# Construction of Multi-Modal Transcriptome-Small Molecule Interaction Networks from High-Throughput Measurements to Study Human Complex Traits

*Vaha Akbary Moghaddam[1], Sandeep Acharya[2], Michaela Schwaiger-Haber[3], Shu Liao[4], Wooseok J. Jung[4], Bharat Thyagarajan[5], Leah P. Shriver[3], E. Warwick Daw[1], Nancy L. Saccone[1], Ping An[1], Michael R. Brent[1,4], Gary J. Patti[3], Michael A. Province[1*]*

[1] *Department of Genetics, School of Medicine, Washington University in St. Louis, MO, USA*
[2] *Division of Computational & Data Sciences, McKelvey School of Engineering, Washington University in St. Louis, MO, USA*
[3] *Department of Chemistry, School of Arts & Sciences, Washington University in St. Louis, MO, USA*
[4] *Department of Computer Science & Engineering, McKelvey School of Engineering, Washington University in St. Louis, MO, USA*
[5] *Department of Laboratory Medicine & Pathology, School of Medicine, University of Minnesota, MN, USA*

## Abstract

Small molecules (SMs) are integral to biological processes, influencing metabolism, homeostasis, and regulatory networks. Despite their importance, a significant knowledge gap exists regarding their downstream effects on biological pathways and gene expression, largely due to differences in scale, variability, and noise between untargeted metabolomics and sequencing-based technologies. To address these challenges, we developed a multi-omics framework comprising a machine learning-based protocol for data processing, a semi-supervised network inference approach, and network-guided analysis of complex traits. The ML protocol harmonized metabolomic, lipidomic, and transcriptomic data through batch correction, principal component analysis, and regression-based adjustments, enabling unbiased and effective integration. Building on this, we proposed a semi-supervised method to construct transcriptome-SM interaction networks (TSI-Nets) by selectively integrating SM profiles into gene-level networks using a meta-analytic approach that accounts for scale differences and missing data across omics layers. Benchmarking against three conventional unsupervised methods demonstrated the superiority of our approach in generating diverse, biologically relevant, and robust networks. While single-omics analyses identified 18 significant genes and 3 significant SMs associated with insulin sensitivity (IS), network-guided analysis revealed novel connections between these markers. The top-ranked module highlighted a cross-talk between fiber-degrading gut microbiota and immune regulatory pathways, inferred by the interaction of the protective SM, N-acetylglycine (NAG), with immune genes (*FCER1A*, *HDC*, *MS4A2*, and *CPA3*), linked to improved IS and reduced obesity and inflammation. Together, this framework offers a robust and scalable solution for multi-modal network inference and analysis, advancing SM pathway discovery and their implications for human health. Leveraging data from a population of thousands of individuals with extended longevity, the inferred TSI-Nets demonstrate generalizability across diverse conditions and complex traits. These networks are publicly available as a resource for the research community.

## Introduction

The rapid advancements in high-throughput mass spectrometry and the emergence of metabolomics as a key field in omics research, have uncovered new and complex roles of small molecules (SMs) in biological processes and their impact on health and diseases [1]. SMs, typically defined as organic compounds with a molecular weight of less than 1,500 Da, can be classified based on their origin as either endogenous or exogenous [2][3]. Endogenous SMs, also known as primary metabolites, are synthesized internally through an organism's metabolic processes [4], whereas exogenous SMs, often referred to as chemical exposures, encompass a wide range of substances such as medications, dietary supplements, pollutants, and more [5][6].

SMs play a pivotal role in biological systems, influencing processes such as metabolism, homeostasis, and the regulation of proteins and gene expression [7][8]. Their effect is predominantly mediated through interactions with biological macromolecules, notably proteins [9]. These interactions can be physical, such as enzymatic reactions or regulatory interactions [10][11]. However, in spite of the extensive research on physical protein-SM interactions, significant gaps remain in our understanding of the downstream effects of SMs on biological pathways and the modulation of gene expression [12][13]. Recent studies utilizing RNA-sequencing (RNA-seq) have, in fact, demonstrated diverse differential gene expression patterns upon administration of metabolic compounds in animal models [14-16].

Multi-omics approaches have emerged as a new paradigm in systems biology. Omics integration has been successful in improving the accuracy of predictive models for clinical outcomes [17], increasing statistical power in biomarker discovery, and identifying biological pathways and networks influencing human complex traits. Systematic integration of multiple omics layers-such as epigenomic, transcriptomic, and proteomic profiles- with SM profiles and chemical information has shown success in predicting patient drug responses and personalized drug repurposing for a range of clinical outcomes [21][22]. In more recent years, computational and instrumental advances in untargeted metabolomics have facilitated the integration of high-throughput SM profiles with sequencing-based technologies. Prominent metabolomics platforms, such as MetaboAnalyst and XCMS, now support omics integration at the pathway level for pathway-outcome relationships [23][24]. Beyond pathway-level integration, network inference from SM and gene expression profiles provides a holistic and often novel view on gene-SM relationships. Such networks can be knowledge-guided, inferred via aggregating protein-SM interactions across multiple pathway databases [25]. Alternatively, interaction networks can also be data-driven, constructed through statistical modeling of metabolomic and genomic profiles [23][26].

Recent multi-omics measurements in large-scale human population studies, such as the NIA’s Long Life Family Study (LLFS) or NHLBI’s Framingham Heart Study (FHS), have successfully identified individual SM markers for various complex traits through metabolome-wide association studies (MWAS) conducted on thousands of samples [27][28]. However, integrating SM profiles with other omics data types remains challenging and is subject to major limitations. Untargeted metabolomics profiles are generally measured by liquid chromatography / mass spectrometry (LC/MS), which captures the mass-to-charge ratios (m/z) of compounds [29]. In contrast, most other omics measurements, such as RNA-seq, are based on sequencing technologies. These fundamentally different technologies produce distinct distributions, variabilities, and noise characteristics. These differences can lead to model overfitting and false associations during omics integration if not properly accounted for [30][31]. Furthermore, RNA-seq experiments generally captures at least over 10,000 gene expression profiles, while most LC/MS experiments capture around 1,000 SMs. This disparity in scale can introduce bias, favoring the larger dataset when jointly modeling omics data [32].

Lastly, when it comes to network inference, it is crucial to assess the reliability and biological interpretability of these multi-modal networks. Previous approaches have predominantly relied on pathway annotations from databases such as KEGG for network evaluation and biological interpretation [33][26][34]. While effective, these annotations are biased toward well-characterized compounds, leaving a significant knowledge gap for many newly identified metabolites and chemical exposures. Additionally, many metabolic processes are tissue-specific. However, LC/MS measurements in large-scale human studies typically use easily accessible biospecimen, such as blood, which may not accurately reflect the full range of metabolic status [35][36]. Therefore, there remains the need for unbiased and generalizable approaches to evaluate these multi-modal networks and increase their interpretability.

To address these challenges, we propose a comprehensive framework for integrating untargeted LC/MS profiles with transcriptomics data to construct Transcriptome-SM Interaction Networks (TSI-Nets), and apply them to study human complex traits. Using the data of the NIA's LLFS, one of the largest human multi-omics studies to date [37], the framework includes steps for processing LC/MS and RNA-seq profiles for omics integration, multi-modal network inference, and network-guided analysis of complex traits. Our framework enables the discovery of biologically meaningful interactions and provides insights into the intricate relationships between genes, SMs, and their contribution to metabolic health.

## 2. Materials & Methods

### *2.1. Participants*

Procedures and criteria for eligibility and recruitment of the LLFS participants are described in detail by *Wojczynski et al* [37]. For this study, data from the first clinical exam containing 4953 total participants from 539 families was used. Glucose (mg/dL), insulin (pmol/L), total triglyceride (TG; mg/dL), interleukin 6 (IL6; pg/mL), hemoglobin (g/dL) and glycosylated hemoglobin were measured by the LLFS central laboratory at the University of Minnesota. Participants taking diabetic medications or diagnosed with type 2 diabetes mellitus (T2DM) characterized by fasting glucose levels >= 126 mg/dL or glycosylated hemoglobin >= 6.5% were excluded. Additionally, all non-diabetic participants with fasting time < 8h were also excluded to avoid any metabolic bias. Insulin sensitivity (IS) was calculated by HOMA2 software using fasting and glucose measurements [38]. BMI was calculated as weight (kg) / height ($m^2$). All traits were adjusted for age, age-squared ($age^2$), age-cubed ($age^3$), sex, clinical field centers, and the top 20 genetic principle components (PCs) using a stepwise regression model. IS, TG, and IL6 were also ln-transformed prior to covariate adjustments.

## *2.2. LC/MS workflow*

In the LC/MS workflow, each batch generally consisted of 92 research samples, 2 quality control (QC) samples, and 2 blank samples. QC samples were prepared by pooling a subset of the research samples. Peak lists were generated through MS feature detection, background subtraction, and adduct selection. Following peak list generation, compounds were identified by annotating mass, MS/MS fragmentation patterns, and retention times to both in-house and online libraries. Peak areas were then obtained for the identified compounds. To account for technical variability in raw peak areas, a random forest-based method was applied [29], leveraging QC variability within each batch. These peak areas, adjusted for technical variability, were further processed to account for potential biological confounders. The variables include chronological age, age$^2$, sex, smoking status, and medication usage. A stepwise regression model coupled with principal component analysis (PCA) were used for covariate adjustment. All steps in the workflow were conducted separately for polar SMs and lipids. Detailed protocol information is provided in the Supplementary Text S1.

## *2.3. RNA-seq protocol*

RNA extraction and sequencing were performed by the McDonnell Genome Institute (MGI) at Washington University. Total RNA was obtained from PAXgene™ Blood RNA tubes through the Qiagen PreAnalytiX PAXgene Blood miRNA Kit (Qiagen, Valencia, CA). RNA-Seq data processing was carried out by version 3.3 of the nf-core/RNASeq pipeline with STAR/RSEM, applying default parameters (https://zenodo.org/records/5146005). Genes with fewer than three counts per million in over 98.5% of samples were excluded, and samples with more than 8% intergenic reads were also removed. The remaining data were transformed by the variance stabilizing transformation (VST) function in DESeq2 [39]. The transformed gene expression levels were then adjusted through a stepwise regression model for age, age$^2$, sex, field centers, percent of intergenic reads, and the counts of red blood cells, white blood cells, platelets, monocytes, and neutrophils as baseline covariates. Furthermore, RNA-seq batch information and the top 10 principal components (PCs) of gene expression were incorporated into the model as additional covariates.

## *2.4. Multi-modal network inference*

### *2.4.1. Conventional methods*

Traditionally, network inference in single-omic profiles (e.g., RNA-seq) relies on unsupervised learning, where feature relationships (commonly refer to as edge weights) are first quantified using similarity metrics. An unsupervised clustering algorithm is then applied to the resulting edge weight matrix to define modules of densely connected features. We refer to this as the 'conventional' network inference approach in this study.

In the conventional methods applied here, adjusted SM profiles (comprising polar metabolites, lipids, and identified exposures) were combined with adjusted gene expression profiles from the start of the network inference workflow. Edge weights of gene-gene, gene-SM, and SM-SM pairs were computed using two established methods: WGCNA [40] and GENIE3 [41]. For WGCNA, the optimal soft-threshold (β = 2.3) was selected based on an approximate scale-free topology index of 0.9 and mean connectivity of 15. The resulting adjacency matrix was converted into a topological overlap matrix. For GENIE3, a forest-based model was applied with default parameters.

Following edge weight computation, clustering was performed using MONET software, which implements two top-performing methods from the 2019 DREAM Challenge for gene-level module detection: modularity optimization (MO) and kernel-based (KB) clustering [42][43]. The conventional network inference workflow thus consists of WGCNA and GENIE3 in combination with KB and MO clustering approaches. A detailed description of these methods is provided in Supplementary Text S2.

### *2.4.2. Semi-supervised approach*

The proposed model combines the conventional unsupervised network inference with a supervised approach to selectively integrate SM and RNA-seq profiles, creating multi-modal networks. First, gene-level modules were constructed from the adjusted RNA-seq data of the LLFS using conventional approaches as described earlier. These modules serve as a baseline for the supervised integration of SM profiles, guided by statistical associations and a meta-analysis strategy. Upon construction of gene-level clusters, SMs were connected to the genes in modules based on a two-step significance assessment.

Initially, associations between each SM and individual gene across co-expression modules were calculated using linear mixed models (LMMs). After multiple testing correction for SM associations with each gene in the co-expression network, SMs with significant associations were directly connected to their corresponding genes. In the second step, a correlated meta-analysis (CMA) framework was applied based on the gene-SM associations, leveraging the co-expression structure of modules to further refine the integration of SMs. For each SM, CMA combines gene-SM associations of genes within each module, while accounting for the inter-correlation of these genes using a variance-covariance matrix derived from gene-SM associations [44]. To perform CMA, a pre-defined gene-SM association threshold (p-value = 0.0025) was used to select nominally significant SMs for CMA. SMs with significant meta-analytic p-value after multiple testing correction were then connected to the genes in each module. The detailed statistical and algorithmic framework of the semi-supervised approach is described in the Supplementary Text S3.

### *2.4.3. Knowledge-guided network inference*

In addition to the semi-supervised multi-modal network inference based on data-driven coexpression networks from the LLFS data, knowledge-guided TSI-Nets were constructed using external and independent sources. Specifically, protein-protein interaction (PPI) networks from STRINGdb and InWeb, as well as coexpression networks derived from GEO were used as the baseline gene-level networks. Modules of these networks were selected from those generated by the winners of the 2019 DREAM challenge for unsupervised network clustering [42]. SMs were then integrated into these baseline networks using the same two-step semi-supervised approach described earlier.

## *2.5. TSI-Net inference evaluation*

To benchmark network inference approaches, we evaluated TSI-Nets based on external biological evidence and network diversity metrics. Independent knowledge sources were used to assess the biological relevance of inferred modules. For functional validation, metabolic-set enrichment analysis (MSEA) was performed using MetaboAnalyst 6.0 [23]. Additionally, Gene Ontology (GO) enrichment analysis was conducted using GOATOOLS [45]. To further assess the biological coherence of modules, gene-gene interaction support was evaluated using STRINGdb PPI networks [46], where support was defined as the proportion of genes within a module that interact with at least one other gene in the same module. Similarly, transcription factor (TF) co-regulation support was assessed using human tissue-specific gene regulatory networks (GRNs) for blood cell lineages [47], where support was defined as the proportion of genes sharing a common TF with at least one other gene in the module. Lastly, the diversity of TSI-Nets was evaluated by examining the number of modules, as well as the number of genes and SMs participating in the networks.

### *2.6. "Omics-wide" and network association tests*

LMMs were used for all genome-wide, transcriptome-wide, metabolome-wide, and lipidome-wide association tests to account for familial relatedness based on the LLFS kinship matrix (details in Supplementary Text S3). To control for inflation factors for each "omics"-wide association study (OWAS), the BACON method was applied if the inflation factor ($\lambda$) $\geq 1.2$ [48]. Lastly, to test the association of multi-modal modules with complex traits, Pascal was used [49]. This method computes the sum of chi-squared statistics upon ranking all omics units across the entire network based on their significance for a complex trait.

### *2.7. Framingham Heart Study*

Replication of the transcriptomic analysis for each trait was conducted using data from the Framingham Heart Study (FHS) cohort. FHS is a multi-generational, family-based study investigating genetic, molecular, and environmental factors influencing cardiovascular and related traits [50]. For this study, we utilized data from the second examination of the third-generation cohort, which includes the largest number of RNA-seq measurements available. Eligible participants were selected based on the same criteria as those used for the LLFS cohort, and all traits were adjusted in a manner consistent with the methods outlined in Section 2.1. RNA-seq measurements were processed and adjusted following the same method described in Section 2.3. After applying these criteria and adjustments, a maximum of 1,248 subjects were included in the transcriptomic analysis.

## 3. Results

### *3.1. Overview of Multi-Modal TSI-Net Construction Framework*

This study is structured into three main stages to build a comprehensive multi-omics integration framework with applications for complex human traits (Figure 1):

**1. LC/MS data processing:** To address the challenges of integrating metabolomics and transcriptomics data, we developed a ML-based protocol for processing untargeted LC/MS profiles in conjunction with RNA-seq data. This protocol leverages ML-based batch effect correction, PCA, and regression-based adjustments for biological confounders to harmonize multi-omics data while preserving true biological signals.

**2. Transcriptome-SM interaction network (TSI-Net) inference:** Building on the processed RNA-seq and LC/MS profiles, we benchmarked conventional unsupervised network inference methods against a proposed semi-supervised approach for constructing TSI-Nets. In the semi-supervised framework, homogeneous gene-level baseline networks are built from RNA-seq profiles and SM profiles are selectively integrated into these networks based on statistical associations and meta-analysis. The proposed method ensures meaningful multi-modal network construction by accounting for differences in scale, variability, and missing data across the omics layers. Additionally, it enables the integration of SM profiles into the established gene networks, such as STRINGdb PPI networks or GEO coexpression networks, to construct knowledge-guided TSI-Nets.

**3. Application of TSI-Nets for analysis of complex traits:** To demonstrate the applicability of the established TSI-Nets for biological interpretations, they were applied to study complex metabolic traits, IS, BMI, TG, and IL6. Network association tests and module-level gene-SM interaction analyses demonstrated significant gene-SM interactions relevant to metabolic health. In particular, the top-ranked modules revealed a novel interplay between gut fiber-degrading microbiome metabolism and immune regulation, that is protective for metabolic health.

## *3.2. Enhanced multi-omics integration with LC/MS data processing protocol*

### *3.2.1. Data distribution*

The proposed LC/MS protocol is designed to account for technical, demographic, and biological confounders. It showed marked improvement in data quality, making it suitable for integration with sequencing-based technologies (e.g., RNA-seq) and complex trait analysis. Initially, raw LC/MS peak areas exhibited a highly skewed exponential distribution with a sharp decay, characterized by high kurtosis and standard deviation (SD) (Figure 2a). Following the processing pipeline, SM peak residuals approximated a normal distribution, with most SMs exhibiting significantly reduced kurtosis and SD values constrained between 0 to 1 (Figure 2a). Moreover, the processed LC/MS data aligned the SM profiles with the distribution patterns and ranges of adjusted RNA-seq data and complex traits, despite no direct influence from either category during processing (Figure 2a). This alignment is critical for unbiased data integration in downstream analyses [51]. A similar pattern was observed for SMs in the lipid category (Supplementary Figure 1).

While corrections for technical variables reduced the SD of raw peak areas, this alone proved insufficient for preparing data for multi-omics integration (Supplementary Figure 1), whereas accounting for demographic and biological confounders played a pivotal role in achieving high-quality data suitable for integration. Additional data distribution patterns across different SM categories are detailed in Supplementary Figure 1.

### *3.2.2. Heritability analysis*

Heritability ($h^2$) analysis of the fully processed lipid and polar SMs was conducted using SOLAR [52]. Across both lipid and polar SM categories, strong $h^2$ was observed ($h^2$ = 0.30 for lipids and $h^2$ = 0.294 for polars; Figure 2b), which is consistent with findings from other populations [53]. $h^2$ patterns varied among lipid categories, with most lipids exhibiting high $h^2$ (average $h^2$ = 0.34). However, triglycerides showed relatively lower $h^2$ (average $h^2_{triglycerides}$ = 0.21; Figure 2b). This reduced $h^2$ may reflect a greater environmental contribution for triglycerides, as they are strongly influenced by diet and lifestyle [54]. Supporting this, we also observed a strong correlation between BMI and TG levels among the LLFS participants (r = 0.32; Supplementary Figure 2c). Notably, similar $h^2$ patterns of lipids were observed during the second clinical visit of the LLFS (Supplementary Figure 3), further validating the consistency of the findings.

For polar SMs, compounds derived from dietary, gut microbiome, endogenous, or mixed dietary/endogenous sources exhibited high heritability overall (Figure 2b). Conversely, most drugs profiled in the LLFS cohort showed lower heritability (average $h^2_{drugs}$ = 0.198). Interestingly, chemical exposures also demonstrated heritability patterns similar to those of endogenous and dietary SMs, which might be due to environmental similarities within families and the influence of genetic variation on exposure metabolism [55]. The consistent and relatively strong heritability observed across various SM categories and clinical visits highlights the reliability of the LC/MS data measurement and processing protocols.

### *3.2.3. Stable OMICS integration following data processing*

Following LC/MS data processing, gene-SM association tests were performed to evaluate p-value distribution patterns and inflation factors (λ) across the association scans. As shown in Figure 2c, gene-SM associations based on the processed data exhibited a distribution that closely aligned with uniform expectations. Notably, most scans centered around λ = 1, indicating stable integration of the omics profiles. This represents a remarkable improvement over raw omics profiles, which showed substantial deviations from uniformity (Figure 2c). For trait-gene and trait-SM associations, larger inflation factors were observed. However, applying the BACON approach effectively corrected this inflation, resulting in p-value distributions that better adhere to the expected uniform pattern under the null hypothesis (Supplementary Figure 4). The combination of robust data processing across the omics scans and inflation control enables stable multi-omics integration and reliable association testing.

## *3.3. Semi-supervised network inference outperforms conventional approaches*

### *3.3.1. Benchmarking TSI-Net construction: Support from independent knowledge*

Following the data processing protocol, the processed LC/MS and RNA-seq profiles were integrated using the methods described in Section 2.4 to construct TSI-Nets. Data-driven networks of the LLFS generated by conventional approaches (WGCNA-MO, WGCNA-KB, and GENIE3-KB) were benchmarked against the proposed semi-supervised model based on CMA. Seven evaluation metrics were employed, focusing on support for multi-modal components from independent knowledge sources and network diversity.

MSEA of SMs across multi-modal modules demonstrated the superior performance of the CMA model, with 57.47% of modules showing enrichment for at least one metabolic term, compared to 38.46% for the second best-performing method, WGCNA-MO. (Figure 3a). Next, Gene-gene interactions of TSI-Net modules were matched against STRINGdb PPI networks. The CMA model exhibited the highest degree of support, with an average of 59.33% of gene-level interactions within modules supported by STRINGdb (Figure 3b). The CMA model similarly excelled in capturing TF co-regulation of genes in the TSI-Net modules, as assessed using blood-specific GRNs. On average, 62.85% of genes across the modules were supported by TF co-regulation (Figure 3c). However, for GO enrichment analysis, WGCNA-MO emerged as the best-performing approach, with 48.71% of modules showing enrichment for at least one GO term. The CMA model followed as the second-best method with 24% enrichment for GO terms (Figure 3d).

### *3.3.2. Benchmarking TSI-Net construction: Network diversity*

Next, we evaluated the diversity of TSI-Nets in terms of gene and SM representation, as well as the total number of inferred modules. The CMA model provided far superior performance compared to all conventional approaches and captured a broader range of gene-SM interactions. TSI-Net inferred by CMA retains approximately 99% of all SMs and 96% of all genes profiled in the LLFS within modules. For genes, this demonstrates an almost 38-fold increase compared to the second-best method for independent support, WGCNA-MO, which retained only 2.5% of genes (Figure 3e). Similarly, SM representation within TSI-Nets inferred by conventional approaches ranged from 38% (GENIE3-KB) to 74% (WGCNA-KB) (Figure 3f). Lastly, The CMA model inferred a total of 388 multi-modal modules across the entire network—an almost 10-fold increase compared to WGCNA-MO (Figure 3g). The diversity metrics, combined with the external support for genes and SMs across TSI-Nets, demonstrate the ability of the CMA approach to vastly improve network coverage and representation, while increasing their independent support.

### *3.3.3. TSI-Net properties*

After benchmarking TSI-Nets inferred by different methods, we examined the properties of the networks inferred by the CMA model in greater detail. TSI-Nets inferred by CMA demonstrated consistent integration of SMs into the baseline gene-level networks regardless of their source. On average, SMs constitute ~21% and 16% of the modules in LLFS-derived and knowledge-guided networks, respectively (Figure 3a). In contrast, modules inferred by the conventional methods exhibit a higher proportion of SMs (Supplementary Figure 5). This difference is primarily due to the smaller number of modules and reduced gene diversity in networks constructed by conventional methods (Figure 2e & 2g), limiting their capacity to distribute SMs across diverse modules.

SMs within modules exhibit strong intra-module coherence, despite SM inclusion in gene-level modules being based solely on their relationships with genes. Module-wise PCA on SMs revealed that, on average, PC1 explains 44% of the variance among SMs within TSI-Net modules, which is greater than the variance explained by PC1 for genes (Figure 3b). This result aligns with our expectations, as SMs generally constitute a smaller proportion of modules, and the variance explained by PC1 is inversely correlated with the number of SMs in a module (Supplementary Figure 6). Additionally, the PCA patterns observed in knowledge-guided TSI-Nets were similar and consistent to those in the LLFS TSI-Nets, regardless of the sources of the baseline gene-level modules (STRING, InWeb, and GEO; Supplementary Figure 7). Pairwise SM-SM association tests further demonstrated that SMs within TSI-

Net modules are frequently associated with one another. Specifically, 91% of modules showed an average -$\log_{10}$(p-value) > 1.3 (corresponding to p-value < 0.05 ) for SM-SM associations (Figure 3c).

Gene-SM connections across modules showed consistent interaction patterns in LLFS and knowledge-guided TSI-Nets (Supplementary Figures 8 & 9) with the median number of genes interacting with a given SM being 13 in LLFS networks and 14 in knowledge-guided networks. While these interactions are distributed across different modules, they exhibit an underlying biological relevance supported by independent knowledge sources. On average, 38.24% of genes interacting with an SM also interact with each other based on STRINGdb PPI networks. Additionally, 58.9% of genes share TF co-regulation with other genes interacting with the same SM across all modules (Figure 3d), further highlighting the effectiveness of the semi-supervised integration of SMs within gene-level networks.

### *3.4. Single Omic and Network Association Studies on Insulin Sensitivity*

Upon developing and assessing methodologies for data processing and multi-modal network inference, the established TSI-Nets were applied to investigate the molecular signatures of IS and other metabolic traits. To this end, we first conducted single-omics association analyses, followed by network-guided investigation of key genes-SM interactions and their relationship with metabolic health.

#### *3.4.1. Transcriptome-wide association study (TWAS)*

TWAS identified 18 significant genes associated with IS, 15 of which were successfully replicated in the FHS cohort with consistent regression coefficients as those from the LLFS cohort (Table 1). While none of these genes have been previously reported for IS in TWAS literature, 6 genes (*CPA3*, *GATA2*, *HDC*, *MS4A2*, *AKAP12*, and *PTGER2*) were previously identified as significant markers of fasting glucose or insulin measurements according to the TWAS Atlas [56]. GO enrichment analysis did not identify any significant biological processes or functions associated with these 18 genes. However, 7 genes (*FCER1A*, *CPA3*, *GATA2*, *HDC*, *SLC45A3*, *MS4A2*, and *ENPP3*) were found to interact with each other based on the STRING PPI network, with all of them coherently exhibiting positive coefficients for IS, suggesting a protective relationship. The TWAS QQ-plot for IS is provided in Supplementary Figure 4a.

#### *3.4.2. Metabolome-wide association study (MWAS)*

MWAS identified three SMs significantly associated with IS. Two polar metabolites: N-acetylglycine (NAG; $p$ = 3.90E-5) and dimethylguanidino valeric acid (DMGV; $p$ = 8.71E-09), and one lipid: phosphatidylcholine 35:1 ($p$ = 1.65E-4). The opposing directions of association for NAG and DMGV with IS in the LLFS cohort are consistent with their previously reported roles in metabolic health. NAG is positively associated with IS in our analysis. It has been previously linked to improved glucose homeostasis and overall metabolic health [57-59]. In contrast, DMGV is negatively associated with IS, and it has been implicated in increased risks of T2DM and non-alcoholic fatty liver disease [60][27]. For lipids, phosphatidylcholine 35:1 is also positively associated with IS. While little is known about its roles as a metabolic marker, *Julve et al*. reported increased levels of phosphatidylcholine 35:1 in post-therapy subjects with type 1 diabetes mellitus [61]. MWAS QQ-plots for lipidome and polar metabolome associations with IS are provided in Supplementary Figure 4b-c.

### *3.4.3. Network association study*

Network association studies were performed on the knowledge-guided and LLFS TSI-Nets using Pascal, as described in Section 2.6, expanding upon single-omic analyses by capturing significant modules enriched for multi-modal interactions relevant to IS. Seven significant knowledge-guided TSI-Net modules ($p$ < 8.88E-5; Table 2) and 20 significant LLFS TSI-Net modules ($p$-value < 1.29E-4; Supplementary Table 1) were identified. Most of these modules contained at least one transcriptome-wide or metabolome-wide significant node for IS. Moreover, they were also enriched with suggestive nodes ($p$ < 0.05) that are interacting with the significant nodes. QQ-plots and p-value distribution of network association tests for IS are provided in Supplementary Figures 10 & 11.

## *3.5. TSI-Net in Practice: Uncovering a Cross-Talk between Gut Microbiome Metabolism and Immune System for Metabolic Health*

The top-ranked TSI-Net module, derived from the GEO co-expression network ($p$ = 4.10E-10), highlights a potential cross-talk between gut microbiome metabolism and the immune system for metabolic health. The module contains 4 transcriptome-wide significant genes (*FCER1A*, *HDC*, *MS4A2*, and *CPA3*). It also includes 2 metabolome-wide significant SMs (NAG and DMGV) directly connected to all genes. This pattern was also observed in the most significant LLFS module (Supplementary T1).

*FCER1A*, *HDC*, *MS4A2*, and *CPA3* are known for their roles in IgE-mediated inflammatory responses in mast cells [62][63]. Contrary to their pro-inflammatory functions in mast cells, these genes exhibit protective associations with IS in the LLFS cohort, which were also replicated in the FHS (Table 1). *HDC*, *MS4A2*, and *CPA3* have also been associated with lower blood glucose levels in non-diabetic populations [64]. Additionally, *FCER1A* has been linked to anti-inflammatory functions in white blood cells, including IgE clearance [65], IL10 production [66]. Consistently, these genes were inversely associated with BMI, TG, and IL6 in the LLFS, further reinforcing their protective evidence in metabolic health (Figure 5b).

NAG, a metabolite derived from gut microbiome activity and linked to fiber metabolism [67][68], has been previously associated with improved metabolic health [57-59]. In addition, its protective effects were further validated in an in-vivo study where NAG supplementation improved weight loss in diet-induced obese mice [69]. NAG is associated with higher IS and lower BMI and TG in the LLFS cohort. Conversely, DMGV, a pro-inflammatory metabolite increased in oxidative stress [70], is associated with lower IS and higher levels of BMI, TG, and IL6 in the LLFS. This aligns with prior reports linking DMGV to increased risk of metabolic and cardiovascular complications [27][61]. However, the molecular roles of DMGV and NAG remain to be elucidated.

In the top-ranked TSI-Net module, NAG is positively connected to *FCER1A*, *HDC*, *MS4A2*, and *CPA3*, aligning with their protective associations for IS, BMI, TG, and IL6. In contrast, DMGV is negatively connected to these genes and is associated with adverse metabolic and inflammatory outcomes (Figure

5). These findings are consistent with the established role of fiber intake and fiber-degrading microbiome activity in reducing inflammation across various conditions [71][72]. Supporting this, *rs2251746*, an intronic variant in *FCER1A*, has also been associated with *Ruminococcaceae* abundance [73], a key fiber-degrading bacterial family in gut [74]. DMGV levels exhibit a strong inverse association with NAG levels ($p$ = 4.65E-34). Notably, higher DMGV levels have also been linked to reduced fruit and vegetable consumption [61].

In conclusion, TSI-Net analysis reveals a novel cross-talk between gut microbiome metabolism and the immune system, driven by the interaction of NAG with *FCER1A*, *HDC*, *MS4A2*, and *CPA3*. This axis is strongly associated with improved IS and reduced obesity and markers of inflammation, including DMGV and IL6.

## 4. Discussion

The present study consists of 3 separate, yet complementary sections to address the critical challenges of mutli-omics integration, network-based analyses, and uncovering novel biologically relevant gene-SM relationships: a machine learning-based LC/MS processing protocol, a semi-supervised network inference model (CMA) for multi-modal TSI-Net construction, and the analysis of TSI-Nets for complex metabolic traits (Figure 1).

The LC/MS data processing protocol combines a forest-based method for batch correction, PCA for noise reduction, and a stepwise regression model to adjust for biological confounders. Common normalization approaches for omics integration, such as min-max or z-score normalization, can effectively shift the omics data to the same distribution [75][76]. However, this can oversimplify complex multi-omics relationships and distort true associations. In the proposed protocol, both LC/MS and RNA-seq profiles are transformed and subsequently adjusted for covariates using regression-based ML models, resulting in residuals that approximate a normal distribution. In addition, while a variety of methods have been developed for processing LC/MS profiles [29], their downstream implications for optimal omics integration are often overlooked. Our results demonstrate that while batch effect correction reduces variability, comprehensive adjustments for confounders are required to prevent inflation in gene-SM associations (Supplementary Figure 1). Although effective for the LLFS cohort with large-scale measurements, this ML protocol may be less suitable for smaller populations where simpler models could suffice. Regardless of the strategy, thorough sensitivity analyses-- such as distribution patterns, heritability assessments, and QQ-plots of multi-omics associations-- are essential to ensure the robustness of protocols used for reliable multi-omics integration. (Figure 2).

Following processing RNA-seq and LC/MS profiles, conventional unsupervised network inference approaches were benchmarked against our semi-supervised approach based on CMA to construct TSI-Nets. The conventional methods face challenges in multi-modal integration due to scale differences and missing data across omics layers. In fact, optimal multi-modal modules for each conventional method were inferred by down-scaling gene expression profiles through keeping genes with highest edge weights for SMs, which in turn led to a lack of diversity across the networks (Figure 3e-g). The semi-supervised CMA model addresses these limitations by leveraging gene-level networks to selectively incorporate SMs based on statistical significance. This approach ensures maximal utilization of RNA-seq data while accounting for the smaller scale of SM profiles. Additionally, combination of gene-SM associations with meta-analysis allows us to obtain combined SM association while accounting for missing data across different omics profiles. Lastly, CMA prevents inflation in gene-SM meta-analysis outcomes by adjusting them for gene correlations within modules. CMA has also been implemented for “omics”-wide association studies in correlated or overlapping populations [18][77][78].

Evaluation of TSI-Nets is crucial for their credibility and utilization. Our evaluation framework incorporates diversity metrics, PPI, and TF co-regulation support alongside the widely-used enrichment analysis [34] to reduce the bias towards well-studied genes and SMs (Figure 3). Interestingly, while PPI and TF co-regulation primarily pertain to transcriptomic evaluation, we observed that genes connected to the same SM in CMA-derived TSI-Nets exhibit both PPI and TF co-regulation support, further validating the network's biological coherence (Figure 4d).

TSI-Nets constructed from both LLFS and knowledge-guided inputs exhibit consistent properties (Figure 4) but yield distinct results in association tests (Supplementary Figure 10 & 11). LLFS modules concentrated significant omics associations within a few highly correlated clusters, whereas knowledge-guided modules displayed more uniformly distributed significance patterns, reflecting their independence from the LLFS population. Given the complexity of TSI-Nets, prioritizing top-ranked modules and nodes with higher significance offers a practical strategy for deriving biological insights. In this study, the top-ranked module for IS highlighted a cross-talk between gut microbiome metabolism and the immune system, characterized by the interaction between the microbiome-drived NAG with immune genes *FCER1A*, *HDC*, *MS4A2*, and *CPA3* with protective effects for metabolic health and inflammation (Figure 5).

The TSI-Net framework provides a promising paradigm for bridging the knowledge gap on the molecular relationships of the SMs. However, it is built on statistical relationships, which, while robust, cannot establish causality. This highlights the need for functional validation of the identified interactions to confirm their biological importance. In addition, the current study focuses on transcriptome-SM interactions inferred from gene expression and SM abundance, which do not directly capture physical protein-SM interactions. Future efforts could integrate proteomic data and protein-SM molecular docking with the TSI-Net framework to create comprehensive protein-SM interaction networks. Lastly, applying graph neural networks to TSI-Nets could broaden their applications by enhancing the predictive power for complex traits.

## 5. Conclusions

This study presents a comprehensive framework for multi-omics integration, network inference, and analysis by addressing challenges in LC/MS data processing, multi-modal network construction, and network-guided biological interpretations. Through utilizing data from the healthy LLFS population with exceptional longevity, the inferred TSI-Nets demonstrate generalizable applicability for studying varied conditions and complex traits. Furthermore, the integration of SMs originating from diverse sources highlights the potential to advance SM pathway discovery, particularly for understudied metabolites. Overall, this framework sets a promising foundation for future advancements in multi-omics research.

## Data and code availability

The LLFS data is available at the Exceptional Longevity Translational Resources portal (https://prod.eliteportal.synapse.org/Explore/Projects/DetailsPage?shortName=LLFS). LLFS and knowledge-guided TSI-Net modules and their respective gene-SM interactions are provided as supplemental information. TWAS, MWAS, and network association summary statistics for IS are provided as supplemental information. All code implementations will be provided in later revisions of the manuscript.

## Acknowledgements

We extend our gratitude to the Long Life Family Study consoritum, including the participants, investigators, and administrative and clinical staff

## Funding

This work was funded by grant AG063893 from the US National Institute on Aging.

## Declaration of Competing interest

G.J.P. is a scientific advisory board member for Cambridge Isotope Laboratories and has a collaborative research agreement with Agilent Technologies. G.J.P. is the Chief Scientific Officer of Panome Bio.

## Author contributions

V.A.M conceptualized the project, developed the methodologies, processed the LLFS data, performed all formal analyses, wrote the manuscript, and revised the mansucript. S.A assisted with project conceptualization, RNA-seq data processing, and FHS data processing. M.S.H assisted with the development of LC/MS protocol and supervised the LC/MS data processing. S.L processed FHS data. W.J.J assisted with the development of CMA algorithm. B.T supervised all of the LLFS blood assays. L.P.S supervised the development of LC/MS protocol. E.W.D and N.L.S assisted with the development of statistical methodologies. P.A assisted with phenotype adjustments and criteria for selection of the participants. M.R.B supervised the RNA-seq protocol and assisted with project conceptualization. G.J.P supervised the LC/MS protocol and assisted with project conceptualization. M.A.P conceptualized the project, supervised the statistical methodologies, and revised the manuscript.

## References

[1]. C. B. Clish, "Metabolomics: an emerging but powerful tool for precision medicine," Cold Spring Harbor Molecular Case Studies, vol. 1, no. 1, pp. a000588, 2015, doi: 10.1101/mcs.a000588.

[2]. D. S. Wishart et al., "HMDB: the Human Metabolome Database," Nucleic Acids Research, vol. 35, no. Database issue, pp. D521–D526, 2007, doi: 10.1093/nar/gkl923.

[3]. S. Qiu et al., "Small molecule metabolites: discovery of biomarkers and therapeutic targets," Signal Transduction and Targeted Therapy, vol. 8, no. 132, pp. 1-11, 2023, doi: 10.1038/s41392-023-01399-3.

[4]. J. H. Wang, J. Byun, and S. Pennathur, "Analytical approaches to metabolomics and applications to systems biology," Seminars in Nephrology, vol. 30, no. 5, pp. 500-511, 2010, doi: 10.1016/j.semnephrol.2010.07.007.

[5]. B. Dréno et al., "The influence of exposome on acne," Journal of the European Academy of Dermatology and Venereology, vol. 32, no. 5, pp. 812-819, 2018, doi: 10.1111/jdv.14820.

[6]. S. Zarei et al., "Using heat shock protein (HSP) inducers as an approach to increase the viability of sterlet (Pisces; Acipenseridae; Acipenser ruthenus) cells against environmental diazinon toxicity," Journal of Hazardous Materials, vol. 465, 2024, Art. no. 133194, doi: 10.1016/j.jhazmat.2023.133194.

[7]. L. Lin and J. Zhang, "Role of intestinal microbiota and metabolites on gut homeostasis and human diseases," BMC Immunology, vol. 18, no. 2, pp. 1-25, 2017, doi: 10.1186/s12865-016-0187-3.

[8]. R. Browaeys, W. Saelens, and Y. Saeys, "NicheNet: modeling intercellular communication by linking ligands to target genes," Nature Methods, vol. 17, no. 2, pp. 159–162, 2020, doi: 10.1038/s41592-019-0667-5.

[9]. X. Li, X. Wang, and M. Snyder, "Systematic investigation of protein–small molecule interactions," IUBMB Life, vol. 65, no. 1, pp. 2-8, 2013, doi: 10.1002/iub.1111.

[10]. M. Taheri et al., "Synthesis, in vitro biological evaluation and molecular modelling of new 2-chloro-3-hydrazinopyrazine derivatives as potent acetylcholinesterase inhibitors on PC12 cells," BMC Chemistry, vol. 16, no. 7, 2022, doi: 10.1186/s13065-022-00799-w

[11]. V. Akbary Moghaddam et al., "A novel sulfamethoxazole derivative as an inhibitory agent against HSP70: A combination of computational with in vitro studies," International Journal of Biological Macromolecules, vol. 189, pp. 194-205, 2021, doi: 10.1016/j.ijbiomac.2021.08.128

[12]. X. Du et al., "Insights into protein-ligand interactions: Mechanisms, models, and methods," International Journal of Molecular Sciences, vol. 17, no. 2, p. 144, 2016, doi: 10.3390/ijms17020144

[13]. S. B. King and M. Singh, "Primate protein-ligand interfaces exhibit significant conservation and unveil human-specific evolutionary drivers," PLOS Computational Biology, vol. 19, no. 3, Art. no. e1010966, 2023, doi: 10.1371/journal.pcbi.1010966

[14]. E. I. Hartig et al., "Cortisol-treated zebrafish embryos develop into pro-inflammatory adults with aberrant immune gene regulation," Biology Open, vol. 5, no. 8, pp. 1134-1141, 2016, doi: 10.1242/bio.020065

[15]. M. J. Drummond et al., "Leucine differentially regulates gene-specific translation in mouse skeletal muscle," The Journal of Nutrition, vol. 147, no. 9, pp. 1616-1623, 2017, doi: 10.3945/jn.117.251181

[16]. S. L. Fanalli et al., "RNA-seq transcriptome profiling of pigs' liver in response to diet with different sources of fatty acids," Frontiers in Genetics, vol. 14, Art. no. 1053021, 2023, doi: 10.3389/fgene.2023.1053021

[17]. O. B. Poirion et al., "DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data," Genome Medicine, vol. 13, no. 112, 2021, doi: 10.1186/s13073-021-00930-x

[18]. S. Acharya et al., "A methodology for gene level omics-WAS integration identifies genes influencing traits associated with cardiovascular risks: the Long Life Family Study," Human Genetics, vol. 143, no. 1, pp. 1-20, 2024, doi: 10.1007/s00439-024-02701-1

[19]. A. Singh et al., "DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays," Bioinformatics, vol. 35, no. 17, pp. 3055–3062, 2019, doi: 10.1093/bioinformatics/bty1054

[20]. J. Yan et al., "Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data," Briefings in Bioinformatics, vol. 19, no. 6, pp. 1370-1381, 2018, doi: 10.1093/bib/bbx066

[21]. Y. Wang, Y. Yang, S. Chen, and J. Wang, "DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration," Briefings in Bioinformatics, vol. 22, no. 5, pp. 1-10, 2021, doi: 10.1093/bib/bbab048

[22]. R. L. Allesøe et al., "Discovery of drug–omics associations in type 2 diabetes with generative deep-learning models," Nature Biotechnology, vol. 41, pp. 399–408, 2023, doi: 10.1038/s41587-022-01520-x

[23]. Z. Pang et al., "MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation," Nucleic Acids Research, vol. 52, no. Web Server issue, pp. W398–W406, 2024, doi: 10.1093/nar/gkae253

[24]. E. M. Forsberg et al., "Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online," Nature Protocols, vol. 13, no. 4, pp. 633-651, 2018, doi: 10.1038/nprot.2017.151

[25]. D. Chen et al., "Identification and characterization of robust hepatocellular carcinoma prognostic subtypes based on an integrative metabolite-protein interaction network," Advanced Science, vol. 8, no. 17, 2021, Art. no. 2100311, doi: 10.1002/advs.202100311

[26]. E. Mastej et al., "Identifying protein–metabolite networks associated with COPD phenotypes," Metabolites, vol. 10, no. 4, Art. no. 124, 2020, doi: 10.3390/metabo10040124

[27]. J. F. O'Sullivan et al., "Dimethylguanidino valeric acid is a marker of liver fat and predicts diabetes," Journal of Clinical Investigation, vol. 127, no. 12, pp. 4394-4402, 2017, doi: 10.1172/JCI95995

[28]. L. Wang et al., "Novel loci for triglyceride/HDL-C ratio longitudinal change among subjects without T2D," Journal of Lipid Research, vol. 66, no. 1, Art. no. 100702, 2025, doi: 10.1016/j.jlr.2024.100702

[29]. E. Stancliffe et al., "An untargeted metabolomics workflow that scales to thousands of samples for population-based studies," Analytical Chemistry, vol. 94, no. 50, pp. 17370-17378, 2022, doi: 10.1021/acs.analchem.2c01270

[30]. M. Kang, E. Ko, and T. B. Mersha, "A roadmap for multi-omics data integration using deep learning," Briefings in Bioinformatics, vol. 23, no. 1, pp. 1–16, 2022, doi: 10.1093/bib/bbab454

[31]. B. B. Misra et al., "Integrated omics: tools, advances and future approaches," Journal of Molecular Endocrinology, vol. 62, no. 1, pp. R21–R45, 2019, doi: 10.1530/JME-18-0055

[32]. M. Picard et al., "Integration strategies of multi-omics data for machine learning analysis," Computational and Structural Biotechnology Journal, vol. 19, pp. 3735–3746, 2021, doi: 10.1016/j.csbj.2021.06.030

[33]. S. Basu et al., "Sparse network modeling and Metscape-based visualization methods for the analysis of large-scale metabolomics data," Bioinformatics, vol. 33, no. 10, pp. 1545–1553, 2017, doi: 10.1093/bioinformatics/btx012

[34]. E. Horgusluoglu et al., "Integrative metabolomics-genomics approach reveals key metabolic pathways and regulators of Alzheimer's disease," Alzheimer's & Dementia, vol. 18, no. 6, pp. 1260–1278, 2022, doi: 10.1002/alz.12468

[35]. A. Schultz and A. A. Qutub, "Reconstruction of tissue-specific metabolic networks using CORDA," PLOS Computational Biology, vol. 12, no. 3, Art. no. e1004808, 2016, doi: 10.1371/journal.pcbi.1004808

[36]. K. Diamanti et al., "Organ-specific metabolic pathways distinguish prediabetes, type 2 diabetes, and normal tissues," Cell Reports Medicine, vol. 3, no. 10, Art. no. 100763, 2022, doi: 10.1016/j.xcrm.2022.100763

[37]. M. K. Wojczynski et al., "NIA Long Life Family Study: Objectives, design, and heritability of cross-sectional and longitudinal phenotypes," Journals of Gerontology: Biological Sciences, vol. 77, no. 4, pp. 717–727, 2022, doi: 10.1093/gerona/glab333

[38]. T. M. Wallace, J. C. Levy, and D. R. Matthews, "Use and Abuse of HOMA Modeling," Diabetes Care, vol. 27, no. 6, pp. 1487–1495, 2004, doi: 10.2337/diacare.27.6.1487

[39]. M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," Genome Biology, vol. 15, Art. no. 550, 2014, doi: 10.1186/s13059-014-0550-8

[40]. P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," BMC Bioinformatics, vol. 9, Art. no. 559, 2008, doi: 10.1186/1471-2105-9-559

[41]. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods," PLoS ONE, vol. 5, no. 9, Art. no. e12776, 2010, doi: 10.1371/journal.pone.0012776

[42]. S. Choobdar et al., "Assessment of network module identification across complex diseases," Nature Methods, vol. 16, no. 9, pp. 843–852, 2019, doi: 10.1038/s41592-019-0509-5

[43]. M. Tomasoni et al., "MONET: a toolbox integrating top-performing methods for network modularization," Bioinformatics, vol. 36, no. 12, pp. 3920–3921, 2020, doi: 10.1093/bioinformatics/btaa236

[44]. M. A. Province and I. B. Borecki, "A Correlated Meta-Analysis Strategy for Data Mining 'OMIC' Scans," Pacific Symposium on Biocomputing, 236-246, 2013, doi: 10.1142/9789814447973_0022

[45]. D. V. Klopfenstein et al., "GOATOOLS: A Python library for Gene Ontology analyses," Scientific Reports, vol. 8, Art. no. 10872, 2018, doi: 10.1038/s41598-018-28948-z

[46]. D. Szklarczyk et al., "The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest," Nucleic Acids Research, vol. 51, no. D1, pp. D638–D646, 2023, doi: 10.1093/nar/gkac1000

[47]. D. Marbach et al., "Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases," Nature Methods, vol. 13, no. 4, pp. 366–370, 2016, doi: 10.1038/nmeth.3799

[48]. M. van Iterson et al., "Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution," Genome Biology, vol. 18, Art. no. 19, 2017, doi: 10.1186/s13059-016-1131-9

[49]. D. Lamparter et al., "Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics," PLOS Computational Biology, vol. 12, no. 1, Art. no. e1004714, 2016, doi: 10.1371/journal.pcbi.1004714

[50]. G. L. Splansky et al., "The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: Design, Recruitment, and Initial Examination," American Journal of Epidemiology, vol. 165, no. 11, pp. 1328–1335, 2007, doi: 10.1093/aje/kwm021

[51]. Y. Zheng et al., "Multi-omics data integration using ratio-based quantitative profiling with Quartet reference materials," Nature Biotechnology, vol. 42, pp. 1133–1149, 2024, doi: 10.1038/s41587-023-01934-1

[52]. L. Almasy and J. Blangero, "Multipoint Quantitative-Trait Linkage Analysis in General Pedigrees," American Journal of Human Genetics, vol. 62, pp. 1198–1211, 1998, doi: 10.1086/301844

[53]. F. A. Hagenbeek et al., "Heritability estimates for 361 blood metabolites across 40 genome-wide association studies," Nature Communications, vol. 11, Art. no. 39, 2020, doi: 10.1038/s41467-019-13770-6

[54]. R. Zechner et al., "Adipose triglyceride lipase and the lipolytic catabolism of cellular fat stores," Journal of Lipid Research, vol. 50, no. 1, pp. 3–21, 2009, doi: 10.1194/jlr.R800031-JLR200

[55]. R. Vermeulen, E. L. Schymanski, A.-L. Barabási, and G. W. Miller, "The exposome and health: Where chemistry meets biology," Science, vol. 367, no. 6476, pp. 392–396, 2020, doi: 10.1126/science.aay3164

[56]. M. Lu et al., "TWAS Atlas: a curated knowledgebase of transcriptome-wide association studies," Nucleic Acids Research, vol. 51, no. D1, pp. D1179–D1187, 2023, doi: 10.1093/nar/gkac821

[57]. W. Perng et al., "Metabolomic Determinants of Metabolic Risk in Mexican Adolescents," Obesity, vol. 25, no. 9, pp. 1594–1602, 2017, doi: 10.1002/oby.21926

[58]. C. Papandreou et al., "Plasma metabolites predict both insulin resistance and incident type 2 diabetes: A metabolomics approach within the PREDIMED study," American Journal of Clinical Nutrition, vol. 109, no. 3, pp. 626–634, 2019, doi: 10.1093/ajcn/nqy262

[59]. C. L. Axelrod et al., "Metabolomic fingerprints of medical therapy versus bariatric surgery in patients with obesity and type 2 diabetes: The STAMPEDE trial," Diabetes Care, vol. 47, no. 11, pp. 2024–2032, 2024, doi: 10.2337/dc24-0859.

[60]. F. Ottosson et al., "Dimethylguanidino Valerate: A Lifestyle-Related Metabolite Associated With Future Coronary Artery Disease and Cardiovascular Mortality," Journal of the American Heart Association, vol. 8, no. 20, pp. e012846, 2019, doi: 10.1161/JAHA.119.012846

[61]. J. Julve et al., "Circulating metabolomic and lipidomic changes in subjects with new-onset type 1 diabetes after optimization of glycemic control," Diabetes Research and Clinical Practice, vol. 197, p. 110578, 2023, doi: 10.1016/j.diabres.2023.110578

[62]. E. Inage et al., "Critical Roles for PU.1, GATA1, and GATA2 in the Expression of Human FcεRI on Mast Cells: PU.1 and GATA1 Transactivate FCER1A, and GATA2 Transactivates FCER1A and MS4A2," *The Journal of Immunology*, vol. 192, no. 8, pp. 3936–3946, Apr. 2014, doi: 10.4049/jimmunol.1302366.

[63]. H. Ohtsu et al., "Mice lacking histidine decarboxylase exhibit abnormal mast cells," FEBS Letters, vol. 502, no. 1–2, pp. 53–56, 2001, doi: 10.1016/S0014-5793(01)02663-1

[64]. B. H. Chen et al., "Peripheral Blood Transcriptomic Signatures of Fasting Glucose and Insulin Concentrations," Diabetes, vol. 65, no. 12, pp. 3794–3804, Dec. 2016, doi: 10.2337/db16-0470

[65]. A. M. Greer et al., "Serum IgE clearance is facilitated by human FcεRI internalization," The Journal of Clinical Investigation, vol. 124, no. 3, pp. 1187–1198, 2014, doi: 10.1172/JCI68964

[66]. N. Novak, T. Bieber, and N. Katoh, "Engagement of FcεRI on Human Monocytes Induces the Production of IL-10 and Prevents Their Differentiation in Dendritic Cells," The Journal of Immunology, vol. 167, no. 2, pp. 797–804, July 2001, doi: 10.4049/jimmunol.167.2.797

[67]. B. D. Badal et al., "Substitution of One Meat-Based Meal With Vegetarian and Vegan Alternatives Generates Lower Ammonia and Alters Metabolites in Cirrhosis: A Randomized Clinical Trial," *Clinical and Translational Gastroenterology*, vol. 15, p. e00707, 2024, doi: 10.14309/ctg.0000000000000707

[68]. M. S. Lustgarten et al., "Metabolites related to gut bacterial metabolism, peroxisome proliferator-activated receptor-alpha activation, and insulin sensitivity are associated with physical function in functionally-limited older adults," Aging Cell, vol. 13, no. 5, pp. 918–925, 2014, doi: 10.1111/acel.12251

[69]. K.-J. Su et al., "Systematic metabolomic studies identified adult adiposity biomarkers with acetylglycine associated with fat loss in vivo," Frontiers in Molecular Biosciences, vol. 10, p. 1166333, Apr. 2023, doi: 10.3389/fmolb.2023.1166333

[70]. X. Guo, Y. Xing, and W. Jin, "Role of ADMA in the pathogenesis of microvascular complications in type 2 diabetes mellitus," Frontiers in Endocrinology, vol. 14, p. 1183586, Apr. 2023, doi: 10.3389/fendo.2023.1183586

[71]. S.-M. Kuo, "The interplay between fiber and the intestinal microbiome in the inflammatory response," Advances in Nutrition, vol. 4, no. 1, pp. 16–28, Jan. 2013, doi: 10.3945/an.112.003046

[72]. W. Ma et al., "Dietary fiber intake, the gut microbiome, and chronic systemic inflammation in a cohort of adult men," Genome Medicine, vol. 13, no. 102, 2021, doi: 10.1186/s13073-021-00921-y

[73]. M. C. Rühlemann et al., "Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome," Nature Genetics, vol. 53, no. 2, pp. 147–155, Feb. 2021, doi: 10.1038/s41588-020-00747-1

[74]. A. El Kaoutari, F. Armougom, J. I. Gordon, D. Raoult, and B. Henrissat, "The abundance and variety of carbohydrate-active enzymes in the human gut microbiota," Nature Reviews Microbiology, vol. 11, no. 7, pp. 497–504, Jul. 2013, doi: 10.1038/nrmicro3050

[75]. L. Tong, J. Mitchel, K. Chatlin, and M. D. Wang, "Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis," BMC Medical Informatics and Decision Making, vol. 20, no. 225, pp. 1–12, 2020, doi: 10.1186/s12911-020-01225-8

[76]. R. Duan et al., "Evaluation and comparison of multi-omics data integration methods for cancer subtyping," PLOS Computational Biology, vol. 17, no. 8, p. e1009224, Aug. 2021, doi: 10.1371/journal.pcbi.1009224

[77]. S. J. Hartman et al., "A microbiome-directed therapeutic food for children recovering from severe acute malnutrition," Science Translational Medicine, vol. 16, no. 725, p. eadn2366, Oct. 2024, doi: 10.1126/scitranslmed.adn2366

[78]. M. F. Feitosa et al., "Discovery of genomic and transcriptomic pleiotropy between kidney function and soluble receptor for advanced glycation end products using correlated meta-analyses: The Long Life Family Study," Aging Cell, vol. 23, p. e14261, 2024, doi: 10.1111/acel.14261

## Figure 1

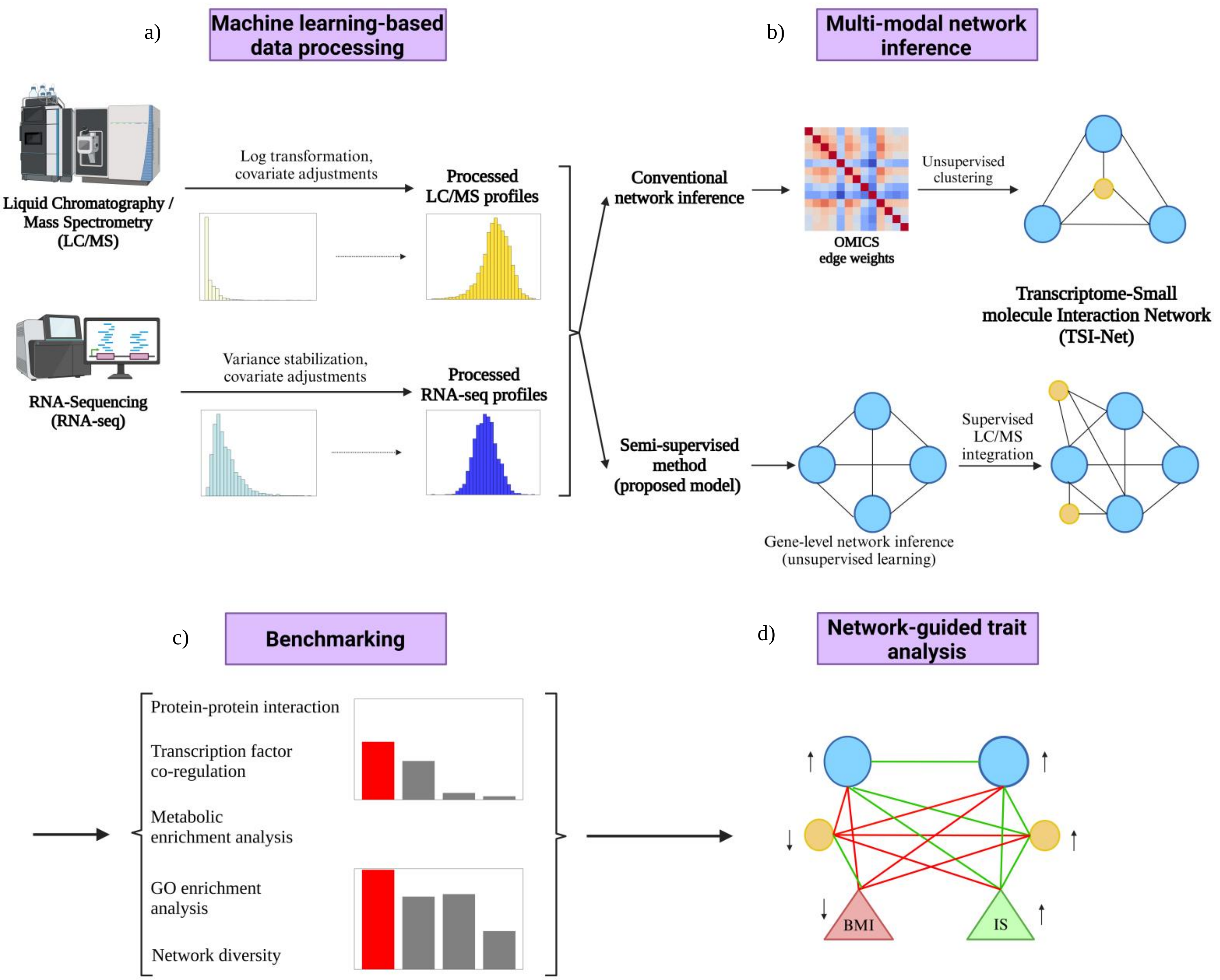

a)
Machine learning-based data processing
Liquid Chromatography / Mass Spectrometry (LC/MS)
Log transformation, covariate adjustments
Processed LC/MS profiles
RNA-Sequencing (RNA-seq)
Variance stabilization, covariate adjustments
Processed RNA-seq profiles
b)
Multi-modal network inference
Conventional network inference
OMICS edge weights
Unsupervised clustering
Transcriptome-Small molecule Interaction Network (TSI-Net)
Semi-supervised method (proposed model)
Gene-level network inference (unsupervised learning)
Supervised LC/MS integration
c)
Benchmarking
Protein-protein interaction
Transcription factor co-regulation
Metabolic enrichment analysis
GO enrichment analysis
Network diversity
d)
Network-guided trait analysis
BMI
IS

**Figure 1. General workflow of the study. a)** In the first section of the study, a ML-based protocol is proposed, composed of data transformation and regression-based adjustments to prepare the LC/MS and RNA-seq profiles for integration. **b)** In the second part, conventional unsupervised network inference approaches as well as a newly proposed semi-supervised method were used to construct TSI-Nets. The semi-supervised method includes unsupervised construction of gene-level networks from the RNA-seq profiles with the conventional approaches, followed by supervise integration of SMs into the baseline networks. **c)** The semi-supervised method is benchmarked against the conventional approaches using multiple evaluation metrics. **d)** Upon benchmarking the resulting TSI-Nets, they were used to study metabolic traits, such as IS or BMI, in a network-guided manner.

## Figure 2

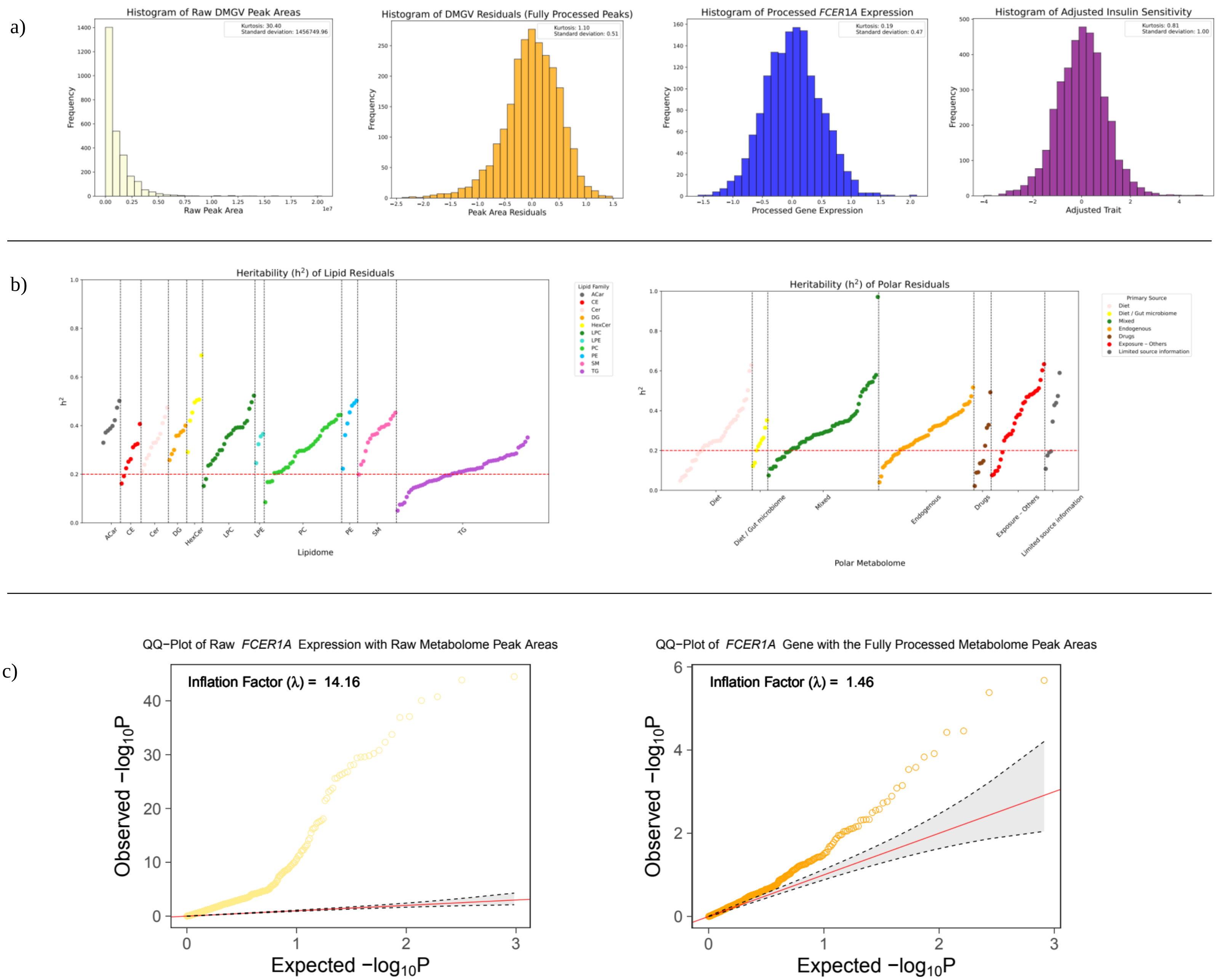


**Figure 2. LC/MS data processing results. a)** From left to right, the first plot illustrates the distribution of raw peak areas for dimethtylguanidino valeric acid (DMGV). Second plot demonstrates the distribution of DMGV upon adjustments for technical, demographical, and biological covariates. Third plot shows the distribution of *FCER1A* gene upon data processing. Finally, the fourth plot represents the distribution of insulin sensitivity upon covariate adjustment. **b)** The plot on the left illustrates the heritibality ($h^2$) of lipids across different lipid categories. The plot on the right illustrates the $h^2$ of polar SMs grouped by their primary source. **c)** QQ-plot of *FCER1A* gene with raw metabolome peak areas (left) and processed peak areas (right). Lipid categories include: “Acar”: acetylcarnitines, “Cer”: ceramides, “CE”: cholesterol esters, “DG”: diglycerides, “HexCer”: hexosylceramides, “LPC”: lysophosphatidylcholines, “LPE”: lysophosphatidylethanolamines, “PC”:

phosphatidylcholines, “PE”: phosphatidylethanolamines, “SM”: sphingomyelins. The primary source of polars include: “Diet”: from dietary sources, “Diet / Gut microbiome”: from gut microbiome metabolism, “Mixed”: from dietary and endogenous sources, “Endogenous”: synthesized internally within the body, “Drugs”, “Exposures – Others”, and “Limited source information”.

**Figure 3**

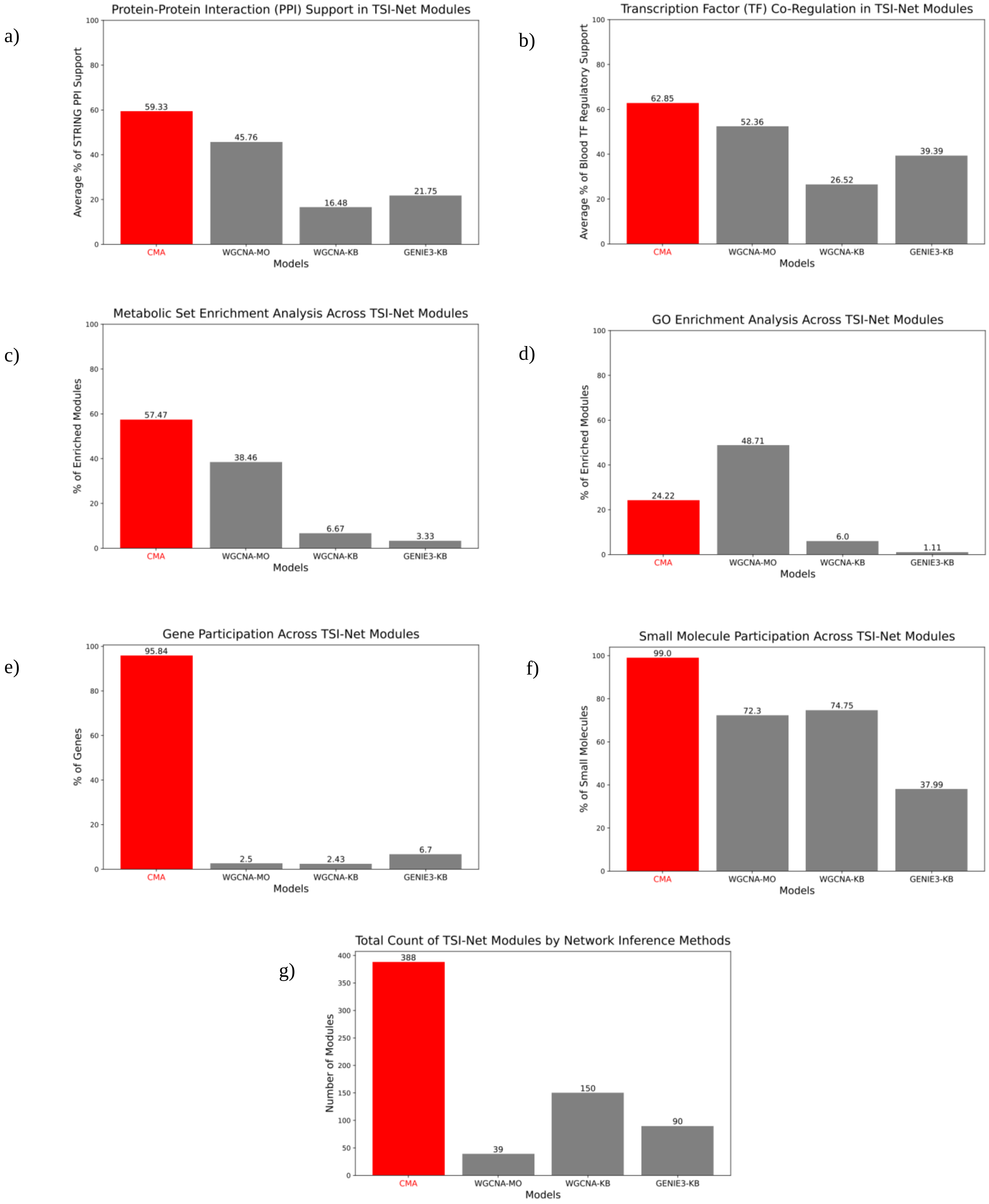

a)
Protein-Protein Interaction (PPI) Support in TSI-Net Modules
Average % of STRING PPI Support
0
20
40
60
80
100
59.33
45.76
16.48
21.75
CMA
WGCNA-MO
WGCNA-KB
GENIE3-KB
Models
b)
Transcription Factor (TF) Co-Regulation in TSI-Net Modules
Average % of Blood TF Regulatory Support
62.85
52.36
26.52
39.39
CMA
WGCNA-MO
WGCNA-KB
GENIE3-KB
Models
c)
Metabolic Set Enrichment Analysis Across TSI-Net Modules
% of Enriched Modules
57.47
38.46
6.67
3.33
CMA
WGCNA-MO
WGCNA-KB
GENIE3-KB
Models
d)
GO Enrichment Analysis Across TSI-Net Modules
% of Enriched Modules
24.22
48.71
6.0
1.11
CMA
WGCNA-MO
WGCNA-KB
GENIE3-KB
Models
e)
Gene Participation Across TSI-Net Modules
% of Genes
95.84
2.5
2.43
6.7
CMA
WGCNA-MO
WGCNA-KB
GENIE3-KB
Models
f)
Small Molecule Participation Across TSI-Net Modules
% of Small Molecules
99.0
72.3
74.75
37.99
CMA
WGCNA-MO
WGCNA-KB
GENIE3-KB
Models
g)
Total Count of TSI-Net Modules by Network Inference Methods
Number of Modules
0
50
100
150
200
250
300
350
400
388
39
150
90
CMA
WGCNA-MO
WGCNA-KB
GENIE3-KB
Models

**Figure 3. Benchmarking network inference approaches for TSI-Net construction.** a) Represents the proportion of TSI-Net modules with significant hits from MSEA. b) Represents the average proportion of gene-gene interactions supported by STRINGdb PPI networks. For each module, proportion of support was calculated, which was averaged across all modules. c) Represents the average proportion of genes co-regulated by same TFs based on the blood-specific GRNs. d) Represents the proportion of TSI-Net modules with significant GO terms. e) Representation of genes profiled in the LLFS across TSI-Nets inferred by each method. f) Representation of SMs profiled in the LLFS across TSI-Nets inferred by each method. g) Total number of TSI-Net modules inferred by each method.

## Figure 4

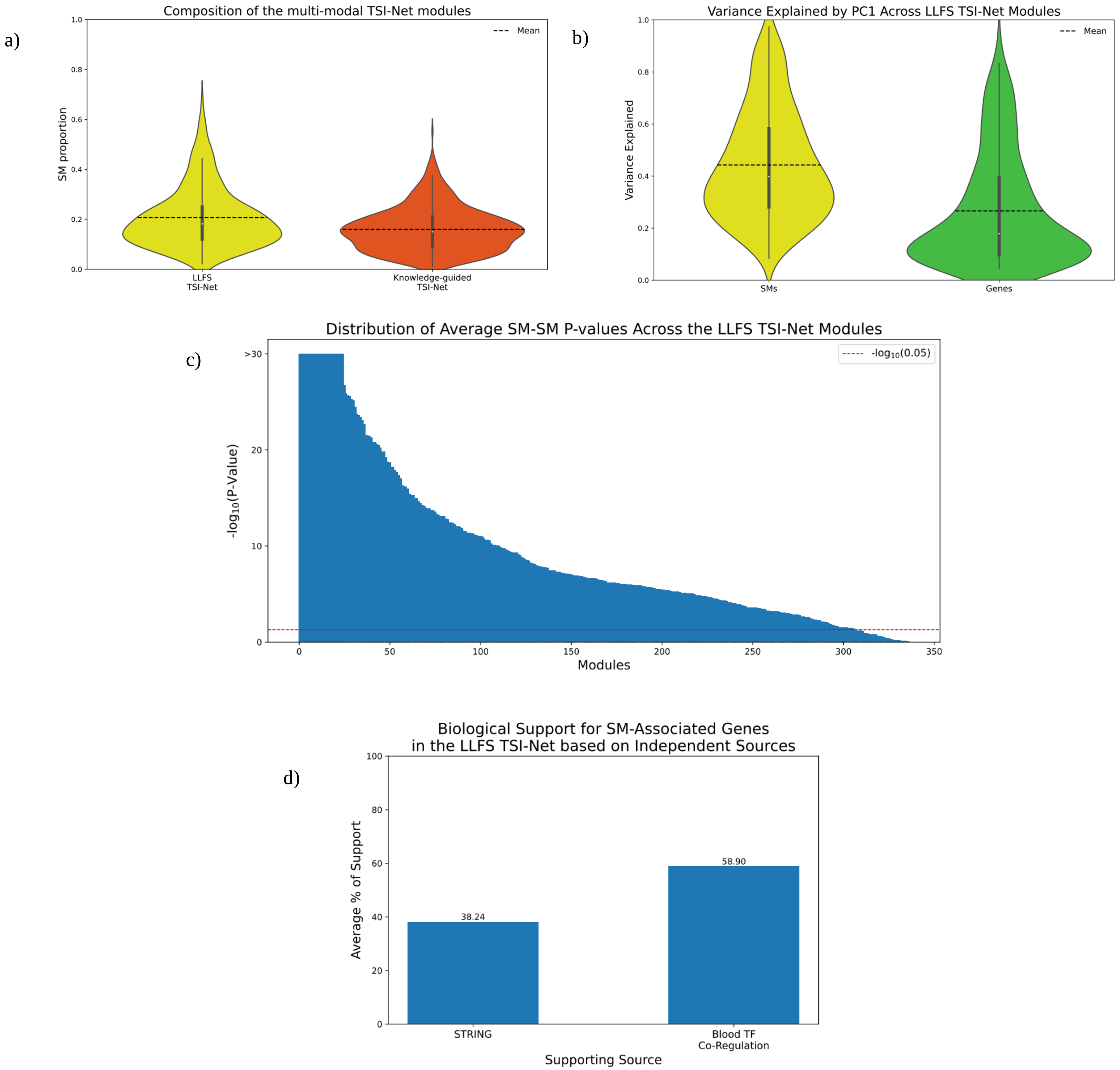
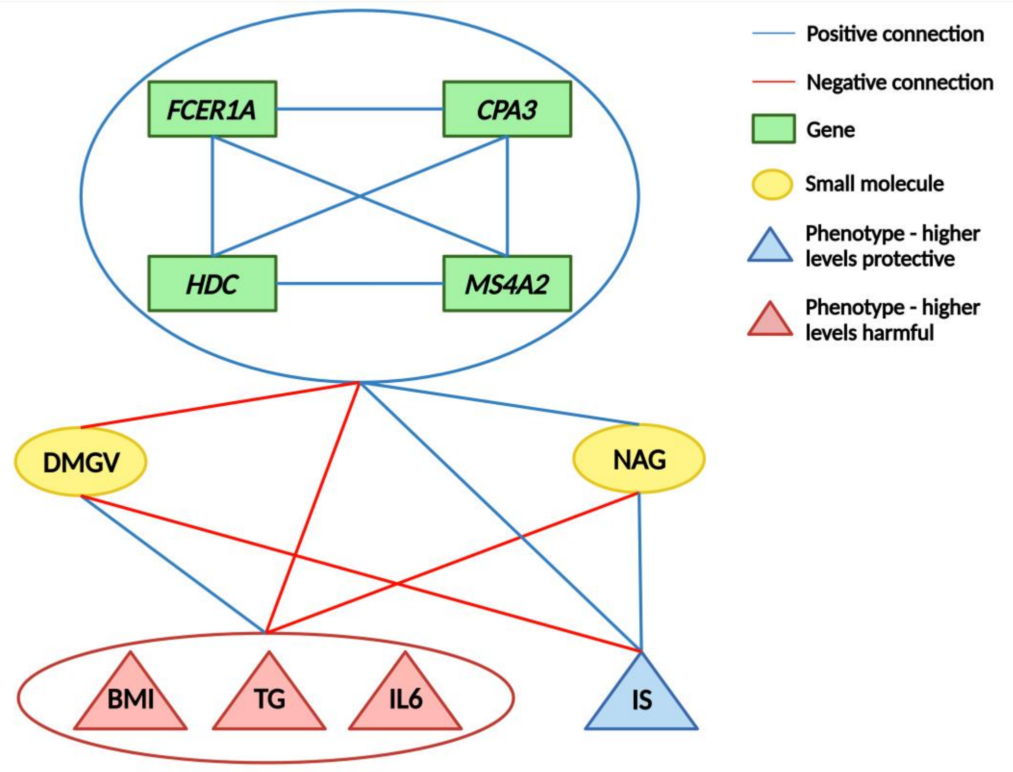


**Figure 4. CMA-derived TSI-Net properties. a)** Violin plots representing the composition of the TSI-Net modules based on the proportion of SMs across modules in the LLFS and knowledge-guided TSI-Nets. **b)** Violin plots of the variance explained by PC1 across the LLFS TSI-Net modules. For each module, separate PCA were performed on genes and SMs participating in the module. **c)** Histogram of the average $-\log_{10}$(p-value) of SM-SM associations in the TSI-Net modules. In each module, pairwise SM-SM associations were assessed and subsequently averaged for the number of pairs. The maximum range of the plot was set to $-\log_{10}(\text{p-value}) \leq 30$ for readability. **d)** Average percentage of support for SM-molecule interacting genes across the LLFS TSI-Nets

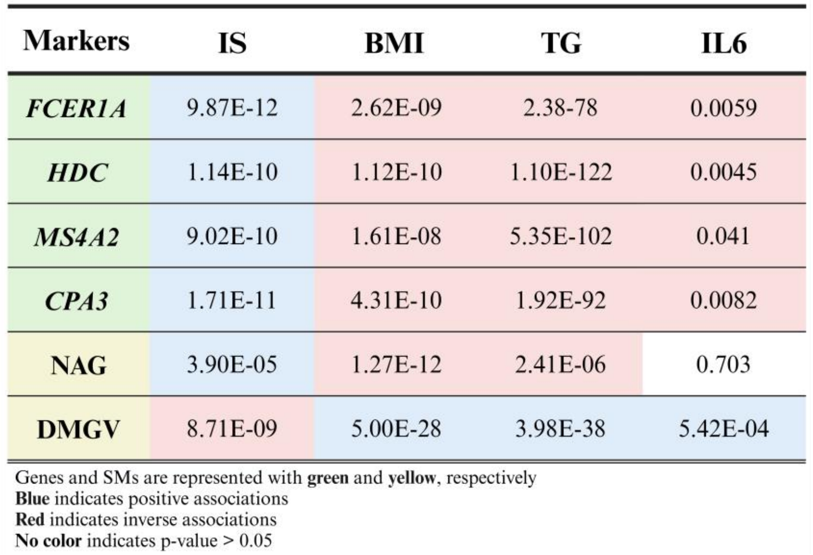
based on the STRINGdb PPI and blood GRN. Percentage of support for the gene set of each SM was calculated and averaged across all SMs.

## Figure 5

a)



b)



| Markers | IS | BMI | TG | IL6 |
|---|---|---|---|---|
| ***FCER1A*** | 9.87E-12 | 2.62E-09 | 2.38-78 | 0.0059 |
| ***HDC*** | 1.14E-10 | 1.12E-10 | 1.10E-122 | 0.0045 |
| ***MS4A2*** | 9.02E-10 | 1.61E-08 | 5.35E-102 | 0.041 |
| ***CPA3*** | 1.71E-11 | 4.31E-10 | 1.92E-92 | 0.0082 |
| **NAG** | 3.90E-05 | 1.27E-12 | 2.41E-06 | 0.703 |
| **DMGV** | 8.71E-09 | 5.00E-28 | 3.98E-38 | 5.42E-04 |

Genes and SMs are represented with **green** and **yellow**, respectively
**Blue** indicates positive associations
**Red** indicates inverse associations
**No color** indicates p-value > 0.05

**Figure 5. Demonstration of the significant gene-SM interactions in the top-ranked module for IS. a)** NAG and DMGV are connected to all 4 genes illustrated in the sub-module. NAG is also connected to BMI and TG (not IL6), DMGV is connected to BMI, TG, and IL6, and the 4 genes are connected to BMI and TG. However, only *FCER1A*, *HDC*, and *CPA3* are connected to IL6. The large blue and red ellipses are used for clear illustration. **b)** Association summary of the significant nodes of the top-ranked module for the metabolic traits.

**Table 1. Significant TWAS summary statistics for IS**

| Gene Symbol | LLFS P-value | LLFS Beta | FHS P-value | FHS Beta | TWAS Atlas |
|---|---|---|---|---|---|
| *FCER1A* | 9.87E-12 | 0.449 | 9.65E-14 | 0.375 | - |
| *CPA3* | 1.71E-11 | 0.400 | 6.71E-15 | 0.336 | Fasting Glucose |
| *GATA2* | 7.17E-11 | 0.236 | 1.41E-13 | 0.254 | Fasting Glucose, Fasting Insulin |
| *HDC* | 1.14E-10 | 0.277 | 1.84E-13 | 0.244 | Fasting Glucose |
| *SLC45A3* | 1.52E-10 | 0.368 | 3.69E-10 | 0.287 | - |
| *ABCG1* | 2.94E-10 | 0.549 | 1.49E-08 | 0.458 | - |
| *MS4A2* | 9.02E-10 | 0.354 | 9.12E-11 | 0.337 | Fasting Glucose |
| *AKAP12* | 9.73E-10 | 0.355 | 8.48E-14 | 0.337 | Fasting Glucose |
| *ITGB8* | 1.83E-08 | 0.353 | N/A | N/A | - |
| *CX3CR1* | 1.85E-07 | -0.646 | 8.64E-06 | -0.4444 | - |
| *MBNL3* | 3.06E-07 | 0.434 | 0.782 | -0.0385 | - |
| *PTGER2* | 5.32E-07 | -0.805 | 1.55E-08 | -0.768 | Fasting Glucose |
| *LINC02458* | 5.39E-07 | 0.418 | 1.51E-12 | 0.474 | - |
| *CCDC71* | 9.71E-07 | 1.087 | 7.15E-5 | 0.809 | - |
| *KLHDC8B* | 1.77E-06 | 0.419 | 5.62E-5 | 0.291 | - |
| *LTF* | 2.27E-06 | -0.175 | 0.0925 | -0.0434 | - |
| *GCSAML* | 2.52E-06 | 0.446 | 9.70E-12 | 0.503 | - |
| *ENPP3* | 2.62E-06 | 0.413 | 2.61E-11 | 0.502 | - |

**Table 2. Network association summary of the significant knowledge-guided TSI-Net modules**

| Module ID | Module P-value | Available Nodes | Significant Nodes* | Suggestive Nodes** |
|---|---|---|---|---|
| **GEO97** | 4.10E-10 | 63 | 7 | 9 |
| **STRING193** | 5.52E-08 | 68 | 4 | 10 |
| **GEO89** | 2.50E-06 | 101 | 2 | 20 |
| **GEO149** | 8.31E-06 | 145 | 1 | 26 |
| **GEO104** | 1.87E-05 | 33 | 1 | 10 |
| **InWeb47** | 2.25E-05 | 48 | 1 | 10 |
| **GEO102** | 6.78E-05 | 61 | 1 | 15 |

* Transcriptome significance threshold: p = 3.05E-06, polar metabolome significance threshold: p = 2.27E-04, lipidome significance threshold: p = 2.66E-04

** Suggestive association: p < 0.05