Robust Inference of Kinase Activity Using Functional Networks

Serhan Yılmaz [1,*], Marzieh Ayati [2], Daniela Schlatzer [3], A. Ercüment Çiçek [4,5], Mark R. Chance [3,6] and Mehmet Koyutürk [1,3]

[1]Department of Computer and Data Sciences, Case Western Reserve University
[2]Department of Computer Science, University of Texas Rio Grande Valley
[3]Center for Proteomics and Bioinformatics, Case Western Reserve University
[4]Department of Computer Engineering, Bilkent University
[5]Department of Computational Biology, Carnegie Mellon University
[6]Department of Nutrition, Case Western Reserve University
[*]serhan.yilmaz@case.edu

Abstract

Mass spectrometry enables high-throughput screening of phospho-proteins across a broad range of biological contexts. When complemented by computational algorithms, phospho-proteomic data allows the inference of kinase activity, facilitating the identification of dysregulated kinases in various diseases including cancer, Alzheimer's disease and Parkinson's disease. To enhance the reliability of kinase activity inference, we present a network-based framework, RoKAI, that integrates various sources of functional information to capture coordinated changes in signaling. Through computational experiments, we show that phosphorylation of sites in the functional neighborhood of a kinase are significantly predictive of its activity. The incorporation of this knowledge in RoKAI consistently enhances the accuracy of kinase activity inference methods while making them more robust to missing annotations and quantifications. This enables the identification of understudied kinases and will likely lead to the development of novel kinase inhibitors for targeted therapy of many diseases. RoKAI is available as web-based tool at http://rokai.io.

1 Introduction

Protein phosphorylation is a ubiquitous mechanism of post-translational modification observed across cell types and species, and plays a central role in cellular signaling. Phosphorylation is regulated by networks composed of kinases, phosphatases, and their substrates. Characterization of these networks is becoming increasingly important in many biomedical applications, including identification of novel disease specific drug targets, development of patient-specific therapeutics, and prediction of treatment outcomes (Rikova *et al.*, 2007; Cohen, 2001).

In the context of cancer, identification of kinases plays a key role in the pathogenesis of specific cancers and their subtypes, leading to the development of kinase inhibitors for targeted therapy (Butrynski *et al.*, 2010; Zhou *et al.*, 2011; Perrotti and Neviani, 2013; Neviani and Perrotti, 2014). Disruptions in the phosphorylation of various signaling proteins have also been implicated in the pathophysiology of various other diseases, including Alzheimer's disease (Neddens *et al.*, 2018; Reese *et al.*, 2011), Parkinson's disease (Koyano *et al.*, 2014), obesity and diabetes (Choi *et al.*, 2010; Copps and White, 2012), and fatty liver disease (Puri *et al.*, 2008), among others. As a consequence, there is increased attention to monitoring the phosphorylation levels of phospho-proteins across a wide range of biological contexts and inferring changes in kinase activity under specific conditions.

Mass spectrometry (MS) provides unprecedented opportunities for large-scale identification and quantification of phosphorylation levels (Dephoure *et al.*, 2013). Typically, thousands of sites are identified in a single MS run. Besides enabling the characterization of the changes in the activity

of phospho-proteins, MS-based phospho-proteomic data offers insights into kinase activity based on changes in the phosphorylation of known kinase substrates (Drake *et al.*, 2012; Casado *et al.*, 2013). Observing that phosphorylation levels of the substrates of a kinase offer clues on kinase activity, Drake *et al.* (2012) use a Kolgomorov-Smirnov statistic to compare the phosphorylation distributions of substrate sites and all other phosphosites. Building on this idea, kinase substrate enrichment analysis (KSEA) (Casado *et al.*, 2013) infers kinase activity based on aggregates of the phosphorylation levels of substrates and assess the statistical significance using Z-test. Mischnik *et al.* (2015) develop these ideas further by introducing a heuristic machine learning method, IKAP, which additionally models the dependencies between kinases that phosphorylate the same substrate. Other approaches (Suo *et al.*, 2014; Ochoa *et al.*, 2016) adapt the widely-used gene set enrichment analysis (GSEA) (Subramanian *et al.*, 2005) for kinase activity inference problem. In parallel to these, a new branch of computational approaches focus on single samples to infer kinase activity (Drake *et al.*, 2016; Wilkes *et al.*, 2017; Beekhof *et al.*, 2019; Krug *et al.*, 2019).

Despite the development of algorithms that utilize relatively sophisticated models, KSEA remains one of the most-widely used tools for kinase activity inference (Wiredja *et al.*, 2017a). This can be largely attributed to the constraints posed by limited comprehensiveness of available data, prohibiting the utility of such sophisticated models. Available kinase annotations still provide very little coverage (less than 10%) for phosphosites identified in MS experiments (Needham *et al.*, 2019). The coverage of MS-based phospho-proteomics is also limited, and many sites existing in sample may be unidentified due to technical factors (Liu and Chance, 2014). Computationally predicted kinase-substrate associations (Linding *et al.*, 2007; Horn *et al.*, 2014) are successfully utilized to expand the scope of kinase activity inference (Wiredja *et al.*, 2017b). However, the coverage of computationally predicted associations is also limited (Ayati *et al.*, 2019) and most algorithms can only make predictions for well-studied kinases (Deznabi *et al.*, 2019).

With a view to expanding the scope of kinase activity inference, we develop a framework that comprehensively utilizes available functional information on kinases and their substrates. We hypothesize that biologically significant changes in signaling manifest as hyper-phosphorylation or de-phosphorylation of multiple functionally related sites. Therefore, having consistently hyper-phosphorylated (or de-phosphorylated) sites in the functional neighborhood of a phosphosite can provide further evidence about the changes in the phosphorylation of that site. Our framework, Robust Kinase Activity Inference (RoKAI), uses a heterogeneous network model to integrate relevant sources of functional information, including: (i) kinase-substrate associations from PhosphositePlus (Hornbeck *et al.*, 2015), (ii) co-evolution and structural distance evidence between phosphosites from PTMcode (Minguez *et al.*, 2012), and (iii) protein-protein interactions (PPI) from STRING (Szklarczyk *et al.*, 2014) for interactions between kinases. On this heterogeneous network, we propagate the quantifications of phosphosites to compute representative phosphorylation levels capturing coordinated changes in signaling. We develop a network propagation (Cowen *et al.*, 2017) algorithm that is specifically designed to accommodate missing sites not identified by MS. To predict changes in kinase activity, we use the resulting representative phosphorylation levels in combination with existing kinase activity inference methods.

A recent study by Hernandez-Armenta *et al.* (2017) benchmarks substrate based inference approaches using a comprehensive atlas of human kinase regulation (Ochoa *et al.*, 2016), encompassing more than fifty perturbations. Using this dataset, we systematically benchmark the improvement provided by RoKAI on the performance of a variety of kinase activity inference methods. In our computational experiments, we observe that the benchmark data is substantially biased in favor of "rich kinases" with many known substrates. Our results show that methods that appear to provide superior performance (e.g., methods that utilize statistical significance) accomplish this by increasing bias toward such rich kinases (since statistical power goes up with increasing number of observations). Motivated by this observation, we systematically evaluate the robustness of kinase activity inference methods using Monte Carlo simulations with varying levels of missingness. The results of this analysis shows that methods biased toward rich kinases are more vulnerable to incompleteness of available kinase-substrate annotations.

Next, we characterize the contribution of each source of functional information on enhancing kinase-activity inference. Our results show that incorporation of "shared kinase associations" (i.e., transferring information between sites that are targeted by the same kinase) significantly improves

kinase activity inference. We observe that, other sources of functional information considered (PPI, co-evolution and structure distance evidence) also provide statistically significant information for kinase activity inference. However, their contribution is smaller in comparison due to either (i) limited coverage or (ii) redundancy with existing kinase-substrate annotations. Finally, we systematically investigate the performance of RoKAI in improving the performance of kinase activity methods. Results of these computational experiments show that RoKAI consistently improves the accuracy, stability, and robustness of all kinase activity inference methods that are benchmarked.

Overall, our results clearly demonstrate the utility of functional information in expanding the scope of kinase activity inference and establish RoKAI as a useful tool in pursuit of reliable kinase activity inference. RoKAI is available as a web tool [1], as well as an open source MATLAB package [2].

2 Results

2.1 Robust inference of kinase activity with RoKAI

With a view to rendering kinase activity inference robust to missing data and annotations, we develop RoKAI, a network-based algorithm that utilizes available functional associations to compute refined phosphorylation profiles. We hypothesize that biologically significant changes in signaling manifest as hyper-phosphorylation or de-phosphorylation of multiple functionally related sites. Therefore, having consistently hyper-phosphorylated (or de-phosphorylated) sites in the functional neighborhood of a phosphosite can provide further evidence about the changes in the phosphorylation of that site. Conversely, inconsistency in the change in the phosphorylation levels of sites in a functional neighborhood can serve as negative evidence that can be used to reduce noise.

Based on this hypothesis, we develop a heterogeneous network model (with kinases and phosphosites as nodes) to propagate the phosphorylation of sites across functional neighborhoods. In this model, each edge has a conductance allowing some portion of the phosphorylation to be carried to the connecting nodes (illustrated in Fig. 1). Therefore, the propagated phosphorylation level of a site represents an aggregate of the phosphorylation of the site and the sites that are (directly or indirectly) functionally associated with it. Consequently, the propagated phosphorylation profiles are expected to capture coordinated changes in signaling, which are potentially less noisy and more robust.

In order to increase the coverage of network propagation, we develop an electric circuit based algorithm (Cowen *et al.*, 2017; Doyle and Snell, 1984) that is specifically designed to incorporate missing sites not identified by MS. While RoKAI does not impute phosphorylation levels for unidentified sites (i.e., it is not intended to fill in missing data), it uses these sites to bridge the functional connectivity among identified sites.

It is important to note that, we do not use network propagation to directly infer kinase activity. Rather, we use it to generate refined phosphorylation profiles that are subsequently used as input to a kinase activity inference method. Thus, the framework of RoKAI can be used together with any existing or future inference methods.

2.2 Experimental Setup

In this section, we describe our benchmarking setup for assessing the performance and robustness of kinase activity inference methods. First, we demonstrate the bias in the gold standard benchmarking data and show how this bias can lead to misleading conclusions on the performance of existing methods. Next, we introduce a robustness analysis in order to (i) overcome the effect of bias on performance estimations, and (ii) to assess the reliability of these algorithms in the presence of missing data. To characterize the value added by RoKAI, we start by assessing the utility of different sources of functional information in inferring kinase activity. Next, by focusing on a baseline kinase activity inference method (mean substrate phosphorylation), we systematically assess

[1] http://rokai.io

[2] http://compbio.case.edu/omics/software/rokai

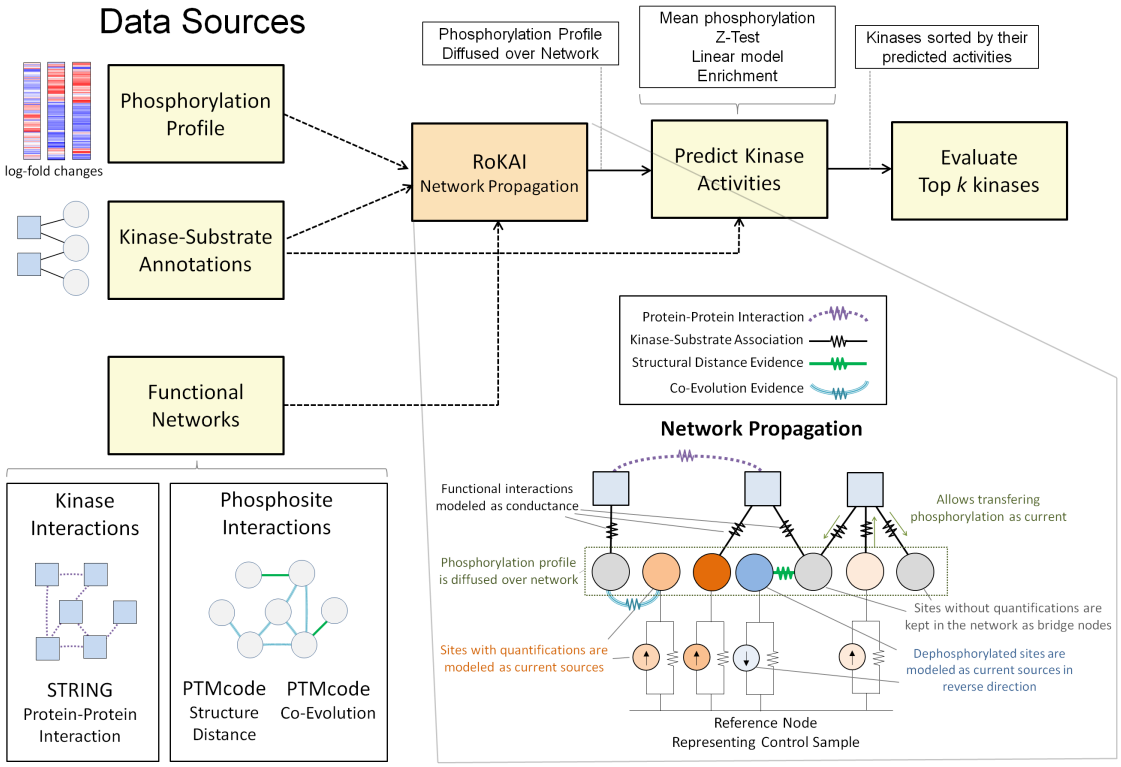

Figure 1. **The workflow and the key idea of RoKAI.** Traditional algorithms for kinase activity inference use condition-specific phosphorylation data and available kinase-substrate associations to identify kinases with differential activity in each condition. RoKAI integrates functional networks of kinases and phosphorylation sites to generate robust phosphorylation profiles. The network propagation algorithm implemented by RoKAI ensures that unidentified sites that lack quantification levels in a condition can still be used as bridges to propagate phosphorylation data through functional paths.

the incorporation of various networks with RoKAI in enhancing the accuracy and robustness of the inference. We then assess the generalizability of these results to a broad range of kinase activity inference methods. Finally, we investigate whether RoKAI's ability to incorporate missing sites in its functional network contributes to the improvement of kinase activity inference.

2.3 Benchmarking Setup

Benchmarking data: Ochoa *et al.* (2016) compiled phospho-proteomics data from a comprehensive range of perturbation studies and used these data to comprehensively benchmark the performance of kinase activity inference methods (Hernandez-Armenta *et al.*, 2017). This benchmark data brings together 24 studies spanning 91 perturbations that are annotated with at least one up-regulated or down-regulated kinase. In each of the studies, the phosphorylation levels of phosphosites are quantified using mass spectrometry. In our computational experiments, we also use this dataset to assess the robustness of existing kinase activity inference methods and validate our algorithms.

Kinase-substrate annotations: We obtain existing kinase-substrate associations from PhosphositePlus (Hornbeck *et al.*, 2015). PhosphositePlus contains a total of 10476 kinase-substrate links for 371 distinct kinases and 7480 sites. Among these annotated sites, 2397 have quantifications in the perturbation data. These sites have a total of 3877 kinase-substrate links with 261 kinases.

Benchmarking metric: The purpose of kinase activity inference is to prioritize kinases for further computational and/or experimental validation. However, in practice, it is typically costly and infeasible to experimentally validate more than a few kinases (Cichonska *et al.*, 2017). Taking this consideration into account, we use a metric, "top-k-hit", that focuses on the top-k kinase

predictions for small values of k. Since the gold standard dataset is incomplete, this metric serves as a minimum bound on the expected probability of discovering an up-regulated or down-regulated kinase if top k kinases predicted by the inference method were to be experimentally validated.

2.3.1 Existing Inference Methods

Kinase activity inference methods differ from each other in terms how they integrate the phosphorylation levels of the substrates of a kinase to estimate its activity. These methods range from simple aggregates and enrichment analyses to more sophisticated methods that take into account the interplay between different kinases. We benchmark the following commonly used inference methods:

Mean (baseline method): One of the simpliest kinase activity inference methods employed by KSEA (Casado *et al.*, 2013). This method represents the activity of a kinase as the mean phosphorylation of its substrates.

Z-score: To assess the statistical significance of inferred activities, KSEA uses z-scores, normalizing the total log-fold change of substrates with the standard deviation of the log-fold changes of all sites in the dataset.

Linear model: The linear model, considered by Hernandez-Armenta *et al.* (2017), aims to take into account of the dependencies between kinases that phosphorylate the same site. In this model, the phosphorylation of a site is modeled as summation of the activities of kinases that phosphorylate the site. A similar (but more complex) approach is also utilized by IKAP (Mischnik *et al.*, 2015).

GSEA: Suo *et al.* (2014) and Ochoa *et al.* (2016) adopt gene set enrichment analysis (GSEA), a widely used method in systems biology (Subramanian *et al.*, 2005), to infer kinase activity by assessing whether the target sites of a kinase exhibit are enriched in terms of their phosphorylation fold change compared to other phosphosites.

2.4 Bias and robustness of existing inference methods

Previous benchmarking by Hernandez-Armenta *et al.* (2017) suggests that methods that rely on statistical significance (Z-Score and GSEA) are superior to their alternatives. However, as shown in Fig. 2(a)), we observe that there is substantial bias in the benchmarking data: "rich" kinases (i.e., kinases with many known substrates) are significantly over-represented among the 25 annotated kinases that have at least one perturbation (median number of substrates: 29 for annotated and 4 for not-annotated kinases, K-S test p-value$<$3.5e-7 for the comparison of annotated kinases with others in terms of their distribution of number of substrates).

Since methods that rely on statistic significance have a positive bias for kinases with many substrates (statistical power is improved with number of observations), we hypothesize that this is the reason behind their observed superior performance. To test this hypothesis, we benchmark two additional inference methods that are artificially biased for kinases with many substrates: (i) *Sum:* Sum of phosphorylation (log-fold changes) of substrates, and (ii) *Num:* Number of substrates, used directly as the predicted activity of a kinase (clearly, this method does not use the phosphorylation levels of sites, thus, it always generates the same ranking of kinases regardless of the phosphorylation data). As shown in Fig. 2(b), methods that are artifically biased for rich kinases appear to have better predictivity over the alternatives.

In order to overcome the effect of this bias on evaluation, we perform a robustness analysis where we hide a percentage of the known substrates of the 25 annotated kinases from the inference methods. The results of this analysis are shown in Fig. 2(c). As seen in the figure, even though methods biased for rich kinases appear to have higher predictivity when all of the available kinase substrate annotations are used, they are not robust to increasing rate of missingness in kinase-substrate annotations. The performance of artificially biased methods fall below that of the low-biased methods (e.g., *Mean* and *Linear Model*) at around 50% missingness. At around 80% missingness, the effect of the bias on evaluation is mitigated i.e., the difference between number of substrates of 25 annotated kinases and the remaining kinases is not at a statistically detectable level anymore. Thus, the performance of biased (e.g., statistical significance based) methods fall below the low-bias methods at around 80% missingness. These observations make the reliability

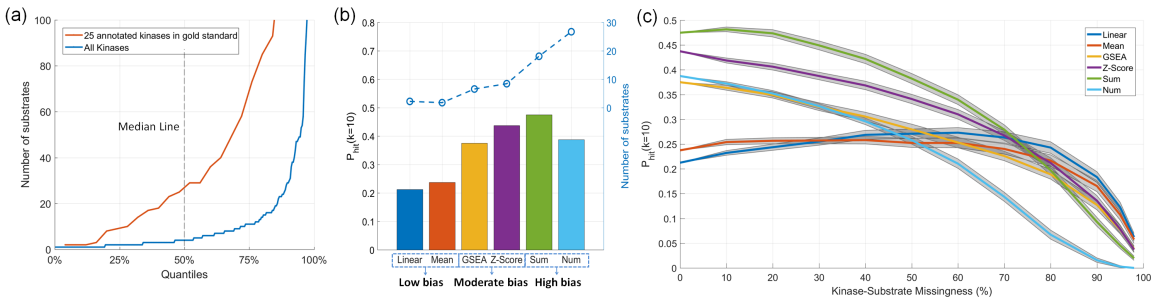

Figure 2. **Existing benchmark data for kinase-activity inference is biased toward kinases with high number of substrates and can be misleading in assessing the performance of inference methods.** (a) Inverse cumulative distribution of the number of substrates for the 25 kinases that are annotated with a perturbation in gold standard benchmarking data compared to all kinases. The x-axis indicates the quantiles. For example, the value on the y-axis that corresponds to $x = 50\%$ indicates the median number of substrates. (b) Performance and bias of baseline kinase activity inference methods. The bars show the probability of identifying an annotated "true" kinase in top 10 predicted kinases (PHit10). The dashed line indicates the average number of substrates of the top 10 predicted kinases for the corresponding method. The high-bias methods (*Sum*: total substrate phosphorylation, and *Num*: number of substrates) are not used in the literature, but are shown here to illustrate the effect of bias on performance assessment. (c) The robustness analysis of the methods for missingness in kinase-substrates links. The x-axis shows the percentage of (randomly selected) kinase-substrates links of 25 gold standard kinases hidden from the kinase activity inference methods. The gray areas indicate the 95% confidence intervals for the mean performance across 100 runs.

of biased methods highly questionable since the available kinase-substrate annotations are largely incomplete.

2.5 Utility of functional networks for inferring kinase activity

To improve the predictions of kinase activity inference methods in a robust manner, our approach is to utilize available functional or structural information. We hypothesize that phosphorylation of sites that are related to the kinase substrates (whether functionally or structurally) would be predictive of kinase activity. Specifically, we investigate the predictive ability of following functional networks:

Known Kinase-Substrates (baseline network): This network comprises of the kinase-substrate associations obtained from PhosphoSitePlus. This is the (only) network that is utilized by all kinase activity inference methods and serves as our baseline.

Shared-Kinase Interactions: Here, we consider two phosphosites to be *neighbors* if both are phosphorylated by the same kinase. We hypothesize that phosphorylation of neighbor sites of kinase-substrates would be predictive of kinase activity. Note that in RoKAI's heterogeneous functional network, there are no additional edges that represent shared-kinase interactions. Instead, RoKAI's network propagation algorithm propagates phosphorylation levels across shared-kinase sites through paths composed of kinase-substrate associations.

STRING Protein-Protein Interactions (PPI): We hypothesize that the phosphorylation levels of the substrates of two interacting kinases will be predictive of each other's activity.

PTMcode Structural Distance Evidence: We hypothesize that phosphorylation of sites that are structurally similar to a kinase's substrates will be predictive of that kinase's activity.

PTMcode Co-Evolution Evidence: We hypothesize that phosphorylation of sites that show similar evolutionary trajectories to a kinase's substrates will be predictive of that kinase's activity.

For each of the functional or structural networks described above, we compute a network activity prediction score for each kinase based on the mean phosphorylation of sites that are considered of interest for the corresponding network (illustrated in Fig. 3). Note that, except for the baseline

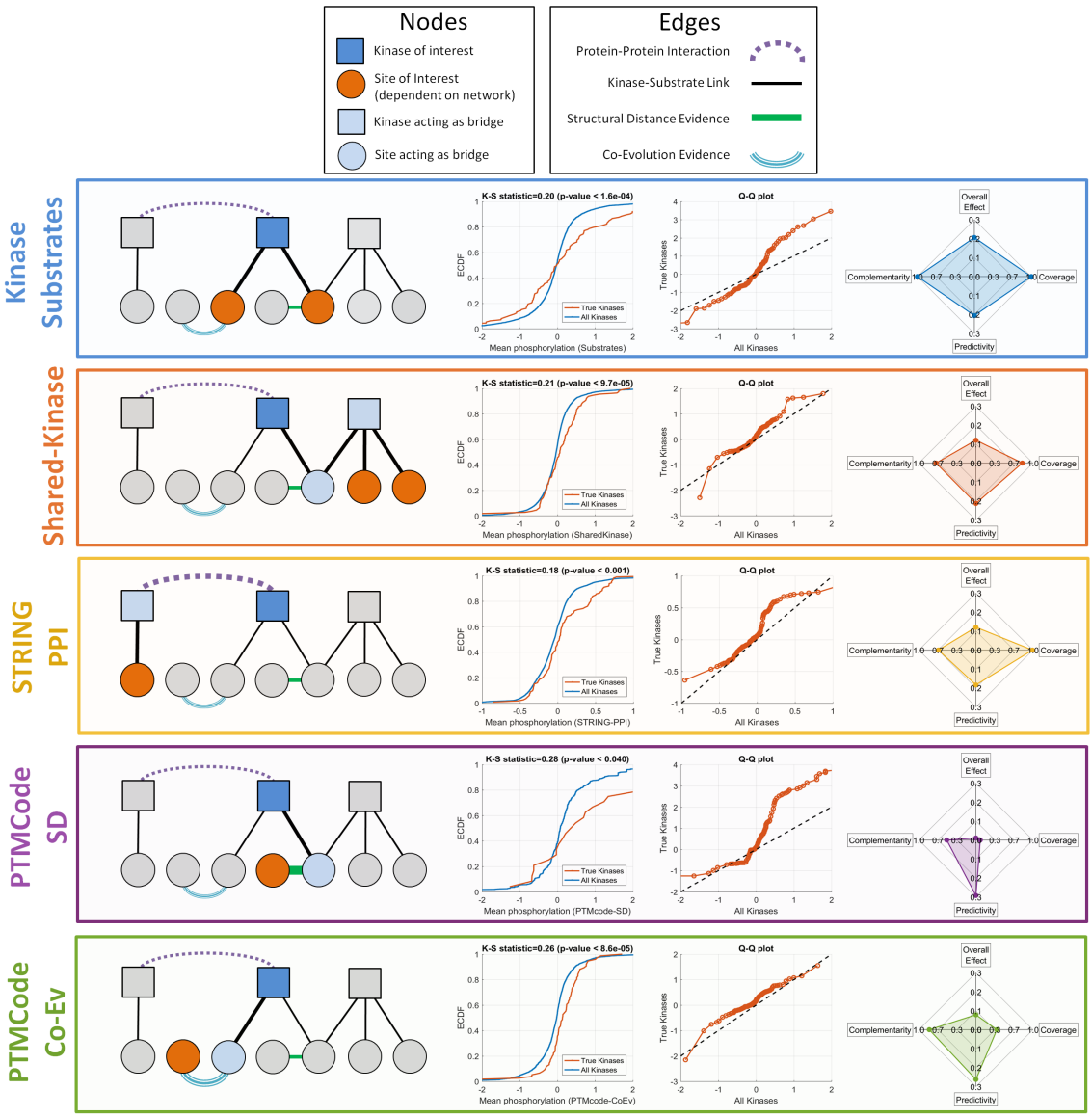

Figure 3. **Utility of available functional or structural information in providing information on kinase activity.** Each row represents a different information source. The upper-most row (Kinase-Substrates) represents the information source that is utilized by all existing kinase activity inference methods, the other four rows represent the information sources introduced here. In each row, the relationship between a kinase (blue square) and the site(s) (red circles) that provide(s) information on the activity of the kinase is illustrated. The center-left plot compares the empirical cumulative distribution (ECDF) of the phosphorylation levels of the "information-providing" sites for "true" perturbed kinases in the benchmark data against all kinases. The center-right plot compares these two distributions using a Q-Q plot. The right-most plot shows the *predictivity* (accuracy in predicting kinase activity), *complementarity* (information provided in addition to the substrates of the kinase), and *coverage* (fraction of kinases that are affected) of the information source. The positive y-axis on the right-most plot shows the overall effect of the information source calculated as the product of the scores shown on the other axes.

network (known kinase-substrates), we do not use the phosphorylation levels of the kinase's own substrates to compute the scores for each network.

To characterize the contribution of each source of functional information on enhancing kinase-

activity inference, we consider the following metrics:

Predictivity: To assess the utility of functional networks in predicting the "true" perturbed kinases in gold standard dataset, we use Kolmogorov-Smirnov (K-S) test (Massey Jr, 1951) comparing the distribution of network scores for true kinases with the distribution of all other kinases. For each functional network, we consider the K-S statistic as the *predictivity score* of the corresponding network.

Coverage: The network scores contain missing values for kinases without any edges in the corresponding functional networks. Thus, while assessing predictivity (as explained above), we utilize only the kinases with a valid network score. To take missing data into account, we compute a *coverage score* which is equal to the percentage of kinases with a valid network score with respect to that functional network.

Complementarity: We aim to utilize the functional networks as an information source that complements available kinase-substrate associations. If there is statistical dependency between functional network scores and the activity inferred by the kinase's own sites, the information provided by the network would be redundant. We use *complementarity score* as one minus absolute linear (Pearson) correlation between the score of each network scores and kinase activity inferred based on the kinase's own substrates. Since the kinase-substrate association network serves as our baseline, we consider it to have 100% complementarity.

Overall Effect: To quantify the overall contribution of the functional networks for improving the predictions of kinase activity, we combine the predicity, coverage and complementarity scores and obtain an *overall effect score*:

$$\text{Overall Effect} = \text{Predictivity} \times \text{Coverage} \times \text{Complementarity} \tag{1}$$

The results of this analysis are shown in Fig. 3. As seen in the figure, all considered functional information sources exhibit statistically significant predictivity of the kinase-perturbations: Known kinase-substrates (K-S statistic = 0.20, p-value<1.6e-4), Shared-kinase interactions (K-S statistic = 0.21, p-value<9.7e-5), Protein-protein interactions (K-S statistic = 0.18, p-value<0.001), Structure distance evidence (K-S statistic = 0.28, p-value<0.04), Co-evolution evidence (K-S statistic = 0.26, p-value<8.6e-5). We observe that the incorporation of "shared kinase associations" in addition to the known kinase substrates has the most overall contribution to the inference of kinase activities (Fig. 3, first and second rows), followed by kinase-kinase interactions (Fig. 3, third row). Even though co-evolution and structural distance networks exhibit strong predictivity, their overall contribution is relatively low due to their limited coverage and redundancy with existing kinase-substrate annotations (Fig. 3, fourth and fifth rows).

2.6 Benchmarking RoKAI-enhanced inference methods

Motivated by the utility of the functional networks for predicting kinase activity, we gradually explore a set of heterogeneous networks with RoKAI by adding sources of functional information based on their overall effect observed in the previous section:

Kinase-Substrate (KS) network: The network used by RoKAI consists only of the known kinase-substrate interactions. Use of this network allows RoKAI to utilize sites with shared-kinase interactions (illustrated in Fig. 3, 2nd row), i.e., sites that are targeted by the same kinase cTontribute to their refined phosphorylation profiles.

KS+PPI network: In addition to KS, this network includes weighted protein-protein interactions between kinases. This allows propagation of phosphorylation levels between substrates of interacting kinases (illustrated in Fig. 3, 3rd row).

KS+PPI+SD network: In addition KS+PPI, this network includes interactions between phosphosites with structural distance (SD) evidence obtained from PTMcode. This allows the utilization of sites that are structually proximate to the substrates of a kinase (illustrated in Fig. 3, 4th row).

KS+PPI+SD+CoEv (combined) network: In addition KS+PPI+SD, this network includes interactions between phosphosites with co-evolution evidence obtained from PTMcode. This allows the utilization of sites that are evolutionarily similar to the substrates of a kinase (illustrated in Fig. 3, 5th row).

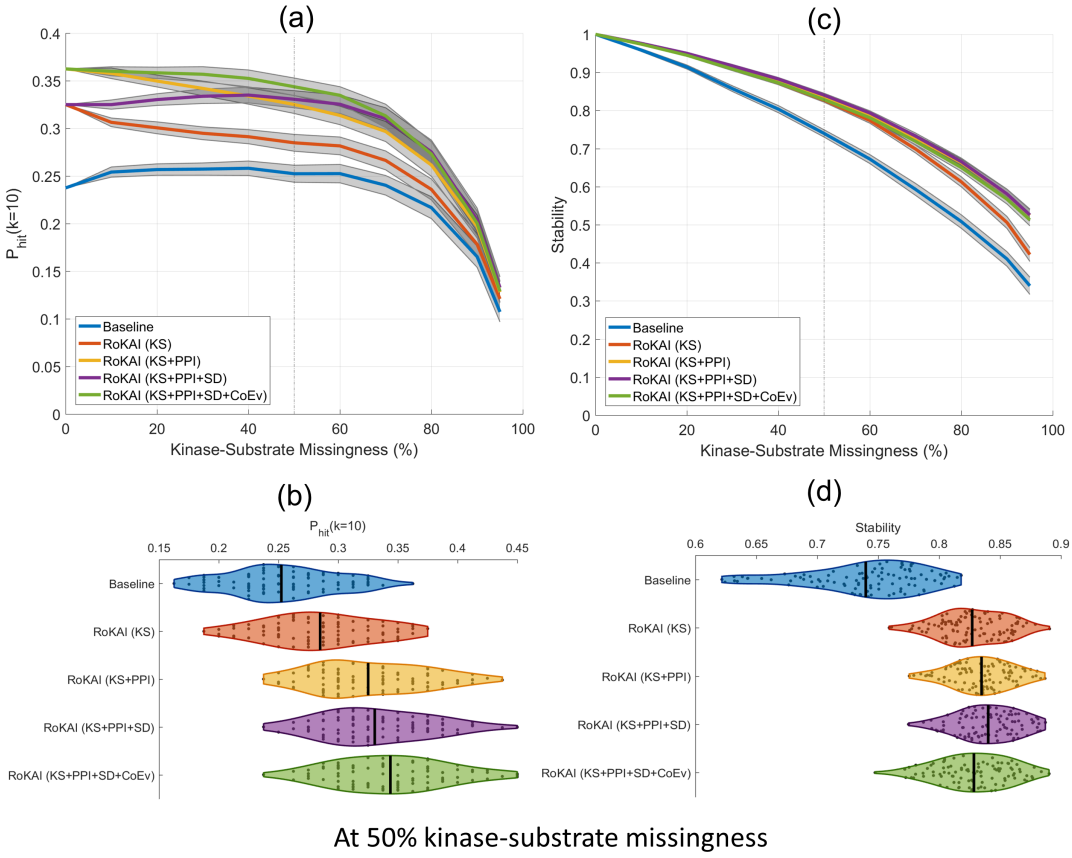

Figure 4. **Comparison of the accuracy and stability of mean substrate phoshorylation and its RoKAI-enhanced versions using various functional or structural networks.** (a) The hit-10 performance (the probability of ranking a true perturbed kinase in the top ten), as a function of missingness (the fraction of kinase-substrate associations that are hidden). (b) The distribution of hit-10 probabilities for 100 instances at 50% missingness. (c) Stability of the inferred activities (measured by the average squared correlation between inferred activities when different portions of kinase-substrate associations are hidden from the inference methods) as a function of missingness. (d) The distribution of stability for 100 instances at 50% missingness.

To assess the performance of RoKAI with these networks, we use the benchmarking data from the atlas of kinase regulation. As previously discussed, this dataset is heavily biased toward kinases with many known substrates. To overcome the effect of this bias on evaluation, we perform robustness analyses where we hide a portion of known kinase-substrate interactions of the 25 kinases that have perturbations. For predicting kinase activity, we use the mean substrate phosphorylation (baseline inference method) and compare the performance of original predictions and RoKAI-enhanced predictions. As shown in Fig. 4(a) and Fig. 4(b), RoKAI consistently and significantly ($p < 0.05$) improves the predictions in a robust manner for varying levels of missing data.

The functional networks that contribute most to the improvements in prediction performance of RoKAI are respectively: KS network (modeling shared-kinase interactions) followed by PPI (for including kinase-kinase interactions) followed by co-evolution evidence. Compared to these, including structural distance evidence in the network has a minor effect on prediction performance. This is in line with the overall effect size estimations (shown in Fig. 3). Since structural distance network has relatively small number of such edges, it provides low coverage and a minor effect size even though the existing edges are estimated to be more predictive of kinase activity compared to other networks.

To further evaluate the robustness of the predictions, we assess the *stability* i.e., the expected

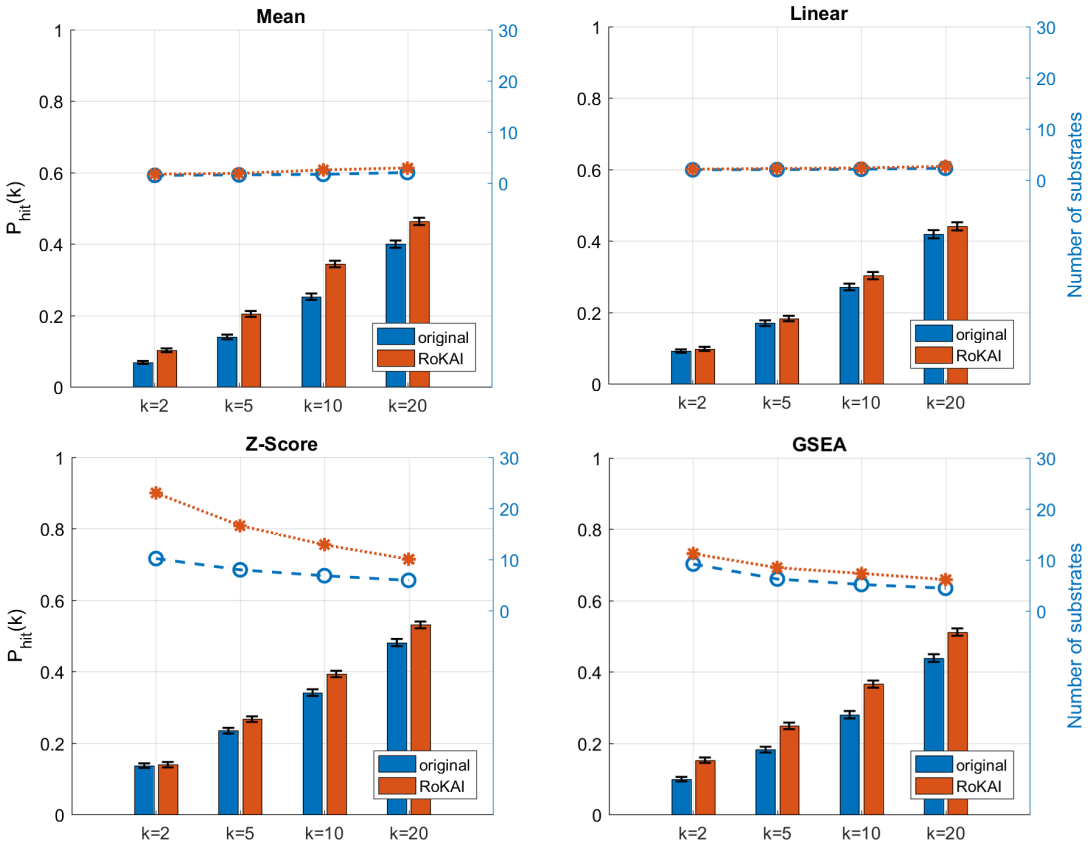

Figure 5. **Contribution of RoKAI (combined network) in improving the performance of different kinase activity inference methods for predicting the true (annotated) kinase in the top k kinase predictions for various k.** The bars show the mean probability of predicting a true kinase among the top k kinases at 50% kinase-substrate missingness. The blue bars indicate the prediction performance using the original (unmodified) phosphorylation profiles and red bars indicate the performance of using RoKAI-enhanced profiles for inferring kinase activity. The dashed lines indicate the average number of substrates of the top k kinases predicted by the corresponding inference method. The black error bars indicate the 95% confidence intervals for the mean performance across 100 runs.

degree of aggreement between the predicted kinase activity profiles when different kinase substrates are used (e.g., because some sites are not identified by a MS run) to infer the activity of a kinase. We measure the stability by computing average squared correlation between different runs of robustness analysis (where a different portion of kinase-substrate links are used for inferring kinase activity in each run). As shown in Fig. 4(c) and Fig. 4(d), predictions made by RoKAI-enhanced phosphorylation profiles are significantly ($p < 0.05$) more stable in addition to being more predictive.

2.6.1 Improvement of RoKAI over a broad range of methods

Since RoKAI provides refined phosphorylation profiles (propagated by functional networks), it can be used in conjunction with any existing (or future) kinase activity inference algorithms. Here, we benchmark the performance of RoKAI when used together with existing inference methods. For each of these methods, we use the refined phosphorylation profile (obtained by RoKAI) to obtain the RoKAI-enhanced kinase activity predictions. To assess the prediction performance while addressing the bias for rich kinases, we perform robustness analysis at 50% kinase-substrate missingness and measure the top-k hit performance for $k = 2, 5, 10$ and 20. As shown in Fig. 5, RoKAI consistently improves the predictions of all inference methods tested.

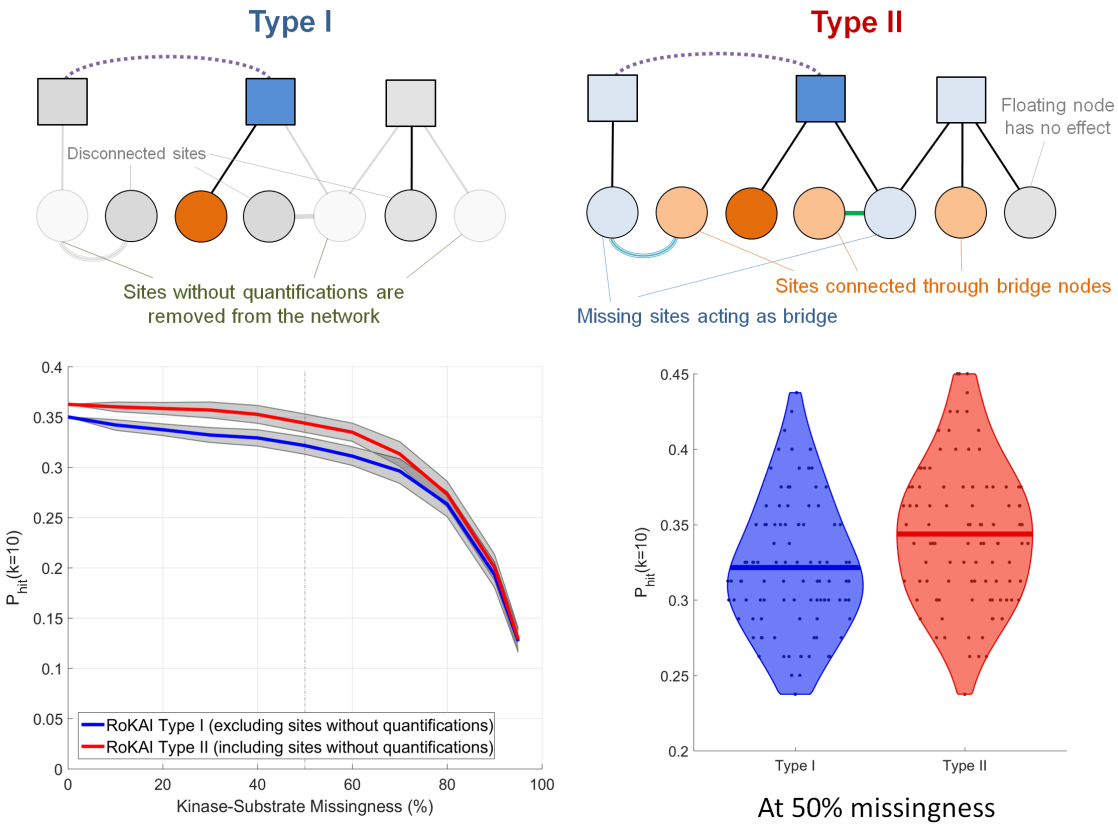

Figure 6. **RoKAI improves kinase activity inference by enabling utilization of the unidentified sites (without quantifications) for predicting the activity of kinases.** In type I (illustrated in top left), the network consists only of sites with quantifications. Whereas, in type II (illustrated in top right), the network includes sites without quantifications to utilize them as bridge nodes. (Bottom Left) Robustness analysis with respect to missingness of kinase-substrate links. The shaded area shows the 95% confidence interval for the mean performance on 100 randomized runs where different kinase-substrate links are removed. (Bottom Right) The performance of RoKAI Type I and Type II at 50% missingness. Each point indicate the performance on a different run. The colored lines indicate the mean performance.

2.6.2 Effect of incorporating unidentified sites in RoKAI

An important feature of RoKAI's network propagation algorithm is its ability to accommodate unidentified sites (i.e., sites that do not have quantified phosphorylation levels in the data) in the functional network. While RoKAI does not impute phosphorylation levels for unidentified sites (i.e., it is not intended to fill in missing data), it uses these sites to bridge the functional connectivity among identified sites. To assess the value added by this feature, we compare two versions of RoKAI: One that removes unidentified sites from the network (Type I) and one that utilizes unidentified sites as bridges (Type II). The results of this analysis are shown in Fig. 6. The kinase activity inference activity method we use in these experiments is mean phosphorylation level. As seen in the figure, retention of unidentified sites in the network consistently improves the accuracy of kinase activity inference. We observe a similar improvement for all other kinase activity inference methods that are considered.

3 Discussion

By comprehensively utilizing available data on the functional relationships among kinases, phosphoproteins, and phosphorylation sites, RoKAI improves the robustness of kinase activity inference to the missing annotations and quantifications. We expect that this will facilitate the identification of understudied kinases with only a few annotations and lead to the development novel kinase inhibitors for targeted therapy of many diseases such as cancer, Alzheimer's disease, and Parkinson's disease. As additional functional information on cellular signaling becomes available, the inclusion of these information in functional networks utilized by RoKAI will likely further enhance the accuracy and robustness of kinase activity inference.

4 Methods

4.1 Problem Definition

Kinase activity inference can be defined as the problem of predicting changes in kinase activity based on observed changes in the phosphorylation levels of substrates. Formally, let $K = \{k_1, k_2, ..., k_m\}$ denote a set of kinases and $S = \{s_1, s_2, ..., s_n\}$ denote a set of phosphorylation sites. For these kinases and phosphosites, a set of annotations are available, where $S_i \subseteq S$ denotes the set of substrates of kinase k_i, i.e., $s_j \in S_i$ if kinase k_i phosphorylates site s_j.

In addition to the annotations, we are given a phosphorylation data set representing a specific biological context. This data set can be represented as a set of quantities q_j for $1 \leq j \leq n$, where q_j denotes the change in the phosphorylation level of phosphosite $s_j \in S$. Usually, q_j represents the log-fold change of the phosphorylation level of the site between two sets of samples representing different conditions, phenotypes, or perturbations. The objective of kinase activity inference is to integrate the annotations and the specific phosphorylation data to identify the kinases with significant difference in their activity between these two sets of samples. In the below discussion, we denote the inferred change in the activity of kinase k_i as $\hat{a}_i$. Since existing kinase activity inference methods are unsupervised, many activity inference methods also compute a p-value to assess the statistical significance of $\hat{a}_i$ for each kinase.

4.2 Background

Kinase activity inference methods differ from each other in terms how they integrate the phosphorylation levels of the substrates of a kinase k_i to estimate its activity $\hat{a}_i$. These methods range from simple aggregates and enrichment analyses to more sophisticated methods that take into account the interplay between different kinases.

Mean (baseline): One of the simpliest kinase activity inference methods employed by KSEA (Casado *et al.*, 2013), this method represents the activity of a kinase as the mean phosphorylation (log-fold change) of its substrates:

$$\hat{a}_i^{(\text{mean})} = \frac{\sum_{s_j \in S_i} q_j}{|S_i|}. \tag{2}$$

where $|S_i|$ is the number of substrates of kinase k_i.

Z-score: To assess the statistical significance of inferred activity, KSEA normalizes the total log-fold change of substrates with the standard deviation of the log-fold changes of all sites in the dataset:

$$\hat{a}_i^{(\text{z-score})} = \frac{\sum_{s_j \in S_i} q_j}{\sigma} = \frac{|S_i|}{\sigma} \hat{a}_i^{(\text{mean})}, \tag{3}$$

where σ denotes the standard deviation of phosphorylation across all phosphosites.

Linear model: The linear model, considered by Hernandez-Armenta *et al.* (2017), aims to take into account of the dependencies between kinases that phosphorylate the same site. A similar (but more complex) approach is also utilized by IKAP (Mischnik *et al.*, 2015). In this model, the phosphorylation of a site is modeled as summation of the activities of kinases that phosphorylate

the site:

$$q_j = \sum_{\substack{\text{for all kinases } i \\ \text{phosphorylating site } j}} a_i \tag{4}$$

where a_i is variable representing the activity of kinase k_i. To infer the kinase activities, least squares optimization function with ridge regularization is used:

$$\hat{a}^{(\text{linear})} = \text{argmin}_a \left\{ \sum_{s_j \in S} (q_j - \sum_{k_i \in K_j} a_i)^2 + \lambda ||a||^2 \right\}, \tag{5}$$

where K_j denotes the set of kinases that phosphorylate site s_j, and λ is an adjustable regularization coefficient. The first term in the objective function (squared loss) ensures that the inferred kinase activities are consistent with the phosphorylation levels of their substrates, whereas the second term (regularization) aims to minimize the overall magnitude of inferred kinase activities. In all experiments, we utilize a regularization coefficient of $\lambda = 0.1$ as previously done in Hernandez-Armenta *et al.* (2017).

GSEA: Gene set enrichment analysis (GSEA) (Subramanian *et al.*, 2005) is a widely used method in systems biology. It uses a weighted Kolmogorov-Smirnov statistic to assess whether a specific set of genes (or in general molecular entities) are significantly over-represented among differentially expressed genes (or in general differential activity/abundance). Ochoa *et al.* (2016) adopt GSEA to infer kinase activity by assessing whether the target sites of a kinase exhibit are enriched in terms of their phosphorylation fold change compared to other phosphosites. For this purpose, the sites are first ranked based on their absolute fold changes. For each kinase k_i, a running sum is computed based on the ranked list of sites. The running sum increases for each site $s_j \in S_i$ (i.e., s_j is a known substrate of k_i), and decreases for each site $s_j \notin S_i$ (i.e., s_j is not a known substrate of k_i. The maximum deviation of this running sum from zero is used as the enrichment score of a kinase. The statistical significance of this enrichment score is assessed using a permutation test. Namely, fold changes of sites are permuted $10,000$ times and enrichment scores are computed for each. The p-value for a kinase is then computed as the number of permutations with higher enrichment score than observed. As the predicted activity of a kinase, -log10 of this p-value is used.

4.3 Phospho-proteomics data preprocessing

Following the footsteps of previous studies (Ochoa *et al.*, 2016; Hernandez-Armenta *et al.*, 2017), we apply some quality control steps to the phospho-proteomics data that is used for benchmarking: (i) we restrict the analysis to mono-phosphorylated peptides that are mapped to canonical transcripts of Ensembl, (ii) we average the log fold changes of technical replicates as well as peptides that are mapped to the same Ensembl position (even if the exact peptides sequences are not identical), and (iii) we filter out the peptides that are identified in only a single study to reduce the amount of false-positive phosphosites, and (iv) we restrict the analysis to perturbations in the gold standard with more than 1000 phosphosite identifications. As a result of these steps, we obtain 53636 sites identified in at least one of 80 perturbations. For these 80 perturbations, there are 128 kinase-perturbation annotations for 25 different kinases.

4.4 Computing benchmarking metric (top-k-hit)

To compute the $\text{P}_{\text{hit}}(k)$ metric (read "top-k-hit"), we apply the following procedure:

1. For each perturbation separately, we rank the kinases based on their absolute activities predicted by the inference method.
2. For each perturbation, we consider the top k kinases with highest predicted activity and compare them with the "true" perturbed kinases in gold standard.
3. If any of the top k kinases is a true kinase (i.e., a kinase that is perturbed in the experiment), we consider the inference method to be successful (i.e., a hit) for that perturbation.

4. We compute the percentage of perturbations with successful predictions and report this quantity as $\mathrm{P}_{\mathrm{hit}}(k)$. Since the gold standard dataset is incomplete, $\mathrm{P}_{\mathrm{hit}}(k)$ metric serves as a minimum bound on the expected probability of discovering an up-regulated or down-regulated kinase if top k kinases predicted by the inference method were to be experimentally validated.

4.5 Robust kinase activity inference (RoKAI)

4.5.1 Heterogeneous network model

RoKAI uses a heterogeneous network model in which nodes represent kinases and/or phosphosites. The edges in this network represent different types of functional association between kinases, between phosphosites, and between kinases and phosphosites. Namely, RoKAI's functional network consists of the following types of edges:

Kinase-Substrate Associations: An edge between a kinase k_i and site s_j indicates that k_i phosphorylates s_j. These kinase-substrate associations obtained from PhosphositePlus (Hornbeck *et al.*, 2015), representing 3877 associations between 261 kinases and 2397 sites.

Structure Distance Evidence: This type of edge between phosphosites s_i and s_j represents the similarity of s_i and s_j on the protein structure. We obtain structure distance evidence from PTMcode (Minguez *et al.*, 2012), which contains 7821 unweighted edges between 8842 distinct sites. Note that, in this network, a large portion of the edges (7037 edges) are intra-protein.

Co-Evolution Evidence: This type of edge between phosphosites s_i and s_j indicates that the protein sequences straddling s_i and s_j exhibit significant co-evolution. We obtain this co-evolution network from PTMcode which contains 178029 unweighted edges between 19122 distinct sites. After filtering the sites for rRCS ≥ 0.9 provided by PTMcode, 41799 edges between 8342 distinct sites remain. Note that, 3516 of these edges overlap with the structural distance network. Thus, when co-evolution and structural distance networks are used together, these 3516 overlapping edges are considered to have a weight of 2.

Protein-Protein Interactions: An edge between kinases k_i and kinase k_j represents a protein-protein interaction between k_i and k_j. We use the protein-protein interaction (PPI) network obtained from STRING (Szklarczyk *et al.*, 2014). As the edge weights, we utilize the combined scores provided by STRING. Overall, the kinase-kinase interaction network contains 13031 weighted edges (weights ranging from 0 to 1) between 255 distinct kinases.

4.5.2 Network propagation

Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ represent RoKAI's heterogeneous functional network, where $\mathcal{V} = K \cup S$ and $\mathcal{E}$ contains four types of edges as described above. To propagate phosphorylation levels of sites over $\mathcal{G}$, we utilize an electric circuit model (illustrated in Fig. 1). In this model, each node $n_i \in \mathcal{V}$ (kinase or phosphosite) has a node potential V_i. Each edge $e_{ij} \in \mathcal{E}$ (which can be a kinase-substrate association, kinase-kinase interaction or association between a pair of phosphosites) has a conductance c_{ij} that allows some portion of the node potential V_i of node n_i to be transferred to node n_j in the form of a *current* I_{ij}:

$$I_{ij} = (V_i - V_j)\, c_{ij} \tag{6}$$

As seen in the equation, the current I_{ij} carried by an edge is proportional to its condundance and the difference in node potentials. In our model, we use the weights available in the corresponding networks to assign conductance values to the edges.

We model the phosphorylation level of a site s_j that is identified in the experiment as a current source $I_j = q_j$ connected to the reference node (representing the control sample) with a unit conductance. This ensures that the node potential V_j of site s_j is equal to its phosphorylation level q_j if it is not connected to any other nodes. This is because the current incoming to a node is

always equal to its outgoing current:

$$\begin{gathered} \textit{Incoming current} = \textit{Outgoing current} \\ q_i = V_i + \sum_{(i,j)\in E} (V_i - V_j)c_{ij}, \qquad \text{if } n_i \text{ has quantification} \\ 0 = \sum_{(i,j)\in E} (V_i - V_j)c_{ij}, \qquad \text{if } n_i \text{ does not have quantification} \end{gathered} \tag{7}$$

Observe that, in this model, the nodes without measured phosphorylation levels (sites that are not identified in an MS run or kinases) act as a bridge for connecting (and transferring phosphorylation levels between) other nodes. This is an important feature of RoKAI as it allows incorporation of unidentified phosphosites in the network model.

To compute the node potentials for all nodes in the network, we represent Equation 7 as a linear system:

$$CV = b \tag{8}$$

$$C_{ij} = \left\{ \begin{array}{ll} 1 & \text{if } i = j \text{ and } n_i \text{ has quantification} \\ c_{ij} & \text{if } i \neq j \text{ and } n_i n_j \in \mathcal{E} \\ 0 & \text{otherwise} \end{array} \right\} \tag{9}$$

$$b_i = \left\{ \begin{array}{ll} q_i & \text{if } n_i \text{ has quantification} \\ 0 & \text{otherwise} \end{array} \right\} \tag{10}$$

Thus, the node potentials V can be computed using linear algebra as follows:

$$V = (C^T C)^{-1} C^T b \tag{11}$$

Note that, to make the matrix inversion numerically stable, we add a small $\tau = 10^{-8}$ to the diagonals of the matrix C.

Once node potentials are computed, we output the propagated phosphorylation levels for identified sites as:

$$\hat{q}_j = V_j. \tag{12}$$

These propagated phosphorylation levels $\hat{q}_j$ are used as input to kinase activity inference algorithms to obtain the inferred activity of kinases.

Acknowledgements

This work has been supported in part by National Institutes of Health grant R01-LM012980 from the National Library of Medicine.

Data Availability

We obtain the benchmarking data from publicly available datasets of previous studies (Ochoa *et al.*, 2016; Hernandez-Armenta *et al.*, 2017) [3]. We obtain the kinase-substrate annotations from PhosphositePlus [4]. We obtain the human protein-protein interaction network from STRING (Szklarczyk *et al.*, 2014) [5]. We obtain the co-evolution and structure distance evidence between phosphosites from PTMcode (Minguez *et al.*, 2012) [6]

[3] http://phosfate.com/download.html
[4] http://www.phosphosite.org/staticDownloads
[5] http://string-db.org/cgi/download.pl
[6] http://ptmcode.embl.de/data.cgi

References

Ayati, M. *et al.* (2019). Cophosk: A method for comprehensive kinase substrate annotation using co-phosphorylation analysis. *PLoS computational biology*, **15**(2), e1006678.

Beekhof, R. *et al.* (2019). Inka, an integrative data analysis pipeline for phosphoproteomic inference of active kinases. *Molecular systems biology*, **15**(4).

Butrynski, J. E. *et al.* (2010). Crizotinib in alk-rearranged inflammatory myofibroblastic tumor. *New England Journal of Medicine*, **363**(18), 1727–1733.

Casado, P. *et al.* (2013). Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal.*, **6**(268), rs6–rs6.

Choi, J. H. *et al.* (2010). Anti-diabetic drugs inhibit obesity-linked phosphorylation of pparγ by cdk5. *Nature*, **466**(7305), 451–456.

Cichonska, A. *et al.* (2017). Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. *PLoS computational biology*, **13**(8), e1005678.

Cohen, P. (2001). The role of protein phosphorylation in human health and disease. the sir hans krebs medal lecture. *European journal of biochemistry*, **268**(19), 5001–5010.

Copps, K. and White, M. (2012). Regulation of insulin sensitivity by serine/threonine phosphorylation of insulin receptor substrate proteins irs1 and irs2. *Diabetologia*, **55**(10), 2565–2582.

Cowen, L. *et al.* (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, **18**(9), 551.

Dephoure, N. *et al.* (2013). Mapping and analysis of phosphorylation sites: a quick guide for cell biologists. *Molecular biology of the cell*, **24**(5), 535–542.

Deznabi, I. *et al.* (2019). Deepkinzero: Zero-shot learning for predicting kinase-phosphosite associations involving understudied kinases. *BioRxiv*, page 670638.

Doyle, P. G. and Snell, J. L. (1984). *Random walks and electric networks*, volume 22. American Mathematical Soc.

Drake, J. M. *et al.* (2012). Oncogene-specific activation of tyrosine kinase networks during prostate cancer progression. *Proceedings of the National Academy of Sciences*, **109**(5), 1643–1648.

Drake, J. M. *et al.* (2016). Phosphoproteome integration reveals patient-specific networks in prostate cancer. *Cell*, **166**(4), 1041–1054.

Hernandez-Armenta, C. *et al.* (2017). Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics*, **33**(12), 1845–1851.

Horn, H. *et al.* (2014). Kinomexplorer: an integrated platform for kinome biology studies. *Nature methods*, **11**(6), 603.

Hornbeck, P. V. *et al.* (2015). Phosphositeplus, 2014: mutations, ptms and recalibrations. *Nucleic acids research*, **43**(D1), D512–D520.

Koyano, F. *et al.* (2014). Ubiquitin is phosphorylated by pink1 to activate parkin. *Nature*, **510**(7503), 162–166.

Krug, K. *et al.* (2019). A curated resource for phosphosite-specific signature analysis. *Molecular & cellular proteomics*, **18**(3), 576–593.

Linding, R. *et al.* (2007). Systematic discovery of in vivo phosphorylation networks. *Cell*, **129**(7), 1415–1426.

Liu, Y. and Chance, M. R. (2014). Integrating phosphoproteomics in systems biology. *Computational and structural biotechnology journal*, **10**(17), 90–97.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, **46**(253), 68–78.

Minguez, P. *et al.* (2012). Ptmcode: a database of known and predicted functional associations between post-translational modifications in proteins. *Nucleic acids research*, **41**(D1), D306–D311.

Mischnik, M. *et al.* (2015). Ikap: A heuristic framework for inference of kinase activities from phosphoproteomics data. *Bioinformatics*, **32**(3), 424–431.

Neddens, J. *et al.* (2018). Phosphorylation of different tau sites during progression of alzheimer's disease. *Acta neuropathologica communications*, **6**(1), 52.

Needham, E. J. *et al.* (2019). Illuminating the dark phosphoproteome. *Sci. Signal.*, **12**(565), eaau8645.

Neviani, P. and Perrotti, D. (2014). Setting op449 into the pp2a-activating drug family. *Clinical Cancer Research*, **20**(8), 2026–2028.

Ochoa, D. *et al.* (2016). An atlas of human kinase regulation. *Molecular systems biology*, **12**(12).

Perrotti, D. and Neviani, P. (2013). Protein phosphatase 2a: a target for anticancer therapy. *The lancet oncology*, **14**(6), e229–e238.

Puri, P. *et al.* (2008). Activation and dysregulation of the unfolded protein response in nonalcoholic fatty liver disease. *Gastroenterology*, **134**(2), 568–576.

Reese, L. C. *et al.* (2011). Dysregulated phosphorylation of ca2+/calmodulin-dependent protein kinase ii-α in the hippocampus of subjects with mild cognitive impairment and alzheimer's disease. *Journal of neurochemistry*, **119**(4), 791–804.

Rikova, K. *et al.* (2007). Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell*, **131**(6), 1190–1203.

Subramanian, A. *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**(43), 15545–15550.

Suo, S.-B. *et al.* (2014). Psea: Kinase-specific prediction and analysis of human phosphorylation substrates. *Scientific reports*, **4**, 4524.

Szklarczyk, D. *et al.* (2014). String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, **43**(D1), D447–D452.

Wilkes, E. H. *et al.* (2017). Kinase activity ranking using phosphoproteomics data (karp) quantifies the contribution of protein kinases to the regulation of cell viability. *Molecular & Cellular Proteomics*, **16**(9), 1694–1704.

Wiredja, D. D. *et al.* (2017a). The ksea app: a web-based tool for kinase activity inference from quantitative phosphoproteomics. *Bioinformatics*, **33**(21), 3489–3491.

Wiredja, D. D. *et al.* (2017b). Phosphoproteomics profiling of nonsmall cell lung cancer cells treated with a novel phosphatase activator. *Proteomics*, **17**(22), 1700214.

Zhou, C. *et al.* (2011). Erlotinib versus chemotherapy as first-line treatment for patients with advanced egfr mutation-positive non-small-cell lung cancer (optimal, ctong-0802): a multicentre, open-label, randomised, phase 3 study. *The lancet oncology*, **12**(8), 735–742.