Supporting Information

Computational Prediction and Analysis for Tyrosine Post-Translational Modifications via Elastic Net

Man Cao†, *Guodong Chen*†, *Lina Wang*‡, *Pingping Wen*‡, *Shaoping Shi*†,*

† Department of Mathematics and Numerical Simulation and High-Performance Computing Laboratory, School of Sciences, Nanchang University, Nanchang 330031, China

‡ College of Chemistry, Nanchang University, Nanchang 330031, China

* To whom correspondence should be addressed: shishaoping@ncu.edu.cn

Table of Content

Table S6. The tyrosine PTM optimization parameter with elastic net.

Table S7. The comparison elastic net with other feature selection methods.

Table S8. Comparison of the prediction performance of independent test between our method and other tools in single kinase.

Table S9. Comparison of the prediction performance of independent test between our method and other tools in kinase family.

1. Supplemental Illustration

AAC

Amino acid composition feature is the most popular coding method and widely used for prediction PTMs sites, which reflects protein sequences amino acid occurrence frequencies information. In this work, we calculated the amino acid frequencies in the sequence surrounding the query site (the site itself is not counted). There are 20 types of amino acids, and thus 20 frequencies are calculated, the sum of which is 1. For a protein sequence fragment n, let $p_n(i)$ represents the occurrence times of the i-th amino acid in the protein sequence fragment n. Thus, the occurrence frequencies $f_n(i)$ is calculated by

$$f_n(i) = \frac{p_n(i)}{2*L}$$

Where L represents the number of up-stream or down-stream amino acids flanking each side of the target tyrosine.

BE

The BE method can reflect the type and position information of the amino acid

residues in protein sequence. BE is used an orthogonal binary coding scheme to transform each amino acid into a 20-dimensional binary vector. Herein, we added a vector 'O' to represent other specific amino acid (e.g., B, Z, and X). Thus, there are 21 amino acids ordered as ACDEFGHIKLMNPQRSTVWYO. Briefly, each amino acid is represented by a 21-dimensional binary vector. For example, amino acid A expressed as 100000000000000000000, Y as 000000000000000000010, and so on. Therefore, if the length of a protein sequence is n, the dimension of the numeric vector is 21*n. For example, in this tyrosine nitration work, the length of a protein sequence is 15 and the final dimension of binary encoding vector is 21*15=315.

K-spaced

Additionally, K-spaced could reflect the characteristics of the residues surrounding modification sites, and it has been successfully used for predicting phosphorylation sites. Therefore, we took into account K-spaced amino acid pair compositions of the tyrosine modification sequence to convert these training sets into numerical series. The K-spaced feature encoding were considered the amino acid pairs that separated by K other amino acids within a protein sequence fragment (k is a natural numbers). Generally, we would add a vector 'O' when the residues are not enough or to represent other specific amino acid (e.g., B, Z, and X). Therefore, there are 441 possible amino acid pair types, (e.g. AA, AC, AD ... AO ... OO). For instance, for k=0, there are 441 0-spaced residue pairs, a feature vector can be defined as

$$(N_{AA}, N_{AC}, N_{AD}, \cdots, N_{OO})_{441}$$

The value of each feature denotes the number of occurrences of the corresponding

residue pair in the fragment. In this work, $k = 0,1,2,3,4$ were jointly considered, so the total dimension of the proposed feature vector is 441*5=2205.

PWAA

To avoid losing the sequence-order information, we presented a PWAA to extract the sequence order information of amino acid residues around nitrotyrosine sites, sulfated tyrosine sites and kinase-specific tyrosine phosphorylation sites. Given an amino acid residuea_i $(i = 1,2,\cdots,20)$, we can express the position information of amino acid a_i in the protein sequence fragment P with $2 * L + 1$ amino acids by the following formula:

$$C_i = \frac{1}{L(L+1)} \sum_{j=-L}^{L} X_{i,j} \left(j + \frac{|j|}{L}\right)$$

Where L denotes the number of upstream residues or downstream residues from the central site in the protein sequence fragment P, $X_{i,j} = 1$ if a_i is the j-th position residue in protein sequence fragment P, otherwise $X_{i,j} = 0$ $(j = -L,\cdots,0,\cdots,L)$. In general, residue a_i is closer to the central site (0 position), the absolute value of C_i is smaller.

EBGW

In the previous work, we found that prediction model of tyrosine sulfation achieved a better performance only by using the EBGW encode feature vector. Based on that, we adopted this encoding scheme of the amino acid sequence considering the hydrophobicity and charged character of amino acid residues. The encoding method based on grouped weight is effective in representing the protein physicochemical properties information from protein sequences, which divides the 20 amino acid

residues into four different classes on the basis of their hydrophobicity and charged character. The four groups as follows:

$$\begin{cases} The\ hydrophobic\ group: C1 = \{A, F, G, I, L, M, P, V, W\} \\ The\ polar\ group: C2 = \{C, N, Q, S, T, Y\} \\ The\ positively\ charged\ group: C3 = \{K, H, R\} \\ The\ negatively\ charged\ group: C4 = \{D, E\} \end{cases}$$

So we can divide the amino acid residues into the following disjoint groups: $C1 + C2$ versus $C3 + C4$, $C1 + C3$ versus $C2 + C4$, and $C1 + C4$ versus $C2 + C3$.

For a given protein p, we calculate three binary sequences:

$$H_1(p_j) = \begin{cases} 1 & if\ p_j \in C_1 + C_2 \\ 0 & if\ p_j \in C_3 + C_4 \end{cases}$$
$$H_2(p_j) = \begin{cases} 1 & if\ p_j \in C_1 + C_3 \\ 0 & if\ p_j \in C_2 + C_4 \end{cases}$$
$$H_3(p_j) = \begin{cases} 1 & if\ p_j \in C_1 + C_4 \\ 0 & if\ p_j \in C_2 + C_3 \end{cases}$$

We divide each binary sequence into J sub-sequences increasing in length. For example, for H_1, the feature value of the j-th sub-sequence is defined as:

$$X_1(j) = \frac{Sum(j)}{D(j)} \qquad j = 1,2,\cdots,J$$
$$D(j) = Int\left(\frac{j*L}{J}\right)$$

Where the function $Sum(j)$ gives the number of 1 in the j-th sub-sequence, $D(j)$ denotes the length of the j-th sub-sequence, the $Int()$ rounds a number to the nearest integer and L is the length of the protein p. So, we create J features for H_1, H_2, H_3 respectively and then we concatenate these three vectors. That is to say, we can transform a protein sequence into a $3J$-dimension vector

$X = [X_1, X_2, X_3] = [X_1(1), \cdots, X_1(J), X_2(1), \cdots, X_2(J), X_3(1), \cdots, X_3(J)]$

We name X as the EBGW string of protein sequence p. Preliminary tests indicated that $J = 5$ was the appropriate number of sub-sequences for predicting tyrosine modification sites.

SVM probability estimates

Chang and Lin[1] discussed the LIBSVM implementation for extending SVM to give probability estimates. Given k classes of data, for any x, the goal is to estimate

$$p_i = P(y = i|x), i = 1 \dots k.$$

Following the setting of the one-against-one (i.e., pairwise) approach for multiclass classification, they first estimate pairwise class probabilities

$$r_{ij} = P(y = i|y = i \; or \; j, x)$$

using an improved implementation of Platt: If $\hat{f}$ is the decision value at x, then we assume

$$r_{ij} \approx \frac{1}{1 + e^{A\hat{f}+B}}$$

where A and B are estimated by minimizing the negative log likelihood of training data.

In addition, Wu et al.[2] used their approaches to acquire $p_i$from all these r_{ij}'s. It solves the following optimization problem:

$$\min_p \frac{1}{2}\sum_{i=1}^{k}\sum_{j:j\neq i}\left(r_{ji}p_i - r_{ij}p_j\right)^2 \quad subject\ to \sum_{i=1}^{k} p_i = 1,\ p_i \geq 0, \forall i$$

REFERENCES

(1) Chang, C. C.; Lin, C. J., LIBSVM: A Library for Support Vector Machines. *Acm T Intel Syst Tec.* **2011**, *2*, 27.

(2) Wu, T. F.; Lin, C. J.; Weng, R. C., Probability Estimates for Multi-class Classification by Pairwise Coupling. *J MACH LEARN RES.* **2004**, *5*, 975-1005.

2. Supplementary Figures

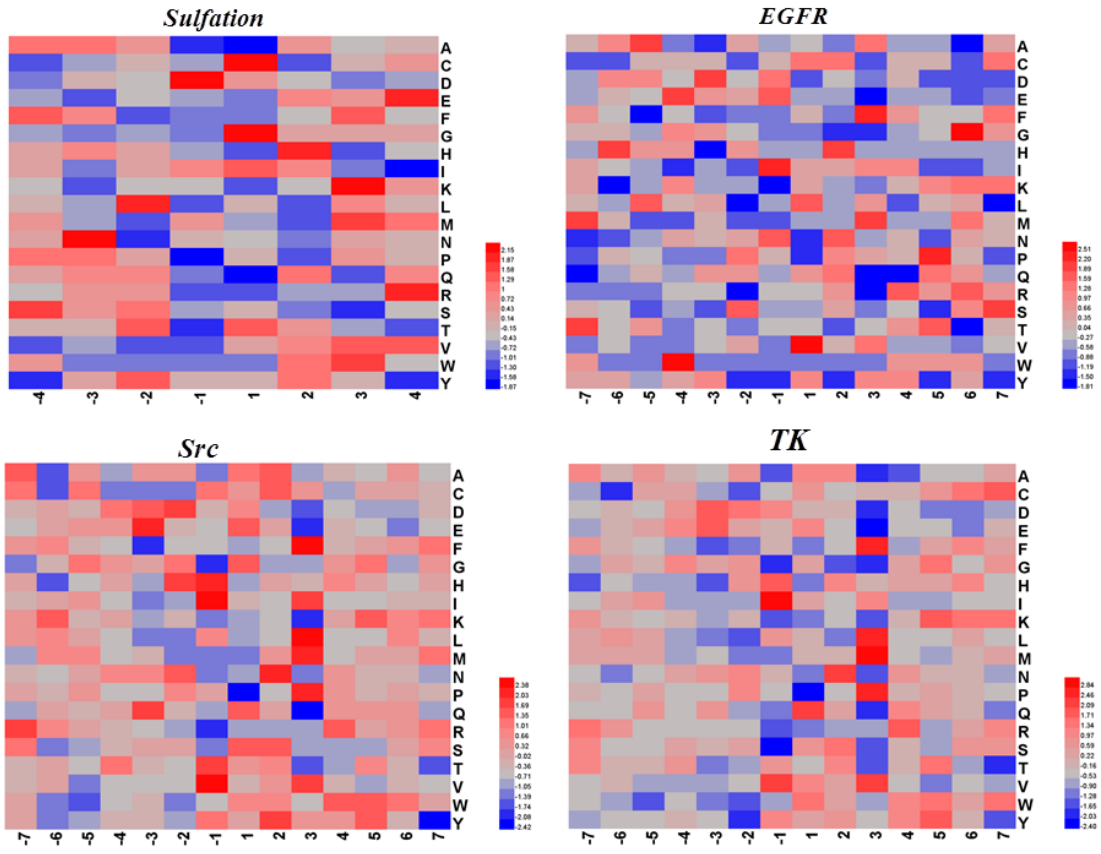

Figure S1. Heat map indicated that position distribution of BE scores for amino acid composition.

Figure S2. Comparisons of AAC in positive and negative datasets. The vertical axis represents the log2 ratio of amino acid frequencies surrounding nitrotyrosine, sulfotyrosine and phosphotyrosine

and non-nitrotyrosine and non-sulfotyrosine and non- phosphotyrosine sites. The horizontal axis represents the 20 amino acids sorted in descending order by the mean log2 ratio tyrosine post-translational modification sequence.

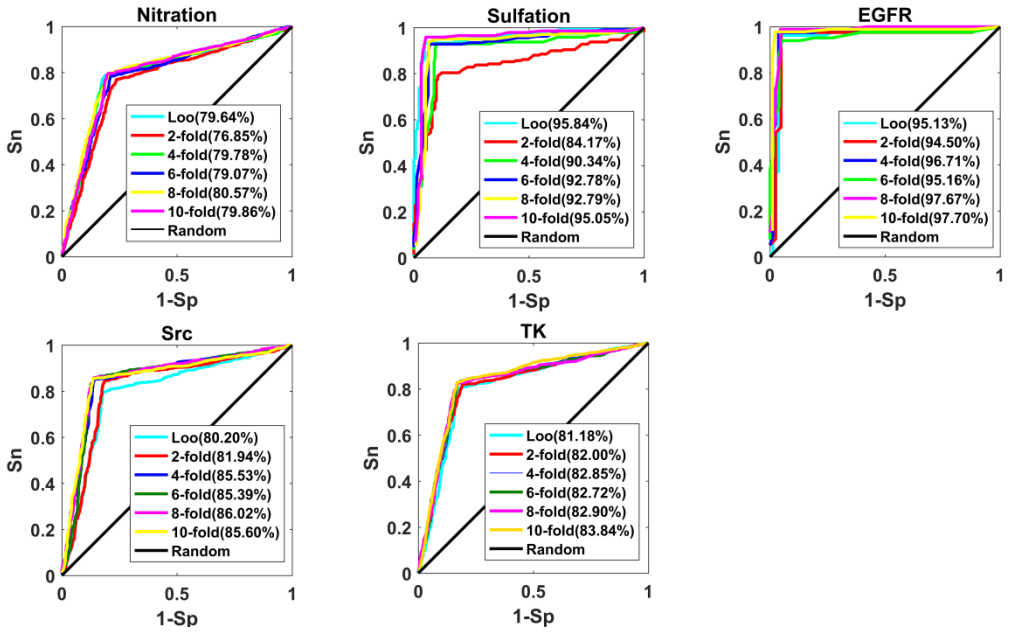

Figure S3. The LOO validation and 2-, 4-, 6-, 8- and 10-fold cross-validations were performed on each data set.

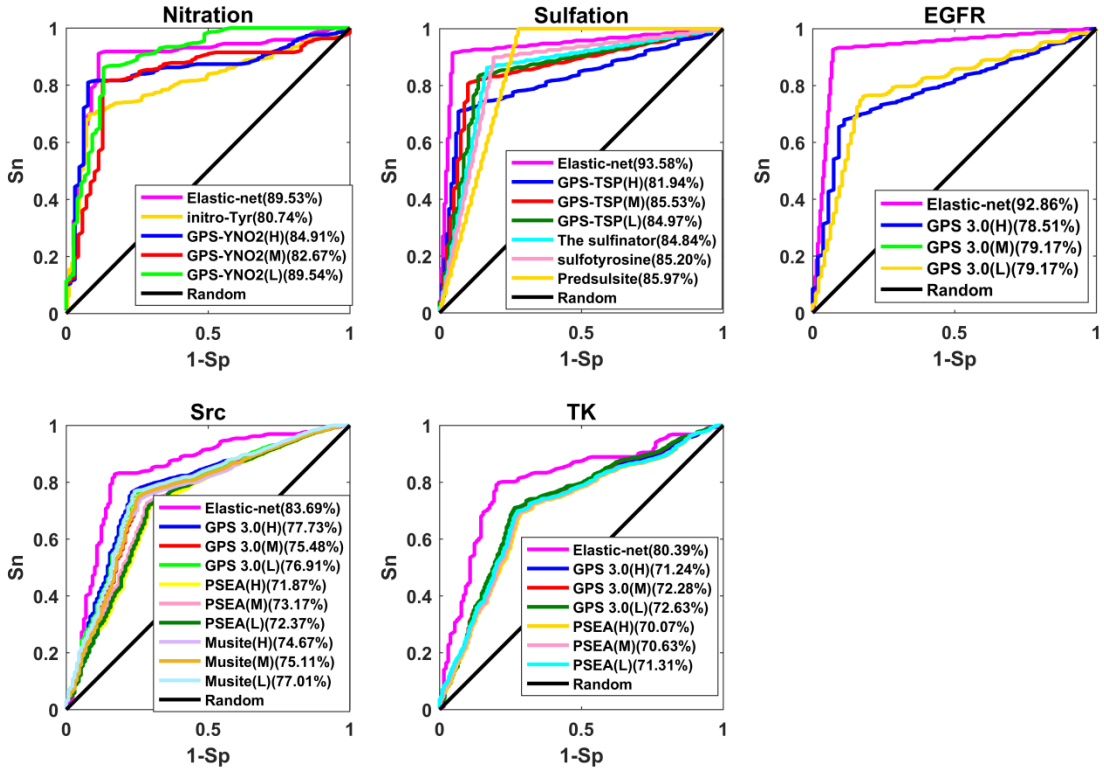

Figure S4. ROC curves of Tyrpred comparison with other tools in prediction of tyrosine nitration,

sulfation, and kinase-specific phosphorylation models, respectively. Each curve represents the average sensitivities and specificities for different thresholds over 10-fold cross-validation.

3. Supplementary Tables

Table S1. The statistics of tyrosine nitration datasets in this study.

Tyrosine modification	eliminate homology before (sites)		eliminate homology after(sites)		Training dataset(sites)		Testing dataset(sites)	
	positive	negative	positive	negative	positive	negative	positive	negative
Nitration	1155	8842	1114	8061	1038	1038	76	76

Table S2. The statistics of tyrosine sulfation datasets in this study.

Tyrosine modification	eliminate homology before (sites)		eliminate homology after(sites)		Training dataset(sites)		Testing dataset(sites)	
	positive	negative	positive	negative	positive	negative	positive	negative
Sulfation	189(365)	1388	90(155)	675	75(132)	132	15(23)	23

Table S3. The statistics of kinase-specific tyrosine phosphorylation datasets in this study.

kinase-specific phosphorylation	eliminate homology before (sites)		eliminate homology after(sites)		Training dataset(sites)		Testing dataset(sites)	
	positive	negative	positive	negative	positive	negative	positive	negative
Single-kinase	--	--	--	--	--	--	--	--
Abl	191	1512	177	1421	149	149	28	28
Lck	120	716	113	689	96	96	17	17
EGFR	105	846	93	765	79	79	14	14
FYN	184	1467	168	1391	141	141	27	27
INSR	78	459	68	378	57	57	11	11
JAK2	69	462	62	384	52	52	10	10
LYN	112	653	113	606	96	96	17	17
Src	660	5291	568	4768	482	482	86	86
Syk	76	432	62	406	52	52	10	10
Kinase-family	--	--	--	--	--	--	--	--
Abl	221	1565	187	1459	158	158	29	29
EGFR	128	1069	113	971	96	96	17	17
InsR	103	596	83	480	70	70	13	13
JakA	110	671	94	596	79	79	15	15
Src	1171	7317	862	6526	732	732	130	130

Syk	110	556	89	485	75	75	14	14
Kinase-group	--	--	--	--	--	--	--	--
TK	2318	12136	1504	10322	1278	1278	226	226

Table S4. Comparison of model performance before and after dimension reduction in tyrosine single-kinase phosphorylation.

Single	**Before**					**After**				
kinase	**Dim**	**Acc(%)**	**Sn(%)**	**Sp(%)**	**Mcc(%)**	**Dim**	**Acc(%)**	**Sn(%)**	**Sp(%)**	**Mcc(%)**
Abl	2581	71.33	70.00	72.67	43.85	169	95.50	97.17	93.83	91.32
FYN	2581	70.88	73.81	67.95	46.27	144	97.17	96.48	97.86	94.43
InsR	2581	68.33	63.33	73.33	42.82	72	98.33	98.33	98.33	96.90
JAK2	2581	61.00	68.00	54.00	24.90	78	96.00	98.00	94.00	92.66
Lck	2581	71.00	59.00	83.00	44.57	135	97.00	98.00	96.00	94.09
LYN	2581	66.00	59.00	73.00	36.00	127	97.50	99.00	96.00	95.14
Src	2581	73.96	75.00	72.92	47.93	276	88.00	88.51	87.48	76.18
Syk	2581	79.57	69.14	90.00	61.68	31	96.00	98.00	94.00	92.66

Table S5. Comparison of model performance before and after dimension reduction in tyrosine Kinase-group phosphorylation.

Kinase	**Before**					**After**				
group	**Dim**	**Acc(%)**	**Sn(%)**	**Sp(%)**	**Mcc(%)**	**Dim**	**Acc(%)**	**Sn(%)**	**Sp(%)**	**Mcc(%)**
Abl	**2581**	78.13	81.25	75.00	56.36	**218**	95.94	97.50	94.38	92.13
EGFR	**2581**	71.50	70.00	73.00	42.62	**131**	97.00	97.00	97.00	94.27
InsR	**2581**	76.43	67.14	85.71	55.13	**119**	98.57	98.57	97.14	97.32
JakA	**2581**	65.09	55.36	74.82	32.88	**121**	96.70	97.14	96.25	93.91
Syk	**2581**	82.08	76.67	87.50	65.81	**62**	98.12	98.75	97.50	96.46

Table S6. The tyrosine PTM optimization parameter with elastic net.

Modification type	λ_1	λ_2	dim
Nitration	0.10	0.12	470
Sulfation	0.10	0.24	144
Single-kinase	--	--	--
Abl	0.10	0.25	169
Lck	0.50	0.20	135
EGFR	0.10	0.20	72
FYN	0.20	0.25	144
InsR	0.30	0.20	72
JAK2	0.50	0.16	78
LYN	0.40	0.20	127
Src	0.30	0.12	276
Syk	0.10	0.16	31
Kinase-family	--	--	--

Abl	0.30	0.24	218
EGFR	0.20	0.24	131
InsR	0.40	0.24	119
JakA	0.50	0.20	121
Src	0.10	0.15	497
Syk	0.10	0.20	62
Kinase-group	--	--	--
TK	0.10	0.10	396

Table S7. The comparison elastic net with other feature selection methods.

Modification type	Method	the performance of prediction			
		Acc(%)	Sn(%)	Sp(%)	Mcc(%)
Nitration	IG	70.07	70.89	69.25	40.22
	F-score	72.05	71.86	72.24	44.20
	mRMR	72.10	72.63	71.57	44.27
	Elastic net	79.67	79.76	79.57	59.40
Sulfation	IG	82.92	81.59	84.26	66.88
	F-score	84.21	84.36	84.05	69.06
	mRMR	89.82	90.97	88.67	80.50
	Elastic net	94.82	94.15	95.49	90.12
Src	IG	78.26	78.65	77.88	56.76
	F-score	79.29	78.39	80.20	58.83
	mRMR	79.36	78.94	79.78	58.96
	Elastic net	85.78	86.46	85.10	71.66

Table S8. Comparison of the prediction performance of independent test between our method and other tools in single kinase.

Single-Kinase	**Method**	**stringency**	**the performance of prediction**			
			Acc(%)	**Sn(%)**	**Sp(%)**	**MCC(%)**
Abl	PSEA	High	88.89	77.78	100.00	79.77
		Medium	87.03	77.78	96.30	75.38
		Low	90.74	88.89	92.59	81.53
	Our work	--	**85.19**	**92.59**	**77.78**	**71.16**
EGFR	GPS	High	75.00	92.86	57.14	53.53
		Medium	57.14	92.86	21.43	20.41
		Low	53.57	92.86	14.29	11.55
	Our work	--	**92.86**	**92.86**	**92.86**	**85.71**
FYN	PSEA	High	83.33	77.78	88.89	67.08
		Medium	83.33	77.78	88.89	67.08
		Low	81.48	81.48	81.48	62.96
	GPS	High	81.48	70.37	92.59	64.58

		Medium	79.63	77.78	78.57	59.30
		Low	77.78	81.48	74.07	55.71
	Our work	--	**81.48**	**88.89**	**74.07**	**63.67**
InsR	GPS	High	86.36	72.72	100.00	75.59
		Medium	86.36	81.81	90.91	73.03
		Low	86.36	90.91	81.81	73.03
	Our work	--	**86.36**	**81.82**	**90.91**	**73.03**
JAK2	GPS	High	59.09	27.27	90.91	23.57
		Medium	63.64	36.36	90.91	32.54
		Low	40.91	63.64	18.18	20.41
	Our work	--	**85.00**	**80.00**	**90.00**	**70.35**
	PSEA	High	76.47	58.82	94.12	56.58
		Medium	79.41	64.71	94.12	61.55
		Low	82.35	70.59	94.12	66.58
Lck	GPS	High	85.29	70.59	100.00	73.85
		Medium	85.29	76.47	94.12	71.71
		Low	91.18	88.24	94.12	82.50
	Our work	--	**85.29**	**82.35**	**88.24**	**70.71**
LYN	PSEA	High	67.65	35.29	100.00	46.29
		Medium	70.59	41.18	100.00	50.91
		Low	67.65	41.18	94.12	41.60
	GPS	High	79.41	64.71	94.12	61.55
		Medium	76.47	70.59	82.35	53.31
		Low	73.53	76.47	70.59	47.14
	Our work	--	**85.29**	**94.12**	**76.47**	**71.71**
Syk	GPS	High	75.00	70.00	80.00	50.25
		Medium	75.00	80.00	70.00	50.25
		Low	75.00	80.00	70.00	50.25
	Our work	--	**85.00**	**90.00**	**80.00**	**70.35**
Src	PSEA	High	73.53	58.82	88.24	49.24
		Medium	71.76	60.00	83.53	44.79
		Low	72.94	62.35	83.53	46.95
	GPS	High	72.94	52.94	92.94	50.06
		Medium	72.35	60.00	84.71	46.14
		Low	72.35	70.59	74.12	44.73
	Musite	High	66.47	35.29	97.65	42.14
		Medium	72.35	52.94	91.76	48.51
		Low	70.59	56.47	84.71	42.92
	Our work	--	**85.88**	**83.53**	**88.24**	**71.84**

Table S9. Comparison of the prediction performance of independent test between our method and other tools in kinase family.

Kinase-family	**Method**	**stringency**	**the performance of prediction**

			Acc(%)	Sn(%)	Sp(%)	MCC(%)
Abl	PSEA	High	78.57	67.86	89.29	58.50
		Medium	78.57	71.43	85.71	57.74
		Low	80.36	78.57	82.14	60.75
	Our work	--	**89.29**	**92.86**	**85.71**	**78.77**
Src	PSEA	High	75.38	76.15	74.62	50.78
		Medium	75.38	77.69	73.08	50.82
		Low	74.23	81.54	66.92	50.00
	GPS	High	80.00	73.08	86.92	60.58
		Medium	80.77	84.12	76.92	61.72
		Low	77.69	90.77	64.62	57.38
	Musite	High	50.77	61.18	94.12	58.56
		Medium	53.46	74.12	89.41	64.29
		Low	55.77	90.59	80.00	70.99
	Our work	--	**82.69**	**83.08**	**82.31**	**65.39**
EGFR	GPS	High	64.71	58.82	70.59	29.62
		Medium	73.53	82.35	64.71	47.81
		Low	67.65	88.24	47.06	38.73
	Our work	--	**91.18**	**94.12**	**88.24**	**82.50**
InsR	GPS	High	61.54	38.46	84.62	26.01
		Medium	61.54	46.15	76.92	24.25
		Low	61.54	46.15	76.92	24.25
	Our work	--	**88.46**	**84.62**	**92.31**	**77.15**
JakA	PSEA	High	56.67	20.00	93.33	19.61
		Medium	60.00	26.67	93.33	26.83
		Low	60.00	33.33	86.67	23.64
	GPS	High	70.00	50.00	80.00	40.82
		Medium	66.67	50.00	73.33	33.63
		Low	66.67	73.33	50.00	33.63
	Our work	--	**80.00**	**86.67**	**73.33**	**60.54**
Syk	PSEA	High	92.86	92.86	92.86	85.71
		Medium	92.86	92.86	92.86	85.71
		Low	82.14	92.86	71.43	65.81
	GPS	High	78.57	85.71	71.43	57.74
		Medium	71.43	85.71	57.14	44.72
		Low	64.29	92.86	35.71	34.82
	Our work	--	**82.14**	**92.86**	**71.43**	**65.81**