Inferring a directed acyclic graph of phenotypes from GWAS summary statistics

Rachel Zilinskas[1,*], Chunlin Li[2,*], Xiaotong Shen[3], Wei Pan[4,], Tianzhong Yang[4,***]**

[1]Statistics and Data Corporation, Tempe, Arizona 85288, U.S.A.

[2]Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.

[3]School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.

[4]Division of Biostatistics and Health Data Science, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.

* Equal contribution

** *Email:* panxx014@umn.edu

*** *Email:* yang3704@umn.edu

Summary: Estimating phenotype networks is a growing field in computational biology. It deepens the understanding of disease etiology and is useful in many applications. In this study, we present a method that constructs a phenotype network by assuming a Gaussian linear structure model embedding a directed acyclic graph (DAG). We utilize genetic variants as instrumental variables and show how our method only requires access to summary statistics from a genome-wide association study (GWAS) and a reference panel of genotype data. Besides estimation, a distinct feature of the method is its summary statistics-based likelihood ratio test on directed edges. We applied our method to estimate a causal network of 29 cardiovascular-related proteins and linked the estimated network to Alzheimer's disease (AD). A simulation study was conducted to demonstrate the effectiveness of this method. An R package `sumdag` implementing the proposed method, all relevant code, and a Shiny application are available at `https://github.com/chunlinli/sumdag`.

Key words: Alzheimer's disease (AD), directed acyclic graph (DAG), genome-wide association study (GWAS), likelihood ratio test, proteomics

1. Introduction

Network analysis has deepened our understanding of biological mechanisms and disease etiologies (Zhang and Itan, 2019). Specifically, protein-protein interaction (PPI) networks that capture the interplay of proteins in the biomolecular systems are vital for normal cell functions (Snider et al., 2015). Disturbing of the normal pattern in the PPI network can be causative to or indicative of a disease state. Studies have linked co-regulatory networks of proteins to a variety of complex diseases (Ross and Poirier, 2004; Emilsson et al., 2018). Recently, a network-based method modeling PPI boasted high accuracy rates in cancer prediction (Id et al., 2021). Cheng et al. (2021) further showed that disease-associated variants were significantly enriched in the sequences coding PPI interfaces compared to variants in healthy individuals. Their work also demonstrated associations of PPIs with drug resistance and overall survival, highlighting the use of protein networks for informing genotype-based therapy. Network-based analyses have shown their potential in advancing precision medicine for complex diseases over traditional approaches which focus on monogenic mutations and independent assessment of risk factors (Napoli et al., 2020).

Network analyses can be categorized into two groups. One utilizes only phenotypic data to construct networks. For example, weighted gene network co-expression analysis estimates an undirected network which is further characterized using dimension reduction techniques (Zhang and Horvath, 2005). Graphical lasso formulation employs penalized methods to estimate a Gaussian graphical model for a large number of variables (Witten et al., 2012). Bayesian network analysis (Friedman et al., 2000) estimates directed acyclic graphs (DAG), which are widely accepted in biological systems (Ashburner et al., 2000), and its recent improvements in computational approaches have led to much shortened computational time (Liu et al., 2016). The other group of methods exploits the use of instrumental variable (IV) techniques to estimate a DAG, assuming a linear structural equation model. Chen et al.

(2018) developed a penalized two-stage least squares method to estimate a DAG, assuming known intervention targets. Li et al. (2023) further extended the work to accommodate unknown intervention targets commonly encountered in biological applications.

Individual-level data are required for all the methods above, which, however, can be difficult to obtain, especially for human studies, due to logistic limitations and privacy concerns. On the other hand, many genome-wide association studies (GWAS) have shared their summary statistics publicly, generating a rich and valuable data resource. Thus, we propose adapting the network estimation and inference methods of Li et al. (2023) to rely only on GWAS summary statistics and a genetic reference panel, both much more easily accessible. We will show how a DAG can be estimated for cardiovascular-related proteins using a large-scale proteomic GWAS summary dataset, and then link the protein network to Alzheimer's disease (AD). The algorithm for the proposed work is packaged in R. Our work represents one of the initial attempts to utilize GWAS summary statistics in the construction of a DAG. We expect that our work can facilitate more comprehensive network analysis in studying biological and medical relationships. In addition to inferring PPI network, our method is readily applicable to understand the interplay of many other molecular and non-molecular phenotypes, as long as the corresponding GWAS summary statistics are available.

2. Methods

2.1 *Network modeling and data*

2.1.1 *Directed phenotype network.* Our goal is to use genotypes as external interventions to construct and infer a DAG that describes the directed relationships among a set of phenotypes. In the framework of interventional Gaussian DAG (Li et al., 2023), we assume

$$\boldsymbol{Y} = \boldsymbol{Y}\mathbf{U} + \boldsymbol{X}\mathbf{W} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P)$ is the $N \times P$ data matrix of P phenotypes, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_Q)$ is the $N \times Q$ data matrix of Q genotypes serving as IVs, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_P)$ is the $N \times P$ error matrix with each row sampled from $N\{\mathbf{0}, \mathrm{Diag}(\omega_1^2, \ldots, \omega_P^2)\}$, and N is the sample size. Note that Equation (1) lacks an intercept because we assume phenotype and genotype are centered at mean 0, which could be easily done with individual-level GWAS data.

In Equation (1), $\mathbf{U}$ and $\mathbf{W}$ are unknown parameters to be estimated. The $P \times P$ matrix $\mathbf{U} = (\mathrm{u}_{kj})$ specifies the network structure such that $\mathrm{u}_{kj} \neq 0$ indicates a directed relation from phenotype k to phenotype j. The $Q \times P$ matrix $\mathbf{W} = (\mathrm{w}_{qp})$ specifies the targets and strengths of interventions in that $\mathrm{w}_{qp} \neq 0$ indicates an interventional relation from genotype q to phenotype p. Let $\mathcal{E} = \{(k, j) : \mathrm{u}_{kj} \neq 0\}$ be the set of directed relations and $\mathcal{I} = \{(q, p) : \mathrm{w}_{qp} \neq 0\}$ be the set of interventional relations.

2.1.2 *Summary statistics and reference panel.* In Equation (1), the data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ contain individual information such that each row represents the variables measured on an individual. On the other hand, GWAS summary statistics aggregate N observations into a single measure for each single nucleotide polymorphism (SNP) across the whole genome. This measure is the average effect of having one copy of the effect allele of the SNP on the phenotype being studied. It is estimated by $\widehat{\beta}_{qp} \approx (\boldsymbol{x}_q^\top \boldsymbol{x}_q)^{-1} \boldsymbol{x}_q^\top \boldsymbol{y}_p$, often reported along with accompanying statistics such as the corresponding standard error $\mathrm{SE}(\widehat{\beta}_{qp})$, z-score z_{qp}, sample size N, reference allele (REF), minor allele frequency (MAF) and p-value. The summary statistics of the Q SNPs in $\mathbf{W}$ are included in the GWAS summary data.

As a complement to the summary-level data, a reference panel comprising genotypic data of individuals from a general population provides the correlation structure among the genotypes. Many existing resources can be used for such a reference panel (The 1000 Genomes Project Consortium, 2015; International HapMap Consortium, 2005; Taliun et al., 2021; Bycroft et al., 2018). Given an $N_r \times Q$ (centered) reference panel $\boldsymbol{X}_r$ of N_r individuals,

we follow the conventional suggestion (Mak et al., 2017) to regularize the genetic correlation matrix $\boldsymbol{R}$ such that $\boldsymbol{R} = (1 - s_p)\boldsymbol{X}_r^\top \boldsymbol{X}_r + s_p \boldsymbol{I}$, where $0 \leqslant s_p \leqslant 1$ is a real number controlling the degree of regularization.

From the summary statistics and the reference panel, we compute the following quantities that are used for the construction and inference of the directed phenotype network. The subsequent computation also assumes $\boldsymbol{X}$ and $\boldsymbol{Y}$ are centered, which does not influence $\widehat{\beta}_{qp}$ and the accompanying statistics.

- The covariance matrix of genotypes $\frac{1}{N}\boldsymbol{X}^\top \boldsymbol{X}$ is estimated by $\frac{1}{N}\widehat{\boldsymbol{X}^\top \boldsymbol{X}} = \frac{1}{N_r}\boldsymbol{R}$.
- Let $s_q^2 = \frac{1}{N}\boldsymbol{x}_q^\top \boldsymbol{x}_q$. Then s_q^2 is estimated by $\widehat{s}_q^2 = 2 \times \mathrm{MAF} \times (1 - \mathrm{MAF})$ provided that MAF is reported in the summary statistics, or otherwise estimated by $\widehat{s}_q^2 = \frac{1}{N_r}(\boldsymbol{X}_r^\top \boldsymbol{X}_r)_{qq}$, the qth diagonal element of $\frac{1}{N_r}\boldsymbol{X}_r^\top \boldsymbol{X}_r$.
- Given $\widehat{s}_q^2$, $\frac{1}{N}\boldsymbol{X}^\top \boldsymbol{y}_p$ is estimated by $\frac{1}{N}\widehat{\boldsymbol{X}^\top \boldsymbol{y}_p} = \left(\widehat{s}_1^2\widehat{\beta}_{1p}, \widehat{s}_2^2\widehat{\beta}_{2p}, \ldots, \widehat{s}_Q^2\widehat{\beta}_{Qp}\right)^\top$.
- For $\frac{1}{N}\boldsymbol{y}_p^\top \boldsymbol{y}_p$, we use the median estimate $\frac{1}{N}\widehat{\boldsymbol{y}_p^\top \boldsymbol{y}_p} = \underset{1 \leqslant q \leqslant Q}{\mathrm{med}} \left\{N \times \widehat{s}_q^2 \times \mathrm{SE}(\widehat{\beta}_{qp})^2 + \widehat{s}_q^2 \times \widehat{\beta}_{qp}^2\right\}$.
- Finally, $\frac{1}{N}\boldsymbol{y}_k^\top \boldsymbol{y}_i$ is estimated using the null SNPs from GWAS summary statistics, i.e., SNPs not marginally associated with $\boldsymbol{y}_k$ or $\boldsymbol{y}_i$. Following Kim et al. (2015), $\mathrm{Corr}(\boldsymbol{y}_k, \boldsymbol{y}_i) \approx \mathrm{Corr}(\boldsymbol{z}_k, \boldsymbol{z}_i)$ where $\boldsymbol{z}_k$ and $\boldsymbol{z}_i$ are vectors of z-scores for the null SNPs. Thus, we can rearrange the sample correlation formula for (centered) phenotype variables and plug in our approximation to obtain $\frac{1}{N}\widehat{\boldsymbol{y}_k^\top \boldsymbol{y}_i} = \frac{1}{N}\,\mathrm{Corr}(\boldsymbol{z}_k, \boldsymbol{z}_i)\sqrt{\widehat{\boldsymbol{y}_k^\top \boldsymbol{y}_k}\,\widehat{\boldsymbol{y}_i^\top \boldsymbol{y}_i}}$. In practice, the SNPs with p-values larger than 0.05 are considered as null SNPs. An alternative method to consider for estimating $\frac{1}{N}\boldsymbol{y}_k^\top \boldsymbol{y}_i$ is also feasible to use for GWAS (Bulik-Sullivan et al., 2015), although not used herein.

Next we extend the framework of interventional Gaussian DAG to leverage large-scale GWAS summary statistics.

2.2 *Method for network construction*

The estimation of interventional Gaussian DAG consists of three steps.

(E1) First, we use penalized regressions to estimate the genotype-phenotype association matrix $\mathbf{V} := \mathbf{W}(\mathbf{I} - \mathbf{U})^{-1}$ in the following equation

$$\boldsymbol{Y} = \boldsymbol{X}\mathbf{V} + \boldsymbol{\epsilon}(\mathbf{I} - \mathbf{U})^{-1}. \tag{2}$$

(E2) Next, we employ the peeling algorithm (Li et al., 2023) to learn a super-DAG, i.e., a directed super-graph without cycles, based on $\widehat{\mathbf{V}}$ obtained in Step (E1).

(E3) Finally, we estimate $\mathbf{U}$ and $\mathbf{W}$ through penalized regressions based on the estimated super-DAG in Step (E2).

Now, we elaborate on our extensions to accommodate summary statistics.

2.2.1 *Estimation of* $\mathbf{V}$ *by truncated Lasso penalized regressions.* In Equation (2), the matrix $\mathbf{V}$ can be estimated column-wise from

$$\boldsymbol{y}_p = \boldsymbol{X}\mathbf{v}_p + \boldsymbol{\xi}_p, \quad \boldsymbol{\xi}_p \sim N(\mathbf{0}, \sigma_p^2\mathbf{I}), \tag{3}$$

where $\boldsymbol{y}_p$ is the data vector of phenotype p, $\boldsymbol{X}$ is the data matrix of genotypes, and vector $\mathbf{v}_p = (\mathrm{v}_{1p}, \ldots, \mathrm{v}_{Qp})^\top$ is the pth column of $\mathbf{V}$. Given the summary statistics, we expand the squared error function $\|\boldsymbol{y}_p - \boldsymbol{X}\mathbf{v}_p\|^2 = \boldsymbol{y}_p^\top\boldsymbol{y}_p + \mathbf{v}_p^\top\boldsymbol{X}^\top\boldsymbol{X}\mathbf{v}_p - 2\boldsymbol{y}_p^\top\boldsymbol{X}\mathbf{v}_p$, and replace the quantities $\boldsymbol{y}_p^\top\boldsymbol{y}_p$, $\boldsymbol{X}^\top\boldsymbol{y}_p$, and $\boldsymbol{X}^\top\boldsymbol{X}$ in Equation (3) with their estimates $\widehat{\boldsymbol{y}_p^\top\boldsymbol{y}_p}$, $\widehat{\boldsymbol{X}^\top\boldsymbol{y}_p}$, and $\widehat{\boldsymbol{X}^\top\boldsymbol{X}}$, respectively. As a result, we estimate $\mathbf{v}_p$ through regressions with the Truncated Lasso Penalty (TLP) (Shen et al., 2012) to minimize

$$\mathbf{v}_p^\top\widehat{\boldsymbol{X}^\top\boldsymbol{X}}\mathbf{v}_p - 2\mathbf{v}_p^\top\widehat{\boldsymbol{X}^\top\boldsymbol{y}_p} \quad \text{s.t.} \quad J(\mathbf{v}_p, \tau_p) \leqslant \kappa_p, \tag{4}$$

where $\kappa_p > 0$ is an integer tuning parameter and $J(\mathbf{v}_p, \tau_p) = \sum_{q=1}^{Q}\min(|\mathrm{v}_{qp}|/\tau_p, 1)$ is the TLP function, which does not penalize the parameters over the threshold τ_p. We use the R package "glmtlp" (Li et al., 2022) to fit the summary-level data regression (4).

For implementation, we fix $\tau_p = 0.5\sqrt{\log(Q)/N}$ and choose $\kappa_p \in \{1, \ldots, Q\}$ individually for each of the P penalized regressions by minimizing the pseudo-BIC (Pattee and Pan, 2020), which is defined as $\mathrm{pBIC}(\lambda_p, \tau_p) = \widehat{\mathrm{SSE}}_p(\widehat{\mathbf{v}}_p)/\widehat{\sigma}_p^2 + \|\widehat{\mathbf{v}}_p\|_0 \times \log(N)$, where $\widehat{\mathrm{SSE}}_p(\widehat{\mathbf{v}}_p) =$

$\widehat{\boldsymbol{y}_p^\top \boldsymbol{y}_p} - 2\widehat{\mathbf{v}}_p^\top \widehat{\boldsymbol{X}^\top \boldsymbol{y}_p} + \widehat{\mathbf{v}}_p^\top \widehat{\boldsymbol{X}^\top \boldsymbol{X}} \widehat{\mathbf{v}}_p$ is the (estimated) sum of squared error of $\widehat{\mathbf{v}}_p$, $\|\widehat{\mathbf{v}}_p\|_0$ is the number of nonzero coefficients in $\widehat{\mathbf{v}}_p$, $\widehat{\mathbf{v}}_p$ is the estimate in (4) with tuning parameters (λ_p, τ_p), N is the sample size (when N differs, the median is taken), and $\widehat{\sigma}_p^2$ is the estimated residual variance for phenotype p in Equation (3). When Q is small compared to N as in our application, a consistent estimate for σ_p^2 can be obtained from the ordinary least squares using all Q genotypes, $\widehat{\sigma}_p^2 = (\widehat{\boldsymbol{y}_p^\top \boldsymbol{y}_p} - 2\widehat{\mathbf{v}}_{p,olse}^\top \widehat{\boldsymbol{X}^\top \boldsymbol{y}_p} + \widehat{\mathbf{v}}_{p,olse}^\top \widehat{\boldsymbol{X}^\top \boldsymbol{X}} \widehat{\mathbf{v}}_{p,olse})/(N - Q)$, where $\widehat{\mathbf{v}}_{p,olse}$ is the estimate in Equation (4) with $\kappa_p = P$. Letting $\widehat{\mathbf{v}}_p$ $(1 \leqslant p \leqslant P)$ be the estimates with the optimally chosen tuning parameters, the final estimate of $\mathbf{V}$ is $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_P)$.

2.2.2 *Estimation of super-DAG by the peeling algorithm.* Given $\widehat{\mathbf{V}}$, the peeling algorithm (Li et al., 2023) can be used to construct a super-DAG with phenotype edge set $\mathcal{E}^+$ (a superset of $\mathcal{E}$) and interventional edge set $\mathcal{I}^+$ (a superset of $\mathcal{I}$). The key idea is that the sparse pattern of matrix $\mathbf{V}$ characterizes the orientations of the relations among the phenotypes. Specifically, it is demonstrated in Li et al. (2023) that $\mathrm{v}_{qp} \neq 0$ implies that genotype q intervenes on phenotype p or an ancestor node of phenotype p in the DAG. Thus, if $\mathrm{v}_{qp} \neq 0$ and $\mathrm{v}_{qi} = 0$ for $i \neq p$, then phenotype p is a leaf node in the DAG, that is, there is no directed edge from phenotype p to the others. On this basis, we can sequentially identify and remove (i.e., peel) the leaf node in the DAG, and construct supersets $\mathcal{E}^+$ and $\mathcal{I}^+$.

Since the peeling algorithm solely depends on $\widehat{\mathbf{V}}$, no modification is needed to extend the existing method to accommodate summary-level data.

2.2.3 *Estimation of* $\mathbf{U}$ *and* $\mathbf{W}$. The peeling algorithm yields supersets $\mathcal{E}^+ \supseteq \mathcal{E}$ and $\mathcal{I}^+ \supseteq \mathcal{I}$. To remove the extra edges in $\mathcal{E}^+$ and $\mathcal{I}^+$, we consider fitting $\mathbf{U}$ and $\mathbf{W}$ within a restricted model defined by $\mathcal{E}^+$ and $\mathcal{I}^+$.

From Equation (1), for phenotype p, we have

$$\boldsymbol{y}_p = \sum_{k \in \mathcal{E}^+(p)} \boldsymbol{y}_k \mathrm{u}_{kp} + \sum_{q \in \mathcal{I}^+(p)} \boldsymbol{x}_q \mathrm{w}_{qp} + \boldsymbol{\epsilon}_p, \quad \boldsymbol{\epsilon}_p \sim N(\mathbf{0}, \omega_p^2 \mathbf{I}), \tag{5}$$

where $\mathcal{E}^+(p) = \{k : (k,p) \in \mathcal{E}^+\}$ and $\mathcal{I}^+(p) = \{q : (q,p) \in \mathcal{I}^+\}$. As in Section 2.2.1, we replace the corresponding quantities with the summary-level data estimates and fit the TLP regression based on Equation (5),

$$\min_{\boldsymbol{\gamma}} \ \boldsymbol{\gamma}^\top \begin{Bmatrix} (\widehat{\boldsymbol{Y}^\top \boldsymbol{Y}})_{\mathcal{E}^+(p),\mathcal{E}^+(p)} & (\widehat{\boldsymbol{Y}^\top \boldsymbol{X}})_{\mathcal{E}^+(p),\mathcal{I}^+(p)} \\ (\widehat{\boldsymbol{X}^\top \boldsymbol{Y}})_{\mathcal{I}^+(p),\mathcal{E}^+(p)} & (\widehat{\boldsymbol{X}^\top \boldsymbol{X}})_{\mathcal{I}^+(p),\mathcal{I}^+(p)} \end{Bmatrix} \boldsymbol{\gamma} - 2\boldsymbol{\gamma}^\top \begin{Bmatrix} (\widehat{\boldsymbol{Y}^\top \boldsymbol{Y}})_{\mathcal{E}^+(p),p} \\ (\widehat{\boldsymbol{X}^\top \boldsymbol{Y}})_{\mathcal{I}^+(p),p} \end{Bmatrix} \quad \text{s.t.} \quad J(\boldsymbol{\gamma}, \tau_p) \leqslant \kappa_p, \tag{6}$$

where $\boldsymbol{\gamma} = \{(\mathrm{u}_{kp})_{k\in\mathcal{E}^+(p)}, (\mathrm{w}_{kp})_{k\in\mathcal{I}^+(p)}\}$ is the parameter vector and $J(\boldsymbol{\gamma}, \tau_p) = \sum_k \min(|\gamma_k|/\tau_p, 1)$. We fix $\tau_p = 0.5\sqrt{\log\{|\mathcal{E}^+(p)| + |\mathcal{I}^+(p)|\}/N}$ and the tuning parameter $\kappa_p \in \{1, \ldots, |\mathcal{E}^+(p)| + |\mathcal{I}^+(p)|\}$ are selected by pseudo-BIC as described in Section 2.2.1. The estimated $\widehat{\mathbf{u}}_p = \{(\widehat{\mathrm{u}}_{kp})_{k\in\mathcal{E}^+(p)}, \mathbf{0}\}$ and $\widehat{\mathbf{w}}_p = \{(\widehat{\mathrm{w}}_{qp})_{q\in\mathcal{I}^+(p)}, \mathbf{0}\}$ $(1 \leqslant p \leqslant P)$ are aggregated to form the the final estimate $\widehat{\mathbf{U}} = (\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_P)$ and $\widehat{\mathbf{W}} = (\widehat{\mathbf{w}}_1, \ldots, \widehat{\mathbf{w}}_P)$.

Due to penalization, we recommend following the common practice to standardize the variables so that the phenotypes and genotypes are on a comparable scale, which is straightforward to do as $\widehat{\boldsymbol{y}_p^\top \boldsymbol{y}_p}$ and $\widehat{\boldsymbol{X}^\top \boldsymbol{X}}$ are obtained. Moreover, if only $\mathbf{U}$ is of interest, penalization of $\mathbf{W}$ is optional.

2.3 *Likelihood-based inference for a DAG*

We extend the likelihood ratio inference (Li et al., 2023) to quantify the uncertainty of the network structures. As in Li et al. (2023), we consider two types of hypothesis testing.

- **Testing of multiple directed relations.** The null hypothesis $H_0 : \mathrm{u}_{kj} = 0$ for each $(k,j) \in \mathcal{H}$ and alternative hypothesis $H_a : \mathrm{u}_{kj} \neq 0$ for some $(k,j) \in \mathcal{H}$. Rejecting H_0 indicates evidence for the presence of some hypothesized relationships in the network.
- **Testing of a directed pathway.** The null hypothesis $H_0 : \mathrm{u}_{kj} = 0$ for some $(k,j) \in \mathcal{H}$ and alternative hypothesis $H_a : \mathrm{u}_{kj} \neq 0$ for each $(k,j) \in \mathcal{H}$. Rejecting H_0 indicates evidence for the presence of the entire directed pathway in the phenotype network.

The procedure for testing multiple directed relations comprises five steps.

(T1) Estimate $\mathbf{V}$ and use the peeling algorithm to obtain $\mathcal{E}^+$ and $\mathcal{I}^+$ as in Section 2.2.

(T2) Identify the set of nondegenerate edges $\mathcal{D}$ (Li et al., 2023), which contains $\mathcal{D}_p$, the nondegenerate edges pointed to phenotype p.

(T3) Estimate the parameters $\mathbf{U}$ and $\mathbf{W}$ under H_0 and H_a, respectively. Specifically, denote by $\widehat{\mathbf{U}}^{(0)}$ and $\widehat{\mathbf{W}}^{(0)}$ the estimates under H_0. Then $\widehat{\mathbf{U}}^{(0)}$ and $\widehat{\mathbf{W}}^{(0)}$ are computed as in the regression (6) with an additional constraint that $\mathrm{u}_{kj} = 0$ for $(k, j) \in \mathcal{H}$. Let $\widehat{\mathbf{U}}^{(1)}$ and $\widehat{\mathbf{W}}^{(1)}$ be the estimates under H_a. Then $\widehat{\mathbf{U}}^{(1)}$ and $\widehat{\mathbf{W}}^{(1)}$ are computed from the restricted models ($1 \leqslant p \leqslant P$), $\boldsymbol{y}_p = \sum_{k\in\mathcal{E}^+(p)\cup\mathcal{D}_p} \boldsymbol{y}_k \mathrm{u}_{kp} + \sum_{q\in\mathcal{I}^+(p)} \boldsymbol{x}_q \mathrm{w}_{qp} + \boldsymbol{\epsilon}_p$, $\boldsymbol{\epsilon}_p \sim N(\mathbf{0}, \omega_p^2\mathbf{I})$, via regression (6), where the penalties become $J(\mathbf{u}_p, \tau_p) = \sum_{k\in\mathcal{E}^+(p)\setminus\mathcal{D}_p} \min(|\mathrm{u}_{kp}|, \tau_p)$ and $J(\mathbf{w}_p, \tau_p) = \sum_{q\in\mathcal{I}^+(p)} \min(|\mathrm{w}_{qp}|, \tau_p)$.

(T4) Compute $\widehat{\omega}_p^2$ ($1 \leqslant p \leqslant P$) from the residual sum of squares of $(\widehat{\mathbf{u}}_p^{(1)}, \widehat{\mathbf{w}}_p^{(1)})$.

(T5) Compute the test statistic $T = 2L\left\{\widehat{\mathbf{U}}^{(1)}, \widehat{\mathbf{W}}^{(1)}, \widehat{\omega}_1^2, \ldots, \widehat{\omega}_P^2\right\} - 2L\left\{\widehat{\mathbf{U}}^{(0)}, \widehat{\mathbf{W}}^{(0)}, \widehat{\omega}_1^2, \ldots, \widehat{\omega}_P^2\right\}$, where L is the log-likelihood of the model (Equation (1)). By Li et al. (2023), T is approximately chi-squared distributed with degrees of freedom $|\mathcal{D}|$ when the size $|\mathcal{D}|$ is less than 50; $(T - |\mathcal{D}|)/\sqrt{2|\mathcal{D}|}$ is approximately standard normal when $|\mathcal{D}|$ exceeds 50. Thus, the p-value is calculated as $\mathrm{PV} = \mathbb{P}_{Z\sim\chi^2_{|\mathcal{D}|}}(Z > T)$ when $|\mathcal{D}| < 50$ and $\mathrm{PV} = \mathbb{P}_{Z\sim N(0,1)}\{Z > (T - |\mathcal{D}|)/\sqrt{2|\mathcal{D}|}\}$ when $|\mathcal{D}| \geqslant 50$.

The procedure for testing a directed pathway is similar, with minor modifications.

(P1) Estimate $\mathbf{V}$ as in Step (T1).

(P2) First, we decompose H_0 into each nongenerate edge $\left\{H_0^{(k,j)} : \mathrm{u}_{kj} = 0\right\}_{(k,j)\in\mathcal{H}}$. For each $H_0^{(k,j)}$, implement Steps (T2)–(T5) above to obtain the corresponding p-value $\mathrm{PV}_{(k,j)}$. The final p-value is computed as the maximum of the p-values for the sub-hypotheses, $\mathrm{PV} = \max\left\{\mathrm{PV}_{(k,j)} : (k, j) \in \mathcal{H}\right\}$.

Of note, testing a directed pathway concerns a composite (null) hypothesis. Fixing $0 < \alpha < 1$, we have $\lim_{n\to\infty} \sup\left\{\mathbb{P}_{\boldsymbol{\theta}}(\mathrm{PV} \leqslant \alpha) : \boldsymbol{\theta} = (\mathbf{U}, \mathbf{W}) \text{ satisfies } H_0\right\} = \alpha$ (Li et al., 2023). In other

words, the test asymptotically achieves exactly the α significance level for the composite null hypothesis.

3. Inferring cardiovascular-related protein-protein interaction network

The role of cardiovascular diseases has been recognized as an important etiologic hallmark of AD (de Bruijn and Ikram, 2014). There are different hypotheses on the various mechanisms underlying the association between AD and cardiovascular diseases (Tini et al., 2020). In this real data application, we constructed a directed PPI network of some cardiovascular-related proteins based on a GWAS of 83 plasma protein biomarkers. We further connected the PPI network to AD through MR analyses.

3.1 *GWAS summary statistics for cardiovascular-related proteins*

The GWAS summary statistics on 83 cardiovascular-related proteins which came from Wald tests for the association between each SNP and the standardized residuals among 3394 European individuals by Folkersen et al. (2017) were used. Five proteins were excluded from the analysis as their corresponding protein-encoding genes are located on the sex chromosome. The summary statistics were first processed to remove (a) indels, (b) SNPs located within one base pair of an indel, (c) SNPs with imputation quality score INFO $\leqslant$ 0.8, and (d) SNPs with MAF $\leqslant$ 0.05. We then used the following steps to select putative IVs for the proteins:

- SNPs were clumped at an r^2 value of 0.01 using 3000 uncorrelated individuals (individuals with kinship coefficients less than 0.084) from UK Biobank of European ancestry as the reference panel such that SNPs were independent of each other for each protein (Bycroft et al., 2018);
- Only the SNPs in the clumped data files located within $\pm$1MB of each protein-encoding genes were considered. In general, cis-regulatory changes will be less pleiotropic (Signor

and Nuzhdin, 2018) and thus these SNPs located close to the genes are more likely to be valid IVs due to the exclusion assumption (Swerdlow et al., 2016; Hemani et al., 2018; Li et al., 2023) (i.e., an IV only directly intervenes on one primary variable);

- To ensure the relevance assumption (Li et al., 2023) was satisfied (i.e., IV intervenes on at least one primary variable), we only selected SNPs whose p-values were below the GWAS significance threshold (5×10^{-8}). This filtering process led to a total number of 33 SNPs and 23 proteins with at least one putative IV in the final network analysis.

The genetic correlation matrix for the included IVs was estimated based on the same reference panel used in clumping. We calculated the empirical correlation of each pair of proteins as the correlation coefficient of the z-scores of the null SNPs, i.e., all autosomal SNPs with MAF $\geqslant 0.05$, INFO $\geqslant 0.8$, and GWAS p-values $\geqslant 0.05$ for both proteins. The number of null SNPs for each pair of proteins ranged from 1,191,204 to 1,223,357. All preparation of the reference panel and GWAS data for both the DAG estimation and MR analysis was done using PLINK version 1.9 (Purcell et al., 2007).

3.2 *GWAS summary statistics for AD*

We explored the relationship between each of the 23 proteins in Folkersen et al. (2017) and AD. We used the summary statistics of the GWAS for AD from a most recent study totaling 111,326 clinically diagnosed or "proxy" AD cases and 677,663 controls (Bellenguez et al., 2022). We removed SNPs with MAF < 0.05, SNPs not included in the GWAS of Folkersen et al. (2017), and clumped SNPs at $r^2 = 0.01$. Among the remaining SNPs, we selected IVs only with GWAS p-value $< 5 \times 10^{-8}$ in the MR analyses.

3.3 *Results*

We constructed a DAG of the 23 proteins as described in Section 2.5.1. As Folkersen et al. (2017) shared MAF in the summary statistics, we compared them with those in UK Biobank,

the reference panel for clumping and estimating genetic correlation matrix. The absolute difference of the MAF of all IVs ranged from 0.001 to 0.055 with a mean 0.02, while the correlation was 0.99 (Table S1). We further performed MR analysis on each protein to evaluate their relationship with AD using the TwoSampleMR package (Hemani et al., 2018). We used Egger's test of intercept for examining the exclusion assumption: If a protein had a p-value of the Egger's test of intercept $> 0.05/23$, there was no evidence against no direct/pleiotropic effects, and we'd go with the more powerful MR-IVW method; otherwise, we used MR-Egger (to allow pleiotropic effects of IVs). In any case, we used the p-value cut-off $< 0.05/23$ to declare statistical significance. Table S2 contains a complete list of MR results. The protein IL18, which showed marginal significance in both Egger's test of intercept (p-value = 0.06) and MR-Egger (p-value = 0.07), was a parent node for several proteins related to AD, including ADM, IL1RL1, CTSD, CXCL6, and CXCL16. We further performed the likelihood ratio test on each edge. Edges with p-value $< 0.05/(23 \times 22 - 56)$ were considered as significant and were in solid line in Figure 2. The number of tests in the Bonferonni correction is bounded by the sum of possible edges among all the nodes minus the total number of edges after the peeling algorithm, which is justified in Supplementary Material S2.1. Each edge from IL18 to the five AD-associated proteins was highly significant in the likelihood ratio test, thus suggesting that simultaneous testing of the pathway from IL18 to the five genes would be significant. Previous studies detected increased levels of pro-inflammatory IL18 in both cardiovascular diseases and in brain regions of AD patients (Sutinen et al., 2014). IL18 is known to increase the level of Cdk5 and GSK-3β, which are involved in Tau hyperphosphorylation, and the inhibition of Cdk5 was known to improve AD subject's condition (Calabrò et al., 2021). Our work suggests a possible regulatory role of IL18 on multiple AD-associated proteins. According to OpenTargets.org (Ochoa, 2021) for current pharmaceuticals either approved or in development with IL18, this protein is

currently a target of an antibody drug to treat diabetes mellitus and a few other conditions; diabetes has long been linked to AD with epidemiological and biological evidence (Barbagallo and Dominguez, 2014). Lastly, we provide a Shiny application that allows users to test any selected proteins in this cardiovascular-related PPI network.

4. Simulation studies

4.1 *Simulation settings*

We simulated the data assuming a fixed $\mathbf{U}$, $\mathbf{W}$, standardized genotype matrix $\boldsymbol{X}$, and sampled each row of $\boldsymbol{\epsilon}$ independently from $N\{\mathbf{0}, \mathrm{Diag}(\omega_1^2, \ldots, \omega_P^2)\}$. Then we generated $\boldsymbol{Y}$ from equation: $\boldsymbol{Y} = \mathbf{V}^\top \boldsymbol{X} + \boldsymbol{\epsilon}(\mathbf{I} - \mathbf{U})^{-1}$. Without loss of generality, no intercept was modeled, i.e., $\boldsymbol{Y}$ was centered at mean 0. The values of $\mathbf{U}$, $\omega_1^2, \ldots, \omega_P^2$, and $\mathbf{W}$ are provided in Supplemental Material S1.1, where the structure of the relationship of $\boldsymbol{Y}$ followed a DAG of 15 nodes/phenotypes (Figure 1). The effect sizes of the non-zero components of $\mathbf{U}$ ranged from 0.002 to 1.16 with a median of 0.06. All phenotypes had at least one valid IV. Twenty-six SNPs were included in the model, with their effect sizes ranging from -2.2 to 2.5 with a median of -0.11. Two SNPs violated the relevance assumption, while the rest were valid IVs. We also varied the effect sizes to be 1/3 and 1/15 of $\mathbf{U}$ while keeping $\mathbf{W}$ fixed.

The standardized genotype matrix $\boldsymbol{X}$ was obtained from unrelated individuals of European ancestry in UK Biobank (Bycroft et al., 2018). We then calculated the summary statistics using a linear model of each phenotype on each standardized genotype, and inputted the summary statistics into the proposed algorithm. The reference panel was obtained from the UK Biobank European samples, which were not correlated with the simulated samples used to derive the summary statistics. SNPs on chromosome 22 with a Hardy-Weinberg Disequilibrium test p-value > 0.0001, missing call rate < 0.05, and MAF > 0.05 were pruned to have $r^2 < 0.01$. We then randomly selected 26 SNPs for $\boldsymbol{X}$. Missing values of SNPs were

imputed by their mean. Null SNPs were directly simulated to be independent of each other and had no relationship with $\boldsymbol{Y}$.

[Figure 1 about here.]

We evaluated the performance of both the network construction and statistical inference for the proposed method. To evaluate the performance of network construction, we examined the false positive (TP) and false negative (FN) rates for $\widehat{\mathbf{U}}$ over 200 replications. The sample size of the summary statistics was varied at 3000, 6000, 9000, and 12,000, and the sample size of the reference panel was fixed at 3000. Null SNPs were simulated to have the same sample size as the GWAS summary statistics.

In terms of testing, we examined the empirical Type I error/power of the likelihood ratio tests with increasing sample sizes and varying strength of $\mathbf{U}$ for the following five scenarios over 1000 replications for the two types of testing.

I. Testing one or more directed relations:

- A1. Type 1 error: testing one edge when in truth it was null with $H_0 : \mathrm{u}_{1,14} = 0$ vs $H_1 : \mathrm{u}_{1,14} \neq 0$;
- A2. Power: testing one edge when in truth it was not null with $H_0 : \mathrm{u}_{1,6} = 0$ vs $H_1 : \mathrm{u}_{1,6} \neq 0$;
- A3. Power: testing two edges together when in truth both were not null with $H_0 : \mathrm{u}_{7,15} = \mathrm{u}_{1,6} = 0$ vs $H_1 : \mathrm{u}_{7,15} \neq 0$ or $\mathrm{u}_{1,6} \neq 0$;

II. Testing of a directed pathway:

- B1. Type 1 error: testing whether at least one edge was not null when in truth only one was not null with $H_0 : \mathrm{u}_{1,14} = 0$ or $\mathrm{u}_{6,12} = 0$ vs $H_1 : \mathrm{u}_{1,14} \neq 0$ and $\mathrm{u}_{6,12} \neq 0$;
- B2. Power: testing whether at least one edge was not null when in truth both were not null with $H_0 : \mathrm{u}_{1,6} = 0$ or $\mathrm{u}_{6,12} = 0$ vs $H_1 : \mathrm{u}_{1,6} \neq 0$ and $\mathrm{u}_{6,12} \neq 0$;

The true strengths of the tested edges were: $u_{1,14} = 0$, $u_{1,6} = 0.27$, $u_{7,15} = 0.44$, and $u_{6,12} = 1.36$.

[Figure 2 about here.]

4.2 *Simulation results*

With the increase of sample size, the constructed networks became closer to the true graph (Figure 3, Numeric values in Figure 3. A are in Table S3 - S6). More specifically, FP of $\widehat{\mathbf{U}}$ was around 0.05, and FN of $\widehat{\mathbf{U}}$ decreased with the increase of sample size when 15000 null SNPs were used to estimate the 15×15 matrix of $\boldsymbol{Y}^\top \boldsymbol{Y}$. Unsurprisingly, the performance of using $\widehat{\boldsymbol{Y}^\top \boldsymbol{Y}}$ estimated by 15,000 null SNPs (denoted as $\mathbf{U}'$ in Figure 3) was slightly worse than that of using $\boldsymbol{Y}^\top \boldsymbol{Y}$ (denoted as $\mathbf{U}$ in Figure 3). In addition, the FP of $\widehat{\mathbf{U}}$ clearly decreased with the increase of effect sizes (from $\mathbf{U}/15$ to $\mathbf{U}/3$ to $\mathbf{U}$). To check the validity of our method, we compared pBIC estimated from summary statistics with BIC estimated from the individual-level data for the same set of penalized regression coefficients. We found the two sets of values were highly concordant, and the results from one iteration were plotted in Figure S2.

In terms of testing, we observed well-controlled Type I error rates for scenarios A1 and B1 (Numeric values present in Figure 3, B-D are in Table S7 - S10), using $\boldsymbol{Y}^\top \boldsymbol{Y}$. We note that the empirical Type I error rates might become conservative when $\boldsymbol{Y}^\top \boldsymbol{Y}$ is replaced by its estimate $\widehat{\boldsymbol{Y}^\top \boldsymbol{Y}}$ derived from 15000 null SNPs (Table S10). In real data analysis of GWAS summary level data, typically a much larger number of null SNPs and various other strategies can be used to better estimate $\boldsymbol{Y}^\top \boldsymbol{Y}$ (Bulik-Sullivan et al., 2015; Kim et al., 2015; Li et al., 2021); however, the investigation in this direction is out of our scope. Furthermore, empirical power was high for scenarios A2, A3, and B2 and increased with sample size and effect size. The empirically power of jointly testing two edges $u_{1,6}$ and $u_{7,15}$ (scenario A3) was larger than testing one edge $u_{1,6}$ alone (scenario A2).

[Figure 3 about here.]

5. Discussion

In this paper, we present a method to estimate an interventional DAG of phenotypes utilizing linear structural equation models, applicable to GWAS summary statistics in the absence of individual-level data. We demonstrated satisfactory performance in terms of the FP and FN rates in network construction and high empirical power and of well-controlled Type I error rates of the likelihood ratio tests. We applied this method to a large-scale proteomic GWAS summary dataset to obtain an estimated DAG of 23 cardiovascular-related proteins and further illustrated the effects of these proteins on AD by MR analysis. These results can be useful in understanding the disease etiology, drug repurposing, and other applications for AD.

We note that the choice of a proper reference panel is just as important for our method as many other summary-statistics-based methods (Deng and Pan, 2018; Chen et al., 2021; Privé et al., 2022). When constructing the cardiovascular-related PPI network, we used an ancestry-matched reference panel of uncorrelated individuals from UK Biobank with a sample size of 3000, which is close to the sample size of the GWAS of cardiovascular proteins. Furthermore, we clumped SNPs around the cis-region of the gene and only used the genome-wide significant SNPs as IVs. This step not only aimed at selecting at least one valid IV for each protein, but also achieving better estimation of the genetic correlation matrix for the IVs as the number of total IVs became much smaller than the sample size of the reference panel, i.e. $Q \ll N_r$. Our analysis was constrained to a single ancestry group and unrelated samples. As the collection of multi-ancestry and related samples increases, it will be of significant research interest to establish networks among these populations, a challenge we anticipate addressing in our future work.

In recent years, a large amount of summary-level data has become widely accessible. On

the molecular level, many studies published their summary statistics for SNP-molecular phenotype associations. For example, variant-gene associations in 49 tissues can be directly downloaded from GTExPortal (Consortium, 2020). Beyond the molecular phenotypes, UK Biobank alone provides GWAS summary statistics on more than 7000 traits, including but not limited to cognitive functions, early life factors, health and medical history, and physical measurement (Bycroft et al., 2018). Our proposed method provides a computational and analytical tool to explore the relationships among multiple phenotype variables by taking advantage of rapid advances in GWAS and other association mappings.

Acknowledgement

R. Zilinskas and C. Li have contributed equally to this work. The authors would like to thank the associate editor and the reviewer for their valuable comments. This research was supported by NIH grants U01 AG073079, R01 AG065636, R01 AG069895, RF1 AG067924, R01 HL116720, and R01 GM126002 and by the Minnesota Supercomputing Institute at the University of Minnesota. T. Yang would like to further acknowledge the Children's Cancer Research Funds and the St. Baldrick's Foundation Scholar Award.

Supplementary Materials

Web Appendices, Supplementary Tables S1–S10, and Figure S1 referenced in Sections 3.3 and 4.2 are available with this paper in the Supplementary Materials. Data and code (sumDAG algorithm, Shiny app, and the real data analysis code) can also be found on `https://github.com/chunlinli/sumdag`.

Data availability statement

We downloaded the summary level GWAS data in Section 3.1 from `https://zenodo.org/record/264128/`. The algorithm for the proposed work is packaged in R, available at `https:`

`//github.com/chunlinli/sumdag`, along with code used for the simulation studies and real data application. The processed summary-level GWAS data that were used as input for the algorithm for the real data application are also included on GitHub.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). The gene ontology consortium gene ontology: tool for the unification of biology. *Nature Genetics* **25,** 25–29.

Barbagallo, M. and Dominguez, L. J. (2014). Type 2 diabetes mellitus and Alzheimer's disease. *World J Diabetes* **5,** 889–893.

Bellenguez, C., Küçükali, F., Jansen, I. E., Kleineidam, L., Moreno-Grau, S., Amin, N., Naj, A. C., Campos-Martin, R., Grenier-Boley, B., Andrade, V., et al. (2022). New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics* **54,** 412–436.

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R., Patterson, N., Robinson, E. B., et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47,** 1236–1241.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209.

Calabrò, M., Rinaldi, C., Santoro, G., and Crisafulli, C. (2021). The biological pathways of Alzheimer disease: A review. *AIMS Neuroscience* **8,** 86–132.

Chen, C., Ren, M., Zhang, M., and Zhang, D. (2018). A two-stage penalized least squares method for constructing large systems of structural equations. *Journal of Machine Learning Research* **19,** 1–34.

Chen, W., Wu, Y., Zheng, Z., Qi, T., Visscher, P. M., Zhu, Z., and Yang, J. (2021). Improved

analyses of gwas summary statistics by reducing data heterogeneity and errors. *Nature Communications* **12,** 7117.

Cheng, F., Zhao, J., Wang, Y., Lu, W., Liu, Z., Zhou, Y., Martin, W. R., Wang, R., Huang, J., Hao, T., Yue, H., Ma, J., Hou, Y., Castrillon, J. A., Fang, J., Lathia, J. D., Keri, R. A., Lightstone, F. C., Antman, E. M., Rabadan, R., Hill, D. E., Eng, C., Vidal, M., and Loscalzo, J. (2021). Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nature Genetics* **53,** 342–353.

Consortium, G. (2020). The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* **369,** 1318–1330.

de Bruijn, R. F. and Ikram, M. A. (2014). Cardiovascular risk factors and future risk of alzheimer's disease. *BMC Medicine* **12,** 1–9.

Deng, Y. and Pan, W. (2018). Improved use of small reference panels for conditional and joint analysis with gwas summary statistics. *Genetics* **209,** 401–408.

Emilsson, V., Ilkov, M., Lamb, J. R., Finkel, N., Gudmundsson, E. F., Pitts, R., Hoover, H., Gudmundsdottir, V., Horman, S. R., Aspelund, T., Shu, L., Trifonov, V., Sigurdsson, S., Manolescu, A., Zhu, J., Örn Olafsson, Jakobsdottir, J., Lesley, S. A., To, J., Zhang, J., Harris, T. B., Launer, L. J., Zhang, B., Eiriksdottir, G., Yang, X., Orth, A. P., Jennings, L. L., and Gudnason, V. (2018). Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361,** 769–773.

Folkersen, L., Fauman, E., Sabater-Lleal, M., Strawbridge, R. J., Frånberg, M., Sennblad, B., Baldassarre, D., Veglia, F., Humphries, S. E., Rauramaa, R., de Faire, U., Smit, A. J., Giral, P., Kurl, S., Mannarino, E., Enroth, S., Åsa Johansson, Enroth, S. B., Gustafsson, S., Lind, L., Lindgren, C., Morris, A. P., Giedraitis, V., Silveira, A., Franco-Cereceda, A., Tremoli, E., study group, I., Gyllensten, U., Ingelsson, E., Brunak, S., Eriksson, P., Ziemek, D., Hamsten, A., and Mälarstig, A. (2017). Mapping of 79 loci for 83 plasma

protein biomarkers in cardiovascular disease. *PLOS Genetics* **13,** e1006706.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using bayesian networks to analyze expression data. In *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 127–135.

Hemani, G., Bowden, J., and Smith, G. D. (2018). Evaluating the potential role of pleiotropy in mendelian randomization studies. *Human Molecular Genetics* **27,** 195–208.

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7,** e34408.

Id, J. Q., Chen, K., Zhong, C., Zhu, S., and Id, X. M. (2021). Network-based protein-protein interaction prediction method maps perturbations of cancer interactome. *PLOS Genetics* **17,** e1009869.

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437,** 1299–1320.

Kim, J., Bai, Y., and Pan, W. (2015). An adaptive association test for multiple phenotypes with GWAS summary statistics. *Genetic Epidemiology* **39,** 651–663.

Li, C., Shen, X., and Pan, W. (2023). Inference for a large directed acyclic graph with unspecified interventions. *Journal of Machine Learning Research* **24,** 1–48.

Li, C., Yang, Y., and Wu, C. (2022). Package 'glmtlp'. `https://cran.r-project.org/web/packages/glmtlp/glmtlp.pdf`.

Li, T., Ning, Z., and Shen, X. (2021). Improved estimation of phenotypic correlations using summary association statistics. *Frontiers in Genetics* **12,** 665252.

Liu, F., Zhang, S.-W., Guo, W.-F., Wei, Z.-G., and Chen, L. (2016). Inference of gene regulatory network based on local bayesian networks. *PLoS computational biology* **12,** e1005024.

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology* **41,** 469–480.

Napoli, C., Benincasa, G., Donatelli, F., and Ambrosio, G. (2020). Precision medicine in distinct heart failure phenotypes: Focus on clinical epigenetics. *American Heart Journal* **224,** 113–128.

Ochoa, D. e. a. (2021). Open targets platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Research* **49,** D1302–D1310.

Pattee, J. and Pan, W. (2020). Penalized regression and model selection methods for polygenic scores on summary statistics. *PLOS Computational Biology* **16,** e1008271.

Privé, F., Arbel, J., Aschard, H., and Vilhjálmsson, B. J. (2022). Identifying and correcting for misspecifications in gwas summary statistics and polygenic scores. *Human Genetics and Genomics Advances* **3,**.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81,** 559–575.

Ross, C. A. and Poirier, M. A. (2004). Protein aggregation and neurodegenerative disease. *Nature Medicine* **10,** S10–S17.

Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107,** 223–232.

Signor, S. A. and Nuzhdin, S. V. (2018). The evolution of gene expression in cis and trans. *Trends in Genetics* **34,** 532–544.

Snider, J., Kotlyar, M., Saraon, P., Yao, Z., Jurisica, I., and Stagljar, I. (2015). Fundamentals of protein interaction network mapping. *Molecular Systems Biology* **11,** 848.

Sutinen, E. M., Korolainen, M. A., Häyrinen, J., Alafuzoff, I., Petratos, S., Salminen,

A., Soininen, H., Pirttilä, T., and Ojala, J. O. (2014). Interleukin-18 alters protein expressions of neurodegenerative diseases-linked proteins in human SH-SY5Y neuron-like cells. *Frontiers in Cellular Neuroscience* **8,** 214.

Swerdlow, D., Kuchenbaecker, K., Shah, S., Sofat, R., Holmes, M., White, J., Mindell, J., Kivimaki, M., Brunner, E., Whittaker, J., Casa, J., and Hingorani, A. (2016). Selecting instruments for mendelian randomization in the wake of genome-wide association studies. *International Journal of Epidemiology* **45,** 1600–1616.

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., et al. (2021). Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature* **590,** 290–299.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* **526,** 68–74.

Tini, G., Scagliola, R., Monacelli, F., La Malfa, G., Porto, I., Brunelli, C., and Rosa, G. M. (2020). Alzheimer's disease and cardiovascular disease: a particular association. *Cardiology Research and Practice* **2020,** 2617970.

Witten, D. M., Friedman, J. H., and Simon, N. (2012). New insights and faster computations for the graphical lasso view. *Journal of Computational and Graphical Statistics* pages 892 – 900.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* **4,**.

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

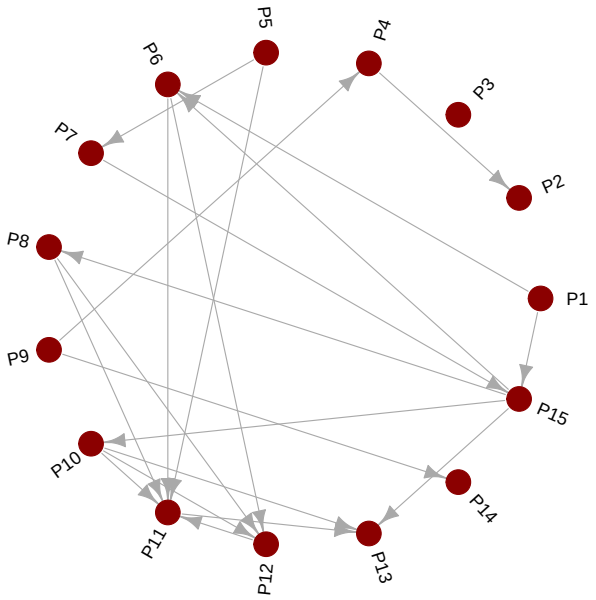

Figure 1. True DAG for the simulation study with 15 phenotypes

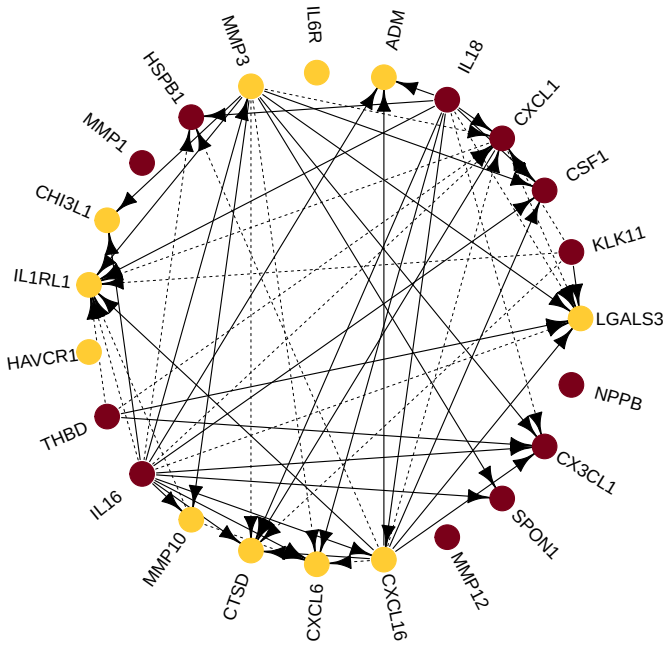

Figure 2. Estimated DAG for 23 proteins based on the GWAS summary statistics of Folkersen et al. (2017); Proteins significantly associated with AD in MR analysis are colored gold. A solid line represents an edge that is statistically significant by the likelihood ratio test whereas a dashed line represents an edge that is not significant.

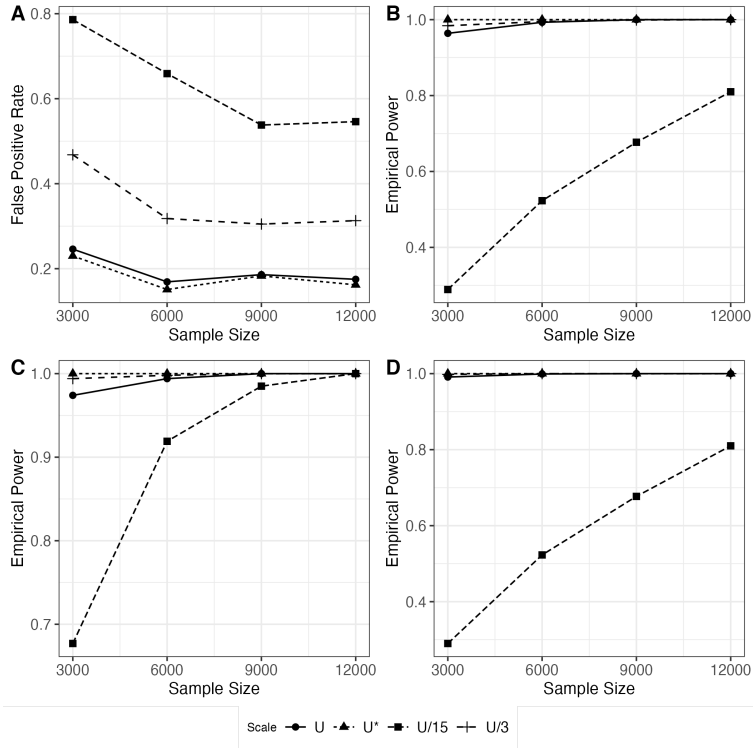

Figure 3. Performance of network construction (A) and likelihood ratio tests (B-D) in simulation with varying sample sizes and true effect sizes at **U**, **U**/3, and **U**/15. Figures B, C, and D represent scenarios A2, A3, and B2 respectively.