ROBUST, FAIR AND ACCESSIBLE: ALGORITHMS FOR ENCHANCING PROTEOMICS AND UNDER-STUDIED PROTEINS IN NETWORK BIOLOGY

SERHAN YILMAZ

Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Thesis Advisor: Mehmet Koyutürk

Department of Computer and Data Sciences

Case Western Reserve University

August, 2023

Robust, Fair and Accessible: Algorithms for enhancing proteomics and under-studied proteins in network biology

Case Western Reserve University
Case School of Graduate Studies

We hereby approve the thesis[*] of

Serhan Yılmaz

for the degree of

Doctor of Philosophy

Dr. Mehmet Koyutürk

Committee Chair, Advisor — June 28, 2023
Department of Computer and Data Sciences

Dr. Vincenzo Liberatore

Committee Member — June 28, 2023
Department of Computer and Data Sciences

Dr. Michael Lewicki

Committee Member — June 28, 2023
Department of Computer and Data Sciences

Dr. Kevin Xu

Committee Member — June 28, 2023
Department of Computer and Data Sciences

Dr. Mark Chance

Committee Member — June 28, 2023
Department of Nutrition, School of Medicine

[*] We certify that written approval has been obtained for any proprietary material contained therein.

Dedicated to those who need it
and those who are in need

Table of Contents

List of Figures

Acknowledgements

I would like to express my deep gratitude and appreciation to the individuals who have played a pivotal role in the completion of my thesis and provided invaluable support throughout my academic journey.

First and foremost, I am immensely grateful to my advisor, Mehmet Koyutürk, for his exceptional guidance, invaluable insights, and continuous encouragement. Throughout my academic journey, he has consistently shown kindness, understanding, and a genuine commitment to my success. His expertise, mentorship, and unwavering support have significantly shaped the direction and quality of my research. I consider myself incredibly lucky to have had the opportunity to work with such an outstanding advisor. I would also like to extend my appreciation to Günnur Karakurt, who has been a significant influence in our collaborative projects and has provided invaluable support throughout. Their contributions have been instrumental in my academic journey, and I am truly grateful for their guidance and presence.

I would also like to express my sincere appreciation to Mark Chance for his role as the principal investigator in the Alzheimer's research project. His expertise, guidance, and support have significantly influenced my understanding of the field and its connection to my thesis. Our regular R01 research meetings have provided valuable insights and enriched my perspective.

Furthermore, I would like to express my gratitude to other esteemed committee members, Vincenzo Liberatore, Kevin Xu, and Michael Lewicki. Their guidance, expertise, and critical feedback have been instrumental in shaping the trajectory of my research and ensuring the rigor of my work. I would also like to mention Vincenzo Liberatore specifically for his continuous support and involvement since the early stages of my academic journey, from the first semester course to the qualification exam, graduate student award ceremony, thesis proposal, and finally to the thesis defense.

I would like to give special recognition to Ahmet, a dear friend and research collaborator. Our close collaboration and mutual support, particularly during my initial years in Cleveland, have been instrumental in overcoming challenges and achieving significant milestones. Ahmet's commitment, dedication, and collaborative spirit have profoundly influenced the outcomes of our research endeavors. I am deeply grateful for his contributions and the positive impact he has had on me, both academically and personally. The support and understanding he has shown during challenging times have been truly remarkable.

I am also grateful to Kaan, with whom I have worked closely on a research project that became an integral chapter of my thesis and interacted frequently. His insightful perspectives, critical thinking, and collaborative efforts significantly contributed to the research findings and enriched the overall quality of my work.

To Filipa, Danie, and Marzieh, I want to express my profound appreciation for their involvement and collaboration in phospho-proteomic research. Filipa's close collaboration, valuable feedback, and support, particularly regarding the RokaiXplorer project, have significantly influenced my research and the final chapter of my thesis. I would also like to acknowledge Marzieh for her contributions to the co-phosphorylation project, which became a separate chapter in my thesis. Their dedication and expertise have greatly enriched the quality and depth of my work.

I would also like to thank Ercüment Çiçek and Öznur Taştan for their supervision and support on a research project undertaken during my time here. Working with Mohamad, a former undergraduate student from their university, was an exceptionally fruitful and productive experience. I am immensely grateful for Mohamad's dedication and invaluable contributions, which not only led to a published paper but also a cherished friendship.

Çerağ, Mengzhen, Julian, Alex, Morgan, Sean, Tyler, and Zach, current and former members of our lab, I am thankful for their collaborative efforts, shared knowledge, and support. Their contributions have created a stimulating research environment and helped enrich my understanding of the subject matter.

Leo, Ayush, Can, Akif, Çağlar, and Orhan, former fellow graduate students and friends, I want to express my appreciation for their friendship, support, and meaningful interactions. Our discussions and shared experiences have provided valuable inspiration and encouragement throughout my academic journey. I am particularly grateful to Leo and Ayush for our close and productive discussions, and Orhan for his invaluable assistance during my initial days in Cleveland, helping me navigate and settle into a new environment.

I would also like to extend my gratitude to Baran, Doğa, Altan, Denver, Görkem, Furkan, Alper, Efe, and other friends who supported me during this journey and blessed me with encouragement and moments of respite. Your presence and friendship made this academic journey more meaningful and enjoyable. Thank you.

Before I finish, I would also like to offer my thanks to the funding organizations who have generously supported my research. R01-LM012980 from National Library Of Medicine, their financial assistance has been crucial in enabling the successful completion of this work.

Last but not least, to my family, whose unwavering love, encouragement, and understanding have been my foundation throughout my academic pursuits, I am forever grateful. My mother Sevim Yılmaz, my father Hilmi Yılmaz, and my sister Hülya Yılmaz, their belief in my abilities and their continuous support have given me the strength to overcome obstacles and achieve my goals.

Thank you all for believing in me and for being an integral part of my success.

Abstract

Robust, Fair and Accessible: Algorithms for Enhancing Proteomics and Under-Studied Proteins in Network Biology

SERHAN YILMAZ

This dissertation presents a comprehensive approach to advancing proteomics and under-studied proteins in network biology, emphasizing the development of reliable algorithms, fair evaluation practices, and accessible computational tools. A key contribution of this work is the introduction of RoKAI, a novel algorithm that integrates multiple sources of functional information to infer kinase activity. By capturing coordinated changes in signaling pathways, RoKAI significantly improves kinase activity inference, facilitating the identification of dysregulated kinases in diseases. This enables deeper insights into cellular signaling networks, supporting targeted therapy development and expanding our understanding of disease mechanisms. To ensure fairness in algorithm evaluation, this research carefully examines potential biases arising from the under-representation of under-studied proteins and proposes strategies to mitigate these biases, promoting a more comprehensive evaluation and encouraging the discovery of novel findings. Additionally, this dissertation focuses on enhancing accessibility by developing user-friendly computational tools. The RoKAI web application provides a convenient and intuitive interface to perform RoKAI analysis. Moreover, RokaiXplorer web tool simplifies proteomic and phospho-proteomic data analysis for researchers without specialized expertise. It enables tasks such as normalization, statistical testing, pathway enrichment, provides interactive visualizations, while also offering researchers the ability to deploy their own data browsers, promoting the sharing of findings and fostering collaborations. Overall, this interdisciplinary research contributes to proteomics and network biology by providing robust algorithms, fair evaluation practices, and accessible tools. It lays the foundation for further advancements in the field, bringing us closer to uncovering new biomarkers and potential therapeutic targets in diseases like cancer, Alzheimer's, and Parkinson's.

1 Introduction

1.1 Goals and focus of this work

This work focuses on the development of algorithms and computational tools various inference and predictive modeling tasks in the context of cellular signaling with the ultimate goal of expanding the existing knowledge-base on proteomics, phospho-proteomics and network biology. Throughout this work, the following aspects are given a notable emphasis in the design of the developed methods:

- **Robust**: Building reliable models that are resistant to missing data and gaps in collective human knowledge. Chapter 2 and 3 has a particular focus on this aspect.
- **Fair**: Ensuring fairness in the validation of the algorithms, uncovering any potential biases, and making sure under-studied proteins are not under-represented in the evaluation. This is the main focus in Chapter 4.
- **Accessible:** Making the analyses and the developed algorithms accessible, easily understandable and applicable, facilitating the adoption of new techniques. For this purpose, several web tools[†,‡,§] are developed in this work. Particularly, Chapter 5 focus on this aspect.

[†] https://rokai.io

[‡] https://yilmazs.shinyapps.io/colipe

[§] http://explorer.rokai.io

1.2 Organization of the thesis

The organization of the proposed thesis is as follows:

- Chapter 1 (this chapter): Describes the motivation and significance behind the work, and summarizes the contents of each chapter.
- Chapter 2: Describes an algorithm named RoKAI for performing robust inference of kinase activity using functional networks. This work is published in Nature Communications[145].
- Chapter 3: Describes the co-phosphorylation (Co-P for short) networks as an additional resource that can be helpful in the context of proteomics for the prediction functional networks such as kinase-substrate interactions and pathway memberships. This work is published in Bioinformatics[6].
- Chapter 4 (titled *"Are under-studied proteins under-studied? How to fairly evaluate link prediction algorithms"*) focuses on link prediction tasks in network biology. It demonstrates a bias toward "rich", well-studied proteins in the commonly used evaluation settings of link prediction algorithms and describes strategies to prevent the under-studied proteins from being under-represented in the evaluation.
- Chapter 5 (titled *"Making Proteomics Accessible: RokaiXplorer for interactive analysis of phospho-proteomic data."*). This chapter presents RokaiXplorer, an interactive web tool aimed at addressing the lack of user-friendly solutions for analyzing and visualizing proteomics and phospho-proteomics data. RokaiXplorer's objective is to enhance accessibility to statistical analyses in this field by providing a streamlined processing and visualization platform. It offers a comprehensive set of modules for data normalization, statistical testing, enrichment analysis, kinase activity inference, and interactive visualizations, catering to researchers without specialized expertise.

1.3 Background and Significance

This section serves as a brief introduction to the key concepts that are necessary to understand the cellular signaling systems and phospho-proteomics.

1.3.1 Post-translational modifications

Post-translational modification (PTMs) are one of the prominent mechanisms regulating the cell activity through dynamic coordination of the signaling networks[134]. A PTM occurs when an aminoacid on a protein is modified through a (typically enzymatic) chemical reaction after the protein is translated by the ribosome[15]. So far, hundreds of different PTM types are identified[9]. Among all known PTM types, some notable ones include: phosphorylation, glycosylation, ubiquitination, methylation and acetylation.

1.3.2 Phosphorylation

Phosphorylation is one of the most studied PTMs with the most available resources due to (i) its prevalence in the eukaryotic signaling system, (ii) availability of effective biochemical enrichment techniques, (iii) and the historical aspect of being one of the first identified PTMs[9]. There are several key functions of protein phosphorylation including glycolysis, protein degradation, regulating protein-protein interactions, and enzyme inhibition/activation[64].

Phosphorylation occurs with the addition of a phosphoryl group (PO_3^-) to an amino acid (Figure 1.1a). The amino acid position in which the phosphorylation occurs is called a phosphorylation site or *phosphosite* for short. Like most PTMs, the protein phosphorylation requires the presence of some specialized enzymes called *kinases* catalyzing the chemical reaction. Note that, these kinases themselves are also proteins and they can contain phosphosites which may alter the activity of the kinase. Whereas, the amino acid on a phosphosite that a kinase phosphorylates is commonly referred as a *substrate*. Although there are technical differences between a substrate and a phosphosite, we generally use them interchangeably in this study. For clarity, we refer a phosphosite as a substrate only in the context of a kinase.

Note that, a kinase does not act on all phosphosites on the phospho-proteome. There are hundreds of known kinase classes, each phosphorylating only a small subset of substrates. Moreover, our knowledge of human phospho-proteome is still far from complete[97], even though there are thousands of known kinase-substrate interactions[47]. Thus, to be able to fully understand the signaling pathways regulated though phosphorylation, it is imperative to first characterize the kinases and the substrates they act upon.

1.3.3 Mass spectrometry

Mass spectrometry (MS) is an analytical technique that allows the large-scale identification and quantification of phosphorylation profiles of cells and tissues[53]. MS offers a rapid way of obtaining rich phosphorylation data, typically with thousands of identified sites[28]. Moreover, in addition to measuring phosphorylation levels, mass spectrometry can also be used to measure the protein abundance[150].

1.3.4 MS-based phospho-proteomics data

Raw Counts: The raw (single sample) MS-data contains a count for each identified phosphosite indicating the abundance of a mono-phosphorylated peptide (Figure 1.1b). Thus, a higher count indicates a greater evidence of phosphorylation. However, these values are not normalized and there is no baseline level to compare to. Therefore, with the raw counts, it is hard to interpret what constitutes a sufficient evidence for a site to be considered phosphorylated.

Fold Changes: To obtain a baseline, the phosphorylation levels of a case sample (e.g., from a cancer patient) are typically compared to a control sample (e.g., from a healthy patient). The normalized phosphorylation profile of MS-data (Figure 1.1c) is typically provided as fold change of the phosphorylated peptide abundances between case and control samples. For example, a fold change of 2 indicates two times as high abundance in the case sample compared to the control sample.

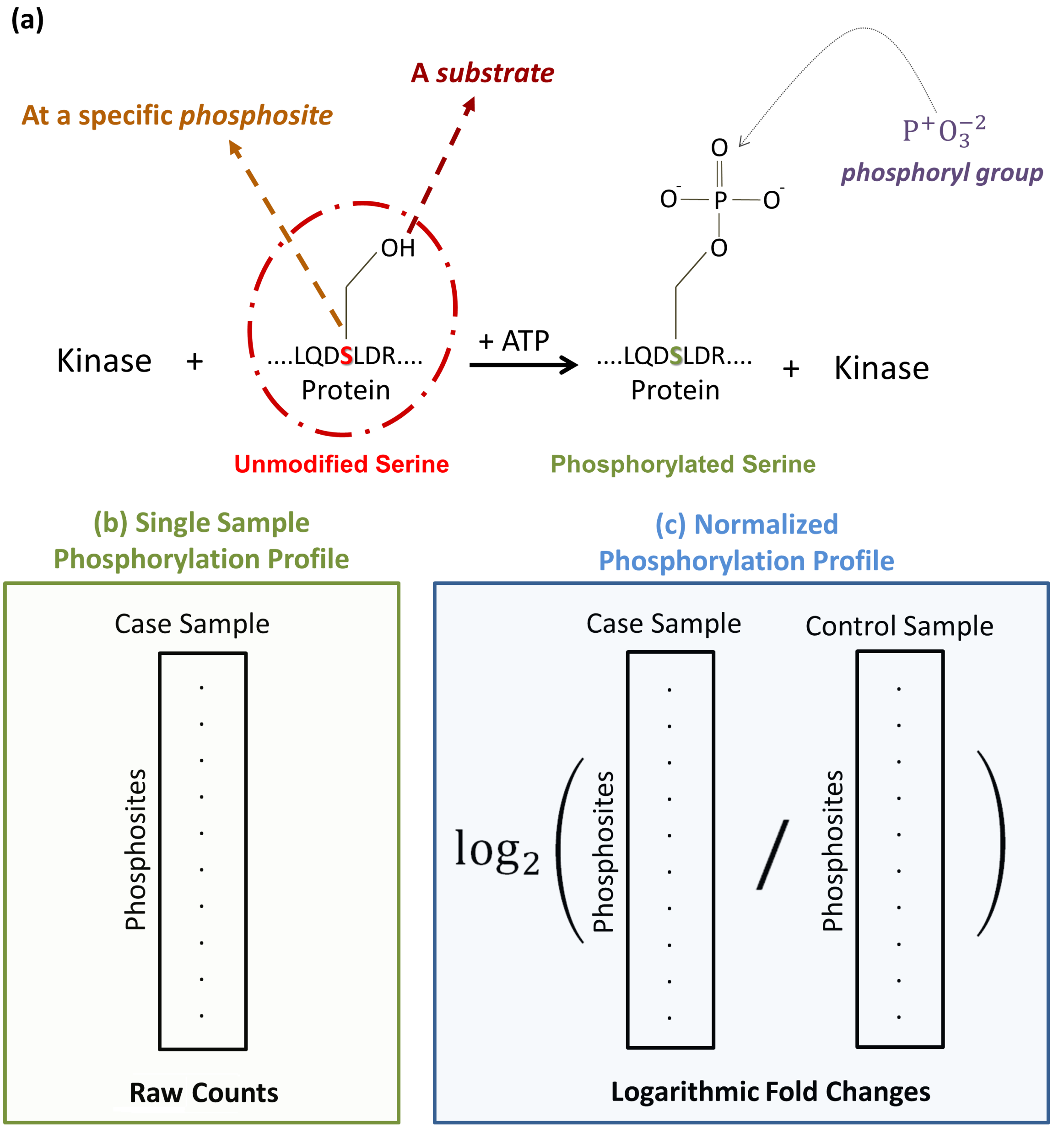


Figure 1.1. (a) Illustrating the phosphorylation of serine amino acid. (b) Single sample phosphosphorylation profile, containing raw counts (in the range $[0, \infty]$) obtained from mass spectometry. (c) Normalized phosphorylation profile for a case sample of interest. The normalized phosphorylation is typically provided as logarithmic fold changes (in the range $[-\infty, \infty]$) relative to a control sample.

1.3.5 Available Resources in the context of Proteomics

This section describes the main datasets and resources that are used in the context of proteomics and phospho-proteomic data analysis.

Kinases and Kinase-Substrate Interactions. The up-to-date information about the known kinases can be obtained from KinBase[80]. Whereas, there are several databases containing annotated and/or predicted kinase-substrate interactions such as PhosphoSitePlus (PSP)[47], Phospho.ELM[32], PhosphoPoint[141], HPRD[106] and Signor[105]. Some of these databases (like PSP) are quite stringent in the interactions they include i.e., they include only annotated interactions that are experimentally validated. Whereas, some (like Phospho.ELM) also include predicted interactions. The predicted interactions are typically obtained by using NetworkKin algorithm[46,78]. The NetworkKin algorithm predicts new kinase-substrates based on motif similarity and proximity on protein-protein interaction (PPI) network. The intuition here is that if two sites are functionally and/or structurally similar, then a kinase phosphorylating one of them may also phosphorylate the other.

Protein-centric functional information. There are several datasets that provide functional information on the protein level: STRING[128] provides protein-protein interaction (PPI) networks indicating proteins that physically interact. Whereas, KEGG[56,57], Reactome[25], Netpath[55] and MSigDB[75] databases provide information about pathways. That is, these pathway databases describe causal links between proteins (or other functional units) that are typically experimentally validated.

Phosphosite-centric functional information. There are a few datasets providing functional information on the phosphosite level: PTMsigDB[65] provides information about sites that are on the same pathway, PTMcode[90] provides information about co-evolved sites, and Phomics[94] provides information about sites that are on an activation loop of a kinase. The sites of the activation loop of a kinase are known to activate that kinase (i.e., cause the kinase to phosphorylate its substrates).

1.3.6 Kinase Activity Inference Problem

Here, the main goal is to infer kinases that are activated or deactivated by a condition of interest (this may be a perturbation like a drug or it could be a disease phenotype like cancer). More specifically, the ultimate goal is to uncover *causal links* between the condition and kinases. However, this is difficult (if not impossible) to achieve computationally since the available MS-data is observational in nature. Thus, a more feasible goal is to uncover the correlations between kinases and the condition of interest. These correlations can then be used to generate informed hypotheses (e.g., by ranking kinases) that may be validated by performing controlled *a posteriori* experiments (e.g., randomized trials). Chapter 2 focuses on this problem and proposes a new algorithm named RoKAI to infer the kinase activities in a robust manner.

1.3.7 Challenges in proteomic and phospho-proteomic data analysis

The main challenges of proteomic and phospho-proteomic data analysis (e.g., in kinase activity inference) can be summarized in three points: (i) The available data (mainly MS-based phosphorylation levels) are noisy and incomplete, (ii) Even though there is data available from multiple sources, it is often unclear how to combine them appropriately, and (iii) The available phosphorylation data is observational in nature which makes it challenging to answer causal questions.

Incomplete data and missingness. In a typical MS run, only around 50% of all known phosphosites are identified (i.e., the phosphorylation levels are obtained). Also, knowledge about kinase-substrate interactions are limited: Barely around 20% of all phosphosites are annotated to be a substrate of a specific kinase[94]. Moreover, these annotations are not distributed uniformly among the kinases i.e., there are a few kinases with many known substrates and many kinases with only a few known substrates. This makes it challenging to infer kinase activity especially for kinases with only a few identified substrates (or without any identified substrates) in the MS run. Moreover, it is not straightforward how to perform this in a fair/unbiased manner, giving equal chance to all kinases (and, for example, prevent the results from being dominated with a few, highly studied kinases regardless of the condition of interest).

Data integration. Even though there are several types of data available from multiple sources, it is often unclear how to incorporate them appropriately. Moreover, the protein reference databases like UniProt[22], NCBI RefSeq[103] and Ensemble[149] all use distinct accession number schemes and there is typically no one-to-one mapping between them, which makes it difficult to combine data from different sources.

1.3.8 Use of machine learning in network biology

As in other applications of machine learning, biological applications face important challenges regarding fairness, bias-awareness, interpretability, generalizibility, and accessibility. However, as compared to other popular applications, biology is unique in terms of what these considerations mean in practice. Besides clinical/translational applications, an important part of predictive tasks in network biology focus on discovery of novel biological knowledge. The knowledge generated can range from functional annotation to identification of regulatory interactions, discovery of cell types, characterization of interactions between biomolecules, and so on.

Network models and graph machine learning in biomedical applications. Network biology has been an important pillar of computational biology research in the last two decades[8]. Network models have been commonly used to describe biological processes and pathways, and represent interactions among biomolecules as well as higher-level associations among biological entities. In the context of predictive tasks using omic datasets, biological networks serve as templates that represent the functional relationships between features, variables, or samples[24]. Machine learning algorithms utilize these functional relationships to integrate data from different modalities, reduce dimensionality, extract and select features, identify latent patterns, and fill missing data gaps[14]. As graph machine learning advanced in the last decade, sophisticated techniques involving node, edge, and graph embeddings[98], graph neural networks[153], and graph convolution[59] have been increasingly applied to predictive tasks in systems biology.

The disconnect between the development vs. the utilization of algorithms. Despite the explosion in algorithm development efforts, there is ongoing debate on the extent of these algorithms' contribution to advancing biology. In their review of the application of machine learning in biology, David Jones[54] writes: "*The problem we face is that many of the papers that we are now seeing as a result of the AI revolution are not advancing the field, because these techniques are not appropriately used: either they do not offer any improvement in comparison with existing methods or their experimental design is flawed; often both.*"

Trustworthy machine learning and the importance of validation. An underlying reason for this distrust in the methods is validation. Computational validation techniques use established benchmark data and assess the accuracy of predictions with respect to these benchmarks. However, there is a lack of a gold standard and these benchmarking data themselves can be biased toward well-studied biological entities. Thus, computational validation efforts, if not done carefully, would only assess the ability of algorithms and predictive models in rediscovering what is already known: saying little, if anything, about whether the predictions can lead to novel discoveries. An overarching goal of this thesis (and particularly the Chapter 4) is to develop strategies to assess a graph machine learning algorithm's potential *to drive new discoveries in new biology*, filling the gaps in the dark side of proteome that is left under-studied[97].

1.3.9 Under-studied proteins and fairness in knowledge discovery

In the context of biological knowledge discovery, an important aspect of *fairness* is the ability to identify biological entities that are relatively less studied (e.g., when a scientist is looking to identify a kinase that phosphorylates a specific phosphorylation site they discovered, does the algorithm give equal consideration to all kinases regardless of how well-studied they are?[30]). Matthew's effect (also commonly referred to as "rich gets richer") is quite pronounced in biology - according to the Understudied Protein Initative that was announced in May 2022[68], “95% of all life science publications focus on a group of 5,000 particularly well-studied human proteins". This effect is also a critical source of bias during the evaluation of machine learning algorithms for biological problems.

In machine learning applications involving network biology, while new algorithms are being evaluated, Matthew’s effect is rarely considered. For example, there is no established notion of fairness in terms of an algorithm’s ability to discover new information on well-studied vs. under-studied biological entities. However, network models are particularly vulnerable to amplifying the Matthew’s effect, since the networks utilized have skewed degree distributions[77] where the degree of a node can depend on how much the respective biological entity is studied in the literature[29]. To make things worse, as we[145] and other groups[36] have demonstrated in previous studies, the benchmarking data that is used to evaluate algorithms also tend to be biased toward well-studied biological entities (this is demonstrated in Chapter 2 for kinase activity inference problem). As a consequence, for a wide range of biomedical problems that utilize networks that represent the accumulated human knowledge, the algorithms that learn the degree bias in the networks can appear to be superior when evaluated on the biased benchmarking data, even when information related to the inherent biological mechanisms themselves may not contribute much to the observed performance. This topic is described more in depth in Chapter 4.

1.4 Summaries of the chapters

1.4.1 Chapter 2: Robust inference of kinase activity using functional networks

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

1.4.2 Chapter 3: Co-Phosphorylation networks as an additional resource for proteomic data analysis

Protein phosphorylation is a ubiquitous regulatory mechanism that plays a central role in cellular signaling and the characterization of phosphorylation dynamics is integral for understanding many critical diseases like cancer and Alzheimer's disease. With evergrowing number of studies utilizing mass-spectrometry based technologies, the availability of high throughput phosphorylation datasets provides unprecedented opportunities to examine signaling landscapes using computational approaches. Here, we comprehensively investigate the functional relevance of "co-phosphorylation", defined as the correlated phosphorylation of a pair of phosphosites. Our results across 9 phospho-proteomic datasets consistently show that functionally associated sites tend to exhibit significant positive or negative co-phosphorylation. We show that this enables high precision predictions of sites that are on the same pathway or targeted by the same kinase. These results establish co-phosphorylation as a useful resource for analyzing phospho-proteins in a network context, which will likely help extend our knowledge on cellular signaling and its dysregulation.

1.4.3 Chapter 4: Fair evaluation of link prediction algorithms in network biology

In the context of biomedical applications, new link prediction algorithms are continuously being developed and these algorithms are typically evaluated computationally, using test sets generated by sampling the edges uniformly at random. However, as we demonstrate, this creates a bias in the evaluation towards "the rich nodes", i.e., those with higher degrees in the network. More concerningly, we demonstrate that this bias is prevalent even when different snapshots of the network are used for evaluation as recommended in the machine learning community. This leads to a cycle in research where newly developed algorithms generate more knowledge on well-studied biological entities while the under-studied entities are commonly ignored. To overcome this issue, we propose a weighted validation setting focusing on under-studied entities and present strategies to facilitate bias-aware evaluation of link prediction algorithms. These strategies can help researchers gain better insights from computational evaluations and promote the development of new algorithms focusing on novel findings and under-studied proteins. We provide a web tool to assess the bias in evaluation data at: https://yilmazs.shinyapps.io/colipe/

1.4.4 Chapter 5: Making Proteomics Accessible: RokaiXplorer for interactive analysis of phospho-proteomic data

Compared to the multitude of tools available for analyzing various omics data types, there remains a notable scarcity of user-friendly solutions specifically tailored for the analysis and visualization of proteomics and phospho-proteomics data. This limitation often poses a barrier, particularly for researchers lacking specialized training in proteomics or data science. To address this critical gap and foster accessibility to statistical analyses in this domain, we present RokaiXplorer—an interactive web tool that streamlines the processing, analysis, and visualization of proteomic and phospho-proteomic data through an intuitive online interface. RokaiXplorer offers a comprehensive suite of modules designed to facilitate the analysis of key aspects such as phosphosites, proteins, kinases, and gene ontology terms. The tool encompasses a diverse range of functionalities, including data normalization, statistical testing, enrichment analysis, and interactive visualizations. Moreover, RokaiXplorer allows researchers to effortlessly deploy their own data browsers, enabling the sharing of research data and findings interactively and promotes increased transparency in research. Overall, we anticipate that RokaiXplorer will be widely embraced by the scientific community as a valuable tool for the analysis of phospho-proteomic data, owing to its simplicity and efficiency in enabling multi-level data analysis within a single application. To access RokaiXplorer, please visit: http://explorer.rokai.io.

1.5 A layman's summary - Explain like I'm 5 (ELI5)

This research is about studying proteins in our bodies to understand how they work together. Scientists in this work have developed a new tool called RoKAI that helps them figure out how certain proteins called kinases are behaving. Kinases are like the actors in a play, and RoKAI helps us know if any of these actors are misbehaving and causing problems in our body. To make sure the tool is fair, the scientists looked at potential biases and found ways to make it more accurate. They also made the tool easy to use, like a game, so other scientists can play with it too. It helps them look at a lot of data and see if there are any patterns or important information that can help us find new ways to treat diseases like cancer, Alzheimer's, and Parkinson's. Overall, this research helps us understand how proteins work in our body and find new ways to help people who are sick!

1.5.1 Exploring Proteins: A Big Book with Exciting Chapters!

This thesis is like a big book and it has different chapters. In Chapter 1, the writer talks about why their work is important and what each chapter is about.

In Chapter 2, the writer explains a special computer program called RoKAI. This helps scientists understand how some proteins called kinases work in our bodies.

In Chapter 3, the writer talks about another tool called Co-P networks. These networks help scientists predict how different proteins interact and work together in our bodies.

In Chapter 4, the writer talks about a problem with how scientists test their tools. They usually focus on studying well-known proteins, but this might not be fair to the proteins we don't know much about. The writer suggests ways to make the tests more fair.

In Chapter 5, the writer want to make it easier for scientists to study proteins by creating a special website called RokaiXplorer. It helps them look at data about different diseases and discover important information.

Overall, the thesis is about finding new ways to understand proteins, hoping to help scientists study diseases better.

2 Robust Inference of Kinase Activity using Functional Networks

2.1 Introduction

Protein phosphorylation is a ubiquitous mechanism of post-translational modification observed across cell types and species, and plays a central role in cellular signaling. Phosphorylation is regulated by networks composed of kinases, phosphatases, and their substrates. Characterization of these networks is becoming increasingly important in many biomedical applications, including identification of novel disease specific drug targets, development of patient-specific therapeutics, and prediction of treatment outcomes[21,113].

In the context of cancer, identification of kinases plays a key role in the pathogenesis of specific cancers and their subtypes, leading to the development of kinase inhibitors for targeted therapy[13,99,108,154]. Disruptions in the phosphorylation of various signaling proteins have also been implicated in the pathophysiology of various other diseases, including Alzheimer's disease[95,112], Parkinson's disease[63], obesity and diabetes[19,23], and fatty liver disease[110], among others. As a consequence, there is increased attention to monitoring the phosphorylation levels of phosphoproteins across a wide range of biological contexts and inferring changes in kinase activity under specific conditions.

Mass spectrometry (MS) provides unprecedented opportunities for large-scale identification and quantification of phosphorylation levels[28]. Typically, thousands of sites are identified in a single MS run. Besides enabling the characterization of the changes in the activity of phospho-proteins, MS-based phospho-proteomic data offers insights into kinase activity based on changes in the phosphorylation of known kinase substrates[17,34]. Observing that phosphorylation levels of the substrates of a kinase offer clues on kinase activity,[34] use a Kolgomorov-Smirnov statistic to compare the phosphorylation distributions of substrate sites and all other phosphosites. Building on this idea, kinase substrate enrichment analysis (KSEA)[17] infers kinase activity based on aggregates of the phosphorylation levels of substrates and assess the statistical significance using Z-test.[93] develop these ideas further by introducing a heuristic machine learning method, IKAP, which additionally models the dependencies between kinases that phosphorylate the same substrate. Other approaches[102,125] adapt the widely-used gene set enrichment analysis (GSEA)[124] for kinase activity inference problem. In parallel to these, a new branch of computational approaches focus on single samples to infer kinase activity[10,35,65,136].

Despite the development of algorithms that utilize relatively sophisticated models, KSEA remains one of the most-widely used tools for kinase activity inference[139]. This can be largely attributed to the constraints posed by limited comprehensiveness of available data, prohibiting the utility of such sophisticated models. Available kinase annotations still provide very little coverage (less than 10%) for phosphosites identified in MS experiments[97]. The coverage of MS-based phosphoproteomics is also limited, and many sites existing in sample may be unidentified due to technical factors[79]. Computationally predicted kinase-substrate associations[46,78] are successfully utilized to expand the scope of kinase activity inference[138]. However, the coverage of computationally predicted associations is also limited[4] and most algorithms can only make predictions for well-studied kinases[30].

With a view to expanding the scope of kinase activity inference, we develop a framework that comprehensively utilizes available functional information on kinases and their substrates. We hypothesize that biologically significant changes in signaling manifest as hyper-phosphorylation or de-phosphorylation of multiple functionally related sites. Therefore, having consistently hyper-phosphorylated (or de-phosphorylated) sites in the functional neighborhood of a phosphosite can provide further evidence about the changes in the phosphorylation of that site. Our framework, Robust Kinase Activity Inference (RoKAI), uses a heterogeneous network model to integrate relevant sources of functional information, including: (i) kinase-substrate associations from PhosphositePlus[49], (ii) co-evolution and structural distance evidence between phosphosites from PTMcode[90], and (iii) protein-protein interactions (PPI) from STRING[128] for interactions between kinases. On this heterogeneous network, we propagate the quantifications of phosphosites to compute representative phosphorylation levels capturing coordinated changes in signaling. To predict changes in kinase activity, we use these resulting representative phosphorylation levels in combination with existing kinase activity inference methods.

In order to increase the coverage of network propagation, we develop an electric circuit based model [24,33] that is specifically designed to incorporate missing sites not identified by MS. While RoKAI does not impute phosphorylation levels for unidentified sites (i.e., it is not intended to fill in missing data), it uses these sites to bridge the functional connectivity among identified sites. Similar electric circuit based models have been employed in the analysis of expression quantitative trait loci (eQTL) to identify causal genes and dysregulated pathways [58,126]. However, one important distinction is that the electric circuit model in RoKAI does not aim to uncover intermediate nodes between select target nodes, rather, it propagates all available quantifications over the network in order to reduce the noise by capturing consistent changes in the functional neighborhood of every node.

A recent study by[44] benchmarks substrate based inference approaches using a comprehensive atlas of human kinase regulation[102], encompassing more than fifty perturbations. Using this dataset, we systematically benchmark the improvement provided by RoKAI on the performance of a variety of kinase activity inference methods. In our computational experiments, we observe that the benchmark data is substantially biased in favor of "rich kinases" with many known substrates. Our results show that methods that appear to provide superior performance (e.g., methods that utilize statistical significance) accomplish this by increasing bias toward such rich kinases (since statistical power goes up with increasing number of observations). Motivated by this observation, we systematically evaluate the robustness of kinase activity inference methods using Monte Carlo simulations with varying levels of missingness. The results of this analysis shows that methods biased toward rich kinases are more vulnerable to incompleteness of available kinase-substrate annotations.

Next, we characterize the contribution of each source of functional information on enhancing kinase-activity inference. Our results show that incorporation of "shared kinase associations" (i.e., transferring information between sites that are targeted by the same kinase) significantly improves kinase activity inference. We observe that, other sources of functional information considered (PPI, co-evolution and structure distance evidence) also provide statistically significant information for kinase activity inference. However, their contribution is smaller in comparison due to either (i) limited coverage or (ii) redundancy with existing kinase-substrate annotations. Finally, we systematically investigate the performance of RoKAI in improving the performance of kinase activity methods. Results of these computational experiments show that RoKAI consistently improves the accuracy, stability, and robustness of all kinase activity inference methods that are benchmarked.

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

[†] http://rokai.io

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

2.2 Results

2.2.1 Robust inference of kinase activity with RoKAI

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

Based on this hypothesis, we develop a heterogeneous network model (with kinases and phosphosites as nodes) to propagate the phosphorylation of sites across functional neighborhoods. In this model, each edge has a conductance allowing some portion of the phosphorylation to be carried to the connecting nodes (illustrated in Figure 2.1). Therefore, the propagated phosphorylation level of a site represents an aggregate of the phosphorylation of the site and the sites that are (directly or indirectly) functionally associated with it. Consequently, the propagated phosphorylation profiles are expected to capture coordinated changes in signaling, which are potentially less noisy and more robust.

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

2.2.2 Experimental Setup

In this section, we describe our benchmarking setup for assessing the performance and robustness of kinase activity inference methods. First, we demonstrate the

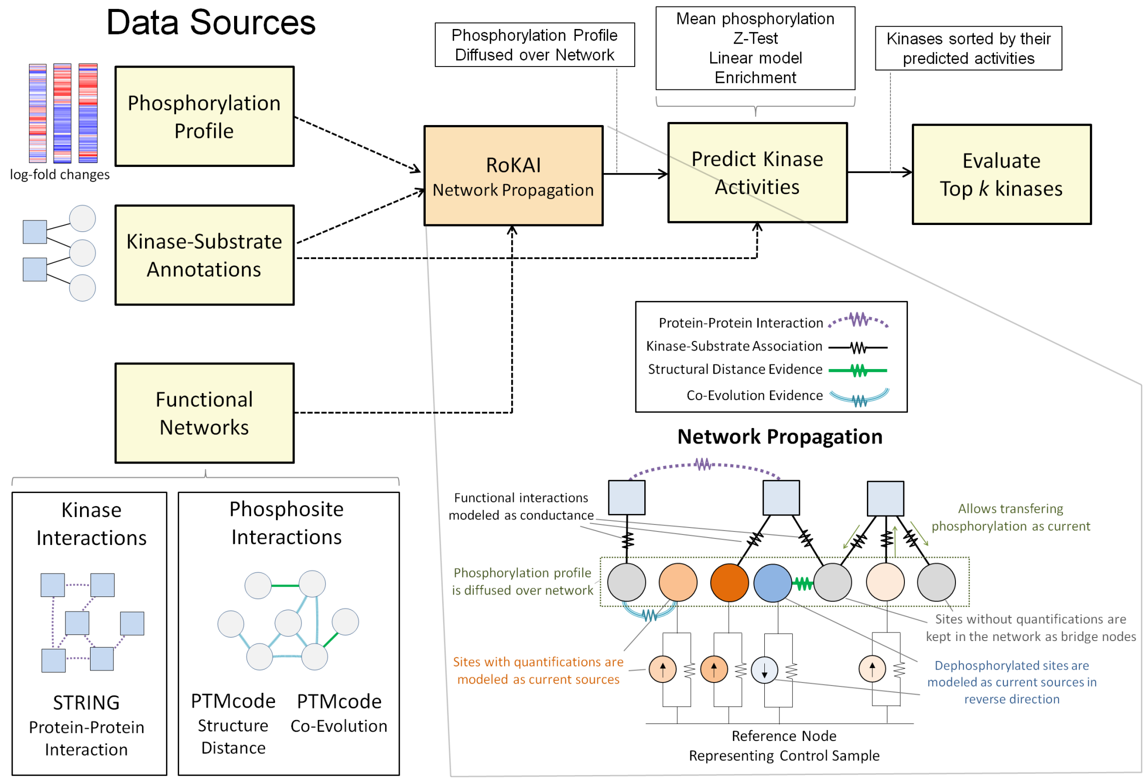

Data Sources
log-fold changes
Phosphorylation Profile
Kinase-Substrate Annotations
Functional Networks
Kinase Interactions
STRING
Protein-Protein Interaction
Phosphosite Interactions
PTMcode
Structure Distance
PTMcode
Co-Evolution
Phosphorylation Profile Diffused over Network
RoKAI
Network Propagation
Mean phosphorylation
Z-Test
Linear model
Enrichment
Predict Kinase Activities
Kinases sorted by their predicted activities
Evaluate Top *k* kinases
Protein-Protein Interaction
Kinase-Substrate Association
Structural Distance Evidence
Co-Evolution Evidence
Network Propagation
Functional interactions modeled as conductance
Allows transfering phosphorylation as current
Phosphorylation profile is diffused over network
Sites without quantifications are kept in the network as bridge nodes
Sites with quantifications are modeled as current sources
Dephosphorylated sites are modeled as current sources in reverse direction
Reference Node
Representing Control Sample

Figure 2.1. The workflow and the key idea of RoKAI. Traditional algorithms for kinase activity inference use condition-specific phosphorylation data and available kinase-substrate associations to identify kinases with differential activity in each condition. RoKAI integrates functional networks of kinases and phosphorylation sites to generate robust phosphorylation profiles. The network propagation algorithm implemented by RoKAI ensures that unidentified sites that lack quantification levels in a condition can still be used as bridges to propagate phosphorylation data through functional paths.

bias in the gold standard benchmarking data and show how this bias can lead to misleading conclusions on the performance of existing methods. Next, we introduce a robustness analysis in order to (i) overcome the effect of bias on performance estimations, and (ii) to assess the reliability of these algorithms in the presence of missing data. To characterize the value added by RoKAI, we start by assessing the utility of different sources of functional information in inferring kinase activity. Next, by focusing on a baseline kinase activity inference method (mean substrate phosphorylation), we systematically assess the incorporation of various networks with RoKAI in enhancing the accuracy and robustness of the inference. We then assess the generalizability of these results to a broad range of kinase activity inference methods. Afterwards, we investigate whether RoKAI's ability to incorporate missing sites in its functional network contributes to the improvement of kinase activity inference. Finally, we explore the effect of including predicted kinase-substrate associations within the RoKAI's framework.

2.2.3 Benchmarking Setup

Benchmarking data: [102] compiled phospho-proteomics data from a comprehensive range of perturbation studies and used these data to comprehensively benchmark the performance of kinase activity inference methods [44]. This benchmark data brings together 24 studies spanning 91 perturbations that are annotated with at least one up-regulated or down-regulated kinase. In each of the studies, the phosphorylation levels of phosphosites are quantified using mass spectrometry. After applying quality control steps (as described in the methods), we analyze a subset of this dataset encompassing 80 perturbations and 53 636 phosphosites identified in at least one of these 80 perturbations. Overall, for these 80 perturbations, there are 128 kinase-perturbation annotations (which is considered gold standard) for 25 different kinases (listed in Supplementary Data 1). In our computational experiments, we use this dataset to assess the robustness of existing kinase activity inference methods and validate our algorithms.

Kinase-substrate annotations: We obtain existing kinase-substrate associations from PhosphositePlus [49]. PhosphositePlus contains a total of 10476 kinase-substrate links for 371 distinct kinases and 7480 sites. Among these annotated sites, 2397 have quantifications in the perturbation data. These sites have a total of 3877 kinase-substrate links with 261 kinases.

Benchmarking metric: The main purpose of kinase activity inference is to prioritize kinases for additional consideration and ideally for experimental validation. However, in practice, it is typically very costly to experimentally validate more than a few kinases [20] and it is infeasible to manually inspect more than a couple dozen. Whereas, benchmarking approaches that are employed in the literature like area-under receiver operating characteristics curve (AUROC) and precision at recall

0.5 consider high number of predictions (k), including kinases that are less significant/active. We find such measures problematic because, even though they include the performance for a high number of predicted kinases in their calculation, it would not be practical for a potential user to inspect or use that many predictions. To take this consideration into account, we use a metric, "top-k-hit", that focuses on the top-k kinase predictions for small values of k. Since the gold standard dataset is incomplete, this metric essentially serves as a minimum bound on the expected probability of discovering an up-regulated or down-regulated kinase if top k kinases predicted by the inference method were to be experimentally validated. In our experiments, we use k=10 (unless otherwise specified) since it represents a reasonable number of kinases to be put to additional scrutiny before experimental validation

2.2.4 Existing Inference Methods

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

Mean (baseline method): One of the simpliest kinase activity inference methods employed by KSEA[17]. This method represents the activity of a kinase as the mean phosphorylation of its substrates.

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

Linear model: The linear model, considered by[44], aims to take into account of the dependencies between kinases that phosphorylate the same site. In this model, the phosphorylation of a site is modeled as summation of the activities of kinases that phosphorylate the site. A similar (but more complex) approach is also utilized by IKAP[93].

GSEA:[125] and[102] adopt gene set enrichment analysis (GSEA), a widely used method in systems biology[124], to infer kinase activity by assessing whether the target sites of a kinase exhibit are enriched in terms of their phosphorylation fold change compared to other phosphosites.

2.2.5 Bias and robustness of existing inference methods

Previous benchmarking by[44] suggests that methods that rely on statistical significance (Z-Score and GSEA) are superior to their alternatives. However, as shown in Figure 2.2(a)), we observe that there is substantial bias in the benchmarking data: “rich" kinases (i.e., kinases with many known substrates) are significantly over-represented among the 25 annotated kinases that have at least one perturbation (median number of substrates: 29 for annotated and 4 for not-annotated kinases, K-S test p-value$<$3.5e-7 for the comparison of annotated kinases with others in terms of their distribution of number of substrates).

Since methods that rely on statistic significance have a positive bias for kinases with many substrates (statistical power is improved with number of observations), we hypothesize that this is the reason behind their observed superior performance. To test this hypothesis, we benchmark two additional inference methods that are artificially biased for kinases with many substrates: (i) *Sum:* Sum of phosphorylation (log-fold changes) of substrates, and (ii) *Num:* Number of substrates, used directly as the predicted activity of a kinase (clearly, this method does not use the phosphorylation levels of sites, thus, it always generates the same ranking of kinases regardless of the phosphorylation data). As shown in Figure 2.2(b), methods that are artifically biased for rich kinases appear to have better predictivity over the alternatives.

In order to overcome the effect of this bias on evaluation, we perform a robustness analysis where we hide a percentage of the known substrates of the 25 annotated kinases from the inference methods. The results of this analysis are shown in Figure 2.2(c). As seen in the figure, even though methods biased for rich kinases appear to have higher predictivity when all of the available kinase substrate annotations are used, they are not robust to increasing rate of missingness

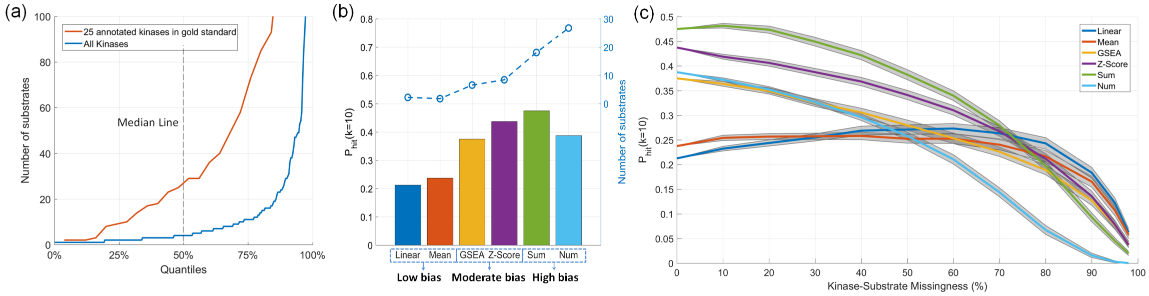


Figure 2.2. Existing benchmark data for kinase-activity inference is biased toward kinases with high number of substrates and can be misleading in assessing the performance of inference methods. (*a*) Inverse cumulative distribution of the number of substrates for the 25 kinases that are annotated with a perturbation in gold standard benchmarking data compared to all kinases. The x-axis indicates the quantiles. For example, the value on the y-axis that corresponds to $x = 50\%$ indicates the median number of substrates. (*b*) Performance and bias of baseline kinase activity inference methods. The bars show the probability of identifying an annotated "true" kinase in top 10 predicted kinases (PHit10). The dashed line indicates the average number of substrates of the top 10 predicted kinases for the corresponding method. The high-bias methods (*Sum*: total substrate phosphorylation, and *Num*: number of substrates) are not used in the literature, but are shown here to illustrate the effect of bias on performance assessment. (*c*) The robustness analysis of the methods for missingness in kinase-substrates links. The x-axis shows the percentage of (randomly selected) kinase-substrates links of 25 gold standard kinases hidden from the kinase activity inference methods. The gray areas indicate the 95% confidence intervals for the mean performance across 100 runs.

in kinase-substrate annotations. The performance of artificially biased methods fall below that of the low-biased methods (e.g., *Mean* and *Linear Model*) at around 50% missingness. At around 80% missingness, the effect of the bias on evaluation is mitigated i.e., the difference between number of substrates of 25 annotated kinases and the remaining kinases is not at a statistically detectable level anymore. Thus, the performance of biased (e.g., statistical significance based) methods fall below the low-bias methods at around 80% missingness. These observations make the reliability of biased methods highly questionable since the available kinase-substrate annotations are largely incomplete.

2.2.6 Utility of functional networks for inferring kinase activity

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

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

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

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

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

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

For each of the functional or structural networks described above, we compute a network activity prediction score for each kinase based on the mean phosphorylation of sites that are considered of interest for the corresponding network (illustrated in Figure 2.3). Note that, except for the baseline network (known kinase-substrates), we do not use the phosphorylation levels of the kinase's own substrates to compute the scores for each network.

To characterize the contribution of each source of functional information on enhancing kinase-activity inference, we consider the following metrics:

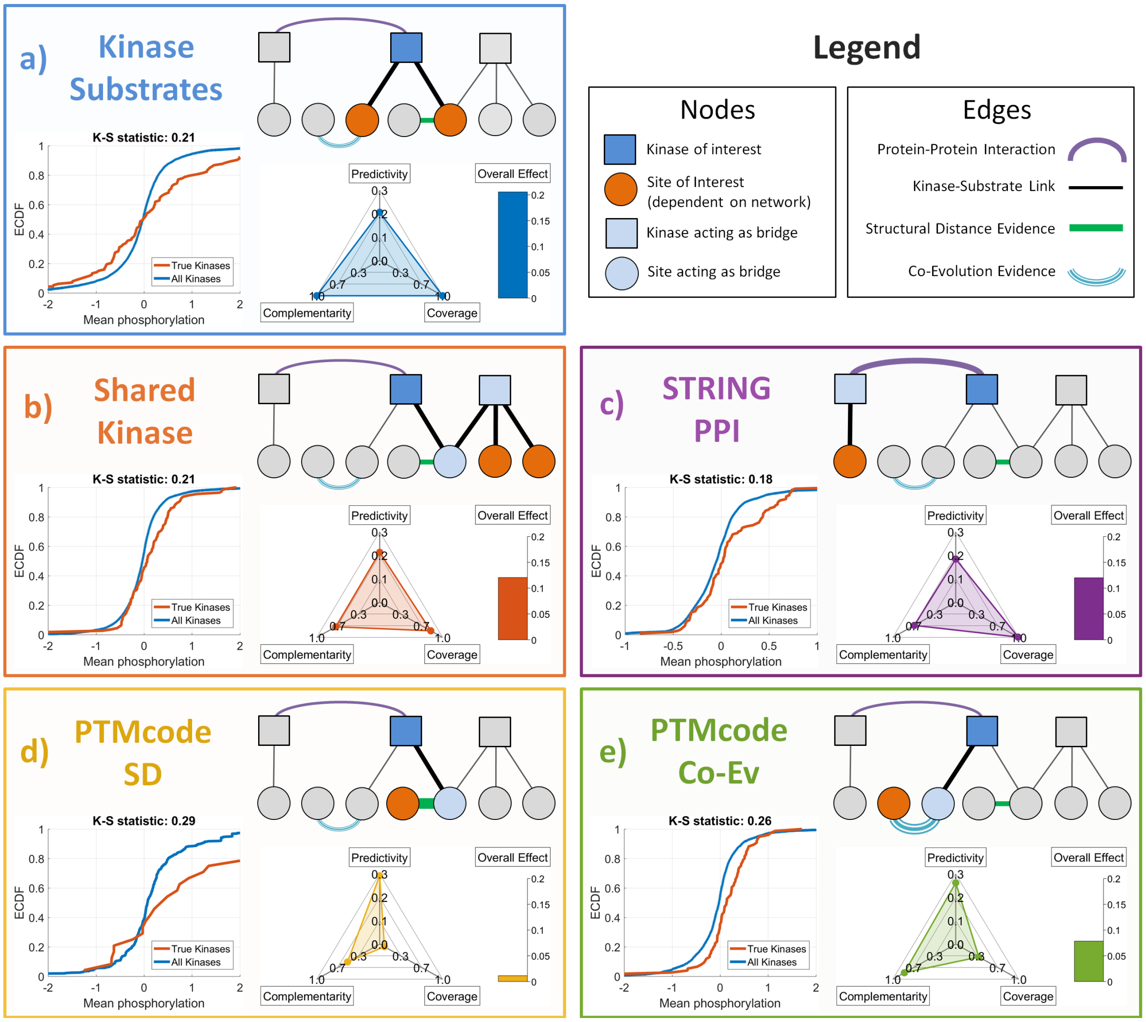


Figure 2.3. Utility of available functional or structural information in providing information on kinase activity. Each panel (labeled a to e) represents a different information source. The first panel (Kinase-Substrates) represents the information source that is utilized by all existing kinase activity inference methods, whereas, the other four panels represent the information sources introduced here. In each panel, the relationship between a kinase (blue square) and the site(s) (red circles) that provide(s) information on the activity of the kinase is illustrated. The bottom-left plot compares the empirical cumulative distribution (ECDF) of the phosphorylation levels of the "information-providing" sites for “true" perturbed kinases in the benchmark data against all kinases. The bottom-right plot shows the *predictivity* (accuracy in predicting kinase activity), *complementarity* (information provided in addition to the substrates of the kinase), and *coverage* (fraction of kinases that are affected) of the information source. The bars represent the overall effect of the information source calculated as the product of the scores shown on the other axes.

- Predictivity: To assess the utility of functional networks in predicting the "true" perturbed kinases in gold standard dataset, we use Kolmogorov-Smirnov (K-S) test[83] comparing the distribution of network scores for true kinases with the distribution of all other kinases. For each functional network, we consider the K-S statistic as the *predictivity score* of the corresponding network.
- Coverage: The network scores contain missing values for kinases without any edges in the corresponding functional networks. Thus, while assessing predictivity (as explained above), we utilize only the kinases with a valid network score. To take missing data into account, we compute a *coverage score* which is equal to the percentage of kinases with a valid network score with respect to that functional network.
- Complementarity: We aim to utilize the functional networks as an information source that complements available kinase-substrate associations. If there is statistical dependency between functional network scores and the activity inferred by the kinase's own sites, the information provided by the network would be redundant. We use complementarity score as one minus absolute linear (Pearson) correlation between the score of each network scores and kinase activity inferred based on the kinase's own substrates. Since the kinase-substrate association network serves as our baseline, we consider it to have 100% complementarity.
- Overall Effect: To quantify the overall contribution of the functional networks for improving the predictions of kinase activity, we combine the predicity, coverage and complementarity scores and obtain an overall effect score:

$$\text{Overall Effect} = \text{Predictivity} \times \text{Coverage} \times \text{Complementarity} \tag{2.1}$$

The results of this analysis are shown in Figure 2.3. As can be seen, all considered functional information sources exhibit statistically significant predictivity of the kinase-perturbations according to two-sample Kolmogorov-Smirnov (K-S) test:

Known kinase-substrates (K-S statistic = 0.21, p-value≤1.3e-4), Shared-kinase interactions (K-S statistic = 0.21, p-value≤7.3e-5), Protein-protein interactions (K-S statistic = 0.18, p-value≤8.7e-4), Structure distance evidence (K-S statistic = 0.29, p-value≤0.03), Co-evolution evidence (K-S statistic = 0.26, p-value≤5.5e-5). We observe that the incorporation of "shared kinase associations" in addition to the known kinase substrates has the most overall contribution to the inference of kinase activities (Figure 2.3, panels a and b), followed by kinase-kinase interactions (Figure 2.3, panel c). Even though co-evolution and structural distance networks exhibit strong predictivity, their overall contribution is relatively low due to their limited coverage and redundancy with existing kinase-substrate annotations (Figure 2.3, panels d and e).

2.2.7 Benchmarking RoKAI-enhanced inference methods

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

Kinase-Substrate (KS) network: The network used by RoKAI consists only of the known kinase-substrate interactions. Use of this network allows RoKAI to utilize sites with shared-kinase interactions (illustrated in Figure 2.3, panel b), i.e., sites that are targeted by the same kinase contribute to their refined phosphorylation profiles.

KS+PPI network: In addition to KS, this network includes weighted protein-protein interactions between kinases. This allows propagation of phosphorylation levels between substrates of interacting kinases (illustrated in Figure 2.3, panel c).

KS+PPI+SD network: In addition KS+PPI, this network includes interactions between phosphosites with structural distance (SD) evidence obtained from PTMcode. This allows the utilization of sites that are structurally proximate to the substrates of a kinase (illustrated in Figure 2.3, panel d).

KS+PPI+SD+CoEv (combined) network: In addition KS+PPI+SD, this network includes interactions between phosphosites with co-evolution evidence obtained

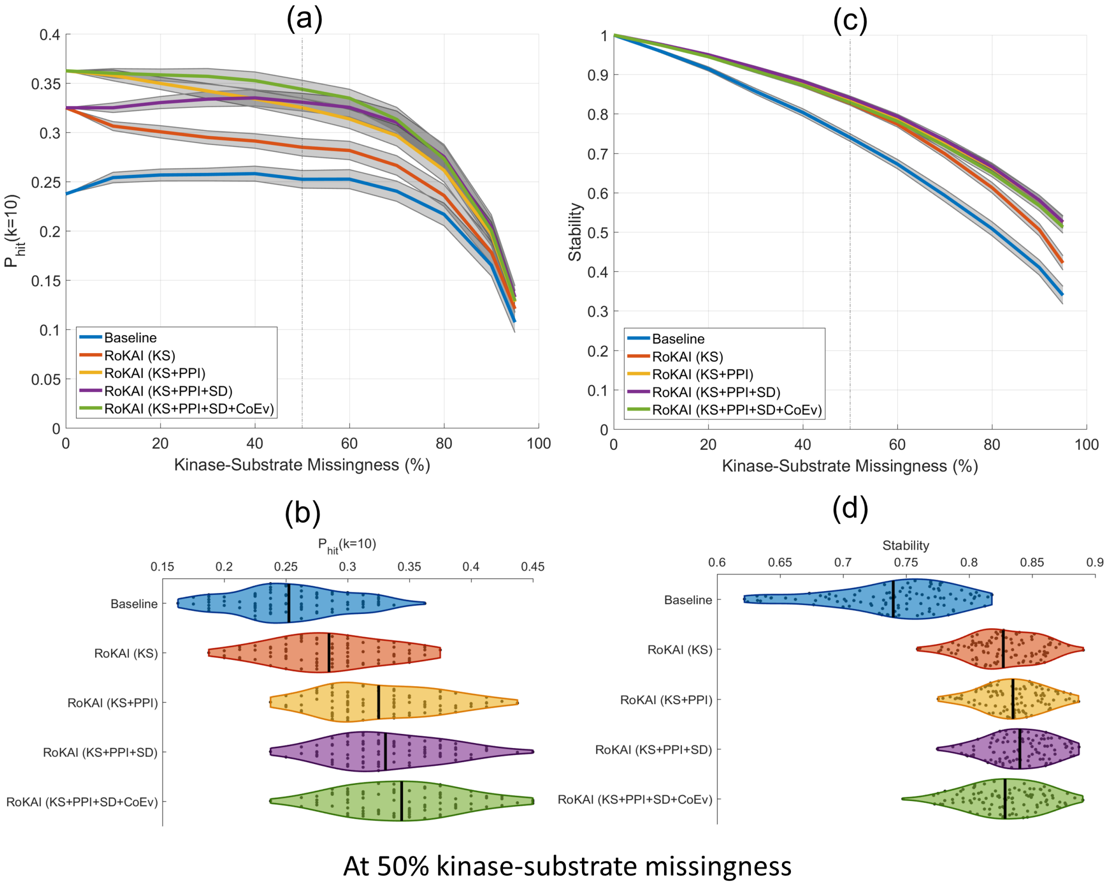


Figure 2.4. Comparison of the accuracy and stability of mean substrate phoshorylation and its RoKAI-enhanced versions using various functional or structural networks. (a) The hit-10 performance (the probability of ranking a true perturbed kinase in the top ten), as a function of missingness (the fraction of kinase-substrate associations that are hidden). The shaded areas indicate the 95% confidence intervals for the mean performance across 100 randomized runs. (b) The distribution of hit-10 probabilities for 100 runs at 50% missingness. (c) Stability of the inferred activities (measured by the average squared correlation between inferred activities when different portions of kinase-substrate associations are hidden from the inference methods) as a function of missingness. The shaded areas indicate 95% confidence intervals for the mean across 100 runs. (d) The distribution of stability for 100 runs at 50% missingness.

from PTMcode. This allows the utilization of sites that are evolutionarily similar to the substrates of a kinase (illustrated in Figure 2.3, panel e).

To assess the performance of RoKAI with these networks, we use the benchmarking data from the atlas of kinase regulation. As previously discussed, this dataset is heavily biased toward kinases with many known substrates. To overcome the effect of this bias on evaluation, we perform robustness analyses where we hide a portion of known kinase-substrate interactions of the 25 kinases that have perturbations. For predicting kinase activity, we use the mean substrate phosphorylation (baseline inference method) and compare the performance of original predictions and RoKAI-enhanced predictions. As shown in Figure 2.4(a) and Figure 2.4(b), RoKAI consistently and significantly ($p < 0.05$) improves the predictions in a robust manner for varying levels of missing data.

The functional networks that contribute most to the improvements in prediction performance of RoKAI are respectively: KS network (modeling shared-kinase interactions) followed by PPI (for including kinase-kinase interactions) followed by co-evolution evidence. Compared to these, including structural distance evidence in the network has a minor effect on prediction performance. This is in line with the overall effect size estimations (shown in Figure 2.3). Since structural distance network has relatively small number of such edges, it provides low coverage and a minor effect size even though the existing edges are estimated to be more predictive of kinase activity compared to other networks.

To further evaluate the robustness of the predictions, we assess the *stability* i.e., the expected degree of aggreement between the predicted kinase activity profiles when different kinase substrates are used (e.g., because some sites are not identified by a MS run) to infer the activity of a kinase. We measure the stability by computing average squared correlation between different runs of robustness analysis (where a different portion of kinase-substrate links are used for inferring kinase activity in each run). As shown in Figure 2.4(c) and Figure 2.4(d), predictions made by RoKAI-enhanced phosphorylation profiles are significantly ($p < 0.05$) more stable in addition to being more predictive.

2.2.8 Improvement of RoKAI over a broad range of methods

Since RoKAI provides refined phosphorylation profiles (propagated by functional networks), it can be used in conjunction with any existing (or future) kinase activity inference algorithms. Here, we benchmark the performance of RoKAI when used together with existing inference methods. For each of these methods, we use

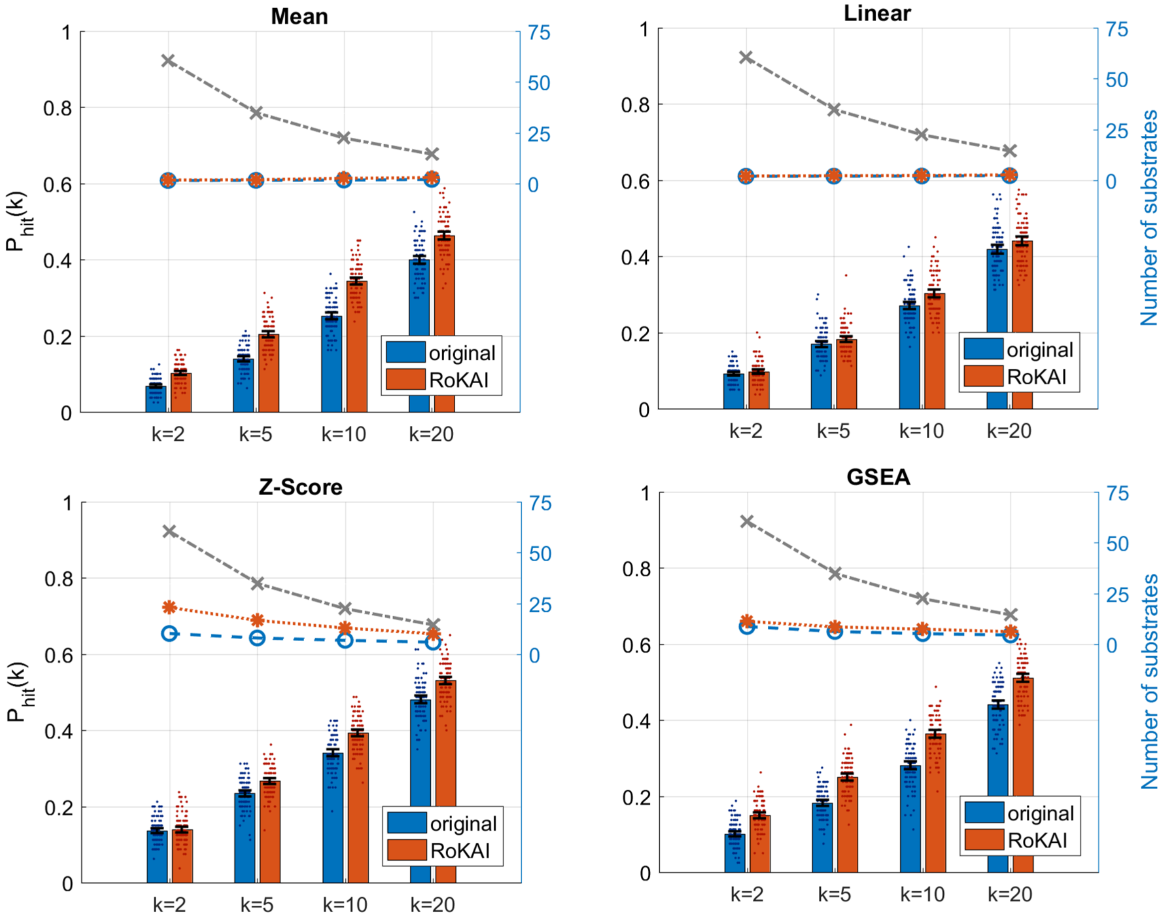


Figure 2.5. Contribution of RoKAI (combined network) in improving the performance of different kinase activity inference methods for predicting the true (annotated) kinase in the top k kinase predictions for various k. The bars show the mean probability of predicting a true kinase among the top k kinases at 50% kinase-substrate missingness. The blue bars indicate the prediction performance using the original (unmodified) phosphorylation profiles and red bars indicate the performance of using RoKAI-enhanced profiles for inferring kinase activity. The colored dashed lines indicate the average number of substrates of the top k kinases predicted by the corresponding inference method (the gray dashed line shows the maximum possible). The black error bars indicate the 95% confidence intervals for the mean performance across 100 randomized runs. The colored points around each bar indicate the performance on different runs.

the refined phosphorylation profile (obtained by RoKAI) to obtain the RoKAI-enhanced kinase activity predictions. To assess the prediction performance while addressing the bias for rich kinases, we perform robustness analysis at 50% kinase-substrate missingness and measure the top-k hit performance for $k = 2, 5, 10$

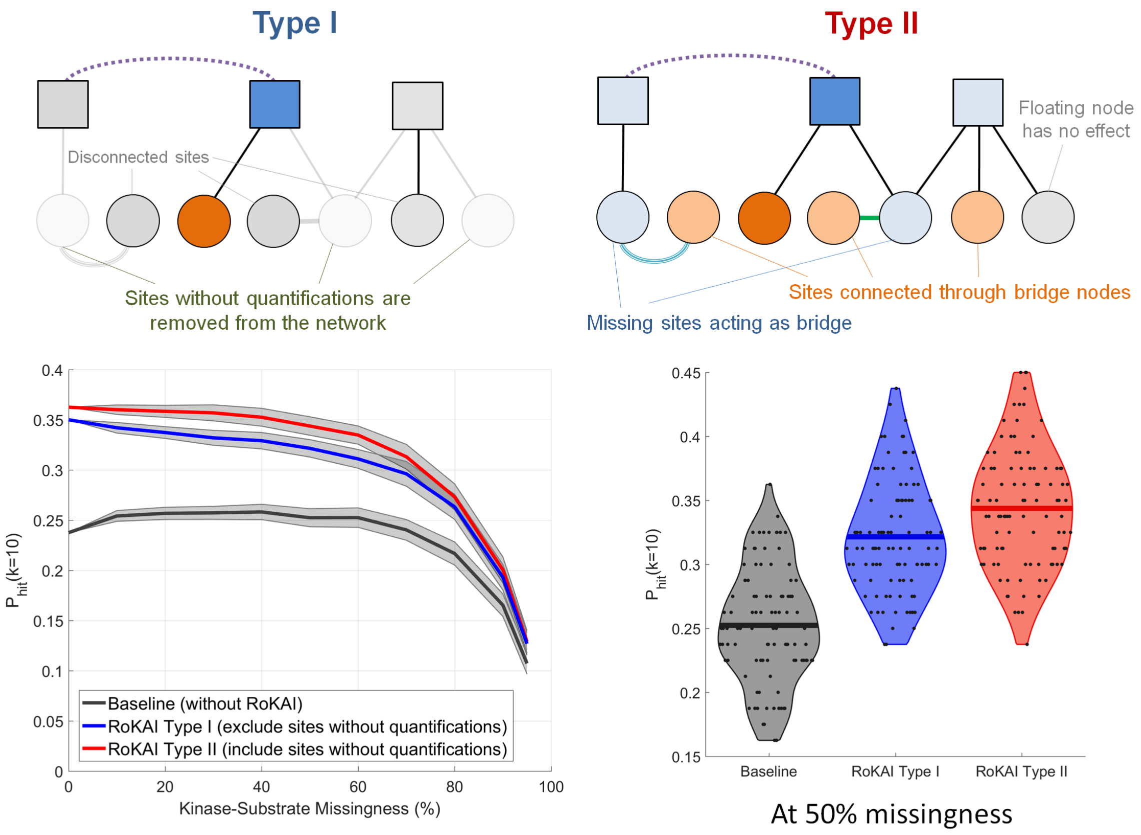


Figure 2.6. RoKAI improves kinase activity inference by enabling utilization of the unidentified sites (without quantifications) for predicting the activity of kinases. In type I (illustrated in top left), the network consists only of sites with quantifications. Whereas, in type II (illustrated in top right), the network includes sites without quantifications to utilize them as bridge nodes. (Bottom Left) Robustness analysis with respect to missingness of kinase-substrate links. The shaded area shows the 95% confidence interval for the mean performance on 100 randomized runs where different kinase-substrate links are removed. (Bottom Right) The performance of RoKAI Type I and Type II at 50% missingness. Each point indicate the performance on a different run. The lines indicate the mean performance across 100 runs.

and 20. As shown in Figure 2.5, RoKAI consistently improves the predictions of all inference methods tested.

2.2.9 Effect of incorporating unidentified sites in RoKAI

An important feature of RoKAI's network propagation algorithm is its ability to accommodate unidentified sites (i.e., sites that do not have quantified phosphorylation levels in the data) in the functional network. While RoKAI does not impute phosphorylation levels for unidentified sites (i.e., it is not intended to fill in missing data), it uses these sites to bridge the functional connectivity among identified sites. To assess the value added by this feature, we compare two versions of RoKAI: One that removes unidentified sites from the network (Type I) and one that utilizes unidentified sites as bridges (Type II). The results of this analysis are shown in Figure 2.6. The kinase activity inference activity method we use in these experiments is mean phosphorylation level. As seen in the figure, retention of unidentified sites in the network consistently improves the accuracy of kinase activity inference although the magnitude of this improvement is rather modest (in comparison to the overall improvement of RoKAI to the baseline). We observe a similar improvement for all other kinase activity inference methods that are considered.

2.2.10 Effect of incorporating predicted kinase-substrate associations

Next, we investigate the utility of using predicted kinase-substrate associations within the RoKAI's framework. For this purpose, we use NetworKIN [46], which lists its predictions separately as motif-based (NetPhorest), interaction-based (STRING), or combined (using both motif and interaction informations). To incorporate these predictions in RoKAI's framework, we consider two strategies: (i) Include the predicted kinase-substrate interactions (in addition to known substrates in PhosphositePlus) during the kinase activity inference but do not alter the RoKAI's functional network, and (ii) Include the predicted interactions in both RoKAI's functional network and during the kinase activity inference (this strategy is annotated RoKAI+).

For this analysis, we use the baseline method (mean phosphorylation) for the inference. To make the results comparable with our previous analysis (using only the known substrates in PhosphositePlus), we limit the analysis to the kinases with at least one known substrate identified in the perturbation experiments (this way, we keep the kinase set same as before). The results of this analysis are shown in Figure 2.7. Here, the x axis shows the number of predicted interactions included in the inference (i.e., as we go right on the x axis we apply a more relaxed threshold

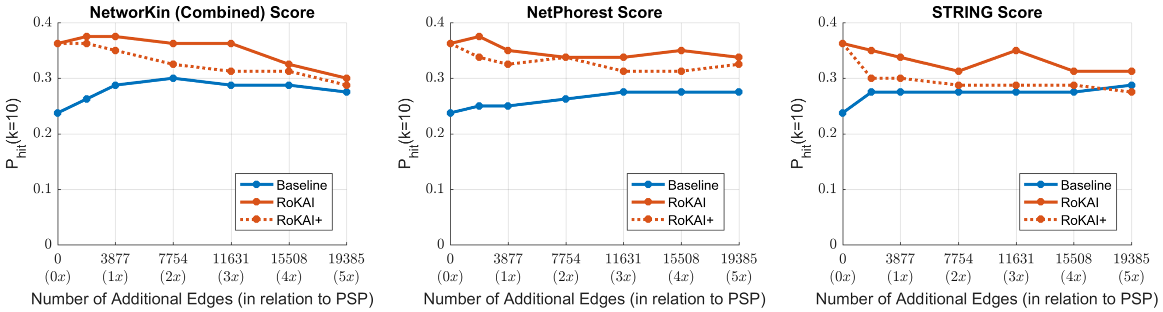


Figure 2.7. The effect of including kinase-substrate links predicted by NetworKin on kinase activity inference. Each panel shows the results for a difference scoring (used for kinase-substrate edges). In each panel, the x-axis shows the number of edges used by the inference methods in addition to the kinase-substrate annotations from PhosphositePlus (PSP). The colored blue and orange lines indicate the performance of baseline-method (mean substrate phosphorylation) and its RoKAI-enhanced version respectively. The dashed-orange line (RoKAI+) indicate the performance when the functional network of RoKAI additionally includes the predicted kinase-substrate edges by NetworKin.

on the prediction score). Thus, the leftmost point (0 at x-axis) corresponds to the case where only confirmed interactions (PhosphositePlus) are used.

As expected, the inclusion of predicted interactions in kinase activity inference improves the performance for the baseline algorithm and the performance of RoKAI-enhanced inference stays above the baseline. However, we observe that the use of predicted interactions together with RoKAI does not improve the inference further (while there is some increase in performance with the inclusion of high-confidence predictions, the inclusion of lower-confidence predictions degrades the performance). In addition, the inclusion of predicted interactions within the RoKAI's functional network always results in less accurate inference. Taken together, these observations suggest that, since RoKAI already includes functional and structural information to compute propagated phosphorylation levels, the inclusion of predicted interactions that use similar information does not further enhance the accuracy of the inference.

2.3 Discussion

By comprehensively utilizing available data on the functional relationships among kinases, phospho-proteins, and phosphorylation sites, RoKAI improves the robustness of kinase activity inference to the missing annotations and quantifications. Its implementation is available as open-source in Matlab as well as a web tool (http://rokai.io) for easy accessibility. We expect that this will facilitate the identification of understudied kinases with only a few annotations and lead to the development of novel kinase inhibitors for targeted therapy of many diseases such as cancer, Alzheimer's disease, and Parkinson's disease. As additional functional information on cellular signaling becomes available, the inclusion of these information in functional networks utilized by RoKAI will likely further enhance the accuracy and robustness of kinase activity inference.

The introduced benchmarking setup provides the opportunity to explore and compare the predictions of a variety of inference algorithms in terms of their robustness to missing annotations. It also allows the estimation of how utilization of different functional networks would influence the inference process. These features can help enable researchers to understand the trade-offs between different kinase activity inference algorithms in terms of their robustness, accuracy, and biases. As a potential resource, we provide the materials (code and data) to reproduce our analysis results in figshare (doi:10.6084/m9.figshare.12644864) that the users can adapt to test different inference methods and/or networks. Using such a framework, we believe the users can make more informed decisions for follow-up studies.

A noteworthy complication in perturbation studies that concern kinase activity inference is the effect of off-target kinases. While recent studies systematically identify off-target kinases in perturbation studies [45,60], the extension of kinase activity inference algorithms and tools like RoKAI to distinguish off-target effects remains an open problem that can advance many important applications like drug development.

An important consideration in kinase activity inference is the dependencies between phosphorylation levels of sites. Some inference methods take into account the dependency between sites that are targeted by the same kinase [44,93]. On the other hand, recent studies utilize protein expression to take into account the dependency between sites on the same protein by normalizing phosphorylation levels of the sites, but results on the effectiveness of this approach are not conclusive [3,151]. Whereas, RoKAI implicitly considers the dependencies between sites using a functional network model. We recognize the explicit modeling of the dependencies as an important open problem that can further enhance the performance and reliability of kinase activity inference.

A key motivation in developing RoKAI was to utilize the missing sites without quantifications by keeping them as bridges in the network (thus, increasing the overall coverage of the network). In our experiments, we indeed observe a consistent improvement for incorporating missing sites (as compared to disregarding them completely). However, contrary to our expectation, the magnitude of this improvement is rather modest. We hypothesize that this may be because of (i) biological redundancy i.e., sites that are reached by missing, bridge nodes may already be covered by other paths consisting of identified nodes, (ii) our incomplete knowledge of functional networks e.g., kinase-substrate annotations. To this end, construction of more comprehensive and detailed networks can potentially enhance the utility of missing sites in improving kinase activity inference. Overall, we recognize this as an important direction for future research.

2.4 Methods

2.4.1 Problem Definition

Kinase activity inference can be defined as the problem of predicting changes in kinase activity based on observed changes in the phosphorylation levels of substrates. Formally, let $K = \{k_1, k_2, ..., k_m\}$ denote a set of kinases and $S =$

$\{s_1, s_2, ..., s_n\}$ denote a set of phosphorylation sites. For these kinases and phosphosites, a set of annotations are available, where $S_i \subseteq S$ denotes the set of substrates of kinase k_i, i.e., $s_j \in S_i$ if kinase k_i phosphorylates site s_j.

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

2.4.2 Background

Mean (baseline): This is a simple method that represents the activity of a kinase as the mean phosphorylation (log-fold change) of its substrates:

$$\hat{a}_i^{\text{(mean)}} = \frac{\sum_{s_j \in S_i} q_j}{|S_i|}. \tag{2.2}$$

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

Z-score: This method normalizes the mean phosphorylation of the substrates to reflect statistical significance:

$$\hat{a}_i^{\text{(z-score)}} = \frac{\sum_{s_j \in S_i} q_j}{\sigma \sqrt{|S_i|}} = \frac{\sqrt{|S_i|}}{\sigma} \hat{a}_i^{\text{(mean)}}, \tag{2.3}$$

where σ is the standard deviation of phosphorylation (log-fold changes) across all phosphosites.

Linear model: In this model, the phosphorylation of a site is modeled as summation of the activities of kinases that phosphorylate the site:

$$q_j = \sum_{\substack{\text{for all kinases } i \\ \text{phosphorylating site } j}} \Bigg(a_i \tag{2.4}$$

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

$$\hat{a}^{(\text{linear})} = \text{argmin}_a \left\{ \sum_{s_j \in S} \Bigg(q_j - \sum_{k_i \in K_j} a_i)^2 + \lambda ||a||^2 \right\} \Bigg(\tag{2.5}$$

where K_j denotes the set of kinases that phosphorylate site s_j, and λ is an adjustable regularization coefficient. The first term in the objective function (squared loss) ensures that the inferred kinase activities are consistent with the phosphorylation levels of their substrates, whereas the second term (regularization) aims to minimize the overall magnitude of inferred kinase activities. In all experiments, we utilize a regularization coefficient of $\lambda = 0.1$ as previously done in[44].

GSEA: To infer the activity of a kinase, this method assesses whether the substrates of the kinase are more enriched compared to other phosphosites in terms of their phosphorylation. To compute the enrichment score, the sites are first ranked based on their absolute fold changes. For each kinase k_i, a running sum is computed based on the ranked list of sites. The running sum increases for each site $s_j \in S_i$ (i.e., s_j is a known substrate of k_i), and decreases for each site $s_j \notin S_i$ (i.e., s_j is not a known substrate of k_i. The maximum deviation of this running sum from zero is used as the enrichment score of a kinase. The statistical significance of this enrichment score is assessed using a permutation test. Namely, fold changes of sites are permuted $10,000$ times and enrichment scores are computed for each. The p-value for a kinase is then computed as the number of permutations with higher enrichment score than observed. As the predicted activity of a kinase, -log10 of this p-value is used.

2.4.3 Phospho-proteomics data preprocessing

Following the footsteps of previous studies[44,102], we apply some quality control steps to the phospho-proteomics data that is used for benchmarking: (i) we restrict the analysis to mono-phosphorylated peptides that are mapped to canonical transcripts of Ensembl, (ii) we average the log fold changes of technical replicates as well as peptides that are mapped to the same Ensembl position (even if the exact peptides sequences are not identical), and (iii) we filter out the peptides that are identified in only a single study to reduce the amount of false-positive phosphosites, (iv) we restrict the analysis to perturbations in the gold standard with more than 1000 phosphosite identifications (which leaves 81 perturbations). Finally, we exclude a hybrid perturbation (i.e., a mixture of both an activator and an inhibitor) from our analysis. As a result of these steps, we obtain 53636 sites identified in at least one of 80 perturbations. For these 80 perturbations, there are 128 kinase-perturbation annotations for 25 different kinases.

2.4.4 Computing benchmarking metric (top-k-hit)

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

(1) For each perturbation separately, we rank the kinases based on their absolute activities predicted by the inference method.

(2) For each perturbation, we consider the top k kinases with highest predicted activity and compare them with the "true" perturbed kinases in gold standard.

(3) If any of the top k kinases is a true kinase (i.e., a kinase that is perturbed in the experiment), we consider the inference method to be successful (i.e., a hit) for that perturbation.

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

kinases predicted by the inference method were to be experimentally validated.

2.4.5 Robust kinase activity inference (RoKAI)

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

Kinase-Substrate Associations: An edge between a kinase k_i and site s_j indicates that k_i phosphorylates s_j. These kinase-substrate associations obtained from PhosphositePlus[49], representing 3877 associations between 261 kinases and 2397 sites.

Structure Distance Evidence: This type of edge between phosphosites s_i and s_j represents the similarity of s_i and s_j on the protein structure. We obtain structure distance evidence from PTMcode[90], which contains 7821 unweighted edges between 8842 distinct sites. Note that, in this network, a large portion of the edges (7037 edges) are intra-protein.

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

Protein-Protein Interactions: An edge between kinases k_i and kinase k_j represents a protein-protein interaction between k_i and k_j. We use the protein-protein interaction (PPI) network obtained from STRING[128]. As the edge weights, we utilize the combined scores provided by STRING. Overall, the kinase-kinase interaction network contains 13031 weighted edges (weights ranging from 0 to 1) between 255 distinct kinases.

Network propagation. Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ represent RoKAI's heterogeneous functional network, where $\mathcal{V} = K \cup S$ and $\mathcal{E}$ contains four types of edges as described above. To propagate phosphorylation levels of sites over $\mathcal{G}$, we utilize an electric circuit model (illustrated in Figure 2.1). In this model, each node $n_i \in \mathcal{V}$ (kinase or phosphosite) has a node potential v_i. Each edge $e_{ij} \in \mathcal{E}$ (which can be a kinase-substrate association, kinase-kinase interaction or association between a pair of phosphosites) has a conductance c_{ij} that allows some portion of the node potential v_i of node n_i to be transferred to node n_j in the form of a *current* I_{ij}:

$$I_{ij} = (v_i - v_j)\, c_{ij} \tag{2.6}$$

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

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

$$\begin{aligned} &\textit{Incoming current} = \textit{Outgoing current} \\ q_i = v_i + \sum_{(i,j)\in E} &\bigg((v_i - v_j)c_{ij}, \qquad \text{if } n_i \text{ has quantification} \\ 0 = \sum_{(i,j)\in E} &\bigg((v_i - v_j)c_{ij}, \qquad \text{if } n_i \text{ does not have quantification} \end{aligned} \tag{2.7}$$

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

To compute the node potentials for all nodes in the network, we represent Equation (2.7) as a linear system:

$$\mathbf{C}\mathbf{v} = \mathbf{b} \tag{2.8}$$

$$C_{ij} = \begin{cases} 1 & \text{if } i = j \text{ and } n_i \text{ has quantification} \\ c_{ij} & \text{if } i \neq j \text{ and } n_i n_j \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

$$b_i = \begin{cases} q_i & \text{if } n_i \text{ has quantification} \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

Thus, the node potentials $\mathbf{v}$ can be computed using linear algebra as follows:

$$\mathbf{v} = (\mathbf{C}^\top \mathbf{C})^{-1} \mathbf{C}^\top \mathbf{b} \tag{2.11}$$

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

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

$$\hat{q}_j = v_j. \tag{2.12}$$

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

3 Co-Phosphorylation: An additional resource for proteomic data analysis and functional network prediction

3.1 Introduction

Protein phosphorylation is a ubiquitous mechanism of post-translational modification observed across cell types and species. Recent estimates suggest that up to 70% of cellular proteins can be phosphorylated[135]. Phosphorylation is regulated by networks composed of kinases, phosphatases, and their substrates. Characterization of these networks is increasingly important in many biomedical applications, including identification of novel disease-specific drug targets, development of patient-specific therapeutics, and prediction of treatment outcomes[21,113].

Phosphorylation is particularly important in the context of cancer, as down-regulation of tumor suppressors and up-regulation of oncogenes (often kinases themselves) by dysregulation of the associated kinase and phosphatase networks are shown to have key roles in tumor growth and progression[42,119]. To this end, characterization of signaling networks enables exploration of the interconnected targets[38,140,145] and identification of causal pathways[156], leading to the development of kinase inhibitors to treat a variety of cancers[13,154]. Disruptions in the phosphorylation of various signaling proteins have also been implicated in the pathophysiology of various other diseases, including Alzheimer's disease[95], Parkinson's disease[63], obesity and diabetes[19], and fatty liver disease[110], among others. As a consequence, there is increased attention to cellular signaling in biomedical applications, motivating researchers to study phosphorylation at larger scales[44].

In response to the growing need for large-scale monitoring of phosphorylation, advanced mass spectrometry (MS)-based phospho-proteomics technologies have exploded[28]. These technologies enable simultaneous identification and quantification of thousands of phosphopeptides and phosphosites from a given sample[144]. These developments result in the generation of data representing the phosphorylation levels of hundreds of thousands of phosphosites under various conditions across a range of biological contexts, including samples from human patients, cell lines, xenografts, and mouse models[79]. As compared to the widespread availability and sharing of genomic and transcriptomic data, public repositories of phospho-proteomic data are sparse, but growing. As a consequence, secondary or integrative analyses of phospho-proteomic data are less common. Despite tremendous advances in the last decade, a majority of the human phosphoroteome has not been annotated to date[96]. Technical issues such as noise, lower coverage, lower number of samples, and low overlap between studies further complicate the analysis of phospho-proteomic data from a systems biology perspective[79].

In order to facilitate large-scale utilization of phospho-proteomic data, we introduced the notion of co-phosphorylation (Co-P)[4]. The motivation behind this approach is to represent phosphorylation data in the form of relationships between pairs of phosphosites. Defining co-phosphorylation as the correlation between pairs of phosphosites across a range of biological states within a given study, we alleviate such issues as batch effects between different studies and missing identifications, while integrating phosphorylation data across multiple studies. Recently, we applied Co-P to the prediction of kinase-substrate associations)[4] and unsupervised identification of breast cancer subtypes[5], showing that co-P enables effective integration of multiple datasets and enhances the reproducibility of predictions.

Co-phosphorylation is similar in spirit, but distinct and complementary to the notion of co-occurrence[72]. Co-occurence qualitatively assesses the relationship between the identification patterns of phosphosites in a broad range of studies. Co-P, on the other hand, quantitatively assesses the relationship between the phosphorylation levels of sites across a set of biological states (within a single study or by integrating different studies). Thus, co-occurrence captures high-level functional associations among phosphosites, whereas Co-P can also discover context-specific associations and provide insights into the dynamics of signaling interactions.

In this paper, we comprehensively characterize the relationship between co-phosphorylation and functional associations/interactions among protein phosphorylation sites. For this purpose, we systematically compare Co-P networks to networks that represent other functional relationships between proteins and phosphosites. These analyses serve two purposes: (i) Validation of Co-P as a relevant and useful tool for inferring functional relationships between proteins, (ii) Generation of knowledge on the basic principles of post-translational regulation of proteins and the functional relationships between them.

3.2 Materials and Methods

3.2.1 Phospho-Proteomic Datasets

We analyze 9 different MS-based phospho-proteomics data representing cancer and non-cancer diseases.

- **BC1 (Breast Cancer):** Huang et al.[50] used the isobaric tags for relative and absolute quantification (iTRAQ) to identify 56874 phosphosites in 24 breast cancer PDX models.
- **BC2 (Breast Cancer):** This dataset was generated to analyze the effect of delayed cold Ischemia on the stability of phosphoproteins in tumor samples using quantitative LC-MS/MS. The phosphorylation level of the tumor samples was measured across 3 time points[85]. The dataset includes 8150 phosphosites mapping to 3025 phosphoproteins in 18 breast cancer xenografts.
- **BC3 (Breast Cancer):** The NCI Clinical Proteomic Tumor Analysis Consortium (CPTAC) conducted an extensive MS based phosphoproteomics analysis of TCGA breast cancer samples[86]. After selecting the subset of samples to have the highest coverage and filtering the phosphosites with

missing intensity values in those tumors, the remaining data contained intensity values for 11018 phosphosites mapping to 8304 phosphoproteins in 20 tumor samples.

- **OC1 (Ovarian Cancer):** This dataset was generated by the same study as BC2, using the same protocol. The dataset includes 5017 phosphosites corresponding to 2425 phosphoproteins in 12 ovarian tumor samples.
- **OC2 (Ovarian Cancer):** The Clinical Proteomic Tumor Analysis Consortium conducted an extensive MS based phosphoproteomic of ovarian HGSC tumors characterized by The Cancer Genome Atlas[151]. We filtered out the phosphosites with missing data and also selected a subset of tumors to maximize the number of phosphosites. This resulted in a total of 5017 phosphosites from 2425 proteins in 12 tumor samples.
- **CRC (Colorectal Cancer):** Abe et al.[2] performed immobilized metal-ion affinity chromatography-based phosphoproteomics and highly sensitive pY proteomic analyses to obtain data from 4 different colorectal cancer cell line. The dataset included 5357 phosphosites with intensity values cross 12 different conditions. These phosphosites map to 2228 phosphoproteins.
- **LC (Lung Cancer):** Wiredja et al.[137] performed a time course label-free phospho-proteomics on non-small lung cancer cell lines across 1, 6 and 24 hrs after applying two different treatments of PP2A activator and MK-AZD, resulting in total of 6 samples. They reported phosphorylation levels for 5068 phosphosites, which map to 2168 proteins.
- **AD (Alzheimer's Disease):** LC-MS/MS phosphoproteomics was performed on eight individual AD and eight age-matched control postmortem human brain tissues. The dataset contains 5569 phosphosites mapping to 2106 proteins[26].
- **RPE (Retinal Pigmented Epithelium):** MS-based phosphoproteomics was performed on three cultured human-derived RPE-like ARPE-19 cells which were exposed to photoreceptor outer segments (POS) for different

time periods (0, 15, 30, 60, 90, and 120 min)[18]. The dataset contains 1016 phosphosites mapping to 619 proteins in 18 samples.

3.2.2 Functional Networks

To assess the functional relevance of co-phosphorylation, we use networks of functional relationships/associations between pairs of phosphorylation sites. For this purpose, we consider four types of functional networks:

Kinase-Substrate Associations (KSAs). We use PhosphoSitePLUS (PSP)[48] as a gold-standard dataset for kinase-substrate associations. PSP reports 9699 associations among 347 kinases and 6906 substrates. We use these associations to constructed a “shared kinase network" of phosphorylation sites, in which nodes represent phosphosites and edges represent the presence of at least one kinase that phosphorylated both sites. The associations obtained from PSP lead to a shared kinase network of 6906 phosphosite nodes and 881685 shared kinase edges.

Protein-Protein Interaction (PPI). We use the PPIs that are provided in STRING database[127] with high confidence (combined score$\geq$0.95). Overall, there are 61452 high-confidence interactions among 8987 proteins. For each of the 9 datasets, we use these PPIs to construct an interaction network among the sites identified in that dataset. In this network, each node represents a phosphosite and each edge represents an interaction between the two proteins that harbor the respective sites.

Evolutionary and Functional Associations. PTMCode is a database of known and predicted functional associations between phosphorylation and other post-translational modification sites[92]. The associations included in PTMCode are curated from the literature, inferred from residue co-evolution, or are based on the structural distances between phosphosites. We utilize PTMcode as a direct source of functional, evolutionary, and structural associations between phosphorylation sites. In the PTMcode network, there are 96519 phosphosite nodes and 4661285 functional association edges between these phosphosites.

Phosphosite-Specific Signaling Pathways. We use PTMsigDB as a reference database of site-specific phosphorylation signatures of kinases, perturbations, and signaling pathways[66]. While PTMSigDB provides data on all post-translational

Table 3.1. Descriptive statistics of the phospho-proteomic datasets used in our computational experiments and their overlap with functional networks. For each dataset, the number of samples, the number of phosphorylation sites that were identified and the number of proteins that are spanned by these sites are shown. For each dataset and functional network pair, the number in the first row shows the number of sites with at least one interaction in the functional network and the second row shows the number of interactions in the functional network with both sites present in the corresponding dataset.

	Descriptive Statistics			*Overlap with Functional Networks*			
Dataset	**# Samples**	**# Phosphosites**	**# Proteins**	**Shared Kinase**	**PPI**	**PTMCode**	**PTMSigDB**
BC1	24	15780	4539	805	7632	4437	138
				27791	142077	15335	2547
BC2	18	8150	3025	243	1639	1007	54
				2723	16541	1811	429
BC3	20	11472	3312	553	4491	3014	119
				13123	45911	9127	2226
OC1	12	5017	2425	414	2450	1318	74
				7174	17584	2580	1032
OC2	12	4802	2230	157	818	510	32
				1114	4764	685	158
CRC	12	5352	2228	320	1663	1240	51
				6237	17573	2715	421
LC	6	5068	2168	380	2036	1238	64
				6493	17884	2919	588
AD	8	5569	1559	238	1743	941	44
				3637	19075	3182	228
RPE	18	1016	619	120	371	193	31
				931	1667	320	216

modifications, we here use the subset that corresponds to phosphorylation. There area 2398 phosphosites that are associated with 388 different perturbation and signaling pathways. We represent these associations as a binary network of signaling-pathway associations among phosphosites, in which an edge between two phosphosites indicates that the phosphorylation of the two sites is involved in the same pathway. The resulting network contains 6276 edges between 2398 phosphosite nodes. For each functional network, the number of nodes/edges edges that overlap with our 9 phospho-proteomic datasets are shown in Table 3.1.

3.2.3 Assessment of Co-Phosphorylation

For a given phosho-proteomic dataset, we define the vector containing the phosphorylation levels of a phosphosite across a number of biological states as the *phosphorylation profile* of a phosphosite. For a pair of phosphosites, we define the co-phosphorylation of the two sites as the statistical association of their phosphorylation profiles. To assess statistical association, we refer to the rich literature on the assessment of gene co-expression based on mRNA-level gene expression[16], and consider Pearson correlation[7], biweight-midcorrelation[121], and mutual information[87]. Since our experiments suggest that the different measures of association lead to similar results (data not shown), we use Pearson correlation as a simple measure of statistical association in our experiments.

We use the datasets described in the previous section to characterize co-phosphorylation in relation to the functional, structural, and evolutionarily relationships between sites and proteins encoded in the functional networks. For this analysis, we investigate correspondence between co-phosphorylation in each individual MS-based phospho-proteomics dataset and each functional network.

3.2.4 Integration of Co-Phosphorylation Networks Across Datasets

Since co-phosphorylation can potentially capture context-specific, as well as universal functional relationships among phosphorylation sites, we also investigate the functional relevance of co-phosphorylation across different datasets. While integrating co-phosphorylation across multiple datasets, the number of samples (i.e., the number of dimensions used to compute the correlation) in each dataset is different. For this reason, we use the adjusted R-squared[89] (denoted R_d^2) to remove the effect of number of dimensions from dataset-specific co-phosphorylation between pairs of phosphosites:

$$R_d^2(i,j) = 1 - \frac{n_d - 1}{n_d - 2}(1 - c_d(i,j)^2). \tag{3.1}$$

Here, $c_d(i,j)$ denotes the co-phosphorylation (measured by Pearson correlation) in dataset $d \in D$ with n_d samples.

In mass-spectrometry based phospho-proteomics, the overlap between the phosphorylation sites that are identified across different studies is usually low[79]. Specifically, for the 9 datasets we use in our computational experiments, there are only 17 phosphosites that are identified in all studies. Consequently, to preserve the scope of our cross-dataset analysis, we use all sites that are identified in at least one study. For this purpose, we develop a measure of cross-dataset co-phosphorylation that can integrate the co-phosphorylation scores computed on an arbitrary number of datasets. To handle missing data without introducing bias, we set $R_d^2(i,j) = 0$ if phosphosite i or phosphosite j is not present in dataset d. Subsequently, we compute the integrated Co-P between sites i and j as follows:

$$c_{integrated}(i,j) = 1 - \prod_{d \in D} \Big(1 - R_d^2(i,j)\Big). \tag{3.2}$$

Observe that, $0 \leq c_{integrated}(i,j) \leq 1$, where the minimum value is realized if the two sites are never identified in the same dataset or their phosphorylation levels have zero correlation if they are identified together. If the phosphorylation levels of two sites exhibit perfect correlation in at least one dataset, then $c_{integrated} = 1$. Finally, as the number of datasets on which both sites are identified goes up, $c_{integrated}$ also tends to go up. Thus $c_{integrated}$ can be thought of as a measure of both co-occurrence[72] and co-phosphorylation[4], since it captures both the tendency of the sites being identified in similar contexts, as well as the relationship between their dynamic ranges of phosphorylation.

3.3 Results and Discussion

3.3.1 Statistical Significance of Co-phosphorylation

To understand whether the notion of co-phosphorylation (co-P) is biologically relevant, we first investigate the distribution of co-P levels across all pairs of phosphosites identified within a study. The results of this analysis for 9 datasets are shown in Figure 3.1. As seen in the figure, co-P follows a normal distribution with mean close to zero (as would be expected if phosphorylation levels were drawn from a normal distribution) and the distribution is narrower (and likely less noisy) if more biological states (dimensions) are available. Based on the premise that co-P can capture functionally relevant relationships, we hypothesize that distribution of co-phosphorylation on real datasets would contain more positively and negatively correlated phosphosite pairs than would be expected at random. To test this hypothesis, we conduct permutation tests by permuting phosphorylation levels across the entire data matrix, and compute the co-P distribution on these randomized datasets. As seen in the figure, co-P is concentrated more on strongly positive or strongly negative correlation levels for all datasets. For all datasets, the Kolmogorov-Smirnov (KS) test p-values for the difference between the observed co-P distribution and permuted co-P of distribution are $<< 1E-9$. Similarly, the t-test p-values for the difference between the means of these distributions are $<< 1E-9$ for all datasets except CRC. The mean difference and the 95% confidence interval for each dataset are provided below the histograms in the figure.

Furthermore, for most datasets (BC2, BC3, OC1), we observe that the mean co-P is clearly shifted to the right, as also indicated by the effect size and the significance of the t-statistic.. For other datasets (BC1, CRC), the difference between the means is close to zero and the corresponding t-statistics are less significant. However, even for these datasets, the KS-test indicates that the difference between the distributions is significantl, and visual inspection of the historgrams suggests that the histogram for observed Co-P values is always more spread. This observation suggests that these datasets also contain a large number of site pairs with negatively correlated phosphorylation levels. Clearly, as with positive correlation, negative correlation can also be indicative of a functional relationship between two phosphorylation sites

Taken together, for all studies considered, there are more pairs of phosphosites with (positively or negatively) correlated phosphorylation levels than would expected at random – hence a large fraction of these strong correlations likely stem from functional or structural relationships between the phosphosites.

3.3.2 Co-Phosphorylation of Intra-Protein Sites

Results of previous studies indicate that the phosphorylation of different sites of the same protein can lead to different functional outcomes[100,101]. Here, with a view to characterizing the functional diversity of the phosphorylation sites on a single protein, we compare the Co-P distribution of pairs of phosphosites that reside on the same protein (intra-protein sites) against the Co-P distribution of pairs of phosphosites that reside on different proteins (inter-protein sites). We also

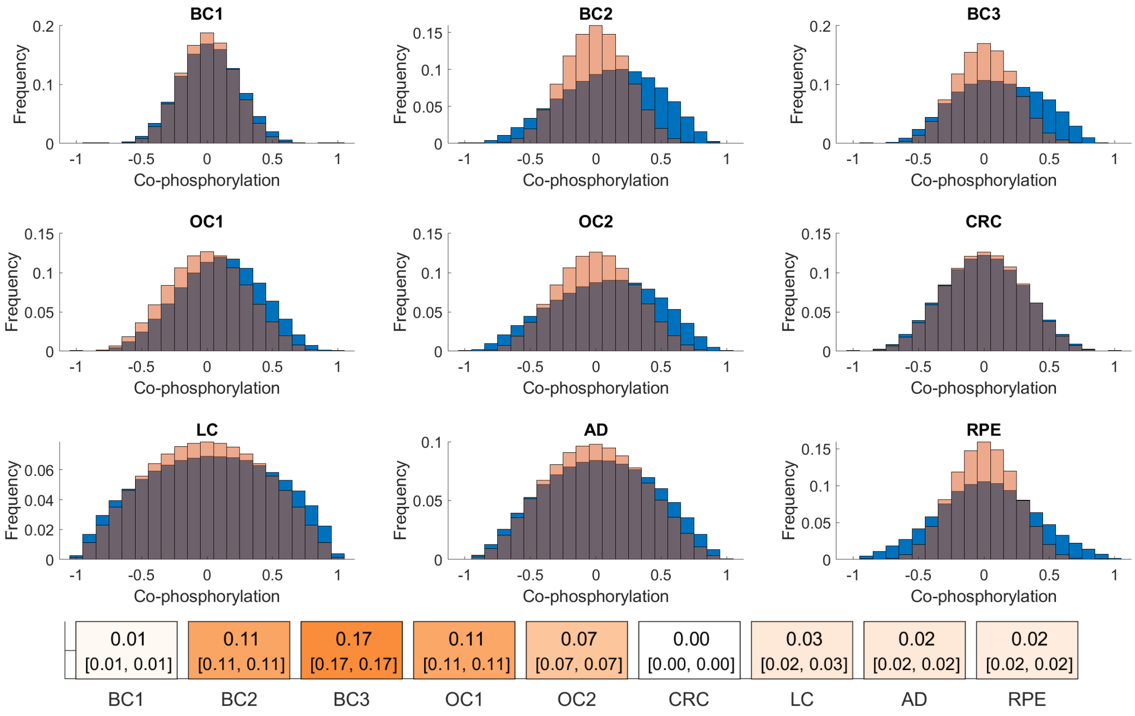


Figure 3.1. Statistical significance of co-phosphorylation. Each panel compares the distribution of co-phosphorylation computed on a specific dataset against that computed on randomly permuted data for each dataset. The blue histogram shows the distribution of co-phosphorylation (the correlation between the phosphorylation levels) of all pairs of phosphosites identified in the corresponding study, the pink histogram in each panel shows the average distribution of co-phosphorylation of all pairs of phosphosites across 100 permutation tests. The permutation tests are performed by randomly permuting all entries in the phosphorylation matrix. The difference between the means of each pair of distributions is given on the colored boxes below. The 95% confidence intervals for the difference are provided in brackets.

investigate the effect of proximity between phosphorylation sites on the functional relationship between the sites. The results of this analysis are shown in Figure 3.2.

As seen in Figure 3.2(a), the distribution of co-P for pairs of intra- and inter-protein sites are significantly different for most of the datasets (the mean differences and confidence intervals are provided in the figure, the p-values for the t-test as well as the KS-test are $<< 1E-9$ for all datasets except RPE). We consistently observe that the co-phosphorylation of intra-protein sites (orange histogram) is shifted towards high co-phosphorylation values. In other words,the phosphorylation levels of sites on the same protein are substantially more positively correlated as compared to the phosphorylation levels of sites on different proteins. While this observation can be partially explained by the impact of protein expression levels, a recent study showed that the protein abundance is overall not a strong indicator of phosphorylation fold-changes[3]. Thus, we hypothesize that intra-proteins pairs exhibit higher co-phosphorylation because those pairs are more likely to be targeted by the same kinase/phosphates, or that they are more likely to be functionally associated by being part of the same signaling pathways.

Note that, the differences between the datasets in terms of the difference of intra- and inter-protein pairs are highly prononunced (e.g., we observe strong difference for BC1, BC3, OC1 while difference is more modest for BC2, CRC, and AD). While there can be biological reasons for this difference, it is important to note that each of these datasets come from different platforms, different sample types (e.g., patient-derived xenografts vs. cell lines), different data collection procedures (e.g., protein degradation due to proteases in the sample), and are highly divergent interms of availability of data (number of identified sites and number of samples). For this reason, the observed differences between the datasets can also be attributed to experimental, technological, or statistical reasons. Further investigation is needed to elucidate potential biological differences between the systems that are represented by these datasets.

Next, we investigate whether the proximity on the protein sequence has any effect on the co-phosphorylation between two intra-protein sites. Since previous studies suggest that closely positioned sites tend to be phosphorylated by the same kinase[118], we expect a positive relation between sequence proximity and co-phosphorylation (i.e., we expect higher co-phosphorylation between close sites). To investigate this, we plot the relationship between the sequence proximity of intra-protein sites, and their co-phosphorylation (Co-P). Figure 3.2(b) shows that the closely positioned intra-protein sites have higher Co-P. Thus, we observe that as the phosphosites get far away from each other, their Co-P typically reduces.

3.3.3 Co-phosphorylation and Functional Association

Li et al.[72] show that phosphorylation sites that are modified together tend to participate in similar biological processes. Here, focusing on the dynamic range of phosphorylation, we hypothesize that phosphosite pairs with correlated phosphorylation profiles are likely to be functionally associated with each other. To test this hypothesis, we investigate the relationship between Co-P and a broad range of functional associations. Since our results in Figure 3.2 suggest that there is a considerable difference between intra-protein and inter-protein sites in terms of their co-phosphorylation, we perform stratified analyses for intra- and inter-protein pairs. The results of this analysis are shown in Figure 3.3.

Shared-Kinase Pairs. First, we consider the Co-P of the substrates of the same kinase (i.e., shared-kinase pairs) as annotated by PhosphositePlus. As seen in the Figure 3.3, in all datasets, the Co-P distribution of shared-kinase pairs is significantly shifted upwards, i.e., sites that are targeted by the same kinase are likely to exihibit stronger correlation of phosphorylation as compared to arbitrary pairs. While this difference is more pronounced for intra-protein pairs, it is also evident for inter-protein pairs. This observation is also in line with previous findings in the literature[3,4].

Phosphorylation Sites on Interacting Proteins. Protein-Protein Interaction networks (PPI) encode physical and functional associations among proteins, thus have been used commonly for various inference tasks in cellular signaling. These tasks include identification of signaling pathways[133], identification of pathways that are mutated in cancers[116], prediction of the effect of mutations on protein interactions[115], and prediction of kinase-substrate associations[46]. It is also well-established that proteins that are coded by co-expressed genes are likely to interact with each other[111]. Here, we compare the PPI network and Co-P network to investigate the pattern of Co-P of pairs of phosphosites on interacting proteins. Note that, by definition, we only have this type of functional interaction for inter-protein sites. As seen in the Figure 3.3, in most of the datasets we consider (including BC1, BC3, OC1, OC2, LC, RPE), there is a clear upward shift of co-P for sites on interacting proteins. This suggests that sites on interacting proteins are likely to be co-phosphorylated. Identification of the specific protein-protein interactions (PPIs) that are associated with co-phosphorylation can be potentially useful in elucidating the mechanisms of these PPIs.

Co-evolution of Phosphorylation Sites. The conservation status of the phosphosites has been used as a tool to measure PTM activity[12]. It has been shown that co-evolving PTMs are likely to be functionally associated[91]. Here, we investigate the relationship between co-evolution and co-phosphorylation of phosphosites. The results of this analysis are shown in Figure 3.3. As seen in the figure, the association between co-evolution and co-phosphorylation is relatively weak compared to the association of co-P with other functional networks.

Phosphorylation Sites with Common Signaling Pathways. Identifying the signaling pathways that are dysregulated in any perturbation and disease is crucial for understanding the underlying mechanism of diseases. While most databases for signaling pathways are limited to gene or protein-centric information, PTM-sigDB[66] provides a collection of PTM site-specific signatures that have been assembled and curated from public datasets. Using PTMsigDB, we investigate the Co-P of phosphosites that are involved in the same pathway. As seen in Figure 3.3, there is considerable difference between the Co-P distribution of the phosphosites that are involved in the same signaling pathway as compated to that of other phosphosite pairs. Similar to the results for shared-kinase pairs, this difference is more pronounced for intra-protein sites.

3.3.4 Predictive Power of Co-phosphorylation

Our results indicate that phosphosites involved in a common pathway or targeted by a common kinase are likely to be co-phosphorylated across different biological states. Motivated by this observation, we quantitatively assess the effectiveness of Co-P in predicting shared-kinase and shared-pathway associations between phosphorylation sites. While doing so, we also assess the contribution of Co-P evidence supported by multiple datasets to the reliability of predictions on functional association. For this purpose, we assess the predictive ability of Co-P computed using each individual dataset as well as the integrated Co-P computed using cross-dataset analysis. The results of this analysis are shown in Figure 3.4.

While constructing the co-P networks, we compute a co-P score for each pair of phosphosites, namely $c_d(i, j)$ for individual dataset d and $c_{integrated}(i, j)$ for the integrated network. We then sort the pairs according to this co-P score and apply a moving threshold to generate a series of co-P networks with increasing number of edges. In the left panel of Figure 3.4, the precision-recall curves for the ability of this network in predicting shared-kinase interactions (top-left panel) and shared-pathway interactions (bottom-left panel) are shown. In this context, recall is the defined as the fraction of edges in the corresponding functional network that also exist in the co-P network, whereas precision is defined as the fraction of edges in the co-P network that also exist in the functional network. To provide a baseline for the predictive ability of the co-P network, we also visualize the mean precision and 95% confidence interval for given recall for a random ranking of phosphosite pairs across 20 runs. As seen in the figure, the precision provided by the co-P network is significantly higher than random ordering for both functional networks. We also observe that co-P delivers higher precision for the shared-pathway network as compared to the shared-kinase network. This is likely because the information in PTMSigDB is sparser than the information in PhosphositePLUS.

The right panel of Figure 3.4 shows the odds ratio of a pair of sites being connected in the functional network as a function of the number of edges in the co-P network. Namely, in these plots, a point on the x axis corresponds to a co-P network with a given number of edges. For this network, the value on the y-axis shows the odds ratio of the event that two sites are connected in the functional network given that they are connected in the co-P network, as compared to a random pair of sites. As seen in the figure, for both shared-kinase and shared-pathway networks, the odds-ratio provided by the integrated co-P network is consistently higher than that provided by any individual network. While the odds-ratio of sharing a kinase goes up to 100 and the odds-ratio of being involved in the same pathway goes up to 30 for pairs of sites with co-P, these odds-ratios respectively converge to 4 and

2 as more edges are added to the integrated co-P network. Overall, these results suggest that co-P networks provide valuable information on the functional association of phosphorylation sites and this information becomes more reliable as co-P information from more datasets are included in the co-P network.

3.4 Conclusion

Mass-spectrometry techniques are advancing and more MS-based quantitative phosphoproteomics data are generated at high volumes. However, integration of these data may be challenging since the data is generated in different labs and in different contexts. By focusing on the relationships between pairs of phosphosites as opposed to their individual phosphorylation levels, co-phosphorylation networks can alleviate the dependency of computational and statistical methods on these factors. In this paper, we systematically investigated the relationship between co-phosphorylation and broad range of known functional associations between proteins and phosphorylation sites. Our results showed that the sites that are functionally associated tend to exhibit higher levels of co-phosphorylation. Our results also showed that the integration of co-phosphorylation networks across different datasets can improve the predictivity of co-phosphorylation, as compared to analyzing the datasets in isolation. These results highlight the power of network models and network-based analyses of phosphorylation data in predicting the functional relationships among phospho-proteins, kinases, and phosphatases in the context of cellular signaling.

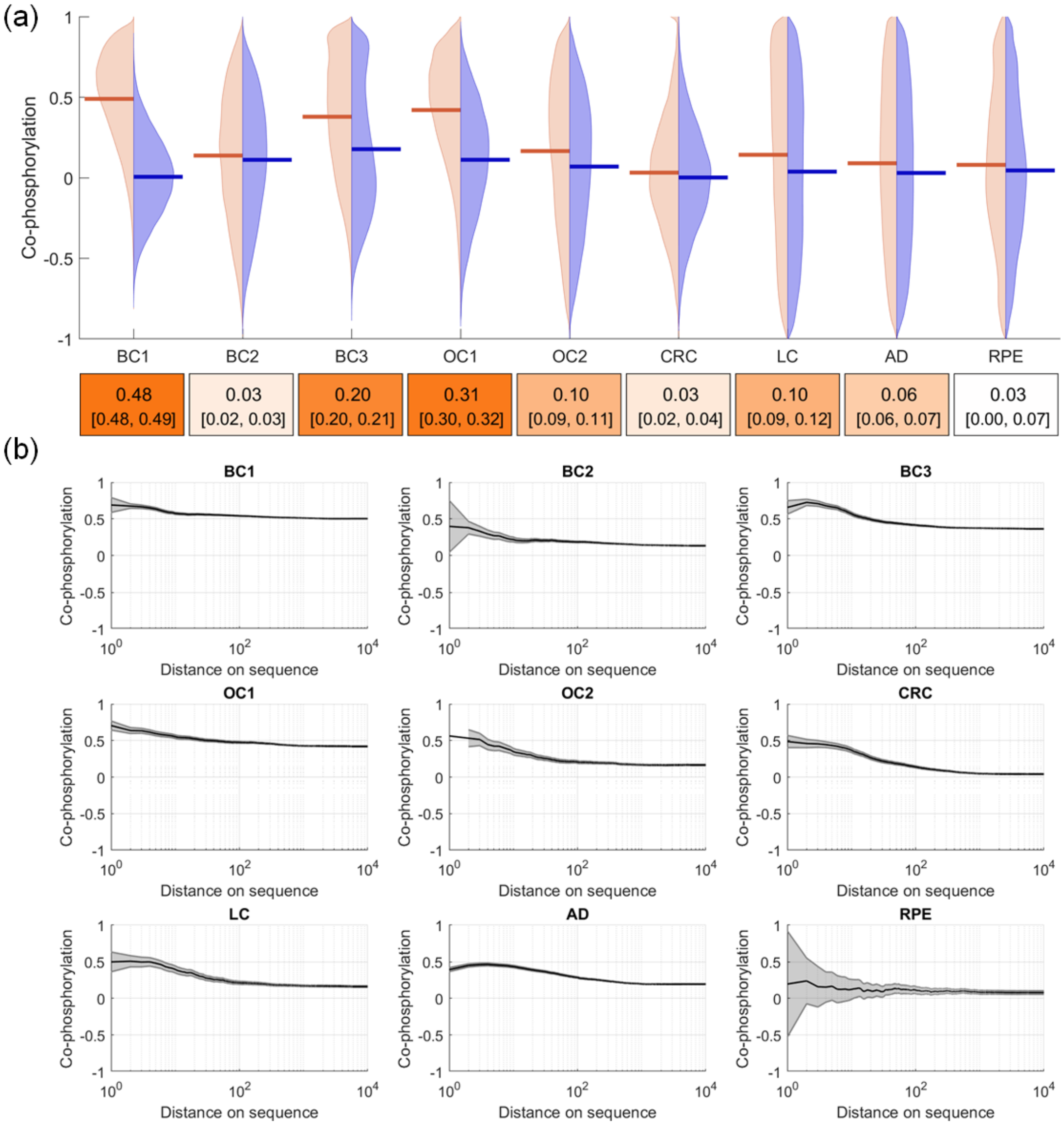
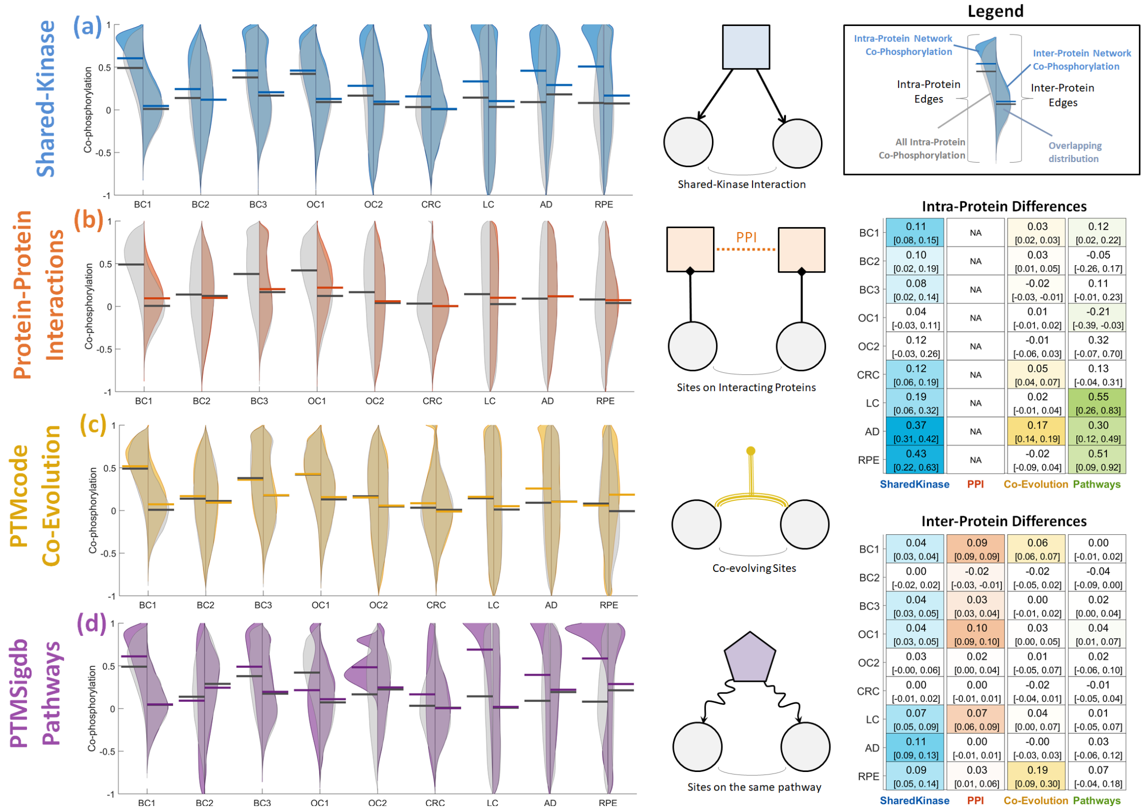


Figure 3.2. Co-phosphorylation of phosphorylation sites on the same protein. (a) Comparison of the distribution of Co-P for all site pairs that are on the same protein (orange histogram) vs. co-P for all pairs of sites on different proteins (blue histogram). Each violin plot represents a different dataset. Colored boxes below indicate the mean difference between the intra-protein pairs and inter-protein pairs. Within brackets, 95% confidence interval for the mean Co-P difference are provided. (b) The relationship between co-P and sequence proximity for pairs of sites that reside on the same protein. Each panel shows a different dataset, the x-axis in each panel shows the distance between sites on the protein sequence (in terms of number of residues) and the y-axis shows the co-phosphorylation between pairs of sites in close proximity (up to the corresponding distance in x-axis). The curve and shaded area respectively show the mean Co-P and its 95% confidence interval.



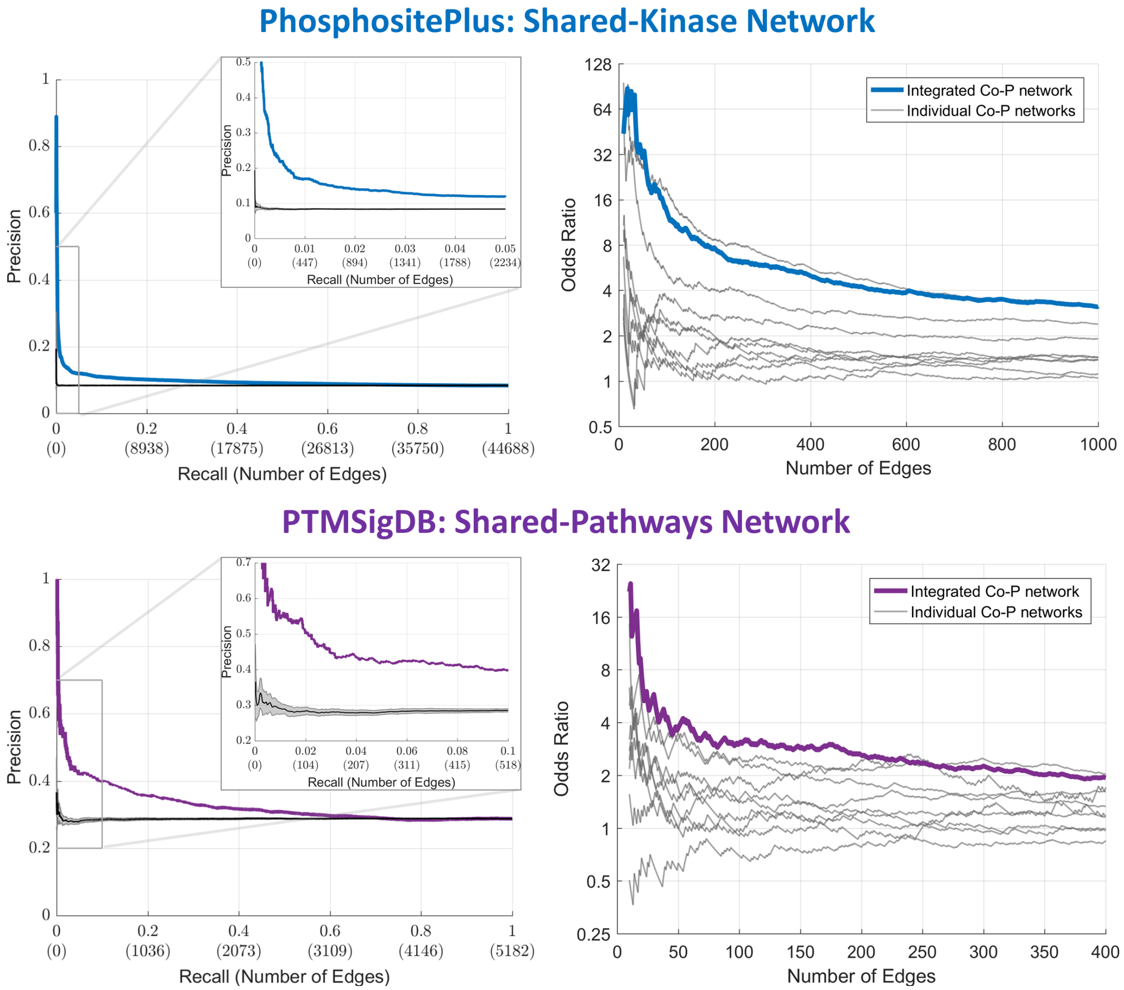
Intra-Protein Differences

	SharedKinase	PPI	Co-Evolution	Pathways
BC1	0.11 [0.08, 0.15]	NA	0.03 [0.02, 0.03]	0.12 [0.02, 0.22]
BC2	0.10 [0.02, 0.19]	NA	0.03 [0.01, 0.05]	-0.05 [-0.26, 0.17]
BC3	0.08 [0.02, 0.14]	NA	-0.02 [-0.03, -0.01]	0.11 [-0.01, 0.23]
OC1	0.04 [-0.03, 0.11]	NA	0.01 [-0.01, 0.02]	-0.21 [-0.39, -0.03]
OC2	0.12 [-0.03, 0.26]	NA	-0.01 [-0.06, 0.03]	0.32 [-0.07, 0.70]
CRC	0.12 [0.06, 0.19]	NA	0.05 [0.04, 0.07]	0.13 [-0.04, 0.31]
LC	0.19 [0.06, 0.32]	NA	0.02 [-0.01, 0.04]	0.55 [0.26, 0.83]
AD	0.37 [0.31, 0.42]	NA	0.17 [0.14, 0.19]	0.30 [0.12, 0.49]
RPE	0.43 [0.22, 0.63]	NA	-0.02 [-0.09, 0.04]	0.51 [0.09, 0.92]

Inter-Protein Differences

	SharedKinase	PPI	Co-Evolution	Pathways
BC1	0.04 [0.03, 0.04]	0.09 [0.09, 0.09]	0.06 [0.06, 0.07]	0.00 [-0.01, 0.02]
BC2	0.00 [-0.02, 0.02]	-0.02 [-0.03, -0.01]	-0.02 [-0.05, 0.02]	-0.04 [-0.09, 0.00]
BC3	0.04 [0.03, 0.05]	0.03 [0.03, 0.04]	0.00 [-0.01, 0.02]	0.02 [0.00, 0.04]
OC1	0.04 [0.03, 0.05]	0.10 [0.09, 0.10]	0.03 [0.00, 0.05]	0.04 [0.01, 0.07]
OC2	0.03 [-0.00, 0.06]	0.02 [0.00, 0.04]	0.01 [-0.05, 0.07]	0.02 [-0.06, 0.10]
CRC	0.00 [-0.01, 0.02]	0.00 [-0.01, 0.01]	-0.02 [-0.04, 0.01]	-0.01 [-0.05, 0.04]
LC	0.07 [0.05, 0.09]	0.07 [0.06, 0.09]	0.04 [0.00, 0.07]	0.01 [-0.05, 0.07]
AD	0.11 [0.09, 0.13]	0.00 [-0.01, 0.01]	-0.00 [-0.03, 0.03]	0.03 [-0.06, 0.12]
RPE	0.09 [0.05, 0.14]	0.03 [0.01, 0.06]	0.19 [0.09, 0.30]	0.07 [-0.04, 0.18]

Figure 3.3. The relationship between co-phosphorylation and functional association between pairs of phosphorylation sites. In each panel, the violin plots compares the distribution of co-P for phosphosite pairs with an edge in the respective functional association network (colored histograms) against all phosphosite pairs (gray-colored histograms), across the 9 datasets that are considered. For each dataset, the left/right violin plots respectively show intra-/inter-protein pairs. The black horizontal lines show the mean Co-P for all (intra- or inter-protein) phosphosite pairs, the colored horizontal lines show the mean Co-P for functionally associated pairs. The four type of functional association networks that are considered are illustrated on the right side of the corresponding violin plot. On the rightmost side, the colored tables show the mean difference between functionally associated pairs and all phosphosite pairs (corresponding to the gap between colored and black horizontal lines in the violinplots) for 9 datasets and 4 functional networks. In each cell, the 95% confidence intervals for the mean difference is given within brackets.



PhosphositePlus: Shared-Kinase Network

PTMSigDB: Shared-Pathways Network

Figure 3.4. The utility of co-phosphorylation in predicting the functional association of phosphorylation sites. (Left) Precision-Recall curve showing the functional predictivity of the Co-P network obtained by integrating 9 different phospho-proteomic datasets. The shaded gray area shows the 95% confidence interval for the mean precision-recall curve for permutation tests obtained by randomly ranking pairs of phosphosites (across 20 runs). (Right) Comparison of the predictive performance of the integrated Co-P network against the 9 individual Co-P networks obtained using each dataset separately. The x-axis shows the number of pairs that are included in the co-P network, the y-axis shows the odds ratio of being connected in the respective functional network given that the sites are connected in the co-P network. (Top) Predicting shared-kinase associations. (Bottom) Predicting shared-pathway associations.

4 Are under-studied proteins under-represented? How to fairly evaluate link prediction algorithms for biomedical applications

4.1 Introduction

Background and related literature. In the context of network biology, link prediction is commonly applied to discover previously unknown associations or interactions[148]. Many biomedical prediction tasks are formulated as link prediction problems, including prediction of drug–disease associations (DDAs)[73], drug response prediction[122], disease gene prioritization[37], prediction of drug-drug interactions (DDI)[152], protein-protein interactions (PPIs)[62], transcription factor regulatory relationships[71], kinase-substrate associations[4], and kinase-kinase interactions[51].

Early research on link prediction focused on computing a score to assess the likelihood of the existence of an edge between two nodes[155]. These include local measures based on guilt-by-association, including common neighbors and preferential attachment[74]. Global approaches, such as random walks, generalize this principle to the notion that nodes that are "proximate" are likely to acquire an edge[70]. More recently, graph embedding models, which map each node to a vector in a lower-dimensional embedding space, allow machine learning methods to be utilized seamlessly in link prediction[41,107]. With the availability of various types of omic data, along with rapid advances in machine learning, more sophisticated learning algorithms, including graph convolutional networks, are increasingly applied to link prediction problems in systems biology[31,123,147].

Evaluation of link prediction algorithms. For evaluating link prediction algorithms, a recommended strategy is to perform the validation on an independent test dataset, using different snapshots of the network (e.g., taken from different data sources or different points in time) as training and test sets[82,142]. In the absence of multiple snapshots, the evaluation is typically performed by generating training and test instances from a single network, sampling the edges to be removed from the network uniformly at random[39,41,69]. With the availability of more link prediction algorithms with ever-increasing sophistication, research on the evaluation of algorithms has also gained attention[82,98,148]. Although there has been significant attention to algorithmic bias and fairness[84] as well as the reproducibility and comparability of the results[81] in graph machine learning, studies investigating fairness and sources of bias in the evaluation of link prediction algorithms are relatively scarce, particularly in the context of network biology[36].

Motivation and significance in systems biology. For biological knowledge discovery, *fairness* can be considered as the ability to identify biological entities that are relatively less studied (e.g., when a scientist is looking to identify a kinase that phosphorylates a specific phosphorylation site they discovered, does the algorithm give equal consideration to all kinases regardless of how well-studied they are?). Matthew's effect (also known as "rich gets richer") is quite pronounced in biology - according to the Understudied Protein Initative that was announced in May 2022[68], "95% of all life science publications focus on a group of 5,000 particularly well-studied human proteins". This effect is also a critical source of *bias* during the evaluation of link prediction algorithms in biology.

We[37,145] and other groups[36] have documented the degree bias in biological networks and its consequences in the context of specific applications in network biology. However, little attention is paid to the effect of bias in evaluating new link prediction algorithms, leading to the development of algorithms that continuously reinforce what is already known about well-studied proteins[54]. In this paper, we show that *both the benchmarking data and standard evaluation techniques for link prediction favor well-studied biological entities.* Specifically, we demonstrate that (i) randomly sampling edges to generate a test set creates bias in which edges that connect high-degree nodes are over-represented, (ii) this bias also exists in settings that utilize different snapshots of a network as training/test sets as opposed to a randomized sampling. In turn, link prediction algorithms that make biased predictions are disproportionately rewarded for favoring high-degree nodes. This results in a serious barrier in making new discoveries involving under-represented biological entities.

Contributions of this study. We argue that successful prediction of interactions and associations that involve low-degree nodes can be more valuable as they can offer more insight about the biological mechanisms under study[68]. Therefore, the evaluation of a link-prediction algorithm in biology needs to account for degree bias throughout analysis. Motivated by this consideration, using prediction of protein-protein interactions (PPIs) as a case example, we first investigate the typical evaluation settings used in the literature. We demonstrate how the current evaluation settings incentivize algorithms to bring forward well-studied proteins in their predictions. To address this issue and faciliate the bias-aware evaluation of link prediction algorithms, we propose multiple strategies organized in five views: (i) quantifying the bias in predictions, (ii) quantifying bias in benchmarking data (and the incentive toward high-bias predictors), (iii) a weighted validation setting that aims to ensure that under-studied proteins are not under-represented in the evaluation, (iv) a stratified analysis that decomposes the prediction performance based on how well-studied the nodes are, and (v) a summary view to outline the main characteristics of an algorithm.

Finally, we survey additional problems to show that the issues we demonstrate in the context of PPI prediction generalize to other link prediction problems in biology: 1) kinase-substrate associations, 2) transcription factor-target interactions, 3) drug-drug interactions, 4) drug-disease associations. These results suggest that, for a broad range of problems in network biology, under-studied entities are severely under-represented in traditional evaluation settings. The proposed framework can be helpful to perform a balanced evaluation, facilitating the development of algorithms focusing on novel findings and new interactions between under-studied biological entities.

4.2 Results

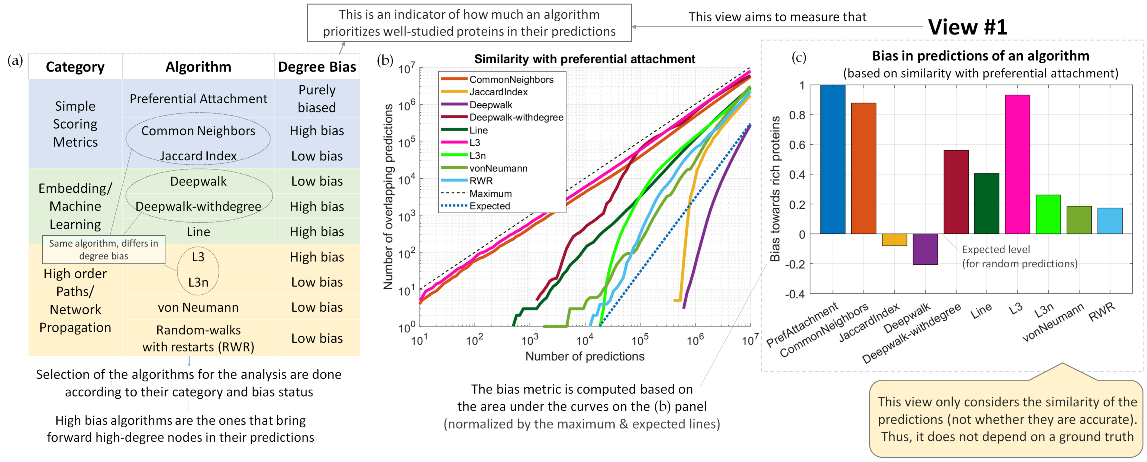
Experimental setting. In this work, to bring to light some issues in standard evaluation settings that are a result of a severe imbalance in the gathered knowledge for biological networks, and to demonstrate the strategies we propose to resolve these issues, we primarily focus on the problem of PPI predictions and the human PPI network obtained from Biogrid[104]. As a final part of our analysis, we analyze a broad range of networks and problems in the context of biomedical applications to show that our observations on PPI network are generalizable to other domains. For simplicity, we mainly consider and refer to the nodes in a protein context, although the developed techniques are not specific to proteins or PPI network. For the evaluation, we obtain the required training/test sets either by random sampling of the edges or by utilizing multiple versions of the Biogrid network (taken at 2020 and 2022). Link prediction algorithms use these training portions of the network to produce prediction scores for pairs of nodes and to obtain a ranking for pairs that are most likely to have an interaction between. These rankings are then compared against the known interactions in the test set to evaluate the prediction performance of the algorithms.

Selected algorithms for the analysis. To select the link prediction algorithms (Figure 4.1(a)) for inclusion in our analysis, we consider two criteria: (i) to include representative methods for different classes of algorithms (e.g., scoring metrics, network propagation methods, embedding/machine learning based methods), and (ii) to include algorithms with differing levels of bias towards high-degree nodes. Namely, we include at least two versions from each category: One version that prioritizes high-degree nodes (high-bias methods) and another, normalized version with lower bias. For example, in the scoring metrics category, common neighbors is the high-bias version, whereas Jaccard index (intersection divided by union) is the normalized, lower-bias version. In both cases, the information source is the same (number of shared interactions), the only difference is the normalization based on node degrees, and therefore, disposition of the method toward high-degree nodes. For Deepwalk[107], a low-bias embedding algorithm that generates and uses feature vectors with a low correlation to the node degrees (S. Figure 4.7), we create a biased version (Deepwalk-withdegree) by adding node degree to the embeddings as an additional feature. Similarly, since L3[62] is a high bias algorithm (as it counts the paths of length 3 and only applies a soft normalization), we create a lower-bias version, L3n, by applying a stronger normalization based on node degrees. Besides these, we also consider preferential attachment (a purely-biased baseline that only considers node-degree information), LINE[130] a neural-network based embedding algorithm with high bias (since its learning process captures the node degree information in the embeddings, S. Figure 4.8), two network propagation algorithms von Neumann[67] and random walks with restarts (RWR)[131] (both with low-bias due to strong normalization based on node degrees in their formulation). For embedding algorithms, we train a logistic regression model to obtain the predictions.

Note that, our aim for the analysis (and the selection of algorithms) is not to determine the best performing method among the state-of-the-art methods for PPI prediction problem. Instead, our aim is to (i) investigate the benchmark data and evaluation process as a function of degree distribution, elucidating the effect of the imbalance in the network on commonly used evaluation settings, and (ii) demonstrate how these evaluation settings can reward the development of algorithms that are biased toward high-degree proteins, which often correspond to well-studied proteins[40].

View #1: Bias of link prediction algorithms toward high-degree proteins

- **Proposed strategy:** To understand the disposition of an algorithm toward well-studied proteins, measure its similarity with preferential attachment (biased baseline).



This is an indicator of how much an algorithm prioritizes well-studied proteins in their predictions

This view aims to measure that

View #1

(a)

Category	Algorithm	Degree Bias
Simple Scoring Metrics	Preferential Attachment	Purely biased
	Common Neighbors	High bias
	Jaccard Index	Low bias
Embedding/ Machine Learning	Deepwalk	Low bias
	Deepwalk-withdegree	High bias
	Line	High bias
High order Paths/ Network Propagation	L3	High bias
	L3n	Low bias
	von Neumann	Low bias
	Random-walks with restarts (RWR)	Low bias

Same algorithm, differs in degree bias

Selection of the algorithms for the analysis are done according to their category and bias status

High bias algorithms are the ones that bring forward high-degree nodes in their predictions

The bias metric is computed based on the area under the curves on the (b) panel (normalized by the maximum & expected lines)

This view only considers the similarity of the predictions (not whether they are accurate). Thus, it does not depend on a ground truth

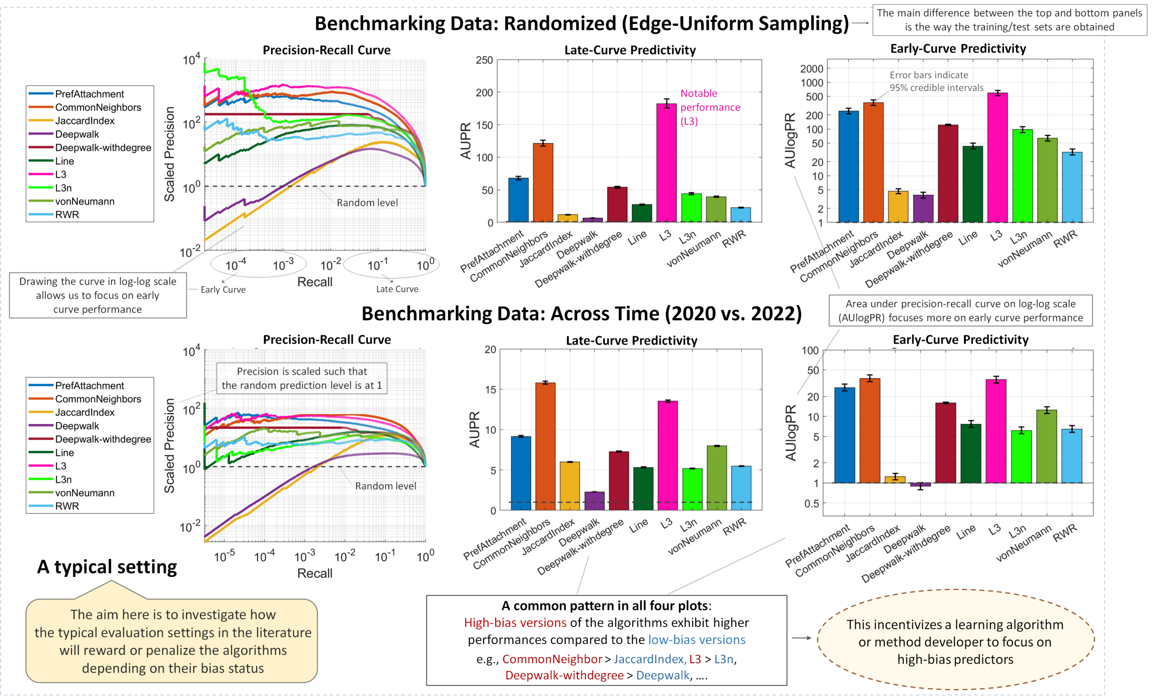
Figure 4.1. Quantifying the bias towards well-studied proteins in the predictions of an algorithm based on similarity with preferential attachment model on Biogrid PPI network. (a) The algorithms selected for the analysis, their categorization and affinities towards degree bias. (b) Overlap of the predictions of the algorithms with preferential attachment. (c) The quantified bias of the algorithms.

- Provides information about how much an algorithm prioritizes high-degree nodes. Node degree is considered an indicator of how well-studied a protein is.

Here, we aim to investigate and quantify the bias in the predictions of the algorithms toward well-studied proteins. For this purpose, we use preferential attachment as a *biased baseline* (since it scores pairs of nodes by multiplying their degrees) and quantify the similarity in the predictions of the algorithms with that of preferential attachment by measuring the overlaps for k predictions (for varying k, Figure 4.1(b)). To obtain a normalized score (where +1/0/-1 indicates bias towards high degrees/no bias/anti-bias towards low degrees), we compute the area under these functions for each algorithm in log-log scale (so that top predictions are given emphasis) and normalize the area according to the maximum possible overlap (k) and the expected overlap (for random predictions). The results of this analysis (Figure 4.1(c)) are mostly as expected: Common neighbors and L3 exhibit the highest bias, followed by Deepwalk-withdegree and Line. Other algorithms exhibit relatively lower bias, while Jaccard index and Deepwalk are slightly biased toward low-degree nodes.

Standard settings for evaluating link prediction algorithms

Here, our aim is to investigate the standard evaluation settings and demonstrate how they can favor bias in predictions towards well-studied entities and how this can lead to conclusions that are counter-productive to the goals of algorithm development in the context network biology and proteomics[68]. For this purpose, following the recommendations of the machine learning community[82,142], we consider two ways to generate train/test splits: (i) Edge-Uniform: We randomly sample the edges in the network (in 2020 version) uniformly at random and include 10% of the edges in the test set and use the remaining as the training set (ii) Across-Time: We use a more recent, 2022 snapshot of the network as the test set and the older 2020 version as the training set.



Benchmarking Data: Randomized (Edge-Uniform Sampling)

The main difference between the top and bottom panels is the way the training/test sets are obtained

Precision-Recall Curve

Late-Curve Predictivity

Early-Curve Predictivity

Notable performance (L3)

Error bars indicate 95% credible intervals

Drawing the curve in log-log scale allows us to focus on early curve performance

Benchmarking Data: Across Time (2020 vs. 2022)

Area under precision-recall curve on log-log scale (AUlogPR) focuses more on early curve performance

Precision-Recall Curve

Precision is scaled such that the random prediction level is at 1

Late-Curve Predictivity

Early-Curve Predictivity

A typical setting

The aim here is to investigate how the typical evaluation settings in the literature will reward or penalize the algorithms depending on their bias status

A common pattern in all four plots:
High-bias versions of the algorithms exhibit higher performances compared to the low-bias versions
e.g., CommonNeighbor > JaccardIndex, L3 > L3n, Deepwalk-withdegree > Deepwalk, ...

This incentivizes a learning algorithm or method developer to focus on high-bias predictors

Figure 4.2. Results of a typical evaluation setting investigating the prediction performance of the link prediction algorithms in the context of PPI predictions. Two types of benchmarking data is considered: (Top panel) Randomized edge-uniform sampling and (Bottom panel) different snapshots of the network across time are used to generate the training and test instances.

In Figure 4.2, the precision-recall curve is shown for all algorithms for both benchmarking datasets. Here, precision is scaled so that the precision of random prediction is 1 (e.g., an algorithm having a scaled precision of 100 indicates $100\times$ more precise predictions compared to random). We also compute two metrics: The area under precision-recall curve (AUPR) and area under precision-recall curve in log-log scale (AUlogPR). Since link prediction problems involve a large background set (i.e., possible node pairs for n nodes is $\Theta(n^2)$), even 10% recall corresponds to a very high number of predictions (in the order of $\approx 10^5/10^6$ for edge-uniform/across-time data; S. Figure 3). Thus, AUPR in linear scale, whose more than 90% of effective region consists of high number of predictions ($> 10^5$), can be considered a measure of *late curve predictivity*. AUlogPR, on the other hand, puts more emphasis on lower recall values (in logarithmic intervals) corresponding to lower number of predictions, thus providing a measure of *early curve predictivity*. While other metrics like early precision are used in recent literature to evaluate early curve predictivity[109], an advantage of AUlogPR over early precision is that it does not require a fixed threshold that defines “early".

As seen in Figure 4.2, the algorithms that are biased toward high-degree nodes seem to outperform other algorithms according to this evaluation setting, where the high-bias versions of the algorithms (CommonNeighbors, L3, Deepwalk-withdegree) exhibit considerably higher prediction performance compared to their low-bias versions (JaccardIndex, L3n, Deepwalk). The differences based on the degree bias seem more pronounced in the early curve (AUlogPR), i.e., the algorithms are typically penalized more strictly if they rank low-degree node pairs higher in their predictions. Overall, these results show that the standard evaluation settings for evaluating link prediction algorithms can incentivize an algorithm or method developer to focus on high-bias predictors that bring forward well-studied biological entities at the expense of the under-studied ones.

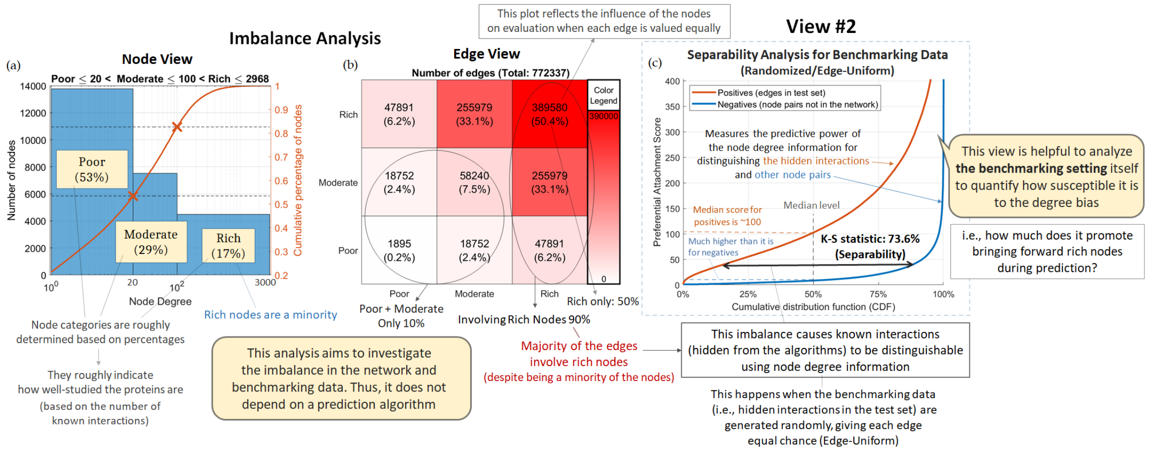
View #2: Bias in benchmarking data and evaluation framework

- **Aim:** To assess bias in an evaluation setting (e.g., training/test sets) in terms of the incentive it provides toward high-bias predictors (that prioritize well-studied entities).
- **Proposed strategy:** Measure the informedness of node degree information in distinguishing the positives from negatives in the test set, using preferential attachment as a representative model for node degree information.

Having observed that the standard evaluation setting favors algorithms that are biased toward high-degree nodes, we next aim to understand the reasons that underlie this observation. For this purpose, we assess the imbalance in a given network or benchmarking data (i.e., train/test splits) in terms of the degree distribution and quantify the predictive power provided by this imbalance for separating the "positives" (known interactions hidden from the algorithms) from the "negatives" (set of possible node pairs without a known interaction). For this purpose, we start by categorizing the nodes based on their connectivity in the PPI network (Figure 4.3(a)): Poor nodes ($<= 20$ interactions), Moderate nodes (degree between 20 and 100), and Rich nodes (> 100 interactions). Note that, we assign these categories by considering the cumulative degree distribution so that Poor and Rich nodes roughly comprise 50% and 15% of all nodes in the network.

Once nodes are categorized into three groups, we categorize the interactions in the network into nine (3x3) groups involving all possible combinations of categories of the incident nodes. We report the number of edges in each of these nine categories (Figure 4.3(b)). This analysis highlights the drastic imbalance in the distribution of the edges between different node groups: Although Poor and Moderate nodes together comprise about 85% of all nodes in the network, 50% of all edges are between two Rich nodes and 90% of the edges in the network involve at least one Rich node. A concerning consequence of this imbalance is that, when all edges are valued equally in the evaluation metrics (as typically the case in standard settings) and when the edges in test set are sampled uniformly at random (Edge-Uniform), this guarantees most of the attention in the evaluation to be given to high-degree-nodes (70% expected influence for Rich nodes). In other words, the evaluation setting pays almost no attention to the ability of algorithms to predict interactions that involve low-degree nodes (5% influence for Poor nodes despite being a 53% majority of all nodes). This situation makes it lucrative for the algorithms to prioritize prediction of new interactions for well-studied proteins at the expense of under-studied ones, even though uncovering a new interaction between under-studied proteins may very well be more beneficial for biological knowledge generation[68].

To quantify the degree to which the algorithms are incentivized to prioritize high-degree nodes, we perform an analysis that we refer as separability analysis and assess the predictive power provided by node degree information. For this purpose, we compare the cumulative distribution functions (CDFs) of the preferential attachment scores for the positive and negative sets (known interactions in test set vs. other node pairs) and use the Kolmogorov-Smirnov (K-S) statistic to quantify the separability of the CDFs (which corresponds to the informedness[146] of the preferential attachment model at its best possible prediction point). Figure 4.3(c)



Rich nodes are a minority

Node categories are roughly determined based on percentages

They roughly indicate how well-studied the proteins are (based on the number of known interactions)

This analysis aims to investigate the imbalance in the network and benchmarking data. Thus, it does not depend on a prediction algorithm

This plot reflects the influence of the nodes on evaluation when each edge is valued equally

Poor + Moderate Only 10%

Involving Rich Nodes 90%

Rich only: 50%

Majority of the edges involve rich nodes (despite being a minority of the nodes)

Measures the predictive power of the node degree information for distinguishing the hidden interactions and other node pairs

Median score for positives is ~100

Much higher than it is for negatives

This view is helpful to analyze **the benchmarking setting** itself to quantify how susceptible it is to the degree bias

i.e., how much does it promote bringing forward rich nodes during prediction?

This imbalance causes known interactions (hidden from the algorithms) to be distinguishable using node degree information

This happens when the benchmarking data (i.e., hidden interactions in the test set) are generated randomly, giving each edge equal chance (Edge-Uniform)

Figure 4.3. Investigating the imbalance in the benchmarking data and the incentive towards high-bias predictors by quantifying the predictive power of node degree information in distinguishing the known interactions on Biogrid PPI network. (a) Assigned node categories indicating how well-studied a protein is based on node degree information. (b) The distribution of the edges in the network across these categories. (c) Separability analysis for the randomized/edge-uniform setting.

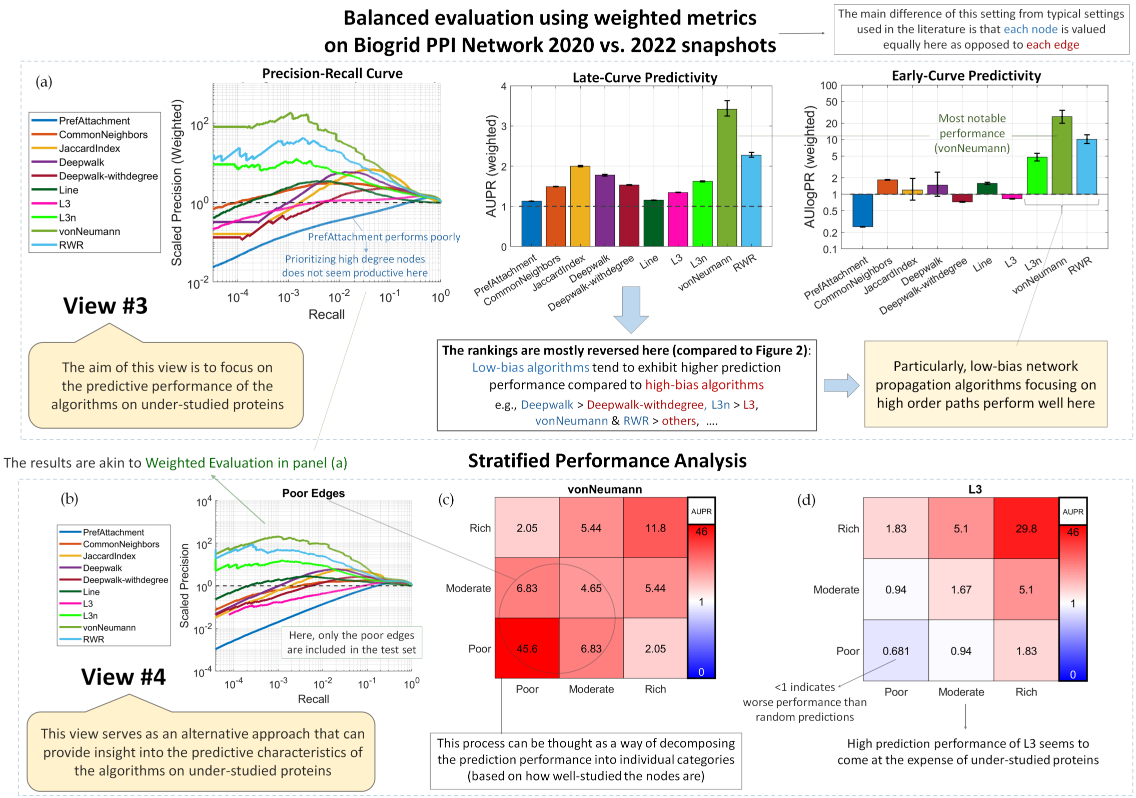
shows that the edges in the positive set generated by random (edge-uniform) sampling are largely distinguishable from negative pairs using the node degrees (K-S statistic: 73.6%). When multiple snapshots (across time) of the network are used for evaluation, the positive edges are still largely separable from negative pairs (K-S statistic: 59.3%, S. Figure 4.10), though to a lesser degree than it is for edge-uniform sampling. This suggests that using a different snapshot of the network as a test set instead of a randomly sampled test set *does not address* the issue of favoring algorithms that make biased predictions. In contrast, this observation reinforces the notion that research continues to generate knowledge that involves well-studied proteins[68], as the nodes that gain interactions over time are those that have high-degree in the earlier network. Thus, we conclude that an alternative evaluation style is needed to prevent the under-representation of the under-studied proteins on evaluation. For this purpose, in the next view, we focus on a simple idea: Valuing each node equally, as opposed to each edge.

View #3: Weighted evaluation setting focusing on under-studied entities

- **Aim:** To ensure that under-studied proteins are not under-represented while assessing the prediction performance for the link prediction algorithms.
- **Proposed strategy:** Apply weights to explicitly value the importance of discovering different interactions based on the degree of the involved nodes. The weights are optimized to balance the influence of the nodes on evaluation.

As we demonstrate in the previous section, standard evaluation settings provide little information on an algorithm's ability to make successful predictions involving under-studied entities, even though making successful predictions involving under-studied entities are at least as important as making successful predictions involving well-studied entities[68]. To fill this important gap in the evaluation pipeline, we propose a weighted setting that aims to balance the influence of each node on evaluation to be roughly equal (hence *node-uniform evaluation,* as opposed to each edge in standard settings). To obtain such weights, we formulate this as an optimization problem, where the objective is to make the weighted node degrees as close as possible to an input degree distribution (which we set as uniform distribution). We develop an iterative algorithm (Algorithm 1) to solve this optimization problem and assign the optimized weights to edges as instance weights during the computation of evaluation metrics (an alternative option to this, that we do not tackle in detail here, is to use the weights as probabilities to generate node-uniform sampled test sets for evaluation). We show that these weighted metrics can roughly balance the influence of the nodes in the evaluation process (S. Figure 4.11, 40%/35%/25% expected influence for Poor/Moderate/Rich nodes in weighted setting) and mitigate the degree bias in the benchmarking data by reducing the predictive power of node degree information (S. Figure 4.11(d), K-S statistic for pref. attachment is 33.9%/18.4% for edge-uniform/across-time data in weighted setting).

The evaluation of the link prediction algorithms using the weighted metrics is shown in Figure 4.4 for across-time data (results for sampled data are given in S. Figure 4.12). As seen in the figure, the performance comparisons suggested by this setting is quite different from that suggested by the standard settings (Figure 4.2) and low-bias versions of the algorithms tend to exhibit higher performances compared to high-bias versions here. Note that, while biased algorithms are not favored in this setting, anti-biased algorithms (that bring forward low-degree nodes indiscriminately) are not favored either. For example, anti-preferential attachment model (i.e., ranking the pairs in the opposite order for preferential attachment) performs just as worse as preferential attachment in this setting (S. Figure 4.13), which suggests that the weighting mitigates the degree bias in the evaluation without causing an anti-bias by inflating the weights of low-degree nodes beyond necessary. Overall, we observe that the best performing algorithms according to this setting are low-biased network propagation algorithms, von Neumann and RWR (whose performance levels are mostly consistent with the standard, unweighted setting, while the other algorithms' have dropped).



Balanced evaluation using weighted metrics on Biogrid PPI Network 2020 vs. 2022 snapshots

The main difference of this setting from typical settings used in the literature is that each node is valued equally here as opposed to each edge

View #3

The aim of this view is to focus on the predictive performance of the algorithms on under-studied proteins

PrefAttachment performs poorly

Prioritizing high degree nodes does not seem productive here

Most notable performance (vonNeumann)

The rankings are mostly reversed here (compared to Figure 2): Low-bias algorithms tend to exhibit higher prediction performance compared to high-bias algorithms

e.g., Deepwalk > Deepwalk-withdegree, L3n > L3, vonNeumann & RWR > others, ...

Particularly, low-bias network propagation algorithms focusing on high order paths perform well here

The results are akin to Weighted Evaluation in panel (a)

Stratified Performance Analysis

View #4

This view serves as an alternative approach that can provide insight into the predictive characteristics of the algorithms on under-studied proteins

Here, only the poor edges are included in the test set

This process can be thought as a way of decomposing the prediction performance into individual categories (based on how well-studied the nodes are)

<1 indicates worse performance than random predictions

High prediction performance of L3 seems to come at the expense of under-studied proteins

Figure 4.4. Balanced evaluation setting focusing on the prediction performance of the algorithms on under-studied proteins with the use of weighted metrics (node-uniform) and stratified analysis on Biogrid PPI Across-Time (2020 vs. 2022) data. Prediction performance of the algorithms for (a) weighted analysis and (b) stratified analysis. (c, d) Late curve predictivity (AUPR) stratified by node categories for von Neumann and L3 algorithms.

View #4: Stratified analysis to focus on under-studied proteins

- **Aim:** To assess the prediction performance of the algorithms for uncovering new interactions depending on how well-studied the involved proteins are.
- **Proposed strategy:** Stratify the prediction performance into individual edge categories (based on the degrees of incident nodes) by keeping only the interactions from one category in the test set during evaluation.

An alternate and perhaps more direct way to investigate the prediction performance of the algorithms for discovering new interactions on under-studied proteins is to decompose the prediction performance into individual categories based on node degrees, measuring the predictivity in each category by only keeping the edges in that category in the test set. For this purpose, we stratify the edges into 3x3 categories based on node degrees (similar to how it is done in Figure 4.3b) and further group them into two categories: Poor edges (edges between Poor+Moderate nodes), and Rich edges (remaining edges involved with a rich node). As seen in Figure 4.4(b) for across-time data, the precision-recall curves for Poor edges is quite similar to the ones obtained by weighted analysis in Figure 4.4 although quite different from the standard setting in Figure 4.2 (this is not surprising since poor edges are given 54% influence in the weighted setting as opposed to 8% in unweighted setting, S. Figure 4.14). Similarly, the curves for Rich edges (S. Figure 4.15) are akin to the ones in standard setting (as these edges were given 92% influence there). In Figure 4.4(c) and (d), we show the results of 3x3 stratified analysis for the best performing algorithms, vonNeumann and L3, respectively for Poor and Rich edges. Here, we observe that vonNeumann's predictivity is relatively balanced across different edge categories, whereas the high predictive performance achieved by L3 on Rich-Rich and Rich-Moderate interactions seems to come at the cost of severely diminished predictivity for edges involving under-studied proteins.

View #5: Simple but comprehensive summary for prediction performance

- **Aim:** To make a comprehensive and bias-aware evaluation in a simple manner.
- **Proposed strategy:** Measure five aspects, early/late curve prediction performance for under-studied/well-studied nodes, and the disposition of an algorithm regarding degree bias.

While inspecting the performance curves (as in Figure 4.2(a), and Figure 4.4(a)) or the results of stratified analysis (as in Figure 4.4(c) and (d)) are in general more informative than looking at individual summary metrics, the use of such metrics is still critical for making quick assessments and comparisons. As a reasonable compromise between simplicity and comprehensiveness, we propose the use of five-metrics to summarize the prediction characteristics of a given algorithm in a bias-aware manner (Figure 4.5), measuring: fairness of the predictions (defined as 1 minus absolute value of the bias metric), early and late curve predictivity (measured by AUPR and AUlogPR respectively) for well-studied entities (results of the standard, unweighted evaluation setting valuing each edge equally) and the same for under-studied entities (results of the weighted setting, valuing each node equally).

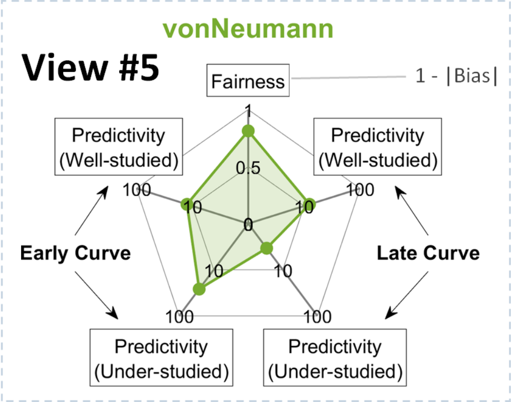


Figure 4.5. 5-metric summary for von Neumann algorithm.

Bias in benchmarking data for other link prediction problems in biology

The results presented so far demonstrate the bias toward well-studied proteins in benchmarking data and the evaluation setting in the context of PPI predictions using the Biogrid network. To assess the generalizibility of our conclusions and motivate the application of the proposed strategies to a broader range of problems, we here analyze the bias in benchmarking data for additional PPI datasets

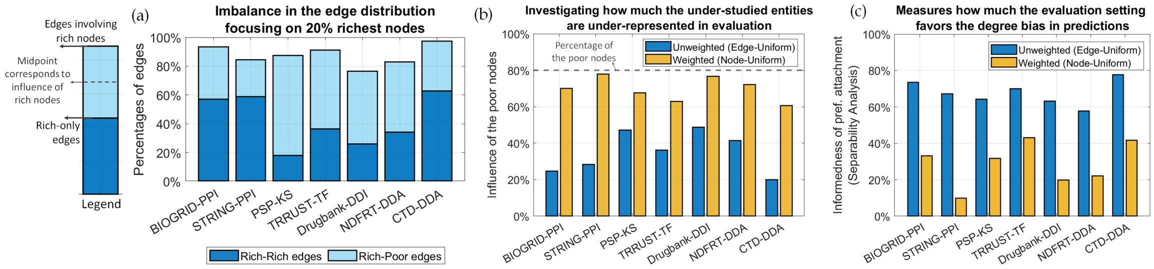


Figure 4.6. Investigating different datasets and link prediction problems in the context of network biology in terms of the imbalance in the network, under-representation of the under-studied entities in the evaluation, and the incentive towards high-bias predictors. (a) The percentage of edges involved with the 20% richest nodes in the network. (b) Influence of the under-studied (i.e., 80% the lowest degree) nodes in the evaluation for weighted (node-uniform) and unweighted (edge-uniform) settings. (c) Predictive power of the node degree information (measured by the separability analysis). Please visit https://yilmazs.shinyapps.io/colipe to inspect the imbalance in these datasets in an interactive manner.

and other prominent link prediction problems in biomedical applications for the weighted (node-uniform) and unweighted (standard, edge-uniform) settings (Figure 4.6). Specifically, we investigate the STRING human PPI network[129], PhosphositePlus kinase-substrate interactions (PSP-KS)[49], TRRUST transcription factor regulatory interactions[43], DrugBank drug-drug interactions (Drugbank-DDI)[61], and drug-disease association (NDRFT-DDA[148] and CDA-DDA[27]) networks. For each of these datasets, we examine the imbalance in the edge distribution for the top 20% of the nodes with highest degrees (Figure 4.6a), quantify how much the under-studied entities are under-represented in the evaluation (by measuring the expected influence of 80% of the nodes with lowest degrees, Figure 4.6b), and the incentive provided by the evaluation setting towards high-bias predictors (by measuring the predictive power of node degree information based on separability analysis, Figure 4.6c). Overall, we observe that for a broad range of network datasets that are commonly used for benchmarking link prediction algorithms, there is large degree of imbalance in the edge distributions, and as a result, standard settings that equally value each edge in the evaluation reward algorithms that bring forward high-degree, well-studied entities in their predictions. To overcome this issue, node-uniform weighting can help balance the influence of the nodes and prevent the under-studied proteins from being under-represented in the evaluation.

4.3 Conclusions

Overall, to facilitate bias-aware evaluation of link prediction algorithms and to promote the discovery of new interactions involving under-studied entities, we suggest the following approaches to developers and evaluators of link prediction algorithms:

- **Bias in algorithms:** Investigate the disposition of the algorithms towards well-studied nodes based on similarity with preferential attachment model (View 1).
- **Bias in benchmarking data:** Examine the training/test splits (and the benchmarking setting itself) to see how much they incentivize high-bias predictors by quantifying the predictive power of node degree information (View 2).
- **Evaluating the prediction performance:** Adopt a weighted setting, valuing each node as opposed to each edge equally (View 3), or perform a stratified analysis (View 4) to assess the prediction performance on under-studied proteins.
- **Summarize the findings:** Consider five aspects to give an outline for the main characteristics of an algorithm, regarding the early curve/late curve predictivity, well-studied/under-proteins, as well as the bias in predictions (View 5).

4.4 Methods Summary

In this section, we briefly describe the methodology we propose, focusing on evaluation metrics and the proposed weighted validation setting. Technical details, formal descriptions, and other information regarding the methods used in this work are provided as Supplementary Materials.

Evaluation metrics. To measure the late-curve prediction performance, we utilize the area under the precision-recall curve (AUPR). For the early-curve performance, we compute the area under the precision-recall curve in log-log scale (AUlogPR) through numerical integration (after normalizing the logarithmic x-axis such that the resulting unit for AUlogPR is precision). Note that, to make the evaluation results comparable with different networks or settings, we scale both metrics to have an expected value of 1 for random predictions. To account for the variance in the estimation of these meaures, we construct 95% credible intervals following a Bayesian approach[52].

Optimization algorithm for node-uniform edge weights. To obtain a set of edge weights (denoted as $\mathbf{W}$ matrix) that establishes node-wise uniformity (i.e., for the row and column sums of $\mathbf{W}$ to be equal for all nodes), we formulate this as an optimization problem and develop an algorithm that iteratively performs multiplicative updates (ensuring the uniformity of the rows in one step, and for the columns in another) until the uniformity of the row and column sums are established simultaneously at an acceptable level. The formulation of the optimization problem and a simplified pseudo-code of the developed algorithm (omitting some details) is given in Algorithm 1. A more detailed description of the algorithm (including a complete pseudo-code) specifying the technical details (e.g., regarding matrix initialization, termination conditions, controlling the step size during updates and so on) are provided in the Supplementary Methods. Note that, we denote $\vec{Q_r}$ and $\vec{Q_c}$ to indicate the desired weights for rows and columns instead of assuming uniformity for the sake of generalizability (i.e., node-uniform when $\mathbf{Q}_r = \mathbf{Q}_c = \mathbf{1}$).

Algorithm 1 Optimization Problem and Algorithm (Simplified Pseudo-Code)

```
Require: Graph $\mathbf{G} = \{\mathcal{V}, \mathcal{E}\}$, desired node weights $\vec{Q_r}$ and $\vec{Q_c}$ for rows and columns,
    step size $\alpha \leq 1$
Ensure: Edge weighting matrix $\mathbf{W}$ such that $\mathbf{W}_{ij} = 0$ if $(i, j) \notin \mathcal{E}$ and $\mathbf{W}_{ij} \geq$
    $0\,\forall(i, j) \in \mathcal{E}$
    Initialize $\mathbf{W} \leftarrow \mathbf{W}_{init}$ and Normalize $\mathbf{W}$ to sum up to 1
    while maximum iteration limit is not reached do
        $\mathbf{D_r} \leftarrow$ row sums of $\mathbf{W}$
        $\mathbf{D_c} \leftarrow$ column sums of $\mathbf{W}$
        $\mathbf{W} \leftarrow \mathbf{W} \oslash (\mathbf{D_r} \oslash \mathbf{Q_r})^{\alpha}$        ▷ $\oslash$ indicates Hadamard (element-wise) division
        $\mathbf{W} \leftarrow \mathbf{W} \oslash (\mathbf{D_c} \oslash \mathbf{Q_c})^{\alpha}$
        Normalize $\mathbf{W}$ to sum up to 1
    end while
Optimization Problem: Compute $\mathbf{W}$ to minimize $||\mathbf{Q_r} - \mathbf{D_r}||_2$ + $||\mathbf{Q_c} - \mathbf{D_c}||_2$
```

Weighted evaluation. In the weighted evaluation setting, we use the optimized edge weights (the matrix $\mathbf{W}$ computed by the algorithm above) as the weight for each positive instance (existing edge in the test set). For this purpose, we generalize the computation of AUPR and AUlogPR to assign weights to instances (positives in the test set) while counting the number of true positives (TPs) and false negatives (FNs). For example, an edge in the test set that is weighted worth of 3 unweighted edges, if included in the predictions of an algorithm, would increase the number of TPs by 3 as opposed to 1. Performance measures are then computed based on these weighted counts.

Influence of a node category on evaluation. We quantify the influence of a node category (rich, moderate, or poor) on evaluation as the total weight of the edges (percentage of edges for standard evaluation) that are incident to the nodes in that category, counting between-category edges as half such that total influence for all categories adds up to 1.

4.5 Methods

4.5.1 Link Prediction Algorithms - Verbal Descriptions

Methods based on scoring metrics: We consider the preferential attachment model to represent a purely biased model (where node pairs are ranked based on the product of the degrees of the endpoints). Common neighbors represents a high-bias algorithm that considers paths of length 2 (high bias since the number of paths in correlated with the node degrees). Jaccard Index represents a low bias version of common neighbors where a normalization is applied based on node degrees.

High order paths/Network propagation algorithms: L3 is a method that counts the paths of length 3 to make predictions. For this purpose, in this work, we consider the formulation given in[62] that applies a soft normalization (based on square root of degrees, this is what we consider the high-bias version). We also introduce a low bias version of it, L3-Normalized (L3n) that applies a stronger normalization based on node degrees. Whereas, von Neumann[67] and random walks with restarts (RWR)[131] are network propagation algorithms that consider a weighted combination of paths of different lengths. Formulation of both algorithms involve a strong normalization based on node degrees (the main difference between them is the style of the normalization, whether it is done symmetrically or based on column normalization). Thus, we consider both to be low-bias algorithms.

Embedding/Learning Methods: We consider two types of embedding methods: Deepwalk[107] (Random walk based) and Line[130] (Neural network based). For each of these methods, we train a logistic regression model using the embeddings as features. Here, deepwalk represents a low-bias algorithm (since the embedding dimensions are uncorrelated with node degrees, likely by design, S. Figure 4.7) and Line is a higher bias algorithm (since its embeddings pick up the node degree info during learning, S. Figure 4.8). For deepwalk-withdegree, we include the node degrees as an additional dimension (as if it is part of the embeddings matrix) to construct a high-bias version of the deepwalk algorithm.

4.5.2 Link Prediction Algorithms - Mathematical Formulations

Methods based on simple scoring metrics. *Preferential Attachment:*

$$\sigma_{PA}(u,v) = \sqrt{|\mathit{\Gamma}(u)||\mathit{\Gamma}(v)|} \tag{4.1}$$

where $\mathit{\Gamma}(u)$ denotes the set containing the neighbors of u. In matrix form, the preferential attachment score is equal to:

$$\sigma_{PA} = \sqrt{D_r \odot D_c} \tag{4.2}$$

where $\odot$ indicates the element-wise (Hadamard) product and D_r and D_c are respectively row and column degrees in matrix form:

$$\begin{aligned} D_r(u,v) &= |\mathit{\Gamma}(u)| \\ D_c(u,v) &= |\mathit{\Gamma}(v)| \end{aligned} \tag{4.3}$$

Common Neighbors:

$$\sigma_{AA}(u,v) = |\mathit{\Gamma}(u) \cap \mathit{\Gamma}(v)| \tag{4.4}$$

In matrix form, this is simply equal to:

$$\sigma_{AA} = A^2 \tag{4.5}$$

where A is the adjacency matrix of the network.

Jaccard Index:

$$\sigma_{JI}(u,v) = \frac{|\mathit{\Gamma}(u) \cap \mathit{\Gamma}(v)|}{|\mathit{\Gamma}(u) \cup \mathit{\Gamma}(v)|} \tag{4.6}$$

In matrix form,

$$\begin{aligned} \sigma_{JI} &= A^2 \oslash N \\ N &= D_r + D_c - A^2 \end{aligned} \tag{4.7}$$

where $\oslash$ indicates the element-wise (Hadamard) divide operation.

Higher-order paths and Network propagation based methods. *L3:* In matrix form,

$$\begin{aligned} \sigma_{L3} &= A' \times A' \times A \\ A' &= A \oslash \sqrt{D_r} \end{aligned} \tag{4.8}$$

L3-Normalized (L3n):

$$\begin{aligned} \sigma_{L3n} &= A_n^3 \\ A_n &= A \oslash D_r \end{aligned} \tag{4.9}$$

von Neumann:

$$\begin{aligned} A_s &= A \oslash \sqrt{D_r \odot D_c} \\ \sigma_{VN} &= \alpha A_s + \alpha^2 A_s^2 + \alpha^3 A_s^3 \ldots \\ &= \sum_{i=1}^{l} \Bigg(\alpha^i A_s^i \end{aligned} \tag{4.10}$$

Here, we use $\alpha = 0.5$ and go up to path lengths of $l = 4$ for computational efficiency reasons.

Random walks with restarts (RWR):

$$\begin{aligned} A_n &= A \oslash D_r \\ \sigma_{RWR} &= \alpha A_n + \alpha^2 A_n^2 + \alpha^3 A_n^3 \ldots \\ &= \sum_{i=1}^{l} \Bigg(\alpha^i A_n^i \end{aligned} \tag{4.11}$$

Similar to von-Neumann method, we use $\alpha = 0.5$ and go up to $l = 4$.

Embedding-based methods. In addition to these methods that compute a single score for each candidate pair, recent link prediction algorithms commonly use node embeddings to facilitate supervised learning. Node embeddings map the nodes in a network to a lower-dimensional embedding space, such that adjacent nodes are mapped to points that are close to each other in this embedding space[41]. Subsequently, using these embeddings as feature vectors and existing edges as training data, machine learning models are trained to predict new edges. We consider two embedding methods that are representative of common approaches to the computation of node embeddings.

Deepwalk (Random walk based node embedding): Deepwalk[107] uses random walks to generate a list of paths in the network as its corpus and then uses Word2Vec[88], a natural language processing algorithm for word embedding, to compute node embeddings by treating the list of paths as text and nodes as words. In our experiments, we use the implementation used in BioNEV[148] repository (which is based on OpenNE[1]) for all embedding methods.

LINE (Neural network based embedding): LINE[130] is one of the earliest algorithms to incorporate neural networks into the computation of node embeddings. It uses a single layer MLP to estimate first and second order proximity of nodes and produces the embedding vectors using a variational auto-encoder.

Unless otherwise specified, we use the default value of 128 in the OpenNE implementation as the embedding dimension (i.e., number of embeddings) for both embedding methods.

In addition to the above two, we consider a version of deepwalk (***Deepwalk-withdegree***) where the node degree information is appended as an additional dimension to the embedding matrix.

Logistic regression as prediction model: For each of three embedding approaches described above, we train a logistic regression using the embeddings as features. For each embedding dimension $x^{(i)}$, we add the following three terms to the logistic regression model corresponding to the prediction for edge (u,v):

$$\text{logit}(Y_{uv}) \propto \sum_i \left(\beta_r^{(i)} x_u^{(i)} + \beta_c^{(i)} x_v^{(i)} + \beta_{rc}^{(i)} x_u^{(i)} x_v^{(i)}\right) \tag{4.12}$$

While training the model, to ensure a balanced training set, we randomly sample the edges with negative labels (i.e., not in the training set) to have the same size as the edges with positive labels.

4.5.3 Evaluation Metrics for Prediction Performance

Precision and Scaled Precision:

$$\text{Precision} = r = \frac{TP}{TP + FP} \tag{4.13}$$

where TP denotes the number of true positives, and FP the number of false positives. The expected precision for random predictions is equal to the prevalence of positive labels:

$$E[\text{Precision}] = \text{Prevalence} = \frac{N_P}{N_{\text{total}}} \tag{4.14}$$

where N_P is the number of positive labels and N_{total} is the total number of edges that are to be predicted (approximately $O(N^2)$).

$$\text{Scaled Precision} = \frac{\text{Precision}}{E[\text{Precision}]} = \frac{TP}{TP + FP}\frac{N_{\text{total}}}{N_P} \tag{4.15}$$

Recall:

$$\text{Recall} = \frac{TP}{N_P} \tag{4.16}$$

Computing the area under precision-recall curve (AUPR). To compute the area under the precision-recall (PR) curve, we use numerical integration. Suppose we have m measurement points. Let TP_i denote the number of true positives, FP_i the number of false positives, $N_i = TP_i + FP_i$ the number of predictions, X_i the recall, and Y_i the precision corresponding to the ith measurement point. In general, m is less than N_{total} since edges having the same prediction score (e.g., because the link prediction method uses discrete scoring like common neighbors) correspond to a single measurement point. Also, without loss of generality, consider that the first point is the $TP = 0$ and $FP = 0$ point and all points are sorted by the number of predictions ($TP_i + FP_i$) in ascending order. With these in mind, we compute the area under precision-recall curve through numerical integration as follows:

$$\text{AUPR} = \frac{\sum_{i=1}^{m-1} \mathit{\Delta} X_i f Y_i}{\sum_{i=1}^{m-1} \mathit{\Delta} X_i} \tag{4.17}$$

where $\mathit{\Delta} X_i$ is the gap between two consecutive points:

$$\mathit{\Delta} X_i = |X_{i+1} - X_i| \tag{4.18}$$

Whereas, fY_i is an interpolating function that returns the normalized area under two consecutive points Y_i and Y_{i+1} (thus, it is a type of averaging for two given points and is always between $[Y_i, Y_{i+1}]$). For example, a simple function for this purpose can be $\frac{Y_i + Y_{i+1}}{2}$ (interpolating the precision values linearly). However, this type of interpolation suffers from inaccuracy when there are large gaps between two consecutive points X_i and X_{i+1}, which is particularly relevant for link prediction methods with discrete scoring. To demonstrate the inaccuracy, suppose the first point is at 1/1 ($TP = 1$, $FP = 0$) with precision 1 and the next point is in 100/10000 ($TP = 100$, $FP = 9900$) with precision 0.01. In this example, although linear interpolation suggests that the average precision would be ≈0.5, observing one TP in the beginning hardly gives any evidence that the precision will be ≈0.5 at the $TP = 5000$ point. To overcome this type of inaccuracy, we use an interpolation function tailored for the precision-recall curve detailed below:

Interpolating the curve during numerical integration for computing the area under. For the intermediate points between two consecutive points X_i and X_{i+1}, we assume that both the true positives and false positives are scaled linearly:

$$\begin{aligned} TP_x &= TP_i + x(TP_{i+1} - TP_i) \\ FP_x &= FP_i + x(FP_{i+1} - FP_i) \end{aligned} \tag{4.19}$$

where x is a normalized variable between [0, 1] indicating which endpoint the point is closest to (e.g., 1 indicates the point is right on the i+1th point). Thus, the precision for the intermediate points is given by the ratio $r_i(x)$:

$$\begin{aligned} r_i(x) &= \frac{TP_i + x(TP_{i+1} - TP_i)}{TP_i + x(TP_{i+1} - TP_i) + FP_i + x(FP_{i+1} - FP_i)} \\ &= \frac{TP_i + x(TP_{i+1} - TP_i)}{N_i + x(N_{i+1} - N_i)} \end{aligned} \tag{4.20}$$

To compute the area under this curve (denoted fY), we need the integral:

$$fY_i = \int_0^1 r_i(x)dx = \int_0^1 \frac{TP_i + x(TP_{i+1} - TP_i)}{N_i + x(N_{i+1} - N_i)}dx \tag{4.21}$$

Solving this integral gives:

$$fY_i = \frac{(TP_iN_{i+1} - TP_{i+1}N_i)\log\left(\frac{N_{i+1}}{N_i}\right)\Bigg(+ (N_{i+1} - N_i)(TP_{i+1} - TP_i)}{(N_{i+1} - N_i)^2} \tag{4.22}$$

Thus, we use the above function for interpolating while computing the area under the PR curve. To give some insight into what this function results in: For the example before (one point at TP/N = 1/1 while the other is at 100/10000), this integral results in ≈ 0.011 precision which is much closer to the latter point (as it should be).

Note that, although this integral (and Equation (4.22)) is not defined at $N_i = 0$ point (since precision is not defined at 0 predictions), the limit from above converges to $\frac{TP_{i+1}}{N_{i+1}} = Y_{i+1}$. Thus, as the first interpolated area, we use:

$$fY_1 = Y_2 \tag{4.23}$$

where Y_2 (i.e., the second point) corresponds to the first measured precision value (since the 0/0 point is specified in the $i = 1$th point in this notation).

Overall, this interpolation is helpful for reducing the inaccuracy when there are large gaps in between, which is particularly relevant for methods with discrete scoring or for computing the area under the PR curve in logarithmic scale.

Early-curve performance, the area under log-scale precision-recall curve (AUlogPR). For computing the area under log-log scale PR curve, the process is similar. We use the interpolating function fY given in Equation (4.22) and perform numerical integration as follows:

$$\begin{gathered}\text{AUlogPR} = \exp\left(\frac{\sum_{i=1}^{m-1} \Delta X_i' \log fY_i}{\sum_{i=1}^{m-1} \Delta X_i'}\right) \\ \Delta X_i' = \log X_{i+1} - \log X_i\end{gathered} \tag{4.24}$$

Note that, while computing AUlogPR, we start the curve at $TP = 10$ point to reduce the variance in the estimation (since the initial points between TP=[1, 10] are considerably volatile).

Note that, for both AUPR and AUlogPR metrics, after computing the area under the curve, we scale them to have an expected value of 1 (for random predictions) by dividing with the prevalence of the positive labels (Equation (4.14)), similar to how it is done in Equation (4.15).

Computing credible intervals for the variance in estimation. We follow a Bayesian approach to estimate the expected variance in the evaluation metrics (e.g., precision and AUPR). Our view here is akin to the "checking whether a coin is fair" problem. We assume that there is an unknown, but fixed probability r (corresponding to precision). Based on this probability, we suppose that we have made k trials (corresponding to predictions) and observed TP number of hits and FP number of misses. Now, we ask the question "Based on these observations, what can we say about the posterior probability of the ratio r?"

If we assume uniform prior (i.e., all r values in [0, 1] are equally likely), the answer to the above question is specified by the beta distribution:

$$r \sim \text{Beta}(TP + 1, FP + 1) \tag{4.25}$$

Thus, we obtain the distribution for the posterior probability of the ratio r (i.e., precision) after k predictions. Based on this distribution, we can easily construct a credible interval containing the 95% of the variance using the inverse cumulative distribution function Beta^{-1}. Note that, in general, there is not a single credible interval unique to a given posterior distribution. Thus, among the alternatives, we choose the equal-tailed interval where the probability of being below the interval is as likely as being above it.

This process gives us a 95% interval for the precision at fixed number of predictions k point. To obtain 95% intervals for the area under metrics (AUPR and AUlogPR), we simply construct the intervals for all k points and compute the area under the precision-recall curves formed by the maximum/minimum bounds.

4.5.4 Weighted Validation Setting Focusing on Under-Studied Entities

Optimization algorithm for obtaining edge weights based on node valuations. We formulate this problem as follows: Suppose we are given as set of node valuations q. Let $q_r(u)$ and $q_c(u)$ denote the desired expected number of edges coming into and going out of $u \in \mathcal{V}$ (i.e., the desired row and column sums). Let $\mathbf{W}$ represent the weights of the edges as a sparse matrix where $\mathbf{W}_{ij} = 0$ if $(i, j) \notin \mathcal{E}$. Here, our aim is to estimate a set of edge weights/values $\mathbf{W}$ such that the row and column sums of $\mathbf{W}$ are respectively equal to v_r and v_c. For this purpose, we will use an expectation-maximization based optimization algorithm with multiplicative steps.

Algorithm 2 Optimization Algorithm for Edge Weights

```
Require: Node valuations $\mathbf{Q_r}$ and $\mathbf{Q_c}$, max. number of iterations, convergence
    threshold $\epsilon_{convergence}$, maximum step size $\alpha_{max}$, step size increment $\gamma_{increment}$, step
    size decrement $\gamma_{decrement}$
Ensure: Edge weighting matrix $\mathbf{W}$
    Initialize $\mathbf{W} \leftarrow \mathbf{W}_{init}$
    Normalize $\mathbf{W}$ to sum up to 1
    Set initial step size $\alpha \leftarrow \alpha_{max}$
    $\mathbf{W}_{best} \leftarrow \mathbf{W}$
    bestError $\leftarrow \infty$
    while maximum iteration limit is not reached do
        $\mathbf{D_r} \leftarrow$ row sums of $\mathbf{W}$
        $\mathbf{D_c} \leftarrow$ column sums of $\mathbf{W}$
        error $\leftarrow$ sum of squared error for $\mathbf{W}$
        Measure $\Delta_{change}$ and $\Delta_{improvement}$
        if $\Delta_{improvement} > 0$ then                                          ▷
            $\mathbf{W}_{best} \leftarrow \mathbf{W}$
            bestError $\leftarrow$ error
            Increase step size $\alpha \leftarrow \min(\alpha\gamma_{increase}, \alpha_{max})$
        else
            Restore $\mathbf{W} \leftarrow \mathbf{W_{best}}$
            Decrease step size $\alpha \leftarrow \alpha\gamma_{decrease}$
        end if
        if $(\Delta_{change} + \Delta_{improvement}) \leq \epsilon_{convergence}$ then
            Stop the optimization and return $\mathbf{W}_{best}$
        end if
        $\mathbf{W} \leftarrow \mathbf{W} \oslash (\mathbf{D_r} \oslash \mathbf{Q_r})^{\alpha}$      ▷ $\oslash$ indicates Hadamard (element-wise) division
        $\mathbf{W} \leftarrow \mathbf{W} \oslash (\mathbf{D_c} \oslash \mathbf{Q_c})^{\alpha}$
        Normalize $\mathbf{W}$ to sum up to 1
    end while
```

The pseudo-code of the algorithm is given in Algorithm 2. Below, we describe each step of the algorithm:

Initialization: In the beginning of the algorithm, we set W to be equal to an initial, approximate solution:

$$\mathbf{W}_{init}(i,j) = \begin{cases} \sqrt{\dfrac{q_r(i)q_c(j)}{d_r(i)d_c(j)}} & \text{if } (i,j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases} \tag{4.26}$$

where $d_r(i)$ and $d_c(j)$ are the row and column degrees of nodes i and j in the network respectively. After setting $\mathbf{W} = \mathbf{W}_{init}$, we normalize the weights $\mathbf{W}$ to sum up to 1 as follows:

$$\mathbf{W} \leftarrow \frac{\mathbf{W}}{\sum_{i,j} \mathbf{W}(i,j)} \tag{4.27}$$

Update steps of the algorithm: Here, to ensure that the updated weights remain positive, we use multiplicative update steps based on row/column normalizations:

$$\begin{aligned} \mathbf{W} &\leftarrow \mathbf{W} \oslash (\mathbf{D_r} \oslash \mathbf{Q_r})^{\alpha} \\ \mathbf{W} &\leftarrow \mathbf{W} \oslash (\mathbf{D_c} \oslash \mathbf{Q_c})^{\alpha} \end{aligned} \tag{4.28}$$

where α is a multiplicative step size parameter and $\mathbf{D_r}$ and $\mathbf{D_c}$ are row/column sum matrices of $\mathbf{W}$ respectively:

$$\mathbf{D_r}(i,j) = \sum_j \mathbf{W}(i,j) \quad \text{and} \quad \mathbf{D_c}(i,j) = \sum_i \mathbf{W}(i,j) \tag{4.29}$$

Similarly, input node valuation vectors v_r and v_c are organized as matrices $\mathbf{V_r}$ and $\mathbf{V_c}$ after being normalized:

$$\mathbf{Q_r}(i,j) = \frac{q_r(i) n_r}{\sum_{i'} q_r(i')} \quad \text{and} \quad \mathbf{Q_c}(i,j) = \frac{q_c(j) n_c}{\sum_{j'} q_c(j')} \tag{4.30}$$

where n_r and n_c are scalars indicating the number of rows and columns in the network. In each step, after updating $\mathbf{W}$ according to Equation (4.28), $\mathbf{W}$ is normalized again to sum up to 1 as in Equation (4.27).

Termination of the algorithm: To determine the convergence of the algorithm, we look at two criteria. The first one focuses on the amount of change in $\mathbf{W}$:

$$\varDelta_{change} = \frac{||\mathbf{W} - \mathbf{W}_{best}||_2}{||\mathbf{W}_{best}||_2} \tag{4.31}$$

The second one focuses on the amount of improvement. For this purpose, we first quantify the error using sum of squares:

$$error(\mathbf{W}') = \sum_i \left(Q_r(i) - \sum_j \mathbf{W}'(i,j) \right)^2 + \sum_j \left(Q_c(j) - \sum_i \mathbf{W}'(i,j) \right)^2 \tag{4.32}$$

Thus, at each step, we measure the improvement $\mathbf{W}$ provides over $\mathbf{W}_{best}$ as follows:

$$\varDelta_{improvement} = \max \left(1 - \frac{error(\mathbf{W})}{error(\mathbf{W}_{best})}, 0 \right) \tag{4.33}$$

Overall, we terminate the algorithm when the amount of change plus the improvement is less than a predefined threshold $\epsilon_{convergence}$:

$$\text{Terminate if } \Delta_{change} + \Delta_{improvement} \leq \epsilon_{convergence} \tag{4.34}$$

In addition to the convergence threshold, we terminate the optimization if the maximum iteration limit is reached. Unless otherwise specified, we use $\epsilon = 10^{-2}$ and 100 maximum iterations for the termination of the algorithm.

Updating the step sizes: When there is no improvement at any point (i.e., $\Delta_{improvement} \leq 0$), we conclude that step size is too large and need to be reduced. For this purpose, we restore $\mathbf{W}$ to $\mathbf{W}_{best}$ (i.e., the best weights with lowest error up to this point) and decrease step size α by a factor of $\gamma_{decrease}$:

$$\alpha \leftarrow \alpha\gamma_{decrease} \tag{4.35}$$

Conversely, when $\Delta_{improvement} > 0$, we restore α by increasing it with a factor $\gamma_{increase}$ and truncating it to α_{max}:

$$\alpha \leftarrow \min\left(\alpha\gamma_{increase}, \alpha_{max}\right) \tag{4.36}$$

Unless otherwise specified, we set $\alpha_{max} = 0.999$, $\gamma_{decrease} = 0.6$, and $\gamma_{increase} = 1.25$.

Note about sparse matrices and efficiency: Here, we have described the update steps (Equation (4.28)) in terms of $\mathbf{D_r}/\mathbf{D_c}$ and $\mathbf{Q_r}/\mathbf{Q_c}$ in matrix format for the sake of brevity and clarity. While implementing the algorithm, the element-wise divide ($\oslash$) operation can be efficiently applied on vectors and sparse matrices without ever storing the full matrices.

Weighted evaluation metrics. After obtaining the weighting matrix $\mathbf{W}$ using the optimization algorithm, let weighting vector $w \in \mathbb{R}^{N_{total} \times 1}$ be organized in such a way that w_i indicates the edge weight corresponding to the ith prediction (after all edges are sorted based on the prediction scores of a method). Using this vector, we can compute the weighted true positives for k predictions as follows:

$$TP_w = \frac{\sum_{i=1}^{k} w_i I_i}{w_{\text{norm}}} \tag{4.37}$$

where I_i is an indicator variable that is equal to 1 if the ith prediction is a true positive and is equal to 0 otherwise and w_{norm} is a normalization factor:

$$w_{\text{norm}} = \frac{\sum_{i=1}^{N_{\text{total}}} w_i I_i}{N_P} \tag{4.38}$$

where N_P is the number of positive labels in the test set.

After obtaining the weighted true positives, the weighted versions of the AUPR and AUlogPR metrics are computed as described in the previous sections (this time using TP_w instead of TP).

$$\text{Precision (Weighted)} = r_w = \frac{TP_w}{TP_w + FP} \tag{4.39}$$

Computing credible intervals for the weighted metrics. Previously in *"Computing intervals for the variance in estimation"* section, we obtained the posterior distribution of the ratio r corresponding to unweighted precision (Equation (4.25)). Here, we will transform this for the weighted precision. For this purpose, we start by defining a weighting factor w_f equal to the ratio of weighted and unweighted true positives:

$$w_f = \frac{TP_w}{TP} \tag{4.40}$$

Using this, we can write the equation for weighted precision in terms of the unweighted ratio r:

$$\begin{aligned} r_w = \frac{TP_w}{TP_w + FP} &= \frac{w_f TP}{w_f TP + FP} \\ = \frac{\frac{w_f TP}{TP + FP}}{\frac{w_f TP + FP}{TP + FP}} &= \frac{w_f r}{(w_f - 1)r + 1} \end{aligned} \tag{4.41}$$

Thus, we transform the distribution given in Equation (4.25) according to the above equation to obtain the posterior distribution of the weighted ratio r_w. After that, we compute the 95% credible intervals as detailed before.

4.5.5 Quantifying the bias towards high-degree nodes in method predictions

Here, for each algorithm, we count the number of overlapping edges with the predictions from preferential attachment (i.e., the number of pairs that are predicted as positive by both algorithms) for varying values of k (where k indicates the number of predictions for both algorithms). Next, we compute the bias metric based on the area under this curves and normalize it with respect to the maximum & expected values to confine in [-1, 1] region (the maximum is equal to k, whereas the expected value for random predictions is k^2 divided by the total number of pairs to be predicted).

4.5.6 Separability analysis quantifying the imbalance in benchmarking data

Here, we first compute the preferential attachment scores (using the node degrees in training data) for the positives (the hidden interactions in test set) and negative node pairs (pairs without a known interaction). Next, we make use of the kolmogorov-smirnov statistic (which corresponds to the maximum distance in the cumulative distribution functions of) to quantify the predictive power of node degree information. Note that, this way of quantifying the separability is equivalent to computing informedness at the best prediction point (i.e., the maximum vertical difference in the ROC curve and the diagonal line corresponding to random predictions) for the predictions of preferential attachment model. For the weighted version of the separability analysis, we simply use the optimized edge weights as instance weights while estimating the cumulative distribution function (CDF) for the positives and compute the kolmogorov-smirnov statistic as usual (this time using the CDF for the weighted positives).

4.5.7 Computing the influence of the nodes on evaluation

Influence of the node categories on the evaluation is computed based on the percentages given in Figure 4.3(b) and S. Figure 4.11(b) (i.e., based on the number of edges or the weights of edges). Here, the influence of mixed category edges are counted as as half for each category (e.g., a Poor-Rich edge provides half of its weight as influence to poor category, and the other half to rich category). Note that, this way of computing the node influence ensures that their total adds up to 1.

4.5.8 Performing stratified analysis

For the stratified analysis, we first obtain the node categories (Poor, Moderate, Rich) as shown in Figure 3a. Next, we assign each edge in the test set into one of six categories (e.g., Rich-Rich, Poor-Rich, Poor-Moderate and so on). For each of these categories separately, we repeat the evaluation (and compute performance metrics like AUPR), keeping only the edges in the corresponding category in the test set (in other words, considering the prediction of only the edges in that category to be true positives). Note that, the background set of possible node pairs is not affected by this stratification (i.e., the negative set includes pairs from all categories).

4.5.9 Datasets used in this work

The bulk of the experiments done in this paper uses BioGRID[104] Human Protein-Protein Interaction network for two versions obtained at different times: (i) 2020 version (v4.0.189) contains 464,003 interactions between 25,776 proteins, (ii) 2022 version (v4.4.210) contains 784,774 interactions between 27,408 proteins. For constructing the training/test sets across time (2020 for training, new edges in 2022 for testing), we filter for the proteins that exist in the 2020 version and use the 308,334 new interactions for 16305 proteins in 2022 version as the test set (for a total of 772,337 interactions between 25,776 proteins, training & test sets combined).

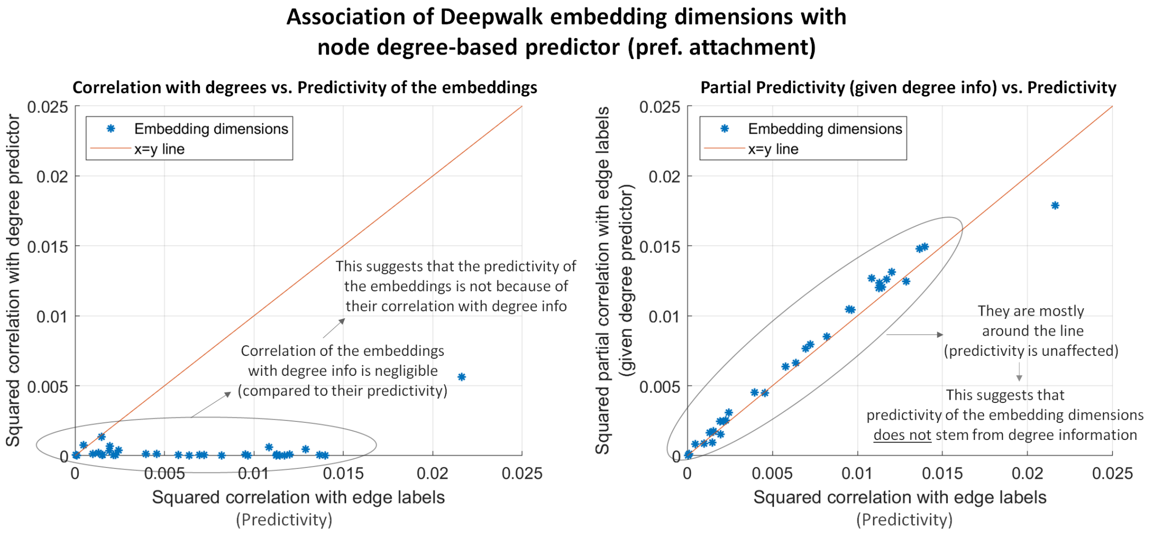
The final part of our analysis includes six other networks listed below. Some of them were obtained and parsed from the source databases directly, while others are taken from BioNEV[148] repository as pre-processed edgelists.

- STRING PPI[129] contains 359,776 interaction between 15,131 proteins. Taken from BioNEV repository as an unweighted undirected network.
- PhosphoSitePlus Kinase-Substrate (PSP-KS) dataset[49] contains 13,664 Kinase-phosphosite pairings. Taken from source (PhosphoSitePlus) and parsed by us as an unweighted undirected heterogeneous bipartite network. Filtered only to contain pairs observed in Human tissue.
- TRRUST[43] (Transcriptional Regulatory Relationships Unraveled by Sentence based Text mining) dataset contains 3,149 transcription-factor relationships between 1,621 genes. Taken from source and parsed by us as

an unweighted undirected network. Filtered only to contain Activation relationships (as opposed to repression or unknown).

- Drugbank[61] Drug-Drug Interaction dataset contains 242,027 interactions between 2,191 drugs. Taken from BioNEV repository as an unweighted undirected network.
- NDFRT is a Disease-drug association dataset containing 56,515 associations between 13,545 diseases and drugs. Taken from BioNEV repository as an unweighted undirected heterogeneous bipartite network.
- CTD[27] is a Disease-drug association dataset containing 92,813 associations between 12,765 diseases and drugs. Taken from BioNEV repository as a pre-processed unweighted undirected heterogeneous bipartite network.

4.6 Supplementary Figures



Association of Deepwalk embedding dimensions with node degree-based predictor (pref. attachment)

Correlation with degrees vs. Predictivity of the embeddings

This suggests that the predictivity of the embeddings is not because of their correlation with degree info

Correlation of the embeddings with degree info is negligible (compared to their predictivity)

Partial Predictivity (given degree info) vs. Predictivity

They are mostly around the line (predictivity is unaffected)

This suggests that predictivity of the embedding dimensions does not stem from degree information

Embedding dimensions
x=y line
Squared correlation with degree predictor
Squared correlation with edge labels
(Predictivity)
Squared partial correlation with edge labels
(given degree predictor)
0
0.005
0.01
0.015
0.02
0.025

Figure 4.7. Investigating the embedding dimensions of Deepwalk in terms of their association with node degrees. The analysis suggests that the embeddings of Deepwalk does not depend on the degree information.

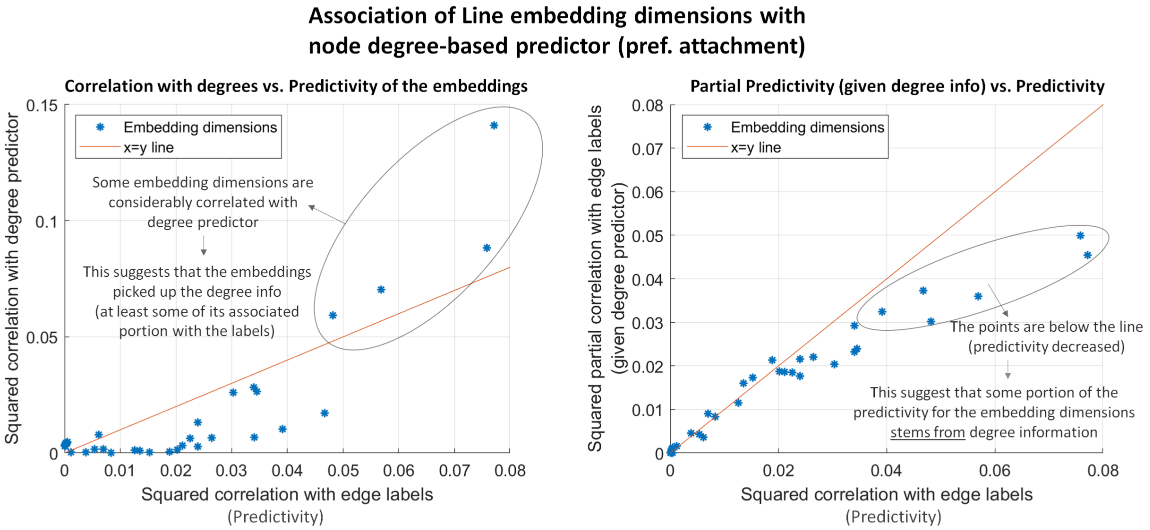


Figure 4.8. Investigating the embedding dimensions of Line in terms of their association with node degrees. The analysis suggests that Line embeddings picked up the node degree information and the predictivity of some of its embeddings dimensions stems from their correlation with node degrees.

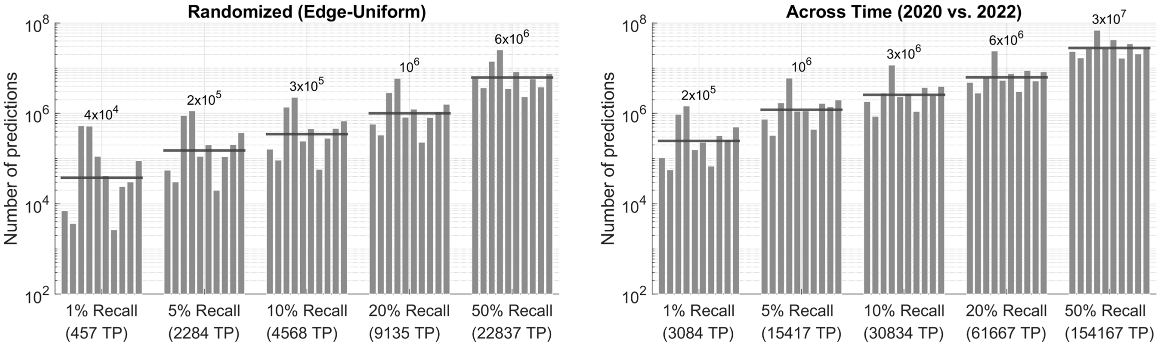


Figure 4.9. Number of predictions required to reach a particular recall threshold for Biogrid PPI predictions. (Left) The randomized (edge-uniform) evaluation. (Right) The across time evaluation setting. For both panels, the bars represent different link prediction algorithms. The horizontal lines and the numbers on the top indicate the geometric average of the number of predictions for each recall threshold.

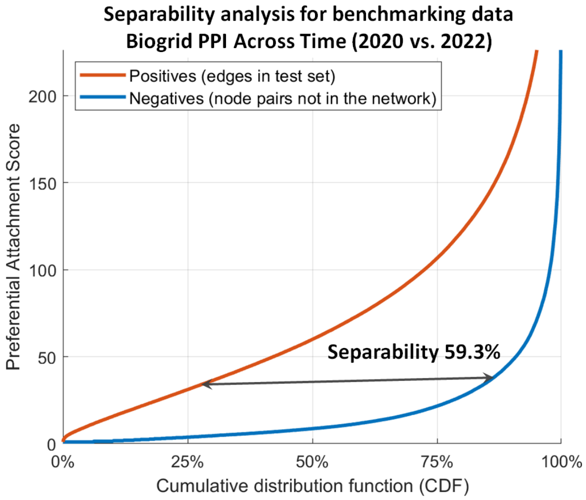


Figure 4.10. Separability analysis investigating the informedness of node degree information for Biogrid PPI across time (2020 vs. 2022) setting.

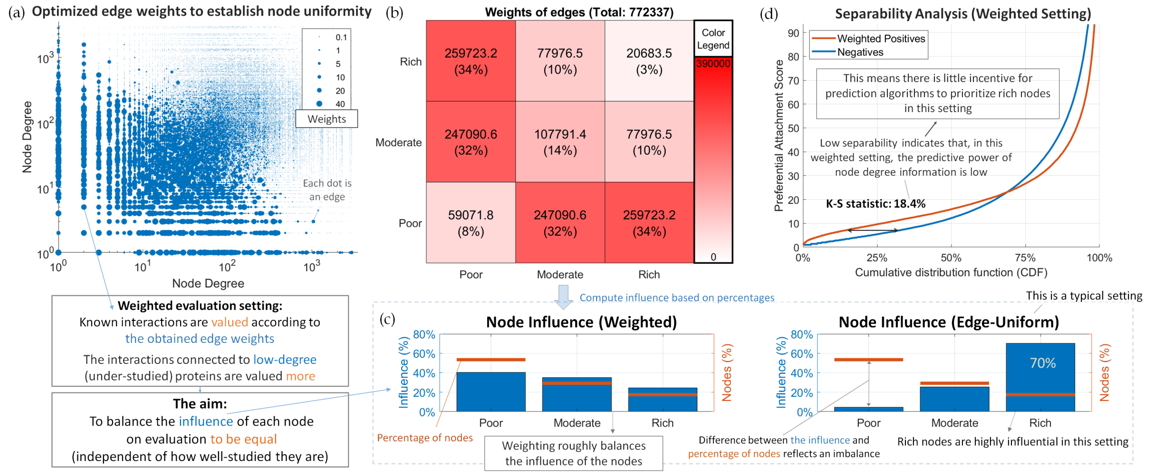
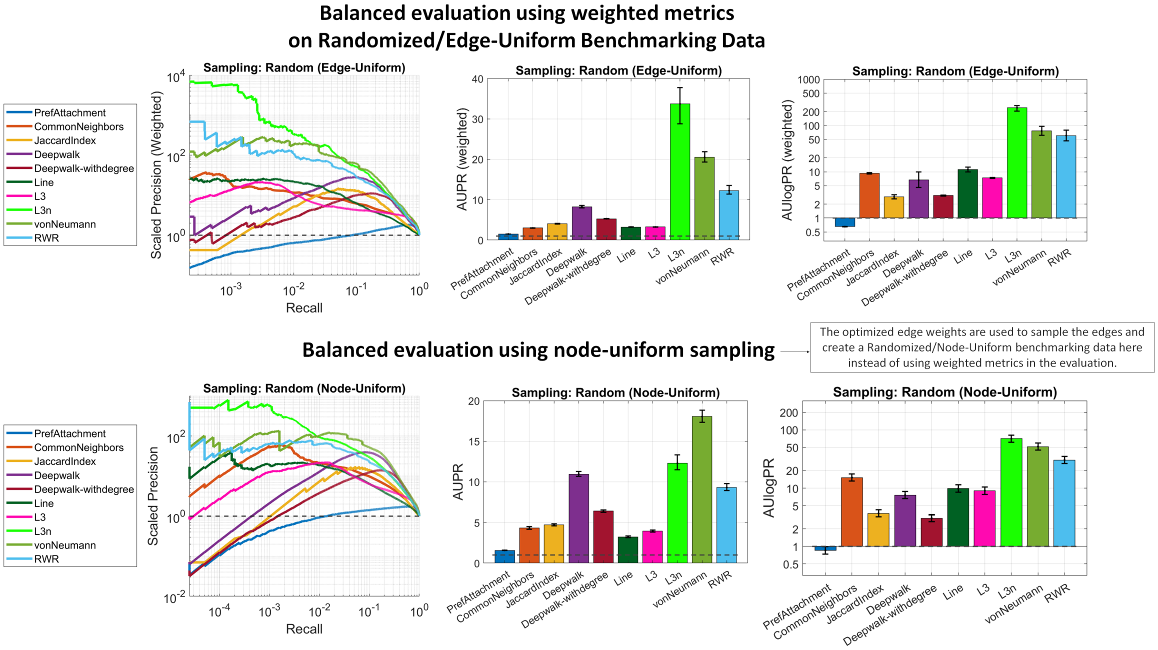


Figure 4.11. Mitigating degree bias in the evaluation of link prediction algorithms by assigning weights to edges during evaluation. Assignment of optimized edge weights establishes node-uniformity and balances the influence of nodes on evaluation. (a) Visualization of the optimized edge weights with respect to the degrees of incident nodes. The size of each point reflects the assigned weight of the corresponding edge. (b) The total weight of the edges by node category. (c) Influence of the nodes on evaluation (shown as bars) with respect to the node categories for weighted (node-uniform) and unweighted (edge-uniform) settings. (c) Separability analysis for the weighted (node-uniform) evaluation setting.



The optimized edge weights are used to sample the edges and create a Randomized/Node-Uniform benchmarking data here instead of using weighted metrics in the evaluation.

Figure 4.12. Balanced/Weighted evaluation results on randomized (sampled) benchmarking data for Biogrid PPI predictions. (Top) Balanced evaluation using weighted metrics. (Bottom) Balanced evaluation via node-uniform sampling (using the weights as sampling probabilities)

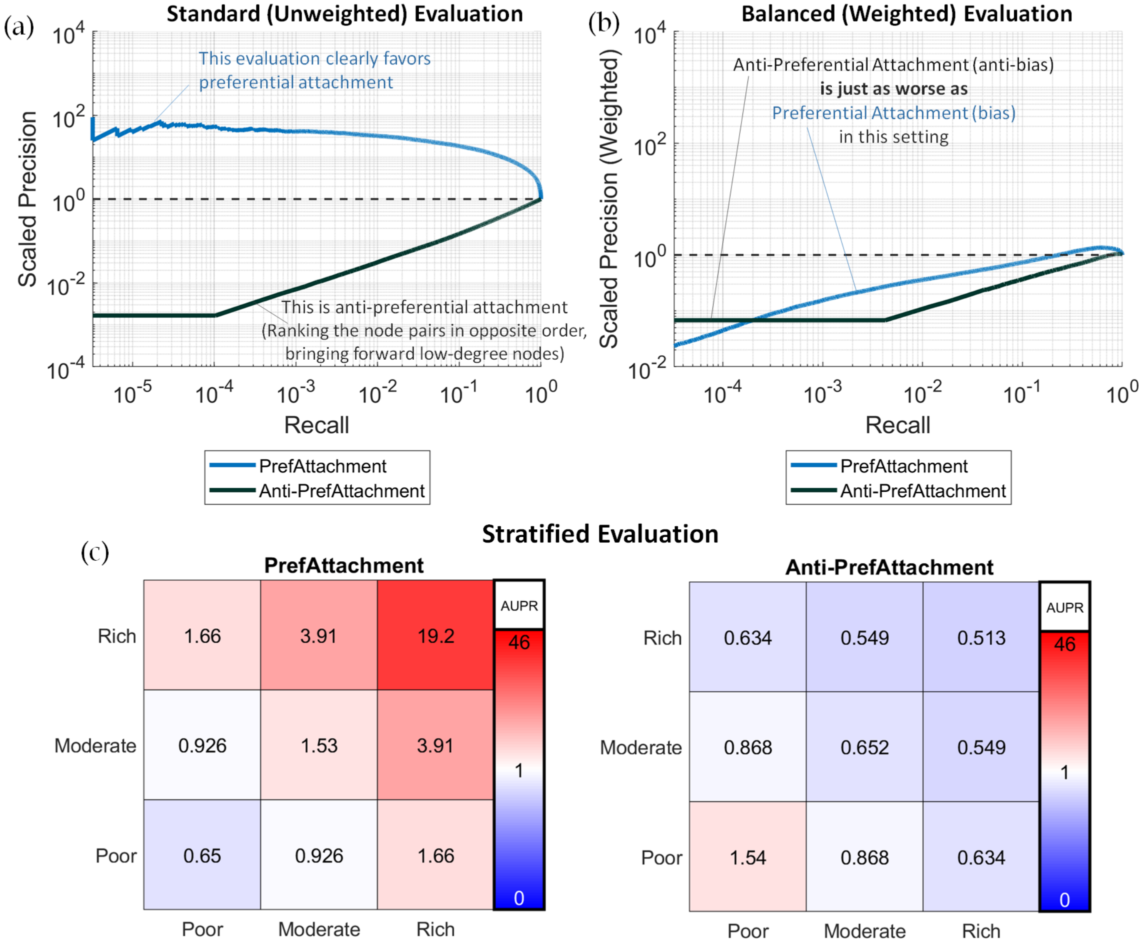
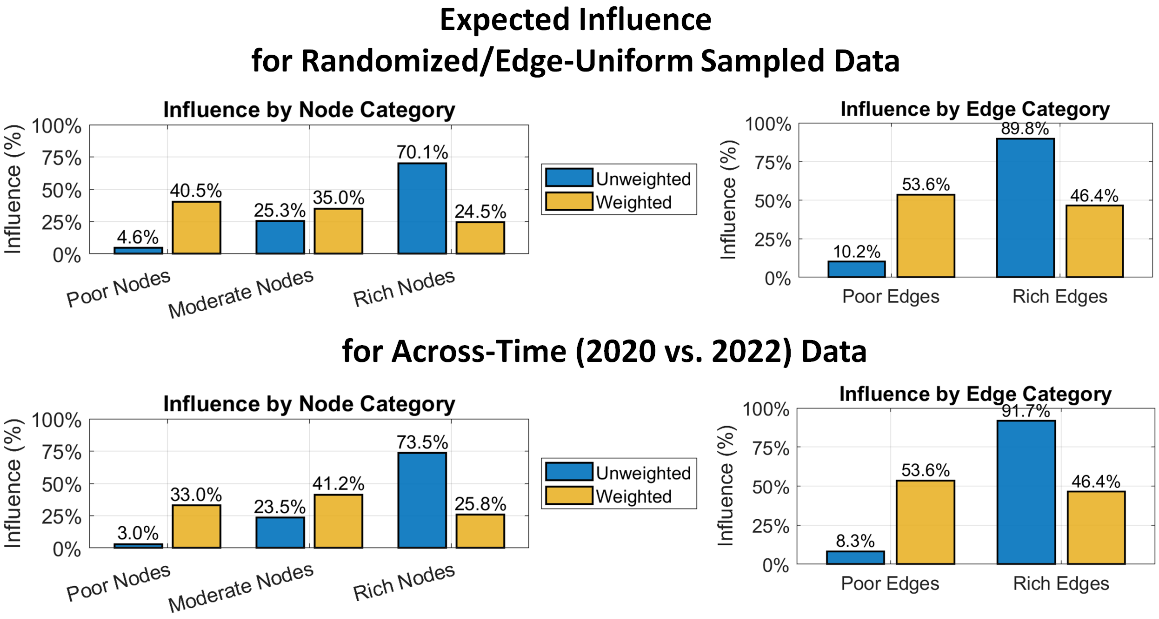
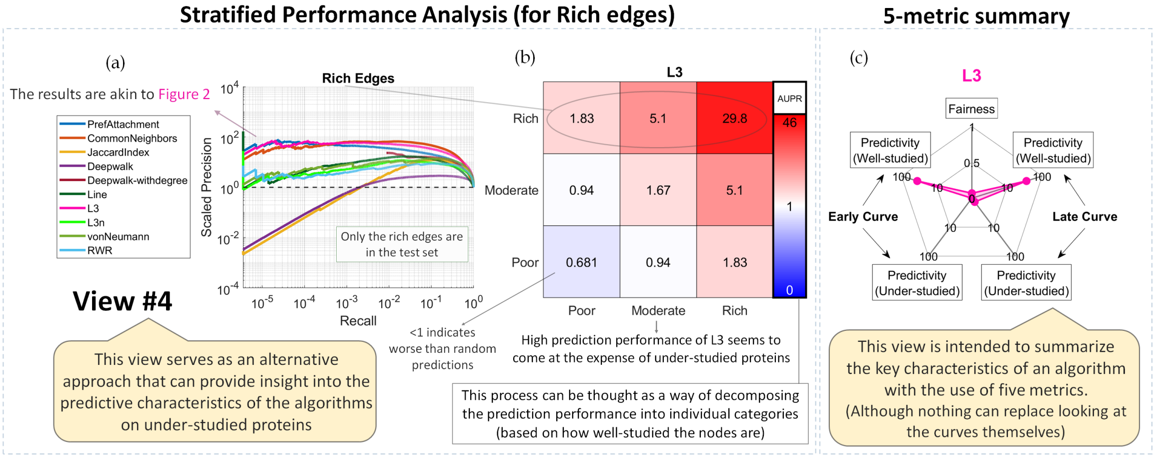


Figure 4.13. Comparison of preferential attachment (biased baseline) and anti-preferential attachment (anti-biased baseline) in different evaluation settings on Biogrid PPI predictions. Across-time (2020 vs. 2022) snapshots of the network are used as the benchmarking data (i.e., train/test splits) in this analysis. (a & b) Precision-Recall curves for the preferential attachment and anti-preferential attachment models respectively for standard (unweighted) and balanced (weighted) evaluation settings. (c) Stratified performance analysis results for preferential attachment and anti-preferential attachment algorithms. Each cell indicates the prediction performance of the algorithms for the corresponding edge category (e.g., for Poor-Rich category, only the edges that are between poor and rich nodes are included in the test set).


Expected Influence
for Randomized/Edge-Uniform Sampled Data

for Across-Time (2020 vs. 2022) Data

Figure 4.14. Expected influence for different categories of nodes or edges based on node degrees for randomized/edge-uniform and across-time bencharking data.


Stratified Performance Analysis (for Rich edges)

5-metric summary

The results are akin to Figure 2

View #4

This view serves as an alternative approach that can provide insight into the predictive characteristics of the algorithms on under-studied proteins

Only the rich edges are in the test set

<1 indicates worse than random predictions

High prediction performance of L3 seems to come at the expense of under-studied proteins

This process can be thought as a way of decomposing the prediction performance into individual categories (based on how well-studied the nodes are)

This view is intended to summarize the key characteristics of an algorithm with the use of five metrics. (Although nothing can replace looking at the curves themselves)

Figure 4.15. Stratified performance analysis for Rich edges connected to well-studied nodes and the 5-metric summary for the best performing method on rich edges. (a) Precision-Recall performance curves for rich edges in log-log scale. (b) Late curve prediction performance (AUPR) stratified by node categories for L3 algorithm. (c) 5-metric summary for L3.

5 Making Proteomics Accessible: RokaiXplorer for interactive analysis of phospho-proteomic data

5.1 Introduction

In the field of proteomics and phospho-proteomics, there is a growing need for user-friendly tools that enable researchers to analyze and visualize data with minimal training. To address this need and make proteomics data analysis easily accessible to researchers without expertise in computer and data sciences, we introduce RokaiXplorer, a comprehensive framework for performing exploratory analysis on proteomic and phosphorylation data in an interactive environment (Figure 5.1).

RokaiXplorer offers a range of functionalities that operate at five levels: Phosphosite, Phospho-protein, Protein expression, Pathway Enrichment, and Kinases. It allows for the identification of significant dysregulation at each level and presents the top findings through various visualizations, including volcano plots, heatmaps, bar plots, tables, and a network view (Figure 5.2). One of the distinguishing features of RokaiXplorer is its interactivity, which enables users to click on selected items in the visualizations to access an Inspection Window. This window provides comprehensive information about the selected items, including the source of evidence for dysregulation, quantifications and raw data for all samples.

Getting started with RokaiXplorer is straightforward and user-friendly. The application provides an interactive tutorial that guides users through the initial steps, making it easy to familiarize themselves with the tool. To begin using RokaiXplorer,

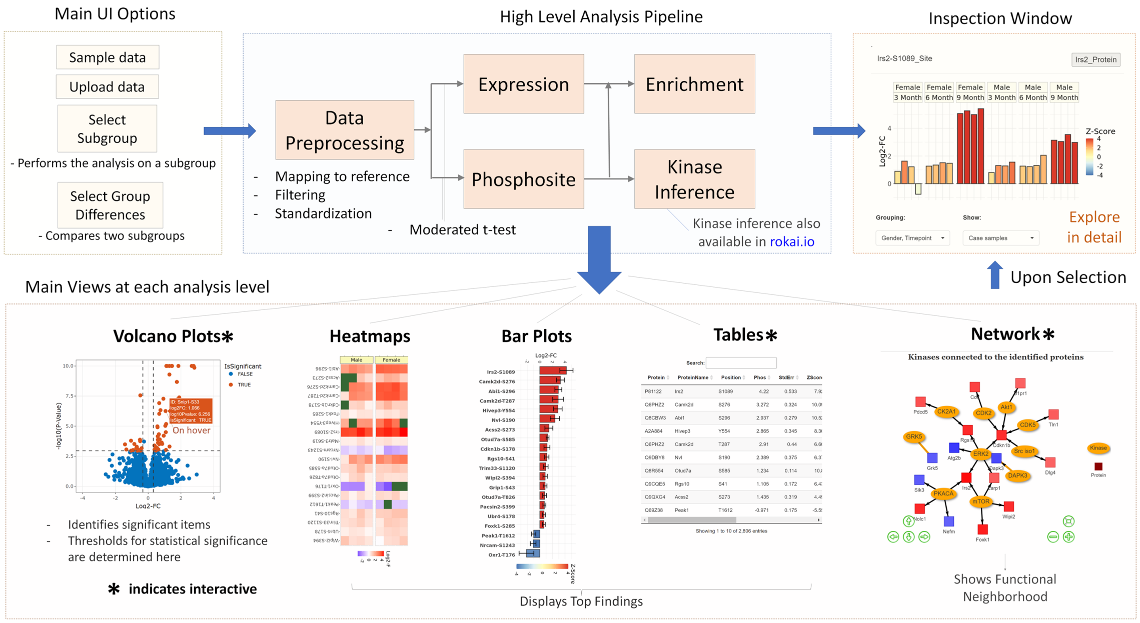


Figure 5.1. The workflow and key idea of RokaiXplorer.

users only need two types of input data: quantification data and meta data. The quantification data is a .csv file that contains the phosphorylation levels of each phosphosite/peptide, with each row representing a specific site and the columns containing quantification values for multiple samples. The meta data file complements the quantification data by providing additional information about the samples, such as their grouping. The main group field, which is mandatory, specifies the case/control status of the samples, while optional additional groups can be utilized to focus the analysis on specific subgroups if desired. RokaiXplorer also supports the input of protein expression data, enhancing its versatility for comprehensive analyses.

5.2 Results and Main Features

RokaiXplorer offers a comprehensive suite of modules that enable researchers to perform various analyses on their datasets. One of its key functionalities is dataset normalization, which ensures accurate and reliable comparisons between samples. By normalizing the data, RokaiXplorer minimizes potential biases and enhances the statistical power of subsequent analyses.

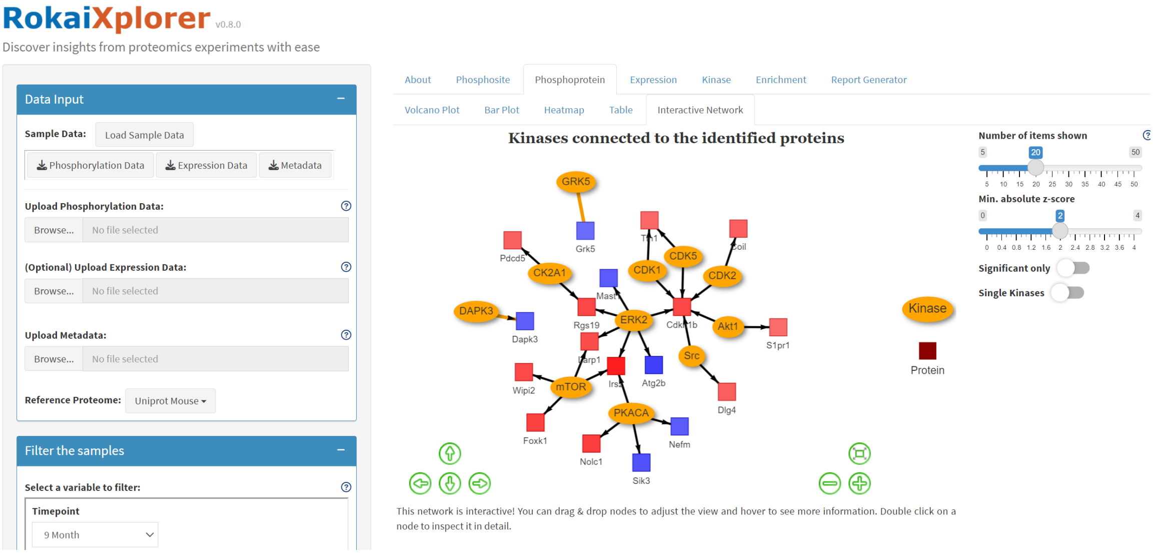
In addition to dataset normalization, RokaiXplorer facilitates the identification of dysregulated proteins, peptides, and potential biomarkers. It employs statistical tests, such as moderated t-tests, to assess dysregulation with precision, aiding in the identification of molecular signatures associated with specific conditions or diseases. Additionally, RokaiXplorer extends its analysis beyond individual proteins by providing insights into kinase activities. Through the utilization of the RoKAI algorithm, it infers kinase activities based on observed dysregulation patterns of phosphosites, contributing to a deeper understanding of cellular processes and signaling pathways.

To uncover the biological context of the dysregulated phospho-proteomic profiles, RokaiXplorer incorporates enrichment analysis of gene ontology (GO) terms. By identifying over-represented biological processes and molecular functions, researchers can gain valuable insights into the functional implications of their data. The GO enrichment analysis is performed using a chi-squared test with Yate's correction, ensuring reliable and statistically significant results.

In addition, RokaiXplorer goes beyond traditional data analysis tools by offering a robust Report Generator feature. This feature simplifies the analysis of data for multiple subgroups and streamlines the process of exporting results as formatted Excel tables. With the Report Generator, researchers can effortlessly investigate the impact of variables such as gender or tissue type by performing separate analyses for each subgroup of interest. By selecting the desired analysis type and defining grouping variables, users can generate customized reports with a single click. This convenient and user-friendly functionality enhances the efficiency of data analysis and facilitates the dissemination of research findings.

5.2.1 Customization options in RokaiXplorer

Customization options in RokaiXplorer go beyond simple data processing and analysis. The tool provides researchers with the ability to tailor their analysis to specific subgroups and species, allowing for a more targeted investigation. By utilizing



Sample Data:

Upload Phosphorylation Data: No file selected

(Optional) Upload Expression Data: No file selected

Upload Metadata: No file selected

Reference Proteome: Uniprot Mouse

Select a variable to filter:

Timepoint: 9 Month

Number of items shown: 20

Min. absolute z-score: 2

Significant only

Single Kinases

This network is interactive! You can drag & drop nodes to adjust the view and hover to see more information. Double click on a node to inspect it in detail.

Figure 5.2. A snapshot of the user interface for RokaiXplorer as of v0.8.0. The active tab in the figure displays the results of interactive network visualization for phospho-protein analysis.

these options, researchers can focus their analysis on particular subpopulations or target organisms of interest.

One of the customization features is the ability to filter the analysis based on specific subgroups. For instance, users can select a subgroup such as "Gender → Male" to restrict the analysis to male samples only. By filtering the dataset in this manner, researchers can obtain results that are specific to the chosen subgroup, allowing for subgroup-specific insights and comparisons.

Additionally, RokaiXplorer supports multiple reference proteomes, accommodating various species of interest. The tool currently provides reference proteomes for Human (Homo sapiens), Mouse (Mus musculus), and Rat (Rattus norvegicus). Researchers working with proteomic data from these species can leverage the respective reference proteomes to enhance the accuracy and relevance of their analyses. This species-specific customization ensures that the results obtained from RokaiXplorer are aligned with the biological context and characteristics of the target organisms.

Furthermore, RokaiXplorer enables researchers to explore group differences between two subgroups. This functionality is particularly valuable for comparative studies, where researchers want to investigate the patterns or dysregulation profiles between two specific groups. By selecting the desired subgroups for comparison, users can gain a deeper understanding of the molecular distinctions and uncover potential biomarkers or targets that are unique to each subgroup.

5.2.2 Interactive data browser: Share your discoveries feature

To foster collaboration and knowledge sharing, RokaiXplorer offers the "Share Your Discoveries" feature, which enables researchers to deploy their own interactive applications showcasing their data and analysis results. With this feature, the applications can be accessed online with the user data and settings already preloaded, allowing collaborators and other researchers to explore the data and gain valuable insights. This feature can enhance the impact of proteomic discoveries and facilitates interdisciplinary collaborations.

Deploying RokaiXplorer with preloaded input data is a straightforward process. Researchers can easily prepare and deploy their applications by following a few steps using the provided R scripts in the Github repository [*]. These steps include installing R, RStudio, and Rtools (for Windows users), creating an RStudio project, downloading the RokaiXplorer source code, and installing the required R libraries. Once the setup is complete, researchers can run RokaiXplorer in deployment mode, customize the application for their specific data and configuration, and make modifications to the application's title, descriptions, and about page. Additionally, RokaiXplorer allows users to export configuration files, enabling them to set desired analysis parameters for the online application and ensure reproducibility of results. Finally, researchers can deploy their application to shinyapps.io, a popular platform for hosting and sharing Shiny applications. By setting up a shinyapps.io account, connecting it to RStudio, and deploying the application, researchers can freely and effortlessly share their interactive RokaiXplorer application with others through a unique link, making their findings accessible to a wider audience.

[*] https://github.com/serhan-yilmaz/RokaiXplorer

5.2.3 Example application on Alzheimer's disease

In this study, we utilize the capabilities of RokaiXplorer to analyze proteome and phospho-proteome data from a mouse hippocampus tissue study on Alzheimer's disease (AD). The data includes variables such as time (3, 6, and 9 months), sex (male and female), and genetic background (5XFAD versus wild type), which correspond to specific AD phenotypes such as Aβ42 plaque deposition, memory deficits, and neuronal loss. By applying RokaiXplorer, we aimed to explore the temporal and sex-linked variations in AD, focusing on biomarker discovery and the identification of potential clinical targets.

This study involved various analyses to understand the phosphoproteome changes in the hippocampus of 5XFAD mice during Alzheimer's Disease progression. We investigated the temporal and sex-linked patterns in phosphorylation, aiming to estimate the disease burden and identify trends over time, followed by a statistical analysis to identify specific dysregulated phosphopeptides between the WT and 5XFAD mice groups. In addition, we compared phosphorylation patterns to protein expression levels to assess the complementarity of phosphorylation to protein expression.

In addition, we identified consistent phosphoproteins that could potentially serve as markers for Alzheimer's Disease, and investigated regulatory mechanisms involved in phosphorylation events through kinase inference analysis.

Finally, pathway enrichment analysis was conducted to understand the biological pathways and networks impacted by the observed phosphoproteome changes. Together, these analyses provided a multi-faceted approach to uncovering the complex dynamics of phosphorylation and its implications in Alzheimer's Disease progression. To facilitate the interpretation of our findings and promote free exploration of the data and results by other researchers, we utilized the RokaiXplorer application to develop the interactive tool AD-Xplorer. The findings are presented in the form of a live data browser with intact analysis capabilities, which can be accessed online at: https://yilmazs.shinyapps.io/ADXplorer/

In the browser, the analysis can further be specified to focus on a particular subgroup. For example, selecting "9 Month" and "Female" on the left panel displays the findings for that group by performing the analysis after filtering the samples that fit the criteria. All presented capabilities are made generic and can be readily applied on other datasets abd studies, including the deployment of the dataset online as a live browser. The groups to customize the analysis are specified in a metadata file and deploying a dataset online only requires a configuration file (that can be generated via the online interface), a markdown file (to specify the descriptions on the front page) and the input data files. RokaiXplorer supports data from all proteomics quantification methods (e.g., label-free, SILAC, isobaric labeling).

Overall, we anticipate that RokaiXplorer will be an appreciated tool in the community to analyze phospho-proteomic data because of its simplicity and speed, enabling the analysis of data at different levels in one application. RokaiXplorer is available at: http://explorer.rokai.io

5.3 Methods

5.3.1 Data input

The input data required for RokaiXplorer consists of two data files and one additional data file which is optional. The following provides detailed information on the data formats for each file:

- **Phosphorylation Data:**

  The phosphosite quantification data should be provided in CSV format. The file should contain the following columns:

  - **Protein (first column):** This column should contain the Uniprot protein identifier.
  - **Position (second column):** This column specifies the position of the modified phosphosite on the protein.
  - **Samples (multiple columns):** Each column represents a sample, and the values in each column indicate the phosphorylation intensity

of the corresponding phosphosite for that sample. The intensities should not be log-transformed, as this step is performed within the application.

- **Metadata:**

  The metadata file should also be in CSV format and contain the following information:

  - **RowName (first column):** This column provides the name of the group specifier.
  - **Samples (multiple columns):** Each column represents a sample, and the values in each column indicate the group identity for that sample.
  - **Group (first row):** This row is necessary and specifies the main group that determines the case/control status of the samples.
  - **Other Groups (multiple rows):** You can use the optional rows in the metadata file to specify additional groups for the samples. These additional groups allow you to filter the samples and focus the analysis on a particular subgroup of interest.

  Please ensure that your metadata file is in CSV format. The main group, which determines the case/control status, is required, while other group specifications are optional.

- **Expression Data (Optional):**

  If available, you can include protein expression data in CSV format. The file should have the following columns:

  - **Protein (first column):** This column contains the Uniprot protein identifier.
  - **Samples (multiple columns):** Each column represents a sample, and the values in each column indicate the expression intensity of the corresponding protein for that sample. The intensities should not be log-transformed, as this step is performed within the application.

  Including protein expression data is optional, but if provided, it should be in CSV format.

5.3.2 Data preprocessing

Notation. Let $\boldsymbol{V} \in \mathbb{R}^{n \times m}$ denote the input data matrix for phosphorylation, where the rows denote phosphosites and the columns denote the samples, and let $V[i, j]$ refer to an entry of this matrix corresponding to phosphosite i and sample j. Let $\boldsymbol{V}_{case} \in \mathbb{R}^{n \times m_{case}}$ and $\boldsymbol{V}_{ctrl} \in \mathbb{R}^{n \times m_{ctrl}}$ denote the subsets of this matrix correspond to case and control samples respectively, having n phosphosites, m_{case} case samples and m_{ctrl} control samples.

Optional step: Filtering samples for a subgroup. If desired, the users have the option to filter the samples (columns) to narrow down the analysis to a specific subgroup. This step is carried out before any other analysis steps. By doing this, only the data related to the selected subgroup will be used for the analysis. It is essentially equivalent to excluding the data for other subgroups from the input altogether.

Log-transformation and normality assumption. As a first step of preprocessing, we apply a log transformation on the quantification matrix $\boldsymbol{V}$ to make sure that it approximately follows a normal distribution:

$$\tilde{\boldsymbol{V}} = \log_2 \boldsymbol{V} \tag{5.1}$$

where $\tilde{\boldsymbol{V}}$ the matrix after the transformation. After the transformation, we assume that each column $\boldsymbol{v}[:, j]$ of $\tilde{\boldsymbol{V}}$ follows a normal distribution $\mathcal{N}(\mu_j, \sigma_j)$.

Optional step: Centering for variance stabilization. As an optional step after the log transformation, we center each sample (column) to have 0 mean value by substracting the sample means $\hat{\mu}_j$:

$$\begin{gathered} \tilde{\boldsymbol{v}}[:, j] = \boldsymbol{v}[:, j] - \hat{\mu}_j \\ \hat{\mu}_j = \sum_{i:1}^{n} \frac{\boldsymbol{v}[i, j]}{n} \end{gathered} \tag{5.2}$$

Note that, the missing values are omitted during the computation of the sample mean $\hat{\mu}_j$. This step is enabled by default and is recommended to balance out potential systematic differences that may occur between the samples

5.3.3 Statistical inference at phosphosite level

As a first to identify phosphosites that are significantly different between the case and control samples, we compute the fold changes. Let $q[i]$ denote the log fold change for phosphosite i, which is equal to the following:

$$q[i] = \sum_{j:1}^{m_{case}} \frac{\boldsymbol{v}_{case}[i,j]}{m_{case}} - \sum_{j:1}^{m_{ctrl}} \frac{\boldsymbol{v}_{ctrl}[i,j]}{m_{ctrl}} \tag{5.3}$$

Optional step: Centering the fold changes. As alternative approach to balance out any potential systematic bias between the case and control groups, the option to center the log fold changes are provided by subtracting the mean across all phosphosites:

$$\begin{aligned} \tilde{q}[i] &= q[i] - \mu_z \\ \mu_z &= \sum_{i}^{n} \frac{q[i]}{n} \end{aligned} \tag{5.4}$$

Statistical tests. To determine the statistical significance of the log fold changes $\boldsymbol{q}$, we consider various models that differ in how the standard errors are estimated.

Z-test: The first and the simplest option is to estimate the standard errors based on the standard deviation across the phosphosites and perform a z-test based on this.

Let s_z be sample standard deviation of $\boldsymbol{q}$ across all phosphosites. Here, if we assume the standard error $\sigma[i]$ of each phosphosite to be the same and equal to s_z, the z-score $z[i]$ for each phosphosite follows a normal distribution:

$$z[i] = \frac{q[i]}{s_z} \tag{5.5}$$

Based on the inverse normal distribution and the z-scores $\boldsymbol{z}$, the corresponding p-values are computed to the statistical significance. In addition, Benjamini-Hochberg[11] procedure is applied to limit the false discovery rate (FDR) of the findings.

Note that, this test is the simplest option with the weakest assumptions. It should only be applied in cases where the number of samples for each (case or control) group are too low and the standard deviations cannot be measured accurately (e.g., when is only a single sample). Otherwise, a t-test is more appropriate.

Pooled t-test: The second option is to estimate the standard error based on the standard deviation across the samples and perform a pooled t-test based on this.

Let $s_{case}[i]$ and $s_{ctrl}[i]$ be sample standard deviations estimated across the samples (columns) for each phosphosite i, and let $s_{pooled}[i]$ denote the pooled standard deviation for phosphosite i which is given as follows:

$$s_{pooled}[i] = \frac{(m_{case}[i] - 1)\, s^2_{case}[i] + (m_{ctrl}[i] - 1)\, s^2_{ctrl}[i]}{m_{case}[i] + m_{ctrl}[i] - 2} \tag{5.6}$$

where $m_{case}[i]$ and $m_{ctrl}[i]$ represent the number of case/control samples with quantifications (i.e., having non-missing data) for phosphosite i. Assuming normality, independence between the samples, and equal variances between two groups (i.e., case and control), the t-statistic $t[i]$ for phosphosite i follows a t-distribution with degrees of freedom $df[i]$ such that:

$$\begin{aligned} t[i] &= \frac{q[i]}{\sigma[i]} \\ \sigma[i] &= s_{pooled}[i]\sqrt{\Big(\frac{1}{m_{case}[i]} + \frac{1}{m_{ctrl}[i]}} \\ df[i] &= m_{case}[i] + m_{ctrl}[i] - 2 \end{aligned} \tag{5.7}$$

and the statistical significance and p-values are computed accordingly based on the t-distribution. Note that, this test requires at least $m_{case} \geq 2$ and $m_{ctrl} \geq 2$ to be performed.

Moderated t-test: If desired, as a potential improvement to pooled t-test, the moderated t-test[120] can be performed, which utilizes an empirical Bayes method to shrink the pooled sample variances towards a common value and to augment the degrees of freedom for the individual variances. For this purpose, we utilize the implementation in limma package of R[114]. Specially, we utilize the *SquuezeVar* function which takes the pooled standard deviations $s_{pooled}[i]$ as input and returns the moderated standard deviations $s_{mod}[i]$ and the extra degrees of freedom gained

df_{ext}. Based on these, the moderated t-test is performed:

$$\begin{aligned} \tilde{t}[i] &= \frac{q[i]}{\tilde{\sigma}[i]} \\ \tilde{\sigma}[i] &= s_{mod}[i]\sqrt{\left(\frac{1}{m_{case}[i]} + \frac{1}{m_{ctrl}[i]}\right.} \\ \tilde{df}[i] &= m_{case}[i] + m_{ctrl}[i] - 2 + df_{ext} \end{aligned} \tag{5.8}$$

where the moderated t-statistic $\tilde{t}[i]$ follows a t-distribution having $\tilde{df}[i]$ degrees of freedom.

Defaults: By default, if there are at least $m_{case} \geq 2$ and $m_{ctrl} \geq 2$ samples, a moderated t-test is performed. If that is not case, or if the analysis is to performed for a single sample (e.g., for heatmaps), a z-test is performed.

To ensure generalization and to simplify notation in the following sections, we will adopt the assumption that a t-test is performed. Additionally, we will consider the z-test as a special case of the t-test, where the parameters $\sigma[i]$ are represented as s_z and the degrees of freedom $df[i]$ are treated as ∞.

5.3.4 Statistical inference at phospho-protein level

After assessing the significance of phosphosites, we combine their results to perform statistical inference at the protein level, again comparing the case samples with the control samples. Let $q[i]$ be the resultant log2 fold change, and $\sigma[i]$, $df[i]$ be the corresponding standard error and degrees of freedom obtained from t-test for phosphosite i.

To perform the inference at the protein level, we first compute the mean log-fold changes $q_p[j]$ for each protein j:

$$q_p[j] = \frac{\sum_{i \in \mathcal{V}_j} q[i]}{\left(\mathcal{V}_j\right|} \tag{5.9}$$

where $\mathcal{V}_j$ denotes the set of phosphosites corresponding to protein j.

To estimate the pooled standard error $\sigma_p[j]$ and the corresponding degrees of freedom $df_p[j]$ in the estimation of the mean log-fold changes for each protein j, we use the Satterthwaite approximation[117]:

$$\begin{aligned} \sigma_p[j] &= \frac{\sqrt{\sum_{i\in\mathcal{V}_j} \sigma^2[i]}}{|\mathcal{V}_j|} \\ df_p[j] &= \frac{\left(\sum_{i\in\mathcal{V}_j} \sigma^2[i]\right)^2}{\sum_{i\in\mathcal{V}_j}\left(\frac{\sigma^4[i]}{df[i]}\right)} \end{aligned} \tag{5.10}$$

Based on these estimations, to compute the significance of a protein j, a t-test is performed with the t-statistic $t_p[j]$:

$$t_p[j] = \frac{q_p[j]}{\sigma_p[j]} \tag{5.11}$$

which follows a t-distribution with $df_p[j]$ degrees of freedom under the null hypothesis.

5.3.5 Optional: Statistical inference for protein expression

The statistical inference for protein expression follows the same methodology employed in the phosphosite level analysis, which includes pooled/moderated t-tests or z-tests as described in the *Statistical inference at phosphosite level* section. The key difference is that, in this case, the analysis is performed at the protein level instead of the phosphosite level.

5.3.6 Statistical inference at kinase level

We use the notation $\boldsymbol{W}_{ks}$ to represent the kinase-substrate network, which consists of interactions between n_{kin} kinases and n phosphosites. We obtain this network from either the PhosphoSitePlus[49] or Signor[76] databases. Typically, this network is sparse, with a value of 1 in the entry $w_{ks}[i, j]$ indicating that kinase i targets phosphosite j.

To identify dysregulated kinases that exhibit significant differences between case and control samples, we employ two approaches for inferring kinase activities. The first approach is a simple one, involving the calculation of mean substrate phosphorylation. This approach considers the phosphorylation (log-FC) of the known targets of a kinase and takes the mean value as the inferred activity of that kinase. In contrast, the RoKAI algorithm is a more comprehensive approach that utilizes a functional network to enhance the accuracy and robustness of kinase activity inference[145].

Mean substrate phosphorylation (without RoKAI). To perform the kinase activity inference based on the mean substrate phosphorylation, we first compute the mean log-fold changes $q_k[i]$ for each kinase i:

$$q_k[i] = \frac{\sum_{j \in \boldsymbol{w}_{ks}[i,:]} q[j] - \mu_z}{|\boldsymbol{w}_{ks}[i,:]|} \tag{5.12}$$

where $\boldsymbol{w}_{ks}[i,:]$ denotes the set of phosphosites that are known targets of kinase i and μ_z denotes the mean log fold change across all phosphosites (see Equation (5.4)).

To estimate the pooled standard error $\sigma_k[i]$ and the corresponding degrees of freedom $df_k[i]$ in the estimation of the mean log-fold changes for each kinase i, we use the Satterthwaite approximation[117]:

$$\begin{aligned} \sigma_k[i] &= \frac{\sqrt{\sum_{j \in \boldsymbol{w}_{ks}[i,:]} \sigma^2[j]}}{|\boldsymbol{w}_{ks}[i,:]|} \\ df_k[i] &= \frac{\left(\sum_{j \in \boldsymbol{w}_{ks}[i,:]} \sigma^2[j]\right)^2}{\sum_{j \in \boldsymbol{w}_{ks}[i,:]} \left(\frac{\sigma^4[j]}{df[j]}\right)} \end{aligned} \tag{5.13}$$

Based on these estimations, to compute the significance of a kinase j, a t-test is performed based on the t-statistic $t_k[i]$:

$$t_k[i] = \frac{q_k[i]}{\sigma_k[i]} \tag{5.14}$$

which follows a t-distribution with $df_k[i]$ degrees of freedom under the null hypothesis.

Inference of kinase activities using RoKAI algorithm. The RoKAI algorithm[145] is a method that propagates phosphorylation levels in a functional network. The network includes kinase-substrate associations, protein-protein interactions between kinases, and structure distance and co-evolution evidence for interactions between phosphosites. Using an electric circuit model, the algorithm transfers node potentials through the network using a conductance matrix $\boldsymbol{C}$. By solving a linear system, the algorithm computes node potentials, which enables the propagation of phosphorylation levels. The algorithm then infers kinase activities based on mean phosphorylation of known targets of the kinase using the propagated values.

Since RoKAI algorithm employs a linear model, the inferred activity of a kinase can be expressed as a weighted summation of phosphorylation levels (i.e., log fold changes) of known targets of a kinase, along with other phosphosites in the kinase's functional neighborhood. In this section, we discuss the statistical methods used to determine the significance of the inferred kinase activities based on RoKAI. The following section cover the process of obtaining the weights that express the underlying formula in the RoKAI algorithm.

Let $\boldsymbol{W}$ denote the weighting matrix between n_{kin} kinases and n phosphosites, where $w[i,j]$ represents the weight of phosphosite j in the inferred activity $q_a[i]$ of kinase i in the RoKAI inference such that:

$$q_a[i] = \sum_{j:1}^{n} w[i,j]q[j] \tag{5.15}$$

Expressing this in matrix form yields:

$$\boldsymbol{q}_a = \boldsymbol{W}\boldsymbol{q} \tag{5.16}$$

Note that, without loss of generality, we assume that the weights are scaled such that they add up to 1 for each kinase. This scaling ensures that the weights represent a weighted average. To estimate the standard error $\sigma_a[i]$ and the corresponding degrees of freedom $df_a[i]$ in the inferred activity $q_a[i]$ of kinase i, we use the Satterthwaite approximation:

$$\begin{aligned} \sigma_a[i] &= \sqrt{\sum_{j:1}^{n} w^2[i,j]\sigma^2[j]} \\ df_a[i] &= \frac{\left(\sum_{j:1}^{n} w[i,j]\sigma^2[j]\right)^2}{\sum_{j:1}^{n} \frac{w^2[i,j]\sigma^4[j]}{df[j]}} \end{aligned} \tag{5.17}$$

Expressing this in matrix form yields:

$$\begin{aligned}\boldsymbol{\sigma}_a &= \sqrt{\boldsymbol{W}^{\odot 2}\boldsymbol{\sigma}^{\odot 2}} \\ \boldsymbol{df}_a &= \left(\boldsymbol{W}\boldsymbol{\sigma}^{\odot 2}\right)^{\odot 2} \oslash \left(\boldsymbol{W}^{\odot 2}\left(\boldsymbol{\sigma}^{\odot 4} \oslash \boldsymbol{df}\right)\right)\end{aligned} \tag{5.18}$$

To clarify, the symbol $^{\odot k}$ represents the operation of taking the element-wise kth power (also known as the Hadamard power) of a matrix or vector. For example, $^{\odot 2}$ corresponds to the element-wise square operation. Similarly, $\oslash$ represents element-wise division.

After $\boldsymbol{q}_a$, $\boldsymbol{\sigma}_a$ and $\boldsymbol{df}_a$ are estimated, to assess the statistical significance of a kinase i, a t-test is performed based on the t-statistic $t_a[i]$:

$$\begin{aligned} t_a[i] &= \frac{q_a[i]}{\sigma_a[i]} \\ \boldsymbol{t}_a &= \boldsymbol{q}_a \oslash \boldsymbol{\sigma}_a \end{aligned} \tag{5.19}$$

which follows a t-distribution with $df_a[i]$ degrees of freedom under the null hypothesis.

Note that, although the open formulas (such as in Equation (5.17)) are provided for clarity, in the implementation, their matrix correspondences (such as in Equation (5.18)) are performed using efficient sparse matrix operations for improved computational performance.

Obtaining weights expressing the underlying formula for RoKAI inference. The RoKAI algorithm utilizes a heterogeneous network, denoted as $W_{in} \in \mathcal{R}^{n_{ks} \times n_{ks}}$, where the nodes represent kinases and/or phosphosites. The total number of nodes in the network is denoted as $n_{ks} = n_{kin} + n$, which is the sum of the number of kinases (n_{kin}) and the number of phosphosites (n). The edges in this network capture various functional associations between kinases, phosphosites, and their combinations. To propagate the phosphorylation values across this functional network, the RoKAI algorithm employs an electric circuit model and solves a system of equations, as described in the reference[145] and outlined below:

$$\boldsymbol{C}\boldsymbol{v} = \boldsymbol{b} \tag{5.20}$$

In the given equation, the matrix $\boldsymbol{C} \in \mathcal{R}^{n_{ks} \times n_{ks}}$ represents the conductance between nodes in the network, allowing a portion of phosphorylation to be transfered to nearby nodes in the form of current. The vector $\boldsymbol{b} \in \mathcal{R}^{n_{ks} \times 1}$ indicates the phosphorylation levels (log fold changes), while $\boldsymbol{v} \in \mathcal{R}^{n_{ks} \times 1}$ represents the node potentials to be computed. These node potentials $\boldsymbol{v}$ reflect the phosphorylation levels after the information has propagated through the network.

It is important to emphasize that not all nodes in the network are required to be quantified. Even nodes that do not have a computed fold change value, such as those with missing values in the experimental data, can still be retained in the network. In the case of an unquantified node denoted as i, its corresponding entry in the vector $\boldsymbol{b}$ is assigned a value of $b[i] = 0$ to indicate that it does not contribute to the fold change calculation and only kept as bridge node connecting other nodes.

The solution vector $\boldsymbol{v}$ of this system can be obtained efficiently using standard linear algebra solvers. However, in addition to finding the the solution vector $\boldsymbol{v}$, our goal here is to compute the weights $\boldsymbol{W}_p \in \mathcal{R}^{n_{ks} \times n_{ks}}$ that represent the underlying solution of this system such that:

$$\begin{aligned} \boldsymbol{v} &= \boldsymbol{W}_p \boldsymbol{b} \\ \boldsymbol{W}_p &= \boldsymbol{C}^{-1} \end{aligned} \tag{5.21}$$

Obtaining the weighting matrix $\boldsymbol{W}_p$ explicitly through the matrix inversion operation described above becomes computationally expensive, especially for typical large-scale networks with thousands or more nodes (phosphosites or kinases). This approach becomes even more challenging when working with limited computational resources, such as those available on a web server. To overcome this challenge, we have implemented several optimizations for the RoKAI algorithm to improve computational efficiency that involves computing a partial inverse. These optimizations ensure that the algorithm remains feasible and scalable, even for large-scale networks.

First, we introduce the concept of *relevant nodes*. These are phosphosites or kinases that have a functional annotation in the network, meaning they are associated with at least one edge in the functional network $\boldsymbol{W}_{in}$. We denote the number of relevant nodes as n_{rel}, and I_{rel} represents the indices of these nodes. To focus on the relevant nodes, we define $\hat{\boldsymbol{W}}_p$ as the subset of the network $\boldsymbol{W}_p$ that only includes the relevant nodes, represented by the indices I_{rel}. In other words, $\hat{\boldsymbol{W}}_p$ corresponds to the submatrix $\boldsymbol{W_p}[I_{rel}, I_{rel}]$, which has a size of $n_{rel} \times n_{rel}$. Similarly, we define $\hat{\boldsymbol{C}}$ as the subset of the conductance matrix that corresponds to the relevant nodes.

In addition to the concept of relevant nodes, we introduce the notion of *quantified nodes*. These nodes refer to the phosphosites that are explicitly identified in the dataset being processed. In other words, they are phosphosites for which a fold change value is computed (i.e., does not have a missing value). In general, we are only interested in computing the further subset of $\hat{\boldsymbol{W}}_p$ that involves the quantified nodes. We refer to this subset as $\hat{\hat{\boldsymbol{W}}}_p$ which is of size $n_q \times n_q$ where n_q is the number of nodes that are both quantified and relevant. The remaining nodes that are relevant but not quantified play a crucial role as bridges connecting other nodes in the network. While these are necessary for obtaining the solution of the system as they contribute to the overall connectivity and information flow, they need not be explicitly expressed in the final solution or underlying formula behind the inferred kinase activities.

We can partition the conductance matrix $\hat{\boldsymbol{C}}$ into 2×2 blocks based on on the quantification status of the nodes:

$$\hat{\boldsymbol{C}} = \begin{bmatrix} \hat{\boldsymbol{C}}_{11} & \hat{\boldsymbol{C}}_{12} \\ \hat{\boldsymbol{C}}_{21} & \hat{\boldsymbol{C}}_{22} \end{bmatrix} \Bigg(\tag{5.22}$$

In this partitioning scheme, we assign a label of "1" to the block corresponding to the quantified nodes, indicating their quantification status, and a label of "2" to the block representing the remaining nodes. Similarly, we can express the linear system in Equation (5.20) using the partioning as follows:

$$\hat{\boldsymbol{C}} \begin{bmatrix} \hat{\boldsymbol{v}}_1 \\ \hat{\boldsymbol{v}}_2 \end{bmatrix} \Bigg(= \begin{bmatrix} \hat{\boldsymbol{b}}_1 \\ \hat{\boldsymbol{b}}_2 \end{bmatrix} \Bigg(\tag{5.23}$$

Since the unquantified nodes in block "2" does not have computed fold changes, $\hat{\boldsymbol{b}}_2 = 0$ and the above equation can be rewritten as:

$$\hat{\boldsymbol{C}} \begin{bmatrix} \hat{\boldsymbol{v}}_1 \\ \hat{\boldsymbol{v}}_2 \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{b}}_1 \\ 0 \end{bmatrix} \tag{5.24}$$

What we are looking for here is a matrix $\hat{\hat{\boldsymbol{W}}}_p$, a partial inverse of $\hat{\boldsymbol{C}}$ that will satisfy:

$$\hat{\hat{\boldsymbol{W}}}_p \hat{\boldsymbol{b}}_1 = \hat{\boldsymbol{v}}_1 \tag{5.25}$$

Given that the block matrix $\hat{\boldsymbol{C}}_{11}$ is invertible, we can obtain a partial inverse of matrix $\hat{\boldsymbol{C}}$ by inverting $\hat{\boldsymbol{C}}_{11}$ and replacing the corresponding block $\hat{\boldsymbol{C}}_{22}$ with the Schur complement $\hat{\boldsymbol{C}}/\hat{\boldsymbol{C}}_{11}$ and adjusting the off-diagonal elements of the resulting matrix accordingly[132]:

$$\mathrm{inv}_1\hat{\boldsymbol{C}} = \begin{bmatrix} \mathrm{inv}_1\hat{\boldsymbol{C}}_{11} & \mathrm{inv}_1\hat{\boldsymbol{C}}_{12} \\ \mathrm{inv}_1\hat{\boldsymbol{C}}_{21} & \mathrm{inv}_1\hat{\boldsymbol{C}}_{22} \end{bmatrix} = \begin{bmatrix} \left(\hat{\boldsymbol{C}}_{11}\right)^{-1} & -\left(\hat{\boldsymbol{C}}_{11}\right)^{-1}\hat{\boldsymbol{C}}_{12} \\ \hat{\boldsymbol{C}}_{21}\left(\hat{\boldsymbol{C}}_{11}\right)^{-1} & \hat{\boldsymbol{C}}_{22} - \hat{\boldsymbol{C}}_{21}\left(\hat{\boldsymbol{C}}_{11}\right)^{-1}\hat{\boldsymbol{C}}_{12} \end{bmatrix} \tag{5.26}$$

This partial inversion corresponds to a rotation of the matrix and satisfies the following property[132]:

$$\mathrm{inv}_1\hat{\boldsymbol{C}} \begin{bmatrix} \hat{\boldsymbol{b}}_1 \\ \hat{\boldsymbol{v}}_2 \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{v}}_1 \\ \hat{\boldsymbol{b}}_2 \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{v}}_1 \\ 0 \end{bmatrix} \tag{5.27}$$

Thus, this produces two main equations:

$$\begin{aligned} \mathrm{inv}_1\hat{\boldsymbol{C}}_{11}\hat{\boldsymbol{b}}_1 + \mathrm{inv}_1\hat{\boldsymbol{C}}_{12}\hat{\boldsymbol{v}}_2 &= \hat{\boldsymbol{v}}_1 \\ \mathrm{inv}_1\hat{\boldsymbol{C}}_{21}\hat{\boldsymbol{b}}_1 + \mathrm{inv}_1\hat{\boldsymbol{C}}_{22}\hat{\boldsymbol{v}}_2 &= 0 \end{aligned} \tag{5.28}$$

Reorganizing the second equation above yields:

$$\begin{aligned} \mathrm{inv}_1\hat{\boldsymbol{C}}_{22}\hat{\boldsymbol{v}}_2 &= -\mathrm{inv}_1\hat{\boldsymbol{C}}_{21}\hat{\boldsymbol{b}}_1 \\ \hat{\boldsymbol{v}}_2 &= -\left(\mathrm{inv}_1\hat{\boldsymbol{C}}_{22}\right)^{-1}\mathrm{inv}_1\hat{\boldsymbol{C}}_{21}\hat{\boldsymbol{b}}_1 \end{aligned} \tag{5.29}$$

Substituting $\hat{\boldsymbol{v}}_2$ into the first equation results in:

$$\begin{aligned} \text{inv}_1\hat{\boldsymbol{C}}_{11}\hat{\boldsymbol{b}}_1 + \text{inv}_1\hat{\boldsymbol{C}}_{12}\hat{\boldsymbol{v}}_2 &= \hat{\boldsymbol{v}}_1 \\ \text{inv}_1\hat{\boldsymbol{C}}_{11}\hat{\boldsymbol{b}}_1 - \text{inv}_1\hat{\boldsymbol{C}}_{12}\left(\text{inv}_1\hat{\boldsymbol{C}}_{22}\right)^{-1}\text{inv}_1\hat{\boldsymbol{C}}_{21}\hat{\boldsymbol{b}}_1 &= \hat{\boldsymbol{v}}_1 \\ \left(\text{inv}_1\hat{\boldsymbol{C}}_{11} - \text{inv}_1\hat{\boldsymbol{C}}_{12}\left(\text{inv}_1\hat{\boldsymbol{C}}_{22}\right)^{-1}\text{inv}_1\hat{\boldsymbol{C}}_{21}\right)\hat{\boldsymbol{b}}_1 &= \hat{\boldsymbol{v}}_1 \end{aligned} \tag{5.30}$$

Thus,

$$\hat{\hat{\boldsymbol{W}}}_p = \text{inv}_1\hat{\boldsymbol{C}}_{11} - \text{inv}_1\hat{\boldsymbol{C}}_{12}\left(\text{inv}_1\hat{\boldsymbol{C}}_{22}\right)^{-1}\text{inv}_1\hat{\boldsymbol{C}}_{21} \tag{5.31}$$

Substituting the values for the entries of the partial inverse, $\text{inv}_1\hat{\boldsymbol{C}}$ matrix results in:

$$\hat{\hat{\boldsymbol{W}}}_p = \left(\hat{\boldsymbol{C}}_{11}\right)^{-1} + \left(\hat{\boldsymbol{C}}_{11}\right)^{-1}\hat{\boldsymbol{C}}_{12}\left(\text{inv}_1\hat{\boldsymbol{C}}_{22}\right)^{-1}\hat{\boldsymbol{C}}_{21}\left(\hat{\boldsymbol{C}}_{11}\right)^{-1} \tag{5.32}$$

Here, the computation of the inverse $\left(\text{inv}_1\hat{\boldsymbol{C}}_{22}\right)^{-1}$ is still computationally costly. Fortunately, we do not have to explicitly compute it. The multiplication $\hat{\boldsymbol{C}}_{12}\left(\text{inv}_1\hat{\boldsymbol{C}}_{22}\right)^{-1}$ corresponds to the solution $\boldsymbol{S}$ of the following linear system, which can be efficiently solved (e.g., using *mrdivide* function in Matlab or *solve* function in R):

$$\begin{aligned} \boldsymbol{S}\text{inv}_1\hat{\boldsymbol{C}}_{22} &= \hat{\boldsymbol{C}}_{12} \\ \boldsymbol{S} &= \hat{\boldsymbol{C}}_{12}\left(\text{inv}_1\hat{\boldsymbol{C}}_{22}\right)^{-1} \end{aligned} \tag{5.33}$$

where

$$\text{inv}_1\hat{\boldsymbol{C}}_{22} = \hat{\boldsymbol{C}}_{22} - \hat{\boldsymbol{C}}_{21}\left(\hat{\boldsymbol{C}}_{11}\right)^{-1}\hat{\boldsymbol{C}}_{12} \tag{5.34}$$

Substituting the solution matrix S yields the equation:

$$\hat{\hat{\boldsymbol{W}}}_p = \left(\hat{\boldsymbol{C}}_{11}\right)^{-1} + \left(\hat{\boldsymbol{C}}_{11}\right)^{-1}\boldsymbol{S}\hat{\boldsymbol{C}}_{21}\left(\hat{\boldsymbol{C}}_{11}\right)^{-1} \tag{5.35}$$

Note that, the weighting matrix $\hat{\hat{\boldsymbol{W}}}_p$ of size $n_q \times n_q$ is defined for quantified and relevant nodes. However, it can be easily mapped to the space of all phosphosites $\boldsymbol{W}_p$ (of size $n \times n$) by setting the remaining values for all other nodes as 0.

After the weighting matrix for phosphosites $\boldsymbol{W}_p$ is obtained, the weights kinase inference $\boldsymbol{W}$ are obtained for mean substrate phosphorylation using the kinase-substrate network $\boldsymbol{W}_{ks}$:

$$\boldsymbol{W} = \boldsymbol{W}_{ks}\boldsymbol{W}_p \tag{5.36}$$

Here, the matrix $\boldsymbol{W}$ represents a weighted network of size $n_{kin} \times n$, where n_{kin} denotes the number of kinases and n represents the number of phosphosites. On the other hand, the matrix $\boldsymbol{W}_{ks}$ represents an unweighted network of size $n_{kin} \times n$, specifically indicating the interactions between kinases and phosphosites that are known kinase targets.

Options for kinase activity inference. RokaiXplorer offers several options specific for kinase activity inference. These are outlined below:

- **Kinase-substrate dataset:** The options are PhosphoSitePlus (PSP) or PSP + Signor. The former only uses the known kinase targets W_{ks} from PSP[49], and the latter additionally include known targets from Signor[76] as well.
- **RoKAI Network:** Determines what kind of interactions are included in the RoKAI functional network W_{in}. *KinaseSubstrate* option only includes the known kinase targets W_{ks} network. *KS+PPI* option additionally adds the protein-protein interactions (PPI) between the kinases. *KS+PPI+SD* also includes interactions between phosphosites based on structure distance evidence. *KS+PPI+SD+CoEv* further includes interactions between phosphosites based on co-evolution evidence.
- **Use sites in functional network:** A binary flag that determines whether the network propagation through RoKAI should be performed or not.
- **Min. number of substrates:** Determines the minimum number of phosphosites that are known targets of a kinase needs to be identified in the dataset for that kinase to be included in the analysis. In other words, the kinase will be considered only if it has at least the specified minimum number of identified phosphosites in W_{ks}.

5.3.7 Statistical inference at pathway level

In addition, RokaiXplorer provides the functionality to perform pathway or gene ontology (GO) term enrichment analysis based on peptides or proteins that have been identified as significant in previous analyses involving phosphorylation or protein expression. Pathway analysis is a valuable tool for gaining insights into the biological pathways and networks affected by the observed changes in the phosphoproteome. By conducting pathway analysis, users can further understand the broader biological context and functional implications of the identified phosphorylation events.

The enrichment analysis in RokaiXplorer is conducted using an over-representation analysis (ORA) approach. This involves assessing the enrichment of significant phosphosites or proteins in specific pathways or gene ontology terms. It is important to note that the analysis is performed at the protein level, even when examining phosphorylation at the phosphosite level. Specifically, for enrichment analysis on phosphosites, we map the significant phosphosites to their corresponding proteins. A protein is considered significant if it contains at least one significant phosphosite. To determine the statistical significance of the enrichment, we employ the chi-squared test with Yate's correction[143]. This test helps evaluate whether the observed distribution of significant phosphosites or proteins across pathways or gene ontology terms deviates significantly from what would be expected by chance alone and produces the p-values. The test produces p-values, which indicate the strength of evidence for enrichment. We apply the Benjamini-Hochberg procedure[11] to alleviate the multiple comparisons issue by limit the false discovery rate (FDR) of the findings.

Estimating magnitude of enrichment — Bayesian log risk ratio. To estimate the magnitude of enrichment, RokaiXplorer utilizes a Bayesian approach to estimate log risk ratio. These estimates are primarily used to visualize the magnitude of enrichment, such as in bar plots in the inspection window. Let $\mathcal{P}_i^+$ denote the set of significant proteins within pathway i (e.g., the *hits*), and $\mathcal{P}_i^-$ denote the set of non-significant proteins within the pathway (i.e., the *misses*). Similarly, $\mathcal{P}_{\text{all}}^+$ represents the set of all significant proteins, and $\mathcal{P}_{\text{all}}^-$ represents the set of all non-significant proteins. Furthermore, $n^+[i]$ corresponds to the number of significant proteins in pathway i, while $n^-[i]$ corresponds to the number of non-significant proteins within the pathway. Similarly, n_{all}^+ denotes the total number of significant proteins, and n_{all}^- denotes the total number of non-significant proteins. Without using a

Bayesian prior, the log risk ratio (logRR) can be computed as follows:

$$\text{logRR}[i] = \log_2 \frac{n^+[i] / \left(n^+[i] + n^-[i]\right)}{\left(n^+_{all} - n^+[i]\right) / \left(n^+_{\text{all}} + n^-_{\text{all}} - n^+[i] - n^-[i]\right)} \tag{5.37}$$

This estimate presents several issues. For instance, when a pathway contains only 1 significant protein (1 hit and 0 misses), the risk ratio approaches the maximum value possible (equal to 1 divided by the significance ratio $\frac{n^+_{\text{all}}}{n^+_{\text{all}} + n^-_{\text{all}}}$), even though the presence of only 1 significant protein provides weak evidence that cannot be reliably attributed to anything other than chance alone. In contrast, a pathway with 100 significant proteins out of 100 proteins yields a similar risk ratio, despite providing much stronger and more confident evidence. To address this issue, we employ a Bayesian estimate that incorporates a prior belief, assuming that the pathway's enrichment (hit ratio) is equal to the significance ratio r^+_{all} when no additional evidence is available:

$$r^+_{\text{all}} = \frac{n^+_{\text{all}}}{n^+_{\text{all}} + n^-_{\text{all}}} \tag{5.38}$$

For this purpose, our perspective is similar to that of the coin flip problem. Imagine that we conduct a series of independent trials, where each trial can result in a hit or a miss. After performing these trials, we observe a total of n^+ hits and n^- misses. Now, our goal is to determine the likelihood that the coin is fair, or in other words, what is our posterior belief about the probability of getting a hit, denoted as r?

Assuming uniform prior for all possible r values, the answer lies in the beta distribution:

$$r \sim \text{B}(n^+ + 1, n^- + 1) \tag{5.39}$$

Here, we can generalize this approach by introducing α^+ and α^- , which represent the "prior number of hits" and "prior number of misses" respectively. These values act as pseudotrials and represents the prior belief in the estimation process:

$$r \sim \text{B}(n^+ + \alpha^+, n^- + \alpha^-) \tag{5.40}$$

where $r^+_\alpha = \frac{\alpha^+}{\alpha^+ + \alpha^-}$ represents the prior mean. To select appropriate prior values, we consider three principles. First, we aim to ensure that r^+_α matches the population significance ratio r^+_{all}. Second, we want α^+ and α^- to be at least equal to 1 or higher. Lastly, we want the prior to be equivalent to uniform prior when $r^+_{\text{all}} = 0.5$ (where hits and misses are equally likely). To incorporate these principles, we utilize the following prior values:

$$\begin{aligned} \alpha^+ &= 1 \\ \alpha^- &= \frac{n^-_{all}}{n^+_{all}} \end{aligned} \tag{5.41}$$

It is important to note that these prior values are chosen under the assumption that the number of significant proteins n^+_{all} is smaller than the number of non-significant proteins n^-_{all}. However, in the unlikely scenario where n^-_{all} is larger than n^+_{all}, we swap the two groups and set $\alpha^- = 1$ and $\alpha^+ = \frac{n^+_{\text{all}}}{n^-_{\text{all}}}$ to satisfy the desired criteria and ensure appropriate estimation.

Based on these parameters, we calculate the medians $m[i]$ and $m_{\text{out}}[i]$ of the posterior distribution for the hit ratios $r[i]$ and $r_{\text{out}}[i]$ respectively. The hit ratio $r[i]$ represents the ratio of significant proteins to all proteins in pathway i, while the hit ratio $r_{\text{out}}[i]$ represents the ratio of significant proteins to all proteins outside of pathway i. The medians $m[i]$ and $m_{\text{out}}[i]$ are obtained based on the quantile function for beta distribution:

$$\begin{aligned} m[i] &= F^{-1}_{\text{Beta}}(0.5; n^+[i] + \alpha^+, n^-[i] + \alpha^-) \\ m_{out}[i] &= F^{-1}_{\text{Beta}}(0.5; n^+_{all} - n^+[i] + \alpha^+, n^-_{all} - n^-[i] + \alpha^-) \end{aligned} \tag{5.42}$$

where $F^{-1}_{\text{Beta}}(p; a, b)$ denotes the quantile function of the beta distribution for quantile p and parameters a and b. Thus, based on the provided parameters and Bayesian prior, we derive an estimate of the log risk ratio $\log\hat{\text{R}}\text{R}[i]$ for pathway i as follows:

$$\log\hat{\text{R}}\text{R}[i] = \log_2 \frac{m[i]}{m_{out}[i]} = \log_2 \frac{F^{-1}_{\text{Beta}}(0.5; n^+[i] + \alpha^+, n^-[i] + \alpha^-)}{F^{-1}_{\text{Beta}}(0.5; n^+_{all} - n^+[i] + \alpha^+, n^-_{all} - n^-[i] + \alpha^-)} \tag{5.43}$$

Parameters for pathway enrichment. RokaiXplorer offers several options for the pathway enrichment. These are outlined below:

Inclusion criteria (enrichment terms): These options determine the inclusion of enrichment terms in the analysis. The first option categorizes the terms based on their category, such as *Biological Process*, *Cellular Process*, or *Molecular Function* for GO terms. Additionally, there are options to filter out pathways based on the number of proteins identified in the input dataset. This can be done either by specifying a minimum number of observed proteins or by setting a minimum ratio of observed proteins to the total number of proteins in the pathway. Furthermore, there is an option to filter out highly similar pathways based on the Jaccard index. Pathways that exhibit a highly similar set of observed proteins are filtered out, retaining only one of them. In cases of duplication, the smaller pathway is retained Furthermore, there is an option to filter out highly similar pathways based on the Jaccard index. When this option is enabled, if multiple pathways exhibit a highly similar set of observed proteins, only one of them is retained, prioritizing the smaller pathway with less number of proteins.

Background set (proteins): These options determine the set of significant proteins to be used for the enrichment analysis. The first option allows for selecting the data source, which can be either the results from the *Phosphosite*, *Phosphoprotein*, or *Protein expression* analyses. If the *Phosphosites* option is chosen, the background set will include proteins that have at least one significant phosphosite. The subsequent options define the criteria for determining significance, such as the cut-off based on p-values or log2 fold changes. Additionally, there is a binary setting to indicate whether the Benjamini-Hochberg procedure should be applied to control the false discovery rate (FDR) of the findings. Finally, an additional option is provided to restrict the set of significant proteins to either positive or negative log fold changes, if desired.

5.3.8 Inspection window: Performing sample-wise inferences

In addition to the specific statistical analyses described earlier (e.g., phosphosites, phosphoproteins, protein expression, kinase inference, and pathway enrichment), RokaiXplorer offers sample-wise analyses at the individual sample level for more detailed inspection in the inspection window. These sample-wise analyses compare each sample in the Case group to all samples in the Control group, providing granular results that capture the variance between samples. The inspection window allows for visualization of these results through bar plots or box plots, providing a comprehensive view of the individual sample-level analysis.

For the sample-wise analysis at the phosphosite and protein expression level, we utilize the z-test as a simpler alternative to the t-test. The z-test estimates the standard error by considering the variance across the phosphosites or proteins. Unlike the t-test, which requires multiple samples to measure the variance between them, the z-test can be applied without such a requirement. Furthermore, in the kinase activity inference, we perform the analysis without utilizing network propagation through RoKAI's functional network. This decision is made to conserve the computational resources on the web server and streamline the analysis process.

References

[1] Openne: An open source toolkit for network embedding. `https://github.com/thunlp/OpenNE`, 2019.

[2] Yuichi Abe, Maiko Nagano, Takahisa Kuga, Asa Tada, Junko Isoyama, Jun Adachi, and Takeshi Tomonaga. Deep phospho-and phosphotyrosine proteomics identified active kinases and phosphorylation networks in colorectal cancer cell lines resistant to cetuximab. *Scientific reports*, 7(1):1–12, 2017.

[3] Osama A Arshad, Vincent Danna, Vladislav A Petyuk, Paul D Piehowski, Tao Liu, Karin D Rodland, and Jason E McDermott. An integrative analysis of tumor proteomic and phosphoproteomic profiles to examine the relationships between kinase activity and phosphorylation. *Molecular & Cellular Proteomics*, 18(8 suppl 1):S26–S36, 2019.

[4] Marzieh Ayati, Danica Wiredja, Daniela Schlatzer, Sean Maxwell, Ming Li, Mehmet Koyutürk, and Mark R Chance. Cophosk: A method for comprehensive kinase substrate annotation using co-phosphorylation analysis. *PLoS computational biology*, 15(2):e1006678, 2019.

[5] Marzieh Ayati, Mark R Chance, and Mehmet Koyuturk. Co-phosphorylation networks reveal subtype-specific signaling modules in breast cancer. *Bioinformatics (Oxford, England)*, page btaa678, 2020.

[6] Marzieh Ayati, Serhan Yılmaz, Mark R Chance, and Mehmet Koyuturk. Functional characterization of co-phosphorylation networks. *Bioinformatics*, 38 (15):3785–3793, 2022.

[7] Sara Ballouz, Wim Verleyen, and Jesse Gillis. Guidance for rna-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*, 31(13):2123–2130, 2015.

[8] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[9] Karl W Barber and Jesse Rinehart. The abcs of ptms. *Nature chemical biology*, 14(3):188, 2018.

[10] Robin Beekhof, Carolien van Alphen, Alex A Henneman, Jaco C Knol, Thang V Pham, Frank Rolfs, Mariette Labots, Evan Henneberry, Tessa YS Le Large,

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

[11] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[12] Jos Boekhorst, Bas van Breukelen, Albert JR Heck, and Berend Snel. Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome biology*, 9(10):R144, 2008.

[13] James E Butrynski, David R D'Adamo, Jason L Hornick, Paola Dal Cin, Cristina R Antonescu, Suresh C Jhanwar, Marc Ladanyi, Marzia Capelletti, Scott J Rodig, Nikhil Ramaiya, et al. Crizotinib in alk-rearranged inflammatory myofibroblastic tumor. *New England Journal of Medicine*, 363(18):1727–1733, 2010.

[14] Diogo M Camacho, Katherine M Collins, Rani K Powers, James C Costello, and James J Collins. Next-generation machine learning for biological networks. *Cell*, 173(7):1581–1592, 2018.

[15] Matt Carter and Jennifer C Shieh. *Guide to research techniques in neuroscience*. Academic Press, 2015.

[16] Scott L Carter, Christian M Brechbühler, Michael Griffin, and Andrew T Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–2250, 2004.

[17] Pedro Casado, Juan-Carlos Rodriguez-Prados, Sabina C Cosulich, Sylvie Guichard, Bart Vanhaesebroeck, Simon Joel, and Pedro R Cutillas. Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal.*, 6(268):rs6–rs6, 2013.

[18] Cheng-Kang Chiang, Aleksander Tworak, Brian M Kevany, Bo Xu, Janice Mayne, Zhibin Ning, Daniel Figeys, and Krzysztof Palczewski. Quantitative phosphoproteomics reveals involvement of multiple signaling pathways in early phagocytosis by the retinal pigmented epithelium. *Journal of Biological Chemistry*, 292(48):19826–19839, 2017.

[19] Jang Hyun Choi, Alexander S Banks, Jennifer L Estall, Shingo Kajimura, Pontus Boström, Dina Laznik, Jorge L Ruas, Michael J Chalmers, Theodore M

Kamenecka, Matthias Blüher, et al. Anti-diabetic drugs inhibit obesity-linked phosphorylation of pparγ by cdk5. *Nature*, 466(7305):451–456, 2010.

[20] Anna Cichonska, Balaguru Ravikumar, Elina Parri, Sanna Timonen, Tapio Pahikkala, Antti Airola, Krister Wennerberg, Juho Rousu, and Tero Aittokallio. Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. *PLoS computational biology*, 13(8): e1005678, 2017.

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

[22] UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212, 2014.

[23] KD Copps and MF White. Regulation of insulin sensitivity by serine/threonine phosphorylation of insulin receptor substrate proteins irs1 and irs2. *Diabetologia*, 55(10):2565–2582, 2012.

[24] Lenore Cowen, Trey Ideker, Benjamin J Raphael, and Roded Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9):551, 2017.

[25] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477, 2013.

[26] Eric B Dammer, Andrew K Lee, Duc M Duong, Marla Gearing, James J Lah, Allan I Levey, and Nicholas T Seyfried. Quantitative phosphoproteomics of alzheimer's disease reveals cross-talk between kinases and small heat shock proteins. *Proteomics*, 15(2-3):508–519, 2015.

[27] Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Roy McMorran, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. The comparative toxicogenomics database: update 2019. *Nucleic acids research*, 47(D1):D948–D954, 2019.

[28] Noah Dephoure, Kathleen L Gould, Steven P Gygi, and Douglas R Kellogg. Mapping and analysis of phosphorylation sites: a quick guide for cell biologists. *Molecular biology of the cell*, 24(5):535–542, 2013.

[29] Kapil Devkota, James M Murphy, and Lenore J Cowen. Glide: combining local methods and diffusion state embeddings to predict missing interactions in biological networks. *Bioinformatics*, 36(Supplement_1):i464–i473, 2020.

[30] Iman Deznabi, Busra Arabaci, Mehmet Koyutürk, and Oznur Tastan. Deepkinzero: Zero-shot learning for predicting kinase-phosphosite associations involving understudied kinases. *BioRxiv*, page 670638, 2019.

[31] Iman Deznabi, Busra Arabaci, Mehmet Koyutürk, and Oznur Tastan. Deepkinzero: zero-shot learning for predicting kinase–phosphosite associations involving understudied kinases. *Bioinformatics*, 36(12):3652–3661, 2020.

[32] Holger Dinkel, Claudia Chica, Allegra Via, Cathryn M Gould, Lars J Jensen, Toby J Gibson, and Francesca Diella. Phospho. elm: a database of phosphorylation sites—update 2011. *Nucleic acids research*, 39(suppl_1):D261–D267, 2010.

[33] Peter G Doyle and J Laurie Snell. *Random walks and electric networks*, volume 22. American Mathematical Soc., 1984.

[34] Justin M Drake, Nicholas A Graham, Tanya Stoyanova, Amir Sedghi, Andrew S Goldstein, Houjian Cai, Daniel A Smith, Hong Zhang, Evangelia Komisopoulou, Jiaoti Huang, et al. Oncogene-specific activation of tyrosine kinase networks during prostate cancer progression. *Proceedings of the National Academy of Sciences*, 109(5):1643–1648, 2012.

[35] Justin M Drake, Evan O Paull, Nicholas A Graham, John K Lee, Bryan A Smith, Bjoern Titz, Tanya Stoyanova, Claire M Faltermeier, Vladislav Uzunangelov, Daniel E Carlin, et al. Phosphoproteome integration reveals patient-specific networks in prostate cancer. *Cell*, 166(4):1041–1054, 2016.

[36] Fatma-Elzahraa Eid, Haitham A Elmarakeby, Yujia Alina Chan, Nadine Fornelos, Mahmoud ElHefnawi, Eliezer M Van Allen, Lenwood S Heath, and Kasper Lage. Systematic auditing is essential to debiasing machine learning in biology. *Communications biology*, 4(1):1–9, 2021.

[37] Sinan Erten, Gurkan Bebek, Rob M Ewing, and Mehmet Koyutürk. Da da: degree-aware algorithms for network-based disease gene prioritization. *BioData mining*, 4(1):1–20, 2011.

[38] Doriano Fabbro, Sandra W Cowan-Jacob, Henrik Möbitz, and Georg Martiny-Baron. Targeting cancer with small-molecular-weight kinase inhibitors. In *Kinase Inhibitors*, pages 1–34. Springer, 2012.

[39] Ming Gao, Leihui Chen, Xiangnan He, and Aoying Zhou. Bine: Bipartite network embedding. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 715–724, 2018.

[40] Jesse Gillis, Sara Ballouz, and Paul Pavlidis. Bias tradeoffs in the creation and analysis of protein–protein interaction networks. *Journal of proteomics*, 100: 44–54, 2014.

[41] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of KDD*, pages 855–864, 2016.

[42] Vincentius A Halim, Monica Alvarez-Fernandez, Yan Juan Xu, Melinda Aprelia, Henk WP van den Toorn, Albert JR Heck, Shabaz Mohammed, and Rene H Medema. Comparative phosphoproteomic analysis of checkpoint recovery identifies new regulators of the dna damage response. *Sci. Signal.*, 6(272): rs9–rs9, 2013.

[43] Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, et al. Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, 46(D1):D380–D386, 2018.

[44] Claudia Hernandez-Armenta, David Ochoa, Emanuel Gonçalves, Julio Saez-Rodriguez, and Pedro Beltrao. Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics*, 33(12):1845–1851, 2017.

[45] Maruan Hijazi, Ryan Smith, Vinothini Rajeeve, Conrad Bessant, and Pedro R Cutillas. Reconstructing kinase network topologies from phosphoproteomics data reveals cancer-associated rewiring. *Nature Biotechnology*, 38 (4):493–502, 2020.

[46] Heiko Horn, Erwin M Schoof, Jinho Kim, Xavier Robin, Martin L Miller, Francesca Diella, Anita Palma, Gianni Cesareni, Lars Juhl Jensen, and Rune Linding. Kinomexplorer: an integrated platform for kinome biology studies. *Nature methods*, 11(6):603, 2014.

[47] Peter V Hornbeck, Bin Zhang, Beth Murray, Jon M Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. Phosphositeplus, 2014: mutations, ptms and recalibrations. *Nucleic acids research*, 43(D1):D512–D520, 2014.

[48] Peter V Hornbeck, Bin Zhang, Beth Murray, Jon M Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. Phosphositeplus, 2014: mutations, ptms and recalibrations. *Nucleic acids research*, 43(D1):D512–D520, 2015.

[49] Peter V Hornbeck, Bin Zhang, Beth Murray, Jon M Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. Phosphositeplus, 2014: mutations, ptms and recalibrations. *Nucleic acids research*, 43(D1):D512–D520, 2015.

[50] Kuan-lin Huang, Shunqiang Li, Philipp Mertins, Song Cao, Harsha P Gunawardena, Kelly V Ruggles, DR Mani, Karl R Clauser, Maki Tanioka, Jerry Usary, et al. Proteogenomic integration reveals therapeutic targets in breast cancer xenografts. *Nature communications*, 8(1):1–17, 2017.

[51] Brandon M Invergo, Borgthor Petursson, Nosheen Akhtar, David Bradley, Girolamo Giudice, Maruan Hijazi, Pedro Cutillas, Evangelia Petsalaki, and Pedro Beltrao. Prediction of signed protein kinase regulatory circuits. *Cell systems*, 10(5):384–396, 2020.

[52] Edwin T Jaynes and Oscar Kempthorne. Confidence intervals vs bayesian intervals. In *Foundations of probability theory, statistical inference, and statistical theories of science*, pages 175–257. Springer, 1976.

[53] Connie R Jimenez and HM Verheul. Mass spectrometry-based proteomics: from cancer biology to protein biomarkers, drug targets, and clinical applications. In *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting*, pages e504–10, 2014.

[54] David T Jones. Setting the standards for machine learning in biology. *Nature Reviews Molecular Cell Biology*, 20(11):659–660, 2019.

[55] Kumaran Kandasamy, S Sujatha Mohan, Rajesh Raju, Shivakumar Keerthikumar, Ghantasala S Sameer Kumar, Abhilash K Venugopal, Deepthi Telikicherla, J Daniel Navarro, Suresh Mathivanan, Christian Pecquet, et al. Netpath: a public resource of curated signal transduction pathways. *Genome biology*, 11(1):R3, 2010.

[56] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[57] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 2016.

[58] Yoo-Ah Kim, Stefan Wuchty, and Teresa M Przytycka. Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput Biol*, 7 (3):e1001095, 2011.

[59] KC Kishan, Rui Li, Feng Cui, and Anne R Haake. Predicting biomedical interactions with higher-order graph convolutional networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(2):676–687, 2021.

[60] Susan Klaeger, Stephanie Heinzlmeir, Mathias Wilhelm, Harald Polzer, Binje Vick, Paul-Albert Koenig, Maria Reinecke, Benjamin Ruprecht, Svenja Petzoldt, Chen Meng, et al. The target landscape of clinical kinase drugs. *Science*, 358(6367), 2017.

[61] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, et al. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, 39(suppl_1):D1035–D1041, 2010.

[62] István A Kovács, Katja Luck, Kerstin Spirohn, Yang Wang, Carl Pollis, Sadie Schlabach, Wenting Bian, Dae-Kyum Kim, Nishka Kishore, Tong Hao, et al. Network-based prediction of protein interactions. *Nature communications*, 10(1):1–8, 2019.

[63] Fumika Koyano, Kei Okatsu, Hidetaka Kosako, Yasushi Tamura, Etsu Go, Mayumi Kimura, Yoko Kimura, Hikaru Tsuchiya, Hidehito Yoshihara, Takatsugu Hirokawa, et al. Ubiquitin is phosphorylated by pink1 to activate parkin. *Nature*, 510(7503):162–166, 2014.

[64] Nicole Kresge, Robert D Simoni, and Robert L Hill. The process of reversible phosphorylation: the work of edmond h. fischer. *Journal of Biological Chemistry*, 286(3):e1–e2, 2011.

[65] Karsten Krug, Philipp Mertins, Bin Zhang, Peter Hornbeck, Rajesh Raju, Rushdy Ahmad, Matthew Szucs, Filip Mundt, Dominique Forestier, Judit Jane-Valbuena, et al. A curated resource for phosphosite-specific signature analysis. *Molecular & cellular proteomics*, 18(3):576–593, 2019.

[66] Karsten Krug, Philipp Mertins, Bin Zhang, Peter Hornbeck, Rajesh Raju, Rushdy Ahmad, Matthew Szucs, Filip Mundt, Dominique Forestier, Judit Jane-Valbuena, et al. A curated resource for phosphosite-specific signature analysis. *Molecular & cellular proteomics*, 18(3):576–593, 2019.

[67] Jérôme Kunegis, Ernesto W De Luca, and Sahin Albayrak. The link prediction problem in bipartite networks. In *International Conference on Information*

Processing and Management of Uncertainty in Knowledge-based Systems, pages 380–389. Springer, 2010.

[68] Georg Kustatscher, Tom Collins, Anne-Claude Gingras, Tiannan Guo, Henning Hermjakob, Trey Ideker, Kathryn S Lilley, Emma Lundberg, Edward M Marcotte, Markus Ralser, et al. Understudied proteins: opportunities and challenges for functional proteomics. *Nature Methods*, pages 1–6, 2022.

[69] Yi-An Lai, Chin-Chi Hsu, Wen Hao Chen, Mi-Yen Yeh, and Shou-De Lin. Prune: Preserving proximity and global ranking for network embedding. *Advances in neural information processing systems*, 30, 2017.

[70] Chengwei Lei and Jianhua Ruan. A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity. *Bioinformatics*, 29(3):355–364, 2013.

[71] Hao Li, Yu Sun, Hao Hong, Xin Huang, Huan Tao, Qiya Huang, Longteng Wang, Kang Xu, Jingbo Gan, Hebing Chen, et al. Inferring transcription factor regulatory networks from single-cell atac-seq data based on graph neural networks. *Nature Machine Intelligence*, 4(4):389–400, 2022.

[72] Ying Li, Xueya Zhou, Zichao Zhai, and Tingting Li. Co-occurring protein phosphorylation are functionally associated. *PLoS computational biology*, 13(5):e1005502, 2017.

[73] Xujun Liang, Pengfei Zhang, Lu Yan, Ying Fu, Fang Peng, Lingzhi Qu, Meiying Shao, Yongheng Chen, and Zhuchu Chen. Lrssl: predict and interpret drug–disease associations based on data integration using sparse subspace learning. *Bioinformatics*, 33(8):1187–1196, 2017.

[74] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, 2003.

[75] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[76] Luana Licata, Prisca Lo Surdo, Marta Iannuccelli, Alessandro Palma, Elisa Micarelli, Livia Perfetto, Daniele Peluso, Alberto Calderone, Luisa Castagnoli, and Gianni Cesareni. Signor 2.0, the signaling network open resource 2.0: 2019 update. *Nucleic acids research*, 48(D1):D504–D510, 2020.

[77] Gipsi Lima-Mendez and Jacques Van Helden. The powerful law of the power law and other myths in network biology. *Molecular BioSystems*, 5(12):1482–1493, 2009.

[78] Rune Linding, Lars Juhl Jensen, Gerard J Ostheimer, Marcel ATM van Vugt, Claus Jørgensen, Ioana M Miron, Francesca Diella, Karen Colwill, Lorne Taylor, Kelly Elder, et al. Systematic discovery of in vivo phosphorylation networks. *Cell*, 129(7):1415–1426, 2007.

[79] Yu Liu and Mark R Chance. Integrating phosphoproteomics in systems biology. *Computational and structural biotechnology journal*, 10(17):90–97, 2014.

[80] Gerard Manning, David B Whyte, Ricardo Martinez, Tony Hunter, and Sucha Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

[81] Alexandru Mara, Jefrey Lijffijt, and Tijl De Bie. Evalne: A framework for evaluating network embeddings on link prediction. *arXiv preprint arXiv:1901.09691*, 2019.

[82] Alexandru Cristian Mara, Jefrey Lijffijt, and Tijl De Bie. Benchmarking network embedding models for link prediction: Are we making progress? In *2020 IEEE 7th International conference on data science and advanced analytics (DSAA)*, pages 138–147. IEEE, 2020.

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

[84] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[85] Philipp Mertins, Feng Yang, Tao Liu, DR Mani, Vladislav A Petyuk, Michael A Gillette, Karl R Clauser, Jana W Qiao, Marina A Gritsenko, Ronald J Moore, et al. Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Molecular & cellular proteomics*, 13(7):1690–1704, 2014.

[86] Philipp Mertins, DR Mani, Kelly V Ruggles, Michael A Gillette, Karl R Clauser, Pei Wang, Xianlong Wang, Jana W Qiao, Song Cao, Francesca Petralia, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, 534(7605):55–62, 2016.

[87] Patrick E Meyer, Frederic Lafitte, and Gianluca Bontempi. minet: Ar/bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics*, 9(1):461, 2008.

[88] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[89] Jeremy Miles. R squared, adjusted r squared. *Wiley StatsRef: Statistics Reference Online*, 2014.

[90] Pablo Minguez, Ivica Letunic, Luca Parca, and Peer Bork. Ptmcode: a database of known and predicted functional associations between post-translational modifications in proteins. *Nucleic acids research*, 41(D1):D306–D311, 2012.

[91] Pablo Minguez, Luca Parca, Francesca Diella, Daniel R Mende, Runjun Kumar, Manuela Helmer-Citterich, Anne-Claude Gavin, Vera Van Noort, and Peer Bork. Deciphering a global network of functionally associated post-translational modifications. *Molecular systems biology*, 8(1), 2012.

[92] Pablo Minguez, Ivica Letunic, Luca Parca, Luz Garcia-Alonso, Joaquin Dopazo, Jaime Huerta-Cepas, and Peer Bork. Ptmcode v2: a resource for functional associations of post-translational modifications within and between proteins. *Nucleic acids research*, 43(D1):D494–D502, 2015.

[93] Marcel Mischnik, Francesca Sacco, Jürgen Cox, Hans-Christoph Schneider, Matthias Schäfer, Manfred Hendlich, Daniel Crowther, Matthias Mann, and Thomas Klabunde. Ikap: A heuristic framework for inference of kinase activities from phosphoproteomics data. *Bioinformatics*, 32(3):424–431, 2015.

[94] Stephanie Munk, Jan C Refsgaard, Jesper V Olsen, and Lars J Jensen. From phosphosites to kinases. In *Phospho-Proteomics*, pages 307–321. Springer, 2016.

[95] Joerg Neddens, Magdalena Temmel, Stefanie Flunkert, Bianca Kerschbaumer, Christina Hoeller, Tina Loeffler, Vera Niederkofler, Guenther Daum, Johannes Attems, and Birgit Hutter-Paier. Phosphorylation of different tau sites during progression of alzheimer's disease. *Acta neuropathologica communications*, 6(1):52, 2018.

[96] Elise J. Needham, Benjamin L. Parker, Timur Burykin, David E. James, and Sean J. Humphrey. Illuminating the dark phosphoproteome. *Science Signaling*, 12(565), 2019. ISSN 1945-0877. doi: 10.1126/scisignal.aau8645. URL `https://stke.sciencemag.org/content/12/565/eaau8645`.

[97] Elise J Needham, Benjamin L Parker, Timur Burykin, David E James, and Sean J Humphrey. Illuminating the dark phosphoproteome. *Sci. Signal.*, 12 (565):eaau8645, 2019.

[98] Walter Nelson, Marinka Zitnik, Bo Wang, Jure Leskovec, Anna Goldenberg, and Roded Sharan. To embed or not: network embedding as a paradigm in computational biology. *Frontiers in genetics*, 10:381, 2019.

[99] Paolo Neviani and Danilo Perrotti. Setting op449 into the pp2a-activating drug family. *Clinical Cancer Research*, 20(8):2026–2028, 2014.

[100] Hafumi Nishi, Alexey Shaytan, and Anna R Panchenko. Physicochemical mechanisms of protein regulation by phosphorylation. *Frontiers in genetics*, 5:270, 2014.

[101] Hafumi Nishi, Emek Demir, and Anna R Panchenko. Crosstalk between signaling pathways provided by single and multiple protein phosphorylation sites. *Journal of molecular biology*, 427(2):511–520, 2015.

[102] David Ochoa, Mindaugas Jonikas, Robert T Lawrence, Bachir El Debs, Joel Selkrig, Athanasios Typas, Judit Villén, Silvia DM Santos, and Pedro Beltrao. An atlas of human kinase regulation. *Molecular systems biology*, 12(12), 2016.

[103] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2015.

[104] Rose Oughtred, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Chris Stark, Andrew Willems, Lorrie Boucher, Genie Leung, Nadine Kolas, Frederick Zhang, et al. The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30 (1):187–200, 2021.

[105] Livia Perfetto, Leonardo Briganti, Alberto Calderone, Andrea Cerquone Perpetuini, Marta Iannuccelli, Francesca Langone, Luana Licata, Milica Marinkovic, Anna Mattioni, Theodora Pavlidou, et al. Signor: a database of

causal relationships between biological entities. *Nucleic acids research*, 44 (D1):D548–D554, 2015.

[106] Suraj Peri, J Daniel Navarro, Troels Z Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, TKB Gandhi, KN Chandrika, Nandan Deshpande, Shubha Suresh, et al. Human protein reference database as a discovery resource for proteomics. *Nucleic acids research*, 32(suppl_1): D497–D501, 2004.

[107] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of KDD*, pages 701–710, 2014.

[108] Danilo Perrotti and Paolo Neviani. Protein phosphatase 2a: a target for anticancer therapy. *The lancet oncology*, 14(6):e229–e238, 2013.

[109] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.

[110] Puneet Puri, Faridoddin Mirshahi, Onpan Cheung, Ramesh Natarajan, James W Maher, John M Kellum, and Arun J Sanyal. Activation and dysregulation of the unfolded protein response in nonalcoholic fatty liver disease. *Gastroenterology*, 134(2):568–576, 2008.

[111] Arun K Ramani, Zhihua Li, G Traver Hart, Mark W Carlson, Daniel R Boutz, and Edward M Marcotte. A map of human protein interactions derived from co-expression of human mrnas and their orthologs. *Molecular systems biology*, 4(1):180, 2008.

[112] Lindsay C Reese, Fernanda Laezza, Randall Woltjer, and Giulio Taglialatela. Dysregulated phosphorylation of ca2+/calmodulin-dependent protein kinase ii-α in the hippocampus of subjects with mild cognitive impairment and alzheimer's disease. *Journal of neurochemistry*, 119(4):791–804, 2011.

[113] Klarisa Rikova, Ailan Guo, Qingfu Zeng, Anthony Possemato, Jian Yu, Herbert Haack, Julie Nardone, Kimberly Lee, Cynthia Reeves, Yu Li, et al. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell*, 131(6):1190–1203, 2007.

[114] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[115] Carlos HM Rodrigues, Yoochan Myung, Douglas EV Pires, and David B Ascher. mcsm-ppi2: predicting the effects of mutations on protein–protein interactions. *Nucleic acids research*, 47(W1):W338–W344, 2019.

[116] Matthew Ruffalo, Mehmet Koyutürk, and Roded Sharan. Network-based integration of disparate omic data to identify" silent players" in cancer. *PLoS computational biology*, 11(12):e1004595, 2015.

[117] Franklin E Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114, 1946.

[118] Regev Schweiger and Michal Linial. Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data. *Biology direct*, 5 (1):6, 2010.

[119] Sreenath V Sharma, Daphne W Bell, Jeffrey Settleman, and Daniel A Haber. Epidermal growth factor receptor mutations in lung cancer. *Nature Reviews Cancer*, 7(3):169–181, 2007.

[120] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

[121] Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics*, 13(1):328, 2012.

[122] Zachary Stanfield, Mustafa Coşkun, and Mehmet Koyutürk. Drug response prediction as a link prediction problem. *Scientific reports*, 7(1):1–13, 2017.

[123] Chang Su, Jie Tong, Yongjun Zhu, Peng Cui, and Fei Wang. Network embedding in biomedical data science. *Briefings in bioinformatics*, 21(1):182–197, 2020.

[124] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[125] Sheng-Bao Suo, Jian-Ding Qiu, Shao-Ping Shi, Xiang Chen, and Ru-Ping Liang. Psea: Kinase-specific prediction and analysis of human phosphorylation substrates. *Scientific reports*, 4:4524, 2014.

[126] Silpa Suthram, Andreas Beyer, Richard M Karp, Yonina Eldar, and Trey Ideker. eqed: an efficient method for interpreting eqtl associations using protein networks. *Molecular systems biology*, 4(1):162, 2008.

[127] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1):D447–D452, 2014.

[128] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1):D447–D452, 2014.

[129] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613, 2019.

[130] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of WWW*, pages 1067–1077, 2015.

[131] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Random walk with restart: fast solutions and applications. *Knowledge and Information Systems*, 14(3):327–346, 2008.

[132] Michael J Tsatsomeros. Principal pivot transforms: properties and applications. *Linear Algebra and its Applications*, 307(1-3):151–165, 2000.

[133] Mitchell J Wagner, Aditya Pratapa, and TM Murali. Reconstructing signaling pathways using regular language constrained paths. *Bioinformatics*, 35(14): i624–i633, 2019.

[134] Yu-Chieh Wang, Suzanne E Peterson, and Jeanne F Loring. Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell research*, 24(2):143, 2014.

[135] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M Savitski, Emanuel Ziegler, Lars Butzmann,

Siegfried Gessulat, Harald Marx, et al. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587, 2014.

[136] Edmund H Wilkes, Pedro Casado, Vinothini Rajeeve, and Pedro R Cutillas. Kinase activity ranking using phosphoproteomics data (karp) quantifies the contribution of protein kinases to the regulation of cell viability. *Molecular & Cellular Proteomics*, 16(9):1694–1704, 2017.

[137] Danica Wiredja. *Phosphoproteomic Characterization of Systems-Wide Differential Signaling Induced by Small Molecule PP2A Activation*. PhD thesis, Case Western Reserve University, 2018.

[138] Danica D Wiredja, Marzieh Ayati, Sahar Mazhar, Jaya Sangodkar, Sean Maxwell, Daniela Schlatzer, Goutham Narla, Mehmet Koyutürk, and Mark R Chance. Phosphoproteomics profiling of nonsmall cell lung cancer cells treated with a novel phosphatase activator. *Proteomics*, 17(22):1700214, 2017.

[139] Danica D Wiredja, Mehmet Koyutürk, and Mark R Chance. The ksea app: a web-based tool for kinase activity inference from quantitative phosphoproteomics. *Bioinformatics*, 33(21):3489–3491, 2017.

[140] Peng Wu, Thomas E Nielsen, and Mads H Clausen. Small-molecule kinase inhibitors: an analysis of fda-approved drugs. *Drug discovery today*, 21(1): 5–10, 2016.

[141] Chia-Ying Yang, Chao-Hui Chang, Ya-Ling Yu, Tsu-Chun Emma Lin, Sheng-An Lee, Chueh-Chuan Yen, Jinn-Moon Yang, Jin-Mei Lai, Yi-Ren Hong, Tzu-Ling Tseng, et al. Phosphopoint: a comprehensive human kinase interactome and phospho-protein database. *Bioinformatics*, 24(16):i14–i20, 2008.

[142] Yang Yang, Ryan N Lichtenwalter, and Nitesh V Chawla. Evaluating link prediction methods. *Knowledge and Information Systems*, 45(3):751–782, 2015.

[143] Frank Yates. Contingency tables involving small numbers and the χ 2 test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217–235, 1934.

[144] John R Yates III, Shabaz Mohammed, and Albert JR Heck. Phosphoproteomics, 2014.

[145] Serhan Yılmaz, Marzieh Ayati, Daniela Schlatzer, A Ercüment Çiçek, Mark R Chance, and Mehmet Koyutürk. Robust inference of kinase activity using functional networks. *Nature communications*, 12(1):1–12, 2021.

[146] William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[147] Bin Yu, Cheng Chen, Xiaolin Wang, Zhaomin Yu, Anjun Ma, and Bingqiang Liu. Prediction of protein–protein interactions based on elastic net and deep forest. *Expert Systems with Applications*, 176:114876, 2021.

[148] Xiang Yue, Zhen Wang, Jingong Huang, Srinivasan Parthasarathy, Soheil Moosavinasab, Yungui Huang, Simon M Lin, Wen Zhang, Ping Zhang, and Huan Sun. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics*, 36(4):1241–1251, 2020.

[149] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, et al. Ensembl 2018. *Nucleic acids research*, 46(D1): D754–D761, 2017.

[150] Guoan Zhang, Beatrix M Ueberheide, Sofia Waldemarson, Sunnie Myung, Kelly Molloy, Jan Eriksson, Brian T Chait, Thomas A Neubert, and David Fenyö. Protein quantitation using mass spectrometry. In *Computational biology*, pages 211–222. Springer, 2010.

[151] Hui Zhang, Tao Liu, Zhen Zhang, Samuel H Payne, Bai Zhang, Jason E McDermott, Jian-Ying Zhou, Vladislav A Petyuk, Li Chen, Debjit Ray, et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell*, 166(3):755–765, 2016.

[152] Wen Zhang, Yanlin Chen, Dingfang Li, and Xiang Yue. Manifold regularized matrix factorization for drug-drug interaction prediction. *Journal of biomedical informatics*, 88:90–97, 2018.

[153] Xiao-Meng Zhang, Li Liang, Lin Liu, and Ming-Jing Tang. Graph neural networks and their current applications in bioinformatics. *Frontiers in genetics*, 12, 2021.

[154] Caicun Zhou, Yi-Long Wu, Gongyan Chen, Jifeng Feng, Xiao-Qing Liu, Changli Wang, Shucai Zhang, Jie Wang, Songwen Zhou, Shengxiang Ren, et al. Erlotinib versus chemotherapy as first-line treatment for patients with advanced egfr mutation-positive non-small-cell lung cancer (optimal, ctong-0802): a multicentre, open-label, randomised, phase 3 study. *The lancet oncology*, 12(8):735–742, 2011.

[155] Tao Zhou. Progresses and challenges in link prediction. *Iscience*, 24(11): 103217, 2021.

[156] Özgün Babur, Augustin Luna, Anil Korkut, Funda Durupinar, Metin Can Siper, Ugur Dogrusoz, Alvaro Sebastian Vaca Jacome, Ryan Peckner, Karen E. Christianson, Jacob D. Jaffe, Paul T. Spellman, Joseph E. Aslan, Chris Sander, and Emek Demir. Causal interactions from proteomic profiles: Molecular data meet pathway knowledge. *Patterns*, page 100257, 2021. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2021.100257. URL `https://www.sciencedirect.com/science/article/pii/S2666389921000830`.