THE UNIVERSITY OF QUEENSLAND

Molecular interaction motifs in a system-wide network context: Computationally charting transient kinase-substrate phosphorylation events

Ralph Patrick
Bachelor of Science (Hons) in Mathematics and Computational Biology

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2015

The School of Chemistry and Molecular Biosciences

UNIVERSITY OF QUEENSLAND

Abstract

Molecular interaction motifs in a system-wide network context: Computationally charting transient kinase-substrate phosphorylation events

by Ralph Patrick

Protein phosphorylation is the most ubiquitous of post-translational modifications, regulating a wide variety of essential functions from cell-cycle progression through to DNA damage repair. Phosphorylation is regulated by the kinases – a super-family of proteins that comprise the third largest protein family in the human genome. While advances in high-throughput mass spectrometry have resulted in the identification of hundreds of thousands of phosphorylation sites, the identification of the kinases that regulate these phosphorylation events has largely remained elusive. Understanding the kinases responsible for phosphorylation events is often crucial for understanding the function of the modification, however the transient nature of kinase binding means that identifying genuine kinase-binding events *in vivo* is both difficult and expensive.

The vast majority of methods for computationally predicting kinase binding targets rely primarily on sequence features. A lack of specificity in many kinase-binding motifs means that valid binding patterns can be found randomly throughout the proteome – leaving such methods susceptible to high false-positive rates. However, the determinants of phosphorylation are not limited to the sequence; kinases are regulated through various cellular processes including mediating/activating proteins, localisation and cell cycle-specific expression. While such information has increasingly become accessible through proteomic databases, incomplete coverage, variable certainty and the heterogeneous nature of context and sequence information means that the integration of relevant features into a computational model is non-trivial.

In this thesis I present a method for the probabilistic integration of these two aspects of kinase regulation – context and sequence – into a Bayesian network model that can accurately predict kinase substrates. In the first part of the thesis I demonstrate how a model that incorporates knowledge of kinase-substrate phosphorylation, protein interactions and protein abundance

across the cell cycle can be used to classify kinase substrates. The model achieves high level of prediction accuracy as determined by cross-validation, obtaining an average AUC of 0.86 across all kinases tested. When applying the model to complement sequence-based kinase-specific phosphorylation site prediction using previously published methods, I find it improves prediction performance for most comparisons made. As a validation of these ideas, I also show how protein interaction networks can be coupled with gene expression data to predict changes in phosphorylation status in response to varying cell treatment conditions.

To integrate kinase-binding affinity into the modelling framework, I present a method for classifying kinase-binding sites from sequence, which captures features from the linear motifs surrounding known kinase-specific phosphorylation sites. This method incorporates observed position-specific amino acid frequencies and counts of co-occurring neighbouring amino acids into a Bayesian network model. The model is trained to discriminate between a kinase's binding profile, that of its family members, and a phosphorylation background. I show how this sequence model can be integrated as a module into the larger context model, allowing for a comprehensive description of the factors that influence kinase binding. This seamless integration of context and sequence increases kinase-substrate prediction accuracy, when compared to the first context model, by over 50% at low false-positive levels. I find that this system of predicting kinase substrates, coupled with predicting kinase binding sites from sequence, convincingly outperforms existing kinase-specific phosphorylation site classifiers; a comparison of prediction accuracy at strict specificity levels shows that my method predicts kinase-specific phosphorylation sites with an average of 9-22% greater sensitivity (at a strict specificity level of 99.9%) than the alternatives. The method, named PhosphoPICK, has been made freely available as a web-service.

Possessing a predictor that ably integrates the context and sequence conditions that regulate phosphorylation allows an approach to problems in phosphorylation that were not feasible previously. Non-synonymous single nucleotide polymorphisms (nsSNPs) have the potential to disrupt (or introduce) kinase binding sites through the modification of key amino acids that mediate kinase activity. To validate that PhosphoPICK accurately represents the biological characteristics determining phosphorylation occurrence, I developed a method applying PhosphoPICK to predict variant-causing phosphorylation loss and gain. The method quantifies the

expected effect of a nsSNP on phosphorylation based on predictions from the sequence model, and the probability that a query kinase will target the variant protein. Employing distributions of predicted variants across the proteome, the method can provide a measure of the significance of novel variants. Evaluating the method on known examples of variants causing phosphorylation loss or gain from the literature, I show that PhosphoPICK can detect the positive examples at strict specificity levels.

While the methodology presented in this work was developed for phosphorylation, it should be considered a framework that could be applied to alternative biological processes. Sequence motifs and protein interactions are necessary elements for a spectrum of biology, including post-translational modifications other than phosphorylation. The short ubiquitin-like modifier (SUMO), for example, operates on defined sequence motifs, but is also highly dependent on the context factors that SUMO substrates operate in. The methods I describe allow an approach to alternative protein prediction problems, such as SUMOylation, where the integration of context and sequence characteristics can provide a comprehensive description of the relevant regulatory features.

Declaration by Author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

1. **Patrick R**, Lê Cao KA, Kobe B and Bodén M (2015) PhosphoPICK: modelling cellular context to map kinase-substrate phosphorylation events. *Bioinformatics*, 31(3), 382-389

2. Mehdi AM, **Patrick R**, Bailey TL and Bodén M (2014) Predicting the dynamics of protein abundance, *Molecular & Cellular Proteomics*. May;13(5):1330-40

Publications included in the thesis

Patrick R, Lê Cao KA, Kobe B and Bodén M (2015) PhosphoPICK: modelling cellular context to map kinase-substrate phosphorylation events. *Bioinformatics*, 31(3), 382-389 – incorporated as Chapter 2

Contributor	Statement of contribution
Author RP (Candidate)	Designed experiments (30%) Carried out experiments Wrote the paper
Author KAL	Designed experiments (30%)
Author BK	Designed experiments (10%)
Author MB	Designed experiments (30%)

Contributions by others to the thesis

The work reported in this thesis was carried out under the primary supervision of Associate Professor Mikael Bodén, and under the secondary supervision of Dr. Kim-Anh Lê Cao and Professor Bostjan Kobe. As such, they contributed to the design of the project, and the design of the experiments carried out as part of it.

Statement of parts of the thesis submitted to qualify for the award of another degree

None

Acknowledgements

I am very thankful to many people who have made this PhD thesis not just possible, but a fun and rewarding experience. First and foremost, I owe a debt of gratitude to my primary supervisor Mikael Bodén, who has been a patient teacher, and guided my research and writing efforts for many years now. I am also thankful to my co-supervisors, Kim-Anh Lê Cao and Bostjan Kobe. Kim-Anh, for her support and guidance in weekly meetings over the years, for her statistics advice, and for the many paper drafts that she has read and provided comment on. I'm also thankful to Bostjan for his advice on the biology side of things, and his invaluable contributions to papers. I also need to thank the past and present members of the Boden lab. Working environment can make a big difference, and I have been fortunate to be a part of a supportive and friendly group. While I have not spent a lot of time with the Kobe lab, I am very appreciative for the times I have been able to present at their group meetings and get feedback from people who have a different perspective to bioinformaticians.

Thanks is also due to the University of Queensland for providing a scholarship. Without their financial assistance, this thesis would not have been possible. I am also thankful to the School of Chemistry and Molecular Biosciences for generous financial assistance towards attending numerous conferences, both in Australia and overseas. In addition, I am thankful to the graduate school for providing an international travel scholarship to visit Burkhard Rost's lab at the Technical University of Munich. This allowed me to spend some time with an overseas lab, and work on some interesting biological problems that would not have been possible without the travel scholarship. I am also thankful to Burkhard Rost and his lab for their welcome, and allowing me to be a part of their group for a month.

On a personal note, I want to thank all the friends and family who have supported me throughout the duration of my PhD. I owe a special thanks to my long-suffering wife, Lauren, who patiently endured as a "thesis widow" during the final months of bringing the thesis to an end. Thanks also to my parents for being supportive throughout the PhD, and the many friends and other family members who have made life during the course of the PhD much more fun than it would be without them. The greatest thanks is owed to the God who "makes all things beautiful in their time". Including, I believe, a PhD thesis.

Keywords

bioinformatics, machine learning, bayesian networks, data integration, phosphorylation, kinases, systems biology

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060102, Bioinformatics, 60%
ANZSRC code: 060109, Proteomics and Intermolecular Interactions, 20%
ANZSRC code: 060114, Systems Biology, 20%

Fields of Research (FoR) Classification

FoR code: 0601, Biochemistry and Cell Biology, 100%

Contents

List of Figures

List of Tables

Abbreviations

ANN	**A**rtificial **N**eural **N**etwork
AUC	**A**rea **U**nder the **C**urve
BAC	**B**alanced **Ac**curacy
BN	**B**ayesian **N**etwork
CPT	**C**onditional **P**robability **T**able
CDF	**C**umulative **D**istribution **F**unction
KNN	**K**-**N**earest **N**eighbour
NB	**N**uclear **B**ody
NN	**N**eural **N**etwork
PDB	**P**rotein **D**ata **B**ank
PML	**P**ro**m**yelocytic **L**eukemia
PSSM	**P**osition **S**pecific **S**coring **M**atrix
PPI	**P**rotein-**P**rotein **I**nteraction
PTM	**P**ost-**T**ranslational **M**odification
PWM	**P**osition **W**eight **M**atrix
RF	**R**andom **F**orest
ROC	**R**eceiver **O**perating **C**haracteristic
SNP	**S**ingle **N**ucleotide **P**olymorphism
SUMO	**S**hort **U**biquitin-like **Mo**difier
SVM	**S**upport **V**ector **M**achine
TF	**T**ranscription **F**actor

Chapter 1

Introduction and overview

1.1 Introduction

Advanced "omics" technologies are rapidly transforming the proteomic research landscape, with a variety of proteomic and genomic databases recording hundreds of thousands of molecular interactions (1–3), post-translational modifications (4–9) and functional annotations (10–13). This increasingly massive and diverse amount of data requires development of computational methods that can integrate and analyse complex biological data – thus allowing for the kind of observations and hypothesis testing that only a systems approach to biology allows.

Many important biological processes and functions involve, or are regulated by a variety of post-translational modifications. For example, acetylation is a co-regulator of major cellular functions including chromatin remodelling, nuclear transport and protein degradation (14, 15) while glycosylation is involved in protein folding, localisation and trafficking amongst other things (16). Protein phosphorylation is the most ubiquitous post-translational modification and has regulatory roles in a wide array of biologically important functions from DNA damage repair (17) through to the control of cell-cycle progression. As a consequence of this, there has been great interest in identifying protein phosphorylation events, with advanced phosphoproteomic technologies successful in identifying hundreds of thousands of phosphorylation sites across multiple proteomes. The protein enzymes – kinases – responsible for these modifications have generally remained elusive; however it is the assignment of kinases to phosphorylation sites that can give insight into the biological pathway that a site may be involved in. This has resulted in many attempts to build computational methods for predicting phosphorylation sites and the kinases that are responsible for them (18).

Broadly speaking, there are two domains of information that need to be considered when seeking to understand the regulation of kinase-mediated phosphorylation. The first is linear motifs – short regions of amino acids that allow interactions between proteins and are necessary for kinases to bind their target substrates. The second domain of information concerns what may be termed the "context factors" that regulate kinase activity at the wider cellular level. Given the highly specific functions that kinases regulate, it is essential that the activity of kinases themselves be tightly controlled (19). Kinases are subject to a range of regulatory mechanisms, such as activating or mediating proteins (20), cell cycle stage-specific expression and sub-cellular localisation (21). Furthermore, "cross-talk" between post-translational modifications adds another element of regulatory complexity; for example, phosphorylation can act as a promoter or inhibitor of ubiquitiniation (22), and likewise glycosylation can act as an inhibitor of phosphorylation (16). These diverse factors, in concert with the sequence-specificity of kinases, all contribute towards ensuring kinase-substrate fidelity.

The majority of existing methods for predicting kinase-specific phosphorylation sites have primarily focussed on modelling features within the linear motifs that surround phosphorylation sites. However, many motifs are non-specific and can be found at random in protein sequences, leading to the identification of numerous false-positives. In addition to motifs, there are several examples of predictors complementing the sequence data with other types of information contained within the protein. Protein features such as disorder (23) and surface accessibility (24) have been shown to improve model accuracy for predicting phosphorylation sites, while protein structure has been used to both inform the design of predictive methods (25), and supplement motif-based predictions of phosphorylation (26). While such approaches can identify valid kinase binding locations within a protein, even the presence of a perfect kinase binding motif is no guarantee that a kinase will come into contact with the protein (27). Despite these limitations, there has been very little work invested in developing methods to analyse the context factors that regulate kinases.

There are numerous examples of context information, with high coverage across the proteome, that could be leveraged to build computational methods for predicting kinase substrates. Protein-protein interactions (PPIs) are relevant to essentially all biological processes, and huge numbers of them have been recorded in databases. In particular, PPI networks are an excellent source of information on the molecular context that proteins operate in. PPI networks contain a unique capacity to identify proteins that interact with, and perhaps mediate between, kinases and their substrates. Other information concerning the "association" between proteins can be gleaned from gene co-expression studies, which are incorporated into the STRING database (2). Similarly, as phosphorylation is involved in cell cycle-specific processes, the incorporation

of information relating to cell-cycle progression would be useful in identifying cell cycle-specific kinase activity (28). The increasing availability of such proteome-wide data in publicly available databases provides a unique opportunity to leverage such information in computational methods.

The challenge here is two-fold. Firstly, the context information that could be used to describe a kinase's regulation at the systems level is highly diverse, and the information will not be available for all proteins. Secondly, modelling the context that a kinase operates in, and modelling its binding specificity are two very different problems. Therefore, the integration of context and sequence into a single model of phosphorylation is non-trivial.

This thesis proposes a novel computational framework based on probabilistic modelling to bridge the gap between these diverse sequence and context aspects of kinase regulation. I show that Bayesian networks are an ideal tool for such a task, allowing for the seamless integration of diverse types of information, and the handling of uncertain or missing data. The method works across species, with the ability to predict kinase substrates with high accuracy in three model organisms: human, mouse and yeast. While this work describes a method for the integration of information relevant to phosphorylation, the framework that I propose should be considered generic, with the potential to be applied to alternative post-translational modifications, or other biological functions where both linear motifs and context factors are relevant.

1.2 Kinase-mediated phosphorylation

Protein phosphorylation was first described as an enzyme-regulated process in 1954, when a liver enzyme was observed to catalyse the phosphorylation of caesin (29). Since then, the many studies involving phosphorylation have pointed to the modification as a central control mechanism underlying every essential biological process that cells undertake (30). While earlier studies estimated that approximately 30% of human proteins could be phosphorylation substrates, more recent work in phosphoproteomics has indicated a much higher figure of at least 70% (28), making phosphorylation a highly ubiquitous post-translational modification. A consequence of the central importance and pervasive nature of phosphorylation is that many diseases and cancers are related to aberrant phosphorylation events, with kinases having emerged as key drug targets (31).

Kinases are a protein superfamily containing over 500 identified members in human, and comprising the third most populated family of proteins (32). Kinases phosphorylate their target

substrates through the transfer of a phosphate group from an adenosine triphosphate (ATP) donor to (primarily) a serine (S), threonine (T) or tyrosine (Y) residue on the protein substrate. It is well documented that histidine residues also undergo phosphorylation in bacterial cells (33), though this is not known to be a common occurrence in eukaryotes. As the focus of this work is on eukaryotic phosphorylation, only S/T/Y phosphorylation will be considered.

Phosphorylation is likely a significant factor in understanding complex organisms, with phosphorylation of eukaryotic proteins showing a significant increase compared to prokaryotes in terms of numbers of phosphorylation sites (6). Indeed, the presence of large numbers of phosphorylation sites on a protein can result in high levels of regulatory flexibility, with a protein containing n phosphorylation sites having 2^n potential phosphorylation "states" that it can exist in (34). There are a plethora of examples of complex biological processes that could be explored to illustrate the important regulatory role of phosphorylation. The mitotic cell cycle is particularly illustrative example of regulation by kinase-meditated phosphorylation, and has been well studied (35).

1.2.1 Phosphorylation as a key regulator of the cell cycle

The mitotic cell-cycle is a highly regulated process in multi-cellular organisms, where correct cell numbers must be maintained and damaged cells restrained from replicating. Indeed, the definition of cancer is the situation where this process has become impaired, with cells containing irreparable DNA damage continuing to replicate unchecked. The cell cycle is divided up in to 5 main cell-cycle stages, the progression through which are controlled tightly by phosphorylation and mediated by a variety of kinases. The main drivers of the cell cycle are the cyclin-dependent kinases (CDKs), which perform key (though potentially overlapping) functions at specific stages of the cell cycle (36). There are a variety of other kinases that at specific stages, or under specific conditions, act to inhibit or activate the CDKs; there are further kinases that respond to damage and enable the organisation of the cell prior to the completion of mitosis.

The initial phase is the growth 1 (G1) phase, where the cell increases in size in preparation of DNA replication. G1 progression is mediated by the CDK4 and CDK6 kinases (37). The transition between G1 and synthesis (S) phase is crucial. A key driver of the G1/S phase transition is CDK2 in complement with cyclin E (38). As the cell transitions into S phase the expression of cyclin E decreases and cyclin A increases – forming the cyclin A-CDK2 complex that phosphorylates the DNA replication machinery. As S phase involves the duplication of the chromosomes, a highly delicate process, the DNA can suffer damage during replication (39).

In situations of DNA damage during replication the ATR kinase interacts with the replication machinery to halt S phase progression (40). The phase following S phase, growth 2 (G2), specifically checks the newly replicated DNA for damage prior entry into M phase There are a number of kinases that regulate the DNA damage response during G2, but key kinases are ATR and ATM, which activate Checkpoint kinase 1 (CHK1) and CHK2.

Phosphorylation is particularly ubiqitous during mitosis where many complex operations are required to take place in order to separate sister chromatid into separating cells. Various processes to facilitate this such as spindle formation, centrosome maturation/separation and chromosome attachment to the spindle are controlled by kinases (41, 42). The central driver behind mitosis is the cyclin B1-CDK1 kinase complex, whose activity can trigger different mitotic events (43). The number of cyclin B1-CDK1 complexes increases prior to mitosis, but they are kept inactive by the phosphorylation of CDK1 by the MYT1 and WEE1 kinases. The rapid dephosphorylation of CDK1 by phosphatases is a key signal for the start of mitosis, causing the complex to activate Polo kinase 1 (PLK1), the most tightly periodically expressed gene in the genome (44), which is essential for mitotic progression (if DNA damage is detected PLK1 will be deactivated until the damage is repaired). There are several other kinases such as Aurora kinases A and B that are involved in ordering and condensing chromosomes, and organising the mitotic spindle (45).

Even this brief overview of the role of phosphorylation in cell cycle progression should illustrate the fact that kinase activity must be highly specific, with kinases maintaining tight selectivity for target selection. While there is some level of redundancy that can be tolerated, aberrant functioning of several kinases has been linked to cancers – particularly kinases involved in DNA damage repair pathways (such as ATM/ATR) and those involved in arresting cell–cycle progression in case of irreparable DNA damage. The consequences of kinase malfunction should further underline the importance of understanding kinase activity and selectivity. I now turn to consider how it is that kinase-substrate specificity is maintained.

1.2.2 Kinase specificity and regulation

Given the role of kinases in regulating a wide array of biological processes through phosphorylation, it is critical that the kinases themselves be strictly regulated to maintain substrate fidelity. There are several characteristics of kinases and the wider cell that contribute to ensuring that kinases phosphorylate the correct substrates under the correct conditions. For the purpose of

this work, I specify two main categories of interest: the binding affinity of kinases for their substrates and the context factors that regulate kinase activity.

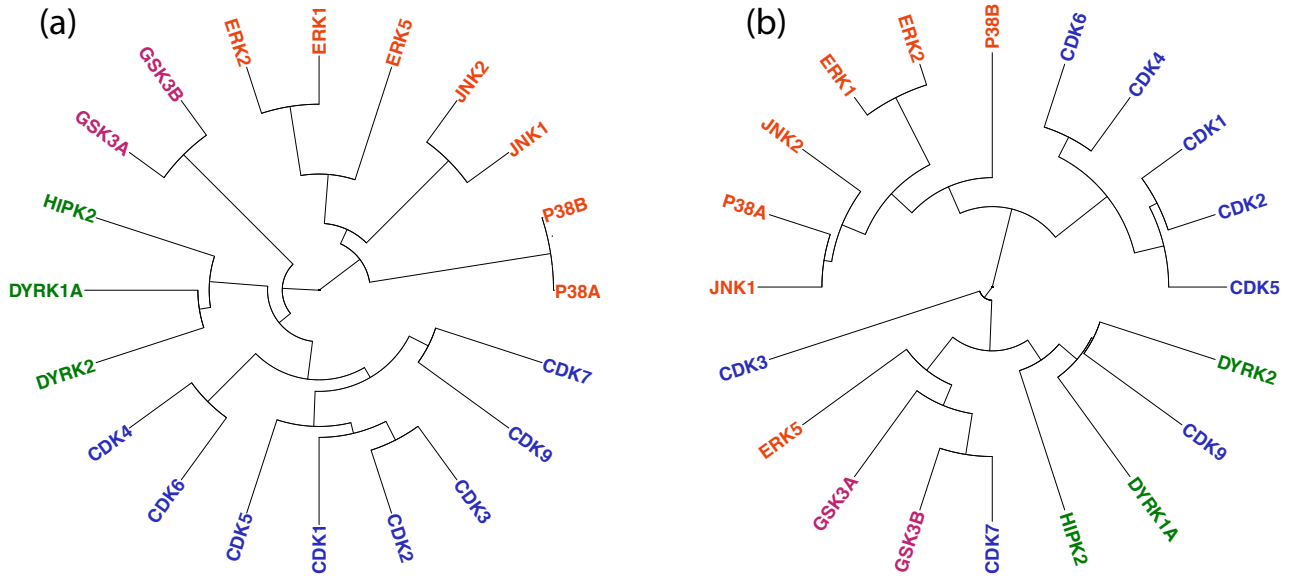

FIGURE 1.1: Dendrograms showing sequence similarity of kinases from their catalytic domains (a), and binding-specificity from their target peptides (b). Kinase domains were sourced from Uniprot (`www.uniprot.org`) and phosphorylation peptides were sourced from PhosphoSitePlus® (4). Kinases have been coloured according to the four sub-families clustered by catalytic domain sequence similarity (a).

Kinase-substrate binding affinity concerns the propensity of kinases to preferentially phosphorylate sites on proteins that contain a short pattern of amino acids, what is often termed a *linear motif*. The preference for a kinase to bind to a linear motif is determined by the catalytic domain of the kinase. Analysis of the 3D structure of kinases has shown that kinases contain short catalytic domains that bind to certain sequences on substrate proteins (46). Sequence similarity between kinases in these domains allows them to be organised into families and sub-families (32), with closely related kinases having a tendency to bind to similar sites. Figure 1.1 shows dendrograms of kinases within the CMGC family. In Figure 1.1(a) the kinases have been clustered according to the sequence similarity in their binding domains, and coloured according to sub-family. Figure 1.1(b) shows a dendrogram where the kinases have been clustered according to sequence similarity in their known phosphorylation target peptides. While there is not a perfect overlap between the two dendrograms, they demonstrate that kinases within the same sub-family will have a tendency to bind to similar sequence patterns.

Figure 1.2 shows examples of sequence logos for various kinase binding sites and the surrounding amino acids. Some kinase binding motifs appear to be unspecific – the proline-dependent kinases

such as CDK1 and ERK1 have a proline in the +1 position after the phosphorylation sites as their main recognition symbol. An [S/T]P motif can easily be found at random throughout the proteome, with almost 90% of human proteins containing the motif. Protein kinase A (PKA), which appear to preferentially bind to a motif of the form [RK][RK]X[ST] can be found in 53% of human proteins. The ATM motif, with a glutamine at the +1 position, can be found in 88% of humans proteins. The AurB motif, which appears to have a preference for an arginine at the -2 position, is seen in 87% of proteins. While a fixed motif is not the ideal way to predict kinase binding sites (as explored in Section 1.5), this illustrates the fact that amino acid motifs that could represent valid kinase binding sites can be found in a large proportion of the proteome.

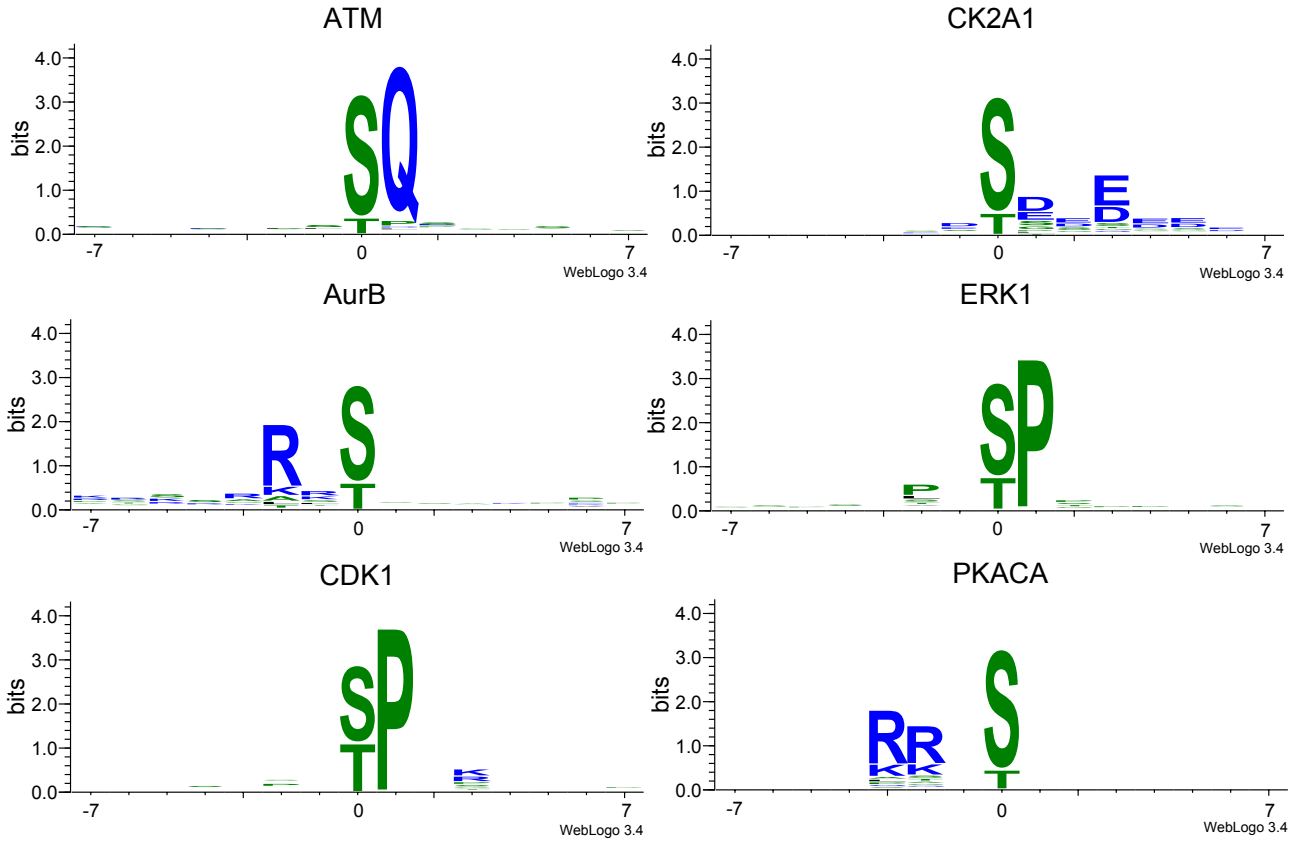

FIGURE 1.2: Sequence logos representing various kinase binding specificities in a 15 residue window surrounding the phosphorylation site (position 0). Phosphorylation peptides were sourced from PhosphoSitePlus® (4) and logos were generated using WebLogo 3 (47). Logos are listed alphabetically according to kinase name.

The presence of a valid kinase binding site on a protein is no guarantee that a kinase will come into contact with the protein however (27). It has long been recognised that the activity of kinases can be regulated through upstream processes. The activation of kinases can be controlled through interacting proteins, complex formation, or phosphorylation events (48). Protein-protein interaction (PPI) networks can influence phosphorylation, or themselves be regulated

by phosphorylation, in a variety of ways. For example, kinase activity can be regulated by mediating proteins that help target their substrates (27, 49). In addition, phosphorylation can act as a switch within PPI networks, to enable or disable specific protein-protein interactions within a network (50). Phosphorylation has previously been found to be associated with proteins at the centre of PPI "hubs", with a broad array of interaction partners (51). There is also a wide variety of processes throughout the cell to activate or mediate kinases. Cell-cycle control kinases are activated at specific stages in the cell cycle, with the cell cycle-specific expression of cyclins coupled to the activation of their cyclin dependent kinase (CDK) counterparts to regulate their activity at the correct cell-cycle stage (52). An example of proteins that mediate the interaction between kinases and their substrates can be seen in scaffold proteins, which are integral to intracellular signalling networks and phosphorylation – in particular through coordinating kinase cascades (20).

In Section 1.2.1 I described the importance of kinase activity over the cell cycle, and some of the regulatory mechanisms involved in ensuring cell cycle stage-specific kinase activation and deactivation. In addition to protein interactions and the cell cycle, sub-cellular localisation plays a role in regulating kinase activity. Kinase CK2, which has large numbers of identified substrates, regulates different processes depending on its location in the cell (53). Critical cellular functions like apoptosis can be associated with kinase sub-cellular location. CDK2 has been shown to localise to the nucleus in proliferating cells, and to the cytoplasm in cells undergoing apoptosis (54). Sub-cellular localisation can also act in concert with cell-cycle progression to regulate kinase activity – for example the sub-cellular localisation of Wee1 kinase is dependent on cell-cycle stage (55).

There are many examples of context factors that regulate kinase activity. The ones that have been described here have the capacity to be modelled, using the available data: huge numbers of protein-protein interactions across multiple species have been catalogued in databases such as BioGRID (1), and less direct "associations" in the STRING database (56). Protein-protein interaction (or association) networks can feasibly be used to model the "interaction" context that kinases and their substrates operate it. Furthermore, the sub-cellular activity of kinases could be indirectly captured through the interaction networks of a kinase and its substrates.

1.3 Experimental identification of phosphorylation

There are two separate problems to consider when attempting to identify the phosphorylation status of a protein. The first is identifying whether a "phosphorylateable" residue actually

undergoes a phosphorylation modification. The second is the identification of the kinase that catalyses the modification. As is outlined below, while the identification of phosphorylated residues has become easier, and therefore the data extensive, the same has generally not occurred for the identification of kinases. This has lead to a large disparity between the number of known phosphorylation sites, and the sites that are annotated with a kinase. As phosphorylation is reversible, another consideration is the phosphatases that are responsible for removing phosphorylation modifications. However, the focus in my work will be on the identification of phosphorylation and kinase targets.

In initial studies on phosphorylation, protein phosphorylation sites were detected using ^{32}P labelling – a radioactive isotope of phosphorous. An alternative method for detecting phosphorylation sites is phospho-antibodies, which can recognise the phosphorylated form of a protein. In recent years however, the introduction of high-throughput mass spectrometry has resulted in phosphoproteomic studies that have identified tens of thousands of phosphorylation sites (28, 57).

While the identification of *in vivo* phosphorylation sites has become easier, identifying the kinases responsible for regulating the sites has in most cases remained elusive. Many experiments to identify the kinases regulating phosphorylation sites are performed using *in vitro* assays. Such experiments generally involve purifying the kinase and potential substrate to be tested and adding them in solution with ATP. Phosphorylated forms of the potential substrate can then be tested for using anti-bodies (for phosphorylation-specific forms of a protein or peptide) or mass spectrometry. For example, putative ATM substrates were identified through mutagenesis experiments on a known substrate to characterise the kinase's optimal binding motif (58). This motif was then used to identify proteins containing similar motifs, and again subjected to *in vitro* assays to confirm that they can be phosphorylated by the kinase. While *in vitro* experiments certainly provide valuable information about the likely kinases to be catalysing a phosphorylation modification, they are no guarantee that the kinase will phosphorylate the site *in vivo*.

There are several methods for using a combination of *in vivo* and *in vitro* experiments to detect kinase targets. Kinases can be transfected with the protein under study, and the phosphorylation levels measured; if inhibition of the kinase leads to a reduction in phosphorylation, this is a strong indication that the site is phosphorylated by the kinase (59). Another related method used frequently is to show through an *in vitro* assay that a kinase binds to the site of interest, then show *in vivo* that when the kinase is inactivated (through the use of a kinase inhibitor or transfection of kinase-specific siRNA for example), the target site does not get phosphorylated

(60–63). Alternative or additional evidence that can be provided is to demonstrate that the kinase interacts with the putative substrate *in vivo* – something that can be shown through co-immunoprecipitation experiments (64–67). There are also kinase phosphorylation-specific antibodies that can be used, through immunoprecipitation experiments, to identify whether a specific phosphorylation event occurs (68). Such experiments can also be combined with kinase inhibitors to confirm that the site is down-phosphorylated in the absence of the kinase.

There are also examples of "global" screening for kinase substrates across the proteome. Such methods can include *in vitro* screening of kinases simply for the purpose of deciphering their binding specificity (69). There are also methods that aim to identify putative kinase substrates. One method takes advantage of a small genetic modification of a kinase that allows it to use bulky ATP that wild-type kinases would not be able to bind to. Mass spectrometry can then be used to identify peptides that contain the heavier phosphate form. This method has been employed using *in vitro* kinase assays with cell lysates to identify putative substrates for CDK2 (70), and a similar method has been employed for CDK1 in both a human cell line (71) and in yeast (72). Such methods are useful for the identification of putative substrates, but require further work to confirm that the kinase targets the substrates and sites *in vivo*.

1.4 Phosphorylation databases

The inherent difficulty associated with identifying *in vivo* kinase substrates means that while phosphorylation sites are regularly detected, the kinase responsible generally remained unknown. For databases that catalogue phosphorylation sites, there is a substantial gap between the number of phosphorylation sites, and the number of phosphorylation sites annotated with a kinase. I present here a brief overview of the main eukaryotic phosphorylation databases, and their data collection for phosphorylation sites and kinase annotations. Figure 1.3 shows counts for four different databases that catalogue phosphorylation sites. PhosphoGRID is a yeast-specific phosphorylation site database, Phospho.ELM catalogues vertebrate phosphorylation sites, HPRD catalogues human modifications and PhosphoSitePlus contains modifications from a variety of mammalian organisms.

The Phospho.ELM, HPRD and PhosphoGRID databases collate phosphorylation data from the primary literature. The PhosphoGRID data collection process involves searching abstracts identified through a PubMed search using phosphorylation-related keywords (73). An examination of the experimental technique used for identifying the phosphorylation site (or sites)

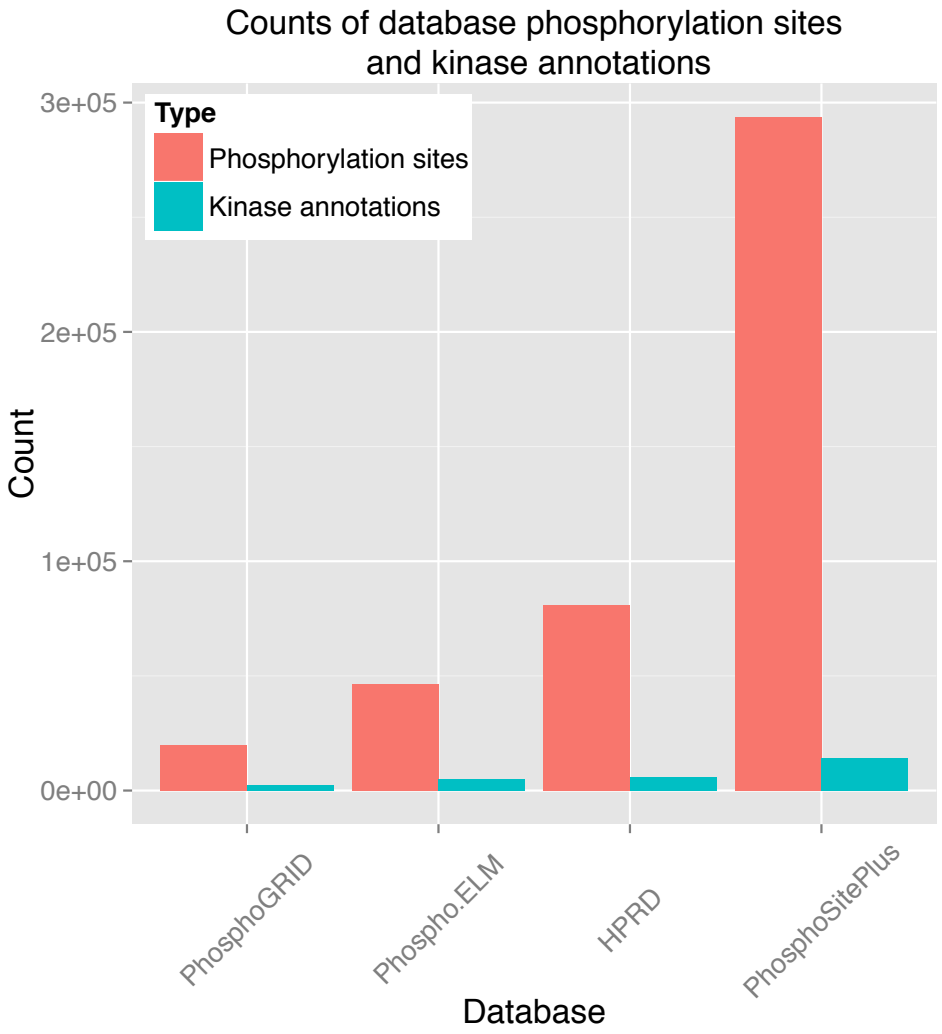

FIGURE 1.3: Counts of the total number of phosphorylation sites, and the number of sites annotated with a kinase for the four databases used in this work. Databases shown are PhosphoGRID (73), Phospho.ELM (74), HPRD (7) and PhosphoSitePlus® (4)

informs the confidence assigned to a particular phosphorylation sites; phosphorylation sites recorded in PhosphoGRID typically have multiple examples of experimental evidence supporting their inclusion. Where a kinase has been experimentally shown to target the site, that information is included.

Phospho.ELM also compiles phosphorylation sites (and kinase annotations) through manual searches of the literature. In addition, the database web-site contains the capacity for researchers to upload their phosphorylation data for inclusion in the database. HPRD collates large amounts of human protein information, including protein-protein interactions, sub-cellular localisation data, as well as multiple PTMs. Similar to the previous database, the information in HPRD is manually curated through the search and analysis of the primary literature. The

PhosphoSitePlus database, while also relying on the manual curation of phosphorylation sites and kinase interactions, contains a far larger number of phosphorylation sites compared to the alternatives (Figure 1.3). The PhosphoSitePlus® database has grown substantially as a result of phosphorylation data generated through high-throughput mass spectrometry. Cell signalling Technology (CST), which operates PhosphoSitePlus, also generates its own phosphoproteomic data that is incorporated into the database. The phosphorylation sites contained in PhosphoSitePlus are all recorded with the experimental techniques used to identify them, and whether the experiments were performed *in vitro* or *in vivo*. Where kinases are known for the sites, the experimental methodology is also listed. This ensures that users can choose what data they are willing to trust, based on the methods of experimental validation.

As can be seen from Figure 1.3, there is a large disparity between the number of phosphorylation sites recorded in these databases, and the number of sites that are annotated with a kinase. The number of kinase-specific phosphorylation sites in PhosphoSitePlus is barely 5% of the total number of phosphorylation sites, and for Phospho.ELM the percentage is 10%. Due to the expansive gap between known phosphorylation sites and their kinase annotations, there has been a great interest in developing computational methods that can not only predict phosphorylation sites, but also the kinases that mediate the modification.

1.5 Computational prediction of phosphorylation

Since Blom and colleagues published their phosphorylation site predictor, NetPhos, in 1999 (75), the field of computational eukaryotic phosphorylation prediction has grown tremendously, with over 50 methods published to date (Figure 1.4). While some methods aim only to predict phosphorylation sites (75–77), the majority of these predictors are *kinase-specific*. That is, the predictor scans a potential phosphorylation substrate to identify the most likely positions for some query kinase to bind to. As the focus of this thesis is on predicting kinase targets, rather than phosphorylation sites generally, this section will focus on the methods for predicting phosphorylation sites in a kinase-specific manner. The different methods have various coverage of kinases, with the approach and the training data impacting what kinases are available to the method. Some methods will make predictions for kinase families or sub-families in addition to, or instead of, individual kinases.

Typically, phosphorylation site predictors, whether kinase-specific or not, are *primarily* sequence-based – that is, the predictor mostly relies on sequence information in the form of motifs surrounding phosphorylation sites to make predictions. In some cases however, there are methods

that complement the amino acid sequence with additional information such as structure, disorder or context. In addition to the variety of information types, there have been different kinds of tools that have been employed for predicting phosphorylation sites. In this section I give an overview of how these tools have been applied to build the various phosphorylation predictors, and how different types of information have been used. Table 1.1 contains a summary of the available kinase-specific phosphorylation prediction tools, with information concerning their availability and usability.

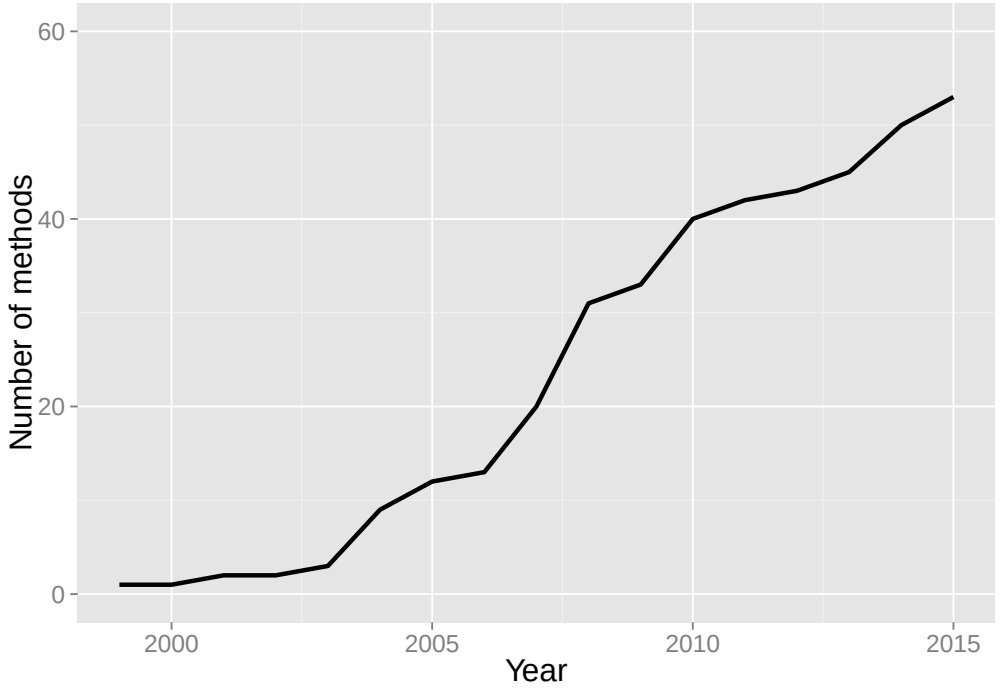

FIGURE 1.4: Count of eukaryotic phosphorylation prediction methods published between 1999 and 2015.

1.5.1 Scoring matrices

There are several kinds of sequence scoring matrices that have been used in phosphorylation site prediction – position specific scoring matrices (PSSMs), position weight matrices (PWM) and substitution matrices. A substitution matrix is a 20×20 matrix representing all possible pairwise combinations of amino acids, and is used to score the substitution of one amino acid for another. A classic example is the BLOSUM62 matrix often used to calculate sequence similarity. The original Group Phosphorylation Site predictor (GPS 1.0) used the BLOSUM62 substitution matrix to make predictions (78). The method relies on the hypothesis that phosphorylation

sites with high sequence similarity in the surrounding peptides are more likely to be targeted by the same kinase (or kinase family). They define a "Phosphorylation Site Peptide", or PSP(m,n) to represent a phosphorylation site with m residues upstream and n residues downstream from the site. To score a putative phosphorylation site for a given kinase, they take the known phosphorylation sites of the kinase and its immediate family. A similarity score is calculated between the putative site PSP and the known binding sites by summing the substitution scores from the BLOSUM62 matrix that occur in each position of the PSP(m,n). In an updated version of the method (GPS 2.0), the authors used matrix mutation to modify the BLOSUM62 matrix and found increased sensitivity and specificity in the performance of the method (79). In the current version of the method (GPS 2.1), the authors optimised the number of upstream and downstream residues that are used for the PSP(m,n) (80). Instead of a fixed m and n, the values were optimised for each kinase family.

A different kind of scoring matrix is the position weight matrix (PWM). A PWM is a $20 \times m$ matrix, where each row in the matrix represents an amino acid and m is the length of some sequence motif. Each value in the matrix represents the weight – based on observed frequencies in training data – that an amino acid contributes towards classification at a given position. An example of a method that uses PWMs is Predikin, which guides its generation of PWMs based on an understanding of kinase structure and substrate-binding sites from X-ray structures of kinase-bound phosphorylation substrates (25). Predikin is based on the concept of specificity determining residues (SDRs): conserved sequence regions in the binding domain of kinases that determine whether they will bind to certain substrates. Given a kinase, Predikin identifies other kinases with similar substrate binding sites. Substrates and phosphorylation sites for these kinases are identified from the Predikin database PredikinDB (81, 82), which itself sources kinase substrates from UniProt (83). A frequency matrix is constructed by counting the occurrences of amino acids in a heptapeptide (7 residue window) centred on the phosphorylation sites, and is converted into a PWM. Predikin is therefore able to build PWMs for any kinase if it has a binding region similar to a kinase in the PredikinDB, giving it a wide scope over many kinases.

The MIMP (mutation impact on phosphorylation) predictor was implemented as part of a method for scoring the expected effect of protein variants on phosphorylation (84). The authors collected kinase-specific phosphorylation data from a variety of databases, and constructed kinase-specific PWMs by calculating amino acid frequencies at each position within a 15 residue window around the phosphorylation site. To optimise the performance of their PWMs the authors performed an iterative refinement process, whereby they constructed an initial PWM

based on the full set of positive samples. Based on random sampling of negatives, they generated distributions of scores using the PWM, scoring both the negatives and the positives used to generate the PWM. If a positive sample fell within the 90th percentile of the negative distribution, it was discarded, and a new PWM was built based on the remaining set of positives. This process was repeated until positives were no longer discarded (subject to a lower bound of ten positive sequences remaining).

The above methods use phosphorylation data from known kinase-protein targets to train their algorithms, but an alternative method is to estimate the binding specificity of kinases from *in vitro* peptide array experiments. Scansite scores kinase binding locations on proteins using kinase-specific PSSMs constructed from data generated in peptide library experiments (85). These experiments involve incubating a kinase domain of interest with peptides that have a fixed phosphorylatable (S/T/Y) residue and sequencing the peptides that are found to be phosphorylated (86). These experiments yield relative levels of amino acid at positions relative to the central phosphorylation residue, which can be normalised and converted into a PSSM (87). The NetPhorest predictor (88) also builds PSSMs from *in vitro* peptide array data, though its scoring matrices are complemented with artificial neural network classifiers (discussed below). Due to the fixed structure of PSSMs, they train PSSMs on the peptide array data that is unlikely to offer information regarding relationships between residues.

There are several drawbacks to using scoring matrices for predicting kinase targets. Firstly, on account of their fixed structure, scoring matrices do not have the capacity to recognise subtle sequence patterns; co-occurrence of certain amino acids acids would not be picked up by a scoring matrix for example. Secondly, on account of only considering the linear motifs, scoring matrices are blind to additional factors both within the substrate (e.g. surface accessibility), and outside the protein, which could have an important regulatory impact on the kinases that target it. Due to these issues, many phosphorylation predictors use machine learning methods that are not only able to detect patterns in the sequences surrounding kinase binding sites, but have the capacity to incorporate information in addition to that contained in the amino acid sequence.

1.5.2 Machine learning methods

There are several machine learning methods that have been applied to phosphorylation prediction. By far the most popular method is the support vector machine (SVM), though there have been several examples of the use of neural networks. In addition, hidden Markov models

TABLE 1.1: Table of kinase-specific phosphorylation site predictors. For tools with multiple versions published, the most recent is cited. Availability specifies whether the tool can be used through the web (Webtool), with a downloadable graphical user interface (GUI), or with a downloadable command line (CL) tool that can be operated through a terminal. If there is no available tool, or if a method's website is no longer accessible, it is listed as "unavailable"; alternatively, the authors may offer a downloadable dataset. Batch specifies whether users can feasibly input and run large numbers (e.g. over one thousand) of protein sequences through the tool. Some tools (e.g. Scansite) instead have options to search an existing sequence database, instead up a large upload. The methods have been listed according to year published.

Name	Availability	Batch	Website	Year	Ref.
Scansite	Web	No	`scansite3.mit.edu`	2003	(87)
NetPhosK	Web	No	`cbs.dtu.dk/services/NetPhosK`	2004	(16)
PPSP	Web	No	`ppsp.biocuckoo.org`	2006	(89)
KinasePhos	Web	No	`kinasephos2.mbc.nctu.edu.tw`	2007	(90)
CRPhos	CL tool	Yes (CL)	`ptools.ua.ac.be/CRPhos`	2008	(91)
NetPhorest	Web, CL tool	Yes (CL)	`netphorest.info`	2008	(88)
Phos3D	Web	No	`phos3d.mpimp-golm.mpg.de`	2009	(26)
PredPhospho	Unavailable	N/A	N/A	2009	(92)
Musite	Web, GUI	Yes (GUI)	`musite.net`	2010	(93)
N/A	dataset	N/A	`bioinfo.bjmu.edu.cn/phospho`	2010	(94)
Predikin	Web, CL tool	Yes (Web/CL)	`predikin.biosci.uq.edu.au`	2011	(95)
GPS	Web, GUI	Yes (GUI)	`gps.biocuckoo.org`	2011	(80)
PhosK3D	Web	No	`csb.cse.yzu.edu.tw/PhosK3D`	2013	(24)
PKIS	Web	No	`bioinformatics.ustc.edu.cn/pkis`	2013	(96)
NetworKIN	Web, CL tool	Yes (CL)	`networkin.info`	2014	(97)
MIMP	Web, CL tool	Yes (CL)	`mimp.baderlab.org`	2015	(84)

(HMMs) and conditional random fields have been employed. In this section I firstly give a review of the use of support vector machines in kinase-specific phosphorylation site prediction; this is followed by an overview of the additional tools that have been employed.

Support vector machines

The most popular machine learning method that has been applied to phosphorylation is the support vector machine (SVM). An SVM is a discriminatory method of binary classification, that solves an optimisation problem to separate two labelled sets of inputs (98). A key concept in SVM training is the use of kernel functions to transform training data into more easily separable dimensions. This is useful for the classification of high dimensional data, such as protein sequences. The other key concept in SVMs is margins. SVMs use decision boundaries to separate classes, such that a margin separating the data points in the two classes is maximised; often the data points associated with the classes will not be perfectly separable however. The idea of a "soft margin" SVM is to allow some flexibility in the margin around the decision boundary,

whereby points close to the decision boundary are ignored – allowing for the placement of a decision boundary that provides greater separation of the two classes.

SVMs have been applied to phosphorylation prediction in several ways. The simplest involves a binary encoding of amino acid occurrences at positions within a window surrounding a phosphorylation site; other methods provide more sophisticated profiles of the amino acid content of the motifs. There are also publications demonstrating how SVMs can be used to incorporate into a model different types of protein information, such as protein disorder, evolutionary information (in a kinase-generic predictor) and 3D structure, to supplement sequence data (90, 93, 99).

PredPhospho is an SVM phosphorylation predictor that was developed as part of a database for scoring the potential effect of missense mutations on protein phosphorylation (92). The authors represented phosphorylation site motifs using a binary encoding of amino acid occurrence within some m length window surrounding the phosphorylation site. Amino acids were encoded using a vector of length 20 where, for example, methionine was encoded as 10000000000000000000, isoleucine as 010...000 and so forth. A motif was therefore represented as a vector of such binary vector encodings. The vectors were used as input features for training SVM models in a kinase-specific manner. Another method for encoding kinase binding peptides for use in an SVM is the composition of monomer spectrum (CMS) technique, used by the protein kinase identification server (PKIS) (96). Given some phosphorylated peptide, a single monomer spectrum is defined as a vector of amino acid counts for that peptide; i.e. the peptide sequence EQEESPLRR could be encoded as the vector 01000300110000201000, where each position in the vector represents the count of an amino acid occurrence in the peptide. The CMS method encodes the sequence information in an m length sequence window by computing monomer spectrum vectors for all windows from size 3 to m, where the window is centred on the phosphorylation site.

KinasePhos2 (90) is the successor to the hidden Markov model (HMM) methodology of KinasePhos (100), and incorporates information on sequence content and protein-coupling patterns from phosphorylation motifs into SVM classifiers. For the KinasePhos2 predictor, a kinase binding motif (taken as a 9-residue window surrounding the phosphorylation site) is represented by two features: an encoding of the protein sequence in the motif, and a profile of amino acid couplings. Given some pair of amino acids X and Z, and a distance d between them, Wong and colleagues defined an amino acid coupling, C_{XdZ}, as the frequency of XdZ divided by the frequency of Z, as observed in training data. In order to select couplings relevant for phosphorylation, they calculated the difference in values of C_{XdZ} for both the set of phosphorylation sites and a background set of all phosphorylatable residues. If the difference in C_{XdZ} passed a certain threshold it was used as a feature in the SVM model. In addition to amino acid coupling

profiles, the authors defined three different encodings of the amino acid content of the phosphorylation motifs: (1) an encoding based on the row numbers of amino acids in the BLOSUM62 matrix, (2) a reduced alphabet based on the amino acid properties (hydrophobicity, polarity etc.) and (3) a vector encoding where each amino acid was defined using a 20-dimensional vector.

The SVM methods described so far use a variety of motif representations as input features in SVM classifiers. However, SVMs are able to perform classification on the basis of multiple protein characteristics – in addition to the sequence – when these are represented as numerical values in feature vectors. This is something taken advantage of by the Musite predictor, which combines k-nearest neighbour (KNN) scores (representing the distance between the motif surrounding a query site and positive or negative examples of phosphorylation) with amino acid frequencies and protein disorder scores into SVM classifiers (93). Given positive and negative phosphorylation data, a distance metric is calculated between a query site and the sites within the positive and negative data sets. Some k number of the top scoring neighbours in both positive and negative sets are then identified, and a ratio of positives to negatives is calculated. The authors defined five increasing levels of k such that the KNN score for a phosphorylation site is represented as a vector of 5 elements. Amino acid frequencies were calculated within at 13 residue window around the phosphorylation site, and represented as a size 20 vector of frequencies. Disorder predictions for query proteins were made using the VSL2b disorder predictor (101), and disorder scores for a phosphorylation peptide (as defined by some m length window around the phosphorylation site) were defined as the average VSL2b disorder prediction across the residues contained in the peptide. The authors define 3 values of m (1, 5 and 13) to calculate disorder scores for a peptide. These scores were included as inputs into the feature vectors used to train the SVM.

The Phos3D predictor proposed that some elements of kinase-substrate recognition may lie in amino acids that are spatially close, though sequence distant, from phosphorylation sites (26). They used sequence content from phosphorylation motifs encoded with physical-chemistry properties (hydrophobicity, disorder indices, solvent accessibility etc.), and supplemented this with spatial data from 3D structures. To incorporate spatial information into the model they identified amino acids that were in a range of 2 to 10 Å to the phosphorylation residue. For each amino acid they calculated a ratio between the number of times the amino acid was in range of the phosphorylation site, and all other amino acids that were in range. They were therefore able to build a profile of amino acid frequencies in spatial closeness to the phosphorylation site. After comparing the prediction accuracy of their model using just sequence features verse

the combination of sequence and spatial information, they found that the inclusion of spatial information increased prediction accuracy by approximately 5%.

Separate (though similarly named) to the Phos3D predictor is the PhosK3D predictor, which also incorporates amino acid motifs and spatial information from PDB structures into kinase-specific SVM models (24). The authors incorporated several sequence characteristics into SVM models. After extracting phosphorylation peptides of length 13 from around phosphorylation sites, the authors constructed PWMs to represent the raw sequence content of the motifs. In addition, they calculated surface area accessibility of the residues within the phosphorylation peptides with the tool RVP-Net (102), and secondary structure using PSIPRED (103). Similar to Phos3D they used the 3D protein structures to calculate spatial amino acid frequencies, though the authors of PhosK3D calculated amino acid frequencies at varying distances from the phosphorylation site, ranging from 3 to 12 Å. Su and colleagues found that a model incorporating spatial and sequence information obtained an average 10% increase in prediction accuracy over using sequence alone, up from the 5% increase seen with the Phos3D predictor.

The results from the Phos3D and PhosK3D methods demonstrate that spatial information improves phosphorylation prediction accuracy over using linear motifs alone; however, a major drawback is the lack of availability of 3D structure information for many proteins. This imposes restrictions both for the proteins that predictions can be made on, and the kinases that the method can make predictions for; there needs to be 3D structures available for a kinase's substrates in order to train a kinase-specific predictor. The PhosK3D predictor offers prediction for 127 kinases or kinase sub-families using its sequence method, but only 21 for the method incorporating spatial information. This illustrates a drawback of using SVMs to train phosphorylation predictors on selectively available data: they are unable to handle proteins where that data is missing.

The SVM methods presented in this section have an advantage over scoring matrices in that they allow for more flexible representations of the phosphorylation peptides, and have the capacity to incorporate multiple information types. There are limitations to using SVMs for phosphorylation prediction, however, as seen by the restraints of data availability imposed on the Phos3D and PhosK3D methods. An additional limitation concerns the discriminatory approach of SVMs, which make them unsuitable for problems that are not of a binary classification nature. As was shown in Section 1.2.2, kinases within the same family can share binding site characteristics; indeed one phosphorylation site can be targeted by multiple kinases. The binding preference of kinases is therefore not best represented as a binary discrimination problem, as it will be in an SVM model.

Additional methods

While scoring matrices and SVMs are the tools underlying the majority of phosphorylation predictors, there are several alternative methods that have been proposed. Neural networks (NN) are an earlier machine learning method that have been applied in phosphorylation site prediction. A NN is a network consisting of layers of nodes, with input at the top of the layer and output at the bottom. Weights associated with the inputs and subsequent layers of nodes – set using training data – can potentially learn non-linear sequence features, such as relationships between positions in a motif. The kinase-generic phosphorylation predictor NetPhos (75), its kinase-specific successor NetPhosK (16), as well as the Yeast-specific (though kinase-generic) variation NetPhosYeast (104) each train three layer neural networks to predict phosphorylation sites on the basis of sequence data. NetPhos and NetPhosK represent motifs surrounding a phosphorylation site as a vector of binary encodings of amino acids (105), similar to that described for the PredPhos method previously; i.e. an encoding of an amino acid takes on the form 100..000. The phosphorylation peptides represented using the binary encoding scheme were then presented as feature vectors for training the NN models.

As mentioned previously, the NetPhorest predictor has an alternate approach to either building PSSMs or training neural networks depending on the type of phosphorylation data being considered (88). *In vitro* peptide array data, which offers a quantitative representation of amino acid frequencies, but can not be used to glean correlations between positions in a motif, was used to construct PSSMs. The protein phosphorylation data generally used for training phosphorylation predictors was applied to training NNs. They trained NNs by representing the phosphorylation peptide data using the binary amino acid encoding scheme employed by Blom and colleagues (75), and used cross-validation testing to optimise several parameters (peptide window size, number of hidden neurons in the model and the learning rate) for the kinase-specific models.

KinasePhos (the predecessor to the SVM-based predictor KinasePhos2 described above) trained profile hidden markov models (profile HMMs) to predict kinase-specific phosphorylation sites (100). HMMs are a form of graphical modelling used for labelling sequential data, and profile HMMs are a specialisation whereby a sequence alignment is used to build a position-specific scoring model. The authors sourced phosphorylation data from the PhosphoBase (106) and Uniprot (83) databases. The phosphorylation sites were labelled according to kinase annotations, with non-phosphorylated S/T/Y sites contained in the substrate sequences used as negative. Extracted phosphorylation peptides were then used to construct the profile HMMs on a kinase-specific basis.

Dang and colleagues proposed that conditional random fields (CRFs) could be used to model the amino acid characteristics in kinase binding sites (91). CRFs are similar to HMMs in that they are used for labelling sequential data and are a form of graphical modelling. In contrast to HMMs however, CRFs use a conditional probability approach to assign labels to observations. The authors used phosphorylation data from Phospho.ELM to construct sets of kinase-specific phosphorylation peptides (using a window of length 9 centred around the phosphorylation site). They built feature vectors representing a variety of amino acid characteristics in the motif, such as co-occurrence of amino acids, and co-occurrence of grouped amino acids according to chemical classes defined by Wong and colleagues (90). For each kinase, a set of feature vectors from the positive phosphorylation examples was used to build a CRF model, and the negative data used to obtain false-positive rate thresholds for predictions made by the model.

Another methodology that has been proposed is Bayesian decision theory, which was employed in the PPSP (prediction of PK-specific phosphorylation site) predictor (89). Xue and colleagues defined two classes, C_1(phosphorylated) and C_2 (unphosphorylated). The application of Bayesian decision theory in this scenario is given some unclassified sample x, x will be considered phosphorylated (i.e. belonging to class C_1) if $P(C_1|x) > P(C_2|x)$, and unphosphorylated otherwise. The authors obtained kinase-specific phosphorylation sites from Phospho.ELM, and as with other methods defined negatives to be S/T/Y sites within the retrieved proteins that were not phosphorylated. From these sequences, peptides of length 9 were retrieved. They defined a sample peptide as $\vec{x} = (x_1, x_2, ..., x_9)$ and used Bayes theorem to denote $P(C_1|x_j) = \frac{P(x_i|C1)P(C_1)}{P(x_j)}$, where $j \in [1, 2, ..., 9]$ positions in the motif. The values in the equation are therefore determined based on observations in the training data. In addition to an error function calculated on the basis of bio-chemical similarities between amino acids from a BLOSUM62 matrix, they were able to use the above function to predict the phosphorylation status of given peptides based on the known phosphorylation examples.

The methods described so far only rely on information that is contained within a protein. There are many different ways that the sequence context within a phosphorylation site motif can be represented. Typically some m length window around a phosphorylation site is chosen and the amino acid content of the resulting peptide can then be represented in several ways. The amino acid content can be modelled in a position specific manner as in scoring matrices or in some machine learning methods such as NetPhosK and KinasePhos. Alternatively, or additionally, the amino acid content can be modelled in a non-position specific manner, for example the composition of monomer spectrum approach used by PKIS. As we have seen, machine learning classifiers also allow for additional information contained in the protein (such as structure) to be incorporated as features in predictive models.

As was outlined in Section 1.2.2, the factors that determine kinase substrates are not limited to those contained in the protein. While the methods outlined here may be able to ascertain the validity of a kinase binding location on a protein, they cannot say anything about the probability of a kinase coming into contact with the protein in the first place. A separate problem, and one that cannot be addressed by modelling the binding characteristics of kinases alone, is how phosphorylation substrates come into contact with the required kinase (27). The cellular context that the protein exists in needs to be considered to determine candidate kinases. In the following section I describe the previous work that has been carried out in applying context information to phosphorylation prediction.

1.5.3 Context-based methodology

While the vast majority of phosphorylation predictors only consider information contained in the protein, there have been three studies that have supplemented models of phosphorylation motifs with context information. One of the studies, by Li and colleagues, attempted to integrate phosphorylation motifs with a variety of different functional or context annotations (94); this approach was also adopted by Fan and colleagues, and applied to additional kinase families (107). The context and functional information included protein-protein association scores from the STRING database, gene ontology (GO) annotations (molecular functions, cellular components and biological processes), and other structural or pathway data. Li and colleagues used SVMs to build classifiers for 8 kinase families and compared prediction accuracy between using sequence alone, and sequence with various other data sources. Sourcing phosphorylation data from Phospho.ELM, they generated feature vectors representing peptides from a 9 residue window surrounding phosphorylation sites using the binary encoding scheme employed by Blom and colleagues (16).

The authors experimented with adding a variety on functional data types to their SVM classifier. Most relevant to the issue of context is their use of protein-protein associations obtained from the STRING database, and labelling of protein cellular components from gene ontology (GO) annotations. STRING contains binary protein-protein "association scores" – a probability of two proteins being functionally related, if not interacting directly, on the basis of various sources such as gene co-expression data, literature searching and protein-protein interactions (108). To incorporate STRING data into their model they identified proteins in the STRING database that were over-represented (based on a hypergeometric test) in their positive data compared to the negative unphosphorylated data. The STRING associations were also encoded using a binary format; if a query substrate interacted with a protein according to

STRING it was designated a 1, and 0 otherwise. Comparing the accuracy of phosphorylation prediction using sequence alone against incorporating STRING, the authors' results show that the use of STRING scores resulted in virtually no average increase in prediction accuracy. While one kinase family (PKC) obtained a moderate increase in accuracy from 78.26% to 83.46%, on average they obtained an accuracy of 87.33% using sequence alone, and 87.98% when the sequence was supplemented with STRING scores. As the STRING database contains a range of scored protein-protein associations, from low confidence to high confidence, converting all occurrences to a binary format is likely to introduce a large amount of noise, perhaps partially explaining the results. Similar to their results with STRING, there appeared to be one kinase family (GSK3) that benefitted from the cellular component annotations, with an increase of accuracy from 77.69% to 87.24%. However, if the GSK3 kinase is excluded, the cellular component annotations actually result in an average decrease in accuracy, from 78.34% to 77.62%. While the authors found their addition of context information useful in some regards, it appears that their approach has little generalisability among kinase families. This should illustrate that modelling the context that kinases operate in is not a trivial exercise, and an ad hoc approach of simply adding various sources to a model is unlikely to result in a system with generalised predictive power.

To date, the most promising approach to using context information to improve kinase-specific phosphorylation prediction was made with the NetworKIN predictor (109). NetworKin improved upon motif based scoring by including a "context score", which was calculated on the basis of protein-protein association scores contained in the STRING database (108). The original NetworKIN algorithm consisted of two main stages. In the first stage, one or more proteins were submitted along with known, or suspected, phosphorylation sites. Using PSSMs and NN-based sequence prediction (i.e. the NetPhorest methodology described earlier), one or more kinase families were assigned to the sites. In the second stage, candidate kinases from the families predicted in the first stage are scored on the basis of an "association network" constructed using the STRING database. As STRING associations are represented with a probability based on the strength of the underlying data, a network of protein-protein associations can be constructed with varying path lengths. To calculate a context score on the basis of such an association network, the Floyd-Warshall algorithm was used to find the shortest path between the query protein and a kinase that is a member of the predicted kinase families. The final score was the product of the STRING and sequence scores.

The latest version of NetworKIN updated the distance algorithm used to assign a kinase to a potential substrate (97). The current score includes penalties based on path length, and the number of connections in intermediate association hubs. In addition, rather than pre-screening

the query substrates to identify potential kinases from sequence, the sequences are scanned separately with NetPhorest (88) to allow a probability to be assigned for any kinase. The final score for a kinase-specific phosphorylation site is a naive Bayes product of the sequence and context scores. In contrast to the method by Li and colleagues described above, Linding and colleagues showed that their use of the STRING network provides a more generalised level of increased phosphorylation prediction accuracy over using sequence alone. They evaluated their method for its ability to correctly predict kinase-specific phosphorylation sites on a set of 38 kinases for which 10 or more phosphorylation sites were known. This evaluation yielded an average area under the curve (AUC) of 0.78 using sequence (NetPhorest) alone, and an average of 0.83 when NetPhorest was complemented with STRING scores. Despite this, 12 out of the 38 kinases recorded a *decrease* in AUC when the STRING score was included, and an additional 4 recorded an AUC increase of under 0.01, meaning that for over 40% of the kinases the STRING score resulted in a negligible or negative impact on prediction accuracy.

The methods described here have used context data to supplement sequence scores, but the factors that contribute towards a kinase targeting a protein at the systems level are complex. The context that surrounds a phosphorylation event will likely be more sophisticated than what can be represented with a shortest path search; indeed it appears that there are strong limitations in how much accuracy can be gained through the sequence and context models described above. There are three problems that need to be addressed: (1) modelling how kinases come into contact with their substrates, (2) modelling the binding of kinases to target sites and (3) integrating these divergent elements into a unifying model of phosphorylation. While much work has been done on (2), even the little work that has been done on (1) and (3) has focussed less on modelling context to understand how kinases target substrates, but rather using context as a supplement to sequence scores.

1.6 Research aims and project overview

The computational methods described in Section 1.5 rely almost exclusively on information that is contained within the protein. These predictors primarily model amino acid content within a fixed window surrounding a phosphorylation site, though some incorporate additional information such as protein disorder or 3D structure. The NetworKIN method, though it is the first step towards the use of context information, essentially identifies the “closest” kinase to a protein substrate it can find in the STRING database. Furthermore, separating out the problem of the sequence specificity of kinases and the context factors that regulate their activity

means that potential influences between the two domains of information will be neglected. A comprehensive model of phosphorylation would provide a seamless integration of the context and sequence factors that influence how kinases target their substrates. My hypothesis is that by integrating of two aspects of kinase regulation, context and sequence, I can build methods that predict kinase substrates with greater accuracy than if considering context or sequence in isolation.

The **general aim** of this project was to develop a framework for integration of cellular context data such as protein-protein interaction information with sequence data in order to solve biological problems where these two domains of information are of high relevance. As phosphorylation is an event regulated both through the sequence binding affinity of kinases, and mediating protein interactions, it has been a prime candidate for this study. In the final part of the study, the prediction tool was applied to a biological problem: detecting the effect of single nucleotide polymorphisms (SNPs) on protein phosphorylation status. The more **specific aims** of the project are outlined below.

1. Integrate cellular context information in the form of protein-protein interactions, cell-cycle progression and kinase-specific phosphorylation events into a model that can classify the kinase, or kinases, responsible for phosphorylating a putative phosphorylation substrate. As part of this aim I also address the following questions:

   (a) Can the model be used to improve sequence-based phosphorylation site prediction through combining its output with that of existing phosphorylation predictors?

   (b) Can context can be used to predict phosphorylation status change in proteins (part of an sbv IMPROVER competition)?

2. Develop a probabilistic model for predicting kinase binding sites from sequence. Incorporate this sequence model into the context model.

   (a) How generalisable is the modelling approach when applied to kinases from different species?

   (b) How accurate is the model at predicting kinase-specific phosphorylation sites compared to alternative methods?

3. The final aim is to apply the model to a biological problem: detecting the effect of nsSNPs on protein phosphorylation status.

(a) Is context an important contributor for understanding the effect of nsSNPs on phosphorylation?

(b) How reliable is the system for detecting known examples of variant-causing differential phosphorylation?

The first part of the project, described in Chapter 2, focussed on understanding the context that kinases operate in and how to leverage available data in order to design a method that could predict kinase substrates based on context. This chapter describes the design and implementation of a Bayesian network model that incorporates experimentally confirmed instances of kinase-substrate phosphorylation, protein-protein interaction/association data and cell-cycle data in order to predict kinase substrates. Through cross-validation evaluation I show that the model obtains reliable prediction accuracy, with an average AUC of 0.86 across the 59 kinases tested. Chapter 2 also demonstrates that the accuracy of previously-published sequence-operating methods for predicting kinase-specific phosphorylation sites can be improved by complementing their scores with context-based predictions from the Bayesian network. The method has been implemented as a tool accessible to the scientific community. The web server of PhosphoPICK (Phosphorylation in a Protein Interaction Context for Kinases) is publicly available at `http://bioinf.scmb.uq.edu.au/phosphopick`.

I had the opportunity to participate in the sbv IMPROVER (systems biology verification for Industrial Methodology for PROcess VErification in Research) species translation challenge. The purpose of the challenge was to the predict protein phosphorylation status change in response to varying treatment conditions, given gene expression data as measured under the same conditions. The first sub-challenge was to develop a method for predicting phosphorylation status change in rat cells, and the second sub-challenge was to predict phosphorylation status change in human cells using data from rat cells. The challenge allowed me to investigate whether a phosphorylation model based on protein-protein interaction data could use condition-dependent knowledge of protein expression levels to predict changes in protein phosphorylation. Chapter 3 describes a method for overlaying the protein-protein interaction networks of phosphoproteins with gene expression data in order to predict phosphorylation status change. The method obtained promising prediction accuracy, being ranked 6 out of 21 competitors in the first sub-challenge, and 7 out of 13 in the second sub-challenge.

While Aim 1 focussed on the problem of modelling context to predict kinase substrates, Aim 2 focussed on developing an algorithm for predicting kinase binding sites from sequence, and incorporating this algorithm into the larger context model. to obtain a more complete model of

kinase-protein phosphorylation. Chapter 4 describes a method that considers position-specific amino acid frequencies and the occurrence of co-occurring neighbouring amino acids (specifically dimers and trimers) within a window surrounding a phosphorylation site. The model was defined using a Bayesian network structure that allows the model to discriminate between a kinase's binding pattern, that of it's family members, and a phosphorylation background. Incorporating the sequence and context models enabled this "combined model" to predict kinase substrates with higher accuracy than by using context alone. The final system employed by PhosphoPICK involves using the combined model to obtain a prediction for whether a kinase will phosphorylate a substrate, and the sequence model to score the potential binding sites within the protein. When comparing the ability of PhosphoPICK and alternative methods to predict kinase-specific phosphorylation sites, I found that PhosphoPICK outperformed the alternatives for most comparisons made; PhosphoPICK obtained an average increase in sensitivity of between 9 and 22% over the alternatives at a 99.9% specificity level. Employing this system, PhosphoPICK is currently able to make predictions for 107 human kinases.

While I have primarily been working on human data, I was interested in testing PhosphoPICK on additional species. After obtaining phosphorylation data for mouse and yeast, I was able to build models for mouse covering 24 kinases, and models for yeast covering 26 kinases. When testing the mouse and Yeast models for predicting kinase substrates, I found that the combined model offered greater performance gains over using context alone than for the human version. This likely reflects the diminished availability of context data for mouse and yeast – protein abundance information over the cell-cycle was not available like it was for human, and the protein-protein interaction networks are smaller than for human.

Non-synonymous SNPs (nsSNPs) have the potential to cause loss or gain of protein phosphorylation sites through amino acid variants that either disrupt, or introduce, kinase-substrate binding sites. The final aim of the project was to use PhosphoPICK to build a method for predicting the effect of nsSNPs on phosphorylation. This method is described in Chapter 5. Using the Bayesian network models presented in Chapter 4, I build distributions of predicted variant effects over all protein-altering variants contained in the UniProt database. These distribution could then be used to quantify the significance of a novel variant's effect on phosphorylation. Using a set of phosphorylation-loss or phosphorylation gain-causing variants collected from the primary literature, I show that the method is able to detect known phosphorylation-altering variants at high levels of specificity.

Chapter 2

PhosphoPICK: Modelling cellular context to map kinase-substrate phosphorylation events[1]

2.1 Abstract

The determinants of kinase-substrate phosphorylation can be found both in the substrate sequence and the surrounding cellular context. Cell cycle progression, interactions with mediating proteins and even prior phosphorylation events are necessary for kinases to maintain substrate specificity. While much work has focussed on the use of sequence-based methods to predict phosphorylation sites, there has been very little work invested into the application of systems biology to understanding phosphorylation. Lack of specificity in many kinase substrate binding motifs means that sequence methods for predicting kinase binding sites are susceptible to high false-positive rates.

We present here a model that takes into account protein-protein interaction information, and protein abundance data across the cell cycle to predict kinase substrates for 59 human kinases that are representative of important biological pathways. The model shows high accuracy for substrate prediction (with an average AUC of 0.86) across the 59 kinases tested. When using the model to complement sequence-based kinase-specific phosphorylation site prediction, we found that the additional information increased prediction performance for most comparisons made, particularly on kinases from the CMGC family. We then used our model to identify functional

[1]Chapter reproduced from the paper published in *Bioinformatics*, 2015

overlaps between predicted CDK2 substrates and targets from the E2F family of transcription factors. Our results demonstrate that a model harnessing context data can account for the short-falls in sequence information and provide a robust description of the cellular events that regulate protein phosphorylation.

2.2 Introduction

Regulation of cellular processes occurs on multiple levels, with epigenetic modifiers and transcription factors (TFs) controlling gene expression, while various post-translational modifications regulate many protein functions (14–16). The most ubiquitous of post-translational modifications is phosphorylation, with at least 70% of human proteins estimated to be phosphorylation substrates (28). Phosphorylation is likely a significant factor in regulating the function of complex organisms, with a significant increase in the numbers of phosphorylation sites in eukaryotic compared to prokaryotic proteins (6). Phosphorylation is known to have numerous regulatory roles across the cell cycle, and specific kinases have been implicated in the regulation of G1 phase (110), the G1/S phase transition (111) and DNA replication and damage repair (112). Phosphorylation is particularly ubiquitous during mitosis where many complex operations such as spindle formation, centrosome maturation/separation and chromosome attachment to the spindle are controlled by kinases (41).

While advanced phosphoproteomic technologies have succeeded in identifying thousands of phosphorylation sites across multiple proteomes (28, 113), there has been an ever widening gap between known phosphorylation sites and the kinases responsible for those sites (114). Currently just over 10% of the phosphorylation sites recorded in the eukaryotic phosphorylation site database Phospho.ELM are annotated with a kinase. There have been examples of *in vitro* studies identifying kinase-substrate binding events (69), and while these studies offer interesting insights into the consensus motifs of kinase binding sites, it is unknown whether the binding events observed *in vitro* would occur *in vivo*. Determining kinase-substrates *in vivo* is non-trivial however, though there have been promising results from combining *in vitro* kinase detection assays with *in vivo* phosphoproteomics (115). As a result of the inherent difficulty in determining *in vivo* kinase substrates, there has been a great interest in developing computational tools to predict kinase-specific phosphorylation sites, with over forty phosphorylation site prediction methods published (18). While some methods aim only to predict phosphorylation sites (75, 104), the majority predict kinase-specific phosphorylation sites.

Historically, phosphorylation site predictors have operated primarily on protein amino acid sequences, relying on the information contained in the sequence region surrounding phosphorylation sites. It has long been recognised that short sequence motifs alone are insufficient for achieving respectable accuracy in predicting kinase-specific phosphorylation sites. As a result, prediction methods have often complemented sequence information with other types of data such as knowledge of 3D structure (26, 81), sequence disorder (93) and kinase family similarity (80). While such additional data typically improves prediction performance to an extent, they do not reflect the wider cellular regulatory mechanisms that cause kinases to target their correct substrates – a protein with an appropriate kinase binding site will not necessarily come into contact with that kinase (27).

The phosphorylation of a target substrate by a kinase is not determined solely by its binding affinity, but by various context factors that determine how a kinase comes into contact with its substrates (46). This is recognised by the NetworKIN predictor (97), which combines sequence-based scores with a score generated on the basis of a STRING network (108). Context factors can include cellular location (21), mediating and activating proteins such as scaffold proteins (20), cyclins (116), and cell cycle-specific expression of kinases and their substrates. Protein-protein interaction data can certainly be used to represent such context factors; though while there is vast amounts of protein-protein interaction data currently available in databases such as BioGRID (1) and STRING, incomplete coverage and variable certainty means that the integration of context features into a model is non-trivial.

In this work we explore a probabilistic model to accommodate missing values, seamless combination of protein interactions and cell-cycle expression, and to provide flexible options for querying potential kinase substrates. The model we present here, named PhosphoPICK (Phosphorylation in a Protein Interaction Context for Kinases), integrates known kinase-substrate relationships, protein-protein interactions (PPI), and cell-cycle data to predict kinase substrates for 59 human kinases. PhosphoPICK shows high prediction accuracy, with a mean AUC of 0.86 across the 59 kinases. We then demonstrate how our method can boost the prediction accuracy of kinase-specific phosphorylation site prediction by combining PhosphoPICK predictions with the phosphorylation site predictions from three previously published methods. We find that PhosphoPICK improves kinase-specific phosphorylation site prediction for most comparisons made, though greater performance increases were noticed on CMGC kinases – in particular cyclin dependant kinases (CDKs), where we observed substantial performance gains as measured by AUC50. We show that proteins predicted to be CDK2 substrates by PhosphoPICK have GO terms consistent with known CDK2 substrates, and investigate the functional overlap between

known and predicted CDK2 substrates, and the targets of specific E2F TFs using ChIP-Seq data.

2.3 Methods

2.3.1 Bayesian network model

We used a Bayesian network (BN) to design our model. Bayesian networks differ from machine learning tools that have previously been used for phosphorylation site prediction in several important ways. Bayesian networks are transparent, allowing for an understanding of how the variables in the model influence the final outcome (117). Furthermore, the probabilistic nature of a Bayesian network means that even in the absence of missing data, the model can still infer the most likely value of the unknown variables on the basis of the known data (118, 119). We represent observations about protein interactions, kinase-specific phosphorylation events and cell-cycle profiles as Boolean variables in a BN model (Figure 2.1). The model represents observations about a phosphorylation substrate - the kinases that bind to it, protein interactions, and whether it is up-regulated during the cell-cycle phases. The kinase nodes are linked to protein-protein interaction events that are believed to be relevant for the kinase to phosphorylate substrates. A latent variable is used to capture information from the cell-cycle data, and the kinase nodes are then conditioned on this latent variable.

2.3.2 Data resources

Known kinase-substrate relationships.

We obtained kinase substrates from Phospho.ELM and HPRD, after converting HPRD IDs to Uniprot identifiers. In order to identify protein interactions between kinases and their substrates, we selected kinases for which we found greater than 10 substrates. In total, we use 59 human kinases along with a total of 1,210 substrates. Table 2.1 shows the numbers of substrates that were identified for each of the 59 kinases. The 1,210 substrates contained 2,964 unique phosphorylation sites that were annotated with at least one kinase.

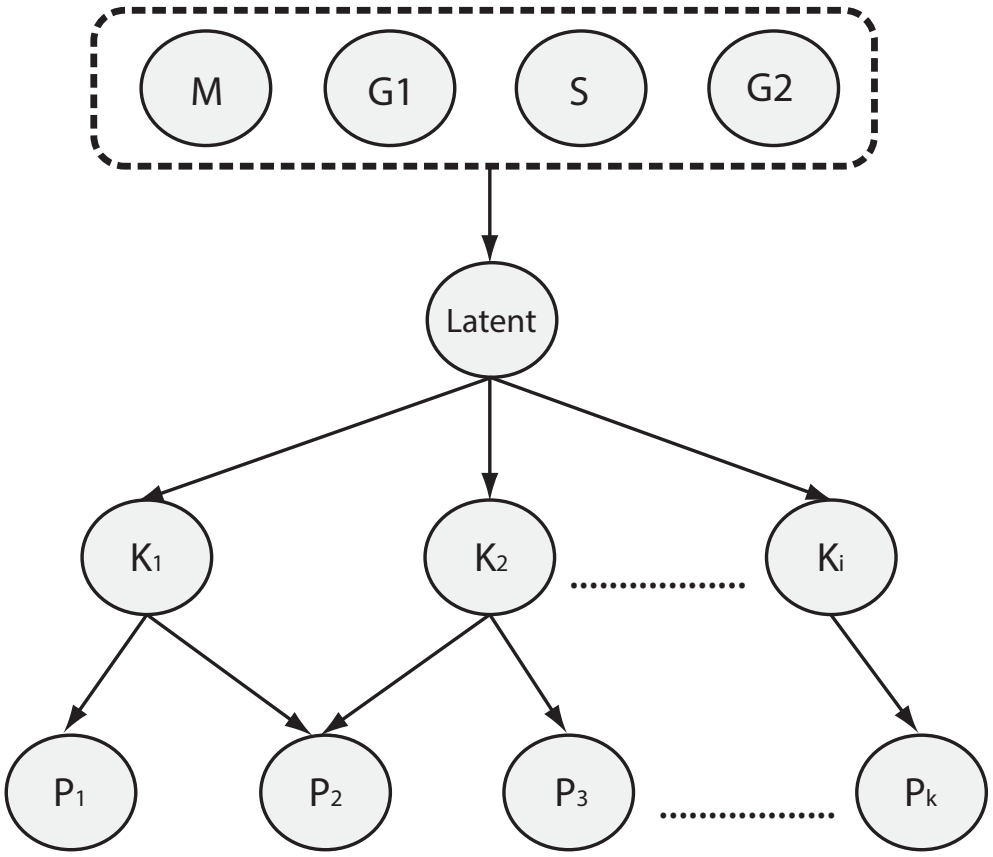

FIGURE 2.1: The PhosphoPICK Bayesian network model. Each of the kinase (K) nodes, representing a phosphorylation event by that kinase, are conditioned on a latent variable incorporating protein abundance across four stages of the cell cycle: Mitosis (M), G1, S and G2. The "leaf" nodes represent protein-interaction (P) events between the proteins represented in the nodes and a potential substrate. These nodes are conditioned on relevant kinase-specific phosphorylation events.

Protein-protein interaction and association data.

To identify and model interaction networks of kinases and their substrates, we used PPI data. In cases where physical interaction data is unavailable, associations inferred on the basis of other sources such as gene co-expression or literature mining may be informative, and such information is available in the STRING database. PPI information was taken from the Biological General Repository for Interaction Datasets (BioGRID) (1) by selecting entries that were of type "direct interaction" or "physical association". As protein-protein interactions are represented in binary format, this information was incorporated into the model as a Boolean value. The STRING database scores an association probability between two proteins, with a score of 0.4 defined as medium confidence. To convert this probability into a Boolean value

we defined cut-off probabilities, such that given some cut-off θ, any association with a probability $\geq \theta$ was classified as `true`, and any association with a probability $< \theta$ was classified as `false`. We tested three cut-off probabilities, starting at the medium confidence level of 0.4 and increasing in increments of 0.2. We found that a cut-off probability of 0.6 provided the best overall performance (Table A.1), and is the cut-off used in this work.

To identify relevant connections between kinases and protein-protein interaction events, the following steps were taken. Substrates were first grouped according to their kinase (one substrate could be assigned to multiple kinases). BioGRID was then searched for proteins that interacted both with a substrate and with its kinase – these proteins were added to a pool of potential protein interaction connections. For each kinase, the proteins in the pool were ranked in descending order according to the number of interactions that were observed with the kinase's substrates. An observation is defined as a substrate-protein interaction occurring in BioGRID and/or the STRING database. A count c was defined, so that for each kinase only the top c protein-interactions were used to form connections. To ensure that there would be enough observations of substrate-protein interactions for setting model parameters, a lower bound of 10 was set such that for a given kinase, at least 10 substrate-protein observations were required for the protein to be considered as a connection to that kinase. We tested three different upper-bounds of c: 25, 40 and 50 to determine the effect of varying sized interaction networks on prediction performance.

Protein cell-cycle data.

In order to model the availability of substrates during the cell cycle, we used data obtained from the experiments by (28), who measured the abundance of proteins at six stages throughout the cell cycle - M phase, G1 phase, the transition between G1 and S phase (G1/S), early S phase, late S phase and G2 phase. An asynchronous population of cells was also measured, and the signal used to log_2 normalise the measurements from the cells arrested during the six stages. A protein with a value of 0 during a stage of the cell cycle has an abundance equivalent to the asynchronous population, while a negative value indicates down-regulation and a positive value indicates up-regulation. To avoid fitting the model too strongly to data generated from a single cell type, we represented proteins' cell-cycle profiles in a simple binary format across four stages – M, G1, S and G2. We collapsed the G1 and G1/S stages in to the single variable "G1" and the early S and late S stages into the variable "S". If a protein has a value greater than 0 that stage is labelled as `true`; otherwise it is labelled as `false`. The G1 and S variables were set to `true` if at least one of their respective collapsed stages had a value greater than 0.

2.3.3 Model parameters and training

The variables in the network were represented with two kinds of probability tables. A conditional probability table (CPT) represents all possible values that a variable X can take given the set of parents, $pa(X)$, it is conditioned on. Parameters are set during training by calculating the frequency of occurrence of all possible configurations of $pa(X)$. If X does not have parents, the CPT simply represents the observed frequency from training data of X being true.

For situations where a variable is conditioned on greater than six parents, we used a variation of the Noisy-OR approximation (120). In order to set the parameters of the Noisy-OR table during training, each row (representing a parent variable) in the table was calculated as follows: each training sample where the parent is observed as being true was identified. A weighted frequency for each parent pa was calculated such that

$$\text{freq(pa)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{t}{(t+f)\text{pconf}_i}\right), \tag{2.1}$$

where n is the number of configurations of parent variables where pa is observed to be true, $pconf_i$ is the number of parents set to `true` in configuration i, t is the count of the variable the Noisy-OR node is representing being `true` during the ith configuration of parents, and f is the count of it being `false`.

For the latent variable, and variables that are conditioned on it, parameters are calculated using the expectation-maximisation (EM) algorithm on a training set (121).

2.3.4 Evaluation and definition of negative test sets

A common problem to phosphorylation-site prediction is that of defining a negative test set (18). However, as our model is not trained using sequence data, we were able to use a sequence-scoring method to define negative test sets for each of the 59 kinases in the model. To score protein sequences for kinase binding sites we used the Predikin web server (95) to obtain position weight matrices (PWMs) for 53 of the kinases in the model. For the remaining six, we constructed PWMs using phosphorylation sites from curated data (Section 2.3.5). For a given kinase, we scored each substrate in the training data-set by obtaining the highest scoring potential phosphorylation site. We then ranked the substrates based on the highest-scoring site from lowest to highest, and assigned an equal number of positive and negative substrates for

that kinase. As very low scores indicate a protein that the kinase cannot phosphorylate, this gives us a high-confidence negative test set for each of the kinases in the model.

We evaluated the model for each kinase for its ability to correctly predict known substrates compared to the negative set. To score the probability of a kinase phosphorylating a query protein, all nodes in the network were set according to the relevant data for the query protein except for the kinase that we were inferring. Model performance was evaluated using receiver operating characteristic (ROC) analysis by calculating the area under the ROC curve (AUC) (122). We used 15-fold cross-validation, and performed the cross-validation 10 times with different data-set splits. To avoid the possibility of the model gaining information about the test data during training, we ensured that each protein interaction variable was only connected to a kinase if, within the training fold, there were 10 (our previously defined lower bound) or more kinase substrates interacting with that protein. The data sets used to train and test the model are available in the supplementary material.

2.3.5 Generating position weight matrices

For most kinases we were able to obtain position weight matrices (PWMs) from the Predikin web-server, but for kinases CSNK2A1, CSNK2A2, ATM, ATR, CSNK2B and PRKDC we constructed PWMs based on known phosphorylation sites from Phospho.ELM and HPRD. The PWMs were constructed by taking a seven-residue window surrounding the phosphorylation site as described previously (81), and calculating the weight of each amino acid within each position in the window by calculating

$$w(a, j) = log_2 f(a, j)/0.05,$$

where $w(a, j)$ represents the weight of amino acid a at position j, $f(a, j)$ represents the frequency, and 0.05 represents a uniform distribution of amino acids.

2.3.6 Setting non-query kinase nodes on the basis of sequence data

We tested the ability of the model to classify for a query kinase when the remaining kinase variables in the model were set on the basis of sequence data. The PWMs were used to scan the sequences and ascertain the highest scoring potential phosphorylation site for each kinase.

For each training fold during cross-validation, we calculated the median scores for the kinases' negative sequences (the proteins they are not known to be phosphorylating). When evaluating the model on the test fold, we took the median of the negative scores and for each test substrate, a kinase variable was set to `false` if its PWM score for that sequence was below the median PWM score, and was left un-instantiated otherwise. This will result in a rough estimate of what kinases are *not* phosphorylating a query protein, with the model able to infer the probability of the remaining kinases phosphorylating the protein.

2.3.7 Testing the effect of STRING text mining on kinases

In order to test whether the use of text mining in the STRING database could be inflating the performance of PhosphoPICK, we repeated our cross-validation tests for each kinase as follows. When constructing a data file of input feature vectors for some kinase K, we first re-calculated the STRING score as described in (56) for each association involving K, omitting the text mining score. For each substrate Sub of K in the data file, if an interaction between Sub and K had been observed (as defined in Section 2.3.2) previously, but now was not being observed, the interaction between Sub and K was defined as `null` – the Bayesian network will consider this to be unobserved.

2.3.8 Applying model to sequence-based predictions of phosphorylation sites

From our curated set of kinase substrates, we identified 2,964 kinase-specific phosphorylation sites. In order to perform a fair comparison of how PhosphoPICK can improve the performance for predictions of novel proteins, we again performed 15 fold cross-validation with 10 data set splits, but retained the predictions for each protein in the test set. We then took the mean kinase scores for each protein across the 10 data set splits. In order to measure the ability of the methods being tested to predict phosphorylation sites, we took every potential phosphorylation site (serine/threonine or tyrosine) in the substrate set, and tested the methods' ability to predict known kinase-specific phosphorylation sites out of all these potential sites. We compared the performance of Predikin, GPS 2.1 and NetworKIN with the addition of PhosphoPICK by using two metrics: the AUC50 (an ROC curve calculated up to the first 50 false positives), and the sensitivity calculated at the threshold that yielded the fiftieth false positive. These metrics indicate the performance of the methods at a false-positive rate of 0.0005 (i.e. specificity of

0.9995) for serine/threonine kinases, and a false-positive rate of 0.002 (specificity of 0.998) for tyrosine kinases.

Predikin:

To make predictions for potential phosphorylation sites using Predikin, we used the PWMs that we were able to download from the Predikin web-server (95). The PWMs were used to score each potential phosphorylation site in our substrate set. For the comparison with PhosphoPICK, we first normalised the predictions on a per-kinase basis, by taking the minimum (min) and maximum (max) scores for each kinase. Each kinase-specific phosphorylation site prediction ($pred$), was then normalised by calculating $score = pred - min/max - min$. We made two comparisons by taking the product and the sum of the normalised PWM score, and the PhosphoPICK score for the substrate.

GPS:

We downloaded the current version of the GPS predictor (GPS 2.1) and adjusted the threshold setting to "none" so that we could make predictions for all phosphorylation sites in our set. For cases where GPS did not have a specific selection option for a kinase, we made predictions using the sub-family of the kinase: Akt was selected for the prediction of AKT1 phosphorylation sites, Abl for the prediction of ABL1 sites, and CK2a for predictions of CSNK2A1 and CSNK2A2 phosphorylation sites. The only kinase we were unable to make predictions for was PRKDC. For comparison with PhosphoPICK, we again calculated the sum and the product of the normalised site predictions (normalisation was performed as described above for Predikin) made by GPS, and the substrate predictions given by PhosphoPICK.

NetworKIN:

We downloaded the NetworKIN 3.0 (97) software provided for running on a local machine. We were able to make predictions for most kinases, with the exceptions of MAPK14, AKT1, PDPK1, PRKG1, RSK1, CSK, JAK1, JAK2, RET, CHK1, MAPKAPK2, AURKB, CSNK2B, PLK1 and PRKDC. As described above, to combined a NetworKIN score for a kinase, we first normalised the scores (normalisation was performed as described above for Predikin), then took alternatively the sum and the product of the PhosphoPICK score.

2.3.9 GO term enrichment analyses

We first obtained a background set of human proteins from UniProt (`http://www.uniprot.org`) by downloading all reviewed canonical human proteins. Of this set of 20,209 proteins, we identified 18,469 that were annotated with GO terms in the QuickGO web-service (`http://www.ebi.ac.uk/QuickGO`). Statistical significance of GO term enrichments was determined using Fisher's exact test, with a Bonferroni correction.

2.3.10 Transcription factor analysis

To obtain a set of putative E2F binding sites, chromatin immunoprecipitation sequencing (ChIP-Seq) data was downloaded from the ChIP-Seq experiment matrix provided by the ENCODE consortium (123). We downloaded ChIP-Seq narrow peak files for TFs E2F1, E2F4 and E2F6 in HeLa cells – the same cell type used for generating the cell-cycle data used in this study (28). In order to map the ChIP-Seq peaks to likely gene promoter regions, we also downloaded refSeq annotated genes of human genome 19 with a 2000 base pair region upstream of each of the genes. If a ChIP-Seq peak overlapped with an upstream region from a gene, the TF was considered to be targeting that gene.

2.4 Results

2.4.1 Model performance for predicting kinase substrates

We generated five BN models by grouping kinases according to their family similarities (32): CMGC, AGC, TK, CAMK and a combined model that incorporated kinases from the CK1, STE, atypical and other families. We tested the ability of the model to classify kinase substrates with varying numbers of protein interaction connections, and under three conditions. To gauge the level of influence that substrate abundance during the cell cycle has on prediction performance, we evaluated a version of the model excluding the cell-cycle variables (PPI only model), and compared the performance to the full model. When making inferences about a kinase-substrate phosphorylation event, the model relies on the knowledge of other potential kinases phosphorylating that substrate. However, for the majority of proteins there is little, if any, experimental information on any known kinase-specific phosphorylation events. Therefore, to determine whether the model could be reliably extended to the wider proteome, we tested

model performance when setting non-query kinase nodes to `false` on the basis of their sequence binding motifs (Section 2.3.6).

TABLE 2.1: Evaluation of model performance with median AUC on all kinases in the model, as tested under three different conditions: interactions only (int. only), full model, and kinase variables approximated using sequence data (seq. approx.). Also shown is the number of substrates (positive test set) that were identified for each kinase. Results are shown for 15-fold cross validation across 10 data-set splits. The best result for each kinase is highlighted in bold. CDK, MAPK and PRKC represent a family of kinases – the average values of their family members are included in the table. Kinases are listed according to the family-specific BN that they were incorporated into, where the "combined" model contained kinases from the CK1, STE, atypical and other families of kinases.

	kinase	substrates	int. only	full model	seq. approx.
CMGC	CDK	247	0.88±0.011	0.87±0.015	**0.91±0.01**
	GSK3B	58	0.81±0.01	0.82±0.009	**0.88±0.005**
	MAPK	136	0.84±0.016	0.88±0.016	**0.92±0.015**
AGC	AKT1	79	0.89±0.007	0.89±0.004	**0.91±0.001**
	GRK2	14	0.86±0.022	0.87±0.035	**0.87±0.01**
	PDPK1	23	**0.95±0.011**	0.94±0.011	0.91±0.021
	PRKACA	154	0.94±0.003	0.93±0.006	**0.96±0.002**
	PRKC	394	0.73±0.005	**0.86±0.006**	0.82±0.006
	PRKG1	26	0.86±0.014	0.86±0.009	**0.90±0.01**
	ROCK1	21	**0.80±0.006**	0.80±0.01	0.79±0.011
	RSK1	27	0.91±0.027	0.89±0.019	**0.93±0.008**
	RSK2	22	0.67±0.012	**0.77±0.027**	0.71±0.036
TK	ABL1	40	0.89±0.017	0.88±0.014	**0.97±0.006**
	BTK	14	0.79±0.056	**0.83±0.091**	0.69±0.11
	CSK	18	0.87±0.012	**0.95±0.034**	0.91±0.036
	EGFR	38	0.84±0.01	0.84±0.016	**0.95±0.001**
	FYN	38	0.83±0.033	0.85±0.049	**0.96±0.01**
	HCK	16	0.94±0.01	**0.96±0.032**	0.95±0.046
	INSR	23	0.92±0.011	**0.96±0.012**	0.93±0.002
	JAK1	11	0.65±0.12	0.69±0.078	**0.76±0.098**
	JAK2	17	0.95±0.013	0.95±0.026	**0.97±0.036**
	LCK	23	0.93±0.004	0.94±0.004	**0.96±0.011**
	LYN	39	0.76±0.028	0.77±0.028	**0.87±0.02**
	RET	16	0.60±0.11	**0.82±0.07**	0.69±0.096
	SRC	125	0.85±0.01	0.87±0.009	**0.89±0.003**
	SYK	27	**1.00±0.0**	**1.00±0.0**	0.98±0.004
	ZAP70	12	**0.95±0.064**	0.92±0.019	0.94±0.059

Continued on next page

	kinase	substrates	int. only	full model	seq. approx.
				Continued from previous page	
CAMK	CAMK1A	12	0.22±0.074	**0.75±0.075**	0.56±0.083
	CAMK2A	41	0.84±0.014	**0.89±0.022**	0.81±0.021
	CAMK2G	26	0.97±0.006	0.96±0.001	**0.98±0.012**
	CHK1	11	0.88±0.045	0.52±0.05	**0.91±0.038**
	LKB1	17	0.86±0.023	**0.90±0.054**	0.88±0.03
	MAPKAPK2	21	0.91±0.008	0.93±0.025	**0.93±0.01**
Combined	ATM	46	**0.99±0.001**	**0.99±0.001**	0.98±0.003
	ATR	14	**0.99±0.029**	0.98±0.018	0.92±0.047
	AURKB	16	0.94±0.004	**0.95±0.017**	0.91±0.03
	CSNK1A1	25	0.88±0.014	**0.89±0.011**	0.86±0.017
	CSNK1D	13	0.64±0.13	**0.69±0.074**	0.63±0.147
	CSNK2A1	135	0.87±0.002	**0.9±0.004**	0.89±0.005
	CSNK2A2	67	**0.96±0.001**	**0.96±0.005**	0.95±0.004
	CSNK2B	20	0.86±0.005	**0.88±0.017**	0.87±0.012
	PAK1	27	**0.59±0.03**	0.58±0.043	0.49±0.029
	PAK2	12	0.21±0.13	**0.53±0.12**	0.40±0.115
	PLK1	23	**0.92±0.005**	**0.92±0.006**	**0.92±0.006**
	PRKDC	11	0.74±0.064	0.76±0.053	**0.81±0.068**

The AUC results (shown in Table 2.1, with averaged ROC curves for the five models and three conditions shown in Figure 2.2) for 10 cross-validation runs evaluated on all 59 kinases in the model for the three different conditions demonstrate that the prediction accuracy of the full model is quite high, with most kinases having median AUCs surpassing 0.8. The average AUC over all of the kinases is 0.86. The generally low standard deviation indicates that these results are consistent regardless of the breakup of training/test data that is presented to the model. We tested three different values for maximum number of protein interactions that could be connected to a kinase variable (25, 40 and 50), but found that increasing the number of protein interaction events connected to the kinase variables had very little effect on the performance of the model (Table A.2), indicating the model's ability to make classifications based on a relatively small number of connections to the individual kinase nodes.

When comparing the performance of the PPI only model to the full model, we found that on average the inclusion of protein abundance data collected across the cell cycle offered modest improvements to prediction performance. For some kinases there was a greater performance improvement – for example a 15% increase for PRKC kinases, a nearly 10% increase for the tyrosine kinase CSK – but for many other kinases the inclusion of cell-cycle data seemed to

have little effect. This demonstrates that while the protein-protein interaction data provides the main source of information for the model, the use of cell-cycle data can offer improved prediction performance for some kinases. This performance increase occurs despite the fact that we only have cell-cycle data for less than half of the substrates in our set: the model infers the cell-cycle profiles for the remaining proteins. This indicates that the model, when trained on cell-cycle data, can still be applied to query proteins that have no associated cell-cycle data.

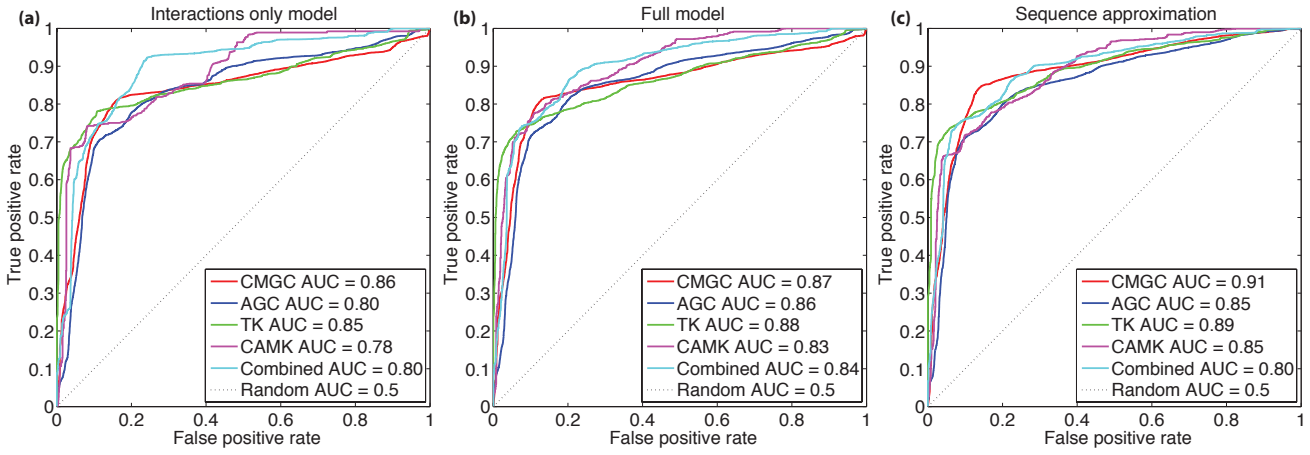

FIGURE 2.2: ROC plots showing prediction accuracy of the Bayesian network model as tested under the three different conditions: (a) interactions only, (b) full model, and (c) kinase variables approximated using sequence data. ROC curves represent the averaged ROC values of the kinases within each of the five Bayesian network models. Results are shown for 15-fold cross validation across 10 data-set splits.

Table 2.1 also shows a comparison between setting the kinase nodes with database ("full model") versus sequence information ("seq. approx."). We found that for many kinases, using sequence to set the kinase nodes actually resulted in an increase in performance. The median AUC for the CMGC kinases went from 0.87 using database information to 0.91 using sequence, and the median AUC for the tyrosine kinases increased from 0.88 to 0.94.

The possibility was raised that as the kinase-substrate data from HPRD and Phospho.ELM is sourced from the literature, and the STRING database also includes text mining from the literature, a system of circular logic could be inflating the performance values. To determine whether such an effect was occurring, we re-ran our simulations for each kinase with the text mining information for that kinase removed (Section 2.3.7). We found that while for some kinases this information appeared to have a large impact on prediction capability, it was not the case for the majority of kinases (Table A.3).

2.4.2 Improving sequence-based prediction of phosphorylation sites

For the remainder of this paper, the results were generated using the full model, with a PPI count of 25, and setting non-query kinase nodes on the basis of their sequence binding motifs. We tested the ability of PhosphoPICK to complement two phosphorylation site predictors that operate on sequence data: Predikin (95) and GPS (80). We also tested NetworKIN (97), which combines sequence scores with a context score generated on the basis of STRING associations. Comparisons were made by normalising the values of the methods being tested against, and summing the PhosphoPICK prediction (Section 2.3.8). Figure 2.3 shows the AUC50 (the AUC obtained when calculating ROC up to the fiftieth false positive) comparison for Predikin, GPS and NetworKIN across the five BNs, where the highest false-positive rate for serine/threonine kinases was 0.0005, and the highest for tyrosine kinases was 0.002. Individual results for each kinase are shown in Tables A.4–A.6. The results show that across all kinase families, there is an average increase in performance when the Predikin and GPS scores are complemented with PhosphoPICK predictions, with largest performance increases observed with kinases from the CMGC family. We found that the performance of GPS improved by 2-fold for predicting CMGC sites when combined with PhosphoPICK, and that the performance of Predikin was improved by over 6-fold. The smallest performance increases were observed with tyrosine kinases, where we found a 15% performance increase for GPS and a 40% increase for Predikin.

We found that in most cases, PhosphoPICK was unable to improve the performance of the NetworKIN predictions. As Figure 2.3 shows, the differences in AUC50 between classifying phosphorylation sites with NetworKIN alone and NetworKIN+PhosphoPICK are minor. However, as the NetworKIN score is already a combination of a STRING and sequence-based score, it is possible that a simple summing of scores cannot yield further performance increases.

2.4.3 Understanding E2F and CDK2 regulation

To evaluate the ability of the predictions made by PhosphoPICK to provide biological insights, we used CDK2 as a case study for a proteome-wide analysis. To determine whether predictions were consistent with what is known about CDK2, several GO enrichment analyses were performed (Section 2.3.9), comparing significantly over-represented GO terms (Fisher's exact test, Bonferroni correction, E-value$<$0.05) obtained for known CDK2 substrates with those obtained for the predicted substrates. We found that the known CDK2 substrates were enriched most strongly in various terms related to the G1/S transition of the cell cycle, such as DNA

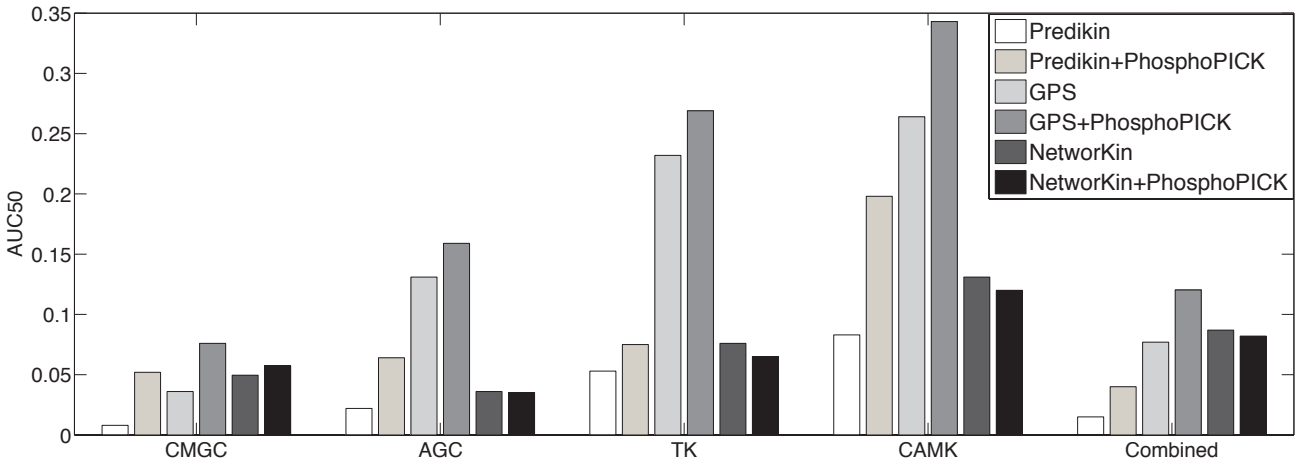

FIGURE 2.3: Comparison between predicting kinase-specific phosphorylation sites with three alternative scoring methods, and when the methods are informed by PhosphoPICK. AUC50 was calculated for each kinase as a measure of the predictive performance at low false positive levels. Shown here are the average values for each individual BN. The comparison is made by normalising the scores of the alternative methods to a value between 0 and 1, then summing this value with the PhosphoPICK prediction for a substrate.

damage response and DNA repair (Table A.7). This is consistent with the role of CDK2 in the regulation of the transition from G1 to S phase in response to DNA damage (112).

To investigate the agreement of PhosphoPICK predictions with known CDK2 substrates, we performed a proteome-wide scoring for CDK2 and took the top 300 novel predictions, excluding known CDK2 substrates from the set of predicted substrates. We again performed a GO enrichment analysis, and compared the values of the prediction terms with the significant terms that were found during the analysis on the known substrates. Table A.7 shows the GO terms found to be significantly over-represented among known CDK2 substrates, ranked from most significant to least significant. Over half of the terms (31/59) were found to be significantly over-represented among the novel substrates predicted by PhosphoPICK.

CDK2 is known to be a regulator of the TF E2F1 (124), a member of the E2F family, that is known to play a role in the G1/S transition, and DNA replication during the S phase (125–127). The E2F family is comprised of three classes of TFs: transcriptional activators, retinoblastoma (Rb)-dependent repressors, and Rb-independent repressors. What is currently lacking is an understanding of what specific roles in the S phase are controlled at transcriptional level by E2F, and at the post-translational level by CDK2.

In order to investigate what overlapping functions may exist between E2F-regulated transcription and CDK2-mediated phosphorylation, we took ChIP-Seq data (123) for E2F1 (activator),

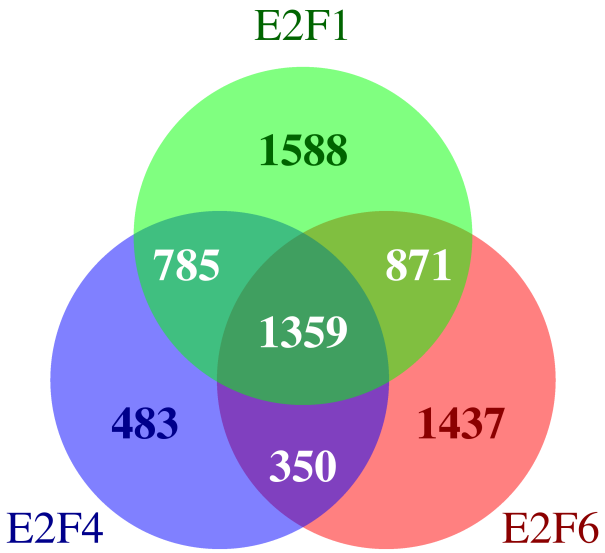

FIGURE 2.4: Venn diagram showing overlapping targets between E2F1, E2F4 and E2F6. CDK2 substrates within the unique E2F1 and unique E2F6 target groups had an over-representation of GO terms relating to apoptosis and ubiquitination. CDK2 substrates within the central (all overlapping E2F targets) group were enriched in terms relating to DNA replication and DNA repair.

E2F4 (Rb-dependent repressor), and E2F6 (Rb-independent repressor). Figure 2.4 shows a Venn diagram of the unique and overlapping gene targets that exist among the three TFs (Section 2.3.10). It has been shown previously that overlapping targets between E2F1 and E2F4 are enriched in DNA replication and repair GO terms (128). We found that the overlapping targets of all three TFs are also enriched in such GO terms.

We then combined our set of predicted CDK2 substrates with known CDK2 substrates and identified proteins from this combined set of substrates that were in the unique and overlapping groups of E2F targets. GO enrichment tests were performed with the CDK2 substrates as the foreground and the remainder of the TF target group as the background. This allowed us to detect what role CDK2 plays within these TF target groups. Tables A.8–A.14 contain the GO terms found to be significantly over-represented (E-value<0.05) among CDK2 substrates in the TF target groups. While we found significantly over-represented GO terms in all TF target groups, we noticed a larger number of process-specific terms among the unique E2F1 targets (Table A.8), unique E2F6 targets (Table A.10) and the overlapping targets among all three TFs (Table A.11). We found that CDK2 substrates among unique E2F1 targets and unique E2F6 targets were enriched in several terms relating to the regulation of apoptosis, as well as

ubiquitination. Substrates in the overlapping group of targets of all three TFs were enriched in terms relating to DNA replication and DNA damage repair.

2.5 Discussion

Protein phosphorylation is a highly regulated process, being controlled by the binding specificity to the protein kinase catalytic site, as well as various cellular processes that further enhance the kinase-substrate fidelity (27, 46). We have demonstrated how a probabilistic model of protein-protein interactions and cell-cycle data can be used to accurately classify kinase substrates. Importantly, we found that our model, when combined with sequence-operating methods, was able to improve the accuracy of kinase-specific phosphorylation site prediction at false positive levels below 0.002 for tyrosine kinases and below 0.0005 for serine/threonine kinases.

One potential point of concern in our current approach is that we only had access to cell-cycle data for a single cell type, and whether this could result in a tissue-specific influence that impeded predictions in some cell types. However, as the phosphorylation site data we obtained from Phospho.ELM originates from multiple cell types (including, for example, HeLa cells (129, 130), HEK 293T cells (131), MELN cells (132), and T98G glioma cells (133)), the performance of the model across these varying cell types validates the appropriateness of the data we used. We attribute this largely to the simple representation of the cell-cycle data as four Boolean variables, which would be unlikely to result in a cell type-specific bias. Somewhat counter-intuitively, we found that the cell-cycle data did not improve prediction performance for the CDKs – kinases whose activity is strongly linked to cell-cycle progression – while offering performance increase to other kinases. Though this work focussed on the use of protein abundance data for representing protein cell-cycle profiles, we note that dynamic gene-expression data across the cell-cycle also exists for human proteins (44). Further work could investigate what influence dynamic gene-expression data can provide to kinase-substrate prediction.

We observed some variation among the performance evaluations for the individual kinases, indicating that the model works better on certain kinases. However, we found that the performance for prediction of kinase substrates (Table 2.1) was not necessarily an indicator of what improvement would be seen when applying the model to phosphorylation site prediction. For example, the PhosphoPICK algorithm had excellent performance when classifying tyrosine kinase substrates – in several cases with AUCs greater than 0.9. However, when predicting

phosphorylation sites of tyrosine kinases using Predikin and GPS, we found that the prediction of tyrosine kinase phosphorylation sites benefitted the least from the addition of the PhosphoPICK score, and the score appeared to be detrimental to predictions made by NetworKIN.

The kinase family where PhosphoPICK consistently demonstrated the most powerful prediction performance was the CMGC family - principally CDK and MAPK kinases. We found that PhosphoPICK generally improved the prediction of phosphorylation sites for the CMGC kinases as tested across each of the three methods, though there were some cases where PhosphoPICK resulted in a decrease in the accuracy of NetworKIN predictions. As the kinases in these families have very similar binding patterns, it is likely that mediating proteins captured by the PhosphoPICK model make a greater contribution in the correct assignment of a kinase to a substrate. These results lend support to the intuitive notion that the addition of context information would support sequence-based predictions most powerfully when the kinase binding patterns are less specific, or are very similar among family members – or both, as is the case with CDK and MAPK kinases.

It was interesting to note that the putative CDK2 substrates within the overlapping E2F1, E2F4 and E2F6 targets groups were over-represented with GO terms related to DNA replication and DNA damage repair. Considering this group of genes was itself already enriched in such terms (when compared to the proteome), this underscores the importance that CDK2 has in regulating DNA replication and DNA damage repair (134, 135). There are several potential responses to DNA damage, but in some cases cells may undergo apoptosis (136, 137). We also noticed that putative CDK2 substrates within the unique E2F1 and E2F6 target groups were both over-represented with terms relating to the regulation of apoptosis and ubiquitination. These are both processes that CDK2 has previously been implicated in (138, 139), and ubiquitination is also known to play an important role in regulating apoptopic proteins (140). E2F1 is known to be a regulator of apoptosis (141), and similarly E2F6 can negatively regulate apoptosis (142), so it was interesting to find that the putative CDK2 substrates within the unique E2F1 and E2F6 target groups were enriched in apoptosis and ubiquitination GO terms. These results seem to suggest a dynamic regulatory interplay between the E2F family at the transcriptional level, and the CDK2 kinase at the post-translational level.

Chapter 3

Cross-species differential phosphorylation prediction: The sbv IMPROVER species translation challenge

3.1 Summary

In the previous chapter I demonstrated how context information such as protein-protein interaction networks could be used to predict phosphorylation substrates. An interesting application that could be derived from a phosphorylation model is understanding how the phosphorylation status of substrates change under different conditions; i.e. switching between phosphorylated and unphosphorylated states. A dynamic representation of phosphorylation change under different conditions can be obtained with a method that integrates protein-protein interaction information with condition-dependent protein expression levels.

The aim of the sbv IMPROVER (systems biology verification for Industrial Methodology for PROcess VErification in Research) species translation challenge was to investigate whether, after perturbing signalling pathways in rat, the effects on phosphorylation substrates can be predicted in human (143). Challenge participants were provided with differential phosphorylation data sets (for a set of phosphoproteins), as well as gene expression data sets, as measured under a range of conditions in rat and human epithelial cell lines. Participants were asked to

design a models that could, firstly, predict phosphorylation change in rat cell lines from gene expression data, and secondly, predict phosphorylation change in human cell lines using rat data. This challenge therefore provided a unique opportunity to investigate whether a phosphorylation model based on protein-protein interaction data could be used to predict phosphorylation status in response to changes in gene expression levels. Furthermore, it allowed me to test whether such a model would be able to predict phosphorylation change cross-species.

In this chapter I present a method that cross-references the protein-protein interaction networks of the phosphoproteins analysed by sbv IMPROVER with genes found to be differentially expressed under the same conditions as differentially phosphorylated phosphoproteins. Support vector machine (SVM) and random forest (RF) classifiers were employed to train computational models to predict phosphorylation status change based on the expression level of the identified proteins. The hypothesis was that a model based on the protein-protein interaction network of a phosphorylation substrate would predict phosphorylation status change more accurately than a model based on randomly selected genes. I show that in 86% of the tests performed, the protein-protein interactions models did indeed out-perform the random gene models, substantiating the validity of the approach. Based on the challenge to predict rat phosphorylation status change from gene expression, the method obtained an average balanced accuracy of 65% as measured on a blind independent test set kept from the challenge participants by sbv IMPROVER. For the first sub-challenge my method was ranked 6 of 21 participants, and 7 out of 13 for the second sub-challenge. These results demonstrate that the PhosphoPICK method could be applied to predicting change in the phosphorylation status of substrates based on condition-specific gene expression levels.

3.2 Introduction

Phosphorylation is the main regulatory switch used for modulating protein function, where change in the phosphorylation status of proteins can determine the activity of biological pathways. An example of such a pathway is the MAPK (mitogen-activated protein kinase) pathway, which links extracellular signals to activity in the nucleus, and is activated by phosphorylation (144). Understanding how the phosphorylation state of these pathways can change in response to treatments is important for drug development and discovery, however limited availability of data is a confounding difficulty. As a result, the development of predictive models that can use more readily available high-throughput data such as gene expression levels to infer the phosphorylation state of proteins is needed.

A further consideration is the level to which observations made by such models can be translated between species. Animal models are regularly used in a diverse range of biomedical settings, where treatments for cancers and diseases are tested. Animal models such as mouse have proved invaluable in gaining a greater understanding of tumour progression (145). While the purpose of animal models is to be able to infer the response that may be seen in a human, there are severe limitations on how much can be inferred between animal models and human. A treatment may work in the animal model, but the results do not translate across in human clinical trials. There are numerous reasons why translation between an animal model and human trials do not work, but an important consideration is whether stimuli and treatments affect animal models in comparable ways to how they affect humans.

The sbv IMPROVER Species Translation Challenge was created to probe the limits of translatability of biological observations between rat and human. Of particular relevance, was the focus of two of the challenges on predicting differential phosphorylation in response to stimuli. Epithelial cell lines from rat and human were exposed to a series of identical treatments, with gene expression (genome-wide) and phosphorylation levels of 16 important regulatory proteins (kinases and transcription factors) measured under the various conditions. The first aim, or challenge, of the competition was to create an algorithm that could predict the change of phosphorylation status of the given proteins based on gene expression data. The second challenge concerned the "translatability" between rat and human – predicting phosphorylation status in human proteins based on rat phosphorylation and gene expression data.

There are several ways that a treatment could result in differential protein phosphorylation. If a treatment affects the expression of a kinase or a phosphatase, it is probable that the ordinary phosphorylation of the kinase's substrates will be perturbed. In the case of the kinase, phosphorylation levels would be expected to decrease in response to down-regulation, and in response to phosphatase down-regulation, substrate phosphorylation levels would be expected to be higher than a control. Alternatively, if a key activator or mediator involved in regulating kinase/phosphatase activity is differentially expressed, there is the potential for downstream effects resulting in a change in phosphorylation levels. The aim is therefore to identify what genes are relevant to modulating the phosphorylation of the proteins of interest – whether these genes have a direct impact (kinases/phosphatases), or an indirect impact (activating/mediating proteins). The PhosphoPICK algorithm presented in Chapter 2 introduced a method for using protein-protein interaction and association networks to predict kinase substrates. A different application of this would be to use the interaction networks to identify the proteins relevant to determining a substrate's phosphorylation status; changes in such proteins expression levels

would be expected to effect the phosphorylation status of the substrate. Therefore, I hypothesised that by leveraging protein-protein interaction and association networks in combination with the data provided by sbv IMPROVER, a classification method could be trained to predict changes in phosphorylation status from the gene expression data.

This chapter presents a method for identifying the genes most relevant to the phosphorylation status of a phosphoprotein by cross-referencing differentially expressed genes with the protein-protein interaction network of the phosphoprotein. By observing what genes (in the interaction network) are differentially expressed under the same conditions that the phosphoprotein is differentially phosphorylated, the identified gene set was used to train a machine learning classifier. I show that by incorporating these gene sets in support vector machine or random forest classifiers, differential phosphorylation could be predicted with high accuracy as measured by AUC. I also perform a statistical analysis to demonstrate that the use of our identified "relevant genes" in a classifier performs significantly better than if a set of randomly selected genes are trained on the same classifier.

A phosphoprotein classifier trained on gene expression data could be used to infer the phosphorylation status of the phosphoprotein under differing treatment conditions. I trained such classifiers for the phosphoproteins in sub-challenge 1 (SC1), and used them to predict changing phosphorylation status from the test gene expression data provided. For sub-challenge 2 (SC2) I trained phosphoprotein classifiers on rat data, and used the classifiers to predict differential phosphorylation in the test set of human data.

3.3 Methods

3.3.1 Data provided by sbv IMPROVER

A detailed description of the data generated for the competition is available at (146), but an overview of the relevant data is provided here. The experiments were performed on normal human bronchial epithelial (NHBE) and normal rat bronchial epithelial (NRBE) cell lines. Phosphorylation levels of 16 phosphoproteins and genome-wide measurement of gene expression were measured under 52 different stimuli, or Dulbecco's Modified Eagle's Medium (DME) as a control condition. In order to divide the data into training data provided to competition participants and hold-out data, the experiment was divided into two sets of stimuli: 26 for the training set and 26 for the hold-out set.

Gene expression was measured using an Affymetrix™ chip, six hours after the cells had been exposed to a treatment. There were two or three biological replicates for each of the 52 treatments, and 4 or 5 replicates for the DME control. The experiment yielded gene expression levels for 13,841 rat genes and 20,110 human genes. Orthologs between rat and human proteins (totalling 12,458 orthologs) were also provided to the competition participants based on HGNC Comparison of Orthology Predictions.

Phosphorylation data was generated using the Luminex xMap™ technology (147), which uses beads coated in antibodies that can bind specific proteins – in this case, phosphorylated proteins. The phosphorylation status of the proteins was measured at two time intervals, 5 minutes and 25 minutes. Phosphorylation signals at the two time intervals were measured in triplicates for each stimulus, and in sextuplets for the DME control. As defined for the competition, a phosphoprotein was considered "activated" (i.e. differentially phosphorylated) if the absolute difference in Luminex xMap™ signal between the DME control and a treatment was greater than 3. A distinction was not made between phosphoproteins being "down" phosphorylated or "up" phosphorylated, however given the small number of differentially phosphorylated proteins (61/416 of the measured rat phosphoproteins in the training data were differentially phosphorylated, and 35/416 of the human) it was not feasible to divide the differentially phosphorylated proteins into two groups.

3.3.2 Additional data and classification tools

In addition to the data provided for the challenge, we sourced protein-protein interaction (PPI) data for both human and rat from the BioGRID database (1), as well as protein-protein association data from the STRING database (148). BioGRID provides binary interactions between proteins, while STRING scores associations between proteins based on multiple streams of evidence such as gene co-expression and literature mining.

I made use of two machine learning tools that have had great success when applied to a variety of biological problems: support vector machines (SVM) (98, 149) and random forest (RF). As the prediction problem here is a two-class discriminatory one, I was interested in using machine learning methods that could optimally separate the differentially phosphorylated samples from the non-differentially phosphorylated samples. SVMs and RFs are better suited to such a task than Bayesian networks, which are generative models. A description of SVMs is provided in Section 1.5.2. The SVM models presented in this chapter use the radial basis kernel function, which defines the distance between two input features x and x' as

$$\psi(x, x') = exp(-\alpha||x - x'||^2), \tag{3.1}$$

where α is a variable that can be adjusted.

RFs are based on decision classification trees. Employing a bootstrapping method, RFs construct an ensemble of trees that establish classification rules as determined by training data. A prediction on test data is made by averaging the results obtained across the ensemble of trees.

I used SVM and RF implementations from the Python machine learning toolkit (MILK, `http://luispedro.org/software/milk/`). Depending on the phosphoprotein, either the SVM or RF classifier might provide the best performance. Therefore each phosphoprotein was trained using both an SVM and an RF, and the optimal classifier was selected that obtained the highest AUC from a cross-validation experiment. The classifier that was selected for each of the phosphoproteins in SC1 and SC2 is listed in Table 3.1 and Table 3.2, respectively.

3.3.3 Predicting differential phosphorylation with gene expression

The purpose of sub-challenge 1 was to predict differential phosphorylation in the rat cell lines using GEx data also generated from rat. I identified relevant genes to each phosphoprotein by investigating the overlap between the protein-protein interaction and association networks of the phosphoproteins, and the genes observed to be differentially expressed under the same stimuli conditions that the phosphoproteins were differentially phosphorylated under.

Identification of genes relevant to phosphoproteins

To create vectors of gene expression values that could be used to train the SVM or RF, genes relevant to a phosphoprotein of interest first needed to be identified. This was done according to the following procedure. Firstly, the interaction network of the phosphoprotein was extracted from the BioGRID or STRING databases. As the rat BioGRID PPI network is limited compared to the human one, the human PPI was used after converting the human proteins to rat using the provided ortholog mappings. For the STRING database, scored associations were converted into a binary format using a cut-off threshold: for each phosphoprotein an association was retained if it obtained a score greater than 50%.

Secondly, the phosphoprotein interaction network was cross-referenced with genes differentially expressed under the same treatment conditions that the phosphoprotein was differentially phosphorylated (as defined in Section 3.3.1). To identify differential gene expression, Gaussian distributions of normalised (log-ratio of signal to control) gene expression values across treatment conditions were constructed. Applying the cumulative distribution function with a threshold probability γ, a gene was considered differentially expressed if the CDF-derived probability fell below γ. The value of γ will yield differing numbers of significant genes; too few genes could result in important information being missed, while too many could cause over-fitting of the model. Therefore, for each phosphoprotein, different values of γ were tested to determine the optimal value. Once the relevant gene set was identified for a phosphoprotein, feature vectors of normalised gene expression values across the treatment conditions were constructed for use as input features in training a classifier.

Model training and classification

Prediction accuracy of the classifiers was determined from leave-one-out cross-validation, and measured by calculating area under the receiver operating characteristic curve (AUC). Three of the phosphoproteins had zero or only one positive (differentially phosphorylated) samples under any of the treatment conditions, and were excluded from the analysis. The 13 remaining phosphoproteins were evaluated with different combinations of data source, classification tool and γ value to determine the optimal configurations, which are listed in Table 3.1.

To determine whether the prediction accuracy of the classifiers was due to the gene selection process, the accuracy was compared to a "random gene" model. A random gene model contained the same number of genes and other model parameters, but the gene set was chosen randomly. For each phosphoprotein, 1000 random gene models were generated, and a count made of the number of times the random model outperformed the interactions-based model, as measured by AUC. An empirical P-value was derived from the count divided by 1000 – the number of tests run. The P-value was therefore an indication of the reliability of the accuracy measure for a phosphoprotein model; i.e. a low P-value shows that the genes derived form the interaction/association networks are enabling the model's prediction accuracy.

Once the optimal configuration of classification method, data source and γ value was determined for each of the phosphoproteins, the method was used to predict phosphorylation status from the test gene expression data for SC1.

TABLE 3.1: Parameters for the phosphoprotein models for predicting phosphorylation status change in human proteins. If a phosphoprotein is marked as "N/A", that indicates that there were not sufficient (two or more) positive samples to train a model.

Protein	**Data source**	**Classifier**	**CDF sig. (γ)**
AKT1	STRING	SVM	0.00001
CREB1	STRING	SVM	0.0001
FAK1	BioGRID	SVM	0.001
GSK3B	STRING	RF	0.00001
HSPB1	N/A	N/A	N/A
IKBA	BioGRID	RF	0.0001
KS6A1	N/A	N/A	N/A
KS6B1	BioGRID	SVM	0.001
MK03	BioGRID	RF	0.001
MK09	STRING	SVM	0.0001
MK14K11	N/A	N/A	N/A
MP2K1	BioGRID	RF	0.0001
MP2K6	STRING	SVM	0.001
PTN11	STRING	RF	0.0001
TF65	BioGRID	SVM	0.0001
WNK1	BioGRID	SVM	0.001

3.3.4 Predicting human phosphorylation change from rat data

The goal of SC2 was to predict differential phosphorylation in human cell lines based on gene expression and/or phosphorylation levels from the rat cell lines. SC2 therefore presented the opportunity to test whether the method could work cross-species. Building on the method presented in Section 3.3.3, my approach was to first use human data to identify genes relevant to the human phosphoproteins, and then use the orthologous rat gene expression data to train models on predicting human phosphorylation status change. In addition, at the completion of SC1 participants were provided with the hold-out rat phosphorylation data. This provided the opportunity to evaluate the impact of supplementing the gene expression models with rat phosphorylation data in order to predict human phosphorylation status change.

Feature selection and model training

Human phosphoprotein gene sets were obtained based on the method described in Section 3.3.3, with a few adjustments to account for predicting phosphorylation status in the human cell line. Firstly, the human gene expression data was used to identify differentially expressed genes. Secondly, as the human STRING association networks were much larger than their

TABLE 3.2: Parameters for the phosphoprotein models for predicting phosphorylation status change in human proteins from rat data. If a phospho-protein is marked as "N/A", that indicates that there were not sufficient (two or more) positive samples to train a model. "+ Phos" indicates that phosphorylation status from the gold standard rat phosphorylation data was used.

Protein	**Data Source**	**Classifier**	**CDF Sig. (γ)**
AKT1	BioGRID	SVM	0.001
CREB1	BioGRID + Phos	RF	0.001
FAK1	N/A	N/A	N/A
GSK3B	BioGRID + Phos	RF	1.0E-5
HSPB1	N/A	N/A	N/A
IKBA	STRING (60)	SVM	0.0001
KS6A1	BioGRID + Phos	SVM	0.01
KS6B1	BioGRID + Phos	SVM	0.01
MK03	STRING (70) + Phos	RF	1.0E-9
MK09	N/A	N/A	N/A
MK14K11	N/A	N/A	N/A
MP2K1	BioGRID + Phos	RF	0.01
MP2K6	N/A	N/A	N/A
PTN11	BioGRID + Phos	RF	0.001
TF65	N/A	N/A	N/A
WNK1	N/A	N/A	N/A

rat counterparts, stricter cut-off thresholds of 60% or 70% were tested. For SC2 seven of the phosphoproteins had zero or only one example of differential phosphorylation, meaning that those phosphoproteins were excluded from the analysis.

Given a set of relevant genes for a phosphoprotein, the gene expression data from the rat orthologs was used to construct feature vectors for classifier training. The feature vectors could also be extended to contain the median Luminex xMap^{TM} signals (of protein phosphorylation status) for the rat cells for the two time points, 5 and 25 minutes. Cross-validation evaluation was used to determine if the phosphorylation data provided additional predictive power compared to using the gene expression data alone (Table 3.2).

Table 3.2 details what data source, classifier algorithm and γ value were used for each of the nine phosphoproteins in SC2. There were also parameters related to the SVM model that were tuned. The level of flexibility allowed in the SVM margins is defined by a parameter, C, which can be optimised. Cross-validation testing showed that setting the C value to 6 gave the optimal prediction accuracy across the phosphoproteins. In addition, the α parameter in the radial basis kernel function (Equation 3.1) was set to 2 after testing.

TABLE 3.3: Prediction accuracy (measured using AUC) for the phosphoprotein classifiers of phosphorylation status change for sub-challenge 1 and sub-challenge 2. P-values represent the observed frequency of a random-gene model outperforming the model derived from phospho-protein protein-protein interaction networks. If a phospho-protein is marked as "N/A", that indicates that there were not sufficient (two or more) positive samples to train a model.

	Challenge 1		Challenge 2	
Kinase	**AUC**	**P-value**	**AUC**	**P-value**
AKT1	0.86	0.075	0.88	0.01
CREB1	0.93	0.080	0.85	0.08
FAK1	0.65	0.30	N/A	N/A
GSK3B	0.88	0.020	0.74	0.04
HSPB1	N/A	N/A	N/A	N/A
IKBA	0.96	0.085	0.96	0.06
KS6A1	N/A	N/A	0.96	0.10
KS6B1	0.71	0.060	1.00	0.00
MK03	0.77	0.200	0.86	0.08
MK09	0.86	0.200	N/A	N/A
MK14K11	N/A	N/A	N/A	N/A
MP2K1	0.71	0.090	0.93	0.02
MP2K6	0.71	0.200	N/A	N/A
PTN11	1.00	0.090	0.96	0.02
TF65	0.91	0.180	N/A	N/A
WNK1	0.94	0.01	N/A	N/A

As a final step, the optimised phosphoproteins were evaluated on the test gene expression and phosphorylation data to predict phosphorylation status in the human cell lines.

3.4 Results

3.4.1 Predicting phosphorylation status change in rat cells

I evaluated the method for its ability to predict differential phosphorylation for 13 out of 16 of the phosphoproteins. Table 3.3 contains the set of AUC values found from performing leave-one-out cross-validation on the training data for each of the phosphoproteins, as well as the P-value representing the probability that the AUC value could be due to chance, rather than the set of genes included in the model. The results showed promising performance, with an average AUC of 0.84 across the proteins for the cross-validation test. For challenge 1, all but one of the phosphoproteins obtained P-values < 0.2, though I only found two examples of phosphoproteins (WNK1 and GSK3B) in sub-challenge 1 where the P-values obtained statistical significance (P

< 0.05). The average P-value across the 13 phosphoproteins was 0.14, meaning that in 86% of tests, the interaction models out-performed the random-gene models. The fact that the interaction models had greater accuracy for the vast majority of tests is a strong indication that the prediction accuracy of the method is mostly due to the gene selection process.

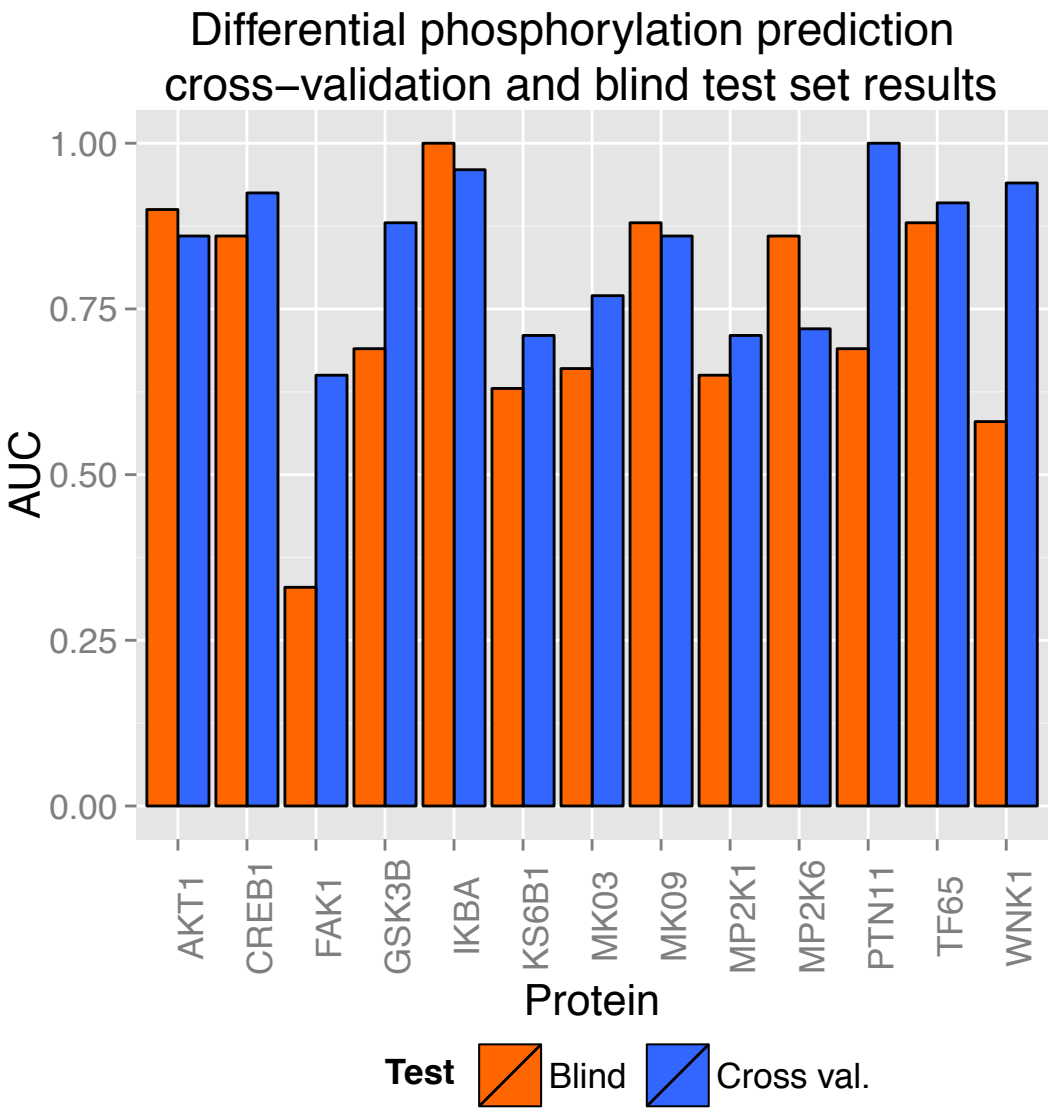

FIGURE 3.1: Comparison between prediction accuracy as measured by AUC for cross-validation testing, and the hold-out test set kept by SBV Improver.

I calculated the AUC values for each of the phosphoprotein test data sets provided after SC1, and compare the prediction accuracy obtained from the cross-validation tests to the accuracy on the blind test set (Figure 3.1). The method maintained a promising level of prediction accuracy, obtaining an average AUC of 0.74 for the blind test set, down from an average AUC of 0.84

TABLE 3.4: Rankings of participants in challenge 1. My entry is highlighted in bold font. Prediction accuracy was measured with three metrics: area under the precision-recall curve (AUPRC), Pearson correlation coefficient and balanced accuracy (BAC)

Rank	Team	AUPRC	Pearson	BAC
1	50	0.38	0.72	0.68
1	75	0.38	0.71	0.72
1	49	0.42	0.71	0.68
4	93	0.37	0.70	0.61
5	111	0.35	0.64	0.67
6	**89**	**0.31**	**0.65**	**0.65**
6	61	0.35	0.68	0.6
8	112	0.29	0.63	0.66
9	116	0.27	0.62	0.59
10	64	0.23	0.59	0.58
11	90	0.24	0.59	0.56
12	100	0.23	0.60	0.56
13	0.78	0.28	0.56	0.55
14	0.72	0.15	0.55	0.58
15	105	0.19	0.56	0.53
16	82	0.14	0.54	0.55
17	106	0.13	0.53	0.55
18	71	0.14	0.49	0.45
19	52	0.13	0.49	0.46
20	84	0.10	0.48	0.49
21	99	0.07	0.43	0.50

across the proteins for the cross-validation test. Many of the AUC values were fairly consistent between the cross-validation and blind tests, with some notable exceptions. Interestingly, while the prediction performance of some kinases decreased when tested on the blind test set – which is expected – there were a few cases where the performance was observed to increase when evaluated on the blind test set. MP2K6 was the most notable example, showing an increase in AUC from 0.72 to 0.86. WNK1 showed the sharpest decrease in prediction performance, from an AUC of 0.94 to a near-random value of 0.58. Surprisingly, WNK1 was also the phosphoprotein with the most significant P-value obtained from comparing the gene-selection accuracy with the random-gene model – this would intuitively make it the most likely phosphoprotein to maintain a similar level of prediction accuracy on the hold-out set.

Table 3.4 shows the rankings from challenge 1, where prediction performance was measured using a variety of accuracy metrics. My method was ranked 6^{th} place out of 21 teams that submitted predictions to the competition. The ranking demonstrates that the method was nonetheless amongst the best performing in the competition. Most importantly, as a proof of concept, the

results demonstrate the potential for using context information such a protein-protein interaction networks to inform the design of methods that can predict protein phosphorylation status change.

3.4.2 Predicting human phosphorylation change from rat data

For sub-challenge 2, I evaluated the ability of the method to predict differential phosphorylation in the human cell lines based on the rat data. Table 3.3 shows the AUC values (and level of significance) for evaluating the prediction accuracy of the method through leave-one-out cross-validation. The lack of differential phosphorylation examples in the human cell lines meant that I was only able to obtain results for 9/16 of the phosphoproteins. I found that the prediction performance for the 9 models was generally quite high, with all but one (GSK3B) of the phosphoproteins obtaining AUCs of greater than 0.85. Furthermore, for 8 of the phosphoproteins, the AUCs I obtained were found to be significant ($P < 0.05$). The high level of prediction accuracy and the confidence (based on low empirical P-values) that the performance is not due to chance, is an indicator that the method can be used to predict human phosphorylation status from observations made in rat cells.

For many of the phosphoproteins, the prediction accuracy observed for sub-challenge 1 was not reflected in sub-challenge 2. KS6B1 provided a particularly stark example, with an AUC of 0.71 in sub-challenge 1, but perfect prediction accuracy (with an AUC of 1 and a P-value of 0) for challenge 2. It is possible that the increased prediction accuracy is at least partly due to the inclusion of rat phosphorylation data – it is certainly counter-intuitive that the prediction accuracy for rat to human is generally higher than for sub-challenge 1 where the predictions were only in rat. For the phosphoproteins where I did not include phosphorylation data (AKT1 and IKBA), the prediction accuracy (measured by AUC) remained almost identical between sub-challenge 1 and sub-challenge 2 (Table 3.3).

While the competition participants were provided with the hold-out phosphorylation rat data after SC1 was completed, the hold-out human phosphorylation data was not made available. I was therefore unable to ascertain the prediction accuracy of the individual phosphoproteins on the holdout data for challenge 2. Table 3.5 shows the overall rankings and accuracy metrics that were provided by sbv IMPROVER. 13 teams entered SC2 and my method was ranked 7^{th}. The overall prediction accuracy of the method as measured on the test data was low compared to SC1, with a balanced accuracy of 0.58. However this could partly be a reflection of the fact

TABLE 3.5: Rankings of participants in challenge 2. My entry is highlighted in bold font. Prediction accuracy was measured with three metrics: area under the precision-recall curve (AUPRC), Pearson correlation coefficient and balanced accuracy (BAC)

Rank	Team	AUPRC	Pearson	BAC
1	50	0.54	0.75	0.77
2	111	0.41	0.68	0.76
3	49	0.33	0.72	0.68
4	61	0.34	0.70	0.66
5	52	0.31	0.69	0.71
6	93	0.29	0.67	0.59
7	**89**	**0.18**	**0.57**	**0.58**
7	116	0.28	0.68	0.55
9	112	0.13	0.55	0.56
10	97	0.10	0.54	0.57
11	90	0.12	0.55	0.53
12	105	0.06	0.47	0.45
13	84	0.06	0.45	0.42

that 7/16 of the phosphoproteins for SC2 did not have sufficient positive samples for training classifiers.

3.5 Discussion

Phosphorylation is an important regulatory mechanism for controlling protein function. Being able to predict changing phosphorylation states in proteins based on observation in gene expression level would be highly valuable; being able to infer the phosphorylation state of proteins in human based on observations in a rodent model even more so. The sbv IMPROVER challenge presented the opportunity to evaluate whether concepts underlying PhosphoPICK – that of using cellular context in the form of protein-protein interaction networks to predict phosphorylation substrates – could be used to predict changes in protein phosphorylation status based on gene expression. I have described in this chapter a method that can combine knowledge of a phosphoprotein's interaction and association networks with gene expression data to predict the changing state of phosphorylation levels with promising accuracy. The results validate a proof-of-concept that the interaction networks of phosphoproteins are useful features in designing methods that can predict phosphorylation status change given the expression levels of interacting proteins.

It was noted that even the top teams were only able to achieve a balanced accuracy of 70%, indicating that there are perhaps inherent limitations in the ability to predict phosphorylation status from gene expression (143). There were three teams that ranked first place in sub-challenge 1: teams 49, 50 and 75. The three teams employed quite diverse methodologies despite the similarity of their prediction accuracy on the test set. Team 49 ranked the gene expression data in a phosphoprotein-specific manner according to a moderated t-test P-value (143), identifying genes that underwent significant fold change in the same treatment conditions as the phosphoproteins. They used a linear discriminant analysis (LDA) model, which was fit to the training data by taking the top identified genes within a given fold change threshold – the threshold was optimised through cross-validation runs. Similar to my approach, they did not train classifiers for phosphoproteins with 1 or 0 examples of differential phosphorylation in the training set – for those phosphoproteins, all samples in the test set were set to 0. Team 50 employed two methods; one involved the calculation of mutual information between the phosphorylation status of each of the 16 phosphoproteins, and differential gene expression (binarised based on the results of a t-test). The second method employed a principal components analysis (PCA) on the gene expression data to identify leading principal components (PCs). An LDA model was fit to the training data as with team 49, with leave-one-out cross-validation used to optimise the number of PCs in the model. The final prediction was a weighted average of the scores generated from the two models. Team 75 created support vector regression models for each of the 16 phosphoproteins, with a feature selection process to select genes that should be included in the models. Cross-validation on the training set was used to identify the optimal number of genes for inclusion in the models.

While these methods obtained moderately better prediction accuracy then our own method, there are likely intrinsic limits to what can be achieved without considering additional information about the proteins under study. This is similar to observations that have been made in other areas of proteomics; for example it is recognised that only a small percentage of protein abundance levels can be explained from gene expression levels alone. Rather, predicting protein abundance from gene expression benefits from additional information from the protein and RNA level (150). Similarly, a method for predicting phosphorylation status that incorporates additional relevant information from the protein level would provide a better description of how phosphorylation levels can change in response to gene expression values.

The post-competition analysis also discussed the possibility that while it may be feasible to predict human phosphorylation status from rat phosphorylation data, the current limitations in computational tools mean that predictions do not benefit from the inclusion of gene expression data (151). This is in agreement with our results, which found that almost all the human

phosphoproteins benefited from the use of phosphorylation data in addition to gene expression for predicting phosphorylation in human cell lines from rat cell lines. My results for predicting human phosphorylation change from rat data seemed promising based on the training data, but I was unable to determined the performance of the individual phosphoproteins on the hold-out test data. As the data only allowed me to train models for half of the phosphoproteins, it is difficult to gauge whether the low balanced accuracy (58%) obtained from the hold-out evaluation performed by sbv IMPROVER is indicative of the accuracy of the trained models.

A major limitation in this study was the number of training samples available. We found that for many of the phosphoproteins there were only a couple of positive examples of differential phosphorylation that could be used for training. In challenge 2 there were 7 examples – nearly half – of the phosphoproteins that did not have any examples of differential phosphorylation associated with them. In such cases it was not possible to train and evaluate our model. Nonetheless, even with the data limitations, the results presented in this chapter demonstrate the potential for using methodology derived from PhosphoPICK to predict changes in protein phosphorylation states.

Chapter 4

Prediction of kinase-specific phosphorylation sites through an integrative model of protein context and sequence[1]

4.1 Abstract

The identification of kinase substrates and the specific phosphorylation sites they regulate is an important factor in understanding protein function regulation and signalling pathways. Computational prediction of kinase targets – assigning kinases to putative substrates, and selecting from protein sequence the sites that kinases can phosphorylate – requires the consideration of both the cellular context that kinases operate in, as well as their binding affinity.

We report here a novel probabilistic model for the classification of kinase-specific phosphorylation sites from sequence across three model organisms: human, mouse and yeast. The model incorporates position-specific amino acid frequencies, and counts of co-occurring amino acids from kinase binding sites in a kinase- and family-specific manner. We show how this model can be seamlessly integrated with protein interactions and cell-cycle abundance profiles. When evaluating the prediction accuracy of our method, PhosphoPICK, on an independent hold-out set of kinase-specific phosphorylation sites, we found it achieved an average specificity of 97%

[1]Chapter reproduced from paper of the same name currently in submission.

while correctly predicting 32% of true positives. We also compared PhosphoPICK's ability, through cross-validation, to predict kinase-specific phosphorylation sites with alternative methods, and found that at high levels of specificity PhosphoPICK outperforms alternative methods for most comparisons made.

We investigated the relationship between experimentally confirmed phosphorylation sites and predicted nuclear localisation signals by predicting the most likely kinases to be regulating the phosphorylated residues immediately upstream or downstream from the localisation signal. We show that kinases PKA, Akt1 and AurB have an over-representation of predicted binding sites at particular positions downstream from predicted nuclear localisation signals, demonstrating an important role for these kinases in regulating the nuclear import of proteins.

4.2 Introduction

Kinases regulate a wide variety of essential biological processes through protein phosphorylation, including transcription factor activity (152), the control of DNA damage repair pathways (153), the progression of cells through mitosis (57), and protein import into the nucleus (154). Knowledge of the kinases that regulate phosphorylation substrates is therefore a significant factor in understanding the functional consequences of protein phosphorylation events. While hundreds of thousands of phosphorylation sites have been identified across thousands of proteins (4), the kinases that regulate these sites in most cases remain unknown. Computational methods that predict kinase-specific phosphorylation sites are therefore an important contributor to understanding the role of phosphorylation events in biological processes (155). Such methods contribute to the guidance of phosphorylation experiments (156) and provide information about the likely signalling pathways that phosphorylation sites may be involved in (157).

Kinase-mediated phosphorylation is regulated by several important factors that can be leveraged to build predictive models. One is the sequence-level motifs surrounding phosphorylation sites that interact with kinase binding domains. The protein sequence determines whether a kinase can bind to the protein; previous studies have shown that local motifs surrounding a phosphorylation site interact with the binding domain of kinases to allow phosphorylation (25, 46). There are numerous kinase-specific phosphorylation site predictors that take advantage of the sequence specificity of kinases to predict kinase-specific phosphorylation sites (80, 95, 158) as well as phosphorylation sites in a non-kinase specific manner (16, 104).

The presence of valid kinase-binding motifs on a protein is no guarantee that a kinase will phosphorylate a substrate however (27). The targeting of phosphorylation substrates by kinases is subject to, and controlled by, a wide variety of processes within the cell – what may be called the "context factors" that ensure kinase-substrate fidelity. Context factors can include proteins that mediate the interaction between kinases and their substrates (20), activating proteins such as cyclins (52), sub-cellular compartmentalisation (159) and the various stages within the mitotic cell cycle (160).

We have shown previously that context information (in the form of protein-protein interaction and association data, as well as protein abundance levels across the cell cycle) can be incorporated into a probabilistic model that maps kinases to putative substrates (161). This model not only provides an accurate predictor of kinase substrates, but importantly, the sequence-level prediction of kinase-specific phosphorylation sites can be greatly enhanced by the model's additional predictive power. While this model was able to use context alone to predict kinase substrates, we hypothesised that the incorporation of sequence and context into a single model would provide better explanatory power of the factors that describe kinase targets.

In this paper, we present a novel probabilistic method for predicting kinase-specific phosphorylation sites that incorporates position-specific amino acid frequencies and counts of co-occurring neighbouring amino acids in a family-specific manner across three model organisms: human, mouse and yeast. We demonstrate that this sequence model can be used as a module within a larger Bayesian network that describes the context factors that influence how a kinase targets a protein substrate. The seamless integration of these two domains of information – context and sequence – allows for a comprehensive model of kinase-protein phosphorylation. We compare the ability of our method, PhosphoPICK, to predict kinase-specific phosphorylation sites against alternative phosphorylation predictors, and show that PhosphoPICK has a superior ability to predict kinase-specific phosphorylation sites for most comparisons made.

As we now have a predictor that ably integrates the context and sequence conditions that regulate phosphorylation, we are in a position to investigate phosphorylation-dependent functions and probe the kinases that are involved in regulating these functions. The nuclear import of proteins is a highly-specific process, involving the binding of importin proteins to cargo proteins that contain a relevant nuclear localisation signal (NLS) (162, 163). It has been shown that the binding of importin proteins to their cargo can be controlled (promoted or inhibited) by the presence of phosphorylation adjacent to the NLS (164). We therefore investigated the relationship between nuclear localisation signals and phosphorylation by cross-referencing experimentally identified phosphorylation sites with predicted NLSs. We used PhosphoPICK to

identify the most likely candidate kinases for NLS-adjacent phosphorylation sites, and performed a statistical analysis to identify sites relative to NLSs that have an over-representation of kinase binding sites. We identify several kinases as candidates to regulate phosphorylation sites at sites downstream from the NLSs, most notably protein kinase A (PKA), Akt1 and Aurora kinase B (AurB). We also identify kinases that regulate sites upstream from the NLS, including cyclin dependent kinase 2 (CDK2). Gene ontology (GO) term enrichment analyses indicate that the phosphorylation of specific sites close to the NLS by these kinases regulates distinct biological functions.

4.3 Methods

4.3.1 Data resources

We obtained kinase-specific phosphorylation data for human and mouse from PhosphoSitePlus®, www.phosphosite.org (4) and for yeast (Saccaromyces cerevisiae) from PhosphoGRID (73), which is a database of *in vivo* phosphorylatio sites. For data collected from PhosphoSitePlus®, we ensured that phosphorylation sites used were known to occur *in vivo*, but for both databases, the kinase annotations are often informed by *in vitro* or *in vivo* experiments. We chose phosphorylation site data for kinases where there were greater than 5 unique kinase substrates, resulting in 5,209 kinase-specific phosphorylation sites across 1,826 proteins for human, 956 kinases-specific phosphorylation sites across 417 proteins for mouse, and 2,219 kinase-specific phosphorylation sites across 722 substrates for yeast. In order to have a more extensive background of phosphorylation events for training a sequence model, we also used phosphorylation sites that did not have a kinase assigned to them. We used phosphorylation sites from PhosphoSitePlus® that were generated using low-throughput methods; similarly for PhosphoGRID, sites were included if they were identified using more than one method, or if the single detection method was not mass spectrometry. This resulted in an additional 5,939 phosphorylation sites for human, 2,865 additional phosphorylation sites for mouse and 674 additional phosphorylation sites for yeast.

Protein-protein interaction (PPI) data was sourced from BioGRID (1), protein-protein association data from STRING (148), and protein abundance data across the cell cycle from the work by Olsen and colleagues (28). As the cell-cycle information was only available for human, cell-cycle data was not incorporated into the mouse or yeast kinase models. A detailed description of how this data was curated and processed is available in (161).

In order to evaluate the prediction accuracy of our method on completely novel data, we created a hold-out set for kinases for which there were more than 100 known substrates – there were nine such human kinases. For each of the nine kinases, we selected a random set of substrates equal to 10% of that kinase's substrates that were *not* in the original set of substrates used for developing the model (161). These substrates were excluded from all analyses and simulations, and were used only for a final evaluation of model accuracy. This resulted in a hold-out set of 145 proteins – containing 416 phosphorylation sites specific to the nine kinases. After removing the hold-out set, a set of 1,671 human proteins and 4,907 kinase-specific human phosphorylation sites remained for training and testing.

In addition, we built similarity-reduced sets of the phospho-peptide sequences obtained from PhosphoSitePlus and PhosphoGRID in order to determine whether sequence similarity could be inflating prediction accuracy. The BLASTP program (165) was used to perform a pairwise sequence similarity comparison of each of the phospho-peptides, using 15-residue sequences centred on the phosphorylation site. All 15-residue pairs obtaining a BLASTP E-value under 0.05, with sequence identity of at least 30%, were retained. Similar pairs within the same kinase category were reduced through the arbitrary removal of one of the phospho-peptides; phospho-peptides that were similar, but phosphorylated by different kinases, were not reduced. The similarity reduction was also applied to the background set of peptides.

4.3.2 PhosphoPICK method and workflow

Building on our existing context model, we developed a model for predicting kinase-specific phosphorylation sites from sequence, as well as a model that incorporates this sequence model into the context model described in our previous work.

Sequence model

We present a Bayesian network model for modelling various sequence features of a kinase binding motif (Figure 4.1). We represent potential amino acid residues in an n length sequence motif surrounding a phosphorylation site as discrete variables conditioned on two Boolean variables. The first represents the event that some kinase of interest, K, binds to the site, the second represents the event that a family member (i.e. any family member of K) binds to the site. Each variable – R_{-m} to R_{+m}, where R_0 represents the site for which phosphorylation is predicted – contains three distributions of amino acid frequencies. These represent (1) the probability

of each amino acid occurring at the position where K is seen to be phosphorylating, (2) the amino acid frequencies for binding sites from the family members of K, and (3) the amino acid frequency background as seen across all other phosphorylation sites in the training set.

In addition to position-specific amino acid frequencies, we included k-mers of k=2 (dimers) and k=3 (trimers) to encode the frequency of co-occurring neighbouring amino acids. This should allow the model to capture some paired dependencies that may exist between amino acids. In order to avoid over-parameterising the sequence model with all possible combinations of dimers and trimers, we only added the k-mers that were observed in some θ percentage of kinase binding motifs from a training set. During cross-validation, the training set of kinase-binding motifs was taken, and k-mers observed within the motifs were counted. If a k-mer occurred in more than the θ percentage threshold of substrates, the k-mer was added to the model. We tested three cut-offs of θ: 5, 10 and 20, and found that 5 gave the best prediction accuracy across the full set of kinases (see Table B.1 for results across the set of human kinases, Table B.2 for mouse kinases and Table B.3 for yeast kinases). As shown in Figure 4.1, the k-mers are represented as a series of n Boolean variables, $Kmer_1$ to $Kmer_n$, where a k-mer is considered to be `true` if it is observed in the amino acid motif surrounding the phosphorylation site. The k-mer nodes were trained to capture the probability of each k-mer occurring within a kinase's binding motif, that of its family members and the background set of phosphorylation sites.

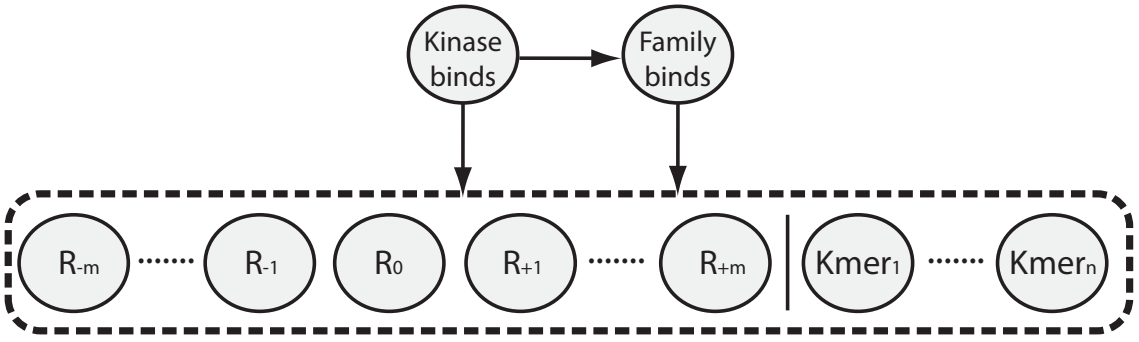

FIGURE 4.1: Sequence model. R nodes represent positions in a motif surrounding the phosphorylation site, where R_0 is the potential phosphorylation site. $Kmer_1$ to $Kmer_n$ represent the dimer and trimer configurations incorporated into the model.

It has been shown previously that varying the motif length in predicting kinase binding sites improves prediction accuracy (80). Therefore, for each kinase we tested five different window sizes centred around the phosphorylated residue: 7, 9, 11, 13 and 15. For each kinase we selected the window size that gave the best prediction accuracy as measured within a cross-validation test (Tables B.4, B.5 and B.6).

Combined model

The combined model retains the structure of the "context" Bayesian network described previously (161), but with the sequence model incorporated into it. This model represents observations about kinase-substrate phosphorylation events, protein-protein interaction/association events believed to be relevant to kinases encoded in the model, and cell-cycle profiles of substrates as Boolean variables. A connection between a kinase and a PPI event is defined if the protein is interacting with at least 5 of the kinase's substrates. Up to 25 connections between a kinase and a PPI event can be defined.

The sequence model was incorporated into the larger context model in a kinase-specific manner, such that for each kinase the kinase target variable in the sequence model is conditioned on the variable in the context model representing the kinase phosphorylating a substrate (Figure 4.2). We created models based on sets of kinases as they are classified into family similarity (32). For human, we created eight family-specific models comprising kinases from the CMGC (cyclin-dependent, mitogen-activated, glycogen synthase and Cdc2-like), AGC (protein kinase A, G and C families), CAMK (Ca^{2+}/calmodulin-dependent kinase), TK (tyrosine kinase), "other", STE, CK1 (cell kinase 1) and atypical kinase families. For mouse, we created three models with kinases from the CMGC, AGC and TK families; and for yeast we created four models from the CMGC, AGC, CAMK and other kinase families.

4.3.3 Setting non-query kinase nodes

The model relies partly on the expected activity of alternative kinases that are encoded in the Bayesian network. However, there is no experimental information on kinase binding events for the majority of proteins, and negative evidence (a protein *not* being phosphorylated by a particular kinase) is non-existent. Therefore we employ the amino acid sequence of a query protein to estimate what kinases in the model will not bind to the protein, and can therefore be set to `false`. In order to decide when kinase variables in the model should be set to `false`, the following steps were followed for each non-query kinase. Within a training fold, the positive training samples for that kinase were set aside. 75% of the substrates within the negative set were selected randomly, and each phosphorylation site within this set was added to the training data, while the remaining substrates were set aside as a test set.

The sequence model was trained using the selected training samples, and used to scan over each of the substrates within the test set, with the highest score for each of the substrates

recorded. The median value of these scores was then taken as a threshold representing the highest expected score for a protein that is not phosphorylated by the kinase. When evaluating the model on a test substrate, for each non-query kinase node, its sequence model was used to scan the substrate and the highest score is recorded. If the score falls below the calculated threshold value, that kinase node is set to `false`, otherwise it remains unspecified.

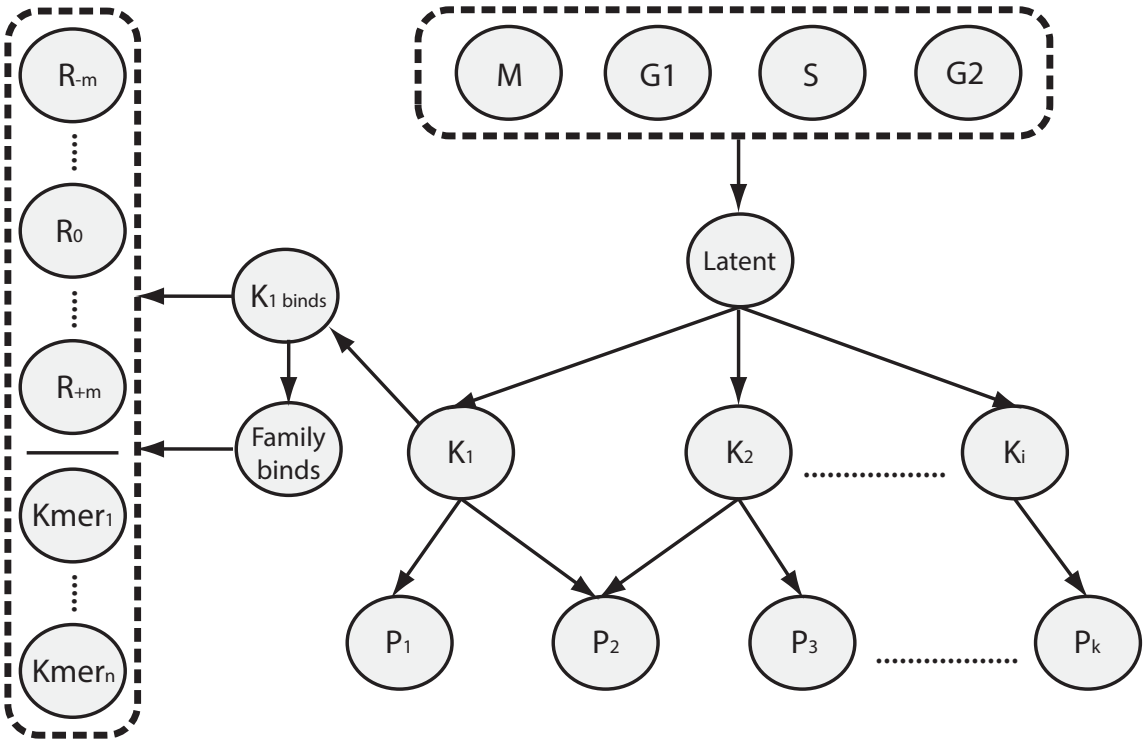

FIGURE 4.2: PhosphoPICK Bayesian network model incorporating both context and sequence data. The bottom layer of nodes (P_1 to P_k) represent protein interactions incorporated into the model. These are conditioned on relevant kinases (K_1 to K_i), which are themselves conditioned on a latent node incorporating variables representing the four cell cycle stages. The K_1 *binds* "sequence" variable is conditioned on its corresponding K_1 "context" variable.

Prediction workflow

A diagram illustrating the PhosphoPICK workflow for generating a prediction is shown in Figure 4.3. To determine the probability of a query kinase phosphorylating a given substrate, the relevant context data are queried and the corresponding nodes in the Bayesian network are instantiated. As there is no experimental information on kinase binding events for the majority of proteins, and negative evidence (a protein *not* being phosphorylated by a particular kinase) is non-existent, the protein sequence is used to provide an estimate of what alternative kinases will *not* phosphorylate the given substrate (Section 4.3.3).

The model is then scanned over the substrate's amino acid sequence, and for every potential phosphorylation site, the n length motif corresponding to the query kinase surrounding the phosphorylation site is used to set the sequence nodes in the network. For every potential phosphorylation site, the node representing the kinase phosphorylating a substrate is queried, and the highest probability for the scan is taken as the score for that substrate. Separately, the potential phosphorylation sites within the substrate are scored using the sequence model. The final score for a kinase-specific phosphorylation site prediction is equal to the average of the substrate score from the combined model, and the site score from the sequence model.

4.3.4 Model training

Sequence model

The nodes in the sequence Bayesian network are defined using conditional probability tables (CPTs), which learn from training data all possible values that a variable can take, given the set of parents it is conditioned on. If a variable does not have parents, the CPT will represent the observed frequency from the training data of it being true. As there may be amino acids or k-mers that do not occur in some of the training data, we added a uniform pseudo-count of 0.05 to all the amino acid and k-mer nodes, ensuring that the model does not consider some amino acids or k-mers impossible to occur.

Combined model

The nodes in the combined model are defined using CPTs and our variation on the NoisyOR node (161), which allows for an approximation of a CPT. The protein interaction nodes were defined using NoisyOR variables, allowing parameters to be inferred even in the case of data sparsity. All other variables in the combined model were defined as CPTs.

As the combined model incorporates data representing different problems – that of predicting kinase substrates, and predicting kinase binding sites, the model was trained in two stages. First, the set of unique substrates was presented for expectation maximisation training (121) in order to set the parameters for the protein-interaction, cell-cycle and kinase nodes in the network. The parameters for these variables were then locked in place. Next, the sequence module within the network was trained using the set of phosphorylation sites contained in the training fold, with the position-specific amino acid nodes and k-mer nodes being set as for the

sequence model. There will be some cases in the phosphorylation site data where a kinase will be phosphorylating a substrate, but not the site. In these cases, the node representing the kinase binding the substrate was set to `false`.

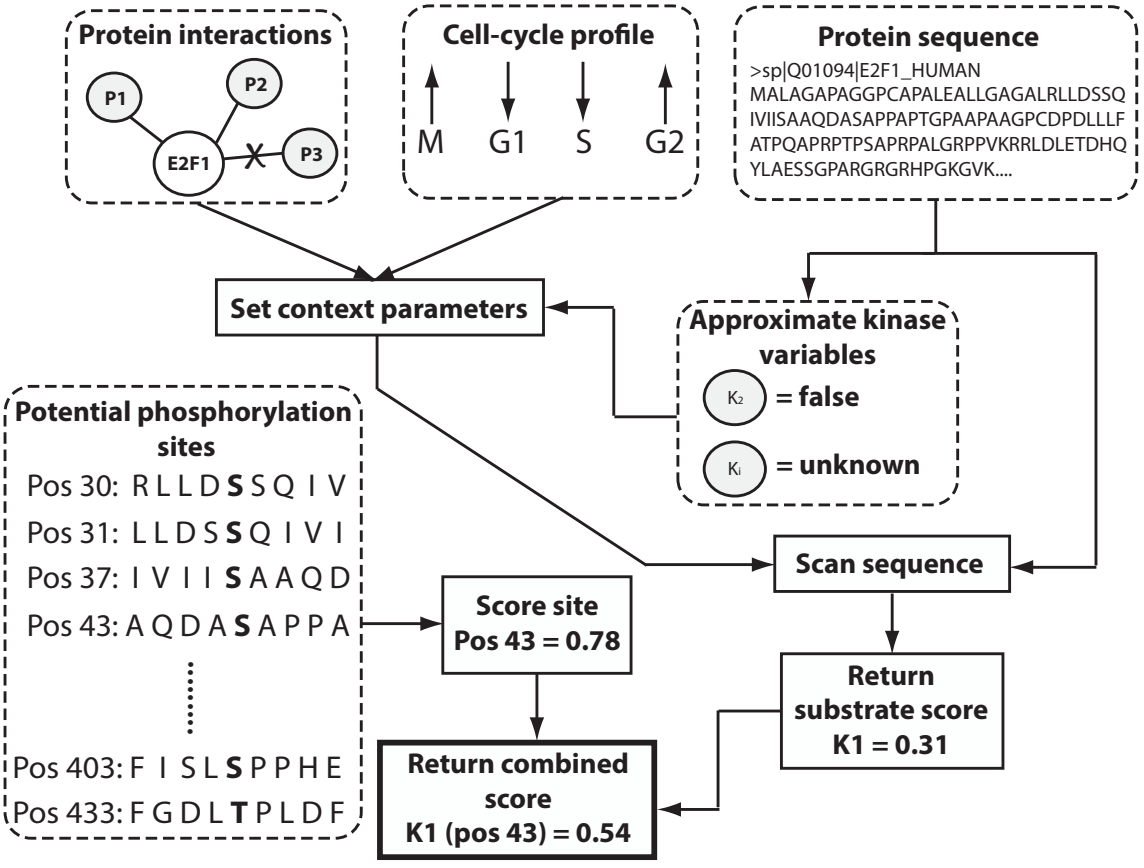

FIGURE 4.3: Diagram showing the workflow involved when a kinase is queried for a protein submitted to the model. BioGRID and STRING are queried to identify what proteins the substrate interacts with, and the protein-interaction variables are set accordingly. If cell-cycle data is available, it will be included also. The substrate sequence is used to estimate what kinases in the model will *not* bind to the substrate, with the remainder left unspecified. The model is then scanned across the sequence to identify the highest probability of the kinase phosphorylating the substrate. Separately, the sequence model is used to score all potential sites in the query substrate. The final prediction for a potential phosphorylation site is the average of the substrate and site score.

4.3.5 Evaluating model prediction accuracy

The prediction accuracy of the models was evaluated across the 107 human kinases, 24 mouse kinases and 26 yeast kinases using ten-fold cross-validation across ten randomised data-set splits. The prediction accuracy of the sequence model was evaluated by its ability to correctly

classify kinase-specific phosphorylation sites out of the set of known kinase-binding sites, and the combined model was evaluated by its ability to correctly classify kinase substrates out of the set of substrates.

To ascertain the effect that our sequence model features have on prediction accuracy, we evaluated the accuracy of a simple baseline sequence model that only contained the position-specific amino acid nodes conditioned on the kinase variable (the family variable was excluded). We also evaluated the prediction accuracy of the context model (the combined model excluding the sequence information) and compared its accuracy with the combined model to ascertain what improvement may be gained from incorporating sequence and context information into a single model. Prediction accuracy was determined using receiver operating characteristic (ROC) and calculation of area under the ROC curve (AUC) as a measure of overall model performance (122). We also calculated area under the ROC curve up to the fiftieth false positive (AUC50) as a measure of performance at low false-positive levels.

Comparisons to alternative methods

We compared the ability of the complete PhosphoPICK work-flow to predict kinase-specific phosphorylation sites out of all potential phosphorylation sites in the substrate sequences. The comparison was performed firstly against the sequence model only, and secondly against three alternative methods that have a larger number of kinases available for making predictions: GPS 2.1 (80), NetPhorest 2.0 (97) and NetworKIN 3.0 (97). We downloaded the standalone prediction software for each of the three methods and ran the set of 1,671 proteins through them. For NetworKIN and NetPhorest, we did not specify the sites we wanted predictions for. We used GPS's batch prediction system to run GPS on the protein set, selecting the "no threshold" option.

In order to compare PhosphoPICK predictions to the alternative methods, we again did a 10x ten-fold cross-validation run of the combined model as well as of the sequence model. As most of the potential phosphorylation sites in the substrates were not in the set of peptides used for training the sequence model (and therefore not part of the cross-validation run), the fully trained sequence model was used to score potential phosphorylation sites outside of the training set.

Due to the large number of potential phosphorylation sites being scored (~170,000 S/T sites and ~30,000 Y sites), we calculated sensitivity for two stringent levels of specificity – 99.9% and

99%. The difference in sensitivity between PhosphoPICK and each alternative was calculated across all ten cross-validation runs.

Calculating significance of predictions

Users of the PhosphoPICK web-server are provided with an option to include empirical P-value calculations alongside their predictions, allowing for a measure of the significance of the predictions. To obtain empirical P-values, we first calculated proteome-wide distributions of predictions; i.e. for all kinases, substrate predictions were obtained for every protein in the relevant proteome (human, mouse or yeast), and site predictions were made for all potential phosphorylation sites in the proteome. To calculate a combined P-value for a prediction, Fisher's method for combining probabilities was applied such that:

$$X = -2(ln(P_{context}) + ln(P_{site})),$$

where $P_{context}$ and P_{site} represent the P-value value calculated for a context score given to a substrate and a motif score given to a site respectively, and X follows a Chi squared distribution with 4 degrees of freedom.

4.3.6 Evaluation on hold-out set

When evaluating the performance of the model on the hold-out set, the full sets of training data was used to train the model. We predicted each potential phosphorylation site (all S/T residues for serine/threonine kinases and all Y residues for the tyrosine kinase Src) in the hold-out sequences, and evaluated the performance of the model for each kinase by its ability to predict the kinases' phosphorylation sites out of all potential sites. In order to evaluate how well the method would be expected to perform using the P-value based thresholding system on the web-server, P-values were calculated for the predictions, and if a P-value for a prediction fell below 0.005 the prediction was considered to be `true`, and `false` otherwise.

We calculated sensitivity, specificity, balanced accuracy (BAC) and Matthews' correlation coefficient (MCC). The metrics are defined as follows, where TP is the number of true positives, FP the number of false positives, TN the number of true negatives, and FN the number of false negatives.

Sensitivity:

$$sens. = \frac{TP}{TP + FN}$$

Specificity:

$$spec. = \frac{TN}{TN + FP}$$

Balanced accuracy:

$$BAC = \frac{sensitivity}{specificity}$$

Matthews' correlation coefficient:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

4.4 Results

4.4.1 Sequence model for classifying kinase binding sites

The sequence model was evaluated by its ability to correctly classify, on a per-kinase basis, kinase-specific phosphorylation sites out of the set of known kinase binding sites. Table 4.1 shows results for an example set of kinases (the CMGC family of kinases), and Table 4.2 contains the averaged prediction accuracy for each of the kinase families across the three tested species. The full set of values are available in Tables B.7, B.8 and B.9. The sequence model has good prediction accuracy over the kinases tested, with an average AUC of 0.79 across all human kinases. We found that 66% of kinases obtained an AUC of greater than 0.75, demonstrating that the model works well for the majority of kinases. We noticed particularly high accuracy for the CMGC kinases, where 17/20 of the kinases in this family obtained an AUC of greater than 0.8 (Table 4.1); and also the atypical kinases, where all of those kinases obtained an AUC greater than 0.8, and 3/4 greater than 0.85 (Table B.7). The worst performing family appeared to be the tyrosine kinase family, where we found an average AUC of 0.62 – substantially lower than the overall average (of 0.79), and much lower than the accuracy from the various serine/threonine kinase families.

We compared the sequence model against a baseline model that only considered the position-specific amino acid frequencies. While the sequence model outperforms the baseline in general, we noticed that there was substantially higher accuracy at low false-positive levels as measured by the AUC50. In the "other" family of kinases, there was a greater than 3-fold increase in the AUC50, and in the CMGC and CK1 families we found a greater than 2-fold increase in AUC50.

TABLE 4.1: Comparison of prediction accuracy across human CMGC kinases between predicting kinase-specific phosphorylation sites with a baseline model that only considers position-specific amino acid frequencies, and the sequence model. Results were generated using ten-fold cross-validation repeated across ten randomised data-set splits. Shown are the average and standard deviation of the AUC and AUC50 values.

	AUC		AUC50	
Kinase	Baseline	Sequence model	Baseline	Sequence model
CDK2	0.86±0.001	0.89±0.001	0.06±0.002	0.10±0.004
CDK1	0.88±0.002	0.89±0.002	0.09±0.004	0.07±0.008
ERK2	0.86±0.002	0.86±0.001	0.05±0.004	0.07±0.010
ERK1	0.86±0.005	0.86±0.005	0.04±0.005	0.07±0.012
GSK3B	0.77±0.009	0.81±0.006	0.09±0.007	0.13±0.014
P38A	0.79±0.007	0.81±0.007	0.12±0.016	0.15±0.017
JNK1	0.83±0.005	0.87±0.004	0.08±0.013	0.15±0.014
CDK5	0.84±0.012	0.84±0.009	0.07±0.009	0.05±0.007
JNK2	0.75±0.015	0.73±0.023	0.03±0.013	0.07±0.015
CDK7	0.77±0.017	0.88±0.019	0.16±0.044	0.31±0.032
GSK3A	0.89±0.014	0.90±0.026	0.26±0.020	0.46±0.045
CDK4	0.85±0.012	0.87±0.012	0.07±0.007	0.18±0.025
P38B	0.79±0.006	0.83±0.014	0.07±0.015	0.26±0.046
HIPK2	0.81±0.016	0.86±0.013	0.23±0.030	0.38±0.043
DYRK1A	0.77±0.034	0.83±0.033	0.01±0.024	0.26±0.043
CDK9	0.78±0.011	0.83±0.015	0.04±0.022	0.32±0.030
DYRK2	0.68±0.032	0.78±0.019	0.00±0.000	0.31±0.043
ERK5	0.79±0.015	0.83±0.016	0.02±0.014	0.32±0.034
CDK6	0.80±0.019	0.86±0.009	0.07±0.016	0.18±0.030
CDK3	0.69±0.031	0.76±0.050	0.00±0.000	0.36±0.045

On the mouse kinases, the model achieved a more moderate average AUC of 0.71, reflecting the diminished availability of positive training data when compared to human or yeast kinases.

Similar to the results seen in the human kinases, however, the CMGC kinases performed the best, with an average AUC of 0.79, and the tyrosine kinases were again the worst performing, with an average AUC of 0.63.

TABLE 4.2: Performance comparisons between predicting kinase-specific phosphorylation sites with a baseline model that only considers position-specific amino acid frequencies, and the sequence model. Results were generated using ten-fold cross-validation repeated across ten randomised data-set splits. Shown are the average and standard deviation of the AUC and AUC50 values.

	AUC		AUC50	
Family	Baseline	Sequence	Baseline	Sequence
Human				
CMGC	0.80±0.013	0.84±0.014	0.08±0.013	0.21±0.026
AGC	0.76±0.017	0.79±0.018	0.15±0.028	0.21±0.029
TK	0.56±0.022	0.62±0.025	0.11±0.021	0.18±0.024
CAMK	0.73±0.023	0.77±0.024	0.11±0.014	0.19±0.027
Other	0.69±0.019	0.80±0.021	0.07±0.013	0.32±0.038
STE	0.71±0.031	0.79±0.052	0.23±0.049	0.38±0.053
CK1	0.75±0.020	0.86±0.025	0.12±0.019	0.30±0.031
Atypical	0.84±0.009	0.87±0.008	0.18±0.008	0.20±0.030
Mouse				
CMGC	0.74±0.016	0.79±0.016	0.14±0.017	0.24±0.029
AGC	0.72±0.025	0.75±0.032	0.17±0.034	0.26±0.051
TK	0.60±0.025	0.63±0.029	0.26±0.032	0.31±0.026
Yeast				
CMGC	0.67±0.028	0.76±0.028	0.11±0.007	0.32±0.030
AGC	0.79±0.020	0.85±0.025	0.24±0.027	0.46±0.034
CAMK	0.64±0.024	0.78±0.024	0.05±0.017	0.34±0.037
Other	0.74±0.017	0.84±0.023	0.10±0.010	0.35±0.035

The yeast kinase models performed quite well, achieving an average AUC of 0.81. In yeast, the best performing kinases were from the AGC family, with an average AUC of 0.85, and an AUC50 exceeding any other kinase family from mouse or human. We noticed that the sequence model had a substantial increase in accuracy when compared to the baseline – particularly at the low false-positive rates as measured by AUC50. The CAMK kinases recorded the sharpest increase, with an average AUC50 of over 6-fold greater than the baseline model. In general, we found that the use of k-mers offered a great advantage over the simpler representation of position-specific amino acid frequencies, and that this was particularly noticeable at low false-positive levels. Our results indicate that our combination of features offers a highly accurate model for predicting kinase phosphorylation sites across diverse kinase families and species.

In order to test whether sequence similarity within the phospho-peptides could be inflating prediction accuracy, we re-trained the sequence model on the similarity reduced data-set. Table B.10, Table B.11 and Table B.12 contain a comparison of the fully trained sequence model and the model trained on the reduced data-set. For the majority of kinases, the similarity reduction did not result in a decrease in AUC. On average, there was a negligible difference in AUC, with an average decrease across all kinases of 0.004 seen with the reduced data set. Similarly, differences in the average AUC50 were slight, and within the margin of error. This demonstrates that the prediction accuracy of the sequence model is not due to homologous phospho-peptides in the training data, and can be applied to unseen samples.

4.4.2 Kinase substrate prediction

We compared the ability of the context model to predict kinase substrates against the combined (context plus sequence) model. Table 4.3 shows AUC and AUC50 values from the CMGC family of kinases, with averaged results across the kinase families summarised in Table 4.4, and the full set of results for all kinases available in Tables B.13, B.14 and B.15. The results demonstrate that across the kinase families, the incorporation of sequence data improved the ability of the model to predict kinase substrates. We noticed larger increases in prediction accuracy for the human CMGC, AGC and CAMK kinase families: the average AUC50 for CMGC increased from 0.31 to 0.43, AGC saw a similar increase from 0.21 to 0.34 and CAMK the largest – from 0.25 to 0.40. Figure 4.4 shows ROC plots, calculated up to 50 false positives (representing the AUC50 score), for kinases from the human CAMK family. ROC plots showing the prediction accuracy of the combined vs context model for the full set of kinases is available in Figure B.1 – Figure B.15. The ROC curves in Figure 4.4 demonstrate that the combined model is able to provide a substantial improvement in prediction accuracy at low false-positive rates for many of the CAMK kinases, when compared to the context model.

Table 4.3: Combined model accuracy across human CMGC kinases when compared to the context only model. Table shows accuracy values for classifying kinase substrates with both models as determined by 10-fold cross-validation across 10 randomised data-set splits. Prediction accuracy is measured using median and standard deviation of the AUC and AUC50 across the data-set splits.

	AUC		AUC50	
Kinase	Context model	Combined model	Context model	Combined model
CDK2	0.69±0.003	0.76±0.002	0.097±0.0016	0.110±0.0024

Continued on next page

Kinase	Context model	Combined model	Context model	Combined model
			Continued from previous page	
CDK1	0.77±0.002	0.79±0.002	0.088±0.0035	0.101±0.0036
ERK2	0.74±0.002	0.78±0.003	0.139±0.0022	0.155±0.0047
ERK1	0.78±0.003	0.81±0.003	0.125±0.0021	0.147±0.0048
GSK3B	0.74±0.002	0.79±0.005	0.151±0.0015	0.178±0.0032
P38A	0.80±0.003	0.80±0.006	0.132±0.0012	0.167±0.0115
JNK1	0.84±0.002	0.87±0.010	0.263±0.0021	0.310±0.0097
CDK5	0.78±0.006	0.82±0.007	0.183±0.0059	0.230±0.0081
JNK2	0.83±0.008	0.89±0.022	0.216±0.0113	0.313±0.0247
CDK7	0.93±0.034	0.95±0.048	0.560±0.0117	0.705±0.0327
GSK3A	0.81±0.042	0.91±0.028	0.378±0.0258	0.610±0.0551
CDK4	0.87±0.002	0.88±0.006	0.309±0.0263	0.494±0.0219
P38B	0.78±0.071	0.75±0.058	0.198±0.0330	0.410±0.0466
HIPK2	0.89±0.033	0.98±0.054	0.365±0.0155	0.780±0.0618
DYRK1A	0.92±0.032	0.90±0.015	0.698±0.0361	0.617±0.0257
CDK9	0.96±0.045	0.90±0.043	0.548±0.0175	0.656±0.0348
DYRK2	0.63±0.038	0.91±0.010	0.363±0.0098	0.849±0.0552
ERK5	0.82±0.078	0.97±0.141	0.549±0.0270	0.709±0.1387
CDK6	0.83±0.012	0.82±0.010	0.539±0.0201	0.698±0.0172
CDK3	0.54±0.047	0.57±0.064	0.284±0.0473	0.407±0.0822

While the context information accounts for the bulk of the accuracy, there were several examples of kinases where including the protein sequence in the model greatly improved prediction accuracy. In a few instances, prediction accuracy was increased from low or even random to a much higher value; for example the PKCI kinase improved from an AUC of 0.50 to an AUC of 0.77 (Table B.13), and DYRK2 obtained a huge increase from an AUC of 0.63 to 0.91 (Table 4.3). There were also several examples of substantial accuracy gains, even when the kinase already had moderate to high accuracy in the context model; we observed that the prediction accuracy of GSK3A increased from 0.81 to 0.91, tyrosine kinase Syk increased from 0.81 to 0.90 and CAMK kinase Pim1 increased from 0.8 to 0.94. While there were examples of prediction accuracy decreasing when sequence information was added, these decreases were slight, indicating that the accuracy gains for incorporating sequence and context information far outweigh any potential losses.

TABLE 4.4: Performance comparisons between predicting kinase substrates with the context Bayesian network model, and with the combined sequence & context model. Results were generated using ten-fold cross-validation repeated across ten randomised data-set splits. Shown are the average and standard deviation of the AUC and AUC50 values.

	AUC		AUC50	
	Context	Combined	Context	Combined
Human				
CMGC	0.80±0.023	0.84±0.027	0.31±0.015	0.43±0.032
AGC	0.74±0.025	0.79±0.029	0.21±0.015	0.34±0.035
TK	0.81±0.027	0.82±0.026	0.31±0.020	0.39±0.039
CAMK	0.66±0.039	0.76±0.032	0.25±0.016	0.40±0.034
Other	0.80±0.034	0.81±0.037	0.36±0.029	0.47±0.044
STE	0.73±0.059	0.80±0.063	0.40±0.043	0.57±0.072
CK1	0.79±0.035	0.81±0.028	0.39±0.032	0.41±0.042
Atypical	0.85±0.015	0.89±0.014	0.36±0.005	0.45±0.015
Mouse				
CMGC	0.73±0.011	0.79±0.020	0.38±0.009	0.45±0.035
AGC	0.48±0.033	0.63±0.043	0.20±0.015	0.31±0.056
TK	0.61±0.045	0.78±0.052	0.25±0.020	0.46±0.052
Yeast				
CMGC	0.65±0.032	0.76±0.042	0.22±0.020	0.44±0.050
AGC	0.57±0.043	0.71±0.048	0.26±0.036	0.48±0.048
CAMK	0.64±0.036	0.70±0.020	0.15±0.029	0.33±0.037
Other	0.60±0.036	0.75±0.045	0.21±0.019	0.40±0.033

In general, the accuracy for mouse kinases was more enhanced by the incorporation of sequence when compared to the accuracy for human kinases (Table B.14). We noticed that the accuracy for mouse AGC kinases was no greater than random for context alone, with a low AUC of 0.48. However, after the incorporation of sequence data, the AUC increased to a much higher value of 0.63 (Table 4.4). This is likely due to the size of the mouse protein-interactome, which is much smaller than the human version. The most substantial gains were made for the tyrosine kinases, where the average AUC for the family increase from 0.61 to 0.78 – a near 30% increase in prediction accuracy. There was a similar increase in the AUC50, from 0.25 to 0.46, indicating that the incorporation of the sequence model also made an important contribution at low false-positive levels.

The yeast kinases benefitted even more than the mouse kinases from the incorporation of sequence, with substantial increases to prediction accuracy observed across the four yeast kinase families (Table B.14). Prediction accuracy for yeast AGC and "other" kinases increased in AUC

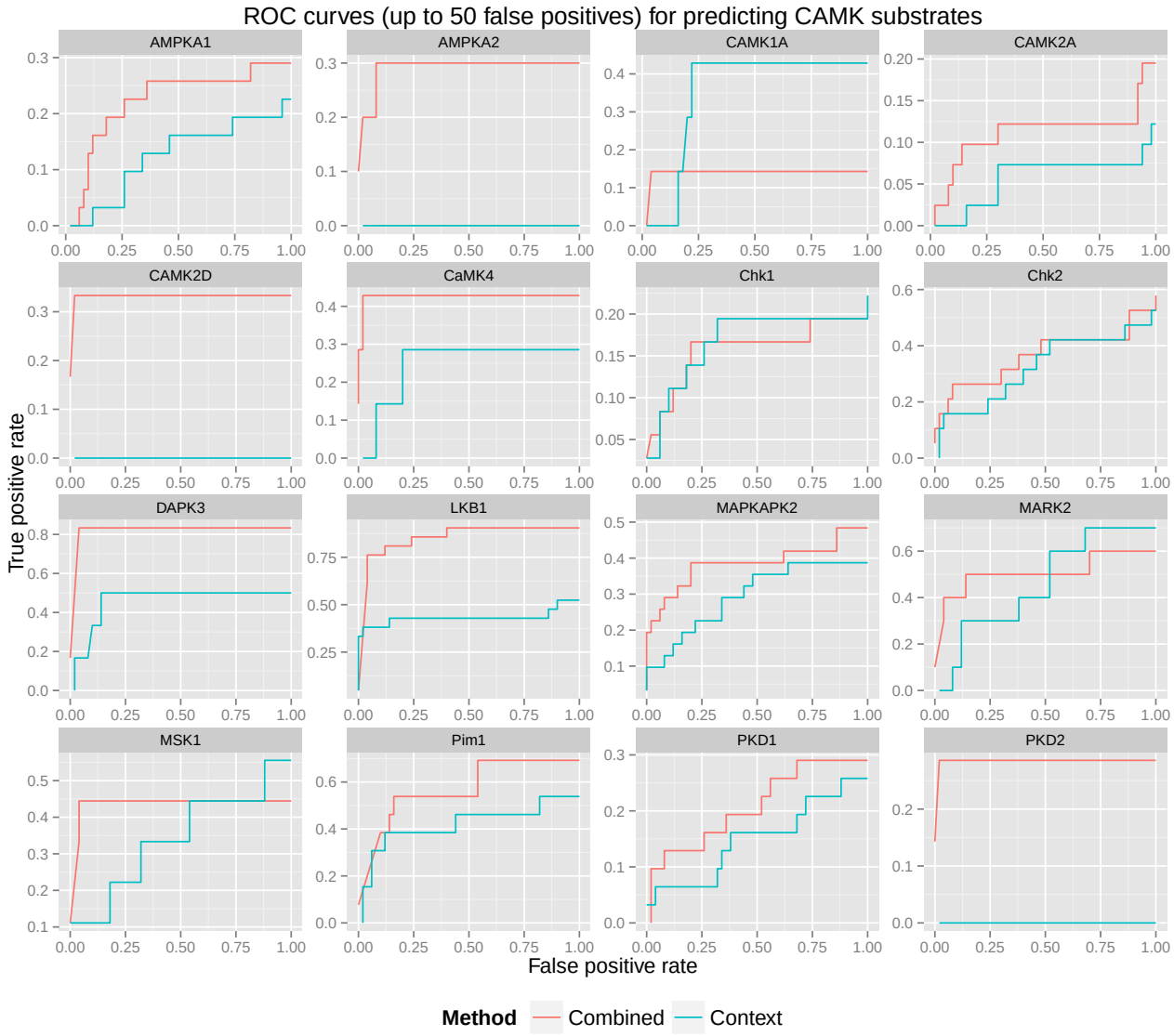

FIGURE 4.4: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the human CAMK family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

value by an average of 0.14 and 0.15 respectively, while CMGC kinases increased by an average of 0.09. We also found that the AUC50 increased by approximately two-fold for each of the four yeast kinase families. The results for mouse and yeast kinases indicate that the model is able to offset the reduced availability of the context information through the sequence data.

4.4.3 Comparisons to alternative methods

We tested the ability of PhosphoPICK (i.e. the full PhosphoPICK workflow described in section "Prediction workflow") to correctly classify the known kinase phosphorylation sites out of all potential sites within our set of phosphorylation substrates. Due to the number of potential phosphorylation sites (~170,000 S/T sites and ~30,000 Y sites), we tested prediction accuracy at more stringent levels of specificity – 99.9% and 99%. We compared the prediction sensitivity of PhosphoPICK with using sequence alone. We found that by combining the substrate score from the combined model with the site score from the sequence model, we were consistently able to improve prediction accuracy when compared to using the sequence model alone (Figure 4.5).

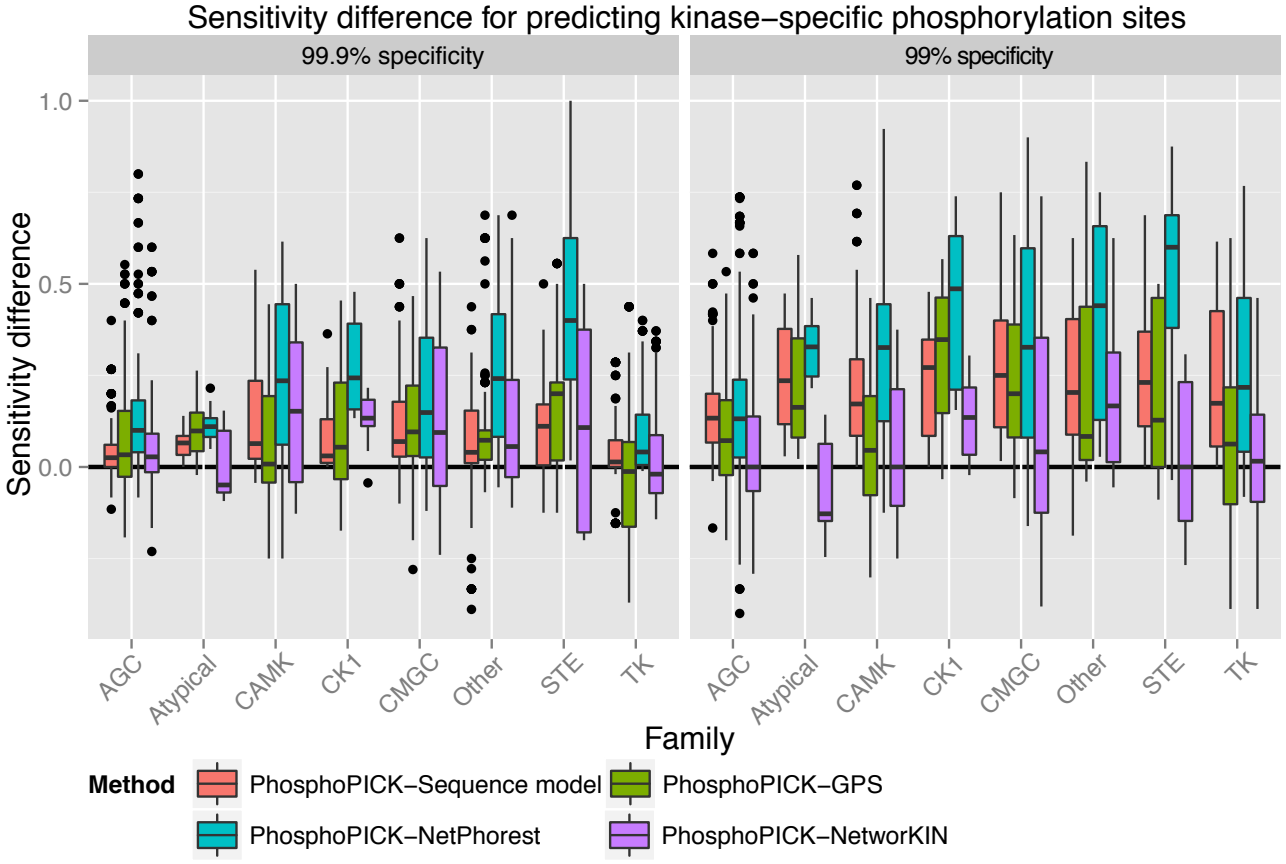

FIGURE 4.5: Sensitivity comparisons for predicting kinase-specific phosphorylation sites out of all potential phosphorylation sites in the protein training set between PhosphoPICK and alternative classification methods. Comparisons were made by performing cross-validation across ten data-set splits for each of the kinases. Sensitivity was calculated for all methods at two levels of specificity: 99.9% and 99%. Comparisons were made between PhosphoPICK and the sequence method alone, and between PhosphoPICK and three alternative predictors: GPS, NetPhorest and NetworKIN.

On average, the use of the combined model offered the greatest level of accuracy increase to kinases from the CMGC family, with an average sensitivity difference of 0.12 at 99.9% specificity and 0.27 at 99% specificity. This is consistent from our previous findings that the use of context offers greater support to phosphorylation site prediction from CMGC kinases. The CAMK kinases gained a similar level of sensitivity at the higher specificity threshold, though there was a smaller average sensitivity difference of 0.22 at the 99% specificity level. The AGC and TK kinases appeared to benefit the least, with a sensitivity difference at 99.9% specificity of 0.045 and 0.042, respectively.

We also compared the ability of PhosphoPICK to predict kinase-specific phosphorylation sites to three alternative methods: GPS 2.1 (80), NetPhorest 2.0 and NetworKIN 3.0 (97). We compared the prediction sensitivity of the different methods at the specificity levels described above. Figure 4.5 shows the sensitivity difference between PhosphoPICK and the compared methods at two levels of specificity: 99.9% and 99%. Tables B.16 and B.17 contain the full set of comparisons for individual kinases at specificity levels 99.9% and 99%, respectively. In addition, Figure 4.6 shows a comparison against the various methods using MCC as the comparison metric. We found that at the stricter level of specificity, PhosphoPICK obtained an increased level of sensitivity over the alternatives for most comparisons made. At the 99.9% specificity level, PhosphoPICK gained an average sensitivity increase of 9% when compared to NetworKIN, 10% compared to GPS and 22% compared to NetPhorest. At the 99% specificity level, PhosphoPICK gained average sensitivity increases of 6%, 18% and 35% when compared against NetworKIN, GPS and NetPhorest, respectively. While PhosphoPICK obtained greater prediction accuracy on average, there were some cases where PhosphoPICK performed worse than the alternatives – for example the tyrosine kinases, where we observed an average sensitivity difference against GPS of -0.014 at the 99.9% specificity level. We also noticed that PhosphoPICK performed worse on the atypical kinases when compared to NetworKIN, with a small difference in sensitivity at 99.9% specificity of -0.004, and a larger difference of -0.076 at 99% specificity.

4.4.4 Evaluation using the hold-out set

PhosphoPICK contains the option to calculate P-values for predictions, representing the likelihood of obtaining a given prediction by chance, given how predictions are distributed over the proteome. To estimate the level of accuracy that is to be expected from using the fully trained model underlying the web-server, we evaluated prediction accuracy using our hold-out set of

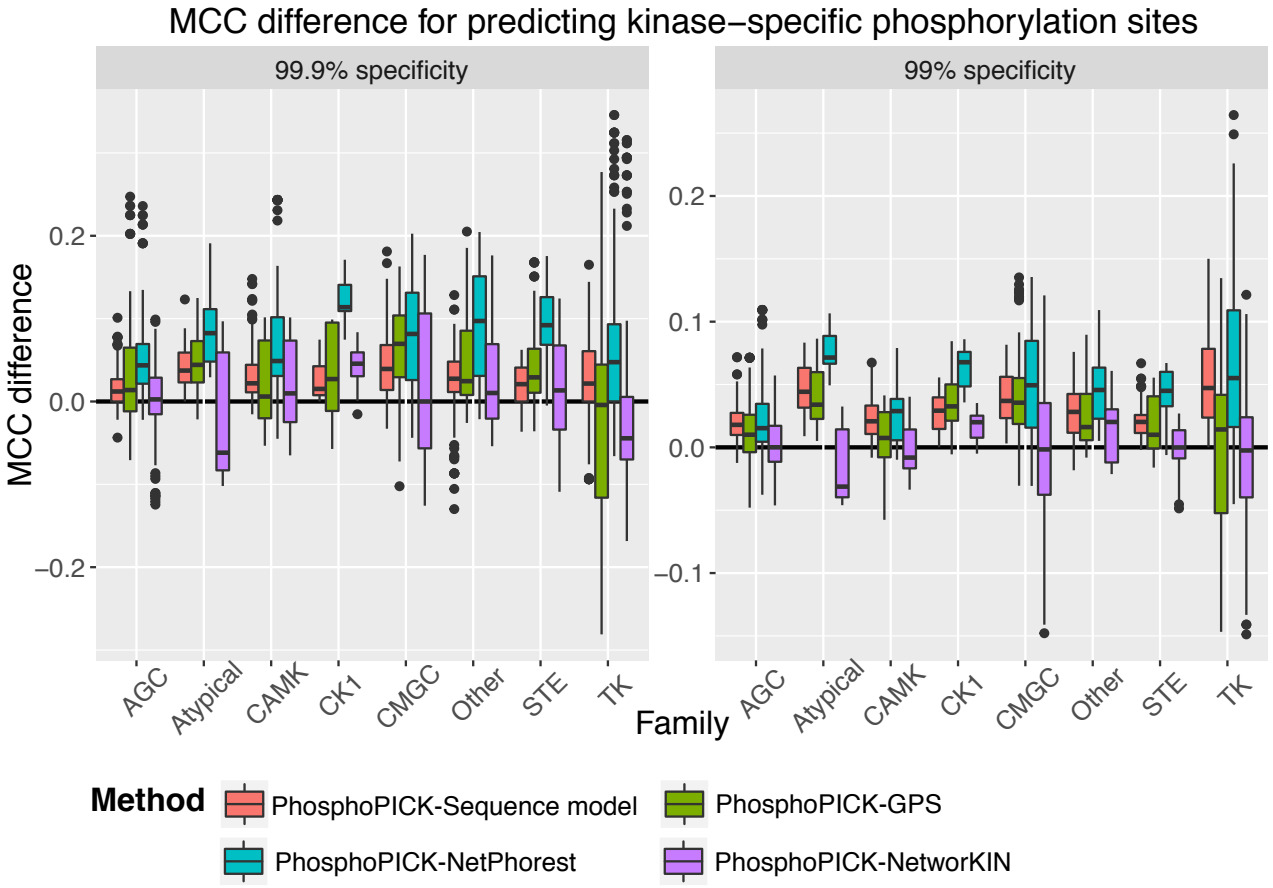

FIGURE 4.6: MCC comparisons for predicting kinase-specific phosphorylation sites out of all potential phosphorylation sites in the protein training set between PhosphoPICK and alternative classification methods. Comparisons were made according to the procedure described for Figure 4.5, with MCC as the comparison metric.

145 substrates (of the kinases listed in Table 4.5) by calculating P-values of the predictions and considering predictions that fell below a P-value threshold of 0.005 (Section 4.3.6).

We found that PhosphoPICK was generally able to maintain a high level of specificity, with an average specificity of 97% across the 9 kinases represented in the hold-out set (Table 4.5). There was a diverse range of sensitivity levels (from 3% for Src to 62% for CK2A1), with an average of 32% – well above what would be expected by chance given the percentage of false-positive predictions. This confident prediction accuracy on completely novel data indicates that PhosphoPICK is a reliable method for uncovering new kinase substrates and kinase-specific phosphorylation sites.

TABLE 4.5: Prediction accuracy on hold-out set for predicting kinase-specific phosphorylation sites (below a P-value threshold of 0.005) as measured by a variety of metrics – sensitivity, specificity, balanced accuracy (BAC) and Matthews' correlation coefficient (MCC). Results were generated by training the model on the full training data set, and evaluating it on the hold-out set. Results represent the ability of PhosphoPICK to correctly predict the known kinase-specific phosphorylation sites out of all potential sites in the set of hold-out substrates. In total there were 14,617 S/T sites and 2,324 Y sites.

Kinase	Positives	Sensitivity	Specificity	BAC	MCC
CDK2	72	0.36	0.96	0.66	0.12
CDK1	39	0.51	0.93	0.72	0.09
ERK2	55	0.22	0.98	0.60	0.08
ERK1	56	0.29	0.98	0.63	0.12
PKACA	53	0.28	0.99	0.64	0.18
PKCA	40	0.15	0.97	0.56	0.04
Akt1	15	0.4	0.98	0.69	0.09
CK2A1	52	0.62	0.95	0.78	0.15
Src	34	0.03	0.99	0.51	0.02

4.4.5 Multiple kinases regulate nuclear localisation

We predicted NLSs using the NucImport predictor (119), a tool for predicting nuclear proteins and the location of their NLSs on the basis of protein interaction and sequence data (NucImport does not explicitly incorporate protein phosphorylation into its predictions). The complete human proteome (including isoforms) was run through NucImport and all proteins that were predicted to contain a type-1 classical NLS were retained – there were 4134 such proteins. The type-1 classical NLS contains an optimal four residue amino acid configuration of KR(K/R)R or K(K/R)RK (166). In order to investigate phosphorylation within a window surrounding the NLS, we defined a centre position, P_0, as the third residue within the predicted NLS (in the literature, this position is usually designated "P4" (162)), and cross-referenced the location of the signals with known phosphorylation sites from PhosphoSitePlus®. We identified 1,830 phosphorylation sites that were within a 20 residue window around P_0. These phosphorylation sites were submitted to PhosphoPICK for analysis (predicting all human kinases), and a P-value threshold of 0.005 was used to return results with a high level of stringency.

In order to test for kinases that were regulating specific positions in relation to the NLS, we counted the number of predicted binding events for kinases at each position within the 20 residue window surrounding P_0. To determine whether the number of predicted kinase binding sites near an NLS was greater than would be expected by chance, we tested for over-representation

against all known phosphorylation sites within the set of predicted nuclear proteins. Over-representation was tested for using Fisher's exact test with Bonferroni correction to obtain E-values (the P-values for the Fisher's exact test were corrected by the total number of tests performed; i.e. the number of kinases multiplied by the number of sites – 2,247).

Figure 4.7 shows the distribution of predicted binding sites for several kinases around the P_0 position of the NLS. We found that there was higher phosphorylation activity downstream from the NLS, where protein kinase A (PKA), aurora kinase B (AurB), and Akt1 in particular were found to have the most significantly over-represented binding locations. At position 3 (P_3), the most significant kinase was PKA (E = $2.03e^{-38}$), which was predicted to be phosphorylating 55/144 of the phosphorylation sites at that position. AurB had a pair of highly significant binding sites at positions 2 (E = $7.32e^{-30}$) and 3 (E = $2.4e^{-21}$).

There were fewer observations of kinases over-represented at phosphorylation sites upstream from the NLS, though we found that cyclic dependent kinase 2 (CDK2) and protein kinase C alpha (PKCa) were significantly over-represented at several upstream positions. At positions -4, -5 -6 and -7, CDK2 was found to have the most significant over-representation of sites compared to any other kinase. CDK2 was predicted to target 28/50 (E = $9.42e^{-13}$) of the phosphorylation sites at position -4, 31/61 (E = $2.1e^{-13}$) at position -5, 27/89 (E = $6.4e^{-10}$) at position -6 and 23/88 (E = $6.0e^{-07}$) at position -7.

To investigate whether the proteins being phosphorylated at these specific sites were involved in similar biological processes, we performed gene ontology (GO) term enrichment analyses. We performed the tests by taking a foreground set of proteins and testing for over-representation (Fisher's exact test, with Bonferroni multiple correction) of terms in the foreground set against a background comprised of our set of phosphorylated nuclear proteins. Significant terms should therefore not simply represent general phosphorylation or nuclear functions, but functions specifically related to the kinase being tested.

We performed GO term enrichment tests on a kinase-specific basis, identifying substrates that were predicted to be phosphorylated within the 20 residue window surrounding P_0. We also tested substrates that were predicted to be phosphorylated at the specific sites that were identified as being over-represented for the kinase being tested (Tables B.18 – B.25). We found that AurB substrates were enriched in the GO terms "chromosome", "nucleosome" and "nucleosome assembly". Interestingly, while the proteins phosphorylated by AurB at the P_3 position were enriched in similar GO terms, the proteins phosphorylated at P_2 returned no significant GO terms. While CDK2 substrates obtained the significant terms "chromosome", "cell cycle",

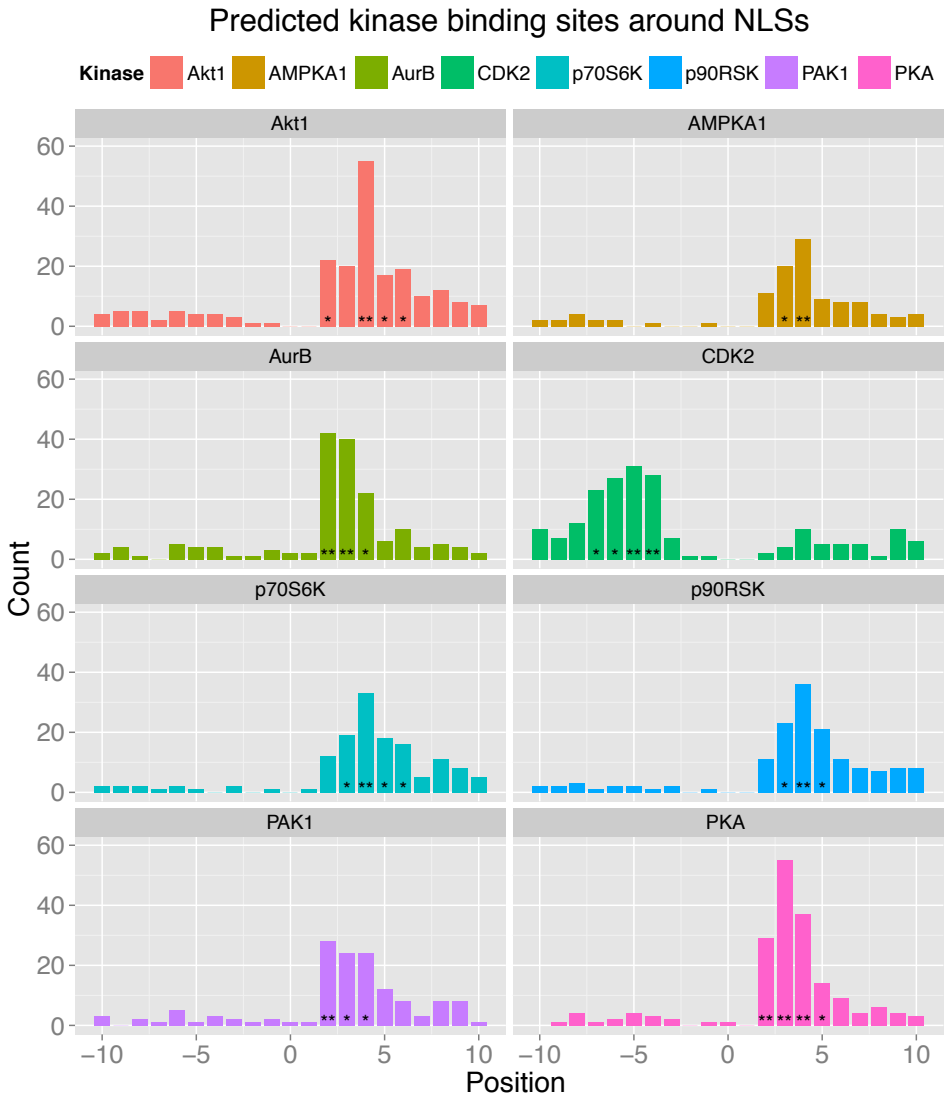

FIGURE 4.7: Distribution of predicted kinase phosphorylation sites surrounding NLSs. The locations of predicted NLSs were cross-referenced with phosphorylation sites from PhosphoSitePlus® and PhosphoPICK was used to assign kinases to the sites. Count represents the number of times a kinase was predicted to phosphorylate a specific site relative to the NLS. Over-representation of a kinase for a particular site was assessed using a Fisher's exact test with a Bonferroni multiple correction. (*) indicates an E-value < 0.05 and (**) an E-value $< 1.0E^{-10}$.

"nucleus" and "DNA repair", none of its significant binding site positions were found to be be associated with enriched GO terms.

We noticed that kinases with an over-representation of binding events at P_4 consistently obtained a number of significant GO terms for substrates phosphorylated at that site. In addition to AurB mentioned above, PKA P_4 substrates had 10 enriched GO terms, Akt1 had 4, AMPKA1

and p70S6K both had 11 and p90RSK had 8. We noticed that there was also some repetition of enriched GO terms among these kinases at P_4 – the term for "fibroblast growth factor receptor (FGFR) signalling pathway" was the most significant P_4 term for each of the AGC kinases (PKA, Akt1, p70S6K and p90RSK), and was the second most significant for AMPKA1 kinase. To determined whether phosphorylation at P_4 in general was associated with specific functions (such as the FGFR signalling pathway) we did a GO term enrichment test with all substrates that were phosphorylated at that position, however no GO terms were found to be significant (Table B.26). This would indicate that the phosphorylation of the site at P_4 does not by itself correspond to a particular function, rather this is dependent on the kinase regulating the site.

4.5 Discussion

The regulation of protein function through kinase-mediated phosphorylation is a complex process involving numerous aspects of cellular behaviour on the systems biology level, and the binding capacity of kinases to substrates on the molecular level. We have presented here a novel method for probabilistically modelling the sequence features that determine kinase binding at a molecular level. We have shown that PhosphoPICK is able to leverage these two diverse types of information and seamlessly integrate them into a model that can identify kinase substrates with high accuracy.

A benefit of the integration of sequence and context data into a single probabilistic model is the ability to take into account interdependance between these heterogeneous sources of information; i.e. the likelihood of seeing certain amino acids or k-mers in a protein may change depending on the context information, and similarly, the expectation of certain protein interactions can be influenced by the protein sequence. Indeed, we have found that the combined model can be used to query expected kinase binding sequence motifs and generate corresponding sequence logos (47) based on context information presented to the model (see Section B.1 for an example).

A counter-intuitive result seen as a part of the integration of sequence and context was that the performance seen in the sequence was not necessarily reflected in the combined model. The tyrosine kinases were a particularly interesting example; we found that while the tyrosine sequence models (for both human and mouse) were the least accurate amongst the sequence models, the mouse combined model benefited greatly from the incorporation of sequence, with

a near two-fold increase seen in the AUC50. This is an indication that while the two individual systems – sequence and context – of predicting kinase binding events may be limited by themselves, the integration of the two can result in a much more powerful predictive model.

It was interesting to note that though the sequence model obtained the greatest accuracy (for phosphorylation site prediction) on the human kinases, the yeast kinases in general saw the highest increases in prediction accuracy (particularly as measured by AUC50) when the sequence model was incorporated into the context model. While the availability of context data (e.g. cell cycle data) is likely a factor in the observed differences in prediction performance between organisms, a uni-cellular organism like yeast would be expected to require less sophistication in the regulation of kinase activity than higher organisms. Consequently, the use of context factors is no doubt more important for understanding kinase targets in higher organisms.

For more complex organisms such as human and mouse, an additional realm of biology to consider in relation to phosphorylation and kinase activity is tissue and cell-type specificity. Protein phosphorylation has the potential to change substantially depending on the cell type, and the biological processes that kinases regulate can also vary depending on cell or tissue type. While there is limited amounts of consolidated tissue-specific phosphorylation data, there is growing amounts of tissue-specific protein expression data (167). In addition to protein expression data, the FANTOM consortium has profiled vast cell-type specific gene expression atlases (168). Such data resources could make it possible to infer more probable candidate kinases based on which ones are available in the tissue or cell type of interest. While outside the scope of the current study, this would certainly make for an interesting avenue of exploration in future work.

A system-wide analysis of biological mechanisms has the potential to reveal functional trends that may not otherwise be apparent. Our analysis of the overlap of NLSs and phosphorylation events has shown that there are several kinases that may be implicated in the regulation of nuclear localisation through the phosphorylation of specific sites close to the NLS. Phosphorylation is a well-documented mechanism of nuclear localisation (154, 163, 164, 169–172). Because classical NLSs are positively charged, introduction of a negatively charged phosphate group in the vicinity of the NLS would in general be expected to inhibit nuclear import, as previously demonstrated for CDK1-mediated phosphorylation at positions "P0" and "P-1" (164) (interestingly, these sites correspond to our P_{-4} and P_{-5} positions, which saw the most significant over-representation of CDK2 binding sites.). However, the effect will depend on the specific position that is phosphorylated, and in some positions phosphorylation can stimulate nuclear import (154, 163, 169, 170, 172, 173).

Several of the kinases identified in our study have previously been implicated in nuclear import. For example, the import of sex-determining factor SOX9 is regulated by PKA, whereby the phosphorylation of two phosphorylation sites (one next to the NLS) enhances SOX9 binding to importin β (174). Adenomatus polyposis coli (APC) is another example of a protein where nuclear import is regulated by phosphorylation (175). In this case, APC contains two identified NLSs and a putative PKA-mediated phosphorylation site is positioned immediately after the second NLS, which leads to a reduction in APC nuclear localisation when the site is active. As a key regulator during mitosis, AurB is involved in several processes such as mitotic chromosome condensation (176), and it has also been shown to phosphorylate residues within the vicinity of NLSs (177). The Akt kinase has been shown to be a regulator of nuclear localisation (178), and phosphorylation by Akt is able to impair the nuclear import of p27 *in vitro* (179). Similarly, CDK2 is known to be a regulator of nuclear localisation (180). While these studies confirm that these kinases are involved in nuclear localisation, our results shed light on specific mechanisms whereby nuclear localisation is controlled by the phosphorylation of key residues close to the NLS.

4.6 Availability

PhosphoPICK is freely available online as a web-server, and can be used in two ways. A user can upload protein sequences, and select any number of kinases to obtain predictions for potential phosphorylation sites on the proteins. Significance of predictions can be gauged through the calculation of empirical P-values, and only results below a chosen level of significance returned. Visualisation of results is also available through a "Protein Viewer" page based on the BioJS (181) package pViz (182). Secondly, the web-server allows for the construction of downloadable proteome-wide sets of kinase-substrate predictions for any of the kinases and species described in this paper. A more detailed description of the web-server workflow is available in Section B.2.

Chapter 5

PhosphoPICK-SNP: Quantifying the effect of nsSNPs on protein phosphorylation[1]

5.1 Abstract

Genome-wide association studies are identifying single nucleotide polymorphisms (SNPs) linked to various diseases, however the functional effect caused by these variants is often unknown. One potential functional effect, the loss or gain of protein phosphorylation sites, can be induced through variations in key amino acids that disrupt or introduce valid kinase binding patterns. Current methods for predicting the effect of SNPs on phosphorylation operate on the sequence content of reference and variant proteins. However, consideration of the amino acid sequence alone is insufficient for predicting phosphorylation change, as context factors determine kinase-substrate selection.

We present here a method for quantifying the effect of SNPs on protein phosphorylation through an integrated system of motif analysis and context-based assessment of kinase targets. By predicting the effect that known variants across the proteome have on phosphorylation, we are able to use this background of proteome-wide variant effects to quantify the significance of novel variants for modifying phosphorylation. We validate our method on a manually curated set of phosphorylation change-causing variants from the primary literature, showing that the method predicts known examples of phosphorylation change at high levels of specificity.

[1]Chapter reproduced from paper of the same name currently pending submission.

5.2 Introduction

The identification of genetic variants linked to disease is transforming the biomedical research landscape. Genome wide association studies (GWAS) have been identifying numerous single nucleotide polymorphisms (SNPs) over-represented in patients with in a wide variety of diseases including cancer. While many SNPs are being discovered, the precise effect that they have on resultant RNA or protein products is generally not known. One of the potential effects of non-synoymous SNPs (nsSNPs) on protein function is the disruption of post-translational modifications (183). As phosphorylation is the most ubiquitous modification, the potential for phosphorylation sites to be affected by amino acid variants is high. For example, the PhosphoSitePlus® database (184) has identified numerous sequence variants that fall within the immediate vicinity of a phosphorylation site, and the recent analysis of cancer driver mutations has implicated phosphorylation as being a major factor in understanding the disruption of signalling pathways caused by amino acid variations (185).

There have been numerous examples of disease-associated naturally occurring variants that impact the phosphorylation status of proteins. The majority of such examples have involved a variant disrupting a phosphorylation site in the reference protein, though there have been at least two examples of missense mutations found to introduce phosphorylation sites (186, 187). While there have been relatively few studies experimentally determining the effect of naturally occurring variants on phosphorylation, there are tens of thousands of nsSNPs that have the potential to impact phosphorylation. The PhosphoSitePlus® PTMVar dataset (184) , which is comprised of missense mutations cross-referenced to post-translational modifications, contains over 19,000 examples of variants falling within a 15-residue window surrounding a known phosphorylation site. Such variants have the potential to disrupt existing phosphorylation sites, but there will be many additional variants with the potential to introduce new phosphorylation sites. In addition to PhosphoSitePlus®, the PTM-SNP database collates SNPs that occur in the vicinity of a number of post-translational modifications, including phosphorylation (183).

There have also been databases developed that catalogue the predicted effect of SNPs on potential phosphorylation sites. Ryu and colleagues defined the term “phosphovariant” to refer to a mutation that impacts the phosphorylation status of an amino acid (92). To predict examples of phosphovariants, they developed PredPhospho, a support vector machine model that predicts kinase-specific phosphorylation sites based on the amino acid motifs surrounding potential phosphorylation sites. Applying PredPhospho to missense mutations obtained from

Swiss-Prot, they predicted examples of phosphovariants and incorporated them into the PhosphoVariant database (92). The PhosSNP database is another example of cataloging variants predicted to modify protein phosphorylation (188). Ren and colleagues employed the GPS 2.0 software, a kinase-specific phosphorylation site predictor that uses optimised substitution matrices (79). The GPS 2.0 predictor was applied to variants from the dbSNP database (189), with the ones predicted to cause a change in phosphorylation status or to cause a change in the kinase targeting the phosphorylation site, were compiled into the PhosSNP database.

Most recently, the MIMP (mutation impact on phosphorylation) method has been developed, which uses position weight matrices and Gaussian mixture models to score the probability that a variant will cause loss or gain of phosphorylation (84). In contrast to the other methods, MIMP provides a prediction service rather than a database. For the purpose of consistency with the most recently published work, we will consider two classes of "phosphovariants": phosphorylation-loss causing variants and phosphorylation-gain causing variants.

The current methods for predicting the effect of nsSNPs on phosphorylation, described above, operate on the sequence content surrounding a potential phosphorylation site. While methods based on linear motifs can predict the potential for a kinase binding site to be disrupted (190), the presence of a valid kinase-substrate binding motif on a protein is no guarantee that a kinase will come into contact with the protein (27). We have previously developed a method, PhosphoPICK, for predicting kinase substrates using protein-protein interaction networks and protein abundance across the cell cycle. The use of such context information can improve the prediction accuracy of kinase-specific phosphorylation site prediction from sequence by over two-fold at low false-positive levels (161). An approach that integrates cellular context information with sequence information should therefore be able to provide a more accurate assessment of the effect of SNPs on phosphorylation than methods that operate on sequence alone.

Building on the properties of PhosphoPICK, we present here a method for quantifying the effect of nsSNPs on protein phosphorylation status. Taking stock of known missense mutations across the proteome, as collected in UniProt, we use PhosphoPICK to build kinase-specific, proteome-wide sets of predicted variant effects on phosphorylation. These sets provide a "background distribution" that can be used to calculate a measure of significance for the predicted effect that a novel variant has on phosphorylation loss or gain.

In order to validate our approach, we searched the literature for naturally occurring variants causing phosphorylation loss or gain, identifying 19 such variants. By comparing the threshold at which our method detects true positives against that of the background, we demonstrate that our method is able to detect over 50% of the known phosphovariants within the first 2% of the

background distribution. This demonstrates the method's reliability in detecting true examples of differential phosphorylation from over one million potential phosphovariants. Applying the method to variants in the vicinity of phosphorylation sites from the PhosphoSitePlus® PTM-Var dataset (184), we find that the predicted phosphovariants are over-represented among the ones with disease annotations. These results support the conclusion that our method, named PhosphoPICK-SNP, is able to detect variants that have functional significance.

5.3 Methods

5.3.1 Data resources

Missense mutation data

We obtained the UniProt index of protein altering variants (191), which maps dbSNP variants (189) to proteins within the UniProt database (downloaded March, 2015). This file contained 752,857 variants mapped to amino acid variants in UniProt proteins. The variants covered 89,909 protein sequences in the UniProt database.

Phosphorylation sites affected by naturally occurring variants

Through a manual search of the literature, we compiled a list of naturally occurring variants that were found experimentally to either disrupt or introduce a phosphorylation site. For the purpose of this work we included variants that were shown either *in vivo* or *in vitro* to affect the phosphorylation of a specific site; although there are examples of studies showing changing phosphorylation levels on the protein, we only recorded examples where the precise phosphorylation site was known. Table 5.1 contains the list of identified genes, with variant and phosphorylation site affected. We found 17 examples of phosphorylation loss and 2 examples of phosphorylation gain in response to nsSNPs. Of the 17 loss-causing variants, 6 of the mutations are on the phosphorylation site.

5.3.2 Building distributions of variant effects

We built distributions of predicted variant effects on phosphorylation in a kinase-specific basis across all protein altering variants. PhosphoPICK employs two Bayesian network models to

TABLE 5.1: Naturally occurring variants that have been shown through *in vivo* or *in vitro* experiments to affect a phosphorylation site either adjacent to, or at the site of, the variant. The effect can be to disrupt an existing phosphorylation site (loss), or introduce a new one (gain).

Gene	Variant	Phos. site	Effect	Reference
Cyclin D1	T286R	T286	loss	(192)
hOG1	S326C	S326	loss	(193)
p53	P47S	S46	loss	(194)
BDNF	V66M	T62	loss	(195)
CDKN1A	D149G	S146	loss	(196)
hERG1	K897T	T897	gain	(186)
PPARγ2	P113Q	S112	loss	(197)
PTP-1B	P387L	S386	loss	(198)
UBE3A	T485A	T485	loss	(199)
PER2	S662G	S662	loss	(200)
MeCP2	R306C	T308	loss	(201)
NKX3-1	R52C	S48	loss	(202)
PLN	R14C	S16	loss	(203)
ABCB4	T34M	T34	loss	(204)
MAF	P59H	T58	loss	(205)
GLUT1	R223W	S226	loss	(206)
AR	R405S	S405	gain	(187)
Gab1	T387N	T387	loss	(207)
STAT1	L706S	Y701	loss	(208)

make predictions. The first model classifies kinase-substrate binding sites from sequence, and incorporates position-specific amino acid frequencies and counts of co-occurring neighbouring amino acids within some m length window surrounding a potential phosphorylation site (paper in submission). This model is henceforth referred to as the *sequence model*. Separately, a Bayesian network model integrates the sequence model with protein-protein interaction and association data sourced from BioGRID (209) and STRING (148), as well as protein abundance data across the cell cycle (28), in order to calculate the probability that a kinase ordinarily targets a given protein. This model is henceforth referred to as the *combined model*. When scoring the effect of a variant we use PhosphoPICK to generate three scores: (1) $R_{substrate}$, the prior probability based on the combined model that the kinase would be expected to target the reference protein, (2) R_{site} the probability according to the sequence model that the kinase will phosphorylate the site of interest on the reference protein, and (3) V_{site} the probability that the kinase will target the site of interest on the variant protein.

Kinases within PhosphoPICK contain different optimal binding site windows that are considered

when making a prediction for a potential phosphorylation site. Therefore, given a query kinase, we checked for variants that fell within a window surrounding a potential phosphorylation site. For each potential phosphorylation site, we recorded a reference peptide and a variant peptide containing the missense mutation. We then used the sequence model to obtain the R_{site} and V_{site} scores from the reference and variant peptides respectively. If the central residue for a peptide is not a valid phosphorylation site (for example a threonine is mutated to an arginine) it will be scored 0. We defined a score difference,

$$D_{site} = V_{site} - R_{site} \tag{5.1}$$

where a negative value of D_{site} indicates the variant is predicted to cause *decreased* probability of phosphorylation, and a positive value represents an *increased* probability of phosphorylation.

We calculated distributions of D_{site} values in a kinase-specific manner across all potential phosphorylation sites that contained a missense mutation within the window for the query kinase. A potential phosphorylation site is defined as any serine (S) or threonine (T) residue for S/T kinases, any tyrosine (Y) residue for Y kinases, or any S/T/Y residue for dual specificity kinases.

5.3.3 Calculating variant significance

The significance of the effect on phosphorylation by a variant is calculated in a kinase-specific manner, as described by the following procedure. Given some kinase K, an m length window corresponding to K is centred on potential phosphorylation sites within the protein sequence, where if the variant falls within a window, m length reference and variant peptides are retained. D_{site} is then calculated from the reference and variant peptides using Equation 5.1. The difference is then compared to the background distribution and a P-value from both tails of the distribution is calculated – representing whether the difference is greater (increased probability of phosphorylation) or less (decreased probability of phosphorylation) than would be expected by chance. The P-values are calculated such that

$$P_{loss} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(D_i \leq D_{site}) \tag{5.2}$$

$$P_{gain} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(D_i \geq D_{site}) \tag{5.3}$$

where n is the number of variants contained in the background distribution for kinase K and $\mathbf{I}(\cdot)$ is the indicator function. The final P-value representing the site, P_{site}, is calculated as the minimum of P_{loss} and P_{gain}.

The $R_{substrate}$ context score for the query protein is then retrieved. As for the D_{site} scores, we have distributions of context scores across the proteome for each kinase. We therefore calculate an empirical P-value for the $R_{substrate}$ score, $P_{substrate}$, based on a count of the proteome-wide context scores that are greater than or equal to $R_{substrate}$, using the same form as Equation 5.3. We then use Fisher's method to combine the two P-values into a combined P-value that represents the confidence of the variant effect size given both the difference in sequence scores and likelihood that the reference protein would ordinarily be a substrate of the query kinase. Given the P-values P_{site} and $P_{substrate}$, we calculate:

$$X = -2(ln(P_{substrate}) + ln(P_{site})) \tag{5.4}$$

where X follows a Chi squared distribution with 4 degrees of freedom. The combined P-value, $P_{combined}$, can then be derived from X. As a single phosphovariant can be scored with all kinases available to PhosphoPICK (currently numbering 107), we correct the P-value for multiple testing using a Bonferroni multiple correction on P_{site} and $P_{combined}$ to obtain E_{site} and $E_{combined}$.

5.3.4 Evaluating method accuracy on known variants

In order to calculate an estimate of the number of potential phosphorylation sites that were affected by the presence of a nearby variation, we used a 10-fold cross-validation approach to build a set of predicted background values. The proteins within the background set were split into 10 partitions, where 9 of the partitions were used to construct distributions for both the context scores and the D_{site} values. These distributions were then used to evaluate and obtain E-values for the variants in the remaining partition. For each variant the lowest E-value was retained as representing the greatest liklihood that the mutation resulted in a change in phosphorylation status.

To evaluate our method on its ability to detect the examples of differential phosphorylation recorded in Table 5.1, we evaluated the known variants on our method using each of the 10 partitions from the cross-validation test to construct the background distributions. For each variant we calculated the median of the E-values generated across the cross-validation runs;

similar to the background, the final E-value assigned to a variant was the minimum of the E-values for all potential kinases. As there is no obvious way to define a true negative set, we compared to the background set the E-value thresholds at which the true positives were identified; i.e. at each E-value threshold calculated for a true positive, we calculated the number and percentage of variants in the background set that were also identified at that threshold. We performed this test using both E_{site} and $E_{combined}$ values to understand the influence of context on predicting phosphorylation change.

We compared our method's ability to detect the known variants against that of the MIMP predictor (84). We downloaded the local version of the software, and ran the background set of protein sequences and variants through it, specifying probability and log thresholds of 0 to enable a comparison over all thresholds.

5.4 Results

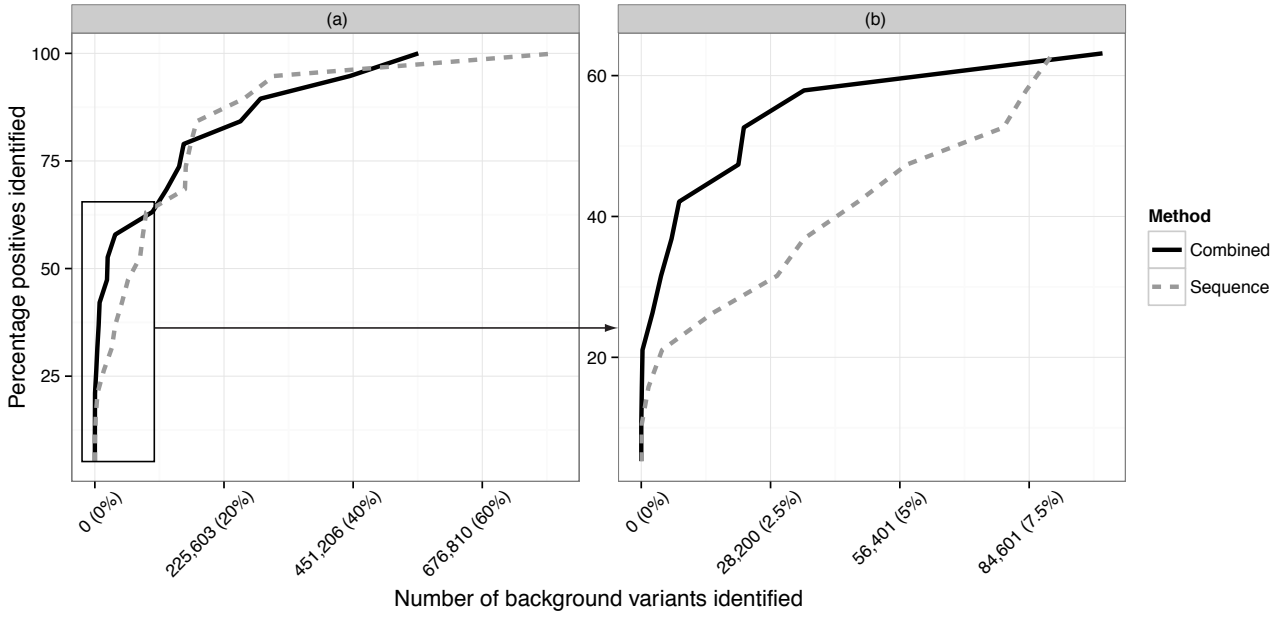

FIGURE 5.1: Line-curves showing the tradeoff between the percentage of positive differential phosphorylation examples identified and the number of variants considered (as the E-value cut-off decreases). Comparison is made between predicting phosphorylation change using sequence alone, and combining sequence with context. Shown is the tradeoff until all positive examples are detected (a), as well as the tradeoff up until 10% of the background variants are detected (b).

The experimentally determined examples of differential phosphorylation listed in Table 5.1 were used to gauge how well our method performed in identifying real examples of phosphorylation

gain and loss. Figure 5.1 shows a tradeoff between the percentage of known positives detected and the background at each E-value threshold a positive was discovered at. When predicting phosphovariants using the combined E-value, we found that the majority (over 50%) of the known positives could be identified within the first 2% of the background distribution. We were able to identify 79% of the experimental examples at an E-value threshold corresponding to 14% of the background. These results demonstrate that the method can identify true positive examples of phosphovariants at high levels of specificity, which represent candidates of real interest to biologists.

We also evaluated the use of sequence only for predicting phosphovariants (i.e. using the E_{site} value), in order to determine if the incorporation of context information was providing an increase in prediction accuracy. When using sequence alone, the majority of variants were not detected until 7% of the background distribution was reached (Figure 5.1). Given the combined method detected the majority of variants at 2% of the background, this represents a 3-fold increase when using the combined E-value. As can be seen from Figure 5.1(a), at the more liberal E-value thresholds there was less difference between sequence alone and the combined E-values. However, these results show that the approach of combining context and sequence information provides the greatest benefit for identifying true variants at higher levels of specificity.

5.4.1 Estimating phosphorylation sites affected by SNPs

In order to investigate the effect of context on predicting differential phosphorylation, we used the methods for calculating E_{site} and $E_{combined}$ to estimate the number of putative phosphorylation sites affected by the nsSNPs contained in the UniProt index of protein altering variants. We performed two tests: firstly, we identified predicted differentially-phosphorylated sites on the basis of E_{site}, where if E_{site} fell below 0.05 the variant was considered to cause differential phosphorylation; i.e. a phosphovariant. In the second test, the $E_{combined}$ value was applied as a filter, where only variants with $E_{combined}$ and E_{site} falling below 0.05 were classified as a phosphovariant.

Based on our cross-validated analysis of the background distribution, we identified the variants that were predicted to be causing differential phosphorylation. In total we found 65,203 variants that were predicted, based on their E_{site} value, to cause differential phosphorylation. When requiring that a variant obtain an E-value < 0.05 for both E_{site} and $E_{combined}$, the number dropped to 41,075. Figure 5.2 shows a histogram of the $E_{combined}$ values calculated for all the

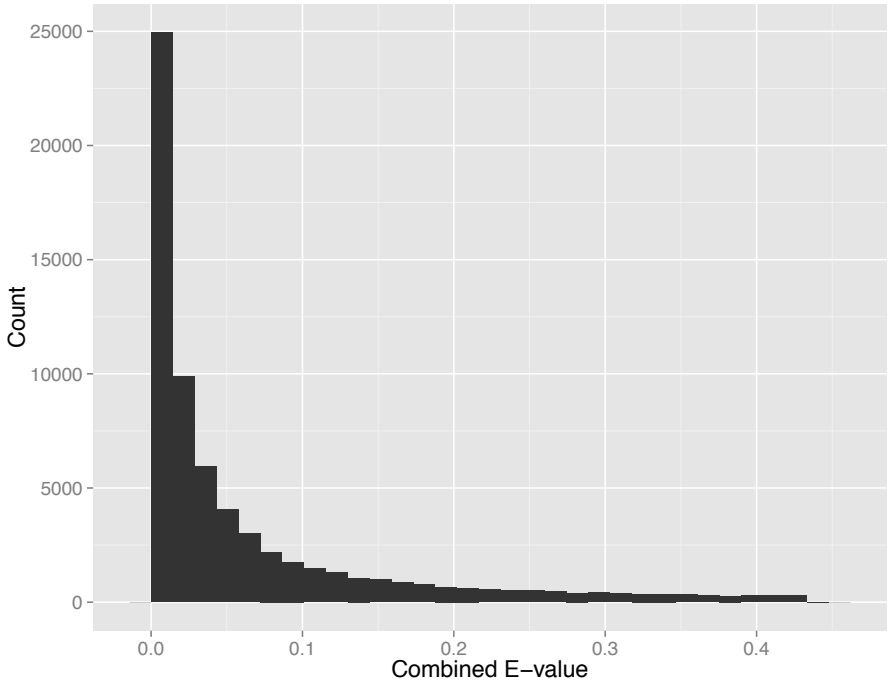

FIGURE 5.2: Histogram showing combined E-value ($E_{combined}$) scores for all variants considered to be significantly likely (E-value < 0.05) to result in differential phosphorylation based on sequence alone.

variants that were found to be significant based on E_{site} alone. While the majority maintain a high level of significance when context is included, nearly 40% of the variants obtained an E-value > 0.05 after context is included. These results illustrate the effect that context has in filtering out spurious examples of phosphovariants where the kinase is unlikely to target the query protein.

5.4.2 Comparison with alternative method

We compared the ability of the MIMP method (84) to predict the set of positives out of the background to our combined method. As MIMP was unable to make predictions for the two phosphorylation gain sites (due to the centre residue of the reference protein being non-phosphorylatable) we performed the comparison using the 17 phosphorylation loss-causing variants. As can be seen from Figure 5.3, at stricter cut-off thresholds our method is able to detect greater numbers of the true positive examples. Within 2% of the background distribution our method is able to detect 47% of the 17 positives, however MIMP does not reach 47% until 3.6% of the background – this corresponds approximately to an additional 17,000 variants.

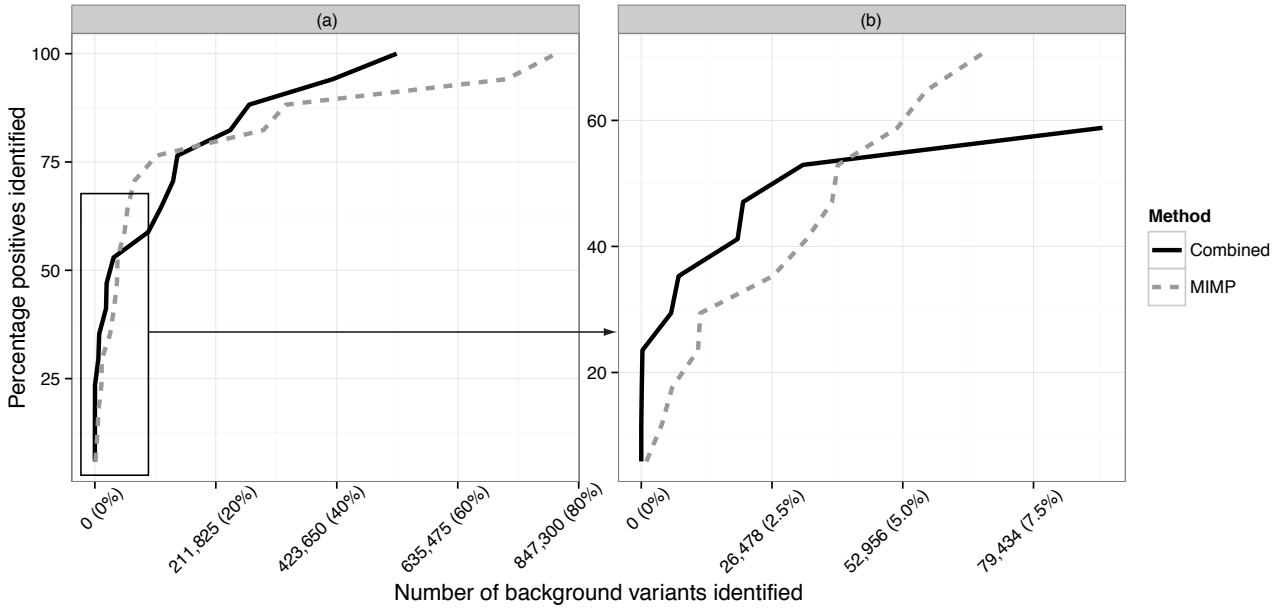

FIGURE 5.3: Line-curves showing a comparison of detecting experimentally confirmed phosphovariants between the combined PhosphoPICK-SNP method and MIMP (84). Shown is the tradeoff until all positive examples are detected (a), as well as the tradeoff up until 10% of the background variants are detected (b).

5.4.3 Phosphorylation loss in disease

We used our method to determine whether the variants that were most confidently predicted to result in a change in phosphorylation status were over-represented among disease-associated variants. We used the PTMVar database from PhosphoSitePlus®, which cross-references post-translational modification information from PhosphoSitePlus® with variant information from the UniProt human variation database. The PTMVar database annotates sites with the classification 'Disease', 'Polymorphism' or 'Unclassified'. Variants that were within the vicinity of phosphorylation sites and were annotated with either 'Disease' or 'Polymorphism' were selected. We then counted the number of times that a variant in each of these classes was predicted to be differentially phosphorylated with a decreased probability of phosphorylation (i.e. it obtained an $E_{combined}$ value < 0.05 in a test for decreased probability), and counted the number of times the variants in both classes were not predicted to be differentially down-phosphorylated.

Fisher's exact test was used to determine the over-representation. We found that variants annotated as disease-associated were significantly over-represented among the variants predicted to result in down-phosphorylation, with a P-value of 0.0002. This indicates that while the presence of a variant in the vicinity of a phosphorylation does not necessarily result in phosphorylation

disruption, our method is able to detect the disease-associated variants that will have a strong impact on phosphorylation.

5.4.4 Prediction of phosphorylation disruption in disease-associated sites

Given that our methods is reliably able to detect phosphorylation loss events, we used it to identify the most likely examples of phosphorylation loss in the PhosphoSitePlus® PTMVar database that were associated with at least one of five cancer types: ovarian, breast, colorectal, liver and pancreatic. These variants were run through our method, and variants that obtained E-values below 0.05 for both E_{site} and $E_{combined}$ were retained. Table 5.2 lists top scoring variants with their disease associations, where the variant has been mapped to the vicinity of a phosphorylation site. The full list of variants is available in Table C.1. In total, we found 52 examples of predicted phosphorylation loss caused by variants related to ovarian cancer, 12 for breast cancer, 8 for colorectal cancer, 19 for liver cancer and 9 for pancreatic cancer.

TABLE 5.2: Cancer-associated variants predicted to cause loss of phosphorylation. Variants are listed according to the cancer or disease they are associated with. Each row contains protein name as UniProt accession, the location of the variant and phosphorylation site, the kinase predicted to target the site, the reference and variant scores for the peptide.

	Protein	Variant	Phos	Kinase	$R_{subst.}$	R_{site}	V_{site}	$E_{combined}$	Peptide
Ovarian	P35222	G555A	T551	Akt2	1.0	1.0	4.95E-05	9.36E-09	QDTQRR**T**$_p$SMG[G/A]TQ
	P26010	Y753H	Y753	FAK	1.0	1.0	0.0	2.23E-08	YRLSVEI[**Y**$_p$/H]DRREYSR
	Q7KZI7	S197N	S197	NEK6	1.0	1.0	0.0	2.3E-07	KIADFGF[**S**$_p$/N]NEFTFGN
	P51813	S212R	S212	GSK3B	0.998	1	0	1.33e-07	PPSSST[**S**$_p$/R]LAQYDS
	P46939	M1256R	T1259	MARK2	1	0.914	0.0005	4.47e-05	R[M/R]KS**T**$_p$EVLP
Breast	P14859	S88F	S88	DNAPK	1.0	1.0	0.0	8.6E-06	SQQPSQP[**S**$_p$/F]QQPSVQA
	P43355	K278T	Y276	Brk	0.998	1	0.000491	0.000103	RALAETS**Y**$_p$V[K/T]VLEYV
	P03372	H6Y	T2	VRK1	0.0492	0.0792	0.00421	0.00111	M**T**$_p$MTL[H/Y]TKA
	Q99490	D816Y	S818	P38B	0.0186	0.587	0.000173	0.00126	CTPSG[D/Y]L**S**$_p$PLSREPP
	P54646	S523G	S527	p90RSK	0.36	0.78	0.00501	0.00214	LTG[S/G]TLS**S**$_p$VSPRLGS
Colorectal	P04637	E271K	S269	CAMK2A	1	0.781	0.011	0.000131	NLLGRN**S**$_p$F[E/K]VRVC
	Q9P253	A913S	S912	ERK5	0.495	0.848	0.0394	0.00475	APPPAKG**S**$_p$[A/S]RAKEAE
	Q9NPD5	I292M	S293	CaMK4	0.832	0.524	3.21e-08	0.00785	ERK[I/M]**S**$_p$LSLH
	Q6ZMN7	G784R	S783	CaMK4	0.792	0.454	1.52e-06	0.00954	TQSS**S**$_p$[G/R]QSS
	Q92953	V450I	S448	ROCK1	0.326	0.845	0.00479	0.012	RAKRNG**S**$_p$I[V/I]SMNL
Liver	P35222	T41A	T41	GSK3A	1	1	0	2.28e-09	GIHSGAT[**T**$_p$/A]TAPSLSG
	P35222	S37F	S37	GSK3A	1	1	0	2.28e-09	YLDSGIH[**S**$_p$/F]GATTTAP
	P35222	T41A	T41	IKKA	1	1	0	7.97e-09	GIHSGAT[**T**$_p$/A]TAPSLSG
	P35222	S37F	S37	IKKA	1	1	0	7.97e-09	YLDSGIH[**S**$_p$/F]GATTTAP
	P35222	T41A	T41	GSK3B	1	0.997	0	8.13e-06	IHSGAT[**T**$_p$/A]TAPSLS

Continued on next page

	Protein	Variant	Phos	Kinase	$R_{subst.}$	R_{site}	V_{site}	$E_{combined}$	Peptide
									Continued from previous page
	Q9BYV9	T519I	T519	p70S6K	1	1	0	1.14e-08	LETRTR[$\mathbf{T}_p$/I]SSSCSS
Pancreatic	P04637	E271K	S269	CAMK2A	1	0.781	0.011	0.000131	NLLGRN$\mathbf{S}_p$F[E/K]VRVC
	Q9BYV9	T519I	S525	p70S6K	1	0.84	0.0917	0.000294	[T/I]SSSCS$\mathbf{S}_p$YSYAED
	P56715	A135V	S137	MARK2	0.848	0.987	2.83e-06	0.000418	IS[A/V]H$\mathbf{S}_p$PPHP
	P05129	P524R	Y521	Brk	0.356	0.227	0.000117	0.000629	TFCGTPD**Y**IA[P/R]EIIA

We found several examples of predicted phosphorylation loss on the β-catenin protein (Uniprot accession number P35222), which was a top candidate for phosphorylation loss for both ovarian and liver cancer. The T41 phosphorylation site, which has been previously identified as a GSK3B target (210), is a known site mutated in cancers (211). It is predicted by PhosphopICK-SNP that the T41A mutation would abolish a GSK3B phosphorylation site at T41 (Table 5.2)

There was also an example of predicted phosphorylation loss on tumour suppressor protein p53 (Uniprot accession number P04637), which has been shown previously to lose phosphorylation as a consequence of a P47S mutation (194). The E271K variant, which was associated with both pancreatic and colorectal cancer (212), was found to have a significant likelihood of disrupting the phosphorylation site at S269. The phosphorylation site at S269 is known to be an important regulator of p53 transcriptional activity (213).

5.5 Discussion

With increasing numbers of disease-associated variants being catalogued, the need for reliable functional annotations is only going to continue to grow. While there are many potential functional effects of gene-coding variants on protein function, such as the perturbation of protein structure or the disruption of one of the many post-translational modifications that proteins undergo, phosphorylation is a high-probability target of disruption due to the ubiquitous nature of this protein modification process. We have presented here a method for quantifying the expected effect of nsSNPs on protein phosphorylation, and have demonstrated that it detects experimentally confirmed examples of phosphovariants at high levels of specificity.

An advantage of our approach is the consideration of the cellular context that kinases and their substrates operate in. We have shown that by incorporating context into the prediction of phosphovariants, we can identify positive examples of phosphovariants at higher levels of specificity than if using sequence alone. There are examples of phosphovariants that represent a trivial loss of phosphorylation; the removal of a phosphorylated serine, threonine or tyrosine residue will

by definition cause loss of phosphorylation. A method that operates only on sequence may be able to correctly predict such cases, but introduce false-positive predictions for cases where the mutation occurs on a residue adjacent to the phosphorylation site. Given the small number of trivial losses contained in our test set, the specificity increase gained by incorporating context into predictions indicates that our method is able to predict a broader spectrum of potential phosphovariants than by using sequence alone. In addition, when comparing our method to an alternative method of predicting the effect of variants on phosphorylation, MIMP, we found that we could predict positive examples of phosphorylation loss at stricter specificity levels than the MIMP method.

While there are over 19,000 examples of missense mutations in the vicinity of a phosphorylation site according to the PhosphoSitePlus® PTMVar dataset (184), we found that the mutations with the strongest propensity for causing phosphorylation loss were associated significantly with disease annotations. While computational analysis of variants has predicted both phosphorylation loss and gain to be associated with disease (214), this study represents an analysis of the predicted effect of variants on experimentally determined phosphorylation sites. However, a greater availability of phosphorylation-gain examples in response to variants would enable a similar analysis to be performed examining the link between phosphorylation gain and disease. There are key residues within a kinase-substrate binding motif that determine the ability of a kinase to catalyse a phosphorylation modification (25, 46). The mutation of these key residues can disrupt the phosphorylation site, and specific effects will depend on the associated kinase. For example, the loss of a proline at the +1 position relative to a phosphorylation site in a proline-directed kinase-substrate binding motif will cause loss of phosphorylation (194, 197, 205). However, the mutation of alternative, non-key, residues within the motif would not be expected to disrupt the phosphorylation site the same extent. As a result, it is to be expected that many missense mutations, even if they are in the vicinity of a phosphorylation site, will not cause a loss of phosphorylation. Our results indicate that PhosphoPICK-SNP is able to detect the mutations that do have an impact on phosphorylation, and therefore have a greater likelihood of being associated with disease.

5.6 Availability

The PhosphoPICK-SNP web-service takes as input protein sequences in Fasta format, and information defining the mutation occurring in the proteins. This follows the format used for missense mutations in Tables 5.1 and 5.2, for example S523G. Users choose which kinases to

make predictions for, and select an E-value threshold for returning results; results that obtain both E_{site} and $E_{combined}$ values below the threshold will be returned. The output is an interactive table of results which details the context score provided to the protein ($R_{substrate}$), the reference and variant scores (R_{site} and V_{site}) obtained from the potential phosphorylation peptide, the $E_{combined}$ value and the peptide itself. More comprehensive information regarding the variant is available in a downloadable tab-delimited text file of the results.

Chapter 6

Conclusion

6.1 Summary

The regulation of much of the molecular functions that proteins are involved in is determined by complex sets of factors. Kinase-mediated protein phosphorylation is a prime example of this, where the determinants of kinase targets can be seen at multiple levels. At the sequence level, protein substrates must contain amino acid sequences suitable to binding by a kinase's catalytic domain. At the cellular level a wide array of processes – localisation, expression, mediating and activating proteins – all contribute towards ensuring kinase-substrate fidelity. The main hypothesis of this thesis was that computational methods for predicting phosphorylation would benefit from a computational framework that can seamlessly integrate the context and sequence determinants of protein phosphorylation. Such a framework would not be unique however, but have the capacity to translate to additional post-translational modifications or protein functions that are regulated through motifs and cellular context.

I this thesis I have proposed a novel computational framework, based on probabilistic graphical modelling, for integrating the sequence and context factors that regulate phosphorylation. Chapter 2 showed how protein context, in the form of protein-protein interaction and association networks, as well as protein abundance across the cell cycle, could be incorporated into a Bayesian network model that predicts kinase substrates. The model, named PhosphoPICK, showed reliable prediction accuracy, with an average AUC of 0.86 across the 59 human kinases tested. An important question at this point was, do the kinase-substrate predictions provide additional predictive power to a sequence-operating method of phosphorylation prediction? By complementing kinase-specific phosphorylation site predictions from existing methods with

PhosphoPICK substrate scores, I found that PhosphoPICK could indeed improve the prediction accuracy of sequence-operating predictors. For some kinases the improvement was substantial: in particular, when the PhosphoPICK model complemented scores from the Predikin and GPS methods, it improved the prediction of CMGC kinase phosphorylation sites by over two-fold. This was therefore a demonstration that the context model could be used to supplement sequence methods of phosphorylation prediction, and improve prediction accuracy.

While Chapter 2 showed that the PhosphoPICK context model could supplement independent sequence models, the main question of the thesis was how to integrate context and sequence information – diverse data types – into a single model of phosphorylation. Chapter 4 presented a probabilistic model of kinase-binding motifs, that incorporates position-specific amino acid frequencies and k-mer frequencies in a way that captures motif sequence context in a kinase- and family-specific manner. The chapter demonstrated how this sequence model could be incorporated into the context model presented in Chapter 2. This seamless integration of features meant that the context information had the capacity to influence the model's expectation of the sequence, and vice versa. Importantly, the combination of context and sequence was found to greatly increase the prediction accuracy of the model when applied to kinase-substrate prediction, with an average 50% increase in prediction accuracy at low false-positive levels (as measured by AUC50). This result was a validation of the driving hypothesis behind the thesis, that phosphorylation prediction methods would gain increased accuracy using a combined model of context and sequence, rather than considering them in isolation. The power of this approach was seen clearly in the improved prediction accuracy of PhosphoPICK compared to alternative methods GPS, NetPhorest and NetworKIN. A comparison of kinase-specific phosphorylation site prediction showed that PhosphoPICK obtained an average sensitivity increase of between 9 and 22% at a 99.9% specificity level; a substantial improvement.

Chapter 4 further demonstrated that the PhosphoPICK methodology is generalisable across species, after it was applied to kinases from mouse and yeast. I found that the combined model was particularly effective with mouse, greatly increasing its prediction accuracy from context alone. The size of available mouse protein-protein interaction networks is much smaller than in human, which indicated a major advantage of the combined model: When one aspect of the data is more limited, such as mouse protein-protein interaction networks, the sequence module within the Bayesian network has the capacity to compensate for the uncertain context information.

One of the potential uses for a model of phosphorylation is to understand how the phosphorylation status of proteins or specific sites can be altered. I have shown at two levels how the

theory underpinning PhosphoPICK, or the algorithm itself, can be used to predict changes in protein phosphorylation status from gene expression, and the effect of SNPs on phosphorylation, respectively. In Chapter 3 I investigated the feasibility of applying the PhosphoPICK method from Chapter 2 to predicting changes in protein phosphorylation status from gene expression. The sbv IMPROVER species translation challenge, which provided participants with gene expression and phosphorylation data collected under various treatment conditions, presented a unique opportunity to evaluate whether PhosphoPICK could be used to predict protein phosphorylation change. Chapter 3 detailed a method to identify genes that explain the changing phosphorylation status of phosphoproteins in response to treatments. By cross-referencing the protein-protein interaction networks of phosphoproteins with genes differentially expressed under the same treatment conditions as differentially phosphorylated proteins, a candidate set of genes could be identified. Using the expression of these genes as input features into SVM and RF classifiers, phosphorylation status change could be predicted from gene expression with promising accuracy. From the first sub-challenge in the competition, the method was able to predict phosphorylation status change in rat cells from rat gene expression with an average AUC of 0.74 on a blind hold-out set, and an average AUC of 0.86 as measured by cross-validation on training data. The method was also ranked 6 out of 21 in the competition, further demonstrating the utility of the approach. These results were an indicator that the PhosphoPICK approach can be extended to predict protein phosphorylation status change in response to treatment conditions.

Building on from the methods for predicting kinase targets (at both the substrate level, and site level) presented in Chapter 4, Chapter 5 showed how PhosphoPICK could be used to analyse the effect of amino acid variations on phosphorylation sites. Non-synonymous single nucleotide polymorphisms (nsSNPs) have the capacity to cause loss or gain of phosphorylation through the modification of key amino acids that determine kinase binding affinity. The studies that have identified examples of phosphorylation change resulting from nsSNPs have invariably found them in the context of disease-associated variants, highlighting the importance of identifying the variants that do cause loss or gain of phosphorylation. Chapter 5 presented a method that uses PhosphoPICK to construct a background of variant effects, based on score differences between a reference and variant peptide according to the sequence model, across missense mutations collected in the UniProt database. Comparing a novel variant against the background allowed a level of significance, as determined by an empirical P-value to be calculated. Combining this with context scores generated by the combined PhosphoPICK Bayesian network model, the method was able to score the effect of a variant based not only on the difference in sequence scores caused by the mutation, but the prior belief of the query kinase

targeting the protein. Comparing this "combined score" with using sequence alone, as well as an alternative method, MIMP, I demonstrated that PhophoPICK is able to identify true positive examples of phosphovariants at higher threshold stringency than either MIMP or the sequence method alone. This demonstrates that just as predicting phosphorylation sites under ordinary circumstances benefits from the combination of context and sequence, understanding the effect of amino acid variants on phosphorylation also benefits from protein context.

6.2 A framework for modelling biological systems

An overarching goal in this work was to design a method that could integrate the sequence and context factors that regulate phosphorylation. This is a concept that is not unique to phosphorylation – there are many biological processes that rely on a combination of linear motifs and interacting proteins to maintain specificity. The method that has been presented here should be considered a framework that could be applied to different biological processes, with alternative post-translational modifications being an obvious choice for candidate studies.

There are many different types of post-translational modifications that proteins can undergo; these can involve structural change to the protein (such as proteolytic cleavage), chemical modifications like phosphorylation, or the linkage of an additional protein. SUMO (short ubiquitin-like modifier) is a modification involving the addition of the small SUMO protein to a lysine residue on a protein, and is involved in the regulation of a diverse range of molecular functions (215). The importance of SUMOylation can be seen in its regulation of protein promyelocytic leukemia (PML) nuclear body (NB) formation. The PML-NB is an important sub-nuclear compartment, found in many tissues, and appears to have a highly dynamic role in regulating an array of processes, including DNA repair and transcription, in response to cellular stresses (216, 217). The importance of the PML-NB is illustrated by the link between its aberrant function and leukaemia, as well as tumours (216). A critical regulatory component of PML-NB is SUMO: PML-NB function is regulated by SUMO (218), and the SUMOylation of PML-NB proteins is required for their localisation to the nucleus, and the correct formation of the PML-NB (219).

There are four SUMO protein paralogues that can be covalently attached to a lysine residue on a substrate protein (220). The process of SUMOylation follows a cascade of enzymes, whereby an activating enzyme (E1) first activates the SUMO protein, which is then transferred to a substrate protein by the E2 enzyme. The conjugation of the SUMO protein to the substrate is

often performed by the E2 enzyme in conjunction with the ligase enzyme, E3. The regulation of SUMO targets is complex, and involves a combination of context and sequence characteristics, as for phosphorylation. SUMOylation by specific SUMO paralogues is regulated by factors such as sub-cellular location and cell-cycle stage (221); in addition, SUMOylation can occur in response to cellular stress (222). The binding of SUMO to a target protein is known to occur within a set of well-defined motifs – some of which are phosphorylation-dependent. The best known consensus motif for SUMO is the ψ**K**XE motif, where **K** is the SUMOylated lysine residue, ψ is a hydrophobic residue and X is any amino acid (223). An example of a phosphorylation-dependent motif follows the form ψ**K**XEXXS$_p$P, where S$_p$ is the phosphorylation site (224). Even more so than the predictors that have been built for phosphorylation, the existing methods for predicting protein SUMOylation sites operate primarily on sequence motif data (225–229).

There are parallels that can be made between phosphorylation and SUMOylation; the modifications are both regulated by the presence of valid motifs and context factors. These parallels are a strong indication that the modelling framework presented in this thesis could feasibly be applied to the SUMO modification. While the number of known protein SUMOylation sites is small compared to phosphorylation, there are sufficient to train and evaluate a predictive model, with over 850 sites currently recorded in the PhosphoSitePlus® database (184). Furthermore, the protein-protein interaction and association databases used in this thesis, BioGRID (209) and STRING (148), should provide the context information necessary to form the basis of a SUMOylation predictor that follows the framework underlying PhosphoPICK. Based on the results presented in this thesis, a SUMOylation predictor that captures the sequence and context conditions that determine SUMOylation would be expected to gain substantial increases in prediction accuracy over the methods that operate on sequence alone.

There are many more examples of PTMs, and motif-based PTM predictors that could be given. Sequence-operating methods have been developed for a variety of post-translational modifications, such as methylation (230, 231), glycosylation (232) and acetylation (233). This thesis has illustrated the power of leveraging not only the sequence information intrinsic to proteins, but the context that the proteins operate in, for predicting kinase-substrate phosphorylation events. But more than that, it has presented a computational framework to enable a comprehensive modelling of the complex factors that regulate the diversity of protein modifications.

Appendix A

Chapter 2 supplementary material

TABLE A.1: Model prediction accuracy (measured using AUC) when varying STRING thresholds are used to add protein interactors to the model.

	kinase	40	60	80
CMGC	CDK1	0.86±0.004	**0.87±0.004**	0.75±0.004
	CDK2	0.89±0.004	**0.91±0.004**	0.79±0.006
	CDK5	0.92±0.006	**0.96±0.004**	0.74±0.018
	CDK7	**0.92±0.022**	0.91±0.029	0.83±0.029
	GSK3B	0.87±0.006	**0.88±0.005**	0.83±0.011
	MAPK1	0.86±0.003	**0.89±0.004**	0.8±0.004
	MAPK3	0.88±0.006	**0.90±0.011**	0.78±0.014
	MAPK8	0.90±0.009	**0.92±0.008**	0.84±0.009
	MAPK9	0.89±0.05	**0.94±0.043**	0.65±0.083
	MAPK14	0.94±0.003	**0.95±0.008**	0.89±0.005
AGC	AKT1	0.88±0.002	**0.91±0.001**	**0.91±0.004**
	GRK2	**0.87±0.01**	**0.87±0.01**	0.43±0.13
	PDPK1	**0.91±0.012**	**0.91±0.021**	0.75±0.027
	PRKACA	**0.97±0.003**	0.96±0.002	0.58±0.009
	PRKCA	**0.77±0.006**	0.76±0.005	0.55±0.007
	PRKCB	**0.86±0.017**	**0.86±0.011**	0.55±0.045
	PRKCD	0.85±0.008	**0.86±0.012**	0.56±0.011
	PRKCE	0.82±0.024	**0.83±0.027**	0.52±0.107
	PRKCG	**0.90±0.014**	**0.90±0.01**	0.55±0.035
	PRKCH	**0.54±0.113**	0.52±0.098	**0.54±0.13**
	PRKCT	0.91±0.028	**0.93±0.031**	0.39±0.103
	PRKCZ	**0.90±0.011**	**0.90±0.009**	0.47±0.025

Continued on next page

	kinase	40	60	80
			Continued from previous page	
	PRKG1	**0.90±0.013**	**0.90±0.01**	0.81±0.011
	ROCK1	0.89±0.01	0.79±0.011	0.29±0.042
	RSK1	**0.93±0.009**	**0.93±0.008**	0.77±0.026
	RSK2	**0.71±0.048**	**0.71±0.036**	0.36±0.066
TK	ABL1	0.96±0.003	**0.97±0.006**	0.90±0.004
TK	BTK	**0.72±0.106**	0.69±0.11	0.52±0.104
TK	CSK	0.89±0.04	**0.91±0.036**	0.57±0.088
TK	EGFR	0.94±0.032	**0.95±0.001**	0.93±0.004
TK	FYN	0.94±0.002	**0.96±0.01**	0.82±0.017
TK	HCK	**0.95±0.023**	**0.96±0.046**	0.82±0.059
TK	INSR	**0.95±0.015**	**0.93±0.017**	0.92±0.033
TK	JAK1	0.75±0.124	**0.76±0.098**	0.52±0.143
TK	JAK2	0.92±0.034	**0.97±0.036**	0.83±0.085
TK	LCK	**0.97±0.01**	0.96±0.011	0.93±0.014
TK	LYN	0.86±0.024	**0.87±0.02**	0.82±0.016
TK	RET	**0.68±0.073**	0.69±0.096	0.55±0.083
TK	SRC	0.87±0.004	**0.89±0.003**	0.86±0.003
TK	SYK	**0.98±0.007**	**0.98±0.004**	0.89±0.009
TK	ZAP70	**0.95±0.06**	0.94±0.059	0.58±0.1
CAMK	CAMK1A	**0.60±0.081**	0.56±0.083	0.61±0.1
CAMK	CAMK2A	**0.85±0.01**	0.81±0.021	0.53±0.031
CAMK	CAMK2G	**0.99±0.006**	0.98±0.012	0.70±0.046
CAMK	CHK1	**0.92±0.041**	0.91±0.038	0.91±0.036
CAMK	LKB1	**0.96±0.022**	0.88±0.03	0.73±0.056
CAMK	MAPKAPK2	**0.93±0.007**	**0.93±0.01**	0.89±0.02
Combined	ATM	0.97±0.005	**0.98±0.003**	0.98±0.004
Combined	ATR	**0.93±0.033**	0.92±0.047	0.72±0.105
Combined	AURKB	**1.00±0.002**	0.91±0.03	0.93±0.015
Combined	CSNK1A1	0.86±0.025	**0.86±0.017**	0.41±0.048
Combined	CSNK1D	**0.63±0.143**	**0.63±0.147**	0.41±0.041
Combined	CSNK2A1	0.87±0.004	**0.89±0.005**	0.69±0.008
Combined	CSNK2A2	**0.95±0.007**	**0.95±0.004**	0.60±0.012
Combined	CSNK2B	**0.88±0.012**	0.87±0.012	0.37±0.04
Combined	PAK1	**0.54±0.023**	0.49±0.025	0.52±0.021
Combined	PAK2	0.38±0.127	**0.40±0.115**	0.39±0.128
Combined	PLK1	**0.92±0.004**	**0.92±0.006**	0.89±0.012
Combined	PRKDC	**0.81±0.07**	**0.81±0.068**	0.63±0.131

TABLE A.2: Model prediction accuracy (measured using AUC) for varying numbers of interaction connections to kinase variables in Bayesian network models.

	kinase	25	40	50
CMGC	CDK1	**0.87±0.004**	**0.87±0.003**	0.86±0.004
	CDK2	**0.91±0.004**	0.89±0.006	0.88±0.007
	CDK5	**0.96±0.004**	**0.96±0.004**	**0.96±0.001**
	CDK7	0.91±0.029	**0.92±0.021**	0.91±0.035
	GSK3B	**0.88±0.005**	0.87±0.008	0.87±0.007
	MAPK1	**0.89±0.004**	0.88±0.004	0.87±0.004
	MAPK3	**0.90±0.011**	0.86±0.016	0.83±0.017
	MAPK8	**0.92±0.008**	**0.92±0.009**	**0.92±0.011**
	MAPK9	**0.94±0.043**	**0.94±0.044**	**0.94±0.043**
	MAPK14	**0.95±0.008**	**0.95±0.009**	**0.95±0.008**
AGC	AKT1	**0.91±0.001**	0.90±0.002	0.90±0.002
	GRK2	0.87±0.01	0.88±0.027	**0.89±0.022**
	PDPK1	**0.91±0.021**	0.91±0.021	0.91±0.021
	PRKACA	**0.96±0.002**	0.95±0.002	0.95±0.002
	PRKCA	0.76±0.005	0.76±0.006	0.77±0.005
	PRKCB	**0.86±0.011**	0.85±0.011	0.85±0.011
	PRKCD	**0.86±0.012**	**0.86±0.011**	**0.86±0.011**
	PRKCE	0.83±0.027	**0.84±0.024**	**0.84±0.028**
	PRKCG	0.90±0.01	**0.91±0.014**	**0.91±0.01**
	PRKCH	0.52±0.098	0.55±0.013	**0.57±0.13**
	PRKCT	**0.93±0.031**	0.90±0.042	0.92±0.037
	PRKCZ	**0.90±0.009**	**0.90±0.011**	**0.90±0.011**
	PRKG1	0.90±0.01	0.89±0.015	**0.91±0.008**
	ROCK1	0.79±0.011	**0.80±0.011**	0.79±0.011
	RSK1	0.93±0.008	**0.94±0.01**	**0.94±0.012**
	RSK2	**0.71±0.036**	0.70±0.029	0.70±0.031
TK	ABL1	**0.97±0.006**	**0.97±0.006**	**0.97±0.006**
	BTK	0.69±0.11	**0.77±0.088**	0.76±0.1
	CSK	**0.91±0.036**	**0.91±0.028**	**0.91±0.046**
	EGFR	**0.95±0.001**	**0.95±0.001**	**0.95±0.001**
	FYN	**0.96±0.01**	**0.96±0.008**	**0.96±0.007**
	HCK	**0.95±0.046**	0.94±0.029	0.94±0.034
	INSR	0.93±0.017	0.92±0.023	**0.94±0.024**
	JAK1	**0.76±0.098**	0.71±0.101	0.74±0.13
	JAK2	**0.97±0.036**	0.96±0.042	0.92±0.033
	LCK	0.96±0.01	**0.97±0.009**	0.96±0.009
	LYN	**0.87±0.02**	0.86±0.022	0.86±0.016

Continued on next page

	kinase	25	40	50
			Continued from previous page	
	RET	0.69±0.096	**0.72±0.09**	0.67±0.05
	SRC	**0.89±0.003**	**0.89±0.003**	0.88±0.004
	SYK	**0.98±0.004**	**0.98±0.004**	0.87±0.003
	ZAP70	0.94±0.059	**0.96±0.055**	0.93±0.058
CAMK	CAMK1A	0.56±0.083	**0.58±0.09**	0.57±0.1
CAMK	CAMK2A	0.81±0.021	**0.82±0.017**	0.81±0.019
CAMK	CAMK2G	0.98±0.012	0.98±0.011	**0.99±0.009**
CAMK	CHK1	0.91±0.038	**0.92±0.037**	0.91±0.04
CAMK	LKB1	0.88±0.03	**0.89±0.034**	**0.89±0.031**
CAMK	MAPKAPK2	**0.93±0.01**	**0.93±0.011**	**0.93±0.007**
Combined	ATM	**0.98±0.003**	0.97±0.001	0.97±0.004
Combined	ATR	0.92±0.047	**0.97±0.053**	0.95±0.049
Combined	AURKB	0.91±0.03	**0.93±0.024**	**0.93±0.021**
Combined	CSNK1A1	0.86±0.017	**0.89±0.016**	**0.87±0.02**
Combined	CSNK1D	0.63±0.147	0.61±0.113	**0.64±0.05**
Combined	CSNK2A1	**0.89±0.005**	**0.89±0.006**	0.86±0.003
Combined	CSNK2A2	0.95±0.004	**0.96±0.004**	0.95±0.004
Combined	CSNK2B	0.87±0.012	0.87±0.015	**0.88±0.013**
Combined	PAK1	**0.49±0.025**	**0.57±0.058**	0.49±0.026
Combined	PAK2	0.40±0.115	**0.49±0.099**	0.43±0.127
Combined	PLK1	0.92±0.006	0.92±0.005	**0.94±0.018**
Combined	PRKDC	**0.81±0.068**	0.77±0.047	**0.80±0.071**

Table A.3: Comparison of model prediction accuracy (measured using AUC) between using STRING with all data sources (normal) and when STRING text mining influence for a test kinase has been removed.

	kinase	normal	text mining removed
CMGC	CDK1	**0.87±0.004**	0.76±0.004
CMGC	CDK2	**0.91±0.004**	0.88±0.005
CMGC	CDK5	**0.96±0.004**	0.94±0.015
CMGC	CDK7	**0.91±0.029**	0.89±0.027
CMGC	GSK3B	**0.88±0.005**	0.82±0.008
CMGC	MAPK1	**0.89±0.004**	0.84±0.005
CMGC	MAPK3	**0.90±0.011**	0.73±0.014
CMGC	MAPK8	**0.92±0.008**	0.91±0.012
CMGC	MAPK9	**0.94±0.043**	0.92±0.02
CMGC	MAPK14	**0.95±0.008**	0.94±0.01

Continued on next page

	kinase	normal	text mining removed
		Continued from previous page	
AGC	AKT1	**0.91±0.001**	0.89±0.003
	GRK2	0.87±0.01	0.86±0.024
	PDPK1	**0.91±0.021**	0.76±0.026
	PRKACA	**0.96±0.002**	**0.96±0.003**
	PRKCA	0.76±0.005	0.65±0.007
	PRKCB	**0.86±0.011**	0.63±0.022
	PRKCD	**0.86±0.012**	0.66±0.012
	PRKCE	**0.83±0.027**	0.72±0.066
	PRKCG	**0.90±0.01**	0.87±0.016
	PRKCH	0.52±0.098	**0.54±0.132**
	PRKCT	0.93±0.031	**1.00±0.008**
	PRKCZ	**0.90±0.009**	0.86±0.008
	PRKG1	**0.90±0.01**	0.88±0.005
	ROCK1	**0.79±0.011**	0.50±0.012
	RSK1	**0.93±0.008**	0.86±0.018
	RSK2	**0.71±0.036**	0.44±0.054
TK	ABL1	**0.97±0.006**	0.94±0.006
	BTK	0.69±0.11	**0.70±0.13**
	CSK	**0.91±0.036**	0.80±0.058
	EGFR	**0.95±0.001**	0.92±0.008
	FYN	**0.96±0.01**	0.91±0.019
	HCK	**0.95±0.046**	0.79±0.058
	INSR	**0.93±0.017**	0.80±0.042
	JAK1	**0.76±0.098**	0.73±0.11
	JAK2	**0.97±0.036**	0.92±0.053
	LCK	0.96±0.011	**0.97±0.006**
	LYN	**0.87±0.02**	**0.87±0.018**
	RET	**0.69±0.096**	0.65±0.135
	SRC	**0.89±0.003**	0.88±0.004
	SYK	**0.98±0.004**	0.94±0.016
	ZAP70	**0.94±0.059**	0.88±0.128
CAMK	CAMK1A	0.56±0.083	**0.59±0.056**
	CAMK2A	**0.81±0.021**	0.66±0.015
	CAMK2G	**0.98±0.012**	0.92±0.036
	CHK1	0.91±0.038	**0.96±0.03**
	LKB1	**0.88±0.03**	**0.88±0.034**
	MAPKAPK2	**0.93±0.01**	0.83±0.01
	ATM	**0.98±0.001**	0.96±0.007

Continued on next page

	kinase	normal	text mining removed
		Continued from previous page	
Combined	ATR	0.92±0.047	**0.94±0.048**
	AURKB	0.91±0.03	0.85±0.022
	CSNK1A1	**0.86±0.017**	0.69±0.049
	CSNK1D	0.63±0.147	0.62±0.105
	CSNK2A1	**0.89±0.005**	0.88±0.004
	CSNK2A2	**0.95±0.004**	0.93±0.006
	CSNK2B	**0.87±0.012**	0.81±0.013
	PAK1	0.49±0.025	**0.50±0.022**
	PAK2	**0.40±0.115**	0.38±0.12
	PLK1	0.92±0.006	**0.93±0.018**
	PRKDC	0.81±0.068	**0.82±0.044**

TABLE A.4: Comparison between classifying phosphorylation sites using Predikin, and classifying phosphorylation sites when Predikin score is combined with PhosphoPICK predictions using two methods – sum and product. Comparisons were made using AUC50 (area under an ROC curve calculated up to the first 50 false positives), and sensitivity (predicted true positives/total true positives) at the threshold that yielded the fiftieth false positive. In case of a tie, an arbitrary order is used to determine the top fifty false positives.

	Kinase	AUC50			Sensitivity		
		Predikin	Combined		Predikin	Combined	
			Sum	Product		Sum	Product
CMGC	CDK1	0.018	0.060	**0.061**	0.030	**0.090**	**0.090**
	CDK2	0.009	0.065	**0.068**	0.017	0.116	**0.124**
	CDK5	0.006	**0.115**	0.084	0.013	**0.160**	0.120
	CDK7	0.000	**0.009**	0.000	0.000	**0.059**	0.000
	GSK3B	0.010	**0.028**	0.018	0.025	**0.042**	0.025
	MAPK1	0.004	0.012	**0.014**	0.010	**0.030**	**0.030**
	MAPK3	0.005	0.017	**0.019**	0.035	0.035	**0.043**
	MAPK8	0.006	0.056	**0.064**	0.018	**0.091**	**0.091**
	MAPK9	0.011	**0.150**	0.024	0.036	**0.286**	0.071
	MAPK14	0.015	**0.022**	0.021	0.021	**0.053**	**0.053**
AGC	AKT1	0.018	**0.048**	0.046	0.048	**0.114**	0.095
	GRK2	0.000	**0.029**	0.000	0.000	**0.080**	0.000
	PDPK1	**0.206**	0.182	0.103	**0.286**	0.262	0.143
	PRKACA	0.008	**0.051**	0.041	0.018	**0.076**	0.058
	PRKCA	**0.013**	0.012	0.013	0.019	0.019	**0.023**
	PRKCB	0.000	0.032	**0.050**	0.000	0.063	**0.079**
	PRKCD	0.025	**0.053**	0.045	0.048	**0.071**	0.060

Continued on next page

	Kinase	Predikin	Combined		Predikin	Combined	
					Continued from previous page		
AGC	PRKCE	0.027	**0.038**	0.000	**0.038**	**0.038**	0.000
	PRKCG	0.000	0.017	**0.025**	0.000	0.050	**0.075**
	PRKCH	**0.058**	**0.058**	0.000	**0.083**	**0.083**	0.000
	PRKCT	0.000	**0.140**	0.000	0.000	**0.400**	0.000
	PRKCZ	0.000	0.068	**0.073**	0.000	**0.103**	**0.103**
	PRKG1	0.000	**0.088**	0.085	0.000	**0.128**	**0.128**
	ROCK1	0.000	**0.077**	0.073	0.000	**0.130**	**0.130**
	RSK1	0.000	**0.058**	0.055	0.000	**0.091**	**0.091**
	RSK2	0.000	0.000	0.000	0.000	0.000	0.000
TK	ABL1	0.012	**0.060**	0.056	0.022	0.089	**0.111**
	BTK	0.002	**0.020**	0.000	**0.048**	**0.048**	0.000
	CSK	0.000	0.000	0.000	0.000	0.000	0.000
	EGFR	0.062	0.118	**0.119**	0.107	**0.161**	**0.161**
	FYN	0.016	**0.018**	0.009	**0.049**	0.033	0.016
	HCK	**0.011**	0.000	0.000	**0.032**	0.000	0.000
	INSR	0.049	**0.172**	0.160	0.094	**0.226**	0.189
	JAK1	0.000	0.000	0.000	0.000	0.000	0.000
	JAK2	**0.027**	0.008	0.000	**0.030**	**0.030**	0.000
	LCK	0.062	0.134	**0.146**	0.099	0.225	**0.239**
	LYN	**0.043**	0.027	0.030	**0.088**	0.035	0.053
	RET	0.000	0.000	0.000	0.000	0.000	0.000
	SRC	0.024	0.037	**0.038**	0.045	**0.072**	**0.072**
	SYK	**0.204**	0.193	0.182	**0.283**	0.245	0.226
	ZAP70	0.201	**0.235**	0.023	0.280	**0.400**	0.040
CAMK	CAMK1A	0.177	**0.180**	0.000	**0.333**	**0.333**	0.000
	CAMK2A	0.000	**0.059**	0.031	0.000	**0.118**	0.044
	CAMK2G	0.000	**0.044**	0.000	0.000	**0.095**	0.000
	CHK1	0.090	**0.138**	0.000	0.105	**0.158**	0.000
	LKB1	0.205	**0.448**	0.000	0.308	**0.538**	0.000
	MAPKAPK2	0.023	**0.224**	0.192	0.031	**0.281**	0.219
combined	AURKB	0.056	**0.124**	0.037	0.081	**0.243**	0.081
	CSNK1A1	0.000	**0.047**	0.033	0.000	**0.104**	0.083
	CSNK1D	0.000	0.000	0.000	0.000	0.000	0.000
	PAK1	0.004	**0.014**	0.000	0.026	**0.053**	0.000
	PAK2	**0.028**	0.027	0.000	**0.057**	**0.057**	0.000
	PLK1	0.000	0.029	**0.029**	0.000	**0.051**	**0.051**

TABLE A.5: Comparison between classifying phosphorylation sites using Predikin, and classifying phosphorylation sites when GPS score is combined with PhosphoPICK predictions using two methods – sum and product. Comparisons were made using AUC50 (area under an ROC curve calculated up to the first 50 false positives), and sensitivity (predicted true positives/total true positives) at the threshold that yielded the fiftieth false positive. In case of a tie, an arbitrary order is used to determine the top fifty false positives.

	Kinase	AUC50			Sensitivity		
		GPS	Combined		GPS	Combined	
			Sum	Product		Sum	Product
CMGC	CDK1	0.000	**0.022**	0.019	0.000	**0.059**	0.040
	CDK2	0.000	**0.070**	0.068	0.000	**0.137**	**0.137**
	CDK5	0.000	**0.094**	0.087	0.000	**0.173**	0.147
	CDK7	0.291	0.323	**0.526**	0.389	0.333	**0.778**
	GSK3B	0.015	0.015	**0.018**	**0.042**	0.034	0.034
	MAPK1	0.006	**0.035**	**0.035**	0.015	**0.049**	**0.049**
	MAPK3	0.008	0.039	**0.039**	0.009	**0.087**	**0.087**
	MAPK8	0.000	**0.043**	**0.043**	0.000	**0.109**	**0.109**
	MAPK9	0.017	**0.099**	0.098	0.0357	**0.179**	**0.179**
	MAPK14	0.020	0.051	**0.053**	0.021	**0.104**	0.094
AGC	AKT1	0.073	0.265	**0.265**	0.114	**0.352**	**0.352**
	GRK2	0.080	0.204	**0.613**	0.111	0.370	**0.815**
	PDPK1	0.3940	0.388	**0.496**	0.476	0.429	**0.667**
	PRKACA	0.010	**0.102**	0.099	0.026	**0.174**	**0.174**
	PRKCA	0.005	0.021	**0.021**	0.013	0.032	**0.039**
	PRKCB	0.005	**0.040**	**0.040**	0.016	**0.047**	**0.047**
	PRKCD	0.027	0.070	**0.080**	0.047	0.094	**0.118**
	PRKCE	0.147	**0.158**	0.060	**0.192**	**0.192**	0.077
	PRKCG	0.125	**0.149**	0.020	0.125	**0.200**	0.025
	PRKCH	**0.156**	0.153	0.000	**0.167**	**0.167**	0.000
	PRKCT	0.000	0.040	**0.108**	0.000	**0.400**	0.200
	PRKCZ	0.148	0.190	**0.196**	0.172	0.207	**0.241**
	PRKG1	0.159	**0.164**	0.110	**0.231**	**0.231**	0.180
	ROCK1	0.073	0.125	**0.131**	0.152	0.152	**0.174**
	RSK1	**0.254**	0.160	0.162	**0.273**	0.182	0.182
	RSK2	**0.711**	0.653	0.138	**0.778**	**0.778**	0.222
	ABL1	0.235	0.270	**0.276**	0.311	**0.378**	**0.378**
	BTK	0.230	**0.232**	0.050	**0.286**	**0.286**	0.190
	CSK	0.245	**0.370**	0.090	**0.500**	**0.500**	0.286
	EGFR	0.221	**0.339**	0.333	0.339	**0.482**	0.464
	FYN	**0.086**	0.053	0.062	**0.115**	0.098	0.098

Continued on next page

	Kinase	GPS	Combined		GPS	Combined	
					Continued from previous page		
TK	HCK	**0.195**	0.028	0.001	**0.290**	0.065	0.032
	INSR	0.121	0.249	**0.282**	0.170	0.302	**0.358**
	JAK1	**0.368**	**0.368**	0.082	**0.368**	**0.368**	0.263
	JAK2	**0.116**	0.122	0.083	**0.212**	**0.212**	0.182
	LCK	0.091	**0.229**	0.221	0.141	**0.352**	0.296
	LYN	**0.216**	0.198	0.212	0.263	0.333	**0.351**
	RET	**0.464**	0.461	0.195	**0.571**	**0.571**	0.429
	SRC	0.077	0.086	**0.089**	0.126	**0.212**	0.167
	SYK	0.288	**0.458**	0.452	0.434	**0.585**	**0.585**
	ZAP70	0.533	**0.557**	0.305	0.720	**0.760**	0.480
CAMK	CAMK1A	**0.653**	0.647	0.000	**0.667**	**0.667**	0.000
	CAMK2A	0.021	0.117	**0.146**	0.074	0.221	**0.235**
	CAMK2G	0.005	**0.083**	0.024	0.048	**0.238**	0.143
	CHK1	0.148	**0.202**	0.051	0.211	**0.263**	0.105
	LKB1	**0.722**	0.706	0.511	**0.769**	**0.769**	0.538
	MAPKAPK2	0.044	0.344	**0.348**	0.093	**0.375**	**0.375**
combined	ATM	0.033	**0.148**	0.146	0.088	**0.221**	**0.221**
	ATR	0.000	**0.025**	0.019	0.000	**0.054**	**0.054**
	AURKB	0.112	**0.178**	0.169	0.162	**0.243**	**0.243**
	CSNK1A1	**0.341**	0.227	0.200	**0.354**	0.271	0.229
	CSNK1D	**0.016**	0.012	0.000	**0.056**	**0.056**	0.000
	CSNK2A1	0.016	**0.055**	0.052	0.031	0.094	**0.100**
	CSNK2A2	0.039	0.187	**0.195**	0.069	0.276	**0.302**
	CSNK2B	0.066	0.069	**0.074**	**0.094**	**0.094**	**0.094**
	PAK1	0.023	**0.025**	0.000	**0.026**	**0.026**	0.0
	PAK2	**0.108**	0.102	0.018	**0.200**	**0.200**	0.057
	PLK1	0.093	0.296	**0.299**	0.128	**0.462**	**0.462**

TABLE A.6: Comparison between classifying phosphorylation sites using NetworKIN, and classifying phosphorylation sites when NetworKIN score is combined with PhosphoPICK predictions using two methods – sum and product. Comparisons were made using AUC50 (area under an ROC curve calculated up to the first 50 false positives), and sensitivity (predicted true positives/total true positives) at the threshold that yielded the fiftieth false positive. In case of a tie, an arbitrary order is used to determine the top fifty false predictions. The specificity at this threshold is 0.9995 for serine/threonine kinases, and 0.998 for the tyrosine kinases.

	Kinase	AUC50			Sensitivity		
		NetworKIN	Combined		NetworKIN	Combined	
			Sum	Product		Sum	Product
CMGC	CDK1	0.031	**0.034**	**0.034**	0.056	**0.061**	**0.061**
	CDK2	0.125	**0.159**	**0.159**	0.242	**0.274**	**0.274**
	CDK5	0.047	**0.173**	0.169	0.080	**0.240**	**0.240**
	CDK7	**0.109**	0.000	0.000	**0.222**	0.000	0.000
	GSK3B	0.008	**0.045**	**0.045**	0.017	**0.067**	**0.067**
	MAPK1	0.051	**0.081**	0.076	0.104	**0.129**	**0.129**
	MAPK3	0.047	0.049	**0.052**	0.096	0.078	**0.087**
	MAPK8	0.004	**0.013**	**0.013**	**0.036**	**0.018**	**0.018**
	MAPK9	**0.024**	0.017	0.000	**0.036**	**0.036**	0.000
AGC	GRK2	**0.018**	0.000	0.000	**0.037**	0.000	0.000
	PRKACA	0.046	**0.067**	0.065	0.093	0.111	**0.115**
	PRKCA	**0.015**	0.011	0.010	**0.026**	**0.016**	**0.013**
	PRKCB	0.000	**0.064**	**0.064**	0.000	**0.079**	**0.079**
	PRKCD	0.000	0.000	0.000	0.000	0.000	0.000
	PRKCE	0.000	0.000	0.000	0.000	0.000	0.000
	PRKCG	**0.039**	0.025	0.003	**0.077**	0.026	0.026
	PRKCH	**0.040**	**0.040**	0.000	**0.083**	**0.083**	0.000
	PRKCT	0.176	0.176	0.176	0.200	0.200	0.200
	PRKCZ	**0.017**	0.000	0.000	0.034	0.034	0.034
	ROCK1	**0.032**	0.019	0.019	0.065	0.065	0.065
	RSK2	**0.054**	0.000	0.029	**0.111**	0.000	**0.111**
TK	ABL1	**0.090**	0.060	0.060	**0.178**	0.111	0.089
	BTK	**0.023**	0.000	0.000	**0.048**	0.000	0.000
	EGFR	0.135	**0.161**	**0.161**	0.250	**0.268**	**0.268**
	FYN	**0.068**	0.050	0.048	**0.131**	0.082	0.082
	HCK	**0.047**	0.000	0.000	**0.097**	0.000	0.000
	INSR	0.072	0.092	**0.119**	0.151	0.132	**0.170**
	LCK	0.118	**0.125**	**0.125**	**0.169**	0.155	0.155
	LYN	**0.106**	0.088	0.086	**0.175**	0.158	0.158
	SRC	0.043	**0.047**	0.046	0.081	**0.095**	**0.095**

Continued on next page

	Kinase	NetworKIN	Combined		NetworKIN	Combined	
					Continued from previous page		
	SYK	0.020	**0.023**	**0.023**	0.058	**0.077**	**0.077**
	ZAP70	**0.120**	0.065	0.000	**0.200**	0.080	0.000
CAMK	CAMK1A	0.203	**0.227**	0.133	0.333	0.333	0.333
CAMK	CAMK2A	**0.072**	0.031	0.007	**0.147**	0.118	0.015
CAMK	CAMK2G	**0.138**	0.123	0.114	0.238	0.238	0.238
CAMK	LKB1	0.113	**0.278**	**0.278**	0.231	**0.308**	**0.308**
combined	ATM	0.100	**0.101**	**0.101**	0.204	0.204	0.204
combined	ATR	**0.040**	0.000	0.000	**0.081**	0.000	0.000
combined	CSNK1A1	0.008	0.010	**0.014**	0.021	0.021	0.021
combined	CSNK1D	0.000	0.000	0.000	0.000	0.000	0.000
combined	CSNK2A1	**0.062**	0.053	0.052	0.102	**0.109**	0.102
combined	CSNK2A2	0.071	0.100	**0.101**	0.273	**0.327**	0.282
combined	PAK1	**0.026**	0.011	0.000	**0.053**	0.026	0.000
combined	PAK2	**0.388**	0.386	0.378	0.400	0.400	0.400

TABLE A.7: Gene ontology (GO) term enrichment analysis for known CDK2 substrates and predicted substrates. The first two columns in the table show GO terms and their descriptions that were found to be significantly over-represented (Fisher's exact test, Bonferroni correction, E-value<0.05) in known CDK2 substrates, with the E-values shown in the third column. The terms are ordered from most to least significant. The final column contains the E-values for the terms that were found when performing the same enrichment test on the top 300 predictions for PhosphoPICK. If a value is listed as "N/A", then the term was not identified with any protein in the set of predictions.

GO term	Description	Substrates	PhosphoPICK
GO:0005654	nucleoplasm	4.38e-20	8.74e-144
GO:0007049	cell cycle	3.62e-14	4.18e-125
GO:0000278	mitotic cell cycle	9.35e-12	6.95e-214
GO:0005634	nucleus	9.91e-11	8.22e-78
GO:0000082	G1/S transition of mitotic cell cycle	2.92e-10	1.95e-126
GO:0005515	protein binding	1.94e-08	2.00e-52
GO:0045893	positive regulation of transcription, DNA-templated	4.09e-07	0.155
GO:0006974	cellular response to DNA damage stimulus	3.70e-06	4.55e-25
GO:0007050	cell cycle arrest	4.92e-06	0.0010
GO:0006978	DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator	1.08e-05	259.34
GO:0031625	ubiquitin protein ligase binding	1.94e-05	0.08
GO:0006260	DNA replication	2.86e-05	1.10e-59
GO:0051726	regulation of cell cycle	3.10e-05	1.80e-28

Continued on next page

GO term	Description	Substrates	PhosphoPICK
		Continued from previous page	
GO:0008285	negative regulation of cell proliferation	4.83e-05	524.32
GO:0019901	protein kinase binding	7.20e-05	2.76e-09
GO:0000083	regulation of transcription involved in G1/S transition of mitotic cell cycle	7.82e-05	8.14e-13
GO:0000079	regulation of cyclin-dependent protein serine/threonine kinase activity	8.54e-05	1.69e-19
GO:0005730	nucleolus	0.0001	7.06e-34
GO:0005667	transcription factor complex	0.0001	0.0007
GO:0003700	sequence-specific DNA binding transcription factor activity	0.0002	390.26
GO:0005737	cytoplasm	0.0003	1.43e-21
GO:0000307	cyclin-dependent protein kinase holoenzyme complex	0.0003	1.84e-06
GO:0000785	chromatin	0.0003	5.75e-15
GO:0006281	DNA repair	0.0003	8.41e-37
GO:0051301	cell division	0.0003	6.49e-72
GO:0071850	mitotic cell cycle arrest	0.0004	N/A
GO:0031571	mitotic G1 DNA damage checkpoint	0.0006	210.79
GO:0050681	androgen receptor binding	0.0007	179.82
GO:0008134	transcription factor binding	0.0007	1.06e-12
GO:0030521	androgen receptor signaling pathway	0.001	811.58
GO:0008284	positive regulation of cell proliferation	0.002	4.85
GO:0030308	negative regulation of cell growth	0.003	506.21
GO:0000086	G2/M transition of mitotic cell cycle	0.003	2.82e-26
GO:0003682	chromatin binding	0.003	1.35e-08
GO:0071158	positive regulation of cell cycle arrest	0.004	41.04
GO:0043433	negative regulation of sequence-specific DNA binding transcription factor activity	0.004	97.84
GO:0006351	transcription, DNA-templated	0.004	114.1
GO:0006357	regulation of transcription from RNA polymerase II promoter	0.004	247.83
GO:0003713	transcription coactivator activity	0.005	284.81
GO:0006977	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	0.007	5.06e-56
GO:0008156	negative regulation of DNA replication	0.009	57.94
GO:0043550	regulation of lipid kinase activity	0.009	55.29
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	0.01	0.001
GO:0006270	DNA replication initiation	0.01	1.14e-22

Continued on next page

GO term	Description	Substrates	PhosphoPICK
		Continued from previous page	
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	0.01	0.01
GO:0007265	Ras protein signal transduction	0.01	46.02
GO:0001836	release of cytochrome c from mitochondria	0.01	N/A
GO:0005829	cytosol	0.01	4.07e-42
GO:0003677	DNA binding	0.01	1.34e-12
GO:0044212	transcription regulatory region DNA binding	0.02	270.76
GO:0043234	protein complex	0.02	0.007
GO:0004860	protein kinase inhibitor activity	0.02	614.09
GO:0000790	nuclear chromatin	0.03	2.24e-09
GO:0007369	gastrulation	0.03	701.09
GO:0045892	negative regulation of transcription, DNA-templated	0.03	2.06
GO:0045668	negative regulation of osteoblast differentiation	0.03	701.09
GO:0006355	regulation of transcription, DNA-templated	0.04	133.62
GO:0043353	enucleate erythrocyte differentiation	0.045	108.82
GO:0090344	negative regulation of cell aging	0.045	N/A

TABLE A.8: Gene ontology (GO) term enrichment analysis for CDK2 substrates within unique E2F1 targets.

GO term	Description	E-value
GO:0051439	regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	2.81e-11
GO:0051437	positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	2.81e-11
GO:0051436	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	2.81e-11
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	2.81e-11
GO:0006977	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	2.23e-10
GO:0000082	G1/S transition of mitotic cell cycle	7.79e-10
GO:0000278	mitotic cell cycle	9.90e-09
GO:0005654	nucleoplasm	1.62e-08
GO:0002474	antigen processing and presentation of peptide antigen via MHC class I	7.40e-08
GO:0002479	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	7.40e-08
GO:0042590	antigen processing and presentation of exogenous peptide antigen via MHC class I	2.20e-07
GO:0006521	regulation of cellular amino acid metabolic process	2.39e-07

Continued on next page

GO term	Description	E-value
	Continued from previous page	
GO:0043066	negative regulation of apoptotic process	8.78e-07
GO:0000502	proteasome complex	1.42e-06
GO:0000209	protein polyubiquitination	2.35e-06
GO:0042981	regulation of apoptotic process	8.43e-06
GO:0010467	gene expression	1.78e-05
GO:0016032	viral process	4.70e-05
GO:0016071	mRNA metabolic process	0.0003
GO:0006915	apoptotic process	0.0003
GO:0016070	RNA metabolic process	0.0004
GO:0034641	cellular nitrogen compound metabolic process	0.0009
GO:0022624	proteasome accessory complex	0.0015
GO:0005634	nucleus	0.0088
GO:0005829	cytosol	0.0130
GO:0000079	regulation of cyclin-dependent protein serine/threonine kinase activity	0.0151
GO:0005637	nuclear inner membrane	0.0300

TABLE A.9: Gene ontology (GO) term enrichment analysis for CDK2 substrates within unique E2F4 targets.

GO term	Description	E-value
GO:0007049	cell cycle	3.29e-16
GO:0000278	mitotic cell cycle	8.16e-14
GO:0051301	cell division	3.39e-11
GO:0007067	mitotic nuclear division	3.39e-11
GO:0005819	spindle	6.60e-07
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	6.21e-05
GO:0005654	nucleoplasm	6.41e-05
GO:0007094	mitotic spindle assembly checkpoint	7.82e-05
GO:0019901	protein kinase binding	0.0001
GO:0051439	regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	0.0005
GO:0051437	positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	0.0005
GO:0005856	cytoskeleton	0.0016
GO:0005829	cytosol	0.0039
GO:0005524	ATP binding	0.0042
GO:0000082	G1/S transition of mitotic cell cycle	0.0062
GO:0005737	cytoplasm	0.0067
GO:0008283	cell proliferation	0.0072

Continued on next page

GO term	Description	E-value
	Continued from previous page	
GO:0000086	G2/M transition of mitotic cell cycle	0.0072
GO:0007080	mitotic metaphase plate congression	0.0078
GO:0000922	spindle pole	0.0133
GO:0007059	chromosome segregation	0.0185
GO:0030496	midbody	0.0430
GO:0051436	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	0.0430

Table A.10: Gene ontology (GO) term enrichment analysis for CDK2 substrates within unique E2F6 targets.

GO term	Description	E-value
GO:0000278	mitotic cell cycle	4.61e-16
GO:0051439	regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	5.02e-16
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	5.02e-16
GO:0051437	positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	5.02e-16
GO:0051436	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	5.02e-16
GO:0000082	G1/S transition of mitotic cell cycle	8.36e-10
GO:0006977	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	5.50e-09
GO:0042590	antigen processing and presentation of exogenous peptide antigen via MHC class I	3.28e-08
GO:0006521	regulation of cellular amino acid metabolic process	3.28e-08
GO:0002479	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	3.28e-08
GO:0005654	nucleoplasm	7.04e-08
GO:0002474	antigen processing and presentation of peptide antigen via MHC class I	1.14e-07
GO:0000502	proteasome complex	6.81e-07
GO:0000209	protein polyubiquitination	6.81e-07
GO:0042981	regulation of apoptotic process	2.47e-06
GO:0005829	cytosol	4.09e-05
GO:0016071	mRNA metabolic process	4.44e-05
GO:0016032	viral process	4.44e-05
GO:0016070	RNA metabolic process	7.97e-05
GO:0034641	cellular nitrogen compound metabolic process	0.0001
GO:0022624	proteasome accessory complex	0.0001

Continued on next page

GO term	Description	E-value
	Continued from previous page	
GO:0007094	mitotic spindle assembly checkpoint	0.0004
GO:0005680	anaphase-promoting complex	0.0004
GO:0010467	gene expression	0.0006
GO:0070979	protein K11-linked ubiquitination	0.0011
GO:0043066	negative regulation of apoptotic process	0.0015
GO:0007049	cell cycle	0.0044
GO:0051301	cell division	0.0156
GO:0006915	apoptotic process	0.0232
GO:0044281	small molecule metabolic process	0.0324
GO:0030163	protein catabolic process	0.0432

TABLE A.11: Gene ontology (GO) term enrichment analysis for CDK2 substrates within overlapping E2F1, E2F4 and E2F6 targets.

GO term	Description	E-value
GO:0000278	mitotic cell cycle	1.02e-34
GO:0006260	DNA replication	3.13e-25
GO:0005654	nucleoplasm	2.35e-24
GO:0000082	G1/S transition of mitotic cell cycle	2.47e-21
GO:0007049	cell cycle	8.60e-21
GO:0005634	nucleus	5.73e-14
GO:0006271	DNA strand elongation involved in DNA replication	1.04e-12
GO:0007067	mitotic nuclear division	9.91e-09
GO:0006281	DNA repair	4.87e-08
GO:0051301	cell division	8.35e-08
GO:0006270	DNA replication initiation	2.91e-07
GO:0032201	telomere maintenance via semi-conservative replication	1.40e-06
GO:0000722	telomere maintenance via recombination	5.61e-06
GO:0005515	protein binding	7.18e-06
GO:0000775	chromosome, centromeric region	7.29e-06
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	1.36e-05
GO:0003677	DNA binding	1.50e-05
GO:0051439	regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	2.79e-05
GO:0051436	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	2.79e-05
GO:0000777	condensed chromosome kinetochore	9.98e-05
GO:0005694	chromosome	0.0001
GO:0000723	telomere maintenance	0.0001

Continued on next page

GO term	Description	E-value
	Continued from previous page	
GO:0017111	nucleoside-triphosphatase activity	0.0003
GO:0051437	positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	0.0004
GO:0006974	cellular response to DNA damage stimulus	0.0007
GO:0000776	kinetochore	0.0009
GO:0005524	ATP binding	0.0010
GO:0006297	nucleotide-excision repair, DNA gap filling	0.0011
GO:0007094	mitotic spindle assembly checkpoint	0.0011
GO:0051726	regulation of cell cycle	0.0019
GO:0008283	cell proliferation	0.0037
GO:0005819	spindle	0.0054
GO:0000083	regulation of transcription involved in G1/S transition of mitotic cell cycle	0.0104
GO:0006302	double-strand break repair	0.0204
GO:0003690	double-stranded DNA binding	0.0409

TABLE A.12: Gene ontology (GO) term enrichment analysis for predicted CDK2 substrates overlapping with E2F1 and E2F4 targets but not E2F6 targets.

GO term	Description	E-value
GO:0007049	cell cycle	6.26e-20
GO:0000278	mitotic cell cycle	2.33e-15
GO:0000082	G1/S transition of mitotic cell cycle	8.67e-11
GO:0051301	cell division	2.05e-10
GO:0007067	mitotic nuclear division	7.50e-08
GO:0005634	nucleus	6.05e-06
GO:0006260	DNA replication	7.18e-06
GO:0005515	protein binding	1.97e-05
GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process	4.91e-05
GO:0006977	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	0.0001
GO:0051439	regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	0.0005
GO:0048015	phosphatidylinositol-mediated signaling	0.0005
GO:0005654	nucleoplasm	0.0007
GO:0005694	chromosome	0.0020
GO:0007051	spindle organization	0.0057
GO:0051437	positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	0.0057
GO:0051726	regulation of cell cycle	0.0127

Continued on next page

GO term	Description	E-value
	Continued from previous page	
GO:0006281	DNA repair	0.0152
GO:0006974	cellular response to DNA damage stimulus	0.0152
GO:0005874	microtubule	0.0215
GO:0005876	spindle microtubule	0.0351
GO:0005667	transcription factor complex	0.0351
GO:0043066	negative regulation of apoptotic process	0.0432

TABLE A.13: Gene ontology (GO) term enrichment analysis for predicted CDK2 substrates overlapping with E2F1 and E2F6 targets but not E2F4 targets.

GO term	Description	E-value
GO:0019905	syntaxin binding	0.0119
GO:0042770	signal transduction in response to DNA damage	0.0119
GO:0006974	cellular response to DNA damage stimulus	0.0375
GO:0006281	DNA repair	0.0375

TABLE A.14: Gene ontology (GO) term enrichment analysis for predicted CDK2 substrates overlapping with E2F4 and E2F6 targets but not E2F1 targets.

GO term	Description	E-value
GO:0000278	mitotic cell cycle	9.44e-09
GO:0007049	cell cycle	8.01e-05
GO:0051301	cell division	0.0004
GO:0007067	mitotic nuclear division	0.0021
GO:0000082	G1/S transition of mitotic cell cycle	0.0028
GO:0000775	chromosome, centromeric region	0.0238
GO:0005654	nucleoplasm	0.0296
GO:0005694	chromosome	0.0360
GO:0007059	chromosome segregation	0.0371

Appendix B

Chapter 4 supplementary material

TABLE B.1: Sequence model accuracy across **human** kinases when different percentages of kinase phosphorylation peptides were used to determine the set of k-mers added to the sequence model. Table shows median AUC and AUC50 values for classifying kinase phosphorylation sites with the sequence model as determined by 10-fold cross-validation across 10 randomised data-set splits. Kinases are grouped according to their family, with the average prediction accuracy for each family shown.

		AUC			AUC50		
	Kinase	5%	10%	20%	5%	10%	20%
CMGC	CDK2	0.89±0.001	0.89±0.001	0.89±0.001	0.100±0.004	0.105±0.003	0.086±0.003
	CDK1	0.89±0.002	0.89±0.001	0.89±0.002	0.071±0.008	0.081±0.011	0.105±0.009
	ERK2	0.86±0.001	0.86±0.001	0.87±0.002	0.067±0.010	0.063±0.007	0.084±0.009
	ERK1	0.86±0.005	0.85±0.005	0.84±0.005	0.066±0.012	0.035±0.006	0.036±0.007
	GSK3B	0.81±0.006	0.80±0.007	0.80±0.007	0.132±0.014	0.137±0.011	0.107±0.007
	P38A	0.81±0.007	0.81±0.007	0.80±0.007	0.151±0.017	0.150±0.018	0.131±0.017
	JNK1	0.87±0.004	0.85±0.005	0.84±0.005	0.155±0.014	0.074±0.013	0.082±0.014
	CDK5	0.84±0.009	0.85±0.009	0.84±0.011	0.050±0.007	0.086±0.011	0.054±0.011
	JNK2	0.73±0.023	0.71±0.022	0.71±0.018	0.068±0.015	0.054±0.011	0.055±0.007
	CDK7	0.88±0.019	0.78±0.017	0.76±0.018	0.310±0.032	0.270±0.018	0.235±0.052
	GSK3A	0.90±0.026	0.88±0.017	0.85±0.022	0.458±0.045	0.351±0.041	0.219±0.033
	CDK4	0.87±0.012	0.85±0.012	0.83±0.014	0.179±0.025	0.055±0.017	0.065±0.021
	P38B	0.83±0.014	0.81±0.014	0.81±0.014	0.260±0.046	0.217±0.040	0.105±0.049
	HIPK2	0.86±0.013	0.84±0.017	0.84±0.017	0.380±0.043	0.224±0.031	0.229±0.034
	DYRK1A	0.83±0.033	0.80±0.039	0.81±0.030	0.260±0.043	0.147±0.070	0.041±0.035
	CDK9	0.83±0.015	0.80±0.010	0.78±0.011	0.320±0.030	0.227±0.056	0.057±0.018
	DYRK2	0.78±0.019	0.76±0.024	0.72±0.029	0.306±0.043	0.197±0.061	0.000±0.006
	ERK5	0.83±0.016	0.81±0.011	0.82±0.009	0.317±0.034	0.148±0.034	0.073±0.026
	CDK6	0.86±0.009	0.85±0.011	0.82±0.011	0.183±0.030	0.163±0.026	0.029±0.010

Continued on next page

	Kinase	5%	10%	20%	5%	10%	20%
							Continued from previous page
	CDK3	0.76±0.050	0.76±0.050	0.66±0.059	0.357±0.045	0.357±0.045	0.000±0.036
	Average	0.84±0.014	0.82±0.014	0.81±0.015	0.21±0.026	0.157±0.027	0.09±0.02
AGC	PKACA	0.89±0.003	0.89±0.003	0.89±0.003	0.120±0.008	0.126±0.007	0.126±0.007
	PKCA	0.84±0.001	0.83±0.002	0.83±0.002	0.133±0.009	0.129±0.006	0.109±0.008
	Akt1	0.92±0.004	0.91±0.004	0.91±0.005	0.181±0.017	0.169±0.014	0.167±0.008
	PKCD	0.70±0.009	0.69±0.009	0.68±0.010	0.043±0.006	0.026±0.006	0.027±0.005
	PKG1	0.86±0.027	0.86±0.026	0.87±0.026	0.203±0.020	0.226±0.014	0.201±0.021
	p90RSK	0.80±0.010	0.77±0.012	0.74±0.015	0.173±0.037	0.024±0.021	0.035±0.013
	PKCE	0.67±0.017	0.65±0.020	0.64±0.020	0.100±0.006	0.098±0.015	0.092±0.022
	PKCZ	0.63±0.020	0.59±0.027	0.56±0.029	0.143±0.029	0.014±0.011	0.015±0.014
	PKCB	0.71±0.019	0.67±0.022	0.65±0.023	0.127±0.028	0.110±0.019	0.136±0.020
	RSK2	0.71±0.023	0.72±0.023	0.69±0.028	0.124±0.017	0.095±0.022	0.069±0.016
	ROCK1	0.76±0.012	0.75±0.011	0.74±0.010	0.146±0.032	0.110±0.025	0.136±0.019
	PDK1	0.84±0.018	0.84±0.018	0.85±0.019	0.499±0.024	0.450±0.015	0.414±0.011
	PKCT	0.77±0.041	0.78±0.030	0.80±0.026	0.125±0.047	0.070±0.045	0.089±0.044
	PKCG	0.65±0.024	0.62±0.026	0.63±0.026	0.108±0.064	0.037±0.013	0.027±0.013
	p70S6K	0.83±0.010	0.82±0.013	0.80±0.014	0.284±0.029	0.155±0.026	0.114±0.016
	SGK1	0.83±0.018	0.82±0.017	0.83±0.022	0.328±0.011	0.270±0.030	0.258±0.025
	Akt2	0.87±0.012	0.89±0.018	0.87±0.020	0.159±0.020	0.169±0.026	0.101±0.034
	GRK2	0.86±0.014	0.84±0.014	0.77±0.017	0.529±0.033	0.371±0.028	0.144±0.015
	ROCK2	0.77±0.015	0.69±0.033	0.76±0.020	0.171±0.002	0.175±0.003	0.140±0.011
	PKCI	0.81±0.023	0.73±0.043	0.78±0.027	0.160±0.049	0.198±0.066	0.227±0.055
	PKCH	0.90±0.026	0.85±0.028	0.83±0.037	0.561±0.038	0.345±0.051	0.327±0.065
	PKN1	0.79±0.058	0.79±0.058	0.65±0.095	0.202±0.108	0.202±0.108	0.150±0.103
	Average	0.79±0.018	0.77±0.021	0.76±0.022	0.21±0.029	0.162±0.026	0.141±0.025
TK	Src	0.56±0.006	0.57±0.007	0.55±0.005	0.102±0.005	0.081±0.007	0.084±0.007
	Abl	0.62±0.009	0.60±0.011	0.60±0.012	0.149±0.016	0.124±0.010	0.108±0.013
	Fyn	0.59±0.009	0.57±0.011	0.56±0.012	0.121±0.009	0.067±0.014	0.084±0.010
	Lck	0.53±0.012	0.54±0.011	0.54±0.013	0.063±0.016	0.050±0.014	0.062±0.015
	Lyn	0.48±0.016	0.48±0.016	0.47±0.017	0.048±0.012	0.053±0.011	0.061±0.014
	EGFR	0.56±0.023	0.53±0.022	0.54±0.021	0.050±0.018	0.024±0.010	0.054±0.016
	Syk	0.81±0.018	0.82±0.016	0.80±0.015	0.266±0.025	0.308±0.024	0.290±0.019
	InsR	0.69±0.026	0.67±0.029	0.67±0.028	0.352±0.025	0.177±0.017	0.156±0.022
	JAK2	0.58±0.028	0.52±0.029	0.52±0.033	0.155±0.030	0.107±0.025	0.072±0.025
	FAK	0.67±0.050	0.50±0.033	0.40±0.017	0.360±0.067	0.071±0.039	0.041±0.014
	Ret	0.54±0.023	0.52±0.018	0.52±0.015	0.193±0.025	0.166±0.020	0.166±0.021
	Arg	0.67±0.036	0.53±0.041	0.66±0.034	0.154±0.017	0.070±0.040	0.193±0.030
	Brk	0.60±0.021	0.53±0.034	0.49±0.032	0.197±0.007	0.079±0.044	0.066±0.018

Continued on next page

	kinase	5%	10%	20%	5%	10%	20%
							Continued from previous page
	ALK	0.57±0.032	0.57±0.032	0.50±0.031	0.000±0.000	0.000±0.000	0.000±0.000
	Btk	0.71±0.033	0.70±0.028	0.70±0.031	0.311±0.053	0.205±0.047	0.152±0.043
	PDGFRB	0.61±0.033	0.60±0.019	0.51±0.017	0.255±0.040	0.143±0.033	0.047±0.019
	JAK3	0.81±0.032	0.72±0.046	0.72±0.056	0.398±0.063	0.158±0.054	0.161±0.051
	Hck	0.58±0.025	0.51±0.032	0.50±0.029	0.089±0.017	0.063±0.017	0.057±0.017
	Pyk2	0.62±0.033	0.62±0.033	0.45±0.076	0.173±0.019	0.173±0.019	0.000±0.000
	Average	0.21±0.0246	0.59±0.025	0.56± 0.026	0.181±0.024	0.112±0.023	0.098±0.019
CAMK	CAMK2A	0.68±0.011	0.67±0.011	0.64±0.011	0.119±0.012	0.093±0.014	0.084±0.014
CAMK	Chk1	0.71±0.017	0.70±0.020	0.69±0.022	0.062±0.022	0.055±0.014	0.060±0.019
CAMK	AMPKA1	0.72±0.016	0.74±0.018	0.75±0.018	0.079±0.014	0.087±0.012	0.094±0.013
CAMK	MAPKAPK2	0.78±0.019	0.79±0.014	0.80±0.016	0.141±0.028	0.089±0.015	0.076±0.021
CAMK	PKD1	0.76±0.010	0.75±0.010	0.74±0.012	0.088±0.012	0.089±0.016	0.063±0.016
CAMK	LKB1	0.81±0.009	0.80±0.011	0.79±0.015	0.579±0.018	0.497±0.005	0.486±0.010
CAMK	MSK1	0.86±0.032	0.83±0.061	0.79±0.048	0.333±0.076	0.259±0.076	0.109±0.050
CAMK	Chk2	0.62±0.020	0.61±0.023	0.59±0.021	0.027±0.010	0.018±0.008	0.017±0.007
CAMK	Pim1	0.84±0.025	0.84±0.029	0.74±0.026	0.353±0.031	0.249±0.054	0.042±0.033
CAMK	AMPKA2	0.86±0.028	0.82±0.028	0.81±0.033	0.116±0.037	0.051±0.018	0.057±0.021
CAMK	MARK2	0.80±0.024	0.73±0.042	0.75±0.030	0.245±0.002	0.267±0.022	0.237±0.047
CAMK	CAMK1A	0.83±0.016	0.83±0.016	0.82±0.019	0.423±0.065	0.423±0.065	0.345±0.062
CAMK	DAPK3	0.67±0.035	0.55±0.054	0.49±0.038	0.194±0.065	0.000±0.016	0.000±0.013
CAMK	CaMK4	0.79±0.032	0.79±0.032	0.71±0.085	0.000±0.000	0.000±0.000	0.000±0.000
CAMK	PKD2	0.80±0.054	0.80±0.054	0.81±0.108	0.075±0.040	0.075±0.040	0.016±0.017
CAMK	CAMK2D	0.83±0.041	0.83±0.041	0.81±0.095	0.250±0.000	0.250±0.000	0.176±0.036
	Average	0.77±0.024	0.75±0.029	0.73±0.037	0.193±0.027	0.156±0.023	0.117±0.024
Other	CK2A1	0.93±0.001	0.93±0.001	0.93±0.001	0.386±0.004	0.374±0.004	0.374±0.004
Other	PLK1	0.78±0.007	0.76±0.009	0.73±0.010	0.121±0.016	0.102±0.014	0.091±0.010
Other	AurB	0.79±0.010	0.78±0.009	0.77±0.010	0.086±0.010	0.077±0.018	0.035±0.005
Other	AurA	0.74±0.012	0.74±0.016	0.74±0.015	0.101±0.012	0.038±0.018	0.015±0.012
Other	PLK3	0.66±0.039	0.61±0.032	0.61±0.020	0.212±0.039	0.040±0.014	0.000±0.000
Other	IKKA	0.69±0.013	0.67±0.015	0.62±0.011	0.241±0.046	0.077±0.028	0.029±0.009
Other	IKKB	0.75±0.021	0.68±0.016	0.63±0.016	0.374±0.022	0.176±0.016	0.123±0.017
Other	TBK1	0.76±0.032	0.73±0.026	0.68±0.027	0.296±0.041	0.218±0.036	0.098±0.030
Other	CK2A2	0.91±0.036	0.85±0.022	0.82±0.020	0.441±0.063	0.188±0.057	0.021±0.015
Other	IKKE	0.96±0.011	0.95±0.015	0.90±0.024	0.690±0.088	0.408±0.043	0.203±0.048
Other	TTK	0.82±0.036	0.66±0.033	0.65±0.037	0.355±0.057	0.049±0.012	0.067±0.020
Other	NEK6	0.78±0.021	0.78±0.021	0.76±0.026	0.309±0.035	0.309±0.035	0.160±0.050
Other	NEK2	0.76±0.041	0.68±0.064	0.69±0.036	0.493±0.064	0.386±0.052	0.283±0.093
	Average	0.80±0.021	0.76±0.022	0.73±0.02	0.32±0.038	0.19±0.027	0.12±0.024

Continued on next page

	kinase	5%	10%	20%	5%	10%	20%
							Continued from previous page
STE	PAK1	0.70±0.013	0.66±0.018	0.65±0.020	0.038±0.009	0.005±0.003	0.011±0.006
	Cot	0.84±0.020	0.80±0.018	0.80±0.026	0.502±0.086	0.462±0.077	0.459±0.088
	MST1	0.75±0.042	0.69±0.032	0.65±0.041	0.204±0.028	0.055±0.022	0.000±0.000
	ASK1	0.82±0.021	0.70±0.028	0.69±0.035	0.392±0.061	0.142±0.059	0.135±0.055
	MKK4	0.90±0.038	0.79±0.014	0.79±0.018	0.642±0.029	0.534±0.009	0.544±0.009
	MST2	0.72±0.052	0.66±0.072	0.64±0.073	0.192±0.047	0.124±0.038	0.121±0.037
	PAK2	0.73±0.074	0.53±0.069	0.45±0.048	0.360±0.078	0.087±0.049	0.000±0.000
	MKK7	0.96±0.084	0.96±0.084	0.84±0.051	0.799±0.054	0.807±0.057	0.629±0.006
	MEK1	0.72±0.050	0.72±0.050	0.66±0.041	0.466±0.009	0.468±0.007	0.478±0.009
	Average	0.79±0.044	0.73±0.043	0.69±0.039	0.40±0.044	0.30±0.036	0.26±0.023
CK1	CK1A	0.78±0.009	0.75±0.009	0.73±0.013	0.195±0.011	0.097±0.018	0.085±0.016
	CK1D	0.90±0.006	0.88±0.008	0.87±0.009	0.232±0.029	0.131±0.023	0.045±0.018
	CK1E	0.87±0.018	0.82±0.026	0.76±0.018	0.415±0.059	0.188±0.050	0.023±0.020
	VRK1	0.87±0.068	0.83±0.075	0.65±0.045	0.348±0.027	0.353±0.030	0.346±0.045
	Average	0.86±0.025	0.82±0.029	0.75±0.021	0.30±0.03	0.19±0.03	0.12±0.025
Atypical	ATM	0.95±0.002	0.95±0.002	0.95±0.002	0.277±0.017	0.267±0.011	0.308±0.015
	ATR	0.86±0.009	0.86±0.008	0.85±0.012	0.114±0.014	0.102±0.009	0.114±0.009
	DNAPK	0.86±0.005	0.86±0.004	0.85±0.005	0.170±0.012	0.161±0.010	0.147±0.011
	mTOR	0.81±0.017	0.77±0.014	0.77±0.016	0.220±0.040	0.091±0.018	0.077±0.019
	Average	0.87±0.008	0.86±0.007	0.85±0.009	0.195±0.021	0.155±0.012	0.162±0.014

TABLE B.2: Sequence model accuracy across **mouse** kinases when different percentages of kinase phosphorylation peptides were used to determine the set of k-mers added to the sequence model. Table shows median AUC and AUC50 values for classifying kinase phosphorylation sites with the sequence model as determined by 10-fold cross-validation across 10 randomised data-set splits. Kinases are grouped according to their family, with the average prediction accuracy for each family shown.

		AUC			AUC50		
	Kinase	5%	10%	20%	5%	10%	20%
CMGC	ERK2	0.83±0.006	0.83±0.006	0.83±0.005	0.194±0.017	0.220±0.016	0.241±0.017
	ERK1	0.82±0.010	0.80±0.011	0.80±0.012	0.164±0.021	0.131±0.013	0.118±0.015
	CDK5	0.80±0.013	0.78±0.013	0.76±0.014	0.167±0.016	0.145±0.013	0.093±0.010
	CDK1	0.79±0.013	0.77±0.013	0.78±0.015	0.184±0.030	0.160±0.026	0.138±0.021
	JNK1	0.78±0.014	0.76±0.014	0.76±0.017	0.219±0.040	0.169±0.027	0.173±0.025
	P38A	0.74±0.017	0.72±0.015	0.69±0.018	0.226±0.028	0.202±0.031	0.117±0.021
	CDK2	0.74±0.034	0.69±0.033	0.68±0.023	0.340±0.033	0.154±0.042	0.075±0.014
	GSK3B	0.83±0.021	0.77±0.020	0.71±0.021	0.414±0.049	0.152±0.035	0.108±0.020

Continued on next page

	Kinase	5%	10%	20%	5%	10%	20%
							Continued from previous page
	Average	0.79±0.016	0.77±0.016	0.75±0.016	0.239±0.029	0.167±0.025	0.133±0.018
AGC	PKACA	0.81±0.007	0.79±0.006	0.79±0.006	0.245±0.014	0.242±0.015	0.251±0.009
AGC	PKCA	0.72±0.010	0.70±0.013	0.69±0.012	0.253±0.016	0.198±0.018	0.192±0.013
AGC	Akt1	0.81±0.011	0.82±0.011	0.81±0.010	0.383±0.047	0.413±0.052	0.348±0.060
AGC	PKCD	0.75±0.028	0.64±0.051	0.68±0.029	0.113±0.037	0.068±0.025	0.080±0.021
AGC	p90RSK	0.87±0.013	0.81±0.020	0.90±0.009	0.216±0.037	0.175±0.044	0.371±0.041
AGC	RSK2	0.79±0.042	0.79±0.042	0.68±0.087	0.283±0.085	0.283±0.085	0.280±0.084
AGC	PKG1	0.66±0.042	0.66±0.042	0.36±0.043	0.000±0.000	0.000±0.000	0.000±0.000
AGC	p70S6K	0.88±0.029	0.88±0.029	0.76±0.045	0.394±0.062	0.394±0.062	0.326±0.078
AGC	PKCZ	0.69±0.095	0.69±0.095	0.47±0.108	0.286±0.114	0.286±0.114	0.071±0.110
AGC	PKCE	0.54±0.043	0.54±0.043	0.47±0.033	0.444±0.102	0.444±0.102	0.000±0.066
	Average	0.75±0.032	0.73±0.035	0.66±0.038	0.262±0.052	0.25±0.052	0.192±0.048
TK	Src	0.61±0.012	0.55±0.016	0.54±0.018	0.267±0.013	0.191±0.020	0.178±0.016
TK	Fyn	0.64±0.018	0.63±0.013	0.66±0.015	0.307±0.027	0.253±0.037	0.335±0.038
TK	Abl	0.52±0.042	0.42±0.040	0.49±0.039	0.151±0.008	0.135±0.013	0.111±0.026
TK	Lyn	0.65±0.027	0.66±0.028	0.65±0.028	0.286±0.029	0.298±0.026	0.247±0.025
TK	Lck	0.64±0.060	0.53±0.072	0.64±0.050	0.271±0.056	0.177±0.066	0.265±0.070
TK	Syk	0.71±0.014	0.60±0.029	0.61±0.022	0.601±0.024	0.299±0.073	0.336±0.026
	Average	0.627±0.029	0.56±0.033	0.60±0.029	0.314±0.026	0.226±0.039	0.245±0.033

TABLE B.3: Sequence model accuracy across **yeast** kinases when different percentages of kinase phosphorylation peptides were used to determine the set of k-mers added to the sequence model. Table shows median AUC and AUC50 values for classifying kinase phosphorylation sites with the sequence model as determined by 10-fold cross-validation across 10 randomised data-set splits. Kinases are grouped according to their family, with the average prediction accuracy for each family shown.

		AUC			AUC50		
	Kinase	5%	10%	20%	5%	10%	20%
CMGC	CDC28	0.93±0.001	0.93±0.001	0.93±0.001	0.295±0.012	0.297±0.013	0.346±0.009
CMGC	CTK1	0.70±0.008	0.69±0.008	0.69±0.007	0.434±0.000	0.432±0.001	0.432±0.000
CMGC	MCK1	0.83±0.009	0.80±0.011	0.74±0.016	0.348±0.024	0.230±0.022	0.127±0.025
CMGC	PHO85	0.71±0.018	0.64±0.014	0.61±0.013	0.172±0.010	0.113±0.023	0.043±0.013
CMGC	SSN3	0.74±0.057	0.67±0.051	0.63±0.041	0.295±0.064	0.027±0.016	0.000±0.000
CMGC	HOG1	0.79±0.047	0.73±0.040	0.67±0.040	0.301±0.052	0.099±0.031	0.080±0.028
CMGC	KNS1	0.93±0.038	0.83±0.032	0.69±0.065	0.591±0.056	0.333±0.085	0.083±0.055
CMGC	SLT2	0.68±0.037	0.58±0.062	0.40±0.040	0.271±0.048	0.215±0.060	0.000±0.032
CMGC	FUS3	0.54±0.035	0.54±0.035	0.53±0.052	0.217±0.004	0.217±0.004	0.048±0.040

Continued on next page

	Kinase	5%	10%	20%	5%	10%	20%
							Continued from previous page
	Average	0.76±0.028	0.71±0.028	0.65±0.03	0.325±0.03	0.218±0.028	0.129±0.022
AGC	TPK1	0.95±0.003	0.95±0.003	0.95±0.003	0.383±0.011	0.336±0.017	0.391±0.010
	TPK3	0.81±0.036	0.76±0.033	0.71±0.039	0.595±0.058	0.426±0.040	0.359±0.048
	YPK1	0.74±0.043	0.68±0.061	0.62±0.046	0.443±0.087	0.327±0.073	0.167±0.078
	PKH2	0.75±0.037	0.75±0.037	0.72±0.100	0.250±0.003	0.250±0.003	0.040±0.048
	PKH1	0.98±0.006	0.98±0.006	0.88±0.026	0.750±0.000	0.750±0.000	0.500±0.037
	PKC1	0.88±0.024	0.84±0.052	0.85±0.037	0.346±0.045	0.338±0.067	0.228±0.088
	Average	0.85±0.025	0.83±0.032	0.79±0.042	0.461±0.034	0.405±0.033	0.281±0.051
CAMK	SNF1	0.78±0.014	0.71±0.014	0.66±0.015	0.162±0.032	0.023±0.009	0.022±0.010
	FRK1	0.75±0.021	0.70±0.043	0.60±0.047	0.424±0.048	0.367±0.087	0.019±0.015
	PSK2	0.74±0.047	0.58±0.026	0.51±0.029	0.413±0.055	0.016±0.013	0.004±0.014
	DUN1	0.85±0.013	0.83±0.018	0.79±0.023	0.379±0.012	0.256±0.015	0.182±0.050
	Average	0.78±0.024	0.71±0.026	0.64±0.029	0.345±0.037	0.167±0.031	0.057±0.023
Other	CKA1	0.89±0.005	0.89±0.006	0.88±0.006	0.313±0.015	0.294±0.017	0.212±0.010
	CKA2	0.91±0.007	0.91±0.007	0.90±0.007	0.355±0.017	0.314±0.011	0.251±0.013
	MPS1	0.86±0.016	0.84±0.014	0.83±0.015	0.231±0.036	0.142±0.025	0.111±0.017
	PTK1	0.67±0.015	0.64±0.025	0.56±0.024	0.139±0.020	0.047±0.010	0.029±0.010
	PTK2	0.89±0.046	0.76±0.037	0.64±0.024	0.755±0.065	0.263±0.043	0.000±0.011
	IPL1	0.91±0.009	0.91±0.008	0.92±0.012	0.276±0.018	0.298±0.028	0.236±0.020
	BUD32	0.73±0.063	0.70±0.072	0.49±0.052	0.385±0.071	0.335±0.064	0.000±0.000
	Average	0.84±0.023	0.81±0.024	0.74±0.02	0.351±0.035	0.242±0.028	0.12±0.012

TABLE B.4: Sequence model accuracy for varying window sizes in **human** kinases, where kinases are grouped according to family. Table shows accuracy values for classifying kinase phosphorylation sites with the sequence model as determined by 10-fold cross-validation across 10 randomised data-set splits. Prediction accuracy is shown using median and standard deviation of the AUC and AUC50 across the data-set splits. Varying window sizes were applied to determine the optimal window size on a kinase-specific basis. The window size determined for a kinase is highlighted through bold text. Optimal window size was determined primarily through AUC50 as a measure of the model's accuracy at low false-positive rates. If accuracy did not increase through increasing window size, the lower window size was chosen.

		AUC					AUC50				
	Kinase	7	9	11	13	15	7	9	11	13	15
CMGC	CDK2	0.89±0.001	0.89±0.001	0.89±0.001	**0.89±0.001**	0.89±0.001	0.066±0.008	0.073±0.003	0.088±0.006	**0.100±0.004**	0.100±0.006
	CDK1	**0.89±0.002**	0.89±0.001	0.89±0.001	0.89±0.001	0.89±0.002	**0.071±0.008**	0.053±0.008	0.055±0.005	0.045±0.004	0.059±0.007
	ERK2	0.87±0.001	0.87±0.001	0.87±0.002	0.86±0.002	**0.86±0.001**	0.048±0.006	0.040±0.003	0.046±0.007	0.044±0.005	**0.067±0.010**
	ERK1	0.87±0.003	0.86±0.004	0.86±0.004	0.86±0.004	**0.86±0.005**	0.034±0.011	0.045±0.009	0.042±0.012	0.059±0.014	**0.066±0.012**
	GSK3B	0.72±0.006	0.80±0.005	0.81±0.007	**0.81±0.006**	0.81±0.008	0.031±0.004	0.096±0.008	0.127±0.014	**0.132±0.014**	0.117±0.013
	P38A	0.83±0.005	0.83±0.005	0.82±0.004	0.81±0.006	**0.81±0.007**	0.091±0.020	0.142±0.017	0.145±0.016	0.135±0.015	**0.151±0.017**
	JNK1	0.87±0.004	0.87±0.005	0.86±0.004	0.85±0.005	**0.87±0.004**	0.092±0.018	0.123±0.021	0.134±0.015	0.118±0.012	**0.155±0.014**
	CDK5	0.85±0.007	0.85±0.009	**0.84±0.009**	0.84±0.008	0.84±0.007	0.016±0.007	0.037±0.007	**0.050±0.007**	0.027±0.006	0.026±0.010
	JNK2	0.79±0.009	0.77±0.012	0.72±0.016	0.69±0.022	**0.73±0.023**	0.045±0.012	0.049±0.014	0.051±0.013	0.048±0.013	**0.068±0.015**
	CDK7	0.70±0.031	0.76±0.024	0.84±0.022	0.89±0.018	**0.88±0.019**	0.094±0.039	0.254±0.067	0.326±0.052	0.307±0.040	**0.310±0.032**
	GSK3A	0.85±0.023	0.90±0.022	0.89±0.022	0.90±0.028	**0.90±0.026**	0.281±0.032	0.405±0.034	0.446±0.033	0.438±0.031	**0.458±0.045**
	CDK4	0.87±0.008	0.87±0.009	0.88±0.010	0.86±0.012	**0.87±0.012**	0.085±0.015	0.078±0.008	0.098±0.024	0.099±0.015	**0.179±0.025**
	P38B	0.83±0.005	0.86±0.010	0.86±0.008	0.85±0.012	**0.83±0.014**	0.097±0.019	0.168±0.022	0.226±0.034	0.222±0.047	**0.260±0.046**
	HIPK2	0.84±0.011	0.85±0.011	0.86±0.010	0.85±0.010	**0.86±0.013**	0.206±0.029	0.222±0.032	0.245±0.039	0.300±0.039	**0.380±0.043**
	DYRK1A	0.76±0.021	0.78±0.020	0.83±0.026	0.84±0.029	**0.83±0.033**	0.000±0.013	0.107±0.017	0.206±0.028	0.248±0.038	**0.260±0.043**
	CDK9	0.77±0.011	0.79±0.009	0.80±0.013	**0.83±0.015**	0.84±0.015	0.220±0.031	0.275±0.023	0.287±0.039	**0.320±0.030**	0.306±0.039
	DYRK2	0.73±0.023	0.76±0.028	0.79±0.021	0.80±0.019	**0.78±0.019**	0.066±0.015	0.159±0.032	0.242±0.053	0.297±0.050	**0.306±0.043**
	ERK5	0.73±0.027	0.79±0.024	0.82±0.020	0.83±0.017	**0.83±0.016**	0.000±0.000	0.043±0.020	0.257±0.045	0.272±0.038	**0.317±0.034**
	CDK6	0.83±0.012	0.84±0.014	0.83±0.014	0.84±0.014	**0.86±0.009**	0.075±0.018	0.077±0.027	0.093±0.030	0.138±0.029	**0.183±0.030**
	CDK3	0.76±0.039	0.77±0.039	0.73±0.031	0.77±0.051	**0.76±0.050**	0.000±0.000	0.065±0.005	0.152±0.035	0.235±0.003	**0.357±0.045**
	PKACA	**0.89±0.003**	0.89±0.003	0.89±0.003	0.89±0.003	0.89±0.003	**0.112±0.007**	0.112±0.008	0.120±0.008	0.115±0.009	0.111±0.006
	PKCA	0.81±0.004	0.83±0.003	0.83±0.003	**0.84±0.001**	0.84±0.001	0.118±0.006	0.120±0.005	0.107±0.009	**0.133±0.009**	0.123±0.009
	Akt1	0.88±0.003	0.87±0.003	0.92±0.004	**0.92±0.004**	0.92±0.003	0.071±0.012	0.077±0.008	0.170±0.014	**0.181±0.017**	0.186±0.013
	PKCD	0.69±0.007	0.70±0.004	0.71±0.006	**0.70±0.009**	0.69±0.008	0.032±0.011	0.039±0.007	0.038±0.009	**0.043±0.006**	0.034±0.008
	PKG1	0.84±0.020	**0.86±0.027**	0.86±0.027	0.84±0.026	0.83±0.027	0.202±0.023	**0.203±0.020**	0.208±0.022	0.209±0.020	0.216±0.023
	p90RSK	0.83±0.016	0.81±0.016	0.81±0.014	0.81±0.011	**0.80±0.010**	0.065±0.010	0.073±0.015	0.131±0.031	0.161±0.037	**0.173±0.037**
	PKCE	0.68±0.015	0.65±0.014	0.65±0.015	0.63±0.015	**0.67±0.017**	0.085±0.013	0.097±0.013	0.096±0.004	0.096±0.000	**0.101±0.006**

Continued on next page

	Kinase	7	9	11	13	15	7	9	11	13	15
										Continued from previous page	
AGC	PKCZ	0.57±0.016	0.61±0.020	0.62±0.021	**0.63±0.020**	0.61±0.022	0.021±0.011	0.067±0.026	0.098±0.029	**0.143±0.029**	0.138±0.026
	PKCB	0.72±0.022	**0.71±0.019**	0.68±0.016	0.70±0.017	0.73±0.016	0.099±0.025	**0.127±0.028**	0.099±0.025	0.122±0.029	0.116±0.023
	RSK2	0.70±0.024	0.66±0.022	0.68±0.020	**0.71±0.023**	0.67±0.025	0.071±0.025	0.044±0.019	0.084±0.026	**0.124±0.017**	0.140±0.025
	ROCK1	0.79±0.008	0.77±0.005	0.78±0.007	**0.76±0.012**	0.75±0.014	0.127±0.030	0.109±0.039	0.133±0.029	**0.146±0.032**	0.155±0.027
	PDK1	**0.84±0.018**	0.81±0.012	0.78±0.011	0.78±0.011	0.78±0.012	**0.499±0.024**	0.476±0.014	0.472±0.017	0.465±0.017	0.461±0.016
	PKCT	**0.77±0.041**	0.71±0.035	0.70±0.039	0.67±0.047	0.62±0.050	**0.125±0.047**	0.124±0.040	0.124±0.039	0.124±0.036	0.124±0.037
	PKCG	0.61±0.022	0.63±0.022	0.65±0.029	0.64±0.029	**0.65±0.024**	0.000±0.025	0.004±0.050	0.035±0.059	0.067±0.054	**0.108±0.064**
	p70S6K	0.79±0.015	0.79±0.014	0.82±0.008	**0.83±0.010**	0.82±0.009	0.037±0.010	0.117±0.023	0.228±0.027	**0.284±0.029**	0.271±0.024
	SGK1	**0.83±0.018**	0.78±0.019	0.84±0.016	0.81±0.016	0.81±0.017	**0.328±0.011**	0.324±0.005	0.299±0.005	0.292±0.005	0.295±0.002
	Akt2	0.84±0.019	0.82±0.012	0.85±0.014	**0.87±0.012**	0.85±0.013	0.162±0.034	0.151±0.026	0.141±0.021	**0.159±0.020**	0.120±0.029
	GRK2	0.80±0.013	0.82±0.013	0.85±0.015	0.86±0.016	**0.86±0.014**	0.301±0.038	0.410±0.036	0.468±0.031	0.510±0.031	**0.529±0.033**
	ROCK2	**0.77±0.015**	0.71±0.018	0.67±0.016	0.65±0.011	0.69±0.009	**0.171±0.002**	0.171±0.002	0.174±0.002	0.173±0.002	0.171±0.002
	PKCI	**0.81±0.023**	0.80±0.021	0.80±0.018	0.82±0.017	0.80±0.018	**0.160±0.049**	0.162±0.048	0.158±0.048	0.158±0.051	0.170±0.047
	PKCH	0.86±0.024	0.86±0.023	0.87±0.022	0.89±0.023	**0.90±0.026**	0.388±0.052	0.378±0.050	0.484±0.049	0.488±0.033	**0.560±0.039**
	PKN1	0.76±0.048	**0.79±0.058**	0.74±0.054	0.69±0.063	0.68±0.057	0.140±0.079	**0.202±0.108**	0.158±0.090	0.202±0.103	0.258±0.130
TK	Src	0.55±0.006	**0.56±0.006**	0.56±0.006	0.57±0.006	0.57±0.008	0.082±0.004	**0.102±0.005**	0.082±0.007	0.096±0.006	0.087±0.006
	Abl	0.62±0.012	**0.62±0.009**	0.61±0.008	0.62±0.011	0.63±0.011	0.132±0.014	**0.149±0.016**	0.132±0.015	0.134±0.012	0.142±0.012
	Fyn	**0.59±0.009**	0.60±0.013	0.59±0.016	0.59±0.017	0.60±0.018	**0.121±0.009**	0.108±0.007	0.114±0.010	0.108±0.014	0.116±0.019
	Lck	0.54±0.012	0.55±0.012	**0.53±0.012**	0.54±0.017	0.56±0.016	0.044±0.009	0.032±0.009	**0.063±0.016**	0.042±0.015	0.039±0.014
	Lyn	0.45±0.010	0.46±0.016	0.45±0.019	0.46±0.019	**0.48±0.016**	0.000±0.002	0.027±0.009	0.027±0.010	0.041±0.010	**0.048±0.012**
	EGFR	0.51±0.017	0.50±0.019	0.51±0.024	**0.56±0.023**	0.54±0.026	0.022±0.009	0.032±0.012	0.036±0.012	**0.050±0.018**	0.030±0.013
	Syk	0.73±0.016	0.74±0.015	0.77±0.020	0.79±0.018	**0.81±0.018**	0.174±0.019	0.178±0.023	0.216±0.024	0.235±0.026	**0.266±0.025**
	InsR	0.68±0.024	**0.69±0.026**	0.64±0.020	0.63±0.014	0.64±0.016	0.229±0.014	**0.351±0.025**	0.349±0.022	0.346±0.020	0.340±0.017
	JAK2	0.52±0.014	0.53±0.021	0.52±0.021	0.56±0.024	**0.58±0.028**	0.086±0.019	0.153±0.033	0.140±0.027	0.135±0.024	**0.156±0.030**
	FAK	0.58±0.056	0.69±0.046	0.65±0.049	0.67±0.045	**0.67±0.050**	0.206±0.039	0.286±0.054	0.316±0.056	0.307±0.063	**0.360±0.067**
	Ret	0.41±0.024	0.44±0.022	0.46±0.019	0.49±0.016	**0.54±0.023**	0.149±0.027	0.159±0.031	0.162±0.031	0.195±0.035	**0.192±0.024**
	Arg	0.57±0.027	0.57±0.041	0.52±0.036	0.63±0.037	**0.67±0.036**	0.107±0.008	0.046±0.017	0.036±0.020	0.122±0.021	**0.154±0.017**
	Brk	0.57±0.019	0.57±0.014	0.56±0.020	0.53±0.021	**0.60±0.021**	0.204±0.017	0.192±0.011	0.198±0.005	0.194±0.003	**0.197±0.007**
	ALK	0.40±0.029	**0.57±0.032**	0.54±0.030	0.46±0.027	0.45±0.024	0.000±0.000	**0.000±0.000**	0.000±0.000	0.000±0.000	0.000±0.006
	Btk	0.68±0.034	**0.71±0.033**	0.67±0.045	0.67±0.044	0.65±0.038	0.315±0.055	**0.311±0.053**	0.320±0.057	0.307±0.053	0.297±0.058
	PDGFRB	0.64±0.031	0.61±0.034	0.62±0.032	0.64±0.031	**0.61±0.033**	0.165±0.013	0.162±0.031	0.154±0.031	0.206±0.036	**0.255±0.040**
	JAK3	0.71±0.028	0.78±0.030	0.80±0.032	**0.81±0.032**	0.78±0.032	0.259±0.041	0.362±0.045	0.381±0.063	**0.398±0.063**	0.369±0.063
	Hck	0.55±0.027	0.51±0.023	0.49±0.025	0.49±0.032	**0.58±0.025**	0.086±0.013	0.078±0.015	0.072±0.017	0.041±0.020	**0.089±0.017**
	Pyk2	0.55±0.041	**0.62±0.033**	0.56±0.025	0.58±0.026	0.44±0.026	0.000±0.000	**0.173±0.019**	0.157±0.023	0.011±0.017	0.000±0.000

Continued on next page

	Kinase	7	9	11	13	15	7	9	11	13	15
											Continued from previous page
CAMK	CAMK2A	0.73±0.011	0.74±0.010	0.72±0.009	**0.68±0.011**	0.69±0.012	0.069±0.015	0.056±0.006	0.100±0.013	**0.119±0.012**	0.112±0.013
	Chk1	0.69±0.023	0.67±0.030	0.68±0.023	0.68±0.022	**0.71±0.017**	0.048±0.012	0.046±0.011	0.055±0.013	0.058±0.020	**0.062±0.022**
	AMPKA1	0.65±0.021	0.68±0.021	**0.72±0.016**	0.70±0.013	0.73±0.015	0.053±0.017	0.065±0.018	**0.079±0.014**	0.070±0.016	0.076±0.012
	MAPKAPK2	0.81±0.020	0.79±0.020	**0.78±0.019**	0.78±0.021	0.77±0.023	0.080±0.026	0.082±0.027	**0.141±0.028**	0.132±0.020	0.121±0.017
	PKD1	0.70±0.018	0.71±0.018	**0.76±0.010**	0.73±0.011	0.70±0.011	0.016±0.010	0.021±0.012	**0.088±0.012**	0.087±0.017	0.085±0.021
	LKB1	0.82±0.008	0.81±0.008	0.81±0.009	0.82±0.010	**0.81±0.009**	0.504±0.022	0.532±0.015	0.561±0.018	0.569±0.017	**0.579±0.017**
	MSK1	0.77±0.033	0.80±0.028	0.83±0.027	0.85±0.031	**0.86±0.032**	0.187±0.046	0.193±0.072	0.238±0.077	0.313±0.082	**0.333±0.076**
	Chk2	0.56±0.022	0.59±0.027	0.61±0.023	0.59±0.020	**0.62±0.020**	0.000±0.000	0.009±0.007	0.018±0.009	0.017±0.007	**0.027±0.010**
	Pim1	0.69±0.021	0.75±0.031	0.85±0.025	0.84±0.024	**0.84±0.025**	0.180±0.055	0.277±0.046	0.324±0.045	0.338±0.035	**0.352±0.032**
	AMPKA2	0.75±0.026	0.79±0.031	0.84±0.028	**0.86±0.028**	0.85±0.029	0.004±0.006	0.038±0.016	0.051±0.017	**0.116±0.037**	0.118±0.040
	MARK2	0.75±0.026	**0.80±0.024**	0.76±0.022	0.76±0.032	0.74±0.031	0.243±0.008	**0.245±0.002**	0.247±0.019	0.245±0.004	0.247±0.013
	CAMK1A	0.82±0.021	0.81±0.016	**0.83±0.016**	0.76±0.018	0.74±0.018	0.397±0.067	0.396±0.067	**0.423±0.064**	0.426±0.044	0.425±0.064
	DAPK3	0.44±0.033	0.60±0.059	0.68±0.038	0.66±0.038	**0.67±0.035**	0.000±0.000	0.005±0.012	0.068±0.034	0.089±0.054	**0.194±0.065**
	CaMK4	0.78±0.018	**0.79±0.032**	0.74±0.028	0.70±0.027	0.65±0.036	0.000±0.000	**0.000±0.000**	0.000±0.000	0.000±0.000	0.000±0.000
	PKD2	0.83±0.040	0.75±0.045	0.83±0.037	0.79±0.039	**0.80±0.054**	0.000±0.000	0.000±0.003	0.029±0.016	0.051±0.026	**0.075±0.040**
	CAMK2D	0.73±0.031	0.70±0.033	0.71±0.038	0.80±0.034	**0.83±0.041**	0.144±0.019	0.240±0.004	0.245±0.002	0.250±0.000	**0.250±0.000**
Other	CK2A1	0.92±0.001	0.93±0.001	0.93±0.001	**0.93±0.001**	0.93±0.001	0.316±0.004	0.356±0.003	0.374±0.004	**0.386±0.004**	0.370±0.004
	PLK1	0.78±0.012	0.76±0.012	0.78±0.010	0.78±0.008	**0.78±0.007**	0.098±0.010	0.093±0.012	0.077±0.016	0.087±0.014	**0.121±0.016**
	AurB	**0.79±0.010**	0.77±0.008	0.76±0.009	0.77±0.011	0.77±0.010	**0.086±0.010**	0.075±0.010	0.067±0.011	0.084±0.011	0.073±0.008
	AurA	**0.74±0.012**	0.74±0.010	0.73±0.011	0.75±0.011	0.72±0.014	**0.101±0.012**	0.093±0.013	0.082±0.019	0.079±0.017	0.070±0.015
	PLK3	0.65±0.032	0.64±0.037	0.64±0.041	0.62±0.039	**0.66±0.039**	0.031±0.020	0.020±0.025	0.066±0.023	0.140±0.031	**0.212±0.039**
	IKKA	0.68±0.019	0.64±0.014	0.64±0.011	0.66±0.012	**0.69±0.013**	0.040±0.012	0.060±0.016	0.110±0.024	0.131±0.028	**0.241±0.046**
	IKKB	0.62±0.013	0.68±0.013	0.73±0.010	0.76±0.018	**0.75±0.021**	0.026±0.006	0.135±0.014	0.243±0.030	0.313±0.026	**0.374±0.022**
	TBK1	0.73±0.029	0.76±0.032	0.77±0.030	**0.76±0.032**	0.74±0.032	0.162±0.018	0.182±0.023	0.269±0.031	**0.296±0.041**	0.298±0.041
	CK2A2	0.88±0.021	0.86±0.019	0.86±0.024	0.89±0.026	**0.91±0.036**	0.241±0.040	0.391±0.069	0.389±0.061	0.426±0.059	**0.441±0.063**
	IKKE	0.96±0.015	0.97±0.012	0.96±0.010	0.96±0.012	**0.96±0.011**	0.206±0.027	0.489±0.080	0.663±0.087	0.669±0.090	**0.690±0.088**
	TTK	0.61±0.025	0.72±0.026	0.82±0.031	**0.82±0.036**	0.81±0.047	0.045±0.019	0.098±0.025	0.266±0.026	**0.355±0.057**	0.351±0.052
	NEK6	0.84±0.016	0.80±0.015	0.79±0.020	0.82±0.015	**0.78±0.021**	0.095±0.032	0.190±0.057	0.173±0.053	0.230±0.056	**0.309±0.035**
	NEK2	0.72±0.032	0.69±0.032	0.66±0.051	0.68±0.045	**0.76±0.041**	0.144±0.022	0.371±0.070	0.356±0.046	0.463±0.054	**0.493±0.064**
STE	PAK1	0.73±0.007	**0.70±0.013**	0.66±0.014	0.70±0.014	0.69±0.012	0.023±0.004	**0.038±0.009**	0.023±0.009	0.037±0.007	0.038±0.008
	Cot	0.82±0.014	0.80±0.017	0.81±0.020	**0.84±0.020**	0.83±0.025	0.496±0.091	0.500±0.088	0.500±0.089	**0.502±0.086**	0.497±0.088
	MST1	0.73±0.028	0.77±0.028	0.76±0.025	0.74±0.040	**0.75±0.042**	0.115±0.001	0.118±0.001	0.161±0.016	0.165±0.018	**0.205±0.028**
	ASK1	0.73±0.018	0.78±0.022	0.79±0.017	0.82±0.020	**0.82±0.021**	0.251±0.055	0.313±0.056	0.362±0.056	0.377±0.059	**0.392±0.061**
	MKK4	0.88±0.030	0.86±0.035	0.89±0.042	**0.90±0.038**	0.87±0.040	0.601±0.004	0.602±0.007	0.618±0.018	**0.646±0.029**	0.652±0.035

Continued on next page

	Kinase	7	9	11	13	15	7	9	11	13	15
										Continued from previous page	
	MST2	0.75±0.055	0.65±0.052	0.65±0.047	0.70±0.052	**0.72±0.052**	0.123±0.035	0.161±0.038	0.161±0.037	0.159±0.037	**0.192±0.047**
	PAK2	0.72±0.056	0.76±0.060	0.79±0.068	0.75±0.073	**0.73±0.074**	0.035±0.019	0.080±0.035	0.180±0.042	0.289±0.068	**0.360±0.078**
	MKK7	0.96±0.084	0.98±0.089	0.98±0.088	0.96±0.083	**0.96±0.084**	0.547±0.016	0.719±0.034	0.736±0.045	0.747±0.053	**0.799±0.057**
	MEK1	0.71±0.050	0.73±0.056	**0.72±0.050**	0.74±0.044	0.75±0.032	0.497±0.040	0.485±0.011	**0.466±0.010**	0.476±0.009	0.476±0.006
CK1	CK1A	0.76±0.014	0.76±0.014	0.77±0.013	0.78±0.011	**0.78±0.009**	0.058±0.010	0.066±0.012	0.100±0.014	0.166±0.013	**0.195±0.011**
CK1	CK1D	0.85±0.007	0.86±0.010	0.87±0.008	0.88±0.007	**0.90±0.006**	0.047±0.017	0.128±0.028	0.118±0.026	0.183±0.030	**0.232±0.029**
CK1	CK1E	0.82±0.021	0.82±0.018	0.83±0.021	0.83±0.019	**0.87±0.018**	0.157±0.027	0.205±0.056	0.303±0.055	0.346±0.047	**0.415±0.059**
CK1	VRK1	0.54±0.024	0.68±0.022	0.77±0.026	0.81±0.051	**0.87±0.068**	0.265±0.011	0.266±0.006	0.342±0.031	0.345±0.017	**0.348±0.027**
Atypical	ATM	0.95±0.002	0.95±0.001	0.95±0.001	**0.95±0.002**	0.95±0.001	0.233±0.015	0.270±0.016	0.273±0.014	**0.277±0.017**	0.275±0.015
Atypical	ATR	**0.90±0.007**	0.88±0.007	0.86±0.009	0.85±0.011	0.82±0.012	**0.106±0.008**	0.106±0.014	0.114±0.014	0.103±0.010	0.099±0.016
Atypical	DNAPK	0.87±0.004	0.87±0.004	0.87±0.005	0.86±0.005	**0.86±0.005**	0.125±0.008	0.132±0.010	0.159±0.010	0.155±0.010	**0.170±0.012**
Atypical	mTOR	0.69±0.015	0.74±0.016	0.76±0.018	0.77±0.020	**0.81±0.017**	0.113±0.034	0.156±0.040	0.184±0.039	0.186±0.040	**0.220±0.040**

TABLE B.5: Sequence model accuracy for varying window sizes in **mouse** kinases, where kinases are grouped according to family. Table shows accuracy values for classifying kinase phosphorylation sites with the sequence model as determined by 10-fold cross-validation across 10 randomised data-set splits. Prediction accuracy is shown using median and standard deviation of the AUC and AUC50 across the data-set splits. Varying window sizes were applied to determine the optimal window size on a kinase-specific basis. The window size determined for a kinase is highlighted through bold text. Optimal window size was determined primarily through AUC50 as a measure of the model's accuracy at low false-positive rates. If accuracy did not increase through increasing window size, the lower window size was chosen.

		AUC					AUC50				
	Kinase	7	9	11	13	15	7	9	11	13	15
CMGC	ERK2	0.85±0.006	0.84±0.006	**0.83±0.007**	0.83±0.007	0.83±0.006	0.165±0.019	0.163±0.022	**0.224±0.024**	0.224±0.024	0.222±0.028
CMGC	ERK1	0.82±0.006	0.81±0.009	0.82±0.009	0.82±0.009	**0.82±0.010**	0.102±0.015	0.152±0.018	0.147±0.021	0.147±0.021	**0.164±0.021**
CMGC	CDK5	0.81±0.006	0.80±0.012	**0.77±0.010**	0.77±0.010	0.72±0.009	0.141±0.014	0.128±0.014	**0.172±0.013**	0.172±0.013	0.184±0.014
CMGC	CDK1	**0.79±0.013**	0.79±0.017	0.78±0.016	0.78±0.016	0.76±0.017	**0.184±0.030**	0.171±0.023	0.165±0.020	0.165±0.020	0.138±0.011
CMGC	JNK1	0.73±0.012	**0.78±0.014**	0.74±0.018	0.74±0.018	0.71±0.023	0.187±0.029	**0.219±0.040**	0.222±0.024	0.222±0.024	0.202±0.020
CMGC	P38A	0.73±0.018	0.72±0.026	0.70±0.020	0.70±0.020	**0.74±0.017**	0.184±0.023	0.136±0.018	0.180±0.017	0.180±0.017	**0.226±0.028**
CMGC	CDK2	0.76±0.025	0.77±0.030	0.77±0.034	0.77±0.034	**0.74±0.034**	0.110±0.024	0.192±0.022	0.314±0.041	0.314±0.041	**0.340±0.033**
CMGC	GSK3B	0.67±0.018	0.76±0.021	0.85±0.020	0.85±0.020	**0.83±0.021**	0.106±0.020	0.196±0.032	0.391±0.059	0.391±0.059	**0.414±0.049**
	PKACA	**0.81±0.007**	0.79±0.009	0.79±0.009	0.78±0.008	0.78±0.008	**0.245±0.014**	0.182±0.012	0.149±0.016	0.163±0.012	0.180±0.013

Continued on next page

	Kinase	7	9	11	13	15	7	9	11	13	15
										Continued from previous page	
AGC	PKCA	0.70±0.014	**0.72±0.010**	0.69±0.007	0.71±0.010	0.71±0.013	0.146±0.012	**0.253±0.016**	0.251±0.021	0.239±0.015	0.244±0.014
	Akt1	0.80±0.019	0.81±0.022	**0.81±0.011**	0.81±0.014	0.81±0.020	0.187±0.027	0.222±0.042	**0.383±0.047**	0.373±0.059	0.358±0.052
	PKCD	**0.75±0.028**	0.74±0.033	0.72±0.037	0.65±0.041	0.69±0.046	**0.113±0.037**	0.087±0.032	0.098±0.034	0.097±0.040	0.052±0.031
	p90RSK	**0.87±0.013**	0.80±0.013	0.76±0.012	0.73±0.016	0.76±0.016	**0.216±0.037**	0.236±0.044	0.237±0.037	0.241±0.049	0.290±0.031
	RSK2	**0.79±0.042**	0.75±0.051	0.68±0.069	0.64±0.073	0.60±0.067	**0.283±0.085**	0.286±0.086	0.284±0.085	0.284±0.085	0.284±0.085
	PKG1	**0.66±0.042**	0.66±0.040	0.56±0.035	0.58±0.024	0.67±0.024	**0.000±0.000**	0.000±0.001	0.000±0.000	0.000±0.000	0.000±0.004
	p70S6K	0.81±0.035	0.79±0.039	0.84±0.033	**0.88±0.029**	0.86±0.035	0.393±0.062	0.396±0.064	0.391±0.039	**0.394±0.062**	0.377±0.074
	PKCZ	0.64±0.076	0.66±0.070	0.69±0.098	0.63±0.087	**0.69±0.095**	0.087±0.041	0.134±0.055	0.137±0.057	0.277±0.111	**0.286±0.114**
	PKCE	0.51±0.032	0.53±0.039	0.55±0.049	0.55±0.052	**0.54±0.043**	0.251±0.072	0.222±0.066	0.324±0.087	0.432±0.100	**0.444±0.102**
TK	Src	0.54±0.016	0.57±0.014	**0.61±0.012**	0.60±0.014	0.57±0.011	0.160±0.019	0.215±0.022	**0.267±0.013**	0.273±0.012	0.248±0.013
	Fyn	**0.64±0.018**	0.66±0.016	0.66±0.015	0.62±0.013	0.63±0.014	**0.307±0.027**	0.284±0.031	0.262±0.034	0.265±0.025	0.283±0.039
	Abl	**0.52±0.042**	0.39±0.039	0.38±0.023	0.38±0.030	0.40±0.029	**0.151±0.008**	0.155±0.004	0.154±0.005	0.151±0.007	0.166±0.005
	Lyn	0.58±0.027	0.61±0.022	0.64±0.021	**0.65±0.027**	0.64±0.031	0.191±0.025	0.281±0.025	0.279±0.026	**0.286±0.029**	0.248±0.028
	Lck	0.65±0.040	0.64±0.056	**0.64±0.060**	0.64±0.070	0.62±0.070	0.199±0.060	0.193±0.059	**0.270±0.056**	0.228±0.059	0.186±0.048
	Syk	0.65±0.035	0.71±0.025	0.69±0.023	0.72±0.015	**0.71±0.014**	0.261±0.025	0.311±0.025	0.411±0.055	0.576±0.037	**0.601±0.024**

TABLE B.6: Sequence model accuracy for varying window sizes in **yeast** kinases, where kinases are grouped according to family. Table shows accuracy values for classifying kinase phosphorylation sites with the sequence model as determined by 10-fold cross-validation across 10 randomised data-set splits. Prediction accuracy is shown using median and standard deviation of the AUC and AUC50 across the data-set splits. Varying window sizes were applied to determine the optimal window size on a kinase-specific basis. The window size determined for a kinase is highlighted through bold text. Optimal window size was determined primarily through AUC50 as a measure of the model's accuracy at low false-positive rates. If accuracy did not increase through increasing window size, the lower window size was chosen.

		AUC					AUC50				
	Kinase	7	9	11	13	15	7	9	11	13	15
CMGC	CDC28	0.93±0.001	0.93±0.001	0.93±0.001	0.93±0.001	**0.93±0.001**	0.243±0.008	0.242±0.008	0.234±0.010	0.262±0.013	**0.295±0.012**
	CTK1	0.72±0.011	0.70±0.012	0.70±0.012	**0.70±0.008**	0.71±0.009	0.418±0.001	0.417±0.002	0.421±0.002	**0.434±0.000**	0.430±0.002
	MCK1	0.73±0.021	0.79±0.014	0.80±0.012	0.82±0.007	**0.83±0.009**	0.141±0.020	0.209±0.027	0.261±0.026	0.324±0.016	**0.348±0.024**
	PHO85	0.66±0.013	0.70±0.013	0.72±0.018	0.71±0.023	**0.71±0.018**	0.097±0.014	0.094±0.023	0.124±0.010	0.154±0.010	**0.172±0.010**
	SSN3	0.69±0.053	0.72±0.051	0.76±0.042	0.73±0.052	**0.74±0.057**	0.044±0.027	0.204±0.051	0.230±0.052	0.296±0.063	**0.295±0.064**
	HOG1	0.66±0.042	0.66±0.046	0.70±0.049	0.75±0.046	**0.79±0.047**	0.042±0.015	0.048±0.027	0.063±0.024	0.208±0.065	**0.301±0.052**
	KNS1	0.88±0.031	0.92±0.024	0.93±0.028	0.92±0.032	**0.93±0.038**	0.268±0.023	0.431±0.044	0.506±0.047	0.521±0.055	**0.591±0.056**

Continued on next page

	Kinase	7	9	11	13	15	7	9	11	13	15
											Continued from previous page
	SLT2	0.61±0.030	0.62±0.043	0.61±0.036	0.63±0.032	**0.68±0.037**	0.008±0.016	0.059±0.002	0.089±0.030	0.263±0.041	**0.271±0.048**
	FUS3	0.54±0.025	**0.54±0.035**	0.51±0.031	0.45±0.025	0.47±0.029	0.108±0.036	**0.217±0.004**	0.220±0.002	0.222±0.000	0.222±0.000
AGC	TPK1	0.95±0.003	**0.95±0.003**	0.95±0.004	0.94±0.004	0.93±0.005	0.355±0.017	**0.383±0.011**	0.373±0.010	0.360±0.010	0.382±0.013
AGC	TPK3	0.72±0.045	0.74±0.034	0.78±0.039	0.76±0.032	**0.81±0.036**	0.206±0.046	0.309±0.062	0.440±0.071	0.525±0.074	**0.595±0.058**
AGC	YPK1	0.76±0.037	0.80±0.039	0.80±0.045	**0.74±0.043**	0.68±0.037	0.241±0.052	0.303±0.076	0.352±0.091	**0.443±0.087**	0.398±0.083
AGC	PKH2	0.74±0.054	**0.75±0.037**	0.72±0.053	0.68±0.048	0.64±0.048	0.240±0.004	**0.250±0.003**	0.250±0.000	0.250±0.000	0.249±2.776e-17
AGC	PKH1	0.95±0.020	0.96±0.014	0.97±0.007	**0.98±0.006**	0.96±0.010	0.738±0.003	0.745±0.003	0.749±0.002	**0.750±0.000**	0.750±0.000
AGC	PKC1	**0.88±0.024**	0.89±0.017	0.87±0.010	0.90±0.016	0.87±0.020	**0.346±0.045**	0.269±0.058	0.192±0.044	0.228±0.044	0.232±0.045
CAMK	SNF1	0.73±0.013	0.73±0.013	0.74±0.020	0.76±0.018	**0.78±0.014**	0.040±0.009	0.040±0.009	0.078±0.027	0.153±0.035	**0.162±0.032**
CAMK	FRK1	0.68±0.026	0.68±0.026	0.68±0.027	0.73±0.021	**0.75±0.021**	0.181±0.038	0.181±0.038	0.371±0.050	0.404±0.050	0.424±0.048
CAMK	PSK2	0.71±0.027	0.71±0.027	0.73±0.040	0.73±0.047	**0.74±0.047**	0.374±0.044	0.374±0.044	0.393±0.049	0.402±0.050	0.413±0.055
CAMK	DUN1	0.87±0.012	0.87±0.012	**0.85±0.013**	0.85±0.015	0.87±0.015	0.273±0.016	0.273±0.016	**0.379±0.012**	0.374±0.011	0.358±0.009
Other	CKA1	0.90±0.003	0.90±0.003	0.90±0.003	0.89±0.005	**0.89±0.005**	0.200±0.019	0.210±0.013	0.248±0.014	0.287±0.018	**0.313±0.015**
Other	CKA2	0.92±0.005	0.91±0.006	0.91±0.006	0.91±0.006	**0.91±0.007**	0.154±0.015	0.199±0.011	0.276±0.015	0.334±0.014	**0.355±0.017**
Other	MPS1	0.83±0.013	0.83±0.016	0.83±0.020	0.86±0.015	**0.86±0.016**	0.078±0.017	0.122±0.017	0.155±0.032	0.174±0.033	**0.231±0.036**
Other	PTK1	0.61±0.015	0.63±0.017	0.62±0.020	0.67±0.013	**0.67±0.015**	0.024±0.011	0.050±0.009	0.048±0.011	0.088±0.013	**0.139±0.020**
Other	PTK2	0.79±0.027	0.81±0.034	0.86±0.045	0.86±0.042	**0.89±0.046**	0.302±0.049	0.419±0.033	0.517±0.049	0.640±0.049	**0.755±0.065**
Other	IPL1	**0.91±0.009**	0.89±0.013	0.87±0.013	0.83±0.012	0.83±0.016	**0.276±0.018**	0.200±0.027	0.232±0.041	0.158±0.031	0.139±0.027
Other	BUD32	0.61±0.076	0.63±0.070	0.72±0.070	**0.73±0.063**	0.74±0.069	0.020±0.029	0.177±0.044	0.315±0.067	**0.385±0.071**	0.310±0.071

TABLE B.7: Comparison of prediction accuracy across **human** kinases between predicting kinase-specific phosphorylation sites with a baseline model that only considers position-specific amino acid frequencies, and the sequence model. Kinases are grouped according to their family, with the average prediction accuracy for each family included. Results were generated using ten-fold cross-validation repeated across ten randomised data-set splits. Shown are the average and standard deviation of the AUC and AUC50 values.

		AUC		AUC50	
	Kinase	Baseline	Sequence model	Baseline	Sequence model
CMGC	CDK2	0.86±0.001	0.89±0.001	0.06±0.002	0.10±0.004
	CDK1	0.88±0.002	0.89±0.002	0.09±0.004	0.07±0.008
	ERK2	0.86±0.002	0.86±0.001	0.05±0.004	0.07±0.010
	ERK1	0.86±0.005	0.86±0.005	0.04±0.005	0.07±0.012
	GSK3B	0.77±0.009	0.81±0.006	0.09±0.007	0.13±0.014
	P38A	0.79±0.007	0.81±0.007	0.12±0.016	0.15±0.017
	JNK1	0.83±0.005	0.87±0.004	0.08±0.013	0.15±0.014
	CDK5	0.84±0.012	0.84±0.009	0.07±0.009	0.05±0.007
	JNK2	0.75±0.015	0.73±0.023	0.03±0.013	0.07±0.015
	CDK7	0.77±0.017	0.88±0.019	0.16±0.044	0.31±0.032
	GSK3A	0.89±0.014	0.90±0.026	0.26±0.020	0.46±0.045
	CDK4	0.85±0.012	0.87±0.012	0.07±0.007	0.18±0.025
	P38B	0.79±0.006	0.83±0.014	0.07±0.015	0.26±0.046
	HIPK2	0.81±0.016	0.86±0.013	0.23±0.030	0.38±0.043
	DYRK1A	0.77±0.034	0.83±0.033	0.01±0.024	0.26±0.043
	CDK9	0.78±0.011	0.83±0.015	0.04±0.022	0.32±0.030
	DYRK2	0.68±0.032	0.78±0.019	0.00±0.000	0.31±0.043
	ERK5	0.79±0.015	0.83±0.016	0.02±0.014	0.32±0.034
	CDK6	0.80±0.019	0.86±0.009	0.07±0.016	0.18±0.030
	CDK3	0.69±0.031	0.76±0.050	0.00±0.000	0.36±0.045
	Average	0.80±0.013	0.84±0.014	0.078±0.013	0.21±0.026
AGC	PKACA	0.89±0.003	0.89±0.003	0.10±0.005	0.12±0.008
	PKCA	0.82±0.004	0.84±0.001	0.10±0.004	0.13±0.009
	Akt1	0.91±0.005	0.92±0.004	0.23±0.014	0.18±0.017
	PKCD	0.67±0.011	0.70±0.009	0.05±0.007	0.04±0.006
	PKG1	0.86±0.019	0.86±0.027	0.25±0.035	0.20±0.020
	p90RSK	0.74±0.022	0.80±0.010	0.05±0.010	0.17±0.037
	PKCE	0.59±0.015	0.67±0.017	0.07±0.018	0.10±0.006
	PKCZ	0.55±0.022	0.63±0.020	0.01±0.007	0.14±0.029
	PKCB	0.64±0.025	0.71±0.019	0.11±0.018	0.13±0.028
	RSK2	0.68±0.031	0.71±0.023	0.08±0.012	0.12±0.017
	ROCK1	0.71±0.008	0.76±0.012	0.15±0.023	0.15±0.032
	PDK1	0.85±0.020	0.84±0.018	0.46±0.009	0.50±0.024

Continued on next page

	Kinase	Baseline	Sequence model	Baseline	Sequence model
				Continued from previous page	
	PKCT	0.80±0.025	0.77±0.041	0.11±0.037	0.12±0.047
	PKCG	0.62±0.023	0.65±0.024	0.01±0.007	0.11±0.064
	p70S6K	0.78±0.013	0.83±0.010	0.11±0.026	0.28±0.029
	SGK1	0.83±0.018	0.83±0.018	0.28±0.045	0.33±0.011
	Akt2	0.82±0.023	0.87±0.012	0.11±0.036	0.16±0.020
	GRK2	0.73±0.014	0.86±0.014	0.09±0.028	0.53±0.033
	ROCK2	0.78±0.015	0.77±0.015	0.13±0.012	0.17±0.002
	PKCI	0.82±0.017	0.81±0.023	0.28±0.049	0.16±0.049
	PKCH	0.83±0.027	0.90±0.026	0.32±0.059	0.56±0.038
	PKN1	0.77±0.021	0.79±0.058	0.29±0.148	0.20±0.108
	Average	0.76±0.017	0.79±0.018	0.154±0.028	0.21±0.029
TK	Src	0.53±0.004	0.56±0.006	0.07±0.004	0.10±0.005
TK	Abl	0.58±0.011	0.62±0.009	0.11±0.007	0.15±0.016
TK	Fyn	0.54±0.011	0.59±0.009	0.10±0.008	0.12±0.009
TK	Lck	0.54±0.014	0.53±0.012	0.05±0.009	0.06±0.016
TK	Lyn	0.50±0.017	0.48±0.016	0.08±0.011	0.05±0.012
TK	EGFR	0.54±0.015	0.56±0.023	0.06±0.005	0.05±0.018
TK	Syk	0.78±0.018	0.81±0.018	0.27±0.020	0.27±0.025
TK	InsR	0.61±0.030	0.69±0.026	0.21±0.020	0.35±0.025
TK	JAK2	0.50±0.018	0.58±0.028	0.10±0.016	0.16±0.030
TK	FAK	0.44±0.025	0.67±0.050	0.09±0.030	0.36±0.067
TK	Ret	0.43±0.026	0.54±0.023	0.17±0.027	0.19±0.025
TK	Arg	0.66±0.039	0.67±0.036	0.15±0.022	0.15±0.017
TK	Brk	0.56±0.016	0.60±0.021	0.15±0.026	0.20±0.007
TK	ALK	0.49±0.021	0.57±0.032	0.04±0.020	0.00±0.000
TK	Btk	0.60±0.036	0.71±0.033	0.14±0.044	0.31±0.053
TK	PDGFRB	0.59±0.017	0.61±0.033	0.09±0.043	0.25±0.040
TK	JAK3	0.63±0.040	0.81±0.032	0.19±0.053	0.40±0.063
TK	Hck	0.51±0.026	0.58±0.025	0.08±0.022	0.09±0.017
TK	Pyk2	0.64±0.027	0.62±0.033	0.00±0.021	0.17±0.019
	Average	0.56±0.022	0.62±0.025	0.11±0.021	0.18±0.024
CAMK	CAMK2A	0.64±0.011	0.68±0.011	0.10±0.011	0.12±0.012
CAMK	Chk1	0.69±0.022	0.71±0.017	0.07±0.017	0.06±0.022
CAMK	AMPKA1	0.75±0.019	0.72±0.016	0.10±0.014	0.08±0.014
CAMK	MAPKAPK2	0.79±0.016	0.78±0.019	0.08±0.020	0.14±0.028
CAMK	PKD1	0.75±0.015	0.76±0.010	0.08±0.014	0.09±0.012
CAMK	LKB1	0.77±0.013	0.81±0.009	0.47±0.003	0.58±0.018
CAMK	MSK1	0.76±0.044	0.86±0.032	0.10±0.049	0.33±0.076

Continued on next page

	Kinase	Baseline	Sequence model	Baseline	Sequence model
				Continued from previous page	
	Chk2	0.59±0.023	0.62±0.020	0.03±0.008	0.03±0.010
	Pim1	0.72±0.018	0.84±0.025	0.01±0.010	0.35±0.031
	AMPKA2	0.81±0.031	0.86±0.028	0.07±0.024	0.12±0.037
	MARK2	0.80±0.024	0.80±0.024	0.26±0.020	0.24±0.002
	CAMK1A	0.86±0.015	0.83±0.016	0.41±0.017	0.42±0.065
	DAPK3	0.47±0.035	0.67±0.035	0.00±0.010	0.19±0.065
	CaMK4	0.76±0.028	0.79±0.032	0.00±0.000	0.00±0.000
	PKD2	0.84±0.038	0.80±0.054	0.02±0.011	0.07±0.040
	CAMK2D	0.72±0.022	0.83±0.041	0.00±0.000	0.25±0.000
	Average	0.73±0.023	0.77±0.024	0.11±0.014	0.19±0.027
Other	CK2A1	0.93±0.002	0.93±0.001	0.36±0.005	0.39±0.004
	PLK1	0.72±0.010	0.78±0.007	0.07±0.013	0.12±0.016
	AurB	0.77±0.010	0.79±0.010	0.05±0.008	0.09±0.010
	AurA	0.73±0.016	0.74±0.012	0.02±0.012	0.10±0.012
	PLK3	0.55±0.019	0.66±0.039	0.00±0.000	0.21±0.039
	IKKA	0.53±0.010	0.69±0.013	0.00±0.005	0.24±0.046
	IKKB	0.52±0.017	0.75±0.021	0.01±0.010	0.37±0.022
	TBK1	0.59±0.038	0.76±0.032	0.04±0.016	0.30±0.041
	CK2A2	0.81±0.015	0.91±0.036	0.08±0.022	0.44±0.063
	IKKE	0.82±0.038	0.96±0.011	0.09±0.015	0.69±0.088
	TTK	0.60±0.025	0.82±0.036	0.05±0.016	0.35±0.057
	NEK6	0.77±0.020	0.78±0.021	0.08±0.033	0.31±0.035
	NEK2	0.63±0.024	0.76±0.041	0.00±0.019	0.49±0.064
	Average	0.69±0.019	0.8±0.021	0.066±0.013	0.32±0.038
STE	PAK1	0.69±0.012	0.70±0.013	0.03±0.007	0.04±0.009
	Cot	0.79±0.022	0.84±0.020	0.48±0.098	0.50±0.086
	MST1	0.61±0.035	0.75±0.042	0.00±0.014	0.20±0.028
	ASK1	0.64±0.048	0.82±0.021	0.14±0.047	0.39±0.061
	MKK4	0.86±0.012	0.90±0.038	0.54±0.035	0.64±0.029
	MST2	0.64±0.035	0.72±0.052	0.12±0.035	0.19±0.047
	PAK2	0.64±0.031	0.73±0.074	0.00±0.000	0.36±0.078
	MKK7	0.78±0.032	0.96±0.084	0.54±0.004	0.80±0.054
	MEK1	0.83±0.039	0.72±0.050	0.50±0.115	0.47±0.009
	Average	0.71±0.031	0.79±0.052	0.228±0.049	0.38±0.053
CK1	CK1A	0.70±0.014	0.78±0.009	0.08±0.014	0.19±0.011
	CK1D	0.84±0.008	0.90±0.006	0.07±0.019	0.23±0.029
	CK1E	0.72±0.027	0.87±0.018	0.05±0.021	0.42±0.059

Continued on next page

	Kinase	Baseline	Sequence model	Baseline	Sequence model
				Continued from previous page	
	VRK1	0.72±0.032	0.87±0.068	0.29±0.022	0.35±0.027
	Average	0.75±0.02	0.86±0.025	0.124±0.019	0.30±0.031
Atypical	ATM	0.95±0.002	0.95±0.002	0.37±0.014	0.28±0.017
Atypical	ATR	0.86±0.008	0.86±0.009	0.14±0.009	0.11±0.014
Atypical	DNAPK	0.83±0.008	0.86±0.005	0.13±0.005	0.17±0.012
Atypical	mTOR	0.72±0.019	0.81±0.017	0.08±0.003	0.22±0.040
	Average	0.84±0.009	0.87±0.008	0.18±0.008	0.20±0.029

Table B.8: Comparison of prediction accuracy across **mouse** kinases between predicting kinase-specific phosphorylation sites with a baseline model that only considers position-specific amino acid frequencies, and the sequence model. Kinases are grouped according to their family, with the average prediction accuracy for each family included. Results were generated using ten-fold cross-validation repeated across ten randomised data-set splits. Shown are the average and standard deviation of the AUC and AUC50 values.

		AUC		AUC50	
	Kinase	Baseline	Sequence model	Baseline	Sequence model
CMGC	ERK2	0.81±0.006	0.83±0.006	0.27±0.011	0.19±0.017
CMGC	ERK1	0.78±0.011	0.82±0.010	0.19±0.018	0.16±0.021
CMGC	CDK5	0.73±0.013	0.80±0.013	0.09±0.015	0.17±0.016
CMGC	CDK1	0.76±0.022	0.79±0.013	0.17±0.018	0.18±0.030
CMGC	JNK1	0.74±0.019	0.78±0.014	0.13±0.018	0.22±0.040
CMGC	P38A	0.67±0.021	0.74±0.017	0.10±0.022	0.23±0.028
CMGC	CDK2	0.76±0.020	0.74±0.034	0.10±0.020	0.34±0.033
CMGC	GSK3B	0.70±0.018	0.83±0.021	0.07±0.011	0.41±0.049
	Average	0.74±0.016	0.79±0.016	0.14±0.017	0.24±0.029
AGC	PKACA	0.78±0.006	0.81±0.007	0.22±0.014	0.25±0.014
AGC	PKCA	0.67±0.014	0.72±0.010	0.15±0.011	0.25±0.016
AGC	Akt1	0.82±0.015	0.81±0.011	0.34±0.049	0.38±0.047
AGC	PKCD	0.71±0.014	0.75±0.028	0.13±0.024	0.11±0.037
AGC	p90RSK	0.90±0.015	0.87±0.013	0.31±0.048	0.22±0.037
AGC	RSK2	0.80±0.056	0.79±0.042	0.29±0.087	0.28±0.085
AGC	PKG1	0.70±0.023	0.66±0.042	0.12±0.049	0.00±0.000
AGC	p70S6K	0.82±0.032	0.88±0.029	0.18±0.062	0.39±0.062
AGC	PKCZ	0.61±0.050	0.69±0.095	0.00±0.000	0.29±0.114
AGC	PKCE	0.38±0.028	0.54±0.043	0.00±0.000	0.44±0.102
	Average	0.72±0.025	0.75±0.032	0.17±0.034	0.26±0.051

Continued on next page

	Kinase	Baseline	Sequence model	Baseline	Sequence model
				Continued from previous page	
TK	Src	0.52±0.021	0.61±0.012	0.17±0.024	0.27±0.013
	Fyn	0.66±0.018	0.64±0.018	0.33±0.030	0.31±0.027
	Abl	0.49±0.035	0.52±0.042	0.15±0.025	0.15±0.008
	Lyn	0.66±0.023	0.65±0.027	0.25±0.026	0.29±0.029
	Lck	0.72±0.030	0.64±0.060	0.32±0.046	0.27±0.056
	Syk	0.57±0.023	0.71±0.014	0.33±0.041	0.60±0.024
	Average	0.60±0.025	0.63±0.029	0.26±0.032	0.31±0.026

TABLE B.9: Comparison of prediction accuracy across **yeast** kinases between predicting kinase-specific phosphorylation sites with a baseline model that only considers position-specific amino acid frequencies, and the sequence model. Kinases are grouped according to their family, with the average prediction accuracy for each family included. Results were generated using ten-fold cross-validation repeated across ten randomised data-set splits. Shown are the average and standard deviation of the AUC and AUC50 values.

		AUC		AUC50	
	Kinase	Baseline	Sequence model	Baseline	Sequence model
CMGC	CDC28	0.93±0.001	0.93±0.001	0.30±0.003	0.29±0.012
	CTK1	0.75±0.009	0.70±0.008	0.47±0.004	0.43±0.000
	MCK1	0.69±0.026	0.83±0.009	0.06±0.009	0.35±0.024
	PHO85	0.64±0.014	0.71±0.018	0.06±0.010	0.17±0.010
	SSN3	0.54±0.035	0.74±0.057	0.00±0.000	0.29±0.064
	HOG1	0.62±0.034	0.79±0.047	0.07±0.022	0.30±0.052
	KNS1	0.78±0.038	0.93±0.038	0.01±0.008	0.59±0.056
	SLT2	0.56±0.039	0.68±0.037	0.00±0.011	0.27±0.048
	FUS3	0.51±0.055	0.54±0.035	0.00±0.000	0.22±0.004
	Average	0.67±0.028	0.76±0.028	0.11±0.007	0.32±0.03
AGC	TPK1	0.94±0.004	0.95±0.003	0.39±0.013	0.38±0.011
	TPK3	0.63±0.040	0.81±0.036	0.19±0.014	0.60±0.058
	YPK1	0.63±0.018	0.74±0.043	0.03±0.024	0.44±0.087
	PKH2	0.77±0.028	0.75±0.037	0.07±0.046	0.25±0.003
	PKH1	0.91±0.013	0.98±0.006	0.55±0.024	0.75±0.000
	PKC1	0.87±0.014	0.88±0.024	0.19±0.039	0.35±0.045
	Average	0.79±0.02	0.85±0.025	0.24±0.027	0.46±0.034
TK	SNF1	0.68±0.011	0.78±0.014	0.01±0.004	0.16±0.032
	FRK1	0.57±0.032	0.75±0.021	0.00±0.000	0.42±0.048
	PSK2	0.59±0.026	0.74±0.047	0.12±0.043	0.41±0.055
	DUN1	0.73±0.027	0.85±0.013	0.07±0.020	0.38±0.012

Continued on next page

	Kinase	Baseline	Sequence model	Baseline	Sequence model
				Continued from previous page	
	Average	0.64±0.024	0.78±0.024	0.05±0.017	0.34±0.037
Other	CKA1	0.90±0.003	0.89±0.005	0.18±0.014	0.31±0.015
	CKA2	0.91±0.006	0.91±0.007	0.17±0.014	0.36±0.017
	MPS1	0.82±0.017	0.86±0.016	0.09±0.021	0.23±0.036
	PTK1	0.58±0.014	0.67±0.015	0.00±0.005	0.14±0.020
	PTK2	0.66±0.019	0.89±0.046	0.00±0.000	0.75±0.065
	IPL1	0.91±0.010	0.91±0.009	0.28±0.017	0.28±0.018
	BUD32	0.39±0.049	0.73±0.063	0.00±0.000	0.39±0.071
	Average	0.74±0.017	0.84±0.023	0.10±0.01	0.35±0.035

Table B.10: Comparison of prediction accuracy across human kinases between predicting kinase-specific phosphorylation sites using the sequence model trained on the full data-set, and when the model is trained on the similarity-reduced data-set. Prediction accuracy is calculated on the similarity-reduced data-set. If a kinase could not be trained on the reduced data-set due to too few positive training samples it was marked as "N/A". Kinases are grouped according to their family, with the average prediction accuracy for each family included. Results were generated using ten-fold cross-validation repeated across ten randomised data-set splits. Shown are the average and standard deviation of the AUC and AUC50 values.

		AUC		AUC50	
	Kinase	Full set	Reduced set	Full set	Reduced set
CMGC	CDK2	0.89±0.001	0.89±0.001	0.10±0.003	0.10±0.009
	CDK1	0.89±0.002	0.90±0.001	0.08±0.008	0.09±0.007
	ERK2	0.86±0.002	0.86±0.002	0.07±0.011	0.06±0.007
	ERK1	0.85±0.005	0.85±0.002	0.06±0.010	0.05±0.008
	GSK3B	0.81±0.006	0.81±0.004	0.14±0.016	0.13±0.007
	P38A	0.81±0.008	0.80±0.007	0.14±0.015	0.11±0.022
	JNK1	0.86±0.005	0.86±0.006	0.17±0.016	0.16±0.023
	CDK5	0.84±0.010	0.83±0.008	0.04±0.004	0.03±0.006
	JNK2	0.72±0.023	0.71±0.017	0.05±0.012	0.02±0.013
	CDK7	0.86±0.022	0.85±0.023	0.23±0.038	0.25±0.060
	GSK3A	0.89±0.028	0.91±0.023	0.46±0.050	0.45±0.052
	CDK4	0.87±0.012	0.86±0.010	0.18±0.025	0.15±0.029
	P38B	0.83±0.014	0.84±0.014	0.26±0.045	0.27±0.022
	HIPK2	0.85±0.015	0.85±0.009	0.34±0.046	0.27±0.044
	DYRK1A	0.83±0.033	0.84±0.027	0.26±0.045	0.27±0.065
	CDK9	0.82±0.017	0.80±0.017	0.34±0.033	0.31±0.033
	DYRK2	0.78±0.019	0.80±0.028	0.31±0.042	0.28±0.066
	ERK5	0.82±0.017	0.84±0.024	0.31±0.034	0.34±0.036

Continued on next page

	Kinase	Full set	Reduced set	Full set	Reduced set
				Continued from previous page	
	CDK6	0.85±0.009	0.83±0.012	0.17±0.027	0.09±0.033
	CDK3	0.73±0.057	0.66±0.061	0.28±0.054	0.14±0.012
	Average	0.84±0.015	0.83±0.015	0.198±0.027	0.178±0.027
AGC	PKACA	0.89±0.003	0.89±0.002	0.12±0.007	0.11±0.004
	PKCA	0.84±0.001	0.83±0.002	0.12±0.009	0.13±0.006
	Akt1	0.91±0.004	0.91±0.003	0.18±0.018	0.18±0.012
	PKCD	0.70±0.009	0.69±0.007	0.04±0.007	0.04±0.009
	PKG1	0.84±0.026	0.82±0.013	0.12±0.023	0.07±0.018
	PKCB	0.71±0.020	0.70±0.029	0.12±0.022	0.12±0.021
	PKCE	0.66±0.018	0.64±0.025	0.06±0.006	0.02±0.012
	p90RSK	0.78±0.011	0.76±0.012	0.16±0.036	0.19±0.041
	PKCZ	0.62±0.020	0.64±0.034	0.13±0.026	0.14±0.022
	ROCK1	0.74±0.012	0.73±0.016	0.15±0.030	0.09±0.024
	GRK2	0.86±0.015	0.85±0.011	0.52±0.034	0.51±0.044
	RSK2	0.72±0.021	0.72±0.028	0.11±0.014	0.11±0.021
	PDK1	0.79±0.024	0.75±0.026	0.35±0.023	0.32±0.014
	p70S6K	0.82±0.010	0.83±0.021	0.26±0.029	0.25±0.052
	PKCG	0.65±0.023	0.62±0.025	0.11±0.064	0.13±0.042
	PKCT	0.77±0.045	0.75±0.020	0.14±0.052	0.14±0.002
	SGK1	0.81±0.020	0.77±0.020	0.27±0.009	0.23±0.029
	Akt2	0.86±0.015	0.84±0.021	0.19±0.023	0.20±0.015
	ROCK2	0.76±0.015	0.77±0.019	0.18±0.000	0.18±0.071
	PKCH	0.88±0.030	0.89±0.034	0.51±0.047	0.49±0.074
	PKCI	0.81±0.022	0.80±0.034	0.16±0.049	0.16±0.064
	PKN1	N/A	N/A	N/A	N/A
	Average	0.78±0.017	0.77±0.019	0.19±0.025	0.181±0.028
TK	Src	0.55±0.006	0.55±0.006	0.09±0.004	0.08±0.009
	Abl	0.62±0.009	0.62±0.011	0.14±0.015	0.15±0.012
	Fyn	0.59±0.009	0.57±0.012	0.12±0.010	0.12±0.008
	Lck	0.53±0.013	0.50±0.016	0.06±0.013	0.05±0.013
	EGFR	0.55±0.023	0.52±0.011	0.04±0.015	0.06±0.020
	InsR	N/A	N/A	N/A	N/A
	Lyn	0.48±0.016	0.48±0.021	0.03±0.008	0.02±0.008
	Syk	0.81±0.018	0.81±0.013	0.28±0.027	0.27±0.019
	JAK2	0.57±0.027	0.55±0.021	0.14±0.027	0.14±0.016
	Ret	0.50±0.027	0.53±0.026	0.12±0.026	0.18±0.036
	PDGFRB	0.61±0.033	0.63±0.048	0.26±0.040	0.26±0.051
	Hck	0.53±0.028	0.54±0.023	0.07±0.018	0.11±0.023

Continued on next page

	Kinase	Full set	Reduced set	Full set	Reduced set
				Continued from previous page	
	Btk	0.74±0.039	0.73±0.024	0.28±0.046	0.20±0.025
	FAK	0.67±0.050	0.63±0.032	0.36±0.067	0.35±0.067
	Arg	0.67±0.037	0.67±0.033	0.16±0.018	0.18±0.054
	JAK3	0.82±0.031	0.79±0.035	0.40±0.062	0.33±0.072
	Brk	0.60±0.021	0.61±0.045	0.20±0.007	0.19±0.052
	ALK	0.57±0.033	0.55±0.069	0.00±0.000	0.00±0.006
	Pyk2	0.62±0.033	0.61±0.091	0.18±0.019	0.18±0.062
	Average	0.61±0.025	0.60±0.03	0.163±0.023	0.160±0.03
CAMK	CAMK2A	0.66±0.011	0.63±0.026	0.11±0.011	0.10±0.022
CAMK	Chk1	0.70±0.018	0.68±0.011	0.04±0.019	0.03±0.013
CAMK	AMPKA1	0.72±0.015	0.70±0.018	0.09±0.014	0.09±0.017
CAMK	MAPKAPK2	0.77±0.019	0.78±0.011	0.09±0.026	0.12±0.013
CAMK	Chk2	0.62±0.020	0.63±0.026	0.03±0.010	0.02±0.011
CAMK	PKD1	0.77±0.011	0.76±0.020	0.07±0.013	0.06±0.021
CAMK	LKB1	0.73±0.013	0.73±0.026	0.41±0.027	0.43±0.035
CAMK	MSK1	0.85±0.026	0.84±0.047	0.36±0.085	0.41±0.055
CAMK	CAMK1A	0.81±0.018	0.79±0.022	0.37±0.037	0.24±0.002
CAMK	Pim1	0.85±0.024	0.87±0.012	0.39±0.036	0.40±0.032
CAMK	CaMK4	0.79±0.036	0.80±0.054	0.00±0.000	0.00±0.000
CAMK	DAPK3	0.67±0.035	0.71±0.028	0.19±0.065	0.22±0.052
CAMK	AMPKA2	0.86±0.030	0.87±0.021	0.13±0.040	0.14±0.025
CAMK	MARK2	0.76±0.029	0.67±0.021	0.10±0.001	0.01±0.017
CAMK	PKD2	0.79±0.054	0.79±0.025	0.08±0.042	0.08±0.027
CAMK	CAMK2D	0.83±0.041	0.84±0.035	0.25±0.000	0.25±0.100
	Average	0.76±0.025	0.76±0.025	0.169±0.026	0.16±0.028
Other	CK2A1	0.93±0.001	0.93±0.002	0.38±0.003	0.38±0.003
Other	PLK1	0.78±0.007	0.77±0.009	0.12±0.016	0.12±0.016
Other	AurB	0.79±0.010	0.78±0.008	0.07±0.010	0.06±0.009
Other	AurA	0.75±0.013	0.74±0.012	0.11±0.013	0.11±0.016
Other	PLK3	0.66±0.039	0.66±0.025	0.22±0.040	0.21±0.044
Other	IKKA	0.69±0.013	0.68±0.015	0.24±0.046	0.23±0.030
Other	IKKB	0.75±0.021	0.75±0.015	0.37±0.022	0.37±0.017
Other	TBK1	0.79±0.034	0.81±0.014	0.31±0.043	0.33±0.053
Other	CK2A2	0.91±0.036	0.92±0.022	0.44±0.063	0.48±0.030
Other	IKKE	0.96±0.011	0.95±0.016	0.69±0.088	0.68±0.038
Other	TTK	0.84±0.038	0.84±0.026	0.38±0.061	0.41±0.041
Other	NEK6	0.78±0.021	0.78±0.016	0.32±0.035	0.33±0.028
Other	NEK2	0.76±0.041	0.77±0.029	0.49±0.065	0.50±0.050

Continued on next page

	Kinase	Full set	Reduced set	Full set	Reduced set
				Continued from previous page	
	Average	0.80±0.022	0.80±0.021	0.32±0.039	0.324±0.029
STE	PAK1	0.69±0.013	0.67±0.013	0.04±0.010	0.03±0.009
	Cot	0.82±0.024	0.80±0.061	0.43±0.083	0.37±0.054
	MST1	0.75±0.042	0.75±0.027	0.20±0.028	0.21±0.029
	ASK1	0.79±0.024	0.77±0.045	0.36±0.056	0.40±0.053
	MKK4	0.82±0.068	0.83±0.015	0.44±0.067	0.42±0.005
	MST2	0.70±0.046	0.62±0.029	0.14±0.034	0.09±0.048
	PAK2	0.73±0.074	0.77±0.082	0.36±0.078	0.37±0.094
	MKK7	0.93±0.152	0.94±0.093	0.76±0.114	0.80±0.121
	MEK1	0.63±0.067	0.67±0.086	0.31±0.006	0.29±0.088
	Average	0.76±0.057	0.76±0.05	0.339±0.053	0.332±0.056
CK1	CK1A	0.77±0.009	0.77±0.010	0.19±0.011	0.17±0.015
	CK1D	0.90±0.006	0.90±0.006	0.23±0.029	0.23±0.033
	CK1E	0.87±0.018	0.87±0.032	0.42±0.059	0.43±0.066
	VRK1	0.87±0.058	0.89±0.032	0.39±0.021	0.32±0.023
	Average	0.85±0.023	0.86±0.02	0.306±0.03	0.288±0.034
Atypical	ATM	0.95±0.002	0.95±0.002	0.29±0.017	0.28±0.015
	ATR	0.86±0.009	0.86±0.008	0.12±0.014	0.11±0.016
	DNAPK	0.86±0.005	0.85±0.006	0.17±0.012	0.19±0.011
	mTOR	0.81±0.018	0.78±0.021	0.22±0.038	0.17±0.039
	Average	0.87±0.008	0.86±0.009	0.201±0.02	0.186±0.02

Table B.11: Comparison of prediction accuracy across **mouse** kinases between predicting kinase-specific phosphorylation sites using the sequence model trained on the full data-set, and when the model is trained on the similarity-reduced data-set. Prediction accuracy is calculated on the similarity-reduced data-set. If a kinase could not be trained on the reduced data-set due to too few positive training samples it was marked as "N/A". Kinases are grouped according to their family, with the average prediction accuracy for each family included. Results were generated using ten-fold cross-validation repeated across ten randomised data-set splits. Shown are the average and standard deviation of the AUC and AUC50 values.

		AUC		AUC50	
	Kinase	Full set	Reduced set	Full set	Reduced set
CMGC	ERK2	0.83±0.006	0.83±0.009	0.19±0.017	0.19±0.019
	ERK1	0.82±0.010	0.83±0.019	0.16±0.021	0.18±0.030
	CDK5	0.79±0.013	0.79±0.011	0.15±0.016	0.13±0.032
	CDK1	0.79±0.013	0.78±0.018	0.19±0.031	0.17±0.026
	JNK1	0.77±0.014	0.78±0.020	0.22±0.040	0.26±0.039

Continued on next page

	Kinase	Full set	Reduced set	Full set	Reduced set
				Continued from previous page	
	P38A	0.74±0.017	0.73±0.021	0.23±0.028	0.21±0.025
	CDK2	0.74±0.034	0.76±0.031	0.34±0.033	0.34±0.044
	GSK3B	0.83±0.021	0.83±0.016	0.41±0.045	0.36±0.037
	Average	0.79±0.016	0.79±0.018	0.237±0.029	0.229±0.031
AGC	PKACA	0.81±0.007	0.81±0.007	0.25±0.014	0.24±0.023
AGC	PKCA	0.72±0.010	0.70±0.014	0.25±0.016	0.25±0.027
AGC	Akt1	0.82±0.012	0.80±0.022	0.40±0.049	0.43±0.037
AGC	PKCD	0.75±0.028	0.73±0.043	0.11±0.037	0.11±0.026
AGC	p90RSK	0.87±0.014	0.82±0.019	0.24±0.041	0.22±0.035
AGC	RSK2	0.76±0.043	0.79±0.040	0.17±0.050	0.00±0.000
AGC	PKG1	0.66±0.042	0.68±0.015	0.00±0.000	0.00±0.000
AGC	p70S6K	0.88±0.029	0.89±0.033	0.39±0.062	0.39±0.063
AGC	PKCZ	0.69±0.095	0.68±0.057	0.29±0.114	0.29±0.086
AGC	PKCE	0.54±0.043	0.53±0.057	0.44±0.102	0.44±0.089
	Average	0.75±0.032	0.74±0.031	0.254±0.049	0.237±0.039
TK	Src	0.59±0.012	0.57±0.014	0.24±0.014	0.22±0.020
TK	Fyn	0.64±0.019	0.63±0.013	0.31±0.027	0.30±0.024
TK	Abl	0.52±0.041	0.49±0.036	0.15±0.008	0.15±0.038
TK	Lyn	0.65±0.027	0.65±0.016	0.29±0.030	0.28±0.027
TK	Lck	0.65±0.064	0.64±0.044	0.31±0.065	0.30±0.061
TK	Syk	0.71±0.014	0.70±0.024	0.60±0.024	0.57±0.055
	Average	0.63±0.03	0.61±0.024	0.318±0.028	0.302±0.038

Table B.12: Comparison of prediction accuracy across mouse kinases between predicting kinase-specific phosphorylation sites using the sequence model trained on the full data-set, and when the model is trained on the similarity-reduced data-set. Prediction accuracy is calculated on the similarity-reduced data-set. If a kinase could not be trained on the reduced data-set due to too few positive training samples it was marked as "N/A". Kinases are grouped according to their family, with the average prediction accuracy for each family included. Results were generated using ten-fold cross-validation repeated across ten randomised data-set splits. Shown are the average and standard deviation of the AUC and AUC50 values.

		AUC		AUC50	
	Kinase	Full set	Reduced set	Full set	Reduced set
CMGC	CDC28	0.93±0.001	0.93±0.001	0.30±0.012	0.30±0.010
CMGC	CTK1	0.53±0.012	0.68±0.018	0.15±0.000	0.16±0.014
CMGC	MCK1	0.83±0.009	0.81±0.015	0.35±0.024	0.33±0.041
CMGC	PHO85	0.71±0.018	0.69±0.023	0.17±0.010	0.18±0.013

Continued on next page

	Kinase	Full set	Reduced set	Full set	Reduced set
				Continued from previous page	
	SSN3	0.73±0.058	0.74±0.051	0.29±0.064	0.29±0.066
	HOG1	0.79±0.046	0.82±0.024	0.30±0.052	0.35±0.030
	KNS1	0.93±0.038	0.95±0.038	0.59±0.056	0.68±0.088
	SLT2	0.68±0.037	0.69±0.041	0.27±0.047	0.28±0.029
	FUS3	0.54±0.035	0.54±0.046	0.22±0.004	0.22±0.065
	Average	0.74±0.028	0.76±0.028	0.293±0.03	0.31±0.04
AGC	TPK1	0.96±0.002	0.96±0.002	0.38±0.012	0.38±0.015
AGC	TPK3	0.80±0.038	0.82±0.034	0.57±0.062	0.62±0.047
AGC	YPK1	0.76±0.049	0.74±0.061	0.49±0.095	0.49±0.068
AGC	PKH2	0.74±0.037	0.74±0.040	0.25±0.003	0.25±0.074
AGC	PKH1	0.98±0.006	0.98±0.030	0.75±0.000	0.75±0.075
AGC	PKC1	0.88±0.024	0.88±0.040	0.35±0.045	0.37±0.062
	Average	0.85±0.026	0.85±0.034	0.464±0.036	0.476±0.057
TK	SNF1	0.78±0.014	0.78±0.019	0.16±0.032	0.16±0.027
TK	FRK1	0.75±0.021	0.74±0.043	0.42±0.048	0.42±0.054
TK	PSK2	0.74±0.047	0.74±0.035	0.41±0.055	0.45±0.066
TK	DUN1	0.85±0.013	0.85±0.013	0.38±0.012	0.35±0.051
	Average	0.78±0.024	0.78±0.027	0.34±0.037	0.35±0.049
Other	CKA1	0.89±0.005	0.90±0.007	0.31±0.015	0.32±0.019
Other	CKA2	0.91±0.007	0.91±0.004	0.36±0.017	0.36±0.023
Other	MPS1	0.86±0.016	0.86±0.021	0.23±0.036	0.22±0.038
Other	PTK1	0.67±0.015	0.67±0.024	0.14±0.020	0.15±0.012
Other	PTK2	0.89±0.046	0.87±0.028	0.75±0.065	0.75±0.087
Other	IPL1	0.91±0.009	0.91±0.016	0.28±0.018	0.29±0.023
Other	BUD32	0.73±0.063	0.78±0.043	0.39±0.071	0.40±0.059
	Average	0.86±0.023	0.84±0.021	0.351±0.035	0.355±0.037

TABLE B.13: Combined model accuracy across **human** kinases when compared to the context only model. Kinases are grouped according to their family, with the average prediction accuracy for each family included. Table shows accuracy values for classifying kinase substrates with both models as determined by 10-fold cross-validation across 10 randomised data-set splits. Prediction accuracy is shown using median and standard deviation of the AUC and AUC50 across the data-set splits.

	AUC		AUC50	
Kinase	Context model	Combined model	Context model	Combined model
CDK2	0.69±0.003	0.76±0.002	0.097±0.0016	0.110±0.0024

Continued on next page

	Kinase	Context model	Combined model	Context model	Combined model
					Continued from previous page
CMGC	CDK1	0.77±0.002	0.79±0.002	0.088±0.0035	0.101±0.0036
	ERK2	0.74±0.002	0.78±0.003	0.139±0.0022	0.155±0.0047
	ERK1	0.78±0.003	0.81±0.003	0.125±0.0021	0.147±0.0048
	GSK3B	0.74±0.002	0.79±0.005	0.151±0.0015	0.178±0.0032
	P38A	0.80±0.003	0.80±0.006	0.132±0.0012	0.167±0.0115
	JNK1	0.84±0.002	0.87±0.010	0.263±0.0021	0.310±0.0097
	CDK5	0.78±0.006	0.82±0.007	0.183±0.0059	0.230±0.0081
	JNK2	0.83±0.008	0.89±0.022	0.216±0.0113	0.313±0.0247
	CDK7	0.93±0.034	0.95±0.048	0.560±0.0117	0.705±0.0327
	GSK3A	0.81±0.042	0.91±0.028	0.378±0.0258	0.610±0.0551
	CDK4	0.87±0.002	0.88±0.006	0.309±0.0263	0.494±0.0219
	P38B	0.78±0.071	0.75±0.058	0.198±0.0330	0.410±0.0466
	HIPK2	0.89±0.033	0.98±0.054	0.365±0.0155	0.780±0.0618
	DYRK1A	0.92±0.032	0.90±0.015	0.698±0.0361	0.617±0.0257
	CDK9	0.96±0.045	0.90±0.043	0.548±0.0175	0.656±0.0348
	DYRK2	0.63±0.038	0.91±0.010	0.363±0.0098	0.849±0.0552
	ERK5	0.82±0.078	0.97±0.141	0.549±0.0270	0.709±0.1387
	CDK6	0.83±0.012	0.82±0.010	0.539±0.0201	0.698±0.0172
	CDK3	0.54±0.047	0.57±0.064	0.284±0.0473	0.407±0.0822
	Average	0.80±0.023	0.84±0.027	0.31±0.015	0.43±0.032
AGC	PKACA	0.65±0.002	0.68±0.003	0.060±0.0004	0.064±0.0027
	PKCA	0.69±0.002	0.71±0.004	0.070±0.0017	0.086±0.0046
	Akt1	0.78±0.002	0.81±0.004	0.181±0.0037	0.225±0.0035
	PKCD	0.65±0.004	0.65±0.008	0.116±0.0023	0.135±0.0053
	PKG1	0.83±0.010	0.84±0.020	0.335±0.0064	0.421±0.0342
	p90RSK	0.88±0.004	0.88±0.010	0.242±0.0083	0.334±0.0205
	PKCE	0.70±0.009	0.76±0.012	0.030±0.0051	0.205±0.0255
	PKCZ	0.71±0.005	0.68±0.012	0.136±0.0084	0.199±0.0182
	PKCB	0.68±0.009	0.76±0.018	0.166±0.0121	0.194±0.0214
	RSK2	0.77±0.006	0.82±0.016	0.290±0.0041	0.364±0.0378
	ROCK1	0.84±0.008	0.89±0.016	0.392±0.0097	0.571±0.0458
	PDK1	0.94±0.029	0.97±0.030	0.402±0.0171	0.767±0.0231
	PKCT	0.80±0.013	0.78±0.011	0.225±0.0056	0.312±0.0499
	PKCG	0.72±0.027	0.79±0.011	0.199±0.0325	0.270±0.0474
	p70S6K	0.89±0.018	0.91±0.030	0.398±0.0035	0.502±0.0234
	SGK1	0.83±0.015	0.87±0.028	0.254±0.0139	0.342±0.0209
	Akt2	0.84±0.049	0.80±0.029	0.119±0.0212	0.237±0.0614
	GRK2	0.88±0.014	0.68±0.050	0.323±0.0113	0.385±0.0874
	ROCK2	0.48±0.061	0.66±0.067	0.181±0.0195	0.193±0.0194

Continued on next page

	Kinase	Context model	Combined model	Context model	Combined model
					Continued from previous page
	PKCI	0.50±0.112	0.77±0.114	0.101±0.0604	0.541±0.0872
	PKCH	0.72±0.042	0.98±0.053	0.376±0.0294	0.738±0.0460
	PKN1	0.46±0.102	0.58±0.085	0.130±0.0507	0.330±0.0777
	Average	0.74±0.025	0.79±0.029	0.21±0.015	0.34±0.035
TK	Src	0.75±0.002	0.78±0.002	0.062±0.0018	0.063±0.0023
	Abl	0.85±0.003	0.86±0.005	0.153±0.0026	0.171±0.0043
	Fyn	0.78±0.005	0.81±0.007	0.110±0.0018	0.118±0.0048
	Lck	0.84±0.004	0.85±0.007	0.172±0.0082	0.190±0.0089
	Lyn	0.77±0.010	0.83±0.010	0.104±0.0031	0.169±0.0188
	EGFR	0.76±0.010	0.84±0.017	0.110±0.0143	0.145±0.0253
	Syk	0.81±0.015	0.90±0.009	0.324±0.0060	0.444±0.0354
	InsR	0.82±0.019	0.87±0.007	0.378±0.0182	0.456±0.0263
	JAK2	0.83±0.019	0.84±0.018	0.422±0.0111	0.476±0.0209
	FAK	0.81±0.033	0.83±0.040	0.295±0.0291	0.533±0.0554
	Ret	0.91±0.001	0.91±0.003	0.493±0.0127	0.598±0.0558
	Arg	0.77±0.068	0.74±0.044	0.400±0.0391	0.434±0.0757
	Brk	0.81±0.038	0.81±0.042	0.613±0.0624	0.556±0.0583
	ALK	0.80±0.104	0.76±0.120	0.296±0.0183	0.375±0.1237
	Btk	0.79±0.010	0.81±0.013	0.469±0.0223	0.654±0.0304
	PDGFRB	0.86±0.010	0.89±0.021	0.532±0.0125	0.764±0.0482
	JAK3	0.71±0.050	0.73±0.033	0.511±0.0549	0.606±0.0430
	Hck	0.84±0.090	0.79±0.054	0.291±0.0205	0.389±0.0408
	Pyk2	0.84±0.014	0.76±0.037	0.243±0.0458	0.320±0.0555
	Average	0.81±0.027	0.82±0.026	0.31±0.02	0.39±0.039
CAMK	CAMK2A	0.67±0.015	0.69±0.011	0.057±0.0108	0.153±0.0212
	Chk1	0.77±0.008	0.78±0.007	0.161±0.0028	0.172±0.0128
	AMPKA1	0.76±0.010	0.79±0.009	0.132±0.0111	0.217±0.0064
	MAPKAPK2	0.81±0.007	0.83±0.013	0.302±0.0062	0.365±0.0313
	PKD1	0.68±0.009	0.70±0.018	0.145±0.0086	0.197±0.0177
	LKB1	0.86±0.009	0.97±0.005	0.446±0.0073	0.840±0.0089
	MSK1	0.78±0.057	0.72±0.031	0.354±0.0399	0.433±0.0538
	Chk2	0.86±0.009	0.89±0.011	0.314±0.0154	0.382±0.0172
	Pim1	0.80±0.039	0.94±0.068	0.422±0.0231	0.564±0.0522
	AMPKA2	0.31±0.075	0.64±0.024	0.000±0.0000	0.291±0.0295
	MARK2	0.88±0.058	0.91±0.060	0.478±0.0236	0.608±0.0464
	CAMK1A	0.72±0.082	0.61±0.056	0.495±0.0482	0.283±0.0429
	DAPK3	0.71±0.063	0.88±0.036	0.442±0.0155	0.825±0.0776
	CaMK4	0.43±0.060	0.62±0.046	0.226±0.0396	0.424±0.0050

Continued on next page

	Kinase	Context model	Combined model	Context model	Combined model
					Continued from previous page
	PKD2	0.42±0.081	0.62±0.081	0.000±0.0000	0.284±0.0693
	CAMK2D	0.16±0.046	0.57±0.038	0.000±0.0000	0.332±0.0495
	Average	0.66±0.039	0.76±0.032	0.25±0.016	0.40±0.034
Other	CK2A1	0.73±0.003	0.77±0.002	0.116±0.0013	0.156±0.0043
	PLK1	0.81±0.010	0.82±0.007	0.143±0.0055	0.131±0.0067
	AurB	0.78±0.014	0.85±0.010	0.183±0.0139	0.168±0.0177
	AurA	0.73±0.011	0.74±0.015	0.175±0.0104	0.198±0.0136
	PLK3	0.89±0.026	0.84±0.025	0.419±0.0230	0.688±0.0414
	IKKA	0.84±0.023	0.81±0.022	0.515±0.0052	0.583±0.0093
	IKKB	0.89±0.005	0.87±0.019	0.322±0.0075	0.530±0.0323
	TBK1	0.99±0.004	0.99±0.001	0.735±0.0329	0.752±0.0320
	CK2A2	0.83±0.071	0.75±0.056	0.324±0.0162	0.625±0.0500
	IKKE	0.78±0.135	0.85±0.125	0.557±0.1298	0.554±0.1165
	TTK	0.69±0.107	0.71±0.108	0.201±0.0920	0.579±0.1296
	NEK6	0.55±0.011	0.67±0.058	0.447±0.0042	0.321±0.0415
	NEK2	0.93±0.028	0.91±0.036	0.552±0.0404	0.762±0.0805
	Average	0.80±0.034	0.81±0.037	0.36±0.029	0.47±0.044
STE	PAK1	0.76±0.025	0.73±0.011	0.191±0.0104	0.182±0.0121
	Cot	0.84±0.103	0.85±0.116	0.159±0.0380	0.593±0.1458
	MST1	0.63±0.047	0.65±0.036	0.436±0.0289	0.307±0.0396
	ASK1	0.88±0.109	0.94±0.118	0.681±0.0982	0.784±0.1428
	MKK4	0.70±0.033	0.90±0.036	0.428±0.0445	0.868±0.0556
	MST2	0.85±0.038	0.84±0.046	0.780±0.0466	0.697±0.0595
	PAK2	0.66±0.053	0.80±0.051	0.143±0.0488	0.423±0.0513
	MKK7	0.65±0.094	0.85±0.129	0.375±0.0615	0.820±0.1301
	MEK1	0.60±0.026	0.60±0.026	0.451±0.0057	0.455±0.0114
	Average	0.73±0.059	0.80±0.063	0.40±0.043	0.57±0.072
CK1	CK1A	0.76±0.019	0.78±0.016	0.290±0.0236	0.204±0.0333
	CK1D	0.75±0.051	0.83±0.031	0.315±0.0278	0.379±0.0544
	CK1E	0.80±0.055	0.95±0.037	0.364±0.0592	0.560±0.0664
	VRK1	0.85±0.014	0.68±0.028	0.583±0.0184	0.493±0.0157
	Average	0.79±0.035	0.81±0.028	0.39±0.032	0.41±0.042
Atypical	ATM	0.83±0.011	0.86±0.013	0.242±0.0042	0.302±0.0054
	ATR	0.90±0.024	0.89±0.027	0.391±0.0081	0.478±0.0161
	DNAPK	0.92±0.003	0.93±0.005	0.314±0.0050	0.404±0.0120
	mTOR	0.75±0.020	0.88±0.011	0.504±0.0037	0.624±0.0275
	Average	0.85±0.015	0.89±0.014	0.36±0.005	0.45±0.015

TABLE B.14: Combined model accuracy across **mouse** kinases when compared to the context only model. Kinases are grouped according to their family, with the average prediction accuracy for each family included. Table shows accuracy values for classifying kinase substrates with both models as determined by 10-fold cross-validation across 10 randomised data-set splits. Prediction accuracy is shown using median and standard deviation of the AUC and AUC50 across the data-set splits.

		AUC		AUC50	
	Kinase	Context model	Combined model	Context model	Combined model
CMGC	ERK2	0.73±0.010	0.77±0.012	0.269±0.0030	0.280±0.0113
	ERK1	0.73±0.013	0.70±0.014	0.301±0.0064	0.341±0.0141
	CDK5	0.61±0.015	0.70±0.017	0.329±0.0076	0.246±0.0333
	CDK1	0.79±0.013	0.79±0.013	0.413±0.0061	0.496±0.0234
	JNK1	0.71±0.009	0.76±0.015	0.414±0.0048	0.453±0.0428
	P38A	0.72±0.011	0.80±0.027	0.350±0.0189	0.446±0.0445
	CDK2	0.86±0.003	0.92±0.047	0.608±0.0223	0.724±0.0795
	GSK3B	0.69±0.015	0.86±0.016	0.377±0.0044	0.576±0.0331
	Average	0.73±0.011	0.79±0.02	0.38±0.009	0.45±0.035
AGC	PKACA	0.45±0.022	0.61±0.009	0.107±0.0044	0.128±0.0209
	PKCA	0.44±0.020	0.54±0.014	0.070±0.0102	0.118±0.0217
	Akt1	0.73±0.006	0.83±0.012	0.141±0.0154	0.459±0.0443
	PKCD	0.65±0.028	0.61±0.029	0.270±0.0141	0.296±0.0441
	p90RSK	0.28±0.052	0.61±0.020	0.000±0.0000	0.222±0.0020
	RSK2	0.49±0.021	0.58±0.079	0.346±0.0056	0.427±0.0852
	PKG1	0.19±0.052	0.41±0.067	0.000±0.0000	0.167±0.0500
	p70S6K	0.42±0.094	0.65±0.094	0.287±0.0865	0.292±0.0965
	PKCZ	0.56±0.020	0.74±0.039	0.435±0.0051	0.490±0.0829
	PKCE	0.55±0.016	0.76±0.070	0.385±0.0101	0.489±0.1074
	Average	0.48±0.033	0.63±0.043	0.20±0.015	0.31±0.056
TK	Src	0.79±0.012	0.85±0.011	0.311±0.0039	0.362±0.0068
	Fyn	0.64±0.011	0.78±0.031	0.151±0.0273	0.553±0.0550
	Abl	0.41±0.025	0.62±0.036	0.176±0.0086	0.211±0.0517
	Lyn	0.83±0.044	0.81±0.033	0.460±0.0269	0.595±0.0354
	Lck	0.81±0.114	0.94±0.162	0.434±0.0521	0.731±0.1625
	Syk	0.18±0.062	0.68±0.037	0.000±0.0000	0.332±0.0000
	Average	0.61±0.045	0.78±0.052	0.26±0.02	0.46±0.052

TABLE B.15: Combined model accuracy across **yeast** kinases when compared to the context only model. Kinases are grouped according to their family, with the average prediction accuracy for each family included. Table shows accuracy values for classifying kinase substrates with both models as determined by 10-fold cross-validation across 10 randomised data-set splits. Prediction accuracy is shown using median and standard deviation of the AUC and AUC50 across the data-set splits.

		AUC		AUC50	
	Kinase	Context model	Combined model	Context model	Combined model
CMGC	CDC28	0.63±0.003	0.76±0.003	0.148±0.0033	0.274±0.0082
	CTK1	0.46±0.021	0.48±0.027	0.041±0.0119	0.079±0.0188
	MCK1	0.73±0.038	0.84±0.034	0.303±0.0104	0.427±0.0264
	PHO85	0.83±0.012	0.81±0.012	0.449±0.0207	0.396±0.0387
	SSN3	0.54±0.018	0.85±0.035	0.176±0.0180	0.667±0.0531
	HOG1	0.85±0.003	0.79±0.020	0.463±0.0121	0.551±0.0375
	KNS1	0.41±0.044	0.77±0.054	0.000±0.0000	0.500±0.0573
	SLT2	0.78±0.116	0.79±0.135	0.211±0.0710	0.571±0.1434
	FUS3	0.66±0.040	0.71±0.055	0.161±0.0299	0.500±0.0667
	Average	0.65±0.033	0.76±0.042	0.22±0.02	0.44±0.05
AGC	TPK1	0.75±0.007	0.73±0.006	0.349±0.0092	0.333±0.0138
	TPK3	0.16±0.026	0.70±0.045	0.000±0.0000	0.583±0.0472
	YPK1	0.46±0.033	0.74±0.033	0.000±0.0000	0.390±0.0367
	PKH2	0.41±0.131	0.44±0.125	0.236±0.0953	0.097±0.0462
	PKH1	0.84±0.048	0.84±0.048	0.820±0.0513	0.818±0.0456
	PKC1	0.81±0.015	0.81±0.029	0.129±0.0609	0.665±0.0991
	Average	0.57±0.043	0.71±0.048	0.26±0.036	0.48±0.048
CAMK	SNF1	0.64±0.018	0.72±0.027	0.183±0.0149	0.217±0.0244
	FRK1	0.57±0.085	0.68±0.021	0.109±0.0350	0.301±0.0387
	PSK2	0.80±0.016	0.77±0.023	0.143±0.0408	0.488±0.0643
	DUN1	0.55±0.025	0.61±0.014	0.150±0.0243	0.328±0.0187
	Average	0.64±0.036	0.70±0.021	0.15±0.029	0.33±0.036
Other	CKA1	0.76±0.033	0.77±0.024	0.253±0.0177	0.280±0.0192
	CKA2	0.79±0.018	0.78±0.011	0.226±0.0074	0.307±0.0257
	MPS1	0.80±0.020	0.79±0.017	0.372±0.0069	0.397±0.0082
	PTK1	0.42±0.025	0.62±0.023	0.036±0.0140	0.165±0.0249
	PTK2	0.54±0.066	0.99±0.084	0.201±0.0605	0.888±0.0651
	IPL1	0.71±0.056	0.72±0.100	0.373±0.0234	0.371±0.0462
	BUD32	0.21±0.037	0.60±0.054	0.000±0.0000	0.426±0.0424
	Average	0.60±0.036	0.75±0.045	0.21±0.019	0.40±0.033

TABLE B.16: Sensitivity differences for kinases at **99.9% specificity**, where kinases are grouped according to their family, with the average sensitivity difference for each family included. The sensitivity difference between PhosphoPICK and each alternative method was measured for predicting kinase-specific phosphorylation sites out of all potential phosphorylation sites in our set of substrates. If we were unable to identify predictions for a kinase, it was marked as "N/A".

		Sensitivity difference between PhosphoPICK and alternative			
	Kinase	Sequence model	GPS	NetPhorest	NetworKIN
CMGC	CDK2	0.0009±0.0031	0.0371±0.0126	0.0205±0.0126	-0.0329±0.0126
	CDK1	0.0625±0.0057	0.0920±0.0115	0.0684±0.0115	-0.0523±0.0115
	ERK2	0.0166±0.0132	0.0238±0.0111	0.0151±0.0111	-0.0746±0.0111
	ERK1	0.0341±0.0193	0.0606±0.0197	0.0312±0.0197	-0.0512±0.0197
	GSK3B	0.0558±0.0114	0.0240±0.0161	0.0473±0.0161	0.0085±0.0161
	P38A	0.0405±0.0186	0.1230±0.0202	0.1500±0.0202	0.1390±0.0202
	JNK1	0.0624±0.0311	0.1680±0.0253	0.1800±0.0253	0.1090±0.0253
	CDK5	0.0161±0.0161	0.0500±0.0168	-0.0145±0.0168	-0.1270±0.0168
	JNK2	-0.0143±0.0293	0.0171±0.0229	0.0457±0.0229	-0.0114±0.0229
	CDK7	0.0160±0.0408	-0.1840±0.0408	-0.0240±0.0408	-0.1440±0.0408
	GSK3A	0.1120±0.0176	0.3820±0.0474	0.3240±0.0474	0.3820±0.0474
	CDK4	0.1520±0.0307	0.1740±0.0307	0.3040±0.0307	0.3040±0.0307
	P38B	0.1830±0.0660	0.2610±0.0660	0.3720±0.0660	0.3720±0.0660
	HIPK2	0.1300±0.0812	N/A	0.5050±0.0723	0.4380±0.0723
	DYRK1A	0.3200±0.0400	0.440±0.0327	N/A	N/A
	CDK9	0.0407±0.0452	N/A	N/A	N/A
	DYRK2	0.4870±0.0875	N/A	N/A	N/A
	ERK5	0.2140±0.0714	0.1140±0.0857	0.3520±0.0857	0.1620±0.0857
	CDK6	0.2830±0.0428	0.3500±0.0373	0.4500±0.0373	0.3830±0.0373
	CDK3	0.2250±0.0935	N/A	0.5620±0.0839	0.3120±0.0839
	Average	0.1220±0.0382	0.1330±0.0310	0.1990±0.0365	0.1250±0.0365
AGC	PKACA	0.0398±0.0121	0.0587±0.0095	0.0180±0.00947	-0.1070±0.0095
	PKCA	0.0104±0.0064	0.0367±0.0138	0.0033±0.0138	-0.0004±0.0138
	Akt1	0.0732±0.0222	N/A	0.0098±0.0147	-0.0098±0.0147
	PKCD	0.0156±0.0124	0.0156±0.0124	0.0267±0.0124	0.0267±0.0124
	PKG1	0.0533±0.0371	0.0400±0.0359	0.1730±0.0359	0.0400±0.0359
	p90RSK	0.0632±0.0268	0.2130±0.0415	0.1610±0.0415	0.1340±0.0415
	PKCE	0.0025±0.0075	-0.0225±0.0075	0.1020±0.0075	0.0275±0.0075
	PKCZ	0.0178±0.0133	-0.0244±0.0306	0.1310±0.0306	0.0867±0.0306
	PKCB	0.0103±0.0235	0.0538±0.0179	0.0538±0.0179	0.0795±0.0179
	RSK2	0.0968±0.0323	-0.0226±0.0355	0.1710±0.0355	-0.0226±0.0355
	ROCK1	0.0260±0.0180	-0.0200±0.0390	0.0800±0.0390	0.0200±0.0390
	PDK1	0.0276±0.0207	-0.0552±0.0442	0.2900±0.0442	-0.0207±0.0442

Continued on next page

	Kinase	Sequence model	GPS	NetPhorest	NetworKIN
					Continued from previous page
	PKCT	0.0000±0.0000	-0.0708±0.0458	0.0542±0.0458	-0.0708±0.0458
	PKCG	0.0231±0.0188	-0.1040±0.0517	0.0885±0.0517	0.0500±0.0517
	p70S6K	0.0636±0.0370	0.1940±0.0411	0.1330±0.0411	-0.0182±0.0411
	SGK1	-5e-14±0.0421	0.2230±0.0414	0.0692±0.0414	-0.1230±0.0414
	Akt2	0.2000±0.0516	0.0333±0.0683	0.1000±0.0683	0.1000±0.0683
	GRK2	0.0026±0.0079	0.4890±0.0376	0.4630±0.0376	0.1740±0.0376
	ROCK2	0.0000±0.0000	N/A	0.1820±0.0000	0.0909±1.39e-17
	PKCI	-0.0167±0.0333	N/A	0.0500±0.0553	0.1330±0.0553
	PKCH	0.2330±0.0745	0.3070±0.0680	0.7070±0.0680	0.5070±0.0680
	PKN1	0.0500±0.0764	N/A	N/A	N/A
	Average	0.0451±0.0261	0.0747±0.0356	0.1460±0.0339	0.0522±0.0339
TK	Src	0.0081±0.0079	-0.0152±0.0084	-0.0011±0.0084	-0.0187±0.0084
	Abl	0.0176±0.0164	-0.0228±0.0130	0.0327±0.0130	0.0438±0.0130
	Fyn	0.0056±0.0124	0.0022±0.00667	0.0022±0.0067	-0.0200±0.0067
	Lck	0.0260±0.0114	-0.0151±0.0207	0.0260±0.0207	-0.0699±0.0207
	Lyn	-0.0020±0.0163	-0.0863±0.00961	0.0314±0.0096	-0.0078±0.0096
	EGFR	0.0122±0.0100	-0.1860±0.0143	0.0184±0.0143	-0.1240±0.0143
	Syk	0.1160±0.0329	-0.0047±0.0357	0.2740±0.0357	0.2740±0.0357
	InsR	0.0343±0.0308	0.2460±0.0343	0.3600±0.0343	0.3310±0.0343
	JAK2	0.0000±0.0000	0.0129±0.0214	N/A	N/A
	FAK	0.1190±0.0519	0.3750±0.0791	N/A	N/A
	Ret	0.0370±0.0331	-0.3110±0.0474	N/A	N/A
	Arg	0.2000±0.0408	0.0182±0.0408	0.1090±0.0408	-0.0727±0.0408
	Brk	0.0000±0.0000	0.0000±0.0000	0.1430±0.0000	-0.0714±1.39e-17
	ALK	0.0000±0.0000	-0.2220±0.0000	N/A	N/A
	Btk	-0.1380±0.0462	-0.2310±2.78e-17	0.0000±0.0000	-0.0769±1.39e-17
	PDGFRB	0.0261±0.0288	0.1610±0.0552	0.1610±0.0552	0.1170±0.0552
	JAK3	0.0312±0.0576	0.1500±0.0800	N/A	N/A
	Hck	0.0850±0.0391	-0.2150±0.0391	0.0850±0.0391	0.0850±0.0391
	Pyk2	0.2290±0.1140	0.0857±0.1140	N/A	N/A
	Average	0.0424±0.0289	-0.0136±0.0326	0.0955±0.0214	0.0300±0.0214
CAMK	CAMK2A	0.0397±0.0191	0.0159±0.0159	0.0450±0.0159	-0.0697±0.0159
	Chk1	0.0102±0.0137	-0.0204±0.0241	N/A	N/A
	AMPKA1	0.0511±0.0195	0.0170±0.0266	-0.0255±0.0266	-0.0894±0.0266
	MAPKAPK2	0.0364±0.0253	-0.0545±0.0396	N/A	N/A
	PKD1	0.0511±0.0217	-0.0319±0.0238	0.0957±0.0238	0.0532±0.0238
	LKB1	0.0290±0.0097	0.1840±0.0207	0.5970±0.0207	0.3660±0.0207
	MSK1	0.3000±0.0856	N/A	N/A	N/A

Continued on next page

	Kinase	Sequence model	GPS	NetPhorest	NetworKIN
					Continued from previous page
	Chk2	0.0560±0.0215	-0.056±0.0307	N/A	N/A
	Pim1	0.0609±0.0651	N/A	0.2700±0.0696	0.1390±0.0696
	AMPKA2	0.2120±0.0288	0.2060±0.0395	0.2060±0.0395	0.2650±0.0395
	MARK2	0.1080±0.0534	N/A	N/A	N/A
	CAMK1A	0.0222±0.0667	0.4440±0.0000	0.4440±0.0000	0.4440±0.0000
	DAPK3	0.4310±0.0705	0.2850±0.0846	0.5150±0.0846	0.2300±0.0846
	CaMK4	0.0500±0.0829	-0.2000±0.0829	-0.2000±0.0829	-0.0750±0.0829
	PKD2	0.2250±0.0500	N/A	0.2250±0.0500	-0.0250±0.0500
	CAMK2D	0.2120±0.0800	N/A	0.3380±0.0800	0.4630±0.0800
	Average	0.1180±0.0446	0.0717±0.0353	0.2280±0.0449	0.1550±0.0449
Other	CK2A1	0.0206±0.0036	0.0714±0.0052	0.0775±0.0052	-0.0435±0.0052
	PLK1	0.0284±0.0120	0.0157±0.0118	N/A	N/A
	AurB	0.0480±0.0148	-0.0040±0.0104	N/A	N/A
	AurA	0.0056±0.0208	-0.0056±0.0299	-0.0056±0.0299	-0.0611±0.0299
	PLK3	0.1320±0.0516	N/A	N/A	N/A
	IKKA	0.0759±0.0371	0.0241±0.0438	0.2310±0.0438	0.0387±0.0438
	IKKB	0.0333±0.0282	0.2130±0.0345	0.3130±0.0345	0.2860±0.0345
	TBK1	0.1690±0.0462	N/A	N/A	N/A
	CK2A2	0.1880±0.0559	0.6130±0.0468	0.5500±0.0468	0.1750±0.0468
	IKKE	-0.0889±0.1430	N/A	N/A	N/A
	TTK	0.2440±0.1660	N/A	0.481±0.1650	0.4810±0.1650
	NEK6	0.1800±0.0748	0.1000±1.39e-17	N/A	N/A
	NEK2	-0.2420±0.1310	0.1170±0.0667	0.1170±0.0667	0.0333±0.0667
	Average	0.0610±0.0605	0.1270±0.0277	0.2520±0.0560	0.1300±0.0560
STE	PAK1	0.0107±0.0143	0.0179±0.0080	0.0357±0.0080	-0.1790±0.0080
	Cot	0.0778±0.0509	0.4830±0.1290	N/A	N/A
	MST1	0.1180±0.0372	N/A	0.2650±0.0395	0.1740±0.0395
	ASK1	0.1070±0.0659	N/A	N/A	N/A
	MKK4	0.3120±0.1010	N/A	0.7750±0.1220	0.1500±0.1220
	MST2	0.1500±0.0500	N/A	0.2630±0.0673	0.0403±0.0673
	PAK2	0.1690±0.0576	0.1920±0.1050	0.4230±0.1050	0.3460±0.1050
	MKK7	0.0500±0.0829	-0.0250±0.0500	0.8500±0.0500	0.4750±0.0500
	MEK1	0.0000±0.0000	0.2000±2.78e-17	0.4000±5.55e-17	-0.2000±2.78e-17
	Average	0.1110±0.0511	0.1740±0.0584	0.4300±0.0560	0.1150±0.0560
CK1	CK1A	0.0133±0.0109	-0.0267±0.0133	0.1510±0.0133	0.1400±0.0133
	CK1D	0.0216±0.0162	0.1860±0.0351	0.2410±0.0351	0.1860±0.0351
	CK1E	0.0652±0.0446	-0.0565±0.0516	0.4220±0.0516	0.0739±0.0516
	VRK1	0.2640±0.0636	0.4360±0.0545	N/A	N/A

	Average	0.0910±0.0338	0.1350±0.0387	0.2710±0.0334	0.1330±0.0334
Atypical	ATM	0.0866±0.0223	0.0779±0.0312	0.1420±0.0312	-0.0616±0.0312
	ATR	0.0885±0.0167	0.1180±0.0161	0.0689±0.0161	-0.0623±0.0161
	DNAPK	0.0242±0.0146	-0.0088±0.0176	0.1230±0.0176	0.1120±0.0176
	mTOR	0.0526±0.0235	0.1950±0.0443	N/A	N/A
	Average	0.0630±0.0193	0.0955±0.0273	0.1110±0.0216	-0.0040±0.0216

TABLE B.17: Sensitivity differences for kinases at **99% specificity**, where kinases are grouped according to their family, with the average sensitivity difference for each family included. The sensitivity difference between PhosphoPICK and each alternative method was measured for predicting kinase-specific phosphorylation sites out of all potential phosphorylation sites in our set of substrates. If we were unable to identify predictions for a kinase, it was marked as "N/A".

		sensitivity difference between PhosphoPICK and alternative			
	Kinase	Sequence model	GPS	NetPhorest	NetworKIN
CMGC	CDK2	0.0406±0.0081	0.0878±0.0242	-0.0023±0.0242	0.0023±0.0242
	CDK1	0.1850±0.0270	0.3290±0.0144	0.1510±0.0144	-0.0318±0.0144
	ERK2	0.1030±0.0173	-0.0570±0.0303	0.0355±0.0303	-0.1180±0.0303
	ERK1	0.1070±0.0244	-2e-15±0.0268	0.1120±0.0268	-0.1290±0.0268
	GSK3B	0.0775±0.0147	0.0814±0.0233	0.0581±0.0233	-0.0349±0.0233
	P38A	0.0910±0.0256	0.0748±0.0304	0.3160±0.0304	0.0743±0.0304
	JNK1	0.2250±0.0309	0.3530±0.0361	0.4820±0.0361	0.0824±0.0361
	CDK5	0.0548±0.0207	0.1060±0.0308	-0.1350±0.0308	-0.3130±0.0308
	JNK2	0.2400±0.0343	0.1110±0.0469	0.3110±0.0469	-0.2890±0.0469
	CDK7	0.4360±0.0631	0.1880±0.0256	0.3880±0.0256	0.2280±0.0256
	GSK3A	0.2530±0.0459	0.4180±0.0668	0.4180±0.0668	0.3000±0.0668
	CDK4	0.4370±0.0471	0.4130±0.0466	0.7610±0.0466	0.6740±0.0466
	P38B	0.2830±0.0678	0.2940±0.0434	0.5720±0.0434	0.5100±0.0434
	HIPK2	0.3100±0.0700	N/A	0.8400±0.0539	0.5730±0.0539
	DYRK1A	0.3730±0.0680	0.5270±0.0200	N/A	N/A
	CDK9	0.3960±0.0598	N/A	N/A	N/A
	DYRK2	0.6190±0.0763	N/A	N/A	N/A
	ERK5	0.5240±0.1280	0.3430±0.1490	0.6760±0.1490	0.0095±0.1490
	CDK6	0.5170±0.0522	0.6070±0.0133	0.8070±0.0133	0.4400±0.0133
	CDK3	0.1750±0.0612	N/A	0.0625±0.0839	-0.1880±0.0839
	Average	0.272±0.0471	0.242±0.0392	0.344±0.0439	0.105±0.0439
	PKACA	-0.0031±0.0163	-0.0938±0.0184	-0.0548±0.0184	-0.1390±0.0184
	PKCA	0.0263±0.0099	0.0230±0.0158	0.0304±0.0158	-0.0437±0.0158
	Akt1	0.0458±0.0155	N/A	-0.0856±0.0264	-0.0268±0.0264
	PKCD	0.0789±0.0153	-0.0322±0.0246	0.0122±0.0246	-0.0433±0.0246

Continued on next page

	Kinase	Sequence model	GPS	NetPhorest	NetworKIN
					Continued from previous page
AGC	PKG1	0.0900±0.0473	-0.0667±0.0365	0.200±0.0365	-0.0333±0.0365
	p90RSK	0.1500±0.0457	0.2110±0.0372	0.1320±0.0372	0.1580±0.0372
	PKCE	0.1950±0.0245	0.1130±0.0301	0.2120±0.0301	0.0125±0.0301
	PKCZ	0.0578±0.0178	0.0133±0.0267	0.0800±0.0267	-0.0089±0.0267
	PKCB	0.0872±0.0515	0.1900±0.0576	0.0615±0.0576	0.2670±0.0576
	RSK2	0.1900±0.0488	0.1740±0.0413	0.2390±0.0413	0.0774±0.0413
	ROCK1	0.3120±0.0634	0.1020±0.0648	0.3420±0.0648	0.0220±0.0648
	PDK1	0.1480±0.0269	0.0448±0.0269	0.1480±0.0269	-0.0241±0.0269
	PKCT	0.0958±0.0267	-0.0833±0.0527	0.0833±0.0527	-0.2080±0.0527
	PKCG	0.3310±0.0734	0.1620±0.0985	0.3150±0.0985	0.3150±0.0985
	p70S6K	0.1640±0.0545	0.1580±0.0182	0.0364±0.0182	0.0667±0.0182
	SGK1	0.1190±0.0607	0.0423±0.0607	0.0808±0.0607	-0.1120±0.0607
	Akt2	0.1930±0.0629	-0.0800±0.0653	-0.2800±0.0653	-0.0133±0.0653
	GRK2	0.2320±0.0349	0.4500±0.0299	0.7130±0.0299	0.1610±0.0299
	ROCK2	0.0909±1.39e-17	N/A	0.1820±0.0000	-0.2730±5.55e-17
	PKCI	0.4170±0.1180	N/A	0.4250±0.1260	0.4250±0.1260
	PKCH	0.1670±0.1090	0.4330±0.0683	0.6330±0.0683	0.3000±0.0683
	PKN1	0.0833±0.1540	N/A	N/A	N/A
	Average	0.1490±0.0489	0.0977±0.0430	0.1670±0.0441	0.0419±0.0441
TK	Src	0.0403±0.0117	-0.0929±0.0163	0.0131±0.0163	-0.0364±0.0163
	Abl	0.0319±0.0199	-0.0995±0.0294	0.0560±0.0294	0.0560±0.0294
	Fyn	0.0411±0.0186	-0.0011±0.0256	0.0433±0.0256	-0.1120±0.0256
	Lck	0.0795±0.0279	-0.1600±0.0253	0.0726±0.0253	-0.2420±0.0253
	Lyn	0.0333±0.0197	-0.2220±0.0197	0.0529±0.0197	-0.0647±0.0197
	EGFR	0.0449±0.0327	-0.3370±0.0345	-0.0306±0.0345	-0.3370±0.0345
	Syk	0.2860±0.0642	0.0140±0.0599	0.6420±0.0599	0.2520±0.0599
	InsR	0.0429±0.0263	0.0143±0.0367	0.3290±0.0367	0.0429±0.0367
	JAK2	0.1290±0.0289	0.0774±0.0214	N/A	N/A
	FAK	0.3190±0.0763	0.5750±0.0673	N/A	N/A
	Ret	0.1040±0.0363	-0.2070±0.0602	N/A	N/A
	Arg	0.4830±0.0972	0.2770±0.1070	0.2770±0.1070	0.0046±0.1070
	Brk	0.4930±0.0500	0.5430±0.0350	0.6860±0.0350	0.1140±0.0350
	ALK	0.4560±0.1160	0.2330±0.1160	N/A	N/A
	Btk	0.4150±0.0923	0.1620±0.0803	0.4690±0.0803	0.2380±0.0803
	PDGFRB	0.2090±0.0543	0.2090±0.0543	0.2520±0.0543	-0.0522±0.0543
	JAK3	0.2870±0.0893	0.1380±0.0545	N/A	N/A
	Hck	0.5000±0.0316	0.1300±0.0400	0.5300±0.0400	0.2300±0.0400
	Pyk2	0.3860±0.0915	0.2430±0.0915	N/A	N/A

Continued on next page

	Kinase	Sequence model	GPS	NetPhorest	NetworKIN
					Continued from previous page
	Average	0.2310±0.0518	0.0787±0.0513	0.2610±0.0434	0.0072±0.0434
CAMK	CAMK2A	0.0302±0.0132	-0.2680±0.0218	-0.0221±0.0218	-0.1040±0.0218
	Chk1	0.1140±0.0410	0.0449±0.0363	N/A	N/A
	AMPKA1	0.0723±0.0255	-0.0447±0.0322	0.0617±0.0322	-0.0872±0.0322
	MAPKAPK2	0.2480±0.0400	0.0250±0.0561	N/A	N/A
	PKD1	0.1040±0.0461	0.1280±0.0404	0.1700±0.0404	-0.0638±0.0404
	LKB1	0.1190±0.0521	0.2030±0.0541	0.5210±0.0541	0.1750±0.0541
	MSK1	0.2000±0.0789	N/A	N/A	N/A
	Chk2	0.1560±0.0496	-0.0700±0.0361	N/A	N/A
	Pim1	0.2870±0.0758	N/A	0.3780±0.0675	0.2480±0.0675
	AMPKA2	0.3410±0.0634	0.1650±0.0353	0.3410±0.0353	0.2240±0.0353
	MARK2	0.2830±0.0764	N/A	N/A	N/A
	CAMK1A	0.0000±0.0000	0.4440±0.0000	0.4440±0.0000	-0.1110±0.0000
	DAPK3	0.6540±0.0788	0.4310±0.0510	0.8920±0.0510	-0.1080±0.0510
	CaMK4	0.4750±0.0500	-0.1250±0.0000	0.1250±0.0000	0.1250±0.0000
	PKD2	0.1380±0.1180	N/A	0.0500±0.1000	-0.0750±0.1000
	CAMK2D	0.2250±0.0750	N/A	0.3500±0.0750	0.3500±0.0750
	Average	0.2150±0.0552	0.0848±0.0330	0.3010±0.0434	0.0520±0.0434
Other	CK2A1	-0.0009±0.0070	0.0206±0.0073	0.0575±0.0073	-0.0348±0.0073
	PLK1	0.0873±0.0203	0.0725±0.0229	N/A	N/A
	AurB	0.1290±0.0260	-0.0080±0.0208	N/A	N/A
	AurA	0.0778±0.0408	0.0556±0.0329	0.1110±0.0329	0.1110±0.0329
	PLK3	0.4730±0.0649	N/A	N/A	N/A
	IKKA	0.2690±0.0530	0.1860±0.0229	0.4540±0.0229	-0.0074±0.0229
	IKKB	0.3380±0.0377	0.5380±0.0324	0.6640±0.0324	0.3130±0.0324
	TBK1	0.4960±0.0631	N/A	N/A	N/A
	CK2A2	0.3440±0.0504	0.4250±0.0375	0.4250±0.0375	0.2370±0.0375
	IKKE	0.0833±0.0756	N/A	N/A	N/A
	TTK	0.3440±0.2440	N/A	0.5750±0.2270	0.4500±0.2270
	NEK6	0.1300±0.0458	0.0000±0.0000	N/A	N/A
	NEK2	0.3170±0.1280	0.7000±0.1190	0.6170±0.1190	0.2830±0.1190
	Average	0.2370±0.0659	0.2210±0.0329	0.4150±0.0684	0.1930±0.0684
STE	PAK1	0.1050±0.0168	-0.0536±0.0179	0.0000±0.0179	-0.2320±0.0179
	Cot	0.1330±0.0619	0.4170±0.1300	N/A	N/A
	MST1	0.2410±0.0412	N/A	0.3740±0.0176	0.1920±0.0176
	ASK1	0.3070±0.0848	N/A	N/A	N/A
	MKK4	0.3620±0.0375	N/A	0.7750±0.1220	-0.1000±0.1220
	MST2	0.5060±0.0813	N/A	0.6810±0.0187	0.2370±0.0188

Continued on next page

	Kinase	Sequence model	GPS	NetPhorest	NetworKIN
					Continued from previous page
	PAK2	0.2850±0.0913	0.4460±0.0462	0.5230±0.0462	0.2920±0.0462
	MKK7	0.0250±0.0750	0.0000±0.0000	0.8750±0.0000	0.0000±0.0000
	MEK1	0.1200±0.0980	0.1200±0.0980	0.5200±0.0980	-0.0800±0.0980
	Average	0.2320±0.0653	0.1860±0.0583	0.5350±0.0458	0.0441±0.0458
CK1	CK1A	0.0378±0.0218	0.0100±0.0265	0.1990±0.0265	0.0211±0.0265
CK1	CK1D	0.2920±0.0315	0.5220±0.0343	0.4950±0.0343	0.1700±0.0343
CK1	CK1E	0.3040±0.0802	0.2610±0.0550	0.6520±0.0550	0.2170±0.0550
CK1	VRK1	0.3090±0.0833	0.4270±0.0582	N/A	N/A
	Average	0.2360±0.0542	0.3050±0.0435	0.4490±0.0386	0.1360±0.0386
Atypical	ATM	0.0895±0.0258	0.0959±0.0203	0.2350±0.0203	-0.1250±0.0203
Atypical	ATR	0.3690±0.0499	0.2480±0.0405	0.3300±0.0405	-0.1790±0.0405
Atypical	DNAPK	0.1560±0.0343	0.0659±0.0269	0.3960±0.0269	0.0769±0.0269
Atypical	mTOR	0.3840±0.0591	0.4890±0.0542	N/A	N/A
	Average	0.2500±0.0423	0.2250±0.0355	0.3200±0.0292	-0.0756±0.0292

TABLE B.18: Gene ontology (GO) term enrichment analysis for predicted Akt1 substrates. Shown are all positions that the kinase was found to be significantly over-represented at.

position	GO term	Description	E-value
All	GO:0019901	protein kinase binding	0.0007
2	N/A	N/A	N/A
3	GO:0010907	positive regulation of glucose metabolic process	0.013
3	GO:0002053	positive regulation of mesenchymal cell proliferation	0.013
4	GO:0008543	fibroblast growth factor receptor signaling pathway	0.0017
4	GO:0019901	protein kinase binding	0.002
4	GO:0007173	epidermal growth factor receptor signalling pathway	0.022
4	GO:0032000	positive regulation of fatty acid beta-oxidation	0.007
5	GO:0090343	positive regulation of cell ageing	0.006
6	GO:0005158	insulin receptor binding	0.003
6	GO:0010907	positive regulation of glucose metabolic process	0.006
6	GO:0032000	positive regulation of fatty acid beta-oxidation	0.037

TABLE B.19: Gene ontology (GO) term enrichment analysis for predicted AMPKA1 substrates. Shown are all positions that the kinase was found to be significantly over-represented at.

position	GO term	Description	E-value
All	GO:0019901	protein kinase binding	0.0001
All	GO:0008543	fibroblast growth factor receptor signaling pathway	0.001
All	GO:0005829	cytosol	0.001
All	GO:0048011	neurotrophin TRK receptor signaling pathway	0.003
All	GO:0008286	insulin receptor signaling pathway	0.003
All	GO:0097149	centralspindlin complex	0.006
All	GO:0005158	insulin receptor binding	0.022
All	GO:0005737	cytoplasm	0.026
All	GO:0007173	epidermal growth factor receptor signaling pathway	0.036
All	GO:0006302	double-strand break repair	0.037
All	GO:0007049	cell cycle	0.039
All	GO:0007265	Ras protein signal transduction	0.042
All	GO:0005515	protein binding	0.048
3	GO:0010907	positive regulation of glucose metabolic process	0.014
3	GO:0002053	positive regulation of mesenchymal cell proliferation	0.014
4	GO:0019901	protein kinase binding	4.26e-05
4	GO:0008543	fibroblast growth factor receptor signaling pathway	0.0004
4	GO:0007173	epidermal growth factor receptor signaling pathway	0.0008
4	GO:0048011	neurotrophin TRK receptor signaling pathway	0.0011
4	GO:0038095	Fc-epsilon receptor signaling pathway	0.0014
4	GO:0048015	phosphatidylinositol-mediated signaling	0.0016
4	GO:0008286	insulin receptor signaling pathway	0.0029
4	GO:0060397	JAK-STAT cascade involved in growth hormone signaling pathway	0.007
4	GO:0005158	insulin receptor binding	0.023
4	GO:0010907	positive regulation of glucose metabolic process	0.029
4	GO:0005829	cytosol	0.036

TABLE B.20: Gene ontology (GO) term enrichment analysis for predicted AurB substrates. Shown are all positions that the kinase was found to be significantly over-represented at.

position	GO term	Description	E-value
All	GO:0005694	chromosome	1.47e-05
All	GO:0000786	nucleosome	0.0003
All	GO:0006334	nucleosome assembly	0.011

Continued on next page

position	GO term	Description	E-value
		Continued from previous page	
All	GO:0043065	positive regulation of apoptotic process	0.02
2	N/A	N/A	N/A
3	GO:0005694	chromosome	7.59e-08
3	GO:0000786	nucleosome	1.17e-06
3	GO:0006334	nucleosome assembly	2.49e-05
3	GO:0046982	protein heterodimerization activity	0.011
4	GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II	0.0007
4	GO:0007018	microtubule-based movement	0.003
4	GO:0097149	centralspindlin complex	0.022
4	GO:0051256	mitotic spindle midzone assembly	0.022
4	GO:0005874	microtubule	0.032

TABLE B.21: Gene ontology (GO) term enrichment analysis for predicted CDK2 substrates. Shown are all positions that the kinase was found to be significantly over-represented at.

position	GO term	Description	E-value
All	GO:0005694	chromosome	5.68e-05
All	GO:0007049	cell cycle	0.0008
All	GO:0005634	nucleus	0.011
All	GO:0006281	DNA repair	0.022
-4	N/A	N/A	N/A
-5	N/A	N/A	N/A
-6	N/A	N/A	N/A
-7	N/A	N/A	N/A

TABLE B.22: Gene ontology (GO) term enrichment analysis for predicted p70S6K substrates. Shown are all positions that the kinase was found to be significantly over-represented at.

position	GO term	Description	E-value
All	GO:0048011	neurotrophin TRK receptor signalling pathway	0.0002

Continued on next page

position	GO term	Description	E-value
		Continued from previous page	
All	GO:0008286	insulin receptor signalling pathway	0.003
All	GO:0007173	epidermal growth factor receptor signalling pathway	0.003
All	GO:0008543	fibroblast growth factor receptor signalling pathway	0.013
All	GO:0019901	protein kinase binding	0.037
3	GO:0010907	positive regulation of glucose metabolic process	0.008
3	GO:0008286	insulin receptor signalling pathway	0.01
4	GO:0008543	fibroblast growth factor receptor signalling pathway	2.78e-05
4	GO:0007173	epidermal growth factor receptor signalling pathway	6.85e-05
4	GO:0008286	insulin receptor signalling pathway	0.0002
4	GO:0038095	Fc-epsilon receptor signalling pathway	0.0027
4	GO:0048015	phosphatidylinositol-mediated signalling	0.0027
4	GO:0019901	protein kinase binding	0.018
4	GO:0048011	neurotrophin TRK receptor signalling pathway	0.031
4	GO:0005158	insulin receptor binding	0.033
4	GO:0090343	positive regulation of cell ageing	0.036
4	GO:0010907	positive regulation of glucose metabolic process	0.036
4	GO:0004871	signal transducer activity	0.049
5	GO:0006974	cellular response to DNA damage stimulus	0.0005
5	GO:0090343	positive regulation of cell ageing	0.0079
5	GO:0031465	Cul4B-RING E3 ubiquitin ligase complex	0.0079
5	GO:0006281	DNA repair	0.049
6	GO:0010907	positive regulation of glucose metabolic process	0.003
6	GO:0032000	positive regulation of fatty acid beta-oxidation	0.015
6	GO:0045725	positive regulation of glycogen biosynthetic process	0.025
6	GO:0043548	phosphatidylinositol 3-kinase binding	0.025
6	GO:0048011	neurotrophin TRK receptor signalling pathway	0.032
6	GO:0046326	positive regulation of glucose import	0.038
6	GO:0008286	insulin receptor signalling pathway	0.048

TABLE B.23: Gene ontology (GO) term enrichment analysis for predicted p90RSK substrates. Shown are all positions that the kinase was found to be significantly over-represented at.

position	GO term	Description	E-value
All	GO:0019901	protein kinase binding	0.0002
All	GO:0045087	innate immune response	0.003
All	GO:0048011	neurotrophin TRK receptor signalling pathway	0.018

Continued on next page

position	GO term	Description	E-value
		Continued from previous page	
All	GO:0038095	Fc-epsilon receptor signalling pathway	0.033
All	GO:0006974	cellular response to DNA damage stimulus	0.033
3	GO:0010907	positive regulation of glucose metabolic process	0.018
3	GO:0019901	protein kinase binding	0.0196
3	GO:0002053	positive regulation of mesenchymal cell proliferation	0.02
3	GO:0042169	SH2 domain binding	0.02
4	GO:0008543	fibroblast growth factor receptor signalling pathway	0.0012
4	GO:0007173	epidermal growth factor receptor signalling pathway	0.0027
4	GO:0019901	protein kinase binding	0.0033
4	GO:0048015	phosphatidylinositol-mediated signalling	0.0042
4	GO:0008286	insulin receptor signalling pathway	0.0077
4	GO:0006974	cellular response to DNA damage stimulus	0.027
4	GO:0010907	positive regulation of glucose metabolic process	0.041
4	GO:0005158	insulin receptor binding	0.042
-5	N/A	N/A	N/A

TABLE B.24: Gene ontology (GO) term enrichment analysis for predicted PAK1 substrates. Shown are all positions that the kinase was found to be significantly over-represented at.

position	GO term	Description	E-value
All	GO:0019901	protein kinase binding	1.1e-05
All	GO:0006915	apoptotic process	0.0003
All	GO:0050731	positive regulation of peptidyl-tyrosine phosphorylation	0.014
-2	N/A	N/A	N/A
3	GO:0010907	positive regulation of glucose metabolic process	0.016
3	GO:0002053	positive regulation of mesenchymal cell proliferation	0.018
4	GO:0019901	protein kinase binding	3.8e-05
4	GO:0032000	positive regulation of fatty acid beta-oxidation	0.0003
4	GO:0005158	insulin receptor binding	0.0078
4	GO:0048015	phosphatidylinositol-mediated signalling	0.01
4	GO:0010907	positive regulation of glucose metabolic process	0.013
4	GO:0032467	positive regulation of cytokinesis	0.014
4	GO:0008286	insulin receptor signalling pathway	0.017

Continued on next page

position	GO term	Description	E-value
		Continued from previous page	
4	GO:0051256	mitotic spindle midzone assembly	0.039
4	GO:0097149	centralspindlin complex	0.039
4	GO:0008543	fibroblast growth factor receptor signalling pathway	0.041

TABLE B.25: Gene ontology (GO) term enrichment analysis for predicted PKA substrates. Shown are all positions that the kinase was found to be significantly over-represented at.

position	GO term	Description	E-value
All	GO:0004871	signal transducer activity	3.33e-05
All	GO:0048011	neurotrophin TRK receptor signalling pathway	0.0001
All	GO:0007165	signal transduction	0.0005
All	GO:0005737	cytoplasm	0.0017
All	GO:0005515	protein binding	0.004
All	GO:0043065	positive regulation of apoptotic process	0.01
All	GO:0019901	protein kinase binding	0.029
All	GO:0007399	nervous system development	0.049
2	GO:0042301	phosphate ion binding	0.0302
3	N/A	N/A	N/A
4	GO:0008543	fibroblast growth factor receptor signalling pathway	0.0001
4	GO:0007173	epidermal growth factor receptor signalling pathway	0.0003
4	GO:0019901	protein kinase binding	0.0006
4	GO:0032000	positive regulation of fatty acid beta-oxidation	0.002
4	GO:0048015	phosphatidylinositol-mediated signaling	0.0073
4	GO:0048011	neurotrophin TRK receptor signalling pathway	0.0087
4	GO:0008286	insulin receptor signalling pathway	0.013
4	GO:0060397	JAK-STAT cascade involved in growth hormone signalling pathway	0.019
4	GO:0042593	glucose homeostasis	0.032
4	GO:0005829	cytosol	0.034
5	GO:0097149	centralspindlin complex	0.015
5	GO:0048008	platelet-derived growth factor receptor signaling pathway	0.049
5	GO:0090399	replicative senescence	0.049

TABLE B.26: Gene ontology (GO) term enrichment analysis for substrates predicted to contain an NLS and a phosphorylation site at the specific position relative to the NLS.

position	GO term	Description	E-value
-10	GO:0005694	chromosome	0.0017
-10	GO:0000786	nucleosome	0.024
			Continued on next page
position	GO term	Description	E-value
			Continued from previous page
-9	GO:0005730	nucleolus	0.045
-8	N/A	N/A	N/A
-7	N/A	N/A	N/A
-6	N/A	N/A	N/A
-5	N/A	N/A	N/A
-4	N/A	N/A	N/A
-3	N/A	N/A	N/A
-2	N/A	N/A	N/A
-1	N/A	N/A	N/A
0	N/A	N/A	N/A
1	N/A	N/A	N/A
2	N/A	N/A	N/A
3	GO:0005694	chromosome	0.0039
4	N/A	N/A	N/A
5	GO:0006974	cellular response to DNA damage stimulus	0.0025
5	GO:0008274	gamma-tubulin ring complex	0.015
5	GO:0097149	centralspindlin complex	0.015
6	GO:0048011	neurotrophin TRK receptor signaling pathway	0.025
7	GO:0000786	nucleosome	5.31e-10
7	GO:0006334	nucleosome assembly	2.02e-08
7	GO:0032982	myosin filament	1.53e-05
7	GO:0005694	chromosome	2.90e-05
7	GO:0005859	muscle myosin complex	0.0001
7	GO:0046982	protein heterodimerization activity	0.0007
7	GO:0030016	myofibril	0.002
7	GO:0016459	myosin complex	0.015
7	GO:0000146	microfilament motor activity	0.019
7	GO:0042742	defense response to bacterium	0.019
7	GO:0005925	focal adhesion	0.041
7	GO:0030049	muscle filament sliding	0.041
8	GO:0000786	nucleosome	1.98e-07
8	GO:0006334	nucleosome assembly	2.42e-06
			Continued on next page

position	GO term	Description	E-value
		Continued from previous page	
8	GO:0005694	chromosome	0.00027
8	GO:0046982	protein heterodimerization activity	0.0087
8	GO:0042742	defense response to bacterium	0.009
9	GO:0006334	nucleosome assembly	8.68e-08
9	GO:0000786	nucleosome	2.12e-07
9	GO:0005694	chromosome	0.0004
10	N/A	N/A	N/A

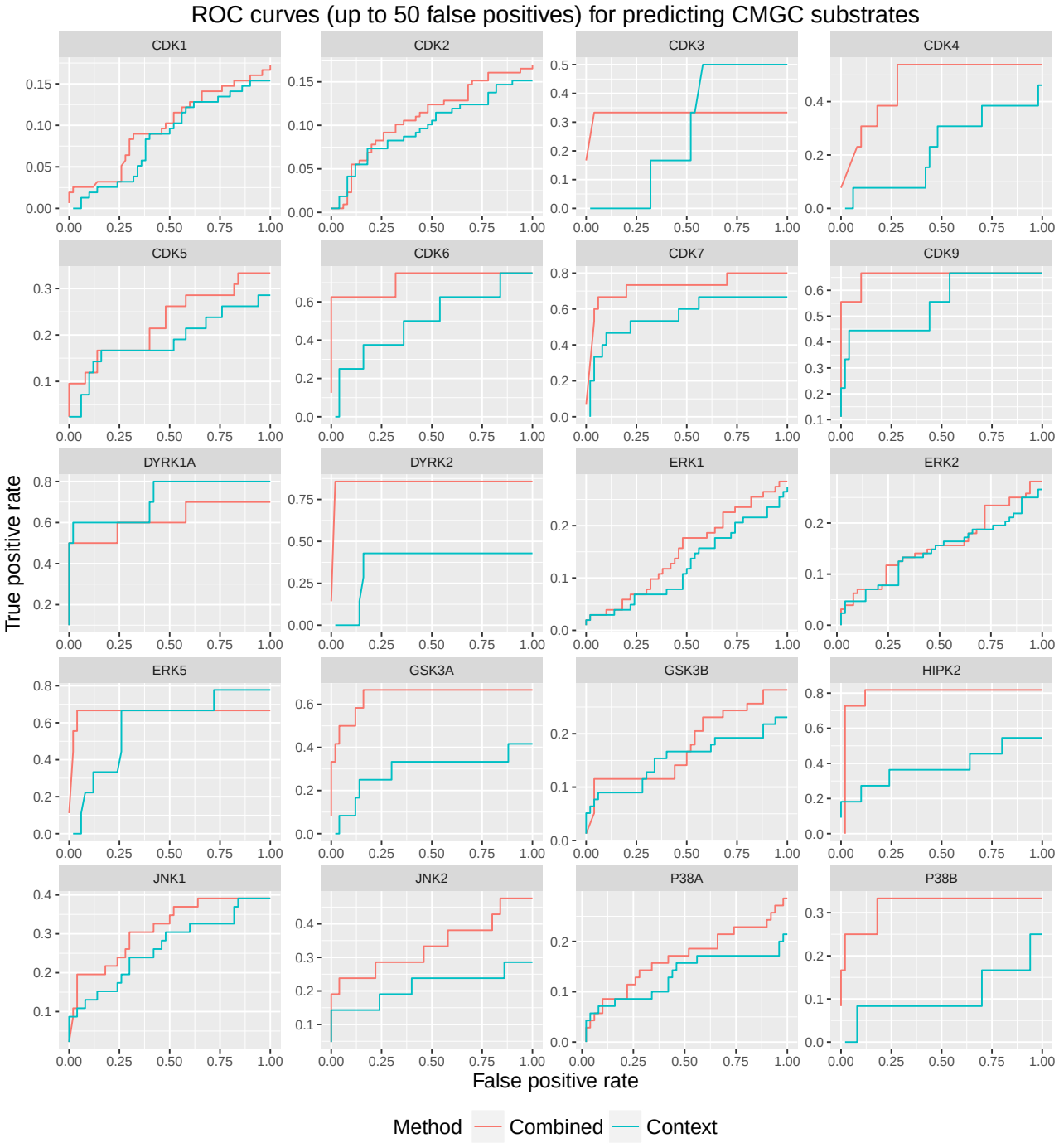

FIGURE B.1: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the human CMGC family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

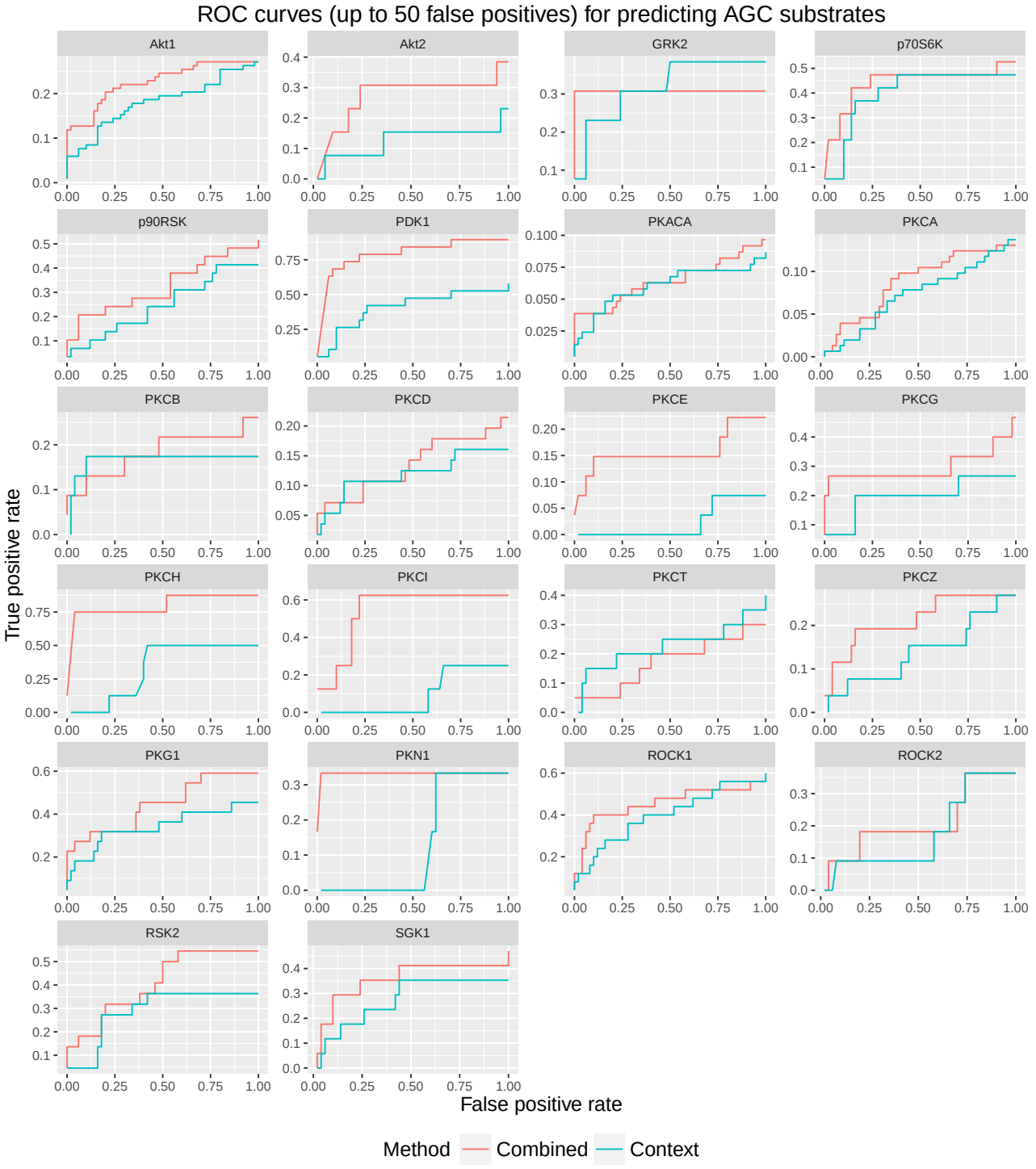

FIGURE B.2: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the human AGC family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

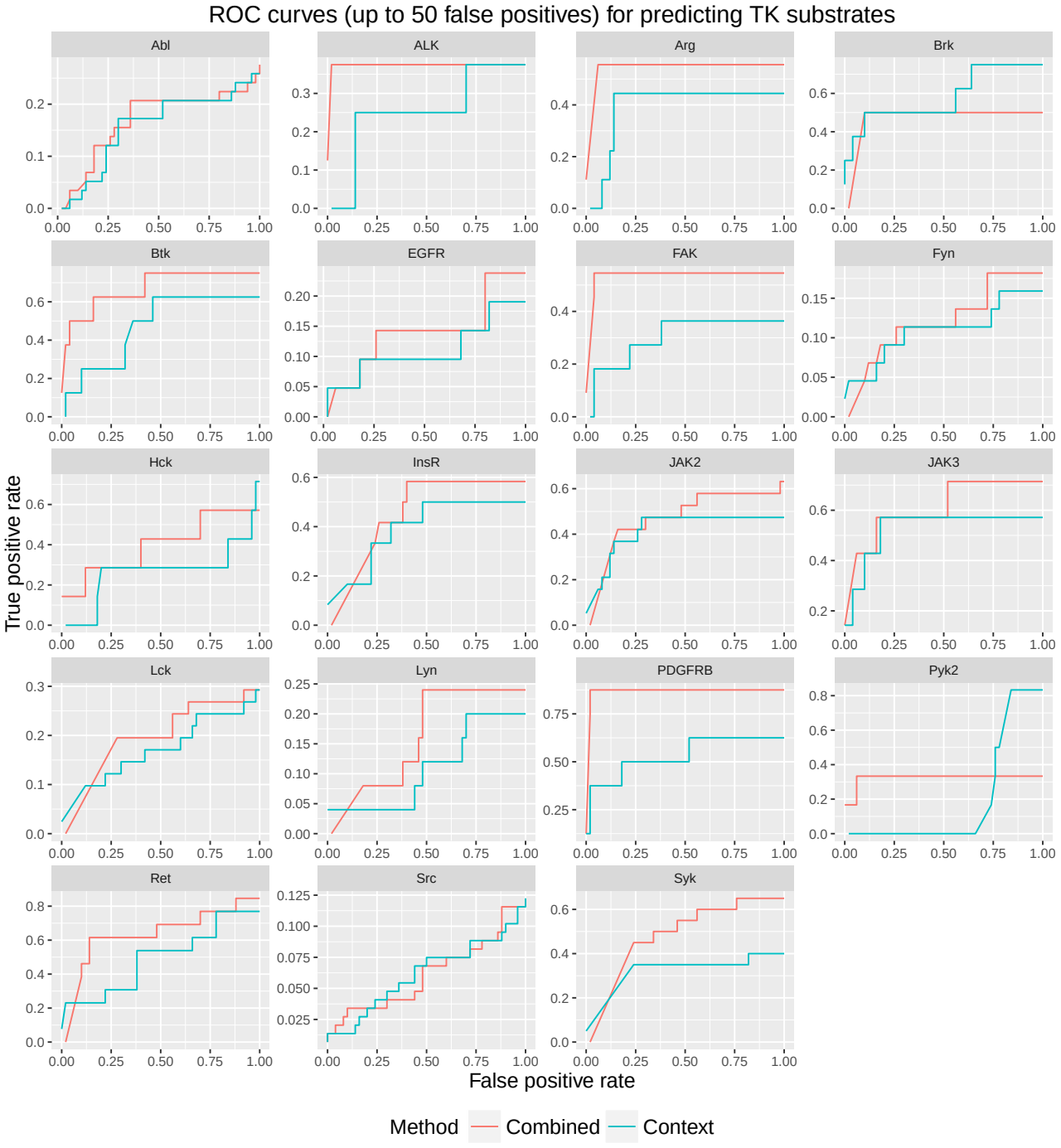

FIGURE B.3: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the human TK family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

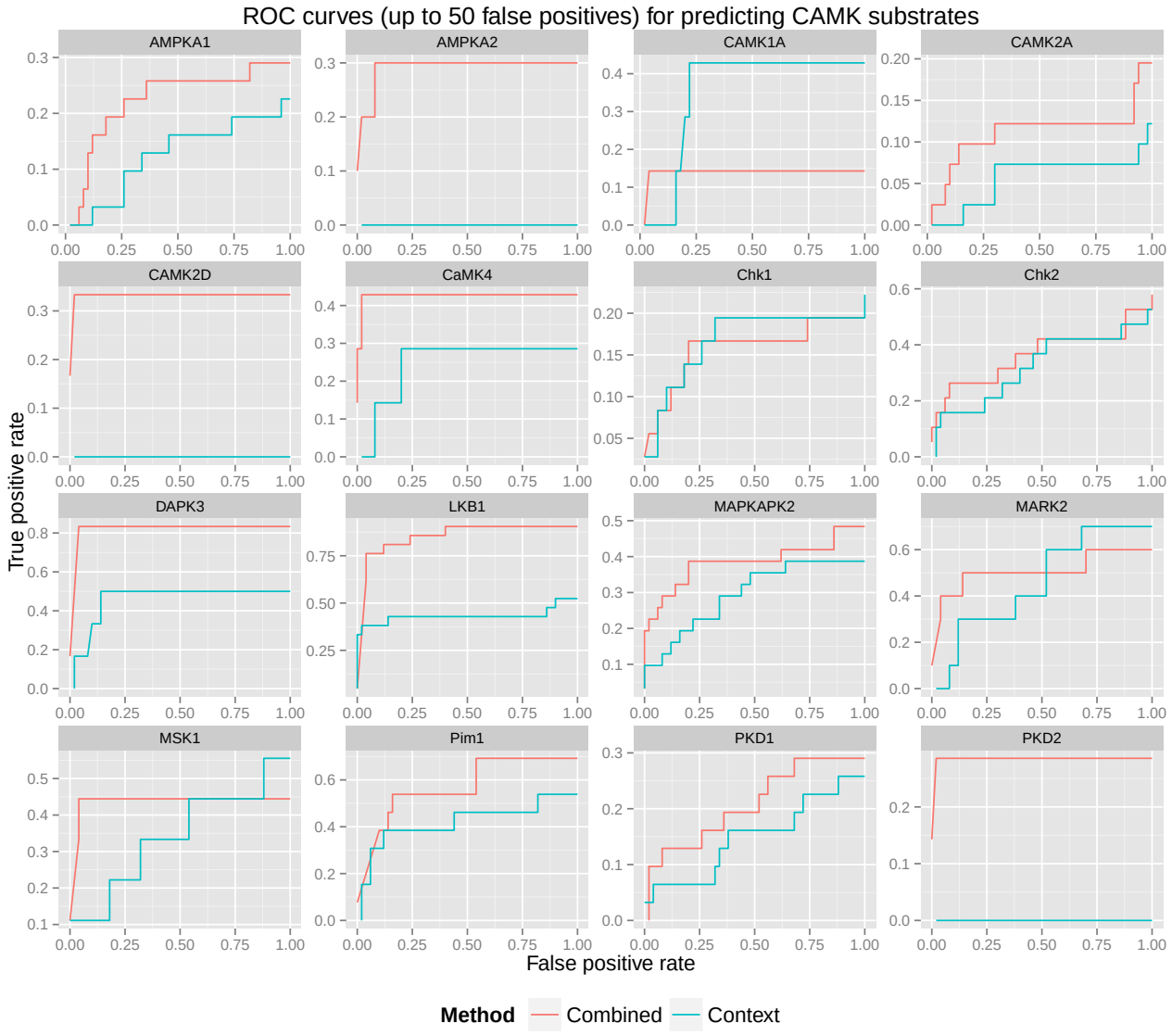

FIGURE B.4: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the human CAMK family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

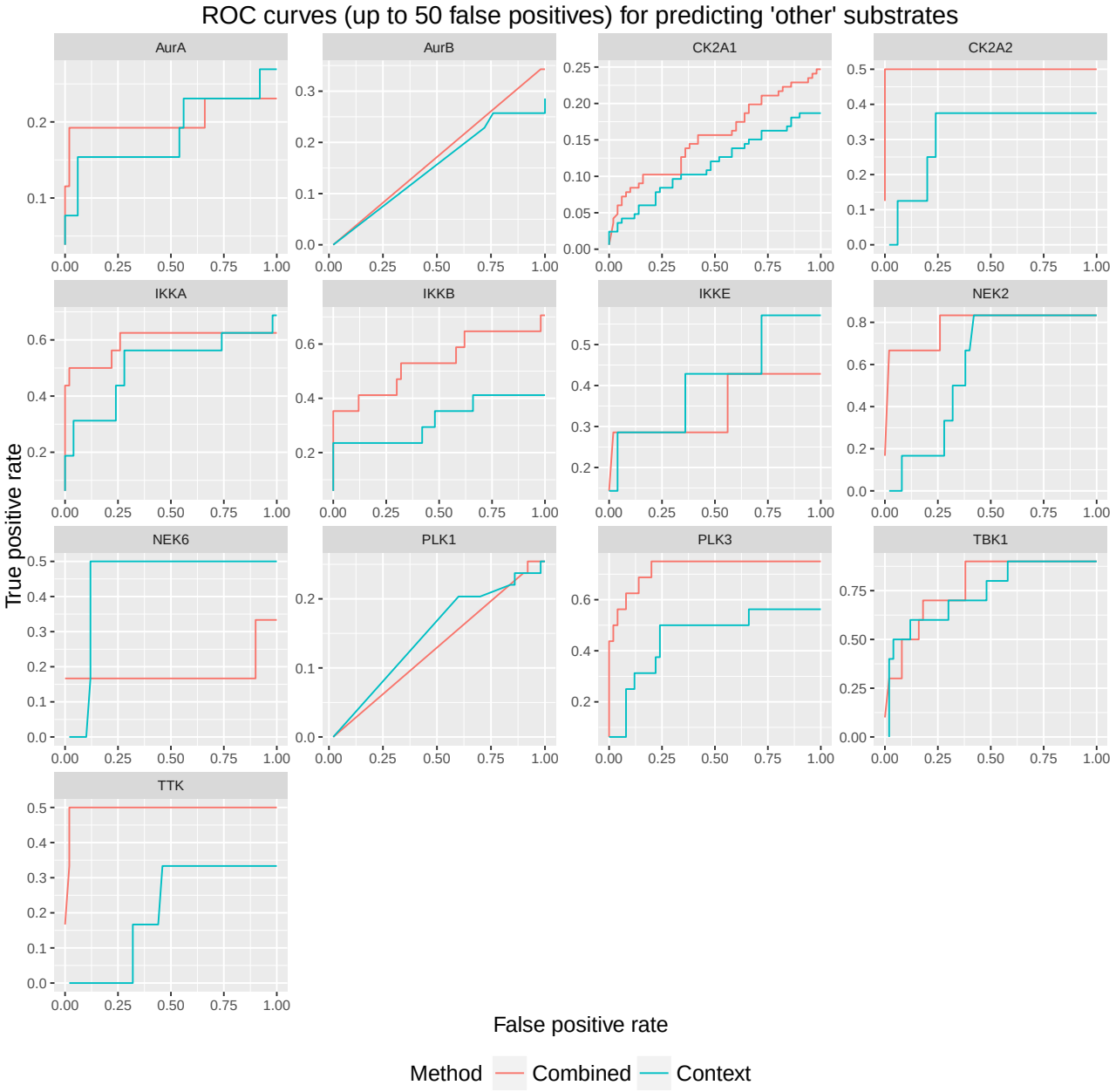

FIGURE B.5: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the human 'other' family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

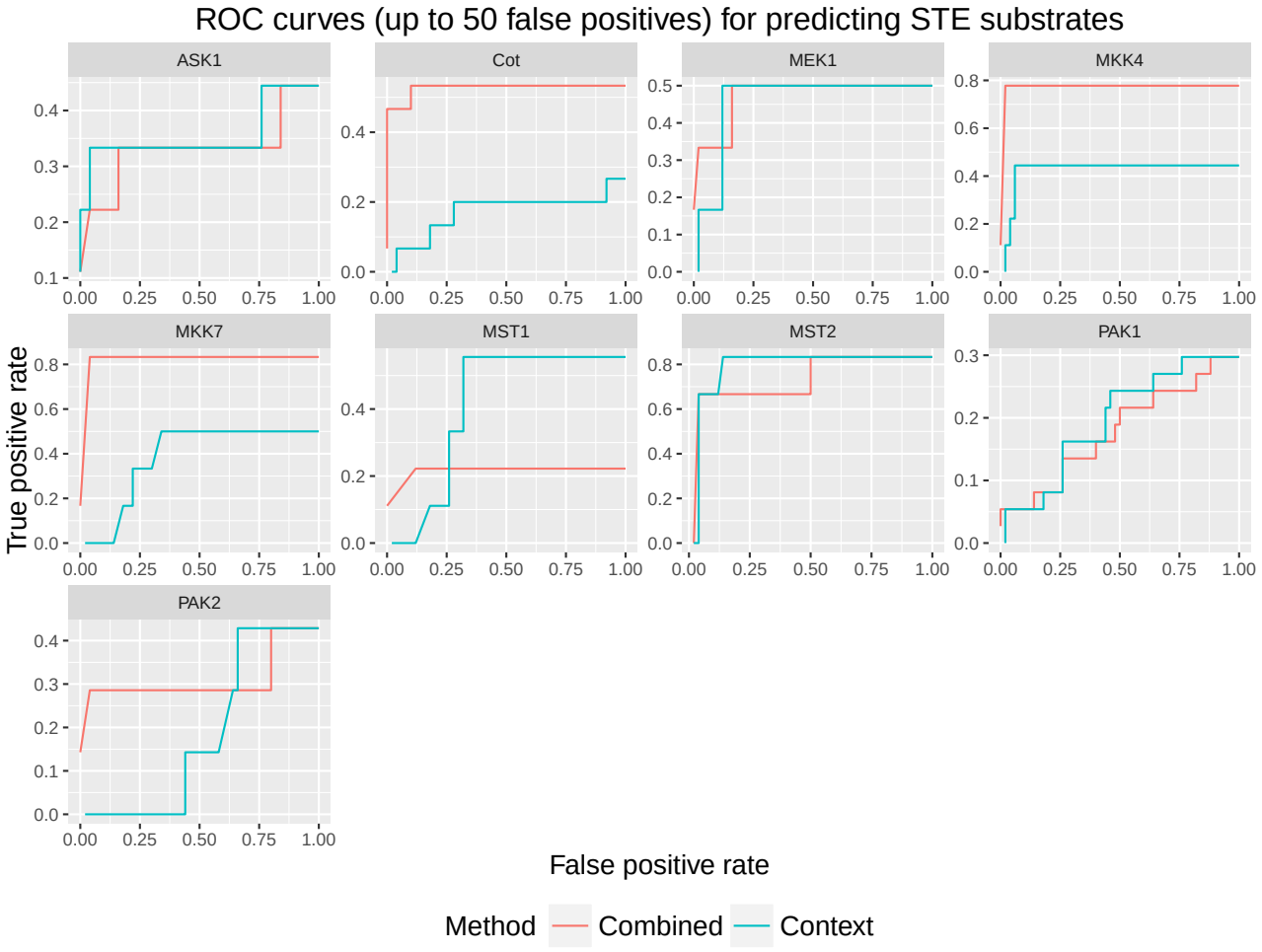

FIGURE B.6: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the human STE family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

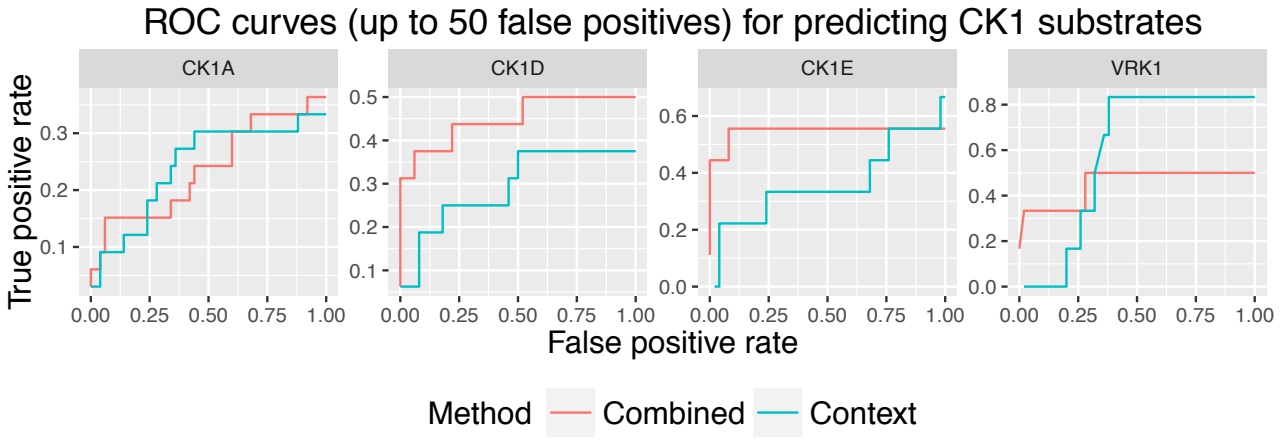

FIGURE B.7: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the human CK1 family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

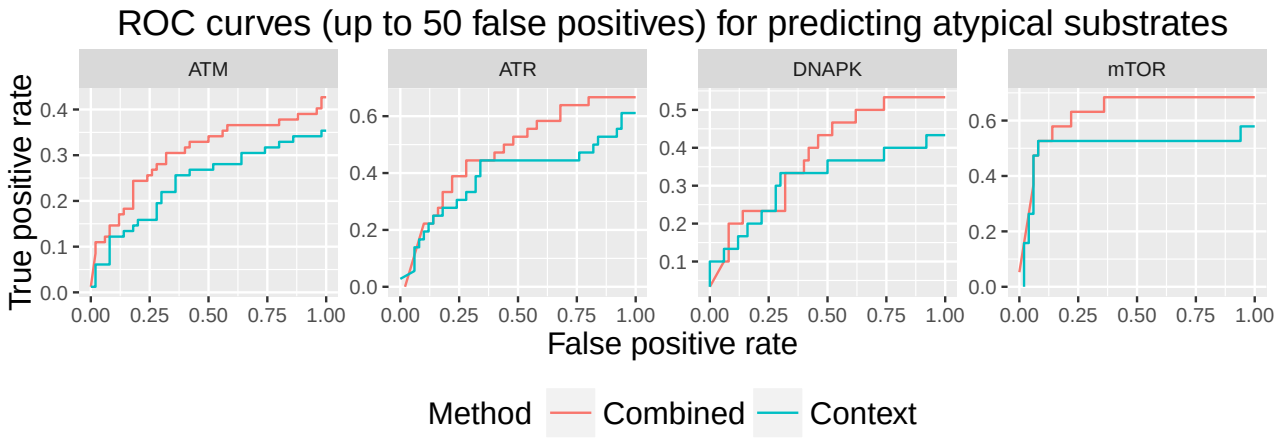

FIGURE B.8: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the human atypical family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

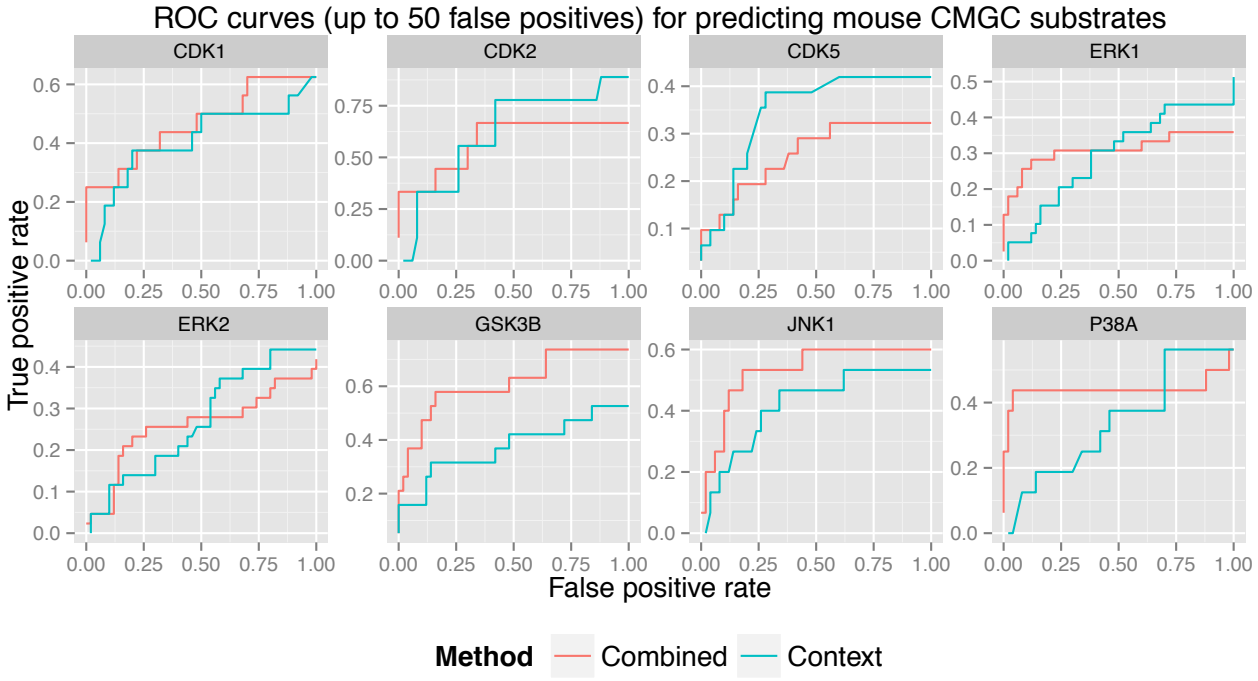

FIGURE B.9: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the mouse CMGC family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

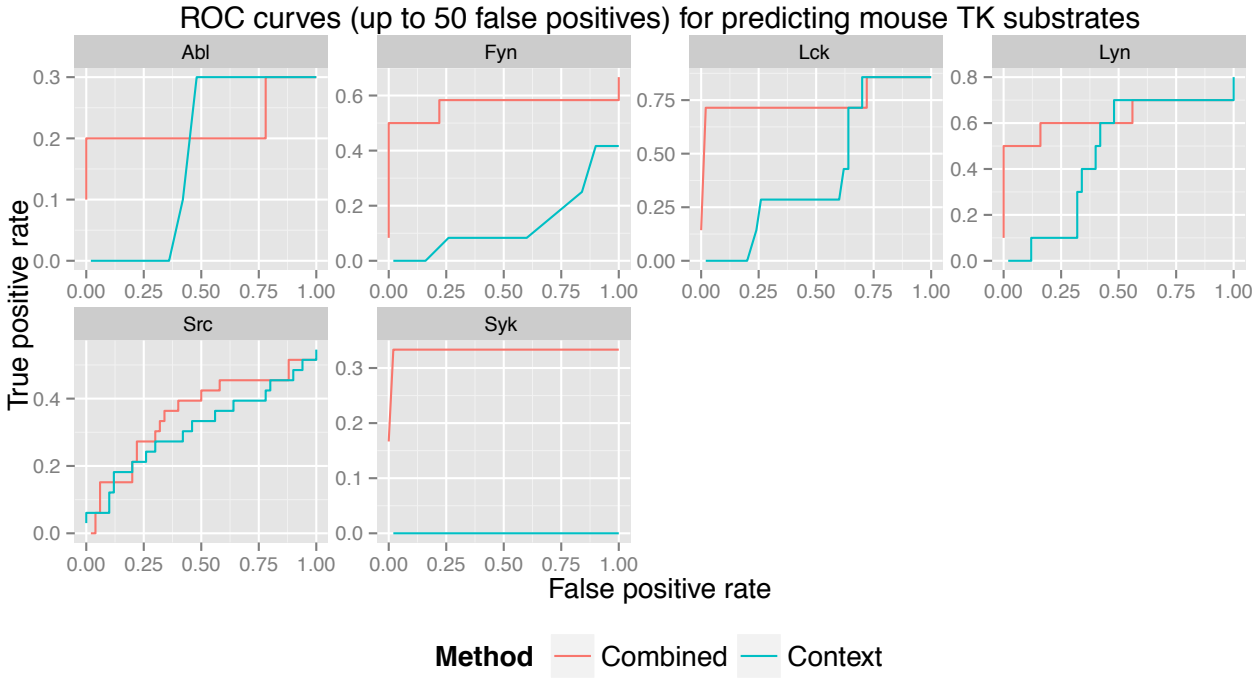

FIGURE B.10: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the mouse TK family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

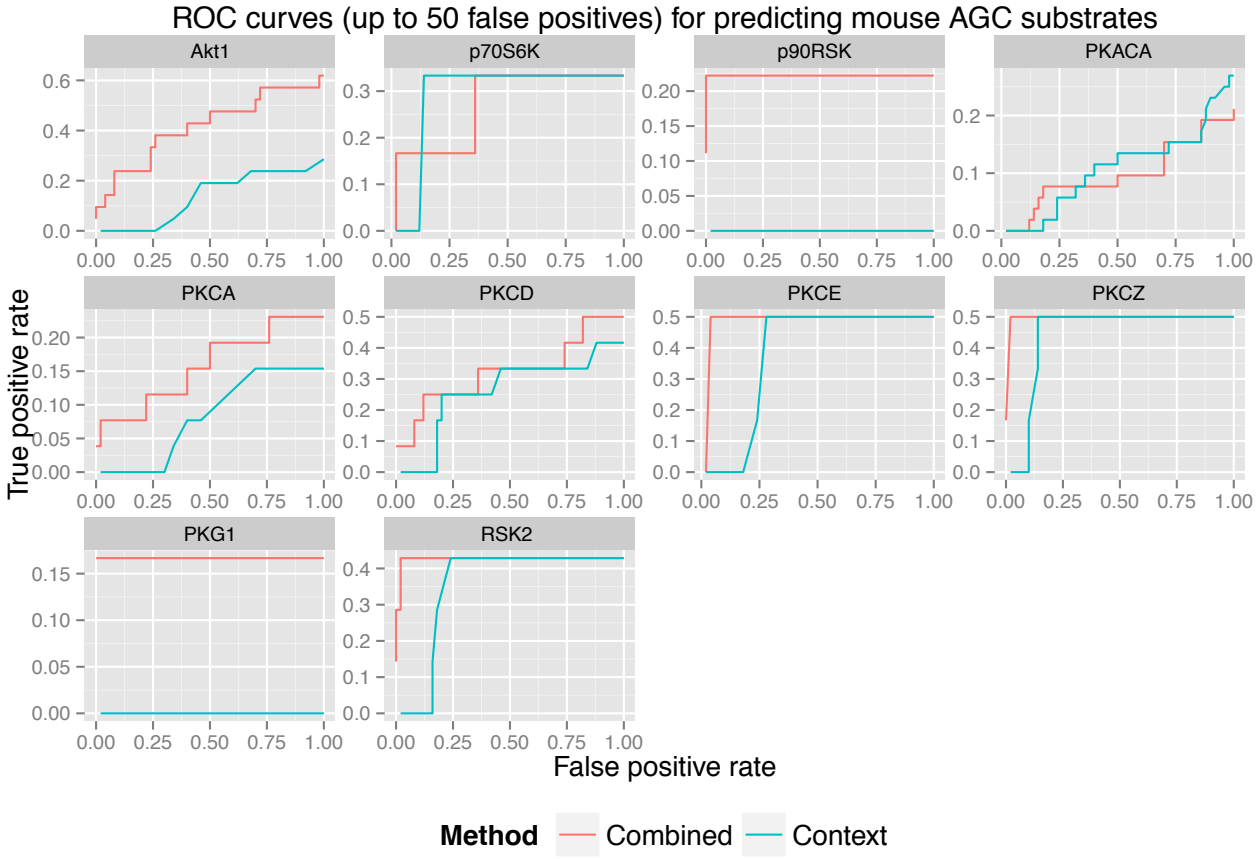

FIGURE B.11: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the mouse AGC family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

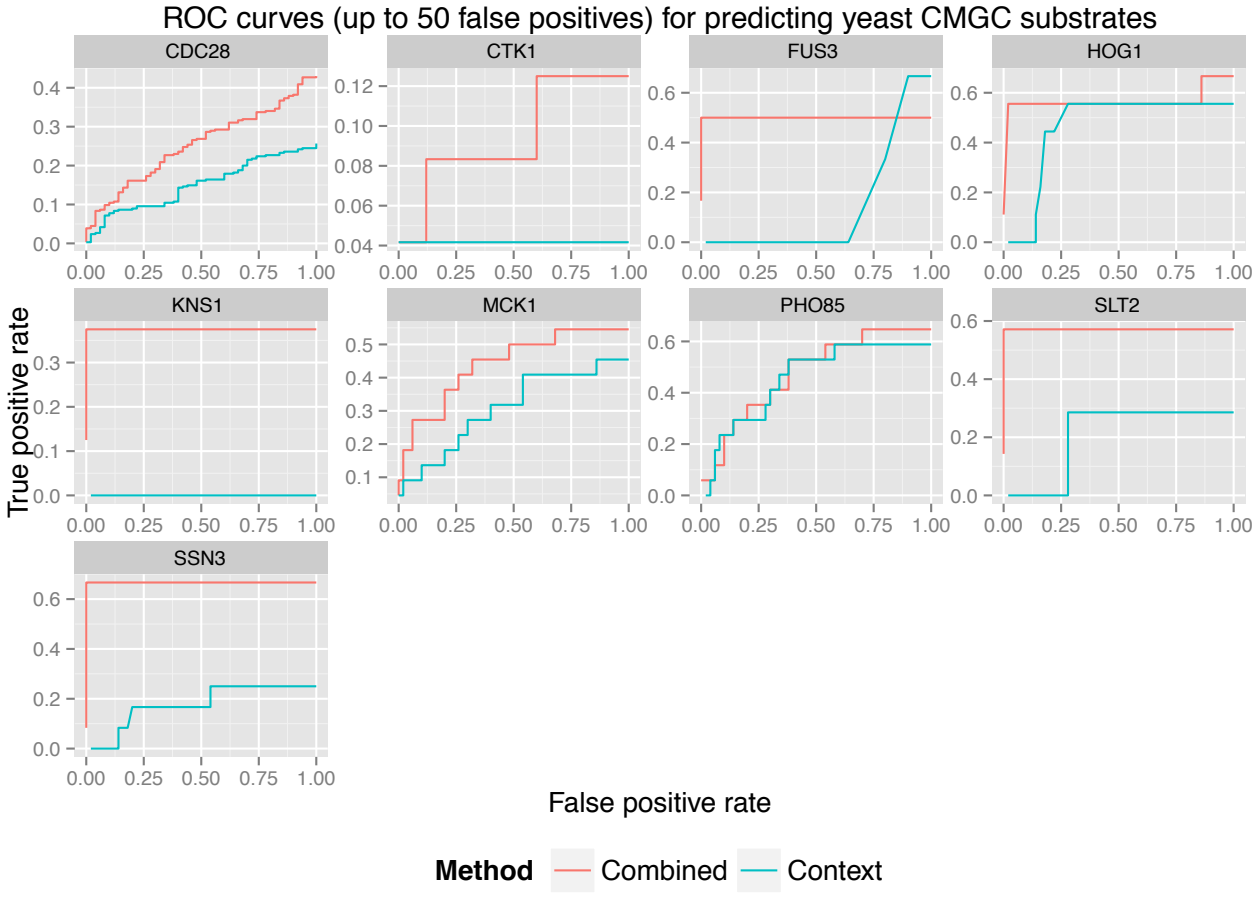

FIGURE B.12: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the yeast CMGC family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

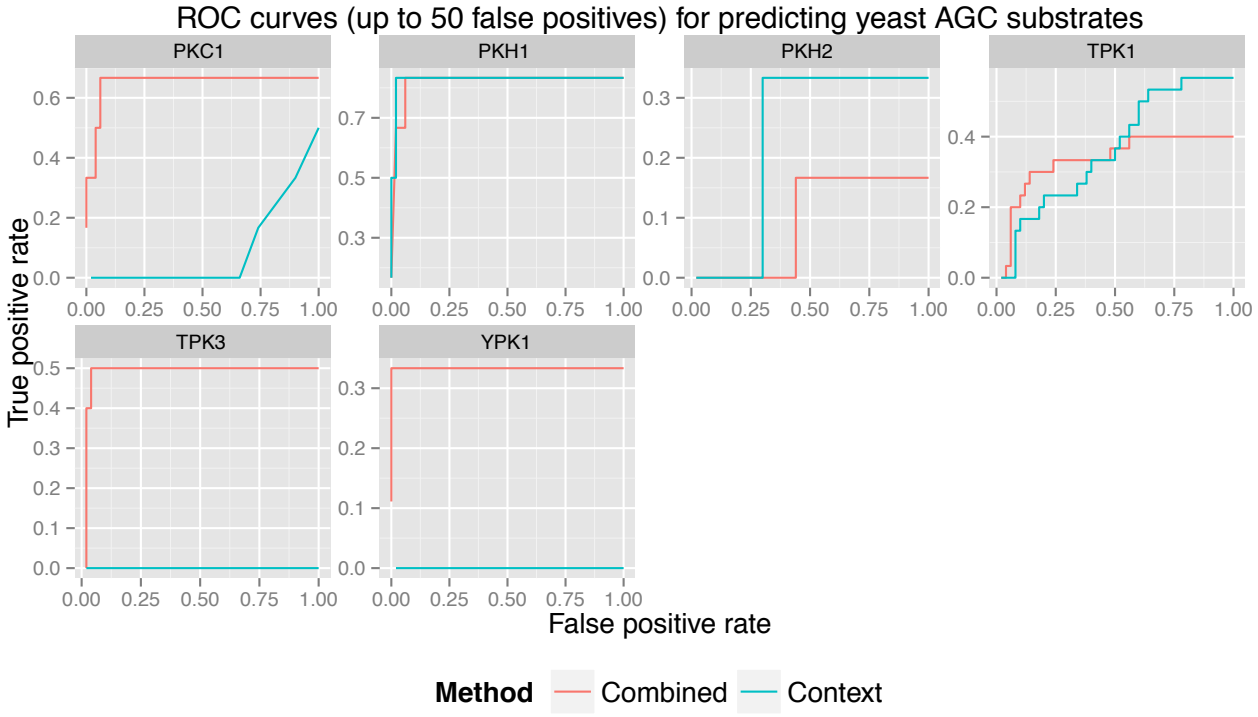

FIGURE B.13: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the yeast AGC family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

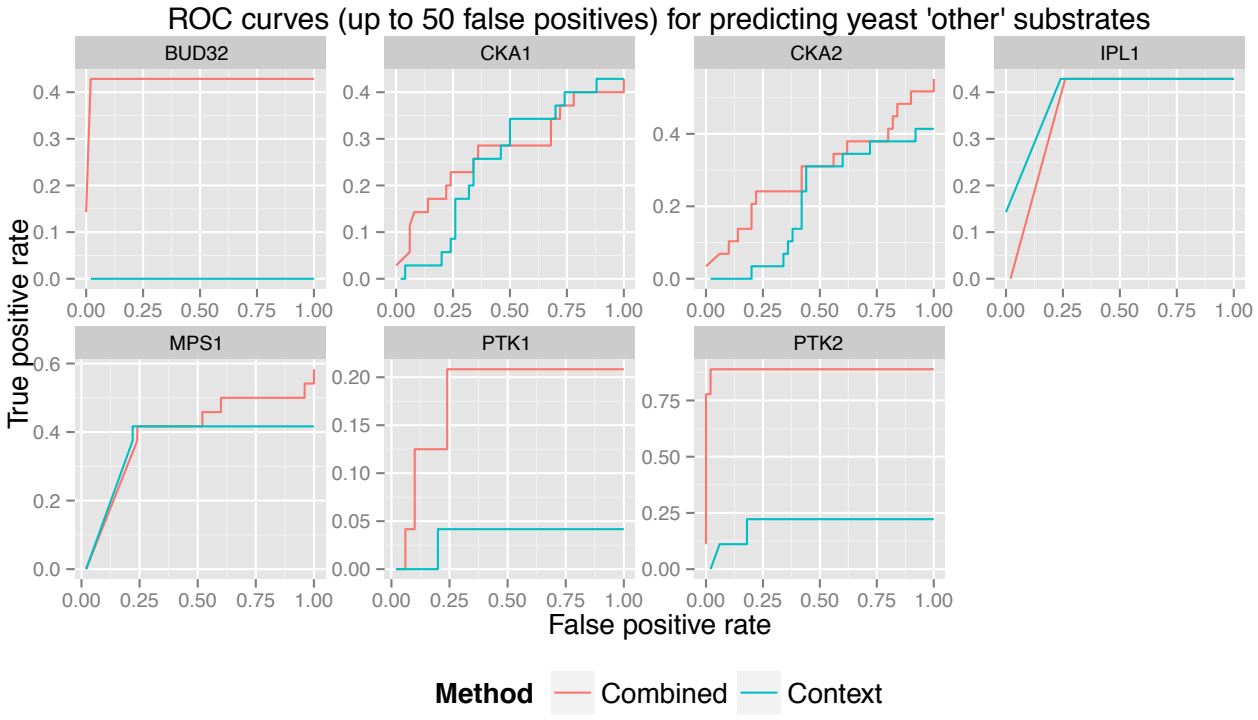

FIGURE B.14: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the yeast 'other' family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

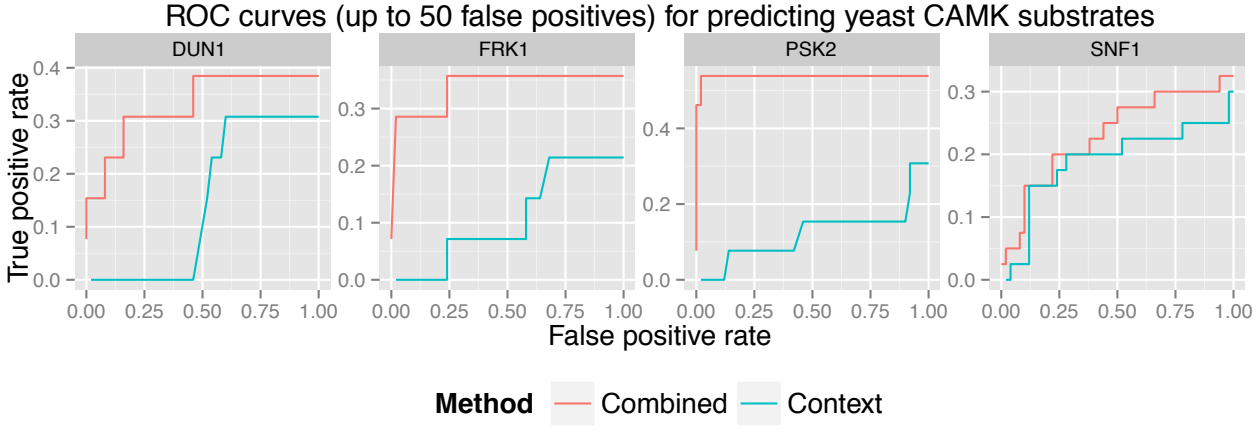

FIGURE B.15: ROC plots showing the prediction accuracy of the combined and context models for predicting kinases substrates from the yeast CAMK family of kinases. The ROC curves are calculated up to the first 50 false positives from a 10-fold cross-validation run.

B.1 Identifying expected sequence motifs from context

As the Bayesian network combined two diverse types of information, we were interested in observing what the model "expects" from a kinase binding motif in response to the protein interaction and cell-cycle data that is presented to it. To do this we took the full set of human proteins from Uniprot (canonical plus isoforms) and obtained their relevant context information.

For each protein, we first set the context parameters in the Bayesian network: the protein interaction nodes, cell-cycle nodes and kinase nodes (except the kinase being queried). We used the most probable explanation (MPE) form of inference to determine the most likely value for the query kinase phosphorylating the substrate, as well as the expected values of the dimer and trimer nodes. If the model at this point did not believe the query kinase to be phosphorylating the protein, the protein was discarded. Otherwise, we then used the expected values of the k-mer variables to set their respective nodes, and queried each of the position-specific amino acid nodes, inferring the probability of each potential amino acid.

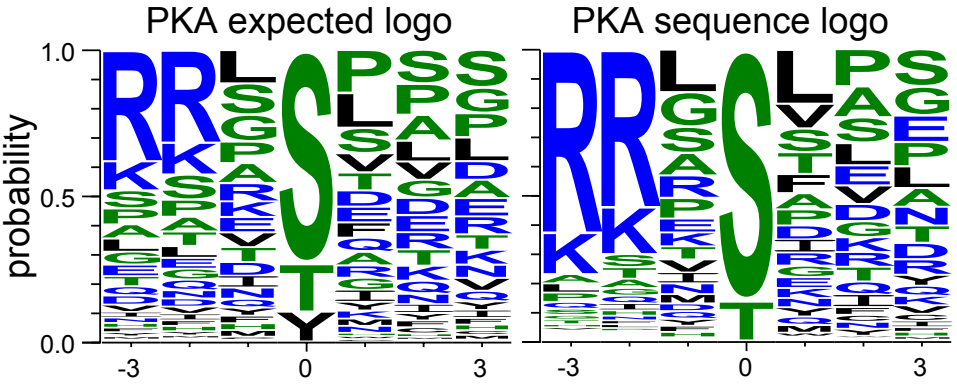

FIGURE B.16: Comparison of sequence logos for PKA kinase. Left logo shows amino acid probabilities expected by the combined model for PKA binding sites when context information for a query substrate indicates that PKA will target the protein. Right logo was made using peptides from actual PKA phosphorylation substrates. Logo generated using WebLogo3 (47).

For each position in the motif, we then took the sum of probabilities for each amino acid across the samples predicted to be phosphorylated by the kinase. This resulted in a position-specific matrix of counts across the 20 amino acids for the kinase. In order to visualise the position-specific amino acid counts we used WebLogo 3 (47) to generate sequence logos from the count matrix.

Figure B.16 shows a sequence logo generated from the probability distributions of amino acids from proteins predicted to be PKA substrates, based only on context data being provided to

the model. We compared this to a sequence logo generated from actual PKA substrates from PhosphoSitePlus®. The comparison shows that there is a high level of similarity between the expected amino acids, given the context information, and the amino acid frequencies from actual PKA substrates. This demonstrates that the model is able to have a prior expectation about what binding site to expect on a protein sequence, before actually seeing the sequence.

B.2 Web-server workflow

Uniprot reviewed (Swissprot) proteins were downloaded for human (July, 2014), mouse (February, 2015) and yeast (February, 2015). The full set of canonical and isoform proteins were downloaded for the three species. For each kinase, the combined model was trained on the full set of training data. Each protein in the relevant proteome was submitted to the model and the probability of it being a substrate of the kinase was queried. The kinase predictions for each substrate were stored in an SQLite3 database.

When a user uploads a Fasta file of protein sequences, they are submitted for a BLASTP query against the proteome of the chosen species (human, mouse or yeast). If an exact match is made for a protein in the database, that protein is retrieved. We also wanted to allow for users to submit isoforms or homologs that are not in the database; i.e. such proteins would obtain a substrate prediction based on the closest relative protein in the database. Therefore, if an exact match is not made, proteins in the database that obtain an E-value < 0.001, and have a sequence identity of at least 90% will be considered. The highest E-value is taken, and all proteins in the database that obtain the E-value are returned. Once proteins in the database have been identified from the BLASTP search, the requested kinase predictions are retrieved.

The user's sequences are then scanned using the sequence model and each potential phosphorylation site is scored. If the user has requested that their predictions be thresholded according to P-value, only the results that fall below the chosen P-value threshold will be returned. The output is an interactive table of results for each potential phosphorylation site in the user's submitted proteins for each kinase that was queried. Users can filter their results by providing a list of protein names, or protein names and sites. The results can also be downloaded as a tab-delimited text file. The results for each protein can be viewed separately by clicking on a desired protein to be redirected to the "Protein Viewer" page, which presents an interactive view of the protein annotated with predicted phosphorylation sites.

In addition to submitting protein sequences for analysis, the option exists to download proteome-wide sets of kinase-substrate predictions. Similar to the submission page, users are able to select sets of kinases from either human, mouse of yeast, though instead of uploading protein sequence, there is an option to choose between downloading predictions for the set of Swissprot canonical or isoform proteins. P-values for predictions can also be calculated.

Results visualisation

In order to create a way for visualising the potential kinase binding sites on a protein, we implemented a "Protein Viewer" page. This was based on the BioJS (181) package pViz (182), which allows the zoomable visualisation of an amino acid sequence with multiple rows of annotations on specified positions on the sequence. For a protein, the visualisation consists of a row of annotations representing potential phosphorylation sites for each kinase that a user queries. Phosphorylation site predictions are presented as coloured circles, where the shade of the circle indicates the strength of the context prediction and the size of the circle indicates the strength of the sequence prediction for that site. When a user clicks on a site, an information box is displayed showing the details of that prediction.

Appendix C

Chapter 5 supplementary material

TABLE C.1: Variants are listed according to the cancer or disease they are associated with. Each row contains protein name as UniProt accession, the location of the variant and phosphorylation site, the kinase predicted to target the site, the reference and variant scores for the peptide.

Cancer	Protein	Variant	Phos.	Kinase	$R_{subst.}$	R_{site}	V_{site}	$E_{combined}$	Peptide
Ovarian	P35222	G555A	T551	Akt2	1	1	4.95e-05	9.36e-09	QDTQRR$\mathbf{T}_p$SMG[G/A]TQ
	P26010	Y753H	Y753	FAK	1	1	0	2.23e-08	YRLSVEI[$\mathbf{Y}_p$/H]DRREYSR
	Q7KZI7	S197N	S197	NEK6	1	1	0	2.3e-08	KIADFGF[$\mathbf{S}_p$/N]NEFTFGN
	P51813	S212R	S212	GSK3B	0.998	1	0	1.33e-07	PPSSST[$\mathbf{S}_p$/R]LAQYDS
	P46939	M1256R	T1259	MARK2	1	0.914	0.0005	4.47e-05	R[M/R]KS$\mathbf{T}_p$EVLP
	P43355	K278I	Y276	Brk	0.998	1	0.000979	0.000103	RALAETS$\mathbf{Y}_p$V[K/I]VLEYV
	P18846	L71I	S72	MSK1	1	0.0525	0.00138	0.000107	RKILKD[L/I]$\mathbf{S}_p$SEDTRGR
	P35222	G555A	T556	Akt2	1	0.0694	7.46e-09	0.000108	RTSMG[G/A]$\mathbf{T}_p$QQQFVE
	Q96BY6	T1347M	T1347	CAMK2D	0.681	0.351	0	0.00011	CFLHIMK[$\mathbf{T}_p$/M]ISYETLI
	Q13009	S170F	S170	PKCA	1	0.982	0	0.00013	SFKKKR[$\mathbf{S}_p$/F]KSADIW
	Q9BQQ3	R106C	S109	MAPKAPK2	0.486	0.999	0.00399	0.000331	SF[R/C]RA$\mathbf{S}_p$EQVWH
	Q08999	P547A	T541	CDK7	1	0.21	0.0142	0.000577	ACCLEVV$\mathbf{T}_p$FSYKP[P/A]G
	Q96KQ7	S119F	S118	CDK2	0.711	0.964	0.00256	0.0013	ATKSFP$\mathbf{S}_p$[S/F]PSKGG
	P50747	S83N	S83	MARK2	0.321	0.87	0	0.00131	SASG[$\mathbf{S}_p$/N]EPAG
	O94988	R648Q	S650	AMPKA1	0.207	0.978	0.0507	0.00132	FMR[R/Q]R$\mathbf{S}_p$SSLGS
	O94988	R648Q	S651	AMPKA1	0.207	0.942	0.0156	0.00132	MR[R/Q]RS$\mathbf{S}_p$SLGSY
	Q96KQ7	S119F	S119	CDK5	0.091	0.806	0	0.00154	KSFPS[$\mathbf{S}_p$/F]PSKGG
	O75182	S130L	S126	IKKB	0.0176	0.979	0.000183	0.0016	NIQSPLT$\mathbf{S}_p$QEN[S/L]HNH
	Q13136	R137Q	S138	MARK2	0.214	0.752	7.89e-09	0.00199	RHE[R/Q]$\mathbf{S}_p$LRMT
	Q8N9Q2	T47R	T47	CK2A1	1	1	0	0.002	VLDVSS[$\mathbf{T}_p$/R]SSEDSD
	Q96D09	F508C	S512	PKG1	0.98	0.975	0.0982	0.00226	[F/C]RST$\mathbf{S}_p$PFGI
	P55209	K276Q	T269	MST1	0.458	0.267	0.0983	0.00233	WKKGKNV$\mathbf{T}_p$LKTIKK[K/Q]
	P04198	P358L	S355	CDK6	0.000617	0.55	0.000508	0.00236	KKIKSEA$\mathbf{S}_p$PR[P/I]LKSV
	Q2M1Z3	N776T	S778	CDK4	0.0007	0.959	0.037	0.00246	VGGPG[N/T]L$\mathbf{S}_p$PPLPPAP
	P20338	A208T	S204	MAPKAPK2	0.0707	0.953	0.00571	0.00268	LRQLR$\mathbf{S}_p$PRR[A/T]Q
	Q14149	V872I	T874	GRK2	0.995	0.982	0.14	0.00279	QTATD[V/I]ST$\mathbf{S}_p$SNIEES
	Q14149	V872I	S875	GRK2	0.995	0.999	0.167	0.00291	TATD[V/I]STS$\mathbf{S}_p$NIEESV
	Q14686	S1349A	S1349	CDK2	0.538	0.939	0	0.0031	SPGRQN[$\mathbf{S}_p$/A]KAPKLT
	Q04206	E127Q	S131	ATR	0.889	0.87	0.0366	0.0032	L[E/Q]QAI$\mathbf{S}_p$QRIQT
	Q86UR5	R1113W	S1116	Akt1	0.473	0.991	0.024	0.00324	DRA[R/W]SA$\mathbf{S}_p$TNCLRP
	O75182	S130L	T125	IKKB	0.0176	0.924	0.00019	0.00438	LNIQSPL$\mathbf{T}_p$SQEN[S/L]HN
	O75182	S130L	S130	IKKB	0.0176	0.921	0	0.00447	PLTSQEN[$\mathbf{S}_p$/L]HNHGDGA
	Q14517	S157P	S150	NEK6	0.00491	0.00376	0.000213	0.00461	NDLRPLF$\mathbf{S}_p$PTSYSV[S/P]
	Q92610	T1024N	S1027	PKCG	0.42	0.885	0.0502	0.00488	QSFH[T/N]PN$\mathbf{S}_p$LRKHIRN
	Q92766	S1140F	S1140	ERK5	0.404	0.795	0	0.00567	ASSPEAA[$\mathbf{S}_p$/F]PTEQGPA
	Q15361	S240W	S240	ATR	0.00249	0.949	0	0.00612	AMPEG[$\mathbf{S}_p$/W]QAGRE
	A0PJX4	Q132H	S130	DNAPK	0.0411	0.995	0.0237	0.00672	CLRPKEP$\mathbf{S}_p$Q[Q/H]PIRFS
	Q8TDJ6	L964P	S960	CaMK4	0.865	0.804	0.316	0.00742	PHSS$\mathbf{S}_p$IAN[L/P]

Continued on next page

	Protein	Variant	Phos.	Kinase	$R_{subst.}$	R_{site}	V_{site}	$E_{combined}$	Peptide
									Continued from previous page
	P12235	R188K	Y187	ALK	0.827	0.754	5.87e-08	0.0075	GIII$\mathbf{Y}_p$[R/K]AAY
	P08151	R637Q	S640	PKD1	0.205	0.855	0.0074	0.00754	VT[R/Q]RA$\mathbf{S}_p$DPAQA
	P62995	R62G	S64	AurB	0.102	0.998	0.00645	0.00762	R[R/G]S$\mathbf{S}_p$RRH
	O94988	R648Q	S650	CAMK1A	0.0375	0.0134	9.08e-06	0.00912	FMR[R/G]R$\mathbf{S}_p$SSLGS
	Q86VZ2	S240R	T247	PKCG	0.206	0.751	0.0191	0.00996	[S/R]RGRCLK$\mathbf{T}_p$YTGHKNE
	P17020	S129Y	S129	RSK2	0.00935	0.82	0	0.0105	GRRLPQ[$\mathbf{S}_p$/Y]LSQEGD
	P09848	V971L	Y974	Lyn	0.0304	0.961	0.163	0.0107	RALK[V/L]KA$\mathbf{Y}_p$FSISWS
	Q15047	S504C	S504	GSK3B	0.174	0.928	0	0.011	SVGSGH[$\mathbf{S}_p$/C]SPTSPA
	P20930	R2018K	S2017	P38B	0.000651	0.0439	6.53e-06	0.0115	QLQSADS$\mathbf{S}_p$[R/K]HSGIGH
	Q9BQQ3	R106C	S104	PKCD	0.0198	0.892	0.0153	0.0121	ASVRFC$\mathbf{S}_p$F[R/C]RASE
	O60237	G422V	S421	CaMK4	0.734	0.316	2.35e-05	0.0122	RRFS$\mathbf{S}_p$[G/V]LFN
	Q8WXG6	R1643W	S1646	PKD2	0.000201	6.86e-05	5.97e-08	0.0125	RTPP[R/W]PV$\mathbf{S}_p$S
	Q8IWI9	S645R	S645	ERK2	0.013	0.802	0	0.0128	STKNTPV[$\mathbf{S}_p$/R]PGSTFPD
	P18583	S1782F	S1780	p70S6K	0.00452	0.903	0.0084	0.0132	SMPERA$\mathbf{S}_p$E[S/F]SSEE
	Q14653	E137K	T135	DNAPK	0.138	0.893	0.0187	0.0137	GGGSTSD$\mathbf{T}_p$Q[E/K]DILDE
	Q9NQL9	S382R	S381	AMPKA1	0.098	0.945	0.268	0.0151	LARSQ$\mathbf{S}_p$[S/R]PFLP
	O14717	G155V	S150	P38B	0.000301	0.0272	3.02e-06	0.0164	QYQEFLL$\mathbf{S}_p$PTSL[G/V]IP
	Q9NUQ6	I136K	S135	CaMK4	0.515	0.204	0.00342	0.0182	EKKI$\mathbf{S}_p$[I/K]LEE
	Q96AV8	S160R	S160	CDK3	1.08e-05	5.36e-06	0	0.0192	KFLARYP[$\mathbf{S}_p$/R]YPLSTEK
	P78312	V259M	S261	mTOR	0.00387	0.955	0.0271	0.0212	RSPPS[V/M]S$\mathbf{S}_p$ASSGSGS
	Q9BTC0	S660N	S660	DNAPK	0.00101	0.967	0	0.022	PGRLGAM[$\mathbf{S}_p$/N]AAPSQPN
	P30533	Q244K	S242	PKG1	0.706	0.656	0.0731	0.0237	LRRV$\mathbf{S}_p$H[Q/K]GY
	P52948-2	R538H	T536	CDK6	2.12e-06	0.234	0.0634	0.0237	TPTHYKL$\mathbf{T}_p$P[R/H]PATRV
	Q9BYJ9	V200I	T202	PKCE	0.0236	0.771	0.0258	0.0249	VSSSA[V/I]K$\mathbf{T}_p$VGSVVSS
	Q8N3K9	T2592I	T2592	PLK1	0.000765	0.842	0	0.0263	SFSLVKA[$\mathbf{T}_p$/I]SVTEKSE
	P50052	R350Q	S353	p90RSK	0.000294	0.586	0.0953	0.0317	SMSC[R/Q]KS$\mathbf{S}_p$SLREMET
	Q86UR5	R1113W	S1116	AMPKA1	0.0404	0.604	0.0128	0.0336	RA[R/W]SA$\mathbf{S}_p$TNCLR
	Q6W4X9	T1911M	T1911	GSK3B	0.0328	0.882	0	0.0347	SPSSFS[$\mathbf{T}_p$/M]AKTSTS
	Q8N4N8	R110C	T113	SGK1	0.17	0.976	0.000545	0.0352	[R/C]TA$\mathbf{T}_p$KWV
	P18583	S1782F	S1782	CK2A1	0.994	1	0	0.0391	PERASE[$\mathbf{S}_p$/F]SSEEKD
	Q5M775	S312P	S312	CK1A	0.323	0.985	0	0.0395	HGNALRT[$\mathbf{S}_p$/P]GSSSSDV
	Q9ULE3	S310A	S310	mTOR	0.00139	0.885	0	0.0427	PPPPLPS[$\mathbf{S}_p$/A]PPPSSVN
	Q9UL68	S237R	S237	CK2A1	0.998	1	0	0.0438	NSLEDD[$\mathbf{S}_p$/R]DKNENL
	O94988	R648Q	S650	AurA	0.0837	0.963	0.0028	0.0456	R[R/Q]R$\mathbf{S}_p$SSL
	Q9ULE3	S310A	S310	GSK3B	0.00684	0.921	0	0.0465	PPPLPS[$\mathbf{S}_p$/A]PPPSSV
Breast	P14859	S88F	S88	DNAPK	1	1	0	1.95e-06	SQQPSQP[$\mathbf{S}_p$/F]QQPSVQA
	P14859-5	S111F	S111	DNAPK	1	1	0	8.66e-06	SQQPSQP[$\mathbf{S}_p$/F]QQPSVQA
	P43355	K278T	Y276	Brk	0.998	1	0.000491	0.000103	RALAETS$\mathbf{Y}_p$V[K/T]VLEYV
	P03372	H6Y	T2	VRK1	0.0492	0.0792	0.00421	0.00111	M$\mathbf{T}_p$MTL[H/Y]TKA
	Q99490	D816Y	S818	P38B	0.0186	0.587	0.000173	0.00126	CTPSG[D/Y]L$\mathbf{S}_p$PLSREPP
	P54646	S523G	S527	p90RSK	0.36	0.78	0.00501	0.00214	LTG[S/G]TLS$\mathbf{S}_p$VSPRLGS
	P43487	E16D	T13	CK2A2	0.0196	0.135	0.021	0.00259	DTHEDHD$\mathbf{T}_p$ST[E/D]NTDE
	Q9NTX9	S95G	S95	CK2A1	1	1	0	0.00313	ADEDSA[$\mathbf{S}_p$/G]DLSDSE
	Q9NTX9	S95G	S95	CK1D	0.697	0.993	0	0.00354	NADEDSA[$\mathbf{S}_p$/G]DLSDSER
	P54646	S523G	S529	p90RSK	0.36	0.693	0.175	0.00412	G[S/G]TLSSV$\mathbf{S}_p$PRLGSHT
	P43487	E16D	S14	CK2A2	0.0196	0.02	0.0116	0.00597	THEDHDT$\mathbf{S}_p$T[E/D]NTDES
	O14681	T319A	S326	CDK6	1.96e-05	0.451	0.0093	0.00642	[T/A]SAEKFP$\mathbf{S}_p$PHPSPAK
	Q96RK0	E104K	S105	JNK1	0.00186	0.98	0.182	0.0112	PGATCP[E/K]$\mathbf{S}_p$PGPGPPH
	Q702N8	L929H	S925	PKD2	0.000176	6.29e-05	1.54e-08	0.0134	SKASERS$\mathbf{S}_p$VQL[L/H]ASC
	Q9H8V3	T833P	T833	PKCA	0.699	0.951	0	0.0173	SRAIKK[$\mathbf{T}_p$/P]SKKVTR
	Q9NTX9	S95G	S93	CK1D	0.697	0.953	0.0528	0.0178	KENADED$\mathbf{S}_p$A[S/G]DLSDS
	Q86UP2	T1316P	T1316	GRK2	0.19	0.736	0	0.0304	NSDVSPE[$\mathbf{T}_p$/P]ESSEKET
	Q86UP2	T1316P	S1319	GRK2	0.19	0.678	0.000182	0.0357	VSPE[T/P]ES$\mathbf{S}_p$EKETMSV
Colorectal	P04637	E271K	S269	CAMK2A	1	0.781	0.011	0.000131	NLLGRN$\mathbf{S}_p$F[E/K]VRVC
	Q9P253	A913S	S912	ERK5	0.495	0.848	0.0394	0.00475	APPPAKG$\mathbf{S}_p$[A/S]RAKEAE
	Q9NPD5	I292M	S293	CaMK4	0.832	0.524	3.21e-08	0.00785	ERK[I/M]$\mathbf{S}_p$LSLH
	Q6ZMN7	G784R	S783	CaMK4	0.792	0.454	1.52e-06	0.00954	TQSS$\mathbf{S}_p$[G/R]QSS
	Q92953	V450I	S448	ROCK1	0.326	0.845	0.00479	0.012	RAKRNG$\mathbf{S}_p$I[V/I]SMNL
	Q6ZMN7	G784R	S781	CaMK4	0.792	0.171	1.38e-05	0.0131	AATQ$\mathbf{S}_p$SS[G/R]Q
	Q92729	R856C	S853	CDK1	0.0472	0.903	0.0725	0.0164	LGG$\mathbf{S}_p$PR[R/C]
	P17936	T7M	T7	PKACA	0.311	0.903	0	0.0205	QRARP[$\mathbf{T}_p$/M]LWAAA
	Q9BTA9	S475L	S475	CDK2	0.055	0.919	0	0.0275	ISTPPV[$\mathbf{S}_p$/L]SQPKVS
Liver	P35222	T41A	T41	GSK3A	1	1	0	2.28e-09	GIHSGAT[$\mathbf{T}_p$/A]TAPSLSG
	P35222	S37F	S37	GSK3A	1	1	0	2.28e-09	YLDSGIH[$\mathbf{S}_p$/F]GATTTAP
	P35222	T41A	T41	IKKA	1	1	0	7.97e-09	GIHSGAT[$\mathbf{T}_p$/A]TAPSLSG
	P35222	S37F	S37	IKKA	1	1	0	7.97e-09	YLDSGIH[$\mathbf{S}_p$/F]GATTTAP
	P35222	T41A	T41	GSK3B	1	0.997	0	8.13e-06	IHSGAT[$\mathbf{T}_p$/A]TAPSLS

Continued on next page

	Protein	Variant	Phos.	Kinase	$R_{subst.}$	R_{site}	V_{site}	$E_{combined}$	Peptide
									Continued from previous page
	P35222	S37F	S37	JNK1	1	0.984	0	1.33e-05	YLDSGIH[**S**$_p$/F]GATTTAP
	P35222	T41A	S45	GSK3A	1	0.11	1.16e-06	2e-05	GAT[T/A]TAP**S**$_p$LSGKGNP
	P35222	S37F	T40	GSK3A	1	1	0.923	2.14e-05	SGIH[S/F]GA**T**TTAPSLS
	P35222	T41A	T41	JNK1	1	0.961	0	2.15e-05	GIHSGA**T**$_p$[T/A]TAPSLSG
	P35222	T41A	S47	GSK3A	1	0.0535	1.43e-06	2.48e-05	T[T/A]TAPSL**S**$_p$GKGNPEE
	P35222	T41A	T40	GSK3A	1	1	0.965	2.95e-05	SGIHSGA**T**$_p$[T/A]TAPSLS
	P35222	S37F	S33	CK2A2	1	0.0545	0.00319	4.62e-05	QQQSYLD**S**$_p$GIH[S/F]GAT
	P35222	T41A	T40	JNK1	1	0.896	0.0106	4.69e-05	SGIHSGA**T**$_p$[T/A]TAPSLS
	Q13950	V203F	T205	Pim1	1	1	0.0302	4.7e-05	TLTIT[V/F]F**T**$_p$NPPQVAT
	Q9Y6B2	S8C	S2	MSK1	0.717	0.799	0.00184	7.28e-05	M**S**$_p$EMAEL[S/C]E
	Q9Y463	K47N	S42	CAMK1A	0.981	0.948	0.00136	8.16e-05	FRDAT**S**$_p$APLR[K/N]
	P35222	T41A	S47	IKKA	1	0.882	0.21	0.000184	T[T/A]TAPSL**S**$_p$GKGNPEE
	P35222	S37F	T40	JNK1	1	0.896	0.198	0.000188	SGIH[S/F]GA**T**TTAPSLS
	P04083	E139V	T132	CK2A2	0.431	0.773	0.494	0.000875	AAMKGLG**T**$_p$DEDTLI[E/V]
	Q92997	S188G	S188	CK1D	0.999	0.954	0	0.00164	SSELETT[**S**$_p$/G]FFDSDED
	Q8IXF0	T461I	S458	p70S6K	0.27	0.964	0.00343	0.00173	PEKTSE**S**$_p$SE[T/I]SDS
	Q8IXF0	T461I	S464	GRK2	0.998	0.999	0.175	0.00208	ESSE[T/I]SD**S**$_p$ESDSKDT
	Q8IXF0	T461I	S459	p70S6K	0.27	0.927	0.00059	0.00262	EKTSES**S**$_p$E[T/I]SDSE
	Q8IXF0	T461I	T461	GRK2	0.998	0.763	0	0.00273	KTSESSE[**T**$_p$/I]SDSESDS
	P27816	S867G	S867	MSK1	0.0005	0.00983	0	0.0057	RPKSTST[**S**$_p$/G]SMKKTTT
	Q8IXF0	T461I	T461	CK2A1	0.998	1	0	0.0212	TSESSE[**T**$_p$/I]SDSESD
	O14647	E167G	S165	CK1D	0.282	0.962	0.0494	0.0265	DEQEQGT**S**$_p$A[E/G]SEPEQ
	O43347	L308Q	T304	NEK6	8.54e-05	6.19e-05	8.07e-11	0.0291	PGSTPSR**T**$_p$GGF[L/Q]GTT
	Q01081	R202H	S206	PAK1	0.0227	0.915	0.0301	0.0359	[R/H]RSR**S**$_p$RDRG
	Q9Y463	K47N	S49	PKG1	0.51	0.454	0.000119	0.0395	LR[K/N]L**S**$_p$VDLI
	Q9Y463	L48F	S49	PKG1	0.51	0.454	0.00388	0.0398	LRK[L/F]**S**$_p$VDLI
Pancreatic	Q9BYV9	T519I	T519	p70S6K	1	1	0	1.14e-08	LETRTR[**T**$_p$/I]SSSCSS
	P04637	E271K	S269	CAMK2A	1	0.781	0.011	0.000131	NLLGRN**S**$_p$F[E/K]VRVC
	Q9BYV9	T519I	S525	p70S6K	1	0.84	0.0917	0.000294	[T/I]SSSCS**S**$_p$YSYAED
	P56715	A135V	S137	MARK2	0.848	0.987	2.83e-06	0.000418	IS[A/V]H**S**$_p$PPHP
	P05129	P524R	Y521	Brk	0.356	0.227	0.000117	0.000629	TFCGTPD**Y**IA[P/R]EIIA
	P05129	P524R	T518	LKB1	1	0.156	0.000956	0.00106	TTRTFCG**T**$_p$PDYIA[P/R]E
	P16333	R63W	S66	CDK3	0.000153	0.000188	4.82e-05	0.00365	KNSA[R/W]KA**S**$_p$IVKNLKD
	O95954	R446W	S449	PKD1	0.184	0.982	0.0623	0.00366	LR[R/W]AV**S**$_p$VPLTL
	Q6P0Q8	K1420N	S1418	CAMK1A	0.145	0.0533	0.00125	0.00433	AALAA**S**$_p$E[K/N]KLA
	Q8NEV4	R1358L	S1355	PKCB	0.831	0.995	0.0408	0.00473	VFIQ**S**$_p$KY[R/L]G
	Q9UQ35	R2530Q	S2532	AurB	0.413	0.993	0.00288	0.00483	E[R/Q]R**S**$_p$SSS
	Q9NZ56	R446Q	S450	RSK2	0.0283	0.934	0.107	0.00676	KR[R/Q]PEP**S**$_p$LSRGSR
	O95935	P122L	S121	CDK1	0.0235	0.821	0.00407	0.0343	PKG**S**$_p$[P/L]AR

Bibliography

[1] Chatr-aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., OÕDonnell, L., Reguly, T., Breitkreutz, A., Sellam, A., Chen, D., Chang, C., Rust, J., Livstone, M., Oughtred, R., Dolinski, K., and Tyers, M. (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.,* **41**(D1), D816–D823.

[2] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., and von Mering, C. (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.,* **43**(D1), D447–D452.

[3] Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G., and Hermjakob, H. (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.,* **42**(D1), D358–D363.

[4] Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.,* **40**(D1), D261–D270.

[5] 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**(suppl 1), D240–D244.

[6] Gnad, F., Gunawardena, J., and Mann, M. (2011) PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res.,* **39**(suppl 1), D253–D260.

[7] Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009) Human Protein Reference Database 2009 update. *Nucleic Acids Res.,* **37**(suppl 1), D767–D772.

[8] Lu, C.-T., Huang, K.-Y., Su, M.-G., Lee, T.-Y., Bretaña, N. A., Chang, W.-C., Chen, Y.-J., Chen, Y.-J., and Huang, H.-D. (2013) dbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Res.,* **41**(D1), D295–D305.

[9] Gupta, R., Birch, H., Rapacki, K., Brunak, S., and Hansen, J. E. (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res.,* **27**(1), 370–372.

[10] Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Lu, F., Marchler, G. H., Mullokandov, M., Omelchenko, M. V., Robertson, C. L., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., Zhang, N., Zheng, C., and Bryant, S. H. (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.,* **39**(suppl 1), D225–D229.

[11] Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M. J., Kitano, H., and Thomas, P. D. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.,* **33**(suppl 1), D284–D288.

[12] Hoffmann, R. and Valencia, A. (2004) A gene network for navigating the literature. *Nat. Genet.,* **36**(7), 664–664.

[13] Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.,* **37**(1), 1–13.

[14] Choudhary, C., Kumar, C., Gnad, F., Nielsen, M. L., Rehman, M., Walther, T. C., Olsen, J. V., and Mann, M. (2009) Lysine Acetylation Targets Protein Complexes and Co-Regulates Major Cellular Functions. *Science,* **325**(5942), 834–840.

[15] Hwang, C.-S., Shemorry, A., and Varshavsky, A. (2010) N-Terminal Acetylation of Cellular Proteins Creates Specific Degradation Signals. *Science,* **327**(5968), 973–977.

[16] 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**(6), 1633–1649.

[17] Matsuoka, S., Ballif, B. A., Smogorzewska, A., McDonald, E. R., Hurov, K. E., Luo, J., Bakalarski, C. E., Zhao, Z., Solimini, N., Lerenthal, Y., Shiloh, Y., Gygi, S. P., and Elledge, S. J. (2007) ATM and ATR Substrate Analysis Reveals Extensive Protein Networks Responsive to DNA Damage. *Science,* **316**(5828), 1160–1166.

[18] Trost, B. and Kusalik, A. (2011) Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics,* **27**(21), 2927–2935.

[19] Ubersax, J. A. and Ferrell Jr, J. E. (07, 2007) Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell Biol.,* **8**(7), 530–541.

[20] Good, M. C., Zalatan, J. G., and Lim, W. A. (2011) Scaffold Proteins: Hubs for Controlling the Flow of Cellular Information. *Science,* **332**(6030), 680–686.

[21] Scott, J. D. and Pawson, T. (2009) Cell Signaling in Space and Time: Where Proteins Come Together and When They're Apart. *Science,* **326**(5957), 1220–1224.

[22] Hunter, T. (2007) The Age of Crosstalk: Phosphorylation, Ubiquitination, and Beyond. *Molecular Cell,* **28**(5), 730 – 738.

[23] Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., and Dunker, A. K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.,* **32**(3), 1037–1049.

[24] Su, M.-G. and Lee, T.-Y. (2013) Incorporating substrate sequence motifs and spatial amino acid composition to identify kinase-specific phosphorylation sites on protein three-dimensional structures. *BMC Bioinformatics,* **14**(Suppl 16), S2.

[25] Brinkworth, R. I., Breinl, R. A., and Kobe, B. (2003) Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl. Acad. Sci. U. S. A.,* **100**(1), 74–79.

[26] Durek, P., Schudoma, C., Weckwerth, W., Selbig, J., and Walther, D. (2009) Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinformatics,* **10**(1), 117.

[27] Zhu, G., Liu, Y., and Shaw, S. (2005) Protein Kinase Specificity: A Strategic Collaboration between Kinase Peptide Specificity and Substrate Recruitment. *Cell Cycle,* **4**, 52 – 56.

[28] Olsen, J. V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M. L., Jensen, L. J., Gnad, F., Cox, J., Jensen, T. S., Nigg, E. A., Brunak, S., and Mann, M. (2010) Quantitative Phosphoproteomics Reveals Widespread Full Phosphorylation Site Occupancy During Mitosis. *Sci. Signal.,* **3**(104), ra3.

[29] Burnett, G. and Kennedy, E. P. (1954) The enzymatic phosphorylation of proteins. *J. Biol. Chem.,* **211**(2), 969–980.

[30] Cohen, P. (2002) The origins of protein phosphorylation. *Nat. Cell Biol.,* **4**(5), E127–E130.

[31] Cohen, P. (2002) Protein kinases – the major drug targets of the twenty-first century?. *Nat. Rev. Drug Discov.,* **1**(4), 309–315.

[32] Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002) The Protein Kinase Complement of the Human Genome. *Science,* **298**(5600), 1912–1934.

[33] Swanson, R. V., Alex, L. A., and Simon, M. I. (1994) Histidine and aspartate phosphorylation: two-component systems and the limits of homology. *Trends Biochem. Sci.,* **19**(11), 485 – 490.

[34] Thomson, M. and Gunawardena, J. (2009) Unlimited multistability in multisite phosphorylation systems. *Nature,* **460**(7252), 274–277.

[35] Doerr, A. (2008) Phosphorylation and the cell cycle. *Nat. Meth.,* **5**(10), 858–859.

[36] Hochegger, H., Takeda, S., and Hunt, T. (2008) Cyclin-dependent kinases and cell-cycle transitions: does one fit all?. *Nat. Rev. Mol. Cell Biol.,* **9**(11), 910–916.

[37] Sherr, C., Kato, J., Quelle, D., Matsuoka, M., and Roussel, M. (1994) D-type cyclincs and their cyclin-dependent kinases – G1 phase integrators of the mitogenic response. *Cold Spring Harbor Symp. Quant. Biol.,* **59**, 11–19.

[38] Bartek, J. and Lukas, J. (2001) Pathways governing G1/S transition and their response to DNA damage. *FEBS Letters,* **490**(3), 117 – 122.

[39] Das-Bradoo, S. and Bielinsky, A. (2010) DNA Replication and Checkpoint Control in S Phase.. *Nature Education,* **3**(9), 50.

[40] Cimprich, K. A. and Cortez, D. (2008) ATR: an essential regulator of genome integrity. *Nat. Rev. Mol. Cell Biol.,* **9**(8), 616–627.

[41] Johnson, L. N. (2011) Substrates of Mitotic Kinases. *Sci. Signal.,* **4**(179), pe31.

[42] Ma, H. T. and Poon, R. Y. C. (2011) How protein kinases co-ordinate mitosis in animal cells. *Biochem. J.,* **435**(1), 17–31.

[43] Gavet, O. and Pines, J. (2010) Progressive Activation of CyclinB1-Cdk1 Coordinates Entry to Mitosis. *Dev. Cell,* **18**(4), 533 – 543.

[44] Gauthier, N. P., Jensen, L. J., Wernersson, R., Brunak, S., and Jensen, T. S. (2010) Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. *Nucleic Acids Res.,* **38**(suppl 1), D699–D702.

[45] Carmena, M. and Earnshaw, W. C. (2003) The cellular geography of Aurora kinases. *Nat. Rev. Mol. Cell Biol.,* **4**(11), 842–854.

[46] Kobe, B., Kampmann, T., Forwood, J. K., Listwan, P., and Brinkworth, R. I. (2005) Substrate specificity of protein kinases and computational prediction of substrates. *Biochim. Biophys. Acta,* **1754**(1-2), 200 – 209.

[47] Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004) WebLogo: A Sequence Logo Generator. *Genome Res.,* **14**(6), 1188–1190.

[48] Pearce, L. R., Komander, D., and Alessi, D. R. (2010) The nuts and bolts of AGC protein kinases. *Nat. Rev. Mol. Cell Biol.,* **11**(1), 9–22.

[49] Colledge, M. and Scott, J. D. (1999) AKAPs: from structure to function. *Trends in Cell Biology,* **9**(6), 216 – 221.

[50] Grossmann, A., Benlasfer, N., Birth, P., Hegele, A., Wachsmuth, F., Apelt, L., and Stelzl, U. (2015) Phospho-tyrosine dependent protein–protein interaction network. *Mol. Syst. Biol.,* **11**(3).

[51] Duan, G. and Walther, D. (2015) The Roles of Post-translational Modifications in the Context of Protein Interaction Networks. *PLoS Comput. Biol.,* **11**, 1–23.

[52] Bloom, J. and Cross, F. R. (2007) Multiple levels of cyclin specificity in cell-cycle control. *Nat. Rev. Mol. Cell Biol.,* **8**(2), 149–160.

[53] Faust, M. and Montenarh, M. (2000) Subcellular localization of protein kinase CK2. *Cell Tissue Res.,* **301**(3), 329–340.

[54] Hiromura, K., Pippin, J. W., Blonski, M. J., Roberts, J. M., and Shankland, S. J. (2002) The subcellular localization of cyclin dependent kinase 2 determines the fate of mesangial cells: role in apoptosis and proliferation. *Oncogene,* **21**(11), 1750–1758.

[55] Baldin, V. and Ducommun, B. (1995) Subcellular localisation of human wee1 kinase is regulated during the cell cycle. *J. Cell Sci.,* **108**(6), 2425–2432.

[56] von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.,* **33**(suppl 1), D433–D437.

[57] Dephoure, N., Zhou, C., Villén, J., Beausoleil, S. A., Bakalarski, C. E., Elledge, S. J., and Gygi, S. P. (2008) A quantitative atlas of mitotic phosphorylation. *Proc. Natl. Acad. Sci. U. S. A.,* **105**(31), 10762–10767.

[58] Kim, S.-T., Lim, D.-S., Canman, C. E., and Kastan, M. B. (1999) Substrate Specificities and Identification of Putative Substrates of ATM Kinase Family Members. *J. Biol. Chem.,* **274**(53), 37538–37543.

[59] Chen, S., Xu, Y., Yuan, X., Bubley, G. J., and Balk, S. P. (2006) Androgen receptor phosphorylation and stabilization in prostate cancer by cyclin-dependent kinase 1. *Proc. Natl. Acad. Sci. U. S. A.,* **103**(43), 15969–15974.

[60] Matthess, Y., Raab, M., Sanhaji, M., Lavrik, I. N., and Strebhardt, K. (2010) Cdk1/Cyclin B1 Controls Fas-Mediated Apoptosis by Regulating Caspase-8 Activity. *Mol. Cell. Biol.,* **30**(24), 5726–5740.

[61] Rajanala, K., Sarkar, A., Jhingan, G. D., Priyadarshini, R., Jalan, M., Sengupta, S., and Nandicoori, V. K. (2014) Phosphorylation of nucleoporin Tpr governs its differential localization and is required for its mitotic function. *J. Cell Sci.,* **127**(16), 3505–3520.

[62] Marazita, M. C., Ogara, M. F., Sonzogni, S. V., Martí, M., Dusetti, N. J., Pignataro, O. P., and Cánepa, E. T. (2012) CDK2 and PKA Mediated-Sequential Phosphorylation Is Critical for p19INK4d Function in the DNA Damage Response. *PLoS ONE,* **7**(4), e35638.

[63] Marais, A., Ji, Z., Child, E. S., Krause, E., Mann, D. J., and Sharrocks, A. D. (2010) Cell Cycle-dependent Regulation of the Forkhead Transcription Factor FOXK2 by CDK·Cyclin Complexes. *J. Biol. Chem.,* **285**(46), 35728–35739.

[64] Ruffner, H., Jiang, W., Craig, A. G., Hunter, T., and Verma, I. M. (1999) BRCA1 Is Phosphorylated at Serine 1497 In Vivo at a Cyclin-Dependent Kinase 2 Phosphorylation Site. *Mol. Cell. Biol.,* **19**(7), 4843–4854.

[65] Narayan, N., Massimi, P., and Banks, L. (2009) CDK phosphorylation of the discs large tumour suppressor controls its localisation and stability. *J. Cell Sci.,* **122**(1), 65–74.

[66] Li, P.-F., Li, J., Müller, E.-C., Otto, A., Dietz, R., and von Harsdorf, R. (2002) Phosphorylation by Protein Kinase CK2: A Signaling Switch for the Caspase-Inhibiting Protein ARC. *Mol. Cell,* **10**(2), 247 – 258.

[67] Mueller, T., Breuer, P., Schmitt, I., Walter, J., Evert, B. O., and Wüllner, U. (2009) CK2-dependent phosphorylation determines cellular localization and stability of ataxin-3. *Hum. Mol. Genet.,* **18**(17), 3334–3343.

[68] Jarrett, S. G., Horrell, E. M. W., Christian, P. A., Vanover, J. C., Boulanger, M. C., Zou, Y., and D'Orazio, J. A. (2014) PKA-Mediated Phosphorylation of ATR Promotes Recruitment of XPA to UV-Induced DNA Damage. *Mol. Cell,* **54**(6), 999 – 1011.

[69] Mok, J., Kim, P. M., Lam, H. Y. K., Piccirillo, S., Zhou, X., Jeschke, G. R., Sheridan, D. L., Parker, S. A., Desai, V., Jwa, M., Cameroni, E., Niu, H., Good, M., Remenyi, A., Ma, J.-L. N., Sheu, Y.-J., Sassi, H. E., Sopko, R., Chan, C. S. M., De Virgilio, C., Hollingsworth, N. M., Lim, W. A., Stern, D. F., Stillman, B., Andrews, B. J., Gerstein, M. B., Snyder, M., and Turk, B. E. (2010) Deciphering Protein Kinase Specificity Through Large-Scale Analysis of Yeast Phosphorylation Site Motifs. *Sci. Signal.,* **3**(109), ra12.

[70] Chi, Y., Welcker, M., Hizli, A., Posakony, J., Aebersold, R., and Clurman, B. (2008) Identification of CDK2 substrates in human cell lysates. *Genome Biol.,* **9**(10), R149.

[71] Blethrow, J. D., Glavy, J. S., Morgan, D. O., and Shokat, K. M. (2008) Covalent capture of kinase-specific phosphopeptides reveals Cdk1-cyclin B substrates. *Proc. Natl. Acad. Sci. U. S. A.,* **105**(5), 1442–1447.

[72] Ubersax, J. A., Woodbury, E. L., Quang, P. N., Paraz, M., Blethrow, J. D., Shah, K., Shokat, K. M., and Morgan, D. O. (2003) Targets of the cyclin-dependent kinase Cdk1. *Nature,* **425**(6960), 859–864.

[73] Stark, C., Su, T.-C., Breitkreutz, A., Lourenco, P., Dahabieh, M., Breitkreutz, B.-J., Tyers, M., and Sadowski, I. (2010) PhosphoGRID: a database of experimentally verified in vivo protein phosphorylation sites from the budding yeast Saccharomyces cerevisiae. *Database,* **2010**.

[74] Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., and Diella, F. (2010) Phospho.ELM: a database of phosphorylation sites - udpate 2011. *Nucleic Acids Res.,* **39**, 1–7.

[75] Blom, N., Gammeltoft, S., and Brunak, S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.,* **294**(5), 1351 – 1362.

[76] Chen, X., Shi, S.-P., Suo, S.-B., Xu, H.-D., and Qiu, J.-D. (2015) Proteomic analysis and prediction of human phosphorylation sites in subcellular level reveal subcellular specificity. *Bioinformatics,* **31**(2), 194–200.

[77] Dou, Y., Yao, B., and Zhang, C. (2014) PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino Acids,* **46**(6), 1459–1469.

[78] Zhou, F.-F., Xue, Y., Chen, G.-L., and Yao, X. (2004) GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem. Biophys. Res. Commun.,* **325**(4), 1443 – 1448.

[79] Xue, Y., Ren, J., Gao, X., Jin, C., Wen, L., and Yao, X. (2008) GPS 2.0, a Tool to Predict Kinase-specific Phosphorylation Sites in Hierarchy. *Mol. Cell. Proteomics,* **7**(9), 1598–1608.

[80] Xue, Y., Liu, Z., Cao, J., Ma, Q., Gao, X., Wang, Q., Jin, C., Zhou, Y., Wen, L., and Ren, J. (2011) GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Eng., Des. Sel.,* **24**(3), 255–260.

[81] Saunders, N., Brinkworth, R., Huber, T., Kemp, B., and Kobe, B. (2008) Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC Bioinformatics,* **9**(1), 245.

[82] Saunders, N. F. W. and Kobe, B. (2008) The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information. *Nucleic Acids Res.,* **36**(suppl 2), W286–W290.

[83] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.,* **31**(1), 365–370.

[84] Wagih, O., Reimand, J., and Bader, G. D. (2015) MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nat. Meth.,* **12**(6), 531–533.

[85] Obenauer, J. C., Cantley, L. C., and Yaffe, M. B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.,* **31**(13), 3635–3641.

[86] Songyang, Z., Blechner, S., Hoagland, N., Hoekstra, M. F., Piwnica-Worms, H., and Cantley, L. C. (1994) Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr. Biol.,* **4**(11), 973 – 982.

[87] Obenauer, J. C., Cantley, L. C., and Yaffe, M. B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.,* **31**(13), 3635–3641.

[88] Miller, M. L., Jensen, L. J., Diella, F., Jørgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S. A., Bordeaux, J., Sicheritz-Ponten, T., Olhovsky, M., Pasculescu, A., Alexander, J., Knapp, S., Blom, N., Bork, P., Li, S., Cesareni, G., Pawson, T., Turk, B. E., Yaffe, M. B., Brunak, S., and Linding, R. (2008) Linear Motif Atlas for Phosphorylation-Dependent Signaling. *Sci. Signaling,* **1**(35), ra2–ra2.

[89] Xue, Y., Li, A., Wang, L., Feng, H., and Yao, X. (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics,* **7**(1), 163.

[90] Wong, Y.-H., Lee, T.-Y., Liang, H.-K., Huang, C.-M., Wang, T.-Y., Yang, Y.-H., Chu, C.-H., Huang, H.-D., Ko, M.-T., and Hwang, J.-K. (2007) KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.,* **35**(suppl 2), W588–W594.

[91] Dang, T. H., Van Leemput, K., Verschoren, A., and Laukens, K. (2008) Prediction of kinase-specific phosphorylation sites using conditional random fields. *Bioinformatics,* **24**(24), 2857–2864.

[92] Ryu, G.-M., Song, P., Kim, K.-W., Oh, K.-S., Park, K.-J., and Kim, J. H. (2009) Genome-wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases. *Nucleic Acids Research,* **37**(4), 1297–1307.

[93] Gao, J., Thelen, J. J., Dunker, A. K., and Xu, D. (2010) Musite, a Tool for Global Prediction of General and Kinase-specific Phosphorylation Sites. *Mol. Cell. Proteomics,* **9**(12), 2586–2600.

[94] Li, T., Du, P., and Xu, N. (2010) Identifying Human Kinase-Specific Protein Phosphorylation Sites by Integrating Heterogeneous Information from Various Sources. *PLoS ONE,* **5**(11), e15411.

[95] Ellis, J. J. and Kobe, B. (2011) Predicting Protein Kinase Specificity: Predikin Update and Performance in the DREAM4 Challenge. *PLoS ONE,* **6**(7), e21169.

[96] Zou, L., Wang, M., Shen, Y., Liao, J., Li, A., and Wang, M. (2013) PKIS: computational identification of protein kinases for experimentally discovered protein phosphorylation sites. *BMC Bioinformatics,* **14**(1), 247.

[97] Horn, H., Schoof, E. M., Kim, J., Robin, X., Miller, M. L., Diella, F., Palma, A., Cesareni, G., Jensen, L. J., and Linding, R. (2014) KinomeXplorer: an integrated platform for kinome biology studies. *Nat. Meth.,* **11**(6), 603–604.

[98] Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (10, 2008) Support Vector Machines and Kernels for Computational Biology. *PLoS Comput. Biol.,* **4**(10), e1000173.

[99] Biswas, A., Noman, N., and Sikder, A. (2010) Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. *BMC Bioinformatics,* **11**(1), 273.

[100] Huang, H.-D., Lee, T.-Y., Tzeng, S.-W., and Horng, J.-T. (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.,* **33**(suppl 2), W226–W229.

[101] Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., and Obradovic, Z. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics,* **7**(1), 208.

[102] Bryson, K., McGuffin, L. J., Marsden, R. L., Ward, J. J., Sodhi, J. S., and Jones, D. T. (2005) Protein structure prediction servers at University College London. *Nucleic Acids Res.,* **33**(suppl 2), W36–W38.

[103] Ahmad, S., Gromiha, M. M., and Sarai, A. (2003) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics,* **19**(14), 1849–1851.

[104] Ingrell, C. R., Miller, M. L., Jensen, O. N., and Blom, N. (2007) NetPhosYeast: prediction of protein phosphorylation sites in yeast. *Bioinformatics,* **23**(7), 895–897.

[105] Blom, N., Hansen, J., Brunak, S., and Blaas, D. (1996) Cleavage site analysis in picornaviral polyproteins: Discovering cellular targets by neural networks. *Protein Science,* **5**(11), 2203–2216.

[106] Blom, N., Kreegipuu, A., and Brunak, S. (1998) PhosphoBase: A database of phosphorylation sites. *Nucleic Acids Res.,* **26**(1), 382–386.

[107] Fan, W., Xu, X., Shen, Y., Feng, H., Li, A., and Wang, M. (2014) Prediction of protein kinase-specific phosphorylation sites in hierarchical structure using functional information and random forest. *Amino Acids,* **46**(4).

[108] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., and Mering, C. v. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.,* **39**(suppl 1), D561–D568.

[109] Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. A., Jörgensen, C., Miron, I. M., Diella, F., Colwill, K., Taylor, L., Elder, K., Metalnikov, P., Nguyen, V., Pasculescu, A., Jin, J., Park, J. G., Samson, L. D., Woodgett, J. R., Russell, R. B., Bork, P., Yaffe, M. B., and Pawson, T. (2007) Systematic Discovery of In Vivo Phosphorylation Networks. *Cell,* **129**(7), 1415 – 1426.

[110] Harbour, J., Luo, R. X., Santi, A. D., Postigo, A. A., and Dean, D. C. (1999) Cdk Phosphorylation Triggers Sequential Intramolecular Interactions that Progressively Block Rb Functions as Cells Move through G1. *Cell,* **98**(6), 859 – 869.

[111] Sherr, C. J. and Roberts, J. M. (1999) CDK inhibitors: positive and negative regulators of G1-phase progression. *Genes & Dev.,* **13**(12), 1501–1512.

[112] Coverley, D., Laman, H., and Laskey, R. A. (2002) Distinct roles for cyclins E and A during DNA replication complex assembly and activation. *Nat. Cell. Biol.,* **4**(7), 523–528.

[113] Huttlin, E. L., Jedrychowski, M. P., Elias, J. E., Goswami, T., Rad, R., Beausoleil, S. A., Villén, J., Haas, W., Sowa, M. E., and Gygi, S. P. (2010) A Tissue-Specific Atlas of Mouse Protein Phosphorylation and Expression. *Cell,* **143**(7), 1174 – 1189.

[114] Diella, F., Cameron, S., Gemund, C., Linding, R., Via, A., Kuster, B., Sicheritz-Ponten, T., Blom, N., and Gibson, T. (2004) Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics,* **5**(1), 79.

[115] Xue, L., Wang, W.-H., Iliuk, A., Hu, L., Galan, J. A., Yu, S., Hans, M., Geahlen, R. L., and Tao, W. A. (2012) Sensitive kinase assay linked with phosphoproteomics for identifying direct kinase substrates. *Proc. Natl. Acad. Sci. U. S. A.,* **109**(15), 5615–5620.

[116] Lim, S. and Kaldis, P. (2013) Cdks, cyclins and CKIs: roles beyond cell cycle regulation. *Development,* **140**(15), 3079–3093.

[117] Patrick, R., Cao, K.-A., Davis, M., Kobe, B., and Bodén, M. (2012) Mapping the stabilome: a novel computational method for classifying metabolic protein stability. *BMC Syst. Biol.,* **6**(1), 60.

[118] Bauer, D. C., Willadsen, K., Buske, F. A., Cao, K.-A. L., Bailey, T. L., Dellaire, G., and Bodén, M. (2011) Sorting the nuclear proteome. *Bioinformatics,* **27**(13), i7–i14.

[119] Mehdi, A., Sehgai, M., Kobe, B., Bailey, T., and Bodén, M. (2011) A probabilistic model of nuclear import of proteins. *Bioinformatics,* **27**(9), 1239–1246.

[120] Oniśko, A., Druzdzel, M. J., and Wasyluk, H. (2001) Learning Bayesian network parameters from small data sets: application of Noisy-OR gates. *Internat. J. Approx. Reason.,* **27**(2), 165 – 182.

[121] Do, C. B. and Batzoglou, S. (2008) What is the expectation maximization algorithm. *Nat. Biotechnol.,* **26**(8), 897–899.

[122] Baldi, P., Brunak, S., Chauvin, Y., Anderson, C. A. F., and Nielsen, H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics,* **16**(5), 412–424.

[123] ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature,* **489**(7414), 57–74.

[124] Lorna Morris, E. A. and Thangue, N. B. L. (2000) Regulation of E2F transcription by cyclin E-Cdk2 kinase mediated through p300/CBP co-activators. *Nat. Cell. Biol.,* **2**(4), 232 – 239.

[125] Attwoll, C., Denchi, E. L., and Helin, K. (2004) The E2F family: specific functions and overlapping interests. *EMBO J.,* **23**(24), 4709 – 4716.

[126] Lammens, T., Li, J., Leone, G., and Veylder, L. D. (2009) Atypical E2Fs: new players in the E2F transcription factor family. *Trends Cell Biol.,* **19**(3), 111 – 118.

[127] Biswas, A. K. and Johnson, D. G. (2012) Transcriptional and Nontranscriptional Functions of E2F1 in Response to DNA Damage. *Cancer Res.,* **72**(1), 13–17.

[128] Lee, B.-K., Bhinge, A. A., and Iyer, V. R. (2011) Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. *Nucleic Acids Res.,* **39**(9), 3558–3573.

[129] Wells, N. J. and Hickson, I. D. (1995) Human Topoisomerase IIα is Phosphorylated in a Cell-Cycle Phase-Dependent Manner by a Proline-Directed Kinase. *Eur. J. Biol. Chem.,* **231**(2), 491–497.

[130] Kraft, C., Herzog, F., Gieffers, C., Mechtler, K., Hagting, A., Pines, J., and Peters, J.-M. (2003) Mitotic regulation of the human anaphase-promoting complex by phosphorylation. *EMBO J.,* **22**(24), 6598–6609.

[131] Johnson, T. K., Schweppe, R. E., Septer, J., and Lewis, R. E. (1999) Phosphorylation of B-Myb Regulates Its Transactivation Potential and DNA Binding. *J. Biol. Chem.,* **274**(51), 36741–36749.

[132] Medunjanin, S., Hermani, A., De Servi, B., Grisouard, J., Rincke, G., and Mayer, D. (2005) Glycogen Synthase Kinase-3 Interacts with and Phosphorylates Estrogen Receptor α and Is Involved in the Regulation of Receptor Activity. *J. Biol. Chem.,* **280**(38), 33006–33014.

[133] Hansen, K., Farkas, T., Lukas, J., Holm, K., Ronnstrand, L., and Bartek, J. (2001) Phosphorylation-dependent and -independent functions of p130 cooperate to evoke a sustained G1 block. *EMBO J.,* **20**(3), 422–432.

[134] Deans, A. J., Khanna, K. K., McNees, C. J., Mercurio, C., Heierhorst, J., and McArthur, G. A. (2006) Cyclin-Dependent Kinase 2 Functions in Normal DNA Repair and Is a Therapeutic Target in BRCA1-Deficient Cancers. *Cancer Res.,* **66**(16), 8219–8226.

[135] Satyanarayana, A. and Kaldis, P. (2009) A dual role of Cdk2 in DNA damage response. *Cell Div.,* **4**(1), 9.

[136] Zhou, B.-B. S. and Elledge, S. J. (2000) The DNA damage response: putting checkpoints in perspective. *Nature,* **408**(6811), 433–439.

[137] Bakkenist, C. J. and Kastan, M. B. (2004) Initiating Cellular Stress Responses. *Cell,* **118**(1), 9 – 17.

[138] Hayami, R., Sato, K., Wu, W., Nishikawa, T., Hiroi, J., Ohtani-Kaneko, R., Fukuda, M., and Ohta, T. (2005) Down-regulation of BRCA1-BARD1 Ubiquitin Ligase by CDK2. *Cancer Res.,* **65**(1), 6–10.

[139] Huang, H., Regan, K. M., Lou, Z., Chen, J., and Tindall, D. J. (2006) CDK2-Dependent Phosphorylation of FOXO1 as an Apoptotic Response to DNA Damage. *Science,* **314**(5797), 294–297.

[140] Zhang, H.-G., Wang, J., Yang, X., Hsu, H.-C., and Mountz, J. D. (2004) Regulation of apoptosis proteins in cancer cells by ubiquitin. *Oncogene,* **23**(11), 2009–2015.

[141] DeGregori, J., Leone, G., Miron, A., Jakoi, L., and Nevins, J. R. (1997) Distinct roles for E2F proteins in cell growth control and apoptosis. *Proc. Natl. Acad. Sci. U. S. A.,* **94**(14), 7245–7250.

[142] Yang, W.-W., Wang, Z.-H., Zhu, Y., and Yang, H.-T. (2006) E2F6 negatively regulates ultraviolet-induced apoptosis via modulation of BRCA1. *Cell Death Differ.,* **14**(4), 807–817.

[143] Dayarian, A., Romero, R., Wang, Z., Biehl, M., Bilal, E., Hormoz, S., Meyer, P., Norel, R., Rhrissorrakrai, K., Bhanot, G., Luo, F., and Tarca, A. L. (2015) Predicting protein phosphorylation from gene expression: top methods from the IMPROVER Species Translation Challenge. *Bioinformatics,* **31**(4), 462–470.

[144] Dhillon, A. S., Hagan, S., Rath, O., and Kolch, W. (2007) MAP kinase signalling pathways in cancer. *Oncogene,* **26**(22), 3279–3290.

[145] Cheon, D.-J. and Orsulic, S. (2011) Mouse Models of Cancer. *Annu. Rev. Pathol.: Mech. Dis.,* **6**(1), 95–119.

[146] Poussin, C., Mathis, C., Alexopoulos, L. G., Messinis, D. E., Dulize, R. H. J., Belcastro, V., Melas, I. N., Sakellaropoulos, T., Rhrissorrakrai, K., Bilal, E., Meyer, P., Talikka, M., Boué, S., Norel, R., Rice, J. J., Stolovitzky, G., Ivanov, N. V., Peitsch, M. C., and Hoeng, J. (06, 2014) The species translation challenge—A systems biology perspective on human and rat bronchial epithelial cells. *Scientific Data,* **1**, 140009 EP –.

[147] Dunbar, S. A. (2006) Applications of Luminex®xMAP^{TM}technology for rapid, high-throughput multiplexed nucleic acid detection. *Clinica Chimica Acta,* **363**(1-2), 71 – 82.

[148] Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L. J. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.,* **41**(D1), D808–D815.

[149] Yang, Z. R. (2004) Biological applications of support vector machines. *Briefings Bioinf.,* **5**(4), 328–338.

[150] Mehdi, A. M., Patrick, R., Bailey, T. L., and Bodén, M. (2014) Predicting the Dynamics of Protein Abundance. *Mol. Cell. Proteomics,* **13**(5), 1330–1340.

[151] Biehl, M., Sadowski, P., Bhanot, G., Bilal, E., Dayarian, A., Meyer, P., Norel, R., Rhrissorrakrai, K., Zeller, M. D., and Hormoz, S. (2015) Inter-species prediction of protein phosphorylation in the sbv IMPROVER species translation challenge. *Bioinformatics,* **31**(4), 453–461.

[152] Mayr, B. and Montminy, M. (2001) Transcriptional regulation by the phosphorylation-dependent factor CREB. *Nat. Rev. Mol. Cell. Biol.,* **2**(8), 599–609.

[153] Liu, C., Srihari, S., Cao, K.-A. L., Chenevix-Trench, G., Simpson, P. T., Ragan, M. A., and Khanna, K. K. (2014) A fine-scale dissection of the DNA double-strand break repair machinery and its implications for breast cancer therapy. *Nucleic Acids Res.,* **42**(10), 6106–6127.

[154] Nardozzi, J., Lott, K., and Cingolani, G. (2010) Phosphorylation meets nuclear import: a review. *Cell Commun. Signaling,* **8**(1), 32.

[155] Hjerrild, M. and Gammeltoft, S. (2006) Phosphoproteomics toolbox: Computational biology, protein chemistry and mass spectrometry. *FEBS Lett.,* **580**(20), 4764 – 4770.

[156] Hjerrild, M., Stensballe, A., Rasmussen, T., Kofoed, C., Blom, N., Sicheritz-Ponten, T., Larsen, M., Brunak, S., Jensen, O., and Ganuneltoft, S. (2004) Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry. *J. Proteome Res.,* **3**(3), 426–433.

[157] Manning, B., Tee, A., Logsdon, M., Blenis, J., and Cantley, L. (2002) Identification of the tuberous sclerosis complex-2 tumor suppressor gene product tuberin as a target of the phosphoinositide 3-kinase/Akt pathway. *Mol. Cell,* **10**(1), 151–162.

[158] Wong, Y.-H., Lee, T.-Y., Liang, H.-K., Huang, C.-M., Wang, T.-Y., Yang, Y.-H., Chu, C.-H., Huang, H.-D., Ko, M.-T., and Hwang, J.-K. (2007) KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.,* **35**(suppl 2), W588–W594.

[159] Ebisuya, M., Kondoh, K., and Nishida, E. (2005) The duration, magnitude and compartmentalization of ERK MAP kinase activity: mechanisms for providing signaling specificity. *J. Cell Sci.,* **118**(14), 2997–3002.

[160] Lapenna, S. and Giordano, A. (2009) Cell cycle kinases as therapeutic targets for cancer. *Nat. Rev. Drug Discovery,* **8**(7), 547–566.

[161] Patrick, R., Lê Cao, K.-A., Kobe, B., and Bodén, M. (2015) PhosphoPICK: modelling cellular context to map kinase-substrate phosphorylation events. *Bioinformatics,* **31**(3), 382–389.

[162] Marfori, M., Mynott, A., Ellis, J. J., Mehdi, A. M., Saunders, N. F., Curmi, P. M., Forwood, J. K., Bodén, M., and Kobe, B. (2011) Molecular basis for specificity of nuclear import and prediction of nuclear localization. *Biochim. Biophys. Acta, Mol. Cell Res.,* **1813**(9), 1562 – 1577.

[163] Christie, M., Chang, C.-W., Róna, G., Smith, K. M., Stewart, A. G., Takeda, A. A., Fontes, M. R., Stewart, M., Vértessy, B. G., Forwood, J. K., and Kobe, B. (2015) Structural Biology and Regulation of Protein Import into the Nucleus. *J. Mol. Biol.,* pp. –.

[164] Róna, G., Borsos, M., Ellis, J. J., Mehdi, A. M., Christie, M., Környei, Z., Neubrandt, M., Tóth, J., Bozóky, Z., Buday, L., Madarász, E., Bodén, M., Kobe, B., and Vértessy, B. G. (2014) Dynamics of re-constitution of the human nuclear proteome after cell division is regulated by NLS-adjacent phosphorylation. *Cell Cycle,* **13**(22), 3551–3564.

[165] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009) BLAST+: architecture and applications. *BMC Bioinf.,* **10**(1), 1–9.

[166] Kosugi, S., Hasebe, M., Matsumura, N., Takashima, H., Miyamoto-Sato, E., Tomita, M., and Yanagawa, H. (2009) Six classes of nuclear localization signals specific to different binding grooves of importin α. *J. Biol. Chem.,* **284**(1), 478–485.

[167] Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, r., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Pontén, F. (2015) Tissue-based map of the human proteome. *Science,* **347**(6220).

[168] Consortium, T. F., the RIKEN PMI, and (DGT), C. (2014) A promoter-level mammalian expression atlas. *Nature,* **507**(7493), 462–470.

[169] Jans, D. A. (1995) The regulation of protein transport to the nucleus by phosphorylation. *Biochem. J.,* **311**(3), 705–716.

[170] Jans, D. A. and Hubner, S. (1996) Regulation of protein transport to the nucleus: central role of phosphorylation. *Physiol. Rev.,* **76**(3), 651–685.

[171] Róna, G., Marfori, M., Borsos, M., Scheer, I., Takács, E., Tóth, J., Babos, F., Magyar, A., Erdei, A., Bozóky, Z., Buday, L., Kobe, B., and Vértessy, B. G. (2013) Phosphorylation adjacent to the nuclear localization signal of human dUTPase abolishes nuclear import: structural and mechanistic insights. *Acta Crystallogr., Sect. D,* **69**(12), 2495–2505.

[172] FONTES, M. R. M., TEH, T., TOTH, G., JOHN, A., PAVO, I., JANS, D. A., and KOBE, B. (2003) Role of flanking sequences and phosphorylation in the recognition of the simian-virus-40 large T-antigen nuclear localization sequences by importin-α. *Biochemical Journal,* **375**(2), 339–349.

[173] Kosugi, S., Hasebe, M., Entani, T., Takayama, S., Tomita, M., and Yanagawa, H. (2008) Design of Peptide Inhibitors for the Importin α/β Nuclear Import Pathway by Activity-Based Profiling. *Chem. Biol.,* **15**(9), 940 – 949.

[174] Malki, S., Nef, S., Notarnicola, C., Thevenet, L., Gasca, S., Mèjean, C., Berta, P., Poulat, F., and Boizet-Bonhourne, B. (2005) Prostaglandin D2 induces nuclear import of the sex-determining factor SOX9 via its cAMP-PKA phosphorylation. *EMBO J.,* **24**(10), 1798–1809.

[175] Zhang, F., White, R. L., and Neufeld, K. L. (2000) Phosphorylation near nuclear localization signal regulates nuclear import of adenomatous polyposis coli protein. *Proc. Natl. Acad. Sci. U. S. A.,* **97**(23), 12577–12582.

[176] Goldenson, B. and Crispino, J. D. (2015) The aurora kinases in cell cycle and leukemia. *Oncogene,* **34**(5), 537–545.

[177] Guise, A. J., Greco, T. M., Zhang, I. Y., Yu, F., and Cristea, I. M. (2012) Aurora B-dependent Regulation of Class IIa Histone Deacetylases by Mitotic Nuclear Localization Signal Phosphorylation. *Mol. Cell. Proteomics,* **11**(11), 1220–1229.

[178] Biggs, W. H., Meisenhelder, J., Hunter, T., Cavenee, W. K., and Arden, K. C. (1999) Protein kinase B/Akt-mediated phosphorylation promotes nuclear exclusion of the winged helix transcription factor FKHR1. *Proc. Natl. Acad. Sci. U. S. A.,* **96**(13), 7421–7426.

[179] Liang, J., Zubovitz, J., Petrocelli, T., Kotchetkov, R., Connor, M. K., Han, K., Lee, J.-H., Ciarallo, S., Catzavelos, C., Beniston, R., Franssen, E., and Slingerland, J. M. (2002) PKB/Akt phosphorylates p27, impairs nuclear import of p27 and opposes p27-mediated G1 arrest. *Nat. Med.,* **8**(10), 1153–1160.

[180] Petersen, B. O., Lukas, J., Sørensen, C. S., Bartek, J., and Helin, K. (1999) Phosphorylation of mammalian CDC6 by Cyclin A/CDK2 regulates its subcellular localization. *EMBO J.,* **18**(2), 396–410.

[181] Gómez, J., García, L. J., Salazar, G. A., Villaveces, J., Gore, S., García, A., Martín, M. J., Launay, G., Alcántara, R., del Toro, N., Dumousseau, M., Orchard, S., Velankar, S., Hermjakob, H., Zong, C., Ping, P., Corpas, M., and Jiménez, R. C. (2013) BioJS: an open source JavaScript framework for biological data visualization. *Bioinformatics,* **29**(8), 1103–1104.

[182] Mukhyala, K. and Masselot, A. (2014) Visualization of protein sequence features using JavaScript and SVG with pViz.js. *Bioinformatics,* **30**(23), 3408–3409.

[183] Kim, Y., Kang, C., Min, B., and Yi, G.-S. (2015) Detection and analysis of disease-associated single nucleotide polymorphism influencing post-translational modification. *BMC Med. Genomics,* **8**(Suppl 2), S7.

[184] 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**(D1), D512–D520.

[185] Reimand, J., Wagih, O., and Bader, G. D. (2013) The mutational landscape of phosphorylation signaling in cancer. *Sci. Rep.,* **3**, 2651.

[186] Gentile, S., Martin, N., Scappini, E., Williams, J., Erxleben, C., and Armstrong, D. L. (2008) The human ERG1 channel polymorphism, K897T, creates a phosphorylation site that inhibits channel activity. *Proc. Natl. Acad. Sci. U. S. A.,* **105**(38), 14704–14708.

[187] Lagarde, W. H., Blackwelder, A. J., Minges, J. T., Hnat, A. T., French, F. S., and Wilson, E. M. (2012) Androgen Receptor Exon 1 Mutation Causes Androgen Insensitivity by Creating Phosphorylation Site and Inhibiting Melanoma Antigen-A11 Activation of NH2- and Carboxyl-terminal Interaction-dependent Transactivation. *J. Biol. Chem.,* **287**(14), 10905–10915.

[188] Ren, J., Jiang, C., Gao, X., Liu, Z., Yuan, Z., Jin, C., Wen, L., Zhang, Z., Xue, Y., and Yao, X. (2010) PhosSNP for Systematic Analysis of Genetic Polymorphisms That Influence Protein Phosphorylation. *Mol. Cel. Proteomics,* **9**(4), 623–634.

[189] Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.,* **29**(1), 308–311.

[190] Kobe, B. and Bodén, M. (2012) Computational Modelling of Linear Motif-Mediated Protein Interactions. *Curr. Top. Med. Chem.,* **12**(14), 1553–1561.

[191] Consortium, T. U. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.,* **43**(D1), D204–D212.

[192] Benzeno, S., Lu, F., Guo, M., Barbash, O., Zhang, F., Herman, J. G., Klein, P. S., Rustgi, A., and Diehl, J. A. (2006) Identification of mutations that disrupt phosphorylation-dependent nuclear export of cyclin D1. *Oncogene,* **25**(47), 6291–6303.

[193] Luna, L., Rolseth, V., Hildrestrand, G. A., Otterlei, M., Dantzer, F., Bjørås, M., and Seeberg, E. (2005) Dynamic relocalization of hOGG1 during the cell cycle is disrupted in cells harbouring the hOGG1-Cys326 polymorphic variant. *Nucleic Acids Res.,* **33**(6), 1813–1824.

[194] Li, X., Dumont, P., Pietra, A. D., Shetler, C., and Murphy, M. E. (2005) The Codon 47 Polymorphism in p53 Is Functionally Significant. *J. Biol. Chem.,* **280**(25), 24245–24251.

[195] Deng, F.-Y., Tan, L.-J., Shen, H., Liu, Y.-J., Liu, Y.-Z., Li, J., Zhu, X.-Z., Chen, X.-D., Tian, Q., Zhao, M., and Deng, H.-W. (2013) SNP rs6265 Regulates Protein Phosphorylation and Osteoblast Differentiation and Influences BMD in Humans. *J. Bone Miner. Res.,* **28**(12), 2498–2507.

[196] Oh, Y.-T., Chun, K., Park, B., Choi, J.-S., and Lee, S. (2007) Regulation of cyclin-dependent kinase inhibitor p21WAF1/CIP1 by protein kinase Cδ-mediated phosphorylation. *Apoptosis,* **12**(7), 1339–1347.

[197] Ristow, M., Müller-Wieland, D., Pfeiffer, A., Krone, W., and Kahn, C. R. (1998) Obesity Associated with a Mutation in a Genetic Regulator of Adipocyte Differentiation. *N. Engl. J. Med.,* **339**(14), 953–959.

[198] Echwald, S. M., Bach, H., Vestergaard, H., Richelsen, B., Kristensen, K., Drivsholm, T., Borch-Johnsen, K., Hansen, T., and Pedersen, O. (2002) A P387L Variant in Protein Tyrosine Phosphatase-1B (PTP-1B) Is Associated With Type 2 Diabetes and Impaired Serine Phosphorylation of PTP-1B In Vitro. *Diabetes,* **51**(1), 1–6.

[199] (2015) An Autism-Linked Mutation Disables Phosphorylation Control of UBE3A. *Cell,* **162**(4), 795 – 807.

[200] Toh, K. L., Jones, C. R., He, Y., Eide, E. J., Hinz, W. A., Virshup, D. M., Ptácek, L. J., and Fu, Y.-H. (2001) An hPer2 Phosphorylation Site Mutation in Familial Advanced Sleep Phase Syndrome. *Science,* **291**(5506), 1040–1043.

[201] Ebert, D. H., Gabel, H. W., Robinson, N. D., Kastan, N. R., Hu, L. S., Cohen, S., Navarro, A. J., Lyst, M. J., Ekiert, R., Bird, A. P., and Greenberg, M. E. (2013) Activity-dependent phosphorylation of MeCP2 threonine 308 regulates interaction with NCoR. *Nature,* **499**(7458), 341–345.

[202] Gelmann, E. P., Steadman, D. J., Ma, J., Ahronovitz, N., Voeller, H. J., Swope, S., Abbaszadegan, M., Brown, K. M., Strand, K., Hayes, R. B., and Stampfer, M. J. (2002) Occurrence of NKX3.1 C154T Polymorphism in Men with and without Prostate Cancer and Studies of Its Effect on Protein Function. *Cancer Res.,* **62**(9), 2654–2659.

[203] Ceholski, D. K., Trieber, C. A., Holmes, C. F. B., and Young, H. S. (2012) Lethal, Hereditary Mutants of Phospholamban Elude Phosphorylation by Protein Kinase A. *J. Biol. Chem.,* **287**(32), 26596–26605.

[204] Gautherot, J., Delautier, D., Maubert, M.-A., Aït-Slimane, T., Bolbach, G., Delaunay, J.-L., Durand-Schneider, A.-M., Firrincieli, D., Barbu, V., Chignard, N., Housset, C., Maurice, M., and Falguières, T. (2014) Phosphorylation of ABCB4 impacts its function: Insights from disease-causing mutations. *Hepatology,* **60**(2), 610–621.

[205] Niceta, M., Stellacci, E., Gripp, K. W., Zampino, G., Kousi, M., Anselmi, M., Traversa, A., Ciolfi, A., Stabley, D., Bruselles, A., Caputo, V., Cecchetti, S., Prudente, S., Fiorenza, M. T., Boitani, C., Philip, N., Niyazov, D., Leoni, C., Nakane, T., Keppler-Noreuil, K., Braddock, S. R., Gillessen-Kaesbach, G., Palleschi, A., Campeau, P. M., Lee, B. H., Pouponnot, C., Stella, L., Bocchinfuso, G., Katsanis, N., Sol-Church, K., and Tartaglia, M. (2015) Mutations Impairing GSK3-Mediated MAF Phosphorylation Cause Cataract, Deafness, Intellectual Disability, Seizures, and a Down Syndrome-like Facies. *Am. J. Hum. Genet.,* **96**(5), 816 – 825.

[206] (2015) A Protein Kinase C Phosphorylation Motif in GLUT1 Affects Glucose Transport and is Mutated in GLUT1 Deficiency Syndrome. *Mol. Cell,* **58**(5), 845 – 853.

[207] Ortiz-Padilla, C., Gallego-Ortega, D., Browne, B. C., Hochgrafe, F., Caldon, C. E., Lyons, R. J., Croucher, D. R., Rickwood, D., Ormandy, C. J., Brummer, T., and Daly, R. J. (2013) Functional characterization of cancer-associated Gab1 mutations. *Oncogene,* **32**(21), 2696–2702.

[208] Dupuis, S., Dargemont, C., Fieschi, C., Thomassin, N., Rosenzweig, S., Harris, J., Holland, S. M., Schreiber, R. D., and Casanova, J.-L. (2001) Impairment of Mycobacterial But Not Viral Immunity by a Germline Human STAT1 Mutation. *Science,* **293**(5528), 300–303.

[209] Chatr-aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., Reguly, T., Nixon, J., Ramage, L., Winter, A., Sellam, A., Chang, C., Hirschman, J., Theesfeld, C., Rust, J., Livstone, M. S., Dolinski, K., and Tyers, M. (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.,* **43**(D1), D470–D478.

[210] van Noort, M., van de Wetering, M., and Clevers, H. (2002) Identification of Two Novel Regulated Serines in the N Terminus of β-Catenin. *Exp. Cell Res.,* **276**(2), 264 – 272.

[211] Sagae, S., Kobayashi, K., Nishioka, Y., Sugimura, M., Ishioka, S., Nagata, M., Terasawa, K., Tokino, T., and Kudo, R. (1999) Mutational analysis of beta-catenin gene in Japanese ovarian carcinomas: frequent mutations in endometrioid carcinomas.. *Jpn. J. Cancer Res.,* **90**(5), 510–515.

[212] Sjöblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S. D., Willis, J., Dawson, D., Willson, J. K. V., Gazdar, A. F., Hartigan,

J., Wu, L., Liu, C., Parmigiani, G., Park, B. H., Bachman, K. E., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., and Velculescu, V. E. (2006) The Consensus Coding Sequences of Human Breast and Colorectal Cancers. *Science,* **314**(5797), 268–274.

[213] Wu, L., Ma, C. A., Zhao, Y., and Jain, A. (2011) Aurora B Interacts with NIR-p53, Leading to p53 Phosphorylation in Its DNA-binding Domain and Subsequent Functional Suppression. *J. Biol. Chem.,* **286**(3), 2236–2244.

[214] Radivojac, P., Baenziger, P. H., Kann, M. G., Mort, M. E., Hahn, M. W., and Mooney, S. D. (2008) Gain and loss of phosphorylation sites in human cancer. *Bioinformatics,* **24**(16), i241–i247.

[215] Flotho, A. and Melchior, F. (2013) Sumoylation: A Regulatory Protein Modification in Health and Disease. *Annu. Rev. Biochem.,* **82**(1), 357–385.

[216] Bernardi, R. and Pandolfi, P. P. (2007) Structure, dynamics and functions of promyelocytic leukaemia nuclear bodies. *Nat. Rev. Mol. Cell Biol.,* **8**(12), 1006–1016.

[217] Zhong, S., Salomoni, P., and Pandolfi, P. P. (2000) The transcriptional role of PML and the nuclear body. *Nat Cell Biol,* **2**(5), E85–E90.

[218] Mao, Y. S., Zhang, B., and Spector, D. L. (2011) Biogenesis and function of nuclear bodies. *Trends Genet.,* **27**(8), 295 – 306.

[219] Zhong, S., Müller, S., Ronchetti, S., Freemont, P. S., Dejean, A., and Pandolfi, P. P. (2000) Role of SUMO-1–modified PML in nuclear body formation. *Blood,* **95**(9), 2748–2752.

[220] Wilkinson, K. A. and Henley, J. M. (2010) Mechanisms, regulation and consequences of protein SUMOylation. *Biochem. J.,* **428**(2), 133–145.

[221] Ayaydin, F. and Dasso, M. (2004) Distinct In Vivo Dynamics of Vertebrate SUMO Paralogues. *Mol. Biol. Cell,* **15**(12), 5208–5218.

[222] Saitoh, H. and Hinchey, J. (2000) Functional Heterogeneity of Small Ubiquitin-related Protein Modifiers SUMO-1 versus SUMO-2/3. *J. Biol. Chem.,* **275**(9), 6252–6258.

[223] Rodriguez, M. S., Dargemont, C., and Hay, R. T. (2001) SUMO-1 Conjugation in Vivo Requires Both a Consensus Modification Motif and Nuclear Targeting. *J. Biol. Chem.,* **276**(16), 12654–12659.

[224] Yang, X.-J. and Grégoire, S. (2006) A Recurrent Phospho-Sumoyl Switch in Transcriptional Repression and Beyond. *Mol. Cell,* **23**(6), 779 – 786.

[225] Xue, Y., Zhou, F., Fu, C., Xu, Y., and Yao, X. (2006) SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res.,* **34**(suppl 2), W254–W257.

[226] Zhao, Q., Xie, Y., Zheng, Y., Jiang, S., Liu, W., Mu, W., Liu, Z., Zhao, Y., Xue, Y., and Ren, J. (2014) GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic Acids Res.,* **42**(W1), W325–W330.

[227] Teng, S., Luo, H., and Wang, L. (2012) Predicting protein sumoylation sites from sequence features. *Amino Acids,* **43**(1), 447–455.

[228] Xu, J., He, Y., Qiang, B., Yuan, J., Peng, X., and Pan, X.-M. (2008) A novel method for high accuracy sumoylation site prediction from protein sequences. *BMC Bioinformatics,* **9**(1), 8.

[229] Chen, Y.-Z., Chen, Z., Gong, Y.-A., and Ying, G. (06, 2012) SUMOhydro: A Novel Method for the Prediction of Sumoylation Sites Based on Hydrophobic Properties. *PLoS ONE,* **7**(6), e39195.

[230] Chen, H., Xue, Y., Huang, N., Yao, X., and Sun, Z. (2006) MeMo: a web tool for prediction of protein methylation modifications. *Nucleic Acids Res.,* **34**(suppl 2), W249–W253.

[231] Shao, J., Xu, D., Tsai, S.-N., Wang, Y., and Ngai, S.-M. (2009) Computational Identification of Protein Methylation Sites through Bi-Profile Bayes Feature Extraction. *PLoS ONE,* **4**(3), e4920.

[232] Steentoft, C., Vakhrushev, S. Y., Joshi, H. J., Kong, Y., Vester-Christensen, M. B., Schjoldager, K. T.-B. G., Lavrsen, K., Dabelsteen, S., Pedersen, N. B., Marcos-Silva, L., Gupta, R., Paul Bennett, E., Mandel, U., Brunak, S., Wandall, H. H., Levery, S. B., and Clausen, H. (2013) Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.,* **32**(10), 1478–1488.

[233] Hou, T., Zheng, G., Zhang, P., Jia, J., Li, J., Xie, L., Wei, C., and Li, Y. (2014) LAceP: Lysine Acetylation Site Prediction Using Logistic Regression Classifiers. *PLoS ONE,* **9**(2), e89575.