Supplementary Information

# Uncovering phosphorylation–based specificities through functional interaction networks

Omar Wagih[1], Naoyuki Sugiyama[2], Yasushi Ishihama[2] and Pedro Beltrao[1]

[1] *European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD*
[2] *Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshidashimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan*

Correspondence should be addressed to pbeltrao@ebi.ac.uk

# 1 Supplementary Results

## Attempting to improve AUC of predicted models

In in the current implementation we exclude all phosphosites with proline at +1 when predicting the specificity for kinases that are not proline directed. This requires prior knowledge regarding the different kinase families, which might not always be available for different PTM types or species. Instead of having to specify the P+1 kinases, we can use all phosphosites as a background for motif enrichment to decrease the importance of P+1 motifs. However, this approach results in a moderate decrease in the mean AUC to 0.61 **(Supplementary Figure 11)** and a smaller number of predictions (32 vs. 85). When using all phosphosites as a background, the importance of prolines and arginines at certain positions is decreased and in most cases are not enriched for, likely resulting in incorrect enrichments and ultimately models **(Supplementary Figure 12)**.

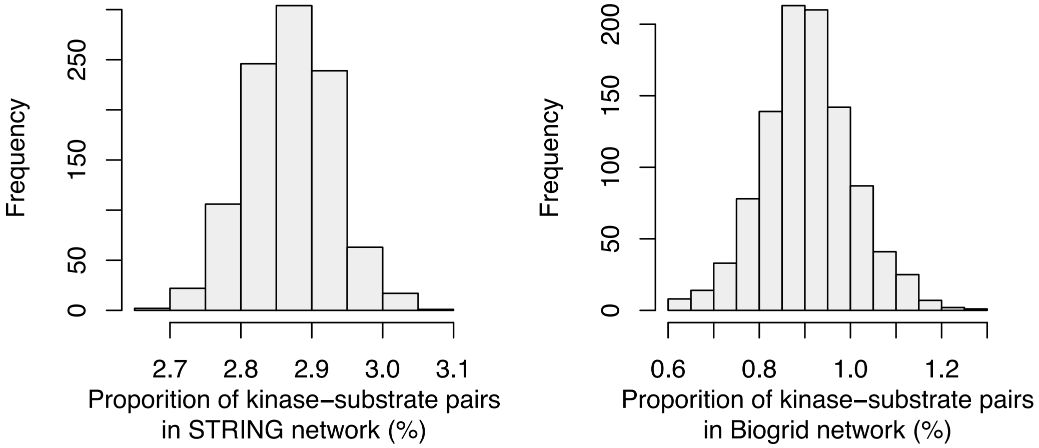
We also tested if improving the phosphorylation site quality could improve results. We filtered the phosphorylation data using two different criteria. First, we selected phosphosites that were annotated to at least two pubmed articles. Second, since MS methods are biased towards highly abundant proteins, we removed phosphosites that occurred in the top 10% abundant proteins as defined in PaxDB[1]. For each criterion, we generated models and measured the performance. Overall, filtering did not appear to improve the performance of the models **(Supplementary Figure 11)**.

## 2 Supplementary Figures

### Figure 1

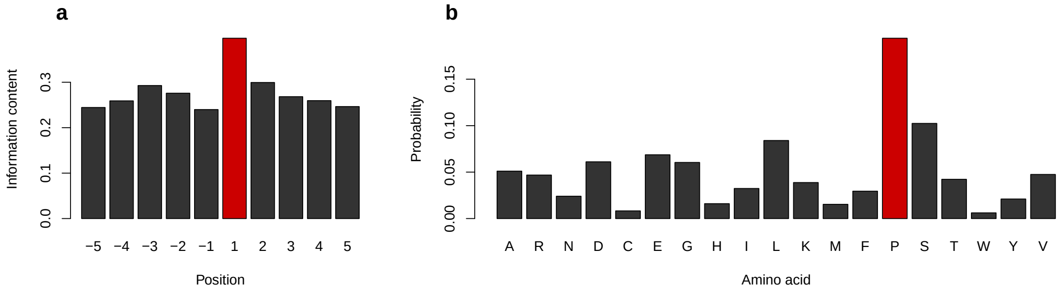
Proportion of kinase-substrate pairs in STRING: 5.5% ($p$-value$<$0.001)

Proportion of kinase-substrate pairs in BioGRID[2]: 7.61% ($p$-value$<$0.001)



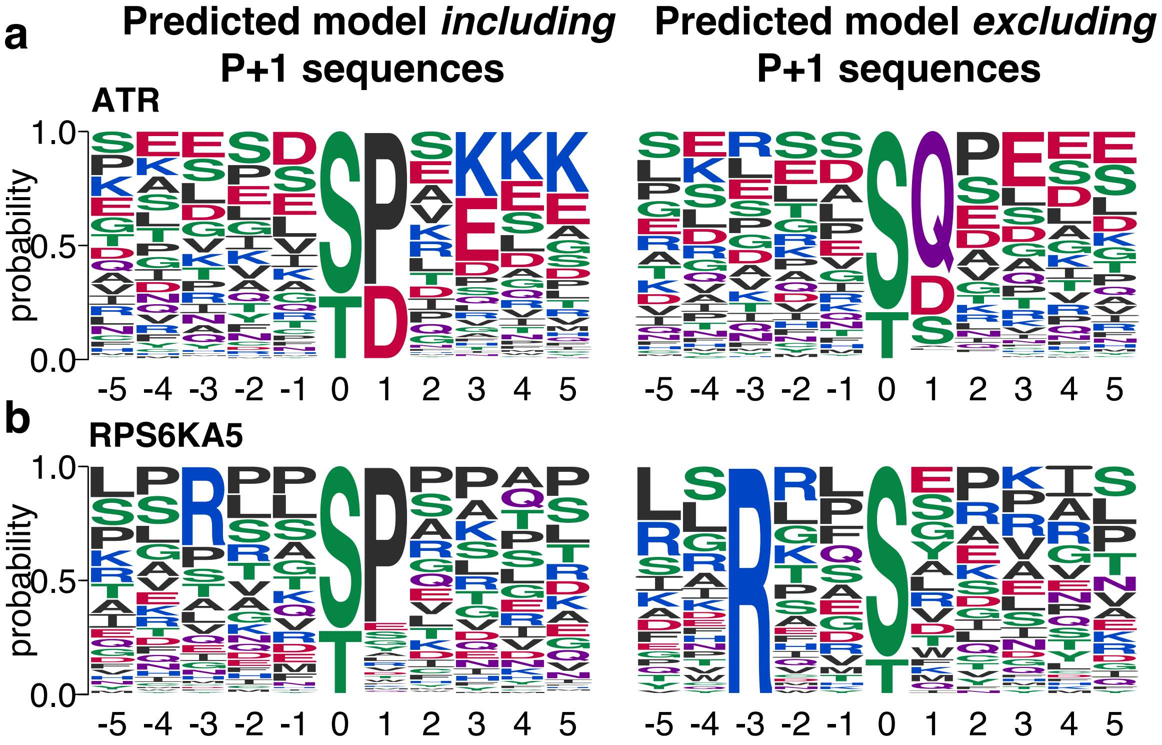
**Enrichment of kinase–substrate pairs in protein interaction networks**. Histogram of In the STRING interaction network for the human kinases 5.5% of the interactions correspond to known kinase–substrate interactions. Similarly, Biogrid contains 7.61% known kinase-substrate interactions. The histograms show the proportion of 1,000 random kinase–substrate pairs in STRING and BioGRID.

# Figure 2



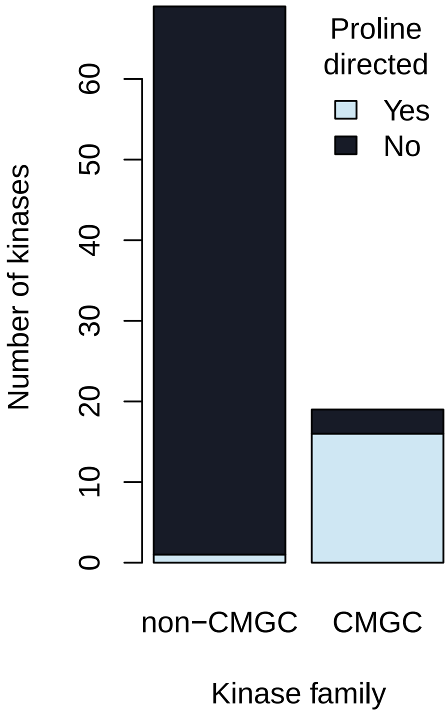
**Proline bias across all phosphosites**. Bar plots showing (a) Information content for each position flanking the central residue for all known phosphosites. The highest information content position is shown in magenta. (b) Amino acid frequencies for each position of all phosphosites. Proline is highlighted in red.

## Figure 3



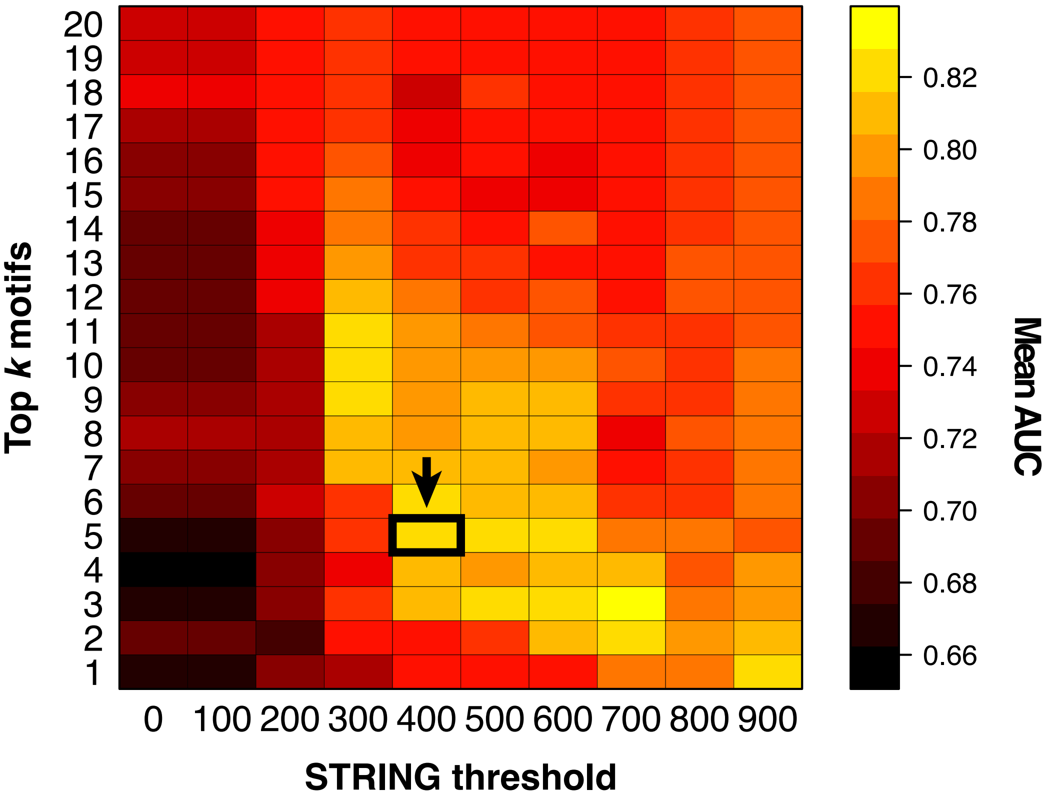
**Effect of removing P+1 phosphosites for kinases**. (a-b) Two examples of non-proline directed kinases, ATR and RPS6KA5. Each example shows the predicted specificity before (left) and after (right) removal of P+1 phosphosites. Consistent enrichment of proline is observed for these cases if P+1 phosphosites are not removed, masking the true specificity of the kinase.

## Figure 4



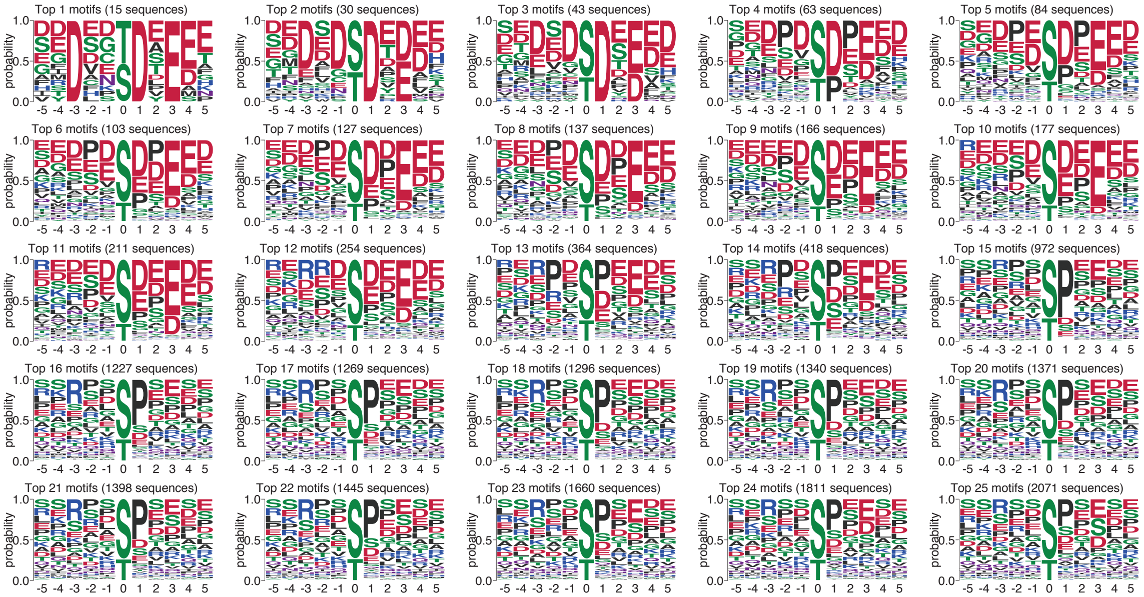
**Proline-directed CMGC kinases**. Bar plot showing CMGC vs. non-CMGC kinases with at least 20 substrates and the proportions of each class, which are proline directed. Kinases were considered proline directed if a predominance of Proline was observed at position +1.

## Figure 5



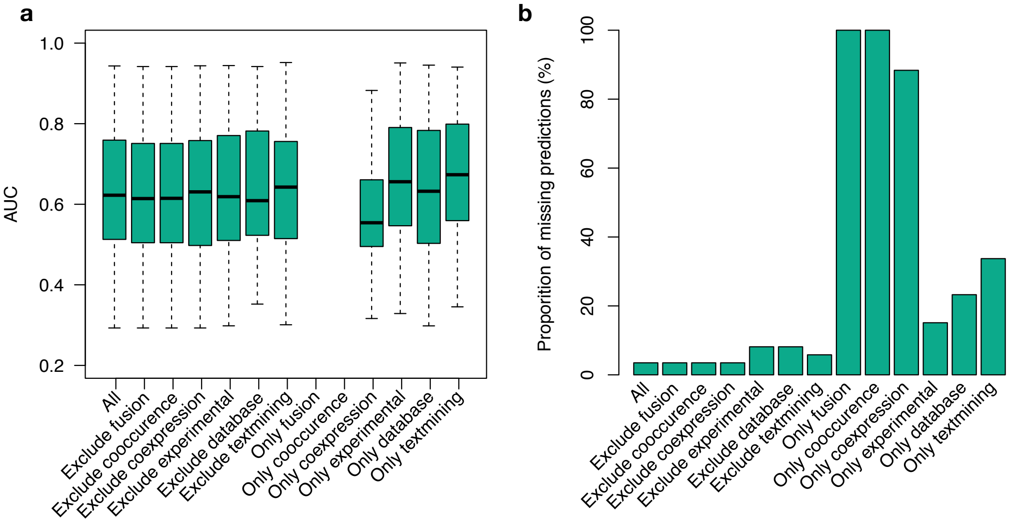
**Benchmarking of the method on nine kinases with well-known specificities**. Optimal parameters of the method were computed by varying the STRING score threshold and the top k significant motifs used in constructing the model. The performance of kinases is measured in each case. The arrow shows the data point corresponding to the selected thresholds.

## Figure 6



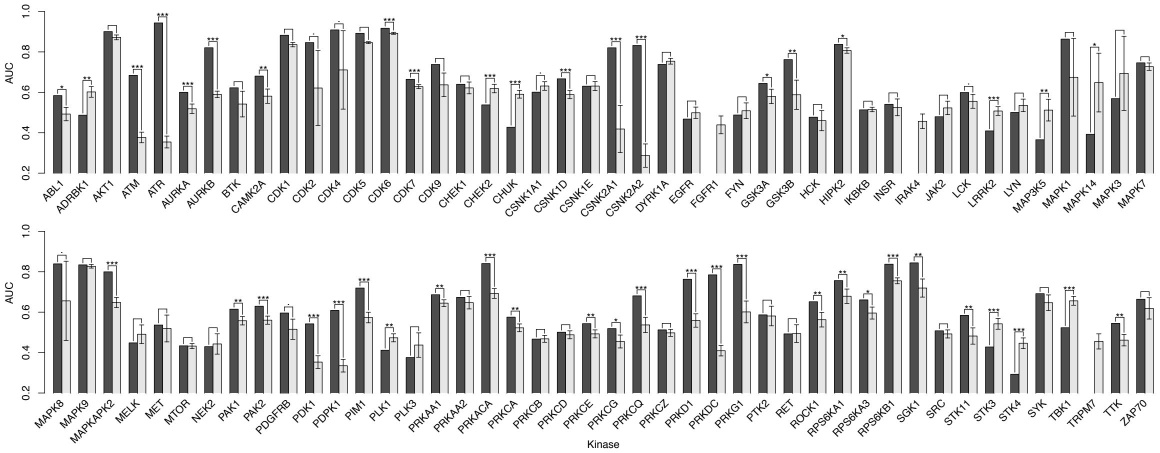
**Masking predicted specificity by over-selecting motifs**. Specificity predictions for CSNK2A1 (CK2) resulting from different top k significant motifs. Over selecting motifs can result in less specific predictions and sometimes the inclusion of other contaminant motifs.

## Figure 7



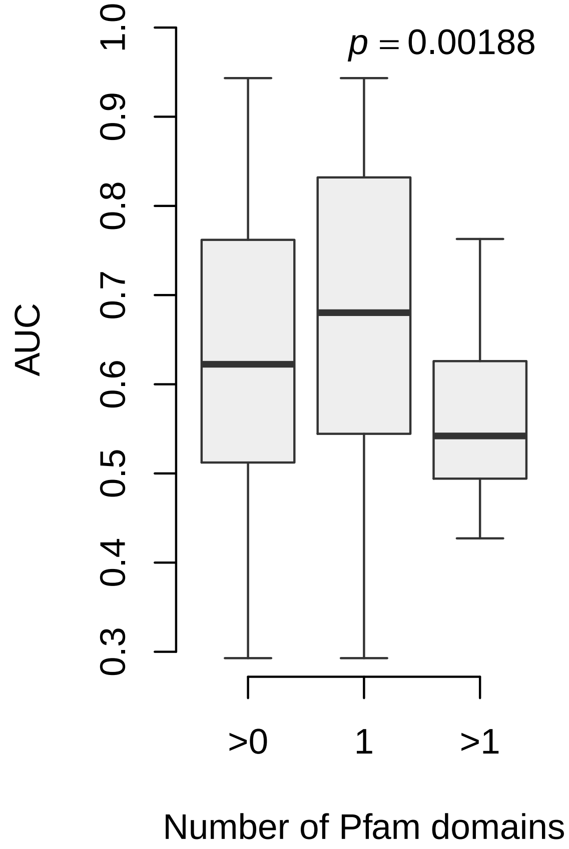
**Distribution of AUCs using different STRING evidences** (a) Distribution of AUCs for predicted models of all kinases with $\geq$ 20 known substrates by either excluding a particular string evidence, or using only that evidence to generate the prediction. (b) Bar plot showing the proportion of kinases with no prediction resulting from lack of interactions when restricting evidences.

## Figure 8



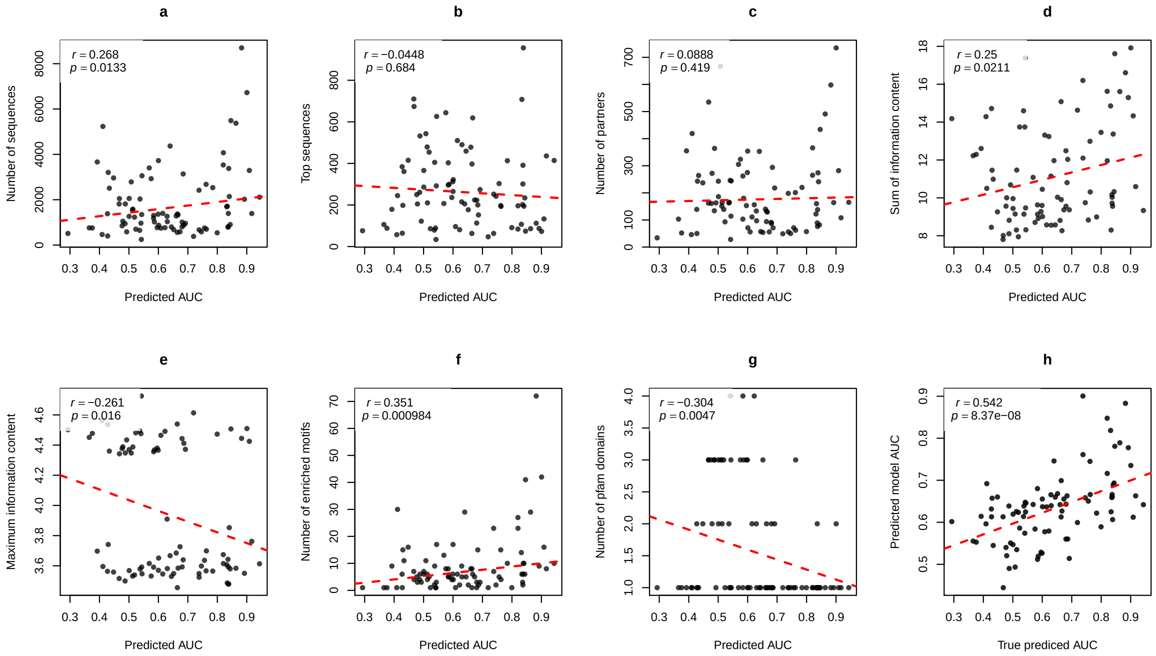
**Performance of predicted vs. random models**. Bar plots showing AUCs of predicted models of 85 kinases with ≥20 known targets vs. that of random models (.p<0.01, *p<0.05, **p<0.01, *** p<0.001, *z*-test). Error bars represent the median absolute deviation of 1000 random models.

**Figure 9**



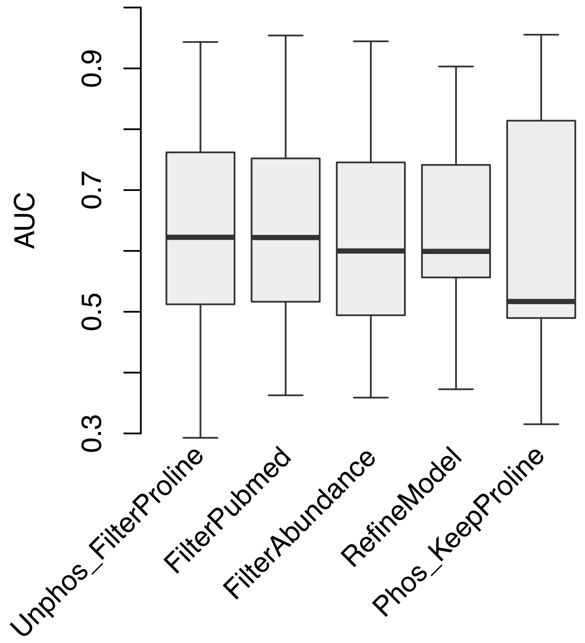
**Impact of the number of domains on predictions**. Bar plot showing the distribution of AUCs for kinases, depending on the number of Pfam[3] domains they contain. Kinases with more than one domain, overall, have significantly lower AUCs.

# Figure 10



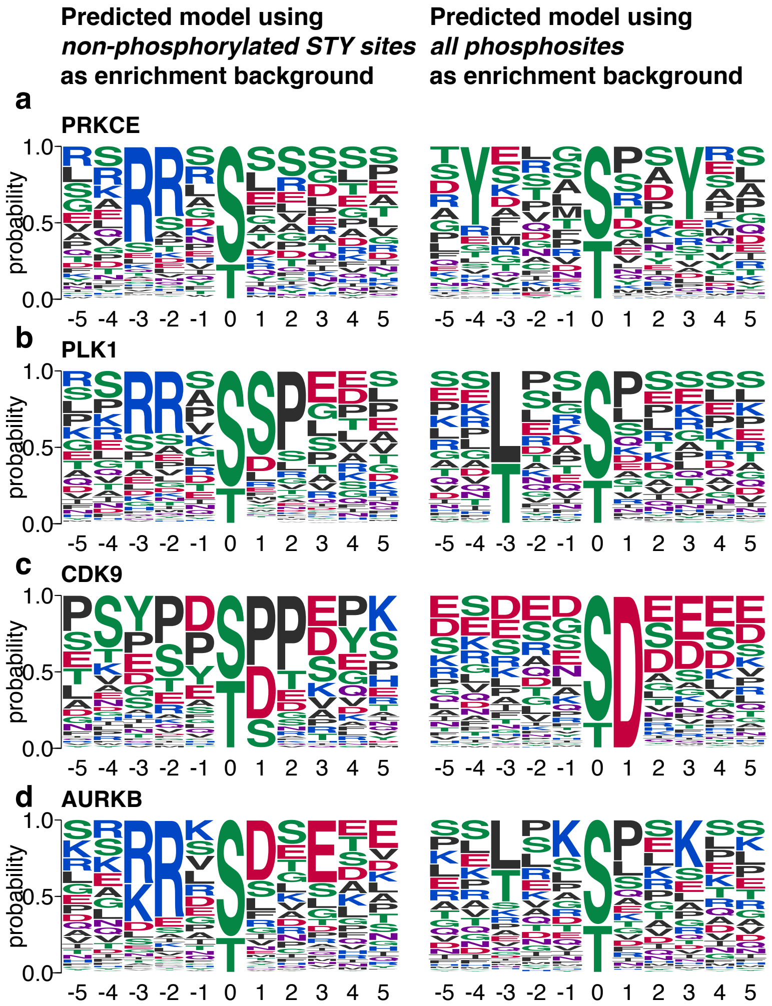
**Feature correlations**. AUCs of predicted models correlated with (a) number of phosphosite sequences on functional partners, (b) number of phosphosite sequences matching the top 5 enriched motifs, (c) number of functional partners, (d) sum of information content across positions of models (e) maximum information content amongst different positions, (f) total number of enriched motifs and (g) number of annotated Pfam domains. (h) A linear regression model built using a combination of features (a,d,e,f) is used to predict the AUC of predicted specificity models, which are correlated against the true AUC. Line of best fit is shown in red.

## Figure 11



**Variations of motif enrichment background sets**. Bar plot showing the distribution of AUCs by (1) using unphosphorylated STY sites as background for motif enrichment, while filtering P+1 phosphosites (our method) versus (2) Filtering phosphosites having less than two associated pubmed IDs (3) Filtering phosphosites occurring in highly abundant proteins, obtained from PaxDB[1] (4) Refining function partner phosphosites using the method described in Reimand *et al.*[4]. (5) Using all phosphorylation sites as a background while retaining P+1 phosphosites in non-proline directed kinases.

## Figure 12

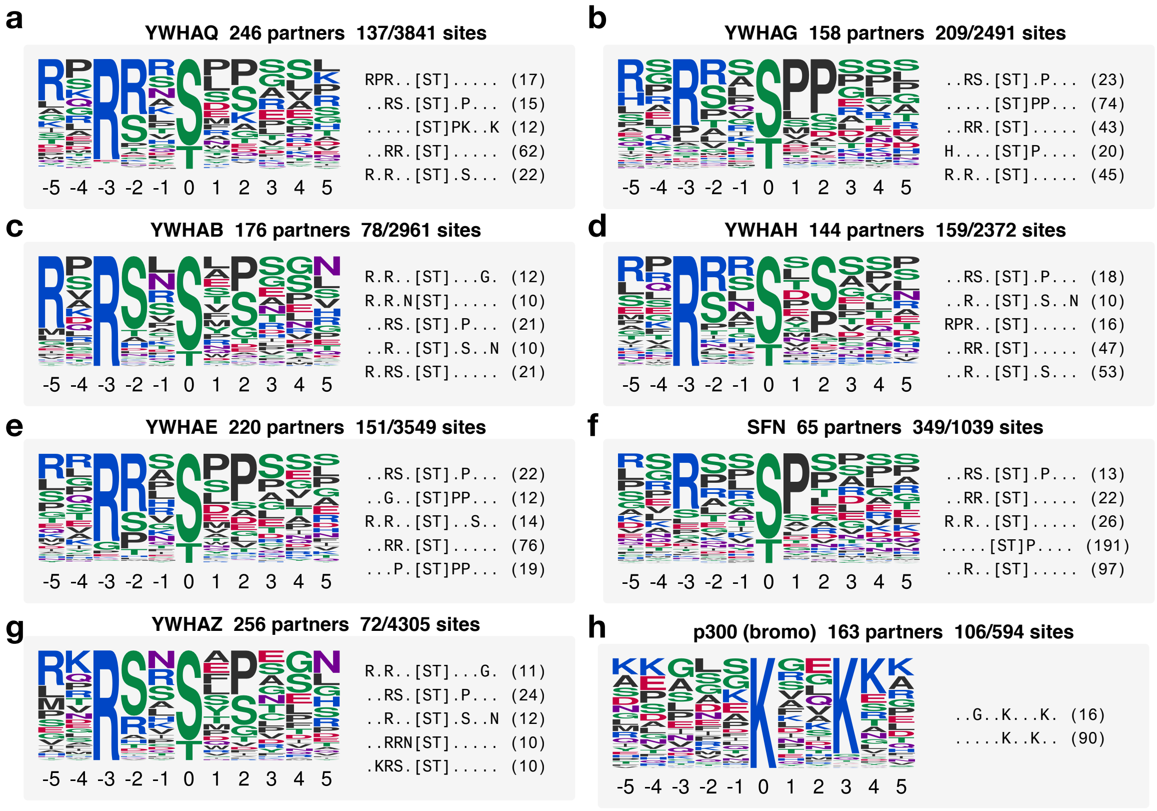


Predicted model using *non-phosphorylated STY sites* as enrichment background

Predicted model using *all phosphosites* as enrichment background

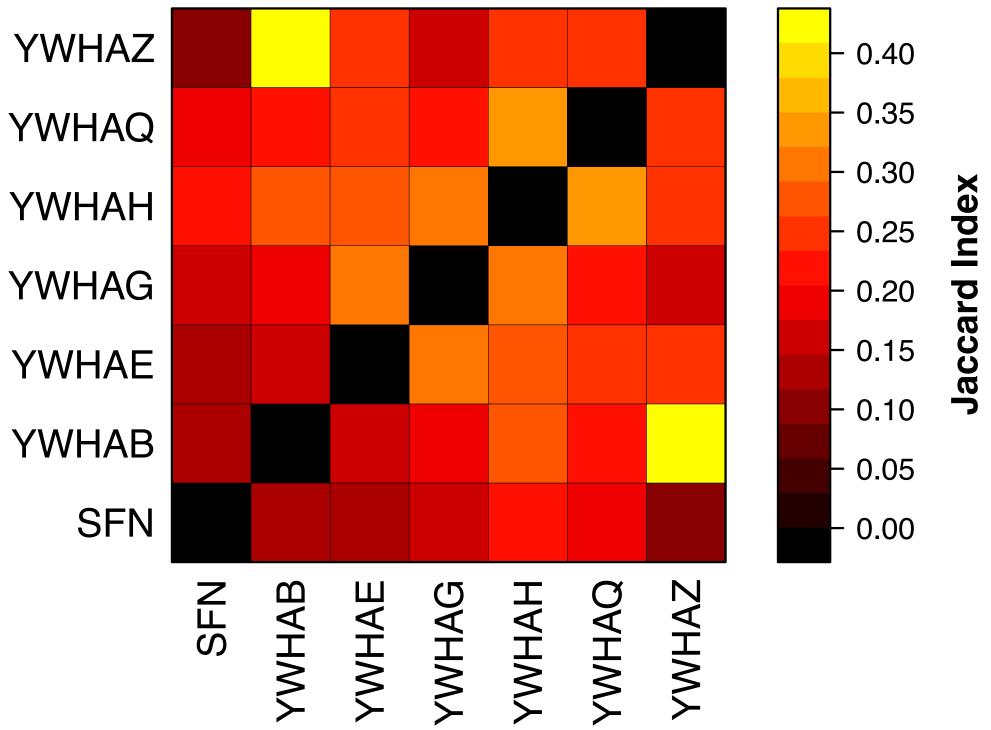
**Using phosphorylated sequences as background for motif enrichment, while retaining P+1 sequences for non proline-directed kinases**. (a-d) Predicted models for non proline-directed kinases, using top 5 significant motifs. The left predicted model is with unphosphorylated sequences as background for motif enrichment, while filtering out P+1 sequences. The right predicted model is using all phosphorylation sites as background for motif enrichment, while retaining P+1 sequences.

# Figure 13



**Predictions of specificities for other PTM types**. Each panel shows the logo representing the predicted specificity (left) and the top five extracted motifs and the number of sites matching them (right). (a-g) Shows predictions for 14-3-3 proteins and (h) shows the prediction for bromodomain-containing protein p300.

## Figure 14



**Overlap between 14-3-3 predictions**. Heatmap showing the Jaccard overlap between sequences used in constructing the different 14-3-3 models. In most cases, there is little overlap, despite the fact that same motifs are recovered.

## References

1. Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schrimpf, S. P., Hengartner, M. O., and von Mering, C. *Mol. Cell Proteomics* **11**(8), 492–500 Aug (2012).
2. Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. *Nucleic Acids Res.* **34**(Database issue), D535–9 Jan (2006).
3. Sonnhammer, E. L., Eddy, S. R., and Durbin, R. *Proteins* **28**(3), 405–20 Jul (1997).
4. Reimand, J., Wagih, O., and Bader, G. D. *Sci Rep* **3**, 2651 (2013).