ShinyDegSEM: an interactive application for pathway perturbation analysis in gene expression studies via structural equation modeling (#118650)

First submission

Guidance from your Editor

Please submit by **2 Jul 2025** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for guidance.

Custom checks
Make sure you include the custom checks shown below, in your review.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

All review materials are strictly confidential. Uploading the manuscript to third-party tools such as Large Language Models is not allowed.
If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files

Download and review all files from the materials page.

9 Figure file(s)
1 Other file(s)

Custom checks

DNA data checks

- Have you checked the authors data deposition statement?
- Can you access the deposited data?
- Has the data been deposited correctly?
- Is the deposition information noted in the manuscript?

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- **Impact and novelty is not assessed.** Meaningful replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

ShinyDegSEM: an interactive application for pathway perturbation analysis in gene expression studies via structural equation modeling

Zhehan Jiang [Corresp., 1, 2], **Jihong Zhang** [3], **Yuanfang Liu** [1], **Jinying Ouyang** [4], **Linlin Sun** [Corresp., 5], **Hao Guo** [6]

[1] Institute of Medical Education, Peking University Health Science Center, Beijing, China

[2] National Center for Health Professions Education Development, Peking University, Beijing, China

[3] Department of Counseling, Leadership, and Research Methods, University of Arkansas at Fayetteville, Fayetteville, Arkansas, United States

[4] School of Public Health, Peking University, Beijing, China

[5] Department of Neurobiology, School of Basic Medical Sciences, Peking University, Beijing, China

[6] School of Health Humanities, Peking University, Beijing, China

Corresponding Authors: Zhehan Jiang, Linlin Sun
Email address: jiangzhehan@bjmu.edu.cn, linlin.sun@pku.edu.cn

Background : Researchers in biology and bioinformatics are increasingly interested in unraveling the complex mechanisms underlying phenotypic variations. A key challenge lies in identifying perturbed biological pathways and understanding how these perturbations propagate through intricate gene regulatory networks.

Results : To address this challenge, we developed ShinyDegSEM, an interactive R Shiny application that leverages structural equation modeling (SEM) to facilitate pathway perturbation analysis in gene expression studies. ShinyDegSEM streamlines identifying differentially expressed genes (DEGs), generating pathway models based on biological knowledge, and evaluating these models to uncover perturbed pathway modules. This article is a tutorial to navigate users through the analysis workflow with detailed explanations and examples. This feature ensures that even novice researchers can quickly grasp the concepts and apply the tool to their datasets.

Conclusions : The application integrates multiple steps, including DEG detection using significance analysis of microarray, perturbed pathway analysis with signaling pathway impact analysis, and SEM-based model refinement and comparison between experimental and control groups. The interactive interface of ShinyDegSEM allows researchers to easily upload their gene expression data, select appropriate criteria for DEG detection and pathway analysis, and visualize the results in intuitive graphs and tables. The tool provides insights into deregulated genes and modified gene-gene relationships within perturbed pathways.

ShinyDegSEM: An Interactive Application for Pathway Perturbation Analysis in Gene Expression Studies via Structural Equation Modeling

Zhehan Jiang*[1, 2], Jihong Zhang[3], Yuanfang Liu[1], Jinying Ouyang[4], Linlin Sun*[5], Hao Guo[6]

[1]Institute of Medical Education, Health Science Center, Peking University, Beijing, China
[2]National Center for Health Professions Education Development, Peking University, Beijing, China
[3]Department of Counseling, Leadership, and Research Methods, University of Arkansas, Fayetteville, AR, USA
[4]School of Public Health, Peking University, Beijing, China
[5]Department of Neurobiology, School of Basic Medical Sciences, Key Laboratory for Neuroscience, Ministry of Education/National Health Commission of China, Neuroscience Research Institute, Peking University, Beijing, China
[6]School of Health Humanities, Peking University, Beijing, China

Corresponding Author:
Zhehan Jiang*, jiangzhehan@bjmu.edu.cn, [1]Institute of Medical Education, Health Science Center, Peking University, 38 Xueyuan Road, Haidian District, Beijing, 100083, China
Linlin Sun*, linlin.sun@pku.edu.cn, [5]Department of Neurobiology, School of Basic Medical Sciences, Peking University, 38 Xueyuan Road, Haidian District, Beijing, 100083, China

Author Note

Zhehan Jiang, https://orcid.org/0000-0002-1376-9439

Abstract

Background: Researchers in biology and bioinformatics are increasingly interested in unraveling the complex mechanisms underlying phenotypic variations. A key challenge lies in identifying perturbed biological pathways and understanding how these perturbations propagate through intricate gene regulatory networks.

Results: To address this challenge, we developed ShinyDegSEM, an interactive R Shiny application that leverages structural equation modeling (SEM) to facilitate pathway perturbation analysis in gene expression studies. ShinyDegSEM streamlines identifying differentially expressed genes (DEGs), generating pathway models based on biological knowledge, and

evaluating these models to uncover perturbed pathway modules. This article is a tutorial to navigate users through the analysis workflow with detailed explanations and examples. This feature ensures that even novice researchers can quickly grasp the concepts and apply the tool to their datasets.

Conclusions: The application integrates multiple steps, including DEG detection using significance analysis of microarray, perturbed pathway analysis with signaling pathway impact analysis, and SEM-based model refinement and comparison between experimental and control groups. The interactive interface of ShinyDegSEM allows researchers to easily upload their gene expression data, select appropriate criteria for DEG detection and pathway analysis, and visualize the results in intuitive graphs and tables. The tool provides insights into deregulated genes and modified gene-gene relationships within perturbed pathways.

Keywords: Structural equation modeling, Shiny, differentially expressed genes, significance analysis of microarray, perturbed pathway analysis

Introduction

Biological networks have been popular in recent years (Scardoni, Petterlini, & Laudanna, 2009; Chin et al., 2014; Omony, 2014; Liu et al., 2020; Wang et al., 2021), stemming from recognizing that biological systems are inherently complex, with numerous interconnected components operating in concert to maintain cellular homeostasis and adapt to environmental stimuli (Goldstein, 2019; Liu et al., 2020). Network biology employs graph-theoretic approaches to represent biological molecules, such as genes, proteins, and metabolites, as nodes in networks, where edges represent the interactions among these components (Alm & Arkin, 2003; Albert, 2007). This paradigm shift has not only enhanced our understanding of biological processes but has also provided a new platform for various applications of analytical frameworks and tools such as machine learning (Muzio, O'Bray, & Borgwardt, 2021], statistical modeling (Lee & Tzou, 2009; Oates & Mukherjee, 2012; Epskamp, Rhemtulla, & Borsboom, 2017; Valdeolivas et al., 2018], and pathway analysis (Isci et al., 2011; Rodchenkov et al., 2019). These tools enable

researchers to unravel the complexities of biological networks, predict behaviors, and identify potential intervention points (Lee & Tzou, 2009).

Structural Equation Modeling (SEM) Framework

Among the analytical frameworks, structural equation modeling (SEM; Kline, 2023) stands out due to its unique capability to handle complex relationships in measurement models and between latent variables. SEM is a statistical method that allows researchers to test complex theories by examining the relationships between multiple variables (Anderson & Gerbing, 1988; Ullman & Bentler, 2013; Kline, 2023). Specifically, SEM combines factor analysis, multiple regression, and path analysis. SEM allows researchers to build and test models demonstrating how different variables are connected and influence each other. The mathematical expressions and notations (Pepe & Grassi, 2014; Kline, 2023) are in the Supplementary Materials [SM].

SEM in Biological Studies

Conventional SEM uses measurement and structural models to examine the relationships between observed and latent variables. The SEM method in this paper focuses on relationships between observed variables (e.g., gene expression) while accounting for unobserved factors and using path diagrams to represent the models visually. This approach is well-suited for analyzing gene expression data and uncovering the underlying mechanisms of biological pathways (Liu, de la Fuente, & Hoeschele, 2008; Neto et al. 2010; Cai, Bazerque, & Giannakis; 2013; Romdhani et al., 2015; Wang, Lu, & Miao, 2016; Igolkina et al., 2018).

Researchers have applied SEM in biological and health studies, especially with biological network techniques (Liu, de la Fuente, & Hoeschele, 2008; Neto et al. 2010; Cai, Bazerque, & Giannakis; 2013; Romdhani et al., 2015; Wang, Lu, & Miao, 2016; Igolkina et al., 2018). For example, Liu, de la Fuente, and Hoeschele (2008) examined using linear SEM to identify sparse

networks to validate genetic network inference through simulation and application to real genetic datasets. The researchers found that SEM was promising for accurately identifying different network edges.–Neto et al. (2010) developed a quantitative trait loci (QTL)-driven phenotype network method called QTLnet to jointly infer causal networks and genetic architecture of sets of phenotypes. They validated this framework through simulations and real data analysis. The QTLnet method incorporates SEM features, using graphical models to illustrate causal relationships between genes and phenotypes and within phenotypes. Likewise, Romdhani et al. (2015) proposed a test to analyze the relationships between genetic variants of gene candidates and correlated traits. They applied this method to real data to examine associations between genes and cardiovascular disease-related traits. Their approach leverages SEM to model complex relationships, providing a robust framework for understanding how genetic variants influence multiple correlated traits simultaneously. Cai, Bazerque, and Giannakis (2013) contributed to developing a sparsity-aware maximum likelihood (SML) algorithm for using sparse structural equation models to model gene regulatory networks. Similarly, Wang, Lu, and Miao (2016) proposed an efficient structural identifiability analysis algorithm for static linear SEM to help examine graphical models of biological networks with latent variables. In addition, Igolkina et al. (2018) examined the SEM to examine gene expression pathway coefficient differences between gene network data from 144 schizophrenia (SCZ) patients and 111 control individuals (without SCZ themselves and no family history of SCZ). They found that the SEM can identify the altered relationships between gene interactions at different statistical significance levels (e.g., $p < .01$). Moreover, various R packages that can apply SEM in biology studies have been developed, such as GenomicSEM (Grotzinger et al., 2019), GW-SEM (Pritikin et al., 2021), SEMgraph (Grassi, Palluzzi, & Tarantino, 2022), and SEMdeep (Grassi & Tarantino, 2025).

Literature shows that although SEM has shown great promise in the biological and health field, its full potential in applied research remains untapped, mainly due to the relatively low collaboration between SEM methodologists and biological researchers. This gap can be attributed to several factors, including the technical complexity of SEM, the distinct backgrounds and terminologies used by researchers from different fields, and the limited exposure of biological researchers to SEM methodologies.

Advantages of SEM in Pathway Analysis

Hypothesized Causal relationships via SEM. Structural equation modeling enhances pathway analysis by addressing the critical limitations of traditional correlation-based methods. Unlike approaches that only identify correlated relationships, SEM evaluates hypothesized causal structures, modeling both direct and indirect regulatory influences (e.g., gene $A \rightarrow$ gene $B \rightarrow$ gene C). This allows researchers to test mechanistic explanations for observed gene expression changes, such as cascading effects or feedback loops.

In genetic pathway analysis, SEM uses directed edges ($\rightarrow$) to represent regulatory relationships (e.g., transcription factor binding) and bidirected edges ($\leftrightarrow$) to account for unmeasured confounders (e.g., environmental factors or latent proteins) that jointly affect multiple genes. While initial pathway models (e.g., from the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa et al., 2002; 2004; 2017) are simplified abstractions of biological networks, SEM provides a framework for validating and iteratively refining these models using empirical data. For example, SEM can test whether adding a hypothesized interaction (e.g., a post-translational modifier) improves model fit, thereby bridging gaps between static pathway maps and dynamic biological reality.

Comparative Analysis of Regulatory Network Dynamics Across Groups in SEM. Multiple group analysis in SEM enables comparative evaluation of regulatory interactions involving pre-identified differentially expressed genes (DEGs). By testing invariance in path coefficients and network structures across groups, SEM reveals context-specific rewiring of regulatory relationships, such as strengthened or weakened causal effects between DEGs in disease conditions.

Structural equation modeling extends beyond transcriptomic correlations by testing hypothesized *directed relationships* between genes, even when their RNA levels lack strong pairwise correlations. By modeling pathways (e.g., Gene $B_1 \rightarrow$ Gene B_2 via latent mediators), SEM can infer regulatory effects masked in simple correlation analyses. While SEM cannot directly measure post-translational modifications (PTMs) or dynamic cascades, it can incorporate latent variables to approximate such mechanisms if supported by auxiliary data. The strength of SEM is evaluating how well a predefined network structure (including indirect or hierarchical relationships) explains observed gene expression patterns, revealing path coefficients that reflect hypothesized regulatory influences.

Multiple Data Sources and Comprehensive Analysis via SEM. The SEM pipeline integrates multi-modal data sources, such as gene expression (microarrays), curated pathway topologies (KEGG), and protein-protein interaction networks (e.g., STRING database (Szklarczyk et al., 2015), to construct biologically plausible regulatory models. This integration enhances robustness by cross-validating hypotheses against orthogonal data types. For example, in a study of frontotemporal lobar degeneration with ubiquitinated inclusions (FTLD-U), SEM analysis of the glutamatergic synapse pathway identified *PSD-95* as a hub gene and revealed altered regulatory relationships involving *SHANK2* and glutamate receptors under progranulin

mutation. The model further suggested context-specific activation or inhibition of connections (e.g., strengthened *PSD-95→SHANK2* interactions in mutant conditions). Similarly, in multiple sclerosis (MS), SEM highlighted dysregulated genes (*ARF6*, *CRKL*, and *PIP5K1C*) within the Fc gamma R-mediated phagocytosis pathway. These findings align with prior studies implicating phagocytic dysfunction in MS pathogenesis (Pepe & Grasssi, 2014). SEM disentangles direct regulatory effects from indirect associations, offering mechanistic insights into neurodegenerative processes by combining pathway and interaction data.

Model Assessment via SEM

Structural equation modeling evaluates model fit using statistical tests and indices (Kline, 2023) such as the chi-square test, root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). Biological evidence from databases like STRING can be incorporated to validate and include known interactions. A well-fitting model is typically indicated by a non-significant χ^2 test p value (though this test is sensitive to sample size), RMSEA $\leq$ .06 (Hu & Bentler, 1999), and SRMR $\leq$ .05 or .10 (West, Taylor, & Wu, 2012; Grotzinger et al., 2021) for adequate or good fit, respectively. These indices evaluate how closely the proposed model aligns with the observed data. To refine the model, modification indices (MI) estimate the potential improvement in fit (quantified by the expected decrease in χ^2) if a constrained parameter (e.g., a path or covariance) is freely estimated (Kline, 2023). Statistical indices, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC), can also be used for SEM model comparisons and selections (Grassi, Palluzzi, & Tarantino, 2022; Kline, 2023). However, modifications are only justified when they align with substantive theory, domain knowledge, or plausible causal mechanisms. Nonsignificant paths may be removed to enhance parsimony if such changes do not compromise theoretical

expectations. Iterative adjustments balancing statistical guidance and substantive rationale are critical to avoid overfitting and support generalizability. See the SM for detailed explanations.

Validation for SEM Results

Comparative analyses based on other methods, such as simple differential expression or correlation-based network analysis, could be conducted to validate SEM results. Benchmark datasets with known ground truth can validate the accuracy and reliability of SEM. Experimental validation of key SEM findings through assays, such as testing the impact of perturbing specific genes or connections, would confirm predicted changes in gene activity. Evaluating the predictive accuracy of SEM models would also strengthen their assessment (e.g., predicting disease progression or treatment response). Lastly, developing more intuitive visualizations that highlight key findings and show network differences between experimental conditions would enhance the understanding and communication of SEM results. We aim to contribute to the use of SEM for pathway analysis by developing a Shiny application (app).

Interactive biological web applications hosted on Shiny servers have been published more recently due to the increasing awareness among researchers of their methodological advances and practical ease. For example, Jia et al. (2022) systematically reviewed biological web applications built with R or Shiny and their basic and advanced features. However, applications specifically designated to handle SEM are less commonly seen; one of the most well-known is *power4SEM*, which is used for power calculations (Jak et al., 2021). Our article serves as a tutorial brief to address this gap by developing an R Shiny software application called *ShinyDegSEM*, which connects bioinformatics with SEM. Although researchers had elegantly applied SEM in gene expression and pathway analysis data (Pepe & Grassi, 2014), to our knowledge, this is the first tool that adopts SEM to investigate perturbed pathway modules

derived from gene expression data. We aim to demystify SEM for biologists by combining a series of analyses with a user-friendly interface, allowing users to execute various computations and functions through a point-and-click interface.

Materials & Methods

Understanding phenotypic variation requires studying perturbations in complex intracellular networks rather than focusing solely on single-gene dysregulation. High-throughput gene expression data enables investigation of changes in gene expression profiles across different conditions. A comprehensive analysis of pathway perturbation via SEM integrates two key components: genome-wide association studies (GWAS) with pathway extensions to identify genetic associations, and SEM-based modeling, evaluation, and refinement to quantify network-level effects. Building on the foundational workflow of Pepe and Grassi (2014), which spans from identifying differentially expressed genes (DEGs) to validating and interpreting perturbed pathway models, we enhance this approach by incorporating recent advancements in GWAS and SEM into *ShinyDegSEM*. Our implementation offers improved flexibility, usability, and analytical precision for pathway-centric studies. See the SM for detailed gene study terminologies and methodologies. The following steps are needed to apply *ShinyDegSEM* for conducting pathway analyses using SEM.

Step 1. In the initial step, users can collect and prepare data for analysis. Three primary genomic data types can be included: (1) gene expression data (including microarray-based transcript abundance quantification (Schena et al., 1995) and RNA-sequencing (RNA-seq) for genome-wide expression profiling with single-nucleotide resolution (Wang, Gerstein, & Snyder, 2009; AlJanahi, Danielsen, & Dunbar, 2018), (2) genomic variation data (e.g., whole-genome or exome sequencing data) capturing nucleotide-level polymorphisms and structural variants DePristo et al., 2011 4), and (3) quantitative real-time PCR (qRT-PCR) data for precise

expression validation (Hendriks-Balk, Michel, & Alewijnse, 2007). Public repositories such as NCBI's Gene Expression Omnibus (GEO; National Center for Biotechnology Information, 2024) and KEGG (Kanehisa et al., 2002; 2004; 2017) may serve as additional data sources. Prepared data (e.g., .txt or .csv formats) are imported for follow-up analysis.

Step 2. In step 2, we identify DEGs to detect significant gene expression level changes between two or more conditions. For microarray data, methods such as significance analysis of microarrays (SAM; Tusher, Tibshirani, & Chu, 2001) are commonly employed. RNA-seq data typically utilize count-based approaches, including normalization and statistical modeling via negative binomial distributions (Rapaport et al., 2013). Alternative strategies combine fold-change (FC) thresholds with non-stringent p-value cutoffs to balance sensitivity and specificity (Shi et al., 2008). Emerging machine learning approaches, including deep learning frameworks, offer additional tools for DEG detection (Tasaki et al., 2020), especially in complex datasets.

Step 3. In step 3, we identify perturbed pathways. Biologically perturbed pathways are identified as functional modules enriched with DEGs, which are indicative of potential disease-associated dysregulation (Pham et al., 2016). Established computational approaches include: (1) enrichment analysis (e.g., over-representation or gene set enrichment; Rahmati et al., 2017), (2) signaling pathway impact analysis (SPIA) that combines topological and statistical metrics (Tarca et al., 2009), and (3) integration with curated pathway databases (e.g., KEGG; Kanehisa et al., 2017). These pathways are subsequently modeled as directed graphs or gene networks, where nodes represent molecular components and edges depict functional interactions, enabling the visualization and topological analysis of perturbed systems (Goh et al., 2007).

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway graphs were converted into directed graphs for SEM analysis. In this representation, nodes represent genes derived from

251 microarray, RNA-seq data, or Protein Information Resource (PIR) superfamilies, which are
252 clusters of evolutionarily related proteins with shared functions. Edges represent directed
253 biochemical interactions between nodes, which are categorized into two primary types:
254 molecular interactions, including protein-protein binding and enzymatic reactions, and regulatory
255 relationships such as transcriptional activation or suppression (Pepe & Grassi, 2014; Grassi &
256 Tarantino, 2022). The directed graph structure encodes the causal dependencies between
257 molecular components, allowing SEM to quantify pathway-wide dysregulation across
258 comparison groups (or between diseased and normal controls). Main advantages of this approach
259 include: (1) maintaining biological interpretability through preservation of established pathway
260 architectures, (2) enabling quantitative assessment of both magnitude and directionality of
261 molecular interactions, and (3) supporting investigation of condition-specific pathway
262 dysregulation through group comparisons.

263 Edges can be further classified into two types by directionality (Pepe & Grassi, 2014;
264 Grassi & Tarantino, 2022). Directed edges (→) indicate a direct influence of one gene on
265 another. The direction of the arrow indicates which gene regulates the other. For example, if
266 gene Y_1 has a directed edge pointing to gene Y_2 ($Y_1 \rightarrow Y_2$), it means that gene Y_1 is an upstream
267 regulator that directly affects the activity of gene Y_2. Bidirected edges (↔) represent covariances
268 between two genes attributable to unmeasured common causes (e.g., latent upstream regulators
269 or shared environmental factors) influencing both genes.

270 These edges in the directed graphs can have signs, which is a crucial aspect of how SEM is
271 used in this context. The strength and direction of the influence between two genes connected by
272 a directed edge (→) are quantified by path coefficients. These coefficients typically range from
273 − 1 to 1 if the data are standardized. Positive path coefficients indicate a net activation or

positive control, meaning that an increase in the activity of the upstream gene is expected to increase the activity of the downstream gene. Negative path coefficients represent net inhibition or negative control, meaning that an increase in the activity of the upstream gene is expected to decrease the activity of the downstream gene. On the other hand, bi-directed edges (↔) represents the covariance between two genes *i* and *j* due to unobserved factors, which is quantified by ψ_{ij}.

Structural equation modeling (SEM) employs linear regression equations in which path coefficients (β_{ij}) quantify both the strength and direction (i.e., positive or negative) of relationships between variables, serving as weights in the model equations. These signed coefficients are essential for determining the nature of gene-gene interactions within pathways, distinguishing between activating (positive), inhibitory (negative), or latent common-cause relationships. The framework enables comparison of these signed effects across experimental or different conditions through parameter contrasts between groups. During model refinement, MI, z-tests, and external biological databases (e.g., STRING; Szklarczyk et al., 2015) can inform the addition of directed or bidirected edges, with database-derived interaction signs directly informing path coefficient directions.

Steps 4 & 5. In step 4, we integrate curated pathway topologies (e.g., from KEGG) with data-driven network filtering using the algorithms proposed by Pepe and Grassi (2014). Canonical pathways are first represented as directed graphs and then pruned using partial correlations derived from gene expression data (e.g., Type I error rates < .05). In step 5, we apply SEM to the refined pathways, where differential analysis of path coefficients identifies statistically perturbed interactions across groups.

The SEM analyses were conducted using a suite of specialized R packages chosen for their complementary capabilities. Specifically, the lavaan package (Rosseel, 2012) served as the primary platform for model specification, parameter estimation, and goodness-of-fit assessment, which provides comprehensive functionality for diverse SEM applications. The lavaan package uses maximum likelihood (ML) estimation by default for continuous and complete data (Rosseel, 2012). For network visualization and manipulation of model components, including latent variable relationships, we employed the igraph package (Csardi & Nepusz, 2006), which facilitates intuitive graphical representation and interpretation of complex model structures. Additional analytical support was provided by the semTools package (Jorgensen et al., 2022), which offered essential utilities for data diagnostics, model comparison, and advanced statistical evaluations. This package enhanced our analytical workflow through its specialized functions, complementing core SEM procedures. A distinctive aspect of our approach involved integrating network analysis with SEM using the SEMgraph package (Grassi, Palluzzi, & Tarantino, 2022). This specialized tool enabled network-based model exploration, including fitting SEM models, pathway identification, detection of initial nodes, and robustness assessment through graph-theoretic and statistical metrics. Combining traditional SEM with network analysis, SEMgraph provided unique insights into model interconnectivity and dynamics.

After estimating the initial SEM model based on the perturbed pathways and gene connections from examined data (e.g., microarray), we obtain the strength of gene-gene connections, also known as path coefficients. The SEM models can be modified based on additional information, such as goodness-of-fit indices (e.g., RMSEA and SRMR), which are used to support the decision on whether to refine them iteratively. After the final structure of the model is determined, the remaining analysis focuses on assessing the appropriateness of group

comparison in SEM through invariance tests, examining whether the models differ significantly between groups (e.g., diseased vs. healthy), and identifying genes and gene-gene interactions that show significant differences in expression or regulation.

The remaining work is interpretation, where researchers should consider correlating the perturbed genes and connections with known biological processes and disease mechanisms. More importantly, like other biological analyses through statistical mining, it is critical to discuss the implications of the findings to understand the phenotype of interest.

Shiny Walkthrough

The Layout of the ShinyDegSEM Application

We first describe the ShinyDegSEM application (app) layout and then explain how to navigate the main screen. The initial screen of the app is displayed in Figure 1. The left panel (in gray) includes five steps for user navigation, while the right panel (in white) shows the outputs of each step. The five steps in the app are: (1) Step 1 Data Input, (2) Step 2 DEG Analysis, (3) Step 3 Enrichment Analysis, (4) Step 4 Network Analysis, and (5) Step 5 SEM Analysis. Specifically, users can click the "*Browse*" button under step 1 to upload a .txt or .csv data file and start the analysis.

Using the ShinyDegSEM Application

We used the same gene expression microarray data as Pepe and Grassi (2014) to demonstrate the app's use. The dataset pertains to MS. It includes genome-wide expression data from peripheral blood mononuclear cells (PBMC) of 12 MS patients and 15 healthy controls, contributed by Kemppinen et al. (2011). The dataset (Kemppinen et al., 2019) is stored in the Gene Expression Omnibus (GEO; National Center for Biotechnology Information [NCBI], 2024) database under ID GSE21942. Figure 2 shows a screen plot after uploading the dataset and the

patient and control group memberships from the label file (in step 1). The right panel shows a data preview with 21026 rows for genes and 27 columns for participant IDs. The application can be downloaded from https://osf.io/kw8zf/. When run, it follows the regular Shiny app execution method.

Procedures Before Conducting SEM

Let's proceed to steps 2 through 4 before performing SEM. In step 2 for DEG analysis, the default delta value (Tusher, Tibshirani, & Chu, 2001) in SAM analysis was set to 1 in the app, and users can adjust it according to their study. For example, we used 0.95 as Pepe and Grassi (2014) did. Step 3 involved enrichment analysis for identifying perturbed pathways. Step 4 is network analysis. Specifically, steps 2 and 3 analyses will be performed automatically after uploading the files. Users can click the "*Run Network Analysis*" button to initiate the analysis related to step 4. After a short wait (depending on the dataset size), results from steps 2 to 4 will gradually appear in the right panel. For example, clicking the "*DEGs acquirement*" button will display the output of the SAM analysis for DEG analysis (see Figure 3). Similarly, clicking the "*Enrichment analysis – Get pathway*" button will display the output of different perturbed pathways. By clicking the "*Network Analysis*", we can see model information and graphs for identified pathways, such as the "B cell receptor signaling", "Fc gamma R-mediated phagocytosis", and "Chagas disease" pathways. For example, Figure 4 shows the identified differentially expressed genes (DEGs) and non-DEGs (NDEGs) within the context of the Fc gamma R-mediated phagocytosis pathway, which is associated with autoimmune dysregulation and inflammation. The DEGs (*CRKL*, *ARF6*, *PLA2G4A*, and *ARPC4*) were identified (corresponding to Entrez IDs 1399, 382, 5321, and 10093, respectively) and matched those shown in Pepe and Grassi's (2014) study. Researchers can identify DEGs based on network

analysis results and understand the direction of gene interactions within a pathway. These findings can then be incorporated into SEM to investigate causal relationships among more genes and their interactions, providing insights into the regulatory mechanisms underlying the pathway.

Results

Running SEM Based on Network Analysis Results

In this paragraph, we describe how to work on SEM based on results from the network analysis and explore gene relationships in step 5 of the app. First, select one or more pathway(s) of interest from the panel, such as the "Chagas disease" pathway. Second, choose the SEM estimator, which is set to ML by default (Rosseel, 2012), and click "*Run Initial SEM*". The initial model output will appear in the right panel, displaying the model summary and model fit indices (see Figure 5), such as the SRMR (Kline, 2023) and the RMSEA (Anderson & Gerbing, 1988). The initial model related to the "Chagas disease " pathway did not fit the data well, with chi-square statistic $\chi^2(36) = 118.92$ and $p < .001$, RMSEA = .292, and SRMR = .308. In addition, we can modify the initial model by selecting an additional path and clicking "*Add the path and run the model again*". Adding six paths, we improved the model fit substantially (see Figure 6), with model 6 having chi-square statistic $\chi^2(30) = 36.71$ and $p = .186$, RMSEA = .091, and SRMR = .120.

Invariance Evaluation. Additionally, we can evaluate model invariance on edge and node concerning group membership in MS disease, which is like evaluating measurement invariance (Meredith, 1993; Vandenberg & Lance, 2000) on factor loadings and intercepts in a measurement model, respectively. We can evaluate the invariance based on model 6 by clicking "*Run Model Invariance*". First, the output (see Figure 7) showed model fit indices for the base

model (e.g., model 6, which did not consider group effects and assumed edge or node invariance), "group effects on edges" model (i.e., a two-group model which examines group effects on edges), and "group effects on node" model (i.e., a common model which examines group effects on nodes). For example, the RMSEA for the three models was .091, .105, and .285, respectively. Second, the analysis of variance (ANOVA) output comparing the "group effects on edge" model and the base model (see Figure 7) showed that edge invariance was not supported for model 6 between the two groups, indicating that the weights for the gene-gene interactions between the two groups are not equal. Third, the chi-square goodness of fit test on "group effects on node" model showed that node invariance was supported for this model ($p > .05$), meaning that the baseline gene expression levels for genes in the Chagas pathway were equal between the two groups, when all upstream regulators in the model were zero. If model invariance is violated, it is recommended that users run the SEM model related to group membership separately. By clicking "*Run Node Analysis*" and "*Run Edge Analysis*", we can evaluate the strengths and directions of gene-gene interactions and the impact of group membership (see part of the results in Figures 8 and 9).

Discussion

Model Validation and Causal Interpretation

When performing SEM, we should consider the accuracy of research and the validity of results. We can examine configural, edge, and node invariances before performing group comparisons. For example, suppose an initial model for a specific pathway does not fit the data well and cannot be improved by adding paths, we may explore the consistency of the SEM structure regarding gene interactions across groups. Specifically, we can examine whether the same relationship patterns (e.g., expressed genes or gene-gene interactions) hold across groups.

We note that edge invariance is stricter and requires equal strength of relationships (e.g., path coefficients or edge weights) across groups (Vandenberg & Lance, 2000). We can also examine node invariance to understand whether the baseline expression of genes is equal across groups. To control inflated Type I error from multiple comparisons, in the app, we used Brown's combined goodness of fit test (Moskvina et al., 2011; Cinar & Viechtbauer, 2022) implemented in SEMgraph to evaluate whether nested SEM models (e.g., base and "group effects on edge" models) fit the data equally well across groups. This approach complemented traditional likelihood ratio tests (LRTs) by aggregating evidence from multiple nested comparisons into a single statistical assessment (Cinar & Viechtbauer, 2022).

In addition, we can assess whether the coefficients of a specific pathway between groups differ statistically (e.g., MS and "Chagas disease") or investigate the relationships between different pathways. The evaluation enables researchers to examine gene regulation and expression differences between disease and control groups, facilitating our understanding of pathophysiology and treatment.

We clarify that the core purpose of SEM is to infer causal relationships rather than merely correlations. While correlation can indicate a relationship, SEM models how the activity of one gene directly influences the activity of another. The model uses path coefficients to quantify the strength and direction of these influences. A directed edge (A → B) indicates that gene A is an upstream regulator that directly affects the activity of gene B. The path coefficient quantifies the expected change in B's activity resulting from a change in A's activity. This influence does not have to be a direct and positive correlation at the transcript level.

Considerations for SEM Data in Biological Applications

The SEM in this paper can use data from gene expression microarrays and incorporate information from other sources to build and refine the initial model. The initial model in the demonstration uses curated biological pathways from databases such as KEGG, which provide information about various gene relationships, including regulatory relationships, protein-protein interactions, and metabolic pathways. According to Pepe and Grassi (2014), the model is further refined by identifying the shortest paths between differentially expressed genes (DEGs), which tailors a model specific to the observed changes in the gene expression data. Genes not differentially expressed but part of the shortest path are grouped into Protein Information Resource (PIR) superfamilies based on evolutionary relationships (2014), potentially highlighting standard functions or regulatory mechanisms.

Databases like STRING can provide information on known and predicted protein-protein interactions and functional associations. That information can be applied to inform model modification by adding new directed or bi-directed edges based on biological evidence.

Phosphorylation and Causal Inference in Structural Equation Modeling of Transcriptomic Data

Phosphorylation-mediated regulation presents a unique challenge in transcriptomic analyses, as the causal influence of gene *A* on gene *B*'s activity may not correlate strongly with their respective RNA levels. Structural equation modeling addresses this limitation by detecting consistent directional relationships between genes, even when their transcript abundances are uncoupled.

When gene *A* phosphorylates gene *B*'s protein product, increased transcription of *A* may lead to elevated *A* protein levels and subsequent changes in *B*'s functional state without necessarily altering *B*'s mRNA abundance. SEM captures this relationship through path coefficients that

reflect the net directional effect of *A* on *B*, incorporating both direct and indirect regulatory influences. The model evaluates whether systematic covariation exists between changes in *A*'s expression and downstream consequences on *B*'s activity, whether measured through functional assays or proxy gene expression patterns.

Model evaluation involves rigorous testing of proposed relationships. A poorly fitting edge (A → B) indicates a mismatch between the modeled relationship and the underlying biological mechanisms. Researchers may refine the model by adding or removing edges based on modification indices and biological plausibility. In addition, multiple-group analysis enables the comparison of model parameters across experimental conditions, revealing context-specific differences in regulatory strength and direction that may reflect condition-dependent phosphorylation states or other post-translational modifications.

This approach provides particular value in cases where post-translational regulation decouples protein activity from transcript abundance. By focusing on systematic patterns of covariation across multiple measurements, SEM can infer causal relationships that would be obscured by examining RNA correlations alone. However, the validity of such inferences depends critically on iterative model refinement and integration of complementary biological evidence.

Conclusion

This study presents ShinyDegSEM, an interactive application that implements a pathway-constrained SEM framework to analyze gene regulatory networks. By incorporating prior pathway knowledge (e.g., KEGG pathways) to guide model structure, the app enables researchers to estimate direct regulatory effects between observed gene expression levels and compare these relationships across experimental or clinical conditions. The tool's user-friendly interface democratizes advanced statistical modeling, eliminating the need for specialized coding

expertise and bridging the gap between computational biology and experimental, clinical, or population research.

We have currently demonstrated the use of ShinyDegSEM modeling to investigate gene interactions within individual pathways, providing a biologically interpretable framework for generating and validating hypotheses. Future study can expand the application's functionality to investigate longitudinal gene data or integrate multi-omics data, further enhancing its utility for dynamic changes or systems-level analyses. By combining accessibility with rigorous statistical methods, ShinyDegSEM has the potential to accelerate discoveries in gene regulatory research and foster interdisciplinary collaboration.

Acknowledgements

Not available.

References

Albert R. Network inference, analysis, and modeling in systems biology. *The Plant Cell* 2007; **19**(11):3327–3338. https://doi.org/10.1105/tpc.107.054700

AlJanahi AA, Danielsen M, Dunbar CE. (2018). An introduction to the analysis of single-cell RNA-sequencing data. *Mol Ther Methods Clin Dev* 2018; **10**, 189–196.

Alm E, Arkin AP. Biological networks. *Curr Opin in Struct Biol* 2003; **13**(2): 193–202.

506 Anderson JC, Gerbing DW. Structural equation modeling in practice: A review and
507 recommended two-step approach. *Psychol Bull* 1988; **103**(3): 411–423.
508 Cai X, Bazerque JA, Giannakis GB. Inference of gene regulatory networks with sparse structural
509 equation models exploiting genetic perturbations. *PLoS Comput Biol* 2013; **9**(5): 1–13.
510 Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and
511 sub-networks from complex interactome. *BMC Syst Biol* 2014; **8**: 1–7.
512 https://doi.org/10.1186/1752-0509-8-S4-S11
513 Cinar O, Viechtbauer W. A comparison of methods for gene-based testing that account for
514 linkage disequilibrium. *Front Genet.* 2022; **13**: 1–14.
515 Csardi G, Nepusz T. The igraph software. *Complex Syst* 2006; **1695**: 1-9.
516 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. (2011). A
517 framework for variation discovery and genotyping using next-generation DNA sequencing data.
518 *Nat Genet* 2011; **43**(5): 491–498.
519 Epskamp S, Rhemtulla M, Borsboom D. (2017). Generalized network psychometrics: combining
520 network and latent variable models. *Psychometrika* 2017; **82**: 904–927.
521 https://doi.org/10.1007/s11336-017-9557-x
522 Goh K, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. (2007). The human disease
523 network. *Proc Natl Acad Sci U S A* 2007; **104**(21): 8685–8690.
524 Goldstein DS. How does homeostasis happen? Integrative physiological, systems biological, and
525 evolutionary perspectives. *Am J Physiol Regul Integr Comp Physiol* 2019; **316**(4): R301-R317.
526 https://journals.physiology.org/doi/full/10.1152/ajpregu.00396.2018
527 Grassi M, Palluzzi F, Tarantino B. SEMgraph: an R package for causal network inference of
528 high-throughput data with structural equation models. *Bioinformatics* 2022; **38**(20): 4829-4830.

Grassi M, Tarantino B. SEMdeep: structural equation modeling with deep neural network and machine learning. R package version 1.0.0. 2025.

Grotzinger AD, Rhemtulla M, Akingbuwa WA, Ip HF, Adams MJ, Lewis CM, et al. Genomic SEM provides insights into the multivariate genetic architecture of complex traits. *Nature Hum Behav* 2021; **5**(12):1577-88. doi:10.1038/s41562-021-01110-y.

Grotzinger AD, Rhemtulla M, de Vlaming R, Ritchie SJ, Mallard TT, Hill WD, et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Na Hum Beha 2019*; **3**(5): 513-525.

Hendriks-Balk MC, Michel MC, Alewijnse AE. Pitfalls in the normalization of real-time polymerase chain reaction data. *Basic Research in Cardiology* 2007; **102**: 195–197.

Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Modeling* 1999; **6**(1):1–55.

Igolkina AA, Armoskus C, Newman JRB, Evgrafov OV, Mclntyre LM, Nuzhdin SV, et al. Analysis of gene expression variance in schizophrenia using structural equation modeling. *Front Mol Neurosci* 2018; **11**: 1–12.

Isci S, Ozturk C, Jones J, Out HH. Pathway analysis of high-throughput biological data within a Bayesian network framework. *Bioinformatics* 2011; **27**(12): 1667–1674.

Jak S, Jorgensen TD, Verdam MGE, Oort FJ, Elffers L. Analytical power calculations for structural equation modeling: a tutorial and Shiny app. *Behav Res Methods* 2021; **53**: 1385–1406

Jia L, Yao W, Jiang Y, Li Y, Wang Z, Li H, et al. Development of interactive biological web applications with R/Shiny. *Brief Bioinform* 2022; **23**(1):1–15.

Jorgensen TD, Pornprasertmanit S, Schoemann AM, Rosseel Y, Miller P, Quick C, et al. semTools: useful tools for structural equation modeling. R package version 0.5-6. 2022.

552 Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on
553 genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017; **45**: D353–D361.
554 Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GeomeNet. *Nucleic*
555 *Acids Res* 2002; **30**(1): 42–46.
556 Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering
557 the genome. *Nucleic Acids Res* 2004; **32**: D277–D280.
558 [Dataset]* Kemppinen AK, Kaprio J, Palotie A, Saarela J. 2019. Expression data from peripheral
559 blood mononuclear cells in multiple sclerosis patients and controls. National Center for
560 Biotechnology Information. Gene Expression Omnibus. Accessed September 30, 2024.
561 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21942
562 Kemppinen AK, Kaprio J, Palotie A, Saarela J. Systematic review of genome-wide expression
563 studies in multiple sclerosis. *BMJ Open* 2011; **1**(1): 1–10.
564 Kline RB. *Principles and practices of structural equation modeling*. 5th ed. New York: The
565 Guilford Press; 2023.
566 Lee WP, Tzou WS. Computational methods for discovering gene networks from expression data.
567 *Brief Bioinform* 2009; **10**(4): 408-423. https://academic.oup.com/bib/article/10/4/408/299503
568 Liu B, de la Fuente A, Hoeschele I. Gene network inference via structural equation modeling in
569 genetical genomics experiments. *Genetics* 2008; **178**(3): 1763–1776.
570 Liu C, Ma Y, Zhao J, Nussinov R, Zhang Y, Chen F, et al. (2020). Computational network
571 biology: data, models, and applications. *Phys Rep* 2020; **846**: 1–66.
572 Meredith W. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*
573 1993; **58**(4): 525–543.

574 Moskvina V, O'Dushlaine C, Purcell S, Craddock N, Holmans P, O'Donovan MC. Evaluation of
575 an approximation method for assessment of overall significance of multiple dependent tests in a
576 genome wide association study. *Genet Epidemiol.* 2011; **35**(8): 861–866.
577 Muzio G, O'Bray L, Borgwardt K. Biological network analysis with deep learning. *Brief*
578 *Bioinform* 2021; **22**(2): 1515–1530.
579 National Center for Biotechnology Information. Gene expression omnibus. Accessed September
580 29, 2024. https://www.ncbi.nlm.nih.gov/geo/query/acc.cg
581 Neto EC, Keller MP, Attie AD, Yandell BS. Causal graphical models in systems genetics: a
582 unified framework for joint inference of causal network and genetic architecture for correlated
583 phenotypes. *Ann Appl Stat* 2010; **4**(1): 320–339.
584 Oates CJ, Mukherjee S. Network inference and biological dynamics. *Ann Appl Stat* 2012; **6**(3):
585 1209–1235.
586 Omony J. Biological network inference: a review of methods and assessment of tools and
587 techniques. *Annu Res Rev Biol* 2014; **4**(4): 577–601.
588 Pepe D, Grassi M. Investigating perturbed pathway modules from gene expression data via
589 structural equation models. *BMC Bioinformatics* 2014; **15**(132): 1–15.
590 Pham LM, Carvalho L, Schaus S, Kolaczyk ED. Perturbation detection through modeling of
591 gene expression on a latent biological pathway network: a Bayesian hierarchical approach. *J Am*
592 *Stat Assoc* 2016; **111**(513): 73–92.
593 Pritikin JN, Neale MC, Prom-Wormley EC, Clark SL, Verhulst B. GW-SEM 2.0: efficient,
594 flexible, and accessible multivariate GWAS. *Behav Genet* 2021; **51**(3): 343–357.

595 Rahmati S, Abovsky M, Pastrello C, Jurisica I. PathDIP: an annotated resource for known and
596 predicted human gene-pathway associations and pathway enrichment analysis. *Nucleic Acids Res*
597 2017; **45**: D419–D426.
598 Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of
599 differential gene expression analysis methods for RNA-seq data. *Genome Biol* 2013; **14**: 1–13.
600 Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, et al. Pathway Commons 2019
601 update: integration, analysis, and exploration of pathway data. *Nucleic Acids Res* 2019; **48**:
602 D489–D497.
603 Romdhani H, Hwang H, Paradis G, Roy-Gagnon M-H, Labbe A. Pathway-based association
604 study of multiple candidate genes and multiple traits using structural equation models. *Genet*
605 *Epidemiol* 2015; **39**(2): 101–113.
606 Rosseel Y. lavaan: an R package for structural equation modeling. *J Stat Softw* 2012; **48**: 1–36.
607 Scardoni G, Petterlini M, Laudanna C. Analyzing biological network parameters with
608 CentiScaPe. *Bioinformatics* 2009; **25**(21): 2857–2859.
609 Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression
610 patterns with a complementary DNA microarray. *Science* 1995; **270**(5235): 467–470.
611 Shi L, Jones WD, Jensen RV, Harris SC, Perkins RG, Goodsaid FM, et al. The balance of
612 reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray
613 studies. *BMC Bioinformatics* 2008; **9**(Suppl 9): 1–19.
614 Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING
615 v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*
616 2015; **43**(D1): D447–D452.

Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J, et al. A novel signaling pathway impact analysis. *Bioinformatics* 2009; **25**(1): 75–82.

Tasaki S, Gaiteri, C, Mostafavi S, Wang Y. Deep learning decodes the principles of differential gene expression. *Nat Mach Intell* 2020; **2**: 376–386.

Tusher VG, Tibshirani R, Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. **Proc Nati Acad Sci U S A** 2001; *98*(9): 5116-5121.

Ullman JB, Bentler PM. Structural equation modeling. In: Schinka JA, Velicer WF, Weiner IB, editors. *Handbook of psychology: Research methods in psychology*. 2nd ed. Hoboken (NJ): John Wiley & Sons, Inc.; **2013**. p. 661–690.

Valdeolivas A, Tichit L, Navarro C, Perrin S, Odelin G, Levy N, et al. (2018). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* 2018; **35**(3): 497–505. https://academic.oup.com/bioinformatics/article/35/3/497/5055408

Vandenberg RJ, Lance CE. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organ Res Methods* 2000; **3**(1): 4–70.

Wang Y, Lu N, Miao H. Structural identifiability of cyclic graphical models of biological networks with latent variables. *BMC Syst Biol* 2016; **10**(41): 497–505. https://doi.org/10.1186/s12918-016-0287-y

Wang YXR, Li L, Li JJ, Huang H. Network modeling in biology: statistical methods for gene and brain networks. *Stat Sci.* 2021; **36**(1): 89–108. https://projecteuclid.org/journals/statistical-science/volume-36/issue-1/Network-Modeling-in-Biology--Statistical-Methods-for-Gene-and/10.1214/20-STS792.full

639 Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev*
640 *Genet* 2009; **10**(1): 57–63.
641 West SG, Taylor AB, Wu W. Model fit and model selection in structural equation modeling. In:
642 Hoyle RH, editor. *Handbook of structural equation modeling*. New York: The Guilford Press;
643 2012. p. 209–231.

644

645

646 **List of Abbreviations**
647 **App** Application
648 **DEGs** Differentially Expressed Genes
649 **FC** Fold-change
650 **FTLD-U** Frontotemporal Lobar Degeneration with Ubiquitinated Inclusions
651 **GEO** Gene Expression Omnibus
652 **GOF** Goodness of Fit
653 **GWAS** Genome-wide Association Studies
654 **KEGG** Kyoto Encyclopedia of Genes and Genomes
655 LRTs Likelihood Ratio Tests
656 **MI** Modification Indices
657 **ML** Maximum Likelihood
658 **MS** Multiple Sclerosis
659 **NCBI** National Center for Biotechnology Information
660 **NDEGs** Not Differentially Expressed Genes
661 **PBMC** Peripheral Blood Mononuclear Cells
662 **PIR** Protein Information Resource
663 PTMs Post-translational modifications
664 **QRT-PCR** Quantitative Real-time Polymerase Chain Reaction
665 **QTL** Quantitative Trait Loci
666 **QTLnet** Quantitative Trait Loci (QTL)-driven Phenotype Network Method
667 **RMSEA** Root Mean Square Error of Approximation
668 **RNA-seq** RNA-sequencing
669 **SAM** Significance Analysis of Microarrays
670 **SCZ** Schizophrenia
671 **SEM** Structural Equation Modeling
672 SM Supplementary Materials
673 **SML** Sparsity-aware Maximum Likelihood

SPIA Signaling Pathway Impact Analysis
SRMR Standardized Root Mean Square Residual
STRING Search Tool for the Retrieval of Interacting Genes/Proteins Database

Availability and Requirements

Project name: ShinyDegSEM application project

Project home page: https://osf.io/kw8zf/

Operating system(s): Platform independent

Programming language: R

Other requirements: Not applicable.

License: Correct citation is needed.

Any restrictions to use by non-academics: Correct citation is needed.

Figure 1

Layout of the ShinyDegSEM application

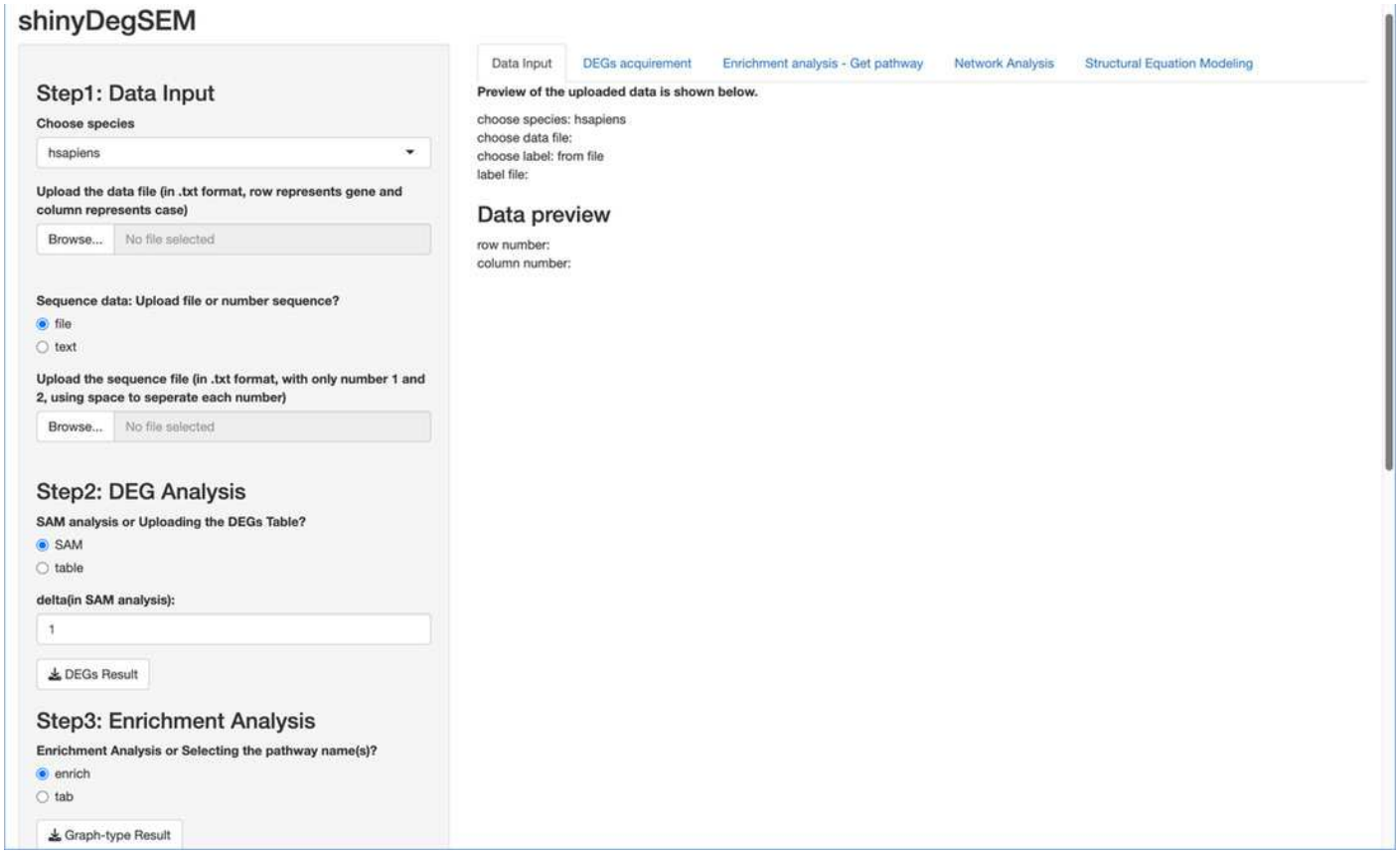


Figure 2

ShinyDegSEM app screen view after uploading data and group membership files

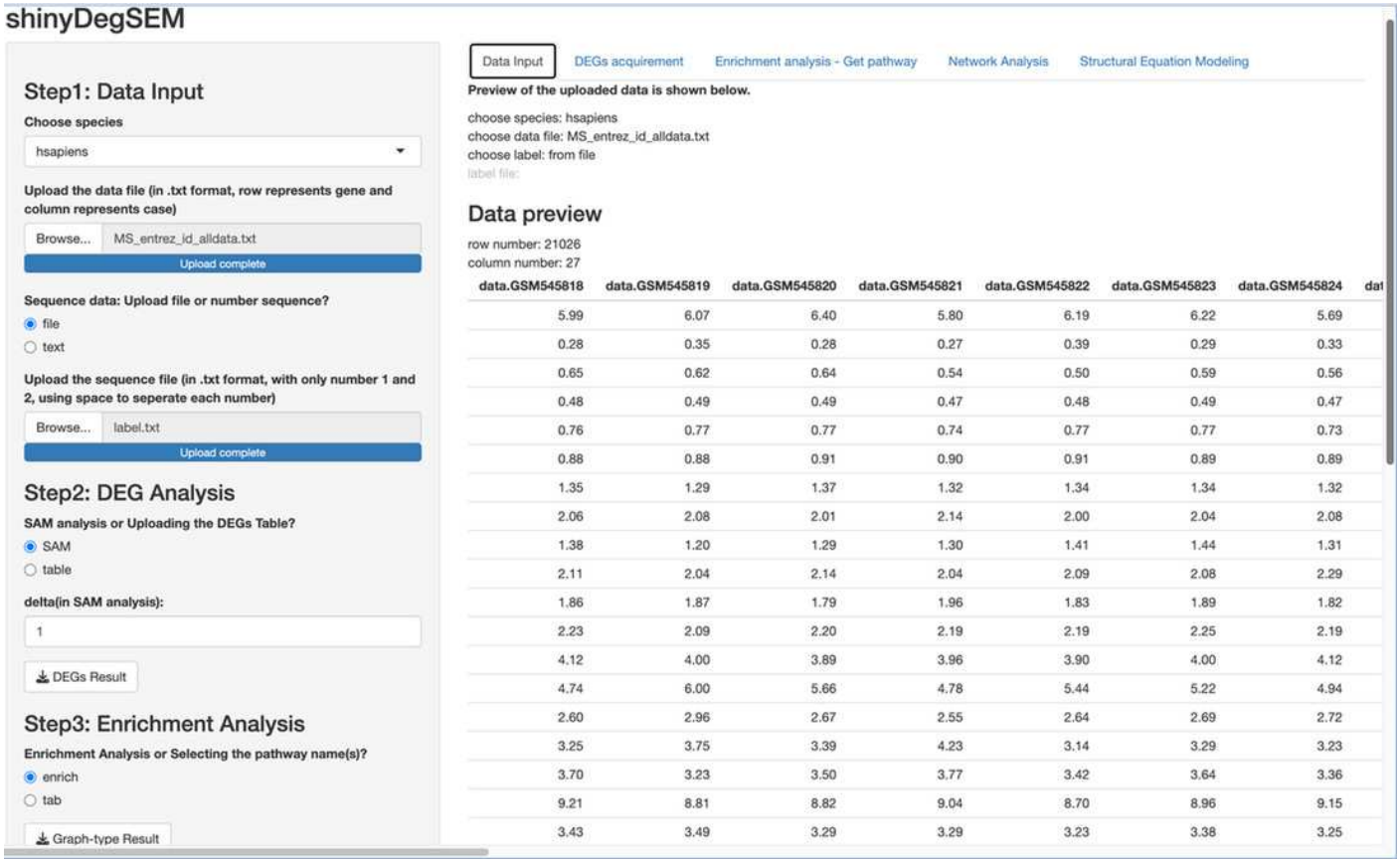
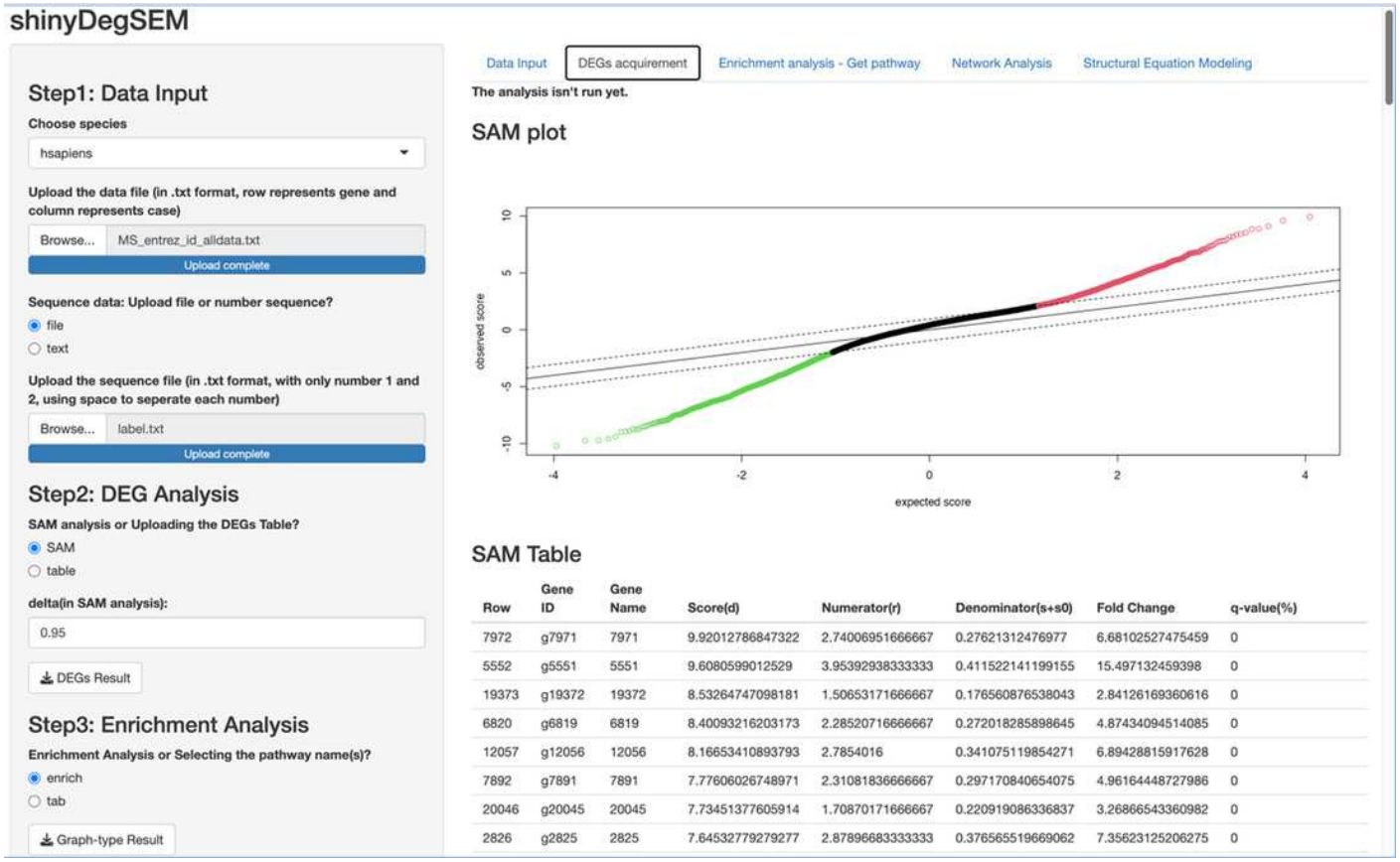


Figure 3

Output of significance analysis of microarrays (SAM) for Multiple Sclerosis *(MS) data*



Row	Gene ID	Gene Name	Score(d)	Numerator(r)	Denominator(s+s0)	Fold Change	q-value(%)
7972	g7971	7971	9.92012786847322	2.74006951666667	0.27621312476977	6.68102527475459	0
5552	g5551	5551	9.6080599012529	3.95392938333333	0.411522141199155	15.497132459398	0
19373	g19372	19372	8.53264747098181	1.50653171666667	0.176560876538043	2.84126169360616	0
6820	g6819	6819	8.40093216203173	2.28520716666667	0.272018285898645	4.87434094514085	0
12057	g12056	12056	8.16653410893793	2.7854016	0.341075119854271	6.89428815917628	0
7892	g7891	7891	7.77606026748971	2.31081836666667	0.297170840654075	4.96164448727986	0
20046	g20045	20045	7.73451377605914	1.70870171666667	0.220919086336837	3.26866543360982	0
2826	g2825	2825	7.64532779279277	2.87896683333333	0.376565519669062	7.35623125206275	0

Figure 4

The Fc gamma R-mediated phagocytosis pathway

The green nodes are DEGs. The yellow nodes are not DEGs (i.e., NDEGs).

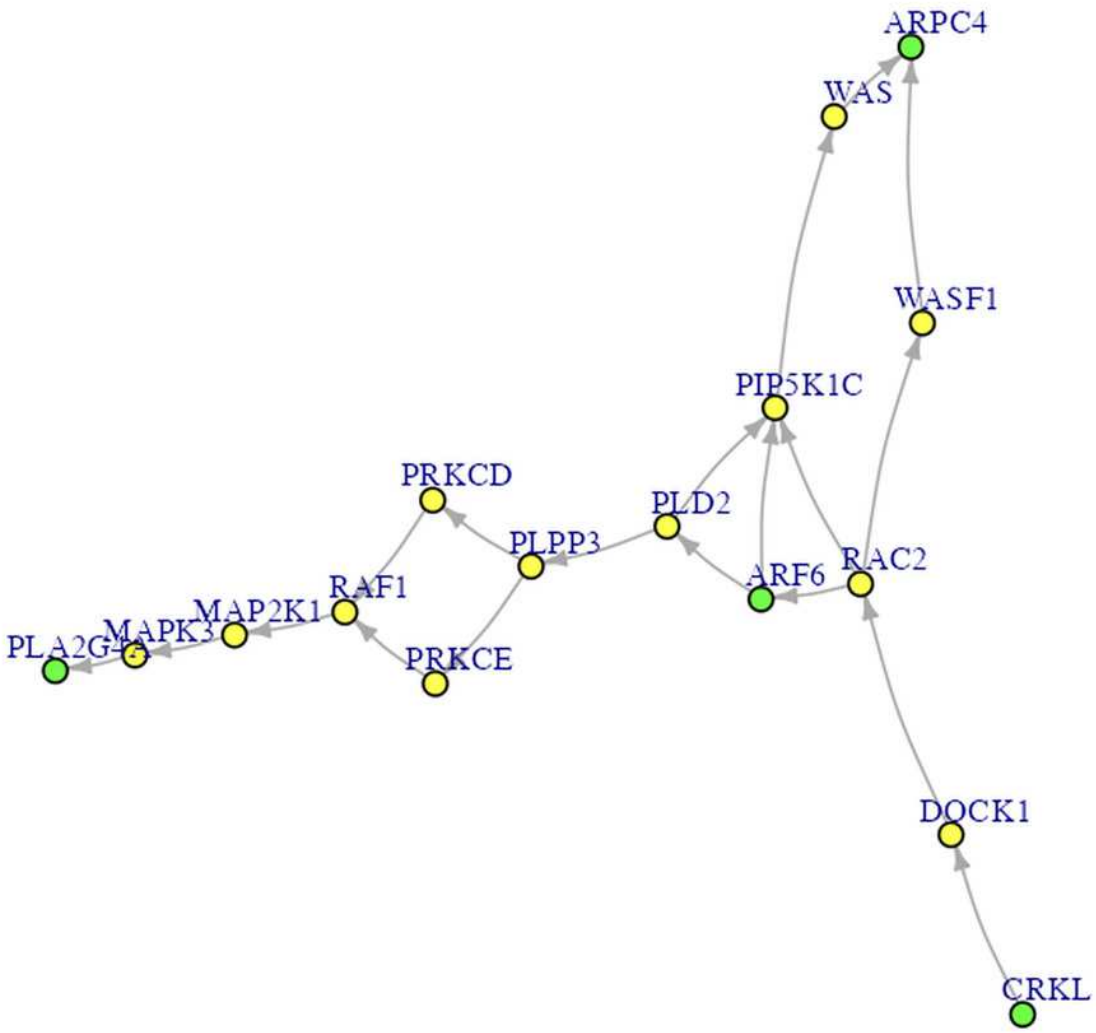


Figure 5

Initial structural equation modeling (SEM) output from network analysis for Chagas Disease pathway

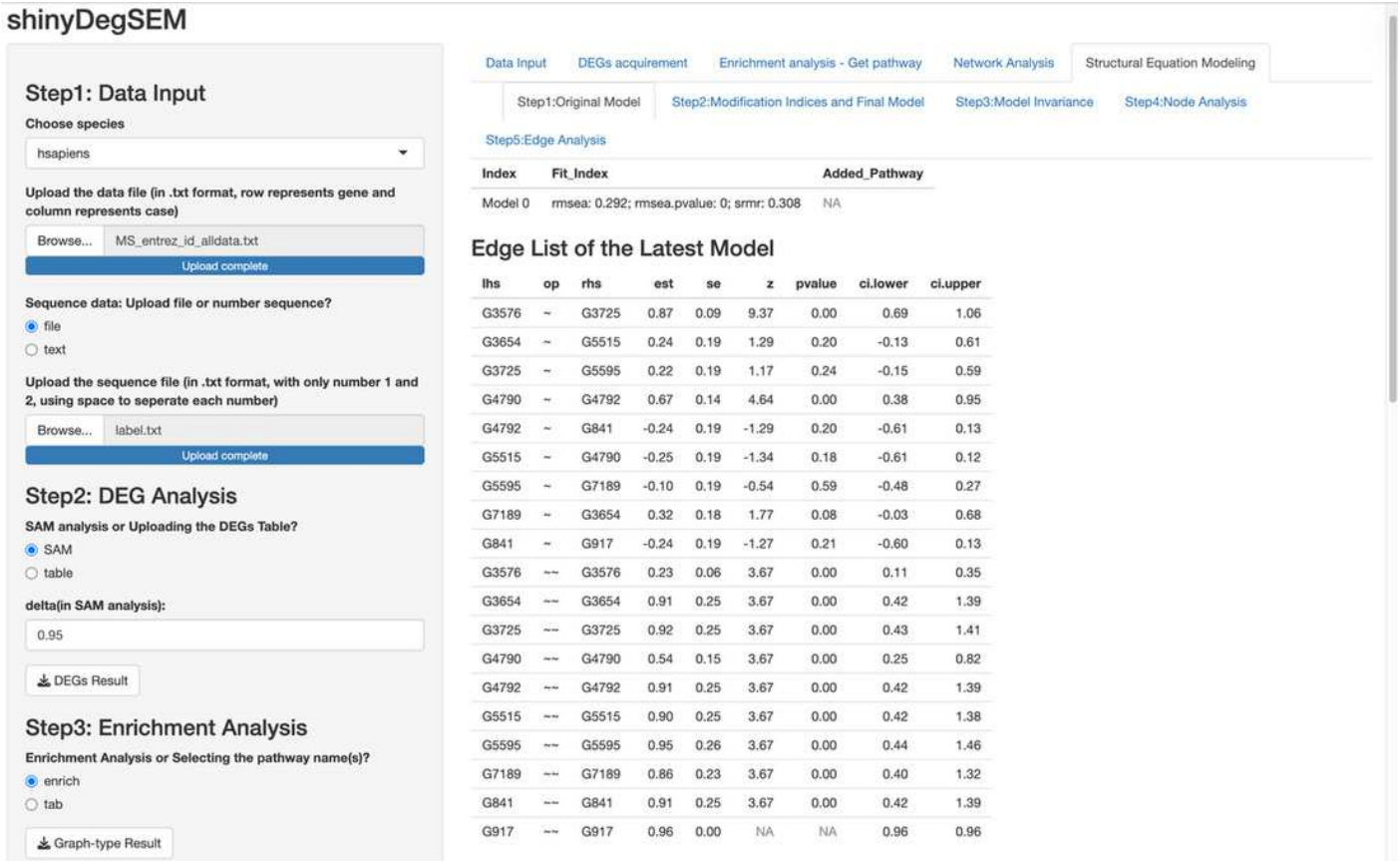
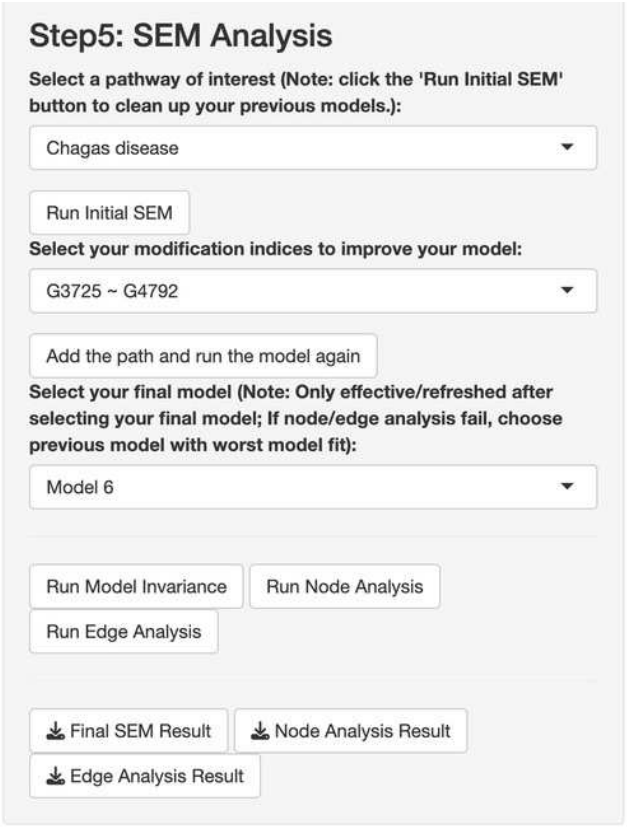


Figure 6

Final structural equation modeling (SEM) output from network analysis for Chagas Disease pathway



G5515	~~	G5515	0.95	0.28	3.44	0.00	0.41	1.49
G5595	~~	G5595	1.57	0.98	1.61	0.11	-0.34	3.48
G7189	~~	G7189	0.48	0.14	3.39	0.00	0.20	0.76
G841	~~	G841	0.91	0.25	3.67	0.00	0.42	1.39
G917	~~	G917	0.35	0.09	3.67	0.00	0.16	0.53

Latest Model | Final Model (refreshed after selecting your final model)

```
      rmsea rmsea.pvalue         srmr
      0.091        0.265        0.120
```

```
lavaan 0.6-18 ended normally after 35 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        25

  Number of observations                            27

Model Test User Model:

  Test statistic                                36.710
  Degrees of freedom                                30
  P-value (Chi-square)                           0.186

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Observed
  Observed information based on                Hessian

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  zG3576 ~
    zG3725            1.088    0.123    8.853    0.000
  zG3654 ~
    zG3576           -1.971    0.853   -2.311    0.021
    zG5515           -0.979    0.602   -1.626    0.104
  zG3725 ~
    zG5515           -0.616    0.276   -2.229    0.026
    zG5595            1.415    0.681    2.079    0.038
  zG4790 ~
    zG4792            0.661    0.150    4.414    0.000
  zG4792 ~
    zG3725            0.539    0.172    3.130    0.002
```

Figure 7

Model invariance result between two groups for the final model on Chagas Disease

The two groups were multiple sclerosis (MS) patients and healthy controls. “fit_base” refers to a model which does not consider group effects and assumes same patterns of relationships (e.g., nodes, edges, and pathways) across groups. “fit_node” refers to a common model which assumes the node baselines (e.g., baseline expression of genes or intercepts when all upstream regulators in the model are 0) are equal across groups. fit_edge” refers to a two-group model which assumes the strength or direction of relationships (e.g., path coefficients and gene-gene interaction or edge weights) are equal across groups. npar = number of model parameters, chisq = chi-square goodness of fit test, df = degrees of freedom, pvalue = *p* value for chi-square goodness of fit test, “cfi” = comparative fit index, “aic” = Akaike information criterion, “bic” = Bayesian information criterion, “rmsea” = root mean square error of approximation, “srmr” = standardized root mean square residual, ANOVA = analysis of variance.

Model Invariance

Model Fit Indices

```
Model Fit Indices of the Base, Group Effects on Node, and Group Effects on Edge Models
         npar     chisq df       pvalue       cfi      aic      bic     rmsea      srmr
fit_base   25  36.70964 30 1.857558e-01 0.9520594 657.7890 690.1850 0.0910138 0.1198040
fit_node   35  38.97399 30 1.263125e-01 0.9489827 634.1102 679.4645 0.1052568 0.1100210
fit_edge   50 125.69341 60 1.474240e-06 0.5481113 735.7547 800.5466 0.2847858 0.2241065
```

Nested Model Comparison (ANOVA)

```
ANOVA (Base vs. Edge Models)

Chi-Squared Difference Test

         Df    AIC    BIC  Chisq Chisq diff   RMSEA Df diff Pr(>Chisq)
fit_base 30 657.79 690.18  36.71
fit_edge 60 735.75 800.55 125.69     88.984 0.26985      30  9.356e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Group Effects on Node Model

```
Chi-square Goodness-of-Fit Test (for Group Effects on Node Model, i.e., fit_node model)
lavaan 0.6-18 ended normally after 24 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        35

  Number of observations                            27

Model Test User Model:

  Test statistic                                38.974
  Degrees of freedom                                30
  P-value (Chi-square)                           0.126
```

Figure 8

Node analysis result for the final model on Chagas disease pathway

"lhs" (left-hand side) denotes the dependent variable in the model. "op" = operator, "rhs" (right-hand side) represents the predictor variable. "est" represents estimated regression coefficient of the predictor variable on the dependent variable (i.e., each gene's expression or activity level). se = standard error of the estimated regression coefficient, z = standardized test statistic (i.e., z score) for the estimated regression coefficient. ci.lower = lower bound of the 95% confidence interval for the estimated regression coefficient, ci.upper = upper bound of the 95% confidence interval for the estimated regression coefficient. The symbol "~" means "is regressed on". group = 1 for patients with *Multiple* Sclerosis (MS), group = 0 for healthy controls.

	lhs	op	rhs	est	se	z	pvalue	ci.lower	ci.upper
1	G3576	~	group	0.308	0.108	2.867	0.004	0.097	0.519
2	G3654	~	group	0.134	0.375	0.356	0.722	−0.602	0.869
3	G3725	~	group	0.306	0.252	1.216	0.224	−0.187	0.800
4	G4790	~	group	−0.189	0.191	−0.992	0.321	−0.563	0.184
5	G4792	~	group	0.510	0.183	2.793	0.005	0.152	0.868
6	G5515	~	group	−0.731	0.144	−5.091	0.000	−1.012	−0.449
7	G5595	~	group	0.371	0.223	1.662	0.097	−0.066	0.807
8	G7189	~	group	−0.241	0.191	−1.261	0.207	−0.616	0.134
9	G841	~	group	−0.355	0.257	−1.381	0.167	−0.859	0.149
10	G917	~	group	0.284	0.157	1.807	0.071	−0.024	0.591

Figure 9

Part of the edge analysis result for the final model on Chagas Disease pathway

"lhs" (left-hand side) denotes the dependent variable in the model. "op" = operator, "rhs" (right-hand side) represents the predictor variable. d_est = path coefficient between two genes, quantifying the strength and direction of the gene influence. d_se = standard error of the path coefficients, d_z = standardized z-scoreof the estimated path coefficient. d_lower = lower bound of the 95% confidence interval for the path coefficient, d_upper = upper bound of the 95% confidence interval for the path coefficient. The symbol "~" means "is regressed on".

	lhs	op	rhs	d_est	d_se	d_z	pvalue	d_lower	d_upper
1	G3576	~	G3725	2.521	3.858	0.653	0.514	-5.042	10.083
2	G3654	~	G3576	-6.252	13.900	-0.450	0.653	-33.496	20.992
3	G3654	~	G5515	-6.430	10.132	-0.635	0.526	-26.289	13.429
4	G3725	~	G5515	-1.199	0.850	-1.410	0.159	-2.866	0.468
5	G3725	~	G5595	-0.982	2.964	-0.331	0.740	-6.792	4.828
6	G4790	~	G4792	-0.141	0.292	-0.483	0.629	-0.714	0.431
7	G4792	~	G3725	0.264	0.384	0.686	0.492	-0.489	1.016
8	G4792	~	G841	0.238	0.381	0.625	0.532	-0.509	0.985
9	G5515	~	G4790	-0.342	0.388	-0.881	0.378	-1.103	0.419
10	G5595	~	G3654	0.068	0.761	0.089	0.929	-1.423	1.559
11	G5595	~	G7189	0.845	0.658	1.285	0.199	-0.444	2.134
12	G7189	~	G3654	0.690	0.497	1.387	0.165	-0.285	1.665
13	G7189	~	G5515	0.349	0.336	1.038	0.299	-0.310	1.009
14	G841	~	G917	0.536	0.368	1.454	0.146	-0.186	1.257
15	G917	~	G5515	-0.821	0.303	-2.710	0.007	-1.415	-0.227