Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Sparse Generalized Tensor Canonical Correlation Analysis for Multi-Omics Network Inference

Weixuan Liu WEIXUAN.LIU@CUANSCHUTZ.EDU
Department of Biotatistics and Informatics
University of Colorado Anschutz Medical Campus
Aurora, CO 80045, USA

Farnoush Banaei-Kashani FARNOUSH.BANAEI-KASHANI@UCDENVER.EDU
Department of Computer Science and Engineering
University of Colorado Denver
Denver, CO 80204, USA

Russell P Bowler BOWLERR@NJHEALTH.ORG
Division of Pulmonary Medicine, Department of Medicineh
National Jewish Health
Denver, CO 80206, USA

Katerina J Kechris KATERINA.KECHRIS@CUANSCHUTZ.EDU
Department of Biotatistics and Informatics
University of Colorado Anschutz Medical Campus
Aurora, CO 80045, USA

Editor:

Abstract

Multiple omics (genomics, proteomics, etc.) profiles are commonly generated to gain insight into a disease or physiological system. Constructing multi-omics networks with respect to the trait(s) of interest provides an opportunity to understand relationships between molecular features but integration is challenging due to multiple data sets with high dimensionality. Methods like sparse multiple canonical correlation network analysis (SmCCNet) construct multi-omics network modules by integrating one or two omics types and a single trait of interest. However, these types of methods may be limited due to (1) lack of accounting for higher-order correlations existing among features, (2) computational inefficiency when extending to more than two omics data when using a penalty term-based sparsity method, and (3) lack of flexibility for focusing on specific correlations (e.g., omics to phenotype correlation versus omics-to-omics correlations). In this work, we have developed a novel multi-omics network analysis pipeline called Sparse Generalized Tensor Canonical Correlation Analysis Network Inference (SGTCCA-Net) that can effectively overcome these limitations. Simulation and real data experiments demonstrate the effectiveness of our novel method for inferring omics networks and features of interest.

Keywords: Multi-omics, Network Inference, Tensor

1. Introduction

1.1 Multi-Omics Data Integration

Integrating multiple datasets with the same set of subjects has been actively explored in the machine learning literature, and analysis of this kind is referred to as multi-view machine learning (Xu et al., 2013). It has a wide range of applications including clustering of subjects based on the consensus of different views (Bickel and Scheffer, 2004) or performing image annotation (Kalayeh et al., 2014). In addition, it has been explored in the biomedical domains (Rappoport and Shamir, 2018; Nicora et al., 2020). Recent advances in biomedical technologies now offer the opportunity to generate high-throughput data at the molecular level, such as the genome, transcriptome, proteome, and metabolome. Investigators are now generating multiple omics data and the collection is called multi-omics data (Subramanian et al., 2020). Traditionally, each omics data is analyzed separately from other omics data (Gilchrist et al., 2006; Monteiro et al., 2013), which may lose information about the connection between omics data. Integrating multi-omics data can help uncover biological mechanisms and interactions at the molecular level. Multi-omics data may be integrated for different purposes, including disease subtyping (Menyhárt and Győrffy, 2021; Duan et al., 2021), variable selection (Lin and Lane, 2017; Pierre-Jean et al., 2020), network analysis (Jiang et al., 2019; Zhou et al., 2020), and biomarker prediction (Wu et al., 2019; Yan et al., 2018; Subramanian et al., 2020).

1.2 Multi-View Canonical Correlation Analysis

Dimensional reduction is one of the more common goals of multi-view machine learning. For example, Liu et al. (2013) proposed to implement non-negative matrix factorization to cluster multi-view data, and Zhang et al. (2016) developed multi-view co-reduction to preserve the within-view locality and between-view consistency in the lower dimensional embedding of each view.

Other types of dimension reduction methods for multi-view data are based on Canonical Correlation Analysis (CCA) developed by Hotelling (1992), which seeks to find the linear combination (canonical weight) that maximizes the correlation between 2 sets of data. Usually, there is more than one solution, which may be referred to as a canonical weight matrix, and thus it is commonly used for dimension reduction by projecting the original data into a shared lower dimensional space between two views (or omics data) with canonical weight matrix. Given two data sets X_1 and X_2, the optimization function is:

$$w_1, w_2 = \arg\max_{w_1,w_2} \frac{w_1^T X_1^T X_2 w_2}{\sqrt{w_1^T X_1^T X_1 w_1 w_2^T X_2^T X_2 w_2}} \tag{1}$$

However, this formulation only considers 2 sets of data and is not generalized to multiple data. Witten and Tibshirani (2009) developed a multiple canonical correlation method that maximizes the summation of pairwise canonical correlations. Suppose there are $k = 1, \ldots, K$ views, then it can be formulated as follow:

$$\max_{w_1,\ldots,w_k} \sum_{i<j} w_i^T X_i^T X_j w_j \quad s.t. w_k^T X_k^T X_k w_k = 1 \tag{2}$$

They further added a penalty term to the formulation above to achieve sparsity, which is called Sparse Multiple Canonical Correlation Analysis (SmCCA). However, Hu et al. (2017) shows that SmCCA results in a problem of "'unfair combination of pairwise covariances" since most of the SmCCA methods maximize the pairwise canonical covariance instead of correlation, and the scale of covariance may be problematic when summing over all pairwise canonical covariance, thus an adaptive version of SmCCA is proposed to effectively solve this problem.

A particular variant of the multiple canonical correlation analysis extends the pairwise relationship to a higher-order relationship by maximizing the tensor canonical correlation (Luo et al., 2015), which captures higher-order correlation among multi-view data and projects the multi-view data into a shared lower dimensional embedding. A further extension to this tensor canonical correlation analysis (Wong et al. (2020)) is to combine deep learning with tensor canonical correlation analysis to maximize the higher-order correlation between views in the nonlinear lower dimensional space for dimensional reduction.

1.3 Using Canonical Correlation Analysis to Identify Networks Associated with a Quantitative Trait

Network analysis for molecular profiles is often used with dimension reduction to identify and visualize connections between features (Jiang et al., 2019; Zhou et al., 2020), in particular, the goal may be to infer underlying true biological processes or relationships in the observed data. In addition to canonical correlation analysis, regression is another approach that is implemented for multi-omics network inference, particularly for gene regulatory network (Haury et al., 2012; Huynh-Thu et al., 2010). However, most methods do not incorporate an outcome or phenotype in the form of a quantitative trait. Canonical correlation analysis-based methods can accomplish this goal by expanding CCA to incorporate a phenotype (Y). SmCCNet, proposed by Shi et al. (2019), partition omics features into different network modules while considering the phenotype(s) of interest. It incorporates the idea of a scaled version of SmCCA, which has the optimization problem in the following form:

$$\begin{aligned} \arg\max_{w_1,w_2} & \; aw_1^T X_1^T X_2 w_2 + bw_1^T X_1^T Y + cw_2^T X_2^T Y \\ & \text{s.t.} ||w_i||^2 = 1, P_i(w_i) = c_i, i = 1, 2, \end{aligned} \tag{3}$$

where “a”, “b”, and “c” are scaling factors to put more importance on pairwise combinations of views. For example, one may want to increase “b” and “c” to increase the influence on the phenotype for determining the canonical weights, and $P(\cdot)$ is the lasso penalty term for sparsity (Tibshirani, 1996), but other types of penalties can be used as well. The optimal penalty parameter is selected through k-fold cross-validation. After that, the canonical weights can be extracted based on the objective function above to construct an adjacency matrix. Finally, to construct a network, hierarchical clustering can be implemented afterward to extract multiple multi-omics network modules. This method has been applied in various contexts including identifying phenotype-specific miRNA-mRNA and proteomics-metabolomics correlations and networks (Mastej et al., 2020). However, the existing SmCCNet method can only be adapted to two molecular profiles plus one single phenotype. When extending

it to 3 or more omics data types, it can become computationally expensive due to the 5-fold cross-validation step to find the optimal penalty parameter for each omics data set.

Another multi-omics analysis and network inference method are called Data Integration Analysis for Biomarker discovery using Latent variable approaches for Omics studies (DIABLO) (Singh et al., 2019), which has a similar formulation to the sparse multiple canonical correlation analysis problems. It optimizes the following equation:

$$\begin{aligned} &\arg\max_{w_1,w_2} \sum_{i,j} c_{ij} cov(X_i w_i, X_j w_j) \\ &\text{s.t.} ||w_i||^2 = 1, ||w_i||_1 < \lambda_i, \end{aligned} \tag{4}$$

where c_{ij} is the indicator of whether to include the correlation structure between views i and j. In the presence of phenotype data, DIABLO simply treats it as another data view X. In the actual implementation of the DIABLO, there is no lasso penalty, but the user can choose how many features to be included as non-sparse for each data view.

Even though the formulation of SmCCNet and DIABLO are similar, they have some differences: (1) SmCCNet allows users to prioritize certain correlation pairs (through the scaling factor); (2) DIABLO allows for any number of molecular profiles and multiple phenotypes; (3) adjacency matrix from SmCCNet is aggregated through canonical weights obtained from subsampling the data multiple times for a more robust solution, while the adjacency matrix from DIABLO is the subset of correlation matrix with only molecular features selected by the canonical correlation analysis.

1.4 Sparse Generalized Tensor Canonical Correlation Analysis

As reviewed the dimension reduction methods based on canonical correlation analysis only consider the summation of all pairwise correlation, including methods like SmCCNet and DIABLO. However, for multi-omics data with more than two molecular profiles, the correlation structure is higher-order rather than pairwise, where higher-order correlation is defined as the simultaneous correlation among 3 or more features. This type of higher-order correlation can be captured by pairwise correlation if and only if all pairwise correlations are strong. However, in multi-omics data, it may not be the case that higher-order correlations also have strong pairwise correlations. In addition, even though tensor canonical correlation analysis (TCCA) is able to capture higher-order correlations, it only allows for one type of correlation, but not all types of lower-order correlation structures. In addition, sparsity is not considered in TCCA, which prevents the model from revealing the structure of the multi-omics correlation between molecular features. Therefore, in this work, we present a new method for identifying multi-omics phenotype specifics network that is based on TCCA, includes both higher and lower order correlations, and incorporates sparsity. We call this method Sparse Generalized Tensor Canonical Correlation Analysis and combine it with network analysis (SGTCCA-Net). We test this method on simulations of multi-view data along with a phenotype, in addition to applying it to multi-omics data for studying chronic obstructive pulmonary disease.

2. Methods

2.1 Pipeline Workflow

The general pipeline workflow of SGTCCA-Net is similar to SmCCNet with three major changes: (1) sparse multiple canonical correlation analysis in the first step is replaced by our novel generalized tensor canonical correlation analysis, (2) subsampling is biased with a higher probability to select features with higher density value in covariance tensors/matrices to ensure sparsity, which is also referred to as Turbo-SMT algorithm (introduce later), and (3) hierarchical clustering is replaced by a novel network trimming algorithm to eliminate noisy nodes. The details of this new pipeline are demonstrated in Figure 1. This end-to-end pipelines- inputs the multi-omics data $X_1, X_2, ..., X_k \in \mathbb{R}^{N\times d_i}, i = 1, 2, ..., k$, and outputs subnetwork adjacency matrix $M_s^{(sub)} \in \mathbb{R}^{(\sum_{j=1}^{k} p_j)\times(\sum_{j=1}^{k} p_j)}$, where $p_j \leq d_j$ for all $j = 1, 2, ..., k$.

2.2 Generalized Tensor Canonical Correlation Analysis

2.2.1 Tensor Canonical Correlation Analysis (TCCA)

Luo et al. (2015) developed the Tensor canonical correlation analysis to directly extract higher-order correlation between more than two sets of data. Let $z_1, z_2, ..., z_k \in \mathbb{R}^{n\times 1}$ be k vectors with the same length, and they are centered, then the higher-order covariance between these vectors can be defined as:

$$\rho(z_1, z_2, ..., z_k) = \frac{1}{n}(z_1 z_2 ... z_k)^T \mathbf{1}, \tag{5}$$

where $z_1 z_2 ... z_k$ is the element-wise multiplication between k vectors. Suppose there are k views of the multi-view data, denoted by $X_p \in \mathbb{R}^{N\times d_p}, p = 1, 2, .., k$, which are centered with mean 0. for each view, there are N observations and d_p features, then the covariance tensor $C_{1,2,...,k} \in \mathbb{R}^{d_1\times d_2\times ...\times d_k}$ for k views can be denoted by:

$$C_{1,2,...,k} = \frac{1}{N}\sum_{j=1}^{N} x_{1j} \circ x_{2j} \circ ... \circ x_{kj}, \tag{6}$$

where $\circ$ denotes the outer product. If the canonical weights for each view are $w_i, i = 1, 2, .., k$. It is simple to show that each entry of $C_{1,2,...,k}$, denoted by $C_{1,2,...,k}(j_1, j_2, ..., j_k)$, can be calculated through equation (5):

$$C_{1,2,...,k}(j_1, j_2, ..., j_k) = \frac{1}{n}[X_1(j_1)X_2(j_2)...X_1(j_1)]^T \mathbf{1}, \tag{7}$$

where $X_i(j_i)$ stands for the j_ith feature of the ith view. This equivalency indicates the correctness of higher-order covariance calculation through equation (6). Then the tensor canonical correlation analysis (TCCA) is formulated as follows:

$$\begin{aligned} \arg\max_{w_i} \rho = C_{1,2,...,k} \times_1 w_1^T \times_2 w_2^T \times ... \times_k w_k^T \\ \text{s.t.} w_i^T C_{ii} w_i = 1, i = 1, 2, ..., k, \end{aligned} \tag{8}$$

where C_{ii} is the covariance matrix in the ith view, the i-mode product of $C_{1,2,...,k}$ and w_i^T, denoted by $C_{1,2,...,k} \times_i w_i$ is defined as follow:

$$(C_{1,2,...,k} \times_i w_i^T)(j_1, j_2, ..., j_{p-1}, 1, j_{p+1}, ..., j_k) = C_{1,2,...,k}(j_1, j_2, ..., j_k) \cdot w_i^T \tag{9}$$

To get the optimal canonical weight w_i, a transformations need to be made on $C_{1,2,...,k}$ so that the C_{ii} in the constraint in the equation above can be eliminated. Otherwise, for high-dimensional data, the constraint would make the optimization computationally intensive. Therefore, an alternative is to directly optimize the canonical covariance by removing C_{ii} from the constraint:

$$\begin{aligned} \arg\max_{h_i} \rho = C_{1,2,...,k} \times_1 h_1^T \times_2 h_2^T \times ... \times_k h_k^T \\ \text{s.t.} h_i^T h_i = 1, i = 1, 2, ..., k, \end{aligned} \tag{10}$$

where $h_i, i = 1, 2, ..., k$ are canonical weight vectors. It has been shown that the optimization problem above is equivalent to the following form (De Lathauwer et al., 2000):

$$\begin{aligned} \min_{\rho, h_1, h_2, ..., h_k} \quad & ||C_{1,2,...,k} - \hat{C}_{1,2,...,k}||_F^2 \\ \text{s.t. } & \hat{C}_{1,2,...,k} = \rho \cdot h_1 \circ h_2 \circ ... \circ h_k \end{aligned} \tag{11}$$

This results in being able to solve for the canonical weights, using either a gradient-based method or alternating least squared.

2.2.2 Generalized Tensor Canonical Correlation Analysis (GTCCA) Optimization Problem

In general, there are two major problems of TCCA: (1) the way the higher-order correlation is calculated will suffer from true effect cancellation with an odd number of views, and (2) the direction of correlation is meaningless for the higher-order setting. An example of the potential effect cancellation when calculating the higher-order covariance between an odd number of vectors is that. for 3 identical vectors with the element (-2,-1,0,1,2), the original calculation gives a higher-order covariance of 0, while theoretically, they are perfectly correlated. In addition, taking the absolute value of the higher-order correlation eliminates the trouble of interpreting the direction of the higher-order correlation. For instance, a positive correlation between 4 vectors may represent all positive correlation, or 2 vectors with positive correlation and 2 with negative correlation, which is difficult to interpret. Therefore, we modified how the higher-order correlation is calculated in Generalized Tensor Canonical Correlation Analysis (GTCCA). Let $z_1, z_2, ..., z_k \in \mathbb{R}^{n \times 1}$ be k centered vectors with the same length, then the higher-order correlation between these vectors can be defined as:

$$|\rho(z_1, z_2, ..., z_k)| = \begin{cases} |\frac{1}{n}(z_1 z_2 ... z_k)^T \mathbf{1}|, \\ \qquad\qquad \text{if } k = 2m, m \in \mathbb{N} \\ \frac{1}{n \cdot k} \sum_{i=1}^{k} |(z_1 ... |z_i| ... z_k)^T \mathbf{1}|, \\ \qquad\qquad \text{if } k = 2m+1, m \in \mathbb{N}, \end{cases} \tag{12}$$

where $\mathbf{1}$ is the all-one vector with a length of n. Based on the new definition of higher-order correlation, the covariance tensor between $X_1, X_2, ..., X_k$ can be formulated as:

$$C_{1,2,...,k} = \begin{cases} |\frac{1}{N}\sum_{j=1}^{N} x_{1j} \circ x_{2j} \circ ... \circ x_{kj}|, \\ \qquad\qquad \text{if } k = 2m, m \in \mathbb{N} \\ \frac{1}{k}\sum_{i=1}^{k} |C_{1,2,...,|i|,...,k}|, \\ \qquad\qquad \text{if } k = 2m+1, m \in \mathbb{N} \end{cases} \tag{13}$$

Where

$$C_{1,2,...,|i|,...,k} = \frac{1}{N}\sum_{j=1}^{N} x_{1j} \circ x_{2j} \circ ... \circ |x_{ij}| \circ ... \circ x_{kj} \tag{14}$$

Theorem 1 *If a covariance tensor $C_{1,2,...,k}$ between $X_1, X_2, ..., X_k$ is calculated by equation (13), then each entry of $C_{1,2,...,k}$, denoted by $C_{1,2,...,k}(j_1, j_2, ..., j_k)$, is equivalent to:*

$$|\rho(X_1(j_1), X_2(j_2), ..., X_k(j_k))| = \begin{cases} |\frac{1}{n}(X_1(j_1)X_2(j_2)...X_k(j_k))^T \mathbf{1}|, \\ \qquad\qquad \text{if } k = 2m, m \in \mathbb{N} \\ \frac{1}{n \cdot k}\sum_{i=1}^{k} |(X_1(j_1)...|X_i(j_i)|...X_k(j_k))^T \mathbf{1}|, \\ \qquad\qquad \text{if } k = 2m+1, m \in \mathbb{N} \end{cases} \tag{15}$$

The theorem above shows that even if the covariance tensor is calculated with equation (13), each entry of the covariance tensor is still the representation of the higher-order correlation defined by equation (12) for corresponding features.

2.2.3 Higher and Lower-Order Correlations in GTCCA

Then to define GTCCA which allows for higher and lower order correlations (e.g., pairwise), let set $S_m = \{(m_1, ..., m_m) : m_i \in \{1, ..., k\}, m_i \neq m_j \forall i = 1, 2, ..., k\}$ be all distinct possible combination from k choose m. Let $S_m(i)$ be the ith element in the set, then

$$\begin{aligned} \arg \max_{h_1, h_2, ..., h_k} & \rho_k(1)^2 + \sum_{j=1}^{\binom{k}{k-1}} a_{k-1,j}\rho_{k-1}(j)^2 + ... \\ & + \sum_{j=1}^{\binom{k}{3}} a_{3,j}\rho_3(j)^2 + \sum_{j=1}^{\binom{k}{2}} a_{2,j}\rho_2(j)^2 \\ & \text{s.t.} \quad h_i^T h_i = 1, i = 1, 2, ..., k, \end{aligned} \tag{16}$$

where $\rho_m(j) = C_{S_m(j)} \times_1 h_{m_1} \times ... \times_m h_{m_m}$ for all $m = 1, 2, ..., k$. Compared with TCCA, this design allows for flexibility with respect to a specific experiment design of interest by allowing a portion of $a_{i,j}$ to be 0 (similar to DIABLO). In addition, the scaling factor $a_{i,j}$ can be set to values other than 1 to prioritize certain correlation structures (similar to SmCCNet). Since it is hard to optimize equation (16), we found the problem equivalency, which is given below:

Theorem 2 *Let $C_{S_m(j)}$ be the covariance tensor of view $(m_1, ..., m_m) \in S_m(j)$ such that $C_{S_m(j)} \in \mathbb{R}^{d_{m_1} \times d_{m_2} \times ... \times d_{m_m}}$, If the optimization goal is formulated as follow:*

$$\begin{aligned} \arg \max_{h_1, h_2, ..., h_k} & \rho_k(1)^2 + \sum_{j=1}^{\binom{k}{k-1}} a_{k-1,j} \rho_{k-1}(j)^2 + ... \\ & + \sum_{j=1}^{\binom{k}{3}} a_{3,j} \rho_3(j)^2 + \sum_{j=1}^{\binom{k}{2}} a_{2,j} \rho_2(j)^2 \\ & \text{s.t.} \quad h_i^T h_i = 1, i = 1, 2, ..., k, \end{aligned} \tag{17}$$

where $\rho_m(j) = C_{S_m(j)} \times_1 h_{m_1} \times ... \times_m h_{m_m}$ for all $m = 1, 2, ..., k$, then the optimization problem above is equivalent to the following:

$$\begin{aligned} \arg \min_{h_1, h_2, ..., h_k} & a_{k,1} ||C_{S_k(1)} - \hat{C}_{S_k(1)}||_F^2 \\ & + \sum_{j=1}^{\binom{k}{k-1}} a_{k-1,j} ||C_{S_{k-1}(j)} - \hat{C}_{S_{k-1}(j)}||_F^2 \\ & + ... + \sum_{j=1}^{\binom{k}{3}} a_{3,j} ||C_{S_3(j)} - \hat{C}_{S_3(j)}||_F^2 \\ & + \sum_{j=1}^{\binom{k}{2}} a_{2,j} ||C_{S_2(j)} - \hat{C}_{S_2(j)}||_F^2 \\ & \text{s.t.} \quad h_i^T h_i = 1, i = 1, 2, ..., k, \end{aligned} \tag{18}$$

where $\hat{C}_{S_m(j)} = \rho_m(j) h_{m_1} \circ h_{m_2} \circ ... \circ h_{m_m}$ is the rank-1 approximation of $C_{S_m(j)}$.

2.2.4 Multi-Omics Example

In a multi-omics data example, suppose there are three types of molecular profiles: transcriptomics (tr), proteomics(pr), and metabolomics (me), with phenotype (ph) data. Define these datasets as X_{tr}, X_{pr}, X_{me} and Y_{ph}. Using GTCCA to find phenotype-related correlation structure, the optimization problem is given by:

$$\begin{aligned} \arg \max_{h_{tr}, h_{pr}, h_{me}, h_{ph}} & \rho^2_{tr,pr,me,ph} + \rho^2_{tr,pr,ph} + \rho^2_{tr,me,ph} + \rho^2_{pr,me,ph} \\ & + \rho^2_{tr,ph} + \rho^2_{pr,ph} + \rho^2_{me,ph} \\ & \text{s.t.} \quad h_i^T h_i = 1, i = g, pr, m, ph, \end{aligned} \tag{19}$$

where $\rho_{tr,pr,me,ph} = C_{tr,pr,me,ph} \times_1 h_{tr} \times_2 h_{pr} \times_3 h_{me} \times_4 h_{ph}$, and the other correlation components are in the same form as well. By Theorem (2), optimizing this objective function is equivalent to:

$$\begin{aligned}\arg \min_{h_{tr},h_{pr},h_{me},h_{ph}} & ||C_{tr,pr,me,ph} - \hat{C}_{tr,pr,me,ph}||_F^2 \\ & + ||C_{tr,pr,ph} - \hat{C}_{tr,pr,ph}||_F^2 \\ & + ||C_{tr,me,ph} - \hat{C}_{tr,pr,ph}||_F^2 \\ & + ||C_{pr,me,ph} - \hat{C}_{pr,me,ph}||_F^2 \\ & + ||C_{tr,ph} - \hat{C}_{tr,ph}||_F^2 + ||C_{pr,ph} - \hat{C}_{pr,ph}||_F^2 \\ & + ||C_{me,ph} - \hat{C}_{me,ph}||_F^2 \\ & \text{s.t. } h_i^T h_i = 1, i = g, pr, m, ph, \end{aligned} \tag{20}$$

where $\hat{C}_{tr,pr,me,ph} = \rho_{tr,pr,me,ph} \cdot h_{tr} \circ h_{pr} \circ h_{em} \circ h_{ph}$ represents the rank-1 approximation of the covariance tensor, and other components are similar. Below is the example gradient calculation for the transcriptome h_{tr},

$$\begin{aligned}\frac{\partial f}{\partial h_{tr}} &= [\hat{C}_{tr,pr,me,ph(1)} - C_{tr,pr,me,ph(1)}] \\ &\cdot (\rho_{tr,pr,me,ph} \cdot h_{tr} \odot h_{pr} \odot h_{me}) \\ &+ [\hat{C}_{tr,pr,ph(1)} - C_{tr,pr,ph(1)}](\rho_{tr,pr,ph} \cdot h_{ph} \odot h_{pr}) \\ &+ [\hat{C}_{tr,me,ph(1)} - C_{tr,me,ph(1)}](\rho_{tr,me,ph} \cdot h_{ph} \odot h_{me}) \\ &+ (h_{tr}\rho_1 h_{ph}^T - C_{1,y})h_{ph}\rho_{tr,ph} + \lambda(h_{tr} - \bar{h}_{ph}) \end{aligned} \tag{21}$$

$$\begin{aligned}\frac{\partial f}{\partial \rho_{tr,pr,me,ph}} &= (\hat{C}_{tr,pr,me,ph} - C_{tr,pr,me,ph}) \\ &\times_1 h_{tr} \times_2 h_{pr} \times_3 h_{me} \times_4 h_{ph}, \end{aligned} \tag{22}$$

where $\odot$ is the Khatri-Rao product, given two matrices $A \in \mathbb{R}^{m_1 \times n}$ and $B \in \mathbb{R}^{m_1 \times n}$, and let $\otimes$ denotes the Kronecker product between two vectors, the Khatri-Rao product is given by:

$$A \odot B = [a_1 \otimes b_1, a_2 \otimes b_2, ..., a_n \otimes b_n], \tag{23}$$

and $\otimes$ is the Kronecker product. $C_{tr,pr,me,ph(i)}$ is the mode-i matricization of tensor $C_{tr,pr,me,ph}$. This is a way to matricize the tensor by mapping the elements of the tensor into a matrix by arranging the mode-i fibers (think of it as the higher-order rows and columns) so that they become the columns of $C_{tr,pr,me,ph(i)}$. The gradient for other hs and ρs can be calculated similarly. After the gradient has been taken, the next step is to concatenate all the gradients into a long vector. All first-order gradient-based methods can be used for optimization, and the nonlinear conjugate gradient method is chosen to solve for the optimal canonical weights (see Acar et al. (2014) for detail).

2.3 Sparse Generalized Tensor Canonical Correlation Analysis

One of the most common ways of ensuring sparsity is to apply a penalty-based method, such as lasso. However, these methods are computationally intensive in this context since they

would require tensor computation on the original covariance tensors. SmCCNet tackles sparsity with the combination of subsampling and cross-validation. However, since the general recommendation of subsampling is 70% of the original features, it is still computationally intensive in the case of covariance tensor when applying the SmCCNet subsampling method. Therefore, we modified and implement a powerful and effective method called Turbo-SMT proposed by Papalexakis et al. (2014), which guarantees both accuracy and effectiveness of the tensor/matrix decomposition. The core idea of Turbo-SMT is to biased subsample features based on the tensor/matrix density value that is calculated for each feature, that is, a feature in tensor/matrix that generally has a larger value in corresponding entries will more likely be selected than other features that have lower values. However, tensor/matrix decomposition often involves the extraction of multiple factors (canonical weight in our model), and the order of solution will likely be distorted between iterations. To address this issue, in subsampling steps, all iterations will share a portion of common features, and those features will be used to adjust the order of factors based on the correlation between factors. For instance, if two iterations (subsamples) are run with Turbo-SMT, and two factors are extracted for each iteration, then to ensure the order of factors is not distorted between iteration 1 and iteration 2, the correlation between factors based on common features should be calculated to adjust for the order of the solution. Below is the definition of tensor density:

Definition 3 *Let $C_{1,2,...,k} \in \mathbb{R}^{d_1 \times d_2 \times ... \times d_k}$ be a tensor with k mode (in covariance tensor, it is equivalent to k views), then the tensor density with respect to mode i, denoted by $I_i \in \mathbb{R}^{d_i \times 1}$ is defined as:*

$$I_i = \sum_{j_1,...,j_{i-1},j_{i+1},...,j_k} C_{1,2,...,k}(j_1, ..., j_{i-1}, \cdot, j_{i+1}, ..., j_k). \tag{24}$$

In general, it means we sum over all other modes except for the mode i. Suppose for Equation (20), define $I_i^{(C)}$ as the density for covariance tensor/matrix C with respect to the ith mode, and Norm() is defined as the vector is normalized by L1 norm of the vector, then, for example, the subsampling density with respect to transcriptomics (tr), proteomics (pr), metabolomics (me) data is defined as:

$$\begin{aligned}
I_{tr} &= \text{Norm}([I_{tr}^{(C_{tr,pr,me,ph})}]^2) + \text{Norm}([I_{tr}^{(C_{tr,pr,ph})}]^2) + \text{Norm}([I_{tr}^{(C_{tr,me,ph})}]^2) + \text{Norm}([I_{tr}^{(C_{tr,ph})}]^2) \\
I_{pr} &= \text{Norm}([I_{pr}^{(C_{tr,pr,me,ph})}]^2) + \text{Norm}([I_{pr}^{(C_{tr,pr,ph})}]^2) + \text{Norm}([I_{pr}^{(C_{pr,me,ph})}]^2) + \text{Norm}([I_{pr}^{(C_{pr,ph})}]^2) \\
I_{pr} &= \text{Norm}([I_{me}^{(C_{tr,pr,me,ph})}]^2) + \text{Norm}([I_{me}^{(C_{tr,me,ph})}]^2) + \text{Norm}([I_{me}^{(C_{pr,me,ph})}]^2) + \text{Norm}([I_{me}^{(C_{pr,ph})}]^2).
\end{aligned} \tag{25}$$

This formulation ensures that each covariance tensor/matrix contributes equally to the subsampling density vector, and taking the square instead of direct summation ensures consistency of covariance density calculation with Equation (19). Detailed information about Turbo-SMT can be found in (Papalexakis et al., 2014).

2.4 Network Construction and Trimming

After obtaining the canonical weight through SGTCCA, the next step is to construct an adjacency matrix. The adjacency matrix construction process is similar to SmCCNet, and

the detailed algorithm is illustrated in algorithm 1 and Shi et al. (2019). The general idea is to take the outer product between concatenated canonical weight and itself, and if multiple set of canonical weights are extracted from SGTCCA, each set of canonical weight itself will form one adjacency matrix. However,even though the adjacency matrix is sparse, it may still contain noisy features/nodes. Therefore, we further trim the global network with the PageRank algorithm (Langville and Meyer, 2011) and principal component analysis (PCA) network summarization score using the first principal component (PC1) The original PageRank algorithm was used by Google to rank web pages based on their importance. Its application to networks (adjacency matrix) is to count the number and strength of the edges for each node to determine how important each node is. PC1 summarization score is obtained by taking the subset of molecular features based on the network module, then perform principal component analysis and obtain the PC score. The general workflow of the network trimming algorithm is as follows:

- calculate node importance from the adjacency matrix with the PageRank algorithm.
- Start from a baseline network size with top nodes based on step 1 (minimum network size that user set), adding the next node based on node importance, calculate PC score for the augmented network, and evaluate PC score correlation with phenotype and PC score correlation with PC score obtained from the baseline network.
- repeat 2 until correlation with phenotype and PC score correlation with baseline network reach a certain threshold.

This design ensures both high summarization score correlation with phenotype and baseline network summarization score, which implies no noise node are kept in the network after trimming.

Algorithm 1: Network Construction and Trimming Algorithm

Input: Canonical weight array $H_j \in \mathbb{R}^{d_j \times S \times I}, j = 1, 2, ..., k$, S:total number of solutions (canonical weight), I: total number of subsamples in Turbo-SMT ;

Initialize: s = 1 ;

Repeat until $s = S$:

(1) Extract the sth canonical weight from each omics data, obtain $H_{s,j} \in \mathbb{R}^{d_j \times I}$, $j = 1, 2, ..., k$;

(2) Scale and concatenate $H_{s,j}$ for all $j = 1, 2, ..., k$, and obtain $H \in \mathbb{R}^{(\sum_{j=1}^{k} d_j) \times I}$

(3) Calculate adjacency matrix $M = [H, H]$, where $[\cdot;\cdot,\cdot]$ is the matrix outer product, and $M_s \in \mathbb{R}^{(\sum_{j=1}^{k} d_j) \times (\sum_{j=1}^{k} d_j)}$;

(4) Trim M_s with PageRank algorithm and PC1 summarization score and obtain sub-networks $M_s^{(sub)} \in \mathbb{R}^{(\sum_{j=1}^{k} p_j) \times (\sum_{j=1}^{k} p_j)}$, where $p_j \leq d_j$ for all $j = 1, 2, ..., k$;

(5) s = s + 1;

Output: Global network $M_1, M_2, ..., M_S$, and sub-networks $M_1^{(sub)}, M_2^{(sub)}, ..., M_s^{(sub)}$

2.5 Simulation Study

To evaluate the performance of our method versus other state-of-the-art methods, we simulate multi-omics data based on latent factors (See Figure 2). Four datasets are simulated to represent transcriptomics, proteomics metabolomics, and phenotype. Five independent clusters are simulated, which include (1) gene-protein-metabolite-phenotype 4-way correlation (red, Figure 2), (2) gene-protein-phenotype 3-way correlation (orange, Figure 2), (3) gene-metabolite-phenotype 3-way correlation (yellow, Figure 2), (4) protein-metabolite-phenotype 3-way correlation (green, Figure 2), and (5) all phenotype-specific 2-way correlation (light blue, dark blue, and purple, Figure 2). In addition, non-phenotype-related correlation structures such as gene-protein-metabolite 3-way correlation (brown, Figure 2) are simulated; Furthermore, we randomly permute rows and columns of a multi-omics data set and use it as the background dataset (see COPD data set below).

- Simulate latent 11 factors $l_1, l_2, ..., l_{11} \in \mathbb{R}^{N\times 1}$ following multivariate normal distribution $\text{MVN}(0, \Sigma)$, where Σ is the 11×11 identity matrix, each latent factor represents a particular correlation structure (different colors in Figure 2).
- Simulate block-wise weight vectors $b_{i,j} \in \mathbb{R}^{1\times P_j}, i = 1, 2, ..., 10; j = 1, 2, 3, 4$, where j is the number of omics data sets.
- Randomly permutes rows and columns of real multi-omics data and obtains background data $E_j \in \mathbb{R}^{N\times P_j}, j = 1, 2, 3, 4$.
- Simulate multi-omics data based on latent factors and randomly-permuted real multi-omics data.

The formula for obtaining the final simulation data is:

$$\begin{aligned}
X_1 &= l_1 \cdot b_{1,1}^T + l_2 \cdot b_{2,1}^T + l_3 \cdot b_{3,1}^T + l_5 \cdot b_{5,1}^T \\
&\quad + l_8 \cdot b_{8,1}^T + l_9 \cdot b_{9,1}^T + l_{10} \cdot b_{10,1}^T + \alpha \cdot E_1 \\
X_2 &= l_1 \cdot b_{1,2}^T + l_2 \cdot b_{2,2}^T + l_4 \cdot b_{4,2}^T + l_6 \cdot b_{6,2}^T \\
&\quad + l_8 \cdot b_{8,2}^T + l_9 \cdot b_{9,2}^T + l_{11} \cdot b_{11,2}^T + \alpha \cdot E_2 \\
X_3 &= l_1 \cdot b_{1,2}^T + l_3 \cdot b_{3,3}^T + l_4 \cdot b_{4,3}^T + l_7 \cdot b_{7,3}^T \\
&\quad + l_8 \cdot b_{8,3}^T + l_{10} \cdot b_{10,3}^T + l_{11} \cdot b_{11,3}^T + \alpha \cdot E_3 \\
X_4 &= \sum_{i=1}^{7} l_i + \beta \cdot E_4
\end{aligned} \tag{26}$$

Where α and β represent the strength of the noise. In summary, the simulated datasets have 3 types of blocks: (1) signal blocks representing all phenotype-specific correlation structures, which are given by latent factors $l_1, ..., l_5$; (2) non-phenotype-specific blocks representing all non-phenotype-specific correlation structures, which are given by latent factors $l_6, ..., l_{11}$; (3) background noise without any correlation structure, which is given by randomly permuted multi-omics data E_1, E_2, E_3, and E_4.

To mimic the true multi-omics correlation structure in most scenarios (the strong connection between omics but the weak connection between omics and phenotype) to test the noise tolerance of each method, we imposed weaker noise on the omics data, and stronger noise on the simulated phenotype data by setting α to 0.2 and β to 1. For each omics data, each phenotype-related correlation block has 20 features, and the non-phenotype-related correlation block has 60 features, with the rest of noisy features (i.e., no relationship between omics or phenotype). To make the simulated data size compatible with the real-world multi-omics data we used to simulate the background noise, we simulate 462 subjects, with 972 features in gene data, 1317 features in protein data, and 995 features in metabolite data.

We compare the performance of our novel pipeline with the following variations on 25 replications: (1) SGTCCA-Net with only the higher order gene-protein-metabolite-phenotype correlation being explored; (2) SGTCCA-Net with the original covariance tensor calculation scheme, (3) SmCCNet with various combinations of scaling factors and sparsity levels, (4) DIABLO from MixOmics with the different number of features kept. All these methods have available adjacency matrices for performance evaluation. The performance is evaluated at the node level, and a node is predicted positive if its maximal connection to other nodes in the adjacency matrix passes a certain threshold, which is consistent with SmCCNet evaluation (Shi et al., 2019), and the AUC score can be calculated through checking prediction result with a series of the threshold value.

To ensure a fair comparison between methods, we only extract one solution (canonical weight) from each of these methods. For SmCCNet, the penalty candidates are 0.25 (high sparsity), 0.5 (medium sparsity), and 0.75 (light sparsity) with the subsampling proportion of 70% and the subsampling number of 100. There are two options for SmCCNet: (1) unweighted SmCCNet which has equal scaling factors for all pairwise correlations, and (2) weighted SmCCNet with scaling factors for all phenotype-related correlation components set to 10 and others set to 1. Setting the scaling factors of the phenotype-specific correlation structure to 10 is the common approach for most of SmCCNet applications (Mastej et al., 2020; Shi et al., 2019), which not only prioritizes the phenotype-specific correlation but also preserves a portion of the between-omics correlation. For DIABLO, the number of features kept from each omics type is set to different levels. In this experiment, we use 100, 150, 200,250, and 300 for evaluation.

2.6 TCGA Breast Cancer Network Analysis

We use multi-omics data from the Cancer Genome Atlas Program (TCGA) breast invasive carcinoma project to demonstrate the analytical result of our novel network analysis pipeline. The dataset used in this experiment contains RNA sequencing data with a normalized count obtained through the Illumina HiSeq platform, micro RNA (miRNA) expression data of tumor samples obtained through the Illumina HiSeq platform at miRgene-level, and log-ratio normalized reverse phase protein arrays (RPPA) expression data from tumor samples at the gene level. The phenotype used in this example is tumor purity, which is defined as the percentage of cancer cells in one sample of tumor tissue (Li et al., 2019). After matching subjects with all molecular profiles and phenotype data available, there are 117 subjects. After filtering genes based on standard deviation to eliminate genes that are less variable, the breast cancer data has 5039 genes, 823 miRNAs, and 175 RPPAs. In SGTCCA-

Net, we assume the correlation structure of gene-miRNA-RPPA-phenotype, gene-miRNA-phenotype, gene-RPPA-phenotype, miRNA-RPPA-phenotype, and all pairwise molecular profiles with phenotype. The percentage of common subsampled features is set to 7% and the percentage of distinct features is set to 2%. The total number of subsamples is set to 10. Only the first solution will be used to construct the adjacency matrix, but to ensure the canonical weight correspondence for the first canonical weight is met, we extract two sets of canonical weight and perform canonical weight matching between subsamples. To evaluate the result from the TCGA breast cancer data, we use both the network summary PC1 correlation with respect to the phenotype as well as individual molecular feature correlation with respect to phenotype for all network molecular features.

2.7 COPD Gene Network Analysis

In this experiment, we use the COPDGene gene data with transcriptomics (RNA-Seq), proteomics (SomaLogic), and metabolomics (Metabolon) to construct multi-omics networks with respect to phenotype forced expiratory volume (FEV1) percent predicted and percent emphysema. FEV1 is defined as the amount of air forced by patients from their lungs in one second, and the percent predicted means patients' FEV1 value divided by the average FEV1 in the population of people with similar demographics. Percent emphysema is quantified with regard to lung density in computed tomography (CT) images.

We use the data from phase II of COPDGene (Regan et al., 2011). The multi-omics data used here is described in more detail in (Gillenwater et al., 2021). The full CODPGene total mRNA sequencing data contained 2655 peripheral blood samples and 65988 genes following data processing, and after matching subjects with molecular profiles (genes, proteins, metabolites) and phenotype data available, there were 462 subjects. In addition, we filtered the transcriptomics data with standard deviation to eliminate genes with less variability (threshold = 0.435). After data preprocessing, there were 462 subjects, with 972 genes, 1317 proteins, and 995 metabolites. To run SGTCCA-Net, as above we assumed a correlation structure of gene-protein-metabolite-phenotype, gene-protein-phenotype, gene-metabolite-phenotype, protein-metabolite-phenotype, and all pairwise molecular profiles with phenotype. For Turbo-SMT, the percentage of common subsampled features is set to 7%, and the percentage of distinct features is set to 2%. The total number of subsamples is set to 10. Same as above, only the first solution is used to construct the adjacency matrix.

3. Result

3.1 Simulation

Table 1 shows that under the most complex correlation structures and weak correlation with the phenotype, SGTCCA-Net significantly outperforms the other two CCA-based network analysis methods SmCCNet and DIABLO in all cases and when only the 4-way correlation structure is assumed (STCCA-Net). SmCCNet and DIABLO in theory may be able to find higher-order correlations but because the objective functions are focused on pairwise correlations they miss many higher-order relationships (false negatives), in addition to finding spurious higher-order correlations that do not have phenotype relationships (false positives. Comparing SGTCCA-Net with STCCA-Net, the latter only finds the 4-way

higher-order correlations and is limited in finding other lower-order phenotype-specific correlation structures. DIABLO generally has poor results compared with other methods because (1) it uses an unweighted version of sparse multiple canonical correlation analysis, with less emphasis on phenotype-specific correlation, and (2) the adjacency matrix of DIABLO is a correlation matrix between all selected variables, not based on the canonical weights as in all the other methods, which generates large amounts of false positive nodes when the between-omics correlation is high but there is no phenotype-specific correlation structure present.

3.2 TCGA Breast Cancer Network Analysis

We applied SGTCCA-Net to the TCGA breast cancer data with 5039 genes, 823 miRNAs, 175 RPPAs, and phenotype of tumor purity. After trimming the network, the final subnetwork contains 53 molecular features with 22 genes, 7 miRNAs, and 24 RPPAs. The PC1 summarization score correlation with respect to phenotype is 0.769 ($p < 1 \times 10^{-3}$).

C16orf54 (Chromosome 16 open reading frame 54) has the highest correlation with respect to tumor purity among all the genes in the network (-0.759) (Table 2) and has been shown to have low expression levels in most of cancer, and it has high predicted power when distinguishing tumor from normal tissue (Du et al., 2022); hsa-mir-150 has the highest correlation with respect to tumor purity among all the miRNAs in the network (-0.643), and is involved in many solid tumors such as breast, lung and gastric (Wang et al., 2015); PCNA (proliferating cell nuclear antigen) has the highest correlation with respect to tumor purity (-0.587), and is associated with DNA replication in cancer cells (Wang et al., 2022). The network module is visualized in Figure 3.

3.3 COPD Gene Network Analysis

We applied SGTCCA-Net to the COPD gene data with 972 genes, 1317 proteins, 995 metabolomics, and phenotype of FEV1 percent predicted (FEV1pp) and percent emphysema. After trimming the FEV1pp network, the final subnetwork contains 27 molecular features with 4 genes, 12 proteins, and 11 metabolites. The PC1 summarization score correlation with respect to phenotype is 0.374 ($p < 1 \times 10^{-3}$). After trimming the percent emphysema network, the final subnetwork contains 55 molecular features, which include 10 genes, 28 proteins, and 17 metabolites. The PC1 summarization score correlation with respect to phenotype is 0.352 ($p < 1 \times 10^{-3}$)

The top ten features for each phenotype are reported in Table 3. Troponin T has the highest correlation with respect to FEV1pp (-0.376), and studies have shown that elevated level of Troponin T during an exacerbation is associated with the death of COPD patients (Blaschko and Lampert, 2008; Elmenawi et al., 2021). C-reactive protein is also negatively correlated with FEV1pp (-0.270) and has been shown to be associated with poor lung function (Aksu et al., 2013). In the emphysema network, C5AR2 (Complement C5a Receptor 2) has the highest correlation with respect to percent emphysema (0.331), and increased levels are associated with a COPD patient's recovery time (Westwood et al., 2016). Finally, leptin is also highly correlated with percent emphysema (-0.277), which is an important biomarker for COPD exacerbation (Masoud et al., 2019) and is associated with reduced lung function (Nilsson et al., 2021). The network module is visualized in Figure 4.

4. Conclusion and Discussion

In this paper, we have developed a novel multi-omics network analysis pipeline called Sparse Generalized Tensor Canonical Correlation Analysis Network Inference (SGTCCA-Net). This method successfully addresses the weakness of existing CCA-based multi-omics network inference methods SmCCNet and DIABLO, two methods on the basis of pairwise correlation, and is able to capture any higher-order/lower-order correlation of interest. In a simulation study, the current version of SGTCCA-Net is capable of identifying features with different types of higher-order/lower-order phenotype-specific correlation structures and significantly outperforms the other two CCA-based multi-omics network inference pipelines under different model setups. The real data analysis demonstrates the ability of our method in capturing important molecular features for cancer and lung disease.

Despite the advantages of our network analysis pipeline, there are still some drawbacks: (1) the current method is capable of capturing correlation structure, but with limited ability to separate and report different correlation structures; in future work we will add steps in the pipeline to effectively separate all different correlation structures (e.g., what combinations of features have two-way versus three-way relationships), (2) the current method compresses information from SGTCCA to a 2D space to construct and adjacency matrix for network inference, however, covariance tensors are the natural representation of a hypergraph (graph that connects more than 2 nodes), in the future we will develop methods to infer networks through the hypergraph directly, and (3) the network summarization using principal component analysis fails to consider the higher-order network topology, and thus we will develop advanced algorithms to obtain a network summarization score that can best represents network topology in the higher-order setting.

Funding

R01HL152735 (KJK, FBK, RPB, WL), R01HL089897 (RPB, KAP), R01HL137995 (RPB, LAG, KJK); R01HL129937 (RPB), U01HL089856 (COPDGene), U01HL089897 (COPDGene).

References

E. Acar, E. E. Papalexakis, G. Gürdeniz, M. A. Rasmussen, A. J. Lawaetz, M. Nilsson, and R. Bro. Structure-revealing data fusion. *BMC bioinformatics*, 15(1):1–17, 2014.

F. Aksu, N. Capan, K. Aksu, R. Ofluoğlu, S. Canbakan, B. Yavuz, and K. O. Akin. C-reactive protein levels are raised in stable chronic obstructive pulmonary disease patients independent of smoking behavior and biomass exposure. *Journal of thoracic disease*, 5 (4):414, 2013.

S. Bickel and T. Scheffer. Multi-view clustering. In *ICDM*, volume 4, pages 19–26, 2004.

M. B. Blaschko and C. H. Lampert. Correlational spectral clustering. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000.

X. Du, W. Xia, W. Fan, X. Shen, H. Wu, and H. Zhang. Integrated analysis of c16orf54 as a potential prognostic, diagnostic, and immune marker across pan-cancer. *Disease markers*, 2022, 2022.

R. Duan, L. Gao, Y. Gao, Y. Hu, H. Xu, M. Huang, K. Song, H. Wang, Y. Dong, C. Jiang, et al. Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLoS computational biology*, 17(8):e1009224, 2021.

K. A. Elmenawi, V. Anil, H. Gosal, H. Kaur, H. C. Ngassa, and L. Mohammed. The importance of measuring troponin in chronic obstructive pulmonary disease exacerbations: A systematic review. *Cureus*, 13(8), 2021.

A. Gilchrist, C. E. Au, J. Hiding, A. W. Bell, J. Fernandez-Rodriguez, S. Lesimple, H. Nagaya, L. Roy, S. J. Gosline, M. Hallett, et al. Quantitative proteomics analysis of the secretory pathway. *Cell*, 127(6):1265–1281, 2006.

L. A. Gillenwater, S. Helmi, E. Stene, K. A. Pratte, Y. Zhuang, R. P. Schuyler, L. Lange, P. J. Castaldi, C. P. Hersh, F. Banaei-Kashani, et al. Multi-omics subtyping pipeline for chronic obstructive pulmonary disease. *PloS one*, 16(8):e0255337, 2021.

A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert. Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6(1):1–17, 2012.

H. Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992.

W. Hu, D. Lin, S. Cao, J. Liu, J. Chen, V. D. Calhoun, and Y.-P. Wang. Adaptive sparse multiple canonical correlation analysis with application to imaging (epi) genomics study of schizophrenia. *IEEE Transactions on Biomedical Engineering*, 65(2):390–399, 2017.

V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

D. Jiang, C. R. Armour, C. Hu, M. Mei, C. Tian, T. J. Sharpton, and Y. Jiang. Microbiome multi-omics network analysis: statistical considerations, limitations, and opportunities. *Frontiers in genetics*, 10:995, 2019.

M. M. Kalayeh, H. Idrees, and M. Shah. Nmf-knn: Image annotation using weighted multi-view non-negative matrix factorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 184–191, 2014.

A. N. Langville and C. D. Meyer. Google's pagerank and beyond. In *Google's PageRank and Beyond*. Princeton university press, 2011.

Y. Li, D. M. Umbach, A. Bingham, Q.-J. Li, Y. Zhuang, and L. Li. Putative biomarkers for predicting tumor sample purity based on gene expression data. *BMC genomics*, 20(1): 1–12, 2019.

E. Lin and H.-Y. Lane. Machine learning and systems genomics approaches for multi-omics data. *Biomarker research*, 5(1):1–6, 2017.

J. Liu, C. Wang, J. Gao, and J. Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 252–260. SIAM, 2013.

Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen. Tensor canonical correlation analysis for multi-view dimension reduction. *IEEE transactions on Knowledge and Data Engineering*, 27(11):3111–3124, 2015.

H. H. Masoud, A. El-Hafeez, M. Ahmed, M. S. Ismail, and N. G. Baharetha. Leptin as a local inflammatory marker in chronic obstructive pulmonary disease acute exacerbation. *Egyptian Journal of Bronchology*, 13(2):139–147, 2019.

E. Mastej, L. Gillenwater, Y. Zhuang, K. A. Pratte, R. P. Bowler, and K. Kechris. Identifying protein–metabolite networks associated with copd phenotypes. *Metabolites*, 10(4):124, 2020.

O. Menyhárt and B. Győrffy. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Computational and structural biotechnology journal*, 19:949–960, 2021.

M. Monteiro, M. Carvalho, M. Bastos, and P. Guedes de Pinho. Metabolomics analysis for biomarker discovery: advances and challenges. *Current medicinal chemistry*, 20(2): 257–271, 2013.

G. Nicora, F. Vitali, A. Dagliati, N. Geifman, and R. Bellazzi. Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. *Frontiers in oncology*, 10:1030, 2020.

U. Nilsson, S. Söderberg, H. Backman, A. Blomberg, and A. Lindberg. Leptin levels are associated with reduced lung function in men with copd, 2021.

E. E. Papalexakis, C. Faloutsos, T. M. Mitchell, P. P. Talukdar, N. D. Sidiropoulos, and B. Murphy. Turbo-smt: Accelerating coupled sparse matrix-tensor factorizations by 200x. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 118–126. SIAM, 2014.

M. Pierre-Jean, J.-F. Deleuze, E. Le Floch, and F. Mauger. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. *Briefings in bioinformatics*, 21(6):2011–2030, 2020.

N. Rappoport and R. Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic acids research*, 46(20):10546–10562, 2018.

E. A. Regan, J. E. Hokanson, J. R. Murphy, B. Make, D. A. Lynch, T. H. Beaty, D. Curran-Everett, E. K. Silverman, and J. D. Crapo. Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1):32–43, 2011.

W. J. Shi, Y. Zhuang, P. H. Russell, B. D. Hobbs, M. M. Parker, P. J. Castaldi, P. Rudra, B. Vestal, C. P. Hersh, L. M. Saba, et al. Unsupervised discovery of phenotype-specific multi-omics networks. *Bioinformatics*, 35(21):4336–4343, 2019.

A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, and K.-A. Lê Cao. Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062, 2019.

I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14: 1177932219899051, 2020.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

F. Wang, X. Ren, and X. Zhang. Role of microrna-150 in solid tumors. *Oncology letters*, 10 (1):11–16, 2015.

Y.-L. Wang, W.-R. Wu, P.-L. Lin, Y.-C. Shen, Y.-Z. Lin, H.-W. Li, K.-W. Hsu, and S.-C. Wang. The functions of pcna in tumor stemness and invasion. *International Journal of Molecular Sciences*, 23(10):5679, 2022.

J.-P. Westwood, A. J. Mackay, G. Donaldson, S. J. Machin, J. A. Wedzicha, and M. Scully. The role of complement activation in copd exacerbation recovery. *ERJ Open Research*, 2 (4), 2016.

D. M. Witten and R. J. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8 (1), 2009.

H. S. Wong, L. Wang, R. Chan, and T. Zeng. Deep tensor cca for multi-view learning. *arXiv preprint arXiv:2005.11914*, 2020.

C. Wu, F. Zhou, J. Ren, X. Li, Y. Jiang, and S. Ma. A selective review of multi-level omics data integration using variable selection. *High-throughput*, 8(1):4, 2019.

C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

J. Yan, S. L. Risacher, L. Shen, and A. J. Saykin. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in bioinformatics*, 19(6):1370–1381, 2018.

C. Zhang, H. Fu, Q. Hu, P. Zhu, and X. Cao. Flexible multi-view dimensionality co-reduction. *IEEE Transactions on Image Processing*, 26(2):648–659, 2016.

G. Zhou, S. Li, and J. Xia. Network-based approaches for multi-omics integration. *Computational Methods and Data Analysis for Metabolomics*, pages 469–487, 2020.

5. Tables and Figures

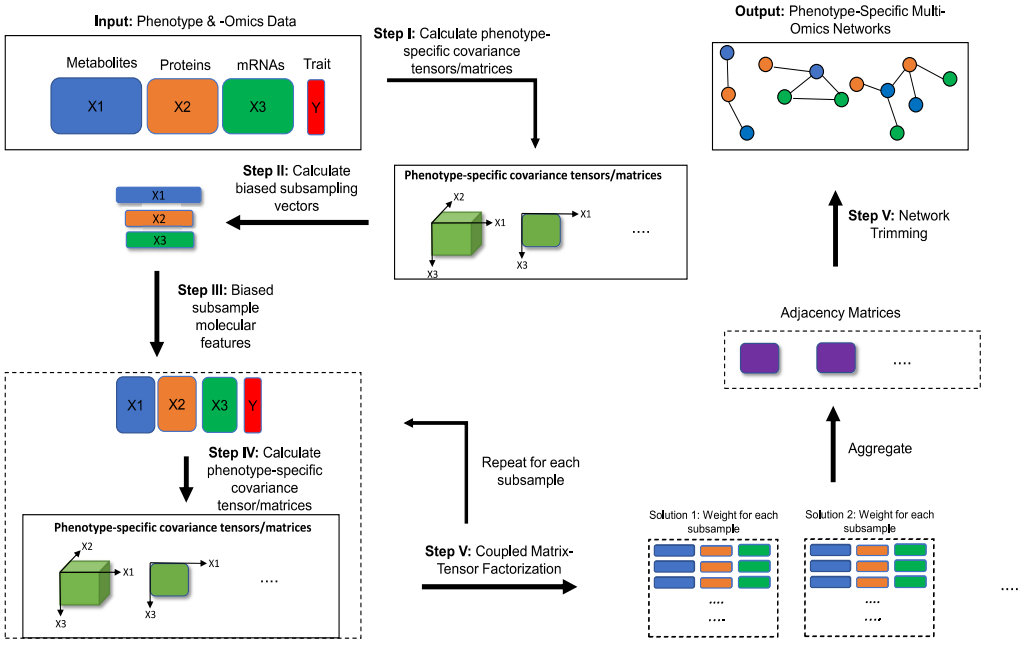

Figure 1: Workflow of SGTCCA-Net pipeline for multi-omics network inference. It consists of three core steps: higher-order correlation extraction, affinity matrix construction, and network trimming. Biased subsampling involves calculating the covariance tensor/matrix density with respect to certain molecular profiles, which is also the importance/probability vector when randomly sampling features from each molecular profile. Adjacency matrix calculation

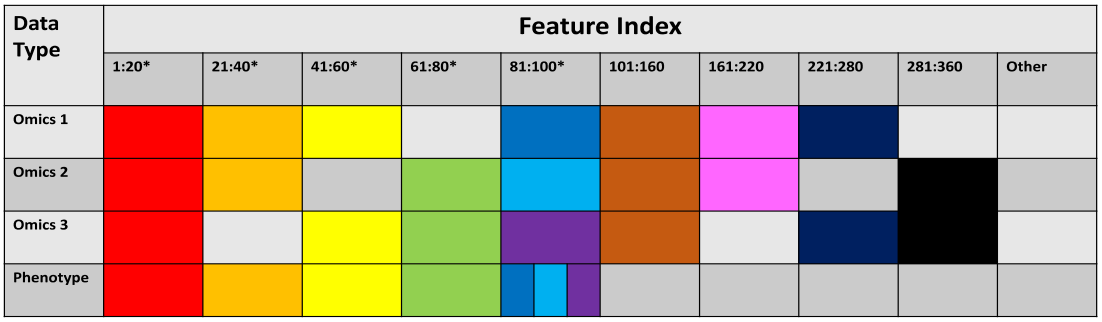

Figure 2: Simulated multi-omics data correlation structure. The same color represents features that are simulated with the same latent factor, and different colors denote that the two latent factors are independent, implying that the two correlation structures are independent. Grey means that no latent factor is used to simulate these features, * means these simulated features are signals features (phenotype-specific correlation structure).

Median AUC (interquartile range)				
Method	Omics 1	Omics 2	Omics 3	Overall
SGTCCA-Net	**0.932 (0.867, 0.942)**	**0.924 (0.894, 0.965)**	**0.912 (0.868, 0.931)**	**0.916 (0.888, 0.931)**
STCCA-Net	0.554 (0.500, 0.629)	0.484 (0.469, 0.542)	0.541 (0.498, 0.573)	0.522 (0.507, 0.549)
SmCCNet (unweighted, light sparsity)	0.545 (0.508, 0.554)	0.539 (0.525, 0.559)	0.533 (0.518, 0.572)	0.546 (0.529, 0.554)
SmCCNet (weighted, light sparsity)	0.686 (0.659, 0.716)	0.684 (0.646, 0.741)	0.521 (0.503, 0.543)	0.634 (0.621, 0.652)
SmCCNet (unweighted, medium sparsity)	0.534 (0.511, 0.577)	0.537 (0.499, 0.572)	0.526 (0.513, 0.576)	0.541 (0.525, 0.562)
SmCCNet (weighted, medium sparsity)	0.700 (0.677, 0.725)	0.693 (0.674, 0.727)	0.541 (0.514, 0.563)	0.648 (0.634, 0.668)
SmCCNet (unweighted, high sparsity)	0.474 (0.469, 0.476)	0.471 (0.428, 0.526)	0.473 (0.462, 0.476)	0.482 (0.449, 0.501)
SmCCNet (weighted, high sparsity)	0.669 (0.475, 0.770)	0.741 (0.659, 0.768)	0.633 (0.524, 0.753)	0.647 (0.597, 0.760)
DIABLO (keep 100 features)	0.457 (0.444, 0.512)	0.460 (0.460, 0.491)	0.445 (0.445, 0.458)	0.462 (0.454 0.493)
DIABLO (keep 150 features)	0.436 (0.416, 0.527)	0.445 (0.439, 0.487)	0.442 (0.424, 0.477)	0.460 (0.437, 0.476)
DIABLO (keep 200 features)	0.463 (0.400, 0.526)	0.442 (0.426, 0.506)	0.458 (0.415, 0.486)	0.463 (0.447, 0.490)
DIABLO (keep 250 features)	0.484 (0.395, 0.543)	0.445 (0.416, 0.531)	0.464 (0.419, 0.492)	0.464 (0.439, 0.514)
DIABLO (keep 300 features)	0.482 (0.408, 0.554)	0.444 (0.416, 0.523)	0.452 (0.418, 0.511)	0.460 (0.443, 0.520)

Table 1: Simulation results. The performance is evaluated through the AUC of the precision-recall curve generated by applying different thresholds to the maximal connection of molecular features to each other. For this simulation, 25 replications are used and the median and interquartile range in parenthesis of the AUC is used to compare the performance of different methods. Overall means taking the weighted average over all 3 omics data.

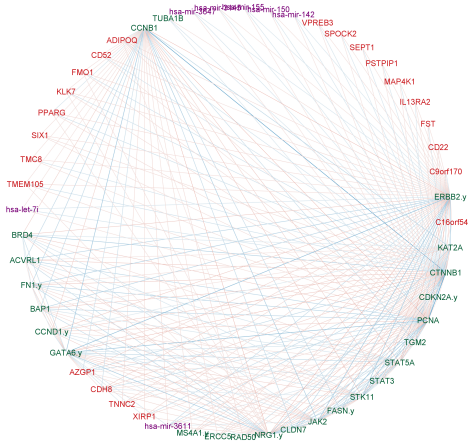

Figure 3: Multi-omics network module for TCGA breast cancer data with respect to tumor purity. The red node stands for gene, the purple node stands for miRNA, and the green node stands for RPPA. The width and color depth of the edge stands for the strength of the connection between two molecular features, and the type of color stands for whether two nodes are positively correlated (red) or negatively correlated (blue).

Gene			miRNA			RPPA		
Name	Type	Correlation	Name	Type	Correlation	Name	Type	Correlation
C16orf54	gene	-0.759	hsa-mir-150	mirna	-0.643	PCNA	rppa	-0.587
TMC8	gene	-0.749	hsa-let-7i	mirna	-0.614	NRG1.y	rppa	-0.505
PSTPIP1	gene	-0.729	hsa-mir-142	mirna	-0.599	STAT5A	rppa	-0.443
CD52	gene	-0.728	hsa-mir-155	mirna	-0.539	BAP1	rppa	0.427
SPOCK2	gene	-0.728	hsa-mir-3647	mirna	-0.388	BRD4	rppa	-0.416

Table 2: Top 5 individual molecular features in tumor purity network for TCGA data.

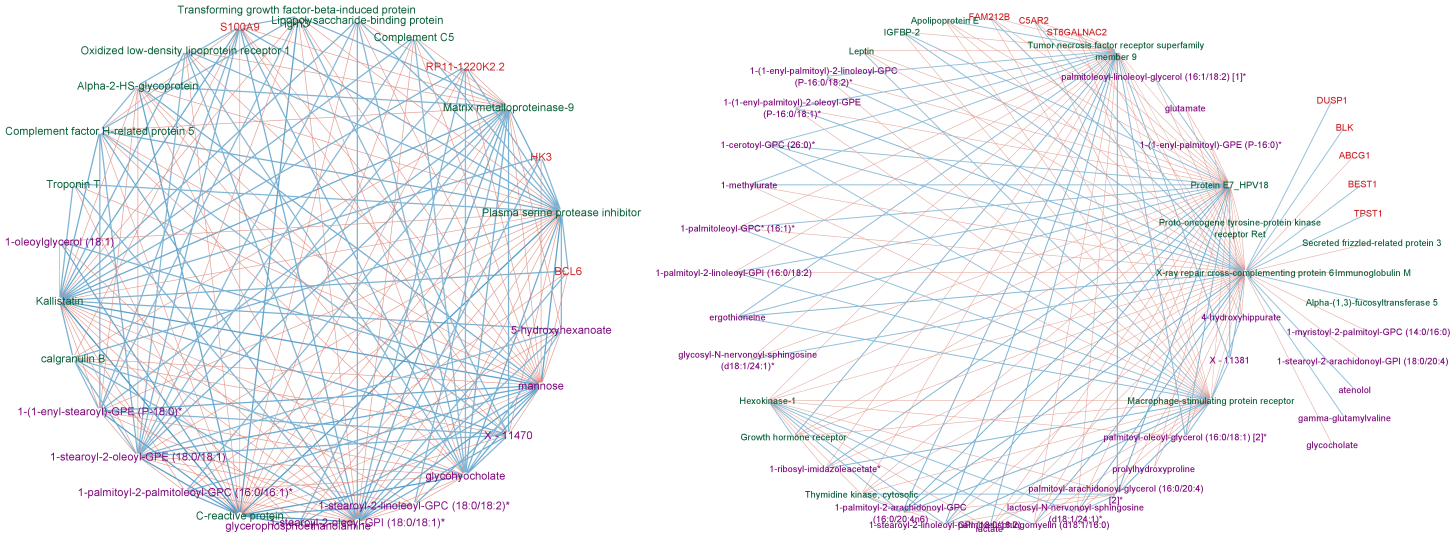

Figure 4: Multi-omics network module for COPDGene data with respect to FEV1 Percent Predicted (left) and Percent Emphysema (right). The red node stands for gene, the purple node stands for metabolite, and the green node stands for protein. The width and color depth of the edge stands for the strength of the connection between two molecular features, and the type of color stands for whether two nodes are positively correlated (red) or negatively correlated (blue). The number of nodes is different from the text because, for visualization, the network went through further edge-cut trimming.

FEV1 Percent Predicted			Percent Emphysema		
Name	Type	Correlation	Name	Type	Correlation
Troponin T	protein	-0.376	C5AR2	gene	0.331
C-reactive protein	protein	-0.270	Leptin	protein	-0.277
Kallistatin	protein	0.256	Growth hormone receptor	protein	-0.276
Plasma serine protease inhibitor	protein	0.229	ST6GALNAC2	gene	0.272
1-stearoyl-2-linoleoyl-GPC (18:0/18:2)*	metabolite	0.221	TPST1	gene	0.265
Oxidized low-density lipoprotein receptor 1	protein	-0.22	ABCG1	gene	0.259
S100A9	gene	-0.214	DUSP1	gene	0.255
5-hydroxyhexanoate	metabolite	-0.213	BEST1	gene	0.245
1-(1-enyl-stearoyl)-GPE (P-18:0)*	metabolite	0.212	BLK	gene	0.244
BCL6	gene	-0.212	SLC31A2	gene	-0.241

Table 3: Top 10 individual molecular features in network module that are highly correlated with phenotype FEV1 percent perdicted (left) and Percent Emphysema (right).

6. Supplement:Proof of Theorem 2

Proof

Based on the formulation above, the augmented Lagrangian of the optimization is given by:

$$\begin{aligned} f = & a_{k,1}||C_{S_k(1)} - \hat{C}_{S_k(1)}||_F^2 \\ & + \sum_{j=1}^{\binom{k}{k-1}} a_{k-1,j}||C_{S_{k-1}(j)} - \hat{C}_{S_{k-1}(j)}||_F^2 \\ & + ... + \sum_{j=1}^{\binom{k}{3}} a_{3,j}||C_{S_3(j)} - \hat{C}_{S_3(j)}||_F^2 \\ & + \sum_{j=1}^{\binom{k}{2}} a_{2,j}||C_{S_2(j)} - \hat{C}_{S_2(j)}||_F^2 + \sum_{i=1}^{k} \alpha_i((||h_i||_2^2 - 1) \\ & \text{s.t.} \quad h_i^T h_i = 1, i = 1, 2, ..., k \end{aligned} \tag{27}$$

Taking the derivative of f with respect to $\rho_m(j), m = 1, 2, ..., k$ and $j = 1, 2, ..., \binom{k}{m}$, and denote i_{m_s} the feature index of the m_s-th view $(s = 1, 2, ..., m)$, yields the following:

$$\begin{aligned} & \sum_{i_{m_1},...,i_{m_s}} C_{S_m(j)}(i_{m_1}, ..., i_{m_m}) \prod_{m_s \in S_m(j)} h_{m_s}(i_{m_s}) \\ & \quad - \rho_m(j) \sum_{i_{m_1},...,i_{m_s}} \prod_{m_s \in S_m(j)} h_{m_s}(i_{m_s})^2 \\ & \quad = 0 \end{aligned} \tag{28}$$

Taking the derivative of f with respect to $\alpha_s, s = 1, 2, ..., k$ yields:

$$\sum_{i_s} h_s(i_s)^2 = 1 \tag{29}$$

then combine the above two equations, we have:

$$\begin{aligned} \rho_m(j) &= \sum_{i_{m_1},...,i_{m_m}} C_{S_m(j)}(i_{m_1}, ..., i_{m_m}) \prod_{m_s \in S_m(j)} h_{m_s}(i_{m_s}) \\ &= C_{S_m(j)} \times_1 h_{m_1} \times_2 h_{m_2} \times ... \times h_{m_m} \end{aligned} \tag{30}$$

With the above result, we can finally derive the following:

$$
\begin{aligned}
&a_{k,1}||C_{S_k(1)} - \hat{C}_{S_k(1)}||_F^2 + \sum_{j=1}^{\binom{k}{k-1}} a_{k-1,j}||C_{S_{k-1}(j)} - \hat{C}_{S_{k-1}(j)}||_F^2 \\
&+ ... + \sum_{j=1}^{\binom{k}{3}} a_{3,j}||C_{S_3(j)} - \hat{C}_{S_3(j)}||_F^2 + \sum_{j=1}^{\binom{k}{2}} a_{2,j}||C_{S_2(j)} - \hat{C}_{S_2(j)}||_F^2 \\
&= a_{k,1}(||C_{S_k(1)}||_F^2 - 2 < C_{S_k(1)}, \hat{C}_{S_k(1)} > + ||\hat{C}_{S_k(1)}||_F^2) \\
&+ \sum_{j=1}^{\binom{k}{k-1}} a_{k-1,j}(||C_{S_{k-1}(j)}||_F^2 - 2 < C_{S_{k-1}(j)}, \hat{C}_{S_{k-1}(j)} > \\
&+ ||\hat{C}_{S_{k-1}(j)}||_F^2) + ... \\
&+ \sum_{j=1}^{\binom{k}{3}} a_{3,j}(||C_{S_3(j)}||_F^2 - 2 < C_{S_3(j)}, \hat{C}_{S_3(j)} > + ||\hat{C}_{S_3(j)}||_F^2) \\
&+ \sum_{j=1}^{\binom{k}{2}} a_{2,j}(||C_{S_2(j)}||_F^2 - 2 < C_{S_2(j)}, \hat{C}_{S_2(j)} > + ||\hat{C}_{S_2(j)}||_F^2), \qquad (31)
\end{aligned}
$$

where $< C_{S_m(j)}, \hat{C}_{S_m(j)} >$ is the inner product between two tensors, and by De Lathauwer et al. (2000), $< C_{S_m(j)}, \hat{C}_{S_m(j)} >$ is evaluated as:

$$
\begin{aligned}
&< C_{S_m(j)}, \hat{C}_{S_m(j)} > = \sum_{i_{m_1},...,i_{m_m}} C_{S_m(j)}(i_{m_1}, ..., i_{m_m}) \hat{C}_{S_m(j)}(i_{m_1}, ..., i_{m_m}) \\
&= \sum_{i_{m_1},...,i_{m_m}} C_{S_m(j)}(i_{m_1}, ..., i_{m_m}) \rho_m(j) h_{m_1}(i_{m_1}) \cdot ... \cdot h_{m_m}(i_{m_m}) \\
&= \rho_m(j) \sum_{i_{m_1},...,i_{m_m}} C_{S_m(j)}(i_{m_1}, ..., i_{m_m}) h_{m_1}(i_{m_1}) \cdot ... \cdot h_{m_m}(i_{m_m}) \\
&= \rho_m(j)^2 \qquad (32)
\end{aligned}
$$

Since $h_i, i = 1, 2, ..., k$ all have the unit norm, combining the equation above with equation 31 we obtain:

$$
\begin{aligned}
&a_{k,1}||C_{S_k(1)} - \hat{C}_{S_k(1)}||_F^2 + \sum_{j=1}^{\binom{k}{k-1}} a_{k-1,j}||C_{S_{k-1}(j)} - \hat{C}_{S_{k-1}(j)}||_F^2 \\
&+ ... + \sum_{j=1}^{\binom{k}{3}} a_{3,j}||C_{S_3(j)} - \hat{C}_{S_3(j)}||_F^2 + \sum_{j=1}^{\binom{k}{2}} a_{2,j}||C_{S_2(j)} - \hat{C}_{S_2(j)}||_F^2 \\
&= a_{k,1}(||C_{S_k(1)}||_F^2 - 2\rho_k(1)^2 + \rho_k(1)^2) \\
&+ \sum_{j=1}^{\binom{k}{k-1}} a_{k-1,j}(||C_{S_{k-1}(j)}||_F^2 - 2\rho_{k-1}(j)^2 + \rho_{k-1}(j)^2) \\
&+ ... + \sum_{j=1}^{\binom{k}{3}} a_{3,j}(||C_{S_3(j)}||_F^2 - 2\rho_3(j)^2 + \rho_3(j)^2) \\
&+ \sum_{j=1}^{\binom{k}{2}} a_{2,j}(||C_{S_2(j)}||_F^2 - 2\rho_2(j)^2 + \rho_2(j)^2) \\
&= (a_{k,1}||C_{S_k(1)}||_F^2 + \sum_{j=1}^{\binom{k}{k-1}} a_{k-1,j}||C_{S_{k-1}(j)}||_F^2 + ... \\
&+ \sum_{j=1}^{\binom{k}{3}} a_{3,j}||C_{S_3(j)}||_F^2 + \sum_{j=1}^{\binom{k}{2}} a_{2,j}||C_{S_2(j)}||_F^2) \\
&- [\rho_k(1)^2 + \sum_{j=1}^{\binom{k}{k-1}} a_{k-1,j}\rho_{k-1}(j)^2 + ... + \sum_{j=1}^{\binom{k}{3}} a_{3,j}\rho_3(j)^2 \\
&+ \sum_{j=1}^{\binom{k}{2}} a_{2,j}\rho_2(j)^2] \\
&= \text{constant} - [\rho_k(1)^2 + \sum_{j=1}^{\binom{k}{k-1}} a_{k-1,j}\rho_{k-1}(j)^2 + ... \\
&+ \sum_{j=1}^{\binom{k}{3}} a_{3,j}\rho_3(j)^2 + \sum_{j=1}^{\binom{k}{2}} a_{2,j}\rho_2(j)^2] \qquad (33)
\end{aligned}
$$

Therefore, to minimize equation 18, we need to maximize equation 17, and thus the equivalency holds.

■