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

Ralph Patrick[1], Coralie Horin[2], Bostjan Kobe[1,3,4], Kim-Anh Lê Cao[5], and Mikael Bodén*[1,3]

[1]School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, 4072, Australia

[2]Polytech Nice-Sophia, Université Nice Sophia-Antipolis, Nice, 06103, France

[3]Institute for Molecular Bioscience, The University of Queensland, St Lucia, 4072, Australia

[4]Australian Infectious Diseases Research Centre, The University of Queensland, St Lucia, 4072, Australia

[5]The University of Queensland Diamantina Institute, Translational Research Institute, Woolloongabba, QLD, 4102, Australia

Corresponding author name: Mikael Bodén

Email: m.boden@uq.edu.au

Phone number: +61 7 3365 1307

Running Title: Predicting kinase-specific phosphorylation sites

*to whom correspondence should be addressed

Abbreviations

- **NLS** Nuclear localisation signal
- **CMGC** Cyclin-dependent, mitogen-activated, glycogen synthase and Cdc2-like
- **AGC** Protein kinase A, G and C families
- **TK** Tyrosine kinase
- **CAMK** Ca^{2+}/calmodulin-dependent kinase
- **CK1** Cell kinase 1
- **CPT** Conditional probability table
- **ROC** Receiver operating characteristic
- **AUC** Area under the (receiver operating characteristic) curve
- **AUC50** Area under the curve up to the 50th false positive
- **GO** Gene ontology

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. This consideration enables investigation of how phosphorylation influences a range of biological processes.

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% 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, indicating a greater role for phosphorylation in regulating the nuclear import of proteins than previously thought.

PhosphoPICK is freely available online as a web-service at `http://bioinf.scmb.uq.`

`edu.au/phosphopick`.

Introduction

Kinases regulate a wide variety of essential biological processes through protein phosphorylation, including transcription factor activity (1), the control of DNA damage repair pathways (2), the progression of cells through mitosis (3), and protein import into the nucleus (4). 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 (5), the kinases that regulate these sites in most cases remains unknown. Computational methods that predict kinase-specific phosphorylation sites are therefore an important contributor to understanding the role of phosphorylation events in biological processes (6). Such methods contribute to the guidance of phosphorylation experiments (7) and provide information about the likely signalling pathways that phosphorylation sites may be involved in (8).

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 (9, 10). There are numerous kinase-specific phosphorylation site predictors that take advantage of the sequence specificity of kinases to predict kinase-specific phosphorylation sites (11, 12, 13) as well as phosphorylation sites in a non-kinase specific manner (14, 15).

The presence of valid kinase-binding motifs on a protein is no guarantee that a kinase will phosphorylate a substrate however (16). 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

(17), activating proteins such as cyclins (18), sub-cellular compartmentalisation (19) and the various stages within the mitotic cell cycle (20).

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 (21). 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 nuclear localisation signal (NLS) (22). It has been shown that the binding of importin proteins to their cargo can be promoted or inhibited by the presence of phosphorylation adjacent to the NLS (23). We therefore investigate the relationship between nuclear localisation signals and phosphorylation by cross-referencing experimentally identified phosphorylation sites with predicted NLSs. We use PhosphoPICK to identify the most likely candidate kinases for NLS-adjacent phosphorylation sites, and perform 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 such as 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 regulate distinct biological functions.

Materials and Methods

Data resources

We obtained kinase-specific phosphorylation data for human and mouse from PhosphoSitePlus®, www.phosphosite.org (5) and for yeast (Sacceromyces cerevisiae) from PhosphoGRID (24), which is a database of *in vivo* phosphorylation sites. For data collected from PhosphoSitePlus®, we ensured that phosphorylation sites used were known to occur *in vivo*. 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 data were sourced from BioGRID (25), protein-protein association data from STRING (26), and protein abundance data across the cell cycle from the work by Olsen and colleagues (27). As the cell-cycle information was only available for human, cell-cycle data were not incorporated into the mouse or yeast kinase models. A detailed description of how this data were curated and processed is available in (21).

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 (21). 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.

PhosphoPICK method and workflow

Building on our existing context model, we describe 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. The data used for training the models are available in S1 Data.

Sequence model: We present a Bayesian network model for modelling various sequence features of a kinase binding motif (Fig. 1(a)). 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 (S1 Table, S2 Table & S3 Table). As shown in Fig. 1(a), 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.

It has been shown previously that varying the motif length in predicting kinase binding sites improves prediction accuracy (13). Therefore, for each kinase we tested five different window sizes centred around the phosphorylated residue: 7, 9, 11, 13 and 15. For each kinase

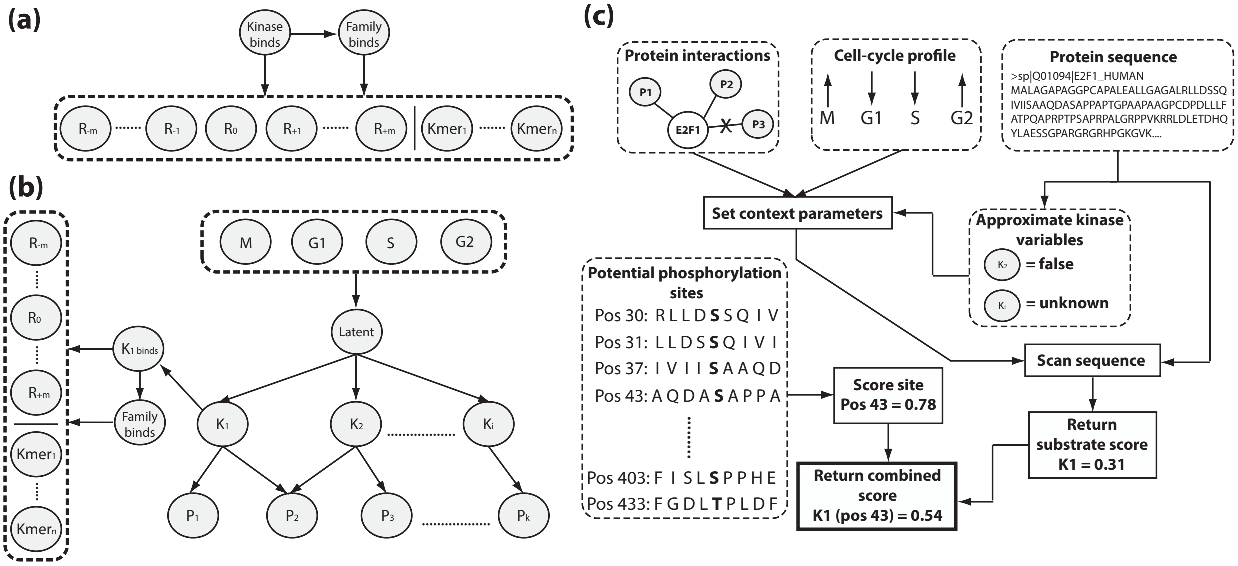


Figure 1: **PhosphoPICK Bayesian networks and workflow.** **(a)** Sequence model. R nodes represent positions in a motif surrounding the phosphorylation site, where R_0 is the potential phosphorylation site. $Kmer_1$ to $Kmer_n$ represent the dimer and trimer configurations incorporated into the model. **(b)** 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. **(c)** 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.

we selected the window size that gave the best prediction accuracy as measured within a cross-validation test (S4 Table, S5 Table & S6 Table).

Combined model: The combined model retains the structure of the "context" Bayesian network described previously (21), 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 (Fig. 1(b)). We created models based on sets of kinases as they are classified into family similarity (28). 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.

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 then trained using the selected training samples, and used to scan over each of the substrates within the test set. The highest score for each of the substrates was 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 is 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.

Prediction workflow: A diagram illustrating the PhosphoPICK workflow for generating a prediction is shown in Fig. 1(c). 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. Non-query kinase nodes are either set to `false` or left unspecified based on the predicted probability that the kinase can bind the substrate sequence.

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.

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 (21), 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 (29) 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`.

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 on 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 (30). 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 (13), NetPhorest 2.0 (31) and NetworKIN 3.0 (31). 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:

$$X = -2(ln(P_{context}) + ln(P_{site})) \tag{1}$$

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.

Evaluation on hold-out set: When evaluating the performance of the model on the hold-out set, the full set of training data were 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:

$$sens. = \frac{TP}{TP+FN} \tag{2}$$

Specificity:

$$spec. = \frac{TN}{TN+FP} \tag{3}$$

Balanced accuracy:

$$BAC = \frac{sens.}{spec.} \tag{4}$$

Matthews' correlation coefficient:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5}$$

Results

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 1 shows the averaged prediction accuracy for each of the kinase families; the full set of values are available in S7 Table for human kinases, S8 Table for mouse kinases, and S9 Table for yeast kinases. 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 (S7 Table); and also the atypical kinases, where all of those kinases obtained an AUC greater than 0.8, and 3/4 greater than 0.85 (S7 Table). The worst performing family appeared to be 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.

Table 1: **Performance comparisons between predicting kinase-specific phosphorylation sites with a baseline model that only considers position-specific amino acid frequencies, and the sequence model.**

	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

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.

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.

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.

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.

Kinase substrate prediction

We compared the ability of the context model to predict kinase substrates against the combined (context plus sequence) model. The results summarised in Table 2 (see S10 Table, S11 Table and S12 Table for the complete set of kinases) 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.

Table 2: **Performance comparisons between predicting kinase substrates with the context Bayesian network model, and with the combined sequence & context model.**

	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

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.

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, and DYRK2 obtained a huge increase from an AUC of 0.63 to 0.91. 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.

In general, the accuracy for mouse kinases was more enhanced by the incorporation of sequence when compared to the accuracy for human kinases. 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. 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. Prediction accuracy for yeast AGC and "other" kinases increased in AUC 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.

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 (Fig. 2).

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 (13), NetPhorest 2.0 and NetworKIN 3.0 (31). We compared the prediction sensitivity of the different methods at the specificity levels described above. Fig. 2 shows the sensitivity difference between PhosphoPICK and the compared methods at two levels of specificity: 99.9% and 99%. The full set of comparisons for individual kinases are available in S13 Table (comparisons at 99.9% specificity) and S14 Table (comparisons at 99% specificity). 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

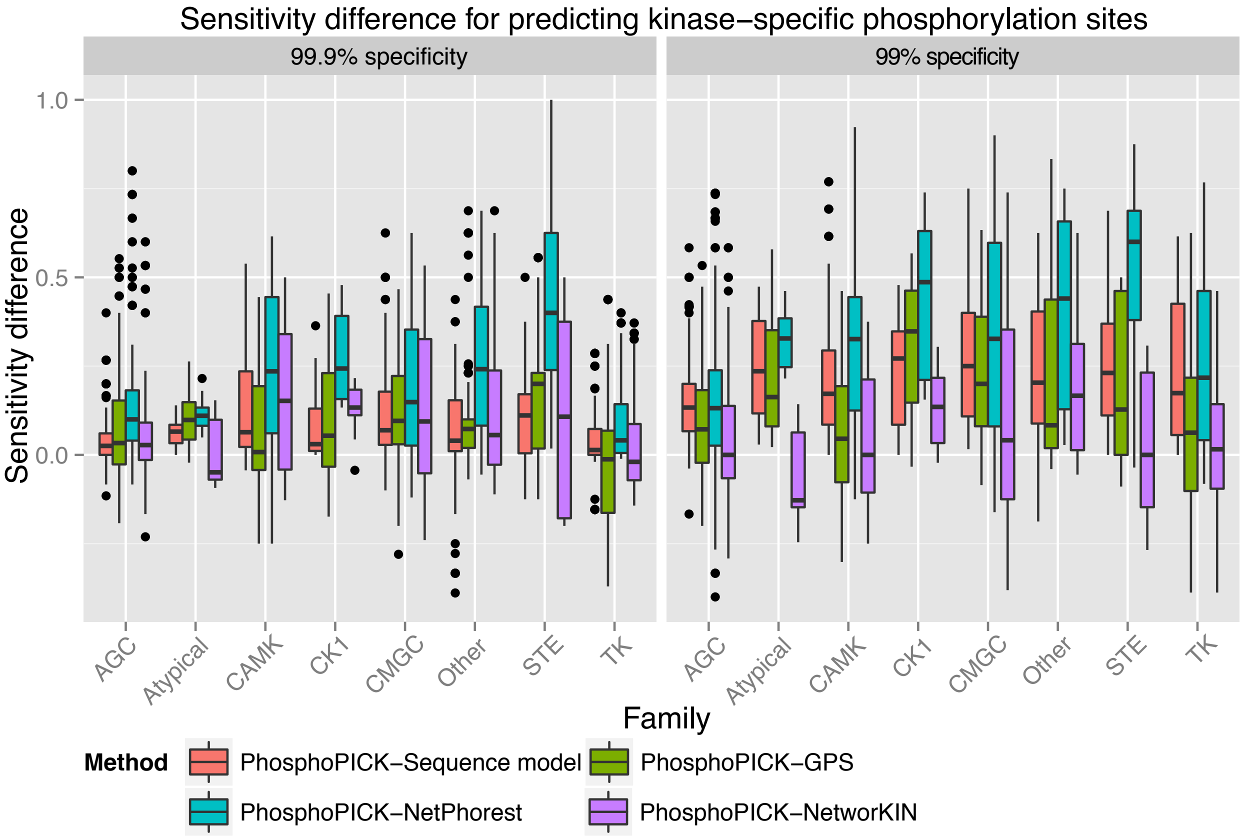


Figure 2: **Sensitivity comparisons for predicting kinase-specific phosphorylation sites out of all potential phosphorylation sites in the protein training set between PhosphoPICK and alternative classification methods.** Sensitivity comparisons for predicting kinase-specific phosphorylation sites out 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.

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. 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.

Evaluation on 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 on our hold-out set of 145 substrates (of the kinases listed in Table 3) by calculating P-values of the predictions and considering predictions that fell below a P-value threshold of 0.005.

Table 3: **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).**

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

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.

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 3). 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.

Multiple kinases regulate nuclear localisation

We predicted NLSs using the NucImport predictor (32), 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 (33). 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 (what can alternatively be designated "P4" (22)), 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).

Fig. 3 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. We found that AurB

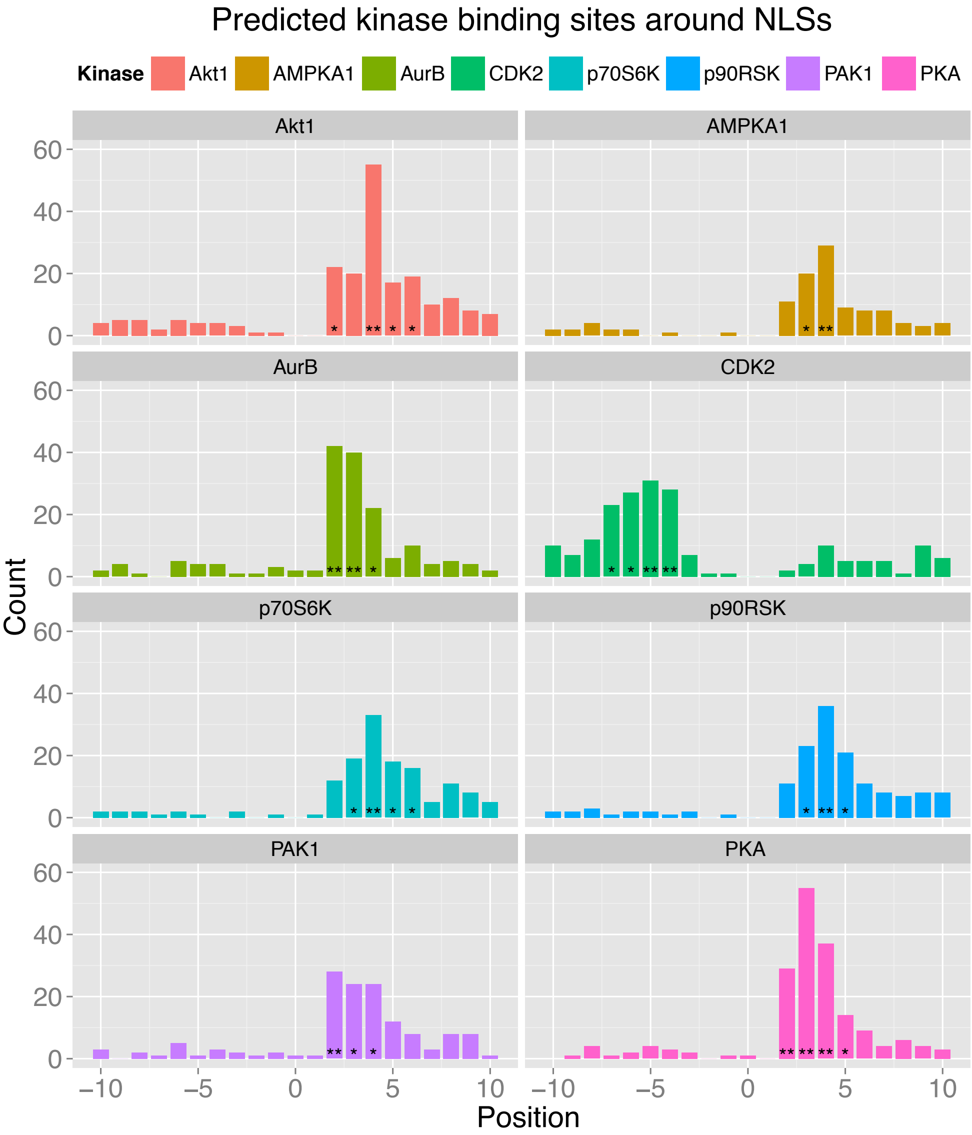


Figure 3: **Distribution of predicted kinase phosphorylation sites surrounding NLSs.** The locations of predicted NLSs were cross-referenced with phosphorylation sites from PhosphoSitePlus® and PhosphoPICK was used to assign kinases to the sites. Count represents the number of times a kinase was predicted 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 $< 1.0E^{-10}$.

substrates were enriched in the GO terms "chromosome", "nucleosome" and "nucleosome assembly" (S15 Table). 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”, “nucleus” and “DNA repair”, none of its significant binding site positions were found to be be associated with enriched GO terms (S16 Table).

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 (S17 Table), Akt1 had 4 (S18 Table), AMPKA1 and p70S6K both had 11 (S19 Table and S20 Table, respectively) and p90RSK had 8 (S21 Table). 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 (S22 Table). 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.

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 (34) based on context information presented to the model (see S1 Text 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 (35). In addition to protein expression data, the FANTOM consortium has profiled vast cell-type specific gene expression atlases (36). 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 (4, 37, 38). 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" (23) (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 (4, 37, 38, 39).

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 (40). Adenomatus polyposis coli (APC) is another example of a protein where nuclear import is regulated by phosphorylation (41). 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 reductions 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 (42), and it has also been shown to phosphorylate residues within the vicinity of NLSs (43). The Akt kinase has been shown to be a regulator of nuclear localisation (44), and phosphorylation by Akt is able to impair the nuclear import of p27 *in vitro* (45). Similarly, CDK2 is known to be a regulator of nuclear localisation (46). 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.

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 (47) package pViz (48). 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 S2 Text.

S1 Table

Sequence model accuracy across human kinases when different percentages of kinase-substrate phosphorylation peptides were used to determine k-mers added to the 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
	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
	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

AGC	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

	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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

	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
	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
	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
	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
	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
	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
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

S2 Table

Sequence model accuracy across mouse kinases when different percentages of kinase-substrate phosphorylation peptides were used to determine k-mers added to the 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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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
	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

S3 Table

Sequence model accuracy across yeast kinases when different percentages of kinase-substrate phosphorylation peptides were used to determine k-mers added to the 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
	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
	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
	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
	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
	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
	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
	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
	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
	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

S4 Table

Sequence model accuracy for varying window sizes in human kinases. 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. Kinases in the table are grouped according to family.

		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**

AGC	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**
	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
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
	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**
	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**
	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**
	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
	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
	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**
	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**

S5 Table

Sequence model accuracy for varying window sizes in mouse kinases. 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. Kinases in the table are grouped according to family.

		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
	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**
	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
	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
	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
	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**
	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**
	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**
AGC	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
	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**

S6 Table

Sequence model accuracy for varying window sizes in yeast kinases. 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. Kinases in the table are grouped according to family.

		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**
	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
	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**
	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
	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
	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
	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**
	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
	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
	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**
	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**
	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**
	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**
	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**
	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

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

S7 Table

Comparison of prediction accuracy across human kinases between sequence model and baseline. 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
	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
	Src	0.53±0.004	0.56±0.006	0.07±0.004	0.10±0.005
	Abl	0.58±0.011	0.62±0.009	0.11±0.007	0.15±0.016
	Fyn	0.54±0.011	0.59±0.009	0.10±0.008	0.12±0.009
	Lck	0.54±0.014	0.53±0.012	0.05±0.009	0.06±0.016
	Lyn	0.50±0.017	0.48±0.016	0.08±0.011	0.05±0.012
	EGFR	0.54±0.015	0.56±0.023	0.06±0.005	0.05±0.018
	Syk	0.78±0.018	0.81±0.018	0.27±0.020	0.27±0.025
	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
	FAK	0.44±0.025	0.67±0.050	0.09±0.030	0.36±0.067
	Ret	0.43±0.026	0.54±0.023	0.17±0.027	0.19±0.025
	Arg	0.66±0.039	0.67±0.036	0.15±0.022	0.15±0.017
	Brk	0.56±0.016	0.60±0.021	0.15±0.026	0.20±0.007
	ALK	0.49±0.021	0.57±0.032	0.04±0.020	0.00±0.000
	Btk	0.60±0.036	0.71±0.033	0.14±0.044	0.31±0.053
	PDGFRB	0.59±0.017	0.61±0.033	0.09±0.043	0.25±0.040
	JAK3	0.63±0.040	0.81±0.032	0.19±0.053	0.40±0.063
	Hck	0.51±0.026	0.58±0.025	0.08±0.022	0.09±0.017
	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
	Chk1	0.69±0.022	0.71±0.017	0.07±0.017	0.06±0.022
	AMPKA1	0.75±0.019	0.72±0.016	0.10±0.014	0.08±0.014
	MAPKAPK2	0.79±0.016	0.78±0.019	0.08±0.020	0.14±0.028
	PKD1	0.75±0.015	0.76±0.010	0.08±0.014	0.09±0.012
	LKB1	0.77±0.013	0.81±0.009	0.47±0.003	0.58±0.018
	MSK1	0.76±0.044	0.86±0.032	0.10±0.049	0.33±0.076
	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
	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

Other	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
	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
	ATR	0.86±0.008	0.86±0.009	0.14±0.009	0.11±0.014
	DNAPK	0.83±0.008	0.86±0.005	0.13±0.005	0.17±0.012
	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

S8 Table

Comparison of prediction accuracy across mouse kinases between sequence model and baseline. 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
	ERK1	0.78±0.011	0.82±0.010	0.19±0.018	0.16±0.021
	CDK5	0.73±0.013	0.80±0.013	0.09±0.015	0.17±0.016
	CDK1	0.76±0.022	0.79±0.013	0.17±0.018	0.18±0.030
	JNK1	0.74±0.019	0.78±0.014	0.13±0.018	0.22±0.040
	P38A	0.67±0.021	0.74±0.017	0.10±0.022	0.23±0.028
	CDK2	0.76±0.020	0.74±0.034	0.10±0.020	0.34±0.033
	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
	PKCA	0.67±0.014	0.72±0.010	0.15±0.011	0.25±0.016
	Akt1	0.82±0.015	0.81±0.011	0.34±0.049	0.38±0.047
	PKCD	0.71±0.014	0.75±0.028	0.13±0.024	0.11±0.037
	p90RSK	0.90±0.015	0.87±0.013	0.31±0.048	0.22±0.037
	RSK2	0.80±0.056	0.79±0.042	0.29±0.087	0.28±0.085
	PKG1	0.70±0.023	0.66±0.042	0.12±0.049	0.00±0.000
	p70S6K	0.82±0.032	0.88±0.029	0.18±0.062	0.39±0.062

	PKCZ	0.61±0.050	0.69±0.095	0.00±0.000	0.29±0.114
	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
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

S9 Table

Comparison of prediction accuracy across yeast kinases between sequence model and baseline. 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
	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

S10 Table

Combined model accuracy across human kinases compared to the context only model. 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
CMGC	CDK2	0.69±0.003	0.76±0.002	0.097±0.0016	0.110±0.0024
	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
	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

AGC	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
	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
	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

S11 Table

Combined model accuracy across mouse kinases compared to the context only model. 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

S12 Table

Combined model accuracy across yeast kinases compared to the context only model. 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

S13 Table

Sensitivity differences for kinases at 99.9% specificity. 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
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

CMGC	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
	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
	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

CAMK	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
	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
	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

Continued on next page

Continued from previous page

	Kinase	Sequence model	GPS	NetPhorest	NetworKIN
STE	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

S14 Table

Sensitivity differences for kinases at 99% specificity. 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
	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

AGC	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
	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
	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
	CK1D	0.2920±0.0315	0.5220±0.0343	0.4950±0.0343	0.1700±0.0343
	CK1E	0.3040±0.0802	0.2610±0.0550	0.6520±0.0550	0.2170±0.0550
	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
	ATR	0.3690±0.0499	0.2480±0.0405	0.3300±0.0405	-0.1790±0.0405
	DNAPK	0.1560±0.0343	0.0659±0.0269	0.3960±0.0269	0.0769±0.0269
	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

S15 Table

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
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

S16 Table

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

S17 Table

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

S18 Table

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

S19 Table

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

S20 Table

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
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

S21 Table

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
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

S22 Table

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. Table shows GO terms identified at each position in the 20-residue window surrounding the NLS.

position	GO term	Description	E-value
-10	GO:0005694	chromosome	0.0017
-10	GO:0000786	nucleosome	0.024
-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
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

S1 Text

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.

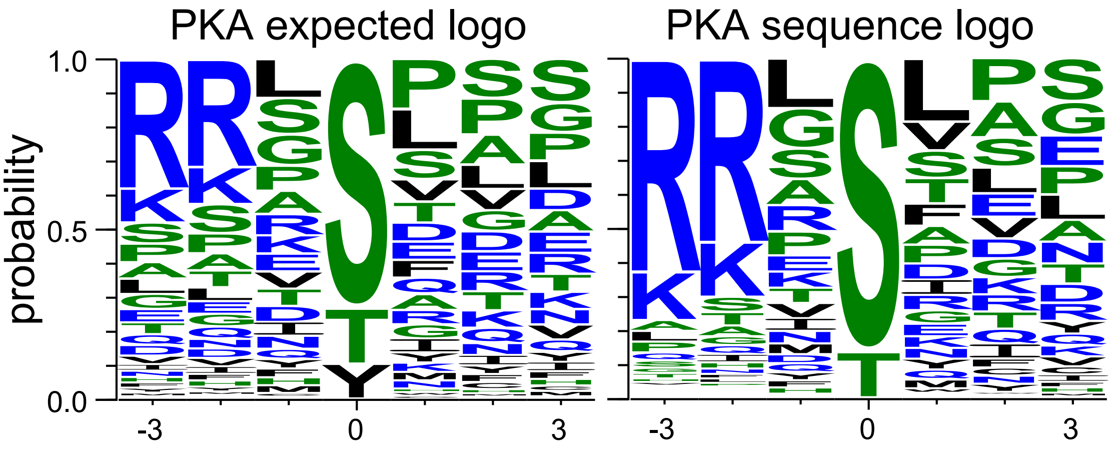
For each protein, we first set the context parameters in the Bayesian network: the protein interaction nodes, cell-cycle nodes and kinase nodes (except the kinase being queried). We used the most probable explanation (MPE) form of inference to determine the most likely value for the query kinase phosphorylating 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.

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 (34) to generate sequence logos from the count matrix.

S1 Fig 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.

S1 Fig

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 (34).

S2 Text

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.

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 (47) package pViz (48), 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.

S1 Data

Training data and model specification files for training the models presented in the paper.

References

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

[2] 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.

[3] 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.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Ellis, J. J. and Kobe, B. (2011) Predicting protein kinase specificity: Predikin update and performance in the DREAM4 challenge. *PLoS ONE,* **6**(7), e21169.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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**.

[25] Chatr-aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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).

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

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

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

[39] 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.

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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