}$	$\frac{\mathbf{tr}((\mathbf{UU^T}-\mathbf{diag(UU^T)})(\mathbf{VV^T}-\mathbf{diag(VV^T)}))}{\sqrt{\mathbf{tr}(\mathbf{UU^T}-\mathbf{diag(UU^T)})^2\mathbf{tr}(\mathbf{VV^T}-\mathbf{diag(VV^T)})^2}}$	$\mathbb{R}^{N\times p}\times\mathbb{R}^{N\times q}\to[-1,1]$	hoggorm [37]
dCOR	$\frac{\mathbf{dCov(U,V)}}{\sqrt{\mathbf{dCov(U,U)dCov(V,V)}}}$ if denominator > 0 otherwise 0	$\mathbb{R}^{N\times p}\times\mathbb{R}^{N\times q}\to[0,1]$	dcor [38]
$^{**}\mathbf{r_M}$	$\frac{\sum_{i,j=1}^{N}(\mathbf{X_{ij}}-\overline{\mathbf{X}})(\mathbf{Y_{ij}}-\overline{\mathbf{Y}})}{\sqrt{\mathbf{var(X)var(Y)}}}=\frac{1}{\mathbf{d-1}}\sum_{\mathbf{i,j=1}}^{\mathbf{N}}\frac{\mathbf{X_{ij}}-\overline{\mathbf{X}}}{\mathbf{s_X}}\frac{\mathbf{Y_{ij}}-\overline{\mathbf{Y}}}{\mathbf{s_Y}}$	$\mathbb{R}^{N\times N}\times\mathbb{R}^{N\times N}\to[-1,1]$	skbio [39]

$\mathbf{diag(UU^T)}$ is an $N\times N$ matrix that only contains the diagonal elements of $\mathbf{UU^T}$, similarly for $\mathbf{diag(VV^T)}$.
* Strictly speaking, a map containing the gene expression level of each gene is an additional input for the modified Jaccard function.
** $\mathbf{d=N(N-1)}$, $\overline{\mathbf{X}}=\frac{\sum_{i,j=1}^{N}\mathbf{X_{ij}}}{\mathbf{d}}$, $\mathbf{s_X}=\sum_{\mathbf{i,j=1}}^{\mathbf{N}}\frac{(\mathbf{X_{ij}}-\overline{\mathbf{X}})^2}{\mathbf{d}}$, $\mathbf{var(X)=d*s_X}$; similarly for $\mathbf{Y}$.

$$\mathbf{dCov(U,V)}=\frac{1}{N^2}\sum_{i,j=1}^{N}\mathbf{X_{ij}Y_{ij}}$$

where $\mathbf{X_{ij}}=|\mathbf{U_{i.}}-\mathbf{U_{j.}}|_p^\alpha-\frac{1}{N}\sum_{j=1}^{N}|\mathbf{U_{i.}}-\mathbf{U_{j.}}|_p^\alpha-\frac{1}{N}\sum_{i=1}^{N}|\mathbf{U_{i.}}-\mathbf{U_{j.}}|_p^\alpha+\frac{1}{N^2}\sum_{i,j=1}^{N}|\mathbf{U_{i.}}-\mathbf{U_{j.}}|_p^\alpha$; $\mathbf{U_{i.}}$ indicates the i-th row (sample) of $\mathbf{U}$; and $|\cdot|_{\mathbf{p}}^{\alpha}$ is an Euclidean norm with exponent α in $\mathbb{R}^p$. $\mathbf{Y_{ij}}$ is defined similarly [34]. Then, the empirical distance correlation is defined as

$$\mathbf{dCor(U,V)}=\begin{cases}\frac{\mathbf{dCov(U,V)}}{\sqrt{\mathbf{dCov(U,U)dCov(V,V)}}} & \text{if } \mathbf{dCov(U,U)dCov(V,V)}>\mathbf{0}\\ 0 & \text{otherwise}\end{cases}$$

Unlike $\mathbf{RV_{mod}}$, **dCov** can detect non-linear associations [12].

Mantel coefficient The inputs for the Mantel function ($\mathbf{r_M}$ in Table 1) should be two Euclidean distance matrices defined as $\mathbf{X_{ij}}=||\mathbf{U_{i.}}-\mathbf{U_{j.}}||_2$ and $\mathbf{Y_{ij}}=||\mathbf{V_{i.}}-\mathbf{V_{j.}}||_2$, where $\mathbf{U_{i.}}$ indicates the i-th row vector of $\mathbf{U}$. $\mathbf{U_{j.}}, \mathbf{V_{i.}}, \mathbf{V_{j.}}$ are defined similarly. The function $\mathbf{r_M}$ can detect non-linear multivariate associations [12,35].

KNN-based association Given a submatrix $\mathbf{U}$ and a metric (e.g., L1, L2, or cosine distance), a complete graph and its KNN subgraph can be built where in the KNN graph, each node represents a sample and k edges connect between the node and its KNNs (excluding the self-loop). We deployed sklearn.neighbors.NearestNeighbors to find and store the k-nearest neighborhoods for each sample (Fig. 1.A) using the "auto" algorithm and $\sqrt{N}$, with "L2" (Euclidean distance) as the default k and metric [36].

Suppose that the KNN graphs for two gene sets have been determined as described above. Friedman and Rafsky proposed a KNN-based coefficient denoted as Γ to measure the cumulative overlaps of the samples' neighborhoods, which in turn reflects the overall resemblance of the multivariate features [22].

$$\Gamma_{\mathbf{U,V}}=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathcal{A}_{ij}^{[\mathbf{U}]}\mathcal{A}_{ij}^{[\mathbf{V}]}$$

$$=\frac{1}{2}\sum_{i=1}^{N}|\mathbf{K_U}(i)\bigcap\mathbf{K_V}(i)|$$

where $\mathcal{A}^{[\mathbf{U}]}, \mathcal{A}^{[\mathbf{V}]}$ are the adjacency matrices for the KNN graphs built on $\mathbf{U}$ and $\mathbf{V}$ [22]; and $\mathbf{K_U}(i), \mathbf{K_V}(i)$ are the sets of KNNs of sample i (excluding itself) in $\mathbf{U}, \mathbf{V}$. We normalized $\Gamma_{\mathbf{U,V}}$ by the possible maximum counts of overlapping neighbors to derive $\mathcal{S}_{\mathbf{KNN}}$ as follows (Fig. 1.B):

$$\mathcal{S}_{\mathbf{KNN}}(\mathbf{U,V})=\frac{2\Gamma_{\mathbf{U,V}}}{kN}\in[0,1]$$

2.5. Simulation and comparison studies

In the first simulation experiment, two multivariate Gaussian random vectors $\mathbf{X}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_p), \mathbf{Y}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_q)$ in two feature spaces are independent, that is, $\mathbf{X}\perp\mathbf{Y}$, and 1000 observations are drawn from the joint distribution of $\mathbf{X},\mathbf{Y}$. We fix p to be 300 and assign q to be $100,200,\cdots,800$ (Fig. 2.A, upper left). In the second experiment, $\mathbf{X}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_p)$, and $\mathbf{Y}_1=\mathbf{X}_1,\cdots,\mathbf{Y}_q=\mathbf{X}_q, q\leq p=300$; that is, realizations of $\mathbf{Y}$ are subvectors of realizations of $\mathbf{X}$. The overlap between $\mathbf{X},\mathbf{Y}$ gradually increases as $q=10,30,\cdots,270$ (Fig. 2.A, upper right). In the third experiment, $\mathbf{X}$ and $\mathbf{Y}$ maintain the same relationship, but now $\mathbf{X}$ is a random vector of a Gaussian mixture distribution such that $\mathbf{X}|\mathbf{Z}=i\sim\mathcal{N}(\mathbf{A}\mu_i,\mathbf{AA}^T), i=1,2,3$, where $\mu_i\in\mathbb{R}^p$ is a random mean vector and $\mathbf{A}\in\mathbb{R}^{p\times p}$ is randomly generated to increase the anisotropy of data in each subpopulation; $\mathbf{Z}$ specifies the component memberships of observations. Observed data are equally drawn from three subpopulations (Fig. 2.A, lower left). In the fourth experiment, $\mathbf{X}$ is the same random vector as in the third experiment, and $\mathbf{Y}=f(\mathbf{X})=\alpha\mathbf{X}+\beta\mathbf{X}^{\circ 3}$, where $\alpha,\beta\in\mathbb{R}$ are scalar constants and $\circ$ indicates the element-wise Hadamard power. p is assigned to be $100,200,\cdots,800$ (Fig. 2.A, lower right). Finally, we measure the similarities of 50 random pairs of ImmuneSigDB gene sets. The cells for computing similarities are 1000 sampled classical monocytes in the blood of the E-MTAB-11536 dataset (Fig. 2.B). $\mathbf{RV_{mod}}$, **dCOR** (exponent = 1), $\mathbf{r_M}$, and $\mathcal{S}_{\mathbf{KNN}}$ with metrics l_1, l_2, cos and $k=31\approx\sqrt{1000}$ are deployed to measure the associations as described in previous sections. The experiment is repeated 50 times for each given p and q.

Meanwhile, we perform a running time comparison of $\mathbf{RV_{mod}}, \mathbf{r_M}$, $\mathbf{JC_{mod}}$, and $\mathcal{S}_{\mathbf{KNN}}$ with the default k and metric that all support parallel computations. The sample size is $100,200,\cdots,800$, and the number of gene sets is $200,400,\cdots,1000$. The cells are sampled from classical monocytes in the blood of the E-MTAB-11536 dataset, and gene sets are sampled from ImmuneSigDB (v2022.1). We record the similarity matrices computed by the tested methods with different sample sizes or numbers of gene sets and their computing time. The Pearson correlation of two flattened upper triangles (above the main diagonals) of similarity matrices is used to quantify the linear association of two similarity measurements [40]. To justify the impact of sample size, given a fixed collection of gene sets, we treat the measurements of $\mathbf{RV_{mod}}$ or $\mathbf{r_M}$ under maximal sample size as the reference and compute the Pearson correlations between them and other similarity measurements (Fig. 2.C). The machine used in this part is a desktop workstation (HP Z640) with 10 CPUs (1.70 GHz Xeon) and 64 GB of RAM. Moreover, to explore the impact of hyperparameters (k and metrics) on the performance of $\mathcal{S}_{\mathbf{KNN}}$, we fix the sample size to be 1000 and randomly choose 100 ImmuneSigDB gene sets, with either one of the metrics (l_2 or cos) and different k values (Fig. 2.D, E).

2.6. Representation of a multiplex network for gene sets

A multiplex network is a special subtype of a multilayer network. The latter is a triple $\mathcal{M}=(\mathrm{Y},\vec{\mathcal{G}},\mathcal{G})$, where $\mathrm{Y}=\{\alpha:\alpha\in\{1,\cdots,M\}\}$ is an index set for the M layers; $\vec{\mathcal{G}}=(\mathcal{G}_1,\cdots,\mathcal{G}_\alpha,\cdots,\mathcal{G}_M)$ is an ordered list of networks (layers); $\mathcal{G}_\alpha=(\mathbf{V}_\alpha,\mathbf{E}_\alpha)$ is an individual network where $\mathbf{V}_\alpha$ is the set of nodes and $\mathbf{E}_\alpha$ contains edges that can be either directed or undirected and weighted or unweighted; and $\mathcal{G}$ is a list of bipartite networks such as $\mathcal{G}_{\alpha,\beta}=(\mathbf{V}_\alpha,\mathbf{V}_\beta,\mathbf{E}_{\alpha,\beta})$, where $\alpha,\beta\in\mathrm{Y},\alpha\neq\beta$. In other words, $\mathcal{G}$ depicts the pairwise connections between different networks in $\vec{\mathcal{G}}$ [13].

When the nodes in $\mathbf{V}_\alpha$ remain the same for $\alpha\in\mathrm{Y}$ (so-called replica nodes) and edges between any pair of distinct layers connect exclusively to the identical replica nodes, the multilayer network $\mathcal{M}$ is called a multiplex network. Specifically, if $\mathcal{M}$ is an unweighted multiplex network, then the $|\mathbf{V}|M\times|\mathbf{V}|M$ matrix $\mathcal{A}$, called the supra-adjacency matrix, can describe its structure compactly as

$$\mathcal{A}=\begin{pmatrix}\mathcal{A}^{[1]} & \mathbf{I} & \cdots & \mathbf{I}\\ \mathbf{I} & \mathcal{A}^{[2]} & \cdots & \mathbf{I}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{I} & \mathbf{I} & \cdots & \mathcal{A}^{[M]}\end{pmatrix}$$

where $\mathcal{A}^{[\alpha]},\alpha\in\mathbf{Y}$ is the adjacency matrix for $\mathcal{G}_\alpha$ [13].

A similarity matrix of gene sets can be treated as an adjacency matrix for a layer if its entries are non-negative, for example, $\rho=\mathcal{S}_{\mathbf{KNN}}$. Furthermore, a preset positive threshold can filter edges with low weights, reducing the unwanted effects of random noise. Given a list of adjacency matrices for gene sets with the edges exclusively linked to identical gene sets, we can implement a multiplex network that satisfies the formal definition above.

2.7. Exploration of structural properties of a gene-set multiplex network

High-level properties of a gene-set multiplex network We first explore the relations among the layers and subcollections of gene sets at a high level. One way to detect the inter-layer connections is to first cluster nodes in each layer of the multiplex network [41], and then use the normalized mutual information (NMI, sklearn.metrics.normalized_mutual_info_score) to quantify the resemblance of cluster structures, which in turn reflects their layer-wise similarities [15]. The seaborn.clustermap displays the layer-wise similarities with a dendrogram showing the hierarchical clusters of layers under the default setting [42,43]. A multiplex community is a collection of gene sets with connections across multiple layers. The Leiden algorithm can be generalized for community detection in a multiplex network. The quality function for a multiplex network is defined as a weighted sum of quality in each layer as $Q(\mathcal{M},P)=\sum_k w_k Q(\mathcal{G}_k,P)$, where P is a partition of nodes that is the same for all layers and w_k is a weight for the k-th layer that can be uniform. We deploy leidenalg.ModularityVertexPartition in Python for this task, and the optimization process is the same as that in a monoplex network [41]. Another important example of a subcollection of gene sets is a family of gene ontology gene sets, including a parent gene set and its descendants. For example, GO-CC (cellular components) gene sets have a tree-like organizational structure that a directed acyclic graph can represent. The difference is that some gene sets may have more than one parent. The connections between two families of gene sets may change drastically in layers of the multiplex network compared to the knowledge-based Jaccard network. To quantify this difference, we propose a fold change of similarities as follows:

$$\mathbf{FC}(\mathcal{F},\mathcal{H})=\frac{\frac{1}{M|\mathcal{F}||\mathcal{H}|}\sum_{i=1}^{M}\sum_{U\in\mathcal{F},V\in\mathcal{H}}\mathcal{S}_{\mathbf{KNN}}(\mathbf{U}_i,\mathbf{V}_i)}{\frac{1}{|\mathcal{F}||\mathcal{H}|}\sum_{U\in\mathcal{F},V\in\mathcal{H}}\mathbf{JC}(U,V)}=\frac{\sum_{i=1}^{M}\sum_{U\in\mathcal{F},V\in\mathcal{H}}\mathcal{S}_{\mathbf{KNN}}(\mathbf{U}_i,\mathbf{V}_i)}{M\sum_{U\in\mathcal{F},V\in\mathcal{H}}\mathbf{JC}(U,V)}$$

where M is the number of layers in $\mathcal{M}$; and $\mathbf{U}_i,\mathbf{V}_i$ are gene expression matrices corresponding to the gene sets U,V in the i$^{\text{th}}$ layer. Alternatively speaking, **FC** is the ratio between the average score of $\mathcal{S}_{\mathbf{KNN}}$ for all pairs of gene sets U,V in two families $\mathcal{F},\mathcal{H}$ across all layers of $\mathcal{M}$ and their average **JC** score in the Jaccard network. When $\mathcal{F},\mathcal{H}$ are the same family of gene sets, we call it a within-family fold change. If the numerator and denominator are zero, we assign their **FC** as a missing value. If only the denominator is zero, **FC** is conservatively evaluated to be 1. To consider the overall context-over-knowledge difference in the connections between a family of gene sets $\mathcal{F}$ and other families, for example, the collections of level-3 GO-CC gene sets and their descendants, we define the between-family fold change as

$$\frac{\sum_{\mathcal{H}\neq\mathcal{F},\mathcal{H}\in\Omega}\mathbf{FC}(\mathcal{F},\mathcal{H})}{|\Omega|-1}$$

where Ω is a collection of families. The python package goatools is used in this task to search level-3 GO-CC gene sets and their descendants [44].

We use the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) and graphical interface to visualize multi-omics networks (Grimon) to visualize the distributions of gene sets. UMAP assumes that the data are distributed uniformly on a manifold, and computationally, it modifies the raw distance/dissimilarity matrix and embeds the newly constructed graph in the low-dimensional space via a force-directed graph layout algorithm [45]. UMAP suits our needs as a fast and straightforward tool without displaying excessive information such as edges. Additionally, we apply Grimon to aggregate and adjust the UMAP coordinates of all the layers in the multiplex network for a unified view [46].

Structural coefficients of gene sets At the scale of individual nodes (gene sets), several structural coefficients have been proposed for a multiplex network with unweighted and undirected layers [15,16], which can be generalized for multiplex networks with weighted (or directed) layers. The multiplex participation coefficient for node i in a multiplex network $\mathcal{M}$ with weighted layers can be

$$P_i=\frac{M}{M-1}\left(1-\sum_{\alpha\in\mathrm{Y}}(\frac{k_i^\alpha}{O_i})^2\right)$$

where $k_i^\alpha=\sum_{j\neq i}\mathcal{A}_{ij}^{[\alpha]}$ is the sum of weights for incident edges of a node i at layer α and $O_i=\sum_{\alpha\in\mathrm{Y}}k_i^\alpha$. A large P_i implies that node i has consistent connections across layers. If P_i is as low as 0, node i has connectivity at a single layer [15].

The (local) clustering coefficient of node i quantifies the tendency of nodes in its (open) neighborhood to form connected links (together as triangles) in a network, which is initially proposed for unweighted and undirected graphs as $\mathbf{C_i}=\frac{\sum_{j\neq i,k\neq i,k\neq j}\mathcal{A}_{ij}\mathcal{A}_{jk}\mathcal{A}_{ki}}{\sum_{j\neq i,k\neq i,k\neq j}\mathcal{A}_{ij}\mathcal{A}_{ki}}=\frac{\sum_{j\neq i,k\neq i,k\neq j}\mathcal{A}_{ij}\mathcal{A}_{jk}\mathcal{A}_{ki}}{K_i(K_i-1)}$, where $\mathcal{A}$ is an adjacency matrix and $K_i=\sum_{j\neq i}\mathcal{A}_{ij}$ [47]. For weighted and undirected networks, $\mathcal{A}_{ij}$ can be the non-negative normalized weight of edge E_{ij} (i.e., $\tilde{w}_{ij}=\frac{w_{ij}}{\max_{ij}w_{ij}}$), and $K_i(K_i-1)$ becomes $(\sum_j\tilde{w}_{ij})^2-\sum_k\tilde{w}_{ki}^2$ [48,49]. In the case of directed and weighted graphs, we may consider a specific pattern of triangles with fixed edge directions (e.g., $\mathbf{C_i}=\frac{\sum_{j\neq i,k\neq i,k\neq j}\tilde{w}_{ij}\tilde{w}_{ik}\tilde{w}_{jk}}{\sum_{j\neq i,k\neq i,k\neq j}\tilde{w}_{ij}\tilde{w}_{ik}}$) [50] or all possible triangles that contain node i as their vertex (i.e., $\mathbf{C_i}=\frac{\sum_{j\neq i,k\neq i,k<j}(\tilde{w}_{ij}+\tilde{w}_{ji})(\tilde{w}_{ik}+\tilde{w}_{ki})(\tilde{w}_{jk}+\tilde{w}_{kj})}{2\sum_{j\neq i,k\neq i,k<j}(\tilde{w}_{ij}+\tilde{w}_{ji})(\tilde{w}_{ik}+\tilde{w}_{ki})}$) [51,52].

An extended version for a multiplex network can be developed by placing the edges of a triangle on two or three different layers, which correspond to coefficients $\mathbf{C}_{i,1}$ and $\mathbf{C}_{i,2}$ [15,16]. Without loss of generality, we assume that $\mathcal{G}_\alpha$ is directed and weighted and has no self-loops for

all $\alpha \in \mathbf{Y}$. The equations in the definitions can be transformed into matrix forms to facilitate vectorization and parallel processing. The proof is in the appendix.

$$\mathbf{C}_{i,1} = \frac{\sum_{\alpha}\sum_{\alpha'\neq\alpha}\sum_{j\neq i,m\neq i,j<m}(\mathcal{A}_{ij}^{[\alpha]}+\mathcal{A}_{ji}^{[\alpha]})(\mathcal{A}_{jm}^{[\alpha']}+\mathcal{A}_{mj}^{[\alpha']})(\mathcal{A}_{mi}^{[\alpha]}+\mathcal{A}_{im}^{[\alpha]})}{2(M-1)\sum_{\alpha}\sum_{j\neq i,m\neq i,j<m}(\mathcal{A}_{ij}^{[\alpha]}+\mathcal{A}_{ji}^{[\alpha]})(\mathcal{A}_{mi}^{[\alpha]}+\mathcal{A}_{im}^{[\alpha]})}$$

$$= \frac{\sum_{\alpha}\sum_{\alpha'\neq\alpha}\mathbf{Diag}((\mathcal{A}^{[\alpha]}+\mathcal{A}^{[\alpha]^T})(\mathcal{A}^{[\alpha']}+\mathcal{A}^{[\alpha']^T})(\mathcal{A}^{[\alpha]}+\mathcal{A}^{[\alpha]^T}))_{ii}}{2(M-1)\sum_{\alpha}(((\mathcal{A}^{[\alpha]}+\mathcal{A}^{[\alpha]^T})\mathbf{1})^{\circ 2}-(\mathcal{A}^{[\alpha]}+\mathcal{A}^{[\alpha]^T})^{\circ 2}\mathbf{1})_i}$$

$$\mathbf{C}_{i,2} = \frac{\sum_{\alpha}\sum_{\alpha'\neq\alpha}\sum_{\alpha''\neq\alpha,\alpha'}\sum_{j\neq i,m\neq i,j\neq m}(\mathcal{A}_{ij}^{[\alpha]}+\mathcal{A}_{ji}^{[\alpha]})(\mathcal{A}_{jm}^{[\alpha'']}+\mathcal{A}_{mj}^{[\alpha'']})(\mathcal{A}_{mi}^{[\alpha']}+\mathcal{A}_{im}^{[\alpha']})}{2(M-2)\sum_{\alpha}\sum_{\alpha'\neq\alpha}\sum_{j\neq i,m\neq i,j\neq m}(\mathcal{A}_{ij}^{[\alpha]}+\mathcal{A}_{ji}^{[\alpha]})(\mathcal{A}_{mi}^{[\alpha']}+\mathcal{A}_{im}^{[\alpha']})}$$

$$= \frac{\sum_{\alpha}\sum_{\alpha'\neq\alpha}\sum_{\alpha''\neq\alpha,\alpha'}\mathbf{Diag}((\mathcal{A}^{[\alpha]}+\mathcal{A}^{[\alpha]^T})(\mathcal{A}^{[\alpha'']}+\mathcal{A}^{[\alpha'']^T})(\mathcal{A}^{[\alpha']}+\mathcal{A}^{[\alpha']^T}))_{ii}}{2(M-2)\sum_{\alpha}\sum_{\alpha'\neq\alpha}(((\mathcal{A}^{[\alpha]}+\mathcal{A}^{[\alpha]^T})\mathbf{1})\odot((\mathcal{A}^{[\alpha']}+\mathcal{A}^{[\alpha']^T})\mathbf{1})-((\mathcal{A}^{[\alpha]}+\mathcal{A}^{[\alpha]^T})\odot(\mathcal{A}^{[\alpha']}+\mathcal{A}^{[\alpha']^T}))\mathbf{1})_i}$$

We refer to a multiplex version of the PageRank coefficient, which reflects the global centrality of a node in a multiplex network, and implement it in Python [17]. The algorithm attempts to integrate centrality information from layers iteratively. Suppose that the adjacency matrices $\{\mathcal{A}^{[\mathcal{L}_k]}\}$ of layers $(\mathcal{L}_1,\cdots,\mathcal{L}_m)$ in a multiplex network are column-normalized (i.e., $\sum_{i=1}^{N}\mathcal{A}_{ij}^{[\mathcal{L}_k]}=1$ for all k,j), and the multiplex PageRank centrality for nodes in layers $(\mathcal{L}_1,\cdots,\mathcal{L}_{k-1})$ denoted as $\mathbf{X}^{k-1}=[x_1^{k-1},\cdots,x_N^{k-1}]^T$ has already been determined. Then, the multiplex PageRank centrality for node i in layers $(\mathcal{L}_1,\cdots,\mathcal{L}_k)$ is

$$x_i^k = \alpha_{\mathcal{L}_k}\sum_{j=1}^{N}(x_i^{k-1})^{\beta}\mathcal{A}_{ij}^{[\mathcal{L}_k]}\frac{x_j^k}{g_j^k}+(1-\alpha_{\mathcal{L}_k})\frac{(x_i^{k-1})^{\gamma}}{\sum_{r=1}^{N}(x_r^{k-1})^{\gamma}}$$

where $k=2,\cdots,m$; $g_j^k=\sum_{r=1}^{N}\mathcal{A}_{rj}^{[\mathcal{L}_k]}(x_r^{k-1})^{\beta}+\delta(0,\sum_{r=1}^{N}\mathcal{A}_{rj}^{[\mathcal{L}_k]}(x_r^{k-1})^{\beta})$; δ is the Kronecker delta function; and $\alpha_{\mathcal{L}_k},\gamma,\beta$ are scalar hyperparameters (defaults are respectively 0.85, 1, and 1 in our experiments) [17]. The equation can be expressed more concisely as

$$\mathbf{X}^k = \alpha_{\mathcal{L}_k}\hat{\mathcal{M}}\mathbf{X}^k+(1-\alpha_{\mathcal{L}_k})\frac{(\mathbf{X}^{k-1})^{\circ\gamma}}{\mathbf{1}^T(\mathbf{X}^{k-1})^{\circ\gamma}}$$

where $\mathcal{M}=\mathbf{D}\mathcal{A}^{[\mathcal{L}_k]}$, $\hat{\mathcal{M}}$ is its column-normalized form (entries in a column remain zeros if divided by a zero column sum), and $\mathbf{D}$ is a diagonal matrix such that $\mathbf{D}_{ii}=(x_i^{k-1})^{\beta}$. The base case is when $k=1$ and $[x_1^1,\cdots,x_N^1]^T$ is defined to be the same as the PageRank centrality of a monoplex network $\mathcal{L}_1$ [53]. The numeric computation is through iterative procedures, and the final converged multiplex PageRank centrality for node i is $x_i=x_i^m$ [17]. We use the ratio of raw centrality to a uniform mass $\frac{1}{\text{number of gene sets}}$ to represent the centrality of nodes in this study.

2.8. Applications in gene set enrichment analysis

GSEA is a popular tool for interpreting the functional meaning of a target gene list. Given an ordered list of n weighted genes $\Omega=\{g_1,\cdots,g_n\}$, for example, differentially expressed genes ranked by a metric for differences or adjusted p-values in hypothesis tests, the enrichment score (**ES**) for a gene set U is defined as follows:

$$\mathbf{S_{hit}}(U,i)=\sum_{\substack{g_k\in U\\ k\le i}}\frac{|\rho_k|^{\alpha}}{\sum_{g_k\in U}|\rho_k|^{\alpha}}$$

$$\mathbf{S_{miss}}(U,i)=\sum_{\substack{g_k\notin U\\ k\le i}}\frac{1}{n-|U|}$$

$$p=\arg\max_i|\mathbf{S_{hit}}(U,i)-\mathbf{S_{miss}}(U,i)|$$

$$\mathbf{ES}(U)=\mathbf{S_{hit}}(U,p)-\mathbf{S_{miss}}(U,p)$$

where p,i are positions in the ordered list of genes, ρ_k is the weight for the gene g_k, and α is a scalar hyperparameter [18]. By permutations, a null distribution can be built. The proposal of the NES aims to reduce the effect of different sizes of gene sets in multiple-hypothesis testing by rescaling the enrichment score [18]. For a gene set U with a highly positive or negative NES, the genes in U are unlikely to be randomly distributed in Ω, which implies the existence of their underlying connection. We use clusterProfiler (version 3.17) in R to conduct GSEA [26].

Practitioners often deal with many enriched gene sets in an experiment. A typical routine to determine the priority of enriched terms is to rank them by their (normalized) enrichment scores, gene overlap ratio (GeneRatio), or adjusted p-values, and only the gene sets with top ranks are emphasized. Those metrics mainly consider the effect size or statistical significance of the distributional deviation of genes. Some researchers may even pick their terms of interest, regardless of the ranks of those gene sets. However, a neurological term enriched in a study about gliomas may have a fundamentally different meaning if it were enriched in a rheumatology study. GSEA treats each gene set independently, and the interpretation of enriched terms does not explicitly consider contextual and relational attributes of gene sets, which the structural coefficients can supplement.

We use the multiplex clustering coefficients (C1 and C2) as indicators to find local clusters of gene sets with significant similarities by searching the "consistent" neighbors of gene sets with high-level clustering coefficients, for example, sharing the neighborhood over the layers ($\mathcal{S}_{\mathbf{KNN}}>0.2$). Then, rather than relying on the knowledge of gene sets' hierarchical organizations [8,21], local representatives of the groups are selected if they have maximal (local) multiplex PageRank centrality in the groups, which produces context-based representatives of the groups in a data-driven manner. The gene sets other than the representatives in the group are removed from the enriched terms. As many terms with high (global) multiplex centrality are gene sets too ambiguous for interpretation, we can concentrate on more specific terms with lower multiplex centrality. Regarding the multiplex participation coefficient, $(\frac{k_i^{\alpha}}{O_i})^2\in[0,1]$ (section 2.7) is a scalar value to reflect the biased "participation" of a gene set i in a particular layer α. By examining gene sets' participation coefficients and per-layer participation, we know the specificity of their connectivities to particular layers. To combine them, we use multiplex clustering coefficients and PageRank centrality to reorganize the raw structures of enriched terms in GSEA to reduce redundancy and skip enriched terms with overly broad functions. Then, we can prioritize gene sets by both NES and multiplex centrality levels. The alternative way is to follow the orders of structural coefficients of gene sets, for example, multiplex participation coefficients, depending on the interest of researchers. We find that the multipartite graph is a valuable tool (networkx.multilayered_graph) to visualize the enriched terms ranked by different standards in a sensible way [54].

3. Results and discussion

3.1. Computing performance of $\mathcal{S}_{\mathbf{KNN}}$

As described in section 2.5, we first conduct a simulation study to test the ability of selected similarity measures to reflect various simulated relations (Fig. 2.A, B). For simulation dataset 1, each data point i has the same probability $\frac{K}{N-1}$ of being the KNN of another point j, and the neighborhoods in the random matrices $\mathbb{X}$ and $\mathbb{Y}$ are also independent. Thus, we can estimate the expectation of $\mathcal{S}_{\mathbf{KNN}}(\mathbb{X},\mathbb{Y})$ as $\mathbf{E}[\frac{2*\frac{1}{2}\sum_{i=1}^{N}|\mathbf{K}_{\mathbb{X}}(i)\bigcap\mathbf{K}_{\mathbb{Y}}(i)|}{KN}]=\frac{1}{KN}\sum_{i=1}^{N}\mathbf{E}[|\mathbf{K}_{\mathbb{X}}(i)\bigcap\mathbf{K}_{\mathbb{Y}}(i)|]=\frac{1}{KN}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{E}[\mathbf{1}_{j\in\mathbf{K}_{\mathbb{X}}(i)\bigcap\mathbf{K}_{\mathbb{Y}}(i)}]=\frac{N(N-1)(\frac{K}{N-1})^2}{KN}=\frac{K}{N-1}$. In this experiment, $K=31\approx\sqrt{N}$, and we observe the scores of $\mathcal{S}_{\mathbf{KNN}}$ fluctuate on its expectation $\frac{K}{N-1}\approx 0.03103$ with a little variance no matter which metric is

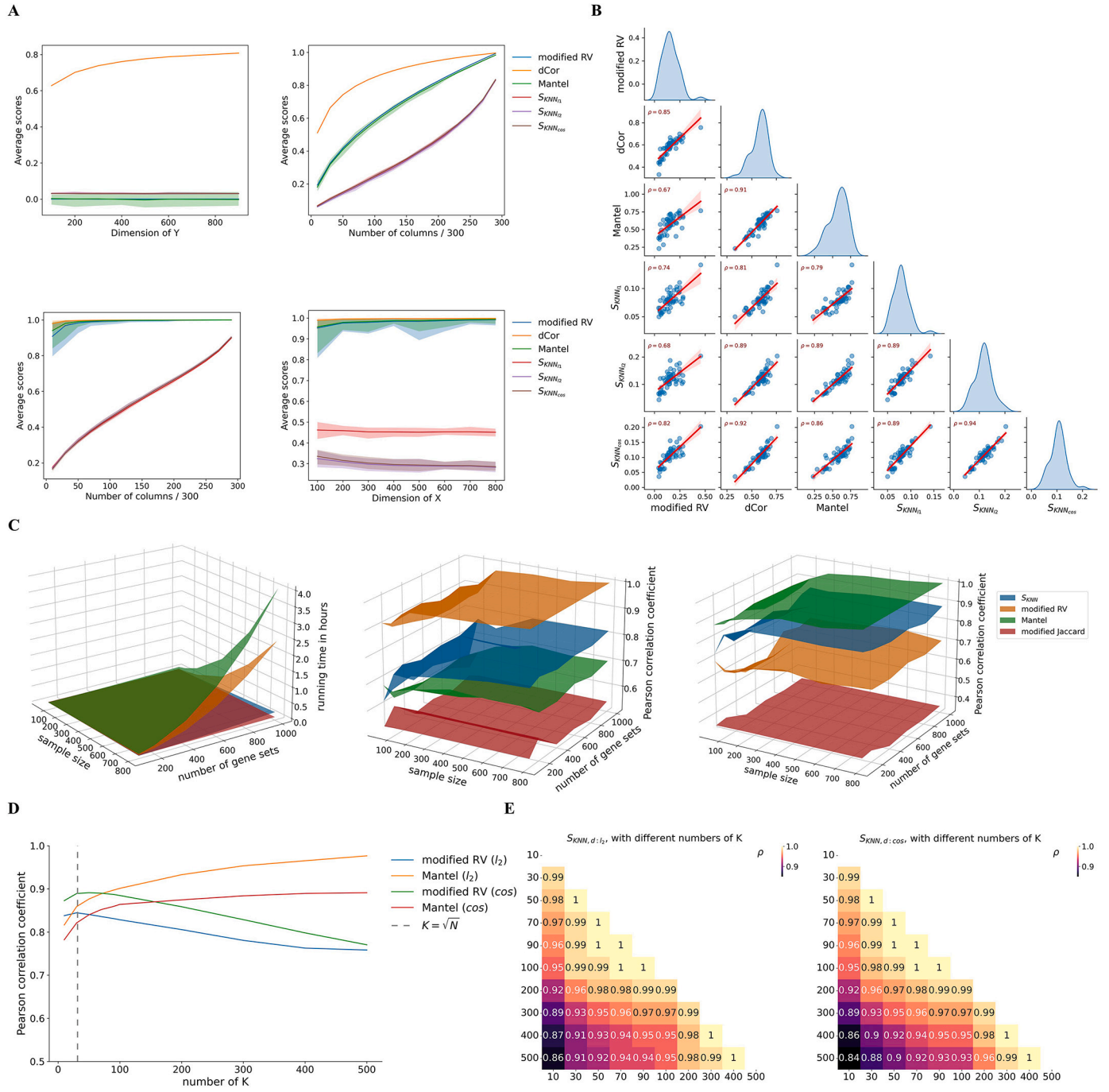

Fig. 2. Results of the simulation and comparison studies. (A) Upper left: Most similarity measures exhibit close-to-zero scores for a pair of independent datasets, except for **dCOR** (exponent = 1). Upper right, lower left: $\mathcal{S}_{\mathbf{KNN}}$ captures the similarity gradient between a dataset and its subset with increasingly overlapped features. Lower right: The similarity corresponding to a non-linear function remains relatively stable with inputs of various dimensions. The widths of colored bands represent the range of scores in 50 repeated experiments. (B) Outcomes by different similarity measures for real gene sets. (C) The running time of $\mathbf{RV}_{\mathbf{mod}}$ and $\mathbf{r}_{\mathbf{M}}$ grows much faster than that of $\mathcal{S}_{\mathbf{KNN}}$ and $\mathbf{JC}_{\mathbf{mod}}$ as the sample size or number of gene sets increases (left). The measurement by $\mathbf{JC}_{\mathbf{mod}}$ does not correlate well with the outcomes of $\mathbf{RV}_{\mathbf{mod}}$ (middle) or $\mathbf{r}_{\mathbf{M}}$ (right), though it has the shortest running time. (D, E) The impact of hyperparameters K and metrics on the performance of $\mathcal{S}_{\mathbf{KNN}}$.

loaded (Fig. 2.A, upper left). In other words, for two completely random and independent datasets, $\mathcal{S}_{\mathbf{KNN}}$ will give a near-zero similarity score in the choice of a reasonable K (e.g., $\sqrt{N}$). The difference between simulation datasets 2 and 3 is that dataset 2 has no internal structures while dataset 3 has three separate clusters. $\mathcal{S}_{\mathbf{KNN}}$ smoothly captures the gradient of similarity between a data matrix and its submatrix as the number of columns increases. However, other similarity measures might be too sensitive to the global structures of datasets as their similarity scores approach 1 as soon as the submatrix recruits about 50 (over 300) columns of the raw data matrix (Fig. 2.A, upper right and lower left). Given a non-linear function, $\mathcal{S}_{\mathbf{KNN}}$ and other similarity measures can assign a non-trivial and consistent score to it regardless of the dimension of the input dataset, though different metrics embedded in $\mathcal{S}_{\mathbf{KNN}}$ may lead to divergent scores (Fig. 2.A, lower right). Fig. 2.B shows the correla-

tions between pairs of different similarity measures for 50 ImmuneSig gene sets. $\mathcal{S}_{\mathbf{KNN}}$ with default metric l_2 and $K = \sqrt{N}$ has an intermediate correlation (0.68) with $\mathbf{RV}_{\mathbf{mod}}$, but a strong correlation (0.89) with $\mathbf{r}_{\mathbf{M}}$.

We then perform a running time comparison experiment on $\mathbf{RV}_{\mathbf{mod}}, \mathbf{r}_{\mathbf{M}}, \mathbf{JC}_{\mathbf{mod}}$, and $\mathcal{S}_{\mathbf{KNN}}$ with the default metric and K. For an execution with 800 samples and 1000 target gene sets, $\mathcal{S}_{\mathbf{KNN}}$ outperforms $\mathbf{RV}_{\mathbf{mod}}$ and $\mathbf{r}_{\mathbf{M}}$ remarkably having more than one order of magnitude less time cost (10.4 minutes versus 2.4 and 4.0 hours). $\mathbf{JC}_{\mathbf{mod}}$ has an even lower time consumption (Fig. 2.C, left) but at the cost of worse performance. For simplicity of discussion, we assume that all L gene sets have the same size as s, and the computational complexity of the multiplication of an $N \times s$ matrix and its transpose is $O(sN^2)$. Then, from section 2.4 and Table 1, the time complexities of $\mathbf{RV}_{\mathbf{mod}}$ and $\mathbf{r}_{\mathbf{M}}$ are $O(sN^2 + N^3)$ and $O(sN^2)$, respectively. For computing the $L \times L$ similarity matrix, their time complexities are $O[(sN^2 + N^3)L^2]$ and $O(sN^2L^2)$, respectively. In contrast, the time complexity of computing the L index matrices (Fig. 1) is $O(sN^2L)$ by the brute algorithm of sklearn.neighbors. Computing $\mathcal{S}_{\mathbf{KNN}}$ based on a pair of index matrices (Fig. 1) would require the complexity of $O(KN)$. Thus, the total complexity for computing the similarity matrix is $O(sN^2L + KNL^2)$. The time complexity of $\mathbf{JC}_{\mathbf{mod}}$ for computing the similarity matrix is $O(sL^2)$ given a map of average gene expressions of genes as its input.

Regarding the performance, the measurements by $\mathcal{S}_{\mathbf{KNN}}, \mathbf{RV}_{\mathbf{mod}}, \mathbf{r}_{\mathbf{M}}$ are mutually associated and influenced by the sample size. With the measurement by $\mathbf{RV}_{\mathbf{mod}}$ or $\mathbf{r}_{\mathbf{M}}$ under the largest sample size as a reference (Fig. 2.C, middle right), the correlations between the reference and the measurements by $\mathbf{RV}_{\mathbf{mod}}$ or $\mathbf{r}_{\mathbf{M}}$ under different sample sizes gradually become stable as the sample size increases. Because the cells in this experiment are sampled from a homogenous population, a requirement for a larger sample size for the convergence of measurements is expected if they were sampled from a heterogeneous population. The correlations between the reference and measurements by $\mathcal{S}_{\mathbf{KNN}}$ lie between $\mathbf{RV}_{\mathbf{mod}}$ and $\mathbf{r}_{\mathbf{M}}$. For example, compared to $\mathbf{r}_{\mathbf{M}}$, $\mathcal{S}_{\mathbf{KNN}}$ is closer to $\mathbf{RV}_{\mathbf{mod}}$ and farther from $\mathbf{r}_{\mathbf{M}}$. The measurements by $\mathbf{JC}_{\mathbf{mod}}$ have poor Pearson correlations with both methods $\mathbf{RV}_{\mathbf{mod}}$ and $\mathbf{r}_{\mathbf{M}}$, which suggests that gene membership and expression levels alone cannot explain the interactions among gene sets.

We further explore the impact of K and metrics on the outcomes of $\mathcal{S}_{\mathbf{KNN}}$. With an increasing number of K ($< \frac{1}{2}N$), $\mathcal{S}_{\mathbf{KNN}}$ has the tendency of approaching the non-linear association function $\mathbf{r}_{\mathbf{M}}$ and deviating from the linear association function $\mathbf{RV}_{\mathbf{mod}}$ with both metrics l_2 and cos (Fig. 2.D). $\mathcal{S}_{\mathbf{KNN}}$ loaded with the l_2 metric is significantly closer to $\mathbf{r}_{\mathbf{M}}$ than when loaded with the cos metric (p-value = 0.002 by Wilcoxon signed-rank test). Oppositely, $\mathcal{S}_{\mathbf{KNN}}$ loaded with the cos metric is significantly closer to $\mathbf{RV}_{\mathbf{mod}}$ than when loaded with l_2 (p-value = 0.002 by Wilcoxon signed-rank test). Thus, the hyperparameters K and metrics can adjust the inclination of $\mathcal{S}_{\mathbf{KNN}}$ towards a linear or non-linear similarity measure. But in general, a change of hyperparameter K does not abruptly switch the effect of measurement: for K in the interval around 100, the outcomes of $\mathcal{S}_{\mathbf{KNN}}$ correlate strongly (Fig. 2.E).

Similarity measures like $\mathcal{S}_{\mathbf{KNN}}$ have some limitations. Unlike the deterministic semantic relations, the quality of gene expression datasets influences the computational results. For example, a scRNA-seq dataset with low coverage depth may have few detected genes for gene sets, resulting in a sparse count matrix that is unreliable in predicting the similarity. Due to the stochastic expression nature of genes, especially regulatory genes, the current scRNA-seq technology may not reliably capture their gene expression levels, which may underestimate their connectivity in the community [55]. In addition, the current exclusion criteria may reject some “true negative” gene sets that are silent under specific contexts.

3.2. The multiplex network model can capture the biological variation of gene sets in context

We use $\mathcal{S}_{\mathbf{KNN}}$ with the default setting to compute similarities among gene sets in GO-CC and ImmuneSig. Five cell populations from human donors (EMBL-EBI) that include memory B cells in spleens, alveolar macrophages in lungs, natural killer cells in lung-draining lymph nodes, naive T cells in jejunum epithelium, and resident memory T cells in bone marrow set up the biological context for the measurements, corresponding to layers in the multiplex network (Fig. 3). We exclude gene sets that are too small or too large, namely, less than 10 or greater than 2000 genes in a gene expression matrix. To mitigate noise in the network, edges whose weights are lower than a positive threshold (e.g., 0.2) are discarded. We add an auxiliary network built on the Jaccard similarity matrix as a knowledge-based reference.

The layer-wise similarities are measured by NMI, and hierarchical clustering is performed (section 2.7). Layers for T cells and natural killer cells are clustered close to the layer for B cells. The layer for macrophages is on another branch of the hierarchical tree (Fig. 3.A). The hierarchical structure of layers in the multiplex network fits the model of hematopoiesis and the classification of peripheral agranular leukocytes in physiological states [56], which demonstrates the efficacy of multiplex networks for capturing a genuine biological variation.

In contrast, the knowledge-based Jaccard network has the least NMI with all layers in the multiplex network. In the UMAP space of the embedded Jaccard layer, GO-CC gene sets barely cover ImmuneSig gene sets. Conversely, these two groups of gene sets often significantly overlap in the UMAP space of embedded layers of the multiplex network (Fig. 3.D). The difference between a knowledge-based and a context-based network is prominent in this example. From a biological standpoint, this is reasonable, as gene sets in ImmuneSig rely on specific molecular machinery to conduct immunological or other elementary functions. Because GO-CC gene sets have a tree-like hierarchical structure, we are interested in how the difference between knowledge-based and context-based networks is decomposed into within-family and between-family portions (section 2.7).

The chosen families of gene sets are level-3 GO-CC gene sets and their descendants. Our finding (Fig. 3.C) shows that the between-family fold changes (median 2, range 0.86-16.7) are remarkably higher than within-family fold changes (median 0.99, range 0.8-1.23). For instance, the family GO:0070161 (anchoring junction) exhibits the highest ratio of these two metrics. It has a within-family fold change of 0.93, implying its within-family relations are, on average, similar to those found in the knowledge-based reference. In contrast, its between-family fold change with other families is 16.74, indicating that their contextual relations are quite different from those in the knowledge-based approach. A specific example is the family GO:0005681 (spliceosomal complex). The proximity of its family members to the GO:0070161 family in the UMAP space is illustrated in Fig. 3.D. In the Jaccard layer, these two families are separated by distance. But in the layers of the multiplex network, many of their descendant gene sets (e.g., GO:0014704 and GO:0071006) become variably close to each other. A biological explanation is that spliceosomes actively process pre-mRNA, and the spliced mRNA will impact the product of many molecules, including those related to the anchoring junction. For immune cells, the anchoring junction relates to, for example, cell migration, cell adherence, and antigen presentation. Thus, unsurprisingly, these two families of gene sets interact in context-based networks. Many of their family members aggregate in the embedded UMAP space (Fig. 3.D) in line with their within-family fold change. Thus, the model of multiplex networks of gene sets can provide insights into context-based gene-set relations that significantly differ from their knowledge-based counterpart.

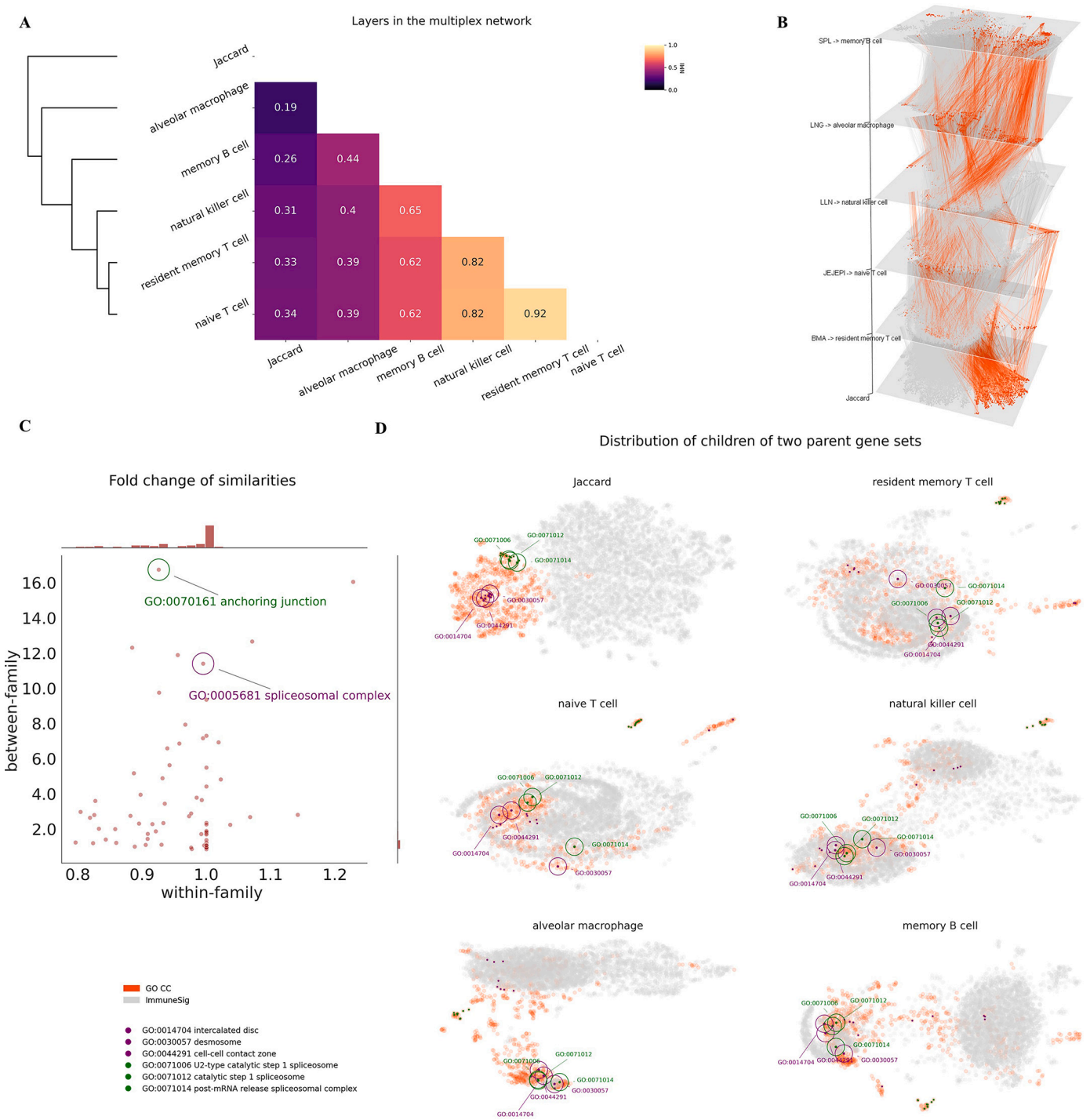

Fig. 3. Difference between the knowledge-based Jaccard network and the context-based multiplex network. (A) NMI scores indicate similarity between pairs of layers. The NMI scores and the hierarchical clustering show that the Jaccard layer differs significantly from all other layers in the multiplex network. (B) View of the distributions of GO-CC and ImmuneSig gene sets by Grimon. Each splice corresponds to a plot in D by its label. The coordinates of gene sets in the slices are UMAP coordinates adjusted by Grimon. There is a pronounced change in terms of the interactions between gene sets of GO-CC and ImmuneSig in the Jaccard network and layers in the multiplex network. (C, D) Most level-3 GO-CC gene sets and their descendants show within-family fold changes around 1. In contrast, their between-family fold changes can be much higher. Families GO:0070161 and GO:0005681 are parent gene sets. Their between-family relations change remarkably from the knowledge-based Jaccard layer to layers in the multiplex network, as visualized by the distributions of their family gene sets in the embedded UMAP space. Some of their descendant gene sets (marked by circles) become close in the multiplex network embedding.

3.3. Structural coefficients can provide multifaceted information for gene sets in a multiplex network

Now, we turn to another study on the multiplex network of GO-BP gene sets. The scRNA-seq dataset for computing $\mathcal{S}_{\mathbf{KNN}}$ is pbmc3k.final, which can be easily fetched by SeuratData. Cell populations for measurements include naive CD4 T cells, memory CD4 T cells, CD8 T cells, B cells, and CD14 monocytes (section 2.2). The threshold for truncating edge weights is relaxed to 0.1 to preserve more network variations, which results in low pairwise NMI scores among the layers in the multiplex network. The benefit is to reduce isolated nodes and include more gene sets in the measurements of their structural properties. Nevertheless, the clusters of T cell subtypes and the affinity between CD8 T cells and monocytes are observable in the dendrogram (Fig. 4.A). The knowledge-based Jaccard layer again shows the lowest layer-wise similarities to context-based layers in the multiplex network (Fig. 4.A).

Multiplex community detection reveals five major clusters after merging small groups (less than 200 over 5336 filtered gene sets) of gene sets into one extensive collection. The two most significant communities (blue and orange, communities 0 and 1 in Fig. 4.D) are visibly separated (Fig. 4.B, leftmost). Their primary difference is that the gene sets in community 0 are mainly large gene sets with weaker connections to other gene sets. In comparison, the gene sets in community 1 are usually small gene sets with stronger relations to others (Fig. 4.D, bottom left). Community 4 (purple) is an extensive collection merged from isolated small clusters of gene sets. Interestingly, members in community 4 are mainly small gene sets that possess high multiplex clustering coefficients (Fig. 4.D), indicating they are somehow borderline and tightly interconnected groups in the community. In the auxiliary Jaccard network, the partitions of multiplex communities disappear as they blend, which can be explained by the fundamental difference between a knowledge-based network and a context-based multiplex network. In summary, the multiplex communities in this experiment are primarily determined by their basic structural properties in the network, that is, gene set size and connections.

Several structural coefficients for individual gene sets can be computed, as described in section 2.7. They reflect the diversity of gene sets in their community from a structural perspective. They are concise summaries of gene sets' particular local, global, or dynamic features, as shown in Fig. 4.B (second from the left to the rightmost), which visualizes the patterns of three different structural coefficients in the multiplex network. We provide the code and data in the GitHub repository for these 3D visualizations. GO:0034368 (protein-lipid complex remodeling) is the gene set with the highest multiplex PageRank centrality (ratio of 30.03). The protein-lipid complex has been reported to be a positive regulator of immune cells [57]. GO:0051240 (positive regulation of multicellular organismal process) is the gene set with the highest PageRank centrality (ratio of 2.74) in the Jaccard layer. Compared with GO:0034368, it gives less contextual information, and regardless of which cell population, it has the same PageRank centrality given the same collection of GO-BP gene sets. GO:0019835 (cytolysis) exhibits the lowest value for the participation coefficient, reflecting its biased connections in specific layers. Further investigation shows that the connections between GO:0019835 and other gene sets are significantly more active in the "B cells" layer. GO:0015867 (ATP transport) has the highest level of multiplex clustering coefficients (C2). Its clustering neighbors include several gene sets that are descendants of GO:0006862 (nucleotide transport), for example, GO:0015865 (purine nucleotide transport) and GO:0051503 (adenine nucleotide transport). Rarely, a gene set may occupy high ranks in more than two categories of structural coefficients (yellow in Fig. 4.C).

The association between two structural features of gene sets is generally complicated and non-monotonous (Fig. 4.D). However, we can notice some clear associations. The gene set size significantly influences the PageRank centrality of a gene set in the Jaccard network, as a disproportionate number of gene sets in community 0 (often large gene sets) have higher-than-average PageRank centrality in the Jaccard network, resulting in a Spearman correlation of 0.83 (Fig. 4.D), which is not valid for multiplex PageRank centrality because both small and large gene sets can have high centrality depending on the context (Fig. 4.D). Clustering coefficients C1 and C2 are associated significantly in this experiment. Large gene sets tend to have lower clustering coefficients than do small gene sets. However, their clustering coefficients are unrelated to centrality, as indicated by the two peaks in their co-plots (Fig. 4.C, D). The irrelevance illustrates the fundamental difference between centrality and clustering coefficients, as the former more often represents a dominant role in a hierarchical system. The latter reflects the tightness of inter-connections among members in the neighborhood of a gene set.

3.4. Using structural coefficients to reorganize and prioritize enriched terms in GSEA

The contextual and relational attributes of gene sets obtained by studying a relevant multiplex network can be transferred to GSEA. We run a typical GSEA on a list of differentially expressed genes. The case-control groups are memory CD4 T cells and other cell populations, including naive CD4 T cells, CD8 T cells, B cells, and CD14 monocytes in pbmc3k.final (Fig. 5.A), which are the same groups of cells used for computing similarities and building the multiplex network in section 3.3. The absolute NES values in descending order are used as the criteria to rank enriched GO-BP gene sets (Fig. 5.A). The structural coefficients can be retrieved from the multiplex network built in section 3.3.

The distribution of three structural coefficients of enriched terms in GSEA is displayed in Fig. 5.B. Gene sets with extreme scores in these structural measures, such as GO:0019886, can be easily identified. The UpSet plot visualizes the exclusive intersections among the 6 collections of gene sets that consist of the top 30 gene sets ranked by different standards (Fig. 5.C). A multipartite graph shows the detailed connections between these collections of high-ranking gene sets (Fig. 5.D). High-ranking gene sets identified by the two centrality coefficients or the two clustering coefficients have significant overlapping (Fig. 5.C-D). But for the two centrality coefficients, the multiplex PageRank centrality emphasizes context more; for example, the ranks of immune-related gene sets GO:0002682 and GO:0002684 increase compared to their positions determined by the PageRank centrality in the Jaccard network. Some noticeable gene sets appear in the lists of top NES values and other structural coefficients. For example, GO:0007155 and GO:0022610 (cell/biological adhesion) have upper-level multiplex PageRank coefficients and NES values, and GO:0002377 (immunoglobulin production) has a high NES and a low participation coefficient. Despite these unique findings, some problems are apparent in this raw arrangement of enriched terms. First, many top-ranked items by the NES have redundant information related to antigen processing and presentation (Fig. 5.A, D). Second, the gene sets with top centrality coefficients mainly consist of general terms that need to be more detailed for an in-depth interpretation. For instance, GO:0002684 (positive regulation of the immune system process) relates to the difference between memory T cells and other immune cells. Still, more specific information is needed for interpreting the functional meaning of differentially expressed genes.

To solve the first problem, we note that these gene sets also have high-ranking clustering coefficients, which suggests they are located in a densely interconnected region whose members share robust relationships across layers (Fig. 5.D), and a representative term alone may suffice to identify them. In other words, the groups of interconnected gene sets in a multiplex network are likely to share common functions in context and are eligible for being compressed into singular terms. By searching consistent neighbors, the first group of interconnected gene sets are found to include GO:0019886, GO:0002478, GO:0002495, GO:0002501, GO:0002503, GO:0002504, GO:0019884, GO:0019882, GO:0048002, GO:0002396, and GO:0002399. In the GO hierarchical

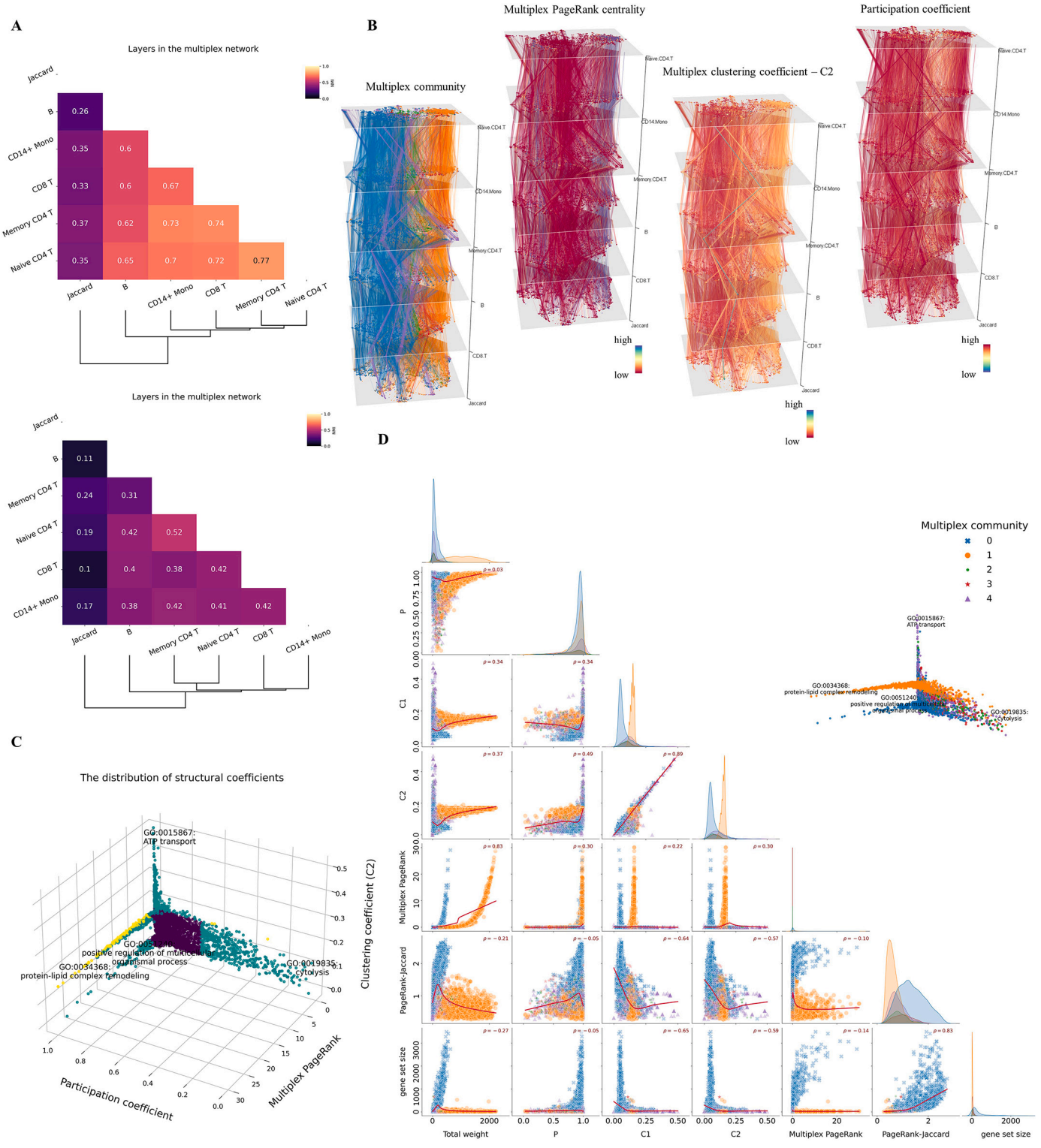

Fig. 4. Structural properties of a multiplex network of GO-BP gene sets. (A) NMI scores indicate layer-wise similarities. The dendrogram shows the hierarchical clustering of layers in the multiplex network. Upper: The threshold is 0.2 for filtering edges. Lower: The threshold is relaxed to 0.1 to preserve variations of inter-gene-set relations. (B) Visualizations of the multiplex communities and distributional patterns of structural coefficients. The coordinates of gene sets in each layer are UMAP coordinates adjusted by Grimon. Each layer corresponds to a cell population in the PBMC3K scRNAseq dataset, where the inter-gene-set similarities are measured. (C) The distribution of three primary structural coefficients of GO-BP gene sets. The ratios of centrality coefficients to the uniform mass ($\frac{1}{\text{number of gene sets}}$) rather than raw coefficients are shown. Gene sets with higher than 90th percentile C2 and multiplex PageRank coefficients or those with lower than 10th percentile participation coefficients are marked by dark cyan; gene sets in more than one category are marked by yellow. (D) Even though the Spearman correlations among attributes of gene sets may be high, the relations between structural coefficients are generally complex and non-monotonous, as indicated by the locally weighted linear regression lines (red).

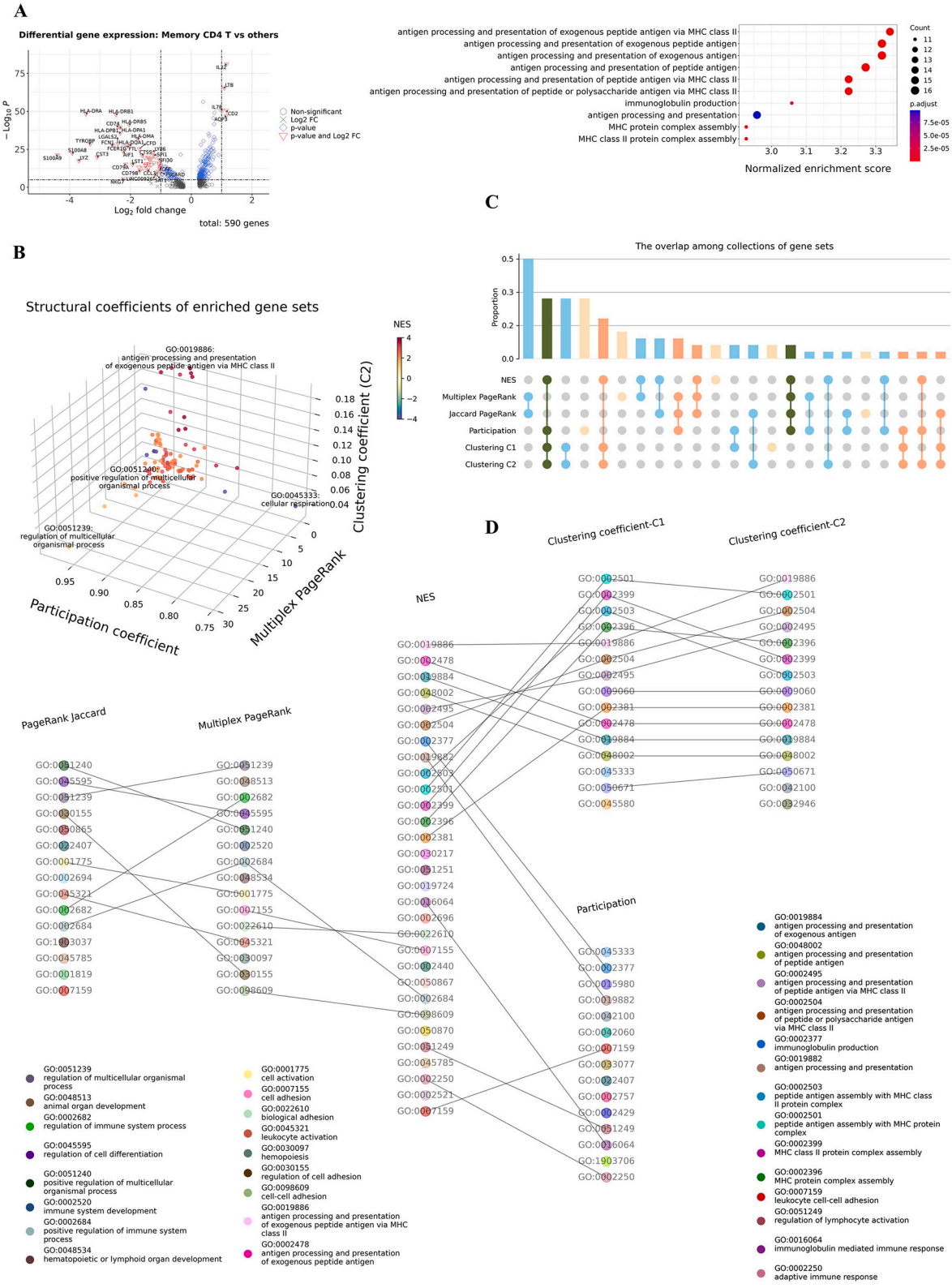

Fig. 5. Application of a gene-set multiplex network to GSEA. (A) Left: GSEA experiment on a list of differentially expressed genes. Hundreds of genes are differentially expressed in the memory CD4 T cells (mostly downregulated) compared with other cell populations. Right: Top enriched terms ranked by NES. (B) Distribution of multiplex PageRank centrality, C2 clustering coefficients, and participation coefficients of enriched gene sets. (C) The UpSet plot shows the exclusive intersections among collections of the top 30 enriched gene sets ranked by NES, multiplex/Jaccard PageRank centrality, multiplex clustering coefficients C1 and C2, and participation coefficients. (D) Raw multipartite graph of enriched terms. Many of the top NES enriched gene sets, such as GO:0019886, also have upper-level multiplex clustering coefficients. Some unique gene sets, such as GO:0002684 and GO:0002377, appear at the top of lists by NES and the multiplex PageRank centrality or participation coefficient. However, the raw arrangement of enriched terms has some problems (section 2.8). The first is information redundancy among enriched terms with high clustering coefficients. Second, gene sets with high-level multiplex centrality, such as GO:0002684, are primarily general terms with ambiguous meanings.

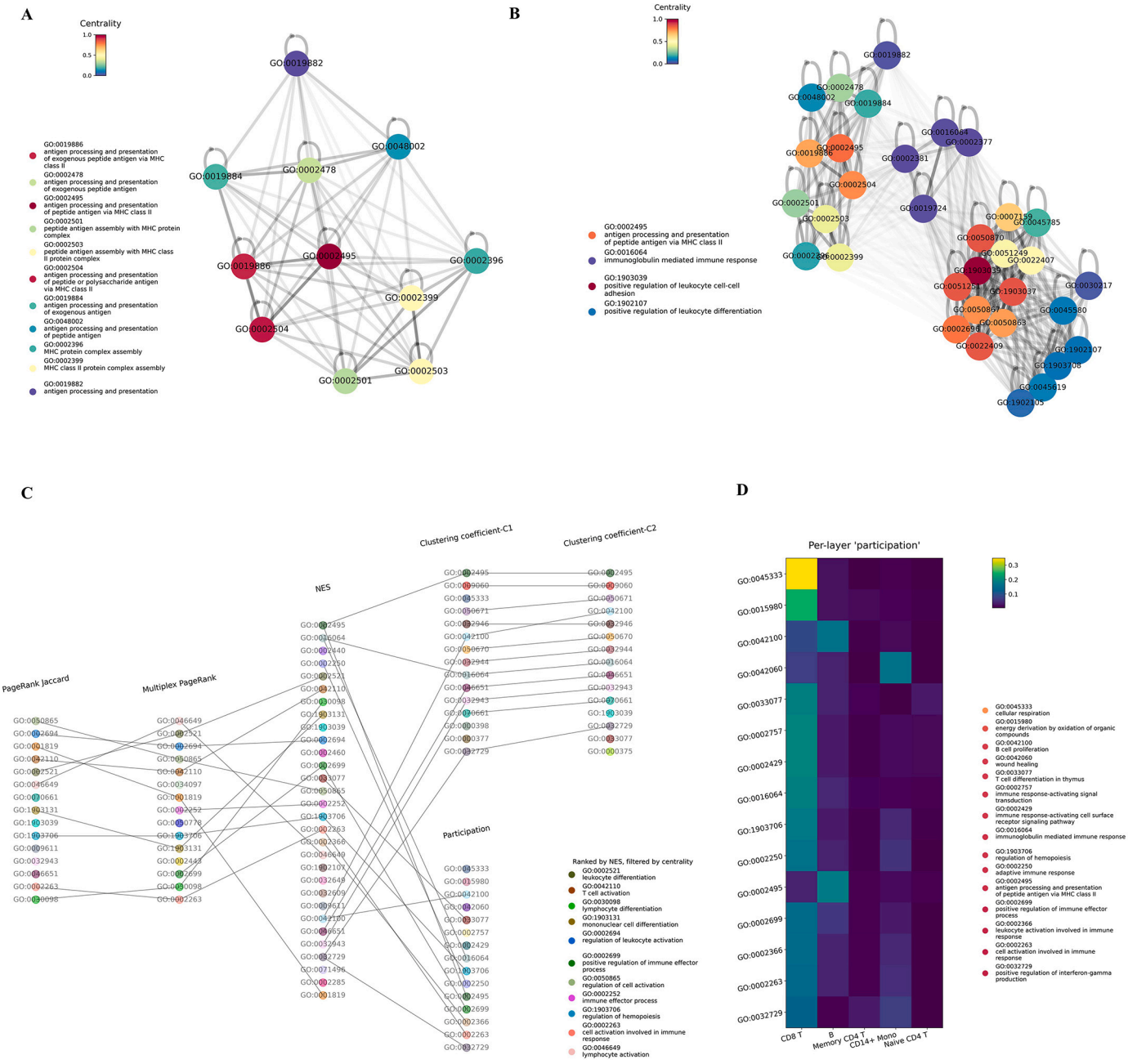

Fig. 6. Reorganizing and prioritizing enriched terms by structural coefficients. (A) One group of densely interconnected gene sets. The graph is the embedding of the average adjacency matrix of the multiplex layers. A lighter hue depicts the lines between gene sets with more minor similarities. The color of a node represents its multiplex PageRank centrality in the local network across layers. GO:0002495 exhibits the maximum multiplex centrality among the group members and thus is eligible to represent the group. (B) In the same way, three additional groups are identified, and the local centrality identifies their representative gene sets. The color of a node in this plot represents its multiplex PageRank centrality among all the groups. (C) Organization of enriched terms after compressing groups of enriched terms above and constraining the scope of multiplex centrality to a lower level. Finally, enriched terms are ranked by NES but skipped if they have low-level multiplex centrality. (D) The heatmap visualizes per-layer "participation" of enriched terms with the lowest participation coefficients. A general tendency towards more connectivity in the layer of CD8 T cells is evident.

tree, GO:0002501, GO:0002503, GO:0002399, and GO:0002396 are descendant gene sets of GO:0065003 (protein-containing complex assembly), and GO:0019884, GO:0048002, GO:0002504, GO:0002478, and GO:0019886 are descendants of GO:0019882 (antigen processing and presentation). We then compute the (local) multiplex PageRank centrality for members in the group and find that GO:0002495 (antigen processing and presentation of peptide antigen via MHC class II) has the top centrality score. From a biological standpoint, the MHC class II mediates the antigen presentation for immune cells. Thus, GO:0002495 is a reasonable representative for this gene-set group in the multiplex network (Fig. 6.A). Then, the members in the group, except the representative gene set, are removed from the enriched terms. Similarly, we find another three clusters of gene sets, though not exhaustively. Each group of gene sets is represented by a particular gene set with the highest local centrality score (Fig. 6.B). Regarding the second problem, we can adjust the window of centrality to analyze enriched terms with lower multiplex centrality. Thus, we skip the top 15 items in the list of multiplex PageRank centrality, followed by many items with more specific information. The priority of enriched terms is assigned to gene sets that appear on the ranking lists of both NES and multiplex PageRank central-

ity, such as GO:0002521 (leukocyte differentiation) and GO:0042110 (T cell activation) (Fig. 6.C), which are biologically more informative than the initial list of top enriched terms. We can gain additional insights into enriched terms by checking the gene sets' per-layer "participation" (section 2.8). For example, there is a general tendency of many gene sets for higher participation in the layer of CD8 T cells, for instance, GO:0045333 (cellular respiration) (Fig. 6.D). GO:0042100 (B cell proliferation) and GO:0002495 have more connectivities in the layer of B cells, which concords with the literature that B cells are classical antigen-presenting cells [58]. GO:0042100 (wound healing) has a layer-specificity to the network of CD14 monocytes, and there is evidence to support the role of monocytes in wound healing [59]. These examples suggest that a gene set may connect more with other gene sets in the cell populations where it functions more actively.

In summary, we propose a heuristic pipeline based on the relational attributes of gene sets to reorganize and rank gene sets (section 2.8):

1. Gene sets with high multiplex clustering coefficients (C1 and C2) are clues to searching densely interconnected groups of gene sets. Terms with maximum (local) multiplex centrality can represent other group members so that each group can be compressed into a singular item.
2. Because gene sets with high multiplex PageRank centrality are often general terms with ambiguous meanings (i.e., difficult to interpret), we can adjust the scope to focus on enriched terms with lower multiplex centrality. However, centrality is a sensible measure for prioritizing gene sets. Thus, if an enriched term appears at the top of the NES rank list but has a low multiplex centrality, we can temporarily skip it to find terms with greater centrality.
3. The participation coefficients can assist in a more comprehensive study of enriched terms, as they reflect the biased connectivity of gene sets in different contexts.

Nevertheless, the pipeline is arbitrary, and it may also make sense to follow the order of multiplex centrality or participation coefficients regardless of the NES values. Moreover, the knowledge-based structural coefficients can sometimes approximate the context-based multiplex network model results. But as we demonstrated in section 3.2, the multiplex network model can capture the biological variation of inter-gene-set relations that the knowledge-based model cannot.

4. Conclusion

In this study, we explored the new application of a classical KNN-based method to computing the similarity of gene sets. The simulation and running time comparison experiments demonstrated its superior computing performance compared to other similarity measures. Facilitated by $\mathcal{S}_{\mathbf{KNN}}$, we built a multiplex network of gene sets to capture biological variation, which better reflects the dynamic and complex nature of gene-set interactions. The structural coefficients of gene sets extract relational characteristics of gene sets in their community, enabling a more thorough understanding of their roles. When researchers perform GSEA, they often have to make a decision on the criteria to prioritize enriched terms to identify the most important ones. Currently, the criteria are monotonous because only the effect size/significance of enrichment is emphasized and gene sets' contextual and relational features are often ignored. However, we can take advantage of these structural coefficients to identify a group of comprehensive criteria to reorganize and prioritize enriched terms, which may enhance the interpretability of GSEA.

Funding

This work was funded by a KAKENHI Grant-in-Aid from the Japan Society for the Promotion of Science (Grant No. 21K21316).

Code availability

The code used in this study is available at https://github.com/flyeous/gene-set-multiplex-network.

CRediT authorship contribution statement

Cheng Zheng: Conceptualization, Formal analysis, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Man Wang:** Conceptualization, Methodology. **Ryo Yamada:** Conceptualization, Methodology. **Daigo Okada:** Conceptualization, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank FORTE Inc. (https://www.forte-science.co.jp/) for proofreading the English in this manuscript.

The graphical abstract and Fig. 1 in this paper were created with BioRender.com.

Appendix A. Proof of equations in $\mathbf{C}_{i,1}, \mathbf{C}_{i,2}$

Recall that $\mathcal{A}^{[\alpha]}, \mathcal{A}^{[\alpha']}, \mathcal{A}^{[\alpha'']}$ are adjacency matrices for three different weighted and directed layers ($\mathcal{G}_\alpha, \mathcal{G}_{\alpha'}, \mathcal{G}_{\alpha''}$, respectively) in a multiplex network $\mathcal{M} = (\mathrm{Y}, \vec{\mathcal{G}}, \mathcal{C})$ with non-negative, normalized edge weights, and no self-loops. We can write $\mathcal{A}'^{[\alpha]} = \mathcal{A}^{[\alpha]} + \mathcal{A}^{[\alpha]^T}$ s.t. $\mathcal{A}'^{[\alpha]}_{ij} = \mathcal{A}^{[\alpha]}_{ij} + \mathcal{A}^{[\alpha]}_{ji}$ and $\mathcal{A}'^{[\alpha]} = \mathcal{A}'^{[\alpha]^T}$. We treat $\mathcal{A}'^{[\alpha']}$ and $\mathcal{A}'^{[\alpha'']}$ similarly in $\mathbf{C}_{i,1}, \mathbf{C}_{i,2}$.

For the numerator of $\mathbf{C}_{i,1}$, it suffices to show $\forall \alpha, \alpha' \in \mathbf{Y}$ and $\alpha \neq \alpha'$

$$\begin{aligned}
\sum_{j\neq i, m\neq i, j<m} \mathcal{A}'^{[\alpha]}_{ij}\mathcal{A}'^{[\alpha']}_{jm}\mathcal{A}'^{[\alpha]}_{mi} &= \frac{1}{2}\sum_{j\neq i, m\neq i, j\neq m} \mathcal{A}'^{[\alpha]}_{ij}\mathcal{A}'^{[\alpha']}_{jm}\mathcal{A}'^{[\alpha]}_{mi} \quad \text{(by the symmetry of } \mathcal{A}'^{[\alpha]}, \mathcal{A}'^{[\alpha']}\text{)} \\
&= \frac{1}{2}\sum_{m\neq i}\Big(\sum_{j\neq m, j\neq i} \mathcal{A}'^{[\alpha]}_{ij}\mathcal{A}'^{[\alpha']}_{jm}\Big)\mathcal{A}'^{[\alpha]}_{mi} \quad \text{(arrangement of terms)} \\
&= \frac{1}{2}\sum_{m\neq i}(\mathcal{A}'^{[\alpha]}\mathcal{A}'^{[\alpha']})_{im}\mathcal{A}'^{[\alpha]}_{mi} \quad (\mathcal{A}'^{[\alpha]}_{ii}, \mathcal{A}'^{[\alpha']}_{mm} = 0) \\
&= \frac{1}{2}(\mathcal{A}'^{[\alpha]}\mathcal{A}'^{[\alpha']}\mathcal{A}'^{[\alpha]})_{ii} = \frac{1}{2}\mathbf{Diag}(\mathcal{A}'^{[\alpha]}\mathcal{A}'^{[\alpha']}\mathcal{A}'^{[\alpha]})_{ii}
\end{aligned}$$

For the denominator of $\mathbf{C}_{i,1}$, it suffices to show $\forall \alpha \in \mathbf{Y}$

$$\begin{aligned}
\sum_{j\neq i, m\neq i, j<m} \mathcal{A}'^{[\alpha]}_{ij}\mathcal{A}'^{[\alpha]}_{mi} &= \frac{1}{2}\sum_{j\neq i, m\neq i, j\neq m} \mathcal{A}'^{[\alpha]}_{ij}\mathcal{A}'^{[\alpha]}_{mi} \quad \text{(by the symmetry of } \mathcal{A}'^{[\alpha]}\text{)} \\
&= \frac{1}{2}\sum_{m\neq i}\Big(\sum_{j\neq m, j\neq i} \mathcal{A}'^{[\alpha]}_{ij}\Big)\mathcal{A}'^{[\alpha]}_{mi} \quad \text{(arrangement of terms)} \\
&= \frac{1}{2}\sum_{m\neq i}((\mathcal{A}'^{[\alpha]}\mathbf{1})_i - \mathcal{A}'^{[\alpha]}_{im})\mathcal{A}'^{[\alpha]}_{mi} \quad (\mathcal{A}'^{[\alpha]}_{ii} = 0) \\
&= \frac{1}{2}((\mathcal{A}'^{[\alpha]}\mathbf{1})_i(\mathbf{1}^T\mathcal{A}'^{[\alpha]})_i - \sum_{m\neq i}\mathcal{A}'^{[\alpha]}_{im}\mathcal{A}'^{[\alpha]}_{mi}) \quad (\mathcal{A}'^{[\alpha]}_{mm} = 0) \\
&= \frac{1}{2}((\mathcal{A}'^{[\alpha]}\mathbf{1})^{\circ 2} - \mathcal{A}'^{[\alpha]^{\circ 2}}\mathbf{1})_i \quad \text{(by the symmetry of } \mathcal{A}'^{[\alpha]}\text{)}
\end{aligned}$$

In a similar way, $\sum_{j\neq i, m\neq i, j\neq m} \mathcal{A}'^{[\alpha]}_{ij}\mathcal{A}'^{[\alpha'']}_{jm}\mathcal{A}'^{[\alpha']}_{mi} = \mathbf{Diag}(\mathcal{A}'^{[\alpha]}\mathcal{A}'^{[\alpha'']}\mathcal{A}'^{[\alpha']})_{ii}$ $\forall \alpha, \alpha', \alpha'' \in \mathbf{Y}$ and $\alpha \neq \alpha', \alpha'' \neq \alpha, \alpha'$ in the numerator of $\mathbf{C}_{i,2}$. For the denominator of $\mathbf{C}_{i,2}$, it suffices to show $\forall \alpha, \alpha' \in \mathbf{Y}$ and $\alpha \neq \alpha'$

$$\begin{aligned}
\sum_{j\neq i, m\neq i, j\neq m} \mathcal{A}'^{[\alpha]}_{ij}\mathcal{A}'^{[\alpha']}_{mi} &= \sum_{m\neq i}\Big(\sum_{j\neq i, j\neq m} \mathcal{A}'^{[\alpha]}_{ij}\Big)\mathcal{A}'^{[\alpha']}_{mi} \quad \text{(arrangement of terms)} \\
&= \sum_{m\neq i}((\mathcal{A}'^{[\alpha]}\mathbf{1})_i - \mathcal{A}'^{[\alpha]}_{im})\mathcal{A}'^{[\alpha']}_{mi} \quad (\mathcal{A}'^{[\alpha]}_{ii} = 0) \\
&= (\mathcal{A}'^{[\alpha]}\mathbf{1})_i(\mathbf{1}^T\mathcal{A}'^{[\alpha']})_i - \sum_{m\neq i}\mathcal{A}'^{[\alpha]}_{im}\mathcal{A}'^{[\alpha']}_{mi} \quad (\mathcal{A}'^{[\alpha']}_{mm} = 0) \\
&= ((\mathcal{A}'^{[\alpha]}\mathbf{1}) \odot (\mathcal{A}'^{[\alpha']}\mathbf{1}) - (\mathcal{A}'^{[\alpha]} \odot \mathcal{A}'^{[\alpha']})\mathbf{1})_i \quad \text{(by the symmetry of } \mathcal{A}'^{[\alpha']}\text{)}
\end{aligned}$$

References

[1] Maleki F, Ovens K, Hogan DJ, Kusalik AJ. Gene set analysis: challenges, opportunities, and future research. Front Genet 2020;11:654.
[2] Kaizer EC, Glaser CL, Chaussabel D, Banchereau J, Pascual V, White PC. Gene expression in peripheral blood mononuclear cells from children with diabetes. J Clin Endocrinol Metab 2007;92(9):3705–11.
[3] Godec J, Tan Y, Liberzon A, Tamayo P, Bhattacharya S, Butte AJ, et al. Compendium of immune signatures identifies conserved and species-specific biology in response to inflammation. Immunity 2016;44(1):194–206.
[4] Keseler IM, Skrzypek M, Weerasinghe D, Chen AY, Fulcher C, Li G-W, et al. Curation accuracy of model organism databases. Database 2014:2014.
[5] Gillis J, Pavlidis P. Assessing identity, redundancy and confounds in gene ontology annotations over time. Bioinformatics 2013;29(4):476–82.
[6] Stoney RA, Schwartz J-M, Robertson DL, Nenadic G. Using set theory to reduce redundancy in pathway sets. BMC Bioinform 2018;19(1):1–11.
[7] Wang G, Oh D-H, Dassanayake M. Gomcl: a toolkit to cluster, evaluate, and extract non-redundant associations of gene ontology-based functions. BMC Bioinform 2020;21:1–9.
[8] Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS ONE 2010;5(11):e13984.
[9] Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 2009;25(8):1091–3.
[10] Rahmatallah Y, Emmert-Streib F, Glazko G. Gene sets net correlations analysis (gsnca): a multivariate differential coexpression test for gene sets. Bioinformatics 2014;30(3):360–8.
[11] Tsai C-A, Chen JJ. Gene set correlation analysis and visualization using gene expression data. Curr Bioinform 2021;16(3):406–21.
[12] Josse J, Holmes S. Measuring multivariate association and beyond. Stat Surv 2016;10:132.
[13] Bianconi G. Multilayer networks: structure and function. Oxford University Press; 2018.
[14] Ding H, Yang Y, Xue Y, Seninge L, Gong H, Safavi R, et al. Prioritizing transcriptional factors in gene regulatory networks with pagerank. iScience 2021;24(1):102017.
[15] Battiston F, Nicosia V, Latora V. The new challenges of multiplex networks: measures and models. Eur Phys J Spec Top 2017;226(3):401–16.
[16] Battiston F, Nicosia V, Latora V. Structural measures for multiplex networks. Phys Rev E 2014;89(3):032804.
[17] Halu A, Mondragón RJ, Panzarasa P, Bianconi G. Multiplex pagerank. PLoS ONE 2013;8(10):e78293.
[18] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci 2005;102(43):15545–50.
[19] Hung J-H, Yang T-H, Hu Z, Weng Z, DeLisi C. Gene set enrichment analysis: performance evaluation and usage guidelines. Brief Bioinform 2012;13(3):281–91.
[20] Zyla J, Marczyk M, Weiner J, Polanska J. Ranking metrics in gene set enrichment analysis: do they matter? BMC Bioinform 2017;18:1–12.
[21] Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g: profiler, gsea, cytoscape and enrichmentmap. Nat Protoc 2019;14(2):482–517.
[22] Friedman JH, Rafsky LC. Graph-theoretic measures of multivariate association and prediction. Ann Stat 1983:377–91.
[23] S. Lab. pbmc3k.SeuratData: 3k PBMCs from 10X genomics, r package version 3.1.4; 2020.
[24] Domínguez Conde C, Xu C, Jarvis L, Rainbow D, Wells S, Gomes T, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science 2022;376(6594):eabl5197.
[25] Carlson M. org.Hs.eg.db: genome wide annotation for human, r package version 3.14.0; 2021.
[26] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterprofiler 4.0: a universal enrichment tool for interpreting omics data. Innovation 2021;2(3).
[27] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25(1):25–9.
[28] Consortium TGO. The gene ontology resource: enriching a gold mine. Nucleic Acids Res 2021;49(D1):D325–34.
[29] Carlson M. GO.db: a set of annotation maps describing the entire gene ontology, r package version 3.14.0; 2021.
[30] Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (msigdb) 3.0. Bioinformatics 2011;27(12):1739–40.
[31] Pękalska E, Harol A, Duin RP, Spillmann B, Bunke H. Non-Euclidean or non-metric measures can be informative. In: Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR). Springer; 2006. p. 871–80.
[32] Levandowsky M, Winter D. Distance between sets. Nature 1971;234(5323):34–5.
[33] Smilde AK, Kiers HA, Bijlsma S, Rubingh C, Van Erk M. Matrix correlations for high-dimensional data: the modified rv-coefficient. Bioinformatics 2009;25(3):401–5.
[34] Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. Ann Stat 2007;35(6):2769–94.
[35] Diniz-Filho JAF, Soares TN, Lima JS, Dobrovolski R, Landeiro VL, Telles MPdC, et al. Mantel test in population genetics. Genet Mol Biol 2013;36:475–85.
[36] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[37] Tomic O, Graff T, Liland KH, Næs T. hoggorm: a python library for explorative multivariate statistics. J Open Sour Softw 2019;4(39). https://doi.org/10.21105/joss.00980.
[38] Carreño CR. dcor: distance correlation and related e-statistics in python. Available from: https://github.com/vnmabus/dcor, 2017.
[39] T. scikit-bio development team. scikit-bio: a bioinformatics library for data scientists, students, and developers. Available from: http://scikit-bio.org, 2020.
[40] Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg 2018;126(5):1763–8.
[41] Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports 2019;9(1):5233.
[42] Waskom ML. seaborn: statistical data visualization. J Open Sour Softw 2021;6(60):3021. https://doi.org/10.21105/joss.03021.
[43] Hunter JD. Matplotlib: a 2d graphics environment. Comput Sci Eng 2007;9(3):90–5. https://doi.org/10.1109/MCSE.2007.55.
[44] Klopfenstein D, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, et al. Goatools: a python library for gene ontology analyses. Sci Rep 2018;8(1):10872.
[45] McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction. Available from: arXiv:1802.03426, 2018.
[46] Kanai M, Maeda Y, Okada Y. Grimon: graphical interface to visualize multi-omics networks. Bioinformatics 2018;34(22):3934–6.
[47] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature 1998;393(6684):440–2.
[48] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005;4(1).
[49] Holme P, Park SM, Kim BJ, Edling CR. Korean university life in a network perspective: dynamics of a large affiliation network. Phys A, Stat Mech Appl 2007;373:821–30.
[50] Miyajima K, Sakuragawa T. Continuous and robust clustering coefficients for weighted and directed networks. Available from: arXiv:1412.0059, 2014.
[51] Fagiolo G. Clustering in complex directed networks. Phys Rev E 2007;76(2):026107.
[52] Clemente GP, Grassi R. Directed clustering in weighted networks: a new perspective. Chaos Solitons Fractals 2018;107:26–38.
[53] Bonald T, de Lara N, Lutz Q, Charpentier B. Scikit-network: graph analysis in python. J Mach Learn Res 2020;21(185):1–6. Available from: http://jmlr.org/papers/v21/20-412.html.
[54] Hagberg A, Swart P, Chult DS. Exploring network structure, dynamics, and function using networkx. Tech. rep. Los Alamos, NM (United States): Los Alamos National Lab. (LANL); 2008.
[55] Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. Scenic: single-cell regulatory network inference and clustering. Nat Methods 2017;14(11):1083–6.
[56] Doulatov S, Notta F, Laurenti E, Dick JE. Hematopoiesis: a human perspective. Cell Stem Cell 2012;10(2):120–36.
[57] Vansarla G, Håkansson AP, Bergenfelz C. Hamlet a human milk protein-lipid complex induces a pro-inflammatory phenotype of myeloid cells. Eur J Immunol 2021;51(4):965–77.
[58] Giles JR, Kashgarian M, Koni PA, Shlomchik MJ. B cell–specific mhc class ii deletion reveals multiple nonredundant roles for b cell antigen presentation in murine lupus. J Immunol 2015;195(6):2571–9.
[59] Rehak L, Giurato L, Meloni M, Panunzi A, Manti GM, Uccioli L. The immune-centric revolution in the diabetic foot: monocytes and lymphocytes role in wound healing and tissue regeneration—a narrative review. J Clin Med 2022;11(3):889.