frontiers in Genetics

ORIGINAL RESEARCH
published: 27 January 2022
doi: 10.3389/fgene.2021.760299

An Augmented High-Dimensional Graphical Lasso Method to Incorporate Prior Biological Knowledge for Global Network Learning

Yonghua Zhuang[1]***, *Fuyong Xing*[1], *Debashis Ghosh*[1], *Farnoush Banaei-Kashani*[2], *Russell P. Bowler*[3] *and Katerina Kechris*[1]***

[1]Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, United States, [2]Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO, United States, [3]National Jewish Health, Denver, CO, United States

OPEN ACCESS

Edited by:
Xiaoping Liu,
University of Chinese Academy of Science, China

Reviewed by:
Le Li,
Cornell University, United States
Pavel Loskot,
Zhejiang University-University of Illinois at Urbana-Champaign Institute (ZJUI), China

****Correspondence:***
Yonghua Zhuang
Yonghua.Zhuang@cuanschutz.edu
Katerina Kechris
Katerina.Kechris@cuanschutz.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *17 August 2021*
Accepted: *08 November 2021*
Published: *27 January 2022*

Citation:
Zhuang Y, Xing F, Ghosh D, Banaei-Kashani F, Bowler RP and Kechris K (2022) An Augmented High-Dimensional Graphical Lasso Method to Incorporate Prior Biological Knowledge for Global Network Learning.
Front. Genet. 12:760299.
doi: 10.3389/fgene.2021.760299

Biological networks are often inferred through Gaussian graphical models (GGMs) using gene or protein expression data only. GGMs identify conditional dependence by estimating a precision matrix between genes or proteins. However, conventional GGM approaches often ignore prior knowledge about protein-protein interactions (PPI). Recently, several groups have extended GGM to weighted graphical Lasso (wGlasso) and network-based gene set analysis (Netgsa) and have demonstrated the advantages of incorporating PPI information. However, these methods are either computationally intractable for large-scale data, or disregard weights in the PPI networks. To address these shortcomings, we extended the Netgsa approach and developed an augmented high-dimensional graphical Lasso (AhGlasso) method to incorporate edge weights in known PPI with omics data for global network learning. This new method outperforms weighted graphical Lasso-based algorithms with respect to computational time in simulated large-scale data settings while achieving better or comparable prediction accuracy of node connections. The total runtime of AhGlasso is approximately five times faster than weighted Glasso methods when the graph size ranges from 1,000 to 3,000 with a fixed sample size (n = 300). The runtime difference between AhGlasso and weighted Glasso increases when the graph size increases. Using proteomic data from a study on chronic obstructive pulmonary disease, we demonstrate that AhGlasso improves protein network inference compared to the Netgsa approach by incorporating PPI information.

Keywords: graphical Lasso, Gaussian graphical model, protein-protein interaction, gene network, systems biology

1 INTRODUCTION

Networks are a useful framework for representing relationships in biological and disease pathways (Zhang et al., 2014). Understanding complex biological networks including protein-protein interaction (PPI) networks is a fundamental and challenging issue in computational and systems biology (Kuchaiev et al., 2009). Known protein-protein interactions have been collected from numerous sources, including experimental data, computational prediction methods, and public text collections to form a global network. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is one of the most comprehensive protein association databases and it includes direct (physical) and indirect (functional) associations (Szklarczyk et al., 2016). Although STRING is continually being updated, the protein-protein interactions are still not complete and accurate due to potential errors and missing information in current high-throughput assays (Huttlin et al., 2017). In addition, protein interactions and pathways associated with specific diseases in STRING may be limited. To build a more complete and specific protein-protein interaction network related to diseases of interest, we need to reconstruct networks based on study-specific co-expression data, in addition to prior knowledge of protein-protein interactions.

Gene networks are commonly inferred using co-expression data (Saelens et al., 2018). One popular expression-based network reconstruction method is weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008). It was originally designed for microarray expression measurements. More recently, it has been extended for sequencing expression data, as well as other data, such as proteomic and metabolomic (DiLeo et al., 2011; Shirasaki et al., 2012; Zhang et al., 2013; Langfelder et al., 2016). While WGCNA has gained popularity, it was originally designed for a single data type but more recently has been extended for integrating multiple data types (Mamdani et al., 2015), by first constructing relevant homogeneous networks in parallel and then combining the separate networks. However, it is not clear how to best combine networks based on different data types and how to incorporate known pathways or protein/genetic interaction information.

Another popular expression-based network reconstruction approach is modeling gene interactions using a Gaussian graphical model (GGM) (Dobra et al., 2004). Under the assumption of multivariate normality of gene expression data, the GGM uses the inverse of the gene covariance matrix as a measure for gene associations. For many genes, the associations are usually very sparse. One popular method to estimate a sparse network is the graphical Lasso algorithm (Mamdani et al., 2015). Similar to WGCNA, GGMs are widely used in biological applications for network graph construction but often ignore the known protein/genetic interaction. Recently, this approach was extended to incorporate partially known information with a weighted graphical Lasso (Li and Jackson, 2015; Zuo et al., 2017). It has been demonstrated that weighted Glasso significantly improved the prediction accuracy of protein-protein interactions. However, the graphical Lasso can be computationally expensive for a large-scale feature space and therefore limited for global network learning using high-dimensional omics data (Fattahi and Sojoudi, 2019).

In addition to the weighted graphical Lasso, Jing Ma *et al.* developed a network-based gene set analysis (Netgsa) approach for network-based enrichment analysis by incorporating prior pathway information (Ma et al., 2016). The Netgsa approach combines the neighborhood selection technique (Meinshausen and Buhlmann, 2006) with constrained maximum likelihood estimation using the graphical Lasso algorithm (Friedman et al., 2008). It exploits the fact that the estimated neighbors of each node using neighborhood selection coincide with the nonzero entries of the inverse covariance matrix, resulting in high accuracy and fast computation. The Netgsa method was designed to take prior binary node interaction information in one or a few pathways and estimate the edge strengths based on the current data. However, the Netgsa approach does not account for edge weights (e.g., interaction strengths) of known but incomplete protein-protein interactions. In addition, whether the hybrid approach of combining neighborhood selection and maximum likelihood estimation outperforms conventional weighted Glasso has not been well studied.

Besides GGM-based network analysis, there are some recent advances for incorporating prior knowledge in protein or gene network reconstruction. These newly developed methods include Multi-Level PPINs reconstruction (MLPR) (Xu et al., 2018), Diffuse2Direct approach to orient a network (Silverbush and Sharan, 2019), Ensemble Deep Neural Networks with Attention Mechanism (EnAmDNN) (Li et al., 2020), and prior network-dependent gene network inference (pGNI) (Wang et al., 2021). The MLPR method was designed for protein complexes detection through a random walk on the fingerprint similarity networks (Xu et al., 2018). Diffuse2Direct is a diffusion-based method to incorporate prior knowledge and orient an undirected or a partially directed network (Silverbush and Sharan, 2019). The EnAmDNN approach is a deep ensemble learning method to combine multiple models for network construction (Li et al., 2020). The pGNI method was designed to incorporate the modular structures for protein-protein network reconstruction (Xu et al., 2018). However, these newly developed methods are not based on conditional correlation and precision matrix for graph construction. As standard correlation networks, these methods do not take into account the conditional dependencies of proteins, which could lead to potential bias.

In our proposed method, Augmented High-Dimensional Graphical Lasso model (AhGlasso), we first extend the Netgsa hybrid approach to incorporate the edge weights of prior known protein-protein relationships and omics data for global network learning. Then, we implement a screening based on the standard Pearson correlation to further speed up computation. We compare our proposed method with Netgsa and weighted Glasso-based methods in terms of computation time and accuracy with simulated data. Of note, we do not compare AhGlasso with MLPR, Diffuse2Direct, EnAmDNN and pGNIC since these four methods do not provide conditional correlation outputs. Finally, we illustrate an application of AhGlasso on protein expression data from the COPDGene study.

2 MATERIALS AND METHODS

2.1 Graph Structure Learning With Augmented Graphical Lasso

There are several methods for network structure learning including correlation networks and Gaussian graphical models

(GGM). The correlation network method is based on the covariance matrix Σ. $\Sigma_{i,j} = 0$ means that X_i and X_j are marginally independent without observing other variables. However, this kind of independence is hard to find in real-world problems. Instead, GGM is more appropriate since it is based on conditional correlations and the precision matrix. Compared with the more standard correlation network, the conditional independence correlation coefficient is a more sophisticated dependence measure and may be more suitable for modeling real-world biological networks.

To obtain the precision matrix, a common assumption is that the precision matrix Q is sparse. For example, genes are only assumed to interact with a small subset of other genes. Based on the above assumption, we could apply a neighbor selection approach developed by Meinshausen and Bühlmann to learn the relationship between two nodes with Lasso regression, which allows zero parameters through a penalty (Meinshausen and Buhlmann, 2006). Here we take node i as an example to illustrate how to identify one node's neighbors.

$$\hat{\beta}_{node_i} = \underset{\beta_{node_i}}{\arg\min} \|\mathbf{Y} - \mathbf{X}\beta_{node_i}\|_2^2 + \lambda\|\beta_{node_i}\|_1, \qquad (1)$$

where **Y** is the expression value of $node_i$, **X** denotes the expression values of all the other nodes, $\hat{\beta}_{node_i}$ is a vector of estimated coefficients from the Lasso regression, and λ is the sparsity parameter for the Lasso regression. The Lasso regression is repeated for each node and determines whether pairwise nodes are conditionally independent or not (Meinshausen and Buhlmann, 2006). Although the neighborhood selection method is remarkably fast, it is an approximation method for estimating sparse networks. Friedman et al., developed the Graphical Lasso to address sparse inverse covariance estimation with constrained maximum likelihood estimation (Friedman et al., 2008).

GGMs including neighborhood-selection algorithm and Graphical Lasso are widely used in biological applications for network graph construction but ignores known protein/genetic interactions. This approach was recently extended to incorporate partially known information with a weighted graphical Lasso (Li and Jackson, 2015; Zuo et al., 2017). However, it has been known that the Graphical Lasso can be computationally intractable for high-dimensional omics data (Zhang et al., 2018; Fattahi and Sojoudi, 2019). In addition, a Network-Based Gene Set analysis (Netgsa) approach was developed to incorporate prior pathway information (Ma et al., 2016). The Netgsa approach combines the neighborhood selection technique (Meinshausen and Buhlmann, 2006) with constrained maximum likelihood estimation (Friedman et al., 2008). The Netgsa approach was designed to incorporate known binary interaction information in one or a few pathways and estimate the edge strengths. However, it does not take edge weights (e.g., interaction strength) into account.

In our proposed method, AhGlasso, we first extend the Netgsa approach and incorporate the edge weights of prior known but incomplete protein-protein relationships from STRING as shown in **Algorithm 1** to reconstruct the global network. The method combines the neighborhood selection strategy with constrained maximum likelihood estimation using Graphical Lasso algorithm to efficiently reconstruct the global network. We also apply Ψ-screening as discussed below to speed up computation. The input for AhGlasso is expression data $\boldsymbol{D}_{n\times p}$, where n denotes the number of samples and p denotes the number of nodes (genes or proteins). In addition, the input is a prior known PPI matrix $\boldsymbol{W}_{p\times p} = [w_{ij}]$, where $i, j = 1, 2, \ldots, p$ and w_{ij} denotes the edge weights, and for diagonal entries $w_{ii} = 0$. $\boldsymbol{P}$ is the set of all nodes in the network graph. $\boldsymbol{J}$ denotes the set of nodes which have at least one connection with other nodes in the prior known PPI matrix and $\boldsymbol{J}^{\mathrm{c}}$ denotes the set of isolated nodes ($\boldsymbol{J}^{\mathrm{c}} = \boldsymbol{P} \setminus \boldsymbol{J}$). For node i in $\boldsymbol{J}$, $\boldsymbol{J}_i$ indicates other nodes which have prior known connections with node i, while $\boldsymbol{J}_i^{\mathrm{c}}$ denotes other nodes that are not connected with node i. Of note, both $\boldsymbol{J}_i$ and $\boldsymbol{J}_i^{\mathrm{c}}$ do not include the ith node. In the expression data ($\boldsymbol{D}$ matrix), $\boldsymbol{Y}^i$ is a column vector ($n \times 1$) of expression data for node i (i.e., $\boldsymbol{D}_{.,i}$) and $\boldsymbol{X}^{\boldsymbol{J}_i}$ denotes the expression data for the nodes in $\boldsymbol{J}_i$, which are connected to node i. The output of the algorithm is the conditional correlation matrix between nodes, $\hat{\boldsymbol{E}}_{p\times p}$.

2.2 Network Structure Learning With Ψ-screening and Ψ Partial Correlation Coefficient

Liang *et al.* proposed an equivalent measure of partial correlation coefficients for GGM under the assumption of the Markov property and adjacency faithfulness (Liang et al., 2015). They defined the set of nodes v where the edge weight to node i is greater than γ as $\hat{E}_i(\gamma) = \{v\colon |\hat{e}_{iv}| > \gamma\}$, where $\hat{e}$ is the pair partial correlation of nodes i and v, and γ denotes a threshold value. Additional sets are defined as $\hat{E}_{i,-k}(\gamma) = \{v\colon |\hat{e}_{iv}| > \gamma\}\backslash\{k\}$, and $\hat{E}_{k,-i}(\gamma) = \{v\colon |\hat{e}_{kv}| > \gamma\}\backslash\{i\}$, where k denotes a node in network graph not equal to i. The partial correlation coefficient Ψ_{ik} was defined by $\Psi_{ik} = [\psi_{ik}]$, where $\psi_{ik} = \hat{E}_{i,-k}$ if $|\hat{E}_{i,-k}| < |\hat{E}_{k,-i}|$ and $\psi_{ik} = \hat{E}_{k,-i}$ otherwise. With the partial correlation coefficients, the network structure could be learned with the Ψ algorithm (Supplement) including correlation screening (Liang et al., 2015). For correlation screening, we could reduce the size of the neighborhood by removing the nodes having a lower correlation (in absolute value). Similar to the *huge* R library (Zhao et al., 2012), we adapted the correlation screening step in the Ψ-algorithm to AhGlasso to reduce the size of potential neighborhood in **Algorithm 1** and speed up the network estimation. For example, for $i \in J^{\mathrm{c}}$, we find $\hat{\beta}^i = \underset{\beta\in\mathbb{R}^i}{\arg\min} \frac{1}{m}\|\mathbf{Y}^i - \mathbf{X}^i\beta^i\|_2^2 + \lambda\|\beta^i\|$. Instead of computing all potential neighbor nodes $\hat{\beta}^i$ with Lasso, we could reduce the size of the i neighborhood by removing the nodes having a low Pearson correlation (in absolute value). In other words, we only need to find the potential neighbor nodes with high Pearson correlation (absolute value > λ) with node i.

Algorithm 1 | Graph structure learning with augmented high-dimensional graphical Lasso.

```
Input ;
$W$: prior known PPI matrix ($p \times p$);
$D$: expression data ($n \times p$)
for every node $i$ do
    if $i \in J$ = {nodes with known connection with other nodes } then
        (a) for $J_i$ nodes with known connection with node $i$, find $\hat{\beta}^{J_i}$ using linear regression (GGM);
        $\hat{\beta}^{J_i} = \underset{\beta\in\mathbb{R}^{J_i}}{\arg\min} \frac{1}{n}\|\mathbf{Y}^i - \mathbf{X}^{J_i}\beta^{J_i}\|_2^2 + \lambda\|(1-w_{J_i})\beta^{J_i}\|$, where $Y^i$ denotes expression value of
         $i$, $X^{J_i}$ denotes expression values of nodes in $J_i$, and $w_{J_i}$ $(1 \times |J_i|)$ denotes the prior known
         weights between node $i$ and other nodes in $J_i$ ;
        Calculate the residual: $\mathbf{R}^i = \mathbf{Y}^i - \mathbf{X}^{J_i}\hat{\beta}^{J_i}$ ;
        (b) for $J_i^{\complement}$ nodes without known connection with node $i$, find $\hat{\beta}^{J_i^{\complement}}$ with
         $\hat{\beta}^{J_i^{\complement}} = \underset{\beta\in\mathbb{R}^{J_i^{\complement}}}{\arg\min} \frac{1}{n}\|\mathbf{R}^i - \mathbf{X}^{J_i^{\complement}}\beta^{J_i^{\complement}}\|_2^2 + \lambda\|\beta^{J_i^{\complement}}\|$ (optional $\Psi$-screening) ;
        Combine $\hat{\beta}^{J_i}$ and $\hat{\beta}^{J_i^C}$ by the original indices to get $\hat{\beta}^i$
    else
        $i \in J^{\complement}$, find $\hat{\beta}^i$ where $\hat{\beta}^i = \underset{\beta\in\mathbb{R}^i}{\arg\min} \frac{1}{m}\|\mathbf{Y}^i - \mathbf{X}^i\beta^i\|_2^2 + \lambda\|\beta^i\|$(optional $\Psi$-screening);
    end
end
if node $\in$ {nodes with $\hat{E} = \{(i,k) : \hat{\beta}_i^k \neq 0$ or $\hat{\beta}_k^i \neq 0\}\}$ then
    estimate inverse covariance matrix $\hat{\Omega} = \underset{\Omega\in\mathbb{R}}{\arg\min}\{trace(\Sigma\Omega) - log(det\Omega)\}$, where $\Sigma$ denotes the
     empirical covariance matrix, and $\Omega$ denotes the inverse of $\Sigma$;
    estimate conditional correlations;
else
    fill 0 for the remaining node pair associations
end
Output the newly learned network structure: partial correlation matrix $\hat{E}$
```

2.3 Hyper-Parameter Tuning and Model Selection

Like other Lasso-based optimization procedures, the sparsity parameter λ is crucial for AhGlasso because it controls the sparsity of network prediction. If the lambda value is greater than the optimum, we could get an over-sparse network estimation. If the lambda value is smaller than the optimum, we could get over-dense network estimation. There is no consensus on how to select the λ hyper-parameter, which is an active research area. The λ parameter could be tuned by log-likelihood, Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), or the Extended Bayesian Information Criteria (EBIC), and with or without cross-validation. AIC and EBIC often lead to an over-sparse network, while the log-likelihood may result in a too dense network since there is no penalty for the number of edges. Although BIC works well in general low-dimensional scale-free networks, it could lead to under-fitting and an over-sparse network, especially when the network graph is large or does not have the scale-free property. Although real-world networks are often claimed to be scale-free, strongly scale-free structure is rare even in biological domains (Broido and Clauset, 2019). Selecting the criteria for model selection is difficult and depends heavily on uncheckable or difficult-to-check assumptions on the data generating process. K fold cross-validation (CV) provides a potential tool to solve this challenge. In this study, we evaluated different cross-validation options (AIC, BIC, EBIC) and the standard BIC for parameter optimization in a scale-free and non-scale-free network. The model fitting error is AIC, BIC or EBIC for the λ parameter selection. The one standard error rule was also adapted to compare models with different numbers of parameters to select the most parsimonious model with low error (Hastie et al., 2019). Specifically, the simplest model whose mean error falls within one standard deviation from the smallest average (e.g., minimal mean of BIC in CV) achieved for the respective metric (e.g., BIC) was chosen (**Algorithm 2**). Of note, the AIC, BIC, and EBIC are maximum likelihood estimate driven and penalize free parameters in an effort to combat overfitting. In Graphical Lasso model selection, AIC, BIC, and EBIC are often calculated based on the log Likelihood but with different penalization strategies for the number of parameters (Li and Jackson, 2015; Ma et al., 2016; Zuo et al., 2017). Since the Glasso estimates the inverse matrix based on penalized log Likelihood, we calculate AIC, BIC, and EBIC based on the penalized log Likelihood instead.

In order to find the optimal regularization parameter, λ, networks were estimated under a sequence of λ values. The upper bound ($\lambda_{\max}$, representing maximum value of λ) of the regularization parameter which makes all estimates equal to 0 was calculated with the *huge* package (Zhao et al., 2012). With predefined λ minimal ratio (such as 0.01), we can calculate $\lambda_{\min}$ (representing minimal value of λ), which is equal to the λ minimal ratio × $\lambda_{\max}$. A sequence of λ candidates with a predefined number (such as 40) were generated starting from $\lambda_{\min}$ to $\lambda_{\max}$ in a log scale for a grid search.

2.4 Data Simulation

Scale-free biological networks often have two properties: 1) the node degree follows a power-law distribution and 2) the interactions of proteins/genes are sparse. Therefore, we simulated sparse networks with the scale-free property using the R packages *huge* (Zhao et al., 2012) and *fastnet* (Dong et al., 2016) to mimic biological networks (**Figure 1**). The simulation procedures are detailed in the **Figure 1** legend. The density of the simulated networks ranged from 2 to 4%, which was similar to observed densities of protein-protein interactions in the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database for many organisms (Szklarczyk et al., 2016; Ashtiani et al., 2018). For example, when the node number is 1,000, the number of connected edges in 2% density of the simulated network = $\frac{1000\times(1000-1)}{2}\times 0.02$ of connections exist. Of note, the simulated scale-free graph with *huge* R package tends to be very sparse (less than 0.005) when the graph size is larger than 500. We simulated the scale-free graph structure with the modified "net.barabasi.albert" function in *fastnet* R package (Dong et al., 2016). Once the scale-free network is built, *huge* creates the true precision matrix Ω and the partial correlation can be calculated based on the precision matrix. The absolute value of the partial correlation serves as the prior knowledge (prior PPI) for upcoming simulations. We use the absolute value since the

Algorithm 2 | λ optimization in augmented high-dimensional graphical Lasso.

```
Result: $\lambda$ optimization with cross-validation.
We illustrate the algorithm using the BIC, but any of the other metrics (AIC, EBIC, etc) can be
  substituted.
(a) Randomly and equally split data $X$ into $K$ folds, denoted as $X_1, X_2, X_3, ....X_K$;
for each $\lambda \in \Lambda$ (regularization parameter set) do
    for each $k \in \{1, 2, \ldots, K\}$ do
        $X_{input} = [..., X_{k-1}, X_{k+1}, ....]$;
        Run AhGlasso algorithm with $X_{input}$ to obtain the inverse covariance matrix $\hat{\Omega}$ (estimated
          precision matrix);
        Calculate $Penalized\, negative\, loglikelihood = \{trace(\Sigma\hat{\Omega}) - log(det\hat{\Omega}) + \lambda||\hat{\Omega}||_1\}$
          where $\hat{\Omega}$ is the estimated precision matrix;
        Calculate $BIC_{X_k}(\lambda) = penalized\, negative\, loglikelihood + \frac{log(m)}{m}|\hat{E}_i(\lambda)|$ where $\hat{E}_i(\lambda)$ is
          the number of estimated edge set that are non-zero;
    end
    Calculate the average and standard error for $BIC$: $\mu_{BIC}(\lambda) = \frac{\sum_k BIC_{X_k}(\lambda)}{k}$, and
      $SE = \frac{\sqrt{var(BIC_{X_k}(\lambda))}}{\sqrt{K}}$;
end
(b) Obtain $\lambda_{min} = \arg\min_{\lambda \in \Lambda} \mu_{BIC}(\lambda)$ that achieves the minimal average of BIC;
(c) Obtain optimal $\lambda$ by moving $\lambda_{min}$ in the direction of increasing regularization to one standard
  error limit, $\lambda_{opt} = \{\lambda : (\mu_{BIC(\lambda_{min})}) + SE(BIC(\lambda_{min}))\}$
Output the optimal $\lambda$
```

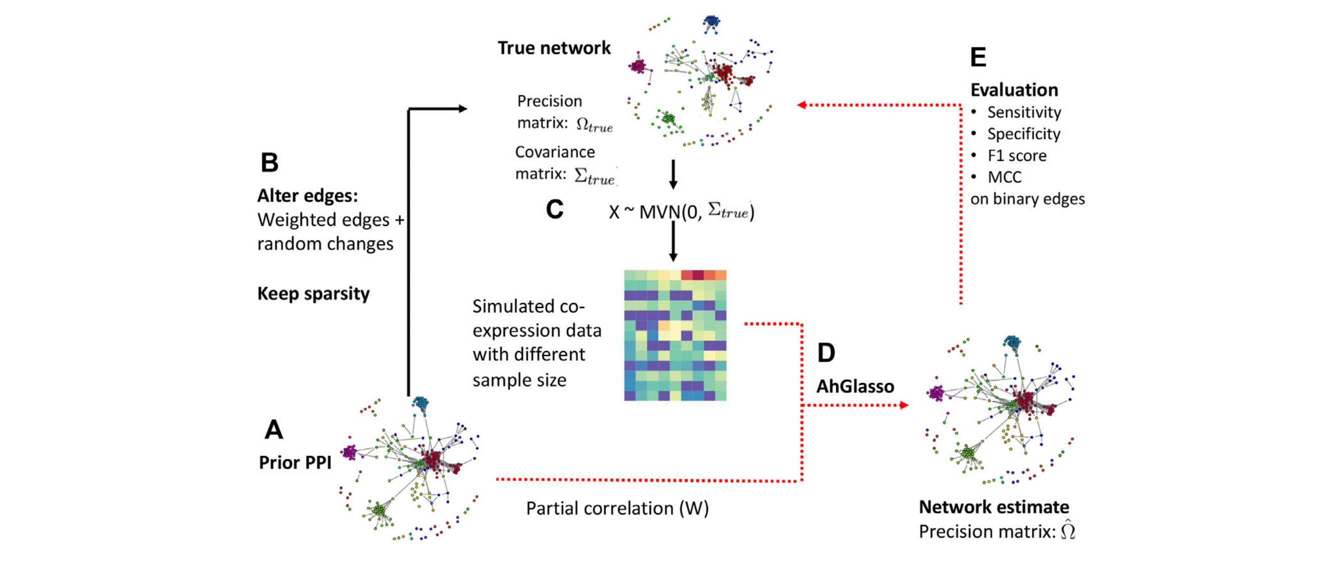

FIGURE 1 | Flowchart of data simulation and method evaluation. Step **(A)**: a scale-free prior network was built by inputting the node number p (graph size). With the prior scale-free network and its precision matrix Ω, the partial correlation can be calculated. The absolute value of partial correlation serves as prior knowledge (prior PPI) for upcoming simulations. Step **(B)**: we altered network connections by adding random changes but the sparsity was kept at a similar level unless otherwise specified. Strong node connections were likely kept due to their higher absolute value in the precision matrix and were less likely to become 0. Step **(C)**: with altered precision matrix (Ω_{true}) and its corresponding covariance matrix (Σ_{true}), we simulated expression data $X_{n \times p} \sim N$ (0, Σ_{true}). Of note, we ensured the covariance matrix to be positive definite based on the Frobenius norm. Step **(D)**: incorporating prior PPI and simulated dataset, we estimated the target network (i.e., adjacency matrix) with our proposed method. Step **(E)**: the prediction performance was evaluated with sensitivity, F1 score, and MCC.

direction of many protein-protein interactions in STRING is not specified.

We altered network connections by adding random changes on the precision matrix but keeping the sparsity at a similar level because the biological network is dynamic. For simplicity, we assume the total number of newly appearing interactions is close to the total number of existing interactions that get lost. The precision matrix controls the magnitude of partial correlations. The original off-diagonal elements of the precision matrix range from 0.2 to 1. The symmetric uniform distributed random values range from −0.5 to 0.5. Strong node connections were likely kept due to their higher values in the precision matrix. If the absolute values of altered elements in the precision matrix were less than 0.2, we reset them to 0. The altered precision matrix and its corresponding covariance matrix served as the target network we would infer and compare to evaluate method performances. We defined them as "true" precision matrix and covariance matrix for simplicity. With the altered precision matrix (Ω_{true}) and its corresponding covariance matrix (Σ_{true}), we could simulate expression data $X_{n \times p} \sim N$ (0, Σ_{true}). Of note, we obtain a positive definite precision matrix with the Frobenius norm through the *huge* R package (Zhao et al., 2012). Specifically, the smallest eigenvalue of the precision matrix Ω_{true}, denoted by λ_1 is computed. Then we set the precision matrix equal to $\Omega + (|\lambda_1| + 0.1)I$. The covariance matrix is then computed to generate multivariate normal data.

We created simulation datasets with various p (graph size) and n (sample size) as well as different overlapping degree between the prior known network and the target true network. The overlapping degree between the prior known network and the target true network was defined as ρ. The ρ was calculated by comparing the prior known precision matrix and target precision matrix in binary format (i.e., non-zero weights were converted to one for simplified calculation). The range of ρ between 0 and 100 was explored. Specially, adding random changes on the precision matrix above leads to a certain degree of ρ change. To allow for a variety of overlapping percentages, we randomly replaced some non-zero edges in precision matrix to become 0 with predefined mutation percentages while the same number of randomly selected zero edges were replaced to be non-zero. In other words, a similar sparsity level of the network was kept. The ρ degree was determined with the altered network and the prior network at the end. Using the prior PPI and simulated dataset, we estimated the target network with our proposed new method. We also estimated the network without prior PPI information as the baseline control. We implemented two published weighted graphical Lasso approaches and the Netgsa method for comparison (Li and Jackson, 2015; Ma et al., 2016; Zuo et al., 2017). In the Netgsa implementation, non-zero weights in prior PPI were converted to 1.

We created simulation datasets with various p (graph size, from 400 to 3000) and n (sample size, from 100 to 1000) for evaluating different λ tuning criteria and comparing different methods. In order to systematically evaluate AhGlasso with prior knowledge, simulations were performed with a variety of overlapping percentages between prior knowledge and target true network (ranging from 0 to 100%).

In reality, the prior PPI is often partially-known but not totally noisy. We also investigated the method performance when the

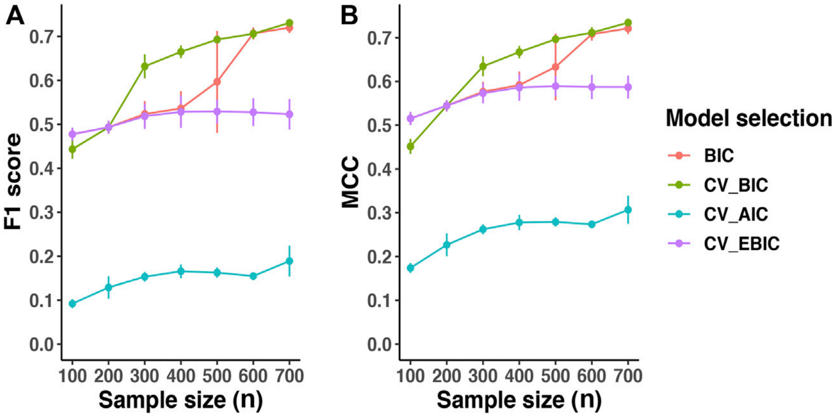

FIGURE 2 | Comparison of Model Selection Criteria. The simulated protein network included 500 (*p*) nodes. The overlapping between prior information and target true network is 50%. With the same true network and its corresponding covariance matrix (Σ_{true}), we created various sizes (*n*) of multiple normal expression data for testing. The λ was optimized with regular BIC, cross-validation-based BIC (CV_BIC), cross-validation-based AIC (CV_AIC) and cross-validation-based EBIC (CV_EBIC). The AhGlasso algorithm with the optimal λs derived from different criteria outputs the predicted network. The F1 score and MCC were calculated based on the estimated network and true network. For each simulation setting, the simulations were repeated 5 times. The lines represent the mean scores for the simulated sample size and the error bars represent the standard error of the mean. Of note, similar results were achieved in various *p* and *n* simulations (data not shown).

prior network is a subset of the true network. We randomly removed the connected edges in the prior network under the predefined subset percentage (ranging from 10 to 100%) of the target network. If the subset percentage for the prior network is 100%, the prior network is the same as the target network. If the subset percentage for prior network is 50%, the density of the prior network is 50% of the density of the target network.

2.5 Model Evaluation

To measure the accuracy of the estimation and compare the method performances, we focused on the accuracy of whether an edge existed in the true graph and was also estimated to be non-zero. That is, we converted non-zero estimated edges to one in the edge matrix and counted whether in the true graph this edge existed (and vice versa) to define negative or positive prediction. We use several metrics including sensitivity, specificity, F1 score, and Matthews correlation coefficient (MCC). The F1 score can be interpreted as a weighted average of the precision and recall and is a suitable measure for imbalanced datasets like sparse networks. The sparse network is one of extremely imbalanced datasets since only a small percent of edges between proteins exist while the majority of elements in the adjacency matrix are 0. F1 score is computed as

$$F1 = \frac{2 * Precision{*}Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}, \quad (2)$$

where Precision = $\frac{TP}{TP+FP}$, and Recall = $\frac{TP}{TP+FN}$. TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively.

MCC is another metric for measuring classification performance and is widely used in network prediction. MCC takes into account all four values in the confusion matrix, and a high value (close to 1) means that both classes are predicted well. MCC was calculated as follows based on 2 × 2 contingency table,

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(FP + FN)(TN + FP)(TN + FN)}}, \quad (3)$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively.

2.6 STRING PPI Database

STRING (http://www.STRING-db.org) is a database of known and predicted protein-protein interactions. It currently covers 5,214,234 proteins from 1133 organisms (Szklarczyk et al., 2016). In STRING, protein-protein pair associations (i.e., the "edge weights" in each network) are represented by confidence scores. The scores indicate the estimated probability that a given interaction is biologically meaningful, specific, and reproducible, given the supporting evidence. There are seven evidence channels in STRING: 1) experiments; 2) database; 3) text-mining; 4) coexpression; 5) neighborhood; 6) fusion; and 7) co-occurrence. The edge scores (weights) between proteins in the STRING PPI database range from 0 to 1, with 1 being the highest possible confidence of interaction. In the COPD study, we retrieved human PPI data from STRING and filtered out the prior interactions with scores less than 0.2.

2.7 Proteomics Data in COPDGene Study

As an application of the methodology to real data, we used the proteomics data generated in the COPDGene Study. COPDGene is a multicenter genetic epidemiology study that enrolled 10,198 participants with and without the chronic obstructive pulmonary disease (COPD) between 2007 and 2011 (Phase I study) to identify genetic factors associated with COPD (Ragland et al., 2019). COPD is a disease characterized by reduced lung function and symptoms

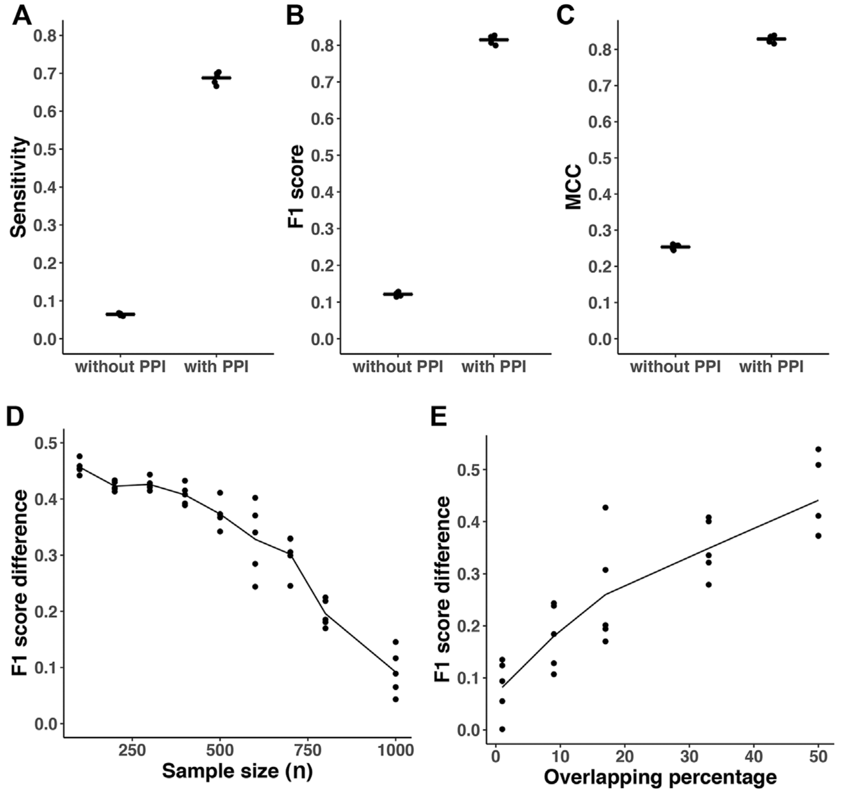

FIGURE 3 | Prediction accuracy comparison with or without incorporating prior information. All results are based on p = 500 nodes and n = 200. The density of simulated graph is around 2%. The λ was optimized with cross-validation-based BIC (CV_BIC). The AhGlasso algorithm with the optimal λs output the predicted networks with or without incorporating prior knowledge. **(A–C)** The overlapping degree between prior information and the target true network is 80%. The sensitivity **(A)**, F1 score **(B)**, and MCC **(C)** were calculated based on the estimated network and true network. For each simulation setting, the simulations were repeated 5 times. The bars represent the means of corresponding statistics metrics. Pair student's t-tests were performed to compare the corresponding metrics. All p values <0.0001. **(D)**, the overlapping degree between prior information and target true network is 50%. A variety of sample sizes of data was simulated as shown in X-axis. The simulation was repeated 5 times for each simulation setting. Y-axis represents the difference of F1 score between the outputs with or without prior knowledge. The lines represent the mean scores and the dots represent the results in each simulation. **(E)**, a variety of overlapping degree between prior information and target true network was simulated as shown in X-axis. Y-axis represents the difference of F1 score between the outputs with or without prior knowledge. The lines represent the mean scores and the dots represent the results in each simulation.

such as shortness of breath. Five-year follow-up visits took place from 2013 to 2017 (Phase II study). Proteomic profiles were constructed on participants who agreed to participate in the ancillary study of Phase II COPDGene study. All analyses were performed on frozen plasma from p100 tubes. After removing observations that did not pass QC or have no phenotype data, were duplicates, whose primary pulmonary diagnosis was not COPD, were a never smoker, and at Phase 2 reported having a lung transplant or lung volume reduction before Phase 2, there was 1206 subject in Phase 1 and 1010 in Phase 2. The Global Obstructive Lung Disease (GOLD) system was used to grade the severity of airflow limitation: GOLD 0 (controls) and GOLD 1–4 (COPD cases). Our study focuses on the 486 COPD cases (GOLD stage >0) in the Phase 2 study since the inherent protein networks between controls and COPD cases might be different (Mastej et al., 2020).

Proteomic profiling was performed using the SomaScan® platform (Boulder, Colorado) (Candia et al., 2017). The Human Plasma SomaScan® 1.3k kit (SL Part Number 900–00011) was used following the manufacturer's recommended protocol. Data from all samples passed quality-control criteria and were fit for analysis. To map with the STRING database, the proteomics expression data without one-to-one mapping to gene symbols were removed. For example, if two SomaScan® aptamers map to the same gene symbol, these two aptamers' corresponding expression data were removed. In addition, some aptamers either detect the expression level of a protein complex or detect the total expression level of several proteins by targeting a shared subunit. These aptamers were removed as well for simplicity. Expression data for 1212 proteins were retained for network construction.

2.8 Gene Ontology Enrichment Analysis on Hub Proteins in COPD Associated Network

Using the prior PPI network retrieved from the STRING database, we applied AhGlasso and Netgsa to construct COPD-associated networks. In addition, we also constructed a COPD-associated network without a prior PPI network for

TABLE 1 | Summary of network learning methods to incorporate prior knowledge.

Method	Published year	Abbreviation	Algorithm	Weight for prior knowledge	Model selection	Screening
Weighted Glasso (wGlasso)	2015	*wGlasso*_2015	Weighted Glasso	Continuous	BIC[a]	No
Weighted Glasso(wGlasso)	2017	wGlasso_2017	Weighted Glasso	Continuous	CV_log likelihood	No
Netgsa	2017	Netgsa	NB and Glasso	Binary	BIC[a]	No
AhGlasso	—	AhGlasso	NB and Glasso	Continuous	CV_BIC[b]	Yes

Notes: BIC, bayesian information criterion; CV, cross-validation; NB, neighbor selection; AhGlasso, augmented high-dimensional Graphical Lasso;

[a], based on log Likelihood;

[b], based on penalized log Likelihood.

comparison. Because there is a lack of ground truth to evaluate the prediction accuracy, we performed Gene Ontology (GO) enrichment analysis with Fisher's exact test using *the topGO* R package (Alexa and Rahnenführer, 2009) on hub proteins in the COPD-associated networks. Specifically, we analyzed the top 40 hub proteins to identify significantly enriched molecular functions in the updated COPD networks. The hub proteins were defined based on the degree of nodes. Gene Ontology (GO) is a well-known framework for supporting the computational representation of biological systems (Ashburner et al., 2000). It defines a set of concepts used to describe the functions of gene products, and relationships between these concepts. It contains three aspects that hold terms defining the basic concepts of molecular function (MF), biological processes (BP), and cellular components (CC), respectively. Specifically, a GO annotation is an association between a specific gene product and a GO concept. GO was well established and has often been used to evaluate the quality of newly constructed or reconstructed protein-protein interaction networks (Xu et al., 2018; Seyyedsalehi et al., 2021). We focused on BP ontology enrichment analysis since we are interested in what biological processes are involved in COPD. The adjusted p values were calculated with Benjamini-Hochberg Procedure for False Positive Rate (FDR) correction. The significant level was set to FDR <0.05.

2.9 Statistical Software

Unless otherwise specified, the data manipulation and data analyses were performed using *RStudio* (version 1.2.5019) (RStudio Team, 2019) and R (version 4.0.3) (R Core Team, 2020). The R packages *ggplot2*_1.8.6 (Wickham, 2009), ***biomaRt***_2.44.4 (Durinck et al., 2005), *topGO*_2.40.0 (Alexa and Rahnenführer, 2009), *plyr*_1.8.6 (Wickham, 2011), *Netgsa*_3.1.0 (Ma et al., 2016), *Glasso*_1.11 (Friedman et al., 2008) and *huge*_1.3.4.1 (Zhao et al., 2012) were used for data preparation and gene network analysis from differential expression data.

3 RESULTS

3.1 Model Selection Criteria

Since λ controls the sparsity of the output network, selecting the λ parameter is crucial for Lasso-based approaches like AhGlasso. The λ parameter can be selected based on a variety of criteria but there is no consensus on the best criteria. Cross-validation provides a general tool to solve this kind of challenge. For $p = 500$, we found that cross-validation with BIC provides higher accuracy in terms of F1 score (**Figure 2A**) and MCC (**Figure 2B**) than BIC without cross-validation when the sample size is between 300 and 600. When the sample size is larger, BIC with or without cross-validation is similar. Cross-validation with BIC outperforms cross-validation with AIC and EBIC. Finally, cross-validation with AIC and EBIC results in an under-fitting model and over-sparse network, leading to a lower F1 score and MCC. Therefore, we chose cross-validation with BIC as model selection criteria to select the λ parameter henceforth.

3.2 Incorporating Prior Knowledge Significantly Improves Prediction Accuracy

In different simulation settings, we found that the predicted networks based on *a priori* information had significantly greater performance than estimations without *a priori* information (**Figure 3**). For simulations with $p = 500$, $density = 2\%$, $overlapping(\rho) = 80\%$, the mean of sensitivity, F1 score and MCC in the estimated networks with prior knowledge are 0.68, 0.81 and 0.82, respectively (**Figures 3A–C**). However, the corresponding mean of sensitivity, F1 score, and MCC without prior knowledge are only 0.07, 0.12, and 0.25, respectively. The differences of corresponding metrics are statistically significant ($p < 0.0001$, paired student's t-test). In addition, with fixed 50% overlapping, we found that the F1 score difference decreases when the sample size gets larger (**Figure 3D**). It is expected that when we have a larger sample size, the advantages of incorporating prior information are diminished compared to smaller sample sizes. The F1 score difference is also sensitive to the amount of overlapping percentage between the prior knowledge and target network (**Figure 3E**). When the overlapping percentage is larger (i.e., the prior information is more accurate), it provides more useful information for network reconstruction. In addition, we also performed simulations with $p = 1000$, $density = 4\%$, $overlapping(\rho) = 50\%$ with similar results (data not shown).

3.3 AhGlasso Outperforms Conventional Methods

Recently, several groups have extended GGM to weighted Graphical Lasso (wGlasso) and network-based gene set

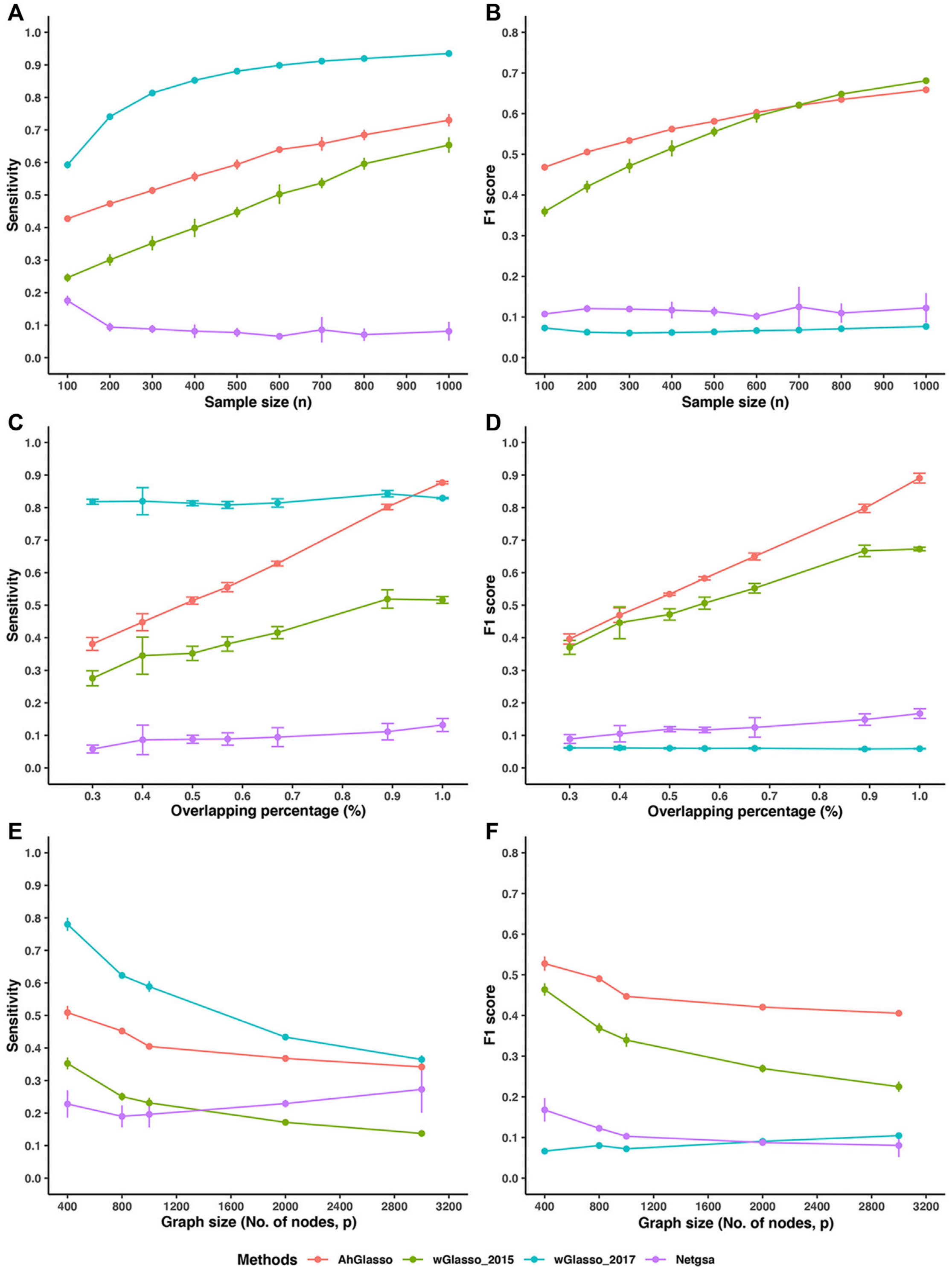

FIGURE 4 | Method performance comparisons. The simulated protein network graph included 1000 (*p*) nodes. The density of the simulated graph is around 4%. **(A–B)** The overlapping between prior information and target true network is 50%. A variety of sample sizes of data was simulated as shown in X-axis. With the same true network and its corresponding covariance matrix (Σ_{true}), we created various sizes (*n*) of multiple normal expression data for testing. We estimated the true network topology by using two weighted graphical Lasso (wGlasso_2015 and wGlasso_2017), Netgsa, and the proposed AhGlasso methods. The sensitivity **(A)** and F1 score **(B)** were calculated based on the estimated network and true network. **(C–D)** A variety of overlapping between prior information and target true network was simulated as shown in X-axis. The sample size was fixed at *n* = 300. The sensitivity **(C)** and F1 score **(D)** were calculated based on the estimated network and true network. **(E–F)** We investigated the effect of different graph sizes (p ranges from 400 to 3000) with a fixed sample size (*n* = 100) and the overlapping between prior network and target network (50%). The sensitivity **(E)** and F1 score **(F)** were calculated as previously. For each simulation setting, the simulations were repeated 5 times. The lines represent the mean scores for the simulated sample size and the error bars represent the standard error of the mean for each method.

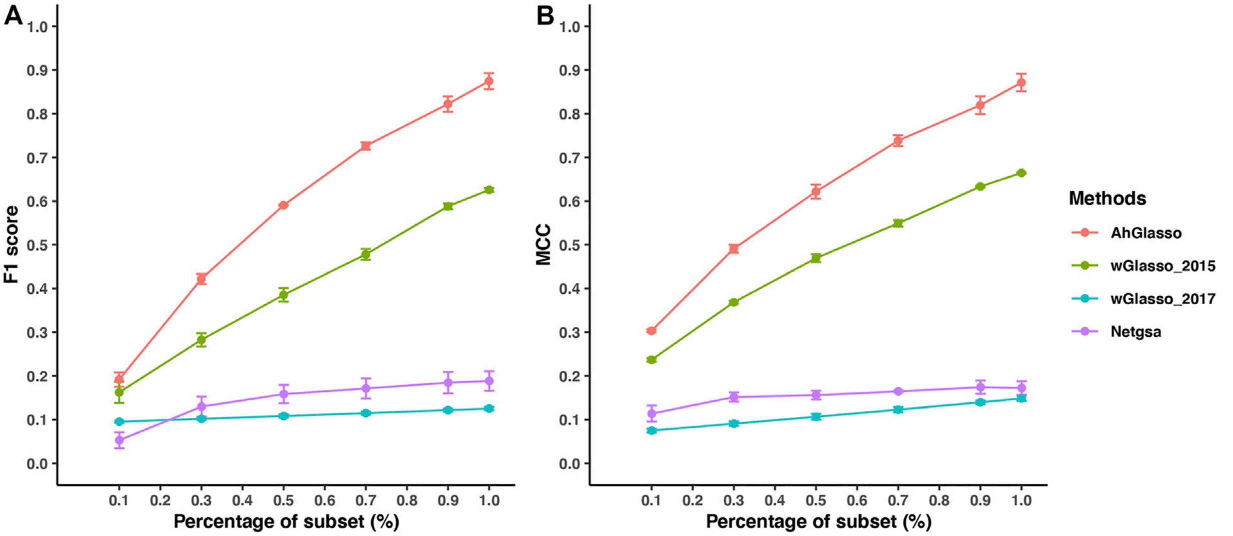

FIGURE 5 | Method performance comparisons with partially known prior information. The simulated protein network graph includes 1000 (p) nodes with fixed sample size (n = 300). The density of the simulated graph is around 4%. The prior information is a subset of the target true network. A variety of subset percentages of the prior network were simulated as shown in X-axis. We estimated the true network topology by using two weighted graphical Lasso (wGlasso_2015 and wGlasso_2017), Netgsa, and the proposed AhGlasso methods. The F1 score **(A)** and MCC **(B)** were calculated based on the estimated network and true network. For each simulation setting, the simulations were repeated 5 times. The lines represent the mean scores for the simulated sample size and the error bars represent the standard error of the mean for each method.

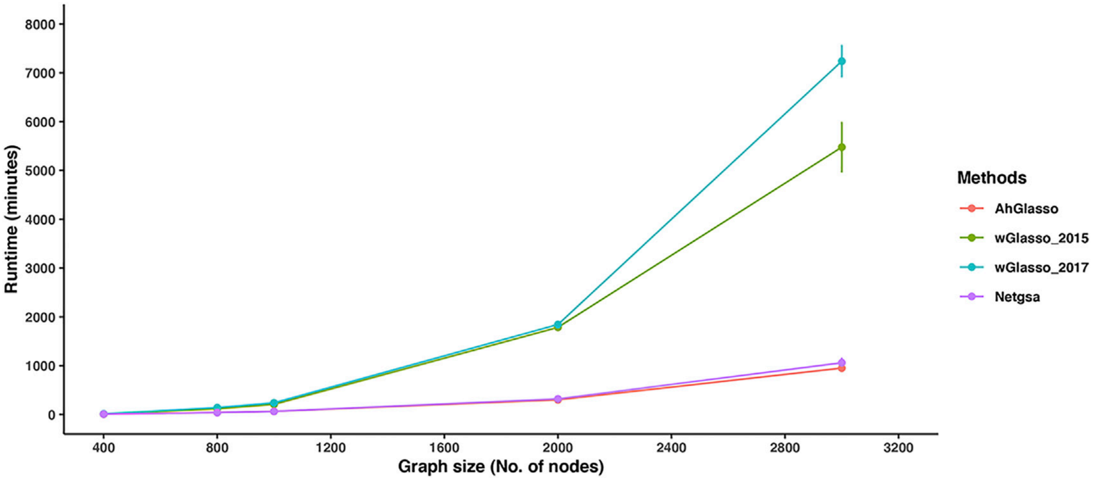

FIGURE 6 | Method runtime comparisons. To compare the computational efficiency of the two weighted graphical Lasso (wGlasso_2015 and wGlasso_2017), Netgsa, and the proposed AhGlasso method in the high-dimensional setting, we simulated datasets with a fixed sample size (n = 200) and 4% graph density, but various graph size (p varies from 400 to 3000). We compared the total run time of parameter tuning (CV was removed if involved for fair comparisons) and the final model fitting with the chosen parameter. The λ minimal ratio was set to 0.01 and 40 λ candidates were searched. The simulation was repeated 5 times for each setting. The lines represent the mean runtime for the simulated graph size and the error bars represent the standard error of the mean for each method.

analysis (Netgsa) to incorporate prior biological information (**Table 1**). These methods incorporate prior knowledge into graphical model-building procedures as well as gene set analysis. Recently, Yi *et al.* and Zuo *et al.* implemented weighted Graphical Lasso (Glasso) algorithms to incorporate prior known network information for network learning and the key difference between the two methods is how to select the λ parameter (Li and Jackson, 2015; Zuo et al., 2017). In Yi *et al.*'s study, λ was optimized by the BIC criteria (wGlasso_2015) while it was tuned with likelihood and cross-validation in Zuo *et* al.'s research (wGlasso_2017). Besides weighted Graphical Lasso, Ma *et al.* developed a Network-Based Gene Set analysis (Netgsa) approach for network learning and gene-set enrichment analysis by incorporating prior pathway information (Ma et al., 2016). The Netgsa approach combines the neighborhood selection technique (Meinshausen and Buhlmann, 2006) with constrained maximum likelihood estimation.

We estimated the true network topology using AhGlasso and the three alternative methods (wGlasso_2015, wGlasso_2017, Netgsa). To make a fair comparison, we tuned the regularization parameter in each method with its designed optimization method. In this comparison study, we mimicked the high-dimensional setting and created simulation datasets with large p (1000) and various sample sizes. We found that our proposed method achieved higher sensitivity and F1 score than wGlasso_2015 when the sample size is small (**Figure 4**). They had similar performances when the sample size is large. These two methods achieved higher F1 scores and had an overall higher accuracy than the other two methods. The F1 score in the wGlasso_2017 method was consistently and extremely low although it achieved very high sensitivity. The specificity of wGlasso_2017 is very low (data not shown) since the weighted Graphical Lasso in Zuo *et a*l.'s study selects the model based on the log-likelihood only but does not penalize the number of edges, which leads to overfitting the model (over-dense output network). Although the Netgsa approach outperforms wGlasso_2017, its sensitivity and F1 scores are still substantially lower than our proposed method and wGlasso_2015. The MCC score patterns were similar to the F1 scores pattern (data not shown). Of note, the improvement of the proposed method decreases when compared with wGlasso_2015 when the sample size increases as shown in **Figure 4B**. However, it is not uncommon to have a limited sample size in high-throughput omics studies, especially in human studies using tissue samples but not blood.

We next investigated the effect of overlapping percentages between the prior network and target network with fixed graph size (p = 1000) and sample size (n = 300). We found that AhGlasso achieved higher sensitivity and F1 score than wGlasso_2015 and the difference increases when the overlapping percentage increases (**Figures 4C,D**). AhGlasso and wGlasso_2015 achieved higher F1 scores and had an overall higher accuracy than the other two methods.

In addition, we also investigated the effect of different graph sizes (p ranges from 400 to 3000) with a fixed sample size (n = 100) and the overlapping between prior network and target network (50%). We found that AhGlasso achieved higher sensitivity and F1 score than wGlasso_2015 and the differences increase when the graph size increases (**Figures 4E,F**). AhGlasso and wGlasso_2015 achieved higher F1 scores and had an overall higher accuracy than the other two methods (**Figure 4F**).

Since strong scale-free networks can be rare in biological contexts, we also compared the method performance on a simulated non-scale-free network. We first simulated a random network with p = 500 and different sample sizes ranging from 200 to 700. The overlapping percentage between prior knowledge and target network was set at 88%. We found that our proposed method has similar sensitivity scores to wGlasso_2015 and wGlasso_2017 (**Supplementary Figure S1**) and it outperforms the other three methods in terms of F1 scores in our tests (**Supplementary Figure S1**). In addition, we simulated with a fixed sample size (n = 300) but different degrees of overlapping between prior knowledge and target truth network. We also found that our proposed method outperforms the other three methods in terms of F1 scores in our tests (**Supplementary Figure S2**). Besides random networks, we also simulated networks with hub or cluster structures for comparison. We found that AhGlasso achieves higher F1 scores than wGlasso_2015, wGlasso_2017, and Netgsa in networks with both cluster network and hub network (data not shown).

In the real world, the prior PPI is often incomplete rather than purely noisy. We next investigated the method performance when the prior is a subset of the true network (i.e., incomplete prior information). We simulated datasets with a fixed sample size (n = 300) and graph density (4%). We randomly removed the connected edges in the prior network under the predefined subset percentage of the target network (**Figure 5**). When the subset percentage of the prior network increases, the F1 score increases in all the tested methods. Under the same subset percentage, our proposed AhGlasso method achieves a higher F1 score (**Figure 5A**) and MCC (**Figure 5B**) than the other three methods. The differences between AhGlasso and wGlasso_2015 increase when the subset percentage increases (**Figure 5**).

3.4 Comparison of Runtimes

To compare the runtimes of the tested methods, we simulated datasets with a fixed sample size (n = 200) but various graph sizes (p varies from 400 to 3000). We evaluated the computation time based on an Intel(R) Xeon(R) Gold 6152 CPU @ 2.10 GHz CentOS Linux 7 (Core) operating system. Since the regularization parameter tuning in each method is a critical step, we compare the total run time of parameter tuning (CV was removed if involved for fair comparisons) and the final model fitting with the chosen parameter. The simulation was repeated 5 times for each setting. With p = 1000, the total running time for AhGlasso, wGasso_2015, wGlasso_2017, and Netgsa was 64 ± 6.64 min, 207.51 ± 36.70 min, 239.97 ± 44.28 min, and 64.34 ± 7.4 min, respectively (**Figure 6**). The runtimes of AhGlasso and Netgsa are comparable and they are around 4 times faster than the two weighted Graphical Lasso-based algorithms when p = 1000. The runtimes of wGlasso_2015 and wGlasso_2017 exponentially increase when the graph size increases. With p = 3000, the total running time for AhGlasso, wGasso_2015, wGlasso_2017, and Netgsa was 951.09 ± 173.51 min, 5477.11 ± 815.34 min, 7238.65 ± 523.54 min, and 1060.65 ± 238.83 min, respectively. The runtime of AhGlasso is around 7 times faster than the two weighted Graphical Lasso-based algorithms when p = 3000. The runtime of AhGlasso is also faster than Netgsa. Of note, the F1 score patterns in these simulations for tested methods were similar to **Figure 4B** (data not shown). In addition, the runtimes of weighted Graphical Lasso-based algorithms are much more sensitive to the λ minimal ratio than AhGlasso and Netgsa. The runtime differences between AhGlasso and weighted Graphical Lasso-based algorithms are greater when decreasing the λ minimal ratio and searching more λ candidates.

When the feature space is large, the graphical Lasso-based methods are computationally expensive and even intractable. AhGlasso preselects the potential neighbor nodes with ψ screening to narrow down the neighbor node space. In addition, the Meinshausen-Bühlmann algorithm (Meinshausen and Buhlmann, 2006) is incorporated in AhGlasso to select neighbor nodes before maximum likelihood estimation. Due to the screening and selection steps, AhGlasso is more efficient and faster than weighted Glasso methods in large-scale graph construction.

3.5 AhGlasso Improves Network Inference for COPD

For a real data application, we used proteomics data generated in the COPDGene Phase II Study. COPD is a disease characterized by reduced lung function and symptoms such as shortness of breath. Protein-protein interactions could play important roles in COPD pathogenesis in COPD development. We collected a large proteomic data from 1010 subjects from the COPDgene cohort using the SomaScan® platform (Candia et al., 2017). We constructed a COPD-associated network on COPD cases (n = 486) with or without PPI data and identified important protein-protein interactions contributing to COPD development.

We incorporated prior known PPI information from STRING and constructed COPD-associated networks with the AhGlasso and Netgsa methods. The top 40 hub proteins were chosen for GO enrichment analysis. In the AhGlasso analysis with known PPI knowledge, we found 35 molecular function pathways that were significantly enriched while only 12 pathways were enriched in the analysis without prior PPI knowledge (**Supplementary Table S1** and **Supplementary Table S1**), and only 17 pathways were enriched in the Netgsa analysis (**Supplementary Table S1**). After multiple testing corrections with FDR, we found 23 molecular function pathways that were significantly enriched in AhGlasso analysis with prior PPI information while only one pathway was enriched in the analysis without prior PPI knowledge, and no pathways were enriched in the Netgsa analysis. AhGlasso also identified six hub genes related to the cadherin pathway, which has been reported to play important roles in COPD development (Nelson and Nusse, 2004; Kneidinger et al., 2011; Eapen and Sohal, 2020), but was not enriched in the results of the Netgsa method. In addition, AhGlasso also uniquely enriched cytokine signaling and chemokine signaling pathways in COPD-associated networks, which have been reported to be important for COPD pathogenesis (Bradford et al., 2017; Henrot et al., 2019).

4 DISCUSSION

In this study, we have developed an augmented high-dimensional Graphical Lasso model (AhGlasso) to incorporate edge weights from known protein-protein interaction networks with omics data for global network learning. In our proposed method, we first extended the Netgsa hybrid approach to incorporate the edge weight of prior known but incomplete protein-protein relation for network reconstruction. To speed up the computation and make it feasible for large-scale data, especially when the number of variables is much larger than the sample size, we also implemented Ψ-screening based on the standard Pearson correlation. We compared our proposed method with Netgsa and two weighted Graphical Lasso approaches in terms of computation time and accuracy based on simulations where "ground truth" for the target network is available.

To systematically evaluate the performance of methods and make fair comparisons, we simulated datasets with a variety of network graph sizes, sample sizes, and overlapping percentages between the prior information and the target network for estimation. For network structure learning with GMM, the λ parameter controls the sparsity of the output network. Tuning the λ parameter is a key step for AhGlasso and other Lasso-based approaches. However, there is no consensus on how to select the λ regularization parameter. Cross-validation (CV) provides a general tool to solve this kind of challenge. Our simulation study suggested that cross-validation with BIC is more stable and outperforms the other criteria when the sample size is medium. When the sample size is small (such as $n < (0.25 \times p)$), conventional BIC has similar or even better performance than BIC with cross-validation, which may be explained that the small sample sizes for the K fold cross-validation lead to large variance. When the sample size is very large, the performance of conventional BIC and BIC with cross-validation are similar, which could be explained that the models with or without cross-validation converge to the unbiased true model when data is sufficient. In addition, we also found that cross-validation with AIC and EBIC often results in the under-fitting of the model and over-sparse networks, leading to lower F1 score and MCC.

Our simulation study found that the AhGlasso output estimated networks with *a priori* information had a significantly greater sensitivity, F1 score, MCC than the estimations without *a priori* information. The F1 score difference decreases for larger sample sizes because the impact of incorporating prior information decreases with more data. The F1 score difference is also sensitive to the amount of overlapping percentage between the prior knowledge and the target network. As expected, when the prior information is less accurate, it provides less useful information for network reconstruction. However, when the overlapping percentage between prior information and the target network is low, we did not observe an increase in the false-positive rate.

The key difference between the two weighted Glasso methods that incorporate prior known network information for network graph learning is how they select the λ parameter (Li and Jackson, 2015; Zuo et al., 2017). In wGlasso_2015, the λ was optimized with the BIC criteria while it was tuned with likelihood and cross-validation in wGlasso_2017 (Li and Jackson, 2015; Zuo et al., 2017). The wGlasso_2017 approach often resulted in high sensitivity but extremely low specificity, which leads to a low F1 score. This could be explained that wGlasso_2017 tunes the λ parameter based on likelihood with a lack of penalty for the number of edges, which leads to an over-dense network estimation. Our study also demonstrated AhGlasso generally outperforms Netgsa in terms of F1 scores, which suggests the advantages of taking edge weights into account. Although wGlasso_2015 could achieve comparable accuracy to AhGlasso, its computation is intractable for large-scale data. It took days when the network graph size is bigger than 1000. Compared with wGlasso_2015, our proposed method, AhGlasso, is computationally scalable and much more efficient than the weighted Glasso based methods when the sample size is large. In summary, the new method, AhGlasso, outperforms wGlasso-based algorithms with respect to computational time in simulated large-scale data settings, while achieving better or comparable prediction accuracy of node connections.

In the COPDgene study, AhGlasso with prior PPI found more enriched GO terms than the estimated network without PPI on the top 40 hub proteins in the resultant networks (23 vs. 1 after multiple testing correction with FDR). We also found AhGlasso discovered more enriched GO terms than Netgsa on the top 40 hub proteins in the resultant networks (23 vs. 0 after FDR correction). After FDR correction, we found significantly enriched cadherin, chemokine signaling, cytokine signaling pathways in AhGlasso, but not with the Netgsa analysis or the analysis without PPI. These three pathways have been demonstrated to play important roles in COPD. Several studies reported that the cadherin/WNT/catenin pathway could be a novel therapeutic target for attenuating airway remodeling in COPD (Nelson and Nusse, 2004; Kneidinger et al., 2011; Eapen and Sohal, 2020). This real data study suggests that AhGlasso improves COPD-related network inference compared to the Netgsa approach in integrating a proteomics dataset and prior PPI with edge weights.

Although we only focused on analyzing single omics data in this study, the partial correlation coefficient derived in AhGlasso can be transformed to a Z score via Fisher's transformation for multiple omics data integration. The Z scores from different single omics data can be easily combined using Stouffer's meta-analysis method (Stouffer et al., 1949).

Although AhGlasso outperforms wGlasso-based algorithms and Netgsa in our simulations and COPD case study, there are some limitations in the method. One limitation is that AhGlasso can be computationally intensive when the graph size is large, but still faster than the alternatives compared. Another limitation is that the optional K-fold cross-validation step can significantly increase the computation burden. Furthermore, K-fold cross-validation is not recommended if the sample size is small. Although we compared AhGlasso with wGlasso_2017 and Netgsa, the latter were not designed for global network learning. The wGlasso_2017 approach is designed for network-based differential gene expression analysis using differentially weighted graphical Lasso on pre-selected differentially expressed genes. The Netgsa approach is designed for network-based pathway enrichment analysis and focused on protein-protein interaction changes in known pathways. Furthermore, to implement Netgsa non-zero weights in the prior PPI were converted to 1, which could cause potential biases.

5 CONCLUSION

We present an augmented method, called AhGlasso, for incorporating prior information in biological network reconstruction. Our study suggests that cross-validation with BIC generally performs better than regular BIC, cross-validation AIC, or EBIC. This new method outperforms wGlasso-based algorithms with respect to computational time in large-scale data settings while achieving better or comparable prediction accuracy of edges. Our method also improves COPD-associated network inference compared to the Netgsa approach. Although demonstrated on one -omics data and prior PPI, our method could be generalized to multi-omics data.

DATA AVAILABILITY STATEMENT

Clinical Data and SOMAScan data in this study are available through COPDGene (https://www.ncbi.nlm.nih.gov/gap/, ID: phs000179. v6. p2).

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YZ and KK developed the statistical method and analysis design. DG, FX, FB-K contributed to method development and improvement. RB provided access to data sets and helped evaluate clinical findings. YZ processed the data and performed the statistical analysis. All authors contributed to writing the final manuscript.

FUNDING

This work was supported by NIH/NHLBI R01 HL152735 (KK, RB, FB-K, and YZ), R01 HL137995 (KK, RB, and FB-K) and R01HL125583 (KK and YZ). Additional support for this work was provided by the National Institutes of Health, National Heart, Lung, and Blood Institute, through the BioData Catalyst program (award 1OT3HL142479-01, 1OT3HL142478-01, 1OT3HL142481-01, 1OT3HL142480-01, 1OT3HL147154-01). Any opinions expressed in this document are those of the author(s) and do not necessarily reflect the views of NHLBI, individual BDCatalyst team members, or affiliated organizations and institutions. The project described was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. 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.

ACKNOWLEDGMENTS

The authors wish to acknowledge the contributions of the consortium working on the development of the NHLBI BioData Catalyst ecosystem.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.760299/full#supplementary-material

REFERENCES

Alexa, M., and Adamson, A. (2009). Interpolatory point Set Surfaces-Convexity and Hermite Data. *ACM Trans. Graph.* 28, 1–10. doi:10.1145/1516522.1516531

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Ashtiani, M., Salehzadeh-Yazdi, A., Razaghi-Moghadam, Z., Hennig, H., Wolkenhauer, O., Mirzaie, M., et al. (2018). A Systematic Survey of Centrality Measures for Protein-Protein Interaction Networks. *BMC Syst. Biol.* 12, 80–17. doi:10.1186/s12918-018-0598-2

Bradford, E., Jacobson, S., Varasteh, J., Comellas, A. P., Woodruff, P., O'Neal, W., et al. (2017). The Value of Blood Cytokines and Chemokines in Assessing Copd. *Respir. Res.* 18, 1–11. doi:10.1186/s12931-017-0662-2

Broido, A. D., and Clauset, A. (2019). Scale-free Networks Are Rare. *Nat. Commun.* 10, 1–10. doi:10.1038/s41467-019-08746-5

Candia, J., Cheung, F., Kotliarov, Y., Fantoni, G., Sellers, B., Griesman, T., et al. (2017). Assessment of Variability in the Somascan Assay. *Sci. Rep.* 7, 1–13. doi:10.1038/s41598-017-14755-5

DiLeo, M. V., Strahan, G. D., den Bakker, M., and Hoekenga, O. A. (2011). Weighted Correlation Network Analysis (Wgcna) Applied to the Tomato Fruit Metabolome. *PLoS One* 6, e26683. doi:10.1371/journal.pone.0026683

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse Graphical Models for Exploring Gene Expression Data. *J. Multivariate Anal.* 90, 196–212. doi:10.1016/j.jmva.2004.02.009

Dong, X., Castro, L. E., and Shaikh, N. I. (2016). Fastnet: An R Package for Fast Simulation and Analysis of Large-Scale Social Networks. *J. Stat. Softw.* Forthcoming.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., et al. (2005). Biomart and Bioconductor: a Powerful Link between Biological Databases and Microarray Data Analysis. *Bioinformatics* 21, 3439–3440. doi:10.1093/bioinformatics/bti525

Eapen, M. S., and Sohal, S. S. (2020). WNT/β-catenin Pathway: A Novel Therapeutic Target for Attenuating Airway Remodelling and EMT in COPD. *EBioMedicine* 62, 103095. doi:10.1016/j.ebiom.2020.103095

Fattahi, S., and Sojoudi, S. (2019). Graphical Lasso and Thresholding: Equivalence and Closed-form Solutions. *J. machine Learn. Res.* 20, 1–44.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics* 9, 432–441. doi:10.1093/biostatistics/kxm045

Hastie, T., Tibshirani, R., and Wainwright, M. (2019). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC.

Henrot, P., Prevel, R., Berger, P., and Dupin, I. (2019). Chemokines in Copd: from Implication to Therapeutic Use. *Ijms* 20, 2785. doi:10.3390/ijms20112785

Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., et al. (2017). Architecture of the Human Interactome Defines Protein Communities and Disease Networks. *Nature* 545, 505–509. doi:10.1038/nature22366

Kneidinger, N., Yildirim, A. Ö., Callegari, J., Takenaka, S., Stein, M. M., Dumitrascu, R., et al. (2011). Activation of the WNT/β-Catenin Pathway Attenuates Experimental Emphysema. *Am. J. Respir. Crit. Care Med.* 183, 723–733. doi:10.1164/rccm.200910-1560oc

Kuchaiev, O., Rašajski, M., Higham, D. J., and Pržulj, N. (2009). Geometric De-noising of Protein-Protein Interaction Networks. *Plos Comput. Biol.* 5, e1000454. doi:10.1371/journal.pcbi.1000454

Langfelder, P., Cantle, J. P., Chatzopoulou, D., Wang, N., Gao, F., Al-Ramahi, I., et al. (2016). Integrated Genomics and Proteomics Define Huntingtin Cag Length-dependent Networks in Mice. *Nat. Neurosci.* 19, 623–633. doi:10.1038/nn.4256

Langfelder, P., and Horvath, S. (2008). Wgcna: an R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* 9, 559. doi:10.1186/1471-2105-9-559

Li, F., Zhu, F., Ling, X., and Liu, Q. (2020). Protein Interaction Network Reconstruction through Ensemble Deep Learning with Attention Mechanism. *Front. Bioeng. Biotechnol.* 8, 390. doi:10.3389/fbioe.2020.00390

Li, Y., and Jackson, S. A. (2015). Gene Network Reconstruction by Integration of Prior Biological Knowledge. *G3: Genes, Genomes, Genet.* 5, 1075–1079. doi:10.1534/g3.115.018127

Liang, F., Song, Q., and Qiu, P. (2015). An Equivalent Measure of Partial Correlation Coefficients for High-Dimensional Gaussian Graphical Models. *J. Am. Stat. Assoc.* 110, 1248–1265. doi:10.1080/01621459.2015.1012391

Ma, J., Shojaie, A., and Michailidis, G. (2016). Network-based Pathway Enrichment Analysis with Incomplete Network Information. *Bioinformatics* 32, 3165–3174. doi:10.1093/bioinformatics/btw410

Mamdani, M., Williamson, V., McMichael, G. O., Blevins, T., Aliev, F., Adkins, A., et al. (2015). Integrating Mrna and Mirna Weighted Gene Co-expression Networks with Eqtls in the Nucleus Accumbens of Subjects with Alcohol Dependence. *PLoS One* 10, e0137671. doi:10.1371/journal.pone.0137671

Mastej, E., Gillenwater, L., Zhuang, Y., Pratte, K. A., Bowler, R. P., and Kechris, K. (2020). Identifying Protein-Metabolite Networks Associated with COPD Phenotypes. *Metabolites* 10, 124. doi:10.3390/metabo10040124

Meinshausen, N., and Bühlmann, P. (2006). High-dimensional Graphs and Variable Selection with the Lasso. *Ann. Stat.* 34, 1436–1462. doi:10.1214/009053606000000281

Nelson, W. J., and Nusse, R. (2004). Convergence of Wnt, SS-Catenin, and Cadherin Pathways. *Science* 303, 1483–1487. doi:10.1126/science.1094291

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ragland, M. F., Benway, C. J., Lutz, S. M., Bowler, R. P., Hecker, J., Hokanson, J. E., et al. (2019). Genetic Advances in Chronic Obstructive Pulmonary Disease. Insights from Copdgene. *Am. J. Respir. Crit. Care Med.* 200, 677–690. doi:10.1164/rccm.201808-1455so

RStudio Team (2019). *RStudioIntegrated Development Environment for R*. Boston, MA: RStudio, Inc.

Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A Comprehensive Evaluation of Module Detection Methods for Gene Expression Data. *Nat. Commun.* 9, 1090. doi:10.1038/s41467-018-03424-4

Seyyedsalehi, S. F., Soleymani, M., Rabiee, H. R., and Mofrad, M. R. K. (2021). Pfp-wgan: Protein Function Prediction by Discovering Gene Ontology Term Correlations with Generative Adversarial Networks. *Plos one* 16, e0244430. doi:10.1371/journal.pone.0244430

Shirasaki, D. I., Greiner, E. R., Al-Ramahi, I., Gray, M., Boontheung, P., Geschwind, D. H., et al. (2012). Network Organization of the Huntingtin Proteomic

Interactome in Mammalian Brain. *Neuron* 75, 41–57. doi:10.1016/j.neuron.2012.05.024

Silverbush, D., and Sharan, R. (2019). A Systematic Approach to orient the Human Protein-Protein Interaction Network. *Nat. Commun.* 10, 3015–3019. doi:10.1038/s41467-019-10887-6

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams, R. M., Jr (1949). The American Soldier: Adjustment during Army Life. *studies Soc. Psychol. World war ii* 1.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2016). The String Database in 2017: Quality-Controlled Protein–Protein Association Networks, Made Broadly Accessible. *Nucleic Acids Res.*, gkw937.

Wang, M.-G., Ou-Yang, L., Yan, H., and Zhang, X.-F. (2021). Inferring Gene Co-expression Networks by Incorporating Prior Protein-Protein Interaction Networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi:10.1109/tcbb.2021.3103407

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *J. Stat. Softw.* 40, 1–29. doi:10.18637/jss.v040.i01

Xu, B., Liu, Y., Lin, C., Dong, J., Liu, X., and He, Z. (2018). Reconstruction of the Protein-Protein Interaction Network for Protein Complexes Identification by Walking on the Protein Pair Fingerprints Similarity Network. *Front. Genet.* 9, 272. doi:10.3389/fgene.2018.00272

Zhang, B., Tian, Y., and Zhang, Z. (2014). Network Biology in Medicine and beyond. *Circ. Cardiovasc. Genet.* 7, 536–547. doi:10.1161/circgenetics.113.000123

Zhang, G., He, P., Tan, H., Budhu, A., Gaedcke, J., Ghadimi, B. M., et al. (2013). Integration of Metabolomics and Transcriptomics Revealed a Fatty Acid Network Exerting Growth Inhibitory Effects in Human Pancreatic Cancer. *Clin. Cancer Res.* 19, 4983–4993. doi:10.1158/1078-0432.ccr-13-0209

Zhang, R., Fattahi, S., and Sojoudi, S. (2018). Large-scale Sparse Inverse Covariance Estimation via Thresholding and max-det Matrix Completion. *Int. Conf. Machine Learn. (Pmlr)*, 5766–5775.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The Huge Package for High-Dimensional Undirected Graph Estimation in R. *J. Mach Learn. Res.* 13, 1059–1062.

Zuo, Y., Cui, Y., Yu, G., Li, R., and Ressom, H. W. (2017). Incorporating Prior Biological Knowledge for Network-Based Differential Gene Expression Analysis Using Differentially Weighted Graphical Lasso. *BMC bioinformatics* 18, 99–14. doi:10.1186/s12859-017-1515-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.