Supporting Information

Title: MS1-level proteome quantification platform allowing maximally increased multiplexity for SILAC and *in vitro* chemical labeling

Authors: Yeon Choi[1,2,†], Kyowon Jeong[1,2,†,§], Sanghee Shin[1,2,†], Joon Won Lee[3,†], Young-suk Lee[1,2], Sangtae Kim[4], Sun Ah Kim[1,2], Jaehun Jung[3], Kwang Pyo Kim[3], V. Narry Kim[1,2,*], and Jong-Seo Kim[1,2,*]

Affiliations:

[1]Center for RNA Research, Institute for Basic Science, Seoul 08826, Korea

[2]School of Biological Sciences, Seoul National University, Seoul 08826, Korea

[3]Department of Applied Chemistry, Kyung Hee University, Yongin 17104, Korea

[4]Illumina, Inc., San Diego, CA 92122, USA

[†] These authors contributed equally to this work.

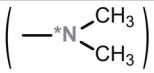
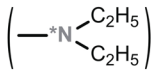
[§] Current address: Applied Bioinformatics, Department for Computer Science, University of Tübingen, Sand 14, 72076 Tübingen, Germany

*Correspondence to: jongseokim@snu.ac.kr and narrykim@snu.ac.kr

Table of Contents

Supplementary Figures

A

Label		The Maximum Multiplexity Achievable by Nominal Mass Differences, for Tryptic Peptide			
		Using ^{13}C, ^{15}N, ≥ 4 Da Mass Spacing	Using D(^{2}H), ^{13}C, ^{15}N, ≥ 4 Da Mass Spacing	Using ^{13}C, ^{15}N, ≥ 2 Da Mass Spacing	Using D(^{2}H), ^{13}C, ^{15}N, ≥ 2 Da Mass Spacing
Di-methyl	C_2H_6 	1	3	2	5
Di-ethyl	C_4H_{10} 	2	4	3	8
Lysine	$C_6H_9N_2H_3O$	3	5	5	9
Arginine	$C_6H_7N_4H_5O$	3	5	6	9

Gray color indicates atoms incapable of isotopic labeling. *N atoms are came from primary amine of labeled peptide.

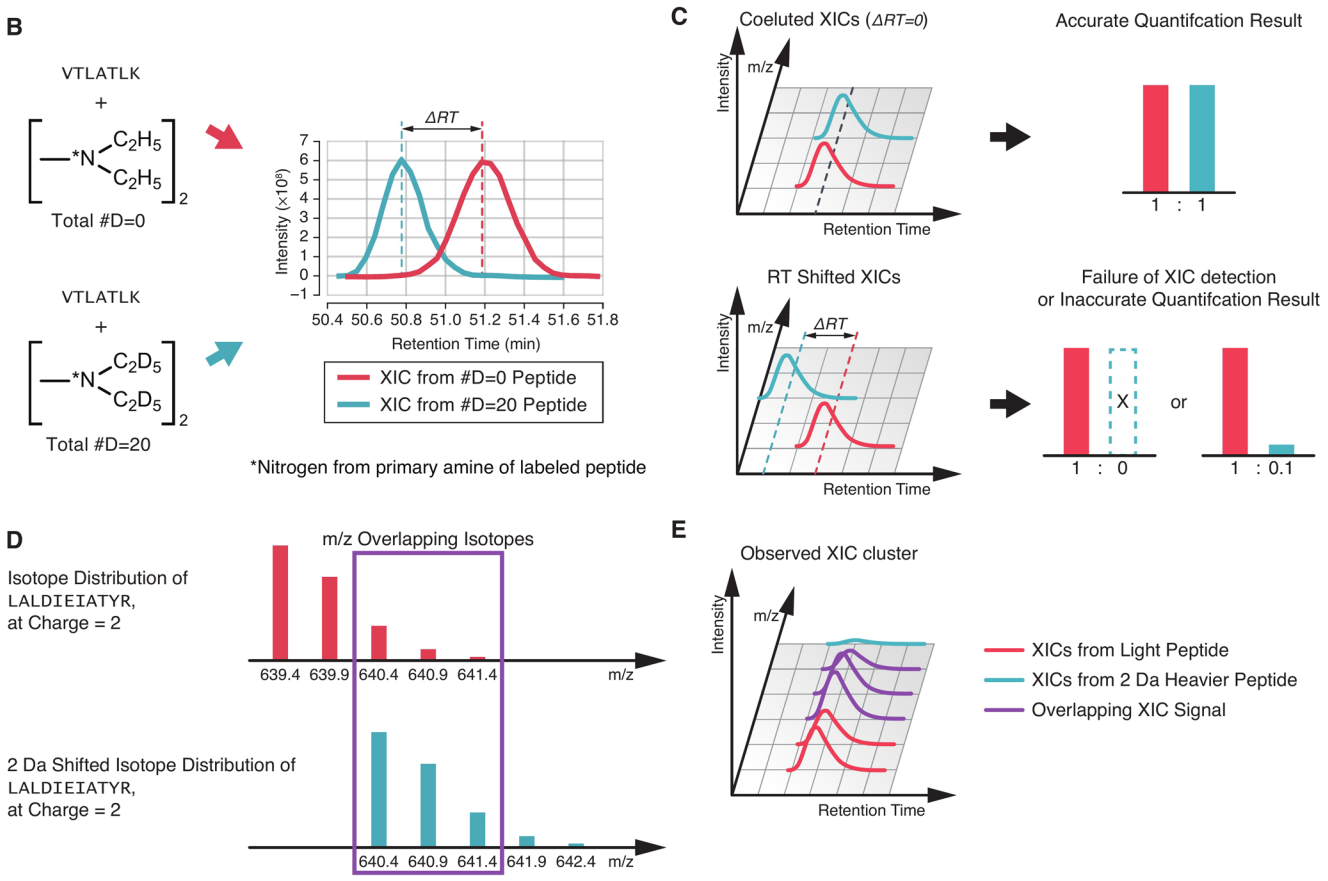


Figure S1. The use of deuteriums and small mass spacing enables far increased multiplexity, in exchange for complicated extracted ion chromatogram (XIC) signal. (A) The maximum multiplexity achievable by nominal mass differences for each type of stable isotopic label (tryptic digestion assumed). Note that far increased multiplexity can be achieved by allowing the use of deuteriums (marked as D (^{2}H)) and ≥2 Da mass spacing together. (B) Retention time (RT) shift for an example peptide VTLATLK. The XICs of non-deuterated (#D=0, red) and highly-deuterated (#D=20, green) peptides exhibit the shifted RT of 0.4 minutes. (C) Quantification of coeluted XICs (upper panel) and RT shifted XICs (lower panel). Unlike coeluted XICs, RT shifted XICs are hard to be detected or accurately quantified since the retention time positions of the XIC peaks are inconsistent. (D) Isotope distributions of an example peptide LALDIEIATYR (red bars) and its labeled counterpart (green

bars) when 2 Da label spacing is assumed. Two isotope distributions have three overlapped m/z values (purple box). Doubly charged peptide ions were shown and m/z values with isotope frequency less than 1% are ignored. (E) A schematic diagram of XIC signals from two forms of the example peptide LALDIEIATYR, with 2 Da mass spacing. XICs from light peptide isotopes (red) and 2 Da heavier peptide isotopes (green) are overlapped at three m/z values (purple).

A Combination of amino acid isotopologues for SILAC-6plex

	Lysine			Arginine		
Channel	**Minimum ΔMass (Da)**	**Isotopologue**	**Commercial Product #***	**Minimum ΔMass (Da)**	**Isotopologue**	**Commercial Product #***
1	+0	K000	ULM-8766	+0	R000	ULM-8347
2	+2	K002	NLM-1554	+2	R002	NLM-395
3	+4	K040	DLM-2640	+4	R004	NLM-396
4	+6	K600	CLM-2247	+7	R070	DLM-541
5	+8	K080	DLM-2641	+9	R072	In-house synthesized
6	+11	K092	DNLM-7545	+11	R074	DNLM-7543

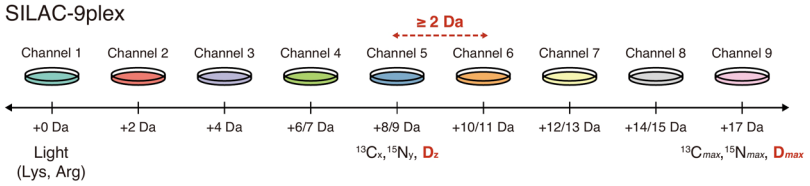
B

Channel	Total Arg PSM Count	Labeled Arg PSM Count	Arg Incorporation Efficiency (%)
2	9855	9725	**98.7**
3	7610	7541	**99.1**
4	7550	7488	**99.2**
5	6782	6691	**98.7**
6	7062	7017	**99.4**

C

Channel	Total Lys PSM Count	Labeled Lys PSM Count	Lys Incorporation Efficiency (%)
2	9065	9016	**99.5**
3	6690	6670	**99.7**
4	6865	6846	**99.7**
5	6438	6403	**99.5**
6	6323	6285	**99.4**

D SILAC-9plex



E Combination of amino acid isotopologues for SILAC-9plex

	Lysine			Arginine		
Channel	**Minimum ΔMass (Da)**	**Isotopologue**	**Commercial Product #***	**Minimum ΔMass (Da)**	**Isotopologue**	**Commercial Product #***
1	+0	K000	ULM-8766	+0	R000	ULM-8347
2	+2	K002	NLM-1554	+2	R002	NLM-395
3	+4	K040	DLM-2640	+4	R004	NLM-396
4	+6	K600	CLM-2247	+6	R600	CLM-2265
5	+8	K080	DLM-2641	+8	R170	In-house synthesized
6	+11	K092	DNLM-7545	+10	R604	DNLM-539
7	+13	K571	In-house synthesized	+12	R174	In-house synthesized
8	+15	K672	In-house synthesized	+14	R572	In-house synthesized
9	+17	K692	CDNLM-6810	+17	R674	CDNLM-6801

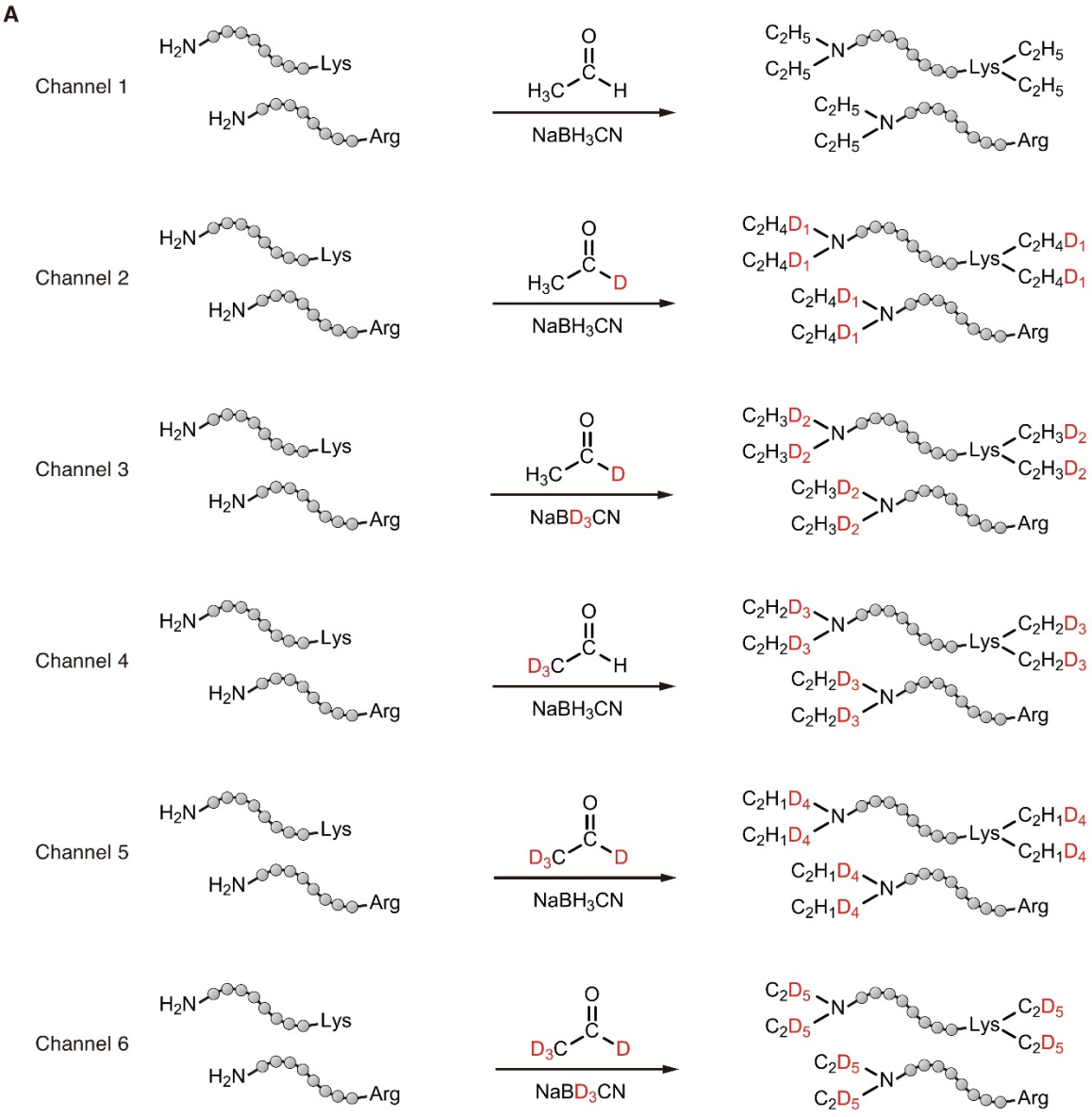
Commercially Available In-house Synthesized

Axyz notation for isotopologues: A: amino acid; x: the number of ^{13}C; y: the number of ^{2}H; z: the number of ^{15}N

*: Commercial products number of Cambridge Isotope Laboratories, Inc

Figure S2. Detail combination tables for SILAC-6plex and SILAC-9plex labeling, and isotope-incorporation efficiency in SILAC-6plex labeling. (A) Combination table of lysine and arginine isotopologues for SILAC-6plex in this paper within 11 Da of total mass window. (B-C) Isotope-incorporation efficiency for individual arginine (B) and lysine isotopologues (C) for SILAC-6plex labeling after > 5 cell cycle passages in the respective SILAC media. (D) Schematic diagram showing maximally-multiplexed SILAC-9plex within 17 Da of total mass window. (E) Combination table of lysine and arginine isotopologues for SILAC-9plex

A

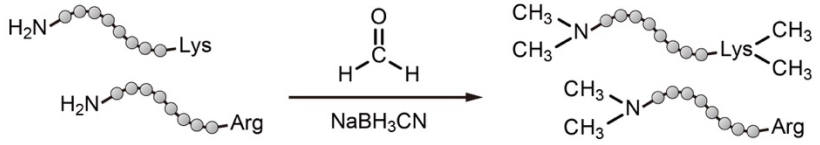


B

Channel	Total PSM Count	Fully Labeled PSM Count	Labeling Efficiency (%)
1	16227	15760	**97.1**
2	18406	17917	**97.3**
3	21278	20743	**97.5**
4	19550	19107	**97.7**
5	17213	16858	**97.9**
6	19258	18885	**98.1**

Figure S3. Labeling scheme of peptide di-ethylation 6plex (DE-6plex). (A) Reaction schemes for DE-6plex labeling of tryptic digests using deuterated isotopologues only. (B) Labeling efficiency of individual DE-6plex labeling channel in tryptic HeLa digests.

A DM-5plex Lageling Reaction Scheme



B DM-5plex Labeling Reagents

Channel	Minimum ΔMass (Da)	Reducing Reagent	Aldehyde	Label
1	+0	$NaBH_3CN$	CH_2O	$(CH_3)_2$
2	+2		$^{13}CH_2O$	$(^{13}CH_3)_2$
3	+4		CD_2O	$(CD_2H)_2$
4	+6	$NaBD_3CN$		$(CD_3)_2$
5	+8		$^{13}CD_2O$	$(^{13}CD_3)_2$

Figure S4. Di-methylation 5plex (DM-5plex) labeling scheme. (A) Reaction scheme for tryptic peptide di-methylation. (B) Table of reagent combination for maximally-multiplexed DM-5plex with 2 Da mass spacing.

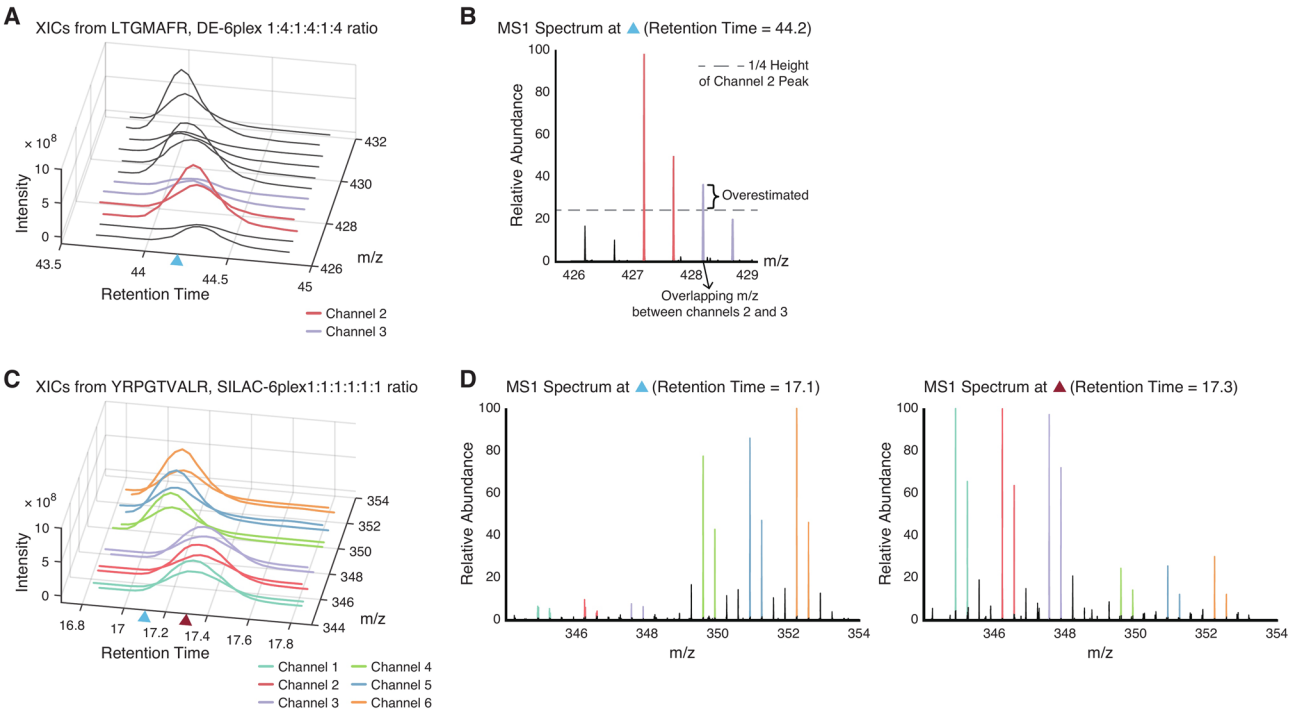


Figure S5. Example XICs and MS1 spectra of DE-6plex (A-B) and SILAC-6plex (C-D) labeled peptide. (A) XICs of DE-6plex labeled peptide LTGMAFR with 1:4:1:4:1:4 mixing ratio. Mass spacing between channels is 2 Da. (B) MS1 spectrum of (A) at retention time 44.2 min. The height of channel 3 peak exceeds the 1/4 height of channel 2, due to the isotope overlapping. (C) XIC of SILAC-6plex labeled peptide YRPGTVALR with 1:1:1:1:1:1 mixing ratio. (D) MS1 spectrum of (C) at retention 17.1 min (left) and 17.3 min (right). Owing to the retention time differences between channels, MS1 spectra do not represent the mixing ratio of the peptide.

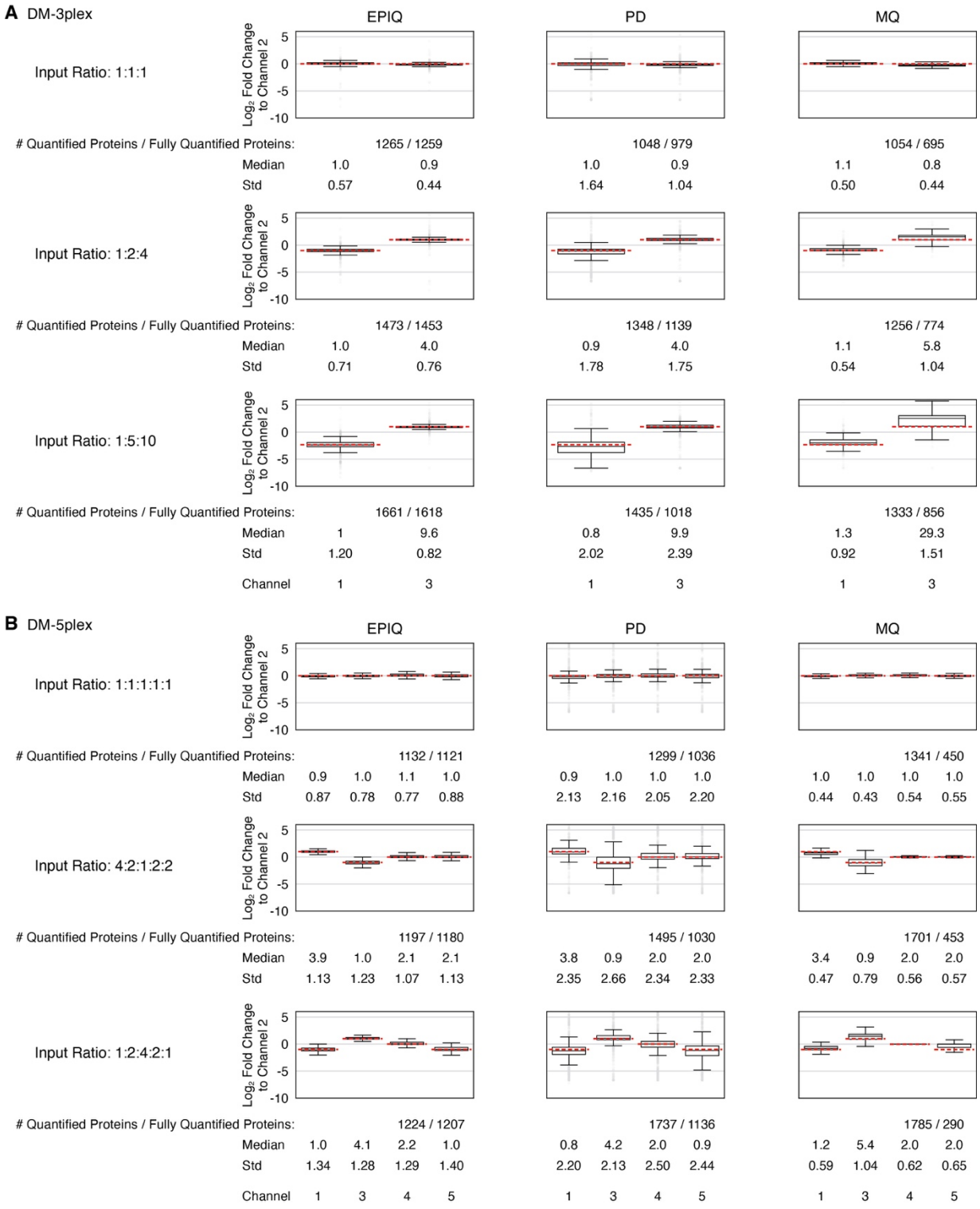


Figure S6. Benchmark tests with DM-3plex or DM-5plex labeled HeLa lysates. (A) The boxplot for protein-level fold changes from the quantities of channel 2, reported by EPIQ (left) and other tools (right), for DM-3plex-labeled HeLa lysate sample[1]. Red dashed lines specify the input ratio. The median (center line), first and third quartiles (lower and upper box limits, respectively), and 1.5 times the interquartile range (whiskers) are shown in boxplots. All quantified proteins were counted, where zero quantities were substituted by 1/100 of the maximum quantity per protein. (B) Analogs of (A) for the DM-5plex-labeled HeLa lysate sample.

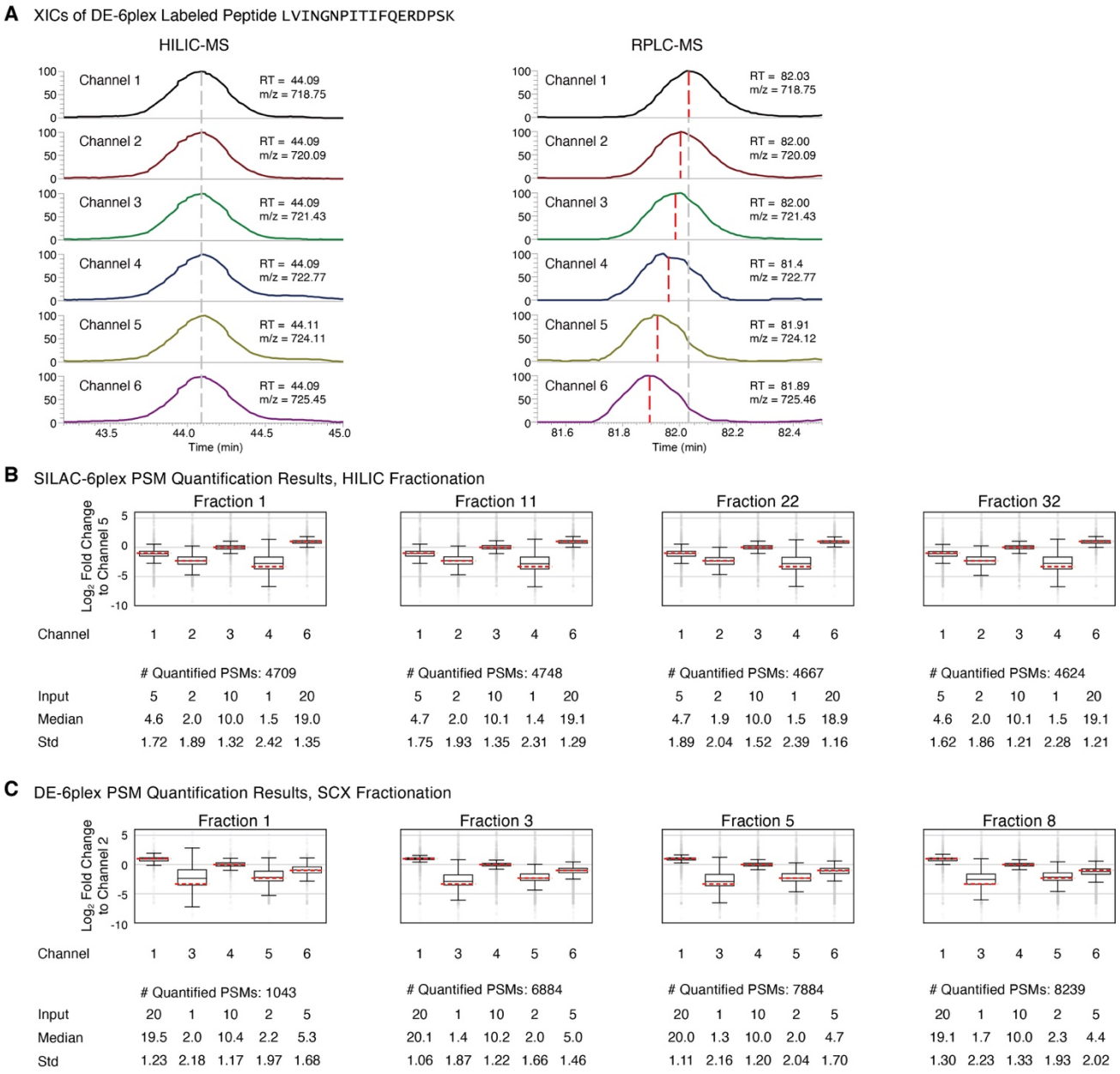


Figure S7. The RT shift-free character of non-RPLC fractionation approaches. The co-eluting character of differentially deuterated peptides in HILIC-MS in comparison with the RPLC-MS data showing significant RT shift (A) and the consistent quantification ratios between different fractions across the whole fractions for both HILIC (B) and SCX (C) fractionations.

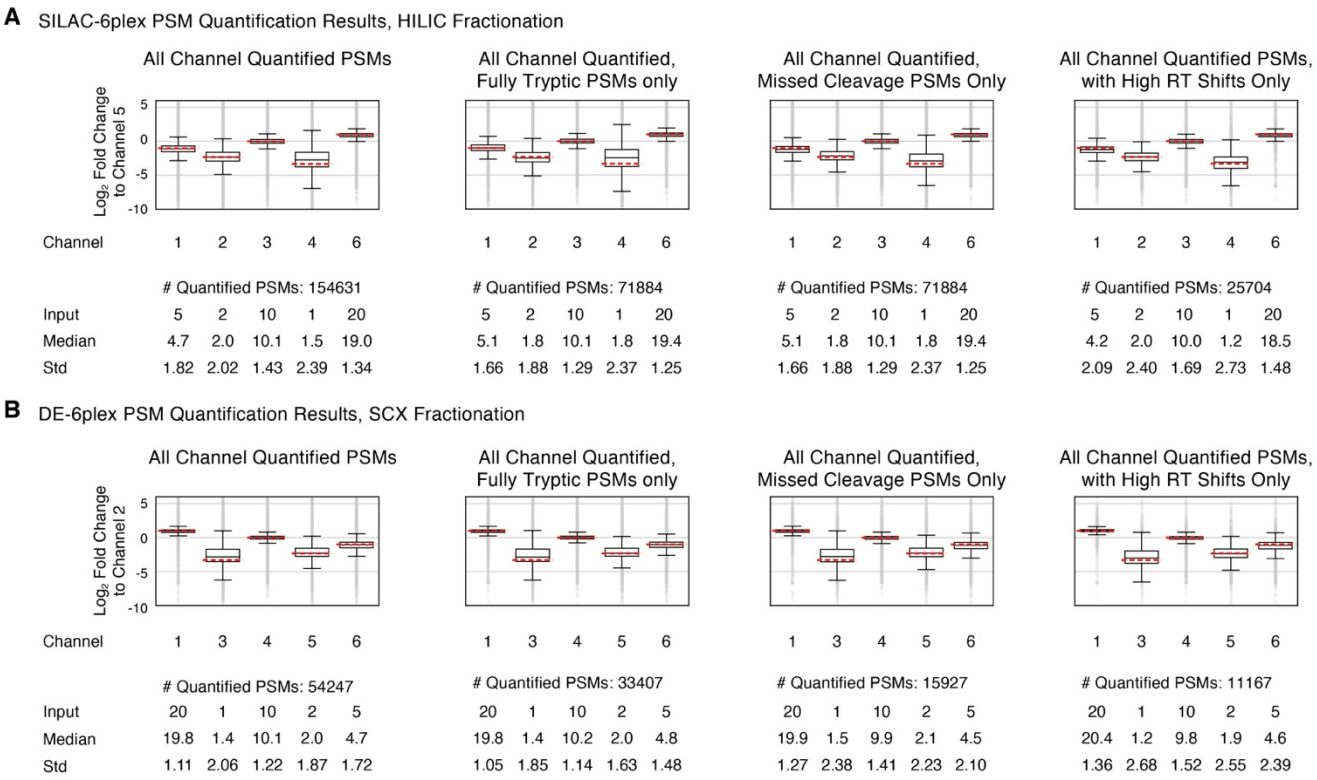


Figure S8. PSM-level quantification results (related to Fig. 3A-D). (A-B) The PSM-level fold change box plots for SILAC-6plex (A) and DE-6plex (B) labeled HeLa lysate sample. The first box plot shows the PSM-level quantification results for all PSMs. The second box plot is for the fully tryptic PSMs with 2-3 Da mass spacing to test whether EPIQ works well for narrow mass spacing. The third box plot is for the all missed-cleavage PSMs. The last box plot is drawn from the PSMs with at least 6 Da mass spacing to test whether EPIQ works well for labels with dozens of deuteriums, that is, causing high RT shifts. For these box plots, only all-channel-quantified PSMs were counted.

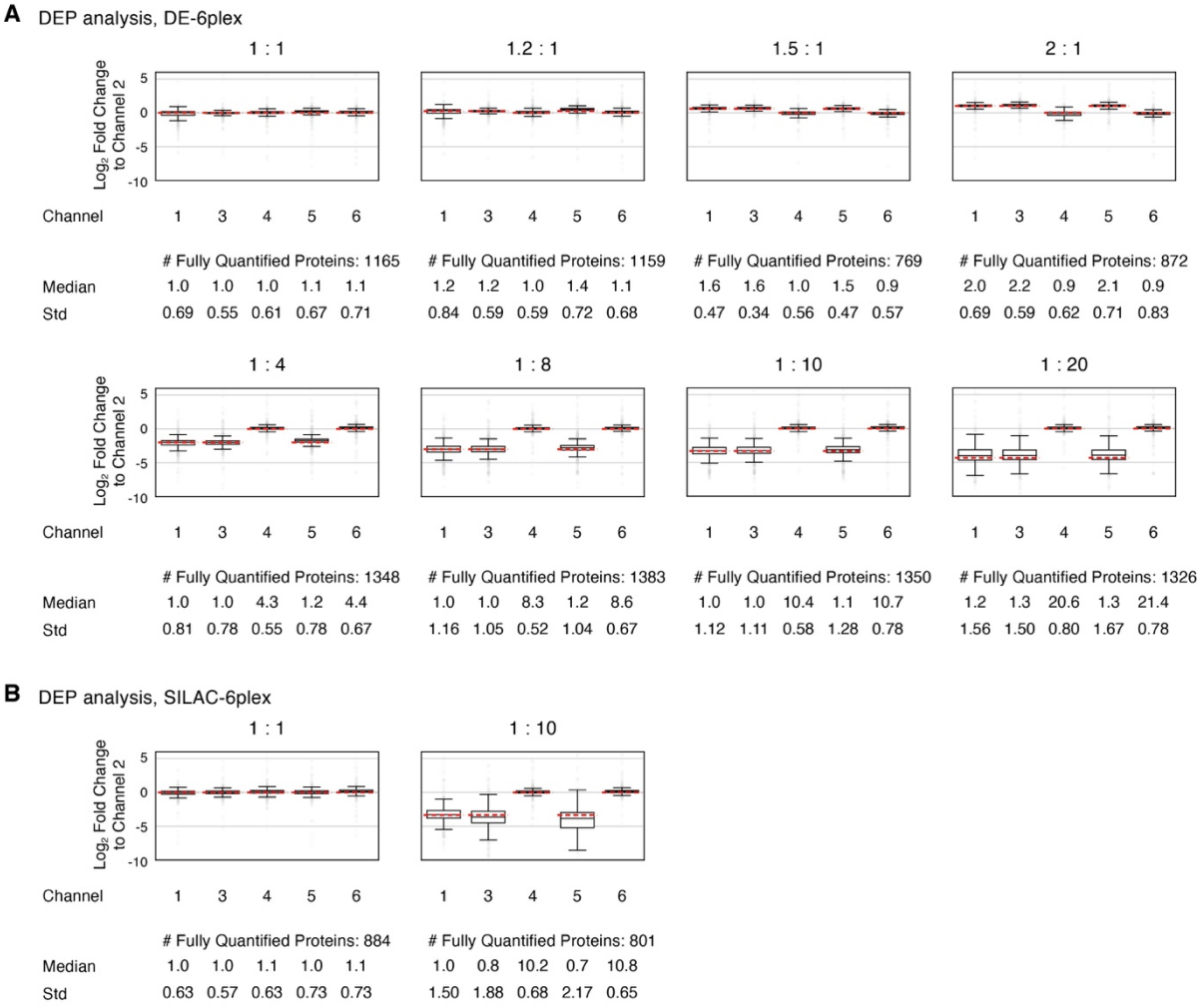


Figure S9. Protein-level fold change boxplots for differentially expressed protein (DEP) analyses (related to Fig. 3E-F). (A-B) The boxplots showing the protein-level fold changes from the label 2 quantities in DE-6plex (A) and SILAC-6plex (B) labeled samples. Red dashed lines specify the input ratio. Only all channel quantified proteins were counted.

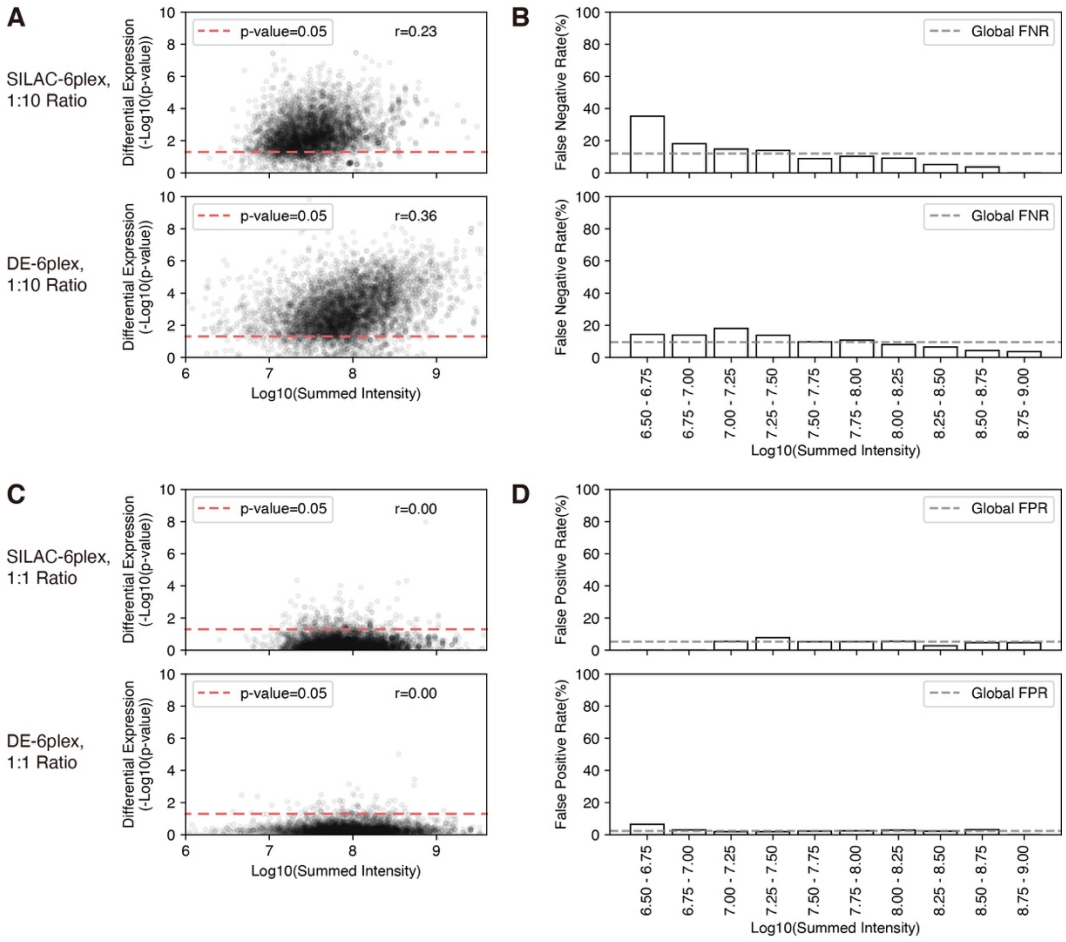
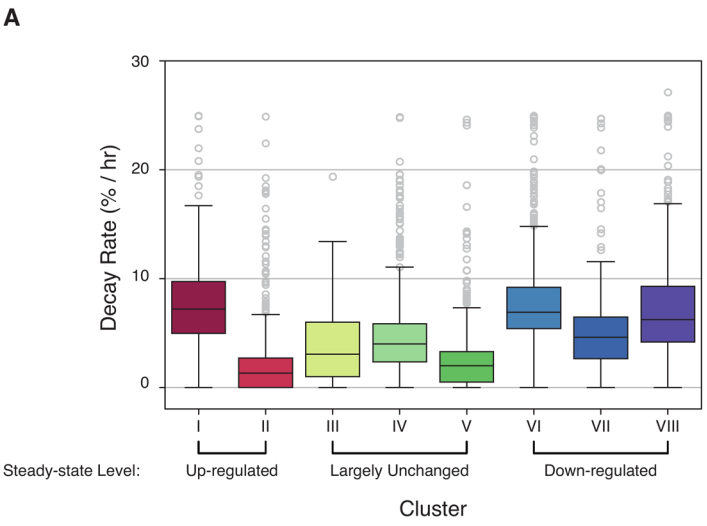


Figure S10. Association between PSM-level precursor intensity and differential expression p-value. (A) 1 to 10 and (B) 1 to 1 test to control triplicate samples used in Figure 3F were used. The intensities of all 6 channels were summed to represent the precursor intensity of each PSM. Scatter plots in the left column show the differential expression p-value (y-axis) and the precursor intensity (x-axis) for the quantified PSMs. Red dashed line shows the p-value of 0.05. Bar plots in the right column display the (A) false negative rate (FNR) and (B) false positive rate (FPR) at the p-value 0.05 cutoff, for each precursor intensity interval. Gray dashed lines indicates the global (A) FNR or (B) FPR among all quantified PSMs.



B

Example)

Less \ Greater	A	B
A	-	p-value of B > A
B	p-value of A > B	-

Decay Rate of Steady-state Level Up-regulated Clusters

Less \ Greater	Cluster I	Cluster II
Cluster I	-	1.0E+00
Cluster II	**2.5E-30**	-

Decay Rate of Steady-state Level Largely Unchanged Clusters

Less \ Greater	Cluster III	Cluster IV	Cluster V
Cluster III	-	8.6E-01	9.9E-01
Cluster IV	1.4E-01	-	1.0E+00
Cluster V	**5.2E-03**	**8.7E-43**	-

Decay Rate of Steady-state Level Down-regulated Clusters

Less \ Greater	Cluster VI	Cluster VII	Cluster VIII
Cluster VI	-	1.0E+00	8.1E-01
Cluster VII	**1.1E-11**	-	**9.6E-07**
Cluster VIII	1.9E-01	1.0E+00	-

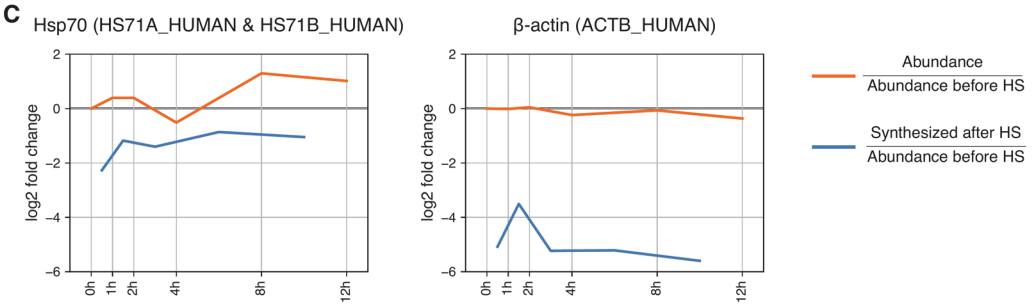


Figure S11. Protein dynamics during the heat shock response (HSR) of HeLa cells (related to Fig. 4.). (A) Boxplots showing the distribution of estimated decay rate for each cluster. The median (center line), first and third quartiles (lower and upper box limits, respectively), and 1.5 times the interquartile range (whiskers) are shown in boxplots. (B) p-values obtained by Welch's t-test between the estimated decay rates of compared clusters. Greater/Less indicates the alternative hypothesis; for example, ‘decay rate of group A is greater than that of group B’ is the hypothesis in the cell for ‘Greater A and Less B’. (C) Protein abundance and synthesis rate dynamics of Hsp70 and β-actin.

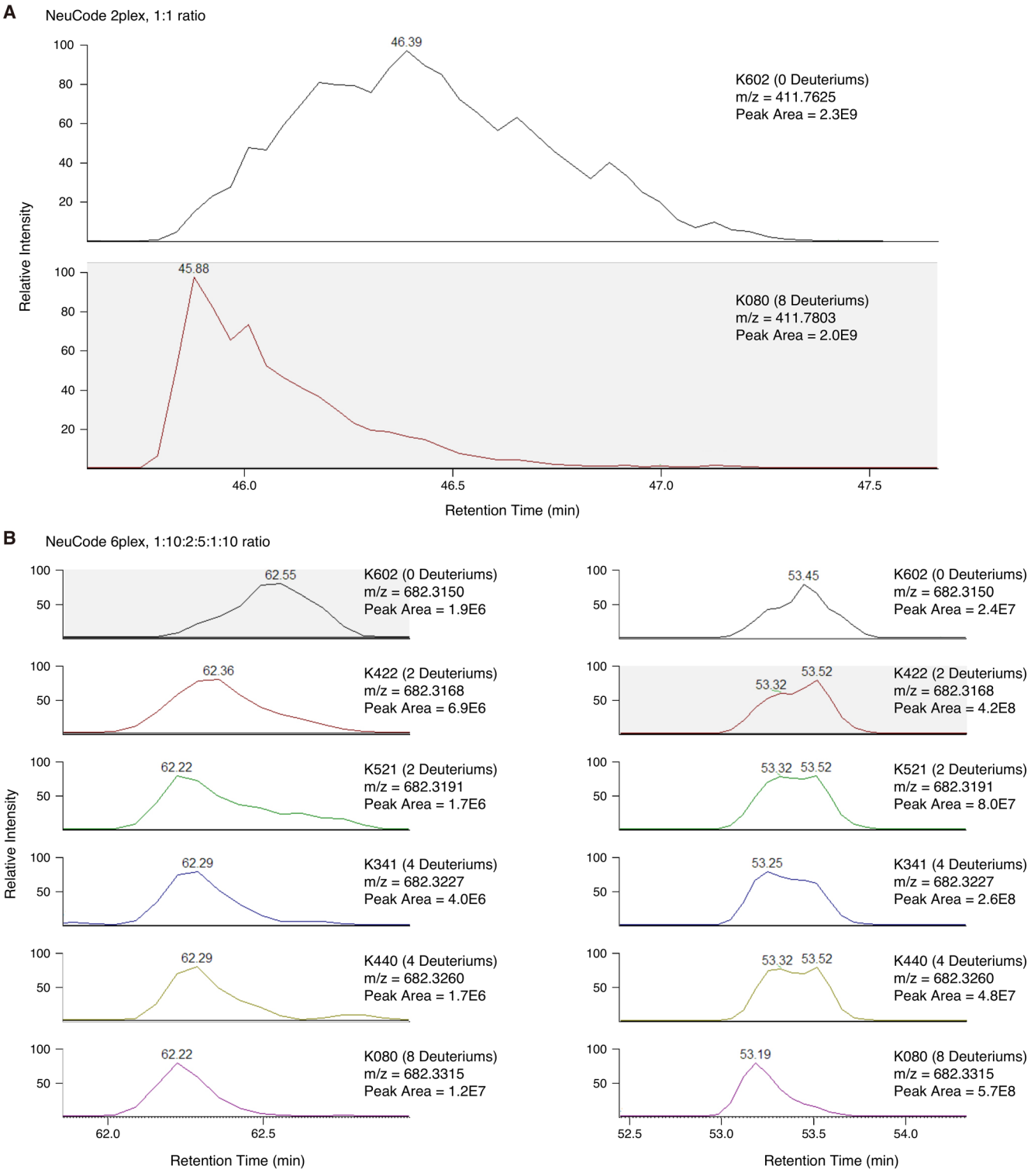


Figure S12. Retention time shift from differential deuteration in NeuCode SILAC (related to discussion). Representative extracted ion chromatograms from NeuCode SILAC 2plex (A) and 6plex (B) experiments in Merrill et. al. (2014)[2]. 30 and 15–20 seconds of retention time shift from differential deuteration were observed for NeuCode 2plex and 6plex labeling.

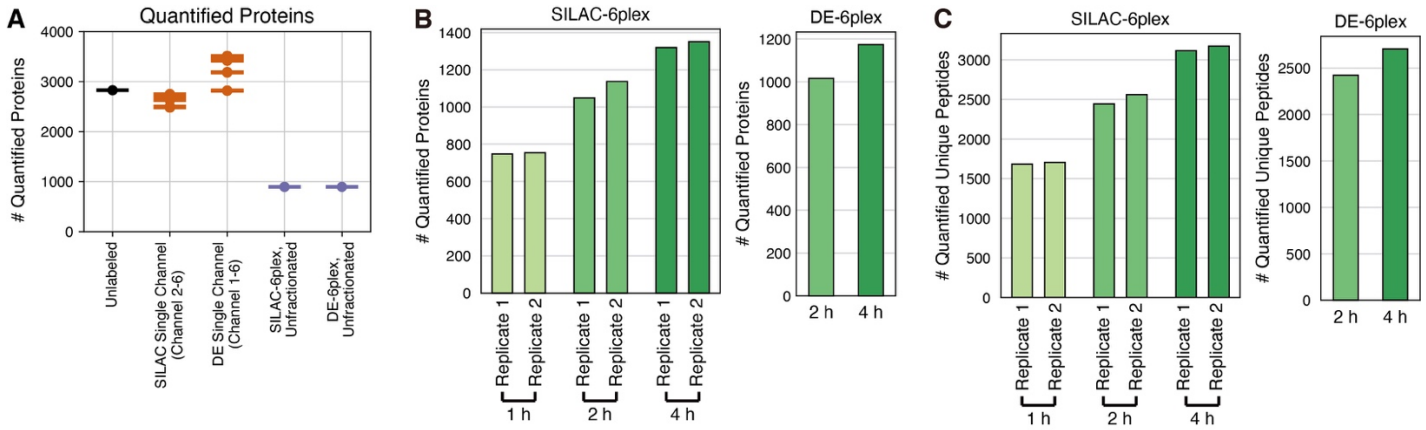


Figure S13. EPIQ number of quantified proteins / unique peptides across various conditions. (A) Number of quantified proteins in unlabeled, SILAC single channel, DE single channel, SILAC-6plex, and DE-6plex analysis. The 6plexed samples were mixed at the same ratio used in Figure 3. (B) Number of quantified proteins and (C) unique peptides in 1, 2, 4 h gradient SILAC-6plex sample and 2, 4 h gradient DE-6plex sample. For B and C, all samples were mixed at an equimolar ratio.

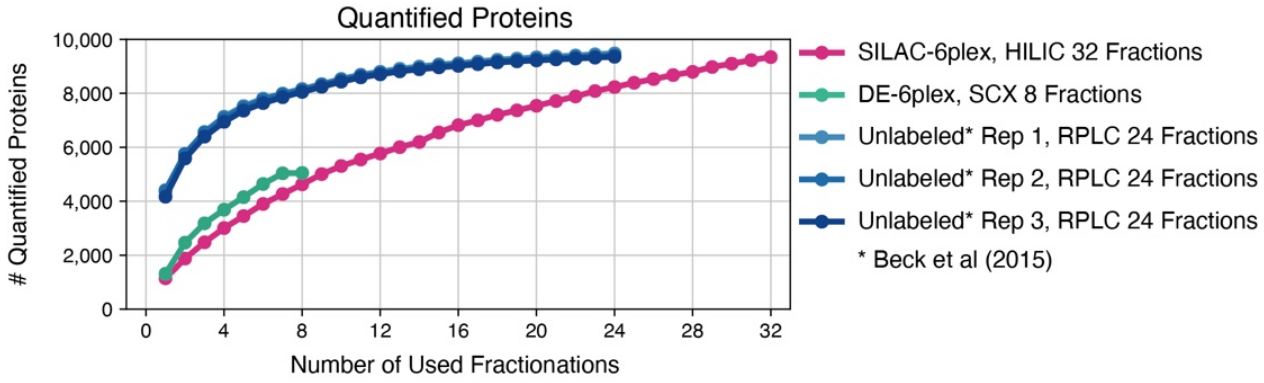


Figure S14. Number of quantified proteins from multidimensional LC-MS/MS analysis. The 6plexed samples were mixed at the same ratio used in Figure 3. The result of Beck et al.[3] was included as a reference result of current in-depth proteomics using the high-pH RPLC based fractionation in HeLa digest.

Supplementary Table List

Table S1. SILAC-6plex 5:2:10:1:10:20 proteins (related to Fig. 3A and Fig. 3C.)

Table S2. DE-6plex 20:10:1:10:2:5 proteins (related to Fig. 3B and Fig. 3D.)

Table S3. Medians and standard deviations of protein-level fold changes (related to Fig. 3C and Fig. 3D.)

Table S4. SILAC-6plex 20:10:1:10:2:5 PSMs, by EPIQ (related to Supplementary Fig. 8A.)

Table S5. DE-6plex 20:10:1:10:2:5 PSMs, by EPIQ (related to Supplementary Fig. 8B.)

Table S6. DEP analysis on protein groups, by EPIQ (related to Fig. 3E-F and Supplementary Fig. 9.)

Table S7. Analysis on protein dynamics of HeLa cells under HSR (related to Fig. 4)

Materials and Methods

Six-plexed SILAC media preparation and HeLa cell culture.

For six-plexed SILAC, lysine and arginine deficient DMEM (Dulbecco's modified Eagle's medium, Thermo Fisher) was used with supplement of isotopically distinct L-lysine:2HCl (K, 0.80 mM) and L-arginine:HCl (R, 0.40 mM) for 6 individual SILAC channels (Supplementary Fig. S2) . The detail ingredients of each SILAC channel are described as follows; light K and R from Cambridge Isotopes Laboratory (CIL) for channel 1, K-$^{15}N_2$ (NLM-1554, CIL) and R-$^{15}N_2$ (NLM-395, CIL) for channel 2, K-D_4 (NLM-2640, CIL) and R-$^{15}N_4$ (NLM-396, CIL) for channel 3, K-$^{13}C_6$ (CLM-2247, CIL) and R-D_7 (DLM-541, CIL) for channel 4, K-D_8 (DLM-2641, CIL) and R-$D_7{}^{15}N_2$ (in-house synthesized, see Supplementary Experiment Note) for channel 5, and K-D_9 (DNLM-7545, CIL) and R-$D_7{}^{15}N_4$ (DNLM-7543, CIL) for channel 6. After supplementing the appropriate lysine and arginine isotopologues, each customized SILAC medium was filtered through a 0.22 μm membrane (Merck) before adding dialyzed fetal bovine serum (FBS) to be 10 % (v/v). HeLa cell line from ATCC was cultured without antibiotics in the customized DMEM media to > 5 cell cycle passages. The normal DMEM (Welgene) containing 10 % FBS was used for HeLa cell culture as well, which cells were used for other labeling experiments.

Peptide sample preparation.

Harvested HeLa cells were lysed using 8 M urea in 50 mM ammonium bicarbonate (ABC) containing protease inhibitors (Pierce). The cysteine residues were alkylated with 40 mM iodoacetamide (Sigma) in the dark after reduction of the disulfide bond with 10 mM 1,4-dithiothreitol (Sigma). The denatured and alkylated samples in 8 M urea were then diluted with 50 mM ABC to less than 1 M of urea and digested by 2 % (w/w) of trypsin (MS grade, Pierce) at 37°C for overnight. The digested peptides were cleaned up using C18 SPE cartridge (SUPELCO). BCA assay was used for protein or peptide assay.

Six-plexed di-ethylation (DE) of tryptic peptides.

De-salted tryptic peptide samples were subject to reductive di-ethylation using acetaldehyde and sodium cyanoborohydride isotopologues. Only deuterated isotopologues were used for DE-6plex labeling, i.e., acetaldehyde-1-d_1 (CH_3CDO), acetaldehyde-2,2,2-d_3 (CD_3CHO), acetaldehyde-1,2,2,2-d_4 (CD_3CDO), and sodium cyanoborodeuteride ($NaBD_3CN$) all from CDN isotopes. The DE-6plex labeling was carried using the following reagent combinations: acetaldehyde (CH_3CHO) and $NaBH_3CN$ for channel 1, CH_3CDO and $NaBH_3CN$ for channel 2, CH_3CDO and $NaBD_3CN$ for channel 3, CD_3CHO and $NaBH_3CN$ for channel 4, CD_3CDO and $NaBH_3CN$ for channel 5, and CD_3CDO and $NaBD_3CN$ for channel 6 (Supplementary Fig. S3). In detail, the lyophilized peptides (30 μg) were dissolved in 100 μL of 100 mM sodium acetate buffer (pH 5.5) and followed by the addition of individual di-ethylating reagents for each channel to be 500 mM acetaldehyde form

and 250 mM sodium cyanoborohydride form with 2 hrs of mild vortexing at room temperature (RT). The identical di-ethylation procedure was carried again to reach quantitative labeling efficiency (> 97 %), and finally followed by quenching with 1 M ABC buffer [4, 5]. The labeled samples were mixed together with various desired input ratios for LC-MS/MS and further cleaned up using C18 cartridge column.

SCX-StageTip fractionation.

Multidimensional LC-MS/MS analysis based on strong-cation exchange (SCX) StageTip (STop-And-Go-Extraction TIPs) fractionation [6, 7] was performed to improve quantitative profiling depth of 6-plexed peptide mixture. Briefly, SCX-StageTip was prepared by stacking three layers of SCX disk (Empore, 3M) into 200 µL tip using 14-gauge syringe. The SCX-StageTip was activated by 100 µL of 100 % acetonitrile (ACN) via centrifugation at 1,000 g. Lyophilized six-plexed peptide sample (30 µg) was resolved with 100 µL of 1% trifluoroacetic acid (TFA) solution and then was loaded into the StageTip. After washing of sample with 0.2 % TFA solution three times, seven-stepwise salt gradient of potassium chloride (KCl) and final elution buffer were employed as follows; 50, 75, 125, 200, 275, 350, and 450 mM KCl in 30 % ACN containing 0.5 % formic acid, and 5 % ammonium hydroxide in 80 % ACN. The resulting eight fractions were subject to desalting process.

Concatenated hydrophilic interaction chromatography (HILIC) fractionation at micro-scale.

For SILAC 6-plex samples, a concatenated HILIC fractionation was carried out for multidimensional LC-MS/MS analysis to improve the quantitative profiling depth. For micro-scale fractionation, a HILIC capillary column (200 µm i.d. x 70 cm) was in-house packed with TSKgel Amide-80 particle (Tosoh, 3 µm)[8, 9]. 30 µg of 6-plexed SILAC peptide samples dissolved in 90% ACN with 0.1% TFA were loaded onto capillary column A linear gradient of solvent A (water with 0.1% TFA) and solvent B (ACN with 0.1% TFA) was applied on nanoAcquity (Waters) at a flow rate of 3 µL/min; 2 to 10 % solvent A for initial 2.5 min, 10 to 12 % solvent A for following 2 mins, 12 to 30% solvent A for next 80 min, 30 to 3 5% solvent A for 5 min and 35 to 80 % solvent A for final 1 min. The eluent was automatically concatenated into 24 or 32 fractions using TriVersa NanoMate (Advion).

LC-MS/MS analysis of labeled HeLa samples.

LC-MS/MS analysis of the labeled samples was carried by Orbitrap Fusion Lumos Tribrid or Q-Exactive Classic mass spectrometer (Thermo Fisher Scientific) coupled with Ultimate 3000 RSLCnano liquid chromatography (Thermo Fisher Scientific), which was equipped with an in-house packed trap (150 µm i.d. x 3 cm) and analytical column (75 µm i.d. x 100 cm) using 3 µm of Jupiter C18 particle (Phenomenex). A linear gradient of solvent A (water with 0.1% formic acid) and solvent B (ACN with 0.1% formic acid) was applied at a flow rate 350 nL/min as follows: 1) as for DE-6plex samples, 5 to 10 % solvent B for initial 5 min, 10 to 30 %

solvent B for next 90 min, and 30 to 40 % for final 20 min 2) as for SILAC-6plex, 5 to 10 % solvent B for initial 5 min, 10 to 30 % solvent B for next 100 min, and 30 to 40 % for final 10 min. Full MS scans (m/z 300 – 1,500) were acquired at a resolution of 60k (or 70k) at m/z 200 with 5E5 (or 1E6) of AGC target value and 50 ms (or 20 ms) of ITmax for Orbitrap Fusion Lumos (or Q-Exactive, respectively). Precursor ions with charge 2-7 were subject to HCD fragmentation under 30 % (or 27 % for Q-Exactive) of NCE via precursor isolation within 1.6 Th of window. The MS/MS scans were acquired at a resolution of 15k (or 17.5k) at m/z 200 with 30 ms (or 60ms) of ITmax and 3E4 (or 1E6) of AGC for for Orbitrap Fusion Lumos (or Q-Exactive, respectively). Dynamic exclusion value was set to 30 sec. 1 sec cycle time and top-12 setting were used for Orbitrap Fusion Lumos and Q-Exactive, respectively.

Used proteome databases for EPIQ (*.fasta files).

For all sample analyses, UniProt reference proteome UP000005640 (last modified on March 15, 2019) was used. Canonical and isoform sequence of all reviewed (SwissProt) and unreviewed (TrEMBL) proteins were downloaded. For the downloaded database, the cRAP (common Repository of Adventitious Proteins) protein sequences version 2012.01.01 (http://www.thegpm.org/crap/index.html) were first appended to remove common contaminant proteins. Decoy sequences were generated by reversing target protein sequences and further concatenated to the target proteome database to estimate FDR. After the searches using MS-GF+, the spectra matched to cRAP sequences are removed before further analysis.

EPIQ parameters.

The search tolerance and quantification tolerance were set to 20 ppm and 5 ppm, respectively. For MS-GF+ search, we set carbamidomethylation of cysteine as static modification, and oxidation of methionine as variable modification. For DE-6plex datasets, the label mass shifts were set at N-terminus of peptide and lysine residue to +56.06260, +58.07515, +60.08771, +62.10026, +64.11281, and +66.12537 from channel 1 to channel 6, respectively. As for SILAC-6plex datasets, the respective label mass shifts of arginine (R) and lysine (K) were assigned as follows; R+0.0/K+0.0 for channel 1, R+1.99407/K+1.99407 for channel 2, R+3.98814/K+4.02511 for channel 3, R+7.04394/K+6.02013 for channel 4, R+9.03801/K+8.05021 for channel 5, and R+11.03208/K+11.05056 for channel 6. For all analyses, we enforced 1% PSM and protein-level FDR threshold (estimated as recommended in [10] and [11]). Each PSM is assigned to a single protein group (see Supplementary Algorithm Notes), and the protein groups with less than two matched/quantified PSMs were removed. The SNR (signal-to-noise ratio) thresholds for peptide/protein quantification were set to 2.

Label impurity bias correction parameters for EPIQ.

PSM quantities were corrected for label impurity bias by procedures described in Supplementary Algorithm Notes. For label impurity correction of DE-6plex multidimensional analysis (Fig. 2 and Supplementary Fig.

S6b), -1 Da isotope abundance ratio of 0%, 1.7%, 4.6%, 8.4%, 6.9%, 9.7% (channel 1 to channel 6) were applied to correct label impurity bias. For DE-6plex DEP analysis (Fig. 2 and Supplementary Fig. S7a), 0%, 1.8%, 4.3%, 8.4%, 5.3%, 7.9% were used. For SILAC-6plex, label impurity of arginine and lysine were set separately. -1 Da isotope abundance ratio of 0%, 4.2%, 4.7%, 12.7%, 9.1%, 10.4% and 0%, 2.8%, 7.0%, 4.6%, 3.4%, 12% were used for arginine and lysine, respectively.

Label incorporation rate bias correction parameters for EPIQ.

For each SILAC-6plex PSM quantity, label incorporation rate bias was corrected as described in Supplementary Algorithm Notes. Per each arginine, quantity correction was done with following incorporation rates: 100%, 98.7%, 99.1%, 99.2%, 98.7%, 99.4% (channel 1 to channel 6). Per each lysine, PSM quantity correction with incorporation rates 100%, 99.5%, 99.7%, 99.7%, 99.5%, 99.4% (channel 1 to channel 6) was applied.

Proteome Discoverer (version 2.2.0.388) parameters.

The same reference proteome used for EPIQ was used for Proteome Discoverer analysis. PSM search was done by Sequest HT, with the same modification and label mass shift setting used for EPIQ. Percolator with default parameters was used for PSM level FDR control. By precursor ion quantifier, only unique peptides were selected and their abundances were calculated based on the area of the signal. Normalization and scaling options were turned off since the total protein amount varies among samples. Master proteins with the number of PSMs less than 2 or protein-level q-value greater than 0.01 were filtered out. Zero quantities in protein abundances were substituted by 1/100 of the maximum quantity per protein, and a master protein was discarded if all abundances are zero.

MaxQuant (version 1.5.3.30) parameters.

The same reference proteome used for EPIQ was used for MaxQuant analysis. Contaminant database embedded in MaxQuant was appended to search space by allowing ‘Include contaminants’ option. For the PSM search, peptide modification and label mass shift setting were the same as EPIQ parameters. Because the current version of MaxQuant supports up to 3 multiplexity, each LC-MS/MS 6-plex dataset was analyzed three times with the following channel sets: channels 1, 2, 3; 2, 4, 5; and 2, 6. Note that all channel sets have channel 2 information that is used to combine fold changes.

To combine the results, finding common protein groups between different runs should be carried out. But because the protein group definition depends on the MS/MS ID results, intersecting protein groups from MaxQuant analyses with different channel configurations is a complicated task. For the sake of simplicity, the representative protein of each protein group was defined as a protein with the largest number of identified peptides. Then the intersection of different MaxQuant runs was done with the representative proteins.

Protein ratios for each channel versus channel 2 were calculated from corresponding 'Ratio M/L', 'Ratio H/M', and 'Ratio H/L' values of MaxQuant output. If the ratios of some channels are missing in MaxQuant results, ratios were calculated from 'Intensity L', 'Intensity M', and 'Intensity H' values instead. If both the ratio and the intensity are unavailable for a channel, the intensity was substituted by 1/100 of the maximum quantity per representative protein. Protein groups with non-positive intensities in all channels were discarded.
Max. labeled AAs parameter was set to 3. Re-quantify option was turned on as the provider recommended for quantifying high ratio samples. Only unique peptides were included in quantification. Protein groups with the number of MS/MS IDs smaller than 2 or q-value greater than 0.01 were discarded. Protein groups comprising only contaminant proteins or containing at least one decoy protein were also removed.

Differentially expressed protein (DEP) analysis.

For multiplexed 3 by 3 t-test DEP analysis (Fig. 2f red line), the two-tailed two-sample t-test was done between log intensities of control and test triplets. Proteins with p-value lower than 0.05 were declared as DEPs. For non-multiplexed (2 channel) DEP determination (Fig. 2f grey and black lines), the fold change per protein was calculated as the ratio between intensities of a test channel and a control channel, chosen randomly. If the absolute log fold-change between these two channels is higher than the log of a given fold-change threshold, a protein was classified as differentially expressed.

Heat shock experimental procedures.

For steady-state SILAC-6plex, fully labeled Hela cells with six channels was incubated at 43°C in each following time; channel 1 - 0 hour (no heat shock control), channel 2 - 1 hour, channel 3 - 2 hours, channel 4 - 4 hours, channel 5 - 8 hours and channel 6 - 12 hours. For pulsed SILAC experiment, culture media of unlabeled Hela cells were exchanged with channel 2 media and cultured at 43°C for 1 hour. After 1 hour incubation, culture media was exchanged into channel 3 media and cultured for another 1 hour at 43°C then, channel 3 media was replaced to channel 4 culture media for further 2 hour incubation at 43°C. Channel 4 media was replaced with channel 5 culture media and incubated for 4 hours. Finally the culture media was replaced with channel 6 media for the final 4 hour incubation at 43°C prior to cell harvest.

Computational analysis on heat shock response of HeLa cells.

Steady-state and pulsed SILAC-6plex spectra were identified and quantified as described above. Steady-state SILAC-6plex protein quantities were normalized by total PSM quantities of each sample to compensate variations among cell culture dishes. On the other hand, we did not perform any between-sample normalization for pulsed SILAC-6plex experiment, because the all pulsed SILAC-6plex channels were obtained from the same cell population. We merged the intersected protein groups of steady-state SILAC and pulsed SILAC data by the representative protein names. Protein groups with less than 2 quantified unique peptides in any of steady-

state SILAC or pulsed SILAC experiment were discarded. To see the abundance dynamics of protein group, we $\log_2$ transformed steady-state SILAC quantities and normalized them by the median quantity of each protein group. For the synthesis rate of protein group, we calculated $\log_2$ fold changes between each pulsed channel and no heat shock control channel, to see the relative protein synthesis rate compared to the steady-state protein level before heat shock. Prior to cluster analysis, the protein abundances and the protein synthesis rates were z-score transformed for feature standardization. The k-means clustering was done by the SciPy package (http://www.scipy.org). GO term enrichment analysis was done by g:profiler, with custom background gene set containing all quantified protein groups in steady-state SILAC and pulsed SILAC (n=9158).

Protein decay rate estimation.

To calculate the decay rate from the steady-state and pulsed SILAC quantities, we adopted discretized protein concentration kinetics model assuming the constant protein change between the sample time points [12]. Also, for simplicity, we assumed that the decay rate of a protein group is the same throughout all time points. With these rational, we formulated the kinetics model per protein group as the following system of equations:

$$T_{n+1} = T_n + i_{n+1}(s_{n+1} - dT_n) \qquad \text{- (1)}$$

$$NP_n = s_n i_n \prod_{k=n+1}^{5}(1 - di_k) \qquad \text{- (2)}$$

$$NP_0 = T_0 \prod_{k=1}^{5}(1 - di_k) \qquad \text{- (3)}$$

where

T_n $(n = 0, 1, 2, \ldots, 5)$ is the observed steady-state protein quantity at each time point

P_n $(n = 1, 2, \ldots, 5)$is the observed pulsed SILAC quantity, which labeled between the time point of T_{n-1} and T_n

P_0 is the observed before heat shock-protein quantity in pulsed SILAC experiment

i_n is the time interval between the time point of T_{n-1} and T_n

s_n is the synthesis rate of a protein group during the time point of T_{n-1} and T_n

d is the decay rate of a protein group

$N$is the normalizing factor between steady-state and pulsed SILAC quantities

Equation (1) models the steady-state protein level change between the sample time points. Equation (2) and (3) model the decay of pulse-labeled protein and protein synthesized before heat shock, respectively. To solve the above system of equations, the fsolve function of the SciPy package (http://www.scipy.org) was used with the following initial values:

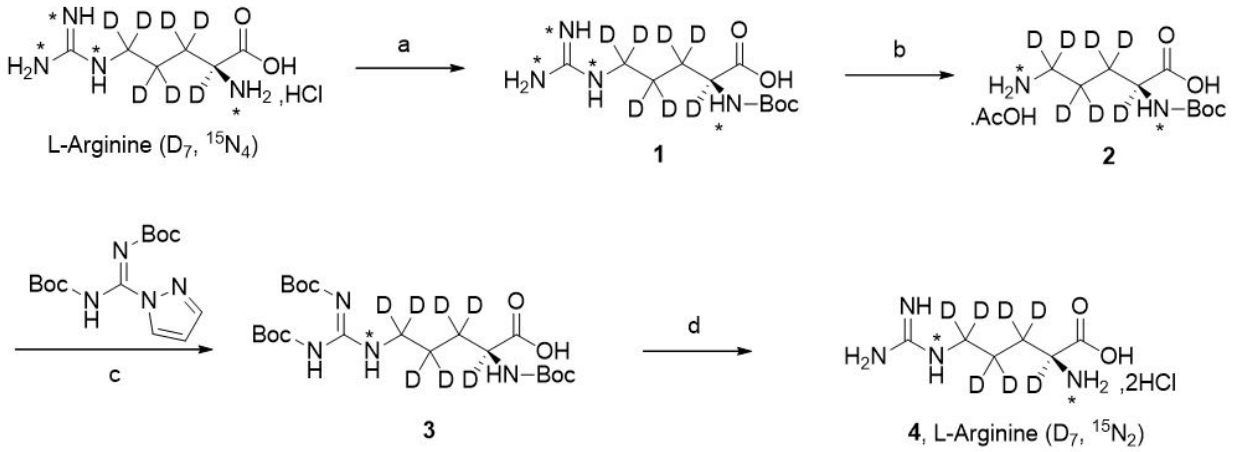
$$d_{init} = 0.2 \qquad \text{- (4)}$$

$$N_{init} = median\left(\frac{T_0}{P_0}\right) \prod_{k=1}^{5}(1 - d_{init} i_k) \qquad \text{- (5)}$$

$$s_{n\_init} = \frac{N_{init} P_n}{i_n} \prod_{k=n+1}^{5}(1 - d_{init} i_k) \qquad \text{- (6)}$$

where $median\ (x)$ indicates the median value of x among all quantified protein groups

Synthesis of R-9 form (L-Arginine-D_7,$^{15}N_2$) for SILAC channel 5



The L-Arginine-D_7,$^{15}N_2$ (**4**) was in-house synthesized via four steps described in above scheme. The synthesis procedures and characterization results for three intermediates (**1-3**) were described separately in the following. In summary, L-Arginine (D_7,$^{15}N_4$), the starting material purchased from Cambridge Isotope Laboratory, was converted to Boc-L-Arginine (D_7,$^{15}N_4$) (**1**), and then decomposed to Boc-L-Orthinine (D_7,$^{15}N_2$) (**2**), and finally restored to arginine form, Boc-L-Arginine (Di-Boc) (D_7,$^{15}N_2$) (**3**).

(1) Synthesis of Boc-L-Arginine (D_7,$^{15}N_4$)

L-Arginine (D_7,$^{15}N_4$) hydrochloride (1.39 g, 6.3 mmol, CIL) was dissolved in 25 mL of water, and sodium bicarbonate (1.58 g, 18.8 mmol) was added followed by di-*tert*-butyl dicarbonate (1.64 g, 7.5 mmol) in acetonitrile (25 mL). The solution was stirred for 18 hr at room temperature and the solvent was evaporated under reduced pressure. The residue was added ethanol and insoluble solid was filter-out. The filtrate was evaporated under reduced pressure to afford 1.8 g Boc-L-Arginine (D_7, $^{15}N_4$) as white solid. The obtained Boc-L-Arginine (D_7, $^{15}N_4$) used without more purification. ^{1}H NMR (400 MHz, DMSO-d_6) d = 8.64 (d, 1H), 7.45 (d, 4H), 6.37 (d, 1H), 1.37 (s, 9H); ^{13}C NMR (100 MHz, DMSO-d_6) d = 175.37 (t), 157.26 (q), 155.09 (d), 77.62, 64.97, 56.05, 28.26, 18.60, 15.22; MS (ESI+) m/z (%) 286.4 ($[M+H]^+$, (100)), 308.4 ($[M+Na]^+$, (10)).

(2) Synthesis of Boc-L-Orthinine (D_7,$^{15}N_2$)

A solution of Boc-L-Arginine (D_7,$^{15}N_4$) (1.8 g) and hydrazine (13.3 mL, 70 % aqueous solution) was heated for 3 hr at 70 ^{o}C. After concentration under reduced pressure, purification was performed by silica-gel chromatography (EA/n-BuOH/AcOH/Water = 4/1/1/1). The product contained fractions were combined, and the solvent was evaporated under reduced pressure. Excess diethyl ether was added and precipitated white solid was filtered and washed with diethyl ether to afford 1.635 g (86 %, 2 steps) Boc-L-Orthinine (D_7,$^{15}N_2$) as a acetic acid salt. ^{1}H NMR (400 MHz, DMSO-d_6) d = 5.95 (d, 1H), 1.78 (s, 3H), 1.37 (s, 9H); ^{13}C NMR (100

MHz, Methanol-d_4) d = 178.79 (br s), 177.88, 157.73 (d), 80.20, 62.66, 35.81, 28.75, 22.50, 20.02, 14.24; MS (ESI+) m/z (%) 242.2 ($[M+H]^+$, (100)), 264.3 ($[M+Na]^+$, (25)).

(3) Synthesis of Boc-L-Arginine (Di-Boc) (D_7,$^{15}N_2$)

Boc-L-Orthinine (D_7, $^{15}N_2$) (1.62 g, 5.4 mmol) was dissolved in 50 mL of methanol, and diisopropylethylamine (2.08 g, 16.1 mmol) was added followed by *N,N'*-Di-Boc-1*H*-pyrazole-1-carboxamidine (1.92 g, 6.2 mmol). The solution was stirred for 3hr at room temperature and the solvent was evaporated under reduced pressure. The mixture was diluted with ethyl acetate and HCl solution (1 N, aq.). The organic layer was separated and washed with brine. The organics dried over anhydrous Na_2SO_4 and concentrated. The residue was purified by silica-gel column chromatography (5 % MeOH/CH_2Cl_2) to afford 1.65 g (63 %) of Boc-L-Arginine (Di-Boc) (D_7, $^{15}N_2$) as white foamy solid. ^{1}H NMR (400 MHz, $CDCl_3$) *d* = 8.42 (d, 1H), 5.34 (d, 1H), 1.49 (s, 9H), 1.48 (s, 9H), 1.45 (s, 9H); ^{13}C NMR (100 MHz, $CDCl_3$) d = 175.58 (t), 163.18, 163.02, 156.32 (d), 155.74 (d), 153.24, 83.40, 80.02, 79.76, 28.39, 28.25, 28.10; MS (ESI+) m/z (%) 184.3 ($[M-3Boc]^+$, (10)), 284.4 ($[M-2Boc]^+$, (40)), 384.4 ($[M-Boc]^+$, (100)), 484.5 ($[M+H]^+$, (50)).

(4) Synthesis of L-Arginine-D_7,$^{15}N_2$

Boc-L-Arginine (Di-Boc) (D_7,$^{15}N_2$) (1.64 g, 3.4 mmol) was dissolved in 30 mL of HCl solution (6 N, aqueous) and was heated for 3 hr at 60 ºC. The mixture was concentrated under reduced pressure. The residue was dissolved in 4 mL of ethanol and was added 50 mL of ethyl acetate. The precipitated solid was filtered and washed ethyl acetate. The solid was dissolved in water and free-dried to afford 745 mg (86 %) of L-Arginine-D_7,$^{15}N_2$ as two hydrochloride salt. ^{1}H NMR (400 MHz, DMSO-d_6) *d* = 8.36 (br s), 7.91(d, 1H), 7.44 (br s), 7.11 (br s); ^{13}C NMR (100 MHz, Methanol-d_4) d = 171.94 (t), 158.62 (dt), 53.29, 41.02, 27.65, 24.71; MS (ESI+) m/z (%) 184.3 ($[M+H]^+$, (100)).

Supplementary Figures & Materials and Methods References

1. Jung, J.; Jeong, K.; Choi, Y.; Kim, S. A.; Kim, H.; Lee, J. W.; Kim, V. N.; Kim, K. P.; Kim, J. S., Deuterium-Free, Three-Plexed Peptide Diethylation for Highly Accurate Quantitative Proteomics. *J Proteome Res* **2019,** *18* (3), 1078-1087.
2. Merrill, A. E.; Hebert, A. S.; MacGilvray, M. E.; Rose, C. M.; Bailey, D. J.; Bradley, J. C.; Wood, W. W.; El Masri, M.; Westphall, M. S.; Gasch, A. P.; Coon, J. J., NeuCode labels for relative protein quantification. *Mol Cell Proteomics* **2014,** *13* (9), 2503-12.
3. Beck, S.; Michalski, A.; Raether, O.; Lubeck, M.; Kaspar, S.; Goedecke, N.; Baessmann, C.; Hornburg, D.; Meier, F.; Paron, I.; Kulak, N. A.; Cox, J.; Mann, M., The Impact II, a Very High-Resolution Quadrupole Time-of-Flight Instrument (QTOF) for Deep Shotgun Proteomics. *Mol Cell Proteomics* **2015,** *14* (7), 2014-29.
4. Kleifeld, O.; Doucet, A.; auf dem Keller, U.; Prudova, A.; Schilling, O.; Kainthan, R. K.; Starr, A. E.; Foster, L. J.; Kizhakkedathu, J. N.; Overall, C. M., Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nat Biotechnol* **2010,** *28* (3), 281-8.
5. Kleifeld, O.; Doucet, A.; Prudova, A.; auf dem Keller, U.; Gioia, M.; Kizhakkedathu, J. N.; Overall, C. M., Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates. *Nat Protoc* **2011,** *6* (10), 1578-611.
6. Kulak, N. A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M., Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat Methods* **2014,** *11* (3), 319-24.
7. Rappsilber, J.; Mann, M.; Ishihama, Y., Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* **2007,** *2* (8), 1896-906.
8. Guo, K.; Ji, C.; Li, L., Stable-isotope dimethylation labeling combined with LC-ESI MS for quantification of amine-containing metabolites in biological samples. *Anal Chem* **2007,** *79* (22), 8631-8.
9. Wu, C. J.; Chen, Y. W.; Tai, J. H.; Chen, S. H., Quantitative phosphoproteomics studies using stable isotope dimethyl labeling coupled with IMAC-HILIC-nanoLC-MS/MS for estrogen-induced transcriptional regulation. *J Proteome Res* **2011,** *10* (3), 1088-97.
10. Jeong, K.; Kim, S.; Bandeira, N., False discovery rates in spectral identification. *BMC Bioinformatics* **2012,** *13 Suppl 16*, S2.
11. Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A.; Working Group on Publication Guidelines for, P.; Protein Identification, D., The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol Cell Proteomics* **2004,** *3* (6), 531-3.
12. Teo, G.; Vogel, C.; Ghosh, D.; Kim, S.; Choi, H., PECA: a novel statistical tool for deconvoluting time-dependent gene expression regulation. *J Proteome Res* **2014,** *13* (1), 29-37.

Supplementary Algorithm Notes

EPIQ pipeline

EPIQ is composed of three steps: spectrum identification, PSM (Peptide-Spectrum Matches) level quantification, and protein level quantification. In this Supplemental Algorithm Notes, each step of EPIQ is described in detail. In the section SAN1, the spectrum identification step is described. In the section SAN2, PSM level quantification via the model-based reconstruction algorithm is presented. The protein level quantification is explained in the section SAN3. The section SAN4 explains sub-algorithms of the model-based reconstruction algorithm. In particular, RT (Retention Time) shift prediction method (the main sub-algorithm of the model-based reconstruction algorithm) is presented in the section SAN4.1.

SAN1 Spectrum identification

To find reliable PSMs, we adopt MS-GF+ [1]. We ran MS-GF+ per *channel* (the index specifying the peptide form from the lightest to the heaviest; e.g., in SILAC, the light form has channel 1 and the heavy channel 2), treating the mass shift of each channel as a fixed modification. After the PSMs are reported from each MS-GF+ run, the PSMs of spectral E-value higher than 1×10^{-8} (a user defined parameter) are filtered out. Note that each PSM conveys the information on not only the peptide sequence but also the channel. We do not discard decoy PSMs at this step as they are to be used to estimate FDR (false discovery rate) afterwards.

SAN2 Model-based reconstruction algorithm for PSM level quantification

Here we describe our model-based reconstruction algorithm that carries out the PSM level quantification. Before presenting the algorithm, we introduce necessary terms first.

Given a peptide P, denote the monoisotopic peptide ion of charge state z and of channel i by P_i^z. The set of P_i^z for all channels $i = 1, 2, \cdots, n$ is written as P^z where n is the number of channels. A *(chromatographic) intensity* on an m/z (mass-to-charge ratio) m and RT t is the observed intensity on the m/z m and RT t, which is proportional to the (summed) abundance of the eluted ion(s) having the m/z m and RT t. The m/z of an ion is determined by the ion mass divided by the ion charge. The *XIC from an ion* is the series of the intensities that are

generated by the ion eluted along the RT axis. Natural isotopologues are considered to be distinct ions. Thus, the XIC from an ion consists of the intensities of the same m/z (within the instrument-specific tolerance), called the *m/z of the XIC*. The intensities in an XIC usually form a bell-shaped peak along the RT axis, and the total area under the bell-shaped XIC is proportional to the abundance of the ion (unless other ions coelute). We define a *component* for P_i^z as the set of the XICs from P_i^z and its natural isotopologues. The *XIC cluster of a peptide ions* P^z is the superimposed collection of the components for P_i^z for $i = 1, \cdots, n$. An XIC cluster, thus, contains multiple bell-shaped (or superimposed bell-shaped) XICs of different m/z values. The *quantity* for channel i and charge z is the summation of the intensities of the component for P_i^z and written as q_i^z.

SAN2.1 XIC cluster generative model

Now we present the XIC cluster generative model. For simplicity, we assume that the m/z and the RT have integer values. We further assume that only two channels exist ($i = 1, 2$) and all ions are singly charged ($z = 1$). The mass spacing is given by 2Da. While the following description is for this simplified case, the description for the practical cases is found at the end of this section.

We focus on the generation of the XIC cluster for P_1^1 and P_2^1 or simply P_1 and P_2 (as the charge state is fixed) with the quantities q_1 and q_2. The model assumes that the following three elements are given from P_i per channel i: i) the position of the monoisotopic XIC from P_i (the m/z and RT position of the apex intensity of the XIC), ii) the shape of the XIC from P_i, and iii) the isotope distribution of P_i. The position consists of m/z position (the m/z of the monoisotopic XIC from P_i) and RT position (the RT of the apex of the monoisotopic XIC from P_i).

Assume the elements for P_1 and P_2 are furnished (by the model) as follows: The position of P_1 is given by $(1001, 4)$, that is, the m/z position 1001 and the RT position 4. P_2 has the position of $(1003, 2)$. The RT shift is thus -2 from channel 1 to 2. The XIC shape of P_1 is expressed by a vector (x_1, x_2, x_3, x_4), where x_t denote the intensity of P_1 for the unit quantity $q_1 = 1$. The intensities x_t are called *shape intensities* of P_1 (as they do not specify the quantity but only the shape). Likewise, the shape intensities of P_2 are written as (y_1, y_2, y_3, y_4). For both shape intensities, assume the second elements (x_1 and y_1, respectively) are the apexes. The isotope ratio of P_1 (and of P_2) is given by $1 : r_1 : r_2$; up to the second isotope we consider.

Given above elements, all XICs in the XIC cluster appear between 1001 (the monoisotopic m/z of P_1) and 1005 (the second isotope m/z of P_2) in m/z. For P_1, the XIC appears between 3 and 6 in RT as its RT position is 4. Likewise, the XIC from P_2 appears between 1 and 4 in RT. Thus, all XICs in the XIC cluster lie between 1 and 6

in RT. Therefore, the XIC cluster of P can be represented by a 5×6 matrix whose element in the ith row and the jth column denotes the intensity at m/z of $(i + 1000)$ and at RT of j. The m/z and RT ranges on which the matrix is defined are together called the *domain* (of the matrix). Note all the XICs components also can be represented by matrices defined on the same domain.

The component for P_1 for unit quantity $q_1 = 1$ can be represented on a single matrix $\mathbf{T}_1$ given by

$$\mathbf{T}_1 := \begin{pmatrix} 0 & 0 & x_1 & x_2 & x_3 & x_4 \\ 0 & 0 & r_1 \cdot x_1 & r_1 \cdot x_2 & r_1 \cdot x_3 & r_1 \cdot x_4 \\ 0 & 0 & r_2 \cdot x_1 & r_2 \cdot x_2 & r_2 \cdot x_3 & r_2 \cdot x_4 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{1}$$

The summation of all elements in $\mathbf{T}_1$ is equal to 1 as $q_1 = 1$. The matrix $\mathbf{T}_1$ is called *template matrix* of P_1. The template matrix represents, thus, the whole shape of the component without the quantity information. The component for P_1 of quantity q_1 can be simply written by $q_1 \cdot \mathbf{T}_1$.

Similarly, the template matrix of P_2 is furnished by

$$\mathbf{T}_2 := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ y_1 & y_2 & y_3 & y_4 & 0 & 0 \\ r_1 \cdot y_1 & r_1 \cdot y_2 & r_1 \cdot y_3 & r_1 \cdot y_4 & 0 & 0 \\ r_2 \cdot y_1 & r_2 \cdot y_2 & r_2 \cdot y_3 & r_2 \cdot y_4 & 0 & 0 \end{pmatrix} \tag{2}$$

The summation of all elements in $\mathbf{T}_2$ is equal to 1. $\mathbf{T}_2$ is the template matrix of P_2, and $q_2 \cdot \mathbf{T}_2$ is the component for P_2.

Finally, the XIC cluster generative model for the peptide P is given by

$$\mathbf{M} := q_1 \cdot \mathbf{T}_1 + q_2 \cdot \mathbf{T}_2. \tag{3}$$

The matrix $\mathbf{M}$ representing the XIC cluster is called the *XIC cluster matrix* of P.

SAN2.2 Reconstruction of observed XIC cluster

Now we describe the reconstruction algorithm. The inputs to the reconstruction algorithm are the spectrum data (the set of all observed intensities) and a PSM from which we want to reconstruct its XIC cluster matrix $\mathbf{M}$. The outputs are the reconstructed XIC cluster matrix along with the quantities. Suppose the input PSM is (P_1, S), where P_1 is the labeled peptide of the above example peptide P. The channel of the identified PSM is therefore

1. We want to estimate the quantities q_1 and q_2 based on the above equation (3).

However, the template matrices $\mathbf{T}_1$ and $\mathbf{T}_2$ should be first estimated to use the equation (3), which can be done by estimating the three elements per channel (the position, the XIC shape, and the isotope distribution) and the domain (for matrix representation). The algorithm begins by inferring the position.

While the m/z positions are readily calculated with the m/z of the PSM and the mass spacing, the RT position inference is not a trivial problem. The RT positions are differently acquired for the identified channel (in our case, $i = 1$) and the other channels. We first start with the identified channel $i = 1$. To determine the RT position for $i = 1$, we find the apex of the XIC from P_1. To find the apex, the intensities having m/z of the most abundant isotope of P_1 are collected. Then, from the RT where P_1 is identified (i.e., the RT of the spectrum S), we search for the apex in the direction of increasing intensity. The RT of the first found apex is regarded as the RT position of the XIC of P_1.

In case of P_2, however, the identification is not given and finding the RT position of P_2 cannot be done as above. Instead, the RT shift $\Delta RT_{1,2}$ between P_1 and P_2 is predicted, and the predicted apex RT for P_2 is obtained by the RT position of P_1 plus $\Delta RT_{1,2}$. If a local XIC apex exist within a small RT window from this predicted apex RT, we take this local XIC as the position of P_2. Otherwise, the predicted apex RT is the RT position of P_2. The RT shift prediction method is described in the section SAN4.1.

After inferring the positions, we estimate the shape intensities (e.g., x_t in $\mathbf{T}_1$) by using log-normal probability density function (pdf) fitting algorithm described in the section SAN4.2. This algorithm is based on our observation that the bell-shapes of XICs fit well with the shapes of log-normal probability density functions (pdfs) with different parameters. The inputs to the fitting algorithm include the RT position that we estimated above, which is to accurately extract the observed XIC to be fitted. The fitting algorithm outputs the shape intensities along with the *RT range* on which the intensities exist. If the algorithm fails to generate any shape intensities for the identified channel ($i = 1$ in this example), we declare the reconstruction fails. If the algorithm fails for other channels ($i = 2$ in this example), the shape intensities of P_1 are instead taken, but the RT position for P_2 is shifted by $\Delta RT_{1,2}$.

The last factor necessary to estimate template matrices is the isotope distribution of the identified peptide. We try to calculate the isotope distribution per peptide molecule rather than to use the pre-calculated isotope distribution of averaged elemental composition per mass (so called *averagine* [2]). However, the complexity to calculate the exact isotope distribution often becomes high making the whole algorithm inefficient. Thus, we developed an efficient approximation algorithm to quickly calculate the isotope distribution given a composition. The algorithm is explained in the section SAN4.3. By using this algorithm, the isotope ratios r_1 and r_2 for P_i can

be provided efficiently.

After estimating the template matrices, we define the domain. The m/z range for P_i can be readily given from the m/z values of P_i (calculated from the identification) and the number of isotopes to consider (an input parameter). The RT range for P_i is already obtained. Let the RT range for P_i be (s_i, e_i) for $i = 1, 2$. The RT range for the domain is given by $(min(s_1, s_2), max(e_1, e_2))$. The m/z range for the domain is defined similarly.

As we know the the position, the shape intensities (x_t and y_t), and the isotope ratios (r_1 and r_2), the estimated template matrices $\hat{\mathbf{T}}_1$ and $\hat{\mathbf{T}}_2$ can be deduced as in (1) and (2), respectively. After obtaining the matrices, the estimated XIC cluster matrix $\hat{\mathbf{M}}$ is defined by taking the observed intensities within the domain. We plug in the estimated matrices $\hat{\mathbf{T}}_1$, $\hat{\mathbf{T}}_2$, and $\hat{\mathbf{M}}$ in the equation (3) to calculate the quantities q_1 and q_2. If the estimation is perfect and the XIC cluster is noise-free, the equation should be solvable by a simple matrix inversion. However, this is rarely the case and $\hat{\mathbf{M}}$ and the equality in the equation (3)) does not hold for any q_1 and q_2 combination. Thus, we take the q_1 and q_2 such that the discrepancy between $\hat{\mathbf{M}}$ and $q_1 \cdot \hat{\mathbf{T}}_1 + q_2 \cdot \hat{\mathbf{T}}_2$ is minimized. More precisely, q_1 and q_2 minimizing $\|\hat{\mathbf{M}} - (q_1 \cdot \hat{\mathbf{T}}_1 + q_2 \cdot \hat{\mathbf{T}}_2)\|_2^2$ are calculated. This is a well defined least-square fitting problem and is readily solvable by using Moore-Penrose pseudo-inverse.

To solve this least-square fitting problem, first each of the matrices $\hat{\mathbf{M}}$, $\hat{\mathbf{T}}_1$, and $\hat{\mathbf{T}}_2$ is converted into a raw vector by concatenating the columns. Denote the converted raw vectors as $\boldsymbol{V\hat{M}}$, $\boldsymbol{V\hat{T}}_1$, and $\boldsymbol{V\hat{T}}_2$, respectively. The solutions for q_1 and q_2 are found by

$$\begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{V\hat{T}}_1 & \boldsymbol{V\hat{T}}_2 \end{pmatrix}^+ \boldsymbol{V\hat{M}}, \tag{4}$$

where $(\ \cdot\ )^+$ denotes the Moore-Penrose pseudo-inverse of a matirx. By solving this problem, we effectively reconstruct the observation $\hat{\mathbf{M}}$ by integrating the estimated template matrices $\hat{\mathbf{T}}_1$ and $\hat{\mathbf{T}}_2$, which is equivalent to performing the deconvolution of the components $q_1 \cdot \mathbf{T}_1$ and $q_2 \cdot \mathbf{T}_2$ from the observation $\hat{\mathbf{M}}$. The negative quantities are converted to zero quantities.

SAN2.3 Practical XIC cluster generative model and reconstruction algorithm

So far we described the model-based reconstruction algorithm with several simplifying assumptions. Here we describe the practical XIC cluster generative model for peptide ions P^z. Continuous m/z and RT values are used for this model. To define the XIC cluster matrix, however, m/z and RT values should be made discrete. In case of m/z values, we are only interested in those corresponding to the masses of the labeled peptides and the

natural isotopes thereof. These ion masses are divided by the input charge z yielding the discrete m/z values to consider; the intensities of these m/z values within tolerance (a user specified parameter) are used. In case of RT values, intensities have arbitrary real valued RTs. Thus, instead of using the raw intensities, the log-linear interpolated (along the RT-axis) intensities are collected every ΔRT such that the resulting XIC cluster matrix has 20-30 columns. The number of channels is not limited to 2 and is denoted by n.

In addition to components that define the XIC from a target peptide ion P_z, we also consider those from coelution noises in our model. The noise XICs from coelution are the XICs from the other ions (than P_i^z for $i = 1, \cdots, n$) present within the domain of interest. They also have the log-normal pdf shapes, but their apex locations are usually different from those of P_i^z. We represent each coelution noise of distinct m/z and distinct apex as a separate matrix (but with the same domain). The number of the coeluted noises in a XIC cluster is not fixed and denoted as m. Each coelution noise is denoted as $w_j \cdot \mathbf{C}_j$, where the matrix $\mathbf{C}_j$ specifies the shape&range of noise (like a template matrix), and w_j specifies the abundance of the noise. The generative model including the coelution noises can be written as

$$\mathbf{M} := \sum_{i=1}^{n} q_i \cdot \mathbf{T}_i + \sum_{j=1}^{m} w_j \cdot \mathbf{C}_j. \tag{5}$$

In the reconstruction step, the noise matrices $\mathbf{C}_j$ should be estimated in addition to the template matrices. For the estimation, we find all apexes within the domain of the XIC cluster. Excluding the apexes of XICs from P_i^z (as they will constitute the templates), each apex is subject to the log-normal fitting algorithm. The output from the fitting algorithm gives the estimated shapes&ranges of the coelution noises $\hat{\mathbf{C}}_j$. For each j, $\hat{\mathbf{C}}_j$ is normalized so that its maximum element has the value of 1.

In the practical algorithm, we also preprocess raw spectrum data. First, for each MS1 spectrum, all low intensity peaks are subject to the following base level correction: given a peak in a spectrum, the base level is given by 3% quantile of the intensities within the 100 m/z window around the peak. If the peak intensity does not exceed the two times the base level, the base level is subtracted from the peak intensity. Second, each XIC is subject to Savitzky-Golay smoothing.

The reconstruction algorithm is not unlike the one for the simplified model. Even if the number of channels increased from 2 to n, the quantities q_i for $i = 1, \cdots, n$ as well as the coelution abundances w_j can be furnished by solving the least square fitting problem as above. Only the q_i values are taken as quantities.

Lastly, we correct the biases from the label isotope impurity and the imperfect label incorporation described

in the following sections SAN2.4 and SAN2.5.

SAN2.4 Correction of isotopic impurity interference (-1 Da correction)

Isotopic labeling reagents generally include isotopic impurities, which are usually lighter isotopologues than the desired form, causing an overestimation of the adjacent light channel.

To correct this interference, we experimentally measure the exact isotope distribution of the target labels. The isotopes except for the -1 Da position from the desired mass are ignored because they generally have quite low occurrences.

To measure the isotope distribution of the label corresponding to channel 1, for example, tryptic peptides labeled solely by the channel 1 label are prepared. Then the peptide sample is subject to LC-MS/MS run. After PSM search by MS-GF+, only the PSMs of extremely low spectral E-values (under $1E^{-12}$) are retained. The PSMs with more than one labeling sites are filtered out as we are interested in the isotope distribution of a single label.

For each of the remaining PSMs, -1 Da *m/z shift* of the labeled peptide is defined as the m/z shift from the heavy to the light isotopes (e.g., the mass of ^{2}H - the mass of ^{1}H) existing in the label. If more than one kinds of isotopes are in the label, the average mass shift is taken. -1 Da *peptide m/z* is calculated by subtracting -1 Da m/z shift from the monoisotopic m/z of the labeled peptide. *Isotope abundance ratio at* -1 Da *m/z* is obtained as the intensity ratio between the peak on the monoisotopic peptide m/z and that on -1 Da peptide m/z found in the MS1 spectrum of the monoisotopic XIC's apex retention time position. The isotope distribution for label of channel 1 is defined as the trimmed median of isotope abundance ratios of all the PSMs (threshold 0.00001). For isotope distributions for other channels are estimated in the same way.

The measured isotope distribution is used to correct the isotope impurity interference in PSM quantification step. For each channel, the discrete convolution between the isotope distribution of the peptide and the isotope distribution of the label is performed. If a PSM includes more than one labels, discrete convolution was repeatedly taken per label. The peptide isotope distribution in template matrix (SAN2.1) is replaced by the convoluted isotope distribution. Simply by performing observed XIC reconstruction with this updated template matrix, the correction of isotope impurity interference is achieved.

SAN2.5 Correction of the differences in isotope incorporation rate

In SILAC, the incorporation of heavy amino acids into cellular proteome is not always perfect. Therefore, original non-labeled peptides should exist at a certain level in SILAC samples. Such nature in SILAC causes to overestimate the quantity of channel 1, which corresponds to the non-labeled peptides.

To correct this bias, we measure the isotope incorporation rate in each label of our SILAC-6plex experiment. Similar to above section, LC-MS/MS analysis of individual label channel is separately performed. For instance, to measure the isotope incorporation rate of the channel 2 label of SILAC-6plex, the tryptic peptides labeled by the channel 2 label are prepared and subject to LC-MS/MS analysis. The PSM search is performed by MS-GF+, while the label mass shift is set as a variable modification. Per labeled amino acid, the isotope incorporation rate is calculated by dividing the number of the labeled PSMs by the total number of the PSMs.

For each quantified PSM, the measured isotope incorporation rate is applied for the correction. For channel i ($i = 1, \cdots, n$) and labeled amino acid aa ($aa = a, \cdots, z$), denote the number of the labeled amino acids as N_{aa}, the incorporation rate as $ir_{aa,i}$, the corrected quantity as q'_i, and the observed quantity as q_i. The corrected quantities $q'_1, \cdots, q'_n$ can be calculated by solving the following matrix equation in (6).

$$\begin{pmatrix} 1 & 1-ir_{a,2} & \cdots & 1-ir_{a,n} \\ 0 & ir_{a,2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & ir_{a,n} \end{pmatrix}^{N_a} \cdots \begin{pmatrix} 1 & 1-ir_{z,2} & \cdots & 1-ir_{z,n} \\ 0 & ir_{z,2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & ir_{z,n} \end{pmatrix}^{N_z} \begin{pmatrix} q'_1 \\ q'_2 \\ \vdots \\ q'_n \end{pmatrix} = \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{pmatrix} \tag{6}$$

SAN2.6 Quantified XIC cluster filtration and FDR estimation

The above formulation also allows us to estimate the signal-to-noise ratio (SNR). The discrepancy between the observation and the reconstruction is given by $\mathbf{N} := \hat{\mathbf{M}} - (\sum_{i=1}^{n} q_i \cdot \hat{\mathbf{T}}_i + \sum_{j=1}^{m} w_j \cdot \hat{\mathbf{C}}_j)$. The noise power is $\|\mathbf{N}\|_2^2$. The signal power, written as $\|\mathbf{S}\|_2^2$, is calculated as $\|\mathbf{S}\|_2^2 = \sum_{i=1}^{n} \|q_i \cdot \hat{\mathbf{T}}_i\|_2^2$. The SNR is given by $\frac{\|\mathbf{S}\|_2^2}{\|\mathbf{N}\|_2^2}$. Any XIC cluster of SNR less then a given threshold (default 2, a user specified parameter) is considered as inaccurately quantified and thus is discarded. The SNR per PSM is also used to define SNR of a peptide/protein. Let a set of l PSMs $\{(P_k, S_k) | k = 1, \cdots, l\}$ is matched to a peptide/protein and the signal and noise powers of a PSM (P_k, S_k) are written as $\|\mathbf{S}_k\|_2^2$ and $\|\mathbf{N}_k\|_2^2$. Then, the SNR of the peptide/protein is defined by $\frac{\sum_{k=1}^{l} \|\mathbf{S}_k\|_2^2}{\sum_{k=1}^{l} \|\mathbf{N}_k\|_2^2}$.

After applying the SNR threshold for all quantified XIC clusters, those having the same domains are compared and discarded except for the one(s) with the lowest spectral E-value. This is done efficiently using an interval tree

algorithm [3]. Then, the *XIC cluster score* is defined per XIC cluster. Given an XIC cluster of an identified PSM (P, S), we find all PSMs having the peptide P within the domain of the XIC cluster matrix. The XIC cluster score is the product of the spectral E-values of the found PSMs. This XIC cluster score is used to calculate PSM level FDR and q-values, exploiting the spatial consistency of the identifications. After calculating the q-values using the target-decoy approach [4], Only the XIC cluster (i.e., PSMs) of q-value lower than 1% (a user specified parameter) are retained.

SAN3 Protein quantification exploiting quantity ratio consistency of matching PSMs

For protein level quantification, the quantified PSMs should be assigned to proteins and the quantities of the PSMs should be merged in the protein level. A PSM is assigned to any protein whose sub-sequence matches to the peptide sequence of the PSM. When a PSM is assigned to proteins, two different cases occur: the peptide in the PSM is matched to a single protein (a *uniquely matched PSM*) or it is matched to multiple proteins (a *shared PSM*). The shared PSMs complicate the protein level analysis as it is unclear which proteins they are from. In addition to the shared PSMs, the PSMs incorrectly quantified and/or identified (*incorrect PSMs*) also hinder the accurate protein level quantification.

To avoid the complication from shared and incorrect PSMs, we iteratively prune the matches between PSMs and proteins based on the assumption that the quantity ratios of the protein and its matching PSMs are expected to be mutually consistent.

The pruning of the matches is done on a bipartite graph $G(U, V, E)$, where U is the set of PSM nodes, V the set of protein nodes, and E the set of edges defined below. Each node $u \in U$ represents a PSM, and node $v \in V$ a protein. Any edge $(u, v) \in E$ is connected if and only if the peptide of the PSM node u matches to the protein node v. For each PSM node u, the quantities, represented by a vector $\mathbf{q}^u := (q_1^u, q_2^u, \cdots, q_n^u)$, are given from the PSM level quantification step. Our goal is to assign the quantities to each protein node $v \in V$.

For a protein node v, let the set of the connected PSM nodes be $U(v)$. Note that $U(v) \subset U$. Denote the summation of vectors $\mathbf{q}^u$ for $u \in U(v)$ by $\mathbf{Q}^v$, that is, $\mathbf{Q}^v := \sum_{u \in U(v)} \mathbf{q}^u$. The vector $\mathbf{Q}^v$ approximates the protein level quantities for the protein node v. The aforementioned mutual ratio consistency requires that the vectors $\mathbf{q}^u$ for $u \in U(v)$ and $\mathbf{Q}^v$ should have similar directions, that is, any two pairs of those vectors should have high cosine

similarity. We first prune the edges between $U(v)$ and v using this cosine similarity. For each PSM node $u \in U(v)$, calculate the cosine between two vectors $\mathbf{q}^u$ and $\mathbf{Q}^v - \mathbf{q}^u$. The calculated cosine reflects the ratio consistency between the peptide level quantity ratio for u and the protein level quantity ratio for v, excluding the contribution from the PSM of u. If the cardinality of $U(v)$ is larger than 1 (i.e., $|U(v)| > 1$), any edge (u, v) is pruned if the cosine for the edge is less than a given cosine threshold (default 0.8, a user specified parameter). After the pruning, $U(v)$ is updated. We update $\mathbf{Q}^v$ accordingly. This pruning step is called *per-protein-pruning*. We repeat per-protein-pruning several times (default 10 times) for all protein nodes $v \in V$. This iteration effectively removes the nodes corresponding to incorrect PSMs or the PSMs with outlying quantity ratios.

Next, we examine each PSM node. Let $V(u)$ be the set of protein nodes connected to the PSM node u. Let the PSM of the node u be a shared PSM, that is, $|V(u)| > 1$. As we already performed the per-protein-pruning, $\mathbf{Q}^v$ is defined for all $v \in V(u)$. We calculate the cosine between two vectors $\mathbf{q}^u$ and $\mathbf{Q}^v$ for each $v \in V(u)$. Out of all edges (u, v), we only retain the edge(s) having the maximum cosine value. If more than one edges remain, the connected protein nodes are grouped forming a *protein group*. Note that only when multiple proteins are connected to exactly the same set of PSMs, they constitute a protein group. This pruning step is called *per-peptide-pruning*. Per-peptide-pruning is done just once for all peptide nodes $u \in U$.

After running the per-peptide-pruning, we again run the per-protein-pruning several times (default 10 times). This is because after the per-peptide-pruning, the effect of shared PSMs is reduced, and the protein level quantities $\mathbf{Q}^v$ for each protein node $v \in V$ should be updated accordingly. After these iterations, any protein node is discarded if the number of matched PSMs is less than a user specified parameter (default 2). Finally, $\mathbf{Q}^v$ for each protein node $v \in V$ are reported by EPIQ as the protein level quantification result for the protein for v.

SAN4 Sub-algorithms for the model-based reconstruction algorithm

In this section, we explain three sub-algorithms: prediction of RT shift, log-normal pdf fitting, and isotope prediction algorithms.

SAN4.1 Prediction of RT shift

For the inference of the template matrices, the accurate prediction of RT shift between peptide ions of different channels is necessary. Even though the main factor causing RT shift is ^{2}H (or deuterium) used in labeling, previous study in [5] suggested that the accurate quantitative prediction of RT shift cannot be achieved solely

with the number of the bound deuteriums. Thus, we try to extract additional features related to RT shift based on empirical analysis of our experimental datasets.

To extract relevant features and also to train the regression model, the training dataset containing "true" RT shifts should be prepared. Let a "true" PSM (P_i, S) be given, where i is the channel of the identified peptide. Let n denote the number of all channels. For each of $j = 1, \cdots, n$, (j sorted by number of deuteriums) the XIC of P_j ion has its (usually unknown) "true" apex RT which we want to measure. Once they are known, the "true" RT shifts between different channels can be deduced by calculating the difference between them. The ideal training dataset consists of the "true" PSMs and the "true" apex RTs associated to each of the "true" PSM.

However, the "true" PSMs are not available. Thus, highly reliable PSMs are collected and used instead. To collect such PSMs, we use DE-6plex and SILAC-6plex (the same labeling used for test experiments) labeled samples and having almost even quantity ratios ($1:1:1:1:1:1$ and $1:0.8:1:0.8:1:0.8$). These ratios are used to minimize possible bias caused by other factors than RT shift itself. For the prepared spectrum datasets, we perform the spectrum identification step. Then, only the PSMs having extremely low spectral E-values (10^{-12}, which corresponds to FDR lower than 10^{-3} in typical searches) are retained in the training dataset.

Next, given a PSM (P_i, S) in the training dataset, we try to find its associated apex RTs. In case of the identified channel i, the apex RT, denoted by t_i, can be measured as described in the section SAN2.2. To ensure t_i is close to the "true" apex RT, we attempt to check the quality of the XIC of P_i ion. To this end, the shape intensities for P_i are acquired by applying the log-normal fitting algorithm. Then the cosine between the shape intensities and the observed intensities is calculated. If the cosine exceeds a high threshold T_{init} (default 0.9), the quality of XIC is regarded as high and so is its apex RT t_i. Such t_i is associated to the PSM (P_i, S). Otherwise, the PSM (P_i, S) is discarded from the training dataset.

If the PSM (P_i, S) is not discarded, we continue to find apex RTs t_j for the channels $j = i + 1$ (or $j = i - 1$). As RT shift prediction is not available at this moment, t_j should be found *without* using the prediction. To do so, the intensities having the m/z of P_j are collected. Then only the intensities around the *reference RT window* $(t_i - w, t_i + w)$ are retained, for a given window width w (about 30 seconds when the whole LC running time is about 2 hours). Then the RT of the apex of the collected intensities defines t_j. This selection is to minimize the chance to select co-eluted peaks' RT. To evaluate the quality of collected intensities, we apply the log-normal pdf fitting algorithm for the collected intensities to obtain the shape intensities for P_j. If the cosine between the shape intensities and the collected intensities exceeds a cosine threshold T_{others} (default 0.9), the apex RT t_j is associated to the PSM (P_i, S). Otherwise, t_j is considered to be undetectable and is discarded. Notice that not

Experiment type	LC-MS/MS runs	PSMs
DE-6plex, 125 min LC	2 replicate 1:1:1:1:1:1 runs, 2 replicate 1:0.8:1:0.8:1:0.8 runs	2739
DE-6plex, 130 min LC	10 replicate 1:1:1:1:1:1: runs	2882
SILAC-6plex	5 replicate 1:1:1:1:1:1: runs	1786

Table SAN1: The number of training PSMs used for different experiments

the PSM but the channel is discarded in this case.

The above procedure is repeated from $j = i + 1$ to $j = n$, and from $j = i - 1$ to $j = 1$, sequentially. When inferring t_j, we find the channel k closest to j such that t_k is retained, take $(t_k - w, t_k + w)$ as the reference RT window, and infer t_j. After gathering all valid apex RTs, we calculate apex RT differences between channels to collect RT shift values. Only if more than one RT shift (i.e., more than two apex RTs) are associated to the PSM, the PSM (P_i, S) is retained in the training dataset.

The above steps describe how to find the RT position and shape intensities *without* using RT shift prediction. Thus, one may ask why RT shift prediction is even necessary. However, this method only works when the XICs have very low noise level and the labeled peptides have even quantity ratio. The RT shift prediction enables the definition of RT positions and shape intensities even when the quality of XIC is poor and the quantity ratio is uneven.

After defining apex RTs of PSMs, we further filter out PSMs to retain the most reliable PSM for each peptide. First, the PSMs with the same peptide sequence (regardless of the identified channels) are grouped. Only a single PSM having the largest number of associated apex RTs is retained per group. If more than one PSMs contain the largest number of associated apex RTs, we choose one with the highest sum of cosines of all channels.

In this study, we generated training dataset containing 7047 PSMs, by processing 19 LC-MS/MS runs as described above. Table SAN1 shows the number of training PSMs used for different experiments.

So far we used the absolute value of RTs and RT shifts, but the total LC running time is different for each LC-MS/MS run. Since the RT values of the same ions are known to be proportional to the total LC running time, the predicted RT shifts also should be scaled according to the LC running time. To avoid such inconvenience, we use *normalized RT* in practice. To define the normalized RT, we rescale RT by the 20th percentile RT t_s and the 80th t_e. The normalized RT t' of absolute RT t is given by

$$t' = \frac{t - t_s}{t_e - t_s} \tag{7}$$

To select features that influence RT shift, we tested several features extracted from the PSMs in the training dataset. For each feature, we calculated the distance correlation([6]) between the feature and the measured RT shift. The distance correlation was adopted as it captures non-linear relations as well as linear relations between variables; some of the informative features (e.g., RT of the PSM) are not linearly correlated to RT shifts (Fig. SAN1). After screening, the following four features were selected.

- Number of deuteriums
- RT where the PSM is identified
- Peptide sequence length
- Proportion of labeled amino acids in a peptide (e.g. proportion of Lysine and proportion of Arginine in SILAC-6plex)

In practice, feature values were standardized so that the sample mean equals zero and sample variance equals one for each feature. Note that the number of deuteriums is defined per channel of each PSM and others are per PSM.

As illustrated in Fig. SAN1, the number of deuteriums (Fig. SAN1 **a**) shows the highest distance correlation (0.66) as expected. However, this correlation is rather low for the accurate prediction of RT shift. The other features have even lower distance correlations than the number of deuteriums. For instance, the proportion of lysine has only the correlation of 0.32. The authors in [5] also defined a similar feature pool and reported low (pearson) correlations between the features in their feature pool and the RT shift. They concluded that peptide or XIC features are scarcely correlated with the RT shift and did not attempted to predict the RT shift.

To test the feasibility of the RT shift prediction, we attempted to find a lower bound of the prediction error before implementing a prediction method. To do so, distinct training datasets from technical replicates were prepared. For each dataset pair, we collected RT shifts from overlapping peptides (the peptides contained in both datasets) and measured their inconsistency. This gives a lower bound of the prediction error. Fig. SAN2 shows the correlation between measured RT shifts in DE-6plex 125m training dataset pairs. Each dot represent an RT shift of a single peptide's single channel. If no error exists between replicates, the dots should be aligned along $y = x$ line. The Root-Mean-Square Error(RMSE) calculated from the pairs was $0.00077 - 0.0012$, which was less than one tenth of the mean XIC range of 0.016. For our purpose, this error bound is sufficiently small.

Next, we benchmarked RT shift prediction performance of various regression methods (Fig. SAN3 **a**), implemented by scikit-learn (http://scikit-learn.org/) and LibSVM ([7]). For each method, the RMSE and correlation

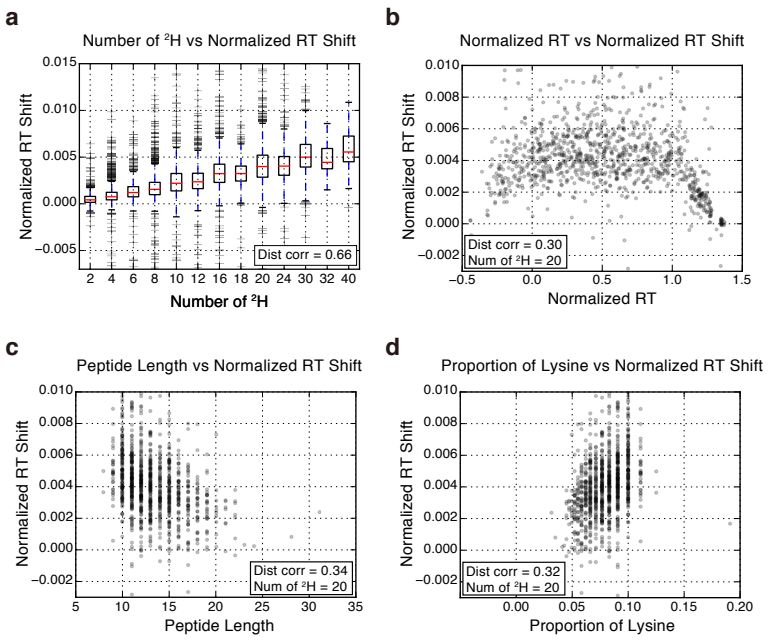


Figure SAN1: Scatter plots for the selected features for training. All features and RT shifts were retrieved from DE-6plex 125m LC-MS/MS runs. Each dot represents RT shift. The distance correlation is specified for each feature (the higher the more correlated). Except for the number of deuteriums in **(a)**, RT shifts from the peptides with 20 deuteriums are used for visualization and distance correlation calculation. **(a)** The number of deuteriums vs. normalized RT shift**(b)**The normalized RT vs. normalized RT shift**(c)**The peptide length vs. normalized RT shift**(d)**The proportion of labeled amino acids vs. normalized RT shift

coefficient were calculated using 10-fold cross-validation. Nu-Support Vector Regression (Nu-SVR) showed the largest correlation coefficient and the smallest RMSE among tested regression methods. We further tested parameters of Nu-SVR to optimize RT shift prediction model performance. The nu-SVR with optimized parameters resulted in a slightly higher RMSE (0.0016) than the lower bound, which is still sufficiently small for our purpose (Fig. SAN3 **b**).

Such small RMSE from nu-SVR method implies that even if each feature is not strongly correlated to RT shift, their combination conveys enough information for the precise prediction. To check if this is the case, we performed RT shift prediction using nu-SVR in which only a single feature is used (single-featured prediction) per feature (Table SAN2). The training was also done separately for each feature. To evaluate the single-featured predictions, we measured RMSE and the correlation coefficient between for each case.

As expected, RT shift prediction by the number of deuteriums showed the best result out of the four single-featured predictions. However, the RMSE values from single-featured training ($1.79 \cdot 10^3 - 2.19 \cdot 10^3$) were larger than the RMSE of all-featured prediction ($1.61 \cdot 10^3$). Also, correlation coefficients of single-featured predictions ($0.17 - 0.59$) were lower than correlation coefficient of all-featured prediction(0.69)

To check whether all features we selected are informative for RT shift prediction, we also benchmarked RT shift prediction with all-except-one features (Table SAN2). Removing any feature from RT shift prediction model reduced the correlation coefficient and increased the RMSE. This indicates all features we selected are conducive for RT shift prediction. However, the effect of removing PSM RT, peptide length and proportion of labeled amino

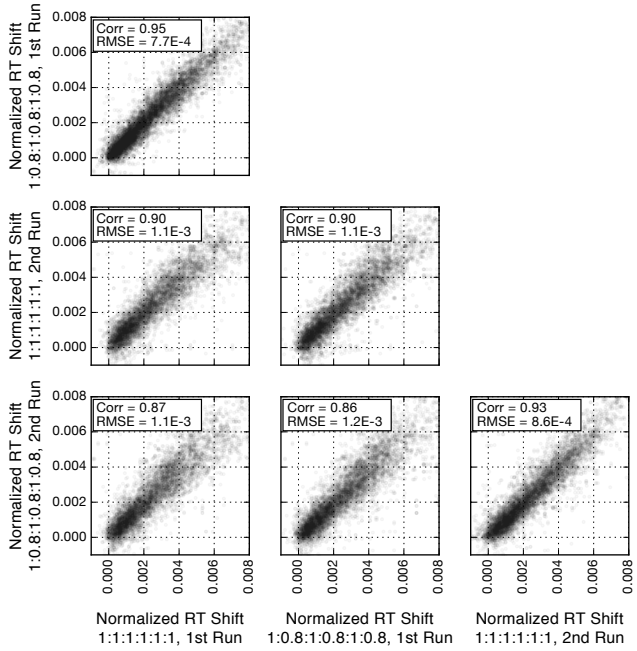


Figure SAN2: Scatter plots for RT shifts measured from DE-6plex 125m LC-MS/MS runs. Each dot represents RT shift of a single peptide form (i.e., a peptide of a single channel), measured from different LC-MS/MS runs. Pearson's correlation coefficient and RMSE (Root-Mean-Square Error) between two runs are specified in each plot.

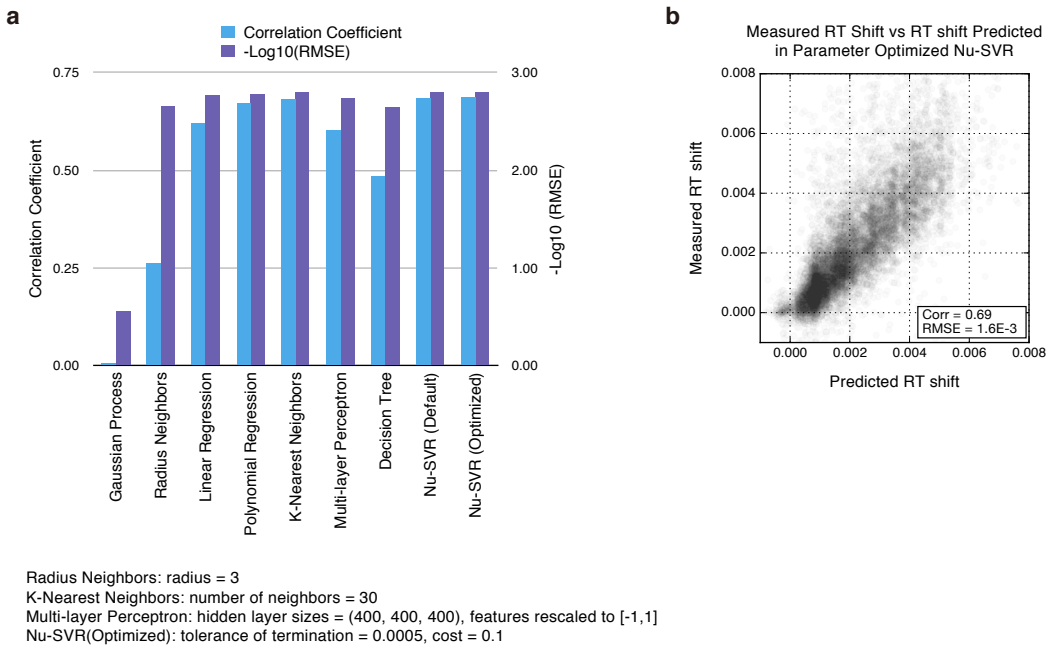
acids were smaller than the effect of removing the number of deuteriums.

Prediction type	Features used	RMSE	Correlation coefficient
Single-featured	# deuteriums	$1.79 \cdot 10^{-3}$	0.59
	PSM RT	$2.14 \cdot 10^{-3}$	0.28
	Peptide length	$2.19 \cdot 10^{-3}$	0.17
	Proportion of labeled amino acids	$2.09 \cdot 10^{-3}$	0.34
All-except-one-featured	All except # deuteriums	$2.00 \cdot 10^{-3}$	0.43
	All except PSM RT	$1.73 \cdot 10^{-3}$	0.63
	All except peptide length	$1.64 \cdot 10^{-3}$	0.67
	All except proportion of labeled amino acids	$1.63 \cdot 10^{-3}$	0.68
All-featured	All four features above	$1.61 \cdot 10^{-3}$	0.69

Table SAN2: Evaluation of predictions with different feature sets

SAN4.2 Log-normal pdf fitting algorithm

The inputs to the log-normal pdf fitting algorithm are the observed intensities (of the same m/z) and a *pivot RT* to which the apex of the log-normal pdfs is to be matched. The pivot RT is usually the apex RT of the observed intensities. The outputs from the algorithm are i) the log-normal pdf shaped intensities that are fit to the input intensities and ii) the RT range where the fitted intensities are located. Denote the input intensities by a vector $(z_0, z_1, \cdots, z_p, \cdots)$, where p is the pivot RT. The log-normal pdf with the location and scale parameters (μ, σ^2) is written as $ln\mathcal{N}(\mu, \sigma^2)$, and the pdf value at x is as $ln\mathcal{N}(x; \mu, \sigma^2)$. We find the largest RT s such that $s < p$ and $z_s \leq z_p/100$. We also search for the smallest RT e such that $e > p$ and $z_e \leq z_p/10$. If we truncate the input



Radius Neighbors: radius = 3
K-Nearest Neighbors: number of neighbors = 30
Multi-layer Perceptron: hidden layer sizes = (400, 400, 400), features rescaled to [-1,1]
Nu-SVR(Optimized): tolerance of termination = 0.0005, cost = 0.1

Figure SAN3: Benchmark of RT shift prediction on DE-6plex 125m LC-MS/MS runs. **(a)** Benchmark of RT shift prediction results from several regression algorithms. Default parameters of scikit-learn were used except for the parameters specified on the bottom side. Also, the features were standardized to zero-mean and unit-variance, except for Multi-Layer Perceptron. Parameter optimized for Nu-SVR showed the best performance concerning both RMSE and correlation coefficient. **(b)** A scatter plot showing the correlation between measured RT shifts and predicted RT shifts. RT shift prediction was done by Nu-SVR with optimized parameters (parameters shown in **(a)**)

intensities from s to e, the truncated intensities form a bell shaped curve starting at s with its apex at p. Define a function $f_\mu(t) := ln\mathcal{N}(\frac{t-s}{p-s}; \mu, \sqrt{\mu})$ for $s \leq t \leq e$. Then, regardless of μ, the functions $f_\mu(t)$ form log-normal pdf curves starting at s with their apexes at p. The value of μ determines the shape of the log-normal pdf curve $f_\mu(t)$. We calculate the cosine between two vectors $(z_s, z_{s+1}, \cdots, z_{e-1}, z_e)$ and $(f_\mu(z_s), f_\mu(z_{s+1}), \cdots, f_\mu(z_{e-1}), f_\mu(z_e))$ for various μ values and take μ that maximizes the cosine value (written by $\hat{\mu}$). If the maximum cosine does not exceed a given threshold (default 0), the algorithm outputs the failure flag. Otherwise, the algorithm outputs the vector $(f_{\hat{\mu}}(z_s), \cdots, f_{\hat{\mu}}(z_e))$ along with the RT range (s, e).

SAN4.3 Isotope distribution calculation algorithm

The inputs to the isotope distribution calculation algorithm are the elemental composition (of a peptide) and the max isotope index N (default 4). The outputs are the isotope ratios from the monoisotope (normalized to 1) to the Nth isotope.

For quick calculation of the distribution, the algorithm uses several tricks. The algorithm only considers the

nominal masses. The high complexity of the isotope distribution calculation is mainly due to the diversity of atoms in the input composition. Thus, the algorithm removes the diversity of the input atoms by generating an *imaginary atom* and substituting all input atoms by the generated imaginary atoms. The isotope profile of the imaginary atom is furnished by taking the nominal-mass-wise average of the profiles of the input atoms. For example, suppose the input composition is $C_{10}O_{30}$. The isotope profile of C can be represented by a vector $(0.9893, 0.0107)$, where its nth element reflects the frequency of the n-th isotope of C. Similarly, the profile of O is given by $(0.99757, 0.00038, 0.00205)$. Then, the imaginary atom denoted by α has the profile given by $(\frac{0.9893\times10+0.99757\times30}{10+30}, \frac{0.0107\times10+0.00038\times30}{10+30}, \frac{0.00205\times30}{10+30}) \approx (0.996, 0.003, 0.001)$. Instead of $C_{10}O_{30}$, the algorithm calculate the isotope distribution of 40 α's, i.e., α_{40}.

The calculation is done efficiently using a dynamic programming algorithm on nominal masses. The algorithm is only described for the above example of α_{40}, but could be readily modified for other cases. We construct a directed acyclic graph (DAG) defined on non-negative integer nodes $0, 1, 2, \cdots, N$. The node n represents the nth isotope of α_{40}. To define edges of the DAG, we examine the isotope profile of α, which is $(0.996, 0.003, 0.001)$. As it has up to the second isotope frequency, two kinds of edges e_1 and e_2 representing the first and the second isotopes are used. Any pair of nodes $n-1$ and n are connected by e_1, and $n-2$ and n are by e_2. Consider a path from node 0 to node n consisting of l_1 e_1 edges and l_2 e_2 edges. The *intensity* of the path is defined as $p(40-l_1-l_2, l_1, l_2; 40; 0.996, 0.003, 0.001)$ where $p(x_1, x_2, x_3; m; p_1, p_2, p_3)$ is the probability mass function of the multinomial distribution (with m number of trials). By using a path-finding algorithm on a DAG [8], one can quickly collect all paths from the node 0 to any node n. The *intensity* of the nth isotope of α_{40} is given by the summation of the intensities of all collected paths. The intensities of all the isotopes (from the 0th to the Nth) of α_{40} are calculated and normalized so the 0th intensity is equal to 1.

Supplementary Algorithm Note References

[1] Sangtae Kim and Pavel A. Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, 5:5277, October 2014.

[2] Michael W. Senko, Steven C. Beu, and Fred W. McLafferty. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 6(4):229–233, April 1995.

[3] H. Edelsbrunner. *Dynamic Data Structures for Orthogonal Intersection Queries*. Forschungsberichte / Inst. f. Informationsverarbeitung. Inst. f. Informationsverarbeitung, TU Graz, 1980.

[4] Kyowon Jeong, Sangtae Kim, and Nuno Bandeira. False discovery rates in spectral identification. *BMC bioinformatics*, 13 Suppl 16:S2, 2012.

[5] Joseph M. Boutilier, Hunter Warden, Alan A. Doucette, and Peter D. Wentzell. Chromatographic behaviour of peptides following dimethylation with H2/D2-formaldehyde: implications for comparative proteomics. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences*, 908:59–66, November 2012.

[6] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, December 2007.

[7] Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

[8] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, 3rd Edition*. The MIT Press, Cambridge, Mass, 3rd edition edition, July 2009.