Science Signaling
AAAS

www.sciencesignaling.org/cgi/content/full/1/35/ra2/DC1

Supplementary Materials for

Linear Motif Atlas for Phosphorylation-Dependent Signaling

Martin Lee Miller, Lars Juhl Jensen, Francesca Diella, Claus Jørgensen, Michele Tinti, Lei Li, Marilyn Hsiung, Sirlester A. Parker, Jennifer Bordeaux, Thomas Sicheritz-Ponten, Marina Olhovsky, Adrian Pasculescu, Jes Alexander, Stefan Knapp, Nikolaj Blom, Peer Bork, Shawn Li, Gianni Cesareni, Tony Pawson, Benjamin E. Turk, Michael B. Yaffe,* Søren Brunak,* Rune Linding*

*To whom correspondence should be addressed. E-mail: brunak@cbs.dtu.dk (S.B.), myaffe@mit.edu (M.B.Y.), and rune.linding@gmail.com (R.L.)

Published 2 September 2008, *Sci. Signal.* 1, ra2 (2008)
DOI: 10.1126/scisignal.1159433

This PDF file includes:

Supplementary Information for:
Linear motif atlas for phosphorylation-dependent signaling

Martin Lee Miller[1,2*], Lars Juhl Jensen[2,3*], Francesca Diella[3], Claus Jørgensen[4], Michele Tinti[5], Lei Li[6], Marilyn Hsiung[4], Sirlester A. Parker[7], Jennifer Bordeaux[7], Thomas Sicheritz-Ponten[1], Marina Olhovsky[4], Adrian Pasculescu[4], Jes Alexander[8], Stefan Knapp[9], Nikolaj Blom[1], Peer Bork[2,10], Shawn Li[6], Gianni Cesareni[5], Tony Pawson[4], Benjamin E. Turk[7], Michael B. Yaffe[8†], Søren Brunak[1,2†] and Rune Linding[4,8,11†]

[1] Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark.

[2] The Novo Nordisk Foundation Centre for Protein Research, University of Copenhagen, Copenhagen, Denmark

[3] European Molecular Biology Laboratory, Heidelberg, Germany

[4] Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada

[5] University of Rome, Tor Vergata, Rome, Italy

[6] University of Western Ontario, London, Ontario, Canada

[7] Department of Pharmacology, Yale University School of Medicine, New Haven, USA

[8] Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, USA

[9] Structural Genomics Consortium, University of Oxford, UK

[10] Max-Delbrück-Centre for Molecular Medicine, Berlin, Germany

[11] Cellular & Molecular Logic Team, The Institute of Cancer Research, London, UK

[*] These authors contributed equally to this work

[†] To whom correspondence should be addressed; E-mail: brunak@cbs.dtu.dk, myaffe@mit.edu and rune.linding@gmail.com

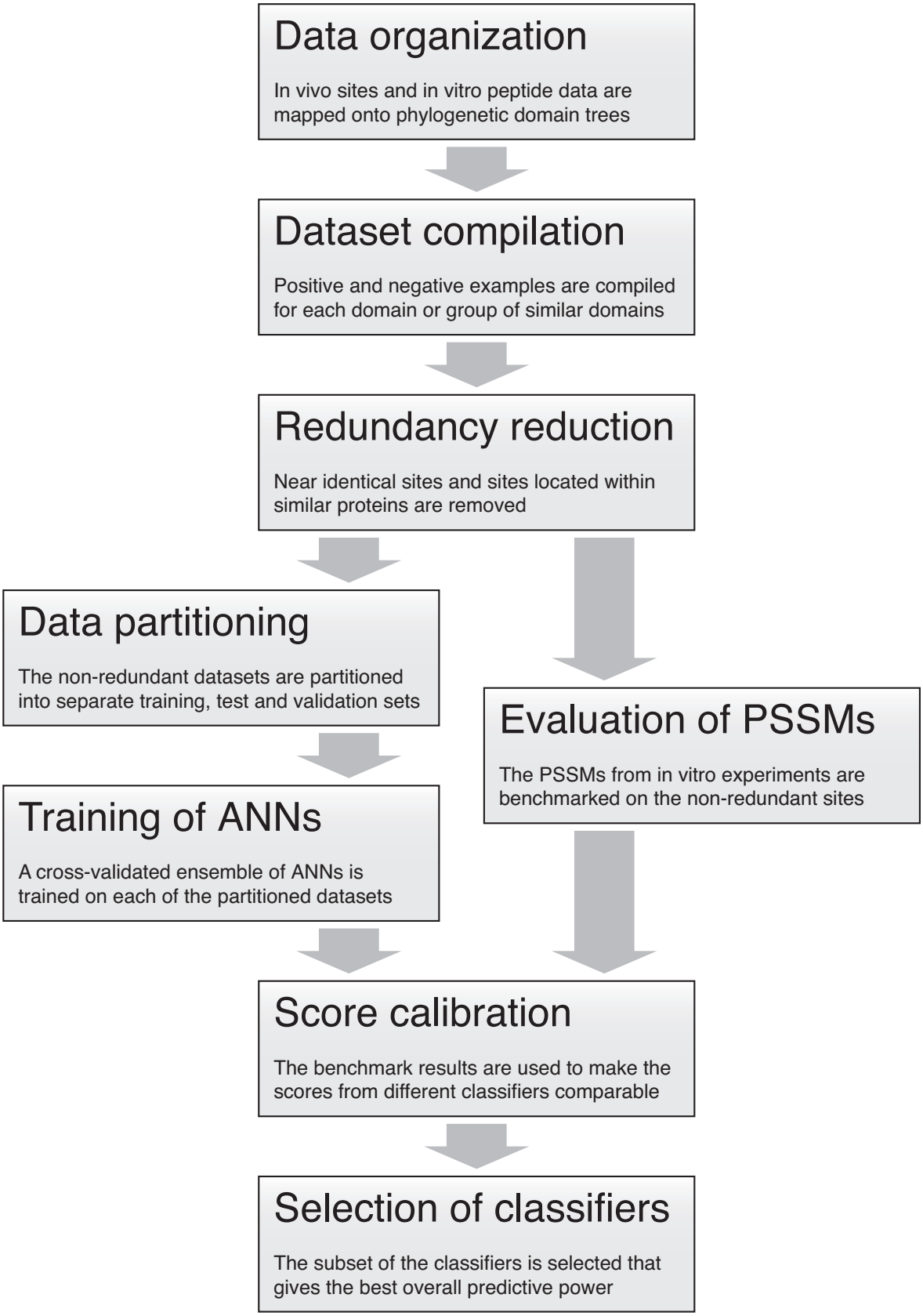


Figure S1: **Overview of the NetPhorest pipeline.** The pipeline consists of a combination of Python, Perl and C-code that takes as input a set of manually organized data sets and phylogenetic domain trees. Most of the steps in the flow chart are shown in more detail in Figures 1 and 2.

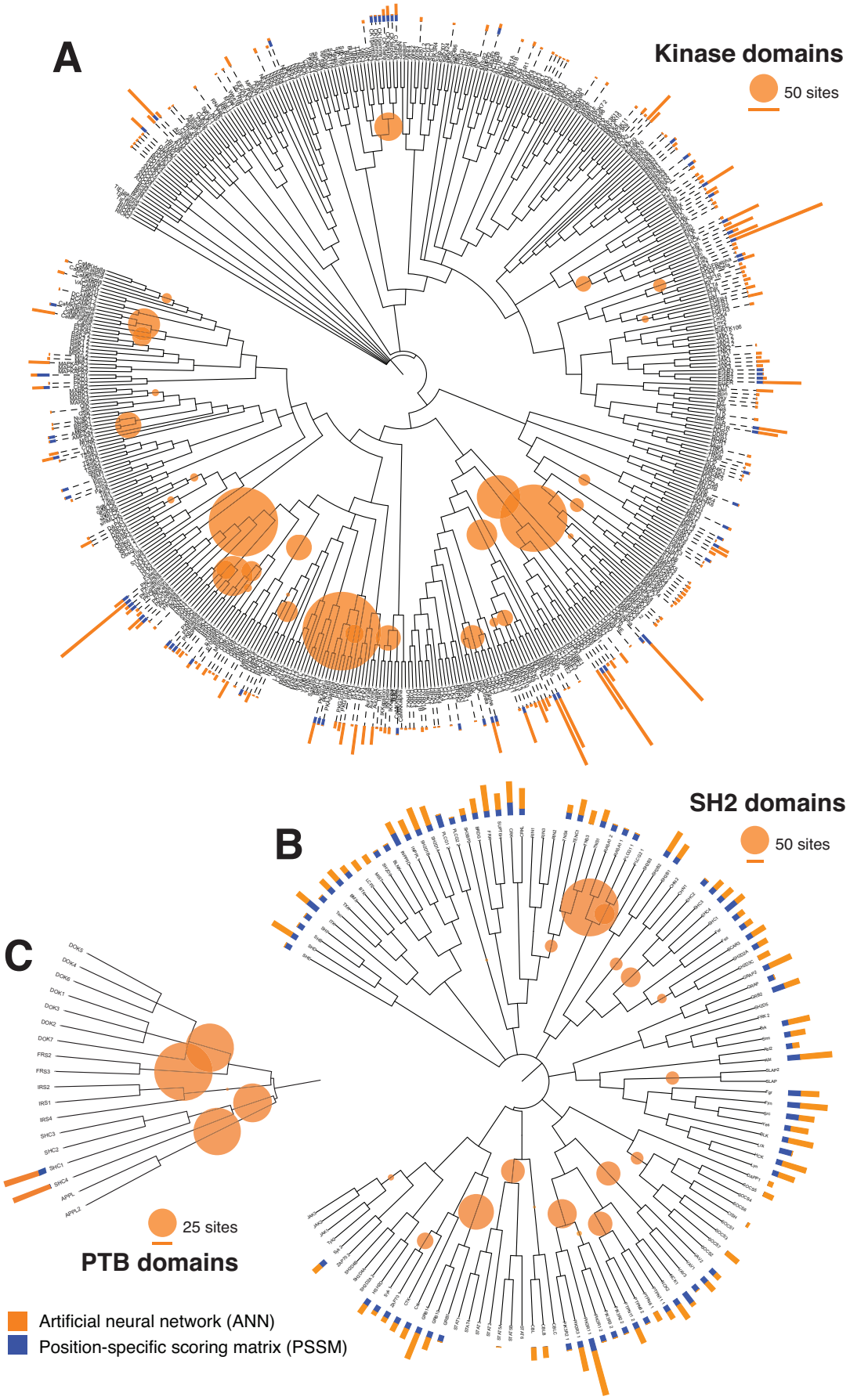


Figure S2: **Phosphorylation data mapped onto domain trees.** The number of phosphorylation sites annotated to be substrates of individual (orange bars) or groups (orange pies, centered on the group in question) of kinases (A), SH2 domains (B) and PTB domains (C) are shown on the respective phylogenetic domain trees. This data was used to develop artificial neural networks (ANNs). The blue bars show NetPhorest collection of position-specific scoring matrices (PSSMs) based on *in vitro* assays. NetPhorest also contains ANNs for the WW domain of PIN1 and PSSMs for 14-3-3 and BRCT domains, all of which bind phosphorylated serines and threonines (not shown).

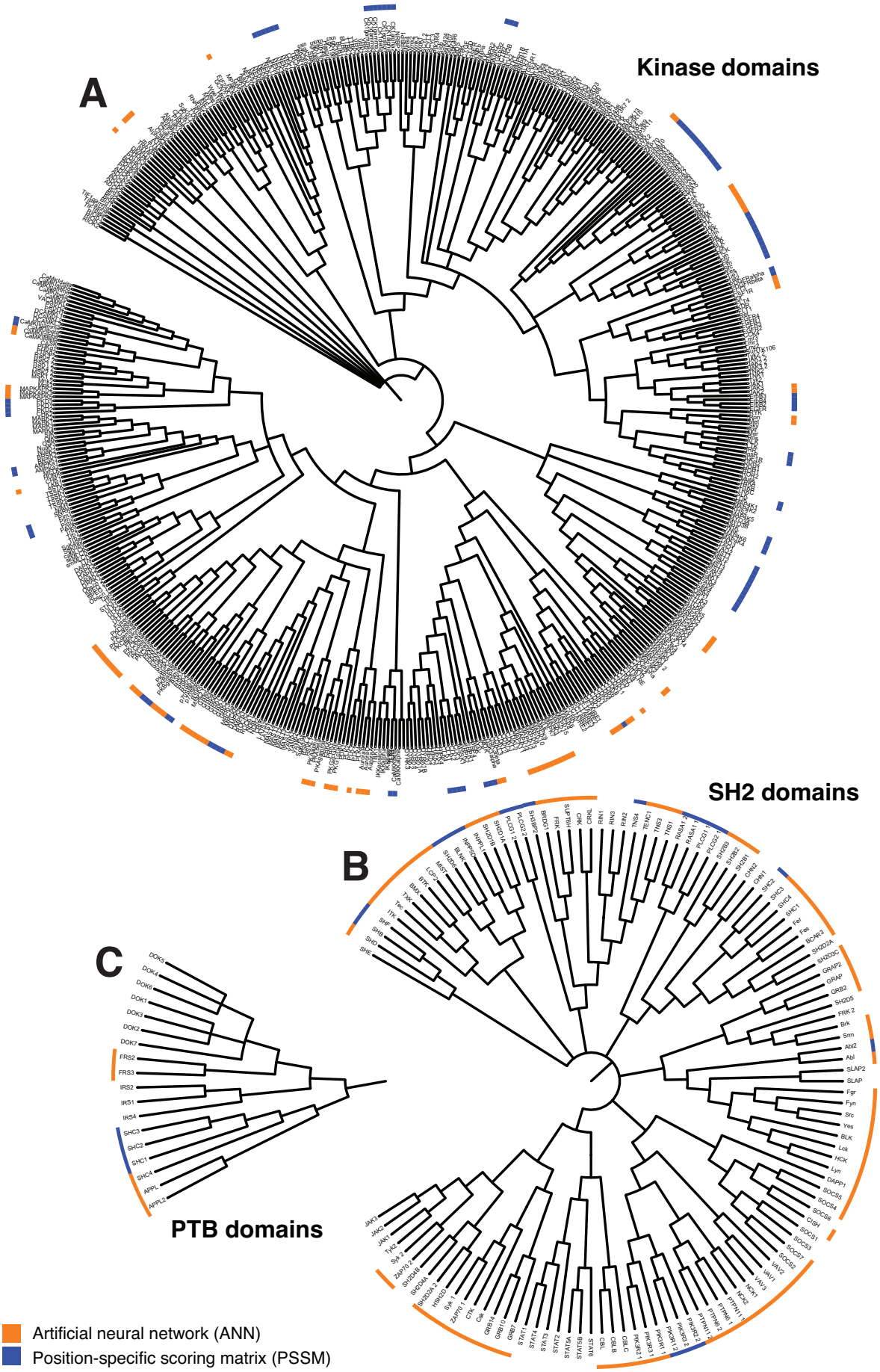


Figure S3: **Coverage of classifiers for targets of kinases, SH2 domains and PTB domains.** Bars show the kinases, SH2 domains and PTB domains that are covered by artificial neural networks (orange) and position-specific scoring matrices (blue). Based on current data, the NetPhorest pipeline yields a non-redundant collection of 125 sequence-based classifiers that cover 179 of 518 kinase domains, 93 of 118 SH2 domains and 8 of 18 phosphotyrosine-binding PTB domains (the figure is best viewed by zooming in a PDF viewer). NetPhorest also contains ANNs for the WW domain of PIN1 and PSSMs for 14-3-3 and BRCT domains, all of which bind phosphorylated serines and threonines (not shown). The trees for kinase and SH2 domains were obtained from Manning et al. [1] and Liu et al. [2], respectively. It should be noted that some of the kinase domains, for example TRRAP, do not appear to have catalytic activity. None of these are covered by the sequence-based classifiers.

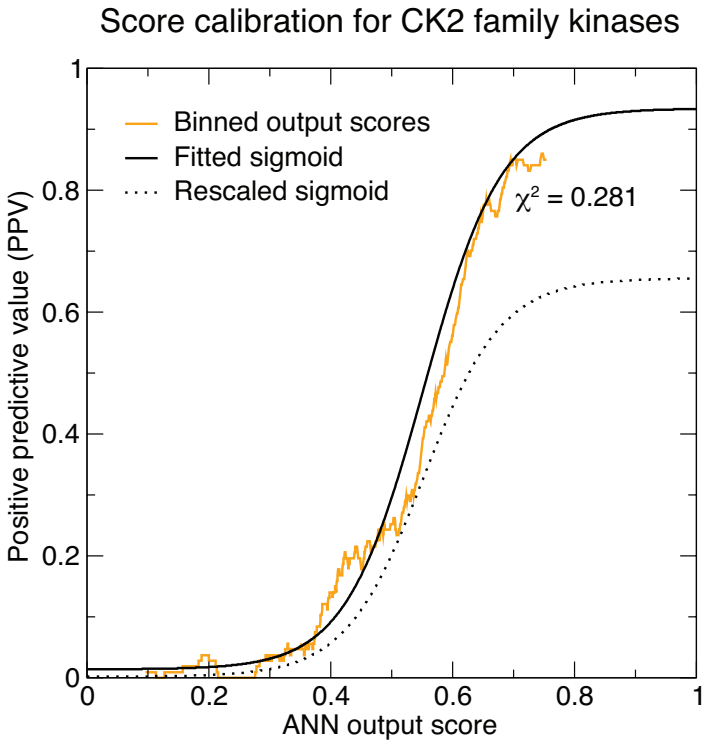


Figure S4: **Score calibration.** To make the output scores from different sequence-based classifiers directly comparable, we calibrated the scores through benchmarking on our compilation of phosphorylation sites. As an example, we show here the score calibration for the CK2 group of kinases, which includes the CK2α1 and CK2α2 isoforms. The fraction of correct predictions (positive predictive value, PPV) is calculated within different score windows (running bins) on the validation set (orange curve). Subsequently, a sigmoid function is fitted to these values, minimizing the sum of squared errors. The resulting calibration curves enables us to estimate the posterior probability that a site is phosphorylated by a particular kinase or group of kinases. However, because the fractions of positive examples within the redundancy-reduced data sets do not reflect the corresponding prior probabilities, the calibration curves has to be rescaled (dotted curve, see Methods for details).

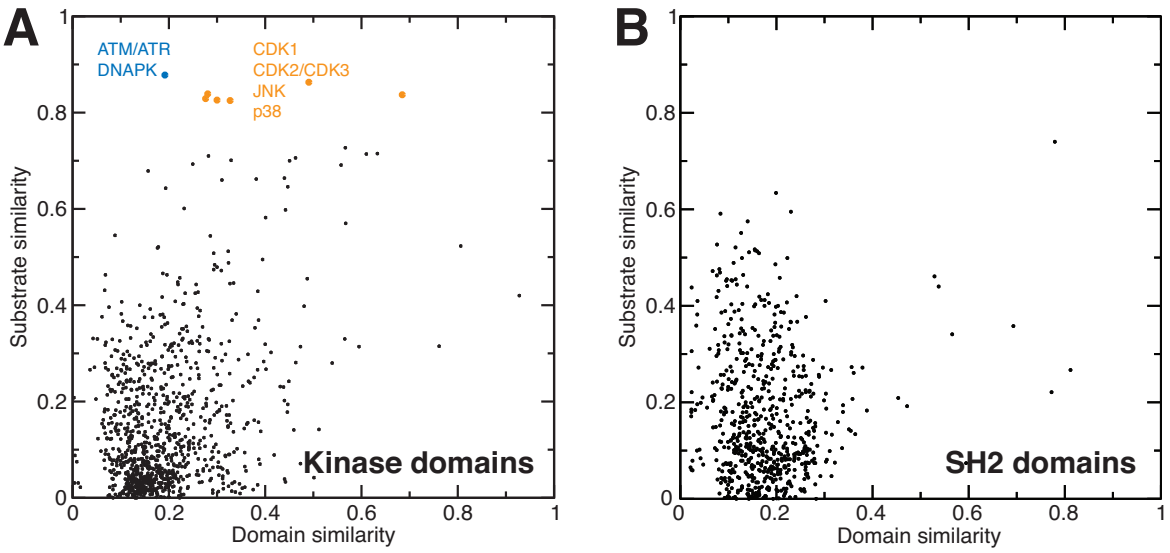


Figure S5: **Correlation between domain similarity and substrate specificity.** Calculating the similarity between two sequence motifs is non-trivial, in particular if the motifs are described by different types of sequence models (ANNs and PSSMs). To assess the similarity between motifs, we created 100,000 random phosphopeptides and scored each of them using all classifiers in NetPhorest. We thereby obtained a set of 100,000 dimensional score vectors, one for each sequence model, and defined the similarity between two motifs as the cosine similarity between their score vectors. The corresponding domain similarities were calculated based on pairwise sequence alignments (self-normalized bit scores, see Methods for details) between all domains in the kinase and SH2 tree. For internal nodes this was calculated by averaging over the pairwise sequence-similarity of each member in the group across all other nodes. Substrate similarity was plotted against domain similarity for kinases (A) and SH2 domains (B). Note that, although significant, the correlation between domain and substrate similarity is far from perfect (R^2 = 0.18, $P < 10^{-12}$ and R^2 = 0.31, $P < 10^{-6}$ for kinases and SH2 domains, respectively; only domain pairs with a self-normalized bitscore of 0.3 or higher are considered). Two groups of kinase domains that have diverged greatly in sequence but have retained their substrate specificity are highlighted in orange and blue, respectively.

A

Y-A-Z-X-X-X-X-S/T-X-X-X-X-A-G-K-K(biotin)	–5 set
Y-A-X-Z-X-X-X-S/T-X-X-X-X-A-G-K-K(biotin)	–4 set
Y-A-X-X-Z-X-X-S/T-X-X-X-X-A-G-K-K(biotin)	–3 set
Y-A-X-X-X-Z-X-S/T-X-X-X-X-A-G-K-K(biotin)	–2 set
Y-A-X-X-X-X-Z-S/T-X-X-X-X-A-G-K-K(biotin)	–1 set
Y-A-X-X-X-X-X-S/T-Z-X-X-X-A-G-K-K(biotin)	+1 set
Y-A-X-X-X-X-X-S/T-X-Z-X-X-A-G-K-K(biotin)	+2 set
Y-A-X-X-X-X-X-S/T-X-X-Z-X-A-G-K-K(biotin)	+3 set
Y-A-X-X-X-X-X-S/T-X-X-X-Z-A-G-K-K(biotin)	+4 set

B

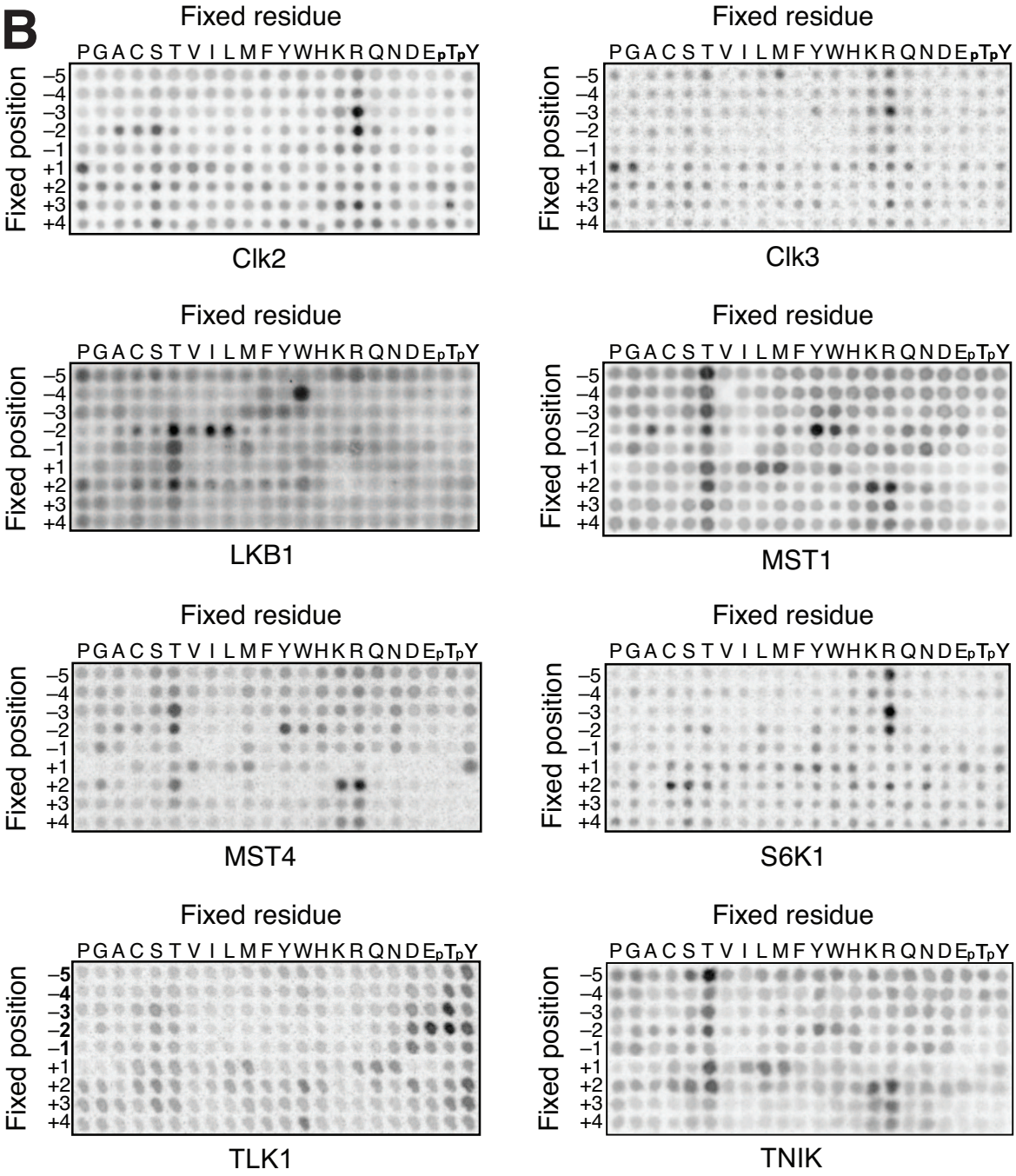
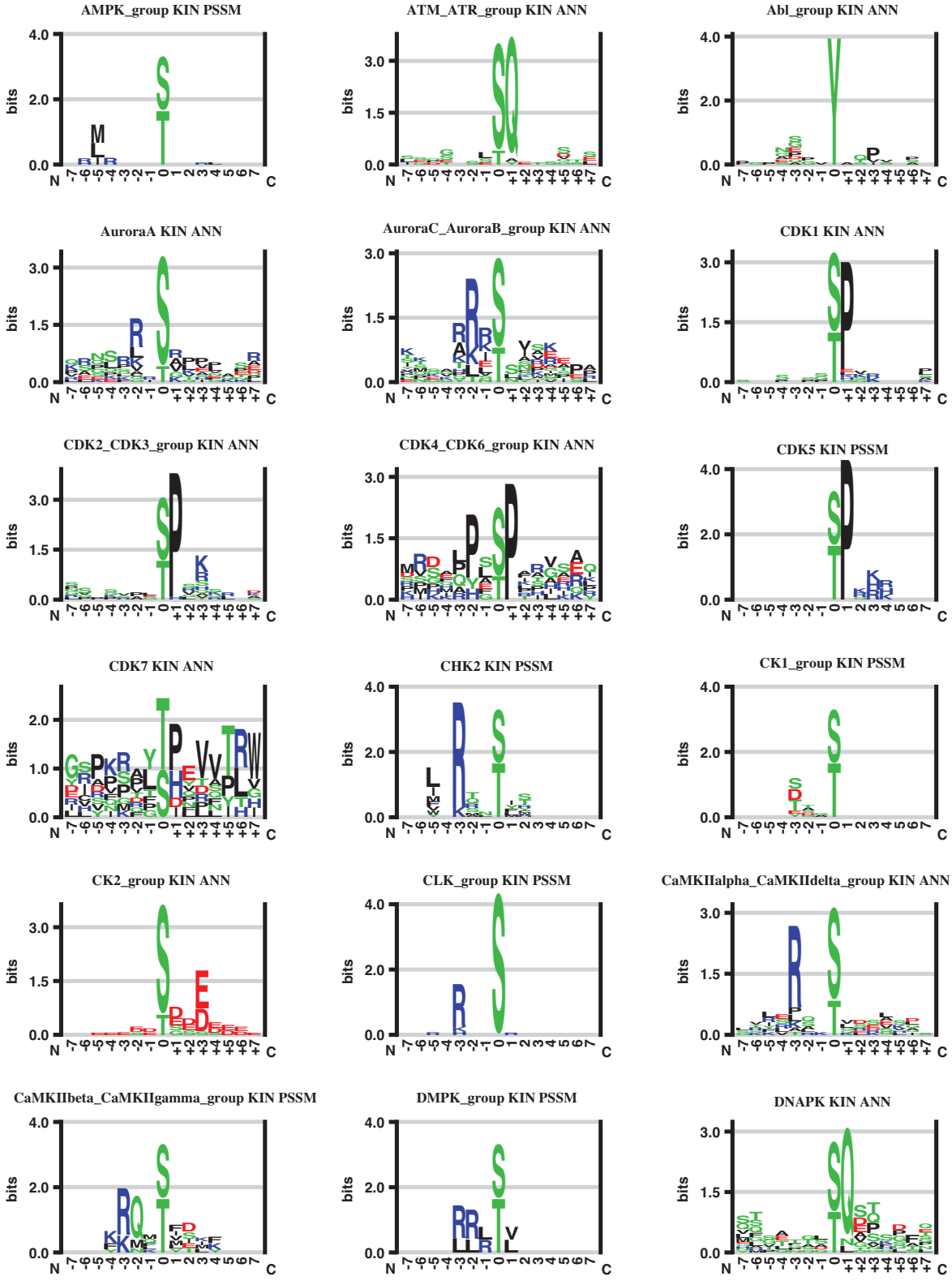
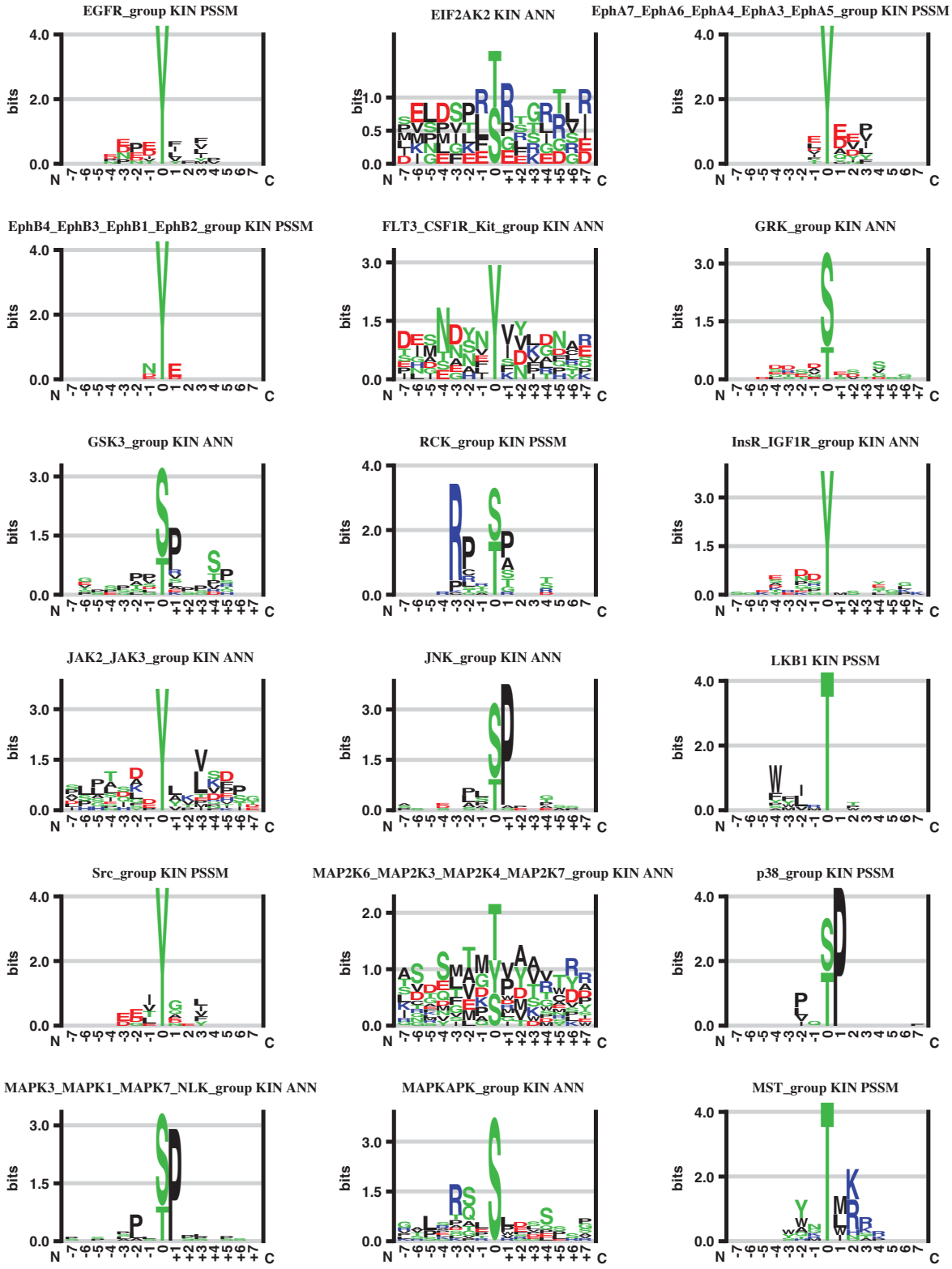
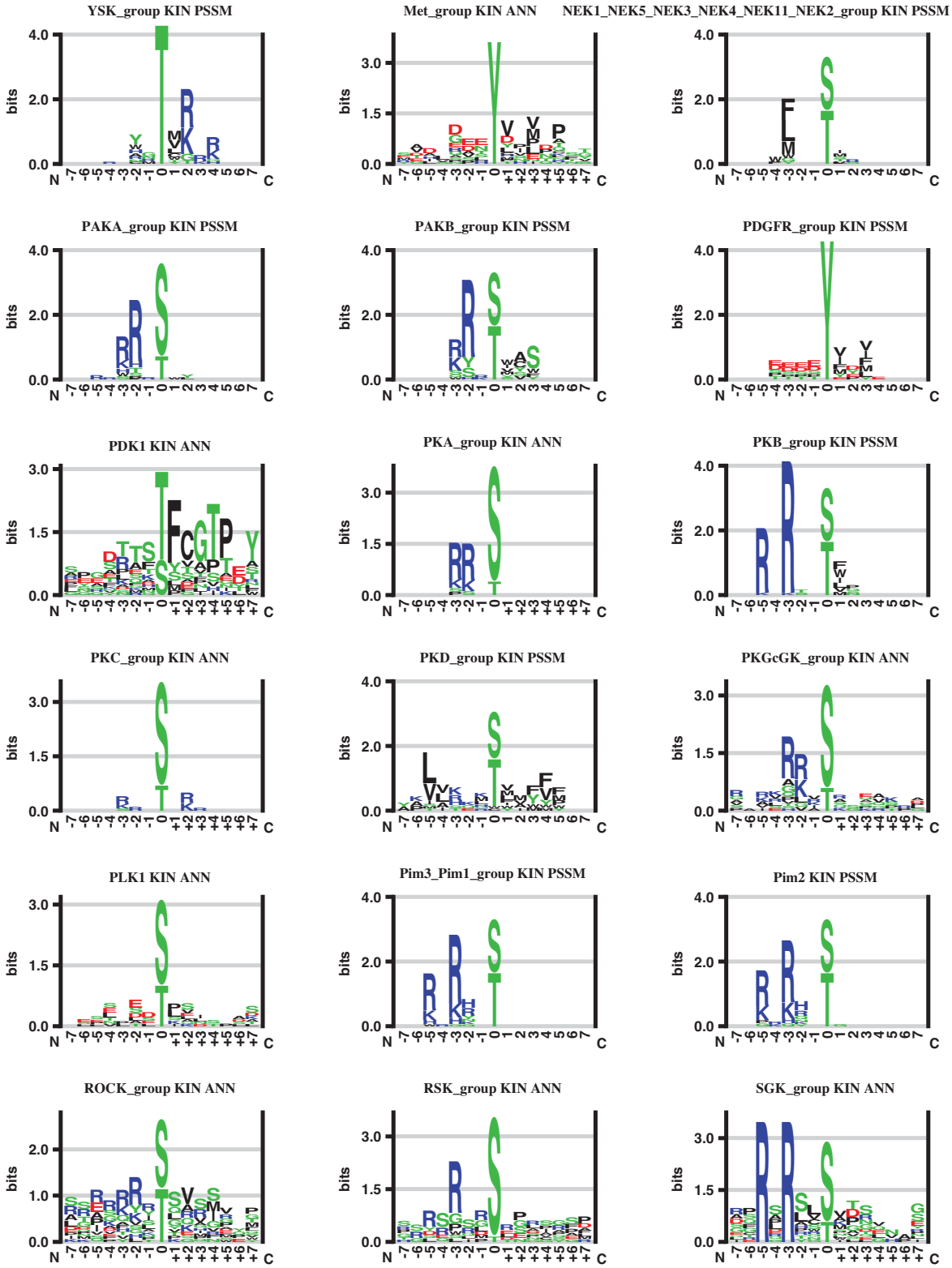


Figure S6: **Kinase matrices from Positional Scanning Peptide Libraries (PSPL).** Protein kinase phosphorylation motifs for Clk2, Clk3, LKB1, Mst1, Mst4, S6K1, Tlk1 and TNIK were determined using arrayed positional scanning peptide libraries as previously described[3]. Briefly, we used a series of biotinylated peptides in which each of nine positions surrounding a central phosphorylation site were systematically substituted with each of the twenty proteogenic amino acids. This set of peptides was subjected in parallel to radiolabel kinase assays followed by capture on streptavidin membranes. Membranes were washed and exposed to a phosphor storage screen.


AMPK_group KIN PSSM
ATM_ATR_group KIN ANN
Abl_group KIN ANN
AuroraA KIN ANN
AuroraC_AuroraB_group KIN ANN
CDK1 KIN ANN
CDK2_CDK3_group KIN ANN
CDK4_CDK6_group KIN ANN
CDK5 KIN PSSM
CDK7 KIN ANN
CHK2 KIN PSSM
CK1_group KIN PSSM
CK2_group KIN ANN
CLK_group KIN PSSM
CaMKIIalpha_CaMKIIdelta_group KIN ANN
CaMKIIbeta_CaMKIIgamma_group KIN PSSM
DMPK_group KIN PSSM
DNAPK KIN ANN
bits
N
C


EGFR_group KIN PSSM
EIF2AK2 KIN ANN
EphA7_EphA6_EphA4_EphA3_EphA5_group KIN PSSM
EphB4_EphB3_EphB1_EphB2_group KIN PSSM
FLT3_CSF1R_Kit_group KIN ANN
GRK_group KIN ANN
GSK3_group KIN ANN
RCK_group KIN PSSM
InsR_IGF1R_group KIN ANN
JAK2_JAK3_group KIN ANN
JNK_group KIN ANN
LKB1 KIN PSSM
Src_group KIN PSSM
MAP2K6_MAP2K3_MAP2K4_MAP2K7_group KIN ANN
p38_group KIN PSSM
MAPK3_MAPK1_MAPK7_NLK_group KIN ANN
MAPKAPK_group KIN ANN
MST_group KIN PSSM
bits
N
C


YSK_group KIN PSSM
Met_group KIN ANN
NEK1_NEK5_NEK3_NEK4_NEK11_NEK2_group KIN PSSM
PAKA_group KIN PSSM
PAKB_group KIN PSSM
PDGFR_group KIN PSSM
PDK1 KIN ANN
PKA_group KIN ANN
PKB_group KIN PSSM
PKC_group KIN ANN
PKD_group KIN PSSM
PKGcGK_group KIN ANN
PLK1 KIN ANN
Pim3_Pim1_group KIN PSSM
Pim2 KIN PSSM
ROCK_group KIN ANN
RSK_group KIN ANN
SGK_group KIN ANN
bits
N
C

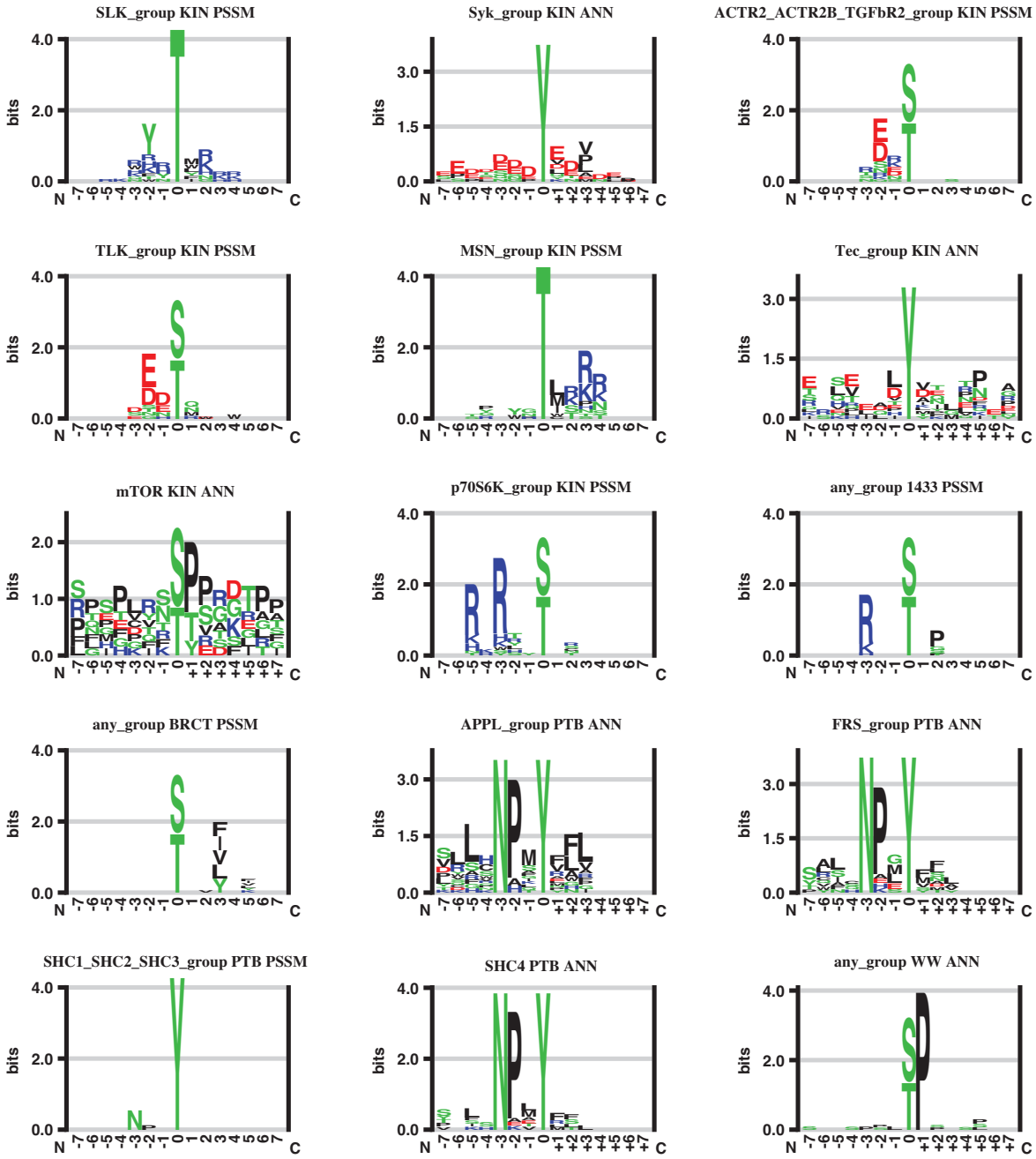
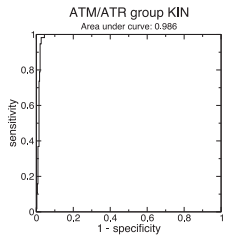
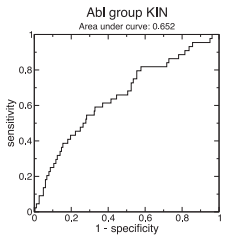
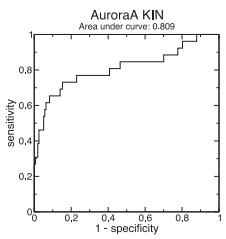
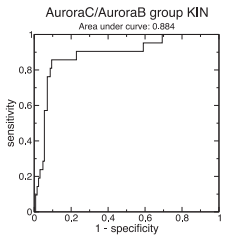


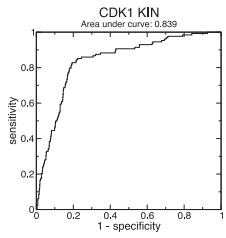
Figure S7: **Sequence logos for kinases and pS/pT-binding domains.** Each logo shows the amino acid preferences at different position relative to the phosphorylation site. The height of the stack of symbols in each position is calculated to be the relative entropy for this position, and the height the individual symbol within a stack is proportional to its frequency. The logos were constructed using enoLOGOS[4]. As new data become available, updated logos will be made available from `http://netphorest.info`.

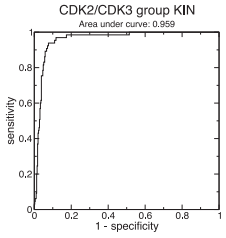

ATM/ATR group KIN
Area under curve: 0.986
sensitivity
1 - specificity

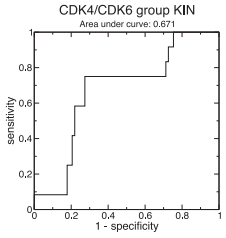

Abl group KIN
Area under curve: 0.652
sensitivity
1 - specificity

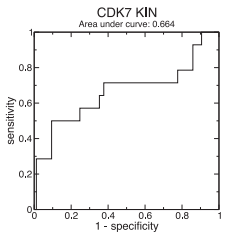

AuroraA KIN
Area under curve: 0.809
sensitivity
1 - specificity

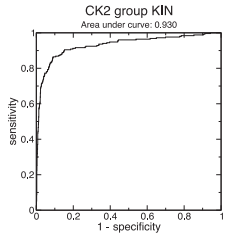

AuroraC/AuroraB group KIN
Area under curve: 0.884
sensitivity
1 - specificity

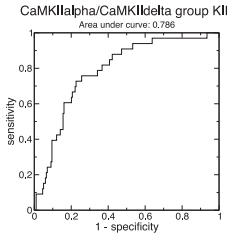

CDK1 KIN
Area under curve: 0.839
sensitivity
1 - specificity

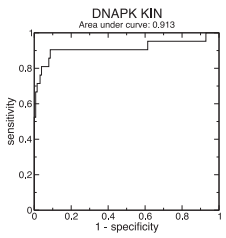

CDK2/CDK3 group KIN
Area under curve: 0.959
sensitivity
1 - specificity

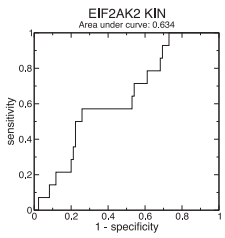

CDK4/CDK6 group KIN
Area under curve: 0.671
sensitivity
1 - specificity

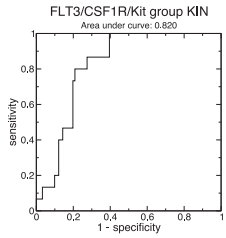

CDK7 KIN
Area under curve: 0.664
sensitivity
1 - specificity

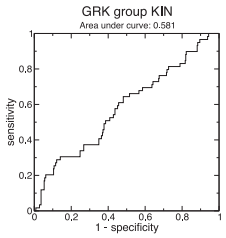

CK2 group KIN
Area under curve: 0.930
sensitivity
1 - specificity

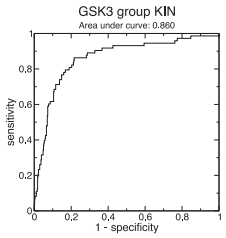

CaMKIIalpha/CaMKIIdelta group KIN
Area under curve: 0.786
sensitivity
1 - specificity

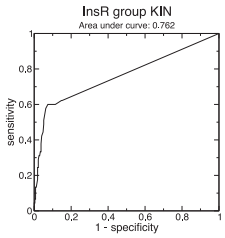

DNAPK KIN
Area under curve: 0.913
sensitivity
1 - specificity

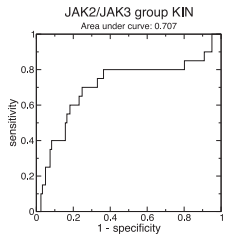

EIF2AK2 KIN
Area under curve: 0.634
sensitivity
1 - specificity

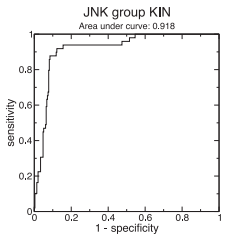

FLT3/CSF1R/Kit group KIN
Area under curve: 0.820
sensitivity
1 - specificity

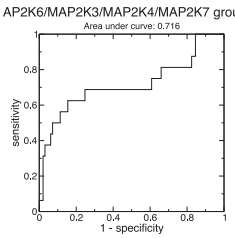

GRK group KIN
Area under curve: 0.581
sensitivity
1 - specificity

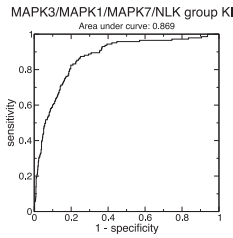

GSK3 group KIN
Area under curve: 0.860
sensitivity
1 - specificity

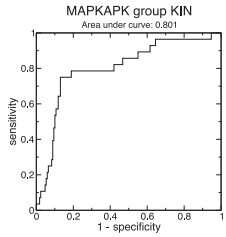

InsR group KIN
Area under curve: 0.762
sensitivity
1 - specificity

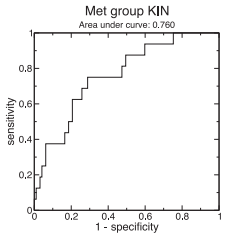

JAK2/JAK3 group KIN
Area under curve: 0.707
sensitivity
1 - specificity

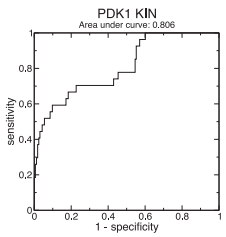

JNK group KIN
Area under curve: 0.918
sensitivity
1 - specificity

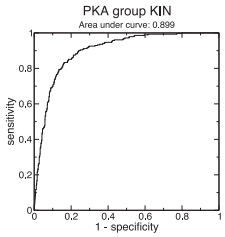

AP2K6/MAP2K3/MAP2K4/MAP2K7 group
Area under curve: 0.716
sensitivity
1 - specificity

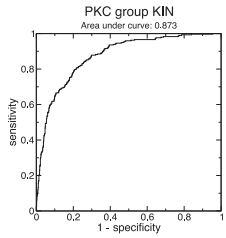

MAPK3/MAPK1/MAPK7/NLK group KIN
Area under curve: 0.869
sensitivity
1 - specificity

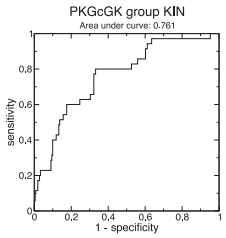

MAPKAPK group KIN
Area under curve: 0.801
sensitivity
1 - specificity

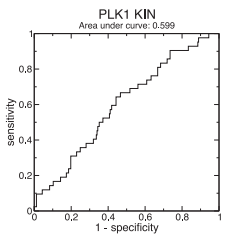

Met group KIN
Area under curve: 0.760
sensitivity
1 - specificity

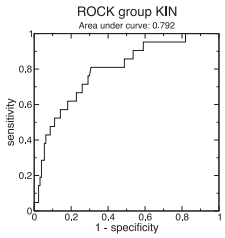

PDK1 KIN
Area under curve: 0.806
sensitivity
1 - specificity

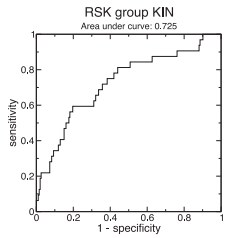

PKA group KIN
Area under curve: 0.899
sensitivity
1 - specificity

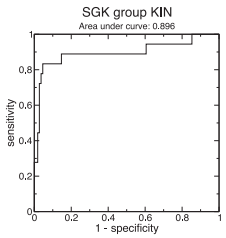

PKC group KIN
Area under curve: 0.873
sensitivity
1 - specificity

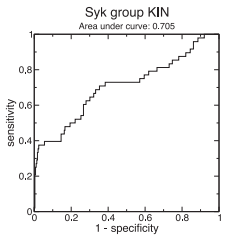

PKGcGK group KIN
Area under curve: 0.761
sensitivity
1 - specificity

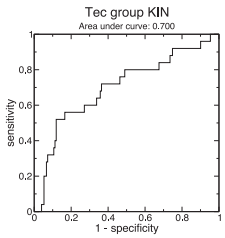

PLK1 KIN
Area under curve: 0.599
sensitivity
1 - specificity

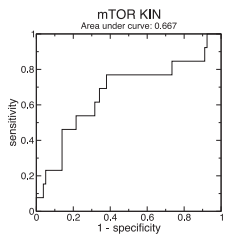

ROCK group KIN
Area under curve: 0.792
sensitivity
1 - specificity

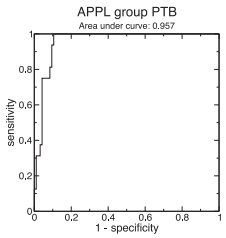

RSK group KIN
Area under curve: 0.725
sensitivity
1 - specificity

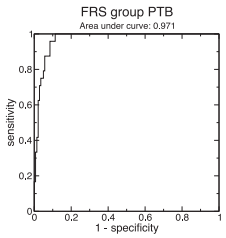

SGK group KIN
Area under curve: 0.896
sensitivity
1 - specificity

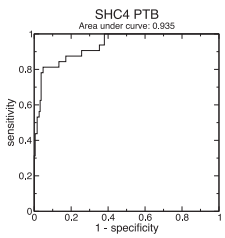

Syk group KIN
Area under curve: 0.705
sensitivity
1 - specificity

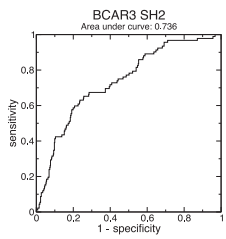

Tec group KIN
Area under curve: 0.700
sensitivity
1 - specificity

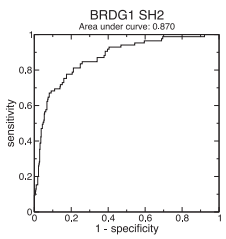

mTOR KIN
Area under curve: 0.667
sensitivity
1 - specificity

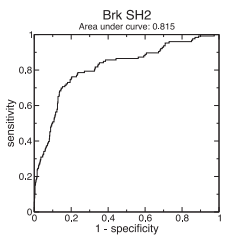

APPL group PTB
Area under curve: 0.957
sensitivity
1 - specificity

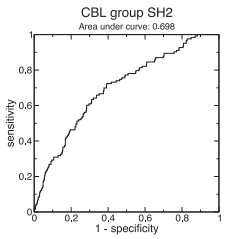

FRS group PTB
Area under curve: 0.971
sensitivity
1 - specificity

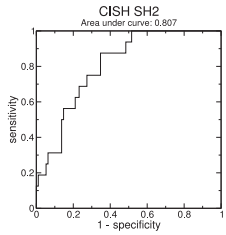

SHC4 PTB
Area under curve: 0.935
sensitivity
1 - specificity

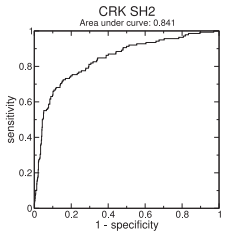

BCAR3 SH2
Area under curve: 0.736
sensitivity
1 - specificity

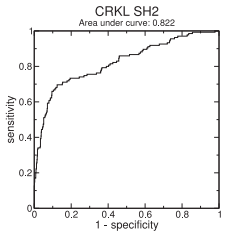

BRDG1 SH2
Area under curve: 0.870
sensitivity
1 - specificity

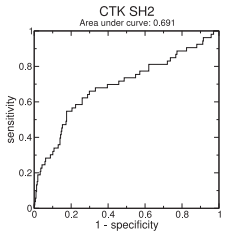

Brk SH2
Area under curve: 0.815
sensitivity
1 - specificity

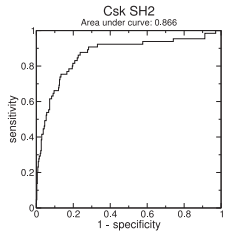

CBL group SH2
Area under curve: 0.698
sensitivity
1 - specificity

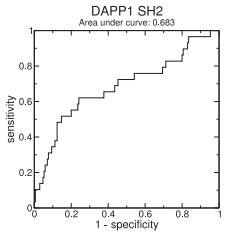

CISH SH2
Area under curve: 0.807
sensitivity
1 - specificity

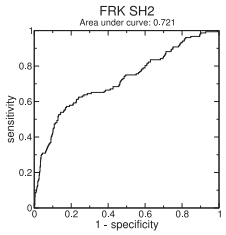

CRK SH2
Area under curve: 0.841
sensitivity
1 - specificity

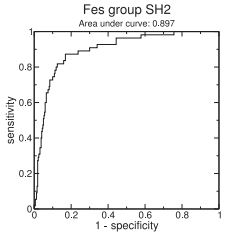

CRKL SH2
Area under curve: 0.822
sensitivity
1 - specificity

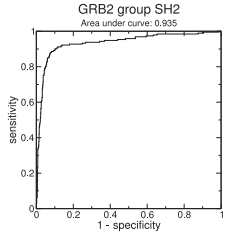

CTK SH2
Area under curve: 0.691
sensitivity
1 - specificity

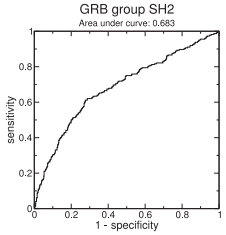

Csk SH2
Area under curve: 0.866
sensitivity
1 - specificity

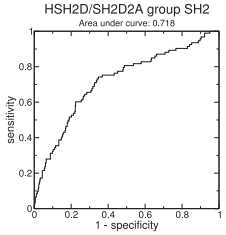

DAPP1 SH2
Area under curve: 0.683
sensitivity
1 - specificity

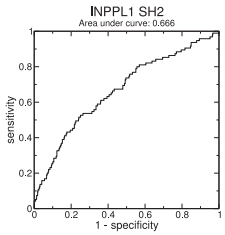

FRK SH2
Area under curve: 0.721
sensitivity
1 - specificity

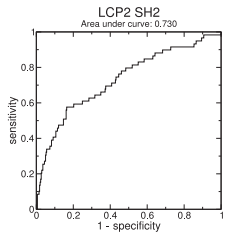

Fes group SH2
Area under curve: 0.897
sensitivity
1 - specificity

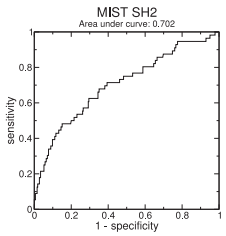

GRB2 group SH2
Area under curve: 0.935
sensitivity
1 - specificity

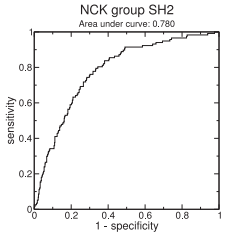

GRB group SH2
Area under curve: 0.683
sensitivity
1 - specificity

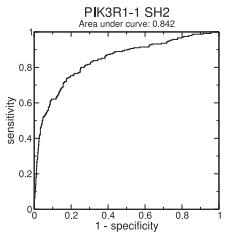

HSH2D/SH2D2A group SH2
Area under curve: 0.718
sensitivity
1 - specificity

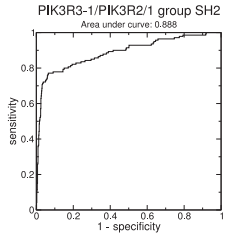

INPPL1 SH2
Area under curve: 0.666
sensitivity
1 - specificity

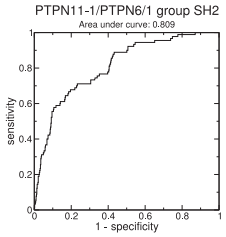

LCP2 SH2
Area under curve: 0.730
sensitivity
1 - specificity

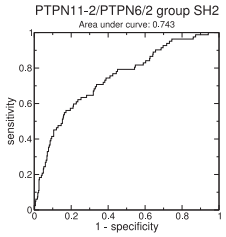

MIST SH2
Area under curve: 0.702
sensitivity
1 - specificity

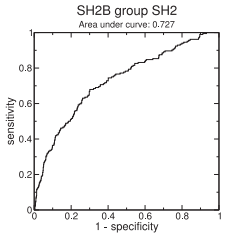

NCK group SH2
Area under curve: 0.780
sensitivity
1 - specificity

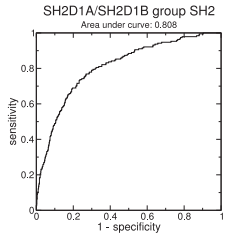

PIK3R1-1 SH2
Area under curve: 0.842
sensitivity
1 - specificity

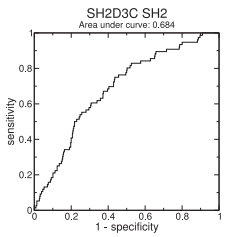

PIK3R3-1/PIK3R2/1 group SH2
Area under curve: 0.888
sensitivity
1 - specificity

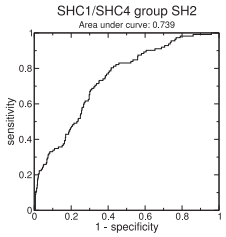

PTPN11-1/PTPN6/1 group SH2
Area under curve: 0.809
sensitivity
1 - specificity

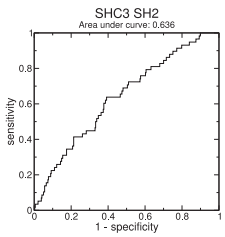

PTPN11-2/PTPN6/2 group SH2
Area under curve: 0.743
sensitivity
1 - specificity

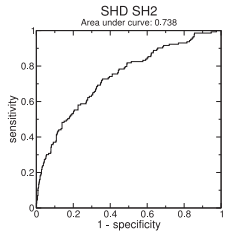

SH2B group SH2
Area under curve: 0.727
sensitivity
1 - specificity

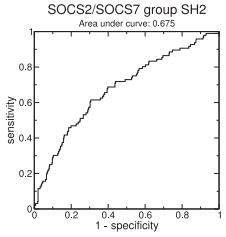

SH2D1A/SH2D1B group SH2
Area under curve: 0.808
sensitivity
1 - specificity

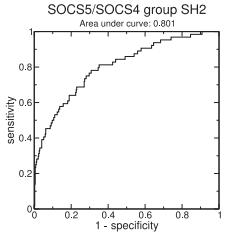

SH2D3C SH2
Area under curve: 0.684
sensitivity
1 - specificity

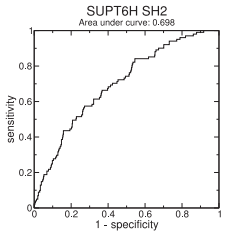

SHC1/SHC4 group SH2
Area under curve: 0.739
sensitivity
1 - specificity

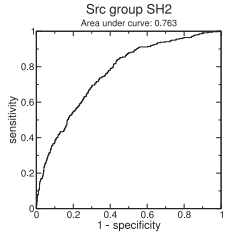

SHC3 SH2
Area under curve: 0.636
sensitivity
1 - specificity

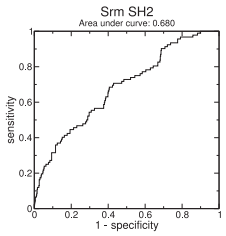

SHD SH2
Area under curve: 0.738
sensitivity
1 - specificity

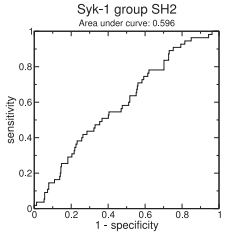

SOCS2/SOCS7 group SH2
Area under curve: 0.675
sensitivity
1 - specificity

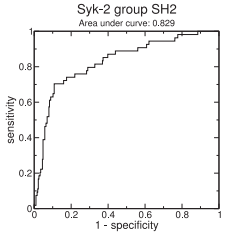

SOCS5/SOCS4 group SH2
Area under curve: 0.801
sensitivity
1 - specificity

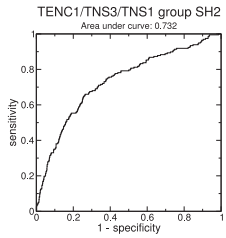

SUPT6H SH2
Area under curve: 0.698
sensitivity
1 - specificity

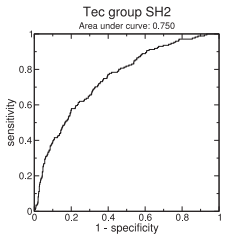

Src group SH2
Area under curve: 0.763
sensitivity
1 - specificity

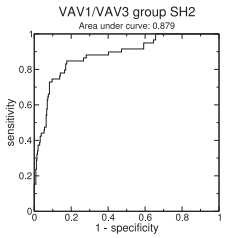

Srm SH2
Area under curve: 0.680
sensitivity
1 - specificity

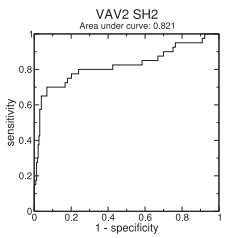

Syk-1 group SH2
Area under curve: 0.596
sensitivity
1 - specificity

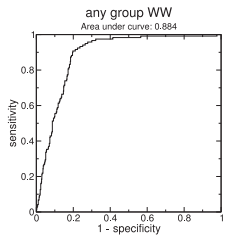

Syk-2 group SH2
Area under curve: 0.829
sensitivity
1 - specificity

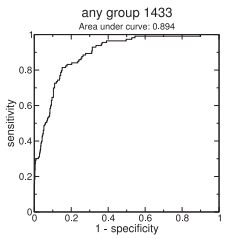

TENC1/TNS3/TNS1 group SH2
Area under curve: 0.732
sensitivity
1 - specificity

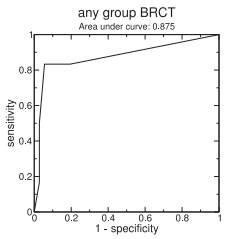

Tec group SH2
Area under curve: 0.750
sensitivity
1 - specificity

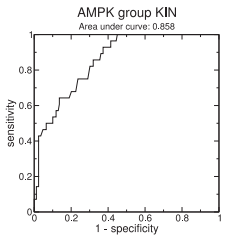

VAV1/VAV3 group SH2
Area under curve: 0.879
sensitivity
1 - specificity


VAV2 SH2
Area under curve: 0.821
sensitivity
1 - specificity


any group WW
Area under curve: 0.884
sensitivity
1 - specificity


any group 1433
Area under curve: 0.894
sensitivity
1 - specificity


any group BRCT
Area under curve: 0.875
sensitivity
1 - specificity


AMPK group KIN
Area under curve: 0.858
sensitivity
1 - specificity

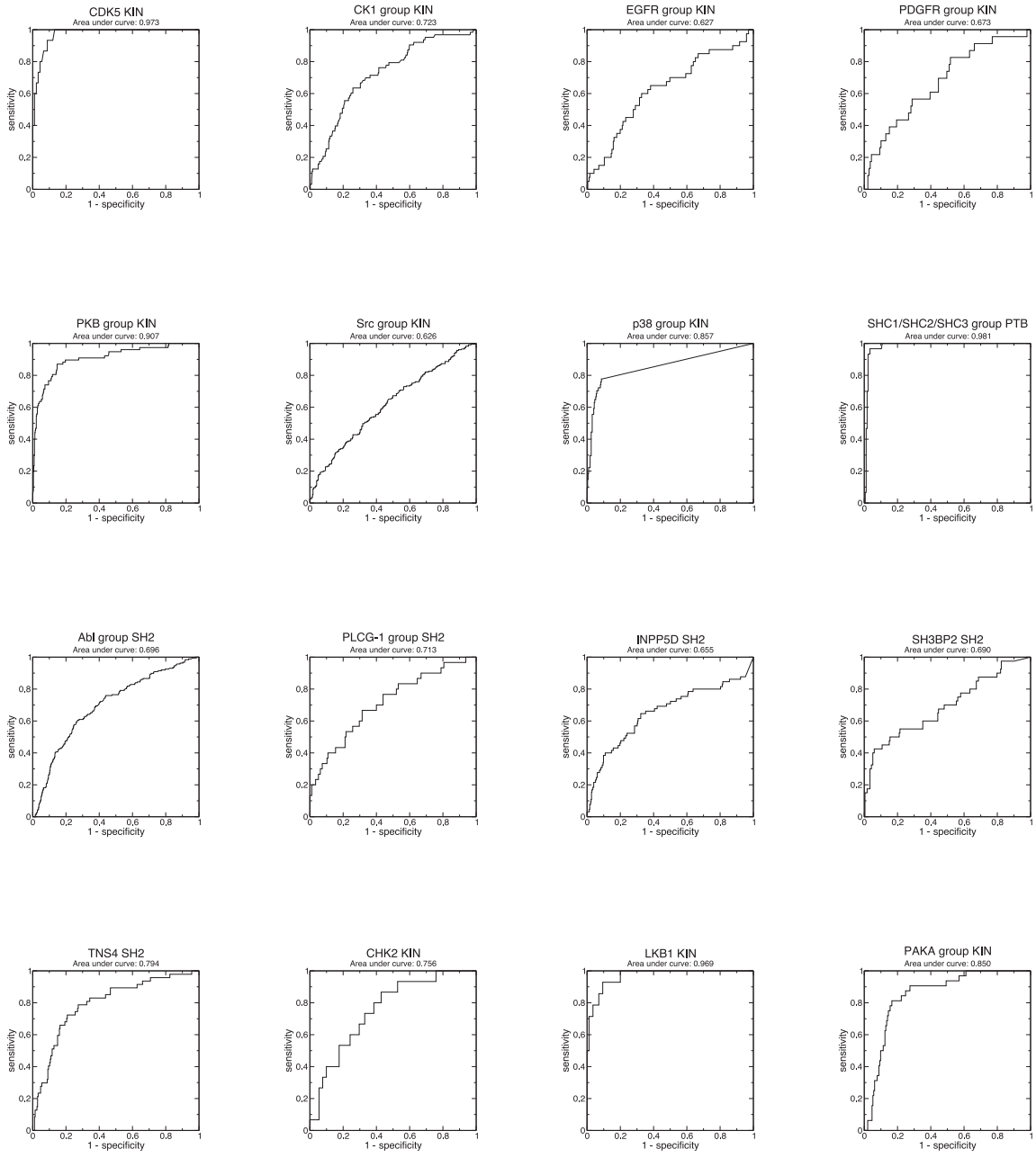


Figure S8: **Receiver output characteristic (ROC) curves for the NetPhorest classifiers.** We benchmarked all classifiers for which at least 12 positive examples were available after homology reduction. Each ROC curve shows the sensitivity as function of the rate of false positivies (1-specificity) for a particular classifier. The performance measures are based on an independent validation set that was not used for training or parameter optimization.

Table S1: **The selected set of NetPhorest classifiers.** The table lists the non-redundant set of classifiers currently incorporated in the NetPhorest atlas. Firstly, we list if the classifier is based on artificial neural networks (ANNs) or a position-specific scoring matrix (PSSM), whether it a classifier for a kinase (KIN) or phospho-binding domain (SH2, PTB, 14-3-3, BRCT or WW) and the types of phosphorylated residues it can recognize (AA). Secondly, we show the number of non-redundant positive sites (POS) used for benchmarking. Finally, we report the performance as area under the receiver operating characteristic curve (AROC); the ROC curves are available in Figure S8 or at `http://netphorest.info`. Due to lack of data, PSSMs could in some cases not be benchmarked. These PSSMs we assumed to have comparable performance to other PSSMs obtained from the same type of assay, and we thus estimate the performance based on the PSSMs that *could* be benchmarked. For the ANNs, the AROC obtained using zero hidden neurons, that is a linear model, is shown in parenthesis.

	Source	**Class**	**AA**	**POS**	**AROC**
14-3-3	PSSM	14-3-3	ST	113	0.894
Abl family	ANN	KIN	Y	44	0.652 (0.610)
Abl family	PSSM	SH2	Y	187	0.696
ACTR2/ACTR2B/TGFbR2	PSSM	KIN	SY	-	0.830
AMPK subfamily	PSSM	KIN	ST	28	0.858
APPL subfamily	ANN	PTB	Y	16	0.957 (0.934)
ATM/ATR	ANN	KIN	ST	57	0.986 (0.985)
AuroraA	ANN	KIN	ST	26	0.809 (0.77)
AuroraC/AuroraB	ANN	KIN	ST	21	0.884 (0.881)
BCAR3	ANN	SH2	Y	92	0.736 (0.737)
BLNK	PSSM	SH2	Y	-	0.749
BRCT	PSSM	BRCT	ST	6	0.875
BRDG1	ANN	SH2	Y	85	0.870 (0.854)
Brk	ANN	SH2	Y	126	0.815 (0.809)
CaMKIIalpha/CaMKIIdelta	ANN	KIN	ST	33	0.786 (0.770)
CaMKIIbeta/CaMKIIgamma	PSSM	KIN	ST	-	0.830
CBL subfamily	ANN	SH2	Y	123	0.698 (0.681)
CDK1	ANN	KIN	ST	128	0.839 (0.82)
CDK2/CDK3	ANN	KIN	ST	65	0.959 (0.959)
CDK4/CDK6	ANN	KIN	ST	12	0.671 (0.810)
CDK5	PSSM	KIN	ST	15	0.973
CDK7	ANN	KIN	ST	14	0.664 (0.532)
CHK2	PSSM	KIN	ST	15	0.756
CISH	ANN	SH2	Y	16	0.807 (0.604)
CK1 family	PSSM	KIN	ST	63	0.723
CK2 family	ANN	KIN	ST	251	0.930 (0.921)
CLK family	PSSM	KIN	ST	-	0.830
CRK	ANN	SH2	Y	138	0.841 (0.823)
CRKL	ANN	SH2	Y	135	0.822 (0.841)
Csk	ANN	SH2	Y	65	0.866 (0.807)
CTK	ANN	SH2	Y	53	0.691 (0.744)
DAPP1	ANN	SH2	Y	29	0.683 (0.694)
DMPK family	PSSM	KIN	ST	-	0.830
DNAPK	ANN	KIN	ST	21	0.913 (0.918)
EGFR family	PSSM	KIN	Y	40	0.627
EIF2AK2	ANN	KIN	ST	14	0.634 (0.560)

	Source	Class	AA	POS	AROC
EphA7/EphA6/EphA4/EphA3/EphA5	PSSM	KIN	Y	-	0.830
EphB4/EphB3/EphB1/EphB2	PSSM	KIN	Y	-	0.830
Fes subfamily	ANN	SH2	Y	55	0.897 (0.846)
FLT3/CSF1R/Kit	ANN	KIN	Y	15	0.820 (0.541)
FRK	ANN	SH2	Y	152	0.721 (0.758)
FRS family	ANN	PTB	Y	24	0.971 (0.939)
GRB2 subfamily	ANN	SH2	Y	193	0.935 (0.923)
GRB family	ANN	SH2	Y	224	0.683 (0.673)
GRK subfamily	ANN	KIN	ST	59	0.581 (0.598)
GSK3 subfamily	ANN	KIN	ST	73	0.860 (0.822)
HSH2D/SH2D2A	ANN	SH2	Y	93	0.718 (0.713)
INPP5D	PSSM	SH2	Y	65	0.655
INPPL1	ANN	SH2	Y	95	0.666 (0.667)
InsR family	ANN	KIN	Y	45	0.762
JAK2/JAK3	ANN	KIN	Y	20	0.707 (0.661)
JNK subfamily	ANN	KIN	ST	49	0.918 (0.940)
LCP2	ANN	SH2	Y	59	0.730 (0.726)
LKB1	PSSM	KIN	T	14	0.969
MAP2K6/MAP2K3/MAP2K4/MAP2K7	ANN	KIN	STY	16	0.716 (0.662)
MAPK3/MAPK1/MAPK7/NLK	ANN	KIN	ST	143	0.869 (0.838)
MAPKAPK family	ANN	KIN	ST	28	0.801 (0.762)
Met family	ANN	KIN	Y	16	0.760 (0.601)
MIST	ANN	SH2	Y	56	0.702 (0.683)
MSN family	PSSM	KIN	ST	-	0.830
MST family	PSSM	KIN	ST	-	0.830
mTOR	ANN	KIN	ST	13	0.667 (0.740)
NCK family	ANN	SH2	Y	117	0.780 (0.785)
NEK1/NEK5/NEK3/NEK4/NEK11/NEK2	PSSM	KIN	ST	-	0.830
p38 subfamily	PSSM	KIN	ST	54	0.857
p70S6K subfamily	PSSM	KIN	ST	-	0.830
PAKA family	PSSM	KIN	ST	32	0.850
PAKB family	PSSM	KIN	ST	-	0.830
PDGFR family	PSSM	KIN	Y	23	0.673
PDK1	ANN	KIN	ST	27	0.806 (0.773)
PIK3R1 1	ANN	SH2	Y	235	0.842 (0.818)
PIK3R1 2	PSSM	SH2	Y	-	0.749
PIK3R2 2	PSSM	SH2	Y	-	0.749
PIK3R3 1/PIK3R2 1	ANN	SH2	Y	140	0.888 (0.881)
PIK3R3 2	PSSM	SH2	Y	-	0.749
Pim2	PSSM	KIN	ST	-	0.830
Pim3/Pim1	PSSM	KIN	S	-	0.830
PKA family	ANN	KIN	ST	282	0.899 (0.878)
PKB family	PSSM	KIN	ST	77	0.907
PKC family	ANN	KIN	ST	296	0.873 (0.867)
PKD family	PSSM	KIN	ST	-	0.830
PKGcGK family	ANN	KIN	ST	35	0.761 (0.681)
PLCG 1 subfamily	PSSM	SH2	Y	30	0.713
PLCG 2 subfamily	PSSM	SH2	Y	-	0.749

	Source	Class	AA	POS	AROC
PLK1	ANN	KIN	ST	42	0.599 (0.498)
PTPN11 1/PTPN6 1	ANN	SH2	Y	90	0.809 (0.792)
PTPN11 2/PTPN6 2	ANN	SH2	Y	82	0.743 (0.703)
RASA1 1	PSSM	SH2	Y	-	0.749
RASA1 2	PSSM	SH2	Y	-	0.749
RCK family	PSSM	KIN	Y	-	0.830
ROCK family	ANN	KIN	ST	21	0.792 (0.787)
RSK family	ANN	KIN	ST	32	0.725 (0.612)
SGK family	ANN	KIN	ST	18	0.896 (0.841)
SH2B family	ANN	SH2	Y	184	0.727 (0.709)
SH2D1A/SH2D1B	ANN	SH2	Y	190	0.808 (0.765)
SH2D3C	ANN	SH2	Y	76	0.684 (0.652)
SH2D6	PSSM	SH2	Y	-	0.749
SH3BP2	PSSM	SH2	Y	40	0.690
SHB	PSSM	SH2	Y	-	0.749
SHC1/SHC2/SHC3	PSSM	PTB	Y	30	0.981
SHC1/SHC4	ANN	SH2	Y	112	0.739 (0.678)
SHC2	PSSM	SH2	Y	-	0.749
SHC3	ANN	SH2	Y	58	0.636 (0.564)
SHC4	ANN	PTB	Y	32	0.935 (0.968)
SHD	ANN	SH2	Y	143	0.738 (0.748)
SHF	PSSM	SH2	Y	-	0.749
SLK family	PSSM	KIN	ST	-	0.830
SOCS2/SOCS7	ANN	SH2	Y	96	0.675 (0.686)
SOCS5/SOCS4	ANN	SH2	Y	64	0.801 (0.708)
Src family	ANN	SH2	Y	339	0.763
Src family	PSSM	KIN	Y	229	0.626 (0.743)
Srm	ANN	SH2	Y	92	0.680 (0.64)
SUPT6H	ANN	SH2	Y	101	0.698 (0.653)
Syk 1 subfamily	ANN	SH2	Y	55	0.596 (0.581)
Syk 2 subfamily	ANN	SH2	Y	54	0.829 (0.862)
Syk family	ANN	KIN	Y	48	0.705 (0.720)
Tec family	ANN	KIN	Y	25	0.700 (0.542)
Tec family	ANN	SH2	Y	171	0.750 (0.716)
TENC1/TNS3/TNS1	ANN	SH2	Y	197	0.732 (0.689)
TLK family	PSSM	KIN	ST	-	0.830
TNS4	PSSM	SH2	Y	47	0.794
VAV1/VAV3	ANN	SH2	Y	59	0.879 (0.869)
VAV2	ANN	SH2	Y	40	0.821 (0.816)
WW	ANN	WW	ST	119	0.884 (0.893)
YSK family	PSSM	KIN	ST	-	0.830

Table S2: **Benchmark of the NetPhorest method.** NetPhorest was compared to four published methods for kinase-specific prediction of phosphorylation sites (Scansite[5], NetPhosK[6], GPS[7] and KinasePhos[8]) and to the simple sequence patterns collected by the ELM[9], PROSITE[10] and HPRD[11] databases. GPS and KinasePhos were benchmarked exclusively on the phosphorylation sites that are dissimilar in sequence to those used for developing the method, whereas the pattern based methods were benchmarked on the full data set (see Methods for details). First, we tested if the method performs significantly ($P < 0.05$) better than random; those which do not are marked with 'R'. Methods performing better than random were subsequently tested to see if they perform significantly poorer than NetPhorest ('P') or comparable to it ('C'). No predictor from any of the tested methods performed significantly better than the corresponding NetPhorest predictor. For each predictor from the other methods, the table lists the area under the receiver operating characteristic curve (AROC), the AROC of the corresponding NetPhorest classifier and the p-value for the comparison of the two. As Scansite[5] and NetPhosK[6] are part of NetPhorest, some comparisons of classifiers from these methods will be self-comparisons; these are marked by an asterisk. It was impossible to compare with other published methods such as PREDIKIN[12] and PredPhospho[13], since these methods do not support batch predictions.

Name	Class	Method	Result	AROC Method	AROC NetPhorest	P-value
Abl	KIN	Scansite	C	0.70	0.70	*
AMPK family	KIN	Scansite	C	0.79	0.79	*
ATM	KIN	NetPhosK	W	0.84	0.99	0
ATM	KIN	Scansite	W	0.95	0.99	0
BLK/Lck/HCK/Lyn	SH2	Scansite	W	0.59	0.73	0
CAMK family	KIN	Scansite	C	0.79	0.81	0.23
CaMKII family	KIN	NetPhosK	W	0.54	0.79	0
CaMKII family	KIN	Scansite	C	0.81	0.81	*
CDK1	KIN	NetPhosK	W	0.75	0.84	0
CDK1	KIN	Scansite	C	0.85	0.85	*
CDK5	KIN	Scansite	C	0.90	0.90	*
CDK family	KIN	NetPhosK	W	0.73	0.86	0
CK1 family	KIN	Scansite	C	0.83	0.83	*
CK2 family	KIN	NetPhosK	W	0.61	0.93	0
CK2 family	KIN	Scansite	C	0.92	0.93	0.32
CRK family	SH2	Scansite	C	0.82	0.83	0.31
DNAPK	KIN	NetPhosK	C	0.88	0.91	0.24
DNAPK	KIN	Scansite	W	0.80	0.91	0.05
EGFR family	KIN	NetPhosK	R	-	0.64	-
EGFR	KIN	Scansite	C	0.63	0.63	*
Fgr/Fyn/Src/Yes	SH2	Scansite	W	0.66	0.73	0
FLT3/CSF1R/Kit/PDGFR	KIN	Scansite	C	0.65	0.65	*
GRB2 family	SH2	Scansite	W	0.89	0.93	0.01
GSK3 subfamily	KIN	NetPhosK	W	0.65	0.86	0
GSK3 subfamily	KIN	Scansite	W	0.73	0.86	0
INPPL1/INPP5D	SH2	Scansite	W	0.56	0.64	0.03
InsR/IGF1R	KIN	NetPhosK	C	0.56	0.76	1
ITK	SH2	Scansite	W	0.76	0.85	0.01
Lck/HCK/Lyn	KIN	Scansite	C	0.59	0.59	0.53
MAPK14	KIN	Scansite	C	0.75	0.79	0.2
MAPK3	KIN	Scansite	W	0.79	0.91	0
NCK family	SH2	Scansite	C	0.74	0.78	0.09

Name	Class	Method	Result	AROC Method	AROC NetPhorest	P-value
p38 subfamily	KIN	NetPhosK	C	0.84	0.84	*
p38 subfamily	KIN	Scansite	C	0.78	0.84	0.11
PIK3R 1 subfamily	SH2	Scansite	W	0.75	0.80	0.03
PKA family	KIN	NetPhosK	W	0.83	0.90	0
PKA family	KIN	Scansite	W	0.86	0.90	0
PKB family	KIN	NetPhosK	W	0.85	0.95	0
PKB family	KIN	Scansite	C	0.95	0.95	*
PKCdelta/PKCtheta	KIN	Scansite	C	0.61	0.64	0.39
PKC family	KIN	NetPhosK	W	0.76	0.87	0
PKC family	KIN	Scansite	W	0.74	0.87	0
PKGcGK family	KIN	NetPhosK	C	0.71	0.76	0.08
RSK family	KIN	NetPhosK	C	0.74	0.74	*
SHC1/SHC4	SH2	Scansite	R	-	0.74	-
Src family	KIN	NetPhosK	C	0.59	0.62	0.37
Src	KIN	Scansite	C	0.56	0.60	0.19
Tec family	SH2	Scansite	W	0.61	0.75	0
14-3-3	14-3-3	HPRD	W	0.73	0.89	0
Abl	KIN	GPS	C	0.74	0.67	0.75
Abl	KIN	HPRD	R	-	0.67	-
Abl	SH2	HPRD	R	-	0.70	-
AGC group	KIN	HPRD	R	-	0.86	-
AMPK subfamily	KIN	GPS	C	0.77	0.86	0.2
AMPK subfamily	KIN	HPRD	C	0.66	0.78	0.07
ATM	KIN	GPS	W	0.85	0.99	0
ATM	KIN	HPRD	W	0.95	0.99	0.03
ATM	KIN	KinasePhos	C	0.91	0.99	0.06
AuroraA	KIN	HPRD	R	-	0.81	-
BRCT	BRCT	HPRD	R	-	0.88	-
CaMKII family	KIN	GPS	R	-	0.79	-
CaMKII family	KIN	HPRD	W	0.74	0.79	0.05
CDK1	KIN	GPS	W	0.7	0.84	0
CDK1	KIN	HPRD	W	0.65	0.84	0
CDK1	KIN	KinasePhos	W	0.7	0.84	0.02
CDK4	KIN	HPRD	R	-	0.82	-
CDK5	KIN	HPRD	W	0.53	0.97	0
CDK family	KIN	ELM	W	0.67	0.86	0
CDK family	KIN	HPRD	W	0.67	0.86	0
CDK family	KIN	KinasePhos	C	0.76	0.87	0.12
CDK family	KIN	Prosite	W	0.67	0.86	0
CK1 family	KIN	ELM	W	0.58	0.72	0.01
CK1 family	KIN	HPRD	W	0.64	0.72	0.05
CK1 family	KIN	Prosite	W	0.58	0.72	0
CK2 family	KIN	ELM	W	0.71	0.93	0
CK2 family	KIN	GPS	W	0.72	0.93	0
CK2 family	KIN	HPRD	W	0.88	0.93	0
CK2 family	KIN	KinasePhos	R	-	0.93	-
CK2 family	KIN	Prosite	W	0.82	0.93	0
CRK family	SH2	HPRD	C	0.78	0.83	0.13

Name	Class	Method	Result	AROC Method	AROC NetPhorest	P-value
DNAPK	KIN	HPRD	C	0.78	0.91	0.11
EGFR family	KIN	HPRD	R	-	0.63	-
EGFR	KIN	HPRD	R	-	0.59	-
Fes	SH2	HPRD	C	0.77	0.83	0.09
Fgr	SH2	HPRD	R	-	0.65	-
Fyn	SH2	HPRD	W	0.51	0.83	0
GRB2	SH2	ELM	C	0.87	0.88	0.45
GRB family	SH2	HPRD	W	0.59	0.68	0
GRK family	KIN	HPRD	R	-	0.58	-
GSK3 subfamily	KIN	ELM	W	0.73	0.86	0
GSK3 subfamily	KIN	HPRD	W	0.77	0.86	0
GSK3 subfamily	KIN	Prosite	W	0.73	0.86	0
INPPL1	SH2	HPRD	R	-	0.67	-
ITK	SH2	HPRD	R	-	0.85	-
JAK2/JAK3	KIN	HPRD	C	0.71	0.71	0.53
JNK subfamily	KIN	HPRD	R	-	0.92	-
LKB1	KIN	HPRD	R	-	0.97	-
MAP2K family	KIN	HPRD	R	-	0.81	-
MAPK3	KIN	HPRD	R	-	0.91	-
MAPKAPK2	KIN	HPRD	R	-	0.86	-
MAPKAPK family	KIN	HPRD	R	-	0.80	-
mTOR	KIN	HPRD	R	-	0.67	-
NCK family	SH2	HPRD	R	-	0.78	-
PAK family	KIN	HPRD	R	-	0.78	-
PDGFR family	KIN	HPRD	R	-	0.67	-
PDK1	KIN	HPRD	R	-	0.81	-
PIKK family	KIN	ELM	C	0.91	0.91	0.32
PIKK family	KIN	Prosite	C	0.91	0.91	0.35
PKA family	KIN	ELM	W	0.80	0.90	0
PKA family	KIN	GPS	W	0.77	0.9	0
PKA family	KIN	HPRD	W	0.86	0.90	0.01
PKA family	KIN	KinasePhos	W	0.72	0.9	0
PKA family	KIN	Prosite	W	0.66	0.90	0
PKB family	KIN	ELM	C	0.85	0.91	0.25
PKB family	KIN	GPS	R	-	0.91	-
PKB family	KIN	HPRD	W	0.86	0.91	0.03
PKB family	KIN	Prosite	C	0.85	0.91	0.14
PKCalpha	KIN	HPRD	R	-	0.82	-
PKC family	KIN	GPS	W	0.67	0.87	0
PKC family	KIN	HPRD	W	0.78	0.87	0
PKC family	KIN	KinasePhos	W	0.66	0.87	0
PKC family	KIN	Prosite	W	0.71	0.87	0
PKGcGK family	KIN	HPRD	R	-	0.76	-
PLCG 1 subfamily	SH2	HPRD	R	-	0.71	-
PLK1	KIN	HPRD	R	-	0.60	-
PTPN11 1	SH2	ELM	W	0.54	0.78	0
PTPN family	SH2	HPRD	W	0.63	0.76	0
SHC1	SH2	HPRD	R	-	0.65	-

Name	Class	Method	Result	AROC Method	AROC NetPhorest	P-value
SHC family	PTB	HPRD	W	0.90	0.97	0.02
Src family	KIN	HPRD	R	-	0.62	-
Src family	SH2	HPRD	W	0.56	0.76	0
Src	KIN	HPRD	C	0.58	0.60	0.37
Src	SH2	ELM	W	0.61	0.74	0
Syk 1	SH2	HPRD	R	-	0.62	-
Syk 2	SH2	HPRD	R	-	0.83	-
Syk	KIN	HPRD	R	-	0.67	-
TK group	KIN	HPRD	R	-	0.79	-
TNS family	SH2	HPRD	R	-	0.70	-
VAV family	SH2	HPRD	R	-	0.84	-
WW	WW	HPRD	C	0.88	0.88	0.24

References

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

[2] B. A. Liu, K. Jablonowski, M. Raina, M. Arcé, T. Pawson, P. D. Nash, The human and mouse complement of SH2 domain proteins—establishing the boundaries of phosphotyrosine signaling. *Mol. Cell* **22**, 851–868 (2006).

[3] J. E. Hutti, E. T. Rarrell, J. D. Chang, D. W. Abbott, P. Storz, A. Toker, L. C. Cantley, B. E. Turk, A rapid method for determining protein kinase phosphorylation specificity. *Nat. Methods* **1**, 27–29 (2004).

[4] C. T. Workman, Y. Yin, D. L. Corcoran, T. Ideker, G. D. Stormo, P. V. Benos, enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res.* **33**, W389–W392 (2005).

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

[6] N. Blom, T. Sicheritz-Pontén, R. Gupta, S. Gammeltoft, S. Brunak, Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4**, 1633–1649 (2004).

[7] Y. Xue, F. Zhou, M. Zhu, K. Ahmed, G. Chen, X. Yao, GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res.* **33**, W184–W187 (2005).

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

[9] P. Puntervoll, R. Linding, C. Gemund, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D. M. Martin, G. Ausiello, B. Brannetti, A. Costantini, F. Ferre, V. Maselli, A. Via, G. Cesareni, F. Diella, G. Superti-Furga, L. Wyrwicz, C. Ramu, C. McGuigan, R. Gudavalli, I. Letunic, P. Bork, L. Rychlewski, B. Kuster, M. Helmer-Citterich, W. N. Hunter, R. Aasland, T. J. Gibson, ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* **31**, 3625–3630 (2003).

[10] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni, C. J. A. Sigrist, The PROSITE database. *Nucleic Acids Res.* **34**, D227–D230 (2006).

[11] R. Amanchy, B. Periaswamy, S. Mathivanan, R. Reddy, S. G. Tattikota, A. Pandey, A curated compendium of phosphorylation motifs. *Nat. Biotechnol.* **25**, 285–286 (2007).

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

[13] J. H. Kim, J. Lee, B. Oh, K. Kimm, I. Koh, Prediction of phosphorylation sites using SVMs. *Bioinformatics* **20**, 3179–3184 (2004).