Leveraging Dynamic Stability to Infer Regulation in Protein-Protein Interaction Networks: A Study of Infectious Vulnerability in COPD.

Joyce Reimer[1], Jeffrey Page[2], Pranta Saha[1], Shichen Shen[3], Xiaoyu Zhu[3], Shuo Qian[3], Manoj Mammen[4], Jun Qu[3], Sanjay Sethi[5], Gordon Broderick[1,6,7*]

[1] Vaccine and Infectious Disease Organization, University of Saskatchewan, Saskatoon, SK, Canada
[2] Complex Exposure Threats (CETC), US Veterans Affairs, Washington DC, USA
[3] Dep. of Pharmaceutical Sciences, University at Buffalo, The State University of New York, Buffalo, NY, USA
[4] Dep. of Medicine, University of Rochester Medical Center, Rochester, NY, USA
[5] Dep. of Medicine, Jacobs School of Medicine & Biomedical Sciences, University at Buffalo, The State University of New York, Buffalo, NY
[6] Dep. of Mathematics and Statistics, College of Arts and Science, University of Saskatchewan, Saskatoon, SK, Canada
[7] Dep. of Pediatrics, College of Medicine, University of Saskatchewan, Saskatoon, SK, Canada

* Corresponding author:
Gordon Broderick, PhD
Principal Scientist, Mathematical Immunology and Immunodynamics Lab. (MIIL)
Vaccine and Infectious Disease Organization (VIDO), University of Saskatchewan
Saskatoon, SK, Canada
gordon.broderick@usask.ca

ABSTRACT

The fourth leading cause of death in the US, Chronic Obstructive Pulmonary Disease (COPD) is punctuated by frequent viral and bacterial infections causing severe acute exacerbations (AECOPD) and increased mortality. In previous work we have shown that altered immune cell signaling may confer increased and persistent susceptibility to infection. Here we continue this investigation by conducting broad-spectrum proteomic profiling of circulating white blood cells to assemble an empirical protein-protein interaction network associated with frequency of infectious exacerbation. In a novel extension of conventional cross-sectional data analyses, we translate these undirected protein-protein interactions into candidate regulatory relationships with both direction and mode of action. The latter are inferred by formulating and solving a constraint satisfaction problem (SAT) whereby predicted dynamic behaviors of any valid regulatory network must support the expected persistent nature of low and high vulnerability phenotypes. Solving this SAT problem produced a set of competing candidate protein regulatory network architectures and signalling rules that unanimously highlighted several novel candidate pathway elements involved in oxidative stress response. Analysis of the overall dynamics supported by these networks, again supported the hypothesis that progression beyond an immune tipping point may confer persistent susceptibility to infection and that this may constitute a stable phenotype or regulatory trap in COPD characterized by a reactive oxygen cascade.

Keywords. Infectious risk; COPD; regulatory networks; tipping point; stable attractors; immune regulation; protein-protein interaction networks

HIGHLIGHTS

- Infectious vulnerability phenotypes in COPD may represent immune regulatory traps.
- A 15-protein oxidative stress response interactome offers a phenotypic signature.
- Homeostasis in cross-sectional data can support inference of co-regulatory actions.
- 8 of 11 documented regulatory actions recovered in candidate model majority vote
- 17 new candidate protein co-regulatory actions proposed unanimously by all models.

1. INTRODUCTION

As our ability to accurately measure broad swaths of the proteome continues to evolve rapidly [1,2], often isolating numerous new proteins that have not yet been annotated. Moreover, even for those proteins that have been annotated, we frequently do not a clear understanding of their function, involvement and role in the broader highly interconnected network of intracellular pathways. To capture the context in which these proteins express, significant efforts have been directed at constructing protein-protein interactions (PPI) maps from experimental data using various measures of association and comparing their structure and connectivity patterns [3]. However, the direction in which a mediator affects a target protein and whether this action activates or inactivates the latter remains challenging to infer limiting current analyses to static comparisons of network structure and connectivity patterns [4]. This is especially true for cross-sectional observational data, which in many cases represents the majority of available data [5]. Numerical strategies have been proposed to generate artificial time course data from cross-sectional data sets that typically involve the re-ordering of pseudo- steady state observations such that they can be supported as an ordered sequence by an underlying forecasting model e.g. Markov Chain (HM) [6,7]. Extensions of such methods have also been proposed that do not require subjects align along a single trajectory but instead that subjects with proximal profiles simply evolve in a shared direction predicted using a set of stochastic differential equations based on Langevin dynamics [8].

Here, we propose a complementary approach where instead of attempting to describe a trajectory of progression, we focus on homeostatic regulatory stability as a defining characteristic by which directed functional relationships between proteins might be inferred from standard undirected PPI maps. We apply this strategy to cross-sectional protein expression data collected as part of a larger ongoing study directed at identifying the mechanistic underpinnings that drive alternative immune responses in subjects with chronic obstructive pulmonary disease (COPD) that are especially vulnerable to infection. The fourth leading cause of death in the US, the course of COPD is punctuated by frequent acute exacerbations (AECOPD), mostly caused by tracheobronchial mucosal infection by bacteria and/or viruses [9]. Paradoxically, acute and chronic infection in COPD is prevalent despite persistent excessive inflammation and vigorous immune response to the pathogens. A ‘Vicious Circle’, where adaptive defects in innate immune response induced by tobacco smoke allow persistent infection, which in turn perpetuates inflammation and dysregulates innate immune response, has been proposed to explain this paradox [10]. As a potential test of this hypothesis, we investigate the role of soluble protein signaling in the peripheral circulation as an accessory to persistent inflammatory signaling in the mucosa [11] and how this might be used as a network signature of increased vulnerability to infection. In a simplifying modification of the above-mentioned methods, like subjects are

expected to progress towards shared one of two stable attractors, namely the persistent phenotypes of low and high vulnerability to infection. Using a one-step discrete difference equation to describe network regulatory dynamics, our analysis suggests that these two phenotypes can be distinguished based on the relative abundance of 15 out of 850 quantifiable protein species and that these phenotypes share a common co-regulatory network structure and state transition logic. As such it may be possible to redirect or de-escalate vulnerability to infection in highly susceptible individuals.

2. METHODS

2.1. Study Population and Sample Collection

As part of a larger ongoing study, serum samples were collected from N = 67 subjects with exacerbation history documented for up to 12 years. All subjects were males between the ages of 46-81, with a median exacerbations per year of approximately 1.8 and a median of 7 years of follow up visits (**Suppl. Table S1**). Most subjects presented with a GOLD symptom severity score corresponding to moderate (GOLD 2; n=32) or severe illness (GOLD 3: n=21). All serum samples were collected during a "well" visit at rest under conditions of stable illness. One subject corresponding to a data entry of 14 exacerbations per year was removed from analysis as a suspected outlier leaving *n*=66 protein expression profiles (subject ID 153).

This study is a sub-study of a larger group of patients with COPD and healthy controls to understand biological determinants of exacerbation frequency and was approved by the Institutional Review Boards of the Veterans Affairs Western New York Healthcare System and University at Buffalo. The participants gave the written consent to the study via an IRB-approved consent form. The studies in this work abide by the Declaration of Helsinki principles. The biological samples and data were accessed for research purposes between 01/07/2019 and 01/05/2024. All data was de-identified and the authors did not have access to any information that could identify individual participants during or after data collection.

2.2. Broad-spectrum LC-MS in serum

Bio-banked samples were transported to the proteomics core facility in liquid nitrogen and stored under -80°C until analysis. Prior to analysis in duplicate by the IonStar LC-MS procedure [1], samples were treated with a detergent-cocktail buffer (0.5% sodium deoxycholate, 2% SDS, 2% IGEPAL® CA-630 and protease/phosphatase inhibitor cocktail) in order to thoroughly denature the proteins and to digest efficiently the samples, followed by a surfactant-aided/precipitation-on-pellet-digestion (SOD) method to achieve quantitative and efficient recovery of peptides [2]. To achieve in-depth profiling and accurate peptide ion-current quantification, digests were separated on a 100-cm-long column with 2-µm-particles by an ultra-high-pressure chromatographic setup with optimized to deliver <15% variation of peptide signal strength [12,13]. An Orbitrap LUMOS ultra-high-field and high-resolution mass spectrometer was used to acquire quantitative ion-current signal and for protein identification. Individual data files were queried against the Swiss-Prot human protein database using the MSGF+ package. A highly stringent set of criteria (e.g. <1% protein FDR and >2 peptides per protein) was employed to ensure detection of small (as low as 20%) changes in low abundance mediators with high confidence while also maintaining excellent depth and sensitivity (>6 orders of magnitudes in protein abundance). Using this strategy has been shown to routinely quantify 5000-6000 proteins (benchmarked in human cell lysates) in up to 150 biological samples in one batch with <1% missing data, <10% quantitative variation (technical replicates) and <5% false-positive

biomarker discovery rate benchmarked against an experimental null method described previously [1].

2.3. Protein interaction network assembly

A protein-protein interaction network was computed and reduced to the component sub-network most relevant to the clinical phenotype according to the steps described in **Figure 1**. To eliminate bias, the abundance values were normalized to unit range for each species independently (Step 1). Interactions between range-adjusted protein species was estimated using mutual information (MI), based on equal bin width, as a more sensitive measure of association. Null model or background MI values were generated concurrently by a random shuffling of abundance values for each species repeated 50 times. At each iteration random background pairwise MI values were generated and stored as a null distribution to be used for significance testing (Step 2). Estimated MI values describing protein-protein interactions need not only be significant compared to random result but should also offer a signal of sufficient magnitude to be *meaningful* when compared to a non-random background. Here we apply Otsu's algorithm [14], typically used in image analysis, to identify a threshold MI value above which a protein-protein interaction can be considered foreground, and below which they can be considered background in the context of this data (Step 3). As a sanity check, the significance of all foreground MI values compared to a random result are verified using a Bonferroni correction for multiple comparisons enforcing a false discovery rate of less than 1% (FDR<0.01) (Step 4). Finally all statistically significant foreground values of MI were also required to be equal or greater in magnitude than twice the average random MI (signal to noise ratio ≥ 2) (Step 5).

The network formed by this subset of only the most important and significant protein-protein interactions was further pruned to include only those proteins most relevant to the clinical variable of interest, namely exacerbation frequency. Proteins indirectly related to exacerbation with a path length exceeding 2 were pruned at this stage of analysis. In other words, proteins that were first and second degree neighbors of exacerbation were retained (Step 6). Although, directionality of these relationships is unknown at this step, protein nodes in this subnetwork with a node degree of 1, cannot in principle be part of a closed loop. As we are focused here on subnetworks supporting closed loop regulatory behavior, only protein nodes with a node degree of 2 or greater were conserved (Step 7). All analyses were conducted using Python 3.12.3 libraries. Estimates of MI were computed using the *feature_selection.mutual_info_regression* routine in the scikit-learn library, with network analyses conducted using NetworkX 3.3. All gene ontology (GO) annotation and enrichment analyses were performed using the Panther database and analysis tools [15].

2.4. *Identifying molecular phenotype for infectious risk*

The existence of a molecular signature in the peripheral blood that might be indicative of a predisposition to repeated infections was explored by examining similarities between subjects in terms of the joint expression of these 15 co-regulating proteins (**Figure 2**). We applied a spectral clustering technique which inherently accounts for the co-expression relationships between these 15 proteins to identify 2 possible groups among the N=66 subjects. Specifically, cluster identification was conducted using *cluster.SpectralClustering* routine from the sklearn Python library, specifying that 2 clusters be identified in a latent space of 2 eigenvectors. The Calinski-Harabasz score was used as a measure of quality of separation between groups with respect to within-group dispersion. A distribution of Calinski-Harabasz scores was also computed as a null

model by assessing clusters generated by repeated random labeling of subjects. Protein co-expression profiles proposed as representative of each exacerbation phenotype were selected as those corresponding to the subject from each group with the expression profile nearest to that cluster's centroid. Once again, all computations were conducted using Python 3.12.3.

2.5. Inferring directed protein-protein co-regulatory actions

The protein-protein interactions shown in **Figure 2** are based on a measure of association and as such are devoid of direction (source to target) and mode of action (activation or inactivation of the target node by the source node). In previous work by our group we demonstrated a method for inferring directionality and mode of action by requiring that information flow through the network enable specific observed and expected dynamic behaviors [16]. First we define at apply to each node a state transition logic whereby the next activation state of a node $s_i(t+1)$ is determined by its current state and its next logical target state or *image* $s_i^*(t+1)$ as determined by the current states and actions of its upstream neighbors (**Eq. 1(a)**). In the current work all activation level of all node species is expressed as a ternary logic variable $s_i(t) \in [0,1,2]$ describing low, medium, or high activation levels. The distinct and sometimes competing actions of upstream neighbors, activated above their respective perception thresholds, are combined using a simple piecewise linear function. In this way, the actions of weak inactivators are weighed against those of strong activators, and vice versa, to recommend an increase or decrease in the activation of the target node be applied in the next iteration [17]. How quickly a node state actually achieves its target value is dependent on the update scheme and the maximum incremental change Δs_i^{max} allowed.

$$s_i^*(t) = \sum_{j \in \Omega_i = 1} w_{j,i} \; p_{j,i} \; s(t) \mathbin{/\!/} t_{j,i} \qquad \textbf{Eq. 1(a)}$$

$$S_i(t+d) = S_i(t) + \min\left([S_i^*(t+1) - S_i(t)], \Delta S_i^{max}\right) \qquad \textbf{Eq. 1(b)}$$

The direction (source to target), mode of action $p_{j,i} \in P$ (activating or inactivating the target) and the perception threshold $t_{j,i} \in T$ for each network interaction and decisional weights $w_{j,i} \in W$ dictating each node's state transition are determined by formulating and solving a Constraint Satisfaction (SAT) problem [18,19]. As a simplification, note that the decay of a node species s_i is represented implicitly by setting a basal state of zero. That is to say that should all upstream activators of a node assume a state of zero, the desired target state for that node will become zero. This parameter search problem was encoded using the open-source Constraint Programming and Modeling in Python (CPMpy) library [20] to invoke the CP-SAT solver [21] in the Google OR-Tools library [22]. CP-SAT a hybrid approach that combines finite domain propagation in Constraint Programming (CP) with the efficiency of Boolean Satisfiability (SAT) solvers. The steady state condition is articulated here as a constraint whereby the image or next target state at time *t+1* is identical to the current state at time *t*. Additional constraints described in Table 1 include network structural considerations whereby we control for the formation of source (without upstream regulators) and sink (without any downstream target) nodes as well as nodes that are devoid of upstream activators and would invariably be downregulated to their floor state (**Table 1**).

2.6. Partial validation of predicted regulatory actions

In an attempt to provide a partial validation of predicted regulatory actions linking protein species in an undirected protein-protein interaction network, we conducted searches of the Elsevier Knowledge Graph database (Elsevier, Amsterdam) [23] using the EmBiology software interface and tools. This database relies on a standardized ontology where groups of synonyms describing 1.4 million biological entities are connected by 15.7 million relationships [24]. These are extracted from peer-reviewed literature using the MedScan natural language processing (NLP) engine [25,26] across multiple sources, including over 34.5 million PubMed abstracts, 430,000 clinical trials, full-text articles from 936 Elsevier journals and 939 non-Elsevier journals, as well as databases including over 200,000 entries from BioGRID, 10,000 from DrugBank and 1.3 million Reaxys drug-target relationships. The basic recognition rules and entity terminology in this ontology are updated annually with new relationships extracted from recently published PubMed. Extracted relationships are interpreted with the source exercising *Unknown, Positive, Negative, or Undefined* effects (control actions) on the target according to standardized functional processes, such as *Direct Regulation, Regulation, Expression, Protein Modification, State Change, Quantitative Change, Molecular Transport*, and others.

3. RESULTS

3.1. A phenotypically relevant protein co-expression space

Applying the sequential selection steps described in **Figure 1** we found a candidate regulatory subnetwork consisting of 15 proteins linked by 25 significant and impactful *undirected* interactions (~12% connection density) (**Figure 2**). The strength of protein-protein associations varied between a mutual information of 0.11 and 0.54, with a median association strength of 0.16 supporting a significance of $p < 0.05$ compared to a null model of randomly sorted values (**Suppl. Table S2**). Of these 15 proteins, 13 were functionally annotated in Gene Ontology (GO) (Gene Ontology Consortium; release date 2024-01-17, DOI:10.5281/zenodo.10536401)(**Suppl. Table S3**). Subsets of genes corresponding to these proteins were significantly over-represented in GO Biological Function classes corresponding almost unanimously to detoxification of hydrogen peroxide and cellular response stress induced by reactive oxygen species (ROS) (**Suppl. Table S4(a)**). Likewise, pathways curated in the Reactome database [27] that were most significantly enriched corresponded to detoxification of ROS (R-HSA-3299685) and response to chemical stress (R-HSA-9711123)(**Suppl. Table S4(b)**).

3.2. Network-informed phenotypic groups

Protein abundance values for each subject were mapped onto an ordinal qualitative scale of Low=0, Nominal=1 and High=2 for each species independently (**Suppl. Table S5**). Spectral clustering of subjects in this discrete qualitative co-expression space of 15 proteins delineated two severity groups separated by a Calinski-Harabasz score of 15.97 ($p << 0.001$) (**Fig. 3**). The first group consisted of n=31 subjects with an average exacerbation frequency of ~2.5 episodes per year whereas the second group consisting of n=35 subjects presented with ~1.5 exacerbation episodes per year (**Fig. 4**). Though average exacerbations differed significantly ($p<0.01$), the noticeable overlap in exacerbation frequencies experienced by individuals assigned to these distinct protein co-expression classes is a reminder of the important contribution of environmental factors to the clinical outcome. Subject 3 (average of 1.47 exacerbations/year) and subject 11 (average of 2.49 exacerbations/year) were identified as representative subjects for each vulnerability group by computing the centroid of each cluster and selecting the subject with the most proximal protein abundance profile. The most important differences separating the

profiles for these representative subjects consist of changes in the expression of 3 proteins. Specifically, expression of catalase (EC 1.11.1.6), and delta-aminolevulinic acid dehydratase (ALADH) (EC 4.2.1.24) (Porphobilinogen synthase), were both increased in the higher risk group, while expression of D-dopachrome decarboxylase (EC 4.1.1.84) was decreased (**Fig. 5**). In contrast, proteins involved in innate immunity were not expressed differently across groups, e.g. Natural killer cell-enhancing factor A and B, Macrophage receptor with collagenous structure, etc…(**Suppl. Table S3**).

3.3. Robust inference of characteristic regulatory control actions

As each of the 25 undirected protein-protein interactions was translated into a pair of opposing directed network edges, the original network architecture consisted of 50 candidate directed edges of unknown mode of action (positive or negative polarity), unknown detection threshold and unknown regulatory weight. We applied the constraints summarized in **Table 1** to a solution space spanning over ~10^{55} possible combinations of parameter settings. As might be expected in such a large underdetermined parameter space, solving the corresponding SAT parameter estimation problem resulted in a large population of competing models describing low and high exacerbation profiles as persistent phenotypes equally well. However in examining a sub-sample of 100 such models we found these to be highly conserved across majority of their structure and decisional logic. Indeed, of the 150 parameters describing network structure and decisional logic, 143 are assigned the same settings in over 80 of the 100 competing models examined here, with 90 such parameter values being agreed upon unanimously (**Fig. 6**). Details of the divergence in parameter values is presented as a phylogenetic tree diagram in **Supplemental Figure S1**. A network model representative of the overall solution set could be defined as that model that is closest to the centroid of all 100 models in the parameter space. Such a representative model is presented in **Figure 7.**

In a first partial verification of predicted regulatory relationships, we used the INDRA programming library [28] to interrogate the BEL Large Corpus database as well as the PathwayCommons database [29], itself an aggregate of over 23 molecular pathway databases including KEGG Pathway, Reactome, PANTHER Pathway, BioGRID, and others. Although these manually curated pathway databases did contain relationships linking some of these 15 proteins to other known regulators or targets, none contained documented relationships linking these proteins to each other. Broadening the search to also include relationships extracted from the peer-reviewed literature using automated text mining and archived in the Elsevier Knowledge Graph, we obtain 11 documented relationships linking these 15 proteins, one of which is reported in two functional categories (i.e. Regulation and Expression). The recovery of these literature informed relationships is summarized in **Table 2**. These 11 documented relationships are supported by a total of 66 peer-reviewed publications, 48 of these confirming a reciprocal regulatory and co-expression relationship between CAT (P04040) and SOD1 (P00441) of indeterminate action. In a majority vote across the current subset of 100 models, 8 of these 11 relationships would have been recovered correctly (~73% recall), with 10 being represented in at least a subset of models. Indeed, only in the case of documented negative co-expressive mediation of PRDX2 (P32119) by CAT (P04040) did all 100 models unanimously disagree with the 2 supporting references, predicting instead that this relationship either doesn't exist or does not contribute towards explaining our 2 representative protein abundance profiles. Conversely of the 39 potential relationships so far undocumented in the Elsevier Knowledge Graph database, 24 relationships were recruited by majority vote including 17 voted upon unanimously by all

100 models to explain both low and high exacerbation frequency representative profiles as persistent signatures. In a worst-case scenario where all 17 unanimously voted relationships were in fact false positives, this would equate to a positive predictive value (PPV) of roughly 32% (if we consider majority ≥ 98 models as near unanimous). Of the remaining 15 undocumented relationships, the current set of 100 models would concur in a majority vote, including 3 by unanimous vote, that these may be non-existent or at the very least not strictly required to explain the data used here as constraints. Detailed results for this model set are presented in **Supplemental Table S3**

4. DISCUSSION

While network associations between markers expressed in peripheral blood have shown promise [30] as illness signatures in COPD, these have focused mainly on the identification of transcript sets. Only recently has this extended to an analysis of protein signaling in from frequent and infrequent exacerbators [31]. Though comparable in size to this study, the latter was conducted using plasma samples as opposed to sera which presents a more dilute and more subtle signature. Moreover, the analysis was again limited to the identification of tightly associated sets of differentially abundant proteins with no consideration of network structural characteristics. In this work we propose a simple numerical strategy to infer possible directed and threshold filtered control actions in an undirected protein-protein interaction network by applying expected dynamic behaviors, in this case homeostatic stability, to cross-sectional data. We demonstrate this strategy in an analysis a high-resolution LC-MS broad-spectrum survey of serum samples collected in a larger group of n=67 participants with stable chronic obstructive pulmonary disease (COPD) with the aim of uncovering illness-mediated changes in protein coregulation that might support increased risk of infection. In previous work by our group, an unsupervised clustering of expression profiles in the over 850 reliably identifiable protein species indicated that a subtle co-expression pattern describing as little as 2% of the variability in a subset of 160 of these proteins could delineate groups of frequent and infrequent exacerbation subjects (Calinski-Harabasz score of 160; p=0.019). The top 15 contributors to this co-expression pattern pointed to key involvement of heme scavenging as telltale Reactome pathway element [32]. Here we revisit the selection of these candidate marker proteins from the perspective of network biology where we (i) use mutual information (MI) as a more sensitive non-linear measure of interaction, (ii) apply more stringent restrictions on the magnitude and significance of association and (iii) enforce relevance to the clinical outcome based on network proximity. Our results suggest that a small network consisting of 25 interactions linking 15 soluble protein mediators may engage in formally maintaining an altered oxidative stress response in a way that is sympathetic and perhaps mutually supportive of persistent inflammation in the mucosa described in earlier work by our group [11], with both being favorable to frequent infection. Interestingly, though there was no direct overlap in these 15 network-informed protein species with those previously identified using unsupervised clustering [32], both sets support pathway activation that predominantly aligns with detoxification of reactive oxygen species. Lack of agreement in individual markers as opposed to functional sets is a long-observed theme in transcriptomic analysis [33] and has been understood as a direct consequence of the interdependency and partial redundancy of individual molecular species in supporting a common overarching biological function. Interestingly, mediators of hemolysis and detoxification have consistently been found as markers distinguishing exposure induced COPD from autoimmune asthma [34], suggesting therapeutic avenues targeting heme scavenging [35,36].

Though useful in identifying functional groups [37], protein-protein interaction networks are for the most part undirected and based on associative relationships. As such they do not describe the flow of regulatory signaling. Here we identified two subjects each with a protein expression profile representative of groups enriched in subjects with a higher or lower frequency of infectious exacerbation. As these individuals were examined during stable illness, we proposed that the corresponding protein expression profile might represent a dynamically stable steady state. An earlier analysis of sputum collected in a subset of participants (n=12) using enzyme-linked immunosorbent assays (ELISA) [11] had suggested that a common immune signaling network of documented regulatory interactions linking 10 cytokines and chemokines was capable of explaining both frequent (≤ 2 exacerbation episodes/year) and infrequent infectious phenotypes. Once again, our analysis of LC-MS protein abundance profiles representative of high and low infectious vulnerability also suggests that these could be explained by the same unifying signaling network and regulatory program. Importantly, by imposing this expected dynamic behavior on the undirected PPI network, we were able to infer not only direction but also a threshold and a mode of regulatory action for 32 of 50 potential interactions based on a majority vote across a subset of 100 competing models (~15% connection density), with 42 regulatory interactions proposed in at least a subset of models (~20% connection density). While larger simulation studies are required to more rigorously assess this approach, it was encouraging to observe that of the 11 relationships documented in the Elsevier Knowledge Graph, a broad literature and database-informed resource, 8 were recovered in a majority vote across 100 competing models with only 1 confirmed false negative. Our previous work with simulated regulatory networks indicated that low false negative rates typically aligned with high false positive rates, with a recall of ~70% typically corresponding to a positive predictive value (PPV) in the order of ~25% [38] when applying some of the more capable reverse engineering methods to perturbation time courses from the DREAM3 challenge [39] in recovering a 10-node network of 25 regulatory interactions. In this limited example, even given the worst-case scenario where all novel predicted relationships were false positives, the results presented here would still be consistent with this performance while relying on cross sectional data only. Given this important distinction, we propose that the integration of expected qualitatively described behaviors have the potential to greatly augment the usefulness of much less costly cross-sectional data sets. This framework also allows for the straightforward inclusion of general topological features as additional universal constraints [40,41], for example our requirement here that any valid network adhere to a closed loop architecture with balanced feedback (no sink nodes, no source nodes, and no strictly unipolar modulation). This has been extended in other work by our group include more complex topological features such as regulatory sub-networks or motifs [42]. Ultimately, by approaching network inference as a constraint satisfaction problem, it is possible to create hybrid initial candidate networks [24] that directly include a priori specific well-documented relationships that further constrain the choice of feasible solutions, with majority agreement in populations of models as opposed to a single unique model having the potential to deliver important improvements in both accuracy and reliability.

5. CONCLUSION

As our ability to broad-spectrum proteomic profiling continues to improve so does the availability of protein-protein interaction network signatures. Such networks are typically extracted from cross-sectional data and as such describe undirected associative rather than source-target co-regulatory relationships. As such, conventional analyses remain limited to a

graph-analytical comparison of changes in connectivity patterns in PPI networks that might arise across phenotypic groups. Inferring the source-target direction and the regulatory action of these interactions using conventional reverse engineering methods require longitudinal time course data, often with stringent sampling requirements. Given the high cost typically involved in conducting longitudinal studies, it is useful to consider alternative approaches that might allow us to infer regulatory actions from resting state profiles, if only to provide a more informed basis for designing time course experiments. Here we propose such an alternative approach where we leverage the otherwise limiting resting state observations to ask what regulatory actions would be required of these protein-protein associations to formally satisfy this steady state condition. Though focused on a specific use case, the analysis presented in this work produced candidate regulatory networks that were broadly overlapping in structure and function. Wherever available, support in the peer-reviewed scientific literature of inferred regulatory actions was also highly favourable. Importantly, the proposed method is not intended to compete with proper time course analysis methods but instead offer an extension to current graph analytical methods when only cross-sectional data is available. In this role, we propose that additional investigation of this approach is merited.

ETHICS STATEMENT

This study is a sub-study of a larger group of patients with COPD and healthy controls to understand biological determinants of exacerbation frequency and was approved by the Institutional Review Boards of the Veterans Affairs Western New York Healthcare System and University at Buffalo. The participants gave the written consent to the study via an IRB-approved consent form. The participants gave the written consent to the study via an IRB-approved consent form. The studies in this work abide by the Declaration of Helsinki principles

MANDATORY DISCLAIMER

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Department of Defense.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENT

The authors thank our veterans and their families for their service and their assistance with this research. Special thanks to the VA Western New York Healthcare System for their continued support of this biorepository (PI Sethi). This article is published with the permission of the Director of VIDO, journal series no. TBD.

FUNDING

This work was supported by Rochester Regional Health in conjunction with the US Department of Defense Congressionally Directed Medical Research Programs (CDMRP) under Peer Reviewed Medical Research Program (PRMRP) award W81XWH1910804 (Broderick - PI; Sethi - Partnering PI). VIDO receives operational funding from the Canada Foundation for

Innovation through the Major Science Initiatives Fund and from the Government of Saskatchewan through Innovation Saskatchewan and the Ministry of Agriculture.

REFERENCES

1. Shen X, Shen S, Li J, Hu Q, Nie L, Tu C, Wang X, Poulsen DJ, Orsburn BC, Wang J, Qu J. IonStar enables high-precision, low-missing-data proteomics quantification in large biological cohorts. Proc Natl Acad Sci U S A. 2018a May 22;115(21): E4767-E4776.

2. Shen S, An B, Wang X, Hilchey SP, Li J, Cao J, Tian Y, Hu C, Jin L, Ng A, Tu C, Qu M, Zand MS, Qu J. Surfactant Cocktail-Aided Extraction/Precipitation/On-Pellet Digestion Strategy Enables Efficient and Reproducible Sample Preparation for Large-Scale Quantitative Proteomics. Analytical Chemistry 2018b; 90 (17): 10350-10359

3. Athanasios A, Charalampos V, Vasileios T, Md. Ashraf G. Protein-protein interaction (PPI) network: recent advances in drug discovery. Current drug metabolism. 2017 Jan 1;18(1):5-10.

4. Manipur I, Giordano M, Piccirillo M, Parashuraman S, Maddalena L. Community detection in protein-protein interaction networks and applications. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2021 Dec 24;20(1):217-37.Sethi S, Murphy TF. Infection in the Pathogenesis and Course of Chronic Obstructive Pulmonary Disease. New England Journal of Medicine 2008;359: 2355-65.

5. Ferraro KF, Kelley-Moore JA. A half century of longitudinal methods in social gerontology: Evidence of change in the journal. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences. 2003 Sep 1;58(5):S264-70.

6. Tucker A, Garway-Heath D. The pseudotemporal bootstrap for predicting glaucoma from cross-sectional visual field data. IEEE transactions on information technology in biomedicine. 2009 Jun 12;14(1):79-85.

7. Li Y, Swift S, Tucker A. Modelling and analysing the dynamics of disease progression from cross-sectional studies. Journal of biomedical informatics. 2013 Apr 1;46(2):266-74.

8. Dutta P, Quax R, Crielaard L, Badiali L, Sloot PM. Inferring temporal dynamics from cross-sectional data using Langevin dynamics. Royal Society open science. 2021 Nov 10;8(11):211374.

9. Sethi S, Murphy TF. Infection in the Pathogenesis and Course of Chronic Obstructive Pulmonary Disease. New England Journal of Medicine 2008;359: 2355-65.

10. Mammen MJ, Sethi S. COPD and the microbiome. Respirology 2016;21: 590-9.

11. Morris MC, Richman S, Lyman CA, Qu J, Mammen MJ, Sethi S, Broderick G. "Hacking the Immune Response to Infection in Chronic Obstructive Pulmonary Disease." In 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 548-555. IEEE, 2020.

12. Nouri-Nigjeh E, Sukumaran S, Tu C, Li J, Shen X, Duan X, DuBois DC, Almon RR, Jusko WJ, Qu J. Highly multiplexed and reproducible ion-current-based strategy for large-scale quantitative proteomics and the application to protein expression dynamics induced by methylprednisolone in 60 rats. Anal Chem. 2014;86(16):8149-57.

13. Tu C, Sheng Q, Li J, Shen X, Zhang M, Shyr Y, Qu J. ICan: an optimized ion-current-based quantification procedure with enhanced quantitative accuracy and sensitivity in biomarker discovery. J Proteome Res. 2014;13(12):5888-97

14. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics, 9(1), 62-66 (1979).

15. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H. PANTHER: Making genome□scale phylogenetics accessible to all. Protein Science. 2022 Jan;31(1):8-22.

16. Sedghamiz, H., Morris, M., Craddock, T.J.A., Whitley, D., Broderick, G.: Bio-ModelChecker: Using Bounded Constraint Satisfaction to Seamlessly Integrate Observed Behavior With Prior Knowledge of Biological Networks. Frontiers in Bioengineering and Biotechnology, 7, 48 (2019a).

17. Sedghamiz, H., Morris, M., Craddock, T.J.A., Whitley, D., Broderick, G.: High-fidelity discrete modeling of the HPA axis: a study of regulatory plasticity in biology. BMC Systems Biology, 12(1):76 (2018).

18. Barták, R. Constraint Programming: In Pursuit of the Holy Grail. Theoretical Computer Science, 17(12), 555-564 (1999).

19. Page J, Oh H, Chacko T, Samuel IB, Lu C, Forsten RD, Reinhard MJ, Costanzo ME, Broderick G. Incorporating Regional Brain Connectivity Profiles into the Inference of Exposure-Related Neurobehavioral Burden in Explosive Ordnance Disposal Veterans. In International Conference on Human-Computer Interaction 2024 Jun 1 (pp. 121-139). Cham: Springer Nature Switzerland.

20. Guns, T.: Increasing modeling language convenience with a universal n-dimensional array, CPpy as python- embedded example. In: The 18th workshop on Constraint Modelling and Reformulation (ModRef 2019), University of Connecticut, Stanmford (2019).

21. Stuckey, P.J.: Lazy Clause Generation: Combining the Power of SAT and CP (and MIP?) Solving. In: Lodi, A., Milano, M., Toth, P. (eds) Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems. CPAIOR 2010. Lecture Notes in Computer Science, vol 6140. Springer, Berlin, Heidelberg (2010)

22. Cuvelier, T., Didier, F., Furnon, V., Gay, S., Mohajeri, S., Perron, L. : OR-Tools' Vehicle Routing Solver: a Generic Constraint-Programming Solver with Heuristic Search for Routing Problems. In: 24e congrès annuel de la société française de recherche opérationnelle et d'aide à la décision, ROADEF, Rennes, France. ⟨hal-04015496⟩ (2023)

23. Kamdar MR, Stanley CE, Carroll M, Wogulis L, Dowling W, Deus HF, Samarasinghe M. Text Snippets to Corroborate Medical Relations: An Unsupervised Approach using a Knowledge Graph and Embeddings. AMIA Jt Summits Transl Sci Proc. 2020 May 30;2020:288-297.

24. Page J, Moore N, Broderick G. A Computational Protocol for the Knowledge-Based Assessment and Capture of Pathologies. In Psychoneuroimmunology: Methods and Protocols 2024 Nov 16 (pp. 265-284). New York, NY: Springer US.

25. Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics. 2003 Sep 1;19(13):1699-706.

26. Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I. Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics. 2004 Mar 22;20(5):604-11.

27. Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, Haw R, Jassal B, Matthews L, May B, Petryszak R. The reactome pathway knowledgebase 2024. Nucleic acids research. 2024 Jan 5;52(D1):D672-8.

28. Bachman JA, Gyori BM, Sorger PK. Automated assembly of molecular mechanisms at scale from text mining and curated databases. Molecular Systems Biology. 2023 May 9;19(5):e11325.

29. Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, Franz M, Siper MC, Cheung M, Wrana M, Mistry H. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic acids research. 2020 Jan 8;48(D1):D489-97.

30. Obeidat M, Nie Y, Chen V, Shannon CP, Andiappan AK, Lee B, Rotzschke O, Castaldi PJ, Hersh CP, Fishbane N, Ng RT, McManus B, Miller BE, Rennard S, Par√© PD, Sin DD. Network-based analysis reveals novel gene signatures in peripheral blood of patients with chronic obstructive pulmonary disease. Respir Res. 2017 Apr 24;18(1):72.

31. Enríquez-Rodríguez CJ, Casadevall C, Faner R, Castro-Costa A, Pascual-Guàrdia S, Seijó L, López-Campos JL, Peces-Barba G, Monsó E, Barreiro E, Cosío BG, Agustí A, Gea J; BIOMEPOC group. COPD: systemic proteomic profiles in frequent and infrequent exacerbators. ERJ Open Res. 2024 Mar 25;10(2):00004-2024.

32. Everingham, E. "Unsupervised Multi-Resolution Analysis of Protein-Protein Abundance Patterns in Infectious Exacerbation of COPD." Master's thesis, Rochester Institute of Technology, 2023.

33. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set?. Bioinformatics. 2005 Jan 15;21(2):171-8.

34. Winter NA, Gibson PG, Fricker M, Simpson JL, Wark PA, McDonald VM. Hemopexin: A Novel Anti-inflammatory Marker for Distinguishing COPD From Asthma. Allergy Asthma Immunol Res. 2021 May;13(3):450-467.

35. Vallelian F, Buehler PW, Schaer DJ. Hemolysis, free hemoglobin toxicity, and scavenger protein therapeutics. Blood. 2022 Oct 27;140(17):1837-1844.

36. Yang, N., Zhang, L., Tian, D., Wang, P., Men, K., Ge, Y., & Zhang, C. Tanshinone increases Hemopexin expression in lung cells and macrophages to protect against cigarette smoke-induced COPD and enhance antiviral responses. Cell Cycle, 2022; 22(6), 645–665.

37. Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, Schneider R, Bagos PG. Using graph theory to analyze biological networks. BioData mining. 2011 Dec;4:1-27.

38. Vashishtha S, Broderick G, Craddock TJ, Fletcher MA, Klimas NG. Inferring broad regulatory biology from time course data: Have we reached an upper bound under constraints typical of in vivo studies?. PloS one. 2015 May 18;10(5):e0127364.

39. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. PloS one. 2010 Feb 23;5(2):e9202.

40. Greenfield A, Hafemeister C, Bonneau R. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinformatics. 2013 Apr 15;29(8):1060-7.

41. Petri T, Altmann S, Geistlinger L, Zimmer R, Küffner R. Addressing false discoveries in network inference. Bioinformatics. 2015 Sep 1;31(17):2836-43.

42. Page JM, Morris MC, Skuse G, Broderick G. Refinement of Biological Regulatory Graphs using Functional Enrichment. In2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2022 Dec 6 (pp. 3580-3586). IEEE.

Table 1. Statement of core constraints in a constraint satisfaction problem.

Role	Constraint
Current state *S(t)* constrained to observed steady state C_l	$s_{i,l}(t) == c_{i,l}$ for *i=1:K state variables* *l = 1:L* observed steady states
Unnecessary edge assigned a polarity of zero implies it is also assigned a weight of zero (and vice versa)	$p_{j,i} == 0 \rightarrow w_{j,i} == 0$ $w_{j,i} == 0 \rightarrow p_{j,i} == 0$
Next target state *S* (t+1)* is constrained to be current state *S(t)* at each constraint *l*	$s^*_{i,l}(t+1) == s_{i,l}(t)$
Range constraints on node state as well as edge weight, polarity and threshold	$s_i \in [0,1,2]$ $w_{j,i} \in [0,1,2]$ $p_{j,i} \in [-1,0,1]$ $t_{j,i} \in [1,2]$
Structural constraint avoiding formation of source (zero indegree)or sink (zero outdegree) nodes	$\sum_{j \in \Omega_i = 1} w_{j,i} \neq 0$
Controlling for only down regulation from upstream nodes	$\lvert p_{j,i} \in P : p_{j,i} > 0 \rvert \neq 0$

Table 2. Summary of constraint-based recovery of 11 relationships extracted from the literature using automated text mining (number of supporting references).

Source ID	P04040	P00441	P04040	P32119	P04040	P30046	Q06830	P04040	P13716	Q06830	P00441
Target ID	P00441	P04040	P32119	P04040	P30046	P04040	P04040	P13716	P04040	P00441	Q13228
Model predicted											
Polarity = -1 (Negative)	0	1	0	0	100	2	100	0	1	100	0
Polarity = 0 (unsupported)	1	98	100	0	0	98	0	0	1	0	0
Polarity = +1 (Positive)	99	1	0	100	0	0	0	100	98	0	100
Embiology											
Source Name	CAT	SOD1	CAT	PRDX2	CAT	DDT	PRDX1	CAT	ALAD	PRDX1	SOD1
Target Name	SOD1	CAT	PRDX2	CAT	DDT	CAT	CAT	ALAD	CAT	SOD1	SELENBP1
Direct Regulation				Unknown (5)							
Regulation	Unknown (26)	Unknown (14)			Unknown (3)	Unknown (3)	Negative (1)	Positive (1)	Positive (1)		
Expression	Unknown (8)		Negative (2)								Positive (1)
Modification										Unknown (1)	
Majority agreement	Y	*N*	*N*	Y	Y	*N*	YY	YY	YY	Y	YY
At least one model?	Y	Y	*N*	Y	Y	Y	YY	YY	YY	Y	YY

CAPTIONS

Figure 1. Overview of data flow and analysis steps involved in the reverse engineering of a closed-loop 15-node subnetwork consisting of significant foreground protein-protein interactions most relevant to infectious exacerbation.

Figure 2. Protein-protein association network (mutual information MI) consisting of 15 protein nodes linked through 25 foreground, impactful, significant MI directly or indirectly relevant to exacerbation frequency (i.e. ≤ 2 degrees of separation). Line width is proportional to the MI value. Nodes are labelled with their UNIPROT protein identifier.

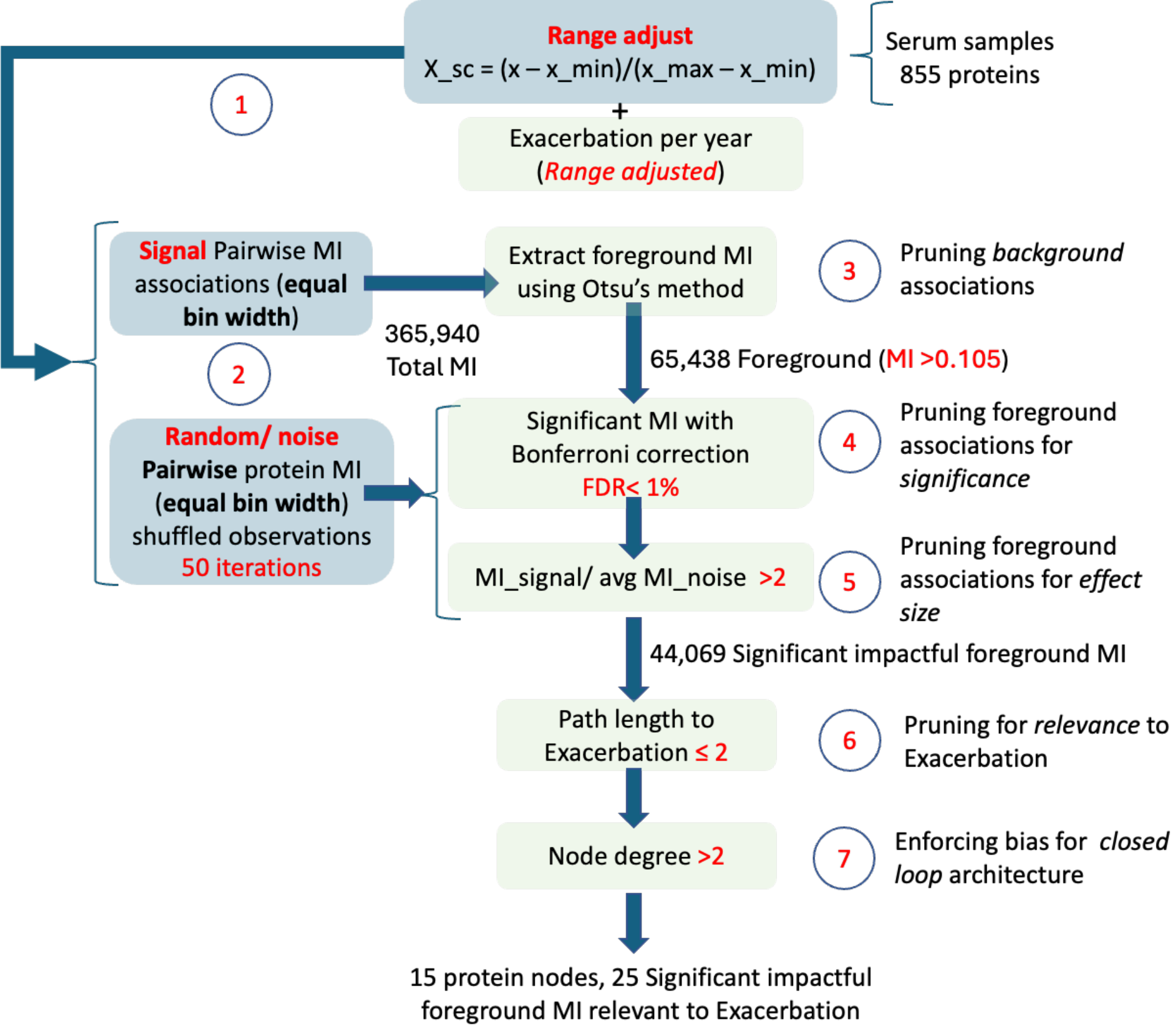
Figure 3. Delineation of subjects into 2 clusters based on the spectral clustering of co-expression patterns in the 15 characteristic proteins shown in Fig. 1. Clusters are separated by a Calinski-Harabasz score of 15.97 ($p<<0.001$ compared to a mean score of ~0.95 for 100 random labeling assignments).

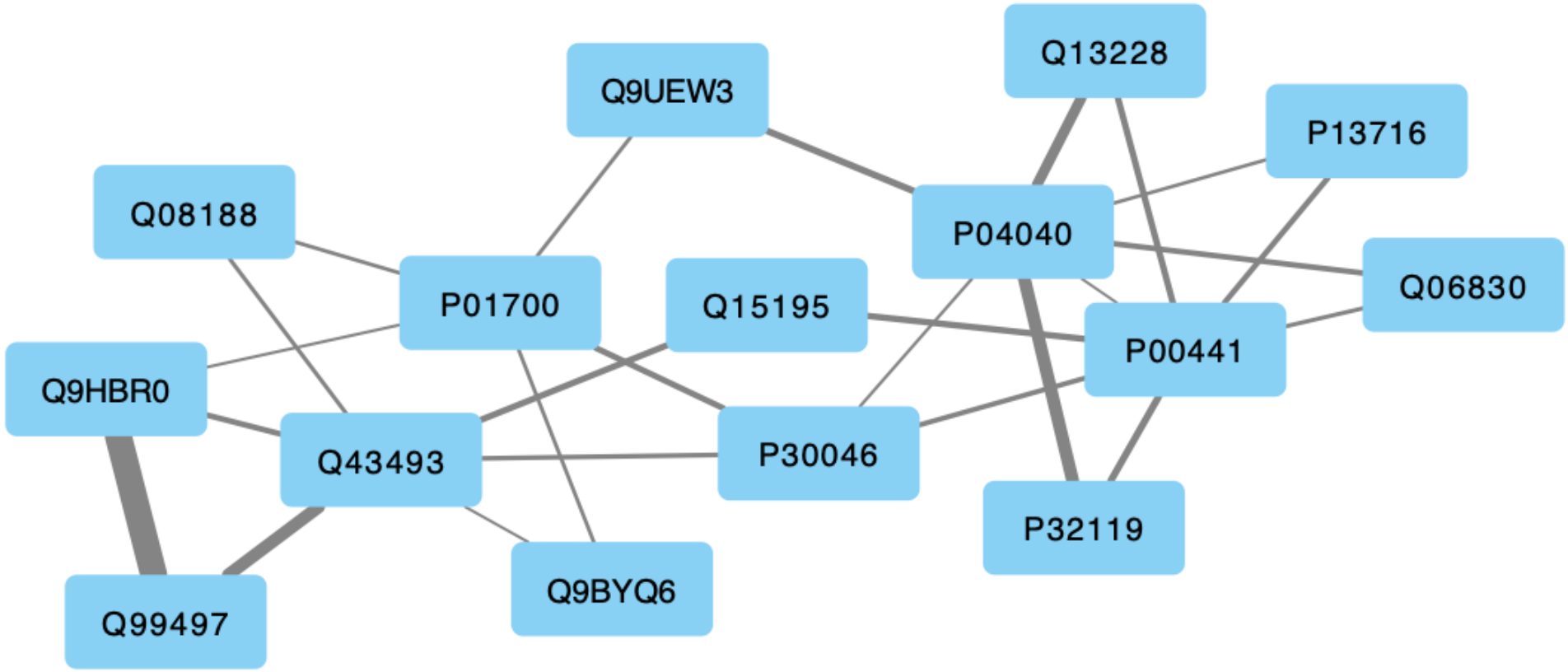
Figure 4. Distribution of exacerbation frequencies for individuals assigned to each protein co-expression cluster from Fig. 3 shows overlapping but significantly different severities on average ($p<0.01$).

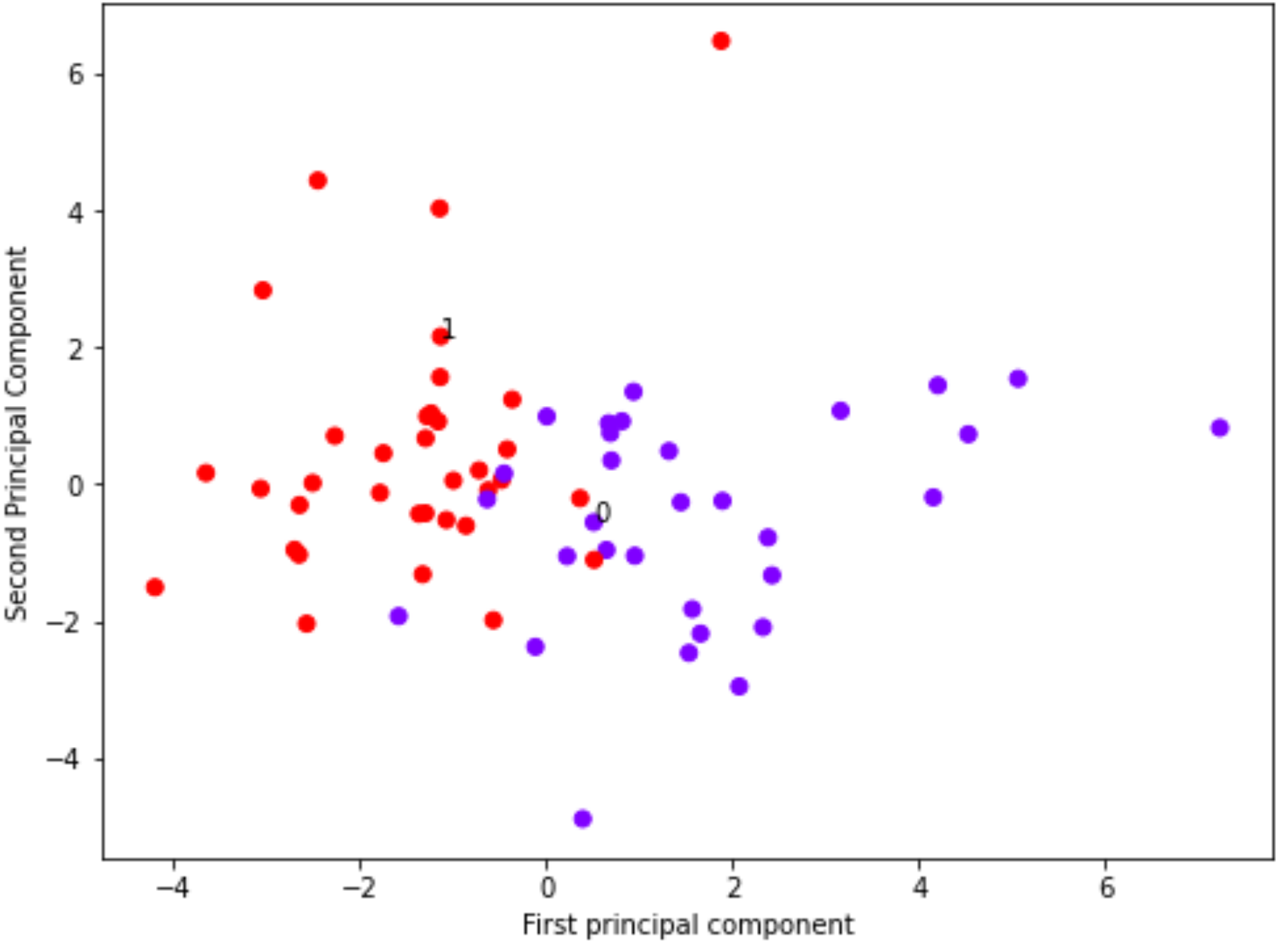
Figure 5. Discrete ordinal scale (Low, Nominal, High) expression profiles for the subjects #3 (C1; 1.47 exacerbations/ year) and #11 (C0, 2.49 exacerbations/ year) closest to the centroid in each severity group.

Figure 6. Evolution in the granularity of regulatory control networks in a subset of 100 competing models describing protein abundance profiles in subjects representative of low and high exacerbation frequency as persistent phenotypes. Parameter values describing network structure and regulatory logic are broadly conserved in over 80% of these models. The least consistent predictions involve possible reciprocal co-regulation of TGM3 (Q08188) and TGOLN2 (Q43493)(relationship 35), so far undocumented.

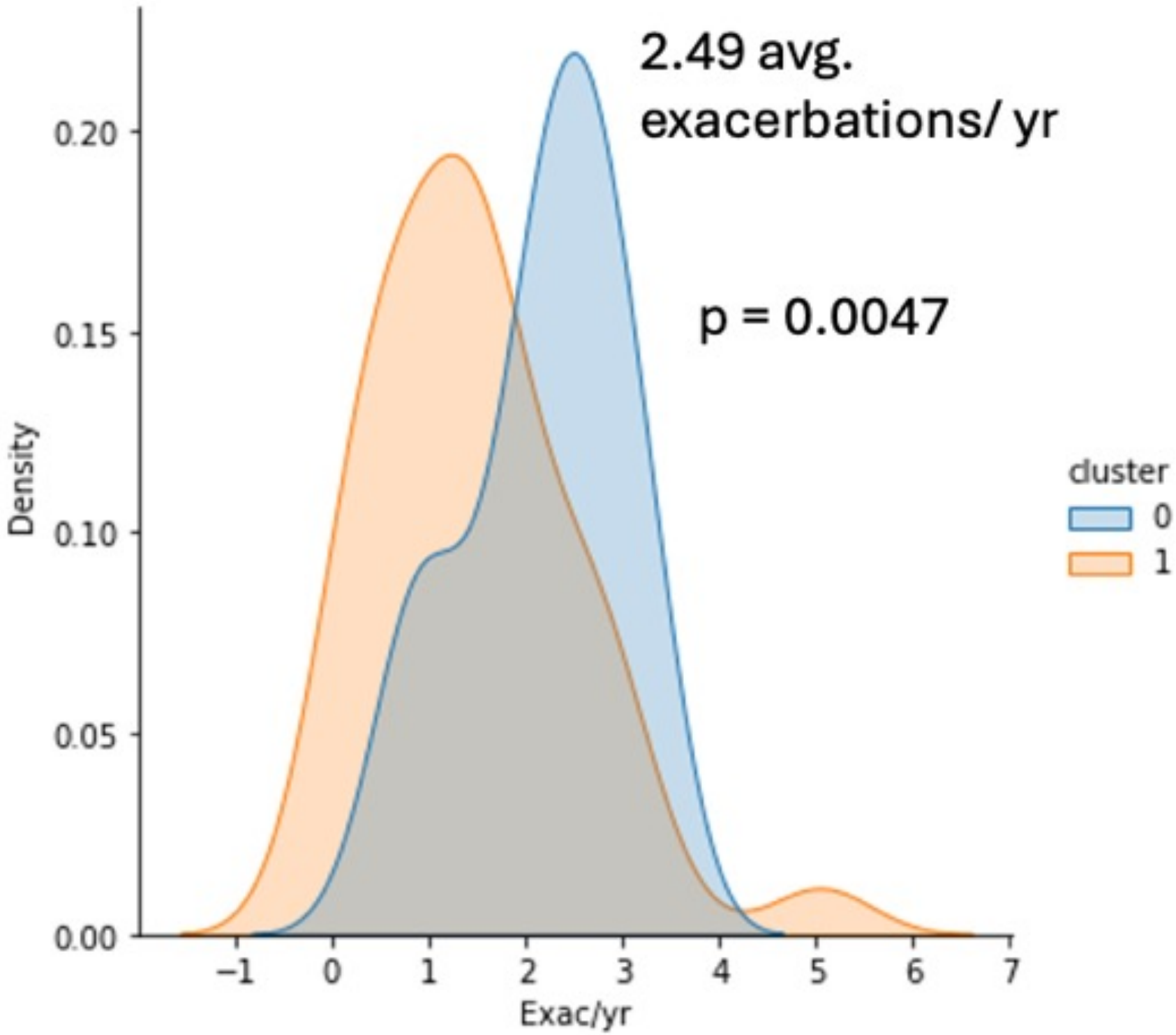
Figure 7. A representative regulatory network, defined as that network closest to the centroid in parameter space, that explains both low and high frequency exacerbation phenotypes as persistent conditions. Green arrows (and red T bars) represent activation (or inactivation) of a target node by an upstream mediator, with line width being proportional to the decision weight (thin w=1, thick w=2).

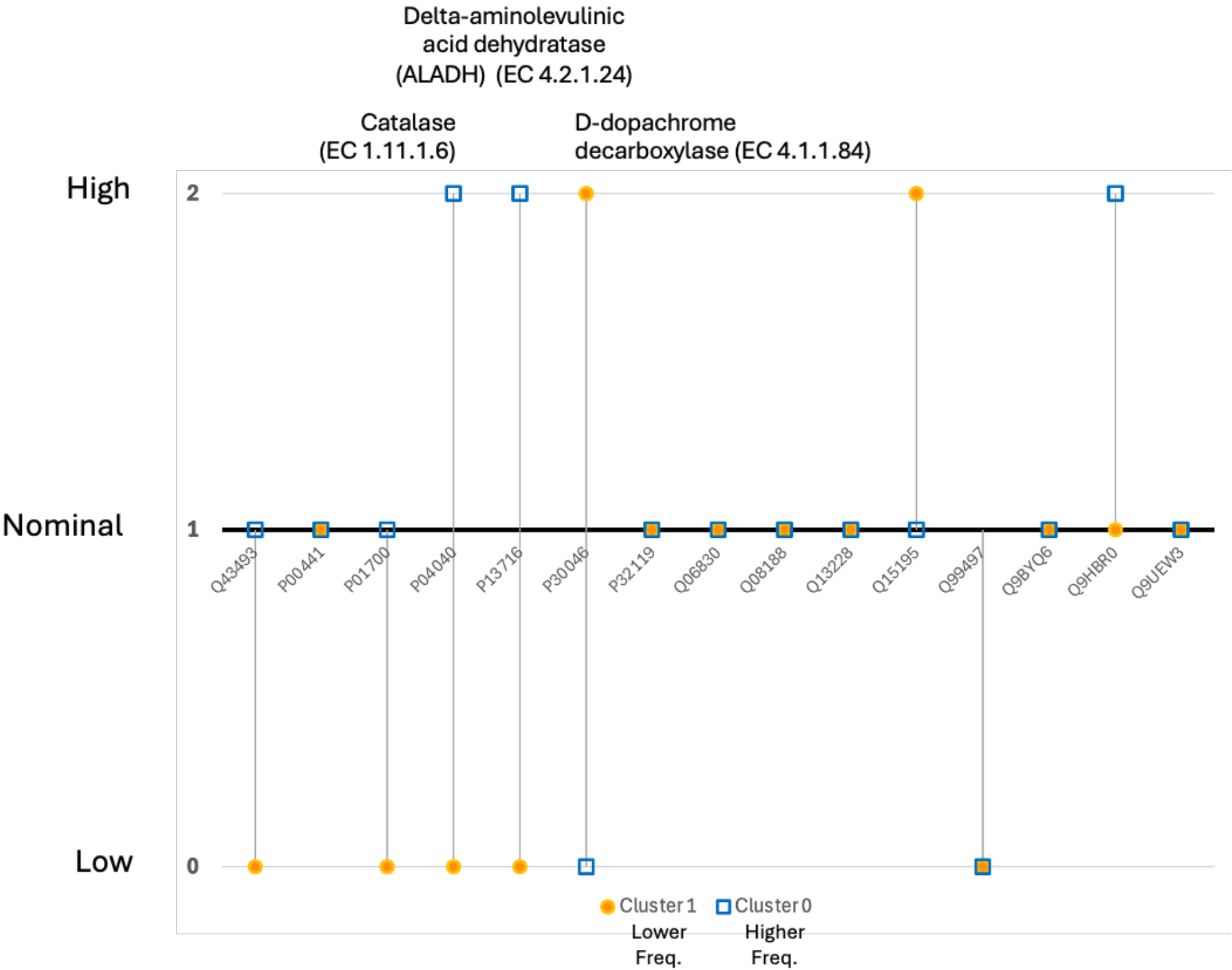

Range adjust
X_sc = (x – x_min)/(x_max – x_min)
Serum samples
855 proteins
1
+
Exacerbation per year
(Range adjusted)
Signal Pairwise MI associations (equal bin width)
Extract foreground MI using Otsu's method
3
Pruning background associations
365,940
Total MI
2
65,438 Foreground (MI >0.105)
Random/ noise
Pairwise protein MI (equal bin width) shuffled observations
50 iterations
Significant MI with Bonferroni correction
FDR< 1%
4
Pruning foreground associations for significance
MI_signal/ avg MI_noise >2
5
Pruning foreground associations for effect size
44,069 Significant impactful foreground MI
Path length to Exacerbation ≤ 2
6
Pruning for relevance to Exacerbation
Node degree >2
7
Enforcing bias for closed loop architecture
15 protein nodes, 25 Significant impactful foreground MI relevant to Exacerbation

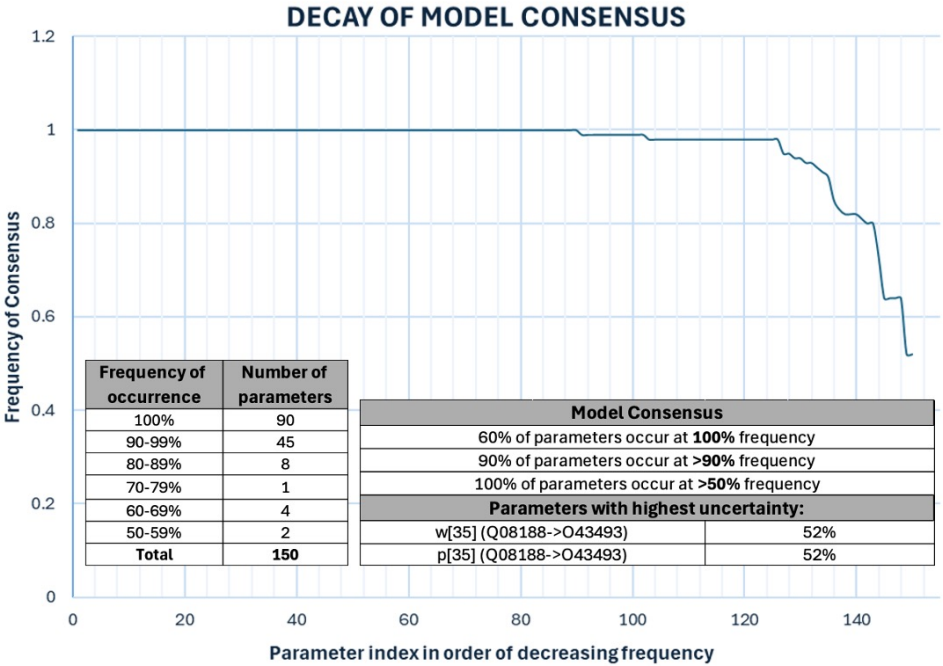

Q13228
Q9UEW3
P13716
Q08188
P04040
Q06830
P01700
Q15195
P00441
Q9HBR0
Q43493
P30046
P32119
Q9BYQ6
Q99497


6
4
2
0
−2
−4
Second Principal Component
1
0
−4
−2
0
2
4
6
First principal component

1.47avg.
Exacerbations/ yr

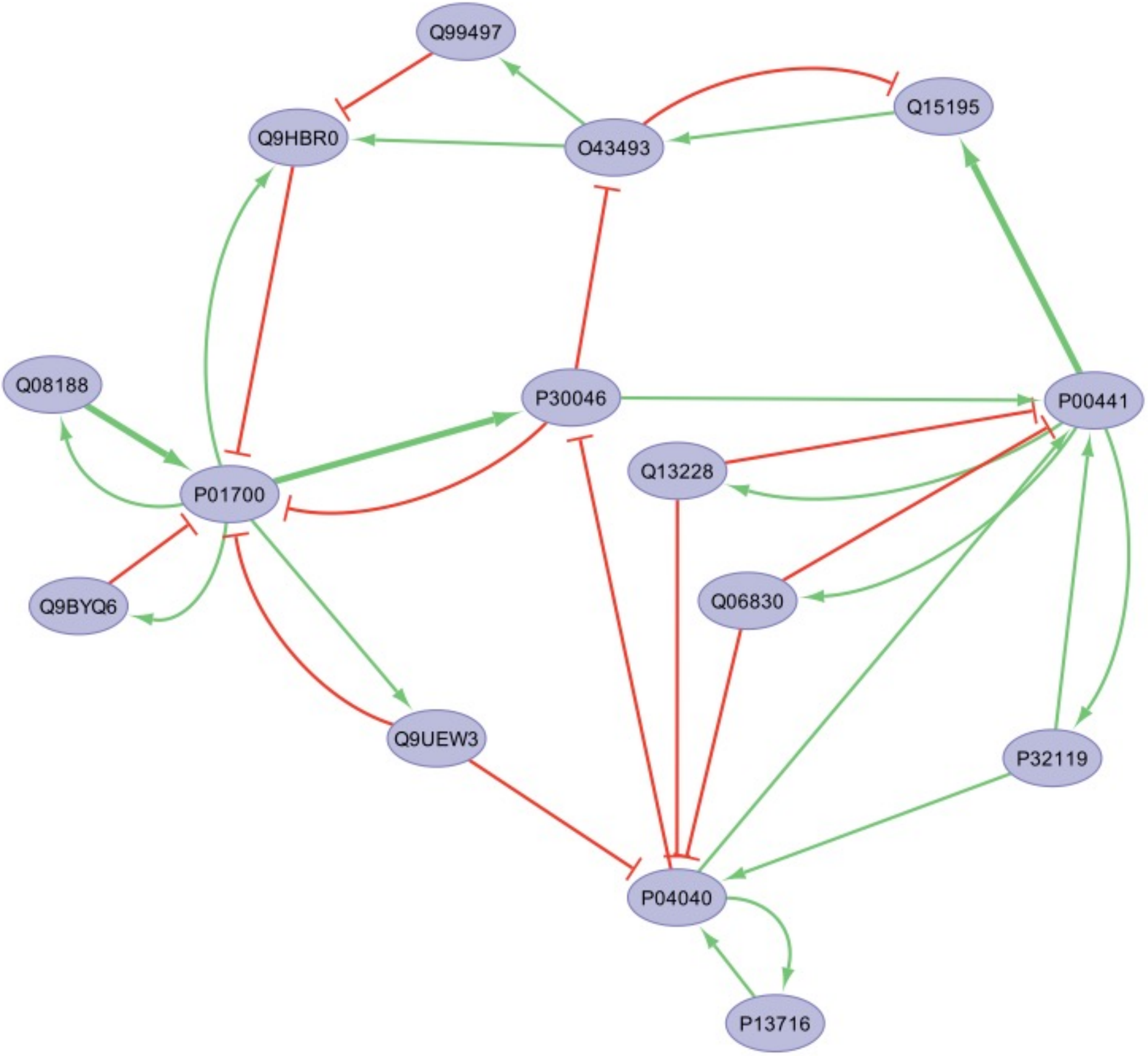


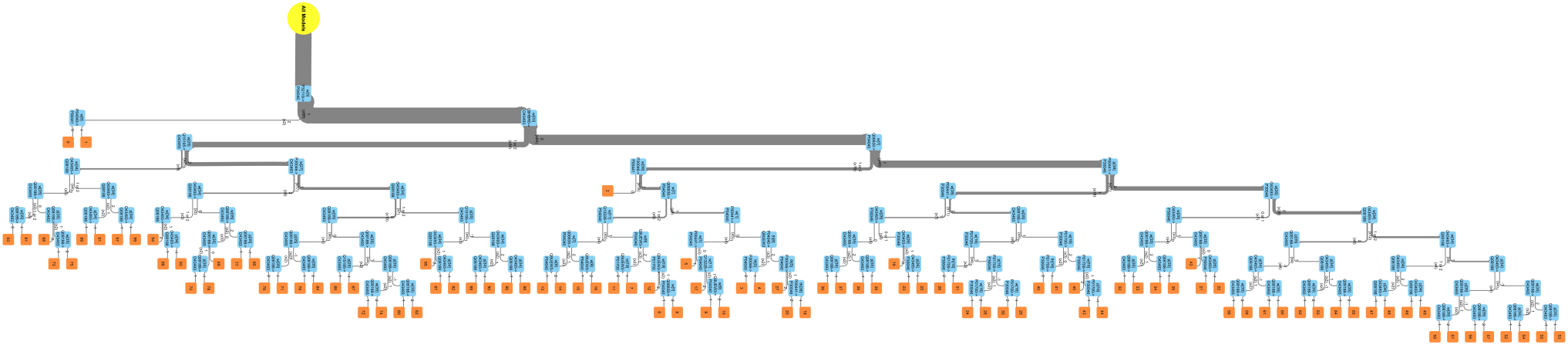

Delta-aminolevulinic acid dehydratase (ALADH) (EC 4.2.1.24)
Catalase (EC 1.11.1.6)
D-dopachrome decarboxylase (EC 4.1.1.84)
High
Nominal
Low
2
1
0
Q43493
P00441
P01700
P04040
P13716
P30046
P32119
Q06830
Q08188
Q13228
Q15195
Q99497
Q9BYQ6
Q9HBR0
Q9UEW3
Cluster 1 Lower Freq.
Cluster 0 Higher Freq.



Frequency of occurrence	Number of parameters
100%	90
90-99%	45
80-89%	8
70-79%	1
60-69%	4
50-59%	2
Total	**150**

Model Consensus	
60% of parameters occur at **100%** frequency	
90% of parameters occur at **>90%** frequency	
100% of parameters occur at **>50%** frequency	
Parameters with highest uncertainty:	
w[35] (Q08188->O43493)	52%
p[35] (Q08188->O43493)	52%


Q99497
Q15195
Q9HBR0
O43493
Q08188
P30046
P00441
Q13228
P01700
Q9BYQ6
Q06830
Q9UEW3
P32119
P04040
P13716



Supplemental Figure S1. Tree diagram decomposition in the parameter space describing structural and decisional logic similarities and differences in a subset of 100 regulatory network models.