phuEGO: A Network-Based Method to Reconstruct Active Signaling Pathways From Phosphoproteomics Datasets

Authors

Girolamo Giudice, Haoqi Chen, Thodoris Koutsandreas, and Evangelia Petsalaki

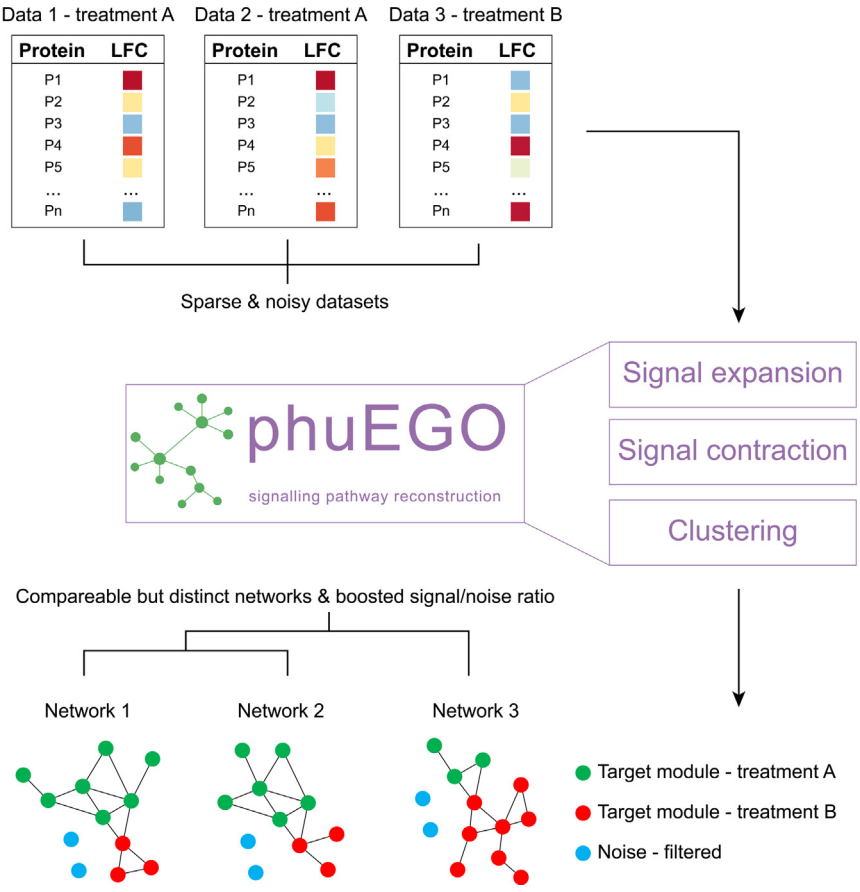
Correspondence

petsalaki@ebi.ac.uk

In Brief

We present phuEGO, a new tool that combines network propagation with ego network decomposition to provide interpretable active network signatures from phosphoproteomics datasets. We demonstrate that phuEGO boosts the signal-to-noise ratio from phosphoproteomics datasets allowing better identification of active signaling processes and improved comparisons and integration across studies. We applied phuEGO on the comparison of phosphoproteomics datasets acquired upon SARS-CoV2 infection after 24 h led to the identification of a network signature that is enriched in known targets for COVID-19.

Graphical Abstract



Highlights

- phuEGO extracts active signaling modules from phosphoproteomics datasets.
- phuEGO improves the signal-to-noise ratio and comparison of such data.
- Open-source package for integration in phosphoproteomics data analyses workflows.
- Active network signature of SARS-CoV-2 infection is enriched in known targets.

2024, Mol Cell Proteomics *23*(*6*), 100771

https://doi.org/10.1016/j.mcpro.2024.100771

phuEGO: A Network-Based Method to Reconstruct Active Signaling Pathways From Phosphoproteomics Datasets

Girolamo Giudice, Haoqi Chen, Thodoris Koutsandreas, and Evangelia Petsalaki*

Signaling networks are critical for virtually all cell functions. Our current knowledge of cell signaling has been summarized in signaling pathway databases, which, while useful, are highly biased toward well-studied processes, and do not capture context specific network wiring or pathway cross-talk. Mass spectrometry-based phosphoproteomics data can provide a more unbiased view of active cell signaling processes in a given context, however, it suffers from low signal-to-noise ratio and poor reproducibility across experiments. While progress in methods to extract active signaling signatures from such data has been made, there are still limitations with respect to balancing bias and interpretability. Here we present phuEGO, which combines up-to-three-layer network propagation with ego network decomposition to provide small networks comprising active functional signaling modules. PhuEGO boosts the signal-to-noise ratio from global phosphoproteomics datasets, enriches the resulting networks for functional phosphosites and allows the improved comparison and integration across datasets. We applied phuEGO to five phosphoproteomics data sets from cell lines collected upon infection with SARS CoV2. PhuEGO was better able to identify common active functions across datasets and to point to a subnetwork enriched for known COVID-19 targets. Overall, phuEGO provides a flexible tool to the community for the improved functional interpretation of global phosphoproteomics datasets.

Signaling pathways regulate the cell's response to external perturbations and modulate some of the most important biological processes such as cell growth, differentiation, and migration (1–3). They function through complex networks with multiple cross-talks with other pathways (4–7) and are highly context-specific; that is, signaling through the same pathway may result in completely different outputs depending on conditions, perturbations, or cell types (8–10). Current pathway annotations as they exist in publicly available databases do not capture this complexity and in addition are highly biased towards well-studied parts of the human signaling network (11, 12).

Mass spectrometry-based technologies allow us to capture in a relatively unbiased way the phosphorylation-based signaling state of a cell, through global phosphoproteomics experiments. This opens the door to data-driven extraction of condition-specific signaling networks that more accurately represent the cell's response than existing annotated pathways.

A limitation associated with using phosphoproteomics experiments is that they are intrinsically noisy, sparse, and lack reproducibility at the peptide level (13–18). The noise can be due to technical reasons (*e.g.* several steps needed for enrichment) but also due to biological reasons, as it is known that not all phosphosites are functional (19). The low abundance of phosphorylated peptides compared to total peptides in the cell is also another source of technical noise and also leads to sparse datasets, and reduced reproducibility at the peptide level compared to other omics modalities. Thus, there is a need for computational approaches that can effectively extract the active network signatures from these datasets.

One class of such methods employs network inference-based techniques, to extract a subnetwork able to explain how the phosphorylation signals propagate (20–23). Bayesian and logic models (24–27), ordinary differential equations (28, 29), linear and nonlinear regression (30) and methods considering pairwise scores based on correlation (31, 32), information theory (33) and others (34, 35), have been developed for inferring causal relationships. The HPN-DREAM network inference challenge (22) found that the best methods typically took advantage of prior knowledge signaling pathways. This means, however, that the results from such methods often suffer from literature bias. This was evident in the inference of cell line-specific edges part of the challenge, where methods tended to perform better in cell lines that better agreed with prior knowledge networks. This bias is mitigated by approaches that combine in-depth

From the European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridgeshire, United Kingdom

*For correspondence: Evangelia Petsalaki, petsalaki@ebi.ac.uk.

https://doi.org/10.1016/j.mcpro.2024.100771

large-scale phosphoproteomics data collection across multiple perturbations and time points with signaling network inference, at the cost, however of requiring extensive context-specific datasets (36, 37).

Another class of algorithms, such as KSTAR (38), KSEA (39), IKAP (40), KinasePA (41), and KEA (42), identify active kinases based on the phosphorylation levels of their substrates. However, these methods typically require a knowledge of site-specific kinase-substrate interactions, which is available only for a small number of well-studied sites. The exception is KSTAR which also accepts predicted kinase-substrate relationships. RoKAI is another interesting method that utilizes functional associations of putative kinase substrates to improve kinase activity prediction (43). PHOTON (44) circumvents these limitations, by integrating a set of significantly functional proteins into a protein–protein interaction (PPI) network and inferring a functionality score that is independent of the fold change of protein phosphorylation. It then uses these to derive active signaling networks from the data. However, PHOTON relies on linking 'terminal' nodes, *i.e.*, the phosphorylated proteins, to a 'source', *i.e.*, the receptor that was stimulated in the experiment through the ANAT method (45). As such the results represent signaling downstream of the 'source' and neglect potential cross talk with other pathways or processes that might also be affected by the stimulus, but not directly linked to the 'source'.

Recently, PPI network-based methods accounting for the global structure of the network have emerged. Distance-based methods such as shortest path and network flow approaches are widely used (46–48). Although most of these methods are applied to transcriptomics data, they can be adapted for use on phosphorylation data. For example, PATHLINKER (49) employs a weighted PPI network and uses a heuristic to maximize the score of the shortest paths between a set of source and target nodes. Other types of distance-based methods such as the prize-collecting Steiner tree (PCST) algorithm (50–52) and the forest variant (PCSF) (53–55) are also used. For example, Tuncbag *et al* (55) employed the PCSF to predict multiple altered pathways in yeast from transcriptomic and proteomic data. As protein interaction networks are starting to be more systematic (56, 57), these approaches start to mitigate the literature bias issue of cell signaling studies. However, the major limitation of the distance-based methods is the assumption that the shortest paths are the most informative or most likely used paths, which may not always be the case (58).

Network propagation-based methods have been developed that boost the signal-to-noise ratio in omics datasets and predict active pathways (59). They have been employed to predict protein functions (60, 61), prioritize candidate disease genes (62–64), detect active modules (65–67), and stratify patients (68, 69). TieDIE (70) performs two propagation computations, from sources and targets, and combines the result rankings to retrieve an active subnetwork. Using this approach Drake *et al* (71) extracted patient-specific network modules and potential drug targets in prostate cancer. Propagation algorithms are a perfect fit for phosphoproteomics data, which tends to be sparse since they can fill the gap between missing values and at the same time reduce the intrinsic noise of such datasets. However, these methods do not explicitly model feedback loops, predict interaction directions, or prioritize the most likely phosphorylation regulators. Additionally, to our knowledge, they tackle the problem of signaling network reconstruction from a global perspective, but they do not consider the effect that a phosphorylated protein has on its direct functional neighbors leading to large and hard-to-interpret network signatures.

To tackle these issues, we present phuEGO, an algorithm for extracting active signaling network signatures from phosphoproteomics data. phuEGO combines a global propagation method with a local approach to extract interpretable signals from phosphoproteomics datasets and allows improved comparison and integration of datasets acquired by different groups albeit in similar conditions.

EXPERIMENTAL PROCEDURES

Datasets

We extracted the log-2 fold change of each phosphosite from the data available at http://phosfate.com (72). Each phosphosite is then associated with a functional score (where available) extracted from Ochoa *et al*, 2019 (19). Each phosphorylated protein can be associated with multiple phosphosites and then to multiple values. To associate a single LFC and functional score to each protein we partitioned each dataset into tyrosine kinases, other kinases, and phosphorylated substrates, and selected the maximum LFC value and functional score per protein, under the assumption that this could represent the functional effect on the neighbors of the protein. Increased and decreased phosphosite sets are treated separately, therefore a phosphorylated protein, exceeding those thresholds, could be present in both sets. To include as many of the modulated kinases measured in the dataset as possible, without keeping kinases that were not modulated at all, we kept phosphorylated tyrosines and all the other kinases with a functional score and log-2 fold change (LFC) greater than the 20th percentile, and we kept all the phosphorylated substrates exceeding both the 80th percentile of LFC and functional score. We opted for a percentile as opposed to a cutoff, so that we do not have very large differences in the number of phosphosites included in the analysis per each dataset, as the distributions of LFC varied greatly. Nonetheless the input is fully customizable and defined by the user, who can adjust it as they deem appropriate for their application. The numbers of total phosphosites in the dataset, functionally annotated or not, those with LFC>1 and number of nodes in resulting phuEGO network are shown for reference in Supplemental Fig. S1.

We also have extracted the LFC from the original publications, described in Supplemental Table S3, where the functional score was not available. We excluded the study of Salek *et al.* (73), since the data were deposited in a database that is no longer available.

The SARS-CoV2 datasets were extracted from the work of Higgins and colleagues (74). In total five datasets comprising 4 different cell types at 24 h post infection were extracted. The datasets comprise the following cell types: A549 (Higgins (74) and Stukalov (75)), Caco-2

human lung epithelial cells (Klann (76)), Vero E6 African Green Monkey kidney cells (Bouhaddou (77)), human induced pluripotent stem cell-derived alveolar epithelial type 2 cells (iAT2, Hekman (78)).

For our analysis, we selected the top 200 increased and decreased phosphorylated proteins. Our results, after selecting alternative numbers of input proteins produced largely similar results, with only small differences (Supplemental Fig. S2). This is a tunable parameter that can be considered by the user depending on their data and application.

Pre-processing of Networks and Datasets

To compile the base network that phuEGO uses for its analysis we did the following: First, we retrieved the entire human protein-protein interaction network from IntAct (79) (version: 4.2.17, last update May 2021). We also added kinase-kinase interactions and kinase-substrate interactions from PhosphoSitePlus (80) (version 6.5.9.3, last update May 2021), OmniPath (81) (last release May 2021) and SIGNOR 2.0 (82) (last release May 2021). Only proteins annotated in Swiss-Prot (83) and those with at least one Gene Ontology term (GO) (84) (last release April 2021) were retained. The resulting protein interaction network (PPI network) comprises 16,407 nodes and 238,035 edges (Supplemental Table S1). The inclusion of protein interactions that are not necessarily signaling-related and that have been collected in more unbiased ways, allows us to mitigate the bias of annotated pathway databases and provides flexibility to search for context-specific solutions during our network identification. Additionally, we modeled edge weights according to simGIC (85) semantic similarity. The Semantic Measures Library (86) was employed to calculate the semantic similarity among the three categories of GO (molecular function, biological process, and cellular component) by adding a virtual root connecting all of them. We also generated 1000 random networks using the configuration model available in the igraph library (method = vl). Briefly, the method (87) implements a Markov chain Monte Carlo algorithm to generate random networks where the node degrees are conserved. Since the edges in the random networks were reshuffled, new random interactions were created and therefore, the edge weights (*i.e.* simGIC semantic similarity) were updated accordingly. We applied the square (or Laplacian) normalization to correct for the hub bias (88). Briefly, the weight of each edge was divided by the square root of the weighted degree of the interacting nodes(1):

$$w_{ij} = \frac{w_{ij}}{\sqrt{d_i d_j}} \quad (1)$$

where w_{ij} indicates the edge weight (*i.e.* semantic similarity) and d_i and d_j represent the weighted degree of node i and node j respectively.

Additionally, we also precalculated the simGIC semantic similarity of each node in the PPI network against all the other nodes and calculated the mean and the standard deviation for each node of the PPI network. These values are used in the next steps of the method to filter the ego networks by calculating the z-score (see paragraph on ego decomposition below). The above-described network was the one used to generate the results in this article; however, the phuEGO package allows the user to input any network that they deem suitable for their application.

Network Propagation by Random-Walk-With-Restart

PhuEGO accepts as input a dataset of phosphorylated UniProtKB entries and the corresponding log-2 fold change (LFC). PhuEGO first assesses the prior input set of nodes, which we will call seeds in this manuscript, meaning the nodes in the PPI network from where random walkers should start. To do so, the input dataset is initially divided into positive and negative LFC. These two partitions are by default subsequently divided into (i) the tyrosine kinases, (ii) the rest of kinases, and (iii) the non-kinase phosphorylated proteins. However, phuEGO also provides the option to alternatively run all the seeds on one or two layers, partitioned as the user prefers. To assess which proteins will be assigned to each partition we retrieved all the human kinases (Clan CL0016) from the Pfam (89) database (Pfam ver 34.0 released in March 2021). Since we wanted to distinguish the tyrosine kinases from the rest of the kinases, we retrieved all the human tyrosine kinases associated with the Pfam domain (PF07714) from UniProtKB (83). In total, 531 kinases are present in our PPI network, of which 127 are tyrosine kinases (Supplemental Table S1).

Each of the partitions corresponds to different restart probability vectors, whose dimension is equal to the number of nodes in the PPI network and the restart probability values are equal to the LFC of the phosphorylated proteins, scaled between 0 and 1. Therefore, we start one distinct RWR (90) run for each partition, with each one involving different sets of prior nodes. As a result, we obtain one probability vector for each partition (three for the default settings), representing the most probable nodes from the perspective of the seed nodes. We recommend maintaining the three-layer partition for phosphoproteomics datasets. The idea behind this procedure is to capture signal propagation in a global manner having as central input nodes the phosphorylated proteins and integrating the signal from these with that from the kinases, as the drivers of cell signaling. Note that "signal" here means a "biologically meaningful data point measurement" and with this procedure, we aim to extract the nodes in the network that are most likely to be causing or be affected by this measurement. To filter out spurious nodes, we repeated the same procedure using the same seed nodes but against 1000 random networks, generated as described above. This allows us to evaluate the percent of its random scores that exceed the real score (*i.e.* the node's empirical *p*-value). At the end of this process only the nodes with a score greater than 95% of its random scores were maintained independently of which partition they have been assessed with. Note that this process is repeated two times, one for the upregulated phosphoproteins and one for the downregulated ones; consequently, two subnetworks are extracted associated with increased and decreased phosphorylation levels respectively. The seed nodes are always included in the subnetworks regardless of their RWR score. Where proteins are included in both the up and downregulated network, they are removed from the one with smaller mean RWR score for that protein from the 3 layers. This is the default option in phuEGO to provide more easy-to-interpret networks, but the user can disable this function and allow inclusion in both networks.

Generation of Functional Ego Networks

From the two subnetworks extracted previously, we extract ego networks as a subgraph centered on a seed/phosphorylated node and comprising all the overrepresented nodes in a 2-step distance from the ego. Since ego networks are still highly interconnected, in theory, they could have the same dimension as the subnetworks extracted from the initial random-walk-with-restart process. To select only those ego neighbors that are most functionally similar to the ego, we computed the z-score associated with each ego neighbor using the precomputed mean and the standard deviation of the simGIC score (see Preprocessing of networks and datasets paragraph) according to (2).

$$z\text{–}score = \frac{simGIC\ (ego, j) - mean_{simGIC\ ego}}{std_{simGIC\ ego}} \quad (2)$$

where the *simGIC(ego,j)* is the semantic similarity between the ego and node *j*, and the mean and std are the means and the

standard deviation of all the semantic similarities between the ego and all the other nodes of the PPI network. The nodes with z-score>1.64 represent the functional ego network since they are also the most similar in terms of semantic similarity to the GO terms in which the ego is involved (95% confidence, one-tail test). Let Γ(ego) represent the first order neighbors of the ego node and $\Gamma_{\Gamma(ego)}$ the second order neighbors of the ego network. The edge weights are updated according to (3):

$$W_{i,j} = \begin{cases} simGIC(i,j), \text{if } i = ego \text{ and } j = \Gamma_{ego} \text{ or } i = \Gamma_{ego} \text{ and } j = \Gamma_{\Gamma_{ego}} \\ \dfrac{simGIC(ego,i) + simGIC(ego,j)}{2}, \text{if } i = \Gamma_{ego} \text{ and } j = \Gamma_{ego} \text{ or } i = \Gamma_{\Gamma_{ego}} \text{ or } j = \Gamma_{\Gamma_{ego}} \end{cases} \quad (3)$$

The ego networks obtained are normalized to correct for hubs using the Laplacian normalization as in (1) (Supplemental Fig. S3, *A* and *B*).

Ego Decomposition

To understand which nodes are more closely related to the ego and, hence, involved in a similar process/pathway, we decomposed each ego network with a number of neighbors greater than 5 into two vectors, one representing the topological distance from the ego, and one the functional distance from the ego.

To calculate the topological proximity, each node of the ego network is the source of a second run of RWR with a damping factor equal to 0.85. The restart probability vector is filled with 0 except in the node under consideration which is equal to 1. To calculate the distance between the ego node and all the other nodes of the ego network, the following formula is used (4):

$$topological\ affinity = 1000 * log_2(2 - jsd(RWR_{ego}, RWR_j)) \quad (4)$$

where *jsd* refers to the Jensen-Shannon distance, representing the similarity between two probability distributions. The RWR_j refers to the RWR probability vector when one of the nodes of the ego network is selected as seed, the RWR_{ego} refers to the RWR probability vector when the ego is the seed node. Nodes with values close to 1 are considered topologically similar to the ego.

The functional vector is defined as the logarithm of the semantic similarity between the ego and any other nodes in the network (5).

$$functional\ distance = 1000 * log_2(1 + simGIC\ (ego, j)) \quad (5)$$

where *simGIC* represents the semantic similarity measure between the ego and the node *j*.

To identify the most similar nodes to the ego we employed the Kernel Density Estimation (KDE) using the Gaussian kernel, where each node is represented as a point in a 2D plane where the x-axis represents the topological affinity to the ego and the y-axis represents the functional similarity to the ego (Supplemental Fig. S3C). The bandwidth for the KDE is estimated using the Silvemann formula (91). KDE estimates the joint probability density function of the topological and semantic similarity vectors obtained at the previous step. We then calculated the joint cumulative distribution function and select only those nodes according to the following formula:

$$F_{XY}(x,y) = P(x \le X < 1, y \le Y < 1)) \quad (6)$$

where x and y are user-defined parameters. For this paper, we set these parameters to 0.85 or 0.9 depending on the application (see respective sections).

Defining the Supernode Network through Merging the ego Networks

Each ego node and the neighbors exceeding the user's selected probability threshold constitute a supernode, that is, a small cluster of proteins that are topologically and functionally related to the ego and therefore potentially affected by its phosphorylation. We then calculated the relationships between all the supernodes to generate the supernode network. To do so for each combination of two ego nodes, we extracted the subnetwork originated by the union of the nodes included in the supernode pair and normalized it according to (1). The two egos, if connected, represent the sources of a third RWR run with a damping factor equal to 0.85. We calculate the weight between supernodes using (7)

$$supernode\ weight = jsd(RWR_{egoA}, RWR_{egoB}) \quad (7)$$

where the RWR_{egoA} refers to the RWR probability vector when the ego_A is selected as seed, the RWR_{egoB} refers to the RWR probability vector when the ego_B is selected as seed node. Edge values close to 0 indicate a strong relationship between supernodes, meaning that they potentially share many neighbors. Note that the link between two supernodes is not necessarily associated with a physical interaction. We then applied the Leiden (92) algorithm to the supernodes network to extract functional modules. Note that the Leiden algorithm is only applied to all the connected components bigger or equal to 4 supernodes. The connected components containing less than 3 supernodes are considered as functional modules (Supplemental Fig. S3). Isolated supernodes are removed.

Evaluation Through Enrichment Analysis and Overlapping Coefficients

Enrichment analysis is a standard approach employed to determine if known biological functions or processes are over-represented (enriched) in a set of genes/proteins of interest. The enrichment analysis is based on Fisher's exact test which assumes that the data is hypergeometrically distributed. We used the nodes in a module as a foreground for the enrichment analysis while the human PPI network is used as the background. Additionally, the *p*-values obtained are Bonferroni corrected. Enrichment analysis can be performed against several databases such as GO (84), KEGG (93), Reactome (94), Bioplanet (95), DisGeNET (96).

In order to assess the similarity between modules and the reference pathways we employed the overlap coefficient or Szymkiewicz–Simpson coefficient (8).

$$Overlap\ coefficient\ (X,Y) = \frac{|X \cup Y|}{min(|X|, |Y|)} \quad (8)$$

where X and Y represent the two sets of proteins under consideration. We also measured the pairwise overlap distance (97) between the following KEGG reference pathways: Cell cycle, EGF, TCR, MAPK, VEGF, TGF, Insulin, and NGF, by employing this formula:

$$Overlap\ distance = 1 - Overlap\ coefficient(X, Y) \quad (9)$$

where X and Y represent the set of proteins involved in the respective reference pathways. Therefore, the distance between a pathway and itself is equal to 0. To explore whether there is a relationship between the similarity of modules in a dataset with the similarity of the respective perturbed pathways we did the following: First, we identified the modules with the best overlapping coefficient in the comparisons between datasets with similar stimulations, considering these as the predominant signal for that pathway. Therefore, each of these datasets had one module identified as representing its predominant signal and these were compared across datasets. For datasets where no such module was identified as there weren't other datasets available with a similar stimulus, we did the comparisons with all the modules and kept the one with the best overlapping coefficient. Since phuEGO extracts on average 4 modules from each dataset, modules comprising less than 10 proteins are discarded to avoid increasing the overlapping coefficient artificially with very small modules against very large ones. For the same reason, we discarded the overlapping coefficients between datasets from the same publication. Additionally, we selected the modules with the best overlapping coefficient regardless of the pathway they could be annotated with.

The performance of phuEGO was compared to the enrichments resulting from a) the seeds b) the network resulting from the initial RWR step and c) the Prize Collecting Steiner Forest algorithm (PCSF) from the omicsintegrator2 package (98). In brief, PCSF works by identifying an optimal forest in a network by maximizing the collected prizes and minimizing the edge costs. We performed a grid search for each dataset to fine-tune the parameters to select the best network from Omicsintegrator2. We selected the following parameter ranges $\omega = [0.25, 0.5, 0.75, 1]$, $\beta = [0.25, 0.5, 0.75, 1, 1.5, 2]$, $\gamma = [3, 3.5, 4, 4.5]$ and selected the network with the best objective function. In particular, ω regulates the number of selected outgoing edges from the root, β is a scaling factor of prizes, and γ controls the edge penalty on hubs.

Calculation of Kinase Activities Using the KSEA Package

We used the KSEA app (99) to extract the kinase-substrate links for the SARS-COV2 datasets (https://casecpb.shinyapps.io/ksea/) using the default parameters.

KSEA requires the phosphosite positions and the modified residues to run, but the latter was not available from Higgins *et al.* To solve this problem, we downloaded the protein sequences and assigned the modified residues accordingly. If the residue position in the protein sequence didn't correspond to a canonical one (serine, threonine, or tyrosine) we selected a window of −/+ 3 residues from the position assessed by Higgins *et al.* and assigned it to the one closer to the center of the window. For each experiment we extracted all the kinase-substrate links extracted by KSEA and assigned a value of 0 or 1 depending on whether they were present or absent in the corresponding experiment. We then calculate the Pearson's correlation coefficient for all the datasets for both increased and decreased phosphorylation.

Comparisons of SARS-CoV2 Networks and Seed Nodes

To compare the networks generated by phuEGO from Higgins *et al.* and the seeds, we assigned to each node in the corresponding network the average RWR values from each of the three partitions or 0 if not present. We performed the same procedure to compare the seed nodes alone with the exception that we employed the LFC values. Then we used the *hclust* package (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/hclust) to perform the hierarchical clustering and *dendsort* (https://cran.rstudio.com/web/packages/dendsort/index.html) to optimize the ordering of leaves in the dendrogram.

Enrichment of Known Targets in SARS-CoV2 Datasets

Known targets for COVID-19 were extracted from Open Targets (February 2023) using the query 'MONDO_0100096'. In total 390 drug targets were extracted, of which 365 were present in the network. Only the SARS-CoV2 networks with a damping factor equal to 0.85 and a KDE threshold $\geq$ 0.85 were selected (Supplemental Table S2). To generate the A549 SARS-CoV2 network we selected the nodes in common between the Higgins (74) and Stukalov (75) network. To assess the overlap between the network nodes and the known targets we used Fisher's exact test considering as background the entire PPI network used in the analysis.

Data Visualization

Plots were generated in Python v3.10 using the *seaborn* and *matplotlib* libraries. Cytoscape (v3.9.1) was used for visualizing networks. Enrichment maps were generated in R with the *clusterprofiler* and *enrichplot* packages. Hierarchical clustering of the SARS-CoV2 datasets was done in R with *pheatmap*.

Calculation of the Run Times for the Method

The workflow of PhuEGO incorporates different computational tasks, to extract the final network signatures and functional modules. In general, they could be classified into four main processes: 1. Loading of networks, 2. RWR for network propagation, 3. Ego-decomposition and interpretation and 4. Identification of modules. The execution time of PhuEGO, as well as that of the above processes, was calculated for different sizes of seed nodes. Overall, five random seed sets were constructed with 100, 150, 200, 150, and 300 proteins respectively. Each set included 30 tyrosine kinases, 30 non-tyrosine kinases, and the rest of seeds were retrieved from the pool of non-kinase proteins. Additionally, the seed nodes were randomly assigned as positively or negatively regulated (LFC values were defined as −1 or 1), creating two approximately equal size subsets. This stratification was adopted in order to run the analysis for all protein layers and for both directions of input signal. PhuEGO ran using the prpack implementation for the RWR and damping factor equal to 0.85. All the experiments were performed on a machine with 8 cores (Intel i7 @ 3GHz) and 16 GB RAM.

Experimental Design and Statistical Rationale

The datasets included in this study were selected as they formed a unified collection by Ochoa and colleagues and included a diverse array of stimulations including well-studied and less commonly studied pathways. They were therefore available both individually from the original studies and as a reanalyzed uniform set, and we could therefore compare the performance on both sets. There was also a functional score annotation available for most peptides included, allowing us to evaluate the ability of phuEGO to improve the signal-to-noise ratio. The nodes included in the final 'global' networks were

chosen as described above and the choice of 1000 random networks for generating the background distributions of the RWR scores was made to allow for a resolution of three decimal points in the calculations while keeping the computational time still reasonable. Network and hub normalizations and statistical tests were done according to common practice as described in the respective sections above.

RESULTS

A Method to Extract Signaling Modules From Phosphoproteomics Data

We developed phuEGO, an algorithm to reconstruct active signaling networks from phosphoproteomics data (Fig. 1). PhuEGO comprises two steps: a) an initial filtering of a global protein interaction network (PIN) compiled from the literature (IntAct (79), SIGNOR (82), PhosphositePlus (80) and OmniPath (81); see Experimental Procedures) to coarsely identify networks associated with increased and decreased phosphorylation using random-walk-with-restart and b) a step to extract the local effect of each differentially abundant phosphosite on its neighbourhood, from these larger networks.

Specifically, phuEGO first generates global networks as a result of random-walk-with-restart performed three times - from (i) tyrosine kinases, (ii) other kinases and (iii) substrates identified in the phosphoproteomics datasets (Experimental Procedures). This reflects our knowledge that kinases are the main drivers of phosphorylation-based signaling responses, with tyrosine kinases typically acting upstream of the global signaling response (100, 101). Upregulated and downregulated phosphosites are treated separately to uncover two networks associated with each class of phosphosites: an upregulated 'active' network and a downregulated one. This parameter is tunable by the user to provide input that takes into consideration, for example, phosphosites known to inactivate proteins.

These coarse networks comprise on average ~2,500 nodes (Fig. 1; Supplemental Fig. S5). To improve the interpretability of the phosphoproteomics datasets and extract more specific up/down-phosphorylated signaling modules phuEGO uses ego network decomposition to capture the functional and topological effect of the phosphosites identified in the datasets locally. Ego networks represent small subnetworks comprising all the nodes that are two steps away from the ego, which phuEGO further reduces by removing nodes that are not functionally similar to the ego (Experimental Procedures). By combining ego network embedding with kernel density estimation (KDE) phuEGO selects the nodes that are most similar to the ego thus generating supernodes, which are small networks comprising the ego and the functionally related neighbours. Then phuEGO generates the supernodes network where the edges weight represents the relationship between supernodes. The Leiden algorithm (92) is employed to partition the supernode network. This procedure generates modules (~3–4 on the datasets tested in this work; Supplemental Fig. S6*A*), comprising the ego and neighboring nodes (~25–50 nodes on average; Supplemental Fig. S6*B*) that are more functional and topologically similar to the ego and, therefore, are more likely to be relevant to the signal represented by the ego. Thus, given a phosphoproteomics dataset, phuEGO extracts interpretable signaling subnetworks, associated with increased and decreased phosphorylation. The full process takes approximately 30 min to run, but if runtime speed is critical for a user's application, this can be reduced to half if the user opts for 500 random networks as the background to identify significantly propagated nodes, and every layer of propagation removed also cuts run time by a third (Supplemental Fig. S7).

phuEGO Boosts the Signal-To-Noise Ratio

As a first step in validating whether our method is indeed able to boost active signals from phosphoproteomics datasets, we evaluated in 46 datasets (Supplemental Table S3),

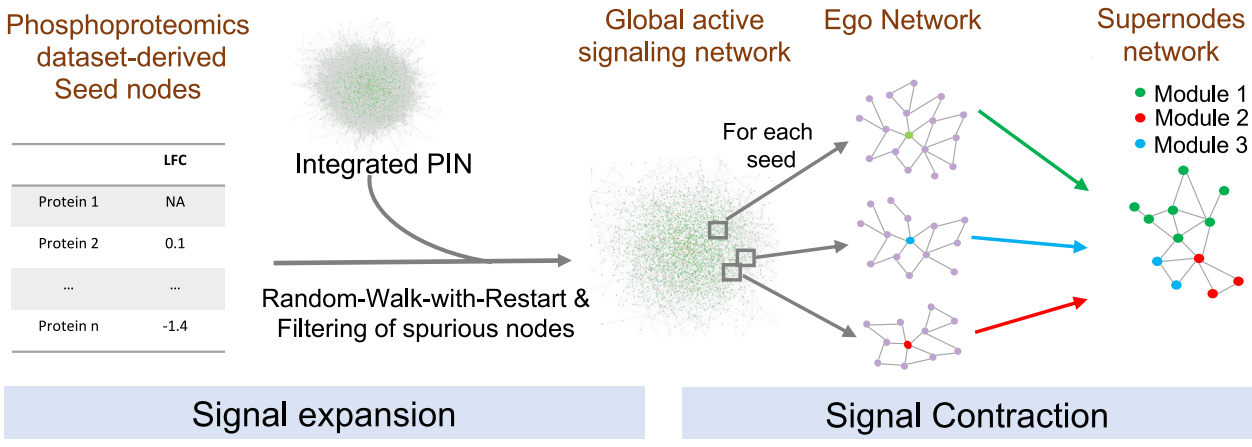


FIG. 1. **Overview of phuEGO's methodology.** PhuEGO first starts by performing a (up-to) three-layer random-walk-with restart on an integrated protein interaction network. It then re-maps the seed nodes as 'egos' and identifies a local network or module that comprises nodes that are most topologically and functionally similar to the ego. Finally, by combining these modules phuEGO generates a network of supernodes that include overlapping functional modules.

whether the phuEGO-extracted active networks were more enriched in the prior knowledge pathways that are expected based on the stimuli (93), than the raw upregulated phosphoproteins in the datasets. We also compared the enrichment ranking to using the RWR alone (Signal expansion stage; Fig. 1) as this would be equivalent to other methods that use network propagation to boost functional signal from omics datasets (44, 70). To our knowledge, there are no other methods that can serve a similar function as phuEGO, with the exception of PHOTON (44), which we were unable to run as it appears to be no longer maintained. The Prize Collecting Steiner Forest algorithm (PCSF) is not a network propagation-based approach, but has been used successfully previously to identify network signatures from phosphoproteomics datasets (102), and we therefore included it in our performance comparison.

We considered pathways as 'more enriched' when they ranked at a higher percentile of the total pathways found (p value<0.05, Bonferroni corrected; Experimental Procedures; Supplemental Data S1) in the phuEGO networks compared to the raw set of differentially abundant phosphosites (seeds). Overall, phuEGO boosts the ranking of the expected pathway for all datasets (Fig. 2*A*). Where the signal is already well-defined in the seeds, it maintains the high ranking and doesn't introduce further noise to dilute it through the diffusion process. Impressively, it can identify and rank highly the correct pathways even in datasets where the signal was initially very weak (*e.g.* Olsen *et al*, 2010, 150 and 180 min) or not present at all among the seed nodes (*e.g.* Olsen *et al*, 2010, 450 min; Fig. 2*A*; Supplemental Fig. S8*A*; Supplemental Table S3).

When comparing to the alternative approaches (RWR and PCSF) phuEGO generally performs better, ranking the relevant enrichment term higher or similar in all but one dataset from D'Souza *et al* 2014 (in two out of the three time points), whereas even for the seeds and PCSF that performed better the ranking was very low (Fig. 2*A*).

One of the main aims of our algorithm is to decrease the intrinsic noise of the phosphorylation datasets and improve their ability to identify the active signaling responses reproducibly. We thus evaluated whether phosphoproteomics experiments treated with the same conditions were more similar to each other before or after phuEGO was applied.

We first computed the overlapping coefficient between the seed nodes and the respective target pathway as the baseline. Each of the clusters that phuEGO identifies represents a unit of signaling, similar to a pathway. We do not expect all cell lines/types to have the exact same global response to the same stimulus, but we do expect at least one of these signaling modules to be similar. We thus extracted the corresponding modules with the highest overlapping coefficient between datasets produced by stimulating the same pathway and compared these to the modules identified from the rest of the datasets (Experimental Procedures). Overall, we found that modules extracted from similarly treated datasets tend to have a higher overlapping coefficient than those that didn't (Fig. 2*B*). Moreover, the similarity of these modules should be roughly analogous to the similarity of the prior knowledge pathways that we expect to be activated with the given stimuli (Experimental Procedures). We found that before phuEGO there is no relationship between the overlap of the phosphoproteins and the similarity of the prior knowledge pathways that we expect to be activated (r^2 = 0.05, p-value << 0.0001). This is also true for the RWR (r^2 = 0.12, p-value << 0.0001) and PCSF (r^2 = 0.05, p-value << 0.0001) approaches (Supplemental Fig. S8*B*). Conversely the dominant modules identified by phuEGO have an overlapping coefficient that is correlated to that of the respective prior knowledge pathways (Fig. 2*B*; r^2 = 0.4 p value << 0.0001). This is true both using the full collection of datasets as reprocessed by Ochoa *et al* (19) and when using the data from the original publications (Supplemental Fig. S6*C*).

To assess whether phuEGO indeed can reduce the inherent noise of phosphoproteomics datasets we evaluated whether phosphosites that survived the process and remained as part of an integrated active signal, *i.e.* supernodes, had a higher functional score compared to those that remained isolated and were therefore filtered out. The functional score was extracted from Ochoa *et al*, 2019 (19) and ranges from 0.0 to 1.0 with higher values representing an increased likelihood that the phosphosite will have a regulatory function on the protein that carries it. Across the 46 datasets (Supplemental Table S3) we found that phuEGO supernodes that remained as part of the active signaling signature were indeed significantly more functional than those that were filtered out (Fig. 2*C*; Mann-Whitney-U p value (damping = 0.5) = 4.8e-9, Mann-Whitney U p value (damping = 0.7) = 5.5e-10, Mann–Whitney U p value (damping = 0.85) = 6.3e-12). Therefore, phuEGO can filter out phosphosites that are less likely to be functional and thus represent noise in the dataset.

Together these analyses demonstrate how phuEGO is able to boost the active signal while reducing the noise in global phosphoproteomics datasets.

phuEGO can Distil the Active Signaling Networks From Diverse Phosphoproteomics Studies of SARS-CoV-2 Infection

As a case study, we compared 5 phosphoproteomics datasets compiled from the literature by Higgins and colleagues (74) at 24 h post infection since this time point was common to all the datasets. The datasets are targeting different cell types: A549 (Higgins (74) and Stukalov (75)), Caco-2 human lung epithelial cells (Klann (76)), Vero E6 African Green Monkey kidney cells (Bouhaddou (77)), human induced pluripotent stem cell-derived alveolar epithelial type 2 cells (iAT2, Hekman (78)). We found that the agreement of increased and decreased phosphorylation abundance LFC (Fig. 3*A*, Supplemental Fig. S9*A*) was very low between

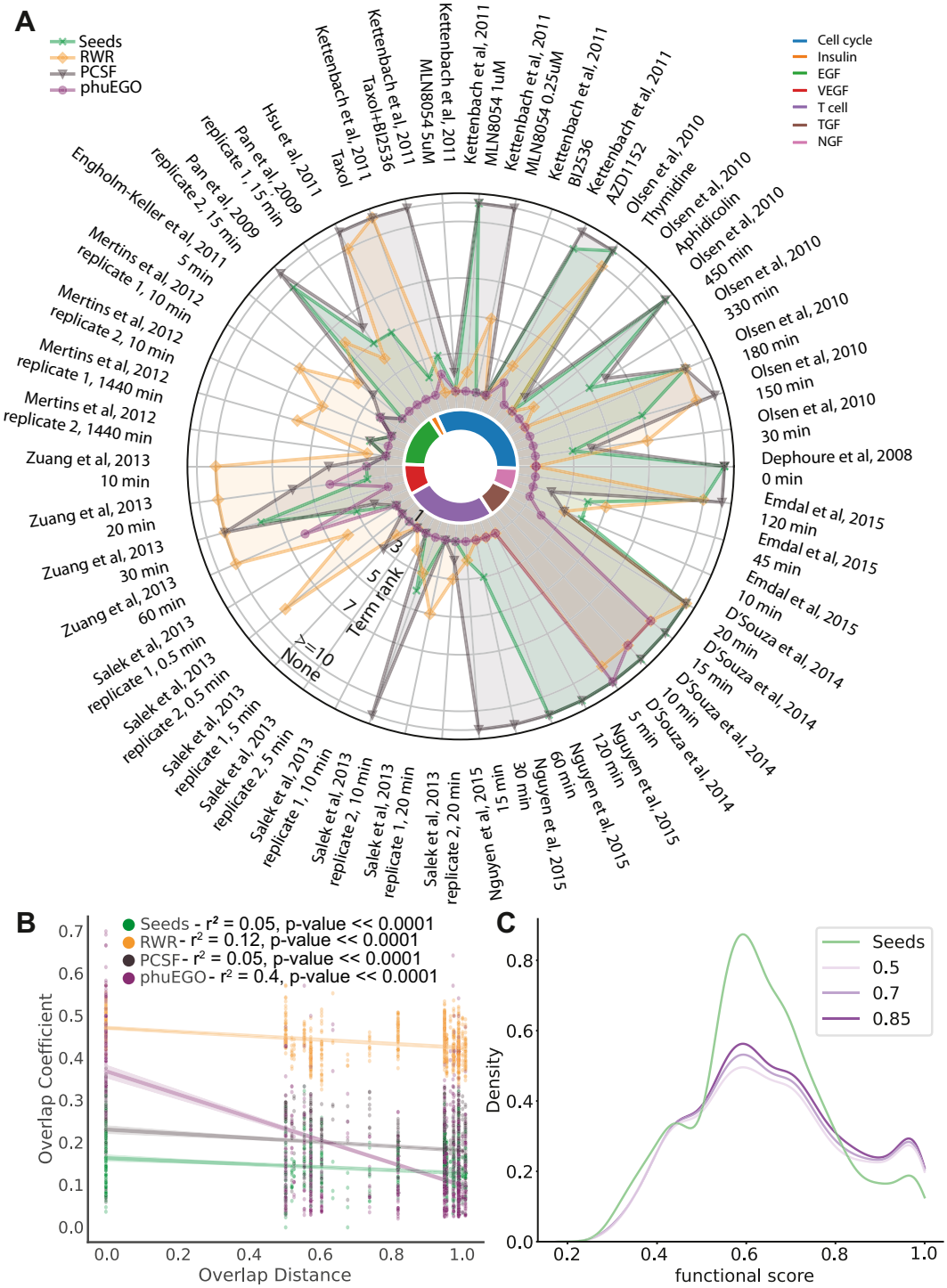


FIG. 2. **Evaluation of phuEGO.** *A*, comparison of seeds, PCSF, RWR and phuEGO with respect to their ability to rank highly the expected dominant signal, as defined by the stimulation used in the relevant dataset, using pathway enrichment analysis. The centre of the circle indicates the relevant pathway ranked first and the perimeter indicates a failure to identify the pathway at any rank. The ranking of the full list of the pathways identified in this enrichment analysis can be found in Supplemental Data S1. *B*, the maximum overlap coefficient of the phuEGO active signature is bigger for datasets that come from similar conditions *versus* those that do not. *C*, phosphosites/nodes retained by phuEGO tend to have a higher functional score indicating an improvement in the signal-to-noise ratio of the active signatures.

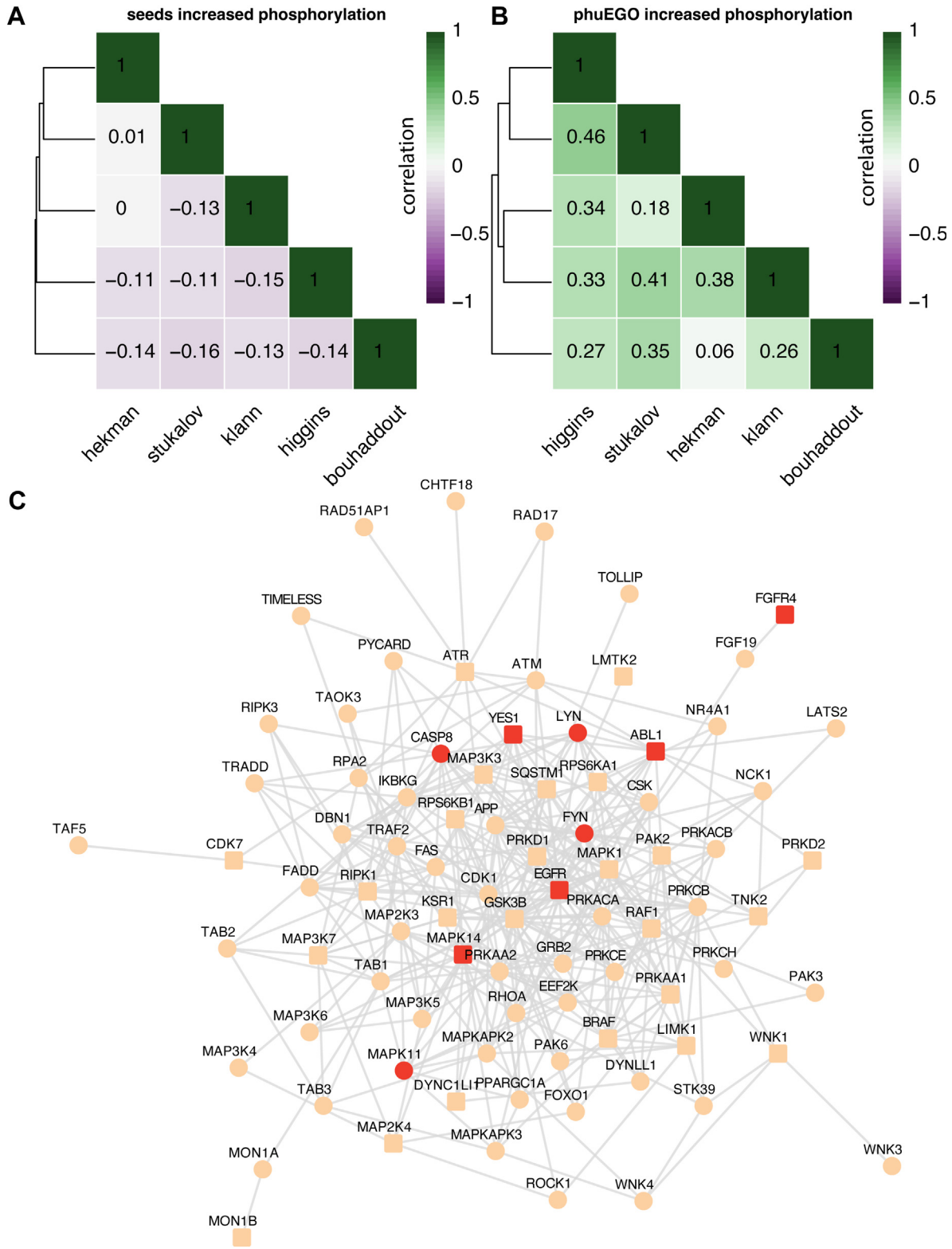


FIG. 3. **PhuEGO extracts active signatures of SARS-CoV-2.** *A*, public phosphoproteomics datasets of SARS-CoV2 infection correlate poorly. *B*, the correlation of public phosphoproteomics datasets upon SARS-CoV2 infection substantially improves after applying phuEGO. *C*, the intersection of phuEGO-derived networks is enriched in known targets for COVID-19.

datasets, as measured by Pearson correlation, even when comparing experiments done on the same cell type. When phuEGO is applied (Supplemental Data S2), the correlation increases and is even higher when comparing the same cell types (Fig. 3*B*, Supplemental Fig. S9*B*). PhuEGO shows improved correlation even when comparing to results from other approaches to improve the signal-to-noise ratio, such as KSEA ((99); Supplemental Fig. S8, *C* and *D*).

We hypothesized that if phuEGO is indeed extracting 'active' signaling signatures, intersecting the signatures across similar datasets, *i.e.* those from the same cell type (Higgins and Stukalov) would result in the enrichment of known targets for COVID-19. Indeed, the resulting network, which comprises 85 nodes and 362 edges included 9 known targets, which is 5-fold more than expected by chance (Fisher's Exact test *p*-value = 1.60e-04, Supplemental Table S2). These include SRC kinases LYN, FYN, and YES1, of which only YES1 was in the original seed set, and p38 MAPK as well as components of the relevant pathway (*e.g.* EGFR and BRAF - which is not a known target). Other interesting proteins include RIPK kinases and ROCK1/RhoA which have been previously shown to be advantageous for SARS-CoV2 infection (103) in relevant genome-wide CRISPR screens.

DISCUSSION

Signaling processes are very important for the physiological function of cells within their environment and they are highly complex and context-specific. This context-specificity is not captured by the current annotated pathways, which are a result of decades of individual studies and represent the consensus network downstream of individual receptors. It is not practical or feasible to delineate and annotate signaling processes in all possible contexts and conditions in which a cell signaling response occurs; a data-driven approach is therefore needed to identify the active signaling processes from context-specific and unbiased omics data.

Phosphoproteomics data are especially suitable for the study of cell signaling as it measures the signaling state of the cell directly, by providing the signature of phosphorylated proteins and sites in a given moment. As discussed, mapping the data on prior knowledge pathways suffers from literature bias and ignores the context and conditions in which the experiment was done. Conversely, purely data-driven network inference is extremely difficult. This is firstly due to the curse of dimensionality, as no available dataset provides as many data points as phosphosites making the problem unsolvable, and secondly, the large understudied signaling space, means that it is anyway very difficult to evaluate methods that do attempt data-driven signaling network inference (35). Here we present phuEGO which uses as its basis protein interaction networks (79), enriched in known signaling regulatory relationships (80–82). Protein interaction networks are continually becoming more unbiased through systematic efforts such as Bioplex (57) or HuRI (56), and therefore they allow us to ground our method on prior knowledge, while at the same time substantially mitigating the severe literature bias that signaling pathways suffer from. As the network is interconnected and no functional units are annotated, our method also allows the data to select the functional modules that are relevant, resulting in functional units that are not restricted by those described by currently annotated pathways and can better capture cross-talk between processes and functional units. In the presented results, we have used a static and universal network for the analysis and the context-specificity of the result stems solely from the data. In addition, while protein interaction networks are far less biassed than pathway databases, there still remains a certain bias, which is further enhanced by the fact that phuEGO uses semantic similarity to model the edges of the network. This means that poorly annotated nodes would be less preferentially used by the method and those without any annotation are indeed excluded (Experimental Procedures). The flexibility of phuEGO, however, means that a user can use any desired network and this can include for example entirely unbiased predicted networks of functional associations and/or context-specific base networks that take into consideration the transcriptome or proteome of the specific cell line or sample provided, wherever this is available.

The use of network propagation to extract active network modules and signatures is quite well-established (59) and indeed highly suitable for the study of cell signaling as it simulates signal propagation by the cell through protein interaction networks. However, currently available methods result in large networks that are very hard to interpret. PhuEGO tackles this issue using firstly the semantic similarity to model the edges, so as to specifically boost the functional signal inherent in the nodes of interest, and secondly applies the local propagation through the ego network deconvolution. The result comprises a much smaller network, organized in distinct functional units/modules that can then be analyzed functionally or examined in more detail either independently or viewed at the systems level through supernode links. This can allow the identification of feedback loops, or the prediction of interaction directions, even though they are not explicitly modeled. Depending on the interest of the user, the parameters can be tuned so that the resulting network is expanded, albeit noisier, or more specifically providing only key signaling processes for the dataset. Providing such precise signaling signatures makes it a lot easier to integrate phosphoproteomics datasets and perform unified analyses as exemplified by the COVID-19 datasets example in this article. At present, phuEGO performs separate analyses for upregulated and downregulated networks for clarity, but in the future, it is possible to integrate the two to extract single signaling network signatures. In particular, as more functional annotations become available for phosphosites, and the network can

include sign and effect of regulatory interactions, phuEGO can provide even more precise signaling signatures from phosphoproteomics data, including increasing the granularity to the phosphosite level, rather than the protein.

Finally, the three-layer propagation that phuEGO performs allows us to capture our knowledge with respect to signal transduction and tune the resulting output based on the seeds that we have the most confidence in to capture the active signal. In this study, we used tyrosine, serine/threonine and non-kinase phosphosites as the three layers, but the method can easily integrate diverse data modalities linking, for example, transcriptomics data, through transcription factor activities, with phosphoproteomics data, through kinase activities and other information.

In conclusion, we present a flexible method, phuEGO that performs (up-to) three-layer network propagation on phosphoproteomics data. We show that it is able to boost the signal-to-noise ratio, enrich functional phosphosites, and provide interpretable active signaling network signatures. It is of note that phuEGO performs well both in the high-quality, uniformly re-analyzed phosphoproteomics datasets in our benchmark (19) and in the datasets extracted from the original papers. It allows us to better compare and integrate global phosphoproteomics (and other omics) datasets, and potentially other sparse and noisy data types, such as single-cell RNAseq. Applying it on five phosphoproteomics datasets derived from cells infected with COVID-19 significantly improved our ability to compare them, and intersecting the two datasets that were collected in A549 cells resulted in significant enrichment of known targets for COVID-19, providing a subnetwork that could point to additional targets. Future improvements of phuEGO include using more unbiased, *e.g.* predicted, and context-specific networks as its basis and integrating functional annotations of phosphosites to improve the active signaling signature extraction. Overall, phuEGO is a useful and versatile tool for the proteomics community and will contribute to the improved study of context-specific cell signaling responses.

DATA AVAILABILITY

All data used in this study is publicly available in the literature, and compiled networks are provided with this work as supplementary tables. phuEGO is freely available as a package through Python Package Index (https://pypi.org/project/phuego/), with source code hosted on: https://github.com/haoqichen20/phuego, and documentation hosted on: https://phuego.readthedocs.io/en/latest/.

Supplemental data—This article contains supplemental data.

Acknowledgments—EMBL-EBI IT Support is acknowledged for provision of computer and data storage servers.

Funding and additional information—This work was supported by: the European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI).

Author contributions—H. C. software; T. K. formal analysis; E. P. writing–review & editing; E. P. and G. G. writing–original draft; E. P. and G. G. visualization; E. P. supervision; E. P. resources; E. P. project administration; E. P. investigation; E. P. funding acquisition; E. P. and G. G.conceptualization. G. G. validation; G. G. methodology; G. G. formal analysis; G. G. data curation.

Conflict of interest—The authors declare that they have no conflicts of interest with the contents of this article.

Abbreviations—The abbreviations used are: GO, Gene Ontology; LFC, Log-2 Fold Change; PCSF, Prize-Collecting Steiner forest; PIN, Protein interaction network; PPI, Protein-protein interactions; RWR, Random-Walk-With-Restart.

Received October 17, 2023, and in revised form, April 8, 2024, Published, MCPRO Papers in Press, April 19, 2024, https://doi.org/10.1016/j.mcpro.2024.100771

REFERENCES

1. Heldin, C.-H., Lu, B., Evans, R., and Silvio Gutkind, J. (2016) Signals and receptors. *Cold Spring Harb. Perspect. Biol.* **8**, a005900
2. Hubbard, S. R., and Miller, W. T. (2007) Receptor tyrosine kinases: mechanisms of activation and signaling. *Curr. Opin. Cell Biol.* **19**, 117–123
3. Zhang, W., and Liu, H. T. (2002) MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res.* **12**, 9–18
4. Natarajan, M., Lin, K.-M., Hsueh, R. C., Sternweis, P. C., and Ranganathan, R. (2006) A global analysis of cross-talk in a mammalian cellular signalling network. *Nat. Cell Biol.* **8**, 571–580
5. Vert, G., and Chory, J. (2011) Crosstalk in cellular signaling: background noise or the real thing? *Dev. Cell* **21**, 985–991
6. Guo, X., and Wang, X. F. (2009) Signaling cross-talk between TGF-beta/BMP and other pathways. *Cell Res.* **19**, 71–88
7. Mendoza, M. C., Er, E. E., and Blenis, J. (2011) The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation. *Trends Biochem. Sci.* **36**, 320–328
8. Hill, C. S., and Treisman, R. (1995) Transcriptional regulation by extracellular signals: mechanisms and specificity. *Cell* **80**, 199–211
9. Barolo, S., and Posakony, J. W. (2002) Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. *Genes Dev.* **16**, 1167–1181
10. Strasen, J., Sarma, U., Jentsch, M., Bohn, S., Sheng, C., Horbelt, D., *et al.* (2018) Cell-specific responses to the cytokine TGFβ are determined by variability in protein levels. *Mol. Syst. Biol.* **14**, e7733
11. [preprint] Moret, N., Liu, C., Gyori, B. M., Bachman, J. A., Steppi, A., Hug, C., *et al.* (2021) A resource for exploring the understudied human kinome for research and therapeutic opportunities. *bioRxiv*. https://doi.org/10.1101/2020.04.02.022277
12. Invergo, B. M., and Beltrao, P. (2018) Reconstructing phosphorylation signalling networks from quantitative phosphoproteomic data. *Essays Biochem.* **62**, 525–534
13. Terfve, C., and Saez-Rodriguez, J. (2012) Modeling signaling networks using high-throughput phospho-proteomics. *Adv. Exp. Med. Biol.* **736**, 19–57
14. Gstaiger, M., and Aebersold, R. (2009) Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat. Rev. Genet.* **10**, 617–627
15. Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B., and Aebersold, R. (2009) Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **138**, 795–806

16. Giudice, G., and Petsalaki, E. (2019) Proteomics and phosphoproteomics in precision medicine: applications and challenges. *Brief. Bioinform.* **20**, 767–777
17. Bell, A. W., Deutsch, E. W., Au, C. E., Kearney, R. E., Beavis, R., Sechi, S., *et al*. (2009) A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods* **6**, 423–430
18. Nilsson, T., Mann, M., Aebersold, R., Yates, J. R., 3rd, Bairoch, A., and Bergeron, J. J. M. (2010) Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods* **7**, 681–685
19. Ochoa, D., Jarnuczak, A. F., Viéitez, C., Gehre, M., Soucheray, M., Mateus, A., *et al*. (2020) The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **38**, 365–373
20. Neuberg, L. G. (2003) Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theor.* **19**, 675–685
21. Freedman, D., and Humphreys, P. (1999) Are there algorithms that discover causal structure? *Synthese* **121**, 29–54
22. Hill, S. M., Heiser, L. M., Cokelaer, T., Unger, M., Nesser, N. K., Carlin, D. E., *et al*. (2016) Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods* **13**, 310–318
23. Samaga, R., and Klamt, S. (2013) Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. *Cell Commun. Signal.* **11**, 43
24. 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**, 1–11
25. Santra, T., Kholodenko, B., and Kolch, W. (2012) An integrated Bayesian framework for identifying phosphorylation networks in stimulated cells. *Adv. Exp. Med. Biol.* **736**, 59–80
26. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–529
27. Zhang, Y., Kweon, H. K., Shively, C., Kumar, A., and Andrews, P. C. (2013) Towards systematic discovery of signaling networks in budding yeast filamentous growth stress response using interventional phosphorylation data. *PLoS Comput. Biol.* **9**, e1003077
28. Hornberg, J. J., Binder, B., Bruggeman, F. J., Schoeberl, B., Heinrich, R., and Westerhoff, H. V. (2005) Control of MAPK signalling: from complexity to what really matters. *Oncogene* **24**, 5533–5542
29. Arkun, Y. (2016) Dynamic modeling and analysis of the cross-talk between Insulin/AKT and MAPK/ERK signaling pathways. *PLoS One* **11**, e0149684
30. Casadiego, J., Nitzan, M., Hallerberg, S., and Timme, M. (2017) Model-free inference of direct network interactions from nonlinear collective dynamics. *Nat. Commun.* **8**, 2192
31. Petsalaki, E., Helbig, A. O., Gopal, A., Pasculescu, A., Roth, F. P., and Pawson, T. (2015) SELPHI: correlation-based identification of kinase-associated networks from global phospho-proteomics data sets. *Nucleic Acids Res.* **43**, W276–W282
32. Savage, S. R., and Zhang, B. (2020) Using phosphoproteomics data to understand cellular signaling: a comprehensive guide to bioinformatics resources. *Clin. Proteomics* **17**, 27
33. Jason, W., and Locasale, A. W.-Y. (2009) Maximum Entropy reconstructions of dynamic signaling networks from quantitative proteomics data. *PLoS One* **4**, e6522
34. Garrido-Rodriguez, M., Zirngibl, K., Ivanova, O., Lobentanzer, S., and Saez-Rodriguez, J. (2022) Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks. *Mol. Syst. Biol.* **18**, e11036
35. Sriraja, L. O., Werhli, A., and Petsalaki, E. (2023) Phosphoproteomics data-driven signalling network inference: does it work? *Comput. Struct. Biotechnol. J.* **21**, 432–443
36. Wilkes, E. H., Terfve, C., Gribben, J. G., Saez-Rodriguez, J., and Cutillas, P. R. (2015) Empirical inference of circuitry and plasticity in a kinase signaling network. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7719–7724
37. Hijazi, M., Smith, R., Rajeeve, V., Bessant, C., and Cutillas, P. R. (2020) Reconstructing kinase network topologies from phosphoproteomics data reveals cancer-associated rewiring. *Nat. Biotechnol.* **38**, 493–502
38. Crowl, S., Jordan, B. T., Ahmed, H., Ma, C. X., and Naegle, K. M. (2022) KSTAR: an algorithm to predict patient-specific kinase activities from phosphoproteomic data. *Nat. Commun.* **13**, 4283
39. Casado, P., Rodriguez-Prados, J. C., Cosulich, S. C., Guichard, S., Vanhaesebroeck, B., Joel, S., *et al*. (2013) Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal.* **6**, rs6
40. Mischnik, M., Sacco, F., Cox, J., Schneider, H.-C., Schäfer, M., Hendlich, M., *et al*. (2015) IKAP: a heuristic framework for inference of kinase activities from Phosphoproteomics data. *Bioinformatics* **32**, 424–431
41. Yang, P., Patrick, E., Humphrey, S. J., Ghazanfar, S., James, D. E., Jothi, R., *et al*. (2016) KinasePA: phosphoproteomics data annotation using hypothesis driven kinase perturbation analysis. *Proteomics* **16**, 1868–1871
42. Alexander Lachmann, A. M. (2009) KEA: kinase enrichment analysis. *Bioinformatics* **25**, 684
43. Yılmaz, S., Ayati, M., Schlatzer, D., Çiçek, A. E., Chance, M. R., and Koyutürk, M. (2021) Robust inference of kinase activity using functional networks. *Nat. Commun.* **12**, 1–12
44. Rudolph, J. D., de Graauw, M., van de Water, B., 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
45. Yosef, N., Zalckvar, E., Rubinstein, A. D., Homilius, M., Atias, N., Vardi, L., *et al*. (2011) ANAT: a tool for constructing and analyzing functional protein networks. *Sci. Signal.* **4**, l1
46. Chasman, D., Ho, Y. H., Berry, D. B., Nemec, C. M., MacGilvray, M. E., Hose, J., *et al*. (2014) Pathway connectivity and signaling coordination in the yeast stress-activated signaling network. *Mol. Syst. Biol.* **10**, 759
47. Basha, O., Mauer, O., Simonovsky, E., Shpringer, R., and Yeger-Lotem, E. (2019) ResponseNet v.3: revealing signaling and regulatory pathways connecting your proteins and genes across human tissues. *Nucleic Acids Res.* **47**, W242–W247
48. Moon, J. H., Lim, S., Jo, K., Lee, S., Seo, S., and Kim, S. (2017) PINTnet: construction of condition-specific pathway interaction network by computing shortest paths on weighted PPI. *BMC Syst. Biol.* **11**, 1–13
49. Ritz, A., Poirel, C. L., Tegge, A. N., Sharp, N., Simmons, K., Powell, A., *et al*. (2016) Pathways on demand: automated reconstruction of human signaling networks. *NPJ Syst. Biol. Appl.* **2**, 16002
50. Bailly-Bechet, M., Borgs, C., Braunstein, A., Chayes, J., Dagkessamanskaia, A., François, J.-M., *et al*. (2011) Finding undetected protein associations in cell signaling by belief propagation. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 882–887
51. Ljubić, I., Weiskircher, R., Pferschy, U., Klau, G. W., Mutzel, P., and Fischetti, M. (2006) An algorithmic framework for the exact solution of the prize-collecting steiner tree problem. *Math. Program.* **105**, 427–449
52. Scott, M. S., Perkins, T., Bunnell, S., Pepin, F., Thomas, D. Y., and Hallett, M. (2005) Identifying regulatory subnetworks for a set of genes. *Mol. Cell. Proteomics* **4**, 683–692
53. Yosef, N., Ungar, L., Zalckvar, E., Kimchi, A., Kupiec, M., Ruppin, E., *et al*. (2009) Toward accurate reconstruction of functional protein networks. *Mol. Syst. Biol.* **5**, 248
54. Huang, S. S., and Fraenkel, E. (2009) Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci. Signal.* **2**, ra40
55. Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S. S., Chayes, J., Borgs, C., *et al*. (2013) Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *J. Comput. Biol.* **20**, 124–136
56. Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., *et al*. (2020) A reference map of the human binary protein interactome. *Nature* **580**, 402–408
57. [preprint] Huttlin, E. L., Bruckner, R. J., Navarrete-Perea, J., Cannon, J. R., Baltier, K., Gebreab, F., *et al*. (2020) Dual proteome-scale networks Reveal cell-specific Remodeling of the human interactome. *bioRxiv*. https://doi.org/10.1101/2020.01.19.905109
58. Cho, D. Y., Kim, Y. A., and Przytycka, T. M. (2012) Chapter 5: network biology approach to complex diseases. *PLoS Comput. Biol.* **8**, e1002820

59. Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017) Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551–562
60. Peña-Castillo, L., Tasan, M., Myers, C. L., Lee, H., Joshi, T., Zhang, C., *et al*. (2008) A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome Biol.* **9 Suppl 1**, S2
61. Cho, H., Berger, B., and Peng, J. (2016) Compact integration of Multi-network Topology for functional analysis of genes. *Cell Syst.* **3**, 540–548.e5
62. Navlakha, S., and Kingsford, C. (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics* **26**, 1057–1063
63. Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **82**, 949–958
64. Guney, E., and Oliva, B. (2012) Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PLoS One* **7**, e43557
65. Reyna, M. A., Leiserson, M. D. M., and Raphael, B. J. (2018) Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* **34**, i972–i980
66. Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., *et al*. (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114
67. Ghiassian, S. D., Menche, J., and Barabási, A.-L. (2015) A DIseAse MOdule detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity Patterns of disease proteins in the human interactome. *PLoS Comput. Biol.* **11**, e1004120
68. Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013) Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115
69. Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., *et al*. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337
70. Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., and Stuart, J. M. (2013) Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion through Interacting Events (TieDIE). *Bioinformatics* **29**, 2757–2764
71. Drake, J. M., Paull, E. O., Graham, N. A., Lee, J. K., Smith, B. A., Titz, B., *et al*. (2016) Phosphoproteome integration reveals patient-specific networks in prostate cancer. *Cell* **166**, 1041–1054
72. Ochoa, D., Jonikas, M., Lawrence, R. T., El Debs, B., Selkrig, J., Typas, A., *et al*. (2016) An atlas of human kinase regulation. *Mol. Syst. Biol.* **12**, 888
73. Salek, M., McGowan, S., Trudgian, D. C., Dushek, O., de Wet, B., Efstathiou, G., *et al*. (2013) Quantitative phosphoproteome analysis unveils LAT as a modulator of CD3ζ and ZAP-70 tyrosine phosphorylation. *PLoS One* **8**, e77423
74. [preprint] Higgins, C. A., Nilsson-Payant, B. E., Kurland, A. P., Ye, C., Yaron, T., Johnson, J. L., *et al*. (2022) SARS-CoV-2 hijacks p38β/MAPK11 to promote virus replication. *bioRxiv*. https://doi.org/10.1101/2021.08.20.457146
75. Stukalov, A., Girault, V., Grass, V., Karayel, O., Bergant, V., Urban, C., *et al*. (2021) Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. *Nature* **594**, 246–252
76. Klann, K., Bojkova, D., Tascher, G., Ciesek, S., Münch, C., and Cinatl, J. (2020) Growth factor receptor signaling Inhibition Prevents SARS-CoV-2 replication. *Mol. Cell* **80**, 164–174.e4
77. Bouhaddou, M., Memon, D., Meyer, B., White, K. M., Rezelj, V. V., Correa Marrero, M., *et al*. (2020) The global phosphorylation landscape of SARS-CoV-2 infection. *Cell* **182**, 685–712.e19
78. Hekman, R. M., Hume, A. J., Goel, R. K., Abo, K. M., Huang, J., Blum, B. C., *et al*. (2021) Actionable Cytopathogenic host responses of human alveolar type 2 cells to SARS-CoV-2. *Mol. Cell* **81**, 212
79. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., *et al*. (2013) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363
80. 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
81. Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016) OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967
82. Licata, L., Lo Surdo, P., Iannuccelli, M., Palma, A., Micarelli, E., Perfetto, L., *et al*. (2020) SIGNOR 2.0, the SIGnaling network open resource 2.0: 2019 update. *Nucleic Acids Res.* **48**, D504–D510
83. UniProt Consortium. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489
84. The Gene Ontology Consortium. (2019) The gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338
85. Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E. N., Falcão, A. O., and Couto, F. M. (2008) Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* **9**, S4
86. Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2014) The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics* **30**, 740–742
87. Viger, F., and Latapy, M. (2005) Efficient and simple generation of random simple connected graphs with prescribed degree sequence. *J. Complex Networks* **3595**, 440–449
88. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* **6**, e1000641
89. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., *et al*. (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419
90. Tong, H., Faloutsos, C., and Pan, J.-Y. (2007) Random walk with restart: fast solutions and applications. *Knowl. Inf. Syst.* **14**, 327–346
91. Silverman, B. W. (2018) *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, UK
92. Traag, V. A., Waltman, L., and van Eck, N. J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233
93. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361
94. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., *et al*. (2022) The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692
95. Huang, R., Grishagin, I., Wang, Y., Zhao, T., Greene, J., Obenauer, J. C., *et al*. (2019) The NCATS BioPlanet - an integrated platform for exploring the Universe of cellular signaling pathways for Toxicology, systems biology, and chemical genomics. *Front. Pharmacol.* **10**, 445
96. Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., *et al*. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855
97. Szymkiewicz, D. (1934) Une conlribution statistique à la géographie floristique. *Acta Soc. Bot. Pol.* **11**, 249–265
98. Tuncbag, N., Gosline, S. J., 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
99. Wiredja, D. D., Koyutürk, M., and Chance, M. R. (2017) The KSEA App: a web-based tool for kinase activity inference from quantitative phosphoproteomics. *Bioinformatics* **33**, 3489–3491
100. Lemmon, M. A., and Schlessinger, J. (2010) Cell signaling by receptor tyrosine kinases. *Cell* **141**, 1117–1134
101. Gocek, E., Moulas, A. N., and Studzinski, G. P. (2014) Non-receptor protein tyrosine kinases signaling pathways in normal and cancer cells. *Crit. Rev. Clin. Lab. Sci.* **51**, 125–137
102. 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
103. Biering, S. B., Sarnik, S. A., Wang, E., Zengel, J. R., Leist, S. R., Schäfer, A., *et al*. (2022) Genome-wide bidirectional CRISPR screens identify mucins as host factors modulating SARS-CoV-2 infection. *Nat. Genet.* **54**, 1078–1089