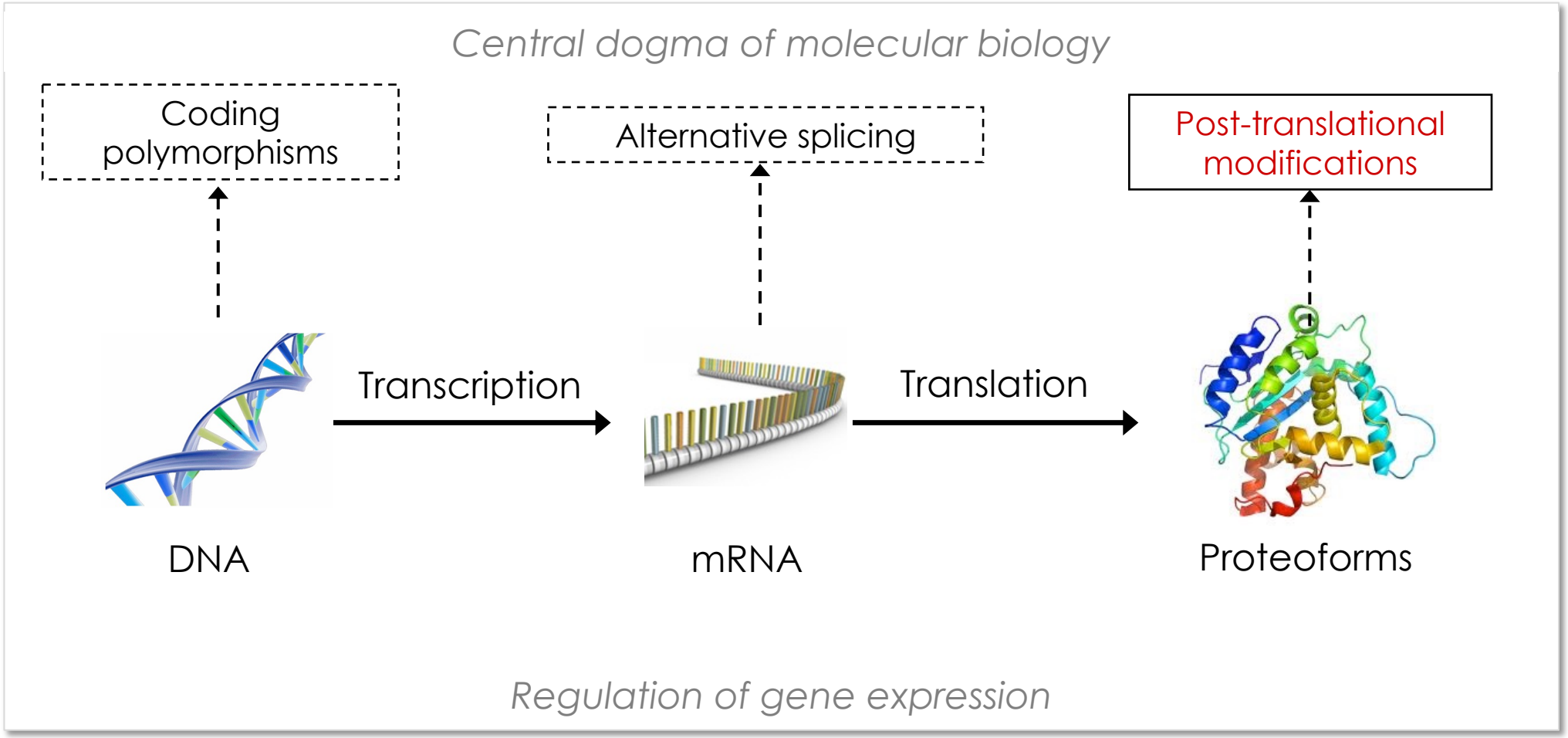
Intact protein analysis
The expanding role of proteoforms in health and disease

Julia Chamot-Rooke

Mass Spectrometry for Biology Unit

Institut Pasteur, Paris (France)

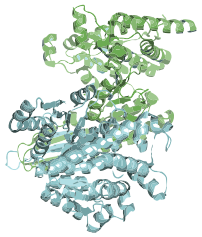
Rationale behind looking at intact proteins



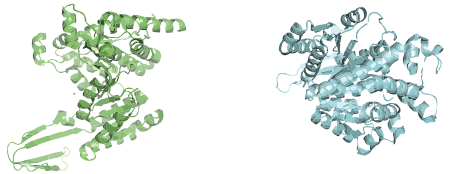
Many protein forms (**proteoforms**) from a single gene

What is a proteoform?

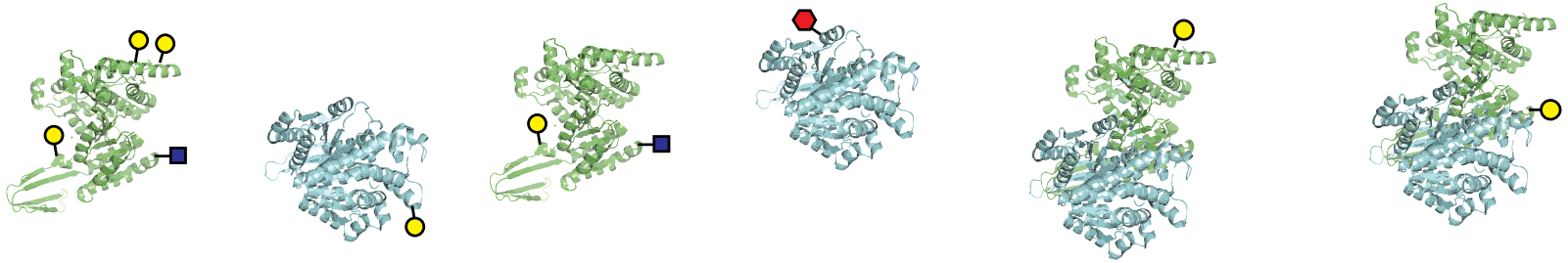
Canonical Sequence (UniProtKB)



Endogenous proteolysis
mRNA splicing
Mutations
SNPs



Site specific features (PTMs):
Govern activity, localization, interactions, half-life



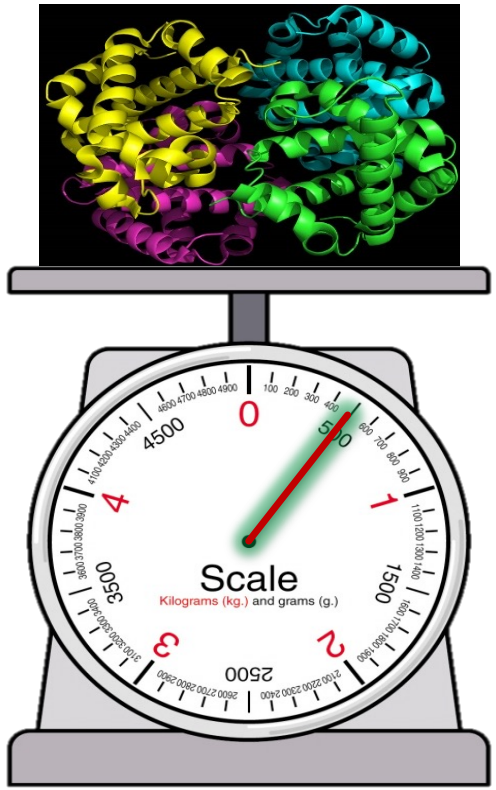
A distinct molecular form of a protein product arising from a single gene

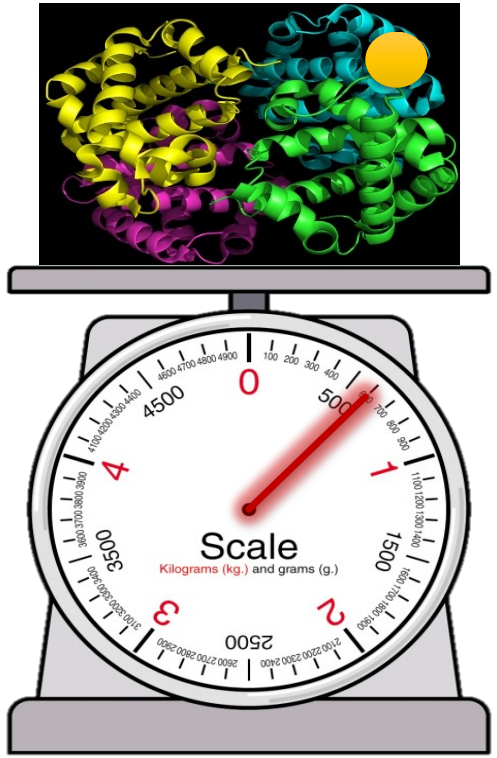
The Consortium for Top-Down Proteomics

CTDP
Consortium for Top-Down Proteomics
Bringing Proteoforms to Life™

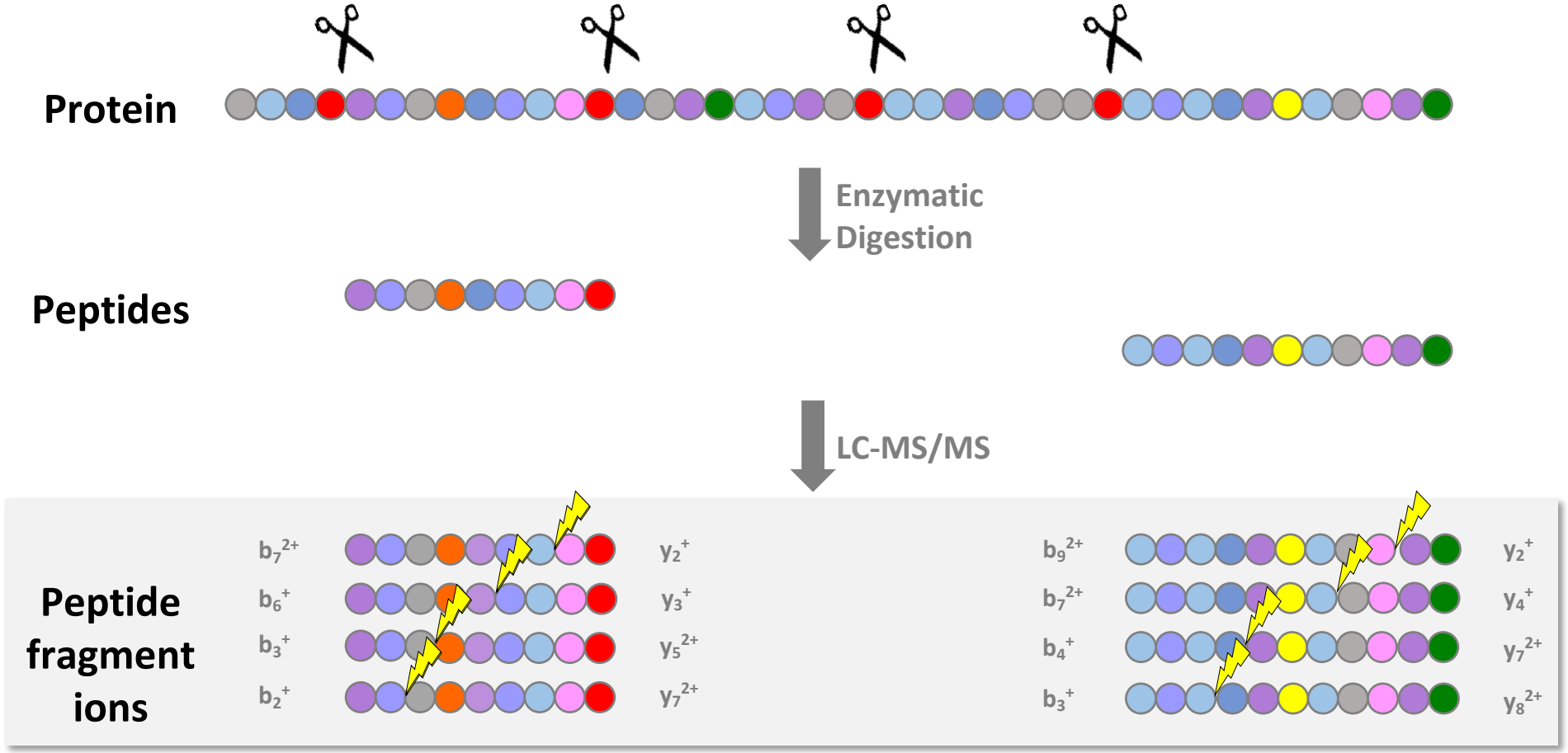
L. Smith *et al. Nat. Methods 10*, 186–187 (2013)

Measuring Intact Proteins (Simple Idea!)



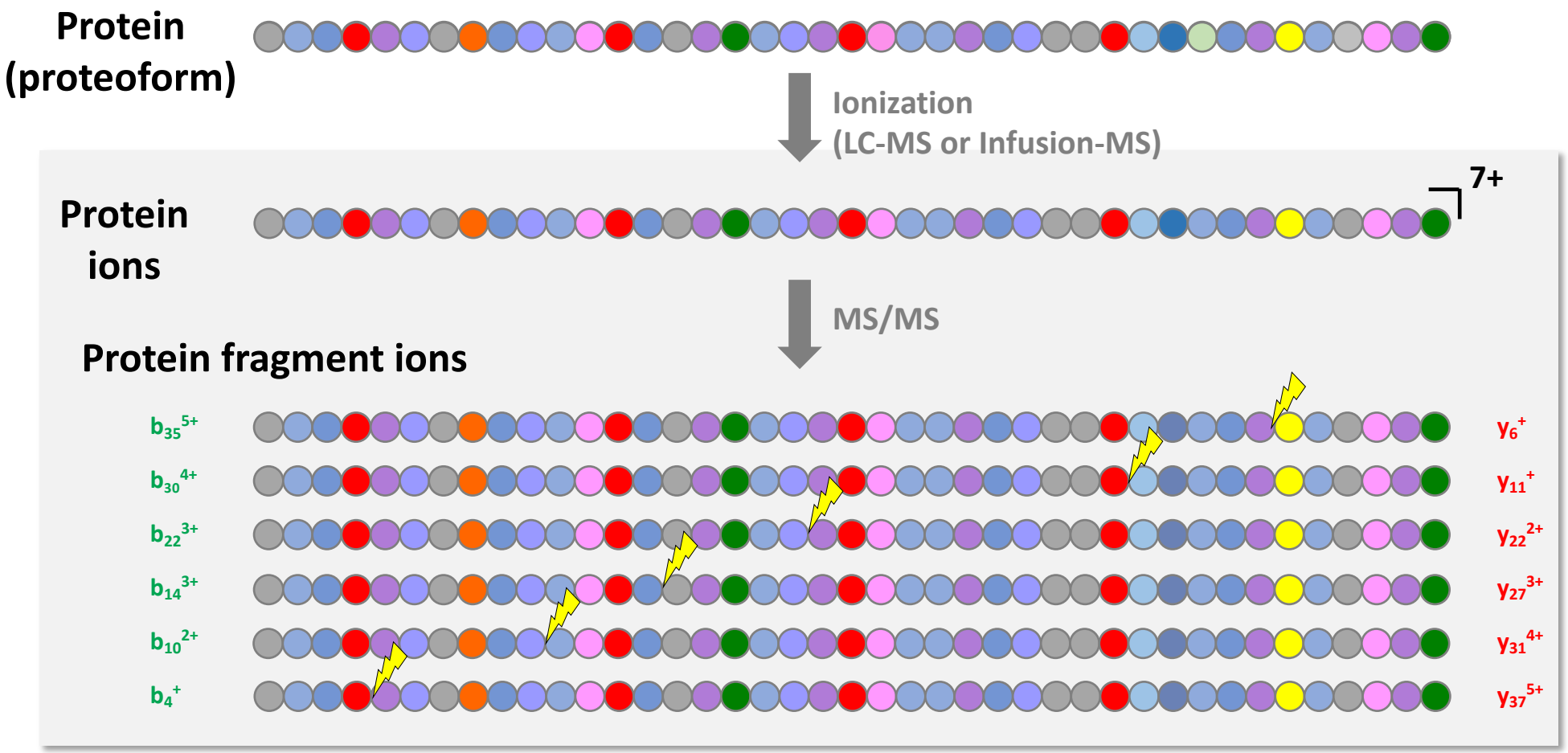


Bottom-Up Proteomics (BUP)



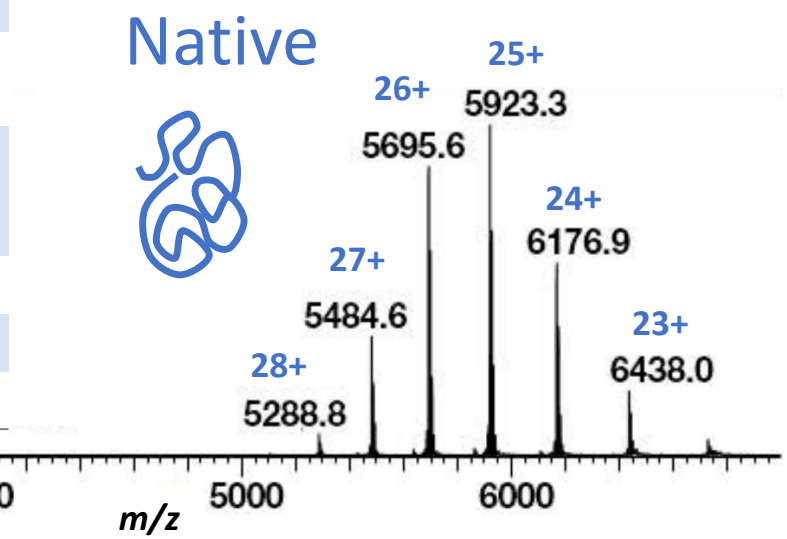
Protein identification based on peptide identification (partial sequence coverage)

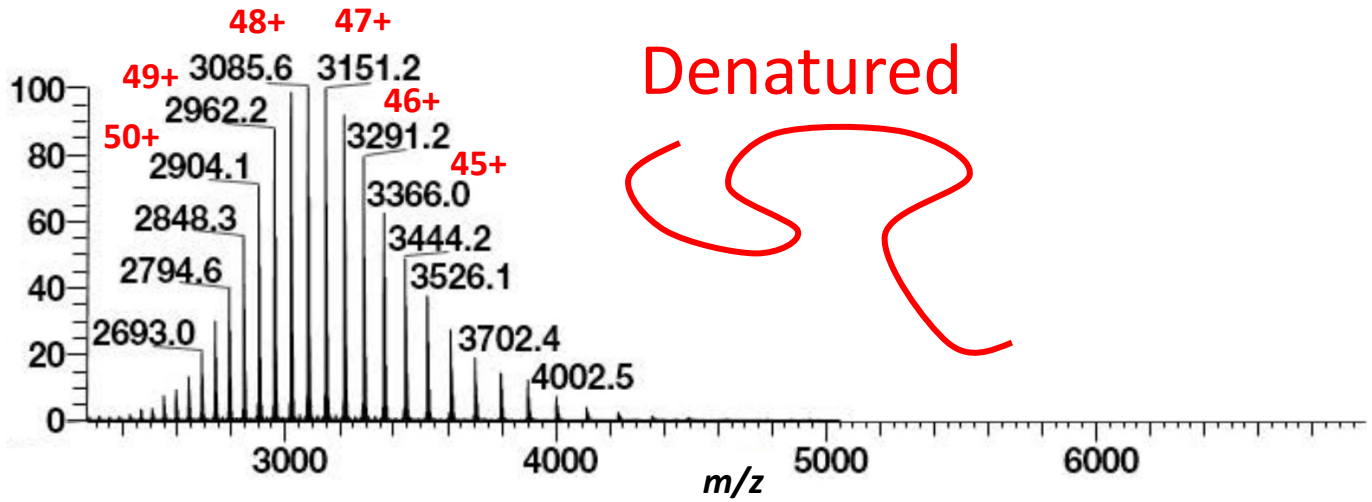
Top-Down Proteomics (TDP)



Denaturing vs native top-down MS

Denaturing Top-down	Native Top-down
Denaturing separation	Native separation (e.g. SEC, IEX)
Partial organic solution (H_20, ACN, Formic Acid, pH 1-2)	Aqueous solution (H_20, Ammonium Acetate, pH 6-8)
MW of single subunit	MW of protein complex and subunits, non-covalent interactions, stoichiometry and structure
Automated analysis (LC-MS)	Manual injection
Complex mixture	Purified samples
Max. MW ~50 kDa (LC-MS/MS)	Max. MW >1,000 kDa

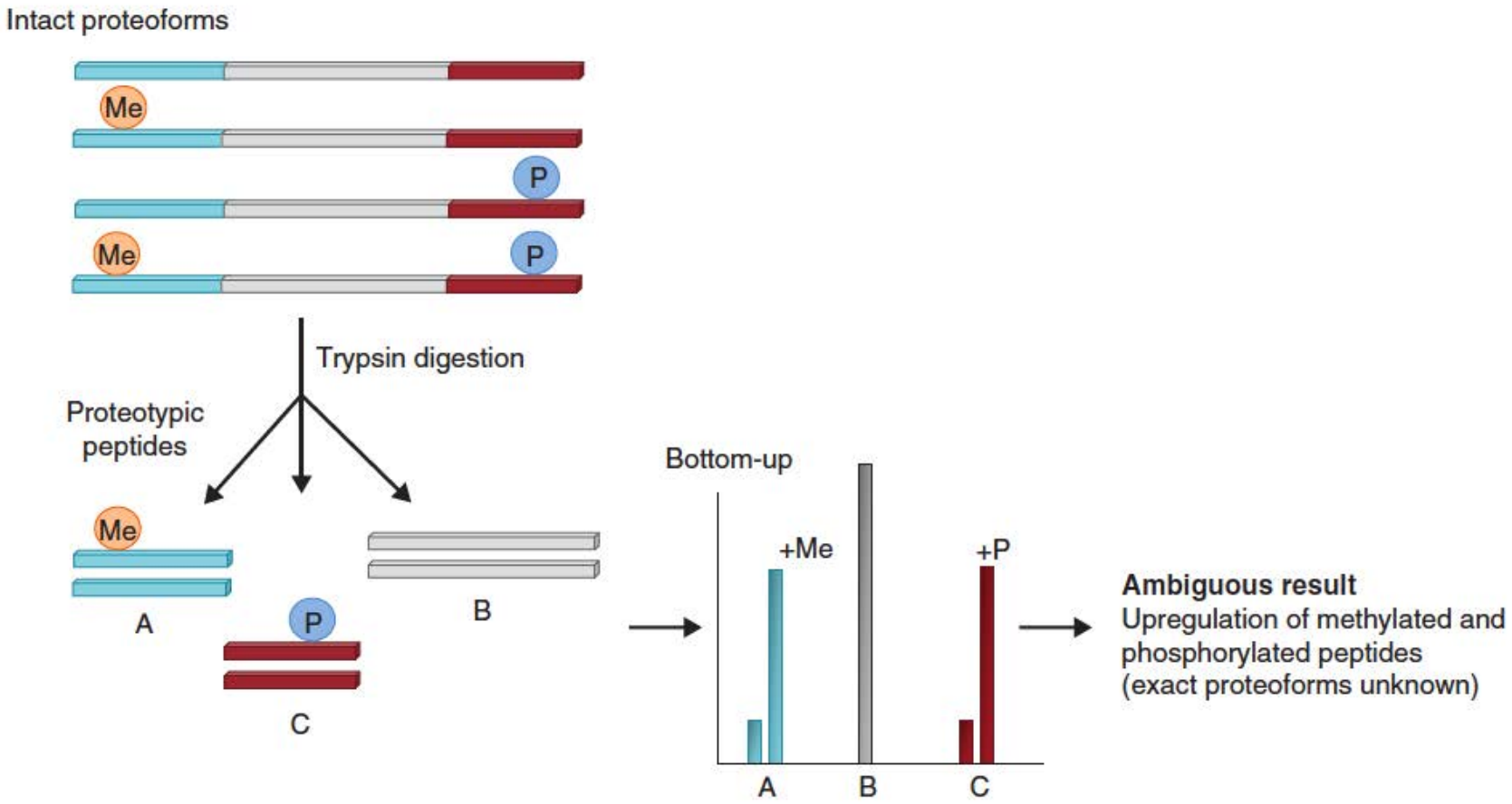




Full MS spectra acquired from intact trastuzumab

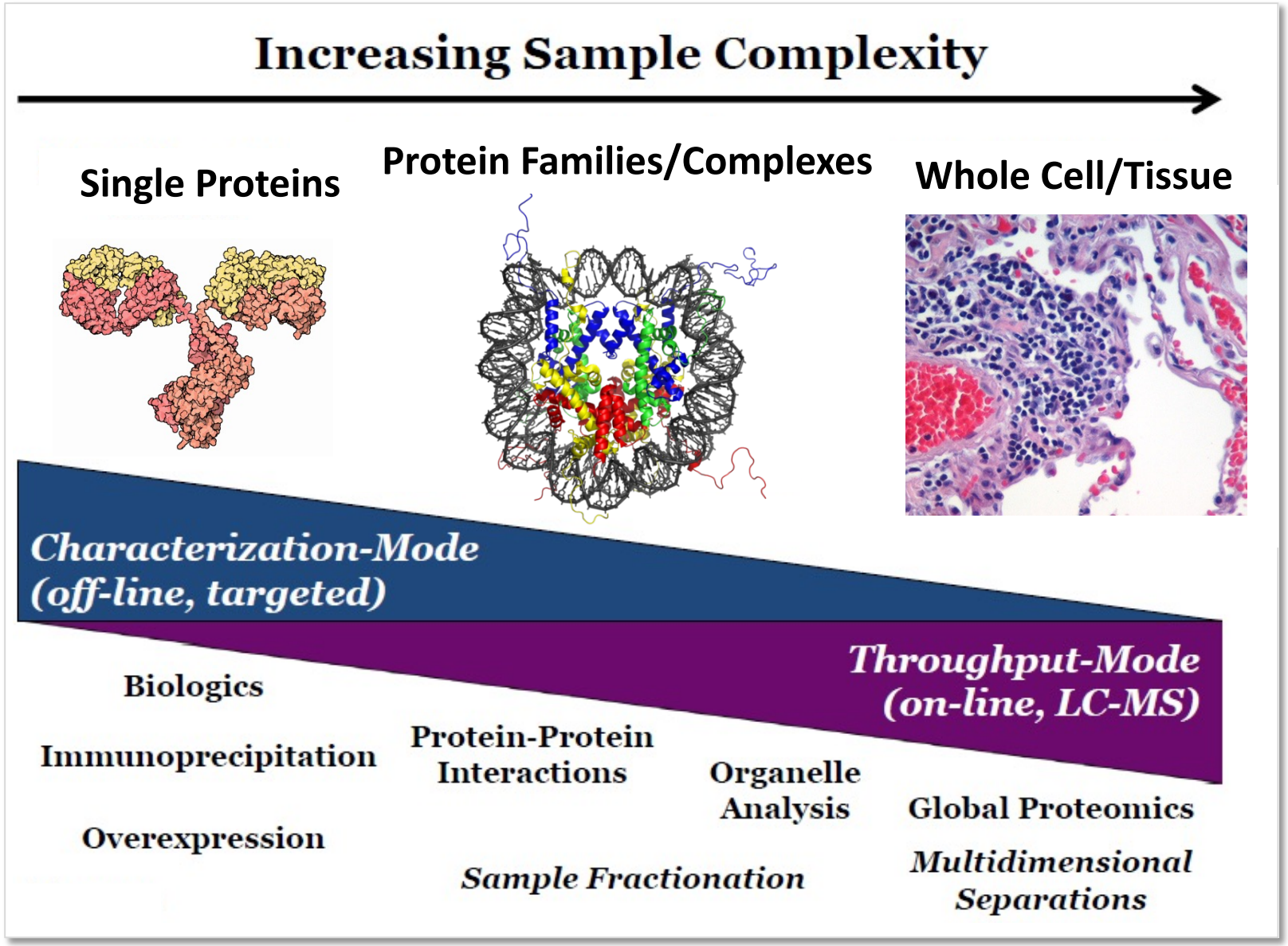
NA. Haverland *et al. J. Am. Soc. Mass. Spectrom.* 28, 1203-1215 (2017)

Proteoforms are closer to phenotypes



J. Savaryn *et al. Genome Medicine 5(53)* (2013)

Broad utility of top-down proteomics



JD. Tipton *et al. J. Biol. Chem.* 286, 25451-25458 (2011)

Guide to intact protein analysis

nature methods

PERSPECTIVE

https://doi.org/10.1038/s41592-019-0457-0

OPEN

Best practices and benchmarks for intact protein analysis for top-down mass spectrometry

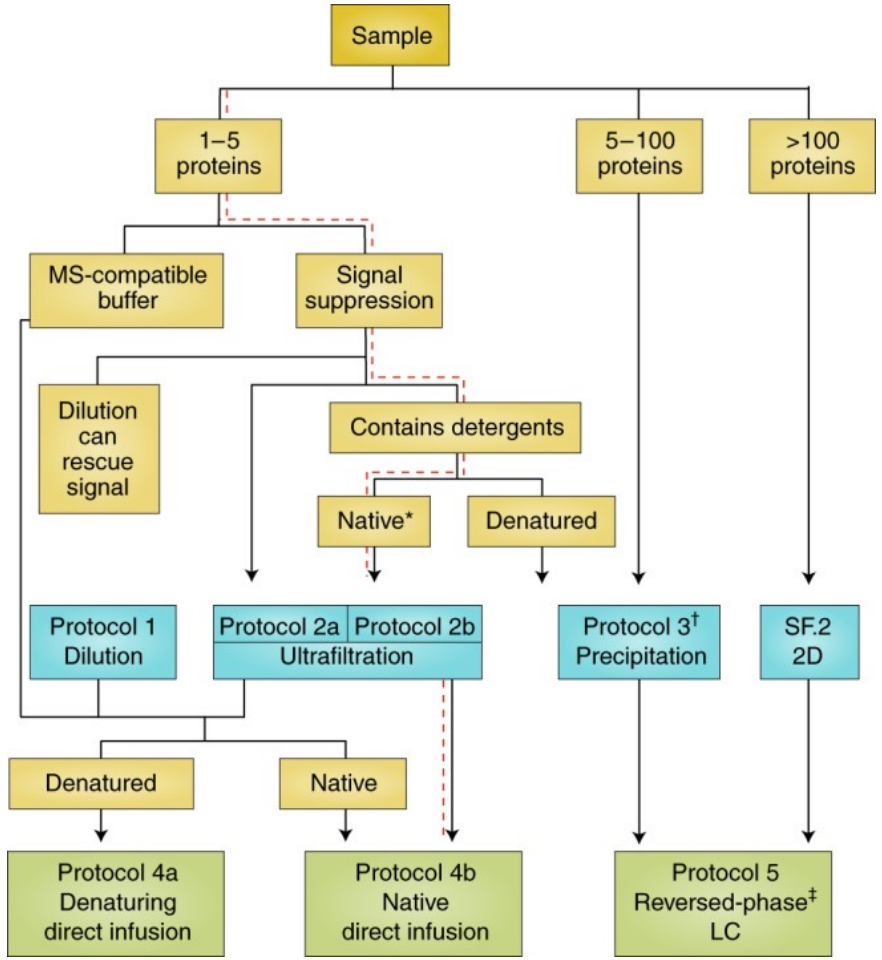
Daniel P. Donnelly[1,16], Catherine M. Rawlins[1,16], Caroline J. DeHart[2], Luca Fornelli[2], Luis F. Schachner[2], Ziqing Lin[3], Jennifer L. Lippens[4], Krishna C. Aluri[1,5], Richa Sarin[1,6], Bifan Chen[3], Carter Lantz[7], Wonhyeuk Jung[7], Kendall R. Johnson[1], Antonius Koller[1], Jeremy J. Wolff[8], Iain D. G. Campuzano[4], Jared R. Auclair[9], Alexander R. Ivanov[1], Julian P. Whitelegge[10], Ljiljana Paša-Tolić[11], Julia Chamot-Rooke[12], Paul O. Danis[13], Lloyd M. Smith[14], Yury O. Tsybin[15], Joseph A. Loo[7], Ying Ge[3], Neil L. Kelleher[2] and Jeffrey N. Agar[1*]

One gene can give rise to many functionally distinct proteoforms, each of which has a characteristic molecular mass. Top-down mass spectrometry enables the analysis of intact proteins and proteoforms. Here members of the Consortium for Top-Down Proteomics provide a decision tree that guides researchers to robust protocols for mass analysis of intact proteins (antibodies, membrane proteins and others) from mixtures of varying complexity. We also present cross-platform analytical benchmarks using a protein standard sample, to allow users to gauge their proficiency.

DP. Donnelly *et al. Nat. Methods* 16(7): 587–594 (2019)

Guide to intact protein analysis

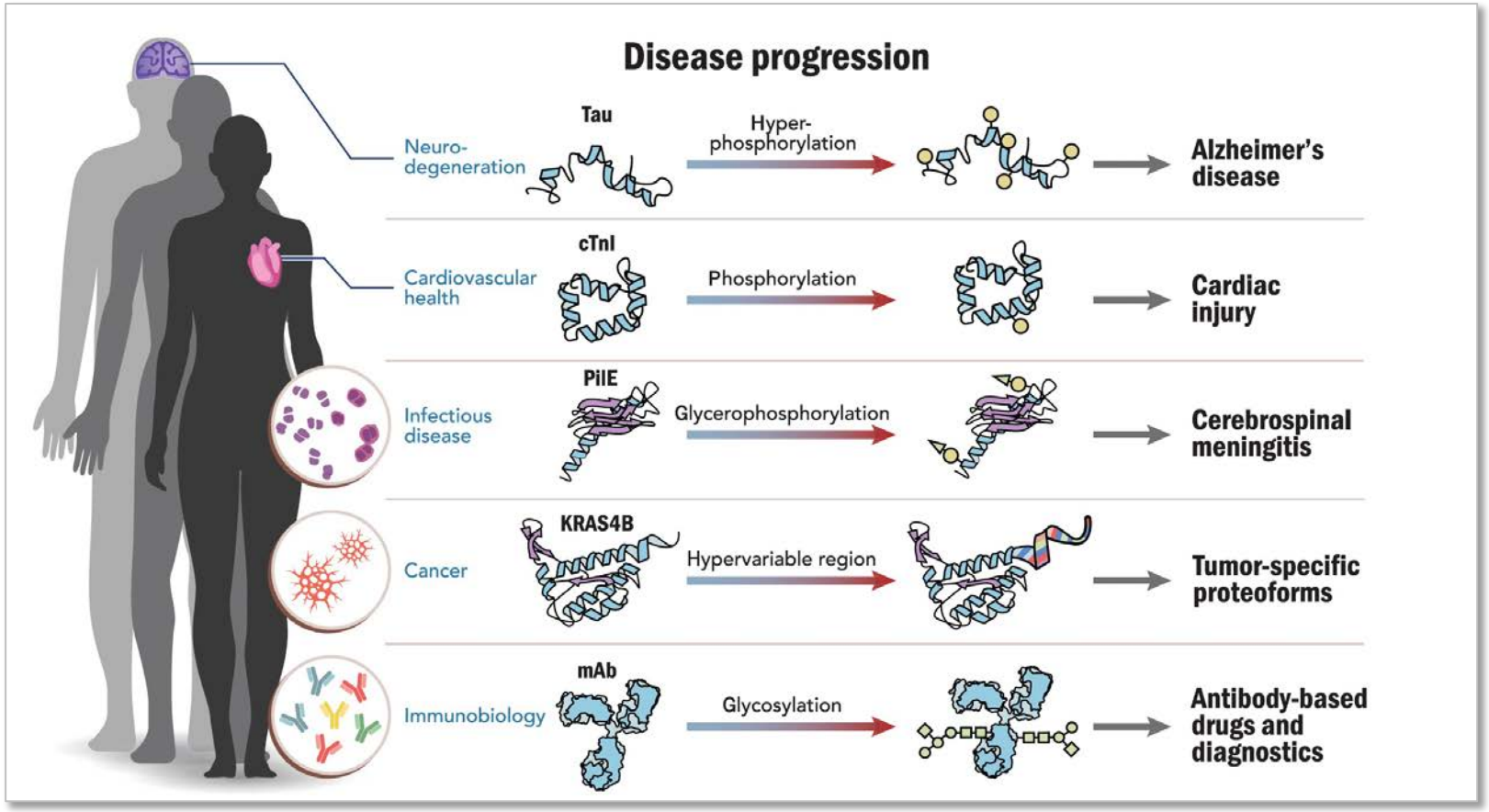
Decision tree for intact protein sample clean-up, preparation and analysis



DP. Donnelly *et al. Nat. Methods* 16(7): 587–594 (2019)

Proteoforms in health and disease

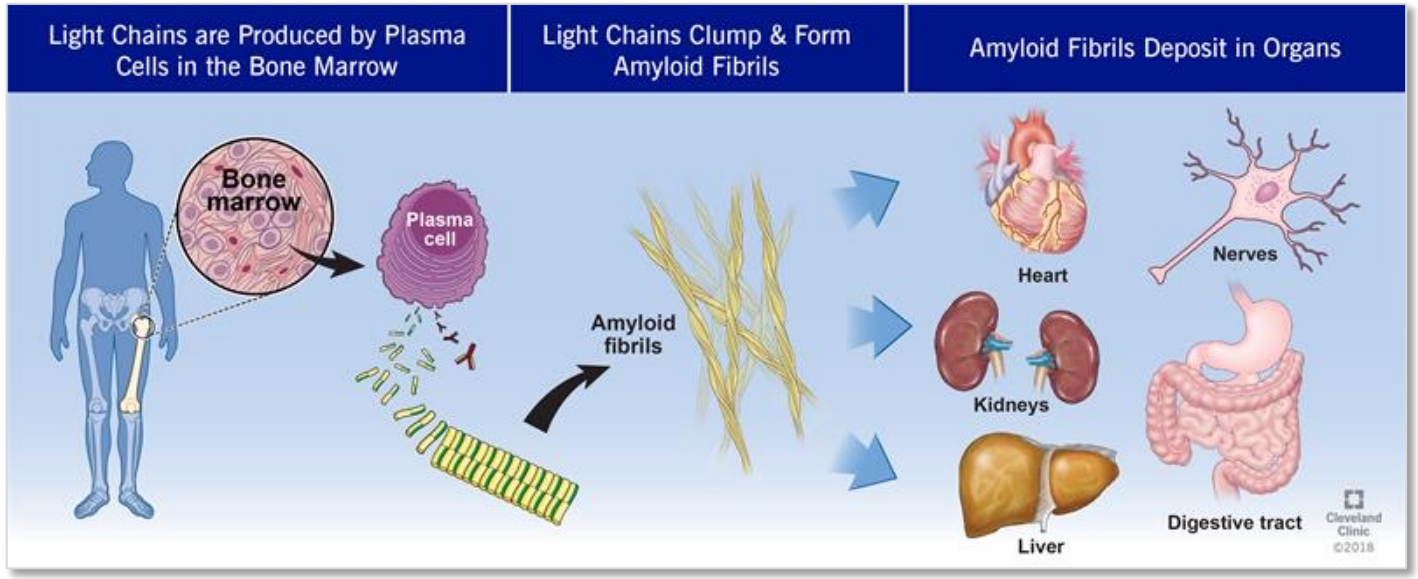
- Proteoform-level knowledge is essential to understand biological function
- Important clinical areas of interest where proteoforms have been identified and linked to disease progression



L. Smith *et al*. *Science Advances* 7, eabk0734 (2021)

Multiple Myeloma & Light Chain Disease

- Multiple myeloma is a malignancy of plasma cells characterized by a clonal expansion of abnormal B-cells
- B-cells accumulate in the bone marrow and secrete large amounts of monoclonal light chains that can deposit in organs such as kidneys leading to:
  - Light Chain Deposition Disease (LCDD) for aggregates of amorphous nature
  - AL-amyloidosis, where aggregates consist of amyloid fibrils

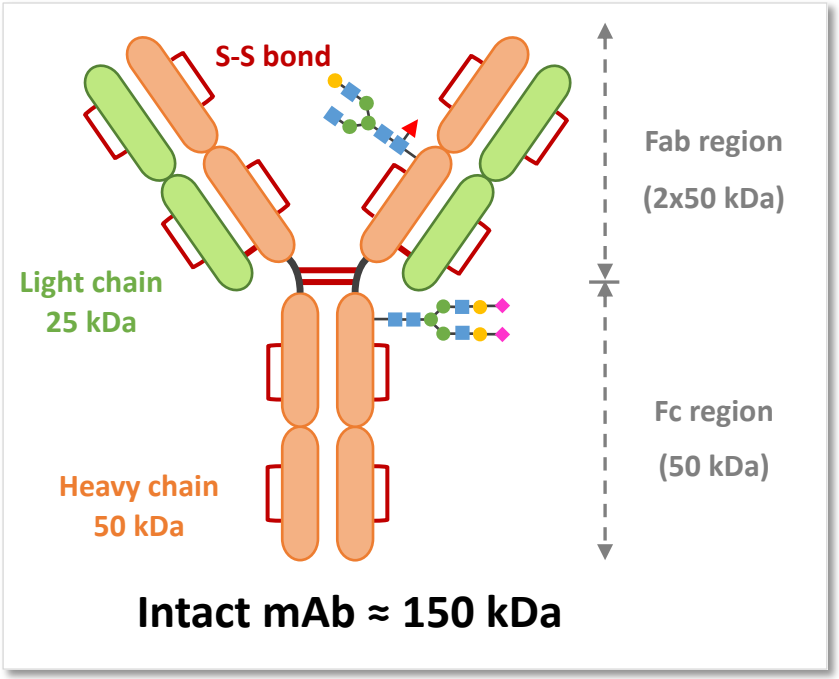


Amyloid light-chain (AL) amyloidosis

https://my.clevelandclinic.org/health/diseases/15718-amyloidosis

Light Chain (LC) Analysis

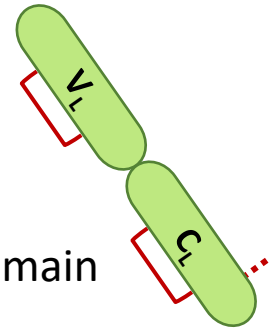
- Currently no possibility to predict the *in vivo* solubility and deposition behavior of a particular LC
- No link between LC abundance and disease severity
- To understand the factors affecting solubility and develop diagnostic tools, LC sequence is essential
  - RNA sequencing of B-cell clones
  - Mass Spectrometry



Light chain

Variable domain
(3 CDRs)

Constant domain



~ 25 kDa

5 cysteines

2 S-S bond

Sequencing of mAbs by MS

- MS extensively used for the characterization of recombinant mAbs (BUP with different enzymes)
- For unknown mAbs:
  - Combination of bottom-up proteomics and intact mAb mass profiling
  - Combination of bottom-up and middle-down proteomics on Fab (50 kDa) after sample fractionation
  - Top-down proteomics on LCs extracted from human sera for classifying plasma cell disorders (21T FT-ICR MS)

Limitation of current methods

- Gaps in sequences
- 70% sequence coverage for LC study (TDP)
- No I/L differentiation, S-S assignment

Our objective

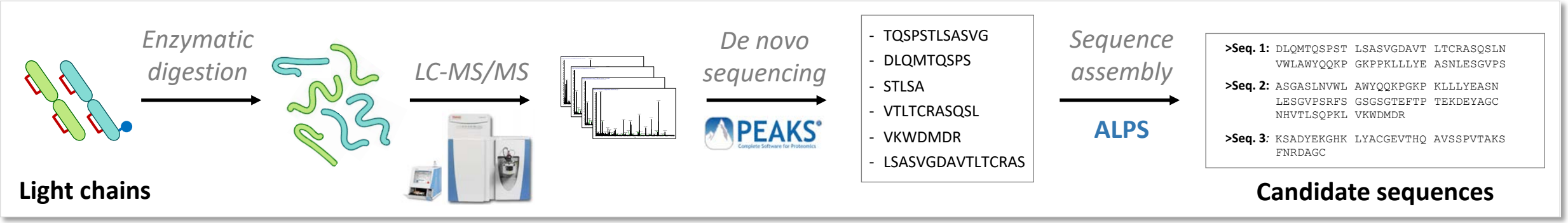
Develop a complete *de novo* MS sequencing workflow for patient-derived LC proteoforms

Collaboration A. Buell (Technical University of Denmark)

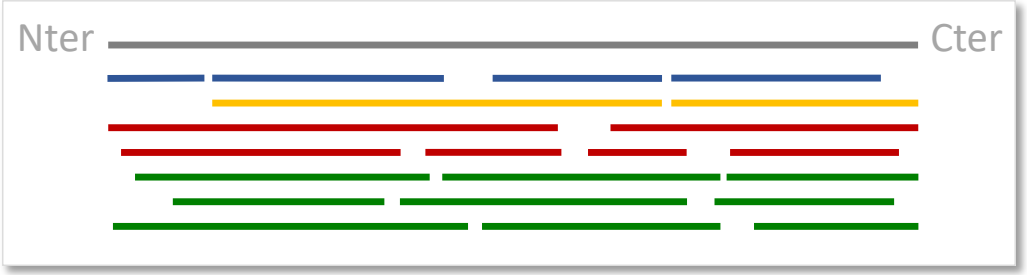
EPIC-XS

Peng *et al. J. Proteome Res.* 20(7):3559-3566 (2021)
He *et al. Anal. Chem.* 91, 3263−3269 (2019)
Bondt *et al.* J. *Cell Systems* 12(12):1131-1143 (2021)

De novo sequencing: step 1 (multiple digestions)



- Enzymatic digestions
  - Specific cleavage — Trypsin (K, R); Lys C (K)
  - Non-specific cleavage — Pepsin; *Nepenthes* fluid



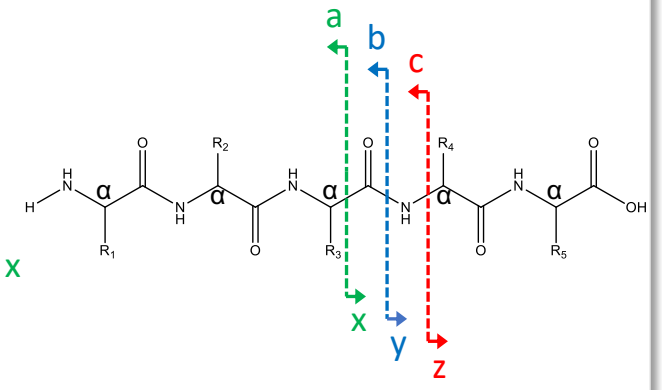
- Sequence assembly

Rey M *et al. Mol. Cell. Proteomics* 12 (2), 464–472 (2013)
Tran N.H. *et al. Scientific Reports* 6:31730 (2016)

De novo sequencing: step 2 (analysis of intact proteins)

- Sample preparation: **no digestion!**
  +/- [reduction, alkylation] of S-S bonds
- LC-MS/MS on Orbitrap Fusion Lumos
  Multiple fragmentation techniques
- Data analysis
  Xtract (FreeStyle), ProsightLite, 5 ppm

Frag. modes	Ion types
CID / HCD	b, y
ETD	c, z
EThcD	b, y, c, z
UVPD	b, y, c, z, a, x



Example for the P15 sample

4 LC-MS/MS runs
(Trypsin, Lys-C, Pepsin, *Nepenthes* fluid)

↓ PEAKS

26,859 unique sequences

↓ ALPS, k-mer 8

Top 2 sequences

Candidate seq. 1

M_{theo}: **23,461.50 Da**

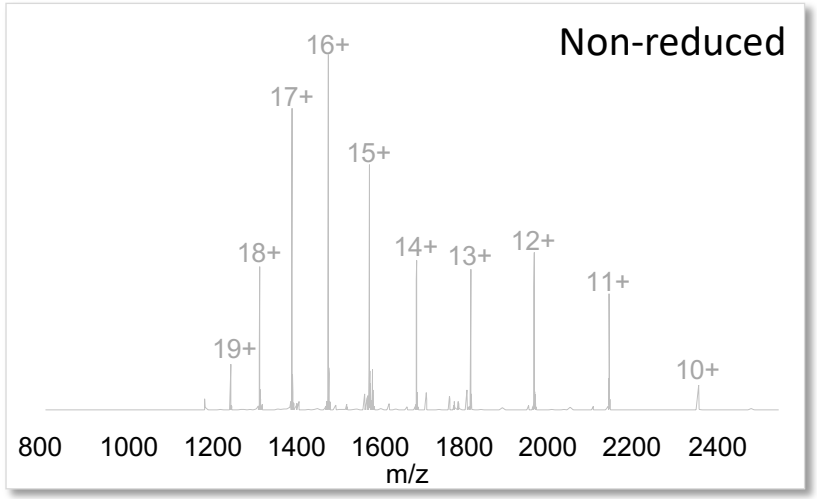
DLQMTQSPST LSASVGDAVT LTCRASQSLN VWLAWYQQKP GKPPKLLLYE AS**N**LESGVPS RFSGSGSGTE
FTLTLSSLQP DDFATYYCQQ YNSYPYTFGQ GAKLELKRTV AAPSVFLFPP SDEQLKSGTA SVVCLLNNFY
PREAKVQWKV DNALQSGNSQ ESVTEQDSKD STYSLSSTLT LSKADYEKHK LYACEVTHQG LSSPVTKSFN
RGEC

Candidate seq. 2

M_{theo}: **23,462.47 Da**

DLQMTQSPST LSASVGDAVT LTCRASQSLN VWLAWYQQKP GKPPKLLLYE AS**D**LESGVPS RFSGSGSGTE
FTLTLSSLQP DDFATYYCQQ YNSYPYTFGQ GAKLELKRTV AAPSVFLFPP SDEQLKSGTA SVVCLLNNFY
PREAKVQWKV DNALQSGNSQ ESVTEQDSKD STYSLSSTLT LSKADYEKHK LYACEVTHQG LSSPVTKSFN
RGEC

P15: Intact Mass Analysis



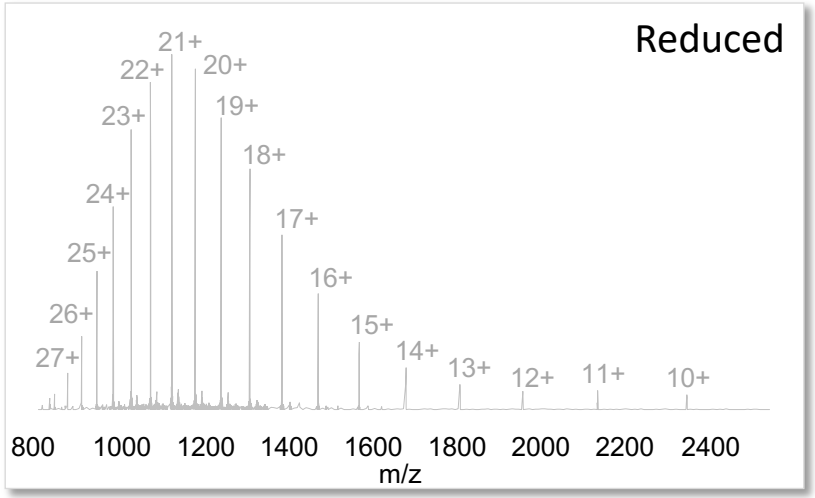
M=23,576.48 Da

Reduction

Δm= -114.93 Da

Δm = +4.03 Da ➔ 2 disulfide bonds

Δm = -118.96 Da ➔ cysteinylation (119.00 Da)



M=23,461.55 Da

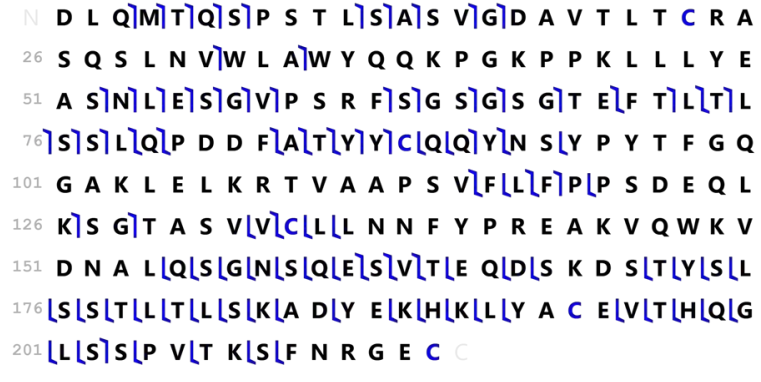
Match with sequence 1

Sequence 1
+ 2 disulfide bonds
+ 1 cysteinylation

P15: Top-down fragmentation maps

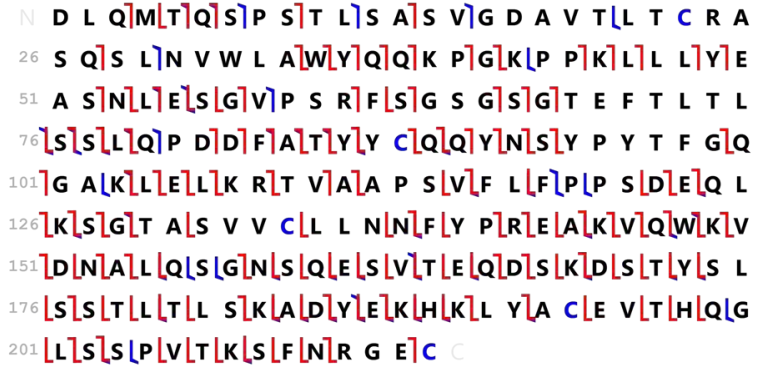
CID

NCE
20%
25%
30%



nr fragments: **114**
Residue cleavage: **46%**

EThcD



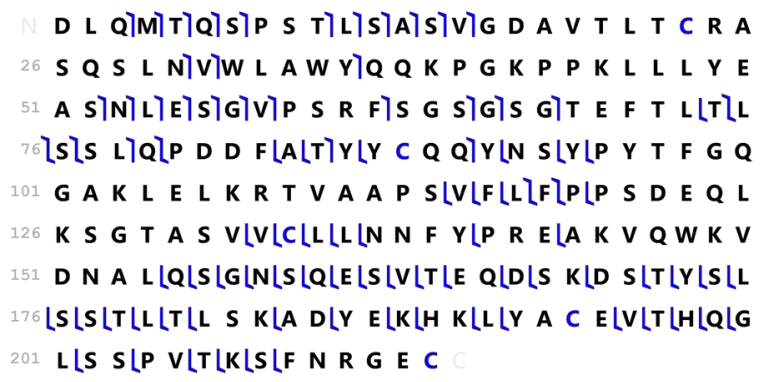
nr fragments: **253**
Residue cleavage: **69%**

Reaction time + NCE
1.5 ms + 5%
5 ms + 5%
10 ms + 10%

⌊ c/z fragment ions
⌊ b/y fragment ions
⌊ a/x fragment ions

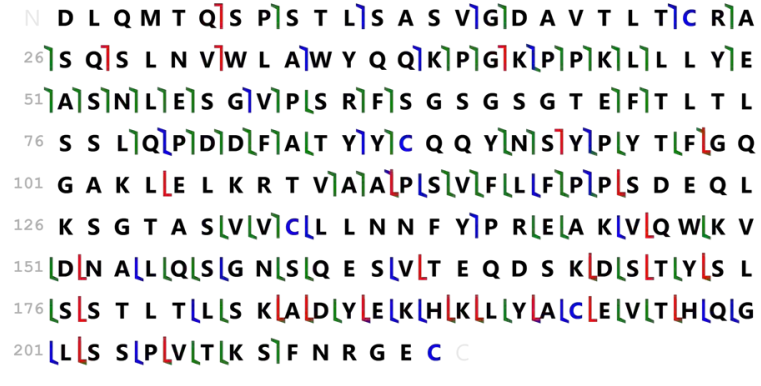
HCD

NCE
10%
12%
15%



nr fragments: **99**
Residue cleavage: **44%**

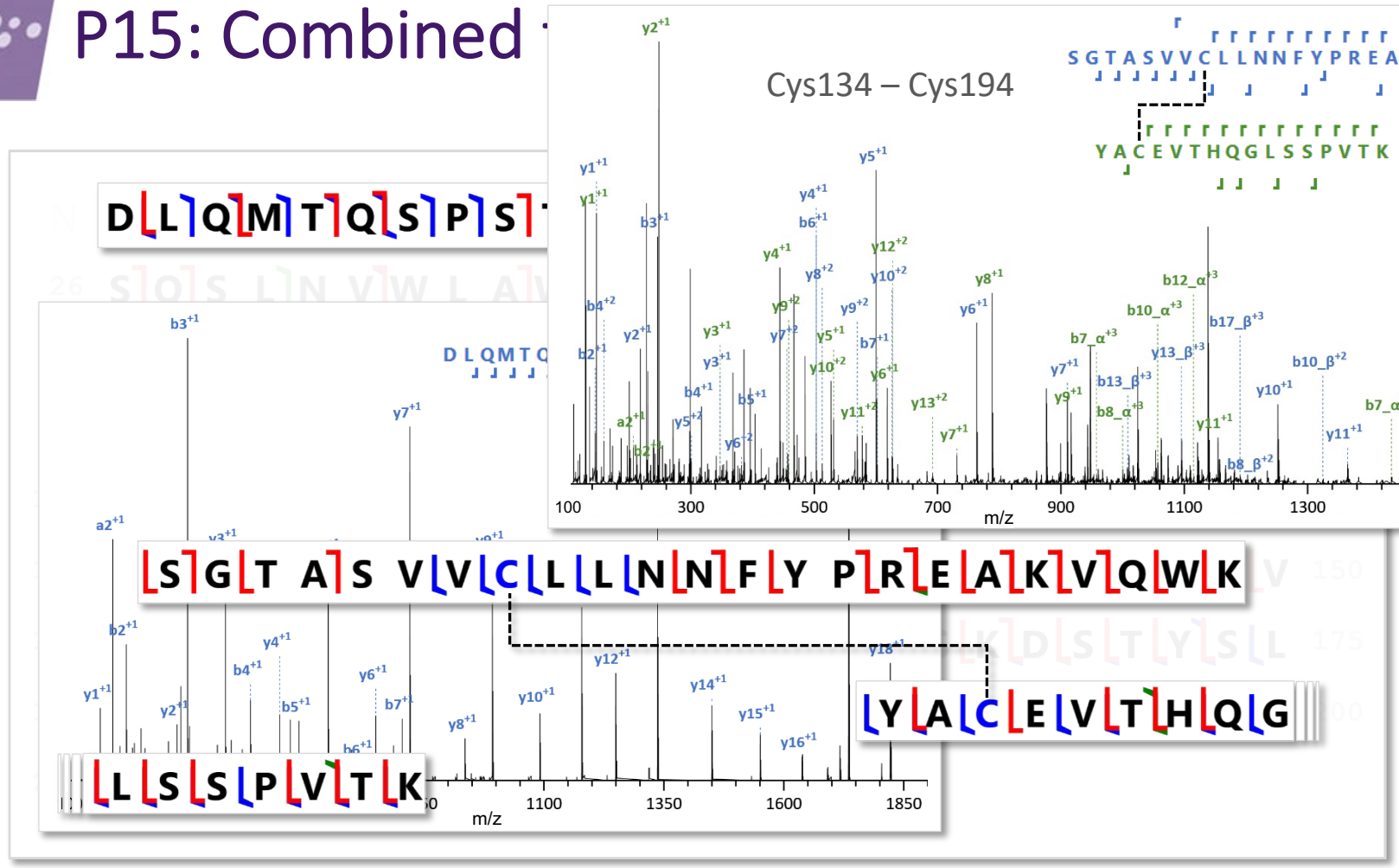
UVPD



nr fragments: **164**
Residue cleavage: **54%**

Reaction time
20 ms
25 ms
30 ms

P15: Combined



Sequence 1

2 S-S + 1 cysteinylation

Δmass (Exp. / Theo): **0.95 ppm**
nr fragments: **375**
Residue cleavage: **85%**

c/z fragment ions
b/y fragment ions
a/x fragment ions

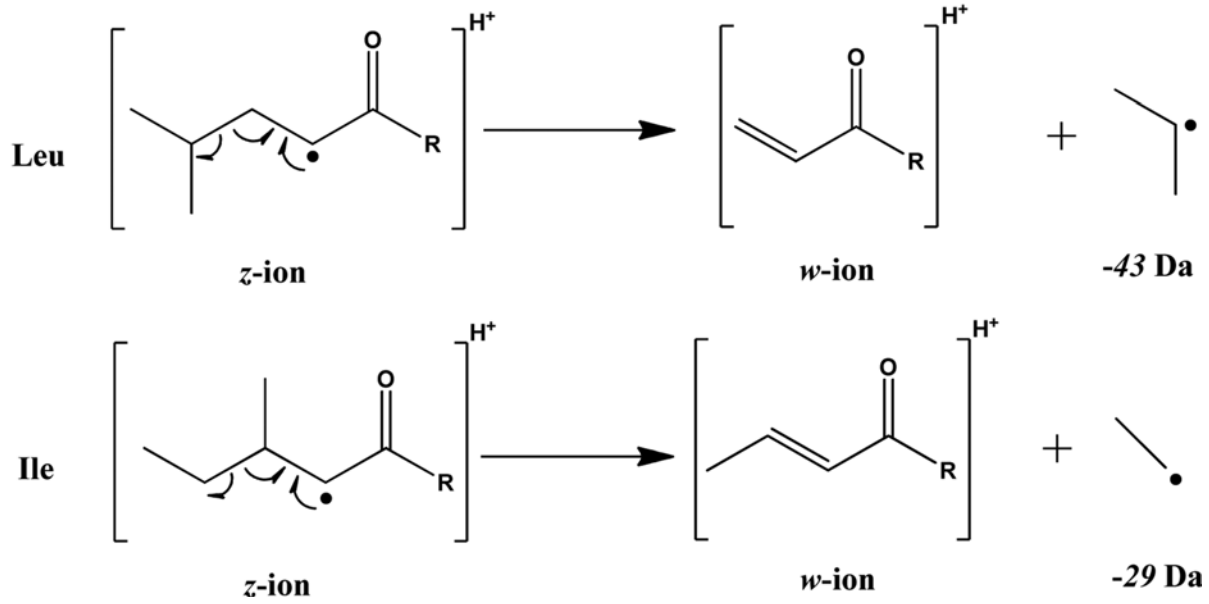
5 ppm fragment mass tolerance

- Analysis of trypsin digest (+/- reduction/alkylation)
  - 100% sequence coverage
  - Disulfides: Cys23 – Cys88, Cys134 – Cys194
  - Cysteinylation on Cys214

P15: Ile/Leu differentiation

z to w ion

EThcD



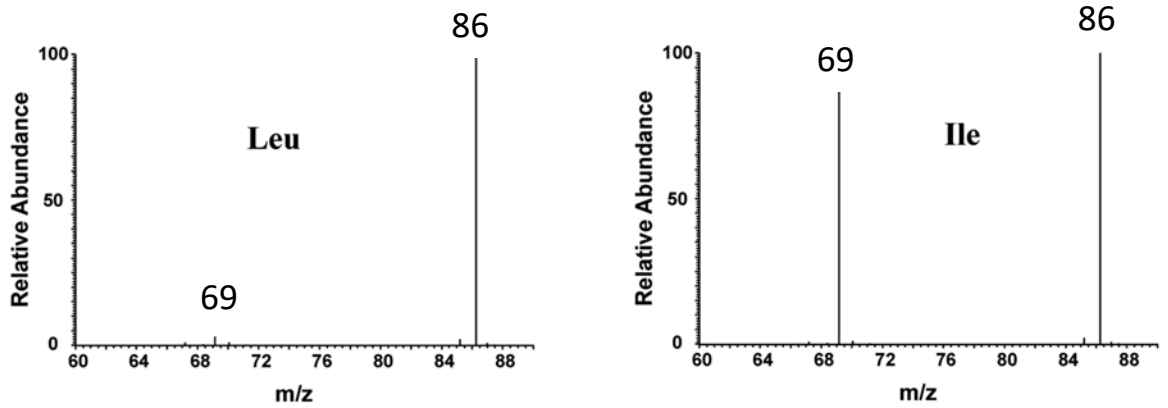
EThcD to generate z and w ions

$\Delta w\text{-}z_L$ = - 43 Da

$\Delta w\text{-}z_I$ = - 29 Da

Immonium ion

HCD



Small b or y ion containing I or L

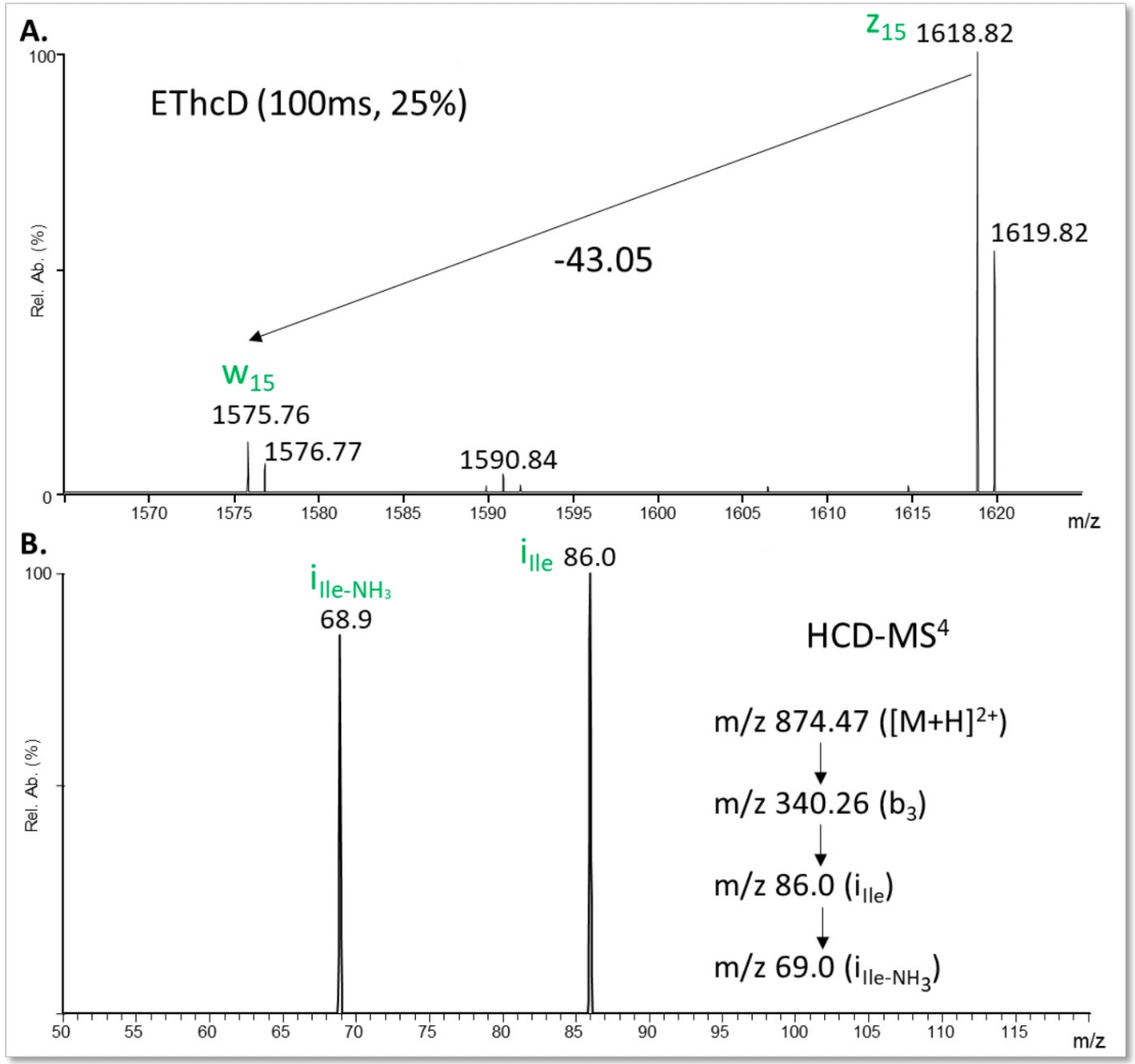
↓ HCD (MS3)

m/z 86

↓ HCD (MS4)

m/z 69

P15: Ile/Leu differentiation



LLIYEASNLESGVPSR

LLIYEASNLESGVPSR

P15: final sequence

κ isotype

CDR1 CDR2

DIQMTQSPSTLSASVGDAVTITCRASQSLNVWLAWYQQKPGKPPKLLIYEASNLESGVP

CDR3

SRFSGSGSGTEFTLTISSLQPDDFATYYCQQYNSYPYTFGQGAKLEIKRTVAAPSVFIF

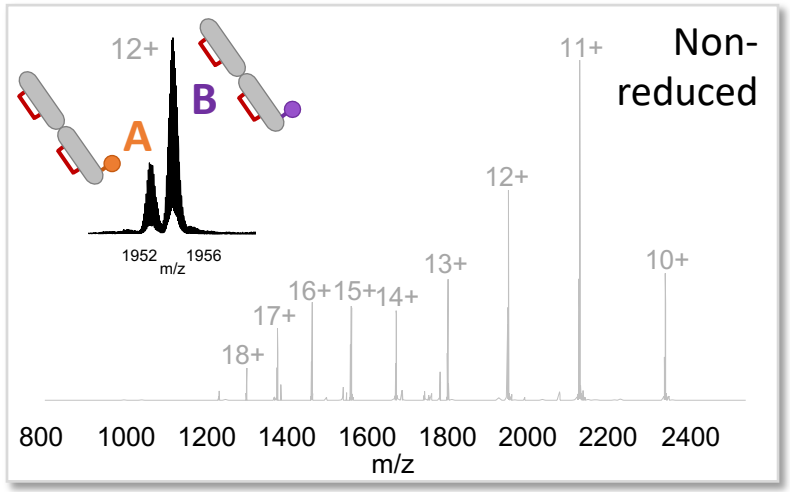
PPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSS

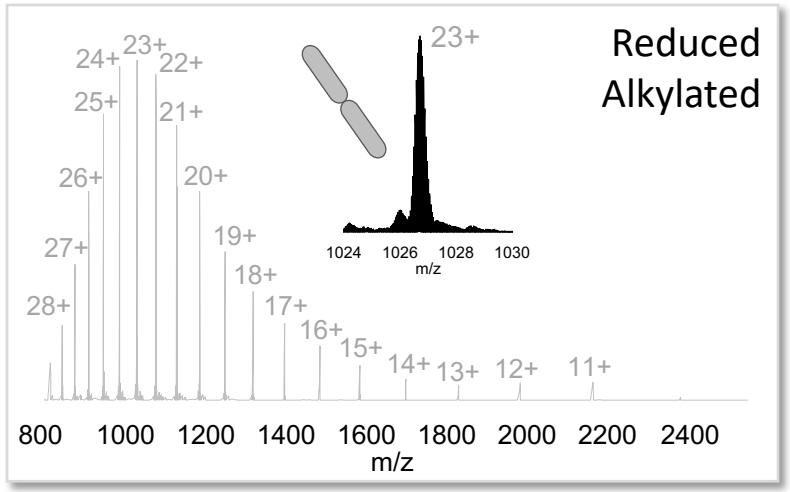
TLTLSKADYEKHKLYACEVTHQGLSSPVTKSFNRGEC*

* Cysteinylation

P8 analysis

Intact mass analysis





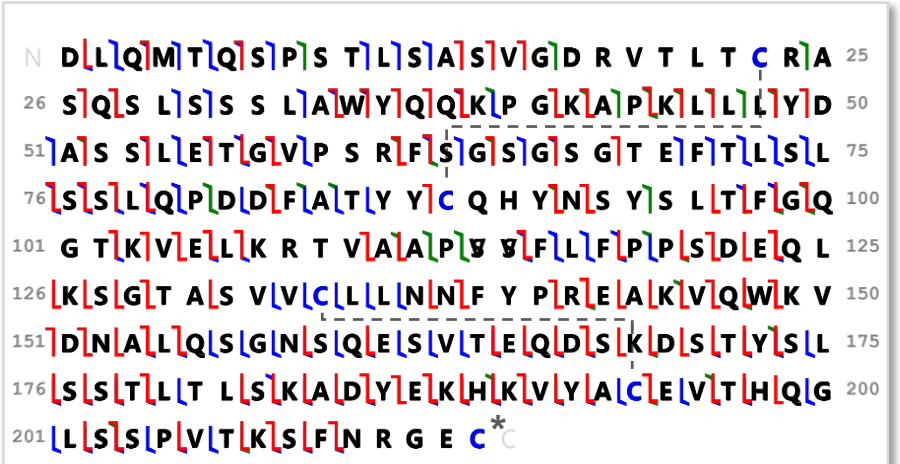
Two proteoforms @:
23,407.35 Da
23,428.39 Da

Δ= +170.13 Da *Reduction*
Δ= +149.09 Da *Alkylation*

One proteoform @:
23,577.48 Da

A single sequence with modified Cys

De novo top down fragmentation analysis



ΔMass (Exp. / Theo): **-1.03 ppm**
nr fragments: **468**
Residue cleavage: **86%**

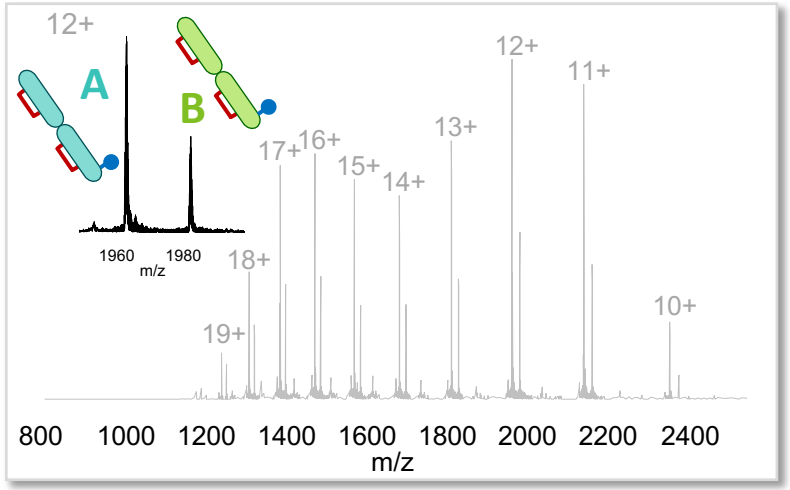
Trypsin digest

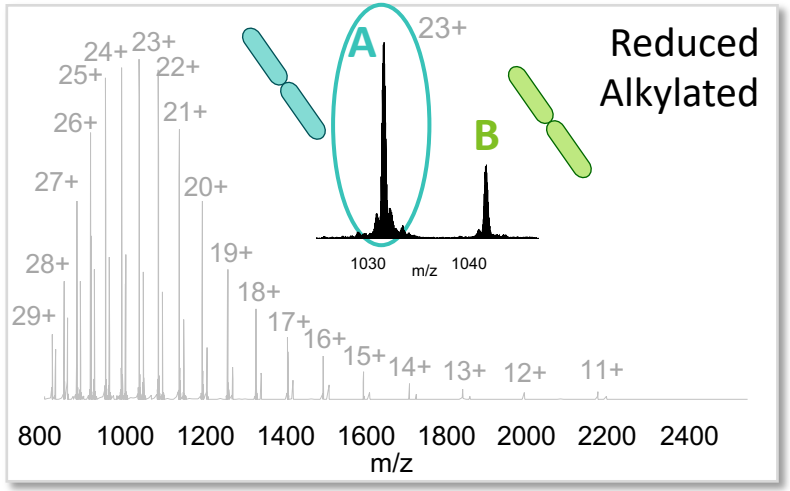
Disulfides: Cys23 – Cys88 & Cys134 – Cys194
Cys214*: Cysteinylation or CoenzymeM (CoM)

CoM: adjuvant in chemotherapy

P5 analysis

Intact mass analysis





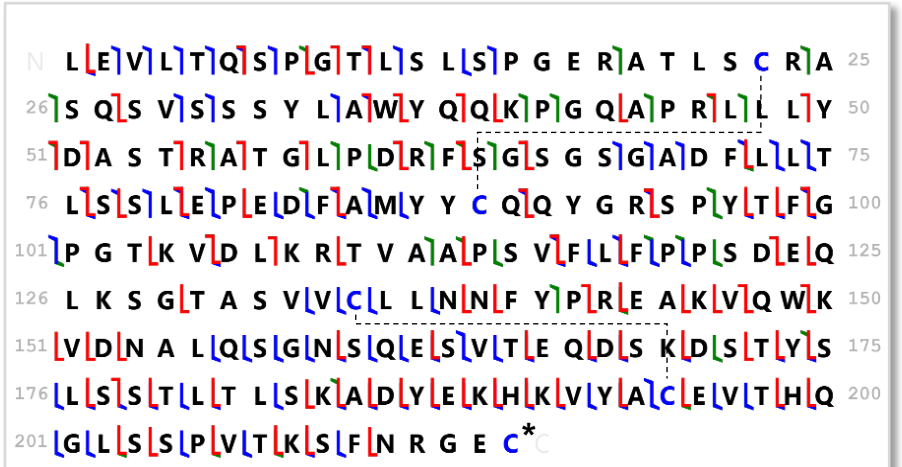
Two proteoforms @:
23,543.49 Da
23,776.61 Da

Reduction
Alkylation

Two proteoforms @:
23,713.62 Da
23,946.81 Da

Two different sequences modified Cys

Top-down fragmentation map (5A)



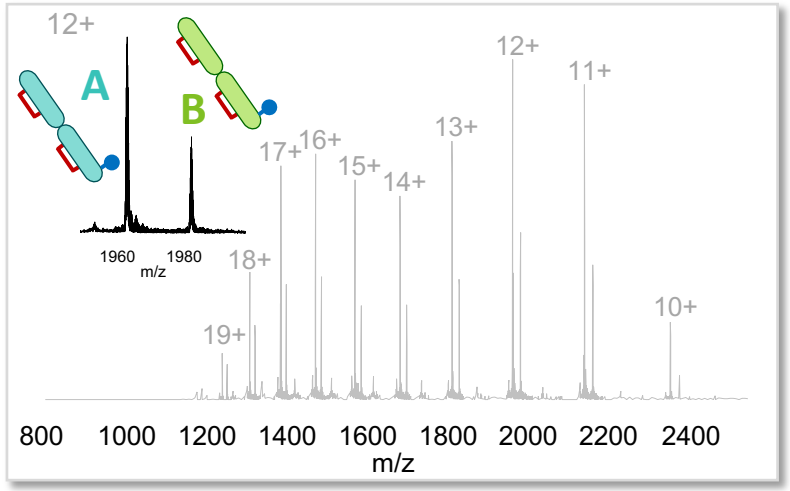
ΔMass (Exp. / Theo): **0.46 ppm**
Residue cleavage: **83%**

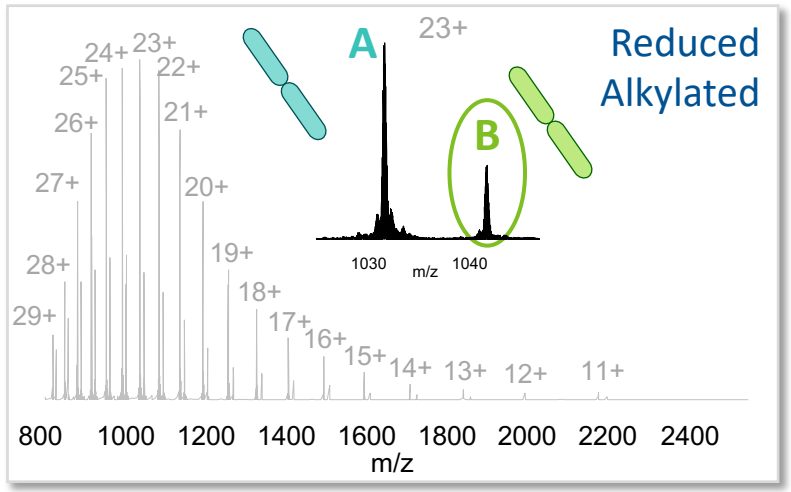
Trypsin digest

Disulfides: Cys23 – Cys89 & Cys135 – Cys195
Cys215*: Cysteinylation

P5 analysis

Intact mass analysis





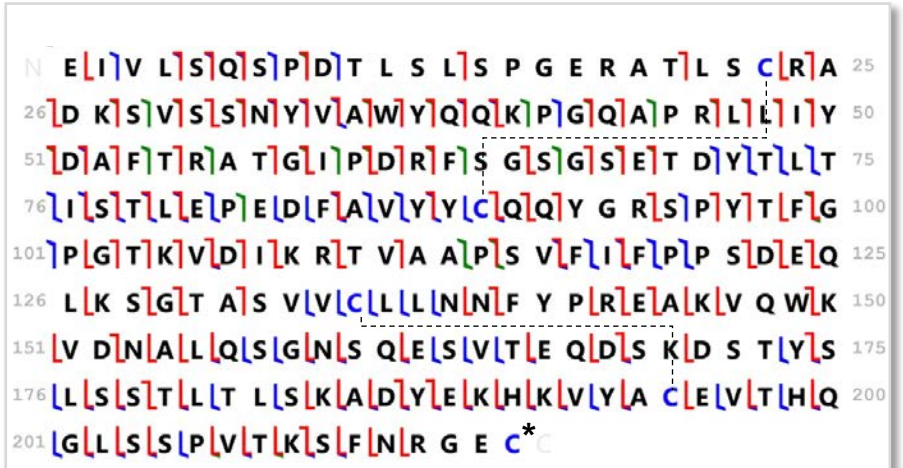
Two proteoforms @:
23,543.49 Da
23,776.61 Da

Reduction
Alkylation

Two proteoforms @:
23,713.62 Da
23,946.81 Da

Two different sequences modified Cys

Top-down fragmentation map (5B)



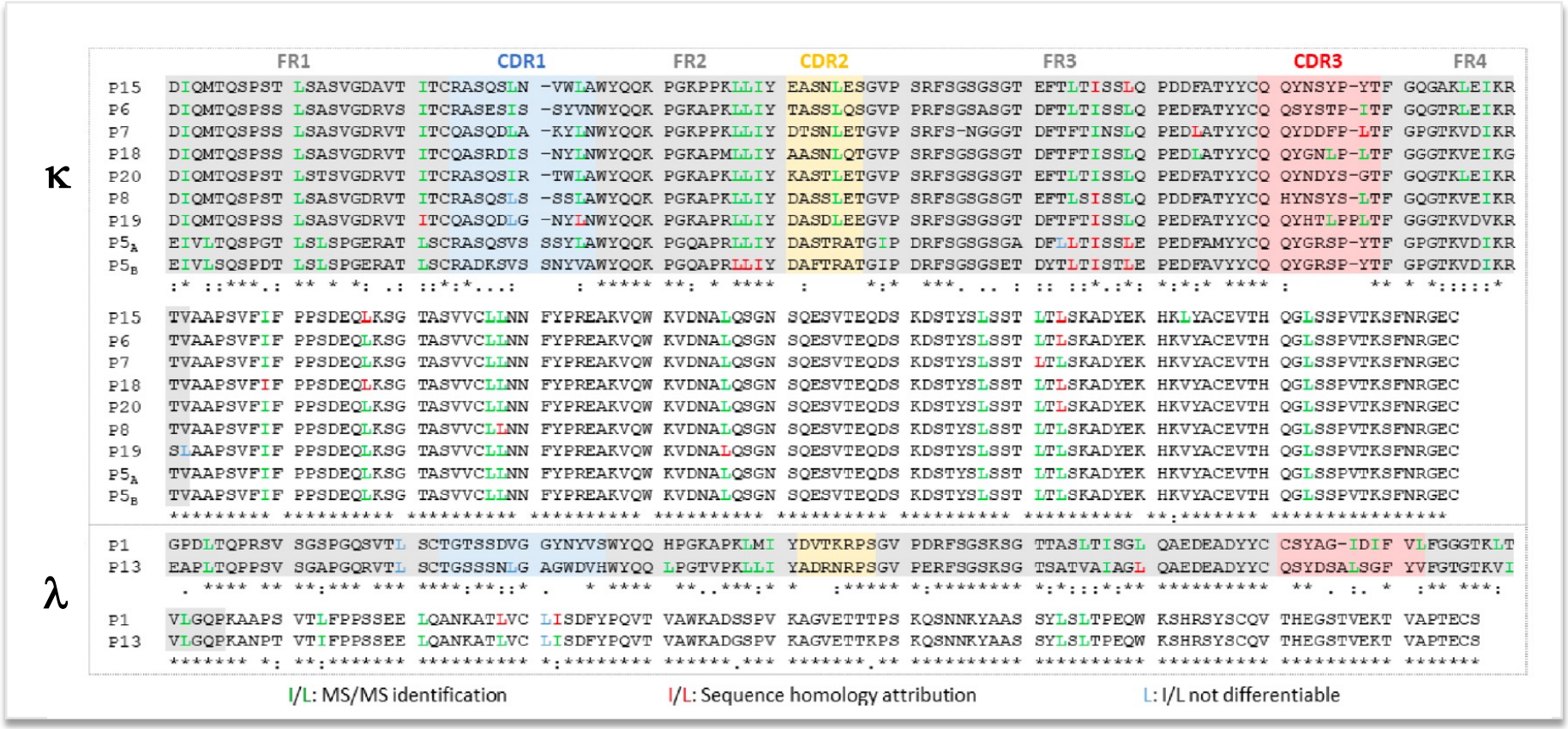
Δmass (Exp. / Theo): **0.6 ppm**
Residue cleavage: **80%**

Trypsin digest

Disulfides: Cys23 – Cys88 & Cys134 – Cys194
Cys215*: Cysteinylation
94% identity between 5A and 5B

Summary

- Light chains extracted for the urine of 10 patients fully de novo sequenced (including all CDRs)



Between 66.7% (κ) and 94% (λ) sequence homology

Unexpected range of modifications
Cysteinylation, CoenzymeM, HexNAc(1)dHex(1)

- Top-down proteomics essential to address samples with multiple proteoforms (could be improved?)

Dupré M. *et al. Anal. Chem.* 93, 10627−10634 (2021)

Improvement with PTCR (Proton Transfer Charge Reduction)

EPIC-XS

- Based on the ion-ion reactions described in the 90's by J. Stephenson & S. McLuckey in ion traps

- Principle of Proton Transfer Reaction (PTR)
  A multiply charged cation is let to react with a singly charged anion (typically originated from a perfluorinated molecule) to generate a cation with reduced charge state

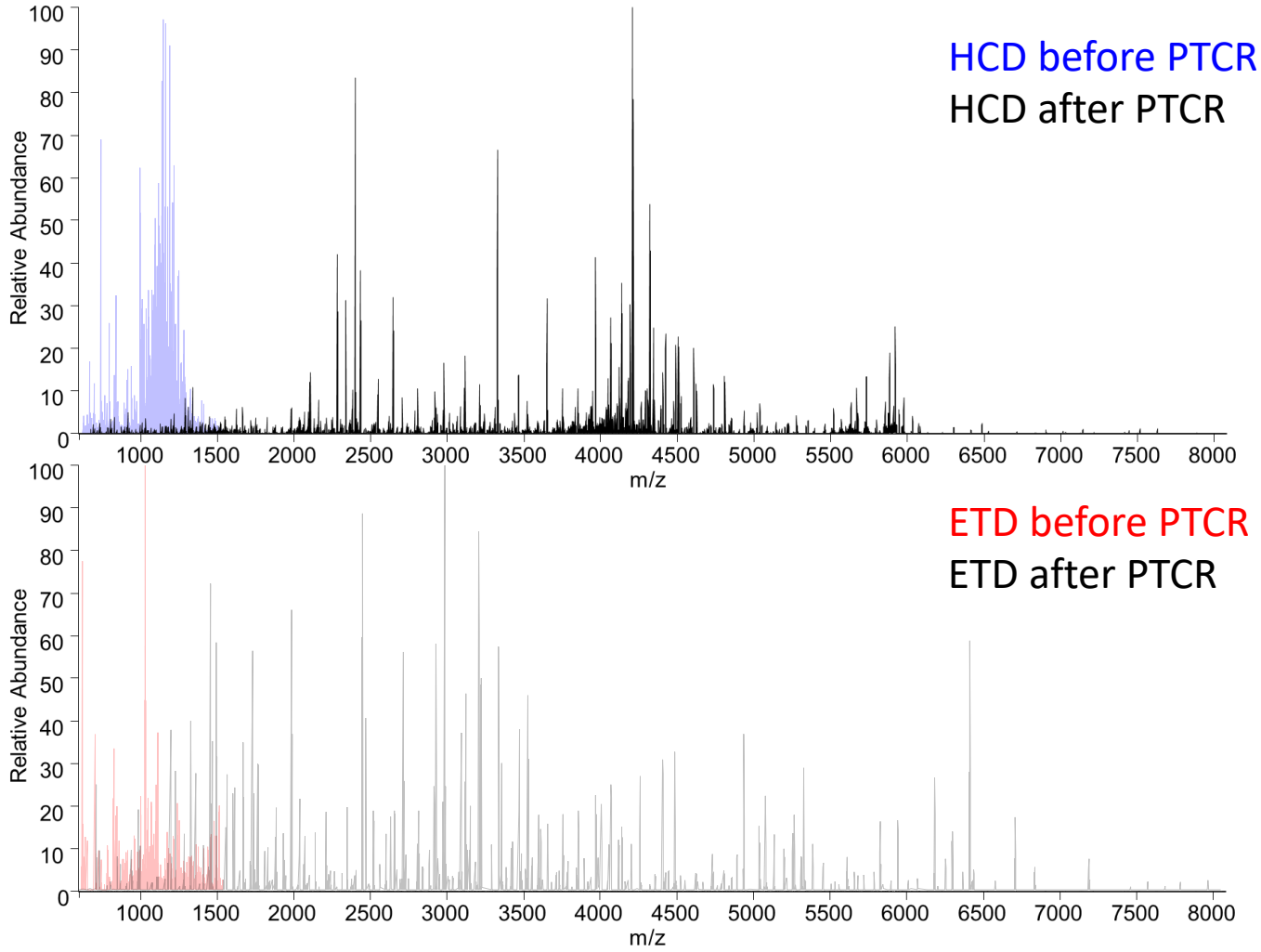
$$[\mathrm{M} + n\mathrm{H}]^{n+} + \mathrm{A}^{-} \rightarrow [M + (n-1)\mathrm{H}]^{(n-1)+} + \mathrm{HA}$$

- Can be used on both MS1 and MS2 level to decongestion spectra and improve S/N
  PTCR performed in the ion trap of tribrid mass spectrometers

Stephenson, J. L.& McLuckey, S. A. *J. Am. Chem. Soc.* 118, 7390 (1996)
Huguet R. *et al. Anal. Chem.* 91, 15732 (2019)

Example of PTCR analysis

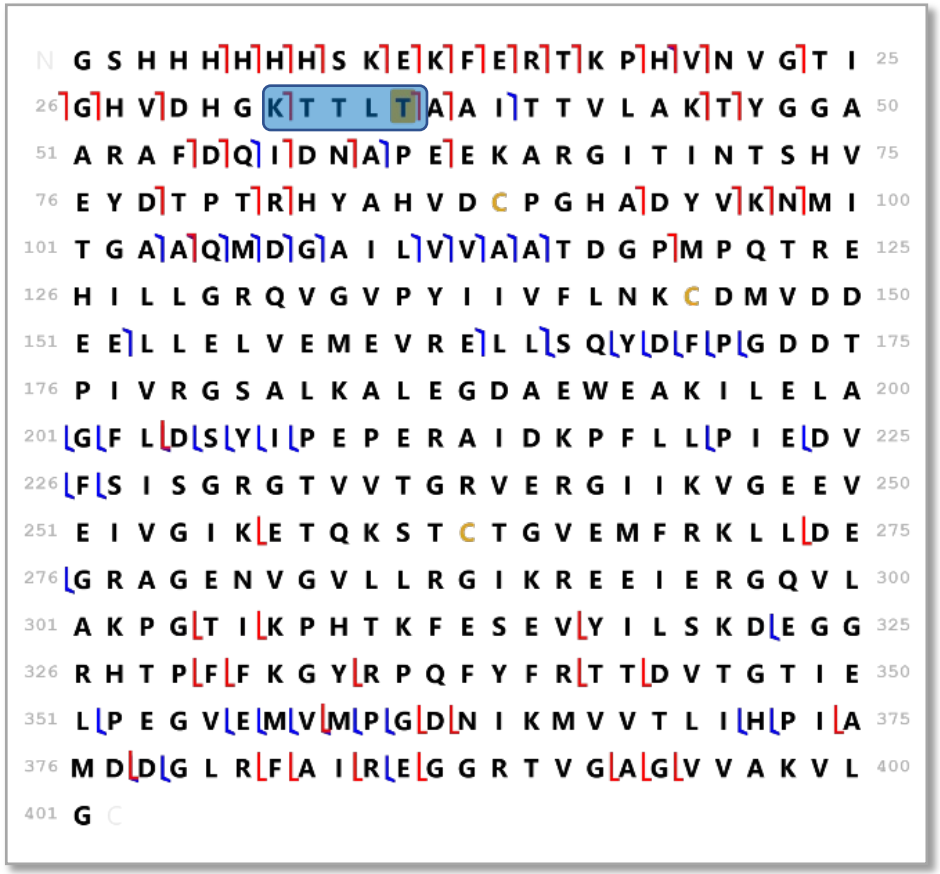
- Localization of an unknown modification on a 44 kDa protein



- Orbitrap Tribrid Eclipse
- PTCR 100 ms, 1E6 reagent target
- PTCR mass range selection: 500 to 1500 m/z
- Mass range: 600-8000 m/z
- Resolution = 120k

Improved sequence coverage with PTCR

- Precise localization of an ADP-ribosylation on Asp29



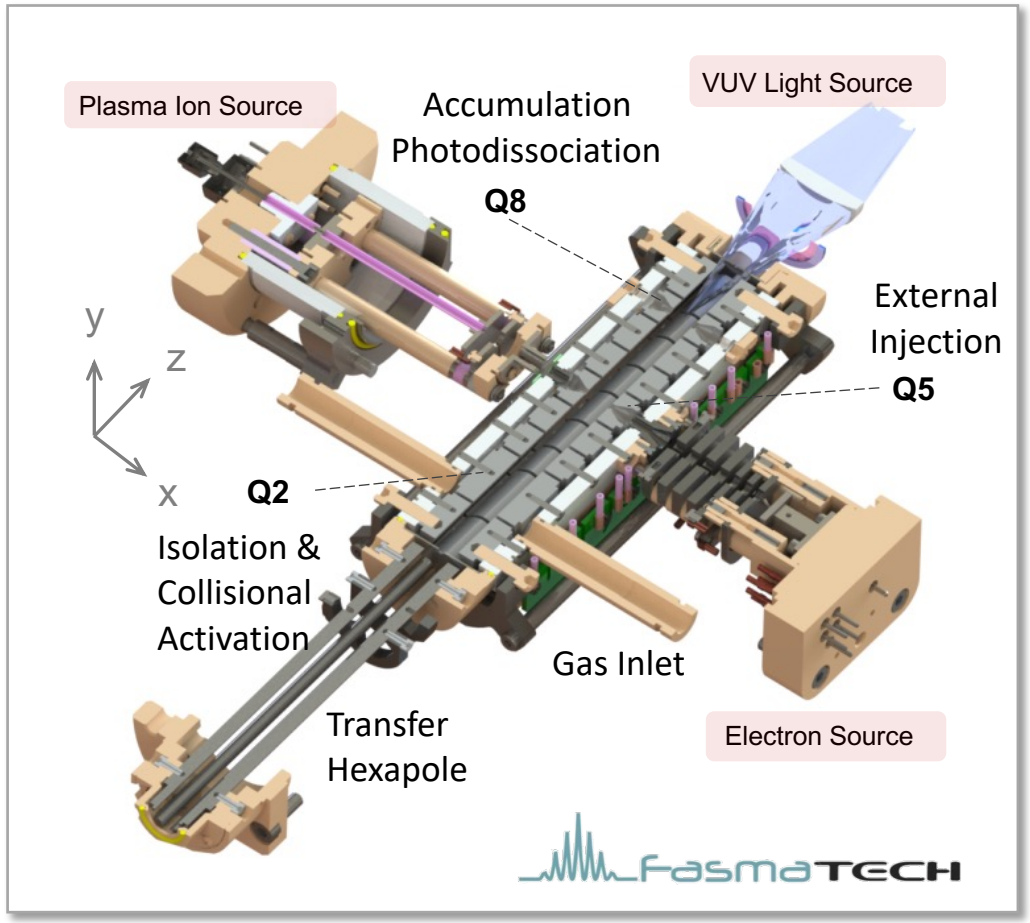
130 unique fragments
30% coverage

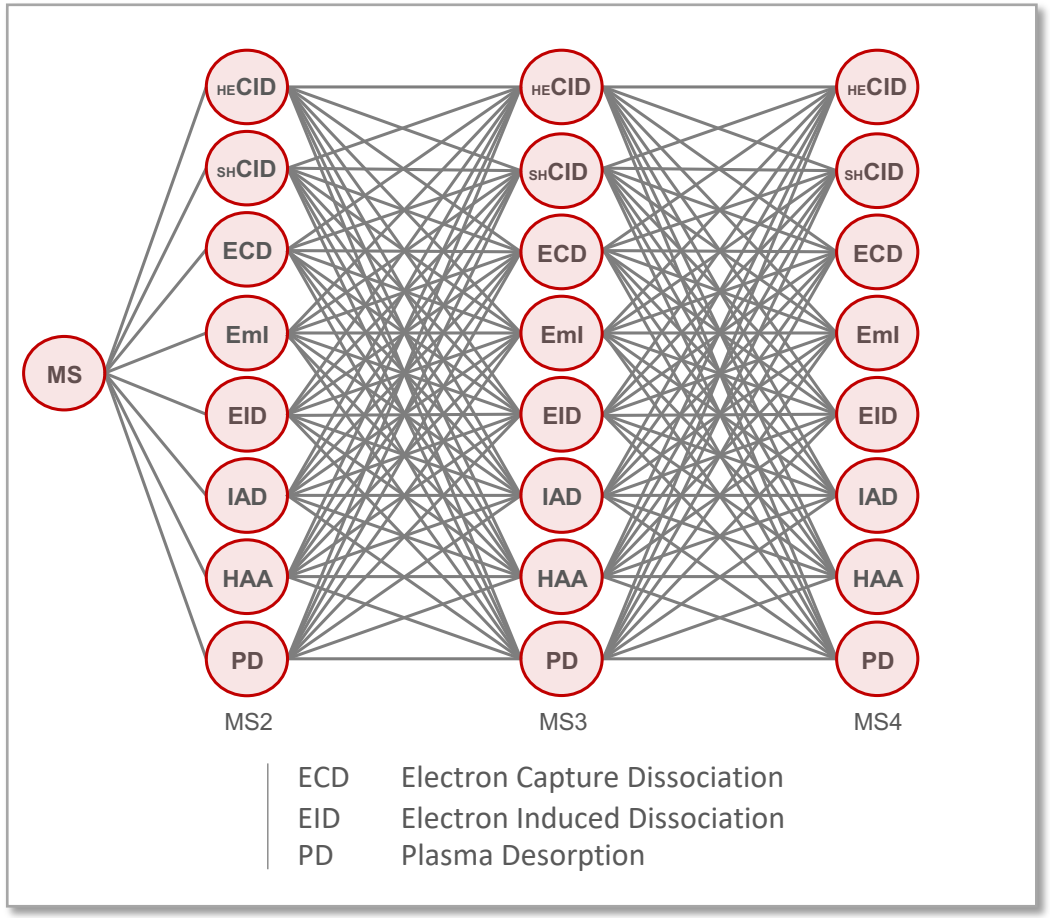
Combined map HCD, ETD, w&w/o PTCR
238 unique fragments
45% coverage

b/y fragment
c/z fragment
ADP-ribosylation

Omnitrap platform for improved top-down analysis

- A novel segmented linear ion trap for enhanced activation and ion storage
- Arsenal of ion activation techniques: CID, ECD, EID (E> 10 eV), photo-dissociation and others…

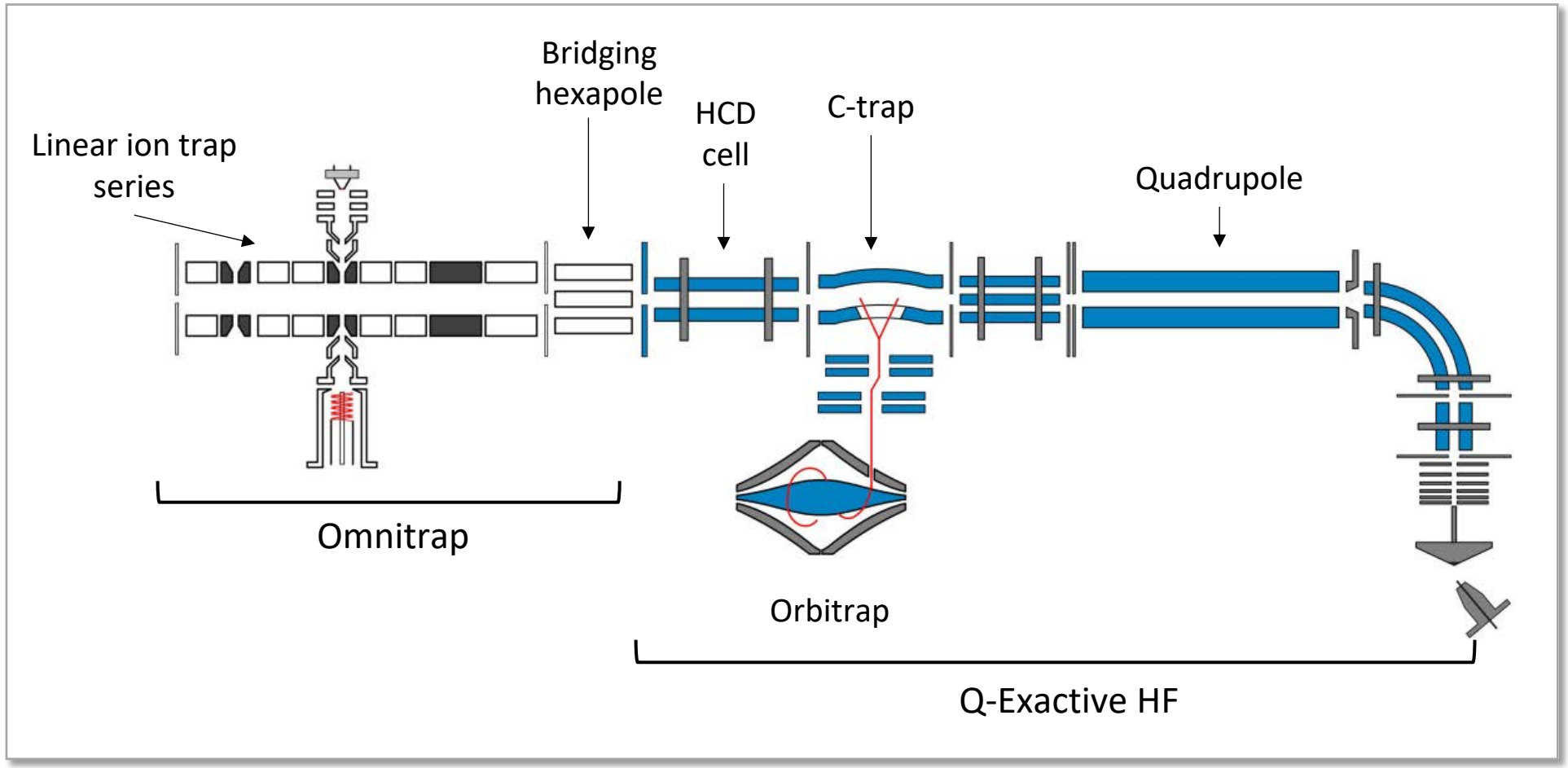




D. Papanastasiou *et al. J. Am. Soc. Mass Spectrom.* (2022)

Q-Exactive HF modified with the Omnitrap platform

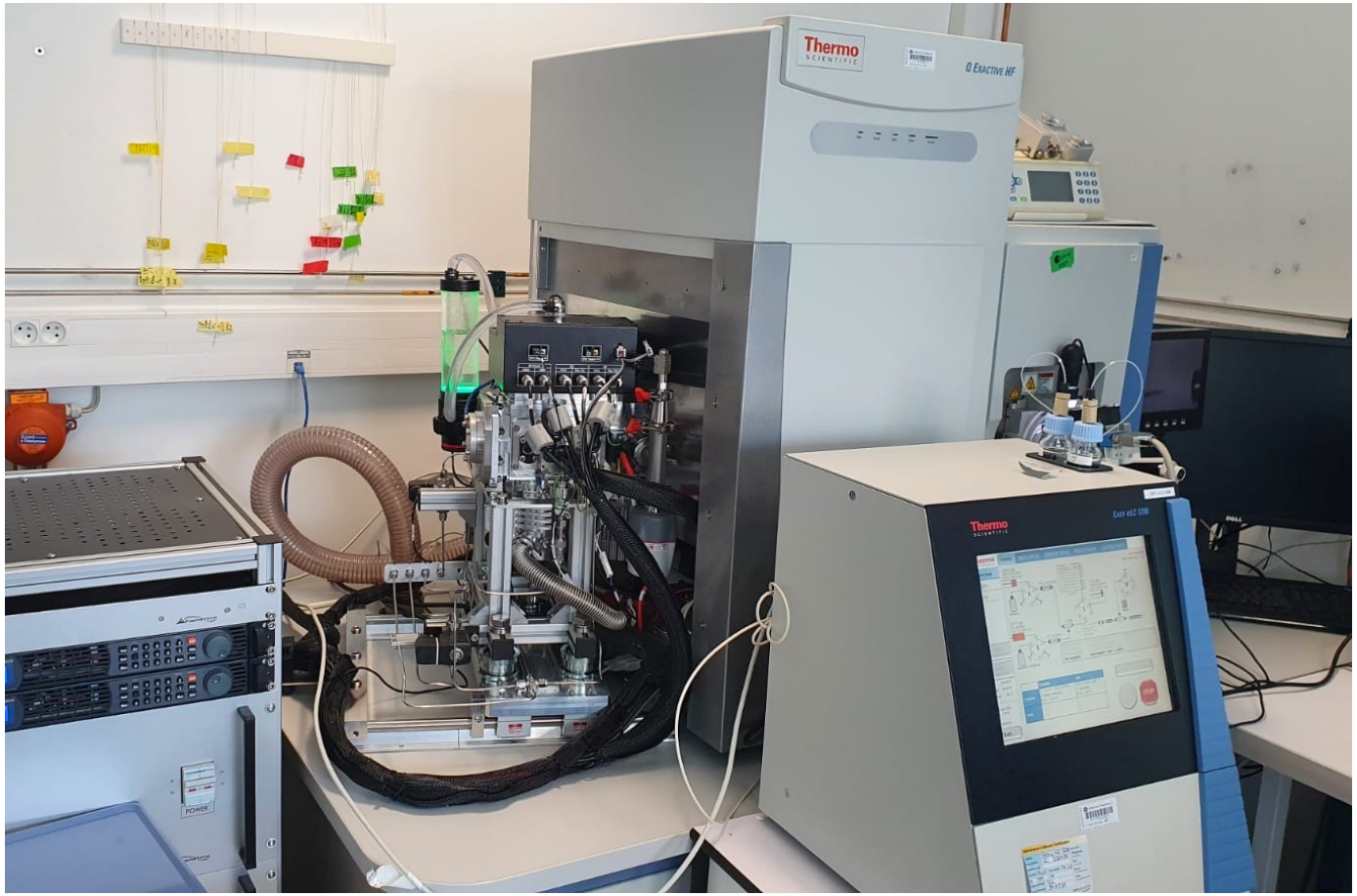
- Currently available as a retrofit to the Q-Exactive/Exploris instrument series



Application areas: top-down/bottom-up proteomics, native MS, glycomics…

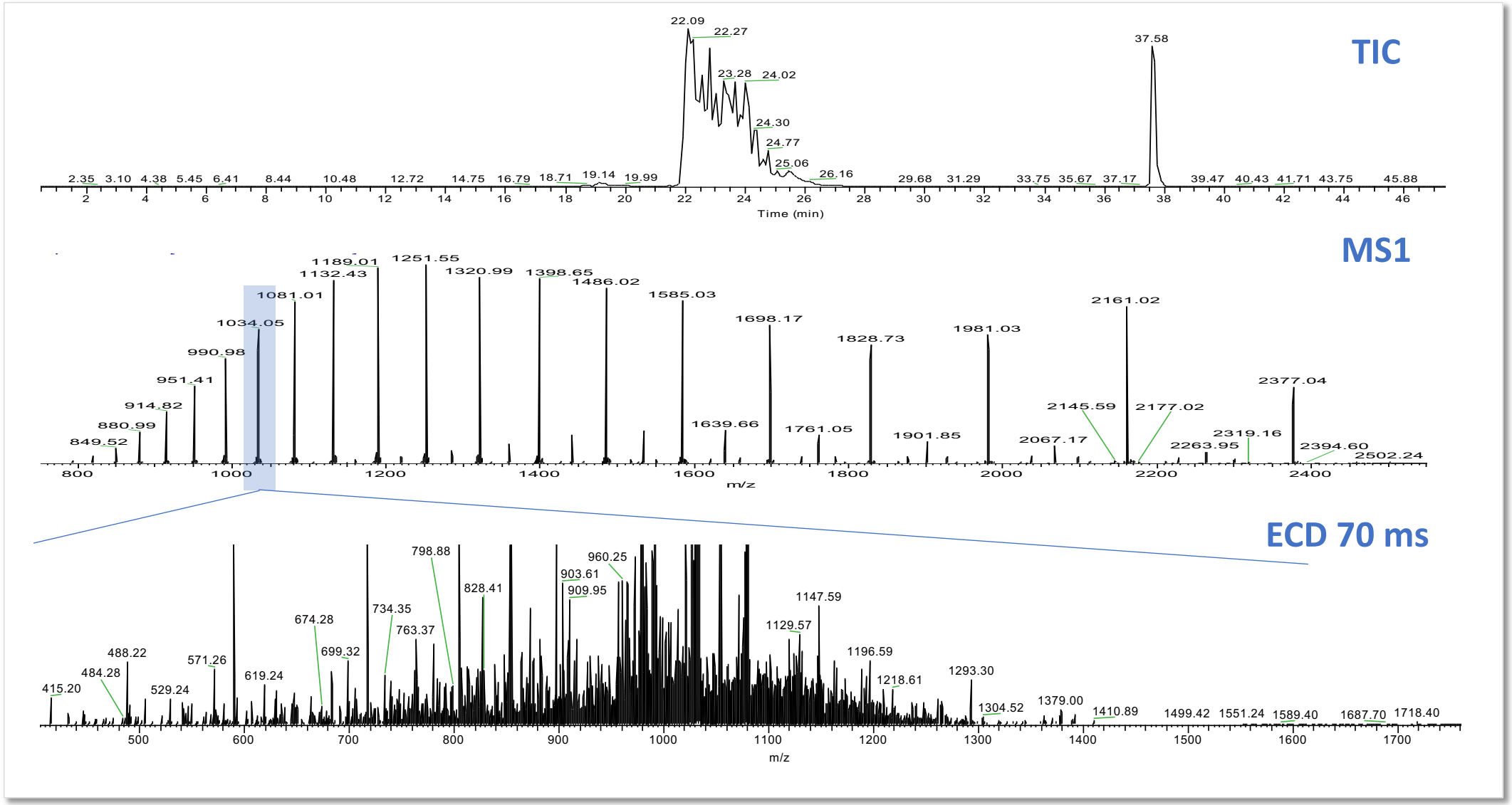
Q-Exactive HF modified with the Omnitrap platform

- Smooth installation (3 days), performance of the Q-Exactive HF (HeLa digest) kept the same



Application areas: top-down/bottom-up proteomics, native MS, glycomics…

P15: ECD 70 ms – MS^2 (nanoLC time scale)



P15: ECD (non-reduced) – MS^2

Residue cleavage: 37%

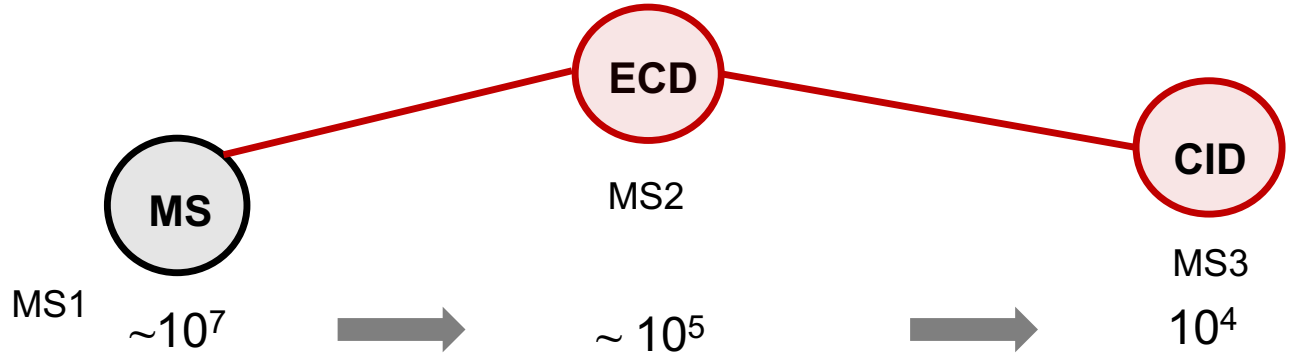
Out of the S-S bridges: 88%

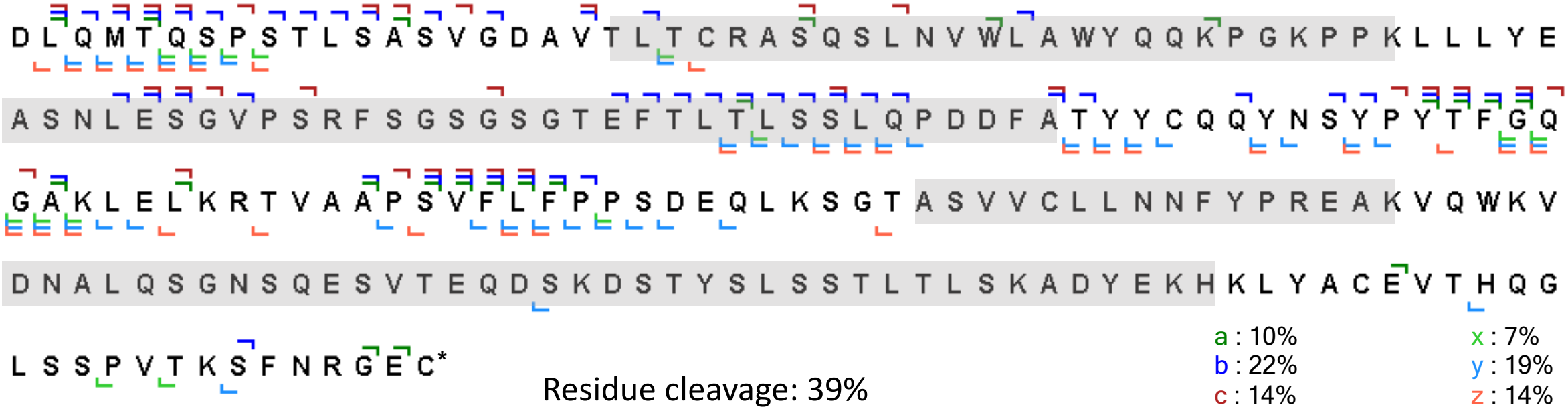
a : 18%
b : 7%
c : 23%

x : 15%
y : 16%
z : 24%

Excellent sequence coverage out of the S-S bridges allowing their easy localization
(Cronus software, no data deconvolution)

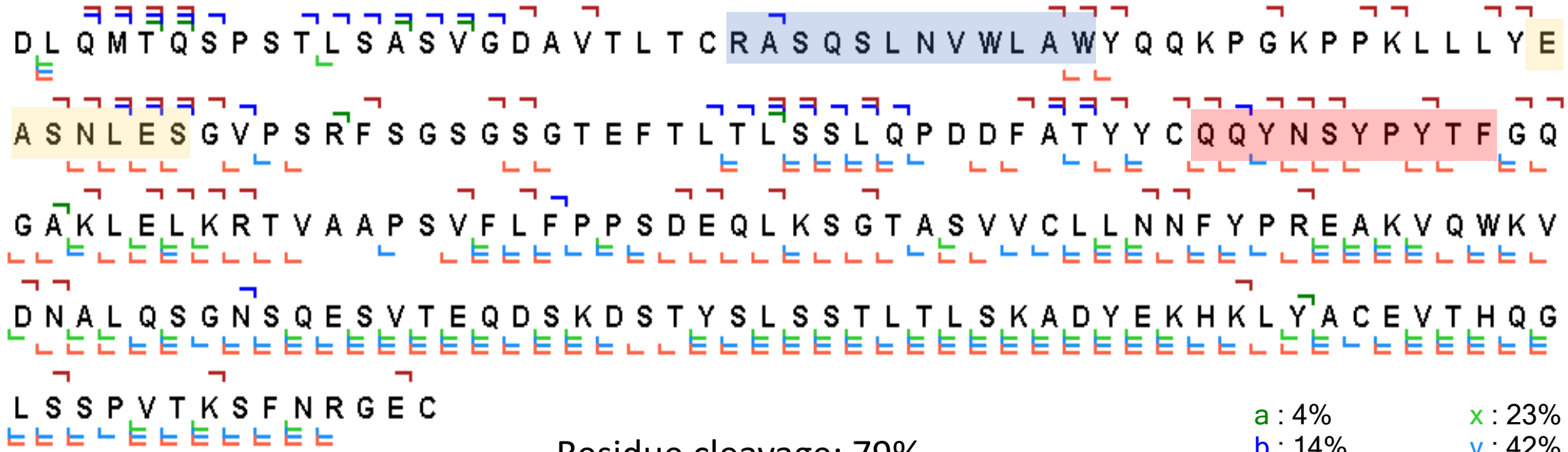
P15: ECD followed by CID (non-reduced) – MS^3





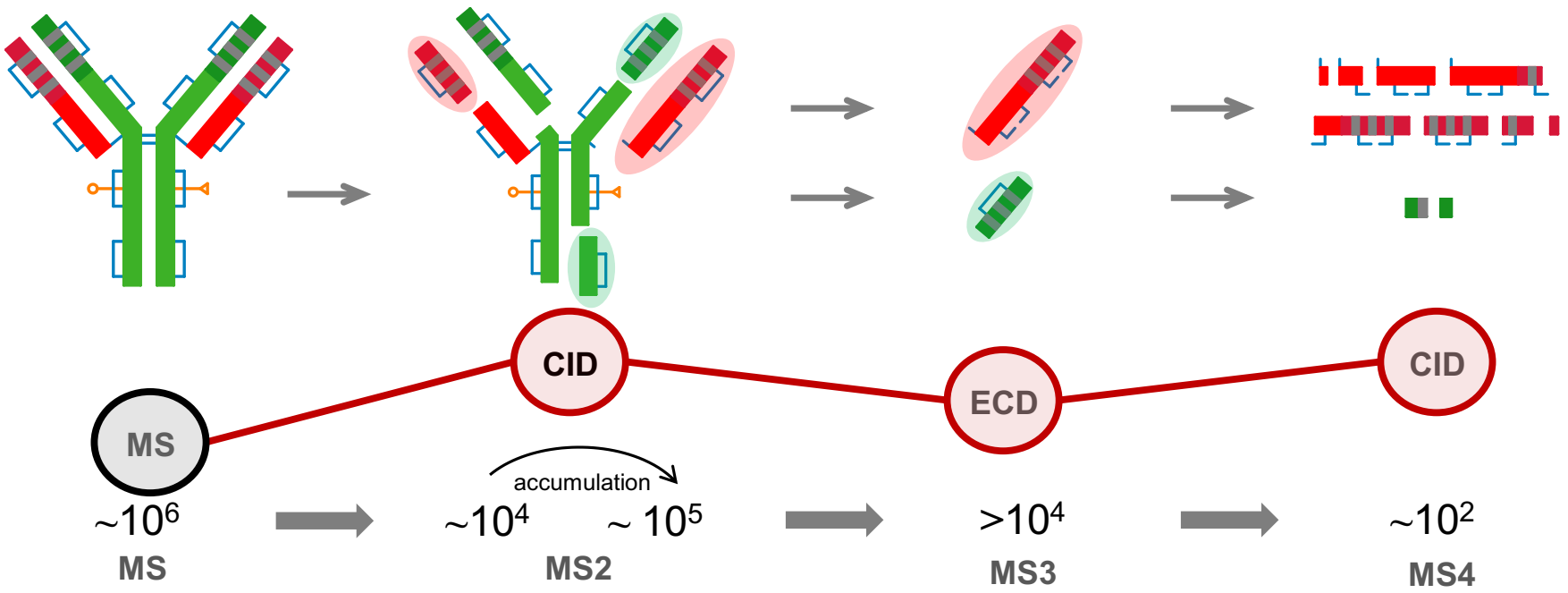
First S-S bond has been cleaved leading to new fragment ions

P15: ECD (reduced) – MS^2



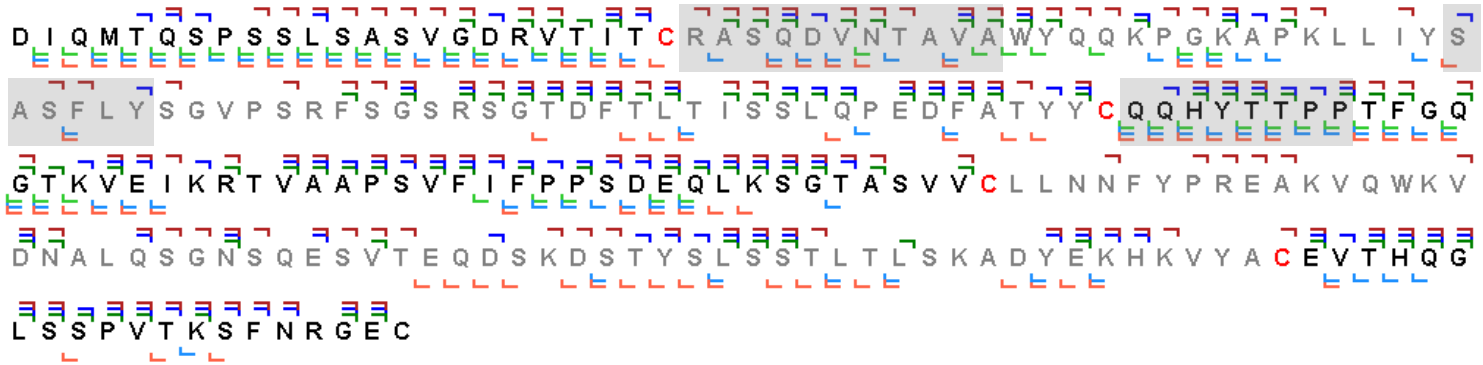
79% sequence coverage in a single experiment

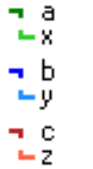
P15: Intact herceptin 150 kDa (infusion) – MS^4



Herceptin Light Chain

Sequence Coverage ~84%



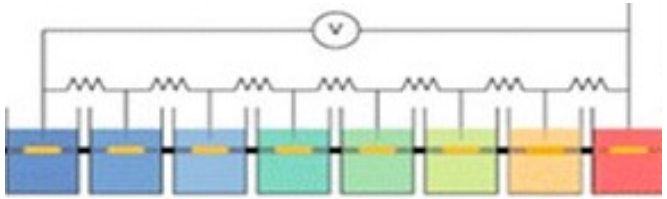


CDR regions

D. Papanastasiou *et al. J. Am. Soc. Mass Spectrom.* (2022)

Conclusion & perspectives

- TDP is essential for samples with multiple proteoforms
- Improvements required at all steps of the experiments
- The Omnitrap platform holds great promise for efficient fragmentation of intact proteins (even large ones)
  More to come: VUV for S-S bond cleavage followed by CID
- Importance of data analysis, new developments required (*de novo* sequencing from top-down data)
- PI trap for proteoform separation (isoelectric focusing)



Proteoforms as the next proteomics currency

PROTEOMICS

Proteoforms as the next proteomics currency

Identifying precise molecular forms of proteins can improve our understanding of function

By **Lloyd M. Smith**[1] *and* **Neil L. Kelleher**[2]

Proteoforms—the different forms of proteins produced from the genome with a variety of sequence variations, splice isoforms, and myriad posttranslational modifications (*1*)—are critical elements in all biological systems (see the figure, left). Yang *et al.* (*2*) recently showed that the functions of proteins produced from splice variants from a given gene—different proteoforms—can be as different as those for proteins encoded by entirely different genes. Li *et al.* (*3*) showed that splice variants play a central role in modulating complex traits. However, the standard paradigm of proteomic analysis, the "bottom-up" strategy provides invaluable information on protein expression in complex systems. However, as many different gene products, isoforms, and proteoforms can contain the same peptide, direct information about the proteoforms present is lost (see the figure, bottom). This issue is the proteomic analog of the problem of "phasing" in genomics (*5*)—determining whether multiple alleles are present on the same segment of DNA. The step of digestion into peptides is essential to the success and robustness of the bottom-up strategy, as well-behaved peptides are more amenable to liquid chromatographic separation and MS analysis than are intact proteins. However, only inferences can be made as to the actual proteoform or proteoforms from which the

L. Smith & N. Kelleher *Science* 359:1106-1107 (2018)

European Proteomics Infrastructure Consortium - Providing Access

MSBio Lab
Institut Pasteur
M. Duchateau
M. Dupré
T. Fu
C. Garcia
M. Rey
D. Tabb

Karolinska Inst.
R. Zubarev
S. Lundström

Thermo Fisher
A. Makarov
K. Fort

Technical University of Denmark
R. Sternke-Hoffmann
A. Buell

Fasmatech
D. Papanastasiou
T. Smyrnakis
M. Kosmopoulou
G. Alevizos

MSVision
A. Langlands
J. Commandeur

https://epic-xs.eu