Analysis of the understudied parts of the phospho-signalome using machine learning methods

Borgthor Petursson

EMBL-EBI
University of Cambridge
This thesis is submitted for the degree of
Doctor of Philosophy

Magdalene College August 2021

Declaration of Originality

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification, except as declared in the preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the Biology Degree Committee.

August 2021

Borgthor Petursson

Abstract

Analysis of the understudied parts of the phospho-signalome using machine learning methods

Borgthor Petursson

In order to make decisions and respond appropriately to external stimuli, cells rely on an intricate signalling system. One of the most important and best studied components of this signalling system is the phospho-signalling network. Phosphorylation relays information through adding phosphoryl groups onto substrates such as lipids or proteins, which in turn leads to changes in substrate function. Crucial components of this system include kinases, which phosphorylate on the substrate molecule and phosphatases that remove the phosphoryl group from the substrate.

To date, even though >100K phosphoproteins have been identified through high throughput experiments, the vast majority of phosphosites are of unknown function, while over a third of kinases have no known substrate (Needham et al., 2019). Furthermore, there is a large study bias in our current knowledge, demonstrated by a disproportionate number of interactions between highly cited kinases and substrates Invergo and Beltrao, 2018. The vast understudied signalling space combined with this study bias make it difficult to understand the general principles underpinning cell signalling regulation and stresses the need to research the phosphoproteomic signalling system in an unbiased manner.

In this thesis the central aim is to use data-driven and unbiased approaches to study the human phosphoproteomic signalling network. The first chapter describes a project where I co-developed a machine learning model to predict signed kinase-kinase regulatory circuits based on kinase specificities and high throughput phosphoproteomics and transcriptomic data. The network was validated using independent high throughput data and used to identify novel kinase-kinase regulatory interactions. This project was done in collaboration with Brandon Invergo, a postdoc in Pedro Beltrao's research group.

In the second chapter I expand upon work done in the first chapter. I used various predictors such as: Co-expression, kinase specificities and different variables characterising kinase-substrate potential target phosphosites to predict kinase-substrate relationships and their signs. I then used independent experimental kinase-substrate predictions to validate the predictions and identify high confidence kinase-substrate relationships. I then combined the kinase-substrate predictions with the kinase-kinase regulatory circuits to identify condition-specific signalling networks. To enable easy use of my method and networks and analyses of phosphoproteomics data by non-expert users I also developed the SELPHI2 server, where the user can extract biological insight from their datasets. SELPHI2 presents a substantial improvement upon the SELPHI server, which was developed in 2015 by my supervisor, Evangelia Petsalaki.

Thirdly, to study the architecture of human cell signalling networks at a whole-cell level and address the limited predictive power of the current models of cell signalling such as pathways found in KEGG (Kanehisa, 2019), Reactome (Jassal et al., 2020) and WikiPathways (Slenter et al., 2018), the third chapter aims to identify signalling modules from phosphoproteomic data. These data-extracted modules were found to have a greater predictive power for independent data sets in terms of number of significant enrichments. Furthermore, we sought to predict the probability of module co-membership from predictors such as membership within data-driven modules, co-phosphorylation and co-expression.

In summary, the work presented here seeks to explore the understudied phospho-signalling systems through system-wide prediction of kinase-substrate regulation and the identification of phospho-signalling modules through data-driven means.

Acknowledgments

Firstly, I would like to thank my supervisor, Evangelia Petsalaki for the wonderful opportunity to study and work at the EMBL-EBI, her enthusiasm for science, her scientific input throughout my journey towards a PhD and all the professional, academic and moral support she has provided me with over the last four years.

In addition. I would like to thank my Thesis Advisory Committee members: Pedro Beltrao, Jasmin Fischer and Wolfgang Huber for their time and invaluable scientific input and support that has improved my research projects. I would also like to thank Pedro Beltrao for our collaboration and his input over the last few years. I would also like to thank Brandon Invergo for fruitful collaboration as well as advice and mentoring early on in my PhD. I would not be the scientist I am today without the enthusiastic community of researchers I have belonged to during my PhD. I am immensely grateful to my fellow lab members: Girolamo Giudice, Sumana Sharma, Charlie Barker, Ioannis Kamzolas, Lourdes Sriraja, Paula Weidemueller, Iguaracy Pinheiro de Sousa, Cansu Dincer, Tao Fang, Grigoriy Nos, Vitalii Kleshchevnikov, Prajna Hebbar, Fadime Oztoprak, Bishoy Wadie, Haoqi Chen, Vivian Robin, Guillermo Calderon and Rzgar Hosseini for their advice, input, collaboration and companionship which proved invaluable over the course of my PhD.

Speaking of community, I am very grateful for the opportunity to befriend and get to know scientists and fellow PhD students from all over the world. In particular, I would like to thank my PhD batch for awesome two months in Heidelberg. Furthermore, I would like to thank Vasileios, Veronika, Kai and Natalie for their ongoing friendship and Silvija for being a great host in Heidelberg. This list would not be complete without mentioning the EMBL-EBI PhD cohort for their companionship in particular my batch at the EMBL-EBI: Aleix, Ally, Conor, Jose and Michael who have provided me with fun experiences and terrific moral and scientific support over the years.

Furthermore, I would like to thank EMBL-EBI for the funding of my project as well as providing me with ample opportunities to network and discuss science at the beautiful EMBL-EBI campus. I would like to thank the EMBL-EBI staff that have made my life easier over the duration of my PhD. In particular, I would like to thank Tracey Andrews, Anna

Alasalmi, Virginie Uhlman, Alex Bateman and Nick Goldman and the staff at human resources.

I would also like to thank Magdalene College for providing roof over my head, community and support over the duration of my PhD. As well as providing me with the opportunity to engage in the Cambridge lifestyle with formals and all.

Last but not least, I would like to thank my family: Anna Borgthorsdottir, Petur Hilmarsson, Asdis Petursdottir, Sigridur Thordis Peturssdottir, Borgthor Olsen Stephansson, Kristjan Borgthorsson, Hilmar Petursson and Asdis Jonsdaottir.

Preface

In chapter 2 I discuss a project I did in collaboration with Brandon Invergo, a former postdoc in the Beltrao group at the EMBL-EBI. My contribution to the project consisted of contributing to the generation of position weight matrices-based predictor in collaboration with Brandon Invergo. Furthermore, I did the testing and development of kinase-kinase regulatory relationship predictions by using different machine learning methods, apart from the Bayesian additive regression trees (BART) method. The clustering of the network was done exclusively by me as well as the assessment of the biological function of the resulting modules.

I also did the Mapping the network onto phosphoproteomic perturbation data and identifying novel kinase-kinase signalling pathways. This work has already been published in Cell systems (Invergo et al., 2020). All work related to signed predictions was done by Brandon Invergo and are thus only mentioned here as a summary for completion but discussed in greater details in the paper. The code generated during this study is available at GitHub (https://github.com/evocellnet/kinase-activity-net/).

Other contributions include: Girolamo Giudice performed analysis on distances between enriched pathways within the IntAct network. David Bradley did all work involving specificity determining residues. Nosheen Akhtar and Petro Cutillas designed the phosphoproteomics experiments needed for the identification of novel pathways. The experiments were executed by Nosheen Akhtar and Maruan Hijazi.

Table of contents

List of figures

List of tables

1 Introduction

1.1 Principles of human cell signalling

Cells need to be able to make appropriate decisions to respond to both environmental and internal stimuli. They achieve this through a cell signalling system that consists largely of complex networks of interacting proteins. An important property of this network is the specificity of its participating proteins such as receptors (Siddle et al., 2001) and other signalling proteins such as kinases and phosphatases in the way they target their substrates. This is in part due to signalling proteins making use of functional units called domain structures that allow them to recognize and bind to specific targets. Domains can be described as specifically folded parts of the protein often around 50-300 amino acids in length (Nair et al., 2019). An example of an important domain that plays a large part in the signalling system is the Src homology domain (SH2) that recognises phospho-tyrosines and docks tightly to some of the proteins that contain a phosphorylated tyrosine residue (Pawson et al., 2001). Another example is the Src homology domains (SH3) that are often found with SH2 (Mayer, 2015) domains and allow for protein-protein interaction through the binding of short proline rich peptides (Ren et al., 1993).

To activate cell signalling a stimulus is needed. These stimuli can be in the form of chemical, electrical, or mechanical signals. In cell biology these stimuli are often chemical in nature, i.e. in the form of a ligand. Such ligands can be proteins, lipids or sugar polymers which specifically bind to a receptor on the cell surface (Nair et al., 2019).

Through ligand binding the receptor typically undergoes conformational changes leading to activation of cytoplasmic enzymatic activity in proteins such as kinases and other proteins that relay information throughout the cell by mechanisms such as post translational modification (PTM) of proteins and other molecules (Knorre et al., 2009). Subsequently PTM leads to changes in molecular structure and function of the target (Karve and Cheema, 2011), which in turn can lead to system wide changes in the cell. A

well-studied example of such cascade of cell signalling upon ligand binding a receptor is the Epidermal growth factor receptor (EGFR) pathway which regulates cell proliferation as well as differentiation and growth and has been known to lead to widespread changes in phospho-regulation in cells (Wee and Wang, 2017).

1.2 Context specificity of cell signalling

Signalling cascades are activated by external and internal stimuli leading to specific responses. However, signalling systems are highly context specific. For instance, the PI3K is known to cause cell growth in adverse conditions such as serum or growth factor removal (Eves Eva M. et al., 1998), while conferring no advantage in a more bountiful environment(Berenjeno et al., 2017; Madsen et al., 2019). In other words, different components of the systems are active depending on conditions. The signalling network is often thought of as being modular. That is, the network contains relatively highly interconnected components that interact with other parts of the network to a lesser extent with their own function that is insulated from other parts of the network. At the same time, the signalling system is highly interconnected with cross talks linking pathways (Vert and Chory, 2011). While this interconnectivity might seem contradictory to modular response to changes in its environment, the cell has mechanisms to improve the specificity and the fidelity of the signal transduction. This is achieved with mechanisms including compartmentalization either through translocation to specific locations within the cell (Rinaldi et al., 2018; Smith and Scott, 2002) and/or through the use of scaffold proteins which bind together two or more components in a pathway, such as a kinase and its substrate (Good et al., 2011), increasing their local concentration and improving the flow of information as a result. Other factors driving context specificity is the variance in protein abundance across tissues and cells (Akbani et al., 2014; Kim et al., 2014). Furthermore, signalling molecules, such as kinases, have been shown to have different activity levels across biological samples and experimental conditions (Ochoa et al., 2016). A direct consequence of this variation in protein abundance and function is the context specificity of cellular signalling, since signalling processes are dependent on their component molecules. Well-studied canonical pathways are likely to be broadly active as they have

been found by multiple experiments, whereas the architecture of active components of the signalling network can vary significantly between conditions, such between healthy and cancerous tissues (Marbach et al., 2016; Saez-Rodriguez et al., 2011). Many examples of single context-specific events in cell signalling have been found. For instance, PI3K activation is known to lead to different cellular responses such as during the insulin driven translocation of GLUT4 from the cytosol to the plasma membrane in 3T3-L1 adipocytes (Tengholm and Meyer, 2002). This translocation does not take place when cells are stimulated with PDGF (Wiese et al., 1995), with the data suggesting that PIP_3 concentration remains too low for GLUT4 translocation (Tengholm and Meyer, 2002). Similarly the Notch pathway has been found to cross talk with other signalling pathways, such as JNK and NF-κB, in a context specific manner in *Drosophila* leading to the development of different biological systems such as immunity and precursor of sensory organs, demonstrating the important role context specificity plays in cellular responses to diverse conditions (Mishra et al., 2021). Often, the same signalling molecule leads to different outcomes in cells. Acetylcholine is a well-known example, where acetylcholine, which binds to similar receptors on the surface of different cell types, is known to lead to different outcomes in different cell types. For instance, in the salivary gland, where acetylcholine binding leads to secretion and in the heart where the binding is interpreted as a signal to reduce rate and force of contraction (Alberts et al., 2002).

1.3 Phospho-signalling

One of the more important and best studied parts of cell signalling is PTM through transient phosphorylation of target proteins or other molecules such as lipids. These modifications are carried out by ‘writer’ proteins, called kinases, which catalyse the covalent addition of a phosphoryl group, ‘readers’ which include phosphorylation-specific binding domains such as SH2 domain that, for example, modulate signal transduction and ‘eraser’ proteins, phosphatases that remove the phosphoryl group from the protein. Phosphorylation of proteins occurs through the addition of ATP’s terminal phosphate group onto the polar group of amino acids most commonly as serine/threonine and

tyrosine. Other phosphorylation events have been reported such as Histidine phosphorylation (Boyer et al., 1962), but its function remains unclear in mammalian cells. Phosphorylation is widespread and is integral to many cell processes such as cell division (Doerr, 2008), localization of proteins and DNA repair (Canovas and Nebreda, 2021). Up to 75% of proteins in the human proteome have known phosphosites and are therefore potentially regulated by phosphorylation (Vlastaridis et al., 2017). Currently, more than 500 protein kinases are known that are divided into seven families of typical kinases and seven atypical kinases (Fabbro et al., 2015). The majority of known protein kinases are serine/threonine kinases (Manning et al., 2002) which is reflected by the fact that most known phosphosites are on serine or threonine residues with only 39,281 of 239,573 (16 %) human phosphosites listed in PhosphoSitePlus being tyrosine residues (Hornbeck et al., 2015). Fewer human phosphatases have been found (>200) (Damle and Köhn, 2019), which are divided into six families (Sacco et al., 2012). Due to the importance of cell signalling in critical processes such as cell division and others, disturbances in the system often cause diseases such as such as diabetes or cancers (Yaffe, 2019) with 37 kinase inhibitors being FDA approved for cancer treatment and with about 150 other kinase-targeting drugs under clinical trial in 2018 (Bhullar et al., 2018) and G protein-coupled receptors also being a common drug target in cancers (Lappano and Maggiolini, 2011). Indeed, mutations in kinases are often drivers of cancers and cancer development (Torkamani et al., 2009). This stresses the need to understand the components of the phosphorylation network and how they propagate signals throughout the network.

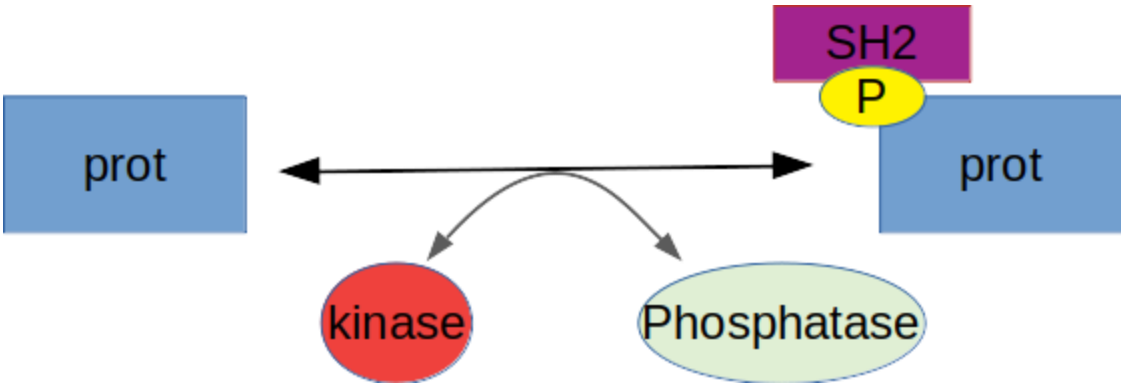

Figure 1.1: *Principles of phospho-signalling***.** *Phospho-signalling is often characterized as a system of writers (kinases) that add phosphoryl group on target substrates, readers (SH2 domains) which recognizes and binds phosphorylated tyrosine and erasers (Phosphatases) which remove the phosphoryl group. Making the PTM reversible.*

1.4 Current models on cell signalling

The state of the art knowledge of the cell's signalling system has been curated and organized into different databases such as SIGNOR (Licata et al., 2020), Reactome (Jassal et al., 2020), KEGG (Kanehisa, 2019), OmniPath (Türei et al., 2016), PhosphositePlus (Hornbeck et al., 2015) and WikiPathways (Slenter et al., 2018). Some of these databases, such as KEGG, Reactome and WikiPathways organize the protein interaction networks further into modules or organizational units within the network that lead to a certain product, cell state or other molecular output. By organizing the network into a set of modules one has a powerful tool that could in theory be used to explain which parts of the signalling network are active or inactive, in a given context. These gene sets are commonly used to analyse high throughput data and compare different phenotypes such as normal to diseased tissues or different tissues with regards to active biological processes (Zhu and Stephens, 2018). Due to their importance, around 70 methods had been proposed to conduct these enrichment analyses in 2019 (Nguyen et al., 2019). One of the more common methods to calculate the overrepresentation is the Fisher's exact test (Fisher, 1935). Another common method of enrichment is the Gene Set Enrichment Analysis (GSEA), which was originally developed to analyse microarray data (Subramanian et al., 2005). These tests are limited in a way that does not account for the

different sizes of the pathways or the different number of pathway annotations for each individual gene. These pitfalls have been addressed with methods such as SetRank, which discards sets that are only significant because of their overlap with other gene sets (Simillion et al., 2017). Another method is Annotation Enrichment Analysis which aims to account for set size and number of sets assigned to each gene (Glass and Girvan, 2014). Another issue with gene set enrichment or pathway enrichment is lack of objectivity (Domingo-Fernández et al., 2018); in other words it is very hard to validate the resulting gene sets unless a certain target pathway is expected to be identified. Organisms are highly complex, and most results can be supported by some references (Nguyen et al., 2019). Furthermore, different pathway databases organize the cell's processes into different sets leading to different results depending on the pathway database (Domingo-Fernández et al., 2018; Mubeen et al., 2019). To counter this problem, integrated pathways have been developed such as the Pathway Commons (Rodchenkov et al., 2020), MSigDB (Liberzon et al., 2011) and ConsensusPathDB (Herwig et al., 2016) and Compath (Domingo-Fernández et al., 2018).

These databases have fallen short when it comes to explaining the various phosphoproteomic data sets and experiments (Köksal et al., 2018; Olsen et al., 2006). At the same time, these models have been successfully used to capture differences in expression levels across samples (Subramanian et al., 2005). These results are surprising as RNA expression profiles correlate poorly with protein abundance (Gry et al., 2009). A recent analysis showed that rather than capturing pathways, the differential expression analyses might be capturing transcription factor modules as differential expression studies did a better job at capturing TF modules than pathways (Szalai and Saez-Rodriguez, 2020). This fact is generally masked, due to the fact that there is a high overlap between TF modules and pathways.

Other databases such as OmniPath (Türei et al., 2016), PhosphoSitePlus (Hornbeck et al., 2015) and SIGNOR (Licata et al., 2020) provide the user with information on kinase-substrate interactions, lists of regulatory sites and phosphosites and causal relationships between proteins. These causal networks provide an invaluable tool to generate predictive and explanatory models. Furthermore, they could be used as a prior network for future discoveries (Hill et al., 2016). However, these databases give limited view on

the phospho-signalling network due to literature bias and as a result, overreliance on these databases might hinder novel discoveries.

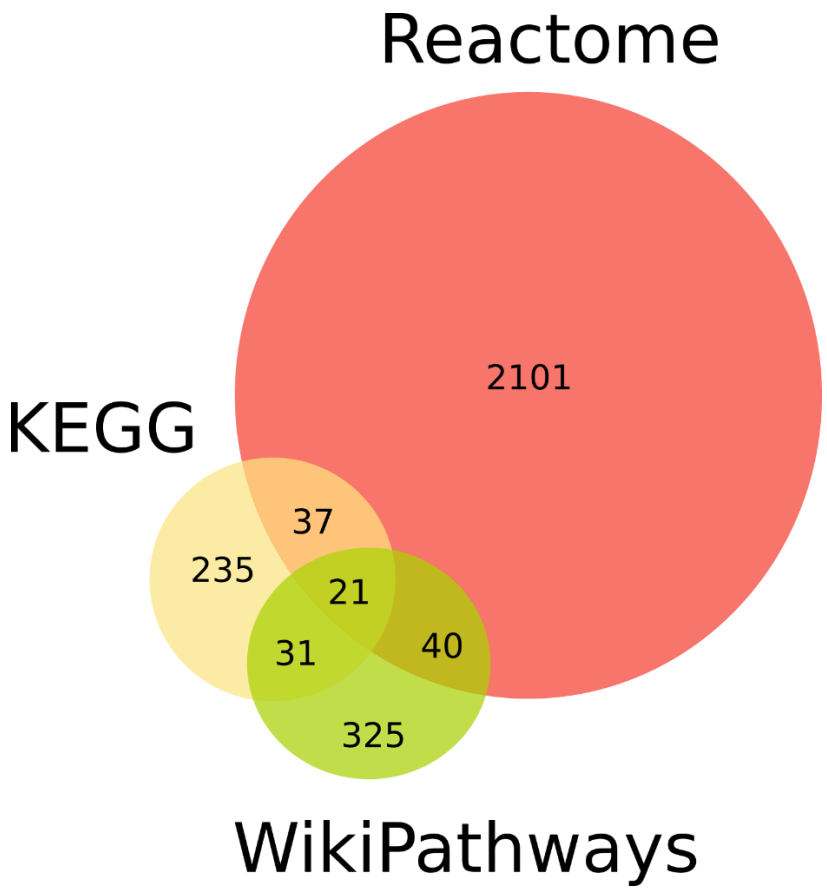

Figure 1.2: *Coherence between established pathway databases. Overlaps of equivalent pathways across three pathway databases. Note the low overlap indicating the non-objective nature of pathway assignment. Adapted from Domingo-Fernández and colleagues (Domingo-Fernández et al., 2018)*

1.5 The dark phosphoproteome

As with other signalling systems, the activity of kinases is context specific which leads to variation in pathway activity across different conditions. This makes this crucial part of the cell machinery inherently elusive and difficult to study. While there are over 100,000 known phosphosites listed in PhosphoSitePlus, only 5% have a known upstream kinase (Needham et al., 2019). Furthermore 90% of the phosphosites with known upstream

kinases have been assigned to around 20% of the kinases with around 150 kinases left without any known downstream substrates (Lappano and Maggiolini, 2011). This large number of phosphosites without a kinase could partially be explained due to non-functional phosphosites that accumulate on proteins throughout evolutionary time (Landry et al., 2009). We have, however, reasons to believe that a large portion of the phosphoproteome is understudied. One seminal study showed that EGF stimulation lead, within minutes, to changes in 14% of the measured phosphosites (Olsen et al., 2006) many of which are not part of the canonical EGFR pathway, indicating that the network is more intricate than pathway databases such as KEGG indicate (Kanehisa, 2019). Indeed, other studies of kinase and phosphatase perturbation data have shown that perturbation leads to system-wide changes in phosphorylation (Bodenmiller et al., 2010) that are not necessarily captured by current knowledge of signalling. Furthermore, the number of substrates assigned to kinases does not correlate with disease relevance according to pathogenic human mutation prevalence and mouse model phenotypes (Needham et al., 2019). These findings indicate that there is a strong bias in the literature and curated databases, where a small portion of the already well-studied part of the phosphoproteome tends to be scrutinized to a greater extent at the expense of the other less studied part of the proteome (Edwards et al., 2011). In fact, most curated kinase-substrate interactions are between highly cited kinases and substrates (Invergo and Beltrao, 2018). Unbiased, high throughput protein-protein interactome studies seem to support this study bias hypothesis as various studies have suggested kinase interaction networks where edges are distributed more evenly across proteins than what the literature suggests (Invergo et al., 2020).

The sheer scale of the phosphoproteome and the context-specific and transient nature of these processes make the systematic study of the phosphoproteome a challenge. The illumination of the dark phosphoproteome will rely heavily on computational and statistical methods to identify kinase-substrate relationships and phospho-signalling modules to prioritize for further experimental validation. To date, many computational methods have been proposed to prioritize kinase-substrate relationships for further analysis by assigning a probability score to potential interactions (Ayati et al., 2018; Horn et al., 2014; Petsalaki et al., 2009). Similarly, methods have been developed to assign functionality scores to

phosphosites to identify sites that might be promising for future functional analysis (Miao et al., 2018; Ochoa et al., 2020; Xiao et al., 2016). These efforts in tandem with continued experimental validation and exploration will be crucial in the effort to map the phospho-interactome and the functional role of the dark phosphoproteome. With increased understanding comes practical application of the knowledge gathered, through development of drugs for new targets and understanding of complex processes such as chronic diseases.

1.6 Modularity of biological systems

Biological systems are often considered to be structured in a modular fashion (Hartwell et al., 1999). This modularity is observed at many different levels: Proteins, cells, organisms and ecosystems (Lorenz et al., 2011). Modularity has also been observed in transcriptomics (Singh et al., 2008), protein-protein interaction networks (Barabási and Oltvai, 2004) and metabolic networks (Spirin et al., 2006). As a result of ubiquity, modularity has been defined differently across different biological fields. Evolutionary biologists might define a module as a conserved sequence, geneticists as co-expressed or co-regulated genes while a network biologist might be defined as a highly interconnected component of a protein interaction network that is less connected to other parts of the networks (Lorenz et al., 2011).

While biological networks such as the signalling network are often assumed to be modular, this is in no way to be certain, due to the large undiscovered regions of the phospho-regulatory networks. Indeed, various experimental results indicate that the human protein-protein interaction is less modular than is often thought and the literature indicates (Luck et al., 2020; Rolland et al., 2014). Furthermore, computational analyses have found that modular networks do not yield the best results or the greatest functional efficiency (Bullinaria, 2007; Kashtan et al., 2009) and smooth fitness models have not been able to capture the benefits of modularity in networks (Orr, 2000). Nevertheless, there seems to be a consensus that biological systems are indeed modular. Some have argued that modularity improves fitness as modularity enhances robustness (Kitano, 2004), that is the ability of the system to maintain functionality in the face of perturbation, as it insulates the system from localized perturbations. Another argument for modularity

is increased evolvability as it allows for experimentation in one component in the system, while maintaining function in other parts of the system (Wagner and Altenberg, 1996). Furthermore, organisms can gain new functions by combining existing modules (Lorenz et al., 2011). These abilities of modular networks make the organism more adaptable and able to respond to environmental changes. Yet another hypothesis, is a more spontaneous emergence in networks through natural growth mechanisms such as duplication and diversification (Solé and Valverde, 2008). Other, more direct, benefits of modularity evolution have also been considered: cells from exotic environments might bring different DNA into another environment and the cells already inhabiting the environment might accommodate the DNA into their genome and benefit from the trait it has to offer. Thus, modularity can arise from horizontal gene transfer (Rainey and Cooper, 2004), which has been shown to have played a large role in *E.coli* transcriptional regulation: Most of *E.coli's* transcription factors has evolved by horizontal gene transfer rather than by duplication of genes (Price et al., 2008).

While theoretical and computational experiments support modularity, empirical studies have also yielded results supporting the modularity hypothesis. Viruses have been found to be divided into separate evolutionary and structural modules (Ferron et al., 2005). Empirical analyses have found that biological networks also exhibit modularity. For instance, it was found that metabolic networks are scale free modular and that networks contained hierarchical modules with clustering coefficients being independent of network size (Barabási and Oltvai, 2004) though these results might stem from the inherent study bias in the databases. Furthermore, it was found that bacteria found in more varied environments had more modular metabolic networks (Parter et al., 2007), demonstrating the evolutionary benefits of modularity. Empirical studies also show that protein-protein interaction networks also exhibit modularity with protein complexes being statistically significantly likely to contain functional modules (Spirin and Mirny, 2003), and with evidence that protein modules are encoded on the genome level (von Mering et al., 2003).

1.7 Modularity of signalling networks

As a biological system, signalling networks are often seen as modular systems divided into functional subunits (Hartwell et al., 1999). Indeed, databases such as KEGG

(Kanehisa, 2019), Reactome (Jassal et al., 2020) and WikiPathways (Slenter et al., 2018) are organized into such modules or pathways including pathways dedicated to signalling. We know, however, that the picture is more complex, due to the high incidence of pathway cross-talks (Vert and Chory, 2011). Other factors, such as the prevalence of undiscovered edges in the signalling networks and context specificity, where there are differences in pathway activities across conditions add to the complexity (Saez-Rodriguez et al., 2011). Olsen et al. found, for instance, that signal travels far and wide throughout the signalling network in a relatively short time frame with around 14% of the measured phosphosites being modulated by the stimuli (Olsen et al., 2006). And pathways as laid out in the databases have been found to have a poor explanative performance when mass spectrometry data is analysed. Many studies analysing changes in phosphorylation of peptides in large scale phosphoproteomic data have been unable to explain changes with canonical pathways. For instance, a recent study found that the EGFR pathway as defined by various databases performed poorly at explaining the changes in phosphorylation levels (Köksal et al., 2018). In fact, only about 5% of the significantly altered phosphopeptides were found in the canonical pathway and most of the proteins listed as members of the reference pathway were not found to be altered, while EGFR pathway reconstruction, made from literature interaction network fitted to the data, fared better. Similar results have been found when other pathways have been analysed. For instance, Humphrey and colleagues found that in a large scale data extracted from insulin treated mice, less than 10% of the regulated phosphosites were found in the insulin pathway (Humphrey et al., 2015).

These experimental results, coupled with the fact that the signalling network appears to be more interconnected than the literature indicates, might paint a picture of a system that is too interconnected and complex to be considered modular. However, findings drawn from forward genetics and cell predictability support the modularity hypothesis (Atay and Skotheim, 2014). For instance, Whi5 has been found to be a robust predictor of cell division in yeast, more so than cell size or time after division (Doncic et al., 2011) and CDK2 has been found to be similarly informative in mammalian cells (Johnson, 2014). This indicates that cell decision can be inferred from a single gene product and is therefore highly predictable, which supports the view that cellular systems are organized

in modular fashion where a single stimulus leads to predictable output. Another example would be the conservation of the MAPK kinase module from yeast to human (Widmann et al., 1999) demonstrating that functional units are conserved throughout evolution. Furthermore, cells can be divided into classes such as tissue of origin based on data such as expression profiles (Jaitin et al., 2014). One way to reconcile this apparent contradiction between high interconnectivity of networks and cross talks between pathways and modularity is that many interactions are non-functional (Atay and Skotheim, 2014). Furthermore, due to the context specificity of signalling protein activities (Ochoa et al., 2016) not all of them will be active at the same time.

It has been found that the feedforward system regulating Far1, which prompts the cell to re-enter the cell cycle is insulated from the cell cycle in S. cerevisiae (Doncic and Skotheim, 2013). This modularity is achieved by the switch-like activity of B-type cyclin-Cdk, a protein that degrades Far1 in cell cycle arrest, providing evidence of modularity through switch-like behaviour of proteins and provides an example of how modularity can be achieved.

1.8 Towards data-driven for biological module identification

The strong arguments for a modular architecture of cell signalling combined with the limitations of the current pathway models stresses the need for the identification of less biased data-driven modules. To date, many approaches and methodologies have been suggested to reconstruct the set of pathways present under a given condition or in a given organism.

An established method to reconstruct biological modules or pathways present in an organism is to generate pathways from its genome by utilizing data available in databases such as GO terms or protein sequences. Examples of such methods are MinPath (Ye and Doak, 2009) and Pathologic (Karp et al., 2016) that reconstruct a set of pathways present in an organism from a set of genes present in an organism. While useful for conservative estimation of pathways that are present, these methods do not contribute to the discovery of new pathways.

1.8.1 Identification of context specific signalling subnetworks

A step towards a more data-driven method is using high throughput phosphoproteomics data to capture biological subnetworks that are active in a given context, which is an already established field. For instance, Temporal Pathway Synthesizer (TPS) has been used to extract active components of the signalling network from time series data (Köksal et al., 2018). And data-driven models have been used to identify active components of the signalling network in cancers by using perturbation data or static data (Ayati et al., 2021; Saez-Rodriguez et al., 2011; Zhu and Stephens, 2018). As an example of a developed application that can be used to analyse signalling data, PHOTON applied mapping onto the interaction network to find active pathways and functional proteins (Rudolph et al., 2016). Such methods do have the limitation of relying on literature-based networks such as IntAct (Orchard et al., 2014) or BioGRiD (Oughtred et al., 2019). Given the large scale of the underexplored phosphoproteome this overreliance on the literature network can be problematic if a more complete picture of phosphoregulation is to be achieved.

1.8.2 Data-driven co-expression modules

The complexity, context specificity and the size of the unknown space within the signalling network stress the need to use statistical methods to identify modules from high throughput dataset without a prior literature defined network. Methods have already been developed to capture transcriptomics modules (Zhou and Altman, 2018). Methods such as the principal component analysis to cluster cell types (žurauskienė and Yau, 2016), weighted gene correlation networks (Langfelder and Horvath, 2008) and independent component analysis have been used to reduce the dimensionality of gene expression data by dividing genes into modules of co-regulated and co-expressed genes.

Many of these studies have been conducted on mRNA data sets. It has been, however, established that mRNA levels have a notoriously low correlation with corresponding protein levels with mRNA levels explaining around 40% of the variation (Gry et al., 2009; Koussounadis et al., 2015). This can be attributed to many factors, including biological regulation that occur between RNA transcription and protein translation. One analysis

showed, however, that significantly differentially expressed mRNA molecules correlate better with their corresponding proteins (Koussounadis et al., 2015). Nevertheless, in order to capture cell signalling on a protein level, large scale proteomics data sets, such as those from mass spectrometry, need to be included to gather a more holistic picture of the biological system.

1.8.3 Limitations of mass spectrometry data

In comparison to RNAseq data, phosphoproteomics data can be challenging to analyse due to data sparsity and the stochastic nature of the peptide sampling by the mass spectrometer. Various methods have been proposed to deal with data sparsity in mass spectrometry data but no consensus or standardization has been reached (Lazar et al., 2016; Webb-Robertson et al., 2015; Wei et al., 2018). There are few issues to consider which make missingness non-random. For instance, despite their importance in signalling, proteins found within the membrane are less likely to be identified by mass spectrometry (Schey et al., 2013) as well as peptides that are close to the detection level, skewing the peptide quantification distribution towards more abundant peptides (Lazar et al., 2016).

Apart from biological issues with the mass spectrometer, there are technical limitations when it comes to peptide and protein identification. The most common protein quantification approaches include bottom-up methods where proteins are digested by enzymes before being introduced to the mass spectrometer. The bottom-up approach has several problems related to the fact that they do not sequence the whole proteome but identify short peptides and map them onto databases, which might introduce errors due to amino acids having similar or the same mass (Timp and Timp, 2020). Additionally, the choice of digestive enzyme is known to impact protein identification (Dau et al., 2020). Furthermore, due to the number of spectra needed to identify peptides, up to 75% of collected spectra are not mapped to a peptide (Griss et al., 2016). Yet another problem arises when a PTM is assigned to a peptide as in many cases more than one site is likely modifiable on each peptide, leading to statistical assignment of phosphosites (Timp and Timp, 2020). Top-down mass spectrometry does not require digestion before measurements and is therefore useful in single protein analysis and the capture of

isoforms and PTMs. However, while it does identify full sequences it is less sensitive than bottom up and has a lower coverage and throughput (Catherman et al., 2014) limiting its utility in proteome wide studies. As a result, mass spectrometry data, while capturing tens of thousands of phosphosites cannot be expected to give the full picture of the cell's proteome.

1.8.4 Use of proteomics data to derive modules

Despite these challenges, the derivation of purely data-driven models to analyse biological samples is quite common in the biological literature (Hoogendijk et al., 2019; Seyfried et al., 2017). Some results have shown that, interestingly, modules derived from proteomics do not necessarily overlap perfectly. For instance an analysis on 129 samples from cortical tissues from 50 individuals with or without Alzheimer's disease conducted by Seyfried et al found that, while some modules were shared across the RNAseq and proteomics modules many were not (Seyfried et al., 2017). This suggests that these different types of data might yield different results and complement each other. Similarly, Hoogendijk and colleagues used modules derived from transcriptomic and proteomic data to capture myeloid differentiation and the development of neutrophil programming (Hoogendijk et al., 2019) using the clustering weighted correlation networks built from the correlation between the RNA and proteomic profiles of their samples. Recently, a purely proteomics-based data-driven module to characterize germ cell maturation in crustacean *Gammarus fossarum* (Degli Esposti et al., 2019).

To my knowledge, modules derived from phosphoproteomic data have not been used to explain independent data sets in the same way pathways are often used for example for pathway enrichment. As previously discussed, a large portion of the phosphoproteome has not been accounted for in the literature (Needham et al., 2019) and similarly, a large fraction of the protein interactome remains to be discovered. Recently, there has been an explosion in publicly available phosphoproteomics data sets. It is therefore feasible to use said data both to generate a set of modules of phosphorylated modules and evaluate them on independent datasets. Similar work has been done on transcriptomics data, where generated modules of genes outperformed genes when faced with a sample classification task when the number of samples is small. This is based on the assumption

that phosphorylation can be used to identify functional modules. Earlier studies have found that phosphosites are more likely to be functionally linked than would be expected at random (Li et al., 2017). Furthermore, co-phosphorylation has been used to identify kinase-substrate relationships (Ayati et al., 2018; Invergo et al., 2020; Petsalaki et al., 2015) either by itself or in concert with other predictive features. These results indicate that it is possible to use data-driven phospho-signalling modules to capture and explain phosphoproteomic data and thus limit our reliance on limited and biased databases.

1.9 Machine learning to address biological problems

Machine learning uses data to generate models of the system in question. In doing so, it differs from other statistical methods because it does not make any stringent assumption about the data, like statistical models do, but instead learns from experience by interacting with the data (Xu and Jackson, 2019). Therefore, machine learning can be useful when patterns need to be derived from complex data sets. Machine learning has been used to identify complex patterns in biological data for predictive modelling. This greatly increases the capability for exploratory research and the identification of novel drug targets (Vamathevan et al., 2019) and the predictions of protein interactions (Invergo et al., 2020; Wang et al., 2019) for instance.

Machine learning is often divided into two groups: Supervised and unsupervised. Supervised machine learning usually entails training models by fitting data to previously defined groups and then using that trained model to make predictions for other independent data sets (Tarca et al., 2007). Supervised methods have been used, for instance, in the case of kinase-substrate relationship and other protein-protein interaction predictions (Horn et al., 2014a; Wang et al., 2020, 2019). On the other hand unsupervised methods base their classification on the data and do not rely on previously defined positive and negative sets (Tarca et al., 2007), such as the various clustering methods that have been successfully employed to divide data sets into categories. For instance, unsupervised methods have been employed to classify cancer categories based on high throughput data sets such as microarrays (Perou et al., 1999). More recently breast cancer samples were clustered into molecular types based on the STRING (Szklarczyk

et al., 2021) protein-protein interaction network and mutational patterns (Rohani and Eslahchi, 2020).

The recent explosion in biological omics data sets has given rise to the need for more sophisticated methods to analyse and integrate the different data sets. Biological data is intrinsically challenging for statistical methods due to its high dimensionality or the number of independent variables, and low number of samples (Xu and Jackson, 2019). Furthermore, the low number of samples only aggravates the challenges brought about by the high dimensionality, such as over-fitting and multiple hypotheses testing. In addition, biological data can be sparse and noisy which complicates its usage. Particularly when various different omics data sets are integrated, which in theory should give the best, most holistic picture of the system under scrutiny (Zitnik et al., 2019), the "curse of dimensionality" arises due to the proliferation of data unrelated to the question (Altman and Krzywinski, 2018).

While machine learning has been proven to be useful in making predictions, various hurdles are still present. In the case of supervised learning, prior knowledge can introduce bias in the discovery process since there is a known bias in the current biological literature. Furthermore the "black box" nature of the models makes interpretation of the resulting models difficult, if not downright impossible (Xu and Jackson, 2019). Furthermore, as touched upon earlier, high dimensionality introduces the potential of overfitting since the flexibility of the equations increases as the size of the training set grows. Even dimension-reducing methods such as PCA are sensitive to the curse of dimensionality (Altman and Krzywinski, 2018). This problem is not fully erased even with more sophisticated algorithms and greater sample sizes. Therefore, it is important to be aware of this problem when a large number of predictions are made for exploratory purposes and false positives tend to accumulate.

1.10 Machine learning and statistical methods for data-driven inference of signalling systems

1.10.1 Use of machine learning to identify signalling circuits

As discussed above a large fraction of the phospho-signalling network is under-studied and unexplored. It therefore stands to reason that relatively unbiased data-driven methods are crucial in the exploration of the understudied space. Phosphoproteomic data has been utilized to predict signalling circuits. For instance, Rudolf et al. used diffusion to capture pathways by using high throughput data mapped on a literature network (Rudolph et al., 2016). A more data-driven example, modular response analysis (MRA) (Kholodenko et al., 2002) has been used to reconstruct an active network from steady state data. MRA, however, is slow and can be difficult to implement on a proteome wide scale. Similarly, methods such as dynamic Bayesian networks (DBN) have been used to construct pathway networks from time series data (Hill et al., 2012) which faces the same problem of scaling. Another commonly used method is to discretize the data and construct the active pathways as logic models even to discover new pathway interactions (Saez-Rodriguez et al., 2009). In this thesis, I present a more integrated method to predict signed kinase signalling circuits based on kinase specificity models, phospho- co-regulation, co-expression and functionality of phosphosites (Invergo et al., 2020).

1.10.2 Computational modelling of kinase specificities

Prediction of kinase-substrate relationships constitutes a well-established field and an important part of the exploration of the dark phosphoproteome (Ayati et al., 2018; Horn et al., 2014; Wang et al., 2020). Kinases are known to target specific motifs surrounding the target phosphosites (Ubersax and Ferrell Jr, 2007) and as a result the modelling of kinase specificities has been used extensively for the computational prediction of kinase-substrates.

One of the more common methods to predict kinase-substrates is the Position Weight matrix (Stormo et al., 1982) which is a model for its respective kinase recognition site. Using the PWM one can quantify whether a potential substrate is significantly more similar to the position weight matrix than the background, and thus more likely to be recognised

by the relevant kinase. Generally, the matrix is constructed for a given kinase from known substrates found in the literature or they could be derived from experimentally predicted substrates.

PWMs are usually constructed by defining a window surrounding the phospho-acceptor site. For each position within the given window the frequency of each amino acid at each position is calculated. In many cases pseudo counts are added to each index in the matrix as some amino acids can be expected to appear zero times at any given position, while not having a zero probability of appearing at the same position.

One of the most apparent limitations of these models is the likely erroneous assumption of independence between positions. Another drawback is that the position weight matrix does not take any structural information into account. Some methods have been proposed to address these problems such as hidden Markov models to solve the independence problem (Huang et al., 2005) and machine learning methods such as sequence Bayesian networks (Patrick et al., 2017) have been utilized to detect more complex patterns underlying kinase specificities as well.

1.10.3 Computational prediction of kinase-substrates

Most kinase-substrate prediction methods base their predictions on known kinase-substrates. The shortcoming of this method is over reliance on known substrates which might limit the novelty of the predictions. Furthermore, some predictions are complemented by biological network databases such as STRING (Szklarczyk et al., 2021), which similarly are over-reliant on the biological literature. There are methods base their prediction, at least partially, on high throughput phosphoproteomic data such as CoPhosK (Ayati et al., 2019) but they are limited to phosphosites captured by LC/MS studies.

To date, various methods have been developed. Below are several examples of kinase-substrate relationship prediction methods. The following list includes several examples:

NetworKIN and NetPhorest (Horn et al., 2014): A popular and widely used kinase-substrate prediction algorithm. NetworKIN utilized cellular context (distances within the STRING network) and kinase specificity, while NetPhorest uses phylogenetic algorithm to classify phosphosites in terms of kinase binding motifs.

Group-based Prediction System (GPS) (Wang et al., 2020)**:** The GPS algorithm has been developed and maintained for the last decade. It is similar to other methods in that GPS based its prediction on the similarity between motifs surrounding potential phosphosites. GPS uses BLOSUM62 (Henikoff and Henikoff, 1992) to calculate the similarity between phosphosites and their surrounding peptides and then clusters the peptides, reasoning that phosphosites clustering together are likely to be phosphorylated by the same kinase.

LinkPhinder (Nováček et al., 2020)**:** This method uses peptides surrounding phosphosites to construct a knowledge graph where the consensus motifs of kinase families are used to construct the graph and then link prediction is used to construct new links or kinase-substrate relationships within the network.

Scansite (Obenauer et al., 2003)**:** Scans for motifs within proteins that are likely to be phosphorylated by a kinase. It then uses the position weight matrix PWM of around 60 kinases to score the phosphosite.

Netphos (Blom et al., 2004)**:** Nethphos like other similar methods employs neural networks and uses experimentally validated S, T and Y sites for kinase specific predictions. These predictors are then used as an input for an ensemble of neural networks. Predictions are only available for 17 kinases.

CoPhosK (Ayati et al., 2019)**:** The algorithm uses naive Bayes model to make predictions for phosphosites found in mass spectrometry proteomics datasets based on kinase-substrate co-phosphorylation associations. Unlike most other kinase-substrate prediction methods, this method does not base their prediction on kinase specificities. It was found that when incorporated with other static methods like KinomeXploreer (Horn et al., 2014), CoPhosK improved their performance.

PhosphoPICK (Patrick et al., 2017): PhosphoPICK uses Bayesian networks to model the amino acid sequence surrounding the phosphosite. PhosphoPICK also incorporates the dimer and trimers surrounding the phospho acceptor sites into its model for prediction. Another Bayesian network is used to make predictions based on information on protein substrates and their availability during the various stages of the cell cycle. Furthermore other variables such as kinase specify models and protein-protein interaction networks (BioGRID (Oughtred et al., 2019), STRING (Szklarczyk et al., 2021)).

One of the main limitations of these methods are the accumulation of false positives and negatives. Due to the scale of the predicted space, as well as our limited knowledge of signalling, validation of novel kinase-substrates is difficult. It therefore becomes clear that independent evaluation is needed for the identification of novel kinase-substrate interactions.

1.11 Experimental prediction of kinase-substrate relationships

Several experimental methods have been proposed and developed to assign kinase to phosphosites experimentally both for single kinases and high throughput screens. A common method to validate a single substrate for a single kinase is to apply antibodies for the target phosphosite to measure phosphorylation level coupled with kinase inhibition (Nováček et al., 2020). Another low throughput method is K-Clasp which binds kinases to biotin tagged peptides that are incubated in a lysate in the presence of ATP-ArN3 (Dedigama-Arachchige and Pflum, 2016). The problem with such low throughput methods is that they are not scalable to the evaluation of large numbers of predicted kinase-substrate relationships.

Various high throughput methods exist such as RNAi inhibition of kinases coupled with measurements of changes in phosphorylation relative to control state (Azorsa et al., 2010; Papageorgiou et al., 2015). The limitation of these methods is that it can be hard to discern if the decreases in phosphorylation are due to the inhibition of a kinase that specifically targets the impacted phosphosite or if changes in phosphorylation are due to the hampering of processes further upstream. Similarly, kinase inhibitors have been employed to predict and identify novel kinase-substrates. Problems with inhibitors include their promiscuity as well as problems with discerning indirect and direct effects. To get around some of these issues, few methods have been put forward. Hijazi and colleagues estimated the most likely upstream kinase by utilizing kinase inhibitor selectivity information (Hijazi et al., 2020) while others have correlated the inhibitor phosphorylation finger print with the inhibition fingerprint of well-known kinases (Watson et al., 2020).

Other high throughput methods to directly link phosphosites with kinases have also been suggested. Sugiyama et al proposed a method where peptides extracted from lysed cells

were introduced with alkaline phosphatase and then exposed to kinases. The resulting phosphorylation levels were then compared with control conditions (Sugiyama et al., 2019). However, like computational studies this method accumulated a large number of unverifiable kinase-substrates and only captured the minority of known kinase-substrate relationships (4% in the case of Suggiyama et al.). This might partly be due to the in vitro conditions which might not capture conditions in living organisms.

1.12 Conclusions

A recurrent theme in the research of the signalling system is the fact that a large portion of the components in the signalling network, both kinases and phosphosites, remain under-studied. This has led to challenges when it comes to tasks such as interpretation of data sets due to insufficient representation of the data on known signalling networks. To this end many methods have been proposed as discussed in Chapter 1.10.3 to make computational predictions based on features such as kinase specificities (Nováček et al., 2020; Wang et al., 2020). While such methods are useful in the exploration of the phosphoproteome, most methods have several limitations. Firstly, basing most of the predictions on established literature will not sufficiently address the issue of study bias as most predictions are based on biased models. High throughput data is therefore needed to aid in the generation of new hypotheses. Similarly, biased models, such as those arising from the aforementioned study bias have proven to be of limited value in the study of high throughput proteomics (Köksal et al., 2018; Olsen et al., 2006). Similarly, data-driven methods are needed to derive more useful models. In light of this, the main aim of this thesis is to explore the less studied portion of the phospho-signalling system. My contributions are divided into the following chapters.

Chapter 1 discusses the development of a method to predict kinase-kinase regulatory networks. This work was done in collaboration with Brandon Invergo a member of the Beltrao group at the EMBL-EBI and has been published (Invergo*, Petursson* et al., 2020). Various predictors were employed to predict kinase-kinase regulatory relationships and their sign. The network was then used to identify potentially novel regulatory pathways by integrating the network with independent experimental data sets (Hijazi et al., 2020; Sugiyama et al., 2019).

Chapter 2 expands upon the work done in Chapter 1 where the modified predictors that were used in Chapter 1 as well as features characterizing the potential target phosphosites (Ochoa et al., 2020) are used to predict kinase-substrate relationships between 367 kinases and more than 80,000 phosphosites found on 8957 proteins, with more than 22 million predictions in all. Signed predictions are made for relationships where the target phosphosite is likely to be functional, comprising over 2 million predictions in all.

The resulting kinase-substrate prediction set assigns higher probability to experimentally predicted kinase-substrate relationships compared to the background after known interactions have been removed from the set. Furthermore, the method makes high confidence predictions for proteins with fewer citations in the literature than the kinase-substrates listed in PhosphoSitePlus (Hornbeck et al., 2015). Moreover, I find that the kinase-substrate predictions perform better than other established methods at discerning between known kinase-substrate relationships as well as experimentally.

Chapter 3 discusses the move towards data-driven modules of phospho-regulation. I discuss the identification of these modules as well as the work done to assign biological function to the modules through enrichment studies as well as GWAS analysis to associate the modules with diseases. I then describe efforts to use the modules to capture known pathways with machine learning methods and compare it with the use of high throughput data, both RNA expression and phosphoproteomics, to do the same.

The overarching theme of this thesis is to analyse the understudied portion of the phosphoproteome by making novel predictions for kinases and their substrates as well as the regulatory sign of these interactions. Furthermore, I present work on the development of data-driven modules that outperform established literature defined modules when it comes to the explanation of high throughput phosphoproteomics data.

2. Inference of kinase-kinase regulatory network

Contributions

The work presented here was done in collaboration with Brandon Invergo, a postdoc member of Pedro Beltrao's group. Pedro Beltrao, Evangelia Petsalaki and Brandon Invergo conceived and supervised the project. Analysis was conducted by Brandon Invergo and me. Girolamo Giudice performed additional analysis on distances between pathways in the IntAct network. David Bradley did the work on specificity determining residues. Nosheen Akhtar and Petro Cutillas conceived the phosphoproteomic experiments that were carried out by Nosheen Akhtar and Maruan Hijazi. More specifically my contribution entailed part of the generation of position weight matrices-based predictors with Brandon Invergo. I did the testing and making predictions by using different machine learning methods apart from the BART method, clustering the network and assessing the biological function of the resulting modules, Mapping the network onto phosphoproteomic perturbation data and identifying novel kinase-kinase signalling pathways. This work has already been published in Cell systems (Invergo et al., 2020). All work related to signed predictions was done by Brandon Invergo and are thus only mentioned here as a summary for completion but discussed in greater details in the paper. The code generated during this study is available at GitHub (https://github.com/evocellnet/kinase-activity-net/).

2.1 Introduction

Phosphorylation in the most studied and most common post translational modification (PTM). PTMs, including protein phosphorylation, lead to changes in the substrate's activity and function. This ability to change protein function allows information to flow through a network of kinase-substrate relationships where the cell response is partly determined by the phosphorylation state of all proteins. This in turn allows the cell to respond to external and internal stimuli and make appropriate decisions, such as when to divide, under any given condition. These processes are discussed in greater detail in the introduction (Chapters 1.1 -1.3).

Kinases comprise a class of proteins that propagates the cell information flow by adding a phosphoryl group on the substrate protein. Therefore, regulation of kinases by other kinases forms an integral part of the phospho-regulation network. A large portion of kinase-kinase regulatory network is understudied with most known kinase-kinase interactions being between highly studied kinases (Invergo and Beltrao, 2018). This stresses the need for a less biased method to predict kinase-kinase interaction and their regulatory signs. Previously data-driven methods have been proposed (Hill et al., 2012; Kholodenko et al., 2002) but they face difficulties when applied kinome-wide (See Intorduction chapter 1.10.1). In this chapter, I describe a machine learning method to predict kinase-kinase regulatory relationships by integrating kinase specificity models, functional score and high throughput data sets such as tissue expression and mass spectrometry data sets.

2.2 Methods

2.2.1 Data

We retrieved list of 504 protein kinases, hereby referred to as the human kinome from the UniProt/Swiss-Prot Protein Knowledgebase, *pkinfam* (accessed 8 November 2017 at https://www.uniprot.org/docs/pkinfam). Two publications provided us with phosphoproteomic data. One included phosphorylation levels of 213 phosphosites on 100 kinases from MCF7 cells treated with 22 kinase inhibitors (Wilkes et al., 2015), while the second had quantifications for 1537 phosphosites on 193 kinases across 83 breast tumour samples (Mertins et al., 2016). RNA expression data from human tissues was acquired from the GTEx project (GTEx Consortium, 2013) as provided by Expression Atlas (E-MTAB-5214, timestamp 26 April 2018) (Papatheodorou et al., 2018) and from the Human Protein Atlas project (accessed from www.proteinatlas.org 1 December 2017) (Uhlén et al., 2015). Lists of kinase-substrate relationships, human phosphosites and kinase regulatory sites were retrieved from PhosphoSitePlus (accessed 1 May 2018) (Hornbeck et al., 2015). We downloaded frequencies of amino acids in the human proteome downloaded from the UniProt proteome database (UniProt Consortium, 2018). Experimental *In vitro* kinase-substrate predictions were downloaded from an earlier publication (Sugiyama et al., 2019). Another experimental kinase-substrate prediction set were acquired from Hijazi and colleague (Hijazi et al., 2020)

2.2.2 Training sets for machine learning

To acquire a high confidence training set for model generations we extracted a set of high confidence kinase-kinase regulatory relationships from the OmniPath (Turei et al., 2016) knowledge base (retrieved Jan 22, 2018). To make certain that the relationships used for training were of high quality we only retained interactions that were found in two or more databases. Altogether 825 interactions were included in the positive set. Negative training sets are harder to define as there is no way of establishing that a kinase-kinase

relationship does not occur in any condition. Nevertheless, by presuming that biological networks are generally sparse, we generated a negative set by randomly sampling sets of possible kinase-kinase regulatory relationships excluding those found in the positive set. When training for BART, working under the assumption that biological networks are sparse we generated a set of negatives that was 8 times as large as the positive set as increasing the size of the training set returned small improvements in performance while increasing memory usage. For the other machine learning methods, I used a balanced training set with equal number of positives and negatives.

For signed predictions we yet again retrieved information from Omni Path. As before, we excluded interactions supported by fewer than two sources. The resulting training set consisted of 394 activating and 109 inhibitory relationships.

2.2.3 Formulation of predictors

2.2.3.1 Construction of kinase specificity models

In order to model kinase specificity, we built position weight matrices. We retrieved information on kinase-substrate relationships and peptides surrounding the phosphosites (+/- 7 AA) from the PhosphoSitePlus database (Hornbeck et al., 2015).

We constructed PWMs only from kinases with at least 10 known substrates. To minimize the effects of redundant substrates we used a weighing method previously described by Henikoff and Henikoff (Henikoff and Henikoff, 1994). The PWMs were constructed as follows:

Given a set of substrate sequences: $S = \{S_1, S_2, \ldots, S_i, \ldots S_{n-1}, S_n\}$ with the residues of the ith substrate is represented as follows: $S_i = \{S_{i1}, S_{i2}, \ldots, S_{i14}, S_{i15}\}$ We give weight to amino acid j in substrate sequence i in the following way:

$$w(a, j) = \frac{1}{c_j \sum_{i=1}^{n} (S_{ij} = a)}$$

Where c_j is the number of unique amino acids found at position j among the substrates S. Subsequently a weight is calculated for each substrate by adding the position specific residue weights together.

$$W(S_i) = \sum_{j=1}^{15} w(S_{ij}, j)$$

Lastly, each sequence weight is normalized by the sum of the other sequence weights.

$$\hat{W}(S_i) = \frac{W(S_i)}{\sum_{k=1}^{n} W(S_k)}$$

Subsequently, PWMs were constructed from these weights. In our case, the size of the matrix is 20 × 15 representing 20 amino acids and 15 positions surrounding the phosphosite. First a matrix r is constructed such that the entry r_{aj} contains weighted counts of the amino acid a at position j in the sequence.

$$r_{aj} = n \sum_{i=1}^{n} V(S_{ij}, a)$$

$$V(S_{ij}, a) = \begin{cases} \hat{W}(S_i), & \text{if } S_{ij} = a \\ 0, & \text{otherwise} \end{cases}$$

Since there is a non-zero probability of observing any amino acid at any given position, pseudo counts are introduced to change the expected probability of amino acids that do not occur at a given position in our samples of substrates from zero to a non-zero value. Our pseudo counts were estimated from amino acid frequencies in the proteome in a position specific manner (Henikoff and Henikoff, 1996). For each column represented as j we define the pseudo count B_j as follows:

$$B_j = m \times c_j$$

Where m is a tune-able parameter set at one in our case and c_j is the number of unique amino acids at position j. Then for each entry in the PSSM the pseudo count for amino acid at a position j becomes:

$$b_{aj} = B_j \times f_a$$

Where f_a represents the proteome-wide frequency of amino acid a. This allows us to construct an empirical PSSM matrix p of probabilities of observing amino acid m at position j.

$$p(a, j) = \frac{b_{aj} + r_{aj}}{B_j + \sum_a r_{aj}}$$

Then the PWM is derived by calculating the $\log_2$ ratio between entry p(a,j) and the frequency of a, f_a

$$PWM_{aj} = \hat{p}(a,j) = log_2\left(\frac{p(a,j)}{f_a}\right)$$

2.2.3.2 PWM assignment to kinases based on family membership or specificity determining residue similarity

We restricted PWM construction to only those kinases that had at least ten annotated substrates in PhosphoSitePlus (Hornbeck et al., 2015). This resulted in PWMs being constructed for 140 kinases. Furthermore, we assigned PWMs to kinases based on families. We built PWM for each family as using families as defined in the KinBase resource (Manning et al., 2002). As a result, family PWMs were assigned to 209 kinases that could not be assigned a PWM individually. Finally, PWMs were assigned to an additional 14 kinases based on the similarity of their specificity determining residues (SDR) (Bradley et al., 2021) with other kinases. In order to assign based on SDR, we systematically explored the association between SDR similarity and PWM similarity measured by Frobenius distance (Ellis and Kobe, 2011). For reference, we calculated the pairwise PWM distance by subsampling 25 kinase-substrates from the same kinase with 25 being the median number of substrates per PWM. We found that the average distance was 1.0 with 1.10 being the 97.5th percentile. Therefore, PWM distance below 1.10 was determined as having the same site specificity. The SDR similarity yielding PWM distances of less than 1.10 in more than 50% of the cases was 0.8. As a result, kinases that had SDR similarity of 0.8 or higher with another kinase were assigned the PWM of the kinase with the greatest SDR similarity.

2.2.3.3 Scoring of phosphosites with PWM

We scored all known phosphosites found on the 504 kinases included in this study against all the PWMs (n = 363) built. The phosphosites were scored by fitting the +/- 7 amino acid residues surrounding the candidate phosphosites to the PWMs. The score, s, was calculated as follows:

$$s = \sum_{j \neq 8} \hat{p}(a, j)$$

Finally, to make the scores equivalent across the different kinases, we min-max normalized the scores for each kinase.

$$s_{min} = \sum_{j \neq 8} \hat{p}(\arg\min_{a} \hat{p}(a, j), j)$$

$$s_{max} = \sum_{j \neq 8} \hat{p}(\arg\max_{a} \hat{p}(a, j), j)$$

$$\hat{s} = \frac{s - s_{min}}{s_{max} - s_{min}}$$

2.2.3.4 Functionality of phosphosites

At the time of this work, the Beltrao group was preparing for publication an algorithm that provides predictions of functionality of phosphosites, which we also used in our study (Ochoa et al., 2020). The predictions were based on a wide variety of phosphosite specific features including structural, evolutionary and biochemical attributes. Since these predictions were done on a defined set of phosphosites based on an analysis on a set of high throughput experiments, some of the phosphosites available in the PhosphoSitePlus database were not represented. We $\log_{10}$-transformed the raw functional scores and min-max normalized the values to arrive at functional scores valued between 0.0 and 1.0, with larger scores reflecting a higher predicted probability of a functional impact of phosphorylation of that site.

2.2.3.5 Connection of PWMs to functional scores

Kinases are often regulated by multiple phosphosites, generating potentially multiple scores per kinase. In our case, we assigned the top PWM score for any given kinase-kinase pair. It is therefore possible that the PWM score assigned for a hypothetical kinase-kinase relationship results from fitting a non-functional phosphosite to the upstream kinase's PWM. Therefore, to predict the upstream kinase's ability to phosphorylate functional sites on the target kinase we generated an independent predictor. The PSSM score, ŝ was calculated for each substrate phosphosite that had an assigned functional

score. Subsequently, we ranked by the n sites by ŝ in descending order, producing a PWM score based ordered ranking of functional scores F = { F_1, F_2,…., F_i,…., F_{n-1}, F_n}. The discounted cumulative gain (DCG) (Järvelin and Kekäläinen, 2002) for the kinase-kinase pair was then defined as:

$$DCG = \sum_{i=1}^{n} \frac{F_i}{log_2(i+1)}$$

Sites assigned a higher PWM score, and therefore lower $rank_i$, make a greater contribution to the DCG. As a result, the more functional the phosphosites with the highest PWM scores, the higher the DCG.

In order to make the scores more comparable across the different kinase-kinase pairs, the DCG scores were min-max normalized. As a result, the normalized DCG was calculated as follows:

$$nDCG = \frac{DCG - DCG_{min}}{DCG_{max} - DCG_{min}}$$

Where DCG_{max} is the maximum DCG achievable by that substrate and DCG_{min} the minimum score.

2.2.3.6 Co-expression and tissue specificity

Tissue expression data from the Human Protein Atlas (Uhlén et al., 2015) and GTEx (GTEx Consortium, 2013) was used to calculate co-expression across kinases across different tissues. We used Spearman's rank sum coefficient to quantify the association.

We calculated the tissue specificity of each kinase or how widely expressed the kinase is across the different tissues by quantifying the skewness of its distribution of expression levels in the Protein Atlas expression database (in transcripts per million, or "TPM") across the samples. The *e1071* package for R was used to calculate the skewness (https://CRAN.R-project.org/package=e1071).

2.2.3.7 Phospho-co-regulation

We assessed the level of phospho-co-regulation of the kinase pairs by measuring the Spearman's correlation between phosphorylation of known regulatory phosphosites in two data sets; one quantified levels of changes in phosphorylation across inhibitor

conditions (Wilkes et al., 2015) and the other contained phosphoproteomic data from breast cancer patients (Mertins et al., 2016). Both data set consisted of a table of $\log_2$ fold-changes for each quantified phosphosite. To ensure samples followed similar distribution, each data set was quantile normalized by condition or sample (Bolstad et al., 2003).

Within each data set, we calculated the correlation between fold-changes of the sites for each phosphosite pair found on the respective kinases across all conditions for samples which had five or more matching quantifications available. Conditions where either of the kinases was under inhibition were removed from the calculation. Spearman's *rho* was used to calculate the correlation and a p-value estimated for the correlation by the asymptotic t approximation. The resulting p-values were $\log_{10}$ transformed; if the estimated p-value was 0, we set the final value to 6. Phosphosites are known to be non-functional in many cases so co-phosphorylation does not necessarily indicate co-regulation. For this reason, the log transformed p-value was then scaled by the functional scores of both sites by multiplying the log transformed p-value with the functional scores of the phosphosites involved. This ensured that only kinase-kinase pairs with highly correlated functional sites received high co-regulation scores. We then used the maximum co-regulation score to quantify co-regulation between the kinases. Finally, the co-regulation scores for all kinase pairs were min-max normalized.

2.2.4 Training of model

All the aforementioned predictors were combined and used to train a machine learning model. We also added a feature indicating if the upstream kinase was threonine or tyrosine. Many supervised learning methods have been proposed to solve classification problems. In our case the following methods were considered:

a) Random forest (Tin Kam Ho, 1995): Briefly put, random forest constructs an ensemble of decision trees based on sub samples of the training set and averages over the results to improve accuracy and prevent over-fitting. I used the RandomForestClassifier() with 150 trees as implemented in scikit-learn (Pedregosa et al., 2011)

b) Support Vector machines (Cortes and Vapnik, 2004): Support vector machines map the examples provided by the training set onto a space to maximize the gap between the positives and negatives. The new data points are then mapped onto the same space and predicted to belong to either category depending on which site of the gap between negatives and positives they land. I used the *svm.SVC()* function as implemented by scikit-learn (Pedregosa et al., 2011). I used radial basis function kernel which uses non-linear division between groups.

c) AdaBoost (Schapire and Singer, 1999): AdaBosst or Adaptive boosting is a meta estimator that fits a classifier to the initial data set and then fits additional classifiers to the same data set where weights are adjusted for incorrectly classified examples to focus more on difficult cases. I used the AdaBoostClassifier function implemented by scikit-learn with 150 estimators. Decision trees were used as base estimators for predictions.

d) Neural networks (Rumelhart et al., 1986): In this project I used Multi-layer perceptron with 100 hidden layers. Here we used the scikit-learn implementation.

e) Logistic regression (Cox, 1958): Briefly put logistic regression models the probability of a variable to belong to a certain class with logistic function. Here I used logistic regression as implemented by scikit-learn with the default parameters.

f) Bayesian additive regression trees (Chipman et al., 2010): Briefly, BART uses the sum of trees method to assign probability. A series of decision trees are fit to the data for data classification. Each tree is constructed from binary decision nodes which make decisions based on one of the features. The leaf nodes of each tree contain values which contribute to the classification value. As BART is a sum-of-trees model, these decision values are summed to produce a final value used for classification. The BART method uses a fixed number of trees, on which it places regularizing priors to ensure that each tree is a “weak learner” so that each tree only makes a small contribution to the final classification value. It does this by limiting the tree depth, shrinking terminal leaf nodes to the median, and adding noise to avoid over-fitting. Bayesian approaches such as Markov-Chain Monte

Carlo (MCMC) backfitting (Chipman et al., 2010) are used for parameter estimates in order to fit the trees to the data.

Due to the random sampling of negatives, 100 different models were generated. The BART predictors were assessed with 20 runs of 3-fold cross validation while the other methods were assessed with 5-fold cross validtion. The AUC scores were used to quantify the predictive power of each predictor. The R package ROCR (Sing et al., 2005) was used to calculate AUC scores and draw ROC curves. The final probability score assigned to each kinase-kinase pair being the average of the 100 different models. In the end we selected BART, in part due its way of incorporating missingness into its predictions

For figures 2.4 (pathway extraction) and 2.3 C (the ranking of known substrates), we built 3 different models for each kinase. The models were trained with a reduced positive set where the kinase's interactions had been removed and a random "negative" set. The mean posterior probability of the kinase's relationships from these 3 models was used as the final prediction value.

2.2.5 Prediction of signed regulatory relationships

This part of the project was exclusively done by my collaborator Brandon Invergo. A more detailed description of the prediction process can be seen in that published paper (Invergo et al., 2020)

We also predicted the sign of kinase-kinase regulatory relationships. The sign prediction can be divided into two steps. The first step of the signed prediction was the prediction of the sign of phosphosites, i.e. whether the phosphorylation of a specific site leads to inhibition or activation of the substrate. The second step predicts the sign of the kinase-kinase relationship. The regulatory phosphosites information from PhosphositePlus was used to predict signs of phosphosites (Hornbeck et al., 2015).

We used the following features for sign prediction: The position in percentage of the site relative to the start/end position of the protein kinase domain (i.e. between 0 and 1 for sites that fall within the domain); the position in percentages of the site along the protein's sequence length; the domain (if any) in which the phosphosite lies, including, but not

limited to, protein kinase domains; the phosphosite residue (serine/threonine or tyrosine); whether or not the substrate is a tyrosine kinase; estimated secondary sequence disorder, as quantified by DISOPRED (Ward et al., 2004); and the -$\log_{10}$(p-value) of the site being in a phosphorylation hot-spot (Strumillo et al., 2019).

To find an appropriate threshold to decide if a phosphosite had activating or inhibiting properties we used the Mathews correlation coefficient.

Subsequently, we made predictions of the sign of the kinase-kinase regulatory relationships. In brief the functional score, the co-regulation score and DCG were modified to indicate the sign of the functional score for which we simply used the sign of the phosphosites. The signed functional score was then used in the signed DCG calculation. The signed co-regulation was calculated in a similar way except the Spearman's *rho* coefficient was used instead of the log transformed p-values. The predictions were made using BART and the model evaluated in the same way as the kinase-kinase regulatory relationships prediction.

2.2.6 Identification of functional modules in network

Biological networks are often seen as being modular in structure. In order to see if the same property applied to our network, I set out to identify and divide our kinase-kinase network into clusters. In order to make the partition I considered three methods: Greedy clustering (Clauset et al., 2004) , Louvain (Blondel et al., 2008) clustering and the Markov (Van Dongen, 2000) cluster algorithm. All these methods were implemented in their respective R packages: *igraph* (Csárdi and Nepusz, 2006) (Greedy clustering and Louvain) and *MCL (https://CRAN.R-project.org/package=MCL).*

Louvain clustering (Blondel et al., 2008): The algorithm is divided into two steps: first, each node is assigned to its own cluster. In the second step each node i is iteratively merged with its neighbors' clusters and the change in the overall network's modularity is assessed. The merge leading to the greatest improvement in modularity is identified followed by a merging node i to the cluster leading to greatest improvement in modularity. This process is repeated until a local maximum has been found. During the second step, a new network is generated from the identified clusters. The edge weights between the

nodes are computed by summing over the weights of the links that connect nodes across each cluster. The first step is then reapplied on the resulting network. These two steps are then repeated iteratively to improve the cluster assignments until no further improvement can be made. The modularity function that the algorithm optimises can be seen below:

$$Q = \frac{1}{2m}\left(\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j)\right)$$

Where m is the sum of weights, A_{ij} is the weight between node i and j and the function δ(u,v) is 1 if u=v, 0 otherwise. C_i is the cluster that node i belongs to, k_i is the sum of the edges attached to node i.

Greedy clustering (Clauset et al., 2004)**:** Similar to Louvain method as a fast and greedy method is used to find a local modularity minimum. It starts in the same way as each node is its own cluster which are merged to maximize improvement in modularity. The merging is then repeated until only one community is left.

Markov clustering (Van Dongen, 2000)**:** The algorithm determines the probability of random walks through the adjacency matrix and from the random walks it constructs a weighted transition matrix. The transition matrix is then iteratively subjected to two operators, expansion and inflation. To transform one set of probabilities into another. Briefly, expansion corresponds to matrix squaring, while inflation corresponds to Hadamard power of a matrix which is followed by a scaling so that values in the matrix continue to represent probabilities. While expansion computes longer random walks, network inflation favours intra cluster walks (shorter distances). This iteration of inflation and expansion leads to graphs that are divided into different segments which are in turn interpreted as different clusters.

2.2.7 Assessment of network modularity

I set out to assess if the high confidence network was modular by comparing it to an empirical distribution of modularity values derived from a randomized network. I applied a cut-off of 0.5 to derive a high confidence network. The remaining edges were

subsequently min-max scaled so that the edges had assigned edge weights on the scale from 0 to 1. A thousand randomized networks were generated with the same degree distribution as the reference network. Randomization was done by using the *sample_degseq()* function in the igraph (Csárdi and Nepusz, 2006) package by employing the “vl" method (Fabien Viger and Matthieu Latapy, 2005). At each randomization, I shuffled the edge weights of the reference network and assigned the shuffled probabilities to the randomized network’s edges as edge weights. I then clustered the resulting random network and the modularity score of the clustering assignment was calculated with *modularity.igraph* as implemented in *igraph* (Clauset et al., 2004; Csárdi and Nepusz, 2006) (Equation in chapter 2.2.6).

2.2.8 Pathway-annotation distances

We retrieved the human protein interaction network from IntAct (version: Oct. 2018) (Orchard et al., 2014). Additionally, we added human phosphorylation events from SIGNOR, PhosphoSitePlus and OmniPath (Turei et al., 2016) as edges to the network, resulting in a network consisting of 17,089 nodes and 166,757 edges. For each pair of pathway annotations, we computed the mean of the length of all shortest paths between the proteins annotated for the pair.

I subsequently divided the distances into two sets: Set of distances that are enriched in the same cluster (n=811) and a set of distances between pathways found enriched in two different clusters (n = 1019). Distances between pathways that shared kinases were excluded from the subsequent analysis which reduced our within-cluster set to 67. Wilcoxon rank sum test was used to determine if the difference between the two distance sets was significant.

2.2.9 Kinase inhibitor experiments

We generated phosphoproteomic data to test the network’s predictions. The data generation was carried out as described previously by Wilkes et al. (Wilkes et al., 2015). Briefly, the Kasumi-1 cell line was grown in RPMI medium supplemented with 10% FBS, and was treated with 1µM trametinib or GDC-0941 for 1 h. The cells were lysed in a urea-based lysis buffer. Subsequently trypsin digestion was applied and phosphopeptides

were enriched using TiO2 chromatography and analysed in a LS-MS/MS system consisting of an Ultimate 3000 ultra-high pressure chromatograph connected to a Q-Exactive Plus mass spectrometer. The Mascot (Cutillas and Vanhaesebroeck, 2007; Perkins et al., 1999) search engine and Pescal (Cutillas and Vanhaesebroeck, 2007) were used for data analysis as described by Wilkes and colleagues (Wilkes et al., 2015).

2.2.10 Support of kinase-kinase predictions with independent experimental kinase-substrate predictions

To see if our predictions could be supported by independent experimental predictions of kinase-substrates I downloaded a list of kinase-substrate predictions from two recent publications (Hijazi et al., 2020; Sugiyama et al., 2019). The two studies were conducted in a different manner. Briefly, Sugiyama and colleagues de-phosphorylated HeLa cells lysate with alkaline phosphatase while spiked phosphatase was inactivated by heat. The lysate was then reacted with recombinant protein kinases (n= 354). The other study conducted by Hijazi and colleagues grew three cell lines (MCF7, NTERA2 and HL60) which were incubated with 61 different kinase inhibitors. This way, Hijazi et al. were able to make kinase-substrate predictions for 103 kinases by analysing the impact of inhibition on phosphorylation levels on phosphosites.

I considered any kinase-kinase relationships that were predicted by either experiment to be experimentally supported. Wilcoxon's rank sum test was used to establish if the probabilities of experimentally supported edges were significantly higher than those for the rest of the unsupported edges.

2.2.11 Identification of phosphosites that are impacted upon kinase inhibition

By analysing phosphoproteomic data derived from Kasumi-1 cells exposed to the kinase inhibitors *trametinib* (MEKi) and *GDC-0941* (PI3Ki), I identified phosphosites that were down-regulated by either inhibitor. For this analysis I considered any peptide, including multi-phosphorylated peptides. The data set was $\log_2$ transformed and quantile normalized to ensure similar distribution across samples. I identified down-regulated

phosphosite in each condition with the *ebayes()* function to fit linear model to the data to estimate modified t-statistics as implemented in the *limma* R package limma (reproducibility-optimized statistical testing) (3.40.6) (Ritchie et al., 2015). The $\log_2$ ratio threshold of less than -1 and false discovery rate of lower than 0.1 was used to identify the down-regulated phosphosites. P-values were adjusted with the Benjamini–Hochberg method (Benjamini and Hochberg, 1995).

2.2.12 Identification of novel kinase-kinase regulatory circuits

To see if any novel signalling pathways could be identified from the kinase perturbation data, we looked for the shortest path from the kinases perturbed by *trametinib* (MAP2K1 and MAP2K2) and GDC-0941 (PI3K) to phosphosites that are down-regulated by their inhibition. PI3K is a lipid kinase and as a result it was not included in our network of protein kinases. I added edges between PI3K and kinases regulated by two PI3K kinases: hsa:5290 (PIK3CA) and hsa:5291 (PIK3CB) or their direct substrate, Phosphatidylinositol-3,4,5-trisphosphate, in the KEGG database (accessed 16. October, 2019)(Kanehisa, 2019) . As a result, edges from PI3K to PRKCD (e.g. hsa:04750),PRKCI (e.g. hsa:04910), PRKCZ (e.g. hsa:04910), SRC (e.g. hsa:04926), AKT1 (e.g. hsa:04151), AKT2 (e.g. hsa:04151), ILK (e.g. hsa:04510), MTOR (e.g. hsa:04150/hsa04910), PDPK1 (e.g. hsa:04150), PDPK2 (e.g. hsa:04068), ITK (e.g. hsa:04062) and PTK2 (e.g. hsa:04062) were added to the network.

To link phosphosites to our kinase-kinase interaction network, I added known kinase-substrate relationships from PhosphoSitePlus (Hornbeck et al., 2015) to the network as well as interactions predicted by both in cell lines (Hijazi et al., 2020) and *in vitro* (Sugiyama et al., 2019) experiments that were considered to be of high enough confidence. Phosphosites that are known substrates of the perturbed kinases were excluded from this analysis. Substrates of kinases manually linked to PI3K were discarded as well. We applied a probability threshold of 0.5 to select high confidence edges. I used the function *all_shortest_paths()* as implemented by the *igraph* R package to identify the shortest directed paths from the impacted phosphosites to the perturbed kinases. I set the parameters as follows: The parameter *mode = "out"* and the edge

weights were set by subtracting the min-max scaled edge probabilities from one. An interaction was deemed novel if it was supported by either cell line or *in vitro* experiment.

2.3 Results

2.3.1 Regulatory relationships can be identified by similar phosphorylation patterns at functional phosphosites and kinase co-expression

The use of high throughput features rests upon the assumption that kinases that are active or inhibited in the same set of tissues and conditions are more likely to be part of the same pathway and therefore more likely to interact in a regulatory manner. Phosphorylation forms an important part of kinase regulation and therefore we calculated the phospho-co-regulation by correlating co-phosphorylation and weighing it with phosphosite functional score. Co-phosphorylation was calculated using two large-scale phosphoproteomic experiments (Mertins et al., 2016; Wilkes et al., 2015).

Overall, known regulatory relationships found in the OmniPath database (Turei et al., 2016) tended to have higher co-regulation scores than the background of unannotated pairs. This held true in both phosphoproteomic experiments, kinase-kinase regulatory pairs tended to have higher co-regulation scores than pairs with without representation in Omni Path (one-sided Wilcoxon rank sum test, $W= 2.8 \times 10^{7}$, $p < 1 \times 10^{-6}$ (Mertins et al., 2016), $W = 9.3 \times 10^{5}$, $p < 1 \times 10^{-6}$ (Wilkes et al., 2015)) (Figures 2.1 D & E).

Co-expression was also considered as a predictor and for this reason, two RNA-seq data sets (GTEx Consortium, 2013, Uhlén et al., 2015) were tested to see if co-expression was an indicator of co-regulation between kinases. In general, co-expression discriminated between known regulatory relationships and the background of unannotated relationships indicating its value as a predictive feature (one-sided Wilcoxon rank sum test, $W= 1.3 \times 10^{8,}$ $p < 1 \times 10^{-6}$ (GTEx Consortium, 2013), $W = 1.3 \times 10^{8,}$ $p < 1 \times 10^{-6}$ (Uhlén et al., 2015)) (Figure 2.1 A & B). Similarly, when tissue specificity is analysed, the absolute difference between tissue specificities between the kinase forming each pair pairs with regulatory relationships tend to have more similar expression profiles than those with no annotated

relationship (one-sided Wilcoxon rank sum test, W = 8.9 × 10^7, p < 1 × 10−6) (Figure 2.1 C)

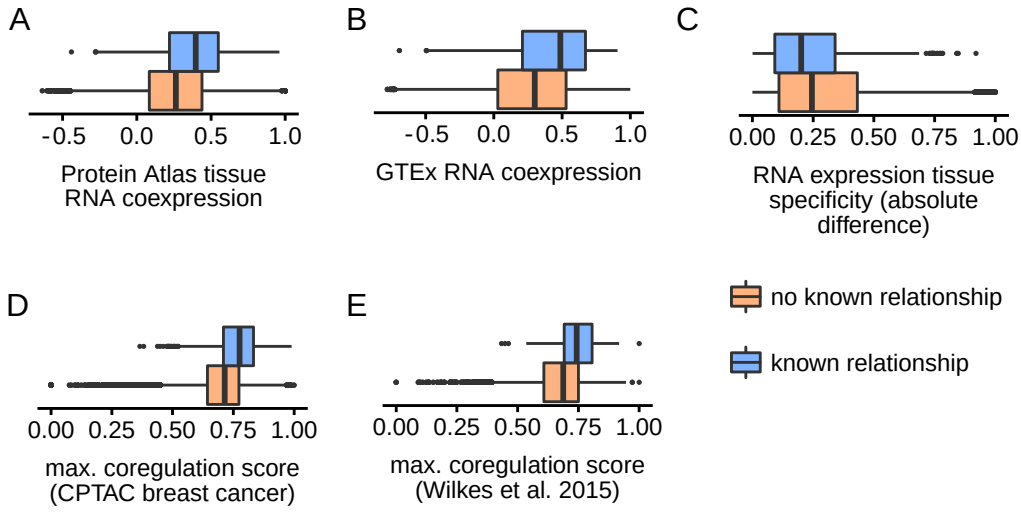

Figure 2.1 *Overview over the associative predictors used in this project. We can see that known relationships tend to have a higher co-expression assigned to them (A & B). While the specificity profile of kinases known to regulate each other are more similar compared to the rest. Similarly, a higher co-regulation scores are assigned to kinases known to have a regulatory relationship (D & E). Figure modified from Invergo & Petursson et al. (Invergo et al., 2020).*

2.3.2 Linking sequence specificity to phosphosite functional impact identifies direct regulation of protein kinase activity

Kinases are known to target specific substrates based on the amino acid sequence surrounding the acceptor phosphosites. By modelling this specificity with a position weight scoring matrix (PWM), we can score a kinase's potential for directly phosphorylating a putative substrate phosphosite. However, high PWM score does not necessarily indicate regulation as the best scoring phosphosite might not be functional. To link PWM scores with phosphosite functionality, we employed the discounted cumulative gain (DCG) which

is often used for information retrieval (Järvelin and Kekäläinen, 2002), where the PWM scores was used as a phosphosite “search function” and the functional score as a “relevance metric” for the substrate phosphosites.

Only 140 kinases had a sufficient number of known substrate sites to build high quality PWMs. It has previously been shown that kinases belonging to the same family tend to have similar specificity profiles, partly due to similar specificity determining residues (SDR) (Bradley and Beltrao, 2018; Bradley et al., 2021). Therefore, we set out to investigate this property to assign PWM to kinases based on family and SDRs. We investigated what level of residue similarity was necessary to make accurate PWM assignments. We found SDR similarity of 0.8 (based on the BLOSUM62 amino-acid substitution matrix) is needed to make assignments that perform significantly better than a randomly assigned PWM (Figure 2.2 A). Nevertheless, this method of assignment did not improve upon assignment by family (Figure 2.2 B). Therefore, we increased the coverage of kinases with PWMs by assigning to assign PWMs in a family wise manner (n = 209) or otherwise assign based SDR similarity (n=14) if a family PWM was not available. As a result, we PWMs assigned to 363 kinases (Figure 2.2 C).

We found that the PWMs of known regulators in OmniPath tend to assign a high PWM score to at least one phosphosite on the substrate kinase (one-sided Wilcoxon rank sum test, $W= 1.0 \times 10^{8}$, $p < 1 \times 10^{-6)}$ (Figure 2.2 D, left panel). Furthermore a highly functional score on the potential substrate kinase seems to indicate regulation and thus has a predictive of a regulatory relationship (one-sided Wilcoxon rank sum test, $W= 6.4 \times 10^{7}$, $p < 1 \times 10^{-6)}$ (Figure 2D, centre panel). By linking these two predictive features together, the DCG also captures known regulatory relationships (one-sided Wilcoxon rank sum test, $W= 4.1 \times 10^{7,}$ $p < 1 \times 10^{-6)}$ (Figure 2.2 D, right panel).

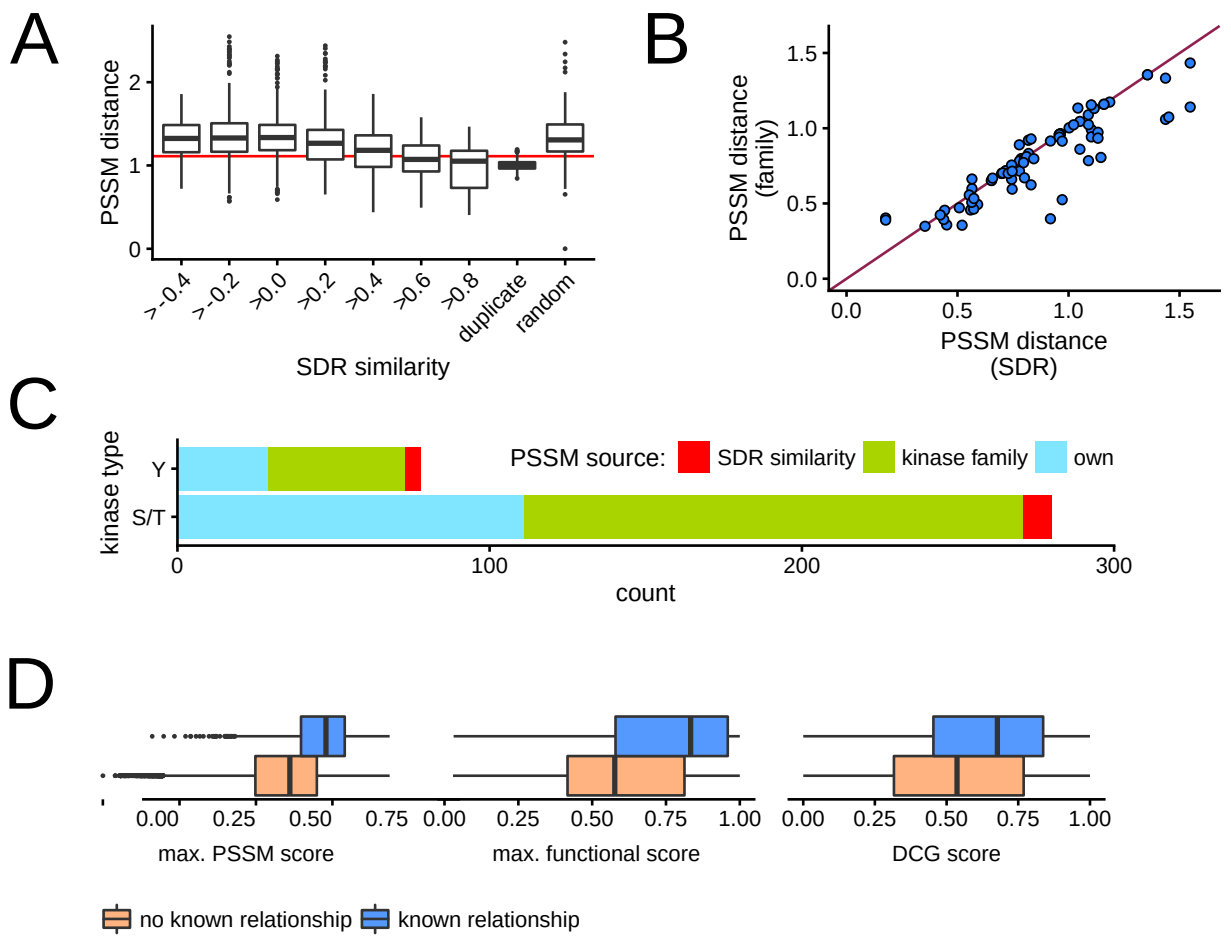

Figure 2.2 *Position weight matrix and functionality score based features*. *Position weight matrices and phosphosite functional scores turned out to be powerful predictors of regulation. To assign PWMs to kinases with few known phosphosites, we assigned family PWM and PWMs based on SDR similarity. For PWMs to be significantly better than random and have a Frobenius distance of less than 1.0 to duplicate PWM, SDR similarity of kinases had to be > 0.8 (A). SDR assignment did not perform better in terms of PWM distance compared with family assignment, therefore family PWMs were assigned before SDR was used (B). 363 PWM were assigned: 140 for kinases with more than 140 known substrates, 209 were assigned based on family and 14 based on SDR. Known relationships had significantly higher PWM score, maximum functional score of substrate and DCG score. Figure modified from Invergo & Petursson et al.* (Invergo et al., 2020).

2.3.4 Choosing machine learning for network prediction

For this project, we tried various different methods for kinase-kinase relationship predictions. The overview of the performance can be seen in **table 2.1** as measured by the area under the ROC curve (AUROC). It should be noted that as BART incorporates missing data into its decisions, while the other methods -as implemented by scikit-learn- do not handle data missingness. As a result, the training sets used to train the models are not equivalent, as data with missing values were removed from the training set. Furthermore, all methods apart from BART used balanced training sets while BART trained with a larger negative set. All methods discerned between relationships found in the literature and the background. Furthermore, due to greater memory efficiency, the methods implemented by scikit-learn were validated with 5-fold cross validation rather than 20 runs of 3-folds as in the case of BART. In the end BART was chosen due to its strong performance as well as the ability to incorporate missing data into its predictions.

Table 2.1: *Performance of different machine learning methods. Methods implemented by scikit learn were evaluated by 100 runs 5-fold cross validation using training set without missing values. BART was evaluated with 20 runs of 3-fold cross validation for each of the 100 models trained. Due to BART's ability to incorporate missing values, a full training set was used.*

Method	**AUC**
Support vector machines	0.75
Logistic regression	0.80
Neural network	0.80
Random Forest	0.81
AdaBoost	0.72
BART	0.88

2.3.5 Description of the resulting probabilistic network

All the predictors were merged and used as an input for BART. While all predictors had a limited but measurable predictive power, the PWMs were the strongest predictor with an AUC of 0.74. However, when combined, these predictors were able to make improved predictions with an average AUC 0.88 across one hundred runs (Figure 2.3 A). The AUC was derived using 20 runs of 3-fold cross validation.

We then investigated whether annotated relationships tended to rank highly among our kinase predictions. The top ranks were significantly better than expected based on a per-kinase random permutations of probabilities (one-sided Wilcoxon rank sum test, regulator: W= W= $5.8 \times 10^{8,}$ $p < 1 \times 10^{-6}$; substrate: W= $7.4 \times 10^{4,}$ $p < 1 \times 10^{-6)}$. Indeed, 50% of kinases had a known regulatory relationship among the top 10 predictions (Figure 2.3 B).

Another way of evaluating our predictor was to see how well it captured low quality regulatory relationships annotated in the literature. That is, relationships found in OmniPath that were mentioned in less than two sources (n =293). We found that these interactions had significantly higher probability assigned to them compared to the background set of edges (Figure 2.3 C). They had, however, edge probabilities lower than the training set (one-sided Wilcoxon rank sum test vs. unannotated: W= $6 \times 10^{7,}$ $p < 1 \times 10^{-6,}$ vs. high-confidence set: W= $8.7 \times 10^{4,}$ $p < 1 \times 10^{-6}$).

We wanted to find out if we could predict substrate kinase and upstream kinases for less studied kinases. (Figure 2.3 D; kinase publication counts were retrieved from (Invergo and Beltrao, 2018)). We found that in our network, kinases in the top three deciles of citation counts (more than 95 publications) accounted for only 31% of the network. Furthermore, 589 regulatory relationships were predicted between kinases in the bottom 50% of publication counts (fewer than 40 publications each).

Overall, there is a large accumulation of novel edges with only around 7% of the high confidence edges being found annotated in the database (**Table 2.2**). However, it could also be because our training set consists largely of well-studied kinases as we can see a significant correlation between citation counts per kinase and top prediction rank. (Figure

2.3 E; Spearman's rank correlation, as regulator: *rho* = -0.34, $p < 1 \times 10^{-6}$; as substrate: *rho* = -0.29, $p < 1 \times 10^{-6}$).

Table 2.2: *Precision and recall at different cut-offs for kinase-kinase network. The number of edges and precision of the network at different cut-offs.*

Cut-off	**No. edges**	**portion annotated in databases**
0.5	4339	0.070
0.6	2113	0.098
0.7	863	0.15
0.8	269	0.25

2.3.6 Signed predictions of kinase-kinase relationships

Prediction of phosphosite sign achieved a Matthews correlation coefficient of 0.42 at cut-off of 0.58 which indicates that phosphosites with higher probability than 0.58 are assumed to be activating while lower values indicate inhibition. Combination of phosphosite sign and interaction features allowed us to predict signs for kinase-kinase regulatory relationships. The maximum Matthews correlation of 0.42 was achieved at a probability cut-off of 0.48

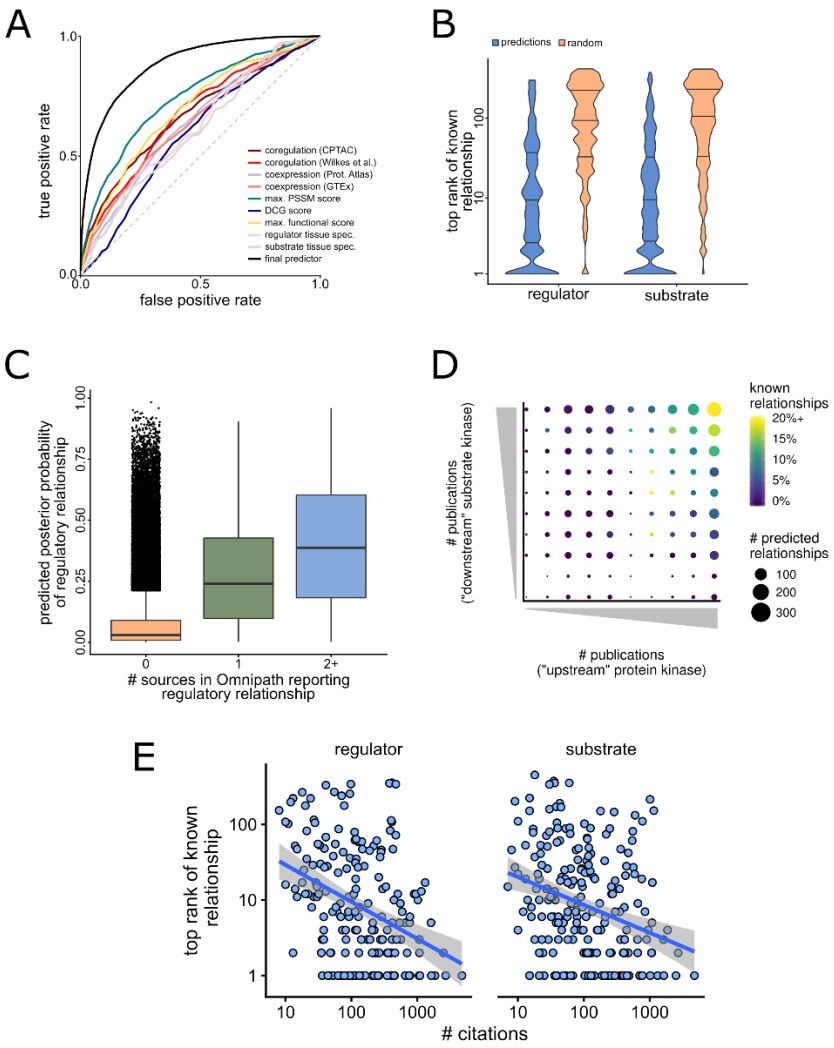

Figure 2.3: *Validation of probabilistic network*. (*A) All predictors were able to discern between known relationships and the rest with the PWMs being the strongest predictor. However, a combination of the different features yielded the best results (AUC =0.88). (B) Looking at the ranking of top predictions we see that our network is significantly more likely to have known relationships among the top ten predictions than random. (C) Our predictor was able to capture relationships only supported by a single source (Which were not included in the training set). However, these were assigned lower probability than the training set. (D) Our predictor is able to make high confidence predictions for less studied kinases and substrates. (E) Nevertheless, highly cited proteins still occupy the top predictions; both for regulators and substrates. Figure modified from Invergo & Petursson et al* (Invergo et al., 2020).

2.3.7 Reconstruction of pathways from probabilistic kinase-kinase network

We wanted to see if pathways could be extracted from the network by looking at a few well-studied kinases. For each pathway, interactions involving any of the kinases in the pathway were removed from the positive set. 0.5 was used as a probability cut-off both for regulatory interaction and sign. Starting with AKT1 and kinases closely functionally related to AKT1 we see that most interactions were recaptured (Figure 2.4 A). The inclusion of less studied kinases and paralogs of these kinases shows, however, that our performance drops suddenly (Figure 2.4 B). Similarly, if we expand the scope and look at the MAPK pathway, we notice an accumulation of unknown edges (Figure 2.4 C).

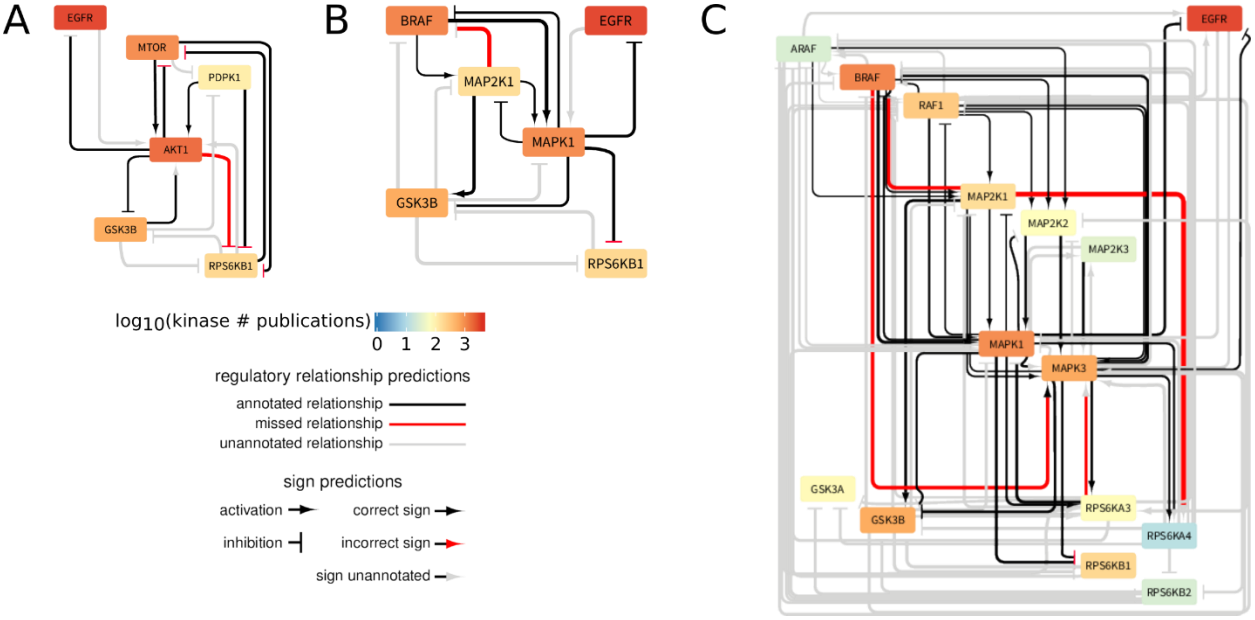

Figure 2.4*: Prediction of signalling pathways. Retrieval of pathways from the network. (A) By only including well studied kinases, we can capture and predict correctly most edges with a relatively low number of unannotated edges. (B) By including less studied kinases and paralogs of kinases used in (A) we get more mistakes and accumulation of unannotated edges. (C) By increasing the scope and including the MAPK pathway we see that while erroneous predictions remain relatively rare, accumulation of unannotated edges increases. Figure taken and modified from Invergo & Petursson et al.* (Invergo et al., 2020).

2.3.8 Identification of functional modules within kinase-kinase network

Biological networks are often thought of as being highly modular and scale-free with relatively few nodes as hubs and most proteins being peripheral in the network interacting with few other proteins. To test if these properties applied to the kinase-kinase network formulated here, I selected high confidence edges (edge probability of > 0.5), 4,339 edges between 317 kinases in all. I applied three different clustering methods on the high confidence network, MCR, greedy clustering and Louvain clustering. The methods identified 7, 5 and 4 clusters that included ten or more genes respectively. MCL did find 3 clusters with biological functional enrichment (48 pathways in all), while greedy clustering was more similar to Louvain, who found 60 enrichments in 3 clusters. Louvain clustering was chosen as giving the most biologically relevant partition with 4 clusters and 72 enrichments in 4 clusters. Louvain returned 4 different clusters each including 193, 45, 107 and 87 kinases respectively. This partition yielded a modularity score of $\mu = 0.325$ which is significantly higher than an empirically derived distribution of modularity scores derived from hundred different randomization of the original unperturbed network where node degree distribution was preserved ($\mu = 0.197$, $\sigma = 0.00339$, $p < 0.001$; Figure 2.5 A). Biological modules are often understood as a functional subunit of the signalling network where each module has a biological function that does not overlap with other modules (Bhattacharyya et al., 2006; Kirschner and Gerhart, 1998). To see if that was the case, I conducted an enrichment analysis on the modules extracted above using the 504 kinases as the background. In order to enrich the modules, the R package ReactomePA (Yu and He, 2016) was used. Each of the modules had a significantly enriched pathway with each of the modules having non overlapping assigned function the top enrichments can be seen below (Figure 2.5 B). Additionally, we found that pathways that clustered together were closer to each other in the literature network indicating that the kinase-kinase network can be divided into modules that correspond to modules within the literature network (Wilcoxon rank sum test, $W = 5.8 \times 10^5$, $p < 1 \times 10^{-6,}$ Figure 2.5 C)

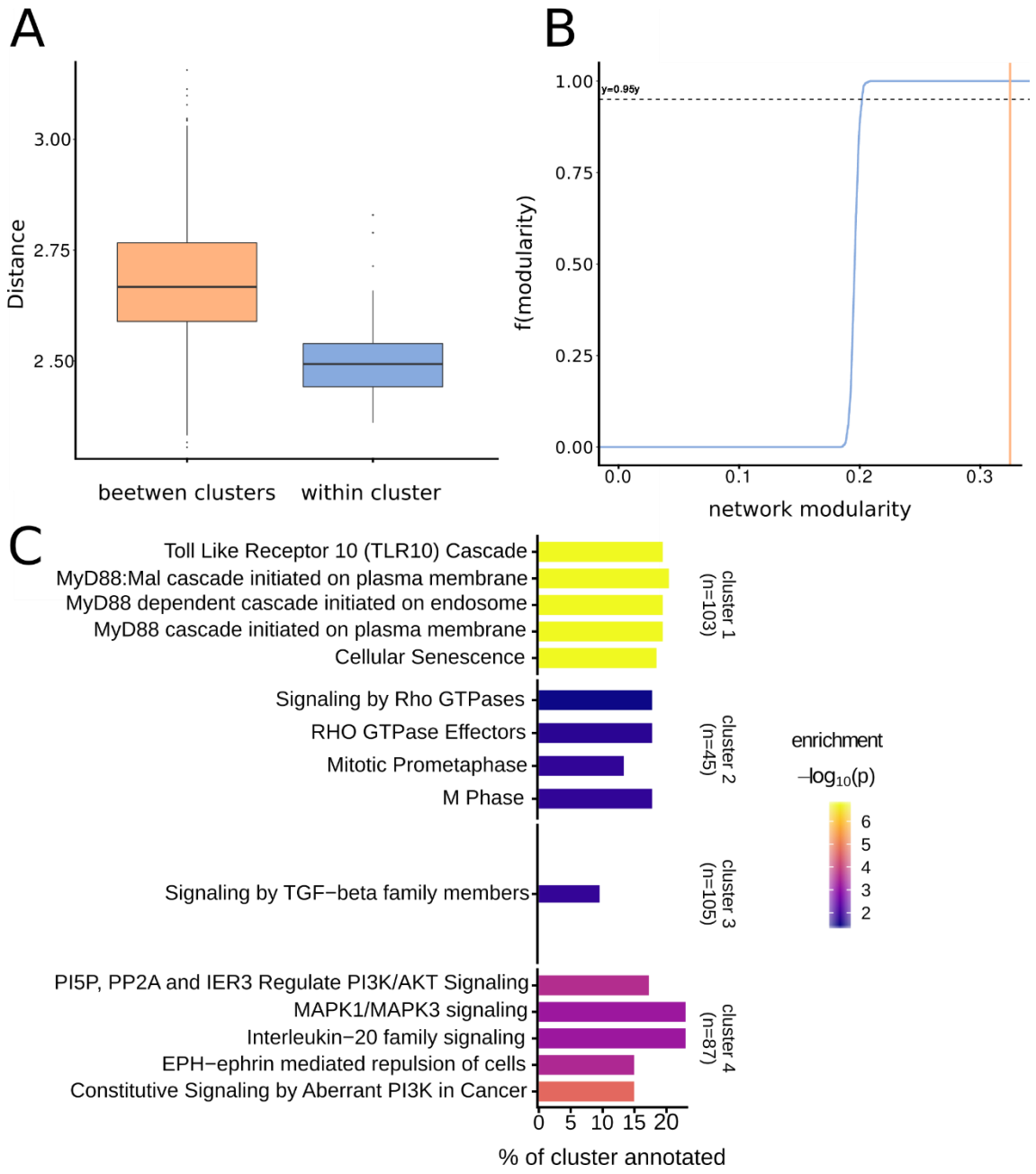

Figure 2.5: *Establishing modularity of the probabilistic network*. (A) *Pathways that clustered together were closer together in the IntAct network than pathways that were not enriched in the same cluster. (B) We found that our network was more modular than an empirical distribution of modularity values made by randomizing our network while maintaining edge degree (C). Each cluster has an assigned function in the form of an enriched pathway. Pathway enrichment did not overlap across clusters. Figure modified from Invergo & Petursson et a*l. (Invergo et al., 2020).

2.3.9 Validation of kinase-kinase relationships with independent experimental kinase-substrate predictions

Since large portions of the signalling system are understudied, any attempt to make less unbiased predictions is bound to run into the problem of accumulation of multiple high confidence predictions without any way of verification. While a significant portion of these unknown edges can be reasonably assumed to be false positives, a number of true novel hits is also likely. In order to evaluate the formulated kinase-kinase interaction network's ability to capture unknown interactions two experimental kinase-substrate predictions were used for evaluation. One study was an *in-vitro* study predicting substrates for 354 kinases by introducing kinases to dephosphorylated peptides (Sugiyama et al., 2019). The other study based its prediction on changes in phosphorylation following kinase inhibition (Hijazi et al., 2020). Both studies have their limitations, but together add confidence to the computationally derived predictions due to the different methodology. Another limitation of this data set for analysis is that these studies predict phosphorylation of phosphosites by an upstream kinase but not kinase regulation.

Probability scores of kinase-kinase interactions that were predicted by either method were compared with predictions that were not. It was found that the experimentally supported edges did have higher probability assigned to them than the ones that were unsupported by either study; (*in-vitro*: median probability of 0.075 ($W = 2.6 \times 10^8$, $p < 1 \times 10^{-6}$)*;* cell line: median probability of 0.12 ($W = 1.6 \times 10^8$, $p < 1 \times 10^{-6}$); background: median probability of 0.030, Figure 2.6 A). When only looking at kinase-substrate relationships, where the functional score of the target phosphosite was high (> 0.5) and thus indicating regulatory relationship, we found that both *in-vitro* sets ($W = 1.0 \text{ X } 10^8$, $p < 1 \text{ X } 10^{-6}$) and cell line set ($W = 8.8 \text{ X } 10^7$, $p < 1 \text{ X } 10^{-6}$) had a significantly higher probability assigned to them compared with the unsupported edges (Figure 2.6A). Experimentally predicted edges were also more likely to be included in our validation set (OR = 3.98 and $p = 1.6 \text{ X } 10^{-4}$ for the cell line study and OR= 4.51 and $p < 1 \text{ X } 10^{-6}$ for the *in vitro* study). Overall, these results indicate that our computationally derived kinase-kinase network has captured real hitherto unknown interactions.

2.3.10 Identification of new regulatory circuits with independent experimental data

Having established that the kinase-kinase regulatory network can discriminate between experimentally supported edges and the background, the ability of the network to discover new pathways was tested. To this end, we generated a phosphoproteomic data set derived from PI3K and MEK inhibited Kasumi-1 cell lines. The data set included 9,183 phosphopeptides quantified upon PI3Ki or MEKi inhibition and a control condition. I found that 112 (PI3K) and 66 (MEKi) phosphosites were downregulated in each condition respectively. By adding known kinase phosphosite relationships from PhosphoSitePlus I identified the weighted shortest path from the inhibited kinases to the down regulated phosphosites. For the final analysis 6 (MEKi) and 11 (PI3K) phosphates were included as the rest either did not have any known upstream kinase or were direct substrate of MEKi and PI3Ki. The distances between the inhibited kinases and the down-regulated phosphosites were calculated as the sum of edge weights across the shortest weighted path. I found that downregulated phosphosites were closer to the inhibited kinases when compared to all other phosphosites included in the network ($W=1.4 \times 10^4$, $p= 0.0028$) (Figure 2.6 B). In order to see if any new pathways could be identified, I selected the shortest paths linking the inhibited kinases and the down-regulated phosphosites and any unknown interaction that was corroborated by either prediction experiment were considered to be a possible new interaction. In this way, possible new interactions were found between CDK1, CDK2 and SRC and between PRPF4B by MAPK1 which was supported by both experiments (Figure 2.6 C).

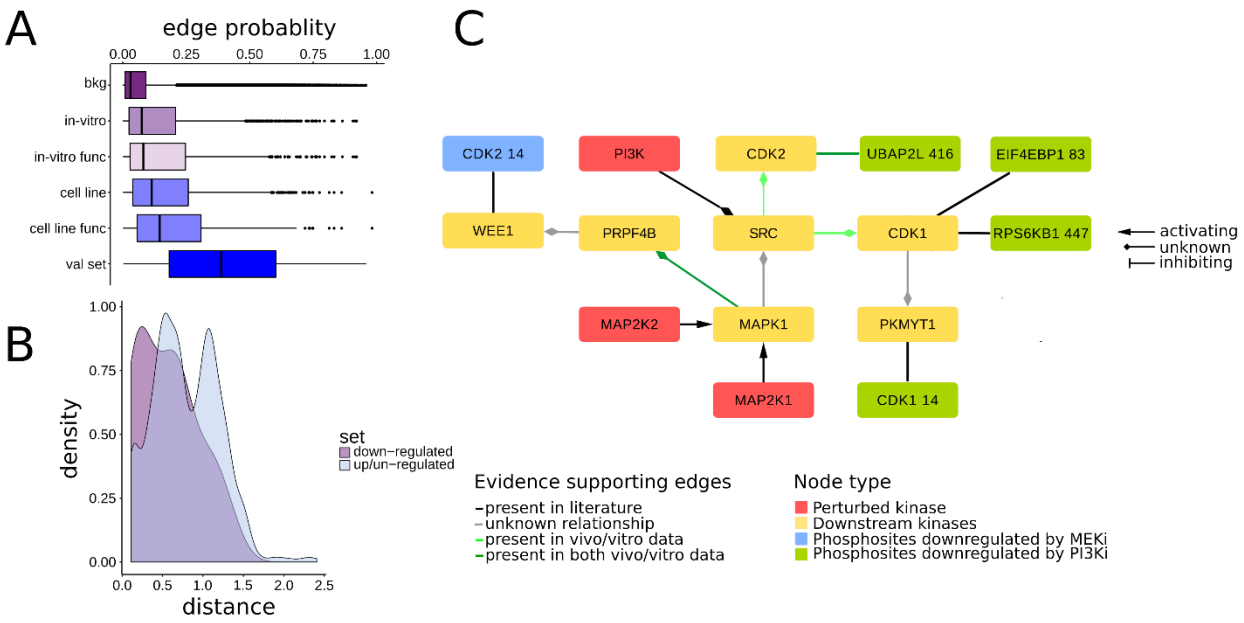

Figure 2.6 *Capturing novel pathways with phosphoproteomics data***.** *(A) Experimentally supported edges had greater probability assigned to them than the background. Experimentally validated edges had, however, lower edge probability than the training set. (B) Phosphosites that were significantly less phosphorylated upon kinase inhibition were closer to the inhibited kinase in our network compared to phosphosites that were less impacted. (C) By linking impact phosphosites to inhibited kinases by the shortest path we found three putative new pathways. The links between SRC and CDK1 and SRC and CDK2 as well as MAPK1 and PRPF4B*, *MAPK1 and PRPF4B are of particular interest due to support by both experiments*

2.4 Discussion

Due to the sheer scale of the possible human kinase regulation network, experimental validation of all potential kinase-kinase regulatory interactions is unfeasible. Here, we have proposed a data-driven machine learning approach to assign probability scores to these regulatory relationships to guide future exploration of the understudied signalling network space. These predictions cannot replace existing experimental methods for relationship confirmation but can be used to reduce the vast space of possible relationships worthy of consideration for the formation of credible hypotheses and to prioritize experiments, for less studied kinases in particular.

Previous efforts to produce kinome-wide predictions of regulatory relationships have depended on existing protein networks to aid data-driven predictions. Rudolph et al. (Rudolph et al., 2016) inferred signalling pathways by employing a network diffusion technique with phosphoproteomic data mapped onto a literature-derived protein-protein interaction network. However, these efforts are heavily impacted by the study bias discussed in the introduction of this thesis which overestimates the importance of well-studied proteins in signal propagation.(Gillis et al., 2014; Invergo et al., 2020; Luck et al., 2020; Rolland et al., 2014). To our knowledge, there has only been one other attempt to predict signed kinase regulatory relationships (Hernandez et al., 2010). This study, based their sign predictions on mapping phosphoproteomics data onto literature network, with its biases onto quantitative phosphoproteomic data. Missing interactions in particular for understudied kinases can reasonably be expected to have a large impact on the results. Our supervised machine learning approach reduces the impact of this bias as predictions can be made for less studied kinases improving the coverage in our predicted network as a result. We do retain, however, a level of bias by using known kinase-substrates to construct the kinase specificity models as well as using literature networks to form the training set. High-throughput methods that measure kinase specificity profiles (see, e.g. (Imamura et al., 2014; Sugiyama et al., 2019) could be used to remedy the former issue. The latter, which may miss highly context specific regulation, might improve as more relationships are experimentally validated.

Many different factors have an impact on the nature of kinase-kinase regulatory relationships and each condition is unique in this way, due the different properties of the kinases involved. Therefore, system-wide generalized prediction for all kinases is an inherently difficult problem. Some features such as phospho-regulation are fundamental and shared by all kinases. However, in order to achieve high quality predictions based on regulation by phosphorylation, a large set of large scale phosphoproteomic experiments across different tissues, cell types and conditions need to be conducted due to the context-specific nature of phospho-signalling. In this study we relied on phosphoproteomic data from breast cancer samples and MCF7 cell lines. Due to the homogeneity of our data sources this might introduce bias into our predictions as the decisions based on the co-regulation score run the risk of being breast cancer specific.

This problem is further compounded by the fact that cancer initiation and progression often leads to dysregulation in the signalling network so regulatory relationships inferred from cancer data might not be representative of healthy tissue. Furthermore, due to the limited number of data sets, only a limited number of kinases were represented in the data. It is therefore clear that in order to make kinase-kinase regulatory predictions, a large number of large-scale phosphoproteomic experiments need to be conducted.

The importance of PWM scores in our predictions further stresses the need for high quality PWM to be constructed for every kinase to add confidence to the predictions made by high throughput data. PWMs will play an important role in kinase-substrate predictions as they help to prune out indirect effects captured by correlative predictors. One weakness of PWMs is that it is constructed from known kinase-substrate relationships and might therefore make overly conservative predictions in some cases.

The predictions made here rested upon the presumption that the signalling network is sparse, that is most kinases regulate a relatively low number of other kinases. Indeed, 75% of our predictions were assigned a probability score of less than 0.09 which is far less than the probability cut-off used in this exercise. However even at a relatively stringent cut-off 0.5 and higher there was a large accumulation of unknown interactions. For instance, at a cut-off of 0.5, only around 7% of the interactions were included in our positive training set. This is partly to be expected as the number of predictions made increases. This leads to a network that is denser and richer in cross-talks across modules and regulatory feedback loops than is typically considered to be the case in the literature. This might be due to the fact that cellular context, such as protein localization, is not considered in this study. In order to validate the new high probability relationships further development in experimental approaches for hypotheses free identification of regulatory relationships are needed.

3 Machine learning-based prediction of kinase-substrate relationships

3.1 Introduction

The sheer size of the phosphoproteome emphasizes the need to explore the set of possible kinase-substrate relationships in a systematic manner. Most kinase-substrate relationships found in data bases such as PhosphoSitePlus (Hornbeck et al., 2015), SIGNOR (Licata et al., 2020) or Phospho.ELM (Diella et al., 2008) are found between well-studied kinases and well-studied substrates (Invergo and Beltrao, 2018). However, high throughput protein interaction mapping analyses (Luck et al., 2020; Rolland et al., 2014) have found that the human interactome is more evenly distributed than databases indicate. This points to a literature bias in current research since the focus lies on well-studied proteins.

So far, many methods have been developed to predict kinase-substrate relationships using predictive features such as kinase specificity models and protein localization. Peer reviewed methods include GPS v5.0 (Wang et al., 2020), KinomeXplorer (Horn et al., 2014), NetPhos v.3.1 (Blom et al., 2004), LinkPhinder (Nováček et al., 2020) and PhosphoPICK (Patrick et al., 2015). One of the shortcomings of these methods is either the limited use of features apart from known kinase-substrates or the use of literature derived networks such as STRING (Szklarczyk et al., 2021) or BioGRID (Oughtred et al., 2019) which inherently increases the study bias in their predictions. Recently, a method titled CoPhosK (Ayati et al., 2019) addressed some of these issues by using phosphoproteomic data and integrating it with another well-known method: KinomeXplorer (Horn et al., 2014). Here I describe a method to make kinase-substrate predictions based on kinase specificity models, high throughput biological data and various features characterizing the potential acceptor phosphosite. I validated the network with external independent experimentally validated kinase-substrate relationships and compare it with established methods. Furthermore, I show that these features can be used to predict the sign of kinase-substrate relationships which, to my knowledge, has

not been done before, other than in our prediction of kinase-kinase regulatory networks discussed in Chapter 2 which were not phosphosite specific.

These predictions were then incorporated into SELPHI2, an expansion upon SELPHI (Petsalaki et al., 2015) which previously based its predictions on correlation analysis to extract associations between kinases and phosphatases and phosphosites. SELPHI2 is a platform that provides biologists with various means to analyse their phosphoproteomic data including enrichment analysis, kinase-substrate relationship predictions and the identification of probabilistic sub networks specific to the conditions under study.

3.2 Methods

3.2.1 Data sets

I downloaded kinase inhibition perturbation phosphoproteomics data from NTERA2, MCF7 and HL60 (23.02.2020) from a previous publication (Hijazi et al., 2020). Kinase-substrate relationship predictions from two earlier experiments were retrieved from two previous publications (Hijazi et al., 2020; Sugiyama et al., 2019). I retrieved a compilation of 435 different perturbation conditions and phosphoproteomics data from breast cancer samples (Mertins et al., 2016) from previous publication (Ochoa et al., 2016). I downloaded tissue RNA expression data (Expression Atlas (Papatheodorou et al., 2020) , 26 April 2018) from the GTEx project (GTEx Consortium, 2013) and tissue and cell line RNA expression data was downloaded from the Human Protein Atlas project (Thul et al., 2017; Uhlén et al., 2015; accessed from www.proteinatlas.org 1 December 2017). Known kinase-substrate interactions were downloaded (04.11.2020) from PhosphoSitePlus (Hornbeck et al., 2015) and signed kinase-substrate relationships were downloaded (13.03.20) from SIGNOR (Licata et al., 2020).

Functional score and a collection of features characterizing phosphosites were retrieved from a previous publication (Ochoa et al., 2020).

3.2.2 Training set for kinase-substrate prediction

As a positive set to predict kinase-substrate relationships, I used kinase-substrate relationships listed in PhosphositePlus (Hornbeck et al., 2015). Since there is no database of true negatives, where it is established that kinase A does not phosphorylate phosphosite B, a set of randomly drawn kinase-substrate relationships ten times as large as the positive set was used. The rationale behind the larger size of the negative set is that biological networks are typically sparse so the training set should reflect this imbalance. The size of the positive set contains more than 5,500 interactions.

In order to make predictions on kinase-substrate interaction signs, I used SIGNOR (Licata et al., 2020) which contains information on the sign of kinase-substrate relationships. Furthermore, for the signed predictions I only retained functional (functional score > 0.5) phosphosites (Ochoa et al., 2020). Here I keep the balance between the two groups relatively even; 673 activating interactions and 497 inhibiting interactions. Therefore, a single model was trained to predict the sign of the kinase-substrate relationships.

To assess the models, ten-fold cross validation as implemented in scikit-learn (Pedregosa et al., 2011) was used to assign probabilities to the kinase-substrate relationships included in the training set by using the *cross_val_predict()* function. To quantify the AUROC and draw ROC curves, the ROCR R package was used (Sing et al., 2005).

3.2.3 Formulation of predictors

In order to train a machine learning model a set of informative predictors were generated. Here I generated a set of predictors similar to those described in Chapter 2. Here, however, predictions are being made on the phosphosite level rather than the protein level. In cases such as co-expression, where we simply have information on protein level the co-expression between the kinase and the protein on which the candidate substrate phosphosite is calculated. The following predictors were used:

Co-expression: Kinases and their substrates can be reasonably assumed to be co-expressed across tissues and cells. I generated the same predictors (co-expression and selectivity) for data derived from tissues and cell lines. Since expression is not measured on the phosphosite level, co-expression between the kinase and its potential substrate

protein was measured. Three data sets were used for this purpose: GTEx (GTEx Consortium, 2013) and from the human protein atlas we used expression data from human cell lines (Thul et al., 2017) and human tissues (Uhlén et al., 2015).

Correlation between kinase activity and putative substrate phosphorylation: Similarly, we can assume that kinase activity correlates with the phosphorylation of its target phosphosite. Earlier we used co-regulation score for kinase-kinase regulatory relationship prediction (Chapter 2.2.3.7). I set out to develop another feature based on co-phosphorylation that was less dependent on the functional score since in this case, I am predicting kinase-substrates rather than kinase-kinase regulation. To this end we used three recent data sets from NTERA2, MCF7 and HL60 cells generated by Hijazi and colleagues (Hijazi et al., 2020) where these cells were introduced to 61 different kinase inhibitors targeting 103 kinases for 1 hour. The activity of the kinases was estimated with the KSEA method (Ochoa et al., 2016) which bases its estimates on the Kolmogorov Smirnov statistical test. To estimate the activities the *ksea_batchKinases()* function was used with 1000 trials to generate empirical p-values. The same method was used to calculate associations between kinases and their putative substrates within the CPTAC breast cancer data set (Mertins et al., 2016). These kinase activities were then correlated with the phosphorylation levels of the site in each condition using the Spearman's rank correlation. The $-\log_2$ ratio of the p-values were used to quantify association.

Functional score: Previously the functionality of 115,000 different phosphosites had been estimated by (Ochoa et al., 2020) by compiling 59 features such as evolutionary age and disorder to predict functional score with machine learning methods. In this case, functionality is defined here as the probability of a given phosphosite is to lead to changes in the functional state of the protein containing the phosphorylated residue. Changes in functionality encompass events such as changes in activity or localisation. The rationale behind this predictor is that kinases are more likely to target highly functional sites in a regulated and predictable way since these would form part of a coordinated cell response. In any case, we are interested in regulatory relationships between kinases and substrates and thus enriching for functional kinase target sites rather than spurious ones is more likely to give us the desired outcome. Furthermore, predictors used to score the

functionality were downloaded from an earlier publication and used as for prediction (Ochoa et al., 2020)

Position weight matrices (PWM) scores: Kinases are known to target specific sequence patterns, or motifs, surrounding their target phosphosites. One way of modelling kinase specificities is by generating position weight matrices where the similarity between the candidate motif and the kinase's consensus motifs are assessed. Here I used the position weight matrices we created for the work presented in Chapter 2.2.3. (Invergo et al., 2020).

Imputation was done on a per-predictor basis in the same manner as done in the original paper. In total 35 predictors, including evolutionary age, disorder as measured by DISOPRED (Jones and Cozzetto, 2015) score and the age of the phosphosite were added as predictors. A complete list of these predictors and their description is provided in **Appendix 3.1**.

One issue that commonly arises when different predictors are added together in this manner is the introduction of missing values in the final matrix. Many different methods have been proposed to impute missing data for machine learning. For the predictors downloaded from a paper published by Ochoa and colleagues (Ochoa et al., 2020) describing the functional score, I used the same imputation methods as used in the paper and are listed in **Appendix 3.1**. For the rest of the predictors that had missing values, i.e. the predictors based on co-expression and co-phosphorylation, I considered three methods for imputation.

Zero imputation: This imputation value replaces all missing values with zero. Generally, this method can introduce bias in the prediction. In this particular case, missing values for association measure were being estimated. In practice, by imputing by zero the assumption of non-association is being made for each kinase-phosphosite pair.

Median imputation: In contrast median imputation replaces the missing values within each predictor with the median value. Similar to zero imputation this method does not factor in correlation between features.

IterativeImputer (Buck, 1960; van Buuren and Groothuis-Oudshoorn, 2011)**:** This is the most sophisticated of the three methods. Iterativeimputer method uses other features in the feature set to impute the missing feature. That is each feature with a missing value is

modelled as a function of the other features. For this purpose *IterativeImputer()* implemented by the python suite scikit-learn was used for the imputation. Bayesian ridge was used to estimate the missing values. The number of iterations run to estimate the missing feature was 10.

Each method was evaluated by their predictive power as measured by mean AUC derived from ten-fold cross validation run on hundred different training sets. The method giving the highest AUC value was selected for the finalized pipeline. Zero imputation resulted in the best model and was therefore used for model training.

3.2.4 Selection of predictive features

To select the optimal number of predictors for classification, I used recursive feature elimination (RFE) with cross validation (Guyon et al., 2002) as implemented in scikit-learn (Pedregosa et al., 2011). RFE uses a learning method, random forest in this case, to assess the relative importance of each feature and the least important feature is then removed. The method then recursively considers a smaller and smaller set of features to train the model. The features used by the model yielding the best AUC were then selected for subsequent prediction. I tuned the following parameters: *Bootstrap,* which decides if the number of samples drawn to train the base estimator, *max_depth,* which sets the maximum depth of the tree, *min_samples_split,* which sets the minimum number of samples needed to constitute a leaf node and *n_estimators,* which set the number of trees included in the forest. The following parameters were considered: *Bootstrap*: True, *max_depth*: [10, 20, 50, 70, 80, 100], *min_samples_split*: [8, 10, 12], *n_estimators*: [150, 300, 400, 500, 1000, 1500]. This procedure is then recursively repeated until the best feature set as measured by cross validation is found.

Due to the fact that the negative set is randomly sampled from the set of kinase-substrate pairs not present in the positive set, optimal features were selected from one hundred training sets and features that were selected in more than half of the runs were kept. For the kinase-substrate prediction 37 of the 49 features were kept and for the signed predictions 45 predictors were kept for model construction. The whole set of features considered, the features used for kinase-substrate prediction and the ones used for sign prediction are listed in **Appendix 3.1.**

3.2.5 Training of predictive model

I used the predictors listed above to train a random forest model as implemented in the scikit-learn package (Pedregosa et al., 2011). As mentioned above, I trained 100 different models for the kinase-substrate relationships with the full feature set. Random forest was used to make predictions based on the predictors listed above. To optimize the model selected I used grid search with cross validation (10-folds) to select models. With the parameters same parameters as considered for the feature elimination or: *Bootstrap*: True, *max_depth*: [10, 20, 50, 70, 80, 100], *min_samples_split*:[8, 10, 12], *n_estimators*: [150, 300, 400 ,500, 1000, 1500].

3.2.6 Comparison with other kinase-substrate prediction methods

Since I have created a web server that provides my kinase-substrate prediction method as a service to users. I will henceforth refer to the kinase-substrate predictions as SELPHI2. In order to assess SELPHI2's predictive power relative to other established methods, we compared SELPHI2's ability to make accurate predictions to five different methods available in the literature:

PhosphoPICK (Patrick et al., 2015)**:** I submitted the sequences of all substrate proteins included in SELPHI2's predictions onto the server (http://bioinf.scmb.uq.edu.au/PhosphoPICK/submit). A p-value threshold for predictions was not used to include all possible predictions. PhosphoPICK makes predictions for 107 human kinases.

GPS v.5.0 (Wang et al., 2020) **:** In this study I used GPS version 5.0. This method makes predictions for 457 kinases. I submitted the sequences of all substrates to a desktop version of GPS v 5.0. I set the threshold (all, medium or stringent) to all to include all possible predictions. GPS was downloaded from (http://gps.biocuckoo.cn/download.php) and batch kinase prediction was run on Ubuntu 18.04

KinomeXplorer (Horn et al., 2014)**:** NetworKIN v 3.0 was downloaded from (http://www.networkin.info/download.shtml) and all substrate sequences were submitted

as a fasta file with an additional file indicating location of relevant phosphosites. KinomeXplorer includes predictions for 193 kinases.

Netphos v.3.1 (Blom et al., 2004)**:** Netphos only allows for the substrate prediction of 17 kinases. All sequences were submitted to the NetPhos v3.1 server (http://www.cbs.dtu.dk/services/NetPhos/). Predictions were made for all residues (Serine, Threonine and Tyrosine). To include all possible predictions, no score threshold was applied.

LinkPhinder (Nováček et al., 2020)**:** Altogether LinkPhinder contains 11,581,940 predictions, which were downloaded from https://linkphinder.insight-centre.org/download. These methods are described in greater detail in Chapter 1.10.3 in the Introduction.

In order to compare SELPHI2 to other peer reviewed methods I compared SELPHI2's ability to detect known annotated kinase-substrate relationships as listed in PhosphoSitePlus (Hornbeck et al., 2015) as well as novel kinase-substrate relationships supported by previous high throughput kinase-substrate prediction experiment (Sugiyama et al., 2019) which is described in greater detail in Chapter 2.2.10.

The comparison between two kinase-substrate prediction methods is a nontrivial problem. Methods differ in their use of training sets and the number of kinases included in the analysis. To address this difference, I made five different comparisons between SELPHI2 and the other methods, only assessing the overlap between SELPHI2 and the comparison method. I quantified the ability of these methods to discern between known or predicted kinase-substrates and the background by calculating the area under the ROC curve as implemented by ROCR (Sing et al., 2005). The same approach was used also for comparing their ability to recover an independent dataset of experimentally determined sites (Sugiyama et al., 2019). The data set is discussed in greater detail in chapter 2.2.10

To compare SELPHI2's ability of predicting known kinase-substrates to the other methods, I generated 100 training sets for each method with known interactions from PhosphoSitePlus (Hornbeck et al., 2015) and included ten times as many random sets of kinase-substrate pairs that were shared by both methods. These sets were used to train and assess SELPHI2's models. Each training set was split ten times into a test set and training set and for each split the training set was used to assign probabilities to the test set.

For each of the hundred training sets, SELPHI2's performance at capturing known kinase-substrates was compared with the other peer reviewed methods.
In order to assess the ability of the methods to capture potentially new kinase-substrate relationships not found in the literature, I compared SELPHI2's ability to discriminate between experimentally predicted interactions (Sugiyama et al., 2019) and the background. To remove any bias, kinase-substrate phosphorylation relationships found in PhosphoSitePlus were excluded from the positive set.

3.2.7 Mapping of kinase-substrate prediction onto phosphoproteomics data

I hypothesized that a way to select for true edges is to prune the kinase-substrate interactions by fitting on independent data. The kinase-substrate predictions were fitted to high throughput phosphoproteomic mass spectrometry data that had been compiled and analysed by Ochoa and colleagues (Ochoa et al., 2016). This compilation contained phosphoproteomic data generated under 436 conditions.
In order to link our kinase-substrate predictions together into a comprehensive network I linked the kinase-substrates together with the kinase-kinase regulatory network (Invergo et al., 2020) created in Chapter 2. In both cases the probability cut-off of 0.5 was used to extract high confidence edges. The Prize collecting Steiner's forest as implemented in the PCFS package (Akhmedov et al., 2017) was used to generate sub-networks to identify kinase-substrate relationships that are likely to be active in a given context. PCSF seeks to optimize profit which is calculated as the sum of node prizes after subtracting the edge costs. Here, I set node prices as the absolute value of the $\log_2$ ratios of the phosphosites included in the data and the cost was set as the edge probability subtracted from one.
PCSF has three tuneable parameters: w which sets the number of trees, b the parameter that tunes the node values and μ the parameters that tunes the edge cost. Different parameter combinations can give different results and therefore all combinations of the following parameters settings were tried: For w anything between one and ten trees were tested. For the node tuning, b, the values: 0.25, 0.5, 0.75, 1.0 1.25 and 1.5 were tested and for μ: 0.000005, 0.00005, 0.0005, 0.005, 0.05. This meant that for each condition,

175 subnetworks were generated. For each condition, we selected the subnetwork that retained the set of kinase-substrate relationships with the highest F1 score.

I also considered a method based on the shortest path. For this analysis, only conditions where the biological sample had been treated with a kinase inhibitor were considered in order to be able to link downregulated phosphosites to a perturbed kinase. This included 42 conditions in all. For each of these inhibitory conditions, phosphosites with higher $\log_2$ ratio than 1 or lower than -1, were identified among the phosphosites included in the data set and the shortest path from the inhibited kinase to these regulated phosphosites were included in a pruned sub network. To calculate the shortest path the *all_shortest_paths()* function as implemented in the R package *igraph* (Csárdi and Nepusz, 2006) was used to identify all shortest paths between the inhibited kinase and the phosphosites.

The third approach that I tried included applying various heat diffusion methods through the combined network. The $\log_2$ ratios of the phosphosites were applied as heat to diffuse throughout the network. I used the various heat diffusion algorithms that were implemented in the R package *diffuStat* (Picart-Armada et al., 2018, 2021):

Raw (Vandin et al., 2011; Zoidi et al., 2015): Takes positive values and sets them as one while non-labeled nodes and nodes with negative values are set as zero. Hereby this vector will be referred to as y. The algorithm then proceeds to smoothen these values with the following formula:

$$f_{raw} = K \cdot y_{raw}$$

Where K is a kernel. In this case I used the default a regularised Laplacian kernel.

Ml (Zoidi et al., 2015)**:** works in the same manner as raw while introducing negative values as -1. Unlabelled nodes are set to zero.

Gm (Mostafavi et al., 2008)**:** Functions in the same way as ml except for the fact that unlabelled nodes that are assigned a bias value that is calculated from the relative size of the three sets of nodes: positive, negative and unlabelled.

Mc (Bersanelli et al., 2016)**:** Mc calculates the score of the node based on an empirical p-value where the node values are permuted n times. The final p-value is in proportion to

how often the diffusion value derived from the initial node values was higher than the permuted values. The p-value is calculated as follows:

$$p_i = \frac{r_i+1}{n.perm+1} \quad (2)$$

Where n.perm is the number of permutation and r_i is the original raw diffusion value of node i. mc is then calculated as follows:

$$f_{mc,i} = 1 - p_i \quad (3)$$

Z (Harchaoui et al., 2013)**:** A parametric alternative to mc and thus provides a faster alternative. The raw value of node i is subtracted by the mean value and divided by the standard deviation.

Ber_s (Bersanelli et al., 2016)**:** Here the nodes are given values based on the change in values between before and after smoothing.

$$f_{ber_s,i} = \frac{f_{raw,i}}{y_{raw,i}+\varepsilon} \quad (4)$$

Where ε is a parameter controlling for the importance of relative change and $f_{raw,i}$ and $y_{raw,i}$ are the same variables as defined in equation 1.

Ber_p (Bersanelli et al., 2016)**:** This scoring combines mc and the raw method.

$$f_{ber_p,i} = -\log_{10}(p_i) \cdot f_{raw,i} \quad (5)$$

Where p_i is the same as in equation (2) and $f_{raw,i}$ is the same as in equation (1)

These diffusions returned a list of node values. In order to select nodes to include in the sub-network, I generated an empirical distribution of node values by randomizing the network with the *vertexsort()*(Jothi et al., 2009) function as implemented in the VertexSort R package which randomizes networks while maintaining the edge distribution. The randomization step was repeated hundred times and nodes with heat values that were higher than 95% of their respective empirical distribution of heat values.

I evaluated the performance of the fitting by measuring the F1 scores of the part of the network included in the fitting. In the case of the PCSF it included phosphosites that were present in the high throughput data and their predicted upstream kinases were compared with the F1 scores of the fitted network. In the case of the shortest path, only phosphosites that were regulated were considered. Since the heat diffusion used an undirected network

as an input which means that signal can traverse all nodes, including phosphosites that are not present in the data, all phosphosites were included in the validation. The validation set in this case was kinase-substrate relationships listed in PhosphositePlus (Hornbeck et al., 2015).

3.2.8 Enrichment of predicted kinase-substrates

Each kinase with more than five predicted substrates had a Reactome (Jassal et al., 2020) enrichment analysis conducted on their substrates. I will refer to the set of predicted substrates of each kinase as a substrate set. For this analysis, only substrate sets larger than five were included. For this analysis kinase-substrate relationships found in PhosphoSitePlus (Hornbeck et al., 2015) were excluded from the SELPHI2 predictions. Different cut-offs were used to select a high confidence network: 0.5, 0.6, 0.7, 0.8 and 0.9. In this manner I analysed SELPHI2 predictions as well as SELPHI2 predictions that were corroborated by either experimental study (Hijazi et al., 2020; Sugiyama et al., 2019) (See chapter 2.2.10). Thirdly, the same enrichment analysis was run on kinase-substrates found in PhosphoSitePlus for comparison. The *ReactomePA* (Yu and He, 2016) R package was used for the enrichment analysis. All proteins included as substrates in the network were used as a background for the analysis. I used 10 as a minimum size of pathway term and 500 as a maximum size. P-values were adjusted with the Benjamini and Hochberg method (Benjamini and Hochberg, 1995) and the p-value cut-off used was 0.05 and q value 0.2.

After significantly enriched pathways for each substrate set were identified, I used the Fisher's exact test (Fisher, 1935) to establish if the pathways that were found to be enriched within each substrate set overlapped significantly with the pathways that the upstream kinase belonged to. Pathways that include any protein included in the SELPHI2 predictions were used as a background.

The resulting p-values were then adjusted with Bonferroni correction (Bonferroni, 1936) equalling the number of kinases with more than five substrates.

3.2.9 Correlation between kinase activities derived from known substrates and activities derived from unknown substrates

I calculated the kinase activities using the KSEA (Ochoa et al., 2016) package (https://github.com/evocellnet/ksea/blob/master/R/ksea.R). KSEA takes known kinase-substrate relationships as input and uses the Kolmogorov Smirnov (Kolmogorov, 1933; Smirnov, 1948) test to assess if known substrates are overrepresented at either the upper or lower end of the $\log_{10}$ ratio distribution found in each condition. To calculate kinase activities, a compilation of perturbation data collected and curated by Ochoa and colleagues (Ochoa et al., 2016) was used and for each condition, kinase activities for all kinases were calculated with the ksea_batchKinases() function as implemented in the KSEA package. The trial parameter which determines the number of permutations conducted on the data in order to generate empirical p-values was set to 1000. The log ratios of the resulting p-values were signed based on the sign of the average $\log_2$ ratios of the substrates. Activities were derived from three different sets of kinase-substrates: known kinase-substrate relationships found in the PhosphoSitePlus database as well as from relationships predicted by SELPHI2, excluding those found in the PhosphoSitePlus database and kinase-substrates and those SELPHI2 predictions that were supported by external experimental kinase-substrate predictions (Hijazi et al., 2020; Sugiyama et al., 2019). In each case different confidence thresholds were tested: 0.5, 0.6, 0.7, 0.8, and 0.9.

The similarity between the kinase activity profile attained by using the interactions extracted from PhosphoSitePlus and the predictive kinase-substrate sets was calculated using the pairwise Spearman's *rho* coefficient for each kinase between the literature derived activities and the activities estimated from the SELPHI2 predictions at different thresholds and for the corresponding kinase.

3.2.10 Development of the SELPHI server

I developed the SELPHI2 server using the *shiny* (Chang et al., 2019) R package for website development. To provide information on which kinase is the most likely to phosphorylate the phosphosites included in the data set, the SELPHI2 kinase-substrate

predictions were used. The user can also choose to predict upstream kinases with correlation analysis similar to the earlier SELPHI version (Petsalaki et al., 2015). In this version phosphosites found on kinases are correlated with all the other phosphosites. Spearman's *rho* is used to calculate the correlation coefficient. In cases where more than one phosphosite is found on each kinase the pair with the highest correlation coefficient is selected.

To conduct an enrichment analysis on the data set a user defined threshold needs to be selected to identify up and downregulated phosphosites. Another way of selecting phosphosites to enrich is to have the data clustered by either k-means(Hartigan and Wong, 1979) clustering or mixed Gaussian models as implemented by the Mclust R package (Scrucca et al., 2016). If k-means is selected, the gap statistic (Tibshirani et al., 2001) as implemented clusterGenomics (Nilsen and Lingjaerde, 2013) R package or the silhouette method (Rousseeuw, 1987) as implemented by ancova(Vidal et al., 2017) R package is used to calculate the optimal number of clusters. The members of each cluster are then used as a gene set to conduct the enrichment on. For the enrichment analysis *enrichR* (Kuleshov et al., 2016) R suite was used. The databases available for the enrichment are: *GO* (Gene Ontology Consortium, 2021), KEGG (Kanehisa, 2019), Reactome (Jassal et al., 2020) and Jensens diseases (Pletscher-Frankild et al., 2015). As an output a dot plot with dot size representing odds ratio and colour representing up/down regulation is produced.

The user is also offered the option to fit the kinase-substrates to their data set. To this end we add the high confidence kinase-substrate predictions to a backbone network either of high confidence edges predicted by the kinase-kinase regulatory network (Invergo et al., 2020) described in Chapter 2 or all kinase-kinase interactions found in OmniPath (Turei et al., 2016). For each condition or sample, the Prize Collecting Steiner's Forest algorithm (Akhmedov et al., 2017) is used to make the fit. The user can select a set of parameters w (number of trees), b (node prize tuning parameter) and μ (edge cost tuning parameter). Furthermore, the user can define a threshold for $\log_2$ ratios to use as prizes for the prize collection.

The user can also generate the sequence logo for each kinase from the substrates that are included in any of the subnetworks generated above. The R package *Logolas* (Dey

et al., 2018) was used for logo generation. If the kinase in question has more than 15 predicted substrates, the user can choose an enrichment depletion logo is drawn using a background frequencies of amino acids as derived from the UniProt proteome database (UniProt Consortium, 2018). If the kinase has less than 15 predicted substrates a standard sequence logo using information content is generated.

Kinase activities were calculated for kinases by selecting predictions from the top 5% of kinase-substrate predictions. For each condition or time point, the overrepresentation of the kinase's substrates were calculated with the Kolmogorov Smirnov test using the ks.test() function in R with the alternative parameter set to greater, indicating that overrepresentation at the upper end of the $\log_2$ ratios in each condition is being calculated. The $-\log_{10}$ of the p-value is then calculated and used as a proxy for activity.

3.3 Results

3.3.1 A probabilistic, data-driven kinase-substrate network

A set of predictive features including kinase specificities, co-expression, and the correlation between kinase activities and phosphorylation of potential phosphosites and disorder in regions surrounding phosphosite were integrated to generate a set of kinase-substrate predictions (details in Methods). These features were used to train a random forest classifier to assign probability to each kinase-substrate relationship. Predictions were made for more than 22 million kinase-substrate relationships between 367 kinase and more than 80,000 phosphosites.

Biological networks are typically considered to be relatively sparse with most proteins only interacting with a relatively low number of other proteins. The predictions made here reflect this view with less slightly more than 1% (286,392) of the predictions being high confidence with probability of equal or higher than 0.5. The median probability of all kinase-substrate predictions is 0.010. The human interactome is often thought of as being largely understudied and undiscovered and due to study bias the same is true for the human phospho-regulatory network. Taking this into consideration, it is unsurprising that the vast minority of the high confidence edges identified by our predictor (1.6 %) were

actually present in PhosphoSitePlus (Hornbeck et al., 2015) **Table 3.1** shows the different precision and recall values, that is what portion of the edges in PhosphoSitePlus are included in the high confidence prediction set at different confidence cut-offs.

Table 3.1:.*Overview the number of edges at different high confidence threshold the portion of those that are present in PhosphoSitePlus as well as the recall (Portion of training set that is captured).*

Probability cut-off	**No. edges**	**% present in PSP**	**% edges in train set**
0.5	286,392	1.6	85
0.6	145,146	2.9	75
0.7	55,034	6.5	65
0.8	14,375	18	46
0.9	2,061	48	18

3.3.2 SELPHI2 captures known relationships while making predictions for less studied proteins

While the overwhelming majority of the high confidence edges identified by SELPHI2 are unknown, known edges are overrepresented at the upper end of the edge probability distribution. For each of the one hundred training sets, we conducted a ten-fold cross validation and found that our average performance as measured in AUROC was 0.956 across the hundred models trained (Figure 3.1 B).

As the aim of this thesis chapter is to investigate the non- and under-studied parts of the phosphoproteome, I looked at how well understudied proteins were represented in this study and compared them with kinase-substrate interactions in PhosphoSitePlus (Hornbeck et al., 2015). Our high confidence (>0.5 probability) kinase-substrate predictions included 367 kinases and 32,566 phosphosites found on 6,787 proteins compared to 388 kinases found in PhosphoSitePlus which phosphorylate 7,255 phosphosites on 2,308 proteins. For comparison, the weighted mean number of citations

per kinases in the PhosphoSitePlus database is 500 and substrates have mean citation number of 115. When looking at high confidence edges my predictions had on average 69 citations per kinase and 23 citations per phospho-acceptor (Figure 3.1 A). This set of high confidence edges includes substrates that have not been assigned an upstream kinase in the literature, stressing the value of this predictor as a tool to explore the space of less studied proteins.

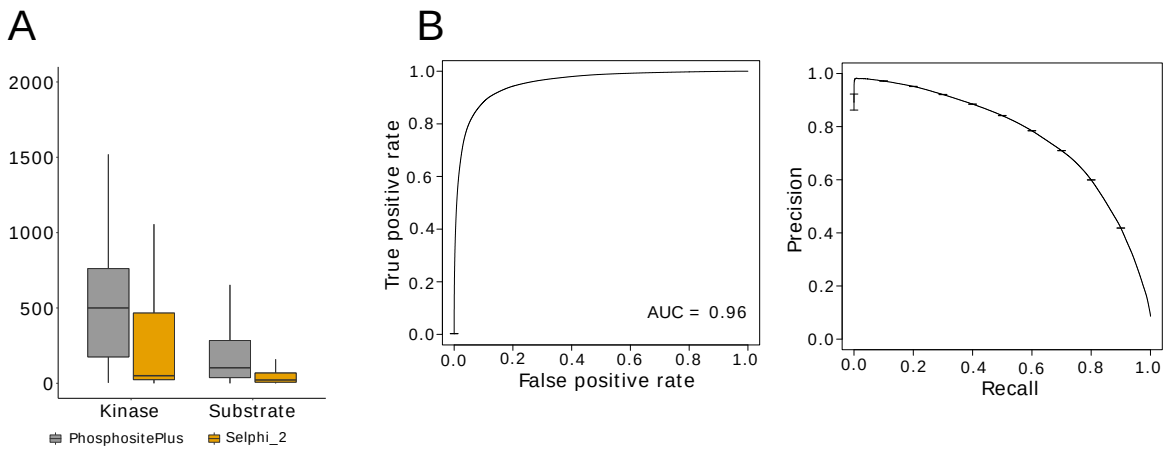

Figure 3.1: *General description of kinase-substrate predictor. The average number of citations per kinase and substrates in SELPHI2 predictions was significantly lower than the average number of citations per kinase and substrates in the PhosphoSitePlus database, indicating that the kinase-substrate predictions have the ability to capture relationships between less studied proteins (A). The average AUROC derived from 100 ten-fold cross-validations run during model training resulting in a high AUC of 0.96 was achieved. Precision-recall curve drawn from the same set of 10-fold cross validation runs (B).*

3.3.3 Independent experimentally supported kinase-substrate relationships have higher probability assigned to them compared to background

One of the main challenges of analysing kinase-substrate predictions is evaluating how much confidence can be assigned to the vast number of edges that are not present in the

current literature. While methods such as cross-validation give us an indication of the ability of the predictor to capture known edges, it is not certain that the computational predictor has the capacity to capture novel edges.

To further validate the SELPHI2 predictions, I set out to compare SELPHI2 to experimental predictions. Currently various high throughput experimental kinase-substrate predictions have been made that could shed light on the power of SELPHI2 to capture novel relationships. For this analysis I used kinase-substrate predictions from an experimental kinase-substrate prediction study made earlier (Hijazi et al., 2020; Sugiyama et al., 2019).

The predictions made by Sugiyama and colleagues were done based on *in vitro* experiments. Phosphopeptides were de-phosphorylated and introduced to a kinase and the resulting phosphorylation levels were compared to a control sample that was not introduced to the kinase. The method developed by Hijazi and colleagues, on the other hand, assigned kinases to phosphosites based on the decrease in phosphorylation in cells that were introduced to 61 different kinase inhibitors, inhibiting the activities of 103 different kinases in three different human cell lines: NTERA2, MCF7 and HL60.

These methods both have their shortcomings. While Sugiyama' study was done *in vitro*, meaning that many of the predicted kinase-substrate relationships cannot be expected to occur in vivo given different environments and the context specificity of kinases. At the same time, many kinase-substrate interactions that take place within cells cannot be expected to be captured *in vitro*. The predictions done by Hijazi were based on data from human cell lines but relied on decrease in phosphorylation levels upon kinase inhibition meaning that it is hard to ascertain if the decrease in phosphorylation is due to inhibition directly.

Given these different methodologies with different limitations, the overlap between the two predictions can be expected to yield a set of relatively high confidence kinase-substrate predictions. With this in mind, I looked at whether kinase-substrates predicted by these two methods had higher probabilities assigned to them compared to the background of unsupported edges. As I was interested in interactions not found in the literature I excluded kinase-substrate relationships found in PhosphoSitePlus. Of the over 22 million predictions made by SELPHI2, 8,154 were found in the Hijazi et al. data set

and 61,980 relationships predicted by Sugiyama et al. and 76 predicted by both studies. Strikingly, the overlap between the two experimental predictions was low (n= 82) (Figure 3.2 A). I found that edges with experimental evidence supporting them had a significantly higher probability assigned to them. The background, defined as edges predicted by neither study, had a median of 0.012 probability of while edges supported by Hijazi's study having a median probability 0.051 ($W = 1.4 \times 10^{11}$, $p < 2.2 \times 10^{-16}$, Wilcoxon rank sum test) and edges corroborated by Sugiyama's study with median probability of 0.068 ($W = 1.1 \times 10^{12}$, $p < 2.2 \times 10^{-16}$, Wilcoxon rank sum test) . Significantly, the kinase-substrate interactions predicted by both methods had yet higher probability assigned to them or 0.28 ($W = 1.5 \times 10^{9}$, $p < 2.2 \times 10^{-16}$, Wilcoxon rank sum test) (Figure 3.2 B).

The overlap between the three prediction sets (SELPHI2, Sugiyama and Hijazi) is rather small with 76 interactions all together which are listed in **Appendix 3.2**. Of these 76 interactions 24 have high confidence of 0.5 or higher. There were 7 interactions with confidence higher than 0.8 which are listed in **Table 3.2**.

Table 3.2: *Experimentally corroborated Interactions* with edge probability > 0.8.

Kinase	**Substrate**	**Probability**	**STRING score**
CDK2	FAM122B 115	0.90	NA
CDK2	NFIC 323	0.86	NA
CDK2	NOP2 732	0.86	0.46
CDK2	NUMA1 2000	0.92	NA
CDK2	PDCD4 94	0.80	NA
CDK2	HNRNPA2B1 259	0.81	NA
MAPK9	EFHD2 74	0.81	NA

Earlier CDK2 has been found to phosphorylate NUMA1 at a different site; 1776. Previous analysis on human kinases in yeast indicate that kinases often target phosphosites in the vicinity of a substrate (Corwin et al., 2017) lending some credence to the hypothesis that NUMA1 is targeted by CDK2. To see if there is some further evidence of interaction between any of these high confidence pairs, I looked them up in the STRING database.

Of the five relationships in question, there is a link between CDK2 and NOP2 in STRING (Szklarczyk et al., 2021) with a combined score of 0.46 with Experimental/Biochemical Data score of 0.41 with three publications mentioning a protein-protein interaction between the three. Furthermore, the proteins co-occur in abstracts of two papers and two papers mention co-expression. The other CDK2 interactions as well as the link between MAPK9 and EFHD2 were not present in the STRING database suggesting little to no evidence that these kinases target these phosphosite. At the same time, it showcases the predictor's ability to capture interactions without any support from prior knowledge. Naturally experimental validation would be needed to be certain that these represent true relationships.

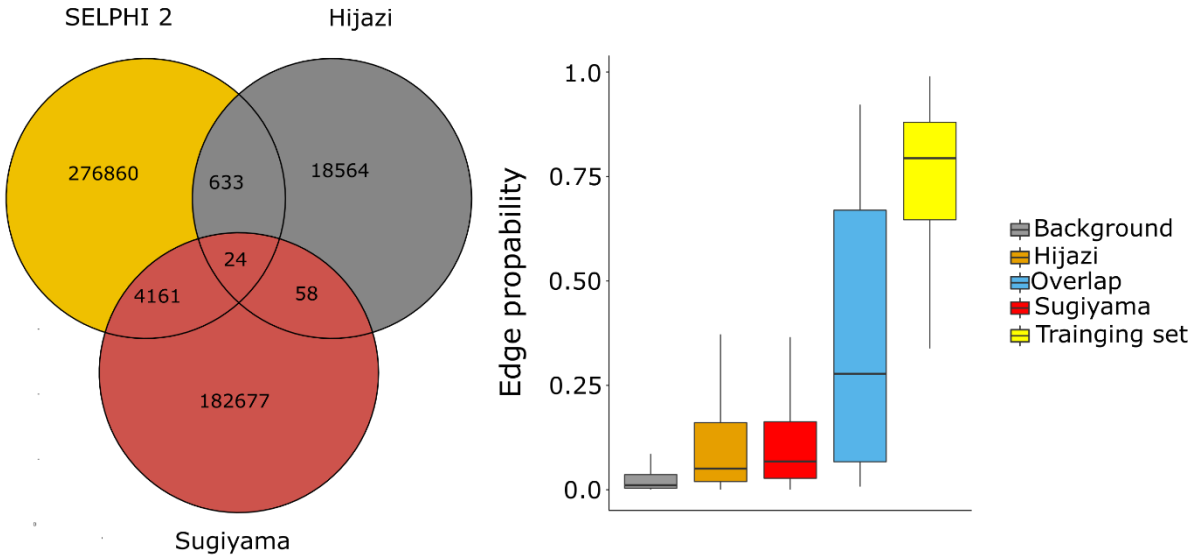

Figure 3.2: *External validation of kinase-substrate predictions. Overlap between the different kinase-substrate predictions sets. While high confidence edges in SELPHI2 share a large overlap in predictions with the other sets, the intersection between the three predictions is small, 24 edges. Given the different methodologies, a relatively high confidence can be allotted to these relationships (A). We find that SELPHI2 assigns higher probability to experimentally validated edges than the background (edges with neither experimental support nor evidence in the literature). Strikingly, edges supported by both experimental studies get assigned an even higher confidence than those supported by either study (B).*

3.3.4 Resulting network is competitive in comparison to other networks

Previously, various methods have been developed to predict kinase-substrates. Therefore, I compared SELPHI2 to five other methods: KinomeXplorer (Horn et al., 2014), GPS (Wang et al., 2020), PhosphoPICK (Patrick et al., 2015), LinkPhinder (Nováček et al., 2020) and Netphos v.3.1 (Blom et al., 2004) These methods all differ in methodology and use different sets of predictors. LinkPhinder, GPS and NetPhos v. 3.1 primarily rely on kinase specificities, that is make predictions based on peptides surrounding candidate phosphosite and how well the surrounding peptide fits into the kinase specificity model. The other methods also base their predictions in part on kinase specificities but also various other features. KinomeXplorer uses the STRING network which scores protein-protein interactions based on evidence found in literature and other sources. PhosphoPICK bases their prediction on PPI from STRING (Szklarczyk et al., 2021) and BioGRID (Oughtred et al., 2019) as well as protein abundance under different phases of the cell cycle.

These different methods were compared in two ways: Their ability to discriminate between known kinase-substrate relationships and unknown edges and, on the other hand, how well they discerned between experimentally predicted edges and the edges not predicted by either study. By comparing the ability of the methods to capture experimentally predicted edges (Sugiyama et al., 2019) we can gauge their abilities of the methods to explore the space of undiscovered kinase-substrate interactions. Such comparisons are complex to carry out. Each method makes predictions for different numbers of kinases and substrates and therefore have different gaps in predictions. For the experimental comparisons, this issue was addressed by only considering the overlap in prediction made by both SELPHI2 and the comparison method. Due to the differences in each method's prediction space (i.e. number of kinases the method is able to make predictions for), five different comparisons were made, one for each method.

The methods' ability to discern between the experimentally predicted edges and the rest were assessed by calculating the area under the ROC curve drawn from each method's prediction score and the label of each edge. In all cases, SELPHI2 performed better at

capturing the predicted kinase-substrates manifesting the method's value in the discovery of unknown kinase-substrates (Figure 3.3).

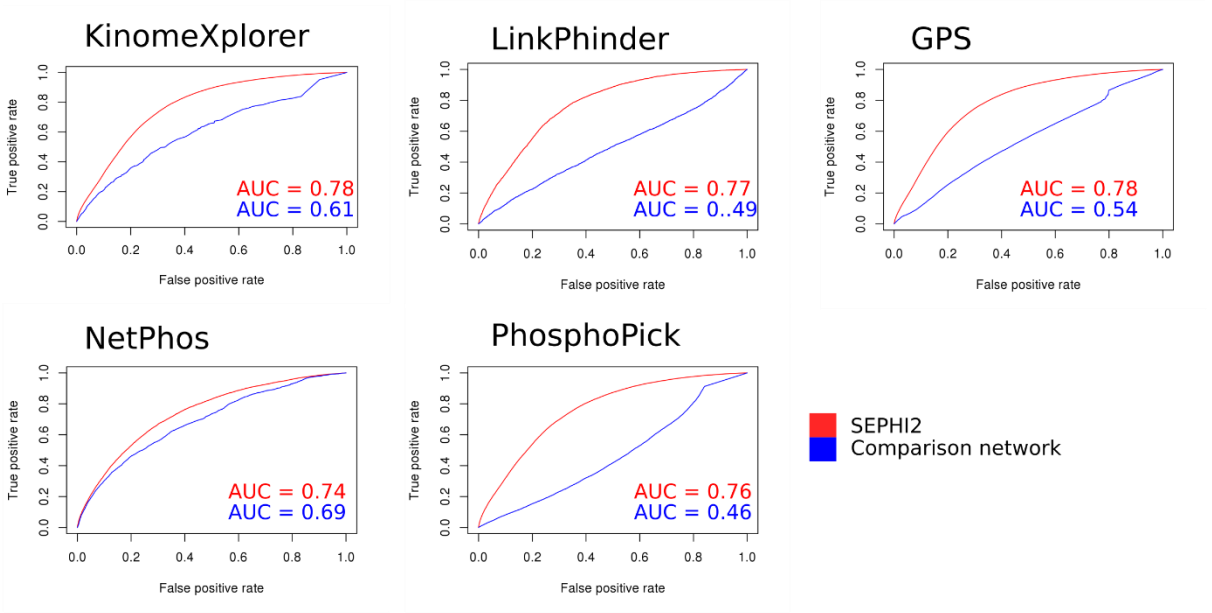

Figure 3.3 *Comparison with other methods; experimental predictions***.** *The different methods' ability to discern between experimentally supported edges and edges without any experimental in each case, SELPHI2 outperforms the comparison method.*

To assess the methods' ability to predict known kinase-substrate relationships a different approach was used as the features included in SELPHI2 were used and cross validation employed to assign probabilities to the interactions used as training set (Methods). Like before five different comparisons were made between SELPHI2 and each of the other methods. Then the AUROC score for each training set was calculated for both SELPHI2 and the comparison method and the scores were averaged over the one hundred runs. SELPHI2 performed the best of any of the tested methods. The results can be seen in Figure 3.4

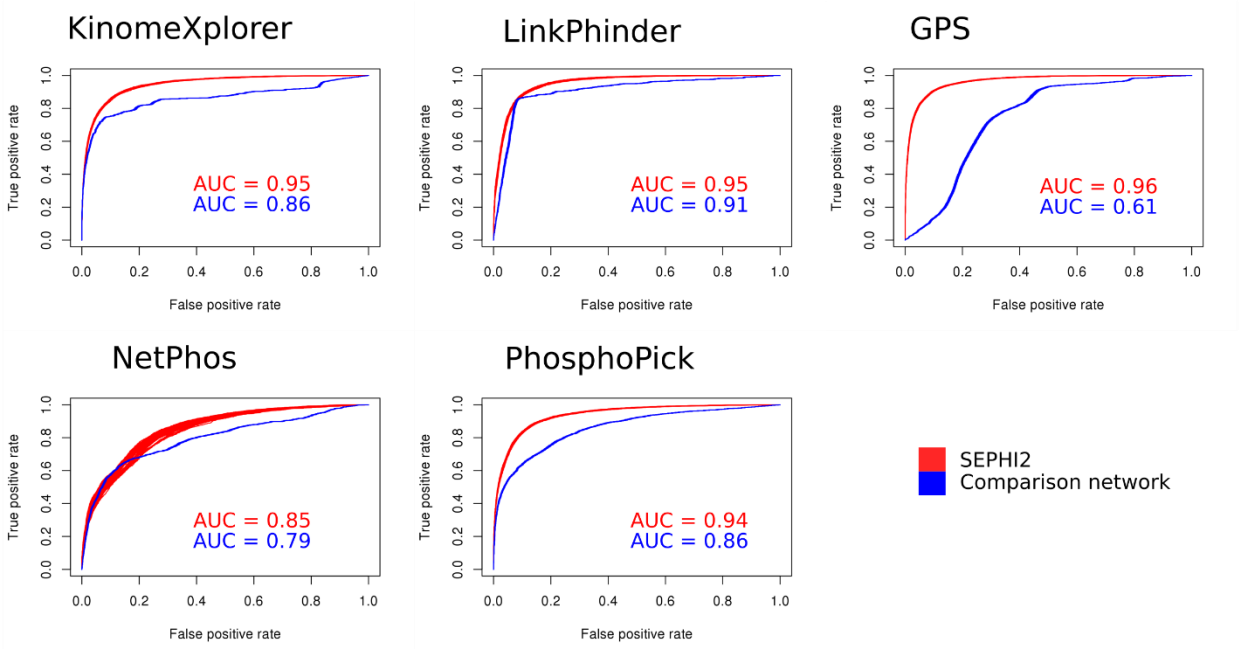

Figure 3.4 *Comparison with other methods in terms of known interactions .AUROC comparison between SELPHI2 and the other methods. ROC curves were generated from ten-fold cross validations conducted on the training set and averaged over hundred different runs. In each case, SELPHI2 outperforms the comparison method.*

3.3.5 Prediction of signed kinase-substrate relationships

PTMs, including phosphorylation, often lead to changes in protein activity and/or localization. To capture these effects on kinase-substrates, I set out to predict the sign of phosphorylation to be used along with the kinase-substrate predictions. With the same initial set of predictive features as in the kinase-substrate predictions, I used recursive elimination feature selection to select the best model. The final 45 used for prediction are listed in **Appendix 3.1**. For the training set I used signed kinase-substrate relationships found in the SIGNOR (Licata et al., 2020) database consisting of 497 inhibiting and 673 activating interactions. For the prediction we selected phosphosites with functional scores of higher than 0.5. Overall, more than 2 million predictions were made between 367 kinases and over 7,000 phosphosites. I found that by using the features selected I was able to achieve an AUROC of 0.83. One issue with the training set is that a large portion of signed interactions are kinases regulating other kinases. Therefore, to assess this

method, I used the same training set and used cross validation to assign probabilities to the training set and looked at the performance only considering interactions between two kinases. I found that there was a small decrease in performance with AUROC of 0.80. The ROC and precision-recall curve generated from ten-fold cross validation can be seen in Figure 3.5.

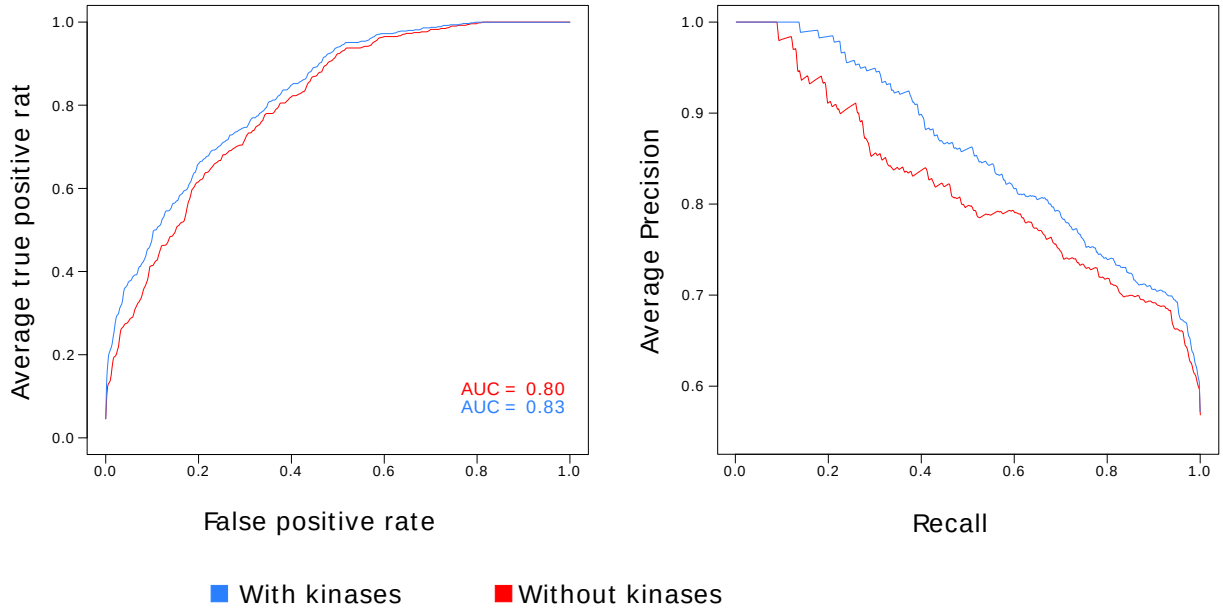

Figure 3.5: *Performance of signed kinase-substrate predictions. ROC and PR curve for signed kinase-substrate predictions. The signed predictor has good performance for signed interactions including for non-kinase substrates. The difference between the two sets is greater in terms of precision-recall with predictions for non-kinase-substrates only having a greater drop in precision as recall increases.*

3.3.6 Mapping kinase-substrate relationships onto data improves precision and F1

In network biology, networks are often fitted to high throughput data sets to extract context specific networks (Hill et al., 2016; Saez-Rodriguez et al., 2011). I therefore investigated if mapping the kinase-substrate relationships to high throughput data could help select for known kinase-substrate relationships as well as single out context specific kinase-substrate relationships. To form a single probabilistic phospho-signalling network the

SELPHI2 predictions were combined with the kinase-kinase regulatory network discussed in Chapter 2, forming a network of kinase-kinase regulatory circuits (Invergo et al., 2020) with predicted substrate phosphosites as nodes with no outgoing edges (Figure 3.6 A). The kinase-kinase regulatory net was used as a backbone as the SELHPI2 predictions predict phosphorylation rather than regulation. To select high confidence edges for both networks a threshold of 0.5 was applied to the combined network. The data used to fit the data was a compilation of phosphoproteomic data sets compiled and described earlier in a previous publication (Ochoa et al., 2016). The combined network was fitted to a total of 436 datasets.

Different mapping methods were tried (description in the methods section 3.2.7) with varying success. The best heat diffusion of those I tested, z, having a median F1 score of 0.046 and median precision of 0.032. The corresponding values for the input network were 0.032 and 0.016. Below in **Table 3.3** an overview over the performance across the different heat diffusion methods can be seen

Table 3.3 *Performance of different heat diffusion methods. Rest of the diffusion methods did not yield results.*

Method	**F1-score**	**precision**
raw	0.028	0.015
z	0.046	0.032
ml	0.028	0.015
gm	0.028	0.015
Ber_s	0.028	0.015

Using the shortest path between the kinases impacted by kinase inhibitors and up and downregulated phosphosites slightly improved performance. While the raw context specific input networks had a median precision of 0.025 the pruned network had a precision of 0.03. The corresponding values for the F1 score were 0.05 and 0.05. In both cases the difference was significant (F1 score: W = 373, p = 0.041, precision: W= 405, p = 0.0077, Wilcoxon rank sum test)

However, the best performing method was Prize Collecting Steiner's Forest, or PCSF (Akhmedov et al., 2017). For each parameter combination (see Methods 3.2.7, second paragraph) we selected the fitting that yielded the best F1 score. The resulting condition-specific subnetworks were found to have a higher proportion of known edges. With the F1 scores of the fitted subnetworks (n = 435) being 0.19 while the unpruned input network had a F1 score 0.034 (Figure 3.6 B). The improvement in precision was even greater with the mean precision of the pruned subnetworks being 0.22 and 0.017 for the unpruned input networks. Both comparisons yielded a significant difference (F1: $W = 1.7 \times 10^5$, $p < 2.2 \times 10^{-16}$, precision: $W = 1.7 \times 10^5$, $p < 2.2 \times 10^{-16}$, Wilcoxon rank sum test). This indicates that if this probabilistic network is fitted to an independent data set the predictions could be used in an exploratory manner to identify kinase-substrate relationships that could be feasible to test further experimentally as well as identifying kinase-substrate relationships that are likely to be active in a given context.

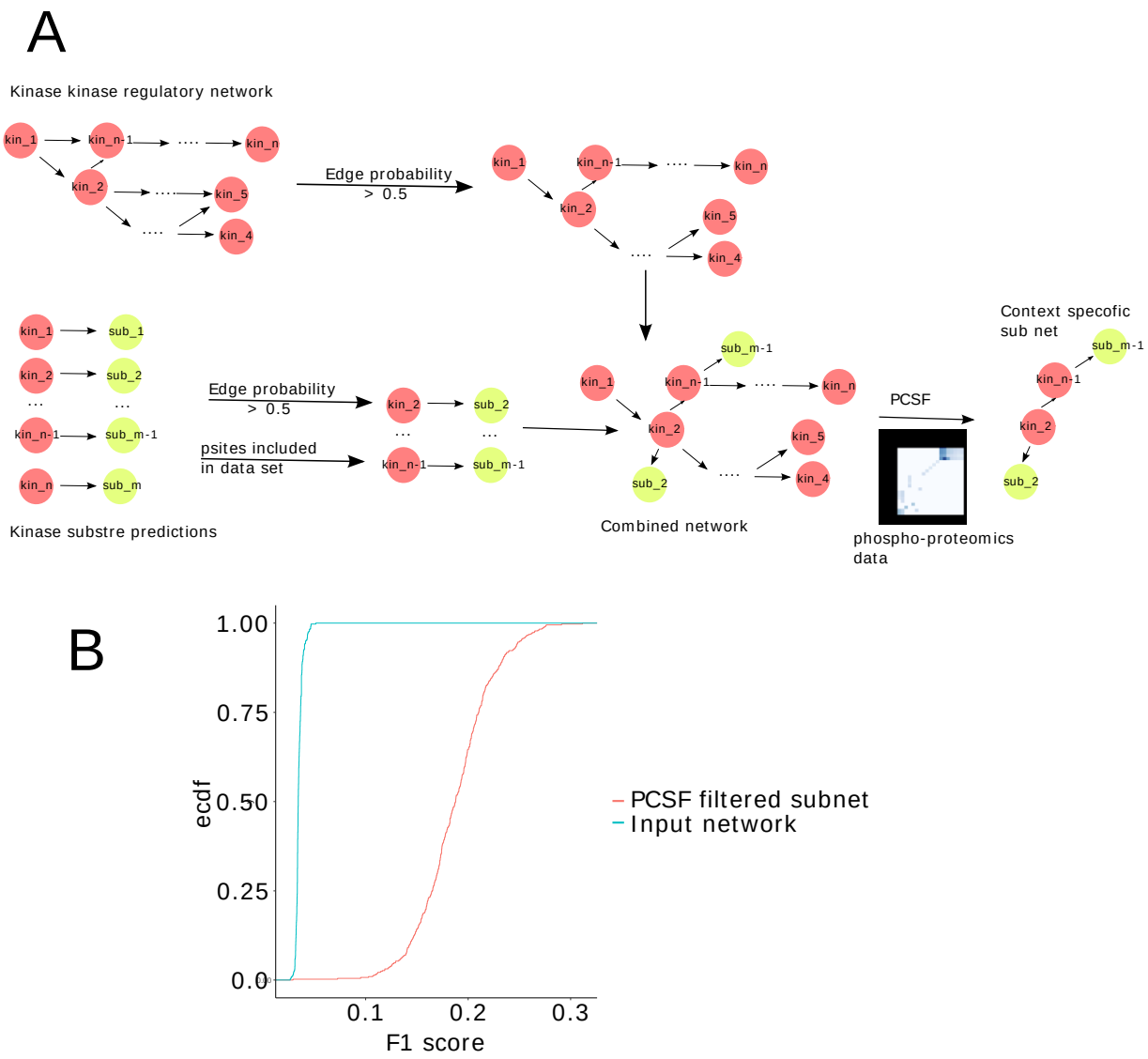

Figure 3.6: *PCFS fitting of predictions to independent data*. *Overview of the methodology, where kinase-kinase regulatory network is combined with kinase-substrate predictions with PCSF used to fit the data to independent phosphoproteomics data (A). Kinase-substrate relationships that are included in the resulting sub networks had a higher F1 scores (B).*

3.3.7 Analysis on the overlap between the functional assignment of kinases and their predicted substrates

The signalling system is often thought of as a modular network organized into a set of different functional modules or pathways that are highly interconnected in contrast to the generally sparse signalling network. Therefore, it stands to reason to assume that kinase-substrates belong to the same modules or pathways as their upstream kinase. I set out to evaluate whether the predictions made by SELPHI2 capture this functional association between kinases and their predicted substrates and compared SELPHI2's results with results obtained by using kinase-substrate relationships from PhosphoSitePlus.

To this end, I conducted a Reactome (Jassal et al., 2020) pathway enrichment analysis on predicted kinase-substrates. Significant pathways from this analysis will be referred to as the substrates' pathways. I then looked at which pathways in Reactome each kinase belonged to. These pathways will be referred to as the kinase's pathway. Then I set out to assess whether there was a significant overlap between the kinases' pathways their respective substrates' pathways. This analysis was conducted with known kinase-substrates retrieved from PhosphoSitePlus (Hornbeck et al., 2015) and compared with high confidence SELPHI2 prediction as well high confidence SELPHI2 prediction supported by either of two independent experimental kinase-substrate predictions at cut-offs of 0.5, 0.6, 0.7, 0.8 and 0.9.

I looked at the proportion of kinases that had a significant pathway overlap with their substrates across the kinase-substrate sets tested. Kinase-substrate relationships extracted from PhosphoSitePlus yielded the highest proportion. Notably I found that the unfiltered set of kinase-substrates yielded significantly higher proportions across all cut-offs than the experimentally supported kinase-substrate prediction sets ($W = 20$, p-value = 0.019). Below in **Table 3.4** that shows the overall results from this analysis including the portion of kinases whose pathways overlap significantly with its substrate's pathways. P-values indicate if the overlap portion differs significantly from the overlap portion achieved by the PhosphoSitePlus set. The literature interactions downloaded from PhosphoSitePlus perform the best. This can partly be explained by the fact that the pathways are built from known interactions. All sets of SELPHI2 predictions tested had a

significantly lower portion of significant pathway overlap than PhosphoSitePlus ($p_{portion}$). Furthermore, the pathway overlap enrichment profile, that is the odd ratios derived from the overlap enrichment analysis, of the different SELHPI2 kinase-substrate interaction sets were quite different across the board from the enrichment profile derived from the literature kinase-substrate interactions as measured by Spearman's *rho* (Figure 3.7). The most similar set was SELPHI2 edges that had higher probability than 0.8 and had been supported by experimental predictions with Spearman's *rho* of 0.39. However, it should be noted that none of the correlations were significant.

One limitation of this analysis is as before the large unexplored space of the phosphoproteome. Therefore, the proteins included in this analysis can be expected to participate in a large number of pathways that they have not been assigned to have either not been discovered. Furthermore, assignment of proteins to pathways can be somewhat arbitrary making enrichment analysis results hard to interpret (Mubeen et al., 2019).

Table 3.4: *Kinase-substrate pathway similarity at different cut-offs*. *Overlap in kinase pathway membership and substrate pathway enrichments. Generally, the experimentally filtered SELPHI2 has a lower overlap in function between kinases and their predicted substrates than the full set of high confidence SELPHI2 predictions.*

Kinase-substrate set	**Probability Cut-off**	**No. overlap enrichments tested**	**Portion of significant pathway overlap**	**$p_{portion}$**
PhosphositePlus	NA	192	0.41	1.00
SELPHI2	0.5	335	0.19	1.23×10^{-7}
SELPHI2	0.6	295	0.21	1.80×10^{-6}
SELPHI2	0.7	244	0.21	7.40×10^{-6}
SELPHI2	0.8	177	0.18	1.27×10^{-7}
SELPHI2	0.9	45	0.18	0.0059
SELPHI2_exp	0.5	103	0.019	1.72×10^{-12}
SELPHI2_exp	0.6	85	0.059	9.09×10^{-9}
SELPHI2_exp	0,7	51	0.059	4.93×10^{-6}
SELPHI2_exp	0.8	17	0.12	0.034

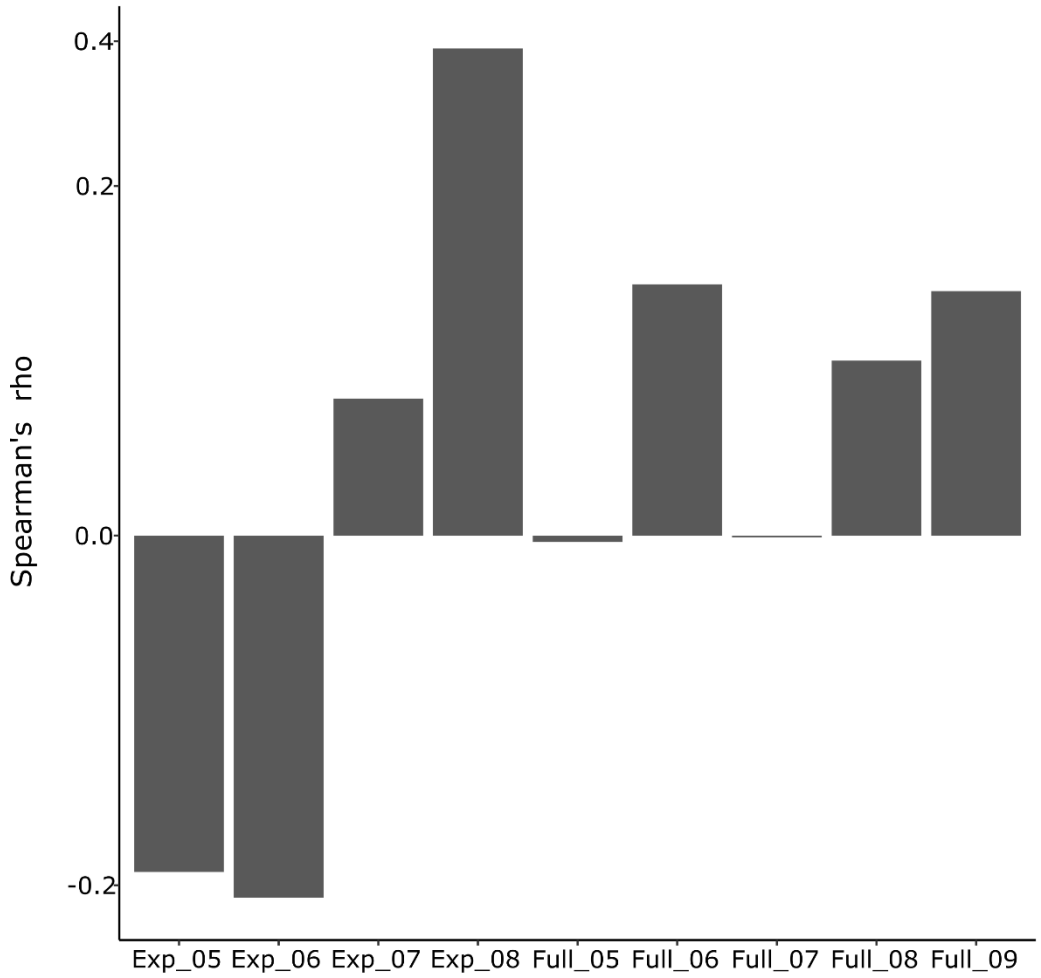

Figure 3.7*: Correlation between kinase activities derived from known interactions and predictions. Enrichment profile similarities between all the kinase-substrate prediction sets and PhosphoSitePlus. Sets represented as Full indicate that SELPHI2 predictions were filtered by confidence threshold and Exp that probability cut-off was applied and that only experimentally supported predictions were retained. For example Full_07 represent SELPHI2 predictions with higher probability assigned than 0.7. All sets yielded a relatively different enrichment profile to PhosphoSitePlus with SELPHI2 edges that had higher probability than 0.8 and had been supported by experimental predictions being the most similar.*

3.3.8 Correlation between kinase activities derived from known substrates and activities derived from unknown predicted substrates

Various methods have been proposed to predict kinase activities. Previously kinase activities under many different conditions had been calculated using known substrates. One would expect that if a kinase is active or inactive that their substrates would be affected. One limitation of this assumption is, though, that many of predicted kinase-substrates could be context specific and therefore not active in a given context. Nevertheless, I conducted a kinase activity estimation by using the kinase-substrate set in PhosphoSitePlus(Hornbeck et al., 2015) and a separate estimation based on my set of predicted kinase-substrate relationships, to establish if estimating kinase activity based on experimentally corroborated kinase-substrate relationships gave estimates more similar to the literature derived estimates.

By using kinase-substrate relationships from PhosphoSitePlus I was able to make predictions for 248 kinases across 396 perturbation conditions. As different cut-offs were used to extract high confidence edges, kinase activity estimates for a different number of kinases could be derived at different cut-offs. An overview of number of kinases available for activity estimates at a different cut-offs can be seen in **Table 3.5** as well as how many of those could have their activity estimated by using literature kinase-substrate relationships.

Table 3.5: *Overview of number of kinase activity estimation at different edge confidence cut-offs. Estimates were calculated for SELPHI2 predictions that were experimentally supported as well as non-experimentally supported predictions.*

Probability cut-off	Corroborated by external experimental predictions	No kinases with activity estimates	Kinase overlap with literature-derived activities
0.5		367	254
0.6		350	243
0.7		326	234
0.8		255	176
0.9		126	92
0.5	✓	187	145
0.6	✓	145	116
0.7	✓	110	90
0.8	✓	71	63
0.9	✓	15	13

There was little similarity between the kinase activity profiles across conditions of the literature derived activities to the ones derived from the predictions. In the case of experimentally supported edges the average Spearman's *rho* coefficient was less than 0.1 for all cut-offs and of 427 kinase activity pairs being correlated, 25 were significant ($p < 0.05$, Bonferroni corrected). For the unfiltered SELPHI2 predictions the highest correlation was achieved at cut-off 0.7 with Spearman's *rho* of 0.16 (Figure 3.8) and 27 of the 234 activity correlations assessed being significant ($p < 0.05$, Bonferroni corrected). The correlation between the cut-off of the kinase-substrate prediction and the median correlation coefficient with the literature derived activities was -0.5 (Spearman's *rho*) with p-value of 0.45 for the full set and a Spearman's *rho* of 0.36 and p-value of 0.55 for the experimentally supported relationships. Thus, the experimentally corroborated high confidence kinase-substrate relationships did not yield kinase activities that were more

similar to the literature derived activities than the unfiltered list of kinase-substrate predictions.

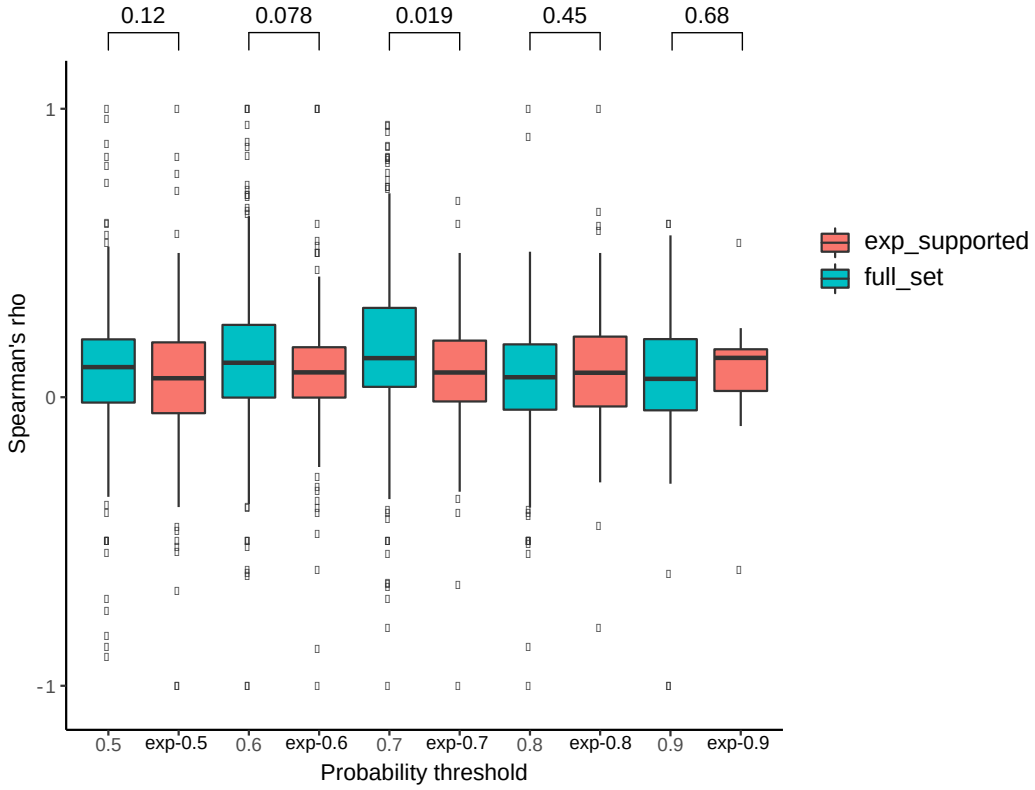

Figure 3.8: *Correlation between kinase activities derived from known interactions and predictions. The average (mean) correlation between kinase activities generated by literature-derived kinase-substrate relationships and high confidence predicted kinase-substrates (blue) and high confidence predictions corroborated by experimental kinase-substrate prediction (red). Correlations are low across the board and the difference between SELPHI2 and experimentally filtered SELPHI2 is insignificant apart from at the probability cut-toff of 0.7. The p-values were calculated using the Wilcoxon rank sum test.*

3.3.9 Overview over the SELPHI2 server

A secondary aim of this project was to rewrite the SELPHI (Systematic Extraction of Linked Phospho-Interactions; (Petsalaki et al., 2015) server which had previously been developed by Petsalaki and colleagues. The SELPHI server provides a platform for users to analyse their phosphoproteomics data. The tool enables the user to cluster and conduct a pathway enrichment on their data set as well as predicting upstream kinases through correlation analysis. Here I rewrote the server and based the kinase-substrate prediction on the SELPHI2 machine learning-based predictions described in this Chapter. To showcase how the SELPHI2 server can be used I tested it on data set generated by Köksal and colleagues (Köksal et al., 2018) where EGFR Flp-In cells were stimulated with EGFR at 8 different time points: 0, 2, 4, 8, 16, 32, 64, or 128 min. The user uploads the data through the upload page (Figure 3.9). The standard input is a $\log_2$ transformed ratio (conditions/control) with phosphosites coded as HGNC gene symbol and a number representing location within protein sequence as row names in the first column. If the input file is in a different format the user can also provide information on which columns contain information on proteins, phosphosite location in protein sequence and columns containing data.

After uploading the phosphoproteomics data, the top kinase-substrate predictions are shown in a table. All kinase-substrate relationships can be downloaded as a .tsv file. Finally the upload page displays a density plot which helps the user visualize the probability distribution of the predicted edges both the ones that are not present in the literature (blue) and those that have previously been discovered (red) (Figure 3.10 B) .

SELPHI2 also provides the user with the ability to conduct an enrichment analysis to identify pathway overrepresentation among up or down regulated proteins. The user can select between several databases: Jensen's diseases (Pletscher-Frankild et al., 2015), KEGG (Kanehisa, 2019), Reactome (Jassal et al., 2020) and GO (Gene Ontology Consortium, 2021).

For the purpose of this overview, the KEGG database is used as a reference pathway to conduct the analysis and the log2 ratio threshold of one is applied to select up (> 1) and down (< -1) regulated phosphosites to conduct the enrichment analysis. Enriched pathways are represented as a dot plot where the size of the dot represents the odds

ratio of the enrichment while the colour represents the sum of the log_2 ratios of the phosphosites involved in the given pathway and the different columns represent different time points or samples. In this case, some pathways are only overrepresented at several of the 8 time points such *as spry regulation of FGF signalling* which is overrepresented at the last three time points: 32,64 and 128 minutes while others like EGFR signalling are found to be overrepresented at all time points (Figure 3.10 A). The user can also cluster the data using the Gaussian mixture model based clustering implemented by the *Mclust* R package (Scrucca et al., 2016) or k-means (Hartigan and Wong, 1979) clustering and conduct an enrichment analysis on the phosphosites belonging to each cluster.

The user can also map the kinase-substrate predictions onto their data to extract a context specific network for each context or time point. A network is constructed to link the kinase-substrate predictions together; the user can either use the probabilistic kinase-kinase network described in Chapter 2 or list of kinase-kinase interactions downloaded from OmniPath (Turei et al., 2016) . Prize collecting Steiner's forest (Akhmedov et al., 2017) is used to identify the optimum sub network. The user can then select their parameters and probability cut-offs for the probabilistic networks. SELPHI2 returns an interactive network where the user can zoom in and select nodes to see their log_2 ratios. The resulting sub-networks for each condition can then be downloaded for further analysis (Figure 3.11A). The resulting kinase-substrates can then be used to generate sequence logo from the kinases and their substrates that are involved in the sub-networks (Figure 3.11 B).

The user can also calculate estimated kinase activities from predicted kinase-substrate relationships. SELPHI2 returns heat maps for the kinases whose substrates are significantly overrepresented at the upper end of the log_2 ratio under the given condition/time point. Activity heat map is also generated from kinase-substrates derived from the literature for comparison. In this case, the predicted activity profiles are quite different from the literature derived activities with SRC being the only active kinase present in both heat maps. No serine/threonine kinases were significantly active based on the predicted kinase-substrate set while several were active based on the kinase-substrates found in PhosphoSitePlus (Figure 3.12 A) Lastly, the user can see how well the predicted kinase-substrates are supported by two independent experimental kinase-substrate predictions (Hijazi et al., 2020; Sugiyama et al., 2019) (Figure 3.12 B).

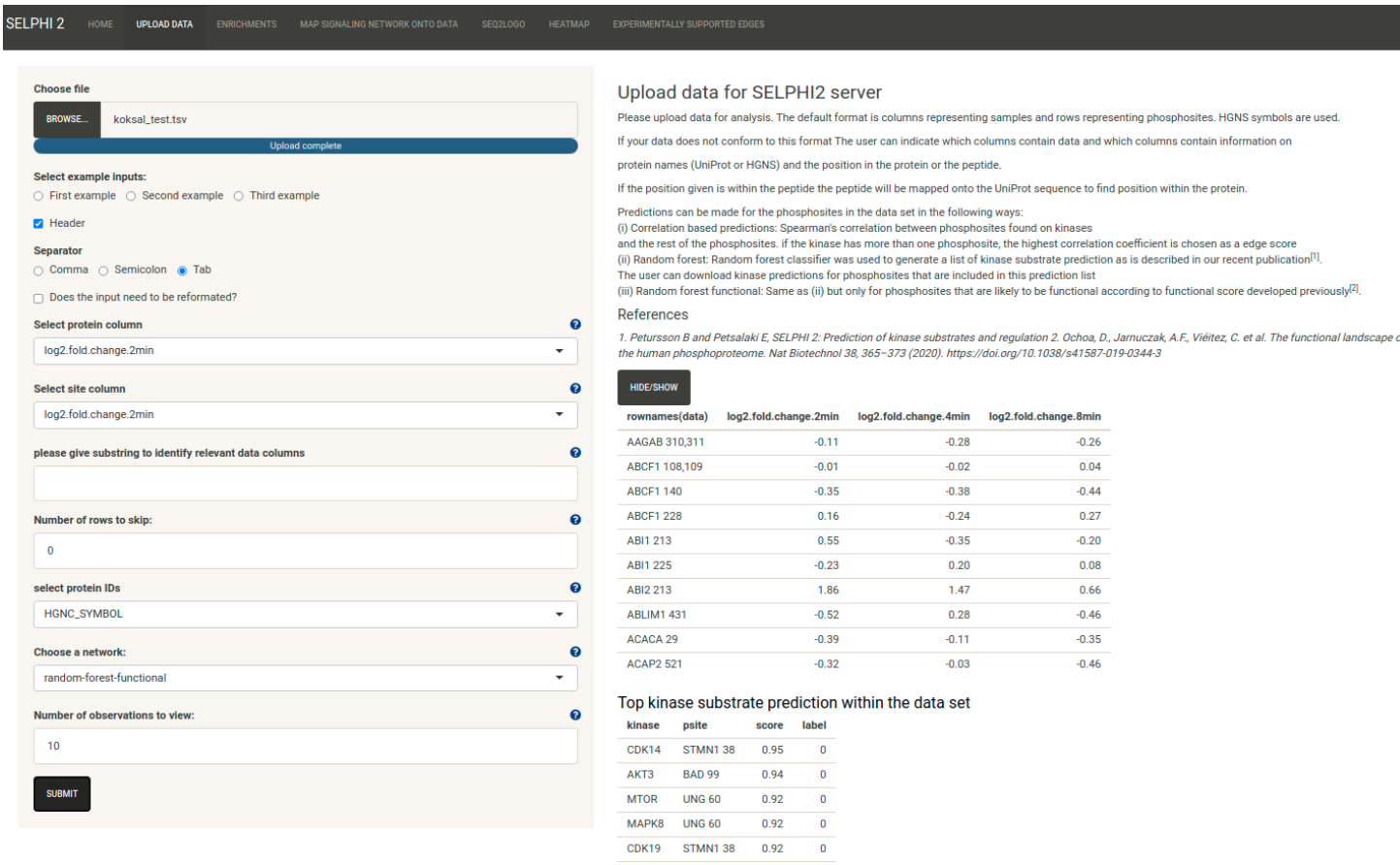

Figure 3.9: *Screenshot from the SELPHI2 server data upload page. The user can upload data for further analysis. In this example and subsequent analysis (Figures 3.10-3.13) time series data from publication by Köksal and colleagues will be used* (Köksal et al., 2018).

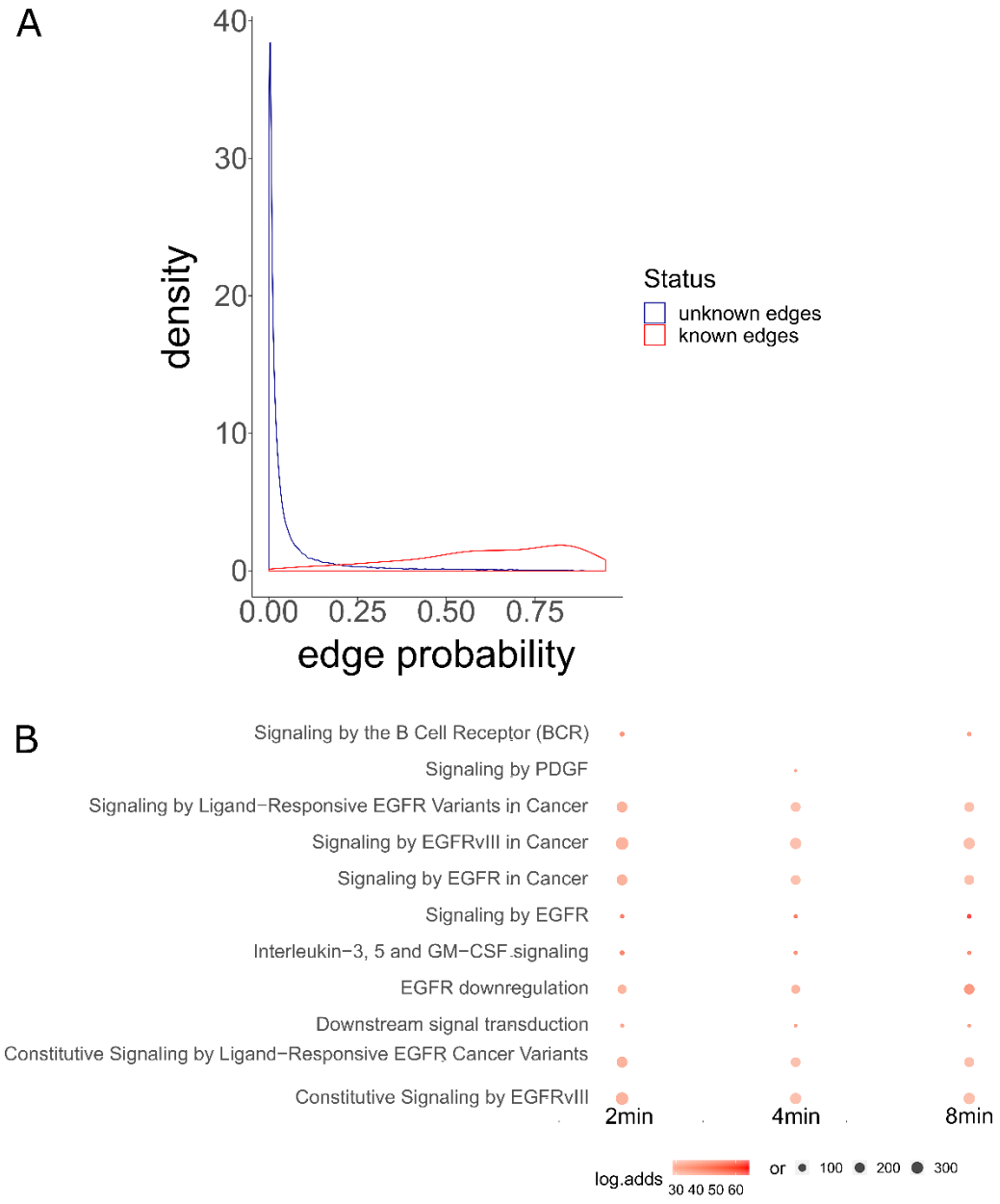

Figure 3.10: *SELPHI2 density and enrichment plots. Example output from a SELPHI2 analysis. Once data has been uploaded the user can view top predictions, download all predictions, and view the probability distribution of kinase-substrate predictions for substrates present in the data set (A). Enrichment analysis returns dot plot showing enrichments among up and down regulated phosphosites. Results from analysis on up regulated phosphosites shown here (B)*

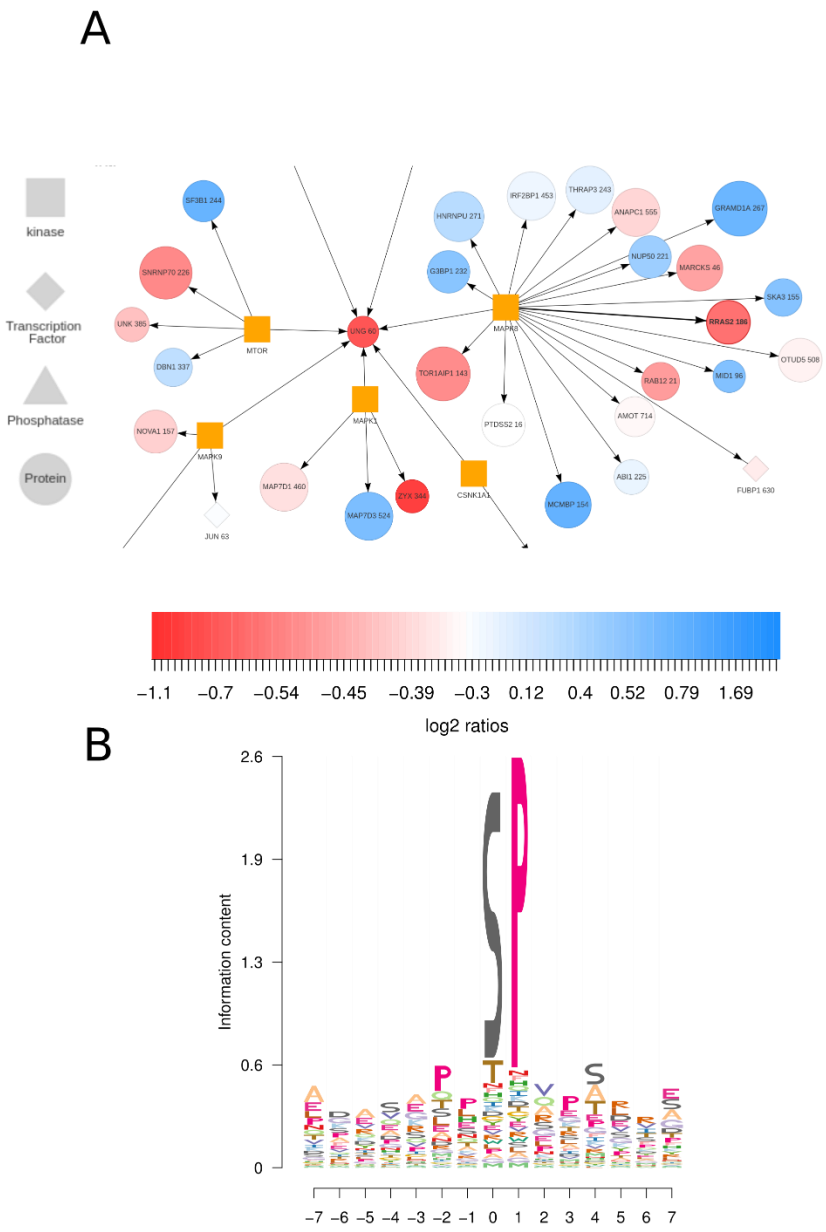

Figure 3.11 *SELPHI2: PCSF fitting and sequence logo. Example output from a SELPHI2 analysis. PCSF is used to identify optimal sub networks obtainable from the phosphoproteomics data (A). Sequence logo for MAPK8 kinase generated from predicted substrates that are included in any of the condition specific sub networks (B).*

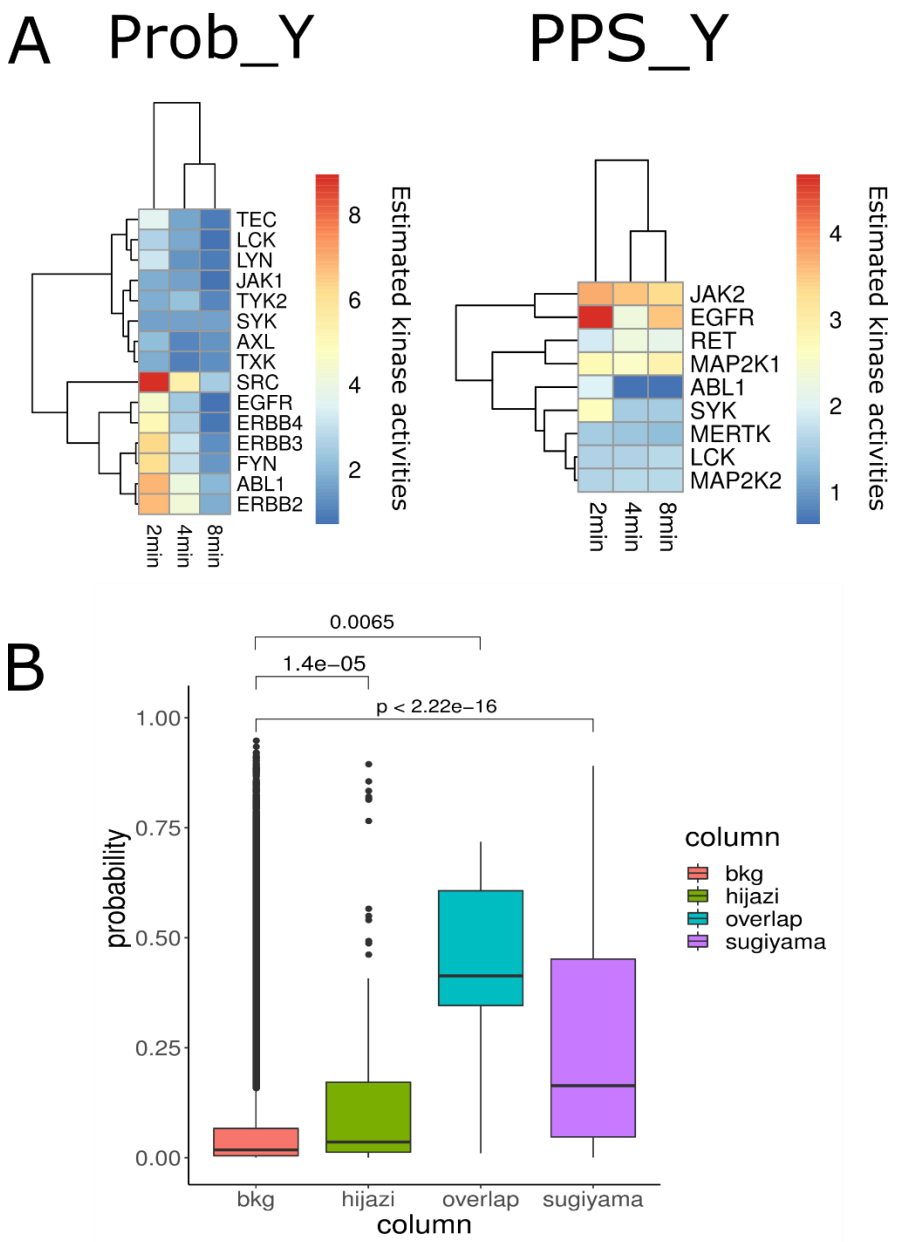

Figure 3.12 *SELPHI2: kinase activities and experimental validation. Kinase activities obtained from the predictions compared to the activities generated from known kinase-substrates. The activity profiles were quite different across the different estimates with low overlap in active tyrosine kinases and no active serine/threonine kinase obtained from the predictions (A). The user can view how well supported predicted kinase-substrates are by two independent experimental kinase-substrate predictions and download those predictions that are corroborated by the two experiments (B).*

3.4 Discussion

In this Chapter I sought to generate kinase-substrate predictions that incorporate kinase specificities as well as high throughput information to predict kinase-substrate relationships as well as the sign of these relationships. While the method described here relies on prior knowledge in the form of position weight matrices, other high throughput measures such as co-phosphorylation and co-expression are integrated as well as other phosphosite specific (Ochoa et al., 2020) features are used. I found that while SELPHI2 is able to predict known kinase-substrates it is also able to capture experimentally predicted kinase-substrates that have not been discovered before. Notably, kinase-substrates that were predicted by two experiments were assigned higher probability compared to kinase-substrates that were predicted by either method. This demonstrates the value of the method in the task of identifying new kinase-substrate relationships. Furthermore, the method is able to make high confidence predictions for less studied kinases and less studied protein substrates when compared with the found in the literature. It can therefore be argued that these predictions provide valuable insight into the less explored parts of the signalling network. While a significant portion of the high confidence predictions can be expected to be hitherto undiscovered relationships, a large part can be reasonably assumed to be false positives due to methodological limitations of kinase-substrate prediction methods, be they computational or experimental. This suggests that despite the value of these predictions for prioritising kinase-substrate relationships and exploring the dark signalling space, by themselves are not sufficient to explore the human cell signalling network architecture as a whole.

In the future, increasing the number of independent high-throughput kinase-substrate experimental datasets could ameliorate this high uncertainty. With our network and literature as it stands, the 24 high confidence kinase-substrate predictions supported by both our method and the two orthogonal experimental studies provide a good starting point for validation of potentially novel components of the human cell signalling network. For instance, of the high probability SELPHI2 predictions, the relationship between CDK2 and NOP2 32 and NUMA1 2000 might be interesting to experimentally validate further given external evidence.

In general, the large accumulation of kinase-substrate relationships that could not be corroborated made further analysis of the network difficult. When the activity profile of kinases derived from SELPHI2 predictions is compared to activities derived from PhosphoSitePlus, it is clear that the difference is quite pronounced. This could either be because of the literature bias in our current knowledge or accumulation of kinase-substrates that do not occur in nature. It must also be pointed out that activity estimates can be noisy as the context specificity of kinase-substrates means that under any given condition only a part of the kinase-substrate relationships are active, even if the kinase is active. The fact that many substrates are targeted by many kinases makes these predictions even harder which in part explains the discrepancy between literatures derived activities and activities derived from kinase-substrate predictions.

Similar issues arose when the similarity of biological function between kinases and their predicted substrates was analysed with kinase-substrates found in the literature having a more similar function to their upstream kinase compared to the predicted substrates. However, it should be noted that due to the large size of the underexplored protein interactome, a large number of proteins might be undiscovered participants in known pathways or partake in pathways that remain undiscovered.

In comparison with five other methods: PhosphoPICK (Patrick et al., 2015), KinomeXplorer (Horn et al., 2014), NetPhos v3.1 (Blom et al., 2004), LinkPhinder (Nováček et al., 2020) and GPS (Wang et al., 2020) we find that SELPHI2 has better performance when it comes to discriminating between unknown kinase-substrate relationships and known ones. Furthermore, SELPHI2 is better at discerning between kinase-substrate relationships that have been independently identified by high throughput experiments. NetPhos performed well both at discerning between known positives and negatives as well as experimentally validated edges. It should be kept in mind though that NetPhos only makes predictions for 17 kinases.

To my knowledge, no other method, neither computational nor experimental, has made predictions on the regulatory status of the predicted kinase-substrate relationships. One challenge is, therefore, that, while cross validation indicates that the predictor effectively discerns between activating and inhibiting relationships, it is difficult to find an

independent set to validate my result and assess its ability to discover new regulatory relationships.

In this Chapter I describe a method to predict kinase-substrate relationships. I find that by incorporating kinase specificity models with high throughput features and features describing potential acceptor phosphosites I am able to capture known kinase-substrate interactions and, crucially, novel experimentally predicted interactions. I find that my method preforms better than other state of the art methods both and capturing known kinase-substrate interactions and experimentally predicted ones. Furthermore, to my knowledge, this is the first method that predicts the sign of kinase-phosphosite interactions which adds to the value of this resource as signs can be incorporated into phosphoproteomic data analysis. These results have also been put together into a new server for SELPHI2 which will be available online once this project has been submitted for peer-review.

4. Towards data-driven modules

4.1 Introduction

Biological systems are often described as being structured in a modular manner. This general principle applies to ecosystems, cell architecture, protein interaction networks and proteins as well as other molecule such as DNA (Lorenz et al., 2011). In the context of signalling, this modularity manifests itself in the modular structure of signalling proteins such as kinases and phosphatases as well as the modular structure of signalling networks (Pawson, 1995). Modularity in signalling networks manifests itself as a network organized into modules of densely interconnected parts of the network with lower density of connections to other parts of the network.

Current models of cell signalling organize the network into pathways, each of which has a certain function or a set of functions as laid out in databases such as KEGG (Kanehisa, 2019), Reactome (Jassal et al., 2020) and WikiPathways (Slenter et al., 2018). While pathways are commonly used in functional analysis of biological data sets and gene sets, there are reasons to believe these static pathways are an overly simplistic representation of the signalling system. In particular, they seem insufficient when it comes to accounting for phosphoproteomic data sets. Olsen and colleagues (Olsen et al., 2006) had found that the signal spreads through the signalling network more widely than one would expect given current knowledge of signalling pathways. Other perturbation studies have reached similar conclusions where perturbation leads to patterns in the phosphoproteomic data sets that differ considerably from the annotated pathways (Köksal et al., 2018).

A part of the reason for these discrepancies is that current literature is biased towards proteins and processes that are already well-studied (Edwards et al., 2011; Luck et al., 2020). More systematic high throughput generation of protein interactome have found that the human protein-protein interactome network is denser with a more equal distribution of edges than the literature indicates (Luck et al., 2020; Rolland et al., 2014). This stresses the need for more data-driven methods to identify signalling modules for unbiased data analysis. Here I describe my work to identify data-driven biological modules and demonstrate their biological robustness and utility. I establish that the modules have properties that should be expected of biological modules such as, In the

case of modules from HL60 and MCF7 cells, proximity between members, correlation with transcription factor and kinase activity and assign traits and functions to the data-driven module. Furthermore, I show that the modules have a greater explanatory power when it comes to independent phosphoproteomic data compared to established pathway modules.

4.2. Methods

4.2.1. Datasets

To maximize the coverage of the data matrix with respect to data available for each peptide, I used 436 phosphoproteomics data sets compiled by David Ochoa and colleagues (Ochoa et al., 2016), extracted from PRIDE and reran with the same parameters through the peptide search protocol. I also used a) a phosphoproteomic data set derived from 50 colorectal cell lines, generated by Roumeliotis and colleagues (Roumeliotis et al., 2017) b) a perturbation data set generated by Hijazi et al. where MCF7, HL60 and NTERA2 cells were grown under different kinase inhibitory conditions (n = 61) with the cells being incubated for 1h under different kinase inhibitor treatments before lysis (Hijazi et al., 2020) c) $\log_2$ transformed and quantile normalized phosphoproteomic and $\log_2$ transformed median normalized expression data from cancer samples and cell lines derived from various sources that had been re-analysed and preprocessed by Abel Sousa and colleagues (Sousa et al., 2021) which included data from 983 samples and cell lines. The data contained data sets from different tissues: brain (Petralia et al., 2020), breast (Koboldt et al., 2012; Lapek et al., 2017; Lawrence et al., 2015; Mertins et al., 2016), colorectal (Cancer Genome Atlas Research Network, 2012; Roumeliotis et al., 2017; Zhang et al., 2014), kidney (Clark et al., 2019), liver (Gao et al., 2019), lung (Gillette et al., 2020), ovary (Network, 2011; Zhang et al., 2016), stomach (Mun et al., 2019) and uterus (Dou et al., 2020) that had been retrieved from Clinical Proteomic Tumour Analysis Consortium (CPTAC) data portal (Edwards et al., 2015). Expression values from cancer cell lines were downloaded from the Cancer Cell line Encyclopedia (Barretina et al., 2012). 180 genome-wide association study (GWAS) SNP

p-values data sets were derived from a DREAM challenge paper on the assessment of biological modules (Choobdar et al., 2019).

4.2.2. Independent component analysis for module extraction

Independent component analysis (ICA) (Jutten and Herault, 1991) divides multivariate signals into additive components assuming they follow a non-Gaussian distribution and are independent of each other. Compared to Principal Component Analysis (PCA), ICA has been found to be able to extract a greater number of biologically relevant components compared to the ICA partly due to the assumption of the PCA that the data is multivariate Gaussian (Lee and Batzoglou, 2003). Already the independent component has been applied to solve biological problems such as to extract transcriptomics modules (Zhou and Altman, 2018) and analyse cancer modules (Sompairac et al., 2019). For this project, I used the ICA as implemented by *MineICA* (Biton et al., 2021). The function *clusterFastICARuns* was used to extract the phosphorylation modules. The function runs the FastICA (Hyvärinen, 1999) algorithm repeatedly with random initialization and then clusters the resulting components and returns the medoids of the resulting clusters as component estimates using the I_q cluster quality index (Himberg et al., 2004) defined as:

$$I_q(C_m) = \frac{1}{|C_m|^2} \sum_{i,j \in C_m} \sigma_{i,j} - \frac{1}{|C_m||C_{-m}|} \sum_{i \in C_m} \sum_{j \in C_{-m}} \sigma_{i,j}$$

Where C_m represents the indices that are part of cluster m, C_{-m} is the set of indices that are not part of cluster m and $|C_m|$ is the number of items in cluster m and $\sigma_{i,j}$ represent the similarities between components as measured by absolute of the mutual correlation coefficients between them. An I_q value of 1 indicates perfect clustering while lower values indicate that the cluster is less isolated and compact.

The cluster quality index was used to assess the quality of the clustering. The final step entails clustering the components with hierarchical clustering and the centrotypes of each cluster are used as estimates for the components.

A large number of independent components have been shown to yield biologically meaningful results (Kairov et al., 2017). Therefore, in this project, in order to extract as much biological information as possible, I extracted m=61 number of clusters as 61 is the number of samples in the different data set. Hierarchical clustering as incorporated into

the *MineICA* package was used to cluster component outcomes and *alg.type* was set to parallel.

4.2.3. Pathway enrichment of phospho modules

To assign biological processes to data-driven modules, I conducted a pathway enrichment on data-driven modules. I used Fisher's exact test (Fisher, 1935) to calculate enrichment odds ratios against the background of the Reactome (Jassal et al., 2020) pathways database. For each module the resulting p-values were adjusted with the Benjamini and Hochberg (Benjamini and Hochberg, 1995) method and finally when the results from all the different modules are pooled together I adjusted the p-values further with the Benjamini and Hochberg method. To define the background, I used proteins that were measured in the three data sets as NTERA2 is missing from the Cancer Cell line encyclopedia. This is done to avoid any enrichment that is cell type-specific.

4.2.4 Calculating distances between proteins in the interaction network

Proteins included in any of the extracted modules were mapped onto the IntAct network (Orchard et al., 2014) (Downloaded 15th May 2019). All distances between the proteins in the modules were calculated by using the *distance*() function in the *igraph* package (Csardi and Nepusz, 2006). An unweighted breadth-first search algorithm was used to identify the shortest distance between any two proteins in an undirected network. Wilcoxon rank sum test (H. B. Mann and D. R. Whitney, 1947) was used to establish if the distance between proteins within modules was closer than protein pairs where the proteins belonged to different modules.

4.2.5. Extracting modules from literature network

To compare my data-driven modules to modules extracted from the literature network I downloaded the phospho-signalling network from OmniPath (Turei et al., 2016). OmniPath only contains signalling networks and is therefore attractive for our purposes as it leaves out interactions that are involved in non-signalling interactions.

To identify modules within the network I used the DEMON (Coscia et al., 2012) network clustering algorithm as implemented in the python library Karate Club (Rozemberczki et al., 2020). DEMON uses the 'local first' method to find overlapping communities within networks. For the minimum community size, I used ten. I tried various merging thresholds from 0 to 1 eventually settling for 0.8 as it performed best in the enrichment of independent data sets (See chapter 4.2.6).

4.2.6. Enrichment for independent data sets

To evaluate the ability of the data-driven modules to describe independent datasets I used a compilation of 436 of different phosphoproteomics data sets compiled by David Ochoa and colleagues (Ochoa et al., 2016). This was done to ensure that the $\log_2$ ratios followed comparable distribution.

When enrichment analysis is conducted it is important to select an appropriate background. Problems can arise when the whole genome is used as a background as in many cases the top enrichments may simply be tissue/sample related terms while in our case I am interested in processes that are activated in response to a perturbation or stimulation. To counter this, I used as background expression data extracted the Cancer Cell line Encyclopedia (Barretina et al., 2012) corresponding to the relevant cell lines that were used in each experiment. I used rpkm 0.5 as a threshold for expression. Due to the fact that not all data sets were extracted from cell lines, I ended up using 270 data sets for the final analysis.

For the enrichment analysis the $\log_2$ ratios greater than 1 were used to indicate up-regulated phosphosites. Subsequently, a Fisher's test was used to calculate overrepresentation of module members. Fisher's exact test was chosen since in many cases the calculations are done on a relatively low number of proteins in which case this test has been found to be effective.

Enrichment (odd ratio) values were calculated for three modules sets: a) data-driven modules, b) pathways in Reactome found at all levels in the hierarchy (Jassal et al., 2020) and c) modules extracted from the OmniPath network (Turei et al., 2016). For each condition and each module set, I corrected the resulting p-values with the Benjamini Hochberg (Benjamini and Hochberg, 1995) method. I pooled the results together and

applied Bonferroni multiple testing correction (Bonferroni, 1936). Subsequently, a p-value cutoff of 0.01 was applied.

To compare the module sets I looked at the number of significant enrichment as well as the odd ratios achieved with the different module sets. To ascertain if the difference was significant, Wilcoxon's rank sum test was used (H. B. Mann and D. R. Whitney, 1947).

4.2.7. Comparison between modules extracted from different data sets

In order to look at the similarities and context specificities, I analysed the clustering assignment across the different data sets, I quantified the similarities between clustering assignments. Since different phosphosites are included in the different modules, only the overlap between the modules were analysed, that is, phosphosites that were included in both clustering assignments. As our clusters are overlapping I used the geometric accuracy metric to assess overlap as implemented function geometric_accuracy() y (Nepusz et al., 2012; Palla et al., 2005) from the python module CluSim (Gates and Ahn, 2019) was used to assess the similarity between the cluster assignments between the different data sets.

To calculate a p-value for each of the measurements I generated a distribution of geometric accuracies by randomly assigning phosphosites to the same number of clusters as in the original assignment. This was repeated a hundred times and the geometric accuracy was calculated between the random assignments. The similarity between the initial assignments was then ranked compared to the random distribution to derive empirical p-values.

Furthermore, I did pairwise comparisons of the clustering assignments across cell lines by calculating the Jaccard Index (Jaccard, 1912) for each cluster pair. The Jaccard index is calculated as follows:

$$Jacc(c_{nx}, c_{my}) = \frac{|A \wedge B|}{|A \vee B|}$$

Where A is the set of phosphosites belonging to c_{nx} (cluster no. n in cell line x) and B is the set of phosphosites belonging to c_{my} (cluster no. m in cell line y)

4.2.8. Association of modules with kinase and transcription factor activities

Next wanted to see if modules correlated with kinase or transcription factor activities across different samples. I used a phosphoproteomics data compilation (Sousa et al., 2021) from the following tissues: brain (Petralia et al., 2020), breast (Koboldt et al., 2012; Lapek et al., 2017; Lawrence et al., 2015; Mertins et al., 2016), colorectal (Cancer Genome Atlas Research Network, 2012; Roumeliotis et al., 2017; Zhang et al., 2014), kidney (Clark et al., 2019), liver (Gao et al., 2019), lung (Gillette et al., 2020), ovary (Cancer Genome Atlas Research Network, 2011; Zhang et al., 2016), stomach (Mun et al., 2019) and uterus (Dou et al., 2020) that had been retrieved from Clinical Proteomic Tumour Analysis Consortium (CPTAC) data portal (Edwards et al., 2015). Kinase activities were calculated with the batch_kinase_predictions() function which calculates activities for all kinases with substrates in the data set as implemented previously in the *KSEA* package (Ochoa et al., 2016). I used kinase-substrate relationships from PhosphoSitePlus and the number of trials run for p-value generation was set to 1000. The resulting p-values were $\log_{10}$ transformed and the sign of the mean $\log_2$ ratio of the kinase's substrate was used to put a sign on the log transformed p-value. A constant of 10^{-6} was added to avoid log transforming zero values.

For transcription factor activities I used the corresponding transcriptomics datasets (Sousa et al., 2021). The *viper* (Alvarez et al., 2016) R package was used to calculate transcription factor activities. A list of transcription factors and their targets was obtained from the *Dorothea* R package (Garcia-Alonso et al., 2019). Interactions with confidence levels of A or B of A-F were retained. A indicates that there are two or more curated resources supporting the TF-target pair or if there are four or more evidences for the interactions. Furthermore, if the interaction is found in a review or the curated interaction is signed and has other evidence it counts as class A. Meanwhile class B indicates likely confidence, which includes:

I. Interactions found in curated databases and supported by CHIP-seq studies.

II. Interactions found in curated databases and further supported by transcription factor binding models and inferred by reverse engineering tissue specific networks from GTEx (GTEx Consortium, 2013).
III. Interactions supported by CHIPseq, GTEx and transcription factor models are also retained.

In order to assess how strongly data-driven modules were associated with distinct transcription factor and kinase activities compared to Reactome pathways (Jassal et al., 2020), I used Fisher's exact test to conduct a pathway enrichment with the data-driven modules and Reactome pathways. This enrichment analysis was conducted on every sample. Up regulated phosphosites were defined as phosphosites with $\log_2$ ratio > 1 in each sample. Subsequently, I correlated the resulting odds ratios for the samples with the activities using Spearman's *rho*.

4.2.9. Association of traits to modules with GWAS

I used the Pascal (Lamparter et al., 2016) software in order to associate modules with variants that have previously been associated with various traits. The Pascal method has previously been developed to score genes and gene modules with SNP summary statistics. Pascal Z-scores are calculated from the number of SNPs associated with a gene assuming that the number of SNPs associated with a gene are normally distributed. To score a pathway, a fusion gene is created which contains all the SNPs for all the genes belonging to the module.

Here I used the chi-squared method. To derive a pathway score the gene level p-values are ranked and divided by the number of genes plus one (n+1) to generate a uniform distribution. This distribution is then converted to a chi-squared distribution and the sum of the m genes belonging to the pathway is compared and tested against the chi-squared score of modules of size m.

I associated our modules with SNP linked to various traits. I used 180 GWAS data sets used earlier to evaluate DREAM challenge submissions (Choobdar et al., 2019). For comparison I scored the data-driven modules and compared them with the scores achieved by the Reactome pathways. To associate traits with Reactome pathways all proteins that participate in a Reactome (Jassal et al., 2020) pathway were used as the

background. The background proteins I used for the data-driven module analyses were all proteins identified in the three phosphoproteomic data sets from MCF7, NTERA2 and HL60. The results from the Reactome and data-driven module associations were pooled together and the p-value was adjusted with the Benjamini Hochberg method (Benjamini and Hochberg, 1995).

4.2.10. Predicting pathway co-occurrence with machine learning

I then set out to see if our modules could be used to predict co-occurrence within pathways for pairs of phosphosites. These predictions were then compared with those made using co-phosphorylation, co-expression and finally if merging of these predictors could improve predictions of pairs of proteins participating in a similar function like they do in pathways. In order to select phosphosites for this analysis I used a set of phosphosites that had previously been assigned a functional score and picked phosphosites with functional score higher than 95% of the phosphosites (Ochoa et al., 2020).

For this purpose I downloaded pathways from KEGG (Kanehisa, 2019), Reactome (Jassal et al., 2020) and WikiPathways (Slenter et al., 2018). Since pathways, in general, may include processes that do not involve phosphorylation in any way, I focused on signalling pathways for subsequent predictions. Therefore, I only included Reactome pathways that were included in the subtree rooted in signalling pathways and for KEGG I included pathways whose entry started with hsa04. All pathways that did not include a kinase were filtered out. To represent module or pathway co-occurrence, I gave phosphosites found on a protein that belongs to the same pathway a value of one while the rest were given a value of zero. The same was done for the modules, that is, module co-occurrence was determined at the protein level.

The training set was generated as follows: all phosphosites found on proteins that are known to partake in the same process were used as a positive set while a random set of phosphosite pairs was used as a negative set. This rests upon the assumption that most proteins do not co-occur in pathways. As a result of this sampling of negatives, a hundred

models were generated and I assigned the final probability score by averaging across the 100 runs.

Other predictors I used in this exercise were co-expression between genes across tissues and cell lines. Data from GTEx (GTEx Consortium, 2013) and The Protein Atlas were used (Thul et al., 2017; Uhlén et al., 2015), as well as co-phosphorylation between phosphosites from cancer cell line data set generated earlier (Roumeliotis et al., 2017). In all cases Spearman's *rho* was used to quantify co-expression and co-phosphorylation. Random forest (Tin Kam Ho, 1995) as implemented by the *scikit-learn* (Pedregosa et al., 2011) python library was used to predict probability of co-occurrence in pathways. I used Grid to find optimal parameters. The same method and parameters were considered as in Chapter 3 (See 3.2.5).The model with the best AUC derived from ten-fold cross validation was used for each of the one hundred runs.

4.2.11. Comparison between model performances based on predictor combinations

To compare the predictive power of each predictor I looked at the performance of the modules, the co-expression and co-phosphorylation and the combination of these predictors. I looked at the intersection of phosphosite pairs between the module, co-phospho and the co-expression and looked at the relative performance as measured by the area under the receiver operator curves. Furthermore, I looked at the combination of all predictors and the combination of co-expression and co-phosphorylation.

4.3. Results

4.3.1. Properties of data-driven modules

To derive data-driven modules phosphoproteomic data generated by a recent publication from three cell lines was used (Hijazi et al., 2020). Altogether, the three perturbation data sets contained information on 21,393 phosphosites. For MCF7 the number was 9,654 phosphosites and NTERA2 and HL60 had measurements assigned to 17,577 and 16,756 phosphosites respectively. The overlap between the three sets can be seen below (Figure 4.1 A.). ICA does not assign a module to each and every phosphosite, rather it extracts subcomponents of a multivariate signal. Here I only extracted variables that contribute the most to each component that is variables whose contribution is greater than 3 standard deviations above the mean. Therefore, not all phosphosites get assigned to a cluster and are therefore not included in subsequent analysis. Furthermore, unlike methods like K-means clustering, modules can overlap. These properties better reflect biological reality as not all phosphosites can be expected to participate in biological processes related to cell signalling and biological modules do indeed overlap. In the case of MCF7, ICA assigned 3144 phosphosites to 61 modules. In the case of HL60 and NTERA2, 5863 and 6952 were assigned to modules respectively indicating that a minority of measured phosphosite contribute to any sub-component. The median, minimum and maximum size of the modules extracted from the three data sets can be seen in **Table 4.1.**

Table 4.1: *Overview over data-driven modules derived from three data sets: NTERA2, HL60 and MCF7.*

Set	Minimum module size	Median module size	Maximum module size
MCF7	104	145	195
NTERA2	235	270	709
HL60	220	269	346

I found that each component was quite dissimilar to all the other components. At the same time each phosphosite was assigned to more than one cluster on average with the median number of clusters per phosphosites being 2 in each data set. The median and maximum number of clusters per phosphosites can be shown in **Table 4.2**.

Table 4.2: *Median number of cluster assignment/phosphosite.*

Data set	Median number of clusters/phosphosites	Maximum number of clusters/phosphosite
NTERA2	2	17
MCF7	2	15
HL60	2	19

Generally speaking, each cell type had a different set of phosphosites contributing to the multivariate signal. In fact, a lower portion of phosphosites assigned to a cluster overlapped across cell types compared to the phosphosites as a whole (Figure 4.1 B).

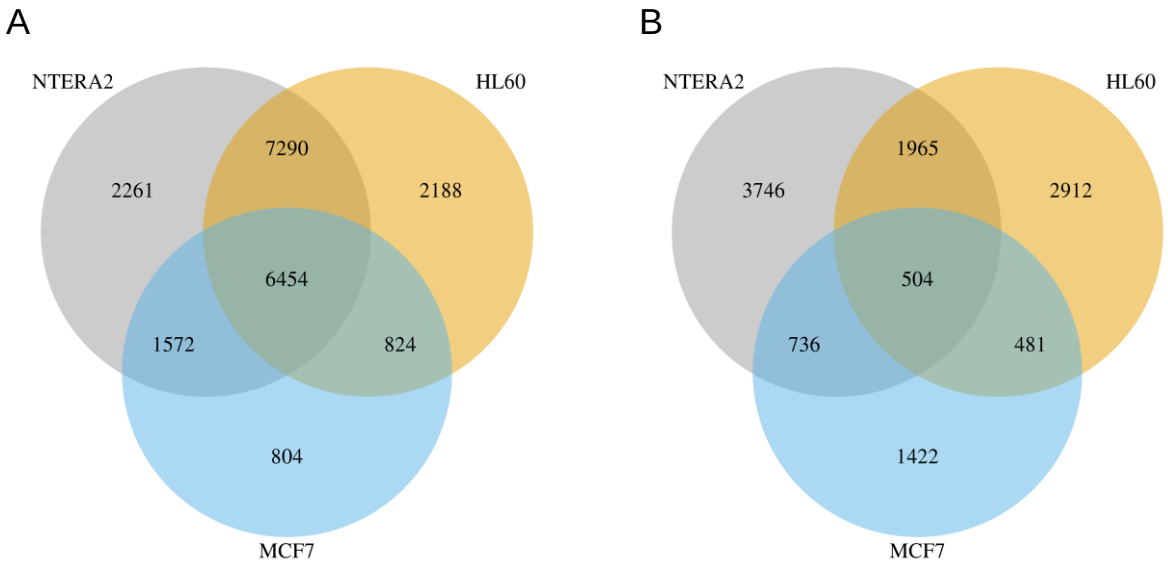

Figure 4.1 *Venn diagram showing phosphosites from different datasets: MCF7, NTERA2 and HL60 and assignments of phosphosites to clusters (A). The overlap between measured phosphosites the three data sets: MCF7, NTERA2 and HL60 (B) Venn diagram showing the overlap between phosphosites assigned to a cluster.*

One would expect interacting proteins to cluster together since phosphorylation of phospho-regulated proteins that interact should in theory correlate. Therefore, I set out to see if known interactions and protein complexes were overrepresented in our modules. In terms of known interactions, I found that protein pairs representing known interactions (Orchard et al., 2014) were over-represented within our modules when compared to across module protein pairs (X-squared = 1573.4, p-value < $2.2\ 10^{-16}$). Likewise, proteins belonging to the same complex (Jassal et al., 2020) had a greater tendency to cluster together (X-squared = 172.9, p-value < $2.2\ 10^{-16}$)

4.3.2. Pathway enrichment of derived modules

Enrichments were conducted for each cell type in three different analyses and the results were subsequently pooled together and p-values adjusted with the Benjamini Hochberg method (Benjamini and Hochberg, 1995). In all, 13,905 significant enrichments were identified for 183 modules with 1142 different pathways yielding a significant hit. A complete list of enrichments for these modules can be seen found uploaded at (https://gitlab.ebi.ac.uk/borgthor/Borgthor_Petursson_EBI_CAM_thesis/-/tree/master). It is therefore clear that all generated modules can be assigned a biological function in terms of annotated pathways. While some previous results that indicate that results from phosphoproteomic analysis do not neatly map onto our current knowledge of signalling processes (Humphrey et al., 2015; Köksal et al., 2018; Olsen et al., 2006) these results indicate that there is at least an overlap between these modules and the literature derived ones.

4.3.3. Distance between proteins within the same module compared to distance across modules

Modules are usually thought of as components of the signalling systems that are relatively highly interconnected compared to other parts of the network and insulated from the rest of the network. Therefore, it stands to reason, that proteins that cluster together in a phospho-signalling module should be closer to each other in the network than the rest.

To assess if this is true, I mapped the protein on which the phosphosites were found onto IntAct (Orchard et al., 2014) and calculated the distance between them and compared the distance between proteins within the same cluster and the distance between proteins that do share a module. I found that proteins belonging to the same module were significantly closer to each other compared to the background in all cases except for NTERA2 (MCF7: $W = 2.9 \times 10^{11}$ p-value $< 2.2 \times 10^{-16}$, HL60 : $W = 2.9 \times 10^{12}$, p-value $< 2.2 \times 10^{-16}$, NTERA2 : $W = 4.3 \times 10^{12}$ p-value =0.8). It should be noted though that the trend was very slight with most proteins being relatively close to each other with median distance being 4 edges in both sets (Figure 4.2).

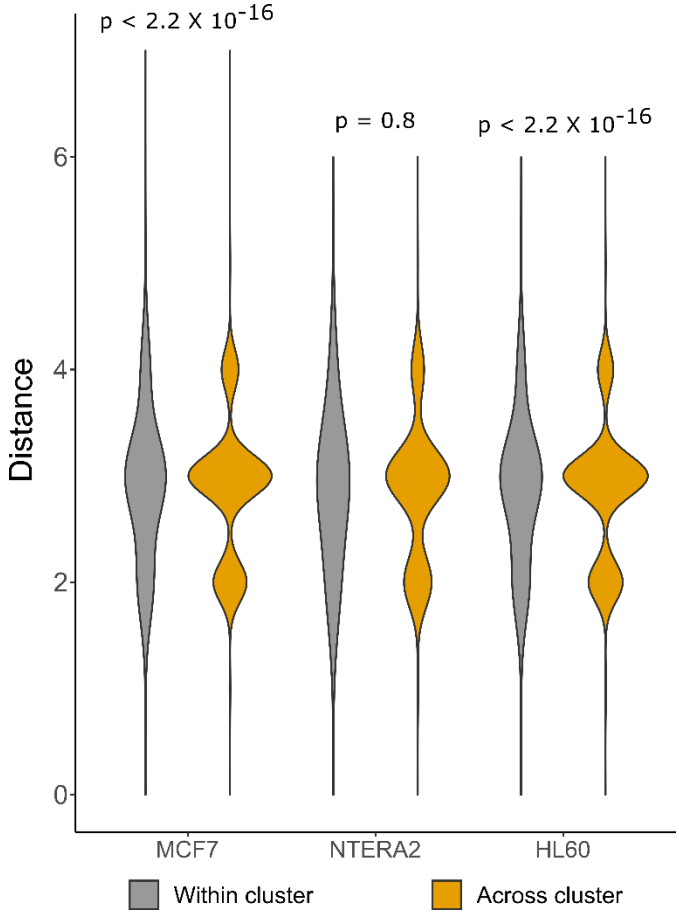

Figure 4.2 *Distances between proteins clustering together and across clusters. While protein pairs that are within the same module tend to be closer to each other, the difference is small.*

4.3.4. Similarity between cluster assignment across the three different data sets

While signalling is context specific one could make the assumption that some core processes are conserved across cell types and conditions. To test this, the similarity between the clustering assignments was assessed across the three large data sets: MCF7, HL60 and NTERA2. Other data sets were not considered due to low overlap in measured phosphosites. In order to assess the similarity between the clustering assignments I used geometric accuracy (GA) since the modules do overlap. GA returns a value between 0 and 1, 1 indicating that the two variables are mutually dependent while 0 indicates no dependence between the variables. Clustering assignments of phosphosites that contributed significantly to a component (were assigned a cluster) in both datasets were compared. In all cases the mutual information was fairly low but in all cases the cluster assignment had higher GA than a set of 100 randomized clustering assignments. The results can be seen in **Table 4.3** below.

Table 4.3: *Similarity of clustering assignments across data sets. In all cases the NMI score is low but significantly higher than NMI achieved by comparing randomized clustering assignments*

Data set comparison	GA	$GA_{randomized}$ (sd)	$p_{empirical}$
MCF7, NTERA2	0.11	0.081 (0.0016)	0.00
NTERA2, HL60	0.093	0.060 (0.00086)	0.00
HL60, MCF7	0.101	0.085 (0.0015)	0.00

In general, these results seem to indicate that the signalling modules drawn from these three different data sets differ to a great extent but less than would be expected by random. I did also explore the pairwise overlap between the clusters as measured by the Jaccard Index (Jaccard, 1912) (Figure 4.3). Overall, the similarity was quite low across all comparisons. The cluster overlap between NTERA2 and HL60, as an average maximum JI for each cluster was 0.070 while module 8 (HL60) and 61 (NTERA2) with JI

0.24 of their phosphosites, which was the highest overlap between the two clustering assignments. For HL60 and MCF7 the corresponding values are 0.088 and 0.16 and 0.085 and 0.22 between MCF7 and NTERA2.

4.3.5. Enrichment of independent data sets with data-driven modules

To compare the data-driven modules with established pathways I conducted an enrichment analysis on 270 independent perturbation phosphoproteomic data sets using the data-driven modules, literature network derived modules and Reactome pathways as a set for comparison. Each module set had a different number of modules and therefore different numbers of hypotheses were tested. The number of hypotheses tested are listed in **Table 4.4.**

Table 4.4*: Number of hypotheses tested and significant hits by module enrichment. List of module sets and the number of enrichment hypotheses tested.*

Module set	Number of hypotheses tested	Number of significant module enrichment
Reactome Pathways, all levels	296666	40
Data-driven modules	49410	9477
Network modules	192188	96

Due to the difference in number of modules, the results were pooled, p-values were Bonferroni corrected and p-values with adjusted value lower than 0.01 were selected. As can be seen in Table 4.4, the data-driven modules yield by far the greatest number of significantly enriched modules, both as a portion of hypotheses tested and in absolute number. These results indicate that generally speaking a large portion of modules are core modules and active across different conditions and due to the interconnectivity of the network, the neat compartmentalized modules as laid out in the Reactome pathway database fail to capture these patterns. Notably, OmniPath clusters generated a greater number of significant hits than Reactome, which suggest that algorithmic methods to

extract modules from the current state of the art knowledge are better suited at capturing biological patterns than the predefined modules present in the data bases.

Among the significant enrichment hits the data-driven modules yielded the highest odds ratio with a median odds ratio of 6.25 while the corresponding values for the network modules and the Reactome pathways was 3.56 and 3.49. The difference between the odds ratios returned by the data-driven modules were significantly higher than the two literature derived sets (OmniPath: $W = 1.9 \times 10^5$ p-value $< 2.2 \times 10^{-16}$, OmniPath network modules: (Reactome: $W = 7.1 \times 10^5$ p-value $< 7.2 \times 10^{-12}$, Wilcoxon rank sum test) while the difference between the OmniPath network modules and Reactome pathways was not significant. These results can be seen in Figure 4.3.

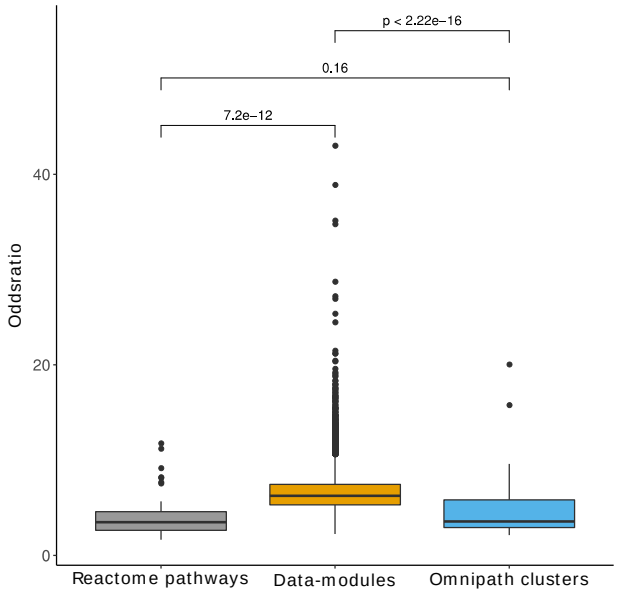

Figure 4.3 *Odds ratios from enrichment analysis from three different module sets: Data-driven modules, Reactome pathways and literature network clusters.The data-driven modules had the greatest performance in terms of odds ratios compared with the literature derived modules.*

Combined with the greater number of significant hits, these results seem to indicate that the databases are insufficient when it comes to capturing phospho-signalling patterns. This is likely due to the incompleteness of our knowledge. Technical limitations of mass

spectrometry data and in particular the enrichment of phosphopeptides could also explain part of the issue as in general, mass spectrometry only captures part of the cellular proteome with data missingness being non-random which introduces biases of its own (Chapter 1.8.3).

4.3.6. Association of phospho-signalling modules with transcription factors modules

Transcription factors regulate gene expression and are known to be regulated by signalling cascades. Due to this link to signalling, I set out to correlate module activity with the activity of transcription factors in matching proteomics and gene expression data sets in order to assign a biological function to the data-driven modules. Module activity was quantified by calculating the modules' overrepresentation among up regulated phosphosites.

For comparison, the corresponding odds ratios were calculated for the Reactome pathways which were also correlated with the transcription factor activities. The data used was retrieved from many different tissues and therefore tissue specific associations were also investigated as well as cross-tissue associations. To assign a transcription factor to each module, I applied a p-value threshold of 0.05 and, subsequently, I looked at the top correlation between the module odds ratios across samples and transcription factor activities (hereby referred to as the top correlation). I found that data-driven modules had higher association with its top transcription factor (Figure 4.4 A). The median top correlation for the data-driven modules was Spearman's *rho* of 0.37 while for the Reactome pathways is the coefficient was 0.26. The difference between the two sets of top correlations was significant ($W = 2.3 \times 10^5$, p-value $< 2.2 \times 10^{-16}$, Wilcoxon rank sum test). Similar results were produced on the tissue level with data-driven modules having the median top correlation of *rho* = 0.34 while Reactome pathways had the median top correlation of *rho* = 0.29, with the difference between them being significant ($W = 1.1 \times 10^7$, p-value $< 2.2 \times 10^{-16}$, Wilcoxon rank sum test). However, these results differ depending on tissue, while data-driven modules yield significantly higher top correlation in brain, lung, liver, stomach, and uterus the difference is non-significant in breast,

colorectal, kidney and ovary. The results can be seen on a per tissue basis in Figure 4.4 B.

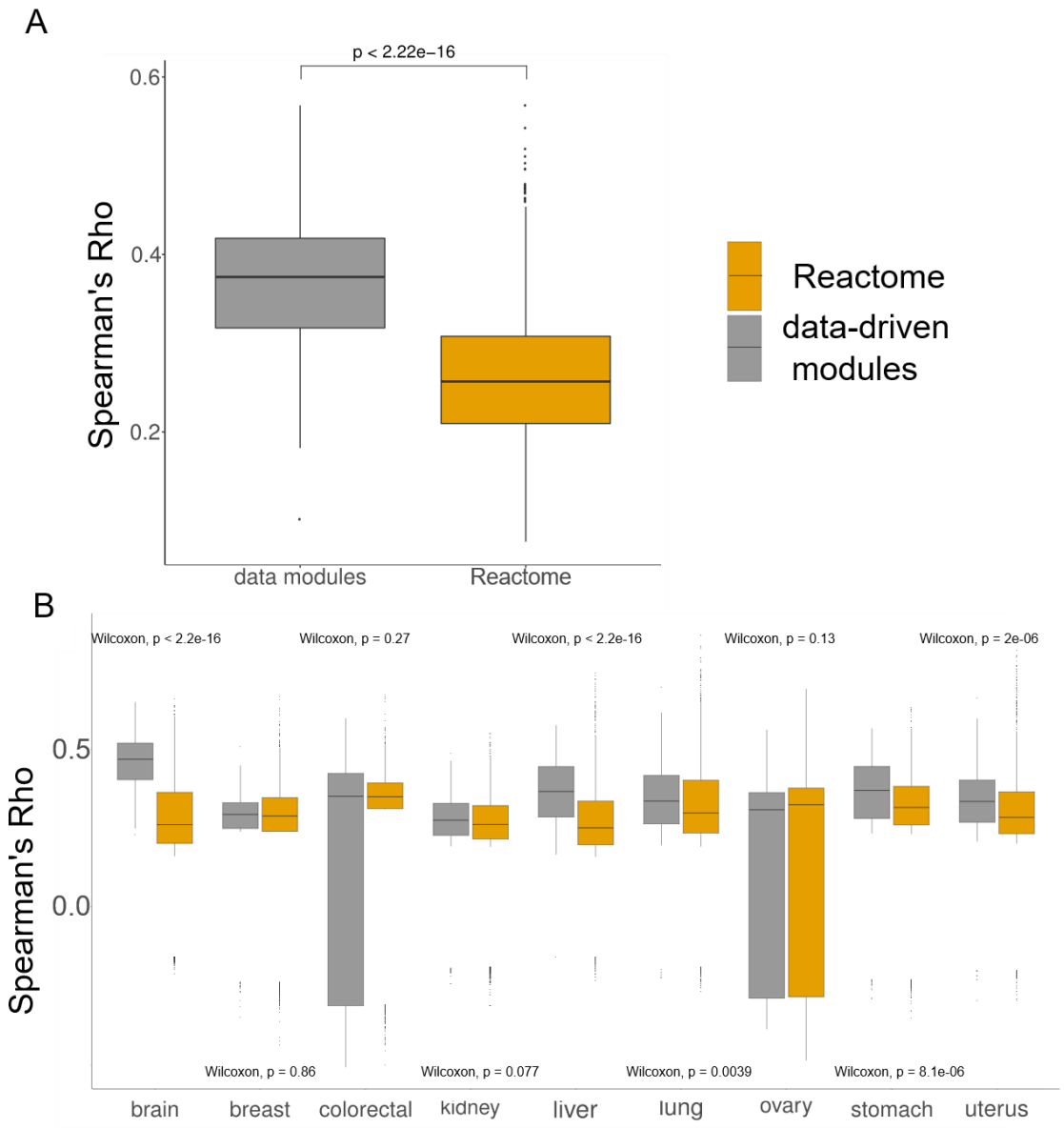

Figure 4.4 *Association between data-driven modules and transcription factors. The top correlation between module enrichment odds ratios and transcription factor activities. Data-driven modules had higher top correlation compared to Reactome derived modules. This was true across tissues while on a per-tissue basis the results vary.*

Interestingly, these results suggest that data-driven phosphorylation modules seem to better capture transcription factor activities than pre-defined Reactome pathways, even though the transcription factor modules are defined by the literature network. The modules are listed with their best matching transcription factor in **Appendix 4.1**. Results on a per tissue basis can be found at: (https://gitlab.ebi.ac.uk/borgthor/Borgthor_Petursson_EBI_CAM_thesis/-/tree/master)

4.3.7. Association of phospho-signalling modules with kinase activities

Similarly to the transcription factor correlation, I wanted to evaluate if a similar pattern emerged when kinase activities were correlated with module enrichment values by using the same phosphoproteomics data set as before. Kinases phosphorylate proteins as part of signal transduction networks so it stands to reason that kinase activities should correlate with the activity of signalling modules. Same as before, I applied p-value threshold of 0.05 and looked at the top correlation for each module. There was a similar pattern that emerged when kinase activities and module odds ratios were correlated across tissues with the data-driven modules having a median top-correlation of 0.30, while Reactome yielded the corresponding value was 0.21 and the difference between the two set was significant ($W= 2.1 \times 10^5$, p-value $< 2.2 \times 10^{-16}$,Wilcoxon rank sum test) (Figure 4.5 A). Across tissues the results vary. Data-driven modules have higher top correlation in brain, colorectal tissue, kidney, lung and ovary, while the Reactome pathways achieved better results in the liver. The rest of the tissues had no significant difference in association between module and kinase. Overall data-driven module-kinase associations across the tissues yielded significantly higher top correlations (*rho* = 0.34) than the Reactome pathways (*rho* = 0.29) ($W= 7.3 \times 10^6$, p-value $< 2.2 \times 10^{-16}$, Wilcoxon rank sum test) (Figure 4.5 B).

This analysis yielded similar results to transcription factors and is surprising since kinase activities are defined by known kinase-substrates. The modules are listed with their best matching kinase in **Appendix 4.2.** The results for the tissue specific analysis can be found at: (https://gitlab.ebi.ac.uk/borgthor/Borgthor_Petursson_EBI_CAM_thesis/-/tree/master)

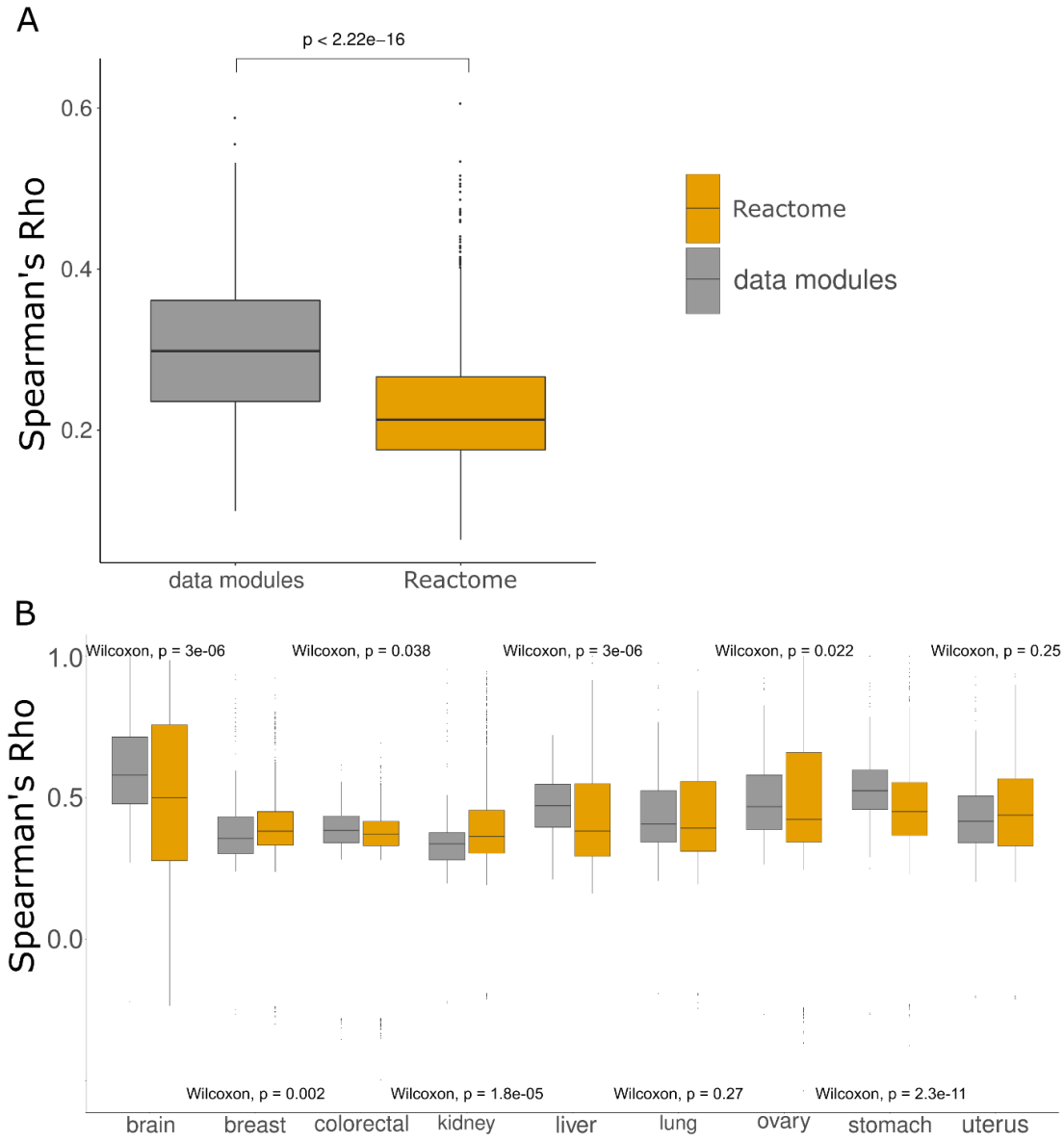

Figure 4.5 *Association between data-driven modules and transcription kinases.The top correlation between module enrichment odds ratios and kinase activities. Data-driven modules had higher top correlation compared to Reactome derived modules (A). The results varied across tissues (B).*

4.3.8. GWAS association of phospho-signalling modules with traits and diseases

One way of linking modules with biological function is to establish whether certain gene variants that have previously been found associated with traits are overrepresented among the proteins present in each module. To this aim I used the Pascal (Lamparter et al., 2016) software to associate variants to each module. I associated traits from 180 SNP-trait association data sets that had previously been used in a DREAM challenge (Choobdar et al., 2019) with the data-driven modules and Reactome pathways. Since my module set and Reactome have vastly different numbers of modules the resulting associations were pooled and the p-values are adjusted with the Benjamin Hochberg method. In all, 183 modules were tested and 21 had at least one trait significantly associated with it (11%). In all 5 traits were assigned to the 21 modules with 21 module-trait associations found overall. A complete list of significant data-driven module trait associations can be seen in **Appendix 4.3**.

For comparison, traits were associated with the Reactome pathways where 2,236 modules were enriched for different traits. 401 significant associations were found between 113 (5 %) pathways and 26 traits. In both cases there was a considerable overlap in trait-module association that is many traits were associated with multiple modules.

While a larger portion of the data-driven modules had significant trait association the Reactome pathways had a greater overall number of associations. Furthermore, the Reactome-trait associations generally had a lower assigned p-value (median = 0.015) compared with the data-driven modules (median = 0.023). However, the difference was not significant.

4.3.9. Use of modules and high throughput data to predict pathway co-membership

Next, I set out to analyse how well the data-driven modules correspond to known signalling pathways and compare this with predictions based on high throughput data. For this aim, I looked at co-phosphorylation from cell line-based phosphoproteomic data

sets (Roumeliotis et al., 2017), co-expression, phosphorylation and co-occurrence in data-driven modules. I found that that co-occurrence in modules predicts pathway co-membership poorly (AUC=0.52). Both co-expression and co-phosphorylation perform better at predicting pathway membership than random with AUC of 0.61 for phosphorylation and 0.93 for co-expression. Interestingly, phosphorylation seems to perform worse than co-expression. Furthermore, I do not find that basing predictions on phosphorylation and co-expression improves predictive performance, and adding the module membership features does not improve predictive performance either. The combination of modules and phosphorylation, however, performs better than either feature (Figure 4.6 A).

Looking at high probability phosphosite pairs (probability > 0.5) from each of the runs, I found that predictions based on expression included a large portion of high probability pairs found by the different feature combinations are not annotated in any of the pathway databases used, while high confidence predictions based on phosphorylation included a relatively high portion of unknown protein pairs. **Table 4.5** lists the number of high probability pairs predicted by each combination and the percentage of those who are unknown.

Table 4.5: *Number of high confidence phosphosite pairs per feature set***.** *Number of high-confidence phosphosite pairs per predictor used for prediction and how many of these pairs are not annotated in the data bases.*

Predictors	No. high confidence predictions	No. unknown high confidence predictions	Percentage unknown
Co-phosphorylation	11586	7553	85
Co-expression	8634	4525	47
Co-phosphorylation & co-expression	7990	3875	49
Co-phosphorylation, co-expression & modules	7767	3653	43

Interestingly, the two sets of undiscovered high probability pairs have a relatively low overlap, that is phosphoproteomics and RNAseq data seem to capture different sets of novel protein pairs associated with the same signalling pathway (Figure 4.6 C). High probability phosphosite interactions by each feature combination can be accessed from **(**https://gitlab.ebi.ac.uk/borgthor/Borgthor_Petursson_EBI_CAM_thesis/-/tree/master**)**.

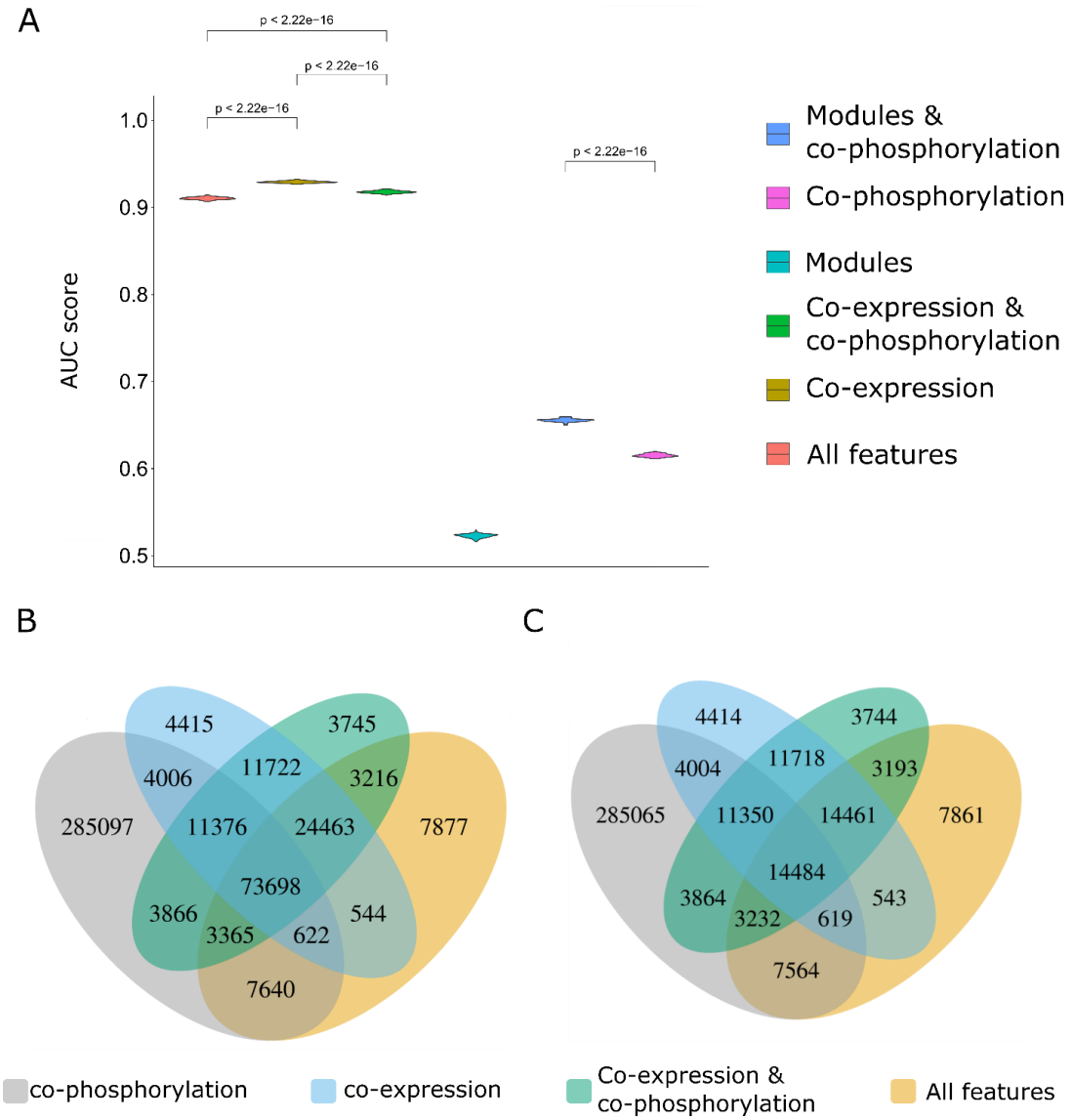

Figure 4.6 *Predictive performance of different features and feature combinations. The features were used to predict signalling pathway co-membership of phosphosites. I find that co-expression outperforms the other features while the data-driven modules do not predict signalling pathway membership by themselves they improve prediction when combined with other features (A). High probability predictions of co-membership differ across the different feature sets with co-phosphorylation differing the most from the other feature sets and with a higher portion of high probability co-memberships that were previously unknown (B). This is particularly the case when previously known co-memberships were excluded (C).*

4.4 Discussion

Various studies have found that pathways as defined in the different data bases such as KEGG (Kanehisa, 2019) or Reactome (Jassal et al., 2020) have been found to have limited value in terms of explanatory power in phosphoproteomic studies (Humphrey et al., 2015; Köksal et al., 2018; Olsen et al., 2006). This is partly due to the fact that the signal disperses quite fast and widely throughout the signalling network but also due to the size of the under-explored part of the signalling network. In this Chapter, the possibility of data-driven methods to identify signalling modules of co-phosphorylated phosphosites is explored. Data-driven modules were derived from three phosphoproteomic data sets extracted from NTERA2, MCF7 and HL60 cell lines (Hijazi et al., 2020) for further analysis. In HL60 and MCF7 proteins that clustered together were found to be closer to each other than proteins that belonged to other clusters the same pattern did not emerge in NTERA2 cells. The difference, while significant, was very small. While one would expect phosphosites that cluster together to be close to each other. These results confirm the hypothesis that biological networks are small world networks with most nodes being only a few edges away from each other (Jeong et al., 2001; Wagner and Fell, 2001). In addition, empirical phosphoproteomic perturbation studies have shown that the signal propagates farther and more widely across the proteome than one would expect if the signalling network was divided into highly insulated compartments as current models (Jassal et al., 2020; Slenter et al., 2018) suggest. This suggests that the idea of a more interconnected signalling network, which is reflected in these results, might be closer to biological reality.

I found that the data-driven modules have a greater number of significant enrichments for independent data sets and a higher odds ratio compared to the protein modules in the Reactome pathway data set, as well as modules extracted from the OmniPath literature interaction network. This indicates that data-driven methods have a greater explanatory power when it comes to the analysis of phosphoproteomic data by better capturing the inherent complexity in the signalling network.

Even though functional enrichment could be made, the assignment of function to each of the modules remains a challenge. Part of the explanation lies in the fact that evidently the modules defined in the literature do not capture the signalling system except to a limited extent. Similarly, the GWAS conducted to establish link between modules and traits yielded mixed results. The association of traits with Reactome pathways yielded more significant results in terms of lower p-values. However, a larger portion of the data-driven modules are assigned a trait than the Reactome pathways. In both cases, a relatively low portion of modules (11 (data-driven) and 5% (Reactome pathways) were assigned a trait indicating that in general, complex traits do not map neatly onto pathways or signalling network modules. The slight advantage the data-driven modules have might be due to the fact that phospho-signals propagate more widely throughout the signalling network than is often depicted in the literature (Humphrey et al., 2015; Köksal et al., 2018; Olsen et al., 2006). Since, they capture this property of the signalling network which might better reflect the distribution of trait associated mutations across the network (Dozmorov et al., 2020).

On the other hand, I found that many of the modules were significantly correlated with transcription factor activities and kinase activities across biological samples and cell lines. Indeed, looking at the top module- transcription factor activity correlation for each module, a higher top correlation was achieved between the data-driven modules then the Reactome pathways. The same was true for kinase activities, indicating a biologically meaningful signal that outperforms current pathway models. At the same time, these results are surprising due to the fact that our current methods of assessing kinase and transcription factor activities are based on known targets of these molecules. These results perhaps reflect that the assignment of protein interactions are compartmentalized in a somewhat arbitrary way in the current pathway annotations (Domingo-Fernández et al., 2018), while cross-talks are common and the interaction network is more interconnected than the pathways databases reflect (Vert and Chory, 2011).

Furthermore, I looked into the ability of the modules to predict co-occurrences in signalling pathways in concert with high throughput data, RNAseq from human tissue (Thul et al., 2017; Uhlén et al., 2015, 2013) and co-phosphorylation from cell line data (Roumeliotis et al., 2017). While the modules had a limited predictive power which reflects the

challenge it has been to assign biological function to the modules, I found that the modules slightly improved predictive performance when combined co-phosphorylation while not improving the predictive power of co-expression when these features are combined. Interestingly, co-expression seems to have a higher predictive power than co-phosphorylation. These results are counter to what would be expected as phosphorylation should in theory better reflect signal propagation through signalling pathways but can partly be explained by the large scale of the understudied phosphoproteome. These results however, agree with analysis done recently (Szalai and Saez-Rodriguez, 2020) which found that pathways as defined in the literature overlap significantly with transcription modules which explains the surprising finding that RNAseq data has higher explanatory power when it comes to known pathways. Another reason could be that phosphosites are usually not assigned to pathways. In other words, while two proteins might share the same signalling pathway only a single phosphosite of many might explain a protein's role in that particular signalling pathway, meaning that phosphorylation on a phosphosite level might add significant noise. Interestingly, the two features seemed to complement each other. In other words, there is low intersection between high probability pairs captured by co-expression and co-phosphorylation, particularly, when looking at phosphosite pairs that have not been previously assigned to the same pathways. These results might indicate that different data sets are needed to fully capture signalling processes as each method has its strengths and limitations: Some proteins in a signalling process might not be regulated by phosphorylation in which case expression data might be more suitable. At the same time, expression profiles might not capture proteins regulated by phosphorylation while remaining expressed across conditions. Limitations include the tendency of expression data to capture transcription factor modules as discussed above while mass spectrometry data has incomplete coverage (Tabb et al., 2010; Timp and Timp, 2020).

The main limitation of the data-driven approach stems from technical limitations of data generation. In order to identify modules of co-regulated phosphosites several challenges need to be overcome. One of the more important ones is that sampling is inherent in mass spectrometry data generation meaning that phosphosites captured in one data set are often missing in another comparable set (Tabb et al., 2010). Furthermore, static

phosphoproteomic data sets often do not capture variation in phospho-signalling across samples but rather changes in protein expression levels (Roumeliotis et al., 2017). Therefore, it is preferable to incorporate either time series data, that is data sets where phosphorylation levels are measured at different time points under otherwise the same condition and in the same biological sample (i.e. same tissue or same cell line). Yet another issue is context-specificity, meaning that phosphorylation profiles differ a lot across different conditions (Saez-Rodriguez et al., 2011). One implication of this is that in order to capture all modules active under all environments, an infinite number of data sets is needed. It is therefore clear that a complete mapping of human signalling modules will remain a challenge in the foreseeable future.

5 Conclusion

The overarching theme of this thesis has been to develop methods to expand our current understanding of human cell signalling networks beyond its well-studied components. My main contributions lie in the development of largely data-driven methods to predict kinase-kinase regulatory networks, which I then expanded also to non-kinase-substrates and the identification of a set of purely data-driven modules of phospho-signalling. These methods, as well as the resulting networks, provide the basis on which to start exploring the dark space of human cell signalling networks and also highlight both the extent to which this is unexplored and the potential for new hypothesis generation and discovery. As more comprehensive datasets become available, these methods can be reapplied to continually improve our insights into the human cell signalling network.

5.1 Summary of results and key findings

In **Chapter 2** I discuss the development, in collaboration with my colleague Brandon Invergo (Beltrao group), of a method to predict signed kinase-kinase regulatory circuits by using features based on kinase specificities, co-expression, co-phosphorylation and phosphosite functionality. We find that our predictions capture known interactions that were not included in the training set and we are able to make predictions across the spectrum of the kinases regardless of how well they are studied. This is in accordance with recent literature from interactomics, which have shown that unbiased protein interaction networks do cover the entire spectrum of the proteome rather than the well-studied portions of it. (Invergo and Beltrao, 2018; Luck et al., 2020; Rolland et al., 2014). Furthermore, we find that our kinase network is modular with modules that are functionally distinct from one another which agrees with the common but not in reality proven- belief that biological systems are modular in structure. Crucially, we find that the kinase-kinase predictions can be validated by novel experimental kinase-substrate predictions (Hijazi et al., 2020; Sugiyama et al., 2019). With the aid of these novel predictions to support our predictors we propose new kinase-kinase regulatory pathways that traverse between the three kinase-kinase regulatory relationships: SRC and CDK1, SRC and CDK2 and the regulation of PRPF4B by MAPK1. SRC is known to phosphorylate CDK1

(PhosphoSitePlus (Hornbeck et al., 2015), downloaded May 2, 2021) while a signed regulatory relationship has not been established between the two. Both of the suggested target sites found on CDK1, Y15 and Y19, have a high functional score of > 0.8 which adds further evidence of regulation between the two proteins. No relationship is reported between the other pairs in OmniPath and there is no association reported in STRING (Szklarczyk et al., 2021), suggesting that these two are entirely novel. The relationship between MAPK1 and PRPF4B is of particular interest as it was supported by both experiments (Hijazi et al., 2020; Sugiyama et al., 2019). In the predictions made by Sugiyama et al, MAPK1 is suggested to phosphorylate PRPF4B on position 578 and 580, both of which are relatively functional with functional values of 0.46 and 0.53 respectively representing the 86th and 93th percentile. Hijazi and colleagues, however, predict phosphorylation at sites 87 (functional score of 0.31, 63th percentile) and 93 (functional score of 0.55, 94th percentile). So while the data sets do not agree on the phosphosites both predict phosphorylation at phosphosites that are predicted to be highly functional.

In **Chapter 3** I discuss an expansion upon the work presented in Chapter 2 where exact kinase-substrate relationships are predicted. Predictions are based on a similar set of features as discussed in Chapter 2 with the addition of features relevant to the potential phospho-acceptor sites. I find that the resulting predictions manage to predict interactions between less studied proteins when compared with the established data bases such as PhosphoSitePlus. The predictor also manages, for the first time, to predict the sign of the kinase-substrate relationships. Furthermore, I find that I am able to capture experimentally-derived edges from completely independent datasets. Furthermore, my method does better at capturing known kinase-substrates and experimentally predicted edges than other state-of-the-art methods (Blom et al., 2004; Horn et al., 2014; Nováček et al., 2020; Patrick et al., 2015; Wang et al., 2020), while having a greater coverage of the human kinome, with the exception of GPS v.5.0 (Wang et al., 2020). This is potentially due to the fact that my method incorporates features based on high throughput data and information characterizing phosphosites which allows for kinase-substrate predictions that are not purely based on kinase specificities.

In **Chapter 4**, I discuss my work on the development of data-driven modules. Current modules of signal transduction, represented by curated pathways, (Jassal et al., 2020;

Kanehisa, 2019; Slenter et al., 2018) are insufficient when it comes to explaining high throughput data. I find that my modules perform better at capturing phosphorylation patterns in independent data sets. However, pinpointing biological function has remained a challenge even with the majority of the modules have a significant pathway enrichment. Meanwhile, most modules could not be associated with traits vie GWAS analysis indicated that most of them were not significantly associated with any trait. It should be stated that relatively few Reactome pathways were associated with traits as well, which might indicate that the traits tested can only be associated with a few modules and that a greater number of traits need to be tested for more general module-trait association. Furthermore, traits such as height are complex which means that associated SNPs can be expected to be distributed among many modules (Dozmorov et al., 2020). However, modules could be associated with kinase and transcription factor activities. The data-driven modules had a higher association with their most highly associated kinase or transcription factor in terms of protein activity compared to the Reactome modules. This seems to indicate that there is in fact a strong link between the biological modules and wider biological regulation in terms of transcription and phosphorylation.

5.2 Limitations of this study

One of the main challenges of this project has been to explore the less studied phosphoproteome without over-relying on the literature and thereby succumbing to its inherent bias. In the case of kinase-substrate predictions and kinase-kinase relationships, we heavily relied on position weight matrices which in turn are constructed from inherently biased knowledgebases. While high-throughput associative features and features that contain descriptive information on phosphosites and kinases do remedy this in part, predictions made by our models are inevitably biased towards previous knowledge. The way forward in the prediction of novel kinase-substrates will be to rely more heavily on high throughput data. Preferably, perturbation phosphoproteomics data should be used as RNA expression is known to poorly correlate with protein levels. Ironically, despite reliance on the literature, the accumulation of false positives is always going to be a challenge when a large number of predictions are made. In the case of this thesis well over 90% of the edges predicted in both chapters will prove hard to validate. Predictions

that involve less studied kinases are particularly hard to assess as they are less likely to have any information on them in the data bases. There have been some efforts to experimentally predict kinase-substrate relationships (Hijazi et al., 2020; Sugiyama et al., 2019). These methods, however, do themselves accumulate a large number of kinase-substrate predictions that cannot be corroborated by any external source. Similarly, there are some methodological issues such as the problems of identifying direct interactions in the case of cell line studies while *in vitro* studies are conducted under conditions that are not present in the cell. Nevertheless, these studies are invaluable in the validation of *in silico* predictions as they provide a way of validation that is independent of the biased literature sources. Edges that are predicted by multiple experimental approaches, provided that the methodology is different enough can be expected to be fairly likely to be accurate.

One thing to consider for the future development of the work described in chapter four is the use of databases that are known to include pathways that are too detailed on one hand and not well defined on the other hand as a reference point against which the data-driven modules are compared. Databases with more standardized and uniformly curated pathway databases such as SIGNOR (Licata et al., 2020) or SignaLink (Csabai et al., 2021) could be more useful as a reference point to assess the performance of the data-driven modules.

The available phosphoproteomic data sets also provide us with a different set of challenges. Currently, diseases are a popular field for phospho-proteomic study and for a valid reason. Cancers and other diseases are known to lead to dysregulation of phospho-signalling and mutations common in cancers are often found in signalling proteins (Yaffe, 2019). In addition, current phosphoproteomics protocols rely on very large numbers of cells to acquire high quality data. Therefore, large-scale phosphoproteomic studies are often conducted in disease models such as cancer cell lines. However, this raises the question of whether modules extracted from these cell lines and protein relationship predictions based on the data are generally valid under normal conditions. Large scale phosphoproteomic studies need to be studied under a diverse set of conditions including healthy tissue for any truly general assumption to be made. In this thesis, data sets extracted from cancer cell lines and cancer samples were the primary

source for module extraction. The independent data sets used to showcase the improved ability of the modules to capture independent regulation were also from cancer cell line data sets. It is therefore not clear, due to the ubiquitous use of cancer data sets, if the results presented here do generally apply in healthy tissue, or indeed in tissues other than breast, that was used in our study. Similarly, cancer data was used to generate the phosphoproteomic based features for kinase-substrate predictions. While other features such as co-expression from tissue data and in particular PWMs do not exhibit this bias, it is conceivable that the inclusion of these features has made the predictions skew towards a more cancer specific network of interactions.

Phospho-enriched mass spectrometry data sets also pose challenges to the aims of this thesis. This is partly due to context specificity but also due to less than full recovery of phospho-peptides. This means that unlike RNAseq data sets, the phosphosites quantified can vary across different measurements. This makes capturing trends and system-wide analyses difficult due to data missingness. As a result, the data-driven modules described in this thesis were generated from different data sets and are therefore all cell type specific to an extent since they were generated from different perturbation states (kinase inhibitors).

5.3 Future directions

At present, our understanding of the human signalling network is limited. Only a fraction of the more than 100,000 phosphosites have a known upstream kinase and around 20% of kinases remain without a single annotated substrate (Needham et al., 2019) with most known kinase-substrate relationships occurring between a well-studied kinase and a well-studied substrate protein (Invergo and Beltrao, 2018). Due to our limited knowledge the current models of signalling pathways fall short when it comes to predicting outcome of perturbations in the signalling system (Humphrey et al., 2015; Köksal et al., 2018; Olsen et al., 2006). At the same time the signalling system is at the centre of many complex chronic disease such as cancers (Yaffe, 2019), with kinases being some of the most commonly dysregulated cancer driver proteins (Zhang et al., 2009). This stresses the need to expand upon our knowledge of the phosphoproteomic system due to its integral role in cell decision making and development and as a result in diseases. At the same

time, the signalling system is complex and context specific (Hill et al., 2017) meaning that not all pathways can be captured at once with the same data set. Furthermore, due the number of potential kinase-substrate interactions and regulatory circuits, computational methods are needed to spearhead the exploration of the human phosphoproteome.

In this thesis I set out to explore the understudied phosphoproteome. I described methods predicting kinase-substrate and kinase-kinase interactions by combining kinase specificities and high throughput data. In this way my collaborators and I managed to make predictions even for less studied kinases and proteins. Additionally, I discuss the generation of data-driven signalling modules that better capture independent phosphoproteomic data sets, and discuss ways to assign function to these novel data-driven modules.

There are, however, remaining issues. While here I describe a more data-driven method for kinase-substrate predictions, predictions still rely on the literature as kinase specificity models feature heavily in predictions, them being the most powerful predictive feature (Chapters 2 and 3). Kinase specificity models have also been commonly used for kinase-substrate predictions by others (Horn et al., 2014; Invergo et al., 2020; Patrick et al., 2015). While literature-derived kinase specificity models are not without problems such as the introduction of bias, the results presented here show that the generation of high-quality specificity models for all kinases are crucial for accurate kinase-substrate predictions. In the case of the inference of kinase-kinase relationships, PWMs provide directionality that would be hard to obtain simply by using features based on high throughput data. Furthermore, they have the ability to capture direct interactions whereas it is hard to distinguish between direct and indirect effects simply from correlative analysis of phosphoproteomics or RNAseq data. A way of getting around the study-bias introduced by applying these models for kinase-substrate predictions would be to obtain a high-quality set of unbiased and experimentally derived relationships. While the accumulation of false positives is an inherent problem in these kinds of high throughput experiments, the intersection of different sets of kinase-substrate predictions, derived by different methods, could be used to produce a high confidence set of kinase-substrates from which specificity models could be constructed. While, to my knowledge, this has not been attempted before it could be a viable strategy once a greater number of such experiments

are made and under greater numbers of conditions. These could also help validate computational predictions and prioritize kinase-substrates for further testing with low-throughput, more accurate methods. Many of these methods depend on mass spectrometry and while some methods have made predictions for a great number of kinases some substrates might not be captured which makes the intersection between such sets challenging. I also found that, when fitting kinase-substrate predictions to phosphoproteomic data sets. Therefore, hypothetically using multiple data sets we could assign higher confidence to edges that are included in multiple sub-networks derived from multiple high-throughput data sets.

Assignment of biological function to novel data-driven modules is also a major challenge even though indicative functional enrichment can be found, since our current state of the art knowledge on functional units in the network is quite limited when it comes to explaining high throughput phosphoproteomic data. While association with biological molecules such as transcription factors and kinases give us an idea of their function a better functional characterization is needed. On way of addressing this issue could be to apply forward genetics such as CRISPR-based assays (Shalem et al., 2015) for functional annotation and phenotypic screening. Already, such methods have been used to simultaneously modifying the expression of multiple genes, providing us with a promising way to experimentally validate the signalling modules generated in this project by analysing the phenotypic impact of module activation or inhibition (Konermann et al., 2015).

Central to the limitations of this thesis are the limitations of the use of mass spectrometry for phosphosite quantification. Recent technologies in particular nanopores have been used to capture primary structure of unfolded proteins and experiments such as the one conducted by Rosen (Rosen et al., 2014) and colleagues showed that phosphorylation could be captured. Rosen et al. identified phosphorylation on thioredoxin on two adjacent sites and were able to differentiate between single and double phosphorylation using α-hemolysin pore. Similarly, AeL and FraC nanopores have been used to identify phosphorylation on peptides (Meng et al., 2019; Restrepo-Pérez et al., 2019). This approach works on the single molecule level and therefore addresses the inability of mass spectrometers to detect low abundance proteins. None of these experiments sequenced

the proteins or peptides, however, nanopores have already been used for DNA sequencing by passing single strand DNA through the pores (Manrao et al., 2012). The development of similar approaches for protein sequencing is a very 'hot' field currently. Proteins, however, are more complex due to the 20 different amino acids and need to be unfolded to be able to be sequences by nanopores. While proteins have been passed through nanopores (Kennedy et al., 2016), differentiating between the 20 different amino acids remains a challenge partly because of their smaller size. Individual amino acids have been identified as well as phosphotyrosine (Ohshiro et al., 2014; Zhao et al., 2014) but many have remained indistinguishable. In 2019, Ouldali and colleagues found that aerolysin nanopore could be used to identify 13 of the 20 amino acids while being able to detect even more by chemically modifying the nanopore. A way of sequencing proteins in this manner might address the difficulty inherent in assigning phosphoryl group to a specific site in 10 AA long peptide that arises due to the fact that multiple serine/threonine and tyrosine residues could be present in the peptide. These advances could provide us with a more accurate and complete coverage of the cellular phosphoproteome.

Nanopores' ability to capture low abundance protein might also solve another problem; that of data sparsity. With more complete proteomic data sets, we might be able to extract more general modules and better identify modules that are active across most conditions, which would further aid in the mapping of the architecture of phospho-signalling. Furthermore, nanopore technologies are relatively cheap compared to mass spectrometers. Another method that has been proposed as an alternative to the mass spectrometer in protein identification and quantification is protein fingerprinting where certain amino acids are tagged with fluorescent reporters. Recently, Swaminathan and colleagues coupled fluorescent tagging with Edman's degradation for sequencing (Swaminathan et al., 2018). Their method works for thousands to millions of protein molecules in parallel. This method, however is slow and there still challenges that have to be addressed if they are to be used for PTM detection as PTMs need to be specifically labelled (Swaminathan et al., 2018). To date, labelling chemistry is only available for a small number of PTMs. Put together, these advances promise alternatives to mass spectrometers that address several of the problems inherent in mass spectrometry data generation, namely, low specificity, high cost, low reproducibility due to sampling and

difficulty in pinpointing the locations of PTMs in the sequence. As a result, better informed hypotheses can be generated for testing as well as a more robust conclusions can be drawn from the data which in turn makes data-driven exploration of the “dark” phosphoproteome a less daunting task.

Nevertheless the work discussed in this thesis is a valuable first step towards more data-driven approaches in the identification of signalling circuits signalling modules which in turn could prove useful in the development of targeted drug therapies. Computational and data-driven methods could provide us with the tools necessary both to ignore kinases that are likely to be more important to cell function or participate in a greater number of interactions than previously thought as well as target previously understudied or neglected phosphorylation events.

5.4 Concluding remarks

This study has focused on the exploration of the less studied parts of the phosphoproteome. My results indicate that data-driven methods of kinase-substrate prediction perform well at capturing novel edges when kinase specificity models are combined with high throughput data and various phosphosite related features. This combination works better at capturing both known interactions and experimentally predicted interactions than other state of the art methods that often rely solely on kinase specificity models. We show that machine learning methods can make predictions on regulatory signs when high throughput data is combined with high throughput predictors and other information such as evolutionary age and location within the protein are incorporated into the predictions. This is true both for kinase-kinase relationship predictions as well as kinase-substrate predictions. Furthermore, the computational methods proposed here are able to make high confidence predictions for less studied kinases and substrates. However, heavy reliance on biased data bases remains an issue in this field.

Another take home message from the results presented in this thesis is that data-driven modules, derived from clusters of co-phosphorylated phosphosites, are better suited for describing phosphoproteomic data compared to pathways as they are laid out in the data bases. The challenge of assigning a function to these modules remains, even though

some hints can be drawn from associating the modules with the significant pathway enrichments, the activity profile of kinases and transcription factors. Then there is the question of high-quality high throughput data. For general broadly applicable modules to be generated a large number of high coverage and high throughput data sets need to be generated. Novel technologies such as nanopores could potentially rise to the challenge and meet this need but the use of purely data-driven modules will remain a challenge in the foreseeable future not only due to technical and computational difficulty in generating them in a robust manner but also due to the paradigm shift needed to move away from the static models (pathways) that are currently being used. Nevertheless, due to the fact that these modules do seem to perform better at capturing phosphoproteomic changes that occur under perturbation, there is an indication that data-driven approaches are needed to capture the and analyse and understand the human phospho-signalling network architecture on a system wide scale.

References

1. Akbani, R., Ng, P.K., Werner, H.M., Shahmoradgoli, M., Zhang, F., Ju, Z., Liu, W., Yang, J.Y., Yoshihara, K., and Li, J. (2014). A pan-cancer proteomic perspective on the Cancer Genome Atlas. Nat Commun 5: 3887.
2. Akhmedov, M., Kedaigle, A., Chong, R.E., Montemanni, R., Bertoni, F., Fraenkel, E., and Kwee, I. (2017). PCSF: An R-package for network-based interpretation of high-throughput data. PLOS Comput. Biol. *13*, e1005694.
3. Alberts, B., Johnson, A., and Lewis, J. (2002). Molecular Biology of the Cell (New York, NY, USA: Garland Science).
4. Altman, N., and Krzywinski, M. (2018). The curse(s) of dimensionality. Nat. Methods *15*, 399–400.
5. Alvarez, M.J., Shen, Y., Giorgi, F.M., Lachmann, A., Ding, B.B., Ye, B.H., and Califano, A. (2016). Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat. Genet. *48*, 838–847.
6. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.
7. Atay, O., and Skotheim, J.M. (2014). Modularity and predictability in cell signalling and decision making. Mol. Biol. Cell *25*, 3445–3450.
8. Ayati, M., Chance, M.R., and Koyutürk, M. (2021). Co-phosphorylation networks reveal subtype-specific signalling modules in breast cancer. Bioinformatics *37*, 221–228.
9. Ayati, M., Wiredja, D., Schlatzer, D., Maxwell, S., Li, M., Koyutürk, M., and Chance, M.R. (2019). CoPhosK: A method for comprehensive kinase substrate annotation using co-phosphorylation analysis. PLOS Comput. Biol. *15*, e1006678.
10. Azorsa, D.O., Robeson, R.H., Frost, D., hoovet, B.M., Brautigam, G.R., Dickey, C., Beaudry, C., Basu, G.D., Holz, D.R., Hernandez, J.A., et al. (2010).

High-content siRNA screening of the kinome identifies kinases involved in Alzheimer's disease-related tau hyperphosphorylation. BMC Genomics *11*, 25.

11. Barabási, A.-L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat. Rev. Genet. *5*, 101–113.
12. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. Nature *483*, 603.
13. Benjamini, Y., and Hochberg, Y. (1995). Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing.
14. Berenjeno, I.M., Piñeiro, R., Castillo, S.D., Pearce, W., McGranahan, N., Dewhurst, S.M., Meniel, V., Birkbak, N.J., Lau, E., and Sansregret, L. (2017). Oncogenic PIK3CA induces centrosome amplification and tolerance to genome doubling. Nat. Commun. *8*, 1–15.
15. Bersanelli, M., Mosca, E., Remondini, D., Castellani, G., and Milanesi, L. (2016). Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. Sci. Rep. *6*, 34841.
16. Bhattacharyya, R.P., Reményi, A., Yeh, B.J., and Lim, W.A. (2006). Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signalling circuits. Annu. Rev. Biochem. *75*, 655–680.
17. Bhullar, K.S., Lagarón, N.O., McGowan, E.M., Parmar, I., Jha, A., Hubbard, B.P., and Rupasinghe, H.P.V. (2018). Kinase-targeted cancer therapies: progress, challenges and future directions. Mol. Cancer *17*, 48–48.
18. Biton, A., Zinovyev, A., Barillot, E., and Radvanyi, F. MineICA: Independent component analysis of transcriptomic data.
19. Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics *4*, 1633–1649.
20. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. *2008*, P10008.

21. Bodenmiller, B., Wanka, S., Kraft, C., Urban, J., Campbell, D., Pedrioli, P.G., Gerrits, B., Picotti, P., Lam, H., Vitek, O., et al. (2010). Phosphoproteomic Analysis Reveals Interconnected System-Wide Responses to Perturbations of Kinases and Phosphatases in Yeast. Sci. Signal. *3*, rs4–rs4.

22. Bolstad, B.M., Irizarry, R.A., Åstrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193.

23. Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8: 3–62.—Open Access Library [Internet]. 2020 [cited 2020 Jan 24]. Google Sch.

24. Boyer, P.D., DeLuca, M., Ebner, K.E., Hultquist, D.E., and Peter, J.B. (1962). Identification of phosphohistidine in digests from a probable intermediate of oxidative phosphorylation. J. Biol. Chem. *237*, PC3306–PC3308.

25. Bradley, D., and Beltrao, P. (2018). Evolution of protein kinase substrate recognition at the active site. BioRxiv 443945.

26. Bradley, D., Viéitez, C., Rajeeve, V., Selkrig, J., Cutillas, P.R., and Beltrao, P. (2021). Sequence and Structure-Based Analysis of Specificity Determinants in Eukaryotic Protein Kinases. Cell Rep. *34*, 108602.

27. Buck, S.F. (1960). A Method of Estimation of Missing Values in Multivariate Data Suitable for use with an Electronic Computer. J. R. Stat. Soc. Ser. B Methodol. *22*, 302–306.

28. Bullinaria, J.A. (2007). Understanding the Emergence of Modularity in Neural Systems. Cogn. Sci. *31*, 673–695.

29. Cancer Genome Atlas Research Network, C.G.A. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330.

30. Cancer Genome Atlas Research Network, C.G.A.R. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609.

31. Canovas, B., and Nebreda, A.R. (2021). Diversity and versatility of p38 kinase signalling in health and disease. Nat. Rev. Mol. Cell Biol. *22*, 346–366.

32. Catherman, A.D., Skinner, O.S., and Kelleher, N.L. (2014). Top Down

proteomics: Facts and perspectives. Adv. OMICs-Based Discip. *445*, 683–693.

33. Chang, W., Cheng, J., Allaire, J.J., Xie, Y., and McPherson, J. (2019). shiny: Web Application Framework for R.

34. Chipman, H.A., George, E.I., and McCulloch, R.E. (2010). BART: Bayesian additive regression trees. Ann. Appl. Stat. *4*, 266–298.

35. Choobdar, S., Ahsen, M.E., Crawford, J., Tomasoni, M., Fang, T., Lamparter, D., Lin, J., Hescott, B., Hu, X., Mercer, J., et al. (2019). Assessment of network module identification across complex diseases. Nat. Methods *16*, 843–852.

36. Clark, D.J., Dhanasekaran, S.M., Petralia, F., Pan, J., Song, X., Hu, Y., da Veiga Leprevost, F., Reva, B., Lih, T.-S.M., and Chang, H.-Y. (2019). Integrated proteogenomic characterization of clear cell renal cell carcinoma. Cell *179*, 964-983. e31.

37. Clauset, A., Newman, M.E.J., and Moore, C. (2004). Finding community structure in very large networks. Phys Rev E *70*, 066111.

38. Cortes, C., and Vapnik, V. (2004). Support-vector networks. Mach. Learn. *20*, 273–297.

39. Corwin, T., Woodsmith, J., Apelt, F., Fontaine, J.-F., Meierhofer, D., Helmuth, J., Grossmann, A., Andrade-Navarro, M.A., Ballif, B.A., and Stelzl, U. (2017). Defining Human Tyrosine Kinase Phosphorylation Networks Using Yeast as an In Vivo Model Substrate. Cell Syst. *5*, 128-139.e4.

40. Coscia, M., Rossetti, G., Giannotti, F., and Pedreschi, D. (2012). DEMON: a local-first discovery method for overlapping communities (Beijing, China: Association for Computing Machinery).

41. Cox, D.R. (1958). The Regression Analysis of Binary Sequences. J. R. Stat. Soc. Ser. B Methodol. *20*, 215–242.

42. Csabai, L., Fazekas, D., Kadlecsik, T., Szalay-Bekő, M., Bohár, B., Madgwick, M., Módos, D., Ölbei, M., Gul, L., Sudhakar, P., et al. (2021). SignaLink3: a multi-layered resource to uncover tissue-specific signaling networks. Nucleic Acids Research gkab909.

43. Csárdi, G., and Nepusz, T. (2006). The igraph software package for

complex network research. InterJournal *Complex Systems*, 1695.

44. Cutillas, P.R., and Vanhaesebroeck, B. (2007). Quantitative Profile of Five Murine Core Proteomes Using Label-free Functional Proteomics *. Mol. Cell. Proteomics *6*, 1560–1573.

45. Damle, N.P., and Köhn, M. (2019). The human DEPhOsphorylation Database DEPOD: 2019 update. Database *2019*.

46. Dana, J.M., Gutmanas, A., Tyagi, N., Qi, G., O'Donovan, C., Martin, M., and Velankar, S. (2019). SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Res. *47*, D482–D489.

47. Dau, T., Bartolomucci, G., and Rappsilber, J. (2020). Proteomics Using Protease Alternatives to Trypsin Benefits from Sequential Digestion with Trypsin. Anal Chem *92*, 9523–9527.

48. Degli Esposti, D., Almunia, C., Guery, M.-A., Koenig, N., Armengaud, J., Chaumot, A., and Geffard, O. (2019). Co-expression network analysis identifies gonad- and embryo-associated protein modules in the sentinel species Gammarus fossarum. Sci. Rep. *9*, 7862.

49. Dey, K.K., Xie, D., and Stephens, M. (2018). A new sequence logo plot to highlight enrichment and depletion. BMC Bioinformatics *19*, 473.

50. Diella, F., Gould, C.M., Chica, C., Via, A., and Gibson, T.J. (2008). Phospho.ELM: a database of phosphorylation sites—update 2008. Nucleic Acids Res. *36*, D240–D244.

51. Dinkel, H., Van Roey, K., Michael, S., Kumar, M., Uyar, B., Altenberg, B., Milchevskaya, V., Schneider, M., Kühn, H., Behrendt, A., et al. (2016). ELM 2016--data update and new functionality of the eukaryotic linear motif resource. Nucleic Acids Res. *44*, D294–D300.

52. Doerr, A. (2008). Phosphorylation and the cell cycle. Nat. Methods *5*, 858–859.

53. Domingo-Fernández, D., Hoyt, C.T., Bobis-Álvarez, C., Marín-Llaó, J., and Hofmann-Apitius, M. (2018). ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. Npj Syst. Biol. Appl. *4*, 43.

54. Doncic, A., and Skotheim, J.M. (2013). Feedforward regulation ensures stability and rapid reversibility of a cellular state. Mol. Cell *50*, 856–868.

55. Doncic, A., Falleur-Fettig, M., and Skotheim, J.M. (2011). Distinct interactions select and maintain a specific cell fate. Mol. Cell *43*, 528–539.

56. Dou, Y., Kawaler, E.A., Zhou, D.C., Gritsenko, M.A., Huang, C., Blumenberg, L., Karpova, A., Petyuk, V.A., Savage, S.R., and Satpathy, S. (2020). Proteogenomic characterization of endometrial carcinoma. Cell *180*, 729-748. e26.

57. Dozmorov, M.G., Cresswell, K.G., Bacanu, S.-A., Craver, C., Reimers, M., and Kendler, K.S. (2020). A method for estimating coherence of molecular mechanisms in major human disease and traits. BMC Bioinformatics *21*, 473.

58. Edwards, A.M., Isserlin, R., Bader, G.D., Frye, S.V., Willson, T.M., and Yu, F.H. (2011). Too many roads not taken. Nature *470*, 163–165.

59. Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., and Ketchum, K.A. (2015). The CPTAC Data Portal: A Resource for Cancer Proteomics Research. J. Proteome Res. *14*, 2707–2713.

60. Ellis, J.J., and Kobe, B. (2011). Predicting Protein Kinase Specificity: Predikin Update and Performance in the DREAM4 Challenge. PLOS ONE *6*, e21169.

61. Eves Eva M., Xiong Wen, Bellacosa Alfonso, Kennedy Scott G., Tsichlis Philip N., Rosner Marsha Rich, and Hay Nissim (1998). Akt, a Target of Phosphatidylinositol 3-Kinase, Inhibits Apoptosis in a Differentiating Neuronal Cell Line. Mol. Cell. Biol. *18*, 2143–2152.

62. Fabbro, D., Cowan-Jacob, S.W., and Moebitz, H. (2015). Ten things you should know about protein kinases: IUPHAR Review 14. Br. J. Pharmacol. *172*, 2675–2700.

63. Fabien Viger and Matthieu Latapy (2005). Fast generation of random connected graphs with prescribed degrees. ArXiv.

64. Ferron, F., Rancurel, C., Longhi, S., Cambillau, C., Henrissat, B., and Canard, B. (2005). VaZyMolO: a tool to define and classify modularity in viral proteins. J. Gen. Virol. *86*, 743–749.

65. Fisher, R.A. (1935). The Logic of Inductive Inference. Journal of the Royal Statistical Society *98*, 39–82.

66. Gao, Q., Zhu, H., Dong, L., Shi, W., Chen, R., Song, Z., Huang, C., Li, J., Dong, X., and Zhou, Y. (2019). Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. Cell *179*, 561-577. e22.

67. Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. *29*, 1363–1375.

68. Gates, A.J., and Ahn, Y.-Y. (2019). CluSim: a python package for calculating clustering similarity. J Open Source Softw *4*, 1264.

69. Gene Ontology Consortium (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res *49*, D325–D334.

70. Gillette, M.A., Satpathy, S., Cao, S., Dhanasekaran, S.M., Vasaikar, S.V., Krug, K., Petralia, F., Li, Y., Liang, W.-W., and Reva, B. (2020). Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. Cell *182*, 200-225. e35.

71. Gillis, J., Ballouz, S., and Pavlidis, P. (2014). Bias tradeoffs in the creation and analysis of protein-protein interaction networks. J. Proteomics *100*, 44–54.

72. Glass, K., and Girvan, M. (2014). Annotation Enrichment Analysis: An Alternative Method for Evaluating the Functional Properties of Gene Sets. Sci. Rep. *4*, 4191.

73. Good, M.C., Zalatan, J.G., and Lim, W.A. (2011). Scaffold proteins: hubs for controlling the flow of cellular information. Science *332*, 680–686.

74. Griss, J., Perez-Riverol, Y., Lewis, S., Tabb, D.L., Dianes, J.A., Del-Toro, N., Rurik, M., Walzer, M., Kohlbacher, O., and Hermjakob, H. (2016). Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. Nat. Methods *13*, 651–656.

75. Gry, M., Rimini, R., Strömberg, S., Asplund, A., Pontén, F., Uhlén, M., and Nilsson, P. (2009). Correlations between RNA and protein expression profiles in 23 human cell lines. BMC Genomics *10*, 365.

76. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx)

project. Nat. Genet. *45*, 580–585.

77. Guerois, R., Nielsen, J.E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J. Mol. Biol. *320*, 369–387.

78. Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. Mach. Learn. *46*, 389–422.

79. H. B. Mann and D. R. Whitney (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann. Math. Stat. *18*, 50–60.

80. Harchaoui, Z., Bach, F., Cappe, O., and Moulines, E. (2013). Kernel-Based Methods for Hypothesis Testing: A Unified View. Signal Process. Mag. IEEE *30*, 87–97.

81. Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. Nature 402, C47–C52.

82. Hartigan, J.A., and Wong, M.A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. J. R. Stat. Soc. Ser. C Appl. Stat. *28*, 100–108.

83. Henikoff, J.G., and Henikoff, S. (1996). Using substitution probabilities to improve position-specific scoring matrices. Bioinformatics *12*, 135–143.

84. Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. *89*, 10915–10919.

85. Henikoff, S., and Henikoff, J.G. (1994). Position-based sequence weights. J. Mol. Biol. *243*, 574–578.

86. Hernandez, M., Lachmann, A., Zhao, S., Xiao, K., and Ma'ayan, A. (2010). Inferring the Sign of Kinase-Substrate Interactions by combining quantitative phosphoproteomics with a literature-based mammalian kinome network. In 2010 IEEE International Conference on BioInformatics and BioEngineering, (IEEE), pp. 180–184.

87. Herwig, R., Hardt, C., Lienhard, M., and Kamburov, A. (2016). Analyzing and interpreting genome data at the network level with ConsensusPathDB. Nat. Protoc. *11*, 1889–1907.

88. Hijazi, M., Smith, R., Rajeeve, V., Bessant, C., and Cutillas, P.R. (2020). Reconstructing kinase network topologies from phosphoproteomics data reveals cancer-associated rewiring. Nat. Biotechnol.

89. Hill, S.M., Heiser, L.M., Cokelaer, T., Unger, M., Nesser, N.K., Carlin, D.E., Zhang, Y., Sokolov, A., Paull, E.O., Wong, C.K., et al. (2016). Inferring causal molecular networks: empirical assessment through a community-based effort. Nat. Methods *13*, 310.

90. Hill, S.M., Lu, Y., Molina, J., Heiser, L.M., Spellman, P.T., Speed, T.P., Gray, J.W., Mills, G.B., and Mukherjee, S. (2012). Bayesian Inference of signalling Network Topology in a Cancer Cell Line. Bioinformatics *28*, 2804–2810.

91. Hill, S.M., Nesser, N.K., Johnson-Camacho, K., Jeffress, M., Johnson, A., Boniface, C., Spencer, S.E.F., Lu, Y., Heiser, L.M., Lawrence, Y., et al. (2017). Context Specificity in Causal signalling Networks Revealed by Phosphoprotein Profiling. Cell Syst. *4*, 73-83.e10.

92. Himberg, J., Hyvärinen, A., and Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. NeuroImage *22*, 1214–1222.

93. Hoogendijk, A.J., Pourfarzad, F., Aarts, C.E.M., Tool, A.T.J., Hiemstra, I.H., Grassi, L., Frontini, M., Meijer, A.B., van den Biggelaar, M., and Kuijpers, T.W. (2019). Dynamic Transcriptome-Proteome Correlation Networks Reveal Human Myeloid Differentiation and Neutrophil-Specific Programming. Cell Rep. *29*, 2505-2519.e4.

94. Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Schärfe, C.P.I., Springer, M., Sander, C., and Marks, D.S. (2017). Mutation effects predicted from sequence co-variation. Nat. Biotechnol. *35*, 128–135.

95. Horn, H., Schoof, E.M., Kim, J., Robin, X., Miller, M.L., Diella, F., Palma, A., Cesareni, G., Jensen, L.J., and Linding, R. (201a). KinomeXplorer: an integrated platform for kinome biology studies. Nat. Methods *11*, 603–604.

96. Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: Mutations, PTMs and

recalibrations. Nucleic Acids Res. *43*, D512-20.

97. Huang, H.-D., Lee, T.-Y., Tzeng, S.-W., Wu, L.-C., Horng, J.-T., Tsou, A.-P., and Huang, K.-T. (2005). Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites. J. Comput. Chem. *26*, 1032–1041.

98. Humphrey, S.J., Azimifar, S.B., and Mann, M. (2015). High-throughput phosphoproteomics reveals in vivo insulin signalling dynamics. Nat. Biotechnol. *33*, 990–995.

99. Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw. *10*, 626–634.

100. Imamura, H., Sugiyama, N., Wakabayashi, M., and Ishihama, Y. (2014). Large-scale identification of phosphorylation sites for profiling protein kinase selectivity. J. Proteome Res. *13*, 3410–3419.

101. Invergo, B.M., and Beltrao, P. (2018). Reconstructing phosphorylation signalling networks from quantitative phosphoproteomic data. Essays Biochem. *62*, 525–534.

102. Invergo, B.M., Petursson, B., Akhtar, N., Bradley, D., Giudice, G., Hijazi, M., Cutillas, P., Petsalaki, E., and Beltrao, P. (2020). Prediction of Signed Protein Kinase Regulatory Circuits. Cell Syst. *10*, 384-396.e9.

103. Jaccard, P. (1912). THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. New Phytol. *11*, 37–50.

104. Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776–779.

105. Järvelin, K., and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. *20*, 422–446.

106. Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway knowledgebase. Nucleic Acids Res. *48*, D498–D503.

107. Jeong, H., Mason, S.P., Barabási, A.-L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41–42.

108. Johnson, A. (2014). Onset of CDK2 activity determines passage through the restriction point in primary cells (Stanford University).

109. Jones, D.T., and Cozzetto, D. (2015). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinforma. Oxf. Engl. *31*, 857–863.

110. Jothi, R., Balaji, S., Wuster, A., Grochow, J.A., Gsponer, J., Przytycka, T.M., Aravind, L., and Babu, M.M. (2009). Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. Mol. Syst. Biol. *5*, 294.

111. Jutten, C., and Herault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal Process. *24*, 1–10.

112. Kairov, U., Cantini, L., Greco, A., Molkenov, A., Czerwinska, U., Barillot, E., and Zinovyev, A. (2017). Determining the optimal number of independent components for reproducible transcriptomic data analysis. BMC Genomics *18*, 712.

113. Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. Protein Sci. Publ. Protein Soc. *28*, 1947–1951.

114. Karp, P.D., Latendresse, M., Paley, S.M., Krummenacker, M., Ong, Q.D., Billington, R., Kothari, A., Weaver, D., Lee, T., and Subhraveti, P. (2016). Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. Brief. Bioinform. *17*, 877–890.

115. Karve, T.M., and Cheema, A.K. (2011). Small Changes Huge Impact: The Role of Protein Posttranslational Modifications in Cellular Homeostasis and Disease. J. Amino Acids *2011*, 207691.

116. Kashtan, N., Mayo, A.E., Kalisky, T., and Alon, U. (2009). An Analytically Solvable Model for Rapid Evolution of Modular Structure. PLOS Comput. Biol. *5*, e1000355.

117. Kel, A.E., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., and Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res *31*, 3576–3579

118. Kennedy, E., Dong, Z., Tennant, C., and Timp, G. (2016). Reading the primary structure of a protein with 0.07 nm 3 resolution using a subnanometre-diameter pore. Nat. Nanotechnol. *11*, 968–976.

119. Kholodenko, B.N., Kiyatkin, A., Bruggeman, F.J., Sontag, E., Westerhoff, H.V., and Hoek, J.B. (2002). Untangling the wires: A strategy to trace functional interactions in signalling and gene networks. Proc. Natl. Acad. Sci. *99*, 12841.

120. Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

121. Kirschner, M., and Gerhart, J. (1998). Evolvability. Proc Natl Acad Sci USA *95*, 8420.

122. Kitano, H. (2004). Biological robustness. Nat. Rev. Genet. *5*, 826–837.

123. Knorre, D.G., Kudryashova, N.V., and Godovikova, T.S. (2009). Chemical and functional aspects of posttranslational modification of proteins. Acta Naturae *1*, 29–51.

124. Koboldt, D., Fulton, R., McLellan, M., Schmidt, H., Kalicki-Veizer, J., McMichael, J., Fulton, L., Dooling, D., Ding, L., and Mardis, E. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.

125. Köksal, A.S., Beck, K., Cronin, D.R., McKenna, A., Camp, N.D., Srivastava, S., MacGilvray, M.E., Bodík, R., Wolf-Yadlin, A., Fraenkel, E., et al. (2018). Synthesizing signalling Pathways from Temporal Phosphoproteomic Data. Cell Rep. *24*, 3607–3618.

126. Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. Inst Ital Attuari Giorn *4*, 83–91.

127. Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., et al. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature *517*, 583–588.

128. Koussounadis, A., Langdon, S.P., Um, I.H., Harrison, D.J., and Smith, V.A. (2015). Relationship between differentially expressed mRNA and mRNA-

protein correlations in a xenograft model system. Sci. Rep. *5*, 10775.

129. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97.

130. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., and Bergmann, S. (2016). Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLOS Comput. Biol. *12*, e1004714.

131. Landry, C.R., Levy, E.D., and Michnick, S.W. (2009). Weak functional constraints on phosphoproteomes. Trends Genet. *25*, 193–197.

132. Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

133. Lapek, J.D., Greninger, P., Morris, R., Amzallag, A., Pruteanu-Malinici, I., Benes, C.H., and Haas, W. (2017). Detection of dysregulated protein-association networks by high-throughput proteomics predicts cancer vulnerabilities. Nat. Biotechnol. *35*, 983–989.

134. Lappano, R., and Maggiolini, M. (2011). G protein-coupled receptors: novel targets for drug discovery in cancer. Nat. Rev. Drug Discov. *10*, 47–60.

135. Lawrence, R.T., Perez, E.M., Hernández, D., Miller, C.P., Haas, K.M., Irie, H.Y., Lee, S.-I., Blau, C.A., and Villén, J. (2015). The proteomic landscape of triple-negative breast cancer. Cell Rep. *11*, 630–644.

136. Lazar, C., Gatto, L., Ferro, M., Bruley, C., and Burger, T. (2016). Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. J. Proteome Res. *15*, 1116–1125.

137. Lee, B., and Richards, F.M. (1971). The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. *55*, 379-IN4.

138. Lee, S.-I., and Batzoglou, S. (2003). Application of independent component analysis to microarrays. Genome Biol. *4*, R76.

139. Li, Y., Zhou, X., Zhai, Z., and Li, T. (2017). Co-occurring protein phosphorylation are functionally associated. PLoS Comput. Biol. *13*, e1005502–

e1005502.

140. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinforma. Oxf. Engl. *27*, 1739–1740.

141. Licata, L., Lo Surdo, P., Iannuccelli, M., Palma, A., Micarelli, E., Perfetto, L., Peluso, D., Calderone, A., Castagnoli, L., and Cesareni, G. (2020). SIGNOR 2.0, the signalling Network Open Resource 2.0: 2019 update. Nucleic Acids Res. *48*, D504–D510.

142. Lorenz, D.M., Jeng, A., and Deem, M.W. (2011a). The emergence of modularity in biological systems. Phys. Life Rev. *8*, 129–160.

143. Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020a). A reference map of the human binary protein interactome. Nature *580*, 402–408.

144. Madsen, R.R., Knox, R.G., Pearce, W., Lopez, S., Mahler-Araujo, B., McGranahan, N., Vanhaesebroeck, B., and Semple, R.K. (2019). Oncogenic PIK3CA promotes cellular stemness in an allele dose-dependent manner. Proc. Natl. Acad. Sci. *116*, 8380.

145. Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science *298*, 1912–1934.

146. Manrao, E.A., Derrington, I.M., Laszlo, A.H., Langford, K.W., Hopper, M.K., Gillgren, N., Pavlenok, M., Niederweis, M., and Gundlach, J.H. (2012). Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. Nat. Biotechnol. *30*, 349–353.

147. Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nat. Methods *13*, 366–370.

148. Mayer, B.J. (2015). The discovery of modular binding domains: building blocks of cell signalling. Nat. Rev. Mol. Cell Biol. *16*, 691–698.

149. Meng, F.-N., Ying, Y.-L., Yang, J., and Long, Y.-T. (2019). A Wild-Type Nanopore Sensor for Protein Kinase Activity. Anal. Chem. *91*, 9910–9915.

150. Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature *534*, 55–62.

151. Miao, B., Xiao, Q., Chen, W., Li, Y., and Wang, Z. (2018). Evaluation of functionality for serine and threonine phosphorylation with different evolutionary ages in human and mouse. BMC Genomics *19*, 431.

152. Mishra, A.K., Sharma, V., Mutsuddi, M., and Mukherjee, A. (2021). signalling cross-talk during development: Context-specific networking of Notch, NF-κB and JNK signalling pathways in Drosophila. Cell. Signal. *82*, 109937.

153. Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. Nat. Methods *10*, 47–53.

154. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. *9*, S4.

155. Mubeen, S., Hoyt, C.T., Gemünd, A., Hofmann-Apitius, M., Fröhlich, H., and Domingo-Fernández, D. (2019). The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. Front. Genet. *10*, 1203.

156. Mun, D.-G., Bhin, J., Kim, S., Kim, H., Jung, J.H., Jung, Y., Jang, Y.E., Park, J.M., Kim, H., and Jung, Y. (2019). Proteogenomic characterization of human early-onset gastric cancer. Cancer Cell *35*, 111-124. e10.

157. Nair, A., Chauhan, P., Saha, B., and Kubatzky, K.F. (2019). Conceptual Evolution of Cell signalling. Int. J. Mol. Sci. *20*, 3292.

158. Needham, E.J., Parker, B.L., Burykin, T., James, D.E., and Humphrey, S.J. (2019). Illuminating the dark phosphoproteome. Sci Signal *12*, eaau8645.

159. Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods *9*, 471–472.

160. Nguyen, T.-M., Shafi, A., Nguyen, T., and Draghici, S. (2019). Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biol. *20*, 203.

161. Nilsen, G., and Lingjaerde, O.C. (2013). clusterGenomics: Identifying clusters in genomics data by recursive partitioning.

162. Nováček, V., McGauran, G., Matallanas, D., Vallejo Blanco, A., Conca, P., Muñoz, E., Costabello, L., Kanakaraj, K., Nawaz, Z., Walsh, B., et al. (2020). Accurate prediction of kinase-substrate networks using knowledge graphs. PLOS Comput. Biol. *16*, e1007578.

163. Obenauer, J.C., Cantley, L.C., and Yaffe, M.B. (2003). Scansite 2.0: proteome-wide prediction of cell signalling interactions using short sequence motifs. Nucleic Acids Res. *31*, 3635–3641.

164. Ochoa, D., Jarnuczak, A.F., Viéitez, C., Gehre, M., Soucheray, M., Mateus, A., Kleefeldt, A.A., Hill, A., Garcia-Alonso, L., Stein, F., et al. (2020). The functional landscape of the human phosphoproteome. Nat. Biotechnol. *38*, 365–373.

165. Ochoa, D., Jonikas, M., Lawrence, R.T., El Debs, B., Selkrig, J., Typas, A., Villen, J., Santos, S.D., and Beltrao, P. (2016). An atlas of human kinase regulation. Mol. Syst. Biol. *12*, 888.

166. Ohshiro, T., Tsutsui, M., Yokota, K., Furuhashi, M., Taniguchi, M., and Kawai, T. (2014). Detection of post-translational modifications in single peptides using electron tunnelling currents. Nat. Nanotechnol. *9*, 835–840.

167. Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006). Global, In Vivo, and Site-Specific Phosphorylation Dynamics in signalling Networks. Cell *127*, 635–648.

168. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., et al. (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res *42*, D358-63.

169. Orr, H.A. (2000). Adaptation and the cost of complexity. Evolution *54*, 13–20.

170. Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., et al. (2019). The BioGRID interaction database: 2019 update. Nucleic Acids Res. *47*, D529–D541.

171. Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. Nature *435*, 814–818.

172. Papageorgiou, A., Rapley, J., Mesirov, J.P., Tamayo, P., and Avruch, J. (2015). A Genome-Wide siRNA Screen in Mammalian Cells for Regulators of S6 Phosphorylation. PLOS ONE *10*, e0116096.

173. Papatheodorou, I., Fonseca, N.A., Keays, M., Tang, Y.A., Barrera, E., Bazant, W., Burke, M., Fullgrabe, A., Fuentes, A.M.-P., George, N., et al. (2018). Expression Atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Res. *46*, D246–D251.

174. Parter, M., Kashtan, N., and Alon, U. (2007). Environmental variability and modularity of bacterial metabolic networks. BMC Evol. Biol. *7*, 169.

175. Patrick, R., Kobe, B., Lê Cao, K.-A., and Bodén, M. (2017). PhosphoPICK-SNP: quantifying the effect of amino acid variants on protein phosphorylation. Bioinformatics *33*, 1773–1781.

176. Patrick, R., Lê Cao, K.-A., Kobe, B., and Bodén, M. (2015). PhosphoPICK: modelling cellular context to map kinase-substrate phosphorylation events. Bioinformatics *31*, 382–389.

177. Pawson, T. (1995). Protein modules and signalling networks. Nature *373*, 573–580.

178. Pawson, T., Gish, G.D., and Nash, P. (2001). SH2 domains, interaction modules and cellular wiring. Trends Cell Biol. *11*, 504–511.

179. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

180. Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. ELECTROPHORESIS *20*, 3551–3567.

181. Perou, C.M., Jeffrey, S.S., van de Rijn, M., Rees, C.A., Eisen, M.B., Ross, D.T., Pergamenschikov, A., Williams, C.F., Zhu, S.X., Lee, J.C., et al. (1999). Distinctive gene expression patterns in human mammary epithelial cells and

breast cancers. Proc. Natl. Acad. Sci. U. S. A. *96*, 9212–9217.

182. Petralia, F., Tignor, N., Reva, B., Koptyra, M., Chowdhury, S., Rykunov, D., Krek, A., Ma, W., Zhu, Y., and Ji, J. (2020). Integrated proteogenomic characterization across major histological types of pediatric brain cancer. Cell *183*, 1962-1985. e31.

183. Petsalaki, E., Helbig, A.O., Gopal, A., Pasculescu, A., Roth, F.P., and Pawson, T. (2015). SELPHI: correlation-based identification of kinase-associated networks from global phospho-proteomics data sets. Nucleic Acids Res. *43*, W276—82.

184. Petsalaki, E., Stark, A., García-Urdiales, E., and Russell, R.B. (2009). Accurate prediction of peptide binding sites on protein surfaces. PLoS Comput. Biol. *5*, e1000335.

185. Picart-Armada, S., Thompson, W.K., Buil, A., and Perera-Lluna, A. (2018). diffuStats: an R package to compute diffusion-based scores on biological networks. Bioinformatics *34*, 533–534.

186. Picart-Armada, S., Thompson, W.K., Buil, A., and Perera-Lluna, A. (2021). The effect of statistical normalization on network propagation scores. Bioinformatics *37*, 845–852.

187. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J.X., and Jensen, L.J. (2015). DISEASES: text mining and data integration of disease-gene associations. Methods San Diego Calif *74*, 83–89.

188. Pollastri, G., Baldi, P., Fariselli, P., and Casadio, R. (2002a). Prediction of coordination number and relative solvent accessibility in proteins. Proteins Struct. Funct. Bioinforma. *47*, 142–153.

189. Pollastri, G., Przybylski, D., Rost, B., and Baldi, P. (2002b). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins *47*, 228–235.

190. Price, M.N., Dehal, P.S., and Arkin, A.P. (2008). Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. Genome Biol. *9*, R4.

191. Rainey, P.B., and Cooper, T.F. (2004). Evolution of bacterial diversity and

the origins of modularity. Genome Plast. Evol. Microb. Genomes *155*, 370–375.

192. Ren, R., Mayer, B.J., Cicchetti, P., and Baltimore, D. (1993). Identification of a ten-amino acid proline-rich SH3 binding site. Science *259*, 1157–1161.

193. Restrepo-Pérez, L., Wong, C.H., Maglia, G., Dekker, C., and Joo, C. (2019). Label-Free Detection of Post-translational Modifications with a Nanopore. Nano Lett. *19*, 7957–7964.

194. Rinaldi, L., Delle Donne, R., Borzacchiello, D., Insabato, L., and Feliciello, A. (2018). The role of compartmentalized signalling pathways in the control of mitochondrial activities in cancer cells. Biochim. Biophys. Acta BBA - Rev. Cancer *1869*, 293–302.

195. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

196. Rodchenkov, I., Babur, O., Luna, A., Aksoy, B.A., Wong, J.V., Fong, D., Franz, M., Siper, M.C., Cheung, M., Wrana, M., et al. (2020). Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Res. *48*, D489–D497.

197. Rohani, N., and Eslahchi, C. (2020). Classifying Breast Cancer Molecular Subtypes by Using Deep Clustering Approach. Front. Genet. *11*, 553587–553587.

198. Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014a). A proteome-scale map of the human interactome network. Cell *159*, 1212–1226.

199. Rosen, C.B., Rodriguez-Larrea, D., and Bayley, H. (2014). Single-molecule site-specific detection of protein phosphorylation with a nanopore. Nat. Biotechnol. *32*, 179–181.

200. Roumeliotis, T.I., Williams, S.P., Gonçalves, E., Alsinet, C., Del Castillo Velasco-Herrera, M., Aben, N., Ghavidel, F.Z., Michaut, M., Schubert, M., Price, S., et al. (2017). Genomic Determinants of Protein Abundance Variation in Colorectal Cancer Cells. Cell Rep. *20*, 2201–2214.

201. Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation

and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65.

202. Rozemberczki, B., Kiss, O., and Sarkar, R. (2020). Karate Club: an API oriented open-source python framework for unsupervised learning on graphs. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 3125–3132.

203. Rudolph, J.D., de Graauw, M., van de Water, B., Geiger, T., and Sharan, R. (2016). Elucidation of signalling Pathways from Large-Scale Phosphoproteomic Data Using Protein Interaction Networks. Cell Syst. *3*, 585-593.e3.

204. Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. Nature *323*, 533–536.

205. Sacco, F., Perfetto, L., Castagnoli, L., and Cesareni, G. (2012). The human phosphatase interactome: An intricate family portrait. FEBS Lett. *586*, 2732–2739.

206. Saez-Rodriguez, J., Alexopoulos, L.G., Epperlein, J., Samaga, R., Lauffenburger, D.A., Klamt, S., and Sorger, P.K. (2009). Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. Mol. Syst. Biol. *5*, 331.

207. Saez-Rodriguez, J., Alexopoulos, L.G., Zhang, M., Morris, M.K., Lauffenburger, D.A., and Sorger, P.K. (2011). Comparing signalling networks between normal and transformed hepatocytes using discrete logical models. Cancer Res. *71*, 5400–5411.

208. Schapire, R.E., and Singer, Y. (1999). Improved Boosting Algorithms Using Confidence-rated Predictions. Mach. Learn. *37*, 297–336.

209. Schey, K.L., Grey, A.C., and Nicklay, J.J. (2013). Mass spectrometry of membrane proteins: a focus on aquaporins. Biochemistry *52*, 3807–3817.

210. Scrucca, L., Fop, M., Murphy, T.B., and Raftery, A.E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. R J. *8*, 289–317.

211. Seyfried, N.T., Dammer, E.B., Swarup, V., Nandakumar, D., Duong, D.M., Yin, L., Deng, Q., Nguyen, T., Hales, C.M., Wingo, T., et al. (2017). A Multi-

network Approach Identifies Protein-Specific Co-expression in Asymptomatic and Symptomatic Alzheimer's Disease. Cell Syst. *4*, 60-72.e4.

212. Shalem, O., Sanjana, N.E., and Zhang, F. (2015). High-throughput functional genomics using CRISPR–Cas9. Nat. Rev. Genet. *16*, 299–311.

213. Siddle, K., Ursø, B., Niesler, C.A., Cope, D.L., Molina, L., Surinya, K.H., and Soos, M.A. (2001). Specificity in ligand binding and intracellular signalling by insulin and insulin-like growth factor receptors. Biochem. Soc. Trans. *29*, 513–525.

214. Simillion, C., Liechti, R., Lischer, H.E.L., Ioannidis, V., and Bruggmann, R. (2017). Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinformatics *18*, 151.

215. Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 3940–3941.

216. Singh, A.H., Wolf, D.M., Wang, P., and Arkin, A.P. (2008). Modularity of stress response evolution. Proc. Natl. Acad. Sci. U. S. A. *105*, 7500–7505.

217. Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S.L., Digles, D., et al. (2018a). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. *46*, D661–D667.

218. Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. Ann. Math. Stat. *19*, 279–281.

219. Smith, F.D., and Scott, J.D. (2002). signalling complexes: junctions on the intracellular information super highway. Curr. Biol. *12*, R32–R40.

220. Solé, R.V., and Valverde, S. (2008). Spontaneous emergence of modularity in cellular networks. J. R. Soc. Interface *5*, 129–133.

221. Sompairac, N., Nazarov, P.V., Czerwinska, U., Cantini, L., Biton, A., Molkenov, A., Zhumadilov, Z., Barillot, E., Radvanyi, F., Gorban, A., et al. (2019). Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. Int. J. Mol. Sci. *20*, 4414.

222. Sousa, A., Dugourd, A., Memon, D., Petursson, B., Petsalaki, E., Saez-Rodriguez, J., and Beltrao, P. (2021). Pan-Cancer landscape of protein activities

identifies drivers of signalling dysregulation and patient survival. BioRxiv 2021.06.09.447741.

223. Spirin, V., and Mirny, L.A. (2003). Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. U. S. A. *100*, 12123–12128.

224. Spirin, V., Gelfand, M.S., Mironov, A.A., and Mirny, L.A. (2006). A metabolic network in the evolutionary context: multiscale structure and modularity. Proc. Natl. Acad. Sci. *103*, 8774–8779.

225. Stormo, G.D., Schneider, T.D., Gold, L., and Ehrenfeucht, A. (1982). Use of the “Perceptron” algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res. *10*, 2997–3011.

226. Strumillo, M.J., Oplová, M., Viéitez, C., Ochoa, D., Shahraz, M., Busby, B.P., Sopko, R., Studer, R.A., Perrimon, N., Panse, V.G., et al. (2019). Conserved phosphorylation hotspots in eukaryotic protein domain families. Nat Commun *10*, 1977.

227. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550.

228. Sugiyama, N., Imamura, H., and Ishihama, Y. (2019). Large-scale Discovery of Substrates of the Human Kinome. Sci. Rep. *9*, 10503.

229. Swaminathan, J., Boulgakov, A.A., Hernandez, E.T., Bardo, A.M., Bachman, J.L., Marotta, J., Johnson, A.M., Anslyn, E.V., and Marcotte, E.M. (2018). Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. Nat. Biotechnol. *36*, 1076–1082.

230. Szalai, B., and Saez-Rodriguez, J. (2020). Why do pathway methods work better than they should? FEBS Lett. *594*, 4189–4200.

231. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., et al. (2021). The STRING database in 2021: customizable protein-protein networks, and functional

characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 49, D605–D612.

232. Tabb, D.L., Vega-Montoto, L., Rudnick, P.A., Variyath, A.M., Ham, A.-J.L., Bunk, D.M., Kilpatrick, L.E., Billheimer, D.D., Blackman, R.K., and Cardasis, H.L. (2010). Repeatability and reproducibility in proteomic identifications by liquid chromatography− tandem mass spectrometry. J. Proteome Res. *9*, 761–776.

233. Tarca, A.L., Carey, V.J., Chen, X., Romero, R., and Drăghici, S. (2007). Machine Learning and Its Applications to Biology. PLOS Comput. Biol. *3*, e116.

234. Tengholm, A., and Meyer, T. (2002). A PI3-Kinase signalling Code for Insulin-Triggered Insertion of Glucose Transporters into the Plasma Membrane. Curr. Biol. *12*, 1871–1876.

235. Thul, P.J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L.M., et al. (2017). A subcellular map of the human proteome. Science *356*.

236. Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B Stat. Methodol. *63*, 411–423.

237. Timp, W., and Timp, G. (2020). Beyond mass spectrometry, the next step in proteomics. Sci. Adv. *6*, eaax8978.

238. Tin Kam Ho (1995). Random decision forests. In Proceedings of 3rd International Conference on Document Analysis and Recognition, pp. 278–282 vol.1.

239. Torkamani, A., Verkhivker, G., and Schork, N.J. (2009). Cancer driver mutations in protein kinase genes. Cancer Lett. *281*, 117–127.

240. Turei, D., Korcsmaros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signalling pathway resources. Nat. Methods *13*, 966–967.

241. Ubersax, J.A., and Ferrell Jr, J.E. (2007). Mechanisms of specificity in protein phosphorylation. Nat. Rev. Mol. Cell Biol. *8*, 530–541.

242. Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P.,

Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419.

243. UniProt Consortium, T. (2018). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *46*, 2699–2699.

244. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. (2019). Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. *18*, 463–477.

245. Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R.

246. Van Dongen, S.M. (2000). Graph clustering by flow simulation.Ph.D.thesis,Universtiy of Utrecht.

247. Vandin, F., Upfal, E., and Raphael, B.J. (2011). Algorithms for detecting significantly mutated pathways in cancer. J. Comput. Biol. J. Comput. Mol. Cell Biol. *18*, 507–522.

248. Vert, G., and Chory, J. (2011). Crosstalk in cellular signalling: background noise or the real thing? Dev. Cell *21*, 985–991.

249. Vidal, M.C., Sato, J.R., Balardin, J.B., Takahashi, D.Y., and Fujita, A. (2017). ANOCVA in R: A Software to Compare Clusters between Groups and Its Application to the Study of Autism Spectrum Disorder. Front. Neurosci. *11*, 16.

250. Vlastaridis, P., Kyriakidou, P., Chaliotis, A., Van de Peer, Y., Oliver, S.G., and Amoutzias, G.D. (2017). Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes. GigaScience *6*, 1–11.

251. Von Mering, C., Zdobnov, E.M., Tsoka, S., Ciccarelli, F.D., Pereira-Leal, J.B., Ouzounis, C.A., and Bork, P. (2003). Genome evolution reveals biochemical networks and functional modules. Proc. Natl. Acad. Sci. U. S. A. *100*, 15428–15433.

252. Wagner, A., and Fell, D.A. (2001). The small world inside large metabolic networks. Proc. R. Soc. Lond. B Biol. Sci. *268*, 1803–1810.

253. Wagner, G.P., and Altenberg, L. (1996). PERSPECTIVE: COMPLEX ADAPTATIONS AND THE EVOLUTION OF EVOLVABILITY. Evol. Int. J. Org.

Evol. *50*, 967–976.

254. Wang, C., Xu, H., Lin, S., Deng, W., Zhou, J., Zhang, Y., Shi, Y., Peng, D., and Xue, Y. (2020). GPS 5.0: An Update on the Prediction of Kinase-specific Phosphorylation Sites in Proteins. Genomics Proteomics Bioinformatics *18*, 72–80.

255. Wang, L., Wang, H.-F., Liu, S.-R., Yan, X., and Song, K.-J. (2019). Predicting Protein-Protein Interactions from Matrix-Based Protein Sequence Using Convolution Neural Network and Feature-Selective Rotation Forest. Sci. Rep. *9*, 9848.

256. Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S.P., Hengartner, M.O., and von Mering, C. (2012). PaxDb, a database of protein abundance averages across all three domains of life. Mol. Cell. Proteomics MCP *11*, 492–500.

257. Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F., and Jones, D.T. (2004). The DISOPRED server for the prediction of protein disorder. Bioinformatics *20*, 2138–2139.

258. Watson, N.A., Cartwright, T.N., Lawless, C., Cámara-Donoso, M., Sen, O., Sako, K., Hirota, T., Kimura, H., and Higgins, J.M.G. (2020). Kinase inhibition profiles as a tool to identify kinases for specific phosphorylation sites. Nat. Commun. *11*, 1684.

259. Webb-Robertson, B.-J.M., Wiberg, H.K., Matzke, M.M., Brown, J.N., Wang, J., McDermott, J.E., Smith, R.D., Rodland, K.D., Metz, T.O., Pounds, J.G., et al. (2015). Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. J. Proteome Res. *14*, 1993–2001.

260. Wee, P., and Wang, Z. (2017). Epidermal Growth Factor Receptor Cell Proliferation signalling Pathways. Cancers *9*, 52.

261. Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., and Ni, Y. (2018). Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. Sci. Rep. *8*, 663.

262. Widmann, C., Gibson, S., Jarpe, M.B., and Johnson, G.L. (1999).

Mitogen-activated protein kinase: conservation of a three-kinase module from yeast to human. Physiol. Rev. *79*, 143–180.

263. Wiese, R.J., Mastick, C.C., Lazar, D.F., and Saltiel, A.R. (1995). Activation of Mitogen-activated Protein Kinase and Phosphatidylinositol 3′-Kinase Is Not Sufficient for the Hormonal Stimulation of Glucose Uptake, Lipogenesis, or Glycogen Synthesis in 3T3-L1 Adipocytes (∗). J. Biol. Chem. *270*, 3442–3446.

264. Wilkes, E.H., Terfve, C., Gribben, J.G., Saez-Rodriguez, J., and Cutillas, P.R. (2015). Empirical inference of circuitry and plasticity in a kinase signalling network. Proc. Natl. Acad. Sci. U. S. A. *112*, 7719–7724.

265. Xiao, Q., Miao, B., Bi, J., Wang, Z., and Li, Y. (2016). Prioritizing functional phosphorylation sites based on multiple feature integration. Sci. Rep. *6*, 24735.

266. Xu, C., and Jackson, S.A. (2019). Machine learning and complex biological data. Genome Biol. *20*, 76.

267. Yaffe, M.B. (2019). Why geneticists stole cancer research even though cancer is primarily a signalling disease. Sci. Signal. *12*, eaaw3483.

268. Ye, Y., and Doak, T.G. (2009). A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes. PLOS Comput. Biol. *5*, e1000465.

269. Yu, G., and He, Q.-Y. (2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol BioSyst *12*, 477–479.

270. Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., and Kim, S. (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382–387.

271. Zhang, H., Liu, T., Zhang, Z., Payne, S.H., Zhang, B., McDermott, J.E., Zhou, J.-Y., Petyuk, V.A., Chen, L., Ray, D., et al. (2016). Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell *166*, 755–765.

272. Zhang, J., Yang, P.L., and Gray, N.S. (2009). Targeting cancer with small molecule kinase inhibitors. Nat. Rev. Cancer *9*, 28–39.

273. Zhao, Y., Ashcroft, B., Zhang, P., Liu, H., Sen, S., Song, W., Im, J., Gyarfas, B., Manna, S., Biswas, S., et al. (2014). Single-molecule spectroscopy

of amino acids and peptides by recognition tunnelling. Nat. Nanotechnol. *9*, 466–473.

274. Zhou, W., and Altman, R.B. (2018). Data-driven human transcriptomic modules determined by independent component analysis. BMC Bioinformatics *19*, 327.

275. Zhu, X., and Stephens, M. (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nat. Commun. *9*, 4361.

276. Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M.M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Inf. Fusion *50*, 71–91.

277. Zoidi, O., Fotiadou, E., Nikolaidis, N., and Pitas, I. (2015). Graph-Based Label Propagation in Digital Media: A Review. ACM Comput Surv *47*.

278. žurauskienė, J., and Yau, C. (2016). pcaReduce: hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics *17*, 140.

Appendix 3.1

Features considered for kinase-substrate model training (Chapter 3.2.3). Apart from: Residue, PWM score, Coreg_293, Functional_score, GTEx, RNA_tissue, RNA_cell, Kinase_selectivity, Substrate_selectivity, NTERA2_coreg, MCF7_coreg and HL60_coreg. The features were downloaded from previous publication by Ochoa et al. (Ochoa et al., 2020)

Name of feature	**Description**	**Imputation**	**Used for kinase-substrate prediction**	**Used for sign prediction**
Residue	phospho acceptor residue: S/T/Y	No missing values	Y	Y
PWM score	Score generated by fitting phosphosite to kinases' PWM (Invergo et al., 2020)	No missing values	Y	Y
Coreg_293	Association between kinase activity and phosphorylation of phosphosite across 86 cancer samples (Mertins et al., 2016)	0	Y	Y
Functional_score	Score denoting the probability of phosphosite to be functional(Ochoa et al., 2020)	No missing values	Y	Y
GTEx	Co-expression between kinase and putative substrate gene across tissues GTEx(GTEx Consortium ,2013)	0	Y	Y

RNA_tissue	Co-expression between kinase and putative substrate gene across tissues (Uhlén et al., 2015)	0	Y	Y
RNA_cell	Co-expression between kinase and putative substrate gene across cell lines (Thul et al., 2017)	0	Y	Y
Kinase_selectivity	kinase skewness in expression distribution across tissues (Uhlén et al., 2015) (See chapter 2)	0	Y	Y
Substrate_selectivity	Substrate protein skewness in expression distribution across tissues (Uhlén et al., 2015) (See chapter 2)	0	Y	Y
NTERA2_coreg	Association between kinase activity and phosphorylation of phosphosite across 63 kinase inhibition conditions, NTERA2 cell line (Hijazi et al., 2020).	0	Y	Y
MCF7_coreg	Association between kinase activity and phosphorylation of phosphosite across 63 kinase inhibition conditions, MCF7 cell line (Hijazi et al., 2020).	0	Y	Y

HL60_coreg	Association between kinase activity and phosphorylation of phosphosite across 63 kinase inhibition conditions, HL60 cell line (Hijazi et al., 2020).	0	Y	Y
Is_DISOPRED	Indicates if phosphosite is disordered. If Disopred v2 score is lower than 0.5, the site is considered disordered (Jones and Cozzetto, 2015).	Phosphosites with missing values are assumed to be disordered	N	Y
DISOPRED score	The phosphosite disorder score as given by DISOPRED v2 (Jones and Cozzetto, 2015).	Median imputation	Y	Y
Exp3d_ala_ddG_effect	Discretized changes in Gibbs energy when residue is mutated into alanine. Measured by FoldX v4 (Guerois et al., 2002)	Missing data set to unknown	N	Y
Exp3d_acid_ddG_effect	Discretized average changes in Gibbs energy when residue is mutated into acidic residue. Measured by FoldX v4 (Guerois et al., 2002)	Missing data set to unknown	Y	Y

Log_10_of_hotspot_pval_min	The feature quantifies occurrence of phosphorylation events within the structural region. The p-value indicates if the phosphosite is found within a region within which phosphosites are significantly enriched (Strumillo et al., 2019)	0	Y	Y
Is_hotspot	If the phosphosite has hotspot (Strumillo et al., 2019) enrichment p_value of <3.36e-07 and has been found in more than 10 MS data sets, the site is considered a hotspot.	Missing values set to FALSE	N	Y

Is_Interface	Data on experimentally resolved or modelled interaction interfaces were downloaded from Interactome3d (Mosca et al., 2013). NACCESS (Lee and Richards, 1971) was used to calculate relative solvent accessibility of atoms. I the relative solvent accessibility changed between interacting and non-interacting form the residue was considered to be on the interface	Missing values are set to FALSE	Y	Y
Adj_ptms w21	Number of phosphosites within +/- 10 residues on either side of the phosphosite.	Zero	Y	Y
Netpho_max_all	Highest posterior probability derived from all models in NetPhorest v2.1 (Horn et al., 2014).	Median imputation	Y	Y
Netpho_max_KIN	Highest posterior probability derived from kinase models in NetPhorest v2.1 (Horn et al., 2014).	Median imputation	Y	Y

Paxdb_abundance_log10	Consensus protein abundance as deposited in the PaxD(Wang et al., 2012) data base	Median imputation	Y	Y
W0_mya	Age of the phosphosite. The age was estimated by combining phylogenetic data with cross species phosphoproteomics (Ochoa et al., 2020).	0	Y	Y
W3_mya	Age of the +/- 3 amino acid region surrounding the phosphosite. The age was estimated by combining phylogenetic data with cross species phosphoproteomics (Ochoa et al., 2020).	0	Y	Y
Quant_top1	Number of times the phosphosite was in the 1% regulated phosphosites across 435 conditions (Ochoa et al., 2016).	0	Y	Y
Quant_top5	Number of times the phosphosite was in the 1% regulated phosphosites across 435 conditions (Ochoa et al., 2016)	0	Y	Y

PWM_max_mss	The best fit to the PSSM of 143 kinases with at least 10 known. Measured as a score from 0 to 1 with the MATCH algorithm (Kel et al., 2003).	Median imputation	Y	Y
ACCpro	Solvent accessibility as predicted by ACCpro (Pollastri et al., 2002a)	Missing values set to unknown	Y	Y
SSpro	Secondary structures estimates form SSpro(Pollastri et al., 2002b)	Missing values set to unknown	N	Y
SSpro8	One of the following class: alpha-helix, 3-10 helix, pi-helix, extended strand, beta-bridge, turn, bend and the rest as predicted by SSpro8 (Pollastri et al., 2002b)	Missing values set to unknown	N	Y
SIFT_min_score	SIFT (Dana et al., 2019) score is used as a proxy for conservation as it predicts functional impact of variants. This scores calculates minimum score across all variants	Median imputation	Y	Y

SIFT_mean_score	SIFT (Dana et al., 2019) score is used as a proxy for conservation as it predicts functional impact of variants. This scores calculates mean score across all variants	Median imputation	Y	Y
SIFT _ala_score	SIFT (Dana et al., 2019) score is used as a proxy for conservation as it predicts functional impact of variants. This scores calculates score of alanine variants	Median imputation	Y	Y
SIFT_acid_score	SIFT (Dana et al., 2019) score is used as a proxy for conservation as it predicts functional impact of variants. This score calculates average score across variants leading to negative charge.	Median imputation	Y	Y
IsProteinDomain	Indicates whether phosphosite is in a protein domain, data derived from Uniprot (UniProt Consortium, 2018).	Missing values set to FALSE	Y	Y

IsProteinKinaseDomain	Indicates whether phosphosite is in a kinase domain, data derived from Uniprot (UniProt Consortium, 2018).	Missing values set to FALSE	N	Y
IsUniprotRegion	Indicates whether phosphosite is in a protein domain, data derived from Uniprot (UniProt Consortium, 2018).	Missing values set to FALSE	N	Y
IsUniprotCompBias	Indicates whether phosphosite is found in a compositionally biased region, data derived from UniProt (UniProt Consortium, 2018).	Missing values set to FALSE	N	N
IsUniprotRepeat	Indicates whether phosphosite is found within a repeated motif, data derived from Uniprot (UniProt Consortium, 2018).	Missing values set to FALSE	N	N
IsUniprotZnFinger	Indicates whether phosphosite is in a zinc finger, data derived from Uniprot (UniProt Consortium, 2018).	Missing values set to FALSE	N	N
IsCytoplasmic	Indicates whether phosphosite's protein is found in the cytoplasm, data derived from Uniprot (UniProt Consortium, 2018).	Missing values set to FALSE	Y	Y

IsMotif	Indicates whether phosphosite is found in any other curated motif, data derived from Uniprot.	Missing values set to FALSE	N	N
IsELMLinearMotif	Indicates if flanking motif is found in a linear motif listed in ELM (Dinkel et al., 2016).	Missing values set to FALSE	Y	Y
IsEV_ala_prediction_epistatic5	Indicates If ala mutation leads to epistatic effects. Calculated with the EVmutation algorithm (Hopf et al., 2017).	Missing values set to FALSE	N	Y
IsKinaseCoreg	Indicates if phosphosite is co-regulated with a kinase. Data used for analysis from compilation of phosphoproteomic data with 435 conditions (Ochoa et al., 2016).	Missing values set to FALSE	Y	Y

Appendix 3.2

Kinase-substrate relationships that are predicted by experimental kinase-substrate prediction (Chapter 3.3.3).

Kinase	Substrate	Probability
AKT1	ABLIM1 452	0.1995
AKT1	ARHGEF12 1288	0.1942
AKT1	MEPCE 152	0.2902
AKT2	ARHGEF12 1288	0.0611
AKT2	HDGFL2 454	0.0469
AKT2	NDRG1 367	0.0623
AKT2	PLEC 4386	0.7158
AKT3	MEPCE 152	0.032
CAMK2A	ARFGAP2 432	0.5819
CAMK2D	ARFGAP2 432	0.132
CAMK2D	CEP170 881	0.0158
CAMK2D	EIF4B 207	0.031
CAMK2D	HNRNPM 528	0.0848
CAMK2G	ARFGAP2 432	0.7265
CAMK2G	EIF4B 207	0.4385
CDK1	PDCD4 94	0.8022
CDK2	AHNAK 177	0.043
CDK2	DDX21 71	0.7653
CDK2	EFHD2 74	0.7721
CDK2	FAM122B 115	0.8951
CDK2	HNRNPA2B1 259	0.8051
CDK2	MEF2D 180	0.3348
CDK2	NFIC 323	0.8574
CDK2	NOP2 732	0.8644
CDK2	NUMA1 2000	0.9232
CDK2	PDCD4 94	0.626

CDK2	UBAP2L 416	0.785
CDK5	MARK2 619	0.6949
CDK5	PDCD4 94	0.6782
CDK6	MEPCE 213	0.2002
CDK6	RBMX 208	0.2908
CDK9	EFHD2 74	0.0756
CDK9	MEPCE 213	0.1441
CDK9	MEPCE 217	0.4568
CDK9	RBMX 208	0.1526
CDK9	ZC3HAV1 378	0.0314
CSNK1E	EEF1D 147	0.0669
CSNK1E	PDXDC1 718	0.0666
CSNK2A2	FXR2 411	0.0075
CSNK2A2	UPF3B 169	0.0145
MAP3K1	PRKAR2A 99	0.1698
MAP4K4	RPLP1 104	0.184
MAP4K4	RPLP2 105	0.2103
MAP4K4	UFD1 299	0.0628
MAP4K5	EIF3A 584	0.0793
MAP4K5	IMPDH2 416	0.4189
MAPK1	EEF1D 162	0.2224
MAPK1	WAPL 221	0.5611
MAPK1	WAPL 226	0.3193
MAPK1	XRCC1 453	0.7276
MAPK3	WAPL 221	0.4199
MAPK3	WAPL 226	0.4219
MAPK9	ARID1A 363	0.5973
MAPK9	EFHD2 74	0.8148
MAPK9	RANBP2 1396	0.4317
MAPK9	WBP11 237	0.2654
MELK	RANBP1 60	0.7485

MINK1	EIF3A 584	0.0228
MINK1	HNRNPAB 242	0.0294
MINK1	HNRNPM 365	0.0092
MINK1	NPM1 260	0.056
MINK1	UFD1 299	0.0211
PAK1	EIF3A 584	0.0566
PAK1	NUCB1 369	0.1064
PAK3	EIF3A 584	0.0213
PLK1	KPNA3 60	0.7032
PLK1	UBA1 820	0.4098
PRKAA1	PPP1R12A 445	0.3224
PRKACA	HDGFL2 454	0.3345
PRKACB	HDGFL2 454	0.1738
PRKCI	PRKAR2A 99	0.2646
RPS6KA2	EIF4H 21	0.341
RPS6KA2	NCBP1 22	0.7327
RPS6KA2	NDRG3 331	0.7722
RPS6KA3	NCBP1 22	0.6674
SRPK3	HNRNPK 284	0.1012

Appendix 4.1

Top association between module enrichment odds ratio and transcription factor activities (Chapter 4.3.6).

Module	**TF**	**Spearman's *rho***	**p-value**
1	E2F1	0.2839	4.34E-19
2	E2F1	0.44819	3.56E-48
3	E2F1	0.50353	2.82E-62
4	E2F1	0.35848	3.25E-30
5	E2F1	0.37734	1.52E-33
6	E2F4	0.47739	2.71E-55
7	E2F1	0.38339	1.16E-34
8	E2F4	0.49037	1.10E-58
9	E2F1	0.28622	2.18E-19
10	E2F1	0.40358	1.47E-38
11	E2F1	0.44118	1.44E-46
12	E2F1	0.35405	1.82E-29
13	E2F1	0.35096	5.99E-29
14	E2F1	0.37464	4.70E-33
15	E2F4	0.39572	5.21E-37
16	E2F1	0.45556	6.60E-50
17	E2F1	0.33224	6.08E-26
18	E2F1	0.38042	4.13E-34
19	E2F1	0.43719	1.14E-45
20	E2F4	0.44689	7.12E-48
21	E2F4	0.49415	1.06E-59
22	E2F4	0.45269	3.15E-49
23	E2F1	0.4151	6.63E-41
24	E2F1	0.45386	1.67E-49
25	E2F4	0.50994	4.40E-64
26	E2F1	0.3677	8.15E-32

27	E2F4	0.42722	1.79E-43
28	E2F1	0.40789	2.00E-39
29	E2F4	0.45919	8.91E-51
30	FOXM1	0.40671	3.47E-39
31	E2F1	0.36964	3.69E-32
32	FOXM1	0.40609	4.63E-39
33	E2F1	0.43162	1.96E-44
34	E2F1	0.41255	2.24E-40
35	E2F1	0.42121	3.47E-42
36	E2F1	0.45559	6.50E-50
37	E2F1	0.39537	6.11E-37
38	E2F1	0.40634	4.12E-39
39	E2F1	0.41392	1.17E-40
40	E2F1	0.43104	2.62E-44
41	E2F1	0.41101	4.62E-40
42	E2F4	0.55496	6.38E-78
43	E2F1	0.43966	3.19E-46
44	E2F4	0.56799	2.45E-82
45	E2F1	0.40104	4.72E-38
46	E2F1	0.41961	7.53E-42
47	E2F1	0.4198	6.88E-42
48	E2F4	0.42698	2.01E-43
49	E2F1	0.45439	1.25E-49
50	E2F1	0.40835	1.61E-39
51	E2F1	0.46981	2.23E-53
52	E2F1	0.46734	9.15E-53
53	FOXM1	0.40762	2.27E-39
54	E2F1	0.38956	8.01E-36
55	E2F1	0.49449	8.57E-60
56	E2F1	0.45929	8.41E-51
57	E2F1	0.36627	1.45E-31

58	E2F1	0.4245	6.90E-43
59	FOXM1	0.4015	3.82E-38
60	E2F1	0.39685	3.15E-37
61	E2F1	0.42908	7.03E-44
62	E2F4	0.34816	1.74E-28
63	E2F1	0.37465	4.67E-33
64	E2F1	0.35515	1.19E-29
65	FOXM1	0.39422	1.02E-36
66	E2F1	0.27969	1.50E-18
67	E2F4	0.41072	5.31E-40
68	FOXM1	0.34078	2.74E-27
69	E2F1	0.2867	1.89E-19
70	E2F1	0.30712	3.20E-22
71	E2F1	0.36149	9.87E-31
72	E2F1	0.31887	6.44E-24
73	E2F1	0.36299	5.42E-31
74	FOXM1	0.35933	2.32E-30
75	E2F1	0.27516	5.55E-18
76	E2F1	0.38734	2.11E-35
77	E2F4	0.44946	1.80E-48
78	E2F4	0.27232	1.24E-17
79	E2F1	0.26103	2.80E-16
80	E2F1	0.22877	9.30E-13
81	E2F1	0.33544	1.93E-26
82	FOXM1	0.30136	2.04E-21
83	PRDM14	0.19306	1.94E-09
84	E2F1	0.23057	6.09E-13
85	E2F1	0.41845	1.33E-41
86	E2F4	0.46023	5.02E-51
87	MYC	0.21659	1.47E-11
88	FOXM1	0.34676	2.95E-28

89	FOXM1	0.31395	3.38E-23
90	E2F1	0.37273	1.04E-32
91	MYC	0.31813	8.29E-24
92	FOXM1	0.26778	4.44E-17
93	FOXM1	0.30191	1.71E-21
94	E2F1	0.33014	1.29E-25
95	E2F1	0.29931	3.90E-21
96	E2F1	0.24998	5.14E-15
97	E2F1	0.28494	3.19E-19
98	SOX2	0.10227	0.00159
99	E2F4	0.36372	4.05E-31
100	E2F4	0.34814	1.75E-28
101	E2F1	0.30409	8.52E-22
102	E2F4	0.379	7.55E-34
103	E2F4	0.35831	3.47E-30
104	E2F4	0.48264	1.20E-56
105	E2F1	0.31632	1.53E-23
106	E2F1	0.3041	8.49E-22
107	E2F1	0.33288	4.85E-26
108	E2F1	0.33852	6.29E-27
109	FOXM1	0.37537	3.46E-33
110	E2F4	0.37523	3.67E-33
111	E2F4	0.24672	1.18E-14
112	E2F1	0.22755	1.23E-12
113	E2F1	0.33812	7.28E-27
114	MYC	0.3127	5.13E-23
115	E2F1	0.23872	8.65E-14
116	E2F1	0.31002	1.24E-22
117	PRDM14	0.39361	1.34E-36
118	PRDM14	0.26871	3.42E-17
119	E2F1	0.31391	3.42E-23

120	E2F4	0.29543	1.31E-20
121	FOXM1	0.27723	3.06E-18
122	FOXM1	0.35551	1.04E-29
123	E2F1	0.36524	2.20E-31
124	E2F1	0.38263	1.61E-34
125	FOXM1	0.36707	1.05E-31
126	E2F1	0.35981	1.92E-30
127	FOXM1	0.4181	1.57E-41
128	E2F1	0.38776	1.76E-35
129	SOX2	0.30266	1.35E-21
130	E2F1	0.46463	4.25E-52
131	E2F1	0.40001	7.52E-38
132	E2F1	0.37556	3.20E-33
133	E2F1	0.32365	1.25E-24
134	E2F1	0.45027	1.17E-48
135	E2F1	0.455	8.97E-50
136	E2F1	0.40174	3.42E-38
137	E2F1	0.26574	7.78E-17
138	E2F1	0.18165	1.69E-08
139	E2F1	0.31374	3.62E-23
140	E2F1	0.36728	9.63E-32
141	E2F1	0.40455	9.41E-39
142	MYC	0.35335	2.39E-29
143	E2F1	0.31397	3.36E-23
144	FOXM1	0.32666	4.38E-25
145	E2F1	0.27658	3.69E-18
146	E2F1	0.4139	1.17E-40
147	FOXM1	0.34943	1.08E-28
148	E2F1	0.35438	1.61E-29
149	E2F1	0.39335	1.50E-36
150	FOXM1	0.2744	6.89E-18

151	E2F1	0.41418	1.03E-40
152	E2F1	0.37169	1.59E-32
153	E2F1	0.43379	6.52E-45
154	E2F1	0.31109	8.72E-23
155	E2F1	0.46964	2.46E-53
156	E2F1	0.40978	8.26E-40
157	E2F1	0.27123	1.69E-17
158	E2F1	0.40098	4.86E-38
159	FOXM1	0.4575	2.28E-50
160	E2F1	0.24162	4.23E-14
161	E2F1	0.29995	3.18E-21
162	E2F1	0.50734	2.41E-63
163	SOX2	0.31855	7.17E-24
164	E2F1	0.42533	4.56E-43
165	E2F1	0.40105	4.70E-38
166	E2F1	0.4378	8.33E-46
167	E2F1	0.3481	1.78E-28
168	E2F1	0.34063	2.90E-27
169	E2F1	0.32445	9.47E-25
170	E2F1	0.26481	1.01E-16
171	E2F1	0.36261	6.32E-31
172	E2F1	0.27222	1.28E-17
173	E2F1	0.39643	3.79E-37
174	E2F1	0.39328	1.55E-36
175	E2F4	0.41155	3.59E-40
176	E2F1	0.34804	1.82E-28
177	E2F1	0.47436	1.60E-54
178	E2F1	0.40063	5.67E-38
179	E2F1	0.3663	1.43E-31
180	E2F1	0.33589	1.64E-26
181	E2F1	0.33615	1.49E-26

182	E2F1	0.42947	5.81E-44
183	E2F1	0.30201	1.66E-21

Appendix 4.2

Top correlation between modules and kinase activities (Chapter 4.3.7).

Module	**Kinase**	**Spearman's *rho***	**p-value**
1	AURKC	0.32553	0.0291
2	CDK2	0.22971	3.35E-13
3	WEE1	0.31973	1.09E-07
4	AURKC	0.31683	0.03396
5	WEE1	0.26487	1.29E-05
6	AURKC	0.48167	0.00081
7	HIPK1	0.27235	2.88E-09
8	CDK2	0.24939	2.32E-15
9	AURKC	0.36565	0.01351
10	CDK2	0.23309	1.47E-13
11	CSNK2A1	0.21456	1.14E-11
12	AURKC	0.35656	0.01621
13	AURKC	0.32658	0.02856
14	AURKC	0.34371	0.02079
15	AURKC	0.3866	0.00871
16	AURKC	0.32309	0.0304
17	MKNK1	0.21052	0.01222
18	AURKC	0.41691	0.00439
19	HIPK1	0.25642	2.43E-08
20	CDK2	0.22789	5.19E-13
21	CDK2	0.26033	1.21E-16
22	WEE1	0.28496	2.53E-06
23	AURKC	0.36268	0.01435
24	AURKC	0.40808	0.00539
25	AURKC	0.36802	0.01287
26	CSNK2A1	0.23465	1.00E-13
27	AURKC	0.2943	0.04972
28	WEE1	0.24176	7.23E-05
29	HIPK1	0.2491	6.17E-08
30	EIF2AK2	0.25281	0.00861
31	EIF2AK2	0.24426	0.01123
32	CDK2	0.26948	9.12E-18
33	CDK7	0.24663	7.43E-13
34	AURKC	0.3196	0.03235

35	AURKC	0.31334	0.03609
36	WEE1	0.26469	1.31E-05
37	CDK7	0.30593	2.87E-19
38	AURKC	0.40175	0.00623
39	AURKC	0.35583	0.01644
40	AURKC	0.30339	0.04277
41	WEE1	0.18875	0.00207
42	AURKC	0.31854	0.03296
43	CSNK2A1	0.31162	1.64E-23
44	CDK2	0.34874	2.10E-29
45	CDK2	0.22468	1.11E-12
46	CSNK2A1	0.15952	5.18E-07
47	CDK2	0.26952	9.00E-18
48	AURKC	0.32632	0.02869
49	HIPK1	0.24246	1.40E-07
50	HIPK1	0.268	5.23E-09
51	CSNK2A1	0.29796	1.53E-21
52	AURKC	0.3644	0.01386
53	AURKC	0.3364	0.02386
54	WEE1	0.26464	1.32E-05
55	AURKC	0.35583	0.01644
56	AURKC	0.35932	0.01534
57	CDK2	0.15637	8.70E-07
58	WEE1	0.32771	5.02E-08
59	AURKC	0.47284	0.00104
60	AURKC	0.31024	0.03807
61	CDK7	0.26557	9.80E-15
62	AURKC	0.37131	0.01203
63	AURKC	0.4144	0.00465
64	AURKC	0.47073	0.0011
65	WEE1	0.25992	1.90E-05
66	SRPK1	0.17858	0.00261
67	EIF2AK2	0.31449	0.00097
68	HIPK1	0.27647	1.63E-09
69	CDK2	0.18174	1.00E-08
70	WEE1	0.23921	8.66E-05
71	PTK2	0.23937	0.03997
72	WEE1	0.18665	0.00233
73	CDK7	0.21441	5.29E-10

74	HIPK1	0.33348	2.07E-13
75	AURKC	0.37988	0.01006
76	AURKC	0.43384	0.00291
77	AURKC	0.44181	0.00238
78	HIPK1	0.18959	4.27E-05
79	AURKC	0.55552	7.46E-05
80	AURKC	0.32849	0.02759
81	EIF2AK2	0.21131	0.0289
82	AURKC	0.42283	0.00381
83	BCR/ABL	0.18979	0.00642
84	AURKC	0.29562	0.04866
85	AURKC	0.43298	0.00297
86	EIF2AK2	0.2115	0.02875
87	AURKC	0.36005	0.01512
88	CSNK2A1	0.23903	3.37E-14
89	YES1	0.21388	0.00365
90	WEE1	0.23192	0.00014
91	AURKC	0.4175	0.00432
92	AURKC	0.3364	0.02386
93	AURKC	0.2995	0.04564
94	LATS2	0.17163	0.04185
95	AURKC	0.35913	0.0154
96	CDK7	0.09951	0.00429
97	AURKC	0.36618	0.01337
98	PTK2	0.3127	0.00668
99	YES1	0.27107	0.00021
100	AURKC	0.36683	0.01319
101	AURKC	0.38383	0.00924
102	TTK	0.23619	1.84E-06
103	AURKC	0.33903	0.02271
104	ILK	0.20352	0.04556
105	HIPK1	0.25676	2.33E-08
106	AURKC	0.35886	0.01549
107	CDK7	0.21093	1.02E-09
108	AURKC	0.36321	0.01419
109	HIPK1	0.28654	3.84E-10
110	HIPK1	0.31057	9.63E-12
111	PTK2	0.23479	0.04405
112	ATM	0.16992	1.04E-06

113	AURKC	0.32098	0.03157
114	AURKC	0.398	0.00678
115	AURKC	0.39259	0.00764
116	CSNK2A1	0.2048	9.68E-11
117	LATS2	0.24367	0.00359
118	CAMKK1	0.22117	0.00994
119	CSNK2A2	0.19438	3.06E-07
120	HIPK1	0.212	4.50E-06
121	CAMKK1	0.21077	0.01414
122	EIF2AK2	0.24054	0.01257
123	AURKC	0.38759	0.00852
124	HIPK1	0.24941	5.94E-08
125	AURKC	0.34167	0.02161
126	CSNK2A2	0.2734	3.57E-13
127	WEE1	0.28382	2.78E-06
128	CDK7	0.15301	1.05E-05
129	AURKC	0.31123	0.03743
130	AURKC	0.29509	0.04908
131	CDK7	0.18788	5.73E-08
132	WEE1	0.28196	3.25E-06
133	AURKC	0.37125	0.01205
134	AURKC	0.35755	0.0159
135	CDK7	0.22279	1.05E-10
136	AURKC	0.37922	0.0102
137	WEE1	0.13062	0.03389
138	GRK6	0.15501	0.00695
139	AURKC	0.44866	0.00199
140	AURKC	0.35432	0.01694
141	AURKC	0.353	0.01738
142	HIPK1	0.28588	4.22E-10
143	HIPK1	0.24885	6.37E-08
144	AURKC	0.49939	0.00048
145	AURKC	0.39134	0.00785
146	AURKC	0.5882	2.15E-05
147	HIPK1	0.28671	3.74E-10
148	HIPK1	0.18841	4.77E-05
149	AURKC	0.43173	0.00306
150	MAPK7	0.25516	5.51E-11
151	CSNK2A1	0.19839	3.72E-10

152	WEE1	0.30157	5.93E-07
153	AURKC	0.36683	0.01319
154	YES1	0.30339	2.98E-05
155	AURKC	0.3671	0.01312
156	CDK2	0.22619	7.78E-13
157	EIF2AK2	0.21653	0.02508
158	AURKC	0.31841	0.03303
159	CSNK2A1	0.2607	1.09E-16
160	AURKC	0.37915	0.01021
161	EPHA2	0.3111	0.01166
162	AURKC	0.36321	0.01419
163	CHUK	0.12354	0.00086
164	CDK2	0.18957	2.22E-09
165	AURKC	0.39306	0.00756
166	CDK7	0.29825	2.38E-18
167	WEE1	0.1631	0.00792
168	AURKC	0.31215	0.03684
169	WEE1	0.27416	6.17E-06
170	PKM	0.19071	1.57E-05
171	AURKC	0.30747	0.03992
172	BCR/ABL	0.14236	0.04173
173	CDK7	0.23834	4.42E-12
174	AURKC	0.44148	0.0024
175	CDK7	0.25231	2.10E-13
176	AURKC	0.36242	0.01442
177	AURKC	0.42369	0.00373
178	AURKC	0.36216	0.0145
179	CDC7	0.22902	3.66E-12
180	AURKC	0.31953	0.03239
181	AURKC	0.39747	0.00686
182	AURKC	0.53194	0.00017
183	AURKC	0.44991	0.00193

Appendix 4.3

Traits that are significantly associated with data-driven modules (Chapter 4.3.8).

Trait	Module	P-value
HIP	152	0.03837381009
Height	44	0.00237109901
Height	58	0.02156365714
Height	159	0.02223752143
Height	130	0.004024890756
Height	141	0.01417822836
Height	178	0.006069039041
Height	42	0.0232802247
Height	123	0.007881653165
Height	1	0.004371478571
Height	151	0.02525776172
Height	8	0.03912647324
Height	160	0.04684224939
Height	125	0.02931724698
Height	150	0.03769608333
Height	112	0.04237465416
Height	166	0.04575691563
HIP (female)	32	0.0180170285
BMI (male, WC-adj)	10	0.04079071978
Weight women	161	0.0232802247
Neuroticism	138	0.0383459003