



Systems biology

# NetSHy: network summarization via a hybrid approach leveraging topological properties

Thao Vu <sup>1,\*</sup>, Elizabeth M. Litkowski<sup>2,3</sup>, Weixuan Liu<sup>1</sup>, Katherine A. Pratte<sup>4</sup>, Leslie Lange<sup>3</sup>, Russell P. Bowler<sup>5</sup>, Farnoush Banaei-Kashani<sup>6</sup> and Katerina J. Kechris <sup>1,\*</sup>

<sup>1</sup>Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA, <sup>2</sup>Department of Epidemiology, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA, <sup>3</sup>Division of Biomedical Informatics & Personalized Medicine, School of Medicine, Colorado University Anschutz Medical Campus, Aurora, CO 80045, USA, <sup>4</sup>Department of Biostatistics, National Jewish Health, Denver, CO 80206, USA, <sup>5</sup>Division of Pulmonary Medicine, Department of Medicine, National Jewish Health, Denver, CO 80206, USA and <sup>6</sup>Department of Computer Science and Engineering, College of Engineering, Design and Computing, University of Colorado Denver, Denver, CO 80204, USA

\*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on May 26, 2022; revised on August 30, 2022; editorial decision on December 10, 2022; accepted on December 20, 2022

## Abstract

**Motivation:** Biological networks can provide a system-level understanding of underlying processes. In many contexts, networks have a high degree of modularity, i.e. they consist of subsets of nodes, often known as subnetworks or modules, which are highly interconnected and may perform separate functions. In order to perform subsequent analyses to investigate the association between the identified module and a variable of interest, a module summarization, that best explains the module's information and reduces dimensionality is often needed. Conventional approaches for obtaining network representation typically rely only on the profiles of the nodes within the network while disregarding the inherent network topological information.

**Results:** In this article, we propose NetSHy, a hybrid approach which is capable of reducing the dimension of a network while incorporating topological properties to aid the interpretation of the downstream analyses. In particular, NetSHy applies principal component analysis (PCA) on a combination of the node profiles and the well-known Laplacian matrix derived directly from the network similarity matrix to extract a summarization at a subject level. Simulation scenarios based on random and empirical networks at varying network sizes and sparsity levels show that NetSHy outperforms the conventional PCA approach applied directly on node profiles, in terms of recovering the true correlation with a phenotype of interest and maintaining a higher amount of explained variation in the data when networks are relatively sparse. The robustness of NetSHy is also demonstrated by a more consistent correlation with the observed phenotype as the sample size decreases. Lastly, a genome-wide association study is performed as an application of a downstream analysis, where NetSHy summarization scores on the biological networks identify more significant single nucleotide polymorphisms than the conventional network representation.

**Availability and implementation:** R code implementation of NetSHy is available at <https://github.com/thaovu1/NetSHy>

**Contact:** thao.3.vu@cuanschutz.edu or katerina.kechris@cuanschutz.edu

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

Complex diseases are rarely a consequence of an abnormality on one single molecule, but rather the result of complex interactions and perturbations involving large sets of molecular components, which gives rise to the emergence of network-based approaches to gain a system-level understanding of the underlying biological processes

(Valentini *et al.*, 2014). In particular, the informative patterns revealed by biological networks have been employed to gain insights into disease mechanisms (Zhang and Itan, 2019), study comorbidities (Hu *et al.*, 2016), facilitate therapeutic drugs and their targets (Fiscun *et al.*, 2018) and discover network-associated biomarkers (Sevimoglu and Arga, 2014). For instance, Pujana *et al.* (2007) generated a

network consisting of 118 genes, in which a novel candidate gene, hyaluronan-mediated motility receptor, was demonstrated to associate with a higher risk of breast cancer in humans. In another study, Shu et al. (2017) constructed shared gene networks to uncover key drivers for cardiovascular disease and type 2 diabetes, which in turn offered important insights for the development of therapeutic avenues targeting both diseases simultaneously.

Network analysis simplifies the complex biological systems to constituents (nodes) and their interactions (edges). Networks can be constructed directly based on gene expression data such as transcriptional regulatory networks (Chen et al., 2006) and co-expression networks (Zhang and Horvath, 2005) or can be built using the integration of multi-omics data (Hawe et al., 2019). For example, in protein-protein interaction (PPI) networks, nodes are individual proteins and pairwise physical interactions are characterized by edges. Similarly, in co-expression networks, genes serve as nodes and their corresponding connecting edges are defined by the correlation between expression patterns. Utilizing the integration of multi-omics data, Bartel et al. (2015) captured the relationships between all pairs of transcripts and metabolites through a transcriptome-metabolome network.

While biological global networks provide a big picture of the underlying cellular processes, they are often too large to be considered as a whole. It has been shown that molecular networks have a high degree of modularity, i.e. they consist of subsets of nodes which are highly interconnected and may perform separate functions (Alexander et al., 2009). Such collections of nodes are often known as modules or subnetworks. Ravasz et al. (2002), for example, focused on studying the metabolic networks of 43 distinct organisms to uncover the hierarchical modularity property, which was shown to closely overlap with known metabolic functions. Additionally, acknowledging the advantages of network modular structure (Caetano-Anollés et al., 2019) regarding evolvability and robustness, Choobdar et al. (2019) launched the community-driven challenge promoting assessment of different methods in identifying disease-relevant modules across a diversity of network types such as PPI, homology and cancer-gene networks. Once identified, the subnetworks are related to external information in downstream analyses to obtain biologically meaningful interpretations. For instance, individual genes in a key module were modeled simultaneously in a LASSO-Cox regression framework to identify signature genes which were predictive of the overall survival of patients with lung adenocarcinoma (Wu et al., 2022). Conversely, one can summarize a module into a feature which best explains the module's behavior, referred to as a module representation, which then serves as direct input for downstream analyses. Such module-centric approaches allow the collective impact of all entities in the identified module on an outcome of interest to be investigated (Langfelder et al., 2013; Schlosser et al., 2020).

Denote a molecular profile of  $n$  subjects and  $p$  features as  $X_{n \times p}$ , a network highlighting the relationships between the features is represented by an adjacency matrix  $A_{p \times p}$ . Supplementary Figure S1 outlines existing approaches for network representation, which utilize either  $X_{n \times p}$  or  $A_{p \times p}$ . In the popular weighted correlation network analysis, Zhang and Horvath (2005) represented the gene expression profiles of a given network by the first principal component (PC) of  $X_{n \times p}$ , namely 'eigengene', denoted as  $Z_{n \times 1}$ . The 'eigengene' can be thought of as a weighted average expression of all individual genes in the network, in which the corresponding weights are defined such that the resulting 'eigengene' can explain the most variation in the data. Similarly, Schlosser et al. (2020) summarized each metabolite module into 'eigenmetabolite', i.e. the first PC of the network metabolic profile. The 'eigenmetabolite' was subsequently used to identify significant genetic associations to make inferences about shared biochemical pathways. Alternatively, one can use the molecular profile of the most highly connected intramodular node, known as a hub node, as the network representation (Langfelder and Horvath, 2008). The rationale for this approach is that hub nodes are more relevant to the functionality of networks than other nodes since they are central of the network's architecture. For instance, in protein knockout experiments, hub proteins were shown to be essential for

survival in lower organisms (Langfelder et al., 2013). While these approaches are capable of summarizing networks at the subject level, neither of them fully takes advantage of network topological properties. Specifically, the 'eigengene' approach only focuses on the direction maximizing the variation in the measurements associated with nodes in the network which does not necessarily reflect the underlying connectivity structure between nodes. The 'hub node' approach, on the other hand, projects the whole module information onto the profile of the single node with the most connections while disregarding the roles of the remaining entities.

Networks are often defined as graphs from the graph theory perspective. There exist many graph embedding techniques which are designed to learn the graph topology directly using a network adjacency matrix  $A_{p \times p}$ . In particular, matrix factorization-based graph embeddings such as Graph Laplacian eigenmaps, multidimensional scaling (Hofmann and Buhmann, 1995) and Isomap (Tenenbaum et al., 2000) exploited the network topology to create an interpoint distance matrix on which spectral decomposition was performed to extract a representation capturing the network structure at the node level. Furthermore, the emergence of deep learning in graph data has widened the scope of graph representation techniques. Deep learning-based methods such as DeepWalk (Perozzi et al., 2014) and node2vec (Grover and Leskovec, 2016) deployed the truncated random walks (Spitzer, 2013), which were essentially the sets of paths sampled from the input graph to maximize the co-occurrence probability of the observing node's neighborhood. In a different manner, autoencoders and deep neural networks can be applied directly to the proximity matrix of the whole graph rather than following random walk paths. More specifically, graph autoencoder (Vincent et al., 2010) approaches such as structural deep network embedding (Wang et al., 2016) and sparse autoencoder (Tian et al., 2014) minimized the reconstruction error of the representation output and the network input through encoder and decoder steps, such that nodes with similar neighborhood would have similar embeddings. The survey by Cai et al. (2018) comprehensively reviewed each of these methods.

The aforementioned embedding techniques show promising results regarding reducing the dimension of input graphs while preserving topology information at the node level, i.e. transforming  $A_{p \times p}$  to  $Z_{p \times d}$  such that  $d < p$ . However, typical analyses linking module-specific features to clinical traits of interest (e.g. disease status, survival time, etc.) require a subject-level representation, i.e.  $Z_{n \times d}$ , with  $d < p$ . That becomes our motivation to propose an approach, NetSHy, that is capable of summarizing a network at a subject level while capturing the network topological properties. Specifically, NetSHy creates a latent matrix by combining the feature profile matrix  $X_{n \times p}$  and network topology stored in a Laplacian matrix  $L_{p \times p}$  prior to performing a principal component analysis (PCA) to obtain a summarization score for each subject. NetSHy is evaluated using inferred biological networks from a study on chronic obstructive pulmonary disease (COPD) (Mastej et al., 2020) as well as simulated networks at different levels of network sparsity. The performance of the proposed approach, NetSHy, is compared to the conventional approach of just using the molecular profile without network information, which we refer to as NoNet, based on: (i) correlation with continuous phenotype and (ii) the variance of data explained by the resulting network summarization. We find that NetSHy outperforms NoNet in recovering the true correlation with the phenotype and maintaining a higher level of explained variation in the data when the networks are relatively sparse. Furthermore, NetSHy is proved to be more robust than NoNet when the sample size of the biological networks decreases. Lastly, we illustrate an example of a downstream analysis by performing a genome-wide association study (GWAS) using the results of the network summarization and find that there are stronger signals when using network information through NetSHy compared to NoNet.

## 2 Materials and methods

### 2.1 NetSHy

A weighted, non-negative, undirected network of  $p$  nodes can be represented by an adjacency matrix  $A = \{a_{kl}\}_{k,l=1}^p$ , where  $a_{kl}$  reflects

the similarity between nodes  $k$  and  $l$  in the network. Denote the corresponding feature profile of all nodes in the network as  $X_{n \times p}$  with  $n$  and  $p$  representing numbers of subjects and features, respectively. Direct connection between any two nodes in the network can be reflected using the Laplacian matrix (Belkin and Niyogi, 2003) as

$$L = D - A,$$

where,  $A_{p \times p}$  is defined as above,  $D$  is a diagonal degree matrix such that  $D_{kk} = \sum_l a_{kl}$ ,  $k, l = 1, \dots, p$ . The symmetric Laplacian matrix  $L_{p \times p}$  records the direct connection of any two nodes as well as the node degree distribution in the network.

With  $L_{p \times p}$  capturing the network topology, we define  $X^*$ , a transformation of  $X$ , as a combination of both node feature profiles and network topology, such that

$$X^* = XL.$$

We then perform PCA on  $X^*$  to extract the first PC of dimension  $n \times 1$ , as a representation capturing the variability in both directions of feature data and topology. For the rest of the article, we will refer to the first PC obtained from  $X$  and  $X^*$  as NoNet and NetSHy summarization, denoted as  $Z_{\text{NoNet}}$  and  $Z_{\text{NetSHy}}$ , respectively. Simultaneously, the corresponding first eigenvectors of size  $p \times 1$ , which store the direction and relative contribution of each node in the network to the summarization, are denoted as  $\phi_{\text{NoNet}}$  and  $\phi_{\text{NetSHy}}$ , respectively.

## 2.2 Simulation scenarios

Two simulation scenarios were designed to evaluate the performance of NoNet and NetSHy summarization [Columns (1) and (2) of Supplementary Fig. S2].

- **Scenario (1):** Given a number of nodes  $p$  and a graph sparsity  $\alpha_0 \in [0, 1]$ , a network, denoted as  $A$ , was generated from a random model Renyi-Erdos (Erdos et al., 1960) such that the probability of a node connecting to another node within a network was approximately  $\alpha_0$ . This was accomplished using the R package *igraph* (Csardi et al., 2006). The edge weights  $\{w_{kl}\}_{k,l=1}^p$  connecting nodes  $k$  and  $l$  were simulated from the uniform distribution such that  $w_{kl} \sim \text{Unif}(0.1, 0.8)$ . Three network sizes of  $p = 30, 60$  and  $100$ , and three levels of sparsity  $\alpha_0 = 0.3, 0.6$ , and  $0.9$  were included in the study.
- **Scenario (2):** An adjacency matrix  $A$  was obtained directly from a previously published metabolite-protein (M-P) network by Mastej et al. (2020). Different from Scenario (1), the network size and sparsity level were fixed at  $p = 20$  and  $\alpha_0 = 0.51$ , respectively. However, by applying hard thresholding to remove weak edges, we were able to additionally assess the impact of network sparsity at  $\alpha_0 = 0.25$ .

Once the network adjacency matrix  $A$  was obtained, each off-diagonal element of the symmetric matrix  $A$  can be thought of a conditional relationship between the two corresponding features. According to a Gaussian graphical model,  $A$  served as a precision matrix  $\Sigma^{-1}$  to simulate the feature data, with additional estimation steps demonstrated in Danaher et al. (2014) to ensure  $\Sigma^{-1}$  a positive definite matrix. The feature data matrix  $X_0$  was generated such that  $X_0 \sim N(0, \Sigma)$ . The phenotype vector  $Y_0$  of size  $n \times 1$  was obtained as a linear function of  $X_0$ , as  $Y_0 = \beta_0 + X_0\beta + \epsilon$  with  $\epsilon = \{\epsilon_i\}_{i=1}^n$ , and  $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ . With the assumption that important nodes (i.e. nodes with high connectivity) had a large influence on the phenotype, we set  $\beta_0 \sim N(0, 1)$ , and  $\beta_{p \times 1} = (\beta_1, \dots, \beta_p)^T = (\sum_{l=1}^p a_{1l}, \dots, \sum_{l=1}^p a_{pl})^T$ , with  $a_{kl}$  as the  $(kl)$ th element of the adjacency matrix  $A$ . We further extended the simulation by perturbing the true data matrix to obtain the observed data matrix  $X$  such that  $X = X_0 + E$ , where  $E = \{e_{ij}\}_{i=1}^n \{j=1}^p$ , and  $e_{ij} \sim N(0, \sigma_e^2)$  denoted the noise matrix. This simulation setup is to mimic the real contexts, as follows. Given a sub-network of size  $p$ , we obtain the corresponding adjacency matrix

$A_{p \times p}$  by directly subsetting the global network instead of re-estimating it (say,  $A'_{p \times p}$ ) using the feature profile  $X_{n \times p}$  of only the nodes within the subnetwork. In other words,  $A_{p \times p}$  and  $A'_{p \times p}$  are different in a sense that  $A$  captures the global signals shared by all features in the dataset while  $A'$  only reflects local signals. Sequentially,  $X_0$  (true profile) of dimension  $n \times p$  and  $X_{n \times p}$  (observed profile) are induced from  $A$  and  $A'$ , respectively. However, in reality, we only observe  $X_{n \times p}$  which can be thought of as the feature profile being contaminated with measurement errors. By leveraging the network information inherent in  $A_{p \times p}$ , we would expect to recover some true underlying signals which might have been lost due to such measurement perturbations.

Furthermore, across the two scenarios, we rigorously investigated the impact of sample size on each method performance. More specifically, we started at the sample size of  $n = 1000$  subjects, and random subsamplings were iterated 1000 times for each sample size of 500, 300, 200, 100 and 50, respectively, to evaluate the robustness of each summarization regarding both criteria detailed in the next section. Note that at each sample size except for  $n = 1000$ , mean and standard deviation were calculated from the 1000 iterations.

### 2.2.1 Evaluation criteria

With the available observed data matrix  $X_{n \times p}$ , phenotype vector  $Y_0$  of size  $n \times 1$ , and the network adjacency matrix  $A_{p \times p}$ , we obtained the subject-level score vectors using NoNet and NetSHy approaches (Section 2), denoted as  $Z_{\text{NoNet}}$  and  $Z_{\text{NetSHy}}$ , respectively. We then evaluated the performance of the two scores using the following criteria:

- Correlation of each summarization score with the true phenotype  $Y_0$  was calculated as:

$$\begin{aligned} \rho_{\text{NoNet}} &= \text{Corr}(Z_{\text{NoNet}}, Y_0) \\ \rho_{\text{NetSHy}} &= \text{Corr}(Z_{\text{NetSHy}}, Y_0) \end{aligned}$$

- Proportion of variance explained (PVE) by each of the two summarization versions NoNet and NetSHy, using the associated first eigenvectors  $\phi_{\text{NoNet}}$  and  $\phi_{\text{NetSHy}}$ , respectively was computed as follows:

$$\begin{aligned} \text{PVE}_{\text{NoNet}} &= \frac{\sum_{i=1}^n \sum_{j=1}^p \{x_{ij}^{(0)} \phi_{\text{NoNet}}^j\}^2}{\sum_{i=1}^n \sum_{j=1}^p \{x_{ij}^{(0)}\}^2} \\ \text{PVE}_{\text{NetSHy}} &= \frac{\sum_{i=1}^n \sum_{j=1}^p \{x_{ij}^{(0)} \phi_{\text{NetSHy}}^j\}^2}{\sum_{i=1}^n \sum_{j=1}^p \{x_{ij}^{(0)}\}^2}, \end{aligned}$$

with  $x_{ij}^{(0)}$  as the  $(ij)$ th element of  $X_0$ ;  $\phi_{\text{NoNet}}^j$  as the  $j$ th element of  $\phi_{\text{NoNet}}$ ; and  $\phi_{\text{NetSHy}}^j$  as the  $j$ th element of  $\phi_{\text{NetSHy}}$ .

The two quantities above were compared to the optimal correlation and PVE, denoted as  $\rho_{\text{opt}}$  and  $\text{PVE}_{\text{opt}}$ , respectively, which were computed directly from the true data matrix  $X_0$ . In particular, the first PC and first eigenvector, denoted as  $Z_{\text{opt}}$  and  $\phi_{\text{opt}}$ , respectively, obtained from  $X_0$  were used to compute  $\rho_{\text{opt}}$  and  $\text{PVE}_{\text{opt}}$ , as follows:

$$\begin{aligned} \rho_{\text{opt}} &= \text{Corr}(Z_{\text{opt}}, Y_0) \\ \text{PVE}_{\text{opt}} &= \frac{\sum_{i=1}^n \sum_{j=1}^p \{x_{ij}^{(0)} \phi_{\text{opt}}^j\}^2}{\sum_{i=1}^n \sum_{j=1}^p \{x_{ij}^{(0)}\}^2}, \end{aligned}$$

with  $\phi_{\text{opt}}^j$  as the  $j$ th element of  $\phi_{\text{opt}}$ . The closer the values to  $\rho_{\text{opt}}$  and  $\text{PVE}_{\text{opt}}$ , the better the performance.

## 2.3 Biological networks

The applicability of NetSHy was further validated using biological networks [Column (3) of Supplementary Fig. S2] specific for COPD, regarding performance robustness and interpretable results. We used a M-P network for robustness assessment and a protein (P) network as a GWAS use case of the method. Note that the observed

data  $X$  and phenotype  $Y$  were used for the analysis directly without any simulation involved.

### 2.3.1 COPDGene and COPD phenotype

The COPDGene study is a multi-center study that enrolled 10 198 participants including non-Hispanic whites and African Americans with and without COPD between 2007 and 2011 (Visit 1). Five-year follow-up visits took place from 2013 to 2017 (Visit 2). Study participants from Visit 2, after removing individuals with lung transplant or lung reduction surgery and never smokers, provided consent; and their blood samples were used for -omic analyses.

COPD was defined by spirometric evidence of airflow obstruction, which was computed as a ratio of post-bronchodilator forced expiratory volume at one second (FEV1) to forced vital capacity. FEV1% is the amount of air one can forcibly exhale in one second divided by the predicted FEV1 adjusted for age, height, race and sex (Hankinson *et al.*, 1999). The global obstructive lung disease (GOLD) system is used to grade COPD. More information on the GOLD system can be found in [Supplementary Section S.2](#).

### 2.3.2 COPDGene genotyping

COPDGene subjects were of self-reported non-Hispanic white or African-American ancestry, and genotyped as previously described by Cho *et al.* (2014). Briefly, genotyping was performed using the HumanOmniExpress array, and BeadStudio quality control, including reclustering on project samples was performed following Illumina guidelines. Subjects and markers with a call rate of  $<95\%$  were excluded. Population stratification exclusion and adjustment on self-reported white subjects was performed using EIGENSTRAT (EIGENSOFT Version 2.0).

### 2.3.3 Proteomic data

The following two platforms were used to quantify proteomic data in Visit 2 of COPDGene. SOMAScan v1.3: P100 plasma was profiled using SOMAScan<sup>®</sup> Human Plasma 1.3K assay (SomaLogic, Boulder, CO, USA) at National Jewish Health. SOMAScan is a multiplex proteomic assay quantified by microarrays. This assay measured 1317 SOMAmers which are short single-stranded deoxyoligonucleotides (aptamers) binding with high affinity and specificity to specific protein structures (Gold *et al.*, 2010). SOMAScan v4.0: This SOMAScan platform used 4979 different SOMAmers to quantify 4776 unique proteins with 4720 unique Uniprot numbers. Details on the preprocessing steps of the proteomic data are given in [Supplementary Section S.2](#).

### 2.3.4 Metabolomic data

The same P100 plasma was profiled using the Metabolon (Durham, NC, USA) Global Metabolomics platform to quantify 1392 metabolites. After filtering for missing values, 995 metabolites were used in the analysis. More details can be found in [Supplementary Section S.2](#).

### 2.3.5 M-P network construction

We used a subset of the COPDGene participants who had both metabolomic and proteomic data available at Visit 2 to construct a M-P network via sparse multiple canonical correlation network (SmCCNet) introduced by Shi *et al.* (2019). The two -omic data were adjusted for white blood cell count, percent eosinophil, percent lymphocytes, percent neutrophils and hemoglobin as these covariates may influence metabolite and protein abundance in human blood studies. Then, SmCCNet was applied to the adjusted metabolomic ( $p_1 = 995$  metabolites) and proteomic ( $p_2 = 1317$  proteins) data to construct multi-omic networks correlated with the phenotype FEV1% ( $n = 994$  subjects) via multiple canonical correlation approach. In essence, SmCCNet maximized the correlation between the two omics datasets (i.e. metabolomics and proteomics) and the phenotype FEV1% while imposing a sparsity to de-emphasize the impacts of metabolites and proteins which did not contribute to the overall correlation. After hierarchical clustering and hard

thresholding to filter out weak edges, strongly connected subnetworks that were well correlated with FEV1% were identified. More details can be found in Mastej *et al.* (2020). In this work, we used a M-P network for FEV1% consisting of 7 metabolites and 13 proteins.

**2.3.5.1 Robustness assessment.** We assume that the collected metabolomics and proteomics data (i.e.  $X$ ) are perturbed measurements due to instrument error of the true but non-observed metabolite and protein (i.e.  $X_0$ ) levels. With  $X_0$  not available, the comparison of NetSHy and NoNet relative to the optimal level (Section 2.2) was not obtainable. We instead focused on assessing the robustness of the two approaches regarding the correlation with the observed phenotype as the sample size decreased. The observed data corresponding to the identified M-P network  $X_{994 \times 20}$  and phenotype  $Y_{994 \times 1}$  were randomly sampled at decreasing sizes: 500, 300, 200, 100 and 50, respectively, and repeated for 1000 iterations. Mean and standard deviation of the correlation of each summarization:  $Z_{\text{NetSHy}}$ ,  $Z_{\text{NoNet}}$  with the observed phenotype were recorded at each sample size, except for the full sample size  $n = 994$ . The robust performances of NetSHy and NoNet were assessed using the dropping rates as the sample size decreased.

### 2.3.6 P network construction

We used a subset of the COPDGene non-Hispanic white participants with proteomic data available at Visit 2 to construct protein networks. The proteomic data collected from SOMAScan v4.0 platform (Section 2.3.3) had larger sample size to perform a GWAS since the data did not need to be matched with the metabolomic data. Similar to M-P network construction, SmCCNet (Shi *et al.*, 2019) was applied to the proteomic data ( $p = 4776$  proteins) to construct protein (P) networks maximizing correlation with the phenotype FEV1% ( $n = 1660$  subjects). A 5-fold cross-validation was used on a set of sparsity parameters from 0.1 to 0.5 with a step size of 0.1 to select an optimal value that minimized the prediction error. After hierarchical clustering and weak edge trimming, a strongly connected network of 16 proteins and well correlated with FEV1% was identified.

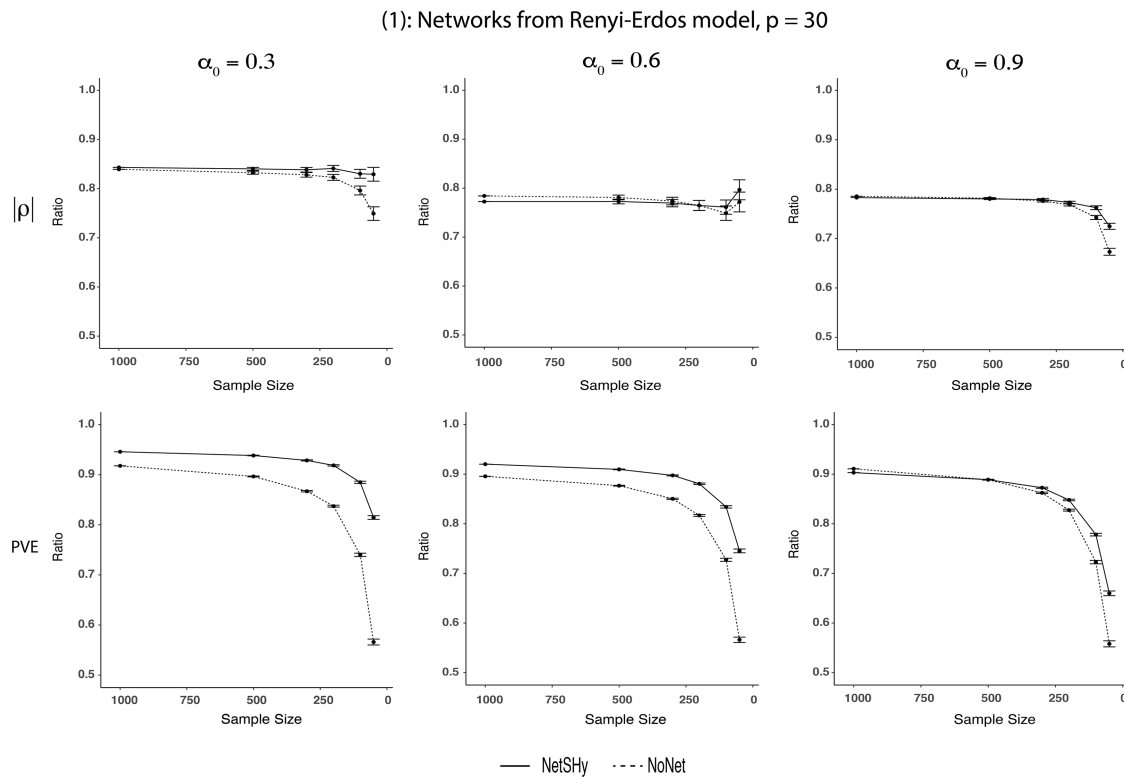
**2.3.6.1 GWAS analysis.** GWAS analysis was used to demonstrate the applicability of the summarization methods (i.e. NetSHy and NoNet) in an example downstream analysis. Specifically, by applying each approach to the identified protein network ( $X_{1660 \times 16}$ ,  $Y_{1660 \times 1}$ ), we obtained summarization scores,  $Z_{\text{NetSHy}}$  and  $Z_{\text{NoNet}}$ . The resulting summarization scores were inverse-normalized prior to linearly regressing on the genotype while adjusting for covariates including age, body mass index (BMI), gender, smoking status and five genetic PCs (Sun *et al.*, 2016). The genetic PCs were obtained from previously performed analysis including only COPDgene participants (Cho *et al.*, 2014). In total, there were 14 553 332 variants tested for significant association with the protein network across the subjects. [Supplementary Section S.2](#) includes our detailed GWAS analysis.

## 3 Results

### 3.1 Simulation results

[Figure 1](#) depicts the performances of NetSHy and NoNet summarization scores in terms of correlation ( $\rho$ ) with true phenotype (top row) and proportion of variance explained (PVE) in true data matrix  $X_0$  (bottom row) with respect to the optimal quantities  $\rho_{\text{opt}}$  and  $\text{PVE}_{\text{opt}}$ , respectively. The closer the values to 1, the better the performances. In addition, the network size was fixed at  $p = 30$  while the level of sparsity increased from  $\alpha_0 = 0.3$  to  $\alpha_0 = 0.9$ . Across the three sparsity levels, both approaches deviated from the optimal level as the sample size decreased. In general, the trend was observed that NoNet dropped at a faster rate compared to NetSHy. Interestingly, NetSHy had higher PVE in comparison with NoNet regardless of sample size and network sparsity. However, as the





**Fig. 1.** Results of simulation scenario (1),  $p = 30$ : Fixing network size at  $p = 30$  while varying sparsity levels from  $\alpha_0 = 0.3$  to  $\alpha_0 = 0.9$ , NetSHy and NoNet were assessed using correlation with phenotype ( $\rho$ ) and proportion of variance explained (PVE) relative to the optimal level, as sample size decreased. Specifically, the sample size was started at 1000 subjects, and random subsamplings were iterated 1000 times for each sample size of 500, 300, 200, 100 and 50, respectively. The closer the value to 1, the better the performance. The error bars summarize the standard deviations of  $\rho$  and PVE from the 1000 iterations at each sample size except for  $n = 1000$ . The range of  $\rho$  and PVE ratio in the y-axis is between 0 and 1. However, we have zoomed in between 0.5 and 1 for better visualization

nodes were connected more densely, i.e. larger  $\alpha_0$ , the deviation in PVE between the two approaches became less apparent. More specifically, at  $\alpha_0 = 0.3$  and  $n = 50$ , the ratio of  $\text{PVE}_{\text{NetSHy}}$  to the optimal  $\text{PVE}_{\text{opt}}$  was 0.81 while such ratio of NoNet was 0.57. However, when  $\alpha_0 = 0.9$ , the PVEs of NetSHy and NoNet with respect to the optimal PVE were 0.66 and 0.56, respectively. A similar pattern was observed for  $\rho$ , but at a more subtle level. For instance, at  $\alpha_0 = 0.3$  and  $n = 50$ , the correlations  $\rho_{\text{NetSHy}}$  and  $\rho_{\text{NoNet}}$  with respect to the optimal correlation  $\rho_{\text{opt}}$  were 0.83 and 0.75, respectively. When the sparsity level increased to  $\alpha_0 = 0.9$ , the ratio of  $\rho_{\text{NetSHy}}$  to  $\rho_{\text{opt}}$  was 0.72 while such ratio of NoNet was 0.67. [Supplementary Figures S3 and S4](#) show the same set of results for network size  $p = 60$  and  $p = 100$ , respectively. Slightly different from the previous case, NetSHy still showed some improvement over NoNet in both  $\rho$  and PVE at sparsity level  $\alpha_0 = 0.3$  when  $p = 100$ . However, in the case of a densely connected network at  $\alpha_0 = 0.9$ , NetSHy and NoNet performed almost identical.

**Figure 2** illustrates the performances of NetSHy and NoNet summarization scores in the empirical-based simulation [Simulation scenario (2) in [Supplementary Fig. S2](#)] at two levels of network sparsity  $\alpha_0 = 0.25$  and  $0.51$  across decreasing sample sizes. Similar to simulation scenario (1), the two approaches deviated from the optimal level as sample size decreased. Though at  $\alpha_0 = 0.25$ , NetSHy suffered a little in recovering true correlation with the phenotype at sample sizes of 1000 and 500, it was more robust at more extreme sizes of 100 and 50. Specifically, at  $n = 1000$ , the ratio of NetSHy correlation  $\rho_{\text{NetSHy}}$  to the optimal level  $\rho_{\text{opt}}$  was 0.63 while that ratio of NoNet was 0.68. However, at the smallest sample size of  $n = 50$ , the correlation of NoNet scores with the phenotype decreased greatly, which caused its ratio with the optimal correlation  $\rho_{\text{opt}}$  to drop to 0.47 while such ratio of NetSHy remained around 0.62. At  $\alpha_0 = 0.51$ , NetSHy and NoNet performed almost identical when sample sizes were large. In particular, at  $n = 1000$ ,

the ratios of NetSHy and NoNet correlations with respect to the optimal level were 0.66 and 0.67, respectively. The improvement of NetSHy over NoNet was more appreciable towards the extreme sizes of 100 and 50. More precisely, at  $n = 50$ , NetSHy maintained the correlation ratio to the optimal level at around 0.66 while that ratio using NoNet scores reduced to 0.55. Regarding PVE, NetSHy outperformed NoNet approach at all sample sizes. Similar to what had been observed, the improvement of NetSHy over NoNet was more substantial towards smaller sample sizes.

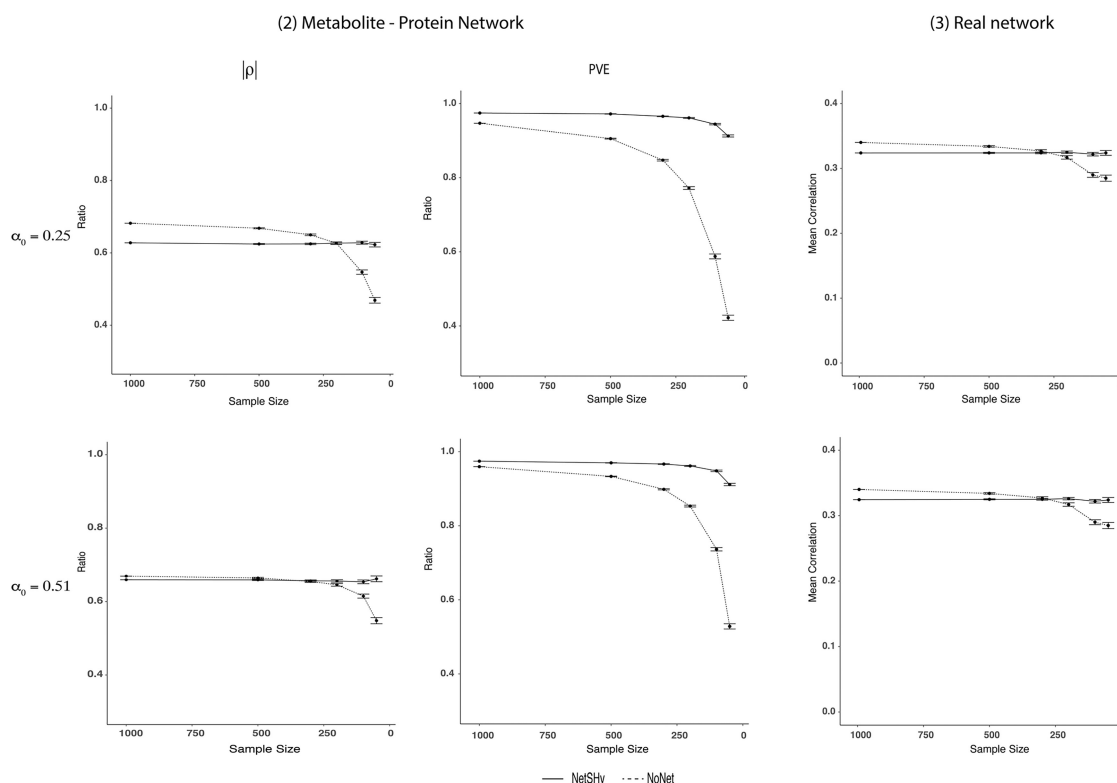
## 3.2 Application results

The evaluation of NetSHy and NoNet was further validated using biological networks, M-P and P networks, with respect to performance robustness and GWAS results.

### 3.2.1 Robustness

Last column of **Figure 2** presents the mean correlation with observed phenotype  $\rho$  using the M-P network of size  $p = 20$  at two sparsity levels  $\alpha_0 = 0.25$  and  $\alpha_0 = 0.51$ . Similar patterns were observed across the two  $\alpha_0$  levels that the correlation  $\rho$  dropped as the sample size got smaller. At the original sample size  $n = 994$ , the observed NoNet correlation with the phenotype ( $|\rho_{\text{NoNet}}| = 0.34$ ) was slightly higher than the corresponding NetSHy correlation ( $|\rho_{\text{NetSHy}}| = 0.32$ ). The difference between these two correlations was not significant with  $P$ -value of 0.49 using bootstrapping. Interestingly, the overall trajectory of NetSHy remained relatively stable even at the small sample size ( $n \leq 100$ ) while NoNet suffered a substantial drop. For instance, at  $n = 50$  and  $\alpha_0 = 0.51$ , the mean correlation of NetSHy was 0.32 while that of NoNet reduced to 0.28.

During the subsampling process, as we selected 500 subjects out of 994 without replacement, the subjects across iterations



**Fig. 2.** Results of simulation scenario (2) and real network (3): *First two columns:* simulation based on a published M-P network with fixed  $p=20$  at two sparsity levels of  $\alpha_0 = 0.25$  (top row)  $\alpha_0 = 0.51$  (bottom row). NetSHy and NoNet were assessed using correlation with phenotype ( $\rho$ ) and proportion of variance explained (PVE) relative to the optimal level, as sample size decreased. Specifically, the sample size was started at 1000 subjects, and random subsamplings were iterated 1000 times for each sample size of 500, 300, 200, 100 and 50, respectively. The closer the value to 1, the better the performance. The range of  $\rho$  and PVE ratio in the y-axis is between 0 and 1. However, we have zoomed in between 0.4 and 1 for better visualization. *Last column:* Published M-P network with fixed  $p=20$  at two sparsity levels of  $\alpha_0 = 0.25$  (top)  $\alpha_0 = 0.51$  (bottom) was used directly for evaluation. Without knowledge of true underlying data matrix  $X_0$ , proportion of variance explained (PVE) was not assessed. Instead, the robustness of each approach regarding observed correlation with phenotype was of interest as the sample size decreased. Specifically, random subsamplings were iterated 1000 times for each sample size of 500, 300, 200, 100 and 50, respectively. The lower the dropping rate of a method's trajectory, the more robust the performance. The error bars summarize the standard deviations of  $\rho$  from the 1000 iterations at each sample size except for  $n = 994$

overlapped to different degrees due to random chance. Intuitively, the overlap was greater for larger sample sizes, leading to less variation across iterations. As such, we observed larger computed standard deviations of correlation with phenotype ( $\rho$ ) and proportion of variance explained (PVE) as we decreased sample size from  $n = 500$  to  $n = 50$  (Fig. 2). Similar patterns were also observed in the simulation studies (Fig. 1).

### 3.2.2 GWAS results and interpretation

To demonstrate a downstream application of network summarization, we tested whether any single nucleotide polymorphisms (SNPs) had a significant association with the protein network across the subjects. This analysis would be useful for identifying potential regulators of the network. Figure 3a and b show the GWAS results for NetSHy and NoNet, respectively. At a threshold level of  $5 \times 10^{-8}$ , NetSHy identified 24 significant SNPs while NoNet detected only one. The top SNP, rs1017301 (Chromosome 12: 9210335), was discovered using NetSHy score ( $p = 2.38 \times 10^{-13}$ , minor allele frequency (MAF)=0.33) whereas the same SNP did not reach significance ( $p = 4.26 \times 10^{-4}$ ) using the NoNet approach. For NoNet summary score, the top SNP was rs118028480 (Chromosome 22: 39592172,  $p = 4.35 \times 10^{-8}$ ). Supplementary Section S.5 provides a full interpretation of the significant SNPs obtained using NetSHy and NoNet summarization scores.

Recall that NetSHy summary score is a weighted average abundance of all proteins in the network with the relative weights determined by performing PCA on the combination of network topology and the corresponding node feature profiles (i.e.  $X^*$  in Section 2.1). Supplementary Figure S5 shows weights of five proteins which

contribute the most to the NetSHy summary score. The five proteins included fructose-bisphosphate aldolase B, fructose-1,6-bisphosphatase 1, argininosuccinate lyase, ferritin and ferritin light chain. The correlation between the NetSHy summary score and FEV1% is 0.36. Interestingly, by checking the correlation of each individual protein with FEV1%, we noticed that the absolute values of the correlation ranged from 0.14 to 0.30. In other words, by using a summary score as an aggregate of all proteins in the network, we saw an increase in the correlation with the phenotype. A detailed description of the relationship between the top proteins with lung diseases, particularly COPD, is given in Supplementary Section S.5.

## 4 Discussion and conclusions

Biological networks provide a system-level understanding of the underlying cellular processes, but they are often too large to be considered as a whole. As a result, subsets of nodes (i.e. modules) which are highly connected to each other may be considered. Furthermore, a purpose of many network analyses is to relate the resulting modules to external sample information in downstream analyses, depending on the research question of interest. However, due to the multidimensional nature of networks, they need to be summarized prior to subsequent analyses. Conventional approaches rely on the feature profiles of the within-network entities while disregarding the inherent connectivity properties to obtain a network representation. As such, the summarization results do not truly reflect the roles of individual biological entities in the network. This motivates us to propose NetSHy, a hybrid approach which is capable of reducing the dimension of networks while accounting for both node profiles

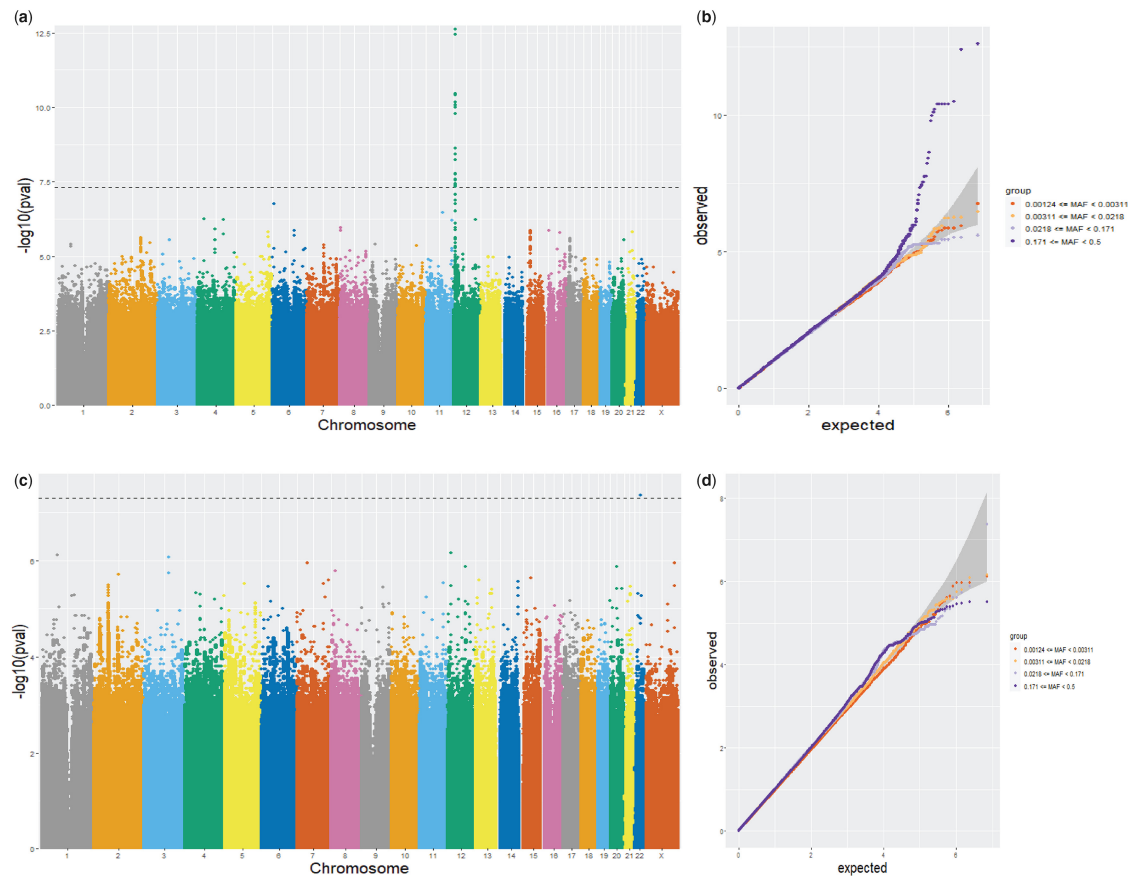


Fig. 3. Results from GWAS studies on protein network specific to FEV1. Top row: NetSHy summarization was used to linearly regress on genotype to identify significantly associated SNPs. (a) Manhattan plot using NetSHy summarization. (b) QQ plot using NetSHy summarization. Bottom row: NoNet summarization was used to linearly regress on genotype to identify significantly associated SNPs. (c) Manhattan plot using NoNet summarization. (d) QQ plot using NoNet summarization

and topological properties. In our preliminary analysis (Supplementary Section S.6), we explored two methods to incorporate topology in network summarization, i.e. a diffusion process (Leiserson *et al.*, 2015; Dimitrakopoulos *et al.*, 2018) and a weighted approach accounting for a secondary proximity embedded in a topology overlap matrix (Zhang and Horvath, 2005). However, we did not pursue further with the comparisons due to instability or suboptimal results. Thus, we only compare the performance of NetSHy with NoNet (i.e. not including network information) through simulation scenarios based on random and empirical networks at varying levels of network size and sparsity, with regard to the ability to recover true correlation with the phenotype of interest and the amount of true variation explained. Furthermore, the robustness of the two approaches is assessed using biological networks via repeated subsamplings at a decreasing level. Finally, we validate the applicability of NetSHy and NoNet approach using the GWAS analysis.

NetSHy outperforms the NoNet approach regarding both correlation with true phenotype ( $\rho$ ) and proportion of variance explained (PVE), when the networks are relatively small and sparse. However, when networks increase in size and the nodes are more densely connected, the improvement of NetSHy over NoNet is not as pronounced. This is not unexpected as when almost every node in the network is interconnected, the connectivity roles of individual nodes are similar. Thus, leveraging topological properties in this scenario might provide no additional gain for NetSHy, as compared to the NoNet approach. In applications on biological networks, the robustness of both approaches is brought into focus, due to lacking true underlying relationship between phenotype and feature data. In the M-P network, the observed correlation with the COPD phenotype FEV1% of NetSHy is slightly lower than that of NoNet at full

sample size (0.34 versus 0.32). Though the difference was insignificant ( $p = 0.49$ ), it is still worth noting. However, by random subsampling at reducing sizes, NetSHy's trajectory of observed correlation with the phenotype drops at a slower rate compared to NoNet, indicating NetSHy is more robust to small sample sizes. Finally, in the GWAS analysis of a protein network, NetSHy and NoNet summarization scores are used as response variables in a linear regression framework with genotype and other covariates. NetSHy identifies more significant SNPs associated with a given network, compared to NoNet approach.

We have presented promising results of NetSHy in representing networks at the subject level; however, we have still relied on the linearity assumption of the classical dimension reduction PCA. Additionally, the topological properties stored in the Laplacian matrix might not be sufficient for capturing the local neighborhood structure when the networks grow bigger and/or denser, as seen in the 'large  $p$ , large  $\alpha_0$ ' simulation scenario. We could potentially leverage the Isomap approach (Tenenbaum *et al.*, 2000) to modify  $X^*$ . Particularly, for any two nodes in a network, instead of their direct connection, the geodesic distances computing their shortest path distances could be used to represent the connectivity measure. Such connectivity matrix would then replace the  $L$  matrix in the calculation of  $X^*$ . Lastly, kernel PCA (Jin *et al.*, 2015) could be applied on  $X^*$  to extract the low-dimensional non-linear representation. Alternatively, we have considered different techniques extracting information contained in large PPI networks such as FUSE (Bhowmick and Seah, 2016), VoG (Koutra *et al.*, 2014), GraSS (LeFevre and Terzi, 2010), SNAP and k-SNAP (Tian *et al.*, 2008), CANAL (Zhang *et al.*, 2010). However, the summaries acquired from these approaches are themselves graphs. We could potentially use these approaches in place of the thresholding counterpart to simultaneously

trim off weak edges and simplify the networks prior to summarizing them. This work is currently under investigation.

## Acknowledgements

The COPDGene® project is supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer-Ingelheim, Genentech, GlaxoSmithKline, Novartis and Sunovion.

## Funding

This work was supported by the National Heart, Lung, and Blood Institute, National Institutes of Health (R01 HL152735, R01 HL137995, U01 HL089897 and U01 HL089856). Any opinions expressed in this document are those of the author(s) and do not necessarily reflect the views of National Heart, Lung, and Blood Institute, National Institutes of Health or affiliated organizations and institutions. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health.

*Conflict of Interest:* none declared.

## Data availability

COP DGene Genotype data and SomaScan can be found on dbGaP for COP DGene (phs000179). COP DGene metabolomic data are available at the NIH Common Fund's National Metabolomics Data Repository (NMDR) website, the MetabolomicsWorkbench, (<https://www.metabolomicsworkbench.org>).

## References

Alexander,R.P. *et al.* (2009) Understanding modularity in molecular networks requires dynamics. *Sci. Signal.*, **2**, pe44.

Bartel,J. *et al.* (2015) The human blood metabolome-transcriptome interface. *PLoS Genet.*, **11**, e1005274.

Belkin,M. and Niyogi,P. (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, **15**, 1373–1396.

Bhowmick,S.S. and Seah,B.S. (2016) Clustering and summarizing protein-protein interaction networks: a survey. *IEEE Trans. Knowl. Data Eng.*, **28**, 638–658.

Caetano-Anollés,G. *et al.* (2019) Emergence of hierarchical modularity in evolving networks uncovered by phylogenomic analysis. *Evol. Bioinform. Online*, **15**, 1176934319872980.

Cai,H. *et al.* (2018) A comprehensive survey of graph embedding: problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.*, **30**, 1616–1637.

Chen,X. *et al.* (2006) BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. *Bioinformatics*, **22**, 2952–2954.

Cho,M.H. *et al.*; NETT Genetics, ICGN, ECLIPSE and COPDGene Investigators. (2014) Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir. Med.*, **2**, 214–225.

Choobdar,S. DREAM Module Identification Challenge Consortium. *et al.* (2019) Assessment of network module identification across complex diseases. *Nat. Methods*, **16**, 843–852.

Csardi,G. *et al.* (2006) The igraph software package for complex network research. *InterJ. Complex Syst.*, **1695**, 1–9.

Danaher,P. *et al.* (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B*, **76**, 373–397.

Dimitrakopoulos,C. *et al.* (2018) Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics*, **34**, 2441–2448.

Erdos,P. *et al.* (1960) On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, **5**, 17–60.

Fiscon,G. *et al.* (2018) Network-based approaches to explore complex biological systems towards network medicine. *Genes*, **9**, 437.

Gold,L. *et al.* (2010) Aptamer-based multiplexed proteomic technology for biomarker discovery. *Nat. Prec.*, **1**–1.

Grover,A. and Leskovec,J. (2016) node2vec: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA*, pp. 855–864.

Hankinson,J.L. *et al.* (1999) Spirometric reference values from a sample of the general us population. *Am. J. Respir. Crit. Care Med.*, **159**, 179–187.

Hawe,J.S. *et al.* (2019) Inferring interaction networks from multi-omics data. *Front. Genet.*, **10**, 535.

Hofmann,T. and Buhmann,J. (1995) Multidimensional scaling and data clustering. In: Leen,T.K. *et al.* (eds) *Advances in Neural Information Processing Systems*. Neural Information Processing Systems Foundation, Inc. (NeurIPS), Denver, CO, USA. pp. 459–466.

Hu,J.X. *et al.* (2016) Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.*, **17**, 615–629.

Jin,T. *et al.* (2015) Low-rank matrix factorization with multiple hypergraph regularizer. *Patt. Recogn.*, **48**, 1011–1022.

Koutra,D. *et al.* (2014) VoG: summarizing and understanding large graphs. In: *Proceedings of the 2014 SIAM international Conference on Data Mining*, pp. 91–99. SIAM.

Langfelder,P. and Horvath,S. (2008) WGCNA: an r package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 1–13.

Langfelder,P. *et al.* (2013) When is hub gene selection better than standard meta-analysis? *PLoS One*, **8**, e61505.

LeFevre,K. and Terzi,E. (2010) GraSS: graph structure summarization. In: *Proceedings of the 2010 SIAM International Conference on Data Mining*, pp. 454–465. SIAM.

Leiserson,M.D. *et al.* (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, **47**, 106–114.

Mastej,E. *et al.* (2020) Identifying protein-metabolite networks associated with COPD phenotypes. *Metabolites*, **10**, 124.

Perozzi,B. *et al.* (2014) Deepwalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–710.

Pujana,M.A. *et al.* (2007) Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat. Genet.*, **39**, 1338–1349.

Ravasz,E. *et al.* (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Schlosser,P. GCKD Investigators. *et al.* (2020) Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. *Nat. Genet.*, **52**, 167–176.

Sevimoglu,T. and Arga,K.Y. (2014) The role of protein interaction networks in systems biomedicine. *Comput. Struct. Biotechnol. J.*, **11**, 22–27.

Shi,W.J. *et al.* (2019) Unsupervised discovery of phenotype-specific multi-omics networks. *Bioinformatics*, **35**, 4336–4343.

Shu,L. *et al.*; Cardiogenics Consortium. (2017) Shared genetic regulatory networks for cardiovascular disease and type 2 diabetes in multiple populations of diverse ethnicities in the United States. *PLoS Genetics*, **13**, e1007040.

Spitzer,F. (2013) *Principles of Random Walk*, Vol. 34. Springer Science & Business Media, New York, NY.

Sun,W. *et al.*; COPDGene Investigators. (2016) Common genetic polymorphisms influence blood biomarker measurements in COPD. *PLoS Genet.*, **12**, e1006011.

Tenenbaum,J.B. *et al.* (2000) A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2323.

Tian,F. *et al.* (2014). Learning deep representations for graph clustering. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28.

Tian,Y. *et al.* (2008) Efficient aggregation for graph summarization. In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp. 567–580.

Valentini,G. *et al.* (2014) An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. *Artif. Intell. Med.*, **61**, 63–78.

Vincent,P. *et al.* (2010) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, **11**, 3371–3408.

Wang,D. *et al.* (2016) Structural deep network embedding. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1225–1234.

Wu,Y. *et al.* (2022) Identification of a four-gene signature associated with the prognosis prediction of lung adenocarcinoma based on integrated bioinformatics analysis. *Genes*, **13**, 238.

Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, **4**.

Zhang,N. *et al.* (2010) Discovery-driven graph summarization. In: *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pp. 880–891. IEEE.

Zhang,P. and Itan,Y. (2019) Biological network approaches and applications in rare disease studies. *Genes*, **10**, 797.