SUPPORTING INFORMATION

ProSelection: A Novel Algorithm to Select Proper Protein Structure Subsets for in Silico Target Identification and Drug Discovery Research

Nanyi Wang [†], Lirong Wang[†, *], and Xiang-Qun Xie[†, *]

[†]Department of Pharmaceutical Sciences and Computational Chemical Genomics Screening Center, School of Pharmacy; NIH National Center of Excellence for Computational Drug Abuse Research; Drug Discovery Institute; University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States

***Corresponding Author: Xiang-Qun (Sean) Xie, MBA, Ph.D.**

Professor of Pharmaceutical Sciences/Drug Discovery Institute

Director of CCGS and NIDA CDAR Centers

335 Sutherland Drive, 206 Pavilion, School of Pharmacy

University of Pittsburgh

Pittsburgh, PA15261, USA

412-383-5276 (Phone)

412-383-7436 (Fax)

Email: xix15@pitt.edu

***Co-corresponding Author: Lirong Wang, Ph.D.**

Assistant Professor of Pharmaceutical Sciences, School of Pharmacy

University of Pittsburgh

517 Salk Hall, 3501 Terrace Street

Pittsburgh, PA15261, USA

412-624-8118 (Phone)

Email: liw30@pitt.edu

SUPPLEMENTARY

TABLE

Table S1. Number of available crystal structures for 19 protein targets (data from RCSB Protein Data Bank).

Protein name	Number of crystal structures
ABL1_HUMAN	19
PIM2_HUMAN	2
MTOR_HUMAN	6
AKT3_HUMAN	1
AKT2_HUMAN	16
AKT1_HUMAN	12
PK3CA_HUMAN	17
SIR2_HUMAN	19
SIR1_HUMAN	8
LRRK2_HUMAN	2
INSR_HUMAN	18
HDAC6_HUMAN	4
CATM_HUMAN	3
CATD_HUMAN	6
CATB_HUMAN	11
CASP1_HUMAN	27
B2CL1_HUMAN	33
BCL2_HUMAN	11
PDPK1_HUMAN	53

The result shown above is a searching result of November 2016. The number of PDB structures available online is ever changing. Abbreviations: PDPK1, 3-phosphoinositide-dependent protein kinase 1; B2CL1, Bcl-2-like protein 1; CASP1, Caspase-1; CATB, Cathepsin B; CATD, Cathepsin D; HDAC6, Histone deacetylase 6;

INSR, Insulin receptor; SIR1, NAD-dependent protein deacetylase sirtuin-1; SIR2, NAD-dependent protein deacetylase sirtuin-2; PK3CA, Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform; AKT1, RAC-alpha serine/threonine-protein kinase; AKT2, RAC-beta serine/threonine-protein kinase; MTOR, Serine/threonine-protein kinase mTOR; ABL1, Tyrosine-protein kinase ABL1.

Table S2. Suggested docking score threshold for active ligands (SDA) of 142 "strong selector" protein structures.

PDB ID_Chain	Target	Threshold	TPR	FPR	PDB ID_Chain	Target	Threshold	TPR	FPR
1csb_B	CATB	6.51	0.11	0.04	3rwp_A	PDPK1	8.46	0.23	0.03
1gag_A	INSR	6.37	0.15	0.04	3rwq_A	PDPK1	8.39	0.20	0.03
1huc_B	CATB	6.57	0.15	0.04	3sc1_A	PDPK1	8.76	0.10	0.03
1i44_A	INSR	6.54	0.21	0.04	3spf_A	B2CL1	6.27	0.05	0.03
1ir3_A	INSR	6.50	0.16	0.04	3zim_A	PK3CA	8.44	0.13	0.04
1irk_A	INSR	6.09	0.22	0.04	3zln_A	B2CL1	9.59	0.56	0.03
1lya_B	CATD	5.30	0.49	0.04	3zlo_A	B2CL1	8.25	0.44	0.04
1lyb_B	CATD	5.67	0.35	0.04	4a06_A	PDPK1	7.61	0.21	0.03
1lyw_B	CATD	5.31	0.31	0.04	4a07_A	PDPK1	7.76	0.28	0.03
1oky_A	PDPK1	8.82	0.15	0.03	4a1u_A	B2CL1	6.40	0.14	0.03
1okz_A	PDPK1	9.23	0.13	0.03	4a1w_A	B2CL1	6.76	0.12	0.04
1p14_A	INSR	6.70	0.19	0.04	4aw0_A	PDPK1	7.34	0.27	0.03
1rqq_A	INSR	6.20	0.18	0.04	4aw1_A	PDPK1	8.99	0.14	0.03
1rwk_A	CASP1	5.04	0.26	0.04	4bpk_A	B2CL1	5.92	0.08	0.04
1uu3_A	PDPK1	7.96	0.20	0.03	4c52_A	B2CL1	5.00	0.20	0.04
1uu7_A	PDPK1	7.85	0.22	0.03	4c5d_A	B2CL1	5.31	0.10	0.02
1uvr_A	PDPK1	7.40	0.36	0.03	4cin_A	B2CL1	4.92	0.15	0.04
2auh_A	INSR	6.91	0.12	0.04	4ct1_A	PDPK1	7.82	0.16	0.03
2b4s_B	INSR	7.62	0.07	0.04	4ehr_A	B2CL1	5.20	0.12	0.04
2biy_A	PDPK1	7.86	0.22	0.03	4ejn_A	AKT1	5.98	0.16	0.05
2hbz_A	CASP1	5.23	0.22	0.04	4ekk_A	AKT1	8.11	0.19	0.05
2ipp_B	CATB	6.68	0.12	0.03	4ekl_A	AKT1	8.98	0.17	0.05
2pe0_A	PDPK1	8.21	0.14	0.03	4gv1_A	AKT1	7.59	0.27	0.05
2pe2_A	PDPK1	8.44	0.11	0.03	4ibm_A	INSR	6.32	0.27	0.04
2r7b_A	PDPK1	8.51	0.14	0.03	4jps_A	PK3CA	7.93	0.16	0.04

2rd0_A	PK3CA	6.74	0.14	0.04	4jsn_A	MTOR	8.78	0.06	0.04
2x39_A	AKT2	8.06	0.29	0.04	4jsp_A	MTOR	7.00	0.17	0.04
2xch_A	PDPK1	9.14	0.06	0.03	4jsv_A	MTOR	7.04	0.16	0.04
2xck_A	PDPK1	8.46	0.13	0.03	4jsx_A	MTOR	8.05	0.17	0.04
2yj1_A	B2CL1	5.66	0.15	0.03	4jt5_A	MTOR	8.42	0.12	0.04
2yq6_A	B2CL1	5.14	0.10	0.03	4jt6_A	MTOR	7.89	0.17	0.04
2yxj_A	B2CL1	5.42	0.19	0.04	4l1b_A	PK3CA	7.25	0.12	0.04
2z8c_A	INSR	6.34	0.22	0.04	4l23_A	PK3CA	8.24	0.13	0.04
3ai8_B	CATB	6.87	0.05	0.04	4l3o_A	SIR2	8.88	0.03	0.04
3bu3_A	INSR	6.20	0.25	0.04	4obz_B	CATD	6.16	0.30	0.04
3bu5_A	INSR	6.12	0.19	0.04	4oc6_B	CATD	5.38	0.48	0.04
3bu6_A	INSR	6.77	0.17	0.04	4od9_B	CATD	5.65	0.39	0.04
3cbj_A	CATB	7.50	0.14	0.04	4ovu_A	PK3CA	6.94	0.20	0.04
3cbk_A	CATB	7.55	0.09	0.04	4ovv_A	PK3CA	7.05	0.10	0.04
3cqu_A	AKT1	8.80	0.13	0.05	4ppi_A	B2CL1	3.76	0.25	0.04
3cqw_A	AKT1	8.62	0.11	0.05	4qve_A	B2CL1	7.62	0.07	0.03
3ekk_A	INSR	6.55	0.16	0.04	4qvf_A	B2CL1	5.61	0.14	0.03
3ekn_A	INSR	7.54	0.08	0.04	4qvx_A	B2CL1	8.82	0.46	0.04
3eta_A	INSR	6.75	0.21	0.04	4rqk_A	PDPK1	8.31	0.17	0.03
3h9o_A	PDPK1	8.53	0.23	0.03	4rqv_A	PDPK1	8.34	0.11	0.03
3hhm_A	PK3CA	7.88	0.08	0.04	4rrv_A	PDPK1	7.57	0.20	0.03
3hiz_A	PK3CA	6.67	0.06	0.04	4tuh_A	B2CL1	9.22	0.26	0.03
3hrc_A	PDPK1	7.63	0.28	0.03	4tuu_A	PK3CA	6.93	0.14	0.04
3hrf_A	PDPK1	8.17	0.16	0.03	4tv3_A	PK3CA	7.87	0.15	0.04
3ion_A	PDPK1	8.14	0.32	0.03	4waf_A	PK3CA	7.47	0.13	0.03
3iop_A	PDPK1	8.56	0.22	0.03	4xlv_A	INSR	5.93	0.30	0.04
3nay_A	PDPK1	8.99	0.11	0.03	4xx9_A	PDPK1	8.49	0.08	0.03
3o96_A	AKT1	6.40	0.17	0.05	4z9v_A	B2CL1	5.68	0.10	0.03
3orx_A	PDPK1	8.29	0.16	0.03	4zop_A	PK3CA	8.42	0.08	0.04
3orz_A	PDPK1	8.47	0.22	0.03	4zzi_A	PK3CA	9.30	0.46	0.04
3otu_A	PDPK1	8.16	0.21	0.03	5ack_A	PDPK1	7.21	0.42	0.03
3ow4_A	AKT1	7.06	0.31	0.05	5b8d_A	HDAC6	7.04	0.07	0.04
3pwy_A	PDPK1	7.63	0.20	0.03	5c3g_A	B2CL1	6.28	0.10	0.02
3qcq_A	PDPK1	8.57	0.17	0.03	5dxh_A	PK3CA	7.15	0.18	0.04
3qcs_A	PDPK1	8.59	0.17	0.03	5dxt_A	PK3CA	8.88	0.12	0.04
3qcx_A	PDPK1	8.64	0.20	0.03	5e1s_A	INSR	7.73	0.12	0.03
3qcy_A	PDPK1	9.94	0.06	0.03	5fi4_A	PK3CA	7.84	0.14	0.04
3qd0_A	PDPK1	9.21	0.12	0.03	5fmj_A	PK3CA	6.80	0.17	0.04
3qd3_A	PDPK1	9.23	0.09	0.03	5fmk_A	B2CL1	5.52	0.25	0.04
3qd4_A	PDPK1	9.72	0.10	0.00	5hkm_A	PDPK1	8.50	0.17	0.00
3qkd_A	B2CL1	5.99	0.10	0.04	5ho7_A	PDPK1	8.16	0.07	0.03

3qkk_A	AKT1	8.19	0.07	0.05	5ho8_A	PDPK1	10.05	0.15	0.03
3qkl_A	AKT1	7.70	0.23	0.05	5kcv_A	AKT1	5.71	0.21	0.05
3qkm_A	AKT1	8.30	0.20	0.05	5kh3_A	HDAC6	6.92	0.15	0.04
3r85_A	B2CL1	5.20	0.10	0.04	5kh7_A	HDAC6	6.66	0.15	0.04
3rcj_A	PDPK1	7.87	0.33	0.03	5kh9_A	HDAC6	7.03	0.06	0.04

Abbreviations: TPR, true positive rate; FPR, false positive rate. PDPK1, 3-phosphoinositide-dependent protein kinase 1; B2CL1, Bcl-2-like protein 1; CASP1, Caspase-1; CATB, Cathepsin B; CATD, Cathepsin D; HDAC6, Histone deacetylase 6; INSR, Insulin receptor; SIR1, NAD-dependent protein deacetylase sirtuin-1; SIR2, NAD-dependent protein deacetylase sirtuin-2; PK3CA, Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform; AKT1, RAC-alpha serine/threonine-protein kinase; AKT2, RAC-beta serine/threonine-protein kinase; MTOR, Serine/threonine-protein kinase mTOR; ABL1, Tyrosine-protein kinase ABL1.

FIGURE

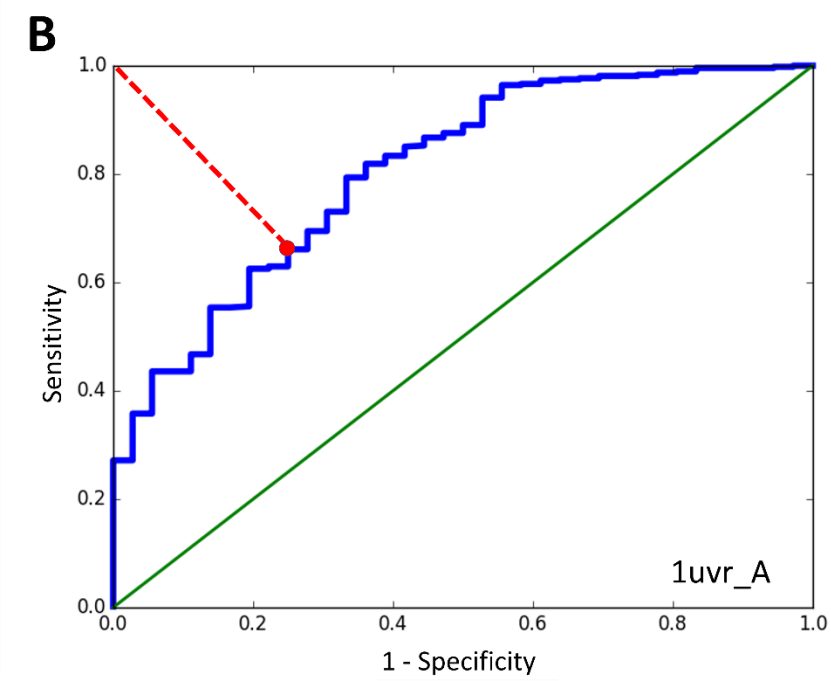

Figure S1. SDA generation. Optimal cutoff point calculation for PDB protein structure 1uvr. Optimal cutoff point is calculated for each ROC curve using Pathagoras' theorem, which minimizing the distance (red dash line) between the top-left corner in the ROC plot and the selected points on the ROC. we require the $FPR^3 \leqslant 0.0001$ for each point to be considered as a "valid point" before using the Pathagoras' theorem. The docking score threshold corresponding to the Optimal cutoff point is named the Suggested Docking Score Threshold for Active Ligands (SDA). For structure 1uvr, the red dot in the figure is selected as the optimized point. It's corresponding docking score is 7.40 and the false positive rate (FPR) is 0.04. This point has the lowest distance (optimized sensitivity and specificity) to the top-left corner among all the points on the ROC curve which have $FPR \leqslant 0.0001$.

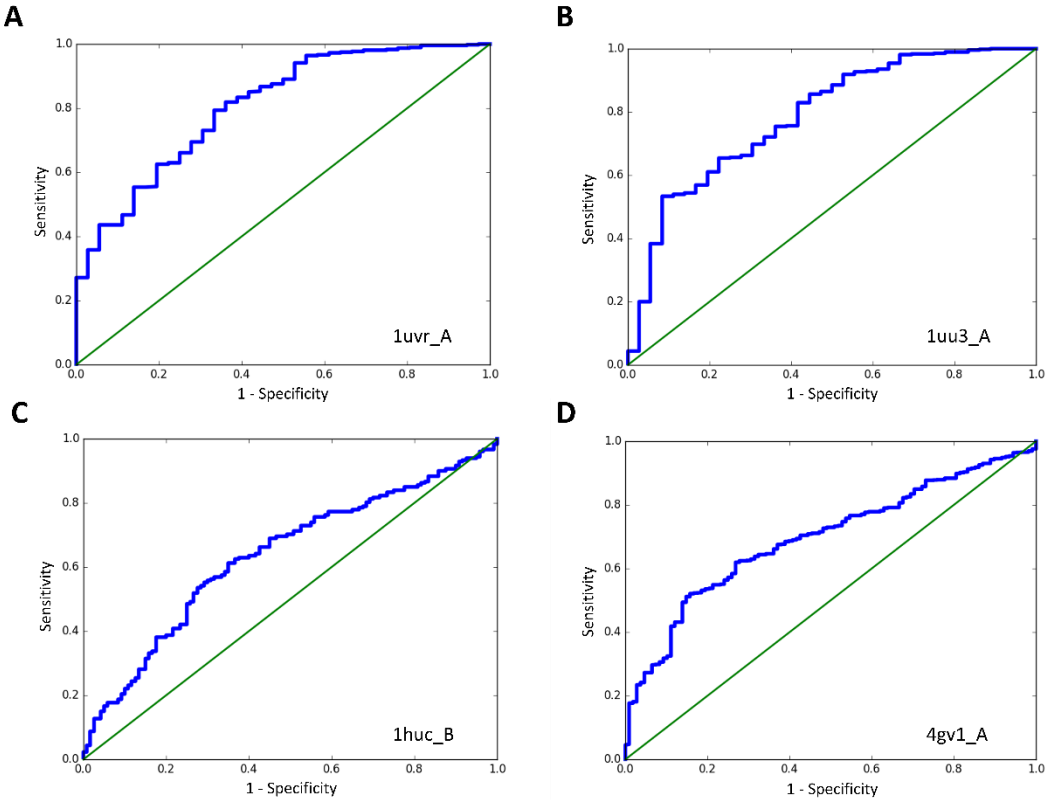

Figure S2. ROC curve of 4 "strong selector" structures. The receiver operating characteristic (ROC) curve of one of the PDB structures of (A) 3-phosphoinositide-dependent protein kinase 1 (PDBid 1uvr, chain A, SDA = 7.40). (B) 3-phosphoinositide-dependent protein kinase 1 (PDBid 1uu3, chain A, SDA = 7.96). (C) Cathepsin B (PDBid 1huc, chain B, SDA = 6.57). (D) RAC-alpha serine/threonine-protein kinase (PDBid 4gv1, chain A, SDA = 7.59).

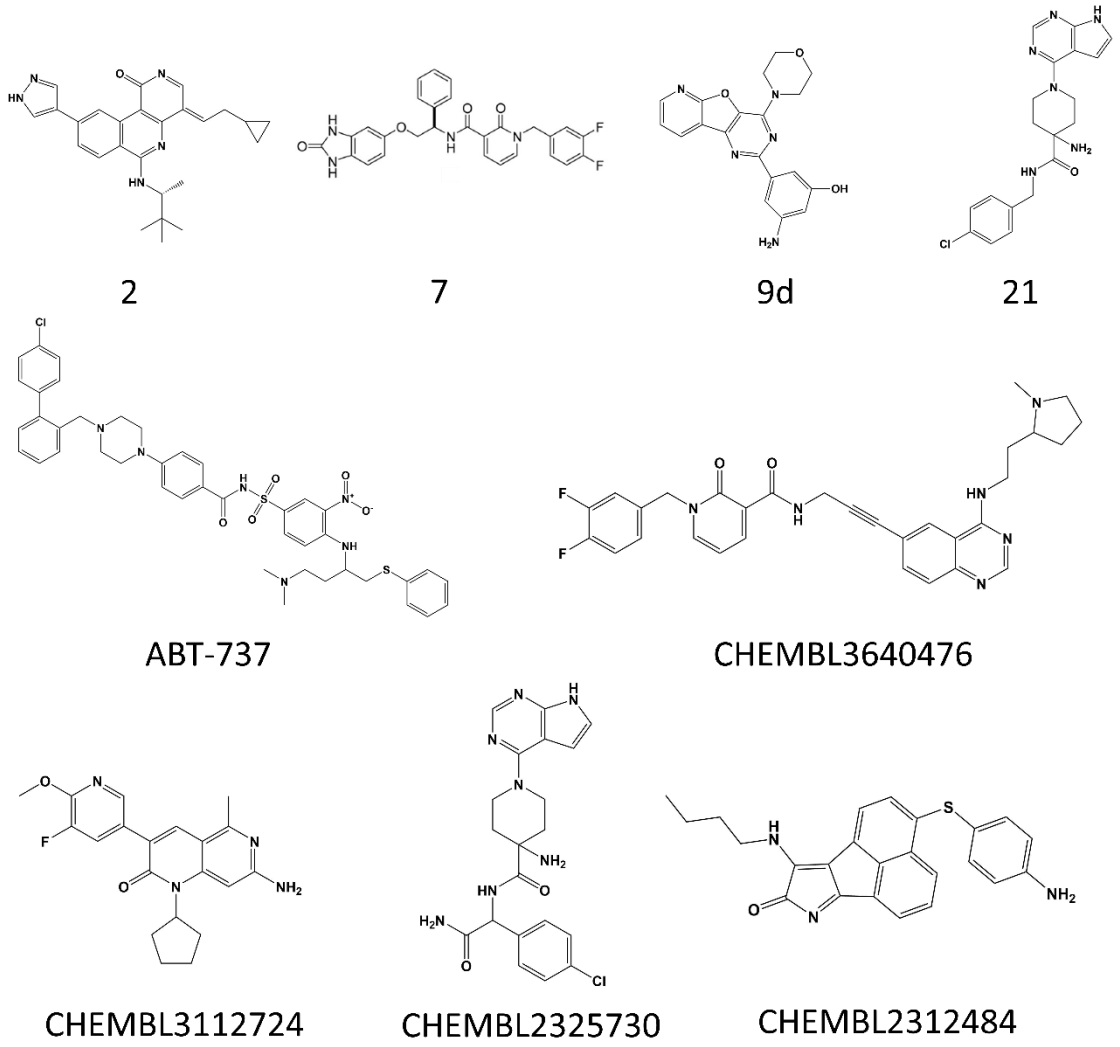

Figure S3. Structures of the inhibitors. Chemical structures of five protein co-crystalized inhibitors and four known protein inhibitors. Compound 2, compound 7, compound 9d, and compound 21 are named using the code number in their original paper. ABT-737 is a known Bcl-xL inhibitor. CHEMBL3640476 is a known PDPK1 ATP competitive inhibitor. CHEMBL3112724 is a known PI3Kα ATP competitive inhibitor. CHEMBL2325730 is a known AKT2 ATP competitive inhibitor. CHEMBL2312484 is a BH3 mimics which inhibit the Bcl-xL.

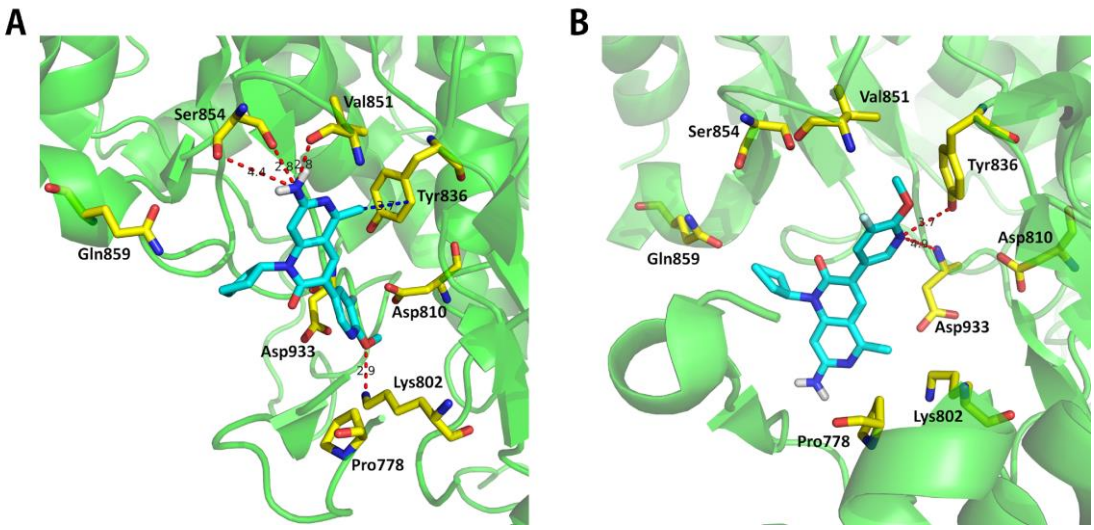

Figure S4. Structures of the compound CHEMBL3112724 bind to the kinase domain in the PI3Kα. (A) CHEMBL3112724 binds to the PI3Kα crystal structure 4L23 (Docking score 9.64). (B) CHEMBL3112724 binds to the PI3Kα crystal structure 4L2Y (Docking score 5.87). Blue sticks represent the inhibitor compound CHEMBL3640476. Yellow sticks represent the important residues which are also labeled. H-bonds to the important residues are displayed as dashed red lines. 4L2Y and 4L23 are derived from the X-ray structure of the protein kinase PI3Kα. Structure 4L2Y is co-crystallized with compound 9d (Supplementary figure 3), a PI103 derivative and known inhibitor of PI3Kα[36]. 4L23 is the native complex of PI103 to PI3Kα. Compound 9d would cause a decreased potency due to the electrostatic repulsion between the incoming NH2 substitute and Lys802[36]. However, in the in vitro experiment, compound 9d was as potent as PI103 against the PI3Kα due to the extra flexibility of Lys802 given by compound 9d. This flexibility also induces additional space at the catalytic site when it is compared to the original protein structure like 4L23[36]. A known PI3Kα ATP competitive inhibitor, CHEMBL3112724, was used to dock both 4L2Y and 4L23 in our research. In 4L23, CHEMBL3112724 has H-bond to Lys802, Val851, and two H-bonds to Ser854. However, in 4L2Y, only two H-bonds are formed between CHEMBL3112724 and PI3Kα. In our structure selection, 4L2Y is the only structure for PI3Kα which is labeled as a "weak selector" ($p = 0.57$) while 4L23 is labeled as a "strong selector" structure ($p = 3.6\times10^{-5}$). The high p-value of 4L2Y may be caused by the structural change of this protein because

of the co-crystallized compound 9d. As is described before, compound 9d induces a more flexible Lys802, which is one of the most important residues inside the docking pocket. Consistent with the failure of molecular docking to predict the experimental potency of 4L2Y in Zhang's group[36], the disability of our selection procedure to label this structure as a "strong selector" is not surprising.

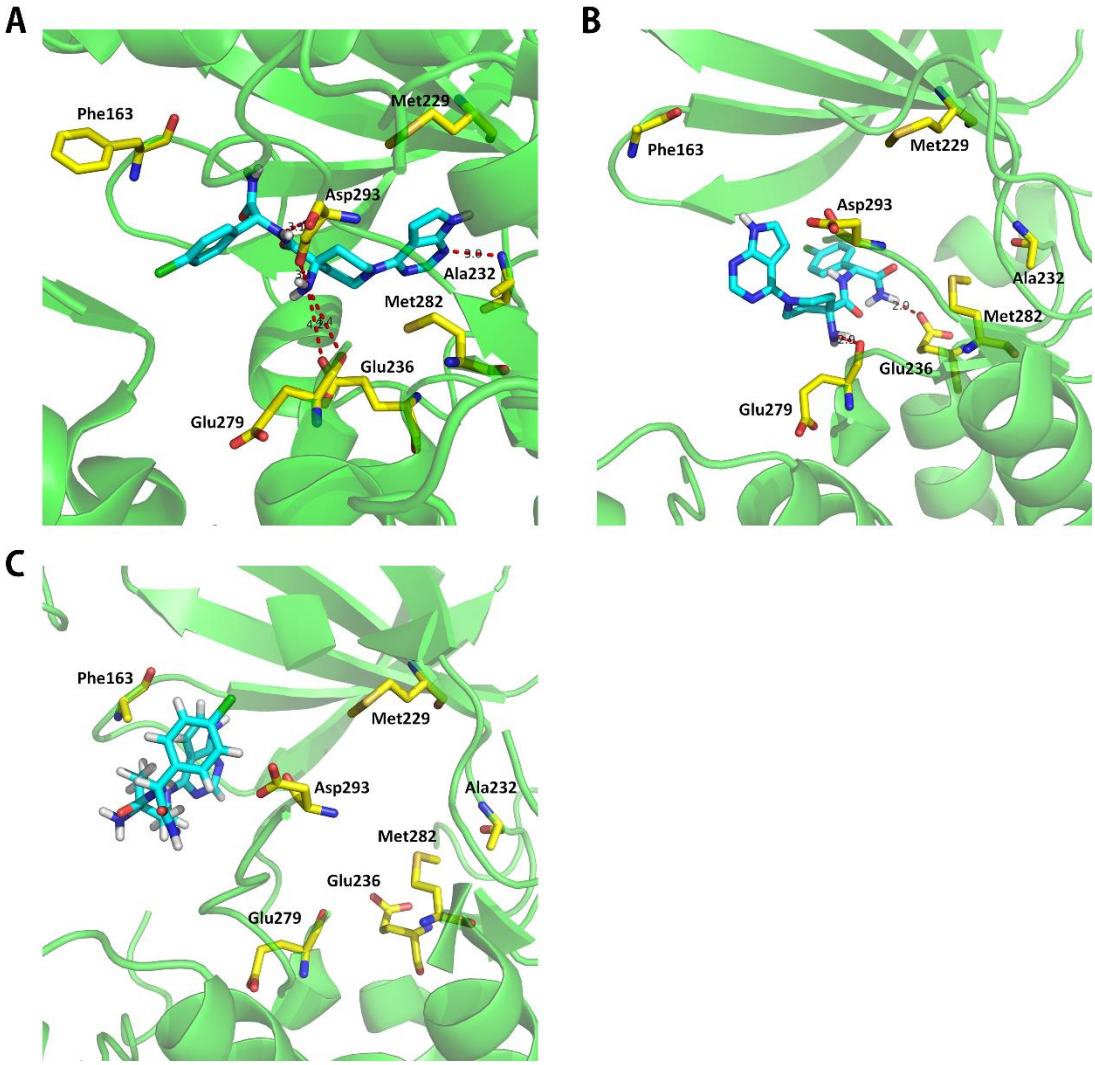

Figure S5. Structures of the compound CHEMBL2325730 bind to the ATP binding site within the kinase domain in AKT2. (A) CHEMBL2325730 binds to the AKT2 crystal structure 2X39 (Docking score 9.11). (B) CHEMBL2325730 binds to the AKT2 crystal structure 1MRV (Docking score 3.53). (C) CHEMBL2325730 binds to the AKT2 crystal structure 1MRY (Docking score 2.66). Blue sticks represent

the inhibitor compound CHEMBL2325730. Yellow sticks represent the important residues which are also labeled. H-bonds to the important residues are displayed as dashed red lines. 2X39, 1MRY, and 1MRV are structures of protein kinase B (PKB or AKT2). Structure 2X39 is the active AKT2 co-crystallized with the known inhibitor compound 21[37] (Supplementary figure 3). 1MRY and 1MRV are crystal structures of an inactive AKT2 kinase domain[38]. In the structure selection, 2X39 is a "strong selector" structure ($p = 0.012$) while 1MRY ($p = 1$) and 1MRV ($p = 1$) are "weak selector" structures. In the docking study, known AKT2 ATP competitive inhibitor CHEMBL2325730 has H-bonds to Ala232, Glu236, Glu279 and Asp293. However, in 1MRV, only two H-bonds are formed between CHEMBL2325730 and AKT2. No H-bond is formed in 1MRY. Structure 1MRV even has higher docking scores for inactive ligands when compared to active ligands. Although the reason behind this "reversed" distribution of docking scores remains unknown, ProSelection has proved once again to be able to distinguish active protein conformations from inactive conformations.