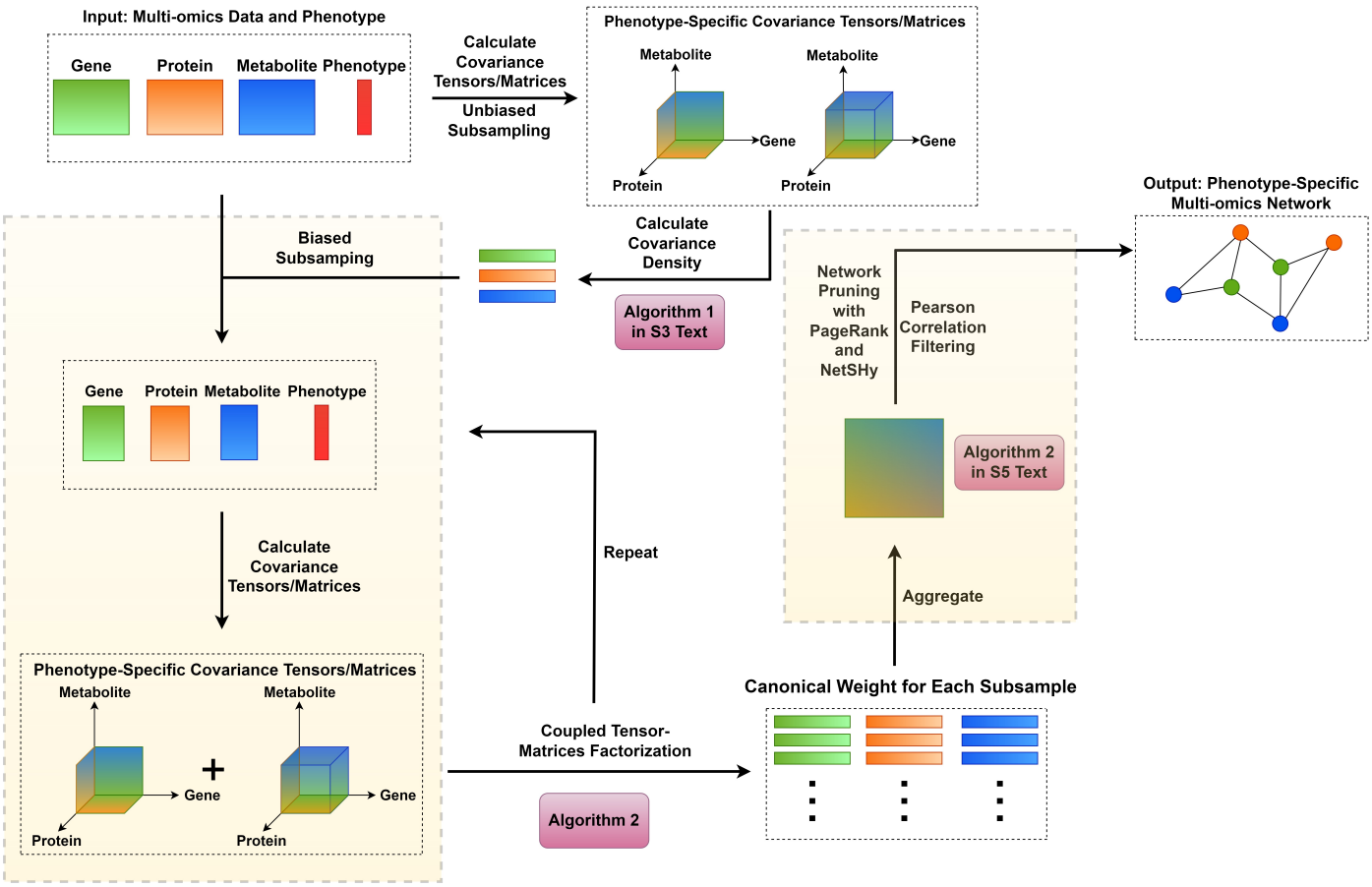
Higher-order Correlation Analysis Framework for Multi-omics Network Inference

Sparse Generalized Tensor Canonical Correlation Network Analysis (SGTCCA-Net) is a novel and powerful multi-omics network inference pipeline that identifies the higher-order relationship between molecular profiles and phenotype (Liu et al. 2024). It constructs multi-omics networks that are associated with the phenotype of interest based on all higher/lower-order correlation of interests. The preprint of this paper is now available at: https://www.biorxiv.org/content/10.1101/2024.01.22.576667v1.

It stems from a well-known method SmCCNet (Shi et al. 2019), where the multi-omics network modules are constructed with respect to phenotype of interest based on the summation of all omics-phenotype correlation and omics-omics correlation (Sparse multiple Canonical Correlation Analysis), along with adjacency matrix construction algorithm and clustering algorithm. There are various different applications of using SmCCNet to identify multi-omics modules that are associated with disease-related traits such as Chronic Obstructive Pulmonary Disease (COPD) (Mastej et al. 2020; Zhuang et al. 2021). It is initially proposed for double-omics integration with quantitative phenotype of interest, a recent upgrade (SmCCNet) enhanced the algorithm by allowing it for analyzing more than two omics data, or single omics data, along with either quantitative or binary phenotype, along with a automated pipeline that execute the complex pipeline with a single line of code (Liu et al. 2023).

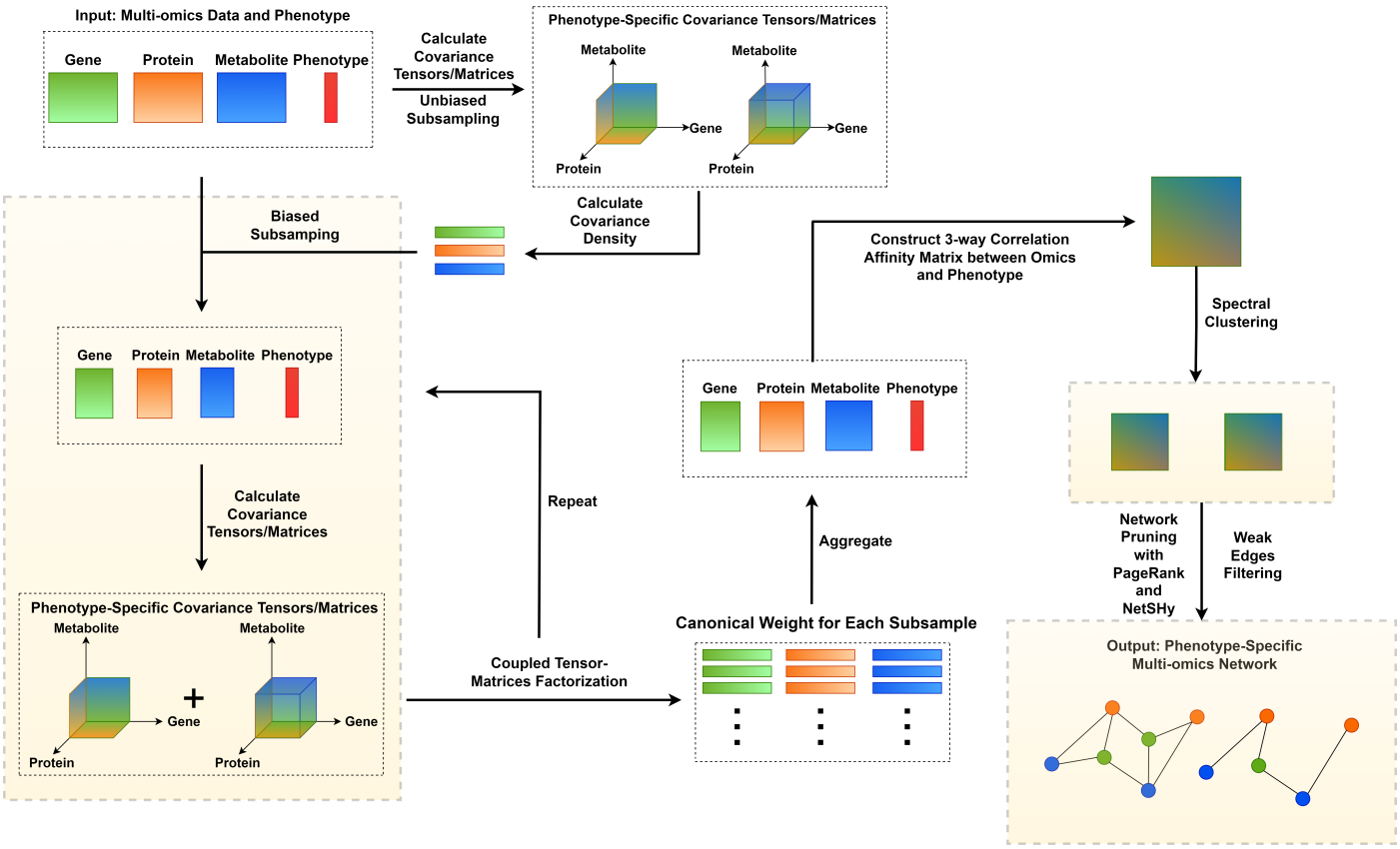


The general principle of higher-order correlation is to identify the simultaneous relationship between more than two nodes, that is, express the relationship between more than two nodes with a unifying quantifiable

measurement. The step-by-step procedure sparse generalized tensor canonical correlation analysis pipeline is given as below:

- Use sparse generalized tensor canonical correlation analysis to identify molecular features that are associated with one or more specified correlation structure of interest (Note that the sparsity is not guaranteed by penalty-based method, but a feature-wise biased subsampling algorithm that selects high correlation density features through a finite amount of iterations).
- Construct feature-wise adjacency matrix between selected molecular features, which measures the affinity between molecular features.
- Use PageRank algorithm ("PageRank Algorithm, 1998; Brin, Page," n.d.) and NetSHy network summarization score (Vu et al. 2022) to prune the adjacency matrix and obtain the final subnetwork (PageRank is used here to rank nodes based on connectivity, and NetSHy is a principal component analysis that takes network topology into account), the network pruning algorithm will yield a final network that has both a high network connectivity and a high correlation to phenotype.

This is the first generation of SGTCCA-Net, which is now available as preprint online, it sets up the frameworks, but with some steps that need to be upgraded. Therefore, we proposed an upgraded version of SGTCCA-Net, called Sparse and geNeralized Tensor Approach for Multi-omics Network Inference with Canonical Correlation Analysis (SaNTA-MoNICCA/SGTCCA-Net 2.0). This approach is currently partially developed, and there is the upgraded pipeline:



Compared to the first generation SGTCCA-Net, this method currently has two major changes:

- Global adjacency matrix construction: in the first generation of SGTCCA-Net, we keep using a design that only measures the affinity between molecular features, excluding phenotype. The second-generation SaNTA-MoNICCA construct adjacency matrix based on the 3-way correlation between molecular features and phenotype.
- Network clustering: in the first generation of SGTCCA-Net, we directly prune the global adjacency matrix, while SaNTA-MoNICCA implements spectral clustering to partition features into different network modules, and each network module may infer some specific biological processes.

We now apply this partially-built SaNTA-MoNICCA algorithm to the new born data to identify the higher-order multi-omics networks that are associated with the preterm birth in low- and middle-income countries (Espinosa et al. 2023).

Demo: Multi-Omics Integration towards Pre-term Birth

We ran our SaNTA-MoNICCA model with lipidome, metabolome, and proteome data, with the phenotype of birth time (sample-to-birth). We assume the following correlation structures:

- lipid-metabolite-protein-phenotype 4-way correlation
- lipid-metabolite-phenotype 3-way correlation
- lipid-protein-phenotype 3-way correlation
- metabolite-protein-phenotype 3-way correlation
- lipid/metabolite/protein-phenotype pairwise correlation

For simplicity, we keep all the default parameter setup for SGTCCA computation, and after the SGTCCA computation, we obtained the sparse feature-wise canonical weight vector for each molecular profile, and we execute the network analysis based on the following steps:

- Canonical weight filtering: filter out canonical weight with low values.
- Construct global adjacency matrix: global adjacency matrix is constructed based on 3-way correlation between 2 molecular features and the phenotype.
- Weak edges filtering: filter out weak edges from the global adjacency matrix.
- Spectral clustering: conduct spectral clustering with eigenheuristics to select the optimal number of clusters, and partition molecular features into different subnetwork modules
- Network prune: prune each subnetwork based on PageRank algorithm and NetSHy network summarization score.

After the network analysis step, we identified 4 final subnetwork modules, which are summarized in the table below:

Final multi-omics network summary table. The table summarize the network size, number of molecular features in each molecular profile, and the highest NetSHy summarization score correlation to birth time.

Network Modules	Network Size	Highest NetSHy Correlation	Number of Lipids	Number of Metabolites	Number of Proteins
1	208	-0.634	4	148	56
2	60	-0.262	0	60	0
3	148	0.435	0	104	44
4	233	-0.306	104	107	22

We observed from the table above that all 4 networks have relatively high correlation to the birth time, while only the second network contains nodes that are only from metabolites, and it has the lowest correlation to birth time compared to other subnetworks. In addition, the correlation between network module 1 and birth time is -0.634, which is significantly higher than other network modules.

As a demonstration of how to interpret the results, we conducted a joint enrichment analysis on all the molecular features that has the non-zero canonical weight. The platform we selected is called ImPaLA ("Impala" 2005), which conducts joint enrichment analysis between genes/proteins and metabolites/lipids. We use MetaboAnalyst (Panjeton et al. 2012) to identify the HMDBs associated with each lipid and metabolite, and use them as the input. After the enrichment analysis, there are 1379 identified pathways. We further impose a stringent filtering step to ensure there are at least 2 enriched genes/proteins and 2 enriched metabolites/lipids, resulting in 37 unique pathways, in which 13 of them are significant (FDR < 0.1). We demonstrate the top 10 of them in the following table. Specifically, the pathway of immune system is known to be related to preterm birth through many studies (Melville and Moss 2013). In addition, we found some GPCR-related pathways, and it is known to be a potential target for drug discovery towards the treatment of preterm labor (Walker et al. 2022).

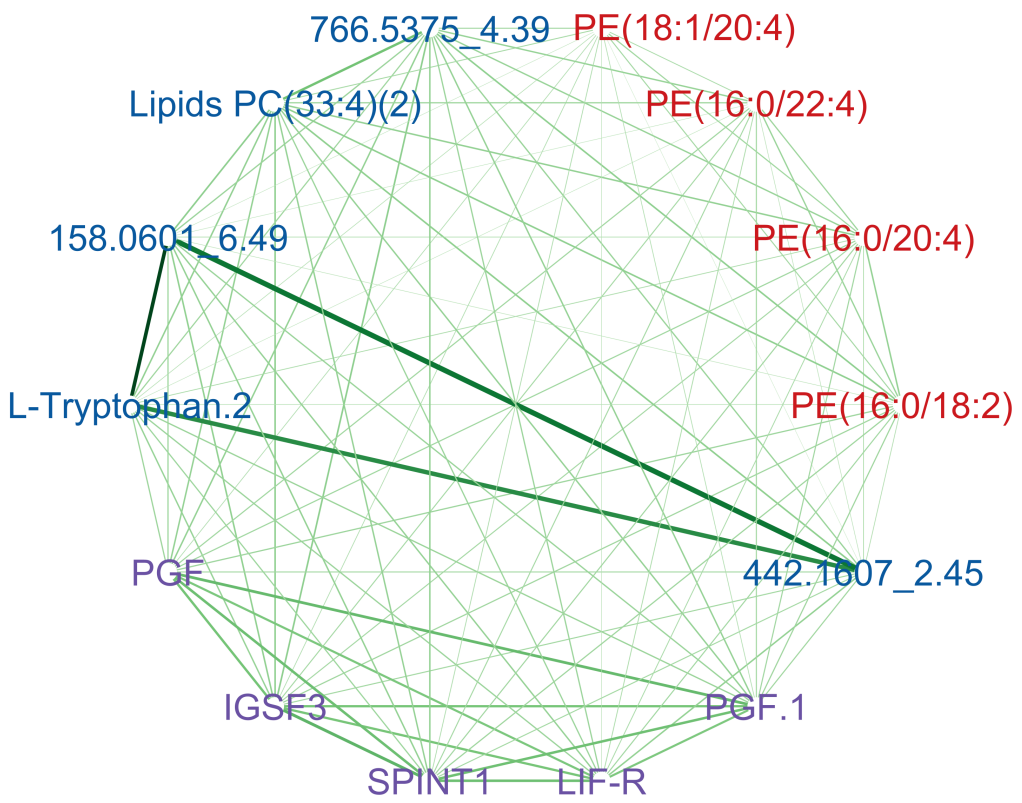
Joint enrichment analysis between metabolites, lipids, and proteins based on ImPaLA platform.

Pathway Name	Pathway Source	Enriched Proteins	Enriched Metabolites	Joint P-value
Metabolism	Reactome	13	39	1.73e-05
Metabolism of lipids	Reactome	2	23	9.64e-05
Immune System	Reactome	27	3	3.08e-03
Amino Acid metabolism	Wikipathways	2	9	6.00e-03
Metabolism of proteins	Reactome	14	11	8.94e-03
GPCR downstream signalling	Reactome	6	10	8.94e-03
Signaling by GPCR	Reactome	6	10	2.28e-02
Signal Transduction	Reactome	17	12	3.82e-02

Pathway Name	Pathway Source	Enriched Proteins	Enriched Metabolites	Joint P-value
Sudden Infant Death Syndrome (SIDS) Susceptibility Pathways	Wikipathways	2	3	6.50e-02
ADORA2B mediated anti-inflammatory cytokines production	Reactome	4	2	8.61e-02

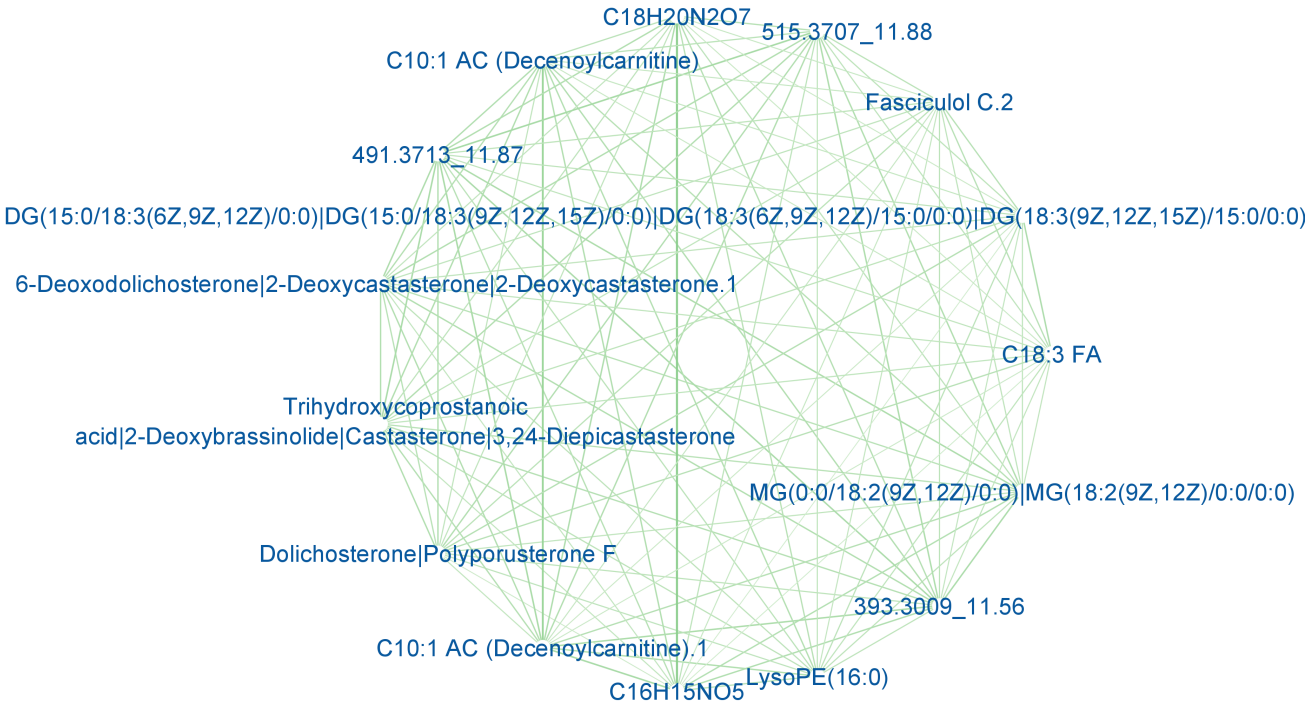
To visually see the 3-way interaction between lipids, metabolites, and proteins, for each final subnetwork module, we use PageRank to identify the top 5 molecular features in each molecular profile, and visualize them in Cytoscape (Verma and Mangla 2023) (the only exception is network module 2, where we include the top 15 metabolites from the network). Since the edge width and color depth are at the same scale across all subnetworks, we found that the first network module has more stronger connection than the other network modules. Specifically, we identified a potential strong 4-way relationship between 158.0601_6.49, L-Tryptophan, and 442.1607_2.45, in which 2 of them are unidentified metabolites and may potentially be the novel findings. Interestingly, we found that L-tryptophan only has a 0.28 correlation with birth time, but coupling it with these two other metabolites makes a higher correlation with birth time.

Network Module 1 Visualization



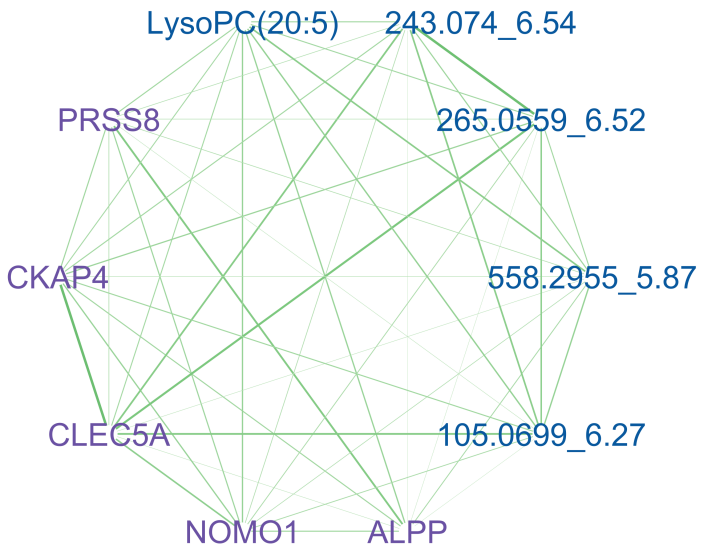
Three-way interaction between top molecular features and birth time for network module 1. Each network includes the top 5 molecular features from each molecular profile (less than 5 if there are not enough molecular features in the final subnetwork for certain molecular profile). Red features are lipids, purple features are proteins, and blue features are metabolites. The width and depth of color for edges represent their relative strength.

Network Module 2 Visualization



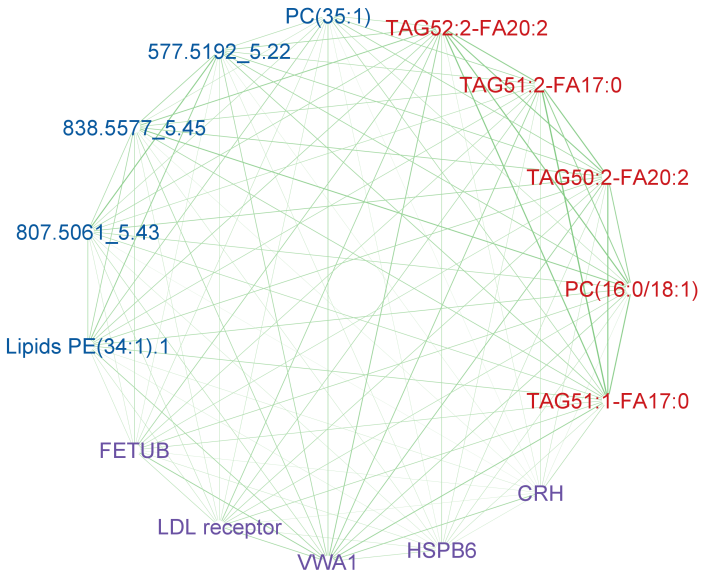
Three-way interaction between top molecular features and birth time for network module 2. Each network includes the top 5 molecular features from each molecular profile (less than 5 if there are not enough molecular features in the final subnetwork for certain molecular profile). Red features are lipids, purple features are proteins, and blue features are metabolites. The width and depth of color for edges represent their relative strength.

Network Module 3 Visualization



Three-way interaction between top molecular features and birth time for network module 3. Each network includes the top 5 molecular features from each molecular profile (less than 5 if there are not enough molecular features in the final subnetwork for certain molecular profile). Red features are lipids, purple features are proteins, and blue features are metabolites. The width and depth of color for edges represent their relative strength.

Network Module 4 Visualization



Three-way interaction between top molecular features and birth time for network modules 4. Each network includes the top 5 molecular features from each molecular profile (less than 5 if there are not enough molecular features in the final subnetwork for certain molecular profile). Red features are lipids, purple features are proteins, and blue features are metabolites. The width and depth of color for edges represent their relative strength.

Potential Downstream Applications

Due to the limited accessbility to the complete data, many downstream analysis wouldn't be able to run. But there are some potential applications. For instance, the network summarization score can be used to identify the mediation effect of cell types or BMI on the relationship between multi-omics networks and birth time. If genomics data is available, then network summarization score can also be used to conduct GWAS or network-QTL analysis. In addition, enrichment analysis can also be conducted through different platform such as Metascape (Zhou et al. 2019). Furthermore, each subnetwork can also be used to predict the phenotype. I'm currently developing a multi-view XGBoost to achieve this task.

This is the initial version of the SGTCCA-Net 2.0/SaNTA-MoNICCA, it is providing some fundamental ways to construct phenotype-specific higher-order multi-omics networks. The next version of it will have more functionalities such as longitudinal multi-omics network, network comparison algorithm, and graph fusion algorithm to identify and separate unique 4-way, 3-way, and pairwise correlation.

Final Subnetwork Files Output Structure

- correlation_sub: Pearson's correlation matrix between molecular features.
- M: adjcency matrix between molecular features, each edge is the 3-way correlation between 2 corresponding molecular features and the phenotype.
- omics_correlation_data: univariate feature correlation to phenotype
- pc_correlation: correlation between the first 3 NetSHy PC and phenotype
- pc_loading: NetSHy PC loading
- pca_x1_pc1: NetSHy summarization score and phenotype data.
- mod_size: final network module size
- rank_value: PageRank score for each molecular feature (sorted)
- sub_type: type of data for each molecular feature

Code Availability

The source code of SGTCCA-Net and the code of SGTCCA-Net application to preterm labor data is available at https://github.com/liux4283/SparseGTCCANet. The final subnetwork result is also available at this repository. The network pruning algorithm is available at the SmCCNet package, which is available at GitHub and CRAN.

References

Espinosa, Camilo A., Waqasuddin Khan, Rasheda Khanam, Sayan Das, Javairia Khalid, Jesmin Pervin, Margaret P. Kasaro, et al. 2023. "Multiomic Signals Associated with Maternal Epidemiological Factors Contributing to Preterm Birth in Low- and Middle-Income Countries." *Science Advances* 9 (21). https://doi.org/10.1126/sciadv.ade7692.

"Impala." 2005, April. https://doi.org/10.1093/acref/9780195301731.013.41789.

Liu, Weixuan, Katherine A. Pratte, Peter Castaldi, Craig Hersh, Russell P. Bowler, Farnoush Banaei-Kashani, and Katerina Kechris. 2024. "A Generalized Higher-Order Correlation Analysis Framework for Multi-Omics Network Inference." http://dx.doi.org/10.1101/2024.01.22.576667.

Liu, Weixuan, Thao Vu, Iain R Konigsberg, Katherine A Pratte, Yonghua Zhuang, and Katerina J Kechris. 2023. "SmCCNet 2.0: An Upgraded r Package for Multi-Omics Network Inference." http://dx.doi.org/10.1101/2023.11.20.567893.

Mastej, Emily, Lucas Gillenwater, Yonghua Zhuang, Katherine A. Pratte, Russell P. Bowler, and Katerina Kechris. 2020. "Identifying Proteinmetabolite Networks Associated with COPD Phenotypes." *Metabolites* 10 (4): 124. https://doi.org/10.3390/metabo10040124.

Melville, Jacqueline M., and Timothy J. M. Moss. 2013. "The Immune Consequences of Preterm Birth." *Frontiers in Neuroscience* 7. https://doi.org/10.3389/fnins.2013.00079.

"PageRank Algorithm, 1998; Brin, Page." n.d. https://doi.org/10.1007/springerreference_57796.

Panjeton, Geoffrey D., Matthew A. Remz, Hannah J. Allen, David H. Powell, and Peggy R. Borum. 2012. "Use of MetaboAnalyst as a Tool to Study Metabolomics Data for a Dietary Intervention Study." *The FASEB Journal* 26 (S1). https://doi.org/10.1096/fasebj.26.1_supplement.637.2.

Shi, W Jenny, Yonghua Zhuang, Pamela H Russell, Brian D Hobbs, Margaret M Parker, Peter J Castaldi, Pratyaydipta Rudra, et al. 2019. "Unsupervised Discovery of Phenotype-Specific Multi-Omics Networks." Edited by Inanc Birol. *Bioinformatics* 35 (21): 4336–43. https://doi.org/10.1093/bioinformatics/btz226.

Verma, Srishti, and Lakshay Mangla. 2023. "Network Analysis Using Cytoscape." http://dx.doi.org/10.21203/rs.3.rs-2487773/v1.

Vu, Thao, Elizabeth M Litkowski, Weixuan Liu, Katherine A Pratte, Leslie Lange, Russell P Bowler, Farnoush Banaei-Kashani, and Katerina J Kechris. 2022. "NetSHy: Network Summarization via a Hybrid Approach Leveraging Topological Properties." Edited by Pier Luigi Martelli. *Bioinformatics* 39 (1). https://doi.org/10.1093/bioinformatics/btac818.

Walker, Abigail R., Camilla B. Larsen, Samit Kundu, Christina Stavrinidis, Sung Hye Kim, Asuka Inoue, David F. Woodward, et al. 2022. "Functional Rewiring of G Protein-Coupled Receptor Signaling in Human Labor." *Cell Reports* 40 (10): 111318. https://doi.org/10.1016/j.celrep.2022.111318.

Zhou, Yingyao, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K. Chanda. 2019. "Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets." *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-09234-6.

Zhuang, Yonghua, Brian D Hobbs, Craig P Hersh, and Katerina Kechris. 2021. "Identifying miRNA-mRNA Networks Associated with COPD Phenotypes." *Frontiers in Genetics* 12 (October). https://doi.org/10.3389/fgene.2021.748356.