Phosphoproteomics data-driven signalling network inference: Does it work?

Lourdes O. Sriraja [a,1], Adriano Werhli [a,b,1], Evangelia Petsalaki [a,*]

[a] *European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Genome Campus, Hinxton CB10 1SD, UK*
[b] *Centro de Ciências Computacionais - Universidade Federal do Rio Grande - FURG, Avenida Itália, km 8, s/n, Campus Carreiros, 96203-900 Rio Grande, Rio Grande do Sul, Brazil[2]*

ARTICLE INFO

ABSTRACT

The advent of global phosphoproteome profiling has led to wide phosphosite coverage and therefore the opportunity to predict kinase-substrate associations from these datasets. However, the regulatory kinase is unknown for most substrates, due to biased and incomplete database annotations. In this study we compare the performance of six pairwise measures to predict kinase-substrate associations using a data driven approach on publicly available time resolved and perturbation mass spectrometry-based phosphoproteome data. First, we validated the performance of these measures using as a reference both a literature-based phosphosite-specific protein interaction network and a predicted kinase–substrate (KS) interactions set. The overall performance in predicting kinase-substrate associations using pairwise measures across both these reference sets was poor. To expand into the wider interactome space, we applied the approach on a network comprising pairs of substrates regulated by the same kinase (substrate-substrate associations) but found the performance to be equally poor. However, the addition of a sequence similarity filter for substrate–substrate associations led to a significant boost in performance. Our findings imply that the use of a filter to reduce the search space, such as a sequence similarity filter, can be used prior to the application of network inference methods to reduce noise and boost the signal. We also find that the current gold standard for reference sets is not adequate for evaluation as it is limited and context-agnostic. Therefore, there is a need for additional evaluation methods that have increased coverage and take into consideration the context-specific nature of kinase–substrate associations.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Signalling networks orchestrate the cell's responses to changes in their environment via stimuli-induced post translational modifications (PTMs) and are often deregulated in disease, including cancer, diabetes and other disorders [1]. Unravelling the connectivity of cell signalling networks can enable the understanding of the underlying mechanisms at play.

Protein phosphorylation is the best studied PTM regulating cell signalling networks and is mediated by protein kinases that transfer the *γ-phosphate* from ATP to Ser, Thr or Tyr residue sites on the substrate. Over 500 kinases have been identified in human cells [2], many of which are able to phosphorylate multiple proteins and sites within the same protein.

Current knowledge about downstream targets of kinases follows a Pareto optimal power law in which 20 % of the kinases accounts for 87 % of substrates in the compiled interactome databases [3]. These databases are geared towards well-studied kinases resulting in an inherent bias in which certain kinases are over-represented in pathway annotations [4]. In addition, there is limited coverage of kinase substrate binding due to the difficulty in capturing interactions that are transient in nature by affinity purification methods or Y2H assays [5]. Finally, kinase substrate interactions depend on the cell state, type, mutation background and environmental context. As such, even if all putative kinase–substrates were characterised, it is impossible to systematically characterise signalling networks across all potential contexts. There is thus a need for data-driven kinase–substrate regulatory network inference from unbiased context-specific whole-cell data, such as mass spectrometry-based global phosphoproteomics data.

To explore the wider kinome space, several kinase substrate prediction methods have been developed based on amino acid position specific scoring matrices (PSSMs) of annotated substrates [6], neural networks that combine probabilistic network from the

* Corresponding author.
   E-mail address: petsalaki@ebi.ac.uk (E. Petsalaki).
[1] These authors contributed equally to this work.
[2] Current address.

STRING database (NetworKIN) [7] or the use of Bayesian additive regression trees that incorporate kinase substrate edge prediction with the kinase mode of regulation (activating or inhibiting) on the phosphosite [8]. More recently prediction methods have evolved to incorporate a wider kinome space using knowledge graph-based embeddings of kinase consensus sites across individual kinases, kinase groups and families [9]. This supervised learning approach utilises predictive models that generate probabilistic scores for each prediction. These approaches rely on prior knowledge and therefore assessing the performance of these methods poses a challenge as vast regions of the signalling network remain unexplored. A more recent study by Ayati et al [10] integrated 9 publicly available phosphoproteome datasets across different disease contexts to identify functionally associated phosphosites in the same pathway or regulated by the same kinase using a purely data driven approach based on the co-phosphorylation metric defined in CoPhosK [11].

High throughput phosphoproteome profiling via mass spectrometry workflows have led to increased global coverage with the identification of phosphosites exceeding 30,000 per experiment [12]. As they depict a relatively unbiased snapshot of a cell's signalling state, they are, in theory, an excellent starting point for deriving data-driven and context-specific signalling networks.

Indeed, several methods have been developed, largely based on approaches already used in gene regulatory network inference from transcriptomics data. These fall in 6 different categories: pairwise association measures [13,14], linear regression [15], ordinary differential equations [16], ensemble approaches [17,18] and Bayesian networks [19]. These methods have been typically developed for specific datasets and whether they are generally suitable for phosphoproteomics data-driven network inference remains an open question. The HPN DREAM challenge benchmarked many methods for the task of inferring a causal network using dynamic RPPA (reverse phase protein array) datasets across different cell lines and perturbations (addition of inhibitor) [14]. The challenge [20] concluded that the best performing methods generally take advantage of prior knowledge networks. However, RPPA data is limited compared to global phosphoproteomics, in that there are much fewer phosphosites that are profiled, and all lie in the 'well studied' space of cell signalling given that there are available antibodies for them. Therefore, the networks that are inferred based on this data are limited in their coverage of the global cell signalling networks.

In this study we explore the potential of pairwise association and statistical dependency measures to derive unbiased signalling networks directly from global phosphoproteomics datasets, i.e., networks that don't depend on the availability of prior knowledge pathways. We focus on six pairwise and non-parametric model-based measures that are commonly used or form the basis for many other methods for data-driven inference of signalling networks from global phosphoproteomics datasets. These consist of Pearson, Spearman and Kendall correlation, mutual information, Gaussian graphical model (GGM) and FunChiSq (Methods) [21]. We provide a systematic comparison of the performance of these measures on both time series and perturbation global phosphoproteomics datasets to predict kinase–substrate associations and substrate–substrate associations (phosphosite substrates regulated by the same kinase) using a purely data driven approach (Fig. 1).

2. Results

2.1. Simulation-based analysis suggests that number and timing of phosphoproteomics samples is critical for network inference method performance

As a first step, to set the baseline of performance for the datasets we performed a simulation study. For this, we generated syn-

thetic data using a multivariate vector autoregressive model based on the Raf signalling pathway from [22], comprising 19 edges and 11 nodes (Fig. 2A). The synthetic data was simulated with varying lengths of duration and sampled intermittently across different time intervals (t = 1,2,3,4 timepoints). The performance of the six measures mentioned above were evaluated in this study to recover the underlying network.

Overall, we found that the AUC for all methods improves as the number of time points in the synthetic data increases (Fig. 2B), as expected and in line with prior published analyses [21]. The curse of dimensionality (CoD), wherein a large number of proteins/genes are measured for each sample and there are more variables (nodes in the network) than timepoints, confounds the performance of any association measure. The CoD is inherent in the synthetic data at smaller time points (t = 4,6,8), however the AUC performance begins to plateau after t = 20 once the number of samples exceeds the number of proteins.

In addition, intermittent sampling was performed on synthetic datasets with 100 time points (Fig. 2C). The data was either sampled every 1,2,3, or 4 time points for each of the 10,000 synthetically generated datasets with 100 time points. We observed a steep reduction in performance as the lag between the sampled time points increases. At lags of 3 and 4, the AUC value for all methods is close to or less than 0.5. More frequent sampling results in higher AUC values.

The same analysis was applied to time lagged data to generate pairwise associations between pairs of samples from t = 1, …, n-1 and t = 2, …, n where n is the length of the simulated time series datasets (Fig. S1). Despite the underlying time dependency from the model that generates the data, all six measures were unable to recapitulate the underlying network. Increasing the length of the time series had no effect in detecting time lagged associations, similarly intermittent sampling of time points generated AUC value close to 0.5.

For the simulated datasets, GGM had the highest performance. The GGM removes spurious connections due to the conditional dependence assumption which removes the effect of all other variables to generate highly related pairwise associations. However, it should be noted that the data is simulated as a Gaussian process and therefore the GGM's top performance in this analysis is unsurprising.

2.2. Data-driven prediction of kinase–substrate associations

We next predicted kinase–substrate associations by applying these six measures on publicly available phosphoproteomics datasets. In total 17 datasets were compiled from 11 published papers that were acquired using label-free, TMT, SILAC or iTRAQ MS based quantification [23–33] (Methods; Table S1). For the 13 time series phosphoproteome datasets, the number of sampled time points varied from four to ten. In addition, there were 4 drug perturbation datasets compiled from two studies [30,31]. Only differentially expressed phosphosites were used in this analysis (p value < 0.05; $\log_2$ fold change > 2; Methods & Table S1) and evaluation results are presented only where 5 or more true positive associations were available. Three metrics were used to evaluate the performance of the six measures. Area under the curve (AUC), ratio of the area under the precision recall curve/baseline (AURPC/baseline) and Fisher's exact odds ratio of true positive enrichment (TP) within the top 20 % of the predictions.

Each of the performance metrics were initially evaluated against an interactome-derived phosphosite specific S/T kinase substrate gold network (Methods). In general, the performance was poor with no method surpassing a mean AUC of 0.6 for the S/T phosphorylation predictions. The best performing measure in detecting direct 1:1 S/T kinase substrate associations is GGM with
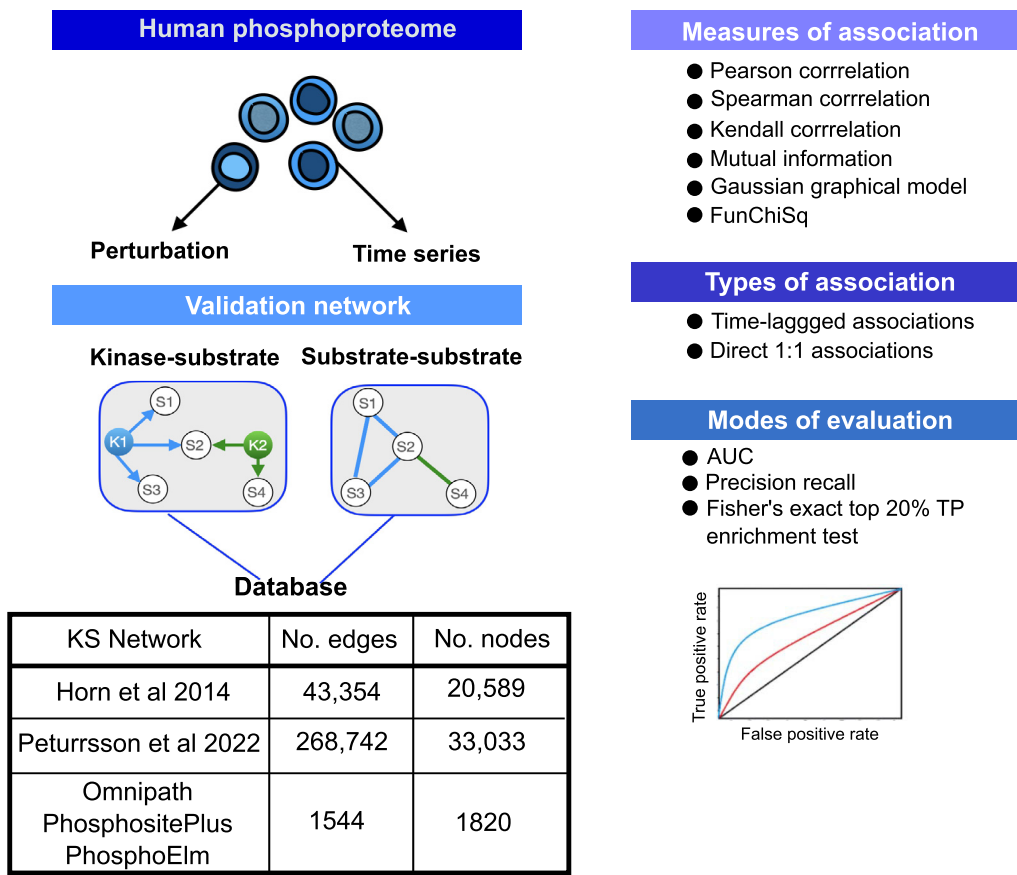
Fig. 1. Workflow of study. We used human drug perturbation and dynamic phosphoproteome datasets and evaluated the ability of 6 metrics to identify kinase substrate associations and substrate–substrate associations. For the latter a positive edge is formed between pairs of substrates that are regulated by the same kinase. A phosphosite-specific interactome was compiled from Omnipath, PhosphositePlus and PhosphoElm. The performance of six different measures of association were evaluated on 1:1 associations for time series and drug perturbation datasets, and time lagged associations for time series datasets. This was done using AUC, precision recall and an enrichment of the number of true positives (TPs) in the top 20% of predictions.
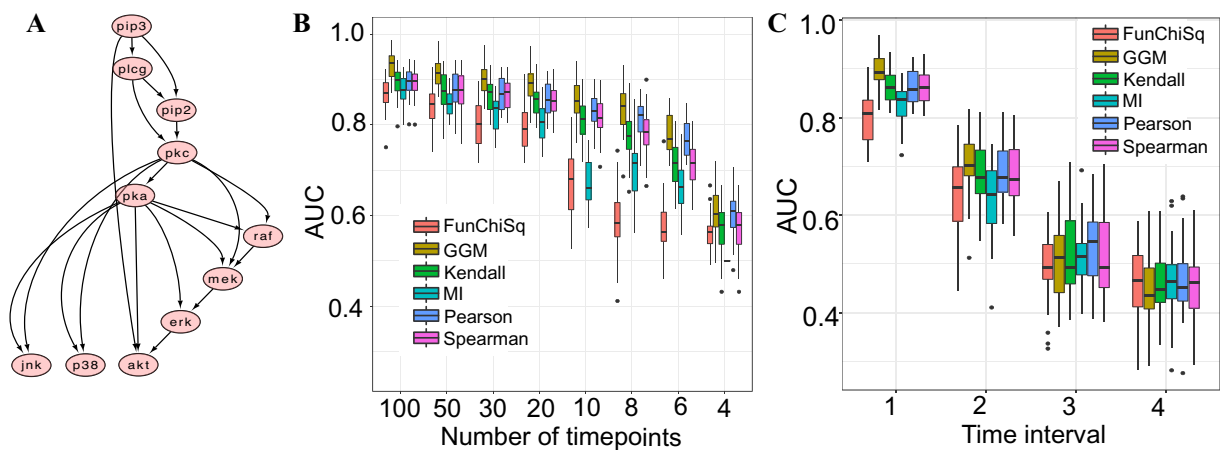

Fig. 2. Results of kinase substrate relationship predictions on simulated data A. RAF signalling pathway from [22] used to generate synthetic data from a multivariate autoregressive model B. 10,000 datasets are generated with varying lengths of duration. The box plot shows the distribution of the AUC scores for each of the predictors across varied lengths of time C. Synthetically generated data with time series length of 100, is sampled intermittently every 1, 2, 3, or 4 time intervals. The distribution of AUC for each of the measures is compared across the different time intervals.

mean AUC $\sim$ 0.6, mean $\log_{10}$(AUPRRC/baseline) = $\sim$0.7, and mean odds ratio of TP enrichment in top 20 % hits = 2.5 (Fig. 3A, Figs. S2A & S3A). All methods performed slightly worse on the perturbation datasets than the time series ones, except for the FunChiSq method which performed better, though still close to random.

For Y kinase substrate association predictions, the average performance is similar to that of the S/T ones, however there is a more bimodal distribution, whereby 2 datasets Reddy et al. 2016 - 10 nmol [29], Hijazi et al 2020 - DMSO [24]) consistently perform relatively well with AUCs around 0.7 whereas the rest mostly
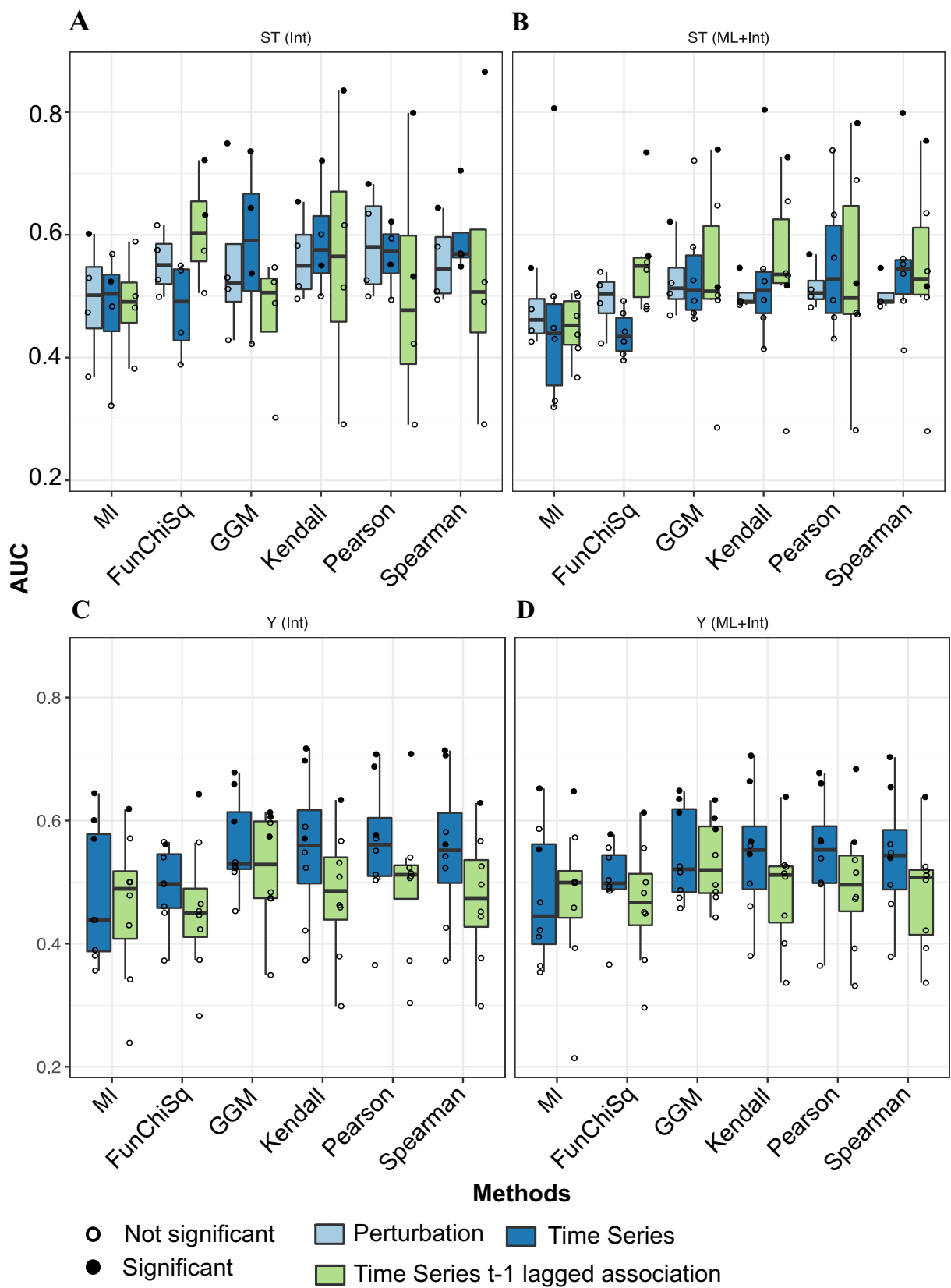
Fig. 3. Evaluation results of kinase–substrate predictions A. AUC distribution profile of S/T kinase substrate associations evaluated against database derived phosphosite specific interactions B. AUC distribution profile of S/T kinase substrate associations evaluated against the interactome and machine learning generated kinase substrate predictions C. AUC distribution profiles of Y kinase substrate associations evaluated against database derived interactome D. AUC distribution profile of Y kinase substrate associations evaluated against the interactome and machine learning generated kinase substrate predictions.

perform worse than random (Figs. 3B & S2B, S3B). Here the GGM, Kendall, Pearson and Spearman methods performed similarly with Kendall slightly outperforming the other methods.

For time lagged S/T kinase substrate associations the FunChiSq method was the best performer (∼0.6, ∼0.6, ∼2 respectively for the evaluation metrics; Fig. 3A, Figs. S2A & S3A), while GGM performed close to random. For time lagged Y phosphorylation predictions, the performance was generally random or worse than random, except for one dataset that had a higher percentage of true positives among the possible associations Reddy et al 2016 - 10nmol [29]) and achieved an AUC of 0.7 with the Pearson correlation metric (Table S2).

One of the caveats of this analysis is the low number of true positive interactions in many datasets. This meant that very few datasets were suitable for evaluation (Table S2). We hypothesised that increasing the number of true positives could help us better evaluate the performance of these methods by including additional potential interactions. We thus compiled kinase substrate networks from two machine learning methods that specialise in predicting kinase substrate associations at the phosphosite level (NetworKin [34] and Selphi2.0 [35]). These predictions were combined with the gold interactome to expand the potential space for evaluation, allowing us to include 2 additional datasets in the evaluation of S/T kinase substrate association predictions.

Using the lenient cutoff for the machine learning-inferred interactions (0.5), the 2 additional datasets (Schmutz et al. 2013 [32] and Koksal et al. 2018 [23]) performed relatively well with the best AUCs being 0.8 with Spearman and Kendall correlation method and 0.63 with the Pearson correlation method respectively. However, the remaining datasets showed on average slightly reduced performance (Fig. 3C, S2E, S3E), including the perturbation datasets, probably due to the relatively high level of noise expected to be included in predicted kinase–substrate associations. There was no effect on the performance of the Y kinase substrate association predictions (Fig. 3D, S2F, S3F).

The poor performance in detecting kinase substrate associations using pairwise measures is not entirely unexpected partially due to limited interactome coverage. We asked whether these approaches can be useful to prioritise the correct kinase for a given substrate. We thus ranked kinase substrate associations predicted using GGM and Pearson correlation (Fig. S5A, S5B) against randomised pairwise associations. A Wilcoxon-rank sum test was performed to generate p-values between the inferred and randomised distributions. For tyrosine kinase substrate associations, Pearson correlation performs best for prioritising kinases, whilst GGM generates statistically significant ranked associations for S/T. There are small distributional shifts for both measures, however the difference is not sufficient to be used as tools for kinase prioritisation.

2.3. Data-driven predicting substrates regulated by the same S/T kinase

Given the small number of true positives in kinase substrate association networks, we asked whether predicting substrate phosphosites regulated by the same kinase would yield better performance. Pairs of substrates regulated by the same kinase were combined to form a substrate–substrate association network (Fig. 1). The compiled gold phosphosite specific interactome for S/T phosphosites has 36,034 interactions, whilst the Y phosphosite interactome is limited to 6,641 interactions. For most datasets there were not sufficient Y phosphosites to evaluate the methods, so we decided to focus on S/T phosphosites.

We found that a purely statistical association-based prediction of substrate–substrate associations performs similarly to the kinase–substrate predictions above, with maximum mean AUCs achieved by GGM, Kendall and Spearman associations (∼0.6). The

performance on the time lagged exercise or the perturbation datasets was random (Fig. 4A, S5A, S6A).

We next sought to reduce the space of false positives using our knowledge that kinases tend to recognise sites that are more similar in sequence, often matching specific motifs. By comparing the sequence similarity between known interactions and random pairs at the interactome wide level we found a small but significant increase for 100k sampled substrates of S/T kinases from a median of 0.142 to a median of 0.166 (Fig. S7A). To evaluate whether the median changes based on the phosphosites specific to each dataset, we calculated the median of sampled 100k TP and TN (randomly assigned pairs of substrates) for each dataset after removing the dataset specific phosphosites (Supplementary Table S2). We thus filtered all predicted association networks by setting the association score of any substrate pairs that had a sequence similarity less than the computed dataset specific median to 0 and evaluating them as predicted negatives.

The application of sequence similarity significantly boosts the performance in generating direct substrate–substrate associations (AUCs ∼ 0.7–0.85), with all datasets becoming statistically significant based on the permutation analysis across all three-evaluation metrics (Fig. 4B, S5B, S6B). The best performance here was achieved by mutual information and the FunChiSq metrics, which had been the worst performers thus far. Some AUCs achieved greater than 0.85 (Table S2). In fact, using the sequence similarity filtered network without any association metric and evaluating the match to the gold interactome gave the best performance across datasets (Fig. 4C-D and Figs. S5 & S6B, D, F). This brings to question the value of association measures for identifying regulatory relationships. However, even across high, medium, or low scoring associations, the difference in sequence similarity between true positives and true negatives was the same. Despite the expectation that high sequence similarity is associated with high pairwise associations, this is clouded by large numbers of false positives (Fig. S7C&D). This means that regardless of how correlated or statistically associated two substrates are, the sequence similarity is equally likely to predict them as associated in all datasets and conditions. Given that the gold set itself is agnostic to context/conditions, the sequence similarity filter can be used to identify the starting network on which association methods can be applied to identify condition/context-specific associations, while reducing the search space and thus the number of false positives.

3. Discussion

A cell's response to cell communication and external perturbation is highly context specific and is regulated by cell signalling processes. Our current understanding of these processes is based on biased and context-agnostic annotated pathways which are not accurately representative. Network inference from global phosphoproteomics data is an attractive way to extract condition-specific and relatively unbiased signalling networks.

In this study, we performed a systematic comparison of measures commonly used as the basis of most network inference methods to detect kinase substrate relationships, and pairwise substrate associations regulated by the same kinase using dynamic and perturbation phosphoproteome datasets. The goal was not to be comprehensive but to identify advantages and disadvantages of different approaches and suggest the best strategies for using them.

To create a baseline of our expectation, we started with a simulated study on synthetic data that demonstrated the known potential pitfalls of inferring associations with time series data. In particular, the length of the time series determines the performance accuracy, and intermittent sampling across the simulated
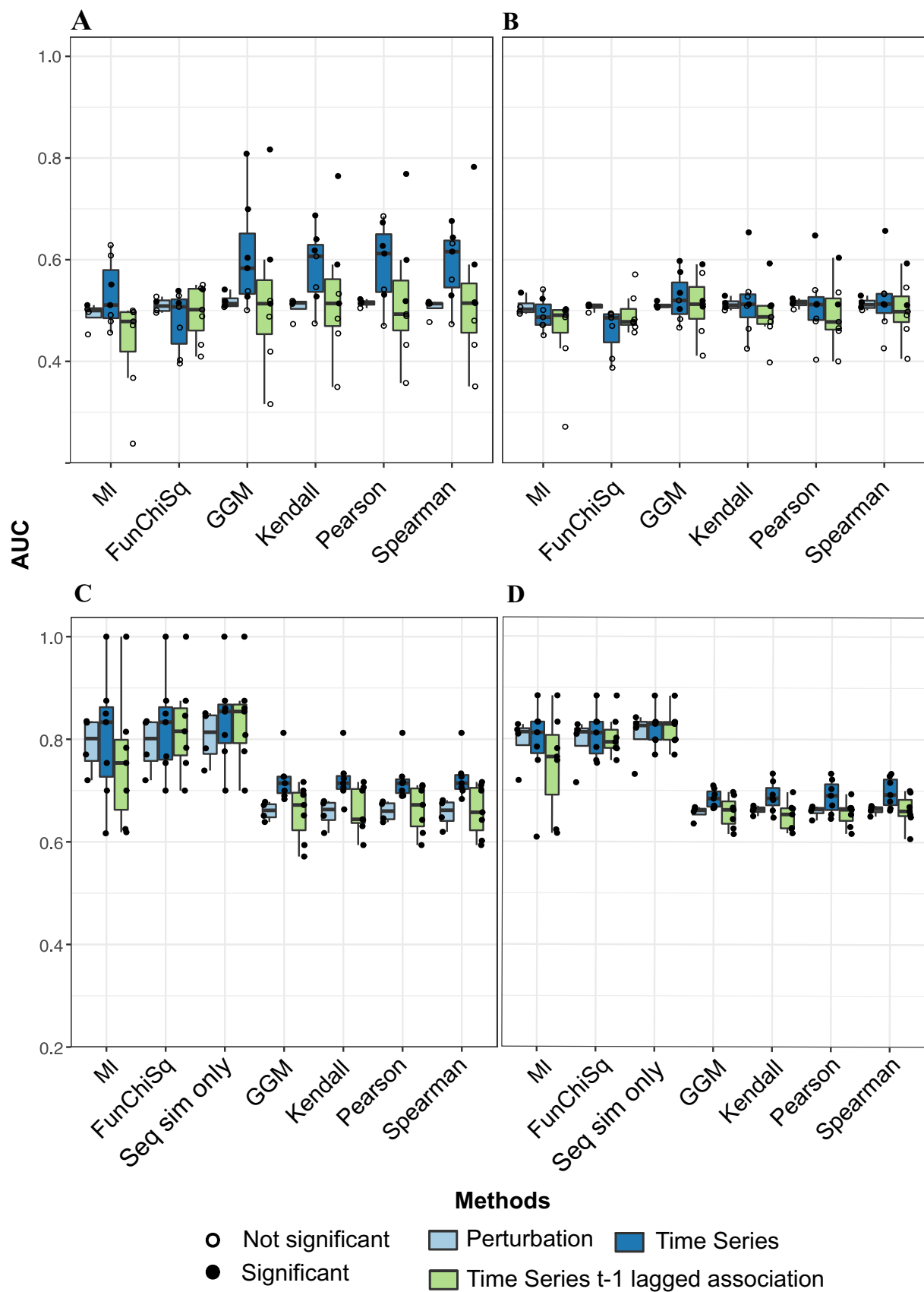
Fig. 4. Evaluation results of substrate–substrate predictions A. AUC distribution profile of substrate–substrate associations evaluated against database derived phosphosite specific interactions B. AUC distribution profile of substrate–substrate associations evaluated against the interactome and machine learning generated substrate–substrate predictions C. AUC distribution profiles of substrate–substrate associations with a sequence similarity filter evaluated against database derived interactome D. AUC distribution profiles of substrate–substrate associations with a sequence similarity filter evaluated against database derived interactome combined with ML derived substrate–substrate predictions.

data makes it difficult to recover the connections from the underlying network especially when the sampled intervals increase. This highlights the importance of time point selection when it comes to dynamic phosphoproteome profiling. Curiously, trying to identify statistical associations using a time lagged strategy was difficult both in the simulated data and in real datasets. From a biological point of view, this might be because phosphorylation reactions occur at varied timescales and the sampled timepoints don't always reflect these reactions. In particular, tyrosine phosphorylation occurs within seconds to minutes of cell stimulation [36], whilst serine/threonine phosphorylation occurs over slightly longer and more varied timescales (minutes rather than seconds).

Despite the incomplete coverage and limited ground truth to evaluate the phosphoproteomics datasets, we based the evaluation of the metrics on a gold set extracted from prior knowledge databases. Although FunChiSq was one of the best performing measures in the DREAM challenge, we found that it does not perform well in predicting associations from global phosphoproteomics datasets and is consistently the worst performing measure. Mutual information is also a poor performer for smaller time series lengths although this effect is minimised as the sample length increases in the simulated data at least. Both measures require the data to be discretized, and this is challenging for datasets that have smaller numbers of samples, due to information loss in which differences/patterns between data points can become obscured once discretized. GGM, Kendall, Pearson and Spearman correlations were typically performing similarly with GGM often performing marginally better. Throughout the study time series datasets tended to perform similarly or better than perturbation datasets, implying that collection of time series might be preferable for network inference studies.

There were instances in which the performance of certain datasets was acceptable, however overall, the performance of all metrics was poor. AUCs were in the 0.6 range and approximate $\log_{10}$ AUPR/baseline ratio of 0.7 for S/T kinase substrate predictions and only marginally higher than zero for Y kinase substrate predictions. Whilst true positive interactions were indeed associated with higher scores of associations, it was clouded by the overwhelming number of false positive hits. This is because most of these association metrics don't necessarily reflect causal relationships, even though it is expected that if they indeed are causal, they will also have high association scores. A major consideration is also the suitability of the prior knowledge-derived gold dataset that is commonly used in the evaluation of network inference methods. We observed that the majority of phosphosites had very low numbers of true positives available for discovery which resulted in only 4–5/13 of our time series datasets being used in the evaluation. This reflects the vast dark space of the human signalling networks [3,4], and makes evaluating connectivity using noisy phosphoproteome datasets challenging as it is impossible to know whether high scoring associations are indeed true or false positives. This was true even when we attempted to generate networks of correlated putative kinase substrates: whilst we were able to increase the number of true positives available for evaluation, the potential false positives also increased subsequently. Using machine learning based predicted kinase substrate relationships to extend our gold set resulted in the inclusion of additional datasets in the evaluation that performed relatively well, but the introduction of the noise of these predicted sets, mostly reduced the performance of datasets that already had a high coverage of true positives.

The introduction of a sequence similarity filter between associated substrates, enhanced the performance leading to AUCs in the order of 0.8 or even higher. This was only possible for S/T kinase substrates as there was no shift in the distribution of the sequence similarity of Y kinase substrates (Fig. S7A&B), possibly due to lack

of extensive motif specificity [37] or alternative substrate specificity strategies through e.g. SH2 domains [38].

The FunChiSq and mutual information were top performing metrics in par with the sequence similarity filter as a sole measure. A possible hypothesis for this boost in performance could be due to the inefficiency of the metrics as sole measures in predicting substrate–substrate associations (Fig. 4A & B), both MI and FunChiSq perhaps become non-redundant once the filter is applied. MI and FunChiSq are discretized measures of association and rely on many samples to work effectively. In order to estimate the joint distribution for MI, the underlying distribution needs to be sampled extensively especially for a 'plug-in' MI estimator based on the frequency of discretized samples. Insufficient number of samples leads to overestimation of MI and therefore introduces a bias in the estimate of statistical dependency between two variables [39–41]. Although the number of samples was limited for time series datasets (approximately 4–8 samples) in this study, the number of samples for perturbation data was higher on the scale of approximately 40–60 samples across the 4 datasets. However, the performance did not improve for the perturbation datasets, and the AUC distribution remained around 0.5 (Fig. 4A & B).

Recently, a method that performs dual data and phosphosite motif clustering showed promising results in network inference from phosphoproteomics data [42]. This corroborates the value of introducing a sequence similarity-based feature in network inference strategies and highlights the importance of reducing the noise in a dataset before applying network inference methods. In addition, efforts have recently been made to identify the substrate recognition motifs for all kinases [43,44], which will further improve our ability to reduce noise and match kinases to the correct substrates. Additional ways to reduce noise in the dataset could include filters to include phosphosites that are predicted to be functional [45], and introducing very strict p-value and fold-change cut-offs for the identification of differentially abundant phosphosites. Interestingly the filtered sequence similarity metric as a sole measure performed better than in combination with an association metric. This reveals an additional issue with the current ways of evaluating these methods using prior knowledge gold sets, as they are context-agnostic and represent a sum of putative relationships that have been observed in vastly different cell types, stimuli, and other conditions. Therefore, any method to identify associations in each phosphoproteomics dataset assumes that these associations exist if the relevant proteins are there, regardless of the conditions. However, this assumption is not always true and could contribute towards the apparent poor performance of the methods.

Indeed, we found that the sequence similarity filter metric was similar across all levels of association scores (Fig S7C & D), even though true positives tend to have higher such scores. Thus, it seems suitable to combine association metrics that provide the data-specific context, with strategies to reduce noise such as such functionality, sequence similarity or motif-based filtering to extract data driven and context specific kinase–substrate networks from phosphoproteomics datasets.

In conclusion, network inference approaches based on the six association metrics evaluated in this study don't perform well when used as a sole measure and evaluated against prior knowledge-derived gold sets of kinase substrate relationships. However, methods that can a priori reduce the noise and boost the signal, such as a sequence similarity filter boosts the performance substantially.

A major issue when evaluating these methods is the sparsity of the gold set. There are several efforts underway to shed light on the dark space of kinase substrate regulatory networks [3,8,35,43,44,46,47], which should largely mitigate the limitation of low coverage in prior knowledge-derived gold sets to evaluate

network inference methods. However, these gold sets will remain context agnostic, making it difficult to identify the true performance of network inference methods. There is thus a need for approaches that combine performance metrics on more comprehensive gold sets, with evaluation based on orthogonal datasets.

4. Materials & methods

4.1. Selection of datasets and methods to evaluate

We selected the kinase substrate association methods to evaluate in this study as follows: The most commonly utilised pairwise expression measure omic-wide is Pearson correlation, which is used to generate co-expression networks of functionally relevant proteins/genes [14,48]. Rank based Spearman and Kendall correlation were found to be the best performing measures in identifying co-expressed genes from the same pathway and were therefore selected as a comparative measure in this analysis [49]. Mutual information-based methods aim to capture nonlinear dependence and have been shown to perform well in recapitulating gene networks [50,51] and frequently used to reconstruct gene regulatory networks. FunChiSq is a more recently derived directional statistical test of association based on the asymptotic null chi-squared distribution [52]. As part of the HPN-DREAM challenge [20], FunChiSq was highly ranked in detecting protein associations from RPPA experimental data as well as the top performer in predicting associations from simulated model data. Finally, the partial correlation based Gaussian graphical model (GGM) is able to detect the association of two genes/proteins given this pair are conditionally dependent on the remaining genes/proteins in the data [53].

All phosphoproteome datasets were selected based on being derived from human cell lines to ensure wider interactome coverage, and sufficient number of samples for evaluation of associations. For dynamic time series datasets, the data needed to be sampled across 4 or more time points, the perturbation datasets used in this analysis had sufficient samples for evaluation. Both the time series and perturbation datasets were filtered for log2 fold change in expression in relation to control or t = 0.

4.2. Data processing

Individual datasets (phosphoproteome and proteome) were processed differently depending on the MS profiling approach used. TMT datasets were column median and row mean normalized to minimise batch effect variations across the replicates. SILAC datasets were normalised to control samples as ratios. Label free datasets were available pre-processed. In addition, some of the published datasets did not have replicates whilst other datasets only included the average intensity across the replicates. The detailed explanation on data processing specific to each dataset is provided in Table S1.

Post-processing batch effects for data with biological or technical replicates were tested using the procedures outlined in Nusinow et al. 2020 [54] (PCA, Hierarchical clustering and linear models). Datasets that exhibited batch effects post-processing were not included in this analysis. Dynamic phosphoproteome datasets with time points that exceeded 1 h were normalized to the corresponding proteome as phosphoproteome to proteome ratios.

The spline based linear mixed model (LMMS package in R) package in R was used to identify differentially expressed phosphosites across time series datasets. Only datasets from human cell lines or patients were analysed in this study.

Differential expression analysis was performed using the *lmmsDE* function in R for datasets that include replicates [55],

where a linear mixed model spline is fit to the data and compared to a null model that fits to the intercept. The function returns Benjamini-Hochberg corrected p-values and an adjusted p-value threshold of 0.05 was used to filter for significant profiles. The phosphosites across replicates were averaged and an additional fold change (FC) filter to each phosphosite response is applied in which phosphosite log2 ratios to t = 0 has at least one time point with |log2 FC| > 2. Published datasets that only included the average phosphosite signal or datasets with no replicates were filtered with |log2 FC| > 2. This was to ensure that only actively regulated phosphosites were included to test for associations and to reduce the false positives.

Two perturbation datasets were also used as a comparison. The published perturbation datasets from Wilkes et al 2015 and Hijazi et al. 2020 [30,31] were already pre-processed as log2 ratios with respect to controls and each phosphosite in each condition has an associated p-value. The perturbation datasets were filtered based on the following: each phosphosites that had 20 % or more p-values < 0.05 and 2 or more phosphosites with |log2 FC| > 2.

4.3. Association measures

Pearson correlation coefficients are computed from continuous data and capture linear associations. Given two time series $x = (x_1, \cdots, x_k)$ and $y = (y_1, \cdots, y_k)$ each of length k, the Pearson correlation coefficient between these variables is:

$$Pearson(x,y) = \frac{\frac{1}{k}\sum_{i=1}^{k}(x_i - x)^-(y_i - y)^-}{\left(\sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - x^-)^2}\right)\left(\sqrt{\frac{1}{k}\sum_{i=1}^{k}(y_i - y^-)^2}\right)}$$

where $\bar{x}$ and $\bar{y}$ are the empirical means of x and y respectively.

Kendall's Tau association is a non-parametric rank-based measure between two variables that is able to capture nonlinear dependencies. Given x_i and y_i are unique, any pair of observations (x_i, y_i) and (x_j, y_j) are classified as concordant if the ranks of both elements agree, i.e., $rg(x_i) = rg(x_j)$ and $rg(y_i) = rg(y_j)$. Where $rg(.)$ denotes the rank and is discordant if they do not agree. The number of concordant and discordant pairs is then employed to calculate the Kendall association score:

$$Kendall(x,y) = \frac{Nro\ concordant - Nro\ discordant}{\binom{n}{2}}$$

Spearman correlation coefficient is a non-parametric rank correlation score between two variables. Spearman correlation is obtained by calculating the Pearson correlations between the rank of the variables:

$$Spearman\ (x,y) = \frac{\frac{1}{k}\sum_{i=1}^{k}(rg(x_i) - rg(x^-))(rg(y_i) - rg(y^-))}{\left(\sqrt{\sum_{i=1}^{k}(rg(x_i) - rg(x^-))^2}\right)\left(\sqrt{\sum_{i=1}^{k}(rg(y_i) - rg(y))^2}\right)}$$

Mutual information is a non-parametric measure of linear or nonlinear association derived from information theory. Both x and y variables are discretized, where $p(x = i, y = i)$ is the joint probability of x and y. Given that variables x and y are discretized in r levels. The MI between these variables can be defined as:

$$MI(x,y) = \sum_{i=1}^{r}\sum_{j=1}^{r}P(x = i, y = j)log\frac{P(x = i, y = j)}{P(x = i)P(y = j)}$$

The functional Chi-squared test is a more recent measure of non-parametric causal dependency based on the normalized Chi-square test [21,52]. Each time series is first discretized by 1d k-means clustering using the Ckmeans.1d.dp function from the Ckmeans.1d.dp package in R [56]. A contingency table is generated

and used as an input into the fun.chisq.test function that generates a normalized chi square statistic as a measure of pairwise dependence. This is the only non-symmetric measure used in this study that is able to detect a directional association as it is based on the asymptotic null chi-square distribution.

In the context of relevance networks, a $n \times n$ matrix represents all the interactions between n nodes, and the entries are computed as either $Pearson(m_i, m_j), Kendall(m_i, m_j), Spearman(m_i, m_j), MI(m_i, m_j)$.

All the above measures are pairwise associations, however a Graphical Gaussian model (GGMs) is a multivariate approach with the assumption that the data is normally distributed. This forms the basis of undirected probabilistic graphical models with conditional dependencies across all the variables. GGMs are based on a (stable) estimation of the covariance matrix under the assumption that the data derives from a multivariate distribution. The element C_{ik} of the covariance matrix C is related to the correlation coefficient between nodes X_i and X_k. A high correlation coefficient between two nodes may indicate a direct interaction, an indirect interaction, or a joint regulation by a common (possibly unknown) factor. However, only the direct interactions are used to generate associations. The strengths of these direct interactions are measured by the partial correlation coefficient ρ_{ik}, which describes the correlation between nodes X_i and X_k conditional on all the other nodes in the network.

Based on the theory of normal distributions, the matrix ρ of partial correlation coefficients ρ_{ik} is the inverse of the covariance matrix C, C^{-1} (with elements C_{ik}^{-1}):

$$\rho_{ik} = \left(\frac{C_{ik}^{-1}}{C_{ii}^{-1} C_{kk}^{-1}} \right)$$

The covariance matrix is estimated using a regularized shrinkage approach outlined in Schaefer et al 2005 [53]. The R package GeneNet was used to estimate the covariance matrix (available at https://cran.r-project.org/package=GeneNet).

Direct (1:1) and time lagged associations (one unit of time) were generated for time series data, and only direct (1:1) associations were generated for the perturbation datasets.

4.4. Kinase substrate ranking

For substrates where there existed a known kinase, we calculated the rank of this true association compared to the rest of the predicted associations and compared this rank to those achieved by randomised associations. This analysis was only performed on Pearson correlation and GGM as both metrics generated associations that were on a continuous scale. Both Kendall and Spearman are rank based methods, whilst MI and FunChiSq discretize the data generating pairwise associations that are all difficult to rank.

4.5. Simulation analysis

In order to assess the performance of the various measures used in this study on a controlled dataset and to evaluate the effect of the number and timing of samples collected, a multivariate model with fixed coefficients was used to generate dynamic synthetic data. The model assumes that the data is sampled from a white noise Gaussian process. An example of a set of equations used to describe a simple three node network for nodes A, B and C $(X_A \rightarrow X_B, X_A \rightarrow X_C, X_B \rightarrow X_C)$ is shown below:

$$X_A^t = \omega_A X_A^t + \varepsilon_A$$

$$X_B^t = \omega_B X_B^t + \omega_{AB} X_A^t + \varepsilon_B$$

$$X_C^t = \omega_C X_C^t + \omega_{Ac} X_A^t + \omega_{BC} X_B^t + \varepsilon_C$$

The strength of an interaction is defined by ω and the added noise to the simulated data is defined by ε. The data simulated in this paper is based on the Raf signalling pathway from Sachs et al 2005 [22], which has 11 nodes and 19 edges (Fig. 2A). The general linear Gaussian distribution function used to generate the simulated data is based on the following equation:

$$X_i N \left(\sum_k \omega_{ik} X_k, \sigma^2 \right)$$

in which $N(.)$ denotes the normal distribution and the sum extends over all parents of node X_i, X_k represents the interacting node, with ω_{ik} representing the strength of an interaction between the interacting nodes whilst σ^2 represents noise that has been added to the data. Here, the strength of the interactions, $|\omega_{ik}|$, is sampled from the uniform distribution over the interval [0.5, 2], and the sign of ω_{ik} is randomly chosen.

4.6. Evaluation of performance

The first adjacency matrix is phosphosite specific at the substrate level and consists of kinase uniprot id's and substrate Uniprot id's with the phosphorylation site and is compiled from PhosphositePlus [57], Omnipath [58] and PhosphoElm [59].

The adjacency matrix was filtered for kinase–substrate interactions only, in which an interaction in the adjacency matrix has an entry of 1 or 0 for tyrosine kinase phosphosite rows that intersect with tyrosine phosphosites only, the remaining entries in the row are set to NA. Similarly, serine/threonine kinases have an entry of 1 (if this interaction exists in the interactome) or 0 (true negative) at the intersection to serine/threonine phosphosite substrates. Non-kinase – non-kinase interactions were assigned NAs in the adjacency matrix and were not included in the final evaluation. Performance of each association method was evaluated using receiver operating curve (ROC) precision recall (AUROC) and enrichment of true positives in top 20 % predictions using Fisher's exact test.

To overcome the limitations of understudied regions of the interactome, a kinase substrate network was compiled from two machine learning algorithms using the phosphosites across all the datasets used in this study. NetworKin from KinomeXplorer [34] uses a naive Bayes approach that combines consensus sequence motifs of kinases and their binding domains based on a model derived probabilistic score [7] with a string database protein proximity score to generate a final score that represents how likely the phosphorylation interaction would occur.

These predictions were combined with kinase substrate predictions from SELPHI2, an algorithm with increased coverage of kinases compared to NetworKin [35]. This method utilises a random forest model that uses position weight matrices for the kinases compiled from PhosphositePlus [57], coexpression from human atlas and GTEX, co-phosphorylation from CPTAC breast cancer samples and kinase inhibition profiles across different cell lines [31]. In addition, this method uses features from the phosphosite functional score from Ochoa et al 2019 [45] to generate kinase substrate predictions with a probability score of how likely the kinase will phosphorylate the substrate. The network was generated using a probability cut-off > 0.5.

Since phosphosites regulated by the same kinase have been shown to cluster together [60], substrate–substrate interactions were generated to account for associations between pairs of substrates regulated by the same kinase. These associations were compiled using the same interactome databases and machine learning derived predictions as described above. Only datasets with 5 or more TPs were used for evaluation. Due to the lack of sufficient

TPs not all of the 17 datasets were used for evaluation across the different modes of associations.

4.7. Sequence similarity calculations

Phosphosites regulated by the same kinase have been shown to cluster together in time series phosphoproteome datasets [60]. This could be due to multiple substrates in close proximity to the kinase, but also due to the specificity of certain kinases to its substrate phosphosites. In order to test kinase specificity to substrates, substrate phosphosites compiled from the interactome were divided into two groups: Y phosphopeptides and S/T phosphopeptides. TP interactions are generated by pairwise combinations of the substrates regulated by the same kinase, and TN interactions are generated from substrate phosphosite pairs that are not known to be regulated by the same kinase (Fig. S7A & B).

The sequence similarity between the substrate phosphopeptides across both sets (TPs and TNs) was calculated using the parSeqSim function from the protr package in R [61], based on the BLOSUM62 substitution matrix using a local alignment of pairs of sequence peptides each of length 15 amino acids long. The distribution of sequence similarity between pairs of substrate phosphopeptides was compared across 100,000 randomly sampled TPs from PhosphositePlus and TN interactions. For Y phosphopeptides there was an overlap between the distributions of sequence similarity, however for S/T phosphopeptide pairs there was a shift to the right in the distribution of sequence similarity for TP interactions. Due to the larger difference in medians between the TP and TN interactions, substrate–substrate associations were only calculated for S/T phosphosites. To account for differences in the measured phosphosites across all the datasets, the dataset specific median cut-off was recalculated after excluding the TPs. 100 k TP and TN interactions were then randomly sampled and the median was recalculated. The median value changed per dataset after exclusion of the TP substrate–substrate associations which led to a data specific threshold being applied. The threshold was defined as the TP median of each dataset's sampled distribution (Supplementary table S2).

Based on the median TP values in Supplementary table S2, data-specific sequence similarity thresholds were applied to filter for pairs of substrate–substrate S/T associations that exceeded the dataset specific thresholds. Associations that were less than the threshold are assigned a pairwise association value of 0 across all six measures and evaluated as a TN.

4.8. Significance testing

To assess the performance of different evaluation methods compared to a null hypothesis, an edge list that consisted of row column indices, an interaction column (1/0) and a measure of association between the indexed phosphosites was generated. The association measure column was randomly sampled without replacement 10,000 times to generate 10,000 randomly sampled networks. The performance of these randomly sampled networks was evaluated using the same procedures as the original datasets. A distribution of AUC values was generated for each of these randomly sampled edge lists that correspond to a dataset, measure and adjacency matrix (for e.g Biggelaar et al 2014 - Pearson correlation - S/T kinase substrate adjacency matrix). A p-value was generated using the following equation:

$$p - value = \frac{(Number\,of\,randomly\,sampled\,AUC\,values > AUC\,for\,dataset)}{10,000}$$

5. Code and data

All code required for reproducing this work is freely available at https://sandbox.zenodo.org/record/1100183#.YxZ229LMJhE.

CRediT authorship contribution statement

Lourdes O. Sriraja: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Adriano Werhli:** Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Funding acquisition. **Evangelia Petsalaki:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.12.010.

References

[1] Chen L, Liu S, Tao Y. Regulating tumor suppressor genes: post-translational modifications. Sig Transduct Target Ther 2020;5:1–25. https://doi.org/10.1038/s41392-020-0196-9.

[2] Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science 2002;298:1912–34. https://doi.org/10.1126/science.1075762.

[3] Needham EJ, Parker BL, Burykin T, James DE, Humphrey SJ. Illuminating the dark phosphoproteome. Sci Signal 2019;12:eaau8645. https://doi.org/10.1126/scisignal.aau8645.

[4] Invergo BM, Beltrao P. Reconstructing phosphorylation signalling networks from quantitative phosphoproteomic data. Essays Biochem 2018;62:525–34. https://doi.org/10.1042/EBC20180019.

[5] Davey NE, Seo M-H, Yadav VK, Jeon J, Nim S, Krystkowiak I, et al. Discovery of short linear motif-mediated interactions through phage display of intrinsically disordered regions of the human proteome. FEBS J 2017;284:485–98. https://doi.org/10.1111/febs.13995.

[6] Miller ML, Jensen LJ, Diella F, Jørgensen C, Tinti M, Li L, et al. Linear motif atlas for phosphorylation-dependent signaling. Sci Signal 2008;1:ra2. https://doi.org/10.1126/scisignal.1159433.

[7] Linding R, Jensen LJ, Ostheimer GJ, van Vugt MATM, Jørgensen C, Miron IM, et al. Systematic discovery of in vivo phosphorylation networks. Cell 2007;129:1415–26. https://doi.org/10.1016/j.cell.2007.05.052.

[8] Invergo BM, Petursson B, Akhtar N, Bradley D, Giudice G, Hijazi M, et al. Prediction of signed protein kinase regulatory circuits. Cell Syst 2020;10:384–396.e9. https://doi.org/10.1016/j.cels.2020.04.005.

[9] Nováček V, McGauran G, Matallanas D, Blanco AV, Conca P, Muñoz E, et al. Accurate prediction of kinase-substrate networks using knowledge graphs. PLoS Comput Biol 2020;16:e1007578. https://doi.org/10.1371/journal.pcbi.1007578.

[10] Ayati M, Yılmaz S, Chance MR, Koyuturk M. Functional characterization of co-phosphorylation networks. Bioinformatics 2022;38:3785–93. https://doi.org/10.1093/bioinformatics/btac406.

[11] Ayati M, Wiredja D, Schlatzer D, Maxwell S, Li M, Koyutürk M, et al. CoPhosK: A method for comprehensive kinase substrate annotation using co-phosphorylation analysis. PLoS Comput Biol 2019;15:e1006678. https://doi.org/10.1371/journal.pcbi.1006678.

[12] Skowronek P, Thielert M, Voytik E, Tanzer MC, Hansen FM, Willems S, et al. Rapid and in-depth coverage of the (phospho-)proteome with deep libraries and optimal window design for dia-PASEF 2022:2022.05.31.494163. https://doi.org/10.1101/2022.05.31.494163.

[13] Liu Z-P. Quantifying gene regulatory relationships with association measures: A comparative study. Front Genet 2017;8:96. https://doi.org/10.3389/fgene.2017.00096.

[14] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf 2008;9:559. https://doi.org/10.1186/1471-2105-9-559.

[15] Dong Z, Song T, Yuan C. Inference of gene regulatory networks from genetic perturbations with linear regression model. PLoS One 2013;8:e83263. https://doi.org/10.1371/journal.pone.0083263.

[16] Ma B, Fang M, Jiao X. Inference of gene regulatory networks based on nonlinear ordinary differential equations. Bioinformatics 2020;36:4885–93. https://doi.org/10.1093/bioinformatics/btaa032.

[17] Ruyssinck J, Huynh-Thu VA, Geurts P, Dhaene T, Demeester P, Saeys Y. NIMEFI: Gene regulatory network inference using multiple ensemble feature importance algorithms. PLoS One 2014;9:e92709. https://doi.org/10.1371/journal.pone.0092709.

[18] Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS One 2010;5: e12776. https://doi.org/10.1371/journal.pone.0012776.

[19] Xing L, Guo M, Liu X, Wang C, Wang L, Zhang Y. An improved Bayesian network method for reconstructing gene regulatory network based on candidate auto selection. BMC Genomics 2017;18:844. https://doi.org/10.1186/s12864-017-4228-v.

[20] Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, Carlin DE, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. Nat Methods 2016;13:310–8. https://doi.org/10.1038/nmeth.3773.

[21] Zhong H, Song M. A fast exact functional test for directional association and cancer biology applications. IEEE/ACM Trans Comput Biol Bioinform 2018:10.1109/TCBB.2018.2809743. https://doi.org/10.1109/TCBB.2018.2809743.

[22] Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science 2005;308:523–9. https://doi.org/10.1126/science.1105809.

[23] Köksal AS, Beck K, Cronin DR, McKenna A, Camp ND, Srivastava S, et al. Synthesizing signaling pathways from temporal phosphoproteomic data. Cell Rep 2018;24:3607–18. https://doi.org/10.1016/j.celrep.2018.08.085.

[24] Vemulapalli V, Chylek LA, Erickson A, Pfeiffer A, Gabriel K-H, LaRochelle J, et al. Time-resolved phosphoproteomics reveals scaffolding and catalysis-responsive patterns of SHP2-dependent signaling. Elife 2021;10:e64251. https://doi.org/10.7554/eLife.64251.

[25] Ji Q, Ding Y, Salomon AR. SRC homology 2 domain-containing leukocyte phosphoprotein of 76 kDa (SLP-76) N-terminal tyrosine residues regulate a dynamic signaling equilibrium involving feedback of proximal T-cell receptor (TCR) signaling. Mol Cell Proteomics 2015;14:30–40. https://doi.org/10.1074/mcp.M114.037861.

[26] Song Y, Subramanian K, Berberich MJ, Rodriguez S, Latorre IJ, Luria CM, et al. A dynamic view of the proteomic landscape during differentiation of ReNcell VM cells, an immortalized human neural progenitor line. Sci Data 2019;6:. https://doi.org/10.1038/sdata.2019.16190016.

[27] van den Biggelaar M, Hernández-Fernaud JR, van den Eshof BL, Neilson LJ, Meijer AB, Mertens K, et al. Quantitative phosphoproteomics unveils temporal dynamics of thrombin signaling in human endothelial cells. Blood 2014;123: e22–36. https://doi.org/10.1182/blood-2013-12-546036.

[28] de Graaf EL, Kaplon J, Mohammed S, Vereijken LAM, Duarte DP, Redondo Gallego L, et al. Signal transduction reaction monitoring deciphers site-specific PI3K-mTOR/MAPK pathway dynamics in oncogene-induced senescence. J Proteome Res 2015;14:2906–14. https://doi.org/10.1021/acs.jproteome.5b00236.

[29] Reddy RJ, Gajadhar AS, Swenson EJ, Rothenberg DA, Curran TG, White FM. Early signaling dynamics of the epidermal growth factor receptor. Proc Natl Acad Sci U S A 2016;113:3114–9. https://doi.org/10.1073/pnas.1521288113.

[30] Wilkes EH, Terfve C, Gribben JG, Saez-Rodriguez J, Cutillas PR. Empirical inference of circuitry and plasticity in a kinase signaling network. PNAS 2015;112:7719–24. https://doi.org/10.1073/pnas.1423344112.

[31] Hijazi M, Smith R, Rajeeve V, Bessant C, Cutillas PR. Reconstructing kinase network topologies from phosphoproteomics data reveals cancer-associated rewiring. Nat Biotechnol 2020;38:493–502. https://doi.org/10.1038/s41587-019-0391-9.

[32] Schmutz C, Ahrné E, Kasper CA, Tschon T, Sorg I, Dreier RF, et al. Systems-level overview of host protein phosphorylation during Shigella flexneri infection revealed by phosphoproteomics. Mol Cell Proteomics 2013;12:2952–68. https://doi.org/10.1074/mcp.M113.029918.

[33] Chiang C-K, Tworak A, Kevany BM, Xu B, Mayne J, Ning Z, et al. Quantitative phosphoproteomics reveals involvement of multiple signaling pathways in early phagocytosis by the retinal pigmented epithelium. J Biol Chem 2017;292:19826–39. https://doi.org/10.1074/jbc.M117.812677.

[34] Horn H, Schoof EM, Kim J, Robin X, Miller ML, Diella F, et al. KinomeXplorer: an integrated platform for kinome biology studies. Nat Methods 2014;11:603–4. https://doi.org/10.1038/nmeth.2968.

[35] Petursson B, Petsalaki E. Data-driven extraction of human kinase-substrate relationships from omics datasets 2022:2022.01.15.476449. https://doi.org/10.1101/2022.01.15.476449.

[36] June CH, Fletcher MC, Ledbetter JA, Samelson LE. Increases in tyrosine phosphorylation are detectable before phospholipase C activation after T cell receptor stimulation. J Immunol 1990;144:1591–9.

[37] Corwin T, Woodsmith J, Apelt F, Fontaine J-F, Meierhofer D, Helmuth J, et al. Defining human tyrosine kinase phosphorylation networks using yeast as an in vivo model substrate. Cell Syst 2017;5:128–139.e4. https://doi.org/10.1016/j.cels.2017.08.001.

[38] Miller WT. Determinants of substrate recognition in nonreceptor tyrosine kinases. Acc Chem Res 2003;36:393–400. https://doi.org/10.1021/ar020116v.

[39] Panzeri S, Treves A. Analytical estimates of limited sampling biases in different information measures. Network 1996;7:87–107. https://doi.org/10.1080/0954898X.1996.11978656.

[40] Samengo I. Estimating probabilities from experimental frequencies. Phys Rev E Stat Nonlin Soft Matter Phys 2002;65:. https://doi.org/10.1103/PhysRevE.65.046124046124.

[41] Hernández DG, Samengo I. Estimating the mutual information between two discrete, asymmetric variables with limited samples. Entropy (Basel) 2019;21:623. https://doi.org/10.3390/e21060623.

[42] Creixell M, Meyer AS. Dual data and motif clustering improves the modeling and interpretation of phosphoproteomic data. Cell Reports Methods 2022;2:. https://doi.org/10.1016/j.crmeth.2022.100167100167.

[43] Johnson JL, Yaron TM, Huntsman EM, Kerelsky A, Song J, Regev A, et al. A global atlas of substrate specificities for the human serine/threonine kinome 2022:2022.05.22.492882. https://doi.org/10.1101/2022.05.22.492882.

[44] Li A, Voleti R, Lee M, Gagoski D, Shah NH. High-throughput profiling of sequence recognition by tyrosine kinases and SH2 domains using bacterial peptide display 2022:2022.08.01.502334. https://doi.org/10.1101/2022.08.01.502334.

[45] Ochoa D, Jarnuczak AF, Viéitez C, Gehre M, Soucheray M, Mateus A, et al. The functional landscape of the human phosphoproteome. Nat Biotechnol 2020;38:365–73. https://doi.org/10.1038/s41587-019-0344-3.

[46] Berginski ME, Moret N, Liu C, Goldfarb D, Sorger PK, Gomez SM. The Dark Kinase Knowledgebase: an online compendium of knowledge and experimental results of understudied kinases. Nucleic Acids Res 2021;49: D529–35. https://doi.org/10.1093/nar/gkaa853.

[47] Moret N, Liu C, Gyori BM, Bachman JA, Steppi A, Hug C, et al. A resource for exploring the understudied human kinome for research and therapeutic opportunities. 2021:2020.04.02.022277. https://doi.org/10.1101/2020.04.02.022277.

[48] Mantini G, Vallés AM, Le Large TYS, Capula M, Funel N, Pham TV, et al. Co-expression analysis of pancreatic cancer proteome reveals biology and prognostic biomarkers. Cell Oncol 2020;43:1147–59. https://doi.org/10.1007/s13402-020-00548-v.

[49] Kumari S, Nie J, Chen H-S, Ma H, Stewart R, Li X, et al. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. PLoS One 2012;7:e50411. https://doi.org/10.1371/journal.pone.0050411.

[50] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinf 2006;7:S7. https://doi.org/10.1186/1471-2105-7-S1-S7.

[51] Lachmann A, Giorgi FM, Lopez G, Califano A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. Bioinformatics 2016;32:2233–5. https://doi.org/10.1093/bioinformatics/btw216.

[52] Zhang Y, Song M. Deciphering interactions in causal networks without parametric assumptions. ArXiv:13112707 [q-Bio] 2013.

[53] Schafer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics 2005;21:754–64. https://doi.org/10.1093/bioinformatics/bti062.

[54] Nusinow DP, Gygi SP. A guide to the quantitative proteomic profiles of the cancer cell line encyclopedia. Syst Biol 2020. https://doi.org/10.1101/2020.02.03.932384.

[55] Straube J, Gorse A-D, Huang BE, Lê Cao K-A. A linear mixed model spline framework for analysing time course 'omics' data. PLoS One 2015:10. https://doi.org/10.1371/journal.pone.0134540.

[56] Wang H, Ckmeans SM. 1d.dp: Optimal k-means clustering in one dimension by dynamic programming. R J 2011;3:29–33.

[57] Hornbeck PV, Kornhauser JM, Latham V, Murray B, Nandhikonda V, Nord A, et al. 15 years of PhosphoSitePlus®: integrating post-translationally modified

sites, disease variants and isoforms. Nucleic Acids Res 2019;47:D433–41. https://doi.org/10.1093/nar/gky1159.

[58] Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods 2016;13:966.

[59] Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, et al. Phospho.ELM: a database of phosphorylation sites—update 2011. Nucleic Acids Res 2011;39: D261–7. https://doi.org/10.1093/nar/gkq1104.

[60] Domanova W, Krycer J, Chaudhuri R, Yang P, Vafaee F, Fazakerley D, et al. Unraveling kinase activation dynamics using kinase-substrate relationships from temporal large-scale phosphoproteomics studies. PLoS One 2016;11: e0157763. https://doi.org/10.1371/journal.pone.0157763.

[61] Xiao N, Cao D-S, Zhu M-F, Xu Q-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics 2015;31:1857–9. https://doi.org/10.1093/bioinformatics/btv042.