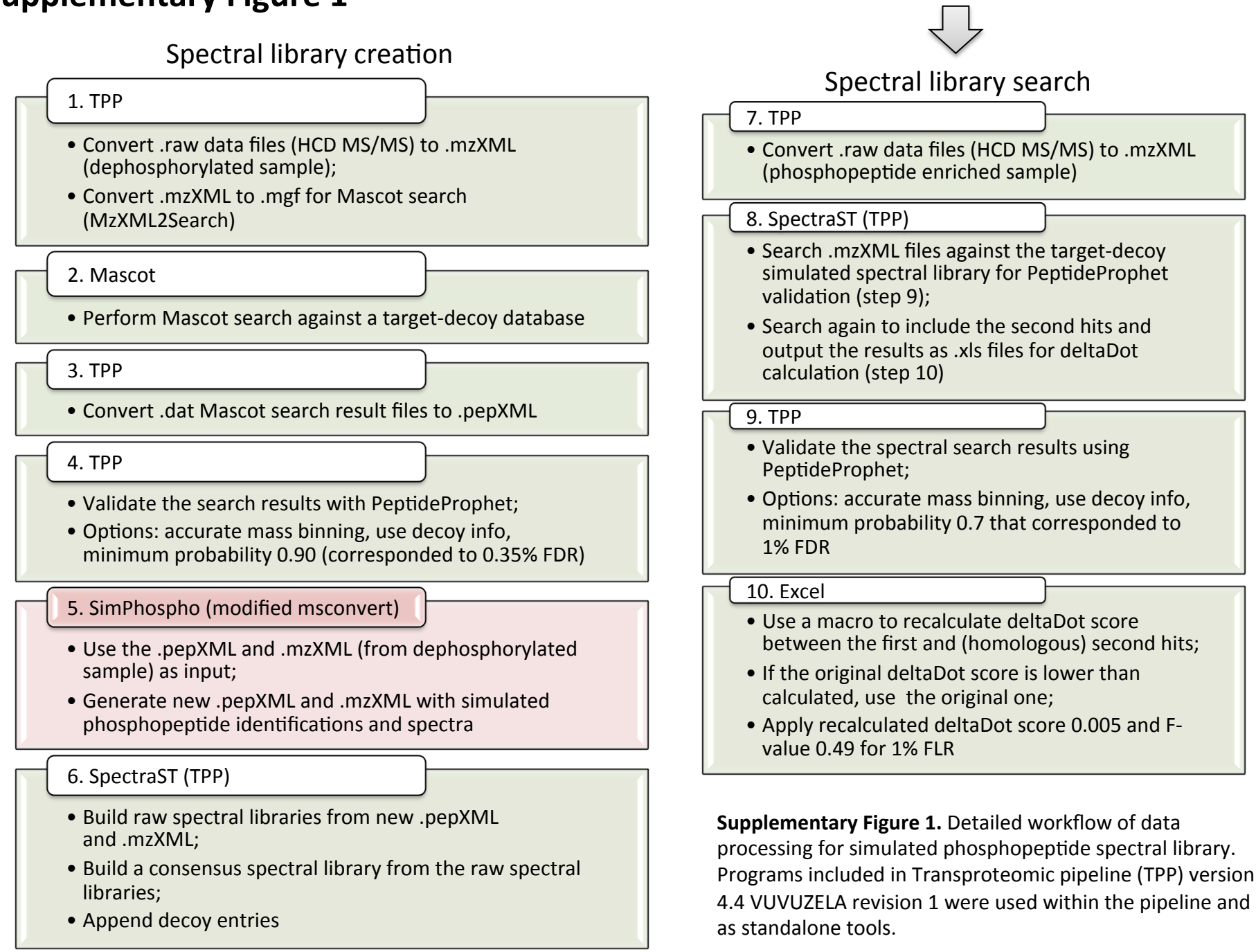
Supplementary Figures 1-13

Confident site localization using a simulated phosphopeptide spectral library

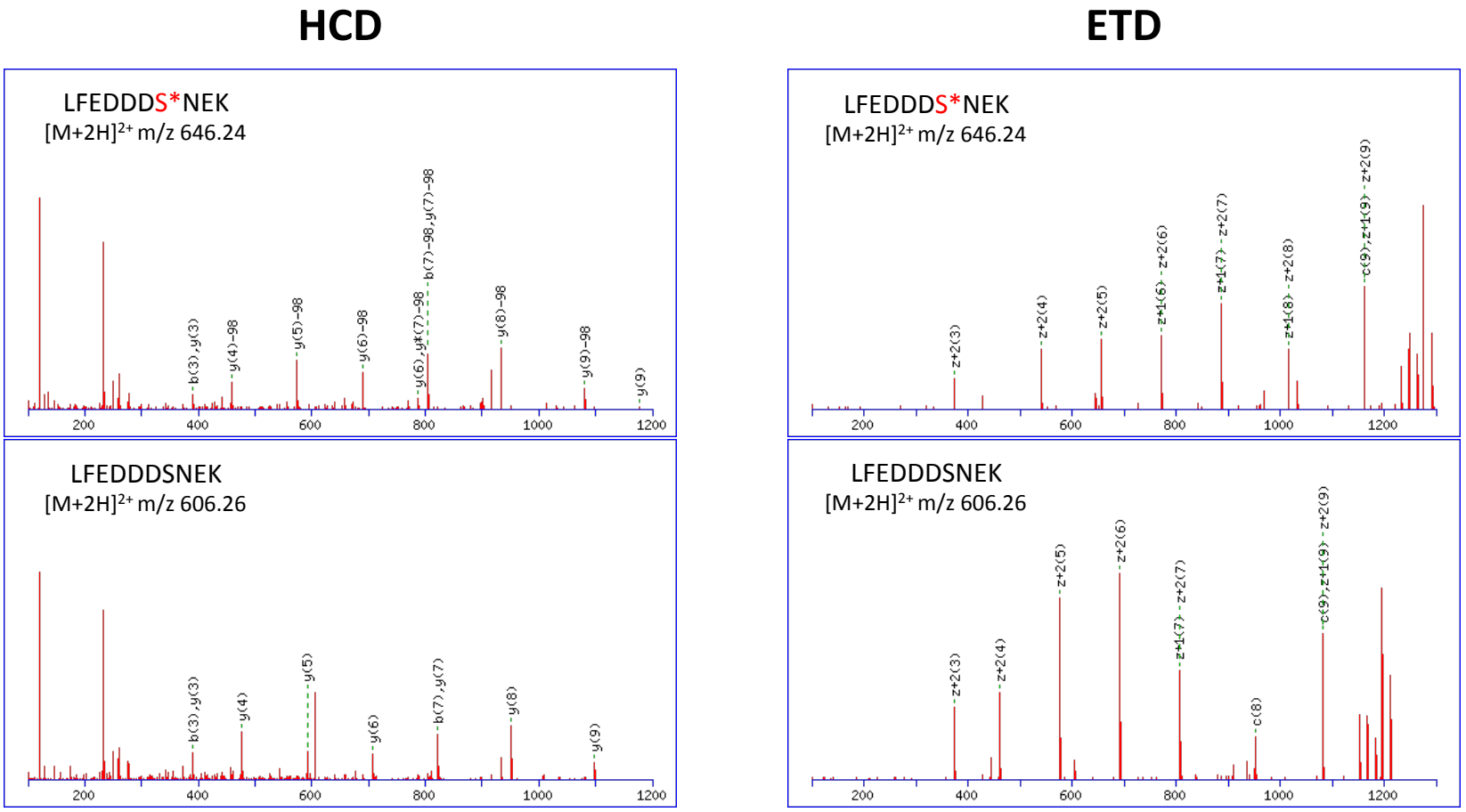
Veronika Suni, Susumu Y. Imanishi, Alessio Maiolica, Ruedi Aebersold, and Garry L. Corthals

Supplementary Figure 1



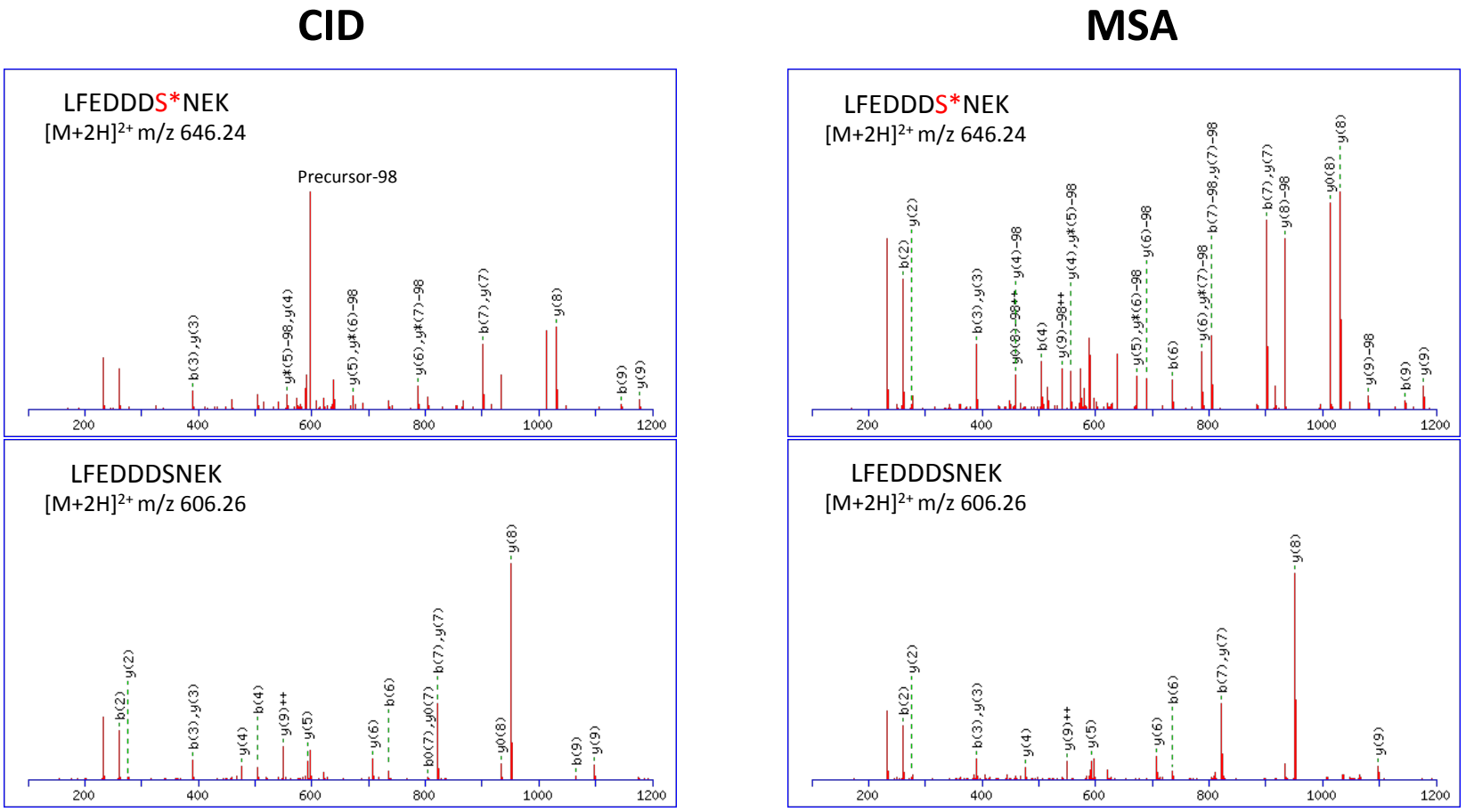
Supplementary Figure 1. Detailed workflow of data processing for simulated phosphopeptide spectral library. Programs included in Transproteomic pipeline (TPP) version 4.4 VUVUZELA revision 1 were used within the pipeline and as standalone tools.

Supplementary Figure 2A



Supplementary Figure 2. Representative MS/MS spectra of phosphorylated and dephosphorylated peptides obtained with different fragmentation methods of LTQ Orbitrap Velos. HeLa phosphopeptide and dephosphorylated peptide samples were analyzed with four different fragmentation methods, HCD, ETD, CID, and MSA (see Table 1). A phosphopeptide (LFEDDDS*NEK; S*: phosphoserine) and its dephosphorylated form with the same charge state were identified by Mascot database searching. (A) Examples of HCD and ETD spectra. Striking similarity can be noted in HCD spectra of phosphorylated and dephosphorylated peptides in terms of intensity and presence of fragment ion peaks (left panel). Likewise, ETD fragmentation of phosphorylated and dephosphorylated peptides shows similarity (right panel).

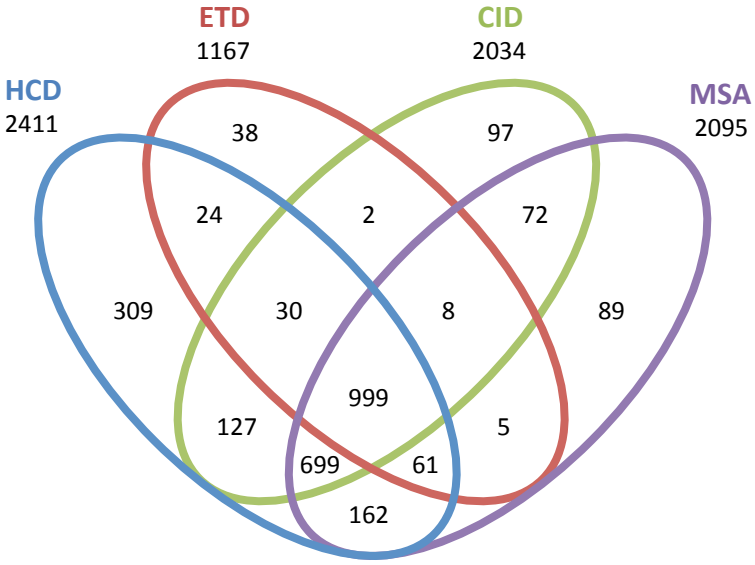
Supplementary Figure 2B



Supplementary Figure 2. (B) Examples of CID and MSA spectra. The complexity of CID spectra of phosphopeptides is higher than in HCD spectra. In many cases the neutral loss of H_3PO_4 from a phosphopeptide precursor ion (-98 Da) is a prominent peak in CID spectra (left). MSA spectra of phosphopeptides are even more complex than CID spectra as tandem fragmentation generates fragment ion peaks with and without the neutral loss of H_3PO_4 (right).

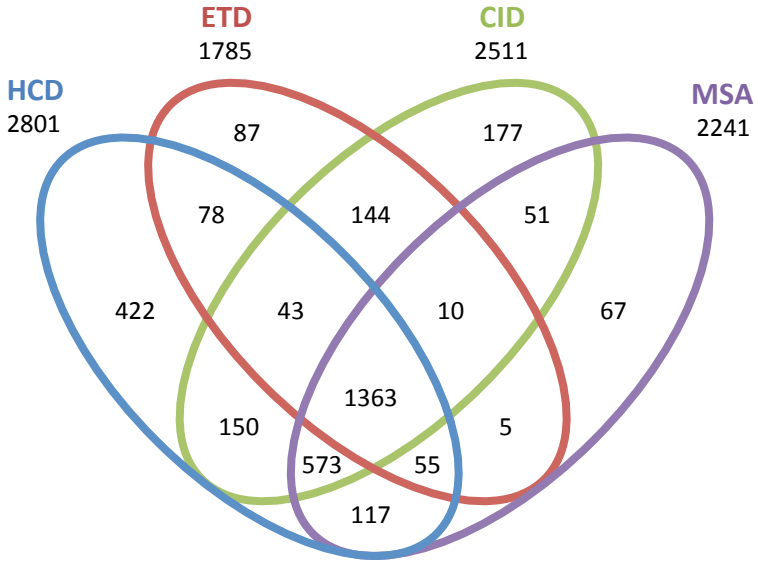
Supplementary Figure 3

A HeLa phosphopeptide data



2722 unique phosphopeptides
(*phosphorylation sites not considered*)

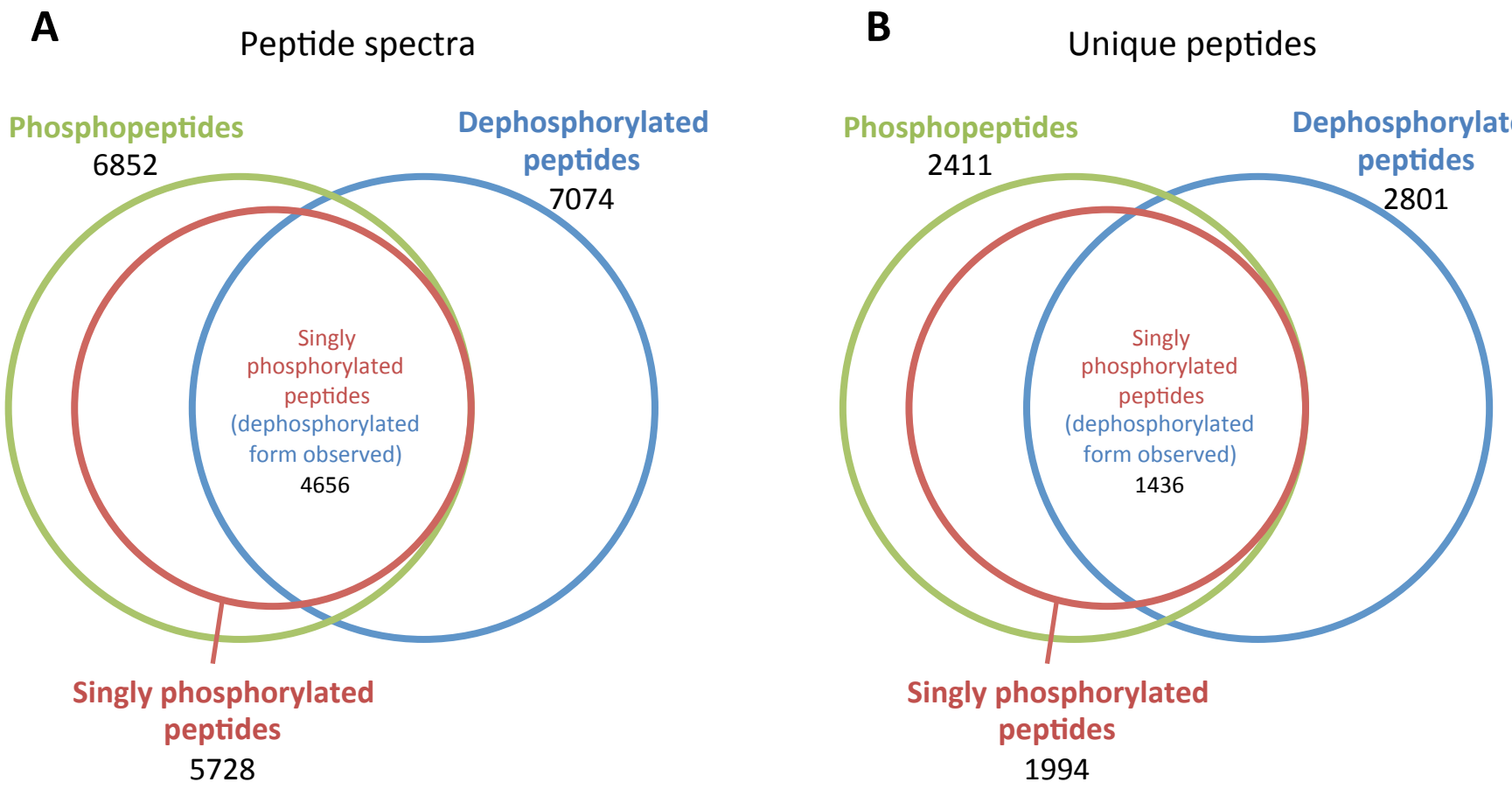
B HeLa dephosphorylated peptide data



3342 unique non-phosphorylated peptides

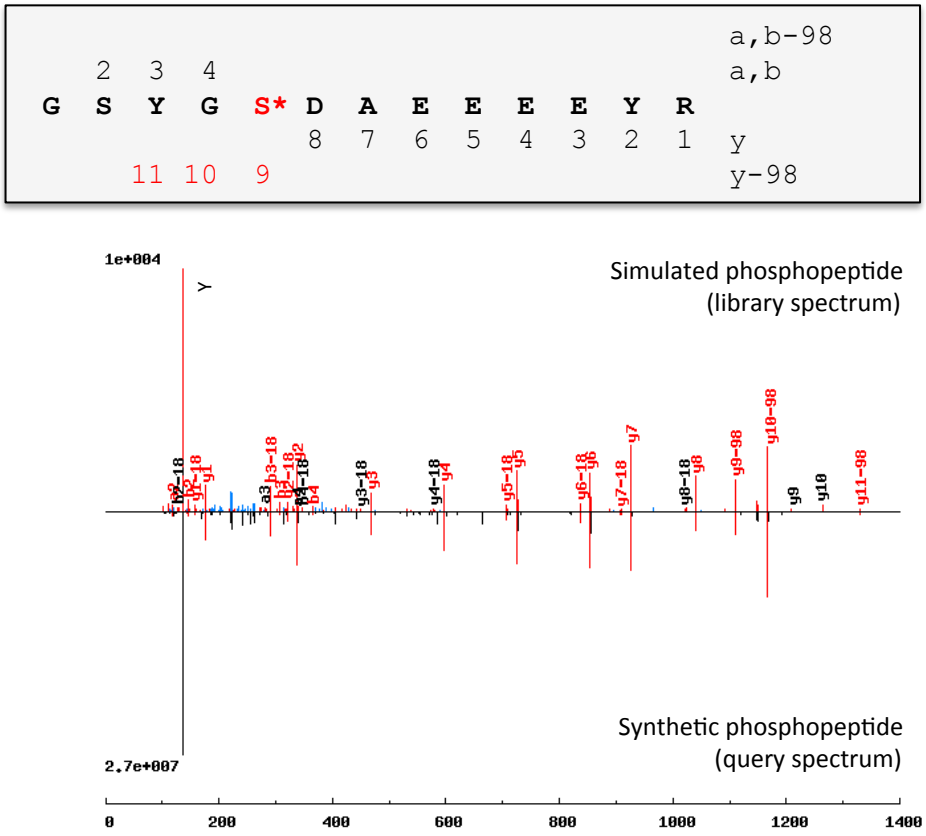
Supplementary Figure 3. Phosphorylated and dephosphorylated peptides identified from HeLa cells by different fragmentation setups of LTQ Orbitrap Velos. HCD, ETD, CID, and MSA data of the HeLa phosphopeptide and dephopshorylated peptide samples were acquired in an Orbitrap mass analyzer (see Table 1). From those data, phosphorylated and non-phosphorylated peptides were identified by Mascot. The number of unique peptides at PeptideProphet estimated FDR of 1% is shown, where a peptide phosphorylated at different sites is counted as 1 (due to the low confidence of the site localization) and also a peptide with and without methionine oxidation is counted as 1. (A) The highest number of phosphopeptide identifications was achieved by HCD, which covered 89% of all the phosphopeptides identified using the four fragmentation methods. (B) In case of the analysis of dephosphorylated peptides, HCD alone produced 84% of all the non-phosphorylated peptides.

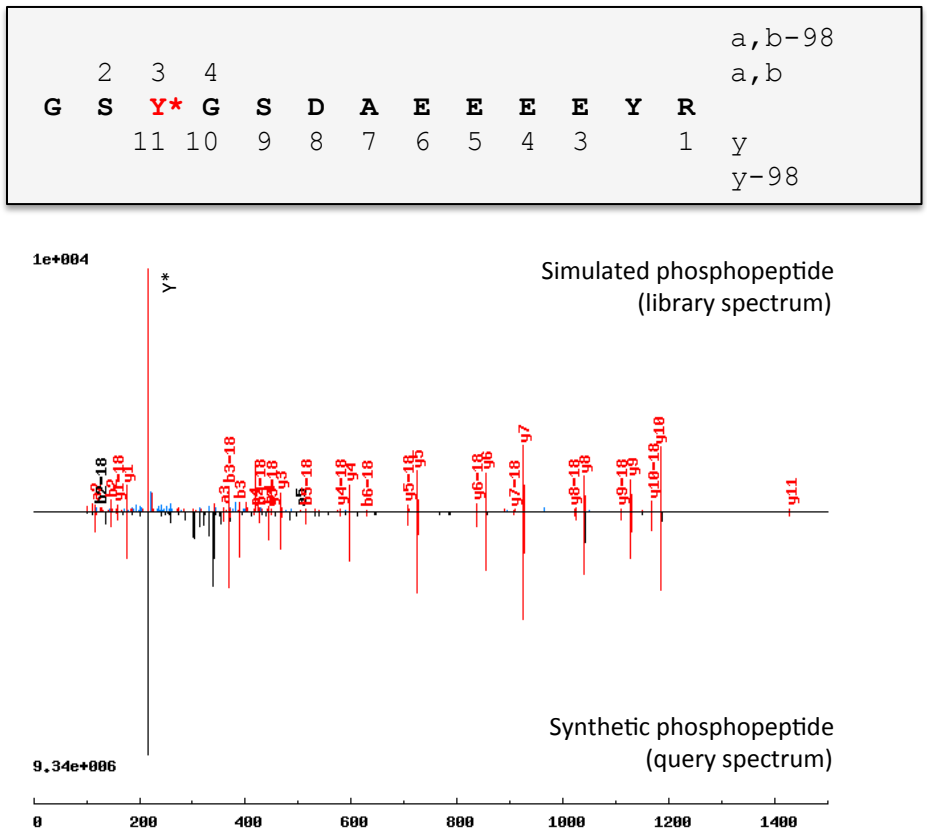
Supplementary Figure 4



Supplementary Figure 4. HeLa phosphopeptides and dephosphorylated peptides observed by Orbitrap HCD. The HeLa phosphopeptides and dephosphorylated peptides were identified from their respective HCD data by Mascot at 1% FDR (refer to Supplementary Fig. 3; see also Table 1 for the used data). (A) 84% of phosphopeptides (5728 of 6852 spectral matches) were identified as singly phosphorylated peptides, of which 81% (4656 spectral matches) was observed also in their dephosphorylated form. Phosphopeptide enrichment selectivity was 90% (6852 of 7624 spectral matches), and dephosphorylation efficiency was 99% (7074 of 7138 spectral matches; not shown in the figure). (B) The number of unique peptides is shown. 72% of singly phosphorylated peptides (1436 of 1994) were observed in the dephosphorylated form. As mentioned in Supplementary Fig. 3, a peptide phosphorylated at different sites is counted as 1 due to the low confidence of the site localization, and also a peptide with and without methionine oxidation is counted as 1.

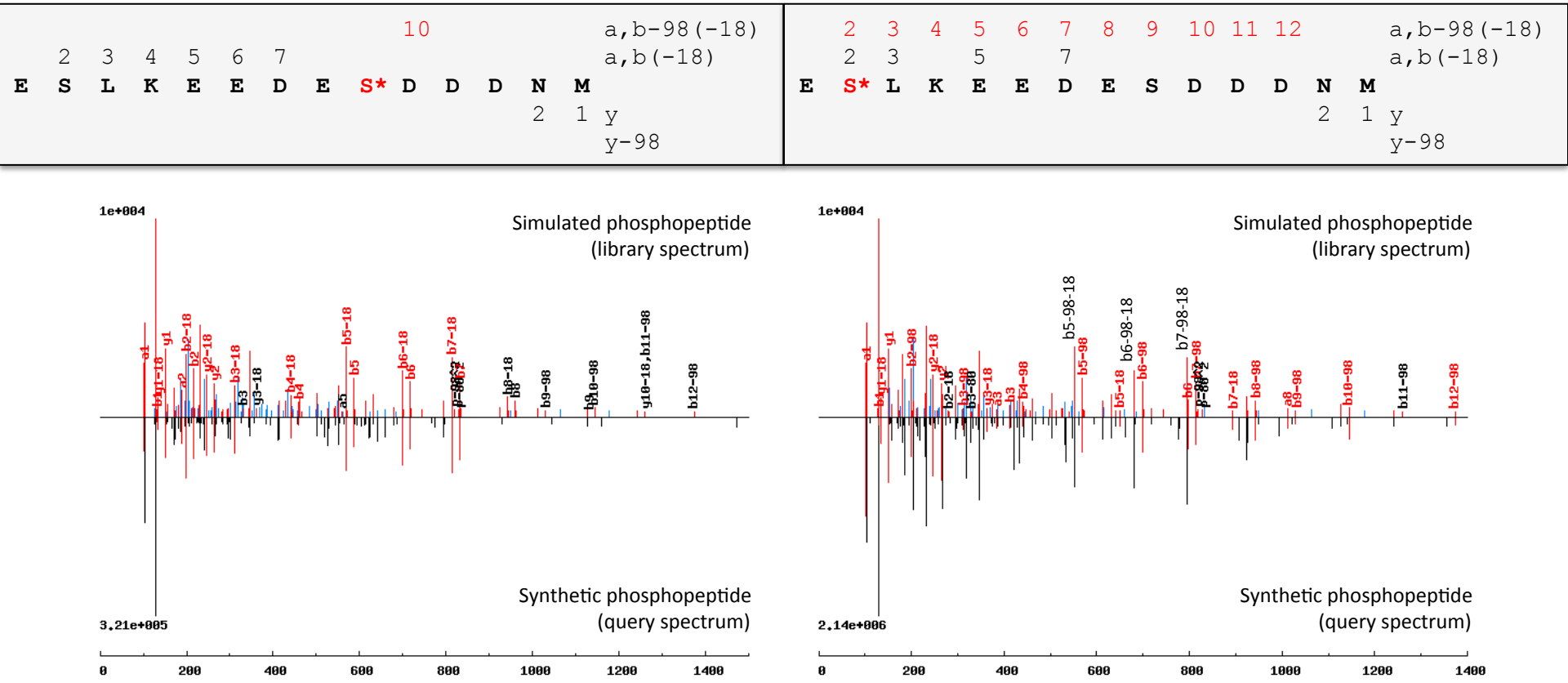
Supplementary Figure 5A





Supplementary Figure 5. Representative HCD spectral matches by SimSpectraST. A simulated phosphopeptide spectral library SimHeLa was built from the Orbitrap HCD spectra of the dephosphorylated HeLa sample (see Table 2). HCD spectra of 20 synthetic phosphopeptides with known phosphorylation sites were acquired using a Q Exactive (query spectra, see also Table 1 for the data). Amino acid residues followed by an asterisk represent phosphorylation sites. A representative spectral match is also shown in Fig. 1C. (A) In HCD spectra of tryptic peptides, mainly y-series fragment ions are observed with relatively small a- and b-series ions. In case of a serine-phosphorylated synthetic peptide GSYGS*DAEEEEYR (left), the predominant neutral loss of H_3PO_4 (-98 Da) was observed on y-ions containing the phosphorylation site. However, a tyrosine phosphorylated form of the peptide, GSY*GSDAEEEEYR (right), maintained H_3PO_4 on fragment ions as well as on the tyrosine immonium ion, without the neutral loss. These typical features of phosphopeptide fragmentation were well predicted in simulated spectra, which contributes to correct spectral matches.

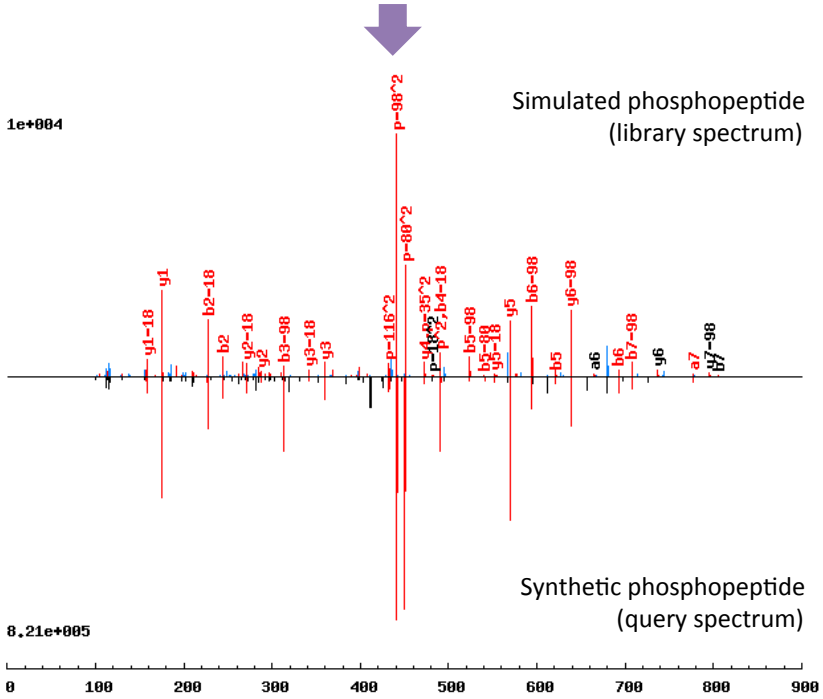
Supplementary Figure 5B



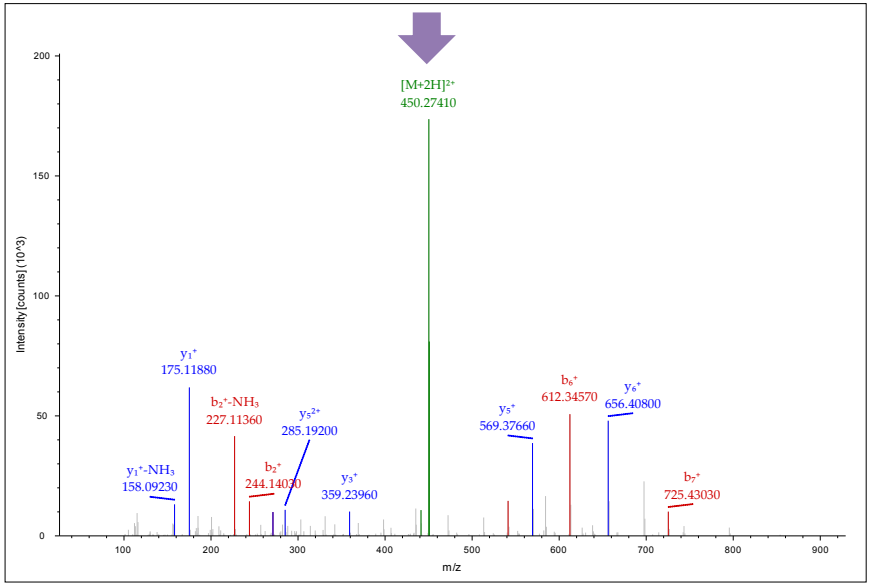
Supplementary Figure 5. (B) A synthetic phosphopeptide ESLKEEDES*DDDNM, which is a protein C-terminal tryptic peptide, generated more b-ions than y-ions due to the lack of C-terminal lysine/arginine and also the presence of internal lysine (left). The b-ions were accompanied by a predominant neutral loss of H_2O (-18 Da). In case of the other phosphorylation isoform ES*LKEEDESDDDNM (right), the loss of H_3PO_4 was observed on b-ions as well as on the H_2O loss ions, with maintained relative abundance. Since the H_2O loss is identical to the loss of H_3PO_4 from phosphorylated serine/threonine ($+HPO_3$ $-H_3PO_4$ = $-H_2O$), the predominant H_2O loss may induce false localization of phosphorylation sites by misrecognizing those losses, e.g. in the localization by phosphoRS 3 considering the H_3PO_4 loss (see Fig. 2D). In simulated spectra, the combination of the neutral losses were well predicted as shown in this figure, which contributes to correct spectral matching and site localization.

Supplementary Figure 5C

		3		5	6	7		a,b-H_3PO_4
	2			5	6	7		a,b
S	**R**	**S***	**P**	**L**	**A**	**I**	**R**	
			5	4	3	2	1	y
		6						y-H_3PO_4

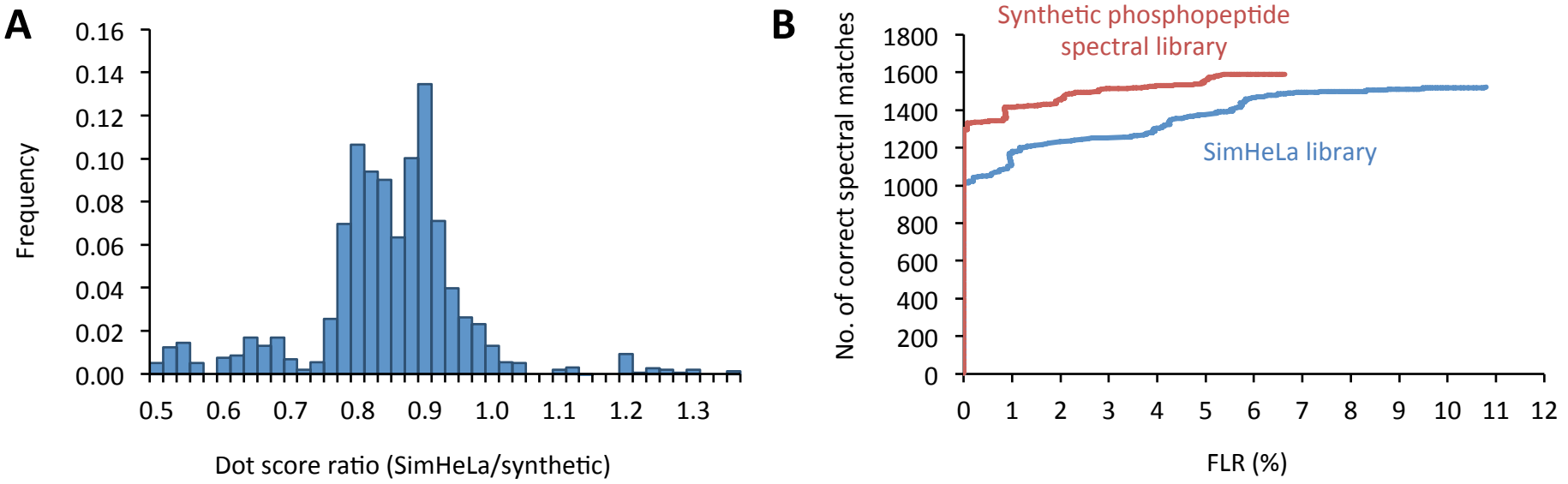


SRSPLAIR
(dephosphorylated peptide HCD spectrum)



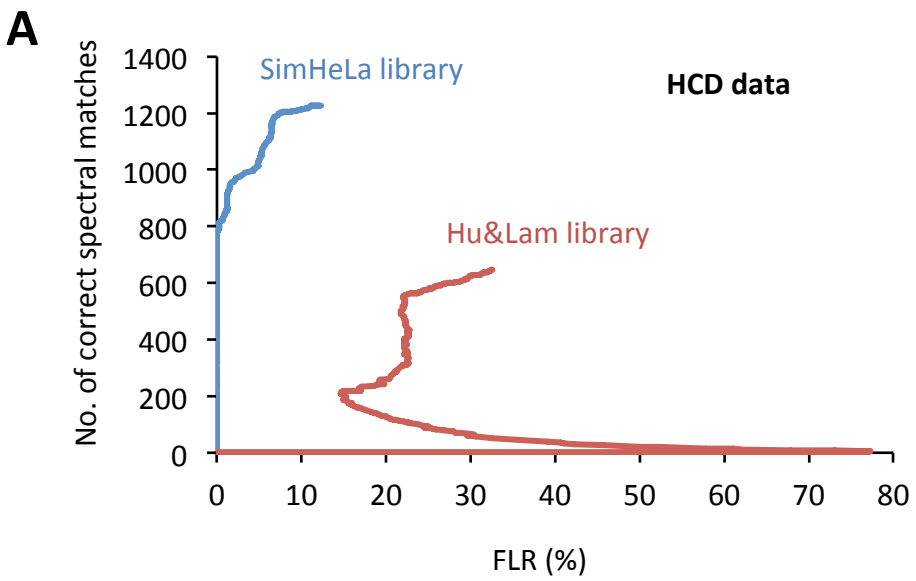
Supplementary Figure 5. (C) SRS*PLAIR. The most intense matched peaks came from the neutral loss of H_3PO_4 from the precursor ion (left). In the HCD spectrum of the dephosphorylated peptide SRSPLAIR used for the simulation (right), the precursor ion was not fragmented completely and remained as a dominant ion peak, of which relative abundance was maintained in the phosphopeptide spectrum as the dominant neutral loss peak. In contrast to CID spectra (Supplementary Fig. 2B), the H_3PO_4 loss from precursor ions in HCD spectra is predictable as well as from fragment ions. (However, peaks in a precursor *m/z* region are ignored in SpectraST searching.)

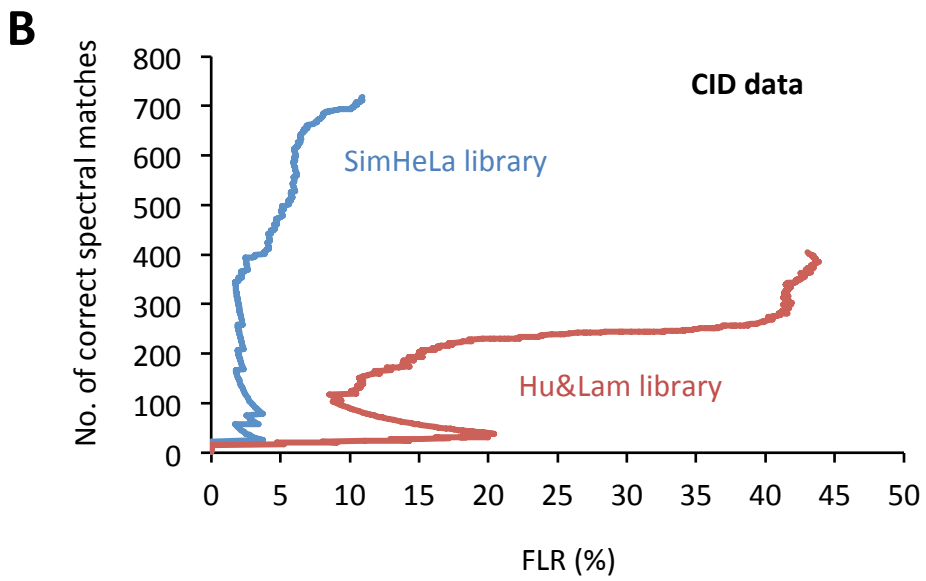
Supplementary Figure 6



Supplementary Figure 6. Comparison of SimHeLa library with the synthetic phosphopeptide spectral library in SpectraST searching. (A) Spectral similarity between simulated and synthetic phosphopeptides was investigated. SpectraST searching of the 20 synthetic phosphopeptide HCD data was performed with a spectral library of the same synthetic phosphopeptides, as well as with SimHeLa library (see Tables 1 and 2). Dot scores of correct spectral matches were compared between those searches. A ratio of dot scores (SimHeLa/synthetic) was 0.83 ± 0.11 (mean ± S.D.). (B) The SpectraST results obtained with these two libraries were compared in phosphorylation site localization. The number of correct spectral matches as a function of FLR is shown. In this data analysis, 2 of the 20 synthetic phosphopeptides were excluded since no alternative site for those phosphopeptides was covered by the synthetic phosphopeptide spectral library (see Supplementary table 1). The data were sorted by the discriminant score F-value. A quantitative ratio of correct spectral matches (SimHeLa/synthetic) was 0.83 at 1% FLR.

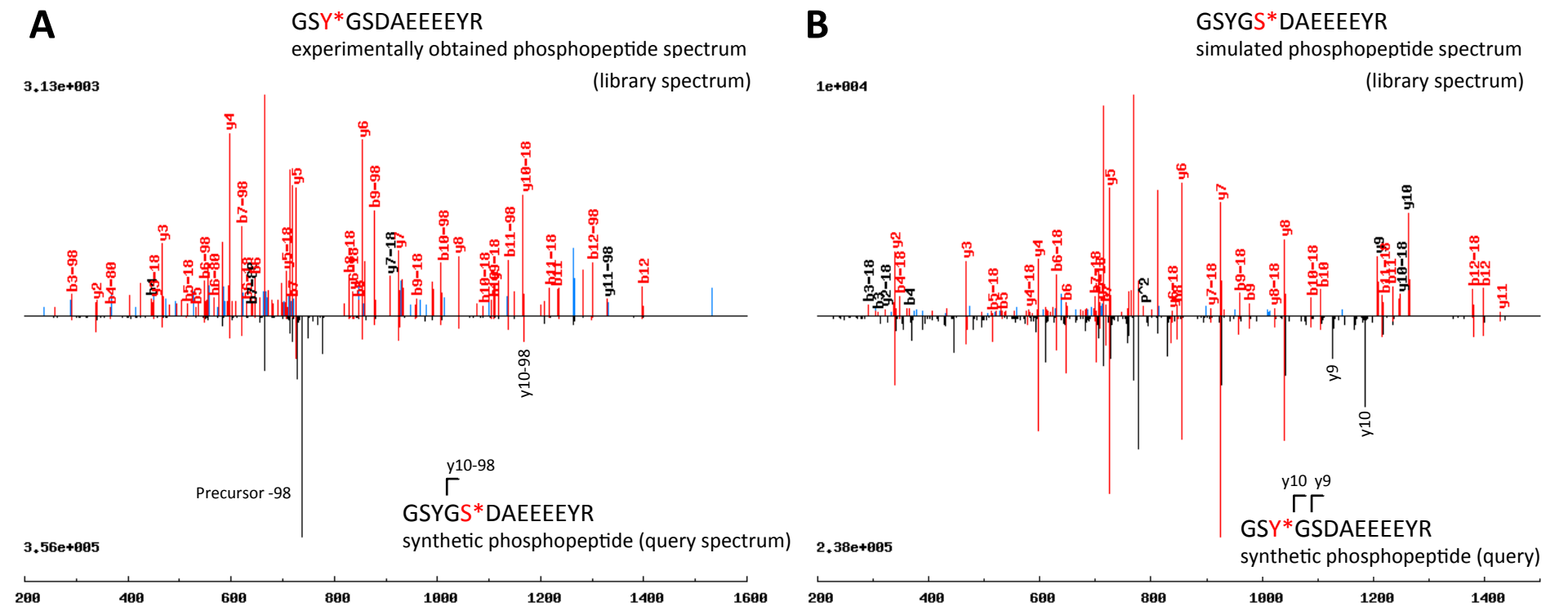
Supplementary Figure 7





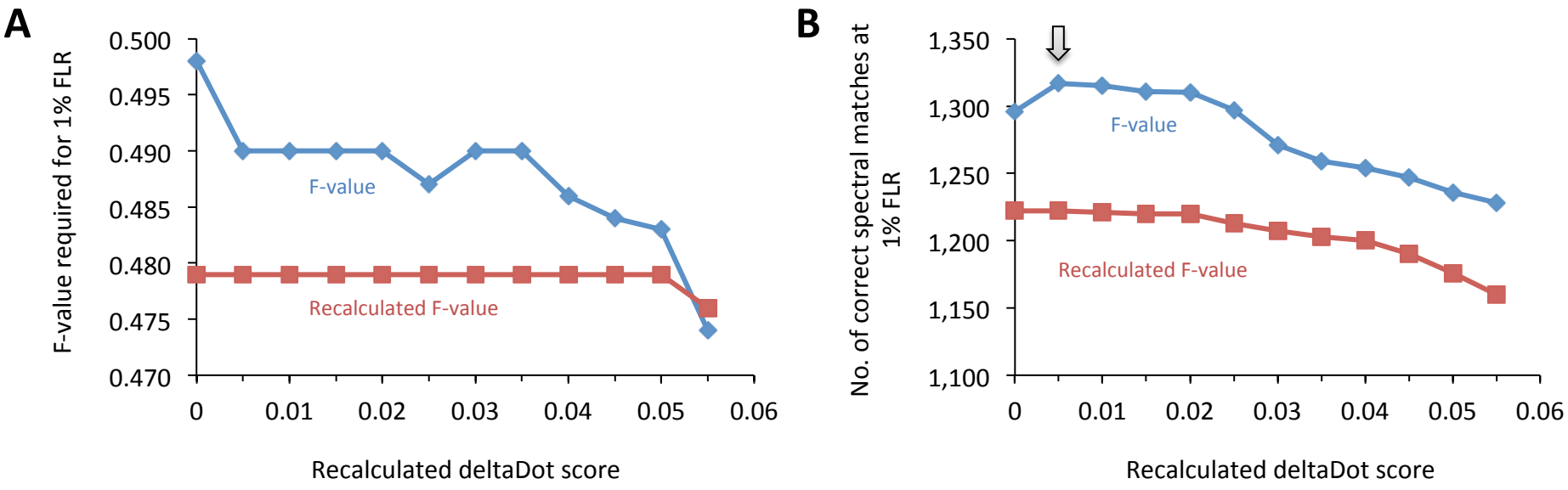
Supplementary Figure 7. Comparison of SimHeLa library with Hu&Lam library in phosphorylation site localization by SpectraST searching. Hu&Lam library consisted of experimentally obtained and simulated CID spectra of phosphopeptides (see Table 2). For comparing two libraries, (A) HCD and (B) CID spectral data of the 20 synthetic phosphopeptides were used (see Table 1). Out of those, 14 synthetic phosphopeptides, of which sequences were covered by both the libraries, were taken into account, and the data were sorted by F-value. SinHeLa library outperformed Hu&Lam library in the localization sensitivity.

Supplementary Figure 8



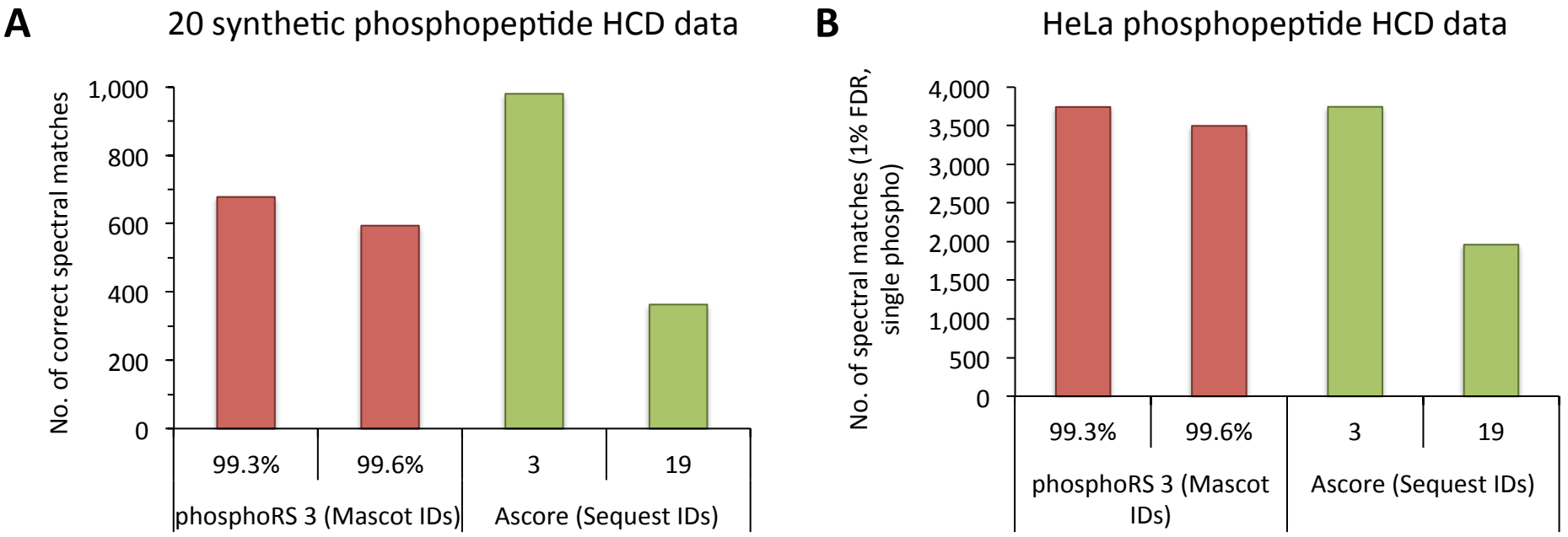
Supplementary Figure 8. Examples of false spectral matches by SpectraST searching with a library created by Hu and Lam (Hu&Lam library). Hu&Lam library consists of two components: experimentally obtained CID spectra of phosphopeptides identified and localized, and also complementary phosphopeptide spectra simulated (semi-empirical predicted) from publically available CID spectra of non-phosphorylated peptides (see Table 2). However, as shown in Supplementary Fig. 7A,B, SpectraST searching with Hu&Lam library could not localize phosphorylation sites as well as with SimHeLa library. One possible reason is the incomplete coverage of alternative phosphorylation sites, which may mislead localization of all uncovered sites into covered sites. Another reason is the false spectral matching as illustrated in this figure. Amino acid residues followed by an asterisk represent phosphorylation sites. (A) A false match of a CID spectrum of a synthetic phosphopeptide GSYGS*DAEEEEYR to an experimentally obtained library spectrum of GSY*GSDAEEEEYR is shown. See Table 1 for the 20 synthetic phosphopeptide CID data used for this figure. Actually, these spectra appear to be well matched except for the mass range around the precursor (ion peaks in that range seemed to be removed from this library spectrum). This indicates that the phosphopeptide used for building the library was false localized, which may constantly result in the false localization in spectral library searching. (B) In case of a false match of a synthetic phosphopeptide spectrum GSY*GSDAEEEEYR to a simulated library spectrum of GSYGS*DAEEEEYR, these spectra are well matched except for the site determining fragment ions, y9 and y10. This false match was most likely due to the inaccurate prediction of phosphopeptide spectra, particularly the neutral loss of H_3PO_4 from phosphorylated serine/threonine.

Supplementary Figure 9



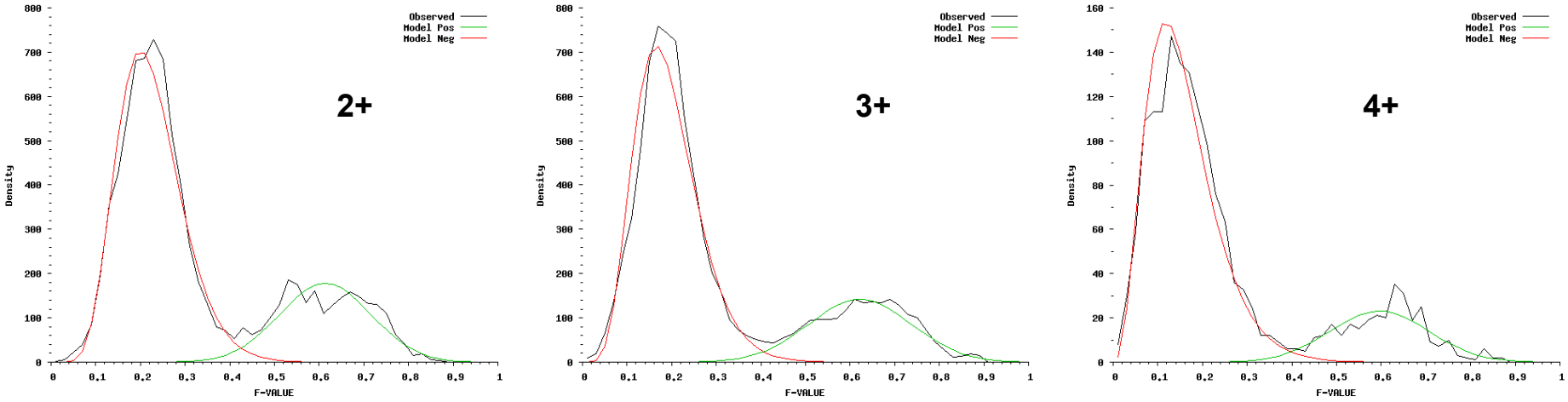
Supplementary Figure 9. Optimization of recalculated deltaDot cutoff. Since deltaDot score used in SpectraST considers difference from a nonhomologous peptide as the second best hit, difference between phosphopeptide isoforms is often ignored. Therefore, we recalculated deltaDot score by comparing the first and second best hits even if they represent the same peptide sequence. As deltaDot score contributes to the discriminant score F-value for every spectral match, F-value was also recalculated based on recalculated deltaDot score. These parameters were evaluated in SimSpectraST searching of the 20 synthetic phosphopeptide HCD data with SimHeLa library (see Tables 1 and 2). (A) F-value and recalculated F-value thresholds required for 1% FLR are shown as a function of recalculated deltaDot score. Fig. 2B illustrates how the F-value thresholds at recalculated deltaDot score of 0.005 were obtained as an example. (B) The number of correct spectral matches at 1% FLR as a function of recalculated deltaDot score is shown, where the corresponding F-value and the recalculated F-value thresholds were applied. This data indicates that SimSpectraST searching using F-value was more sensitive than with recalculated F-value at 1% FLR on the synthetic phosphopeptide HCD dataset. The maximum number of the correct spectral matches (1317 out of 1953 spectral matches) was obtained at the recalculated deltaDot of 0.005 and the corresponding F-value threshold of 0.49. An FLR cutoff will be applied only to phosphopeptides of which serine/threonine/tyrosine are not fully occupied by phosphorylation, i.e. the number of these residues is >1 in case of singly phosphorylated peptides.

Supplementary Figure 10



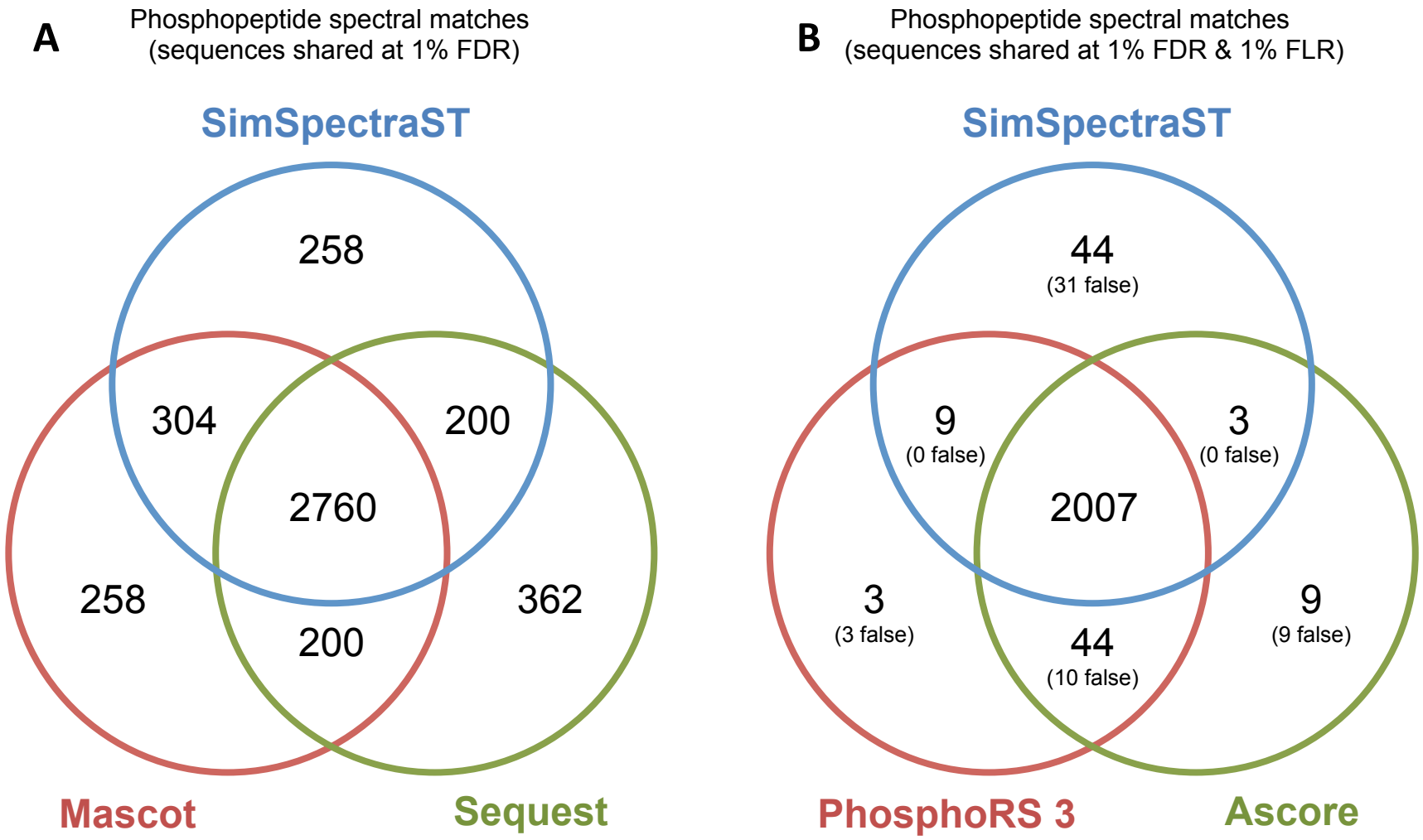
Supplementary Figure 10. Ascore cutoffs for 1% FLR. Ascore of 3 obtained as a 1% FLR cutoff (see Fig. 2C) and the previously reported one (Ascore of 19, based on ion trap CID data, ref. 21) were applied to Sequest search results of (A) the 20 synthetic phosphopeptide HCD data and (B) the HeLa phosphopeptide HCD data. See Table 1 for the used HCD data. Also, as reference, phosphoRS 3 probabilities of 99.3% obtained as a 1% FLR cutoff (Fig. 2D) and 99.6% suggested for the Marx dataset (Fig. 3A) were applied to Mascot search results of the same HCD data. On both the synthetic and HeLa phosphopeptide data, Ascore of 19 clearly reduced the numbers of spectral matches compared to the other tested conditions.

Supplementary Figure 11



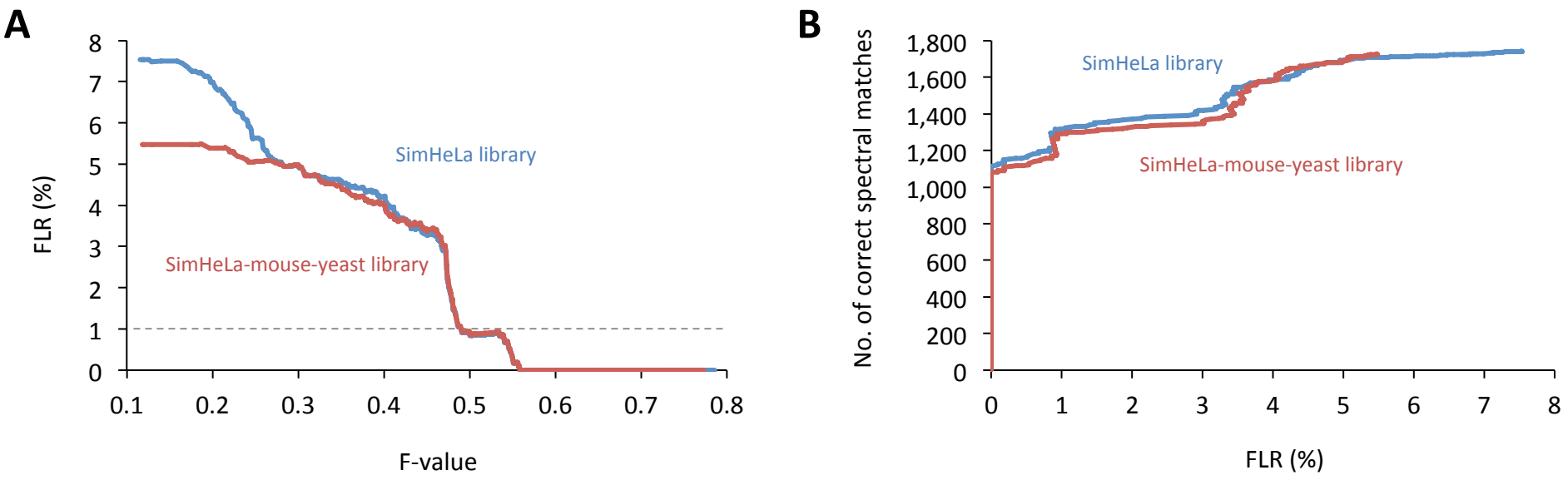
Supplementary Figure 11. PeptideProphet validation of SimSpectraST search results. SimSpectraST searching of the HeLa phosphopeptide HCD data was performed using SimHeLa library (see Tables 1 and 2). FDR was estimated using the target-decoy strategy by PeptideProphet. For each HCD spectral match, a discriminant score F-value is calculated, which is in turn used to assign probability values for peptide identifications. Fitted frequency curves on the SimSpectraST search results (2+, 3+, and 4+ charged peptides) are shown in the figure. Correct identifications were well discriminated from incorrect identifications. Minimum probability of 0.70 was selected as a cutoff for estimated 1% FDR.

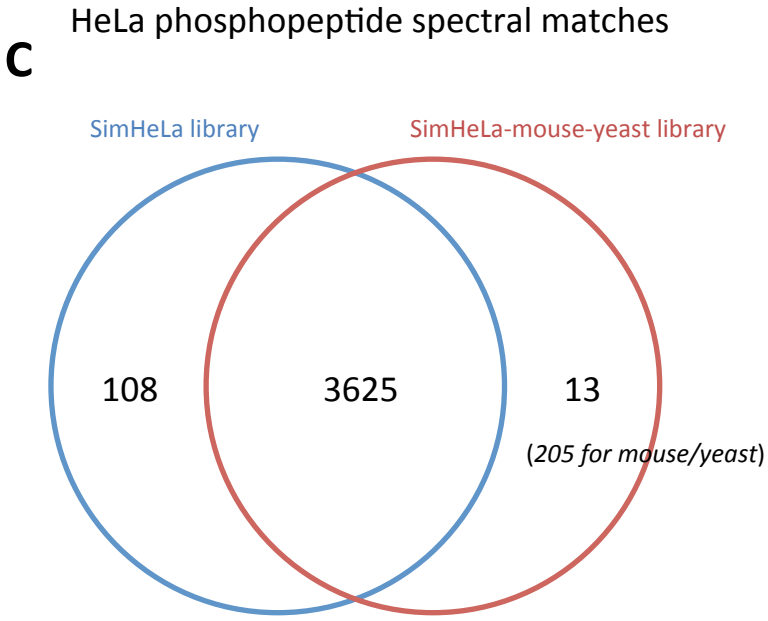
Supplementary Figure 12



Supplementary Figure 12. Differentially localized phosphorylation sites on HeLa tryptic peptides. Out of the singly phosphorylated peptides identified at 1% FDR from the HeLa phosphopeptide HCD data (see Fig. 4A), spectral matches shared by SimSpectraST (SimHeLa library), Mascot, and Sequest were extracted (see also Tables 1 and 2 for the used data and library, respectively). Differential sequence identifications (0.45% of spectral matches) were excluded. (A) The three methods agreed on 78.4% of phosphorylation sites (2760 of 3522 spectral matches). As shown in the figure, no clear bias in the localization was observed among the three methods. (B) After applying the respective 1% FLR cutoffs of SimSpectraST, phosphoRS 3, and Ascore (see Fig. 2), the site agreement was increased to 97.3% (2007 of 2063 spectral matches). All the 56 site disagreement spectra were manually interpreted (3 spectra ambiguous). When we assume that all the agreed sites are correct, estimated FLRs for SimSpectraST, phosphoRS 3, and Ascore are 1.5%, 0.63%, and 0.92%, respectively, which are well approximated to the expected FLR (1%).

Supplementary Figure 13





Supplementary Figure 13. SimSpectraST searching with an increased size of SimHeLa library. SimHeLa library was supplemented with publically available phosphopeptide spectral libraries (mouse and yeast, ion trap CID spectra), which resulted in a 7 times larger library than SimHeLa library (SimHeLa-mouse-yeast library, see Table 2). Decoy entries were added for the FDR estimation. Recalculated deltaDot ≥0.005 was applied. (A) In SimSpectraST searching of the 20 synthetic phosphopeptide HCD data (see Table 1), both SimHeLa and SimHeLa-mouse-yeast libraries required the same F-value (0.49) for 1% FLR. (B) The searching with SimHeLa-mouse-yeast library showed 2.1% less spectral matches than with SimHeLa at 1% FLR; however no significant difference in the sensitivity was observed. The data were sorted by F-value. (C) In SimSpectraST searching of the HeLa phosphopeptide HCD data (see Table 1), the use of SimHeLa-mouse-yeast library provided 95 less spectral matches (2.5% less) than that of SimHeLa library at 1% FDR.