Supplementary Materials for:

Automated Identification of Binding Sites for Phosphorylated Ligands in Protein Structures

Dario Ghersi[#] and Roberto Sanchez*

Department of Structural and Chemical Biology, Mount Sinai School of Medicine, New York, New York, USA

[#]Current address: Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08540, USA.

* To whom correspondence should be addressed:
1425 Madison Avenue, New York, NY 10029, USA
Phone: 212 659 8648
E-mail: roberto@sanchezlab.org or roberto.sanchez@mssm.edu

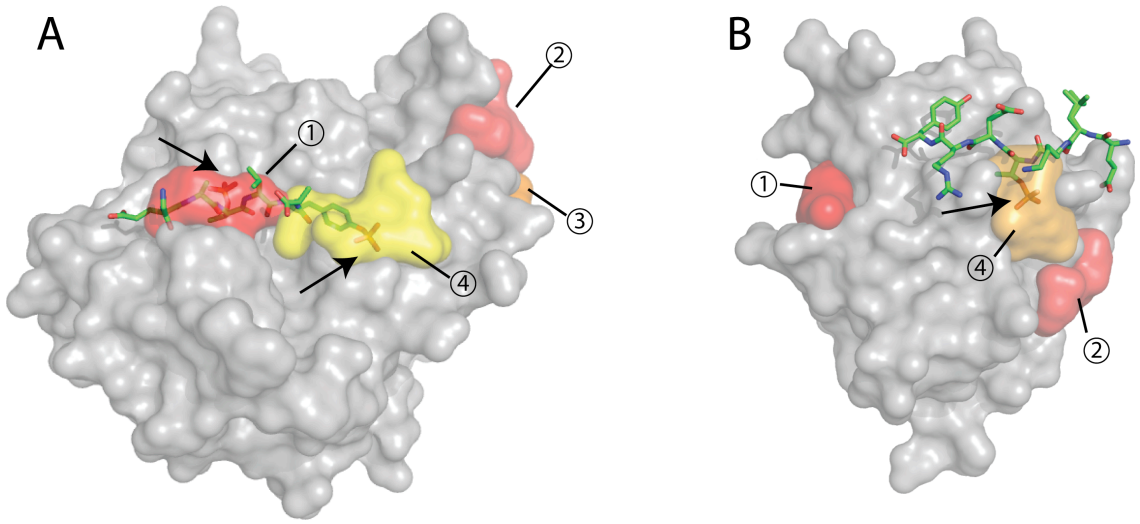

Figure S1. Problematic phospholigand binding site identification cases. The ranking of clusters is indicated in color (higher ranking clusters are more red) and number, and arrows indicate the location of phosphate groups. (A) Phosphothreonine lyase (2z8p) (Chen, et al., 2008): the peptide contains two phosphorylated residues, the first ranking cluster identifies one site, but only the fourth ranking cluster identifies the second site. (B) FHA domain of RNF8 in complex with a phosphopeptide (2pie) (Huen, et al., 2007): the fourth ranking cluster correctly identifies the binding site, whereas the first three clusters probably represent decoy sites (the third-ranking cluster is on the opposite side of the structure, and not visible in the figure). Conservation-based reranking of these clusters moves the fourth cluster to the top. PDB codes are indicated in parentheses.

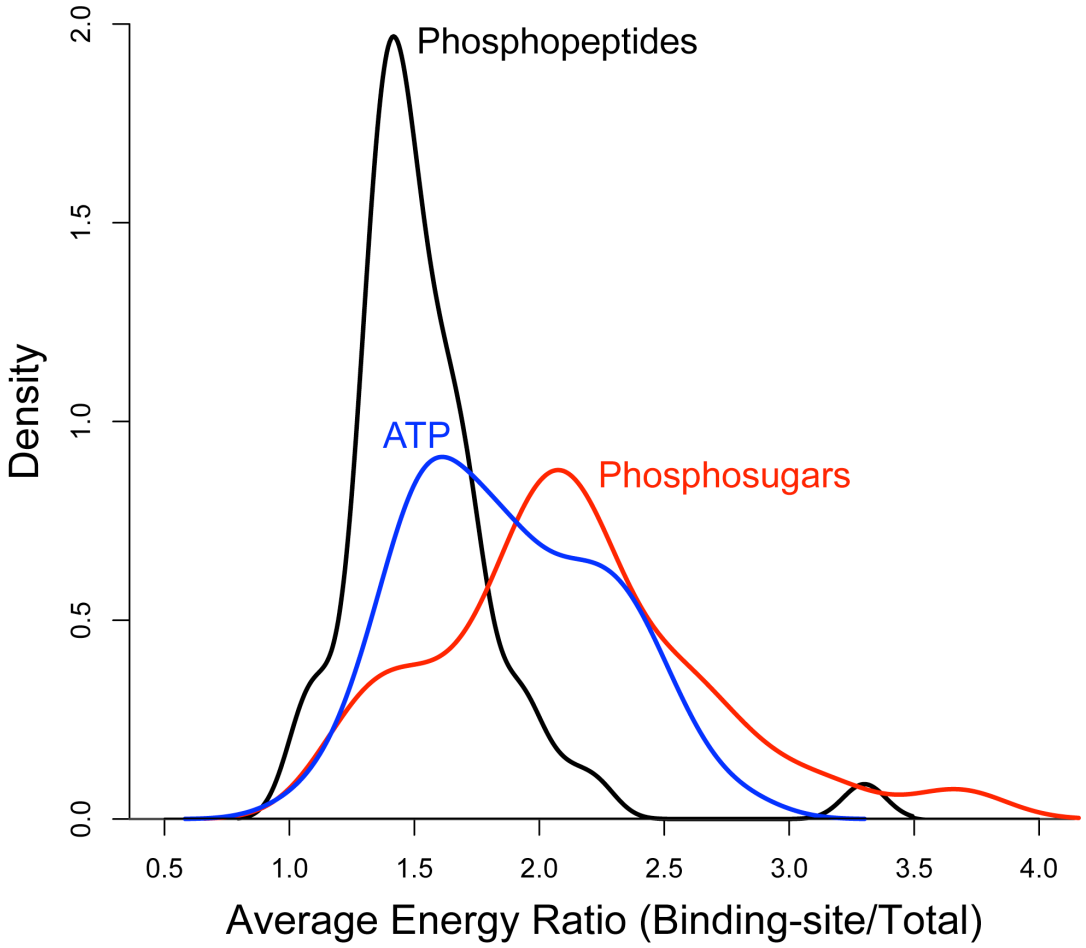

Figure S2. **Relative strength of the energy-based signal in the three types of phospholigand binding sites.** Density plot with the ratio between the average interaction energy in the binding site and the average interaction energy on the whole protein surface (black: phosphopeptides; blue: ATP; red: phosphosugars). The strongest signal comes from the binding sites involved in the recognition of phosphosugars followed by ATP binding sites, while phosphopeptide binding sites present a comparatively weaker signal.

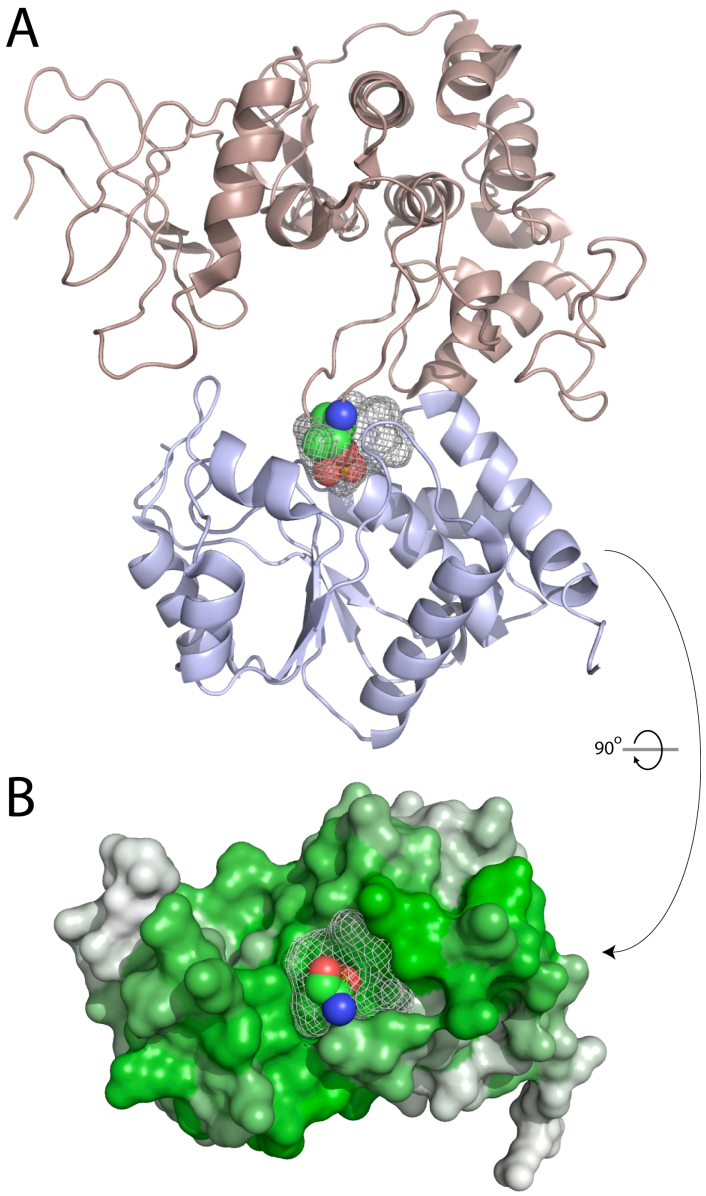

Figure S3. Example of increased specificity of the energy-based over the conservation-based approach. (**A**) Kinase associated phosphatase (light blue) in complex with phospho-cdk2 (light brown) (pdb code: 1fq1) (Song, et al., 2001). The phospho-threonine residue is shown in space-filling representation and first-ranking cluster identifying the phospho-threonine binding site is shown as a gray mesh. (**B**) The surface of the kinase associated phosphatase is colored by conservation (dark green: highly conserved, white: non conserved). The phospho-threonine residue of cdk2 and first- ranking cluster are depicted as described in (A).

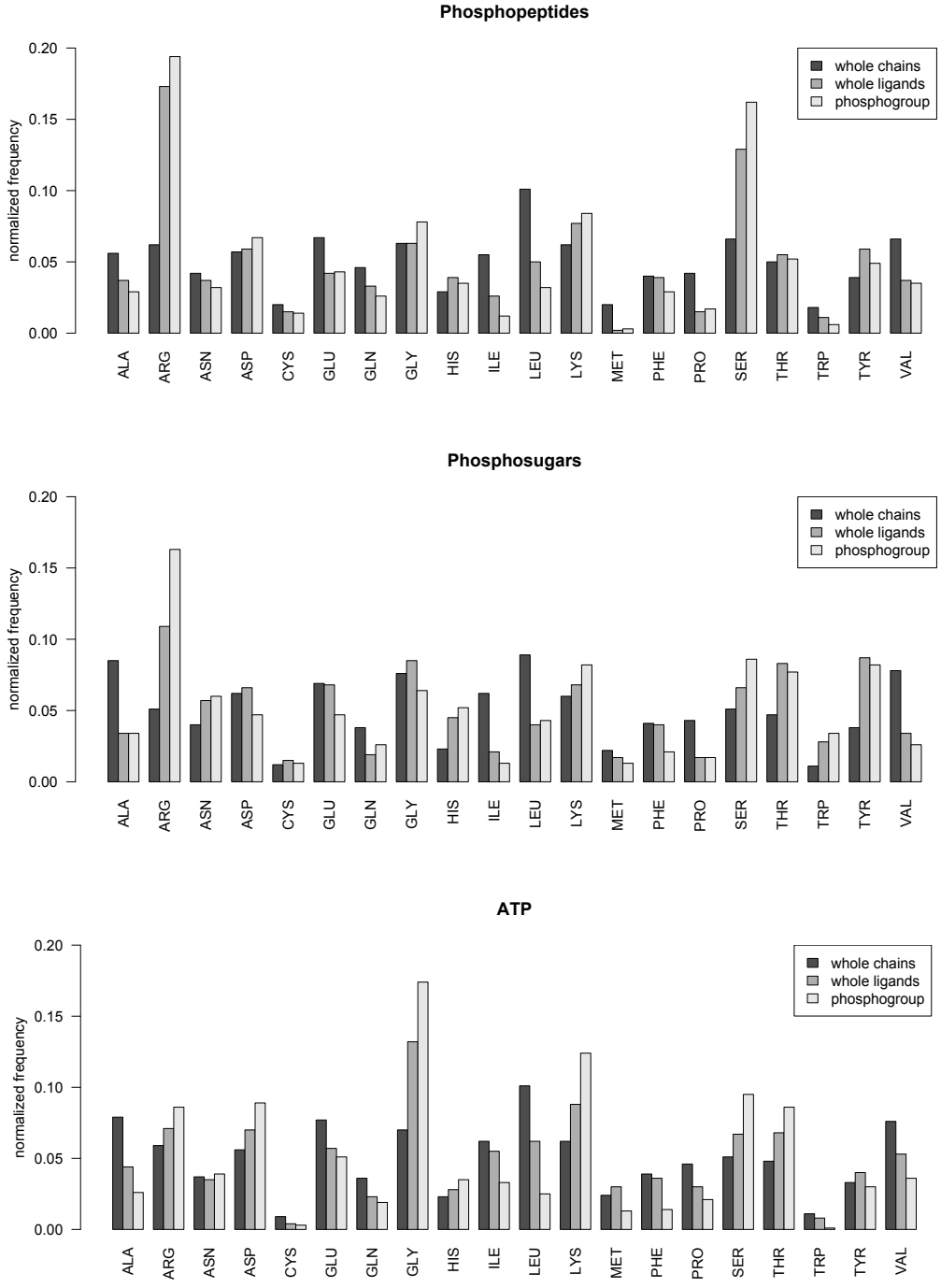

Figure S4. Amino acid distribution in the three phospholigand datasets. The bar plots show the amino acid distribution of the protein chains, of the binding sites for the whole ligand and of the part of the binding sites in contact with the phosphogroup only. Commonalities emerge (e.g., aliphatic amino acids are underrepresented, positively charged amino acids and serine/threonine are overrepresented), as well as differences (higher serine and arginine in phosphopeptides; higher arginine in phosphosugars, and higher glycine and lysine in the ATP dataset).

Pfam ID	Pfam Name	PDB codes
PF00017	SH2 domain	1aycA 1d4wA 1h9oA 1i3zA 1jyrA 1shaA 1yvlA 2ciaA 2hdxA 2hmhA 2iuhA 2oq1A 2vifA
PF00089	Trypsin	2hwlD
PF00102	Protein-tyrosine phosphatase	1eeoA 1fprA 1xxpA 1ygrA 2i6oA
PF00129	Class I Histocompatibility antigen, domains alpha 1 and 2	3bgmA
PF00155	Aminotransferase class I and II	1lc7A
PF00244	14-3-3 protein	2br9A
PF00385	Chromodomain	2b2tA
PF00397	WW domain	2q5aA
PF00400	WD domain, G-beta repeat	1nexB 1p22A 2ovrB
PF00498	FHA domain	1g6gA 1gxcA 2pieA
PF00533	BRCA1 C Terminus (BRCT) domain	1t15A 2vxcA
PF00583	Acetyltransferase (GNAT) family	1puaA
PF00659	POLO box duplicated region	1umwA
PF00702	Haloacid dehalogenase-like hydrolase	1l7pB
PF00782	Dual specificity phosphatase, catalytic domain	1j4xA 1oheA
PF01111	Cyclin-dependent kinase regulatory subunit	2astC
PF01331	mRNA capping enzyme, catalytic domain	1p16A
PF02762	CBL proto-oncogene N-terminus, SH2-like domain	3buxB
PF03031	NLI interacting factor-like phosphatase	2ghtA
PF03166	MH2 domain	1u7fB
PF03536	Salmonella virulence-associated 28kDa protein	2z8pA
PF05706	Cyclin-dependent kinase inhibitor 3 (CDKN3)	1fq1A
PF07475	HPr Serine kinase C-terminal domain	1kkmA

Table S1. Pfam domain distribution in the phosphopeptides dataset.

Pfam ID	Pfam Name	PDB codes
PF00004	ATPase family associated with various cellular activities (AAA)	1j7kA 1nsfA
PF00005	ABC transporter	1b0uA 1ji0A 1r0xC
PF00012	Hsp70 protein	1kaxA
PF00022	Actin	1qz5A 1tyqA
PF00063	Myosin head (motor domain)	1fmwA
PF00069	Protein kinase domain	1csnA 1ol6A 1phkA 1u5rB 1zydA 2biyA 2cchA
PF00078	Reverse transcriptase (RNA-dependent DNA polymerase)	2iajA
PF00118	TCP-1/cpn60 chaperonin family	1sx3J
PF00142	4Fe-4S iron sulfur cluster binding proteins, NifH/frxC family	2c8vA
PF00158	Sigma-54 interaction domain	1ojlE 2c96A
PF00162	Phosphoglycerate kinase	1vjdA
PF00293	NUDIX domain	1vc9A
PF00294	pfkB family carbohydrate kinase	1lhrB 2f02A
PF00483	Nucleotidyl transferase	1yp3A
PF00543	Nitrogen regulatory protein P-II	2gnkA
PF00571	CBS domain	2j9lC
PF00579	tRNA synthetases class I (W and Y)	1m83A 1yidB
PF00580	UvrD/REP helicase	1qhgA
PF00582	Universal stress protein family	1mjhA
PF00586	AIR synthase related protein, N-terminal domain	2hs0A
PF00613	Phosphoinositide 3-kinase family, accessory domain (PIK domain)	1e8xA
PF00619	Caspase recruitment domain	2a5yC
PF00626	Gelsolin repeat	2fghB
PF00680	RNA dependent RNA polymerase	2ilyA
PF00733	Asparagine synthase	1mb9B
PF00749	tRNA synthetases class I (E and Q), catalytic domain	1n75A
PF00764	Arginosuccinate synthase	1kp2A
PF00931	NB-ARC domain	2a5yC
PF01068	ATP dependent DNA ligase domain	1a0iA 2hixA
PF01121	Dephospho-CoA kinase	1jjvA 1uf9C 1vhtC
PF01163	RIO1 family	1zaoA
PF01202	Shikimate kinase	1ko5A 2iywA
PF01293	Phosphoenolpyruvate carboxykinase	1xkvA
PF01336	OB-fold nucleic acid binding domain	1b8aB
PF01411	tRNA synthetases class II (A)	1yfrA
PF01467	Cytidylyltransferase	1f9aE 1yunB
PF01743	Poly A polymerase head domain	1miwB
PF01909	Nucleotidyltransferase domain	1r8bA
PF01923	Cobalamin adenosyltransferase	2nt8A
PF01948	Aspartate carbamoyltransferase regulatory chain, allosteric domain	4at1D
PF02503	Polyphosphate kinase	1xdpA

PF02569	Pantoate-beta-alanine ligase	2a84A
PF02572	ATP:corrinoid adenosyltransferase	1g5tA
PF02750	Synapsin, ATP binding domain	1pk8D
PF02769	AIR synthase related protein, C-terminal domain	2hs0A
PF02786	Carbamoyl-phosphate synthase L chain, ATP binding domain	1dv2A
PF02872	5'-nucleotidase, C-terminal domain	1hp1A
PF03099	Biotin/lipoate A/B protein ligase family	2aruA
PF03205	Molybdopterin guanine dinucleotide synthesis protein	2npiB
PF03237	Terminase-like family	2o0hA
PF07931	Chloramphenicol phosphotransferase-like protein	1qhxA
PF08543	Phosphomethylpyrimidine kinase	2ddoA
PF08544	GHMP kinases C terminal	1kvkA

Table S2. Pfam domain distribution in the ATP dataset

Pfam ID	Pfam Name	PDB codes
PF00171	Aldehyde dehydrogenase family	1uxtA
PF00232	Glycosyl hydrolase family 1	4pbgA
PF00248	Aldo/keto reductase family	2acqA
PF00316	Fructose-1-6-bisphosphatase	1nuwA 2ox3A
PF00328	Histidine phosphatase superfamily (branch 2)	1nt4A
PF00342	Phosphoglucose isomerase	1t10A
PF00343	Carbohydrate phosphorylase	1l5vA 2priA
PF00365	Phosphofructokinase	4pfkA
PF00408	Phosphoglucomutase/phosphomannomutase, C-terminal domain	1p5dX
PF00483	Nucleotidyl transferase	1g23G
PF00532	Periplasmic binding proteins and sugar binding domain of LacI family	1bykB 2nzuG
PF00702	Haloacid dehalogenase-like hydrolase	1z4oA 2b0cA
PF00878	Cation-independent mannose-6-phosphate receptor repeat	1sz0B
PF01380	SIS domain	1moqA
PF01791	DeoC/LacD family aldolase	1w8sI
PF01872	RibD C-terminal domain 2obcA	2obcA
PF02056	Family 4 glycosyl hydrolase	1u8xX 1up7H
PF02157	Mannose-6-phosphate receptor	1c39A
PF02781	Glucose-6-phosphate dehydrogenase, C-terminal domain	1e77A 2bhlA
PF02878	Phosphoglucomutase/phosphomannomutase, alpha/beta/alpha domain I	1p5dX
PF02879	Phosphoglucomutase/phosphomannomutase, alpha/beta/alpha domain II	1p5dX
PF02880	Phosphoglucomutase/phosphomannomutase, alpha/beta/alpha domain III	1p5dX
PF02887	Pyruvate kinase, alpha/beta domain	2vgbA
PF03332	Eukaryotic phosphomannomutase	2fueA
PF06026	Ribose 5-phosphate isomerase A (phosphoriboisomerase A)	1o8bB
PF06560	Glucose-6-phosphate isomerase (GPI)	2gc3A

Table S3. Pfam domain distribution in the phosphosugars dataset.

Name	HET code	PDB codes
1-O-phosphono-alpha-D-galactopyranose	GL1	1z4o
2-deoxy-glucose-6-phosphate	D6G	2pri
5-O-phosphono-beta-D-ribofuranose	RP5	2obc
alfa-D-glucose-1-phosphate	G1P	1g23 1l5v 1nt4 1p5d 1uxt 2b0c
alfa-D-glucose-6-phosphate	G6P	1u8x 1up7 2acq
alfa-D-mannose 1-phosphate	M1P	2fue
alpha-D-mannose-6-phosphate	M6P	1sz0 2gc3
beta-D-arabinofuranose-5'-phosphate	ABF	1o8b
beta-fructose-1,6-diphosphate	FBP	1w8s 2vgb
beta-D-glucose-6-phosphate	BG6	1e77 2bhl 2nzu
beta-galactose-6-phosphate	BGP	4pbg
fructose-6-phosphate	F6P	1nuw 1t10 2ox3 4pfk
glucosamine 6-phosphate	GLP	1moq
pentamannosyl 6-phosphate	P3M	1c39
trehalose-6-phosphate	T6P	1byk

Table S4. Phosphosugars included in the analysis. For each different phosphosugar, the name, the PDB "HET" code and the corresponding PDB entries are reported.

	# of cases with MCC >= 0.3 in top cluster	median MCC	# of cases with MCC >= 0.3 in top three cluster	median MCC	# of cases with MCC >= 0.3 in top five cluster	median MCC
CTP	12/18 (67%)	0.71	17/18 (94%)	0.73	17/18 (94%)	0.74
GTP	30/43 (70%)	0.72	41/43 (95%)	0.71	41/43 (95%)	0.71
TTP	5/14 (36%)	0.61	7 /14 (50%)	0.78	11/14 (79%)	0.78

Table S5. Performance of SiteHound on other three phosphorylated nucleotide datasets. scPDB-2011 was screened for entries containing CTP (20 entries), GTP (73 entries), and TTP (23 entries). Only the entries with a binding site composed by only one chain were retained, obtaining 19 entries for CTP, 57 for GTP and 18 for TTP. The chains were further clustered at 50% sequence identity, obtaining 18 entries for CTP, 43 for GTP and 14 for TTP.