

# Structural Equation Modeling of In silico Perturbations

Jianying  ${\rm Li^1}$ , Pierre R. Bushel<sup>1</sup>, Lin  ${\rm Lin^2}$ , Kevin  ${\rm Day^3}$ , Tianyuan  ${\rm Wang^1}$ , Francesco J. Demayo<sup>4</sup>, San-Pin  ${\rm Wu^{4^*}}$ , Jian-Liang  ${\rm Li^{1^*}}$ 

<sup>1</sup>National Institute of Environmental Health Sciences (NIEHS), United States, <sup>2</sup>Department of Family Health Care Nursing, School of Nursing, University of California, San Francisco, United States, <sup>3</sup>Duke University, United States, <sup>4</sup>Reproductive and Developmental Biology Laboratory, National Institute of Environmental Health Sciences (NIEHS), United States

Submitted to Journal: Frontiers in Genetics

Specialty Section: Computational Genomics

Article type: Technology and Code Article

Manuscript ID: 727532

Received on: 18 Jun 2021

Journal website link: www.frontiersin.org



#### Conflict of interest statement

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

#### Author contribution statement

JL and PB designed the framework, performed the analyses and drafted the paper. LL provided the guidance on SEM, KD developed and draft the R shiny code, TW provided essential components for T-Score calculation and prepared gene signatures and processed public data, FD, SW and JL conceived the idea, provided overall guidance and oversaw the project progression.

#### Keywords

Structural Equation Modeling, Gene Expression, In silico perturbation, molecular interaction, r, Shiny

#### Abstract

Word count: 142

Structural Equation Modeling (SEM) is a statistical approach for studying complex cause-effect hypotheses in a "closed system" of latent (hidden) endogenous variables. SEM has been widely used in various fields involving perturbations and measurable outcomes. We developed an R Shiny application, termed "Structural Equation Modeling of In silico Perturbations (SEMIPs)" to aid in the transfer of perturbations in gene expression pathways from one system to another for determining casual inference of molecular interactions in silico. SEMIPs computes a two-sided t-statistic (T score) to rank signature gene activities for modeling. It implements a basic SEM model and then performs bootstrap random sampling for statistical significance. As a use case example for SEMIPs, we showed that putative direct downstream genes of the GATA2 transcription factor are sufficient to infer GATA2's activities in silico for the conserved PGR-GATA2-SOX17 genetic network in the human uterine endometrium.

#### Contribution to the field

This manuscript described an R Shiny application, termed "Structural Equation Modeling of In silico Perturbations (SEMIPs)" that allows the transfer of perturbations in gene expression pathways from one system to another for determining causal inference of molecular interactions in silico. It provides a convenient interface to allow bench scientists to explore the expression data and generate testable hypotheses. SEMIPs computes a two-sided t-statistic (T score), applies a basic SEM model, and performs bootstrap random sampling for statistical significance. It has the capability to calculate response activities using large datasets representative of biological systems, to leverage the power of SEM to test the relationship among end points in a study, to provide users with the flexibility for testing new hypotheses, and to integrate a non-parametric testing procedure for assessing statistical significance. Overall, SEMIPs enables researchers to rapidly examine a given gene signature across systems and generate the functional hypothesis. This tool will be of broad interest to the translational research community.

#### Funding statement

This research was supported by the Intramural Research Program of the National Institute of Environmental Health Sciences Z1AES103311 (FJD), Z99-ES999999 (SPW), Z01-ES102345 (PRB) and by the Gaine Research Foundation GRF-2018-01 (LL).

#### Ethics statements

#### Studies involving animal subjects

Generated Statement: No animal studies are presented in this manuscript.

#### Studies involving human subjects

Generated Statement: No human studies are presented in this manuscript.

#### Inclusion of identifiable human data

 $\label{thm:continuous} Generated \ Statement: \ No \ potentially \ identifiable \ human \ images \ or \ data \ is \ presented \ in \ this \ study.$ 

#### Data availability statement

Generated Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40659 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58144 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39920.



# Structural Equation Modeling of In silico Perturbations

- Jianying Li<sup>1,2,3</sup>†, Pierre R. Bushel<sup>3,4</sup>†, Lin Lin<sup>5</sup>, Kevin Day<sup>6</sup>, Tianyuan Wang<sup>1,2</sup>, Francesco J.
- 2 DeMayo<sup>7</sup>, San-Pin Wu<sup>7</sup>\*, and Jian-Liang Li<sup>1</sup>\*
- 3 <sup>1</sup> Integrative Bioinformatics, Epigenetics and Stem Cell Biology Laboratory, Division of Intramural
- 4 Research, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709,
- 5 USA
- 6 <sup>2</sup> Kelly Government Solutions, Research Triangle Park, NC 27709, USA
- 7 Massive Genome Informatics Group, National Institute of Environmental Health Sciences, Research
- 8 Triangle Park, NC 27709, USA
- 9 <sup>4</sup> Biostatistics and Computational Biology Branch, Division of Intramural Research, National
- 10 Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA
- <sup>5</sup> Department of Family Health Care Nursing, University of California at San Francisco, San
- 12 Francisco, CA 94143, USA
- 13 <sup>6</sup> Duke University, Durham NC 27713
- <sup>7</sup> Reproductive and Developmental Biology Laboratory, National Institute of Environmental Health
- 15 Sciences, Research Triangle Park, NC 27709, USA
- <sup>†</sup>These authors have contributed equally to this work and share first authorship
- 18 \* Correspondence:
- 19 Jian-Liang Li (jianliang.li@nih.gov) and San-Pin Wu (steve.wu@nih.gov)
- 20 Running Title: Structural Equation Modeling In silico
- 21 Keywords: Structural Equation Modeling, gene expression, *in silico* perturbation, molecular
- 22 interaction, R, Shiny
- 23 Word count: 2249
- 24 Number of Figures: 2 Number of Supplementary Figures: 4
- Number of Tables: 0 Number of Supplementary Tables: 2

16

#### Abstract

Structural Equation Modeling (SEM) is a statistical approach for studying complex cause-effect hypotheses in a "closed system" of latent (hidden) endogenous variables. SEM has been widely used in various fields involving perturbations and measurable outcomes. We developed an R Shiny application, termed "Structural Equation Modeling of In silico Perturbations (SEMIPs)" to aid in the transfer of perturbations in gene expression pathways from one system to another for determining casual inference of molecular interactions *in silico*. SEMIPs computes a two-sided t-statistic (T score) to rank signature gene activities for modeling. It implements a basic SEM model and then performs bootstrap random sampling for statistical significance. As a use case example for SEMIPs, we showed that putative direct downstream genes of the GATA2 transcription factor are sufficient to infer GATA2's activities *in silico* for the conserved PGR-GATA2-SOX17 genetic network in the human uterine endometrium.

#### Introduction

Although gene expression data in public repositories provide a valuable resource for investigators to infer regulatory processes (Edgar et al. 2002), the variables of interest are not always directly measurable in a causal response model system. Moreover, it is challenging to test the knowledge obtained from experimental model systems in humans due to undetermined clinical outcomes and ethical considerations. Structural equation modeling (SEM) offers a statistical framework to make casual inferences about the causality of latent (hidden) endogenous variables in a system (Grace 2006). We were motivated to develop a Structural Equation Modeling of In silico Perturbations (SEMIPs) Shiny application to facilitate casual inference from *in silico* alterations of gene expression pathways. SEMIPs enables quantification of a projected activity metric (two-sided t-statistic, i.e. T score) calculated from gene expression activity upon exposure to a perturbation (Wu et al. 2015), thus allowing users to fit desired SEM models using selected endogenous and exogenous variables.

51 This application also provides two different bootstrap random sampling procedures (elimination with

or without replacement) for testing the significance of a model based a non-parametric distribution.

Previously, SEM was applied to gene expression data to evaluate an alteration of latent gene

interactions that disrupts the progesterone receptor pathway in the uterus of pregnant mice and the

model was then transferred in silico to a human reproductive system (Rubel et al. 2016). SEMIPs

streamlines this process and allows bench scientists to perform the computations and analysis

through a user-friendly interface.

#### **Materials and Methods**

#### Overview of SEMIPs

52

53

54

55

56

57

58

59

62

63

64

65

66

67

68

69

70

71

72

73

74

This SEMIPs RShiny App allows users to compute a two-sided t-statistic (T score) from gene expression data to infer the activities of genes of interest in a quantitative manner. Shown in Figure 1,

SEMIPs App, which is highlighted in the orange dotted rectangle, facilitates the hypothesis generation

and testing framework, this app also provides a 3-node model fitting function using structural equation

modeling to test the joint regulation of a target gene by two upstream regulators in silico. In addition,

for hypothesis generation purposes, a two-way bootstrap method, elimination with replacement or

elimination without replacement, is included in the app to examine the impact of removing genes that

belong to the same signaling cascade from the downstream targets of the gene of interest.

#### **T-score Calculation**

The biological hypothesis generation relies on results obtained from the model animal perturbation

system, normally with experimental mice or rats, then projects into human or other animal systems

when either direct perturbation is not possible or the variables of interest are not directly measurable.

Under an experimental perturbation, the genomics system response was measured and exemplified

through the significantly changed gene set. Such an information was projected into another system

via gene orthologues and the activities of genes of interest will be calculated by a two-side t-statistic

75 (T-score) (Rubel et al. 2016). The SEMIPs RShiny App provides an automated route to calculate 76 such T-score with a separated tab "T-Scores" shown in Figure 2. It requires two components: (1) A 77 list of gene signature (in Entrez gene symbol format) obtained from a study of interest; and (2) A 78 gene expression data matrix that consists of gene expression profiles in a given context. The 79 application will conduct the analysis and produce inferred activity results reflected as T scores that 80 can be used in subsequent downstream analyses. 81 **Structural Equation Modeling** 82 The second feature of SEMIPs App is the structural equation modeling (SEM). SEM is a statistical 83 modeling approach that focuses on the study of complex cause-effect hypotheses about the 84 mechanisms operating in systems, it useful when some variables are not directly measurable often 85 used in clinical psychology research (Lin et al. 2013). We implement the SEM with lavaan package 86 (Rosseel 2018) to provide a 3-node model fitting function using structural equation modeling to test 87 the joint regulation of a target gene by two upstream regulators in silico. The app comes with a 88 sample data file "app installation dir/dataSEM/sampleDAT.txt". When the SEM tab is selected 89 (Figure 2), this data will be loaded, users can select three variables from the drop-down list to test the 90 SEM model. The SEMIPs App also provides a data file template 91 "app installation dir/dataSEM/ sampleDAT.txt" that users can modify and save as 92 "sampleDAT.txt" to overwrite the default data. As a result, users' data will be loaded when the App 93 is launched next time. Users can save the modeling figure and all fitting statistics from the app. 94 **Bootstrap Simulation** 95 The third feature (the bootstrap tab shown in Figure 2) assesses the potential impact from a 96 perturbation on any downstream system. We implemented a two-class (elimination with or without 97 replacement) bootstrap resampling for statistical inference (Supplementary Figure 1), which 98 eliminates unrelated signatures and provides statistical significance to the SEM fitting. For this

feature, it is assumed that the users have successfully run a T score analysis. The users also need to

99

enter the signatures associated with the downstream system of interest to evaluate. To improve the rigor of the statistical test, it is recommended to run the bootstrap a minimum of 1,000 times. This feature involves bootstrapping simulation, it needs multicore hardware and can take more times depending on how many iterations users choose.

#### Sample Data

The SEMIPs App comes with four sets of testing data and data templates for user to use the application and further modify their own data for any customized research projects. They are located at app\_installation\_dir/testData.

### **Hardware and Software Requirement**

SEMIPs was written in R with the Shiny package (Rstudio 2014) that is known for its light weight web development framework with shiny-related features. The lavaan package (Rosseel 2018) was used for the SEM, other depending packages will be checked at the installation and need to be installed if not already available. The application requires modern multicore CPUs for the backend parallel processes. SEMIPs was developed under Linux CentOS7 and has been successfully tested on MacOS (v. 10.14.6) and Windows10. To install and run this application, users can follow the detailed instructions provided in the README.txt file. The SEMIPs Shiny app and source code are freely available at <a href="https://github.com/NIEHS/SEMIPs">https://github.com/NIEHS/SEMIPs</a> under the MIT license.

#### Results

#### An Integrated Hypothesis Generation and Testing Framework

As shown in Figure 1, the SEMIPs workflow depicts a biological question initially tested in an animal model and then applied to a human system. A biological hypothesis is tested in a model animal system (mouse) on relationship between two interacting factors (Fac1 & Fac2) and their endpoints. The hypothesis is translated to another species (i.e. human in our research) via T-score computation and verified with SEM model. Based on the SEM model, a presumed relationship can

be tested in humans by determining the significance of the inference via a non-parametric bootstrap resampling framework. The resulting perturbed pathways can be eventually tested in the animal model. These workflow steps are shown within the dotted rectangle on the right side of Figure 1 with three major features implemented in the SEMPIPs App as the function tabs when the Shiny App is launched (Figure 2).

The Signature Analysis component conducts the T-Score calculation that ultimately helps to translate

#### **T-Score Calculation Assisted Translational Research**

the knowledge obtained from the experimental animal study, as an example into human system.

Users can test this feature by uploading: (1) A list of gene signature (in Entrez gene symbol format) obtained from a study of interest (i.e. Human Sig.xlsx); and (2) A gene expression data matrix that consists of gene expression profiles in a given context (i.e. HumanArray2Shiny.xlsx) located under "/app\_installation\_dir/testData/t-score/". Once successfully uploaded, top few lines of data will be visible for preview (Figure 2). For illustration purposes, we provide both mouse signature (i.e. Mouse Sig.xlsx) and human signature files, the proper matched specie needs to be selected. T-scores will be calculated by clicking the green "Go!" button, the top 10 rows of the T-scores will be shown for preview. The users are encouraged to download the T-Scores for further analysis. Since the T-Scores are calculated from two-side T-test, the corresponding p-values are also reported (the second column in T-Scores results shown in Figure 2).

# **Flexible Structural Equation Modeling**

A three-node structural equation model can be hypothesized by selecting the desired endogenous and exogenous variables. Once the SEM tab is selected, the default data

("app\_installation\_dir/dataSEM/sampleDAT.txt") will be loaded, and all features are available for users to choose from the drop-down windows. Two variables are hypothesized as "causal variable" and one variable will be the "endpoint". The tool reports model fitting statistics in a compressed (zipped) file that can be downloaded, the three-node SEM figure can also be downloaded. This

feature also allows users to test a separate system by uploading their relevant dataset. The dataset requires the same format as the example data. As shown in Figure 1, the SEM model fitting results especially those important statistics will provide valuable information to the hypothesis of interest, which can be further validated from the bench experiment. Sometime, the results can help researchers to proposed new hypothesis.

### **Two-class Bootstrap Simulation**

This feature was designed to assesses the potential impact from a perturbation on any downstream system. For a gene signature list obtained from the perturbation, any gene or gene sets that are biologically associated can be tested with this two-class (elimination with or without replacement) bootstrap resampling for statistical inference (Supplementary Figure 1). In the test data folder "/app\_installation\_dir/testData/bootstrap/", four downstream gene sets are available. Under the "Bootstrap" tab, users can navigate to this location and run the bootstrap simulation analysis. The impact on the downstream system can be assessed by either elimination without replacement or with replacement. To ensure the rigor of the statistical test, it is recommended to run the bootstrap a minimum of 1,000 times. Depending on the hardware configuration, this analysis can take a considerable amount of time. Users can download the zipped results after the analysis is completed.

### **A User Case Application**

Previously we demonstrated that the mouse gene signatures of GATA2 and PGR allow inference of the interaction between GATA2 and PGR for regulation of SOX17 expression in the human endometrial tissues (Rubel et al. 2016). The full GATA2 gene signature consists of both direct and indirect downstream genes of GATA2 in the uterus (Rubel et al. 2016). Since GATA2 is known as a transcription factor that occupies cis-acting elements and confers genomic actions, we hypothesize that expression levels of GATA2's direct downstream targets reflect its activities *in silico*. Here, a GATA2 direct downstream target is defined as a GATA2 regulated gene with GATA2 genome occupancy within 2-kilobase vicinity of the said gene's transcription start site in the uterus (Gene Expression

Omnibus (GEO) accession: GSE40659, (Rubel et al. 2016)). This stringent criterion led to the identification of 634 genes (Supplementary Table 1), which is termed "GATA2 direct signature". The GATA2 activity, as represented by the GATA2 direct signature in a T-score, was quantified by the SEMIPs app from gene expression data of the endometrium tissue for each individual human subject (GEO accession: GSE58144, (Koot et al. 2016)). T scores for the uterine GATA2 in all 115 patients were calculated by the app with the GATA2 direct signature and the data matrix of GEO accession: GSE58144 (Supplementary Table 2). Similarly, T scores for the uterine PGR (termed PGR signature) were obtained using the GEO accession: GSE39920 dataset (Rubel et al. 2016) on the same data matrix via the application's T score calculation function. To test whether the GATA2 direct signature fits the model of the 3-node PGR-GATA2-SOX17 genetic network, the application was fed with T scores of GATA2 direct signature and PGR signature as exogenous variables and the SOX17 expression levels as the endogenous variable under the "SEM" function. The output data shows that, with GATA2 direct signature in place of the full gene signature, the model significantly fits the GEO accession: GSE58144 dataset with all proposed paths (Supplementary Figure 2) and this model is considered not rejected by the human data. This finding suggests that the expression levels of GATA2 direct downstream targets, a subset of the full GATA2 regulated genes, can mathematically serve as surrogate reporters of the GATA2 activities in the human endometrium tissues, which supports our hypothesis. Results of this analysis not only reduce the number of reporter genes for GATA2 activities to 634, but also implicate possibilities of a further reduction with additional filtering criteria on the gene list. A small and manageable panel of markers for GATA2 activities could serve as a future diagnostic tool for pregnancy failure (Díaz-Gimeno et al. 2011).

#### Discussion

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

The SEMIPs R Shiny app offers an easy to use *in silico* perturbation testing system with several advantages. First, it has capability to calculate response activities using large datasets representative

198 of biological systems. Second, it leverages the power of SEM to test the relationship among end 199 points in a study and provides users with the flexibility for testing new hypotheses. Lastly, it 200 integrates a non-parametric testing procedure for assessing statistical significance. 201 This user-friendly app allows quick assessments on genetic interactions and subsequent hypothesis 202 generation without the requirement of extensive knowledge on computation languages and statistical 203 analyses. Due to its simplicity in design, this app is limited to a 3-node model fitting capability. 204 Models of higher complexity can be tested using the R package MplusAutomation that focuses on 205 automating the SEM modeling which was originally implemented in Mplus (Hallquist and Wiley 206 2018), a commercial software. 207 Currently, the two-class bootstrap analysis can only be conducted separately. Integration of these into 208 the SEMIPs methodology for formulation into a single test will be investigated for future design, 209 development and implementation. As noted in the manuscript and mentioned previously, the SEMIPs 210 app has been adopted by researchers in the field with a few papers published recently (Liu et al. 211 2019, Wetendorf et al. 2020). We hope that it can serve a wider research community to address 212 additional scientific questions. 213 **Author Contributions** 

JL and PB designed the framework, performed the analyses and drafted the paper. LL provided the guidance on SEM, KD developed and draft the Rshiny code, TW provided essential components for T-Score calculation and prepared gene signatures and processed public data, FD, SW and JL conceived the idea, provided overall guidance and oversaw the project progression.

#### **Funding**

214

215

216

217

218

219

220

221

This research was supported by the Intramural Research Program of the National Institute of Environmental Health Sciences Z1AES103311 (FJD), Z99-ES999999 (SPW), Z01-ES102345 (PRB) and by the Gaine Research Foundation GRF-2018-01 (LL).

# 222 Acknowledgments

- The authors would like to thank our colleagues Drs. Hamed Bostan, Eric Thomson, James Ward and
- 224 Matt Wheeler for kindly testing SEMIPs and for providing valuable feedbacks to improve the
- application. We also thank for Drs. John House and Rong Li for their critique of the draft of this
- 226 manuscript.

#### 227 Conflict of Interest

- 228 The authors would 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.

## 230 Reference

- Díaz-Gimeno, P., et al. (2011). "A genomic diagnostic tool for human endometrial receptivity based
- on the transcriptomic signature." Fertil Steril 95(1): 50-60, 60.e51-15.
- Edgar, R., et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array
- data repository." Nucleic Acids Res 30(1): 207-210.
- Grace, B. J. (2006). <u>Structural Equation Modeling and Natural Systems</u>, Cambridge University Press.
- Hallquist, M. N. and J. F. Wiley (2018). "MplusAutomation: An R Package for Facilitating Large-
- 237 Scale Latent Variable Analyses in Mplus." <u>Structural Equation Modeling: A Multidisciplinary</u>
- 238 Journal **25**(4): 621-638.
- Koot, Y. E. M., et al. (2016). "An endometrial gene expression signature accurately predicts recurrent
- implantation failure after IVF." Scientific reports 6: 19411-19411.
- Lin, L., et al. (2013). "Uncertainty, mood states, and symptom distress in patients with primary brain
- 242 tumors: analysis of a conceptual model using structural equation modeling." Cancer 119(15): 2796-
- 243 2806.

254

255

- Liu, J., et al. (2019). "JNK(1/2) represses Lkb(1)-deficiency-induced lung squamous cell carcinoma
- 245 progression." Nat Commun **10**(1): 2148.
- 246 Rosseel, Y. (2018). "Latent Variable Analysis."
- 247 Rstudio, I. (2014). "Shinny: Easy web applications in R."
- Rubel, C. A., et al. (2016). "A Gata2-Dependent Transcription Network Regulates Uterine
- 249 Progesterone Responsiveness and Endometrial Function." Cell Rep 17(5): 1414-1425.
- Wetendorf, M., et al. (2020). "Constitutive expression of progesterone receptor isoforms promotes
- 251 the development of hormone-dependent ovarian neoplasms." Sci Signal 13(652).
- Wu, S. P., et al. (2015). "Increased COUP-TFII expression in adult hearts induces mitochondrial
- 253 dysfunction resulting in heart failure." Nat Commun **6**: 8245.

#### Figure Legends

Figure 1. The workflow and application of SEMIPs. The left four rectangles and arrows indicate our hypothesis testing and generation schema. A biological hypothesis is tested in a model animal model system (mouse) on relationship between two interacting factors (Fac1 & Fac2) and their endpoints. The hypothesis is translated to another species (i.e. human in our research) via T-score computation and verified with SEM model. This process is accomplished with our shinyapp indicated by two curved arrows.  $\gamma 11$  and  $\gamma 21$  are correlation efficient and  $\xi$  are model residuals. The two-class bootstrap analysis is shown in the red rectangle box. Hypothesis generating and exploring steps are explained by the bottom two rectangles. Figure 2. The user interface is shown when it is launched. The main panel contains four tabs: "T-Scores", "SEM", "Bootstrap", and "Instruction". The right panel shows the screen when the "Tscores" is selected and generated. In the left panel, the application accepts two inputs, 1) a list of signatures (in Entrez gene symbol format) and 2) a data matrix of expression measurement with the top lines shown for viewing. The green "Go!" button is clicked to launch the T-score generation and grayed out to denote the process is running. The first 10 rows of the T-scores matrix are shown, which can be downloaded by clicking the "Download T-Scores" button.

256

257

258

259

260

261

262

263

264

265

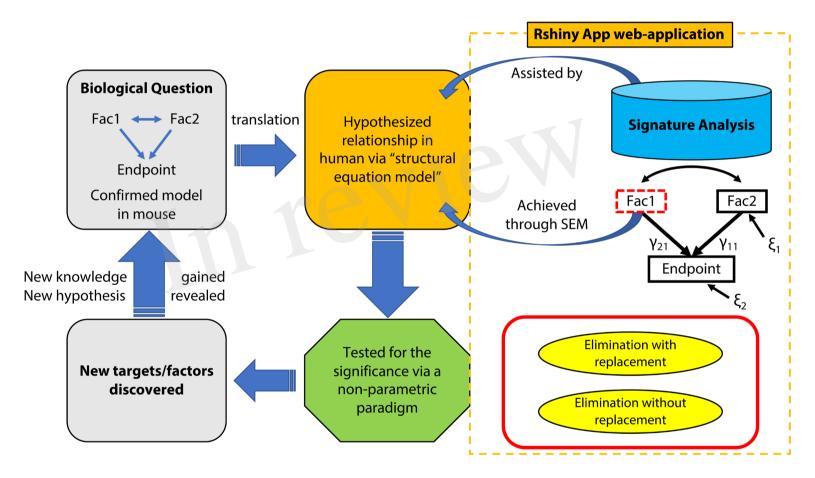
266

267

268

269

270



# **SEMIPs**

