Article

An atlas of substrate specificities for the human serine/threonine kinome

https://doi.org/10.1038/s41586-022-05575-3

Received: 1 May 2022

Accepted: 17 November 2022

Published online: 11 January 2023

Open access

Check for updates

Jared L. Johnson[1,2,28], Tomer M. Yaron[1,2,3,4,5,28], Emily M. Huntsman[1,2], Alexander Kerelsky[1,2,3], Junho Song[1,2], Amit Regev[1,2], Ting-Yu Lin[1,2,6], Katarina Liberatore[1,2], Daniel M. Cizin[1,2], Benjamin M. Cohen[1,2], Neil Vasan[7,8], Yilun Ma[1,2], Konstantin Krismer[9,10], Jaylissa Torres Robles[11,12], Bert van de Kooij[10], Anne E. van Vlimmeren[10], Nicole Andrée-Busch[13], Norbert F. Käufer[13], Maxim V. Dorovkov[14], Alexey G. Ryazanov[14], Yuichiro Takagi[15], Edward R. Kastenhuber[1,2], Marcus D. Goncalves[1,16], Benjamin D. Hopkins[17], Olivier Elemento[3,4], Dylan J. Taatjes[18], Alexandre Maucuer[19], Akio Yamashita[20], Alexei Degterev[21], Mohamed Uduman[22], Jingyi Lu[22], Sean D. Landry[22], Bin Zhang[22], Ian Cossentino[22], Rune Linding[23], John Blenis[1,24,25], Peter V. Hornbeck[22], Benjamin E. Turk[11]✉, Michael B. Yaffe[10,26,27]✉ & Lewis C. Cantley[1,2]✉

Protein phosphorylation is one of the most widespread post-translational modifications in biology[1,2]. With advances in mass-spectrometry-based phosphoproteomics, 90,000 sites of serine and threonine phosphorylation have so far been identified, and several thousand have been associated with human diseases and biological processes[3,4]. For the vast majority of phosphorylation events, it is not yet known which of the more than 300 protein serine/threonine (Ser/Thr) kinases encoded in the human genome are responsible[3]. Here we used synthetic peptide libraries to profile the substrate sequence specificity of 303 Ser/Thr kinases, comprising more than 84% of those predicted to be active in humans. Viewed in its entirety, the substrate specificity of the kinome was substantially more diverse than expected and was driven extensively by negative selectivity. We used our kinome-wide dataset to computationally annotate and identify the kinases capable of phosphorylating every reported phosphorylation site in the human Ser/Thr phosphoproteome. For the small minority of phosphosites for which the putative protein kinases involved have been previously reported, our predictions were in excellent agreement. When this approach was applied to examine the signalling response of tissues and cell lines to hormones, growth factors, targeted inhibitors and environmental or genetic perturbations, it revealed unexpected insights into pathway complexity and compensation. Overall, these studies reveal the intrinsic substrate specificity of the human Ser/Thr kinome, illuminate cellular signalling responses and provide a resource to link phosphorylation events to biological pathways.

Phosphorylation of proteins at serine, threonine, tyrosine and histidine residues controls nearly every aspect of eukaryotic cellular function[1,2,5,6]. Misregulation of protein kinase signalling commonly results in human disease[7]. Deciphering the cellular roles of any protein kinase requires the elucidation of its downstream effector substrates. However, the majority of kinase–substrate relationships that have been published to date involve a relatively small number of well-studied protein kinases, while few, if any, substrates have been identified for the majority of the approximately 300 human protein Ser/Thr kinases within the human kinome[8–10]. This lack of knowledge of kinase–substrate relationships limits the interpretation of large mass spectrometry (MS)-based phosphoproteomic datasets, which to date have collectively reported over 200,000 Ser and Thr phosphorylation sites on human proteins[3,4,11–13]. The specific kinases that are responsible for these phosphorylation events have been reported for less than 4% of these sites[3], substantially limiting the understanding of cellular phosphorylation networks.

Well-studied Ser/Thr kinases are generally known to recognize specific amino acid residues at multiple positions surrounding the site of phosphorylation[14–17]. This short linear motif, which is characteristic of a given protein kinase, ensures fidelity in signalling pathways regulating phosphorylation at a given Ser or Thr residue. Knowledge of kinase-recognition motifs can facilitate the discovery of new substrates, for example, by scanning phosphoproteomics data for matching sequences. However, to date, phosphorylation-site sequence motifs are known for only a subset of the human protein Ser/Thr kinome. We have previously described combinatorial peptide library screening methods that enable the rapid determination of specificity for individual kinases

A list of affiliations appears at the end of the paper.

based on phosphorylation of peptide substrates[18,19]. Here we apply those methods to experimentally determine the optimal substrate specificity for the large majority of the human Ser/Thr kinome, characterize the relationship between kinases on the basis of their motifs, and computationally use these data to identify the protein kinases that are likely to phosphorylate any site identified using MS or other techniques. Finally, we show how this information can be applied to capture complex changes in signalling pathways in cells and tissues after genetic, pharmacological, metabolic and environmental perturbations.

Profiling kinase substrate specificity

Substrate-recognition motifs across the human Ser/Thr kinome were determined by performing a positional scanning peptide array (PSPA) analysis. We used a previously reported combinatorial peptide library that systematically substitutes each of 22 amino acids (20 natural amino acids plus phosphorylated Thr and phosphorylated Tyr) at nine positions surrounding a central phospho-acceptor position containing equivalent amounts of Ser and Thr[19] (Fig. 1a). Using purified recombinant kinase preparations, we successfully obtained phosphorylation-site motifs for 303 Ser/Thr kinases, covering every branch of the human Ser/Thr kinase family tree as well as a collection of atypical protein kinases (Fig. 1b, Supplementary Fig. 1 and Supplementary Tables 1 and 2). Profiling of the large majority of these kinases, including 83 understudied 'dark' kinases, was lacking[8].

Position-specific scoring matrices (PSSMs) derived from quantified PSPA data were analysed by hierarchical clustering to compare kinase substrate motifs across the kinome (Fig. 2 and Supplementary Table 2). As expected, kinases sharing substantial sequence identity displayed a high degree of similarity in their optimal substrate motifs. However, we found many cases in which clustering by PSSM did not strictly recapitulate the evolutionary phylogenetic relationships between kinases inferred from their primary sequences (Fig. 2). Instead, members of most of the major kinase groups were distributed throughout the dendrogram, reflecting numerous examples in which kinases with low overall sequence identity have converged to phosphorylate similar optimal sequence motifs. For example, we found that a number of distantly related kinases (in the YANK, CK1 and 2, GRK and TGFβ receptor families) converged to phosphorylate similar sequence motifs despite their disparate locations on the kinome tree (Fig. 2 (cluster 3)).

Overall inspection of sequence motifs associated with various branches of the motif-based dendrogram revealed that approximately 60% of the Ser/Thr kinome could be represented by simple assignment to one of three previously observed motif classes: selectivity for basic residues N terminal to the phosphorylation site (cluster 1; Fig. 2); directed by a proline residue at the +1 position (cluster 2); or a general preference for negatively charged (acidic and phosphorylated) residues at multiple positions (cluster 3)[15,20,21]. Notably, more than half of all of the reported phosphorylation sites observed by MS could be assigned to one of these three signatures (Extended Data Fig. 1). However, each of these motif classes could be further subcategorized on the basis of selectivity both for and against distinct sets of residues at other positions, reflecting considerable diversity within these clusters (Extended Data Figs. 2–4). The remaining approximately 40% of the Ser/Thr kinome comprised many smaller groups that displayed unique sequence determinants (Fig. 2; clusters 4–17). For example, motifs for the PIKK family kinases (ATM, ATR, DNA-PK and SMG1) clustered into a group that primarily selected a Gln residue at the +1 position (cluster 5), consistent with previous studies[22,23]. Notably, several clusters displayed shared consensus motifs that have not been well recognized previously, such as the group including the IRAK, IRE, WNK, SNRK and RIP kinases (cluster 13), of which the substrate motifs contained basic residues both N and C terminal to the phosphorylation site with dominant selection for aromatic residues at the +3 position. As another example, the kinases LKB1, CAMKK, PINK1 and PBK (cluster 14) primarily recognized hydrophobic residues N terminal to the phosphorylation site in combination with selection for turn-promoting residues (Gly or Asn) at the +1 position. Structural modelling of kinase–peptide complexes revealed complementary features within the kinase catalytic clefts that are probably responsible for the recognition of these motifs (Extended Data Fig. 5a,b).

An important and less generally recognized feature that dominated the clustering was strong negative selection against either positively or negatively charged residues at distinct positions within a motif, suggesting that electrostatic filtering strongly influences kinase substrate selection throughout the kinome[24]. We identified additional classes of amino acids, such as hydrophobic residues, that are selected against by a variety of kinases. These trends suggest that substrate avoidance has a fundamental role in dictating correct kinase–substrate interactions[25,26].

Unexpectedly, we observed that many kinases (129 out of 303) selected either a phosphorylated Thr or a phosphorylated Tyr as the preferred amino acid in at least one position within the motif (Supplementary Fig. 1; where selectivity for phosphorylated Ser was assumed to be equivalent to phosphorylated Thr). In addition to kinases of which the dependence on phospho-priming was previously known (GSK3, CK1 and CK2 families; cluster 3), this phenomenon was particularly evident for the GRK- and YANK-family kinases (Extended Data Fig. 4), both of which have complementary basic residues within their catalytic domains (Extended Data Fig. 5c,d). Notably, individual GRK-family members showed unique and specific selection for the location of the phosphorylated Thr or phosphorylated Tyr residue within their substrate peptides. GRKs are best known for their role in the desensitization of G-protein-coupled receptors (GPCRs), whereby multisite phosphorylation induces the binding of arrestin proteins to inhibit signalling[27,28]. Our findings suggest that the capacity for only seven GRKs to differentially regulate 800 distinct GPCRs probably involves a complex interplay between initial sequence-specific phospho-priming of GPCRs by other Ser/Thr and Tyr kinases, followed by a second level of specificity resulting from GRK-dependent phosphorylation and subsequent recognition by a small number of β-arrestins.

Features of substrate-recognition motifs across the entire kinome could be structurally rationalized on the basis of the presence of specificity-determining residues at particular positions within the kinase catalytic domain[29–32], leading to both expected and unexpected findings. For example, we found that half of the kinases display some degree of selectivity for either a Ser or a Thr as the phospho-acceptor residue (Extended Data Figs. 6 and 7). Consistent with our previously published observations[33], Ser or Thr phospho-acceptor site preference strongly correlated with the identity of the 'DFG+1' residue (that is, the residue immediately after the canonical Asp-Phe-Gly motif within the kinase activation loop), with bulky residues (Phe, Trp, Tyr) at this position in Ser-selective protein kinases and β-branched residues (Val, Ile, Thr) at this position in Thr-selective kinases. However, for some DFG+1 residues, Ser versus Thr selectivity was unexpectedly context dependent. For example, a Leu residue at the DFG+1 position was observed in both Ser-selective and dual-specificity kinases, whereas a DFG+1 Ala residue resulted in a preference for Thr phosphorylation in the context of some kinases (for example, the mitogen-activated protein kinase kinase kinases (MAP3Ks)), but a preference for Ser specificity in others (the IκB kinases). These observations, notable only within the context of the complete Ser/Thr kinome, indicate that additional residues beyond the previously established DFG+1 position can influence Ser/Thr specificity in a context-dependent manner.

Annotation of the human phosphoproteome

Comprehensive knowledge of human Ser/Thr kinase specificity has the potential to 'deorphanize' the large number of reported phosphorylation sites with no associated kinase. To do so, we generated

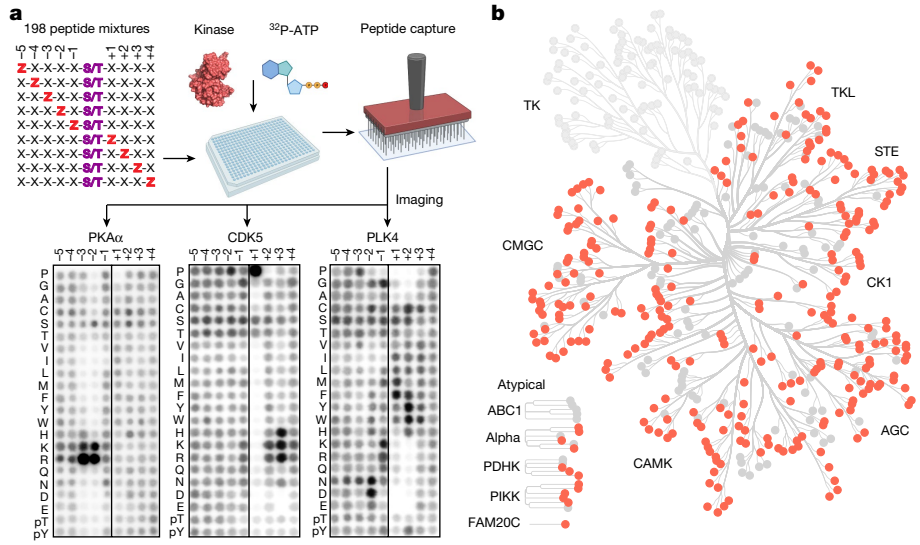


Fig. 1 | Profiling the substrate specificity of the human serine/threonine kinome. a, Experimental workflow for the PSPA analysis and representative results. The schematic was created using BioRender. Z denotes fixed positions containing one of the 20 natural amino acids, or either phosphorylated Thr (pThr) or phosphorylated Tyr (pTyr). X denotes unfixed positions containing randomized mixtures of all natural amino acids except Ser, Thr and Cys. Darker spots indicate preferred residues. **b**, Dendrogram of the human protein kinome, highlighting the Ser/Thr kinases analysed in this study.

a kinome-wide annotation of the human Ser/Thr phosphoproteome, comprising a previously curated set of 82,735 sites that were detected using high-throughput MS[4] plus an additional 7,017 sites that were identified using only low-throughput methods[3]. Using approaches adapted from previously published research, we computationally ranked these 89,752 sites against each Ser/Thr kinase motif[34,35] (Fig. 3a and Supplementary Table 3). Notably, approximately 99% of these phosphorylation sites ranked favourably for at least one kinase that we profiled (that is, the site scored in the top 10% of all sites in the human phosphoproteome for that kinase). These annotations were strongly concordant with sites for which protein kinases involved have been previously identified. For phosphorylation sites of which the upstream kinase has been previously verified by at least three independent reports, encompassing 969 sites and over one third of the kinome, our motif-based approach yielded a median percentile of 95% (that is, the reported site received a higher score than 95% of all putative phosphorylation sites in the phosphoproteome for its established kinase) (Extended Data Fig. 8a). Furthermore, when we back-mapped the motifs of all 303 profiled kinases onto the literature-reported phosphorylation sites, our approach yielded a median reported kinase percentile of 92% (that is, the reported kinase scored more favourably than 92% of all profiled kinases in our atlas for its established substrate) (Extended Data Fig. 9a). These rankings further improved when we considered kinase–substrate pairs with higher numbers of previous reports (Extended Data Figs. 8 and 9), suggesting that, in a large majority of cases, the linear sequence context of phosphorylation sites contributes substantially to kinase–substrate relationships.

Notably, motif predictions alone successfully identified numerous prominently studied kinase–substrate relationships. For example, phosphorylase kinases PHKG1 and PHKG2 emerged as the top two hits (out of 303 kinases) for phosphorylating Ser15 of glycogen phosphorylase (Fig. 3b). This phosphoregulatory event, the very first to be discovered[36], opened up the entire field of phosphorylation-dependent signal transduction. The most highly cited kinase–substrate interaction reported to date is the phosphorylation of the tumour suppressor p53 at Ser15 by the DNA-damage-activated kinase ATM, which scored among the top-ranking kinases associated with that site (Fig. 3c). Notably, other kinases reported to phosphorylate the same site—ATR, SMG1 and DNAPK—scored within the top four predicted kinases[3].

Our approach could also correctly identify kinases for phosphorylation events driven by substrate co-localization or non-catalytic docking interactions, for which we expected less dependence on the phosphorylation-site motifs of their kinases. For example, we correctly identified both the mitochondrial-localized phosphorylation of pyruvate dehydrogenase by the pyruvate dehydrogenase kinases (Extended Data Fig. 10a) and the docking-driven phosphorylation of the MAP kinase ERK by MEK[37] (Extended Data Fig. 10b). Notably, the phosphorylation site on ERK was selected against by nearly every human protein kinase that we profiled except for MEK, explaining how ERK can be exclusively regulated by MEK while avoiding phosphorylation by the kinome at large. Finally, our approach could tease apart kinase subfamilies with similar motifs and correctly assign them to their established substrates. For example, we could distinguish between the CDK family kinases that assume classical roles in cell cycle progression (that is, CDK1, CDK2, CDK3, CDK4 and CDK6) and the subset of CDKs that govern gene transcription (that is, CDK7, CDK8, CDK9, CDK12, CDK13 and CDK19)[38,39] (Extended Data Fig. 11).

Functional annotation of the human phosphoproteome enabled us to examine global trends in kinase–substrate interactions. We found that most phosphorylation sites could be assigned to a small number of putative kinases (that is, BRAF–MEK1, ATM–p53 and CDK4–Rb; Fig. 3d and Supplementary Table 3). However, approximately one-third of all sites lacked unique negative sequence-discriminating features and, instead, matched well to the optimal phosphorylation motifs for a greater number of kinases[21,40,41,42] (that is, Ser119 of CREB, Ser9 of GSK3B and Thr1079 of LATS1; Fig. 3d). This could suggest the importance of other kinase-determining factors (scaffolds, localization and so on) for proper kinase substrate recognition, or may indicate that these specific phosphorylation sites are points of convergence for multiple signalling pathways. For example, cAMP response element binding protein (CREB) is canonically phosphorylated at Ser119 by cAMP-dependent protein kinase (PKA); however, numerous previous reports demonstrate that a broad range of cellular stimuli and drug perturbations impinge on the phosphorylation of this site by no less than ten distinct kinases[3]. Taken together, these findings suggest that the presence of negative-selectivity elements flanking a putative phosphorylation site can be used to insulate a substrate from inappropriate phosphorylation by dozens of related kinases, whereas the absence of such negative

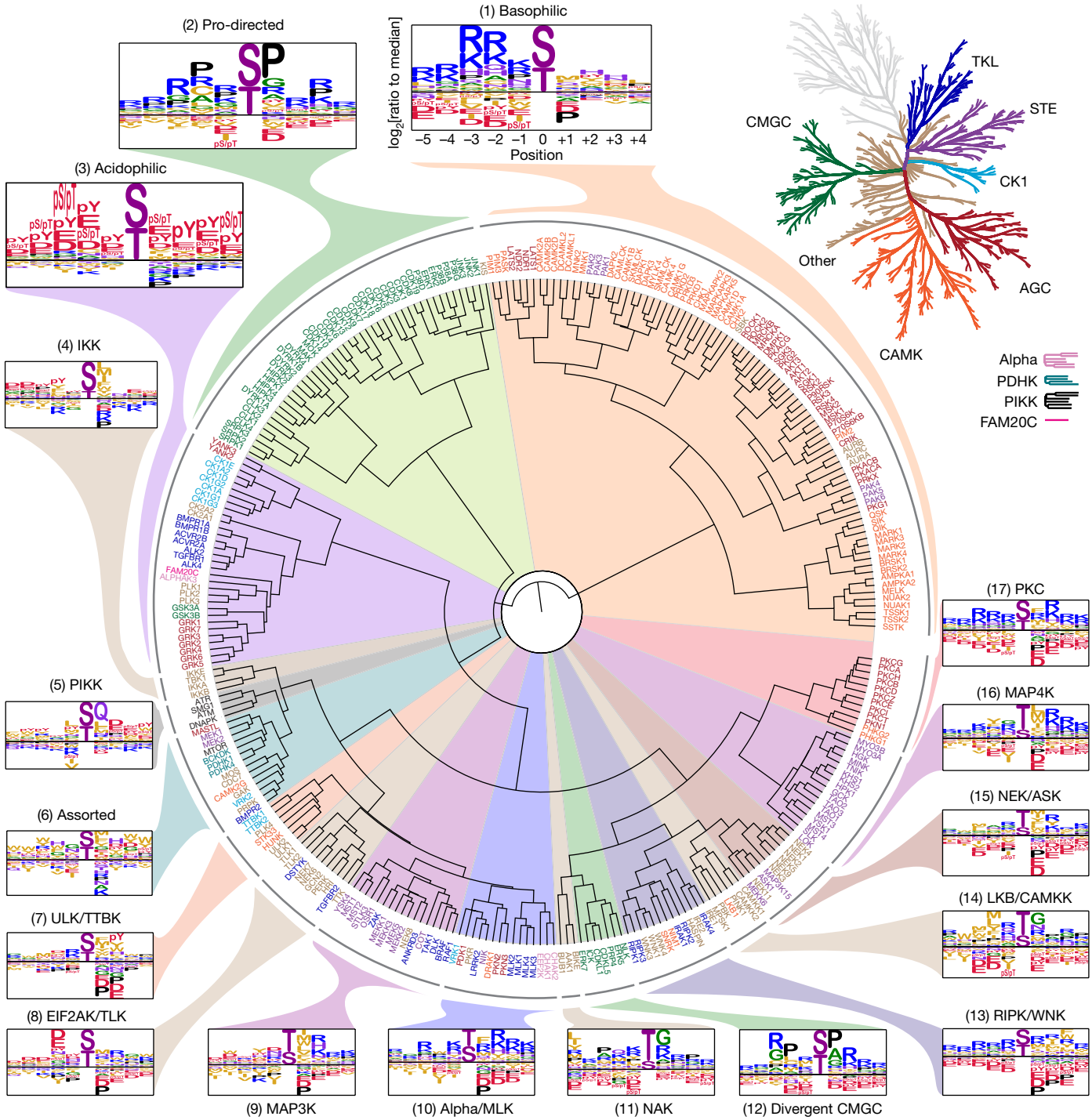


Fig. 2 | Phosphorylation-site motif tree of the human Ser/Thr kinome. Hierarchical clustering of 303 Ser/Thr kinases on the basis of their amino acid motif selectivity (PSSMs). Kinase names are colour labelled according to their phylogenetic relationships (top right)[2].

selectivity can enable protein kinases in distinct pathways to converge on the same target.

Motif-enrichment analysis

Cell signalling networks are complex and dynamic. Perturbation of kinase signalling pathways by genetic manipulations, treatment with growth factors and ligands, environmental stress or small-molecule inhibitors reshapes the phosphoproteome through both direct and indirect effects as a consequence of secondary signalling responses and/or off-target effects from the experimental treatment[43]. Owing to the interconnected and dynamic nature of phosphorylation networks, distinguishing between initial signalling events and those that result from the subsequent activation of additional signalling pathways is a common and challenging problem. We reasoned that kinases underlying both primary and secondary phosphorylation events could potentially be revealed by a global motif-based analysis of changes in the corresponding phosphoproteome. To test this idea, we used publicly available MS datasets from cells collected in the absence or presence of various perturbations and scored all phosphorylation sites with our atlas of Ser/Thr kinase motifs. Kinase motifs that were significantly enriched or depleted after experimental treatment were

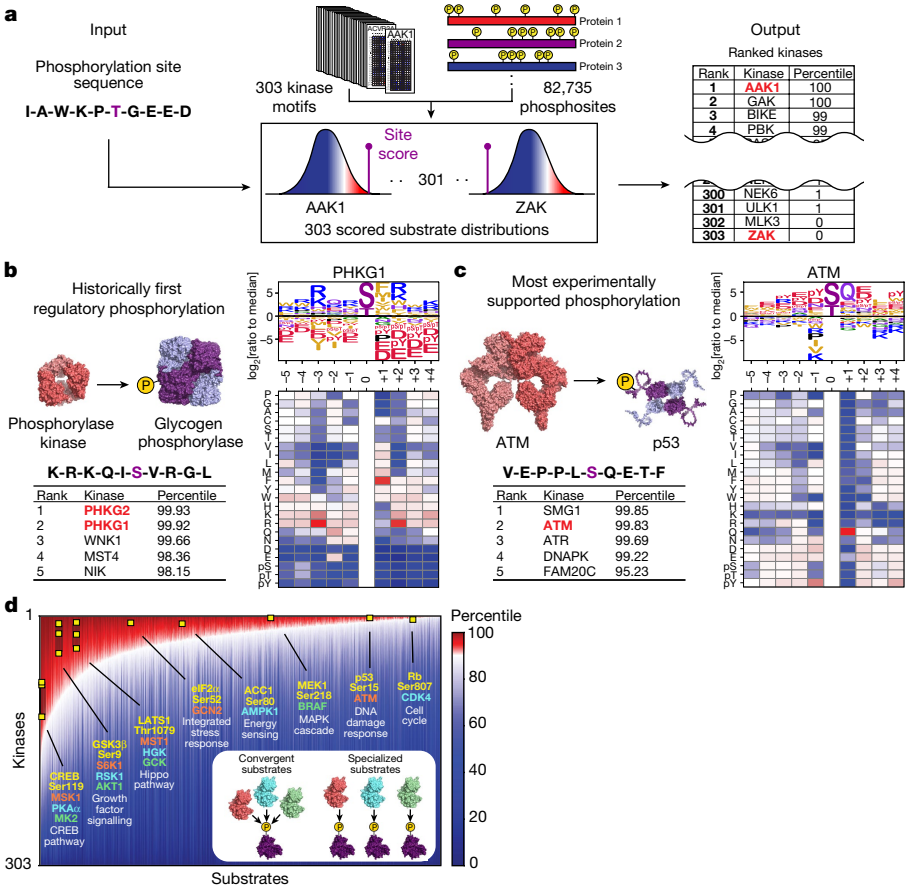


Fig. 3 | Phosphorylation motifs for the human Ser/Thr kinome enable comprehensive scoring and annotation of the human phosphoproteome. **a**, Schematic of the substrate-scoring process[4]. **b**, Results for Ser15 on glycogen phosphorylase alongside PSSM and the substrate motif logo of its established kinase glycogen phosphorylase kinase. **c**, The results for Ser15 of p53 along side its established kinase ATM. **d**, Annotation of the human Ser and Thr phosphoproteome by percentile scores from 303 Ser/Thr kinases performed as shown in **a**. A total of 89,752 phosphorylation sites that were identified using high-throughput approaches[4] and/or low-throughput approaches[3] were sorted along the *x*-axis by their numbers of kinases with percentile scores higher than 90. On the *y*-axis, kinase percentile scores were sorted by rank separately for each site and represented in the heat map. Examples of well-studied kinase–substrate relationships are highlighted (yellow squares). Inset: phosphorylation sites on the left end of the plot scored favourably for many kinases, whereas sites on the right end scored favourably for fewer kinases.

then represented as volcano plots of motif frequencies and adjusted *P* values (Fig. 4a).

Using this approach, we found that sequence motifs corresponding to the most direct target of a genetic or chemical perturbation were among the most significantly regulated, as seen, for example, for the genetic deletion of the secreted primordial casein kinase FAM20C (Fig. 4b). When quantitative phosphoproteomics data from HepG2 cells lacking FAM20C[44] were analysed using our kinome-wide dataset, the most downregulated kinase-recognition motif corresponded to that of FAM20C. Similarly, when skeletal-muscle-like myotube cells were stimulated for 30 min with isoproterenol[45], the most upregulated phosphorylation motifs corresponded to multiple isoforms of cAMP-dependent protein kinase (PKA)—canonical effector kinases downstream of the β_1 and β_2 adrenergic receptors (Fig. 4c). Notably, PKA motifs are highly similar to those of several other basophilic kinases, yet we could identify their enrichment in this scenario. Moreover, our comprehensive Ser/Thr kinome motif collection elucidated secondary signalling events in a dataset from HeLa cells arrested in mitosis using the PLK1 inhibitor BI 2536 (Fig. 4d)[46]; here, in addition to observing a notable downregulation of substrates containing the optimal PLK1 motif, we also noted upregulation of substrates phosphorylated by ATM and ATR. This finding is in good agreement with previous reports that PLK1 can suppress DNA damage signalling in mitotic cells[47,48].

Our motif-based analysis could also be used to reveal key signalling events resulting from more complex interventions. For example, we examined phosphoproteomic data from A549 cells treated with 6 Gy of ionizing radiation[49] (Fig. 4e). Our analysis revealed the up- and downregulation of numerous signalling pathways, including upregulation of canonical kinases involved in the DNA-damage response (ATM, ATR, DNA-PK) and downregulation of canonical kinases involved in cell cycle progression (CDK1, CDK2, CDK4 and CDK6), consistent with G1/S and G2/M arrest. Furthermore, we found up- and downregulation of less-appreciated DNA-damage-responsive kinases (MAPKAPK2[50,51], PLK3[52] and LRRK2[53]).

The full collection of Ser/Thr kinome motifs also enabled the temporal dynamics of signalling to be resolved from time-resolved phosphoproteomic datasets. For example, motif-based analysis of phosphoproteomic data from insulin-treated 3T3-L1 adipocytes[54] revealed rapid activation of the phosphoinositide 3-kinase signalling pathway within 1 min after insulin stimulation followed by subsequent activation of the MAPK pathway, together with downregulation of AMP-activated protein kinases within 60 min (Fig. 4f). Similarly, phosphoproteomic data analysis from lipopolysaccharide-stimulated dendritic cells[55] suggested marked upregulation at 30 min of a set of stress-activated kinases including the IKKs, JNK and p38 MAPKs, along with the MAPKAPK family of p38 effector kinases. This was followed within 4 h by the subsequent upregulation of the PIM kinases

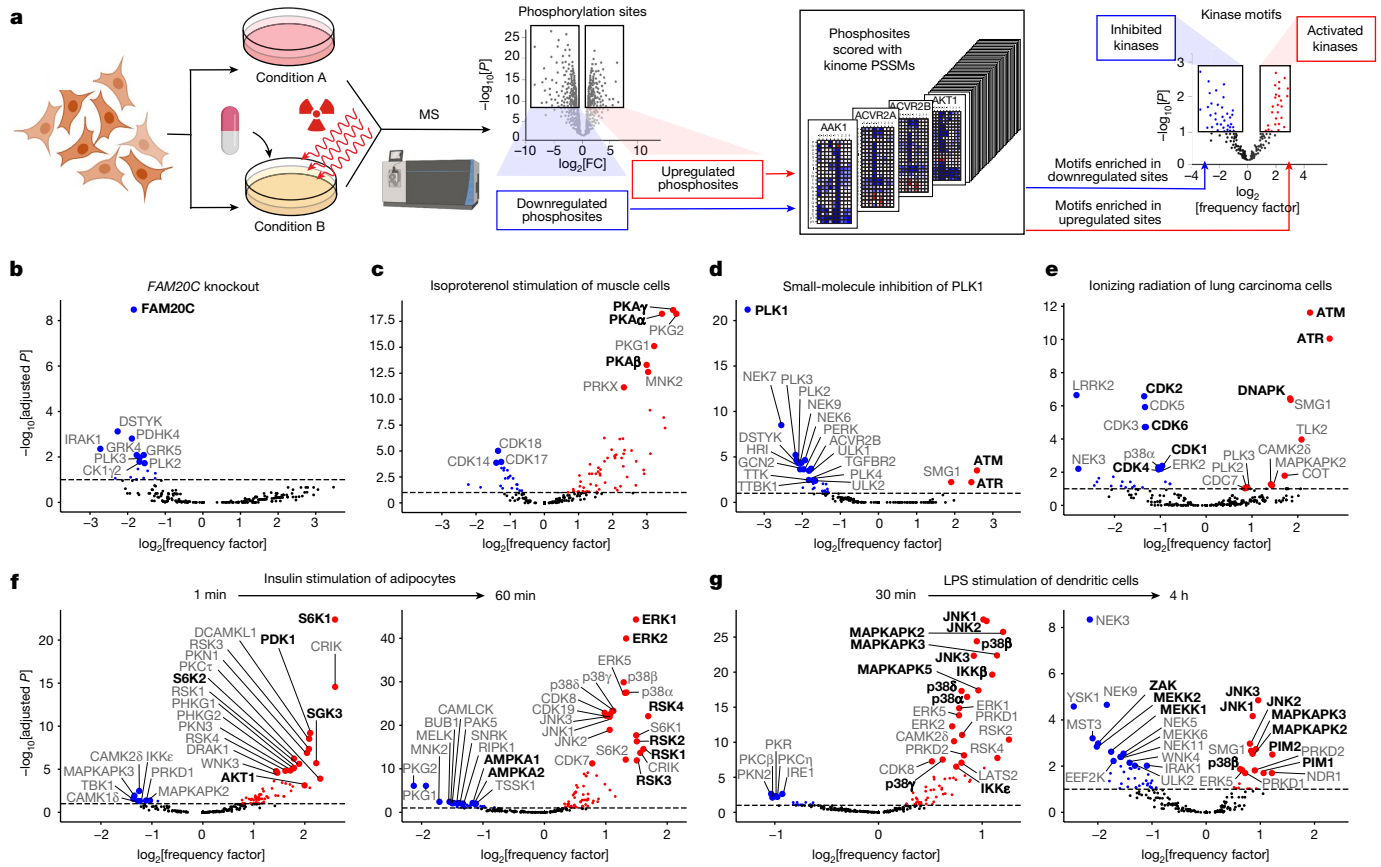


Fig. 4 | Global motif analysis reveals how kinase perturbations and pathway rewiring reshape the phosphoproteome. a, Workflow of the motif enrichment analysis of phosphoproteomics data. The schematic was created using BioRender. **b**–**g**, Results from published datasets. **b**, Conditioned medium of HepG2 cells after genetic deletion of *FAM20C*[44]. **c**, Cultured myotubes after 30 min treatment with 2 µM isoproterenol[45]. **d**, HeLa cells after mitotic arrest by treatment for 45 min with 0.1 µM PLK1 inhibitor BI 2536 (ref. [46]). **e**, A549 cells 2 h after exposure to 6 Gy of ionizing radiation[49]. **f**, 3T3-L1 adipocytes after serum starvation and then 1 min and 60 min treatment with 100 nM insulin[54]. **g**, C57BL/6J mouse bone-marrow-derived dendritic cells after 30 min and 4 h treatment with 100 ng ml^{-1} lipopolysaccharide (LPS)[55]. The enrichments in **b**–**g** were determined using one-sided exact Fisher's tests and corrected for multiple hypotheses using the Benjamini–Hochberg method. Fully annotated versions of these plots are presented in Supplementary Fig. 2.

and suppression of the MAPKs in parallel with the downregulation of their upstream MAPK3Ks (MEKK1, MEKK2 and ZAK)[56], suggestive of a negative-feedback loop (Fig. 4g). Thus, comprehensive motif-based approaches, when applied to time-resolved phosphoproteomics experiments, can decipher the distinct temporal dynamics of different groups of kinases.

Discussion

Here we present the full spectrum of substrate motifs of the human serine/threonine kinome and provide an unbiased comprehensive framework to further explore their cellular functions. Globally, these motifs are substantially more diverse than expected, suggesting a broader substrate repertoire of the kinome. Hierarchical clustering of this dataset reorganized the kinome into at least 38 motif classes and introduced several shared motif features (Fig. 2 and Extended Data Figs. 2–4).

The Ser/Thr kinases that we profiled were, almost without exception, strongly discriminatory against specific motif features. These findings suggest that fidelity in kinase signalling pathways is largely achieved through selective pressure on substrates to avoid phosphorylation by the majority of irrelevant kinases, and that this may occur by tuning the amino acid sequences surrounding the phosphorylation sites to be disfavoured by non-cognate kinases. As this negative selection contributes substantially to proper substrate recognition, accurate identification of kinase–substrate relationships requires a comprehensive knowledge of kinase phosphorylation motifs—not only for an individual kinase of interest, but also for all other kinases in the human kinome that might compete for the same substrate pool.

When this kinome-wide dataset was used to predict the specific kinases that are responsible for substrate phosphorylation solely based on the amino acid sequence surrounding the phosphorylation site, the results were highly accurate at identifying correct kinase–substrate relationships, even without knowledge of tissue specificity, scaffolding effects or subcellular localization. Including such additional information will probably further improve these predictive approaches[57,58]. A limitation of using first-order peptide arrays in these experiments is that they do not directly measure the contributions of interpositional contacts within the substrate peptides, which we have previously shown can affect substrate selection for some tyrosine kinases[32], albeit less so for Ser/Thr kinases[59]. Moreover, we were unable to differentiate between positional selection of Ser or Thr residues and direct phosphorylation of neighbouring residues (for example, peptides containing more than one phospho-acceptor). Structural modelling approaches guided by kinase substrate motif data will potentially decipher this additional information to further improve predictions[60,61].

The examination of MS phosphoproteomic datasets using this global collection of motifs yielded potential biological insights and putative kinase substrates (Fig. 4). For example, in cells undergoing exposure to ionizing radiation (Fig. 4e), ATM was predicted to target 37 of the phosphorylation sites that were upregulated, most of which have never been associated as substrates for ATM (Supplementary Table 4).

As the application of phosphoproteomics to human clinical samples and disease model systems continues to advance, our comprehensive motif-based approach will be uniquely equipped to unravel the complex signalling that underlies human disease progressions, mechanisms of cancer drug resistance, dietary interventions and other important physiological processes. In summary, we foresee that this will provide a valuable resource for a broad spectrum of researchers who study signalling pathways in human biology and disease.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05575-3.

1. Cohen, P. The origins of protein phosphorylation. *Nat. Cell Biol.* **4**, E127–E130 (2002).
2. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912–1934 (2002).
3. Hornbeck, P. V. et al. 15 years of PhosphoSitePlus®: integrating post-translationally modified sites, disease variants and isoforms. *Nucleic Acids Res.* **47**, D433–D441 (2019).
4. Ochoa, D. et al. The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **38**, 365–373 (2020).
5. Fuhs, S. R. & Hunter, T. pHisphorylation: the emergence of histidine phosphorylation as a reversible regulatory modification. *Curr. Opin. Cell Biol.* **45**, 8–16 (2017).
6. Hunter, T. Why nature chose phosphate to modify proteins. *Philos. Trans. R. Soc. B* **367**, 2513–2516 (2012).
7. Lahiry, P., Torkamani, A., Schork, N. J. & Hegele, R. A. Kinase mutations in human disease: interpreting genotype–phenotype relationships. *Nat. Rev. Genet.* **11**, 60–74 (2010).
8. Berginski, M. E. et al. The Dark Kinase Knowledgebase: an online compendium of knowledge and experimental results of understudied kinases. *Nucleic Acids Res.* **49**, D529–D535 (2021).
9. Edwards, A. M. et al. Too many roads not taken. *Nature* **470**, 163–165 (2011).
10. Needham, E. J., Parker, B. L., Burykin, T., James, D. E. & Humphrey, S. J. Illuminating the dark phosphoproteome. *Sci. Signal.* **12**, eaau8645 (2019).
11. Lemeer, S. & Heck, A. J. The phosphoproteomics data explosion. *Curr. Opin. Chem. Biol.* **13**, 414–420 (2009).
12. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
13. Riley, N. M. & Coon, J. J. Phosphoproteomics in the age of rapid and deep proteome profiling. *Anal. Chem.* **88**, 74–94 (2016).
14. Kemp, B. E., Graves, D. J., Benjamini, E. & Krebs, E. G. Role of multiple basic residues in determining the substrate specificity of cyclic AMP-dependent protein kinase. *J. Biol. Chem.* **252**, 4888–4894 (1977).
15. Kemp, B. E. & Pearson, R. B. Protein kinase recognition sequence motifs. *Trends Biochem. Sci.* **15**, 342–346 (1990).
16. Marin, O., Meggio, F., Marchiori, F., Borin, G. & Pinna, L. A. Site specificity of casein kinase-2 (TS) from rat liver cytosol: a study with model peptide substrates. *Eur. J. Biochem.* **160**, 239–244 (1986).
17. Clark-Lewis, I., Sanghera, J. S. & Pelech, S. Definition of a consensus sequence for peptide substrate recognition by $p44^{mpk}$, the meiosis-activated myelin basic protein kinase. *J. Biol. Chem.* **266**, 15180–15184 (1991).
18. Songyang, Z. et al. Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr. Biol.* **4**, 973–982 (1994).
19. Hutti, J. E. et al. A rapid method for determining protein kinase phosphorylation specificity. *Nat. Methods* **1**, 27–29 (2004).
20. Mok, J. et al. Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. *Sci. Signal.* **3**, ra12 (2010).
21. Pearce, L. R., Komander, D. & Alessi, D. R. The nuts and bolts of AGC protein kinases. *Nat. Rev. Mol. Cell Biol.* **11**, 9–22 (2010).
22. Kim, S.-T., Lim, D.-S., Canman, C. E. & Kastan, M. B. Substrate specificities and identification of putative substrates of ATM kinase family members. *J. Biol. Chem.* **274**, 37538–37543 (1999).
23. O'Neill, T. et al. Utilization of oriented peptide libraries to identify substrate motifs selected by ATM. *J. Biol. Chem.* **275**, 22719–22727 (2000).
24. Shah, N. H. et al. An electrostatic selection mechanism controls sequential kinase signaling downstream of the T cell receptor. *eLife* **5**, e20105 (2016).
25. Zhu, G. et al. Exceptional disfavor for proline at the P + 1 position among AGC and CAMK kinases establishes reciprocal specificity between them and the proline-directed kinases. *J. Biol. Chem.* **280**, 10743–10748 (2005).
26. Alexander, J. et al. Spatial exclusivity combined with positive and negative selection of phosphorylation motifs is the basis for context-dependent mitotic signaling. *Sci. Signal.* **4**, ra42 (2011).
27. Reiter, E. & Lefkowitz, R. J. GRKs and β-arrestins: roles in receptor silencing, trafficking and signaling. *Trends Endocrinol. Metab.* **17**, 159–165 (2006).
28. Moore, C. A., Milano, S. K. & Benovic, J. L. Regulation of receptor trafficking by GRKs and arrestins. *Annu. Rev. Physiol.* **69**, 451–482 (2007).
29. Bradley, D. et al. Sequence and structure-based analysis of specificity determinants in eukaryotic protein kinases. *Cell Rep.* **34**, 108602 (2021).
30. Taylor, S. S. & Kornev, A. P. Protein kinases: evolution of dynamic regulatory proteins. *Trends Biochem. Sci.* **36**, 65–77 (2011).
31. Creixell, P. et al. Unmasking determinants of specificity in the human kinome. *Cell* **163**, 187–201 (2015).
32. Miller, M. L. et al. Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.* **1**, ra2 (2008).
33. Chen, C. et al. Identification of a major determinant for serine-threonine kinase phosphoacceptor specificity. *Mol. Cell* **53**, 140–147 (2014).
34. Yaffe, M. B., Leparc, G. G., Lai, J., Obata, T., Volinia, S., & Cantley, L. C. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat. Biotechnol.* **19**, 348–353 (2001).
35. Yaron, T. M. et al. Host protein kinases required for SARS-CoV-2 nucleocapsid phosphorylation and viral replication. *Sci. Signal.* **15**, eabm0808 (2022).
36. Fischer, E. H., Graves, D. J., Crittenden, E. R. S. & Krebs, E. G. Structure of the site phosphorylated in the phosphorylase *b* to *a* reaction. *J. Biol. Chem.* **234**, 1698–1704 (1959).
37. Xu, B.-e, Wilsbacher, J. L., Collisson, T. & Cobb, M. H. The N-terminal ERK-binding site of MEK1 is required for efficient feedback phosphorylation by ERK2 in vitro and ERK activation in vivo. *J. Biol. Chem.* **274**, 34029–34035 (1999).
38. Malumbres, M. et al. Cyclin-dependent kinases: a family portrait. *Nat. Cell Biol.* **11**, 1275–1276 (2009).
39. Eick, D. & Geyer, M. The RNA polymerase II carboxy-terminal domain (CTD) code. *Chem. Rev.* **113**, 8456–8490 (2013).
40. Cohen, P. & Frame, S. The renaissance of GSK3. *Nat. Rev. Mol. Cell Biol.* **2**, 769–776 (2001).
41. Meng, Z. et al. MAP4K family kinases act in parallel to MST1/2 to activate LATS1/2 in the Hippo pathway. *Nat. Commun.* **6**, 8357 (2015).
42. Shaywitz, A. J. & Greenberg, M. E. CREB: a stimulus-induced transcription factor activated by a diverse array of extracellular signals. *Annu. Rev. Biochem.* **68**, 821–861 (1999).
43. Rigbolt, K. T. & Blagoev, B. Quantitative phosphoproteomics to characterize signaling networks. *Semin. Cell Dev. Biol.* **23**, 863–871 (2012).
44. Tagliabracci, V. S. et al. A single kinase generates the majority of the secreted phosphoproteome. *Cell* **161**, 1619–1632 (2015).
45. Needham, E. J. et al. Phosphoproteomics of acute cell stressors targeting exercise signaling networks reveal drug interactions regulating protein secretion. *Cell Rep.* **29**, 1524–1538 (2019).
46. Kettenbach, A. N. et al. Quantitative phosphoproteomics identifies substrates and functional modules of Aurora and Polo-like kinase activities in mitotic cells. *Sci. Signal.* **4**, rs5 (2011).
47. van Vugt, M. A. et al. A mitotic phosphorylation feedback network connects Cdk1, Plk1, 53BP1, and Chk2 to inactivate the G2/M DNA damage checkpoint. *PLoS Biol.* **8**, e1000287 (2010).
48. Macůrek, L. et al. Polo-like kinase-1 is activated by aurora A to promote checkpoint recovery. *Nature* **455**, 119–123 (2008).
49. Winter, M. et al. Deciphering the acute cellular phosphoproteome response to irradiation with X-rays, protons and carbon ions. *Mol. Cell. Proteom.* **16**, 855–872 (2017).
50. Reinhardt, H. C., Aslanian, A. S., Lees, J. A. & Yaffe, M. B. p53-deficient cells rely on ATM-and ATR-mediated checkpoint signaling through the p38MAPK/MK2 pathway for survival after DNA damage. *Cancer Cell* **11**, 175–189 (2007).
51. Reinhardt, H. C. & Yaffe, M. B. Kinases that control the cell cycle in response to DNA damage: Chk1, Chk2, and MK2. *Curr. Opin. Cell Biol.* **21**, 245–255 (2009).
52. Xie, S. et al. Plk3 functionally links DNA damage to cell cycle arrest and apoptosis at least in part via the p53 pathway. *J. Biol. Chem.* **276**, 43305–43312 (2001).
53. Gonzalez-Hunt, C. et al. Mitochondrial DNA damage as a potential biomarker of LRRK2 kinase activity in LRRK2 Parkinson's disease. *Sci. Rep.* **10**, 17293 (2020).
54. Humphrey, S. J. et al. Dynamic adipocyte phosphoproteome reveals that Akt directly regulates mTORC2. *Cell Metab.* **17**, 1009–1020 (2013).
55. Mertins, P. et al. An integrative framework reveals signaling-to-transcription events in toll-like receptor signaling. *Cell Rep.* **19**, 2853–2866 (2017).
56. Johnson, G. L. & Lapadat, R. Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. *Science* **298**, 1911–1912 (2002).
57. Miller, C. J. & Turk, B. E. Homing in: mechanisms of substrate targeting by protein kinases. *Trends Biochem. Sci* **43**, 380–394 (2018).
58. Linding, R. et al. Systematic discovery of in vivo phosphorylation networks. *Cell* **129**, 1415–1426 (2007).
59. Joughin, B. A., Liu, C., Lauffenburger, D. A., Hogue, C. W. & Yaffe, M. B. Protein kinases display minimal interpositional dependence on substrate sequence: potential implications for the evolution of signalling networks. *Philos. Trans. R. Soc. B* **367**, 2574–2583 (2012).
60. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
61. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

[1]Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. [2]Department of Medicine, Weill Cornell Medicine, New York, NY, USA. [3]Englander Institute for Precision Medicine, Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. [4]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [5]Tri-Institutional PhD Program in Computational Biology & Medicine, Weill Cornell Medicine, Memorial Sloan Kettering Cancer Center and The Rockefeller University, New York, NY, USA. [6]Weill Cornell Graduate School of Medical Sciences, Cell and Developmental Biology Program, New York, NY, USA. [7]Department of Medicine, Division of Hematology/Oncology, Columbia University Irving Medical Center, New York, NY, USA. [8]Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, New York, NY, USA. [9]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. [10]Center for Precision Cancer Medicine, Koch Institute for Integrative Cancer Biology, Departments of Biology and Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [11]Department of Pharmacology, Yale School of Medicine, New Haven, CT, USA. [12]Department of Chemistry, Yale University, New Haven, CT, USA. [13]Institute of Genetics, Technische Universität Braunschweig, Braunschweig, Germany. [14]Department of Pharmacology, Rutgers Robert Wood Johnson Medical School, Piscataway, NJ, USA. [15]Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN, USA. [16]Division of Endocrinology, Weill Cornell Medicine, New York, NY, USA. [17]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [18]Department of Biochemistry, University of Colorado, Boulder, CO, USA. [19]SABNP, Univ Evry, INSERM U1204, Université Paris-Saclay, Evry, France. [20]Department of Investigative Medicine, Graduate School of Medicine, University of the Ryukyus, Nishihara-cho, Japan. [21]Department of Developmental, Molecular and Chemical Biology, Tufts University School of Medicine, Boston, MA, USA. [22]Department Of Bioinformatics, Cell Signaling Technology, Danvers, MA, USA. [23]Rewire Tx, Humboldt-Universität zu Berlin, Berlin, Germany. [24]Department of Pharmacology, Weill Cornell Medicine, New York, NY, USA. [25]Department of Biochemistry, Weill Cornell Medicine, New York, NY, USA. [26]Divisions of Acute Care Surgery, Trauma, and Surgical Critical Care, and Surgical Oncology, Department of Surgery, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. [27]Surgical Oncology Program, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [28]These authors contributed equally: Jared L. Johnson, Tomer M. Yaron. ✉e-mail: ben.turk@yale.edu; myaffe@mit.edu; LCantley@med.cornell.edu

Methods

Cell lines

Expi293 (Thermo Fisher Scientific) and HEK293T (ATCC) cells were obtained directly from vendors that perform short tandem repeat genotyping for authentication of human cells and were verified to be mycoplasma-free. Sf9 insect cells were obtained from Thermo Fisher Scientific.

Plasmids

For expression and purification from bacteria, the DNA sequences for the human Ser/Thr kinases, binding partners and chaperones listed below were codon-optimized for expression in *Escherichia coli* using GeneSmart prediction software (GenScript). Optimized coding sequences were synthesized as gBlocks (Integrated DNA Technologies) carrying 16 bp overhangs at the 5′ and 3′ ends to facilitate in-fusion cloning (Clontech) into pET expression vectors (EMD Millipore).

pCDFDuet1 constructs were as follows: HSP90AA1-His_6(full length), hereafter referred to as 'HSP90'; untagged HSP90(full length); His_6-MO25a(full length); His_6-ALPHAK3/ALPK1 N-terminal domain (1–474); and His_8-CCNC(full length) in tandem with MED12-His_8(1–100); untagged MEK5/MAP2K5(S311D, T315D; full length); and untagged CK2B(full length). pET28a constructs were as follows: His_6-PDPK1(full length); His_6-PRP4/PRPF4B(519–end); GST-CAMK1A(full length); GST-CHAK2/TRPM6(1699–end); His_6-caMLCK/MYLK3(490–end); His_6-CAMKK1 (124–411); His_6-ERK7/MAPK15(full length); His_6-SUMO-ALPHAK3/ALPK1 CTD(959–end); MYO3A-His_6(1–308); ERK5/MAPK7-His_6(1–405); His_6-NIK/MAP3K14(327–673); and BMPR2-His_6(172–504). pETDuet1 constructs were as follows: His_6-CDK8(1–360, fusion with C-tail of CDK19(360–end)), His_6-CDK19(full length); His_6-AAK1(27–365); His_6-BIKE(37–345); CK2A1-His_6(full length); CK2A2-His_6(full length); His_{10}-MBP-MEKK1/MAP3K1(1174–end); His_6-CLK1(128–end); His_8-PLK2(57–360); His_{10}-MAP3K15(631–922); His_6-SUMO-ASK1/MAP3K5(659–951); and His_6-TAO2(1–350). The pACYDuet1 construct was as follows: untagged CDC37(full length).

For enhanced expression in mammalian lines cells, the DNA sequences of His_6-GST-SBK1(full length) and Flag-His_6-WNK3(1–434) were optimized for expression in *Homo sapiens* using GeneSmart (GenScript) and synthesized as gBlocks (Integrated DNA Technologies) carrying 16 bp overhangs to facilitate in-fusion cloning into digested pCDNA3.4 (Thermo Fisher Scientific).

To generate a mammalian expression construct for the TAK1/MAP3K7, the coding sequence for this kinase (GE Healthcare Dharmacon, MHS6278-202756930) and its binding partner TAB1 (GE Healthcare Dharmacon, MHS6278-202760135) were PCR-amplified and ligated as a fusion construct (TAK1(1–303)-TAB1(451–end)) into the mammalian expression vector pLenti-X by in-fusion.

Expression constructs purchased or obtained from other laboratories or Addgene were as follows: bacterial expression constructs for GST-VRK1(full length) and GST-VRK2(full length), in pGEX-4T, were received as gifts from P. Lazo[62]. The bacterial expression construct for mouse CDKL5-His_6(1–352), in pET23a+, was received as a gift from S. Katayama[63]. Bacterial expression constructs for His_6-SUMO-PDHK1(full length), His_6-SUMO-PDHK4(full length), pGroESL (GroEL/GroES) and MBP-BCKDK(full length) were received as gifts from D. Chuang, S.-C. Tso and R. Wynn[64,65]. pProEx HTa-BRAF_16mut V600E(444–721) was a gift from M. Therrien at Université de Montréal[66]. Mammalian expression constructs for Flag-ATR(S1333A) and HA-ATRIP were provided by D. Cortez[67]. Bacterial expression constructs for DMPK1, CAMK1G, CAMK2G, PHKG2, CDKL1, GAK and lambda phosphatase were purchased from Addgene (Addgene, 1000000094)[68].

Expression and Purification from bacteria

Transformations were performed using BL21 Star cells (Thermo Fisher Scientific) unless specified otherwise. Antibiotic concentrations used were as follows: carbenicillin (100 mg l^{-1}), kanamycin (50 mg l^{-1}), spectinomycin (25 mg l^{-1}) and chloramphenicol (25 mg l^{-1} in ethanol, prepared fresh). Transformed cells were grown in 1 l Terrific Broth by shaking at 190 rpm at 37 °C until the optical density (λ = 600 nm) reached 0.7–0.8, at which point 1 mM IPTG was added to induce expression. The cells were then transferred to a refrigerated shaker and shaken at 220 rpm at 18 °C for 16–20 h. Cells were centrifuged at 6,000*g*, and the pellets were snap-frozen in liquid nitrogen and stored at −80 °C.

All of the steps for protein purification were performed at 4 °C. Cell pellets were solubilized in lysis buffer (described below) and lysed by probe sonication. The lysate was centrifuged at 20,000*g* for 1 h and the supernatant was combined with affinity purification resin, nickel NTA (Qiagen) or glutathione Sepharose (GE Health) that had been rinsed in base buffer. The supernatant-bead slurry was agitated for 30 min. Resin was washed with 1 l base buffer and eluted in 10 bed volumes of elution buffer. Eluted protein was concentrated using the Ultra Centrifugal Filter Units (Amicon), supplemented with 1 mM DTT and 25% glycerol, and snap-frozen in liquid nitrogen and stored at −80 °C.

The buffers were as follows. Standard lysis buffer: 50 mM Tris pH 8.0, 100 mM NaCl, 2 mM $MgCl_2$, 2% glycerol, HALT EDTA-free phosphatase and protease inhibitor cocktail (Life technologies), 5 mM β-mercaptoethanol and 1–3 g of lysozyme (Sigma-Aldrich). Standard base buffer: 50 mM Tris pH 8.0, 100 mM NaCl, 2 mM $MgCl_2$, 2% glycerol (50 mM imidazole was included for purifications involving polyhistidine tags). Standard wash buffer: 50 mM Tris pH 8.0, 500 mM NaCl, 2 mM $MgCl_2$, 2% glycerol (50 mM imidazole was included for purifications involving polyhistidine tags). Polyhistidine-tag elution buffer: 50 mM Tris pH 8.0, 100 mM NaCl, 2 mM $MgCl_2$, 2% glycerol, 350 mM imidazole. GST-tag elution buffer: 50 mM Tris pH 8.0, 100 mM NaCl, 2 mM $MgCl_2$, 2% glycerol, 10 mM glutathione (pH was adjusted after addition of glutathione).

CDK8 was co-purified with CCNC/MED12. CDK19 was co-purified with CCNC/MED12. CK2A1 and CK2A2 were co-purified with CK2B. ERK5 was co-expressed with MEK5DD. The kinases BRAF and NIK were co-expressed with untagged HSP90–CDC37 complex. ALPHAK3 N- and C-terminal domains were co-purified. DMPK1, CAMK1G, CAMK2G, PHKG2, CDKL1 and GAK were co-expressed with lambda phosphatase in Rosetta 2 cells (Novagen). PDHK1, PDHK4 and BCKDK were co-expressed with GroeL/GroeS and purified with the following buffers: lysis buffer (100 mM potassium phosphate pH 7.5, 10 mM L-arginine, 500 mM KCl, 0.1 mM EDTA, 0.1 mM EGTA, 0.2% Triton X-100, lysozyme), wash buffer (50 mM potassium phosphate pH 7.5, 10 mM arginine, 500 mM NaCl, 0.1% Triton X-100, 2 mM $MgCl_2$), and elution buffer (25 mM Tris pH 7.5, 120 mM KCl, 0.02% Tween-20, 50 mM arginine, 350 mM imidazole for PDHK1 and PDHK4, 20 mM maltose for BCKDK). BCKDK was purified by its MBP tag on amylose resin (NEB). CDKL5 was expressed in BL21-codonplus(DE3)-RIL cells. KIS (full length) was purified as described previously[69].

Expression and purification from mammalian cells

Expi293F cells (Thermo Fisher Scientific) were cultured in 500 ml Expi293 Expression Medium (Thermo Fisher Scientific) in 2 l spinner flasks on a magnetic stirring platform at 100 r.c.f. at 36.8 °C under 8% CO_2. For transfection, 500 μg of expression constructs was diluted in Opti-MEM I Reduced Serum Medium (Thermo Fisher Scientific). ExpiFectamine 293 Reagent (Thermo Fisher Scientific) was diluted with Opti-MEM separately then combined with diluted plasmid DNA for 10 min at room temperature. The mixture was then transferred to the cells (3×10^6 cells per ml) and stirred. Then, 20 h after transfection, ExpiFectamine 293 Transfection Enhancer 1 and Enhancer 2 (Thermo Fisher Scientific) were added to the cells. Two days later, the cells were centrifuged at 300*g* for 5 min, snap-frozen in liquid nitrogen and stored at −80 °C (3 days after transfection).

All of the steps for protein purification were performed at 4 °C. Cell pellets were solubilized in lysis buffer and lysed by dounce

homogenization (20 strokes). The lysate was centrifuged at 100,000*g* for 1 h and the supernatant was combined with affinity purification resin, nickel NTA (Qiagen), glutathione Sepharose (GE Health) or anti-Flag M2 affinity gel (Sigma-Aldrich), and agitated for 30 min (nickel and glutathione beads) or 1 h (anti-Flag beads). Resin was washed with 1 l base buffer and eluted in 10 bed volumes of elution buffer. For elution of Flag tagged-proteins, beads were immersed in elution buffer (0.15 μg ml^{-1} 3× Flag peptide (Sigma-Aldrich)) and agitated for 1 h before elution. Eluted protein was concentrated using Ultra Centrifugal Filter Units (Amicon), supplemented with 1 mM DTT and 25% glycerol, and snap-frozen in liquid nitrogen and stored at −80 °C.

Buffers were as follows. Standard lysis buffer: 50 mM Tris pH 8.0, 150 mM NaCl, 2 mM $MgCl_2$, 5% glycerol, 1% Triton X-100, 5 mM β-mercaptoethanol, HALT protease inhibitors. Standard base buffer: 50 mM Tris pH 8.0, 100 mM NaCl, 2 mM $MgCl_2$, 2% glycerol. Standard wash buffer: 50 mM Tris pH 8.0, 500 mM NaCl, 2 mM $MgCl_2$, 2% glycerol.

His_6–GST-tagged SBK was purified sequentially on nickel and then glutathione resins. The buffers were as follows: the first wash buffer: 25 mM imidazole. SBK1 elution buffer for polyhistidine tag: 50 mM Tris pH 8.0, 100 mM NaCl, 2 mM $MgCl_2$, 2% glycerol, 250 mM imidazole. SBK1 elution buffer for GST tag: 50 mM Tris pH 8.0, 100 mM NaCl, 2 mM $MgCl_2$, 2% glycerol, 10 mM glutathione. Flag–TAK1–TAB1 elution buffer: 50 mM Tris pH 8.0, 100 mM NaCl, 2 mM $MgCl_2$, 2% glycerol, 0.15 μg ml^{-1} 3× Flag peptide.

Flag–His_6–WNK3 was purified sequentially on nickel and then anti-Flag resins. The buffers were as follows: the first wash buffer: 25 mM imidazole. Flag-tag elution buffer (chloride-free): 50 mM Tris pH 7.5, 2 mM magnesium acetate, 2% glycerol, 0.15 μg ml^{-1} 3× Flag peptide.

Flag–ATR(S1333A) (350 μl) and HA–ATRIP (150 μg) were co-transfected into Expi293 cells and incubated for one additional day after addition of enhancers (4 days after transfection). The buffers were as follows. ATR lysis buffer: 50 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 0.25% Tween-20, 2 mM $MgCl_2$, DTT. ATR wash buffer: 50 mM HEPES pH 7.4, 150 mM NaCl, 0.01% Brij-35, 2 mM $MgCl_2$, 5 mM ATP, DTT. ATR elution buffer: 20 mM HEPES pH 7.4, 150 mM NaCl, 0.01% Brij-35, DTT, 0.15 μg ml^{-1} 3× Flag peptide.

Eluates were concentrated to 1 ml in 100 kDa MWCO Amicon tubes and resolved using the MonoS column in a 0–1 M NaCl gradient (buffer: 25 mM Bis-Tris pH 6.9, 0.01% Brij-35 and 5 mM TCEP). A total of 1 ml of each fraction was collected. Fractions 1–4 were combined and concentrated to 1 ml using a 100 kDa MWCO filter and resolved using size-exclusion (Superose 6) in 20 mM HEPES pH 7.4, 200 mM NaCl, 0.01% Brij-35 and 5 mM TCEP. A total of 1 ml of each fraction was collected. Fractions 11–14 were verified to be pure ATR–ATRIP complex on SDS–PAGE.

SMG1–SMG9 complexes were purified from HEK293T cells as described previously[70]. RIPK1, RIPK2 and RIPK3 were purified from insect cells (Sf9) as described previously[71]. The following recombinant active kinases obtained from other laboratories. Recombinant active CDK12–CycK, CDK13–CycK and CDK9–CycT complexes were provided as gifts from M. Geyer[72,73]. Recombinant active DCAMKL1/DCLK1 and MELK were provided as gifts from N. Gray, H.-T. Huang and K. Westover, Y. Liu and W. Harshburger[74–76]. Recombinant active PRPK(full length)–CGI121/TPRKB(full length) complex was provided as a gift from L. Wan and F. Sicheri[77]. Recombinant active HASPIN(452–798) was provided as a gift from A. Musacchio[78]. Recombinant active YSK1 was provided as a gift from X. Luo[79]. Recombinant CK1G2 was provided as a gift from S. Knapp. A list of catalogue and lot numbers of purchased recombinant kinases is provided in Supplementary Table 1.

PSPA analysis

Recombinant kinase was added to a 384-well plate containing peptide substrate library mixtures in solution phase at 50 μM (Anaspec, AS-62017-1 and AS-62335). The reaction was initiated with the addition of 50 μM ATP (50 μCi ml^{-1} γ-^{32}P-ATP, Perkin-Elmer) and incubated for 90 min. The assay conditions for each kinase are described in Supplementary Table 1 (refs. [80–84]). After completion of the reaction, the solutions were spotted onto streptavidin-conjugated membranes (Promega, V2861), where the peptides tightly associated through their C-terminal biotinylation. The membranes were rinsed and then imaged using the Typhoon FLA 7000 phosphorimager (GE) to measure the extent of peptide phosphorylation. Raw data (GEL file) were quantified using ImageQuant (GE) to generate densitometry matrices (Supplementary Table 2). For the kinase ALPHAK3, spots were normalized to the surrounding background, owing to spatial variation in background signal. PDHK1 and PDHK4 showed dual specificity for serine and tyrosine. For these kinases, we used a customized peptide substrate library devoid of tyrosine residues at randomized positions.

In total, 283 human kinase motifs, one motif from a mouse kinase orthologue (CDKL5), one motif from a rat kinase orthologue (KIS) and one motif from an arthropod *Pediculus humanus corporis* kinase orthologue (PINK1), were combined with 17 human kinase motifs that we previously published, including AKT1[85], SRPK1[35], SRPK2[35], SRPK3[35], CK1D[35], DYRK1A[86], DYRK2[86], GSK3A[86], GSK3B[86], CK1A[86], CK1E[86], CK1G1[86], CDK10[87], CDK2[88], CDK3[88], CDK18[88] and CDK7[89].

For the zero-control experiments in Extended Data Fig. 6, biotinylated peptides were synthesized containing only serine or threonine as the phospho-acceptor, where all nine surrounding positions contained degenerate mixtures of the 20 natural amino acids excluding serine, threonine, tyrosine and cysteine.

Matrix processing

The densitometry matrices were column-normalized at all positions by the sum of the 17 randomized amino acids (excluding serine, threonine and cysteine), to yield PSSMs (Supplementary Table 2). PDHK1 and PDHK4 were normalized to the 16 randomized amino acids (excluding serine, threonine, cysteine and additionally tyrosine), corresponding to the uniquely customized peptide library that profiled these kinases. The cysteine row was scaled by its median to be 1/17 (1/16 for PDHK1 and PDHK4). The serine and threonine values in each position were set to be the median of that position. The ratio of serine versus threonine phospho-acceptor favourability (S_0 and T_0, respectively) was determined by summing the values of the serine and threonine rows in the densitometry matrix (S_S and S_T, respectively), accounting for the different serine versus threonine composition of the central (1:1) and peripheral (only serine or only threonine) positions (S_{ctrl} and T_{ctrl}, respectively), and then normalizing to the higher value among the two (S_0 and T_0, respectively, Supplementary Note 1).

Matrix clustering

The dendrogram in Fig. 2 was generated using the normalized matrices with the 20 unmodified amino acids, as well as phosphothreonine and phosphotyrosine. The linkage matrix was computed using the SciPy package in Python (v.3.7.6), using the Ward method. Results were converted to the Newick tree format and plotted using FigTree (v.1.4.4).

Substrate scoring

For scoring substrates, the values of the corresponding amino acids in the corresponding positions were multiplied and scaled by the probability of a random peptide (Supplementary Note 2).

For the percentile score of a substrate by a given kinase, we first computed the a priori score distribution of that kinase PSSM by scoring a reference Ser/Thr phosphoproteome comprising 82,735 identified sites[4] using the method discussed above (Fig. 3a). The percentile score of a kinase–substrate pair is defined as the percentile ranking of the substrate within the score distribution of each kinase[34]. This value was used when analysing all of the detected phosphorylation sites for kinase enrichment.

Kinase enrichment analysis

The single phosphorylation sites (not including multi-phosphorylated peptides) in the analysed phosphoproteomics studies were scored by

all of the characterized kinases (303 Ser/Thr kinases), and their ranks in the known phosphoproteome score distribution were determined as described above. For every non-duplicate, singly phosphorylated site, kinases that ranked within the top 15 kinases for the Ser/Thr kinases were considered to be biochemically favoured kinases for that phosphorylation site. For assessing kinase motif enrichment in phosphoproteomics datasets, we compared the percentage of phosphorylation sites for which each kinase was predicted among the upregulated/downregulated (increased/decreased, respectively) phosphorylation sites (sites with $|\log_2[\text{fold change}]|$ equal or greater than the log[fold change] threshold), versus the percentage of biochemically favoured phosphorylation sites for that kinase within the set of unregulated (unchanged) sites in this study (sites with $|\log_2[\text{fold change}]|$ less than the log[fold change] threshold). The log-transformed fold change threshold was determined to be 1.5 for all panels in Fig. 4, except for Fig. 4e, in which the threshold was set to 0.5 owing to the low range of the log[fold change] in the data. Contingency tables were corrected using Haldane correction (adding 0.5 to the cases with zero in one of the counts). Statistical significance was determined using one-sided Fisher's exact tests, and the corresponding *P* values were adjusted using the Benjamini–Hochberg procedure. Kinases that were significantly enriched (adjusted $P \leq 0.1$), or depleted ($\log_2[\text{frequency factor}] < 0$) for both upregulated and downregulated analysis were excluded from downstream analysis. Then, for every kinase, the most significant enrichment side (upregulated or downregulated) was selected on the basis of the adjusted *P* value and presented in the volcano plots.

Sequence logos

Sequence logos were made using logomaker package in Python[90]. For individual kinases, the normalized matrix was used, where the height of every letter is the ratio of its value to the median value of that position. The serine and threonine heights in the central position (position zero) were set to the ratio between their favourability. For clustered groups of kinases, the average matrix was calculated and presented as sequence logo as described above.

Comparative analyses between amino acids in the kinase domains and their substrate specificities

For Extended Data Fig. 7, kinases were sorted by their $\log_2[S_0/T_0]$ values. For the sequence logo, kinase domains of 290 available kinases were obtained from previously aligned kinase sequences[91]. The alignments to residues Met1–Leu296 in CDK2 (Protein Data Bank (PDB): 1QMZ) were obtained for each kinase, and the frequencies of amino acids in increments of 15-kinases were calculated and plotted as a sequence logo.

Known kinase–substrate pairs

Experimentally validated kinase–substrate relationships were obtained from PhosphoSitePlus (July 2021). The number of reports for each pair was determined by the sum of the in vivo and in vitro reports.

Illustrations

Experimental schema and illustrative models were generated using BioRender (https://biorender.com/). Kinome tree images were generated and modified using Coral (http://phanstiel-lab.med.unc.edu/CORAL/). Structural illustrations were generated using PyMOL. Generic kinase domains in Figs. 1 and 3 were as follows: PKAα (PDB: 1ATP). The kinase and substrate structures in Fig. 3 were as follows: ATM (PDB: 7SIC)[92] and p53 (chimera of AlphaFold AF-P04637-F1-model_v2_1 (1–95)[61] and 2ATA(96–292)[92]) (Fig. 3c), and PHKG2 (PDB: 2Y7J)[92] and PYGM (PDB: 1ABB)[92] (Fig. 3b).

Reporting summary

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

Data availability

The data generated (RAW files) and analysed in this study are provided in this paper. All plasmids generated in this study have either been deposited at Addgene or available on request.

Code availability

The analytical tools used in this study and underlying code are available to the public online (https://kinase-library.phosphosite.org).

62. Sanz-García, M. et al. Substrate profiling of human vaccinia-related kinases identifies coilin, a Cajal body nuclear protein, as a phosphorylation target with neurological implications. *J. Proteom.* **75**, 548–560 (2011).
63. Sekiguchi, M. et al. Identification of amphiphysin 1 as an endogenous substrate for CDKL5, a protein kinase associated with X-linked neurodevelopmental disorder. *Arch. Biochem. Biophys.* **535**, 257–267 (2013).
64. Wynn, R. M., Davie, J. R., Cox, R. P. & Chuang, D. T. Chaperonins GroEL and GroES promote assembly of heterotetramers (*α2β2*) of mammalian mitochondrial branched-chain alpha-keto acid decarboxylase in *Escherichia coli*. *J. Biol. Chem.* **267**, 12400–12403 (1992).
65. Song, J.-L., Li, J., Huang, Y.-S. & Chuang, D. T. Encapsulation of an 86-kDa assembly intermediate inside the cavities of GroEL and its single-ring variant SR1 by GroES. *J. Biol. Chem.* **278**, 2515–2521 (2003).
66. Thevakumaran, N. et al. Crystal structure of a BRAF kinase domain monomer explains basis for allosteric regulation. *Nat. Struct. Mol. Biol.* **22**, 37–43 (2015).
67. Luzwick, J. W., Nam, E. A., Zhao, R. & Cortez, D. Mutation of serine 1333 in the ATR HEAT repeats creates a hyperactive kinase. *PLoS ONE* **9**, e99397 (2014).
68. Albanese, S. K. et al. An open library of human kinase domain constructs for automated bacterial expression. *Biochemistry* **57**, 4675–4689 (2018).
69. Manceau, V. et al. Major phosphorylation of SF1 on adjacent Ser-Pro motifs enhances interaction with U2AF65. *FEBS J.* **273**, 577–587 (2006).
70. Melero, R. et al. Structures of SMG1-UPFs complexes: SMG1 contributes to regulate UPF2-dependent activation of UPF1 in NMD. *Structure* **22**, 1105–1119 (2014).
71. Najjar, M. et al. Structure guided design of potent and selective ponatinib-based hybrid inhibitors for RIPK1. *Cell Rep.* **10**, 1850–1860 (2015).
72. Czudnochowski, N., Bösken, C. A. & Geyer, M. Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition. *Nat. Commun.* **3**, 842 (2012).
73. Greifenberg, A. K. et al. Structural and functional analysis of the Cdk13/Cyclin K complex. *Cell Rep.* **14**, 320–331 (2016).
74. Liu, Y. et al. Chemical biology toolkit for DCLK1 reveals connection to RNA processing. *Cell Chem. Biol.* **27**, 1229–1240 (2020).
75. Ferguson, F. M. et al. Discovery of a selective inhibitor of doublecortin like kinase 1. *Nat. Chem. Biol.* **16**, 635–643 (2020).
76. Huang, H.-T. et al. MELK is not necessary for the proliferation of basal-like breast cancer cells. *eLife* **6**, e26693 (2017).
77. Wan, L. C. et al. Proteomic analysis of the human KEOPS complex identifies C14ORF142 as a core subunit homologous to yeast Gon7. *Nucleic Acids Res.* **45**, 805–817 (2017).
78. Villa, F. et al. Crystal structure of the catalytic domain of Haspin, an atypical kinase implicated in chromatin organization. *Proc. Natl Acad. Sci. USA* **106**, 20204–20209 (2009).
79. Bae, S. J., Ni, L. & Luo, X. STK25 suppresses Hippo signaling by regulating SAV1-STRIPAK antagonism. *eLife* **9**, e54863 (2020).
80. Murillo-de-Ozores, A. R., Chávez-Canales, M., de Los Heros, P., Gamba, G. & Castañeda-Bueno, M. Physiological processes modulated by the chloride-sensitive WNK-SPAK/OSR1 kinase signaling pathway and the cation-coupled chloride cotransporters. *Front. Physiol.* **11**, 585907 (2020).
81. Filippi, B. M. et al. MO25 is a master regulator of SPAK/OSR1 and MST3/MST4/YSK1 protein kinases. *EMBO J.* **30**, 1730–1741 (2011).
82. Zhou, P. et al. Alpha-kinase 1 is a cytosolic innate immune receptor for bacterial ADP-heptose. *Nature* **561**, 122–126 (2018).
83. Taipale, M. et al. Quantitative analysis of HSP90-client interactions reveals principles of substrate recognition. *Cell* **150**, 987–1001 (2012).
84. Klatt, F. et al. A precisely positioned MED12 activation helix stimulates CDK8 kinase activity. *Proc. Natl Acad. Sci. USA* **117**, 2894–2905 (2020).
85. Balasuriya, N. et al. Phosphorylation-dependent substrate selectivity of protein kinase B (AKT1). *J. Biol. Chem.* **295**, 8120–8134 (2020).
86. Zheng, Y. et al. Regulation of folate and methionine metabolism by multisite phosphorylation of human methylenetetrahydrofolate reductase. *Sci. Rep.* **9**, 4190 (2019).
87. Robert, T. et al. Development of a CDK10/CycM in vitro kinase screening assay and identification of first small-molecule inhibitors. *Front. Chem.* **8**, 147 (2020).
88. Ferguson, F. M. et al. Discovery of covalent CDK14 inhibitors with pan-TAIRE family specificity. *Cell Chem. Biol.* **26**, 804–817 (2019).
89. Rimel, J. K. et al. Selective inhibition of CDK7 reveals high-confidence targets and new models for TFIIH function in transcription. *Genes Dev.* **34**, 1452–1473 (2020).
90. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).
91. Modi, V. & Dunbrack Jr, R. L. A structurally validated multiple sequence alignment of 497 human protein kinase domains. *Sci. Rep.* **9**, 19790 (2019).
92. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

Acknowledgements We thank M. Begley and T. DeFalco for helpful discussions and technical assistance; and T. Levy, S. Beausoleil, L. D'Amato, P. Lobo, M. Tran and C. Valencia for technical

support. T.M.Y. thanks S. Yaron, N. Yaron, J. R. Haddad and S. Haddad for their support; and J.L.J. thanks M. Bak-Johnson for her support. This research was supported by Leukemia & Lymphoma Society FELLOW Award (to J.L.J. and L.C.C.); National Institute of Health grants P01 CA120964 (to L.C.C.), R35-CA197588 (to L.C.C.), P01-CA117969 (to L.C.C.), R35-ES028374 (to M.B.Y.), R01-CA226898 (to M.B.Y.), R01-GM104047 (to B.E.T. and M.B.Y.), U24 DK116204 (to B.E.T. and L.C.C.) and R35-GM139550 (to D.J.T.); the joint Cancer Research UK and Brain Tumour Charity funded Brain Tumour Award C42454/A28596 (to M.B.Y.); the Charles and Marjorie Holloway Foundation (to M.B.Y.); and the MIT Center for Precision Cancer Medicine. Support was also provided by the Cancer Center Support Grant P30-CA14051 and T32CA203702 (to E.R.K.) from the National Cancer Institute.

Author contributions J.L.J., T.M.Y., L.C.C., M.B.Y. and B.E.T. conceived the project, designed experiments and analysed the data. J.L.J. and T.M.Y. generated figures. J.L.J. performed the PSPA experiments. T.M.Y. led the computational analyses. T.M.Y., E.M.H., A.K., D.M.C., B.M.C., K.K., M.U., J.L., S.D.L., B.Z. and I.C. performed computational analyses. J.L.J., J.S., A.R., T.-Y.L., N.V., K.L., Y.M., A.D., A.Y. and A.M. generated recombinant proteins. J.L.J., T.M.Y., B.E.T., J.T.R., M.B.Y. and L.C.C. performed structural modelling. P.V.H., D.J.T., Y.T., N.A.-B., N.F.K., B.v.d.K., A.E.v.V., M.V.D., A.G.R., E.R.K., M.D.G., B.D.H., O.E., R.L. and J.B. contributed data and participated in discussions. J.L.J., T.M.Y., M.B.Y., B.E.T. and L.C.C. wrote and edited the manuscript with input from all of the authors.

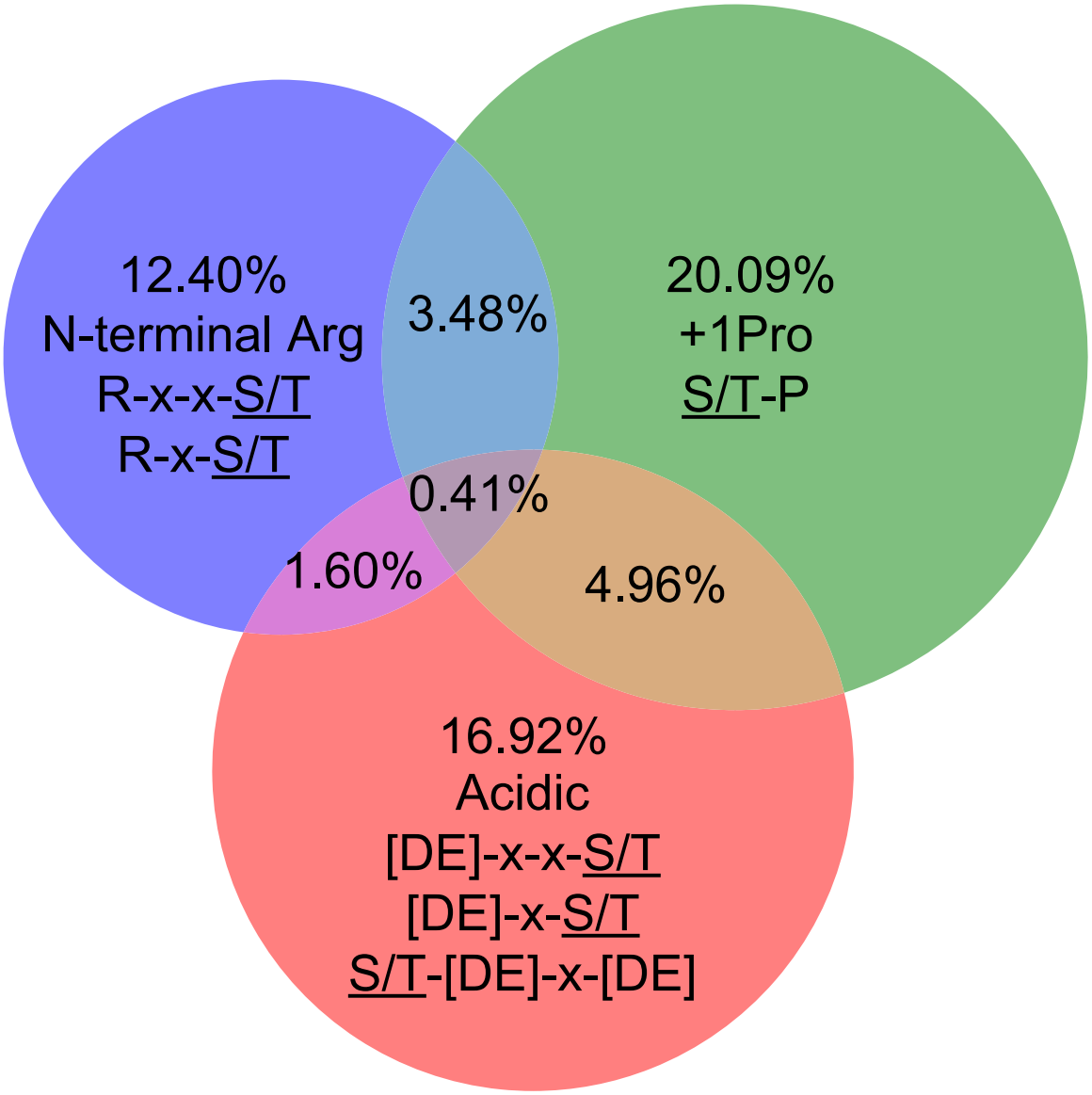
Competing interests L.C.C. is a founder and member of the board of directors of Agios Pharmaceuticals and is a founder and receives research support from Petra Pharmaceuticals; is listed as an inventor on a patent (WO2019232403A1, Weill Cornell Medicine) for combination therapy for PI3K-associated disease 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. M.B.Y receives research support from Cardiff Oncology. T.M.Y. is a co-founder and stockholder and is on the board of directors of DESTROKE, an early-stage start-up developing mobile technology for automated clinical stroke detection. J.L.J has received consulting fees from Scorpion Therapeutics and Volastra Therapeutics. O.E. is a founder and equity holder of Volastra Therapeutics and OneThree Biotech; is a member of the scientific advisory board of Owkin, Freenome, Genetic Intelligence, Acuamark and Champions Oncology; and receives research support from Eli Lilly, Janssen and Sanofi. D.J.T. is a member of the scientific advisory board at Dewpoint Therapeutics. A.D. is an equity holder of Denali Therapeutics; and receives research support from Interline Therapeutics. N.V. reports consulting activities for Novartis and is on the scientific advisory board of Heligenics. M.D.G. is a co-founder and shareholder of Faeth Therapeutics, which is developing dietary and pharmacological therapies for cancer; and has received speaking and/or consulting fees from Pfizer, Novartis, Scorpion Therapeutics and Faeth Therapeutics.

Additional information

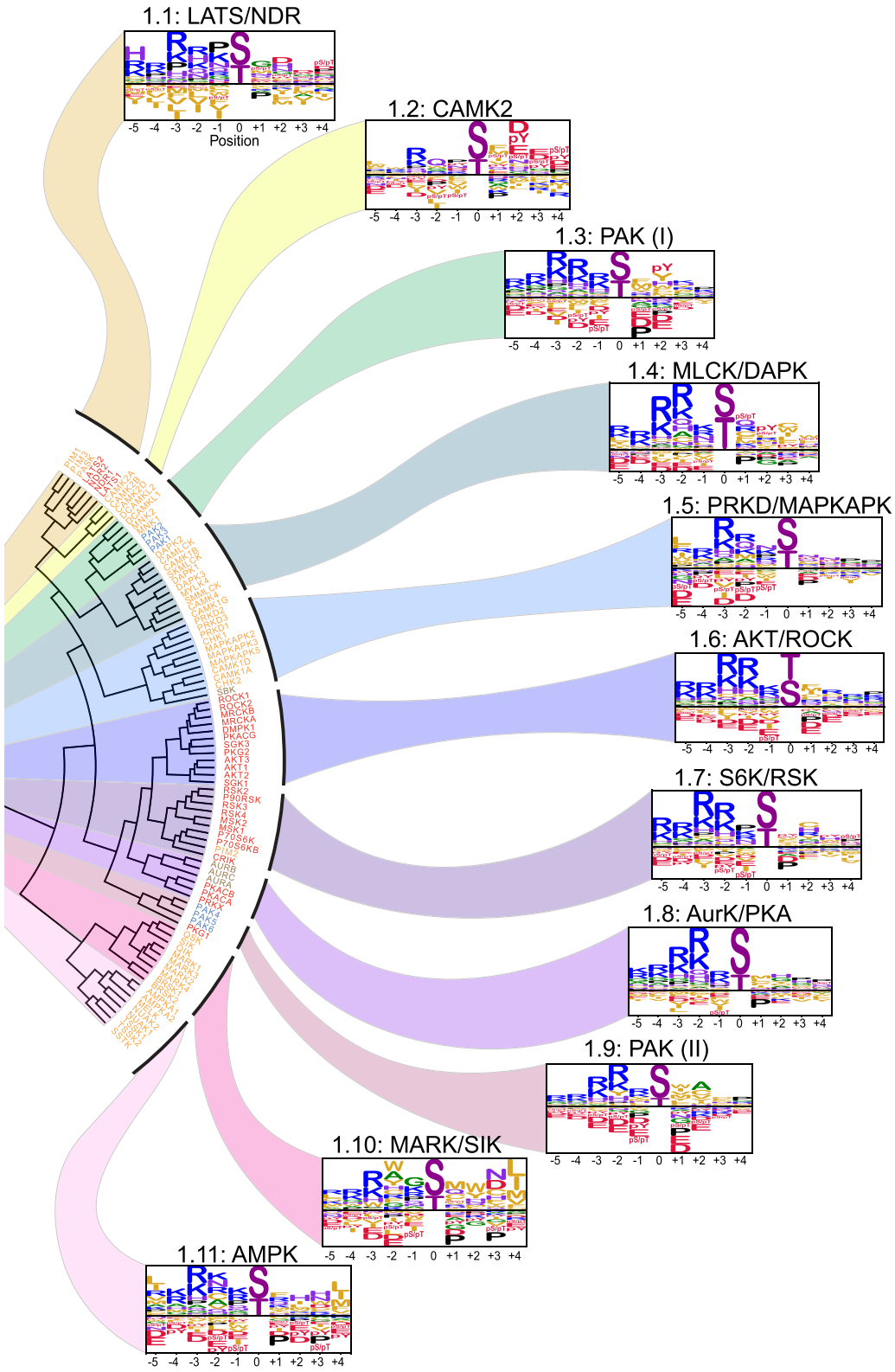
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05575-3.

Correspondence and requests for materials should be addressed to Benjamin E. Turk, Michael B. Yaffe or Lewis C. Cantley.

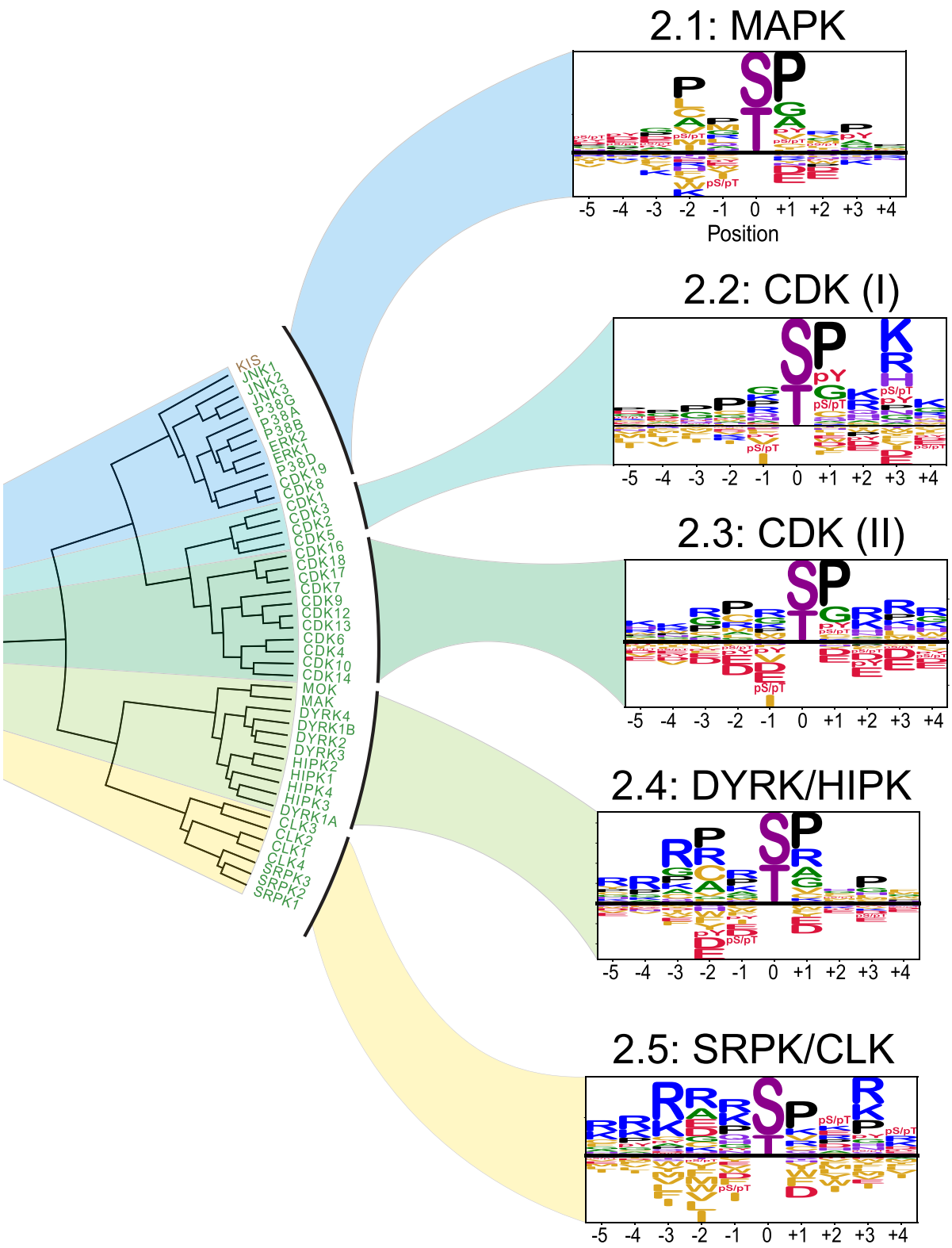
Peer review information *Nature* thanks Tony Hunter and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



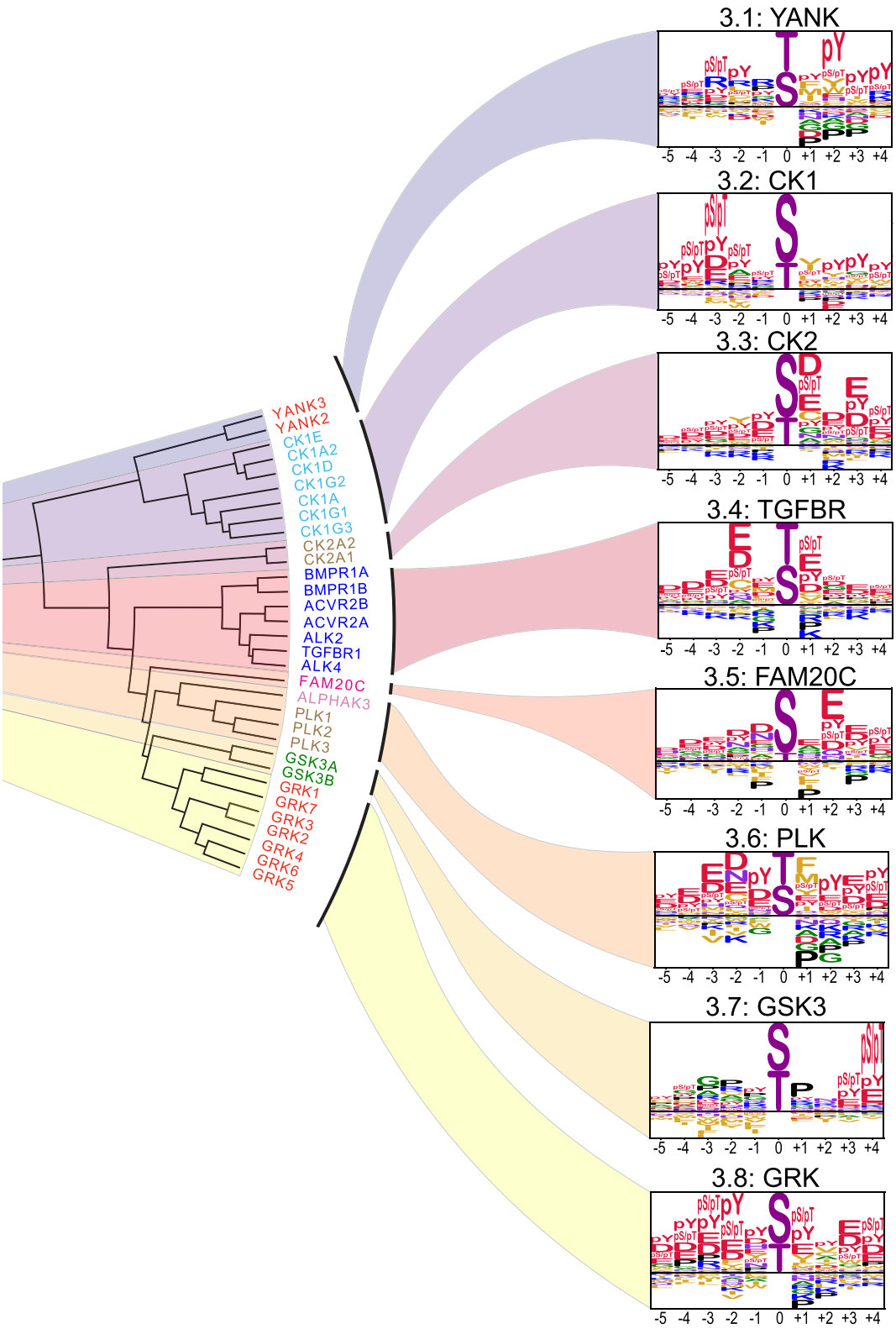
Extended Data Fig. 1 | Representation of phosphorylation site motifs in the human serine and threonine phosphoproteome. Venn diagram representation of the percentages of three prominent Ser/Thr kinase motif features, pertaining to Clusters 1, 2, and 3 in Fig. 2, across 82,735 human serine and threonine phosphorylation sites confidently identified in mass spectrometry experiments[4]. The phosphorylated residues in the logos are represented as S/T.



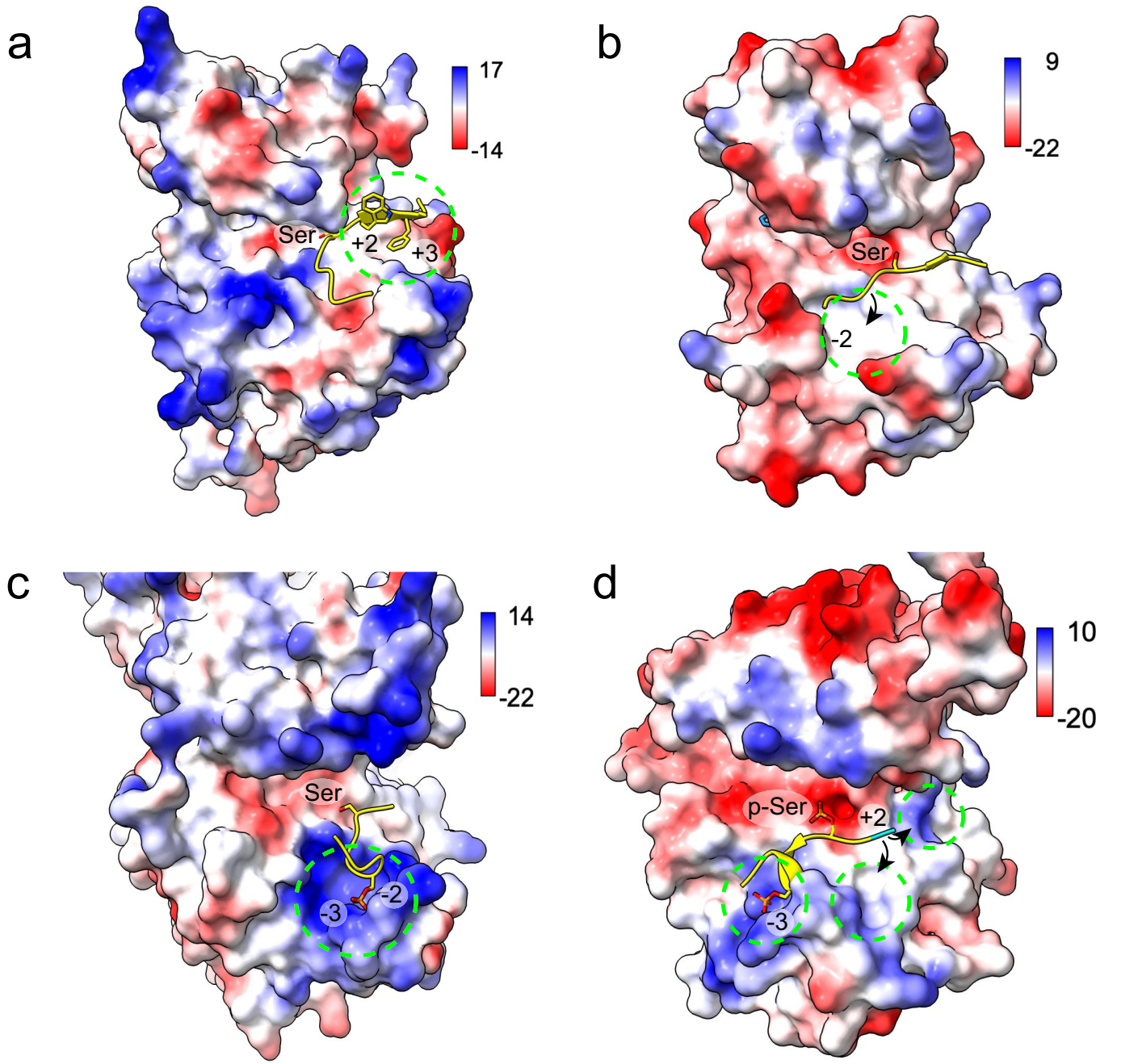
Extended Data Fig. 2 | Subcategorization of the basophilic kinases of Cluster 1. Subcategorization of Cluster 1 from Fig. 2 into 11 motif classes.



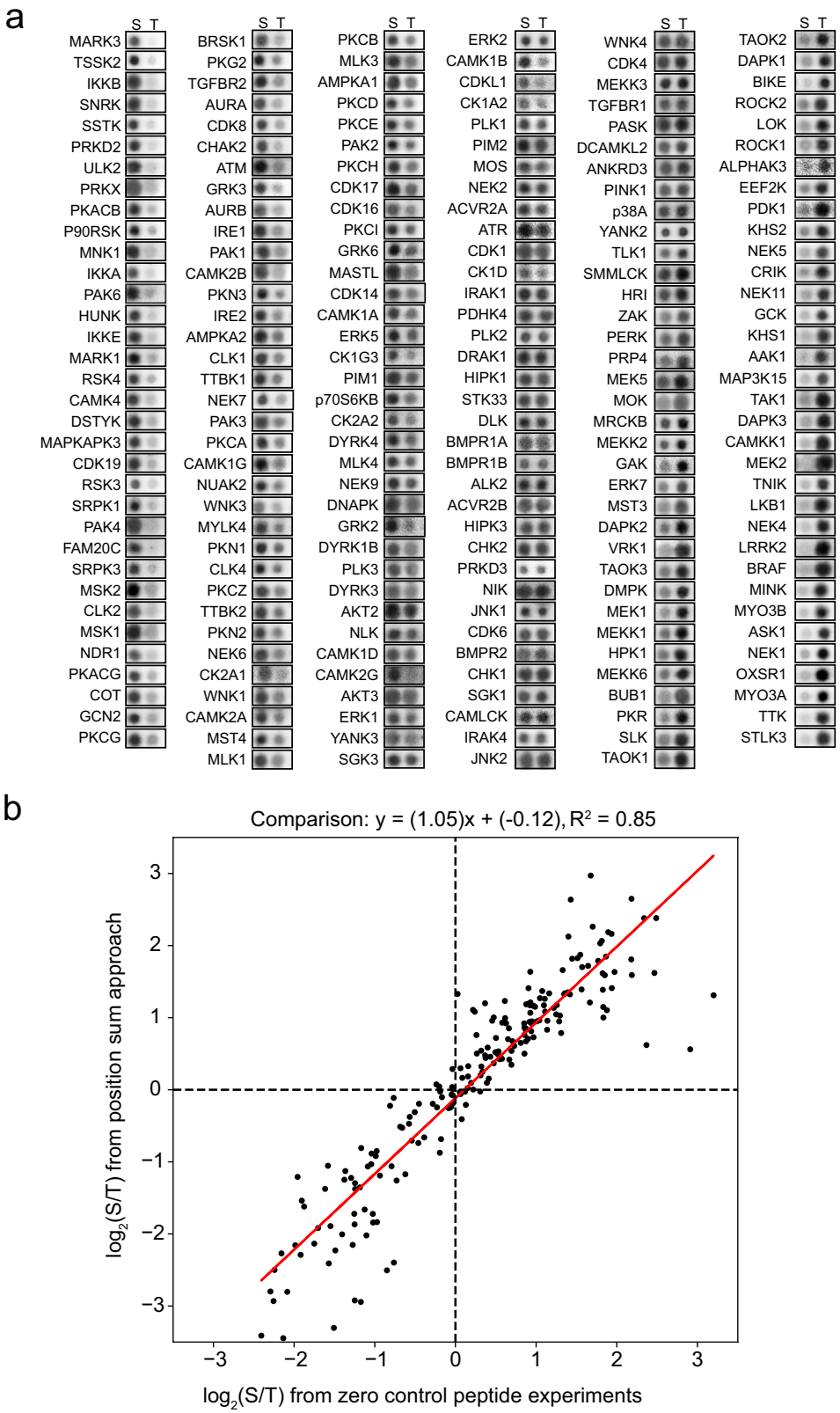
Extended Data Fig. 3 | Subcategorization of the proline-directed kinases of Cluster 2. Subcategorization of Cluster 2 from Fig. 2 into 5 motif classes.



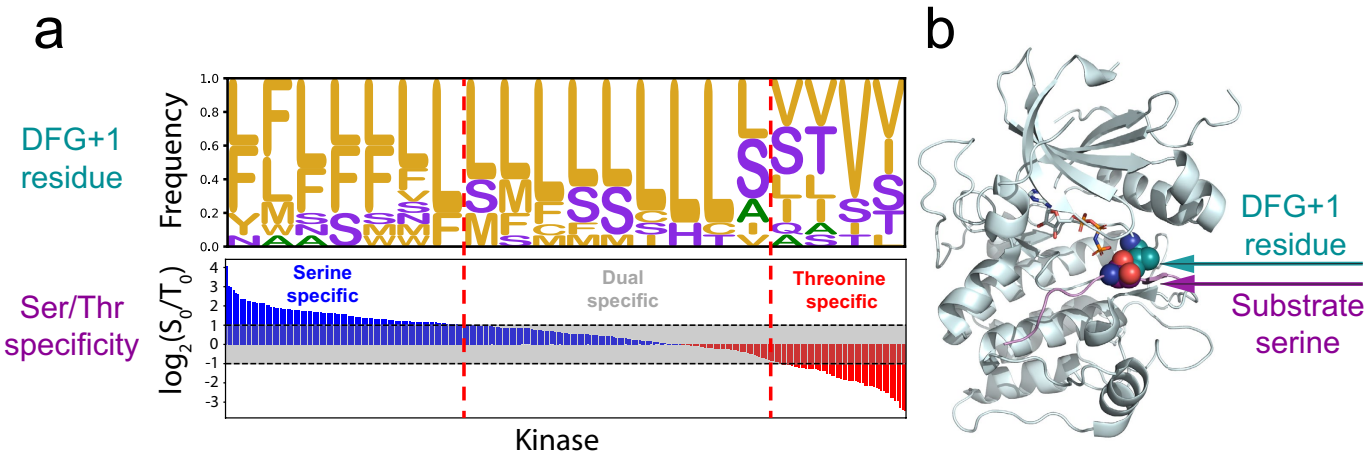
Extended Data Fig. 4 | Subcategorization of the acidophilic kinases of Cluster 3. Subcategorization of Cluster 3 from Fig. 2 into 8 motif classes.



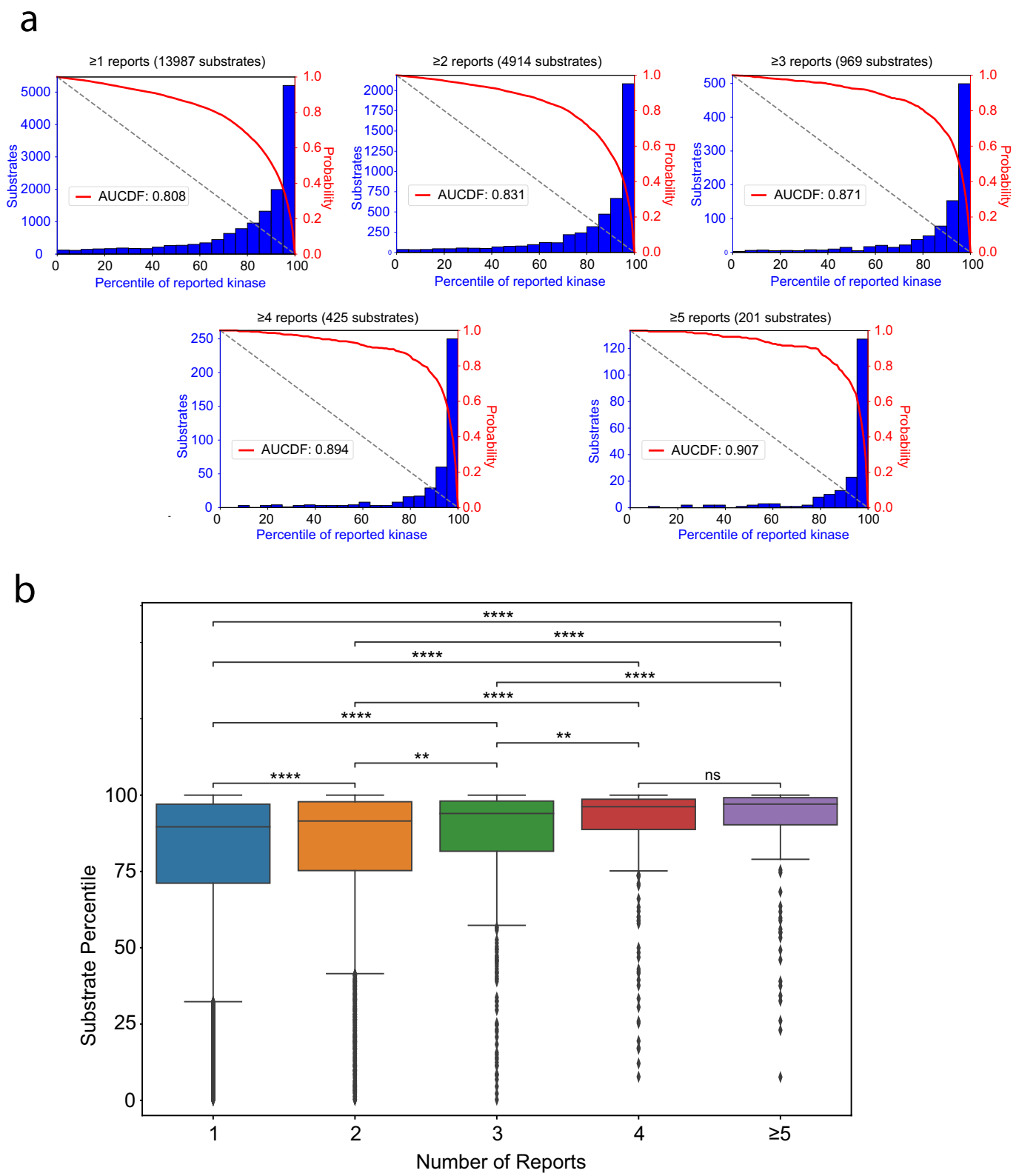
Extended Data Fig. 5 | Structural models of kinase–substrate complexes. **a**, Synthetic peptide from its complex with PAK4 (PDB: 2Q0N) modelled onto WNK3 (PDB: 5O26). Dotted circle highlights a shallow hydrophobic pocket accommodating a +3 Phe residue. **b**, GSK3 peptide from its complex with AKT2 (PDB: 1O6L) modelled onto CAMKK2 (PDB: 2ZV2). Circle indicates a hydrophobic pocket that could accommodate a −2 aliphatic residue. **c**, Monophosphorylated peptide from p63 bound to CK1δ (PDB: 6RU6) modelled onto GRK2 (PDB: 1YM7). Circle shows positive surface potential in the vicinity of the −2 and −3 pSer residues. **d**, p63 peptide bound to CK1δ (PDB: 6RU8) was modelled onto YANK1 (PDB: 4FR4) showing potential binding sites for −3 and +2 phosphorylated residues. Surface electrostatics are represented with Coulombic potential values were computed in ChimeraX and represented by scale bars (kcal/mol·e).



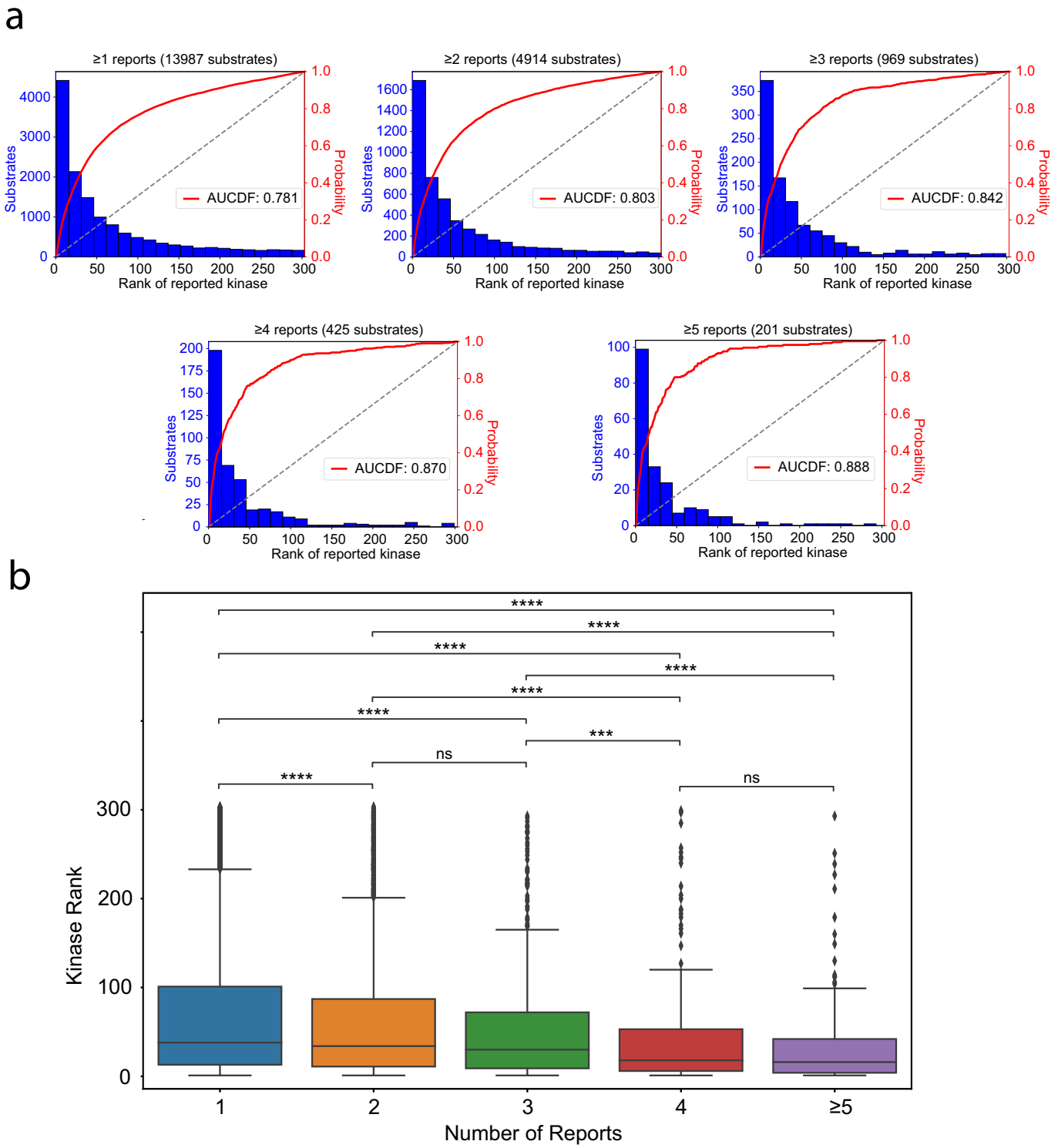
Extended Data Fig. 6 | Profiling phospho-acceptor specificity. a, *in vitro* phosphorylation assays with recombinant kinases and substrate peptides containing either serine or threonine phospho-acceptors. Results shown for 208 recombinant Ser/Thr kinases. **b**, Correlation plot of the experimental results in (**a**) with the position sum approach applied in this study to score Ser/Thr phospho-acceptor preference.



Extended Data Fig. 7 | Global analysis of the relationship between the DFG+1 amino acid and preference for the serine versus threonine phospho-acceptor. a, Bottom, relative preferences for Ser or Thr phospho-acceptor residues for each kinase, arranged in order of decreasing Ser/Thr selectivity. Top, frequency of amino acids at the DGF+1 positions of corresponding kinases (bin size: 15 kinases). **b**, Structural illustration of the proximity between the DFG+1 residue and substrate phospho-acceptor residue, shown with the AKT1 kinase domain bound to substrate (GSK3β) peptide (pdb 1O6K).

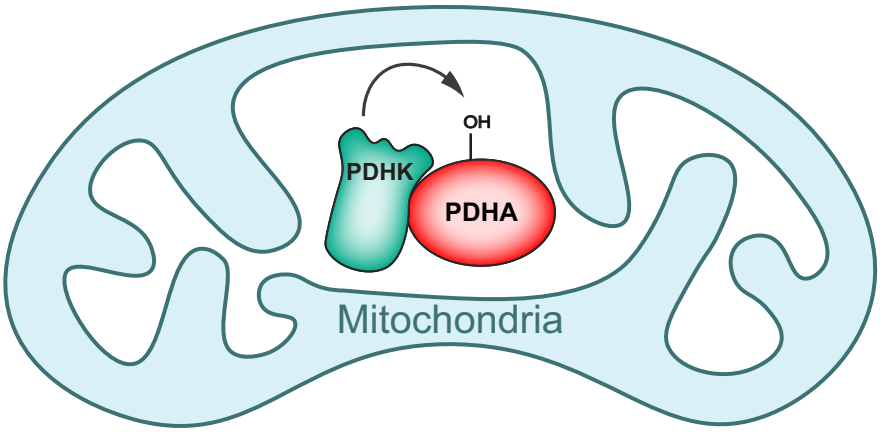


Extended Data Fig. 8 | Global performance analysis of substrate percentile scores for their literature-annotated kinases. a, Percentile-score distributions of substrates for their literature-annotated kinases (AUCDF= area under the cumulative distribution function). **b**, Percentile-score of literature-annotated kinase–substrate pairs as a function of number of reports. Higher number of reports correlates with more favourable percentile-scores between the reported kinase and its substrate. n = 9,073, n = 3,945, n = 544, n = 224, and n = 201 for kinase–substrate relationships with 1, 2, 3, 4, and 5 or more reports, respectively. Statistical analyses were performed using Double-sided Mann-Whitney U-test. Box minima=25th percentile, centre=50th percentile, maxima=75th percentile. Whiskers extend from the box maxima or minima to the largest or smallest value no further than 1.5 x interquartile range. (ns $p > 0.05$, * $p \leq 0.05$, ** $p \leq 10^{-3}$, *** $p \leq 10^{-4}$, **** $p \leq 10^{-5}$).



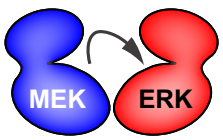
Extended Data Fig. 9 | Global performance analysis of kinase ranks for their literature-annotated substrates. a, Rank distributions of kinases for their literature-annotated substrates (AUCDF= area under the cumulative distribution function). **b**, Rank of the literature-annotated kinase–substrate pairs, as a function of number of reports. Higher number of reports correlates with more favourable ranking of reported kinase for its substrate. n = 9,073, n = 3,945, n = 544, n = 224, and n = 201 for kinase–substrate relationships with 1, 2, 3, 4, and 5 or more reports, respectively. Statistical analyses were performed using Double-sided Mann-Whitney U-test. Box minima=25th percentile, centre=50th percentile, maxima=75th percentile. Whiskers extend from the box maxima or minima to the largest or smallest value no further than 1.5 x interquartile range. (ns $p > 0.05$, * $p \leq 0.05$, ** $p \leq 10^{-3}$, *** $p \leq 10^{-4}$, **** $p \leq 10^{-5}$).

a Inhibition of the pyruvate dehydrogenase complex by PDHK family kinases



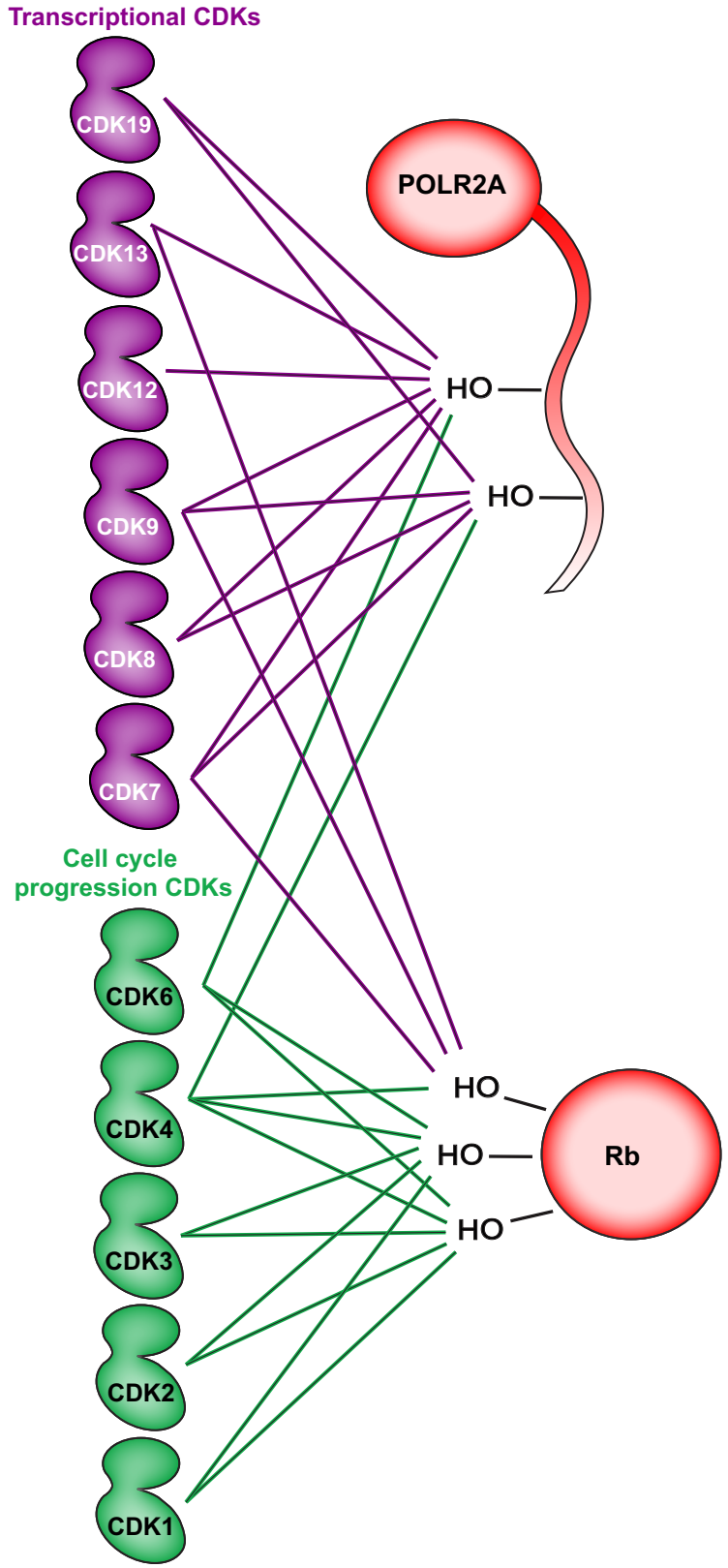
Substrate: PDHA1 Ser293		
Phosphorylation site: RYHGH**S**MSDP		
Rank	Kinase	Percentile
1	**BCKDK**	100.00
2	**PDHK1**	99.86
3	MASTL	99.77
4	**PDHK4**	99.58
5	GRK5	99.15
6	TLK2	98.98
7	HUNK	98.86
8	TTBK2	98.48
9	WNK4	98.46
10	ANKRD3	98.31

b Activation of ERK by MEK



Substrate: ERK1 Thr202/ERK2 Thr203		
Phosphorylation site: HTGFL**T**EYVA		
Rank	Kinase	Percentile
1	**MEK2**	99.81
2	**MEK1**	98.72
3	YANK2	98.71
4	MEK5	98.38
5	MASTL	98.23
6	ASK1	97.73
7	PRP4	95.95
8	VRK2	95.79
9	STLK3	95.11
10	EEF2K	94.03

Extended Data Fig. 10 | Motif-based scoring results for phosphorylation events facilitated by subcellular localization or docking. a, Illustration of the mitochondrial-localized regulation of the pyruvate dehydrogenase complex through phosphorylation by the PDHKs. Scoring results for PDHA1 Ser293, highlighting the PDHK family kinases. **b**, Illustration of docking-driven phosphorylation of ERK1/2 by MEK1/2. Scoring results for ERK1 Thr202 / ERK2 Thr203 (identical sequences), highlighting MEK1 and MEK2.



POLR2A

Substrate: POLR2A Ser1616 (ser2)		
Phosphorylation site: QSPSY**S**PTSP		
Rank	Kinase	Percentile
1	**CDK7**	95.68
2	**CDK13**	91.12
3	ERK5	90.85
4	**CDK12**	90.28
5	**CDK19**	89.89
6	**CDK9**	88.54
7	**CDK8**	88.37
8	CDK17	87.88
9	ERK2	87.86
10	**CDK6**	87.31

Substrate: POLR2A Ser1619 (ser5)		
Phosphorylation site: SYSPT**S**PSYS		
Rank	Kinase	Percentile
1	**CDK7**	99.35
2	ERK2	98.68
3	**CDK8**	98.61
4	**CDK19**	98.38
5	ERK1	97.60
6	PDHK4	97.22
7	**CDK9**	96.21
8	P38A	96.09
9	**CDK4**	95.97
10	P38B	95.37

Rb

Substrate: Rb Ser780		
Phosphorylation site: RPPTL**S**PIPH		
Rank	Kinase	Percentile
1	ERK5	98.74
2	DYRK4	98.49
3	**CDK7**	98.16
4	**CDK4**	98.08
5	DYRK1B	97.75
6	DYRK2	97.35
7	BUB1	97.23
8	PRP4	97.11
9	**CDK9**	97.07
10	**CDK13**	96.51

Substrate: Rb Ser807		
Phosphorylation site: GNIYI**S**PLKS		
Rank	Kinase	Percentile
1	**CDK2**	97.84
2	**CDK6**	94.30
3	**CDK3**	94.57
4	**CDK4**	92.33
5	**CDK1**	89.53
6	CDK5	85.43
7	PINK1	81.46
8	NLK	80.63
9	CDK14	78.38
10	ERK2	78.12

Substrate: Rb Ser811		
Phosphorylation site: ISPLK**S**PYKI		
Rank	Kinase	Percentile
1	NLK	99.50
2	**CDK1**	98.85
3	**CDK3**	98.84
4	**CDK2**	98.78
5	**CDK4**	98.08
6	ERK2	98.06
7	CDK5	97.72
8	**CDK6**	96.60
9	ERK5	96.27
10	CHAK2	96.08

Extended Data Fig. 11 | Scoring comparison of CDK subfamilies. Illustration of the phosphoregulation of RNA Polymerase II (POLR2A) CTD and Retinoblastoma protein (Rb) by their respective canonical CDKs, the transcriptional CDKs (purple) and the cell cycle progression CDKs (green) (left). Links between kinases and substrates correspond to favourable scores between motifs and phosphorylation sites (right).

nature portfolio

Corresponding author(s): Lewis C Cantley
Michael B. Yaffe
Benjamin E. Turk

Last updated by author(s): Nov 3, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Typhoon FLA 7000 phosphorimager (GE Healthcare, Marlborough, MA) was used to collect autoradiography data.
Data analysis	All data analysis was performed using Python (version 3.7.6). The dendrogram in Fig. 2 was displayed using FigTree (version 1.4.4). DNA sequence analysis was performed using SnapGene (version 5.0). Structural analyses were performed using PYMOL (version 2.4.1). Electrostatic calculations were performed using ChimeraX (version 1.4).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data generated and analyzed in the current study are provided with this paper.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No calculations were performed to determine sample sizes. Sample sizes (applicable to extended data figures 8b and 9b) were determined by their number of literature citations annotated on the public database PhosphoSitePlus, which contains the largest known collection of this information.
Data exclusions	No exclusion criteria.
Replication	2-3 replicates. All attempts at replications were successful.
Randomization	This does not apply to this study because the results are quantitative and did not require subjective judgment or interpretation.
Blinding	This does not apply to this study because the results are quantitative and did not require subjective judgment or interpretation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	Expi293 (Thermo-Fisher), HEK293T (ATCC), Sf9 (Thermo-Fisher)
Authentication	Cell lines were obtained from and authenticated by vendors.

Mycoplasma contamination	No mycoplasma contamination was found.
Commonly misidentified lines (See ICLAC register)	None.