SUPPORTING INFORMATION

Searching Post-translational Modifications in Cross-linking Mass Spectrometry Data

Chen Zhou[1,†], Shengzhi Lai[1,†], Shuaijian Dai[1,†], Peize Zhao[2], Ning Li[3,4,⋆] and Weichuan Yu[1,3,⋆]

[1] Department of Electronic and Computer Engineering,
The Hong Kong University of Science and Technology, Hong Kong, China

[2] Interdisciplinary Programs Office,
The Hong Kong University of Science and Technology, Hong Kong, China

[3] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute,
Futian, Shenzhen, China

[4] Guoke-Ningbo Life Science and Health Industry Research Institute, Zhejiang, China

[†] Authors contributed equally to this paper.

[⋆] Correspondence: N.L. (boningli@ust.hk) and W.Y. (eeyu@ust.hk)

Contents

1 Regularized scoring function

Designing a scoring function for identifying post-translational modifications (PTMs) differs from matching regular peptide sequences. It is essential to consider the probability of PTM occurrence within the peptide sequence. Moreover, it is not always true that a better match of peaks leads to a better result.

When there is no penalty for PTM occurrence within the peptide sequence, we have observed that the program often tends to match PTMs even when they do not exist in reality. This is because adding (fake) PTMs to the sequence will result in a higher chance for the mismatched fragmented ions to match the noisy peaks in the spectrum.

To illustrate the extreme case (Fig. S1), let's consider a scenario where we have an incorrect backbone sequence that does not have any theoretical ions matching the peaks in the spectrum. However, by deliberately adding PTMs to each amino acid in this backbone sequence, we can generate new theoretical ions that match the peaks in the spectrum. This is possible if there are no limitations on the choice of PTMs, and their mass can be any real number.

Therefore, we need to add a penalty in the scoring function to restrict the occurrence of the PTMs. Concretely, the scoring function needs to satisfy several requirements:

- As the number of PTMs in the backbone sequence increases, the power should also gradually increase.
- If there are no PTMs on the backbone sequence, there shouldn't be any penalty for the peak matching.
- If all amino acids are assigned PTMs, the score should be 0 regardless of how good the matching result is.
- For peptides with different lengths, the penalty for a single PTM occurrence should vary.

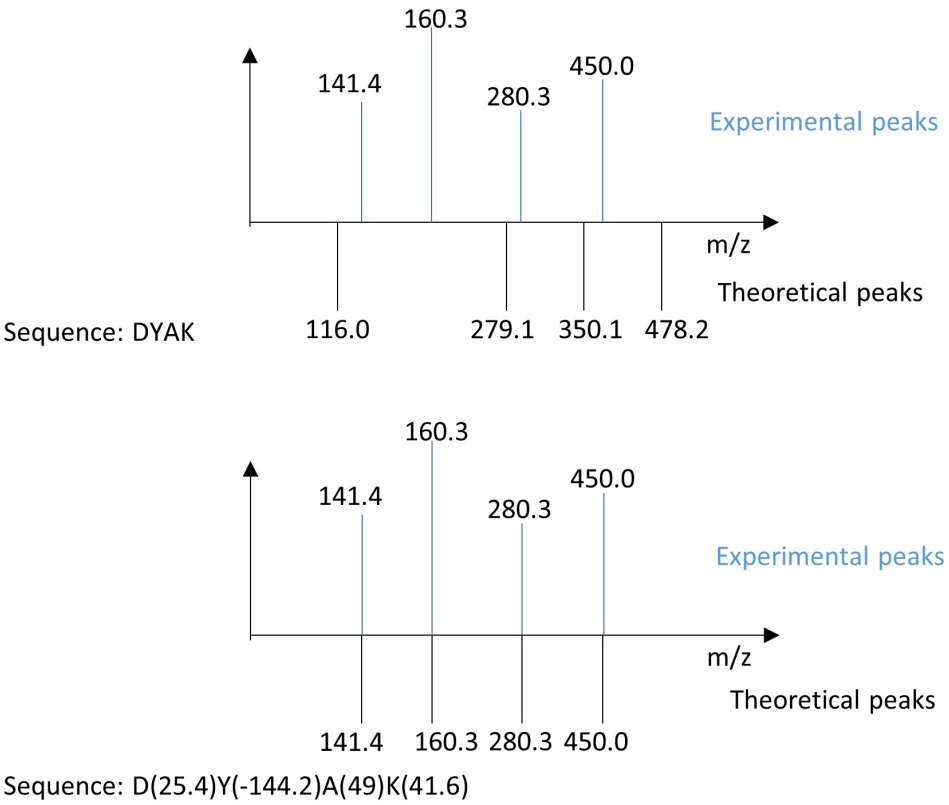

Figure S1: Illustration of the extreme case of fake PTMs on each amino acid to perfectly match the peaks in the spectrum. The sequence 'DYAK' cannot match any experimental peaks in the MS2 spectra. But if we add fake PTMs with masses of 25.4Da, -144.2Da, 49Da, and 41.6Da on amino acids 'D', 'Y', 'A', and 'K', respectively, we can make the theoretical peaks match the experimental ones perfectly.

Longer peptides should incur a higher penalty. The reason for this is that we can consider a scenario where we have two peptide candidates: A, with a sequence length of 100, and B, with a sequence length of 10. Both sequences A and B have one identified PTM that increases their scores compared to their respective unmodified backbones' scores. Since A has 100 possible positions to add a PTM and improve the backbone score, while B only has 10 possible positions, we should penalize A more due to its higher random chance of achieving a better score.

Based on the rationales above, we have designed the regularized Xcorr scoring functions[1]:

$$S_{regularized} = S_{Xcorr} \cdot \left[1 - \left(\frac{\#PTM}{l}\right)^{\frac{\mu_l}{l}}\right] = \vec{e} \cdot \vec{t} \cdot \left[1 - \left(\frac{\#PTM}{l}\right)^{\frac{\mu_l}{l}}\right], \tag{S1}$$

where S_{Xcorr} represents the original Xcorr scoring function, $\vec{e}$ is the digitized experimental peaks, and $\vec{t}$ is the digitized theoretical peaks, $\#PTM$ denotes the number of PTMs, l denotes the length of peptide, and μ_l represents the average length of peptide candidates for the given spectrum.

To demonstrate the properties of the new scoring function, especially for the weight term $\left[1 - \left(\frac{\#PTM}{l}\right)^{\frac{\mu_l}{l}}\right]$, we use some concrete examples and draw a graph.

Suppose we have three peptide candidates for a specific spectrum, with lengths of 5, 10, and 15. Fig. S2 illustrates the curves of the weight changes corresponding to these three lengths. We can observe that the term satisfies our requirements.

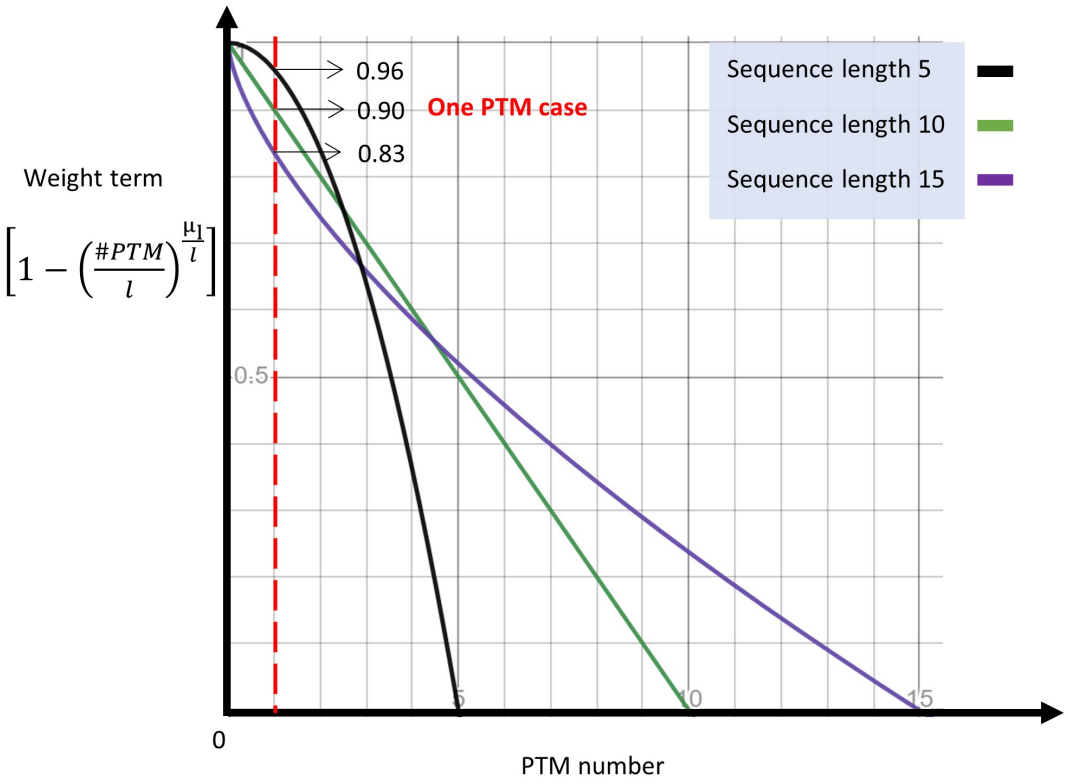

Figure S2: Illustration of the property of the weight term in the scoring function. We generate curves that depict the relationship between the number of PTMs and the weight term in the regularized scoring function. First, all curves are decreasing functions, indicating that as the number of PTMs increases, the weight decreases. Secondly, each curve starts from the point (0,1), indicating that when there are no PTMs (PTM number equals 0), there is no penalty to the score. Third, every curve ends at the point (length, 0), indicating that when all amino acids are assigned PTMs (PTM number equals length), there is a maximum penalty resulting in a weight of 0. Lastly, at the position where the PTM number equals 1, the weight required decreases as the peptide length increases.

2 In vitro chemical cross-linking of bovine serum albumin

To prepare the crosslinked bovine Serum Albumin (BSA) protein, 10 mg of BSA was dissolved in 1 mL of crosslinking buffer containing 40 mM HEPES (pH 8.0), 10% glycerol, and 200 mM NaCl. The crosslinker, CBDPS, was freshly prepared in DMSO to achieve a stock solution of 50 mM. The crosslinker was added to the BSA solution at a concentration of 0.5 μM, and the mixture was subjected to 6 rounds of rotation for 3 minutes per round to achieve a final concentration of 3 μM. The mixture was incubated at room temperature for 1 hour, and the excess crosslinker was quenched by adding 1M Tris-HCl (pH 8.0) to the solution to achieve a final concentration of 100 mM, followed by incubation for 10 minutes at room temperature.

The protein pellet was precipitated by applying 3 volumes (v/v) of pre-cooled 12:1 (v/v) acetone/methanol solution for at least 4 hours at -20 °C. The pellet was air-dried and resuspended in protein resuspension buffer 1 containing 50 mM Tris-HCl pH 8.0, 50 mM NaF, 1% glycerol-2-phosphate, 8 M urea, and 2% glycerol. The solution was treated with 10 mM DTT for 30 minutes, 40 mM IAM for 30 minutes (protected from light), and 10 mM DTT for 10 minutes at room temperature. The solution was then mixed with 3 volumes (v/v) pre-cooled 12:1 (v/v) acetone/methanol solution for at least 4 hours at -20 °C. The protein concentration was quantified using a DC protein assay (Bio-Rad, Hercules, CA, USA).

The protein pellet was dissolved in protein re-suspension buffer 2 containing 40 mM Tris-HCl (pH 8.0) and 6 M urea. The protein solution was diluted with a pre-heated (37 °C) trypsin digestion buffer (40 mM Tris-HCl, pH 8.0) to ensure the final concentration of urea was lower than 1 M. The cross-linked peptides were desalted using C18 Sep-Pak cartridges (Waters, Manchester, UK) and concentrated using SpeedVac (Thermo Scientific Inc., Waltham, MA, USA). The XL-peptides were labeled with light (L) and heavy (H) isotope-coded formaldehyde chemicals, mixed, and desalted using C18 Sep-Pak cartridges.

The CBDPS cross-linked peptides were enriched using high-capacity streptavidin agarose resin

(Pierce, Rockford, IL, USA) and washed with a washing buffer consisting of 50 mM HEPES (pH 7.5). The cross-linked peptides were eluted with 70% ACN and 0.5% FA for 1 hour, and the elution was performed twice. The peptide samples were desalted using Ziptip (MilliporeSigma, Burlington, MA, USA), followed by SpeedVac.

The cross-linked peptides were analyzed using LC-MS/MS. The peptides were separated by a 120-minute gradient elution at a flow rate of 0.3 μL/min with a Thermo-Dionex Ultimate 3000 HPLC system interfaced with a Thermo Orbitrap Fusion Lumos mass spectrometer. The analytical column was the Acclaim PepMapTM RSLC C18 capillary column (75 μm ID, 150 mm length; Thermo Scientific Inc., Waltham, MA, USA). The MS scans were performed in the range of 300 – 2,000 m/z at a resolution of 120 K, and the MS/MS scans were performed at a resolution of 30,000. The 10 most abundant precursors were subjected to a sequential CID-MS/MS and ETD-MS/MS acquisition protocol.

In the CID-MS/MS experiment, the charge state was set as +4 to +8, and the CID normalized collision energy was 30%. For the ETD-MS/MS experiment, the charge-dependent reaction time was enabled, and the Orbitrap resolution was 30 K. For the HCD-MS/MS experiment, the settings were the same as for the CID-MS/MS except the MS/MS activation type was set as HCD and the collision energy was 26%.

3 Synthesized samples from real dataset

The peptides MKTLGR, ALQKSPGPQR, GKWHGDVAVK, KSSSSSEDR, KSS(p)SSSEDRNR, KSS(p)SSSEDR, KSS(p)SSS(p)EDRNR, and GDGGSTTGLSAT(p)PPASLPGSLTNVKALQK, were synthesized by WuXi AppTec (Shanghai, China). Here p stands for phosphorylation. Fmoc-protected N-terminal amine was used to promote the crosslinking between two peptides at the side chains of lysine. The peptide pairs, MKTLGR and KSS(p)SSSEDRNR, ALQKSPGPQR and KSS(p)SSSEDR, MKTLGR and KSS(p)SSS(p)EDRNR, GKWHGDVAVK and KSS(p)SSSEDR, as well as KSSSSSEDR and GDGGSTTGLSAT(p)PPASLPGSLTNVKALQK, were dissolved in phosphate-buffered saline (PBS, pH 7.4) at a concentration of 5mg/mL. Subsequently, 2 mM BS3 crosslinker was added to each peptide mixture, and the incubation was carried out at room temperature for 1 hour. Subsequently, Tris-HCl was added to a final concentration of 20 mM to quench the exceed crosslinker. To remove the N-terminal Fmoc protection, 12.5 μL of piperidine was added to the above-mentioned mixture and the reaction was left to proceed for 2 hours at room temperature. Next, the crosslinked peptide sample was frozen and dried to remove the organic solvents.

The chromatographic enrichment of crosslinked was performed using a HPLC system coupled with a SCX column. Buffer A (7 mM KH2PO4, 30% ACN, pH 3) and Buffer B (7 mM KH2PO4, 30% ACN, 350mM KCl, pH 3) were used as the mobile phases. The separation gradient was set as the following: Buffer A 100% to 90% in 10 min, 90% to 83% in 22 min, 83% to 68% in 8 min, 68% to 20% in 10 min, 20% to 0% in 2 min, 0% to 100% in 12 min and 100% for 30 min. The enriched peptides were collected and further desalted using Ziptip.

The samples were reconstituted in solvent A (0.1% formic acid) and processed using an Easy nLC system with an EASYSPRAY HPLC C18 column (2 µm, 100 Å, 75 µm x 250 mm) at a flow rate of 0.3 µL/min. Peptide separation was achieved over a 60-minute gradient from 2% to 40% solvent B (0.1% formic acid in 80% acetonitrile): from 2% to 30% solvent B (0-53 minutes); from 30% to 40% solvent B (53-59 minutes); and from 40% solvent B to 100% solvent B (59-60 minutes). MS1 resolution was set to 60,000 with a scan range of 350–1500 m/z and a maximum injection time of

50 ms. For MS/MS analysis, the charge state was set to +4 to +8, with a resolution of 15,000 under HCD collision mode, applying a normalized collision energy of 30% for a maximum injection time of 22 ms. The FAIMS mode was configured to 2 CVs (-45 V and -65 V).

4 Linear peptide datasets

Due to the limited availability of PTM-containing cross-linked peptides with known ground truth, we used an alternative approach to validating SeaPIC. We modified the code to accommodate linear peptide PTM identification, disabling the cross-linking reaction site settings in SeaPIC. This adjustment allowed us to exclusively search for linear peptides with potential PTMs. Furthermore, we had access to numerous publicly available datasets of PTM-containing linear peptides with known ground truth. We conducted tests on 21 different PTMs in linear peptides[2] (PXD009449) using SeaPIC and identified all PTMs with the highest PTM scores (supplemental Fig. S3 and S4 and supplemental Tables S1). These experiments demonstrate the reliability of SeaPIC and showcase its applicability in scenarios involving linear peptides. And linear peptide searching mode in SeaPIC shall also benefit the XL-MS datasets because in cleavable cross-linking tasks, MS3 spectra are frequently used to generate the linear peptide data.

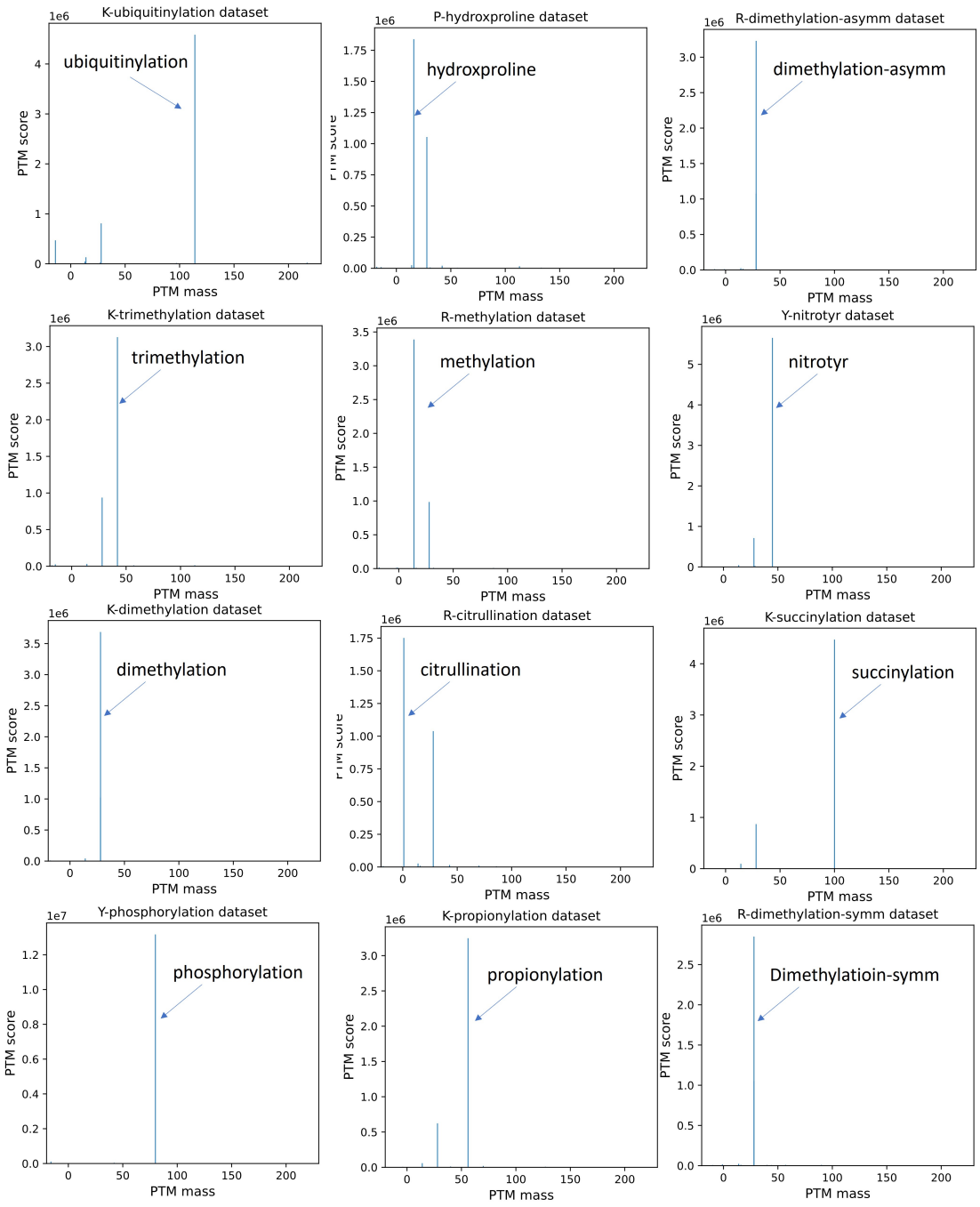

Figure S3: Results of 21 PTMs in the linear datasets. We used SeaPIC to run linear datasets with PTMs and generated bar plots with the horizontal axis PTM mass and the vertical axis PTM score. SeaPIC identified all 21 PTMs as having the highest PTM score.

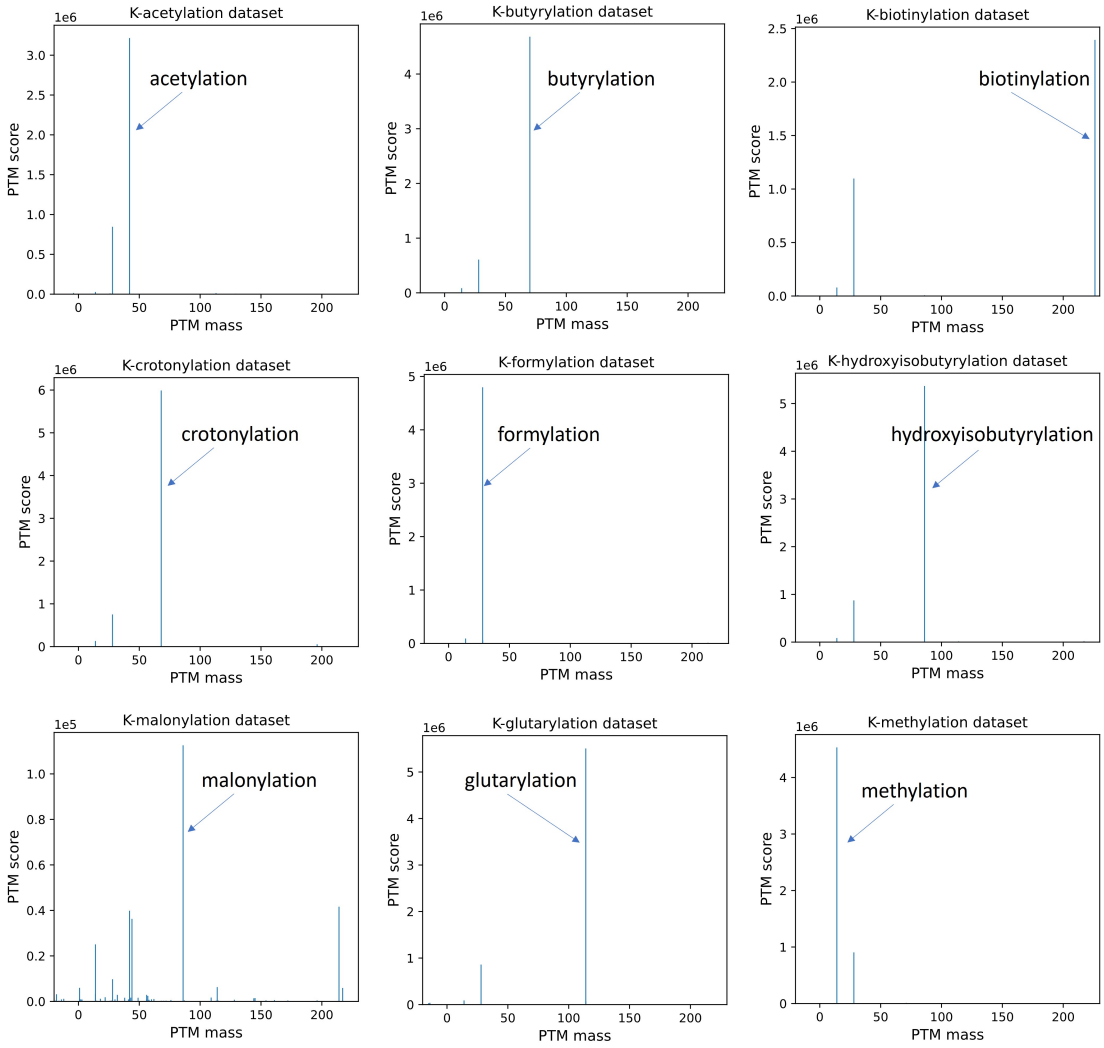

Figure S4: Results of 21 PTMs in the linear datasets. We used SeaPIC to run linear datasets with PTMs and generated bar plots with the horizontal axis PTM mass and the vertical axis PTM score. SeaPIC identified all 21 PTMs as having the highest PTM score.

Table S1: Statistics of 21 PTM datasets results for Fig. S3 and Fig. S4. We provided a list of the PTM information and total scan numbers that SeaPIC successfully identified that contained the relevant PTM information.

file_name	site	PTM name	mass	#scans
Ymod_Phospho_200fmol_3xHCD_R1	Y	phosphorylation	79.966	28756
Kmod_Dimethyl_200fmol_3xHCD_R1	K	dimethylation	28.031	8701
Kmod_Crotonyl_200fmol_3xHCD_R1	K	crotonylation	68.026	12394
Rmod_Methyl_200fmol_3xHCD_R1	R	methylation	14.016	8229
Kmod_Trimethyl_200fmol_3xHCD_R1	K	trimethylation	42.047	7520
Kmod_Hydroxyisobutyryl_200fmol_3xHCD_R1	K	hydroxyisobutyrylation	86.037	11085
Rmod_Dimethyl_asymm_200fmol_3xHCD_R1	R	dimethylation-asymm	28.031	7836
Kmod_Formyl_200fmol_3xHCD_R1	K	formylation	27.995	10488
Kmod_Biotinyl_200fmol_3xHCD_R1	K	biotinylation	226.078	5323
Rmod_Dimethyl_symm_200fmol_3xHCD_R1	R	dimethylation-symm	28.031	6967
Kmod_Propionyl_200fmol_3xHCD_R1	K	propionylation	56.026	6631
Kmod_Ubiquitinyl_200fmol_3xHCD_R1	K	ubiquitinylation	114.043	9546
Rmod_Citrullin_200fmol_3xHCD_R1	R	citrullination	0.984	4182
Kmod_Butyryl_200fmol_3xHCD_R2	K	butyrylation	70.042	10386
Kmod_Succinyl_200fmol_3xHCD_R1	K	succinylation	100.016	9527
Ymod_Nitrotyr_200fmol_3xHCD_R1	Y	nitrotyr	44.985	13930
Kmod_Acetyl_200fmol_3xHCD_R1	K	acetylation	42.011	6685
Kmod_Malonyl_200fmol_3xHCD_R1	K	malonylation	86.0	10228
Kmod_Methyl_200fmol_3xHCD_R1	K	methylation	14.016	10216
Kmod_Glutaryl_200fmol_3xHCD_R1	K	glutarylation	114.032	11553
Pmod_Hydroxyproline_200fmol_3xHCD_R1	P	hydroxproline	15.995	4143

5 Tables and figures

Table S2: We simulated 100,000 spectra and divided them into five datasets, each containing 0, 1, 2, 3, and 4 PTMs. Both pLink2[3] and ECL3[4] were employed to analyze the datasets with default settings, but they failed to identify any results in the remaining datasets except for the 0-PTM dataset. In both pLink2 and ECL3, the identified CSMs in the 1-PTM, 2-PTM, 3-PTM, and 4-PTM datasets are incorrectly matched CSMs.

		Simulated datasets (each contains 20,000 spectra)				
		0-PTM	1-PTM	2-PTM	3-PTM	4-PTM
Identified CSMs number	pLink2	19916	51	113	52	20
	ECL3	19556	44	30	21	18

Table S3: SeaPIC parameters used in the synthetic and real datasets.

	SeaPIC
Enzyme	Trypsin
Miss_cleavages	2
Min_length	5
Known Modifications	C+57.02Da (fixed) M+15.99Da (variable)
MS2 tolerance	0.01Da
Linker info	CBDPS $m_{xl} = 509.097$Da BS2G $m_{xl} = 96.021$Da BS3 $m_{xl} = 138.068$Da
Link site	K

Table S4: Search engine parameters used in the synthetic dataset.

Synthetic dataset	pLink2	ECL3
Enzyme	Trypsin	Trypsin
Miss_cleavages	2	2
Min_length	5	5
Modifications	(fixed) C+57.02Da (variable) M+15.99Da n-term,K+28.03Da n-term,K+34.06Da	(fixed) C+57.02Da (variable) M+15.99Da n-term,K+28.03Da n-term,K+34.06Da
MS1 tolerance	10ppm	10ppm
MS2 tolerance	20ppm	20ppm
Linker info	CBDPS m_{xl} = 509.097Da	
Link site	K	K
FDR	0.01	0.01

Table S5: Search engine parameters used in the PXD042584 dataset.

PXD042584	pLink2	ECL3
Enzyme	Trypsin	Trypsin
Miss_cleavages	2	2
Min_length	5	5
Modifications	(fixed) C+57.02Da (variable) M+15.99Da STY+79.97Da K+156.08Da	(fixed) C+57.02Da (variable) M+15.99Da STY+79.97Da K+156.08Da
MS1 tolerance	10ppm	10ppm
MS2 tolerance	0.02Da	0.02Da
Linker info	BS3 m_{xl} = 138.068Da	
Link site	K	K
FDR	0.01	0.01

Table S6: Search engine parameters used in the PXD045446 dataset.

PXD045446	pLink2	ECL3
Enzyme	Trypsin	Trypsin
Miss_cleavages	2	2
Min_length	5	5
Modifications	(fixed) C+57.02Da (variable) M+15.99Da L+15.01Da K+156.08Da	(fixed) C+57.02Da (variable) M+15.99Da L+15.01Da K+156.08Da
MS1 tolerance	10ppm	10ppm
MS2 tolerance	0.02Da	0.02Da
Linker info	BS3 $m_{xl} = 138.068$Da	
Link site	K	K
FDR	0.01	0.01

Table S7: Search engine parameters used in the PXD023593 dataset.

PXD023593	pLink2	ECL3
Enzyme	Trypsin	Trypsin
Miss_cleavages	2	2
Min_length	5	5
Modifications	(fixed) C+57.02Da (variable) M+15.99Da S+41.03Da K+114.03Da	(fixed) C+57.02Da (variable) M+15.99Da S+41.03Da K+114.03Da
MS1 tolerance	10ppm	10ppm
MS2 tolerance	0.02Da	0.02Da
Linker info	BS2G $m_{xl} = 96.021$Da	
Link site	K	K
FDR	0.01	0.01

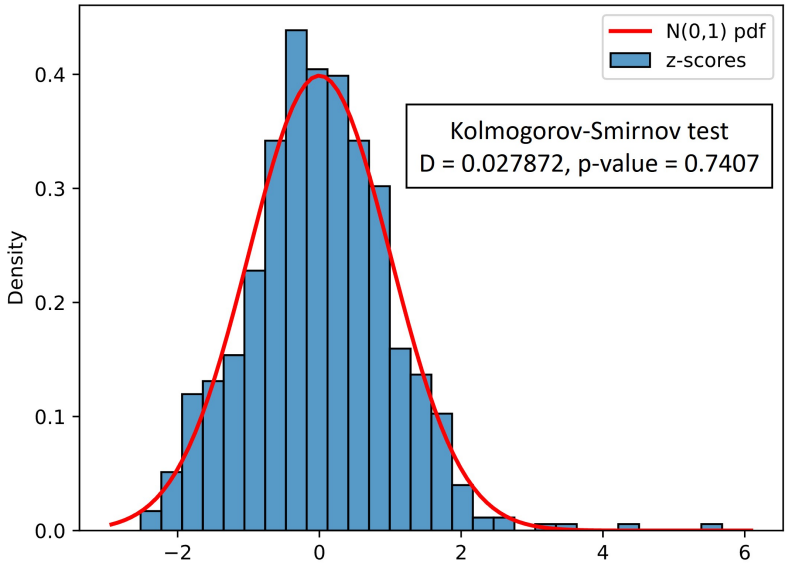

Figure S5: Histogram of z-scores compared with standard normal distribution. We take the logarithmic form of PTM scores and convert them into z-scores by subtracting the mean and dividing by the standard deviation. Then, we plot the histogram of these z-scores and compare it with the standard normal distribution using the K-S test. The p-value indicates that there are no significant differences.

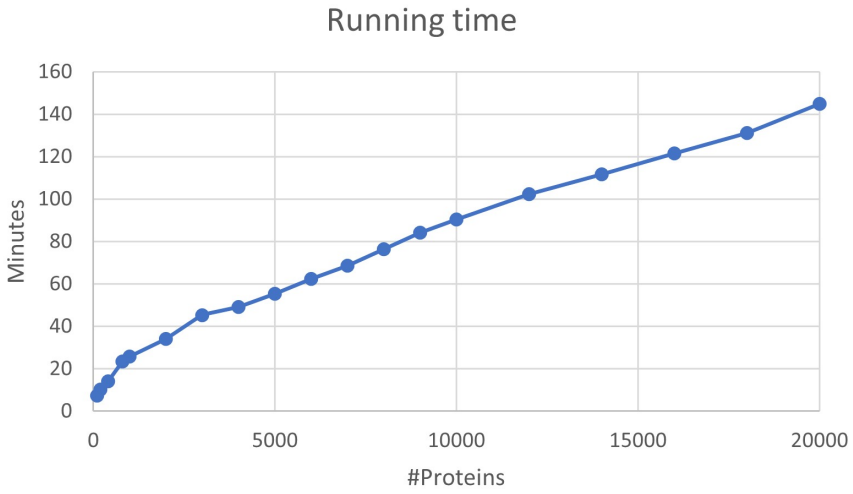

Figure S6: Running time of SeaPIC under different database sizes. SeaPIC runs on an Intel Core i5 2.90 GHz (8 cores/16 processors) Windows desktop computer with 32GB memory. We used different database sizes to run the human dataset with ∼20,000 spectra. The running time varies from several minutes to hours.

References

1. J. K. Eng, B. Fischer, J. Grossmann, and M. J. MacCoss, "A fast sequest cross correlation algorithm," *Journal of Proteome Research*, vol. 7, no. 10, pp. 4598–4602, 2008.

2. D. P. Zolg, M. Wilhelm, T. Schmidt, G. Médard, J. Zerweck, T. Knaute, H. Wenschuh, U. Reimer, K. Schnatbaum, and B. Kuster, "Proteometools: Systematic characterization of 21 post-translational protein modifications by liquid chromatography tandem mass spectrometry (LC-MS/MS) using synthetic peptides," *Molecular & Cellular Proteomics*, vol. 17, no. 9, pp. 1850–1863, 2018.

3. Z.-L. Chen, J.-M. Meng, Y. Cao, J.-L. Yin, R.-Q. Fang, S.-B. Fan, C. Liu, W.-F. Zeng, Y.-H. Ding, D. Tan, *et al.*, "A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides," *Nature Communications*, vol. 10, no. 1, pp. 3404–3415, 2019.

4. C. Zhou, S. Dai, S. Lai, Y. Lin, X. Zhang, N. Li, and W. Yu, "ECL 3.0: a sensitive peptide identification tool for cross-linking mass spectrometry data analysis," *BMC Bioinformatics*, vol. 24, no. 1, p. 351, 2023.