nature communications

Article

https://doi.org/10.1038/s41467-025-57993-2

Deciphering the dark cancer phosphoproteome using machine-learned co-regulation of phosphosites

Received: 7 July 2024

Accepted: 10 March 2025

Published online: 20 March 2025

Check for updates

Wen Jiang[1,2], Eric J. Jaehnig[1,2], Yuxing Liao[1,2], Zhiao Shi[1,2], Tomer M. Yaron-Barir[3,4,5], Jared L. Johnson[6,7], Lewis C. Cantley[6,7] & Bing Zhang[1,2] ✉

Mass spectrometry-based phosphoproteomics offers a comprehensive view of protein phosphorylation, yet our limited knowledge about the regulation and function of most phosphosites hampers the extraction of meaningful biological insights. To address this challenge, we integrate machine learning with phosphoproteomic data from 1195 tumor specimens spanning 11 cancer types to construct CoPheeMap, a network that maps the co-regulation of 26,280 phosphosites. By incorporating network features from CoPheeMap into a second machine learning model, namely CoPheeKSA, we achieve superior performance in predicting kinase-substrate associations. CoPheeKSA uncovers 24,015 associations between 9399 phosphosites and 104 serine/threonine kinases, shedding light on many unannotated phosphosites and understudied kinases. We validate the accuracy of these predictions using experimentally determined kinase-substrate specificities. Through the application of CoPheeMap and CoPheeKSA to phosphosites with high computationally predicted functional significance and those associated with cancer, we demonstrate their effectiveness in systematically elucidating phosphosites of interest. These analyses unveil dysregulated signaling processes in human cancer and identify understudied kinases as potential therapeutic targets.

Protein phosphorylation is a crucial post-translational modification (PTM) that regulates a wide range of cellular processes, such as proliferation, differentiation, motility, and apoptosis[1]. This modification is tightly controlled by protein kinases and phosphatases[2]. Dysregulated protein phosphorylation, often driven by aberrant kinase activity, is implicated in many diseases, including cancer[3–5]. Consequently, kinases have become a promising class of drug targets for developing cancer therapies, with more than 70 kinase inhibitors approved for cancer treatment since the approval of imatinib for chronic myeloid leukemia in 2001[6]. However, these inhibitors only target 50 out of the more than 500 protein kinases encoded by the human genome, leaving the majority of kinases open for further exploration and investigation.

Mass spectrometry (MS)-based phosphoproteomics has transformed the analysis of protein phosphorylation, enabling comprehensive profiling of phosphorylation events across the entire proteome in biological samples[7]. Its application in cancer research has been growing rapidly due to the central role of phospho-signaling in

[1]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX 77030, USA. [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. [3]Meyer Cancer Center, Weill Cornell Medicine, New York, NY 10021, USA. [4]Englander Institute for Precision Medicine, Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY 10021, USA. [5]Columbia University Vagelos College of Physicians and Surgeons, New York, NY 10032, USA. [6]Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA. [7]Dana Farber Cancer Institute, 450 Brookline Ave, Boston, MA 02215, USA. ✉e-mail: bing.zhang@bcm.edu

cancer biology and treatment[8]. A single phosphoproteomic study can uncover tens of thousands of phosphosites, with quantitative comparisons often revealing hundreds to thousands of significantly regulated sites. Annotating phosphosites with existing functional and regulatory information is crucial for interpreting phosphoproteomics data. However, this presents a significant challenge, as less than 5% of the human phosphoproteome has been experimentally linked to specific kinases, and even fewer phosphosites have been functionally characterized. The paucity of knowledge on phosphosites has led to the notion of the dark phosphoproteome[9].

A common task in phosphoproteomic data analysis is to infer upstream kinase activity changes based on phosphoproteomic data. This analysis relies heavily on known kinase-substrate associations (KSAs), excluding the vast majority of phosphosites without annotated kinases, i.e., dark phosphosites. Many computational methods have been developed to predict KSAs, with most leveraging consensus sequence motifs or position-specific scoring matrices (PSSMs) derived from known KSAs for individual kinases[10]. Tools like NetworKIN[11] enhance KSA prediction by integrating sequence-based approaches with protein-protein interaction network data, adding a contextual layer to improve specificity and accuracy. However, because 90% of the annotated phosphosites are associated with just 20% of kinases that are the most well-studied[10], the ability to predict KSAs and infer activities for the understudied kinases is restricted, despite their potential as therapeutic targets. Orthogonal approaches that combine kinase inhibition profiles with computational analysis have been developed to address this limitation[12,13]. For example, the expectancy of being downstream target (EBDT) algorithm utilizes information on kinase-inhibitor selectivity to identify the most probable kinase(s) upstream of phosphosites detected in phosphoproteomics data from cell lines, leading to the identification of thousands of phosphosites as putative downstream targets (PDTs) of 103 kinases[12]. While these methods are valuable, they are time consuming, costly, and restricted by the limited availability and specificity of inhibitors. Importantly, these approaches often include indirect kinase-substrate regulations, making it difficult to deconvolute direct versus indirect phosphosite abundance changes.

Biological networks, such as protein-protein interaction networks and gene co-expression networks, are effective means of propagating information from well-characterized molecular components to understudied ones, leading to a better understanding of their regulation and functions[14]. These networks have proven successful in predicting functions for understudied proteins and identifying cancer genes[15,16]. In the field of protein phosphorylation, known KSAs are often represented as a bipartite network, where edges connect kinase nodes to phosphosite nodes, with no connections within the same node type. This network model forms the foundation of most kinase activity inference algorithms[17] and has also been utilized to predict novel KSAs using missing link prediction methods such as LinkPhinder[18]. To expand this framework, researchers have developed heterogeneous network models by introducing edges between phosphosites and between kinases, incorporating diverse functional information such as coevolution and structural proximity for phosphosites and protein-protein interactions for kinases[19]. For the analysis of phosphoproteomics data, individual phosphosite measurements have been aggregated to the protein level and used as node weights in protein-protein interaction networks to identify active signaling pathways[20]. Alternatively, co-phosphorylation derived from the data has been integrated as edge weights in heterogeneous networks to identify context-specific subnetworks of interest[21]. Phosphosite co-phosphorylation, akin to transcript and protein co-expression, offers valuable insights into their co-regulation. However, despite the rapid expansion of MS-based phosphoproteomics datasets, systematic efforts to infer phosphosite co-regulation relationships absent from existing networks remain scarce. This represents a significant untapped opportunity to uncover deeper insights into the phosphoproteome.

In this study, we aim to construct a comprehensive, data-driven co-regulation network of phosphosites by integrating publicly available large-scale cancer phosphoproteomic datasets. Co-regulated phosphosites are more likely to participate in shared biological processes, making these networks a valuable resource for uncovering functional and regulatory insights, even in the absence of existing knowledge about their regulatory enzymes. Additionally, the latent information embedded in these networks has the potential to improve KSA prediction, thereby advancing kinase activity inference and other applications reliant on accurate kinase-substrate relationship data.

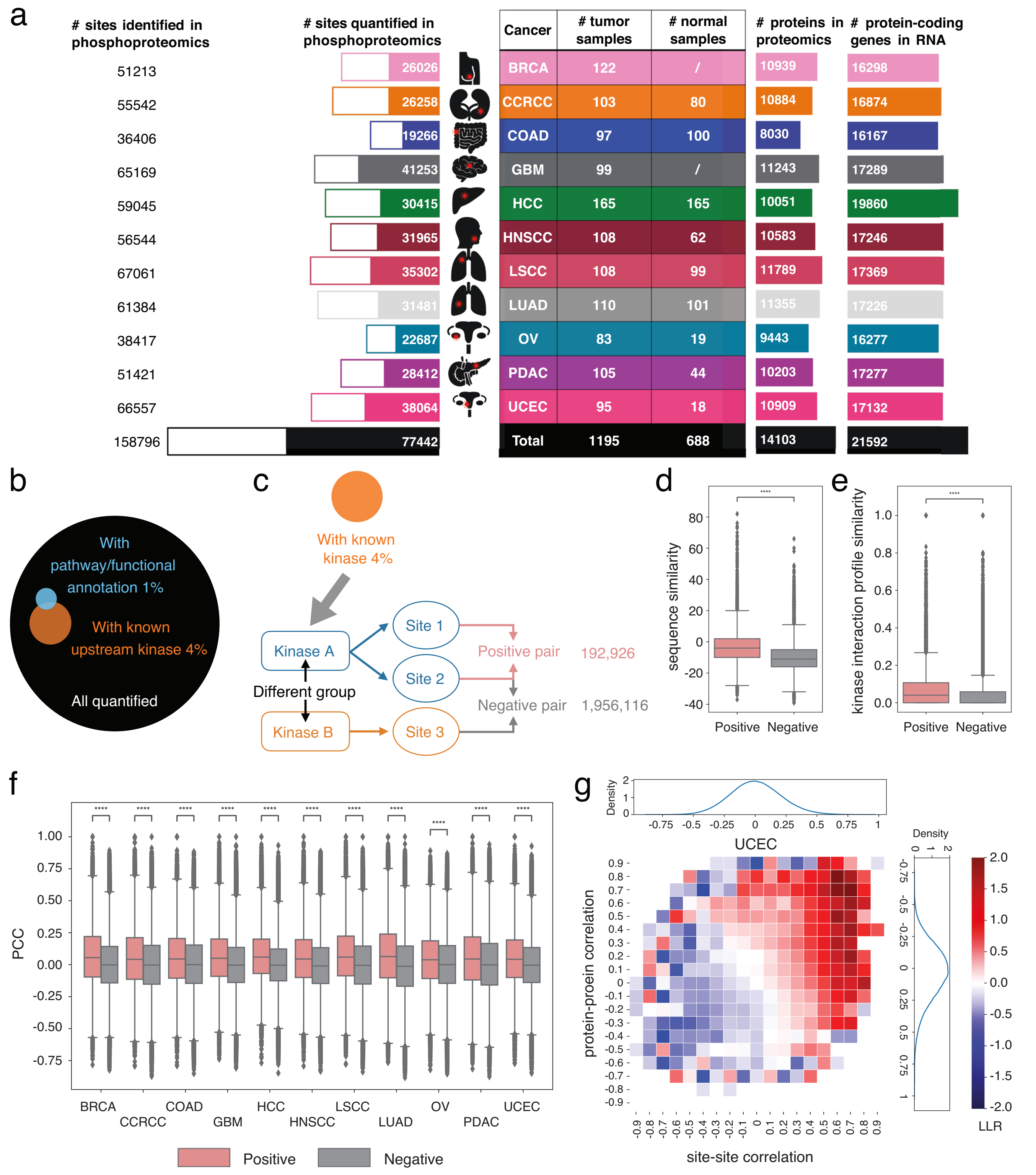
To realize this, we utilize machine learning and phosphoproteomic data from 1195 clinical tumor specimens across 11 cancer types, recently harmonized by the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC), to construct CoPheeMap, a co-regulation network of 26,280 phosphosites. By incorporating network embedding features from CoPheeMap into a second machine learning model called CoPheeKSA, we achieve superior performance in predicting KSAs, uncovering 24,015 associations, many of which involve unannotated phosphosites and understudied kinases. Furthermore, we validate these CoPheeKSA predictions by comparing them to substrate specificities determined experimentally using the Kinase Library (KL) approach[22]. To illustrate the practical utility of CoPheeMap and CoPheeKSA, we present examples demonstrating their ability to systematically shed light on phosphosites of interest, revealing dysregulated signaling processes in human cancer, and identifying understudied kinases as putative therapeutic targets.

Results

Phosphosite co-phosphorylation is an effective predictor of co-regulation

This study utilized recently harmonized pan-cancer (PanCan) proteogenomic data from the CPTAC pan-cancer resource working group[23]. The harmonized dataset contained phosphoproteomic data for 1,195 tumor samples across 11 cancer types, and phosphoproteomic data was also available for normal samples of 9 cancer types (Fig. 1a). In total, the dataset covered 158,796 phosphosites, including 77,442 quantified in at least 20% of samples in one tumor cohort (Supplementary Data 1). Global proteomic and RNA-Seq data were also available for the same cohorts, covering a total of 14,103 and 21,592 protein-coding genes, respectively (Fig. 1a). Among the quantified phosphosites, fewer than 5% were annotated with up-stream kinases or known functions (Fig. 1b), echoing the widely recognized dark phosphoproteome challenge[9].

Phosphosite co-phosphorylation may offer valuable insights into their co-regulation. To formally assess the relationship between phosphosite co-phosphorylation and co-regulation, we constructed a ground-truth dataset of phosphosite co-regulation. This dataset was derived from a previously published database containing 14,679 comprehensively curated KSAs from the literature, covering 362 kinases and 9526 unique sites[24]. In this study, kinases with more than 10 known substrates in this dataset were classified as well-studied, while those with 10 or fewer substrates were considered understudied. Among these KSAs, 4873 KSAs involved phosphosites quantified in our PanCan datasets. Based on these 4873 KSAs, we created a ground-truth dataset that included 192,926 pairs of phosphosites known to be regulated by the same kinase (positives) and 1,956,116 pairs regulated by kinases from different kinase groups (negatives) (Fig. 1c, Supplementary Data 2–3). Since protein kinases are often either Serine-Threonine (S/T) kinases or Tyrosine (Y) kinases, we considered two types of pairs separately: S/T-S/T and Y-Y. This ground-truth dataset included 2410 unique phosphosites, referred to as annotated sites. Phosphosites not included in the ground-truth KSA dataset were defined as unannotated sites for this study.



As expected, compared to negative site pairs, positive site pairs had significantly higher 15-mer peptide sequence similarities (Fig. 1d), and their host proteins had significantly higher kinase interaction profile similarities in the STRING protein-protein association network (Fig. 1e, Methods). In addition, positive site pairs also showed significantly higher abundance correlations than negative pairs across all 11 cancer types (Fig. 1f), largely independent of corresponding host protein correlations as demonstrated in heatmaps co-visualizing the distributions of log likelihood ratios (LLRs) between positive and negative pairs at the site (binned in rows) and protein (binned in columns) levels, respectively (Fig. 1g, Supplementary Fig. 1). These data suggest that phosphosite co-phosphorylation, together with sequence similarity and host protein kinase interaction profile similarity, are effective predictors of co-regulation.

CoPheeMap: a co-regulation map of the human cancer phosphoproteome

To construct a network of co-regulated phosphosites, we utilized the Extreme Gradient Boosting (XGBoost) algorithm[25], which offers several advantages particularly suited to this study. Its robustness in handling

Fig. 1 | Pan-cancer datasets and features associated with phosphosites co-regulation. a Overview of the pan-cancer phosphoproteomics data along with matched global proteomics and RNA-Seq data. BRCA (Breast Cancer), CCRCC (Clear Cell Renal Cell Carcinoma), COAD (Colon Adenocarcinoma), GBM (Glioblastoma), HCC (Hepatocellular Carcinoma), HNSCC (Head and Neck Squamous Cell Carcinoma), LSCC (Lung Squamous Cell Carcinoma), LUAD (Lung Adenocarcinoma), OV (Ovarian Cancer), PDAC (Pancreatic Ductal Adenocarcinoma), and UCEC (Uterine Corpus Endometrial Carcinoma). **b** Proportions of the quantified phosphosites in the pan-cancer data with functional or regulatory information. **c** Construction of a ground-truth dataset containing co-regulated phosphosite pairs (positives) and independently regulated pairs (negatives). **d–f** Positive pairs exhibit significantly higher scores in sequence similarity (**d**), kinase interaction profile similarity (**e**), and Pearson's correlation coefficients (PCC) across cancer types (**f**) compared to negative pairs. The numbers of positive site pairs and negative site pairs are 161,832 and 1,306,973 respectively. P-values derived from a two-sided T test with Bonferroni correction. ****$p < 0.0001$. For boxplots, centerline indicates the median, box limits indicate upper and lower quartiles, whiskers indicate the 1.5 interquartile range. **g** Heatmap illustrating co-regulation levels of phosphosite pairs with specific site co-phosphorylation (x axis) and host protein co-expression (y axis) levels in the UCEC data. Colors indicating co-regulation levels quantified by log-likelihood ratio (LLR). The density plots at the top and right display the distributions of site co-phosphorylation and protein co-expression, respectively. Source data are provided as a Source Data file.

missing values is especially important given the high prevalence of missing data in phosphoproteomics datasets. Additionally, XGBoost effectively manages imbalanced training and prediction data, ensures efficient training and prediction processes, incorporates mechanisms to prevent overfitting, and demonstrates strong performance across diverse applications. Specifically, we developed an XGBoost classifier that integrates sequence similarity, kinase interaction profile similarity between two host proteins, and pairwise phosphosite correlations for each of the 11 cancer types to distinguish the positive and negative phosphosite pairs (Fig. 2a). This classifier was developed using carefully prepared training, validation, and independent test data based on the ground-truth dataset (Supplementary Data 2–3) and 10 repetitions of Monte Carlo validation (Supplementary Fig. 2a–c, Methods). The trained model incorporating all three types of features achieved an Area Under the Receiver Operating Characteristic (AUROC) of 0.78 in the independent test data, outperforming models constructed using only subsets of features classified as static (sequence similarity scores and kinase interaction profile similarity scores) or dynamic (site-site correlations; Fig. 2b). This performance gain remained evident when we applied a stringent False Positive Rate (FPR) threshold of 0.2% (Fig. 2b).

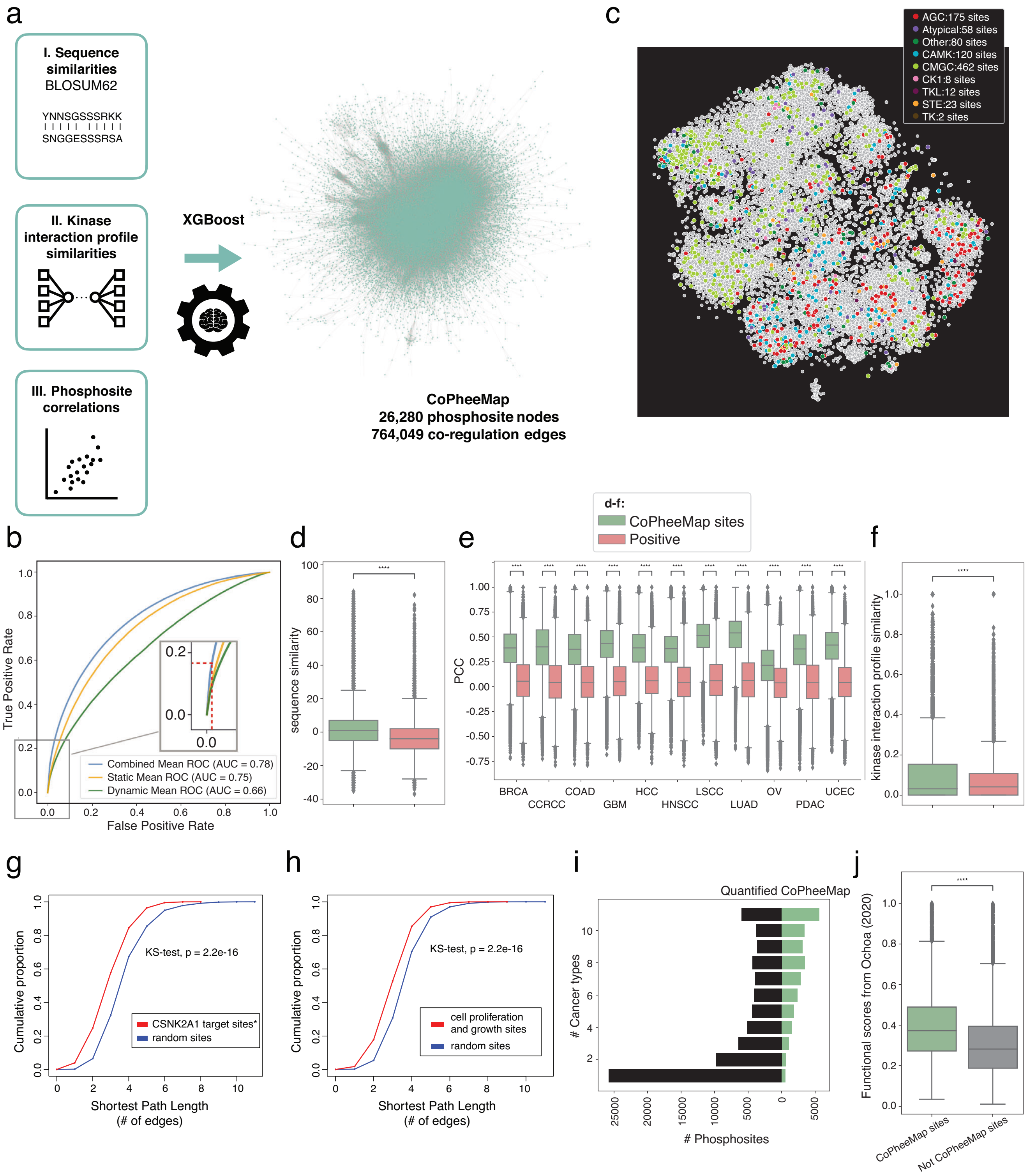
By applying the trained classifier to over 3 billion phosphosite pairs, controlling FPR at 0.2% to account for the large number of possible negatives, we identified 764,049 phosphosite pairs (0.03% of all candidate pairs) as co-regulated phosphosite pairs. These pairs, connecting 26,280 unique phosphosites, constituted a co-regulation network referred to as CoPheeMap (Fig. 2a, Supplementary Data 4). To gain global insights into the organization of CoPheeMap, we generated 16-dimensional vector representations of all phosphosites in the network using the Node2Vec embedding algorithm. Node2Vec captures both local connections between closely related nodes and broader patterns across the network, making it a powerful tool for identifying relationships and community structures within the network[26,27]. For visualization, we further reduced the representations to two dimensions using a multi-scale t-distributed stochastic neighbor embedding (tSNE) approach[28] with parameters recommended for large network analysis[29]. The distance between two phosphosites in the tSNE plot reflects their distance in the CoPheeMap network. In the tSNE plot, phosphosites regulated by kinases in the same group tended to cluster together, further supporting the effectiveness of CoPheeMap in capturing co-regulation relationships between phosphosites (Fig. 2c).

Notably, the predicted co-regulated phosphosite pairs in CoPheeMap showed significantly higher sequence similarity scores and co-phosphorylation levels than the positive pairs in the ground-truth data (Fig. 2d, e). Unlike the other two features, the median kinase interaction profile similarity score for predicted co-regulated phosphosite pairs was relatively lower than that of the positive group (Fig. 2f), indicating that this feature generally had a weaker contribution to the predictions. Nevertheless, the predicted pairs were still enriched with instances showing high kinase interaction profile similarity scores, as reflected by the higher upper quartile. This suggests that while overall contributions were modest, the feature may have played a significant role in certain specific predictions.

Indicative of common targets of the same kinase being better connected than pairs of sites that are phosphorylated by different kinases, the positive site pairs had significantly shorter shortest path lengths (SPLs) than the negative site pairs did (Supplementary Fig. 2d). Moreover, despite the sparsity of CoPheeMap, the known substrates of CSNK1A1, CSNK2A1, CSNK2A2, the three kinases intentionally excluded from the ground truth, were mostly well-connected (Supplementary Fig. 2e–g). Particularly, the distribution of SPLs between all pairs of substrates of CSNK2A1, which had sufficient numbers for statistical assessment, was significantly lower than it was for pairs from a randomly drawn set of sites of the same size (Kolmogorov–Smirnov test, $p = 2.2 \times 10^{-16}$; Fig. 2g), demonstrating the capacity of CoPheeMap to effectively recover missing co-regulation relationships within ground truth data.

Given that sites involved in the same biological function are more likely to be associated with common pathways regulated by common kinases, we reasoned that they would be closer together than random sites in the network. To test this, we collected functional annotations from PhosphoSitePlus for sites in CoPheeMap and parsed them into five categories of related functions: cell growth and proliferation, cellular degradation (apoptosis and autophagy), gene product regulation, microenvironment and mobility, and signaling pathway regulation. Despite the heterogeneity within these high-level classifications, the distribution of SPLs between pairs of sites that are associated with the regulation of cell growth and proliferation is significantly lower than that of pairs of sites from a randomly drawn set of the same size (Kolmogorov–Smirnov test, $p = 2.2 \times 10^{-16}$; Fig. 2h). Similar results were obtained for all other functional categories (Supplementary Fig. 2h–k), demonstrating that functionally related sites are closer together in CoPheeMap than randomly selected sites.

Further analysis showed that the phosphosites in CoPheeMap were more likely to be quantified in multiple cancer types compared with the other quantified phosphosites (Fig. 2i). Sites included in CoPheeMap had higher abundance (Supplementary Fig. 2l) and higher predicted functional scores[30] (Fig. 2j) compared to those not included, suggesting high functional relevance for sites featured in CoPheeMap. The degree distribution plot of CoPheeMap (Supplementary Fig. 2m) revealed that most nodes (phosphosites) had a very small degree (number of connections), while a small number of nodes exhibited degrees exceeding 1000. The presence of these hubs, with degrees several orders of magnitude larger than those of most nodes, is a defining characteristic of a scale-free network, commonly observed in real-world and biological systems[31]. As expected, the annotated sites used for training tended to have higher degrees in the network compared to the unannotated sites; however, the latter also had a substantial number of connections in the network, with a median degree of 3.2, and some unannotated sites even acted as the hubs (Supplementary Fig. 2n). For example, 55 sites had more than 1000 connections, such as RFC1_S368 (1873), API5_S464 (1307), and TOP2A_S1106 (1283). In total, 753,243 edges in CoPheeMap (98.6%) involved at least one unannotated site. This extensive network of connections enabled



the propagation of information from annotated to unannotated sites, an opportunity explored later in this paper.

CoPheeKSA: a network-based generic KSA prediction model

Conventional methods for predicting KSAs typically rely solely on sequence motif information derived from known kinase targets[17]. These approaches are kinase-specific and lack generalizability when it comes to predicting KSAs for understudied kinases. To develop a general KSA prediction model that could be applied to both well-studied and understudied kinases, we hypothesized that biological relationships among phosphosites, kinases, and between kinases and phosphosites could all be utilized to improve generic KSA prediction.

To systematically capture phosphosite co-regulation, we used CoPheeMap. For kinase co-expression or co-function, we used kinase

Fig. 2 | Construction and evaluation of the CoPheeMap. a Three types of features (sequence similarities, kinase interaction profile similarities and phosphosite correlations) were used in the XGBoost machine learning algorithm to predict co-regulated phosphosite pairs for the construction of CoPheeMap. **b** The machine learning model incorporating all three types of features achieved higher AUROC than models using only a subset of features. Red dot line: cut-off of the model controlling False Positive Rate at 0.2% to account for the large number of possible negatives. **c** tSNE visualization of the Node2Vec embedding of the phosphosites in CoPheeMap, with phosphosites color-coded by kinase groups, and the number of phosphosites in each kinase group labeled. **d**–**f** Comparison of sequence similarities (**d**), Pearson's correlation coefficients (**e**), and kinase interaction profile similarities (**f**) between the site pairs included in CoPheeMap and those from the ground-truth positive set. The numbers of CoPheeMap site pairs and the positive site pairs are 764,051 and 103,457, respectively. P-values derived from a two-sided T test with Bonferroni correction. ****$p < 0.0001$. BRCA (Breast Cancer), CCRCC (Clear Cell Renal Cell Carcinoma), COAD (Colon Adenocarcinoma), GBM (Glioblastoma), HCC (Hepatocellular Carcinoma), HNSCC (Head and Neck Squamous Cell Carcinoma), LSCC (Lung Squamous Cell Carcinoma), LUAD (Lung Adenocarcinoma), OV (Ovarian Cancer), PDAC (Pancreatic Ductal Adenocarcinoma), and UCEC (Uterine Corpus Endometrial Carcinoma). **g, h** Comparison of shortest path length distributions between pairs of phosphosites regulated by CSNK2A1 and random site pairs (**g**), or pairs of phosphosites involved in regulating cell growth and proliferation and random site pairs (**h**). **i** Comparison of the frequency of identification of phosphosites in CoPheeMap across cancer types versus other quantified sites. Black: the number of quantified phosphosites in the pan-cancer data set, green: the number of CoPheeMap phosphosites. **j** Comparison of the functional scores of phosphosites in CoPheeMap versus those not in CoPheeMap. The numbers of CoPheeMap sites and non-CoPheeMap sites are 19,538 and 25,759, respectively. The phosphosites and functional scores were downloaded from Ochoa et al. Nature Biotechnology[30]. P-values derived from a two-sided T test. ****$p < 0.0001$. For boxplots, centerline indicates the median, box limits indicate upper and lower quartiles, whiskers indicate the 1.5 interquartile range. Source data are provided as a Source Data file.

co-expression data from the PanCan proteomics datasets and functional associations from the STRING protein-protein association network. The STRING database includes direct protein-protein interactions, pathway membership, and other functional relationships relevant to kinase co-activation within shared signaling pathways and biological processes. Combining these data types, we constructed a kinase network (KMap) with 352 kinases and 3238 edges (Methods, Supplementary Data 5). To generate phosphosite and kinase features for machine learning, we applied the Node2Vec embedding algorithm to CoPheeMap and KMap. This produced low-dimensional vector representations for the 25,028 S/T phosphosites in CoPheeMap and the 352 kinases in KMap (Fig. 3a). Node2Vec models both local and global network contexts, enabling the embeddings to capture the intrinsic characteristics of each phosphosite or kinase. These embeddings add a mechanistic layer of information to improve KSA prediction.

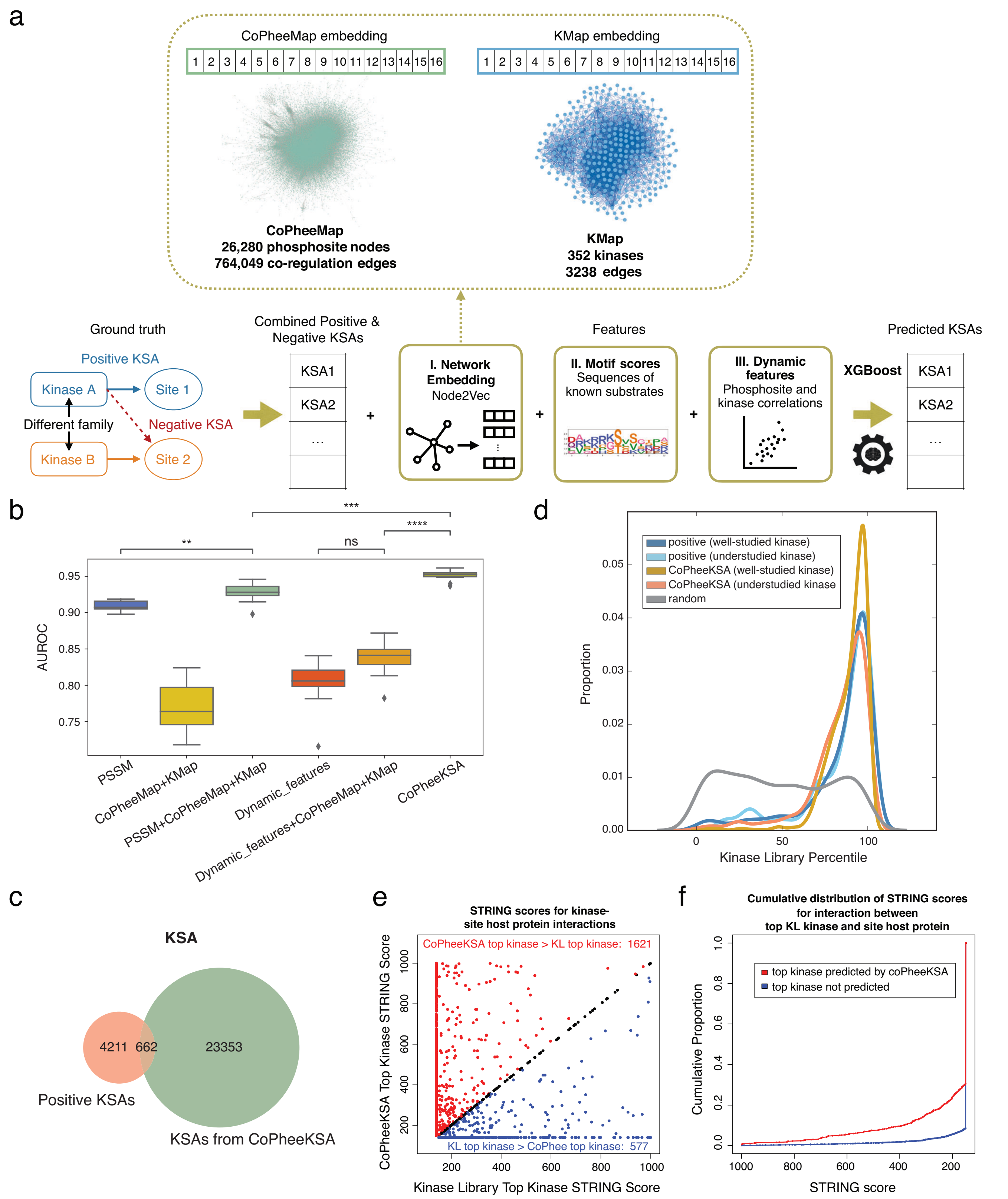
To capture relationships between each kinase and phosphosite pair, we first concatenated the corresponding kinase embedding from KMap with the respective phosphosite embedding from CoPheeMap. Next, we computed PSSM scores for each kinase-phosphosite pair using the consensus motif of known substrates for that kinase (Methods). Moreover, since kinases typically interact with phosphorylation sites in a dynamic manner, utilizing proteomics and phosphoproteomics data, we further computed correlations between kinase protein abundance and phosphosite abundance, as well as between inferred kinase activity and phosphosite abundance, in each CPTAC cohort to capture their dynamic relationships (Methods).

We used these features to train XGBoost models for KSA prediction (Fig. 3a) based on a carefully prepared ground truth KSA dataset (Supplementary Data 6, Methods). Of note, to avoid information leakage, we computed kinase activity scores for each positive KSA separately, removing the corresponding phosphosite in our computation. The model integrating all features achieved a median AUROC of 0.95 in 10 repetitions of Monte Carlo validation, significantly outperforming PSSM-based prediction and XGBoost classifiers using only subsets of features (Fig. 3b, Supplementary Fig. 3a), suggesting that these features held complementary information. To ensure high-quality predictions, we set a stringent prediction threshold of 0.76, corresponding to an LLR of 5.5 (Supplementary Fig. 3b). This means that the selected KSA predictions were 244 times ($e^{5.5} = 244$) more likely to be positive KSAs than negative KSAs within the independent ground-truth test data. Using this threshold, CoPheeKSA predicted 24,015 KSAs involving 9399 phosphosites and 104 kinases, including 26 understudied kinases (Supplementary Data 6). This represents a fivefold increase in the number of KSAs compared to the ground truth positives (Fig. 3c).

To assess the quality of the CoPheeKSA predictions, we leveraged substrate specificities experimentally determined via the KL approach[22]. For each kinase, this method ranked all 82,735 sites in a reference proteome based on their likelihood of being phosphorylated by the kinase. Each site was assigned a percentile score (Supplementary Data 7), with higher scores indicating greater kinase–substrate specificity. For well-studied kinases, KSAs predicted by CoPheeKSA showed significantly higher KL percentile scores than the ground-truth positive KSAs (two-sided T test, $p \leq 0.0001$; Fig. 3d, Supplementary Fig. 3c). For understudied kinases, no significant difference was observed between the score distributions of the two groups (two-sided T test, $p > 0.01$; Fig. 3d, Supplementary Fig. 3c). These results demonstrate the high quality of the CoPheeKSA predictions.

For comparison, we evaluated predicted kinases for all 77,442 quantified phosphosites in our PanCan data using other KSA prediction models, including NetworKIN[11], LinkPhinder[18] and PDT[12]. For both well-studied and understudied kinases, the KSAs predicted by these tools had significantly lower KL percentile scores compared to ground-truth KSAs (two-sided T test, $p \leq 0.0001$; Supplementary Fig. 3c). While LinkPhinder and PDT predicted more regulatory kinases per site than CoPheeKSA, the quality of their predictions was considerably lower, approaching the level of random pairs. NetworKIN predictions were of higher quality than those of LinkPhinder and PDT and closer to CoPheeKSA, but it identified far fewer KSAs compared to CoPheeKSA (Supplementary Fig. 3d). Notably, CoPheeKSA identifications also covered more understudied kinases (Supplementary Fig. 3e). Given that these two KSA prediction approaches employ different methodologies and have different phosphosite coverage scopes, it is not surprising that NetworKIN and CoPheeKSA made positive predictions for different phosphosite sets in the PanCan data (Supplementary Fig. 3f).

Many KSAs with high KL percentile scores (>99%) received low scores in CoPheeKSA predictions (Supplementary Fig. 3g). This discrepancy could be because KL scores represent a kinase's potential to phosphorylate a substrate in vitro, which might not translate to actual interactions in vivo due to cellular conditions or protein structure constraints. To investigate if CoPheeKSA, trained with dynamic tumor data, more accurately predicts in vivo KSAs, we used STRING functional association scores. For phosphosites with different top kinase predictions by CoPheeKSA and KL, CoPheeKSA's top kinase had a higher STRING score for 1621 sites (red points above the black line in Fig. 3e), compared to only 577 sites for KL's top kinase (blue points). Further analysis showed that KSAs predicted by CoPheeKSA generally had higher STRING scores than those not supported by it (Kolmogorov–Smirnov test, $p = 2.2 \times 10^{-16}$; Fig. 3f). Additionally, protein abundance of kinases supported by CoPheeKSA correlated more strongly with their associated sites compared to unsupported kinases (Kolmogorov–Smirnov test, $p = 2.2 \times 10^{-16}$; Supplementary Fig. 3h).



Because the KL approach captures only peptide-based in vitro binding information, these data suggest that, by incorporating dynamic in vivo information in CoPheeMap and CoPheeKSA, our approach can discern between KSAs supported by functional associations from those that are not. This is particularly useful for identifying the subset of high-scoring KSAs predicted by the KL approach that are more likely to manifest in vivo.

Evaluating KSA predictions

To gain more insights into the KSAs predicted by CoPheeKSA, we colored the phosphosites according to the up-stream kinases from prior knowledge and CoPheeKSA in the tSNE plot (Fig. 4a). The co-clustering of the predicted substrates with previously annotated substrates supports the ability of CoPheeMap to connect phosphosites regulated by the same or similar kinases.

Fig. 3 | The CoPheeKSA algorithm and evaluation of prediction results. **a** Overview of the ground-truth data and features used in the XGBoost machine learning algorithm for CoPheeMap-powered KSA (kinase-substrate association) prediction (CoPheeKSA). **b** AUROCs (Area Under the Receiver Operating Characteristic Curve) of the KSA classifiers trained with different feature combinations. The number of AUROCs for each class is 10. P-values derived from a two-sided T test with Bonferroni correction. ns: $0.05 < p \leq 1$, **$0.001 < p \leq 0.01$, ***$0.0001 < p \leq 0.001$, ****$p < 0.0001$. For boxplots, centerline indicates the median, box limits indicate upper and lower quartiles, whiskers indicate the 1.5 interquartile range. PSSM (Position-Specific Scoring Matrix). **c** Venn diagram comparing KSAs predicted by CoPheeKSA and the known KSAs used as ground truth positives. **d** The kinase library percentile score distributions for different groups of kinase-substrate pairs. **e** Scatter plot comparing STRING scores for the associations between the host protein of a phosphosite and two sets of kinases. Each dot represents a unique site, with the y-axis showing the STRING score for the kinase with the highest CoPheeKSA prediction score and x-axis showing the STRING score for the kinase with the highest Kinase Library percentile score. The red and blue dots highlight sites where the two methods predicted different top kinases. **f** Cumulative distributions of STRING scores for associations between the host protein of a phosphosite and the top kinase predicted by the kinase library. The KSAs were separated into two groups for comparison: predicted by CoPheeKSA (red) and not predicted by CoPheeKSA (blue). Source data are provided as a Source Data file.

For each kinase, we further compared the KL percentile score distributions among ground-truth positive KSAs, random KSAs, and CoPheeKSA predicted positive KSAs. For most kinases, both ground-truth positives and predicted positives had median KL percentiles over 90%, clearly higher than those of the random group (Fig. 4b). However, for some kinases, such as CAMK2G, CDK7, DYRK1B and PAK4, the distribution of the KL percentile scores for the ground-truth positives were similar to that for the random (Fig. 4b). Interestingly, for these kinases, the predictions from CoPheeKSA had much better alignments with the KL data, evidenced by significantly improved KL percentile scores.

The number of KSAs predicted by CoPheeKSA for each kinase is visualized in Fig. 4c. Not only did well-studied kinases like CDK8 and MAPK1 obtain more substrates, understudied kinases, such as CDK12, SGK3, SMG1, and NUAK1, were also associated with hundreds of predicted substrates. This could significantly enhance downstream analyses of understudied kinases. Accordingly, we constructed a comprehensive KSA database integrating the KSA information from the ground-truth positives and CoPheeKSA (Methods). Briefly, when a kinase had more than 100 known substrates, no predicted substrates were added. When a kinase had fewer than 100 known substrates, predicted substrates from CoPheeKSA ranked by predicted scores were added until the total reached 500. This KSA database included 7423 unique substrates (Supplementary Data 6).

As an additional method to validate the CoPheeKSA predicted KSAs, we utilized IDPpub, a tool for mining phosphosites from abstracts in PubMed[32]. Sentences extracted by IDPpub from the abstracts may include KSAs missed in the ground-truth dataset. A manual review of the IDPpub extracted evidence sentences confirmed 56 of the KSAs predicted by CoPheeKSA and refuted four (Supplementary Data 8). Some of the confirmed KSAs expanded the known substrates for well-studied kinases such as AKT1, CDK1, MAPK1, PKN1, SGK1, and SIK2. Importantly, supporting evidence was also identified for KSAs involving understudied kinases such as AKT3 and MAP4K1 (A.K.A. HPK1). This analysis shows that CoPheeKSA can capture KSAs supported by literature but not included in our ground-truth dataset.

Illuminating dark functional phosphosites

To illustrate the utility of CoPheeMap and CoPheeKSA for illuminating phosphosites of interest, we investigated phosphosites that had previously been prioritized by a machine learning algorithm as functionally important sites, but without known up-stream kinases. The functional scores of the phosphosites were computed through machine learning-based integration of 59 features indicative of proteomic, structural, regulatory, or evolutionary relevance[30] (Supplementary Data 9). Among the top 50 phosphosites in CoPheeMap with the highest functional scores, 39 had known up-stream kinases (Fig. 5). The higher ratio of phosphosites with known up-stream kinases in this group was expected because of the intrinsic bias toward well-studied phosphosites in the ground truth dataset used to train the machine learning model for functional scoring[30]. Despite the bias, 11 out of the 50 did not have known up-stream kinases.

CoPheeKSA predicted up-stream kinases for seven of these 11 functional sites (Fig. 5). One example is Ser 8 on CD74, a cell-surface receptor and oncogene that has significantly higher protein abundance in tumors compared to normal tissues in multiple cancer types[33–36]. RAC-gamma serine/threonine-protein kinase (AKT3) and serum/glucocorticoid regulated kinase family member 3 (SGK3) were predicted by CoPheeKSA as potential kinases regulating this functional site, and both kinases had KL percentile scores >99% for this phosphosite (Supplementary Data 9). Consistent with the previously reported role for CD74 in immune stimulation[37], AKT3 can be activated when immune cells are stimulated, performing essential functions in both innate and adaptive immune cells[38].

As another example, Thr 497 is located in the activation loop of PRKCA, a Protein Kinase C family member that has been reported to play important roles in different cellular processes[39]. For this phosphosite, Pyruvate Dehydrogenase Kinase 1 (PDK1) was a kinase predicted by CoPheeKSA (Fig. 5), and PDK1 also ranked first in the KL prediction (Supplementary Data 9). Although the database we used to generate ground truth data does not list any up-stream kinases for this site, PDK1 has been reported to phosphorylate Thr 497 on PRKCA, leading to secondary autophosphorylation events and conformational changes in the molecule[40].

The third example is Ser 226 on heat shock protein 90 alpha family class B member 1 (HSP90AB1). CoPheeKSA predicted Casein Kinase 2 Alpha 1 (CSNK2A1, A.K.A. CK2A1) as the regulatory kinase (Fig. 5), which was also ranked among the top by KL (Supplementary Data 9).

For the remaining four dark phosphosites without confident CoPheeKSA predictions, CoPheeMap connected them to neighbors with known up-stream kinases or functional annotation, which could provide useful information for these sites because neighboring sites in CoPheeMap are likely to be co-regulated. For example, Ser 25 on VIM, a type III intermediate filament protein, is connected to substrates of PRKACA, PRKCA, and PRKD1 in CoPheeMap (Fig. 5). The regulatory kinases of these neighboring substrates were derived from both ground truth KSAs and CoPheeKSA predicted KSAs. Supporting this network neighborhood-based inference, 11 AGC kinase family members, including PKCE, PKG2, PKCZ, PKCH, PKCT, PKCI, PKCG, PKCB, PKCD, PKG1, and PKACG ranked among the top 25 kinases predicted by the KL for Ser 25 on VIM (Supplementary Data 9).

Elucidating cancer-associated phosphosites and their regulatory kinases

To demonstrate the utility of CoPheeMap and CoPheeKSA in facilitating the analysis and interpretation of data from phosphoproteomics experiments, we conducted differential analysis of phosphosites between tumor and normal adjacent tissue (NAT) samples in each of the eight CPTAC cohorts with both tumor and NAT samples and then calculated meta-p values to identify phosphosites that were differentially regulated across multiple cancer types (Methods).

First, we examined individual sites among the top 50 with the most significant differential abundance between tumor and NAT samples. Although only 12 of them had an annotated kinase, CoPheeKSA and CoPheeMap provided insights for all 50 sites (Fig. 6a).

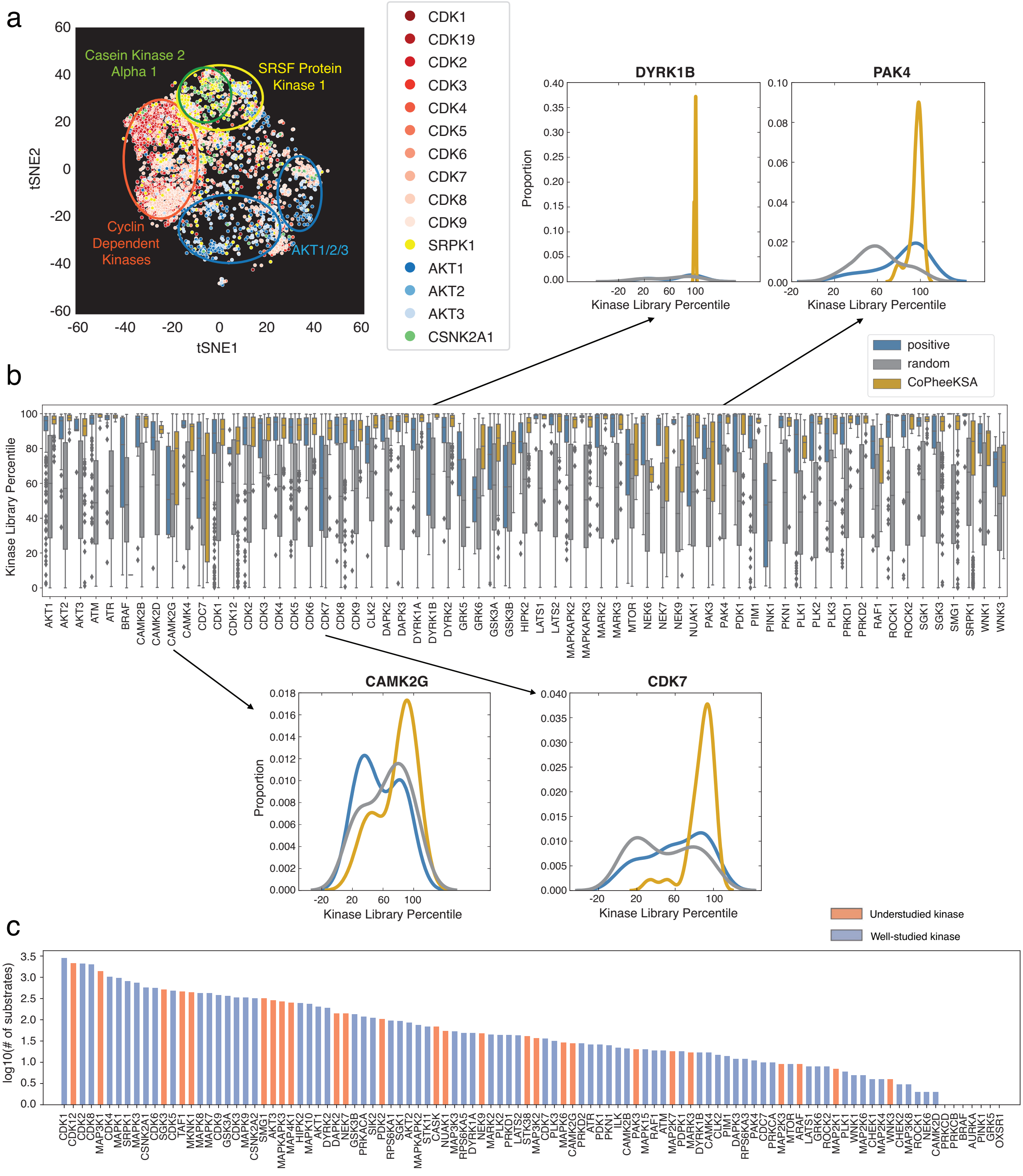


Fig. 4 | KSA predictions from CoPheeKSA. a tSNE visualization of the Node2Vec embedding of the phosphosites in CoPheeMap, with phosphosites color-coded by known or predicted kinase. Phosphosites with multiple predicted kinases are not included in this plot. **b** Kinase library percentile score distributions for individual kinases across three types of KSAs: ground-truth positive KSAs (positive), CoPheeKSA predicted KSAs (CoPheeKSA), and random KSAs. n(positive) = 1205, n(random) = 38,734, n(CoPheeKSA) = 18,439, detailed numbers for each kinase could be found in the Source Data. For boxplots, centerline indicates the median, box limits indicate upper and lower quartiles, whiskers indicate the 1.5 interquartile range. **c** Numbers of substrates predicted for each kinase. Source data are provided as a Source Data file.

Several of these sites, including the most significant ones, Ser 67 on Nucleolin (NCL) and Ser 153 on ESF1 nucleolar pre-rRNA processing protein homolog (ESF1), were predicted by CoPheeKSA to be regulated by CDK2 (Fig. 6a). Consistently, CDK2 was also the top kinase predicted by KL for these two sites (Supplementary Data 10). Although the abnormal activation of cyclin-dependent kinases (CDKs) is widely recognized for its association with tumor cell proliferation and cancer development[41], some of these predicted substrates could potentially act as critical effectors in these processes.

To systematically identify kinases regulating cancer-associated phosphosites across cancer types, we used the meta-p values from our pan-cancer tumor versus normal comparisons as inputs for

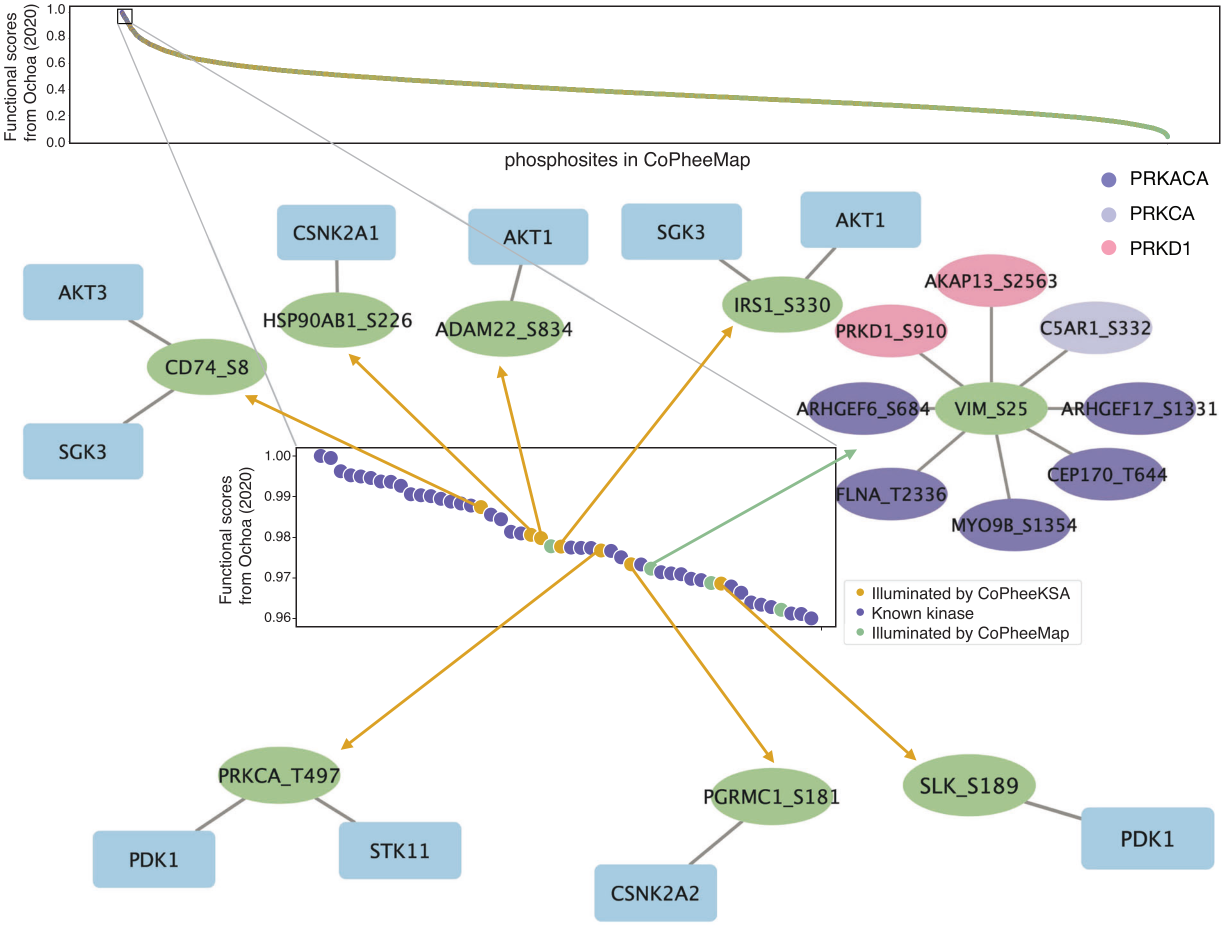


Fig. 5 | Illumination of phosphosites with the highest functional scores by CoPheeKSA and CoPheeMap. The functional scores were obtained from Ochoa et al. Nature Biotechnology[30]. Yellow dot: unannotated sites with kinases predicted by CoPheeKSA. Purple dot: sites with known kinase. Green dot: unannotated sites mapped in CoPheeMap. Source data are provided as a Source Data file.

phosphosite set enrichment analysis based on our expanded KSA database. Consistent with our initial observations that many individual phosphosites highly up-regulated in the PanCan datasets were regulated by CDKs, we confirmed the enrichment of CDK targets in this systematic analysis (Fig. 6b). While hyperactivation of CDK1, CDK2, CDK4, and CDK6 could be inferred without expanding the KSA database to include CoPheeKSA predictions, up-regulated activity for the less well-studied CDKs, including CDK12, CDK9, CDK7, CDK8, and CDK5, was only identifiable by adding targets predicted by CoPheeKSA. Using CDK12 as an example, among the 496 sites that contributed to the enrichment signal, 492 were uniquely associated with CDK12 by CoPheeKSA (Supplementary Fig. 4a). Although the two previously curated target sites of this kinase were up-regulated in cancer samples, without CoPheeKSA predictions, they were insufficient to produce statistically significant enrichment in phosphosite set enrichment analysis. Consistent with the substrate-based hyperactivation inference for CDK12 and other less well-studied CDKs, they all showed elevated protein abundance across cancer types (Supplementary Fig. 4b–f).

Our analysis also identified other kinases that would have not been identified without CoPheeKSA predictions (Fig. 6b). These included serine-arginine protein kinase 1 (SRPK1), TATA-box binding protein associated factor 1 (TAF1), Raf-1 proto-oncogene, serine/threonine kinase (RAF1), MAPK interacting serine/threonine kinase 1 (MKNK1), dual specificity tyrosine phosphorylation regulated kinase 1A (DYRK1A), and several mitogen-activated protein kinases (MAPKs). Many of these inferences were also supported by elevated mRNA and protein abundance of the kinases in tumors compared to NATs (Supplementary Fig. 4g–l). Moreover, some of these kinases, such as SRPK1, have been previously associated with cancer prognosis[42]. Thus, enrichment analysis using our expanded KSA database linked previously unannotated cancer-associated phosphosites to upstream kinases and associated numerous understudied kinases with human cancer, generating viable hypotheses for further experimental validation.

Discussion

The guilt-by-association strategy is commonly used to predict functions of understudied genes and proteins[14,16,43,44]. Leveraging the vast amount of CPTAC pan-cancer phosphoproteomics data, our study expands this strategy to address the dark phosphoproteome challenge through the construction of CoPheeMap, a machine-learned co-regulation network of the human phosphoproteome. Together with CoPheeKSA, which predicts KSAs using latent information embedded within CoPheeMap, these tools provide a comprehensive framework for investigating the regulation and functions of phosphosites, particularly in the context of human cancer.

We demonstrated that co-phosphorylation is an effective predictor of phosphosite co-regulation. It is important to note that the phosphosite abundance data used in this study were not normalized by host protein abundance. Some studies with matched phosphoproteomic and proteomic measurements employ such normalization based on the premise that it may better capture phosphosite abundance changes driven by kinase activity, independent of changes in host protein levels. In our analysis, positive site pairs showed significantly higher abundance correlations than negative pairs across cancer types, largely independent of correlations in their corresponding host protein levels (Fig. 1g, Supplementary Fig. 1). Importantly, many positive pairs displayed both high co-phosphorylation and host protein co-expression, and normalizing by protein abundance significantly reduced the calculated co-

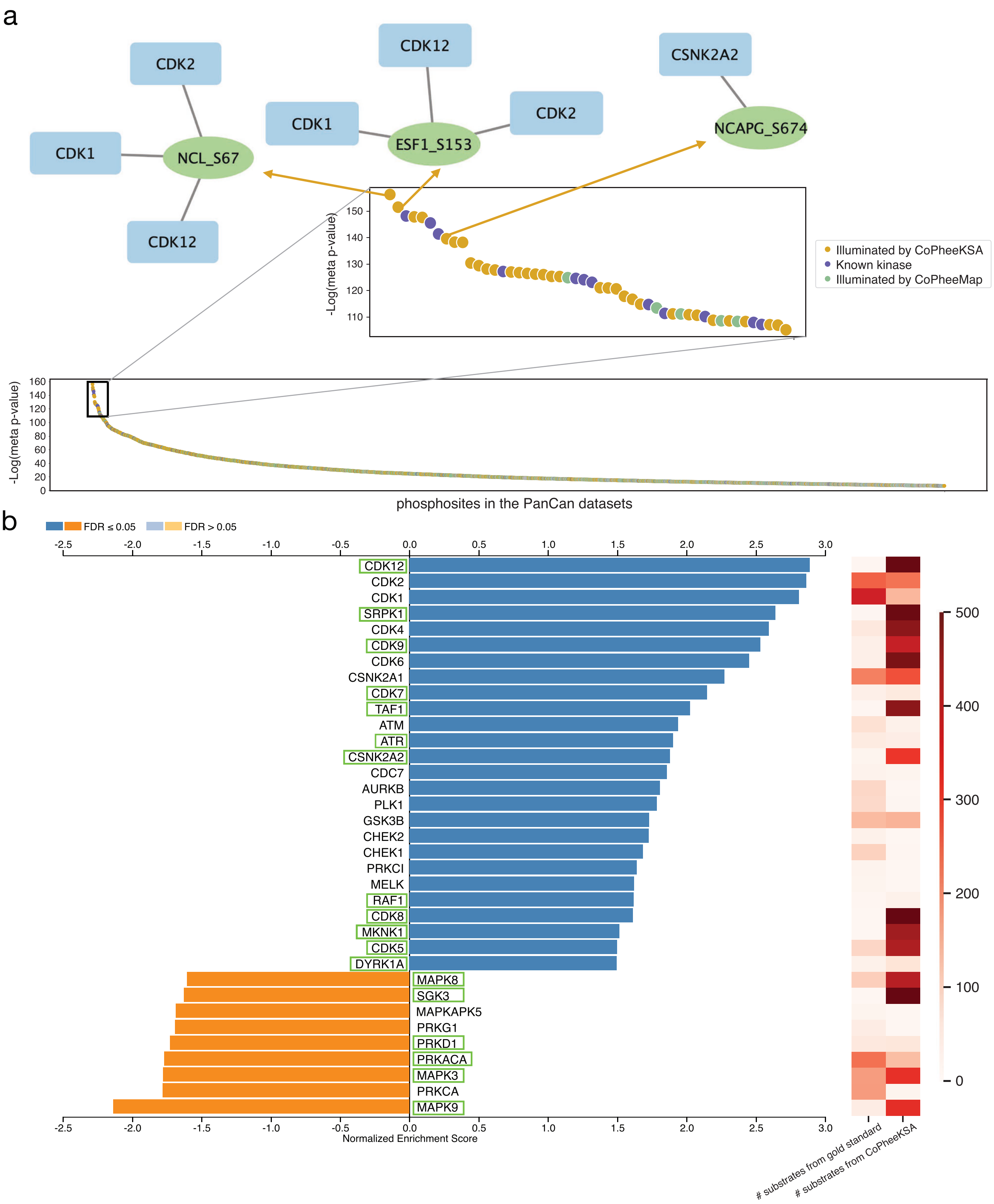


Fig. 6 | Cancer-associated phosphosites and differentially activated kinases identified from a pan-cancer analysis. a Phosphosites with the most significant differential abundance between tumor and normal samples in the pan-cancer analysis (two-sided unpaired Wilcoxon Rank Sum test, with meta p-values calculated using the sumz method). These sites were illuminated using CoPheeKSA and CoPheeMap after identification. Yellow dot: unannotated sites with kinases predicted by CoPheeKSA. Purple dot: sites with known kinase. Green dot: unannotated sites mapped in CoPheeMap. **b** Differentially activated kinases identified from the pan-cancer analysis. The heatmap shows the numbers of substrates for each kinase in the ground-truth dataset and CoPheeKSA predictions, respectively. Kinases with a green box are those identified using the comprehensive KSA database but not by the ground-truth dataset. Source data are provided as a Source Data file.

phosphorylation levels of these pairs. When we normalized phosphosite abundance by protein abundance, the PCC differences between positive and negative pairs decreased markedly compared to the results shown in Fig. 1f. As this study is based on steady-state data from human tumors rather than acute perturbation data, one possible explanation is that certain oncogenic signaling pathways exhibit concordant regulation of all components—kinases and substrates—at multiple levels, including expression, phosphorylation, and activity, to maximize pathway function. In such cases, normalizing phosphosite abundance by host protein abundance may inadvertently remove the impact of kinase activity, as it assumes that similar alterations in protein and phosphosite abundance reflect no change in kinase activity. Supporting this hypothesis, we recently showed in a benchmarking study that normalizing phosphosites by host protein levels in steady-state tumor data reduces kinase activity inference performance[17]. Together, these findings suggest that normalizing phosphosite abundance by host protein abundance may not always be advantageous, particularly in steady-state measurements.

Notably, co-phosphorylation data alone cannot distinguish between direct phosphorylation by the same kinase and indirect relationships mediated by different kinases in the same signaling cascade. To address this, sequence features were incorporated into both the CoPheeMap and CoPheeKSA models to help constrain predictions to direct relationships. The effectiveness of this approach is evident, as CoPheeKSA predictions demonstrated higher agreement with substrate specificities experimentally determined via the KL approach[22] compared to KSAs predicted by other methods (Supplementary Fig. 3c). Conversely, the KSAs included in PDT predictions, which rely solely on kinase inhibition profiles, performed poorly in this evaluation, underscoring the importance of integrating phosphoproteomics with sequence features.

For some kinases, the predictions from CoPheeKSA aligned much better with the KL scores than the ground truth positives, suggesting that our generic model may overcome noise in the ground truth data. Additionally, we observed that many KSAs with high scores in the Kinase Library had low scores in CoPheeKSA, suggesting that CoPheeKSA could be used to refine predictions from the KL approach. This is particularly valuable, as the KL approach evaluates a kinase's ability to phosphorylate a substrate under in vitro conditions, which may not accurately reflect in vivo interactions influenced by protein structure and the cellular environment.

Although comprehensive measurements of protein phosphorylation events are now possible, our ability to interpret phosphoproteomics data remains constrained by limited knowledge of the regulatory mechanisms and functional roles of many phosphosites. Using computationally predicted functional sites from a published study[30] and cancer-associated phosphosites identified through a pan-cancer analysis, we demonstrated how CoPheeMap and CoPheeKSA can generate hypotheses about the regulatory mechanisms and functional roles of biologically significant yet underexplored phosphosites and kinases. It is important to highlight that these are computational predictions, and the resulting hypotheses for individual phosphosites and kinases require further experimental validation. For instance, our analysis of the cancer-associated phosphosites predicted multiple hyperactivated kinases that have been largely overlooked but could be further explored as potential therapeutic targets. One such kinase is CDK12, which is implicated in regulating MYC expression, Wnt/β-catenin signaling, RNA splicing, and ErbB-PI3K-AKT signaling[45]. CDK12 inhibitors have been shown to suppress cancer cell transcription and growth and enhance drug susceptibility[45–48]. Further experimental research into CDK12 inhibitors could validate their therapeutic potential and identify phosphosite biomarkers for monitoring CDK12 activity and the effectiveness of its inhibitors.

We identified several areas for improvement in future research. Firstly, during the development of the ground truth of co-regulated phosphosite pairs for the construction of CoPheeMap, we found that certain kinases can phosphorylate both serine and tyrosine residues. This occurrence, though rare, introduces noise that may affect the training of machine learning models. Therefore, we limited our dataset to pairs of either Ser/Thr-Ser/Thr or Tyr-Tyr sites. Due to the scarcity of tyrosine data from the experimental methods used by CPTAC, particularly across multiple cancers, the phospho-tyrosine subnetwork in CoPheeMap includes only about 300 tyrosines. Consequently, CoPheeKSA was not applied to this subnetwork, limiting insights into tyrosine kinases, which are crucial targets for oncology and therapy development. Future studies should incorporate tyrosine-enriched datasets to enhance CoPheeMap's coverage. Additionally, in our construction of the CoPheeKSA models, the requirement for a kinase to have at least five known substrates left some serine/threonine kinases unexplored. Integrating Kinase Library data into our model training could be a significant enhancement.

Secondly, to minimize bias towards well-researched kinases, CoPheeMap focuses only on co-regulated site pairs without considering specific kinase information. This strategy allowed CoPheeMap to include a substantial number of unannotated phosphosites, and some even acted as hubs. However, sites regulated by well-known kinases are more likely to be the hubs in the network. This intrinsic bias in the training data is difficult to avoid because it is already present in the KSA databases used for ground truth construction. The use of the unbiased Kinase Library data may provide a solution to address this bias.

Thirdly, in constructing CoPheeMap, our ground truth data for co-regulation was derived from known KSAs. Notably, phosphatases also play a pivotal role in modulating phosphorylation dynamics. Because the co-phosphorylation and sequence similarity features utilized in the machine learning model are not constrained by kinase-specific regulation, CoPheeMap has the potential to capture phosphosites co-regulated by phosphatases, even though their regulation was not explicitly modeled. This inherent flexibility underscores the capability of CoPheeMap to identify regulatory relationships beyond kinase-mediated mechanisms, which may provide valuable insights into broader regulatory networks. To enhance the model further, future iterations could incorporate phosphatase interaction profile similarity as an additional feature in the XGBoost framework, alongside kinase interaction profile similarity, to better capture phosphatase-mediated co-regulation. Additionally, as more datasets on phosphatase-substrate relationships become available, these data can be integrated into the ground truth to provide a more balanced representation of phosphorylation dynamics. Such enhancements would allow for more holistic modeling of phosphosite co-regulation, ultimately improving our understanding of the interplay between kinases and phosphatases in complex signaling networks.

Finally, in our study, we utilized data from 11 CPTAC cancer types to construct the pan-cancer CoPheeMap. As additional phosphoproteomics data for other cancer types become available, we plan to incorporate them in our network construction process to improve the network's comprehensiveness. Our approach can also be used to integrate phosphoproteomics datasets from a single cancer type, allowing for the creation of cancer-type-specific CoPheeMaps. Furthermore, this strategy extends beyond cancer to enable the construction of disease-specific networks for various diseases, offering wide-ranging applications in biomedical research.

Methods

Pan-cancer omics data

Data for Breast cancer (BRCA), Clear Cell Renal Cell Carcinoma (CCRCC), Colon Adenocarcinoma (COAD), Glioblastoma (GBM), Head-and-neck squamous cell carcinoma (HNSCC), Lung squamous cell carcinoma (LSCC), Lung adenocarcinoma (LUAD), Ovarian Cancer (OV), Pancreatic Ductal Adenocarcinoma (PDAC) and Uterine Corpus Endometrial Carcinoma (UCEC) were generated by CPTAC and processed by the CPTAC pan-cancer working group, as described previously[23,49]. Processed RNA-Seq, proteomics, and phosphoproteomics data matrices were downloaded from the CPTAC pan-cancer portal at https://pdc.cancer.gov/pdc/cptac-pancancer. For proteomics, the file name is proteome_bcm_gencode_v34_harmonized_v1.zip. For Phosphoproteomics, the file name is phosphoproteome_bcm_gencode_v34_harmonized_v1.zip. For RNA-Seq, the file name is rna_bcm_v1.zip. Samples excluded in the original CPTAC studies[33–35,50–56] for quality and other issues were also excluded in the current study.

RNA-Seq, proteomics, and phosphoproteomics data for hepatocellular carcinoma (HCC) were obtained from a published proteogenomic study[57] and processed using the same CPTAC Pan-Cancer Working Group pipelines[23,49]. The processed data matrices are available at https://zenodo.org/records/14553766.

Ground-truth dataset for CoPheeMap development

Kinase substrate associations were downloaded from the supplementary data in GPS 5.0 (DOI: 10.1016/j.gpb.2020.01.001), which includes 14,679 curated experimentally identified associations. Based on the ground truth kinase substrate associations, we classified kinases as understudied kinases (≤10 known substrates) and well-studied kinases (>10 known substrates). We further annotated the kinases with kinase group information from KinBase, http://kinase.com/web/current/kinbase. Phosphosite pairs regulated by the same protein kinases were defined as the positive pairs. Three kinases—CSNK1A1, CSNK2A1, and CSNK2A2—were intentionally excluded from the ground-truth data for evaluation purposes. Phosphosite pairs regulated by the kinases from different kinase groups were defined as the negative pairs. We used fifteenmer (+/- 7 amino acids) to represent the positive and negative phosphosite pairs (Supplementary Data 2–3). If phosphosites were annotated to be regulated by multiple kinases, no overlapping known up-stream kinase families were allowed in the negative pairs. Only Ser/Thr-Ser/Thr (S/T-S/T) pairs or Tyr-Tyr (Y-Y) pairs were retained in the ground-truth dataset.

Sequence similarity scores

To calculate phosphopeptide sequence similarities, we used the BLOSUM62[58] matrix. The similarity score between the pair of fifteenmers (+/− 7 amino acids) of phosphosite A and phosphosite B was defined as the sum of the log-odds ratios from BLOSUM62 for all flanking positions (−7 to −1 and +1 to +7).

Kinase interaction profile similarity scores

We downloaded the STRING functional protein association network (protein.links.v11.5) and filtered it to retain links with STRING scores greater than 400, creating a dataset referred to as PPI_kinase_400. From this dataset, PPI_kinase_400, we identified, for each phosphosite, a list of kinases associated with its host protein, considering only kinase-protein pairs with STRING scores above 400. To calculate the kinase interaction profile similarity score for phosphosites a and b, we computed the Jaccard index of their respective kinase lists A and B:

Kinase Interaction Profile Similarity Score (a, b) = Jaccard Index (A, B) = J (A, B)

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|},$$

where

A = List of kinases associated with the host protein of phosphosite a in PPI_kinase_400

B = List of kinase associated with the host protein of phosphosite b in PPI_kinase_400

The Jaccard index measures the similarity between the kinase lists, providing a quantitative metric for kinase interaction profile similarity.

Co-phosphorylation versus co-regulation

For each phosphosite pair and their corresponding host protein pair, we required quantification in at least 20 overlapping samples within a single cancer type/cohort to calculate Spearman's correlation coefficient. We then grouped all the phosphosite pairs into 361 bins (19 × 19) based on their site-site correlation and host protein-host protein correlation. Each bin was defined by ranges of correlation values in two dimensions: site-site correlations on one axis and host protein-host protein correlations on the other axis. Both site-site correlations and host protein-host protein correlations ranged from −0.9 to 1, divided into increments of 0.1. This means that each bin contained pairs whose site-site correlations and host protein-host protein correlations fell within specific 0.1-wide intervals (e.g., a bin might include site-site correlations between −0.9 and −0.8 and host protein-host protein correlations between −0.8 and −0.7). The log-likelihood ratios (LLRs) of co-regulation were defined as the fraction of ground-truth positive site pairs divided by the fraction of negative site pairs in each of the 361 bins. Specifically, $LLR=\ln\left(\frac{C(\mathrm{PP}\ \&\ \mathrm{P})/C(\mathrm{PP}\ \&\ \mathrm{N})}{C(P)/C(N)}\right)$, where PP represents the set of positive pairs in a bin (or based on other inclusion criteria, such as predicted scores in later applications in the paper). P and N denote the sets of positive and negative pairs in the background dataset, respectively. Set intersection is denoted by &, while the function $C(\cdot)$ returns the size of a set.

Data for CoPheeMap training, validation, testing, and application

We divided all phosphosites in the ground-truth phosphosite pairs into training, validation, and test groups, ensuring no overlap to prevent information leakage (Supplementary Fig. 2a). Phosphosite pairs within the same group were included in their respective sets. Pairs between training and validation sites were assigned to the validation set, while any pairs involving test sites were assigned to the test set.

For Ser/Thr phosphosites, the ratio between training + validation and test sites, as well as between training and validation sites, was 9:1. For Tyr phosphosites, due to a smaller sample size, the ratio was 5:5. The ratios of positive to negative pairs in the training, validation, and test data were 1:5, 1:5, and 1:50, respectively (Supplementary Fig. 2b).

Over three billion phosphosite pairs had at least one site-level correlation calculated in one or more CPTAC cohorts. All these site pairs were included as the application data.

Extreme Gradient Boosting (XGBoost) classifier for CoPheeMap development

XGBoost was trained on the ground truth data to generate models for classifying each pair of sites as being co-regulated by the same kinase or not. For each pair of sites, there were 14 input features: 11 for pairwise phosphosite correlations from the 11 CPTAC cohorts, one for sequence similarity score, one for kinase interaction profile similarity score, and one label indicating Ser/Thr-Ser/Thr or Tyr-Tyr. Missing values were retained and handled by XGBoost.

We fine-tuned the XGBoost hyper parameters using the validation data. Three classifiers were trained: one using all the 14 features, one using the 11 dynamic features (co-phosphorylation), and one using the two static features (sequence similarity and kinase interaction profile similarity). AUROCs were calculated for each classifier using the test data.

Network embedding and dimensionality reduction

The networks (CoPheeMap and KMap) were embedded using the Node2Vec algorithm[27] with the following parameters: dimensions = 16, window = 1, min_count = 1, batch_words = 4. tSNE was used to reduce the dimensionality of CoPheeMap embedding features from 16 to 2, using parameters recommended for large network analysis[29].

Network visualization

Node link diagrams of the networks were generated using Cytoscape 3.9.1 or Gephi 0.10.1.

Shortest path length (SPL)

The iGraph R (networkx Python) package was used to calculate the SPL between each pair of nodes in CoPheeMap that belong to the indicated group (i.e., targets of CSKN2A1 or sites annotated as regulating the same functional category) or a set of random pairs of the same size.

Analysis of functionally annotated sites

We downloaded the Regulatory_sites database file from PhosphoSitePlus (phosphosite.org) on October 26, 2023. The ON_PROCESS column in this file includes information about cellular processes that are annotated as being regulated by specific phosphosites. After filtering the sites in this file to those that matched the host protein HGNC symbols and fifteenmers of the sites included in the CoPheeMap embedding, sites were grouped into 5 selected categories with sufficient numbers of sites ($n > 100$) associated with functionally distinct cellular processes included in the network. These groupings included cell growth and proliferation (all ON_PROCESS categories including cell growth or cell cycle regulation, regardless of direction of regulation), gene product regulation (ON_PROCESS categories including the key words transcription, translation, chromatin organization, RNA splicing, or RNA stability), cellular degradation (ON_PROCESS categories including apoptosis or autophagy), tumor microenvironment and mobility (ON_PROCESS categories including cytoskeletal reorganization or cell motility), and signal transduction (ON_PROCESS categories including signaling).

Phosphosite functional scores

The functional scores were acquired from the functional landscape of human phosphoproteome paper[30] (DOI: 10.1038/s41587-019-0344-3). The sites were mapped using UniProt IDs and site positions.

Kinase network (KMap)

The links between kinases were defined as the STRING protein-protein association score >400 or at least one protein-protein correlation >0.5 in one CPTAC cohort. This network covered 352 kinases connected by 3,238 edges (Supplementary Data 5).

Ground-truth data for CoPheeKSA development

To construct the positive KSAs for CoPheeKSA development, we overlapped the substrates from the ground truth with CoPheeMap (Ser/Thr) using fifteenmer sequences (+/− 7 amino acids), identifying 2353 positive KSAs.

For negative KSAs, we assigned known substrates of kinase A to kinase B if kinase A and B belonged to different kinase families and both phosphorylated Ser/Thr. If phosphosites were annotated as regulated by multiple kinases, the negative KSAs included only those where the site was not a target of any overlapping upstream kinases from the kinase groups of the kinases known to regulate the site. This approach yielded 114,530 negative KSAs (Supplementary Data 6).

Kinase activity scores

The kinase activity scores for a sample were calculated as the mean values of all known substrates from the ground truth quantified by phosphoproteomics for the sample. Each kinase activity score required the quantification of at least three substrates. To prevent information leakage, phosphosites in the test set were excluded before calculating the kinase activity scores using the known substrates. Additionally, for each KSA in the ground truth where the target site was on the kinase itself, the site abundance was excluded from the calculation of the kinase's activity score.

Kinase motif scores

We derived kinase motifs from the positive ground-truth KSAs by collecting substrate sequences for each kinase, ensuring at least five substrates per kinase. For each kinase, using the 15-mer sequences of the substrates, we constructed a PSSM by calculating the frequency of each amino acid at each position and converting these frequencies to probabilities. The PSSMs were then used to score new substrates by summing the probabilities for each position. To prevent information leakage, test sites were excluded when deriving PSSMs for scoring the test sites.

Correlations between kinase and substrate

For correlations between kinase protein abundance/kinase activity scores and phosphorylation abundance, we require quantification in at least 20 overlapping samples in one cohort to calculate the Spearman's correlation coefficient (PCC).

Data for CoPheeKSA training, validation, testing, and application

The KSAs from different kinases were combined to generate positive and negative KSAs. The positive KSAs were divided into training and test groups for Monte Carlo cross-validation (10 iterations). Each test group contained 235 positive KSAs. The test groups were further split into validation and test subsets, with 25 positive KSAs in each validation set. To minimize the impact of missing values, the ratio of missing values for the dynamic features was maintained consistently across the groups. All potential kinase-substrate associations were prepared as the application data.

XGBoost classifier for CoPheeKSA development

XGBoost was trained using ground truth KSA data to classify each potential KSA (kinase-site pair) as positive or negative. Each KSA was represented by 55 features: 16 features from CoPheeMap embedding, 16 features from KMap embedding, 22 dynamic features representing correlations between the kinase and substrate, and one static feature representing the kinase motif score.

We retained missing values for dynamic features. To address the higher ratio of missing values in the kinase activity scores for the positive KSAs due to the exclusion of site abundance when calculating these scores, we randomly removed some dynamic feature values in the negative data to match the missing value ratio of the positive KSAs to avoid bias. The ratio of positive to negative KSAs in the training, validation, and test data was maintained at 1:10.

Various classifiers were trained using different combinations of features: all features, only motif scores, only network embedding features, network embedding features and motif scores, only dynamic features, and network embedding features and dynamic features. AUROCs were calculated for each classifier. The XGBoost parameters were {max_depth: 2, eta: 0.2, objective: binary: logistic, num_round: 300} with AUC as the evaluation metric.

The kinase library scores

Percentile scores for each phosphosite and kinase pair were computed as described in Johnson, et al.[22]. Briefly, all phosphosites in this study were scored by all the characterized kinases (303 S/T kinases). Their ranks within an a-priori score distribution, based on curated phosphoproteome, were then determined to yield the percentile scores.

NetworKIN prediction

We used the webtool of NetworKIN[11] to predict the up-stream kinases for all the phosphosites quantified in the PanCan datasets, using a threshold of 5.

Comparison of STRING PPI scores and pan-cancer kinase-site correlations for CoPheeKSA and KL KSAs

To assess whether CoPheeKSA better captures KSA relationships that are more likely to occur in vivo compared to predictions from the kinase library, we utilized STRING scores as a measure of the biological relevance of KSAs. Specifically, we extracted functional association scores from the 9606.protein.links.v12.0.txt file from STRING (string-db.org). For each site with a predicted KSA by CoPheeKSA (prediction score > 0.7676), we considered the top-scoring kinase identified by CoPheeKSA or the top-scoring kinase (percentile score) from the KL prediction. In cases where multiple kinases were predicted for the same site by CoPheeKSA, only the highest-scoring kinase was included in the analysis in Fig. 3e.

Since STRING only provided scores for interactions with a minimum value of 150, any interaction not included in the STRING dataset was assumed to have little or no support and set to 149 unless scores were not available for the best kinase nominated by both CoPheeKSA and the KL, in which case the site was excluded. For the analysis in Fig. 3f, a score of 149 was assigned to all KSAs with no STRING scores.

We also compared KSAs from CoPheeKSA and the KL by evaluating the correlations between the kinase protein levels and the predicted target phosphosite levels using the results from a recently published analysis of pan-cancer associations in CPTAC datasets[49]. For the distribution plot shown in Fig. S3h, log10 meta-p values from combining the p-values from Spearman rank correlations between the kinase protein and target phosphosite levels across ten cancer types (calculated as described in the Liao, et al. paper[49]) were signed by direction of correlation; for each site with a KSA prediction from CoPheeMap, the kinase with the top prediction score for that site was chosen to compare with the top kinase from the KL (based on percentile score), and the cumulative distributions for the two sets as the signed log10 p-values decrease are plotted.

IDPpub validation

We downloaded evidence sentence data from IDPpub[32] available at https://idppub.ptmax.org/ and searched for exact matches for substrate HGNC symbols and site amino acid positions corresponding to all KSAs predicted by CoPheeKSA. Each matched sentence was then cross-referenced with the kinases_for_idppub.txt file, which lists any kinases mentioned in the evidence sentences, to search for references to kinases predicted by CoPheeKSA. Given that IDPpub was specifically trained to identify phosphosites rather than KSAs, most of the identified sentences simply co-mention the kinase and the phosphosite without describing their regulatory relationship. To address this, we conducted a manual review of the sentences to determine which explicitly supported or refuted the predicted KSAs. The validated or refuted KSAs from this review were documented in Supplementary Data 8.

KSA database

We constructed the KSA database in the gmt file format for GSEA/ssGSEA analysis (Supplementary Data 6). When a kinase has more than 100 known substrates, no predicted substrates were added. When kinase has fewer than 100 known substrates, the predicted substrates from CoPheeKSA were added up to the maximum number of 500 (ranked by predicted scores).

Tumor versus normal comparison

Tumor and normal samples from 8 CPTAC cancer types (CCRCC, COAD, HNSCC, LSCC, LUAD, OV, PDAC, and UCEC) underwent differential expression analysis using both proteomics and phosphoproteomics data. Proteins were required to be detected in at least 20 tumor samples and 10 normal samples for proteomics and phosphoproteomics datasets. Statistical significance was determined using the unpaired Wilcoxon Rank Sum test. Meta p-values were calculated with the sumz method from the R package metap (V1.4). P-values of individual cohorts were first converted to one-sided p-values and the sign for p-values not consistent with the majority were reversed. The calculated meta p-value was converted back to two-sided p-values and then the major sign of association was added.

Gene set enrichment analysis (GSEA)

We used Webgestalt (https://www.webgestalt.org/) to conduct the GSEA on the phosphosites, using −log10(p-values) as the input[59]. The organism was set to others. The functional database was uploaded using self-defined gmt files.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

For BRCA, CCRCC, COAD, GBM, HNSCC, LSCC, LUAD, OV, PDAC and UCEC, processed RNA-Seq, proteomics, and phosphoproteomics data matrices were downloaded from the CPTAC pan-cancer portal at https://pdc.cancer.gov/pdc/cptac-pancancer. For proteomics, the file name is proteome_bcm_gencode_v34_harmonized_v1.zip. For Phosphoproteomics, the file name is phosphoproteome_bcm_gencode_v34_harmonized_v1.zip. For RNA-Seq, the file name is rna_bcm_v1.zip. For HCC, processed data matrices are available at https://zenodo.org/records/14553766. Source data are provided with this paper.

Code availability

The code for CoPheeMap and CoPheeKSA is available on GitHub at: https://github.com/bzhanglab/CoPheeMap[60].

References

1. Cohen, P. The origins of protein phosphorylation. *Nat. Cell Biol.* **4**, E127–E130 (2002).
2. Hunter, T. Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell* **80**, 225–236 (1995).
3. Blume-Jensen, P. & Hunter, T. Oncogenic kinase signalling. *Nature* **411**, 355–365 (2001).
4. Rikova, K. et al. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell* **131**, 1190–1203 (2007).
5. Zanivan, S. et al. In vivo SILAC-based proteomics reveals phosphoproteome changes during mouse skin carcinogenesis. *Cell Rep.* **3**, 552–566 (2013).
6. Cohen, P., Cross, D. & Jänne, P. A. Kinase drug discovery 20 years after imatinib: progress and future directions. *Nat. Rev. Drug Discov.* **20**, 551–569 (2021).
7. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
8. Gerritsen, J. S. & White, F. M. Phosphoproteomics: a valuable tool for uncovering molecular signaling in cancer cells. *Expert Rev. Proteom.* **18**, 661–674 (2021).
9. Needham, E. J., Parker, B. L., Burykin, T., James, D. E. & Humphrey, S. J. Illuminating the dark phosphoproteome. *Sci. Signal.* **12**, eaau8645 (2019).
10. Savage, S. R. & Zhang, B. Using phosphoproteomics data to understand cellular signaling: a comprehensive guide to bioinformatics resources. *Clin. Proteom.* **17**, 27 (2020).
11. Horn, H. et al. KinomeXplorer: an integrated platform for kinome biology studies. *Nat. Methods* **11**, 603–604 (2014).
12. Hijazi, M., Smith, R., Rajeeve, V., Bessant, C. & Cutillas, P. R. Reconstructing kinase network topologies from phosphoproteomics data reveals cancer-associated rewiring. *Nat. Biotechnol.* **38**, 493–502 (2020).

13. Watson, N. A. et al. Kinase inhibition profiles as a tool to identify kinases for specific phosphorylation sites. *Nat. Commun.* **11**, 1684 (2020).
14. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551–562 (2017).
15. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol. Syst. Biol.* **3**, 88 (2007).
16. Shi, Z., Lei, J. T., Elizarraras, J. M. & Zhang, B. Mapping the functional network of human cancer through machine learning and pan-cancer proteogenomics. *Nat. Cancer* https://doi.org/10.1038/s43018-024-00869-z (2024).
17. Müller-Dott, S. et al. Comprehensive evaluation of phosphoproteomic-based kinase activity inference. bioRxiv https://doi.org/10.1101/2024.06.27.601117 (2024).
18. Nováček, V. et al. Accurate prediction of kinase-substrate networks using knowledge graphs. *PLoS Comput. Biol.* **16**, e1007578 (2020).
19. Yılmaz, S. et al. Robust inference of kinase activity using functional networks. *Nat. Commun.* **12**, 1177 (2021).
20. Giudice, G., Chen, H., Koutsandreas, T. & Petsalaki, E. phuEGO: a network-based method to reconstruct active signaling pathways from phosphoproteomics datasets. *Mol. Cell Proteom.* **23**, 100771 (2024).
21. Ayati, M., Chance, M. R. & Koyutürk, M. Co-phosphorylation networks reveal subtype-specific signaling modules in breast cancer. *Bioinformatics* **37**, 221–228 (2021).
22. Johnson, J. L. et al. An atlas of substrate specificities for the human serine/threonine kinome. *Nature* **613**, 759–766 (2023).
23. Li, Y. et al. Proteogenomic data and resources for pan-cancer analysis. *Cancer Cell* **41**, 1397–1406 (2023).
24. Wang, C. et al. GPS 5.0: an update on the prediction of kinase-specific phosphorylation sites in proteins. *Genomics Proteom. Bioinforma.* **18**, 72–80 (2020).
25. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, USA, 2016). https://doi.org/10.1145/2939672.2939785.
26. Kojaku, S., Radicchi, F., Ahn, Y.-Y. & Fortunato, S. Network community detection via neural embeddings. *Nat. Commun.* **15**, 9446 (2024).
27. Grover, A. & Leskovec, J. node2vec: scalable feature learning for networks. *KDD* **2016**, 855–864 (2016).
28. Maaten, L. V. D. & Hinton, G. E. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* (2008).
29. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, 5416 (2019).
30. Ochoa, D. et al. The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **38**, 365–373 (2020).
31. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet* **5**, 101–113 (2004).
32. Savage, S. R. et al. IDPpub: illuminating the dark phosphoproteome through pubmed mining. *Mol. Cell. Proteom.* **23**, 100682 (2024).
33. Clark, D. J. et al. Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell* **179**, 964–983.e31 (2019).
34. Cao, L. et al. Proteogenomic characterization of pancreatic ductal adenocarcinoma. *Cell* **184**, 5031–5052.e26 (2021).
35. Dou, Y. et al. Proteogenomic characterization of endometrial carcinoma. *Cell* **180**, 729–748.e26 (2020).
36. Zhang, H. et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* **166**, 755–765 (2016).
37. Naujokas, M. F., Morin, M., Anderson, M. S., Peterson, M. & Miller, J. The chondroitin sulfate form of invariant chain can enhance stimulation of T cell responses through interaction with CD44. *Cell* **74**, 257–268 (1993).
38. Cantrell, D. Protein kinase B (Akt) regulation and function in T lymphocytes. *Semin. Immunol.* **14**, 19–26 (2002).
39. Black, A. R. & Black, J. D. The complexities of PKCα signaling in cancer. *Adv. Biol. Regul.* **80**, 100769 (2021).
40. Jones, D. PRKCa: Identification of a Novel Downstream Target of WT1. (The Texas Medical Center Library, 2013).
41. Ghafouri-Fard, S. et al. A review on the role of cyclin dependent kinases in cancers. *Cancer Cell Int* **22**, 325 (2022).
42. Nikas, I. P., Themistocleous, S. C., Paschou, S. A., Tsamis, K. I. & Ryu, H. S. Serine-arginine protein kinase 1 (srpk1) as a prognostic factor and potential therapeutic target in cancer: current evidence and future perspectives. *Cells* **9**, 19 (2019).
43. Mohsen, H. et al. Network propagation-based prioritization of long tail genes in 17 cancer types. *Genome Biol.* **22**, 287 (2021).
44. Wang, J. et al. Proteome profiling outperforms transcriptome profiling for coexpression based gene function prediction. *Mol. Cell Proteom.* **16**, 121–134 (2017).
45. Lui, G. Y. L., Grandori, C. & Kemp, C. J. CDK12: an emerging therapeutic target for cancer. *J. Clin. Pathol.* **71**, 957–962 (2018).
46. Liang, S. et al. CDK12: A potent target and biomarker for human cancer therapy. *Cells* **9**, 1483 (2020).
47. Krajewska, M. et al. CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation. *Nat. Commun.* **10**, 1–16 (2019).
48. Liu, H., Liu, K. & Dong, Z. Targeting CDK12 for cancer therapy: function, mechanism, and drug discovery. *Cancer Res* **81**, 18–26 (2021).
49. Liao, Y. et al. A proteogenomics data-driven knowledge base of human cancer. *Cell Syst.* **14**, 777–787.e5 (2023).
50. Wang, L.-B. et al. Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* **39**, 509–528.e20 (2021).
51. Vasaikar, S. et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell* **177**, 1035–1049.e19 (2019).
52. Satpathy, S. et al. A proteogenomic portrait of lung squamous cell carcinoma. *Cell* **184**, 4348–4371.e40 (2021).
53. Krug, K. et al. Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. *Cell* **183**, 1436–1456.e31 (2020).
54. Huang, C. et al. Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. *Cancer Cell* **39**, 361–379.e16 (2021).
55. Hu, Y. et al. Integrated proteomic and glycoproteomic characterization of human high-grade serous ovarian carcinoma. *Cell Rep.* **33**, 108276 (2020).
56. Gillette, M. A. et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* **182**, 200–225.e35 (2020).
57. Gao, Q. et al. Integrated proteogenomic characterization of hbv-related hepatocellular carcinoma. *Cell* **179**, 561–577.e22 (2019).
58. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* **89**, 10915–10919 (1992).
59. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* **47**, W199–W205 (2019).
60. Jiang, W. Deciphering the dark cancer phosphoproteome using machine-learned co-regulation of phosphosites. *Zenodo* https://doi.org/10.5281/zenodo.14947153 (2025).

Acknowledgements

We gratefully acknowledge contributions from the CPTAC and its Pan-Cancer Analysis Working Group. This work was supported by National Institutes of Health (NIH) grants from the National Cancer Institute (NCI) U24 CA271076 (BZ), R01 CA245903 (BZ), and U01 CA271247 (BZ), by Cancer Prevention & Research Institute of Texas (CPRIT) Award RP220050 (BZ) and RP210027, which supports WJ under the Baylor College of Medicine Comprehensive Cancer Training Program. Funding was also received from the McNair Medical Institute at The Robert and Janice McNair Foundation (BZ). B.Z. is a CPRIT Scholar in Cancer Research and a McNair Scholar. The research carried out in Lewis Cantley's laboratory was supported by grant R35 CA197588. J.L.J. was supported by The Claudia Adams Barr Program for Cancer Research Award.

Author contributions

Conceptualization, B.Z., W.J.; Methodology, W.J.; Formal Analysis, W.J., E.J.J., T.M.Y., J.L.J.; Investigation, W.J., E.J.J., B.Z.; Resources, Y.L., Z.S., T.M.Y., J.L.J., L.C.C.; Data Curation, W.J.; Writing - Original Draft, W.J., E.J.J., B.Z.; Visualization, W.J., E.J.J.; Supervision, B.Z., L.C.C.; Funding Acquisition, B.Z.

Competing interests

B.Z. received research funding from AstraZeneca and consulting fees from Inotiv. L.C.C. is a founder and member of the board of directors of Agios Pharmaceuticals and is a founder and receives research support from Petra Pharmaceuticals; is listed as an inventor on a patent (WO2019232403A1, Weill Cornell Medicine) for combination therapy for PI3K-associated disease or disorder, and the identification of therapeutic interventions to improve response to PI3K inhibitors for cancer treatment; is a co-founder and shareholder in Faeth Therapeutics; has equity in and consults for Cell Signaling Technologies, Volastra, Larkspur and 1 Base Pharmaceuticals; and consults for Loxo-Lilly. J.L.J. has received consulting fees from Scorpion Therapeutics and Volastra Therapeutics. T.M.Y.-B. is a co-founder of DeStroke. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57993-2.

Correspondence and requests for materials should be addressed to Bing Zhang.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.