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

Borgthor Petursson, Evangelia Petsalaki*
*correspondence should be addressed to: petsalaki@ebi.ac.uk

Abstract

Phosphorylation forms an important part of the signalling system that cells use for decision making and regulation of processes such as celll 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 paper, we have created a machine learning-based model to derive a probabilistic kinase-substrate network from omics datasets. We show that our methodology displays improved performance compared to other state of the art kinase-substrate predictions, and provides predictions for more kinases than most of them. 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.

Introduction

Cells relay information through intricate 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[4]. 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] 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], 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.

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[12]. However 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[13]. 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[14], for phosphatases removing phosphosites to shut signals off, regulation of transcription factor translocation or activities[15] and others.

To address these points, in this work, we have extended this machine learning model introducing additional predictors to include non-kinase substrates and to provide predictions at the phosphosite level. We validate our resulting predictions with recent, independent, experimental kinase-substrate predictions and present a list of highly probable novel kinase-substrate relationships[16,17]. Our method, called SELPHI2.0, 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[18–22]. Furthermore, we provide predictions of the sign of the kinase-substrate interactions. Our network's precision, finally, increases when fitted to experimental data, suggesting that it can be used as a prior in data-driven inference of cell signalling networks from phosphoproteomics data.

Materials and methods

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 phosphosites was gathered from PhosphoSitePlus[23] (Downloaded on the 2nd May 2021). Functional scores and predictive features on phosphosites were downloaded from a recent work by Ochoa and colleagues[24]. A list of features that are considered and included in the predictions can be found in Supplementary table 1. Proteomics data sets for co-regulation were downloaded from Mertins and colleagues[25] as well as Hijazi and colleagues[16]. Expression data was gathered from the GTEX[26] (Downloaded on the 26. April 2018) and Human Protein Atlas[27] (Downloaded on the 1. December 2017). Experimentally predicted kinase-substrate relationships were downloaded from two recent publications[16,17].

Predictions were made between 368 kinases, for which we had previously generated Position weight matrices (PWMs)[12], and 80,234 phosphosites found on 9,180 proteins that were listed in PhosphositesPlus[23] and had a functionality score[24] assigned to them.

As a positive training set we used 5,251 kinase-phosphosite relationships extracted from PhosphositePlus[23] (Downloaded 2. May 2021). As there are no databases with information on known negative kinase-substrate pairs and given the fact that biological networks tend to be sparse, a random sample of kinase-substrate relationships ten times as large as the positive set was used as a negative training set. For the prediction of the regulatory sign of kinase-substrate relationships we used 673 activating interactions and 497 inhibiting interactions extracted from the SIGNOR[38] database and focused only on phosphosites likely to lead to functional changes in the protein, i.e. those with a functional score >0.5[24].

We created and acquired 49 predictors considering different features 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 weight matrices[12] (Supplementary table 1). In the case of kinase substrate prediction,to select the most useful predictors we evaluated the performance of different feature combinations using 100 training/testing datasets as described in Supplementary Information. The predictors used in the final model are annotated in Supplementary Table 1. To select the best predictors for the prediction of the sign of kinase substrate relationships a single set containing all activating and inhibiting interactions was used as a training set.

For training the model we used the random forest algorithm[28] as implemented in the *scikit-learn* python library[29] was trained with the positive set and a random sample one hundred times. Parameters were tuned in each run with grid parameter search. Different parameters were considered at each run which were listed in the section above and for each of the one hundred runs the best parameter sets were used (Supplementary Information). 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*[29] python library was used to keep the portion between negatives and positives in each split the same. The average probability across the different 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 studies[16,17] 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[30].

Comparison with other methods

To compare SELPHI2.0 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 work of Sugiyama and colleagues[23], excluding kinase-substrate phosphorylation relationships found in PhosphoSitePlus, to limit literature-based biases. The methods used for this comparison were: PhosphoPICK[22], GPS v.5.0[20], KinomeXplorer[19], NetPhos v.3.1[21] and LinkPhinder[18]. 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. Details are provided in the Supplementary information.

Evaluation of model fit to phosphoproteomic data

In order to capture context specific signalling networks, we constructed a reference 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[12]. 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[31] was fitted to our network. We tried all

combinations of the of the following parameters of the PCSF algorithm: 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 one were used as edge prizes. 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 relationships included in SELPHI2.0 retained after the fitting was compared with the kinase-substrate relationships 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[32]. We searched for related articles in the Pubmed database. Linked publications were retrieved from NCBI Entrez Gene[33] database and publications that mention more than ten kinases were filtered out.

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 as described in the Materials and methods section. By running the prediction algorithm 100 times we achieved an average AUC of 0.96 (Figure 1A). Of the 22,250,869 predictions made, 286,392 edges were high confidence (probability > 0.5). Of those, around 1.7% were found in PhosphoSitePlus[23] while 89% of the known interactions found in our network were assigned a high (> 0.5) probability. A more comprehensive overview of different precision and recall values at different cutoffs can be seen in Supplementary table 2.

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 500 citations per phosphorylating kinase and 115 per substrate respectively, whereas our high confidence predictions the median number of citations are 69 for the kinases and 23 for the substrates. We provide proportionally more predictions between kinases and substrates with significantly lower number of citations per protein on average (Kinases: $W = 2.9 \times 10^{7}$, $p < 2.2 \times 10^{-16}$, one-sided Wilcoxon test. Substrate proteins:

W= 2.6×10^{7} , $p < 2.2 \times 10^{-16}$, 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, establishing the value of this network as a prior to explore the less studied part of the phospho-signalling network.

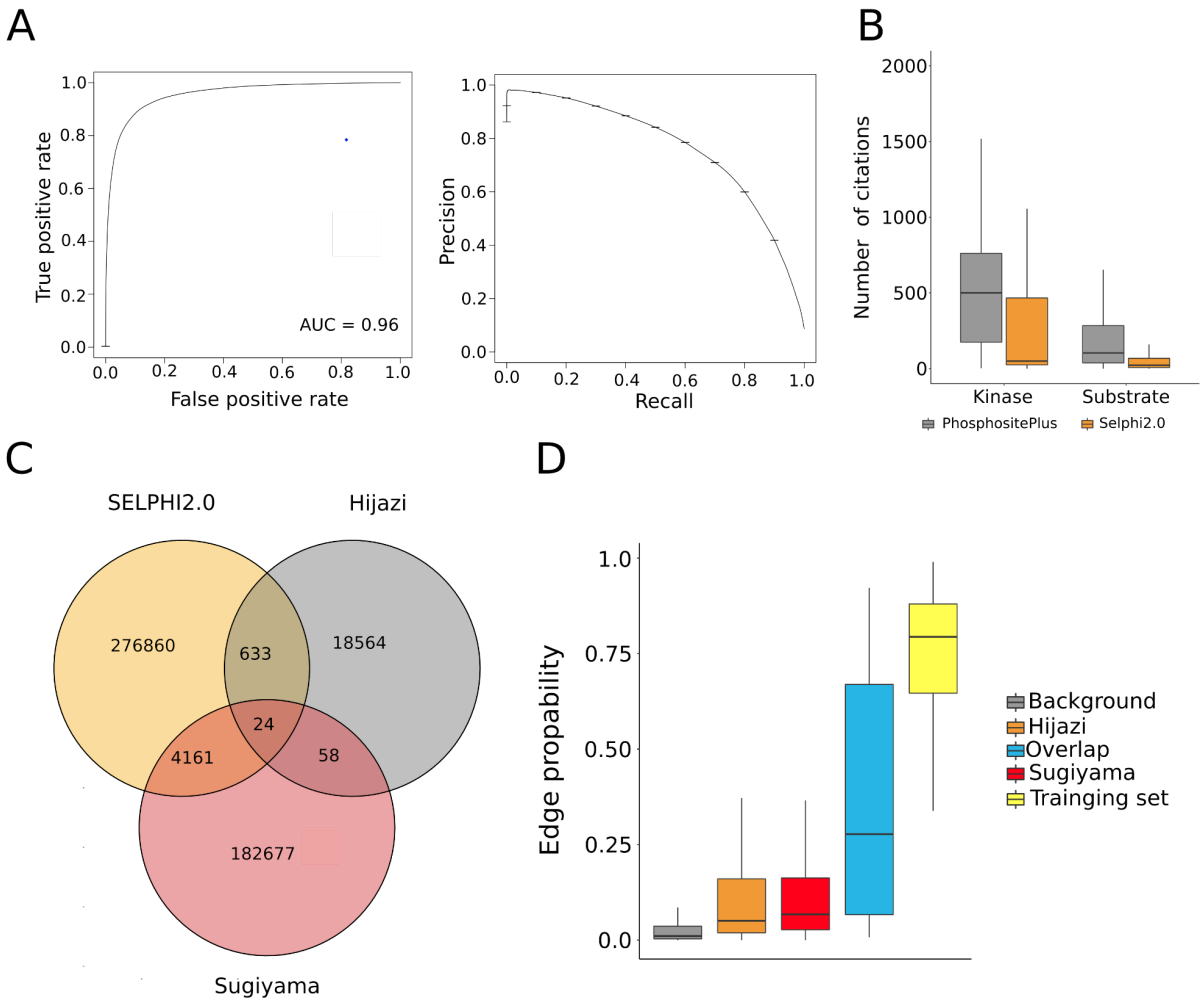

Figure 1: *Precision and recall and different cutoffs (A). The predictive power of our predictive method. With AUC under the ROC curve of 0.96 after 100 runs of cross-validation with a random set of negatives and a positive set from phosphosite plus. Precision recall curve for the same set can be seen below (B). We found that our predictor is able to make high confidence predictions for less cited kinases and less cited substrates compared to kinase-substrates found in PhosphositePlus (C). 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 (D).*

SELPHI2.0 can be used to identify 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[16,17]. In short, one publication introduces kinases to dephosphorylated peptides from cell lysis while the other data set 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 SELPHI2.0 predictions, was relatively low (Figure 1C). Due to the difference 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 in 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. The complete list of kinase-substrate predictions with indicators can be found in Supplementary table 3.

SELPHI2.0 outperforms the state-of-the-art methods for kinase-substrate prediction

In comparison to 5 state-of-the-art kinase-substrate prediction methods (PhosphoPICK[22], GPS v.5.0[20], KinomeXplorer[19], NetPhos v.3.1[21] and LinkPhinder[18]), SELPHI2.0 performs generally better on identifying known kinase-substrate interactions (Supplementary figure 1), while making predictions for more kinases, with the exception of GPS v.5.0[20] which includes 479 kinases. When comparing the performance of these methods using an independent dataset of experimentally supported kinase-substrate pairs[17], having removed those that already exist in PhosphositePlus to remove the effect of literature bias on the methods, we found that our network performs much better than the others with NetPhos3.1 performing the closest with AUC=0.69 compared to our method's 0.74, but making predictions for only 17 (38 kinase genes) kinases. KinomeXplorer achieved AUC of 0.61 compared with SELPHI2.0's AUC of 0.78 for the same set. Other methods performed close to random (Figure 2). Overview over the number of kinase substrate relationships, kinases and the size of the positive set can be seen in Supplementary table 4.

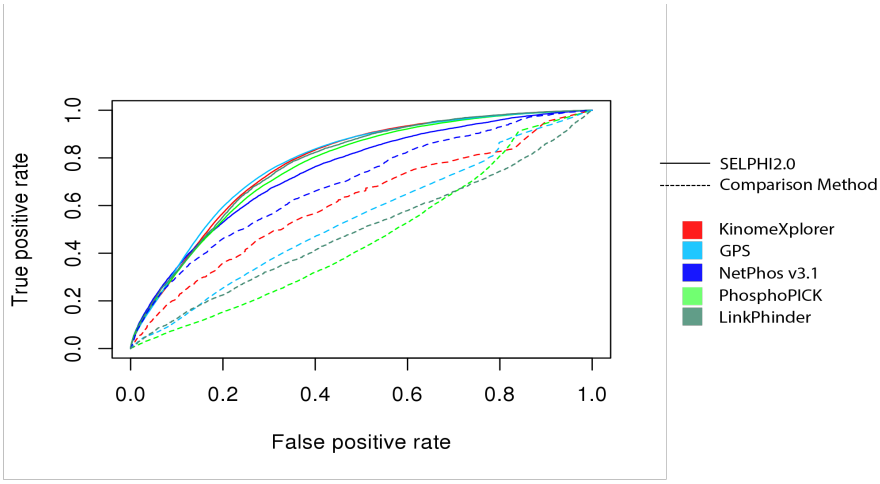

Figure 2: *Comparison between SELPHI2.0 (solid lines) and other state of the art kinase-substrate prediction methods (dashed). The ability of each method to discern between experimentally supported kinase-substrate relationships and the rest of the relationships unsupported by the experiments. In all cases SELPHI2.0 performed better than the state of the art methods.*

Prediction of regulatory status of kinase-substrate relationships

To our knowledge, earlier kinase-substrate prediction methods have not predicted the sign of the relationships. In vivo, phosphorylation may lead to functional changes in the phosphorylated protein. We wanted to capture this behaviour of kinase-substrates by predicting the sign of their interactions (Materials and methods). For this prediction we selected highly functional (functional score > 0.5) phosphosites. We found that our classifier was able to discriminate between kinase-substrate relationships that lead to activation or inhibition of the substrate with AUC of 0.83 from 10-fold cross-validation. Importantly, due to the fact that the SIGNOR training set had a large portion of the signed relationships between kinases we looked at how well the classifier discerned between kinase non-kinase-substrates and saw only a modest decrease with AUC of 0.80 from 10-fold cross-validation (Figure 3).

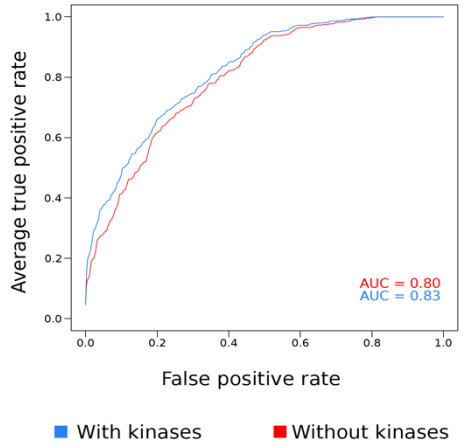

Figure 3: *The AUC from 10-fold cross-validation of signed predictions of kinase-substrates. Using signed kinase-substrate relationships found in SIGNOR as a training set we were able to make high confidence predictions. This high predictive power was retained after kinase non-kinase substrate relationships were assessed.*

Extraction of dataset-specific networks from our prior 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[34]. 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[12], forming a network of kinase-kinase regulatory relationships with phosphosites as nodes without outgoing edges (Materials and Methods). 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.

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[31]. We found that by optimizing 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 0.18 compared to the unpruned input with a F1 score of 0.033. The improvement in precision was even greater with the mean precision of the pruned subnetworks being 0.22 while the precision was 0.017 for the unpruned input networks. Both comparison yielded a significant difference (F1 score: $W = 1.72 \times 10^5$, $p < 2.2 \times 10^{-16}$, precision: $W = 1.72 \times 10^5$, $p < 2.2 \times 10^{-16}$), indicating this combination of kinase-kinase regulatory network and kinase-substrate predictions can be used to extract high probability context-specific subnetworks (Figure 4).

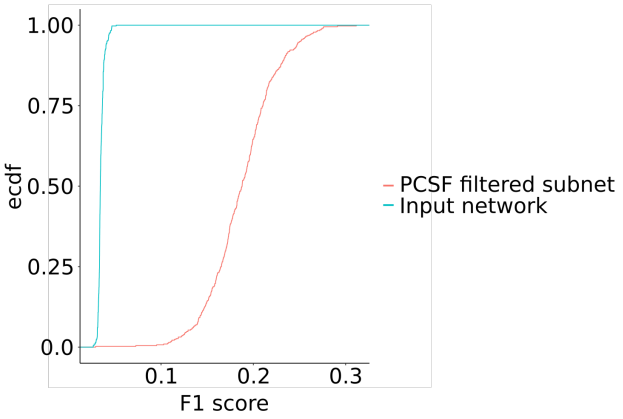

Figure 4: *A Fitting the kinase-substrate predictions to experimental data selects for known kinase-substrate relationships with median F1 score of 0.18 for the fitted sub network compared to 0.033 for unfitted input.*

Discussion

Kinase-substrate networks form the backbone of cell signalling responses and are critical for cell function in health and disease[35]. 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 368 kinases and 80,234 phosphosites.

Our method, called SELPHI2.0, performs better than the five state-of-the-art methods tested[18–22], not only on the benchmark set but importantly on entirely new experimentally supported data[17]. This is true despite including predictions for more kinases than most other methods except for GPSv5.0[20] (479), which however performs worse on the benchmark dataset and close to random in the prediction of new experimentally supported relationships(Fig. 2). This suggests that our network is a good starting point for prioritisation of new kinase-substrate relationships. 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. In addition, no other kinase-substrate prediction method, to our knowledge, provides signed relationships. We provide these for both other kinases and non kinase substrates, making the network suitable for use as a prior not only in standard network inference method, but also in studies of mathematical modelling of signal transduction.

Although we use kinase specificity in the form of PWMs as a predictor, which is a metric that inevitably is somewhat biassed by the literature, most predictors 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. NetPhos[21] 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.

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 both kinases and substrates in the network being more than 5 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.

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 SELPHI2.0 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.

References

1. Pawson, T. & Saxton, T. M. Signaling Networks—Do All Roads Lead to the Same Genes? *Cell* **97**, 675–678 (1999).
2. Hunter, T. Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell* **80**, 225–236 (1995).
3. Bhullar, K. S. *et al.* Kinase-targeted cancer therapies: progress, challenges and future directions. *Mol. Cancer* **17**, 48 (2018).

4. Needham, E. J., Parker, B. L., Burykin, T., James, D. E. & Humphrey, S. J. Illuminating the dark phosphoproteome. *Sci. Signal.* **12**, eaau8645 (2019).
5. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353–D361 (2017).
6. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
7. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966 (2016).
8. Hill, S. M. *et al.* Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods* **13**, 310–318 (2016).
9. Petsalaki, E. *et al.* SELPHI: correlation-based identification of kinase-associated networks from global phospho-proteomics data sets. *Nucleic Acids Res.* **43**, W276-282 (2015).
10. Carlin, D. E. *et al.* Prophetic Granger Causality to infer gene regulatory networks. *PloS One* **12**, e0170340 (2017).
11. Rodchenkov, I. *et al.* Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res.* **48**, D489–D497 (2020).
12. Invergo, B. M. *et al.* Prediction of Signed Protein Kinase Regulatory Circuits. *Cell Syst.* **10**, 384-396.e9 (2020).
13. Piwnica-Worms, H., Saunders, K. B., Roberts, T. M., Smith, A. E. & Cheng, S. H. Tyrosine phosphorylation regulates the biochemical and biological properties of pp60c-src. *Cell* **49**, 75–82 (1987).
14. Pawson, T. & Scott, J. D. Signaling Through Scaffold, Anchoring, and Adaptor Proteins. *Science* **278**, 2075–2080 (1997).
15. Mugabo, Y. & Lim, G. E. Scaffold Proteins: From Coordinating Signaling Pathways to Metabolic Regulation. *Endocrinology* **159**, 3615–3630 (2018).
16. Hijazi, M., Smith, R., Rajeeve, V., Bessant, C. & Cutillas, P. R. Reconstructing kinase network topologies from phosphoproteomics data reveals cancer-associated rewiring. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-019-0391-9.
17. Sugiyama, N., Imamura, H. & Ishihama, Y. Large-scale Discovery of Substrates of the Human Kinome. *Sci. Rep.* **9**, 10503 (2019).
18. Nováček, V. *et al.* Accurate prediction of kinase-substrate networks using knowledge graphs. *PLOS Comput. Biol.* **16**, e1007578 (2020).
19. Horn, H. *et al.* KinomeXplorer: an integrated platform for kinome biology studies. *Nat. Methods* **11**, 603–604 (2014).

20. Wang, C. *et al.* GPS 5.0: An Update on the Prediction of Kinase-specific Phosphorylation Sites in Proteins. *Genomics Proteomics Bioinformatics* **18**, 72–80 (2020).
21. Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S. & Brunak, S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4**, 1633–1649 (2004).
22. Patrick, R., Lê Cao, K.-A., Kobe, B. & Bodén, M. PhosphoPICK: modelling cellular context to map kinase-substrate phosphorylation events. *Bioinformatics* **31**, 382–389 (2015).
23. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512-520 (2015).
24. Ochoa, D. *et al.* The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **38**, 365–373 (2020).
25. Mertins, P. *et al.* Proteogenomics connects somatic mutations to signaling in breast cancer. *Nature* **534**, 55–62 (2016).
26. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).
27. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
28. Ho, T. K. Random decision forests. in *Proceedings of 3rd International Conference on Document Analysis and Recognition* vol. 1 278–282 vol.1 (1995).
29. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825–2830 (2011).
30. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).
31. Akhmedov, M. *et al.* PCSF: An R-package for network-based interpretation of high-throughput data. *PLOS Comput. Biol.* **13**, e1005694 (2017).
32. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinforma. Oxf. Engl.* **25**, 1422–1423 (2009).
33. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33**, D54-58 (2005).
34. Ochoa, D. *et al.* An atlas of human kinase regulation. *Mol. Syst. Biol.* **12**, 888–888 (2016).
35. Cohen, P. The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. *Eur. J. Biochem.* **268**, 5001–5010 (2001).