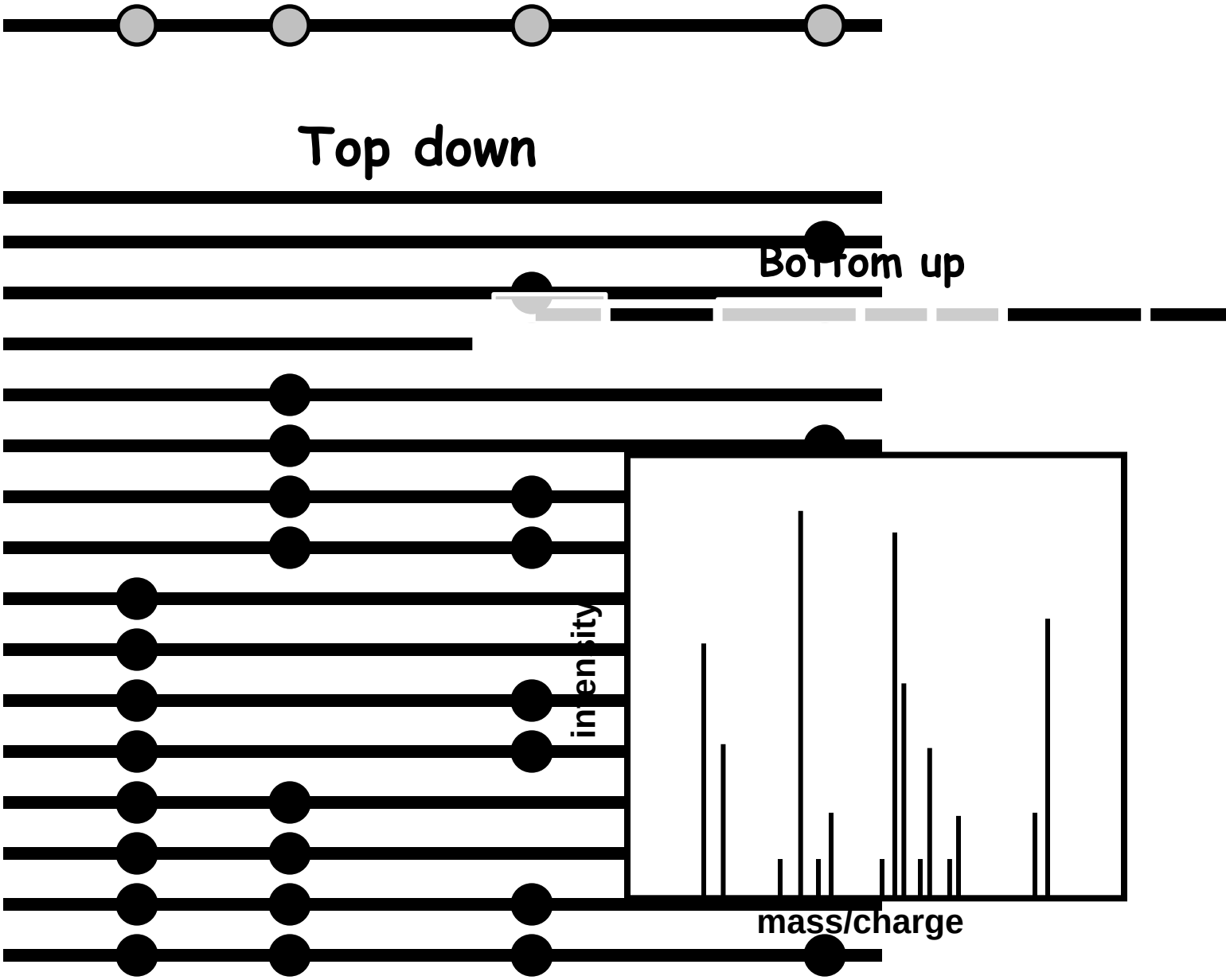
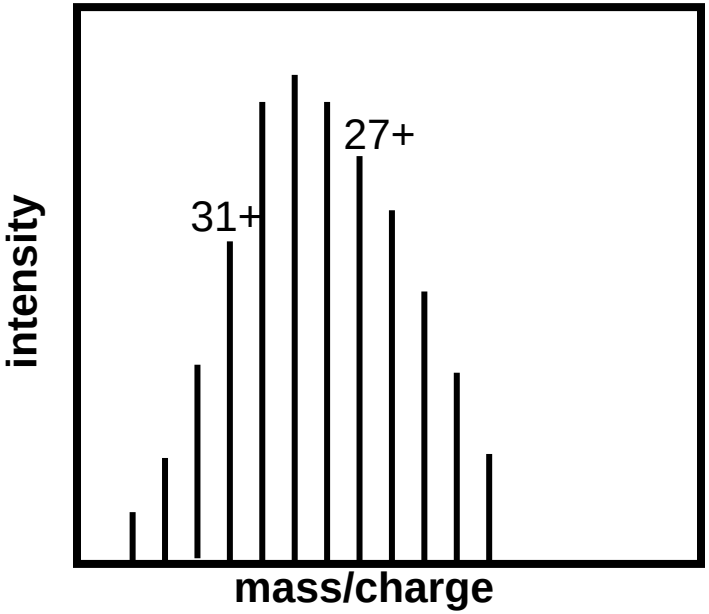
Proteomics Informatics – Protein characterization: post-translational modifications and protein-protein interactions (Week 10)

Top down / bottom up

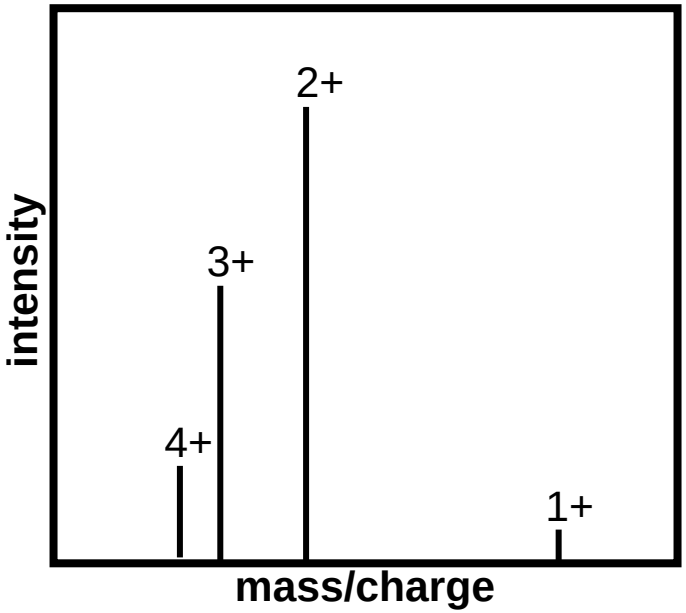


Charge distribution

Top down

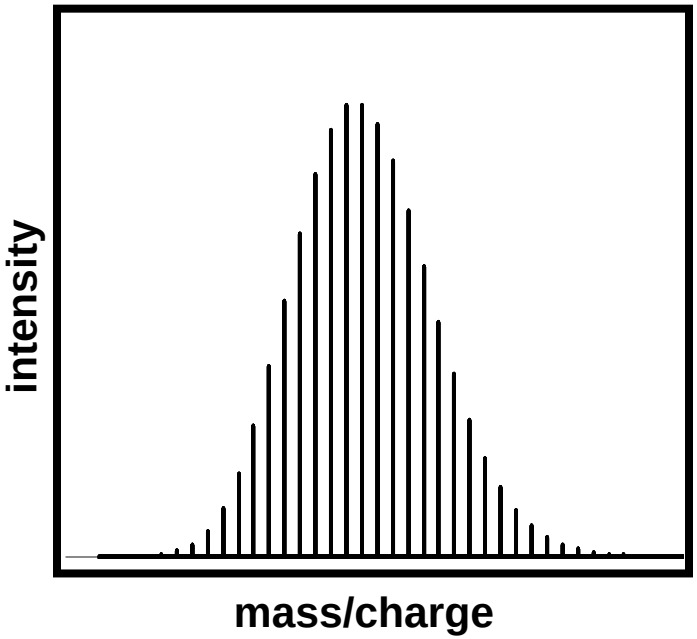


Bottom up

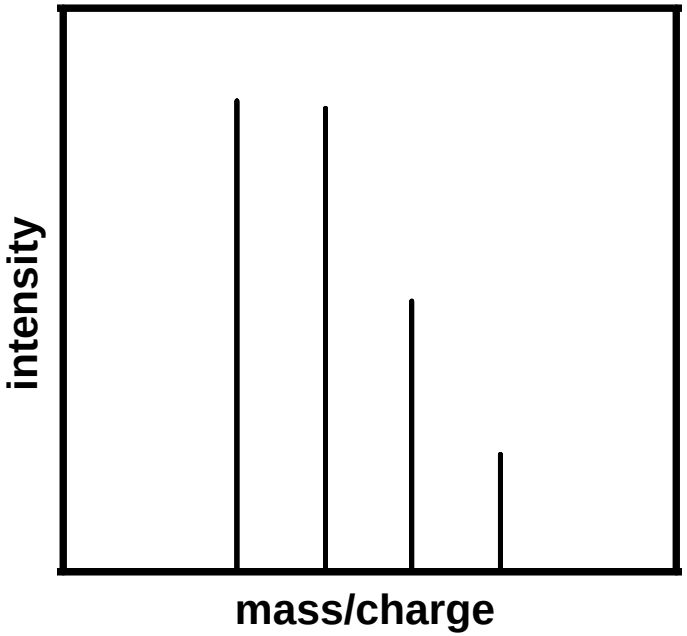


Isotope distribution

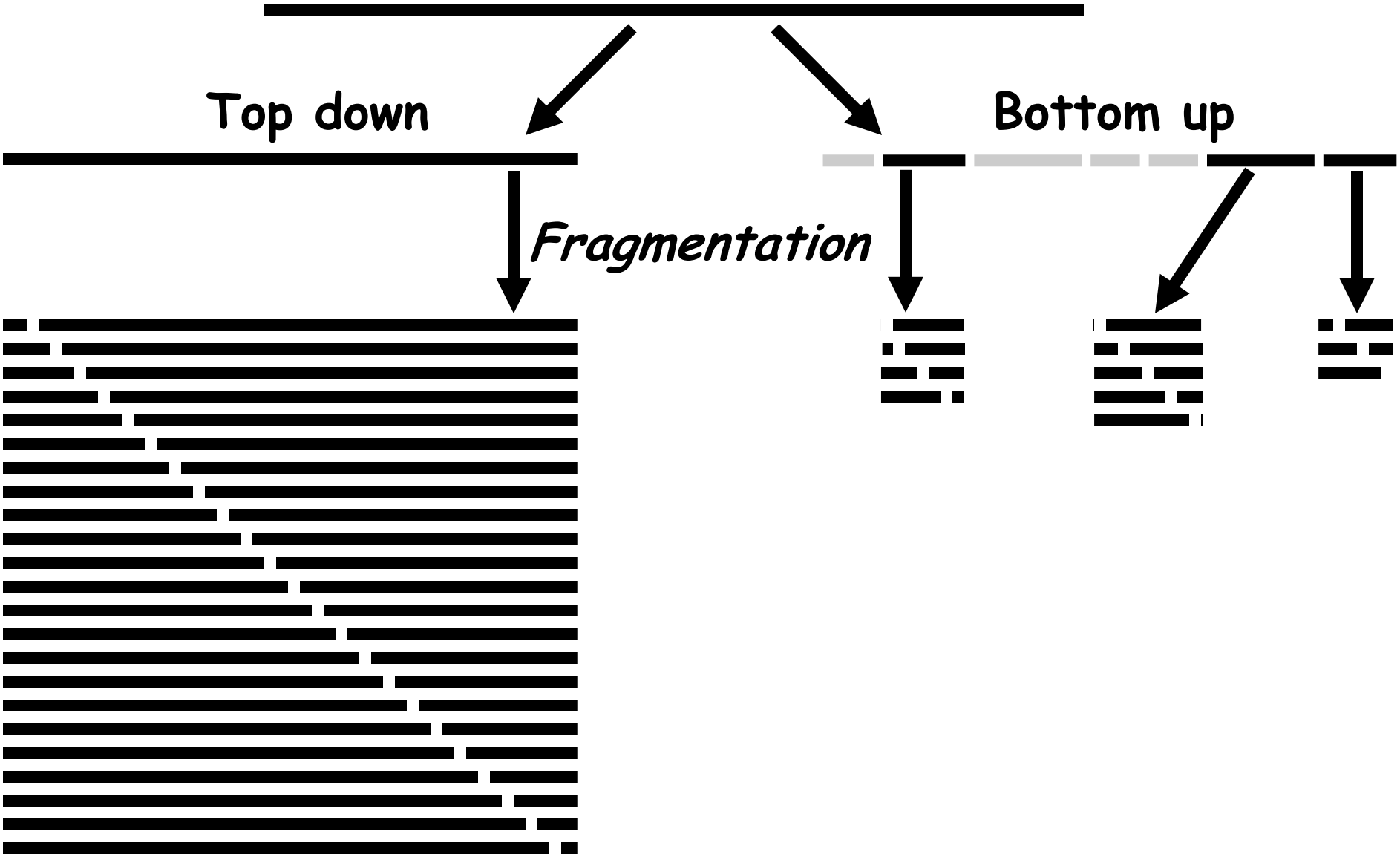
Top down



Bottom up

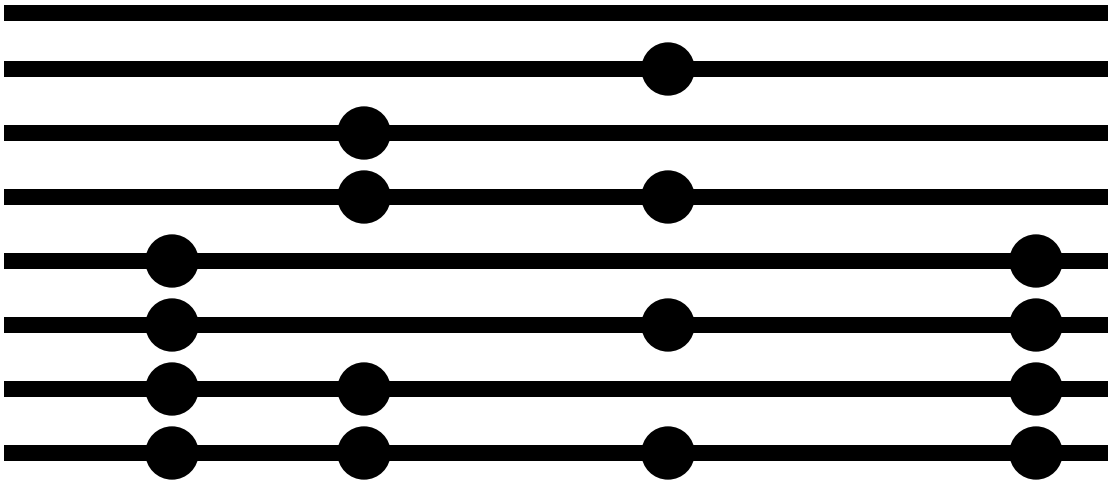


Fragmentation



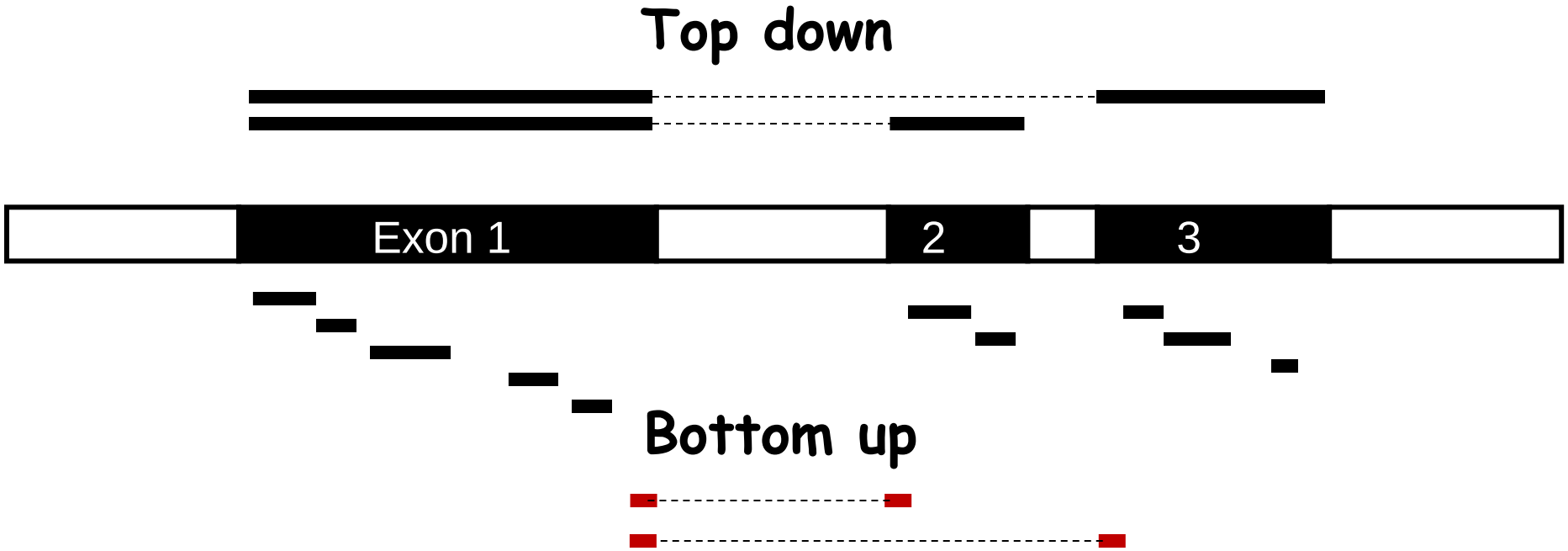
Correlations between modifications

Top down

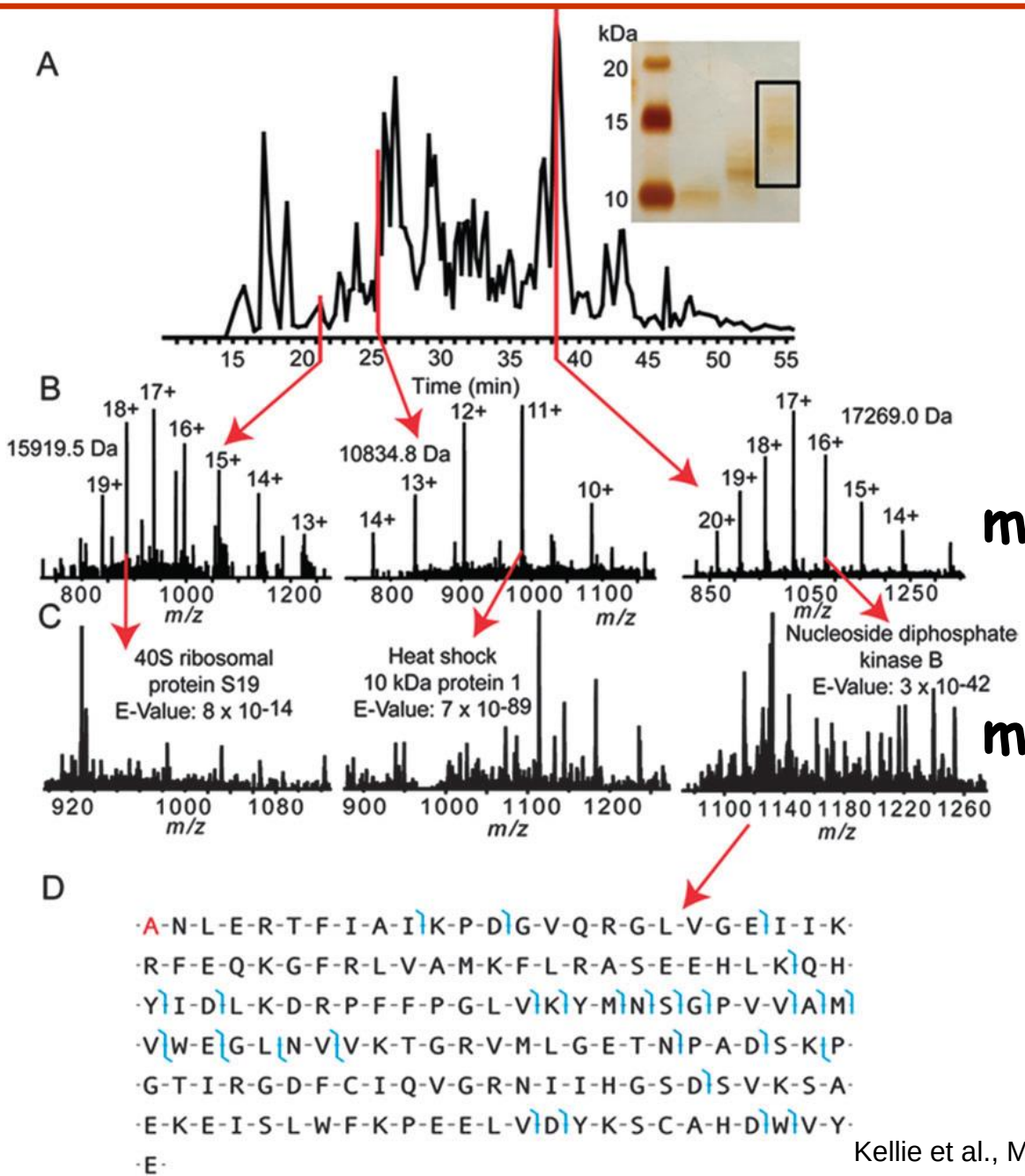


Bottom up

Alternative Splicing



Top down

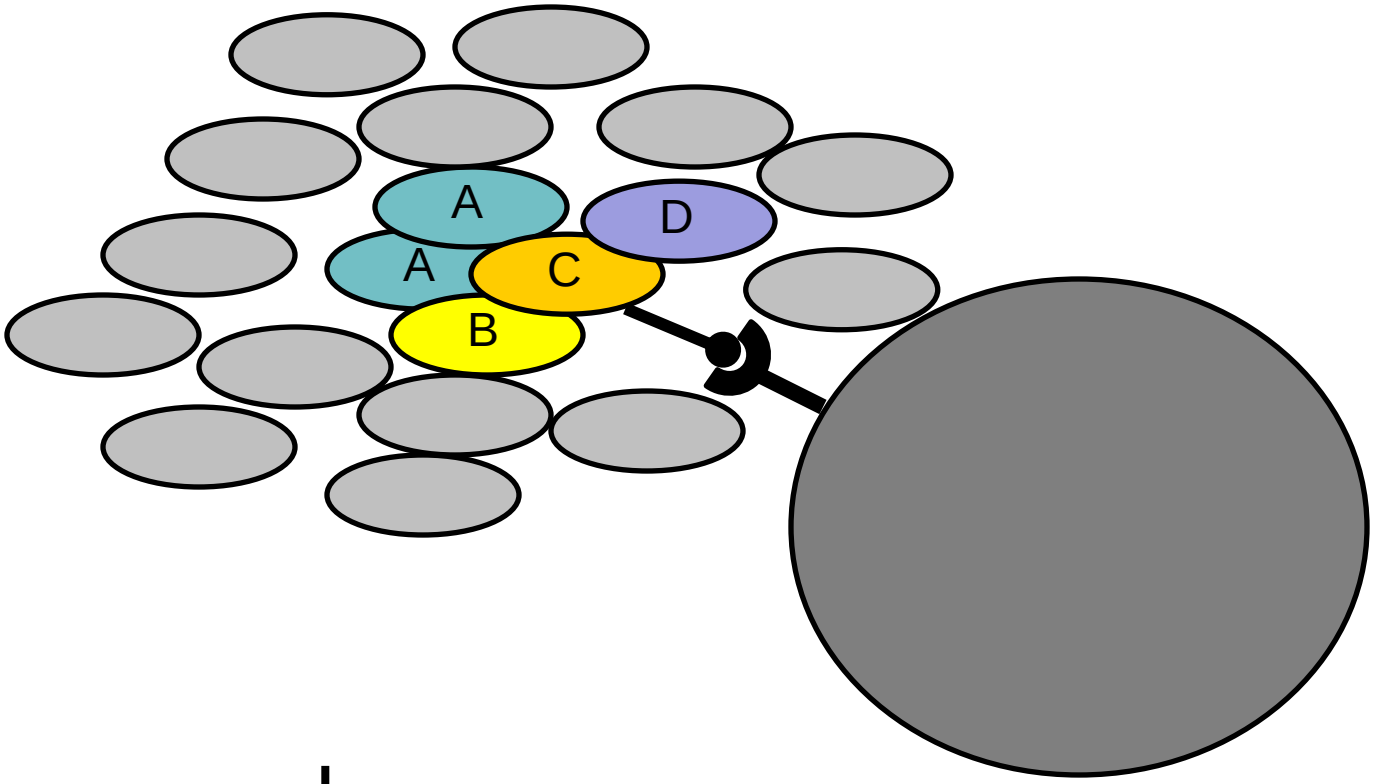


Protein mass spectra

Fragment mass spectra

Kellie et al., Molecular BioSystems 2010

Protein Complexes



↓

Digestion

↓

Mass spectrometry

Protein Complexes – specific/non-specific binding

E Stats Table

	Bait 1	Bait 2	Bait 3	Bait 4	Bait k	
Interactor 1	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	$X_{4,1}$	$X_{k,1}$	$\overline{X}_1$
Interactor 2	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	$X_{4,2}$	$X_{k,2}$	$\overline{X}_2$
Interactor 3	$X_{1,3}$	$X_{2,3}$	$X_{3,3}$	$X_{4,3}$	$X_{k,3}$	⋮
Interactor 4	$X_{1,4}$	$X_{2,4}$	$X_{3,4}$	$X_{4,4}$	$X_{k,4}$	
Interactor m	$X_{1,m}$	$X_{2,m}$	$X_{3,m}$	$X_{4,m}$	$X_{k,m}$	$\overline{X}_m$

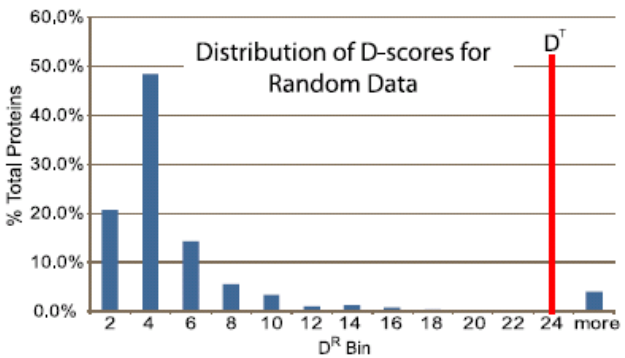
$X_{i,j}$ = total spectral counts for interactor j from bait i

$$\overline{X}_j = \frac{\sum_{i=1,\, j=n}^{i=k} X_{i,j}}{k} \quad ; n = 1,2,\ldots m \qquad \text{(Eq. 1)}$$

$$Z_{i,j} = \frac{X_{i,j} - \overline{X}_j}{\sigma_j} \qquad \text{(Eq. 2)}$$

$$f_{i,j} = \begin{cases} 1 \; ; X_{i,j} > 0 \\ X_{i,j} \end{cases} \qquad p = \begin{cases} \text{number of replicates} \\ \text{runs in which} \\ \text{the interactor is present} \end{cases}$$

$$D^R_{i,j} = \sqrt{\left(\frac{k}{\sum_{i=1}^{i=k} f_{i,j}}\right)^p X_{i,j}} \qquad \text{(Eq. 3)}$$

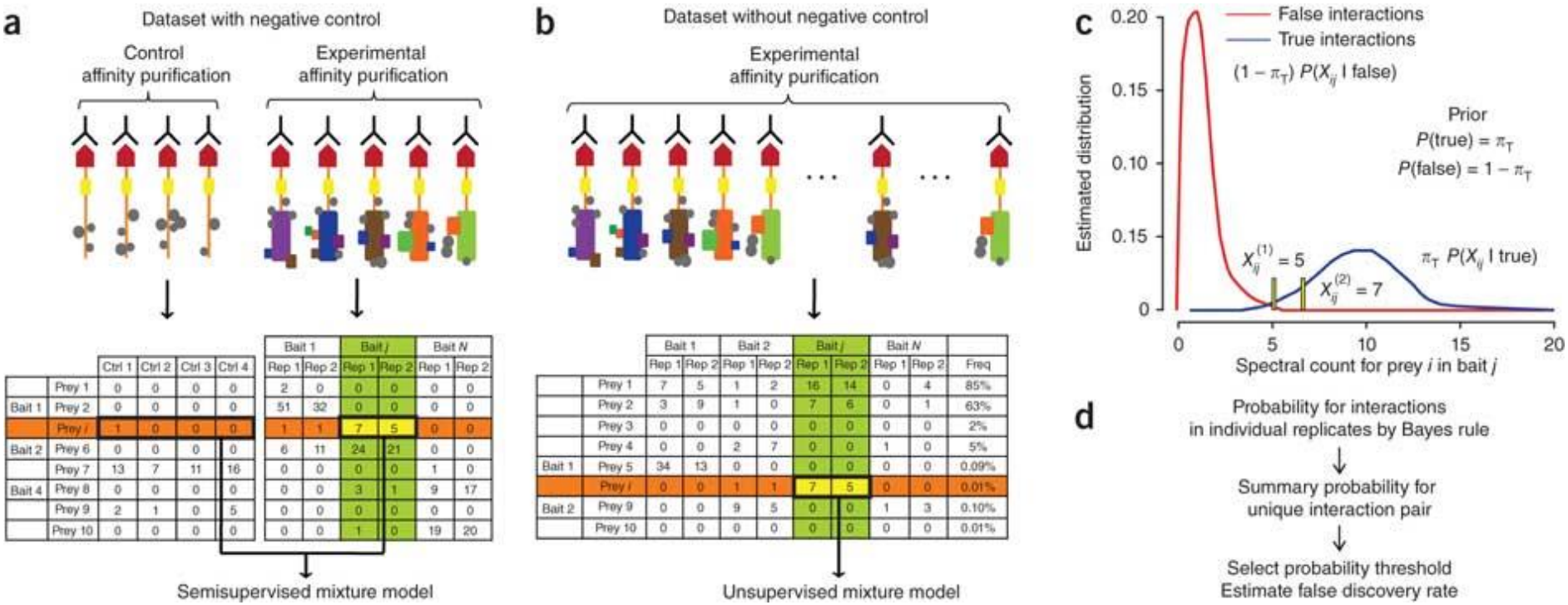


D^T = D-score threshold below which 95% of simulated data falls

$$D^N_{i,j} = D^R_{i,j} \big/ D^T \qquad \text{(Eq. 4)}$$

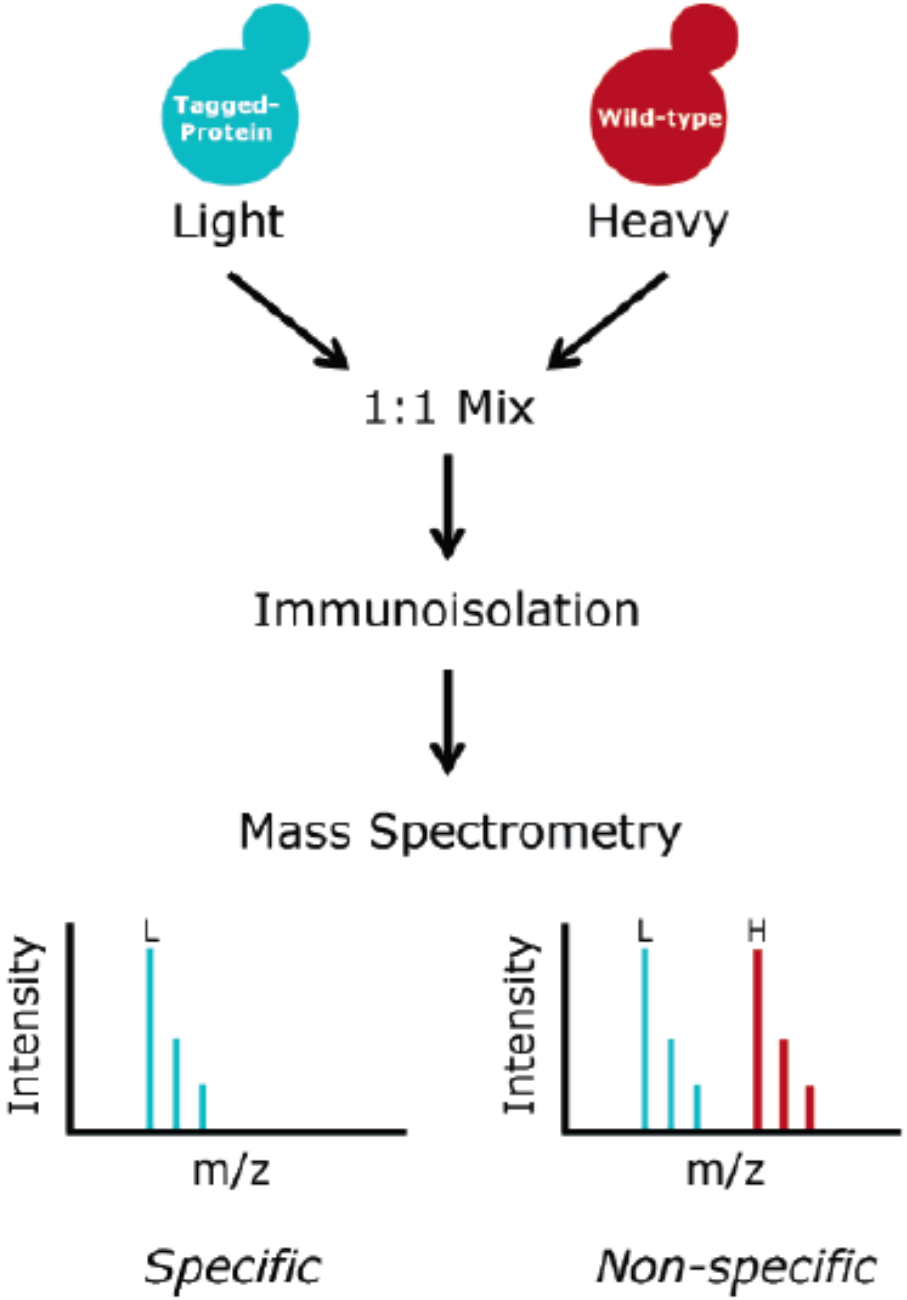
Sowa et al., Cell 2009

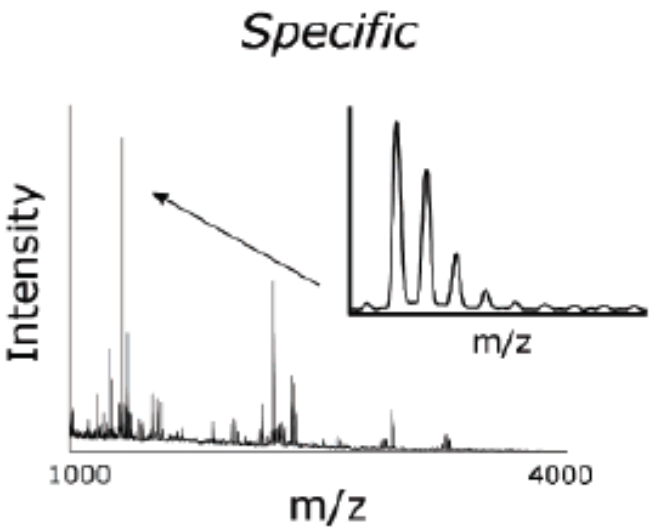
Protein Complexes – specific/non-specific binding

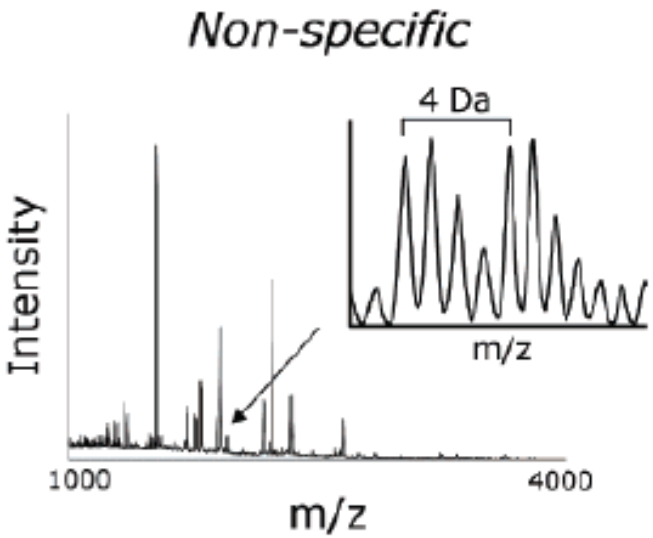


Choi et al., Nature Methods 2010

Protein Complexes – specific/non-specific binding

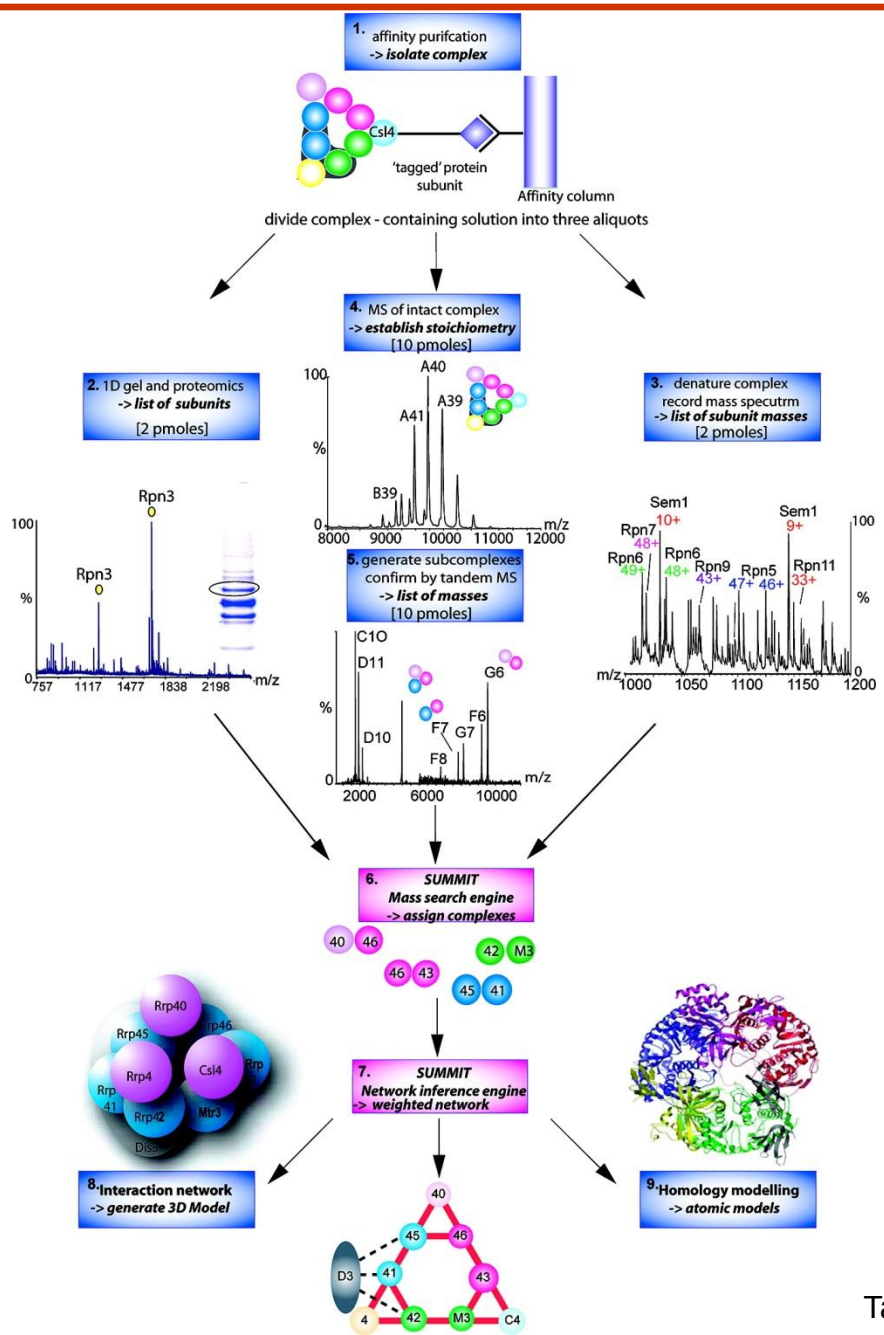






Tackett et al. JPR 2005

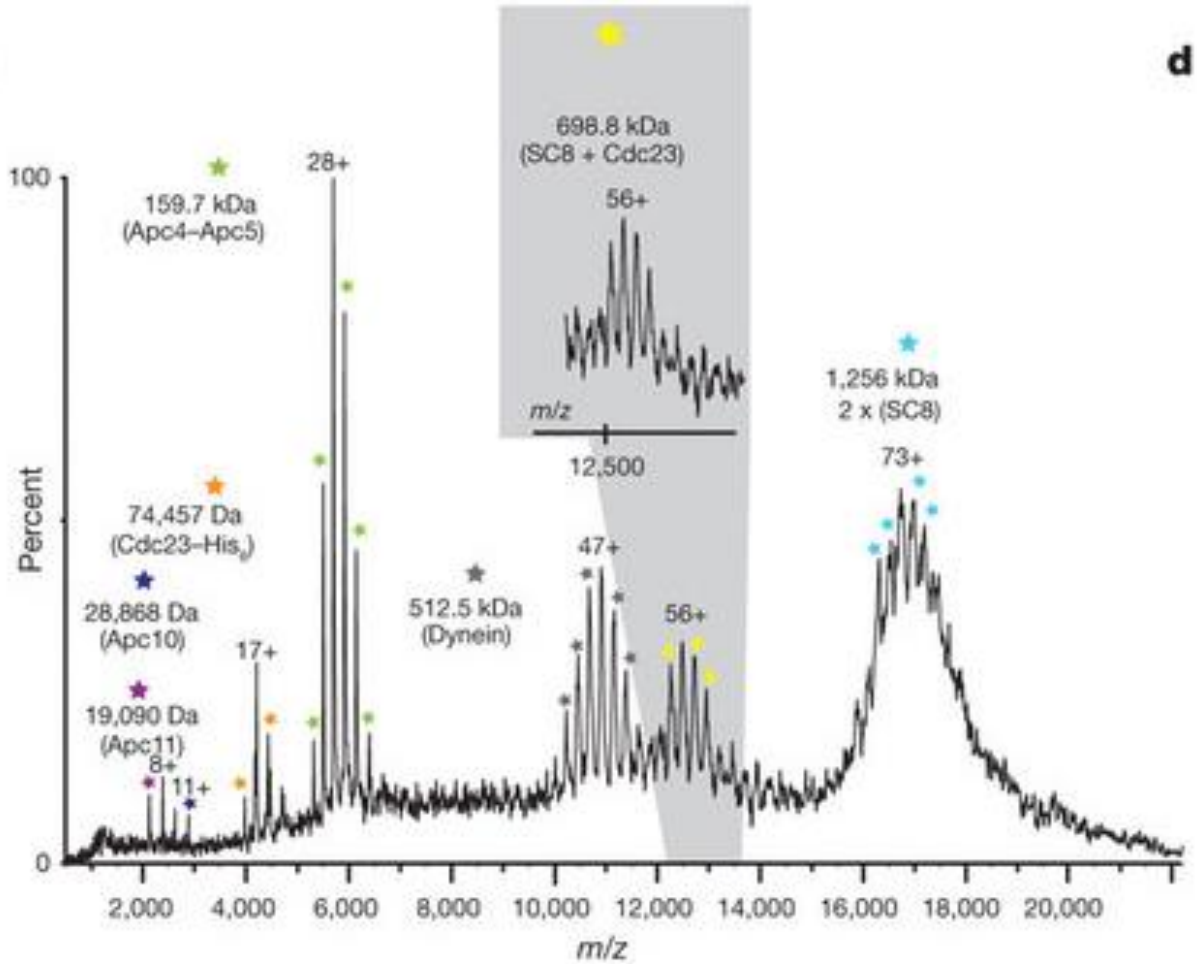
Analysis of Non-Covalent Protein Complexes



Taverner et al., Acc Chem Res 2008

Non-Covalent Protein Complexes

c

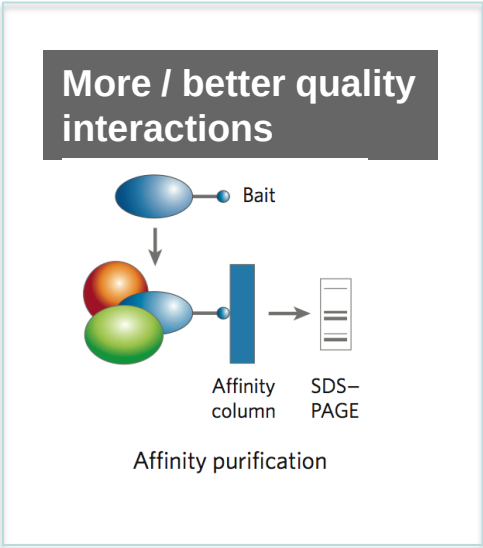


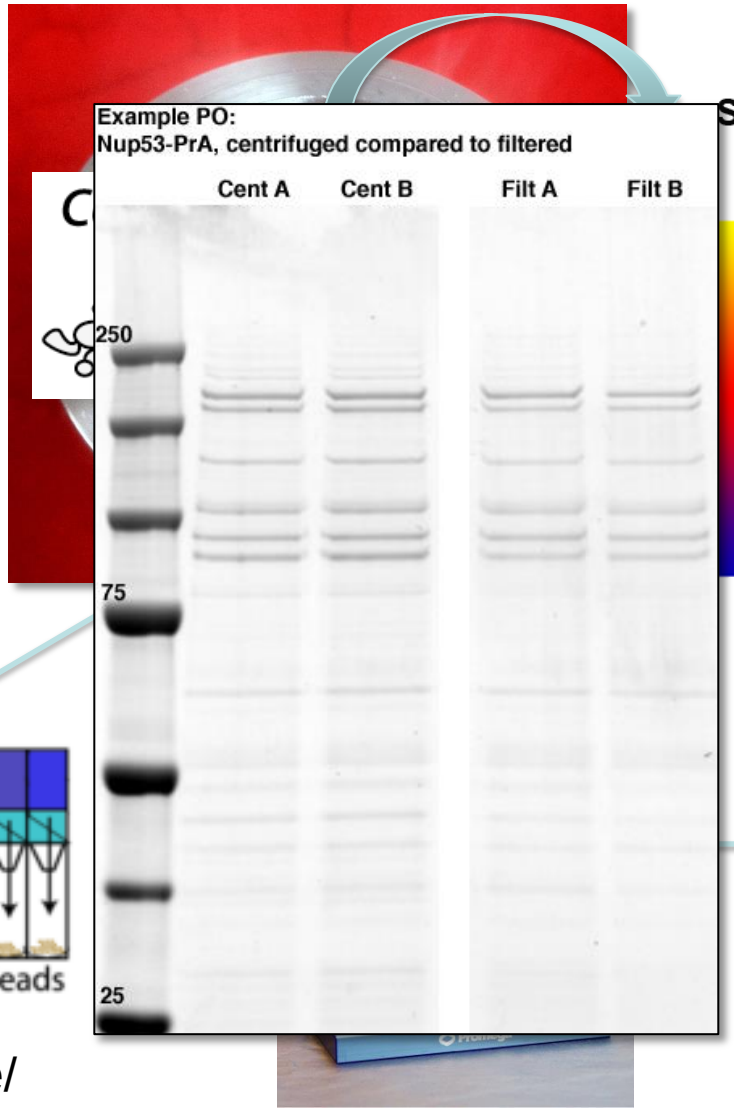
d

S. cerevisiae APC/C subunit	Stoichiometry
Apc1	1
Apc2	1
Apc4	1
Apc5	1
Apc10	1
Apc11	1
Cdc23	2
Cdc16	2
Cdc27	2
Apc9	ND
Mnd2	1
Apc13	1
Cdc26	2
Total MW (kDa)	1,127–1,158

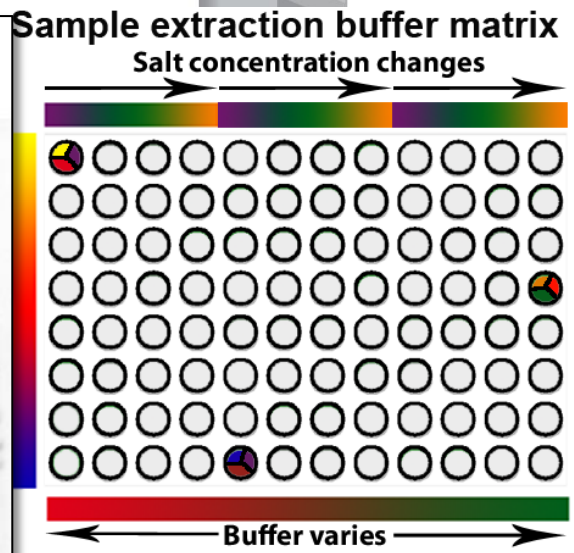
Schreiber et al., Nature 2011

Affinity Capture Optimization Screen

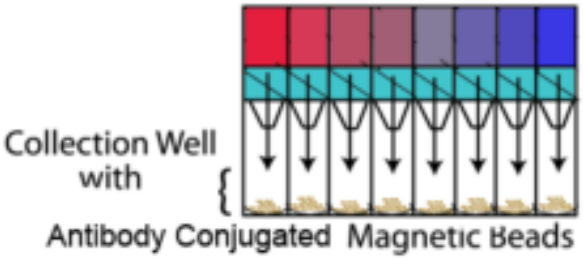




Cell extraction



Filtration



Lysate clearance/
Batch Binding

Binding/Washing/Eluting

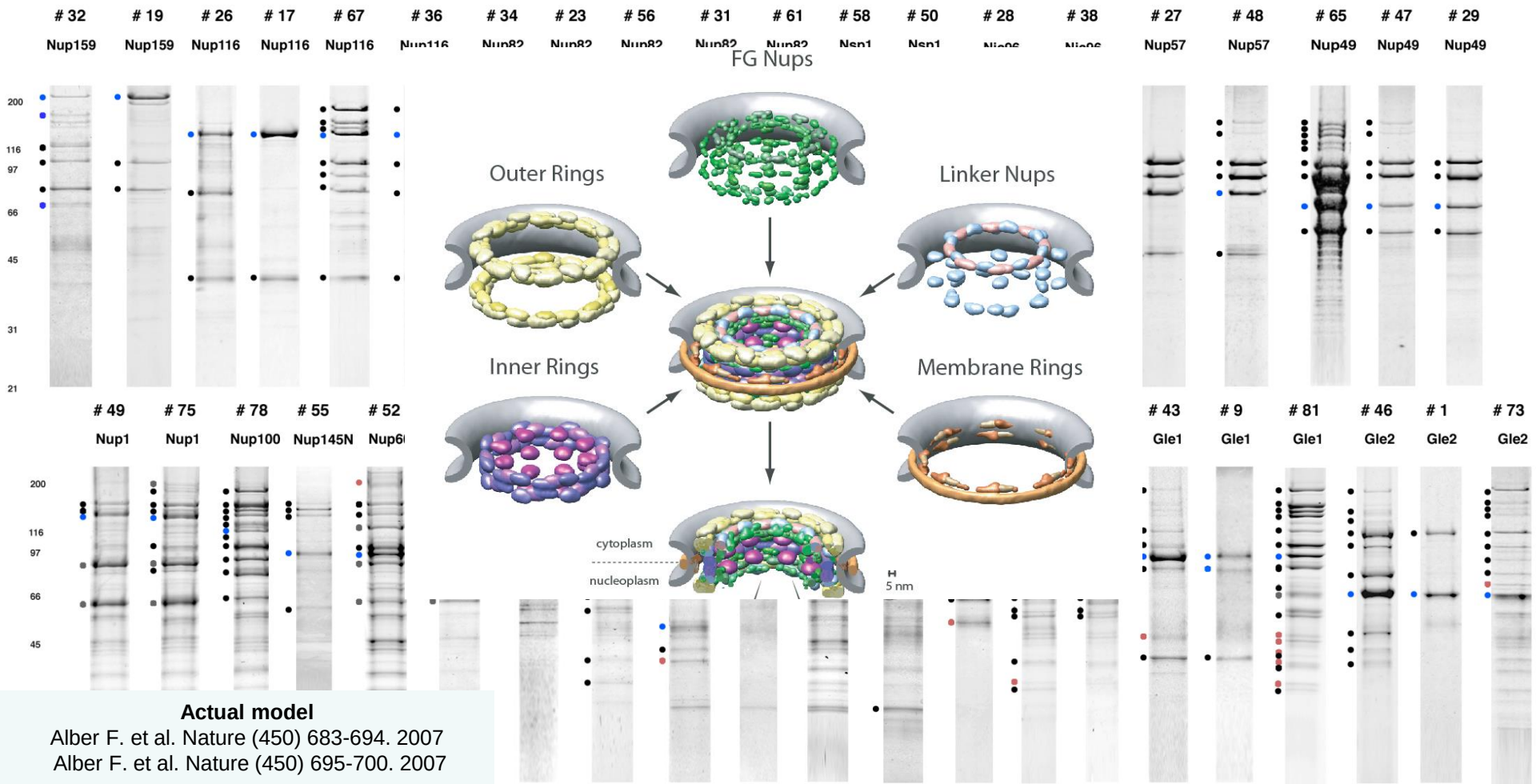
SDS-PAGE

LaCava, Hakhverdyan, Domanski, Rout

Molecular Architecture of the NPC

Ultracentrifugation		Quantitative immunoblotting	Affinity purification	Overlay assay	Electron microscopy	Immuno-electron microscopy	Bioinformatics and machine fractionation
30 S-values	1 S-value	30 relative abundances	75 composites	7 contacts	Electron microscopy maps	10,615 single particles	30 protein sequences

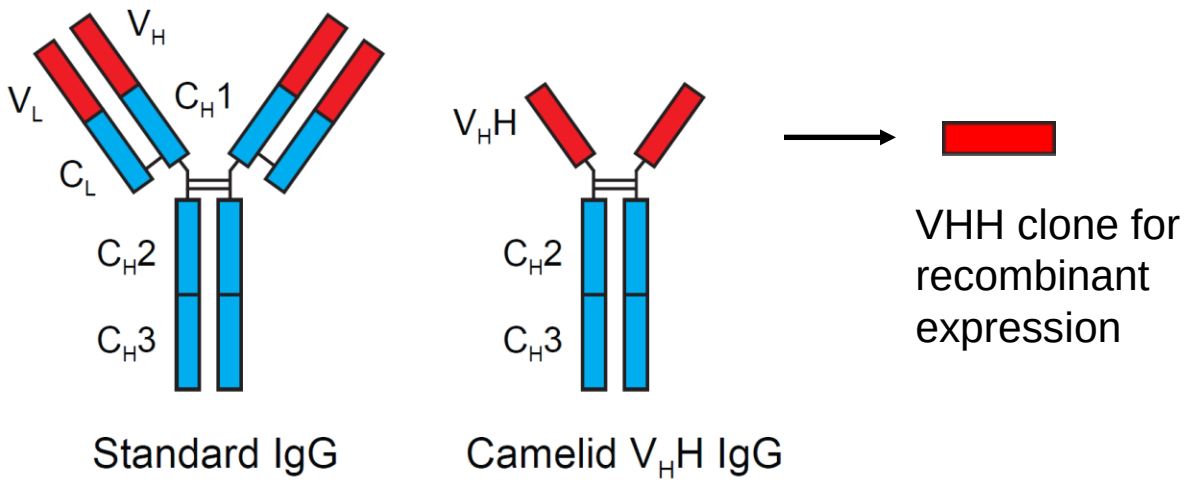
Over 20 different extraction and washing conditions ~ 10 years or art.
(41 pullouts are shown)



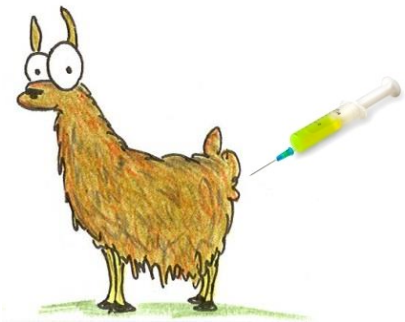
Actual model
Alber F. et al. Nature (450) 683-694. 2007
Alber F. et al. Nature (450) 695-700. 2007

Cloning nanobodies for GFP pullouts

- Atypical heavy chain-only IgG antibody produced in camelid family – retain high affinity for antigen without light chain
- Aimed to clone individual single-domain VHH antibodies against GFP – only ~15 kDa, can be recombinantly expressed, used as bait for pullouts, etc.
- To identify full repertoire, will identify GFP binders through combination of high-throughput DNA sequencing and mass spectrometry



Cloning llamabodies for GFP pullouts



Llama GFP immunization

Bone marrow aspiration

Serum bleed

Lymphocyte total RNA

Crude serum

RT / Nested PCR

IgG fractionation & GFP affinity purification

VHH amplicon

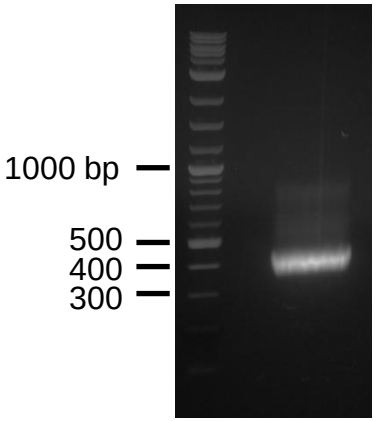
GFP-specific VHH fraction

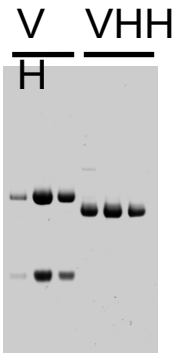
454 DNA sequencing

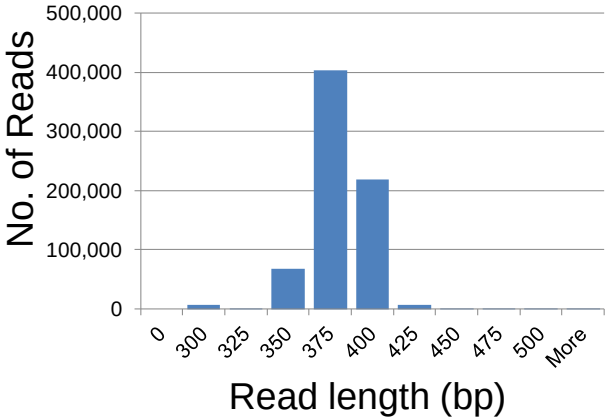
VHH DNA sequence library

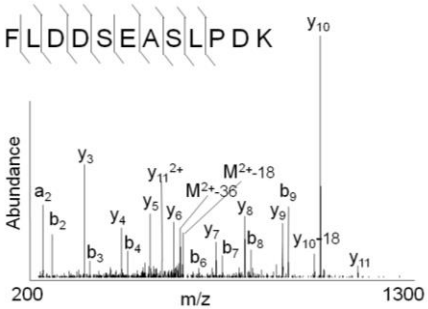
LC-MS/MS

GFP-specific VHH clones









Fridy, Li, Keegan, Chait, Rout

Identifying full-length sequences from peptides

Underlined regions are covered by MS

CDR1 CDR2 CDR3

```
CDR3: 100.0% (14/14); combined CDR: 100.0% (33/33); DNA count: 10
MAQVQLVESGGGLVQAGGSLRLSCVASGRTFSGYAMGWFRQTPGREREAVAAITWSAHSTYYSDSVKDRFTISIDNTRNTGYLQMNSLKPEDTAVYYCTVRHGTWFTTSRYWTDWGQGTQVTVS

CDR3: 100.0% (14/14); combined CDR: 72.7% (24/33); DNA count: 1
MAQVQLVESGGALVQAGASLSVSCAASGGTISKYNMAWFRRAPGRREAVAAITWSAHSTYYSDSVKDRFTISIDNTRNTGYLQMNSLKPEDTAVYYCTVRHGTWFTTSRYWTDWGQGTQVTVS

CDR3: 100.0% (14/14); combined CDR: 72.7% (24/33; DNA count: 1
MADVQLVESGGGLVQSGGSRTLSCAASGRVLATYHLGWFRQSPGRREAVAAITWSAHSTYYSDSVKGRFTISIDNARNTGYLQMNSLKPEDTAVYYCTVRHGTWFTVSRYWTDWGQGTQVTVS

CDR3: 100.0% (14/14); combined CDR: 42.4% (14/33); DNA count: 1
MAQVQLEESGGGLVQAGDSLTLSCSASGRTFTNYAMAWSRQAPGKERELLAAIDAAGGATYYSDSVKGRFTISIDNTRNTGYLQMNSLKPEDTAVYYCTVRHGTWFTTSRYWTDWGQGTQVTVS

CDR3: 100.0% (14/14); combined CDR: 42.4% (14/33); DNA count: 1
MAQVQLVESGGGRVQAGGSLTLSCVGSEGIFWNHVMGWFRQSPGKDREFVARISKIGGTTNYADSVKGRFTISIDNTRNTGYLQMNSLKPEDTAVYYCTVRHGTWFTTSRYWTDWGQGTQVTVS
```

Rank sequences according to:

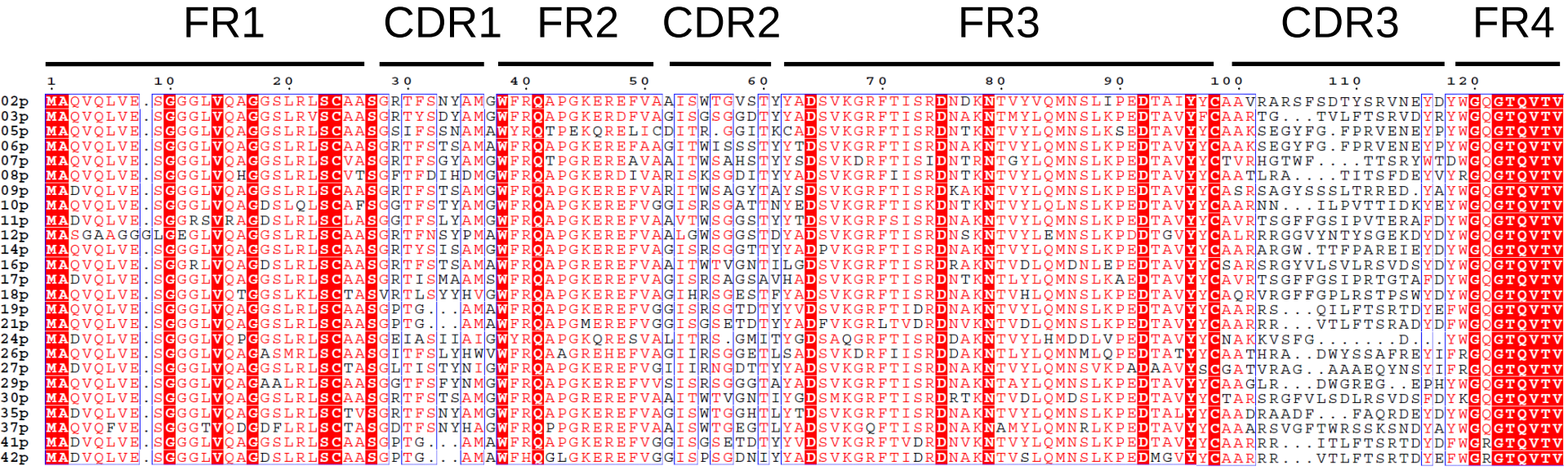
CDR3 coverage; Overall coverage;

Combined CDR coverage; DNA counts;

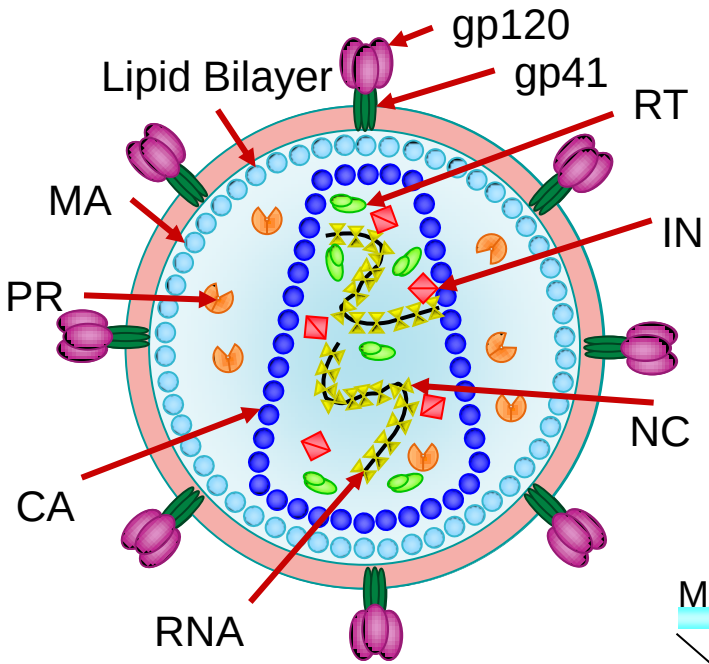
Sequence diversity of 26 verified anti-GFP nanobodies

- Of ~200 positive sequence hits, 44 high confidence clones were synthesized and tested for expression and GFP binding: 26 were confirmed GFP binders.

- Sequences have characteristic conserved VHH residues, but significant diversity in CDR regions.

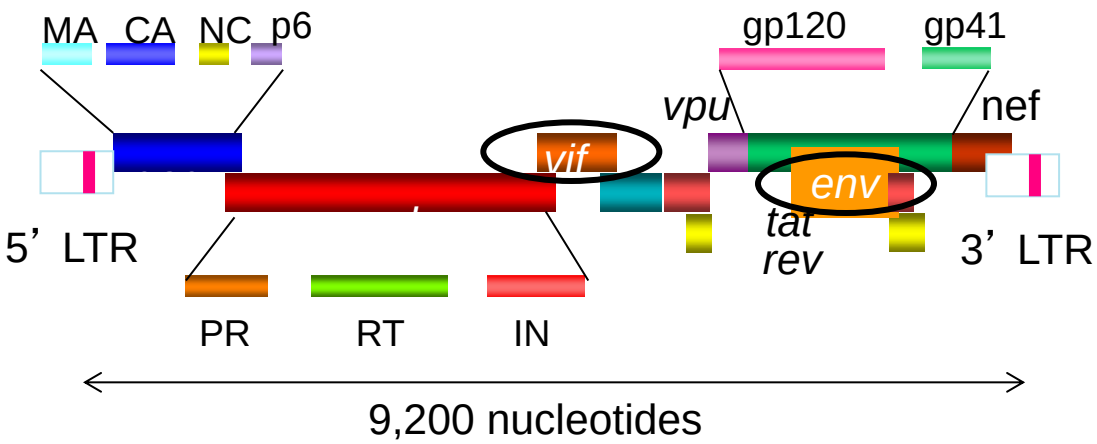


HIV-1

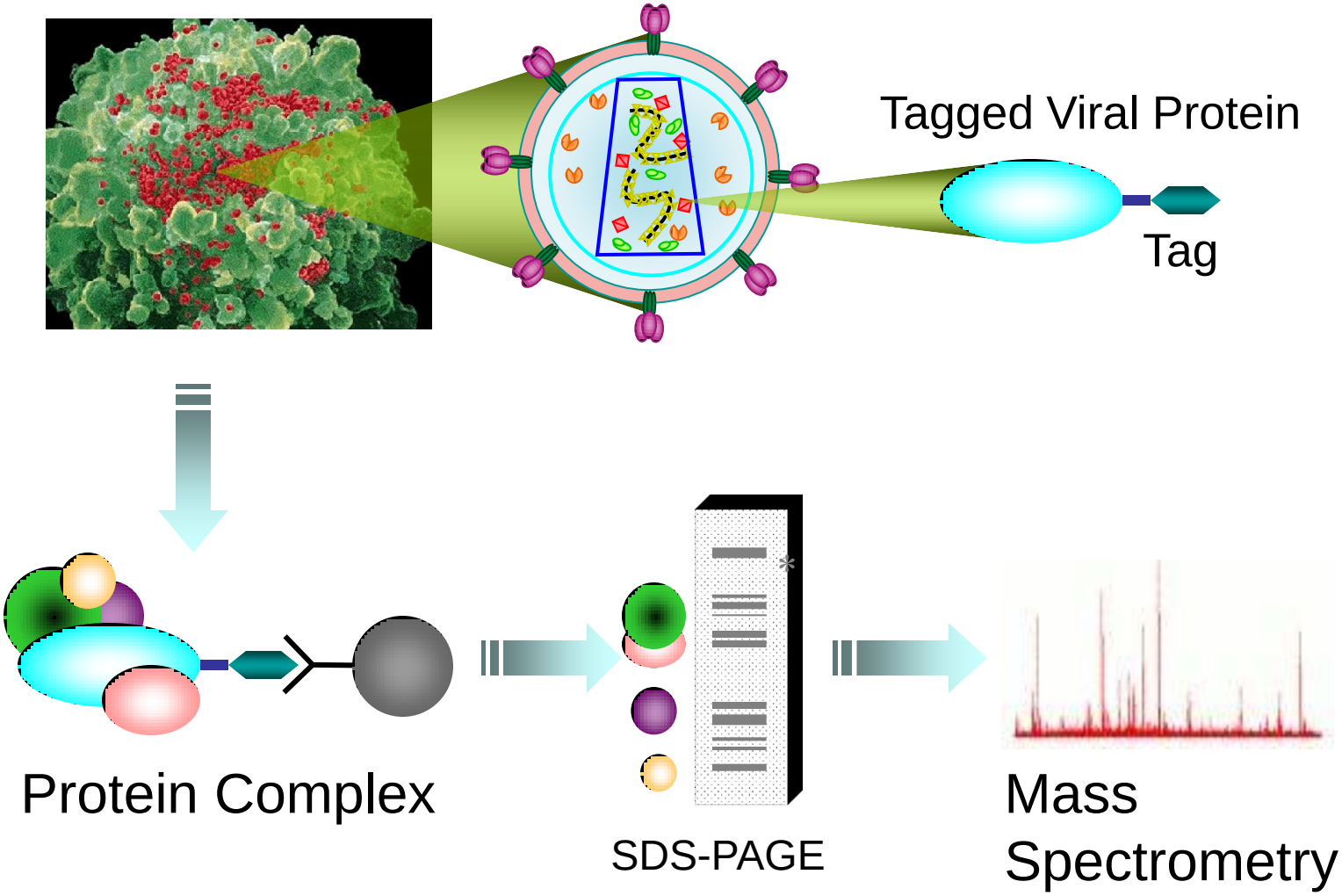


Particle

Genome

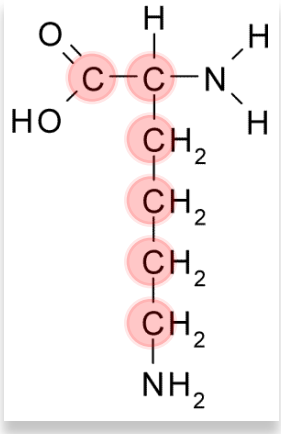


Genetic-Proteomic Approach

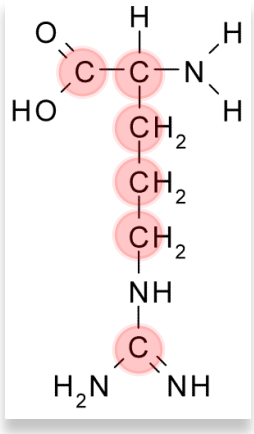


I-Dirt for Specific Interaction

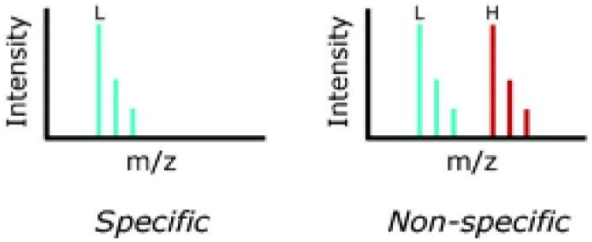
I-DIRT = Isotopic Differentiation of Interactions as Random or Targeted



Lys
(+6 daltons)

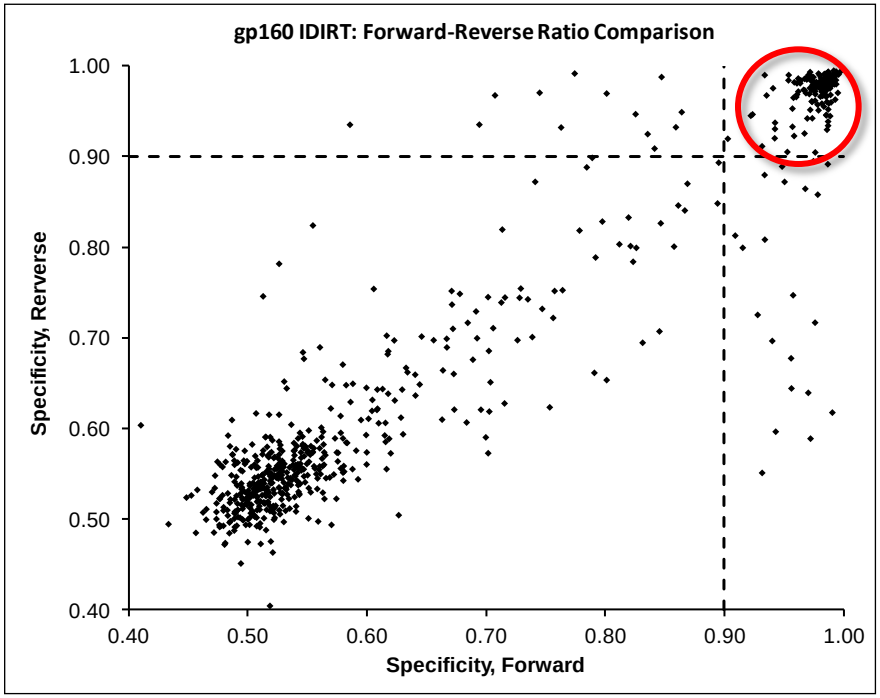


Arg
(+6 daltons)

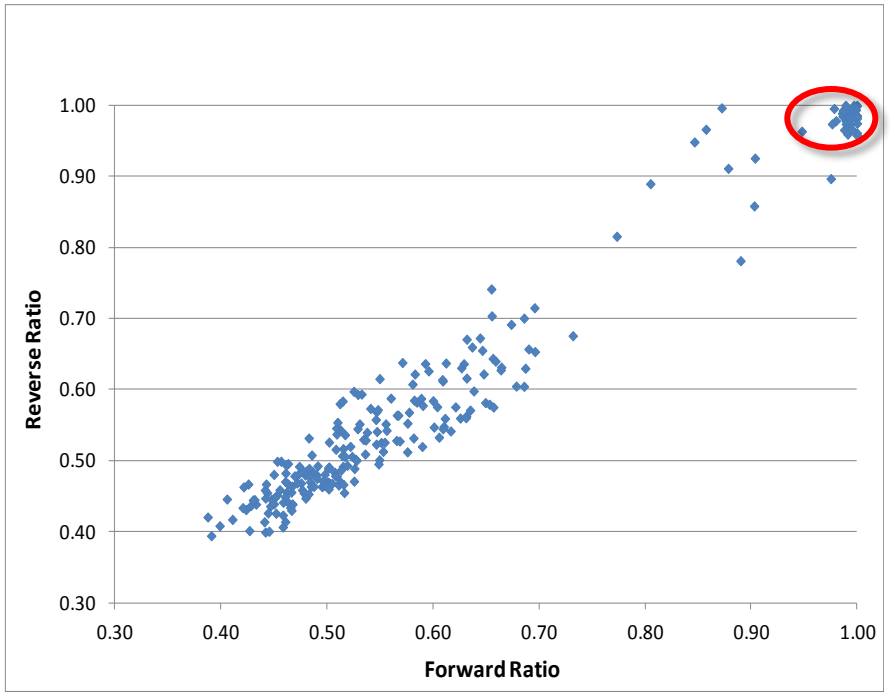


IDIRT and Reverse IDIRT

Env-3xFLAG

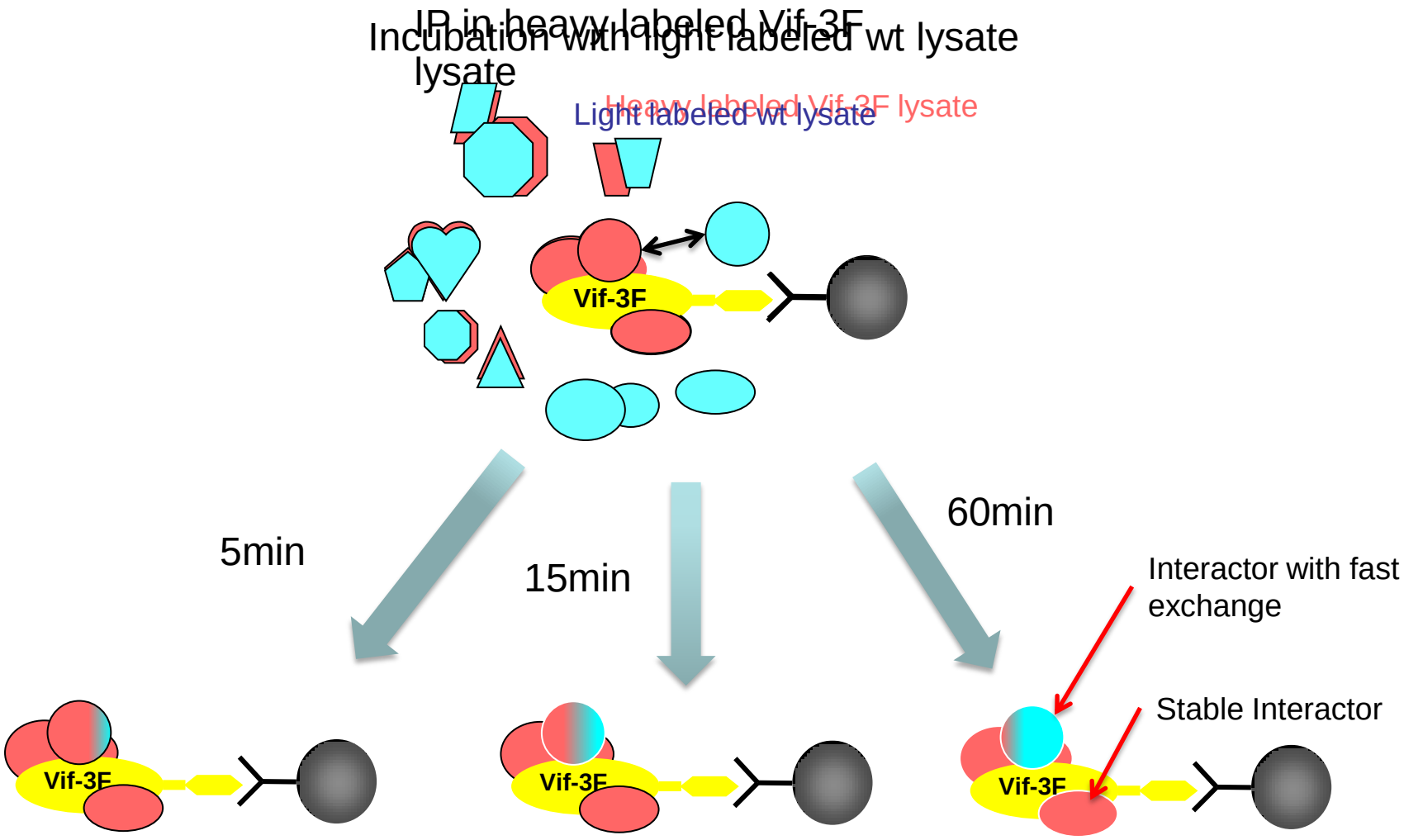


Vif-3xFLAG



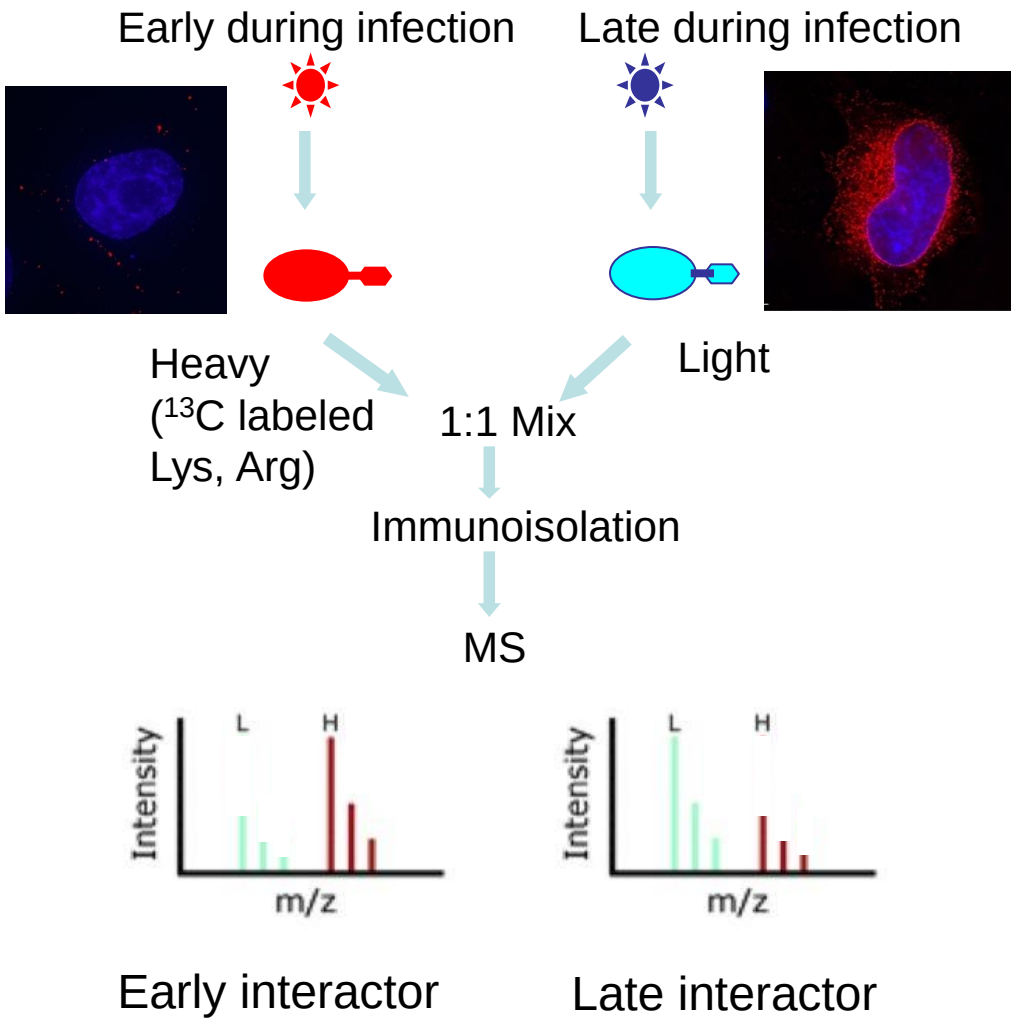
Luo, Jacobs, Greco, Cristae, Muesing, Chait, Rout

Protein Exchange

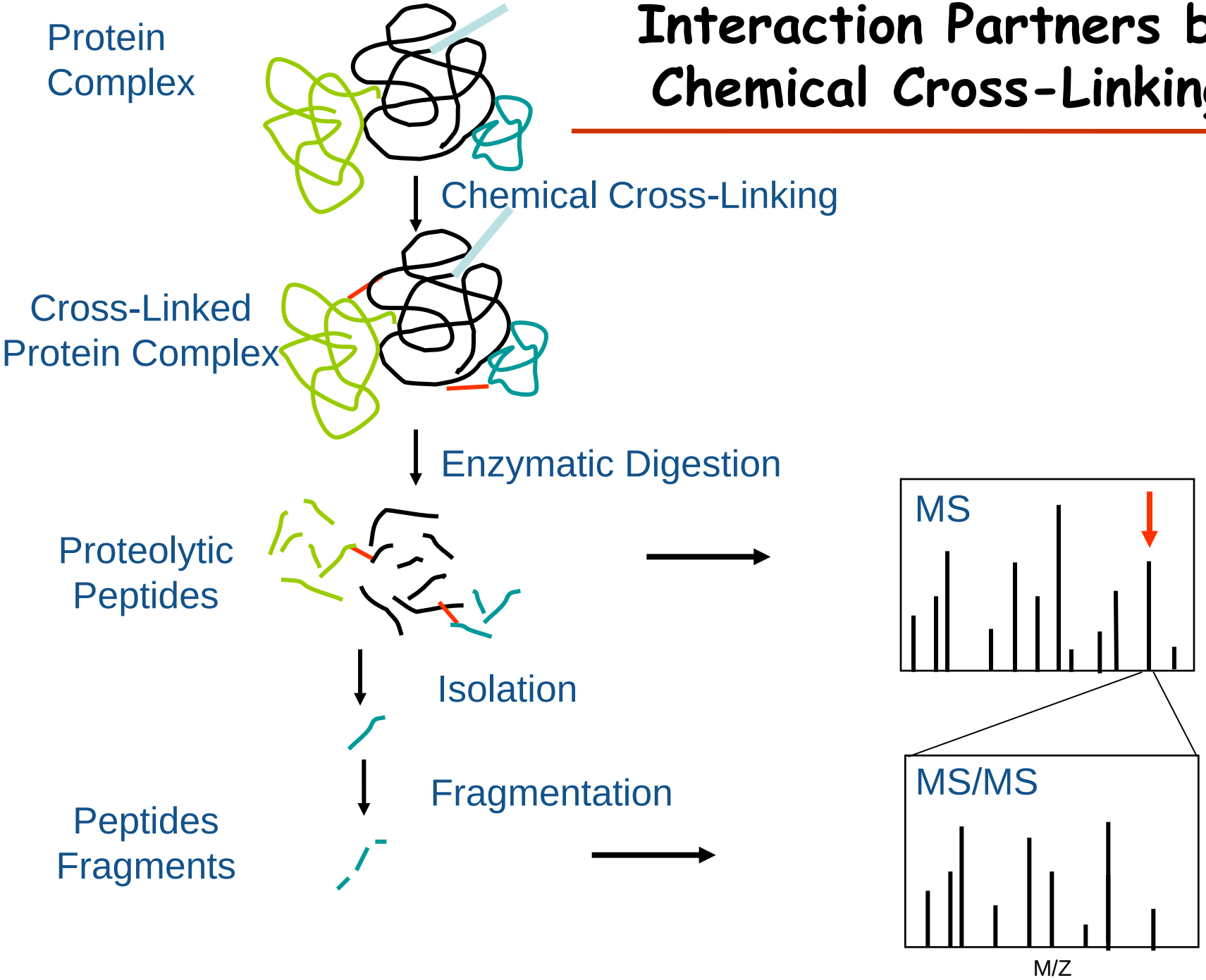


Env Time Course SILAC

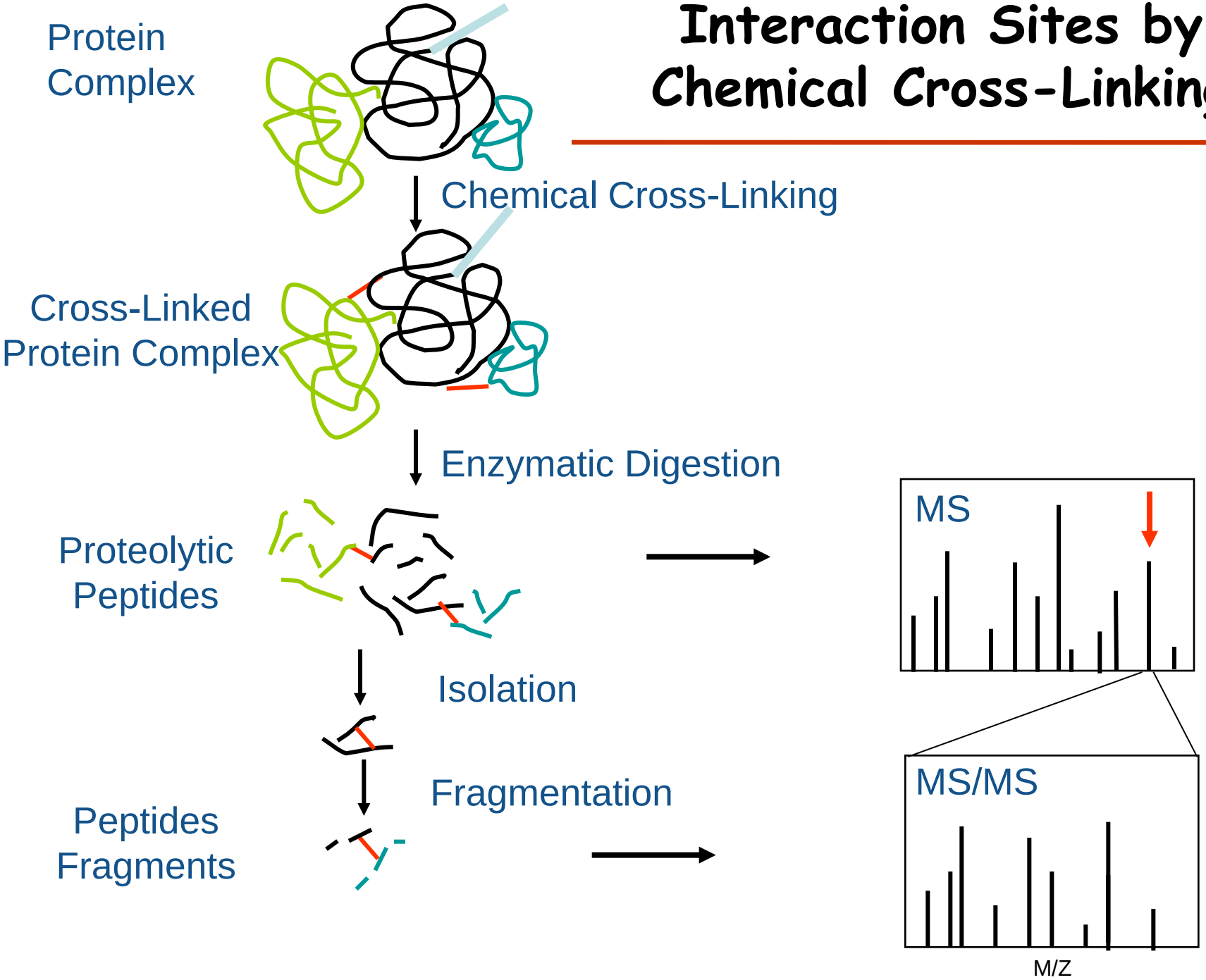
- Differentially labeled infection harvested at early or late stage of infection
- Distinguish proteins that interact with Env at early or late stage during infection



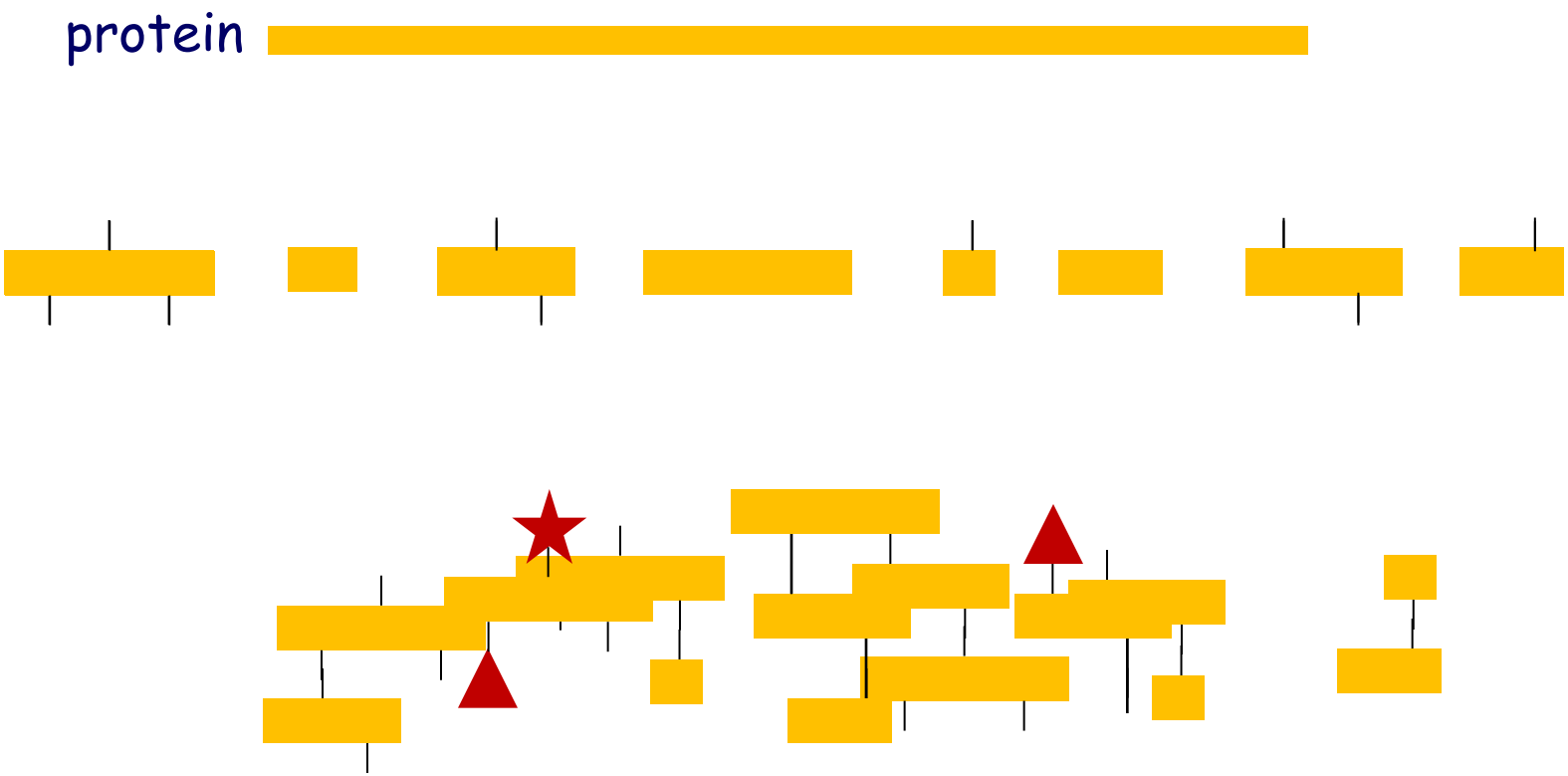
Interaction Partners by Chemical Cross-Linking



Interaction Sites by Chemical Cross-Linking



Cross-linking



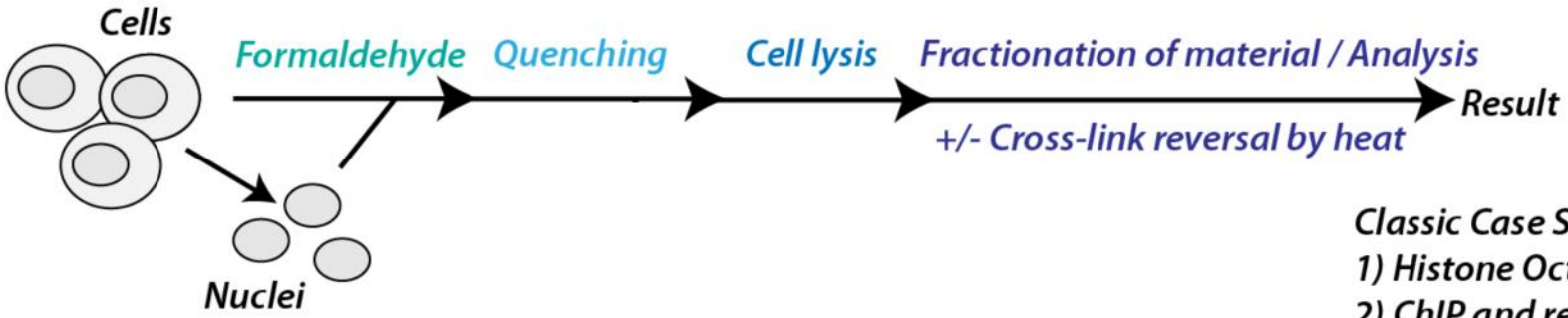
n peptides with reactive groups

(n-1)n/2 potential ways to cross-link peptides pairwise

+ many additional uninformative forms

Protein Crosslinking by Formaldehyde

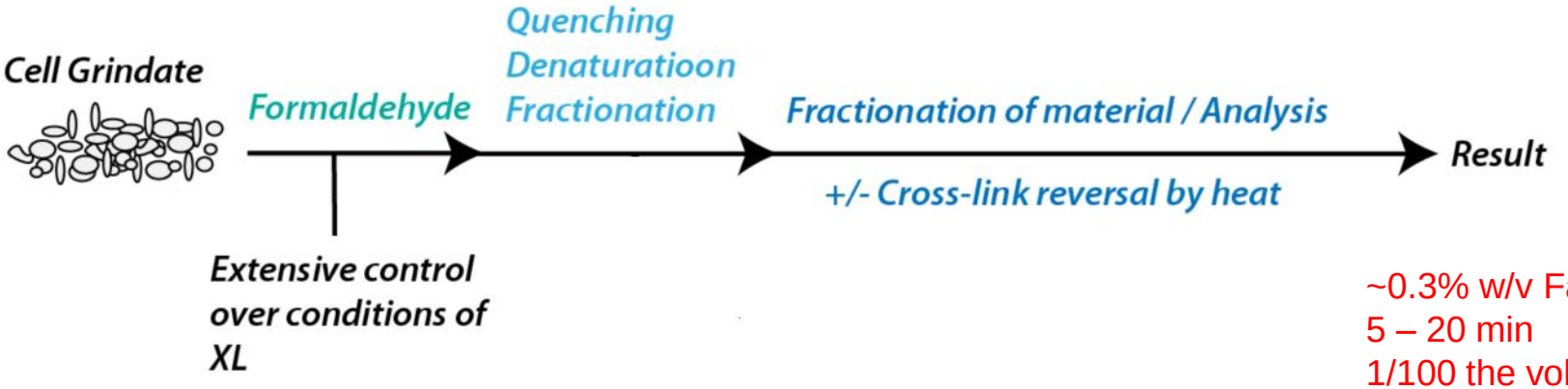
Example Traditional Formaldehyde Cross-linking:



Classic Case Studies:
1) Histone Octamer
2) ChIP and related methods

~1% w/v Fal
20 – 60 min

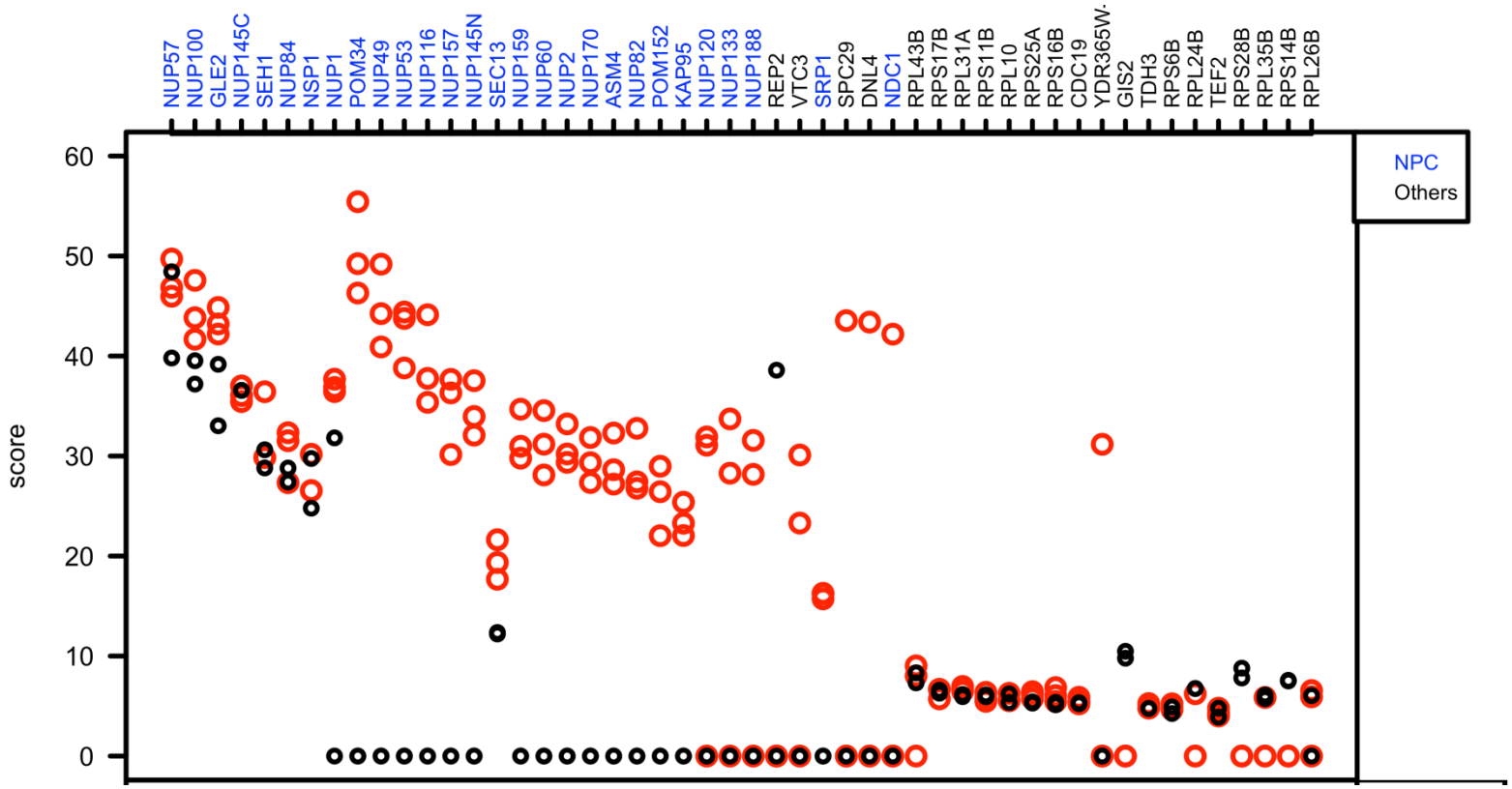
Example Modified Approach:



~0.3% w/v Fal
5 – 20 min
1/100 the volume

LaCava

Protein Crosslinking by Formaldehyde



RED: triplicate experiments, FAI treated grindate
BLACK: duplicated experiments, FAI treated cells (then ground)

SCORE: Log Ion Current / Log protein abundance

Akgöl, LaCava, Rout

Cross-linking

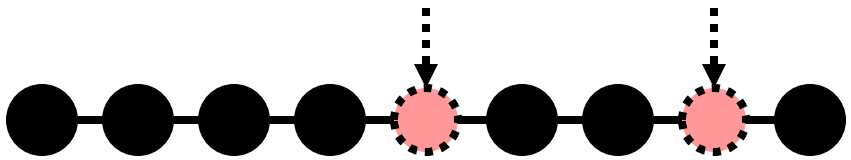
Mass spectrometers have a limited dynamic range and it therefore important to limit the number of possible reactions not to dilute the cross-linked peptides.

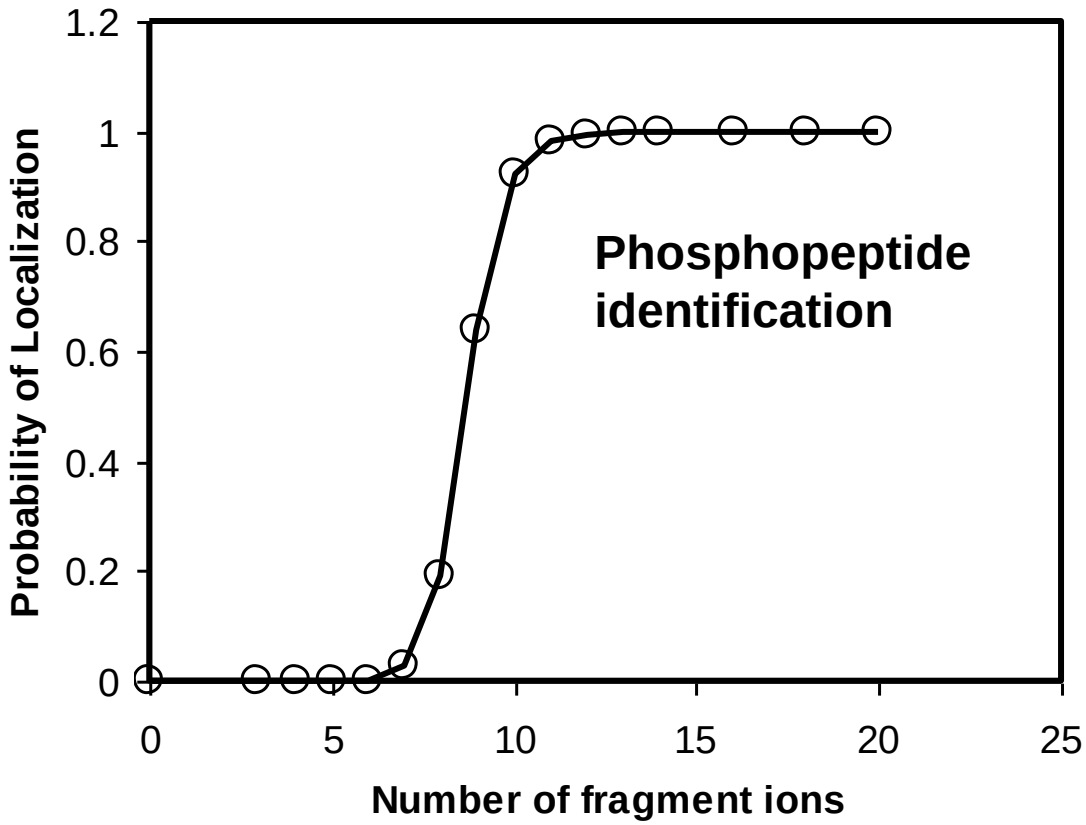
For identification of a cross-linked peptide pair, both peptides have to be sufficiently long and required to give informative fragmentation.

High mass accuracy MS/MS is recommended because the spectrum will be a mixture of fragment ions from two peptides.

Because the cross-linked peptides are often large, CAD is not ideal, but instead ETD is recommended.

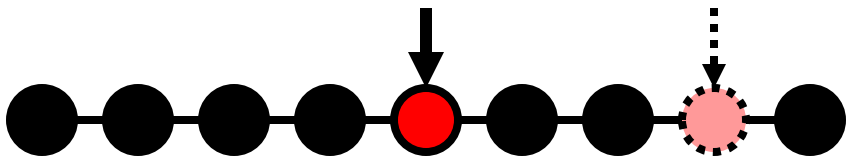
Localization of modifications

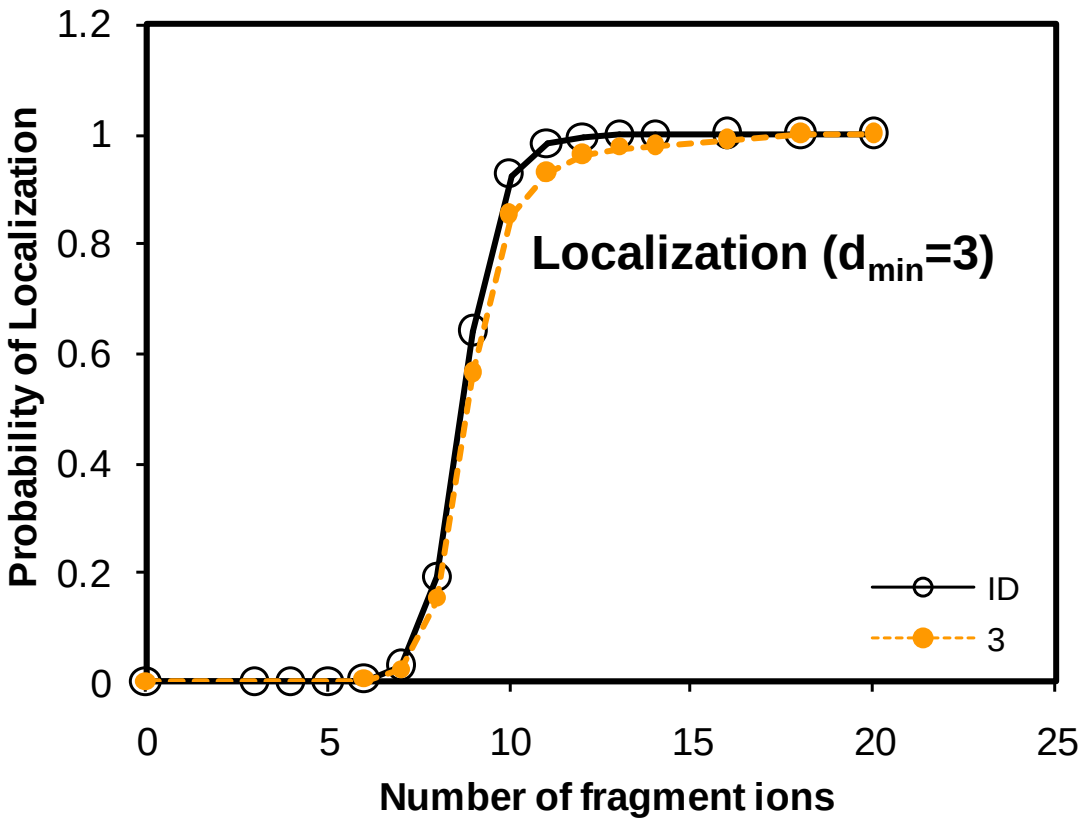




$m_{precursor}$ = 2000 Da
$\Delta m_{precursor}$ = 1 Da
$\Delta m_{fragment}$ = 0.5 Da
Phosphorylation

Localization of modifications

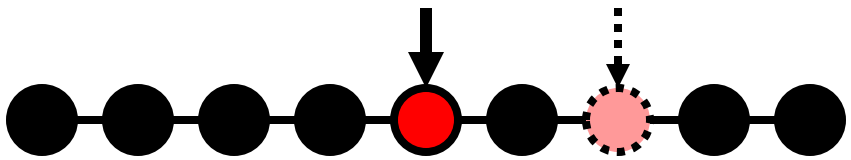


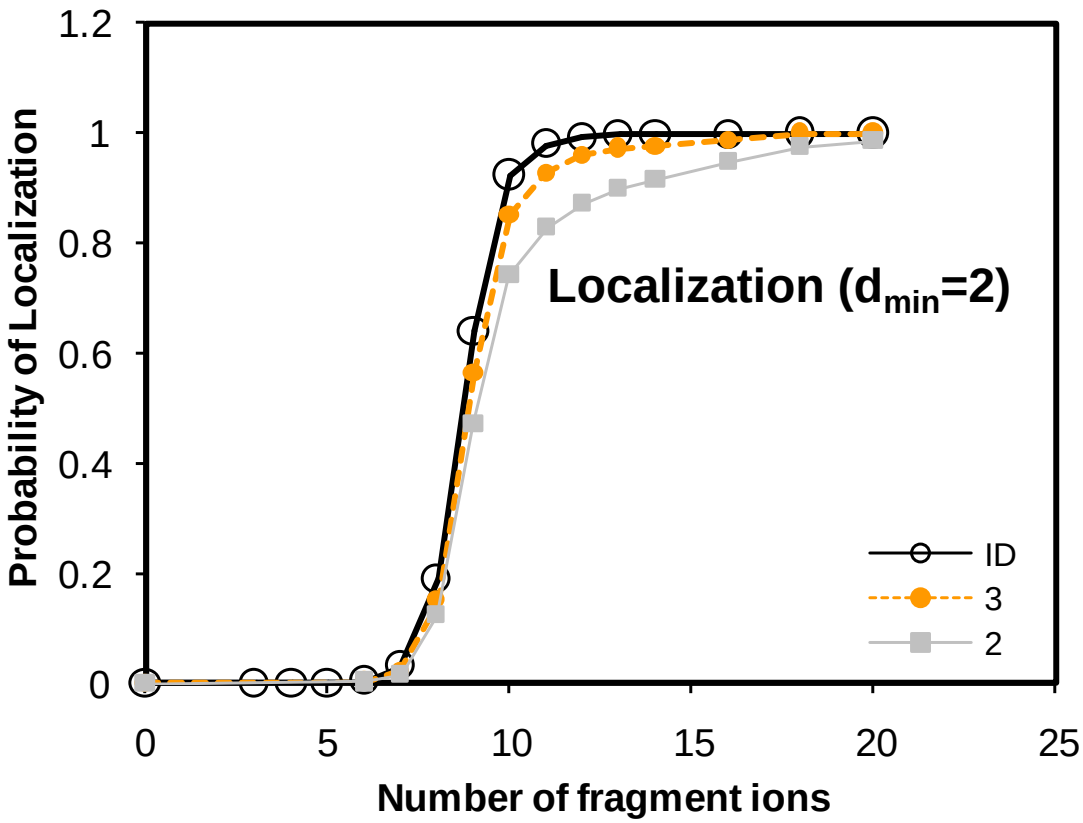


d_{min}>=3 for 47% of human tryptic peptides

$m_{precursor}$ = 2000 Da
$\Delta m_{precursor}$ = 1 Da
$\Delta m_{fragment}$ = 0.5 Da
Phosphorylation

Localization of modifications

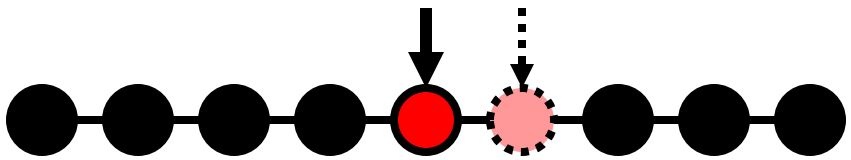


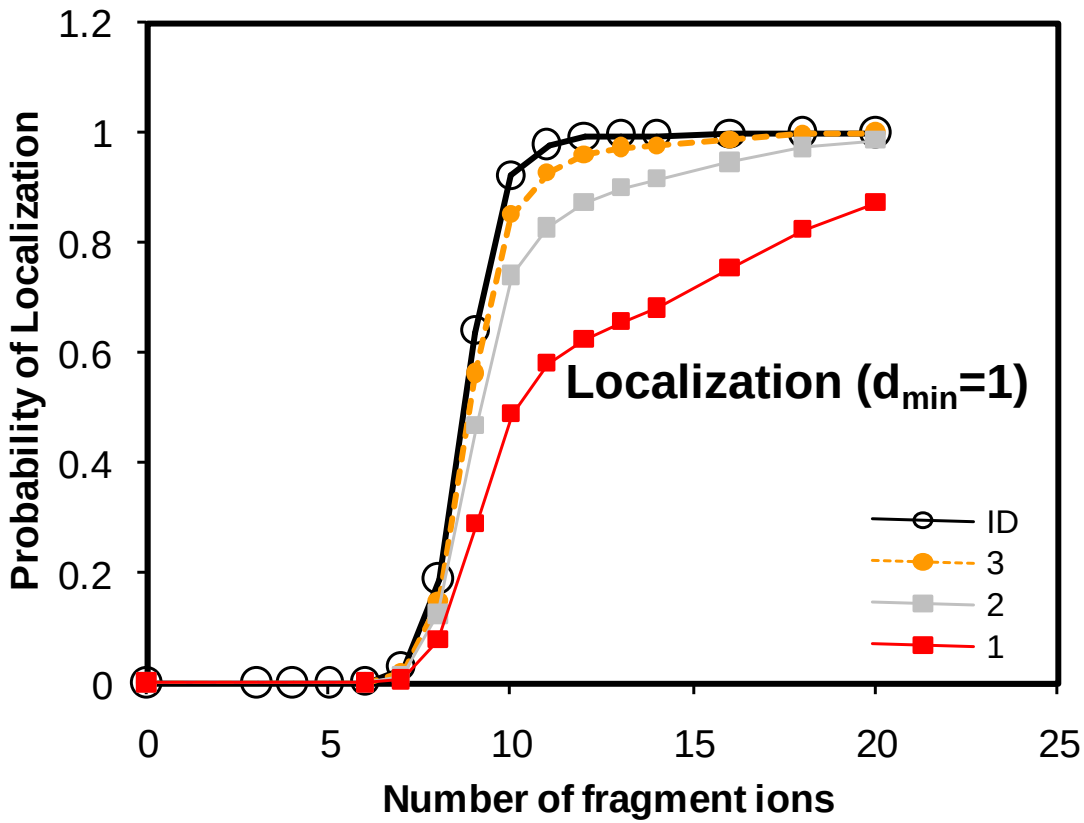


d_{min}=2 for 33% of human tryptic peptides

$m_{precursor}$ = 2000 Da
$\Delta m_{precursor}$ = 1 Da
$\Delta m_{fragment}$ = 0.5 Da
Phosphorylation

Localization of modifications

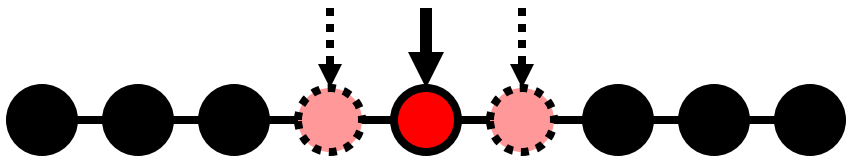


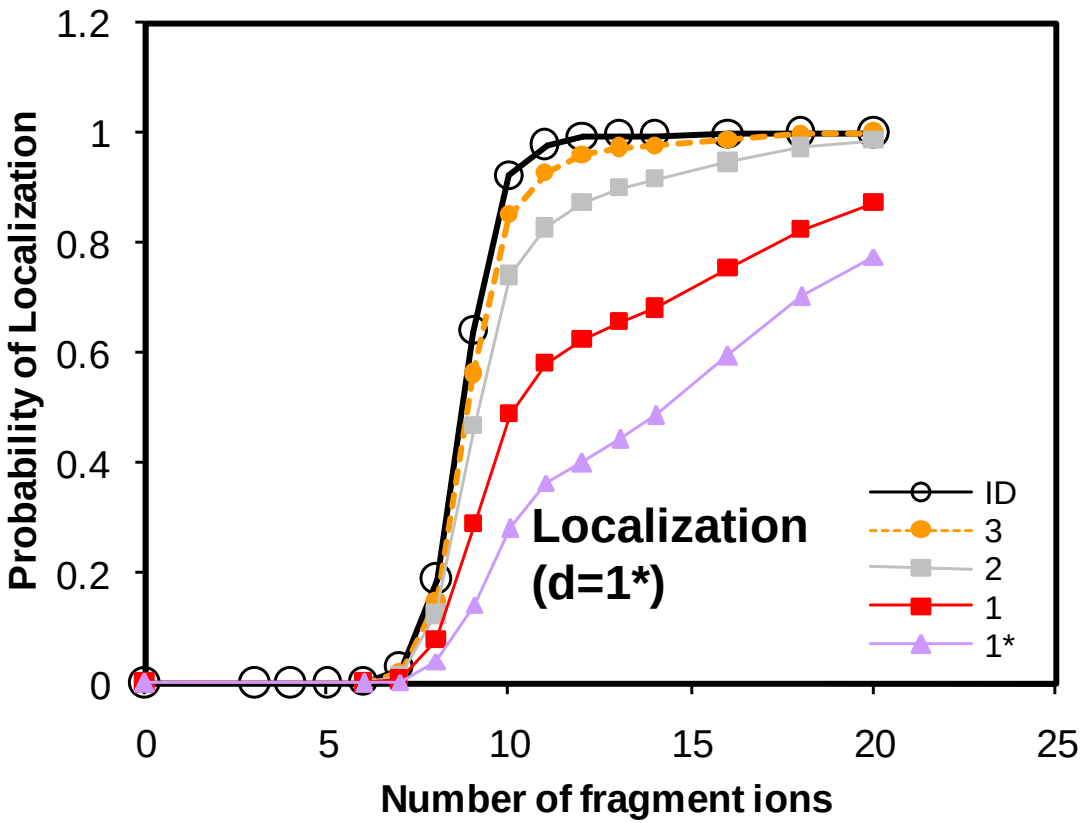


d_{min}=1 for 20% of human tryptic peptides

$m_{precursor}$ = 2000 Da
$\Delta m_{precursor}$ = 1 Da
$\Delta m_{fragment}$ = 0.5 Da
Phosphorylation

Localization of modifications

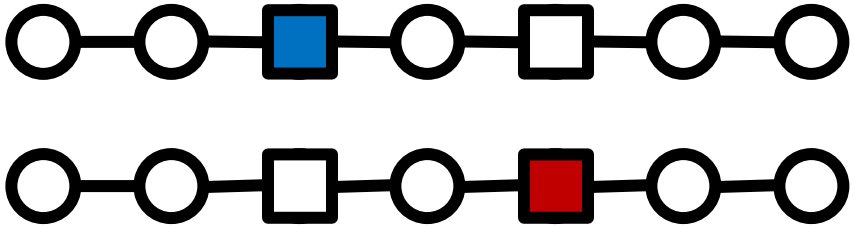




$m_{precursor}$ = 2000 Da
$\Delta m_{precursor}$ = 1 Da
$\Delta m_{fragment}$ = 0.5 Da
Phosphorylation

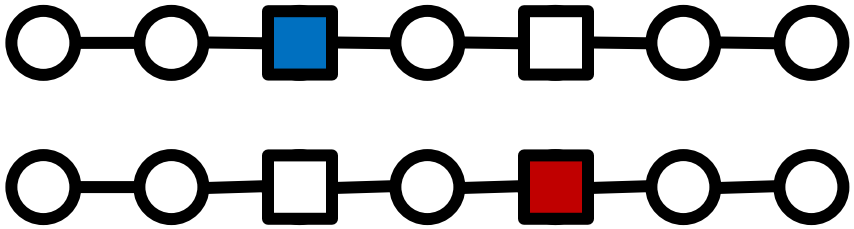
Localization of modifications

Peptide with two possible modification sites

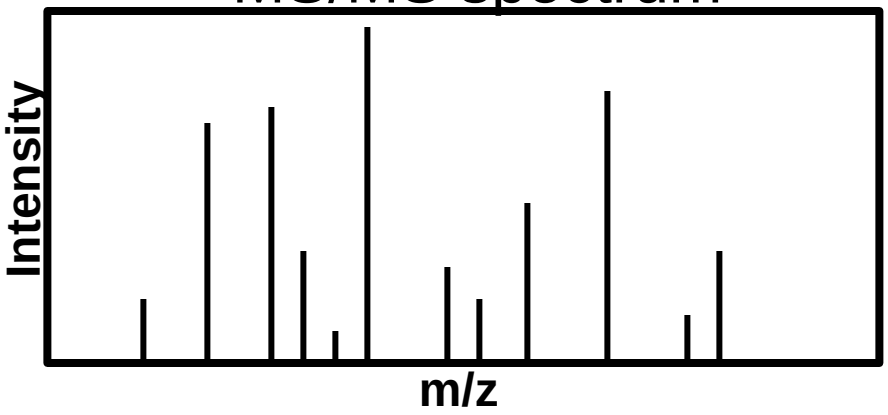


Localization of modifications

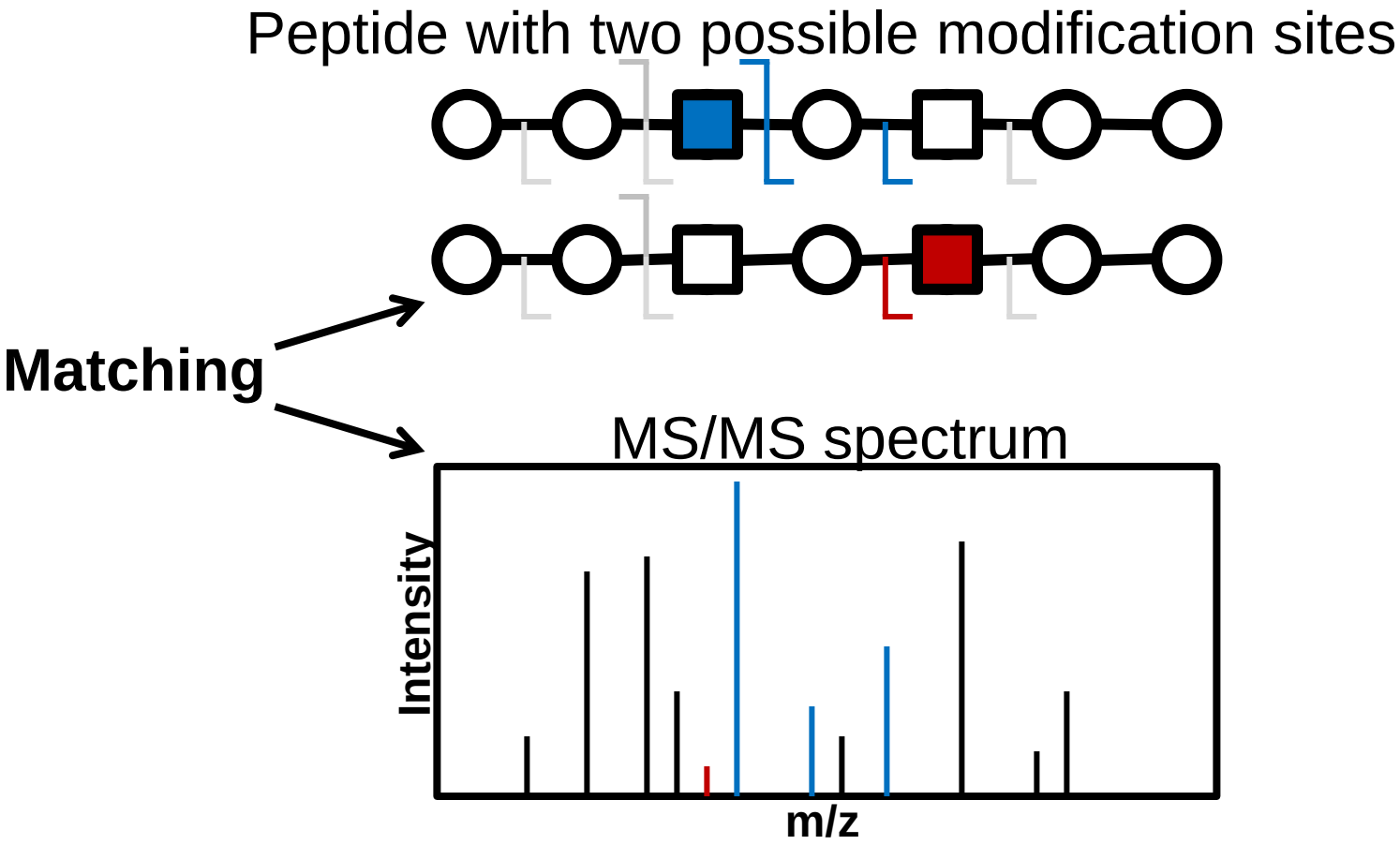
Peptide with two possible modification sites



MS/MS spectrum

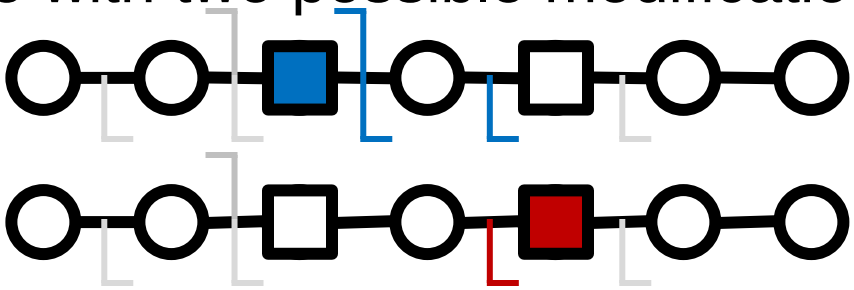


Localization of modifications



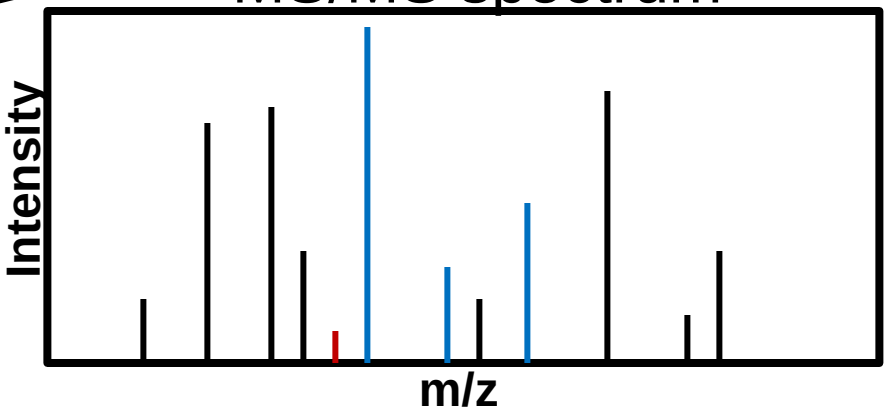
Localization of modifications

Peptide with two possible modification sites



Matching

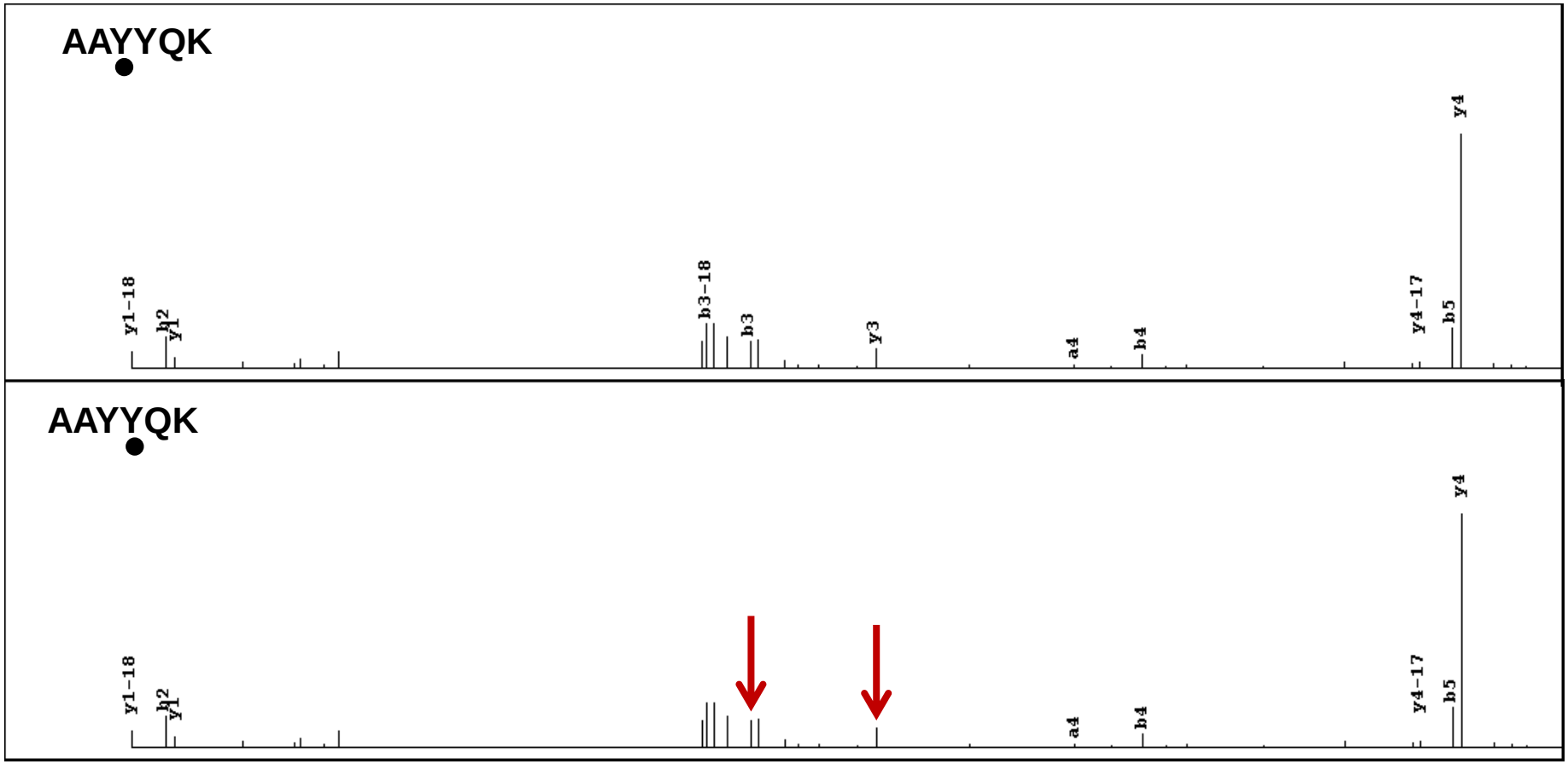
MS/MS spectrum



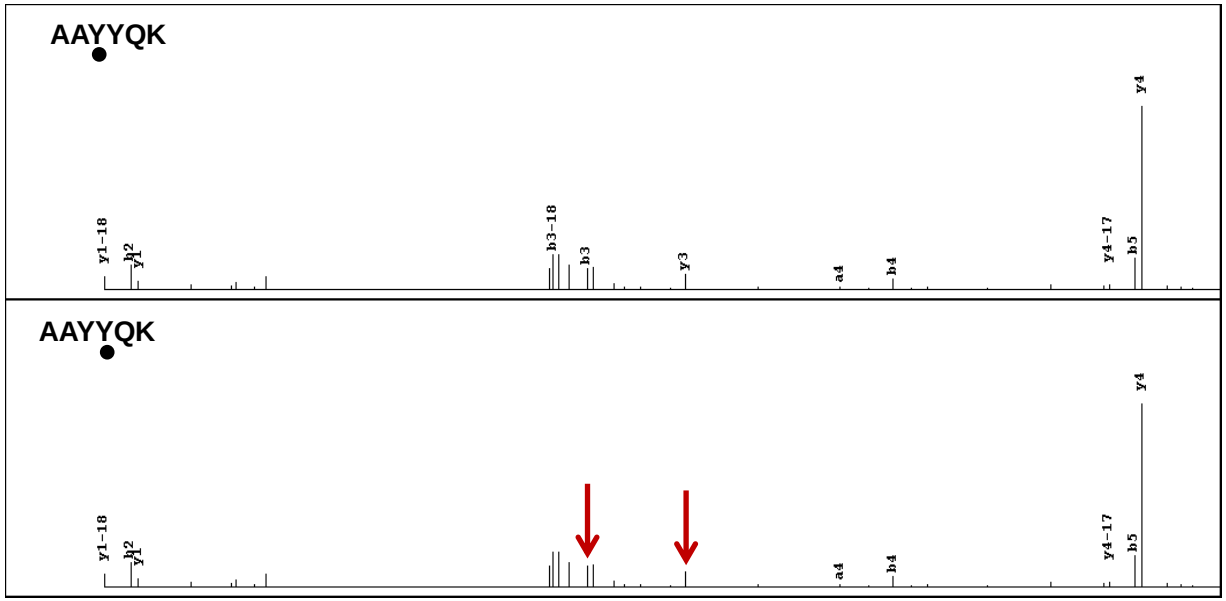
Which assignment does the data support?

1, 1 or 2, or 1 and 2?

Visualization of evidence for localization



Visualization of evidence for localization



Rank	AAYYQK	Total (matched)	Difference (matched)
1	AAYYQK	8	-
2	AAYYQK	6	2 0

Rank	AAYYQK	Total (intensity)	Difference (intensity)
1	AAYYQK	0.809	-
2	AAYYQK	0.735	0.074 0

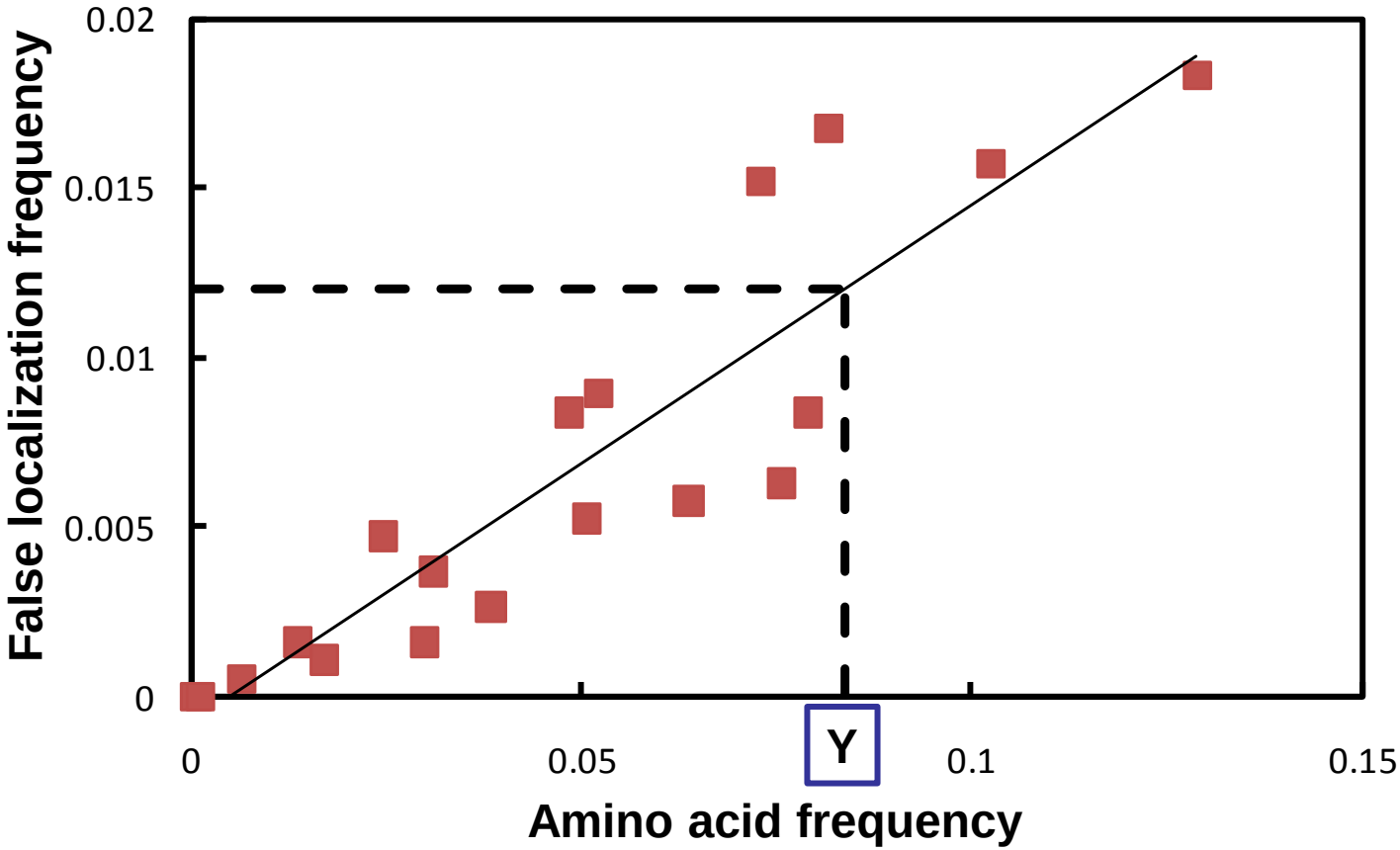
Visualization of evidence for localization

Rank	AAVPSGASTGIYEALELR	Total (matched)	Difference (matched)
1	AAVPSGASTGIYEALELR	15	-
2	AAVPSGASTGIYEALELR	14	3 2
3	AAVPSGASTGIYEALELR	13	4 2

Rank	AAVPSGASTGIYEALELR	Total (intensity)	Difference (intensity)
1	AAVPSGASTGIYEALELR	0.315	-
2	AAVPSGASTGIYEALELR	0.29	0.025 0
3	AAVPSGASTGIYEALELR	0.187	0.167 0.039

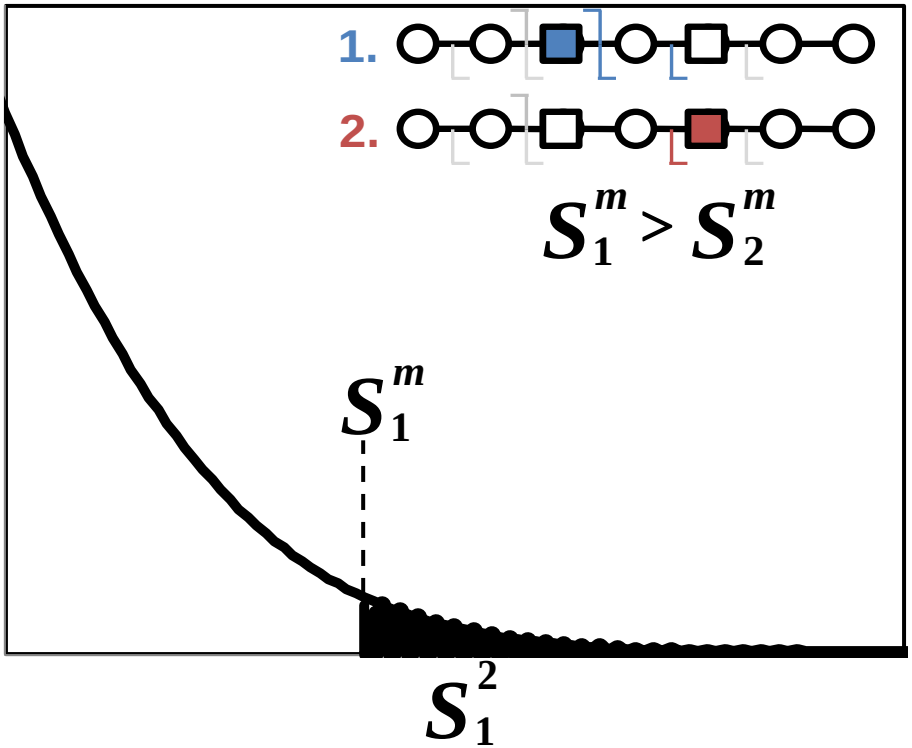
Estimation of global false localization rate using decoy sites

By counting how many times the phosphorylation is localized to amino acids that can not be phosphorylated we can estimate the false localization rate as a function of amino acid frequency.



How much can we trust a single localization assignment?

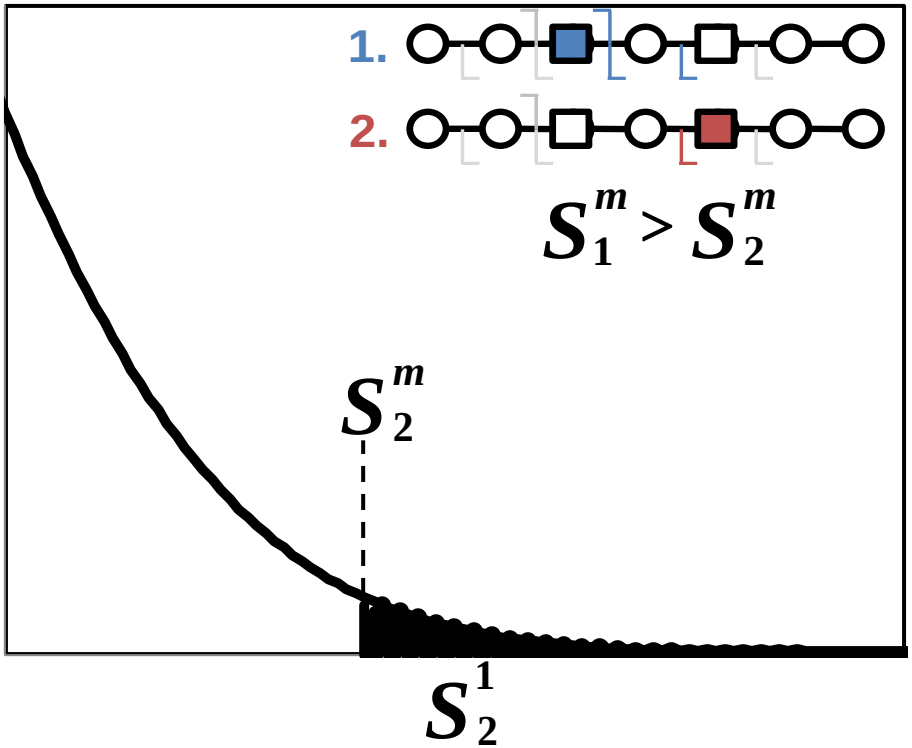
If we can generate the distribution of scores for assignment 1 when 2 is the correct assignment, it is possible to estimate the probability of obtaining a certain score by chance for a given peptide sequence and MS/MS spectrum assignment.



$$p_1^2 = \frac{\int_0^{S_1^m} F^2(S_1^2)\,dS_1^2}{\int_0^{\infty} F^2(S_1^2)\,dS_1^2}$$

Is it a mixture or not?

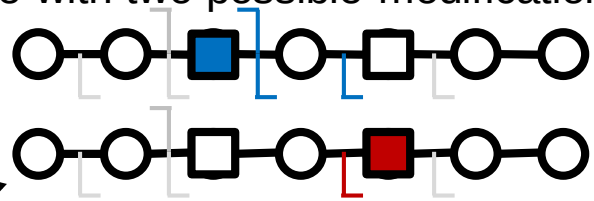
If we can generate the distribution of scores for assignment 2 when 1 is the correct assignment, it is possible to estimate the probability of obtaining a certain score by chance for a given peptide sequence and MS/MS spectrum assignment.



$$p_2^1 = \frac{\int_0^{S_2^m} F^1(S_2^1)\,dS_2^1}{\int_0^{\infty} F^1(S_2^1)\,dS_2^1}$$

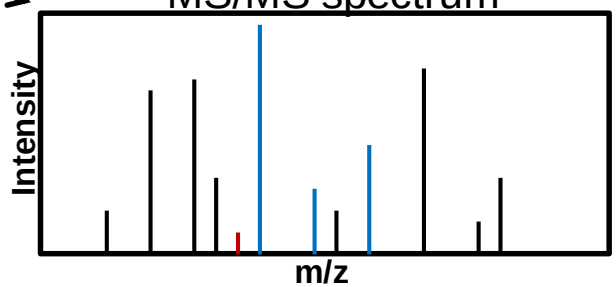
Localization of modifications

Peptide with two possible modification sites



Matching

MS/MS spectrum



Which assignment does the data support?

1, 1 or 2, or 1 and 2?

$$p_1^2 \le p_{th} \text{ and } p_2^1 \le p_{th} \Rightarrow \mathbf{1} \text{ and } \mathbf{2}$$

$$p_1^2 \le p_{th} \text{ and } p_2^1 > p_{th} \Rightarrow \mathbf{1}$$

$$p_1^2 > p_{th} \text{ and } p_2^1 \le p_{th} \Rightarrow \varnothing \quad (S_1^m \ge S_2^m \Rightarrow p_1^2 \le p_2^1)$$

$$p_1^2 > p_{th} \text{ and } p_2^1 > p_{th} \Rightarrow \mathbf{1} \text{ or } \mathbf{2}$$

Proteomics Informatics – Protein characterization: post-translational modifications and protein-protein interactions (Week 10)