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
[3]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, Wisconsin

ABSTRACT

A common way to integrate and analyze large amounts of biological omic data is through pathway finding: creating a subnetwork of a generic background network that represents some process or cellular state using condition-specific omic data. A challenge in pathway finding is that adjusting pathway creation algorithms' parameters produces pathways with drastically different topological properties and biological interpretations. Due to the exploratory nature of pathway finding, 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 finding 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 finding 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 finding 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.[1,2] 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.[3] Thus, it is often preferable to create a customized subnetwork of all known protein interactions, referred to as the interactome, specific to an experimental dataset. We refer to this problem as pathway creation: using condition-specific input omic data to select a subnetwork of a generic background network that represents some process or cellular state.

A variety of pathway creation algorithms exist, which select subnetworks in distinct ways. Existing methods are based on combinatorial optimization problems,[4–6] shortest paths,[7] enrichment analysis,[8]

Send correspondence to AG: gitter@biostat.wisc.edu

network flow,[9–11] and other graph theory algorithms. These methods also take in a variety of inputs. Some, such as the Prize-Collecting Steiner Forest (PCSF) algorithm,[12] accept scores for the biological entities of interest. Other methods, such as PathLinker,[7] 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 creation algorithms almost always require the user to choose parameter settings. 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 (as described in Section A.2.3); 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.

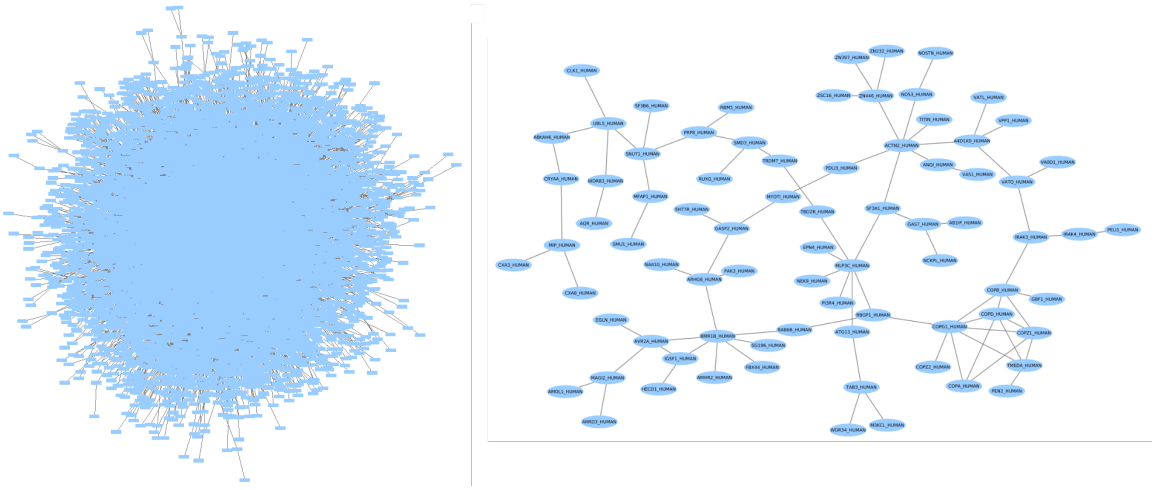

Figure 1. Influenza host factor pathways created using PCSF from RNAi screens (Section 3.3), here showing the largest connected components from ensembling the bottom 100 ranked pathways (left) and the top 100 ranked pathways (right). The only difference in creating the networks was the range of PCSF parameter values.

The simplest way to choose a parameter setting would be to use the default values typically included with pathway creation software. However, it is unlikely that a single parameter setting works for all datasets. The number of proteins, genes, or metabolites of interest can vary based on the experiment, and the effects of input size can be unpredictable for different pathway creation 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. For some methods, it is common to combine multiple runs into an ensemble network.[13–15] In these cases, it is necessary to select multiple parameter settings, which makes using only the default parameters impossible.

Another possible solution could be to apply parameter tuning methods from supervised learning. However, there is no ground truth for supervised parameter tuning, and unsupervised cross-validation is ineffective (Section 3). The objective functions of pathway creation 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[16] and the Bayesian information criterion[17] are not applicable. For instance, the PCSF objective function can be arbitrarily increased by chaging 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.[18] PCSF recommends choosing pathways robust to small random input variation or matching the average degree of input nodes and non-input nodes.[12] 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. Unlike model selection techniques that focus on model accuracy, pathway creation is typically an exploratory analysis. The results are 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, ideally parameter tuning should not focus on traditional notions of accuracy but instead formalize how *useful* generated pathways are to biologists.

One framework for finding optimal parameters in an uncertain setting is parameter advising,[19–21] 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. 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 generated network means that pathway parameter advising is also method agnostic. Pathway parameter advising can tune the parameters of any pathway creation method.

2. METHODS

2.1 Pathway Parameter Advising

Pathway parameter advising is based on the parameter advising[20] framework. 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 creation 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),[22] Reactome,[23] and NetPath,[24] 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,[25] 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 creation 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. A Python implementation of pathway parameter advising is available at https://github.com/gitter-lab/pathway-parameter-advising under the MIT license.

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.[26] However, network motifs typically refer to graphlets that appear in a network significantly more often than expected by chance.

Pathway parameter advising uses the parallel graphlet decomposition library[27] to calculate counts of all graphlets up to size 4 in a pathway. This constitutes 17 possible graphlets, which can be seen in 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.

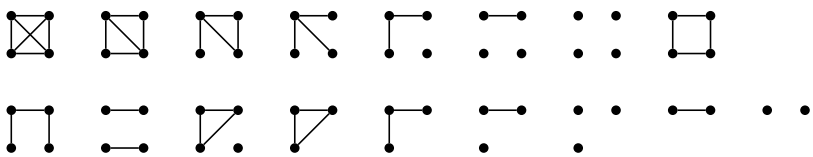

Figure 2. Pathways are decomposed into these 17 graphlets for graphlet frequency distance calculations.

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

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

We considered other graphlet-based metrics such as a variation of relative graphlet frequency distance[28] and graphlet correlation distance,[29] but they performed worse in our preliminary analyses (Figure S1).

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 creation 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 Creation Methods

Table 1 summarizes the 4 methods we used for our implausible pathway detection and reconstruction experiments. Pathway creation algorithms were chosen to have a wide range of methodologies, from NetBox's statistical test to PathLinker's weighted shortest paths algorithm. More detailed information on these methods can be found in Section A.1.

Algorithm	Description	Parameters and Default Values	Range Tested
PathLinker[7]	Connects receptors to transcriptional regulators via weighted k-shortest paths.	**k** (default=100) : Number of shortest paths.	1-1000
NetBox[8]	Hierarchically constructs network using hypergeometric test.	**p** (default=0.05) : p-value threshold for adding an edge	0-1
Prize-Collecting Steiner Forest[4,30]	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[10]	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

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

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.[9] 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. This can be extended to ranking 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,[12] 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.

2.4 Datasets

The interactomes varied by pathway creation algorithm. All parameter tuning was performed with Reactome[23] as the set of reference pathways, while the implausible pathway detection and reconstruction experiments were performed on pathways from the NetPath[24] database. For our case study, we used influenza host factor data gathered from a meta-analysis of RNA interference (RNAi) studies.[31] More details on data used can be found in Section A.2.

2.5 Implausible Network Criteria

In order to examine the ability of pathway parameter advising to avoid parameter settings that lead to implausible or impractical pathways, we created topological criteria that we use to define pathways as plausible or implausible. These are based on previous analyses of biological networks. The specific properties were network size, hub node dependence, clustering coefficient, and assortativity (Section A.5). 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.

2.6 Matthew's Correlation Coefficient

In the NetPath evaluation, we used Matthew's Correlation Coefficient (MCC) to quantify the quality of generated pathways.[32] 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.[33] 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.

3. RESULTS

3.1 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 A.2) to reconstruct the 15 pathways, following the evaluation approach used by PathLinker.[7] This resulted in 60 parameter tuning tasks across the 15 pathways and 4 pathway creation methods. We initially used 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.

We then used precision-recall (PR) curves (Figures S3-S6) to evaluate how well pathway parameter advising and alternative parameter selection strategies distinguish plausible from implausible networks (Section 2.5), treating plausible networks as positive instances. Parameter selection methods that rank plausible networks above implausible networks will have a higher area under the PR curve (AUPR).

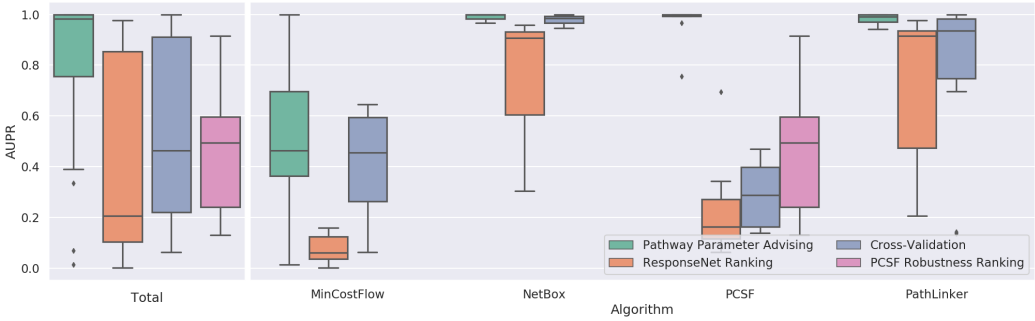

Figure 3. Performance of parameter selection methods on avoiding implausible networks. Boxplots represent the distributions of the AUPRs aggregated for 4 pathway creation methods and 12 test pathway reconstructions from the NetPath database. Degenerate cases where all or no pathways met the plausibility criteria are excluded. Full results, including the 3 validation pathways and degenerate cases, can be found in Figures S3-S6.

Figure 3 shows the distribution of AUPRs across the 4 pathway creation 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 creation 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.

3.2 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.

Figure 4 (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 3). When stratified by pathway creation 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 4 (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.

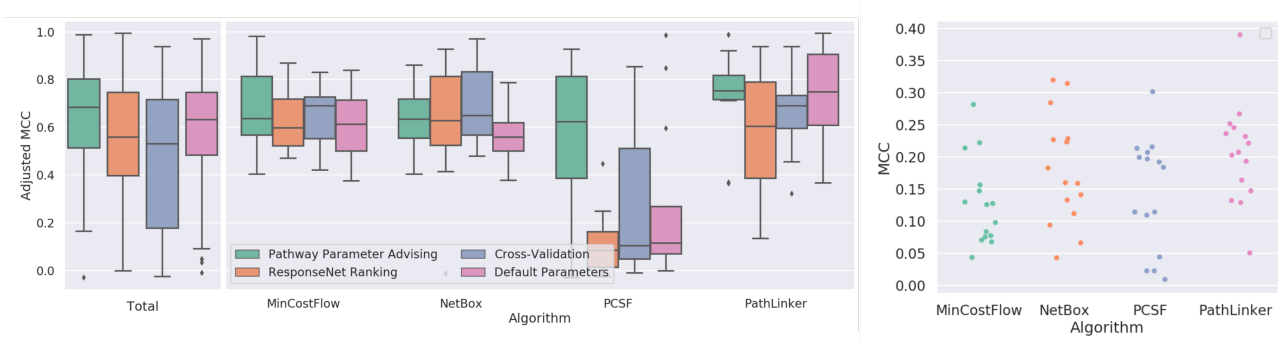

Figure 4. **Left:** Adjusted MCC of parameter selection methods on reconstructing 12 test pathways from the NetPath database across 4 pathway creation methods. MCCs were normalized to the highest possible MCC within a given pathway creation method and pathway. Figure S7 shows MCC values by pathway, including the 3 validation pathways excluded here. **Right:** The highest possible MCC of pathway reconstruction in 60 parameter sweeps across 4 pathway creation methods and 15 NetPath pathways (validation and test). The MCC values are generally low, reflecting low overlap between the predicted and NetPath pathway edges.

3.3 Influenza Host Factor Pathway Prediction

To demonstrate how pathway parameter advising can guide the biological interpretation of omic data, we reconstructed a pathway involving influenza host factors. 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.[31] 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 A.2.3).

After creating the candidate host factor pathways, we ranked the parameter settings using pathway parameter advising. Figure 5 (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.

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, while the middle ranked and low ranked ensemble pathways have 7337 and 15 nodes, respectively (Figure S8). 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[34] (Section A.6). 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

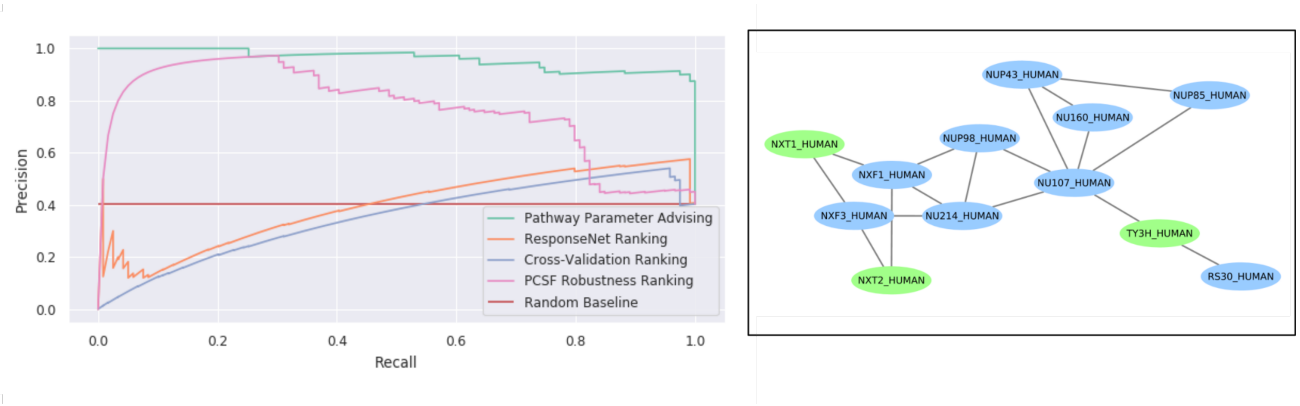

Figure 5. **Left**: Precision-recall curve for implausible networks in PCSF influenza host factor network construction. **Right**: A component of the influenza host factor ensemble pathway created from the top 50 PCSF parameter settings ranked by pathway parameter advising. This component represents 12 of the 86 total nodes in the pathway (Figure S8). Host factor nodes provided as input are shown in blue, while green nodes are "Steiner" nodes that PCSF predicts to connect the host factors.

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.[35] 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 5 (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.[36] This demonstrates the ability of pathways chosen through pathway parameter advising to guide discovery.

4. DISCUSSION

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

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 when performing a pathway analysis. Currently, these types of generated pathways are typically ignored through a process of manual trial and error. Any manual step in the process of 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. 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 other

experiment, this further highlights the effectiveness of pathway parameter advising. Much of the total performance is driven by how well it performs in PCSF, though pathway parameter advising is also the only one to be either the first or second highest performing for all 4 pathway creation 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.

In both experiments, other parameter selection methods especially struggled choosing parameters for PCSF. This is unsurprising as PCSF contains multiple parameters that deeply affect pathway structure and interact with each other. In contrast, a method like PathLinker contains a single parameter, which monotonically increases pathway size. Changing the parameter value has a relatively predictable effect. 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.

There are some drawbacks to pathway parameter advising. Pathway parameter advising requires running 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. Although the popular pathway database Reactome works well in analyses here, if the optimal predicted pathway is reasonable but outside the range of topologies seen among the reference pathways, it would be overlooked. Beyond more complete pathway databases, more tools are needed for biologists to be able to both create and interpret pathways for hypothesis generation. Though pathway parameter advising can select more plausible pathways, pathway interpretation is still difficult.

Our distance metric focuses only on topology and does not include any information about the biological context. ResponseNet[37] and PathLinker[38] extensions consider tissue-specificity and protein localization context, respectively. A possible extension of pathway parameter advising would be to include 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.

Now that we have demonstrated the utility of pathway parameter advising to evaluate predicted pathways, it can be wrapped with hyperparameter optimization to fully automate the process with a Bayesian optimization framework.[39–41] This will allow for complete and standardized automation of the pathway creation 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.

ACKNOWLEDGMENTS

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. 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).

REFERENCES

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

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

[3] Köksal, A. S., Beck, K., Cronin, D. R., McKenna, A., Camp, N. D., Srivastava, S., MacGilvray, M. E., Bodík, R., Wolf-Yadlin, A., Fraenkel, E., Fisher, J., and Gitter, A., "Synthesizing Signaling Pathways from Temporal Phosphoproteomic Data," *Cell Reports* **24**, 3607–3618 (Sept. 2018).

[4] Tuncbag, N., Gosline, S. J. C., Kedaigle, A., Soltis, A. R., Gitter, A., and Fraenkel, E., "Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package," *PLOS Computational Biology* **12**, e1004879 (Apr. 2016).

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

[6] Yosef, N., Ungar, L., Zalckvar, E., Kimchi, A., Kupiec, M., Ruppin, E., and Sharan, R., "Toward accurate reconstruction of functional protein networks," *Molecular Systems Biology* **5**, 248 (Jan. 2009).

[7] Ritz, A., Poirel, C. L., Tegge, A. N., Sharp, N., Simmons, K., Powell, A., Kale, S. D., and Murali, T. M., "Pathways on demand: automated reconstruction of human signaling networks," *npj Systems Biology and Applications* **2**, 1–9 (Mar. 2016).

[8] Cerami, E., Demir, E., Schultz, N., Taylor, B. S., and Sander, C., "Automated Network Analysis Identifies Core Pathways in Glioblastoma," *PLOS ONE* **5**, e8918 (Feb. 2010).

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

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

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

[12] Kedaigle, A. J. and Fraenkel, E., "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 (2018).

[13] MacGilvray, M. E., Shishkova, E., Chasman, D., Place, M., Gitter, A., Coon, J. J., and Gasch, A. P., "Network inference reveals novel connections in pathways regulating growth and defense in the yeast salt response," *PLOS Computational Biology* **13**(5), 1–28 (2018).

[14] Budak, G., Ozsoy, O. E., Son, Y. A., Can, T., and Tuncbag, N., "Reconstruction of the temporal signaling network in Salmonella-infected human cells," *Frontiers in Microbiology* **6**, 730 (2015).

[15] Khurana, V., Peng, J., Chung, C. Y., Auluck, P. K., Fanning, S., Tardiff, D. F., Bartels, T., Koeva, M., Eichhorn, S. W., Benyamini, H., Lou, Y., Nutter-Upham, A., Baru, V., Freyzon, Y., Tuncbag, N., Costanzo, M., San Luis, B.-J., Schöndorf, D. C., Barrasa, M. I., Ehsani, S., Sanjana, N., Zhong, Q., Gasser, T., Bartel, D. P., Vidal, M., Deleidi, M., Boone, C., Fraenkel, E., Berger, B., and Lindquist, S., "Genome-Scale Networks Link Neurodegenerative Disease Genes to α-Synuclein through Specific Molecular Pathways," *Cell Systems* **4**, 157–170.e14 (Feb. 2017).

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

[17] Schwarz, G., "Estimating the Dimension of a Model," *The Annals of Statistics* **6**, 461–464 (Mar. 1978).

[18] Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D., Cashikar, A. G., King, O. D., Auluck, P. K., Geddie, M. L., Valastyan, J. S., Karger, D. R., Lindquist, S., and Fraenkel, E., "Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity," *Nature Genetics* **41**, 316–323 (Mar. 2009).

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

[20] DeBlasio, D. and Kececioglu, J., "Parameter advising for multiple sequence alignment," *BMC Bioinformatics* **16**, A3 (Jan. 2015).

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

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

[23] Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P., "The Reactome Pathway Knowledgebase," *Nucleic Acids Research* **46**(D1), D649–D655 (2018).

[24] Kandasamy, K., Mohan, S. S., Raju, R., Keerthikumar, S., Kumar, G. S. S., Venugopal, A. K., Telikicherla, D., Navarro, J. D., Mathivanan, S., Pecquet, C., Gollapudi, S. K., Tattikota, S. G., Mohan, S., Padhukasahasram, H., Subbannayya, Y., Goel, R., Jacob, H. K. C., Zhong, J., Sekhar, R., Nanjappa, V., Balakrishnan, L., Subbaiah, R., Ramachandra, Y. L., Rahiman, B. A., Prasad, T. S. K., Lin, J.-X., Houtman, J. C. D., Desiderio, S., Renauld, J.-C., Constantinescu, S. N., Ohara, O., Hirano, T., Kubo, M., Singh, S., Khatri, P., Draghici, S., Bader, G. D., Sander, C., Leonard, W. J., and Pandey, A., "NetPath: a public resource of curated signal transduction pathways," *Genome Biology* **11**, R3 (Jan. 2010).

[25] Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., and Mering, C. v., "STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research* **47**, D607–D613 (Jan. 2019).

[26] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U., "Network Motifs: Simple Building Blocks of Complex Networks," *Science* **298**(5594), 824–827 (2002).

[27] Ahmed, N. K., Neville, J., Rossi, R. A., and Duffield, N., "Efficient graphlet counting for large networks," in [*2015 IEEE International Conference on Data Mining*], 1–10 (Nov. 2015).

[28] Prulj, N., Corneil, D. G., and Jurisica, I., "Modeling interactome: scale-free or geometric?," *Bioinformatics* **20**, 3508–3515 (July 2004).

[29] Yaverolu, ö. N., Malod-Dognin, N., Davis, D., Levnajic, Z., Janjic, V., Karapandza, R., Stojmirovic, A., and Prulj, N., "Revealing the Hidden Language of Complex Networks," *Scientific Reports* **4**, 4547 (Apr. 2014).

[30] Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.-S. C., Chayes, J., Borgs, C., Zecchina, R., and Fraenkel, E., "Simultaneous Reconstruction of Multiple Signaling Pathways via the Prize-Collecting Steiner Forest Problem," *Journal of Computational Biology* **20**, 124–136 (Feb. 2013).

[31] Tripathi, S., Pohl, M., Zhou, Y., Rodriguez-Frandsen, A., Wang, G., Stein, D., Moulton, H., DeJesus, P., Che, J., Mulder, L., Yángüez, E., Andenmatten, D., Pache, L., Manicassamy, B., Albrecht, R., Gonzalez, M., Nguyen, Q., Brass, A., Elledge, S., White, M., Shapira, S., Hacohen, N., Karlas, A., Meyer, T., Shales, M., Gatorano, A., Johnson, J., Jang, G., Johnson, T., Verschueren, E., Sanders, D., Krogan, N., Shaw, M., König, R., Stertz, S., Garca-Sastre, A., and Chanda, S., "Meta- and orthogonal integration of influenza omics data defines a role for UBR4 in virus budding," *Cell Host & Microbe* **18**(6), 723 – 735 (2015).

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

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

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

[35] Dou, D., Revol, R., Östbye, H., Wang, H., and Daniels, R., "Influenza a virus cell entry, replication, virion assembly and movement," *Frontiers in Immunology* **9**, 1581–1581 (July 2018).

[36] Han, J., Perez, J. T., Chen, C., Li, Y., Benitez, A., Kandasamy, M., Lee, Y., Andrade, J., tenOever, B., and Manicassamy, B., "Genome-wide CRISPR/Cas9 screen identifies host factors essential for influenza virus replication," *Cell Reports* **23**(2), 596 – 607 (2018).

[37] Basha, O., Mauer, O., Simonovsky, E., Shpringer, R., and Yeger-Lotem, E., "ResponseNet v.3: revealing signaling and regulatory pathways connecting your proteins and genes across human tissues," *Nucleic Acids Research* **47**, W242–W247 (July 2019).

[38] Youssef, I., Law, J., and Ritz, A., "Integrating Protein Localization with Automated Signaling Pathway Reconstruction," in [*2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*], 163–168 (Dec. 2018).

[39] Kandasamy, K., Vysyaraju, K. R., Neiswanger, W., Paria, B., Collins, C. R., Schneider, J., Poczos, B., and Xing, E. P., "Tuning hyperparameters without grad students: Scalable and robust Bayesian optimisation with Dragonfly," *arXiv* (1903.06694) (2019).

[40] Wang, Z., Zoghi, M., Hutter, F., Matheson, D., and De Freitas, N., "Bayesian Optimization in High Dimensions via Random Embeddings," in [*Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*], *IJCAI '13*, 1778–1784, AAAI Press (2013).

[41] Chen, B., Castro, R. M., and Krause, A., "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*, 1379–1386, Omnipress, USA (2012).

[42] Bailly-Bechet, M., Borgs, C., Braunstein, A., Chayes, J., Dagkessamanskaia, A., François, J.-M., and Zecchina, R., "Finding undetected protein associations in cell signaling by belief propagation," *Proceedings of the National Academy of Sciences* **108**, 882–887 (Jan. 2011).

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

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

[45] Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E., "PhosphoSitePlus, 2014: mutations, PTMs and recalibrations," *Nucleic Acids Research* **43**, D512–D520 (Dec. 2014).

[46] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L., "Hierarchical Organization of Modularity in Metabolic Networks," *Science* **297**(5586), 1551–1555 (2002).

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

[48] Newman, M. E. J., "Assortative mixing in networks," *Physical Review Letters* **89**, 208701 (Oct. 2002).

[49] Albert, R., DasGupta, B., Hegde, R., Sivanathan, G. S., Gitter, A., Gürsoy, G., Paul, P., and Sontag, E., "Computationally efficient measure of topological redundancy of biological and social networks," *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* **84**, 036117 (Sept. 2011).

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

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

Chris Magnano and Anthony Gitter

APPENDIX A. SUPPLEMENTARY NOTE

A.1 Pathway Creation Methods

- NetBox: NetBox[8] 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,[4,12,30] nodes are assigned prizes and edges are given costs. The optimal subnetwork, which is found via a message-passing algorithm,[42] 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:

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

  where $p()$ is the positive prize for each node, $c()$ is the cost of each edge, $d()$ is a node's degree, and κ is the number of connected components in the pathway. This 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.

- 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 certain 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.[10] We implemented a version of min-cost flow using the solver provided in Google's OR-Tools*, which solves the min-cost flow problem using the algorithm outlined in Bünnagel, Korte and Vygen.[43] This is a generic version of the algorithm used in ResponseNet.[9] 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.

*https://developers.google.com/optimization/flow/mincostflow

- PathLinker: PathLinker[7] 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.

A.2 Datasets

A.2.1 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 created an interactome by merging protein interactions from the iRefIndex database v13[44] and kinase-substrate interactions from PhosphoSitePlus.[45] 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.

A.2.2 Pathway databases

All parameter tuning was performed with Reactome as the set of reference pathways. Reactome[23] 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.[24] We used 15 NetPath pathways that contain at least 1 receptor and transcriptional regulator and are sufficiently connected, as described by Ritz *et al.*[7] We sampled the NetPath pathways in different ways for each pathway creation 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.*[7]

A.2.3 Influenza host factors

Influenza host factor data was gathered from a meta-analysis of 8 RNAi studies.[31] The meta-analysis used the raw RNAi screen data to calculate a consolidated Z score for a total of 1257 host factor genes.

A.3 Pathway parameter advising implementation

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.

A.4 Metric Choice

We explored 3 graphlet-based metrics for pathway parameter advising: graphlet frequency distance (GFD), normalized graphlet frequency distance (NGFD), and graphlet correlation distance (GCD). For NGFD, we wanted to explore making the metric take 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 graphlets in a pathway.[29] This creates a metric that is solely focused on local topology and has a minimal signal of pathway size or other global topological properties.

Figure S1 shows adjusted MCCs of parameter advising using these 3 graphlet metrics across all 4 pathway creation methods on the 3 NetPath validation pathways: Wnt, TGF Beta, and TNF Alpha. 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 helps in identifying which pathways are similar to reference pathways.

A.5 Plausible Pathway Criteria

In the implausible pathway detection experiment, we created an ad-hoc definition of what topological measurements a reasonable biological pathway would be expected to meet. Our criteria were 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 creation 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.[12] 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;[46] 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.[47] 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.[48,49] 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.

While the criteria laid out for defining a plausible network is useful for comparing networks created by the same method with different parameter settings, it should not be considered as a metric for comparing pathways across pathway creation methods. Different pathway creation methods 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.

A.6 Enrichment Analyses

GO[50] and KEGG pathway[22] enrichment was carried out with DAVID[34] v6.7. 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.

APPENDIX B. SUPPLEMENTARY FIGURES

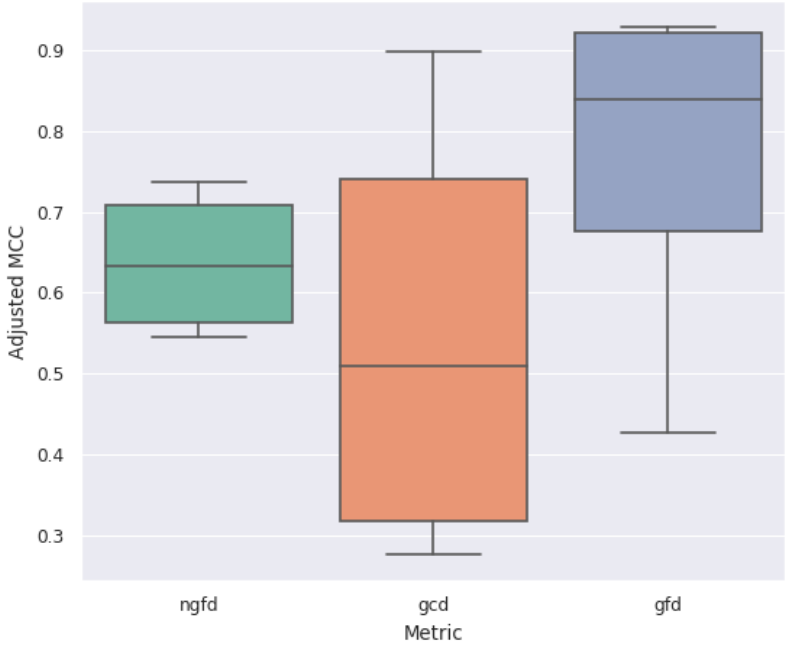

Figure S1. Adjusted MCC on pathway reconstruction of the validation pathways Wnt, TNF Apha, and TGF Beta across 4 pathway creation algorithms. Here, we consider 3 graphlet-based distance metrics: normalized graphlet frequency distance (ngfd), graphlet correlation distance (gcd), and graphlet frequency distance (gfd). We used gfd in pathway parameter advising.

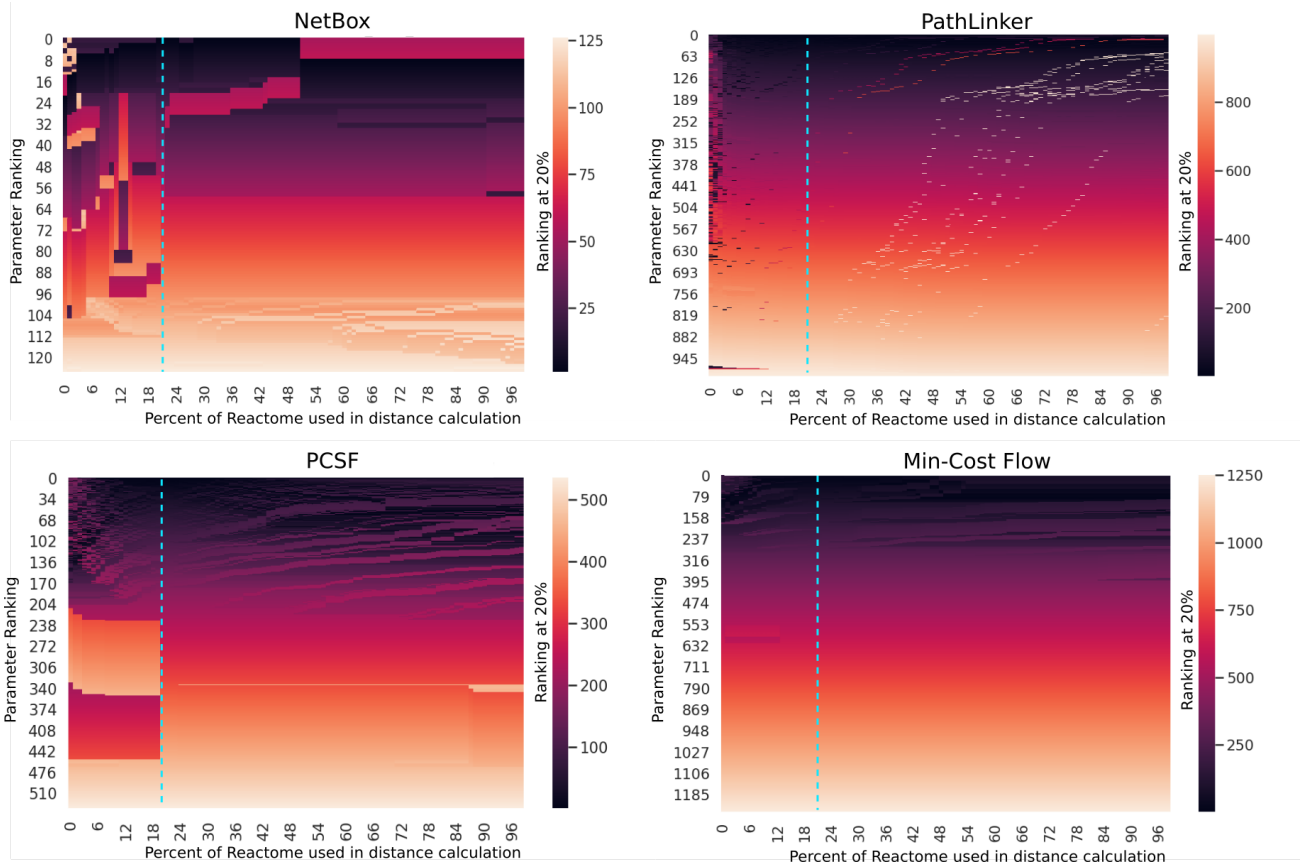

Figure S2. Change in ranking order over different percent thresholds for calculating score in pathway parameter advising on the TGF beta receptor NetPath pathway. Heatmaps are colored based on the ranking at the chosen threshold, 20% (marked by a blue dashed line). While the parameter ranking is unstable 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.

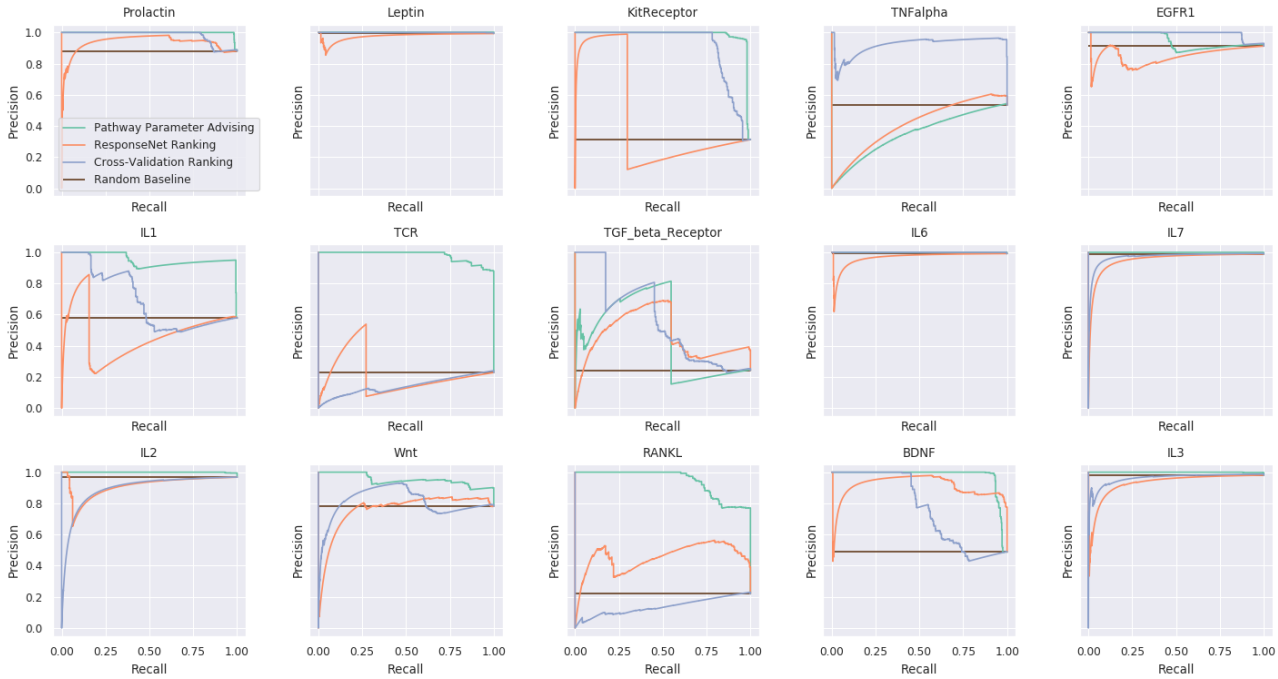

Figure S3. All PR curves for implausible pathway detection using different parameter ranking schemes for pathways created using PathLinker.

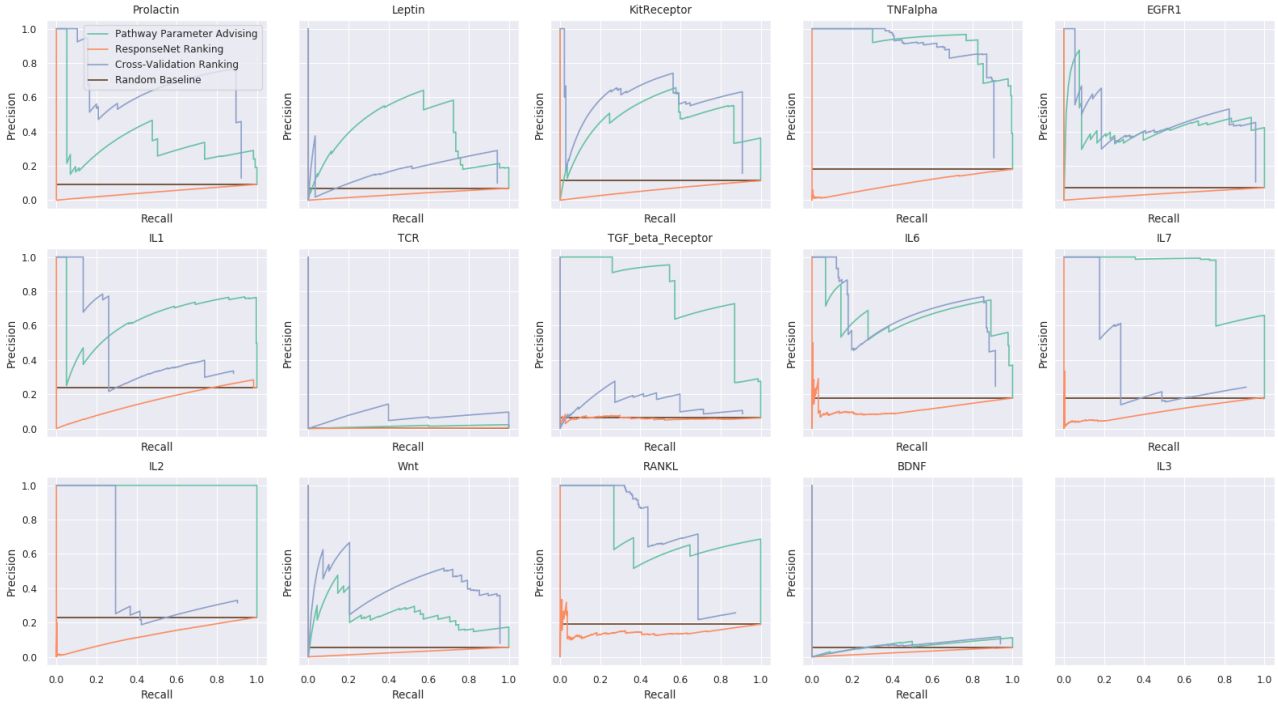

Figure S4. All PR curves for implausible pathway detection using different parameter ranking schemes for pathways created using minimum-cost flow. Blank axes are where no pathways were found to be plausible, occuring in IL3.

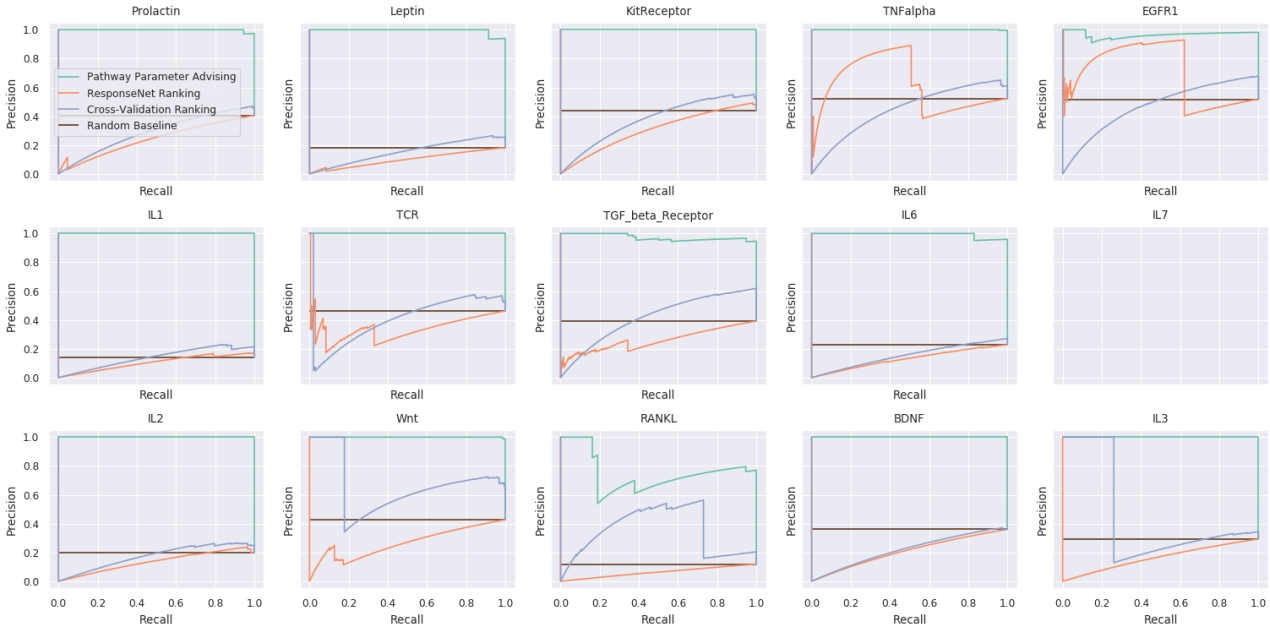

Figure S5. All PR curves for implausible pathway detection using different parameter ranking schemes for pathways created using PCSF. Blank axes are where no pathways were found to be plausible, occurring in IL7.

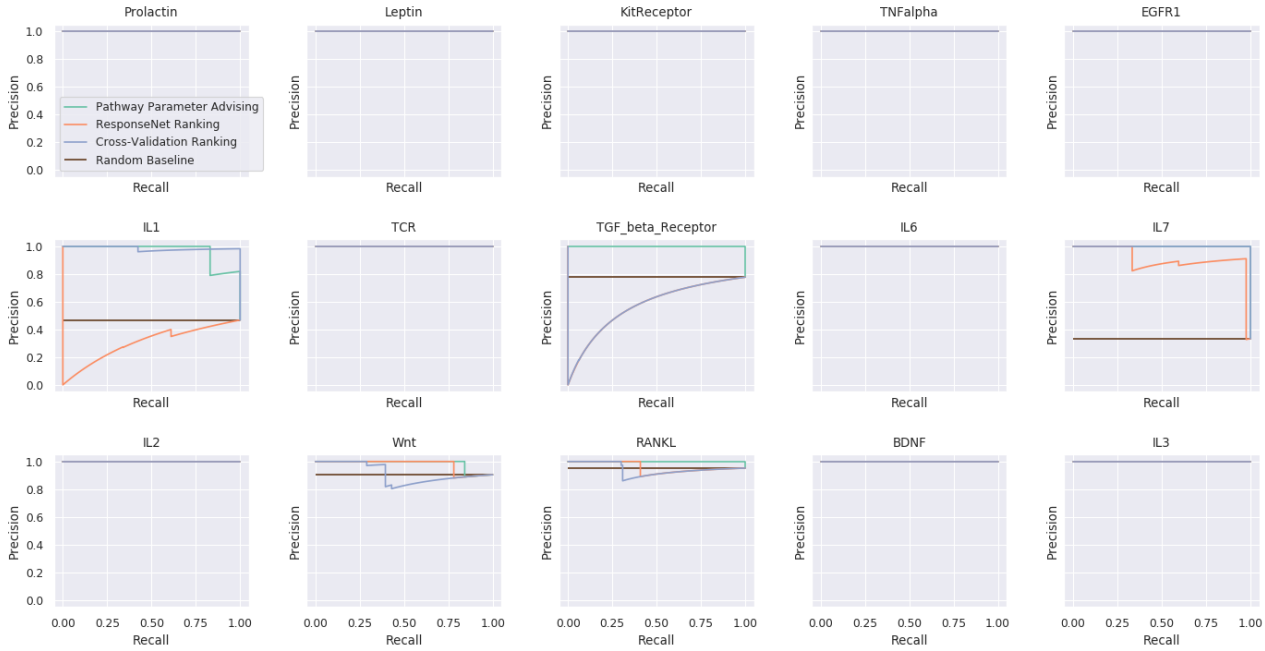

Figure S6. All PR curves for implausible pathway detection using different parameter ranking schemes for pathways created using NetBox. In 10 cases every generated network met the plausible network criteria.

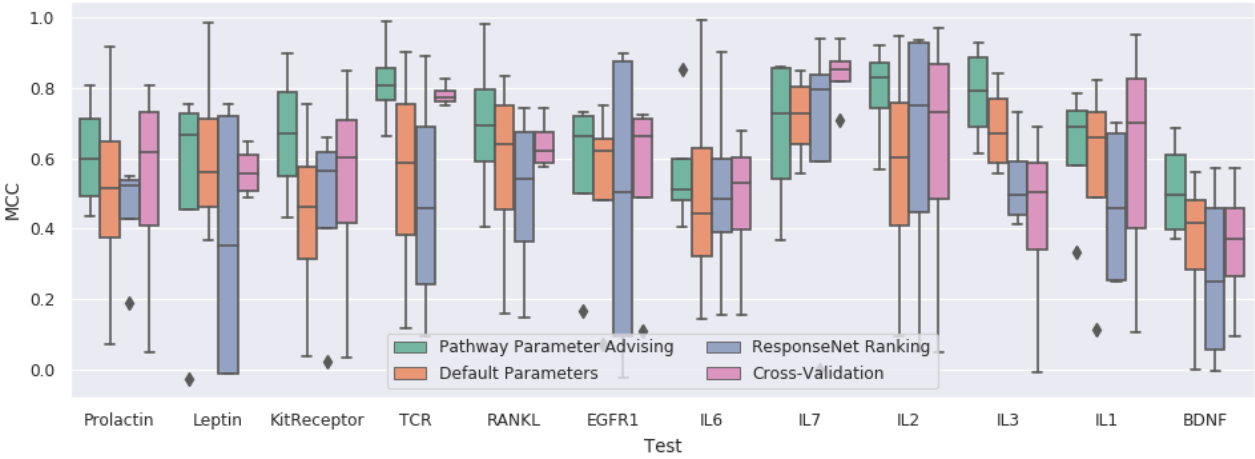

Figure S7. Adjusted MCC values for different parameter selection methods in NetPath pathway construction.

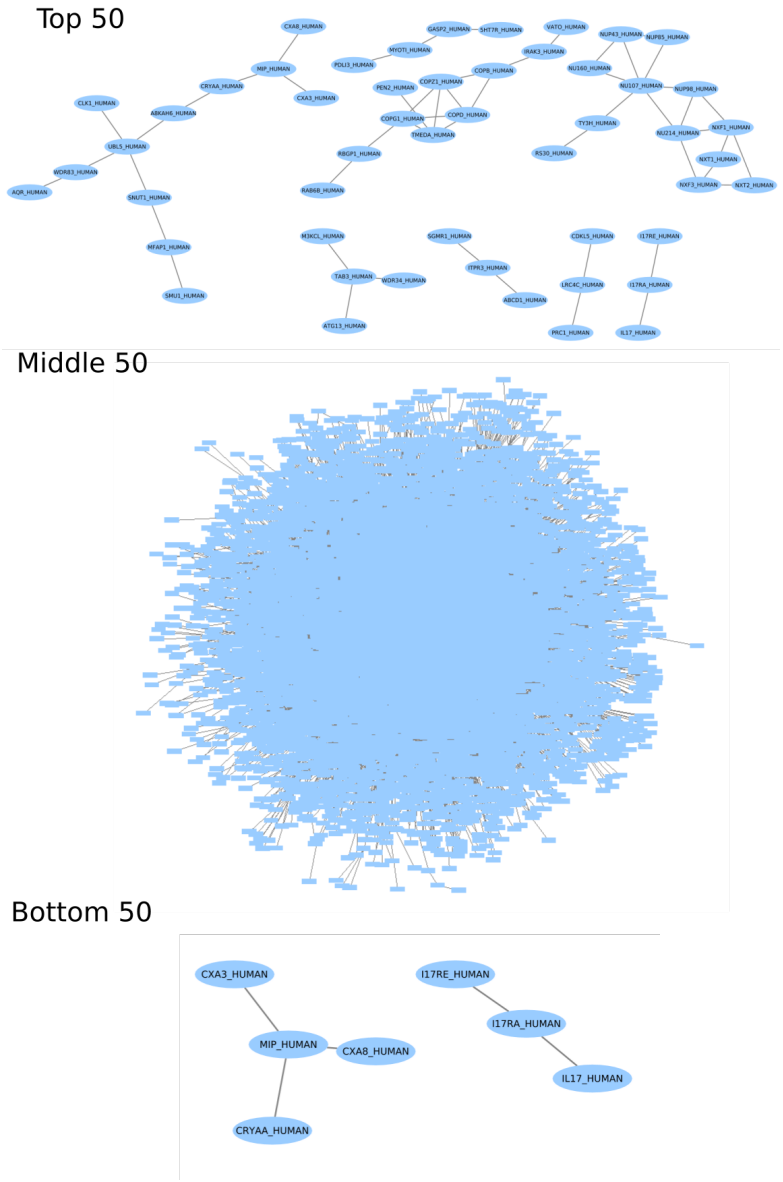

Figure S8. Influenza host factor networks created from ensembling PCSF runs. The resulting pathways from the top 50, middle 50, and bottom 50 parameter settings as ranked by pathway parameter advising. All connected components over 3 nodes are shown.

APPENDIX C. SUPPLEMENTARY TABLES

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 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
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

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.