Received: 17 February 2022 | Revised: 26 April 2022 | Accepted: 3 May 2022

DOI: 10.1002/pmic.202200068

Proteomics
Proteomics and Systems Biology

REVIEW

Computational systems approach towards phosphoproteomics and their downstream regulation

Di Xiao[1,2] | Carissa Chen[1,2] | Pengyi Yang[1,2,3]

[1]Computational Systems Biology Group, Children's Medical Research Institute, The University of Sydney, Westmead, New South Wales, Australia

[2]Charles Perkins Centre, The University of Sydney, Sydney, New South Wales, Australia

[3]School of Mathematics and Statistics, The University of Sydney, Sydney, New South Wales, Australia

Correspondence
Pengyi Yang, Computational Systems Biology Group, Children's Medical Research Institute, The University of Sydney, Westmead, 2145 NSW, Australia.
Email: pengyi.yang@sydney.edu.au

Funding information
National Health and Medical Research Council Investigator Grant, Grant/Award Number: 1173469

Abstract

Protein phosphorylation plays an essential role in modulating cell signalling and its downstream transcriptional and translational regulations. Until recently, protein phosphorylation has been studied mostly using low-throughput biochemical assays. The advancement of mass spectrometry (MS)-based phosphoproteomics transformed the field by enabling measurement of proteome-wide phosphorylation events, where tens of thousands of phosphosites are routinely identified and quantified in an experiment. This has brought a significant challenge in analysing large-scale phosphoproteomic data, making computational methods and systems approaches integral parts of phosphoproteomics. Previous works have primarily focused on reviewing the experimental techniques in MS-based phosphoproteomics, yet a systematic survey of the computational landscape in this field is still missing. Here, we review computational methods and tools, and systems approaches that have been developed for phosphoproteomics data analysis. We categorise them into four aspects including data processing, functional analysis, phosphoproteome annotation and their integration with other omics, and in each aspect, we discuss the key methods and example studies. Lastly, we highlight some of the potential research directions on which future work would make a significant contribution to this fast-growing field. We hope this review provides a useful snapshot of the field of computational systems phosphoproteomics and stimulates new research that drives future development.

KEYWORDS
bioinformatics, cell biology, data processing and analysis, phosphoproteomics, signal transduction, systems biology, technology

1 | INTRODUCTION

Protein phosphorylation is one of the most common post-translational modifications (PTMs) that regulates almost every aspect of protein function, ranging from modulating their dynamics, stability, subcellular localisation to protein- protein interactions [1] (Figure 1). Phosphorylation events act as reversible molecular switches and are controlled by kinases and phosphatases [2]. Together, kinases, phosphatases, and their substrate proteins [3] establish signalling networks that modulate a myriad of biological processes, spanning from cell cycle progression, cell growth, differentiation, and apoptosis [1]. Thus, dysfunctions in phosphorylation-based signalling (phospho-signalling) networks can severely disrupt cellular homeostasis and are involved in many diseases, including metabolic diseases [4] and cancers [5]. The reconstruction and characterisation of phospho-signalling networks therefore have provided invalu able biological knowledge into various cellular processes and contributed promising insights towards therapeutic drug development for the treatment of various diseases [6].

While phospho-signalling has been studied for decades using a variety of experimental approaches, the advances in mass spectrometry

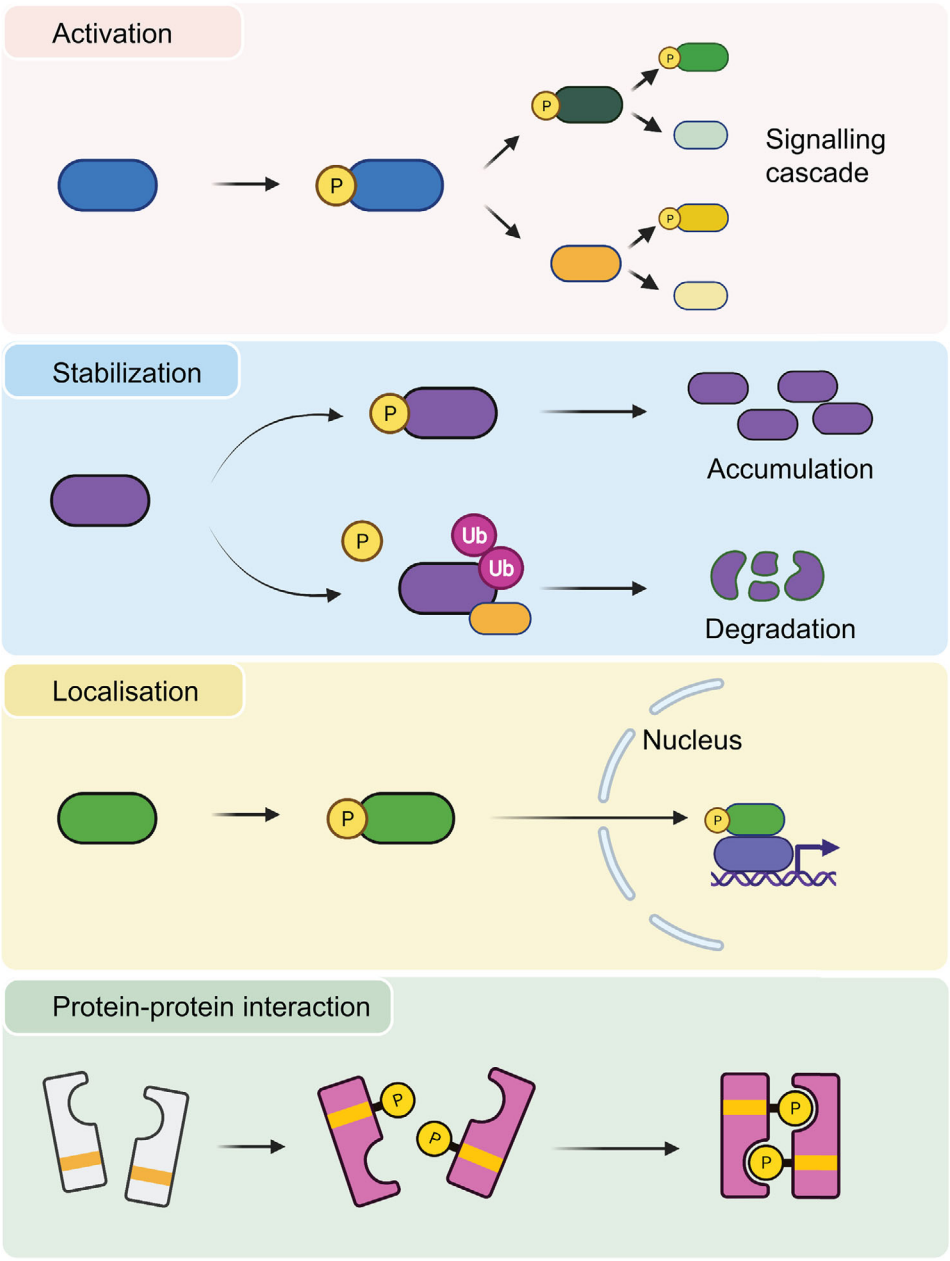

FIGURE 1 A schematic of some key functionalities of phosphorylation events in regulating cellular processes

(MS)-based technologies to measure global phosphorylation events have revolutionised our abil ity to profile phosphorylation events at a large-scale and enables systematic analyses of protein phosphorylation in a high-throughput manner [7]. To date, more than 377,000 non-redundant protein phosphorylation sites from 27 species were collected in PhosphoSitePlus, a curated PTM database [8]. However, phosphoproteomic data analysis is not a trivial matter, due to the experimental complexity which can create various computational challenges when han dling issues such as missing values and batch effects [9]. Furthermore, only a small fraction of phosphosites have been annotated to kinases and phosphatases (amongst which about half to only a handful of well-studied kinases), whereas the majority of phosphosites remain unannotated [10]. These challenges significantly limit the interpretability of phosphoproteomic data and their utility in the functional characterisation of signalling networks.

To this end, a diverse array of computational methods, bioinformatics tools, and databases and resources has been developed for processing and functional analysis of phosphoproteomic data. In this work, we first review the computational strategies and tools for processing phosphoproteomic data, ranging from filtering and imputation to normalisation methods (Figure 2A). Next, we summarise the state-of-the-art computational approaches for the functional analysis of phosphoproteome, including kinase-substrate prediction, kinase activity inference and signalling network reconstruction (Figure 2B). Then, we review systems applications of phosphoproteomic data including their characterisation through annotation and ontology databases

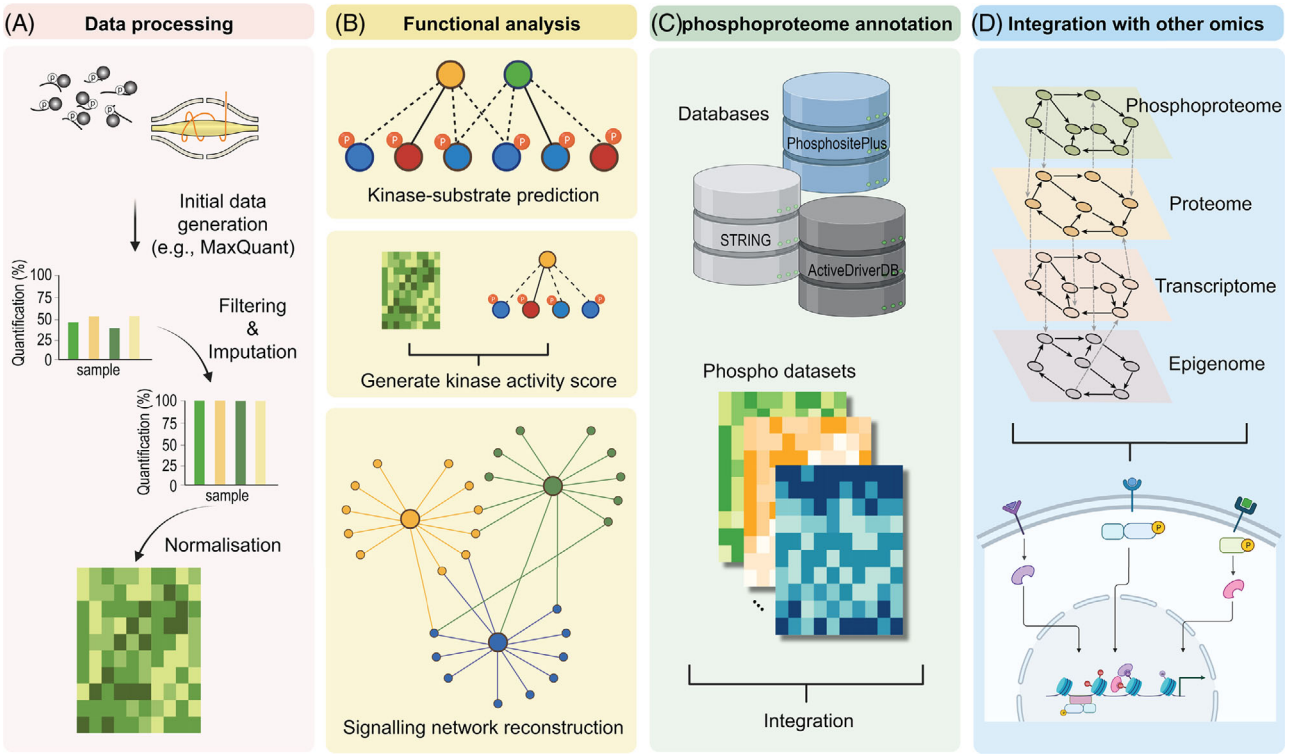

FIGURE 2 A Schematic overview of phosphoproteome analysis pipelines. (A) Phosphoproteomic data processing. (B) Functional analysis of phosphoproteomics data. (C) Characterising the phosphoproteome using databases and other rescources. (D) Integrating phosphoproteomics with other omic datatypes for characterising trans-regulatory networks

(Figure 2C). This is followed by a summary of their integration with other omics data types to understand phospho-signalling and their downstream regulation on trans-regulatory net- works that cut across cellular signalling, transcriptional, translational, and (epi)genetic programmes (Figure 2D). Finally, we discuss the challenges and future directions in this field.

2 | PHOSPHOPROTEOMIC DATA PROCESSING

Mass spectrometry (MS)-based phosphoproteomics has emerged as the dominant tool for identifying and quantifying proteome-wide phosphorylation sites on a global scale and it has become a standard approach to investigate phospho- signalling networks in cellular and biological systems [7]. Typically, in MS-based phosphoproteomics, proteins are first digested into constituent peptide fragments. Then, the phosphopeptides are enriched before measurement by liquid chromatography and tandem mass spectrometry (LC-MS/MS) [61]. The innovations in MS technologies have led to the dramatic increase in the coverage of the phosphoproteome. Such that, the number of identified and quantified phosphorylation sites has increased from a few thousand [62] to routinely over tens of thousands [61, 63] in the last ten years. As increasingly large scale phosphoproteomic datasets are being generated, computational method development has become an essential task for phosphoproteomic data analysis (Table 1). In this section, we will focus on reviewing key computational aspects in phosphoproteomics data processing. This includes steps such as filtering, imputation, normalisation and batch correction, typically performed prior to functional analysis or data integration (Figure 2A).

2.1 | Data filtering

While various quantification strategies have been developed in phosphoproteomics, there remains a large amount of missing values in phosphoproteomic data, where many phosphorylation sites are identified but without quantification. First, the techniques used for quantitative phosphoproteomics (e.g., label-free quantification [LFQ] [64], stable isotope labelling by amino acids in cell culture (SILAC) [65], isobaric tandem mass tags [iTRAQ, TMT] [66, 67]) result in different amount and variability of missing values [68]. Next, the types of data acquisition methods (e.g., data-dependent acquisition [DDA], data independent acquisition [DIA] [69]) also have substantial impact on the missingness in phos phoprotemics data. Other factors contributing to missing values include but are not limited to biological factors such as low protein abundances of transcription factors (TFs), and analytical factors including the mis-cleavage of peptides during digestion, inaccurate peptide-spectrum matches against the protein database, and poor ionisation efficiency and sample loss during preparation [70]. Moreover, the number of missing values present in the same phosphoproteomic dataset varies across different data-processing platforms such as Pro-

TABLE 1 Categorisation of representative tools for phosphoproteomics data analysis

Category	Specific application	Method/Study
Data processing	Imputation	DAPAR [11], DreamAI [12], NAguideR [13]
	Normalisation	Global-centering methods [14], Phosphonormalizer [15], Pairwise normalisation [16]
Functional analysis	Kinase substrate prediction	NetPhosK [17], NetPhorest [18], DeepPhos [19], GPS [20], Musite [21], PrediKin [22], KinomeXplorer [23], PhosphoPICK [24], PhosphoPredict [25], KSP-PUEL [26], CoPhosK [27], SnapKin [28]
	Kinase activity inference	KEA [29], KinasePA [30], IKAP [31], KSEA [32], INKA [33], KARP [34], RoKAI [35]
	Signalling network reconstruction	PathFinder [36], PHOTON [37], CoPPNet [38], PHONEMeS [39], RegPhos [40], INKA [33], PTMsea [41]
	Time-course kinase activity analysis	CLUE [42]
	Time-course phospho-event ordering	Minardo [43]
Data processing & Funcational analysis	Comprehensive suite	PhosR [9], Perseus [44]
	Utilising mutation information	HotPho [45], MIMP [46]
Characterisation	Utilising evolutionary information	Strumillo et al. [47]
	Utilising structural information	Betts et al. [48]
	Integration of multiple features	SAPH-ire [49], SAPH-ire TFx [50], Beltrao et al. [51], Ochoa et al. [52], Xiao et al. [53]
Integration	ESC differentiation	Yang et al. [54], AdaEnsemble [55]
	HRAS signaling	MiNETi [56]
	Yeast pheromone response; Glioblastoma multiforme	PSC Forest algorithm [57]
	Non-small cell lung cancer	Balbin et al. [58]
	Renal cell carcinoma drug targets	COSMOS [59]
	Prostate cancer drug targets	TieDIE [60]

genesis, MaxQuant and Proteios [71]. The presence of missing values significantly affects the completeness of the data and distorts the biological signal. To remedy the missingness, filtering of the data is often applied to reduce missing values (e.g., removing phosphosites with low quantification rates) while minimising the loss of information. For example, Faca et al. showed that data filtering can considerably reduce the aberrant quantitation and unwanted variation while balancing sensitivity and specificity [72]. Similarly, Kim et al. demonstrated that filtering could significantly improve signal-to-noise ratio, leading to better clustering of sample replicates [9].

Given the differences in biological conditions, techniques used in phosphoproteomic profiling, and the varying coverage and depth, the filtering process needs to be optimised according to the experimental design and quantification technique. For instance, as high multiplexing capabilities reduce the abundance of missing values, a single multiplexed TMT batch often has fewer than 1% of missing values [73], negating the need for complex data filtering procedures. However, the abundance of missing values inflate as multiple TMT batches are integrated [74]. To reduce missing values, a user-defined threshold is often set to remove phosphosites that exceed certain percentage of missing values across TMT runs (e.g., filtering out any phosphosites that were not captured in at least 80% of TMT runs [75]). In comparison, LFQ does not allow for sample multiplexing and SILAC only allows up to three channels for simultaneous measurement [68]. For filtering data quantified by LFQ, Yang et al. [54] retained phosphosites that pass pre-defined percentages of quantification across conditions and within each condition across biological replicates. When handling SILAC data, Valdes et al. [76] first removes phosphosites with opposite directions of regulation within a set of biological replicates, and then the phosphosites quantified in all replicates in at least one condition were kept. For the data acquisition methods, DDA only collects precursors of highest abundance whereas DIA systematically collects MS/MS data from

every mass and all detected precursors [77]. As a result of the semi-stochastic nature of precursor selection procedure, DDA often exhibits inconsistent peptide identification in all samples, whereas DIA has high coverage of identified peptides with less missing values. Regardless of the acquisition method, the typical filtering strategy involves setting a data-specific threshold to retain phosphosites that are quantified in a given percentage across replicates and/or conditions [78, 79].

The ambiguity in localisation of the single phosphorylated amino acid within the identified phosphopeptides [80] is another issue commonly addressed in phosphoproteomics data filtering. A localisation score is often assigned to a given phosphosite by various algorithms [81, 82], indicating the probability of the correct localisation of the phosphosite. For example, using a PTM-score, the phosphosites are normally grouped into four classes [83]. Phosphosites which have high localisation confidence (class I: localisation probability $> = 0.75$) are retained. This can be automated using various tools or software (e.g., PhosR [9], Spectronaut [63], Perseus [44] and ProteoViz [84]). Although various ad-hoc data filtering methods are used to pre-process phosphoproteomic data, the field will benefit from designing filtering strategies tailored towards specific technologies and data acquisition methods, given its significant impact on further downstream analyses.

2.2 | Data imputation

Imputation is another key computational technique often performed after data filtering to handle missing values. In essence, imputation strategies replace the missing values with estimates by some computational procedures. The reduction of missing values by well-designed imputation strategies can significantly improve the outcomes of subsequent analyses, such as identifying differentially phosphorylated sites, kinase activity inference and kinase-substrate prediction. Although a large number of imputation approaches have been developed for proteomics data [70] and single-cell RNA-sequencing (scRNA-seq) data [85], very few are specifically designed for phosphoproteomic data. In many cases, imputation methods designed for proteomic datasets are either tuned or directly applied to phosphoproteomic datasets (e.g., [13], DreamAI [12]). This may not be ideal given the experimental differences in profiling proteome and phosphoproteome (e.g., phosphopeptide enrichment). One of the most popular methods implemented in Perseus software [44] for imputing phosphoproteomic data is based on drawing random values from a heuristic distribution created around the lower detection range of the quantified values in label-free phosphoproteomic data. Another tool that enables phospho-specific imputation is the PhosR package which implements a set of heuristic imputation strategies by taking into account quantification rate of each phosphosite and also experimental designs [9]. While a recent study suggests that the best-performing imputation methods differ for proteomic and phosphoproteomic datasets [13], to date, there is still a lack of comprehensive evaluation on how different imputation methods perform across a broad range of phosphoproteomic datasets. A systematic benchmark of existing imputation methods on their accuracy and reliability using an extensive collection of phosphoproteomic datasets will be of great value to the field.

2.3 | Data normalisation

As with any high-throughput biotechnology, systematic biases may be introduced during phosphoproteomic profiling when the experiments are carried out in batches and span across a long period of time. Normalisation is an essential step to capture and correct for those biases. A flexible yet robust normalisation method would enable greater precision in quantitative comparison at phosphorylation levels. Some commonly used normalisation strategies include global centring which centres the peptide abundances to have the same median intensities [14]. Such global centring- based normalisation methods are based on the assumption that the median abundances across all phosphopeptides remain unchanged across different samples and experiments. However, this assumption may not hold in experimental designs that result in a global shift of phosphorylation profiles. Therefore, the application of such a normalisation method needs to be well considered and justified. Various advanced methods have been designed specifically for phosphoproteomic data normalisation. For example, Kauko et al. [16] proposed to normalise label-free quantitative phosphoproteomics based on adjusting phosphopeptide abundances measured before and after the enrichment, as the perturbations in their study introduce unusual unidirectional changes in the phosphopeptide abundance. They demonstrated that the selection of normalisation methods has significant impact on subsequent analyses such as the inference of pathway activities. Similarly, Phosphonormalizer implements a pairwise normalisation procedure incorporating non-enriched phosphopeptide data as a reference, leading to better normalisation results compared to several other global centring methods [15]. Finally, PhosR utilises a removing unwanted variant (RUV) framework [86] along with a set of stably phosphorylated sites (SPSs) [9] as negative controls, which greatly improves the reproducibility of biological replicates amongst samples. Given the need for phosphoproteomic data-specific processing methods for optimal downstream analysis, we anticipate growing methodological innovation on this aspect.

After data normalisation, differential analysis methods such as simple *t*-test and ANOVA test or those developed for gene expression analysis (e.g., Limma [87]) are frequently adapted for quantifying changes in phosphorylation levels in different treatments and conditions or through a time course such as from profiling a differentiation process. The outcomes from these analyses will then feed into functional analysis which we will summarise in the next section.

3 | FUNCTIONAL ANALYSIS OF PHOSPHOPROTEOMICS DATA

Phosphoproteomes are governed by the coordinated regulation of kinases, phosphatases, and their substrates. Functional analysis of

phosphoproteomic data for characterising underlying signalling networks is essential to understand their dynamics in health and dysfunction in disease. In this section, we highlight the existing bioinformatics tools for functional phosphoproteomic analysis, including kinase-specific substrate prediction, kinase activity inference and signalling network reconstruction (Figure 2B).

3.1 | Kinase-substrate prediction

Kinase-specific substrate identification is a key step towards reconstructing signalling networks. However, only a small percentage of identified phosphosites have been annotated with the kinase(s) it is phosphorylated by, and this continues to be the case due to the increasing number of new phosphosites identified in successive large-scale phos phoproteomic studies. Moreover, most kinase-substrate (K-S) relationships are annotated to well-studied kinases, accounting for a very small fraction of all kinases [88]. These challenges have motivated the development of cost- and time-effective bioinformatics tools to systematically predict K-S relationships prior to experimental verification.

Most of the existing computational approaches rely on the evaluation of the flanking sequences of the phosphory lated residues, in which the majority use mainly the primary amino acid sequences, whereas others use various levels of structural information [89]. Specifically, NetPhosK [90], NetPhorest [18], DeepPhos [19] and GPS [20] identify kinase-specific substrates based on amino acid sequences. Nevertheless, kinases recognise the tthree-dimensional structures, instead of the primary sequences, of the peptides surround the phosphosites. To enhance the prediction accuracy, structure information of the phosphopeptide has been incorporated into kinase-substrate prediction [17, 22, 91]. For example, Musite [21] integrates sequence similarity to known phosphosites with protein disorder scores for prediction. Similarly, PrediKin [92] uses the available crystal structures, molecular modelling, and sequence analyses of kinases and substrates. Several tools were developed to predict K-S relationship by jointly analysing protein-protein interactions (PPIs) with kinase recognition motifs. Examples include KinomeXplorer [23], Phospho- PICK [24] and PhosphoPredict [25]. Nevertheless, the coverage of the computationally predicted K-S relationships is far from saturation and most approaches are biased towards well-studied kinases owing to the availability of their annotations.

The advances in MS-based technologies offer excellent opportunities to map the dynamics of each phosphorylation event, which could serve as a rich resource for kinase-specific substrates prediction in a given context. To this end, methods that utilise dynamic phosphoproteomics profiling data together with static information (e.g., sequence motifs) demonstrate the benefit of such complementary strategy in kinase-specific substrates prediction. These including KSP-PUEL [26], a positive-unlabelled ensemble machine learning approach that incorporates dynamic phosphoproteomics data with static sequence information in its prediction, and CoPhosK [27], which utilises correlation analysis to capture collective dynamic signatures of kinase substrates. To overcome the limited number of known substrates for given kinases, Snapkin, an ensemble deep learning model [28], boosts small training datasets by introducing various learning techniques in an ensemble deep learning neural network. While these kinase-substrate prediction methods are useful for identifying K-S relationships, additional efforts are required to improve the accuracy and robustness of such methods given the high noise to signal ratio in large scale phosphoproteomic data.

3.2 | Kinase activity inference

Besides kinase-substrate prediction, large-scale MS-based phosphoproteomic data also provide opportunities to per- form biological system-specific inferences of kinase activities, a vital step to advance our understanding of the functional roles of kinases in biological processes and diseases.

A common approach for kinase activity inference is to assess the activity of a given kinase based on the phosphorylation dynamics of its known and predicted substrates. Although many bioinformatics tools can serve this purpose (e.g., KEA [29, 93], KinasePA [30], IKAP [31], KSEA App [32], KARP [34], INKA [33]), there has been limited effort to compare these tools or determine the optimal set of parameters to use in different scenarios partly due to the difficulty of establishing ground truth for benchmarking. Furthermore, although most kinase activity inference algorithms rely on some type of kinase-substrate enrichment analysis, which estimates the changes of kinase activities based on the coordinated changes of known substrates [94], the applicability of each method may vary depending on their design. For example, while other methods require comparison across groups of samples from different conditions, INKA can be applied to a single sample.

Another key issue of kinase activity inference methods is their reliance on known kinase-substrate relationships. Although comprehensive resources such as PhosphoSitePlus, Signor [95], Phospho.ELM [96], dbPTM 3.0 [97] have collected and curated known relationships between kinases and phosphosites, KSEA-based approaches are limited by the availability of annotated kinases (Figure 3A) and the known K-S relationships (Figure 3B). The incompleteness of substrate discovery of a given kinase potentially affects the kinase activity estimation and biases discoveries towards kinases that are well-studied (Figure 3B). Computational predictions of K-S relationships have been demonstrated to be valuable to enhance the reliability of kinase activity inference by increasing the converge of K-S pairs [98, 93]. However, the quality and scope of computational prediction are still insufficient [27], and again, most methods can only make reliable predictions for well-studied kinases [99]. In addition to computationally increase the number of K-S associates, Yılmaz et al. [35] developed a framework RoKAI, which incorporates functional information of kinases and their corresponding phosphosites. They hypothesised that the biologically significant changes most likely occur on the functionally related phosphosites; wherein the functional environment of a phosphosite can provide further in- formation of the dynamics of the phosphorylation event. To capture the functional networks of phosphosites,

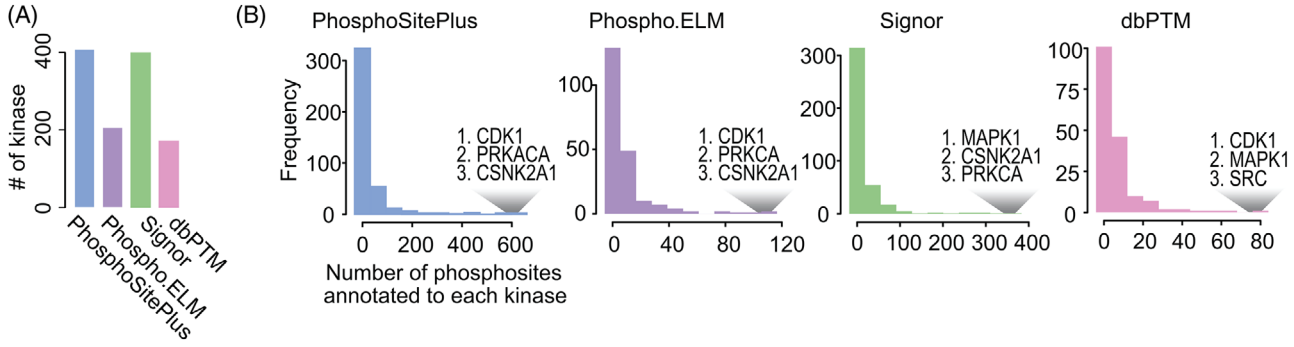

FIGURE 3 The coverage of different phosphosite curation data. (A) The number of annotated kinases in each database. (B) Histograms showing the distribution of the number of annotated phosphosites to kinases in each database. Top three kinases with the most number of annotated substrates are highlighted

RoKAI integrates K-S relationships from PhosphositePlus, coevolution and structural information of phosphosites from PTM- code [100], and PPI from STRING [101]. Comparisons with alternative methods shown that RoKAI outperformed a variety of kinase activity inference approaches. Temporal information of time-course phosphoproteome has also been utilised to infer kinase activity. For example, CLUE [42] firstly partitions phosphosites into optimal clusters according to their temporal profiles and then identifies kinases associated with each cluster. Build on the temporal clusters, Minardo [43] performs further statistical tests on the phospho-signalling events and infers the temporal ordering of these events.

3.3 | Signalling network reconstruction and characterisation

A typical step following kinase-specific substrate identification and kinase activity inference is to reconstruct signalling networks for the systematic characterisation of the interplay amongst activated kinases, their substrates and signalling proteins. Numerous high-throughput studies have revealed the dynamic global architectures of cellular signalling net- works in a wide range of biological contexts (e.g., [102, 103, 88, 104, 54]). These studies characterise phosphorylation- modulated interaction networks based on, not only motif-based predictions but also the dynamics of phosphorylation events, PPIs, genetic interactions, gene expression profiling, and metabolic pathways. Furthermore, many methods have been developed to infer phosphorylation-based networks, including PhosR, RegPhos [40], PathFinder [36], PHO- TON [37], CoPPNet [38], and PHONEMeS [39], INKA [33], and PTMsea [41]. For example, PhosR enables the reconstruction of the signalomes by integrating motif recognition and dynamic profiles of phosphosites across the profiled kinome in the phosphoproteomic data; RegPhos facilitates the reconstruction of intracellular signalling networks by combining the information form experimentally validated K-S associations, PPIs and metabolic pathways; INKA enables reconstruction of signalling network by combining substrate-centric and kinase-centric information and can be applied to a single biological sample rather than comparison across multiple groups; and PTMsea allows the enrichment of signalling pathways in a site-centric manner by taking into account the direction of regulation of phosphosites. It is worth noting that different methods designed for signalling network reconstruction and characterisation as well as those for kinase activity inference may require different input data to perform appropriate analysis. For example, while most methods require only phosphoproteomics data as user input with other resources provided as part of the package (e.g., kinase recognition motifs in PhosR and BioGRID data [105] derived for PHOTON), INKA requires both phosphopeptide and phosphosite to be provided for data analysis. Therefore, users may need to determine the availability of required input when applying each method.

Notably, the constructed networks serve as powerful resources for further biochemical studies. For instance, Saez–Rodriguez and colleagues [106] experimentally validated unexpected signalling events, governing the activation of T cells, identified in the reconstructing networks by their computational models. As another example, INKA was used for identifying targetable kinases as candidates for inhibition in acute myeloid leukaemia cell lines [107] and T cell acute lymphoblastic leukaemia cell lines [6]. The application of kinase activity inference and signalling network reconstruction and characterisation in disease cell lines and clinical settings has revealed that the dysregulation of phospho-signalling plays an important role in disease aetiology.

4 | CHARACTERISING THE PHOSPHOPROTEOME USING DATABASES AND OTHER RESOURCES

Given the increasing availability and improving quality of molecular biological databases, including those dedicated to phosphoproteomics (e.g., PhosphoSitePlus [108] and PHOSIDA [109]), an emerging research direction has been to systematically analyse across large collections of phosphoproteomic datasets and characterise their shared and distinctive features using various databases and resources (Figure 2C). Typically, these studies are not aimed at characterising dataset-specific features, but create general phosphoproteomic knowledge base and resources that can be used in the annotation of individual datasets. This section reviews representative examples within this field of research.

First, genome-wide mutational information has frequently been incorporated to interpret the functional roles of phosphorylation in various contexts, such as cancer. ActiveDriverDB [110] is a database that systematically maps genetic variations and disease mutations to multiple PTMs, and it can be used to characterise phosphosites through the lens of variants and mutations. Similarly, HotPho [45] can systematically identify three-dimensional co-clustering of phosphosites and cancer mutations on known protein structures, revealing potential causal implications of phospho- sites in cancer. MIMP [46] utilises the flanking sequence information around a phosphosite to estimate the disruption of mutations on its phosphorylation. Those tools facilitate the understanding of disease biology by linking kinase networks to disease mutations.

While early studies focused largely on analysing the primary sequence of the phosphosites from a single source/species (e.g., [18, 111]), recent research has extended upon these approaches to characterise the role of the phosphorylation events by comparing the phosphoproteomes within and amongst species, considering the evolutionary and structural features, and interaction with other PTMs. For example, Beltrao et al. [112] showed that functional phosphosites were significantly more constrained across species compared to non-functional ones, highlighting the importance of incorporating evolutionary information when estimating the functional impact of phosphorylation on a given site. Strumillo et al. [47] identified highly conserved regions of phosphorylation by integrating over half a million phospho- sites from 40 eukaryotic species. The identified hotspots were largely mapped to the protein interfaces, suggesting their functional significance, and the regulatory functions of two phosphosites on one of the hotspots were subsequently verified in their experiments. As the phosphorylation events on the protein interfaces often modulate the stability of their interactions [113, 114], the structural information has also been utilised to functionally prioritise phospho-regulatory events. Betts et al. [48] systematically mapped 223,971 phosphosites from five species to PPIs and prioritised about a thousand sites that are potentially involved in enabling or disabling PPIs. A selection of these prioritised sites was subsequently experimentally validated for their impact on protein interactions. Furthermore, the difference in thermal stability between phosphorylated and dephosphorylated proteins can also be predictive of the functional potential of phosphosites and such information could be integrated with other features to characterise phosphorylation events [115, 116].

The integration of multiple features has also been explored to comprehensively characterise phosphoproteome. For example, Dewhurst et al. [49] created a computational framework called SAPH-ire to systematically rank PTMs by integrating phosphorylation dynamics, sequence conservation, structural features and interaction information. The upgraded version, SAPH-ire TFx [50], uses a multi-feature neural network model to prioritise protein PTMs that have highly predicted biological significance. Their methods have demonstrated that incorporating structural and sequence properties to characterise PTMs can improve the confidence in predicting the functional impact of phosphorylation. The functional landscape of the phosphoproteome has also been explored by analysing functionally relevant phosphosites on a genomic scale in diverse biological systems. For example, Beltrao et al. [51] developed a framework to systematically identify functionally significant PTMs (inclusive of phosphorylation), by estimating whether they have other sites altered by PTMs nearby, or whether they are evolutionarily conserved, regulate protein activity, or modulate PPIs. They demonstrated that the protein domain families which harbour conserved PTMs are likely to be regulatory hotspots. Ochoa et al. [52] applied a machine learning approach and identified 59 functionally relevant properties of phosphosites that can be grouped broadly into four categories including MS evidence, phosphosite reg ulation, structural environment and evolutionary conservation. A single functional score was assigned to each of the 119,809 human phosphosites they identified, and the score indicates how likely the phosphosite is functional across different molecular mechanisms, processes and diseases. They further demonstrated that this score was capable of accurately identifying regulatory phosphosites for a wide array of processes and predicting the effect of deleterious mutations. Finally, while most studies focus on the dynamics of phosphorylation, the phosphoproteome that is stable and common across cell tissue types was found to be also functional and its disruptions were linked to cancer development [53]. These works together demonstrate the complexity of the phosphoproteome and the power of integrative analyses in their characterisation. Furthermore, the knowledge generated from these studies across databases and datasets can serve as new resources for future phosphoproteomics data analysis.

5 | INTEGRATING PHOSPHOPROTEOMICS WITH OTHER OMIC DATA TYPES FOR CHARACTERISING TRANS-REGULATORY NETWORKS

Finally, various studies have attempted to link phospho-signalling with downstream transcriptional and translational regulations. These involve using various systematic approaches and integrative methods to reconstruct and characterise 'trans-regulatory networks' that cut across multiple regulatory programmes such as cell signalling, gene transcription, and protein translation (Figure 2D). Here we summarise a few studies towards this direction.

Signalling networks are an integral layer of the trans-regulatory network underlying almost all cellular processes. To comprehensively characterise the roles of various cellular processes in both health and diseases, attempts have been made to integrate phosphoproteomics with other data modalities such as (epi)genomics, transcriptomics, pro teomics, and metabolomics amongst others. One of the commonly used strategies for integrating phosphoproteome is to jointly interpret phosphoproteome with other omics data to link biological processes from signalling to down- stream transcriptional and translational regulations. For example, Yang et al. [54] studied the dynamics of pluripotency in ESCs transitioning from naïve to formative states by integratively analysing phosphoproteome, proteome, transcriptome, and epigenome at matched time points. The K-S networks constructed from this study revealed the key signalling events culminate in the regulation of chromatin landscapes and activation of master TFs. In

their sub- sequent study, the transcriptional and epigenomic regulations underlying cellular signalling were further analysed to unveil the dynamic rewiring of the transcriptional network during pluripotency progression [55]. The publicly avail- able interactome databases, such as PPI database (STRING), protein-small-molecule interaction database (Drugbank), provide a rich resource to reconstruct signalling, transcriptional and translational networks. For example, MiNETi, proposed by Santra et al [56] is an integrative network method that first reconstructs separate protein-protein, K-S, and TF-DNA interaction networks using the protein interactome, phosphoproteomics, and transcriptomics data. Then, these networks are connected using PPI databases, representing the signal transduction from protein complexes to the cell nucleus. MiNETi has been applied to investigate HRAS signalling in the various subcellular compartments of HeLa cells, revealing that variations in HRAS protein interactions led to distinct kinase activation patterns that control gene transcription. This study also demonstrated that HRAS regulates cell migration from the endoplasmic reticulum, and cell survival from the Golgi apparatus. Another method introduced by Tuncbag et al. [57] involves using a Prize–Collecting Steiner (PCS) Forest algorithm to construct functionally meaningful subtrees by integrating phophosproteomic and transcriptomic data. From this, they revealed a novel pathway linking SLT2 to several TFs that maintain cell-wall integrity and controls biosynthesis.

The phosphoproteomics analysis in the multiomic context has also been increasingly applied in cancer studies [117]. In a study by Balbin et al. [58], the authors integrated phosphoproteomics data with transcriptomics and proteomics data to reconstruct signalling networks associated with Ras oncogenes in non-small cell lung cancer (NSCLC). Their integration approach overcame the low-overlap limitation of data integration by defining differently abundant proteins and using the PCS Tree algorithm to construct functional sub-networks. Through the sub-network reconstruc tion, they identified and validated a drug target, LCK, in KRAS-Dep lung cancers. COSMOS, presented by Dugourd et al. [59], is a network contextualisation tool, enabling connection of TF, kinase activity, and metabolite abundance based on causal networks for identifying renal cell carcinoma drug targets. TieDIE, an algorithm designed for linking genomics data with pathway events [118], was used for detecting druggable kinase pathways in prostate cancer pa tients [60]. Taken together, these studies demonstrated that the integration of phosphoproteome with other omic data types enables the reconstruction of trans-regulatory networks that cutting across multiple regulatory programmes (e.g., gene transcription) and broadens our understanding of cross-talk amongst different cell types and across different omic layers in diverse cellular processes and diseases.

6 | CONCLUSION AND FUTURE OUTLOOK

Over the past decade, significant advancement has been made in phosphoproteomics in terms of scale and precision, which greatly facilitated the studies of phosphorylation events, leading to numerous discoveries in cell signalling dynamics and their regulation on downstream biological processes. Given the significant amount of data generated from increasingly large-scale phosphoproteomics experiments, computational systems methods have become the key drivers in translating high-throughput data into biological knowledge. However, many challenges remain in computational analysis and annotation of phosphoproteomics data and integrating them with other data types and resources in a systematic way. In this work, we reviewed some of the most representative methods and studies across diverse applications ranging from data processing, functional analysis, phosphoproteome annotation, to integration of phosphoproteomics data with other omics data types. Here, we summarise a few potential directions that future work would make a significant contribution to this fast growing field.

The advances of next-generation sequencing and mass spectrometry technologies have made the simultaneous collection of information from multiple molecular layers increasingly more applicable in various biological systems. As we have reviewed in this work, many studies have been carried out to investigate the interactions between different molecular levels by integrating multi-omic data that profile multiple molecular programmes (e.g., signalling, transcription, and translation). However, the development of computational methods that are capable of jointly analysing multi-omic data is still at its infancy. We anticipate the future development of advanced tools given the importance of characterising trans-regulatory networks for comprehensive understanding of cellular processes.

Owing to its dynamic and cell type-specific nature, a phosphorylation event cannot be precisely traced in a heterogeneous cell population, highlighting the importance of analysing phosphorylation events at the single-cell level for complex samples such as tissues and organs. Single-cell proteomics has now entered the centre stage [119] and adaptation of phosphoproteomics profiling to single-cell seems to be on the horizon [120]. The maturation of these technologies will revolutionise the field, uncovering the heterogeneity in signalling networks, complementing single- cell genomics and transcriptomics. Computational methods designed for single-cell genomics/transcriptomics data and those developed for integrating multi-omics will form the basis for future development of methods that can re- construct trans-regulatory networks for heterogeneous cells in single-cell multi-omics data.

Finally, the growing bioinformatics toolbox and databases for phosphoproteomics data processing, downstream analysis and integration with other omics data types has made it critical to benchmark these tools and resources for their utility, accuracy, flexibility, robustness, and reproducibility. Very few benchmark studies have been performed so far [121, 13], and the systematic comparison of computational methods using large collections of datasets and evaluation measurements is required. We anticipate the number of such benchmark studies to grow given the increasing reliance on computational methods for data interpretation, drawing conclusions, and guiding further biological analysis.

ACKNOWLEDGMENTS

This work has been supported by a National Health and Medical Research Council Investigator Grant (1173469) to PY, a postgraduate scholarship from Children's medical research institute to DX and CC, and a Research Training Program (RTP) scholarship from Australian government to CC.

AUTHOR CONTRIBUTIONS

Di Xiao, Carissa Chen, and Pengyi Yang wrote and reviewed the manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Pengyi Yang https://orcid.org/0000-0003-1098-3138

REFERENCES

1. Ubersax, J. A., & Ferrell Jr, J. E. (2007). Mechanisms of specificity in protein phosphorylation. *Nature Reviews Molecular Cell Biology*, *8*, 530–541.
2. Hunter, T. (1995). Protein kinases and phosphatases: The yin and yang of protein phosphorylation and signaling. *Cell*, *80*, 225–236.
3. Cohen, P. (2002). The origins of protein phosphorylation. *Nature Cell Biology*, *4*, E127–E130.
4. Su, Z., Burchfield, J. G., Yang, P., Humphrey, S. J., Yang, G., Francis, D., Yasmin, S., Shin, S. Y., Norris, D. M., Kearney, A. L., Astore, M. A., Scavuzzo, J., Fisher-Wellman, K. H., Wang, Q. P., Parker, B. L., Neely, G. G., Vafaee, F., Chiu, J., Yeo, R., & Hogg, P. J. (2019). Global redox proteome and phosphoproteome analysis reveals redox switch in Akt. *Nature Communication*, *10*, 1–18.
5. Rush, J., Moritz, A., Lee, K. A., Guo, A., Goss, V. L., Spek, E. J., Zhang, H., Zha, X. M., Polakiewicz, R. D., & Comb, M. J. (2005). Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nature Biotechnology*, *23*, 94–101.
6. Cordo, V., Meijer, M. T., Hagelaar, R., de Goeij-de Haas, R. R., Poort, V. M., Henneman, A. A., Piersma, S. R., Pham, T. V., Oshima, K., Ferrando, A. A., Zaman, G. J. R., Jimenez, C. R., & Meijerink, J. P. P. (2022). Phosphoproteomic profiling of T cell acute lymphoblastic leukemia reveals targetable kinases and combination treatment strategies. *Nature Communication*, *13*, 1–13.
7. Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, *537*, 347–355.
8. Hornbeck, P. V., Kornhauser, J. M., Latham, V., Murray, B., Nandhikonda, V., Nord, A., Skrzypek, E., Wheeler, T., Zhang, B., & Gnad, F. (2019). 15 years of PhosphoSitePlus®: Integrating post-translationally modified sites, disease variants and isoforms. *Nucleic acids research*, *47*, D433–D441.
9. Kim, H. J., Kim, T., Hoffman, N. J., Xiao, D., James, D. E., Humphrey, S. J., & Yang, P. (2021). PhosR enables processing and functional analysis of phosphoproteomic data. *Cell reports*, *34*, 108771.
10. Needham, E. J., Parker, B. L., Burykin, T., James, D. E., & Humphrey, S. J. (2019). Illuminating the dark phosphoproteome. *Science Signaling*, *12*, eaau8645.
11. Wieczorek, S., Combes, F., Lazar, C., Giai Gianetto, Q., Gatto, L., Dorffer, A., Hesse, A. M., Couté, Y., Ferro, M., Bruley, C., & Burger, T. (2017). DAPAR & ProStaR: Software to perform statistical analyses in quantitative discovery proteomics. *Bioinformatics*, *33*, 135–136.
12. Ma, W., Kim, S., Chowdhury, S., Li, Z., Yang, M., Yoo, S., Petralia, F., Jacobsen, J., Li, J. J., Ge, X., Li, K., Yu, T., Calinawan, A. P., Edwards, N., Payne, S. H., Boutros, P. C., Rodriguez, H., Stolovitzky, G., Zhu, J., … Kang, J. (2021). DreamAI: Algorithm for the imputation of proteomics data. *bioRxiv*, https://doi.org/10.1101/2020.07.21.214205
13. Wang, S., Li, W., Hu, L., Cheng, J., Yang, H., & Liu, Y. (2020). NAguideR: Performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. *Nucleic acids research*, *48*, e83.
14. Välikangas, T., Suomi, T., & Elo, L. L. (2018). A systematic evaluation of normalization methods in quantitative label-free proteomics. *Briefings in Bioinformatics*, *19*, 1–11.
15. Saraei, S., Suomi, T., Kauko, O., & Elo, L. L. (2018). Phosphonormalizer: An R package for normalization of MS-based label-free phosphoproteomics. *Bioinformatics*, *34*, 693–694.
16. Kauko, O., Laajala, T. D., Jumppanen, M., Hintsanen, P., Suni, V., Haapaniemi, P., Corthals, G., Aittokallio, T., Westermarck, J., & Imanishi, S. Y. (2015). Label-free quantitative phos- phoproteomics with novel pairwise abundance normalization reveals synergistic RAS and CIP2A signaling. *Science Reports*, *5*, 13099.
17. Hjerrild, M., Stensballe, A., Rasmussen, T. E., Kofoed, C. B., Blom, N., Sicheritz-Ponten, T., Larsen, M. R., Brunak, S., Jensen, O. N., & Gammeltoft, S. (2004). Identification of phos- phorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry. *Journal of Proteome Research*, *3*, 426–433.
18. Miller, M. L., Jensen, L. J., Diella, F., Jørgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S. A., Bordeaux, J., Sicheritz-Ponten, T., Olhovsky, M., Pasculescu, A., Alexander, J., Knapp, S., Blom, N., Bork, P., Li, S., Cesareni, G., Pawson, T., … Turk, B. E. (2008). Linear motif atlas for phosphorylation-dependent signaling. *Science Signaling*, *1*, ra2–ra2.
19. Luo, F., Wang, M., Liu, Y., Zhao, X. M., & Li, A. (2019). DeepPhos: Prediction of protein phosphorylation sites with deep learning. *Bioinformatics*, *35*, 2766–2773.
20. Wang, C., Xu, H., Lin, S., Deng, W., Zhou, J., Zhang, Y., Shi, Y., Peng, D., & Xue, Y. (2020). GPS 5.0: An update on the prediction of kinase-specific phosphorylation sites in proteins. *Genomics, Proteomics & Bioinformatics*, *18*, 72–80.
21. Gao, J., Thelen, J., Dunker, A. K, & Xu, D. (2010). Musite, a tool for global prediction of general and kinase-specific phospho- rylation sites. *Molecular & Cellular Proteomics*, *9*, 2586–2600.
22. Saunders, N. F. W., & Kobe, B. (2008). The Predikin webserver: Improved prediction of protein kinase peptide specificity using structural information. *Nucleic Acids Research*, *36*, W286–W290.
23. Horn, H., Schoof, E. M., Kim, J., Robin, X., Miller, M. L., Diella, F., Palma, A., Cesareni, G., Jensen, L. J., & Linding, R. (2014). KinomeXplorer: An integrated platform for kinome biology studies. *Nature Methods*, *11*, 603–604.
24. Patrick, R., Lê Cao, K. A., Kobe, B., & Bodén, M. (2015). PhosphoPICK: Modelling cellular context to map kinase-substrate phosphorylation events. *Bioinformatics*, *31*, 382–389.
25. Song, J., Wang, H., Wang, J., Leier, A., Marquez-Lago, T., Yang, B., Zhang, Z., Akutsu, T., Webb, G. I., & Daly, R. J. (2017). PhosphoPredict: A bioinformatics tool for pre- diction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Science Reports*, *7*, 1–19.
26. Yang, P., Humphrey, S. J., James, D. E., Yang, Y. H., & Jothi, R. (2016). Positive-unlabeled ensemble learning for kinase substrate prediction from dynamic phosphoproteomics data. *Bioinformatics*, *32*, 252–259.
27. Ayati, M., Wiredja, D., Schlatzer, D., Maxwell, S., Li, M., Koyutürk, M., & Chance, M. R. (2019). CoPhosK: A method for comprehensive kinase substrate annotation using co-phosphorylation analysis. *Plos Computational Biology*, *15*, e1006678.

28. Lin, M., Xiao, D., Geddes, T. A., Burchfield, J. G., Parker, B. L., Humphrey, S. J., Yang, P.. (2021). SnapKin: A snap- shot deep learning ensemble for kinase-substrate prediction from phosphoproteomics data. *bioRxiv*, 10.1101/2021.02.23.432610
29. Lachmann, A., & Ma'ayan, A. (2009). KEA: Kinase enrichment analysis. *Bioinformatics*, *25*, 684–686.
30. Yang, P., Patrick, E., Humphrey, S. J., Ghazanfar, S., James, D. E., Jothi, R., & Yang, J. Y. H. (2016). KinasePA: Phosphoproteomics data annotation using hypothesis driven kinase perturbation analysis. *Proteomics*, *16*, 1868–1871.
31. Mischnik, M., Sacco, F., Cox, J., Schneider, H.-C., Schäfer, M., Hendlich, M., Crowther, D., Mann, M., & Klabunde, T. (2016). IKAP: A heuristic framework for inference of kinase activities from Phosphoproteomics data. *Bioinformatics*, *32*, 424–431.
32. Wiredja, D. D., Ayati, M., Mazhar, S., Sangodkar, J., Maxwell, S., Schlatzer, D., Narla, G., Koyutürk, M., & Chance, M. R. (2017). Phosphoproteomics profiling of nonsmall cell lung cancer cells treated with a novel phosphatase activator. *Proteomics*, *17*, 1700214.
33. Beekhof, R., Alphen, C., Henneman, A. A., Knol, J. C., Pham, T. V., Rolfs, F., Labots, M., Henneberry, E., Le Large, T. Y., Haas, R. R., Piersma, S. R., Vurchio, V., Bertotti, A., Trusolino, L., Verheul, H. M., & Jimenez, C. R. (2019). INKA, an integrative data analysis pipeline for phosphoproteomic inference of active kinases. *Molecular Systems Biology*, *15*, e8250.
34. Wilkes, E. H., Casado, P., Rajeeve, V., & Cutillas, P. R. (2017). Kinase activity ranking using phosphoproteomics data (KARP) quantifies the contribution of protein kinases to the regulation of cell viability. *Molecular & Cellular Proteomics*, *16*, 1694–1704.
35. Yılmaz, S., Ayati, M., Schlatzer, D., Çiçek, A. E., Chance, M. R., & Koyutürk, M. (2021). Robust inference of kinase activity using functional networks. *Nature Communication*, *12*, 1–12.
36. Bebek, G., & Yang, J. (2007). PathFinder: Mining signal transduction pathway segments from protein-protein interaction networks. *Bmc Bioinformatics [Electronic Resource]*, *8*, 1–13.
37. Rudolph, J. D., De Graauw, M., Van De Water, B., Geiger, T., & Sharan, R. (2016). Elucidation of signaling pathways from large- scale phosphoproteomic data using protein interaction networks. *Cell systems*, *3*, 585–593.e3.
38. Ayati, M., Chance, M. R., & Koyutürk, M. (2021). Co-phosphorylation networks reveal subtype-specific signaling modules in breast cancer. *Bioinformatics*, *37*, 221–228.
39. Gjerga, E., Dugourd, A., Tobalina, L., Sousa, A., & Saez-Rodriguez, J. (2021). PHONEMeS: Efficient modeling of signaling networks derived from large-scale mass spectrometry data. *Journal of Proteome Research*, *20*, 2138–2144.
40. Huang, K. Y., Wu, H. Y., Chen, Y. J., Lu, C. T., Su, M. G., Hsieh, Y. C., Tsai, C. M., Lin, K. I., Huang, H. D., Lee, T. Y., & Chen, Y. J. (2014). RegPhos 2.0: An updated resource to explore protein kinase–substrate phosphorylation networks in mammals. *Database, 2014*, bau034.
41. Krug, K., Mertins, P., Zhang, B., Hornbeck, P., Raju, R., Ahmad, R., Szucs, M., Mundt, F., Forestier, D., Jane-Valbuena, J., Keshishian, H., Gillette, M. A., Tamayo, P., Mesirov, J. P., Jaffe, J. D., Carr, S. A., & Mani, D. R. (2019). A curated resource for phosphosite-specific signature analysis*[S]. *Molecular & Cellular Proteomics*, *18*, 576–593.
42. Yang, P., Zheng, X., Jayaswal, V., Hu, G., Yang, J. Y. H., & Jothi, R. (2015). Knowledge-based analysis for detecting key signaling events from time-series phosphoproteomics data. *Plos Computational Biology*, *11*, e1004403.
43. Kaur, S., Peters, T. J., Yang, P., Luu, L. D. W., Vuong, J., Krycer, J. R., & O'donoghue, S. I. (2020). Temporal ordering of omics and multiomic events inferred from time-series data. *NPJ Systems Biology and Applications*, *6*, 1–7.
44. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., & Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote) omics data. *Nature Methods*, *13*, 731–740.
45. Huang, K., Scott, A. D., Zhou, D. C., Wang, L. B., Weerasinghe, A., Elmas, A., Liu, R., Wu, Y., Wendl, M. C., Wyczalkowski, M. A., Baral, J., Sengupta, S., Lai, C. W., Ruggles, K., Payne, S. H., Raphael, B., Fenyo, D., Chen, K., Mills, G., & Ding, L. (2021). Spatially interacting phosphorylation sites and mutations in cancer. *Nature Communication*, *12*, 1–13.
46. Wagih, O., Reimand, J., & Bader, G. D. (2015). MIMP: Predicting the impact of mutations on kinase-substrate phosphorylation. *Nature Methods*, *12*, 531–533.
47. Strumillo, M. J., Oplová, M., Viéitez, C., Ochoa, D., Shahraz, M., Busby, B. P., Sopko, R., Studer, R. A., Perrimon, N., Panse, V. G., & Beltrao, P. (2019). Conserved phosphorylation hotspots in eukaryotic protein domain families. *Nature Communication*, *10*, 1–11.
48. Betts, M. J., Wichmann, O., Utz, M., Andre, T., Petsalaki, E., Minguez, P., Parca, L., Roth, F. P., Gavin, A. C., Bork, P., & Russell, R. B. (2017). Systematic identification of phosphorylation-mediated protein interaction switches. *Plos Computational Biology*, *13*, e1005462.
49. Dewhurst, H. M., Choudhury, S., & Torres, M. P. (2015). Structural analysis of PTM hotspots (SAPH-ire)–a quantitative infor- matics method enabling the discovery of novel regulatory elements in protein families. *Molecular & Cellular Proteomics*, *14*, 2285–2297.
50. English, N. J., & Torres, M. (2020). SAPH-ire TFx: A machine learning recommendation model and webtool for the predic- tion of functional post-translational modifications. *Faseb Journal*, *34*, 1–1.
51. Beltrao, P., Albanèse, V., Kenner, L. R., Swaney, D. L., Burlingame, A., Villén, J., Lim, W. A., Fraser, J. S., Frydman, J., & Krogan, N. J. (2012). Systematic functional prioritization of protein posttranslational modifications. *Cell*, *150*, 413–425.
52. Ochoa, D., Jarnuczak, A. F., Viéitez, C., Gehre, M., Soucheray, M., Mateus, A., Kleefeldt, A. A., Hill, A., Garcia-Alonso, L., Stein, F., Krogan, N. J., Savitski, M. M., Swaney, D. L., Vizcaíno, J. A., Noh, K. M., & Beltrao, P. (2020). The functional landscape of the human phosphoproteome. *Nature Biotechnology*, *38*, 365–373.
53. Xiao, D., Kim, H. J., Pang, I., & Yang, P. (2022). Functional analysis of the stable phosphoproteome reveals cancer vulnerabilities. *Bioinformatics*, *38*, 1956–1963.
54. Yang, P., Humphrey, S. J., Cinghu, S., Pathania, R., Oldfield, A. J., Kumar, D., Perera, D., Yang, J. Y. H., James, D. E., Mann, M., & Jothi, R. (2019). Multi-omic profiling reveals dynamics of the phased progression of pluripotency. *Cell Systems*, *8*, 427–445.e10.
55. Kim, H. J., Osteil, P., Humphrey, S. J., Cinghu, S., Oldfield, A. J., Patrick, E., Wilkie, E. E., Peng, G., Suo, S., Jothi, R., Tam, P. P. L., & Yang, P. (2020). Transcriptional network dynamics during the progression of pluripotency revealed by integrative statistical learning. *Nucleic acids research*, *48*, 1828–1842.
56. Santra, T., Herrero, A., Rodriguez, J., Von Kriegsheim, A., Iglesias-Martinez, L. F., Schwarzl, T., Higgins, D., Aye, T. T., Heck, A. J. R., Calvo, F., Agudo-Ibáñez, L., Crespo, P., Matallanas, D., & Kolch, W. (2019). An integrated global analysis of compartmentalized HRAS signaling. *Cell Reports*, *26*, 3100–3115.e7.
57. Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.-S. C., Chayes, J., Borgs, C., Zecchina, R., & Fraenkel, E. (2013). Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *Journal of Computational Biology*, *20*, 124–136.
58. Balbin, O. A, Prensner, J. R., Sahu, A., Yocum, A., Shankar, S., Malik, R., Fermin, D., Dhanasekaran, S. M., Chandler, B., Thomas, D., Beer, D. G., Cao, X., Nesvizhskii, A. I., & Chinnaiyan, A. M. (2013). Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nature Communication*, *4*, 2617.
59. Dugourd, A., Kuppe, C., Sciacovelli, M., Gjerga, E., Gabor, A., Emdal, K. B., Vieira, V., Bekker-Jensen, D. B., Kranz, J., Bindels, E. M. J., Costa, A. S. H., Sousa, A., Beltrao, P., Rocha, M., Olsen, J. V., Frezza, C., Kramann, R., & Saez-Rodriguez, J. (2021). Causal integration

of multi-omics data with prior knowledge to generate mechanistic hypotheses. *Molecular Systems Biology*, *17*, e9730.
60. Drake, J. M., Paull, E. O., Graham, N. A., Lee, J. K., Smith, B. A., Titz, B., Stoyanova, T., Faltermeier, C. M., Uzunangelov, V., Carlin, D. E., Fleming, D. T., Wong, C. K., Newton, Y., Sudha, S., Vashisht, A. A., Huang, J., Wohlschlegel, J. A., Graeber, T. G., Witte, O. N., & Stuart, J. M. (2016). Phosphoproteome integration reveals patient- specific networks in prostate cancer. *Cell*, *166*, 1041–1054.
61. Humphrey, S. J., Karayel, O., James, D. E., & Mann, M. (2018). High-throughput and high-sensitivity phosphoproteomics with the EasyPhos platform. *Nature Protocols*, *13*, 1897–1916.
62. Grosstessner-Hain, K., Hegemann, B., Novatchkova, M., Rameseder, J., Joughin, B. A., Hudecz, O., Roitinger, E., Pichler, P., Kraut, N., Yaffe, M. B., Peters, J. M., & Mechtler, K. (2011). Quantitative phospho- proteomics to investigate the polo-like kinase 1-dependent phospho- proteome. *Molecular & Cellular Proteomics*, *10*, M111.008540.
63. Bekker-Jensen, D. B., Bernhardt, O. M., Hogrebe, A., Martinez-Val, A., Verbeke, L., Gandhi, T., Kelstrup, C. D., Reiter, L., & Olsen, J. V. (2020). Rapid and site- specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nature Communication*, *11*, 787.
64. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., & Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & Cellular Proteomics*, *13*, 2513–2526.
65. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., & Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular pro- teomics*, *1*, 376–386.
66. Wiese, S., Reidegeld, K. A., Meyer, H. E., & Warscheid, B. (2007). Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics*, *7*, 340–350.
67. Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., & Hamon, C. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, *75*, 1895–1904.
68. Hogrebe, A., Von Stechow, L., Bekker-Jensen, D. B., Weinert, B. T., Kelstrup, C. D., & Olsen, J. V. (2018). Benchmarking common quantification strategies for large-scale phosphoproteomics. *Nature Communication*, *9*, 1–13.
69. Riley, N. M., & Coon, J. J. (2016). Phosphoproteomics in the age of rapid and deep proteome profiling. *Analytical Chemistry*, *88*, 74–94.
70. Jin, L., Bi, Y., Hu, C., Qu, J., Shen, S., Wang, X., & Tian, Y. (2021). A comparative study of evaluating missing value imputation methods in label-free proteomics. *Science Reports*, *11*, 1760.
71. Välikangas, T., Suomi, T., & Elo, L. L. (2018). A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. *Brief Bioinform*, *19*, 1344–1355.
72. Faca, V. M., Sanford, E. J., Tieu, J., Comstock, W., Gupta, S., Marshall, S., Yu, H., & Smolka, M. B. (2020). Maximized quantitative phosphoproteomics allows high confidence dissection of the DNA damage signaling network. *Science Reports*, *10*, 18056.
73. O'connell, J. D., Paulo, J. A., O'brien, J., & Gygi, S. P. (2018). Proteome-wide evaluation of two common protein quantification methods. *Journal of Proteome Research*, *17*, 1934–1942.
74. Brenes, A., Hukelmann, J., Bensaddek, D., & Lamond, A. I. (2019). Multibatch TMT reveals false positives, batch effects and missing values. *Molecular & Cellular Proteomics*, *18*, 1967–1980.
75. Creixell, M., & Meyer, A. S. (2022). Dual data and motif clustering improves the modeling and interpretation of phospho- proteomic data. *Cell Reports Methods*, *2*, 100167.
76. Valdés, A., Zhao, H., Pettersson, U., & Lind, S. B. (2020). Phosphorylation time-course study of the response during aden- ovirus type 2 infection. *Proteomics*, *20*, 1900327.
77. Hu, A., Noble, W. S., & Wolf-Yadlin, A. (2016). Technical advances in proteomics: New developments in data-independent acquisition. *F1000Research*, 5 419.
78. Lou, R., Liu, W., Li, R., Li, S., He, X., & Shui, W. (2021). DeepPhospho accelerates DIA phosphoproteome profiling through in silico library generation. *Nature Communication*, *12*, 1–15.
79. Zittlau, K. I., Lechado-Terradas, A., Nalpas, N., Geisler, S., Kahle, P. J., & Macek, B. (2022). Temporal analysis of protein ubiquity- lation and phosphorylation during parkin-dependent mitophagy. *Molecular & Cellular Proteomics*, 21, 100191.
80. Chalkley, R. J., & Clauser, K. R. (2012). Modification site localization scoring: Strategies and performance. *Molecular & Cellular Proteomics*, *11*, 3–14.
81. Locard-Paulet, M., Bouyssié, D., Froment, C., Burlet-Schiltz, O., & Jensen, L. J. (2020). Comparing 22 popular phosphopro- teomics pipelines for peptide identification and site localization. *Journal of Proteome Research*, *19*, 1338–1345.
82. Potel, C. M., Lemeer, S., & Heck, A J. R. (2018). Phosphopeptide fragmentation and site localization by mass spectrometry: An update. *Analytical Chemistry*, *91*, 126–141.
83. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., & Mann, M. (2006). Global, in vivo, and site-specific phospho- rylation dynamics in signaling networks. *Cell*, *127*, 635–648.
84. Storey, A. J., Naceanceno, K. S., Lan, R. S., Washam, C. L., Orr, L. M., Mackintosh, S. G., Tackett, A. J., Edmondson, R. D., Wang, Z., Li, H.-Y., Frett, B., Kendrick, S., & Byrum, S. D. (2020). ProteoViz: A tool for the analysis and interactive visualization of phosphoproteomics data. *Molecular omics*, *16*, 316–326.
85. Hou, W., Ji, Z., Ji, H., & Hicks, S. C. (2020). A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biology*, *21*, 218.
86. Molania, R., Gagnon-Bartsch, J. A., Dobrovic, A., & Speed, T. P. (2019). A new normalization for Nanostring nCounter gene expression data. *Nucleic Acids Research*, *47*, 6073–6083.
87. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*, e47–e47.
88. Buljan, M., Ciuffa, R., Van Drogen, A., Vichalkovski, A., Mehnert, M., Rosenberger, G., Lee, S., Varjosalo, M., Pernas, L. E., Spegg, V., Snijder, B., Aebersold, R., & Gstaiger, M. (2020). Kinase interaction network expands functional and disease roles of human kinases. *Molecular Cell*, *79*, 504–520.e509.
89. De Oliveira, P S L., Ferraz, F A N., Pena, D. A., Pramio, D. T., Morais, F. A., & Schechtman, D. (2016). Revisiting protein kinase–substrate interactions: Toward therapeutic development. *Sci Signal*, *9*, re3–re3.
90. Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., & Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, *4*, 1633–1649.
91. Su, M. G., & Lee, T.-Y. (2013). Incorporating substrate sequence motifs and spatial amino acid composition to identify kinase- specific phosphorylation sites on protein three-dimensional structures. *Bmc Bioinformatics [Electronic Resource]*, *14*(16), S2. Suppl.
92. Brinkworth, R. I., Breinl, R. A., & Kobe, B. (2003). Structural basis and prediction of substrate specificity in protein ser- ine/threonine kinases. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 74–79.
93. Kuleshov, M. V., Xie, Z., London, A. B. K., Yang, J., Evangelista, J. E., Lachmann, A., Shu, I., Torre, D., & Ma'ayan, A. (2021). KEA3: Improved kinase enrichment analysis via data integration. *Nucleic Acids Research*, *49*, W304–W316.
94. Casado, P., Rodriguez-Prados, J. C., Cosulich, S. C., Guichard, S., Vanhaesebroeck, B., Joel, S., & Cutillas, P. R. (2013). Kinase-substrate

enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci Signal*, *6*, rs6.
95. Perfetto, L., Briganti, L., Calderone, A., Cerquone Perpetuini, A., Iannuccelli, M., Langone, F., Licata, L., Marinkovic, M., Mattioni, A., Pavlidou, T., Peluso, D., Petrilli, L. L., Pirrò, S., Posca, D., Santonico, E., Silvestri, A., Spada, F., Castagnoli, L., & Cesareni, G. (2015). SIGNOR: A database of causal relationships between biological entities. *Nucleic Acids Research*, *44*, D548–D554.
96. Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., & Diella, F. (2011). Phospho.ELM: A database of phosphorylation sites–update 2011. *Nucleic Acids Research*, *39*, D261–D267.
97. Lu, C. T., Huang, K. Y., Su, M. G., Lee, T.-Y., Bretaña, N. A., Chang, W. C., Chen, Y.-J., Chen, Y.-J., & Huang, H.-D. (2013). DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Research*, *41*, D295–D305.
98. Qi, L., Liu, Z., Wang, J., Cui, Y., Guo, Y., Zhou, T., Zhou, Z., Guo, X., Xue, Y., & Sha, J. (2014). Systematic analysis of the phosphoproteome and kinase-substrate networks in the mouse testis. *Molecular & Cellular Proteomics*, *13*, 3626–3638.
99. Deznabi, I., Arabaci, B., Koyutürk, M., & Tastan, O. (2020). DeepKinZero: zero-shot learning for predicting kinase-phosphosite associations involving understudied kinases. *Bioinformatics*, *36*, 3652–3661.
100. Minguez, P., Letunic, I., Parca, L., & Bork, P. (2013). PTMcode: A database of known and predicted functional associations between post-translational modifications in proteins. *Nucleic Acids Research*, *41*, D306–D311.
101. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., & Von Mering, C. (2014). STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, *43*, D447–D452.
102. Linding, R., Jensen, L. J., Ostheimer, G. J., Van Vugt, M. A. T. M., Jørgensen, C., Miron, I. M., Diella, F., Colwill, K., Taylor, L., Elder, K., Metalnikov, P., Nguyen, V., Pasculescu, A., Jin, J., Park, J. G., Samson, L. D., Woodgett, J. R., Russell, R. B., Bork, P., … Yaffe, M. B. (2007). Systematic discovery of in vivo phosphorylation networks. *Cell*, *129*, 1415–1426.
103. Breitkreutz, A., Choi, H., Sharom, J. R., Boucher, L., Neduva, V., Larsen, B., Lin, Z. Y., Breitkreutz, B. J., Stark, C., Liu, G., Ahn, J., Dewar-Darch, D., Reguly, T., Tang, X., Almeida, R., Qin, Z. S., Pawson, T., Gingras, A. C., Nesvizhskii, A. I., & Tyers, M. (2010). A global protein kinase and phosphatase interaction network in yeast. *Science*, *328*, 1043–1046.
104. Humphrey, S. J., Yang, G., Yang, P., Fazakerley, D. J., Stöckli, J., Yang, J. Y., & James, D. E. (2013). Dynamic adipocyte phosphoproteome reveals that Akt directly regulates mTORC2. *Cell metabolism*, *17*, 1009–1020.
105. Stark, C. (2006). BioGRID: A general repository for interac- tion datasets. *Nucleic acids research*, *34*, D535–D539.
106. Saez-Rodriguez, J., Simeoni, L., Lindquist, J. A., Hemenway, R., Bommhardt, U., Arndt, B., Haus, U. U., Weismantel, R., Gilles, E. D., Klamt, S., & Schraven, B. (2007). A logical model provides insights into T cell receptor signaling. *Plos Computational Biology*, *3*, e163.
107. van Alphen, C., Cloos, J., Beekhof, R., Cucchi, D. G., Piersma, S. R., Knol, J. C., Henneman, A. A., Pham, T. V., van Meerloo, J., Ossenkoppele, G. J., Verheul, H. M. W., Janssen, J. J. W. M., & Jimenez, C. R. (2020). Phosphotyrosine-based phosphopro- teomics for target identification and drug response prediction in AML cell lines. *Molecular & Cellular Proteomics*, *19*, 884–899.
108. Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., & Skrzypek, E. (2015). PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Research*, *43*, D512–D520.
109. Gnad, F., Ren, S., Cox, J., Olsen, J. V., Macek, B., Oroshi, M., & Mann, M. (2007). PHOSIDA (phosphorylation site database): Manage- ment, structural and evolutionary investigation, and prediction of phosphosites. *Genome biology*, *8*, 1–13.
110. Krassowski, M., Paczkowska, M., Cullion, K., Huang, T., Dzneladze, I., Ouellette, B. F. F., Yamada, J. T., Fradet-Turcotte, A., & Reimand, J. (2018). ActiveDriverDB: human disease mutations and genome variation in post-translational modification sites of proteins. *Nucleic Acids Research*, *46*, D901–D910.
111. Eisenhaber, B., & Eisenhaber, F. Prediction of posttranslational modification of proteins from their amino acid sequence. *Data Mining Techniques for the Life Sciences*, *609*, 365–384.
112. Beltrao, P., Trinidad, J. C., Fiedler, D., Roguev, A., Lim, W. A., Shokat, K. M., Burlingame, A. L., & Krogan, N. J. (2009). Evolution of phosphoregulation: Com- parison of phosphorylation patterns across yeast species. *Plos Biology*, *7*, e1000134.
113. Nishi, H., Hashimoto, K., & Panchenko, A. R. (2011). Phosphorylation in protein-protein binding: Effect on stability and function. *Structure (London, England)*, *19*, 1807–1815.
114. Šoštarić, N., O'Reilly, F. J., Giansanti, P., Heck, A. J., Gavin, A. C., & van Noort, V. (2018). Effects of acetylation and phosphorylation on subunit interactions in three large eukaryotic complexes. *Molecular & Cellular Proteomics*, *17*, 2387–2401.
115. Huang, J. X., Lee, G., Cavanaugh, K. E., Chang, J. W., Gardel, M. L., & Moellering, R. E. (2019). High throughput discovery of functional protein modifications by Hotspot Thermal Profiling. *Nature Methods*, *16*, 894–901.
116. Potel, C. M., Kurzawa, N., Becher, I., Typas, A., Mateus, A., & Savitski, M. Impact of phosphorylation on thermal stability of proteins. *Nature Methods*, *18*, 757–759.
117. Mantini, G., Pham, T. V., Piersma, S. R., & Jimenez, C. R. (2021). Computational analysis of phosphoproteomics data in multi- omics cancer studies. *Proteomics*, *21*, 1900312.
118. Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D., & Stuart, J. M. (2013). Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics*, *29*, 2757–2764.
119. Schoof, E. M., Furtwängler, B., Üresin, N., Rapin, N., Savickas, S., Gentil, C., Lechman, E., Keller, U. A. D., Dick, J. E., & Porse, Bo T. (2021). Quantitative single-cell proteomics as a tool to characterize cellular hierarchies. *Nature Communication*, *12*, 1–15.
120. Lun, X. K., & Bodenmiller, B. (2020). Profiling cell signaling networks at single-cell resolution. *Molecular & Cellular Pro- teomics*, *19*, 744–756.
121. Hernandez-Armenta, C., Ochoa, D., Gonçalves, E., Saez-Rodriguez, J., & Beltrao, P. (2017). Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics*, *33*, 1845–1851.

How to cite this article: Xiao, D., Chen, C., & Yang, P. (2022). Computational systems approach towards phosphoproteomics and their downstream regulation. *Proteomics*, e2200068. https://doi.org/10.1002/pmic.202200068