Data-driven extraction of human kinase-substrate relationships from omics datasets

Benjamin Dominik Maier[1#], Borgthor Petursson[1#], Alessandro Lussana[1#] and Evangelia Petsalaki[1,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, United Kingdom

*Correspondence should be addressed to Evangelia Petsalaki, petsalaki@ebi.ac.uk

[#]These authors contributed equally

Running Title: Prediction of human kinase-substrate relationships

Abstract

Phosphorylation forms an important part of the signalling system that cells use for decision making and regulation of processes such as cell division and differentiation. To date, a large portion of identified phosphosites are not known to be targeted by any kinase. At the same time around 30% of kinases have no known target. This knowledge gap stresses the need to make large scale, data-driven computational predictions.

In this study, we have created a machine learning-based model to derive a probabilistic kinase-substrate network from omics datasets. Our methodology displays improved performance compared to other state-of-the-art kinase-substrate prediction methods, and provides predictions for more kinases. Importantly, it better captures new experimentally-identified kinase-substrate relationships. It can therefore allow the improved prioritisation of kinase-substrate pairs for illuminating the dark human cell signalling space.

Our model is integrated into a web server, $SELPHI_{2.0}$, to allow unbiased analysis of phosphoproteomics data, facilitating the design of downstream experiments to uncover mechanisms of signal transduction across conditions and cellular contexts.

Introduction

Cells relay information through signalling networks(1) that are typically regulated by post translational modifications (PTM), with phosphorylation being the best studied one among these(2). Phosphorylation of proteins is catalysed by kinases that target a specific set of substrates, changing their state and/or function. The importance of understanding kinase regulatory networks is highlighted by the fact that a large portion of targeted therapies that are currently used and being developed target kinases(3). Despite this, to date only a small portion (~5%) of the more than 100,000 known phosphosites have a known upstream kinase(4). At the same time 150 kinases have no known substrate and 90% of the phosphosites are assigned to the 20% of the best studied kinases. Given that several studies have demonstrated that the understudied kinases can be just as important for health as the well-studied ones(4), this points to a bias in the literature, where researchers prioritise well-studied proteins. In addition, publicly available databases, such as KEGG(5), Reactome(6), Omnipath(7) and others, that describe our current knowledge of cell signalling pathways, present a static view that represents the 'average' cell and doesn't capture condition-specific signalling networks. As a result these literature-defined static signalling pathways have a limited explanatory value when it comes to the analysis of phosphoproteomics data.

Mass spectrometry-based phosphoproteomics data sets present relatively unbiased views of a cell's signalling state and could be used for extracting unbiased signalling networks from specific experimental conditions(8, 9). However, most well-performing methods that have been developed to this end, base their predictions on prior networks((8, 9) in the form of known pathways mentioned above, thus perpetuating existing literature biases. For example, PropheticGranger, one of the top scoring methods in the HPN-DREAM challenge((8, 9), applies heat diffusion coupled

with L1 penalized regression(10) on a network based on the Pathway Commons database(11). There is thus a need for a more unbiased prior network to be used in methods for network inference of signalling networks from phosphoproteomics data.

Several methods have been developed to predict kinase-substrate interactions(12). Most of them use in some way either kinase specificity maps in the form of Position Specific Scoring Matrices (PSSMs)(13–16), protein interaction or other types of networks(17), structural information(18) or combinations of these(19). More recently co-expression at the protein level(20) or co-phosphorylation has been used(21). A wide range of approaches is used from basic PSSM scoring, log odds ratios(14), profile Hidden Markov Models(16), to machine learning (neural networks(19), SVMs(22), Bayesian decision theory(23)) and more recently deep or transfer learning(24, 25). However, most of these methods are only able to make predictions for a small fraction of the kinome and many of them only make predictions at the kinase family level.

There have recently become available Position Specific Scoring Matrices (PSSMs) describing the specificity preferences of almost all human kinases (26, 27). This information is invaluable for mapping kinase-substrate relationships but it is difficult to deconvolve kinase-substrate relationships when they have similar PSSMs. Integrating co-expression (at the gene and protein level) and other types of information can help resolve this issue, allowing the generation of kinase-substrate regulatory networks in a data-driven way.

In previous work, we used a machine learning approach, combining predictors such as co-phosphorylation, co-expression and kinase specificity models to derive such a data-driven kinase-kinase regulatory network(28). The resulting network provided predictions only at the protein level, whereas it is known that proteins have multiple functional phosphosites, often with entirely different or even opposing functions. E.g. phosphorylation of Y530 found on the Src kinase leads

to its inhibition while dephosphorylation of Y530 and phosphorylation of Y419 leads to its activation(29). In addition, while kinase regulatory networks form the 'skeleton' of the phospho-based signalling networks, non kinase substrates are also critical e.g. in the case of adaptor proteins forming scaffolds to regulate signal propagation(30), for phosphatases removing phosphosites to shut signals off, regulation of transcription factor translocation or activities(31) and others.

To address these points, here, we have extended our previous machine learning model introducing additional predictors to include non-kinase substrates and to provide a list of highly probable novel experimentally predicted kinase-substrate relationships(27, 32, 33) at the phosphosite level. Our method, called $SELPHI_{2.0}$ (Systematic Extraction of Linked PHospho-Interactions 2.0), is able to make predictions for less studied kinases and substrates compared to those found in the literature and perform better than established kinase-substrate prediction methods(19, 20, 34–36). To facilitate the use of this network rather than literature-based pathways as a prior network, we created a web server that extracts context-specific signalling networks from global phosphoproteomics datasets in addition to performing other functional analyses. It substantially improves and extends the functionalities of the previously published SELPHI web server(8) and provides an accurate backbone for network extraction rather than a simple correlation measure for kinase-substrate associations that was used by SELPHI. We have made this service available through the url: https://selphi2.com.

Experimental Procedures

Generation of kinase-substrate probabilistic network

Information on known kinase-substrate relationships as well as a list of phosphosites and the amino acids sequences surrounding them was gathered from PhosphoSitePlus(37) (Downloaded on 03.09.2024). Functional scores and predictive features on phosphosites were downloaded from Ochoa and colleagues(38) (**Supplemental table 1**). Proteomics data sets for co-regulation were downloaded from Mertins and colleagues(39) as well as Hijazi and colleagues(33). Expression data was gathered from GTEx(40) (Downloaded on 03.09.2024) and the Human Protein Atlas version 23.0(41) (Downloaded on 03.09.2024). Experimentally supported kinase-substrate relationships were downloaded from two recent publications(32, 33).

Predictions were made between 421 kinases, and 238,374 phosphosites found on 17,469 proteins that were listed in PhosphositesPlus(37) and had a functionality score(38) assigned to them.

As a positive training set we used 14,542 kinase-phosphosite relationships extracted from PhosphositePlus(37) (Downloaded on 03.09.2024). As there are no databases with information on known negative kinase-substrate pairs and given the fact that biological networks tend to be sparse, random samples of kinase-substrate relationships ten, twenty, and fifty times as large as the positive set were used as a negative training set. There were no differences between the results, we therefore present those where we used a negative set fifty times as large as the positive set.

We created or acquired 48 features (see below and **Supplemental Table 1**) considering different metrics that could affect a kinase-substrate relationship, from co-regulation in high throughput datasets to match of a phosphosite to a kinase specificity model as described by position-specific scoring matrices (PSSMs). To select the most useful predictors we evaluated the performance of

different feature combinations using 100 training/testing datasets. To select the most useful features for kinase-substrate prediction we generated 100 different training sets using our positive set and a set of random kinase-substrate relationships ten times larger than the positive set as a negative set. Random forest(42) classifiers as implemented in *scikit-learn(43)* were used to make predictions. The random forest function has various tunable parameters. *Bootstrap* which enables bootstrap samples to be used when building the trees was set to TRUE. The following parameters were tuned by grid search: *max_depth* which sets the maximum depth of the trees, *min_samples_split* which sets the minimum number of samples needed to split an internal node and *n_estimators* which set the number of trees included in the forest. The following parameters were considered: max_depth: [10, 20, 50, 70, 80, 100], *min_samples_split*:[8, 10, 12], *n_estimators*: [150, 500, 1000, 1500]. All combinations were tested with the *GridSearchCV()* function as implemented in scikit-learn. We then performed feature ranking with recursive feature elimination with the cross-validation algorithm, as implemented in *scikit-learn(43)*. We used ten-fold cross-validation to select the best performing feature composition using the area under the ROC curve to evaluate each model. We selected the 45 features to keep in the final model based on majority voting, *i.e.* selecting features included in >50% of the best performing models (**Supplemental Table 1**).

For training the final model we used the random forest algorithm(42) as implemented in the *scikit-learn* python library(43). The model was trained with the positive set and a random sample one hundred times. Parameters were tuned in each run using grid parameter search as described above. Each model was validated by using ten fold cross-validation. In order to balance the training and test set, a stratified K-fold split as implemented in the *scikit-learn(43)* python library was used to keep the proportion between negative and positive labels in each split the same. The average

probability across the different models' outputs was then calculated to assign probabilities to kinase substrates.

To further evaluate the performance of our model on novel kinase-substrate relationships we used predictions from two recently published experimental studies(32, 33) and tested whether these relationships were assigned a higher probability by our method. To quantify the predictive power of our model in each case, the area under the ROC curve was calculated as implemented in the *ROCR* package(44).

Comparison with other methods

To compare $SELPHI_{2.0}$ with other state-of-the-art peer reviewed methods we assessed how well our method predicted known annotated kinase-substrate relationships (see positive/negative training sets described above) and also novel kinase-substrate relationships supported by experimental procedures. These were derived from the recent works of Sugiyama and colleagues, and Hijazi and colleagues (32, 33), excluding kinase-substrate phosphorylation relationships found in PhosphoSitePlus, to limit literature-based biases.

We compared our method with five other established and most commonly used methods: PhosphoPICK(20), GPS v.5.0(36), GPS v.6.0(24), KinomeXplorer(19), NetPhos v.3.1(34) and LinkPhinder(35). PhosphoPICK and NetPhos v3.1 are available as web servers and sequences of substrate proteins were uploaded onto the servers. For both PhosphoPICK and NetPhos v3.1 we selected no significance threshold to include all predictions. For NetPhos v3.1 we selected all available residues: Serine, threonine and tyrosine. GPS was downloaded from (http://gps.biocuckoo.cn/download.php) and batch kinase prediction was run on Ubuntu 22.04. Confidence threshold was set to "All" to include all predictions. The KinomeXplorer prediction software was downloaded and predictions were made for the phosphosites included in our network.

All 11,581,940 LinkPhinder predictions were downloaded from (https://linkphinder.insight-centre.org/download).

Methods were compared by assessing their ability to capture known kinase substrate relationships from PhosphoSitePlus. As the published methods were able to make predictions for different subsets of kinases, we restricted the comparisons to those predictions that were possible by both methods(**Supplemental Table 2**). For each method under comparison, we generated 100 training sets made from the positive set and a negative set selected from random kinase substrate relationships that were present both in $SELPHI_{2.0}$ as well as the method that was under comparison. The negative set was ten times larger than the positive set.

The ability of $SELPHI_{2.0}$ to discriminate between known interactions and interactions with unknown status was then compared to each of the other methods by calculating the average AUC achieved by $SELPHI_{2.0}$ and the method under comparison for the 100 different training sets only including kinase substrate relationships shared by both methods.

In order to estimate the methods' ability to capture novel kinase substrate interactions we used a set of experimental kinase substrate predictions made previously by Sugiyama and colleagues[8]. For each method, the overlapping predictions between $SELPHI_{2.0}$ and the method under comparison were assessed by how well they captured the experimentally predicted kinase substrate relationships as measured by the AUC. AUC scores were calculated using the *ROCR(44)* package.

Evaluation of model fit to phosphoproteomic data

To capture context specific signalling networks, we constructed a reference $SELPHI_{2.0}$ signalling network by linking the kinase-substrate predictions made in this work to a backbone of a

probabilistic kinase kinase regulatory network that we had previously published(28). A probability cutoff of 0.5 was applied on both networks to select high confidence edges.

High-throughput mass spectrometry-based phosphoproteomic data that had been compiled and analysed by Ochoa and colleagues with 436 conditions(45) was fitted to our network by means of the PCSF algorithm(46). We tried all combinations of the of the following PCSF parameters: For the tree parameter, b, we tried fitting anything from 1 to 10 trees to the data, for parameter w or the node tuning we tried the following values: 0.25, 0.5, 0.75, 1.0, 1.25, 1.5 and for the edge tuning, μ, we tried 0.000005, 0.00005, 0.0005, 0.005, 0.05. Edge probabilities subtracted from 1 were used as edge costs. These parameter combinations generated 436 sub networks that gave the highest F1 score. To evaluate the performance of the fitting, the F1 score of kinase-substrate phosphosite relationships retained after the fitting was compared with those that were used in the input.

Counting number of citations related to proteins

To count the number of citations related to kinases and their substrates we used the Entrez.elink() function from the Biopython module(47). We searched for related articles in the Pubmed database. Linked publications were retrieved from NCBI Entrez Gene(48) database and publications that mention more than ten kinases were filtered out.

Implementation of $SELPHI_{2.0}$ web server

The $SELPHI_{2.0}$ web server has been implemented using the *shiny(49)* R package. Enrichment analyses on the server are performed using the *enrichR(50)* and the clusterProfiler(51) enrichment package and users will be able to select gene sets for enrichment from various databases: Reactome(6), KEGG(5), GO(52) and DISEASES(53). For phosphopeptide clustering, the two clustering algorithms available are Gaussian mixture models as implemented by *mclust(54)* and k-

means clustering(55). If k-means clustering is selected, silhouette index(56) is used to calculate the optimal number of clusters while *mclust* uses Bayesian information criterion to select the best fitting model(57). The output is presented as a dot plot where the colour indicates the sum of the log_2 ratios of the phosphosites involved in the enrichment term and size the odd ratio of the enrichment term.

For the calculation of kinase activities the web server uses the 95th percentile of kinase-substrate probabilities to select high confidence kinase-substrates. We then use the Kolmogorov-Smirnov statistical test to assess if the predicted substrates are overrepresented at the upper end of the log_2 ratio distribution for a given condition. The negative log_{10} transformation of the derived p-value is used to quantify activities. P-values are then multiplied by the average log_2 ratios of the predicted kinase-substrates. For comparison kinase activities using the PhosphositePlus(37) kinase-substrate data set are also calculated.

Results

Generation of a probabilistic human kinase-substrate regulatory network

To create a probabilistic network of kinase-substrate relationships we used the random forest algorithm to combine various predictive variables ranging from co-phosphorylation and co-expression in large datasets, to kinase specificities and features related to the functionality of the phosphosites (**Experimental Procedures**). After running the prediction algorithm 100 times we achieved an average AUC of 0.97 (**Figure 1A**). Of the 73,207,860 predictions made, 721,707 edges were high confidence (probability > 0.5). Of those, around 1.7% were found in PhosphoSitePlus(37) while more than 90% of the known interactions found in our network were assigned a high (> 0.5) probability.

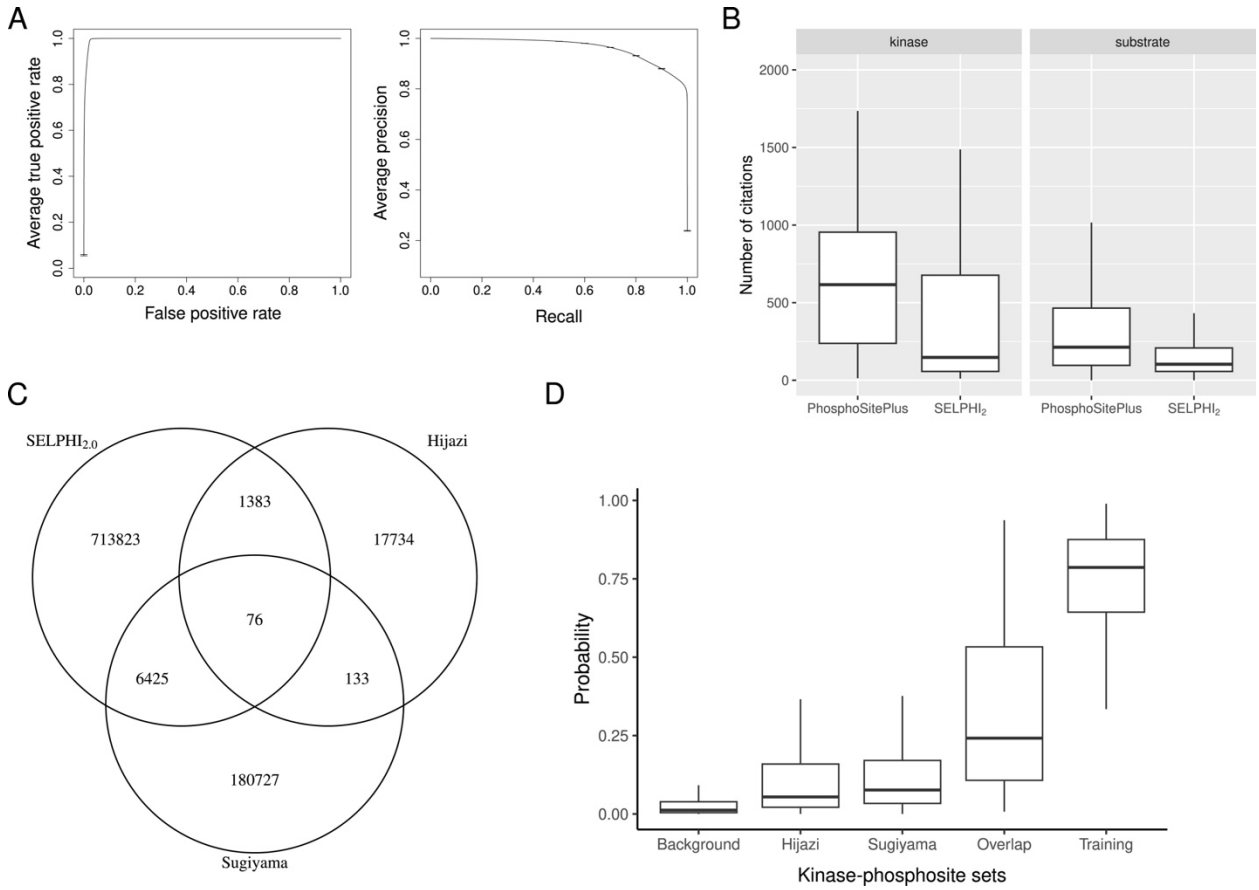

Figure 1. ***Precision and recall and different cutoffs*** *(A). The predictive power of our predictive method. AUC under the ROC curve of 0.97 after 100 runs of cross-validation with a random set of negatives and a positive set from phosphosite plus. Precision recall curve for the same set can be seen on the right (B). We found that our classifier is able to make high confidence predictions for less cited kinases and less cited substrates compared to kinase-substrates found in PhosphositePlus (C). We identified 76 high quality kinase-substrate interactions found by* $SELPHI_{2.0}$ *as well as the two experimental datasets in our study (D). Experimentally validated kinase-substrate relationships were assigned a higher probability compared to the background with kinase-substrate predictions made by both methods being assigned the highest probabilities of all sets.*

Importantly, our network provides high confidence predictions covering the 'dark' or less well studied human signalling network. Specifically, kinase-substrate relationships found in PhosphositePlus have a median number of 685 citations per phosphorylating kinase and 232 per substrate respectively, whereas our high confidence predictions the median number of citations are

194 for the kinases and 96 for the substrates. We provide proportionally more predictions between kinases and substrates with significantly lower number of citations per protein on average (Kinases: $W = 3.37 \times 10^9$, $p < 2.2 \times 10^{-16}$, one-sided Wilcoxon test. Substrate proteins: $W= 3.34 \times 10^9$, $p < 2.2 \times 10^{-16}$, one-sided Wilcoxon test)(**Figure 1B**). These predictions include substrates that have not been mentioned in the literature before and have no known upstream kinase, highlighting the value of this network as a reference to explore the less studied part of the phospho-signalling network.

SELPHI$_{2.0}$ can be used to identify novel high confidence kinase-substrate relationships

To assess the ability of our method to predict novel kinase-substrates, we looked at how well we discerned between a set of experimental kinase-substrate predictions not hitherto found in the literature and the rest of the unsupported predictions. To this end, we used experimentally predicted kinase-substrate relationships from two recent papers(32, 33). In short, one study introduces kinases to dephosphorylated peptides from cell lysis, while the other predicts kinase-substrates based on changes in phosphorylation following kinase inhibition.

We found that the overlap between the three sets, the two experimental sets and the SELPHI$_{2.0}$ predictions, was relatively low (**Figure 1C**). Due to the complementary characteristics between the two experimental methods we reasoned that kinase-substrate relationships identified by both studies should have higher levels of confidence. Our method assigned significantly higher probabilities to experimentally supported edges compared to the rest of the network (**Figure 1D**). Furthermore, we found that kinase-substrate relationships supported by edges found in both data

sets had an even higher probability assigned to them (**Figure 1D**) compared to edges supported by either.

SELPHI$_{2.0}$ outperforms the state-of-the-art methods for kinase-substrate prediction

In comparison to 5 state-of-the-art kinase-substrate prediction methods (PhosphoPICK(20), GPS v.5.0(36) & v.6.0(24), KinomeXplorer(19), NetPhos v.3.1(34) and LinkPhinder(35)), SELPHI$_{2.0}$ performs generally better on identifying known kinase-substrate interactions (**Supplemental Figure 1**), while making predictions for more kinases. When comparing the performance of these methods using an independent dataset of experimentally supported kinase-substrate pairs(32, 33), having removed those that already exist in PhosphositePlus to remove the effect of literature bias on the methods, we found that our network performs better than the others with GPSv.6.0 performing the closest with AUC=0.81 compared to our method's 0.83. NetPhos v3.1 achieved AUC of 0.69 compared to SELPHI$_{2.0}$'s AUC of 0.75, but made predictions for only 17 (38 kinase genes) kinases. KinomeXplorer(19) achieved AUC of 0.61 compared to SELPHI$_{2.0}$'s AUC of 0.81 for the same set (**Supplemental table 2**). Other methods performed close to random (**Figure 2**).

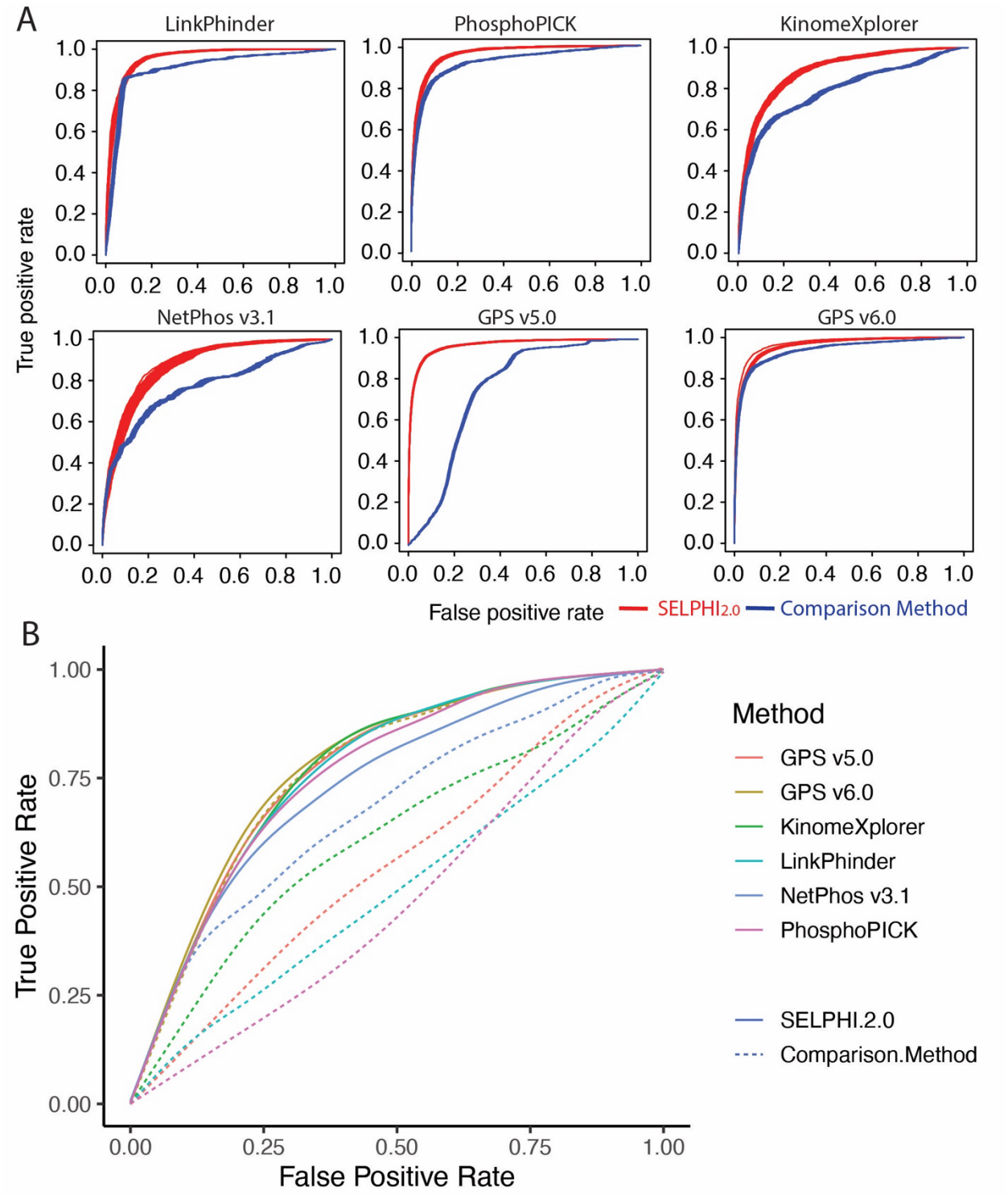

Figure 2: *Comparison between SELPHI$_{2.0}$ (solid lines) and other state of the art kinase-substrate prediction methods (dashed). The ability of each method to discern between test (A) or experimentally supported (B) kinase-substrate relationships and the rest of the relationships unsupported by the experiments. In all cases SELPHI$_{2.0}$ performed better than the state of the art methods.*

Extraction of dataset-specific networks from our reference network selects for known interactions

We fitted our network to a set of mass spectrometry-based global phosphoproteomic data sets generated under different conditions. We used a compilation of mass spectrometry data sets compiled and reanalyzed downloaded from an earlier publication(45). To fit our predictions to high-throughput data, the predicted kinase-substrate edges were combined with a kinase-kinase regulatory network that we had generated previously(28), forming a network of kinase-kinase regulatory relationships with phosphosites as nodes without outgoing edges (**Experimental Procedures**). In order to select high confidence edges, we used edge probability of 0.5 as a threshold for both the kinase-kinase regulatory network and the kinase-substrate predictions.

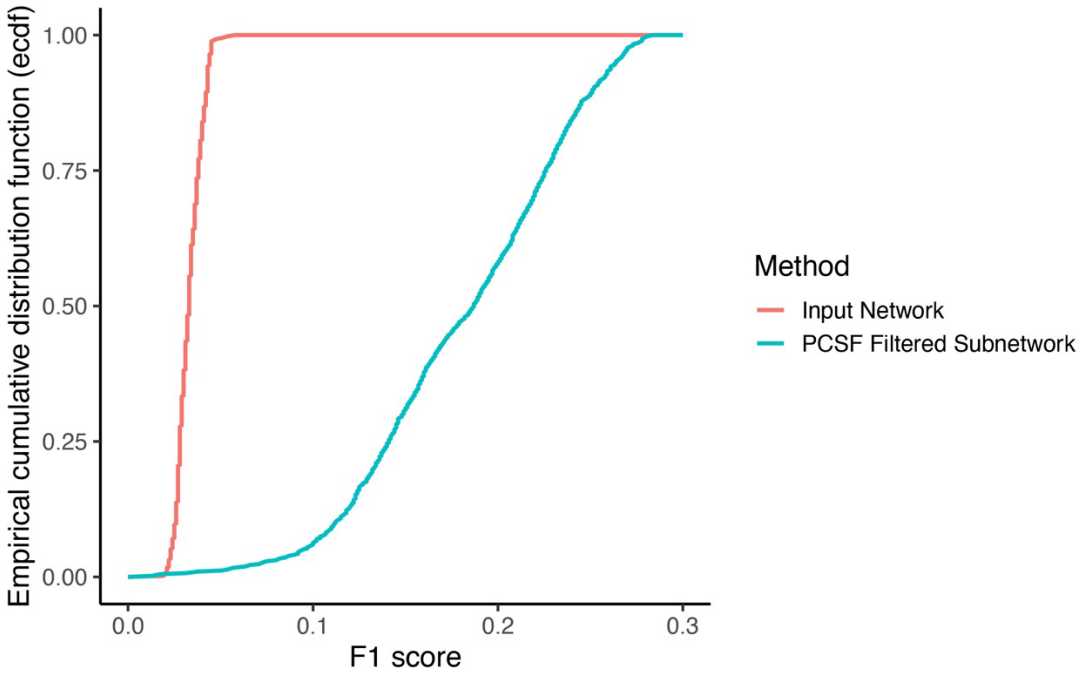

Figure 3: *Fitting the kinase-substrate predictions to experimental data selects for known kinase-substrate relationships with median F1 score of 0.031 for the unfitted sub network compared to 0.19 for the fitted input.*

To fit the combined network to the high-throughput data sets, we used Prize collecting Steiner's forest as implemented in the R package PCSF(46). We found that by optimising the edge cost against the node prizes we were able to select for edges found in the literature. The F1 scores of the fitted subnetworks (n = 415) were 0.19 compared to the unpruned input with a F1 score of 0.031. The improvement in precision was even greater with the mean precision of the pruned subnetworks being 0.22 while the precision was 0.016 for the unpruned input networks. Both comparison yielded a significant difference (F1 score: $W = 2.14 \times 10^6$, $p < 2.2 \times 10^{-16}$, precision: $W = 2.14 \times 10^6$, $p < 2.2 \times 10^{-16}$), indicating this combination of kinase-kinase regulatory network and kinase-substrate predictions can be used to extract high probability context-specific subnetworks (**Figure 3**).

SELPHI$_{2.0}$ web server

We have created a user-friendly web server, SELPHI$_{2.0}$, that allows less biased analyses of global phosphoproteomics datasets, to help users shed light on the mechanisms underpinning their data, including context-specific and understudied signalling processes. The basis of this web server is the data-driven network of kinase substrate predictions described above. SELPHI$_{2.0}$ improves on an earlier resource SELPHI (Systematic Extraction of Linked Phospho-Interactions)(8), which calculated kinase-substrate associations based on correlations and filtering based on prior knowledge.

Users can use SELPHI$_{2.0}$ to retrieve the top kinase-substrate predictions for the phosphosite included in their data and extract condition specific kinase activity profiles (**Figure 4A**). They can also perform standard functional enrichment analysis either on their up/down-regulated phosphosites or on clusters of phosphosites identified through the server. Furthermore, users can

identify the active signalling sub-networks in their data by mapping it to our SELPHI$_{2.0}$ reference kinase-substrate network (**Experimental Procedures; Figure 4B**). Finally, the server also provides higher confidence relationships by overlapping our predictions with the two experimentally-derived data sets published previously(32, 33) . This way the user can prioritise edges that have been corroborated by independent experimental results as well as large scale omics datasets (**Figure 4C**).

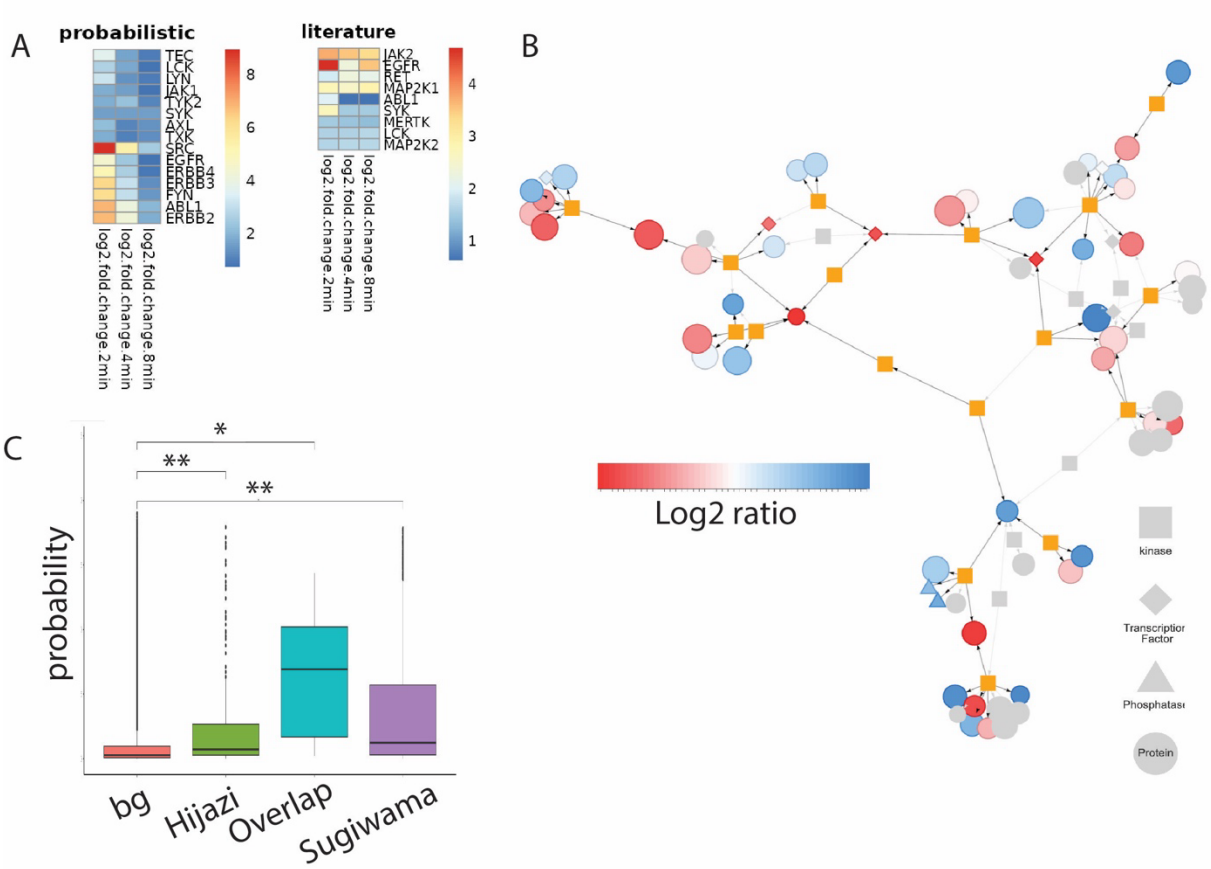

Figure 4: *Examples of the SELPHI$_{2.0}$ web server outputs using as input the first example in the web server. The user can upload phosphoproteomics data and fit the SELPHI$_{2.0}$ reference network to it (A). Kinase activities across the different conditions or time points can be estimated based on kinase substrate predictions (Heatmap titled Prob_Y) and kinase substrates as listed in PhosphoSitePlus (Heatmap, titled PPS_Y) for comparison (B). The user can assess if experimentally predicted edges from independent studies corroborate the prediction made by SELPHI$_{2.0}$. Furthermore, The user can download a list of these experimentally corroborated edges (C).*

Discussion

Kinase-substrate networks form the backbone of cell signalling responses and are critical for cell function in health and disease(58). It is, therefore, important to accurately annotate kinase-substrate relationships, but the vastness of the potential kinase-substrate interaction space makes computational means necessary for their prioritisation. Here, we have used a machine learning model, integrating information largely from high throughput datasets, to generate a probabilistic human kinase-substrate network at the phosphosite level that includes 421 kinases and 238,374 phosphosites.

Our method, called $SELPHI_{2.0}$, performs better than the five state-of-the-art methods tested(19, 20, 34–36), not only on the benchmark set but importantly on entirely new experimentally supported data(32, 33). This is true despite including predictions for more kinases than all other methods. This suggests that our network is a good starting point for prioritisation of new kinase-substrate relationships. NetPhos(34) proved to be an exception, performing well both at discerning between known positives and negatives as well as experimentally validated edges. It should be kept in mind though that NetPhos only makes predictions for 17 kinases (38 kinase genes).

When overlaying our predictions with the two experimentally supported datasets, the resulting 24 high confidence kinase-substrate relationships are supported by both *in vitro* and cell line-based experimental data, and our data-driven machine learning approach, giving them particularly high confidence.

Our predictors, with the exception of the literature-derived PSSMs, were based on unbiased, high throughput datasets including global phosphoproteomics datasets. This, in combination with the coverage of the dark space that our method affords, is a step towards reducing the bias in cell signalling studies. Our network provides much wider coverage of the kinase signalling space than

current knowledge and most other available kinase-substrate prediction methods and is more accurate.

With the median citation number for kinases and substrates in the network being 3.5 and 2.4 times less than the PhosphositePlus database, our network forms a springboard for the exploration of the dark human cell signalling space. Given the importance of a prior network in methods of signalling network inference, this network will significantly contribute to a better understanding of the role of the understudied space in the context of our current knowledge, and will allow methods to generate networks that more accurately reflect the data.

Our method combines the power of PSSMs for kinase-substrate association mapping based on kinase specificity information(26, 27), with functional information from unbiased functional genomics datasets. This can open the door to improved kinase activity estimation from phosphoproteomics datasets improving our ability to describe context-specific signalling processes. For example, tools like PhosX(59)use kinase specificity scores rather than known substrates, can readily take advantage of the $SELPHI_{2.0}$ scores to calculate kinase activities considering both the PSSMs and multiple additional metrics describing the kinase substrate relationships, such as coexpression and others.

We have created a web server called $SELPHI_{2.0}$, which allows the easy analysis of global phosphoproteomics data to design downstream experiments. Global phosphoproteomics datasets provide a unique snapshot of the signalling networks that are active in a given cell type and condition, i.e. context. It is, however, very difficult to extract unbiased 'active' signalling networks from these datasets largely due to the inherent noisiness of both the experimental measurements and cell signalling itself, and due to the sparsity of these datasets(4). In addition to standard analyses such as functional enrichment and clustering, it provides a method for extracting context

specific networks from the data. $SELPHI_{2.0}$ improves on a previous web server, SELPHI, which was using simple correlation between kinases and potential substrates filtered by prior knowledge information for the generation of context-specific signalling networks from phosphoproteomics data. It thus perpetuated the literature bias, for kinases with existing knowledge, while it provided very noisy associations for the understudied ones.

In $SELPHI_{2.0}$ the user can derive dataset-specific networks extracted from our reference probabilistic network using PCSF. As signalling responses vary depending on the conditions and cell properties (e.g. tissue, state etc) and current signalling pathways poorly explain observed phosphoproteomics datasets, methods that infer context-specific signalling networks are necessary to understand signalling in specific contexts of interest. In the future, more sophisticated statistical methods, such as MRA(60) or Integer Linear Programming methods(61), can be implemented as additional options.

Accumulation of false positives is a persistent problem when kinase-substrates are predicted. This is partly due to the large fraction of the signalling network that is currently unknown or understudied but a fraction of these predictions can be assumed to be false positives, even though this can't possibly be confirmed. This is likely true for our method as well and users of the network should take this into consideration when interpreting the results of their analysis. Nevertheless, the vast understudied signalling space requires computational approaches for prioritisation of hypotheses and the performance of our network compared to the state of the art indicates that it provides a good starting point. As more high quality high throughput phosphoproteomic data sets become available our model can be further improved.

We expect that $SELPHI_{2.0}$ will allow the improved prioritisation of kinase-substrate pairs for illuminating the dark human cell signalling space, both through smaller scale signalling studies

and through acting as a relatively unbiased prior in signalling network inference approaches. Our web server also makes this network accessible to non bioinformatics experts and allows the interrogation of large scale phosphoproteomics datasets in an unbiased and context-specific way. We thus expect it to facilitate the design of downstream experiments to illuminate mechanisms of signal transduction across conditions and contexts.

Data and code availability

The SELPHI$_{2.0}$ web server is freely available under: https://selphi2.com and all code for reproducing this project can be found at https://gitlab.ebi.ac.uk/petsalakilab/selphi_2.0 and https://github.com/alussana/selphi2.0-pssms. A Docker image of the web server can be found at: https://gitlab.ebi.ac.uk/petsalakilab/selphi_2. The feature table, positive training set. prediction results and citation matrix can be found in Zenodo: https://zenodo.org/records/13821321?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjIyYjliMThhLWJjNTQtNDMwYy05NTJkLWNiM2E1NjA0MDFkYiIsImRhdGEiOnt9LCJyYW5kb20iOiJiZDk3NmU1MjY4MWEwY2RlOTg4MGJmMmM3YWFkODNhMSJ9.T_-xpVVN5MVVHVTM5xH6YXXcfTZHI94gFIJb92DIAxwtOzCEx5YZXVP_I8o79c9XI72qV9-MxDtsbdMD1t5thg

Author Contributions

BP: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - Original Draft, Visualization. BM & AL: Methodology, Validation, Formal analysis, Investigation, Software, Data curation, Writing - Review & Editing. EP: Conceptualization,

Methodology, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

This work was supported by the European Molecular Biology Laboratory (EMBL), and EMBL-EBI. BP, BM and AL were funded by the EMBL International PhD Programme. EMBL-EBI IT Support is acknowledged for provision of computer and data storage servers.

References

1. Pawson, T., and Saxton, T. M. (1999) Signaling Networks—Do All Roads Lead to the Same Genes? *Cell* 97, 675–678

2. Hunter, T. (1995) Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell* 80, 225–236

3. Bhullar, K. S., Lagarón, N. O., McGowan, E. M., Parmar, I., Jha, A., Hubbard, B. P., and Rupasinghe, H. P. V. (2018) Kinase-targeted cancer therapies: progress, challenges and future directions. *Mol. Cancer* 17, 48

4. Needham, E. J., Parker, B. L., Burykin, T., James, D. E., and Humphrey, S. J. (2019) Illuminating the dark phosphoproteome. *Sci. Signal.* 12,

5. 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

6. Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.* 48, D498–D503

7. Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016) OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* 13, 966–967

8. 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, W276–82

9. 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., Graim, K., Bivol, A., Wang, H., Zhu, F., Afsari, B.,

Danilova, L. V., Favorov, A. V., Lee, W. S., Taylor, D., Hu, C. W., Long, B. L., Noren, D. P., Bisberg, A. J., HPN-DREAM Consortium, Mills, G. B., Gray, J. W., Kellen, M., Norman, T., Friend, S., Qutub, A. A., Fertig, E. J., Guan, Y., Song, M., Stuart, J. M., Spellman, P. T., Koeppl, H., Stolovitzky, G., Saez-Rodriguez, J., and Mukherjee, S. (2016) Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods* 13, 310–318

10. 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

11. Rodchenkov, I., Babur, O., Luna, A., Aksoy, B. A., Wong, J. V., Fong, D., Franz, M., Siper, M. C., Cheung, M., Wrana, M., Mistry, H., Mosier, L., Dlin, J., Wen, Q., O'Callaghan, C., Li, W., Elder, G., Smith, P. T., Dallago, C., Cerami, E., Gross, B., Dogrusoz, U., Demir, E., Bader, G. D., and Sander, C. (2020) Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res.* 48, D489–D497

12. Zhou, F., Xue, Y., Yao, X., and Xu, Y. (2006) A general user interface for prediction servers of proteins' post-translational modification sites. *Nat. Protoc.* 1, 1318–1321

13. Neuberger, G., Schneider, G., and Eisenhaber, F. (2007) pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model. *Biol. Direct* 2,

14. Li, T., Li, F., and Zhang, X. (2008) Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins: Struct. Funct. Bioinf.* 70, 404–414

15. Obenauer, J. C., Cantley, L. C., and Yaffe, M. B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* 31, 3635

16. Huang, H. D., Lee, T. Y., Tzeng, S. W., and Horng, J. T. (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.* 33,

17. Ayati, M., Yilmaz, S., Lopes, F. B. T., Chance, M., and Koyuturk, M. (2023) Prediction of Kinase-Substrate Associations Using The Functional Landscape of Kinases and Phosphorylation Sites. *Pac. Symp. Biocomput.* 28, 73

18. Brinkworth, R. I., Breinl, R. A., and Kobe, B. (2003) Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl. Acad. Sci. U. S. A.* 100,

19. Horn, H., Schoof, E. M., Kim, J., Robin, X., Miller, M. L., Diella, F., Palma, A., Cesareni, G., Jensen, L. J., and Linding, R. (2014) KinomeXplorer: an integrated platform for kinome biology studies. *Nat. Methods* 11, 603–604

20. Patrick, R., Lê Cao, K.-A., Kobe, B., and Bodén, M. (2015) PhosphoPICK: modelling cellular context to map kinase-substrate phosphorylation events. *Bioinformatics* 31, 382–389

21. Ayati, M., Wiredja, D., Schlatzer, D., Maxwell, S., Li, M., Koyutürk, M., and Chance, M. R. (2019) CoPhosK: A method for comprehensive kinase substrate annotation using co-phosphorylation analysis. *PLoS Comput. Biol.* 15, e1006678

22. Kim, J. H., Lee, J., Oh, B., Kimm, K., and Koh, I. (2004) Prediction of phosphorylation sites using SVMs. *Bioinformatics* 20,

23. Xue, Y., Li, A., Wang, L., Feng, H., and Yao, X. (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics* 7,

24. Chen, M., Zhang, W., Gou, Y., Xu, D., Wei, Y., Liu, D., Han, C., Huang, X., Li, C., Ning, W., Peng, D., and Xue, Y. (2023) GPS 6.0: an updated server for prediction of kinase-specific phosphorylation sites in proteins. *Nucleic Acids Res.* 51, W243–W250

25. Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., and Xu, D. (2017) MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics* 33,

26. Yaron-Barir, T. M., Joughin, B. A., Huntsman, E. M., Kerelsky, A., Cizin, D. M., Cohen, B. M., Regev, A., Song, J., Vasan, N., Lin, T.-Y., Orozco, J. M., Schoenherr, C., Sagum, C., Bedford, M. T., Wynn, R. M., Tso, S.-C., Chuang, D. T., Li, L., Li, S. S.-C., Creixell, P., Krismer, K., Takegami, M., Lee, H., Zhang, B., Lu, J., Cossentino, I., Landry, S. D., Uduman, M., Blenis, J., Elemento, O., Frame, M. C., Hornbeck, P. V., Cantley, L. C., Turk, B. E., Yaffe, M. B., and Johnson, J. L. (2024) The intrinsic substrate specificity of the human tyrosine kinome. *Nature* 629, 1174–1181

27. Johnson, J. L., Yaron, T. M., Huntsman, E. M., Kerelsky, A., Song, J., Regev, A., Lin, T.-Y., Liberatore, K., Cizin, D. M., Cohen, B. M., Vasan, N., Ma, Y., Krismer, K., Robles, J. T., van de Kooij, B., van Vlimmeren, A. E., Andrée-Busch, N., Käufer, N. F., Dorovkov, M. V., Ryazanov, A. G., Takagi, Y., Kastenhuber, E. R., Goncalves, M. D., Hopkins, B. D., Elemento, O., Taatjes, D. J., Maucuer, A., Yamashita, A., Degterev, A., Uduman, M., Lu, J., Landry, S. D., Zhang, B., Cossentino, I., Linding, R., Blenis, J., Hornbeck, P. V., Turk, B. E., Yaffe, M. B., and Cantley, L. C. (2023) An atlas of substrate specificities for the human serine/threonine kinome. *Nature* 613, 759–766

28. Invergo, B. M., Petursson, B., Akhtar, N., Bradley, D., Giudice, G., Hijazi, M., Cutillas, P., Petsalaki, E., and Beltrao, P. (2020) Prediction of Signed Protein Kinase Regulatory Circuits. *Cell Syst* 10, 384–396.e9

29. Piwnica-Worms, H., Saunders, K. B., Roberts, T. M., Smith, A. E., and Cheng, S. H. (1987) Tyrosine phosphorylation regulates the biochemical and biological properties of pp60c-src. *Cell* 49, 75–82

30. Pawson, T., and Scott, J. D. (1997) Signaling through scaffold, anchoring, and adaptor proteins. *Science* 278,

31. Mugabo, Y., and Lim, G. E. (2018) Scaffold Proteins: From Coordinating Signaling Pathways to Metabolic Regulation. *Endocrinology* 159, 3615–3630

32. Sugiyama, N., Imamura, H., and Ishihama, Y. (2019) Large-scale Discovery of Substrates of the Human Kinome. *Sci. Rep.* 9, 10503

33. 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.*,

34. Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 4, 1633–1649

35. Nováček, V., McGauran, G., Matallanas, D., Vallejo Blanco, A., Conca, P., Muñoz, E., Costabello, L., Kanakaraj, K., Nawaz, Z., Walsh, B., Mohamed, S. K., Vandenbussche, P.-Y., Ryan, C. J., Kolch, W., and Fey, D. (2020) Accurate prediction of kinase-substrate networks using knowledge graphs. *PLoS Comput. Biol.* 16, e1007578

36. Wang, C., Xu, H., Lin, S., Deng, W., Zhou, J., Zhang, Y., Shi, Y., Peng, D., and Xue, Y. (2020) GPS 5.0: An Update on the Prediction of Kinase-specific Phosphorylation Sites in Proteins. *Genomics Proteomics Bioinformatics* 18, 72–80

37. 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, D512–20

38. Ochoa, D., Jarnuczak, A. F., Viéitez, C., Gehre, M., Soucheray, M., Mateus, A., Kleefeldt, A. A., Hill, A., Garcia-Alonso, L., Stein, F., Krogan, N. J., Savitski, M. M., Swaney, D. L., Vizcaíno, J. A., Noh, K.-M., and Beltrao, P. (2020) The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* 38, 365–373

39. Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J. T., Gatza, M. L., Wilkerson, M., Perou, C. M., Yellapantula, V., Huang, K.-L., Lin, C., McLellan, M. D., Yan, P., Davies, S. R., Townsend, R. R., Skates, S. J., Wang, J., Zhang, B., Kinsinger, C. R., Mesri, M., Rodriguez, H., Ding, L., Paulovich, A. G., Fenyö, D., Ellis, M. J., Carr, S. A., and NCI CPTAC (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534, 55–62

40. GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330

41. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Pontén, F. (2015) Proteomics. Tissue-based map of the human proteome. *Science* 347, 1260419

42. Ho, T. K. (1995) in *Proceedings of 3rd International Conference on Document Analysis and Recognition* (IEEE), pp 278–282 vol.1.

43. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830

44. Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics* 21, 3940–3941

45. Ochoa, D., Jonikas, M., Lawrence, R. T., El Debs, B., Selkrig, J., Typas, A., Villén, J., Santos, S. D., and Beltrao, P. (2016) An atlas of human kinase regulation. *Mol. Syst. Biol.* 12, 888

46. Akhmedov, M., Kedaigle, A., Chong, R. E., Montemanni, R., Bertoni, F., Fraenkel, E., and Kwee, I. (2017) PCSF: An R-package for network-based interpretation of high-throughput data. *PLoS Comput. Biol.* 13, e1005694

47. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423

48. Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 33,

49. Chang, W., Cheng, J., Allaire, J., Sievert, C., and Schloerke, B. shiny: Web Application Framework for R. 2021. R package version 1.6. 0. *Reference Source*,

50. Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., and Ma'ayan, A. (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–7

51. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., and Yu, G. (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov. J.* 2,

52. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29

53. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X., and Jensen, L. J. (2015) DISEASES: text mining and data integration of disease-gene associations. *Methods* 74, 83–89

54. Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016) mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* 8, 289–317

55. Hartigan, J. A., and Wong, M. A. (1979) Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* 28, 100–108

56. Rousseeuw, P. J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65

57. Schwarz, G. (1978) Estimating the Dimension of a Model. *Ann. Stat.* 6, 461–464

58. Cohen, P. (2001) The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. *Eur. J. Biochem.* 268, 5001–5010

59. Lussana, A., and Petsalaki, E. (2024) PhosX: data-driven kinase activity inference from phosphoproteomics experiments. *bioRxiv*, 2024.03.22.586304

60. Halasz, M., Kholodenko, B. N., Kolch, W., and Santra, T. (2016) Integrating network reconstruction with mechanistic modeling to predict cancer therapies. *Sci. Signal.* 9, ra114

61. 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