Automating parameter selection to avoid implausible biological pathway models

Chris S. Magnano [1,2] and Anthony Gitter [1,2,3,*]

[1] Department of Computer Sciences, University of Wisconsin-Madison, Madison, Wisconsin,
[2] Morgridge Institute for Research, Madison, Wisconsin, and
[3] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin.

[*] Send correspondence to AG: gitter@biostat.wisc.edu

Abstract

A common way to integrate and analyze large amounts of biological "omic" data is through pathway reconstruction: using condition-specific omic data to create a subnetwork of a generic background network that represents some process or cellular state. A challenge in pathway reconstruction is that adjusting pathway reconstruction algorithms' parameters produces pathways with drastically different topological properties and biological interpretations. Due to the exploratory nature of pathway reconstruction, there is no ground truth for direct evaluation, so parameter tuning methods typically used in statistics and machine learning are inapplicable. We developed the pathway parameter advising method to tune pathway reconstruction algorithms to minimize biologically implausible predictions. We leverage background knowledge in pathway databases to select pathways whose high-level structure resembles that of manually curated biological pathways. At the core of this method is a graphlet decomposition metric, which measures topological similarity to curated biological pathways. In order to evaluate pathway parameter advising, we compare its performance in avoiding implausible networks and reconstructing pathways from the NetPath database with other parameter selection methods across four pathway reconstruction algorithms. We also demonstrate how pathway parameter advising can guide construction of an influenza host factor network. Pathway parameter advising is method-agnostic; it is applicable to any pathway reconstruction algorithm with tunable parameters. Our pathway parameter advising software is available on GitHub at `https://github.com/gitter-lab/pathway-parameter-advising`.

1 Introduction

Network analysis can integrate and analyze large amounts of biological "omic" data from genomic, transcriptomic, proteomic, or metabolomic assays (Goh *et al.*, 2012; Furlong, 2013). Placing omic data in a network context allows for the discovery of key members of a process that may be missed from a single data source and functional summarization for hypothesis generation and other downstream analyses.

Although biological pathway enrichment can be used to interpret omic data, pathways in curated databases are often incomplete and contain proteins or genes that are not involved in a particular biological context (Köksal *et al.*, 2018). Thus, it is often preferable to infer a customized subnetwork specific to an experimental dataset starting from all known protein interactions, referred to as the interactome. We refer to this problem as pathway reconstruction: using condition-specific input omic data to select nodes and edges from a generic background network that represent some process or cellular state. Pathway reconstruction differs from module detection (Choobdar *et al.*, 2019), which divides a network into functional units or clusters. It is also distinct from network propagation (Cowen *et al.*, 2017), which identifies relevant regions of a larger network but typically does not select specific edges within that region.

Existing pathway reconstruction methods are based on diverse strategies such as combinatorial optimization problems (Tuncbag *et al.*, 2016; Scott *et al.*, 2006; Yosef *et al.*, 2011), shortest paths (Ritz *et al.*, 2016), enrichment analysis (Cerami *et al.*, 2010), network flow (Basha *et al.*, 2013; Goldberg and Tarjan, 1990; Komurov *et al.*, 2010), and other graph theory algorithms. These methods also take in a variety of inputs. Some, such as the Prize-Collecting Steiner Forest (PCSF) algorithm (Tuncbag *et al.*, 2016; Kedaigle and Fraenkel, 2018), accept scores for the biological entities of interest. Other methods, such as PathLinker (Ritz *et al.*, 2016), require the inputs to be split into start-points (sources) and end-points (targets) for the pathway. Despite the different optimization strategies and inputs, pathway reconstruction

algorithms almost always require the user to set parameters. Adjusting the parameters can produce pathways with drastically different topological properties and biological interpretations. For instance, in Figure 1 both pathways were created with the same PCSF algorithm and the same influenza host factor screen data (Section 2.4.3); they only differ in the parameters used. The pathway on the right is reasonably sized and can be interpreted and summarized for downstream analysis. The pathway on the left, however, includes over 7000 nodes and would impractical to interpret or analyze.

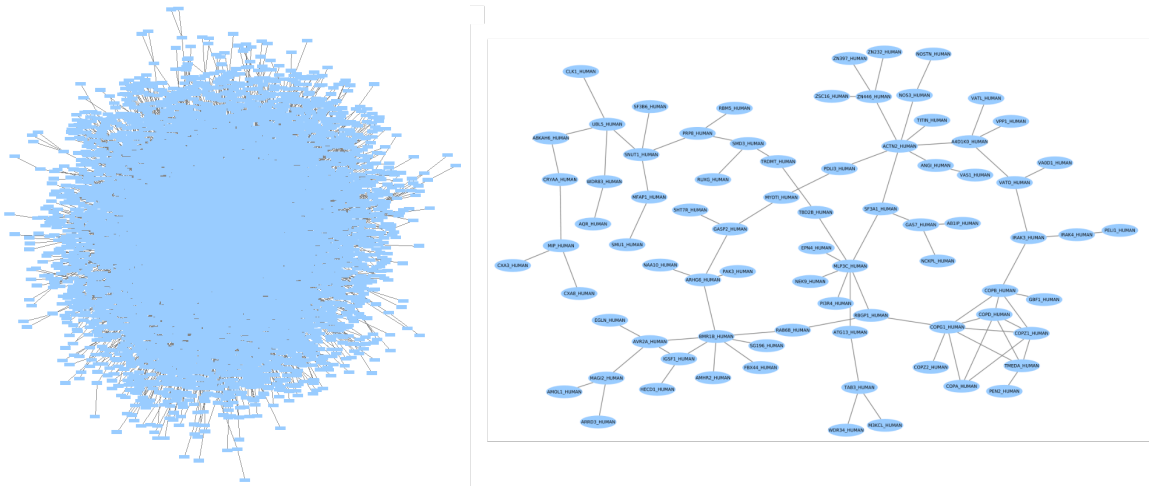

Fig. 1. Influenza host factor pathways created using PCSF from RNA interference (RNAi) screens (Section 3.4), here showing the largest connected components from ensembling the bottom 100 ranked pathways (left) and the top 100 ranked pathways (right). The only difference in creating the networks was the range of PCSF parameter values.

An open challenge is how to configure these critical pathway reconstruction parameters in a manner that is objective and applicable across diverse types of algorithms. Existing approaches tend to be informal and ad hoc, and most best practices in parameter tuning are not applicable to pathway reconstruction. The simplest way to choose a parameters would be to use default values. However, a single parameter setting cannot work for all datasets. The number of input proteins, genes, or metabolites varies based on the experiment, and the effects of input size can be unpredictable for different pathway reconstruction algorithms. For instance, if PathLinker is run with fixed parameters, increasing the number of source and target nodes will often result in a smaller final pathway, which is not necessarily what a user would intend. In addition, some methods are commonly run repeatedly and combined into an ensemble network (MacGilvray *et al.*, 2018; Budak *et al.*, 2015; Khurana *et al.*, 2017), which requires multiple parameter settings.

Parameter tuning methods from supervised learning are also a poor match. There is no ground truth for supervised parameter tuning, and unsupervised cross-validation is ineffective (Section 3). In addition, the objective functions of pathway reconstruction methods only approximate biologically meaningful graph topologies and typically have no probabilistic likelihood. Thus, their values cannot be compared between different parameter settings, and statistical model selection criteria such as the Akaike information criterion (Akaike, 1998) and the Bayesian information criterion (Schwarz, 1978) are not applicable. For instance, the PCSF objective function can be arbitrarily increased by changing the parameter values.

Given the lack of objective, quantitative methods for tuning, parameter settings are often chosen by manual inspection or informal heuristics. ResponseNet recommends choosing parameters that recover at least 30% of inputs while minimizing low confidence edges (Yeger-Lotem *et al.*, 2009). PCSF recommends choosing pathways robust to small random input variation or matching the average degree of input nodes and non-input nodes (Kedaigle and Fraenkel, 2018). We show in Section 3 that these heuristics can perform poorly in practice.

Biologists often have intuition about which pathways are unrealistic or impractical for downstream analysis, such as the 7000 node pathway in Figure 1 or subnetworks with unusual degree distributions. Pathway reconstruction is typically an exploratory analysis used to summarize the input data and generate hypotheses leading to further experiments. In this context, it is important to avoid implausible and uninterpretable pathway topologies. Therefore, parameter tuning should not focus on traditional notions of accuracy but instead formalize how *useful* generated pathways are to biologists. Our major contribution is providing a formal approach based on graph topology that quantifies this biological intuition and can be used to optimize pathway reconstruction in an objective manner.

One framework for finding optimal parameters in an uncertain setting is parameter advising (Kececioglu and DeBlasio, 2013; DeBlasio and Kececioglu, 2015, 2017), which was originally developed for multiple sequence alignment. Parameter advising can be used to adapt the parameter tuning framework in settings where no ground truth tuning set exists. However, parameter advising requires a means to estimate the accuracy of a model with a given set of parameters. Because pathway reconstruction is an exploratory analysis, there is no formal notion of accuracy. We overcome this limitation by leveraging background knowledge to create a ranking metric that prefers pathways in topological agreement with reference pathways. Our parameter tuning method, pathway parameter advising, uses the parameter advising framework in combination with

a distance metric based on graphlet decomposition to measure similarity between generated pathways and pathways from curated databases. Pathway databases may be imperfect and incomplete, but they reflect models that the expert curators consider to be biologically plausible. Only measuring the topology of a generated network means that pathway parameter advising is also method agnostic. Pathway parameter advising can tune the parameters of any pathway reconstruction method.

2 Methods

2.1 Pathway Parameter Advising

Pathway parameter advising is based on the parameter advising framework (DeBlasio and Kececioglu, 2015). A parameter advisor consists of two parts: a set of candidate parameter settings S and an accuracy estimator E. The parameter advisor evaluates each candidate parameter setting in S using E to estimate the optimal parameter set. In order to adapt parameter advising to the pathway reconstruction task, we must choose a function E that can estimate the quality of a generated pathway. While we do not have a direct way to define what criteria an optimal solution satisfies, we do have access to pathways which match biologist intuition of what a biological pathway should look like. Curated pathway databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), Reactome (Fabregat *et al.*, 2018), and NetPath (Kandasamy *et al.*, 2010), contain pathways that have been compiled by biologists. Therefore, we can construct our estimator around these curated pathways. This leads to the key assumption of pathway parameter advising: *generated pathways more topologically similar to manually curated pathways are more useful to biologists*.

Our parameter tuning approach requires the inputs to the pathway generation algorithm, a set of candidate parameter settings, and a set of pathways from a curated pathway database. Pathway generation algorithms' input typically consists of an interactome, such as STRING (Szklarczyk *et al.*, 2019), and a set of biological entities of interest, such as genes or proteins. We refer to the pathways created by the algorithm as *generated* pathways and the curated pathways as *reference* pathways. Pathway parameter advising uses a graphlet distance-based estimator E to score each generated pathway's similarity to the reference pathways. It uses these scores to return a ranking of the generated pathways (or their respective parameter settings).

Pathway parameter advising is designed to be method-agnostic. It can be run with any pathway reconstruction algorithm that generates pathways and has user-specified parameters. Currently, pathway parameter advising is designed to examine undirected graphs, and directed graphs are converted to be undirected while tuning.

2.1.1 Graphlet Decomposition

In order to topologically compare generated and reference pathways, we first decompose all pathways into their graphlet distributions. A graphlet is a subgraph of a particular size within a network. The concept of graphlets is similar to that of network motifs (Milo *et al.*, 2002). However, network motifs typically refer to graphlets that appear in a network significantly more often than expected by chance.

Original work with graphlets only considered connected graphlets to better capture local topology (Pržulj *et al.*, 2004). However, we use both connected and disconnected subgraphs, thus allowing all possible combinations of nodes in a pathway to be considered a graphlet. This allows our parameter ranking to capture global topological properties such as pathway size in addition to local topology. One disadvantage of disconnected graphlet counts is that the counts of disconnected graphlets, such as the graphlet containing 4 unconnected nodes, grow at a much faster rate than those of connected graphlets in sparse networks. However, this disadvantage does not adversely affect our ranking metric (Section 3.1).

Pathway parameter advising uses the parallel graphlet decomposition library (Ahmed *et al.*, 2015) to calculate counts of all graphlets up to size 4 in a pathway. This constitutes 17 possible graphlets (Figure 2). We convert these counts into frequencies and represent each pathway by a vector of 17 values between 0 and 1. This vector, referred to as the graphlet frequency distribution, summarizes the topological properties of a pathway, allowing us to quantify topological similarity.

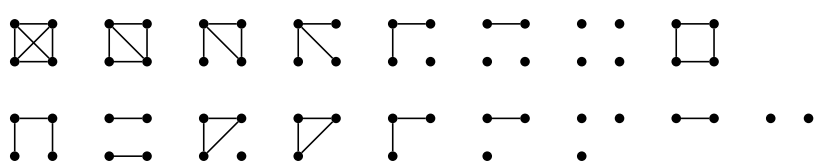

Fig. 2. Pathways are decomposed into these 17 graphlets for graphlet frequency distance calculations.

2.1.2 Distance Calculation

To calculate the topological distance between two pathways, we take the pairwise distance of their graphlet frequency distributions. For pathways G and H, we denote their frequencies of graphlet i as $F_i(G)$ and $F_i(H)$, scalars between 0 and

1. We then define the graphlet frequency distance $D(G, H)$ as

$$D(G, H) = \sum_{i=1}^{17} |F_i(G) - F_i(H)|$$

This differs from relative graphlet frequency distance, which log transforms and scales the raw graphlet counts (Pržulj *et al.*, 2004). We considered other graphlet-based metrics such as a variation of relative graphlet frequency distance and graphlet correlation distance (Yaveroğlu *et al.*, 2014) but found that they performed worse in our preliminary analyses (Figure S1).

Table 1. The 4 pathway reconstruction methods and the parameters tuned for each.

Algorithm	Description	Parameters and Default Values	Range Tested
PathLinker (Ritz *et al.*, 2016)	Connects receptors to transcriptional regulators via weighted k-shortest paths.	**k** (default=100) : Number of shortest paths.	1-1000
NetBox (Cerami *et al.*, 2010)	Hierarchically constructs network using hypergeometric test.	**p** (default=0.05) : p-value threshold for adding an edge	0-1
Prize-Collecting Steiner Forest (Tuncbag *et al.*, 2013, 2016)	Assigns prizes to nodes and costs to edges; solves for highest scoring subnetwork with message passing algorithm.	$\boldsymbol{\beta}$ (default=1): relative weight of the node prizes versus edge costs	0-5
		$\boldsymbol{\omega}$ (default=6): cost of adding an additional tree to the solution network	0-10
		$\boldsymbol{\mu}$ (default=0): degree penalty	0-1
Minimum-Cost Flow (Goldberg and Tarjan, 1990)	Assigns edges costs and nodes as sources and targets of flow. Finds the network that satisfies flow constraints for the least cost.	**f** (default=10): amount of flow pushed through the network from sources to targets	1-50
		c (default=1): edge flow capacity	1-25

2.1.3 Ranking Parameters

After calculating the graphlet frequency distribution for each generated and reference pathway, we can rank parameter settings by their mean graphlet frequency distance to the reference pathways to get E. When calculating this aggregate distance, we only consider the 20% closest reference pathways to the generated pathway. The threshold choice has little impact on the parameter ranking (Figure S2). It is motivated by not requiring a generated pathway to be similar to every reference pathway, but instead similar to at least some reference pathways. Thus, a pathway G's score $E(G)$ is calculated as

$$E(G) = \sum_{r \in R_{top}} D(G, r)$$

where R_{top} is the set of the 20% closest reference pathways to G. The pathways, or equivalently the parameters used to generate those pathways, are sorted by $E(G)$ in descending order. Once the final ranking is created, the top generated pathway can be used for downstream analysis. Alternatively, for pathway reconstruction algorithms where it is common to create an aggregate network of multiple pathways, the top n pathways can be combined in an ensemble.

2.2 Pathway Reconstruction Methods

Pathway reconstruction algorithms were chosen to have a wide range of methodologies, from NetBox's statistical test to PathLinker's weighted shortest paths algorithm. We used the following 4 methods for our implausible pathway detection and reconstruction experiments. These methods and the parameters tested are summarized in Table 1.

PathLinker: PathLinker (Ritz *et al.*, 2016) constructs pathways based on a weighted k shortest paths algorithm. It finds paths between sets of receptors and transcriptional regulators, similar to the source and target nodes in minimum-cost flow. It is controlled by the parameter k, which defines how many paths to return in the final network. We varied k from 1 to 1000 in increments of 1. We used PathLinker version 1.1 for all analyses.

NetBox: NetBox (Cerami *et al.*, 2010) hierarchically constructs networks from a set of input nodes. At each iteration, it searches for nodes that connect two nodes in the current network. It then chooses to add these linker nodes to the network based on the results of a hypergeometric statistical test comparing the degree of the linker node to how many nodes in the pathway it connects. NetBox is controlled by the parameter p, a p-value cut-off, which sets the threshold for whether or not linker nodes should be included. We varied p from 0 to 1 on a log scale from 1×10^{-30} to 1 in increments of half an order of magnitude, giving a total of 60 steps. We used NetBox version 1.0 for all analyses.

Prize-Collecting Steiner Forest: In PCSF (Tuncbag *et al.*, 2013, 2016; Kedaigle and Fraenkel, 2018), nodes are assigned prizes and edges are given costs. The optimal subnetwork, which is found via a message-passing algorithm (Bailly-Bechet *et al.*, 2011), is the pathway F consisting of nodes N_F and edges E_F that best balances collected prizes versus cumulative

edge costs according to the following function:

$$\underset{F}{\operatorname{argmin}} \sum_{n \notin N_F} (\beta \cdot p(n) - \mu \cdot d(n)) + \sum_{e \in E_F} c(e) + \omega \cdot \kappa$$

where $p()$ is the positive prize for each node, $d()$ is a node's degree, $c()$ is the cost of each edge, and κ is the number of connected components in the pathway. The optimal subnetwork is always a tree- or forest-structured graph. We varied 3 PCSF parameters: β, which controls the relative weight of the node prizes versus edge costs was varied from 0 to 5 in increments of 0.5; μ, which affects the penalty for high-degree nodes was varied from 0 to 1 in increments of 0.1; and ω, which controls the cost of adding an additional tree to the solution network was varied from 0 to 10 in increments of 1. We used version 1.3 of the msgsteiner message-passing algorithm and version $0.3.1$ of OmicsIntegrator for all analyses.

Minimum-Cost Flow: The minimum cost flow problem assigns certain nodes in the network to be "sources" and others to be "targets". Edges, which transport the flow from node to node, have a cost associated with using them and a capacity of how much flow they can hold. The solution is the network that satisfies the flow requirements of the source and target nodes while using lowest cost in edges (Goldberg and Tarjan, 1990). We implemented a version of min-cost flow using the solver provided in Google's OR-Tools[1], which solves the min-cost flow problem using the algorithm outlined in Bünnagel *et al.* (1998). This is a generic version of the algorithm used in ResponseNet (Basha *et al.*, 2013). Two parameters control the min-cost flow solution: the total flow through the network, which we vary from 1 to 50 in increments of 1, and the edge flow capacity, which we vary from 1 to 25 in increments of 1. We used Google's OR-Tools version $7.1.6720$ for all analyses.

2.3 Parameter Selection Methods

We consider the following parameter selection strategies from the literature to evaluate our pathway parameter advising approach:

Cross-validation: Cross-validation (CV) involves splitting the input data into training and testing sets multiple times for each parameter setting. A method is then fit or trained on each training set and evaluated on each respective testing set. In this problem setting, we do not have external ground truth with which to evaluate the predictions on test set data. Instead, we performed 5-fold CV on subsets of the input data, choosing the parameter values that generate a pathway from the training set nodes that recover the highest proportion of the test set nodes.

ResponseNet recommendation: We also tested a parameter selection heuristic used by ResponseNet (Yeger-Lotem *et al.*, 2009). The criterion is to select parameters that result in a pathway that includes at least 30% of the input set, while having the lowest proportion of low confidence edges. We extend this to rank the pathways that do include 30% of the inputs by their proportion of low confidence edges, followed by the pathways that include less than 30% of the inputs to form a full ranking.

Randomization stability: As suggested by Kedaigle and Fraenkel (2018), for PCSF we can also rank pathways by their robustness, as measured by how often nodes appeared in multiple runs with small random perturbations to the scores on the input nodes. Ranking pathways this way was only available in PCSF. Although it could be adapted to other pathway reconstruction methods, we decided to use it only in the method for which it was directly implemented.

2.4 Datasets

2.4.1 Interactomes

For both PathLinker and NetBox, we used the interactome included as a part of their software packages. For PCSF and min-cost flow, we used an interactome from Köksal *et al.* (2018) that merged protein interactions from the iRefIndex database v13 (Razick *et al.*, 2008) and kinase-substrate interactions from PhosphoSitePlus (Hornbeck *et al.*, 2014). The interactions from the iRefIndex database include confidence scores, while confidence scores for kinase-substrate interactions were inferred from the number of interactions for each kinase-substrate pair and the type of experiment each interaction was detected in. If an interaction was included in both databases, the PhosphoSitePlus interaction was used. This resulted in a network with 161901 weighted edges.

2.4.2 Pathway databases

All parameter tuning was performed with Reactome as the set of reference pathways. Reactome (Fabregat *et al.*, 2018) is a database of manually curated pathways, including 2287 human pathways. Reactome is open-source, where all contributions must provide literature evidence and are reviewed by an external domain expert before being added. Pathways smaller than 15 nodes were excluded as too small for meaningful interpretation.

The implausible pathway detection and reconstruction experiments were performed on pathways from the NetPath database. NetPath is a collection of 36 manually curated human signal transduction pathways (Kandasamy *et al.*, 2010). We used 15 NetPath pathways that contain at least 1 receptor and transcriptional regulator and are sufficiently connected, as described by Ritz *et al.* (2016). We designated 3 of these NetPath pathways as validation pathways: Wnt, TGF Beta, and TNF Alpha. Validation pathways were used to guide the choice of distance measure, and the remaining 12 pathways were

[1] `https://developers.google.com/optimization/flow/mincostflow`

reserved for a quantitative evaluation. We sampled the NetPath pathways in different ways for each pathway reconstruction algorithm to provide inputs in their expected formats. PCSF and NetBox do not require sources and targets, so we randomly sampled 30% of the pathway nodes as input. We also assigned each input a random prize sampled uniformly between 0 and 5 for PCSF. For PathLinker and min-cost flow, which require sources and targets, we selected all transcription factors and receptors for each pathway as outlined by Ritz *et al.* (2016).

2.4.3 Influenza host factors

Influenza host factor data was gathered from a meta-analysis of 8 RNAi studies (Tripathi *et al.*, 2015). The meta-analysis used the raw RNAi screen data to calculate a consolidated Z score for a total of 1257 host factor genes.

2.5 Implausible Network Criteria

In order to examine the ability of pathway parameter advising to avoid parameter settings that lead to impractical pathways, we created topological criteria that we use to define pathways as plausible or implausible. We use these criteria as a heuristic to label pathways as positive (plausible) or negative (implausible). The labels enable us to evaluate pathway rankings as a classification problem, determining if a method can correctly classify pathways as plausible or implausible. These criteria are based on previous analyses of biological network, and are as follows:

Size: We allowed pathways that had between 10 and 1000 nodes. Pathways whose size was outside this range are not practical for hypothesis generation and downstream analysis.

Hub node dependence: A common issue with pathway reconstruction algorithm is an over-reliance on high-degree or hub nodes, which can create networks consisting almost entirely of a single node and its neighbors with few to no connections between those neighbors (Kedaigle and Fraenkel, 2018). We score this using the ratio of the degree of the highest degree node to the average node degree of the entire pathway. If the highest degree node has over 20 times more edges than the average node in the pathway, we consider that pathway implausible.

Clustering coefficient: Biological networks have been found to have clustering or community structure that is hierarchical (Ravasz *et al.*, 2002); communities within the network exist at multiple scales and are often nested within each other. Thus, it would be reasonable to expect a plausible biological pathway to have at least a moderate level of community structure. We calculate the average clustering coefficient of all nodes in the pathway, a common metric for measuring community structure (Barabási and Oltvai, 2004). The clustering coefficient of a node is the proportion of its neighbors that are also neighbors of each other. This can be averaged over all nodes in the pathway as a measure of the overall level of clustering. We require pathways to have an average clustering coefficient of at least 0.05, as we expect at least some small level of clustering to exist. Because this requirement eliminated all PCSF pathways in 25% of tasks, when evaluating PCSF we excluded this metric in all cases.

Assortativity: A network's level of assortative mixing is defined as the tendency of high degree nodes to be connected to other high degree nodes. Biological networks have been found to be generally dissasortative, meaning that high degree nodes tend to be connected to low degree nodes (Newman, 2002; Albert *et al.*, 2011). Assortativity is measured between -1 and 1, where assortative networks have positive values and dissasortative networks have negative values. This value can be viewed as the correlation between a node's degree and its neighbor's degrees within the pathway. We consider networks with assortativity between -1 and 0.1 plausible to allow for some leeway in pathways being slightly assortative.

We selected these criteria based on attributes it would be reasonable to expect a biological pathway to have, with values supported by the literature where possible. These thresholds were not influenced by the graph topologies in the reference pathway database in order to minimize circularity between the reference pathway-based rankings and the plausibility criteria used to evaluate those rankings.

While the criteria for defining a plausible network are useful for comparing networks created by the same method with different parameter settings, they should not be considered as a metric for comparing pathways across pathway reconstruction methods. Different pathway reconstruction methods are able to use different sources of information and have complex strengths and weaknesses beyond the local topologies they return. For instance, NetBox, which had the highest proportion of plausible pathways, cannot take into account information such as edge confidence or scores on proteins of interest that other methods such as PCSF can. Although they are useful for evaluation, these properties are a binary way to define as a baseline if a pathway is reasonable or unreasonable. Thus, they could not be used to rank pathway reconstruction parameters themselves.

2.5.1 Criteria Grid Search

In order to make sure that our experimental results are not overly sensitive to the specific choice of cut-offs for pathway plausibility, we tested other cut-offs in a grid search. We varied the maximum network size cut-off from 200 to 2000, the hub node dependence measure from 5 to 50, the clustering coefficient cut-off from 0.0 to 0.1, and the assortativity cut-off from -0.5 to 0.5. Each range was divided into 10 intervals, for a total of 10000 sets of plausibililty cut-offs. Figures 5 and S5 evaluate the pathway reconstruction methods across these different thresholds.

2.6 Matthew's Correlation Coefficient

In the NetPath evaluation, we used Matthew's Correlation Coefficient (MCC) to quantify the quality of generated pathways (Matthews, 1975). MCC is a metric that ranges between -1 and 1, where 1 indicates a perfect binary classification

and -1 indicates a complete inverse classification. It can be viewed as the correlation between the predicted and true labels in a classification task. MCC has been shown to be well suited to evaluate classification in imbalanced settings (Boughorbel *et al.*, 2017). When calculating performance, we consider all edges in a NetPath pathway as the positive set and all other edges as the negative net. MCC is defined as

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives. When comparing MCC values of multiple pathways and methods, we normalized MCC values by the best possible MCC among all tested parameter values for that pathway and method. We refer to this value as the adjusted MCC.

2.7 Enrichment Analyses

GO (The Gene Ontology Consortium, 2018) and KEGG pathway (Kanehisa and Goto, 2000) enrichment was carried out with DAVID v6.7 (Huang *et al.*, 2008). Enrichment was performed using GO biological process terms and all KEGG pathways. Thresholds for term inclusion were set to a count of 2 and an EASE score of 0.1.

2.8 Pathway parameter advising implementation

A Python implementation of pathway parameter advising is available at `https://github.com/gitter-lab/pathway-parameter-advising` under the MIT license. While the v0.1.0 release of the pathway parameter advising software supports Python v3.6, the original computations for the results here used Python v2.7.16 and Anaconda version v2019.03. The following package versions were used: pandas v0.24.2, networkx v2.2, numpy v1.16.2, matplotlib v2.2.3, and seaborn v0.9.0. The Parallel Graphlet Decomposition library was pulled from GitHub on April 30, 2019.

2.9 Data Availability

All data used for experiments is publicly available at `http://netpath.org/browse` and from `https://www.pathwaycommons.org/` using the query `https://www.pathwaycommons.org/pc2/search?q=*&type=pathway&datasource=reactome`. Software is available at `https://github.com/gitter-lab/pathway-parameter-advising`.

3 Results

Because pathway reconstruction is an exploratory analysis and our goal is to maximize downstream biological utility, which cannot be directly quantified, we resort to multiple indirect approaches to evaluate pathway parameter advising. We first ensure that the graphlet distance ranking metric has desirable properties. It is not overly sensitive to the disconnected graphlets and shows that reference pathways are similar to one another. We then use the literature to define topological graph properties that make a candidate pathway biologically implausible and show that pathway parameter advising can optimize pathway reconstruction algorithms to avoid implausible pathways. Next, we demonstrate that we can improve the reconstruction of NetPath pathways by comparing predicted pathway edges with the NetPath ground truth. Finally, we show how pathway parameter advising can be applied in practice and generate an influenza host factor pathway from RNAi screens.

3.1 Evaluating the Ranking Metric

In order to determine the validity of our pathway ranking metric, we first examined the overall distribution of graphlet distances ($E(G)$ in Section 2.1.3) across all pathway reconstruction methods and Reactome pathways with added noise. We added noise to pathways by removing a percentage of all edges in the pathway, then adding back that many edges that did not appear in the original pathway. We also calculated the graphlet distance between each Reactome pathway and all of the other Reactome pathways. Figure 3 shows these distributions of graphlet distances. Reactome pathways had a lower mean distance than any set of reconstructed pathways, confirming that our metric ranks curated pathways over reconstructed pathways.

In order to examine the possibility that unconnected graphlets dominate the ranking metric, we first examined the breakdown of total graphlet distance by each of the individual graphlets (Figure 2). Figure S3 shows the contribution of each graphlet to the ranking metric. While a single graphlet, the 4 unconnected nodes graphlet, does have the largest contribution to graphlet distance on average, its median contribution to the ranking metric's value is only 30%. As biological pathways tend to be sparse, this graphlet count scales with pathway size. Thus, this can been seen as pathway size contributing to about 30% of the ranking metric.

To confirm that this graphlet was not negatively impacting our ranking metric, we also examined its contribution to implausible pathway detection performance as described in Section 3.2. We used the NetPath pathways Wnt, TNF alpha, and TGF beta as validation pathways to evaluate different graphlet metrics and develop pathway parameter advising. These 3 validation pathways are excluded from all aggregate results. On these 3 held aside NetPath pathways, we found that using

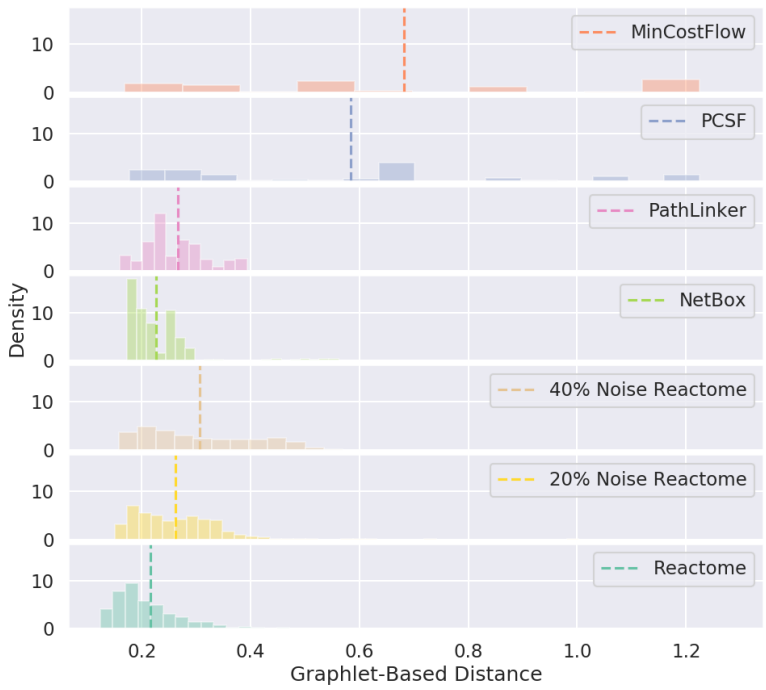

Fig. 3. Distribution of graphlet-based distances ($E(G)$) for all reconstructed pathways, Reactome pathways, and Reactome pathways with added noise. Reactome pathways were excluded from their own distance calculation. Vertical dashed lines show the mean graphlet distance. Reactome pathways were found to have the lowest mean graphlet distance, confirming that our method ranks curated pathways over reconstructed pathways.

the 4 unconnected nodes graphlet alone resulted in poor performance (Figure S4) and that its inclusion moderately boosts performance.

3.2 Implausible Pathway Detection

In order to evaluate pathway parameter advising, we considered its ability to avoid implausible networks. While it is difficult to define a single best pathway in the context of an exploratory analysis, some pathways are clearly biologically unrealistic, infeasible to analyze, or not useful for downstream analysis. Thus, pathway parameter advising should consistently rank parameter settings that lead to plausible networks above those that lead to implausible networks.

We applied the four pathway reconstruction methods to sampled NetPath pathways (Section 2.4) to reconstruct the 15 pathways, following the evaluation approach used by PathLinker (Ritz *et al.*, 2016). This resulted in 60 parameter tuning tasks across the 15 pathways and 4 pathway reconstruction methods.

In these experiments pathways considered plausible, as defined in Section 2.5, are treated as the positive set. Pathways defined as implausible are the negative set. We then used precision-recall (PR) curves (Figures S5-S9) to evaluate how well pathway parameter advising and alternative parameter selection strategies distinguish plausible from implausible networks. Parameter selection methods that rank plausible networks above implausible networks will have a higher area under the PR curve (AUPR).

Figure 4 shows the distribution of AUPRs across the 4 pathway reconstruction methods. Different methods also had varying proportions of networks identified as plausible, with min-cost flow having the lowest mean proportion at 11% and NetBox with the highest at 89%. Among the 12 test pathways, pathway parameter advising has the highest median AUPR for each pathway reconstruction method.

Of the 36 cases where AUPRs could be compared (both plausible and implausible pathways were present), pathway parameter advising had the highest AUPR in 30. Cross-validation had the highest AUPR in the other 6. The impact of the choice of parameter ranking strategy is most stark for PCSF, where graphlet frequency distance has perfect AUPR in almost all pathways and the other approaches struggle. Not only did pathway parameter advising have the highest median AUPR, but its performance was the most consistent; it had the lowest variance in AUPR across all tasks.

In order to make sure that performance was not overly influenced by our specific choice of criteria for plausible and implausible networks, we then varied the plausibility cut-off for each of the 4 topological properties we considered in Section 2.5 in a grid search, resulting in a total of 10000 configurations. Figure 5 shows the AUPR of each parameter ranking method as the plausibility cut-off for each topological feature is varied, and aggregate AUPR values are shown in Figure S5. Pathway parameter advising outperforms all other parameter ranking methods on average at each cut-off value.

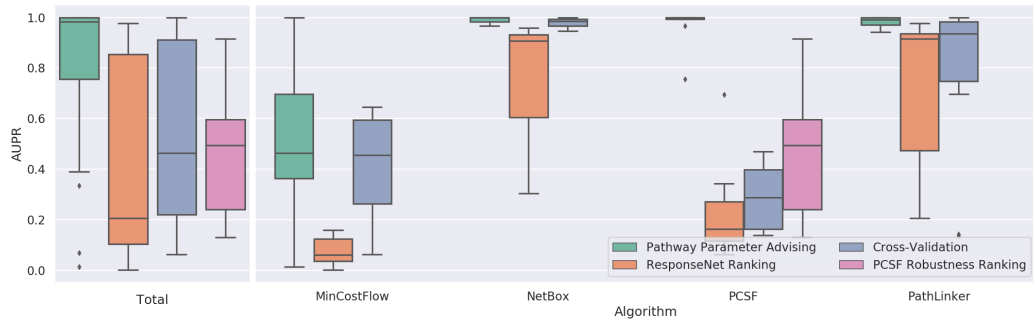

Fig. 4. Performance of parameter selection methods on avoiding implausible networks. Boxplots represent the distributions of the AUPRs aggregated for 4 pathway reconstruction methods and 12 test pathway reconstructions from the NetPath database. Degenerate cases where all or no pathways met the plausibility criteria are excluded. Full results, including the 3 validation pathways and degenerate cases, can be found in Figures S6-S9.

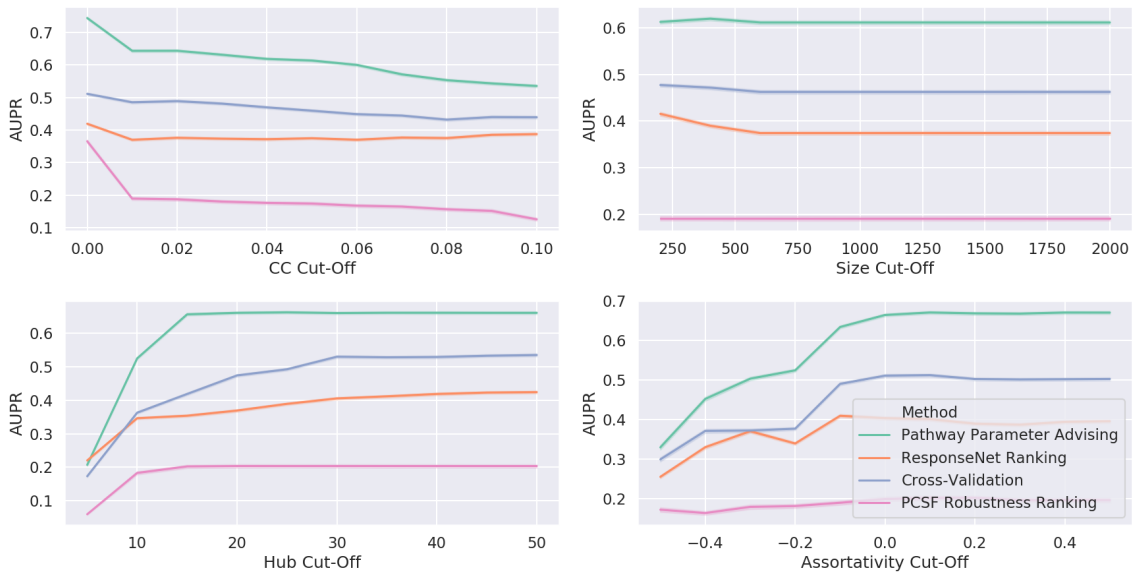

Fig. 5. Performance of parameter selection methods on avoiding implausible networks as the cut-off for plausibility is varied across different topological features – clustering coefficient, pathway size, hub node dependence, and assortativity – as described in Section 2.5. Lines show mean AUPR over the varied cut-offs for the other 3 topological features for all 12 NetPath pathways and 4 pathway reconstruction methods.

3.3 NetPath Pathway Reconstruction

Having achieved our primarily goal of accurately prioritizing parameters that generate plausible pathways, we also evaluated the quality of the pathway reconstructions themselves. We compared pathway parameter advising to the alternative ranking methods and the default parameters. In these experiments we define all pathway edges in the NetPath pathway as a positive set and all other edges as a negative set. We then compared the ability of pathway parameter advising and other parameter ranking methods to promote reconstructed pathways that closely resemble their NetPath equivalent.

Figure 6 (left) shows the adjusted MCCs of all 48 pathway reconstruction tasks. While pathway parameter advising has the highest median adjusted MCC, the parameter selection method has less impact on MCC than it did on pathway plausibility (Figure 4). When stratified by pathway reconstruction algorithm, pathway parameter advising has the highest median adjusted MCC for PCSF and PathLinker. CV has the highest adjusted MCC in min-cost flow and NetBox. Of the 48 reconstruction tasks, pathway parameter advising had the highest median adjusted MCC 21 times, while CV had 13, default parameters had 9, and the ResponseNet ranking had 8, including 3 cases where 2 methods tied.

Figure 6 (right) shows the distribution of best possible MCCs for all reconstruction tasks, including the 3 validation pathways. The best possible MCC was greater than 0.3 in only 4 of the 60 cases, and it was never greater than 0.4. Given these low unadjusted MCC values and the overall objective of our study, the implausible pathway detection experiment is a better indicator of method performance.

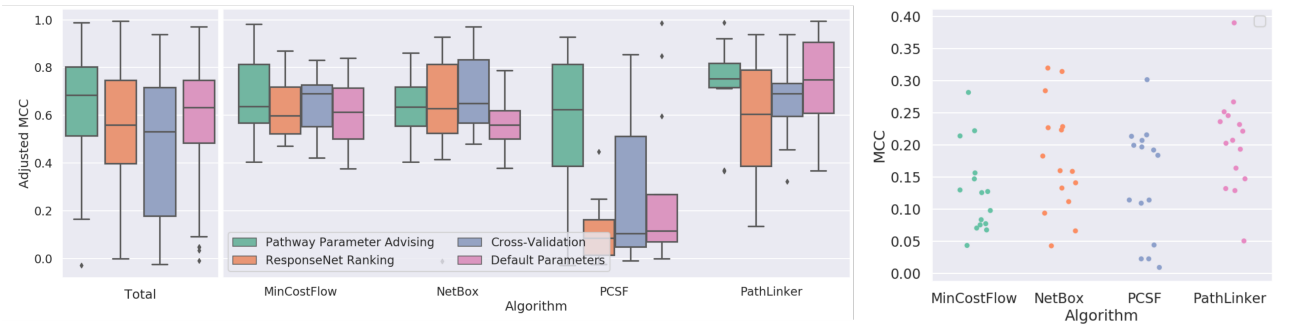

Fig. 6. Left: Adjusted MCC of parameter selection methods on reconstructing 12 test pathways from the NetPath database across 4 pathway reconstruction methods. MCCs were normalized to the highest possible MCC within a given pathway reconstruction method and pathway. Figure S10 shows MCC values by pathway, including the 3 validation pathways excluded here. Right: The highest possible MCC of pathway reconstruction in 60 parameter sweeps across 4 pathway reconstruction methods and 15 NetPath pathways (validation and test). The MCC values are generally low, reflecting low overlap between the predicted and NetPath pathway edges.

3.4 Influenza Host Factor Pathway Reconstruction

To demonstrate how pathway parameter advising can guide the biological interpretation of omic data, we reconstructed an influenza host factor pathway. Our aim was to create a pathway that represents aspects of influenza's infectious activities and could lead to the discovery of new host factors or host factor regulators. We created an influenza host factor network using the 1257 host factors from a meta-analysis of 8 RNAi screens (Tripathi *et al.*, 2015). These host factors were given as input to PCSF, using the same range of possible parameter settings as in the other experiments. We used the magnitude of the consolidated Z scores given in the meta-analysis as node scores (see Section 2.4.3).

After creating the candidate host factor pathways, we ranked the parameter settings using pathway parameter advising. Figure 7 (left) shows the PR curve of different parameter ranking methods' ability to avoid implausible networks. Pathway parameter advising ranked the pathways almost perfectly, with an AUPR of 0.96, while other parameter selection methods had more difficulty separating implausible from plausible networks. CV and the ReponseNet rankings performed worse than the random baseline. This demonstrates that pathway parameter advising performs well not only on simulated data from NetPath but also on data aggregated from real high-throughput experiments.

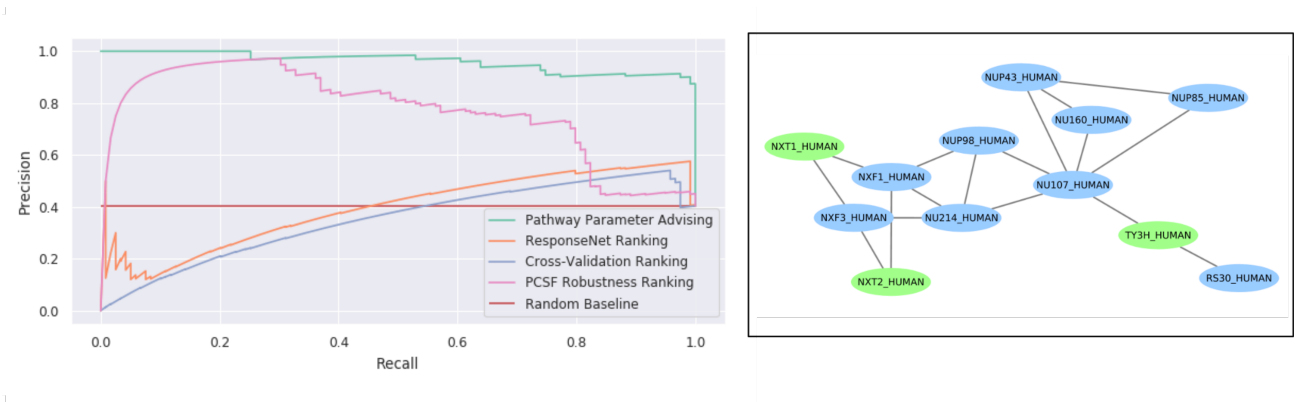

Fig. 7. Left: Precision-recall curve for implausible networks in PCSF influenza host factor network construction. Right: A component of the influenza host factor ensemble pathway created from the top 50 PCSF parameter settings ranked by pathway parameter advising. This component represents 12 of the 86 total nodes in the pathway (Figure S11). Host factor nodes provided as input are shown in blue, while green nodes are "Steiner" nodes that PCSF predicts to connect the host factors.

We also created three ensemble networks from the resultant pathways from the top, middle, and bottom 50 parameter settings ranked by pathway parameter advising. As discussed in Section 1, ensembling networks is a common way to use PCSF. Thus, we expect the top 50 ensemble pathway to be best for downstream analysis and interpreting the input host factor data. Of the 3 constructed ensemble pathways, the pathway made from the 50 highest ranked parameters contains 86 nodes. The middle ranked and low ranked ensemble pathways have 7337 and 15 nodes, respectively (Figure S11). The middle ranked pathway is too large to interpret and does not provide meaningful new insights into the relationships among host factors. The low ranked pathway is too small to illuminate new biological hypotheses. The top ranked pathway, however, is large enough for meaningful enrichment and downstream analyses while remaining small enough to be feasible.

We then performed a gene set enrichment analysis on the top ensemble pathway using DAVID (Huang *et al.*, 2008) (Section 2.7). We tested both Gene Ontology (GO) biological process terms (Table S1) and KEGG pathway enrichment (Table S2). The top 2 KEGG pathways enriched were RNA transport and influenza A. The influenza A pathway being enriched is a confirmation that our ensemble pathway is representing influenza processes well.

The RNA transport pathway enrichment captures one unique aspect of influenza A. It replicates within the nucleus, so it has complex processes for transporting viral RNA in and out of the nucleus (Dou *et al.*, 2018). Similar concepts are observed in the top 2 enriched GO terms, mRNA and tRNA export from nucleus (Table S1). We also see other general viral GO terms, which confirm the top ranked pathway's representation of influenza, such as viral transcription and intracellular transport of virus.

Figure 7 (right) shows one connected component representing 12 of the 86 nodes from the top ranked ensemble pathway. One node in particular, NXT2, was not among the original host factors but was identified as a possible host factor in a later genome-wide CRISPR/Cas9 screen (Han *et al.*, 2018). This demonstrates how pathways chosen through pathway parameter advising could guide new discoveries.

The influenza study also illustrates the danger of running pathway reconstruction methods with default parameters alone. The PCSF network constructed using default parameters had 6676 nodes, which is too large to feasibly interpret. We then looked into the role of NXT2 in the default parameters network. Figure 8 shows the subnetworks of the default parameters network for all nodes reachable within 2 and 3 edges of NXT2. As opposed to a functionally cohesive subnetwork, NXT2 only connects to a high-degree node. This node then connects to APP and SUMO2, which are the 2 highest degree nodes in the interactome. This results in 727 nodes being within 3 edges of NXT2, and is more likely an artifact of the PCSF algorithm run with improper parameters than a meaningful biological pathway structure. This hub based structure gives much less insight into the role of NXT2.

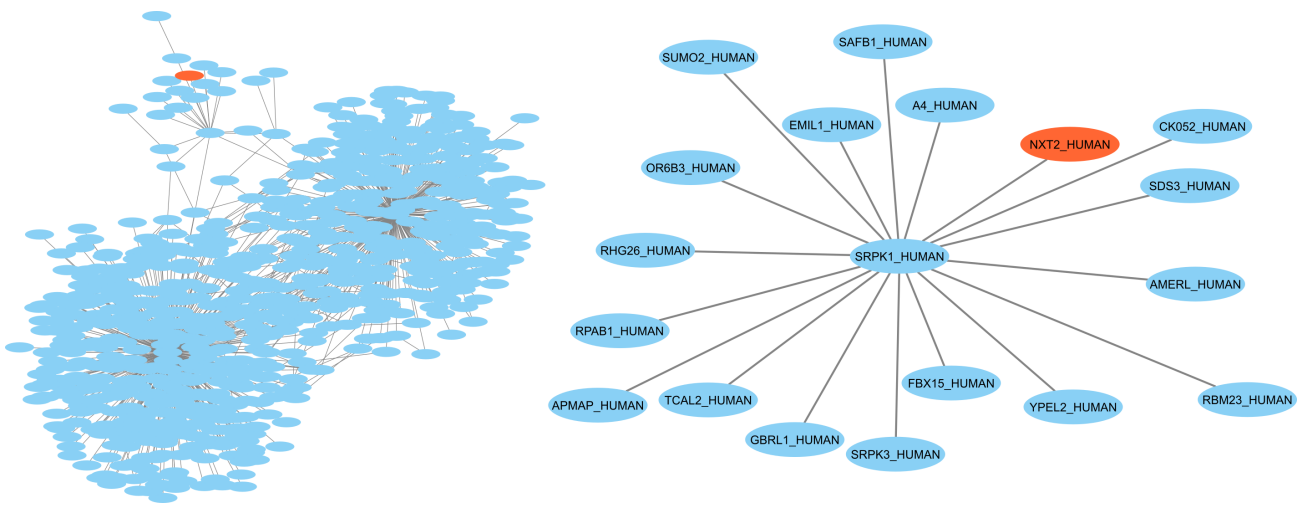

Fig. 8. All nodes within distance 3 (left) and distance 2 (right) of NXT2, highlighted in orange, in the PCSF influenza host factor network constructed from default parameters. Using the default parameters alone resulted in a large hub-node focused network with little useful biological insight.

4 Discussion

Pathway parameter advising selects parameters that lead to useful, plausible networks for a variety of pathway reconstruction algorithms. This parameter tuning approach is algorithm-agnostic and uses background knowledge in the form of pathway databases to succeed in selecting reasonable pathways during pathway reconstruction.

Many of the networks down-ranked by pathway parameter advising, such as pathways with many thousands of nodes or a network consisting only of a single node and its neighbors, seem obvious to avoid. Typically, these types of generated pathways are ignored through a process of manual trial and error. Any manual step in the network analysis could lead to human error, may accidentally introduce bias into the final pathway model, and limits the number of parameter combinations that can be assessed. Therefore, automatically avoiding these poor pathways is important. The specific choice of criteria for defining implausible pathways was inconsequential. Pathway parameter advising excelled at implausible pathway detection compared to other parameter ranking methods for all definitions. Pathway parameter advising quantifies and deprioritizes implausible topologies without any human intervention except for the inclusion of background knowledge.

In addition to avoiding implausible pathways, pathway parameter advising narrowly performed best at reconstructing NetPath pathways. Although it was less clearly dominant than in the implausible pathway detection experiment, this further highlights its effectiveness. Much of the total performance is driven by how well pathway parameter advising performs in

PCSF, though it is also the only one to be either the first or second highest performing for all 4 pathway reconstruction methods. However, raw MCC values in many test cases were so low that differences in MCC were driven by only a few interactions, so this experiment alone does not provide enough evidence to draw strong conclusions.

We also found that no single graphlet dominated our ranking metric. There are likely two causes. In larger networks, such as complete protein interaction networks, disconnected graphlet counts can dominate other graphlets by orders of magnitude (Johansson *et al.*, 2015). In contrast, the networks in biological pathway reconstruction tend to not be large enough for the portion of unconnected graphlets in sparse graphs to completely dominate other graphlets. In addition, our ranking metric only calculates distance from the closest 20% of reference pathways. Thus, the signal of pathway size from the 4 unconnected nodes graphlet guides the selection of the closest reference pathways to pathways of similar size. Within this 20%, other graphlets representing more local topology have a larger contribution.

Although we used different pathway databases in our experiments for reconstructing pathways and the set of reference pathways, it is possible that some pathways are similar across the NetPath and Reactome databases. Cross-database pathway similarity could cause a version of the reconstructed pathway to be used as a reference pathway. However, even if this is the case, the shared pathway would be 1 of the over 1000 Reactome pathways used as a reference set. Thus, its effect on the ranking metric would be negligible.

In both experiments, other parameter selection methods especially struggled choosing parameters for PCSF. Finding a good parameter setting for a method with multiple, complex parameters like PCSF can be especially difficult and important and is where pathway parameter advising is most useful. In contrast, a method like PathLinker contains a single parameter, which monotonically increases pathway size. Changing the parameter value has a relatively predictable effect.

There are some drawbacks to pathway parameter advising. It requires a parameter sweep as opposed to a single run with the default parameter setting. This greatly increases overall runtime for some pathway algorithms. Because pathway parameter advising is algorithm-agnostic, it makes no assumptions about the parameter space it is optimizing. Thus, pathway parameter advising has no way of knowing if the set of parameters considered is broad enough to find the optimal pathway. However, it is worth noting that all other parameter selection methods tested except for the default parameters suffer from this drawback as well.

Another issue is that pathway parameter advising is dependent on a database of reference pathways. The popular pathway database Reactome works well in analyses here. However, if the optimal predicted pathway is reasonable but outside the range of topologies seen among the reference pathways, it would be overlooked.

Our distance metric focuses only on topology and does not include any information about the biological context. ResponseNet (Basha *et al.*, 2019) and PathLinker (Youssef *et al.*, 2018) extensions consider tissue-specificity and protein localization context, respectively. A possible extension of pathway parameter advising would be to account for this information, such as adding a penalty for interactions that occur in different tissues or cellular compartments. Similarly, we used the entire Reactome database as the reference pathways. Limiting the reference pathways to a certain process or function, such as signal transduction or disease, could allow pathway parameter advising to select pathways more similar to a domain of interest. In addition, instead of computing graphlet distance using 20% of all reference pathways, we could first cluster the reference pathways and consider the distances only to pathways in the most topologically similar cluster.

Now that we have demonstrated the utility of pathway parameter advising to evaluate predicted pathways, it could be wrapped with hyperparameter optimization to fully automate the process with a Bayesian optimization framework (Kandasamy *et al.*, 2019; Wang *et al.*, 2013; Chen *et al.*, 2012). This would allow for complete and standardized automation of the pathway reconstruction process. It could reduce the overall time required to select parameters as well because the parameter space can be explored adaptively instead of through an exhaustive grid search. Avoiding a parameter grid search is most valuable for pathway reconstruction methods like PCSF and ANAT (Yosef *et al.*, 2011) that have several tunable parameters.

Acknowledgements

We thank Anna Ritz for providing formatted data and assisting with the NetPath evaluation, Paul Ahlquist and Eunju Park for discussing toplogical features of plausible pathways, Tijana Milenkovic for graphlet distance suggestions, and members of the Gitter research group for their helpful comments.

Funding

This research was supported by National Library of Medicine training grant 5T15LM007359 (CM), NSF CAREER award DBI 1553206 (AG), NIH award 5R21AI125897 (AG), and the John W. and Jeanne M. Rowe Center for Research in Virology (CM and AG).

Competing Interests

The authors declare that there are no competing interests.

Author Contributions

CM and AG developed the methods and experiments, analyzed results, and wrote the manuscript. CM wrote the software and performed the experiments. Both authors read and approved the final manuscript.

References

Ahmed, N. K., Neville, J., Rossi, R. A. and Duffield, N. (2015). Efficient graphlet counting for large networks. In *2015 IEEE International Conference on Data Mining*, pages 1–10.

Akaike, H. (1998). A New Look at the Statistical Model Identification. In E. Parzen, K. Tanabe, and G. Kitagawa, editors, *Selected Papers of Hirotugu Akaike*, Springer Series in Statistics, pages 215–222. Springer New York, New York, NY.

Albert, R., DasGupta, B., Hegde, R., Sivanathan, G. S. *et al.* (2011). Computationally efficient measure of topological redundancy of biological and social networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, **84**(3 Pt 2), 036117.

Bailly-Bechet, M., Borgs, C., Braunstein, A., Chayes, J. *et al.* (2011). Finding undetected protein associations in cell signaling by belief propagation. *Proceedings of the National Academy of Sciences*, **108**(2), 882–887.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, **5**(2), 101–113.

Basha, O., Tirman, S., Eluk, A. and Yeger-Lotem, E. (2013). ResponseNet2.0: revealing signaling and regulatory pathways connecting your proteins and genes–now with human data. *Nucleic Acids Research*, **41**(W1), W198–W203.

Basha, O., Mauer, O., Simonovsky, E., Shpringer, R. *et al.* (2019). ResponseNet v.3: revealing signaling and regulatory pathways connecting your proteins and genes across human tissues. *Nucleic Acids Research*, **47**(W1), W242–W247.

Boughorbel, S., Jarray, F. and El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE*, **12**(6), e0177678.

Budak, G., Ozsoy, O. E., Son, Y. A., Can, T. *et al.* (2015). Reconstruction of the temporal signaling network in Salmonella-infected human cells. *Frontiers in Microbiology*, **6**, 730.

Bünnagel, U., Korte, B. and Vygen, J. (1998). Efficient implementation of the Goldberg-Tarjan minimum-cost flow algorithm. *Optimization Methods and Software*, **10**(2), 157–174.

Cerami, E., Demir, E., Schultz, N., Taylor, B. S. *et al.* (2010). Automated Network Analysis Identifies Core Pathways in Glioblastoma. *PLOS ONE*, **5**(2), e8918.

Chen, B., Castro, R. M. and Krause, A. (2012). Joint Optimization and Variable Selection of High-dimensional Gaussian Processes. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1379–1386, USA. Omnipress.

Choobdar, S., Ahsen, M. E., Crawford, J., Tomasoni, M. *et al.* (2019). Assessment of network module identification across complex diseases. *Nature Methods*, **16**(9), 843–852.

Cowen, L., Ideker, T., Raphael, B. J. and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, **18**(9), 551–562.

DeBlasio, D. and Kececioglu, J. (2015). Parameter advising for multiple sequence alignment. *BMC Bioinformatics*, **16**(2), A3.

DeBlasio, D. and Kececioglu, J. (2017). Learning Parameter-Advising Sets for Multiple Sequence Alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **14**(5), 1028–1041.

Dou, D., Revol, R., Östbye, H., Wang, H. *et al.* (2018). Influenza a virus cell entry, replication, virion assembly and movement. *Frontiers in Immunology*, **9**, 1581–1581.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K. *et al.* (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, **46**(D1), D649–D655.

Furlong, L. I. (2013). Human diseases through the lens of network biology. *Trends in Genetics*, **29**(3), 150–159.

Goh, W. W. B., Lee, Y. H., Chung, M. and Wong, L. (2012). How advancement in biological network analysis methods empowers proteomics. *Proteomics*, **12**(4-5), 550–563.

Goldberg, A. V. and Tarjan, R. E. (1990). Finding Minimum-Cost Circulations by Successive Approximation. *Mathematics of Operations Research*, **15**(3), 430–466.

Han, J., Perez, J. T., Chen, C., Li, Y. *et al.* (2018). Genome-wide CRISPR/Cas9 screen identifies host factors essential for influenza virus replication. *Cell Reports*, **23**(2), 596 – 607.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M. *et al.* (2014). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research*, **43**(D1), D512–D520.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4**, 44.

Johansson, F. D., Frost, O., Retzner, C. and Dubhashi, D. (2015). Classifying large graphs with differential privacy. In V. Torra and T. Narukawa, editors, *Modeling Decisions for Artificial Intelligence*, pages 3–17, Cham. Springer International Publishing.

Kandasamy, K., Mohan, S. S., Raju, R., Keerthikumar, S. *et al.* (2010). NetPath: a public resource of curated signal transduction pathways. *Genome Biology*, **11**(1), R3.

Kandasamy, K., Vysyaraju, K. R., Neiswanger, W., Paria, B. *et al.* (2019). Tuning hyperparameters without grad students: Scalable and robust Bayesian optimisation with Dragonfly. *arXiv*, (1903.06694).

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **28**(1), 27–30.

Kececioglu, J. and DeBlasio, D. (2013). Accuracy Estimation and Parameter Advising for Protein Multiple Sequence Alignment. *Journal of Computational Biology*, **20**(4), 259–279.

Kedaigle, A. J. and Fraenkel, E. (2018). Discovering altered regulation and signaling through network-based integration of transcriptomic, epigenomic and proteomic tumor data. *Methods in Molecular Biology (Clifton, N.J.)*, **1711**, 13–26.

Khurana, V., Peng, J., Chung, C. Y., Auluck, P. K. *et al.* (2017). Genome-Scale Networks Link Neurodegenerative Disease Genes to α-Synuclein through Specific Molecular Pathways. *Cell Systems*, **4**(2), 157–170.e14.

Köksal, A. S., Beck, K., Cronin, D. R., McKenna, A. *et al.* (2018). Synthesizing Signaling Pathways from Temporal Phosphoproteomic Data. *Cell Reports*, **24**(13), 3607–3618.

Komurov, K., White, M. A. and Ram, P. T. (2010). Use of Data-Biased Random Walks on Graphs for the Retrieval of Context-Specific Networks from Genomic Data. *PLOS Computational Biology*, **6**(8), e1000889.

MacGilvray, M. E., Shishkova, E., Chasman, D., Place, M. *et al.* (2018). Network inference reveals novel connections in pathways regulating growth and defense in the yeast salt response. *PLOS Computational Biology*, **13**(5), 1–28.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, **405**(2), 442 – 451.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N. *et al.* (2002). Network Motifs: Simple Building Blocks of Complex Networks. *Science*, **298**(5594), 824–827.

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, **89**, 208701.

Pržulj, N., Corneil, D. G. and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, **20**(18), 3508–3515.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. *et al.* (2002). Hierarchical Organization of Modularity in Metabolic Networks. *Science*, **297**(5586), 1551–1555.

Razick, S., Magklaras, G. and Donaldson, I. M. (2008). iRefIndex: A consolidated protein interaction database with provenance. *BMC Bioinformatics*, **9**(1), 405.

Ritz, A., Poirel, C. L., Tegge, A. N., Sharp, N. *et al.* (2016). Pathways on demand: automated reconstruction of human signaling networks. *npj Systems Biology and Applications*, **2**(1), 1–9.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, **6**(2), 461–464.

Scott, J., Ideker, T., Karp, R. M. and Sharan, R. (2006). Efficient Algorithms for Detecting Signaling Pathways in Protein Interaction Networks. *Journal of Computational Biology*, **13**(2), 133–144.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A. *et al.* (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, **47**(D1), D607–D613.

The Gene Ontology Consortium (2018). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, **47**(D1), D330–D338.

Tripathi, S., Pohl, M., Zhou, Y., Rodriguez-Frandsen, A. *et al.* (2015). Meta- and orthogonal integration of influenza "omics" data defines a role for UBR4 in virus budding. *Cell Host & Microbe*, **18**(6), 723 – 735.

Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.-S. C. *et al.* (2013). Simultaneous Reconstruction of Multiple Signaling Pathways via the Prize-Collecting Steiner Forest Problem. *Journal of Computational Biology*, **20**(2), 124–136.

Tuncbag, N., Gosline, S. J. C., Kedaigle, A., Soltis, A. R. *et al.* (2016). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *PLOS Computational Biology*, **12**(4), e1004879.

Wang, Z., Zoghi, M., Hutter, F., Matheson, D. *et al.* (2013). Bayesian Optimization in High Dimensions via Random Embeddings. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 1778–1784. AAAI Press.

Yaveroğlu, Ö. N., Malod-Dognin, N., Davis, D., Levnajic, Z. *et al.* (2014). Revealing the Hidden Language of Complex Networks. *Scientific Reports*, **4**, 4547.

Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D. *et al.* (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature Genetics*, **41**(3), 316–323.

Yosef, N., Zalckvar, E., Rubinstein, A. D., Homilius, M. *et al.* (2011). ANAT: A Tool for Constructing and Analyzing Functional Protein Networks. *Science Signaling*, **4**(196), pl1.

Youssef, I., Law, J. and Ritz, A. (2018). Integrating Protein Localization with Automated Signaling Pathway Reconstruction. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 163–168.

Supplementary Information: Automating parameter selection to avoid implausible biological pathway models

Chris S. Magnano and Anthony Gitter

S1 Supplementary Figures

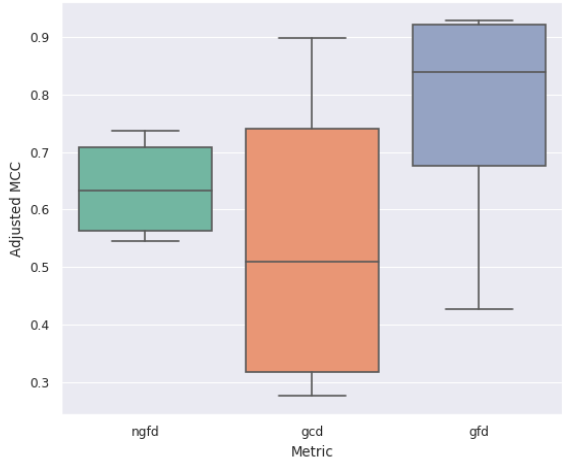

Fig. S1. Examining the effect of different graphlet-based distance metrics on the adjusted MCC of pathway reconstruction. Reconstructions were performed on the validation pathways Wnt, TNF Alpha, and TGF Beta across the 4 pathway reconstruction algorithms. We examined 3 graphlet-based metrics for pathway parameter advising: normalized graphlet frequency distance (NGFD), graphlet correlation distance (GCD), and graphlet frequency distance (GFD). For NGFD, we wanted to explore a metric that takes advantage of all generated pathways being subnetworks of the same interactome. Thus, we normalized all graphlet frequencies by the corresponding graphlet's frequency in the interactome. We also explored GCD, which measures the correlation between connected graphlets in a pathway (Yaveroğlu et al., 2014). This creates a metric that is solely focused on local topology and has minimal information about pathway size or other global topological properties. Adjusted MCC was calculated the same way as in Section 3.3. GFD outperforms the other methods. One possible reason for GFD outperforming more complex methods like GCD is that GCD attempts to eliminate the signal of global topological properties such as size and give information on graphlets only. Some signal of global topology, however, likely aids in identifying which pathways are similar to reference pathways.

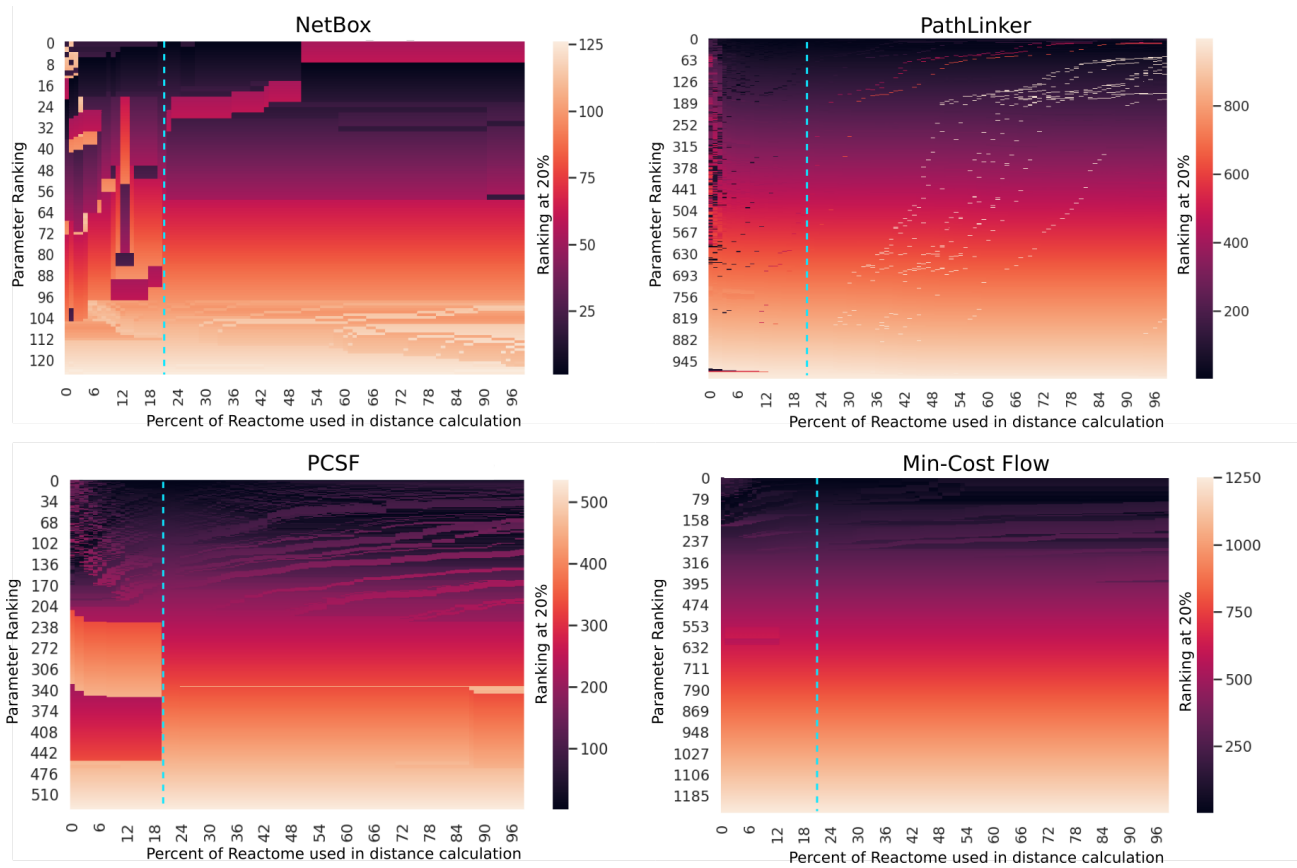

Fig. S2. Change in ranking order over different percent thresholds for calculating score in pathway parameter advising on the TGF beta receptor NetPath pathway. Heatmaps are colored based on the ranking at the chosen threshold, 20% (marked by a blue dashed line). While the parameter ranking is unstable for very small values, by 20% the ranking remains generally unchanged as the threshold continues to increase, as can be seen in the color gradient consistency in the right halves of the figures.

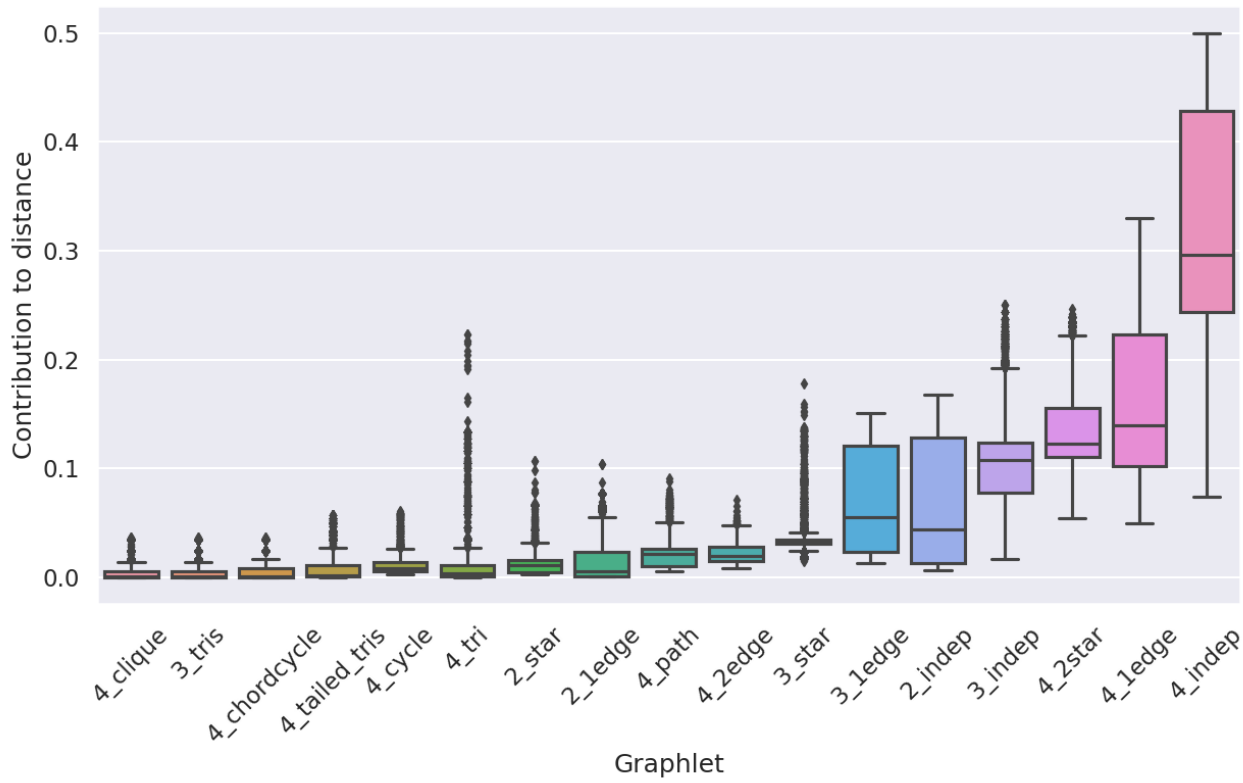

Fig. S3. Contribution of each of the 17 graphlets to graphlet distance across the validation pathways Wnt, TNF Alpha, and TGF Beta and 4 pathway reconstruction algorithms. Graphlets are labeled as according to Ahmed et al. (2015). The 4 disconnected nodes graphlet, 4_indep, contributes to a median of about 30% of the GFD. This can be interpreted as suggesting that about 30% of the GFD is based on network size.

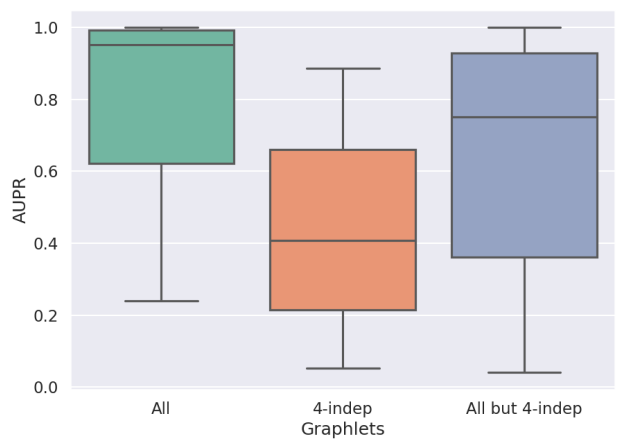

Fig. S4. Performance of parameter selection methods on avoiding implausible networks using different sets of graphlets: all graphlets (All), only the 4 disconnected node graphlet (4-indep), and all graphlets except the 4 independent node graphlet (All but 4-indep). Boxplots represent the distributions of the AUPRs aggregated for 4 pathway reconstruction methods and the validation pathways Wnt, TNF Alpha, and TGF Beta. Using all graphlets yields the best performance. Removing the 4 disconnected node graphlet slightly lowers the AUPR and only using the 4 disconnected node graphlet results in a large performance decrease. This suggests that our ranking method is highly dependent on graphlets other than the 4 disconnected node graphlet.

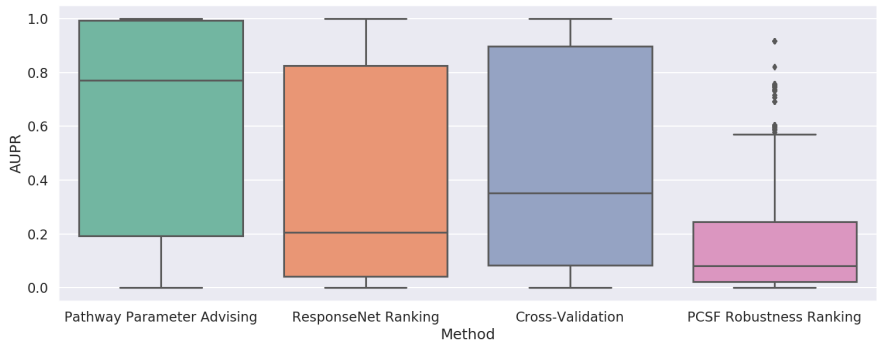

Fig. S5. Performance of parameter selection methods on avoiding implausible networks aggregated for all considered 10000 plausibility criteria. AUPR is shown for all 12 NetPath pathways and 4 pathway reconstruction methods.

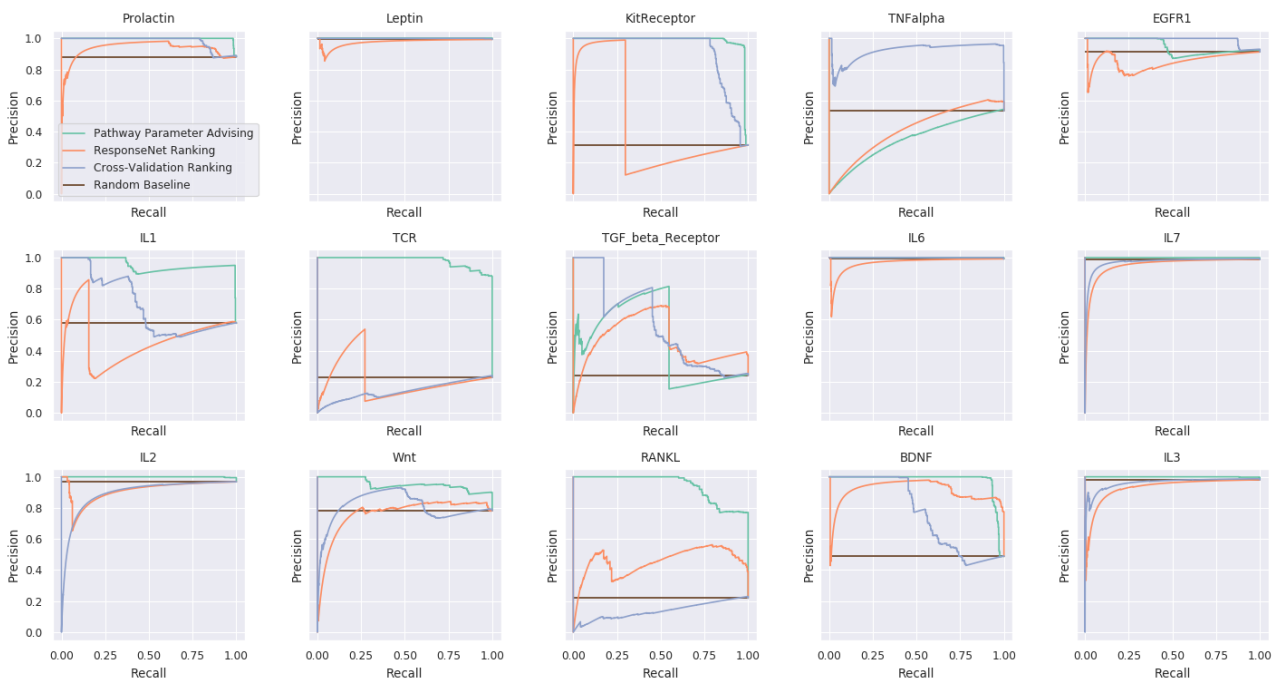

Fig. S6. All PR curves for implausible pathway detection using different parameter ranking schemes for pathways created using PathLinker.

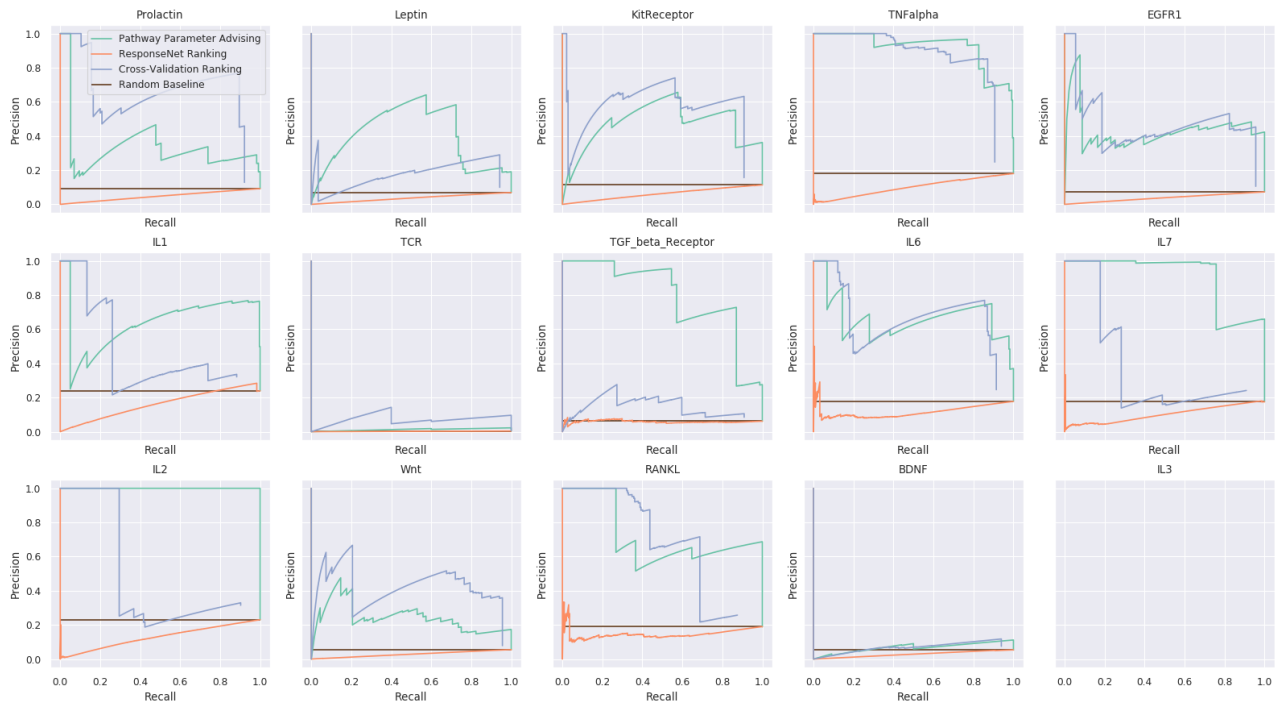

Fig. S7. All PR curves for implausible pathway detection using different parameter ranking schemes for pathways created using minimum-cost flow. Blank axes are where no pathways were found to be plausible, occurring in IL3.

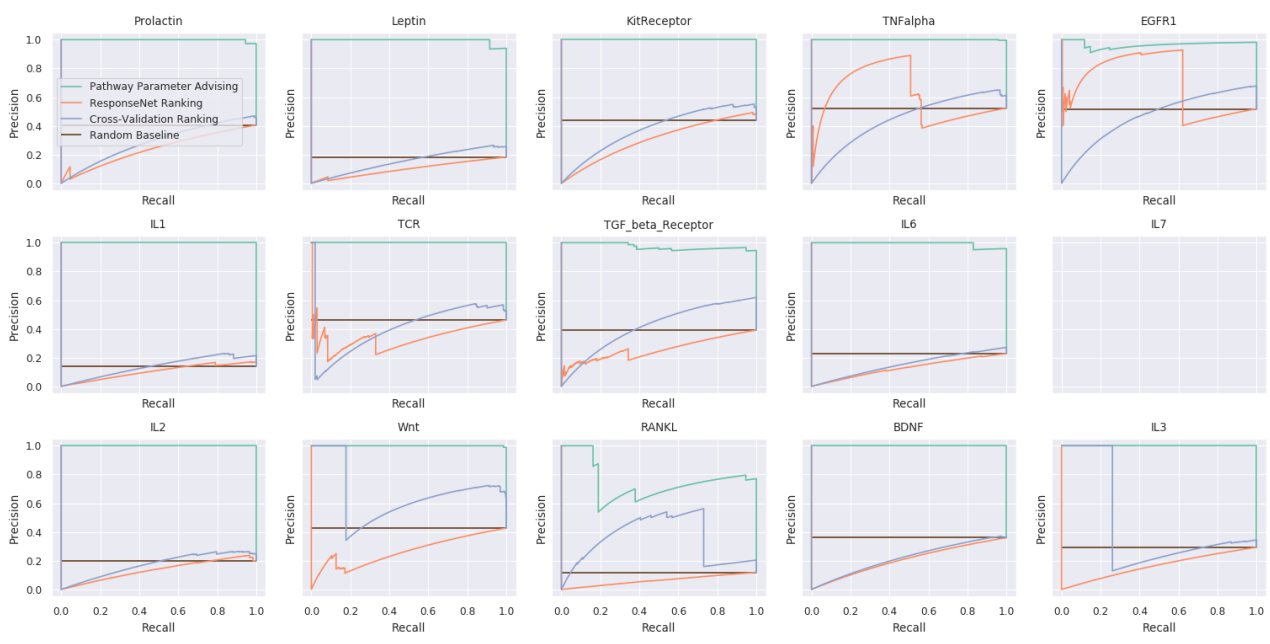

Fig. S8. All PR curves for implausible pathway detection using different parameter ranking schemes for pathways created using PCSF. Blank axes are where no pathways were found to be plausible, occurring in IL7.

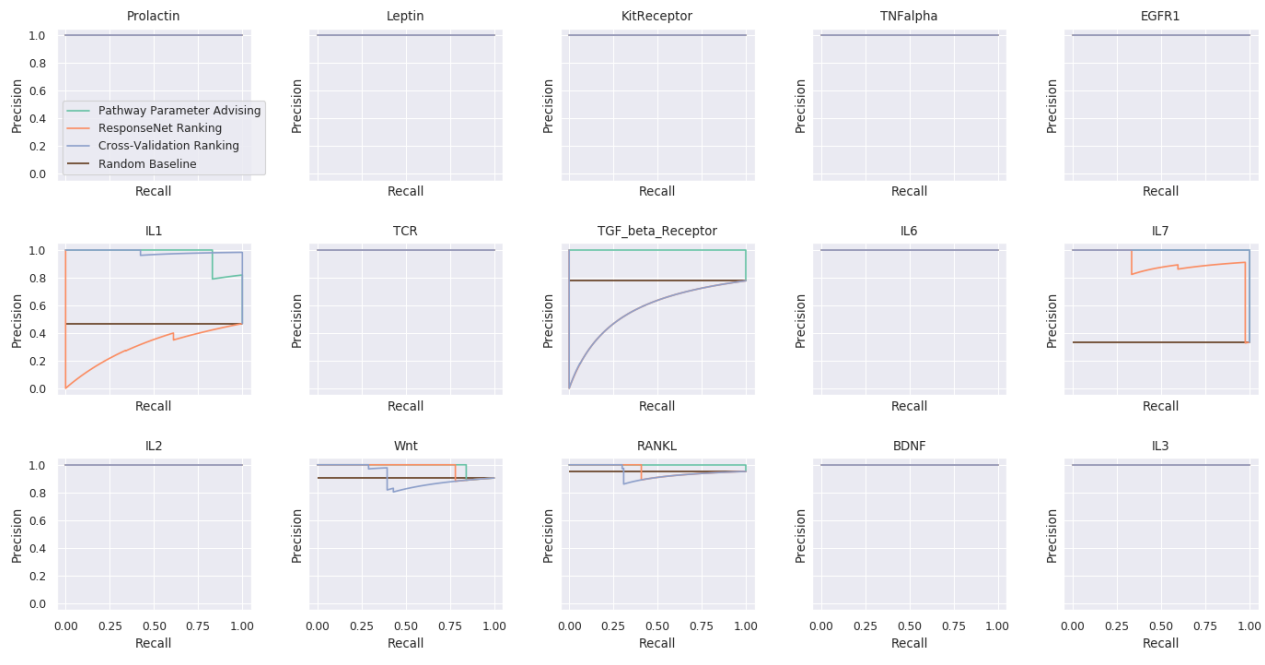

Fig. S9. All PR curves for implausible pathway detection using different parameter ranking schemes for pathways created using NetBox. In 10 cases every generated network met the plausible network criteria.

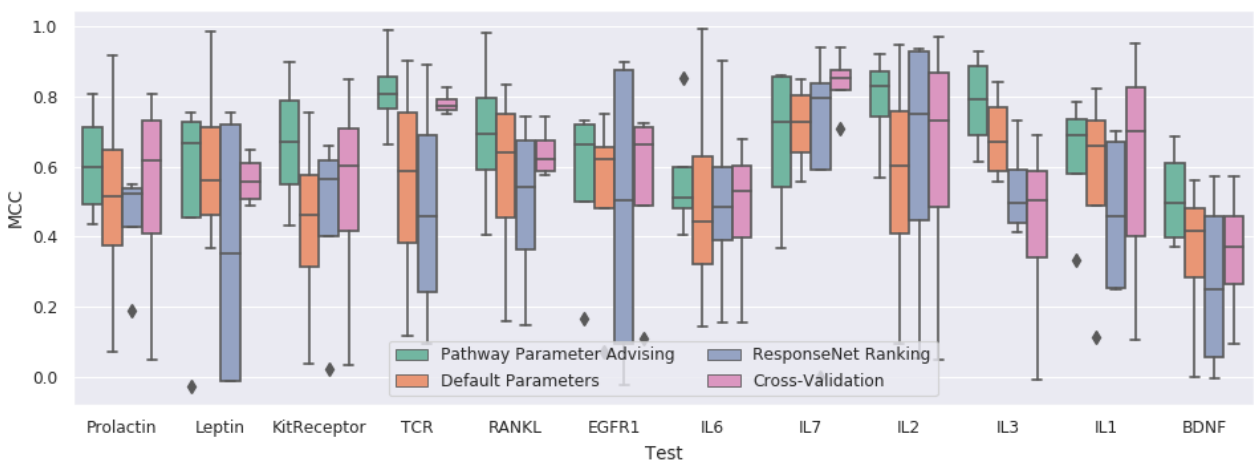

Fig. S10. Adjusted MCC values for different parameter selection methods in NetPath pathway construction.

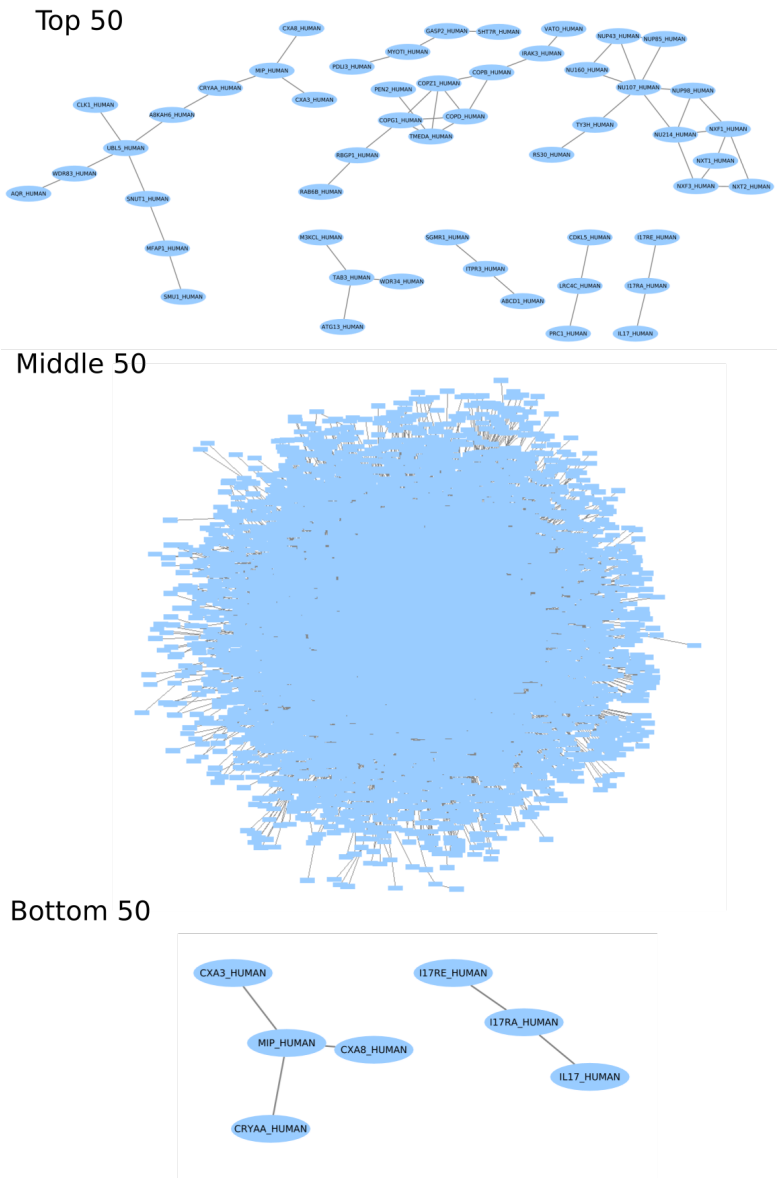

Fig. S11. Influenza host factor networks created from ensembling PCSF runs. The resulting pathways from the top 50, middle 50, and bottom 50 parameter settings as ranked by pathway parameter advising. All connected components over 3 nodes are shown.

S2 Supplementary Tables

Table S1. All GO terms returned from GO-term enrichment with adjusted p-value < 0.05 using DAVID on all nodes in the influenza host factor pathway. The influenza host factor pathway was created using the top 50 pathways ranked by pathway parameter advising constructed using PCSF. The column Benjamini shows p-values corrected using the Benjamini-Hochberg procedure.

Term	Term Name	Count	Benjamini
GO:0006406	mRNA export from nucleus	9	9.07E-06
GO:0006409	tRNA export from nucleus	6	7.65E-05
GO:0010827	regulation of glucose transport	6	5.98E-05
GO:0019083	viral transcription	8	9.81E-05
GO:0007077	mitotic nuclear envelope disassembly	6	0.000157
GO:0075733	intracellular transport of virus	6	0.000277
GO:1900034	regulation of cellular response to heat	6	0.00159
GO:0006890	retrograde vesicle-mediated transport, Golgi to ER	6	0.00215
GO:0031047	gene silencing by RNA	6	0.00806
GO:0016925	protein sumoylation	6	0.00928
GO:0007601	visual perception	7	0.0137
GO:0016032	viral process	8	0.0175
GO:0006886	intracellular protein transport	7	0.0268
GO:0007062	sister chromatid cohesion	5	0.0408

Table S2. The 9 KEGG pathways returned from enrichment using DAVID of all nodes in the influenza host factor pathway. The influenza host factor pathway was created using the top 50 pathways ranked by pathway parameter advising constructed using PCSF. The column Benjamini shows p-values corrected using the Benjamini-Hochberg procedure.

Term	Term Name	Count	Benjamini
hsa03013	RNA transport	10	4.00E-05
hsa05164	Influenza A	6	0.134
hsa03008	Ribosome biogenesis in eukaryotes	4	0.346
hsa05323	Rheumatoid arthritis	4	0.280
hsa03015	mRNA surveillance pathway	4	0.250
hsa05110	Vibrio cholerae infection	3	0.433
hsa00190	Oxidative phosphorylation	4	0.427
hsa04721	Synaptic vesicle cycle	3	0.455
hsa05120	Epithelial cell signaling in Helicobacter pylori infection	3	0.454