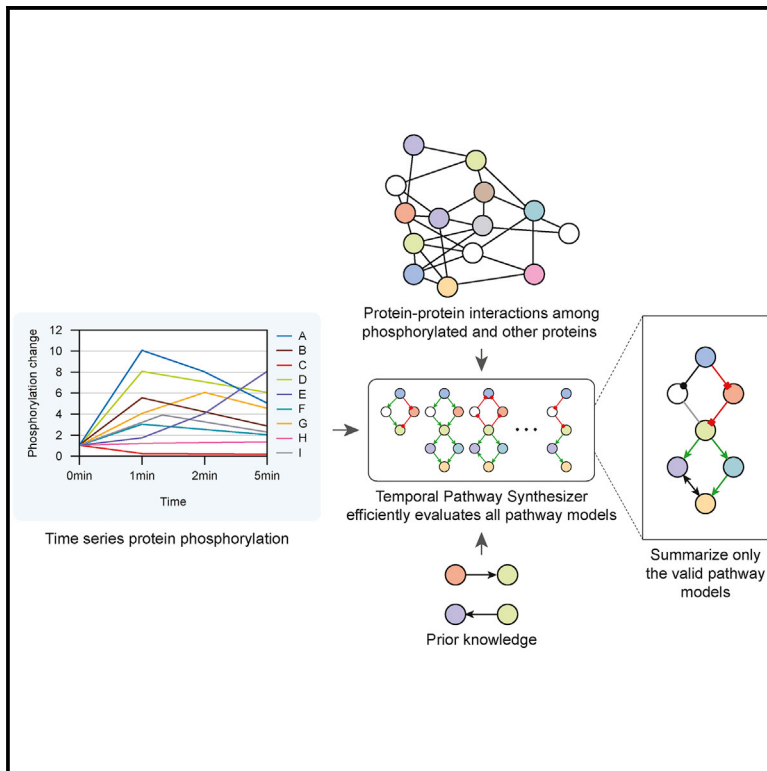


# Cell Reports

## Synthesizing Signaling Pathways from Temporal Phosphoproteomic Data

### Graphical Abstract



### Authors

Ali Sinan Köksal, Kirsten Beck,  
Dylan R. Cronin, ..., Ernest Fraenkel,  
Jasmin Fisher, Anthony Gitter

### Correspondence

gitter@biostat.wisc.edu

### In Brief

Köksal et al. present a computational technique, the temporal pathway synthesizer (TPS), that combines time series global phosphoproteomic data and protein-protein interaction networks to reconstruct the vast signaling pathways that control post-translational modifications.

### Highlights

- TPS predicts signaling pathways that explain large-scale phosphorylation changes
- TPS models dynamics of individual phosphosites in a protein network context
- Existing algorithms do not recover the temporal relationships detected by TPS

### Data and Software Availability

PXD006697



# Synthesizing Signaling Pathways from Temporal Phosphoproteomic Data

Ali Sinan Köksal,<sup>1</sup> Kirsten Beck,<sup>2</sup> Dylan R. Cronin,<sup>3,4</sup> Aaron McKenna,<sup>2</sup> Nathan D. Camp,<sup>2</sup> Saurabh Srivastava,<sup>1</sup> Matthew E. MacGilvray,<sup>5</sup> Rastislav Bodík,<sup>6</sup> Alejandro Wolf-Yadlin,<sup>2,11</sup> Ernest Fraenkel,<sup>7</sup> Jasmin Fisher,<sup>8,9</sup> and Anthony Gitter<sup>3,10,12,\*</sup>

<sup>1</sup>Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA

<sup>2</sup>Department of Genome Sciences, University of Washington, Seattle, WA, USA

<sup>3</sup>Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA

<sup>4</sup>Department of Biological Sciences, Bowling Green State University, Bowling Green, OH, USA

<sup>5</sup>Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI, USA

<sup>6</sup>Paul G. Allen Center for Computer Science and Engineering, University of Washington, Seattle, WA, USA

<sup>7</sup>Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

<sup>8</sup>Microsoft Research, Cambridge, UK

<sup>9</sup>Department of Biochemistry, University of Cambridge, Cambridge, UK

<sup>10</sup>Morgridge Institute for Research, Madison, WI, USA

<sup>11</sup>Present address: Seattle Genetics, Seattle, WA, USA

<sup>12</sup>Lead Contact

\*Correspondence: [gitter@biostat.wisc.edu](mailto:gitter@biostat.wisc.edu)

<https://doi.org/10.1016/j.celrep.2018.08.085>

## SUMMARY

We present a method for automatically discovering signaling pathways from time-resolved phosphoproteomic data. The Temporal Pathway Synthesizer (TPS) algorithm uses constraint-solving techniques first developed in the context of formal verification to explore paths in an interaction network. It systematically eliminates all candidate structures for a signaling pathway where a protein is activated or inactivated before its upstream regulators. The algorithm can model more than one hundred thousand dynamic phosphosites and can discover pathway members that are not differentially phosphorylated. By analyzing temporal data, TPS defines signaling cascades without needing to experimentally perturb individual proteins. It recovers known pathways and proposes pathway connections when applied to the human epidermal growth factor and yeast osmotic stress responses. Independent kinase mutant studies validate predicted substrates in the TPS osmotic stress pathway.

## INTRODUCTION

High-throughput proteomic assays illuminate the amazing breadth and complexity of the signal transduction pathways that cells employ to respond to extracellular cues. These technologies can quantify protein abundance or post-translational modifications (PTMs). Mass spectrometry, in particular, offers a broad view of PTMs, including phosphorylation, ubiquitination, acetylation, and methylation (Choudhary and Mann, 2010), and is not restricted to a predefined list of proteins. Here, we show

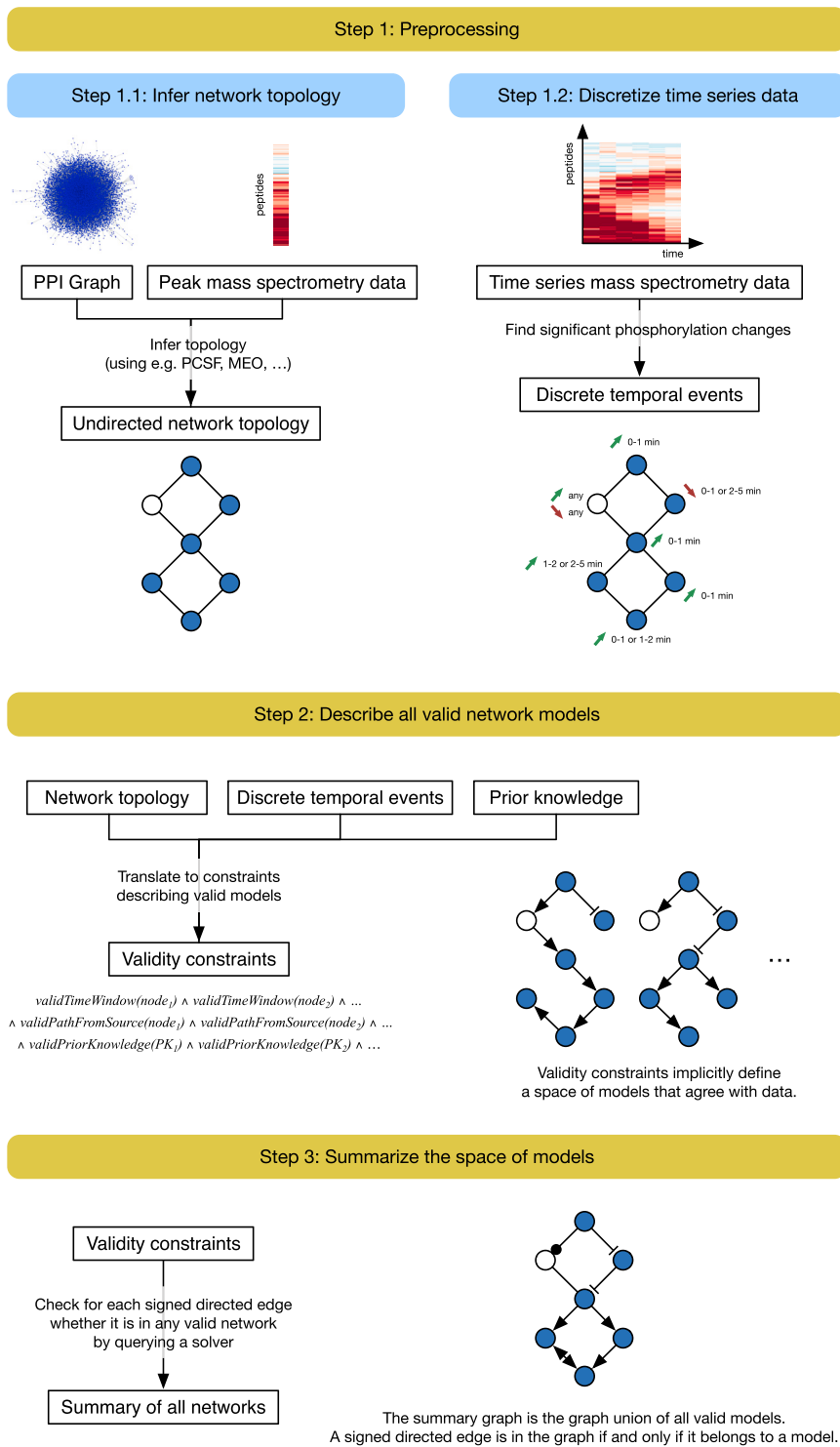
how to discover new facets of signaling cascades from complex proteomic data by integrating observed PTMs with existing knowledge of protein interactions.

Many gaps persist in our understanding of phosphorylation signaling cascades. For example, our mass spectrometry experiments show that nearly all proteins that are significantly (de)phosphorylated when the epidermal growth factor receptor (EGFR) is stimulated are absent from EGFR pathway maps. The low overlap is consistent with previous temporal phosphoproteomic studies of mammalian signaling (Cao et al., 2012; D'Souza et al., 2014; Humphrey et al., 2015). Discordance between mass spectrometry studies and pathway databases can be caused by extensive crosstalk among pathways (Bauer-Mehren et al., 2009), context-specific interactions (Hill et al., 2017), cell- and tissue-specific protein abundance (Kim et al., 2014), and signaling pathway rewiring (Pawson and Warner, 2007).

Network inference algorithms can explain the phosphorylation events that lie outside of canonical pathways and complement curated pathway maps. Specialized algorithms model time series data, which inform the ordering of phosphorylation changes and support causal instead of correlative modeling (Bar-Joseph et al., 2012). Temporal protein signaling information can be used to reconstruct more accurate and complete networks than a single static snapshot of the phosphoproteome.

A complementary challenge to interpreting off-pathway phosphorylation is that the cellular stimulus response includes mechanisms that are not captured in phosphoproteomic datasets. There is an interplay between phosphorylation changes and other integral parts of signaling cascades. Phosphorylation can affect protein stability, subcellular localization, and recognition of interaction partners (Newman et al., 2014). Phosphoproteomic studies measure only one type of PTM, and not all phosphorylated proteins are detected by mass spectrometry. Additional information is required to infer comprehensive signaling cascades that include non-differentially phosphorylated proteins.





**Figure 1. TPS Workflow**

First, the PPI graph is combined with the phosphorylation data to obtain a condition-specific network (step 1.1). This step does not model the temporal information and instead uses the phosphorylation peak, the highest magnitude fold change. Separately, the time series data are converted into discrete timed signaling events (step 1.2). TPS then defines a space of models that agree with the data by transforming the timed events, undirected network topology, and prior knowledge (kinase-substrate interaction directions in this study) into a set of constraints (step 2). It summarizes the solution space by computing the union of all signed, directed graph models that satisfy the given constraints (step 3). The final pathway model predicts how a subset of generic physical protein interactions coordinates to respond to a specific stimulus in a particular cellular context.

data into signaling pathways that extend beyond existing canonical maps. “Synthesizer” refers to applying computational program synthesis techniques (Manna and Waldinger, 1980) to produce pathway models from experimental data (Fisher et al., 2014), not synthetic biology (Benner and Sismour, 2005). TPS overcomes both of the aforementioned challenges in interpreting phosphoproteomic data: modeling signaling events that are not captured by pathway databases and including non-phosphorylated proteins in the predicted pathway structures.

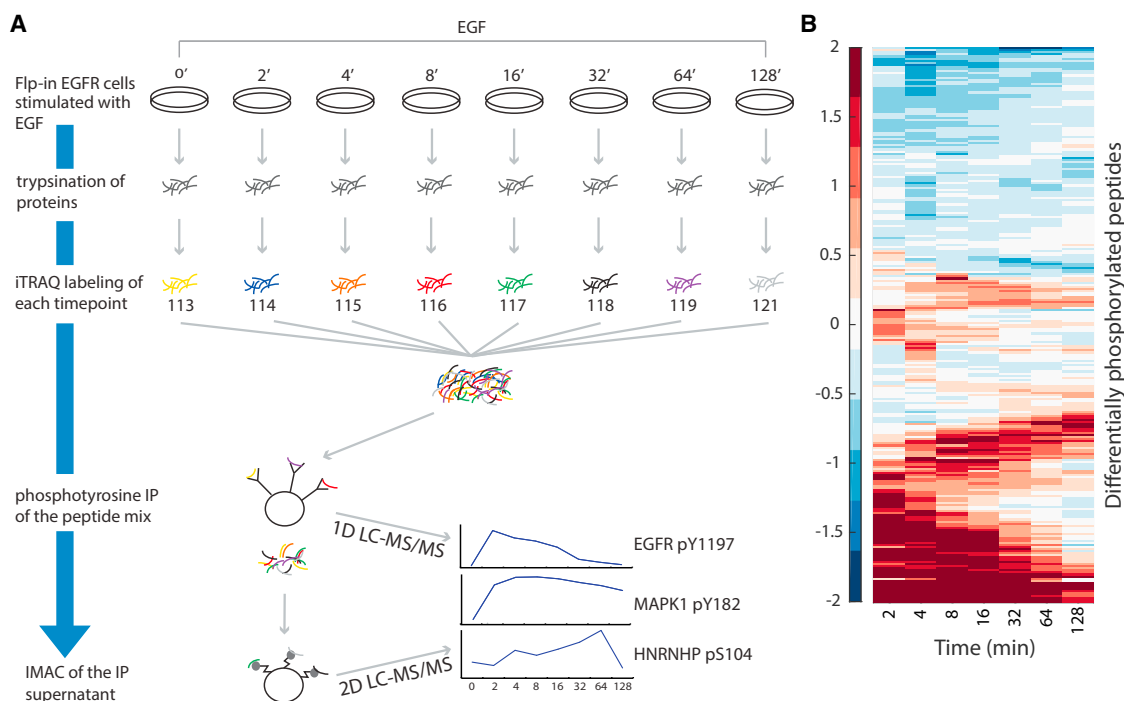
TPS first transforms a PPI graph into a condition-specific network by using mass spectrometry data to filter out irrelevant interactions. Next, TPS finds the orientation and sign of edges in the condition-specific interaction graph based on the order of the phosphorylation events. Phosphorylation timing is modeled separately for each observed phosphorylation site on a protein. TPS systematically explores all signed, directed graphs that may explain how signaling messages propagate from the stimulated source protein. Finally, TPS summarizes the valid graphs into a single aggregate network that explicitly tracks confident and ambiguous predictions. Our temporal pathway visualizer tool interactively visualizes the

Protein-protein interaction (PPI) networks serve this purpose by identifying interactions that connect observed phosphorylation events.

We present the Temporal Pathway Synthesizer (TPS) (Figure 1), a method to assemble temporal phosphoproteomic

summary network alongside the temporal phosphoproteomic data (Köksal et al., 2018).

We study the dynamic signaling responses to human EGF stimulation and yeast osmotic stress. TPS recovers networks that explain how stimulus-responsive proteins are activated or



**Figure 2. Overview of the EGF Response Proteomics Analysis**

(A) Cells are stimulated with EGF for 0, 2, 4, 8, 16, 32, 64, or 128 min and then lysed. Cellular protein content is denatured and digested. Peptides are labeled with iTRAQ and mixed. Tyrosine phosphorylated peptides are enriched by immunoprecipitation, and the flowthrough is passed over immobilized metal affinity chromatography to enrich for phosphorylation events on serine and threonine. The phosphotyrosine-rich fraction is analyzed by 1D-LC-MS/MS. The more complex phospho-serine/threonine-rich fraction is analyzed by 2D-LC-MS/MS. Resulting spectra are identified and quantified using Comet.

(B) The 263 peptides with significant temporal changes in phosphorylation exhibit distinct types of temporal behaviors ( $\log_2$  fold change with respect to pre-stimulation intensity). One group of peptides is activated immediately upon stimulation, whereas others display delayed waves of phosphorylation as signals propagate.

See also [Figures S1 and S2](#) and [Data S1 and S2](#).

inhibited via chains of physical interactions stemming from the upstream receptors. The highest-confidence TPS predictions are well supported by prior knowledge and consistent with kinase perturbations. These insights into well characterized human and yeast pathways exemplify how TPS can produce condition-specific pathway maps.

## RESULTS

### Quantitative Time Series Phosphoproteomics of EGF Response Captures Widespread Signaling Activity

To quantify global EGFR-mediated cellular signaling changes in HEK293 EGFR Flp-In (EGFR Flp-In) cells ([Gordus et al., 2009](#)) with phosphoproteomics, we used in-line two-dimensional high-performance liquid chromatography separation (2D-HPLC) coupled to tandem mass spectrometry (MS/MS) ([Ficarro et al., 2011](#); [Wolf-Yadlin et al., 2006](#)). We stimulated the cells with EGF for 0, 2, 4, 8, 16, 32, 64, or 128 min and collected three biological replicates with two technical replicates each ([Figure 2](#)). We identified 1,068 phosphorylation sites that were detected in all biological replicates (5,442 unique sites detected in at least one replicate), which were then used for TPS network modeling ([Data S1 and S2](#)). Phosphorylation intensities were

well correlated across the three biological replicates ([Köksal et al., 2018](#)).

### Reference Pathway Databases Fail to Explain Phosphorylation Changes

We assessed how much of the observed phosphorylation could be explained by existing pathway databases. To obtain a comprehensive view of EGFR-mediated signaling, we collected eight EGFR-related reference pathways ([Croft et al., 2014](#); [Gough, 2002](#); [Kandasamy et al., 2010](#); [Kanehisa et al., 2012](#); [Layek et al., 2011](#); [Nishimura, 2001](#); [Schaefer et al., 2009](#); [Supplemental Experimental Procedures](#)). Despite the diversity of the pathway diagrams, they all fail to capture the vast majority of significant phosphorylation events triggered by EGF stimulation in our system ([Figures S1 and S2](#)). Among the 203 significantly differentially phosphorylated proteins, typically 5% or fewer are present in a reference pathway. 85% of phosphorylated proteins are missing from all of the EGFR-related pathway maps ([Figure S1B](#)). Additionally, most of the proteins in the EGFR pathway maps are not differentially phosphorylated ([Figure S1A](#)), reflecting a combination of relevant proteins that do not undergo this particular type of PTM, phosphorylation events missed by the mass spectrometry, and interactions that are relevant in

some contexts, but not in EGFR Flip-In cells. The low overlaps agree with phosphoproteomic studies of other mammalian signaling pathways. Less than 10% of insulin-regulated proteins were members of a curated insulin pathway (Humphrey et al., 2015). In a study of T cell receptor signaling, only 21% of phosphorylated proteins were known to be involved in the pathway (Cao et al., 2012). Phosphosites regulated by transforming growth factor  $\beta$  (TGF- $\beta$ ) stimulation were not enriched for the TGF- $\beta$  pathway (D'Souza et al., 2014).

Crosstalk does not explain the low coverage. Most phosphorylated proteins (63%) are not present in the EGFR pathways or any BioCarta, Reactome, or PID pathway (Figure S1B), demonstrating the need for a context-specific representation of EGFR signaling pathway.

### Reconstructing the EGFR Pathway with TPS Explains Temporal Phosphorylation Changes

We applied TPS to model the dynamic signaling response to EGFR stimulation in EGFR Flip-In HEK293 cells. Our workflow consists of three major steps: (1) preprocessing the protein-protein interaction network and temporal phosphorylation data; (2) transforming temporal information, subnetwork structure, and prior knowledge into logical constraints; and (3) summarizing all valid signaling pathway models to discover interactions with unambiguous directions and/or signs (Figure 1).

We first discretized the time series phosphoproteomic data, using Tukey's honest significant difference (HSD) test (Yandell, 1997) to determine whether a peptide exhibits a significant increase, significant decrease, or no change in phosphorylation at each post-stimulation time point. 263 peptides, corresponding to 203 proteins, significantly change at one or more time points (Köksal et al., 2018). Second, we used the prize-collecting Steiner forest (PCSF) (Tuncbag et al., 2013) network algorithm to link the phosphorylated proteins to EGF, the source of stimulation, weighting proteins based on their HSD test significance. PCSF identifies a PPI subnetwork of 316 nodes and 422 edges (Data S3). This subnetwork comprises the interactions through which signaling messages are most likely to propagate. Third, TPS combined the discretized temporal activities of the 263 significantly changing peptides, the PCSF network, and prior knowledge (the orientation of kinase-substrate interactions) to generate a summary of all feasible pathway models (Data S3). Each type of input was translated into logical constraints, which were used to rule out pathway models that are not supported by the data.

In contrast to the reference EGFR pathway diagrams, which capture at most 11% of the differentially phosphorylated proteins, the predicted network from TPS (Figures 3 and S3; Data S3) contains 83% of the responding proteins in its 311 nodes. Each of these proteins is linked to the EGF stimulation with high-confidence protein interactions and has timing that is consistent with the temporal phosphorylation changes of all other proteins in the pathway. These interactions are depicted as directed, signed edges in a graph, where the sign reflects that the proteins have the same (activation) or opposite (inhibition) activity changes. Of the 413 edges in the network, 202 (49%) have a consistent direction in *all of the valid pathway models*, a strong assertion about the confidence in these edge directions.

Thirty-eight of the directed edges have a consistent sign as well. The PPI connections, phosphorylation timing, and prior knowledge of kinase-substrate interaction direction all play distinct, important roles in reducing the number of valid pathway models (Köksal et al., 2018). The timing of protein activation and inactivation in the TPS pathway reveals a rapid spread of signaling post-stimulation (Köksal et al., 2018).

### Prior Evidence Supports EGFR Pathway Predictions

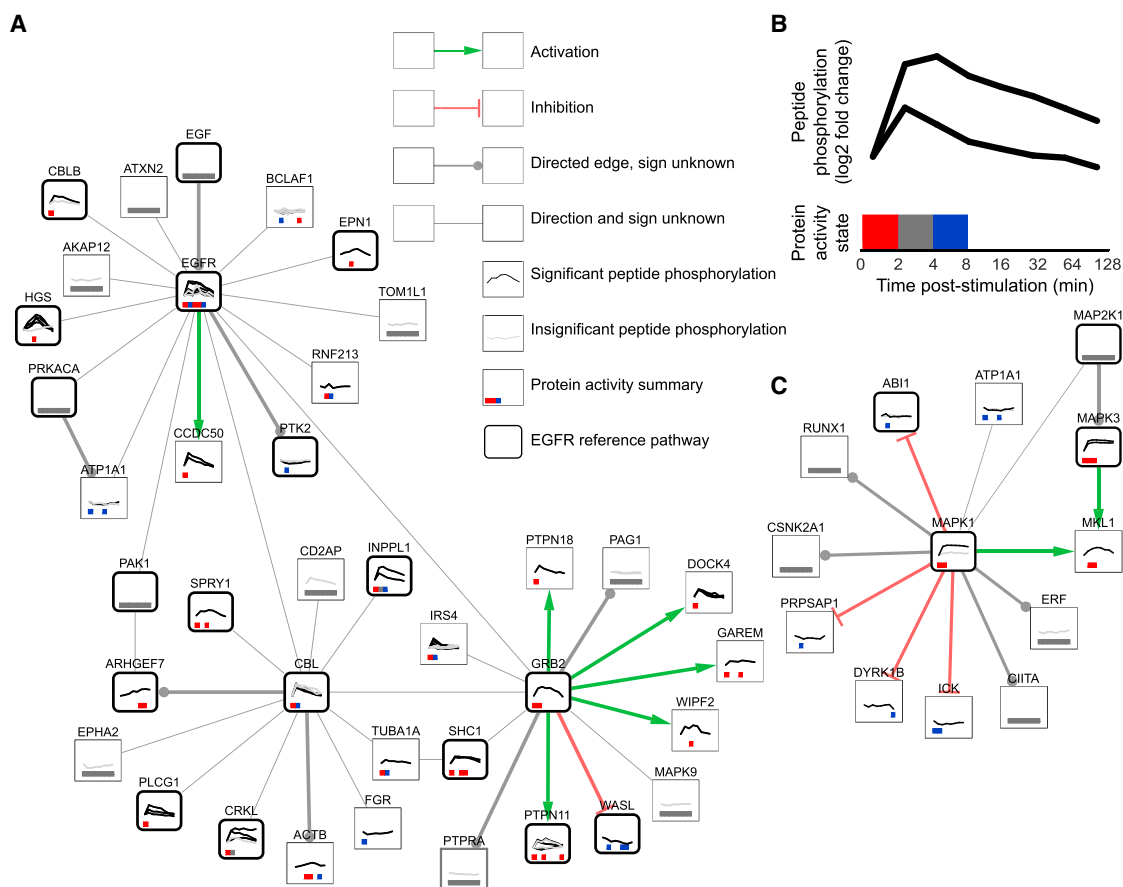
Although nearly all differentially phosphorylated proteins lie outside traditional EGFR pathway representations, 29 (11%) of the 273 phosphorylated proteins and 5 (13%) of the 38 unphosphorylated connective proteins in the TPS network are recognized as EGFR pathway members (Köksal et al., 2018). We find strong evidence for many of the predicted directions as well (Köksal et al., 2018). In total, 82 of 202 interaction directions are supported by our semi-automated evaluations using EGFR reference pathways, the PhosphoSitePlus input data (Hornbeck et al., 2015), and natural language processing software (Chen and Sharp, 2004; Hoffmann and Valencia, 2004; Poon et al., 2014; Data S3 and S4; Supplemental Experimental Procedures). The vast majority of the remaining directions can neither be confirmed nor refuted (Data S3). Our additional analyses (Köksal et al., 2018; Data S3) show that TPS also recovers high-quality pathway models when applied to existing EGF response datasets with lower temporal resolution (Olsen et al., 2006).

### TPS Network Models Can Guide Follow-Up Experiments

The TPS network can be used to prioritize proteins and interactions for additional experimental testing. To illustrate this process, we focused on edges for which the direction or sign were predicted confidently and one of the two proteins is a member of an EGFR reference pathway (Köksal et al., 2018). For each interaction, we inhibited the predicted upstream protein and measured the effect on the predicted target's phosphorylation using western blotting. From our list of ten candidate interactions (Table S1), we selected the three edges for which the antibodies reliably produced clean and quantifiable bands at the right molecular weight: MAPK1-ATP1A1; ABL2  $\rightarrow$  CRK; and AKT1  $\rightarrow$  ZYX (zyxin) (Figures 3C and S4). These proteins are already known to physically interact. The novelty of the TPS predictions is the interactions' relevance to the EGF response. The inhibitors used to inhibit the upstream proteins were SCH772984 for MAPK1, dasatinib for ABL2, and MK-2206 for AKT1. After serum starvation, the cells were treated with an inhibitor for one hour and then stimulated with EGF. We collected data at two time points (denoted short and long; see Figure S4) based on the timing of the phosphorylation events in our mass spectrometry data. Lysates were then assayed by western blot to quantify the level of phosphorylation of the downstream protein.

Dasatinib decreased phosphorylation of CRK (isoform Crk-II) pY221, consistent with the TPS pathway edge (Figure S4). Inhibiting AKT1 increased phosphorylation of Zyxin. In both cases, the predicted interaction direction is supported. MAPK1 inhibition increased ATP1A1 pY10 phosphorylation. The TPS model predicted an inhibitory interaction between these proteins, but the direction was ambiguous. Our data agree with the predicted edge sign and suggest that MAPK1 is upstream of ATP1A1





**Figure 3. TPS EGF Response Pathway Model**

Zoomed regions of the full TPS pathway model visualized with Cytoscape (Shannon et al., 2003).

(A) The EGFR subnetwork (EGFR, GRB2, CBL, and all their direct neighbors) depicts the proteins that first react to EGF stimulation. A substantial portion (18 of 38 proteins) is known to be associated with EGFR signaling. Green and red edges depict activation and inhibition, respectively. Gray edges that terminate in a circle indicate that the interaction is used in the same direction in all possible pathway models, but the sign is ambiguous. Thin, undirected edges are used in different directions in different valid pathway models. Thick, rounded borders show which proteins are present in one or more reference EGFR pathways. Node annotations are detailed in (B).

(B) Line graphs on each protein node show the temporal peptide phosphorylation changes relative to the pre-stimulation level on a log<sub>2</sub> scale. Multiple lines indicate multiple observed phosphopeptides for that protein, where black lines denote statistically significant phosphorylation changes and gray lines indicate insignificant changes. Proteins without line graphs are connective Steiner nodes inferred by PCSF. Colored boxes summarize the TPS inferred activity state across peptides at each time point. Red indicates activation, blue inhibition, gray ambiguity, and white inactivity.

(C) The subnetwork surrounding MAPK1 and MAPK3. TPS correctly determines that MAP2K1 is the kinase that controls both MAPK1 and MAPK3, even though it is not observed in the mass spectrometry data.

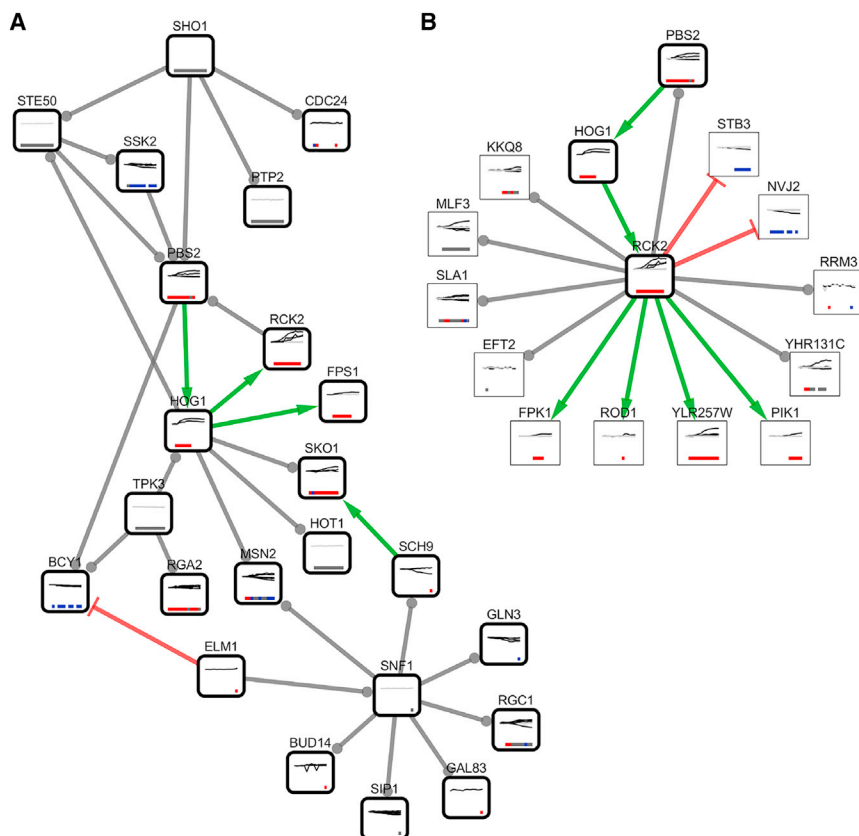
See also Figures S3 and S4, Table S1, and Data S3 and S4.

(Köksal et al., 2018). Truly validating the predicted edges would require more direct manipulation of the relevant kinases because Dasatinib is a multi-target inhibitor (Lindauer and Hochhaus, 2014); SCH772984 inhibits both MAPK1 and MAPK3 (Morris et al., 2013); and MK-2206 inhibits AKT1, AKT2, and AKT3 (Yan, 2009). However, these inhibitor experiments demonstrate how TPS can generate testable predictions from global phosphoproteomic data.

### TPS Makes Network Predictions Not Captured by Alternative Approaches

We compared TPS to two existing methods that combine PPI networks and time series data and a third that uses only the

phosphorylation data (Supplemental Experimental Procedures). The dynamic Bayesian network (DBN) (Hill et al., 2012) infers posterior peptide-peptide interaction probabilities from time series data and network priors. TimeXNet (Patil et al., 2013) formulates pathway prediction as a network flow problem. FunChisq (Zhang and Song, 2013) uses an adapted chi-square test to detect directed relationships between phosphorylated proteins. Comparing the four predicted EGF response pathway models demonstrates the impact of the diverse algorithmic strategies. Almost all of the protein-protein edges are unique to a single method, and no edges are predicted by all four methods (Köksal et al., 2018). Despite greater overlap among the predicted nodes, the four pathways are divergent.



**Figure 4. TPS Osmotic Stress Response Pathway Model**

(A) The portion of the TPS yeast osmotic stress response pathway model for which both proteins are in the osmotic stress reference pathway. TPS correctly recovers the core pathway structure from the Sho1 osmosensor to the primary kinases and transcription factors by ordering proteins based on the phosphorylation timing. Twelve of these pathway interactions are supported by the KEGG high-osmolarity pathway or other literature (Data S4). Node and edge visualizations are as in Figure 3. Note that three interactions (Ste50 → Pbs2, Ste50 → Ssk2, and Rck2 → Pbs2), derived from references (Chasman et al., 2014; Sharifpoor et al., 2011), are not found in other curated versions of the yeast interaction network.

(B) A zoomed view of the TPS pathway depicting Rck2 and the proteins it is predicted to interact with. All four proteins predicted to be activated by Rck2—Fpk1, Pik1, Rod1, and YLR257W—displayed decreased phosphorylation in the RCK2 mutant strain (Romanov et al., 2017), as did predicted targets Mlf3, Sla1, and YHR131C. See also Figure S5 and Data S3 and S4.

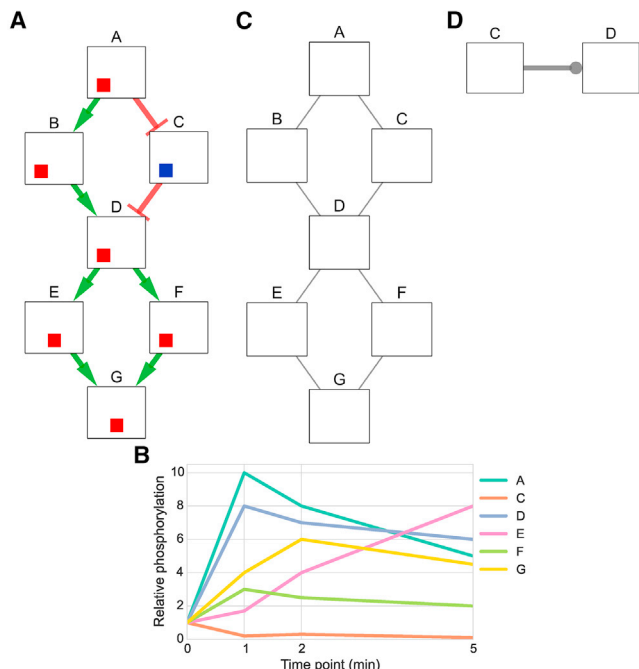
Because most of the differentially phosphorylated proteins are not members of any reference pathway, these pathways cannot be used to assess the overall quality of the predictions. The TimeXNet pathway, the largest of the three predicted networks, generally captures the most reference pathway interactions when ignoring edge direction and sign (Data S4). However, a closer examination that accounts for the predicted interaction direction shows that TPS typically makes the fewest errors, even when controlling for the size of the predicted pathways (Data S4).

### Yeast Osmotic Stress Response Model Recapitulates Known Pathway Structure and Nominates Candidate Rck2 and Cdc28 Substrates

Although they are still not fully characterized, stress-response signaling cascades in the yeast *Saccharomyces cerevisiae* are better understood than their human counterparts and are not subject to cell-type-specific effects. Thus, we applied TPS to model the yeast osmotic stress response to assess its ability to recapitulate this frequently studied pathway and reveal additional interactions. The hyperosmotic stress response is primarily controlled by the high osmolarity glycerol (HOG) pathway. Kanshin et al. (2015) profiled the rapid response to NaCl, an osmotic stressor, measuring phosphorylation changes for 60 s post-stimulation at uniform 5-s intervals. They identified 1,596 phosphorylated proteins, including 1,401 dynamic phosphopeptides on 784 proteins based on their fold changes in the salt stress time series with respect to a control (Table S2). We used

these data to construct a TPS pathway model of the early osmotic stress response (Data S3). The TPS osmotic stress pathway contains 216 proteins and 287 interactions (Figure S5). Thirty-six of these proteins (17%) have been previously annotated as osmotic stress pathway proteins (Kawakami et al., 2016). Focusing on the subset of interactions that connect known HOG pathway members reveals that many of the edges connecting them are correct as well (Figure 4A). TPS recovers the core part of the Kyoto Encyclopedia of Genes and Genomes (KEGG) high-osmolarity pathway, including the interactions Sho1 → Ste50, Sho1 → Cdc24, Sho1 → Pbs2, Ssk2 → Pbs2, and Pbs2 → Hog1 (Data S4). In addition, it correctly places Hog1 as the direct regulator of Rck2 (Romanov et al., 2017) and the transcription factors Hot1, Msn2, and Sko1 (Capaldi et al., 2008). TPS identifies Sch9 as an additional regulator of Sko1 (Pascual-Ahuir and Proft, 2007). Following hyperosmotic shock, Hog1 is recruited to Fps1 (Lee et al., 2013), consistent with the TPS prediction. The predicted feedback from Hog1 to Ste50 is also well supported in osmotic stress (Hao et al., 2008). Many predicted interactions that deviate from the canonical HOG pathway model can be attributed to the input phosphorylation data and background network, not the TPS algorithm (Köksal et al., 2018).

After confirming the TPS osmotic stress model agrees well with existing models, we investigated novel candidate pathway members. The TPS model captured the cascade Hog1 → Rck2 → Eft2 (Romanov et al., 2017; Teige et al., 2001) and predicted additional Rck2 targets (Figure 4B). To test these predictions, we compared them to a recent phosphoproteomic study of an RCK2 mutant subjected to osmotic stress (Romanov et al., 2017). All four proteins that TPS predicts are activated



**Figure 5. Artificial Example Illustrating the Inputs to TPS**

(A) The hypothetical signaling pathway that responds to stimulation of node A. The colored boxes on each node show the time at which the protein is activated or inhibited and begins influencing its downstream neighbors, with the leftmost position indicating the earliest time point. Red boxes are increases in activity, blue boxes are decreases, and white boxes are inactive time points, as in Figure 3B. The left position indicates the activity at 0 to 1 min, the center position at 1 to 2 min, and the right position at 2 to 5 min.

(B) The first input to TPS is time series phosphorylation data of the response to stimulating node A.

(C) The second input is an undirected graph of high-confidence interactions that can recover hidden components that do not appear in the temporal data, such as node B.

(D) The last input, which is optional, is prior knowledge of the pathway interactions expressed as (unsigned) directed edges. We represent unsigned edges with a circular arrowhead.

by Rck2 have defective phosphorylation on at least one phosphosite in *rck2Δ* five minutes after osmotic insult (Romanov et al., 2017). Thus, Rck2 likely directly phosphorylates Fpk1, Pik1, Rod1, and YLR257W upon osmotic stress, as TPS predicts. In addition to the four activated substrates, TPS predicts that Rck2 directly regulates seven additional proteins with an ambiguous sign. Three of these seven predicted targets—Mif3, Sla1, and YHR131C—have a phosphosite that is dependent on Rck2 during osmotic stress (Romanov et al., 2017), supporting the TPS predictions. The three protein-protein edge signs are ambiguous because some phosphosites on the proteins exhibit a significant increase in phosphorylation and others decrease. Similarly, we verified that 67 out of 91 (74%) predicted Cdc28 targets have at least one phosphosite with defective phosphorylation following Cdc28 inhibition (Holt et al., 2009; Kanshin et al., 2017; Köksal et al., 2018).

The high-quality TPS osmotic stress pathway demonstrates the algorithm is broadly useful beyond our own EGF stimulation study. It not only recovers many major elements of the classic

HOG pathway representation but also prioritizes condition-specific kinase targets that are supported by independent perturbations.

## DISCUSSION

The pathway structure illuminated by the phosphorylated proteins in our EGFR Flp-In cells differs considerably from the simple representations in pathway databases. Interpreting signaling data requires reconstructing models specific to the cells, stimuli, and environment being studied. TPS combines condition-specific information—time series phosphoproteomic data and the source of stimulation—with generic PPI networks and optional prior knowledge (Figure 5) to produce custom pathway representations. The predicted EGFR signaling network highlights alternative connections to classic EGFR pathway kinases and extends the pathway with interactions that are supported by prior knowledge in other contexts or kinase inhibition. Combining different constraints on pathway structure from PPI network topology and temporal information is computationally challenging, and we identify predictions that can be obtained only through joint reasoning with all available data (Figure 6).

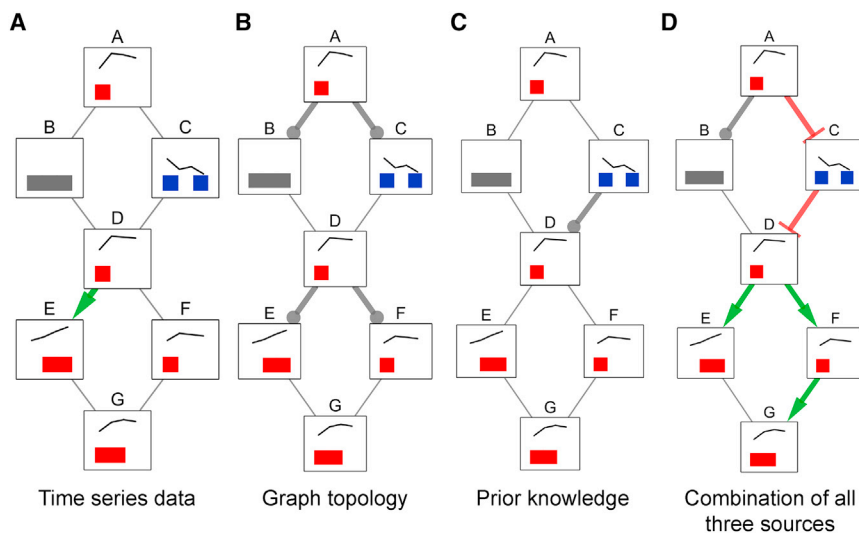
### Contrasting TPS with Related Computational Approaches

TPS integrates information from PPI networks, phosphosite-specific time series phosphoproteomic data, and prior knowledge by introducing a powerful constraint-based approach. Existing classes of signaling pathway inference algorithms do not offer the same functionality as TPS. Methods that identify dependencies in phosphorylation levels (Hill et al., 2012; Zhang and Song, 2013) omit pathway members without observed phosphorylation changes. TPS does not require perturbations to reconstruct pathways (Ciaccio et al., 2015; Molinelli et al., 2013; Terfve et al., 2015). Participants in the HPN-DREAM network inference challenge (Hill et al., 2016) inferred signaling networks from time series data for tens of phosphoproteins, but the top methods either did not scale to our dataset (PropheticGranger; Carlin et al., 2017) or did not perform well (FunChisq; Zhang and Song, 2013). Other algorithms that integrate temporal information with PPI networks (Budak et al., 2015; Gitter and Bar-Joseph, 2013; Jain et al., 2016; Norman and Cicek, 2018; Patil et al., 2013) do not evaluate and summarize all pathway models that are supported by the network and phosphorylation timing constraints. This summarization strategy is what enables TPS to scale to solution spaces (Figure S6) that are substantially larger than those typically considered by declarative computational approaches (Chasman et al., 2014; Dunn et al., 2014; Guziolowski et al., 2013; Köksal et al., 2013; Moignard et al., 2015; Sharan and Karp, 2013). The Supplemental Experimental Procedures contain additional related software beyond these representative examples.

### Future Directions in Pathway Synthesis

TPS offers a powerful framework for combining multiple types of declarative constraints to generate condition-specific signaling pathways. The constraint-based approach could be extended to include additional types of data, such as perturbation data





**Figure 6. TPS Models for Individual versus Combined Data Sources**

Summary graphs obtained by aggregating (via graph union) all possible signed, directed tree models for different constraints obtained from time series data (A), graph topology (B), prior knowledge (in this example, kinase-substrate interaction directions) (C), and all three types of input at the same time (D). If an edge has a unique sign and direction in a summary graph (colored green and red for activations and inhibitions, respectively), this means there are no valid models that assign a different orientation or sign to that edge. Edges that can have any combination of sign and direction in different models are gray without an arrowhead. See also Figure S7.

that link kinase inhibition or deletion to phosphorylation changes. Both temporal (Kanshin et al., 2015) and kinase perturbation (MacGilvray et al., 2018; Romanov et al., 2017) phosphoproteomic data are available for the yeast osmotic stress response. Modeling multiple related conditions (e.g., different ligand stimuli and inhibitor perturbations) could allow TPS to learn not only the signs of interactions but also the logic employed when multiple incoming signals influence a protein. TPS could also accommodate user-defined assumptions or heuristics about pathway properties, such as restrictions on pathway length. Such complex constraints cannot be readily included in approaches like DBN or TimeXNet.

For scalability, TPS requires hard logical constraints instead of probabilistic constraints (Hinton et al., 2006; Katoen et al., 2005). Discrete logic models for noisy biological data require modeling assumptions in order to balance model ambiguity and expressiveness. These tradeoffs and assumptions provide additional opportunities to modify and generalize the TPS model, for instance, a potential TPS extension to infer feedback in networks that is described in the Supplemental Experimental Procedures.

As proteomic technologies continue to improve in terms of depth of coverage (Sharma et al., 2014) and temporal resolution (Humphrey et al., 2015; Kanshin et al., 2015; Reddy et al., 2016), the need to systematically interpret these data will likewise grow. TPS enables reasoning with temporal phosphorylation changes and physical protein interactions to define what drives the vast protein modifications that are not represented by existing knowledge in pathway databases.

## EXPERIMENTAL PROCEDURES

### Temporal Pathway Synthesizer Algorithm Overview

TPS receives three types of input (Figure 1): a time series mass spectrometry phosphoproteomic analysis of a stimulus response; an undirected PPI subnetwork; and optional prior knowledge about interaction directions.

The undirected graph is obtained through a static analysis in which the significantly changing proteins are overlaid on a PPI network. A network algorithm recovers connections among the affected proteins, removing inter-

actions that do not form critical connections between these proteins and nominating hidden proteins that do, even if they are not themselves phosphorylated. We recommend PCSF (Tuncbag et al., 2013) to select the PPI subnetwork but also successfully applied other methods (Gitter et al., 2011; Patil et al., 2013; Yeger-Lotem et al., 2009).

TPS transforms the input data into logical constraints that determine which pathway models can explain the observed phosphoproteomic data. Topological constraints stem from the filtered PPI network and require that phosphorylated proteins are connected to the source of stimulation, such as EGF, by a cascade of signaling events. These signaling events propagate along the edges of the filtered PPI network. Temporal constraints ensure that the order of the signaling events is consistent with the timing of the phosphorylation changes. If protein B is downstream of protein A on the pathway, B cannot be activated or inhibited before A. Prior knowledge constraints guarantee that if the direction or sign of an interaction is known in advance, the pathway may not contain the edge with the opposite direction or sign. Typically, many possible pathways meet all constraints, so TPS summarizes the entire collection of valid pathways and identifies interactions that are used with the same direction or sign across all models. A symbolic solver reasons with these logical constraints and produces the pathway summary without explicitly enumerating all possible pathway models.

To illustrate this process, consider a hypothetical signaling pathway that contains a receptor node A and six other downstream proteins that respond when A is stimulated (Figure 5). The first input is time series mass spectrometry data measuring the response to stimulating the receptor (node A), which quantifies phosphorylation activity for six proteins. Node B is absent from the phosphorylation data because it is post-translationally modified, but not phosphorylated, by A. The second input is an undirected protein-protein interaction graph. These are detected independently of the stimulation condition but filtered based on their presumed relevance to the responding proteins with an algorithm such as PCSF. By combining phosphorylation data with the PPI subnetwork, this topology can recover “hidden” components of the pathway that are not phosphorylated (node B). Finally, TPS accepts prior knowledge of directed kinase-substrate or phosphatase-substrate interactions, such as the edge  $C \rightarrow D$ . Each of these inputs can be used individually to restrict the space of plausible pathway models. Reasoning about them jointly produces more unambiguous predictions than considering each resource separately.

To formulate temporal constraints, we transform the time series data into a set of discrete signaling events (activation or inhibition) for each node, taking an event-based view of the signaling process (Table 1). We determine time points for each node that correspond to statistically significant phosphorylation changes. These discrete events are then used to rule out network models that contain signed, directed paths that violate the temporal ordering of these events no matter which event is chosen for each node. For example, there can

**Table 1. Signaling Timing in the Artificial Example**

Node	Plausible Temporal Signaling Events
A	activated 0 to 1 min
B	activated or inhibited at any time
C	inhibited 0 to 1 min or 2 to 5 min
D	activated 0 to 1 min
E	activated 1 to 2 min or 2 to 5 min
F	activated 0 to 1 min
G	activated 0 to 1 min or 1 to 2 min

Plausible signaling events inferred for each node through a statistical analysis of the time series phosphorylation data. Although B is modified in the 0 to 1 min interval, this is not observed in the phosphoproteomic input data.

be no edge from E to D in any model because D is activated strictly earlier than E regardless of whether E is activated at 1 to 2 min or 2 to 5 min. Because the time series data measure the response to a specific stimulus, we also devise topological constraints that ensure all signaling activity originates from this source. In our example, this asserts that all edges in a solution network must be on a directed path that starts at node A. Finally, our third input, the set of directed interactions, requires that no model violates this prior knowledge by including an edge from D to C.

Figure 6 shows the pathway models that can be learned using each type of constraint alone and in combination. When we enforce only temporal constraints, which corresponds to reasoning locally with phosphorylation data for pairs of nodes to see whether one signaling event strictly precedes another, we obtain a single precise (signed and directed) prediction from D to E (Figure 6A). The topological constraints by themselves are sufficient to orient edges from the source A and from node D because D forms a bottleneck (Figure 6B). The prior knowledge constrains the direction of the edge from C to D, but its sign remains unknown (Figure 6C). Jointly enforcing all of these constraints has a nontrivial impact on the solution space (Figure 6D). For instance, we can infer that F must activate G. If the edge direction was reversed, F would be downstream of E, but the data show that activation of F precedes activation of E. The final model that includes all available data closely resembles the true pathway structure (Figure 5A). The edges incident to node B are ambiguous, and the interaction between E and G cannot be uniquely oriented, but all other interactions are recovered.

The summary for the combination of all constraints produces precise predictions that cannot be obtained by intersecting the summaries for the individual types of constraints. For instance, TPS infers that the relationship between F and G must be an activation from F to G because the sole way G can reach F in a tree rooted at A is through E, but F's activation precedes E's. This inference cannot be made by combining the models in panels A, B, and C. The simple example also highlights the differences in how the TPS constraint-based approach improves upon related methods based on correlation or the time point of maximum phosphorylation change (Köksal et al., 2018). See also Figure S7.

### TPS Pathway Synthesis

TPS takes the undirected network from PCSF and transforms it into a collection of signed, directed graphs that explain dynamic signaling events.

#### Discretization of Time Series Data

To find pathway models that agree with the phosphorylation dynamics, TPS first performs a discretization step that determines time intervals in which each protein may be differentially phosphorylated. The discrete set of activation and inhibition state changes is then used to rule out networks that violate the observed temporal behavior.

The transformation consists of finding time points for each profile where phosphorylation significantly differs from either the baseline (pre-stimulation) or the previous time point. In the baseline comparison, this time point is accepted only if it is not preceded by an earlier, larger change with respect to the baseline. If there is a hypothetical phosphorylation level at which the pro-

tein is activated and acts upon its downstream targets, a signaling event occurs only at the first time this threshold value is reached. This criterion does not apply when comparing to the phosphorylation level at the previous time point. TPS supports missing values in the time series data. The time points for which a phosphopeptide is missing data are assumed to be insignificant in the discretized data.

In our EGF study, we use Tukey's HSD test to find significant differential phosphorylation. If comparing a time point to the baseline or the previous measurement produces a p value below a user-defined threshold, the time point is marked as a possible activation or inhibition event depending on whether the phosphorylation level increased or decreased relative to the earlier time point to which it was compared.

#### Modeling Assumptions

We assume at most one signaling event happens for every node across time points. Our logical solver can explore all possible activation and inhibition events for every node, but the data are often too ambiguous to allow multiple events per node given a single type of stimulation. In the absence of perturbation experiments that test the pathway behavior under different initial conditions, it is impossible to distinguish between different Boolean logic functions governing the behavior of each node and whether a node responds to one or multiple regulators. We therefore formalize pathway models as signed, directed trees, which provide a sufficient basis for explaining the dynamic system behavior under these assumptions.

#### Translating Input into Constraints

TPS transforms each input into a set of constraints that declaratively specify valid signed, directed tree models that agree with the data (Supplemental Experimental Procedures). These constraints are expressed as Boolean formulas with linear integer arithmetic, ranging over symbolic variables that represent choices on edge signs and orientations as well as how the temporal data are interpreted. The constraints can then be solved by a satisfiability modulo theories (SMT) solver to find a network model that satisfies all constraints along with dynamic timing annotations for each interaction in the network.

Using constraints, we restrict the possible orientation and sign assignments to signed, directed tree networks rooted at the source node (e.g., EGF). Furthermore, constraints express how every tree model must agree with the time series data by establishing a correspondence between the order of nodes on tree paths and their temporal order of activity according to the time series data. Finally, we declaratively rule out models that contradict the prior knowledge of kinase-substrate interaction directions. These constraints define a very large space of candidate networks that agree with the data.

#### Pathway Summaries

TPS can reason with large state spaces by summarizing all valid pathways instead of explicitly enumerating them. A summary network is the graph union of all signed, directed tree networks that satisfy the stated constraints (Figure 6). Timing annotations are summarized by computing the set of possible annotations for each node over all solutions. In the graph union, some edges have a unique direction and sign combination, which signifies that this was the only observed signed, directed edge between two given nodes across the solution space. However, this does not guarantee that the edge between the interacting proteins must be present in all valid pathway models. Ambiguous directions or signs in the summary means that there are valid models with different direction or sign assignments.

We compute the summary graph by performing a linear number of SMT solver queries in terms of the size of the input graph. Each query whether at least one signed, directed model contains a specific signed, directed edge. Because individual queries are computationally cheap, we can summarize the entire solution space without enumerating all models, which is typically intractable. The summary graph over-approximates the solution space. It is not possible to recover the exact set of valid models from the summary, only a superset of the models (Figure S7). This tradeoff must be made in order to analyze such a large state space.

#### Using Solvers for Synthesis

TPS uses the Z3 theorem prover (De Moura and Bjørner, 2008) via the ScalaZ3 interface (Köksal et al., 2011) to solve the constraints it generates. It

also provides a custom data flow solver specifically for computing pathway summaries. The custom solver and the symbolic solver produce identical pathway summaries. However, the custom solver is much more scalable because it is specifically designed to address our synthesis task and can handle networks containing more than a hundred thousand edges and phosphosites (Figure S6; Köksal et al., 2018).

### Cell Culture and Mass Spectrometry

We stimulated EGFR Flp-In cells (Gordus et al., 2009) with 23.6 nM EGF (PeproTech) for 0, 2, 4, 8, 16, 32, 64, or 128 min. Cells were lysed and proteins were extracted, denatured, alkylated, and trypsin digested. Following digestion, the tryptic peptides were either lyophilized, stored for future use, or directly processed for mass spectrometry analysis. To quantify dynamic changes in protein phosphorylation, all peptides were isobarically labeled (Ross et al., 2004), enriched using phosphotyrosine-specific antibodies and/or immobilized metal affinity chromatography (IMAC) (Ficarro et al., 2002), and analyzed on a Thermo Fisher Velos Orbitrap mass spectrometer (Ficarro et al., 2011; Wolf-Yadlin et al., 2006) in data-dependent acquisition mode. We determined peptide sequences using Comet (Eng et al., 2013; Data S5) and quantified the iTRAQ signals with Libra (Deutsch et al., 2010). Across three biological replicates, we quantified 5,442 unique peptides in at least one replicate and 1,068 peptides in all replicates and used Tukey's honest significant difference for statistical testing (Data S1). See the Supplemental Experimental Procedures for details and data processing. Also see our p value sensitivity analysis (Köksal et al., 2018).

### Quantitative Western Blotting

We used 25 nM Dasatinib (no. S1021), 400 nM SCH72984 (no. S7101), and 800 nM MK-2206 (no. S1078; all Selleckchem) for kinase inhibition and antibodies pY221-CRK (no. 3491; Crk-II isoform), pY10-ATP1A1 (no. 3060), and pS142/143-Zyxin (no. 8467; all Cell Signaling Technology) for western blotting (Supplemental Experimental Procedures). We normalized loading with  $\beta$ -actin (no. 3700) and imaged blots with an Odyssey Infrared Imaging System (Li-COR Biosciences).

### PCSF

We used the Omics Integrator PCSF implementation (Tuncbag et al., 2016) with msgsteiner (Bailly-Bechet et al., 2011) to recover the most relevant PPIs connecting the phosphorylated proteins. The Supplemental Experimental Procedures describe how we selected parameters, ran PCSF multiple times to identify parallel connections between proteins, generated prizes from the phosphoproteomic data, and created a weighted interaction network from iRefIndex (Razick et al., 2008) and PhosphoSitePlus (Hornbeck et al., 2015).

### DATA AND SOFTWARE AVAILABILITY

The accession number for the raw mass spectrometry proteomics data reported in this paper is PRIDE: PXD006697.

The processed data are in Data S1. TPS (<https://github.com/koksal/tps>) and our visualization tool for TPS output (<https://github.com/koksal/tpv>) are available as MIT-licensed open source software. An archival copy of TPS version 2.2, including instructions for running the software, example data, and scripts for linking PCSF and TPS, is available at <https://doi.org/10.5281/zenodo.1215177>.

### SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, two tables, and five data files and can be found with this article online at <https://doi.org/10.1016/j.celrep.2018.08.085>.

### ACKNOWLEDGMENTS

We thank Nir Piterman for algorithmic discussions, Anthony Soltis and Jennifer Wilson for assistance preparing the human protein-protein interaction network, Alex Hu for phosphoproteomic data processing, Evgeny Kanshin

for assistance with the yeast phosphoproteomic data, Audrey Gasch for yeast osmotic stress response discussions, Sandra Kaplan for copyediting, and Elsevier WebShop Illustration Services for the graphical abstract. Microsoft Research supported A.S.K. and A.G. during the initial stages of this work. A.G. acknowledges support from NSF grant DBI-1553206. D.R.C. acknowledges support from NIH grant U54-AI117924. A.M. acknowledges support from NIH training grant T32-HL007312. M.E.M. acknowledges support from NIH training grant T32-HG002760. E.F. acknowledges support from NIH grants U01-CA184898 and U54-NS091046. R.B. acknowledges support from NSF grants CCF-1139138, CCF-1337415, and ACI-1535191; DOE award FOA-0000619; DARPA grants FA8750-14-C-0011 and FA8750-16-2-0032; as well as gifts from Google, Intel, Mozilla, Nokia, and Qualcomm.

### AUTHOR CONTRIBUTIONS

Conceptualization, A.S.K., S.S., R.B., A.W.-Y., E.F., J.F., and A.G.; Methodology, A.S.K., K.B., R.B., A.W.-Y., E.F., J.F., and A.G.; Software, A.S.K. and A.G.; Validation, K.B. and N.D.C.; Formal Analysis, A.S.K., D.R.C., and A.G.; Investigation, A.S.K., K.B., D.R.C., A.M., N.D.C., M.E.M., and A.G.; Data Curation, A.S.K., K.B., D.R.C., A.W.-Y., and A.G.; Writing, A.S.K., K.B., M.E.M., R.B., A.W.-Y., E.F., J.F., and A.G.; Visualization, A.S.K., E.F., and A.G.; Supervision, R.B., A.W.-Y., E.F., J.F., and A.G.

### DECLARATION OF INTERESTS

R.B. consults with NVIDIA.

Received: July 18, 2016

Revised: April 16, 2018

Accepted: August 29, 2018

Published: September 25, 2018

### REFERENCES

- Bailly-Bechet, M., Borgs, C., Braunstein, A., Chayes, J., Dagkessamanskaia, A., François, J.-M., and Zecchina, R. (2011). Finding undetected protein associations in cell signaling by belief propagation. *Proc. Natl. Acad. Sci. USA* 108, 882–887.
- Bar-Joseph, Z., Gitter, A., and Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* 13, 552–564.
- Bauer-Mehren, A., Furlong, L.I., and Sanz, F. (2009). Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol. Syst. Biol.* 5, 290.
- Benner, S.A., and Sismour, A.M. (2005). Synthetic biology. *Nat. Rev. Genet.* 6, 533–543.
- Budak, G., Eren Ozsoy, O., Aydin Son, Y., Can, T., and Tuncbag, N. (2015). Reconstruction of the temporal signaling network in Salmonella-infected human cells. *Front. Microbiol.* 6, 730.
- Cao, L., Ding, Y., Hung, N., Yu, K., Ritz, A., Raphael, B.J., and Salomon, A.R. (2012). Quantitative phosphoproteomics reveals SLP-76 dependent regulation of PAG and Src family kinases in T cells. *PLoS ONE* 7, e46725.
- Capaldi, A.P., Kaplan, T., Liu, Y., Habib, N., Regev, A., Friedman, N., and O'Shea, E.K. (2008). Structure and function of a transcriptional network activated by the MAPK Hog1. *Nat. Genet.* 40, 1300–1306.
- Carlin, D.E., Paull, E.O., Graim, K., Wong, C.K., Bivol, A., Ryabinin, P., Ellrott, K., Sokolov, A., and Stuart, J.M. (2017). Prophetic Granger causality to infer gene regulatory networks. *PLoS ONE* 12, e0170340.
- Chasman, D., Ho, Y.-H., Berry, D.B., Nemec, C.M., MacGilvray, M.E., Hose, J., Merrill, A.E., Lee, M.V., Will, J.L., Coon, J.J., et al. (2014). Pathway connectivity and signaling coordination in the yeast stress-activated signaling network. *Mol. Syst. Biol.* 10, 759.
- Chen, H., and Sharp, B.M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* 5, 147.

- Choudhary, C., and Mann, M. (2010). Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.* *11*, 427–439.
- Ciaccio, M.F., Chen, V.C., Jones, R.B., and Bagheri, N. (2015). The DIONESUS algorithm provides scalable and accurate reconstruction of dynamic phosphoproteomic networks to reveal new drug targets. *Integr. Biol.* *7*, 776–791.
- Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res.* *42*, D472–D477.
- D'Souza, R.C.J., Knittle, A.M., Nagaraj, N., van Dinther, M., Choudhary, C., ten Dijke, P., Mann, M., and Sharma, K. (2014). Time-resolved dissection of early phosphoproteome and ensuing proteome changes in response to TGF- $\beta$ . *Sci. Signal.* *7*, rs5.
- De Moura, L., and Björner, N. (2008). Z3: an efficient SMT solver. In *Proceedings of the Theory and Practice of Software, 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems* (Springer-Verlag), pp. 337–340.
- Deutsch, E.W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., et al. (2010). A guided tour of the Trans-Proteomic Pipeline. *Proteomics* *10*, 1150–1159.
- Dunn, S.-J., Martello, G., Yordanov, B., Emmott, S., and Smith, A.G. (2014). Defining an essential transcription factor program for naïve pluripotency. *Science* *344*, 1156–1160.
- Eng, J.K., Jahan, T.A., and Hoopmann, M.R. (2013). Comet: an open-source MS/MS sequence database search tool. *Proteomics* *13*, 22–24.
- Ficarro, S.B., McClelland, M.L., Stukenberg, P.T., Burke, D.J., Ross, M.M., Shabanowitz, J., Hunt, D.F., and White, F.M. (2002). Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.* *20*, 301–305.
- Ficarro, S.B., Zhang, Y., Carrasco-Alfonso, M.J., Garg, B., Adelmant, G., Webber, J.T., Luckey, C.J., and Marto, J.A. (2011). Online nanoflow multidimensional fractionation for high efficiency phosphopeptide analysis. *Mol. Cell. Proteomics* *10*, O111.011064.
- Fisher, J., Piterman, N., and Bodik, R. (2014). Toward synthesizing executable models in biology. *Front. Bioeng. Biotechnol.* *2*, 75.
- Gitter, A., and Bar-Joseph, Z. (2013). Identifying proteins controlling key disease signaling pathways. *Bioinformatics* *29*, i227–i236.
- Gitter, A., Klein-Seetharaman, J., Gupta, A., and Bar-Joseph, Z. (2011). Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Res.* *39*, e22.
- Gordus, A., Krall, J.A., Beyer, E.M., Kaushansky, A., Wolf-Yadlin, A., Sevecka, M., Chang, B.H., Rush, J., and MacBeath, G. (2009). Linear combinations of docking affinities explain quantitative differences in RTK signaling. *Mol. Syst. Biol.* *5*, 235.
- Gough, N.R. (2002). Science's signal transduction knowledge environment: the connections maps database. *Ann. N Y Acad. Sci.* *971*, 585–587.
- Guziolowski, C., Videla, S., Eduati, F., Thiele, S., Cokelaer, T., Siegel, A., and Saez-Rodriguez, J. (2013). Exhaustively characterizing feasible logic models of a signaling network using Answer Set Programming. *Bioinformatics* *29*, 2320–2326.
- Hao, N., Zeng, Y., Elston, T.C., and Dohlman, H.G. (2008). Control of MAPK specificity by feedback phosphorylation of shared adaptor protein Ste50. *J. Biol. Chem.* *283*, 33798–33802.
- Hill, S.M., Lu, Y., Molina, J., Heiser, L.M., Spellman, P.T., Speed, T.P., Gray, J.W., Mills, G.B., and Mukherjee, S. (2012). Bayesian inference of signaling network topology in a cancer cell line. *Bioinformatics* *28*, 2804–2810.
- Hill, S.M., Heiser, L.M., Cokelaer, T., Unger, M., Nesser, N.K., Carlin, D.E., Zhang, Y., Sokolov, A., Paull, E.O., Wong, C.K., et al.; HPN-DREAM Consortium (2016). Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods* *13*, 310–318.
- Hill, S.M., Nesser, N.K., Johnson-Camacho, K., Jeffress, M., Johnson, A., Boniface, C., Spencer, S.E.F., Lu, Y., Heiser, L.M., Lawrence, Y., et al. (2017). Context specificity in causal signaling networks revealed by phosphoprotein profiling. *Cell Syst.* *4*, 73–83.e10.
- Hinton, A., Kwiatkowska, M., Norman, G., and Parker, D. (2006). PRISM: a tool for automatic verification of probabilistic systems. In *Tools and Algorithms for the Construction and Analysis of Systems*, H. Hermanns and J. Palsberg, eds. (Springer Berlin Heidelberg), pp. 441–444.
- Hoffmann, R., and Valencia, A. (2004). A gene network for navigating the literature. *Nat. Genet.* *36*, 664.
- Holt, L.J., Tuch, B.B., Villén, J., Johnson, A.D., Gygi, S.P., and Morgan, D.O. (2009). Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* *325*, 1682–1686.
- Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* *43*, D512–D520.
- Humphrey, S.J., Azimifar, S.B., and Mann, M. (2015). High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. *Nat. Biotechnol.* *33*, 990–995.
- Jain, S., Arrais, J., Venkatachari, N.J., Ayyavoo, V., and Bar-Joseph, Z. (2016). Reconstructing the temporal progression of HIV-1 immune response pathways. *Bioinformatics* *32*, i253–i261.
- Kandasamy, K., Mohan, S.S., Raju, R., Keerthikumar, S., Kumar, G.S., Venugopal, A.K., Telikicherla, D., Navarro, J.D., Mathivanan, S., Pecquet, C., et al. (2010). NetPath: a public resource of curated signal transduction pathways. *Genome Biol.* *11*, R3.
- Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* *40*, D109–D114.
- Kanshin, E., Bergeron-Sandoval, L.-P., Isik, S.S., Thibault, P., and Michnick, S.W. (2015). A cell-signaling network temporally resolves specific versus promiscuous phosphorylation. *Cell Rep.* *10*, 1202–1214.
- Kanshin, E., Giguère, S., Jing, C., Tyers, M., and Thibault, P. (2017). Machine learning of global phosphoproteomic profiles enables discrimination of direct versus indirect kinase substrates. *Mol. Cell. Proteomics* *16*, 786–798.
- Katoen, J.-P., Khattri, M., and Zapreev, I.S. (2005). A Markov reward model checker. In *Second International Conference on the Quantitative Evaluation of Systems (IEEE)*, pp. 243–244.
- Kawakami, E., Singh, V.K., Matsubara, K., Ishii, T., Matsuoka, Y., Hase, T., Kul-karni, P., Siddiqui, K., Kodilkar, J., Danve, N., et al. (2016). Network analyses based on comprehensive molecular interaction maps reveal robust control structures in yeast stress response pathways. *npj Syst. Biol. Appl.* *2*, 15018.
- Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. *Nature* *509*, 575–581.
- Köksal, A.S., Kuncak, V., and Suter, P. (2011). Scala to the power of Z3: integrating SMT and programming. In *Automated Deduction – CADE-23*, N. Björner and V. Sofronie-Stokkermans, eds. (Springer), pp. 400–406.
- Köksal, A.S., Pu, Y., Srivastava, S., Bodik, R., Fisher, J., and Piterman, N. (2013). Synthesis of biological models from mutation experiments. In *Proceedings of the 40th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (ACM)*, pp. 469–482.
- Köksal, A.S., Beck, K., Cronin, D.R., McKenna, A., Camp, N.D., Srivastava, S., MacGilvray, M.E., Bodik, R., Wolf-Yadlin, A., Fraenkel, E., et al. (2018). External Supplement to Synthesizing Signaling Pathways from Temporal Phosphoproteomic Data. <https://doi.org/10.6084/m9.figshare.6957461>.
- Layek, R., Datta, A., Bittner, M., and Dougherty, E.R. (2011). Cancer therapy design based on pathway logic. *Bioinformatics* *27*, 548–555.
- Lee, J., Reiter, W., Dohnal, I., Gregori, C., Beese-Sims, S., Kuchler, K., Ammerer, G., and Levin, D.E. (2013). MAPK Hog1 closes the *S. cerevisiae* glycerol channel Fps1 by phosphorylating and displacing its positive regulators. *Genes Dev.* *27*, 2590–2601.
- Lindauer, M., and Hochhaus, A. (2014). Dasatinib. In *Small Molecules in Oncology*, U. Martens, ed. (Springer), pp. 27–65.
- MacGilvray, M.E., Shishkova, E., Chasman, D., Place, M., Gitter, A., Coon, J.J., and Gasch, A.P. (2018). Network inference reveals novel connections in



- pathways regulating growth and defense in the yeast salt response. *PLoS Comput. Biol.* **13**, e1006088.
- Manna, Z., and Waldinger, R. (1980). A deductive approach to program synthesis. *ACM Trans. Program. Lang. Syst.* **2**, 90–121.
- Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A.J., Tanaka, Y., Wilkinson, A.C., Buettner, F., Macaulay, I.C., Jawaid, W., Diamanti, E., et al. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* **33**, 269–276.
- Molinelli, E.J., Korkut, A., Wang, W., Miller, M.L., Gauthier, N.P., Jing, X., Kaushik, P., He, Q., Mills, G., Solit, D.B., et al. (2013). Perturbation biology: inferring signaling networks in cellular systems. *PLoS Comput. Biol.* **9**, e1003290.
- Morris, E.J., Jha, S., Restaino, C.R., Dayananth, P., Zhu, H., Cooper, A., Carr, D., Deng, Y., Jin, W., Black, S., et al. (2013). Discovery of a novel ERK inhibitor with activity in models of acquired resistance to BRAF and MEK inhibitors. *Cancer Discov.* **3**, 742–750.
- Newman, R.H., Zhang, J., and Zhu, H. (2014). Toward a systems-level view of dynamic phosphorylation networks. *Front. Genet.* **5**, 263.
- Nishimura, D. (2001). *BioCarta. Biotech Softw. Internet Rep.* **2**, 117–120.
- Norman, U., and Cicek, A.E. (2018). Spatio-temporal gene discovery for autism spectrum disorder. *bioRxiv*. <https://doi.org/10.1101/256693>.
- Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006). Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635–648.
- Pascual-Ahuir, A., and Proft, M. (2007). The Sch9 kinase is a chromatin-associated transcriptional activator of osmoresponsive genes. *EMBO J.* **26**, 3098–3108.
- Patil, A., Kumagai, Y., Liang, K.C., Suzuki, Y., and Nakai, K. (2013). Linking transcriptional changes over time in stimulated dendritic cells to identify gene networks activated during the innate immune response. *PLoS Comput. Biol.* **9**, e1003323.
- Pawson, T., and Warner, N. (2007). Oncogenic re-wiring of cellular signaling pathways. *Oncogene* **26**, 1268–1275.
- Poon, H., Quirk, C., DeZiel, C., and Heckerman, D. (2014). Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics* **30**, 2840–2842.
- Razick, S., Magkias, G., and Donaldson, I.M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9**, 405.
- Reddy, R.J., Gajadhar, A.S., Swenson, E.J., Rothenberg, D.A., Curran, T.G., and White, F.M. (2016). Early signaling dynamics of the epidermal growth factor receptor. *Proc. Natl. Acad. Sci. USA* **113**, 3114–3119.
- Romanov, N., Hollenstein, D.M., Janschitz, M., Ammerer, G., Anrather, D., and Reiter, W. (2017). Identifying protein kinase-specific effectors of the osmotic stress response in yeast. *Sci. Signal.* **10**, eaag2435.
- Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., et al. (2004). Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3**, 1154–1169.
- Schaefer, C.F., Anthony, K., Krupa, S., Buchhoff, J., Day, M., Hannay, T., and Buetow, K.H. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**, D674–D679.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504.
- Sharan, R., and Karp, R.M. (2013). Reconstructing Boolean models of signaling. *J. Comput. Biol.* **20**, 249–257.
- Sharifpoor, S., Nguyen Ba, A.N., Youn, J.Y., van Dyk, D., Friesen, H., Douglas, A.C., Kurat, C.F., Chong, Y.T., Founk, K., Moses, A.M., and Andrews, B.J. (2011). A quantitative literature-curated gold standard for kinase-substrate pairs. *Genome Biol.* **12**, R39.
- Sharma, K., D'Souza, R.C.J., Tyanova, S., Schaab, C., Wiśniewski, J.R., Cox, J., and Mann, M. (2014). Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep.* **8**, 1583–1594.
- Teige, M., Scheikl, E., Reiser, V., Ruis, H., and Ammerer, G. (2001). Rck2, a member of the calmodulin-protein kinase family, links protein synthesis to high osmolarity MAP kinase signaling in budding yeast. *Proc. Natl. Acad. Sci. USA* **98**, 5625–5630.
- Terfve, C.D.A., Wilkes, E.H., Casado, P., Cutillas, P.R., and Saez-Rodriguez, J. (2015). Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nat. Commun.* **6**, 8033.
- Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.-S.C., Chayes, J., Borgs, C., Zecchina, R., and Fraenkel, E. (2013). Simultaneous reconstruction of multiple signaling pathways via the prize-collecting Steiner forest problem. *J. Comput. Biol.* **20**, 124–136.
- Tuncbag, N., Gosline, S.J.C., Kedaigle, A., Soltis, A.R., Gitter, A., and Fraenkel, E. (2016). Network-based interpretation of diverse high-throughput datasets through the Omics Integrator software package. *PLoS Comput. Biol.* **12**, e1004879.
- Wolf-Yadlin, A., Kumar, N., Zhang, Y., Hautaniemi, S., Zaman, M., Kim, H.-D., Grantcharova, V., Lauffenburger, D.A., and White, F.M. (2006). Effects of HER2 overexpression on cell signaling networks governing proliferation and migration. *Mol. Syst. Biol.* **2**, 54.
- Yan, L. (2009). Abstract #DDT01-1: MK-2206: a potent oral allosteric AKT inhibitor. *Cancer Res.* **69**, DDT01-1.
- Yandell, B.S. (1997). *Practical Data Analysis for Designed Experiments* (Chapman & Hall).
- 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., et al. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat. Genet.* **41**, 316–323.
- Zhang, Y., and Song, M. (2013). Deciphering interactions in causal networks without parametric assumptions. *arXiv*, arXiv:1311.2707v3, <https://arxiv.org/abs/1311.2707>.



**Cell Reports, Volume 24**

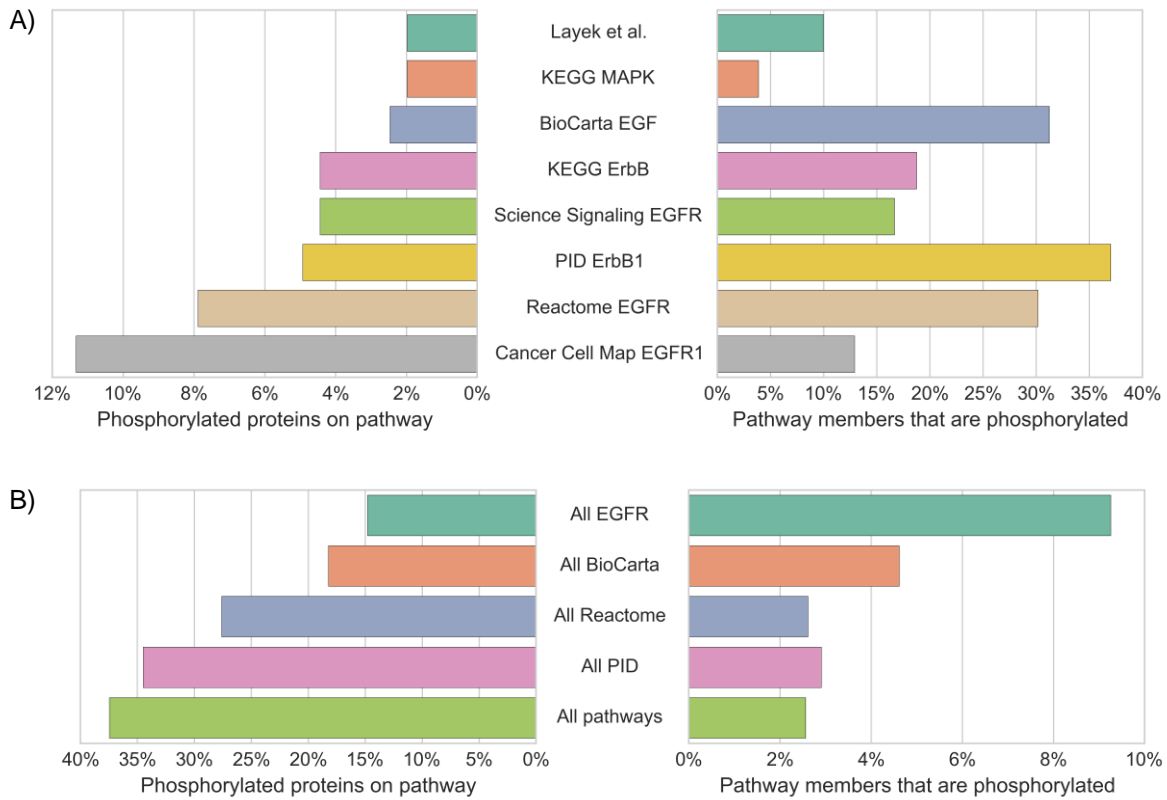
## **Supplemental Information**

### **Synthesizing Signaling Pathways from Temporal**

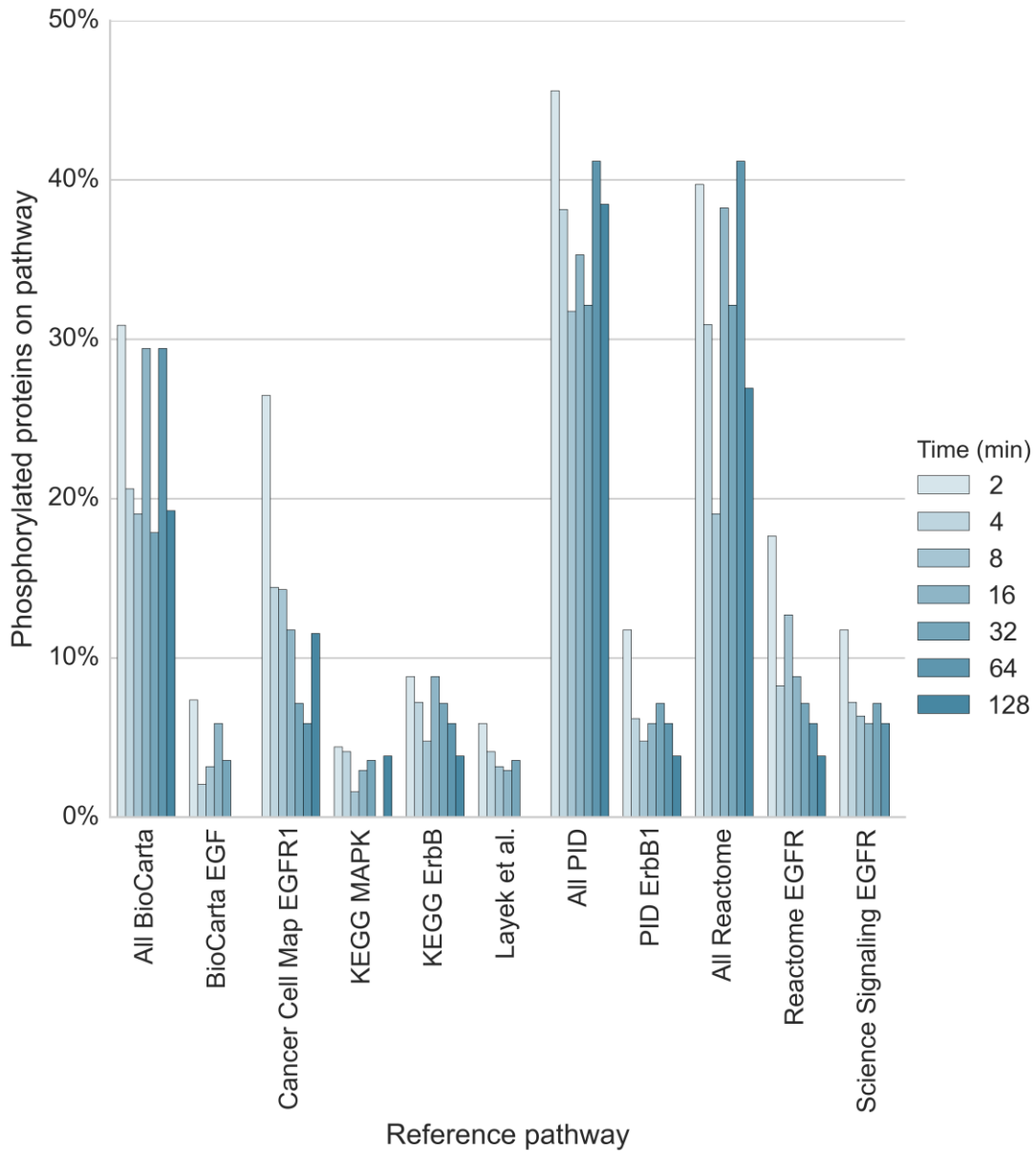
### **Phosphoproteomic Data**

**Ali Sinan Köksal, Kirsten Beck, Dylan R. Cronin, Aaron McKenna, Nathan D. Camp, Saurabh Srivastava, Matthew E. MacGilvray, Rastislav Bodík, Alejandro Wolf-Yadlin, Ernest Fraenkel, Jasmin Fisher, and Anthony Gitter**

## Supplemental Figures



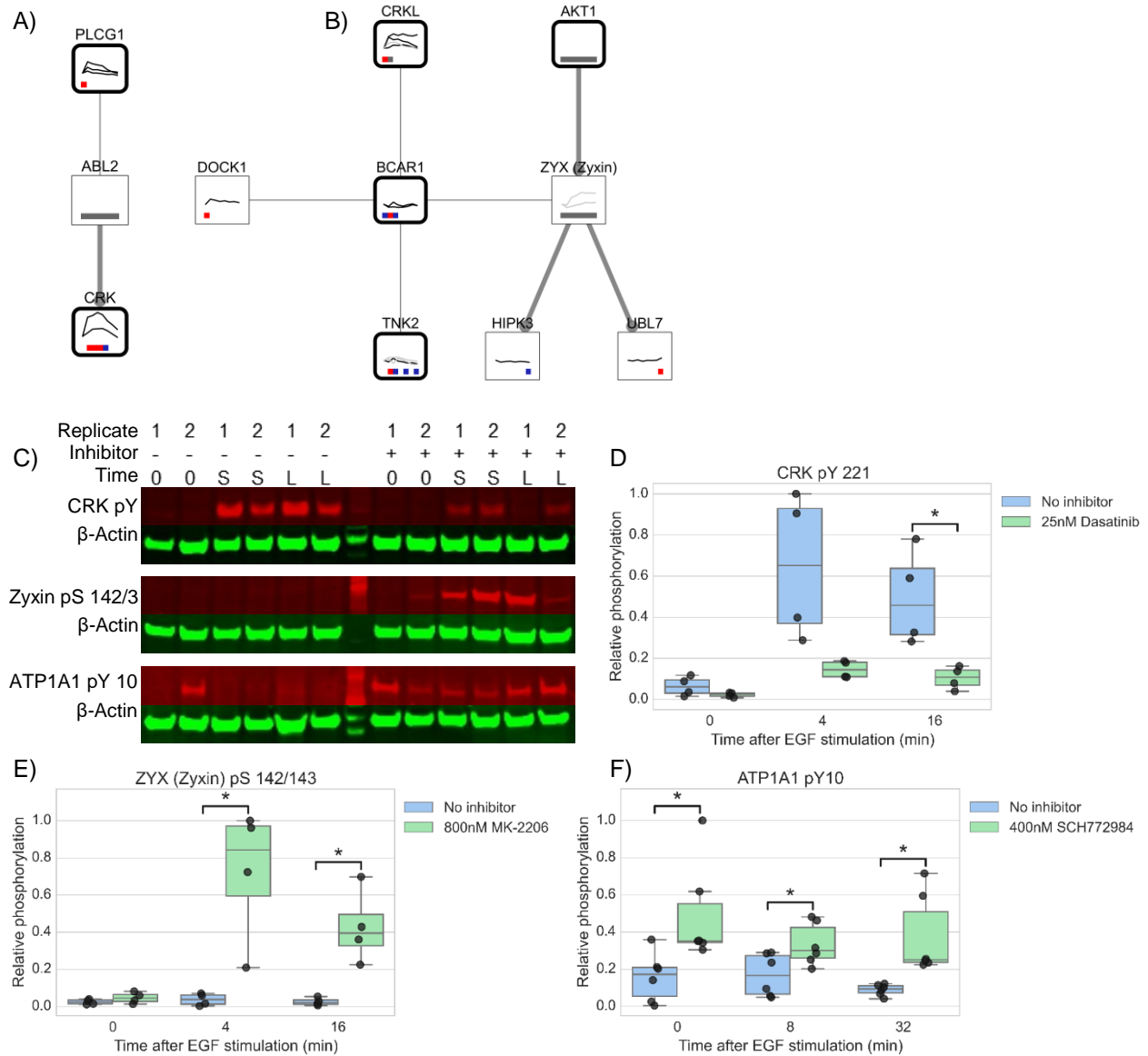
**Figure S1.** Related to Figure 2. A) Over 95% of the significantly differentially phosphorylated proteins in response to EGF stimulation are not included in six of the eight reference pathways. Conversely, the majority of proteins in the reference pathways are not significantly differentially phosphorylated. The Cancer Cell Map pathway achieves the best phosphorylation coverage, but it is still only 11%. B) Most significantly differentially phosphorylated proteins do not appear in any pathway diagram, even when extending the analysis to non-EGFR-specific pathways. All EGFR is the union of the EGFR-related pathways in panel A. All BioCarta, Reactome, and PID reflect the union of all pathways in the respective database. All pathways is the union of all of the above pathways.



**Figure S2.** Related to Figure 2. The percentage of proteins that are significantly differentially phosphorylated and overlap with the reference pathways varies by time point. The percentage is with respect to the number of proteins that are significantly differentially phosphorylated at the given time point: 68 (2 min), 97, 63, 34, 28, 17, and 26 proteins (128 min).

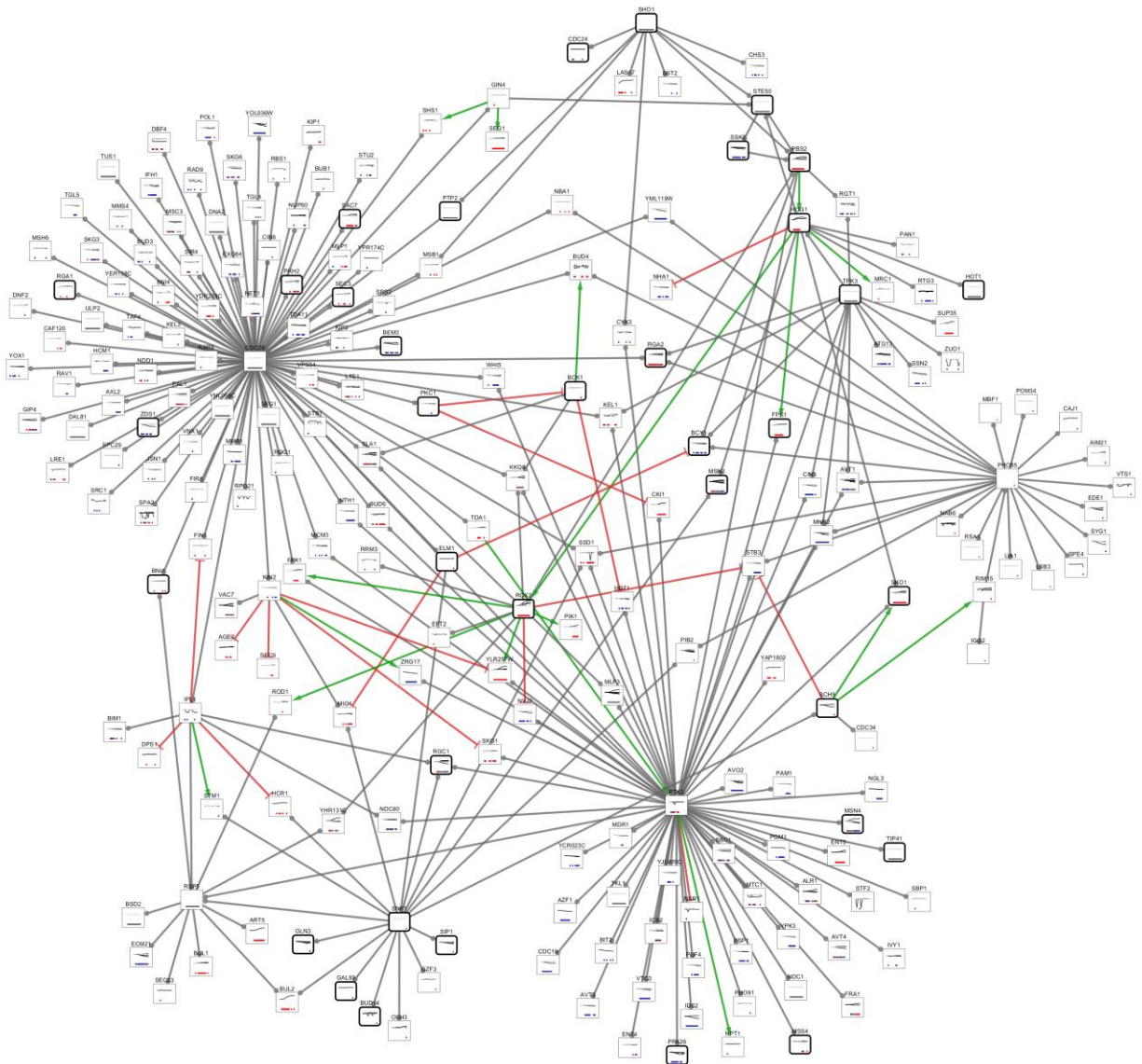


**Figure S3.** Related to Figure 3. The full TPS EGF response pathway summary (Supplemental File 3). Node and edge visualizations are as in Figure 3.

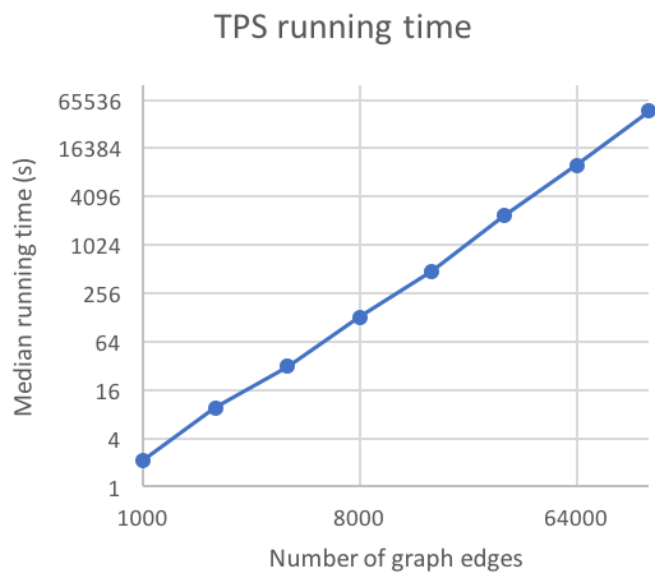


**Figure S4.** Related to Figure 3. Inhibiting predicted pathway edges. A) The predicted pathway context of the ABL2 to CRK interaction. B) The pathway context of the AKT1 to Zyxin interaction, which includes BCAR1, Zyxin, and all of their neighbors. The pathway context of the MAPK1 and ATP1A1 interaction is shown in Figure 3C. C) Western blots for CRK pY 221, Zyxin pS 142/143 and ATP1A1 pY 10 in the presence and absence of small molecule inhibitors targeting their parent node (Dasatinib/ABL2, MK-2206/AKT1, and SCH772984/MAPK1, respectively). The red channel displays detection of the specific phosphorylation sites. The green channel displays detection of  $\beta$ -Actin (a loading control used for normalization of each specific phospho signal). Two biological replicates are shown in each Western blot. Time "0" indicates no EGF stimulation; "S" is short EGF stimulation (four or eight min), and "L" is long EGF stimulation (sixteen or thirty-two min). Absence or presence of inhibitor is shown by "-" and "+", respectively. D) Quantification of CRK pY 221 phosphorylation (four replicates). Phosphorylation levels are relative to the maximum phosphorylation across all conditions and replicates. An asterisk denotes  $p < 0.05$  (two-sided, unpaired, unequal variances t-test). Whiskers show 1.5 times the interquartile range. E) Quantification of Zyxin pS 142/143 phosphorylation (four replicates). F) Quantification of ATP1A1 pY 10 phosphorylation (six replicates).

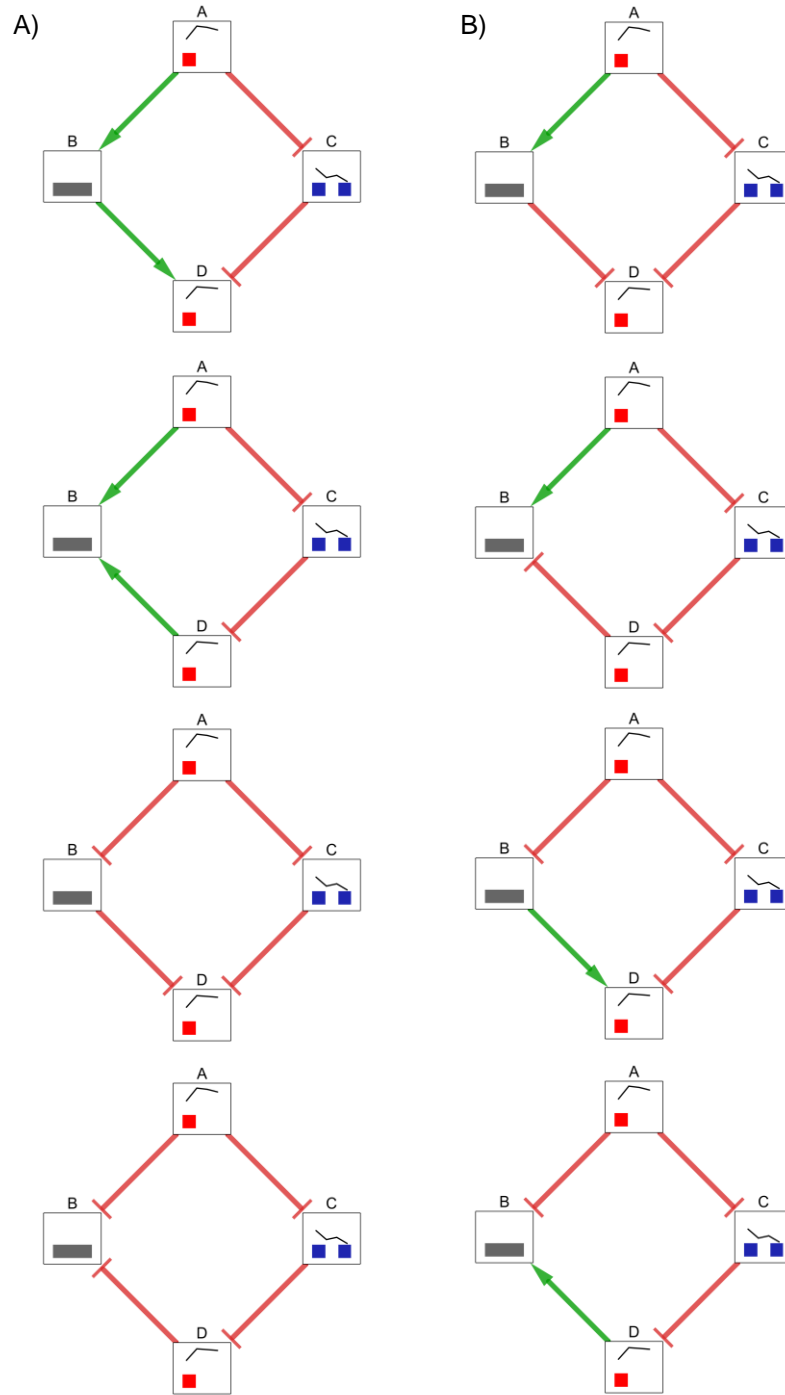




**Figure S5.** Related to Figure 4. The full TPS pathway summary for the yeast osmotic stress response (Supplemental File 3). Node and edge visualizations are as in Figure 3.



**Figure S6.** Related to Experimental Procedures. The median running time of TPS over three replicate runs against the number of input graph edges in log-log scale.



**Figure S7.** Related to Figure 6. An illustration of how the pathway summary graph (Figure 6D) is a generalization of the individual pathway models and can also include invalid models. Not all combinations of direction and sign assignments to the ambiguous edges yield pathways that satisfy all constraints. Here we depict only edges among nodes A, B, C, and D. For simplicity we ignore the models in which one of these edges are absent, even though these models are also included in the summarization. In some cases, B has no valid temporal activity, which we denote with the ambiguous temporal annotation (gray). A) Pathway models in which the signs of edges A-B and B-D are consistent with the constraint that A and D are both activated. B) Pathway models in which these edges have opposing signs and are included in the summary even though they violate a constraint.

## Supplemental Tables

**Table S1.** Related to Figure 3. All predicted interactions tested experimentally. Outcomes are presented in Figure S4 and detailed in (Köksal et al., 2018). MAPK1->MKL1 was tested with two inhibitors. DRP1 is a synonym for DNM1L.

Inhibit	Detect	Compound	Concentration	Antibody	Results
ABL2	CRK	Dasatinib	25 nM	#3491 Phospho-CrkII (Tyr221)	Figure S4
AKT1	RANBP3	MK-2206	800 nM	#9380 Phospho-RanBP3 (Ser58)	No change detected
AKT1	YAP1	MK-2206	800 nM	#13008 Phospho-YAP (Ser127)	No band detected
AKT1	ZYX	MK-2206	800 nM	#8467 Phospho-Zyxin (Ser142/143)	Figure S4
CDK1	MAP2K1	RO-3306	250 nM	#9154 Phospho-MEK1/2 (Ser217/221)	No change detected
EGFR	CCDC50	Erlotinib	200 nM	sc-79367 YMER Antibody (I-12)	No band detected
MAPK1	ATP1A1	SCH772984	400 nM	#3060 Phospho-Na,K-ATPase $\alpha$ 1 (Tyr10)	Figure S4
MAPK1	MKL1	AZD6244	140 nM	Gift from Dr. Ron Prywes MKL1 (Ser454)	No band detected
MAPK1	MKL1	SCH772984	400 nM	Gift from Dr. Ron Prywes MKL1 (Ser454)	No band detected
PRKACA	DNM1L	H-89	480 nM	#4867 Phospho-DRP1 (Ser637)	No band detected
PRKACA	GSK3A	H-89	480 nM	#8452 Phospho-GSK-3 $\alpha$ (Ser21)	No band detected

**Table S2.** Related to Experimental Procedures. The time point-specific log<sub>2</sub> fold changes used to determine significance in the yeast osmotic stress phosphoproteomic dataset.

<b>Time (s)</b>	<b>Threshold for significant dephosphorylation</b>	<b>Threshold for significant phosphorylation</b>
5	-0.4006441	0.3706107
10	-0.3740002	0.3337672
15	-0.3561254	0.3403914
20	-0.3685697	0.3871419
25	-0.4068662	0.3568199
30	-0.4895241	0.4388251
35	-0.5204797	0.4964112
40	-0.4190634	0.4029404
45	-0.3548886	0.3397075
50	-0.4688706	0.4645637
55	-0.4468361	0.3969267
60	-0.4861253	0.4287856



## Supplemental Experimental Procedures

### Cell culture, stimulation, and generation of peptides

Flp-In 293 cells expressing EGFR were described previously (Gordus et al., 2009) and have been used in our previous studies of receptor tyrosine kinase signaling (Wagner et al., 2013). These isogenic cells do not express EGFR heterodimerization receptor partners, and receptor quantities are uniform across cells (~100,000 EGFR/cell). We selected EGFR Flp-In cells for this study because they are easy to manipulate and provide full control of the input signal. We know the number of receptors per cell and thus the ligand concentration necessary to achieve different levels of saturation. Most importantly, because EGFR Flp-In cells are homogeneous with respect to EGFR expression, this system ensures high reproducibility between replicates and minimizes effects of heterogeneous receptor expression between different samples and time points. Although their signaling response may differ from *in vivo* responses in human tissues, this system ensures the phosphorylation changes are EGFR-specific.

Cells were grown using standard cell culture procedures in Dulbecco's modified Eagle's medium supplemented with 10% (v/v) fetal bovine serum, 2 mM glutamine, 100 U/ml penicillin, 100 µg/ml streptomycin, and 150 µg/ml hygromycin B. For the activation of the EGF receptor, the cells were grown in plates to approximately 70% confluency, then washed once with phosphate-buffered saline (PBS) and incubated for 16 hours in serum-free medium. Subsequently cells were stimulated with 23.6 nM EGF (Peprotech) for 0, 2, 4, 8, 16, 32, 64, or 128 minutes. Untreated plates were used for the 0 min time point. After EGF stimulation, cells were lysed on ice with 3 ml of 8 M urea supplemented with 1 mM Na<sub>3</sub>VO<sub>4</sub>. A 10 µl aliquot was taken from each sample to perform the micro bicinchoninic acid protein concentration assay (Pierce) according to the manufacturer's protocol. Cell lysates were reduced with 10 mM DTT for 1 hr at 56 °C, alkylated with 55 mM iodoacetamide for 45 min at room temperature, and diluted to 12 ml with 100 mM ammonium acetate; pH 8.9. 40 µg trypsin (Promega) was added to each sample (~200:1 substrate to trypsin ratio) and the lysates were digested overnight at room temperature. The whole cell digest solutions were acidified to pH 3 with acetic acid (HOAc) and loaded onto C18 Sep-Pak Plus C18 Cartridges (Waters). The peptides were desalted (10 mL 0.1% trifluoroacetic acid (TFA)) and eluted with 10 mL of a solution comprised of 40% acetonitrile (MeCN) with 0.1% TFA. Each sample was divided into ten aliquots and lyophilized overnight to dryness for storage at -80°C.

The peptides were then labeled using 8-plex iTRAQ reagents (Ab Sciex) according to the manufacturer's instructions. 200 µg of lyophilized peptides were resuspended in 30 µl of dissolution buffer, and the corresponding iTRAQ reagent dissolved in 70 µl isopropanol was added. The mixtures were incubated at room temperature for 1 hr and concentrated to ~30 µl. Samples labeled with eight different isotopic iTRAQ reagents were combined and dried to completion. The sample was then rehydrated in 500 µl (0.1% HOAc) and desalted using a Sep-Pack Vac C18 column (Waters). The peptides were eluted with 80% acetonitrile, 0.1% HOAc. The eluate was evaporated to 100 µl in the SpeedVac and lyophilized.

### Phosphopeptide enrichment, mass spectrometry, and identification

Peptides containing phosphotyrosines were enriched using immunoprecipitation (IP). 12 µg of each of the antibodies P-Tyr-1000 (Cell Signaling Technologies), 4G10 (Millipore), and PT-66 (Sigma-Aldrich) were bound to 20 µl of packed protein G Plus agarose beads (Calbiochem) in IP buffer (100 mM Tris-HCl pH 7.4, 0.3% NP-40). The lyophilized peptides were rehydrated with IP buffer, and the pH of the solution was adjusted to 7.4 using 100 mM Tris-HCl pH 8.5. The peptide sample was added to the beads and incubated for 4 hours. The supernatant was then removed and saved for the next step. The beads were washed extensively using IP buffer, 100 mM Tris-HCl and H<sub>2</sub>O. The bound peptides were eluted using 50 µl 15% acetonitrile/0.1% TFA.

Serine and threonine (and remaining tyrosine) phosphorylated peptides were enriched from the IP supernatant using immobilized metal affinity chromatography. Briefly, protein concentration was adjusted to 1 mg/ml protein using wash buffer (80% MeCN/0.1% TFA). 100 µl of Ni-depleted Ni-NTA Superflow beads (Qiagen) were activated with 100 mM FeCl<sub>3</sub>. The supernatant was loaded onto the beads and incubated for 1hr. After washing the beads three times with wash buffer, the bound peptides were eluted twice with 1.4% ammonium hydroxide. The eluates were combined and evaporated in a SpeedVac to 5-10 µl. The sample was then reconstituted to a total volume of 20 µl in 20 mM ammonium formate pH 9.8.

All mass spectrometry experiments were performed on a Thermo Fisher Velos Orbitrap mass spectrometer (Ficarro et al., 2011; Wolf-Yadlin et al., 2006; Zhang et al., 2005) equipped with a nanospray ionsource coupled to a

nanoACQUITY Ultra Performance LC system (Waters) equipped with two binary pumps. Samples were separated using either 1D (IP eluate) or 2D (IMAC eluate) chromatography. For the 1D separation, the sample was loaded onto a 5 cm self-packed (Reliasil, 5  $\mu$ m C18, Orochem) pre-column (inner diameter 150  $\mu$ m) connected to a 20-cm self-packed (ReproSil, 3  $\mu$ m C18, Dr. Maisch) analytical capillary column (inner diameter 50  $\mu$ m) with an integrated electrospray tip ( $\sim$ 1  $\mu$ m orifice). Peptides were separated using a 115-minute gradient with solvents A (H<sub>2</sub>O/formic acid (FA), 99.9:1 (v/v)) and B (MeCN/FA, 99.9:1 (v/v)) as follows: 1 min at 2% B, 84 min from 98 to 40% B, 5 min at 40% B, 20 min at 20% B, and 14 min at 2% B. For the 2D reverse phase chromatography, the sample was first loaded onto a 5 cm self-packed Xbridge column (Waters, inner diameter 150  $\mu$ m) and eluted with a 7-step gradient of 1, 3, 6, 9, 13, 25, and 44% B with solvents A (H<sub>2</sub>O/20 mM ammonium formate pH 9.8) and B (MeCN/20 mM ammonium formate, pH 9.8). The eluted sample was directly loaded onto a 5 cm self-packed precolumn (Reliasil, 5  $\mu$ m C18, Orochem), which was connected to a 20 cm self-packed analytical column (ReproSil, 3  $\mu$ m C18, Dr. Maisch). The peptides were eluted with the same gradient as described above in the second dimension.

Eluted peptides were directly analyzed using the Velos-Orbitrap mass spectrometer operated in data-dependent acquisition (DDA) mode to automatically switch between MS and MS/MS acquisitions. The Top 10 method was used, in which full-scan MS (from m/z 350–2000) was acquired in the Orbitrap analyzer at 120,000 resolution, followed by high-energy, collision-induced dissociation (HCD) MS/MS analysis (from m/z 100–1700) of the top 10 most intense precursor ions with a charge state  $>2$ . The HCD MS/MS scans were acquired using the Orbitrap analyzer at 15,000 resolution at a normalized collision energy of 45%, with the ion selection threshold set to 10,000 counts. Precursor ion isolation width of 2 m/z was used for the MS/MS scans, and the maximum allowed ion accumulation times were set to 500 ms for full (MS) scans and 250 ms for HCD (MS/MS). The standard mass spectrometer tune method settings were as follows: Spray Voltage, 2.2 kV; no sheath and auxiliary gas flow; heated capillary temperature, 325 C; Automatic Gain Control (AGC) enabled. All samples were analyzed by LC-MS/MS in biological triplicates.

MS/MS data files were searched against the human protein database using Comet (Eng et al., 2013). Variable (phosphorylation of serine, threonine, or tyrosine, 79.966331 Da, methionine oxidation, 15.9949 Da) and static (carbamidomethylation of cysteine, 57.02 Da and the iTRAQ modification of 304.205360 Da to peptide N-terminus and lysine side-chains) modifications were used for the search (Supplemental File 5). We applied a 1% false discovery rate threshold at the peptide level. Quantitation of the iTRAQ signals was performed using Libra (Deutsch et al., 2010). We deposited the raw mass spectrometry data to the ProteomeXchange Consortium via the PRIDE (Vizcaíno et al., 2016) partner repository.

### **Mass spectrometry data analysis and temporal phosphorylation significance**

We obtained two technical replicates for each of the three biological replicates, quantifying IP and IMAC eluate separately (Supplemental File 1). To decrease noise, we applied an iTRAQ channel intensity cutoff of 5,000 for IP and 10,000 for IMAC (arbitrary units), removing peptides that did not exceed the threshold in at least one channel. We median centered the data within each channel by dividing the signal by the channel median. For each biological replicate, we concatenated the technical replicates and the IP and IMAC intensities. We removed non-phosphopeptides and aggregated the signal for phosphopeptides with the same sequence (regardless of post-translational modifications) by summing the normalized values (Supplemental File 1). Across the three biological replicates, we quantified 5,442 unique peptides in at least one replicate and 1,068 peptides in all replicates. We focused on these 1,068 peptides for the computational modeling because the repeated observations indicate more reliable quantification and let us assess the significance of phosphorylation changes (Supplemental File 1).

After filtering peptides missing data in one or more biological replicates, we median centered the data again and performed Tukey's Honest Significant Difference (HSD) test (Yandell, 1997) for each peptide (Supplemental File 1). Tukey's HSD test reports the significance in the difference of the mean peptide intensities for all pairs of time points, correcting for the multiple comparisons within each peptide's time course. It is conceptually similar to performing t-tests for all pairs of time points, but it conducts all tests simultaneously and adjusts for multiple testing appropriately. Although we have temporal data, we selected Tukey's HSD test instead of a temporal statistical test like EDGE (Leek et al., 2006) because TPS requires a test statistic for individual time points, not the entire time course.

Tukey's HSD test is already quite conservative (Yandell, 1997). However, it does not correct for the multiple comparisons from testing multiple peptides. As a precaution, we computed q-values (Storey and Tibshirani, 2003)

from the distribution of p-values for all pairs of time points and all peptides. Our p-value threshold  $p < 0.01$  is equivalent to a q-value threshold  $q < 0.188$ , which we consider to be an acceptable false discovery rate for this application. In addition, the predicted network is reasonably robust to the choice of p-value threshold (Köksal et al., 2018).

Our temporal analysis was based on two types of significant changes in phosphorylation: changes relative to the baseline (0 minute) time point and changes relative to the previous time point. For each time point, we took the smaller of these two p-values (comparison to baseline or previous) from Tukey's HSD test as the significance of that time point and produced an aggregate p-value for the peptide as the minimum such p-value over all time points. The aggregate p-value serves as a proxy of whether the peptide significantly changes in response to stimulation at any time point, and 263 peptides responded significantly ( $p < 0.01$ ). In addition, we calculated a summary phosphorylation profile for each peptide by taking the median value at each time point. In order to visualize the  $\log_2$  fold changes (as in Figure 2B), we used the only non-zero replicate for the three peptides where the median value was zero. This substitution was only for visualization; TPS does not operate on fold changes. We mapped peptides to UniProt identifiers, using the best-matching protein identifier for each peptide.

We cannot recover a kinetic model of kinase (or phosphatase) activity from the mass spectrometry data, but our temporal constraints assume that there is an unknown, hidden phosphorylation level at which a kinase (phosphatase) begins to substantially (de)phosphorylate its target. This phosphorylation level may vary by protein, so we do not set a threshold on the phosphorylation fold change. Furthermore, the protein may continue to increase (or decrease) in phosphorylation past this critical level even after it has begun to affect the substrate, so we do not assume that the time of peak phosphorylation is the only relevant time.

### Quantitative Western blotting

The following kinase inhibitors were used at the following concentrations: 25 nM Dasatinib (#S1021), 400 nM SCH772984 (#S7101), and 800 nM MK-2206 (#S1078, all Selleckchem). The Flp-In 293 EGFR cells were serum-starved for 16 hours in growth medium without FBS. Then, if indicated, the kinase inhibitors were added and incubated with the cells for 60 min. Only DMSO was added to control cells. After stimulation with 23.6 nM EGF for the indicated times, the cells were lysed using RIPA buffer (25mM Tris-HCl pH 7.6, 150mM NaCl, 1% NP-40, 1% sodium deoxycholate, 0.1% SDS) with 1 mM sodium orthovanadate. Cells without EGF stimulation were used as controls (0 min time point). After incubation on ice for 15 minutes, the lysates were centrifuged at 21000 g for 10 min, and the protein concentration of the supernatants was determined. Protein amounts were adjusted to 30–50 ug protein/well, and protein phosphorylation was assessed by SDS separation and Western blotting. The membranes were probed with the following antibodies at a dilution of 1:1000: pY221-CRK (#3491, Crk-II isoform), pY10-ATP1A1 (#3060), and pS142/143-Zyxin (#8467, all Cell Signaling Technologies).  $\beta$ -actin (#3700) was used to normalize loading across the gel. Fluorescently labeled secondary antibodies were added according to the manufacturer's instructions at 1:5000 (Goat anti rabbit IRDye 680 and Goat anti mouse IRDye 800, Li-COR Biosciences). Blots were imaged using an Odyssey Infrared Imaging System (Li-COR Biosciences). Quantification of the phosphorylated proteins was performed with the Odyssey analysis software.

### Translating the input into constraints

A symbolic signed, directed graph that assigns a sign and direction to each edge in (a subgraph of) the undirected input graph can be represented by maintaining four Boolean variables per edge, one for each sign-orientation combination. The truth value of a variable denotes whether there is an edge with the corresponding sign and orientation in the solution network. To find a tree network model rooted at the stimulated source node, we need to constrain these truth values. First, we assert that at most one of these four variables can be true. The case where all four variables are false corresponds to the undirected edge being excluded from the solution network. Then, we assert that there are no cycles in the solution graph. To implement the acyclicity constraint, we maintain one integer-valued variable per node and assert that the integer values along all directed paths must monotonically increase. Finally, we assert that if a non-source node has an outgoing edge, it must have an incoming edge as well. This prevents modeling spurious phosphorylation changes that are not caused by the source's stimulation. These constraints together guarantee that we obtain a tree network model in which all edges are on a directed path originating at the source node.

**Example:** The edge ( $A, B$ ) in the undirected graph shown in Figure 5C, like all other edges in the network, is translated into four Boolean variables, *activation-A-B*, *inhibition-A-B*, *activation-B-A*, *inhibition-B-A*. We ensure at

most one variable is true by asserting that if one of the variables is true, the rest must be false. For instance, we assert:

$$activation-A-B ==> !inhibition-A-B \&\& !activation-B-A \&\& !inhibition-B-A$$

which means that if *activation-A-B* is true, the remaining variables must all be false. Both A and B have an associated integer variable, *index-A* and *index-B*, and we state that if there is an edge from A to B, B must be assigned a greater value than A:

$$activation-A-B // inhibition-A-B ==> index-A < index-B$$

A similar constraint is asserted for the opposite direction. Finally, we create a constraint that requires B to have an incoming edge if it has an outgoing edge, based on its neighbors in the undirected graph:

$$activation-B-A // inhibition-B-A // activation-B-D // inhibition-B-D \\ ==> \\ activation-A-B // inhibition-A-B // activation-D-B // inhibition-D-B$$

These constraints together guarantee that a valid solution must be a tree network rooted at A. ■

While the above constraints will ensure that solutions satisfy topological properties, they don't constrain models with respect to the temporal data. Using the temporal events computed from the time series data, TPS requires that the sequence of nodes in each signed, directed path of a tree model must be supported by a corresponding temporally ordered sequence of phosphorylation events. In the example from Figure 5, there can be no models that include an edge from E to D, because it is impossible for E to precede D in a directed path due to all possible activations of E being later than the possible activation of D (Table 1). The same example shows that the temporal ordering along paths can also have an effect beyond pairwise interactions. The sequence of nodes E, G, F cannot appear on a directed path, even though both pairwise interactions are locally consistent, because E can only be activated strictly after F. Concretely, this constraint is enforced by keeping an integer-valued variable for each node, which corresponds to the choice of activation time for that node. The same is done for representing inhibitions, and we assert that at most one of the two events can occur. We restrict the values that the activation variable can take to the time points computed in the discretization step. Finally, we state that if there is an edge from A to B in the signed, directed tree, there must be corresponding choices of time points for A and B that support the interaction. An activation edge from A to B must be supported by the activation (respectively, inhibition) of A, succeeded by the activation (respectively, inhibition) of B; similarly, an inhibition from A to B requires finding an activation (respectively, inhibition) of A, succeeded by the inhibition (respectively, activation) of B.

**Example:** Consider the nodes D and E in Figure 5. We constrain activation choices for D to be 0 (no activation) or 1 (the first time interval):

$$activation-D == 0 // activation-D == 1$$

Similarly, E is either not activated or is active in intervals 2 or 3:

$$activation-E == 0 // activation-E == 2 // activation-E == 3$$

Finally, we assert that if there is an activation from D to E, both nodes must be activated or inhibited, in that order:

$$activation-D-E ==> \\ activation-D != 0 \&\& activation-E != 0 \&\& activation-D <= activation-E$$

We only show constraints for the activation of E by D through their successive activation events, the other cases are similar. ■

The last type of constraint that TPS enforces follows simply from the prior knowledge information. For all known kinase-substrate interactions (given as directed, unsigned edges), no pathway model can include an edge directed in the opposite orientation. This is implemented by ruling out certain values for the edge variables if data is available for a given edge. TPS currently represents kinase-substrate interactions as unsigned but could be trivially extended to treat kinase-substrate interactions as positive edges and phosphatase-substrate interactions as negative edges.

**Example:** In the example from Figure 5D, we are given the kinase-substrate interaction C-D. As a result, we rule out the opposite direction:

$$!activation-D-C \ \&\& \ !inhibition-D-C$$

■

### Extending constraints to model feedback loops

Although TPS does not model feedback loops, it is possible to extend our declarative constraint solving approach to infer network models with feedback. We sketch below one way to allow each node to change its binary activity level up to N times, thereby allowing a node's early activity to influence its late activity through cycles in the network model. The high level idea consists of using multiple copies of the constraint system we describe above, each copy corresponding to a different "layer" in the network. Conceptually, each layer describes temporal activity that comes after all activity in the previous layer. While each layer corresponds to a tree network (using our existing constraint system), their composition represents cyclic graphs that can model feedback.

For the layered system of constraints, we need to formulate constraints that relate each successive layer to its predecessor. Specifically, having an outgoing edge from a node in a given layer requires the same node to have an incoming edge in the same layer or the preceding layer. Furthermore, because each layer corresponds to activity that follows the previous layer, the activation interval of a node in one layer must not come before its chosen activation interval in the preceding layer. Together, these requirements ensure that all activity seen across layers originates from the source node in the first layer, and the order of node activations across layers is correct.

**Example:** We revisit the constraints generated for the edge A-B in the example undirected graph shown in Figure 5C. If we want to model at most two activity intervals per node, we will have two copies of the Boolean variables expressing the sign and orientation of the edge, and we will ensure that at most one Boolean variable is true per layer by asserting the constraint below. We refer to the Boolean variable representing an activation from A to B in layer  $i$  as  $activation_i-A-B$ .

$$activation_1-A-B \ ==> \ !inhibition_1-A-B \ \&\& \ !activation_1-B-A \ \&\& \ !inhibition_1-B-A$$

$$\ \&\&$$

$$activation_2-A-B \ ==> \ !inhibition_2-A-B \ \&\& \ !activation_2-B-A \ \&\& \ !inhibition_2-B-A$$

The modified connectivity constraint for node B will now require B to have an incoming edge in the current layer or the previous layer based on its neighbors in the undirected graph. For instance, we will assert the following for layer 2:

$$activation_2-B-A \ || \ inhibition_2-B-A \ || \ activation_2-B-D \ || \ inhibition_2-B-D$$

$$\ ==>$$

$$activation_2-A-B \ || \ inhibition_2-A-B \ || \ activation_2-D-B \ || \ inhibition_2-D-B$$

$$\ ||$$

$$activation_1-A-B \ || \ inhibition_1-A-B \ || \ activation_1-D-B \ || \ inhibition_1-D-B$$

■

It is possible to further extend this modeling approach to require successive activations of the same node to correspond to distinct time intervals. This can be achieved by asserting that if a node has an incoming edge in layer  $i$  and an outgoing edge in a preceding layer  $j$ , then the activation interval picked for layer  $i$  must come strictly after the interval picked for layer  $j$ .

**Example:** For our running example, we will assert the following constraint for node B. We denote the variable corresponding to the activation interval chosen for B in layer  $i$  as  $activation_{i-B}$ .

$$\begin{aligned}
 & activation_{2-A-B} \parallel inhibition_{2-A-B} \parallel activation_{2-D-B} \parallel inhibition_{2-D-B} \\
 & \quad \&\& \\
 & activation_{1-B-A} \parallel inhibition_{1-B-A} \parallel activation_{1-B-D} \parallel inhibition_{1-B-D} \\
 & \quad ==> \\
 & activation_{1-B} < activation_{2-B}
 \end{aligned}$$

■

This approach to model feedback loops using constraints is similar to the way dynamic Bayesian networks create multiple copies of each variable by "unrolling" the network over each time point. Whereas DBNs have to make  $N$  copies of the network for  $N$  time points, we only need one copy per node activation. Using the above constraint formulation, a solution without feedback loops will still use only one copy of each variable. If we allow two activations per node, we will only need two copies of each variable, irrespective of how many time points there are in the time course dataset.

### Pathway summarization

The space of all valid pathway models with timing annotations defined by the constraints we specified is typically very large, and enumerating all models is not computationally feasible. Given an undirected network  $G$  with  $V$  nodes and  $E$  edges, along with  $T$  time points, there are  $5^E$  ways of assigning a sign and orientation to edges of  $G$  and  $(T*2 + 1)^V$  ways of assigning timing annotations to its nodes. Even for a network with 200 edges, the number of possible sign and orientation assignments is  $6*10^{139}$ . Pathway summarization supports these large state spaces because it is intractable to enumerate all valid pathway models.

The main text describes the edge query procedure used to generate pathway summaries. The example summary in Figure 6 and the summary expansion in Figure S7 illustrate how this summary can contain a superset of the valid pathway models. There exists no valid model that contains an activation from A to B and an inhibition from B to D. The existence of the first edge dictates that B is activated, which implies an inhibition from B to D would decrease D's activity, contrary to what is observed in D's temporal activity profile. However, the knowledge that the edges A-B and B-D must have the same sign is lost through the summarization process.

For visualization and analysis purposes, pathway summaries are depicted as interactions among proteins even though the temporal consistency constraints operate at the level of individual peptides when peptide-level data are available. The protein-level summaries collapse the expanded PPI network, which can introduce ambiguities if there are interactions that are unambiguous at the peptide-level that conflict in terms of direction or sign at the protein-level. TPS is able to detect and report this loss of precision when transitioning to the protein-level network.

A final summarization observation relates to the distinguishability problem between trees and directed acyclic graphs (DAGs) that we discussed in the context of our modeling assumptions. We note that summarizing the space of all tree models as a union graph leads to the same result as summarizing the space of DAGs satisfying the same properties. This stems from the fact that for each DAG model, there exists a set of tree models whose union is the DAG. As a result, the union of all tree models corresponds to the union of all DAGs.

### Tradeoffs between ambiguity, expressiveness, and correctness

The modeling assumptions made when interpreting and translating biological data into logical constraints have complex effects on the degree of ambiguity, expressiveness, and accuracy of the resulting pathway summary. Even with temporal information, many pathway structures can explain the ordered signaling events. This motivates the



reduction of ambiguity with hard logical constraints, where each constraint is fully trusted, instead of with probabilistic constraints (Hinton et al., 2006; Katoen et al., 2005), where a constraint can potentially be violated.

In the PPI network, we allow paths only through chains of experimentally detected PPI. In settings where the PPI network is less complete, we could include edges among highly correlated phosphorylated proteins or predicted interactions based on protein sequence, protein structure, pathway connectivity, or literature mining (Lees et al., 2011; Mosca et al., 2013). The pre-processing step that filters the PPI network operates on a weighted network. These additional edges could be assigned lower weights so that PCSF includes them in the TPS input network only if they are critical for connecting significantly phosphorylated proteins. This would reduce the impact of missing interactions on TPS pathways at the cost of potentially increasing ambiguity because there would be more possible paths through which signal can flow.

Likewise, we observe that some proteins, such as RAS and RAF family members, are not included in the TPS pathway because our mass spectrometry data do not detect their phosphorylation. To increase robustness to potential false negatives in the mass spectrometry, the input PPI network could be modified to include edges from relevant reference pathways with high weights (similar to (Patil et al., 2013)) so that PCSF prefers to include these interactions instead of other high-confidence connections in the PPI network. The weight of these prior knowledge edges would control the tradeoff between condition-specific *de novo* pathway discovery and conformance with prior knowledge.

Unlike single-cell mass cytometry data, where the peak activity times of a small number of phosphoproteins can be resolved precisely (Krishnaswamy et al., 2014), phosphorylation timing in cell population-level mass spectrometry data is inherently ambiguous. Therefore, instead of rigidly determining a protein’s time of activity by selecting the time point at which the greatest phosphorylation change is observed, TPS takes a more general approach. It allows a protein to be activated or inhibited whenever the phosphorylation significantly differs from the level before stimulation or at the immediately preceding time point as long as it is the *first time* at which that phosphorylation level has been observed. We focus on the initial pulse of signaling activity following stimulation, sampling more early time points in our EGF response study because we are more confident that these changes in phosphorylation intensity are due to PTMs instead of changes in protein abundance. Feedback loops cannot be detected when learning a single activation or inhibition time per peptide, a modeling decision we made in our three case studies. However, the TPS framework makes it possible to allow multiple activity changes per peptide in future applications, as we outlined above. Statistical tests of the temporal phosphorylation profiles could determine the number of significant activity changes for each peptide. Then, TPS could search for pathway structures with feedback loops that explain the multiple activation or inhibition events per peptide.

Lastly, we recognize that different phosphopeptides on the same protein can have different phosphorylation changes over time, and we allow each peptide to have its own activation times instead of forcing a single time per protein. This decision can lead to ambiguous edge direction predictions at the protein-level even when the directions are consistent at the peptide level. For example, DOCK1 interacts only with BCAR1 (Figure S4B), yet the direction and sign of the interaction are ambiguous. The uncertainty arises because BCAR1 is phosphorylated on both Y249 and Y387. TPS correctly concludes that the sign cannot be determined because one site could activate DOCK1 and then feed back and affect the other BCAR1 site.

### Prize-collecting Steiner forest

PCSF recovers the sparse subnetwork  $F = (V_F, E_F)$  from a dense PPI network that links proteins of interest by solving

$$\operatorname{argmin}_F \sum_{v \in V_F} (\beta \cdot p(v) - \mu \cdot d(v)) + \sum_{e \in E_F} c(e) + \omega \cdot \kappa$$

where  $p$  is a positive score (prize) that reflects the relevance of a vertex (protein)  $v$ ,  $d$  is the degree (number of neighbors) of  $v$ ,  $c$  is a positive cost for including an edge (interaction)  $e$  in the subnetwork, and  $\kappa$  is the number of disconnected trees in the subnetwork. Parameters  $\beta$ ,  $\mu$ , and  $\omega$  control the size and structure of the solution subnetwork. We swept over many PCSF parameter combinations and identified the parameters that produced reasonably large subnetworks containing many phosphorylated proteins. Including a greater fraction of phosphorylated proteins in the PCSF subnetwork, as opposed to the connective Steiner nodes without observed phosphorylation changes, provides more temporal information for the TPS constraints. An advantage of PCSF is

that it nominates pathway members that are not detected by the mass spectrometry but form critical pathway connections to phosphorylated proteins, like ABL2 and AKT1 in our EGF response study (Figure S4).

In our EGF response analysis, we computed the node prize for each protein using the minimum peptide p-value over all peptides that map to the protein. We computed prizes as  $-\log_{10} pvalue$ , yielding 701 protein prizes. The proteins were not filtered at a particular p-value threshold. There are no replicates in the Olsen et al. EGF response dataset so we used phosphorylation fold changes to define prizes instead of p-values. For each peptide, we computed the  $\log_2$  phosphorylation fold changes at each time point with respect to both the previous time point and the initial 0 minute time point. We extracted the maximum absolute value  $\log_2$  fold change for each peptide as the peptide prize. We then derived protein prizes by taking the maximum peptide prize over all peptides that map to a protein.

The human PPI network we used for both EGF analyses was obtained from two sources: 159,095 undirected interactions from iRefIndex (version 13.0) (Razick et al., 2008) and 4,080 directed kinase-substrate interactions from PhosphoSitePlus (downloaded October 16, 2013) (Hornbeck et al., 2015). The iRefIndex database provides interaction confidence scores that are inspired by the Molecular INteraction database (MINT) (Ceol et al., 2010; Villaveces et al., 2015), which account for the number of publications supporting the interaction, the type of interaction, and the experimental detection methods. Similarly, we scored the directed interactions based on the number of interactions reported for each kinase-substrate pair (across all substrate sites) and whether an interaction was detected *in vitro*, *in vivo*, or both. When merging the iRefIndex and PhosphoSitePlus interactions, if both databases reported an interaction between a pair of proteins we retained the more specific, directed kinase-substrate interaction and its weight, discarding the undirected iRefIndex interaction. The final network contains 15,677 proteins, 157,984 undirected interactions, and 3,917 directed interactions (reciprocal pairs of directed edges with the same score are represented as an undirected edge). It contains 653 of the 701 proteins with mass spectrometry-based prizes from our EGF response data. PCSF respects the direction of the kinase-substrate interactions so that they can only be used in the specified direction.

We derived PCSF protein prizes for the yeast osmotic stress study from the time-point specific salt/control  $\log_2$  fold changes. For each protein, we computed the maximum absolute value  $\log_2$  fold change considering all phosphopeptides mapping to that protein and all time points. This produced 1,596 protein scores from the 4,337 phosphopeptides. Because a complete control time series was available, we also used the temporal salt versus control  $\log_2$  fold changes to obtain the temporal significance scores for TPS. Kanshin et al. used the non-phosphorylated peptides that were not filtered during phosphopeptide enrichment to define a null distribution of  $\log_2$  fold changes at each time point (Kanshin et al., 2015). From these null distributions, they derived time point-specific  $\log_2$  fold change thresholds that correspond to a p-value  $< 0.05$  after adjusting for multiple hypothesis testing (Table S2). We called a phosphopeptide significant at any time point where its  $\log_2$  fold change was more extreme than the time point-specific threshold, which produced 1,401 phosphopeptides and 784 proteins with significant dynamic responses. We did not consider the 0 second fold changes. All analyses considered only the singly phosphorylated (1P) peptides.

Chasman et al. compiled the yeast interaction network we used for the osmotic stress model from many sources (Chasman et al., 2014), with the vast majority of kinase-substrate interactions coming from Ptacek et al. (Ptacek et al., 2005). We removed edges where the interaction evidence was listed as "inferred" or the edge type was "cxcx", "cxorf", or "metapath". To emphasize post-translational modifications, we retained 7,583 directed interactions from this network, which included 4,875 phosphorylation-specific interactions with the edge type "kinasesubstrate", "literaturephos", or "phosphatasesubstrate". Not all source nodes were connected to the rest of the network with these directed edges so we also included 32 undirected PPI between the sources and other proteins. Because the network was originally unweighted, we assigned a default weight of 0.5 to all 7,615 edges.

The PCSF parameters  $\beta$ ,  $\mu$ , and  $\omega$  influence the structure of the optimal PPI subnetwork (Tuncbag et al., 2016). As  $\beta$  grows, the incentive to include (de)phosphorylated proteins increases, outweighing the expense of adding additional edges to connect them. The PCSF solution includes novel proteins, termed Steiner nodes, that are not phosphorylated but are useful for connecting the proteins with positive prizes. To control for hub proteins in the Steiner forests, which can produce non-condition-specific solutions due to their high-connectivity and ability to link many phosphorylated proteins,  $\mu$  penalizes the inclusion of high-degree nodes. Large  $\mu$  results in fewer hubs in the forest.

In order to select the parameters  $\beta$  and  $\mu$  for our EGF response data, we ran PCSF with all combinations of  $\beta$  from 0.05 to 1.0 (step size of 0.05) and  $\mu$  from 0 to 0.01 (step size of 0.001), producing 220 Steiner forests. Because the source of stimulation is known and specific, we rooted the subnetwork at EGF. This forces  $\kappa = 1$ , reducing the problem to the prize-collecting Steiner tree problem and removing the influence of  $\omega$ . We set msgsteiner's parameters  $D = 10$  and  $g = 0.001$  (msgsteiner version 1.1) (Bailly-Bechet et al., 2011). After removing the 34 small or empty solutions with two or fewer Steiner nodes, we selected  $\beta = 0.55$  and  $\mu = 0.008$ , the solution that maximized the fraction of proteins in the Steiner forest that had prizes (were phosphorylated or dephosphorylated in response to EGF). In order to recover alternative connections among the (de)phosphorylated proteins, we generated a collection of 100 PCSF networks. For each PCSF run, we used the previously selected  $\beta$  and  $\mu$  and added random noise to all interaction costs by setting msgsteiner's  $r = 0.01$ . The union of these 100 networks was used as input to TPS.

The Olsen et al. EGF response PCSF analysis used the same source node, parameter sweeping strategy, and msgsteiner parameters (msgsteiner version 1.3). For this dataset, we discarded 47 solutions with two or fewer Steiner nodes, including empty networks, and chose  $\beta = 0.45$  and  $\mu = 0.005$  for generating the final collection of 100 PCSF networks. These parameters produced a network in which the fraction of proteins with prizes, 0.88, was slightly less than the maximum over all solutions, 0.90. However, we preferred this network because it was closer in size to the PCSF network obtained using our EGF response data, facilitating their direct comparison.

For the yeast osmotic stress study, we selected the five sources from the KEGG yeast high osmolarity pathway: Hkr1, Msb2, Opy2, Sho1, and Sln1. We ran PCSF with all combinations of  $\beta$  from 0.25 to 10.0 (step size of 0.25),  $\mu$  from 0 to 0.1 (step size of 0.005),  $\omega$  from 0.5 to 10 (step size of 0.5), generating 16,800 Steiner forests. We discarded 14,562 solutions with two or fewer Steiner nodes, including empty networks. Multiple Steiner forests had the same fraction of proteins with prizes, all of which had the same  $\beta$  and  $\mu$  parameters. From this set of forests, we selected  $\beta = 1.75$ ,  $\mu = 0.095$ , and  $\omega = 4.5$  to generate a collection of 1000 Steiner forests using the same msgsteiner parameters as the EGF analyses (msgsteiner version 1.3).

For all three phosphoproteomic datasets, the temporal phosphorylation data included measurements of individual peptides. TPS expands the subnetwork from PCSF by replacing each protein node by a collection of peptide nodes. These peptide nodes are all peptides that are significantly phosphorylated in the mass spectrometry data and map to the protein. The protein-protein edges are transformed to peptide-peptide edges by connecting all peptide nodes that replace protein A to all peptide nodes that replace protein B if there was a PPI between A and B.

### Phosphoproteomic and network data randomization

We performed three types of data randomization to evaluate the impact on the TPS predictions for our EGF response data. First, we permuted the peptide to protein assignments. Next, we permuted time points within each time course except for the baseline (0 min) time point. Finally, we permuted the PPI network with BiRewire (Iorio et al., 2016). BiRewire supports undirected or directed graphs but not partially directed graphs, so we randomized the directed edges and undirected edges separately and then merged them. For each type of permutation, we ran PCSF and TPS on 100 different versions of the randomized data. Because the phosphorylation timing does not affect the protein prizes, we did not rerun PCSF when permuting the time points. Instead, we reused the PCSF subnetwork from the original data. The PCSF and TPS parameters for the permuted datasets were the same as for the original data. We collected aggregate metrics that compare the permuted data results to the TPS network obtained using the original dataset (Köksal et al., 2018).

In addition, we performed bootstrap analyses by randomly removing 10, 25, or 50% of the significant phosphopeptides. For each fraction of held out data, we generated protein prizes and ran PCSF and TPS 100 times using the original parameters. We again computed aggregate metrics that compare the permuted data results to the TPS network obtained using the original dataset (Köksal et al., 2018).

### Evaluation against pathway databases

In order to compare our predictions against a set of manually curated pathway databases, we developed a framework that collects signed, directed protein-protein interactions from the BioPAX L3 and SIF formats. Our framework allows us to perform a systematic comparison against reference pathways. We extracted pathway interactions from multiple references represented using BioPAX L3. The comparison against SIF networks is straightforward because SIF files are simply lists of pairwise protein-protein interactions.

Extraction from pathway maps is performed in two steps. First, we extract all proteins, complexes, and paths reported between these entities. Then, we query the extracted paths for reachability between pairs of proteins, following rules that allow us to handle a multitude of formalisms used by different references to represent pathways within the BioPAX L3 format.

Not all reference pathway maps use the same types of identifiers. We therefore establish a name correspondence between the protein interaction network and all pathway maps by using protein-gene name mappings and gene synonyms. In order to find entities corresponding to a protein of interest, we query pathway maps with the UniProt protein name and identifier, as well as the corresponding gene name and Entrez gene synonyms retrieved from NCBI at [ftp://ftp.ncbi.nih.gov/gene/DATA/GENE\\_INFO/Mammalia/Homo\\_sapiens.gene\\_info.gz](ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz). This can introduce ambiguity in the analysis because some pathway databases use non-standard gene symbols or a single symbol to represent multiple genes in the same family. For example, ABL2 is a member of KEGG's ErbB signaling pathway, but we did not annotate it as a known EGFR pathway member in Figure S4 due to these gene naming differences. Rather than manually adjust the gene name matching, we use a fully automated process in order to make unbiased claims about pathway converge in Figures S1 and S2.

### ***Extracting proteins, complexes, and paths between them from BioPAX L3***

In the first step, we compute all paths from a protein or complex to another protein or complex without going through an intermediate protein, complex or small molecule. We explicitly handle different representations of interactions including catalysis, control, template reaction regulation, and complex assembly.

A *catalysis* has a controller entity and a controlled conversion. It may optionally have a catalysis direction, which indicates whether the conversion is catalyzed left-to-right or right-to-left. We say that when an entity E is a controller of a conversion C, then the entities at one or both sides of the conversion are reachable from E (depending on the catalysis direction information).

A more general interaction type than catalysis, a *control* has a controller entity and a controlled interaction and does not specify a direction for the controlled interaction. We say that when entity E controls interaction I, both the left and right participants of I can be reached from E.

A *template reaction regulation* expresses the regulation of a template reaction by a physical entity. Template reactions only have a product, which we say is reachable by the entity controlling the regulation.

A *complex assembly* has left and right participants. If P is a left participant, then a right participant complex is reachable from P. We do not use this type of path currently, with a protein start node and complex end node. We instead just look for common complexes that contain two given proteins.

Additionally, we use the reported *molecular interactions* to infer complex assemblies. We say that entities E<sub>1</sub> and E<sub>2</sub> appear in a complex when they are participants in the same molecular interaction.

### ***Computing protein interactions from extracted paths***

Now that we have information about which complexes and which protein/complex paths exist in a resource, we can query the extracted paths for reachability from one protein to another.

Interactions between proteins may be expressed either between the entities directly or indirectly through successive steps of complex formation. We want to use the former when it is available because it usually provides information on interaction directions unambiguously. Considering paths that have complexes at their endpoints, or even simply analyzing which proteins appear in the same complex is essential for mining some interactions, but only considering the more specific, direct interaction information results in a better analysis. For this reason, we perform a staged reachability query with three different types of queries.

In the first stage, we look for paths that have A and B at each end. If such a path can be found for at least one direction, we stop the process and report the direction(s). If an interaction cannot be found in the first stage between A and B, we advance to a second stage of querying, in which we search for complex-to-protein interactions from, for instance, C to B, where A is a member of the complex C. Note that A might be transitively contained in C through another complex contained by C. We further refine this interaction by enforcing that A must have joined complex C no earlier than other non-complex components. If we can find complex assemblies that produce C, we

check whether A is a direct participant of the complex assembly (as opposed to having earlier formed a complex that is a participant in the complex C). If no complex assemblies can be found, we check whether A is a direct participant of the complex C (as opposed to being a participant in a complex that is itself a participant of C).

If neither of the above stages result in inferring a direction, we perform a third and final query, looking for a common complex that contains both A and B. If there is such a complex, we infer an interaction in both directions.

### ***Pathway resources***

We applied our data extraction procedure to eight human pathway references, which are shown with their full names and identifiers when available:

- NCI-Nature Pathway Interaction Database "EGF receptor (ErbB1) signaling pathway" (erbb1\_receptor\_proximal\_pathway): <http://pid.nci.nih.gov/download.shtml>
- Reactome "Signaling by EGFR" (R-HSA-177929): <http://www.reactome.org/ReactomeGWT/entrypoint.html>
- BioCarta "EGF signaling pathway" (h\_egfPathway): <http://pid.nci.nih.gov/download.shtml>
- Cancer Cell Map "EGFR1": [http://www.pathwaycommons.org/pc/dbSnapshot.do?snapshot\\_id=8](http://www.pathwaycommons.org/pc/dbSnapshot.do?snapshot_id=8)
- KEGG "MAPK signaling pathway" (hsa04010) and "ErbB signaling pathway" (hsa04012): <http://ipavs.cidms.org/downloads>
- Layek et al.: Manually reconstructed from Figure 4 of (Layek et al., 2011)
- *Science Signaling* Database of Cell Signaling "Epidermal Growth Factor Receptor Pathway" (CMP\_14987): Originally manually reconstructed from the image at [http://stke.sciencemag.org/cgi/cm/stkecm;CMP\\_14987](http://stke.sciencemag.org/cgi/cm/stkecm;CMP_14987), which was archived in June 2015 and is no longer available. The XML version is available from <http://stke.sciencemag.org/about/help/cm>.

This collection of reference pathways includes: EGFR pathway maps from six popular databases (Croft et al., 2014; Gough, 2002; Kandasamy et al., 2010; Kanehisa et al., 2012; Nishimura, 2001; Schaefer et al., 2009); a Boolean circuit representation of growth factor signaling (Layek et al., 2011); and the related but more general mitogen-activated protein kinase (MAPK) pathway from KEGG. These resources reflect the diverse goals and biases of different pathway curators. BioCarta focuses on the most essential signaling events, containing only 16 proteins. Conversely, Cancer Cell Map, which is part of the NetPath resource (Kandasamy et al., 2010), seeks broader coverage. Its EGFR map contains 178 proteins, approaching the 202 proteins cataloged in a thorough EGFR review (Oda et al., 2005).

For the yeast evaluation, we downloaded the KEGG pathway "MAPK signaling pathway" (sce04011) from [https://www.genome.jp/kegg-bin/show\\_pathway?map=sce04011](https://www.genome.jp/kegg-bin/show_pathway?map=sce04011). We used capDSD (Cao et al., 2014) to parse the XML file and Cytoscape (Shannon et al., 2003) to retain only the edges in the "High osmolarity" connected component of the KEGG pathway. Finally, we exported the Cytoscape network into a SIF format for the evaluation. The SIF file was a multigraph, containing more than one edge for some protein pairs. We retained the most specific direction and sign available for each protein pair.

In addition, we considered an upper and lower bound when calculating the directed precision (Supplemental File 4) because not all KEGG pathway edges have reported directions. Both cases consider edges that are in the KEGG pathway (directed or undirected) and in the TPS network with a specific predicted direction as the positive predictions. For the upper bound on directed precision, the true positives are the edges for which the predicted direction does not conflict with the reported direction in KEGG. For the lower bound, the true positives are the edges for which KEGG reports a unique direction that agrees with the TPS prediction.

### **Natural language processing evaluation**

We used three NLP tools – iHOP (Hoffmann and Valencia, 2004), Chilibot (Chen and Sharp, 2004), and Literome (Poon et al., 2014) – to perform unbiased literature searches for previous evidence of direct interaction for the 54 fringe pathway predictions. Fringe predictions are interactions that are at the periphery of the canonical pathway, meaning one or both of the interacting proteins are members of a reference EGFR pathway map but there is no interaction reported. Each NLP tool takes gene symbols as input so we mapped the UniProt identifiers in our predicted pathway edges to gene symbols. For each pair of genes, the NLP tools report the sentence from a PubMed abstract that is believed to describe an interaction among the corresponding proteins and either the full abstract or

the PubMed Unique Identifier (PMID), which can be used to retrieve the full abstract. All three tools perform entity normalization to automatically reconcile synonyms that refer to the same gene or protein. We did not intend to evaluate the quality of the NLP itself so we manually read the entire abstract to select the sentence that most specifically confirms or denies the predicted interaction, which is not always the sentence identified by the NLP tool. Literome searches for direct interactions and indirect interactions that are mediated by an intermediate third protein, and we only considered the direct interactions. For Chilibot, we considered interactive sentences but not parallel sentences. We classified the degree of support for each interaction using the categories in Supplemental File 3 and summarized the evidence by taking the most specific sentence across the three tools. When evaluating interaction sign, we assessed the interaction's impact on the target protein's phosphorylation, not its function. Because some interactions may have conflicting roles in different contexts and abstracts and each NLP tool has unique strengths, we kept the sentence that best matched our prediction when there were multiple possibilities or conflicts.

### **Osmotic stress evaluation**

Predicted, direct Rck2 targets in the TPS network were compared against a recent phosphoproteomics study on an *RCK2* mutant strain that identified peptides with defective phosphorylation 5 minutes after treatment with 0.5M NaCl (Romanov et al., 2017). We defined defective phosphosites as those that exhibited a wild type osmotic stress response and at least a 1.5-fold defect in *rck2Δ* compared to the wild type. Phosphosites with a wild type osmotic stress response included those with at least a 2-fold increase in phosphorylation in the wild type strain after NaCl treatment in the Romanov et al. study and those that were not quantified in that condition by Romanov et al. but responded significantly to salt stress in the Kanshin et al. study (Kanshin et al., 2015). Cdc28 targets were merged from two *cdc28-as* studies (Holt et al., 2009; Kanshin et al., 2017) using the substrate identification techniques they proposed. Holt et al. required phosphopeptides to match the Cdc28 consensus sequence and a 50% decrease in phosphorylation. Kanshin et al. trained a support vector machine to identify direct Cdc28 substrates. For both Rck2 and Cdc28, we only considered defective phosphosites that were quantified by Kanshin et al. (Kanshin et al., 2015) because those phosphorylated proteins were provided as input to TPS.

### **Running the dynamic Bayesian network, TimeXNet, and FunChisq**

We downloaded the dynamic Bayesian network software (Hill et al., 2012) from <http://mukherjeelab.nki.nl/DBN.html>. We constructed a network prior that indicates, for each pair of peptides in the time series data, whether there exists an edge corresponding to the pair in the protein-protein interaction network we prepared for PCSF. We ran the method with a maximum in-degree of 2 nodes, which is the largest value that the DBN could handle for this dataset. The method chose an optimal prior strength of 1.5 for inference. Edges were generally assigned low probability values across the network with a maximum edge probability of 0.10967. To obtain a final network of size comparable to the TPS output, we filtered the DBN predictions by a minimum probability value of 0.025. We treated opposite directed predictions for the same pair of proteins as one undirected edge in our comparative evaluations.

We downloaded the TimeXNet software (Patil and Nakai, 2014; Patil et al., 2013) from <http://timexnet.hgc.jp/> and ran it with the same weighted protein-protein interaction network we prepared for PCSF. To assign proteins to temporal groups, we selected the time point at which the most significant phosphorylation change occurs, considering all peptides that correspond to a protein. The initial temporal group contained proteins with the most significant change at 2 or 4 minutes. The intermediate group contained the 8, 16, and 32 minute proteins. The 64 and 128 minute proteins composed the late group. As in PCSF, we computed node scores as  $-\log_{10} pvalue$ . Following the parameter selection guidelines in (Patil et al., 2013), we ran TimeXNet using all combinations of  $\gamma_1$  and  $\gamma_2$  from 0 to 5 (step size of 0.5). There were no subnetworks with less than 1% unreliable edges when defining unreliable edges as edges with weight less than 0.5. We instead defined unreliable edges as those with edge weight less than 0.3 and selected  $\gamma_1 = 4.5$  and  $\gamma_2 = 0$  as the optimal parameters. Among the 11 subnetworks with less than 5% unreliable edges, these parameters maximized the number of source nodes and minimized the fraction of unreliable edges. We did not filter the subnetwork by node or edge flow and treated all edges as directed edges based on the flow direction.

We downloaded FunChisq (Zhang and Song, 2013) and the associated data discretization tool Ckmeans.1d.dp as R packages. We wrote a wrapper in Scala that uses the two packages to provide an end-to-end inference pipeline that accepts time series data and produces FunChisq scores (a statistic and a p-value) for every interaction. The wrapper is available publicly at <https://github.com/koksal/funchisq-wrapper>. We discarded time series profiles that were



discretized into only one level by Ckmeans.1d.dp. We mapped the peptide-level output of FunChisq to the protein level, which produced duplicate protein-level edges when multiple peptides mapped to the same protein. We aggregated these duplicate edges by taking the minimum p-value observed for each edge. We filtered the edges by keeping those with p-value  $\leq 0.01$  and limited the network to 1000 edges by descending FunChisq statistic to be comparable in size to the other algorithms. As for the dynamic Bayesian networks, we treated opposite directed predictions for the same pair of proteins as one undirected edge in our comparative evaluations.

### **Contrasting TPS with related computational approaches**

Approaches for building networks from gene expression data alone (reviewed in (De Smet and Marchal, 2010)) can be applied to phosphoproteomic data as well. Extensions of these methods for temporal data introduce time lags and search for dependencies between genes' expression levels over time (Zoppoli et al., 2010). Methods based on Granger causality (Masnadi-Shirazi et al., 2014) identify proteins whose phosphorylation predicts behavior at later time points and provide one type of causal model. However, as we showed in our comparison with the dynamic Bayesian network (Hill et al., 2012) and FunChisq (Zhang and Song, 2013), these and other methods that rely on the phosphorylation data alone (Henriques et al., 2017) miss critical signaling pathway interactions because not all pathway members have observed phosphorylation changes.

Algorithms based on gene and protein perturbations provide an alternative approach toward causal models. Transcriptional regulatory networks have been inferred from expression changes induced by gene knockouts and knockdowns (Anchang et al., 2009; Markowitz et al., 2007; Wang et al., 2014; Yeang et al., 2004). Likewise, signaling networks have been reconstructed by stimulating a pathway and perturbing signaling nodes with kinase inhibitors or RNA interference. Protein activities are observed with antibody-based assays, and pathways are recovered *de novo* (Ciaccio et al., 2015; Fröhlich et al., 2009; Kiani and Kaderali, 2014; Molinelli et al., 2013) or by adapting prior pathway knowledge (Morris et al., 2011). The PHONEMeS method is unique for its ability to handle large-scale phosphoproteomic perturbation data (Terfve et al., 2015).

The HPN-DREAM network inference challenge (Hill et al., 2016) spawned several new approaches for analyzing time series phosphorylation data. Participants predicted signaling pathways from *in silico* time series data and temporal reverse phase protein array data for approximately 45 phosphoproteins in four breast cancer cell lines under various stimuli and inhibitor treatments. In contrast, TPS focuses on the unique demands in modeling and scaling to global phosphoproteomic data with over one hundred thousand phosphosites that are not encountered when modeling only tens of proteins. PropheticGranger (Carlin et al., 2017), the top performer in the HPN-DREAM experimental task, demonstrated the importance of prior knowledge in network inference and modified the standard Granger causality approach to assess dependencies between observed proteins. However, PropheticGranger did not scale when we applied it to our EGF response data with a complete PPI network as the prior. FunChisq (Zhang and Song, 2013), the top performer in the HPN-DREAM *in silico* task, was able to run on our data but did not perform well (Supplemental File 4).

In our EGFR study, the TPS PPI subnetwork input is provided by PCSF, but other network algorithms can also connect phosphorylated proteins using PPI. A related algorithm interpolates between globally optimal (Steiner tree) and locally optimal (shortest path) connections to different proteins (Yosef et al., 2009), and this method has been applied to link functional signaling proteins derived from phosphoproteomics data (Rudolph et al., 2016). Many other approaches connect source and target proteins in a PPI network to identify pathways. ResponseNet (Yeger-Lotem et al., 2009) does so with a maximum flow formulation; SHORTEST (Silverbush and Sharan, 2014) and PathLinker (Ritz et al., 2016) use shortest paths; Maximum Edge Orientation (MEO) (Gitter et al., 2011) orients the undirected edges to produce short, directed paths. Integer programs can express complex optimization preferences with multi-stage objective functions when predicting source-target connections (Chasman et al., 2014; MacGilvray et al., 2018). The predicted networks from any of these methods can be used as input for temporal analysis with TPS.

Among methods that integrate dynamic data and PPI networks, TPS is unique in its ability to assess and summarize all possible pathway structures that are consistent with the input network and the temporal constraints. TPS also considers all possible temporal activations for each peptide instead of mapping proteins to temporal bins in advance like TimeXNet (Patil and Nakai, 2014; Patil et al., 2013). Similarly, Budak et al. use time point-specific PCSF networks to map proteins to times (Budak et al., 2015), TimePath assigns genes to transcriptional phases based on gene expression timing (Jain et al., 2016), Khodaverdian et al. explore theoretical properties of temporal Steiner trees (Khodaverdian et al., 2016), and ST-Steiner allows early subnetworks to influence the structure of subsequent

subnetworks (Norman and Cicek, 2018). The Signaling and Dynamic Regulatory Events Miner (SDREM) models temporal gene expression to infer the timing of transcription factor activity, but the pathway discovery phase does not use any temporal information (Gitter and Bar-Joseph, 2013; Gitter et al., 2013). Vinayagam et al. used temporal phosphorylation to evaluate their predicted PPI directions but did not consider dynamics when making the predictions (Vinayagam et al., 2011). Time series data and interaction networks have also been combined for inferring protein complex dynamics (Park and Bader, 2012), pathway enrichment (Jo et al., 2016), and related problems reviewed in Przytycka et al. (Przytycka et al., 2010).

The key difference between our work and other declarative computational approaches is that TPS operates on networks that are several orders of magnitude larger and summarizes very large solution spaces defined by sparser and less precise experimental data. Model checking and symbolic reasoning have been used to verify properties of manually constructed biological models (Fisher and Piterman, 2014), complete partially specified pathways using perturbation data (Köksal et al., 2013), and synthesize gene regulatory networks directly from data (Dunn et al., 2014; Moignard et al., 2015) (reviewed in (Fisher et al., 2014)). In addition, other types of declarative approaches, such as integer programming (Budak et al., 2015; Chasman et al., 2014; Jain et al., 2016; Ourfali et al., 2007; Sharan and Karp, 2013; Silverbush and Sharan, 2014) and answer set programming (Guziolowski et al., 2013), have been applied to biological pathway analysis. The TPS model summarization strategy, which makes it applicable to comprehensive signaling networks containing more than a hundred thousand edges and phosphosites, sets it apart from these related methods (Figure S6).

### **Random data generation for scalability analysis**

In order to evaluate the ability of TPS to scale to large datasets, we randomly generated graphs, time course phosphorylation profiles, and protein-peptide maps of increasing size, up to 128,000 edges. We tuned the random data generation to approximate the following characteristics of our main case study: median node degree, median number of phosphosites per node, and median number of time points that exhibit a statistically significant change from the baseline. Both the number of phosphosites per node and the number of significant changes per time course were sampled from Poisson distributions.

### **Visualization**

Network figures were created with Cytoscape version 3.2.0 (Shannon et al., 2003). Most other images were generated using the matplotlib (Hunter, 2007), matplotlib-venn, py-upset, and Seaborn Python libraries. The py-upset package (<https://github.com/ImSoErgodic/py-upset>) is a Python port of the UpSet technique (Lex et al., 2014).

## **Supplemental References**

Anchang, B., Sadeh, M.J., Jacob, J., Tresch, A., Vlad, M.O., Oefner, P.J., and Spang, R. (2009). Modeling the Temporal Interplay of Molecular Signaling and Gene Expression by Using Dynamic Nested Effects Models. *Proc. Natl. Acad. Sci.* *106*, 6447–6452.

Cao, M., Pietras, C.M., Feng, X., Doroschak, K.J., Schaffner, T., Park, J., Zhang, H., Cowen, L.J., and Hescott, B.J. (2014). New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics* *30*, i219–i227.

Ceol, A., Chatr-aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.* *38*, D532–D539.

De Smet, R., and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* *8*, 717–729.

Fisher, J., and Piterman, N. (2014). Model Checking in Biology. In *A Systems Theoretic Approach to Systems and Synthetic Biology I: Models and System Characterizations*, V.V. Kulkarni, G.-B. Stan, and K. Raman, eds. (Springer Netherlands), pp. 255–279.

Fröhlich, H., Sahin, Ö., Arlt, D., Bender, C., and Beißbarth, T. (2009). Deterministic Effects Propagation Networks for reconstructing protein signaling networks from multiple interventions. *BMC Bioinformatics* *10*, 322.

- Gitter, A., Carmi, M., Barkai, N., and Bar-Joseph, Z. (2013). Linking the signaling cascades and dynamic regulatory networks controlling stress responses. *Genome Res.* 23, 365–376.
- Henriques, D., Villaverde, A.F., Rocha, M., Saez-Rodriguez, J., and Banga, J.R. (2017). Data-driven reverse engineering of signaling pathways using ensembles of dynamic models. *PLOS Comput. Biol.* 13, e1005379.
- Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* 9, 90–95.
- Iorio, F., Bernardo-Faura, M., Gobbi, A., Cokelaer, T., Jurman, G., and Saez-Rodriguez, J. (2016). Efficient randomization of biological networks while preserving functional characterization of individual nodes. *BMC Bioinformatics* 17, 542.
- Jo, K., Jung, I., Moon, J.H., and Kim, S. (2016). Influence maximization in time bounded network identifies transcription factors regulating perturbed pathways. *Bioinformatics* 32, i128–i136.
- Khodaverdian, A., Weitz, B., Wu, J., and Yosef, N. (2016). Steiner Network Problems on Temporal Graphs. [arXiv:1609.04918 \[cs\]](https://arxiv.org/abs/1609.04918).
- Kiani, N.A., and Kaderali, L. (2014). Dynamic probabilistic threshold networks to infer signaling pathways from time-course perturbation data. *BMC Bioinformatics* 15, 250.
- Krishnaswamy, S., Spitzer, M.H., Mingueneau, M., Bendall, S.C., Litvin, O., Stone, E., Pe'er, D., and Nolan, G.P. (2014). Conditional density-based analysis of T cell signaling in single-cell data. *Science* 346, 1250689.
- Leek, J.T., Mosen, E., Dabney, A.R., and Storey, J.D. (2006). EDGE: extraction and analysis of differential gene expression. *Bioinformatics* 22, 507–508.
- Lees, J.G., Heriche, J.K., Morilla, I., Ranea, J.A., and Orengo, C.A. (2011). Systematic computational prediction of protein interaction networks. *Phys. Biol.* 8, 35008.
- Lex, A., Gehlenborg, N., Strobel, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput. Graph.* 20, 1983–1992.
- Markowitz, F., Kostka, D., Troyanskaya, O.G., and Spang, R. (2007). Nested Effects Models for High-Dimensional Phenotyping Screens. *Bioinformatics* 23, i305–i312.
- Masnadi-Shirazi, M., Maurya, M.R., and Subramaniam, S. (2014). Time-Varying Causal Inference From Phosphoproteomic Measurements in Macrophage Cells. *IEEE Trans. Biomed. Circuits Syst.* 8, 74–86.
- Morris, M.K., Saez-Rodriguez, J., Clarke, D.C., Sorger, P.K., and Lauffenburger, D.A. (2011). Training Signaling Pathway Maps to Biochemical Data with Constrained Fuzzy Logic: Quantitative Analysis of Liver Cell Responses to Inflammatory Stimuli. *PLoS Comput. Biol.* 7, e1001099.
- Mosca, R., Pons, T., Céol, A., Valencia, A., and Aloy, P. (2013). Towards a detailed atlas of protein–protein interactions. *Curr. Opin. Struct. Biol.* 23, 929–940.
- Oda, K., Matsuoka, Y., Funahashi, A., and Kitano, H. (2005). A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol. Syst. Biol.* 1, 2005.0010.
- Ourfali, O., Shlomi, T., Ideker, T., Rupp, E., and Sharan, R. (2007). SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics* 23, i359–i366.
- Park, Y., and Bader, J.S. (2012). How networks change with time. *Bioinformatics* 28, i40–i48.

- Patil, A., and Nakai, K. (2014). TimeXNet: Identifying active gene sub-networks using time-course gene expression profiles. *BMC Syst. Biol.* 8, S2.
- Przytycka, T.M., Singh, M., and Slonim, D.K. (2010). Toward the dynamic interactome: it's about time. *Brief. Bioinform.* 11, 15–29.
- Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R., et al. (2005). Global analysis of protein phosphorylation in yeast. *Nature* 438, 679–684.
- Ritz, A., Poirel, C.L., Tegge, A.N., Sharp, N., Simmons, K., Powell, A., Kale, S.D., and Murali, T. (2016). Pathways on demand: automated reconstruction of human signaling networks. *Npj Syst. Biol. Appl.* 2, 16002.
- Rudolph, J.D., Graauw, M. de, Water, B. van de, Geiger, T., and Sharan, R. (2016). Elucidation of Signaling Pathways from Large-Scale Phosphoproteomic Data Using Protein Interaction Networks. *Cell Syst.* 3, 585–593.e3.
- Silverbush, D., and Sharan, R. (2014). Network orientation via shortest paths. *Bioinformatics* 30, 1449–1455.
- Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* 100, 9440–9445.
- Villaveces, J.M., Jiménez, R.C., Porras, P., del-Toro, N., Duesbury, M., Dumousseau, M., Orchard, S., Choi, H., Ping, P., Zong, N.C., et al. (2015). Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database* 2015, bau131.
- Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., Assmus, H.E., Andrade-Navarro, M.A., and Wanker, E.E. (2011). A Directed Protein Interaction Network for Investigating Intracellular Signal Transduction. *Sci. Signal.* 4, rs8.
- Vizcaíno, J.A., Csordas, A., del-Toro, N., Dienes, J.A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., et al. (2016). 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* 44, D447–D456.
- Wagner, J.P., Wolf-Yadlin, A., Sevecka, M., Grenier, J.K., Root, D.E., Lauffenburger, D.A., and MacBeath, G. (2013). Receptor Tyrosine Kinases Fall into Distinct Classes Based on Their Inferred Signaling Networks. *Sci. Signal.* 6, ra58-ra58.
- Wang, X., Yuan, K., Hellmayr, C., Liu, W., and Markowetz, F. (2014). Reconstructing evolving signalling networks by hidden Markov nested effects models. *Ann. Appl. Stat.* 8, 448–480.
- Yeang, C.-H., Ideker, T., and Jaakkola, T. (2004). Physical Network Models. *J. Comput. Biol.* 11, 243–262.
- Yosef, N., Ungar, L., Zalckvar, E., Kimchi, A., Kupiec, M., Ruppín, E., and Sharan, R. (2009). Toward accurate reconstruction of functional protein networks. *Mol. Syst. Biol.* 5.
- Zhang, Y., Wolf-Yadlin, A., Ross, P.L., Pappin, D.J., Rush, J., Lauffenburger, D.A., and White, F.M. (2005). Time-resolved Mass Spectrometry of Tyrosine Phosphorylation Sites in the Epidermal Growth Factor Receptor Signaling Network Reveals Dynamic Modules. *Mol. Cell. Proteomics* 4, 1240–1250.
- Zoppoli, P., Morganello, S., and Ceccarelli, M. (2010). TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinformatics* 11, 154.