

Understanding kinase-substrate interaction with deep learning and high-throughput scanning

Thesis by
Changhua Yu

In Partial Fulfillment of the Requirements for the
degree of
PhD, Bioengineering



CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2026
Defended May 6, 2026

© 2026

Changhua Yu

ORCID: 0000-0003-4799-4535

All rights reserved except where otherwise noted

ACKNOWLEDGEMENTS

I want to thank my PhD advisor, Dr. David Van Valen, for his unwavering support throughout my journey as a trainee in the lab. Dave taught me to approach problems critically and to always appreciate new knowledge and stay engaged with the field. He is one of the smartest yet most hardworking people I have had the privilege of working with, and he has inspired me to become a self-motivated researcher. I have also — not entirely intentionally — proudly inherited his unhealthy relationship with caffeine.

I want to thank my committee members, Dr. Yisong Yue, Dr. Stephen Mayo, and Dr. Paul Sternberg, for their scientific and career guidance during our meetings. Their insights shaped not only portions of this thesis but also helped me define a timeline to make the most of my training. I also want to thank Dr. Lea Goentoro, Dr. Erik Winfree, and Dr. Matt Thomson for their kind support and encouragement during moments when I felt lost. The generosity of the faculty has helped me trust in myself and pursue my passion on this scientific journey.

I want to thank Edward Pao, Emily Laubscher, and Morgan Schwartz for their patient mentorship in both wet lab and dry lab techniques when I first joined the lab. I would not have been able to pursue the work I envisioned without your guidance. I want to thank Ellen Emerson for being the wonderful desk mate. There were many blue days that turned golden through our conversations about science — and games, and shows.

I want to thank Xuefei Wang, Dr. Ross Barnowski, and all other lab members for their constant encouragement and support, both scientifically and emotionally. The lab feels like home because of all of you, and I will always cherish the many activities we shared and all the laughter we created together. (Special thanks also to Dr. Georgia Gkioxari for so generously conditioning that laughter to my productivity.)

I want to thank my friends at Caltech outside the lab — Yutian Li, Fayth Tan, Shichen Liu, Yujing Yang, Meichen Fang, Fred He, and Aki Ohdera. My lunch breaks would have been far lonelier without you. I am continually inspired by your talents, resilience, and dedication to science.

I want to thank my lifelong friends Yutong Han, Yuki Yu, and Hongzhao Guan, and my undergraduate friends Ying Huang, Josie Gu, Davy Deng, Ning Zhang, and Xiaoyan Kang. Despite the distances and time zones that separate us as we each

follow our own paths, the care we hold for one another has never diminished. I am honored to witness the growth and success of every one of you, and I treasure the long FaceTime calls that carry me back to a younger, more naïve, yet somehow more fearless version of myself.

I also want to thank my dance and table tennis friends — without you, and the joyful doses of dopamine those hobbies provide, I could never have returned to my yeast screens and lines of code feeling recharged.

I want to thank Cheng Shen and Shitian Li for always seeing my better sides and shielding me from my own insecurities. I want to thank my partner, Lawrence Bai, for always being my cushion and my harbor — for understanding my feelings deeply and for planning our travels in ways I never could have managed on my own.

I want to thank my family for the unconditional support that allowed me to pursue the path I chose, and for embracing me wholly and completely. The afternoon calls we shared were the anchors that kept me grounded on this journey so far from home. Dad, thank you for always patiently listening to my projects. I wish I could pick up the context of a new problem as effortlessly as you do — but even in the moments you don't quite follow, it means the world to me that you are simply there.

Lastly, Mom — I wish you all the best on your new journey. I still remember the day we bought a set of fancy measuring cylinders together, the reagent kit we used to diagnose the fish tank, and all the documentaries we watched and museums we visited. It was your curiosity about the world and your creativity in solving problems that first taught me to always ask why and how. It all began on the night we finally managed to fit that wick into the alcohol lamp and watched it light up.

ABSTRACT

Protein phosphorylation, catalyzed by more than 500 human kinases, governs almost every aspect of cellular signaling. Despite decades of research, a profound gap persists between the scale of the kinase signaling network and our ability to characterize it: over 95% of human phosphosites lack a confidently assigned kinase, approximately one-third of the kinome remains functionally understudied, and mechanisms by which kinase domains recognize and select substrates remain under-explored. Consequently, the druggable landscape for targeting phosphorylation rewiring events in disease contexts remains limited. This thesis seeks to address these challenges through complementary advances spanning machine learning, high-throughput interactome screening, and deep mutational scanning.

In Chapter 2, we develop KINBERT, a transformer-based protein language model that jointly encodes paired kinase domain and substrate peptide sequences, trained across 2,304 kinases and 92 organisms to enable genuine generalization to understudied kinases. We demonstrate its utility through disease variant interpretation and identification of host kinase targets during viral infection. In Chapter 3, we develop PhosphoPCA, a barcoded yeast-based platform that couples kinase–substrate phosphorylation to cell growth for high-throughput pooled profiling of kinase specificity, and apply it to identify novel substrates for understudied kinases and to engineer validated live-cell SPARK biosensors for the innate immune kinases such as TBK1. In Chapter 4, we leveraged double-deep phospho-PCA to enable deep mutational scanning of kinase domains with measurable components of fitness from protein stability, phosphorylation activity, and substrate selectivity.

Together, these three chapters establish an integrated framework for systematically decoding the kinase–substrate interaction space, providing a diverse set of novel technologies for illuminating the dark kinome, interpreting pathogenic pSNVs, and enabling generalizable engineering of kinase biosensors and therapeutics.

PUBLISHED CONTENT AND CONTRIBUTIONS

Chapter 3

1. Yu, C., Pao, E., & Van Valen, D. V. (2025). Abstract 1354: Understanding kinase substrate interactions with phospho-PCA and deep mutational scanning. *Journal of Biological Chemistry*, 301(5). <https://doi.org/10.1016/j.jbc.2025.110084>
2. Verma, A., Yu, C., Bachl, S., Lopez, I., Schwartz, M., Moen, E., Kale, N., Ching, C., Miller, G., Dougherty, T., Pao, E., Graf, W., Ward, C., Jena, S., Marson, A., Carnevale, J., Van Valen, D., & Engelhardt, B.E. (2024). Cellular behavior analysis from live-cell imaging of TCR T cell–cancer cell interactions. *bioRxiv* [Preprint]. 2024 Nov 21:2024.11.19.624390. <https://doi.org/10.1101/2024.11.19.624390>. PMID: 39605616; PMCID: PMC11601648.

B.E.E., A.M., S.J., C.W., C.Y., and A.V. conceived Occident. C.Y. developed whole-cell segmentation methodology for Occident and created whole-cell labels for Occident. C.Y. developed the image analysis pipelines used to analyze the data in this work. W.G. and D.V.V. oversaw software engineering for DeepCell Label and Caliban. C.Y. performed cell segmentation and tracking with DeepCell, SAM, and Caliban.

3. Marks, M., Israel, U., Dilip, R., Li, Q., Yu, C., Laubscher, E., Iqbal, A., Pradhan, E., Ates, A., Abt, M., Brown, C., Pao, E., Li, S., Pearson-Goulart, A., Perona, P., Gkioxari, G., Barnowski, R., Yue, Y., & Van Valen, D. (2025). CellSAM: A foundation model for cell segmentation. *Nature Methods*, 22(12), 2585–2593. <https://doi.org/10.1038/s41592-025-02879-w>

C.Y. developed CellSAM integration with live-cell biosensor quantification and lineage tracking. R.D., E. Pradhan, E. Pao, Q.L., C.Y., and E.L. performed data engineering. E.L., C.Y., and U.I. performed CellSAM integration with bioimaging workflows.

Chapter 4

1. Yu, C., Pao, E., & Van Valen, D. (2025). Uncovering the molecular basis of kinase activity and substrate recognition with phospho-PCA. *bioRxiv* [Preprint]. 2025 Nov 13:2025.11.07.687244. <https://doi.org/10.1101/2025.11.07.687244>. PMID: 41278632; PMCID: PMC12637564.

C.Y. and D.V.V. conceived the project. C.Y., E.P., and D.V.V. designed the experiments. C.Y. performed the experiments. C.Y. and D.V.V. designed the sequencing analysis and the method for computationally decomposing phospho-PCA mutational data. C.Y. implemented the data analysis code. C.Y. and D.V.V. analyzed the data. C.Y., E.P., and D.V.V. wrote the paper. D.V.V. supervised the project.

TABLE OF CONTENTS

Acknowledgements	iii
Abstract	v
Published Content and Contributions	vi
Table of Contents	vii
List of Illustrations	ix
List of Tables	xi
Chapter I: Introduction	1
1.1 The Dark Kinome and the Darker Phospho-Interactome	1
1.2 Molecular Basis of Kinase Domain Specificity	2
1.3 Limitations of Existing Experimental Approaches to Characterize the Phospho-Interactome	5
1.4 Gaps in Computational Models for the Phospho-Interactome Prediction	8
1.5 Biological Motivation for Accurate Kinase–Substrate Specificity Profiling	11
Chapter II: Protein language model enables generalizable Kinase-Substrate Interaction (KSI) prediction	15
2.1 Introduction	15
2.2 Results	15
2.3 Discussion	29
2.4 Methods	31
Chapter III: Massively-Parallel phospho-interactome characterization and novel kinase biosensor design enabled via phospho-PCA	35
3.1 Introduction	36
3.2 Results	39
3.3 Discussion	59
3.4 Methods	63
Chapter IV: Uncovering kinase stability, activity, and substrate specificity with double-deep phospho-PCA	71
4.1 Introduction	71
4.2 Results	73
4.3 Discussion	91
4.4 Methods	94
4.5 Supplementary Figures	103
Bibliography	112

LIST OF ILLUSTRATIONS

<i>Number</i>	<i>Page</i>
1.1 Overview of human protein kinase families, functions, disease roles, and understudied landscape.	3
1.2 Current experimental limitations of kinase-substrate specificity studies.	7
1.3 Survey of emerging high-throughput phosphorylation characterization of kinase-substrate specificity studies.	9
1.4 Survey of live-cell kinase biosensor designs.	13
2.1 Overview of the data collection, model architecture, and down-stream application of KINBERT	16
2.2 Details of compiled training corpus for KINBERT.	18
2.3 Model architecture optimization and test data performance benchmarking.	20
2.4 Generalizability benchmark across transformer-based KSI prediction models	22
2.5 Phosphorylation-rewiring event predictions in disease SNV contexts .	24
2.6 Viral phosphoproteomics and drug re-purposing pipeline schematics .	28
3.1 Optimization of phospho-PCA promoter and PAABD design	41
3.2 Bayesian decoupling framework based on Boltzmann energy states .	45
3.3 phospho-PCA shows high precision and reliability in benchmarks, reveals co-evolution of kinase specificity and sequence divergence, and sheds novel insight for understudied kinases	48
3.4 Finetuning KINBERT with phospho-PCA data enables identification of kinase reporter candidates from unseen substrates	52
3.5 CellSAM-Generalist and Occident tracking pipeline enable automated segmentation, tracking, and downstream quantification for live-cell imaging data	55
3.6 SPARK reporters for TBK1 and IKK-beta	57
3.7 Inspection of the substrate-only Libraries and decision boundary for false positive thresholds	62
3.8 Abundance and Activity Decoupling across PAABDs	63

4.1	Phospho-PCA enables deep mutational scanning of kinase domains to understand the sequence-to-function relationship for this enzyme category	75
4.2	Phospho-PCA quantitatively couples kinase-substrate interactions to cell growth	76
4.3	Mutations in conserved regions impact abundance-PCA scores of SFKs	78
4.4	Phospho-PCA reveals activity-biasing mutations and the functional impact of clinical variants	82
4.5	Molecular dynamics reveals allosteric networks governing substrate specificity	88
4.6	Validation of phospho-PCA measurements against published deep mutational scanning data	104
4.7	MoCHI model reveals incomplete decomposition of Fyn mutational data	104
4.8	Protein language model attention analysis identifies a subset of functionally important sites	105
4.9	Evolutionary and biophysical characterization of the V195E gain-of-function mutation	105
4.10	Hierarchical Bayesian model diagnostics demonstrate robust parameter inference	106
4.11	Allosteric hub analysis identifies differentially correlated protein segments in P203R mutant	107
4.12	Pooled abundance and decomposed kinase activity scores across Src-family kinases	108
4.13	Known kinase domain catalytic sites shows negative abundance and activity scores	108
4.14	Illustration of AbundancePCA and PhosphoPCA screening Plasmids .	109
4.15	Substrates and Wild Type kinases in the screenings	109

LIST OF TABLES

<i>Number</i>	<i>Page</i>
2.1 Comparison of training-data scope across KSI prediction methods. . .	17
2.2 Kinase-family and individual-kinase coverage in computational KSI prediction methods.	17
2.3 Distribution of interactions per kinase for database-curated (DB Only) and combined database plus Mass-Spec (DB + Lysate) sources. . . .	17
2.4 Metrics: model architecture optimization and test data performance benchmarking.	20
2.5 Metrics for generalizability benchmark across transformer-based KSI prediction models	23
2.6 Input datasets for the viral phosphoproteomics and drug-repurposing pipeline.	27
3.1 Library processing statistics after QC filtering.	42
3.2 Final constructed library kinase and substrate statistics	42
3.3 Performance metrics across models on phospho-PCA-identified KSIs	51
3.4 SPARK reporter substrate sequences and KINBERT:TAPE scores. . .	54
3.5 phospho-PCA performance statistics on spike-in KSIs.	63
4.1 Sequencing coverage for abundance- and phospho-PCA libraries . . .	110
4.2 Divergent abundance-PCA mutations that disrupt van der Waals in- teraction networks.	110
4.3 Specificity-determining variants.	111

Chapter 1

INTRODUCTION

1.1 The Dark Kinome and the Darker Phospho-Interactome

Protein phosphorylation is one of the most prevalent post-translational modifications in eukaryotic cells and regulates nearly every aspect of cellular physiology. The human genome encodes more than 500 protein kinases capable of modifying a substantial fraction of the proteome (Graham et al., 2004). Protein kinases catalyze the transfer of the γ -phosphate group from ATP onto serine, threonine, or tyrosine residues of substrate proteins, thereby altering enzymatic activity, localization, stability, and molecular interactions (Figure 1.1). Through these mechanisms, phosphorylation governs processes including cell cycle progression, cell proliferation, differentiation, metabolism, and checkpoint signaling. Dysregulation of kinase signaling is a recurring feature of human diseases, particularly cancer, immune disorders, and neurodegeneration (Barefield & Sadayappan, 2010; J. Ji et al., 2015; Vogelstein et al., 2013). In response to upstream stimuli, protein kinases selectively transfer phosphate groups to their substrates, significantly affecting the behavior of the substrate protein by causing changes in enzymatic activity, cellular localization, and affinity for binding partners (Buljan et al., 2020; Roland Akinseye, 2024; Voliotis et al., 2014). Given the central role that protein kinases play in the cell's information-processing capabilities, their ability to control their activity and select a specific set of substrates is essential to their function. Dysregulation of protein kinases can be caused by single missense mutations, or variants, and these variants are important drivers of numerous diseases (Medvedev et al., 2023; Torkamani et al., 2009; L. Wang et al., 2021), including autoimmunity and cancer. Despite decades of research, a large portion of the kinome remains poorly characterized. Approximately one-third of human kinases—around 160 proteins—lack in-depth study of their function, regulation, and substrate relationships, constituting what is collectively termed the "dark kinome" (Gomez et al., 2024).

The "darkness" extends beyond the kinases themselves to their interactions. Large-scale phosphoproteomics has cataloged hundreds of thousands of phosphosites across the human proteome (Huttlin et al., 2010; Lundby et al., 2012), and more than 534,000 experimentally observed phosphorylated residues are now documented.

However, the upstream enzymes responsible for most of these events remain unknown: more than 95% of human phosphosites lack an assigned kinase (Needham et al., 2019). Current phosphoproteomics has cataloged enormous numbers of phosphosites across the proteome, yet the kinases responsible for most of these modification events remain unknown. This disconnect constitutes the darker phospho-interactome: we know the modifications exist, but not the regulatory wiring diagram that produces them. As we would elaborate in 1.2, 1.3, current consortium-scale efforts on profiling phosphorylation motifs and assigning predicted kinase-substrate interactions contains inherent limitations that is hard to transcend onto absolute sequence spaces of kinase-substrate interactions.

Systematic efforts to address this gap have underscored that available resources provide correlational evidence rather than definitive substrate mappings: network inference is not equivalent to causal assignment, and in vitro specificity does not guarantee in vivo substrate selection. Many dark kinases involved in diseases lack functional annotation, validated substrates, and any known downstream targets, creating a compounding bottleneck for translational relevance and therapeutic prioritization (Gomez et al., 2024). Earlier estimates suggested that only 8% of the kinome had been effectively targeted for approved therapeutics; updated analyses through 2024 indicate that approximately 80 FDA-approved small-molecule inhibitors now collectively target around 25 unique human kinases. (Attwood et al., 2021; Roskoski, 2024). This still leaves the majority of all human kinases without a clinical-stage modulator (Attwood et al., 2021). Current phosphorylation omics data derived from genetic, transcriptomic, or proteomic approaches therefore only point indirectly to kinase activities, providing limited value for functional studies of kinases in pathogenesis and for precision medicine development. Improved knowledge of how variants affect catalytic activity of kinase domains as well as substrate recognition will be critical for understanding broader concepts of protein stability, structural determinants of specificity, and mechanisms of allostery, with the potential to generate potent and specific molecules that alter kinase activity without off-target effects (Y. Liu et al., 2025; Martinez et al., 2020; X. Xu et al., 2019).

1.2 Molecular Basis of Kinase Domain Specificity

1.2.1 The Kinase Catalytic Domain Recognizes Phosphosite Flanking Residues

With over half a million experimentally observed human phosphorylation sites, kinases must consistently discriminate among thousands of potential substrates to

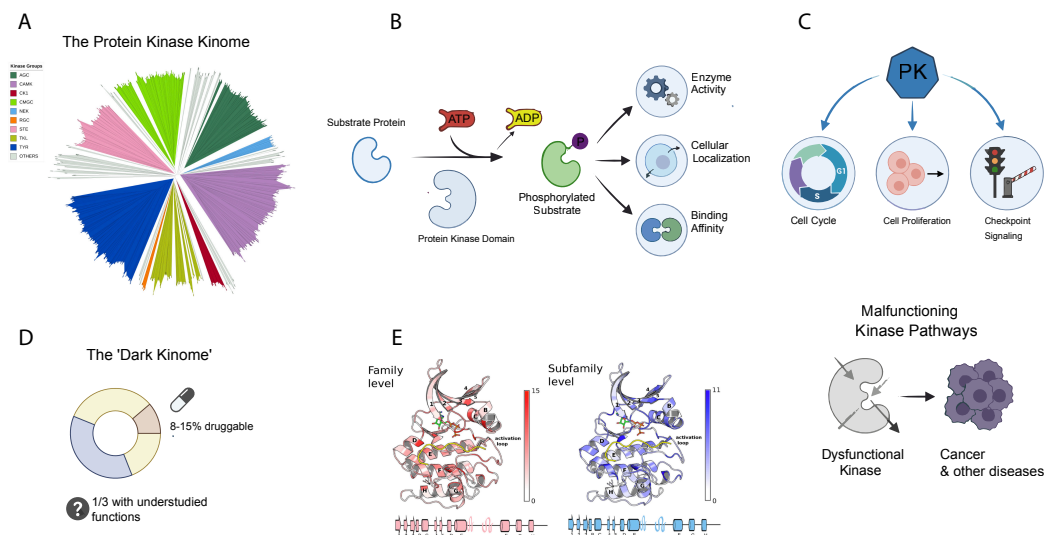


Figure 1.1: Overview of human protein kinase families, functions, disease roles, and understudied landscape. A) Overview of human protein kinase families, adapted from (Modi & Dunbrack, 2019b) B) Diverse biological functions regulated by protein kinase in phosphorylation-dependent manner C) Biological processes regulated by protein kinases and diseases involving malfunctioning kinases D) Overview of the understudied dark kinome E) Evolutionarily Divergent residues on kinase family and subfamily levels implied as specificity determinant candidates, adapted from (Bradley et al., 2021)

maintain signaling fidelity. Computational approaches have succeeded in predicting phosphosite occurrence from sequence context alone (Dou et al., 2014); however, identifying which kinase is responsible for each phosphorylation event remains a far more challenging problem. Although scaffold proteins and subcellular localization contribute to selectivity, the primary molecular determinants reside in an interface between two elements: the approximately 250-residue catalytic domain of the kinase, and the approximately 7–15 amino acid sequence flanking the phosphosite on the substrate (Amanchy et al., 2007).

A key gap in our knowledge is a mechanistic understanding of how kinase domains contribute to substrate specificity. While physical co-localization of kinases with their substrates—often mediated by interactions between peptide-binding domains and short linear interaction motifs—is a key component of substrate specificity (W. T. Miller, 2003), a growing body of work demonstrates that the kinase domain itself makes a significant contribution to substrate recognition (Ubersax & Ferrell Jr, 2007). Domain-swapping experiments on receptor tyrosine kinases have shown

that cells expressing chimeric receptors develop phosphorylation profiles consistent with the substrate preference of the kinase domain rather than the base receptor (Cai et al., 2017; Duplay et al., 1996; Nagao et al., 2021; Nakamura et al., 1996; Piccinini et al., 2002; Santoro et al., 1994). Complementing this work are studies on RET (Songyang et al., 1995), CAMK2 (Özden et al., 2022), c-Src (Cuesta-Hernández et al., 2023; N. H. Shah et al., 2018), PKA (Lubner et al., 2017; Walker et al., 2019), and ERK (Torres Robles et al., 2025), which show that single-point mutations in kinase domains can change their substrate preferences and, in some cases, drive pathogenicity (Lubner et al., 2017; Walker et al., 2019).

1.2.2 Evolutionary Divergence of Specificity-Determining Residues

While classical protein kinases share a comparable bilobal structural framework, they exhibit striking variation in the charge, hydrophobicity, and geometry of the surface residues lining the substrate-binding cleft. This observation, in combination with evolutionary analyses of tyrosine kinases, supports the hypothesis that kinase domains originated from a promiscuous common ancestor and evolved to develop distinct and diverse substrate preferences (Finneran et al., 2020). A paradigmatic example of how individual residues encode substrate discrimination is provided by CaMKII: a single substitution in the kinase domain is sufficient to alter selectivity at the substrate +2 position (Castro-Rodrigues et al., 2018). Disease mutations reinforce this principle at a pathological level. LRRK2 G2019S—one of the most common genetic causes of familial Parkinson’s disease—lies in the DXG ψ motif of the LRRK2 kinase activation loop and stabilizes the DYG-in active conformation, locking the enzyme in a hyperactive state with substantially increased phosphorylation of Rab GTPase substrates (Schmidt et al., 2021). Critically, this single glycine-to-serine substitution not only elevates total catalytic output but also alters the relative phosphorylation efficiency across substrates, illustrating that kinase domain mutations can qualitatively rewire substrate selectivity rather than merely scaling overall activity. Despite the biological and clinical importance of these examples, the effects of missense mutations in kinase catalytic domains on kinase–substrate phosphorylation activity remain characterized only individually, with evidence gathered sporadically across the literature rather than in any systematic framework.

Analysis of phosphorylation data across eukaryotic species reveals that most serine/threonine phosphorylation motifs are broadly present within eukaryotes but absent from prokaryotic phosphoproteomes, underscoring that the substrate recognition grammar of the eukaryotic protein kinase (EPK) superfamily is a eukaryote-

specific evolution (Bradley & Beltrao, 2019). This contrast is sharpest when comparing EPKs with prokaryotic eukaryotic-like kinases (ELKs). ELKs share the bilobal catalytic fold and the conserved DFG/HRD catalytic motifs with their eukaryotic counterparts yet lack the dynamic activation segment and GHI helical subdomain that evolved in EPKs as a platform for allosteric regulation and substrate docking. Consequently, prokaryotic ELKs phosphorylate chemically distinct substrates (Taylor & Kornev, 2011). The existence of a shared catalytic fold that nevertheless supports entirely distinct substrate selectivity illustrates how much of the divergence kinase allosteric landscape remains to be decoded.

1.3 Limitations of Existing Experimental Approaches to Characterize the Phospho-Interactome

Kinases and their target substrates have for the past decades been investigated individually in specific disease and biological pathway contexts. As a result, interactions in literature-curated databases are heavily biased, with many kinases having fewer than ten known substrates (Hornbeck et al., 2015). The major experimental approaches to kinase–substrate interaction (KSI) profiling have each made important contributions, but each carries characteristic limitations that leave the problem currently unsolved (Figure 1.2).

The structural basis for kinase specificity is difficult to interrogate because only approximately 20 co-crystal structures of protein kinases bound to a substrate peptide have been solved (Bradley et al., 2021). This scarcity of kinase–substrate co-complex structures due to the transient nature of phosphorylation events has impeded efforts to derive the principles governing enzyme–substrate recognition from structural data and stands in stark contrast to the hundreds of kinase structures solved in complex with ATP-competitive inhibitors. Even closely related kinases can engage substrates through entirely different binding modes (Bullock et al., 2009), yet the lack of diverse samples of kinase-substrate co-crystal structures limits the direct understanding of substrate recognition mechanism with dynamic and comparative structural evidence.

A major advancement toward understanding kinase selectivity arrived with combinatorial positional scanning peptide libraries. In this approach, a purified recombinant kinase is incubated with ATP and approximately 2.5 billion randomized peptides that systematically substitute each of 20 amino acids at each flanking position. The extent of phosphorylation at each substituted position is detected by phosphoim-

ager, yielding a linear consensus motif describing preferred flanking residues for the screened kinase (Johnson et al., 2023). This strategy was recently applied at atlas scale across the human serine/threonine kinome, profiling 303 kinases to produce the most comprehensive substrate specificity matrix available (Johnson et al., 2023), and subsequently extended to the tyrosine kinome across 84 human tyrosine kinases (Yaron-Barir et al., 2024). However, consensus motifs are at times an oversimplification of kinase specificity and in vivo phosphorylation environment with important practical consequences. The assay is performed cell-free on degenerate synthetic peptides and inherently cannot render sequence-level resolution. Moreover, the mere existence of a consensus phosphorylation site within a protein does not provide an assurance that the protein functions as a substrate within a living organism. For example, past research has identified that 123 proteins with the presence of a full consensus site is not sufficient for Cdk phosphorylation (Ubersax et al., 2003). Vice versa, it is important to note that genuine phosphorylation sites may not always comply with the known consensus. In a study where human tyrosine kinases were profiled against yeast proteome, it was realized that human non-receptor tyrosine kinases are capable of phosphorylating a diverse range of substrates not following the previously defined consensus motifs (Corwin et al., 2017). Some kinases also exhibit multiple distinct selectivity modes that a single consensus motif cannot capture (Mok et al., 2010). Motif-based approaches therefore describe biochemical permissibility rather than true biological interaction. Furthermore, these methods are unable to assess kinase variants in parallel, making it challenging to study the impact of variants on substrate specificity (Johnson et al., 2023; N. H. Shah et al., 2018; Songyang et al., 1994; Yaron-Barir et al., 2024).

Mass spectrometry-based phosphoproteomics provides the complementary, cell-endogenous view of the phosphorylation landscape but faces the opposite limitation: it reveals which sites are phosphorylated without identifying the responsible kinases. Quantitative phosphoproteomics after genetic or pharmacological kinase perturbation can nominate kinase–substrate links indirectly, but identifying which specific kinase is responsible for a given phosphosite remains a challenge and the throughput of screening is limited to the diversity of one cell-type specific peptides and one kinase perturbation at a time (Nováček et al., 2020). Thus, the two most powerful existing approaches—positional scanning for kinase preferences and phosphoproteomics for site discovery—are complementary but non-intersecting: neither alone, nor both combined, reliably profiles which kinase phosphorylates which phosphopeptide with peptide sequence level resolution.

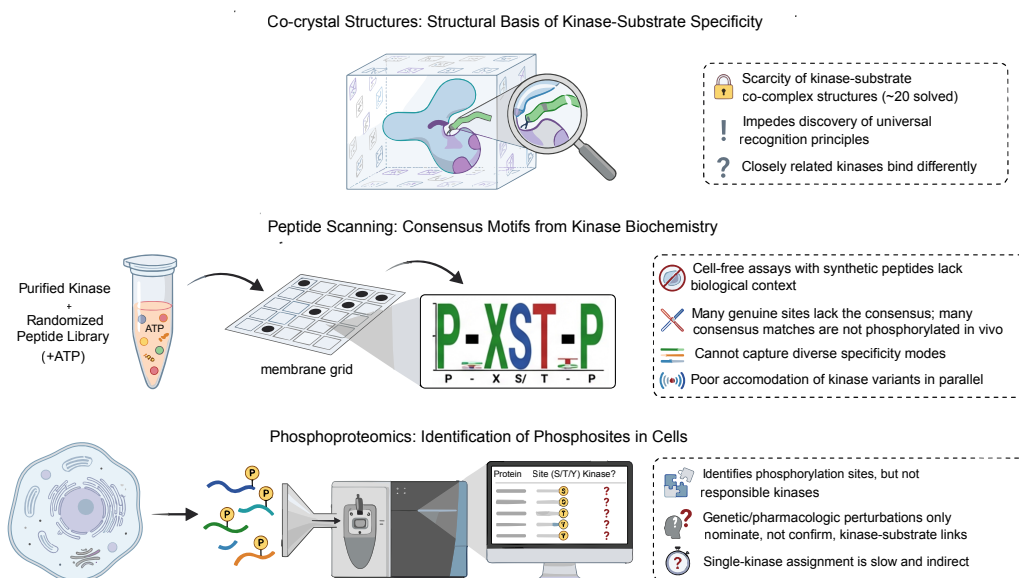


Figure 1.2: **Current experimental limitations of kinase-substrate specificity studies.** A) The scarcity of co-crystal structure with active kinase and substrates. B) Positional scanning results in potentially over-generalized peptide motifs. C) Phospho-proteomics fails to connect kinase to phosphosites.

A third and underappreciated gap is the absence of a scalable, simultaneous, pooled screening strategy capable of profiling large numbers of substrate sequences against multiple kinases in a single experiment. Current specificity profiling is primarily one-kinase-at-a-time: positional scanning, bacterial surface-display peptide libraries (N. H. Shah et al., 2018), and yeast-display platforms each profile a single kinase per experiment, making comprehensive kinase–substrate pairs mapping intractable. Yeast-based human kinase expression arrays represent a significant scaling effort. Breitkreutz et al. (2010) established a resource covering 87 yeast and human kinases for phenotypic readouts, and Jehle et al. (2022) extended this to 266 human kinases in a chemical-genetic platform. However, these array-based approaches test each kinase independently in separate wells and do not resolve individual peptide-level KSI scores for large substrate sets in a pooled format. No current experimental platform supports simultaneous profiling of 1,000 or more substrate sequences against 10 or more kinases variants in a single pooled assay with quantitative, per-substrate KSI resolution. This gap in multiplexing throughput directly limits the training data available for computational predictors and the discovery rate for dark kinase substrates (Figure 1.3).

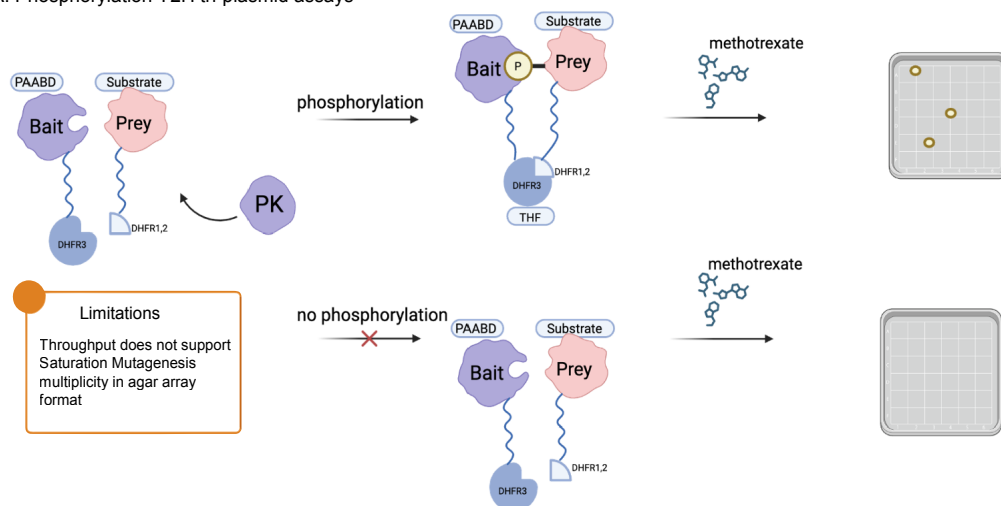
Deep mutational scanning is a technology well suited for understanding the impact

of variants on protein function (Fowler & Fields, 2014; Wei & Li, 2023). In this approach, the impact of thousands of mutations is measured in parallel by coupling protein activity to readable output such as cell growth. This coupling enables quantification using next-generation sequencing to measure the desired properties of protein variants. Deep mutational scanning has been applied to various systems, including viral fitness (Hom et al., 2019; Sourisseau et al., 2019; Starr et al., 2020), photosynthetic enzymes (Prywes et al., 2025), G-protein-coupled receptors (M. K. Howard et al., 2025; Jones et al., 2020), protein–protein interactions (Faure et al., 2022; Weng et al., 2024), as well as kinases such as ABL (Sokirniy et al., 2025), MET (Estevam et al., 2025), EGFR (Hayes et al., 2024), Syk (Hobbs et al., 2022), ERK2 (Brenan et al., 2016), and c-Src (Ahler et al., 2019; Beltran et al., 2024). Although deep mutational scanning is powerful, prior assays are unable to distinguish between increased protein abundance and solubility from kinase activity (Ahler et al., 2019; Hobbs et al., 2022). Two recent preprints suggest that performing separate deep mutational scans for protein activity and abundance can begin to address this limitation (Beltran et al., 2024; C. J. Howard et al., 2025); however, these studies measure overall kinase activity by linking higher kinase activity to increased cytotoxicity from disrupted phosphorylation pathways and are unable to reveal changes in substrate preferences among kinase variants.

1.4 Gaps in Computational Models for the Phospho-Interactome Prediction

Numerous computational tools attempt to predict kinase–substrate interactions. Early approaches such as Scansite employed position-specific scoring matrices (PSSMs) derived from consensus motifs (Obenauer, 2003). While useful, PSSM-based scoring collapses complex recognition into linear residue weights that tend to be overly generalized. Without taking in kinase sequence information, these tools cannot transfer specificity knowledge between kinases. Later methods incorporated interaction databases and machine learning but remained restricted to well-studied kinases with sufficient training examples. Most predictors cover only tens to low hundreds of kinases out of more than 500 (Horn et al., 2014; D. Wang et al., 2017; M. Wang et al., 2017). Knowledge-graph approaches such as LinkPhinder expanded coverage to over 300 kinases through relational learning (Nováček et al., 2020), yet these approaches rely on existing curated interactions and therefore inherit the same literature imbalance of training data. New model architecture is needed: kinase–substrate interaction is a pairwise sequence recognition problem, yet most existing tools either ignore kinase sequence or treat substrate peptides independently

A. Phosphorylation Y2H tri-plasmid assays



B. Deep Mutational Scanning Pipelines

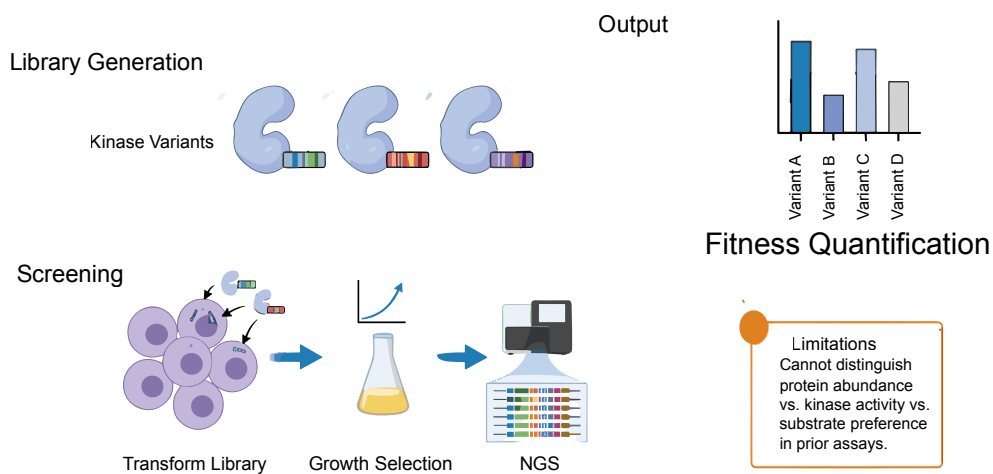


Figure 1.3: Survey of emerging high-throughput phosphorylation characterization of kinase-substrate specificity studies. A) Phospho-Y2H (yeast two-hybrid) pipeline illustration. B) Deep mutational scanning for kinase domain illustration.

of kinase identity. Consequently, they cannot transfer specificity knowledge from well-studied kinases to understudied ones, nor can they generalize to novel biological contexts with novel substrate sequences such as viral infection phosphoproteomics or disease-associated phosphosite variants.

Attempts in using protein language models to predict protein-protein interaction events emerge with the transformer models being more accessible. Protein amino acid sequences are sequential data governed by the physical and evolutionary constraints of protein folding. Two amino acids distant in the one-dimensional sequence may interact strongly in the folded three-dimensional structure. Transformer architectures with multi-headed self-attention learn these long-range dependencies in a parallel and unsupervised manner (Vaswani et al., 2017), making them well suited to protein sequence representation. Pioneering protein language models include UniRep (Alley et al., 2019) and SeqVec (Heinzinger et al., 2019). Subsequent large-scale transformer models such as ESM-2, trained on hundreds of millions of protein sequences, capture structural and functional information without explicit supervision and outperform recurrent architectures on diverse downstream tasks (R. Rao et al., 2019; Rives et al., 2021). Transformer models can represent interactions between two biological sequences by concatenating them into a single input, analogous to question–answer pair modeling in natural language processing. This approach has successfully predicted MHC–peptide binding by treating the MHC allele and peptide as paired input sequences, outperforming state-of-the-art single-sequence tools (Cheng et al., 2021). The same principle applies directly to kinase–substrate interaction, where the kinase catalytic domain and substrate peptide can be jointly encoded, enabling the model to learn pairwise specificity rather than single-sequence properties.

Two recent models illustrate the state of the art and its architectural trade-offs. Phosformer (Z. Zhou et al., 2023) accepts an unaligned kinase catalytic domain sequence paired with a 13-residue substrate peptide centered on the phosphosite. ESM-2 embeddings are extracted independently for the kinase and the peptide, then concatenated and passed through a lightweight transformer classification head that predicts the interaction probability. KolossuS (Jha et al., 2025) adopts a co-embedding architecture that represents both the full-length kinase sequence and a 15-mer substrate peptide in a shared latent space via a multilayer perceptron (MLP) projection layer. Representation embeddings are extracted from ESM-2 at up to 15 billion parameters; average-pooled kinase and substrate embeddings are

then nonlinearly transformed into a shared embedding space in which interacting kinase–substrate pairs are trained to be geometrically close. The model was trained on both the serine/threonine and tyrosine kinome atlas datasets, enabling cross-kinase transfer of specificity information.

Despite these architectural advances, all current protein language model–based KSI predictors training data are dominated by the same well-characterized kinase families that have accumulated the most research documentation over decades such as AGC, CMGC, CAMK, and select tyrosine kinase families. Strong performance on held-out test sets sampled from the same distribution does not guarantee genuine generalization to dark kinases or novel biological contexts. There is a risk that these models learn the statistical regularities of curated biological signaling networks—such as which substrates co-occur with which kinase families in the literature and which motifs are represented in PhosphoSitePlus—rather than the underlying biophysical rules that govern how a kinase catalytic domain recognizes and coordinates a peptide for phosphoryl transfer. A model that has effectively captured the connectivity of well-annotated pathways will perform well when asked to score substrates abiding to the injected biology pathway biases, but will extrapolate poorly when asked to predict substrates for a dark kinase whose signaling context is entirely uncharacterized, or for novel substrates that do not conform to known and well-populated substrates in training data. Benchmark results from KolossuS comparing performance across kinase families with different annotation densities show that prediction performance is markedly lower for kinase families with sparse PhosphoSitePlus representation, consistent with this concern (Jha et al., 2025). Architectural choices that promote learning of biophysical recognition rules, including training on diverse in vitro biochemical datasets, hard negative sampling based on kinase family annotations, and penalization of signaling-context associations using unbiased, cross-kinase screening, are necessary to produce predictors that genuinely extend into the dark kinome.

1.5 Biological Motivation for Accurate Kinase–Substrate Specificity Profiling

1.5.1 Mapping the Phospho-Interactome as a Prerequisite for Engineering Live-Cell Kinase Biosensors

Understanding kinase specificity not only will benefit the understanding of structural determinants on the kinase domain, but also suggesting high-quality substrates for biosensor engineering. Live-cell kinase activity biosensors depend on identifying a substrate peptide that is selectively phosphorylated by a given kinase,

paired with a phospho-binding domain that recognizes that modification (Ni et al., 2018). Several reporter architectures have been developed. FRET-based reporters use donor–acceptor fluorescent protein pairs flanking a phosphorylatable substrate and a phosphopeptide-binding domain (PBD); phosphorylation drives intramolecular conformational change and produces a ratiometric FRET signal (Greenwald et al., 2018; J. Zhang et al., 2001). Kinase Translocation Reporters (KTRs) couple phosphorylation to nucleo-cytoplasmic redistribution of a fluorescent protein, providing a digital-like spatial readout (Regot et al., 2014). Phase-separation-based SPARK sensors leverage multivalent interactions to amplify the phosphorylation signal into bright condensate puncta (Q. Zhang et al., 2018) (Figure 1.4). Single-fluorophore Exrai reporters based on circularly permuted EGFP (cpEGFP) achieve greater dynamic range than classical FRET by directly coupling phosphorylation-induced conformational change to the chromophore environment (Schmitt et al., 2022; J.-F. Zhang et al., 2021). Unlike bulk phosphoproteomics, biosensors probe kinase activity in a natural cellular environment with single-cell, real-time spatiotemporal resolution, deconvoluting heterogeneity masked by population-average measurements. Despite advances in reporter chemistry, biosensor development has concentrated on a small set of well-characterized kinases such as PKA, ERK, CaMKII, and AMPK. The primary limiting factor is not the reporter architecture but substrate identification: without a selective, validated substrate peptide and a corresponding phospho-binding domain that recognizes the phosphorylated peptide, no reporter can be built. As a consequence, kinase biosensors have been iterating for better signal-to-noise ratio only the same set of kinases but remain experimentally inaccessible for most of the kinases without a compatible substrate.

1.5.2 Phosphorylation Rewiring Events in Human Diseases

Large-scale genomic sequencing has revealed that human genetic variation frequently perturbs post-translational regulation including phosphorylation sites. Missense variants are enriched near phosphorylation sites and protein interaction interfaces across diverse disease contexts. Reimand et al. (2013) systematically analyzed cancer genomes and demonstrated that somatic mutations are significantly enriched at and around phosphorylation sites in more than 89% of tumors across 12 cancer types. Phosphorylation-altering single nucleotide variants (pSNVs) were found to be more functionally deleterious than mutations at non-phosphorylation sites. The disease relevance of phosphorylation rewiring extends beyond cancers. In Alzheimer’s disease, tau hyperphosphorylation at more than 40 serine/threonine

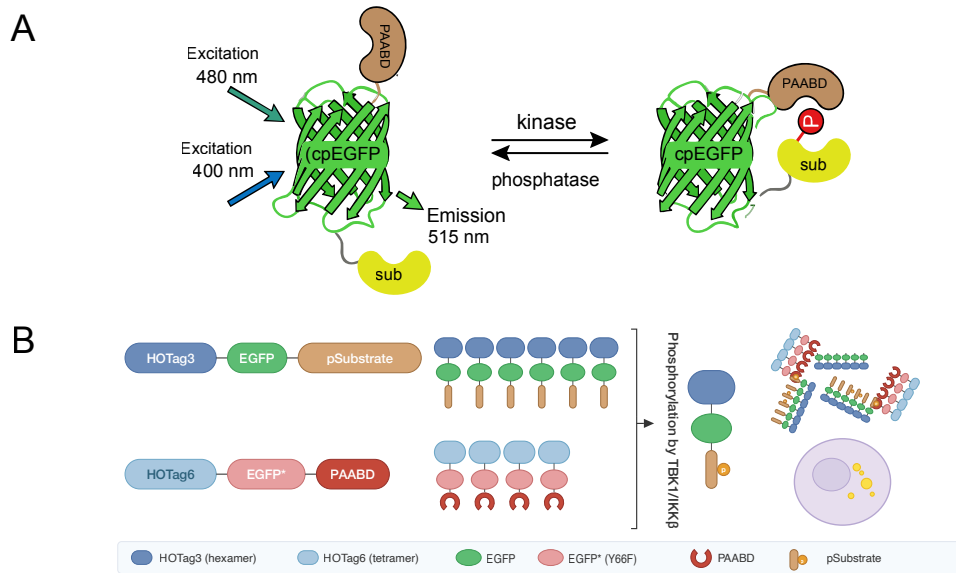


Figure 1.4: **Survey of live-cell kinase biosensor designs.** A) Ratiometric single-fluorescent protein ExRai Design, illustration inspired by (J.-F. Zhang et al., 2021). B) Phase-separation based SPARK reporter design, illustration inspired by (Q. Zhang et al., 2018).

sites—driven largely by CDK5 and GSK-3 β —promotes dissociation from microtubules and the formation of paired helical filaments and neurofibrillary tangles. Phosphorylated tau at Thr231, Ser396, and Ser404 serves as diagnostic biomarker signatures of active disease (Tönnies & Trushina, 2017). In Parkinson’s disease, α -synuclein phosphorylation at Ser129 marks more than 90% of aggregated Lewy body inclusions, directly implicating aberrant kinase activity in aggregate seeding (Tenreiro et al., 2014). Age-dependent upregulation of CDK5 and GSK-3 β activity in the aging mouse brain—detectable by quantitative phosphoproteomics—recapitulates multiple Alzheimer’s-associated phosphorylation signatures and correlates with progressive neuronal dysfunction (Mohallem et al., 2024). In cardiac disease, hypertrophic cardiomyopathy mutations in sarcomeric proteins frequently map to phosphorylation sites normally targeted by PKA and CaMKII during β -adrenergic contractile regulation. Phosphorylation-altering variants at these sites disrupt the phosphorylation-dependent modulation of myosin cross-bridge cycling (Barefield & Sadayappan, 2010). In congenital and developmental disorders, de novo variants in kinase-encoding genes converge on synaptic phosphorylation networks. DYRK1A haploinsufficiency causes intellectual disability and microcephaly via altered phosphorylation of splicing regulators and cytoskeletal proteins, while

CDK5R1 (p35) mutations associated with lissencephaly disrupt CDK5-mediated phosphorylation of substrates critical for cortical neuron migration (Graham et al., 2004; J. Ji et al., 2015). In autoimmune and inflammatory disease, gain-of-function variants in kinase-encoding genes have been identified directly in patients. A striking example is provided by SYK, where monoallelic gain-of-function variants cause multi-organ inflammatory disease including colitis and arthritis (L. Wang et al., 2021). The pattern is reoccurring: dysregulated kinase phosphorylation of immune signaling substrates is a recurrent driver of autoimmune pathology, yet for the vast majority of identified variants the specific kinase–substrate pair responsible for the phenotype remains unassigned. This prevents pSNVs from being translated into mechanistic hypotheses or therapeutic targets.

1.5.3 Viral Hijacking of Host Kinase Networks

Viruses extensively reprogram host kinase signaling to create a permissive cellular environment for replication. RNA viruses relies on co-opted host kinases to phosphorylate viral proteins for nuclear import, particle assembly, and immune evasion, and to remodel host signaling in ways that suppress antiviral defense. In SARS-CoV-2 infection, phosphoproteomics of infected Vero E6 cells revealed dramatic remodeling of host phosphorylation, with CK2 and p38 MAPK identified as major activated kinases (Bouhaddou et al., 2020). In DNA virus Epstein–Barr virus, the viral kinase BGLF4 exhibits CDK1-like substrate preferences and drives widespread phosphorylation changes linked to the DNA damage response and mitotic signaling, while inhibition of mitotic checkpoint signaling suppresses extracellular virus production (R. Li et al., 2015). Similarly, human papillomavirus (HPV) E6 physically interacts with host Aurora kinase B and alters its activity (Boon et al., 2023; Guo et al., 2019). Surveying across the virus infection events suggests that many hijacked kinases remain under-characterized at the substrate level. Accurate kinase–substrate prediction would potentially leverage the drug re-purposing strategy to redirect well-characterized and FDA-approved kinase inhibitors to host-directed antiviral therapeutics.

Chapter 2

PROTEIN LANGUAGE MODEL ENABLES GENERALIZABLE KINASE-SUBSTRATE INTERACTION (KSI) PREDICTION

2.1 Introduction

As thoroughly discussed in Chapter 1, protein phosphorylation, catalyzed by the approximately 500 kinases of the human kinome, governs virtually every aspect of cellular biology, from proliferation and differentiation to immune surveillance and DNA-damage repair. Kinase–substrate interaction (KSI) prediction remains challenging: despite decades of experimental effort, the interaction space is sparsely characterized and coverage of non-human kinases is especially limited. Existing computational methods are constrained by training-data scarcity – motif-based approaches are restricted to families with ≥ 50 –100 annotated sites, instance-based methods cover at most low hundreds of kinases under minimum-site thresholds, and recent deep-learning methods such as Phosformer (Z. Zhou et al., 2023) and DeepKinZero (Deznabi et al., 2020) still fall short in breadth, cross-species transfer, and generalize to poorly annotated kinases (Table 2.2).

We developed KINBERT, a transformer-based protein language model that learns kinase–substrate recognition directly from paired kinase domains and substrate peptide sequences to address these limitations. The model is trained comprehensively on validated KSIs from four complementary databases, expanding training coverage to 2,304 kinases and 92 organisms (Table 2.1)(Table 2.3). It employs a family-aware negative-sampling scheme that creates informative contrastive examples without introducing ambiguity from promiscuous phosphosites. Our benchmarking experiments reveal the optimized model architecture for KSI prediction requires backbone end-to-end with a separator-token architecture that enables joint kinase–substrate encoding. Two proof-of-concept downstream applications further demonstrate the utility of KINBERT as an efficient inference engine for human disease variant interpretation and viral host-kinase identification.

2.2 Results

2.2.1 Comprehensive Database Annotation Integration

KSIs were aggregated from four experimentally curated databases (DeepKinZero (Deznabi et al., 2020), PhosphoSitePlus (Hornbeck et al., 2015), phosphoELM

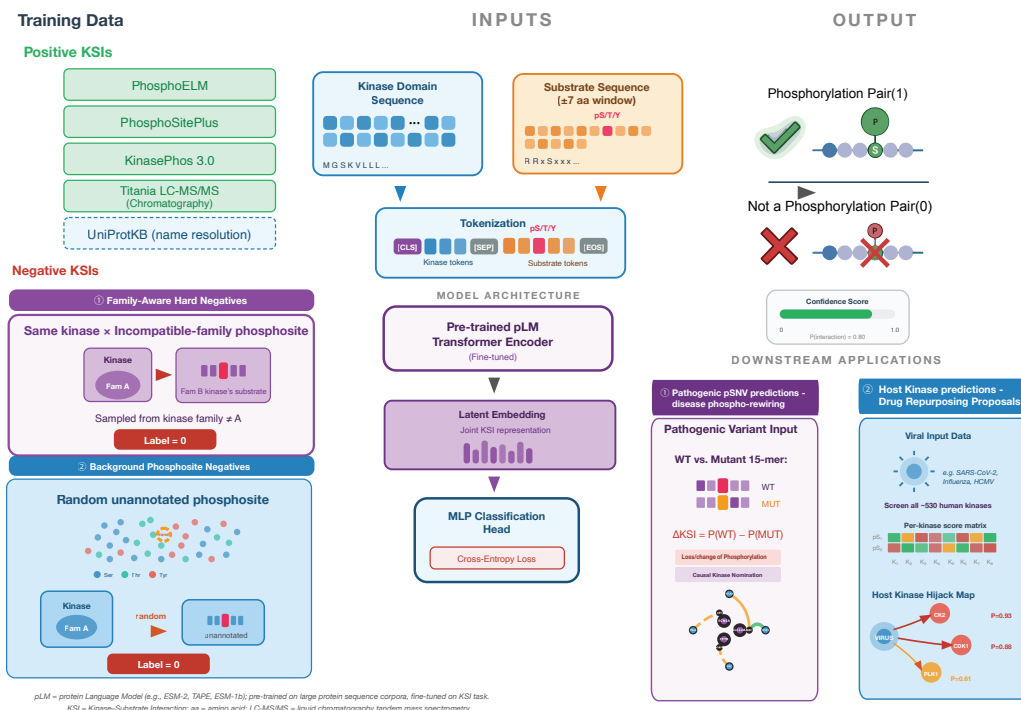


Figure 2.1: Overview of the data collection, model architecture, and downstream application of KINBERT Left: Data collection and negative sample augmentation Middle: Training and model architecture schematics Right: Application of KINBERT to virus drug-repurposing and phosphorylation rewiring predictions

(Diella et al., 2008), and KinasePhos 3.0 (Ma et al., 2023)), standardized to UniProt accessions, and filtered to retain only kinases with a single unambiguous catalytic domain and a resolvable family classification. Phosphosites targeted by more than three kinase families were excluded to reduce promiscuity-driven ambiguity, yielding a class-annotated interaction set of kinase–substrate–phosphosite triples (Figure 2.2). Training negatives were generated using a two-component strategy: family-aware hard negatives, pairing each kinase with residue-matched phosphosites from an incompatible kinase family context, and background negatives drawn from unannotated S/T/Y flanking sequences, together producing a balanced training dataset (Figure 2.1). This mixed negative sampling strategy exposes the model to both fine-grained specificity contrasts and broader non-interacting phosphosite backgrounds, improving generalization across kinase families. On top of this DB Only training set, we ingested further a titania LC-MS/MS characterized dataset where dephosphorylated HeLa cell lysates are used as substrate pools to react with recombinant kinases *in vitro* and results in the more diverse DB+Lysate dataset.

Table 2.1: Comparison of training-data scope across KSI prediction methods.

Model	# of Kinases	# of Organisms	# of KSIs
DeepKinZero	343	1	13,426
Phosformer	800	13	24,534
KINBERT	2,304	92	147,488
KolossuS	Training data statistics not available*		

* KolossuS training set composition was not reported in the preprint (Jha et al., 2025); statistics could not be independently verified.

Table 2.2: Kinase-family and individual-kinase coverage in computational KSI prediction methods.

Method	Number of Kinases or Kinase Families	Criteria for Inclusion
MusiteDeep (D. Wang et al., 2017)	5 families	Families with >100 sites
PhosphoPredict (Song et al., 2017)	8 families	Families with ≥ 50 sites
PhosphoPICK (Patrick et al., 2015)	59 human kinases	Kinases with >10 sites
PKIS (Zou et al., 2013)	56 human kinases	Kinases with >10 sites
KSRPred (M. Wang et al., 2017)	103 human kinases	Kinases with ≥ 15 sites
KinomeXplorer (Horn et al., 2014)	222 kinases covered; accuracy assessed for 14	Kinases with ≥ 20 sites

Table 2.3: Distribution of interactions per kinase for database-curated (DB Only) and combined database plus Mass-Spec (DB + Lysate) sources.

Source	1–5	5–20	20–50	50–100	100–200	>200
DB + Lysate	1,129	290	79	63	66	189
DB Only	1,211	375	113	44	24	23

2.2.2 Model Architecture Optimization and Finalization

A test set comprising 10% of the database-curated-only interactions (DB Test) was withheld for evaluation on more literature-curated interactions, alongside a larger

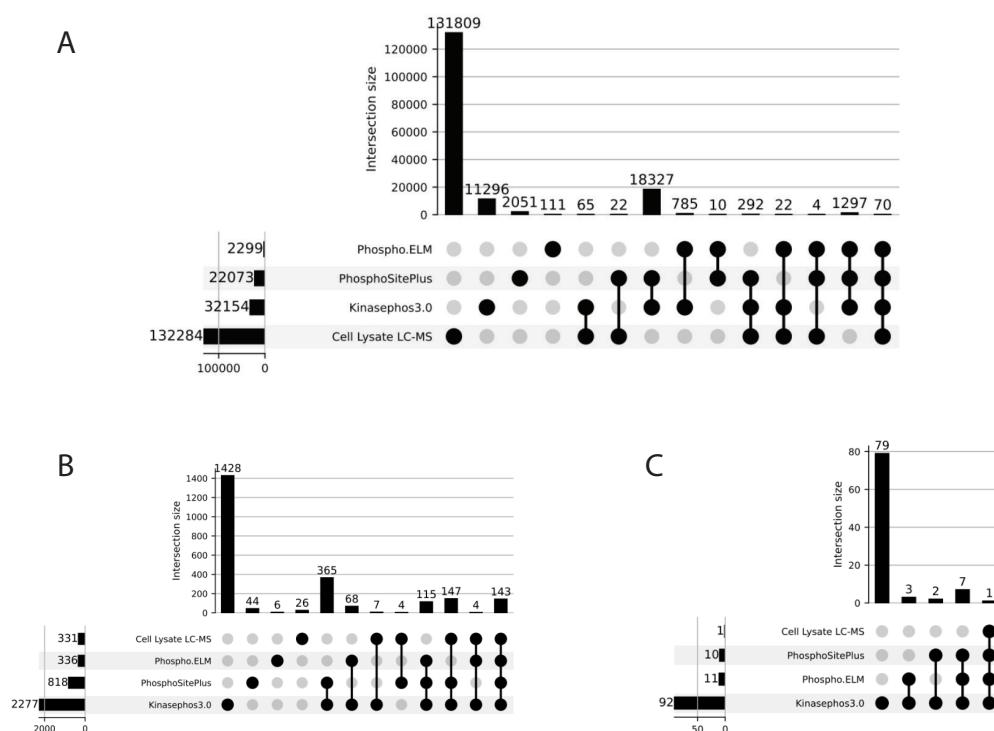


Figure 2.2: Details of compiled training corpus for KINBERT. A) Number of kinase-substrate pairs that overlap between the four ingested stand-alone phospho-interactome sources B) Number of unique kinases that overlap between the four ingested stand-alone phospho-interactome sources C) Number of unique organisms that overlap between the four ingested stand-alone phospho-interactome sources

test set incorporating both database-curated and lysate-based interactions (DB + Lysate Test) that constitutes 10% of all KSI interactions. Between the two external comparators, KolossuS performs better on the DB + Lysate test than Phosformer, which is expected given that KolossuS ingested LC-MS lysate interactions during training. Conversely, Phosformer performs better on the DB Test, consistent with its training being restricted to PhosphoSitePlus interactions.

KINBERT:TAPE models outperformed both Phosformer and KolossuS on both test sets (Figure 2.3), as well as on the 15 aa-mapped versions of Phosformer's training and test sets (Phosformer was originally trained on 11-residue windows; these were mapped back to the original sequence and extended to 15 aa for benchmarking). Within the exploration of KINBERT architectures, having a fine-tunable backbone that enables self-attention between kinase and substrate in the transformer encoder

layers via a separator token produced the best performance on both test sets. A longer 15-residue substrate flanking region and an expanded source of interactions were adopted compared to Phosformer; in contrast to KolossuS, which uses full-length kinase sequences, KINBERT focuses on the kinase domain and substrate prediction, which may be more fitting for the nature of LC-MS cell-lysate characterized interactions.

KolossuS formulates kinase–substrate prediction as a metric learning problem, optimizing cosine similarity between independently encoded protein embeddings. While effective for representation learning, this objective is mismatched to the biology and evaluation of KSIs, which are asymmetric, sparse, and decision-centric rather than symmetric similarity relationships. In contrast, KINBERT directly optimizes interaction likelihood using binary cross-entropy, allowing the model to learn non-metric, asymmetric decision boundaries. Performance improves further when moving from frozen or cached representations to end-to-end fine-tuning with joint sequence encoding, where kinase and substrate are concatenated with an explicit separator token and processed as a single input. This formulation enables token-level cross-conditioning throughout the transformer stack, allowing the backbone to learn pair-specific compatibility features rather than relying on post-hoc embedding comparison. Among backbones, TAPE consistently outperformed ESM-t12 (Rives et al., 2021) under this regime (Table 2.4). This is attributed to TAPE’s pre-training with explicit segment semantics (CLS/SEP tokens) and its greater fine-tuning flexibility with less-conditioned pre-training, which make separator-based joint encoding immediately meaningful and easier to adapt for discriminative pairwise reasoning. By contrast, ESM models are pre-trained almost exclusively on single-chain evolutionary objectives and must relearn boundary semantics and suppress global similarity biases during fine-tuning.

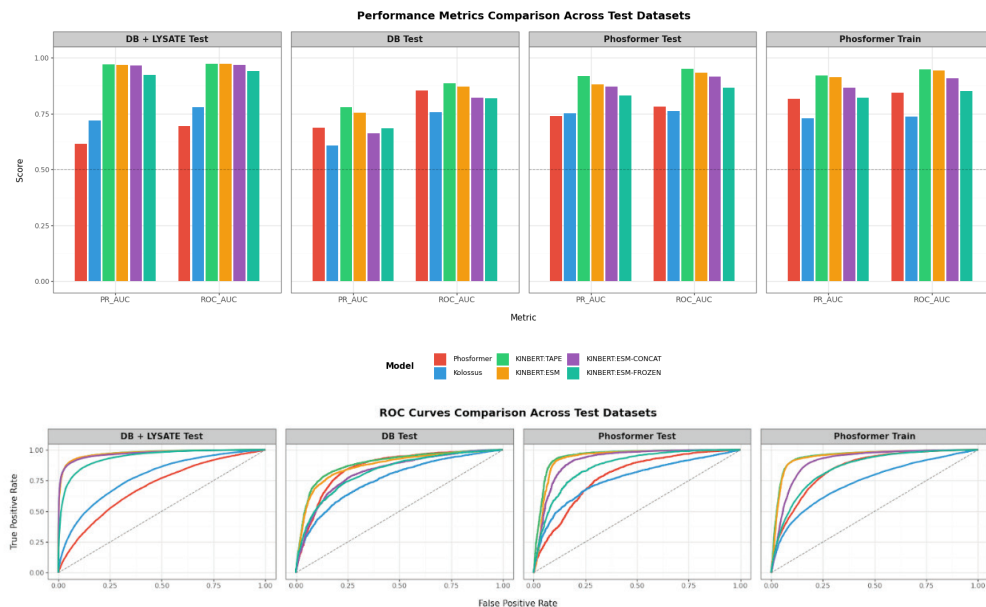


Figure 2.3: **Model architecture optimization and test data performance benchmarking.** Phosformer, KolossuS, and KINBERT with TAPE backbone, ESM-2 backbone, ESM-2 without separator token and ESM-2 with frozen encoder backbone were benchmarked against but Train/Test data of KINBERT and of Phosformer, KolossuS data is currently unavailable from the preprint

Table 2.4: Metrics: model architecture optimization and test data performance benchmarking.

Model	PR_AUC						ROC_AUC					
	Phosformer	KolossuS	KINBERT:TAPE	KINBERT:ESM	KINBERT:ESM-CONCAT	KINBERT:ESM-FROZEN	Phosformer	KolossuS	KINBERT:TAPE	KINBERT:ESM	KINBERT:ESM-CONCAT	KINBERT:ESM-FROZEN
DB + LYSATE Test	0.614	0.720	0.970	0.968	0.964	0.924	0.695	0.779	0.973	0.972	0.968	0.940
DB Test	0.688	0.607	0.780	0.754	0.662	0.686	0.854	0.758	0.887	0.870	0.821	0.819
Phosformer Test	0.719	0.751	0.919	0.882	0.871	0.831	0.782	0.762	0.950	0.934	0.916	0.865
Phosformer Train	0.816	0.729	0.921	0.914	0.866	0.822	0.845	0.737	0.947	0.943	0.908	0.851

2.2.3 Model Performance and Generalization Benchmarking Against Existing Methods

Across all benchmarks, we further evaluate on the performance of the KINBERT:TAPE model trained without access to the corresponding withheld sets, as this setting most directly tests out-of-distribution generalization. Under this strict evaluation, KINBERT:TAPE achieves the highest performance on the Dark Kinase Holdout, despite never observing any dark kinases during training, demonstrating that the model generalizes across kinase families rather than relying on kinase-

specific memorization. It should be noted that many of the dark kinases enriched in the test sets are characterized with documented substrates but remain classified as understudied due to limited functional characterization.

In the Non-Human Kinase Holdout, where all kinases characterized in non-human species are excluded from training, KINBERT:TAPE retains strong PR AUC and ROC AUC, demonstrating substantial cross-species transfer. Performance drops more than in the human dark-kinase holdout, reflecting the additional distribution shift introduced by evolutionary divergence and species-specific signaling contexts. While dark kinases are poorly annotated, they remain part of the human kinome and share conserved catalytic and regulatory features with well-studied kinases, facilitating generalization. In contrast, non-human kinase–substrate interactions differ in both kinase sequence composition and substrate landscape, introducing divergence in motif usage and regulatory constraints. Importantly, KINBERT:TAPE is trained exclusively on kinase domain sequences rather than full-length kinases, which likely contributes to improved transfer by focusing learning on evolutionarily conserved catalytic and substrate-recognition features while discarding species-specific regulatory regions and low-complexity tails. This design reduces reliance on organism-specific context and encourages the model to learn general biochemical compatibility principles.

The Yeast $\times$ Human PK benchmark represents an even more extreme departure from the training distribution. This dataset is based on experiments in which human non-receptor tyrosine kinases were individually expressed in *Saccharomyces cerevisiae*, and their tyrosine phosphorylation activity on the yeast proteome was profiled by LC-MS/MS, yielding thousands of kinase–substrate relationships on a substrate space evolutionarily distant from human signaling networks (Corwin et al., 2017). Because the yeast proteome lacks endogenous tyrosine kinase pathways and phosphotyrosine-binding modules, human kinases acting on this background generate phosphorylation events that are physicochemically permitted but not governed by native co-evolved constraints, compounding the cross-species challenge. Despite these difficulties and the minimal overlap with the training data distribution, KINBERT consistently outperforms both KolossuS and Phosformer, indicating that its predictions are driven by learned biochemical compatibility rather than reliance on organism-specific context or memorization of known interactions. This setting poses a particularly stringent challenge because tyrosine kinases often exhibit broad intrinsic substrate compatibility and require auxiliary regulatory domains (e.g., SH2/SH3

domains) to achieve context-specific substrate selection; when assayed only through their catalytic domains or in a non-native proteome lacking co-evolved pTyr signaling modules, predicting this specificity becomes inherently difficult. Studies of cytoplasmic tyrosine kinases have shown that non-catalytic regions such as SH2 and SH3 domains play important roles in substrate recruitment and localization, shaping substrate specificity beyond the intrinsic catalytic preferences of the kinase domain. Moreover, high-throughput profiling of human tyrosine kinases reveals considerable diversity in substrate sequence preferences, implying that accurate specificity prediction depends on both sequence context and higher-order regulatory interactions (Yaron et al., 2022).

Overall, the separator-augmented TAPE backbone yields the strongest and most consistent predictive performance across diverse and biologically distinct kinase–substrate interaction tasks (Figure 2.4) (Table 2.4).

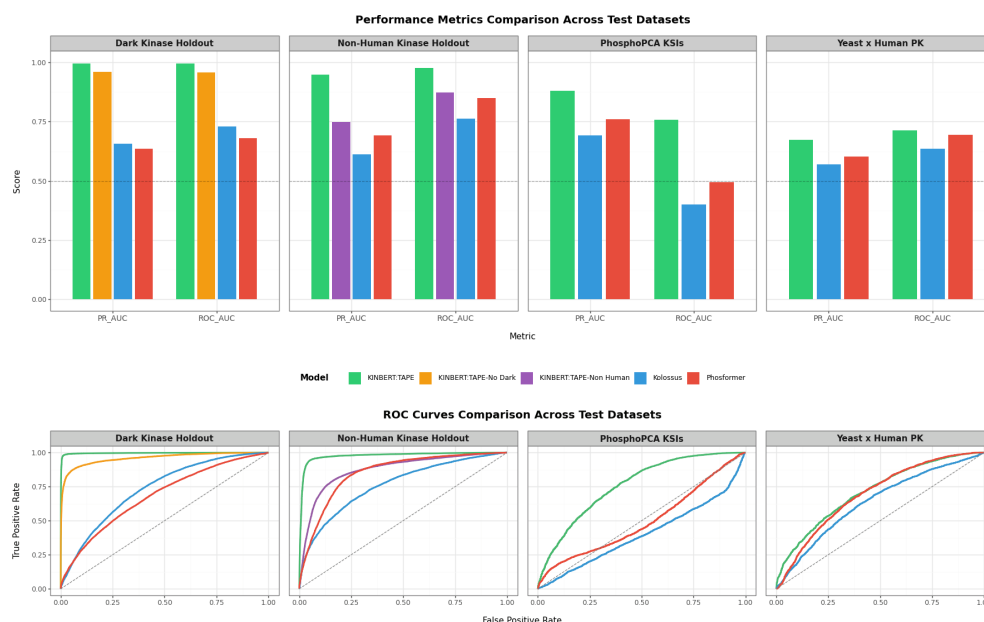


Figure 2.4: Generalizability benchmark across transformer-based KSI prediction models Phosformer, Kolossus, and KINBERT with TAPE backbone were benchmarked against dark kinase KSIs, non-human KSIs, PhosphoPCA-identified KSIs (detailed in Chapter 3), and human kinase reactivity in yeast proteome

Table 2.5: Metrics for generalizability benchmark across transformer-based KSI prediction models

Model	PR_AUC					ROC_AUC				
	KINBERT:TAPE	KINBERT:TAPE-No Dark	KINBERT:TAPE-Non Human	KolossuS	Phosformer	KINBERT:TAPE	KINBERT:TAPE-No Dark	KINBERT:TAPE-Non Human	KolossuS	Phosformer
Dark Kinase Holdout	0.996	0.961		0.658	0.635	0.996	0.959		0.730	0.681
Non-Human Kinase Holdout	0.949		0.750	0.612	0.693	0.976		0.874	0.762	0.851
PhosphoPCA KSIs	0.874			0.692	0.759	0.754			0.401	0.495
Yeast x Human PK	0.674			0.569	0.602	0.713			0.636	0.695

2.2.4 Downstream Application: Inference on Pathogenic PTM Human Variants for Elucidation of Disease-Related Phosphorylation

To connect pathogenic phosphosite variants to their disrupted upstream kinases, we screened missense variants annotated at phosphorylatable residues in PTMVar (Hornbeck et al., 2015) against the full human kinome using KINBERT. Variants were scored in both wild-type and mutant sequence contexts, and kinases showing a significant drop in prediction confidence upon mutation were flagged as disrupted candidates. High-confidence predictions were cross-validated against physical protein–protein interaction evidence from OmniPath (Türei et al., 2016), PhosphoNetworks (Hu et al., 2014), and STRING-DB (Szklarczyk et al., 2023), and retained only for variants with ClinVar-confirmed pathogenic or likely pathogenic classifications (Landrum et al., 2014; Sherry, 2001). Surviving hits were manually curated into evidence tiers based on the directness of experimental support for the specific kinase–site relationship (Figure 2.5). This pipeline yielded a disease-linked kinase–phosphosite network connecting pathogenic variants at phosphorylatable residues to their predicted disrupted upstream kinases and associated clinical phenotypes. We further prioritized variants with literature evidence and novelty with details as described in the Methods section.

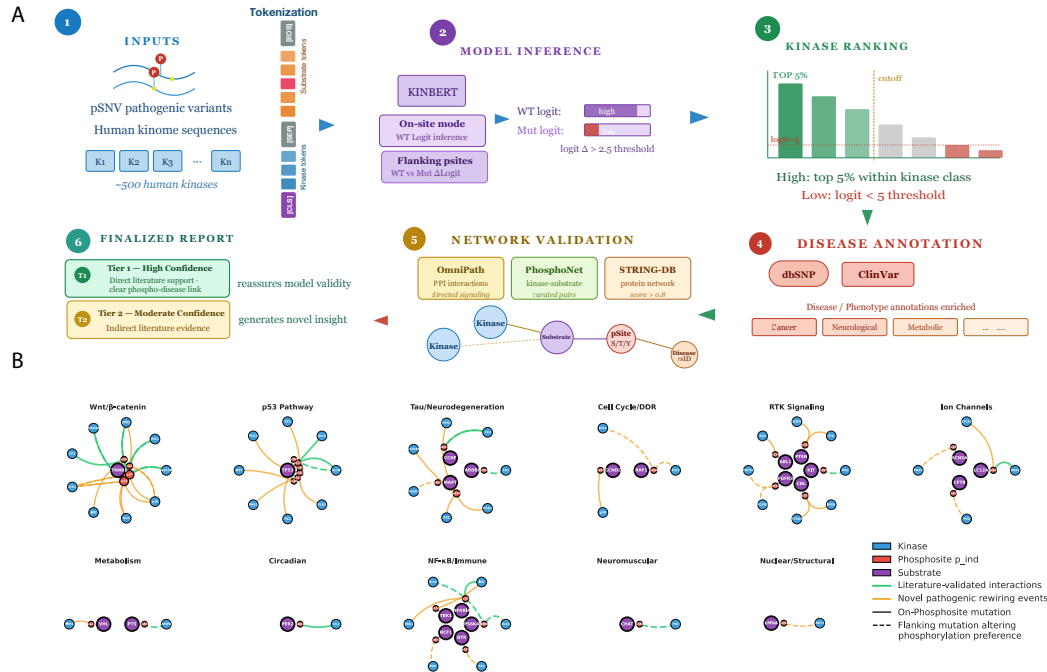


Figure 2.5: Phosphorylation-rewiring event predictions in disease SNV contexts

A) Pathogenic phosphosite variants from PTMVar were screened against the human kinome using KINBERT. Variants were evaluated in wild-type and mutant contexts, and kinases with reduced prediction confidence upon mutation were identified as disrupted candidates. High-confidence hits were filtered using protein–protein interaction evidence (OmniPath, PhosphoNetworks, STRING) and ClinVar pathogenicity annotations, then curated into evidence tiers based on experimental support. This pipeline generated a disease-linked kinase–phosphosite network and prioritized variants with literature support and novelty. B) KINBERT identifies disease-linked kinase–phosphosite disruptions stratified by evidence strength. Solid line recapitulates experimentally validated kinase–substrate–site relationships directly abolished by pathogenic variants, while dashed lines highlights clinically supported but incompletely resolved phosphorylation mechanisms revealed by disease mutations.

The first category comprises kinase–substrate–phosphosite–disrupting variants for which the phosphorylation event is explicitly demonstrated at the residue level in primary literature, with the upstream kinase directly assigned. These include AMPK α 1 phosphorylation of PTPS at Thr58, documented under hypoxic conditions and linking this site to metabolic adaptation (Zhao et al., 2020). Aurora kinases phosphorylate TP53 at defined sites, with Aurora-A targeting Ser215 and Aurora-B targeting Ser269, producing established effects on p53 DNA binding, transcriptional output, and stability (Gully et al., 2012; Q. Liu et al., 2004; L. Wu et al., 2011). Canonical WNT pathway regulation is fully recapitulated, including CK1-family priming

phosphorylation of β -catenin at Ser45, which initiates destruction-box processing, followed by GSK3A/GSK3B phosphorylation of β -catenin at Ser33, Ser37, and Thr41, which together license β -catenin for ubiquitin-mediated degradation (Amit et al., 2002). Variants affecting any of these residues therefore abolish a sequential, experimentally defined phosphorylation cascade.

Additional events include CK1-dependent phosphorylation of PER2 at Ser662, which controls circadian period length and underlies familial advanced sleep-phase syndrome (Shanware et al., 2011); CK2 phosphorylation of cyclin F at Ser621, required for proper SCF–cyclin-F E3 ligase function and genome stability (A. Lee et al., 2017); OSR1/SPAK phosphorylation of SLC12A3 (NCC) at Thr60, a required activation site for the sodium–chloride cotransporter, disrupted in renal salt-wasting disorders (Yang et al., 2013); JNK-mediated phosphorylation of β -catenin at Ser37, documented in stress-activated signaling contexts (M.-H. Lee et al., 2009); and PKC-dependent phosphorylation of KIT at Ser821, which contributes to full receptor signaling output (Blume-Jensen et al., 1995). Across all Tier 1 entries, disease-associated variants either eliminate the phospho-acceptor residue itself or disrupt the immediate local sequence required for kinase recognition, directly abolishing phosphorylation events whose kinase identity, modified residue, and regulatory consequence are experimentally established.

The second category with dashed lines highlights nominate clinically implicated phosphorylation mechanisms that are biologically well supported but incompletely resolved at the kinase–site level, where disease-associated mutations illuminate regulatory failures underlying characteristic patient phenotypes. In the RTK–ubiquitin axis, insulin signaling provides direct support for CBL phosphorylation at Tyr371, and Tyr371 is established as a critical activation switch for CBL E3 ligase activity (Kassenbrock & Anderson, 2004; J. Liu et al., 2002). Clinically, pathogenic and likely pathogenic CBL variants at or near Tyr371 are associated with Noonan syndrome-like developmental phenotypes and juvenile myelomonocytic leukemia, disorders marked by excessive growth-factor signaling. In immune contexts, TAM receptor pathways further implicate CBL-family ubiquitin regulation, as TAM signaling attenuates immune activation through CBL-B-dependent ubiquitination (Chirino et al., 2020). In neurodegeneration, NUA1-dependent tau phosphorylation is strongly linked to tau accumulation and toxicity; reduction of NUA1 lowers both phosphorylated tau and total tau levels in disease models (Lasagna-Reeves et al., 2016). ClinVar-pathogenic MAPT variants mapping to residues annotated as

Ser641 likely reflect isoform-dependent residue numbering and are associated with frontotemporal dementia and related tauopathies. Similarly, tau Ser669 emerges as a potential regulatory hotspot: protein kinase C phosphorylates tau broadly (Hoshi et al., 1987), and PKN1 has been shown to phosphorylate tau with curated site attribution including Ser669 (Taniguchi et al., 2001), with pathogenic variants linked to progressive supranuclear palsy-like syndromes. Beyond neuronal disease, AKT signaling directly phosphorylates lamin A/C, influencing nuclear mechanics, chromatin organization, and cell survival under mechanical stress (Bertacchini et al., 2013). Clinically, pathogenic LMNA variants cause laminopathies including Emery–Dreifuss muscular dystrophy, dilated cardiomyopathy, and premature aging syndromes. Stress signaling further implicates phosphorylation-dependent control of the cell cycle, as p38 activation mediates rapid proteasomal degradation of cyclin D2 under stress conditions (Kida et al., 2007); pathogenic CCND2 variants are associated with megalencephaly and cancer predisposition. Phosphorylation-dependent regulation of CFTR channel activity and trafficking through protein kinase C signaling (Berger et al., 1993; Guggino & Stanton, 2006) provides a further clinically relevant mechanism, with ClinVar variants mapping to incompletely characterized phosphosites and associated with cystic fibrosis. Together, these Tier 2 cases demonstrate how disease-associated mutations illuminate phosphorylation-dependent regulatory mechanisms that are clinically consequential yet incompletely mapped. Rather than representing weak evidence, Tier 2 highlights define a tractable discovery space where human genetics guides refinement of kinase–substrate–site relationships.

2.2.5 Downstream Application: Inference from Viral Phosphoproteomics and Mutational Scanning Datasets to Elucidate Host Kinase Hijack Mechanisms and Propose Drug-Repurposing Strategies

The second downstream application screens viral proteins against the human kinome to identify host kinases potentially hijacked during infection and to propose drug-repurposing strategies. Input data were drawn from five deep mutational scanning (DMS) fitness datasets and one viral phosphoproteomics dataset (Table 2.6).

Table 2.6: Input datasets for the viral phosphoproteomics and drug-repurposing pipeline.

Reference	Viral Protein / System	Dataset Type
Hom et al. (2019)	Matrix protein 1 (M1) of A/Puerto Rico/8/34 (PR8) influenza virus	DMS Virus Fitness
Haddox et al. (2018)	Two HIV transmitted-founder viruses BF520.W14M.C2	DMS Virus Fitness
Thyagarajan and Bloom (2014)	Hemagglutinin from A/WSN/1933 (H1N1) influenza	DMS Virus Fitness
J. M. Lee et al. (2018)	Hemagglutinin from A/Perth/16/2009 (H3N2) influenza virus	DMS Virus Fitness
Sourisseau et al. (2019)	Zika virus E protein	DMS Virus Fitness
Yaron et al. (2022)	Phosphoproteomic analysis of SARS-CoV-2-infected cells	Virus Phosphoproteomics

From DMS datasets, highly conserved pro-fitness phosphorylable residues (not prone to mutation and of type S/T/Y) were prioritized. From virus phosphoproteomics data, consistently phosphorylated nucleocapsid protein or ORF protein residues were selected. KINBERT was then used to perform in silico screening against the human kinome to infer the top kinases potentially phosphorylating each viral phosphorylable site. Confident candidates were linked to kinase inhibitor databases including PKIDB (Carles et al., 2018) and DrugBank (Wishart, 2006), and cross-validated against known protein interaction data.

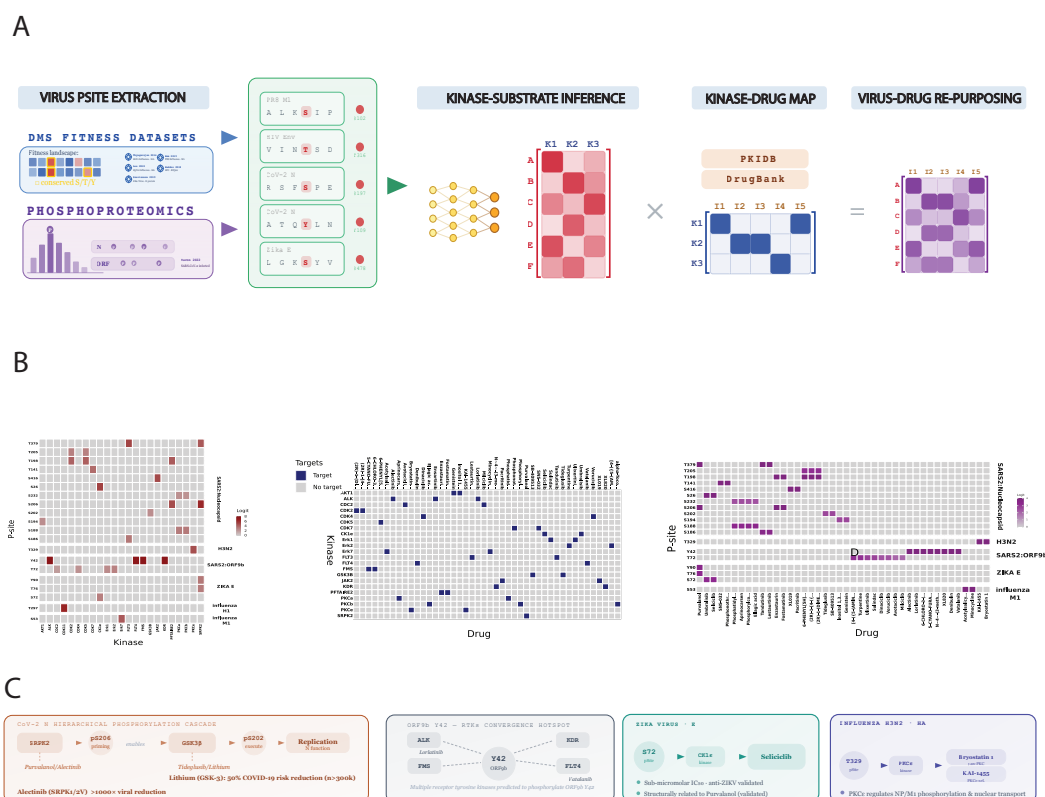


Figure 2.6: Viral phosphoproteomics and drug re-purposing pipeline schematics A) Pipeline detailing DMS and viral phosphoproteomics phosphosites were screened with KINBERT to predict upstream kinases, followed by inhibitor mapping and interaction-based validation. B) Schematics representation and result for re-purposing proposal based on kinase-substrate interaction and kinase-inhibitor relationships C) Summary of novel virus inhibitor prediction highlights

As shown in (Figure 2.6), for the Zika virus E protein, KINBERT nominated CK1 ϵ as the upstream kinase for S72. CK1 ϵ is a pan-viral kinase with documented activity against proteins across multiple virus families, and the predicted inhibitor Seliciclib (Roscovitine) demonstrates strong anti-ZIKV activity at sub-micromolar IC₅₀ values, including at 24 nM in neural progenitor cell models (M. Xu et al., 2016). Its structural relation to Purvalanol, an independently validated CDK-inhibiting anti-ZIKV compound identified in the same screen (M. Xu et al., 2016), further corroborates the prediction, and the convergence of computational nomination with experimental antiviral evidence supports this as a high-confidence therapeutic candidate.

For the SARS-CoV-2 nucleocapsid protein, KINBERT identified a two-kinase phos-

phorylation pattern at the SR domain. SRPK1/2 target S206, a site located within the canonical SRPK recognition sequence of the SR-rich domain (aa 176–246), while GSK-3 β targets the adjacent S202; both sites fall within a region whose phosphorylation state has been associated with N-protein condensate dynamics (Carlson et al., 2020), consistent with a model in which SRPK and GSK-3 β act in concert on the SR domain. GSK-3 β phosphorylation of the N protein SR domain is functionally essential for viral replication, as demonstrated by knockout of *GSK3A* and *GSK3B* abolishing N phosphorylation (X. Liu et al., 2021). The clinical relevance of this axis is supported by a retrospective analysis of over 300,000 patients showing that lithium, a direct GSK-3 inhibitor, reduced COVID-19 risk by approximately 50% (OR = 0.51, $p = 0.005$; (X. Liu et al., 2021)). The nominated inhibitor Tideglusib has completed Phase IIb clinical trials for Alzheimer’s disease with a favorable safety profile, while SRPK1/2 inhibition by Alectinib has been shown to produce greater than 1,000-fold viral reduction in experimental systems (Yaron et al., 2022).

Two novel predictions with no prior experimental precedent were also generated. For the SARS-CoV-2 ORF9b protein, multiple receptor tyrosine kinases – ALK, FMS, KDR, and FLT4 – converge on Y42, which resides in a strong RTK consensus sequence, suggesting this residue constitutes a phosphorylation hotspot warranting experimental follow-up. For the influenza M1 protein, ERK7 (also known as ERK8 or MAPK15) was nominated for S53. ERK7 is an atypical MAPK with unclear physiological functions and no established role in influenza infection; while classical MAPK family members are well-documented participants in influenza signaling (Dey & Mondal, 2024), the specific role of ERK7 remains speculative and this prediction is presented as a hypothesis-generating candidate. The associated inhibitors Aspirin and Minocycline both exhibit antiviral properties in experimental systems, though through non-canonical mechanisms.

2.3 Discussion

KINBERT advances kinase–substrate interaction prediction along several complementary axes. By integrating four curated databases and standardizing identifiers to a common UniProt-based reference, the training corpus reaches 147,488 interactions across 2,304 kinases in 92 organisms, substantially exceeding the scope of prior methods. This breadth is critical for cross-kinase and cross-species generalization, as evidenced by strong performance in the dark-kinase and non-human holdout evaluations.

Focusing on the kinase domain rather than the full-length kinase sequence concentrates learning on the catalytic and substrate-recognition machinery without overfitting on species-specific regulatory context. The use of a 15-residue flanking window rather than the 11-residue window employed by Phosformer provides additional sequence context that may be particularly informative for kinases with extended recognition determinants. The explicit separator token enabling joint kinase–substrate encoding allows the transformer to learn pair-specific compatibility through token-level cross-attention across the full backbone, a capability unavailable to methods that encode kinase and substrate independently and compare resulting embeddings. The contrast with KolossuS further highlights the importance of matching the training objective to the evaluation criterion. Interestingly, we observed overall comparable but consistently better performance using TAPE backbone compared to ESM2-t12, and this observation provides a strong evidence that aligns with previous reports that finetuning ESM2 exhibits instability characterized by spiking losses and abrupt fluctuations in performance, with convergence requiring more than 20–30 epochs (Glaser & Brägelmann, 2025).

The two downstream applications illustrate complementary uses of the inference engine. The pathogenic variant pipeline connects the output of kinase–substrate prediction to human disease genetics, providing mechanistic hypotheses about how missense mutations disrupt phosphorylation-dependent regulatory circuits. Distinguishing site-resolved experimental evidence from biologically supported but incompletely characterized events enables responsible prioritization of predictions for experimental follow-up. The viral phosphoproteomics application demonstrates an orthogonal utility: by screening conserved, pro-fitness viral phosphorylable sites against the human kinome, KINBERT can nominate the host kinases most likely to be exploited by viral replication machinery and link these to existing clinical-stage inhibitors, offering a rational starting point for antiviral drug repurposing.

Several limitations merit acknowledgment. While LC-MS lysate data enriches coverage of less-studied kinases, these interactions capture in-cell phosphorylation under specific conditions and may include indirect events. The exclusion of kinases with multiple conflicting catalytic domain annotations reduces ambiguity but may bias against multi-domain kinases. Cross-species generalization, while demonstrated to be strong, remains more challenging than within-human prediction. The Yeast × Human PK benchmark further highlights this challenge by taking the extreme to fungus cytoplasmic peptide pool. More orthogonally aligned substrate pools in the

future would be required to learn thoroughly the phosphorylation specificity across biases introduced by species imbalance in training.

Future directions include incorporating structural information from AlphaFold kinase–substrate complex predictions, expanding the training corpus with emerging large-scale phosphoproteomics datasets, and extending the downstream variant interpretation pipeline to somatic cancer mutations. The KINBERT framework provides a scalable foundation for systematic, high-coverage kinase–substrate interaction prediction and its downstream application to disease biology.

2.4 Methods

Raw Data Sources

Experimentally supported kinase–substrate interactions were aggregated from six primary sources. **DeepKinZero** provides curated kinase–phosphosite pairs with residue-level annotation; kinase and substrate fields may contain multiple accessions per entry. **PhosphoSitePlus** contributes high-confidence, literature-curated human phosphorylation events, including kinase attribution and ± 7 -residue sequence context. **phosphoELM** supplies curated phosphosites with kinase gene names and organism annotations; kinase accessions were resolved post hoc. **KinasePhos 3.0** provides computationally and experimentally supported kinase–substrate pairs with residue positions, requiring extensive accession normalization; experimentally verified kinase-specific phosphorylation data were collected from PhosphoSitePlus, UniProt, GPS 5.0, and Phospho.ELM, yielding 41,421 verified kinase-specific phosphorylation sites and 1,380 unique kinases, including 753 with known KinBase classification and 627 annotated by evolutionary tree. **PhosD training set** provides kinase–substrate pairs validated across databases but lacking phosphosite positions. **PhosphoNetworks (rawKSI, refKSI, comKSI)** contributes large-scale in vitro and refined kinase–substrate relationships derived from protein microarrays and literature curation; phosphosite positions are generally unavailable.

Catalytic domain coordinates were extracted by parsing UniProt feature annotations, prioritizing canonical protein kinase domains, followed by histidine kinase and Fam20C domains where applicable. Only interactions involving kinases with exactly one unambiguous catalytic domain interval were retained; kinases lacking resolvable domains or exhibiting multiple conflicting domain annotations were excluded.

Family-Aware and Background-Augmented Negative Sampling

To generate informative negative examples for model training, a unified negative sampling dataset was constructed from the class-annotated interaction set (`interaction_classed_221017.pkl`). For each positive kinase–substrate–phosphosite interaction, a family-aware hard negative was generated by pairing the same kinase with a substrate phosphosite of the same residue type (Ser, Thr, or Tyr) that is associated with an incompatible kinase family context. For serine and threonine sites, contrastive negatives were sampled from substrates phosphorylated by kinases belonging to a different protein kinase family, excluding tyrosine kinases; for tyrosine sites, negatives were sampled from substrates phosphorylated by kinases from a different tyrosine kinase subfamily. Phosphosites targeted by more than three kinase families were excluded prior to negative sampling to reduce ambiguity from broadly promiscuous sites. In addition to family-aware contrastive negatives, background S/T/Y negatives were incorporated by pairing kinases with randomly sampled substrate phosphorylatable residue-flanking sequences of the same residue type that are not annotated as phosphorylated substrates. These background negatives capture general sequence-level non-interactions without enforcing family-level incompatibility. Family-aware and background negatives were combined with positive interactions into a single balanced training dataset (`interactions_e2e_1n1p_classed_random.pkl`), used exclusively for model training.

Inference on Pathogenic PTM Human Variants

The pipeline takes as input a set of pathogenic human PTM variants from PTMVar (Hornbeck et al., 2015) – specifically phosphosites harboring disease-associated missense mutations – paired against the full human kinome (~500 kinases). A BERT-based sequence model (TAPE) is run in two complementary inference modes: an on-site mode, which scores each kinase against the 15-amino-acid window centered on the phosphosite directly, and a flanking mode, which scores the wild-type versus mutant peptide in parallel and flags kinases whose logit drops significantly ($\Delta > 2.5$) upon mutation. Kinases are then ranked within their class and filtered to retain only high-confidence predictors (top 5%, logit > 5), while atypical kinases and cell-cycle kinases (CDC, CDK, DYRK) are excluded. Disease annotations are retrieved via the MyVariant.info API, linking each variant’s dbSNP rsID (Sherry, 2001) to ClinVar pathogenicity classifications (Landrum et al., 2014), and only

variants annotated as pathogenic or likely pathogenic are retained. The predicted kinase–substrate pairs are cross-validated against three protein interaction databases – OmniPath (Türei et al., 2016), PhosphoNetworks (Hu et al., 2014), and STRING-DB (Szkłarczyk et al., 2023) (score > 0.8) – to ensure the kinase and substrate are known to physically interact. Finally, all hits are manually curated into evidence tiers.

KINBERT Model Training and Evaluation

KINBERT is a kinase–substrate interaction (KSI) classifier built by fine-tuning the TAPE BERT-base protein language model (`ProteinBertForSequenceClassification`). The TAPE BERT-base backbone consists of a 12-layer transformer encoder pre-trained on large-scale protein sequence corpora using the IUPAC amino acid vocabulary. A two-class linear classification head was appended to the pooled [CLS] representation to produce binary interaction logits. Dropout regularization ($p = 0.1$) was applied within the classification head, and the full backbone was left unfrozen during training to allow task-specific adaptation of all transformer parameters. Each training example consists of a kinase–substrate sequence pair encoded as a single concatenated token sequence padded to a maximum length of 768. The kinase sequence was tokenized first using the TAPE IUPAC tokenizer, followed by a <sep> separator token, and then the tokenized substrate sequence. The resulting concatenated sequence was truncated to a maximum length of 768 tokens and right-padded with <pad> tokens to a uniform length. The labeled dataset was partitioned into training, validation, and test sets using an 80:10:10 ratio. The training split was used for parameter optimization, the validation split for threshold selection, learning rate scheduling, and early stopping, and the held-out test split was reserved exclusively for final performance reporting. The model was trained using stochastic gradient descent (SGD) with Nesterov momentum (momentum = 0.9) at an initial learning rate of 0.001 and no weight decay. Gradient accumulation was set to 1, and the effective batch size was 64 samples per gradient update step. The learning rate was adjusted using a `ReduceLROnPlateau` scheduler monitored on validation loss: when validation loss failed to decrease for 2 consecutive epochs, the learning rate was reduced by a factor of 0.5, down to a minimum of 1×10^{-6} . Training proceeded for a maximum of 40 epochs, with early stopping applied on validation loss with a patience of 10 epochs. Training was performed on four A100 GPUs using bfloat16 automatic mixed precision (AMP) via `torch.autocast`, with TF32 matrix multiplication enabled on Ampere-class GPUs. Multi-GPU

training was supported through PyTorch DistributedDataParallel (DDP) using the NCCL backend, launched via `torchrun`. DistributedSampler ensured non-overlapping data shards across GPUs with epoch-level shuffling, while gradient synchronization was handled automatically by DDP. For the primary benchmark, predicted probabilities from KINBERT:TAPE were compared against Phosformer and KolossuS, with predictions aligned to each panel by matching on kinase domain sequence and 15-amino-acid substrate sequence windows. A fixed threshold of 0.5 was applied uniformly across all models for comparability, with performance reported as Accuracy, F1 Score, ROC-AUC, and PR-AUC.

Model Availability

KINBERT model is available at https://huggingface.co/steveyu323/kinbert_v2_long.

Chapter 3

MASSIVELY-PARALLEL PHOSPHO-INTERACTOME CHARACTERIZATION AND NOVEL KINASE BIOSENSOR DESIGN ENABLED VIA PHOSPHO-PCA

Abstract

Systematic mapping of kinase–substrate interactions (KSIs) at scale is essential for understanding the logic of cellular phosphorylation networks. Here we describe the development and application of a barcoded Y2H-inspired phospho-PCA platform for high-throughput profiling of KSIs across a diverse panel of serine/threonine (ST) and tyrosine (Y) kinases against randomized peptide substrate libraries. The pipeline employs Bayesian hierarchical modeling to decouple kinase-level fitness effects from true phosphorylation interaction signals, and a false-positive control library establishes a promiscuity index as an additional quality control metric. Benchmarking against curated KSI databases and validation with cell-free kinase assays confirm the reliability of the screening approach. Substrate pools are analyzed to reveal co-evolutionary relationships between substrate preference and kinase domain divergence. Novel substrates illuminate previously understudied dark kinases, with case studies including ADCK1, KKCC1, and PHKG1. Finally, fine-tuning a pre-trained transformer model (KINBERT) on phospho-PCA data substantially improved its ability to predict KSIs in the physicochemical space sampled by pooled screening, yielding a high-confidence candidate set for biosensor engineering applications. Specifically, TANK-binding kinase 1 (TBK1) and I κ B kinase β (IKK β) are central effectors of innate immune signaling whose dysregulation underlies autoimmune disease, neurodegeneration, and cancer. Developing live-cell biosensors for these kinases is particularly challenging owing to their delayed and transient activation kinetics, low endogenous abundance, and spatially restricted activation. Here we describe the design and functional characterization of SPARK (Signaling with a Phase-separation Reporter Kinase) reporters targeting TBK1 and IKK β . Optimal substrate peptides were nominated through a multi-stage pipeline integrating curated kinase–substrate databases, high-throughput phospho-PCA screening data, phosphoamino acid-binding domain (PAABD) affinity filtering, and fine-tuned KINBERT phosphorylation selectivity scoring. Reporters were validated in HeLa cells across pharmacological inhibitor and activating stimulus conditions (TNF α , polyIC,

diABZI) using an automated two-stage single-cell quantification pipeline combining CellSAM foundation-model segmentation with Occident tracking. Statistical comparison of peak Δ SPARK distributions confirmed that stimulus treatment produced significant increases in reporter activity above baseline across all quantifiable reporters. The spatial and temporal activation patterns recovered from single-cell traces align with published immunofluorescence and western blot evidence, including Golgi-restricted TBK1 activation and a more pan-cytoplasmic IKK β signal distribution. These results establish a validated, end-to-end pipeline for real-time single-cell reporting of TBK1 and IKK β kinase activity.

3.1 Introduction

Protein phosphorylation, catalyzed by the approximately 500 kinases encoded in the human genome, is among the most pervasive post-translational modifications coordinating cellular signaling. As detailed in Chapter 1, despite decades of biochemical investigation, the full landscape of KSIs remains incompletely characterized. Systematic mapping of KSIs remains a central challenge in cell signaling research, yet several compounding difficulties have impeded progress toward a comprehensive, unbiased atlas. Substrate profiling by applying recombinant kinases *in vitro* does not always recapitulate *in vivo* conditions, limiting the biological relevance of biochemical specificity measurements (Jehle et al., 2022). Eukaryotic model systems offer a more physiologically relevant alternative. Yeast has been widely used in protein–protein interaction studies through yeast two-hybrid (Y2H) protein complementation assay (PCA) screening, in which two haploid yeast strains are transformed with plasmids expressing a bait protein and a prey protein respectively. A drug-resistance protein—DHFR in this case—is split into two fragments (DHFR1,2 and DHFR3), each linked to one interacting partner. Upon mating the haploid strains, the diploid yeast in which prey and bait associate reconstitutes the functional DHFR and survives under methotrexate selection. Positive hits appear as colonies at defined positions in array format, with prey and bait identities inferred from colony position. Because phosphorylation is a transient and unique form of protein–protein interaction, a phospho-Y2H method was developed in which a phospho-amino acid binding domain (PAABD) serves as the bait and the target substrate of a kinase serves as the prey; a third plasmid expressing a human protein kinase is co-transformed. Upon kinase-mediated phosphorylation of the substrate in the same yeast cell, the PAABD binds the phosphorylated peptide, reconstituting DHFR and enabling colony growth under drug selection. Despite demonstrated successes in identifying novel KSIs

through yeast-based screening (Grossmann et al., 2015), the phospho-Y2H format does not accommodate the throughput required for functional deep mutational scanning of kinase domains or for unbiased crossing of more than ten kinases against a substrate library with thousands of peptides.

Next-generation sequencing (NGS) technology has empowered the throughput of functional screening experiments. Rather than inferring variant identity from colony position, NGS enables simultaneous readout of up to hundreds of variants in a pooled, one-pot assay by sequencing the enriched plasmid pool extracted from yeast pellets after drug selection (Faure et al., 2022). However, the three-plasmid architecture of phospho-Y2H—one each for kinase, substrate candidate, and PAABD—is incompatible with NGS-based readout, because the identity of the kinase–substrate pair present in a surviving yeast cell cannot be recovered after pooled screening. To enable kinase–substrate pair identification in a sequencing-compatible format, a single-plasmid construct was designed in which the kinase and substrate are encoded on the same plasmid and connected by a flexible fusion protein linker that permits domain interaction while maintaining spatial separation between the two properly folded moieties (X. Chen et al., 2013). Another challenge arises from the length of the variable insert: because a typical kinase domain is approximately 300 amino acids, the amplicon spanning [kinase domain]–[linker]–[15 amino-acid substrate peptide] requires approximately 1,200 base pairs, exceeding the read-length limit of standard short-read NGS platforms. To overcome this length constraint, high-fidelity long-read amplicon sequencing was applied to a barcoded library in which each kinase–substrate variable region is associated with one or more 19-nucleotide random oligonucleotide barcodes (Wei & Li, 2023). PacBio HiFi circular consensus sequencing, while not providing sufficient depth to quantify relative variant abundance, generates a lookup table linking each kinase–substrate pair to its corresponding barcode (Amorosi et al., 2021). Short-read NGS is then used to deeply sequence barcode-only amplicons, providing the coverage required to quantify each variant’s relative abundance before and after drug selection. In the phospho-interactome context, barcodes overrepresented in the post-selection pool are mapped back to their kinase–substrate pair using the lookup table. A key challenge in such functional screens is disentangling the molecular recognition signal from confounding factors including kinase catalytic efficiency, expression level, and assay geometry.

TANK-binding kinase 1 (TBK1) and I κ B kinase β (IKK β) are central effectors of innate immune signaling. TBK1 is activated downstream of cytosolic nucleic acid

sensors such as cGAS and phosphorylates the transcription factor IRF3 to drive its dimerization, nuclear translocation, and transcription of type I interferons (Fitzgerald et al., 2003). IKK β , in complex with IKK α and the scaffolding subunit NEMO, mediates the canonical NF- κ B activation arm of the STING pathway by phosphorylating I κ B α , targeting it for proteasomal degradation and releasing active NF- κ B to drive pro-inflammatory cytokine gene expression (Fang et al., 2017). Dysregulation of TBK1 and IKK β is implicated in autoimmune conditions including lupus and Sjögren's syndrome, in STING-associated vasculopathy with onset in infancy (SAVI), in TBK1-linked familial ALS, and in oncogenic NF- κ B-driven tumor survival. These kinases are therefore both scientifically important as signal transduction nodes and therapeutically relevant as drug targets. Developing biosensors for TBK1 and IKK β is particularly challenging for three interrelated biological reasons. First, unlike constitutively active kinases such as CK2 or CDK2, TBK1 and IKK β are activated in a stimulus-dependent, pulsatile manner. Western blot time-course experiments measuring pTBK1 Ser172 and downstream pIRF3 following cytosolic dsDNA stimulation consistently show that phosphorylation signal is not detectable before approximately 2 hours post-stimulation, peaks in the 4–6 hour window, and resolves within 12–24 hours (Kemmoku et al., 2022). This narrow and delayed activation window means that endpoint immunofluorescence or immunoblot assays almost inevitably miss the peak of activity unless carefully timed, creating a strong practical motivation for continuous real-time reporters. Furthermore, proteomic datasets from diverse human cell lines consistently place TBK1 in the low-to-moderate expression tier relative to housekeeping kinases such as CK2. Antibody validation studies confirm that detecting endogenous TBK1 by standard immunoblot requires optimized conditions—typically higher amount of total protein and high-sensitivity detection—to obtain reliable signal above background (Alshafie et al., 2022). Lastly, TBK1 activation is not a cytoplasm-wide event; it is spatially confined to the trans-Golgi network (TGN). Live-cell imaging studies demonstrate that upon STING stimulation, STING traffics from the endoplasmic reticulum through the Golgi to the TGN, and that active pTBK1 co-localizes specifically with TGN46-positive subdomains rather than with cis- or medial-Golgi markers (Kemmoku et al., 2022). Recruitment of TBK1 to Golgi membranes for trans-autophosphorylation has also been observed in the RIG-I antiviral pathway, suggesting that Golgi-restricted TBK1 activation is a general feature of RNA sensing as well as DNA sensing (Pourcelot et al., 2016). The spatial restriction means that population-average phosphoproteomic measurements will systematically underestimate TBK1 activity, and that a biosensor

must be capable of reporting subcellular compartment-resolved kinase activity—a requirement that broad-spectrum cytoplasmic reporters cannot fulfill. Systematic identification of high-activity, high-specificity TBK1 and IKK β substrate peptides would directly unlock Exrai-based biosensor development. In the Exrai framework, the substrate peptide and the phosphoamino acid-binding domain (PAABD) are the two modular, independently optimizable components that determine reporter selectivity and dynamic range. The development of Exrai-AKAR for PKA demonstrated that substituting a 14-3-3 domain with an FHA domain significantly enhanced dynamic range (J.-F. Zhang et al., 2021), underscoring that both components must be systematically optimized for each target kinase.

To quantitatively interpret such compartment-resolved signaling readouts at the single-cell level, robust image analysis pipelines are required. Live-cell microscopy provides dynamic measurements of processes including cell division, morphological state transitions, and signaling responses. These analyses require both accurate segmentation and temporally consistent tracking, motivating the development of tools such as CellSAM for accurate segmentation. The Occident pipeline is motivated by long-standing limitations in dynamic-cell analysis datasets and methods, including constrained scale, sparse temporal labels, short trajectories, and limited mitosis annotation (Anjum & Gurari, 2020; Ker et al., 2018; Lugagne et al., 2020; Su et al., 2021; Tsai et al., 2019; Ulicna et al., 2021; Ulman et al., 2017; Zargari et al., 2023).

This chapter describes the library designs and computational pipelines developed to address these challenges, the benchmarking results used to validate screening platform reliability, and the application of the resulting KSI dataset to both finetuning pLM model and kinase reporter development.

3.2 Results

3.2.1 Phospho-PCA Plasmid Construct Optimization

To our knowledge, there is no previously published and available design profiling kinase–substrate interactions using a single-plasmid construct. We therefore began the design with a published Protein Complement Assay (PCA) plasmid backbone and created a prototype phospho-PCA plasmid incorporating the Fyn kinase domain and a known Fyn substrate. The PAABD domain SH2 was linked to DHFR3, and the [substrate–(G-S linker) n –kinase] region (referred to as the “KS region”) was linked to DHFR1,2. To evaluate construct reliability, a catalytically dead kinase version

was included by introducing a mutation in the catalytic loop (K369A). Drawing on findings that expressing human protein kinases under a strong promoter in yeast can compromise cell viability due to excessive phosphorylation of endogenous yeast proteins (Kim et al., 2020), the weaker pCYC promoter was used to drive expression of the [kinase–linker–substrate] fusion. Because no prior report had examined the effect of PAABD promoter strength on yeast-based phosphorylation screening, both pCYC and the stronger pGPD promoter were tested for driving SH2 expression. After cloning the PAABD and KS region into the PCA expression vector, plasmids were transformed into yeast, and single colonies were inoculated at $OD_{600} = 0.005$ in triplicate in 96-well plates. Growth curves were measured over 60 hours at a methotrexate concentration of 200 $\mu\text{g/mL}$. These experiments revealed that driving PAABD expression with pCYC produced a substantially attenuated growth signal relative to pGPD. With pGPD and the optimized GS linker, the screening construct consistently reported phosphorylation-dependent interaction between Fyn and its substrate, while the catalytically inactive Fyn mutant produced no detectable growth (Figure 3.1).

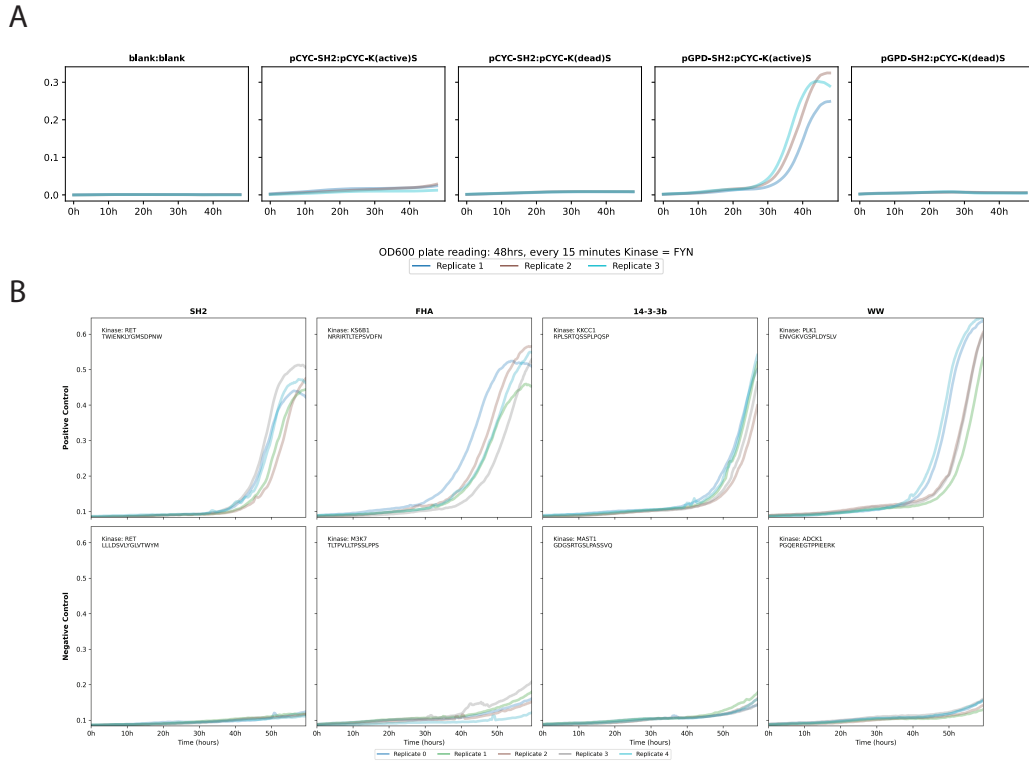


Figure 3.1: Optimization of phospho-PCA promoter strength and PAABD configuration. A) Tuning of promoter strength indicates that a strong promoter driving PAABD yields more distinguishable effects. B) Diversification of PAABD across positive and negative control KSIs enables successful characterization of KSIs with SH2, 14-3-3 β , FHA, and WW domains.

Building on successful promoter and linker optimization, we diversified the PAABD repertoire. Although all PAABDs share a general affinity for phosphorylated residues, individual domains exhibit discriminative binding preferences toward specific phosphorylated residues and the flanking sequence context of the phosphosite. A broad panel of PAABDs was surveyed and applied to screen established positive (phosphorylation-competent) and negative (randomly shuffled) kinase–substrate pairs. This analysis led to the selection of four PAABDs for downstream use: Src-SH2 as a general phosphotyrosine (pY) reader, and FHA, 14-3-3 β , and WW as three phosphoserine/phosphothreonine (pS/pT) readers each with distinct preferences for the residues flanking the phosphorylated site. Together, the four PAABD screening constructs provide theoretical coverage across a diverse pool of phosphorylated 15 amino-acid peptides (Figure 3.1). Library construction details for the 15 serine/threonine kinases and 8 tyrosine kinases are included in the Methods section.

Upon successful cloning, barcoded kinase–substrate libraries were sequenced using PacBio HiFi circular consensus sequencing (Table 3.2). After sequential filtering of quality control and bijective barcode-to-construct assignment, the ST library (blib5) yielded 121,070 (144,716 including non-bijection) unique barcodes from 3,039,052 initial CCS reads (mean 11.27 reads per barcode), and the Y library (blib1) yielded 205,272 (242,424 including non-bijection) unique barcodes from 2,381,935 initial reads (mean 5.31 reads per barcode). Full processing statistics are provided in Table 3.1 and .

Library	Initial CCS reads	Barcode QC	Flank Region QC	After ORF QC	Final unique barcodes	Mean reads per barcode
ST Lib	3,039,052	2,445,887	2,133,880	1,631,649	121,070	11.27
Y Lib	2,381,935	1,935,748	1,691,475	1,286,517	205,272	5.31

Table 3.1: Library processing statistics after QC filtering.

Library	Substrates designed	Kinases	Substrates con-structed (<i>E. coli</i>)	Substrates screened (yeast)	KSIs designed	KSIs con-structed (<i>E. coli</i>)	KSIs screened (yeast)
Y	1,950	8	1,939	1,932	15,600	14,873	13,220
ST	1,978	15	1,968	1,964	29,670	25,214	22,695

Table 3.2: Final constructed library kinase and substrate statistics

During PAABD optimization, we observed one case with negative kinase–substrate pair exhibiting consistent growth. Re-testing with a construct lacking the kinase domain—and with a construct lacking both the kinase domain and the dead kinase—still produced consistent growth in triplicate. Because yeast encode serine/threonine protein kinases homologous to human kinases, this mode of false-positive phosphorylation is attributable to endogenous yeast kinase activity on the substrate construct. Accordingly, a kinase-absent false-positive control library was incorporated into all downstream screening experiments. Because the readout of the screen is yeast growth rather than phosphorylation directly, some peptide substrates can reconstitute DHFR and support proliferation even in the absence of a kinase. To quantify this background, the peptide library was screened in yeast without kinases and per-substrate LFC and FDR were computed from sequencing counts. Substrates with

high substrate-only LFC consistently interacted with a large proportion of kinases, demonstrating that intrinsic growth behavior is a strong predictor of promiscuous KSIs. In contrast, substrates with low LFC rarely showed widespread kinase activity, indicating that their signals were more likely to be kinase-dependent. After filtering, reproducibly growing control substrates were eliminated, yet a large number of novel kinase–substrate interactions remained available for downstream validation (Figure 3.7).

3.2.2 Bayesian Hierarchical Decoupling of Kinase Activity from Phosphorylation Interaction

Inspection of the kinase–substrate screen revealed that enrichment in the DHFR assay reflects more than phosphorylation-dependent recognition. The distribution of log fold change (LFC) scores showed systematic kinase-dependent shifts, indicating a kinase- and PAABD-specific abundance component. Kinase catalytic domains vary in stability, expression, and catalytic turnover, so measured phosphorylation output often reflects enzyme availability and accessibility in addition to intrinsic substrate preference (C. J. Miller & Turk, 2018; Ubersax & Ferrell Jr, 2007). Accordingly, different kinases exhibited consistent shifts in the mean enrichment distribution. In contrast, PAABDs primarily altered the spread of scores by modulating affinity after phosphorylation (Figure 3.8). Phospho-recognition modules have distinct binding strengths and sequence requirements: 14-3-3 and FHA domains typically bind phosphopeptides in the micro-molar range, whereas WW domains recognize short pS/pT-proline motifs more weakly (Durocher et al., 1999; Sudol & Hunter, 2000; Yaffe et al., 1997). Consistent with these properties, WW-domain screens produced tighter, lower-amplitude distributions, while FHA-based assays showed broader dispersion. Thus, kinase domains mainly shift the distribution mean (activity/abundance bias), whereas PAABDs scale the variance (post-phosphorylation binding strength).

To separate kinase activity from molecular recognition, a hierarchical Bayesian model was implemented (Figure 3.2). The diverse substrate library (~2,000 peptides sampling flanking sequence space) allows estimation of kinase-wide baseline behavior independently of peptide identity. Pre-selection counts were incorporated as a log-offset representing expected post-selection abundance in the absence of binding preference, capturing peptide availability rather than recognition. Post-selection counts were modeled with a negative binomial likelihood to account for sequencing over-dispersion. The linear predictor decomposes enrichment into three compo-

nents: a global kinase effect (baseline catalytic or assay efficiency), a replicate-specific deviation, and a substrate-specific phosphorylation term. Phosphorylation effects were centered within each kinase, reflecting the biological constraint that a kinase may be globally strong or weak yet still discriminate among substrates. Statistical power arises from both breadth and replication. Each kinase is evaluated across $\sim 2,000$ peptides, so the activity term is supported by thousands of observations; because most peptides are weak or neutral matches, their aggregate behavior anchors the kinase baseline and yields a precise activity estimate despite noisy individual measurements. Replicates are modeled as a separate random effect, allowing consistent shifts across experiments to reinforce the kinase activity parameter while inconsistent variation is absorbed as experimental noise. This partial pooling stabilizes inference across differences in sequencing depth, selection strength, or batch effects. The posterior mean of the substrate–kinase term is reported as a decoupled binding strength, and confidence is quantified by the posterior probability of a positive interaction and the local false sign rate (lfsr). Conceptually, the model converts the pooled functional screen from a competition dominated by kinase abundance into a measurement of molecular recognition: highly active kinases no longer score broadly merely due to catalytic throughput, and more noisy and less confident measurements would shrink with prior normalization and be flagged with a high LFSR value. The resulting scores approximate intrinsic biochemical selectivity and provide a stable normalization layer for calling substrate-specific interactions.

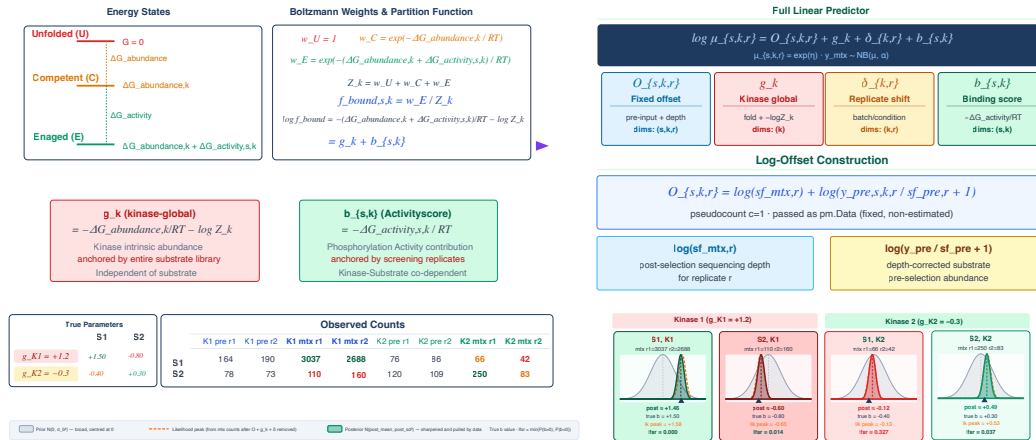


Figure 3.2: **Mathematical derivation of a Bayesian decoupling framework based on Boltzmann-weighted energy states.** Top: Energy states are decomposed into a kinase-global abundance term and a substrate–kinase-specific activity term, forming a linear predictor that absorbs batch-dependent shifts. Bottom: Illustration of the decoupling process from observed counts for two example kinase–substrate pairs.

3.2.3 Benchmarking Against Known KSI Databases and In vitro Validation

To evaluate the performance of the barcoded phospho-PCA platform, screening-positive KSI calls were compared against a curated reference set of experimentally documented interactions. Across the combined library (n=403 annotated pairs), the screen achieves a precision of 0.924. Detailed inspection revealed that a set of false negative substrates are invaluable to phospho-PCA due to the QC removal of false positive substrates and the inability of PAABD to bind to substrates. Restricting the denominator to screening-evaluable pairs yields a substantially higher recall of 0.728. This pattern was consistent across both sub-libraries (Table 3.5).

Inspection of false-negative pairs revealed that a substantial fraction lacked a canonical PAABD motif or were flagged as promiscuous substrates. This suggests that many unrecovered interactions may represent low-affinity or context-dependent events not efficiently captured under the cellular conditions of the yeast screen, as part of the current limitations of the screening platform. The high-precision, lower-recall profile is well-suited to the intended application. Because the primary goal is to generate a compact, high-confidence set of KSIs for downstream mechanistic

validation and biosensor development, minimizing false positives takes precedence over exhaustive interaction recovery. A precision exceeding 0.90 across both sub-libraries means that screening hits can be pursued with high confidence. Missed interactions can be recovered in future rounds through relaxed scoring thresholds, expanded substrate coverage, or orthogonal screening strategies. Of the KSI pairs assessed by both platforms, ~75% of phospho-PCA screening hits were independently confirmed by the Kinexus cell-free kinase activity assay, supporting the high precision of the yeast-based screen on novel KSIs (Figure 3.3).

3.2.4 Substrate Pools Reveal Co-Evolution Between Substrate Preference and Kinase Domain Divergence

Across kinase–kinase sequence pairs, screening-tree distance and kinome-tree distance are significantly correlated (Kendall $\tau = 0.638$, $p = 2.9 \times 10^{-28}$), indicating that kinase-domain evolutionary relationships are a major determinant of substrate selectivity in this panel.

As shown in Figure 3.3, within the tyrosine kinase clade, FRK and FYN form the closest canonical pair, consistent with their shared Src-family architecture. ABL2 clusters with this pair in the kinome tree but diverges in the screening tree. This behavior agrees with ancestral-reconstruction analysis showing that the Src and Abl lineages evolved substantially different substrate specificity after divergence from a common ancestor (Finneran et al., 2020). The phosphomimetic mutants KS6B1-T252D and PLK1-T210D remain within the serine/threonine clade and cluster near their respective wild-type kinases. Rather than altering sequence motif selectivity, activation-loop phosphomimetics primarily stabilize the active kinase conformation and relieve intramolecular autoinhibition—in PLK1, the polo-box-domain-mediated autoinhibited state—while leaving the substrate-binding cleft largely preserved (Chapagai et al., 2025).

Within the serine/threonine clade, an inner basophilic cluster is observed: CaMKK1 pairs most tightly with KS6B1, and KGP1 pairs with PHKG1, consistent with recognition of substrates containing basic residues surrounding the phosphoacceptor site, a defining property of AGC- and CAMK-related kinases (Pearce et al., 2010). Although PLK1 is not classically categorized as basophilic, its large mitotic substrate repertoire creates substantial peptide-space overlap with basophilic kinases (Bibi et al., 2013). Structural alignment analysis further places PLK1 within the CAMK structural group rather than the original “Other” classification, supporting its prox-

imity to CAMK-like kinases in the screening tree (Modi & Dunbrack, 2019b).

Despite being neighbors in the kinome, TBK1 and IKK β segregate in the screen. TBK1 phosphorylates innate immune adaptor proteins including MAVS, STING, and TRIF, promoting IRF3 activation (S. Liu et al., 2015), whereas IKK β phosphorylates I κ B proteins to trigger NF- κ B activation (Ghosh & Karin, 2002). IKK β was expressed with activating S177E/S181E substitutions to ensure constitutive activity (Polley et al., 2013), which may alter conformational regulation. MAST1 clusters with IKK β rather than AGC kinases. MAST kinases differ from canonical AGC kinases in substrate specificity: ENSA is phosphorylated in an acidic DpSG context, and MAST family kinases possess a distinctive activation-loop insertion that does not rely on canonical activation-loop phosphorylation (Lemke et al., 2025). The DUF1908 regulatory domain, required for full MAST2 kinase activity and conserved in MAST1, was absent from the kinase-domain constructs used in the screen and may result in an incomplete rendering of its activity.

MAP3K7/TAK1 lies at the periphery of the serine/threonine clade. Structural re-analysis shows that TAK1 does not cluster with canonical MLK kinases and instead occupies a distinct TKL branch (Modi & Dunbrack, 2019b), consistent with its atypical activation mechanism. Similarly, the structural and biochemical atypicality of STYK1 is well reflected in its substrate screening profile, which placed it as the most distant member from conventional tyrosine kinases in the kinase clustering analysis. Its intracellular kinase domain shares only 30–34% amino acid identity with the FGFR/PDGFR superfamily—its closest known relatives (L. Liu et al., 2004). Unlike typical tyrosine kinases, it is a dual-specificity kinase that phosphorylates both serine/threonine and tyrosine residues (C. Zhou et al., 2020). ADCK1, an ABC1-family atypical kinase, similarly lacks a defined canonical substrate motif and groups distantly from other serine/threonine kinases. Finally, titin kinase functions as a catalytically inactive pseudokinase scaffold rather than an active enzyme (Bogomolovas et al., 2014). Consistent with this, TITIN produced the fewest screening hits and was maximally isolated from other serine/threonine kinases in the screening tree.

Figure 3.3: phospho-PCA shows high precision and reliability in benchmarks, reveals co-evolution of kinase specificity and sequence divergence, and sheds novel insight for understudied kinases A: Left: A juxtaposition of kinase domain phylogeny and substrate selectivity clustering tree reveals the alignments and differences between kinase functional and sequence space. B: phospho-PCA shows high precision with benchmark against literature annotations and additional in vitro validations. C: Information content heatmaps for position-wise AA-property selectivity and substrate GO enrichment for dark kinases ADCK1, KKCC1, and PHKG1.

3.2.5 Novel Substrates Shed New Insight into Understudied Dark Kinases

For the three kinases currently enlisted as dark kinases by The Dark Kinase Knowledgebase, namely CaMKK1 (KKCC1), PHKG1, and ADCK1, we inspected their phosphorylated peptide residue preferences in detail. Given that the randomized peptides were mostly chosen from human phosphoproteomics data, we further examined any meaningful enrichment of biological processes to infer novel involvement of these kinases. It should be noted that CaMKK1 and PHKG1 are comparably better characterized than ADCK1 and would partially serve as pipeline benchmarks for phospho-PCA to reflect known biology findings (Figure 3.3).

CaMKK1 exhibits a strong basic residue preference at position -3 (IC = 0.475), a defining feature of the CaMK kinase family. This motif architecture is consistent with CaMKK1's role as an upstream activator of multiple CaMK family members, enabling recognition of a shared substrate grammar rather than narrow target selectivity. Calcium and calcium channel signaling enrichment directly reflects CaMKK1's canonical activation mechanism as a Ca^{2+} /calmodulin-dependent kinase

(Tokumitsu & Sakagami, 2022). Neuronal and dendritic process enrichment recapitulates well-established $\text{CaMKK} \rightarrow \text{CaMKI/CaMKIV}$ cascades regulating neuronal development and plasticity (Wayman et al., 2008). Nuclear and transcription-related enrichment aligns with CaMKK-mediated pathways linking calcium transients to gene expression programs (Tokumitsu & Sakagami, 2022). Cell population proliferation enrichment is consistent with CaMKK family roles in coupling calcium signaling to growth and metabolism (Sheng et al., 2020). CaMKK1 displays the known CaMK-family substrate grammar and drives calcium-responsive signaling programs that link neuronal activity to transcriptional and developmental outcomes.

PHKG1 (Phosphorylase Kinase $\gamma 1$) displays the strongest basic residue preference at position -3 , with extended enrichment from positions -6 to -3 . This motif precisely matches the canonical motif, demonstrating extreme selectivity consistent with PHKG1's specialized regulatory role. Calcium-mediated signaling enrichment directly reflects PHKG1 activation logic: phosphorylase kinase is activated by Ca^{2+} binding via its calmodulin (δ) subunit, and calcium alone is sufficient for kinase activation (Brushia, 1999). Metabolic and catabolic process enrichment aligns with PHKG1's core role in glycogenolysis, particularly in brain and muscle (Magistretti & Allaman, 2015). GABAergic synapse enrichment suggests neuronal functions beyond classical glycogen metabolism: PHKG1 phosphorylates

neuronal proteins including MAPT/TAU, GAP43, and NRG1 (UniProt Q16816; Pharos database). Structural confirmation that Ca^{2+} /calmodulin binding activates the catalytic γ subunit has been provided by recent crystallographic analysis (Ma et al., 2025). PHKG1 exhibits extreme motif selectivity and directly couples calcium signaling to metabolic output, positioning it as a tightly regulated executor of activity-dependent glycogen breakdown, including in neuronal contexts.

ADCK1 is an atypical kinase with non-canonical architecture and fewer high-confidence substrate hits. Despite sparse hits, ADCK1 shows the highest single-position information content observed: strong basic enrichment at position -6 (IC = 0.755). This potentially suggests an extended substrate-binding interface, consistent with ADCK1's divergent fold and atypical regulatory mechanism. Mitochondrial and metabolic enrichment directly recapitulates ADCK1's established role in mitochondrial structure and bioenergetics: it localizes to the mitochondrial inner membrane and regulates cristae structure and dynamics via OPA1 and IMMT (Yoon et al., 2019). Apoptotic process enrichment aligns with mitochondrial dysfunction upon ADCK1 loss, including reduced membrane potential, ATP depletion, and increased ROS production (Zhuo et al., 2022). Transcriptional and Wnt/ β -catenin signaling enrichment directly mirrors documented ADCK1 function: ADCK1 interacts with TCF4 to activate β -catenin/TCF-dependent transcription and promotes cancer cell growth and migration (Y. Ji et al., 2021). Developmental and spermatogenesis-related enrichment is consistent with energy-intensive differentiation programs and male infertility associations linked to mitochondrial dysfunction. Synaptic and glutamatergic enrichment aligns with high energy demands of excitatory synapses and reported genetic associations with neuropsychiatric disorders (Y. Ji et al., 2021). ADCK1 is a mechanistically distinct kinase whose unique substrate recognition may support its central role in mitochondrial integrity, energy homeostasis, apoptosis resistance, development, and transcriptional regulation.

3.2.6 Fine-Tuning KINBERT on phospho-PCA Data Improves Prediction of Unseen Substrates and nominate optimal SPARK reporter substrates

One of the key limitations discussed in Chapter 2 is the availability of unbiased, high-throughput kinase-substrate interaction data for training sequence-based prediction models. We next evaluated all models on the phospho-PCA KSI benchmark, which comprises ~20 human kinases systematically screened against thousands of substrates in a high-throughput phospho-PCA assay, yielding a dense set of interaction hits that are often physicochemically permissive but not necessarily realized in

endogenous signaling networks.

As expected, given the distributional shift from curated database and HeLa lysate-based training data, all baseline models exhibited markedly reduced performance on this benchmark relative to their standard test sets. Phosformer, Kolossus, and KINBERT:TAPE achieved PR-AUC values of 0.552, 0.469, and 0.547, respectively, and ROC-AUC values of 0.510, 0.452, and 0.610, reflecting their inability to discriminate positive from negative interactions in our orthogonal and unbiasedly-characterized dataset. Inspection of model outputs revealed that pre-trained models defaulted to broadly negative logit predictions with high uncertainty across substrate pairs, consistent with their having learned phosphorylation rules primarily from biologically contextualized data—where non-kinase domain regulation, pathway membership, and cellular co-localization likely contribute implicit filtering—rather than from the unbiased physicochemical space sampled by pooled screening (Figure 3.4).

Fine-tuning KINBERT:TAPE on phospho-PCA training interactions substantially rescued performance, yielding PR-AUC and ROC-AUC of 0.816 and 0.864 on the held-out phospho-PCA test set, while preserving strong performance on the original DB + Lysate benchmark (PR-AUC: 0.953, ROC-AUC: 0.960) (Table 3.3). These results demonstrate that targeted fine-tuning with screening-derived data extends model coverage into the broader physicochemical phosphorylation space without catastrophic forgetting of curated interaction knowledge, and that the phospho-PCA benchmark represents a distinct and nontrivial generalization challenge common to all models evaluated here.

Table 3.3: Performance metrics across models on phospho-PCA-identified KSIs

Model	PR_AUC				ROC_AUC			
	Phosformer	Kolossus	KINBERT:TAPE	KINBERT:TAPE-TUNED	Phosformer	Kolossus	KINBERT:TAPE	KINBERT:TAPE-TUNED
DB + LYSATE Test	0.614	0.720	0.970	0.953	0.695	0.779	0.973	0.960
phospho-PCA KSIs Test	0.552	0.469	0.547	0.816	0.510	0.452	0.610	0.864
phospho-PCA KSIs Train	0.546	0.438	0.534	0.929	0.511	0.421	0.605	0.946

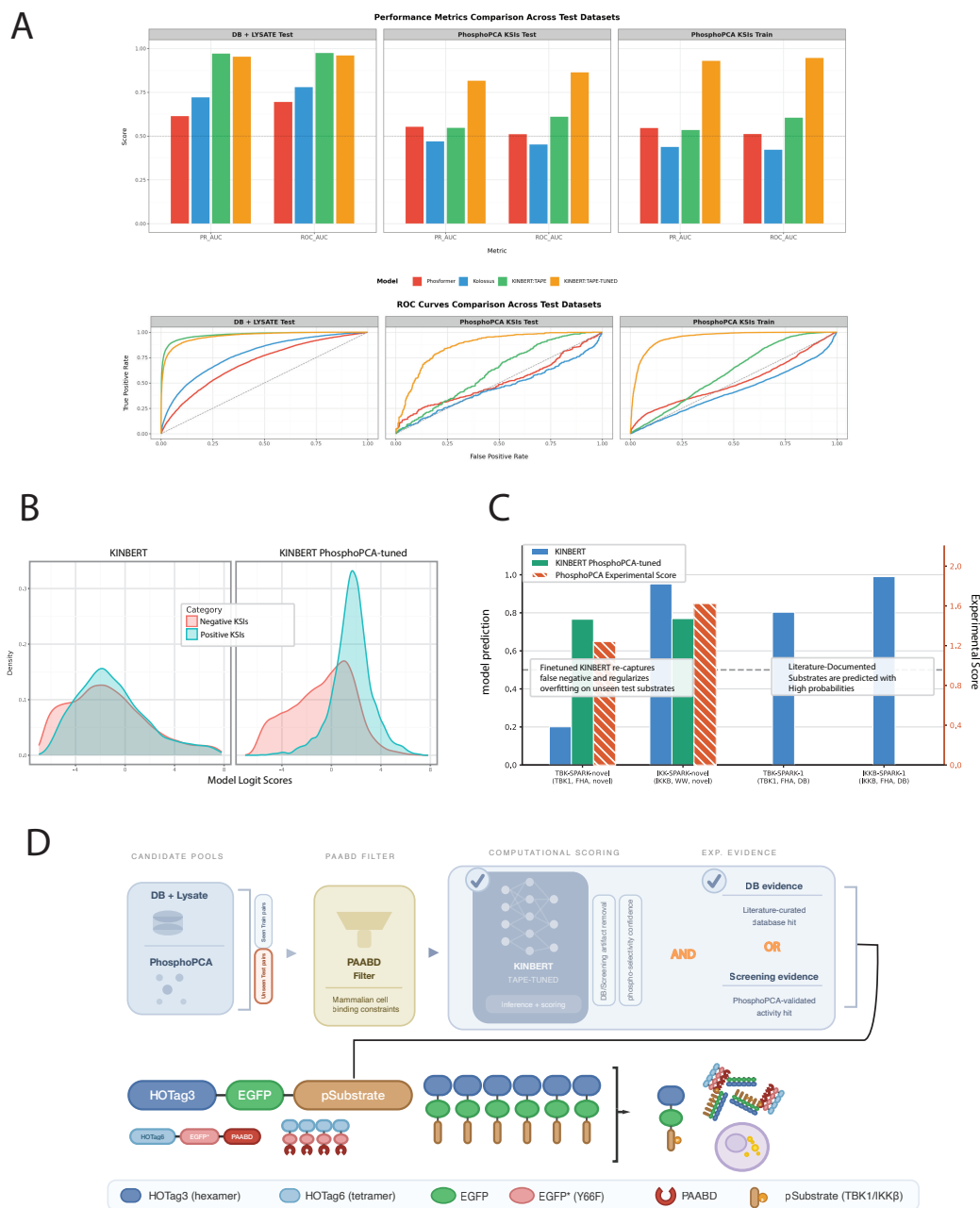


Figure 3.4: Finetuning KINBERT with phospho-PCA data enables identification of kinase reporter candidates from unseen substrates A: Performance metrics of fine-tuned and naive KINBERT on phospho-PCA-identified KSIs. B: Histogram for prediction logic score for naive and fine-tuned KINBERT over the phospho-PCA-KSIs shows improved class segregation. C: Logit score for fine-tuned and naive KINBERT, with TBK-SPARK-novel and IKK-SPARK-novel are unseen substrates removed from fine-tuning. D: Schematics for integrating KINBERT and phospho-PCA as computational and experimental validation sources for downstream biosensor engineering.

To nominate optimal SPARK reporter substrates for IKK β and TBK1, we first assembled two complementary candidate pools: kinase–substrate pairs drawn from the curated DB + Lysate dataset and from the phospho-PCA high-throughput screening dataset, spanning both seen (training) and unseen (test) pairs in each case. Candidates were initially filtered by a strict PAABD binding constraint to retain only pairs with sufficient affinity for the phosphoamino acid-binding domain, a prerequisite for productive intramolecular substrate engagement and reliable reporter dynamic range in mammalian cell contexts. We then applied computational scoring using the fine-tuned KINBERT:TAPE model as an orthogonal layer of evidence—not as a numerical approximation of catalytic activity, since the model is trained on binary phosphorylation classification rather than quantitative activity measurements, but rather as a phosphorylation selectivity classifier whose confidence scores help flag and deprioritize pairs likely to reflect database curation artifacts or substrate specificity arising from non-kinase domain regulatory mechanisms. Independent experimental corroboration was further required from either the literature-curated database or phospho-PCA screening validation, providing biochemical support for each nominated pair. For DB + Lysate-derived candidates, final hit prioritization combined KINBERT selectivity rank with this experimental evidence filter; for phospho-PCA-derived candidates, ranking was performed directly on experimentally characterized activity scores, with KINBERT confidence serving as a secondary consistency check (Figure 3.4-D).

This sequential filtering and prioritization strategy yielded four finalized substrate proposals—one DB-derived and one phospho-PCA-derived candidate for each kinase. Notably, three of these four substrates (TBK1-DB, TBK1-phospho-PCA, and IKK β -phospho-PCA) were not encountered during model training, yet elicited confident positive phosphorylation logits from the fine-tuned model—a marked contrast to the near-chance, broadly negative, and highly uncertain predictions produced by the pre-fine-tuned KINBERT model on the same pairs (Figure 3.4-C). These results demonstrate that fine-tuning on phospho-PCA data extends the model’s ability to confidently assign phosphorylation selectivity to novel kinase–substrate pairings, enabling data-driven substrate nomination for reporter engineering that generalizes beyond the curated training distribution.

Four SPARK reporter variants—IKK-SPK-1, IKK-SPK-novel, TBK-SPK-1, and TBK-SPK-novel—were quantified across conditions spanning vehicle control, kinase inhibitors (BMS-345541 (Burke et al., 2003), GSK-8612 (Thomson et al.,

2019)), activating stimuli (TNF α (Kuo et al., 2009), polyIC (Pourcelot et al., 2016; Reimer et al., 2008), diABZI (Messaoud-Nacer et al., 2022)), and phosphosite T-to-A negative controls. The substrate sequences and computed scores for each reporter construct are summarized in Table 3.4.

Table 3.4: SPARK reporter substrate sequences and KINBERT:TAPE scores.

Reporter	Kinase	Substrate (15 aa)	phospho-PCA	TAPE prob (pre-FT)	TAPE logit (pre-FT)	TAPE prob (FT)	TAPE logit (FT)	PAABD
TBK-SPARK-novel	TBK1_HUMAN	QGPSQTSTPTDVTAI	1.241	0.201	-1.320	0.766	1.427	FHA
IKK-SPARK-novel	IKKB_HUMAN	PHEEPSNSLMAIECR	1.623	0.952	3.000	0.769	1.420	WW
TBK-SPARK-1	TBK1_HUMAN	DHPYASLTLPDDAAA	—	0.803	1.623	—	—	FHA
IKKB-SPARK-1	IKKB_HUMAN	NPQERTLTLVDTGIG	—	0.990	4.620	—	—	FHA

3.2.7 Building a Comprehensive Toolkit for Automated Live-Cell Imaging Quantification: Segmentation, Tracking, and Quantification Using the CellSAM Generalist Module and Occident Pipeline

CellSAM is a SAM-based foundation model for automated cellular instance segmentation trained on a curated multi-archetype dataset (tissue, cell culture, yeast, H&E, and bacteria) with strict train–test decontamination to prevent leakage. To enable fully automated inference, we identified bounding-box prompting as the most effective strategy, then built CellFinder, an Anchor DETR-based detector (Y. Wang et al., 2022) that shares SAM’s ViT features and supplies box prompts for mask enumeration. This combined pipeline achieved state-of-the-art results across ten datasets and strong zero-shot transfer on LIVECell (Edlund et al., 2021). As a result, CellSAM provides a generalist segmentation module that can be reused across diverse biological imaging workflows without redesign. We combined CellSAM with a tracking algorithm (Bochinski et al., 2017) in two representative applications. First, in a HeLa live-cell experiment with transient AMP kinase reporter expression (Schmitt et al., 2022), cells were treated with 20 mM 2-deoxy-D-glucose in a design representative of studies probing signaling dynamics (Purvis & Lahav, 2013). We segmented and tracked cells across 60 frames (120 minutes) to quantify AMP kinase activity trajectories. Second, in budding yeast lineage experiments, we used the same segmentation-tracking pipeline and added division detection to compute cumulative divisions and reconstruct single-cell lineage trees (Figure 3.5-A). Together, these case studies demonstrate that as core segmentation and tracking modules become more generalizable, complete downstream biological analysis pipelines inherit that generality.

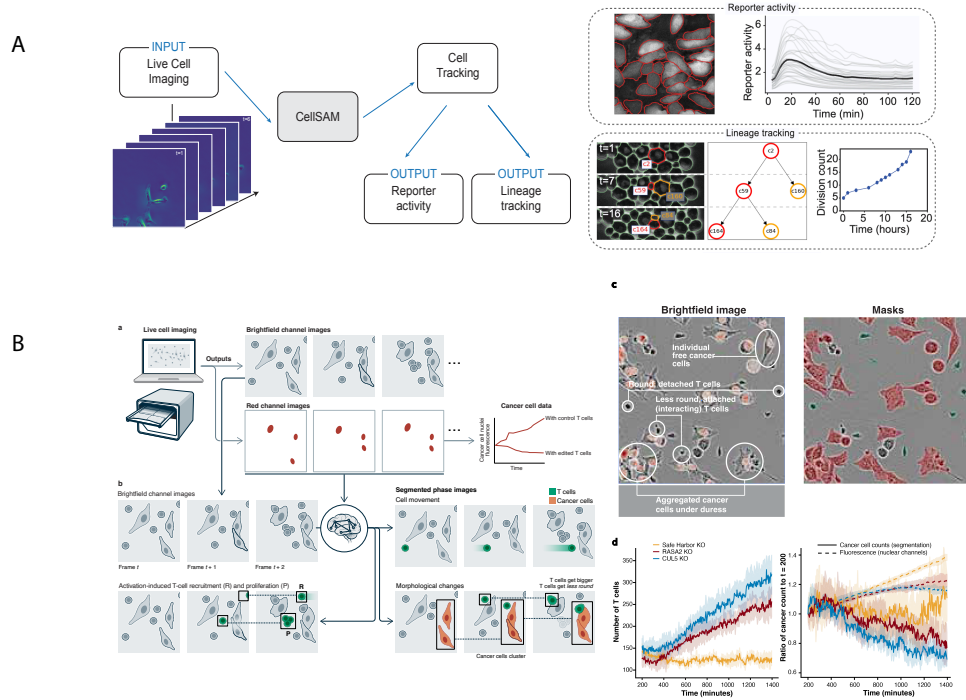


Figure 3.5: CellSAM-Generalist and Occident tracking pipeline enable automated segmentation, tracking, and downstream quantification for live-cell imaging data A: Because CellSAM functions across image modalities and cellular targets, it can be immediately applied across bioimaging analysis workflows without requiring task-specific adaptations. Application of CellSAM to various biological imaging. Segmentations from CellSAM are used to track cells and quantify fluorescent live-cell reporter activity in cell culture. CellSAM segments cells in multiple frames from a video of budding yeast cells. These cells are tracked across frames that ensures consistent identities, enabling accurate lineage construction and cell division quantification. B: a-b): Segmentation pipeline for special cell type masks. A combination of Caliban and SAM was used to get special cancer cell and T cell masks. Red nuclear channel images were used to generate cancer nuclei masks. Phase images were used to generate T cell masks with DeepCell/Caliban and SAM, then ensembled for final masks. Red and phase images were used to generate cancer cell aggregate masks with DeepCell/Caliban. Nuclear masks were used to prompt SAM to generate a second set of masks ensembled with DeepCell to generate individual cancer cell masks. The difference between the aggregate masks and individual cancer masks generated additional attached T cells for inclusion with the previous T cell mask. c) Extracting a catalog of cellular phenotypes with Occident. Morphology is an indicator of cell state. T cells that are interacting with cancer cells change from their round shape to attack cancer cells while cancer cells aggregate together under distress. Segmentation masks are able to capture putative cells and cell aggregates for both cell types. d) Increased T cell expansion with beneficial knockouts. Cancer cell population from segmentation over time reveals greater killing in wells with beneficial T cell perturbations. Fluorescence of the cancer nuclear marker is a lagging indicator of cancer cell population.

To extend segmentation and tracking into behavior-level analysis, we added the Occident pipeline as a downstream quantification layer built on the existing CellSAM + tracking workflow. Within this framework, Caliban provides segmentation and trajectory construction, while Occident aggregates trajectories into interpretable phenotypes for hypothesis testing. We applied this combined pipeline to co-cultures of NY-ESO-1⁺ A375 cancer cells and 1G4 anti-NY-ESO-1 TCR-transgenic T cells across three perturbation states: RASA2 knockout (RASA2 KO), CUL5 knockout (CUL5 KO), and safe-harbor AAVS1 knockout (SH KO). This design is biologically grounded because RASA2 regulates T-cell signaling checkpoints and CUL5 participates in cytokine-linked growth control programs (Bachl et al., 2023). Imaging was acquired every 4 minutes for 24 hours using brightfield plus red fluorescence channels, and outputs were converted into longitudinal cell-state and population measurements. Occident-based quantification recapitulated known functional trends for engineered T cells: RASA2 KO and CUL5 KO conditions showed stronger T-cell expansion and enhanced cancer-cell depletion relative to SH KO controls (Carnevale et al., 2022; Liao et al., 2024). In addition, segmentation-derived cancer counts declined earlier than fluorescence-derived nuclear signal, indicating that reporter intensity can act as a lagging surrogate for true population dynamics (Figure 3.5-B). Together, these results show that adding Occident to the existing generalist segmentation module enables robust segmentation-tracking-quantification analyses of cellular interactions in a single end-to-end workflow.

3.2.8 Spatial and Temporal Activation Pattern of IKK-SPARK and TBK1-SPARK

To characterize IKK β and TBK1 kinase activity at single-cell resolution, we developed a two-stage computational pipeline for automated SPARK reporter quantification. In the first stage (PREDICT module), cells were segmented from time-lapse fluorescence images and a per-cell puncta signal was extracted. In the second stage (VISUALIZATION module), traces underwent quality filtering, and cells were selected to generate population-level overlay traces for each reporter-condition combination. Cell segmentation was performed with CellSAM, a Segment Anything Model variant fine-tuned for fluorescence microscopy, followed by mask post-processing with the Occident tracking pipeline to ensure consistent cell identity across time frames. The SPARK signal at each time frame was computed as the fraction of total cell fluorescence attributable to detected puncta, providing a reporter-expression-normalized measure of kinase activity.

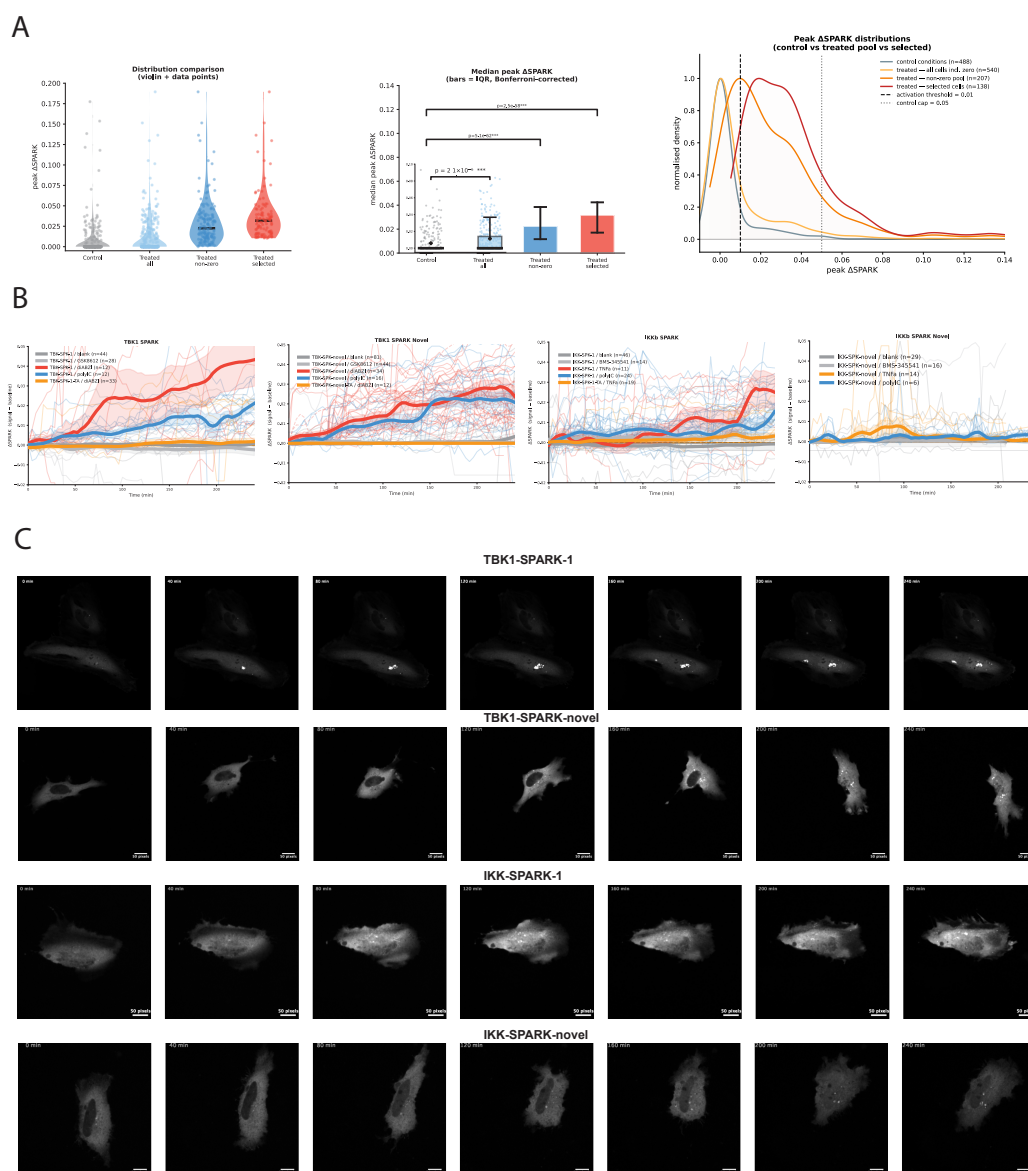


Figure 3.6: KINBERT and phospho-PCA screening results in novel SPARK live-cell reporters for TBK1 and IKK β . A) Left: Violin plots with overlaid jittered data points showing peak Δ SPARK distributions. Mid: Median bar chart with interquartile range whiskers. Right: Kernel density estimates of peak Δ SPARK; dashed and dotted vertical lines mark the activation threshold (0.01) and basal peak cap (0.05), respectively. B) Time-course overlay Δ SPARK traces for selected cells across conditions over 0–240 min. C) Exemplary time-series for all four developed kinase reporters upon activation.

To determine appropriate signal thresholds for cell stratification, we examined the

empirical peak Δ SPARK distributions across all basal-condition and stimulus-treated cells. In basal conditions, the peak Δ SPARK distribution was strongly centered around zero and remained compact below 0.05 across all reporter variants (about 98% on the peak distribution). In treated conditions, approximately half of cells across all pools showed basal spark signal at any timeframe, consistent with the stochastic nature of single-cell NF- κ B and TBK1 activation and the punctum-nucleation threshold behavior of SPARK reporters. A minimum peak Δ SPARK threshold was applied as a signal detection boundary to separate measurable spark activity from residual basal reporter fluctuation. For IKK-SPK-novel in the polyIC condition, automated selection was replaced with manually curated single-cell recordings due to the lower brightness of this reporter variant, and population mean traces were generated from manually validated cells. For TBK-SPK-novel in the diABZI condition, a small number of manually validated cells were appended to the automated selection to supplement the pool. Population mean traces were computed over a common 0–240 min time grid before plotting, with shaded bands indicating the standard error of the mean across contributing cells.

To validate that stimulus treatment produces a measurable increase in SPARK reporter activity above baseline, we compared peak Δ SPARK distributions between control and treated conditions across all reporters excluding the non-quantifiable IKK-SPK-novel. One-sided Mann–Whitney U tests against the control distribution revealed that stimulus treatment produced a significant rightward shift in peak Δ SPARK at every level of the treated cell pool (Figure 3.6-A). Even when considering all treated cells including those with no detectable spark signal, the treated-all distribution was significantly greater than control ($p = 1.7 \times 10^{-7}$, ***). Among the non-zero responding cells, the separation was substantially stronger ($p = 4.2 \times 10^{-63}$, ***, rank-biserial $r = -0.81$, median Δ SPARK = 0.023 vs. 0.000), and was further enriched in the selected sustained-responder subset ($p = 2.1 \times 10^{-59}$, ***, $r = -0.86$, median Δ SPARK = 0.032). Further comparison to previously reported time-resolved measurements demonstrates strong concordance between our dynamic response profiles and established biochemical and imaging evidence. Consistent with prior reports, TBK1 activity exhibits a spatially restricted pattern with Golgi-associated localization and prolonged activation kinetics, persisting on the order of 2–12 hours following stimulation, as observed in both immunofluorescence imaging and western blot analyses (S. Lee et al., 2021; Pourcelot et al., 2016). In contrast, IKK β activation displays a more diffuse, pan-cytoplasmic distribution, aligning with earlier spatial mapping studies and the reported localization of NEMO-

containing signaling assemblies (Malek et al., 2001; Thermo Fisher Scientific, n.d.). The spatial organization captured by our reporters therefore recapitulates known compartment-specific signaling architectures of these kinases (Figure 3.6-B,C).

Notably, our measurements of IKK β dynamics reveal a more sustained and accumulating signal than the rapid activation–dephosphorylation cycles classically inferred from bulk assays. This apparent discrepancy can be explained by two non-mutually exclusive mechanisms. First, the reporter system itself introduces a kinetic bias due to the biophysical properties of phase-separated condensates. As demonstrated in the original SPARK study, condensate maintenance requires a lower threshold of kinase activity than condensate formation, leading to mixed states between activation and dissolution states (Q. Zhang et al., 2018). Specifically, condensates persisted after removal of upstream stimulation in the absence of kinase inhibition, indicating that residual kinase activity—insufficient to initiate new condensates—can nonetheless sustain existing structures. This implies that reporter readouts integrate kinase activity over time and may lag behind biochemical inactivation, resulting in an apparent prolongation of signaling duration. Second, our single-cell measurements resolve substantial cell-to-cell heterogeneity that is inherently obscured in population-averaged assays. Previous work has shown that NF- κ B pathway activation is highly heterogeneous and exhibits digital, all-or-none behavior at the single-cell level, with dose-dependent responses primarily reflecting changes in the fraction of responding cells rather than the duration of signaling within individual cells (Tay et al., 2010). Consequently, the rapid decay observed in bulk measurements arises from averaging asynchronous and heterogeneous responses across the population. More recent studies further emphasize that NF- κ B dynamics are highly variable across individual cells, complicating interpretation of ensemble measurements (Kizilirmak et al., 2022). In this context, our IKK β reporter captures the full distribution of activation durations at single-cell resolution, revealing sustained signaling in subsets of cells that would otherwise be averaged out, thereby reconciling the apparent discrepancy with canonical fast-decay models.

3.3 Discussion

This chapter demonstrates an end-to-end pipeline connecting high-throughput phospho-PCA screening to machine-learning-guided biosensor design and live-cell kinase imaging. The principal strength of the phospho-PCA platform is its ability to generate a dense, high-confidence KSI map without relying on prior substrate annotations. A precision exceeding 0.90 across both sub-libraries means that hits

can be pursued with high confidence, while the ~75% orthogonal confirmation rate by cell-free phosphorylation reaction assays validates the statistical framework on genuinely novel pairs. The inherent trade-off is reduced recall: false-negative interactions are enriched for substrates that either lack a canonical PAABD recognition motif or depend on cellular context—subcellular co-localization, scaffold proteins, or allosteric regulation—that the yeast system does not recapitulate. The Bayesian hierarchical model is critical for making these confidence estimates reliable, because without kinase-level deconvolution, raw enrichment scores conflate enzyme abundance with recognition specificity. The bimodal autonomous-growth distribution observed for the ST sub-library, but not the Y sub-library, highlights a further platform-specific confound arising from endogenous yeast serine/threonine kinase activity; the promiscuity index provides a principled, annotation-free filter against this background. Together, these quality-control layers suggest that phospho-PCA is well suited to producing compact, actionable hit sets for downstream applications.

The substrate-based kinase clustering analysis reveals that phylogenetic distance is a strong predictor of substrate selectivity. This confirms that biochemical substrate profiling captures functional identity complementary to, but not fully predictable from, sequence phylogeny. For understudied dark kinases, the platform’s unbiased substrate pools provide the systematic motif characterization: PHKG1’s extreme basic-residue selectivity, ADCK1’s extended –6 basic preference, and CaMKK1’s canonical CaMK grammar are each consistent with established biology where it exists and generate testable hypotheses.

Fine-tuning KINBERT on phospho-PCA data addresses a previously underappreciated distributional gap in KSI predictors: models trained exclusively on curated database and lysate-derived interactions systematically fail in the broader physico-chemical substrate space probed by pooled screening, performing near chance on the phospho-PCA benchmark before fine-tuning. The improved performance after fine-tuning—without catastrophic forgetting of database-derived knowledge—confirms that the two data modalities are complementary rather than redundant and suggests a hybrid training strategy for next-generation predictors. The downstream utility of this approach is demonstrated directly: the finalized novel SPARK substrate candidates were not seen during training yet received confident positive selectivity scores from the fine-tuned model, in contrast to the broadly negative and uncertain predictions of the pre-fine-tuned model. A limitation is that the model outputs binary phosphorylation classification scores rather than quantitative activity estimates, so

it functions as a selectivity filter rather than a predictor of catalytic efficiency or biosensor dynamic range.

The validation results establish that all four reporter constructs respond to physiologically relevant stimuli above a well-characterized basal distribution, and that their spatial and temporal activation profiles—Golgi-restricted puncta for TBK1 and pan-cytoplasmic signal for IKK β —are concordant with immunofluorescence and biochemical evidence from the literature. The ability to resolve these patterns at single-cell resolution over 240-minute time courses illustrates the practical advantage of continuous live-cell reporters over endpoint assays for kinases whose activation is delayed, transient, and heterogeneous.

Future optimization should focus on tuning HOTag valency, PAABD–phosphopeptide binding kinetics, and expression level to increase dynamic range and reduce the noise floor. Orthogonal validation using non-condensate-based reporters or quantitative phosphoproteomics will be necessary to disentangle reporter-specific kinetic biases from true signaling dynamics. Despite these limitations, the four constructs collectively provide a validated, modular starting point for pharmacological studies, genetic perturbation screens, and further biosensor engineering targeting TBK1 and IKK β .

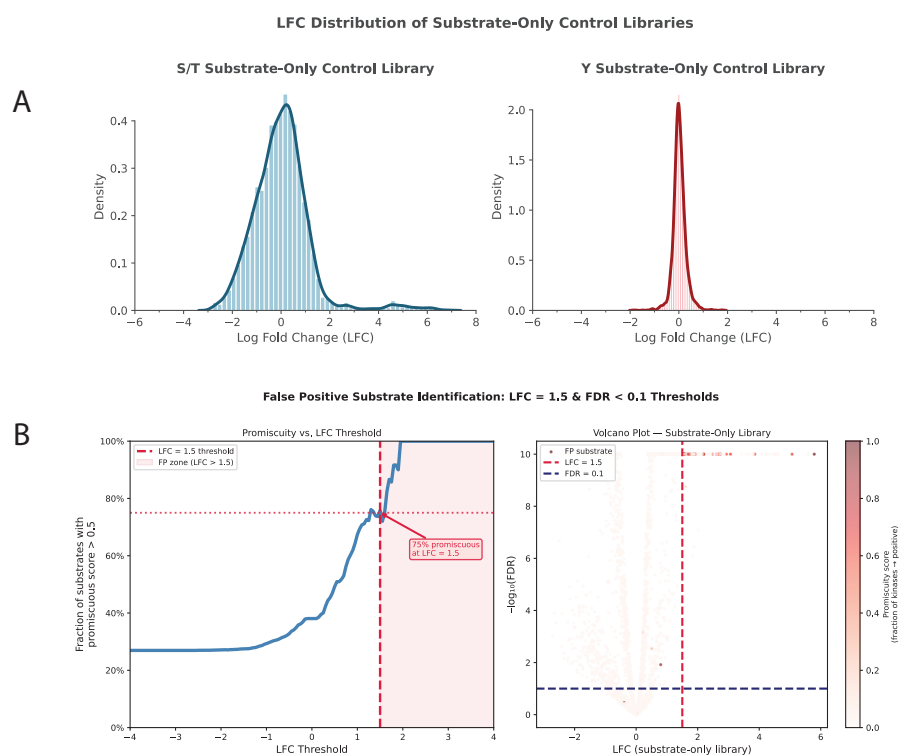


Figure 3.7: Inspection of the substrate-only Libraries and decision boundary for false positive thresholds A: Histogram for LFC distribution of S/T and Y substrate-only libraries B: Elbow Plot and Volcano Plot binning LFC>1.5 as a reasonable FP threshold

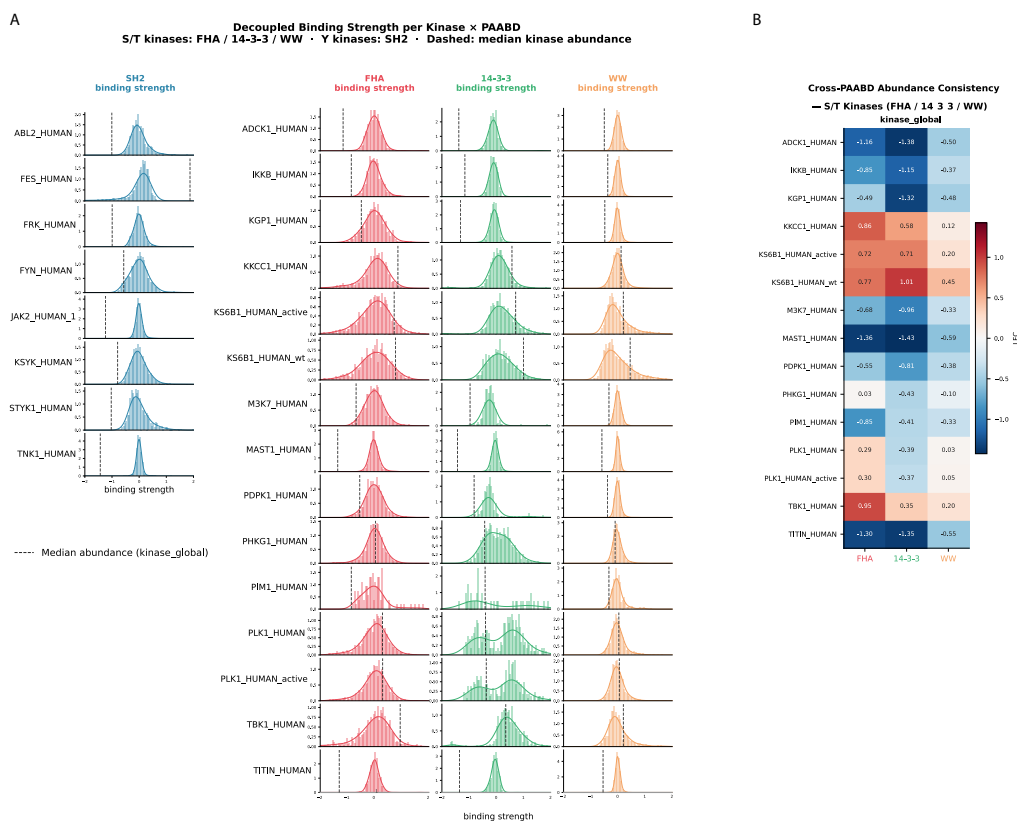


Figure 3.8: **Abundance and Activity Decoupling across PAABDs** A: Histogram for decoupled activity scores for ST/Y kinases, dashed line showing median kinase abundance B: Heatmap for global abundance score for each kinase-PAABD pair

Library	TP	FP	TN	FN _{total}	FP _{sub}	Inactive _{sub}	No PAABD	FN _{eval}	Prec.	Recall
Y	77	5	22	118	4	5	70	39	0.939	0.664
ST	57	6	41	77	8	8	50	11	0.905	0.838
Combined	134	11	63	195	12	13	120	50	0.924	0.728

Table 3.5: phospho-PCA performance statistics on spike-in KSIs.

3.4 Methods

3.4.1 Library Design, Construction, and Screening

Substrate and Kinase Library Composition

A diverse set of 8 tyrosine kinases and 15 serine/threonine kinases was selected for screening. To quality-control pooled screening performance, spike-in control

pairs were included among the screening candidates. From a literature-curated KSI dataset, redundant substrate sequences with a Levenshtein distance less than 7 from any other sequence in the set were first removed; approximately 300 documented kinase–substrate pairs were then sampled from the curated database. To enrich the library with diverse substrates, approximately 1,700 additional candidate substrates were sampled from documented animal phosphoproteomic data, retaining only sequences with a Levenshtein distance greater than 8 from all other library members and centring on either a tyrosine or a serine/threonine phosphoacceptor site.

Library Assembly

Substrates were ordered as oligonucleotide pools and kinases as individual gene blocks synthesized by Twist Biosciences. Kinase gene blocks were pooled equimolarly and crossed with the substrate pool in a Golden Gate reaction; the resulting substrate–linker–kinase fragments were ligated into the yeast expression plasmid. PAABDs were ligated into the same backbone, and 19-nucleotide random barcodes were inserted using Gibson assembly. Transformation efficiency was estimated by 1:1,000 dilution plating, and short-read amplicon sequencing was used to confirm approximately 10× barcode coverage across all variants present in the library.

Yeast Transformation and Selection

Libraries were introduced into yeast strain BY4742 by electroporation and subjected to auxotrophic selection in SC –URA media. A pre-selection cycle (input) was performed by inoculating SC –URA/MET/ADE medium at a starting $OD_{600} = 0.1$ from a saturated culture to allow amplification of the variant pool into exponential growth phase. The competition cycle (output) was initiated by transferring cells from the input culture into competition medium (SC –URA/MET/ADE + 200 $\mu\text{g/mL}$ methotrexate) at a starting $OD_{600} = 0.05$. Plasmids were extracted from pellets collected at the end of both the input and competition cycles, and primers flanking the barcode region were used to generate short-read amplicons for barcode quantification.

Sequencing and Data Analysis

Approximately 20× (10 million reads) PacBio HiFi sequencing coverage was used to generate the variant-to-barcode lookup table, and approximately 100× short-read sequencing coverage was used to quantify barcode abundance in the input and

competition pools. Barcodes were deconvoluted to their original kinase–substrate pair identities, and hits overrepresented in the competition cycle were identified from the read-count table using MAGeCK (B. Wang et al., 2019).

3.4.2 Library Processing and Construction of Barcode–Kinase–Substrate Lookup Tables

PacBio CCS Read Processing

Barcoded kinase–substrate libraries were sequenced using PacBio HiFi circular consensus sequencing (CCS). Reads were imported from compressed FASTA files and standardized to a common orientation. A diagnostic trigger sequence was used to identify reverse-oriented molecules; when present, the entire read was reverse-complemented prior to parsing. This step ensured that all downstream motif searches and coordinate extractions were performed in a consistent direction. Construct architecture was defined by invariant flanking sequences surrounding the barcode and the kinase–substrate (KS) region. Reads were retained only if they contained all expected anchors delimiting: (i) the barcode region, (ii) the KS insert region, and (iii) the internal linker separating substrate and kinase inserts. The nucleotide barcode was extracted between barcode flanks, and the KS region was extracted between KS flanks. The KS region was split at the linker to produce separate substrate and kinase nucleotide inserts. Nucleotide inserts were translated to amino-acid sequences. To ensure library integrity and eliminate sequencing or cloning artifacts, constructs were filtered using the following criteria: (i) barcode length exactly 19 nucleotides, (ii) substrate insert translated to a 15 amino-acid peptide, (iii) absence of internal stop codons in the substrate peptide, and (iv) kinase open reading frame terminated by a single C-terminal stop codon (removed after validation). Reads failing any criterion were discarded. Because a single barcode can occasionally appear associated with multiple inserts due to sequencing or recombination artifacts, a bijective mapping between barcode and (kinase, substrate) pair was enforced. For each barcode, supporting reads for every observed pair were counted and the fraction assigned to the dominant pair was calculated. Barcodes were retained only if $\geq 90\%$ of reads supported a single assignment. The resulting dataset was collapsed to one record per barcode. Validated kinase sequences were matched to a curated kinase reference table to assign kinase identifiers (name, accession, and kinase class such as serine/threonine or tyrosine). The final lookup table therefore linked each barcode uniquely to a substrate peptide and annotated kinase.

Pooled screening libraries were sequenced using Illumina platforms. Per-sample count tables contained observed insert sequences and counts for two experimental conditions: pre-selection (“pre”) and post-selection (“mtx”). Barcodes were extracted from reads using fixed flanking sequences. For certain library variants, a trimmed insert sequence of known length was first recovered and translated to an amino-acid sequence identifier used for aggregation. Low-coverage barcodes are dominated by sampling noise and lead to unstable enrichment estimates. Therefore, for each barcode the combined counts across conditions were required to exceed a minimum threshold: $\text{pre} + \text{mtx} > 100$. Entries below this threshold were removed. When multiple sequencing batches existed, replicate count tables were merged on shared identifiers. Each barcode therefore had replicate-resolved counts: $(\text{pre}_1, \text{mtx}_1, \text{pre}_2, \text{mtx}_2)$.

Barcode count tables were joined to the PacBio-derived lookup tables using the nucleotide barcode as the key. For mapped peptides, motif compatibility annotations were assigned based on the central residue and surrounding positions. These annotations were used only for downstream interpretation and not for model fitting.

3.4.3 False Positive Substrate Identification

To remove non-specific substrates from the KSI dataset, false positive (FP) substrates were identified through two orthogonal criteria applied separately to the serine/threonine (S/T) and tyrosine (Y) arms of the library. For the S/T arm, substrates were first flagged using the substrate-only control library (b1ib7), where constructs expressing substrate peptides without kinases were screened under identical conditions. Only substrates with sufficient read depth ($\text{control_mean} > 10$) were considered, and those showing significant enrichment ($\text{FDR} < 0.1$, $\text{LFC} > 1.5$) were designated as FP candidates (fp_subs_1). This LFC threshold was empirically justified by a promiscuity score analysis: at $\text{LFC} > 1.5$, approximately 75% of flagged substrates could be phosphorylated by more than half of all kinases in the pool, indicating auto-phosphorylation by endogenous yeast kinases rather than specific kinase–substrate interactions. In parallel, a second FP set (fp_subs_2) was derived from the full KSI interaction dataset by computing a promiscuity score for each substrate (defined as the fraction of tested kinases yielding a positive interaction). Substrates with a promiscuity score exceeding 0.9 were excluded as globally non-specific. The final S/T FP list was the union of both sets. For the tyrosine arm, the substrate-only control library (b1ib10) did not exhibit a bimodal LFC distribution, precluding LFC-based thresholding; therefore, only the promiscuity

score criterion was applied. To maintain consistency with the S/T arm—where the LFC = 1.5 cutoff corresponded to ~75% of substrates exceeding a promiscuity score of 0.5—the same 0.75 promiscuity score threshold was applied to the tyrosine library as the equivalent false positive boundary. The combined S/T and Y FP substrate lists were used to annotate all KSI pairs, and a KSI was designated a true positive only if it was screen-positive and its substrate was absent from the FP list.

3.4.4 Statistical Modeling and Decoupling of Enrichment Signals

Barcode counts were aggregated to substrate–kinase pairs by summing across barcodes within each replicate. To account for differences in sequencing depth across replicates and conditions, size factors were computed separately for pre- and post-selection libraries from total counts. A log-offset term was constructed for each substrate–kinase–replicate combination, incorporating both the post-selection size factor and the pre-selection representation of each pair, so that the model baseline reflects expected post-selection expression given pre-selection input.

Post-selection counts were modeled using a hierarchical negative binomial likelihood, with the log mean decomposed as:

$$\log \mu(s, k, r) = \text{offset}(s, k, r) + g_k + \delta(k, r) + b(s, k) \quad (3.1)$$

where g_k is a global kinase effect, $\delta(k, r)$ is a replicate-within-kinase deviation, and $b(s, k)$ is a substrate–kinase binding effect. The kinase-level term absorbs systematic differences in enrichment attributable to kinase expression, catalytic efficiency, or assay conditions, while the binding term captures substrate-specific interaction above or below that baseline. Weakly informative normal priors were placed with replicate deviations ($\sigma_\delta = 0.5$) and kinase or binding effects ($\sigma_g = \sigma_b = 1.0$). Posterior inference was performed using Hamiltonian Monte Carlo with the No-U-Turn Sampler (NUTS) as implemented in PyMC.

3.4.5 Kinexus Peptide Array Kinase Assay

Validation peptides (15 amino acids each) were synthesized via SPOT technology on glycine-modified cellulose membranes using a MultiPep synthesizer (Intavis). Each peptide incorporated a branched Lys(Lys)₂ tag at the N-terminus to promote oriented binding onto glass slides. Peptides were cleaved from cellulose with gaseous ammonia, extracted into DMF/PBS (0.05% PEG-6000), and printed in duplicate onto amino-reactive Nexterion H glass slides (Schott) at 0.1 μ l per spot with 290 μ m spacing. Slides were blocked with borate buffer containing ethanolamine to quench

unreacted surface groups. Each slide was then incubated with 2 μg of kinase in 1.6 μl MOPS buffer supplemented with ATP for 2 hours at 30°C in a humidity chamber. Control slides were treated with MOPS buffer alone or MOPS buffer containing ATP without kinase; additional negative controls consisted of “empty” cellulose spots processed identically to peptide-containing spots. Phosphorylation was detected using Pro-Q Diamond stain (Invitrogen) with 1 hour incubation followed by repeated destaining. Fluorescence signal was acquired at 543 nm and images were exported as TIF files. Spot quantification was performed using ImageGene (BioDiscovery), capturing Mean, Median, and Mode intensity values per spot. For downstream analysis, Mean values from duplicate spots were averaged, background-corrected using negative control signals, and normalized to the maximum signal on each array to yield a relative phosphorylation intensity. A specificity score was further computed by normalizing each peptide’s intensity to its average signal across all kinase-treated arrays processed on the same day.

3.4.6 Live-Cell Imaging Conditions

SPARK reporter constructs were transfected into HeLa cells using TransIT reagent with DNA plasmids. Cells were incubated overnight prior to imaging and stimulus dosing to achieve minimal background signal and capture full activation dynamics. Imaging was performed on a Nikon confocal microscope at 40 $\times$ magnification (pixel size = 0.275 μm).

Dosing conditions

- **diABZI STING agonist-1** (final concentration 2 μM ; MW 849.9 g/mol): stock prepared at 1 mM by dissolving 1 mg in 1,176 μl H₂O (500 $\times$ stock). For dosing: 1.2 μl of 1 mM diABZI added to 100 μl of conditioned DMEM taken from the well.
- **TNF α** (final concentration 1 ng/ml; stock 100,000 $\times$): diluted 1:1,000 in DMEM, then 2 μl added per 200 μl DMEM per well.
- **Poly I:C** (final concentration 600 ng per 200 μl well, i.e., 3 ng/ μl): for 10 wells, 6 μg Poly I:C (2 μl of a 3 $\mu\text{g}/\mu\text{l}$ stock) was combined with 100 μl OptiMEM, 2 μl DNA, and 12 μl TransIT, incubated 30 minutes, and 10 μl of the resulting polyIC:TransIT mix added to 100 μl conditioned DMEM taken from each well.

- **GSK-8612** (TBK1 inhibitor): 1 μ l of 2 mM stock added to 200 μ l culture.
- **BMS-345541** (IKK β inhibitor; final concentration 10 μ M): stock prepared at 1 mM (100 $\times$) by dissolving 1 mg in 3.43 ml H₂O; 1–2 μ l of 2 mM stock added to 200 μ l culture.

3.4.7 Single-Cell SPARK Quantification Pipeline

Single-cell SPARK signal quantification was performed using a two-stage computational pipeline: a PREDICT module for cell segmentation and per-cell signal extraction, and a VISUALIZATION module for quality filtering, cell selection, and population-level overlay trace generation. SPARK reporters targeting IKK β (IKK-SPK-1, IKK-SPK-novel) and TBK1 (TBK-SPK-1, TBK-SPK-novel) were imaged across vehicle control (blank), kinase inhibitor (BMS-345541, GSK-8612), and activating stimulus (TNF α , polyIC, diABZI) conditions, alongside T-to-A phosphosite negative controls (IKK-SPK-1-TA, TBK-SPK-1-TA, TBK-SPK-novel-TA). Cell segmentation was performed using CellSAM (Cell Segment Anything Model), a fine-tuned variant of the Segment Anything Model adapted for fluorescence microscopy, with detection thresholds of `bbox_threshold = 0.38`, `mask_threshold = 0.079`, and `iou_threshold = 0.48`. Raw fluorescence images were intensity-clipped at 3,000 counts prior to inference to normalize for cross-FOV brightness variation.

Raw segmentation masks were post-processed using the Occident tracking pipeline with the following customizations: objects smaller than 5,000 pixels were removed; overlapping masks were resolved using intersection-over-union ($\text{IoU} \geq 0.3$) and intersection-over-query ($\text{IoQ} \geq 0.2$) thresholds. The resulting mask array assigned a consistent integer cell identifier to each pixel across all time frames. Per-cell SPARK signal was quantified at each timeframe by detecting fluorescent puncta within the cell mask. Puncta were identified as pixels exceeding the 99th percentile of the per-cell intensity distribution, and candidate objects were retained if they met minimum area (5 pixels) minimum mean intensity (2,600 units) criteria. The absolute intensity gate was calibrated to the brighter signal range of the standard reporters and served as a specificity filter against background fluctuation. For a subset of IKK-SPK-novel polyIC and TNF α fields of view where automated detection was unreliable due to the lower brightness of this reporter variant, per-cell and per-frame intensity thresholds were set manually following visual inspection of raw puncta. The SPARK signal at each timeframe was computed as the ratio of total puncta fluorescence to total cell fluorescence, providing a measure of kinase-active puncta burden normalized for

cell-to-cell variation in reporter expression level. Quantified single-cell traces from all runs were consolidated into a unified dataset. Frame-rate corrections were applied based on imaging protocol: runs acquired at 5 min/frame (49 frames, 240 min total) and 3 min/frame (81 frames, 240 min total) were corrected accordingly; all remaining runs used a standard 4 min/frame interval. Each cell's SPARK signal was baseline-normalized by subtracting the pre-stimulus frame-0 value, yielding Δ SPARK as the change in puncta burden relative to the resting state. Cells were required to have valid measurements in at least 85% of expected timeframes, ensuring only continuously tracked cells contributed to downstream analysis. Fewer than 3% cells with high pre-stimulus signal prior to stimulus addition were excluded. Cells with artefactual high-frequency signal oscillations were identified and removed prior to selection. Approximately 18% of cells were removed across the full dataset. Among non-zero cells, a minimum peak response threshold was applied to separate true activation from residual basal activity, and up to 30 cells per condition were selected by descending mean signal over the final two-thirds of the time course to prioritize sustained activation. For IKK-SPK-novel - polyIC condition, automated selection was replaced with manually validated recordings due to puncta intensities approaching the noise floor of this reporter variant, likely a consequence of weaker WW domain binding affinity and non-optimized transfection efficiency. To further validate the automated pipeline, for the TBK-SPK-novel diABZI condition, $n = 4$ manually curated traces were appended to the pipeline-selected pool and corresponds well with auto-detected traces.

Data and Materials Availability

Screening sequencing data information and Code for all analysis pipelines is available at https://github.com/steveyu323/kinase_substrate_interaction.git.

Chapter 4

UNCOVERING KINASE STABILITY, ACTIVITY, AND SUBSTRATE SPECIFICITY WITH DOUBLE-DEEP PHOSPHO-PCA

Abstract

Protein kinases relay information to various cellular processes, and their dysregulation underlies numerous human diseases. Despite their importance, our understanding of how kinase domains and their variants impact protein stability, catalytic activity, and substrate recognition is incomplete. In this work, we develop the phosphorylation protein complementation assay (phospho-PCA), which enables quantitative measurements of kinase-substrate interactions by coupling them to the growth of budding yeast, thereby enabling deep mutational scanning of kinase domains. When combined with deep mutational scans targeting folding stability and Bayesian modeling, phospho-PCA can disentangle the relative impact of mutations on kinase domain stability, catalytic activity, and substrate specificity. We demonstrate the accuracy and breadth of phospho-PCA, showing its applicability to both tyrosine and serine/threonine kinase domains. We then apply our method to three closely related protein kinases with distinct substrate preferences, evaluating over 15,000 kinase variants against a panel of three substrates for both catalytic activity and substrate specificity. The resulting dataset constitutes one of the largest variant-to-function maps assembled for this enzyme family to date, revealing numerous mutations that alter kinase activity and substrate specificity. Physics-based modeling reveals how these mutations operate through diverse mechanisms, including long-range allosteric communication, to alter both activity and substrate specificity. Given its scalability, we believe phospho-PCA can measure the functional impact of variants across the entire human kinome.

4.1 Introduction

As thoroughly reviewed in Chapter 1, Protein kinases selectively phosphorylate substrates to regulate enzymatic activity, cellular localization, and protein interactions (Roland Akinseye, 2024). Dysregulation through missense variants drives diseases (Chitluri & Emerson, 2024; Medvedev et al., 2023; Torkamani et al., 2009; L. Wang et al., 2021), motivating the need for precise understanding of how variants affect

catalytic activity and substrate recognition (Y. Liu et al., 2025; Martinez et al., 2020; X. Xu et al., 2019). A key gap remains in understanding how kinase domains themselves contribute to substrate specificity. Domain swapping experiments on receptor tyrosine kinases show that phosphorylation profiles follow the kinase domain rather than the base receptor (Cai et al., 2017; Duplay et al., 1996; Nagao et al., 2021; Nakamura et al., 1996; Piccinini et al., 2002; Santoro et al., 1994), and single-point mutations in kinase domains can alter substrate preferences and drive pathogenicity (Cuesta-Hernández et al., 2023; Lubner et al., 2017; Özden et al., 2022; N. H. Shah et al., 2018; Songyang et al., 1995; Torres Robles et al., 2025; Walker et al., 2019). Evolutionary analyses further suggest kinase domains originated from a promiscuous ancestor and diverged to acquire distinct substrate preferences (Finneran et al., 2020; Gizzio et al., 2022; Taylor & Kornev, 2011). Existing methods for profiling kinase-substrate interactions — including peptide positional scanning libraries (Johnson et al., 2023; N. H. Shah et al., 2018; Songyang et al., 1994; Yaron-Barir et al., 2024) and deep mutational scanning of kinases (Ahler et al., 2019; Beltran et al., 2024; Brenan et al., 2016; Estevam et al., 2025; Hayes et al., 2024; Hobbs et al., 2022; C. J. Howard et al., 2025; Sokirniy et al., 2025) — cannot assess variant effects on substrate specificity in parallel, nor disentangle changes in protein abundance from changes in catalytic activity (Fowler & Fields, 2014; Jehle et al., 2022; Wei & Li, 2023; Xue & Tao, 2013).

In this work, we were able to generate the first, comprehensive maps of how mutations within the kinase domain impact both catalytic activity and substrate recognition. We focus on Fyn, Lck, and c-Src of the Src family kinases (SFKs). These three non-receptor tyrosine kinases share $\approx 70\text{--}80\%$ sequence similarity but play distinct roles in regulating differentiation, immune responses, and cell growth. Fyn has a broad tissue distribution and is involved in T-cell signaling (Groves et al., 1996; Peng & Fu, 2023; L. Wu et al., 2023) and neuronal development (G. Lee et al., 2004; White et al., 2008); it is currently being investigated as a target for neurodegenerative diseases (Matrone et al., 2020; Um et al., 2012). Lck is expressed in T-cells and is a key component of the T-cell activation signaling cascade (Horkova et al., 2023; Iwashima et al., 1994; McNeill et al., 2007; K. Shah et al., 2021). c-Src is one of the most well-studied oncogenes; it is widely expressed and regulates cell adhesion, proliferation, and survival (Reynolds et al., 2014). The different roles these kinases have assumed are partly due to their varying substrate preferences. Variants of SFKs have been linked to numerous clinical conditions, including immunodeficiency and cancer (Hauck et al., 2012; Irby et al., 1999; Lanz et al., 2024; Lui et al., 2024;

Palomero et al., 2014; Turro et al., 2016).

Here, we develop the phosphorylation protein complement assay (phospho-PCA), which uses a growth-based biosensor to quantitatively couple kinase-substrate interactions to the growth of budding yeast. We demonstrate that phospho-PCA enables deep mutational scanning of kinase domains, using the interaction strength with a specific substrate as the readout. Furthermore, we demonstrate that combining mutational scans from phospho-PCA and abundance-PCA, which measures folding stability of protein variants, with statistical modeling enables us to quantify the relative impact of mutations on kinase domain stability, kinase activity, and substrate preference. We perform mutational scans with abundance-PCA and phospho-PCA of Fyn, Lck, and c-Src kinase domains, measuring the impact of ≈ 5000 single amino acid variants ($\approx 98\%$ coverage) for each kinase when paired with representative Fyn, Lck, and c-Src substrates. We demonstrate that evolutionarily conserved sites are associated with significant decreases in abundance-PCA signal, and that the combination of chemical properties and surface accessibility explains the impact of mutations on folding stability. We then identify activity-biasing mutations using decomposed phospho-PCA scores and show that our data can explain the impact of clinical variants. We further utilize these decomposed scores to identify substrate-specificity-determining residues and to conduct a molecular dynamics study to understand, mechanistically, how mutations at these residues affect substrate recognition. This work represents a significant advance for the field, as it provides not only mechanistic insights but also valuable resources in terms of methodology and data. Phospho-PCA and the associated computational methods have the potential to generate mechanistic insights for the entire human kinome. Moreover, our comprehensive dataset can support the development and benchmarking of new methods for variant-effect prediction.

4.2 Results

4.2.1 Phospho-PCA enables deep mutational scanning of kinase domains

Phospho-PCA uses a growth-based biosensor to couple kinase-substrate interactions to the growth of budding yeast cells; a schematic is shown in Figure 4.1. This biosensor consists of two separate proteins expressed in tandem: the first is a kinase domain tethered to its substrate and one-half of a split dihydrofolate reductase (DHFR) enzyme. The second is a phospho-amino-acid binding domain (PAABD) tethered to the other half of the split-DHFR enzyme (Rochette et al., 2015; Shibasaki et al., 2008). When a kinase-substrate interaction occurs, the substrate is

phosphorylated, leading to the binding of the PAABD to the substrate. This binding reconstitutes the DHFR enzyme, conferring resistance to methotrexate. We note that this construct design was inspired by fluorescent kinase biosensors, which use a similar mechanism to couple kinase substrate interactions to changes in fluorescence (Greenwald et al., 2018; J.-f. Zhang et al., 2022). By removing the PAABD, the DHFR fragments complement spontaneously in the cytosol at a rate proportional to the concentration of folded kinase-DHFR1 fusion protein (Figure 4.1a), thus enabling abundance-PCA measurements (Faure et al., 2022). To perform deep mutational scanning, we first cloned a library of kinase variants (made with one-pot saturation mutagenesis) into the kinase portion of our biosensor. Because kinase domains are too long (750–900 bp) for short-read sequencing, we stochastically link the library to short-read nucleotide barcodes and perform long-read sequencing to construct a lookup table (Frank et al., 2022; Larsen et al., 2024; Petersen et al., 2024). Short-read sequencing is used to measure the abundance of each variant after methotrexate selection (Figure 4.1b).

To validate phospho-PCA, we measured growth curves of budding yeast expressing our biosensor construct for a collection of kinase-substrate pairs; the results are shown in Figure 4.2. We observed significantly higher growth for interacting kinase-substrate pairs than non-interacting pairs (Figure 4.2a). We used both catalytically dead kinase domains and mismatched substrates to create non-interacting pairs. We validated that our growth-based biosensors worked for the Fyn, Lck, and c-Src tyrosine kinases, a collection of serine-threonine kinases, and for a panel of phospho-amino acid binding domains (PAABDs) (Figure 4.2a,c). Importantly, by quantitatively extracting growth rates during the exponential phase of cell growth, we observed that our biosensors recapitulated the known preferences of Fyn, Lck, and c-Src for their respective substrates (Figure 4.2b). While our study focuses on the three SFKs, the data for serine-threonine kinases is a crucial demonstration of phospho-PCA's scalability. These results demonstrate that we can utilize our growth-based biosensors to quantify the strength of interaction for almost any kinase and substrate during cell growth.

With our growth-based biosensors validated, we constructed libraries to explore the impact of variants on activity and substrate recognition for the SFKs Fyn, Lck, and c-Src. For each kinase, we generated a library of all possible single-site variants and paired it with one potential substrate (Fyn-sub, Lck-sub, and c-Src-sub). We constructed 18 biosensor libraries that covered phospho-PCA and abundance-PCA

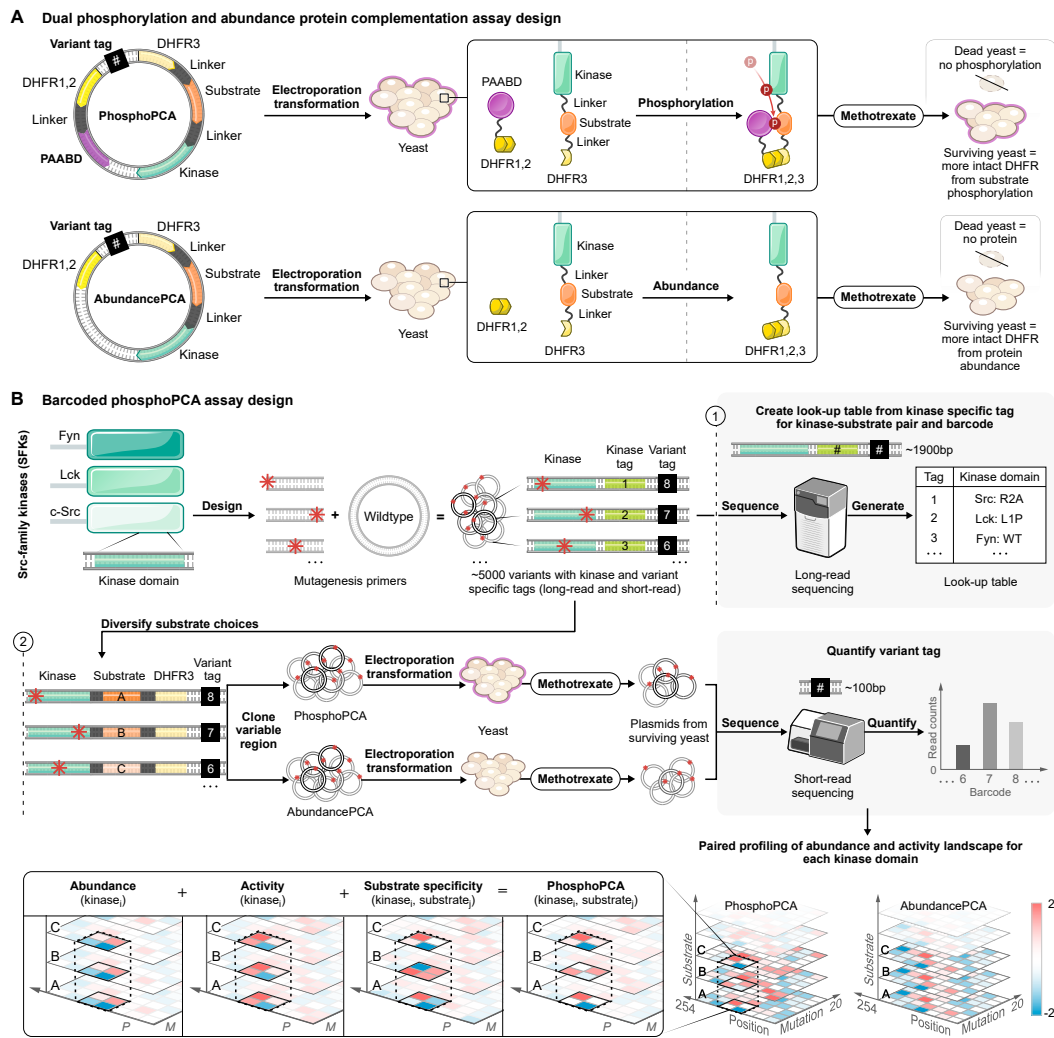


Figure 4.1: Phospho-PCA enables deep mutational scanning of kinase domains to understand the sequence-to-function relationship for this enzyme category. a) Phospho-PCA utilizes a growth-based biosensor to couple kinase-substrate interactions to cell growth. Kinase-substrate interactions reconstitute the DHFR enzyme, conferring resistance to methotrexate selection. Abundance-PCA can be performed with the same construct by removing the PAABD, providing a score of folding stability. b) Deep mutational scanning of kinase domains with phospho-PCA is performed by cloning a kinase variant library into the biosensor construct and barcoding the variants with 16 bp-long barcodes. Long-read sequencing creates a lookup table that links kinase variants to barcodes. Once cloned, the library is transformed into budding yeast cells, and the activity of each variant for phospho and abundance-PCA is measured by quantifying cell growth in the presence of methotrexate with short-read sequencing.

for all possible kinase-substrate pairs. These libraries were screened by measuring growth rates under methotrexate selection using short-read sequencing. In-depth descriptions of our library assembly method, analysis of each library's mutation composition and coverage, methotrexate selection, protocols, and bioinformatic analysis of the sequencing data are given in the Methods section.

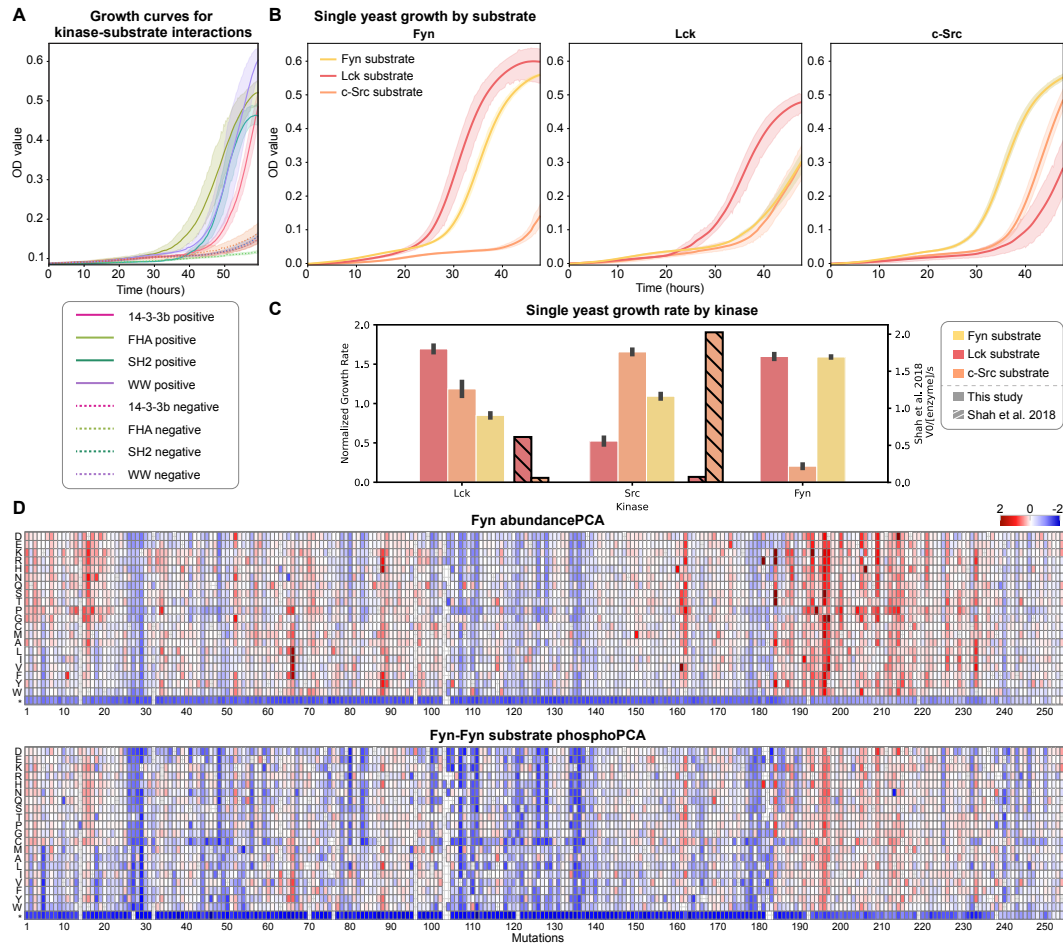


Figure 4.2: Phospho-PCA quantitatively couples kinase-substrate interactions to cell growth. a) Biosensors containing interacting kinase-substrate pairs resist selection, while biosensors with non-interacting pairs display significantly reduced growth. b, c) Quantitative measurement of Fyn, Lck, and c-Src substrate preferences for Fyn-sub, Lck-sub, and c-Src-sub with phospho-PCA. Growth rates were computationally extracted during the exponential phase of cell growth. Scores among all substrates for each kinase were standardized by z-score normalization and then rescaled to a range of 0 to 2. d) Heatmaps for abundance- and phospho-PCA measurements for Fyn and the Fyn/Fyn-sub interaction, respectively.

4.2.2 Structural context determines the impact of mutations on SFK abundance

To investigate the impact of mutations on SFK folding stability from a biophysical perspective, we performed a principal component analysis of our abundance-PCA data. We note that abundance-PCA captures the impact of variants on folding stability, which reduces the amount of folded kinase inside the cell (Faure et al., 2022). For simplicity, we refer to a variant's impact on abundance-PCA scores, and hence folding stability, as “abundance” for the rest of this chapter. c-Src has been screened for stability in the past and the AbundancePCA score shows high alignment and overlapping sets of GoF/LoF residue when compared to previously published work (Figure 4.6). For this analysis, we first grouped each amino acid (excluding proline, cysteine, and glycine) into four biophysical categories: hydrophobic, polar uncharged, positively charged, and negatively charged. We then compiled the biophysical category and abundance data for all mutants across all positions and kinases and performed principal component analysis on the resulting matrix; the results are shown in Figure 4.3a and 4.3b. To simplify our description of the results, we adopt the convention of Modi and Dunbrack for the relevant sections of the kinase domain (Modi & Dunbrack, 2019a). This convention defines 17 key sections throughout the kinase domain—8 in the N-terminal domain (B1N, B1C, B2, B3, HC, B4, B5, HD) and 9 in the C-terminal domain (HE, CL - catalytic loop, ALN - activation loop N terminus, ALC - activation loop C terminus, HF, FL, HG, HH, HI). Each section is named for its dominant secondary structure or its functional role. Throughout this work, residue numbers in the text refer to positions within the isolated 254-residue kinase domain construct, which can be mapped to the full sequence by adding the appropriate offset for each kinase (269 for Fyn, 260 for Lck, and 244 for c-Src).

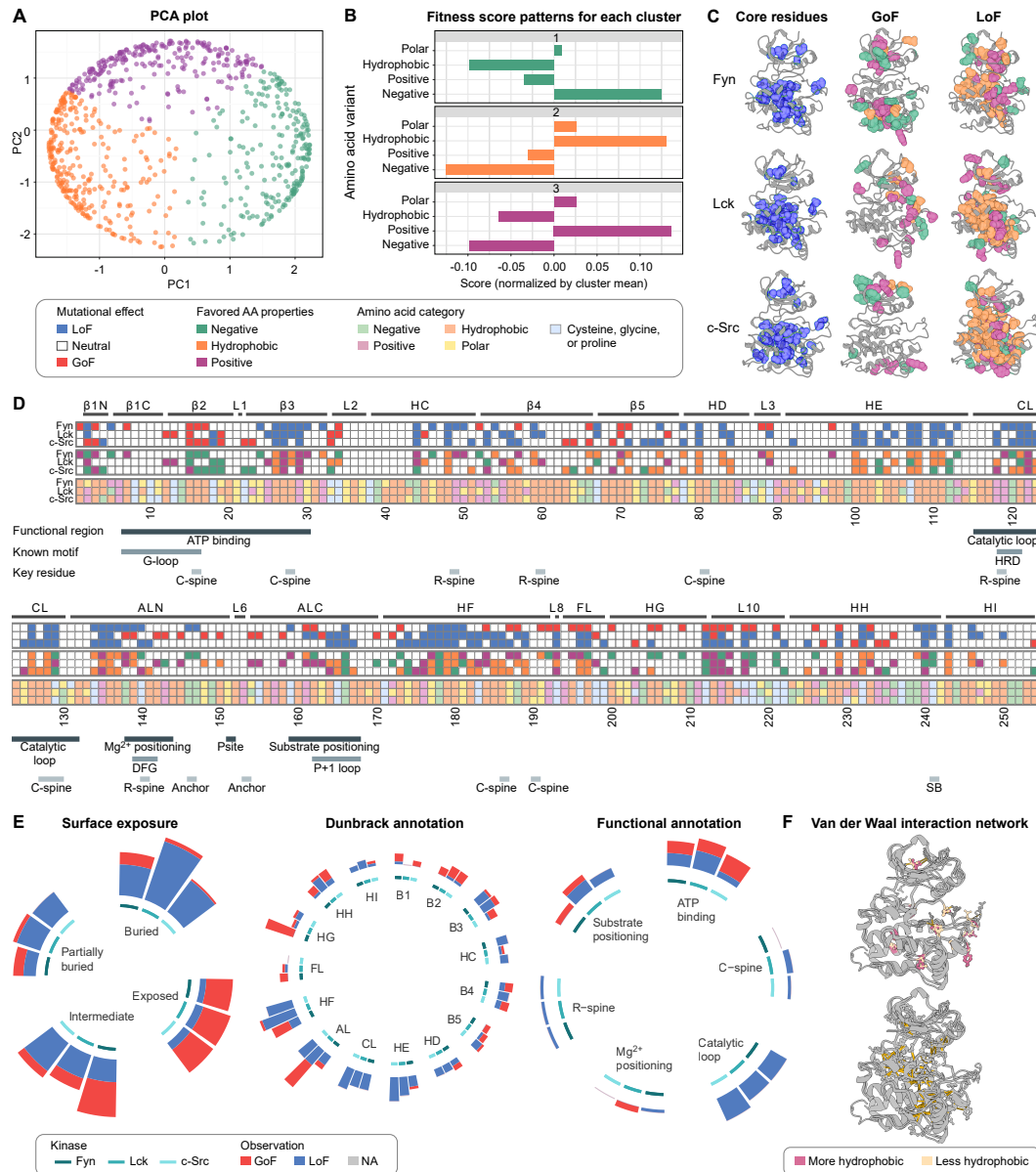


Figure 4.3: Mutations in conserved regions impact abundance-PCA scores of SFKs. a–b) Principal component analysis reveals three distinct responses of positions to mutations. Group 1 and group 3 positions prefer negatively or positively charged residues, respectively, while group 2 positions prefer hydrophobic residues. c) Positional response categories overlaid on the kinase structure. d) Functional annotations and mutational effects for the kinase domains of Fyn, Lck, and c-Src. The six rows display, for each position, the Dunbrack annotation, whether the amino acid property is conserved across the three SFKs, the overall mutational effect, the favored amino acid property, the wild-type amino acid and its category, and a functional annotation. e) Gain-of-function and loss-of-function positions categorized by surface exposure, Dunbrack annotation, and functional annotation. f) Disruption of van der Waals networks explains the impact of mutations on abundance-PCA scores. Top: Kinase-specific van der Waals networks govern solvent accessibility and mutation tolerance. Bottom: Destabilizing sites conserved across SFKs disrupt van der Waals networks in the hydrophobic core.

Using the principal components, we identified three clusters where the presence of negatively charged (1), hydrophobic (2), or positively charged residues (3) leads to a gain-of-function effect in abundance-PCA, whereas alternative properties result in a loss-of-function effect. We then sought to use this clustering, in conjunction with physical and structural reasoning, to explain the impact of convergent and divergent mutations (i.e., mutations with the same or different effects across all three SFKs, respectively). With this goal in mind, we made several observations. First, we observed that most conserved abundance loss-of-function mutations occurred within the hydrophobic core, although some also occurred in structurally distinct regions, such as the HE region and the catalytic loop. Regions with reduced solvent-exposed surface area (SASA, $< 10 \text{ \AA}^2$) were enriched for abundance loss-of-function mutations. In contrast, regions with more solvent exposure were enriched for gain-of-function mutations (Figure 4.3c). These observations were consistent with our clustering analysis; cluster 2 sites were located in the hydrophobic core, while cluster 1 and 3 sites were located at the surface (Figure 4.3c). We also analyzed the impact of mutations on abundance-PCA in functionally important, evolutionarily conserved regions (Figure 4.3d and 4.3e). We observed that each functional region had a different mutational tolerance. The catalytic loop, which is important both functionally and structurally, and residues in the N-terminal portion of the activation loop that are responsible for magnesium positioning, were enriched for loss-of-function mutations. In contrast, the region responsible for ATP binding contained both gain-of-function (B2) and loss-of-function (B3) mutations.

We next sought to understand the impact of divergent abundance mutations. We observed significant divergence for abundance gain-of-function mutations across the three SFKs; these sites were enriched in HF-HG for Fyn, AL for Lck, and B2 for c-Src. We also observed that the sites for these divergent mutations overlapped with regions where the electrostatic charge distribution differed considerably between these three kinases (Figure 4.3d). A closer inspection of these locations revealed a consistent relationship between the physicochemical properties of amino acids and the phenotypic outcomes of mutations. Specifically, at several positions, the introduction of a charged residue that is compatible with the local electrostatic environment of one kinase led to a gain-of-function, while being neutral or detrimental to the other kinases that had a different charge distribution at that site. These sites include positions 172 in cluster 1 (Lck) and 76, 88, 154, and 236 in cluster 3 (Lck and Fyn). Similarly, divergent loss-of-function mutations also reveal location-specific preferences for electrostatic charge. These include mutations in positions 4 in cluster

1 (Src) and 247 in cluster 3 (Lck), suggesting that these kinases evolved to favor negative and positive charges at these positions, respectively.

Finally, we observed that networks of van der Waals interactions were a significant driver of divergence, as they impact the degree of solvent accessibility for specific residues and hence sensitivity to mutations. For example, at position 28, interactions present in Fyn and Lck bury a hydrophobic residue, leading to a loss of function when disrupted. In contrast, c-Src lacks these interactions, leading to increased solvent exposure and greater tolerance to mutation. Identified sites are visualized in Figure 4.3f; a comprehensive list is provided in Table 4.2. When examined together, these data present a consistent picture in which the structural context of a residue determines the impact of mutation on SFK abundance. Disruption of residues within the hydrophobic core reduces abundance, whereas mutations that increase the likelihood of favorable solvent interactions increase abundance. These effects can be brought about by directly altering the physicochemical properties of a residue or by indirectly disrupting van der Waals networks that control surface exposure. We note that these observations are consistent with prior work, which demonstrates that solvent interactions can enhance a protein's thermostability and solubility (Y. Zhou et al., 2019). Overall, these new findings highlight that the effects of variants on SFK abundance are governed by the local structural environment—particularly solvent accessibility, charge compatibility, and van der Waals networks.

4.2.3 A hierarchical Bayesian model separates the contribution of kinase abundance, activity, and substrate specificity to phospho-PCA scores

In this section we developed a novel computational method to understand the mechanistic impact of variants on kinase function from our paired phospho and abundance-PCA data. Understanding the mechanism behind a variant's impact is challenging because there are often multiple pathways a variant can take to perturb a protein's function. For kinases, we envision three mechanistic paths through which mutations can impact the phospho-PCA score, our measure of kinase function. First, a variant could alter folding stability, thereby impacting its abundance in the cell. This would directly affect both the phospho-PCA score (assuming rate-limited kinetics) and the abundance-PCA score. Second, a variant could drive a dynamic shift towards either an active or an inactive conformation. Third, a mutation could alter a kinase's substrate specificity. To understand the contributions of each of these possible mechanisms to our measured phospho-PCA scores, we developed a hierarchical Bayesian model. This model, displayed in Figure 4.4a and described in full detail

in the supplement, posits that phospho-PCA scores can be decomposed into a linear sum of scores that capture a variant's impact on abundance (derived from scaled abundance-PCA measurements), kinase activity (shifts towards active or inactive conformations), and substrate specificity. We reasoned that some of these scores should be shared between conditions. For instance, scores for a variant's impact on abundance should be shared across substrate identity, as the folding stability of the kinase domain should be independent of which short peptide substrate it is tethered to. We confirmed this expectation with exploratory analysis; we note that this effectively allows us to treat the three substrate-specific abundance measurements as technical replicates of the same underlying abundance parameter. Similarly, kinase activity should also be shared across substrate identity. Scores for substrate specificity, however, are expected to vary among all four experimental variables (kinase, substrate, position, and mutation). This expectation can be viewed as a form of regularization that we incorporated into our model. We fit this model to our data using Markov chain Monte Carlo (Abril-Pla et al., 2023); an example decomposition is shown in Figure 4.3b. One advantage of the Bayesian framework for modeling is its ability to quantify uncertainty, which enables the identification of key residues that influence kinase activity and substrate preferences. We note that this decomposition was made possible by the presence of the abundance-PCA data, which was a significant prior advance (Gerasimavicius et al., 2023). Although computational methods for decomposition paired protein complementation assay data exist (Faure & Lehner, 2024; J. Rao et al., 2025), our need to separate the impact of three distinct mechanisms required a bespoke modeling effort. Heatmaps of decomposed scores for kinase abundance, activity, and substrate specificity are shown in Figure 4.4b and Figure 4.12. Our modeling effort is a significant advance, as, to our knowledge, this is the first time an effort has been made to model the relative contributions of variants to abundance, catalytic activity, and substrate specificity for any enzyme.

4.2.4 Computational decomposition reveals conserved activity-biasing variants

The presence of decomposed kinase activity scores for all three SFKs provides an opportunity to explore how variants impact catalytic activity. Indeed, our analysis of these kinase activity scores identified a set of residues that primarily control SFK catalytic activity, distinct from those essential for abundance. We reasoned that these variants exerted their effect either by directly impacting catalysis or by biasing the kinase toward specific conformations, and used this hypothesis to structure our

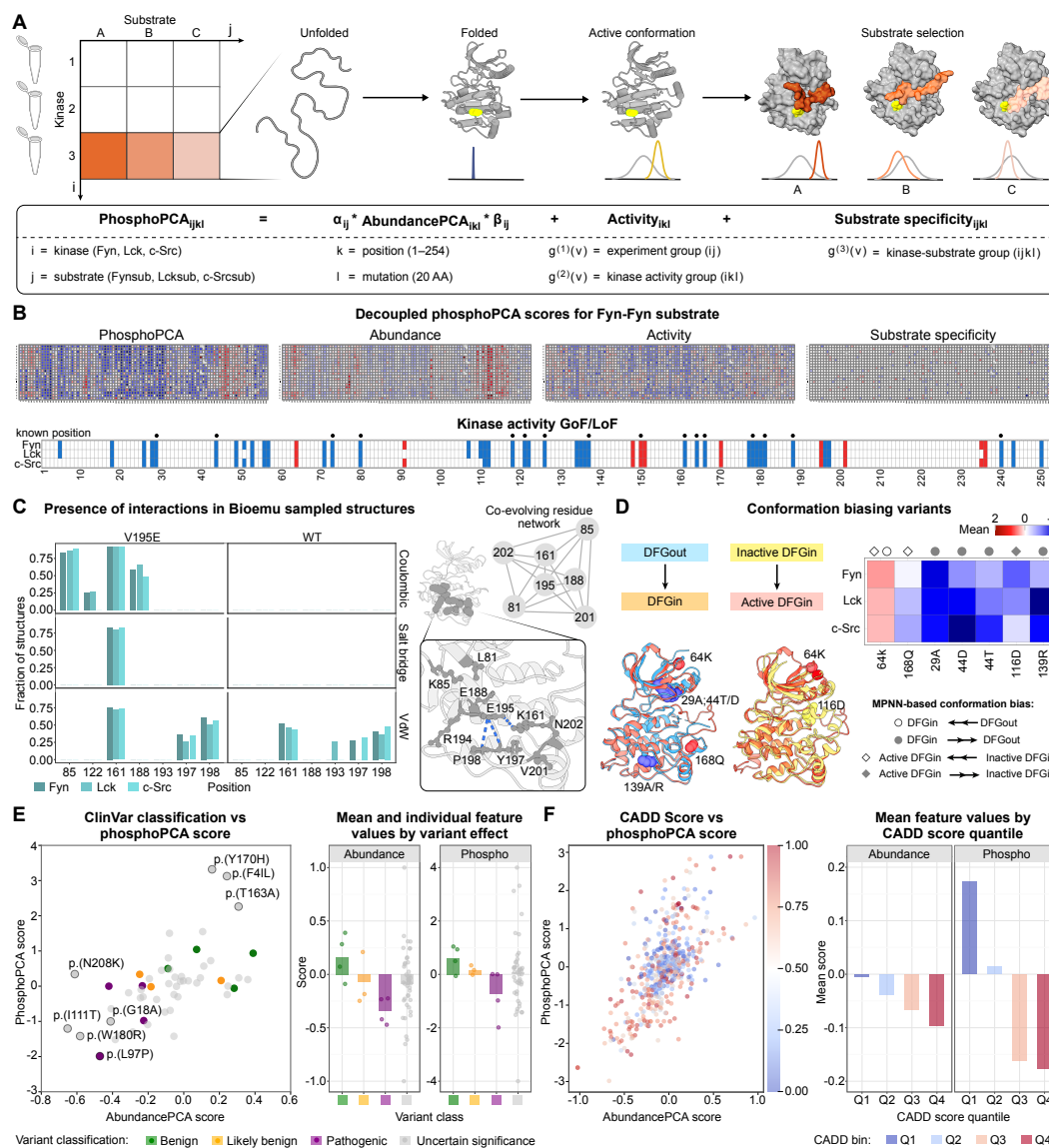


Figure 4.4: Phospho-PCA reveals activity-biasing mutations and the functional impact of clinical variants. a) Hierarchical Bayesian model schematic for decomposing phospho-PCA scores into abundance, kinase activity, and substrate specificity components. b) Example decomposition for Fyn–Fyn substrate pair. c) Conserved gain-of-function mutation V195E reveals a co-evolutionary network. Left: BioEmu conformational sampling reveals interactions among coevolving residues in V195E and wild-type kinases. Right: Network diagram showing V195E establishing new electrostatic interactions that stabilize active conformations. d) ProteinMPNN-based conformational bias analysis identifies mutations favoring different kinase states. e) Clinical variant analysis using ClinVar. Benign variants cluster near the origin; pathogenic variants show significant negative effects in one or both assays. f) Cross-analysis of CADD scores and experimental results.

analysis.

We first sought to categorize residues based on their conserved functional impact across Fyn, Lck, and c-Src. To do so, we examined residues where a loss-of-function (LOF) phenotype was conserved in both abundance-PCA and the decomposed kinase activity scores. We observed that 15 of the 24 conserved LOF positions identified in the abundance screen also showed a conserved LOF impact on kinase activity, suggesting their dual roles in maintaining stability and catalytic function. Of these, ten map to well-known structural motifs and reinforce existing mechanistic insights. For example, V25 and K28 ($\beta 3$) are critical for ATP binding, with K28 forming a key salt bridge with E44 (HC). Other residues with known catalytic roles include E166 (ALC), which is involved in substrate positioning, and R240 (HH), which forms a salt bridge with E166. Our analysis also highlighted residues forming the hydrophobic catalytic spine (C-spine), including L80 (HD), L121, and I126 (Catalytic Loop), whose strong LOF phenotypes underscore the C-spine's architectural importance. D178 (HF) also showed a strong LOF phenotype, consistent with its role in stabilizing the strained backbone conformation of the HRD motif through hydrogen-bonding interactions with the backbone amides of H145 and R146 (Oruganty et al., 2013), a conserved feature critical for maintaining the catalytic loop geometry required for kinase activity. Beyond these motifs, our screen also identified five previously uncharacterized but essential residues, such as I/V136 (ALN) which is adjacent to the R-spine in the Mg^{2+} positioning region of the activation loop, a cluster on the HF helix (V179, W180), and F243 on the HI helix, suggesting they have unappreciated roles in kinase structure and regulation.

To identify residues whose influence is confined to catalytic regulation, we focused on positions exhibiting a conserved LOF phenotype exclusively in the decomposed kinase activity scores. These residues suggest a specialized role in modulating kinase activity or conformation rather than contributing to overall protein folding. We identified a total of 14 residues, 8 of which have known roles in catalysis. These include G18 ($\beta 2$ strand), I28 ($\beta 3$ strand), and E73 ($\beta 5$ strand) for ATP binding; A137 (xDFG motif) for Mg^{2+} coordination; and K161 and A164 (ALC) for substrate alignment in the P+1 loop. Notably, the HF helix, which our analysis identified as enriched for conserved variants, contains critical residues, such as E188 and S181. In PKA, the corresponding mutant to E188 (E230Q) crystallizes in an inactive state even with ATP present, highlighting its role in stabilizing ATP binding (J. Wu et al., 2005). Similarly, S181 on the HF helix helps position the

catalytic loop for efficient transfer of the phosphate group. Our screen also revealed six previously unannotated residues with activity-specific LOF effects. L56 and V57, located near the α C- β 4 loop, likely stabilize the R-spine assembly, a key interface for maintaining an active conformation (Taylor & Kornev, 2011). Other newly implicated residues, including V71 (adjacent to gatekeeper), E112, Y197, and L250, occupy helices in the C-lobe that shape the active-site cleft, suggesting they are important for the dynamic motions required for activation.

While LOF mutations were common, we also identified conserved gain-of-function (GOF) residues that offer insights into activity regulation. For example, at position 150 (Y419 in c-Src), the role of phosphorylation in regulating catalytic activity is more nuanced than simple on/off switching. While Y419 phosphorylation is conventionally associated with activation, non-phosphorylatable mutations retain substantial catalytic activity in certain contexts and can even reduce competition with substrate phosphorylation in single-domain in vitro assays (Spasov et al., 2018; Sun et al., 2002). However, Y419 phosphorylation is critical for higher-order functions, particularly dimerization and assembly of macromolecular complexes that modulate substrate access and selection in vivo (Spasov et al., 2018), explaining why mutations at this site have context-dependent effects on cellular signaling. At position Y170, a known autophosphorylation site, the phospho-mimetic mutation Y170E was enriched with a positive phospho-only score. This suggests that introducing a negative charge may induce the kinase to adopt an active conformation. Novel GOF sites were also identified at 64K and V195E. Interestingly, V195E appeared to stabilize the active state by forming a strong Coulombic interaction and a new hydrogen bond with K161 in the P+1 loop. This interaction functionally offsets the positive charge of K161, a mechanism previously shown to increase Src activity (Huang et al., 2018). We explore both of these novel variants in the following section.

Finally, our analysis of canonical catalytic motifs such as HRD, DFG, and APE revealed that they were enriched with mutations that had a mild effect on abundance but a substantially negative impact on the phospho-PCA score (Nolen et al., 2004). This resulted in negative phospho-only scores for mutations like p.(F139R), p.(E166N), and p.(P165H), confirming their critical and direct role in phosphorylation. In contrast, mutations in other motif residues, such as p.(R119P) and p.(D120P), had strongly negative scores in both assays, indicating a mixed effect on catalysis and abundance. In total, these data highlight the ability of our method

to separate variants that purely disrupt catalysis from those that also compromise abundance.

4.2.5 AI-enabled structural analysis reveals conformation-biasing variants

To understand the structural basis of activity-biasing mutations, we applied recently developed computational methods for conformation stability analysis. These methods utilize the inverse folding model ProteinMPNN to predict a protein's preference for specific conformations (Cavanagh et al., 2025; Dauparas et al., 2022). Here, we utilized ProteinMPNN to investigate whether a specific mutation would introduce a bias towards a particular kinase conformation. For this analysis, we used available crystal structures of c-Src and Lck kinases in active (Src:1YI6, Lck:3LCK), inactive DFG-in (Src:2SRC, Lck:3B2W), and inactive DFG-out conformations (Src:2OIQ, Lck:6PDJ). For each possible single amino acid substitution, we computed the preference for inactive DFG-in versus inactive DFG-out conformations, as well as active DFG-in versus inactive DFG-in conformations. This analysis revealed a set of eight variants that are predicted to favor specific conformational states across both kinases; all have a significant effect on activity in our decomposed data. Of these eight, variants 29A, 44D/T, and 139A/R were predicted to favor the inactive DFG-out conformation over the active DFG-in conformation, consistent with their negative kinase activity scores in our experimental data. Conversely, variant 116D was predicted to favor the inactive DFG-in state over the active DFG-in state, also correlating with reduced activity scores. Notably, variants 168Q and 64K were both predicted to strongly favor the active DFG-in conformation over the inactive DFG-out state. However, only 64K was additionally predicted to favor the active DFG-in conformation over the inactive DFG-in state. This prediction aligns with our experimental observation that 64K represents a novel gain-of-function mutation across all three SFKs, while 168Q shows mild loss-of-function effects. We note that although all eight variants identified by ProteinMPNN were present in our data, it failed to identify most of the allosteric sites that impact kinase activity discovered in our study. This indicates ProteinMPNN has a low recall and high precision for variant effect prediction in this setting. While the precision of these methods enables them to provide useful insights into how variants behave, their limited performance suggests that significantly more data are necessary before they can be used more broadly to investigate the impact of allosteric variants.

We also investigated the molecular basis of the V195E gain-of-function mutation using evolutionary coupling analysis and equilibrium conformation sampling (Fig-

ure 4.4c,d). Evolutionary coupling analysis applied to Fyn, Lck, and c-Src revealed a co-evolutionary network involving residues 81, 85, 161, 188, 195, 201, and 202 (Figure 4.4c, right). To understand how the V195E mutation affects this network, we used BioEmu (Lewis et al., 2025), a recent method for accelerated generation of conformational ensembles for medium-sized proteins, to sample 4,000 conformations for each wild-type kinase and their corresponding 195E mutants. Structural analysis of the generated ensembles revealed that the 195E mutation establishes new Coulombic interactions with residues 85, 122, 161, and 188. Most significantly, a salt bridge between residues 161 and 195 was prevalent only in mutant conformations, bringing these residues into close proximity (Figure 4.4c, left). To explore the impact on kinase activity, we used Kincore to classify conformations as active or inactive. We then evaluated the percentage of structures that were classified as active, depending on whether the salt bridge was present or absent. In all three kinases we observe an increase in the active conformation fraction with the 161-195 salt-bridge present (Figure 4.9). These results demonstrate that the V195E mutation enhances kinase activity by stabilizing active conformations through the formation of a critical salt bridge within a network of co-evolving residues.

4.2.6 Abundance and phospho-PCA scores reveal the impact of clinical variants

A wide range of human diseases can arise from single-point mutations in SFKs, and for many human variants, the potential pathogenicity and function are unknown. We therefore reasoned that integrating abundance and phospho-PCA scores could provide a powerful framework to connect human kinase variants to their molecular and functional consequences. To do so, we first analyzed the behavior of known benign and pathogenic variants across the three SFKs from the ClinVar database within our experimental data (Figure 4.4e) (Landrum et al., 2018). Benign variants displayed near-zero to mildly positive abundance and phospho-PCA scores, consistent with preserved protein function. In contrast, pathogenic variants frequently displayed significantly negative values in one or both metrics, underscoring the essential role of SFKs in human biology. Several variants exhibited negative abundance scores, accompanied by only slight changes in phospho-PCA scores, suggesting that reduced protein stability alone can drive pathogenicity. We speculate that the non-catalytic function of SFKs may explain this observation. Furthermore, several pathogenic variants exhibited significantly negative scores for both abundance- and phospho-PCA. Interestingly, these trends can be used to infer the pathogenicity for variants

of unknown significance (VUS). For instance, the VUS variant N208K displayed a pronounced reduction in abundance-PCA while the VUS variants I111T, W180R, and G18A display significant reductions in both abundance and phospho-PCA, suggesting that all four are pathogenic variants.

To expand our analysis beyond ClinVar's SFK variant collection, we collected population allele frequency data from the Genome Aggregation database (gnomAD) (S. Chen et al., 2024) and annotated them with pre-computed Combined Annotation Dependent Depletion (CADD) scores (Schubach et al., 2024), a computational prediction of how harmful a genetic variant is likely to be. We then used these CADD scores to continue our analysis of the functional impact of SFK variants (Figure 4.4e). When overlaid on our experimental data, where we plot the abundance and phospho-PCA scores for each variant, we observed that variants with low CADD scores (presumably benign) clustered near the origin. However, high CADD variants were concentrated in the bottom-left quadrant, with negative scores for both assays. This finding was consistent with our analysis of the ClinVar database, which revealed a correlation between variant pathogenicity and a negative impact on both scores. Interestingly, the top-left quadrant contained a smaller number of high-CADD variants with reduced abundance-PCA scores but high phospho-PCA scores. This suggests that these rare variants may be deleterious gain-of-function mutations. This analysis demonstrates that the combined abundance and phospho-PCA scores can provide crucial insights into the function of clinically significant variants.

4.2.7 Molecular dynamics reveals allosteric networks perturbed by specificity-determining variants

Allosteric effects, where perturbations at one site propagate through a protein to affect distant functional regions, are ubiquitous in biology; understanding how allosteric communication unfolds and is perturbed by variants is critical for understanding protein evolution and disease. Our computational decomposition of phospho-PCA generated scores for substrate specificity, providing an opportunity to explore the role of allostery on substrate preference. We identified a collection of variants that specifically impact substrate preference for these three SFKs; these residues are shown in Figure 4.5a and described in full detail in the supplement (Table 4.3). From the statistically significant variants we identified, we performed additional filtering by categorizing amino acids based on their chemical properties. We then identified variants in which the amino acid property matched one kinase's

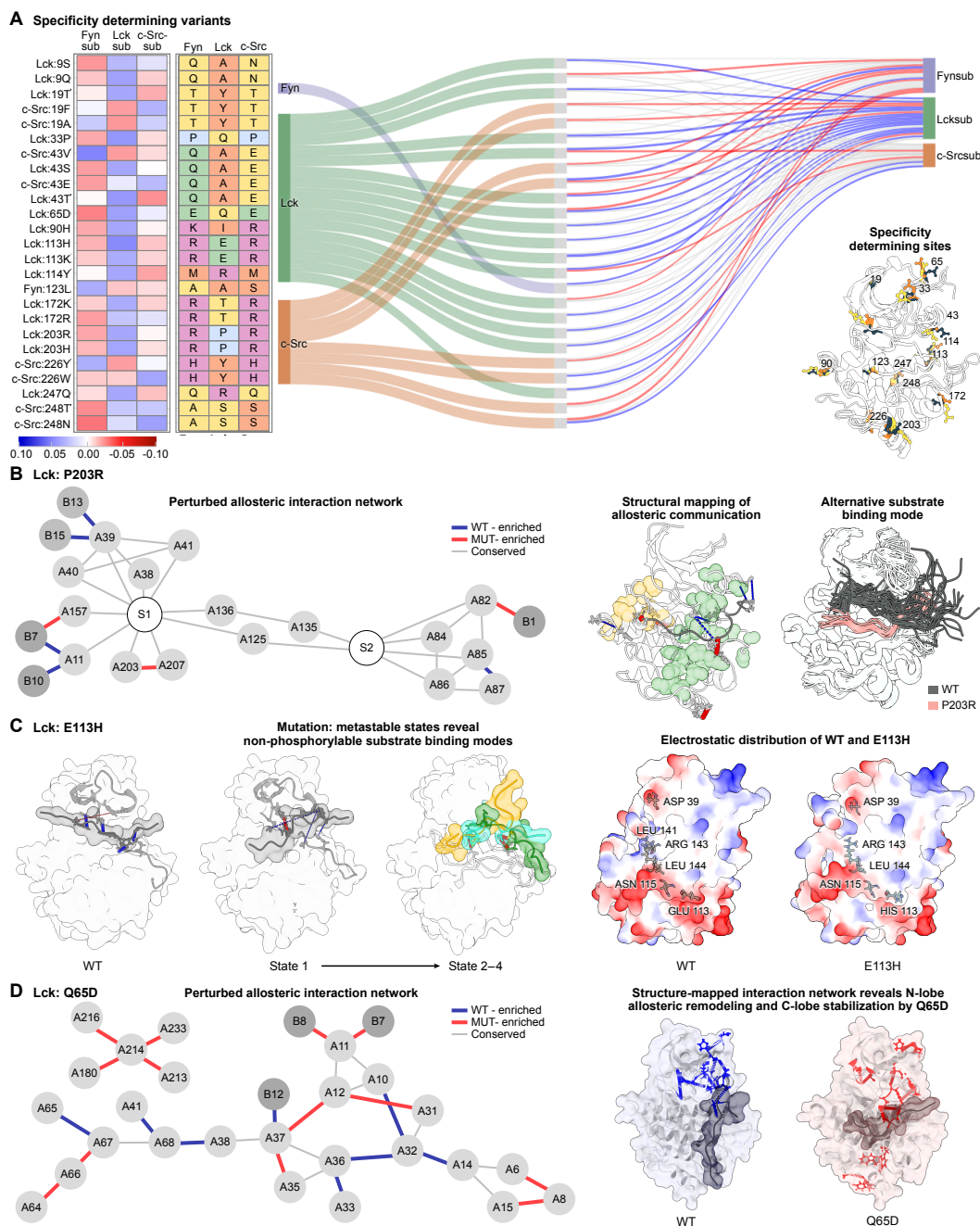


Figure 4.5: Molecular dynamics reveals allosteric networks governing substrate specificity. a) Specificity-determining variants identified through computational decomposition. Left: Heatmaps showing residue chemical properties and substrate specificity scores. Middle: Sankey diagram connecting specificity-determining mutations to affected allosteric sites. Right: Structural mapping showing most specificity-determining residues are distal to the kinase-substrate interface. b) Lck P203R rewires an allosteric network, which leads to altered substrate positioning. c) Lck E113H enables an electrostatic switch reshaping binding modes. d) Lck Q65D reconfigures N-lobe packing to favor Fyn-like substrate.

wild-type property at that site, but differed from the other two; this yielded a list of variants that induced a clear, directional change in the preferred substrate. Surprisingly, when displayed on top of the protein structure, most of the identified residues were distal to the expected binding interface of the substrate (Figure 4.5a, right). Based on this observation, we concluded that allostery is likely a key mechanism underlying how these variants affect substrate specificity.

To understand how distant mutations alter substrate specificity through allosteric mechanisms, we performed molecular dynamics (MD) simulations of several specificity-determining variants bound to their substrates. We used the Chai-1 protein folding model to generate kinase-substrate complexes and simulated both wild-type and variant kinases for 500 ns using GROMACS (Abraham et al., 2015; Chai Discovery et al., 2024). Our analysis focused on Lck and c-Src, as we were unable to obtain confident structural models for Fyn in complex with its substrate. Our central goal was to trace how each distant specificity-determining mutation communicates with the substrate-binding interface—that is, to map the allosteric pathways connecting the mutation site to the residues that directly contact the substrate. To accomplish this, we developed a contact graph framework that identifies plausible routes of structural communication through the protein. First, we constructed residue contact graphs for each simulation, where nodes represent individual residues and edges connect residue pairs that form frequent contacts during the trajectory (Hagberg et al., 2008). We then identified contacts that were strengthened (red edges) or weakened (blue edges) in the variant relative to wild-type. Next, we defined a set of “critical residues” that must be connected by any viable allosteric pathway: (1) the mutated residue itself, (2) substrate residues whose kinase contacts differ between variant and wild-type, (3) kinase residues that directly contact the substrate, and (4) residues in structural regions showing significant changes in correlated motion between variant and wild-type simulations. Finally, we identified minimal pathways connecting all critical residues by solving an approximation to the Steiner tree problem (Kou et al., 1981), which finds the smallest subgraph linking a specified set of nodes. A detailed description of the contact graph construction, critical residue identification, and pathway extraction is provided in the Methods.

The results of this analysis are shown in Figure 4.5 and revealed distinct mechanisms by which different specificity-determining variants use allosteric pathways to modulate substrate preference. Some variants induced new substrate-binding modes

by reorganizing interface contacts, while others introduced new inactive metastable conformations that reduced overall catalytic competence. We describe the specific mechanism for each analyzed variant below.

- *Lck P203R introduces a new, inactive binding mode for the Lck substrate.* Position 203 in Lck represents a critical allosteric site controlling substrate specificity through long-range communication networks within the kinase domain (Figure 4.5b). The P203R mutation reduces Lck's affinity for its native substrate while enhancing activity toward Fyn/c-Src substrates, consistent with prior work showing that the corresponding c-Src residue (R472) coordinates substrates containing negative charges downstream of the phosphorylation site (N. H. Shah et al., 2018). Our molecular dynamics simulations revealed that P203R dramatically repositions the substrate closer to the C-lobe and farther from ATP, accompanied by specific interaction changes at positions +1, +7, +10, +13, and +15—the same positions differing in charge distribution between Lck and c-Src substrates. The mutation propagates its effects through a sophisticated allosteric network: the arginine substitution establishes new intramolecular interactions (sidechain hydrogen bonding between residues 203–207 and increased 205–197 contacts), which alter interaction frequencies along the 203–205–197–196–161–159 pathway, transmitting from HG through the flexible HF-HG loop allosteric hub to the P+1 loop and ultimately to the N-terminal lobe through regulatory spine interactions.
- *Lck E113H creates new, inactive metastable states.* Position 113, located five residues upstream of the HRD motif in the catalytic loop, exhibits divergent electrostatic properties among SFKs (glutamic acid in Lck versus arginine in Fyn/c-Src) and has been previously implicated in controlling substrate specificity through charge-dependent peptide recognition (N. H. Shah et al., 2018). Our molecular dynamics simulations revealed that the E113H charge reversal fundamentally alters Lck's tolerance for positively charged substrates by stabilizing non-productive binding conformations (Figure 4.5c). Across two independent simulations, E113H consistently adopted four distinct phosphorylation-incompetent metastable macrostates with altered substrate positioning compared to wild-type, disrupting critical N-terminal activation loop contacts (Arg143-substrate +12 and Leu141-substrate +11) that normally position the substrate optimally for phosphoryl transfer. Residue interaction graph analysis revealed that the perturbation propagates through the 113–115–

144–143/141 pathway, affecting both the activation loop’s positioning and HC organization. These conformational changes cause looser substrate docking and formation of alternative salt bridges that function as kinetic traps, sequestering the substrate in catalytically incompetent conformations and explaining the experimentally observed reduced phosphorylation efficiency.

- *Lck Q65D dynamically sculpts the substrate-binding interface through allostery.* The Lck Q65D variant exemplifies a specificity switch, causing loss of preference for the Lck substrate while enhancing activity toward the Fyn substrate through both altered surface electrostatics and allosteric network communication (Figure 4.5d). The Q65D substitution at the $\beta 4$ - $\beta 5$ turn creates a new negative charge on the N-lobe surface near the substrate entrance that electrostatically attracts the multiple positive charges upstream of the phosphorylatable tyrosine in the Fyn substrate, enhancing encounter rates and binding affinity. Additionally, Q65D modulates substrate selectivity by altering the kinase’s preference at the -1 position from bulky hydrophobic residues (Lck substrate’s leucine) to polar residues (Fyn substrate’s threonine) through allosteric rewiring of HC/ β -sheet coupling networks. This coordinated remodeling of the N-lobe and C-lobe suggests that the divergent HG-HH region plays a key role in determining kinase-specific substrate preferences.

Collectively, these three variants exemplify how diverse allosteric mechanisms—spanning electrostatic remodeling, metastable-state stabilization, and network rewiring—enable them to exert their influence from distal sites to reshape the substrate-binding interface.

4.3 Discussion

This study represents the most comprehensive experimental characterization of how mutations impact folding stability, catalytic activity, and substrate specificity for any enzyme family to date. We systematically measured the functional impact of over 15,000 kinase domain variants by profiling more than 45,000 kinase-substrate interactions across nine kinase-substrate combinations (three kinases paired with three substrates each), generating phospho-PCA and abundance-PCA measurements for each. This unprecedented scale is matched by the depth of our mechanistic insights: unlike previous deep mutational scanning studies of kinases that measured only overall activity against broad substrate pools, our approach uniquely disentangles the contributions of folding stability, catalytic activity, and substrate preference for

each variant through hierarchical Bayesian modeling. By exploring the sequence-function relationship for three SFKs and their variants, we gained mechanistic understanding of how variants alter kinase conformational ensembles and how allostery shapes the kinase-substrate interface. Critically, we found that exploring conformation landscapes through computational simulations was essential for rationalizing our observations. These simulations, which included conformational ensemble sampling and molecular dynamics, revealed sophisticated allosteric communication networks that connect distal variants to changes in activity and substrate specificity.

Beyond its scale, our work has important implications for protein biology, variant-to-function studies, and computational modeling of allostery. The insights gained from this study on protein kinases are also important for other enzyme families, as the principles governing how structural context determines mutational impact on folding stability, and how allosteric networks propagate functional changes, are likely universal. Our comprehensive dataset serves as an invaluable resource for the community, enabling benchmarking of computational prediction methods and providing ground-truth data for understanding the molecular principles governing enzyme specificity. Indeed, evaluation of existing AI methods, including protein language models and inverse folding approaches, for identifying allosteric sites revealed limited predictive power (Figures 4.4d and 4.8), as the methods we explored had high precision but low recall. This highlights a crucial limitation: the evolutionary information that self-supervised models learn to extract from protein sequences can differ substantially from the functional properties of interest. Our work addresses this critical need by demonstrating a scalable approach to generating high-quality functional data on protein variants that goes beyond measures of activity. Combining such functional and biophysical information with protein language models may offer a promising path forward for improving computational predictions (Gelman et al., 2025). Moreover, our biosensor design, which uses a physical interaction between a modified substrate and a sensor domain to link enzyme activity to cell growth, may be generalized to other enzymes responsible for post-translational modifications (Gonzalez et al., 2025), providing a scalable framework for mapping functional landscapes across entire enzyme families.

In addition to advancing fundamental understanding of enzyme function, our work provides actionable insights for clinical variant interpretation and drug discovery. Our comprehensive functional map enables interpretation of clinical variants, as

demonstrated by our analysis of ClinVar and gnomAD data, which revealed clear signatures that distinguish benign from pathogenic variants that can be used to infer pathogenicity for variants of unknown significance. Beyond variant interpretation, our insights into allosteric networks controlling substrate recognition open new avenues for therapeutic development. Traditional kinase inhibitors target the ATP-binding pocket, often leading to off-target effects due to the structural similarity of kinase domains across the human kinome. Our identification of distal allosteric sites that modulate substrate specificity without broadly affecting catalytic activity suggests a fundamentally different strategy for drug design: developing highly selective kinase modulators that alter substrate preference rather than simply blocking activity. Such allosteric modulators could enable more precise therapeutic interventions by selectively disrupting pathogenic signaling while preserving beneficial kinase functions, potentially reducing the side effects that have limited the clinical utility of many kinase inhibitors.

We close by noting that our work demonstrates that phospho-PCA can be scaled to the entire human kinome. The growth-based biosensor works for both major kinase families, and the method requires only knowledge of kinase-substrate interactions to perform deep mutational scanning. We estimate that each barcoded kinase library costs $\approx \$3,000$ in long-read sequencing costs, and each mutational scan costs an additional $\approx \$1,000$ in short-read sequencing costs. Extending measurements of a similar scale to the entire human kinome would cost $\approx 3.5 \times 10^6$ USD. Although the cost is substantial, it is reasonable considering the value of the data. Such a comprehensive functional map would significantly advance our understanding of how sequence variation shapes cellular signaling networks; the work presented here now places such an advance within reach.

There are several limitations of phospho-PCA and our study that merit discussion. First, we leverage yeast as the model organism for phospho-PCA. The cellular environment (e.g., pH, ionic strength, and cofactor concentrations) in yeast differs from that in mammalian cells, which could impact the translation of our results. Further, yeast express a number of endogenous serine-threonine kinases; this background signal may pose a challenge for future phospho-PCA measurements of serine/threonine kinases. We note that our Bayesian approach to separating the impact of abundance from phospho-PCA scores also has limitations. This model assumes that phospho-PCA scores can be decomposed linearly into abundance, activity, and specificity components, which may oversimplify the relationships between these factors. An-

other limitation is that the difference in scale between activity and specificity scores suggests that phospho-PCA is more sensitive to changes in catalytic rates than to changes in substrate binding affinities. This is likely due to the tethered design of the phospho-PCA biosensor, which creates a high local substrate concentration. Our analysis of clinical variants relies on a connection between the functional scores measured in yeast and human disease phenotypes. Future work should seek to further validate this connection, as this link likely depends on cellular and tissue contexts not captured by our system. While phospho-PCA can identify variants with allosteric effects, it cannot directly map the long-range cooperative networks that mediate allostery. Our study focused on the kinase domain itself and did not examine the effects of mutations in non-catalytic domains, which play a key role in kinase regulation. Indeed, prior work has shown that kinase-domain variants can alter interactions between the kinase domain and other domains (Ahler et al., 2019). Lastly, our emphasis on single amino acid variants means that our study cannot address epistatic interactions between residues. We also leave this to future work.

4.4 Methods

Method details

Yeast strain. All experiments were conducted using *Saccharomyces cerevisiae* strain CEN.PK2-1D (MATa/alpha; ura3-52/ura3-52; trp1-289/trp1-289; leu2-3_112/leu2-3_112; his3 delta1/his3 delta1; MAL2-8C/MAL2-8C; SUC2/SUC2).

Single yeast plasmid construction. Plasmid construction began with two protein complementation assay (PCA) backbones: pGJJ001 (binding-PCA containing DHFR fragments under dual CYC promoters) and pGJJ045 (abundance-PCA containing DHFR fragments under CYC and GPD promoters, convertible into phospho-PCA backbone by phospho-amino acid binding domain insertion). An SH2 domain was first introduced into each backbone using BamHI/SpeI restriction sites, generating pCY001 (CYC/CYC configuration) and pCY002 (GPD/CYC configuration). Active and catalytically inactive Fyn kinase domains were subsequently cloned via NheI/HindIII sites, yielding pCY003–004 from the pCY001 backbone and pCY005–006 from the pCY002 backbone. Using pCY005 as the template, a comprehensive panel of kinase–substrate fusion constructs (with a serine-glycine linker connecting the two) was assembled through an identical NheI/HindIII cloning strategy. This approach generated plasmids pKMS003–008 and pKMS012–016, encompassing all combinations of Fyn, Src, and Lck catalytic domains paired with their respective substrate sequences. Control plasmids containing catalytically dead kinase domains

(pCY006, pKMS015/016) were constructed in parallel to establish baseline measurements.

Yeast transformation. Growth-based biosensors were introduced into yeast through transformation with the corresponding plasmid. Yeast cultures were inoculated in 6 mL YPD medium and grown overnight at 30°C. Cultures were diluted to $OD_{600} = 0.1$ in 180 mL YPD and grown to $OD_{600} = 0.5$ – 0.6 . For lithium acetate/polyethylene glycol transformation, approximately 45 mL of cells were harvested, washed twice in 100 mM lithium acetate, and aliquoted into 1.5 mL microcentrifuge tubes. Salmon sperm DNA was prepared by boiling for 5 minutes followed by ice incubation. Transformation mixtures were prepared by sequential addition to cell pellets: 240 μ L 50% PEG-3350, 36 μ L 1 M lithium acetate, 10 μ L salmon sperm DNA (10 mg/mL), and 74 μ L plasmid DNA (1–5 μ g in a total volume of 360 μ L). After vortexing, samples were incubated at 30°C for 30 minutes, heat-shocked at 42°C for 30 minutes, pelleted, resuspended in 150 μ L water, and plated on SC-URA selective medium for 2–4 days.

Growth-based kinase-substrate interaction assays. Transformed colonies were picked into 96-deep-well plates containing 500 μ L SC-URA/MET/ADE medium per well and grown overnight at 30°C. Cell density was measured by transferring 100 μ L culture plus 100 μ L fresh medium to flat-bottom 96-well plates. The remaining cultures were preserved as glycerol stocks (30% final glycerol concentration) and diluted to $OD_{600} = 0.1$. For kinase activity measurements, 5 μ L diluted cells were added to 95 μ L competition medium (1.053 $\times$ SC-URA/MET/ADE supplemented with 200 μ g/mL methotrexate) to achieve a starting density of $OD_{600} = 0.005$. Growth was monitored for 60 hours at 30°C using a Tecan plate reader with OD_{600} measurements every 15 minutes. Time-course data were smoothed using a centered moving average (25-point window). Exponential growth phases were identified by computing the numerical derivative of smoothed optical density with respect to time. Growth rates were calculated as the slope across selected exponential-phase windows, normalized by the z-score within each kinase and then scaled to the range [0, 2].

Construction of deep mutational scanning libraries. To create barcoded biosensor libraries, we first constructed modular screening plasmids from the abundance-PCA and phospho-PCA plasmids by introducing a BamHI site upstream of the (GGSGG)₄ linker using KpnI and NheI restriction sites. These modifications generated matched kinase-substrate PCA constructs: pDMS001 (abundance-PCA; Fyn

kinase with Fyn substrate) and pDMS002–004 (phospho-PCA; Fyn, Lck, and Src kinases with their respective substrates). Kinase open reading frames were subcloned into the mutagenesis vector pGJJ055 via BamHI/HindIII to create kinase-specific templates pDMS005 (Fyn), pDMS006 (Lck), and pDMS007 (Src). Comprehensive saturation mutagenesis was performed using the one-pot method with NNK primer sets designed for each kinase domain. Following electroporation into TOP10 electrocompetent cells and overnight growth, mutant libraries were harvested by plate scraping, and plasmid DNA was prepared by midiprep. BamHI/HindIII kinase fragments were cloned into pDMS002 to generate phospho-PCA libraries containing the Fyn substrate. Libraries were linearized at the barcode landing pad (AvrII/BsiWI) and barcoded using IDT oligonucleotide cassettes with NEBuilder Assembly. Each barcoded pool was bottlenecked to approximately 3% of transformants to maintain library diversity while ensuring adequate representation. To generate the complete 3×3 kinase-substrate panel, substrate cassettes were exchanged using AvrII/BamHI digestion and ligation with fragments from pDMS003 and pDMS004, yielding nine phospho-PCA libraries (see Figure 4.14 for a construct map). Matched abundance-PCA libraries were constructed by ligating kinase-substrate-barcode cassettes into pGJJ045 via BsiWI/HindIII.

Selection of barcoded biosensor libraries with methotrexate. BY4742 yeast cells were prepared for electroporation by growth to mid-log phase ($OD_{600} \approx 1.6$), followed by washing with ice-cold water and electroporation buffer. Cells were incubated in 0.1 M LiAc/10 mM DTT for 20 minutes at 30°C before final preparation. Libraries (0.3 μ g) were electroporated with salmon sperm DNA (50 μ g) using standard parameters (2.5 kV, 25 μ F, 200 Ω). Transformed cells were rescued in sorbitol/YPD medium and expanded in SC-URA medium for 48 hours. Competition experiments were initiated by transferring cells into medium containing 200 μ g/mL methotrexate at $OD_{600} = 0.05$ and growing for approximately 5 generations. For each library, 12–16 electroporation reactions were pooled across two biological replicates. Cell pellets were harvested by centrifugation and stored at –20°C for subsequent plasmid extraction using ZymoPrep II columns.

Library preparation for next-generation sequencing. Barcode regions were amplified using a two-step PCR protocol. In PCR1, eight parallel 50 μ L reactions per library were performed using KAPA HiFi polymerase with barcode-flanking primers for 15 cycles at 65°C annealing temperature. PCR2 appended Illumina adapter sequences. Final libraries (~225 bp) were sequenced on the Illumina platform with

paired-end 50 bp reads targeting 50 million reads per library. For barcode-variant mapping, the contiguous kinase-substrate-barcode segment (~1.7 kb) was amplified using KAPA HiFi polymerase and sequenced on the PacBio Revio platform, generating approximately 45 million HiFi reads across four SMRT cells.

Quantification and statistical analysis

Next-generation sequencing analysis. We developed bioinformatics pipelines to construct barcode-variant look-up tables from long-read PacBio sequencing and measure barcode abundances in selection experiments from short-read Illumina sequencing.

Barcode-variant look-up table construction. HiFi FASTQ reads from each kinase library were filtered for the exact vector context and the presence of both the DHFR–substrate cassette and a 3' overhang motif. Reads were parsed to extract the landing-pad barcode and the kinase ORF. We validated the kinase library identity via a two-base tag embedded in the barcode (AG, CT, and TA for Fyn, Lck, and Src, respectively). ORFs were length-checked (765 nt), translated, and compared codon-wise to the kinase-specific wild type. The next step was to estimate the consensus variant sequence for each barcode using Bayesian statistics. For each position i in the variable region, Bayes' rule gives:

$$P(b_i = b|D_i) = \frac{P(D_i|b_i = b) P(b_i = b)}{\sum_{b'} P(D_i|b_i = b') P(b_i = b')}, \quad (4.1)$$

where we take the likelihood of observing the correct base to be 0.9 and 0.03 for errors, with a uniform prior $P(b_i = b) = 0.25$. The position-wise consensus selects $\hat{b}_i = \arg \max_b P(b_i = b|D_i)$, and an overall confidence score is computed as $P(\text{consensus correct}) = \prod_i P(b_i = \hat{b}_i|D_i)$. Barcode-variant pairs with a joint posterior probability below 0.99 were excluded from subsequent analyses.

Bioinformatic analysis of deep mutational scanning data. Our bioinformatics pipeline used MAGeCK to compute per-barcode $\log_2$ fold changes for outlier detection, then applied Rosace's Bayesian hierarchical model to the filtered raw count matrix to generate fitness scores. Paired-end FASTQ files were quality-filtered and merged using fastp, with a minimum overlap of 6 base pairs. Barcode abundance was quantified using MAGeCK count, with 34 nucleotides trimmed from the 5' end to remove constant upstream sequences. For each screening condition, enrichment was assessed with MAGeCK test by comparing pre- and post-selection replicates. Only barcodes with minimum pre-selection coverage (>20 reads) were

retained. Analysis was restricted to barcodes corresponding to two or fewer amino acid changes, and barcodes with $|z| > 2$ were excluded as outliers. The resulting 18 screening libraries achieved a barcode-variant coverage of 10–25 $\times$ and included between 95.9% and 99.3% of all possible single amino acid mutations within the 254-residue kinase domain (Table 4.1).

Abundance-PCA score normalization and enrichment analysis. Raw abundance scores were scaled using a relative nonsense baseline, and synonymous variant scores were used to zero-center the distribution. For each kinase, variant-level abundance scores were averaged across the three substrate backgrounds. Variants were classified as GoF if their normalized abundance score exceeded the mean of all positive scores for that kinase, and as LoF if the score fell below 0.45-fold the mean of all negative scores. For each residue position, a one-sided Fisher’s exact test was used with Benjamini-Hochberg correction; positions with $\text{FDR} \leq 0.10$ were designated as enriched.

Solvent-Accessible Surface Area (SASA) computation and stratification. Per-residue SASA was computed for the catalytic kinase domains using crystal structures: Fyn (PDB: 2DQ7), Lck (PDB: 3LCK), and Src (PDB: 1YI6). Per-atom SASA was computed in PyMOL using a 1.4 Å probe radius. Residues were classified into five ordinal exposure categories based on per-residue SASA (strongly exposed $>150 \text{ Å}^2$; moderately exposed $50\text{--}150 \text{ Å}^2$; buried $10\text{--}50 \text{ Å}^2$; strongly buried $2.5\text{--}10 \text{ Å}^2$; core/hydrophobic $<2.5 \text{ Å}^2$).

Decoupled activity score harmonization and enrichment analysis. Variant-specific activity effects decoupled from abundance were extracted from the hierarchical Bayesian model. Within each kinase, positive and negative scores were rescaled independently by the respective kinase-specific means, followed by a between-kinase offset correction. Normalized decoupled activity scores were converted to categorical variant calls: GoF if score > 0.24 , LoF if score < -0.21 , and neutral otherwise. Conserved regulatory positions required significant enrichment in at least two of the three kinases.

Sampling and analysis of kinase conformational ensembles with BioEmu. BioEmu was used to sample 4,000 conformations per sequence for each variant of interest. Side chains were reconstructed and relaxed using BioEmu’s hpacker-openmm workflow. Conformational states were classified using Kincore, iterating frame-by-frame through each trajectory using MDAnalysis to extract per-frame state labels.

Evolutionary coupling analysis of position 195. Long-range coevolution scores for Fyn, Lck, and c-Src were computed and merged into a single table. An undirected graph with residue indices as nodes and EV-coupling probabilities as edge weights was constructed using NetworkX. The local coevolutionary neighborhood of position 195 was extracted by thresholding edges at high confidence (weight > 0.95).

Protein language model analysis of allosteric sites. ESM-1b representations and attention patterns were computed for the kinase-domain sequences using the FAIR-ESM Torch Hub model `esm1b_t33_650M_UR50S`. Per-residue global and active-site attention vectors were derived and putative allosteric residues identified using an active-site attention threshold of 0.005.

Molecular dynamics simulations. Kinase-substrate complex structures were generated using Chai-1 in multimer mode with ligands and geometric constraints enabled. Each complex contained one kinase domain (Chain A), one substrate peptide (Chain B), two magnesium ions (Chains C and D), and ATP (Chain E). Geometric constraints of 4.0 Å were applied between A:R119 and B:Y8, A:D120 and B:Y8, ATP and B:Y8, and ATP and A:L127. ATP and Mg^{2+} coordinates were refined by superimposing PDB structures 1QMZ and 1IR3 onto the Chai-1 models using PyMOL. Molecular dynamics simulations were performed using GROMACS 2025.2 with the CHARMM36-jul2022 force field and the TIP3P water model. Simulation protocols included energy minimization, NVT equilibration, NPT equilibration, and 500 ns production runs. Trajectories were processed using `trjconv` and the MDAnalysis Python package, and stable 100 ns segments were identified based on RMSF analysis and inspected using ChimeraX.

Molecular dynamics trajectory analysis.

Structural alphabet encoding with AllohubPy. MD trajectories were converted to structural alphabet representations using AllohubPy's encoder (alphabet: M32K25, block size = 100, with block randomization). For each fragment pair, $\log_2$ fold changes (mutant vs. wild-type) and Benjamini-Hochberg adjusted p -values were computed. Significant fragments were identified using $FDR < 10^{-4}$ and variant-specific fold-change thresholds ($\log_2 FC > 4$ up-regulated; < -4 down-regulated).

Residue interaction graph construction and pathway analysis. Per-frame interaction frequencies were calculated at residue and atom levels. Differential interactions were identified using variant-specific mean difference thresholds (P203R: 0.115;

Q65D: 0.3). Residue contact networks were constructed with residues as nodes and frequent contacts as edges. For P203R, the full graph G was reduced to a minimal subgraph H that preserves all important residues, collapses conserved intermediates into supernodes, and retains only the minimum intermediates required to maintain connectivity. For E113H, interactions were compared across multiple metastates by binarizing contact frequencies (threshold: 0.3).

Identification of specificity-determining variants. For specificity scores, we required that mutations satisfy both a 94% credible interval test (posterior probability of a directional effect $\geq 97\%$) and a Kolmogorov-Smirnov test comparing the posterior distribution to the prior (Linden-Santangeli et al., 2025; Massey, 1951; Stephens, 2016). This dual-threshold approach guards against cases where the credible interval criterion might be satisfied due to a narrow posterior that has shifted only modestly from the prior.

Data and Materials Availability

All sequencing data, analysis code, and simulation files associated with this chapter are deposited under the title *Uncovering the Molecular Basis of Kinase Activity and Substrate Recognition with Phospho-PCA*. Long-read sequencing data are available at <https://doi.org/10.5281/zenodo.17541166>. Short-read sequencing data are available at <https://doi.org/10.5281/zenodo.17541172>. Analysis pipelines and molecular dynamics simulation trajectories (as compressed .mp4 files) are available at <https://doi.org/10.5281/zenodo.17539481>. Code for all analysis pipelines is available at https://github.com/vanvalenlab/sfk_dms_paper.

Supplementary Text

Decomposition of phospho-PCA data through hierarchical Bayesian modeling

Deep mutational scanning (DMS) experiments often confound several biological effects that—in an ideal world—should be carefully separated to extract meaningful insights. In our kinase studies, measured phosphorylation depends on protein folding stability, the probability the kinase is in an active state, and the specificity of the kinase-substrate interaction. Since protein misfolding reduces kinase abundance, which in turn decreases phosphorylation, folding stability acts as a confounder that must be separated from the other two effects.

Our experimental design uses a dual DMS screen to address this confounding issue: abundance-PCA measures kinase folding and stability, while phospho-PCA

measures the kinase-substrate interaction. To separate these effects, we developed a hierarchical Bayesian regression model that decomposes the observed phospho-PCA signal into three distinct components:

$$\text{phospho-PCA}_{ijkl} = a_{ij} \times \text{Abundance}_{ikl} + b_{ij} + \text{Activity}_{ikl} + \text{Substrate specificity}_{ijkl}, \quad (4.2)$$

with the constraint $\sum_j \text{Substrate specificity}_{ijkl} = 0$ to ensure identifiability. These levels are:

- **Level 1: Kinase-substrate experiment effects.** a_{ij} captures how abundance affects phospho-PCA scores for each specific kinase-substrate pair; b_{ij} is a baseline intercept.
- **Level 2: Kinase activity effects.** Activity_{ikl} captures mutation-induced perturbations to kinase catalytic activity that affect all substrates equally.
- **Level 3: Kinase-substrate specificity effects.** $\text{Substrate specificity}_{ijkl}$ captures effects where mutations differentially impact particular kinase-substrate pairs.

Weakly informative priors are employed:

- $a_{ij} \mid \sigma_a \sim \mathcal{N}(0, \sigma_a^2), \quad \sigma_a \sim \text{HalfNormal}(1)$
- $b_{ij} \mid \sigma_b \sim \mathcal{N}(0, \sigma_b^2), \quad \sigma_b \sim \text{HalfNormal}(1)$
- $f_{ikl}^{(\text{activity})} \mid \sigma_{\text{activity}} \sim \mathcal{N}(0, \sigma_{\text{activity}}^2), \quad \sigma_{\text{activity}} \sim \text{LogNormal}(\log 0.5, 0.5^2)$
- $\delta_{ijkl}^{(\text{specificity})} \mid \sigma_{\text{specificity}} \sim \mathcal{N}(0, \sigma_{\text{specificity}}^2), \quad \sigma_{\text{specificity}} \sim \text{LogNormal}(\log 0.25, 0.5^2)$

with the sum-to-zero constraint $\sum_{(i,j',k,l) \in C_{i,k,l}} \delta_{ij'kl}^{(\text{specificity})} = 0$. The observation model is:

$$\begin{aligned} \mu_{ijkl} &= a_{ij} f_{ikl}^{\text{abundance}} + b_{ij} + f_{ikl}^{(\text{activity})} + \delta_{ijkl}^{(\text{specificity})}, \\ y_{ijkl} \mid \mu_{ijkl}, \sigma &\sim \mathcal{N}(\mu_{ijkl}, \sigma^2), \quad \sigma \sim \text{LogNormal}(\log 0.5, 0.5^2). \end{aligned}$$

Non-centered parameterization ($X = \sigma Z$, $Z \sim \mathcal{N}(0, 1)$) was used for all group effects to improve posterior geometry and reduce divergences. The model was implemented in PyMC and fitted using the No-U-Turn Sampler (NUTS); convergence

was verified using $\hat{r}$ and effective sample size diagnostics. Sampling statistics and exemplary prior and posterior sampling distribution are detailed in Figure 4.10.

A comparison against MoCHI suggests that our decoupling preserved meaning AbundancePCA score signs, while more thoroughly decouple the residual mixed effect from PhosphoPCA (Figure 4.7).

Phospho-PCA reveals the limitations of protein language models

Recent work suggests that protein language models (pLMs) can identify functionally important allosteric sites by identifying distal sites that exhibit high attention scores with known active sites (Dong et al., 2024; Kannan et al., 2024). Following these studies, we used ESM-1b to compute attention scores between each residue and a set of active site residues for Fyn, Lck, and c-Src and compared sites with high scores with our kinase activity scores. The results are shown in Figure 4.8. We found that ESM-1b identified 10 candidate high-attention sites; eight corresponded to variants with significant differences in kinase activity scores in our data, overlapping with loss-of-function variants at positions 53, 71, 73, 112, and 180 as well as the conserved gain-of-function variant at position 236. ESM-1b also identified two sites—125 and 162—adjacent to positions 126 and 161 that displayed significant activity effects. ESM-1b’s analysis was incomplete, as 30 positions containing variants with significant activity scores but minimal impact on abundance were not flagged. Compared to our data, ESM-1b attention had a recall of ≈ 0.21 and a precision of 0.8, indicating that pLMs miss a substantial fraction of functionally important allosteric sites.

Specificity-determining variants

In this section we describe the collection of variants that impact substrate specificity.

- *Fyn*: The S123L mutation decreased activity toward the Fyn substrate by replacing a small polar serine with a hydrophobic leucine. Additionally, position 248 is conserved between Fyn and Lck but differs for c-Src, further supporting its role in substrate specificity determination.
- *Lck*: Lck exhibited the largest collection of specificity-determining mutants, which may reflect its distinct evolutionary position among SFKs (N. Li et al., 2022). The A9Q and A9S mutations decreased specificity toward the Lck substrate while increasing specificity toward the Fyn substrate. The F19T

mutation consistently decreased specificity toward the Lck substrate. The Q33P mutation reduced phosphorylation of the Lck substrate by introducing structural constraints that likely impair the conformational flexibility required for substrate binding. The A43S and A43T mutations reduced specificity toward the Lck substrate, adjacent to the critical E44 substrate-binding residue. The Q65D mutation increases specificity toward the Fyn substrate while reducing activity toward the Lck substrate. The I90H mutation reduced activity toward the Lck substrate. The E113K and E113H variants showed reduced specificity toward the Lck substrate; the opposite charges at position 113 between Lck and c-Src/Fyn suggest this residue separates the SrcA and SrcB subfamilies. The R114Y mutation displayed decreased specificity towards the Lck substrate. The T172R and T172K mutations enhanced activity toward the Fyn substrate while reducing Lck-specific activity. The P203H and P203R mutations reduced activity toward the Lck substrate (N. H. Shah et al., 2018). The R247Q mutation reduced activity toward the Lck substrate.

- *c-Src*: The T19F mutation drove preferential phosphorylation of the Lck substrate, complementing the Lck F19T findings. The Q43E mutation increased activity toward the Fyn substrate, while Q43V decreased Fyn activity but enhanced Lck activity. The H226Y and H226W mutations increased activity toward the Lck substrate while decreasing c-Src substrate activity. The A248T and A248N mutations increased specificity toward the Fyn substrate while reducing specificity toward the c-Src substrate.

4.5 Supplementary Figures

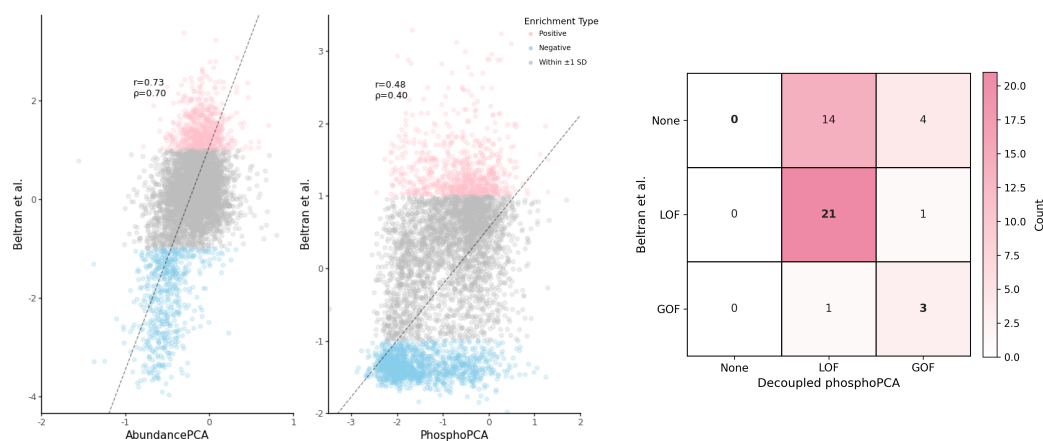


Figure 4.6: Validation of phospho-PCA measurements against published deep mutational scanning data. Left and middle panels: Scatter plots comparing abundance-PCA scores (left) and phospho-PCA scores (middle) from this study with corresponding measurements from Beltran et al. for c-Src kinase. Right panel: Contingency table showing overlap between LOF and GOF variants, demonstrating strong concordance between the two studies.

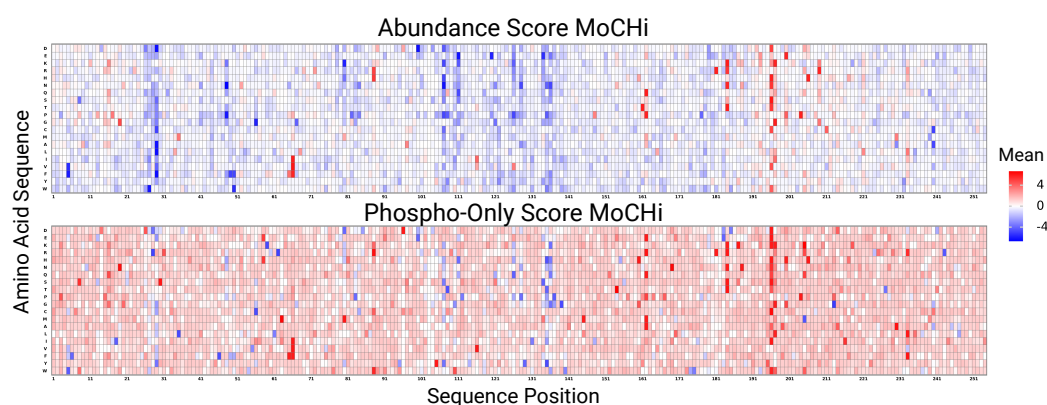


Figure 4.7: MoCHI model reveals incomplete decomposition of Fyn mutational data. Top panel: Abundance scores from MoCHI decomposition. Bottom panel: Phospho-only (activity) scores from MoCHI showing residual mixed effects that motivated the development of our hierarchical Bayesian model.

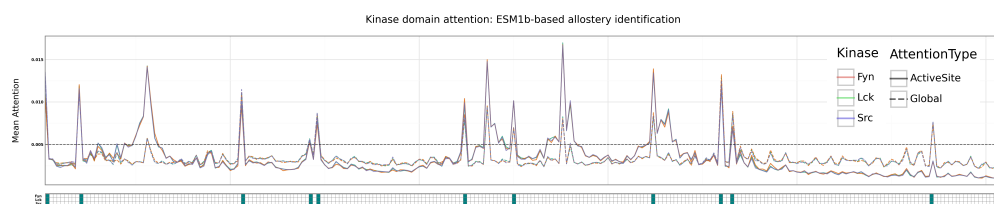


Figure 4.8: **Protein language model attention analysis identifies a subset of functionally important sites.** Top: Line plot showing ESM-1b attention scores across kinase domain positions for Fyn, Lck, and c-Src. The dashed horizontal line indicates the threshold (0.005) for identifying putative allosteric sites. Bottom: Heatmaps showing the distribution of decomposed activity scores and amino acid properties at each position.

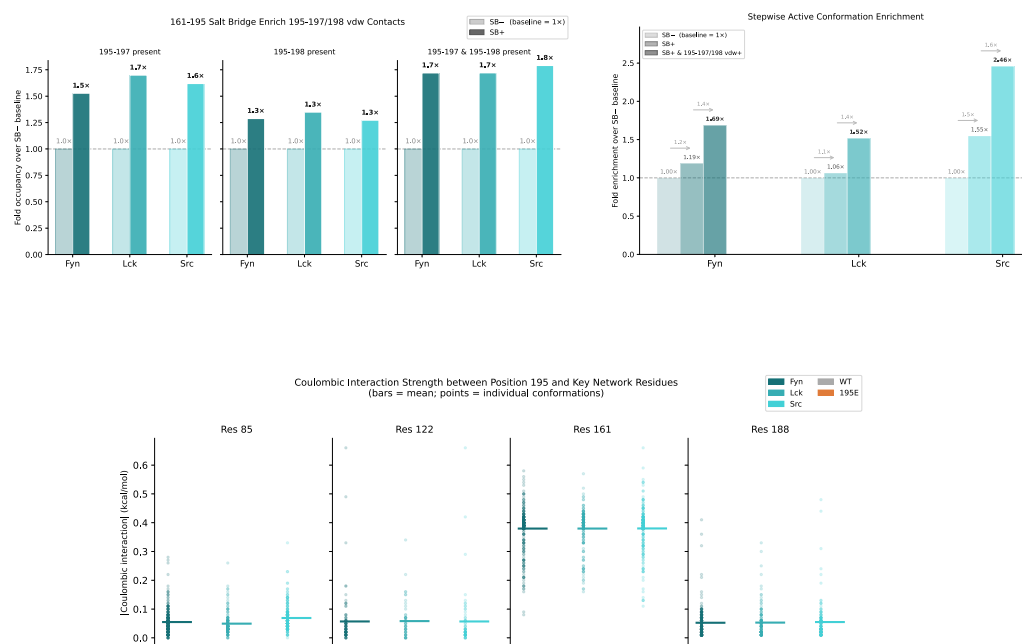


Figure 4.9: **Biophysical characterization of the V195E gain-of-function mutation.** Top: Bar plots showing the fraction of active versus inactive conformations for V195E variants, stratified by presence or absence of the 161–195 salt bridge. Bottom: Scatter plot comparing Coulombic interaction strengths between residue 195 and network partners across wild-type and V195E mutant ensembles.

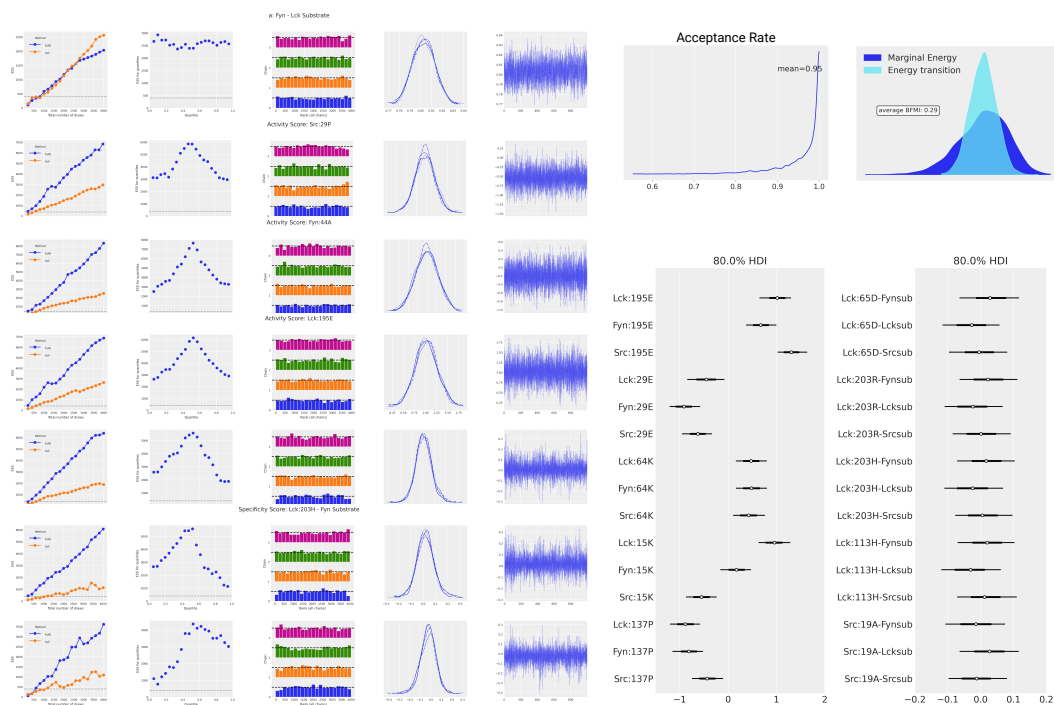


Figure 4.10: **Hierarchical Bayesian model diagnostics demonstrate robust parameter inference.** Diagnostic plots for MCMC sampling including trace plots, rank plots, and energy plots for representative parameters at each hierarchical level. All diagnostics indicate successful convergence with $\hat{r}$ values near 1.0 and adequate effective sample sizes.

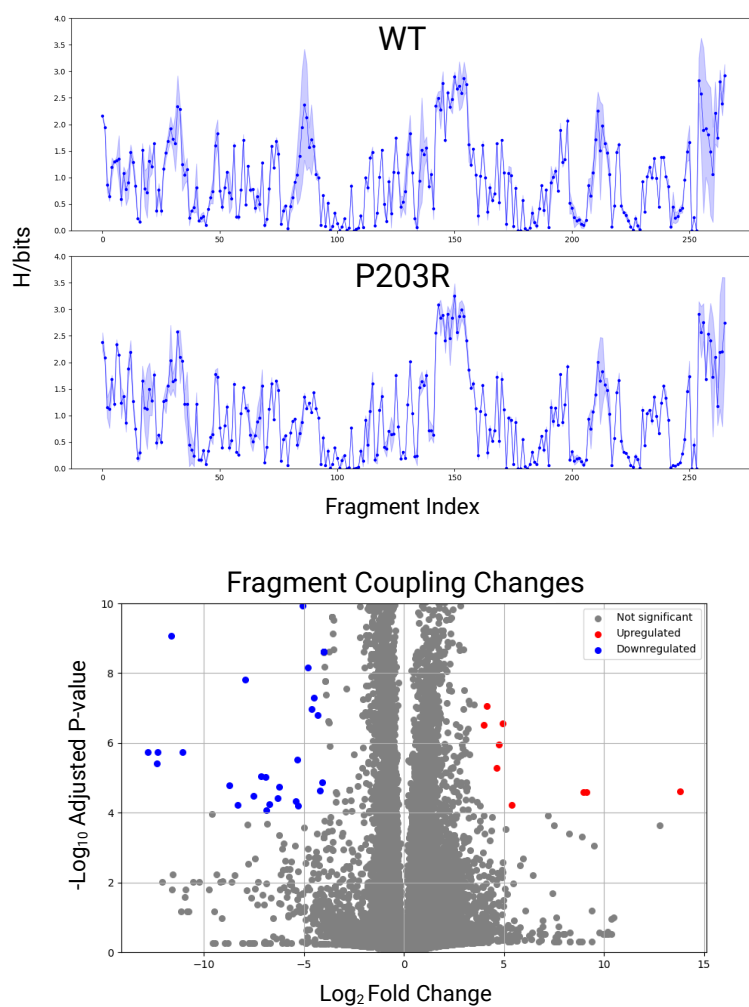


Figure 4.11: Allosteric hub analysis identifies differentially correlated protein segments in P203R mutant. Top: Positional entropy profiles comparing wild-type Lck and P203R mutant. Bottom: Volcano plot displaying $\log_2$ fold changes versus adjusted p -values for all structural fragment pairs.

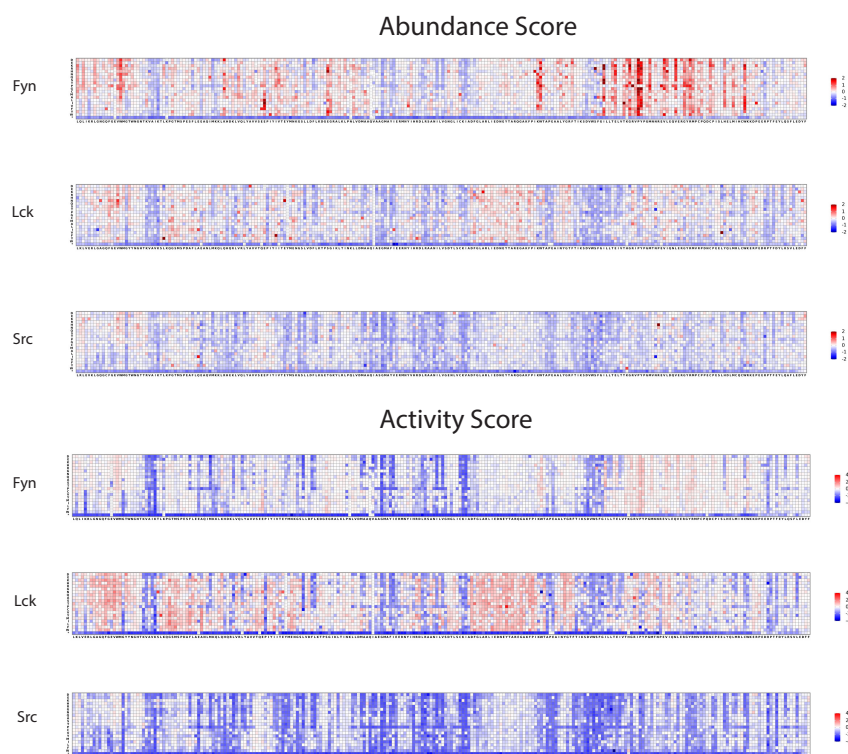


Figure 4.12: **Pooled abundance and decomposed kinase activity scores across Src-family kinases.** Top three panels: Mean abundance-PCA score shared across substrates for Fyn, Lck, and c-Src. Bottom three panels: Decomposed kinase activity scores capturing mutation-induced changes in catalytic function independent of abundance effects.

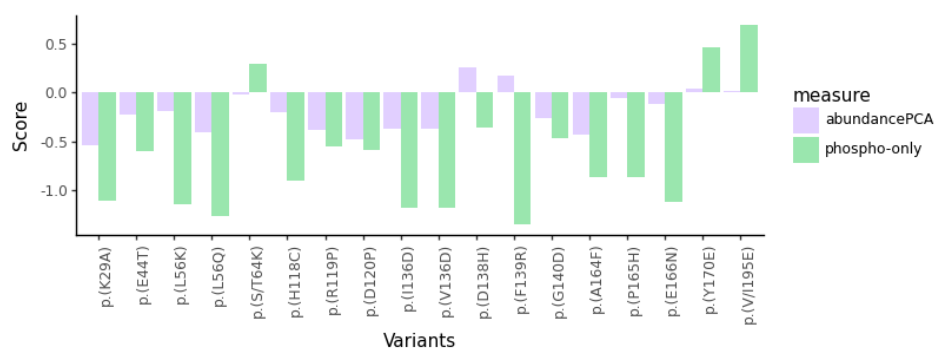


Figure 4.13: **Known kinase domain catalytic sites shows negative abundance and activity scores**

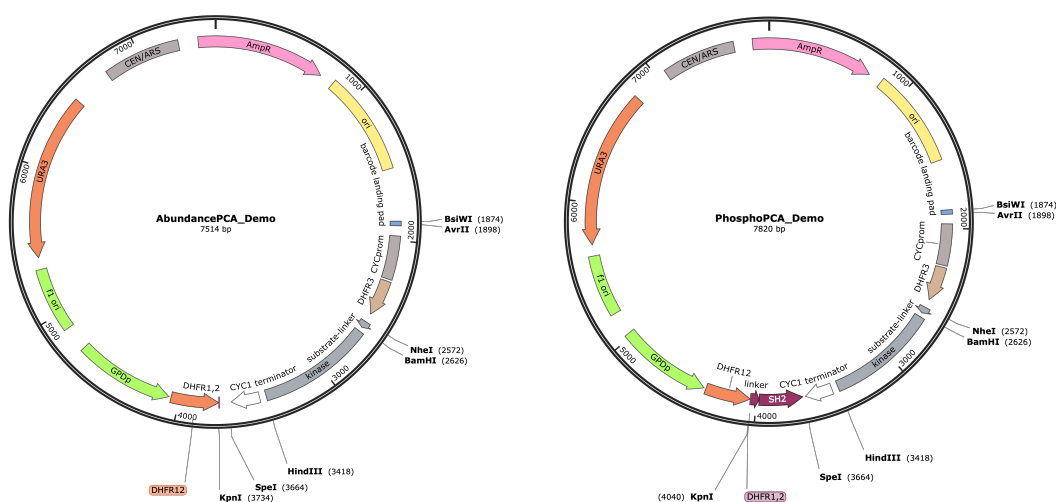


Figure 4.14: Illustration of AbundancePCA and PhosphoPCA screening Plasmids

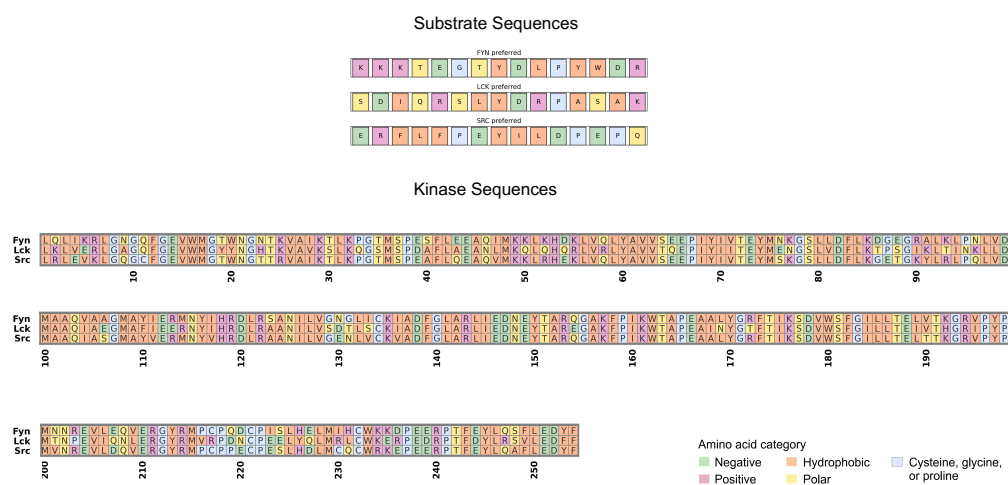


Figure 4.15: **Substrates and Wild Type kinases in the screenings**

lib_type	kinase	substrate	# singleAA barcodes	# singleAA variants	avg barcode coverage	avg variant coverage
PhosphoPCA	Fyn	fynsub	53446	4902	12.126587	0.964961
PhosphoPCA	Lck	fynsub	49204	4898	10.045733	0.964473
PhosphoPCA	Src	fynsub	53495	4990	10.720441	0.982283
AbundancePCA	Fyn	fynsub	106737	4987	21.403046	0.981693
AbundancePCA	Lck	fynsub	82271	4988	16.493785	0.981890
AbundancePCA	Src	fynsub	99511	5046	19.228062	0.993307
PhosphoPCA	Fyn	lcksub	66269	4909	13.489491	0.966339
PhosphoPCA	Lck	lcksub	53772	4908	10.956090	0.966142
PhosphoPCA	Src	lcksub	62900	4992	12.600160	0.982677
AbundancePCA	Fyn	lcksub	61490	4910	12.523422	0.966539
AbundancePCA	Lck	lcksub	40235	4872	8.499043	0.959065
AbundancePCA	Src	lcksub	48966	4945	9.902123	0.973425
PhosphoPCA	Fyn	srcsub	124473	5001	24.888622	0.984449
PhosphoPCA	Lck	srcsub	83174	4987	16.679163	0.981693
PhosphoPCA	Src	srcsub	113412	5062	22.404583	0.996457
AbundancePCA	Fyn	srcsub	106458	4990	21.371706	0.980316
AbundancePCA	Lck	srcsub	67833	4962	13.670496	0.976772
AbundancePCA	Src	srcsub	86335	5039	17.133360	0.991929

Table 4.1: Sequencing coverage for abundance- and phospho-PCA libraries

Position	ss	Fyn interaction	Lck interaction	c-Src interaction	Fyn effect	Lck effect	c-Src effect
28	B3			26:CG2-28:CG1	LOF	LOF	No Effect
28	B3	28:CG2-69:CG	28:CG1-69:CG		LoF	LoF	No Effect
101	HE		101:CB-186:CD2	101:CB-186:CD2	No Effect	LoF	LoF
101	HE		101:O-182:CZ	101:O-182:CZ	No Effect	LoF	LoF
141	ALN	141:O-143:N	141:CA-143:NH1		No Effect	No Effect	LoF
147	ALN		147:C-149:N	147:C-149:N	LoF	No Effect	No Effect
159	ALC		122:NH1-159:CG	122:NH1-159:CG	LoF	No Effect	No Effect
159	ALC	120:OD1-159:CG			LoF	No Effect	No Effect

Table 4.2: Divergent abundance-PCA mutations that disrupt van der Waals interaction networks.

Position	Kinase	WT	Mutant	Fynsub	Lcksub	c-Srcsub
9	Lck	A	Q	0.015	−0.028	0.013
9	Lck	A	S	0.029	−0.021	−0.008
19	Lck	Y	T	0.004	−0.026	0.022
19	c-Src	T	A	−0.015	0.029	−0.014
19	c-Src	T	F	−0.003	0.026	−0.022
33	Lck	Q	P	0.023	−0.032	0.009
43	Lck	A	S	0.026	−0.026	0
43	Lck	A	T	0.001	−0.031	0.03
43	c-Src	Q	E	0.026	−0.006	−0.02
43	c-Src	Q	V	−0.037	0.027	0.01
65	Lck	Q	D	0.034	−0.029	−0.005
90	Lck	I	H	0.023	−0.026	0.003
113	Lck	E	H	0.021	−0.036	0.015
113	Lck	E	K	0.014	−0.025	0.011
114	Lck	R	Y	0.001	−0.026	0.025
123	Fyn	S	L	−0.025	0.015	0.009
172	Lck	T	K	0.013	−0.025	0.012
172	Lck	T	R	0.026	−0.017	−0.008
203	Lck	P	H	0.019	−0.027	0.008
203	Lck	P	R	0.024	−0.026	0.002
226	c-Src	H	W	0.014	0.013	−0.027
226	c-Src	H	Y	−0.024	0.026	−0.002
247	Lck	R	Q	0.009	−0.026	0.017
248	c-Src	A	N	0.039	−0.01	−0.029
248	c-Src	A	T	0.032	−0.016	−0.016

Table 4.3: Specificity-determining variants.

BIBLIOGRAPHY

- Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2, 19–25. <https://doi.org/10.1016/j.softx.2015.06.001>
- Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fannesbeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., Osthege, M., Vieira, R., Wiecki, T., & Zinkov, R. (2023). PyMC: A modern, and comprehensive probabilistic programming framework in python. *PeerJ Computer Science*, 9, e1516. <https://doi.org/10.7717/peerj-cs.1516>
- Ahler, E., Register, A. C., Chakraborty, S., Fang, L., Dieter, E. M., Sitko, K. A., Vidadala, R. S. R., Trevillian, B. M., Golkowski, M., Gelman, H., Stephany, J. J., Rubin, A. F., Merritt, E. A., Fowler, D. M., & Maly, D. J. (2019). A combined approach reveals a regulatory mechanism coupling src's kinase activity, localization, and phosphotransferase-independent functions. *Molecular Cell*, 74(2), 393–408.e20. <https://doi.org/10.1016/j.molcel.2019.02.003>
- Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16(12), 1315–1322. <https://doi.org/10.1038/s41592-019-0598-1>
- Alshafie, W., Fotouhi, M., Shlaifer, I., Ayoubi, R., Edwards, A. M., Durcan, T. M., McPherson, P. S., & Laflamme, C. (2022). Identification of highly specific antibodies for serine/threonine-protein kinase TBK1 for use in immunoblot, immunoprecipitation and immunofluorescence. *F1000Research*, 11, 977. <https://doi.org/10.12688/f1000research.124632.2>
- Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S. G., & Pandey, A. (2007). A curated compendium of phosphorylation motifs. *Nature Biotechnology*, 25(3), 285–286. <https://doi.org/10.1038/nbt0307-285>
- Amit, S., Hatzubai, A., Birman, Y., Andersen, J. S., Ben-Shushan, E., Mann, M., Ben-Neriah, Y., & Alkalay, I. (2002). Axin-mediated CKI phosphorylation of β -catenin at ser 45: A molecular switch for the wnt pathway. *Genes & Development*, 16(9), 1066–1076. <https://doi.org/10.1101/gad.230302>
- Amorosi, C. J., Chiasson, M. A., McDonald, M. G., Wong, L. H., Sitko, K. A., Boyle, G., Kowalski, J. P., Rettie, A. E., Fowler, D. M., & Dunham, M. J. (2021). Massively parallel characterization of CYP2c9 variant enzyme activity and abundance. *The American Journal of Human Genetics*, 108(9), 1735–1751. <https://doi.org/10.1016/j.ajhg.2021.07.001>

- Anjum, S., & Gurari, D. (2020). CTMC: Cell tracking with mitosis detection dataset challenge. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 4228–4237. <https://doi.org/10.1109/CVPRW50498.2020.00499>
- Attwood, M. M., Fabbro, D., Sokolov, A. V., Knapp, S., & Schiöth, H. B. (2021). Trends in kinase drug discovery: Targets, indications and inhibitor design. *Nature Reviews Drug Discovery*, 20(11), 839–861. <https://doi.org/10.1038/s41573-021-00252-y>
- Bachl, S., Talbot, A., Ching, C., Kale, N., Li, Z., Diolaiti, M., Chen, A. P., Muldoon, J., Liu, C., Martin, T., Eyquem, J., & Carnevale, J. (2023). Ablation of cullin-5 in primary human t cells improves tumor killing and persistence in BCMA-targeting CAR-t cells in a multiple myeloma model. *Blood*, 142, 99–99. <https://doi.org/10.1182/blood-2023-182565>
- Barefield, D., & Sadayappan, S. (2010). Phosphorylation and function of cardiac myosin binding protein-c in health and disease. *Journal of Molecular and Cellular Cardiology*, 48(5), 866–875. <https://doi.org/10.1016/j.yjmcc.2009.11.014>
- Beltran, A., Faure, A. J., & Lehner, B. (2024, April 29). The allosteric landscape of the src kinase. <https://doi.org/10.1101/2024.04.26.591297>
- Berger, H., Travis, S., & Welsh, M. (1993). Regulation of the cystic fibrosis transmembrane conductance regulator cl⁻ channel by specific protein kinases and protein phosphatases. *Journal of Biological Chemistry*, 268(3), 2037–2047. [https://doi.org/10.1016/S0021-9258\(18\)53959-4](https://doi.org/10.1016/S0021-9258(18)53959-4)
- Bertacchini, J., Beretti, F., Cenni, V., Guida, M., Gibellini, F., Mediani, L., Marin, O., Maraldi, N. M., De Pol, A., Lattanzi, G., Cocco, L., & Marmioli, S. (2013). The protein kinase akt/PKB regulates both prelamin a degradation and *Lmna* gene expression. *The FASEB Journal*, 27(6), 2145–2155. <https://doi.org/10.1096/fj.12-218214>
- Bibi, N., Parveen, Z., & Rashid, S. (2013). Identification of potential plk1 targets in a cell-cycle specific proteome through structural dynamics of kinase and polo box-mediated interactions (K. S. Lee, Ed.). *PLoS ONE*, 8(8), e70843. <https://doi.org/10.1371/journal.pone.0070843>
- Blume-Jensen, P., Wernstedt, C., Heldin, C.-H., & Rönstrand, L. (1995). Identification of the major phosphorylation sites for protein kinase c in kit/stem cell factor receptor in vitro and in intact cells. *Journal of Biological Chemistry*, 270(23), 14192–14200. <https://doi.org/10.1074/jbc.270.23.14192>
- Bochinski, E., Eiselein, V., & Sikora, T. (2017). High-speed tracking-by-detection without using image information. *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6. <https://doi.org/10.1109/AVSS.2017.8078516>

- Bogomolovas, J., Gasch, A., Simkovic, F., Rigden, D. J., Labeit, S., & Mayans, O. (2014). Titin kinase is an inactive pseudokinase scaffold that supports MuRF1 recruitment to the sarcomeric m-line. *Open Biology*, 4(5), 140041. <https://doi.org/10.1098/rsob.140041>
- Boon, S. S., Lee, Y. C., Yip, K. L., Luk, H. Y., Xiao, C., Yim, M. K., Chen, Z., & Chan, P. K. S. (2023). Interaction between human papillomavirus-encoded e6 protein and AurB induces cell immortalization and proliferation—a potential target of intervention. *Cancers*, 15(9), 2465. <https://doi.org/10.3390/cancers15092465>
- Bouhaddou, M., Memon, D., Meyer, B., White, K. M., Rezelj, V. V., Correa Marroero, M., Polacco, B. J., Melnyk, J. E., Ulferts, S., Kaake, R. M., Batra, J., Richards, A. L., Stevenson, E., Gordon, D. E., Rojic, A., Obernier, K., Fabius, J. M., Soucheray, M., Miorin, L., . . . Krogan, N. J. (2020). The global phosphorylation landscape of SARS-CoV-2 infection. *Cell*, 182(3), 685–712.e19. <https://doi.org/10.1016/j.cell.2020.06.034>
- Bradley, D., & Beltrao, P. (2019). Evolution of protein kinase substrate recognition at the active site (B. E. Turk, Ed.). *PLOS Biology*, 17(6), e3000341. <https://doi.org/10.1371/journal.pbio.3000341>
- Bradley, D., Viéitez, C., Rajeeve, V., Selkrig, J., Cutillas, P. R., & Beltrao, P. (2021). Sequence and structure-based analysis of specificity determinants in eukaryotic protein kinases. *Cell Reports*, 34(2), 108602. <https://doi.org/10.1016/j.celrep.2020.108602>
- 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(5981), 1043–1046. <https://doi.org/10.1126/science.1176495>
- Brenan, L., Andreev, A., Cohen, O., Pantel, S., Kamburov, A., Cacchiarelli, D., Persky, N. S., Zhu, C., Bagul, M., Goetz, E. M., Burgin, A. B., Garraway, L. A., Getz, G., Mikkelsen, T. S., Piccioni, F., Root, D. E., & Johannessen, C. M. (2016). Phenotypic characterization of a comprehensive set of MAPK1/ERK2 missense mutants. *Cell Reports*, 17(4), 1171–1183. <https://doi.org/10.1016/j.celrep.2016.09.061>
- Brushia, R., J. (1999). Phosphorylase kinase: The complexity of its regulation is reflected in the complexity of its structure. *Frontiers in Bioscience*, 4(1), d618. <https://doi.org/10.2741/Brushia>
- 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(3), 504–520.e9. <https://doi.org/10.1016/j.molcel.2020.07.001>
- Bullock, A. N., Das, S., Debreczeni, J. É., Rellos, P., Fedorov, O., Niesen, F. H., Guo, K., Papagrigoriou, E., Amos, A. L., Cho, S., Turk, B. E., Ghosh, G., & Knapp, S. (2009). Kinase domain insertions define distinct roles of CLK kinases in SR protein phosphorylation. *Structure*, 17(3), 352–362. <https://doi.org/10.1016/j.str.2008.12.023>
- Burke, J. R., Pattoli, M. A., Gregor, K. R., Brassil, P. J., MacMaster, J. F., McIntyre, K. W., Yang, X., Iotzova, V. S., Clarke, W., Strnad, J., Qiu, Y., & Zusi, F. C. (2003). BMS-345541 is a highly selective inhibitor of I κ B kinase that binds at an allosteric site of the enzyme and blocks NF- κ b-dependent transcription in mice*. *Journal of Biological Chemistry*, 278(3), 1450–1456. <https://doi.org/10.1074/jbc.M209677200>
- Cai, W., Sakaguchi, M., Kleinriders, A., Gonzalez-Del Pino, G., Dreyfuss, J. M., O'Neill, B. T., Ramirez, A. K., Pan, H., Winnay, J. N., Boucher, J., Eck, M. J., & Kahn, C. R. (2017). Domain-dependent effects of insulin and IGF-1 receptors on signalling and gene expression. *Nature Communications*, 8(1), 14892. <https://doi.org/10.1038/ncomms14892>
- Carles, F., Bourg, S., Meyer, C., & Bonnet, P. (2018). PKIDB: A curated, annotated and updated database of protein kinase inhibitors in clinical trials. *Molecules*, 23(4), 908. <https://doi.org/10.3390/molecules23040908>
- Carlson, C. R., Asfaha, J. B., Ghent, C. M., Howard, C. J., Hartooni, N., Safari, M., Frankel, A. D., & Morgan, D. O. (2020). Phosphoregulation of phase separation by the SARS-CoV-2 n protein suggests a biophysical basis for its dual functions. *Molecular Cell*, 80(6), 1092–1103.e4. <https://doi.org/10.1016/j.molcel.2020.11.025>
- Carnevale, J., Shifrut, E., Kale, N., Nyberg, W. A., Blaeschke, F., Chen, Y. Y., Li, Z., Bapat, S. P., Diolaiti, M. E., O'Leary, P., Vedova, S., Belk, J., Daniel, B., Roth, T. L., Bachl, S., Anido, A. A., Prinzing, B., Ibañez-Vega, J., Lange, S., . . . Marson, A. (2022). RASA2 ablation in t cells boosts antigen sensitivity and long-term function. *Nature*, 609(7925), 174–182. <https://doi.org/10.1038/s41586-022-05126-w>
- Castro-Rodrigues, A. F., Zhao, Y., Fonseca, F., Gabant, G., Cadene, M., Robertson, G. A., & Morais-Cabral, J. H. (2018). The interaction between the drosophila EAG potassium channel and the protein kinase CaMKII involves an extensive interface at the active site of the kinase. *Journal of Molecular Biology*, 430(24), 5029–5049. <https://doi.org/10.1016/j.jmb.2018.10.015>
- Cavanagh, P. E., Xue, A. G., Dai, S., Qiang, A., Matsui, T., & Ting, A. Y. (2025, May 3). Computational design of conformation-biasing mutations to alter protein functions. <https://doi.org/10.1101/2025.05.03.652001>

- Chai Discovery, Boitreaud, J., Dent, J., McPartlon, M., Meier, J., Reis, V., Rogozhnikov, A., & Wu, K. (2024, October 11). Chai-1: Decoding the molecular interactions of life. <https://doi.org/10.1101/2024.10.10.615955>
- Chapagai, D., Strebhardt, K., Wyatt, M. D., & McInnes, C. (2025). Structural regulation of PLK1 activity: Implications for cell cycle function and drug discovery. *Cancer Gene Therapy*, 32(6), 608–621. <https://doi.org/10.1038/s41417-025-00907-7>
- Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., Alföldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., Poterba, T., Wilson, M. W., Tarasova, Y., Phu, W., Grant, R., Yohannes, M. T., Koenig, Z., Farjoun, Y., Banks, E., . . . Karczewski, K. J. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, 625(7993), 92–100. <https://doi.org/10.1038/s41586-023-06045-0>
- Chen, X., Zaro, J. L., & Shen, W.-C. (2013). Fusion protein linkers: Property, design and functionality. *Advanced Drug Delivery Reviews*, 65(10), 1357–1369. <https://doi.org/10.1016/j.addr.2012.09.039>
- Cheng, J., Bendjama, K., Rittner, K., & Malone, B. (2021). BERTMHC: Improved MHC–peptide class II interaction prediction with transformer and multiple instance learning (P. L. Martelli, Ed.). *Bioinformatics*, 37(22), 4172–4179. <https://doi.org/10.1093/bioinformatics/btab422>
- Chirino, L. M., Kumar, S., Okumura, M., Sterner, D. E., Mattern, M., Butt, T. R., & Kambayashi, T. (2020). TAM receptors attenuate murine NK-cell responses via e3 ubiquitin ligase cbl-b. *European Journal of Immunology*, 50(1), 48–55. <https://doi.org/10.1002/eji.201948204>
- Chitluri, K. K., & Emerson, I. A. (2024). The importance of protein domain mutations in cancer therapy. *Heliyon*, 10(6), e27655. <https://doi.org/10.1016/j.heliyon.2024.e27655>
- Corwin, T., Woodsmith, J., Apelt, F., Fontaine, J.-F., Meierhofer, D., Helmuth, J., Grossmann, A., Andrade-Navarro, M. A., Ballif, B. A., & Stelzl, U. (2017). Defining human tyrosine kinase phosphorylation networks using yeast as an in vivo model substrate. *Cell Systems*, 5(2), 128–139.e4. <https://doi.org/10.1016/j.cels.2017.08.001>
- Cuesta-Hernández, H. N., Contreras, J., Soriano-Maldonado, P., Sánchez-Wandelmer, J., Yeung, W., Martín-Hurtado, A., Muñoz, I. G., Kannan, N., Llimargas, M., Muñoz, J., & Plaza-Menacho, I. (2023). An allosteric switch between the activation loop and a c-terminal palindromic phosphomotif controls c-src function. *Nature Communications*, 14(1), 6548. <https://doi.org/10.1038/s41467-023-41890-7>
- Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., De Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H.,

- Kang, A., Sankaran, B., . . . Baker, D. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615), 49–56. <https://doi.org/10.1126/science.add2187>
- Dey, S., & Mondal, A. (2024). Unveiling the role of host kinases at different steps of influenza a virus life cycle (S. Mukhopadhyay, Ed.). *Journal of Virology*, 98(1), e01192–23. <https://doi.org/10.1128/jvi.01192-23>
- Deznabi, I., Arabaci, B., Koyutürk, M., & Tastan, O. (2020). DeepKinZero: Zero-shot learning for predicting kinase–phosphosite associations involving understudied kinases (A. Valencia, Ed.). *Bioinformatics*, 36(12), 3652–3661. <https://doi.org/10.1093/bioinformatics/btaa013>
- Diella, F., Gould, C. M., Chica, C., Via, A., & Gibson, T. J. (2008). Phospho.ELM: A database of phosphorylation sites—update 2008. *Nucleic Acids Research*, 36, D240–D244. <https://doi.org/10.1093/nar/gkm772>
- Dong, T., Kan, C., Devkota, K., & Singh, R. (2024, September 30). Allo-allo: Data-efficient prediction of allosteric sites. <https://doi.org/10.1101/2024.09.28.615583>
- Dou, Y., Yao, B., & Zhang, C. (2014). PhosphoSVM: Prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino Acids*, 46(6), 1459–1469. <https://doi.org/10.1007/s00726-014-1711-5>
- Duplay, P., Alcover, A., Fargeas, C., Sékaly, R. P., & Branton, P. E. (1996). An activated epidermal growth factor receptor/lck chimera restores early t cell receptor-mediated calcium response in a CD45-deficient t cell line. *Journal of Biological Chemistry*, 271(30), 17896–17902. <https://doi.org/10.1074/jbc.271.30.17896>
- Durocher, D., Henckel, J., Fersht, A. R., & Jackson, S. P. (1999). The FHA domain is a modular phosphopeptide recognition motif. *Molecular Cell*, 4(3), 387–394. [https://doi.org/10.1016/S1097-2765\(00\)80340-8](https://doi.org/10.1016/S1097-2765(00)80340-8)
- Edlund, C., Jackson, T. R., Khalid, N., Bevan, N., Dale, T., Dengel, A., Ahmed, S., Trygg, J., & Sjögren, R. (2021). LIVECell—a large-scale dataset for label-free live cell segmentation. *Nature Methods*, 18(9), 1038–1045. <https://doi.org/10.1038/s41592-021-01249-6>
- Estevam, G. O., Linossi, E., Rao, J., Macdonald, C. B., Ravikumar, A., Chrispens, K. M., Capra, J. A., Coyote-Maestas, W., Pimentel, H., Collisson, E. A., Jura, N., & Fraser, J. S. (2025). Mapping kinase domain resistance mechanisms for the MET receptor tyrosine kinase via deep mutational scanning. *eLife*, 13, RP101882. <https://doi.org/10.7554/eLife.101882.3>
- Fang, R., Wang, C., Jiang, Q., Lv, M., Gao, P., Yu, X., Mu, P., Zhang, R., Bi, S., Feng, J.-M., & Jiang, Z. (2017). NEMO-IKK β are essential for IRF3 and NF- κ b activation in the cGAS-STING pathway. *Journal of Immunology*, 199(9), 3222–3233. <https://doi.org/10.4049/jimmunol.1700699>

- Faure, A. J., Domingo, J., Schmiedel, J. M., Hidalgo-Carcedo, C., Diss, G., & Lehner, B. (2022). Mapping the energetic and allosteric landscapes of protein binding domains. *Nature*, 604(7904), 175–183. <https://doi.org/10.1038/s41586-022-04586-4>
- Faure, A. J., & Lehner, B. (2024). MoCHI: Neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis, and allostery from deep mutational scanning data. *Genome Biology*, 25(1), 303. <https://doi.org/10.1186/s13059-024-03444-y>
- Finneran, P., Soucheray, M., Wilson, C., Otten, R., Buosi, V., Krogan, N. J., Swaney, D. L., Theobald, D. L., & Kern, D. (2020, August 12). Bimodal evolution of src and abl kinase substrate specificity revealed using mammalian cell extract as substrate pool. <https://doi.org/10.1101/2020.08.12.248104>
- Fitzgerald, K. A., McWhirter, S. M., Faia, K. L., Rowe, D. C., Latz, E., Golenbock, D. T., Coyle, A. J., Liao, S.-M., & Maniatis, T. (2003). IKKepsilon and TBK1 are essential components of the IRF3 signaling pathway. *Nature Immunology*, 4(5), 491–496. <https://doi.org/10.1038/ni921>
- Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: A new style of protein science. *Nature Methods*, 11(8), 801–807. <https://doi.org/10.1038/nmeth.3027>
- Frank, F., Keen, M. M., Rao, A., Bassit, L., Liu, X., Bowers, H. B., Patel, A. B., Cato, M. L., Sullivan, J. A., Greenleaf, M., Piantadosi, A., Lam, W. A., Hudson, W. H., & Ortlund, E. A. (2022). Deep mutational scanning identifies SARS-CoV-2 nucleocapsid escape mutations of currently available rapid antigen tests. *Cell*, 185(19), 3603–3616.e13. <https://doi.org/10.1016/j.cell.2022.08.010>
- Gelman, S., Johnson, B., Freschlin, C. R., Sharma, A., D’Costa, S., Peters, J., Gitter, A., & Romero, P. A. (2025). Biophysics-based protein language models for protein engineering. *Nature Methods*, 22(9), 1868–1879. <https://doi.org/10.1038/s41592-025-02776-2>
- Gerasimavicius, L., Livesey, B. J., & Marsh, J. A. (2023). Correspondence between functional scores from deep mutational scans and predicted effects on protein stability. *Protein Science*, 32(7), e4688. <https://doi.org/10.1002/pro.4688>
- Ghosh, S., & Karin, M. (2002). Missing pieces in the NF- κ b puzzle. *Cell*, 109(2), S81–S96. [https://doi.org/10.1016/S0092-8674\(02\)00703-1](https://doi.org/10.1016/S0092-8674(02)00703-1)
- Gizzio, J., Thakur, A., Haldane, A., & Levy, R. M. (2022). Evolutionary divergence in the conformational landscapes of tyrosine vs serine/threonine kinases. *eLife*, 11, e83368. <https://doi.org/10.7554/eLife.83368>
- Glaser, M., & Brägelmann, J. (2025, February 7). ESM-effect: An effective and efficient fine-tuning framework towards accurate prediction of mutation’s functional effect. <https://doi.org/10.1101/2025.02.03.635741>

- Gomez, S. M., Axtman, A. D., Willson, T. M., Major, M. B., Townsend, R. R., Sorger, P. K., & Johnson, G. L. (2024). Illuminating function of the understudied druggable kinome. *Drug Discovery Today*, 29(3), 103881. <https://doi.org/10.1016/j.drudis.2024.103881>
- Gonzalez, A., Pon, A., Servage, K., Pawłowski, K., Han, Y., & Sreelatha, A. (2025). A repurposed AMP binding domain reveals mitochondrial protein AMPylation as a regulator of cellular metabolism. *Nature Communications*, 16(1), 7863. <https://doi.org/10.1038/s41467-025-63014-z>
- Graham, M. E., Ruma-Haynes, P., Capes-Davis, A. G., Dunn, J. M., Tan, T. C., Valova, V. A., Robinson, P. J., & Jeffrey, P. L. (2004). Multisite phosphorylation of doublecortin by cyclin-dependent kinase 5. *Biochemical Journal*, 381(2), 471–481. <https://doi.org/10.1042/BJ20040324>
- Greenwald, E. C., Mehta, S., & Zhang, J. (2018). Genetically encoded fluorescent biosensors illuminate the spatiotemporal regulation of signaling networks. *Chemical Reviews*, 118(24), 11707–11794. <https://doi.org/10.1021/acs.chemrev.8b00333>
- Grossmann, A., Benlasfer, N., Birth, P., Hegele, A., Wachsmuth, F., Apelt, L., & Stelzl, U. (2015). Phospho-tyrosine dependent protein–protein interaction network. *Molecular Systems Biology*, 11(3), 794. <https://doi.org/10.15252/msb.20145968>
- Groves, T., Smiley, P., Cooke, M. P., Forbush, K., Perlmutter, R. M., & Guidos, C. J. (1996). Fyn can partially substitute for lck in t lymphocyte development. *Immunity*, 5(5), 417–428. [https://doi.org/10.1016/S1074-7613\(00\)80498-7](https://doi.org/10.1016/S1074-7613(00)80498-7)
- Guggino, W. B., & Stanton, B. A. (2006). New insights into cystic fibrosis: Molecular switches that regulate CFTR. *Nature Reviews Molecular Cell Biology*, 7(6), 426–436. <https://doi.org/10.1038/nrm1949>
- Gully, C. P., Velazquez-Torres, G., Shin, J.-H., Fuentes-Mattei, E., Wang, E., Carlock, C., Chen, J., Rothenberg, D., Adams, H. P., Choi, H. H., Guma, S., Phan, L., Chou, P.-C., Su, C.-H., Zhang, F., Chen, J.-S., Yang, T.-Y., Yeung, S.-C. J., & Lee, M.-H. (2012). Aurora b kinase phosphorylates and instigates degradation of p53. *Proceedings of the National Academy of Sciences*, 109(24). <https://doi.org/10.1073/pnas.1110287109>
- Guo, Y., Zhang, X., Xu, Q., Gong, F., Shi, X., Li, C., Huang, R., Nie, F., Zhu, W., Li, J., Tang, J., Li, R., Zhang, L., Chen, L., & Ma, R. Z. (2019). Human papillomavirus 16 oncoprotein e7 retards mitotic progression by blocking mps1-MAP4 signaling cascade. *Oncogene*, 38(31), 5959–5970. <https://doi.org/10.1038/s41388-019-0851-1>
- Haddox, H. K., Dingens, A. S., Hilton, S. K., Overbaugh, J., & Bloom, J. D. (2018). Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife*, 7, e34420. <https://doi.org/10.7554/eLife.34420>

- Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th python in science conference* (pp. 11–15).
- Hauck, F., Randriamampita, C., Martin, E., Gerart, S., Lambert, N., Lim, A., Soulier, J., Maciorowski, Z., Touzot, F., Moshous, D., Quartier, P., Heritier, S., Blanche, S., Rieux-Laucat, F., Brousse, N., Callebaut, I., Veillette, A., Hivroz, C., Fischer, A., . . . Picard, C. (2012). Primary t-cell immunodeficiency with immunodysregulation caused by autosomal recessive LCK deficiency. *Journal of Allergy and Clinical Immunology*, 130(5), 1144–1152.e11. <https://doi.org/10.1016/j.jaci.2012.07.029>
- Hayes, T. K., Aquilanti, E., Persky, N. S., Yang, X., Kim, E. E., Brenan, L., Goodale, A. B., Alan, D., Sharpe, T., Shue, R. E., Westlake, L., Golomb, L., Silverman, B. R., Morris, M. D., Fisher, T. R., Beyene, E., Li, Y. Y., Cherniack, A. D., Piccioni, F., . . . Meyerson, M. (2024). Comprehensive mutational scanning of EGFR reveals TKI sensitivities of extracellular domain mutants. *Nature Communications*, 15(1), 2742. <https://doi.org/10.1038/s41467-024-45594-4>
- Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., & Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20(1), 723. <https://doi.org/10.1186/s12859-019-3220-8>
- Hobbs, H. T., Shah, N. H., Shoemaker, S. R., Amacher, J. F., Marqusee, S., & Kuriyan, J. (2022). Saturation mutagenesis of a predicted ancestral syk-family kinase. *Protein Science*, 31(10), e4411. <https://doi.org/10.1002/pro.4411>
- Hom, N., Gentles, L., Bloom, J. D., & Lee, K. K. (2019). Deep mutational scan of the highly conserved influenza a virus m1 matrix protein reveals substantial intrinsic mutational tolerance (R. E. Dutch, Ed.). *Journal of Virology*, 93(13), e00161–19. <https://doi.org/10.1128/JVI.00161-19>
- Horkova, V., Drobek, A., Paprckova, D., Niederlova, V., Prasai, A., Uleri, V., Glatzova, D., Kraller, M., Cesnekova, M., Janusova, S., Salyova, E., Tsyklauri, O., Kadlec, T. A., Krizova, K., Platzer, R., Schober, K., Busch, D. H., Weiss, A., Huppa, J. B., & Stepanek, O. (2023). Unique roles of co-receptor-bound LCK in helper and cytotoxic t cells. *Nature Immunology*, 24(1), 174–185. <https://doi.org/10.1038/s41590-022-01366-0>
- 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(6), 603–604. <https://doi.org/10.1038/nmeth.2968>

- 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. <https://doi.org/10.1093/nar/gku1267>
- Hoshi, M., Nishida, E., Miyata, Y., Sakai, H., Miyoshi, T., Ogawara, H., & Akiyama, T. (1987). Protein kinase c phosphorylates tau and induces its functional alterations. *FEBS Letters*, 217(2), 237–241. [https://doi.org/10.1016/0014-5793\(87\)80670-1](https://doi.org/10.1016/0014-5793(87)80670-1)
- Howard, C. J., Abell, N. S., Warneford-Thomson, R. R., Mahdavi, E., Su, A. L., Resnick, C., Mohammed, N., Thompson, E. M., Holzinger, E. R., Catalano, K., Hukku, A., Mintier, G. A., MacKenzie, M., Jiang, B. L., Rabago, D. B., Chan, A., Ntimi, C., Weiler, K. N., Wilson, S. C., . . . Dickel, D. E. (2025, October 12). Deep mutational scanning reveals pharmacologically relevant insights into TYK2 signaling and disease. <https://doi.org/10.1101/2025.10.11.681520>
- Howard, M. K., Hoppe, N., Huang, X.-P., Mitrovic, D., Billesbølle, C. B., Macdonald, C. B., Mehrotra, E., Rockefeller Grimes, P., Trinidad, D. D., Delemotte, L., English, J. G., Coyote-Maestas, W., & Manglik, A. (2025). Molecular basis of proton sensing by g protein-coupled receptors. *Cell*, 188(3), 671–687.e20. <https://doi.org/10.1016/j.cell.2024.11.036>
- Hu, J., Rho, H.-S., Newman, R. H., Zhang, J., Zhu, H., & Qian, J. (2014). PhosphoNetworks: A database for human phosphorylation networks. *Bioinformatics*, 30(1), 141–142. <https://doi.org/10.1093/bioinformatics/btt627>
- Huang, C., Zhang, Z., Chen, L., Lee, H. W., Ayrapetov, M. K., Zhao, T. C., Hao, Y., Gao, J., Yang, C., Mehta, G. U., Zhuang, Z., Zhang, X., Hu, G., & Chin, Y. E. (2018). Acetylation within the n- and c-terminal domains of src regulates distinct roles of STAT3-mediated tumorigenesis. *Cancer Research*, 78(11), 2825–2838. <https://doi.org/10.1158/0008-5472.CAN-17-2314>
- Huttlin, E. L., Jedrychowski, M. P., Elias, J. E., Goswami, T., Rad, R., Beausoleil, S. A., Villén, J., Haas, W., Sowa, M. E., & Gygi, S. P. (2010). A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell*, 143(7), 1174–1189. <https://doi.org/10.1016/j.cell.2010.12.001>
- Irby, R. B., Mao, W., Coppola, D., Kang, J., Loubeau, J. M., Trudeau, W., Karl, R., Fujita, D. J., Jove, R., & Yeatman, T. J. (1999). Activating SRC mutation in a subset of advanced human colon cancers. *Nature Genetics*, 21(2), 187–190. <https://doi.org/10.1038/5971>
- Iwashima, M., Irving, B. A., Van Oers, N. S. C., Chan, A. C., & Weiss, A. (1994). Sequential interactions of the TCR with two distinct cytoplasmic tyrosine kinases. *Science*, 263(5150), 1136–1139. <https://doi.org/10.1126/science.7509083>

- Jehle, S., Kunowska, N., Benlasfer, N., Woodsmith, J., Weber, G., Wahl, M. C., & Stelzl, U. (2022). A human kinase yeast array for the identification of kinases modulating phosphorylation-dependent protein–protein interactions. *Molecular Systems Biology*, 18(3), e10820. <https://doi.org/10.15252/msb.202110820>
- Jha, K., Shonai, D., Parekh, A., Uezu, A., Fujiyama, T., Yamamoto, H., Parameswaran, P., Yanagisawa, M., Singh, R., & Soderling, S. H. (2025, April 28). Deep learning-coupled proximity proteomics to deconvolve kinase signaling in vivo. <https://doi.org/10.1101/2025.04.27.650849>
- Ji, J., Lee, H., Argiropoulos, B., Dorrani, N., Mann, J., Martinez-Agosto, J. A., Gomez-Ospina, N., Gallant, N., Bernstein, J. A., Hudgins, L., Slattery, L., Isidor, B., Le Caignec, C., David, A., Obersztyn, E., Wiśniowiecka-Kowalik, B., Fox, M., Deignan, J. L., Vilain, E., . . . Quintero-Rivera, F. (2015). DYRK1a haploinsufficiency causes a new recognizable syndrome with microcephaly, intellectual disability, speech impairment, and distinct facies. *European Journal of Human Genetics*, 23(11), 1473–1481. <https://doi.org/10.1038/ejhg.2015.71>
- Ji, Y., Liu, Y., Sun, C., Yu, L., Wang, Z., Du, X., Yang, W., Zhang, C., Tao, C., Wang, J., Yang, X., Di, S., & Huang, Y. (2021). ADCK1 activates the β -catenin/TCF signaling pathway to promote the growth and migration of colon cancer cells. *Cell Death & Disease*, 12(4), 354. <https://doi.org/10.1038/s41419-021-03624-9>
- Johnson, J. L., Yaron, T. M., Huntsman, E. M., Kerelsky, A., Song, J., Regev, A., Lin, T.-Y., Liberatore, K., Cizin, D. M., Cohen, B. M., Vasan, N., Ma, Y., Krismer, K., Robles, J. T., Van De Kooij, B., Van Vlimmeren, A. E., Andrée-Busch, N., Käufer, N. F., Dorovkov, M. V., . . . Cantley, L. C. (2023). An atlas of substrate specificities for the human serine/threonine kinome. *Nature*, 613(7945), 759–766. <https://doi.org/10.1038/s41586-022-05575-3>
- Jones, E. M., Lubock, N. B., Venkatakrishnan, A., Wang, J., Tseng, A. M., Paggi, J. M., Latorraca, N. R., Cancilla, D., Satyadi, M., Davis, J. E., Babu, M. M., Dror, R. O., & Kosuri, S. (2020). Structural and functional characterization of G protein–coupled receptors with deep mutational scanning. *eLife*, 9, e54895. <https://doi.org/10.7554/eLife.54895>
- Kannan, G. R., Hie, B. L., & Kim, P. S. (2024, October 3). Single-sequence, structure free allosteric residue prediction with protein language models. <https://doi.org/10.1101/2024.10.03.616547>
- Kassenbrock, C. K., & Anderson, S. M. (2004). Regulation of ubiquitin protein ligase activity in c-cbl by phosphorylation-induced conformational change and constitutive activation by tyrosine to glutamate point mutations. *Journal of Biological Chemistry*, 279(27), 28017–28027. <https://doi.org/10.1074/jbc.M404114200>

- Kemmoku, H., Kuchitsu, Y., Mukai, K., & Taguchi, T. (2022). Specific association of TBK1 with the trans-golgi network following STING stimulation. *Cell Structure and Function*, 47(1), 19–30. <https://doi.org/10.1247/csf.21080>
- Ker, D. F. E., Eom, S., Sanami, S., Bise, R., Pascale, C., Yin, Z., Huh, S.-i., Osuna-Highley, E., Junkers, S. N., Helfrich, C. J., Liang, P. Y., Pan, J., Jeong, S., Kang, S. S., Liu, J., Nicholson, R., Sandbothe, M. F., Van, P. T., Liu, A., . . . Campbell, P. G. (2018). Phase contrast time-lapse microscopy datasets with automated and manual cell tracking annotations. *Scientific Data*, 5(1), 180237. <https://doi.org/10.1038/sdata.2018.237>
- Kida, A., Kakihana, K., Kotani, S., Kurosu, T., & Miura, O. (2007). Glycogen synthase kinase-3 β and p38 phosphorylate cyclin d2 on thr280 to trigger its ubiquitin/proteasome-dependent degradation in hematopoietic cells. *Oncogene*, 26(46), 6630–6640. <https://doi.org/10.1038/sj.onc.1210490>
- Kim, J.-H., Seo, Y., Jo, M., Jeon, H., Lee, W.-H., Yachie, N., Zhong, Q., Vidal, M., Roth, F. P., & Suk, K. (2020). Yeast-based genetic interaction analysis of human kinome. *Cells*, 9(5), 1156. <https://doi.org/10.3390/cells9051156>
- Kizilirmak, C., Bianchi, M. E., & Zambrano, S. (2022). Insights on the NF- κ b system using live cell imaging: Recent developments and future perspectives. *Frontiers in Immunology*, 13. <https://doi.org/10.3389/fimmu.2022.886127>
- Kou, L., Markowsky, G., & Berman, L. (1981). A fast algorithm for steiner trees. *Acta Informatica*, 15(2), 141–145. <https://doi.org/10.1007/BF00288961>
- Kuo, H.-P., Lee, D.-F., Xia, W., Wei, Y., & Hung, M.-C. (2009). TNF α induces HIF-1 α expression through activation of IKK β . *Biochemical and biophysical research communications*, 389(4), 640–644. <https://doi.org/10.1016/j.bbrc.2009.09.042>
- Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Mad-dipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., . . . Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46, D1062–D1067. <https://doi.org/10.1093/nar/gkx1153>
- Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42, D980–D985. <https://doi.org/10.1093/nar/gkt1113>
- Lanz, A.-L., Erdem, S., Ozcan, A., Ceylaner, G., Cansever, M., Ceylaner, S., Conca, R., Magg, T., Acuto, O., Latour, S., Klein, C., Patiroglu, T., Unal, E., Eken, A., & Hauck, F. (2024). A novel biallelic LCK variant resulting in profound t-cell immune deficiency and review of the literature. *Journal of Clinical Immunology*, 44(1), 1. <https://doi.org/10.1007/s10875-023-01602-8>

- Larsen, B. B., McMahon, T., Brown, J. T., Wang, Z., Radford, C. E., Crowe, J. E., Veessler, D., & Bloom, J. D. (2024, April 19). Functional and antigenic landscape of the nipah virus receptor binding protein. <https://doi.org/10.1101/2024.04.17.589977>
- Lasagna-Reeves, C. A., de Haro, M., Hao, S., Park, J., Rousseaux, M. W., Al-Ramahi, I., Jafar-Nejad, P., Vilanova-Velez, L., See, L., De Maio, A., Nitschke, L., Wu, Z., Troncoso, J. C., Westbrook, T. F., Tang, J., Botas, J., & Zoghbi, H. Y. (2016). Reduction of nuak1 decreases tau and reverses phenotypes in a tauopathy mouse model. *Neuron*, 92(2), 407–418. <https://doi.org/10.1016/j.neuron.2016.09.022>
- Lee, A., Rayner, S. L., De Luca, A., Gwee, S. S. L., Morsch, M., Sundaramoorthy, V., Shahheydari, H., Ragagnin, A., Shi, B., Yang, S., Williams, K. L., Don, E. K., Walker, A. K., Zhang, K. Y., Yerbury, J. J., Cole, N. J., Atkin, J. D., Blair, I. P., Molloy, M. P., & Chung, R. S. (2017). Casein kinase II phosphorylation of cyclin f at serine 621 regulates the lys48-ubiquitylation e3 ligase activity of the SCF^(cyclin F) complex. *Open Biology*, 7(10), 170058. <https://doi.org/10.1098/rsob.170058>
- Lee, G., Thangavel, R., Sharma, V. M., Litersky, J. M., Bhaskar, K., Fang, S. M., Do, L. H., Andreadis, A., Van Hoesen, G., & Ksiezak-Reding, H. (2004). Phosphorylation of tau by fyn: Implications for alzheimer's disease. *The Journal of Neuroscience*, 24(9), 2304–2312. <https://doi.org/10.1523/JNEUROSCI.4162-03.2004>
- Lee, J. M., Huddleston, J., Doud, M. B., Hooper, K. A., Wu, N. C., Bedford, T., & Bloom, J. D. (2018). Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human h3n2 influenza variants. *Proceedings of the National Academy of Sciences*, 115(35). <https://doi.org/10.1073/pnas.1806133115>
- Lee, M.-H., Korias, P., Qu, J., & Andreadis, S. T. (2009). JNK phosphorylates β -catenin and regulates adherens junctions. *The FASEB Journal*, 23(11), 3874–3883. <https://doi.org/10.1096/fj.08-117804>
- Lee, S., Shin, J., Kim, J.-S., Shin, J., Lee, S. K., & Park, H.-W. (2021). Targeting TBK1 attenuates LPS-induced NLRP3 inflammasome activation by regulating of mTORC1 pathways in trophoblasts. *Frontiers in Immunology*, 12, 743700. <https://doi.org/10.3389/fimmu.2021.743700>
- Lemke, M. C., Chen, M., Jang, S. S., Rader, M. T., Avala, N. R., Bauchle, C. J., Pluguez, R., Miller, C. J., Butler, J. D., Hargett, S. R., Lank, D. S., Seltzer, B. D., Stornetta, D. S., Fu, Z., Hsu, K.-L., & Harris, T. E. (2025). Characterization of mast2 kinase defines structural features, regulation, and substrates. *Journal of Biological Chemistry*, 301(12), 110922. <https://doi.org/10.1016/j.jbc.2025.110922>

- Lewis, S., Hempel, T., Jiménez-Luna, J., Gastegger, M., Xie, Y., Foong, A. Y. K., Satorras, V. G., Abdin, O., Veeling, B. S., Zaporozhets, I., Chen, Y., Yang, S., Foster, A. E., Schneuing, A., Nigam, J., Barbero, F., Stimper, V., Campbell, A., Yim, J., . . . Noé, F. (2025). Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761), eadv9817. <https://doi.org/10.1126/science.adv9817>
- Li, N., Lin, G., Zhang, H., Sun, J., Gui, M., Liu, Y., Li, W., Liu, J., & Tang, J. (2022). Src family kinases: A potential therapeutic target for acute kidney injury. *Biomolecules*, 12(7), 984. <https://doi.org/10.3390/biom12070984>
- Li, R., Liao, G., Nirujogi, R. S., Pinto, S. M., Shaw, P. G., Huang, T.-C., Wan, J., Qian, J., Gowda, H., Wu, X., Lv, D.-W., Zhang, K., Manda, S. S., Pandey, A., & Hayward, S. D. (2015). Phosphoproteomic profiling reveals epstein-barr virus protein kinase integration of DNA damage response and mitotic signaling (P. D. Ling, Ed.). *PLOS Pathogens*, 11(12), e1005346. <https://doi.org/10.1371/journal.ppat.1005346>
- Liao, X., Li, W., Zhou, H., Rajendran, B. K., Li, A., Ren, J., Luan, Y., Calderwood, D. A., Turk, B., Tang, W., Liu, Y., & Wu, D. (2024). The CUL5 e3 ligase complex negatively regulates central signaling pathways in CD8+ t cells. *Nature Communications*, 15(1), 603. <https://doi.org/10.1038/s41467-024-44885-0>
- Linden-Santangeli, N., Zhang, J., Kramer, B., & Rangamani, P. (2025). Increasing certainty in systems biology models using bayesian multimodel inference. *Nature Communications*, 16(1), 7416. <https://doi.org/10.1038/s41467-025-62415-4>
- Liu, J., Kimura, A., Baumann, C. A., & Saltiel, A. R. (2002). APS facilitates c-cbl tyrosine phosphorylation and GLUT4 translocation in response to insulin in 3t3-l1 adipocytes. *Molecular and Cellular Biology*, 22(11), 3599–3609. <https://doi.org/10.1128/MCB.22.11.3599-3609.2002>
- Liu, L., Yu, X.-Z., Li, T.-S., Song, L.-X., Chen, P.-L., Suo, T.-L., Li, Y.-H., Wang, S.-D., Chen, Y., Ren, Y.-M., Zhang, S.-P., Chang, Z.-J., & Fu, X.-Y. (2004). A novel protein tyrosine kinase NOK that shares homology with platelet-derived growth factor/fibroblast growth factor receptors induces tumorigenesis and metastasis in nude mice. *Cancer Research*, 64(10), 3491–3499. <https://doi.org/10.1158/0008-5472.CAN-03-2106>
- Liu, Q., Kaneko, S., Yang, L., Feldman, R. I., Nicosia, S. V., Chen, J., & Cheng, J. Q. (2004). Aurora-a abrogation of p53 DNA binding and transactivation activity by phosphorylation of serine 215. *Journal of Biological Chemistry*, 279(50), 52175–52182. <https://doi.org/10.1074/jbc.M406802200>
- Liu, S., Cai, X., Wu, J., Cong, Q., Chen, X., Li, T., Du, F., Ren, J., Wu, Y.-T., Grishin, N. V., & Chen, Z. J. (2015). Phosphorylation of innate immune

- adaptor proteins MAVS, STING, and TRIF induces IRF3 activation. *Science*, 347(6227), aaa2630. <https://doi.org/10.1126/science.aaa2630>
- Liu, X., Verma, A., Garcia, G., Ramage, H., Lucas, A., Myers, R. L., Michaelson, J. J., Coryell, W., Kumar, A., Charney, A. W., Kazanietz, M. G., Rader, D. J., Ritchie, M. D., Berrettini, W. H., Schultz, D. C., Cherry, S., Damoiseaux, R., Arumugaswami, V., & Klein, P. S. (2021). Targeting the coronavirus nucleocapsid protein through GSK-3 inhibition. *Proceedings of the National Academy of Sciences*, 118(42), e2113401118. <https://doi.org/10.1073/pnas.2113401118>
- Liu, Y., Zhang, W., Jang, H., & Nussinov, R. (2025). mTOR variants activation discovers PI3k-like cryptic pocket, expanding allosteric, mutant-selective inhibitor designs. *Journal of Chemical Information and Modeling*, 65(2), 966–980. <https://doi.org/10.1021/acs.jcim.4c02022>
- Lubner, J. M., Dodge-Kafka, K. L., Carlson, C. R., Church, G. M., Chou, M. F., & Schwartz, D. (2017). Cushing's syndrome mutant PKA^{L205R} exhibits altered substrate specificity. *FEBS Letters*, 591(3), 459–467. <https://doi.org/10.1002/1873-3468.12562>
- Lugagne, J.-B., Lin, H., & Dunlop, M. J. (2020). DeLTA: Automated cell segmentation, tracking, and lineage reconstruction using deep learning (A. R. Asthagiri, Ed.). *PLOS Computational Biology*, 16(4), e1007673. <https://doi.org/10.1371/journal.pcbi.1007673>
- Lui, V. G., Hoenig, M., Cabrera-Martinez, B., Baxter, R. M., Garcia-Perez, J. E., Bailey, O., Acharya, A., Lundquist, K., Capera, J., Matuszewicz, P., Hartl, F. A., D'Abramo, M., Alba, J., Jacobsen, E.-M., Niewolik, D., Lorenz, M., Pannicke, U., Schulz, A. S., Debatin, K.-M., ... Hsieh, E. W. (2024). A partial human LCK defect causes a t cell immunodeficiency with intestinal inflammation. *Journal of Experimental Medicine*, 221(1), e20230927. <https://doi.org/10.1084/jem.20230927>
- Lundby, A., Secher, A., Lage, K., Nordsborg, N. B., Dmytriiev, A., Lundby, C., & Olsen, J. V. (2012). Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. *Nature Communications*, 3(1), 876. <https://doi.org/10.1038/ncomms1871>
- Ma, R., Li, S., Li, W., Yao, L., Huang, H.-D., & Lee, T.-Y. (2023). KinasePhos 3.0: Redesign and expansion of the prediction on kinase-specific phosphorylation sites. *Genomics, Proteomics & Bioinformatics*, 21(1), 228–241. <https://doi.org/10.1016/j.gpb.2022.06.004>
- Ma, R., Du, B., Shi, C., Wang, L., Zeng, F., Han, J., Guan, H., Wang, Y., & Yan, K. (2025). Molecular basis for the regulation of human phosphorylase kinase by phosphorylation and ca²⁺. *Nature Communications*, 16(1), 3020. <https://doi.org/10.1038/s41467-025-58363-8>

- Magistretti, P. J., & Allaman, I. (2015). A cellular perspective on brain energy metabolism and functional imaging. *Neuron*, 86(4), 883–901. <https://doi.org/10.1016/j.neuron.2015.03.035>
- Malek, S., Chen, Y., Huxford, T., & Ghosh, G. (2001). $I\kappa B\beta$, but not $I\kappa B\alpha$, functions as a classical cytoplasmic inhibitor of NF- κ B dimers by masking both NF- κ B nuclear localization sequences in resting cells*. *Journal of Biological Chemistry*, 276(48), 45225–45235. <https://doi.org/10.1074/jbc.M105865200>
- Martinez, R., Defnet, A., & Shapiro, P. (2020). Avoiding or co-opting ATP inhibition: Overview of type III, IV, v, and VI kinase inhibitors. In P. Shapiro (Ed.), *Next generation kinase inhibitors* (pp. 29–59). Springer International Publishing. https://doi.org/10.1007/978-3-030-48283-1_3
- Massey, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253), 68–78. <https://doi.org/10.1080/01621459.1951.10500769>
- Matrone, C., Petrillo, F., Nasso, R., & Ferretti, G. (2020). Fyn tyrosine kinase as harmonizing factor in neuronal functions and dysfunctions. *International Journal of Molecular Sciences*, 21(12), 4444. <https://doi.org/10.3390/ijms21124444>
- McNeill, L., Salmond, R. J., Cooper, J. C., Carret, C. K., Cassady-Cain, R. L., Roche-Molina, M., Tandon, P., Holmes, N., & Alexander, D. R. (2007). The differential regulation of lck kinase phosphorylation sites by CD45 is critical for t cell receptor signaling responses. *Immunity*, 27(3), 425–437. <https://doi.org/10.1016/j.immuni.2007.07.015>
- Medvedev, K. E., Schaeffer, R. D., Pei, J., & Grishin, N. V. (2023). Pathogenic mutation hotspots in protein kinase domain structure. *Protein Science*, 32(9), e4750. <https://doi.org/10.1002/pro.4750>
- Messaoud-Nacer, Y., Culierier, E., Rose, S., Maillet, I., Rouxel, N., Briault, S., Ryffel, B., Quesniaux, V. F. J., & Togbe, D. (2022). STING agonist diABZI induces PANoptosis and DNA mediated acute respiratory distress syndrome (ARDS). *Cell Death & Disease*, 13(3), 269. <https://doi.org/10.1038/s41419-022-04664-5>
- Miller, C. J., & Turk, B. E. (2018). Homing in: Mechanisms of substrate targeting by protein kinases. *Trends in Biochemical Sciences*, 43(5), 380–394. <https://doi.org/10.1016/j.tibs.2018.02.009>
- Miller, W. T. (2003). Determinants of substrate recognition in nonreceptor tyrosine kinases. *Accounts of Chemical Research*, 36(6), 393–400. <https://doi.org/10.1021/ar020116v>
- Modi, V., & Dunbrack, R. L. (2019a). Defining a new nomenclature for the structures of active and inactive kinases. *Proceedings of the National Academy of Sciences*, 116(14), 6818–6827. <https://doi.org/10.1073/pnas.1814279116>

- Modi, V., & Dunbrack, R. L. (2019b). A structurally-validated multiple sequence alignment of 497 human protein kinase domains. *Scientific Reports*, 9(1), 19790. <https://doi.org/10.1038/s41598-019-56499-4>
- Mohallem, R., Schaser, A. J., & Aryal, U. K. (2024). Molecular signatures of neurodegenerative diseases identified by proteomic and phosphoproteomic analyses in aging mouse brain. *Molecular & Cellular Proteomics*, 23(9), 100819. <https://doi.org/10.1016/j.mcpro.2024.100819>
- Mok, J., Kim, P. M., Lam, H. Y. K., Piccirillo, S., Zhou, X., Jeschke, G. R., Sheridan, D. L., Parker, S. A., Desai, V., Jwa, M., Cameroni, E., Niu, H., Good, M., Remenyi, A., Ma, J.-L. N., Sheu, Y.-J., Sassi, H. E., Sopko, R., Chan, C. S. M., . . . Turk, B. E. (2010). Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. *Science Signaling*, 3(109). <https://doi.org/10.1126/scisignal.2000482>
- Nagao, H., Cai, W., Wewer Albrechtsen, N. J., Steger, M., Batista, T. M., Pan, H., Dreyfuss, J. M., Mann, M., & Kahn, C. R. (2021). Distinct signaling by insulin and IGF-1 receptors and their extra- and intracellular domains. *Proceedings of the National Academy of Sciences*, 118(17), e2019474118. <https://doi.org/10.1073/pnas.2019474118>
- Nakamura, N., Chin, H., Miyasaka, N., & Miura, O. (1996). An epidermal growth factor receptor/jak2 tyrosine kinase domain chimera induces tyrosine phosphorylation of stat5 and transduces a growth signal in hematopoietic cells. *Journal of Biological Chemistry*, 271(32), 19483–19488. <https://doi.org/10.1074/jbc.271.32.19483>
- Needham, E. J., Parker, B. L., Burykin, T., James, D. E., & Humphrey, S. J. (2019). Illuminating the dark phosphoproteome. *Science Signaling*, 12(565), eaau8645. <https://doi.org/10.1126/scisignal.aau8645>
- Ni, Q., Mehta, S., & Zhang, J. (2018). Live-cell imaging of cell signaling using genetically encoded fluorescent reporters. *The FEBS Journal*, 285(2), 203–219. <https://doi.org/10.1111/febs.14134>
- Nolen, B., Taylor, S., & Ghosh, G. (2004). Regulation of protein kinases. *Molecular Cell*, 15(5), 661–675. <https://doi.org/10.1016/j.molcel.2004.08.024>
- Nováček, V., McGauran, G., Matallanas, D., Blanco, A. V., Conca, P., Muñoz, E., Costabello, L., Kanakaraj, K., Nawaz, Z., Walsh, B., Mohamed, S. K., Vandenbussche, P.-Y., Ryan, C. J., Kolch, W., & Fey, D. (2020). Accurate prediction of kinase-substrate networks using knowledge graphs. *PLOS Computational Biology*, 16(12), e1007578. <https://doi.org/10.1371/journal.pcbi.1007578>
- Obenauer, J. C. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Research*, 31(13), 3635–3641. <https://doi.org/10.1093/nar/gkg584>

- Oruganty, K., Talathi, N. S., Wood, Z. A., & Kannan, N. (2013). Identification of a hidden strain switch provides clues to an ancient structural mechanism in protein kinases. *Proceedings of the National Academy of Sciences*, *110*(3), 924–929. <https://doi.org/10.1073/pnas.1207104110>
- Özden, C., Sloutsky, R., Mitsugi, T., Santos, N., Agnello, E., Gaubitz, C., Foster, J., Lapinskas, E., Esposito, E. A., Saneyoshi, T., Kelch, B. A., Garman, S. C., Hayashi, Y., & Stratton, M. M. (2022). CaMKII binds both substrates and activators at the active site. *Cell Reports*, *40*(2), 111064. <https://doi.org/10.1016/j.celrep.2022.111064>
- Palomero, T., Couronné, L., Khiabani, H., Kim, M.-Y., Ambesi-Impiombato, A., Perez-Garcia, A., Carpenter, Z., Abate, F., Allegretta, M., Haydu, J. E., Jiang, X., Lossos, I. S., Nicolas, C., Balbin, M., Bastard, C., Bhagat, G., Piris, M. A., Campo, E., Bernard, O. A., . . . Ferrando, A. A. (2014). Recurrent mutations in epigenetic regulators, RHOA and FYN kinase in peripheral t cell lymphomas. *Nature Genetics*, *46*(2), 166–170. <https://doi.org/10.1038/ng.2873>
- Patrick, R., Lê Cao, K.-A., Kobe, B., & Bodén, M. (2015). PhosphoPICK: Modelling cellular context to map kinase-substrate phosphorylation events. *Bioinformatics*, *31*(3), 382–389. <https://doi.org/10.1093/bioinformatics/btu663>
- Pearce, L. R., Komander, D., & Alessi, D. R. (2010). The nuts and bolts of AGC protein kinases. *Nature Reviews Molecular Cell Biology*, *11*(1), 9–22. <https://doi.org/10.1038/nrm2822>
- Peng, S., & Fu, Y. (2023). FYN: Emerging biological roles and potential therapeutic targets in cancer. *Journal of Translational Medicine*, *21*(1), 84. <https://doi.org/10.1186/s12967-023-03930-0>
- Petersen, B. M., Kirby, M. B., Chrispens, K. M., Irvin, O. M., Strawn, I. K., Haas, C. M., Walker, A. M., Baumer, Z. T., Ulmer, S. A., Ayala, E., Rhodes, E. R., Guthmiller, J. J., Steiner, P. J., & Whitehead, T. A. (2024). An integrated technology for quantitative wide mutational scanning of human antibody fab libraries. *Nature Communications*, *15*(1), 3974. <https://doi.org/10.1038/s41467-024-48072-z>
- Piccinini, G., Bacchiocchi, R., Serresi, M., Vivani, C., Rossetti, S., Gennaretti, C., Carbonari, D., & Fazioli, F. (2002). A ligand-inducible epidermal growth factor receptor/anaplastic lymphoma kinase chimera promotes mitogenesis and transforming properties in 3t3 cells. *Journal of Biological Chemistry*, *277*(25), 22231–22239. <https://doi.org/10.1074/jbc.M111145200>
- Polley, S., Huang, D.-B., Hauenstein, A. V., Fusco, A. J., Zhong, X., Vu, D., Schröfelbauer, B., Kim, Y., Hoffmann, A., Verma, I. M., Ghosh, G., & Huxford, T. (2013). A structural basis for I κ B kinase 2 activation via oligomerization-dependent trans auto-phosphorylation (G. A. Petsko, Ed.). *PLoS Biology*, *11*(6), e1001581. <https://doi.org/10.1371/journal.pbio.1001581>

- Pourcelot, M., Zemirli, N., Silva Da Costa, L., Loyant, R., Garcin, D., Vitour, D., Munitic, I., Vazquez, A., & Arnoult, D. (2016). The golgi apparatus acts as a platform for TBK1 activation after viral RNA sensing. *BMC Biology*, *14*, 69. <https://doi.org/10.1186/s12915-016-0292-z>
- Prywes, N., Phillips, N. R., Oltrogge, L. M., Lindner, S., Taylor-Kearney, L. J., Tsai, Y.-C. C., De Pins, B., Cowan, A. E., Chang, H. A., Wang, R. Z., Hall, L. N., Bellieny-Rabelo, D., Nisonoff, H. M., Weissman, R. F., Flamholz, A. I., Ding, D., Bhatt, A. Y., Mueller-Cajar, O., Shih, P. M., . . . Savage, D. F. (2025). A map of the rubisco biochemical landscape. *Nature*, *638*(8051), 823–828. <https://doi.org/10.1038/s41586-024-08455-0>
- Purvis, J. E., & Lahav, G. (2013). Encoding and decoding cellular information through signaling dynamics. *Cell*, *152*(5), 945–956. <https://doi.org/10.1016/j.cell.2013.02.005>
- Rao, J., Wang, M., Howard, M., Coyote-Maestas, W., & Pimentel, H. (2025, August 1). Cosmos: A position-resolution causal model for direct and indirect effects in protein functions. <https://doi.org/10.1101/2025.08.01.667517>
- Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., & Song, Y. S. (2019). Evaluating protein transfer learning with TAPE. <https://doi.org/10.48550/ARXIV.1906.08230>
- Regot, S., Hughey, J. J., Bajar, B. T., Carrasco, S., & Covert, M. W. (2014). High-sensitivity measurements of multiple kinase activities in live single cells. *Cell*, *157*(7), 1724–1734. <https://doi.org/10.1016/j.cell.2014.04.039>
- Reimand, J., Wagih, O., & Bader, G. D. (2013). The mutational landscape of phosphorylation signaling in cancer. *Scientific Reports*, *3*(1), 2651. <https://doi.org/10.1038/srep02651>
- Reimer, T., Brcic, M., Schweizer, M., & Jungi, T. W. (2008). Poly(i:c) and LPS induce distinct IRF3 and NF- κ b signaling during type-i IFN and TNF responses in human macrophages. *Journal of Leukocyte Biology*, *83*(5), 1249–1257. <https://doi.org/10.1189/jlb.0607412>
- Reynolds, A. B., Kanner, S. B., Bouton, A. H., Schaller, M. D., Weed, S. A., Flynn, D. C., & Parsons, J. T. (2014). SRChing for the substrates of src. *Oncogene*, *33*(37), 4537–4547. <https://doi.org/10.1038/onc.2013.416>
- Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*(15), e2016239118. <https://doi.org/10.1073/pnas.2016239118>
- Rochette, S., Diss, G., Filteau, M., Leducq, J.-B., Dubé, A. K., & Landry, C. R. (2015). Genome-wide protein-protein interaction screening by protein-fragment complementation assay (PCA) in living cells. *Journal of Visualized Experiments*, (97), 52255. <https://doi.org/10.3791/52255>

- Roland Akinseye, O. (2024, December 19). Role of protein kinases in cellular signaling. In Y. Karamanos (Ed.), *Biochemistry* (Vol. 59). IntechOpen. <https://doi.org/10.5772/intechopen.115083>
- Roskoski, R. (2024). Properties of FDA-approved small molecule protein kinase inhibitors: A 2024 update. *Pharmacological Research*, 200, 107059. <https://doi.org/10.1016/j.phrs.2024.107059>
- Santoro, M., Wong, W. T., Aroca, P., Santos, E., Mařoškovā, B., Grieco, M., Fusco, A., & Di Fiore, P. P. (1994). An epidermal growth factor receptor/ *ret* chimera generates mitogenic and transforming signals: Evidence for a ret-specific signaling pathway. *Molecular and Cellular Biology*, 14(1), 663–675. <https://doi.org/10.1128/mcb.14.1.663-675.1994>
- Schmidt, S. H., Weng, J.-H., Aoto, P. C., Boassa, D., Mathea, S., Silletti, S., Torres-Paris, C., Hu, J., Wallbott, M., Komives, E. A., Knapp, S., Herberg, F. W., & Taylor, S. S. (2021). Conformation and dynamics of the kinase domain drive subcellular location and activation of LRRK2. *Proceedings of the National Academy of Sciences*, 118(23), e2100844118. <https://doi.org/10.1073/pnas.2100844118>
- Schmitt, D. L., Curtis, S. D., Lyons, A. C., Zhang, J.-f., Chen, M., He, C. Y., Mehta, S., Shaw, R. J., & Zhang, J. (2022). Spatial regulation of AMPK signaling revealed by a sensitive kinase activity reporter. *Nature Communications*, 13(1), 3856. <https://doi.org/10.1038/s41467-022-31190-x>
- Schubach, M., Maass, T., Nazaretyan, L., Röner, S., & Kircher, M. (2024). CADD v1.7: Using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Research*, 52, D1143–D1154. <https://doi.org/10.1093/nar/gkad989>
- Shah, K., Al-Haidari, A., Sun, J., & Kazi, J. U. (2021). T cell receptor (TCR) signaling in health and disease. *Signal Transduction and Targeted Therapy*, 6(1), 412. <https://doi.org/10.1038/s41392-021-00823-w>
- Shah, N. H., Löbel, M., Weiss, A., & Kuriyan, J. (2018). Fine-tuning of substrate preferences of the src-family kinase lck revealed through a high-throughput specificity screen. *eLife*, 7, e35190. <https://doi.org/10.7554/eLife.35190>
- Shanware, N. P., Hutchinson, J. A., Kim, S. H., Zhan, L., Bowler, M. J., & Tibbetts, R. S. (2011). Casein kinase 1-dependent phosphorylation of familial advanced sleep phase syndrome-associated residues controls PERIOD 2 stability. *Journal of Biological Chemistry*, 286(14), 12766–12774. <https://doi.org/10.1074/jbc.M111.224014>
- Sheng, T., Mao, X.-B., & Zhang, S.-H. (2020). CaMKK β regulates proliferation, apoptosis, and glycolysis of hepatocellular carcinoma via PI3k/AKT pathway. *Annals of Palliative Medicine*, 9(6), 3857–3869. <https://doi.org/10.21037/apm-20-1789>

- Sherry, S. T. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, 29(1), 308–311. <https://doi.org/10.1093/nar/29.1.308>
- Shibasaki, S., Sakata, K., Ishii, J., Kondo, A., & Ueda, M. (2008). Development of a yeast protein fragment complementation assay (PCA) system using dihydrofolate reductase (DHFR) with specific additives. *Applied Microbiology and Biotechnology*, 80(4), 735–743. <https://doi.org/10.1007/s00253-008-1624-x>
- Sokirniy, I., Inam, H., Tomaszewicz, M., Reynolds, J., McCandlish, D., & Pritchard, J. (2025). A side-by-side comparison of variant function measurements using deep mutational scanning and base editing. *Nucleic Acids Research*, 53(14), gkaf738. <https://doi.org/10.1093/nar/gkaf738>
- 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 prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Scientific Reports*, 7(1), 6862. <https://doi.org/10.1038/s41598-017-07199-4>
- Songyang, Z., Blechner, S., Hoagland, N., Hoekstra, M. F., Piwnica-Worms, H., & Cantley, L. C. (1994). Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Current Biology*, 4(11), 973–982. [https://doi.org/10.1016/S0960-9822\(00\)00221-9](https://doi.org/10.1016/S0960-9822(00)00221-9)
- Songyang, Z., Carraway, K. L., Eck, M. J., Harrison, S. C., Feldman, R. A., Mohammadi, M., Schlessinger, J., Hubbard, S. R., Smith, D. P., Eng, C., Lorenzo, M. J., Ponder, B. A. J., Mayer, B. J., & Cantley, L. C. (1995). Catalytic specificity of protein-tyrosine kinases is critical for selective signalling. *Nature*, 373(6514), 536–539. <https://doi.org/10.1038/373536a0>
- Sourisseau, M., Lawrence, D. J. P., Schwarz, M. C., Storrs, C. H., Veit, E. C., Bloom, J. D., & Evans, M. J. (2019). Deep mutational scanning comprehensively maps how Zika envelope protein mutations affect viral growth and antibody escape (J. K. Pfeiffer, Ed.). *Journal of Virology*, 93(23), e01291–19. <https://doi.org/10.1128/JVI.01291-19>
- Spasov, D. S., Ruiz-Saenz, A., Piple, A., & Moasser, M. M. (2018). A dimerization function in the intrinsically disordered n-terminal region of src. *Cell Reports*, 25(2), 449–463.e4. <https://doi.org/10.1016/j.celrep.2018.09.035>
- Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veisler, D., & Bloom, J. D. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, 182(5), 1295–1310.e20. <https://doi.org/10.1016/j.cell.2020.08.012>
- Stephens, M. (2016). False discovery rates: A new deal. *Biostatistics*, kxw041. <https://doi.org/10.1093/biostatistics/kxw041>

- Su, Y.-T., Lu, Y., Liu, J., Chen, M., & Liu, A.-A. (2021). Spatio-temporal mitosis detection in time-lapse phase-contrast microscopy image sequences: A benchmark. *IEEE Transactions on Medical Imaging*, 40(5), 1319–1328. <https://doi.org/10.1109/TMI.2021.3052854>
- Sudol, M., & Hunter, T. (2000). NeW wrinkles for an old domain. *Cell*, 103(7), 1001–1004. [https://doi.org/10.1016/S0092-8674\(00\)00203-8](https://doi.org/10.1016/S0092-8674(00)00203-8)
- Sun, G., Ramdas, L., Wang, W., Vinci, J., McMurray, J., & Budde, R. J. (2002). Effect of autophosphorylation on the catalytic and regulatory properties of protein tyrosine kinase src. *Archives of Biochemistry and Biophysics*, 397(1), 11–17. <https://doi.org/10.1006/abbi.2001.2627>
- Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & von Mering, C. (2023). The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51, D638–D646. <https://doi.org/10.1093/nar/gkac1000>
- Taniguchi, T., Kawamata, T., Mukai, H., Hasegawa, H., Isagawa, T., Yasuda, M., Hashimoto, T., Terashima, A., Nakai, M., Ono, Y., & Tanaka, C. (2001). Phosphorylation of tau is regulated by PKN. *Journal of Biological Chemistry*, 276(13), 10025–10031. <https://doi.org/10.1074/jbc.M007427200>
- Tay, S., Hughey, J. J., Lee, T. K., Lipniacki, T., Quake, S. R., & Covert, M. W. (2010). Single-cell NF- κ b dynamics reveal digital activation and analogue information processing. *Nature*, 466(7303), 267–271. <https://doi.org/10.1038/nature09145>
- Taylor, S. S., & Kornev, A. P. (2011). Protein kinases: Evolution of dynamic regulatory proteins. *Trends in Biochemical Sciences*, 36(2), 65–77. <https://doi.org/10.1016/j.tibs.2010.09.006>
- Tenreiro, S., Eckermann, K., & Outeiro, T. F. (2014). Protein phosphorylation in neurodegeneration: Friend or foe? *Frontiers in Molecular Neuroscience*, 7. <https://doi.org/10.3389/fnmol.2014.00042>
- Thermo Fisher Scientific. (n.d.). Phospho-IKK alpha/beta (ser176, ser180) recombinant superclonal™ antibody (7hcl). https://www.thermofisher.com/order/genome-database/dataSheetPdf?producttype=antibody&productsubtype=antibody_primary&productId=710676&version=Local
- Thomson, D. W., Poeckel, D., Zinn, N., Rau, C., Strohmer, K., Wagner, A. J., Graves, A. P., Perrin, J., Bantscheff, M., Duempelfeld, B., Kasparcova, V., Ramanjulu, J. M., Pesiridis, G. S., Muelbaier, M., & Bergamini, G. (2019). Discovery of GSK8612, a highly selective and potent TBK1 inhibitor. *ACS Medicinal Chemistry Letters*, 10(5), 780–785. <https://doi.org/10.1021/acsmchemlett.9b00027>

- Thyagarajan, B., & Bloom, J. D. (2014). The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife*, 3, e03300. <https://doi.org/10.7554/eLife.03300>
- Tokumitsu, H., & Sakagami, H. (2022). Molecular mechanisms underlying Ca^{2+} /calmodulin-dependent protein kinase kinase signal transduction. *International Journal of Molecular Sciences*, 23(19), 11025. <https://doi.org/10.3390/ijms231911025>
- Tönnies, E., & Trushina, E. (2017). Oxidative stress, synaptic dysfunction, and Alzheimer's disease. *Journal of Alzheimer's Disease*, 57(4), 1105–1121. <https://doi.org/10.3233/JAD-161088>
- Torkamani, A., Verkhivker, G., & Schork, N. J. (2009). Cancer driver mutations in protein kinase genes. *Cancer Letters*, 281(2), 117–127. <https://doi.org/10.1016/j.canlet.2008.11.008>
- Torres Robles, J., Stiegler, A. L., Boggon, T. J., & Turk, B. E. (2025). Cancer hotspot mutations rewire ERK2 specificity by selective exclusion of docking interactions. *Journal of Biological Chemistry*, 301(4), 108348. <https://doi.org/10.1016/j.jbc.2025.108348>
- Tsai, H.-F., Gajda, J., Sloan, T. F. W., Rares, A., & Shen, A. Q. (2019). Usiigaci: Instance-aware cell tracking in stain-free phase contrast microscopy enabled by machine learning. *SoftwareX*, 9, 230–237. <https://doi.org/10.1016/j.softx.2019.02.007>
- Türei, D., Korcsmáros, T., & Saez-Rodriguez, J. (2016). OmniPath: Guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, 13(12), 966–967. <https://doi.org/10.1038/nmeth.4077>
- Turro, E., Greene, D., Wijgaerts, A., Thys, C., Lentaigne, C., Bariana, T. K., Westbury, S. K., Kelly, A. M., Selleslag, D., Stephens, J. C., Papadia, S., Simeoni, I., Penkett, C. J., Ashford, S., Attwood, A., Austin, S., Bakchoul, T., Collins, P., Deevi, S. V. V., . . . Ouwehand, W. H. (2016). A dominant gain-of-function mutation in universal tyrosine kinase *SRC* causes thrombocytopenia, myelofibrosis, bleeding, and bone pathologies. *Science Translational Medicine*, 8(328). <https://doi.org/10.1126/scitranslmed.aad7666>
- Ubersax, J. A., & Ferrell Jr, J. E. (2007). Mechanisms of specificity in protein phosphorylation. *Nature Reviews Molecular Cell Biology*, 8(7), 530–541. <https://doi.org/10.1038/nrm2203>
- Ubersax, J. A., Woodbury, E. L., Quang, P. N., Paraz, M., Blethrow, J. D., Shah, K., Shokat, K. M., & Morgan, D. O. (2003). Targets of the cyclin-dependent kinase cdk1. *Nature*, 425(6960), 859–864. <https://doi.org/10.1038/nature02062>
- Ulicna, K., Vallardi, G., Charras, G., & Lowe, A. R. (2021). Automated deep lineage tree analysis using a bayesian single cell tracking approach. *Frontiers in Computer Science*, 3, 734559. <https://doi.org/10.3389/fcomp.2021.734559>

- Ulman, V., Maška, M., Magnusson, K. E. G., Ronneberger, O., Haubold, C., Harder, N., Matula, P., Svoboda, D., Radojevic, M., Smal, I., Rohr, K., Jaldén, J., Blau, H. M., Dzyubachyk, O., Lelieveldt, B., Xiao, P., Li, Y., Cho, S.-Y., . . . Ortiz-de-Solorzano, C. (2017). An objective comparison of cell-tracking algorithms. *Nature Methods*, *14*(12), 1141–1152. <https://doi.org/10.1038/nmeth.4473>
- Um, J. W., Nygaard, H. B., Heiss, J. K., Kostylev, M. A., Stagi, M., Vortmeyer, A., Wisniewski, T., Gunther, E. C., & Strittmatter, S. M. (2012). Alzheimer amyloid- β oligomer bound to postsynaptic prion protein activates fyn to impair neurons. *Nature Neuroscience*, *15*(9), 1227–1235. <https://doi.org/10.1038/nn.3178>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. <https://doi.org/10.48550/ARXIV.1706.03762>
- Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, *339*(6127), 1546–1558. <https://doi.org/10.1126/science.1235122>
- Voliotis, M., Perrett, R. M., McWilliams, C., McArdle, C. A., & Bowsher, C. G. (2014). Information transfer by leaky, heterogeneous, protein kinase signaling systems. *Proceedings of the National Academy of Sciences*, *111*(3). <https://doi.org/10.1073/pnas.1314446111>
- Walker, C., Wang, Y., Olivieri, C., Karamafrooz, A., Casby, J., Bathon, K., Calebiro, D., Gao, J., Bernlohr, D. A., Taylor, S. S., & Veglia, G. (2019). Cushing's syndrome driver mutation disrupts protein kinase a allosteric network, altering both regulation and substrate specificity. *Science Advances*, *5*(8), eaaw9298. <https://doi.org/10.1126/sciadv.aaw9298>
- Wang, B., Wang, M., Zhang, W., Xiao, T., Chen, C.-H., Wu, A., Wu, F., Traugh, N., Wang, X., Li, Z., Mei, S., Cui, Y., Shi, S., Lipp, J. J., Hinterndorfer, M., Zuber, J., Brown, M., Li, W., & Liu, X. S. (2019). Integrative analysis of pooled CRISPR genetic screens using MAGeCKFlute. *Nature Protocols*, *14*(3), 756–780. <https://doi.org/10.1038/s41596-018-0113-7>
- Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., & Xu, D. (2017). MusiteDeep: A deep-learning framework for general and kinase-specific phosphorylation site prediction (J. Hancock, Ed.). *Bioinformatics*, *33*(24), 3909–3916. <https://doi.org/10.1093/bioinformatics/btx496>
- Wang, L., Aschenbrenner, D., Zeng, Z., Cao, X., Mayr, D., Mehta, M., Capitani, M., Warner, N., Pan, J., Wang, L., Li, Q., Zuo, T., Cohen-Kedar, S., Lu, J., Ardy, R. C., Mulder, D. J., Dissanayake, D., Peng, K., Huang, Z., . . . Muise, A. M. (2021). Gain-of-function variants in SYK cause immune dysregulation and systemic inflammation in humans and mice. *Nature Genetics*, *53*(4), 500–510. <https://doi.org/10.1038/s41588-021-00803-4>

- Wang, M., Wang, T., Wang, B., Liu, Y., & Li, A. (2017). A novel phosphorylation site-kinase network-based method for the accurate prediction of kinase-substrate relationships. *BioMed Research International*, 2017, 1–11. <https://doi.org/10.1155/2017/1826496>
- Wang, Y., Zhang, X., Yang, T., & Sun, J. (2022). Anchor DETR: Query design for transformer-based detector. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3), 2567–2575. <https://doi.org/10.1609/aaai.v36i3.20158>
- Wayman, G. A., Lee, Y.-S., Tokumitsu, H., Silva, A., & Soderling, T. R. (2008). Calmodulin-kinases: Modulators of neuronal development and plasticity. *Neuron*, 59(6), 914–931. <https://doi.org/10.1016/j.neuron.2008.08.021>
- Wei, H., & Li, X. (2023). Deep mutational scanning: A versatile tool in systematically mapping genotypes to phenotypes. *Frontiers in Genetics*, 14, 1087267. <https://doi.org/10.3389/fgene.2023.1087267>
- Weng, C., Faure, A. J., Escobedo, A., & Lehner, B. (2024). The energetic and allosteric landscape for KRAS inhibition. *Nature*, 626(7999), 643–652. <https://doi.org/10.1038/s41586-023-06954-0>
- White, R., Gonsior, C., Krämer-Albers, E.-M., Stöhr, N., Hüttelmaier, S., & Trotter, J. (2008). Activation of oligodendroglial fyn kinase enhances translation of mRNAs transported in hnRNP a2-dependent RNA granules. *The Journal of Cell Biology*, 181(4), 579–586. <https://doi.org/10.1083/jcb.200706164>
- Wishart, D. S. (2006). DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(90001), D668–D672. <https://doi.org/10.1093/nar/gkj067>
- Wu, J., Yang, J., Kannan, N., Madhusudan, Xuong, N.-H., Ten Eyck, L. F., & Taylor, S. S. (2005). Crystal structure of the e230q mutant of cAMP-dependent protein kinase reveals an unexpected apoenzyme conformation and an extended n-terminal α helix. *Protein Science*, 14(11), 2871–2879. <https://doi.org/10.1110/ps.051715205>
- Wu, L., Ma, C. A., Zhao, Y., & Jain, A. (2011). Aurora b interacts with NIR-p53, leading to p53 phosphorylation in its DNA-binding domain and subsequent functional suppression. *Journal of Biological Chemistry*, 286(3), 2236–2244. <https://doi.org/10.1074/jbc.M110.174755>
- Wu, L., Brzostek, J., Sakthi Vale, P. D., Wei, Q., Koh, C. K., Ong, J. X. H., Wu, L.-z., Tan, J. C., Chua, Y. L., Yap, J., Song, Y., Tan, V. J. Y., Tan, T. Y., Lai, J., MacAry, P. A., & Gascoigne, N. R. (2023). CD28-CAR-t cell activation through FYN kinase signaling rather than LCK enhances therapeutic performance. *Cell Reports Medicine*, 4(2), 100917. <https://doi.org/10.1016/j.xcrm.2023.100917>

- Xu, M., Lee, E. M., Wen, Z., Cheng, Y., Huang, W.-K., Qian, X., Tcw, J., Kouznetsova, J., Ogden, S. C., Hammack, C., Jacob, F., Nguyen, H. N., Itkin, M., Hanna, C., Shinn, P., Allen, C., Michael, S. G., Simeonov, A., Huang, W., . . . Tang, H. (2016). Identification of small-molecule inhibitors of zika virus infection and induced neural cell death via a drug repurposing screen. *Nature Medicine*, 22(10), 1101–1107. <https://doi.org/10.1038/nm.4184>
- Xu, X., Chen, Y., Fu, Q., Ni, D., Zhang, J., Li, X., & Lu, S. (2019). The chemical diversity and structure-based discovery of allosteric modulators for the PIF-pocket of protein kinase PDK1. *Journal of Enzyme Inhibition and Medicinal Chemistry*, 34(1), 361–374. <https://doi.org/10.1080/14756366.2018.1553167>
- Xue, L., & Tao, W. A. (2013). Current technologies to identify protein kinase substrates in high throughput. *Frontiers in Biology*, 8(2), 216–227. <https://doi.org/10.1007/s11515-013-1257-z>
- Yaffe, M. B., Rittinger, K., Volinia, S., Caron, P. R., Aitken, A., Leffers, H., Gamblin, S. J., Smerdon, S. J., & Cantley, L. C. (1997). The structural basis for 14-3-3:phosphopeptide binding specificity. *Cell*, 91(7), 961–971. [https://doi.org/10.1016/S0092-8674\(00\)80487-0](https://doi.org/10.1016/S0092-8674(00)80487-0)
- Yang, S.-S., Fang, Y.-W., Tseng, M.-H., Chu, P.-Y., Yu, I.-S., Wu, H.-C., Lin, S.-W., Chau, T., Uchida, S., Sasaki, S., Lin, Y.-F., Sytwu, H.-K., & Lin, S.-H. (2013). Phosphorylation regulates NCC stability and transporter activity in vivo. *Journal of the American Society of Nephrology*, 24(10), 1587–1597. <https://doi.org/10.1681/ASN.2012070742>
- Yaron, T. M., Heaton, B. E., Levy, T. M., Johnson, J. L., Jordan, T. X., Cohen, B. M., Kerelsky, A., Lin, T.-Y., Liberatore, K. M., Bulaon, D. K., Van Nest, S. J., Koundouros, N., Kastenhuber, E. R., Mercadante, M. N., Shobana-Ganesh, K., He, L., Schwartz, R. E., Chen, S., Weinstein, H., . . . Heaton, N. S. (2022). Host protein kinases required for SARS-CoV-2 nucleocapsid phosphorylation and viral replication. *Science Signaling*, 15(757), eabm0808. <https://doi.org/10.1126/scisignal.abm0808>
- Yaron-Barir, T. M., Joughin, B. A., Huntsman, E. M., Kerelsky, A., Cizin, D. M., Cohen, B. M., Regev, A., Song, J., Vasan, N., Lin, T.-Y., Orozco, J. M., Schoenherr, C., Sagum, C., Bedford, M. T., Wynn, R. M., Tso, S.-C., Chuang, D. T., Li, L., Li, S. S.-C., . . . Johnson, J. L. (2024). The intrinsic substrate specificity of the human tyrosine kinome. *Nature*, 629(8014), 1174–1181. <https://doi.org/10.1038/s41586-024-07407-y>
- Yoon, W., Hwang, S.-H., Lee, S.-H., & Chung, J. (2019). Drosophila ADCK1 is critical for maintaining mitochondrial structures and functions in the muscle (B. Lu, Ed.). *PLOS Genetics*, 15(5), e1008184. <https://doi.org/10.1371/journal.pgen.1008184>

- Zargari, A., Lodewijk, G. A., Mashhadi, N., Cook, N., Neudorf, C. W., Aragh-bidikashani, K., Hays, R., Kozuki, S., Rubio, S., Hrabeta-Robinson, E., Brooks, A., Hinck, L., & Shariati, S. A. (2023). DeepSea is an efficient deep-learning model for single-cell segmentation and tracking in time-lapse microscopy. *Cell Reports Methods*, 3(6), 100500. <https://doi.org/10.1016/j.crmeth.2023.100500>
- Zhang, J., Ma, Y., Taylor, S. S., & Tsien, R. Y. (2001). Genetically encoded reporters of protein kinase a activity reveal impact of substrate tethering. *Proceedings of the National Academy of Sciences*, 98(26), 14997–15002. <https://doi.org/10.1073/pnas.211566798>
- Zhang, J.-F., Liu, B., Hong, I., Mo, A., Roth, R. H., Tenner, B., Lin, W., Zhang, J. Z., Molina, R. S., Drobizhev, M., Hughes, T. E., Tian, L., Haganir, R. L., Mehta, S., & Zhang, J. (2021). An ultrasensitive biosensor for high-resolution kinase activity imaging in awake mice. *Nature Chemical Biology*, 17(1), 39–46. <https://doi.org/10.1038/s41589-020-00660-y>
- Zhang, J.-f., Mehta, S., & Zhang, J. (2022). Protocol for reading and imaging live-cell PKA activity using ExRai-AKAR2. *STAR Protocols*, 3(1), 101071. <https://doi.org/10.1016/j.xpro.2021.101071>
- Zhang, Q., Huang, H., Zhang, L., Wu, R., Chung, C.-I., Zhang, S.-Q., Torra, J., Schepis, A., Coughlin, S. R., Kornberg, T. B., & Shu, X. (2018). Visualizing dynamics of cell signaling in vivo with a phase separation-based kinase reporter. *Molecular Cell*, 69(2), 334–346.e4. <https://doi.org/10.1016/j.molcel.2017.12.008>
- Zhao, Q., Zheng, K., Ma, C., Li, J., Zhuo, L., Huang, W., Chen, T., & Jiang, Y. (2020). PTPS facilitates compartmentalized LTBP1 s-nitrosylation and promotes tumor growth under hypoxia. *Molecular Cell*, 77(1), 95–107.e5. <https://doi.org/10.1016/j.molcel.2019.09.018>
- Zhou, C., Qian, X., Hu, M., Zhang, R., Liu, N., Huang, Y., Yang, J., Zhang, J., Bai, H., Yang, Y., Wang, Y., Ali, D., Michalak, M., Chen, X.-Z., & Tang, J. (2020). STYK1 promotes autophagy through enhancing the assembly of autophagy-specific class III phosphatidylinositol 3-kinase complex i. *Autophagy*, 16(10), 1786–1806. <https://doi.org/10.1080/15548627.2019.1687212>
- Zhou, Y., Pérez, B., Hao, W., Lv, J., Gao, R., & Guo, Z. (2019). The additive mutational effects from surface charge engineering: A compromise between enzyme activity, thermostability and ionic liquid tolerance. *Biochemical Engineering Journal*, 148, 195–204. <https://doi.org/10.1016/j.bej.2018.07.020>
- Zhou, Z., Yeung, W., Gravel, N., Salcedo, M., Soleymani, S., Li, S., & Kannan, N. (2023). Phosformer: An explainable transformer model for protein kinase-

- specific phosphorylation predictions (P. L. Martelli, Ed.). *Bioinformatics*, 39(2), btad046. <https://doi.org/10.1093/bioinformatics/btad046>
- Zhuo, B.-b., Zhu, L.-q., Yao, C., Wang, X.-h., Li, S.-x., Wang, R., Li, Y., & Ling, Z.-y. (2022). ADCK1 is a potential therapeutic target of osteosarcoma. *Cell Death & Disease*, 13(11), 954. <https://doi.org/10.1038/s41419-022-05401-8>
- Zou, L., Wang, M., Shen, Y., Liao, J., Li, A., & Wang, M. (2013). PKIS: Computational identification of protein kinases for experimentally discovered protein phosphorylation sites. *BMC Bioinformatics*, 14(1), 247. <https://doi.org/10.1186/1471-2105-14-247>