Label Free kvantifikace, akviziční módy, databáze a FDR

Pavel Talacko

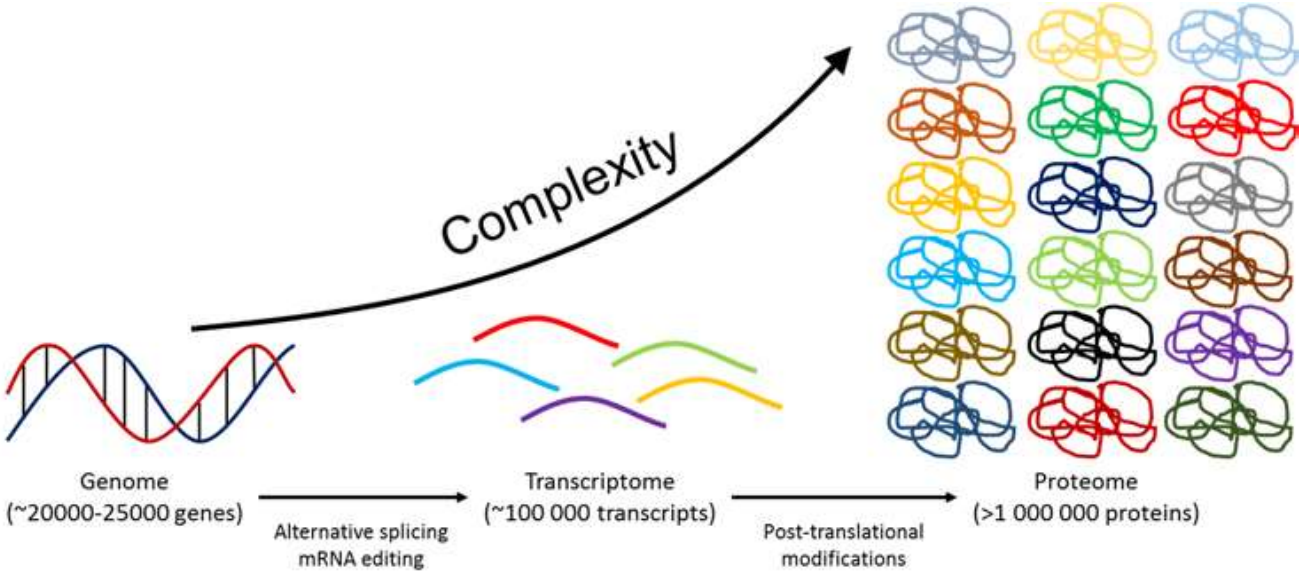
Laboratoř hmotnostní spektrometrie
Proteomics Core Facility
PřF UK, Biocev

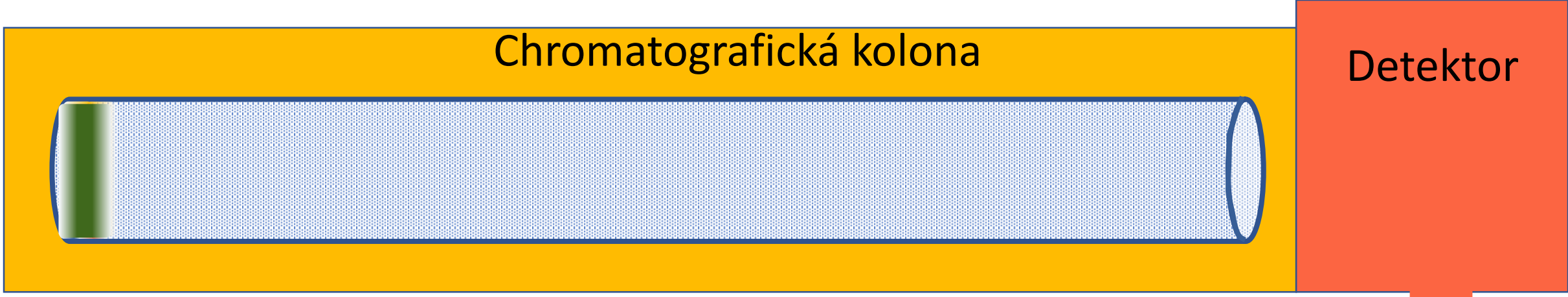
PŘÍRODOVĚDECKÁ
FAKULTA
Univerzita Karlova

BIOCEV

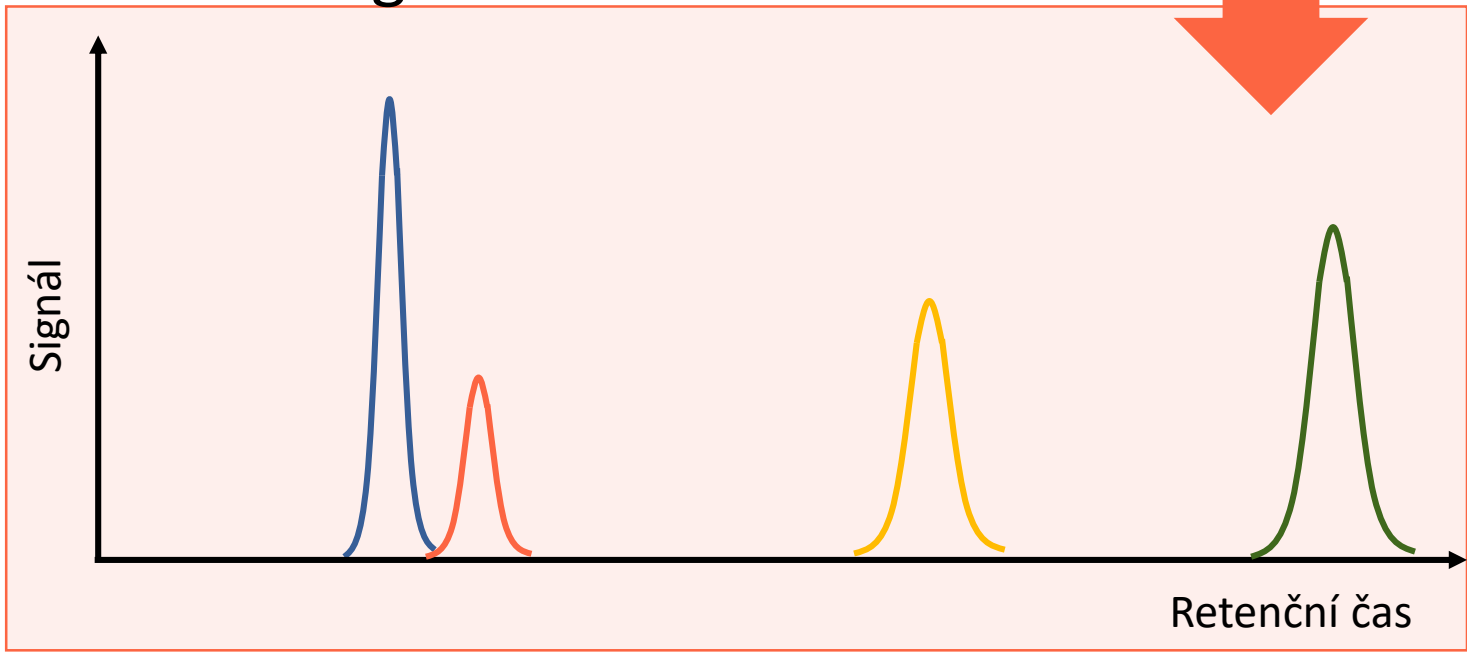
Komplexita proteomu

- Lidský genom obsahuje asi 20 000 genů
- Průměrná lidská buňka exprimuje zhruba 10 000 genů
- Komplexita proteomu výrazně vyšší než komplexita genomu
- Navýšení komplexity v důsledku alternativního splicingu, alternativního startu translace, posttranslačních modifikací…
- Vzorek třeba nejprve separovat pro snížení komplexity



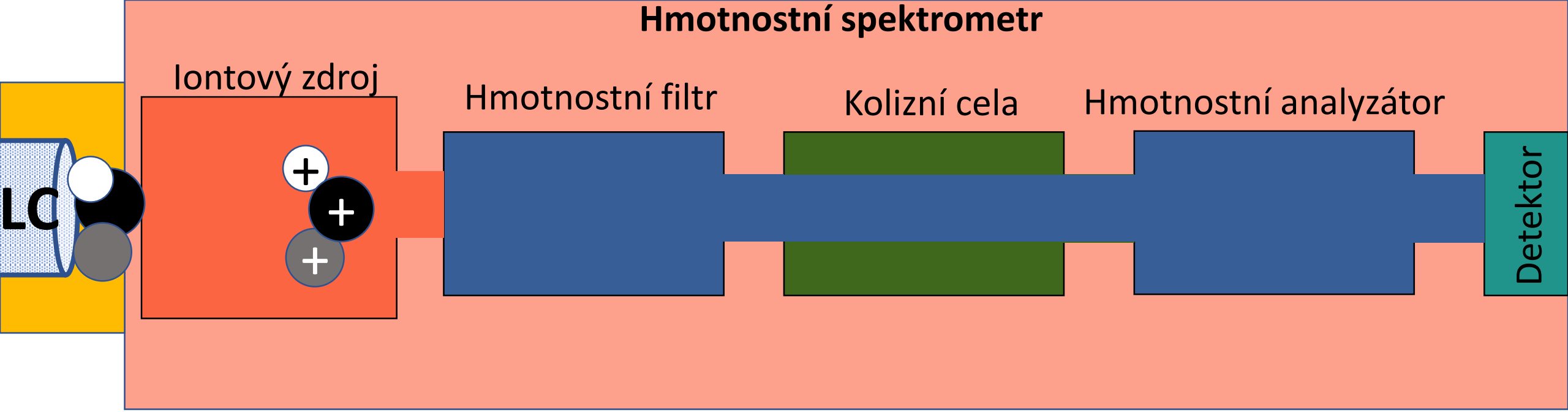


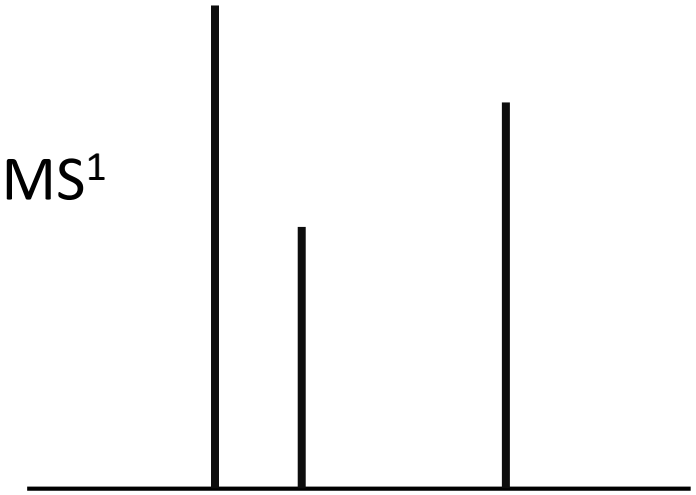
Chromatogram

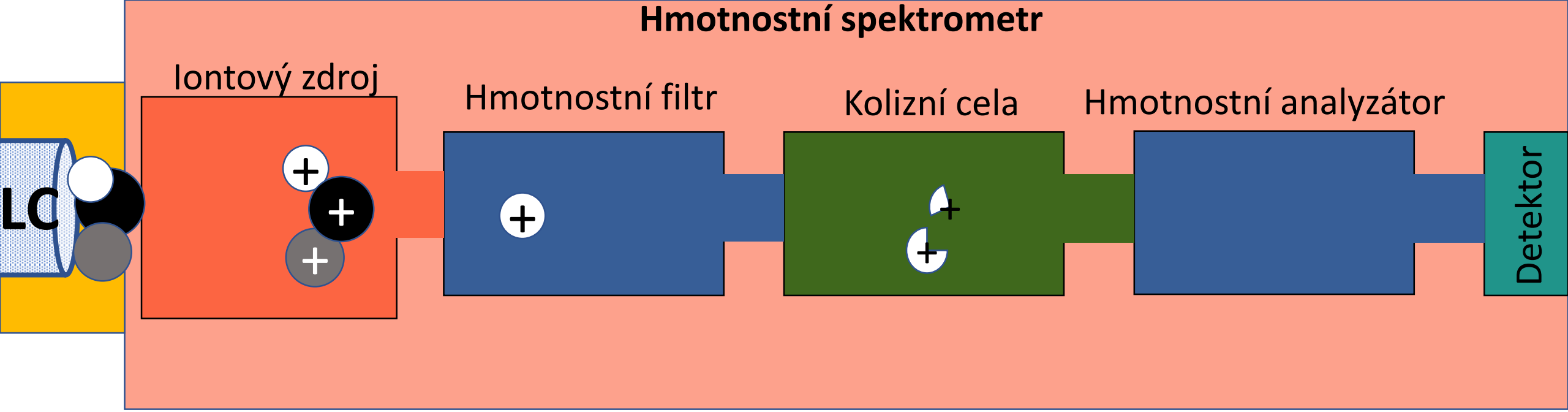


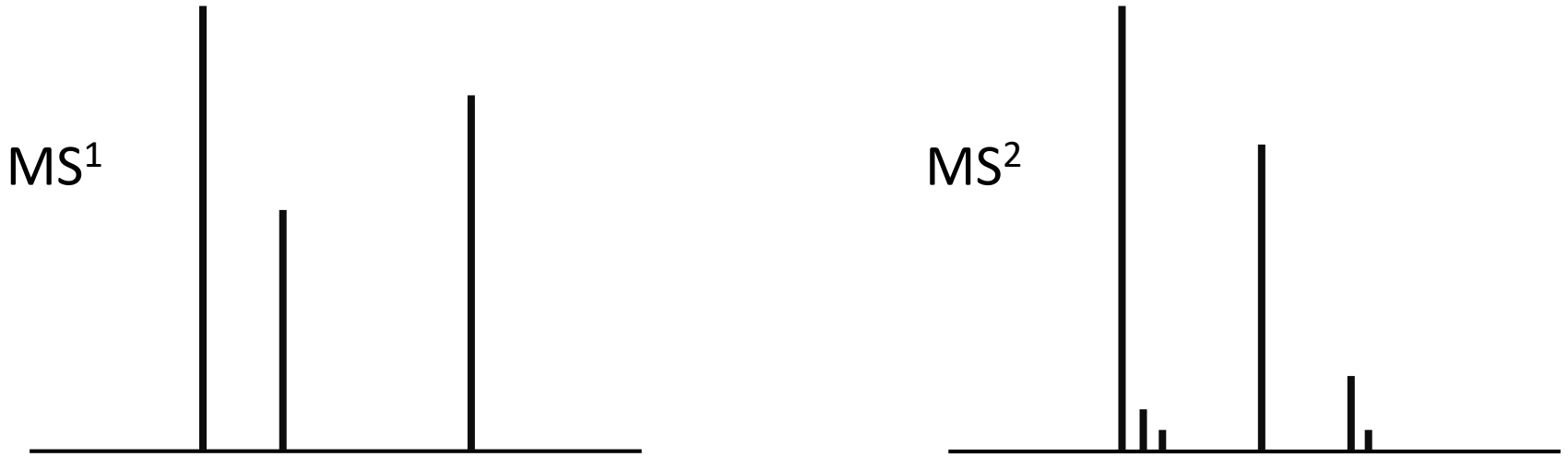
Kapalinová chromatografie (LC)

- šířka chromatografického píku typicky 10 - 20 s


Hmotnostní spektrometr
Iontový zdroj
Hmotnostní filtr
Kolizní cela
Hmotnostní analyzátor
LC
Detektor

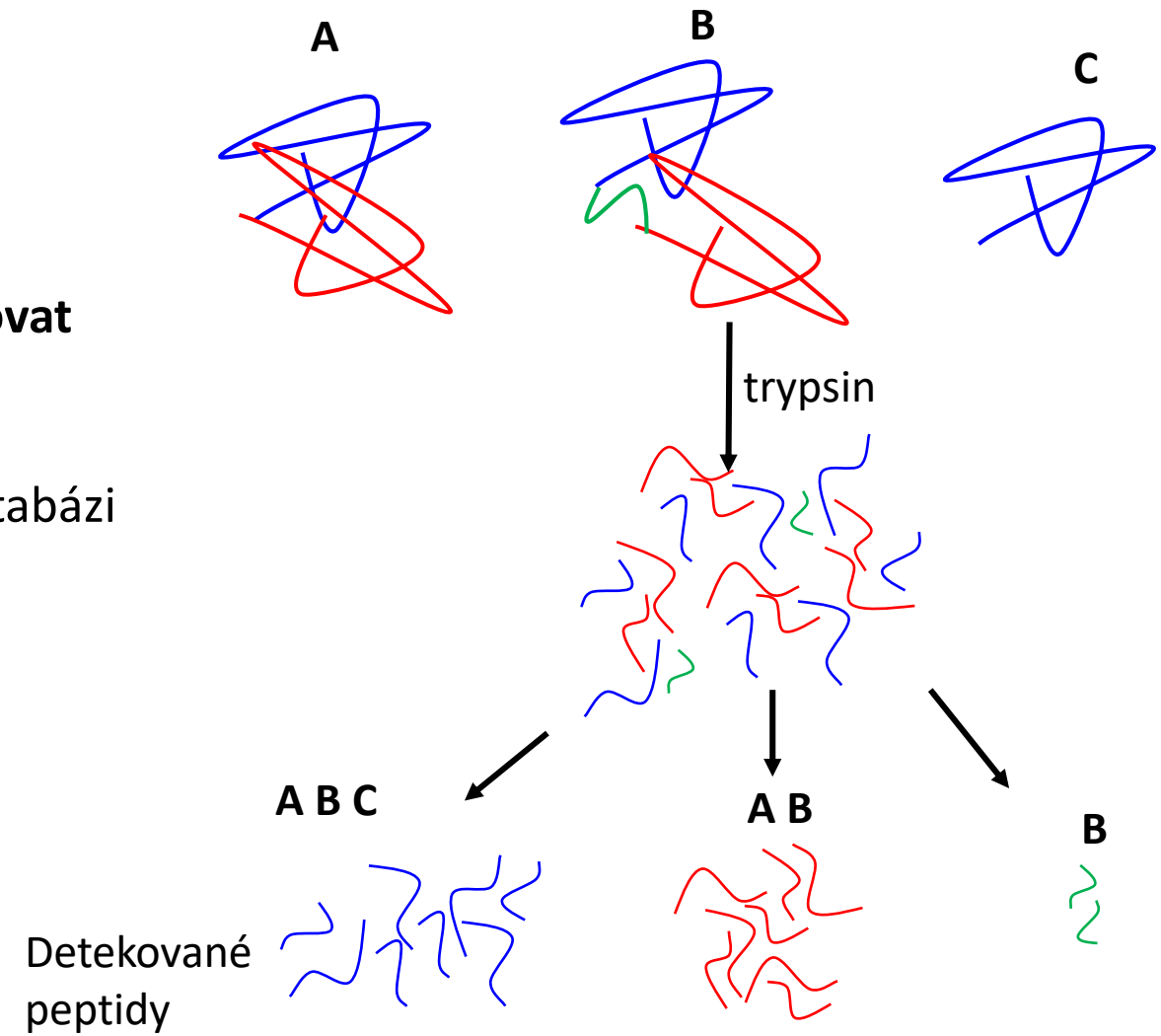

MS1


Hmotnostní spektrometr
Iontový zdroj
Hmotnostní filtr
Kolizní cela
Hmotnostní analyzátor
LC
Detektor

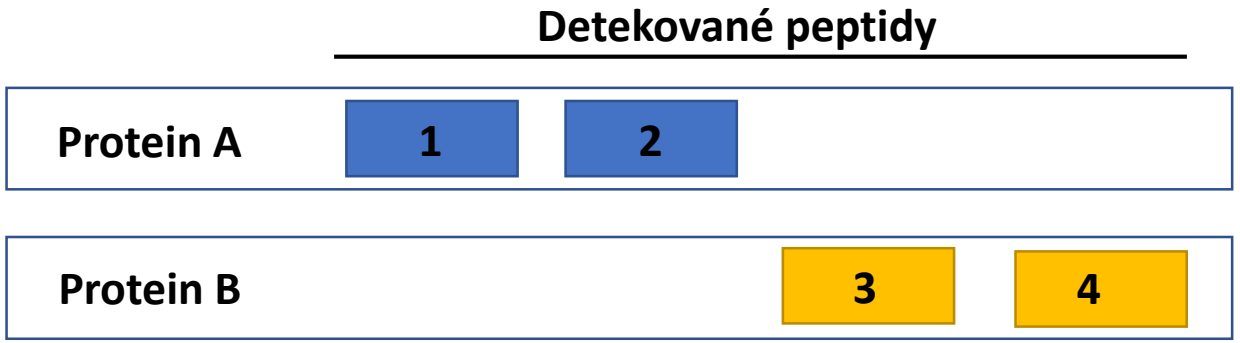

MS1
MS2

Bottom up proteomika

- Detekujeme peptidy, ne proteiny
- Proteolytickým štěpením vzorků ztrácíme informaci o propojení peptid-protein
- **Seznam identifikovaných proteinů musíme rekonstruovat zpětně z identifikovaných peptidů**
- Vedle naměřených dat **potřebujeme** k vyhodnocení databázi proteinů – **neprobíhá de novo sekvenace**
- Problém v případě:
  - sekvenčních homologií
  - alternativního sestřihu mRNA
  - různých variant posttranslačního processingu
  - redundantní databáze



Protein inference – od peptidů zpět k proteinům

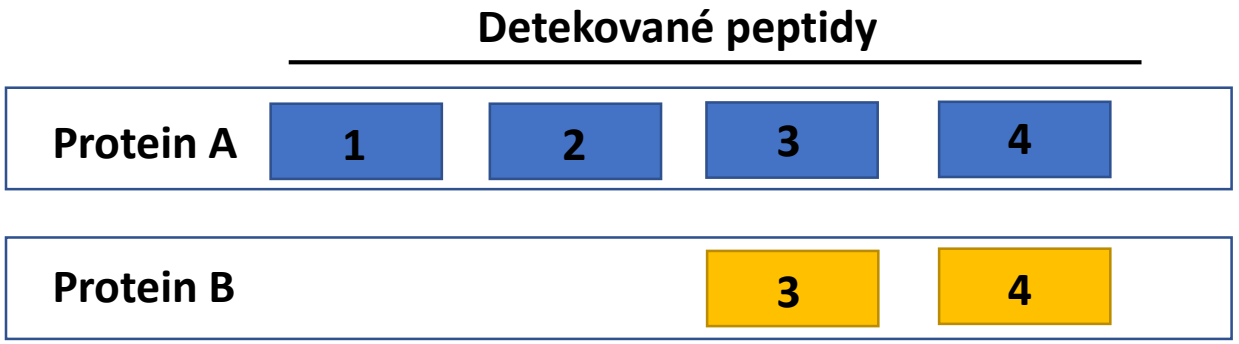


Protein A – 2 unikátní peptidy 1-2

Protein B – 2 unikátní peptidy 3-4

- Dvě jednoznačně odlišitelné protein groups
- **Unikátní (proteotypický) peptid** – lze jednoznačně přiřadit k jednomu konkrétnímu proteinu

Protein inference – od peptidů zpět k proteinům

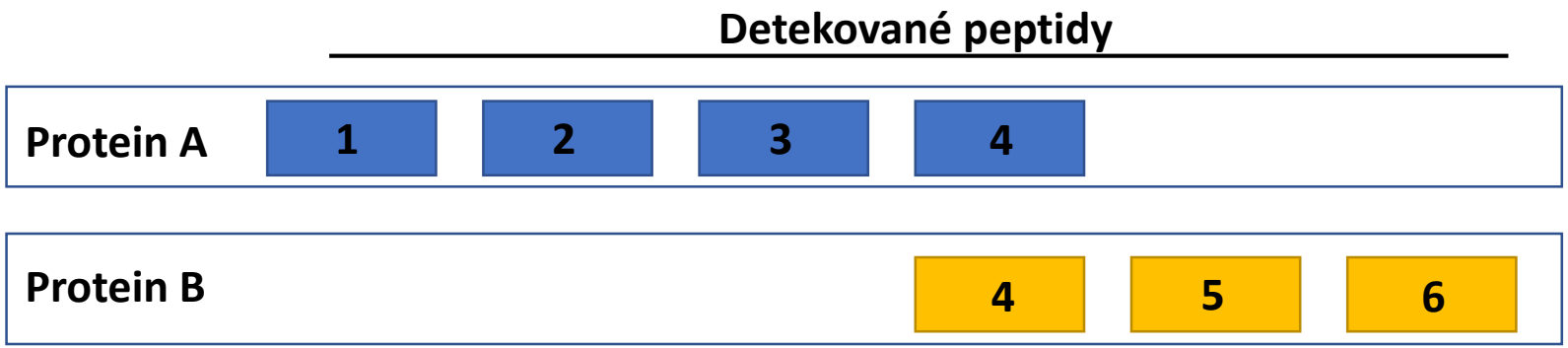


Protein A – 2 unikátní peptidy 1-2

Protein B – žádné unikátní peptidy
Peptidy 3 – 4 sdílené proteiny A a B

- Protein A určitě přítomen, přítomnost proteinu B nelze vyloučit
- Ve výsledku proteiny A a B spojeny do jedné protein group

Protein inference – od peptidů zpět k proteinům



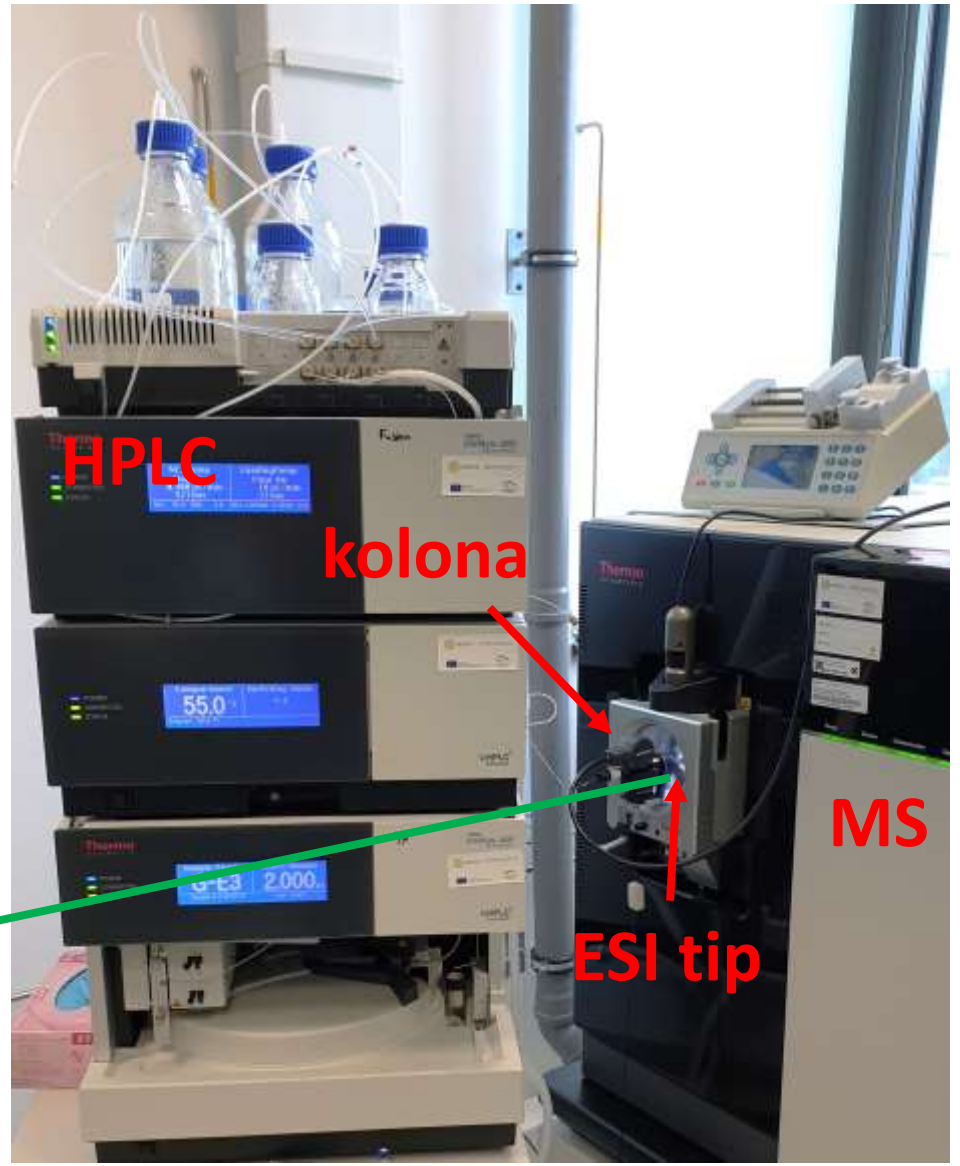
Protein A – 3 unikátní peptidy 1-3

Protein B – 2 unikátní peptidy 5-6
Peptid 4 = razor peptid

- Proteiny A a B jsou identifikovány na základě unikátních peptidů a nemůžeme je tak zařadit do jedné protein group
- Kvantitativní hodnota razor peptidu bude přičtena proteinu o vyšším počtu unikátních peptidů

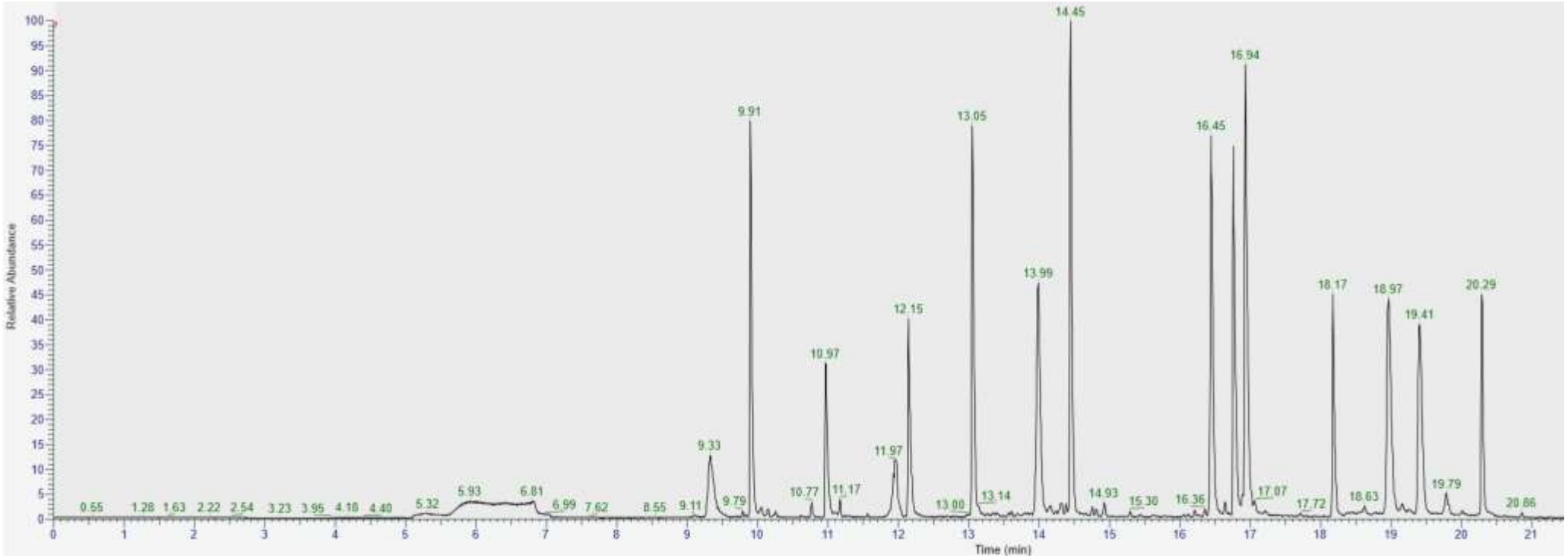
Analýza komplexních vzorků v praxi

- Současná technika umožňuje detekci a kvantifikaci 5 000 – 9 000 proteinů z jednoho nástřiku lyzátu savčích buněk – zhruba polovina celkového proteomu
- HPLC separace online propojená s MS = LC-MS (LC-MS/MS)
- Nutná MS detekce v časové škále odpovídající LC separaci – šířka chromatografického píku obvykle kolem 20 s



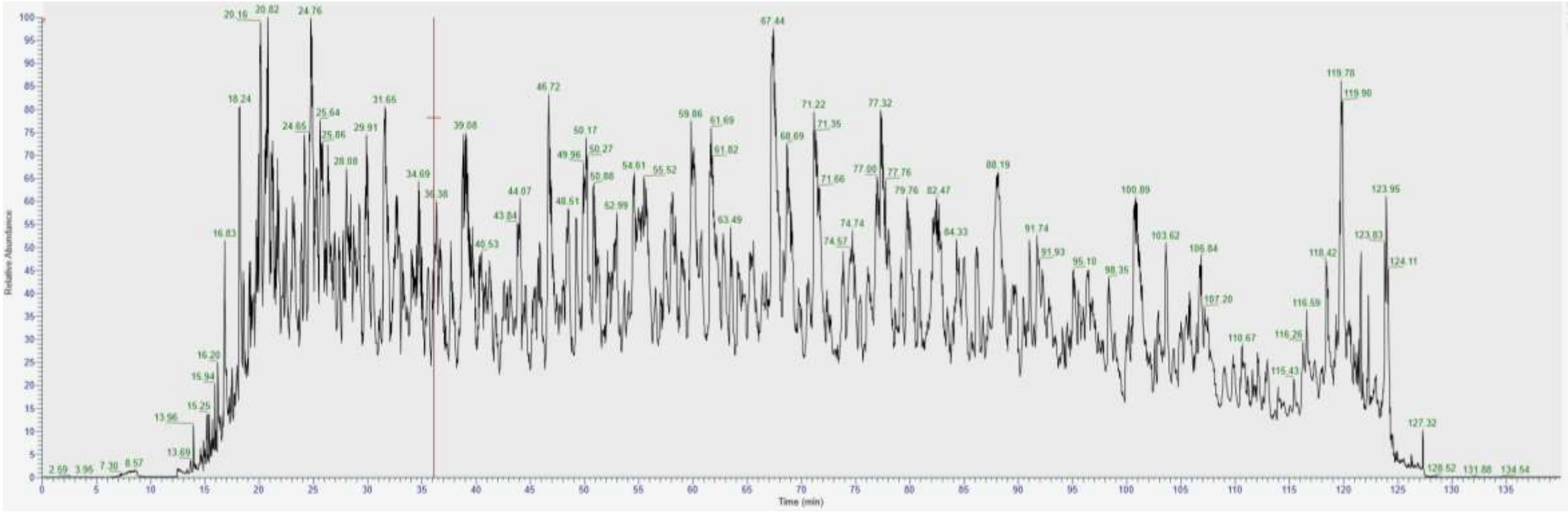
Analýza komplexních vzorků

Směs 14 peptidů – málo komplexní vzorek



Analýza komplexních vzorků

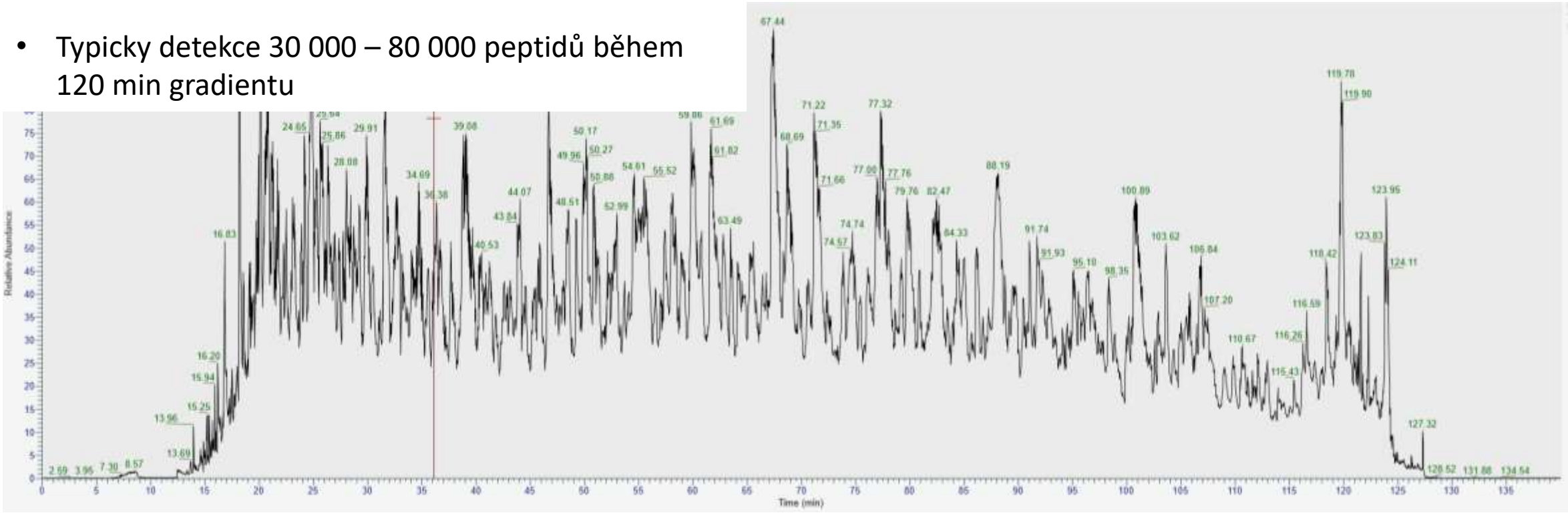
- Analýza 1 µg buněk ovariálního karcinomu (Hela)



Analýza komplexních vzorků

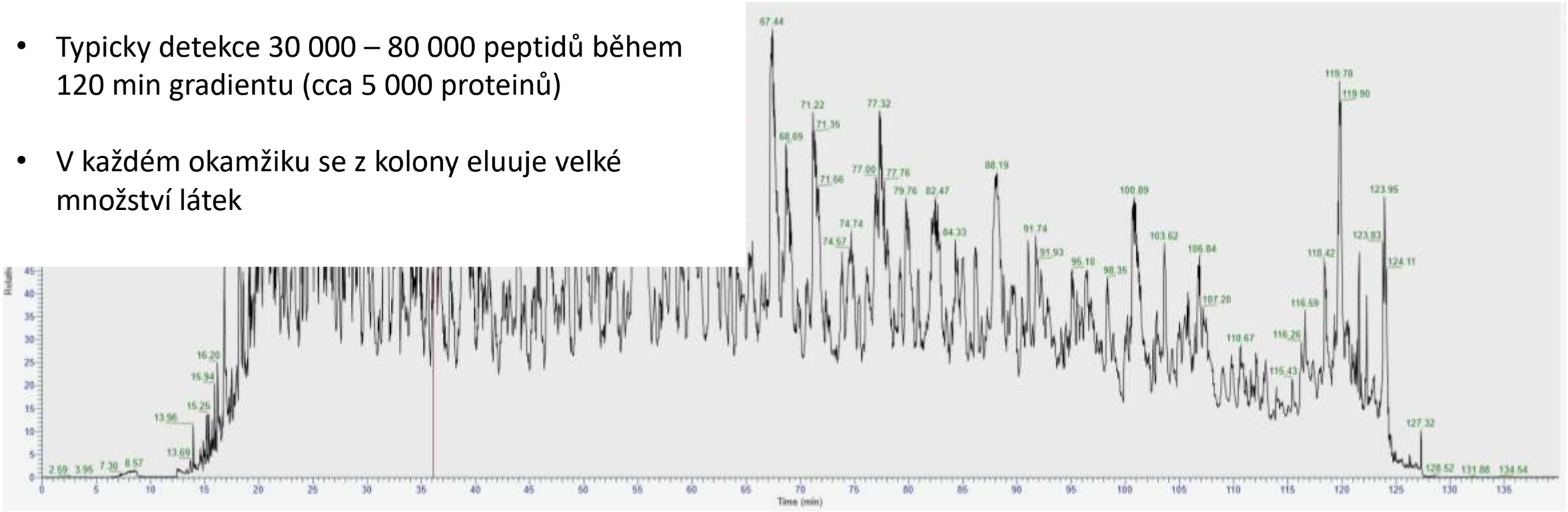
- Analýza 1 µg buněk ovariálního karcinomu (Hela)

- Typicky detekce 30 000 – 80 000 peptidů během 120 min gradientu



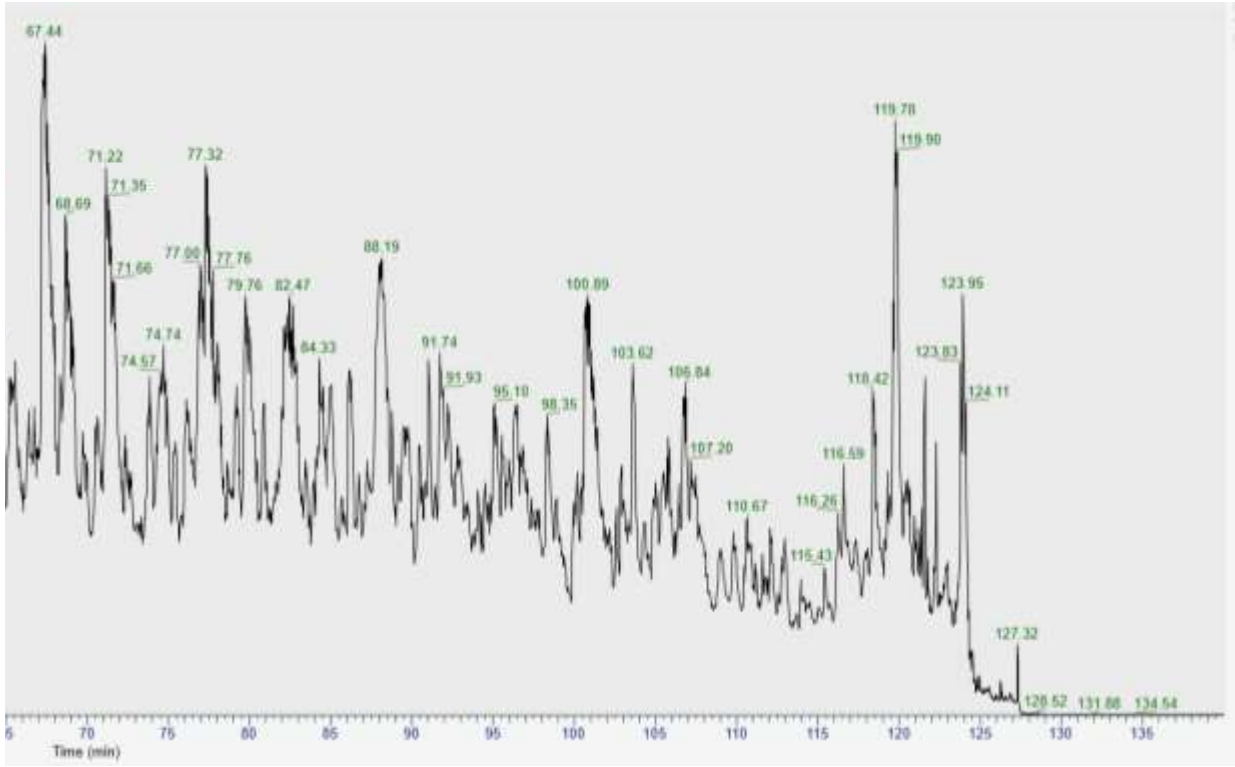
Analýza komplexních vzorků

- Analýza 1 µg buněk ovariálního karcinomu (Hela)
- Typicky detekce 30 000 – 80 000 peptidů během 120 min gradientu (cca 5 000 proteinů)
- V každém okamžiku se z kolony eluuje velké množství látek



Analýza komplexních vzorků

- Analýza 1 µg buněk ovariálního karcinomu (Hela)
- Typicky detekce 30 000 – 80 000 peptidů během 120 min gradientu
- V každém okamžiku se z kolony eluuje velké množství látek
- **Při zachování rozumného množství bodů přes pík (5 – 10) potřeba sběru až desítek spekter za vteřinu – 10 Hz – 40 Hz ->** během 120 min gradientu nasbíráme zhruba 250 000 spekter (30 Hz)
- z 250 000 spekter je následně přiřazeno k peptidu asi 25%



Akviziční módy

- **MS1 a MS2 (MSMS)** scany poskytují různé druhy informací

- Cyklické střídání MS1 a MS2 scanů podle předem daných pravidel

- Cílem je detekovat co nejvíce peptidů (počet MSMS spekter) s co nejkvalitnější kvantitativní informací (počet bodů přes pík)

- **Necílené akviziční módy:**
  - **DDA – data dependent acquisition**

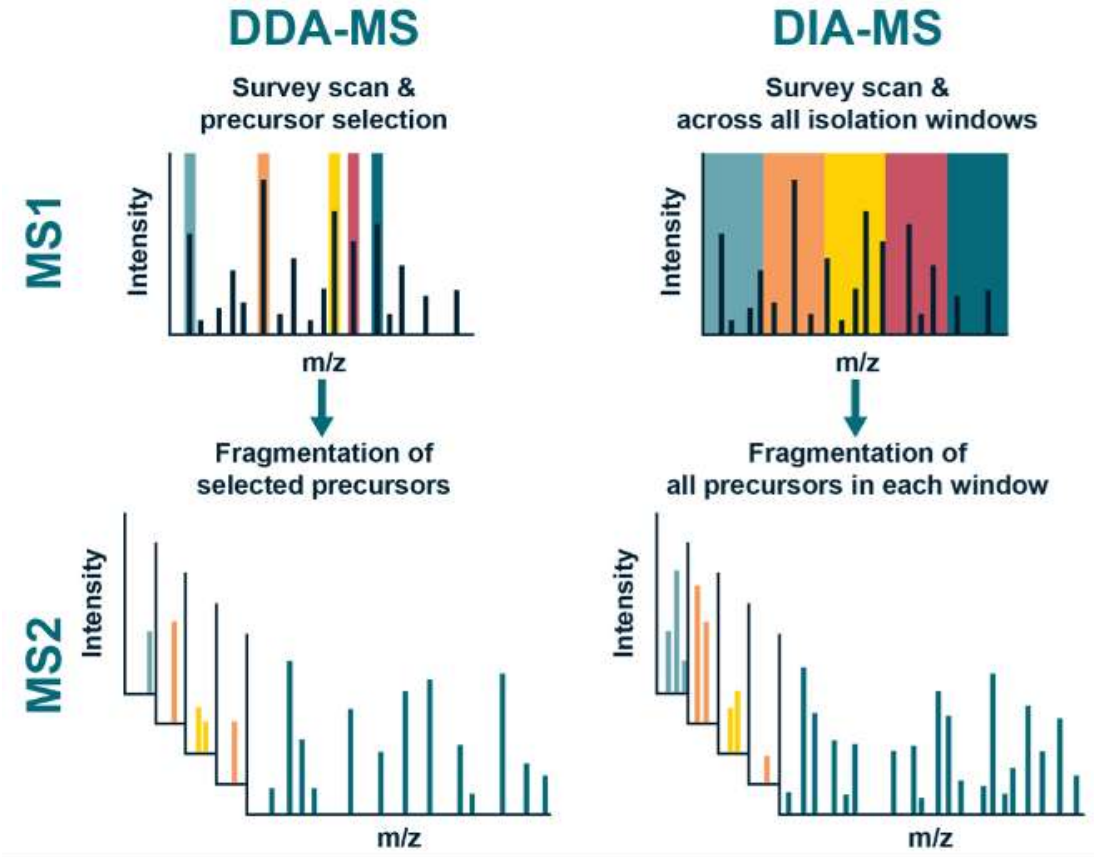
  - **DIA – data independent acquisition**

- **Cílené metody:**
  - **SRM – selected reaction monitoring**

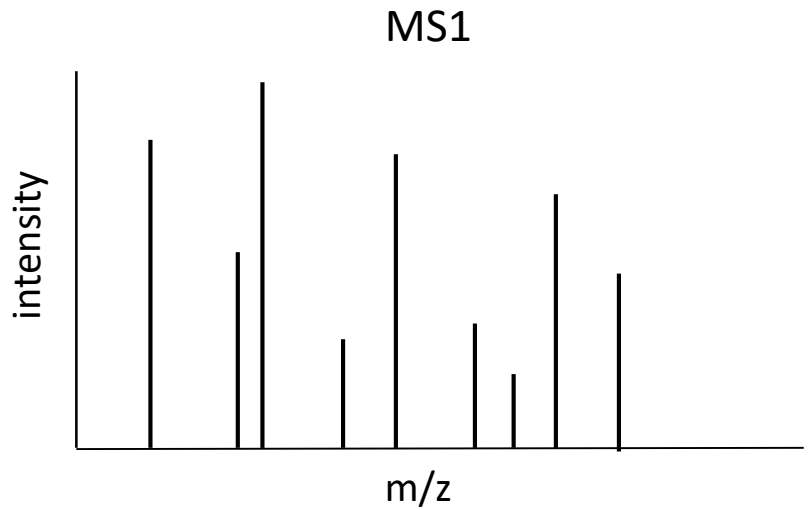
  - **PRM – parallel reaction monitoring**

Akviziční módy

- **MS1 a MS2 (MSMS)** scany poskytují různé druhy informací
- Cyklické střídání MS1 a MS2 scanů podle předem daných pravidel
- Cílem je detekovat co nejvíce peptidů (počet MSMS spekter) s co nejkvalitnější kvantitativní informací (počet bodů přes pík)
- **Necílené akviziční módy:**
  - **DDA – data dependent acquisition**
  - **DIA – data independent acquisition**
- **Cílené metody:**
  - **SRM – selected reaction monitoring**
  - **PRM – parallel reaction monitoring**

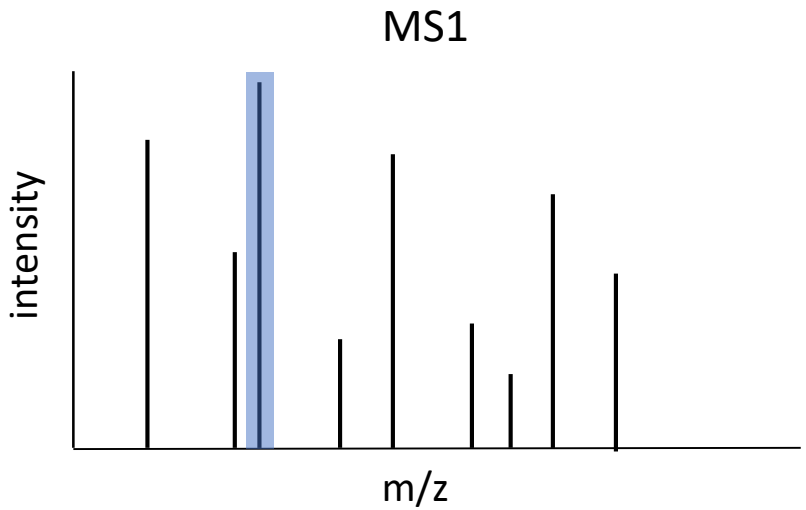


Data Dependent Acquisition



1. MS1 scan (full scan) – detekce prekurzorů s vysokou přesností a rozlišením (300-1600 m/z)

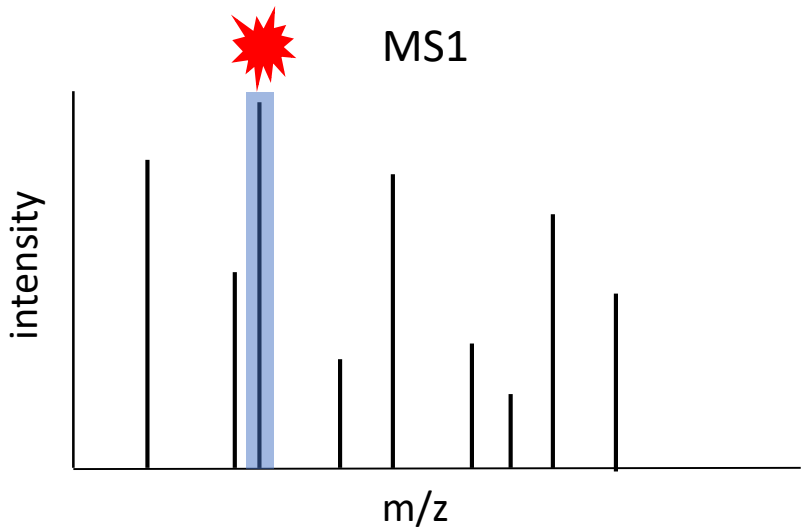
Data Dependent Acquisition



1. MS1 scan (full scan) – detekce prekurzorů s vysokou přesností a rozlišením (300-1600 m/z)

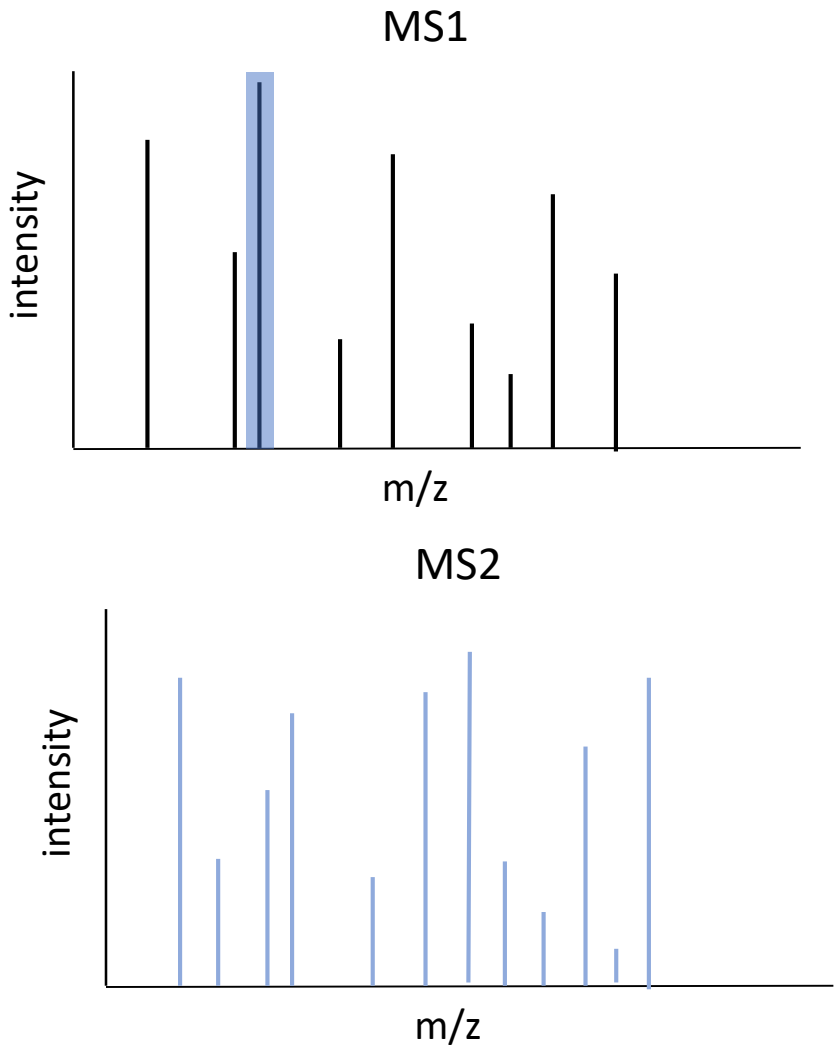
2. Izolace vybraného prekurzoru – kvadrupól, iontová past

Data Dependent Acquisition



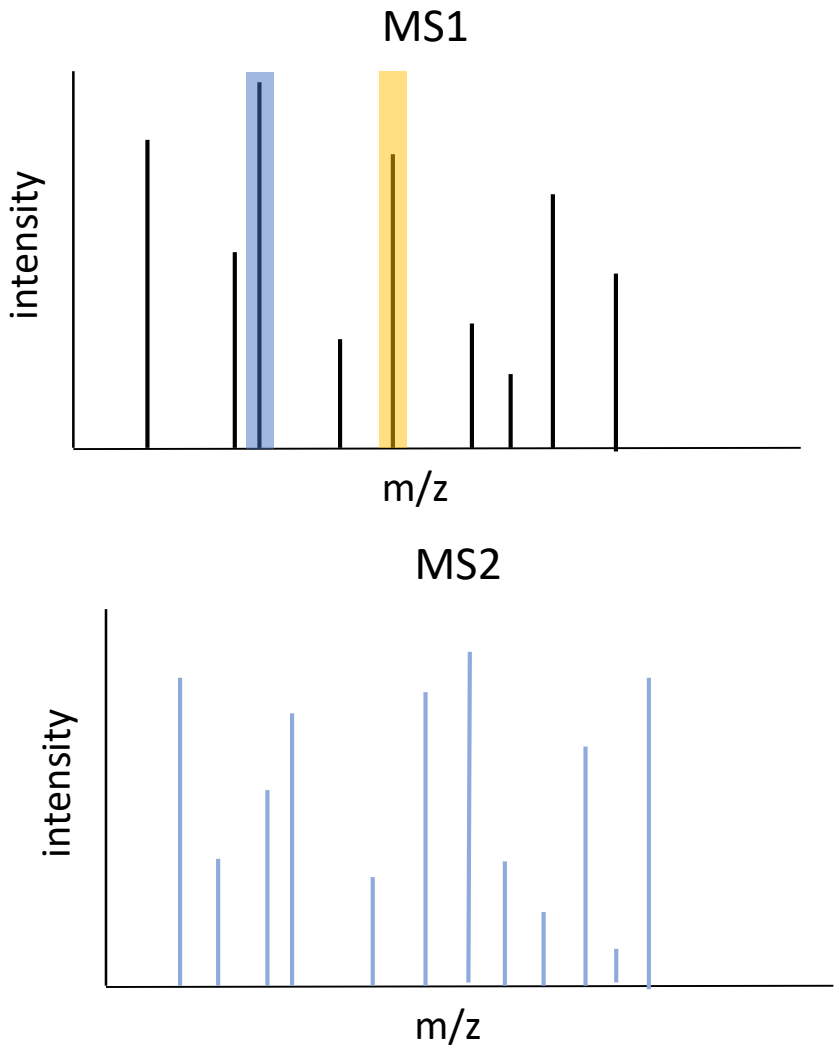
1. MS1 scan (full scan) – detekce prekurzorů s vysokou přesností a rozlišením (300-1600 m/z)
2. Izolace vybraného prekurzoru – kvadrupól, iontová past
3. Fragmentace

Data Dependent Acquisition



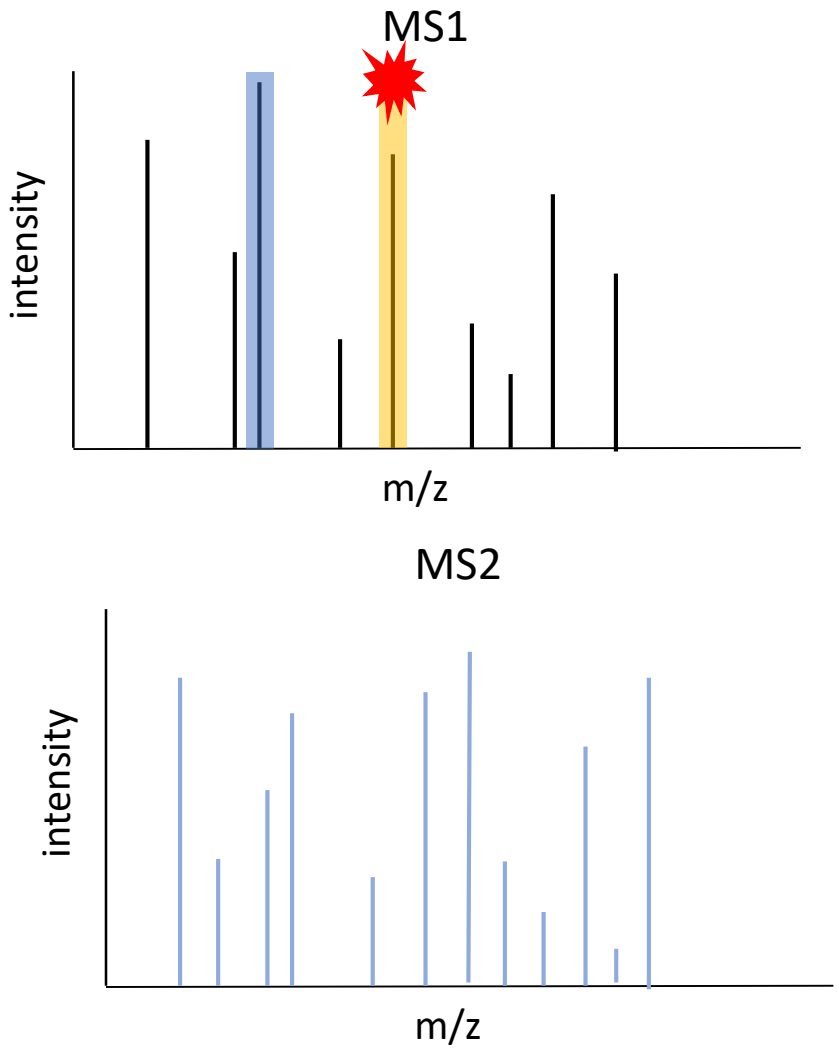
1. MS1 scan (full scan) – detekce prekurzorů s vysokou přesností a rozlišením (300-1600 m/z)
2. Izolace vybraného prekurzoru – kvadrupól, iontová past
3. Fragmentace
4. Detekce vzniklých fragmentů

Data Dependent Acquisition



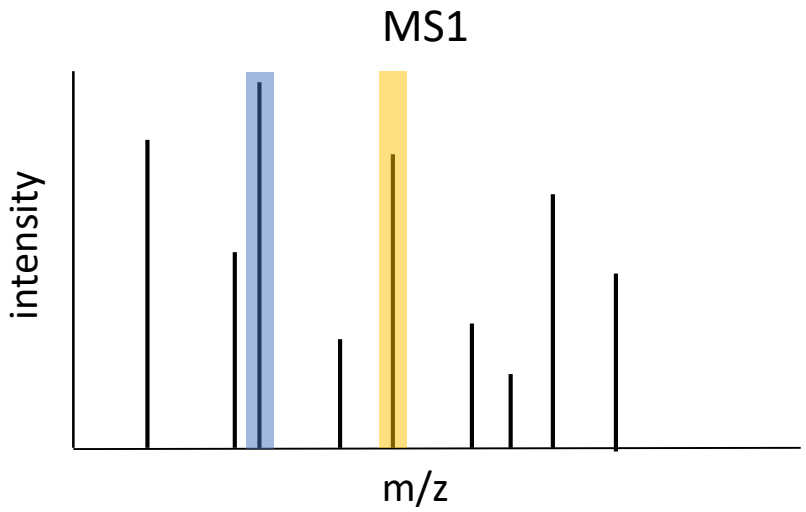
1. MS1 scan (full scan) – detekce prekurzorů s vysokou přesností a rozlišením (300-1600 m/z)
2. Izolace vybraného prekurzoru – kvadrupól, iontová past
3. Fragmentace
4. Detekce vzniklých fragmentů

Data Dependent Acquisition

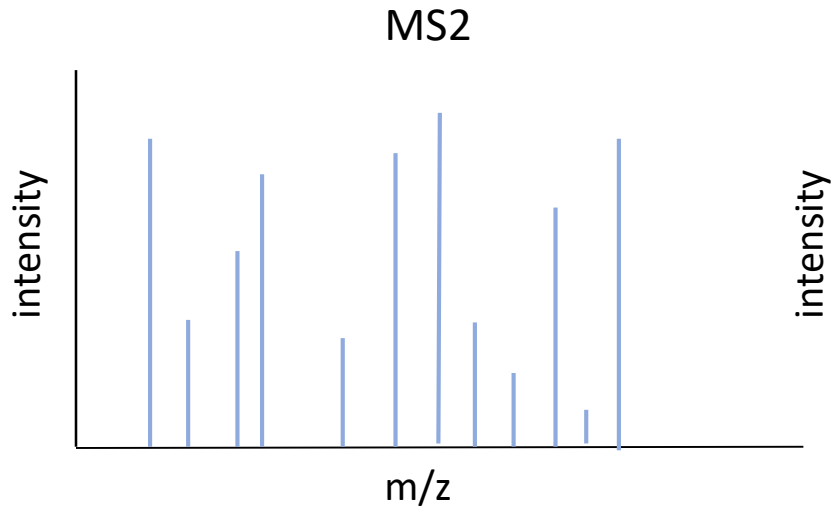


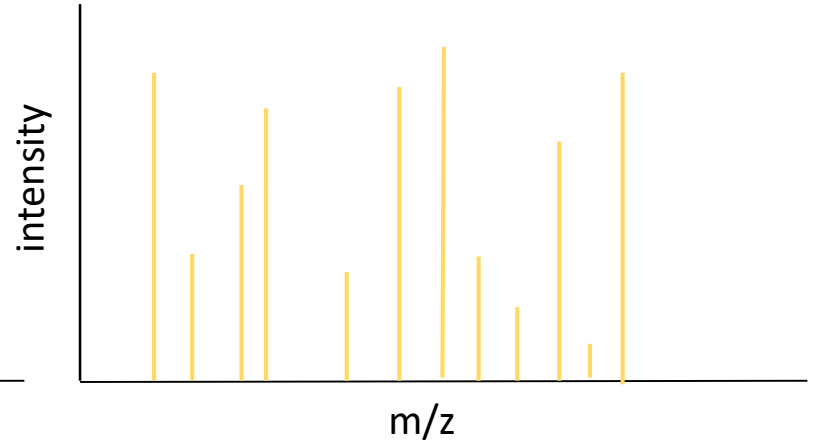
1. MS1 scan (full scan) – detekce prekurzorů s vysokou přesností a rozlišením (300-1600 m/z)
2. Izolace vybraného prekurzoru – kvadrupól, iontová past
3. Fragmentace
4. Detekce vzniklých fragmentů

Data Dependent Acquisition

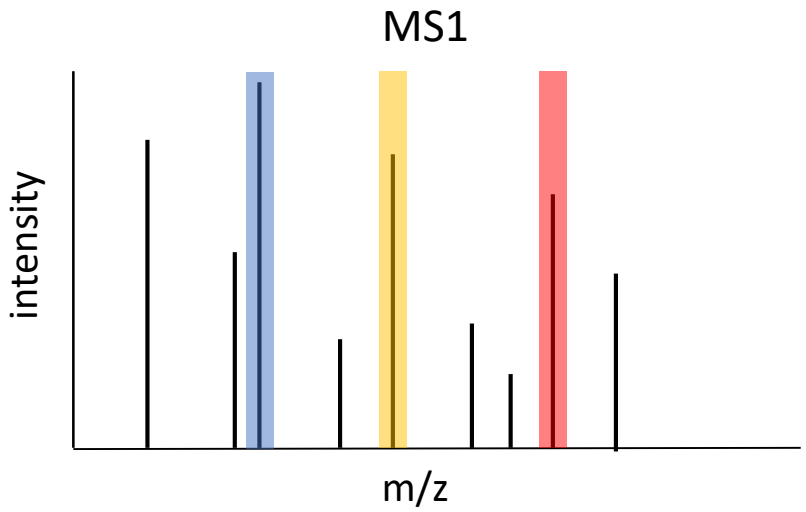


1. MS1 scan (full scan) – detekce prekurzorů s vysokou přesností a rozlišením (300-1600 m/z)
2. Izolace vybraného prekurzoru – kvadrupól, iontová past
3. Fragmentace
4. Detekce vzniklých fragmentů

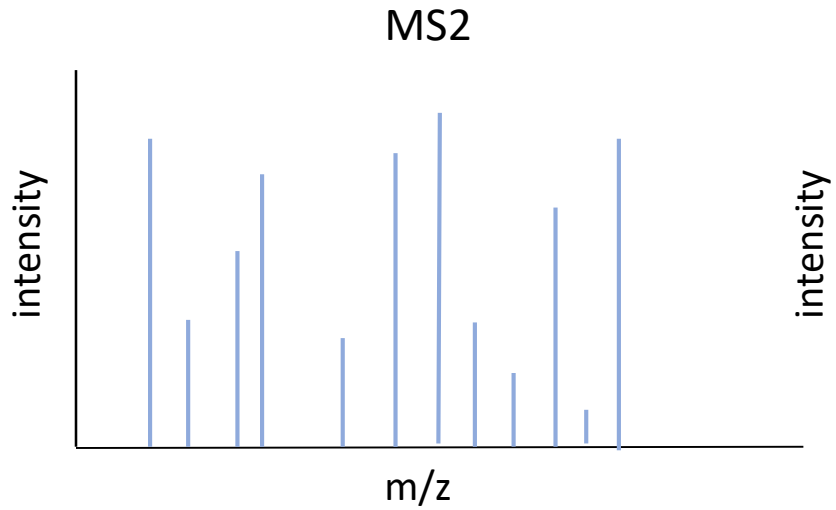


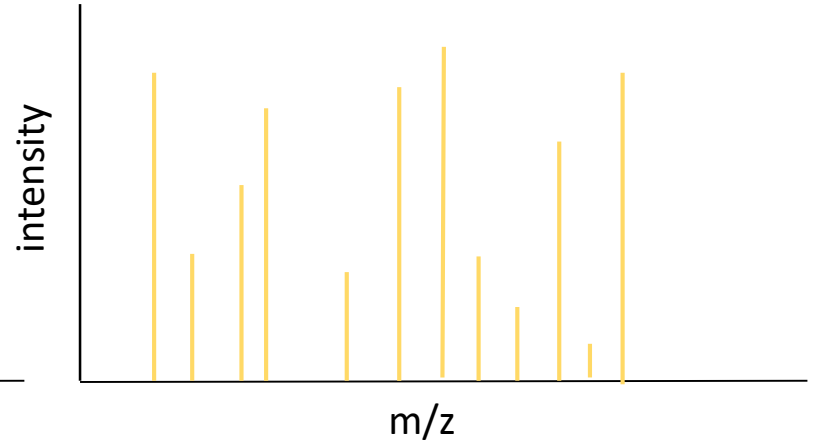


Data Dependent Acquisition

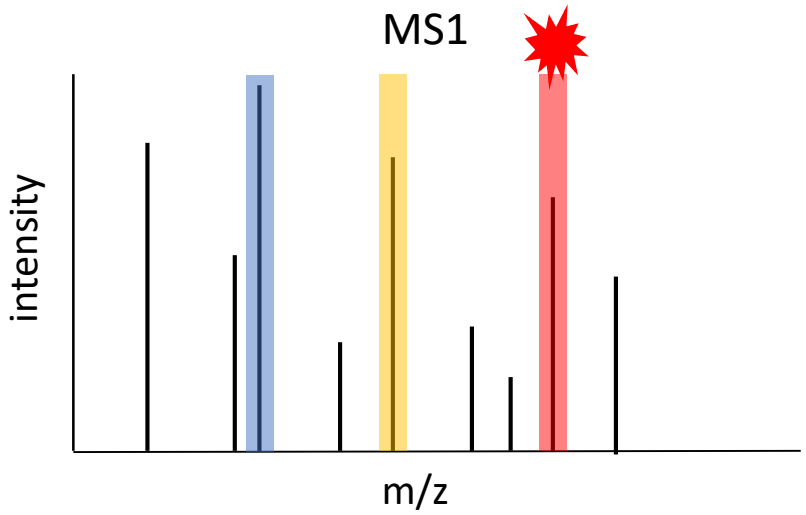


1. MS1 scan (full scan) – detekce prekurzorů s vysokou přesností a rozlišením (300-1600 m/z)
2. Izolace vybraného prekurzoru – kvadrupól, iontová past
3. Fragmentace
4. Detekce vzniklých fragmentů

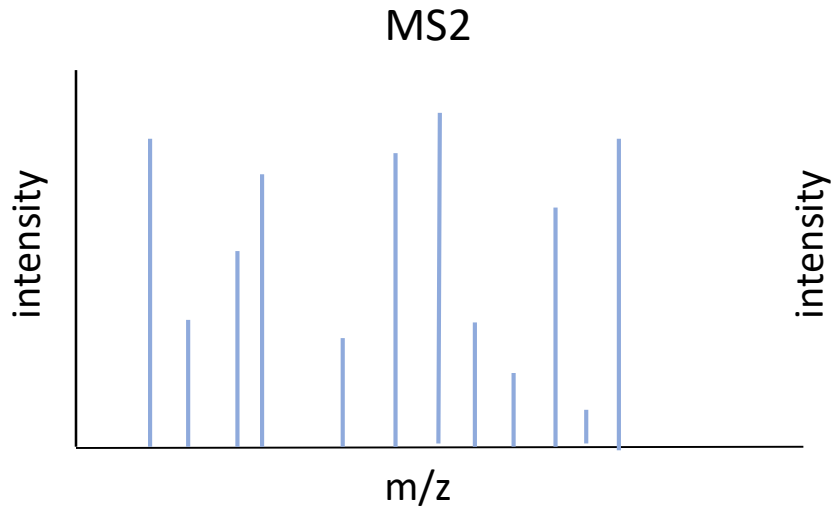


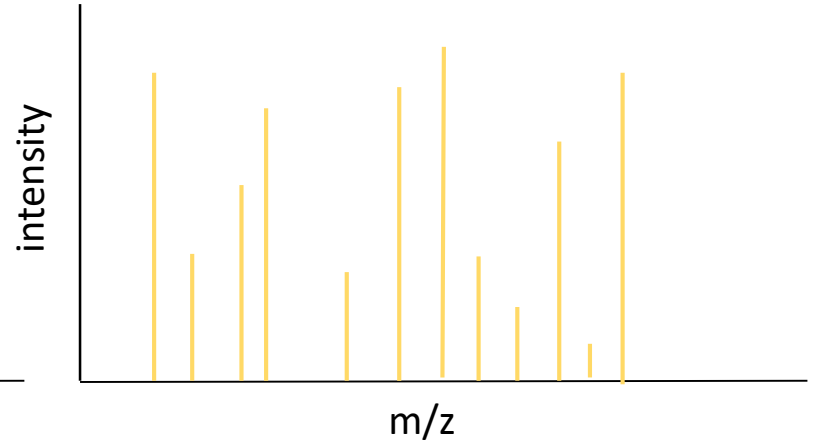


Data Dependent Acquisition

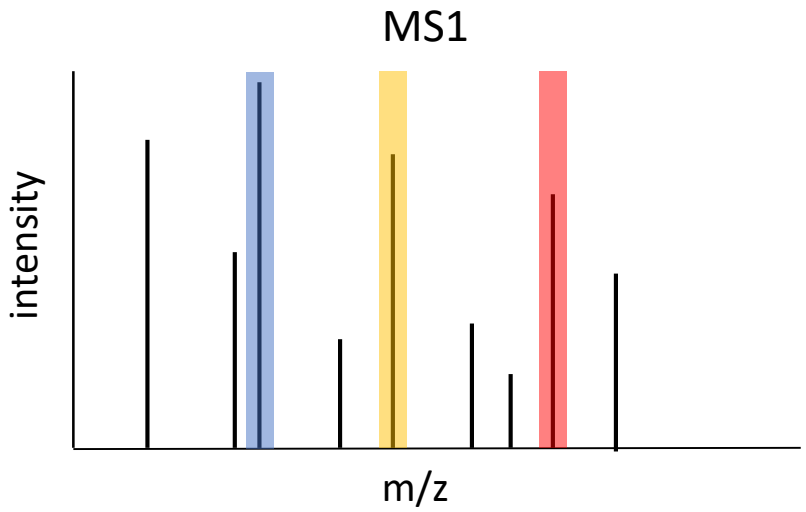


1. MS1 scan (full scan) – detekce prekurzorů s vysokou přesností a rozlišením (300-1600 m/z)
2. Izolace vybraného prekurzoru – kvadrupól, iontová past
3. Fragmentace
4. Detekce vzniklých fragmentů

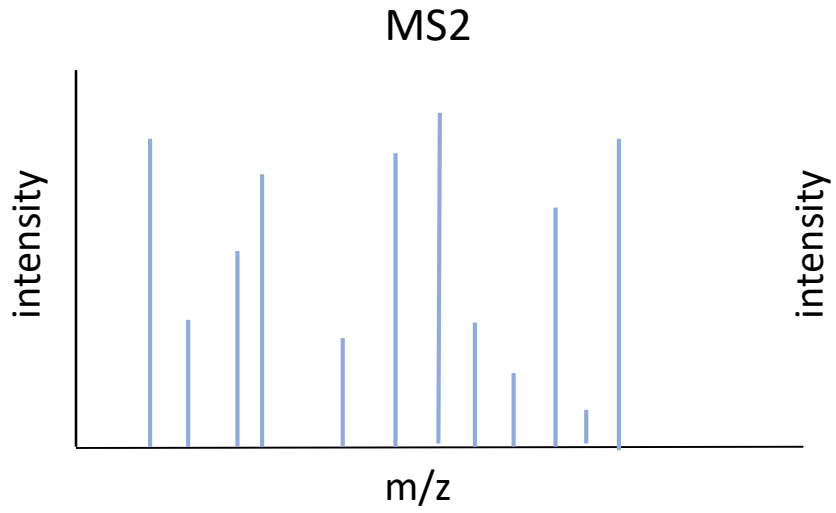


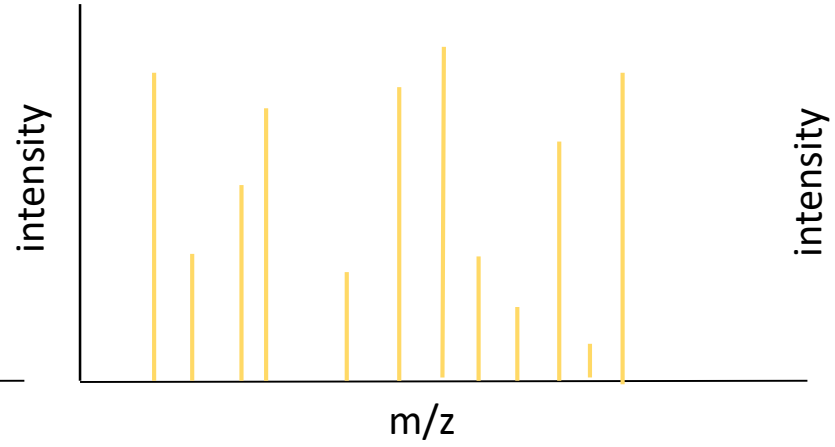


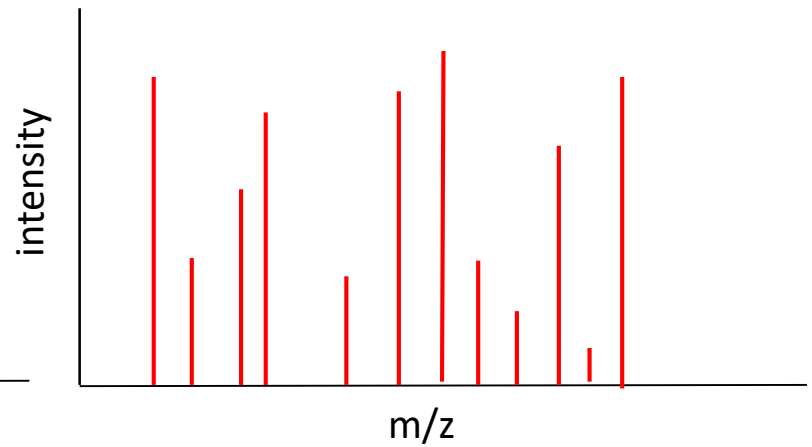
Data Dependent Acquisition



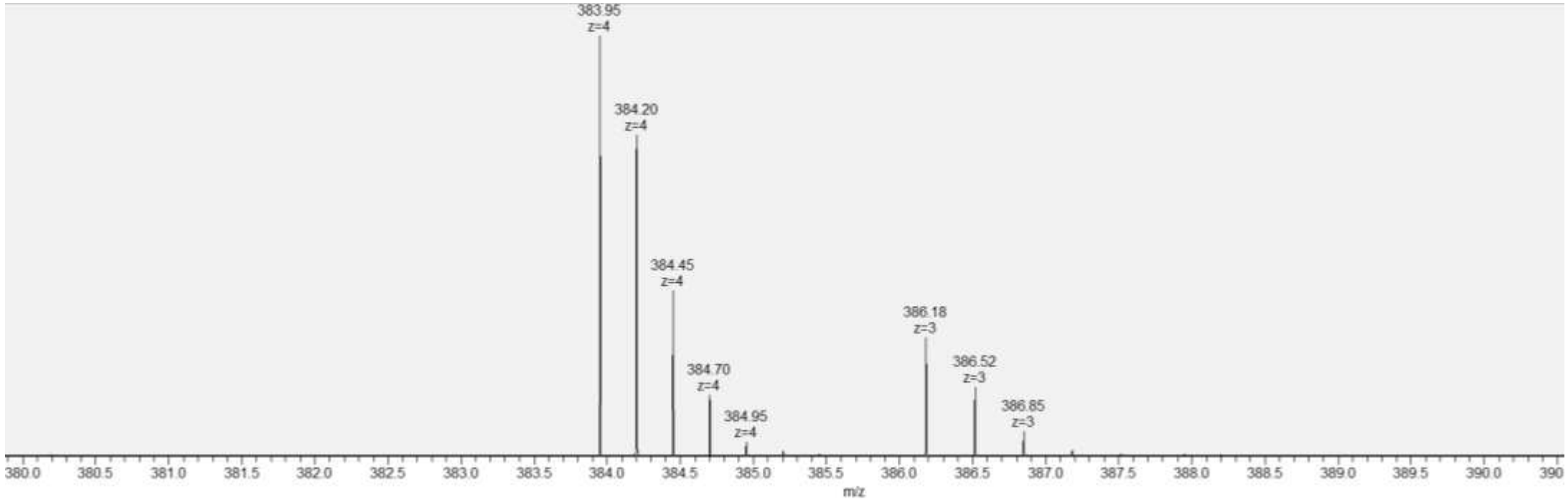
1. MS1 scan (full scan) – detekce prekurzorů s vysokou přesností a rozlišením (300-1600 m/z)
2. Izolace vybraného prekurzoru – kvadrupól, iontová past
3. Fragmentace
4. Detekce vzniklých fragmentů



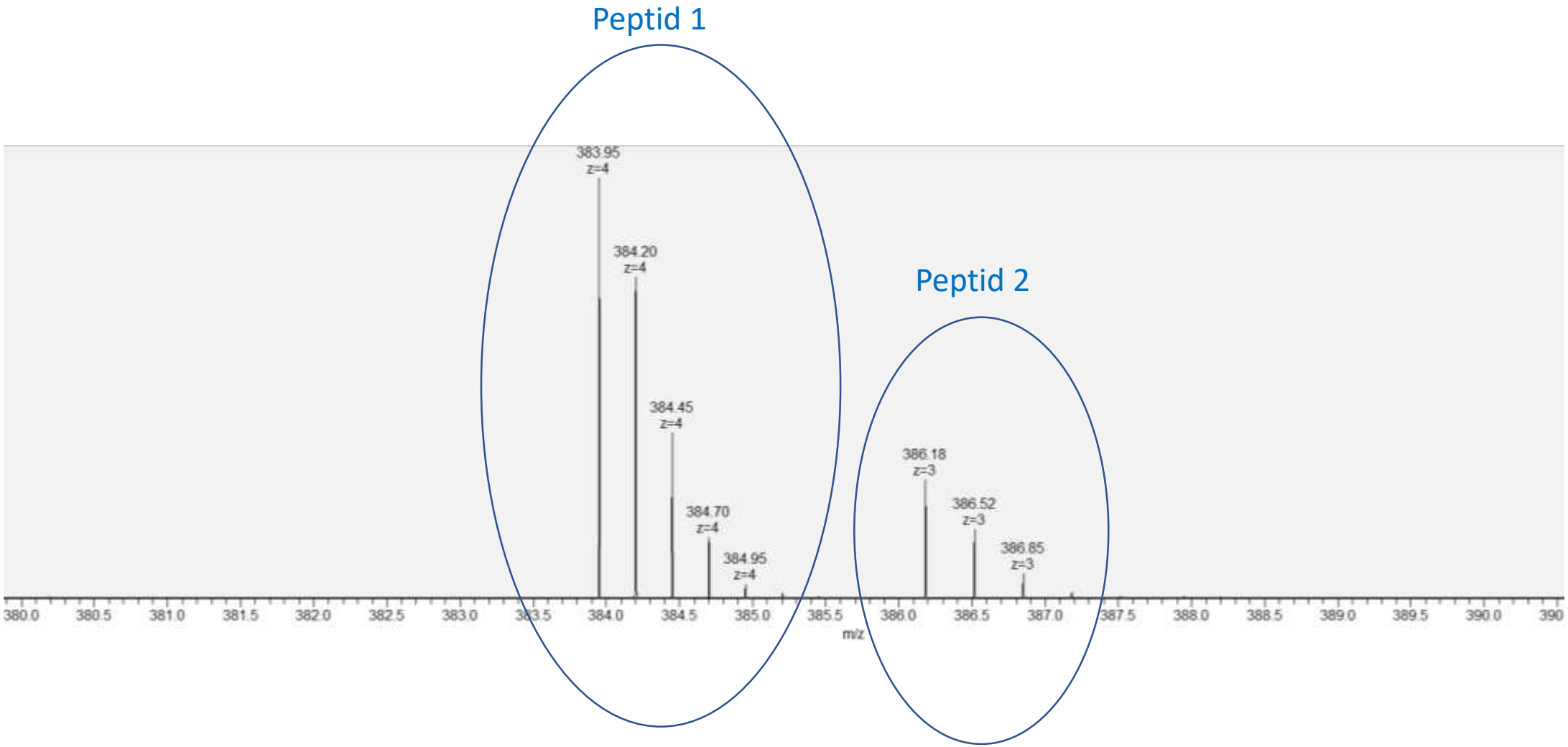




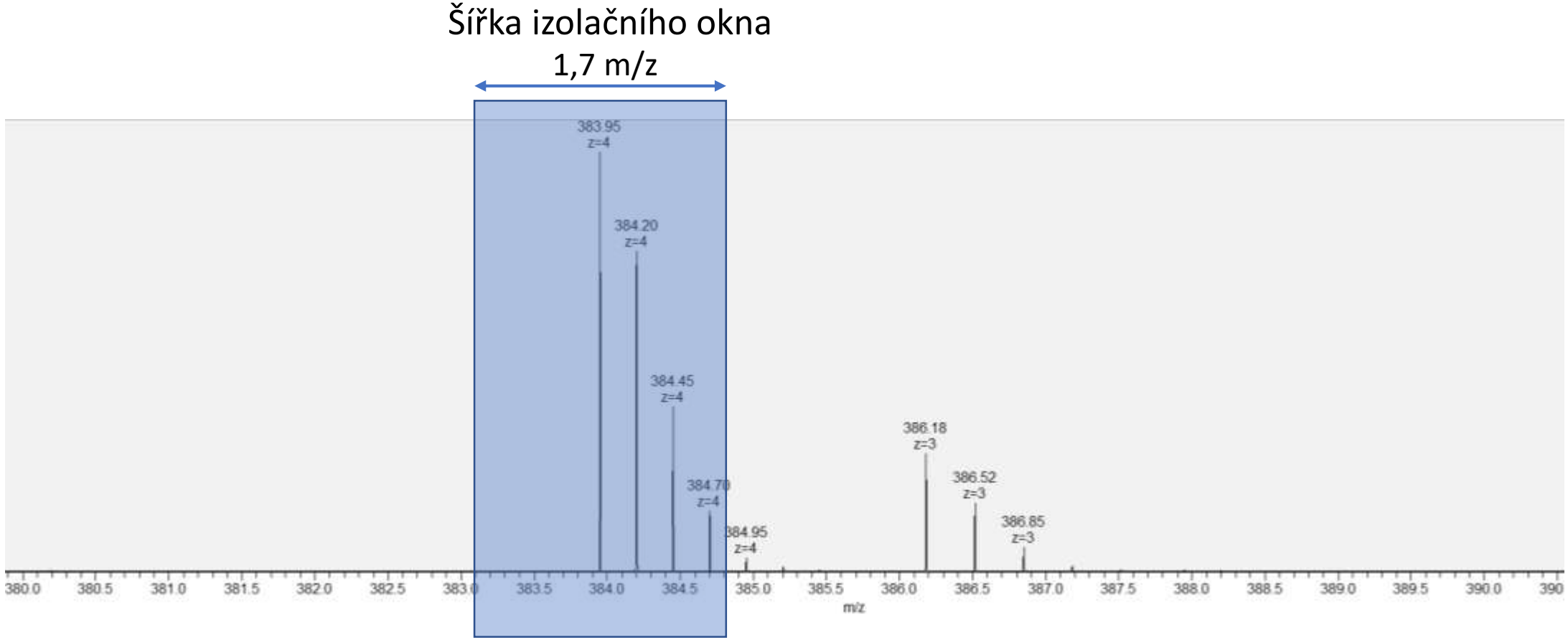
Izolace peptidu



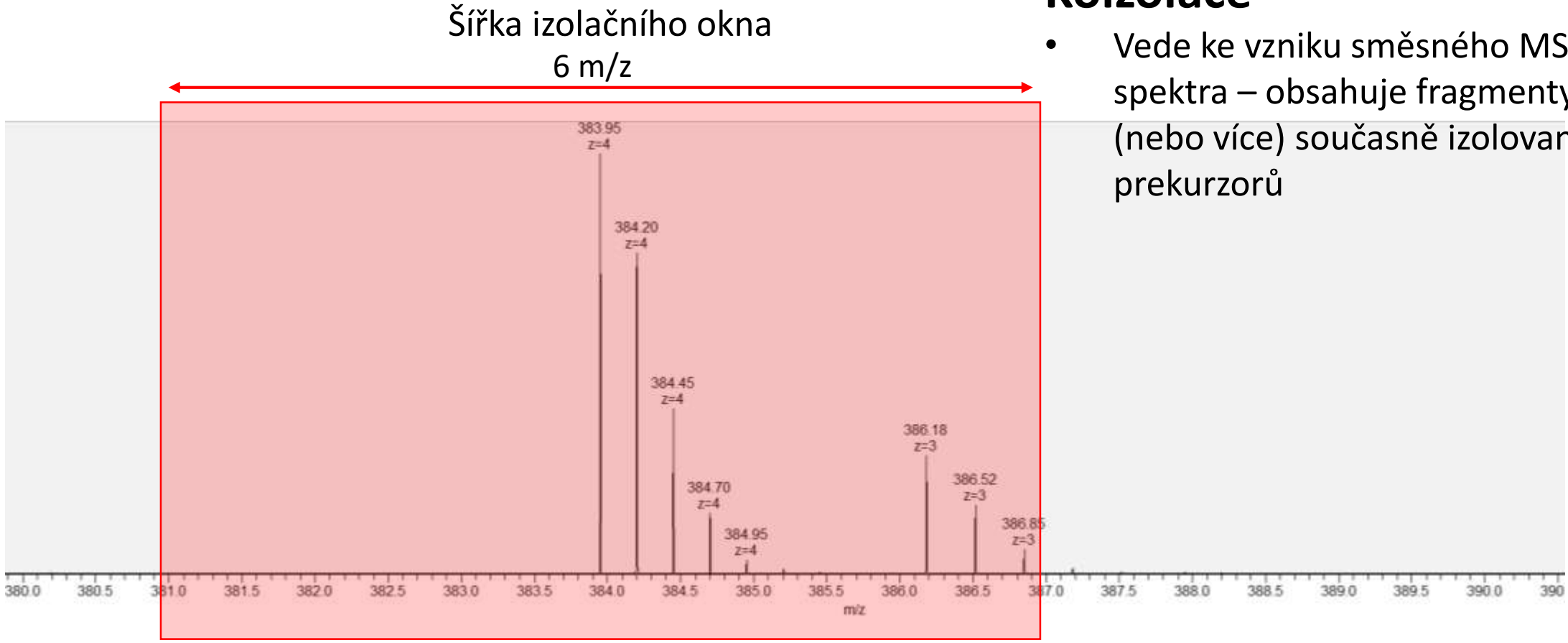
Izolace peptidu



Izolace peptidu



Izolace peptidu

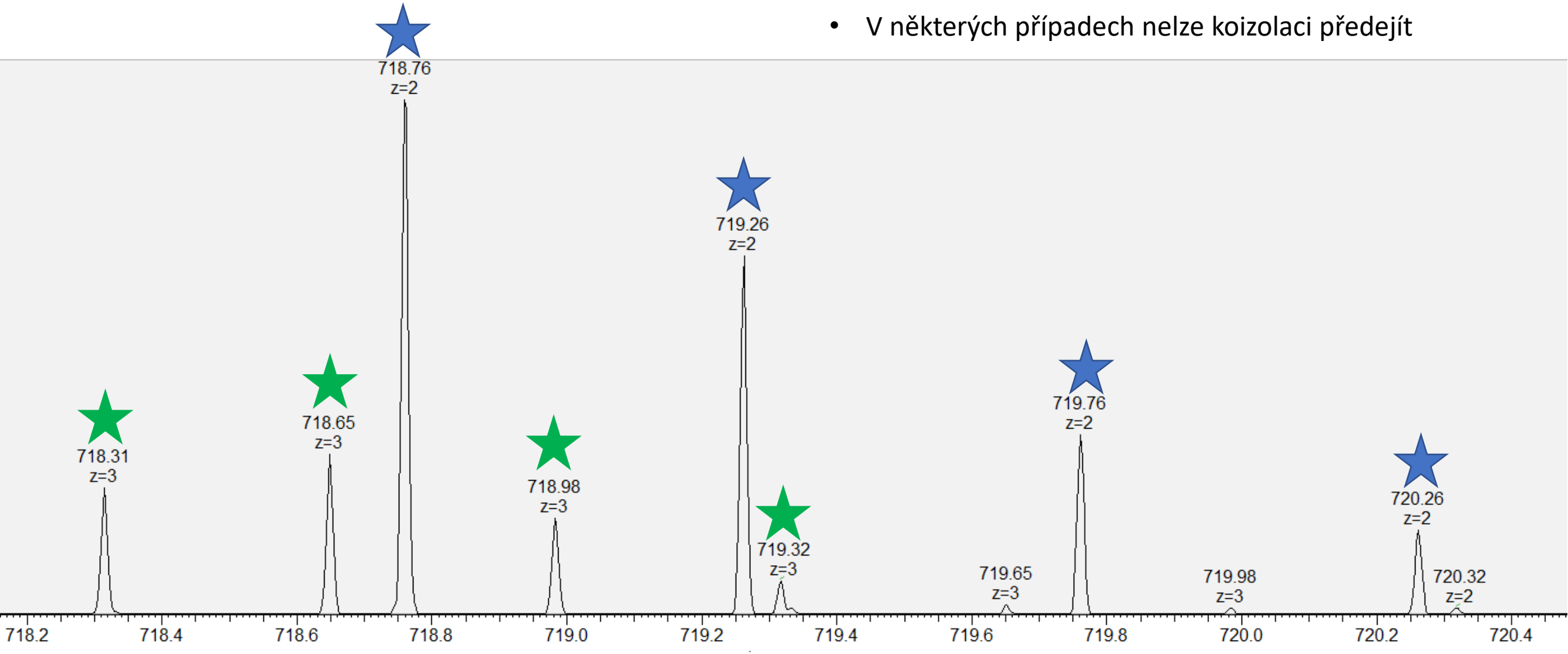


Koizolace

- Vede ke vzniku směsného MSMS spektra – obsahuje fragmenty dvou (nebo více) současně izolovaných prekurzorů

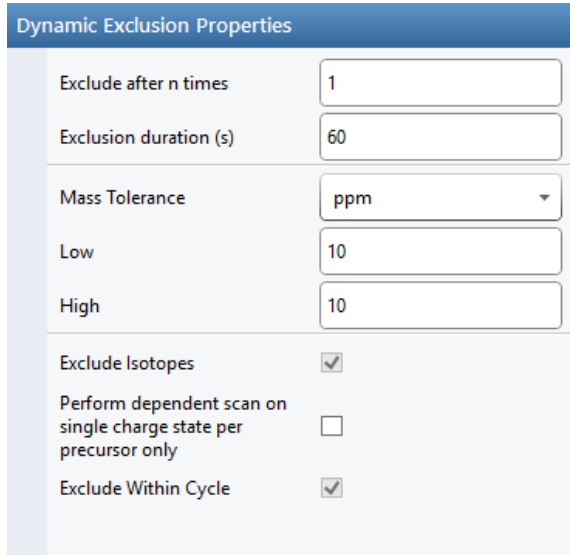
Izolace peptidu

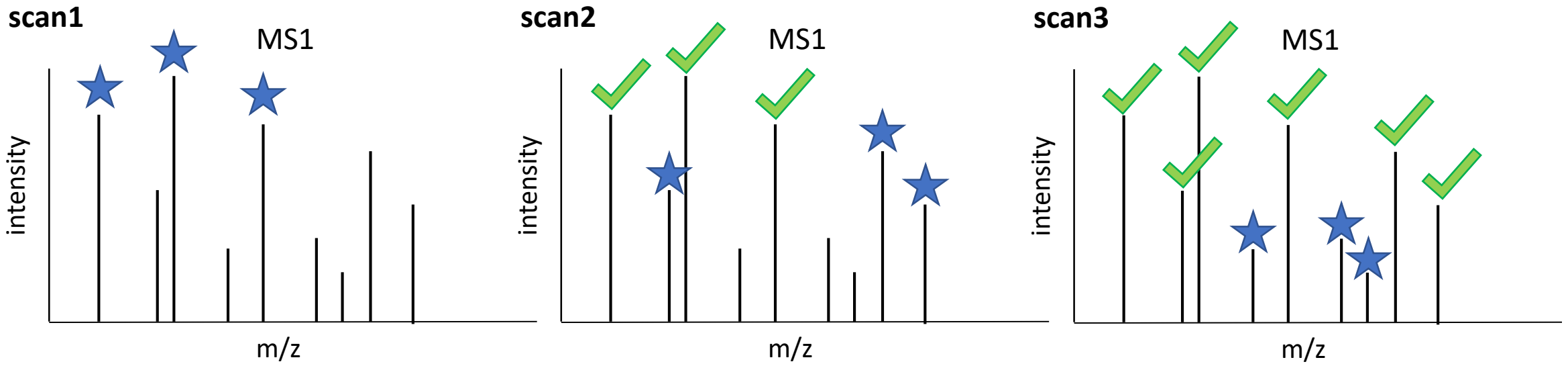
- V některých případech nelze koizolaci předejít



Dynamická exkluze

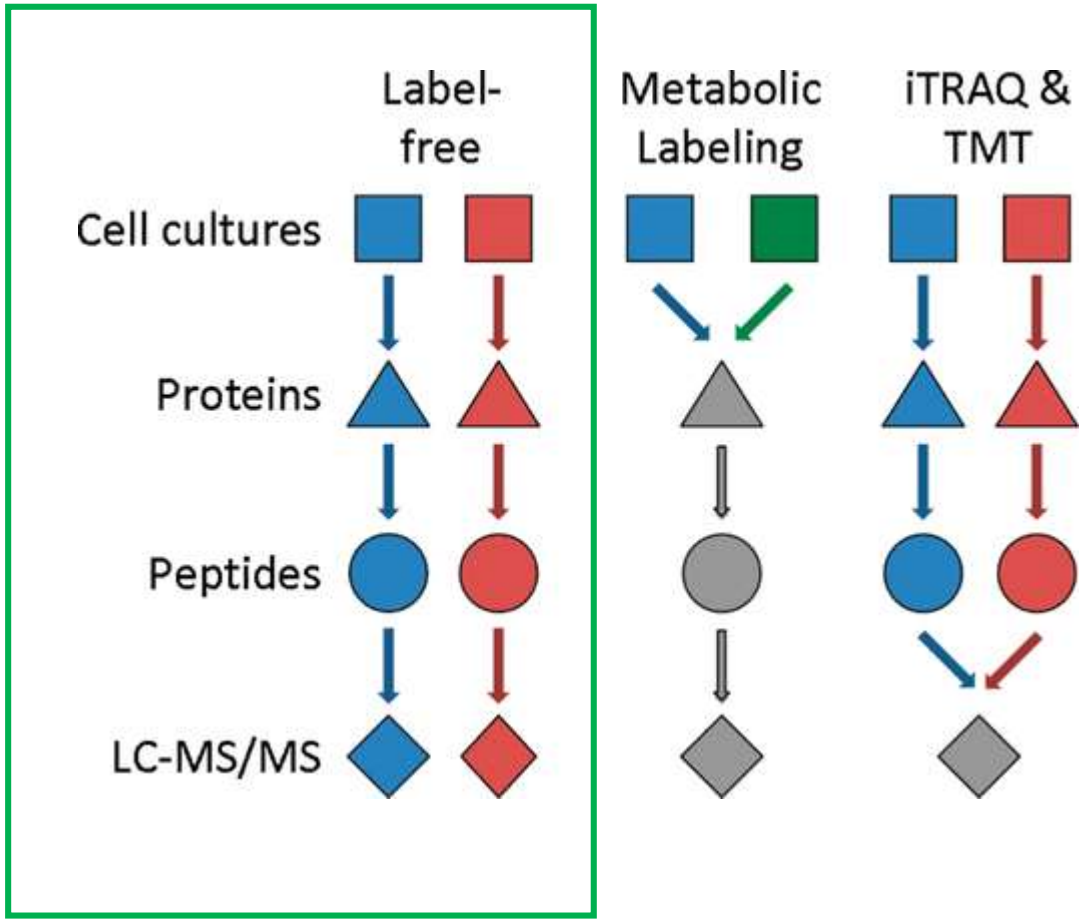
- Obvykle vybíráme pro fragmentaci několik nejintenzivnějších peptidů ve spektru
- Dynamická exkluze předchází opakované fragmentaci stejných peptidů
- Fragmentovaný peptid je po určitý čas vyloučen z výběru pro fragmentaci





Label Free Quantification

- Pro kvantifikaci se využívá výhradně signál samotného peptidu (výška píku/plocha pod píkem)
- Není třeba žádného dodatečného značení na úrovni proteinu/peptidu (SILAC, isobarické značení – viz. následující přednáška)
- Není třeba investovat do metabolických nebo isobarických značek
- Přesnost kvantifikace nižší než u značených přístupů
- Nelze brát v potaz menší kvantitativní rozdíly (10% není z pohledu label free přístupu žádný rozdíl)



Li et al, Systematic Comparison of Label-Free, Metabolic Labeling, and Isobaric Chemical Labeling for Quantitative Proteomics on LTQ Orbitrap Velos, J Prot Res, 2012

Label Free Quantification

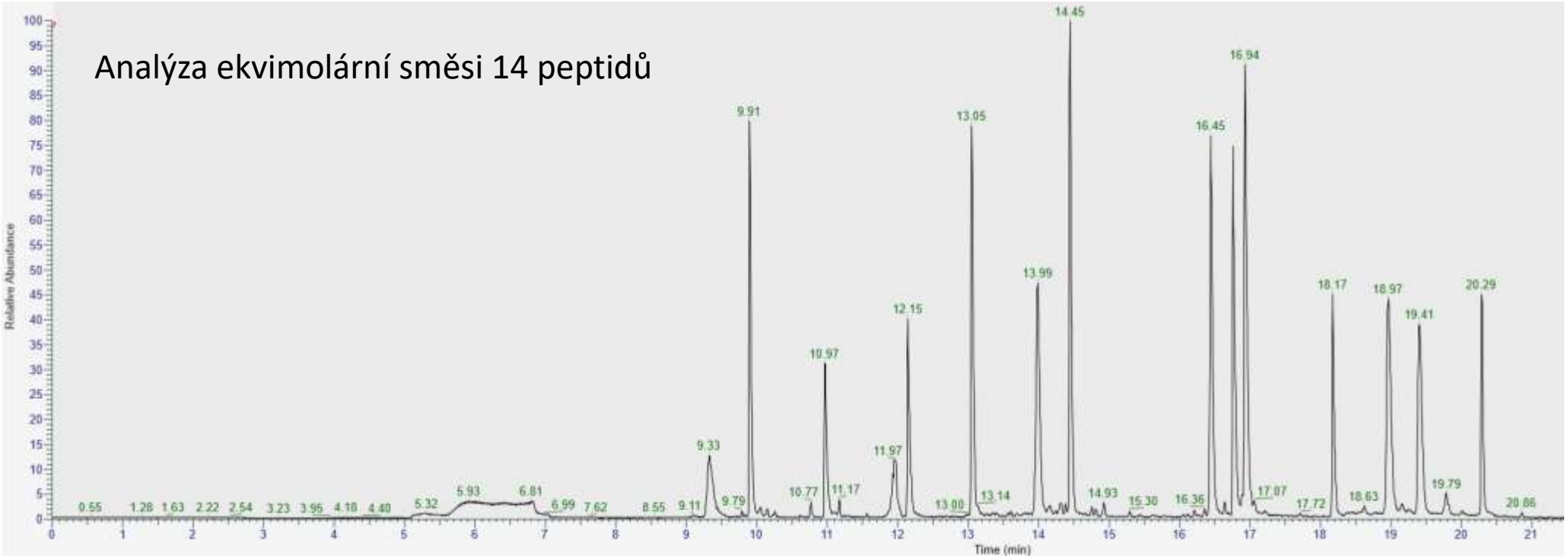
- **MS/MS count (spectral count)**
  - Počet fragmentačních spekter úměrný množství peptidu.
  - Používáno v minulosti

- **Intenzita**
  - Výška píku (plocha pod píkem). Intenzita proteinu daná součtem intenzit peptidů.

- **iBAQ (intensity based absloute quantification)**
  - Suma intenzit peptidů vztažená na počet všech teoretických peptidů (např. tryptických)
  - Vhodné na porovnání intenzit proteinů v rámci jednoho vzorku.

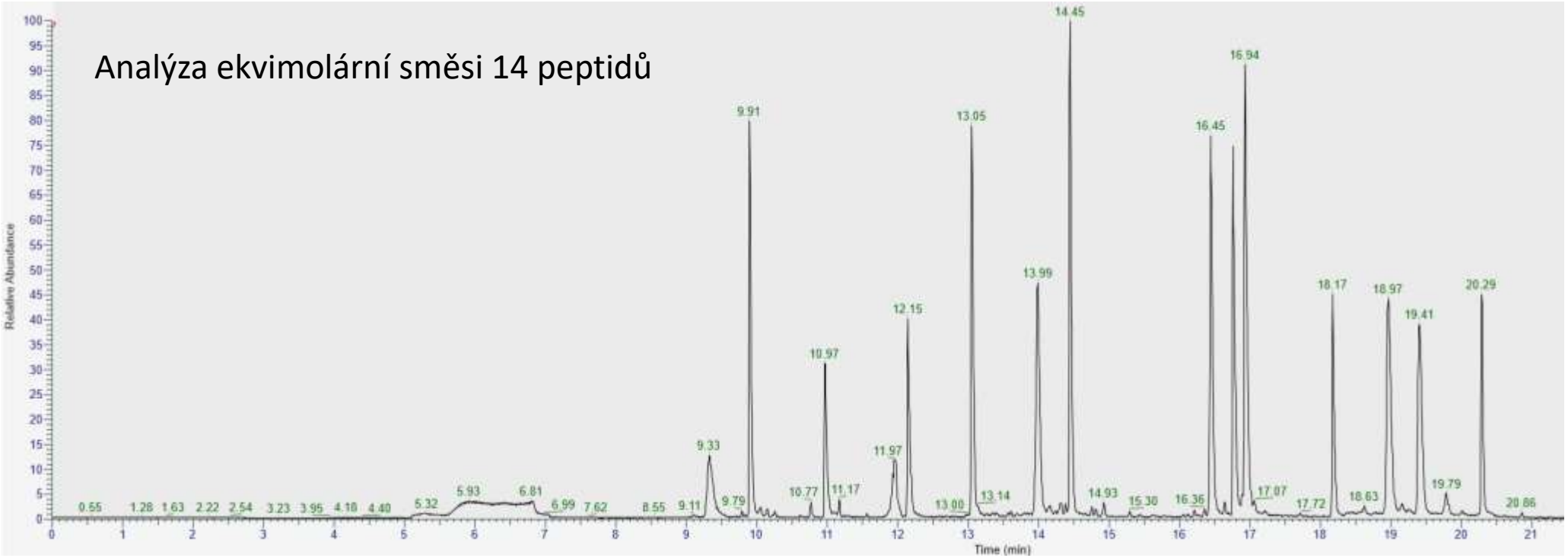
Kvantifikace pomocí MS

- Samotná MS analýza nemusí být kvantitativní
- Nejintenzivnější signál ve spektru nemusí vždy patřit nejvíce zastoupené látce
- Ovlivněno: schopností ionizace, iontovou supresí, stabilitou iontu ve vakuu, matričními efekty, adhezí na povrchy



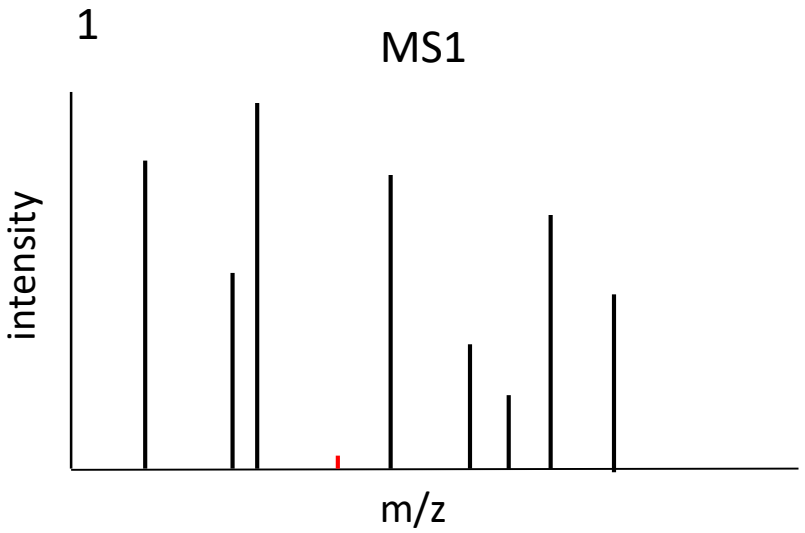
Kvantifikace pomocí MS

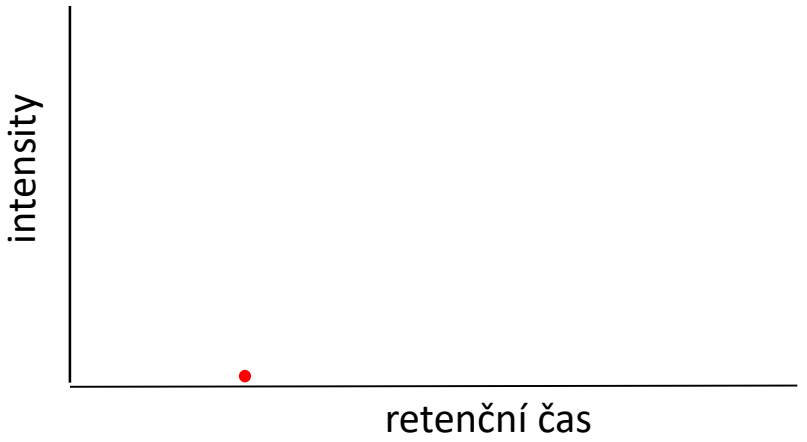
- **Pro kvantitativní porovnání musí být vzorky:**
  - připravované identicky (ideálně jedním člověkem)
  - měřeny ve stejný čas, stejnou metodou
  - reproducibilní chromatografie (stabilní gradient, dostatečná ekvilibrace)



Kde se u DDA bere kvantitativní informace?

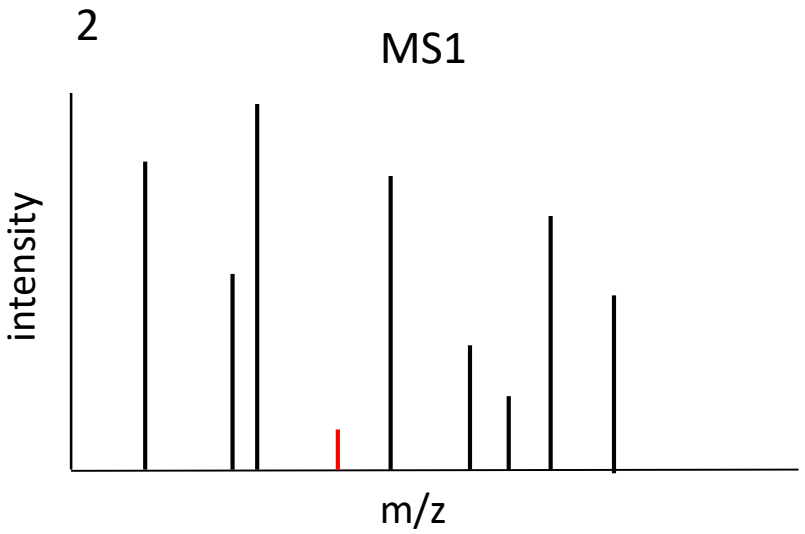
Extracted Ion Chromatogram – XIC

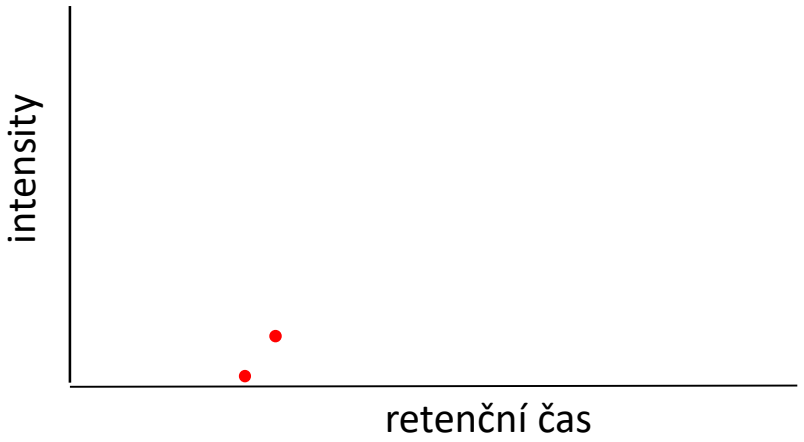




Kde se u DDA bere kvantitativní informace?

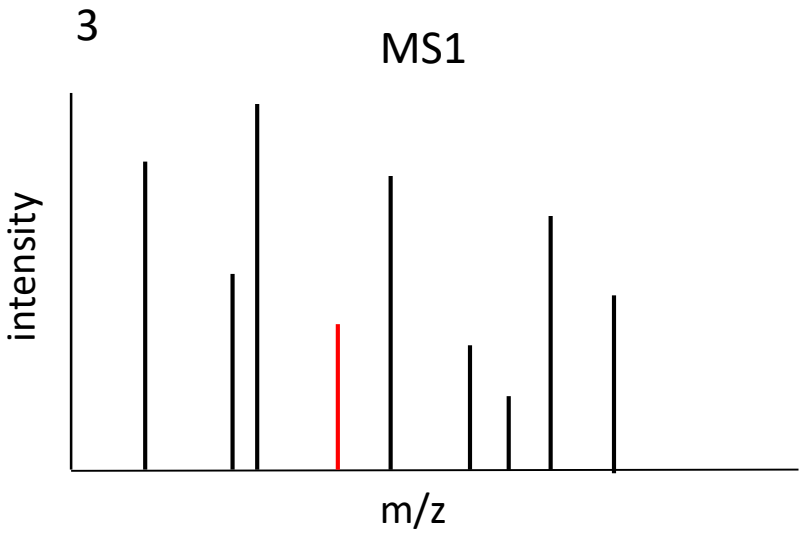
Extracted Ion Chromatogram – XIC

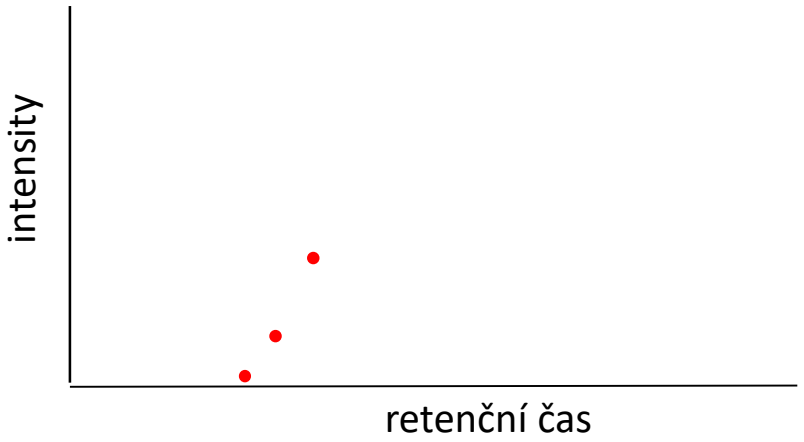




Kde se u DDA bere kvantitativní informace?

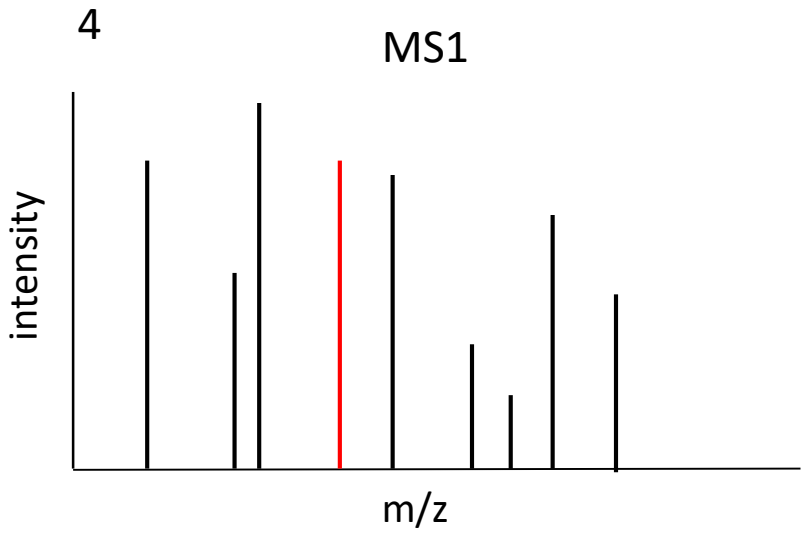
Extracted Ion Chromatogram – XIC

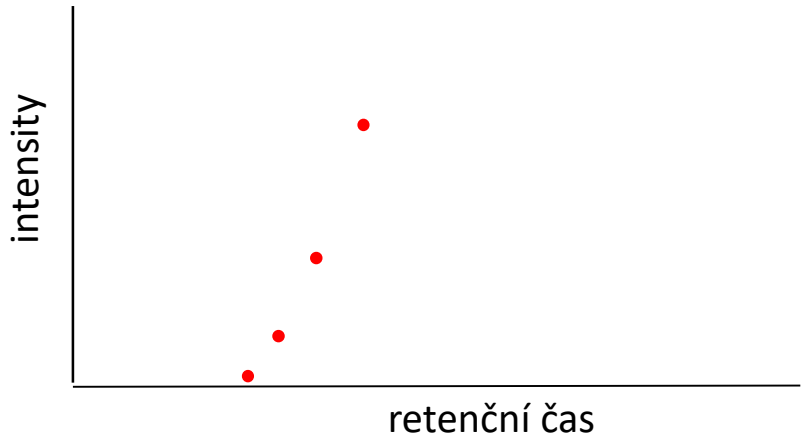




Kde se u DDA bere kvantitativní informace?

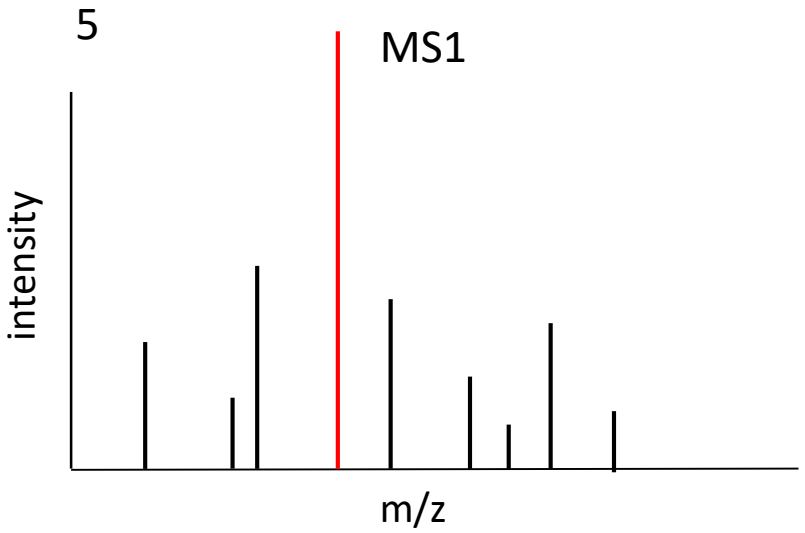
Extracted Ion Chromatogram – XIC

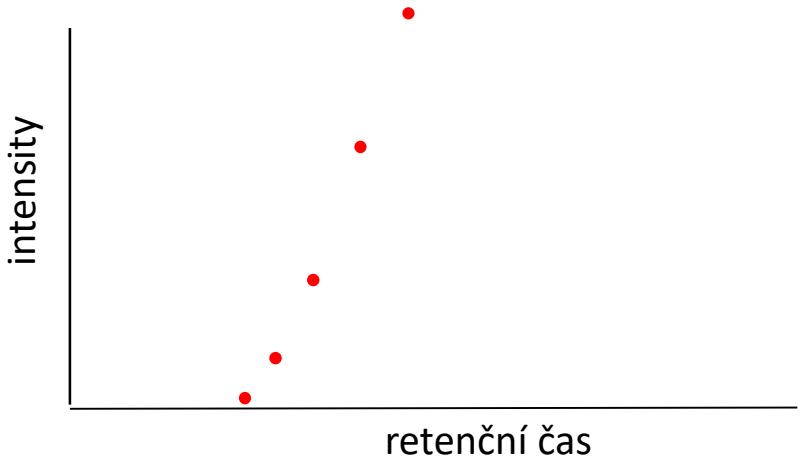




Kde se u DDA bere kvantitativní informace?

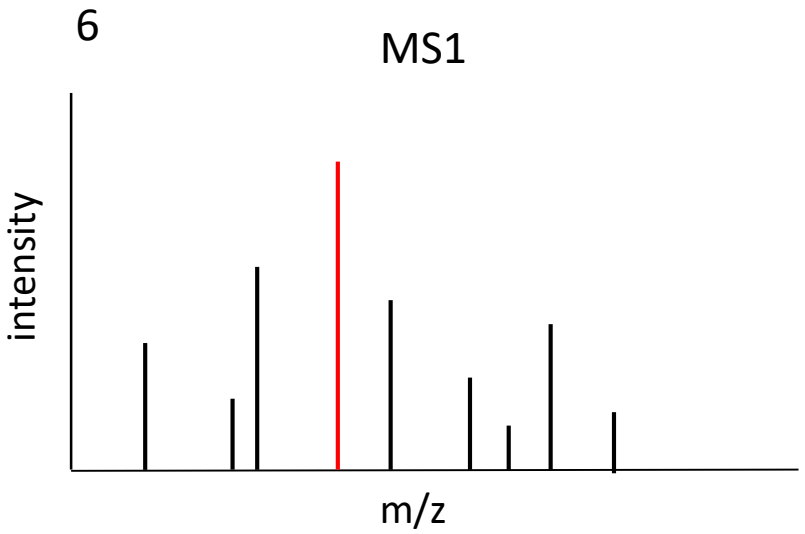
Extracted Ion Chromatogram – XIC

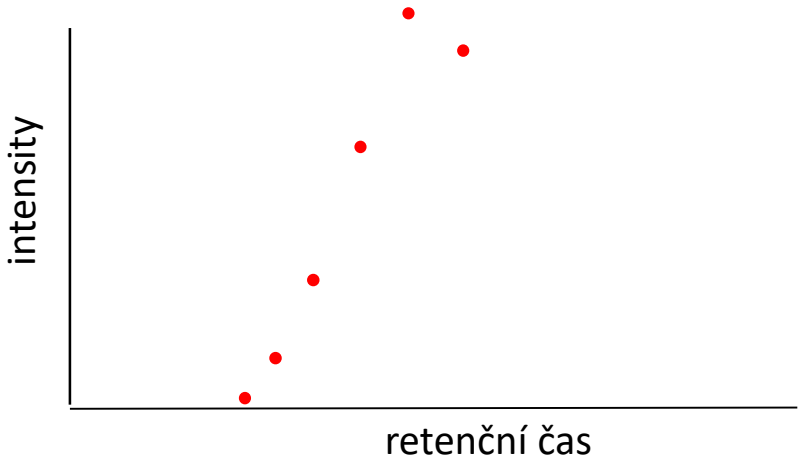




Kde se u DDA bere kvantitativní informace?

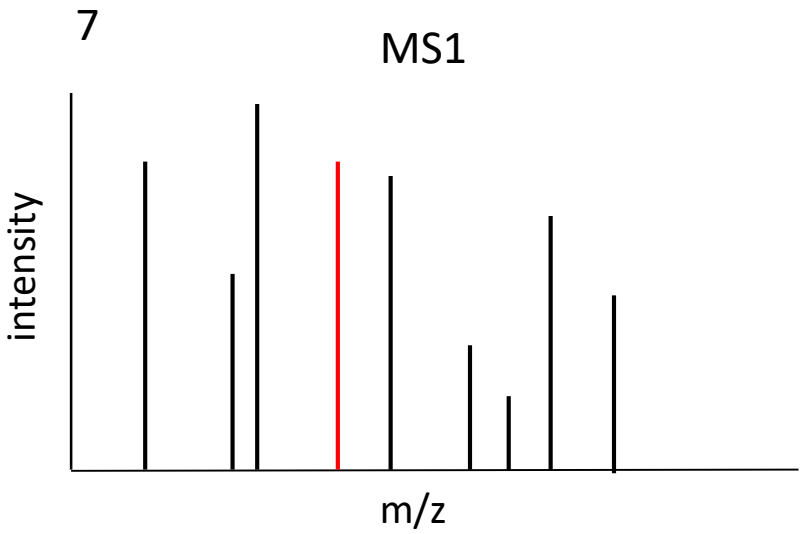
Extracted Ion Chromatogram – XIC

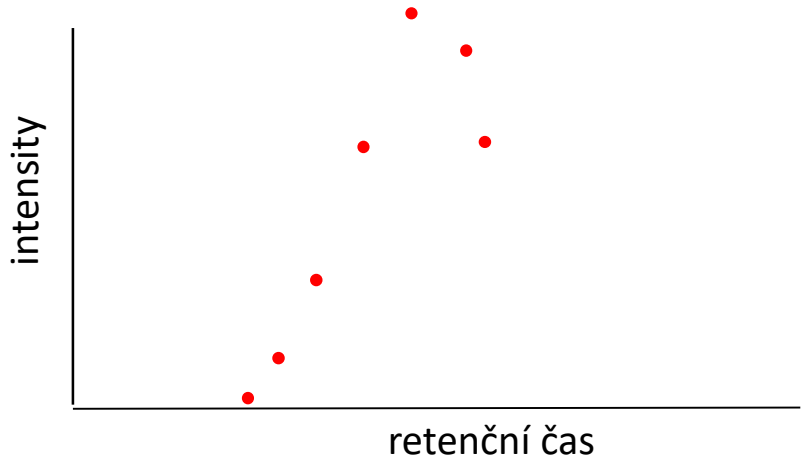




Kde se u DDA bere kvantitativní informace?

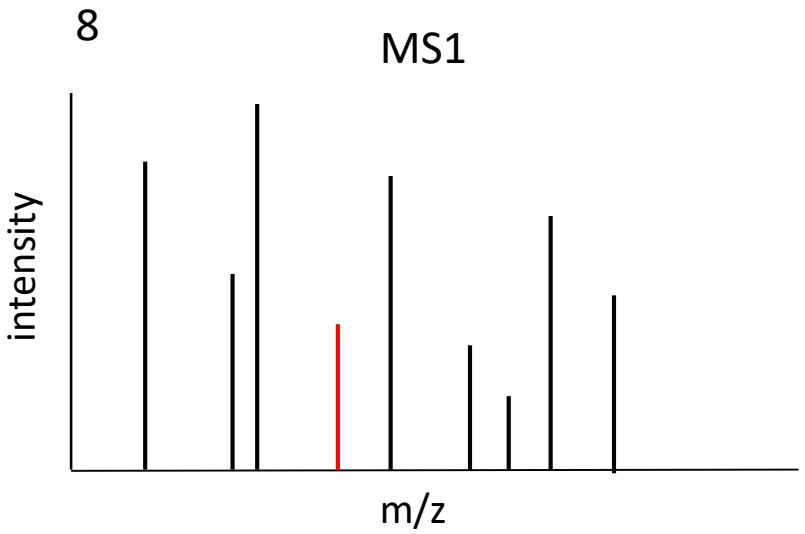
Extracted Ion Chromatogram – XIC

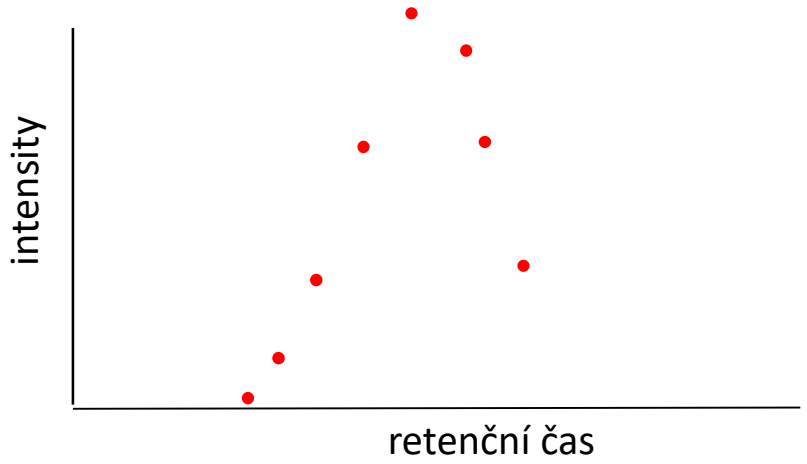




Kde se u DDA bere kvantitativní informace?

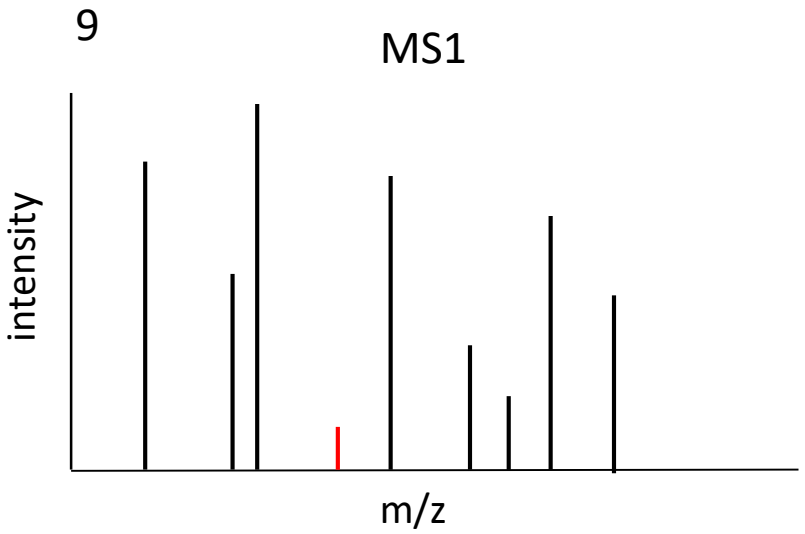
Extracted Ion Chromatogram – XIC

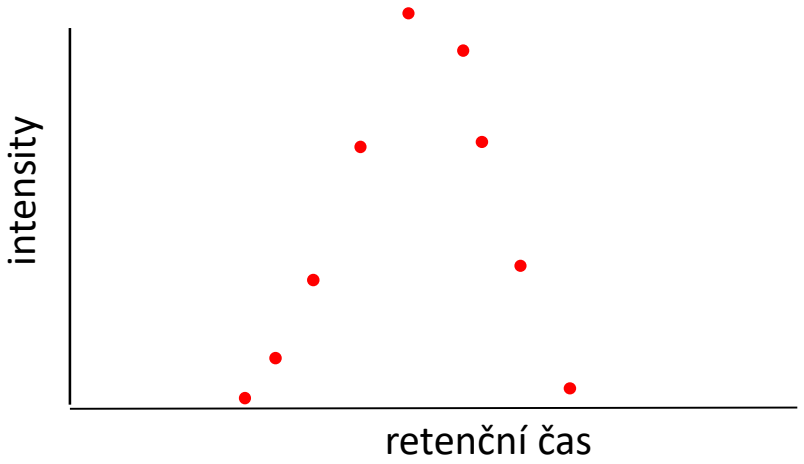




Kde se u DDA bere kvantitativní informace?

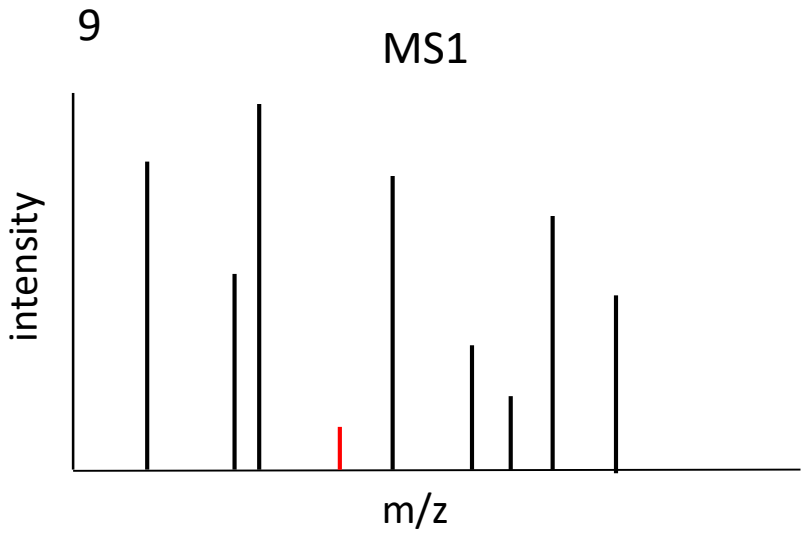
Extracted Ion Chromatogram – XIC

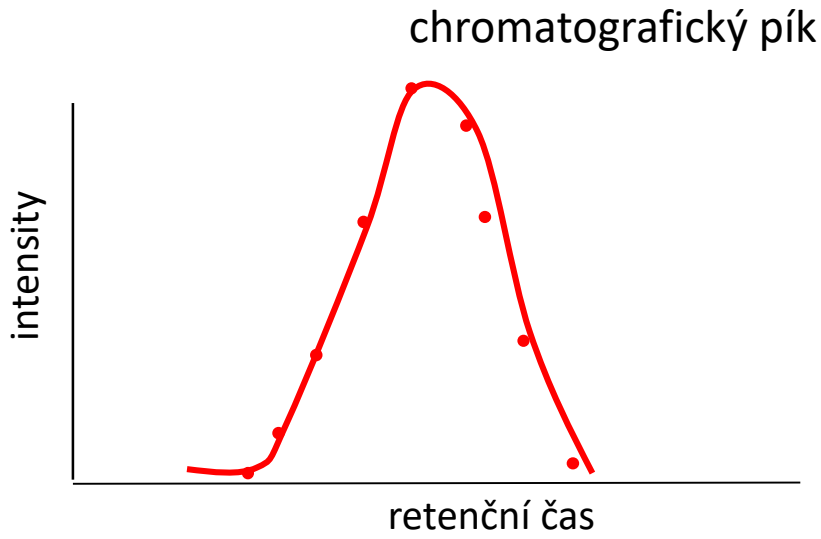




Kde se u DDA bere kvantitativní informace?

Extracted Ion Chromatogram – XIC

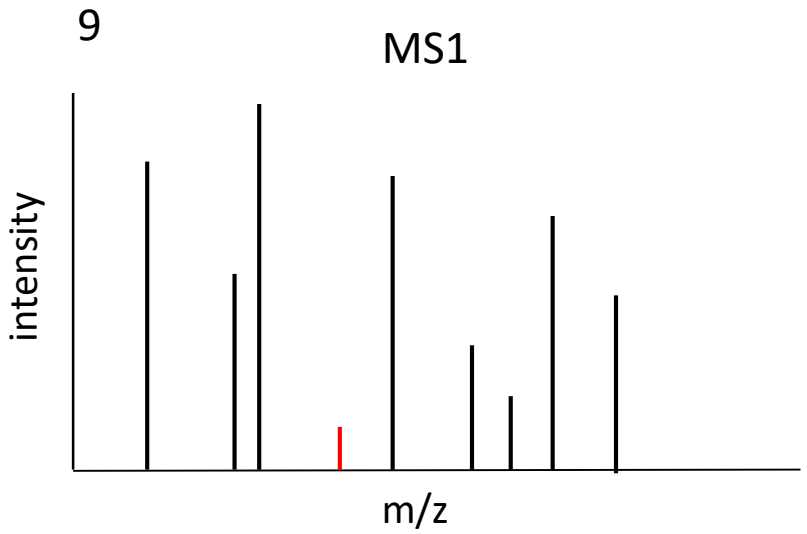


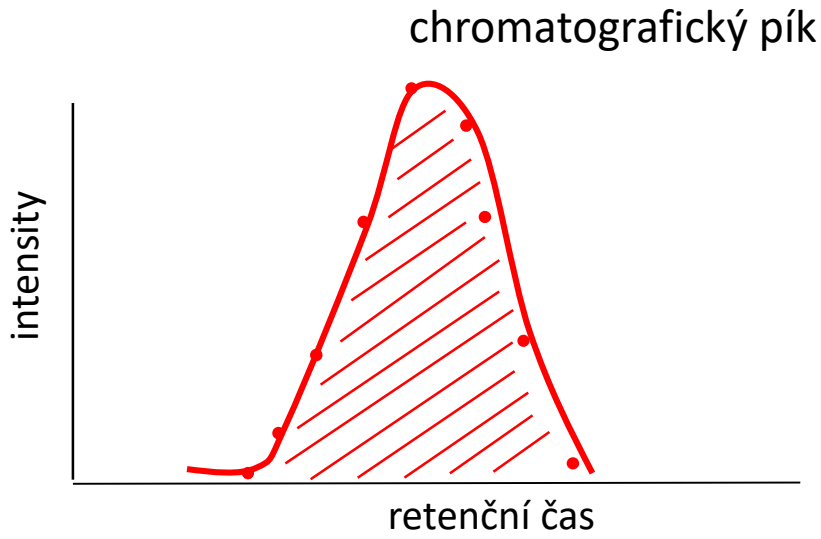


- **Kvantitativní informace obsažena v MS1 signálu**
- V případě DDA analýzy **MS2 spektra neobsahují** kvantitativní informaci
- Základ pro tzv. **Label Free Quantification** (LFQ)

Kde se u DDA bere kvantitativní informace?

Extracted Ion Chromatogram – XIC

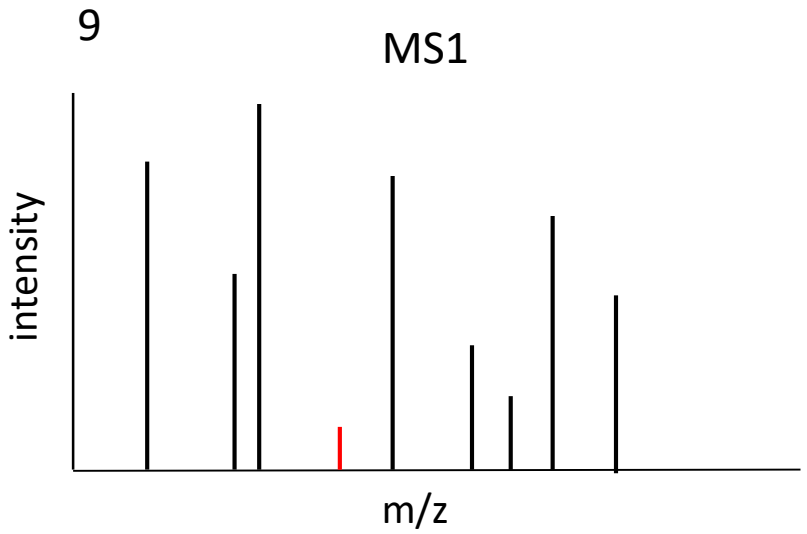


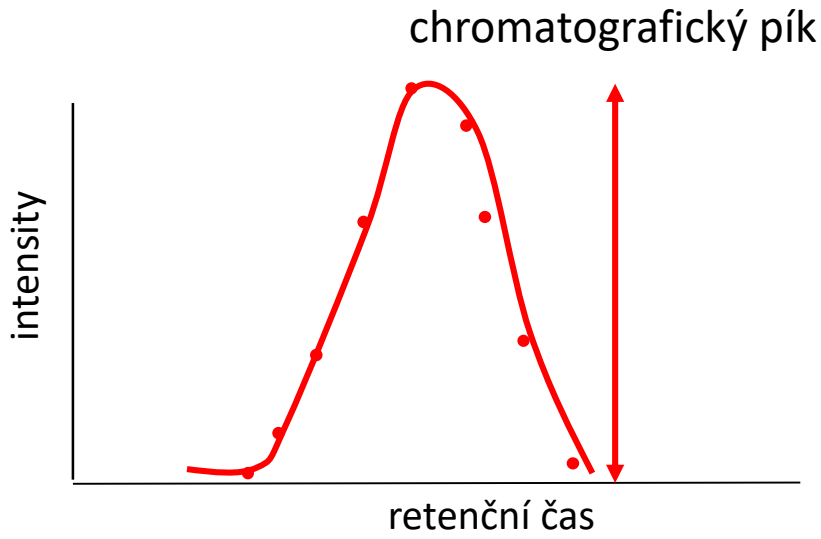


- **Plocha pod píkem**
  - **použití u cílených metod (SRM, PRM)**

Kde se u DDA bere kvantitativní informace?

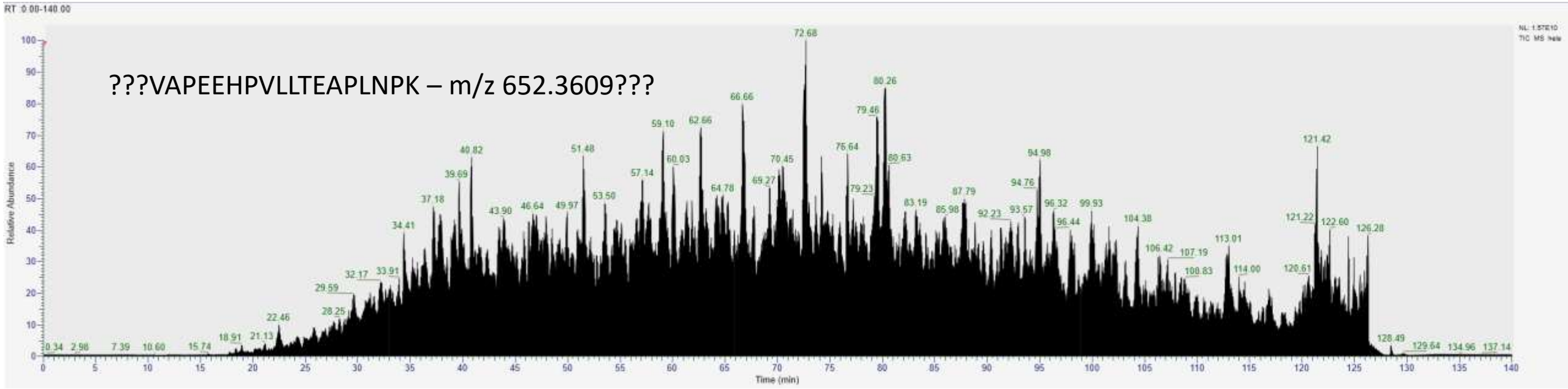
Extracted Ion Chromatogram – XIC





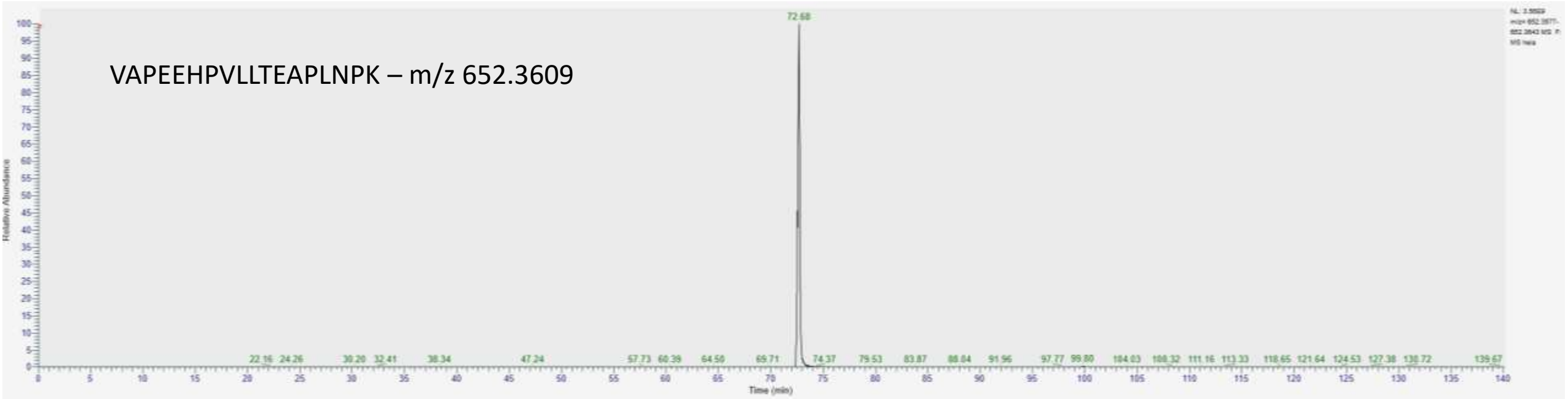
- **Plocha pod píkem**
  - **použití u cílených metod (SRM, PRM)**
- **Výška píku v nejvyšším bodě**
  - **necílená proteomika**
  - **vyšší reprodukovatelnost**

Kde se u DDA bere kvantitativní informace?



Kde se u DDA bere kvantitativní informace?

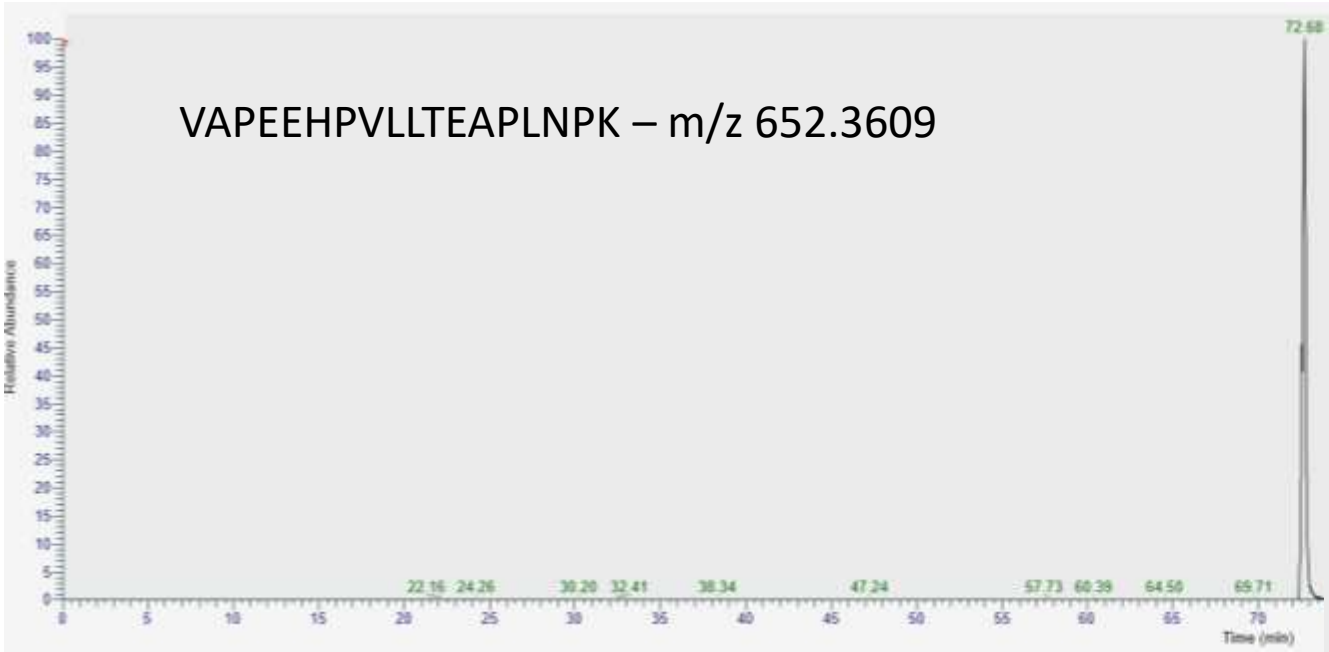
Extracted Ion Chromatogram – XIC

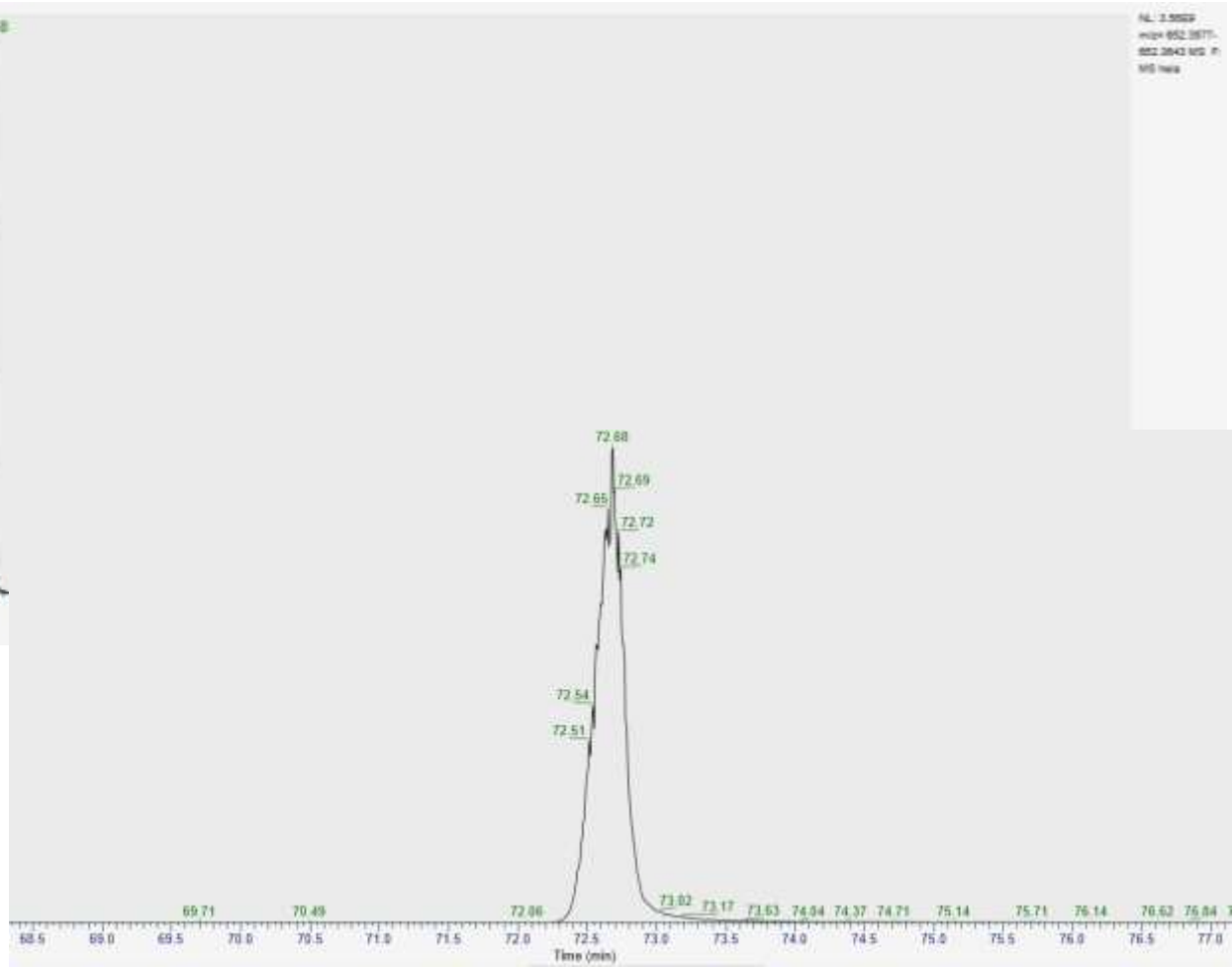


Kde se u DDA bere kvantitativní informace?

Extracted Ion Chromatogram – XIC

VAPEEHPVLLTEAPLNPK – m/z 652.3609

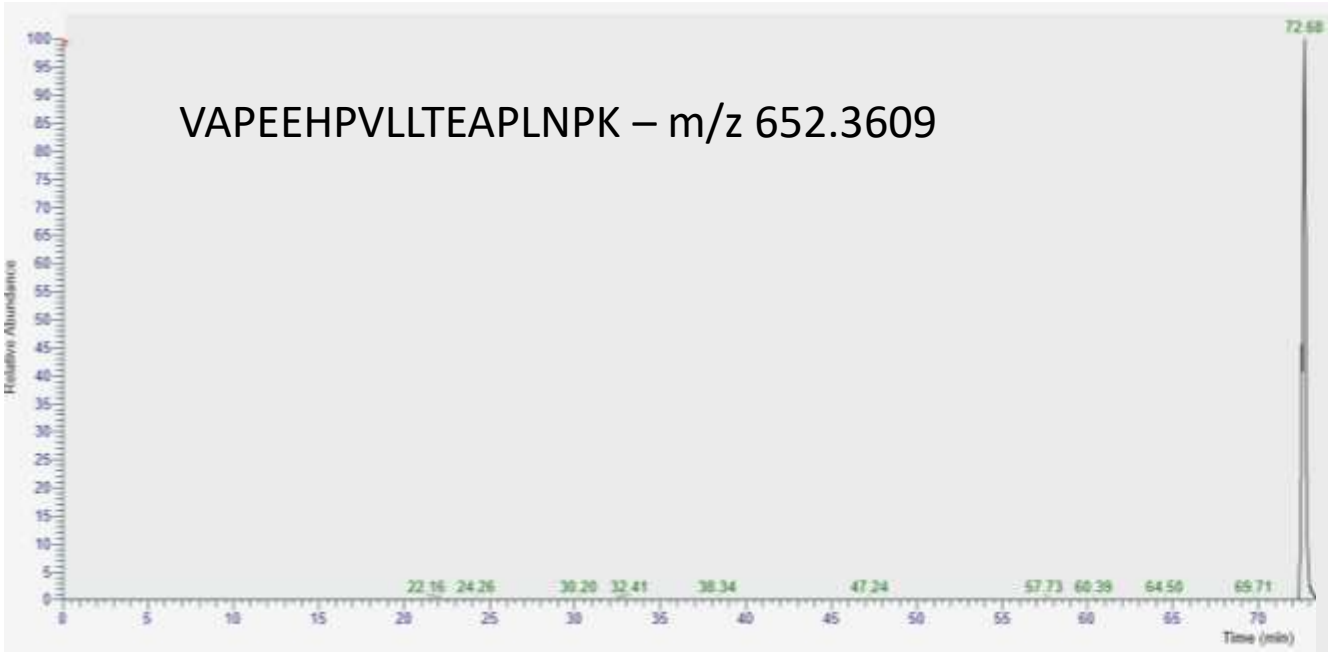


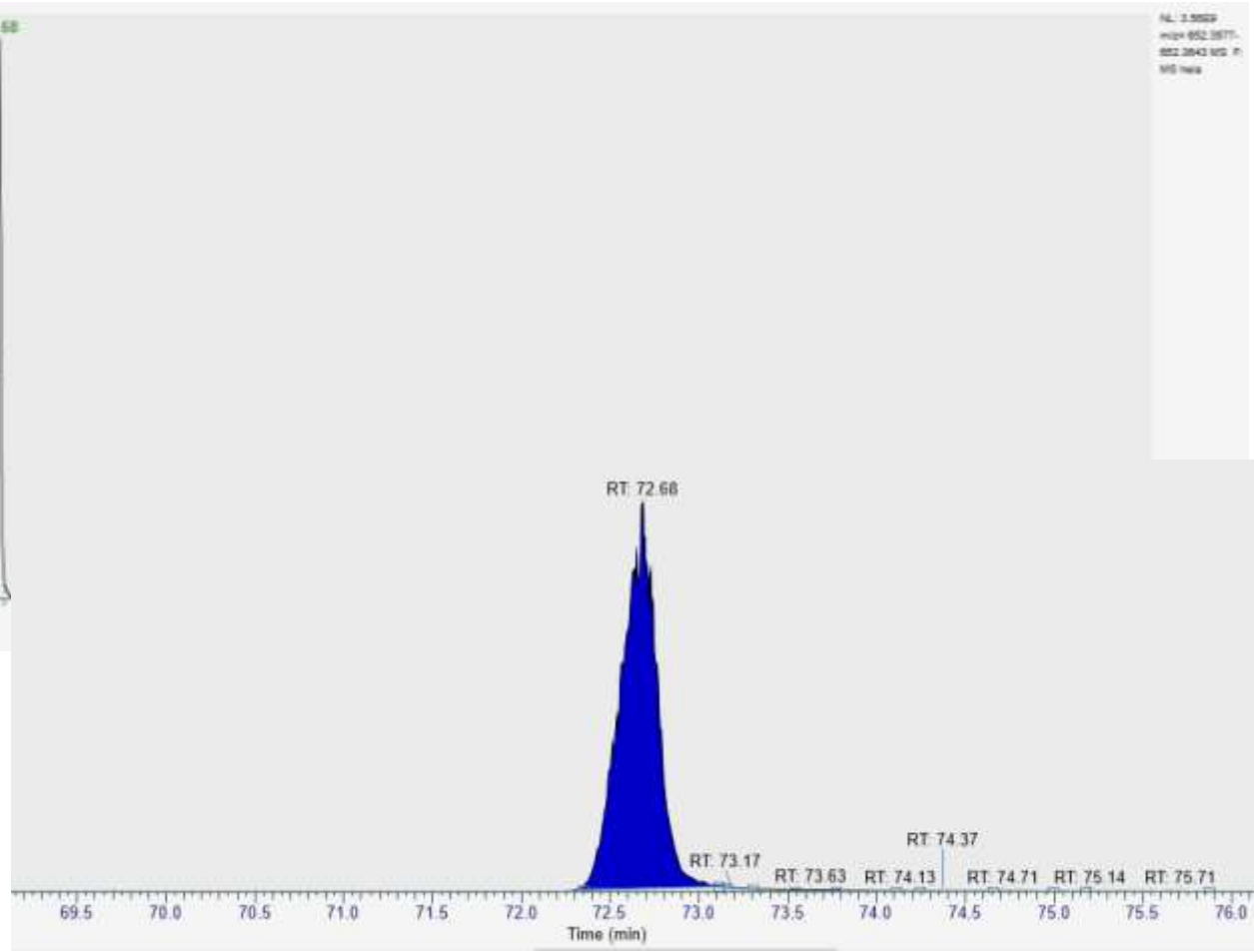


Kde se u DDA bere kvantitativní informace?

Extracted Ion Chromatogram – XIC

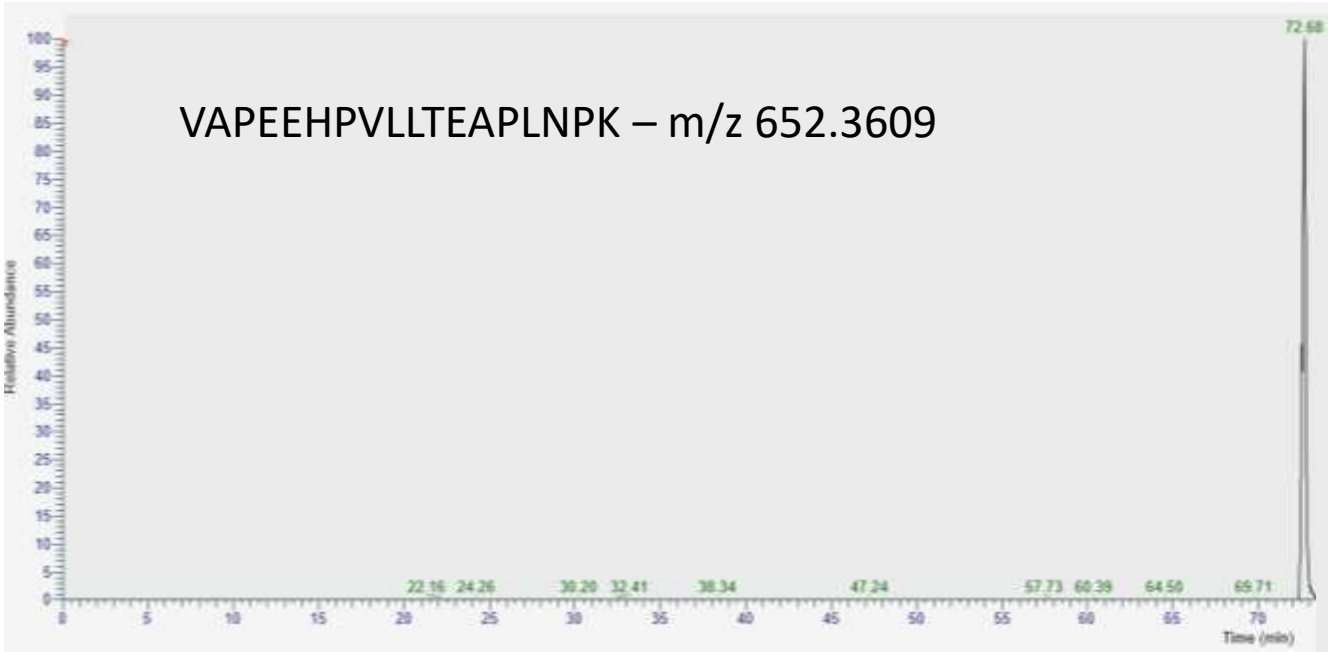
VAPEEHPVLLTEAPLNPK – m/z 652.3609

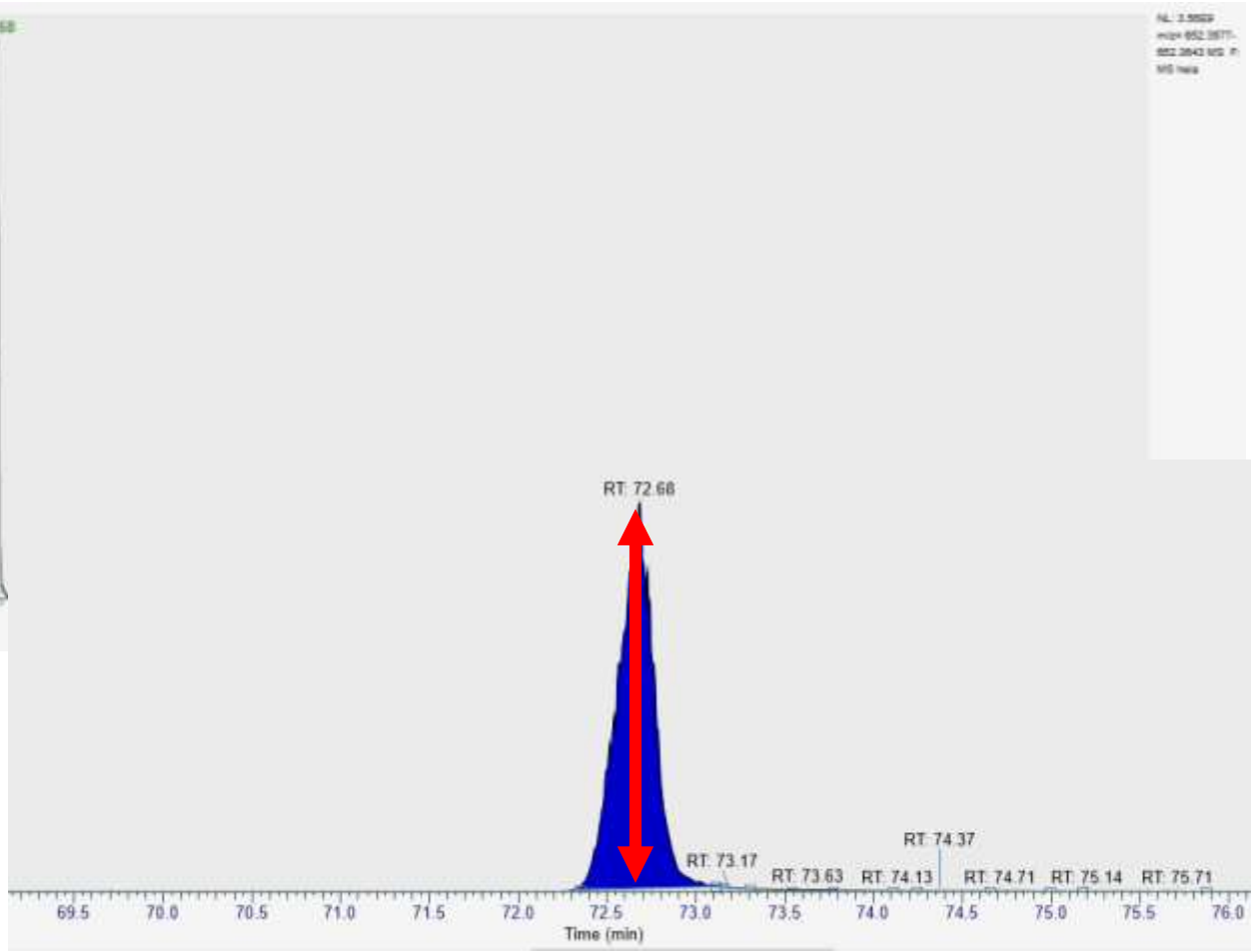




Kde se u DDA bere kvantitativní informace?

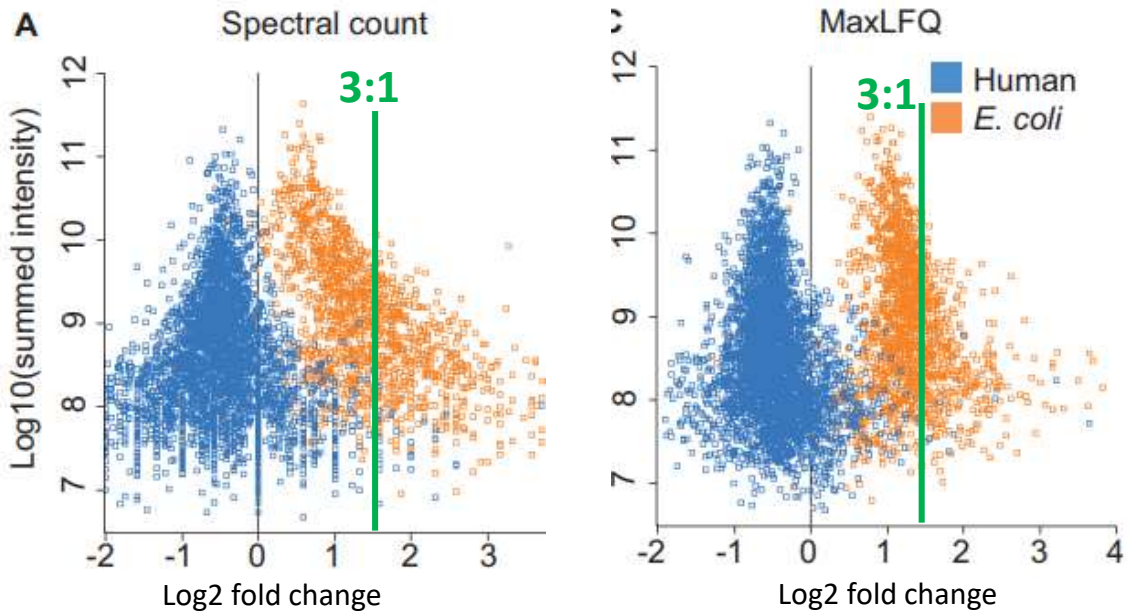
Extracted Ion Chromatogram – XIC





Testování přesnosti kvantifikace

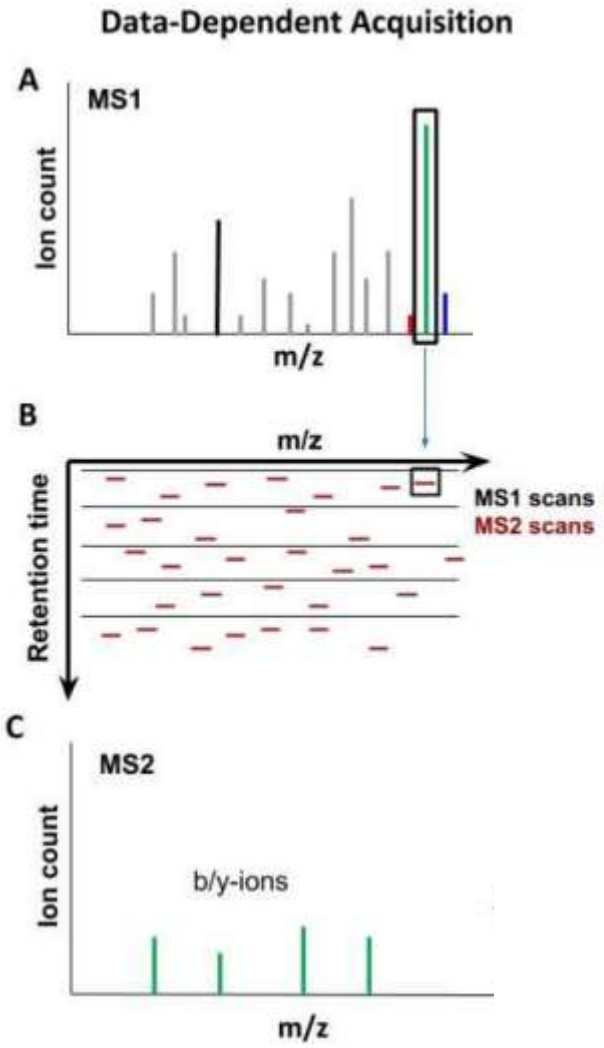
- 2 odlišné proteomy (člověk vs. E. coli)
- smíchány v poměru 1:1 – kontrolní vzorek
- smíchány v dalších známých poměrech (zde 3:1)



Cox et al, Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ, Mol Cell Proteomics 2014

Data Dependent Acquisition

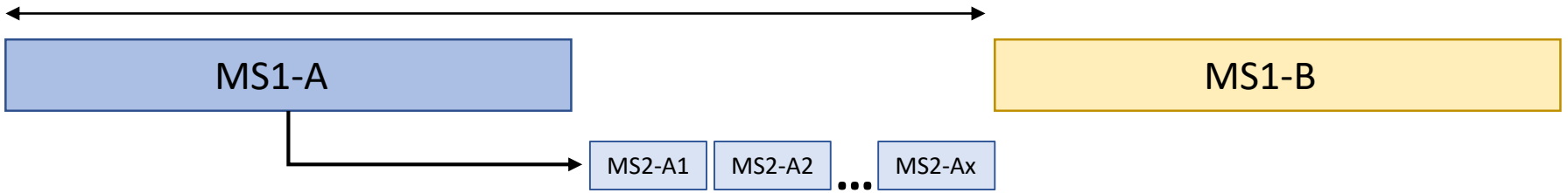
- Stochastický/semistochastický princip – výběr prekurzorů je v podstatě náhodný
- Detekce zhruba 5 000 proteinů ve vzorku savčích buněk při 120 min gradientu
- Máme jednoznačné propojení prekurzor – fragmentační spektrum
- Kvantitativní informace uložena v MS1
- Informace o peptidech, které nebyly vybrány pro fragmentaci je nevratně ztracena
- CV asi 25 %
- Typická je přítomnost značného množství missing values – asi 30 - 40%
  - Chybějící kvantitativní informace ve vzorcích, kde daný peptid nebyl fragmentován

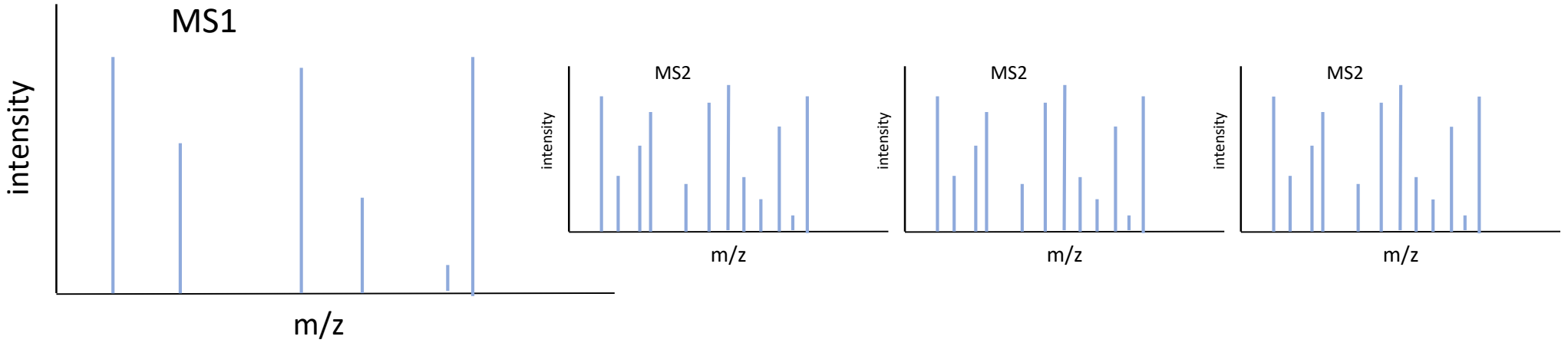


Pappireddi et al, **A Review on Quantitative Multiplexed Proteomics, ChemBioChem 2019**

Délka cyklu vs. šířka píku

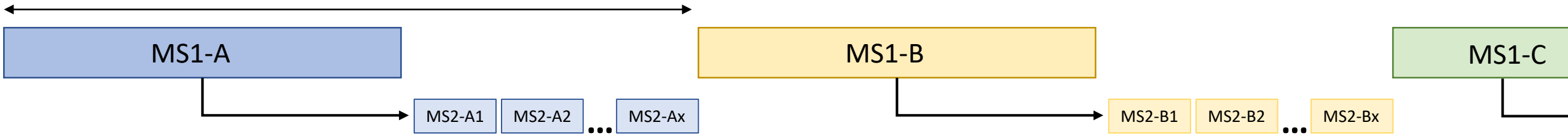
- **Skenovací cyklus**
- Celkový čas věnovaný jednomu cyklu - od počátku jednoho MS1 scanu do začátku následujícího MS1 scanu = cycle time

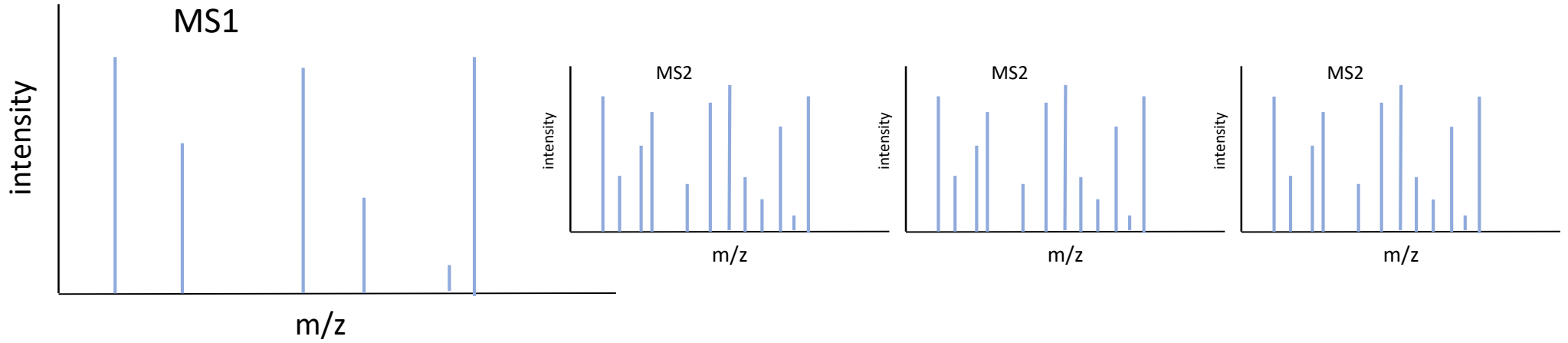




Délka cyklu vs. šířka píku

- **Skenovací cyklus**
  - Celkový čas věnovaný jednomu cyklu - od počátku jednoho MS1 scanu do začátku následujícího MS1 scanu = cycle time





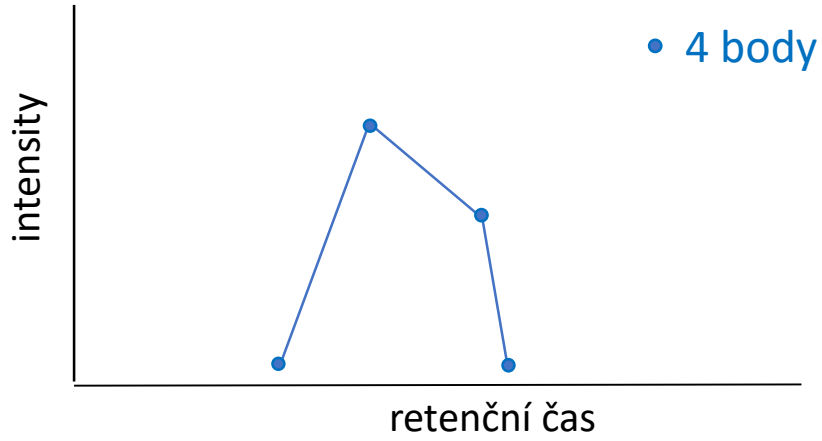
Délka cyklu vs. šířka píku

- Délka cyklu v závislosti na šířce chromatografického píku definuje počet bodů přes pík
- Přímo ovlivňuje přesnost kvantifikace – čím lepší pokrytí píku, tím přesnější kvantifikace

Máme 20 s chromatografický pík

Při délce cyklu 5 s

4 body přes pík



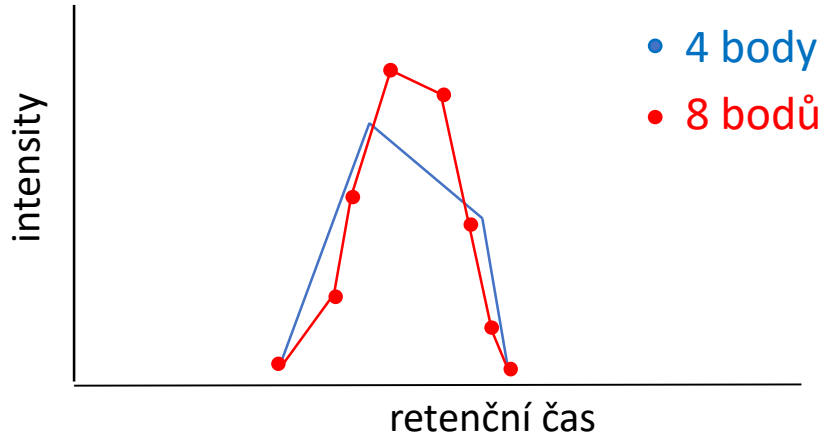
Délka cyklu vs. šířka píku

- Délka cyklu v závislosti na šířce chromatografického píku definuje počet bodů přes pík
- Přímo ovlivňuje přesnost kvantifikace – čím lepší pokrytí píku, tím přesnější kvantifikace

Máme 20 s chromatografický pík

Při délce cyklu 2.5 s

8 bodů přes pík



Délka cyklu vs. šířka píku

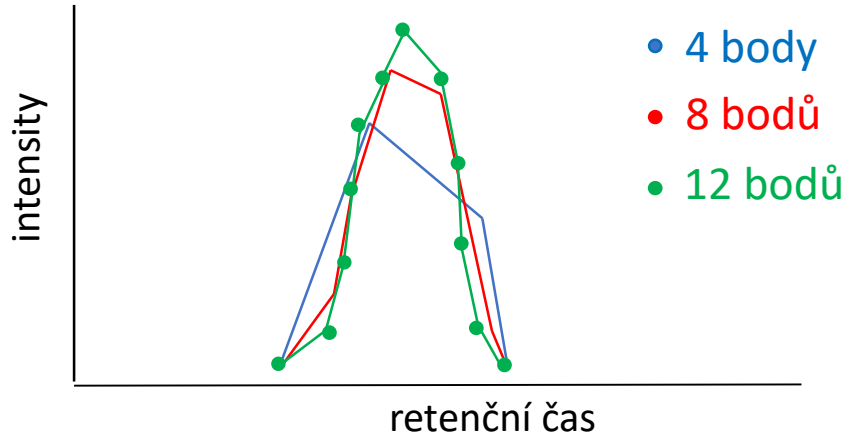
- Délka cyklu v závislosti na šířce chromatografického píku definuje počet bodů přes pík

- Přímo ovlivňuje přesnost kvantifikace – čím lepší pokrytí píku, tím přesnější kvantifikace

Máme 20 s chromatografický pík

Při délce cyklu 1.6 s

12 bodů přes pík



Délka cyklu vs. šířka píku

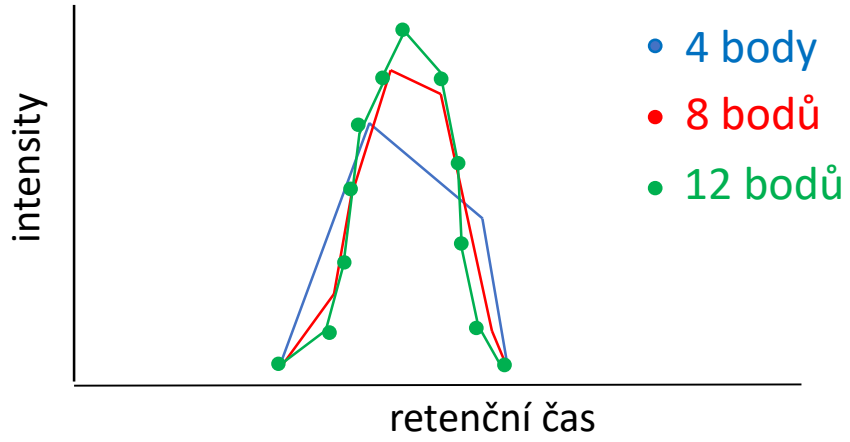
- Délka cyklu v závislosti na šířce chromatografického píku definuje počet bodů přes pík
- Přímo ovlivňuje přesnost kvantifikace – čím lepší pokrytí píku, tím přesnější kvantifikace

- Delší cyklus
  - nasbíráme více MS/MS spekter
  - máme horší pokrytí píku
- Kratší cyklus
  - lepší pokrytí píku
  - méně MS/MS spekter

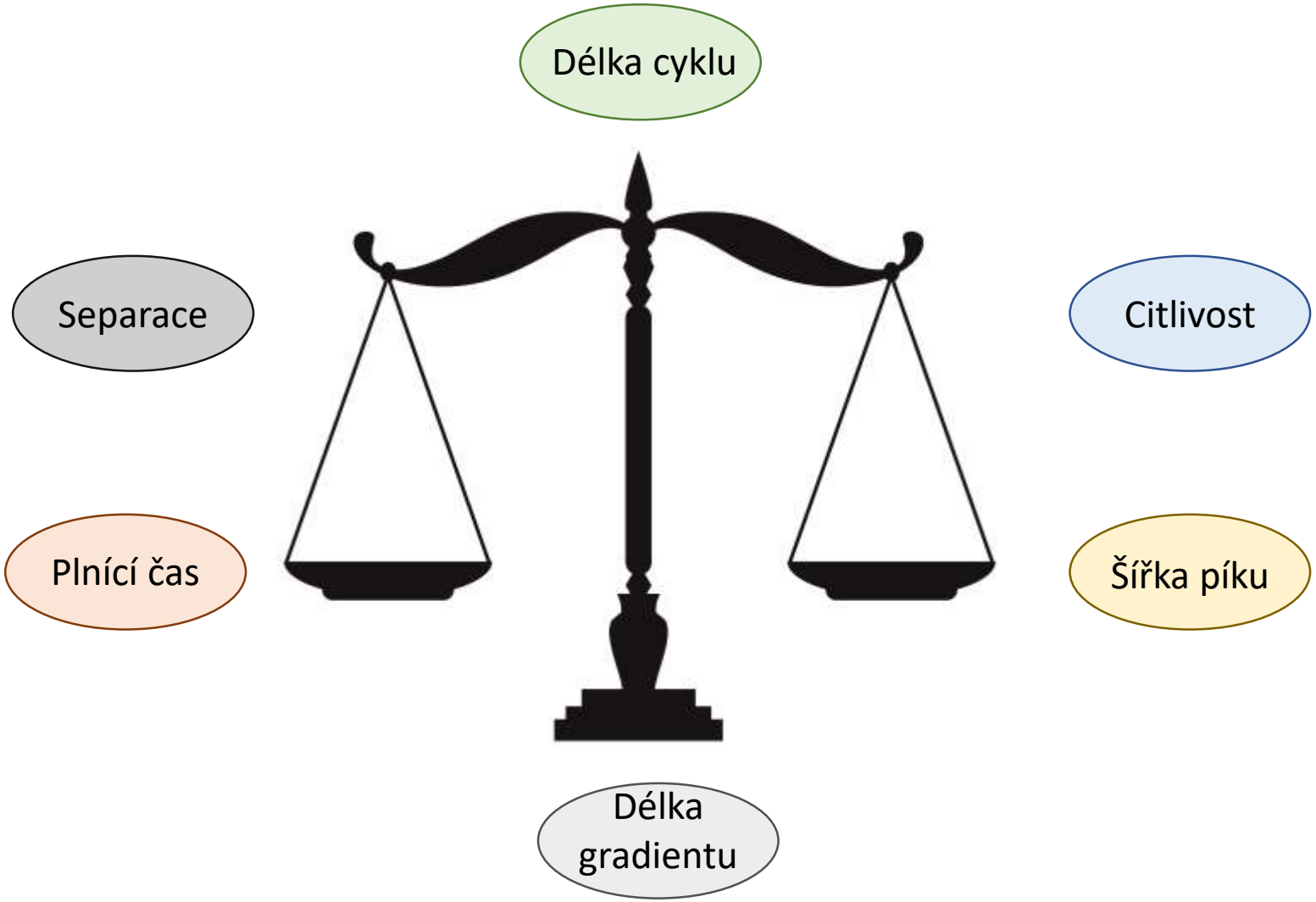
Máme 20 s chromatografický pík

Při délce cyklu 1,6 s

12 bodů přes pík



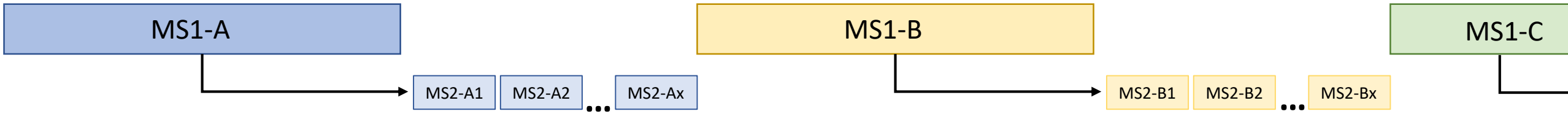
Délka cyklu vs. šířka píku



Paralelizace - tribridy

MS1 sken v orbitrapu asi 250 ms (rozlišení 120 K)
MS2 sken v iontové pasti asi 25 ms

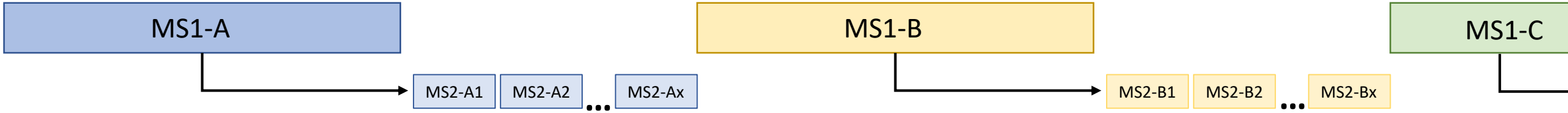
Bez paralelizace



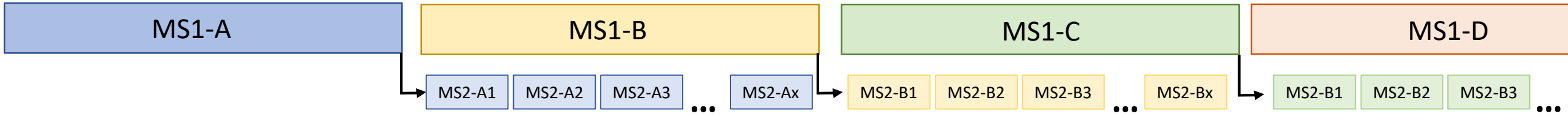
Paralelizace - tribridy

MS1 sken v orbitrapu asi 250 ms (rozlišení 120 K)
MS2 sken v iontové pasti asi 25 ms

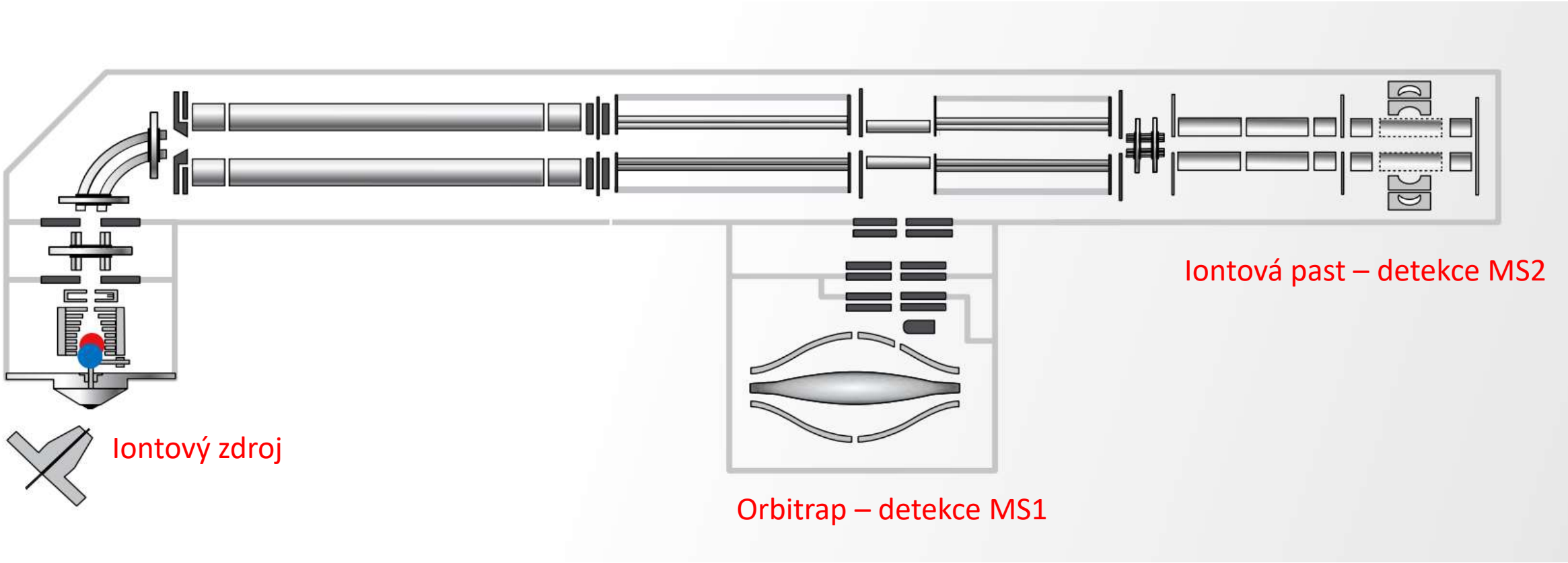
Bez paralelizace



S paralelizací



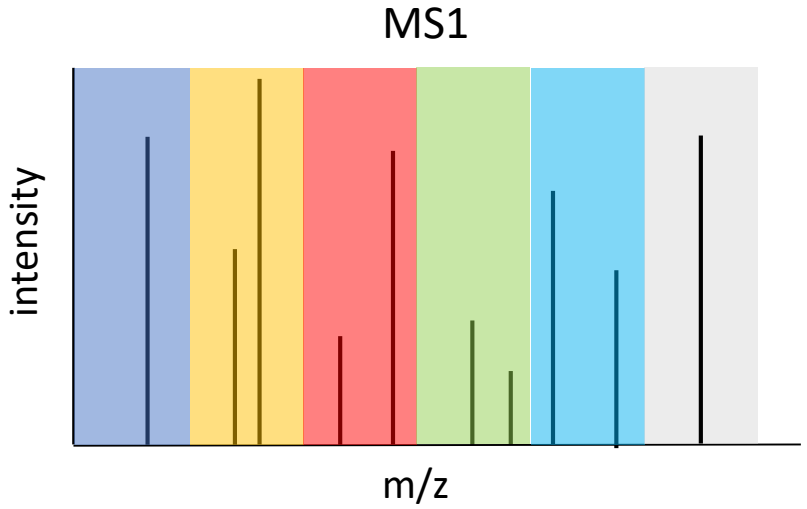
Paralelizace - tribridy



Thermo SCIENTIFIC

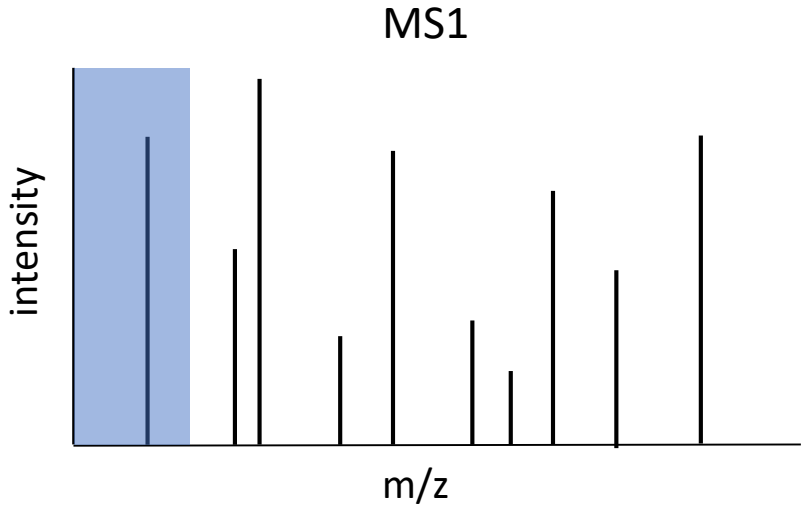
Data Independent Acquisition - DIA

Data Independent Acquisition - DIA

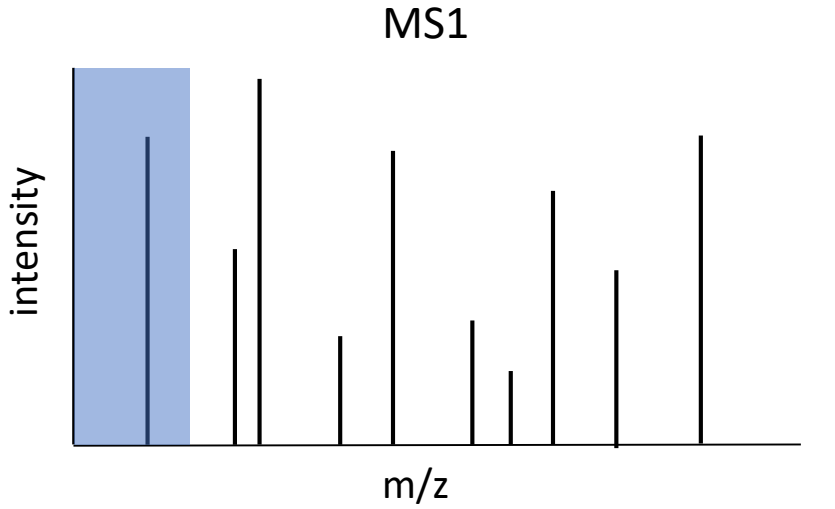


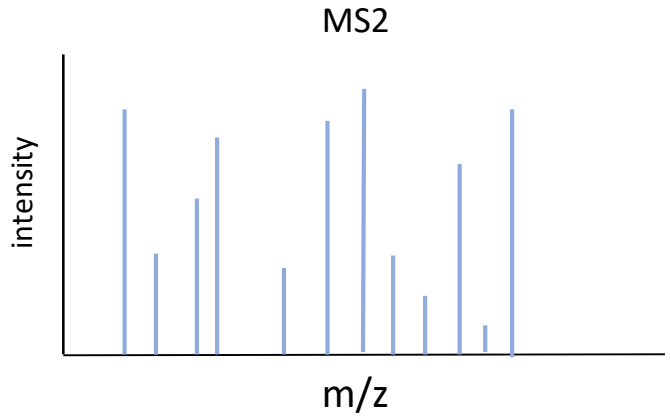
- Několik izolačních oken pokrývajících celý hmotnostní rozsah
- Šířka okna obvykle 10 – 20 Da
- **Izolace a fragmentace všech prekurzorů v daném hmotnostním okně**

Data Independent Acquisition - DIA

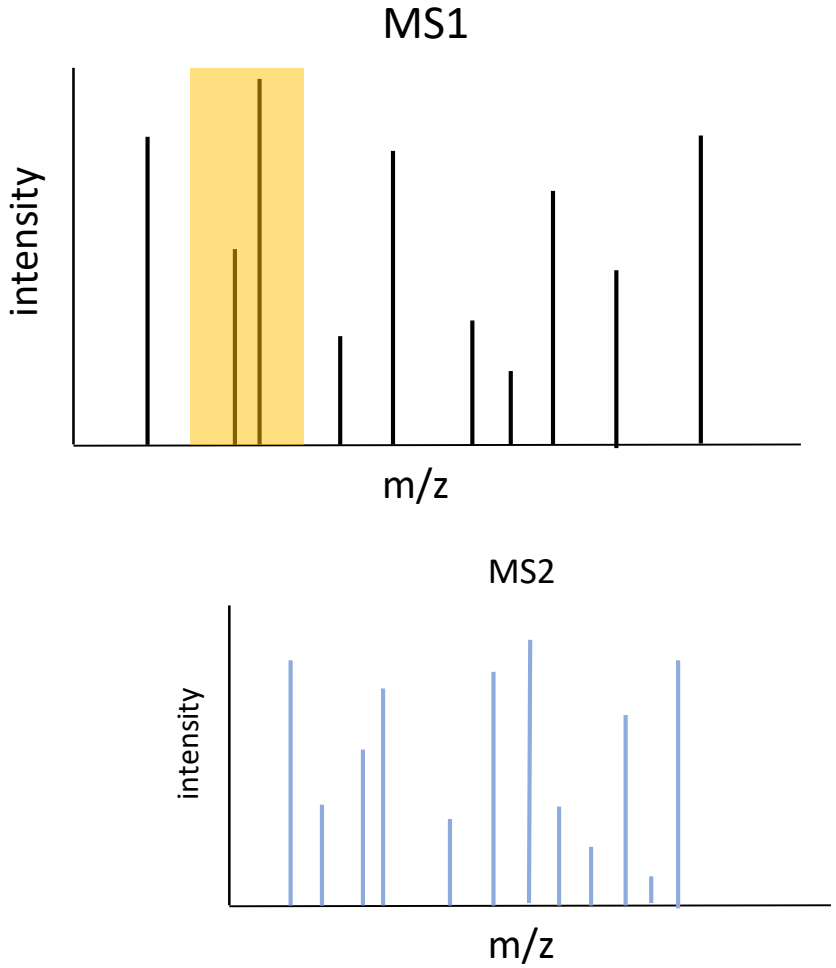


Data Independent Acquisition - DIA

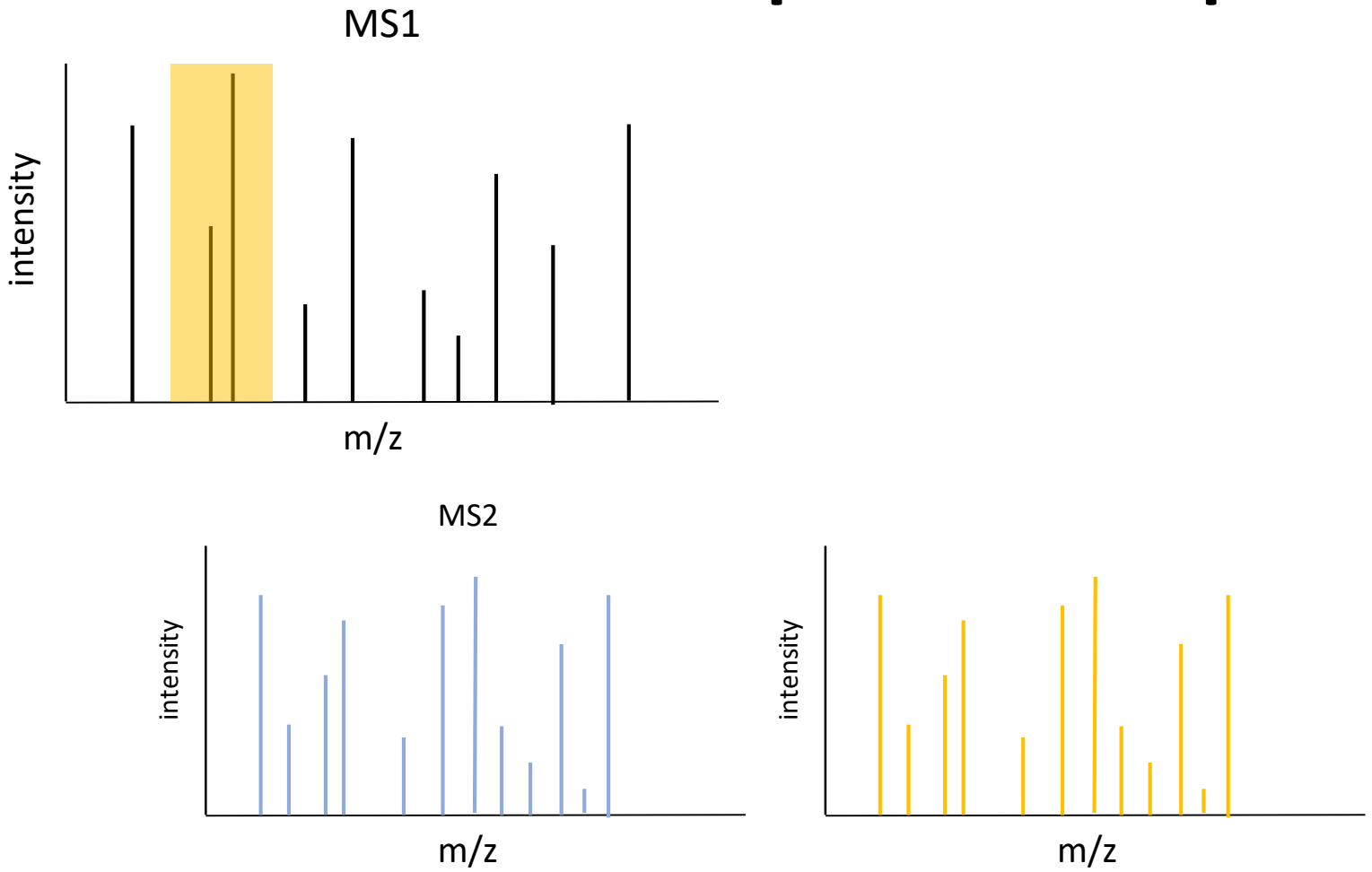




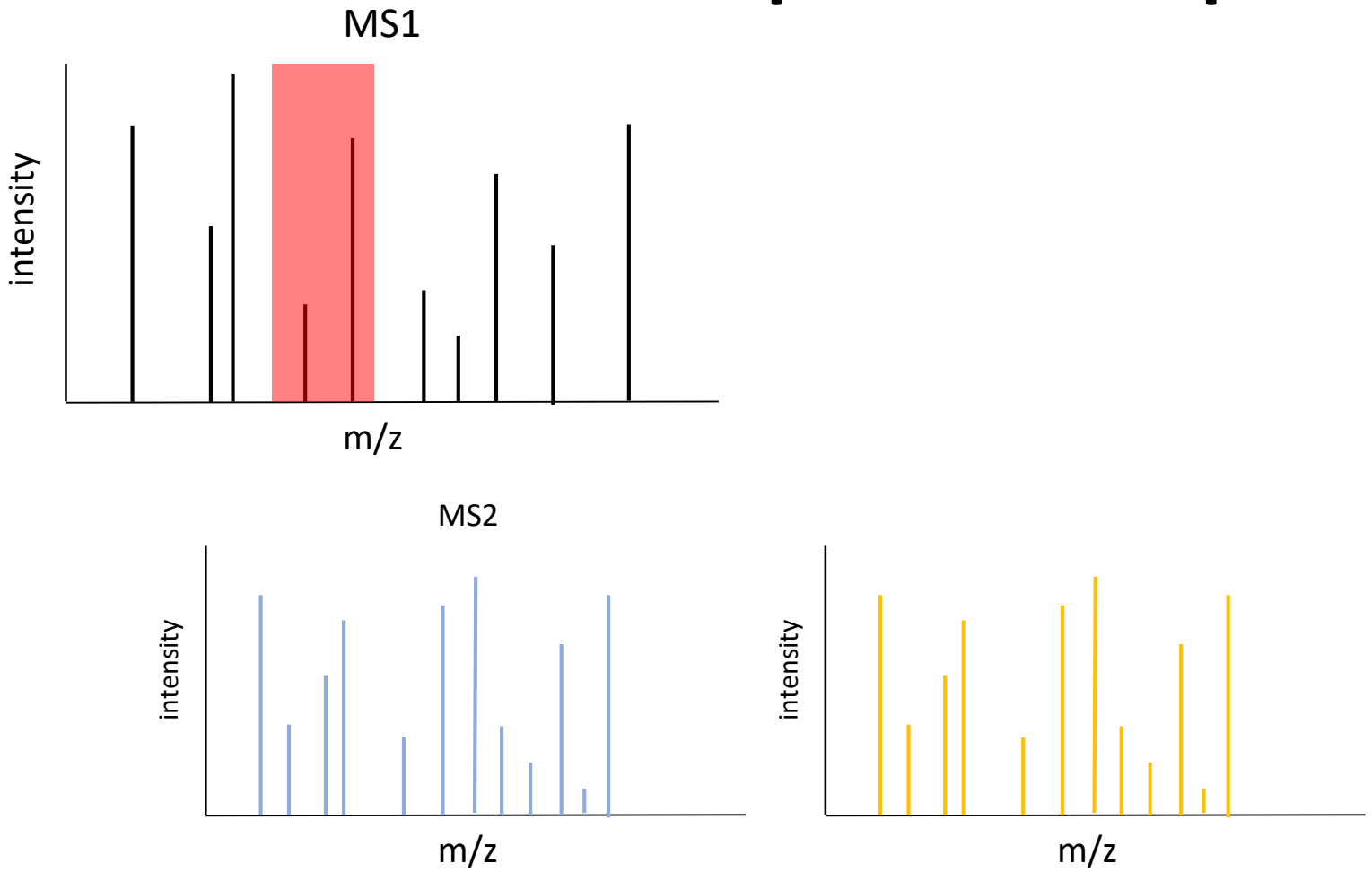
Data Independent Acquisition - DIA



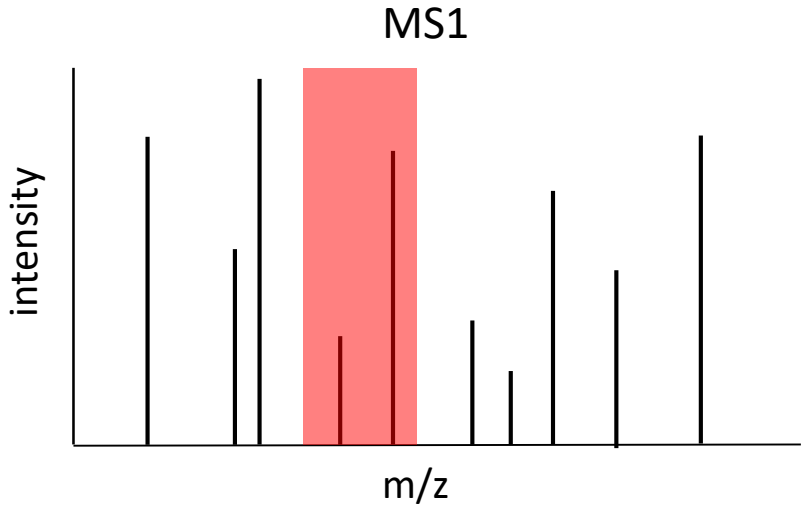
Data Independent Acquisition - DIA

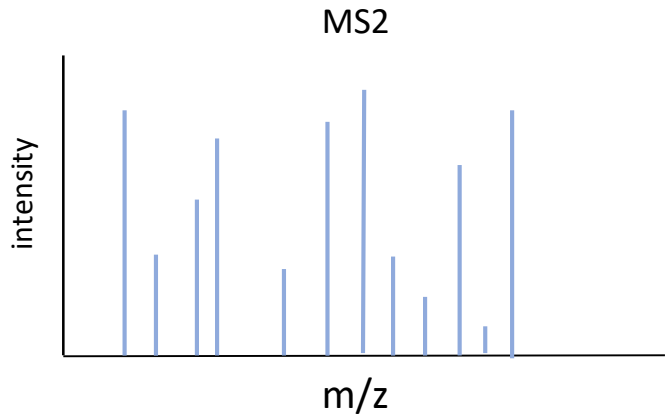


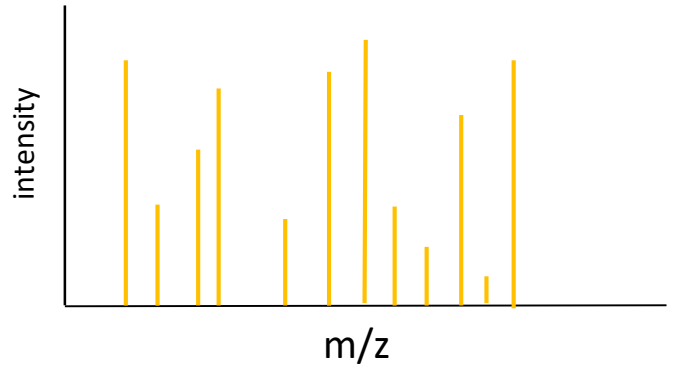
Data Independent Acquisition - DIA

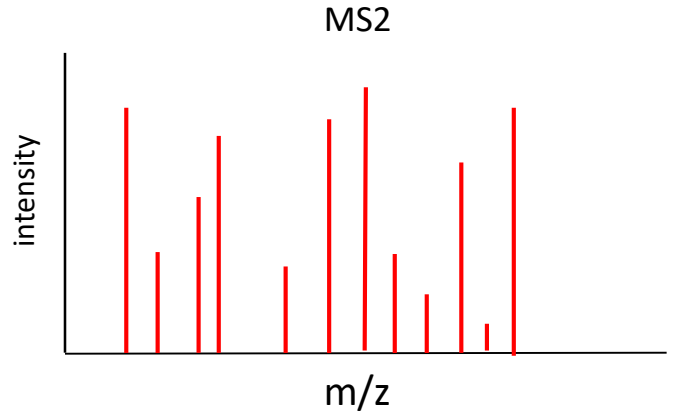


Data Independent Acquisition - DIA

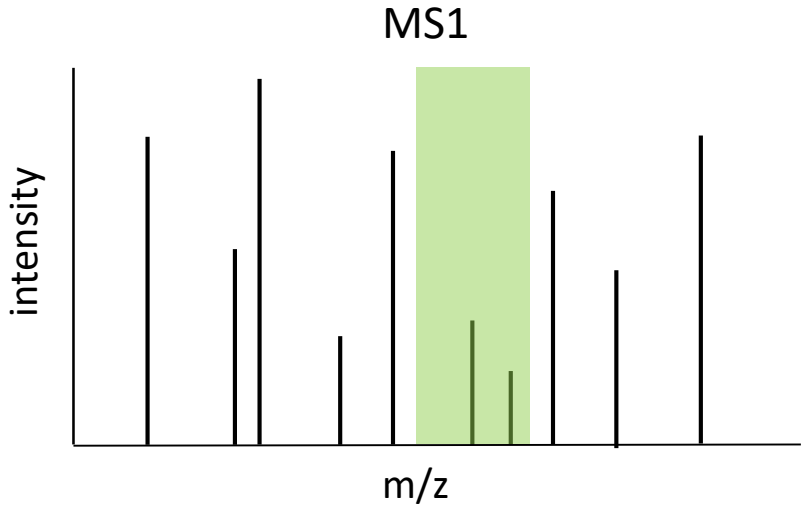


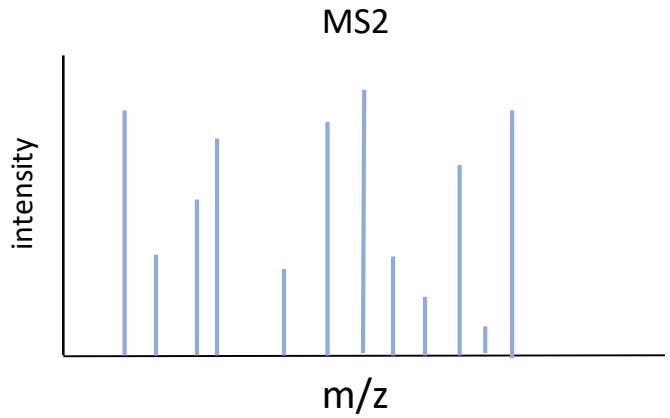


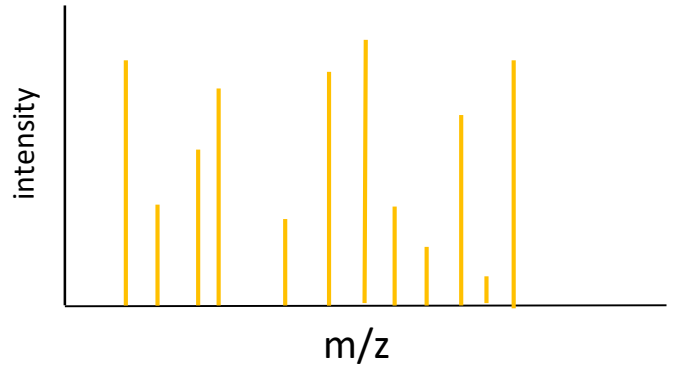


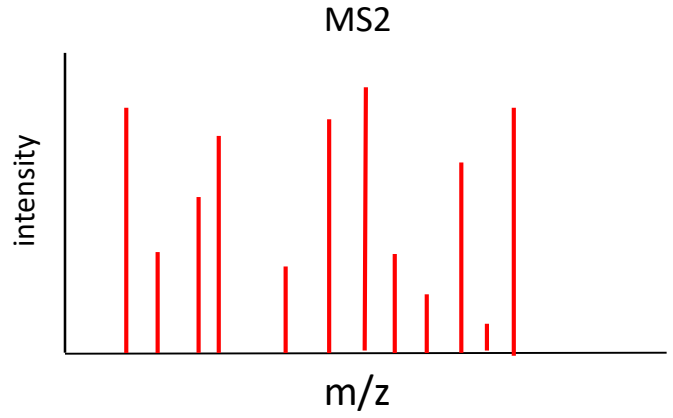


Data Independent Acquisition - DIA

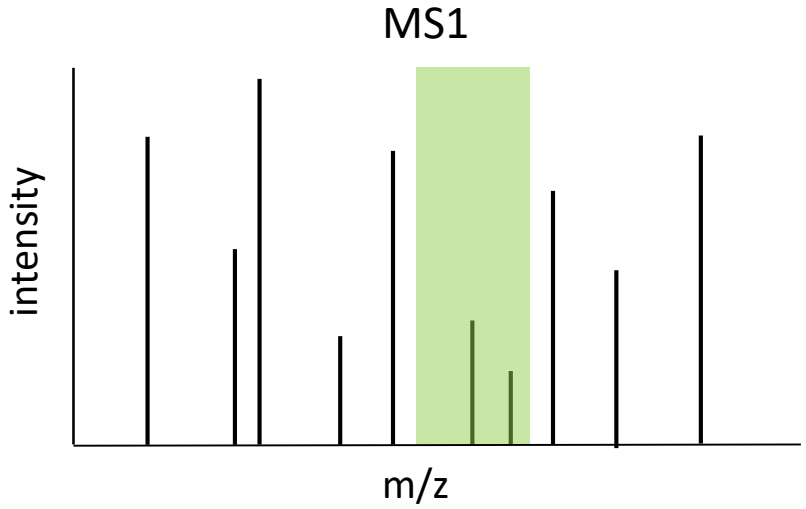


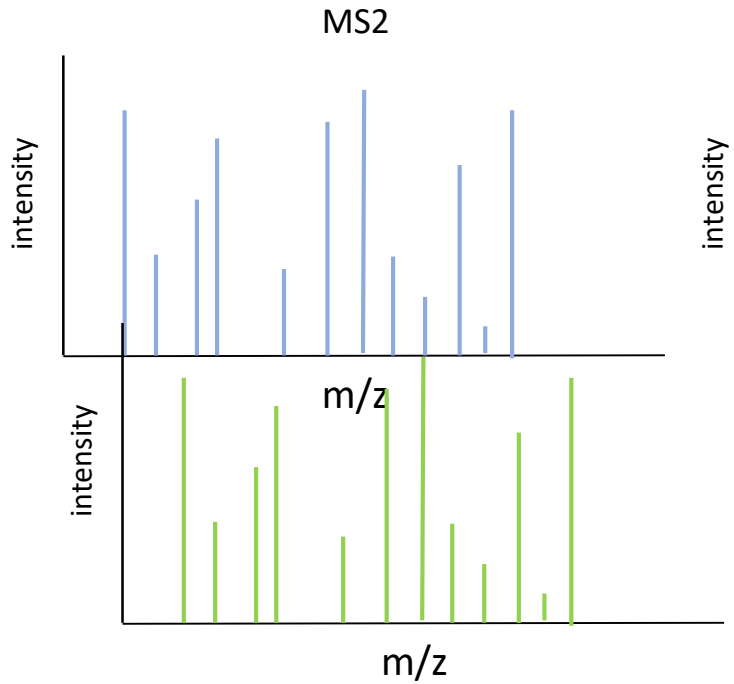


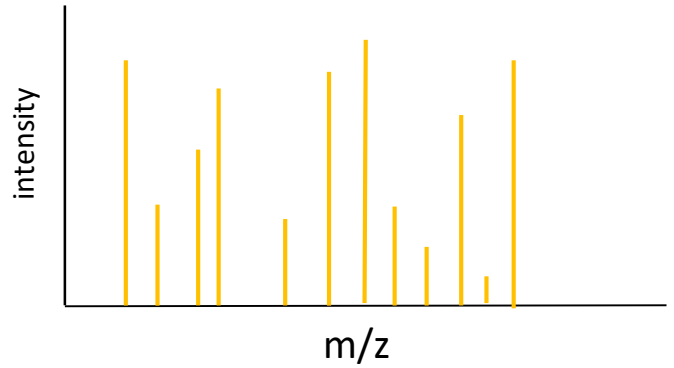


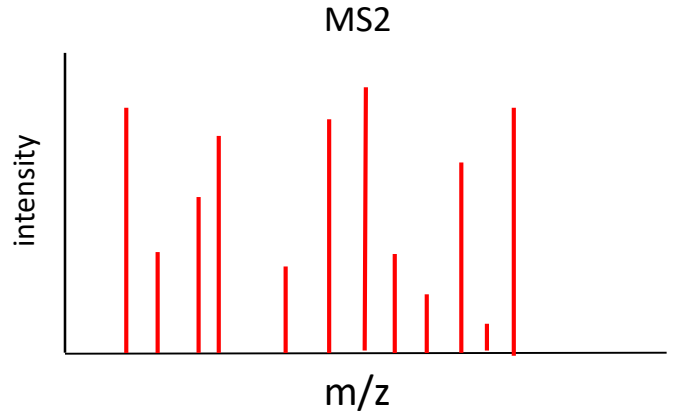


Data Independent Acquisition - DIA

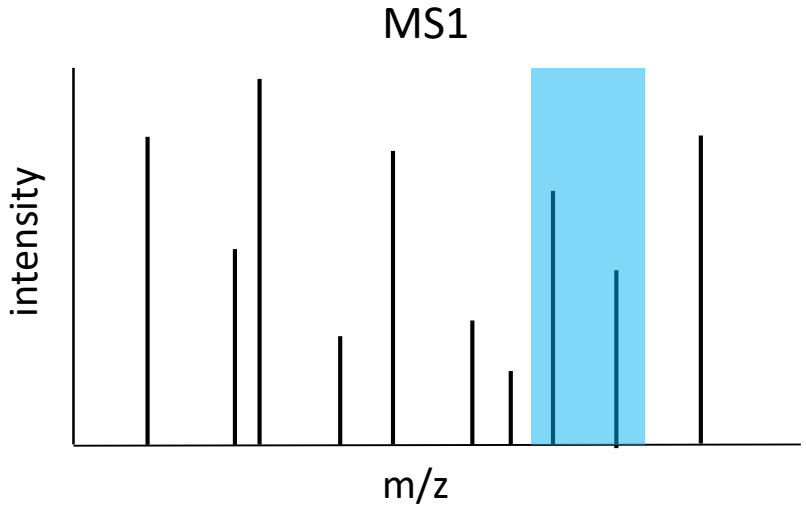


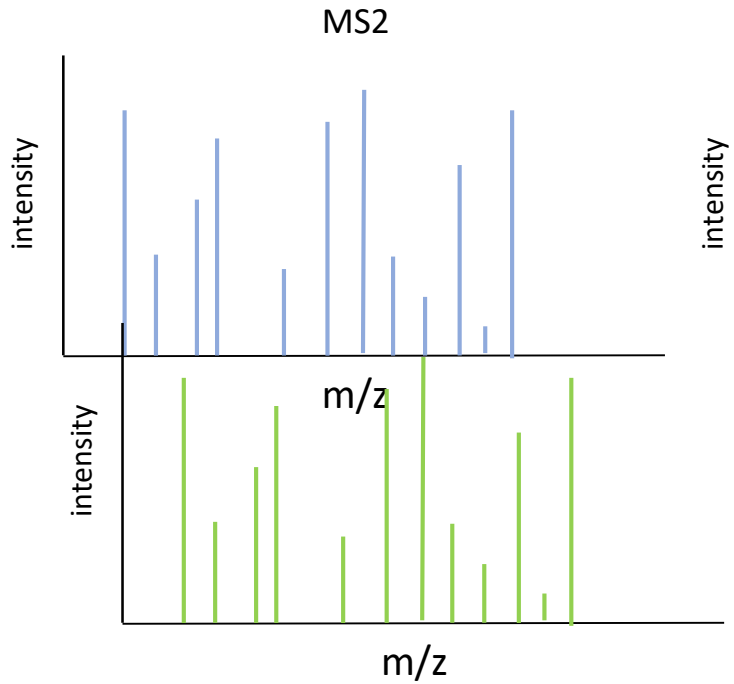


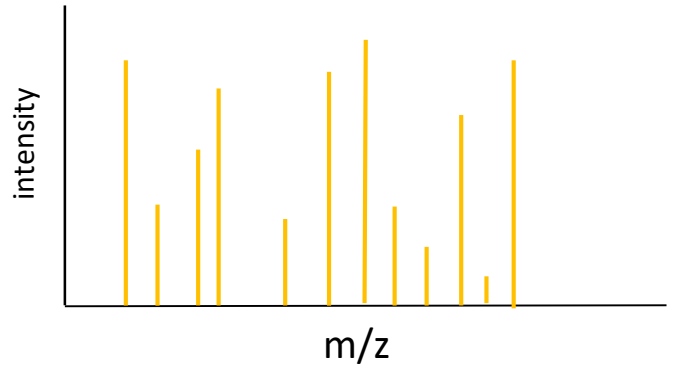


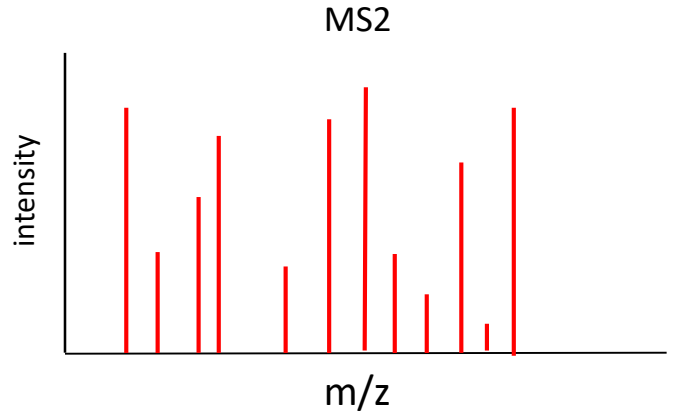


Data Independent Acquisition - DIA

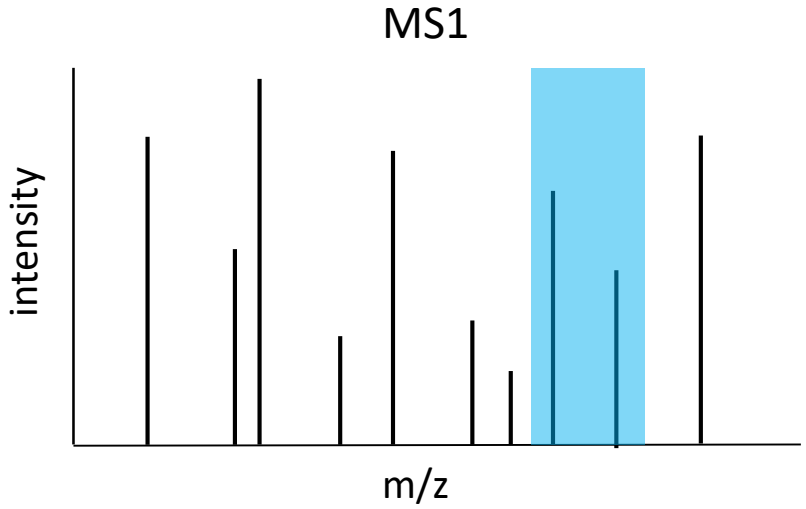


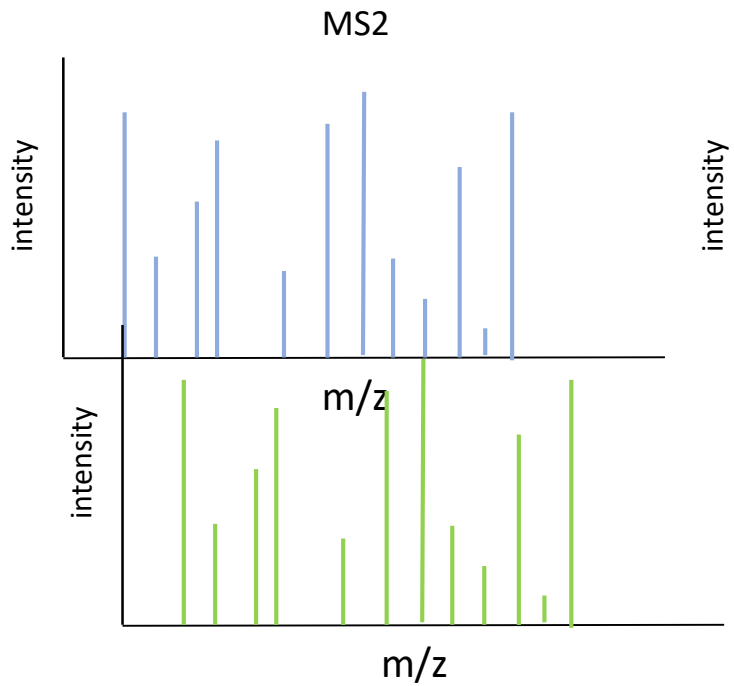


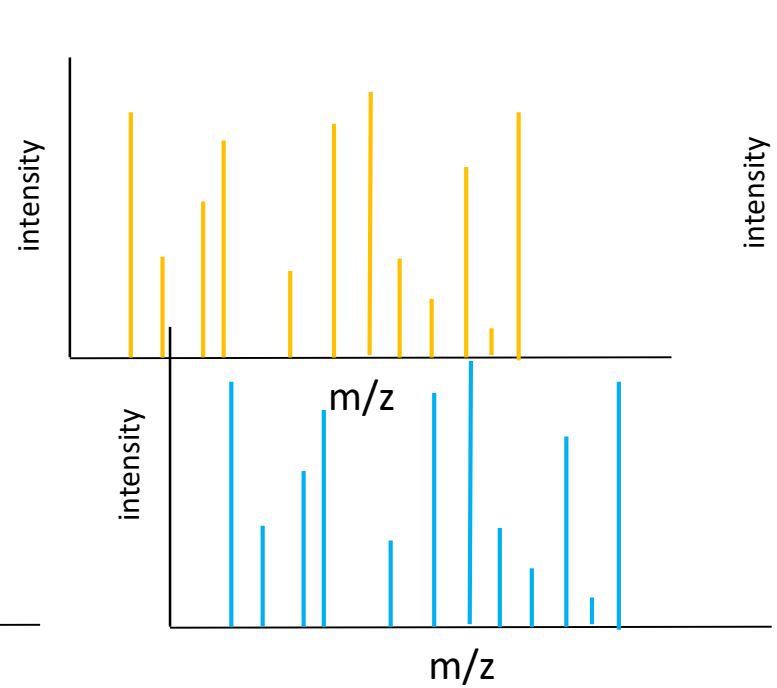


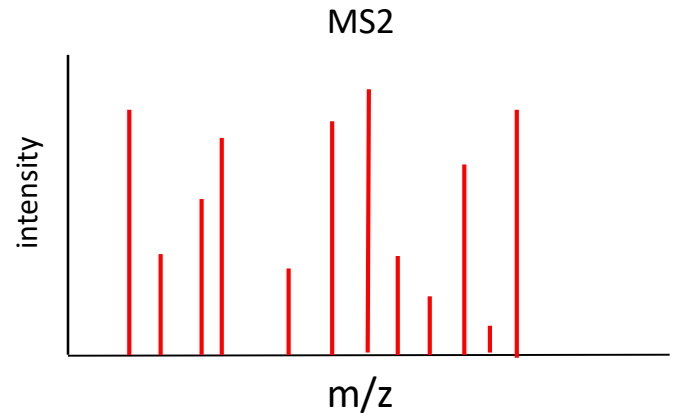


Data Independent Acquisition - DIA

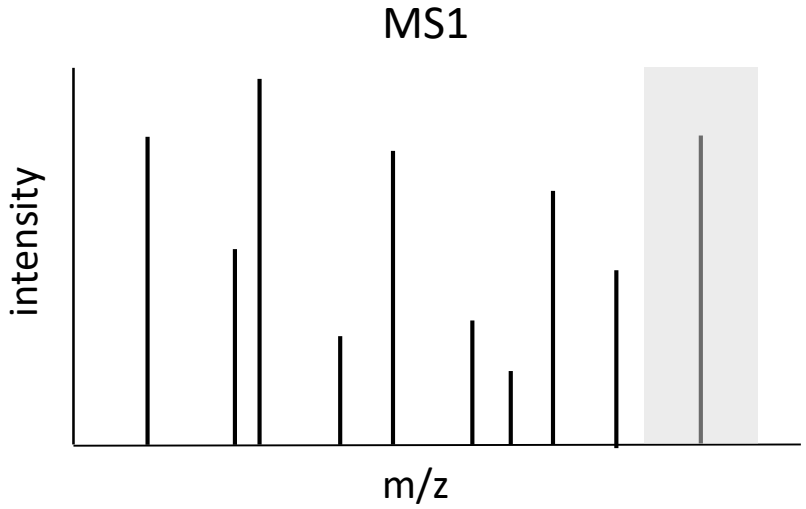


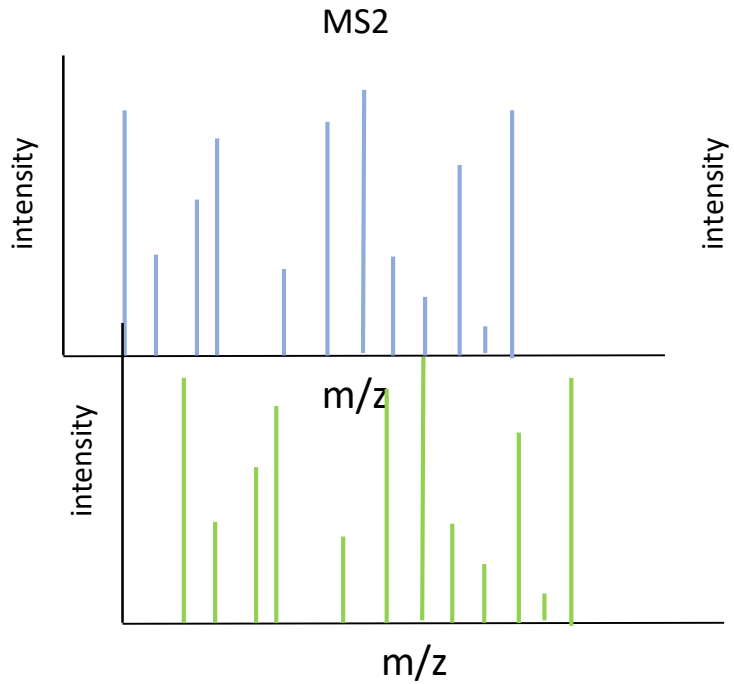


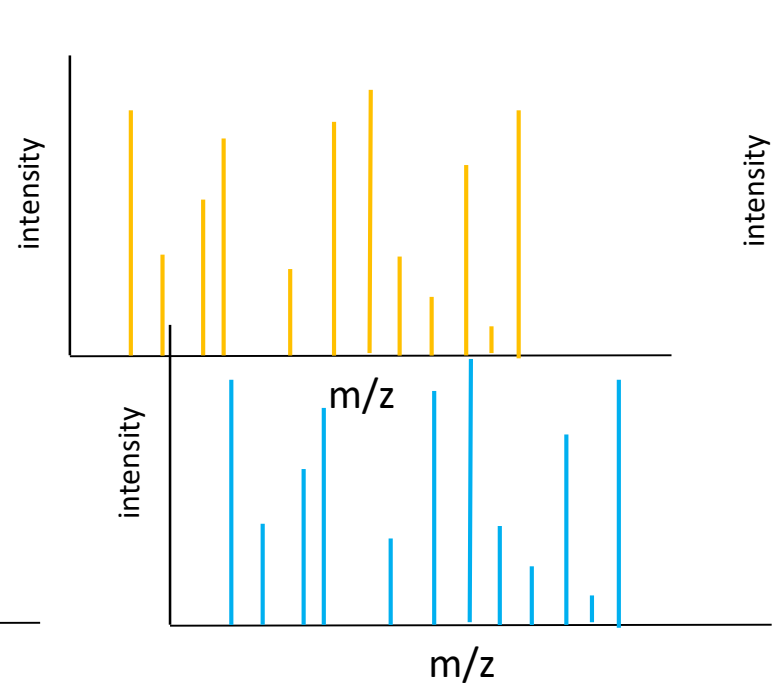


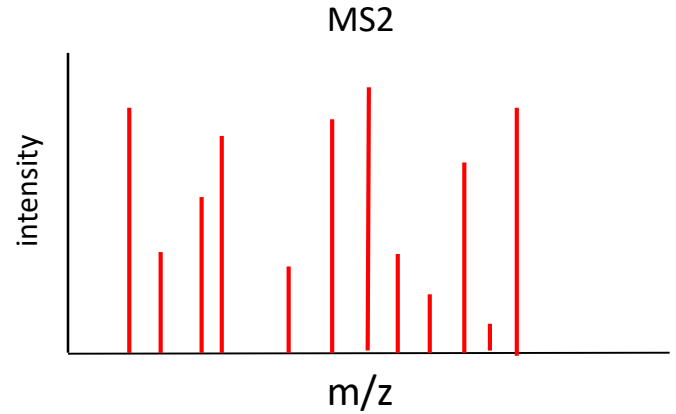


Data Independent Acquisition - DIA

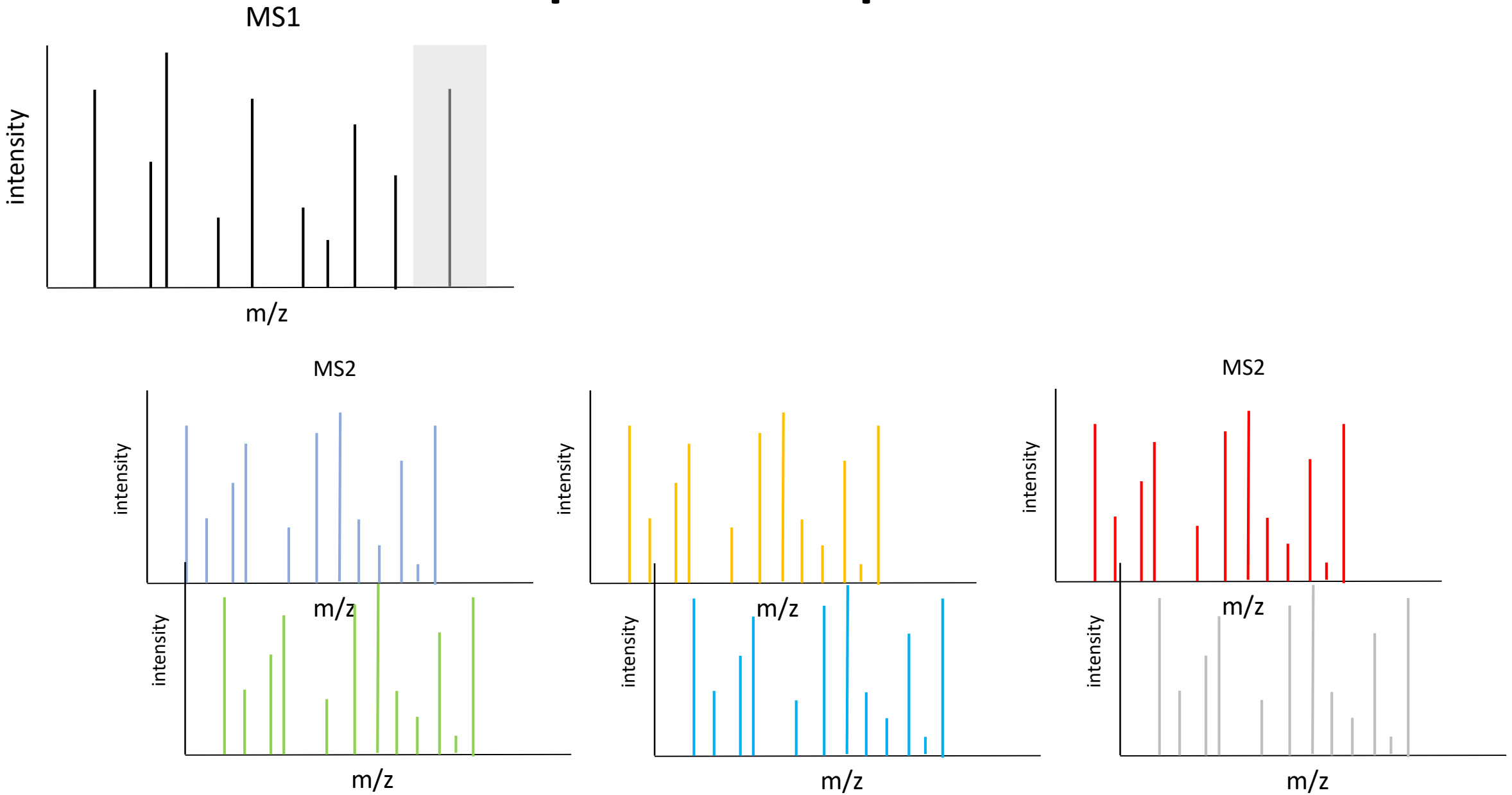






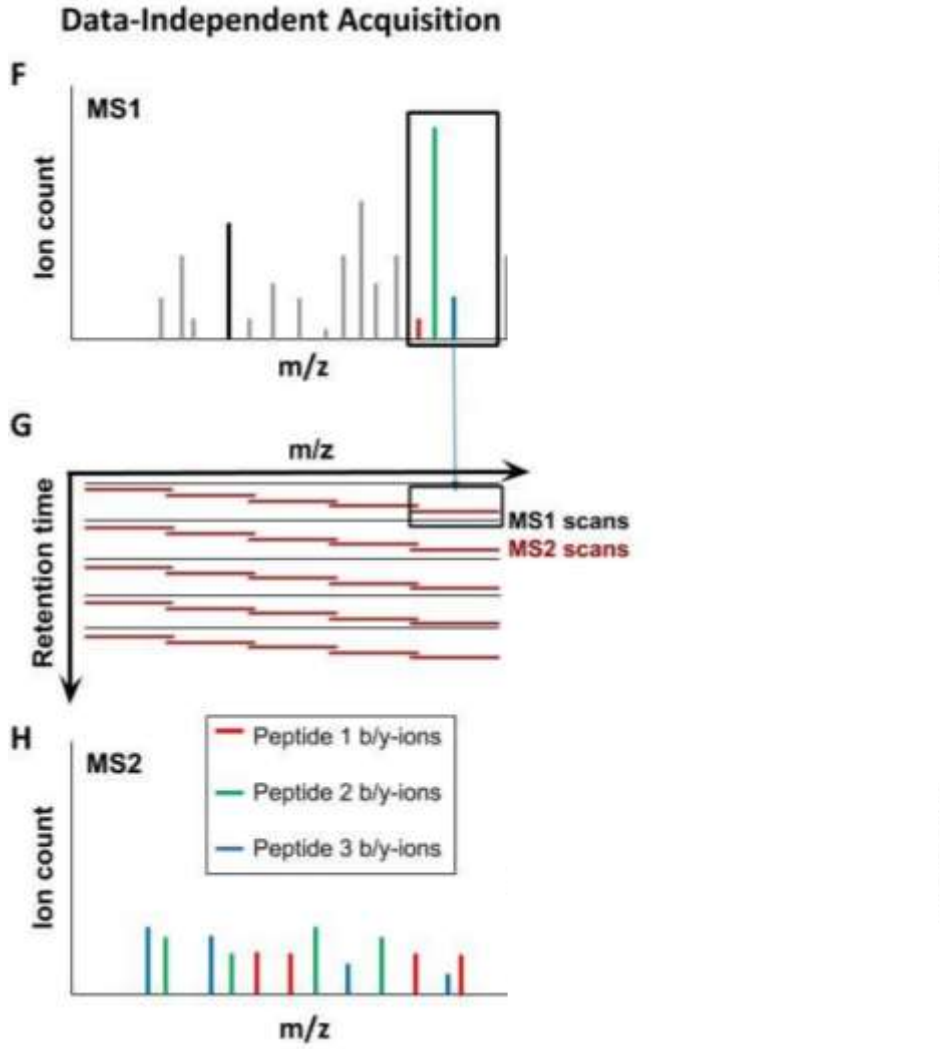


Data Independent Acquisition - DIA



Data Independent Acquisition - DIA

- Fragmentační spektrum obsahuje fragmenty všech peptidů izolovaných v daném hmotnostním okně
- zachovaná informace o všech peptidech
- **kvantitativní informace uložena v MS2**
- **detekce zhruba 8 000 proteinů z lyzátu savčích buněk při 60 min gradientu**
- **přesnější než DDA – CV kolem 5 – 10 %**
- menší počet missing values
    - DDA – 30 – 40 %
    - **DIA – 15 %**
- MS2 spektra výrazně složitější než u DDA – potřeba spektrálních knihoven

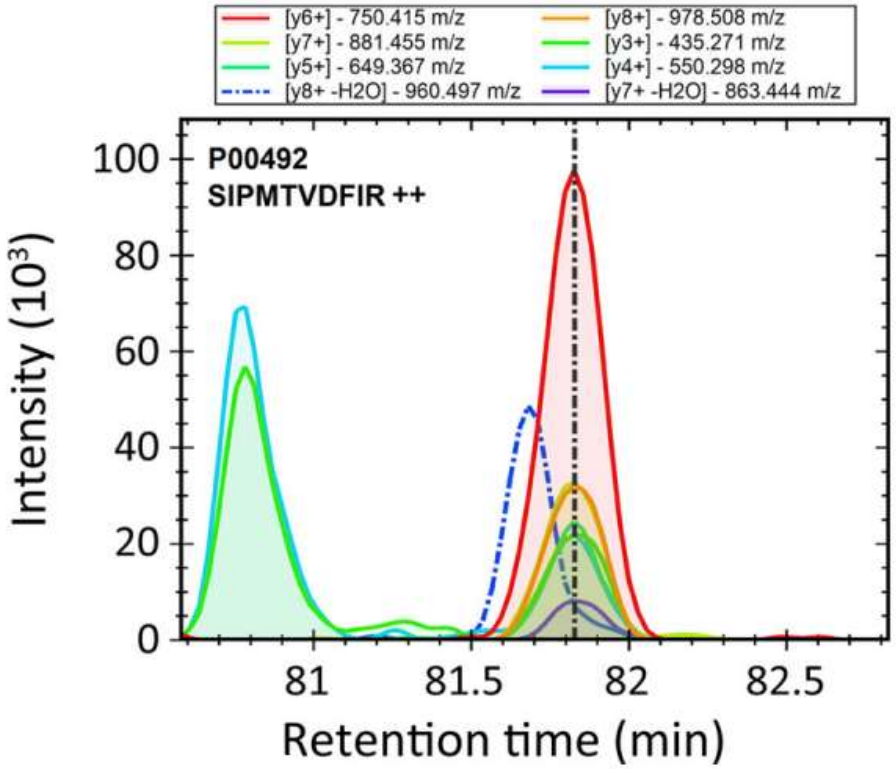


Pappireddi et al, **A Review on Quantitative Multiplexed Proteomics, ChemBioChem 2019**

Kvantifikace v DIA

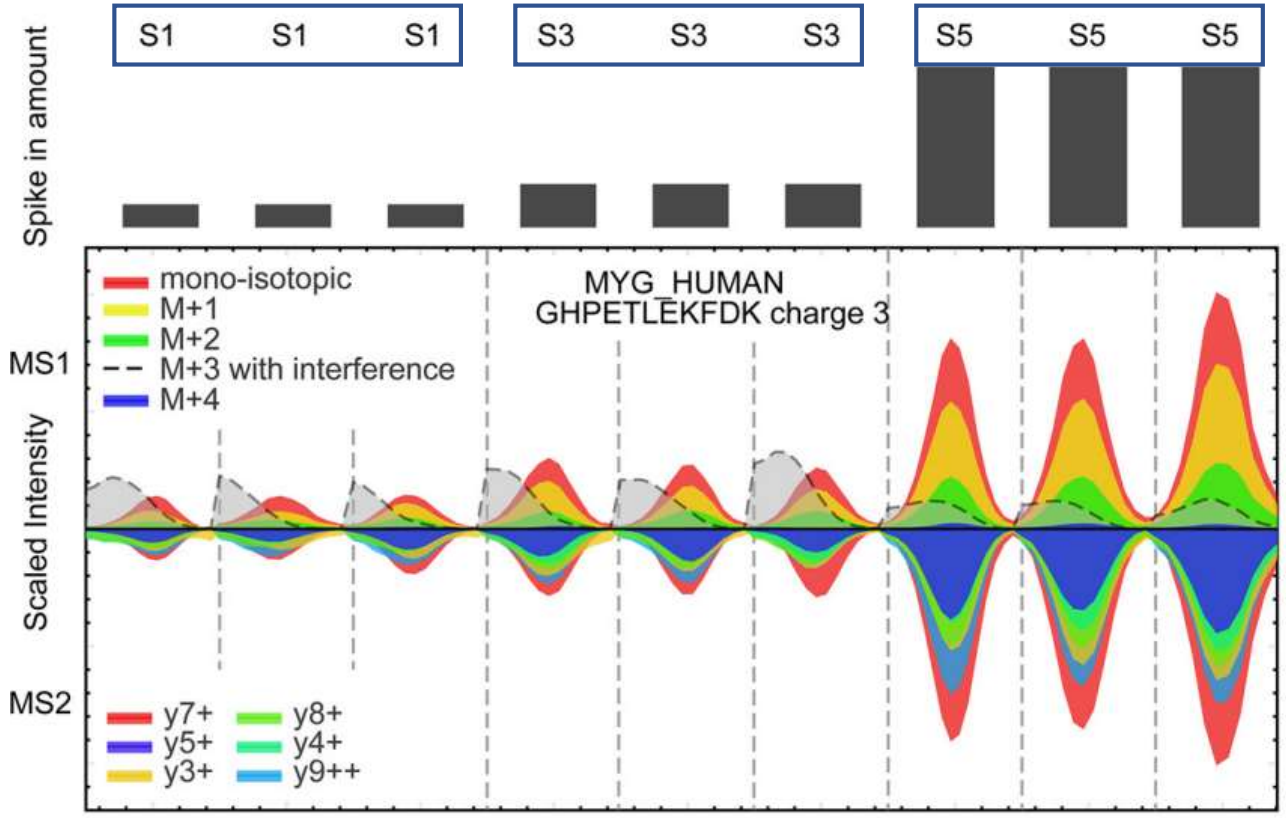
Fragmenty s překrývajícími se elučními profily pochází pravděpodobně z jednoho peptidu.

Interferující signál nekopíruje přesně eluční profil.



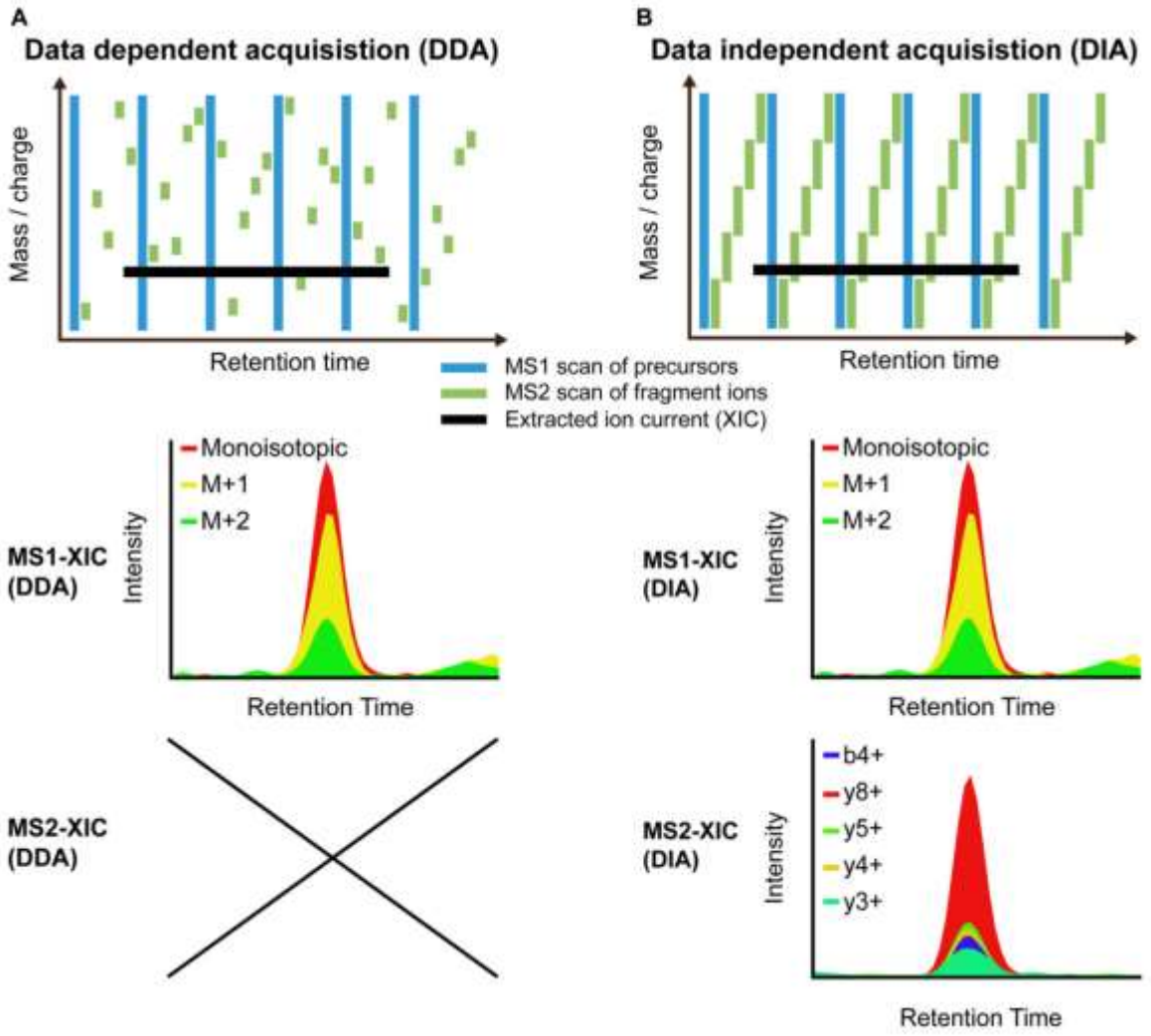
Bruderer et al, Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues, Mol Cell Proteomics, 2015

Kvantifikace v DIA



Huang et al, Combining Precursor and Fragment Information for Improved Detection of Differential Abundance in Data Independent Acquisition, Mol Cell Proteomics, 2020

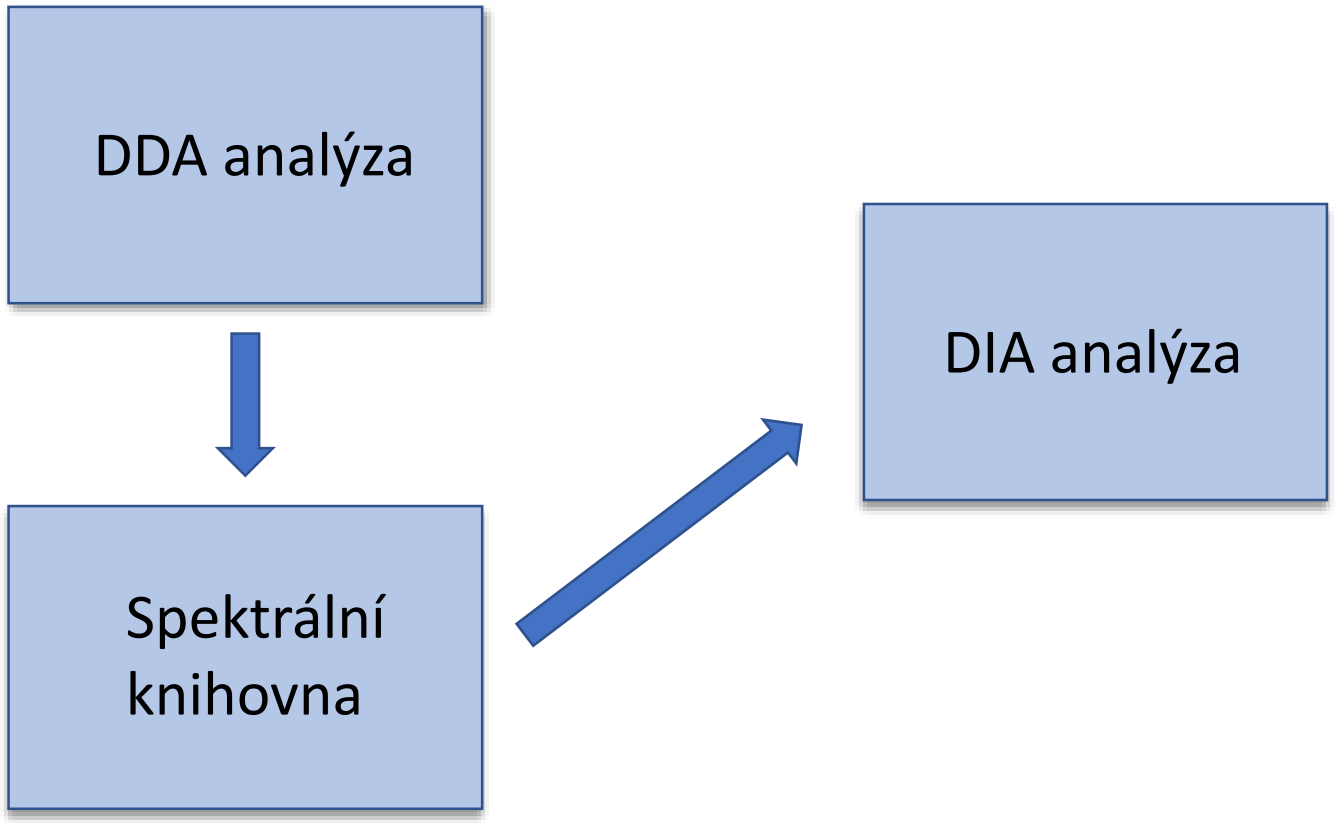
Kvantifikace v DDA vs DIA



Huang et al, Combining Precursor and Fragment Information for Improved Detection of Differential Abundance in Data Independent Acquisition, Mol Cell Proteomics, 2020

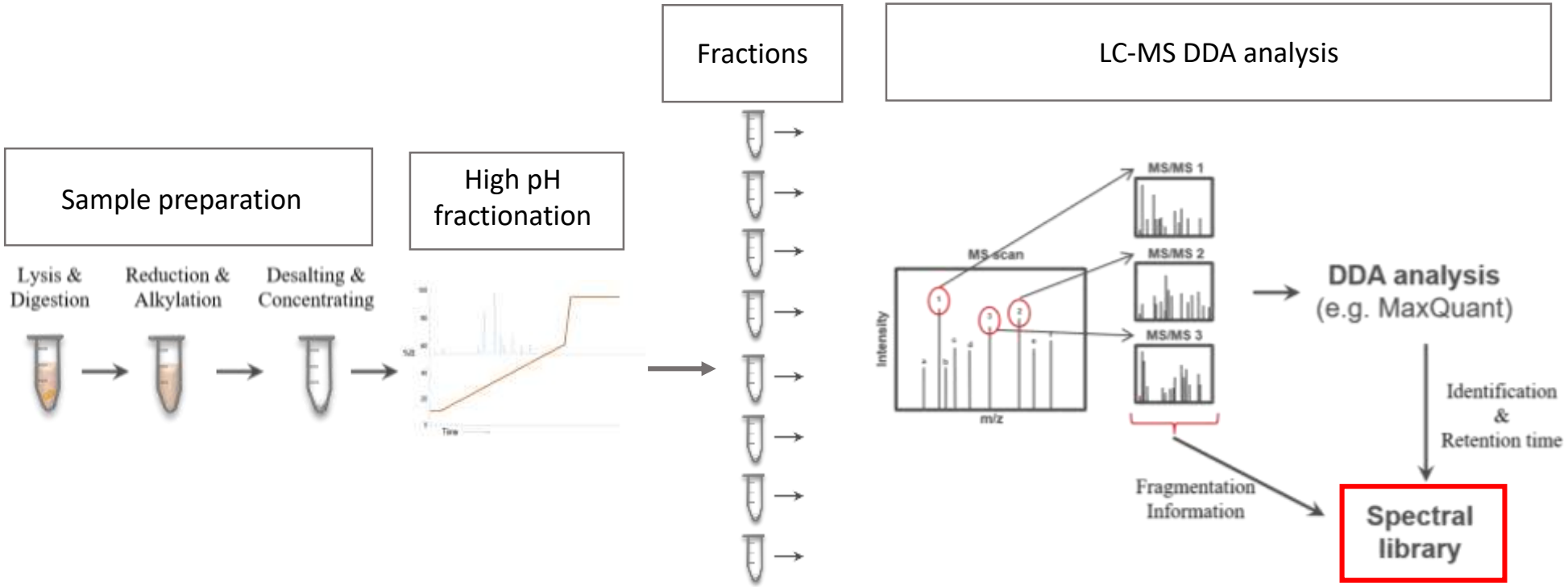
DIA – spektrální knihovny

- Spektrální knihovna
  - obsahuje sekvence peptidů a jejich MSMS spektra
  - měla by být specifická pro daný experiment (např. konkrétní buněčná linie)
  - získáváme pomocí separátní DDA analýzy



DIA – spektrální knihovny

- Spektrální knihovna
  - pro co nejvyšší pokrytí peptidů můžeme vzorek frakcionovat pomocí 2D chromatografie – rozdělení do X frakcí -> **snížení komplexity**
  - 2 následující chromatografické separace – např. C18 v kyselém a zásaditém pH

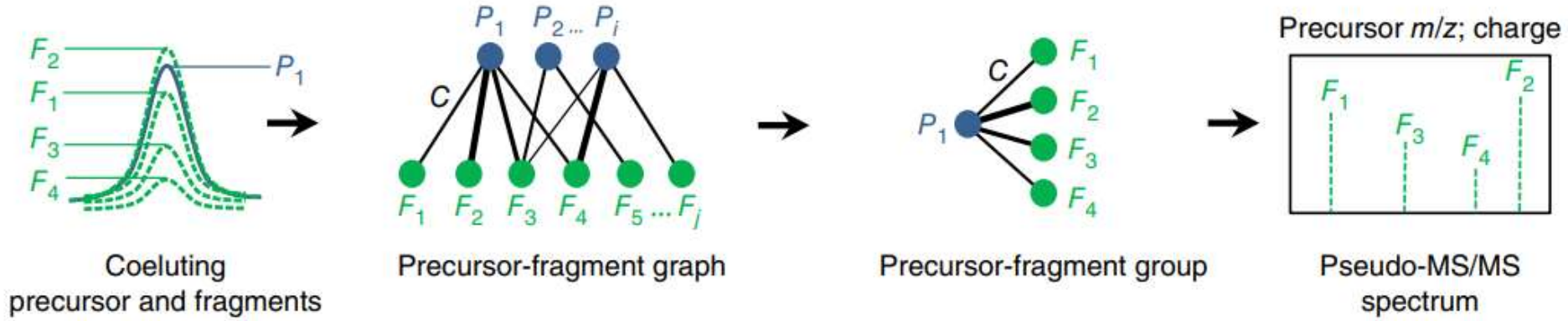


DIA bez spektrálních knihoven

- umožňuje vyhodnocení DIA dat bez potřeby separátního měření spektrálních knihoven

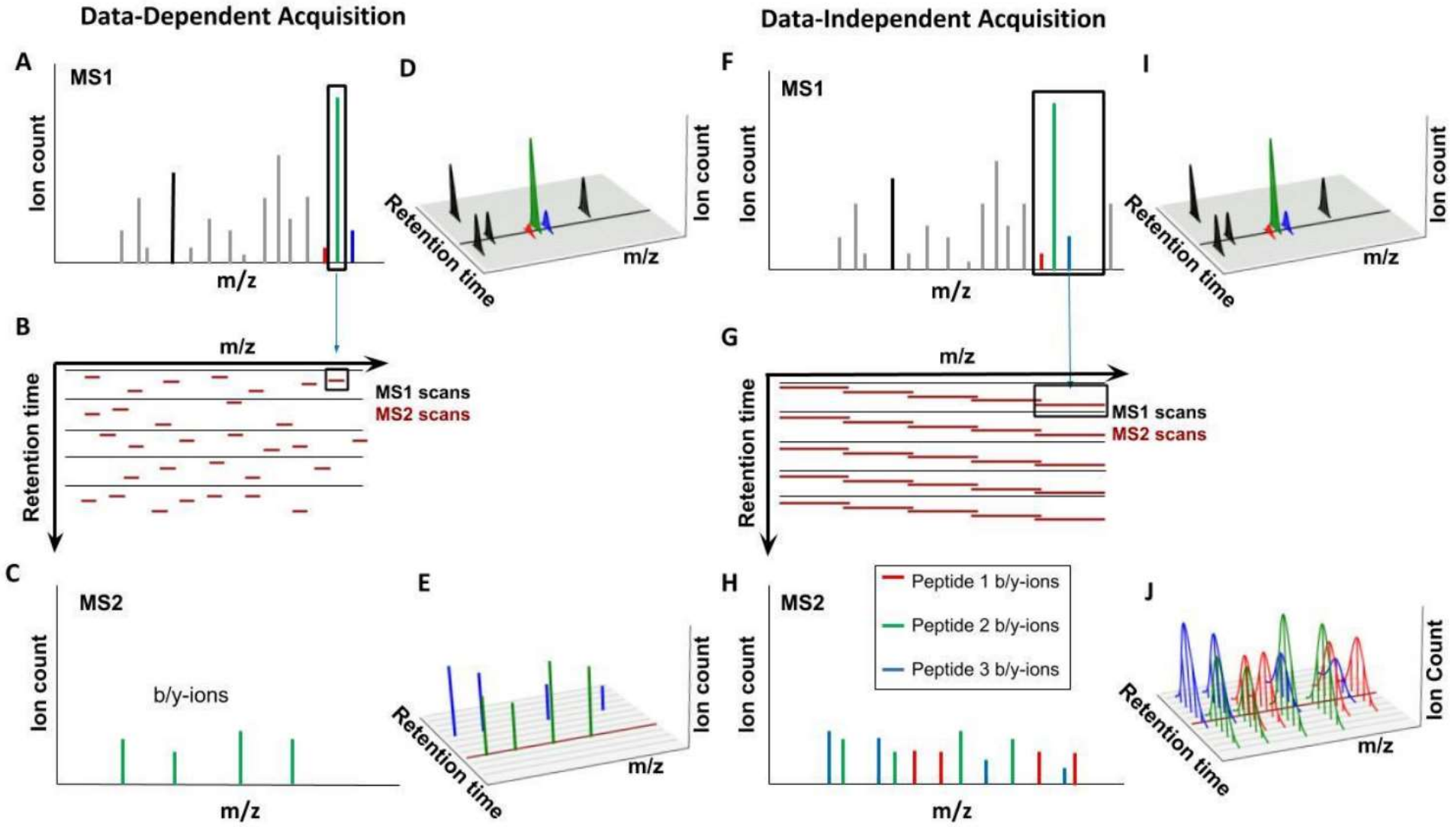
- DDA-like search – v datech jsou hledány spojitosti mezi prekurzorem a fragmenty pomocí machine learning algoritmů

- spektrální knihovna vytvořena *in silico* na základě DIA dat a FASTA databáze



Tsou et al, DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics, Nature Methods 2015

DDA vs. DIA - shrnutí



Pappireddi et al, **A Review on Quantitative Multiplexed Proteomics, ChemBioChem 2019**

DDA vs. DIA - shrnutí

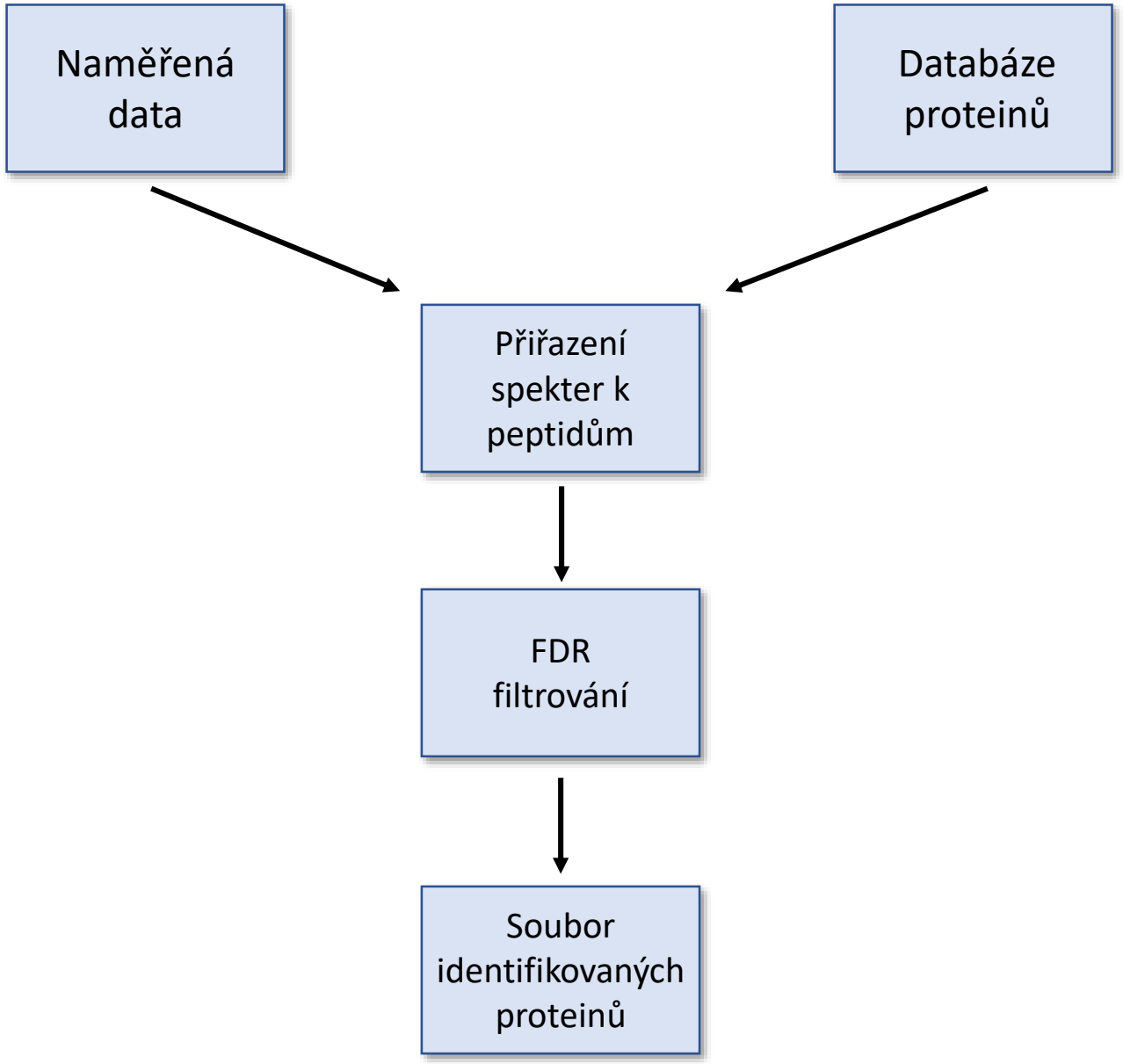
	DDA	DIA
Počet identifikací z lidského lyzátu:	5 000 (120 min gradient)	8 000 (60 min gradient)
Kvantifikace na základě:	MS1	MS2
Potřeba spektrální knihovny:	Ne	Ano/Ne
Počet missing values:	30-40 %	15 %
Přesnost kvantifikace (CV):	25 %	5 – 10 %

DDA vs. DIA - shrnutí

	DDA	DIA
Počet identifikací z lidského lyzátu:	5 000	8 000
Kvantifikace na základě:	MS1	MS2
Potřeba spektrální knihovny:	Ne	Ano/Ne
Počet missing values:	30-40 %	15 %
Přesnost kvantifikace (CV):	25 %	5 – 10 %

Od spekter k souboru proteinů

Jak z 250 000 spekter získat 5000 proteinů?


Naměřená data
Databáze proteinů
Přiřazení spekter k peptidům
FDR filtrování
Soubor identifikovaných proteinů

Databáze

- Vedle naměřených dat **potřebujeme** k vyhodnocení databázi proteinů – **neprobíhá de novo sekvenace**

- FASTA formát: hlavička strukturovaná tak, aby se z ní daly automaticky extrahovat informace pomocí Regex pravidel

>sp|P61206|ARF3_RAT ADP-ribosylation factor 3 OS=Rattus norvegicus OX=10116 GN=Arf3 PE=2 SV=2
MGNIFGNLLKSLIGKKEMRILMVGLDAAGKTTILYKLKLGEIVTTIPTIGFNVETVEYKNISFTVWDVGGQDKIRPLWRHYFQ
NTQGLIFVVDSNDRERVNEAREELMRMLAEDELRDAVLLVFANKQDLPNAMNAAEITDKLGLHSLRHRNWYIQATCATSG
DGLYEGLDWLANQLKNKK

Databáze

- Vedle naměřených dat **potřebujeme** k vyhodnocení databázi proteinů – **neprobíhá de novo sekvenace**

- FASTA formát: hlavička strukturovaná tak, aby se z ní daly automaticky extrahovat informace pomocí Regex pravidel

>sp|P61206|ARF3_RAT ADP-ribosylation factor 3 OS=Rattus norvegicus OX=10116 GN=Arf3 PE=2 SV=2
MGNIFGNLLKSLIGKKEMRILMVGLDAAGKTTILYKLKLGEIVTTIPTIGFNVETVEYKNISFTVWDVGGQDKIRPLWRHYFQ
NTQGLIFVVDSNDRERVNEAREELMRMLAEDELRDAVLLVFANKQDLPNAMNAAEITDKLGLHSLRHRNWYIQATCATSG
DGLYEGLDWLANQLKNKK

Databáze

- Vedle naměřených dat **potřebujeme** k vyhodnocení databázi proteinů – **neprobíhá de novo sekvenace**

- FASTA formát: hlavička strukturovaná tak, aby se z ní daly automaticky extrahovat informace pomocí Regex pravidel

>sp|P61206|ARF3_RAT ADP-ribosylation factor 3 OS=Rattus norvegicus OX=10116 GN=Arf3 PE=2 SV=2

- Obsah databáze by vždy měl odpovídat pravděpodobnému obsahu vzorku = vzorky lidských buněčných linií prohledávat proti databázi lidských proteinů, ne proti databázi všech savců nebo eukaryot

- Použití příliš velké databáze znehodnocuje výsledky

- **I špatná databáze dá vždy nějaký výsledek**

I špatná databáze dá vždy nějaký výsledek!

Vyhodnocení vzorku z lidských Hela buněk proti dvěma různým databázím:

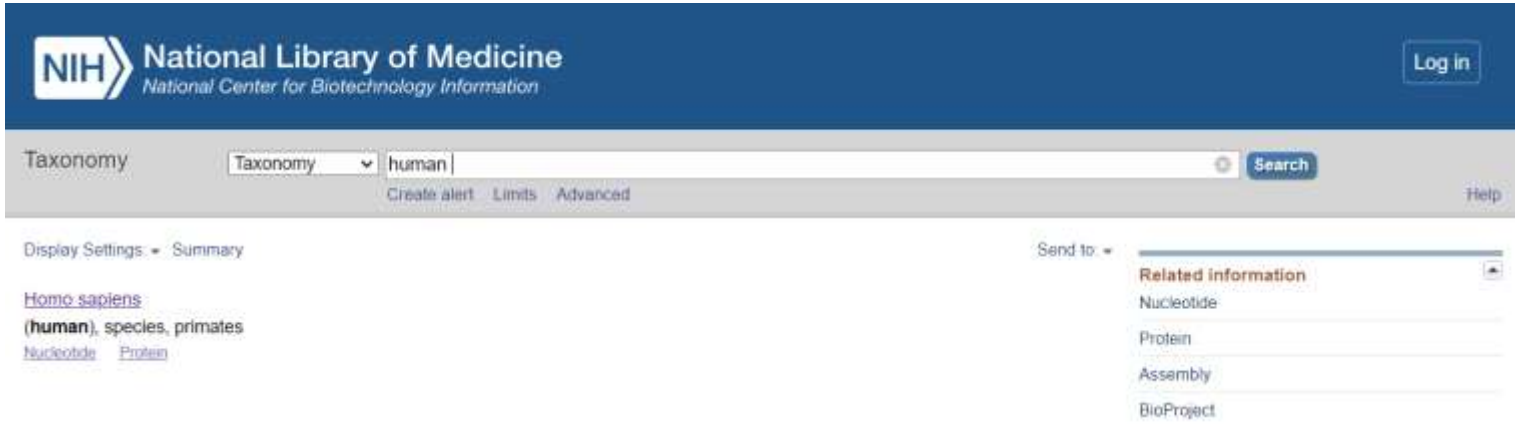
Databáze	Human	Mus musculus
Počet identifikovaných proteinů	5 352	3 634

Když použijeme databázi myši k vyhodnocení dat z lidského vzorku, dostaneme přes 3600 identifikací!

Databáze

- **Zdroje databází:**
  - Uniprot.org

- NCBI - National Center for Biotechnology Information

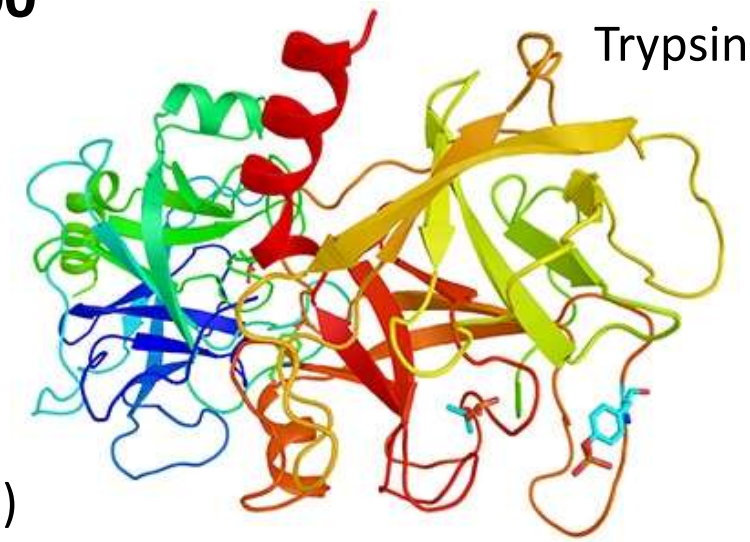


Databáze

- Parametry, které je nutné definovat spolu s použitou databází:
  - Použitá proteasa
  - Modifikace

Databáze – proteasy

- Proč je nutné uvést proteasu použitou k digesci (kombinace proteas)?
  - Počet všech teoretický peptidů o délce 8 – 14 AMK generovaných na základě databáze lidských proteinů je asi **70 000 000**
  - Počet tryptických peptidů o stejných parametrech je asi **900 000**



Trypsin

- Miscleavage
  - Proteasa dané štěpné místo někdy vynechá (sterická blokace, přítomnost modifikací)
  - Specifikujeme kolik vynechání (miscleavage) chceme tolerovat (obvykle 1-2)

Databáze - modifikace

- Společně s databází proteinů a použitou proteasou musíme specifikovat modifikace, které lze ve vzorku očekávat

- **Modifikace obecně se vyskytující ve všech vzorcích:**
  - Modifikace cysteinu – vznik modifikací volných –SH během přípravy vzorku. Záleží na alkylačním činidle (IAA, CAA – carbamidomethyl)
  - Oxidace methioninu
  - Acetylace N-konce

- **Modifikace specifické pro vzorek:**
  - Fosforylace, acetylace, methylace, hydroxylace, ubikvitinylace (GlyGly)

- **Fixní** – počítá se, že modifikace je na dané aminokyselině přítomna vždy (alkylace Cys)
- **Variabilní** – modifikace se na dané aminokyselině vyskytovat může a nemusí

- Software následně generuje varianty peptidů obsahující všechny možné kombinace zadaných modifikací

Databáze - modifikace

Peptid LVSCAGTFK:

1 fixní

LVSC(cam)AGTFK

1 sekvence

Databáze - modifikace

Peptid LVSCAGTFK:

1 fixní

LVSC(cam)AGTFK

1 sekvence

1 fixní + 1 variabilní

LVSC(cam)AGTFK

LVS(ph) C(cam)AGTFK

LVSC(cam)AGT(ph)FK

LVS(ph) C(cam)AGT(ph)FK

4 sekvence

Databáze - modifikace

Peptid LVSCAGTFK:

1 fixní

LVSC(cam)AGTFK

1 sekvence

1 fixní + 1 variabilní

LVSC(cam)AGTFK

LVS(ph) C(cam)AGTFK

LVSC(cam)AGT(ph)FK

LVS(ph) C(cam)AGT(ph)FK

4 sekvence

1 fixní + 2 variabilní

LVSC(cam)AGTFK

LVS(ph) C(cam)AGTFK

LVSC(cam)AGT(ph)FK

LVS(ph) C(cam)AGT(ph)FK

LVS(ph) C(cam)AGTFK(me)

LVSC(cam)AGT(ph)FK(me)

LVS(ph) C(cam)AGT(ph)FK(me)

LVSC(cam)AGTFK(me)

8 sekvencí

Databáze - modifikace

Peptid LVSCAGTFK:

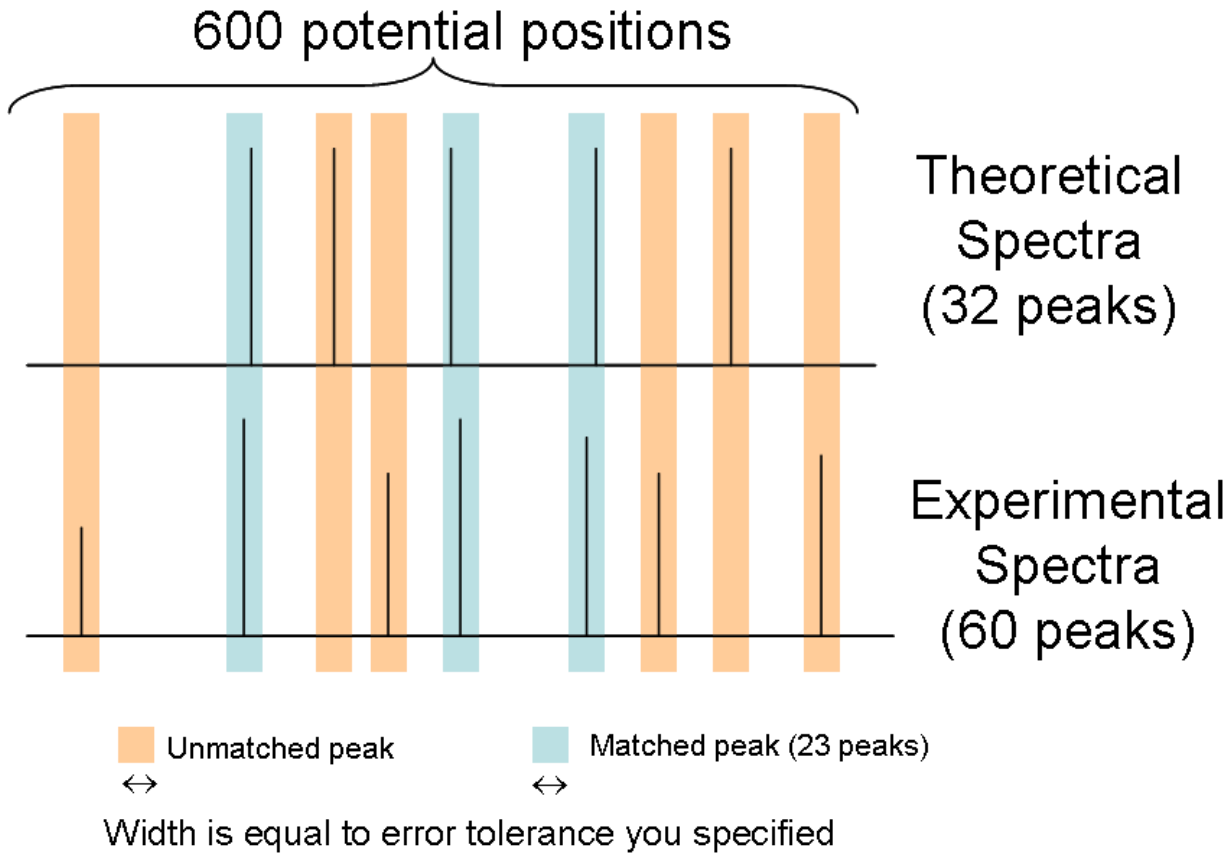
1 fixní	1 fixní + 1 variabilní	1 fixní + 2 variabilní
LVSC(cam)AGTFK	LVSC(cam)AGTFK	LVSC(cam)AGTFK
	LVS(ph) C(cam)AGTFK	LVS(ph) C(cam)AGTFK
	LVSC(cam)AGT(ph)FK	LVSC(cam)AGT(ph)FK
	LVS(ph) C(cam)AGT(ph)FK	LVS(ph) C(cam)AGT(ph)FK
		LVS(ph) C(cam)AGTFK(me)
		LVSC(cam)AGT(ph)FK(me)
		LVS(ph) C(cam)AGT(ph)FK(me)
		LVSC(cam)AGTFK(me)

- Variabilní modifikace výrazně expandují databázi a navyšují tzv. search space a tím i výpočetní náročnost. Použití modifikací, které se ve vzorku nevyskytují nebo se v něm vyskytují jen sporadicky, znehodnocuje výsledek.

- **Vždy by měly být specifikovány jen ty modifikace, které se ve vzorku opravdu a ve výrazné míře vyskytují.**

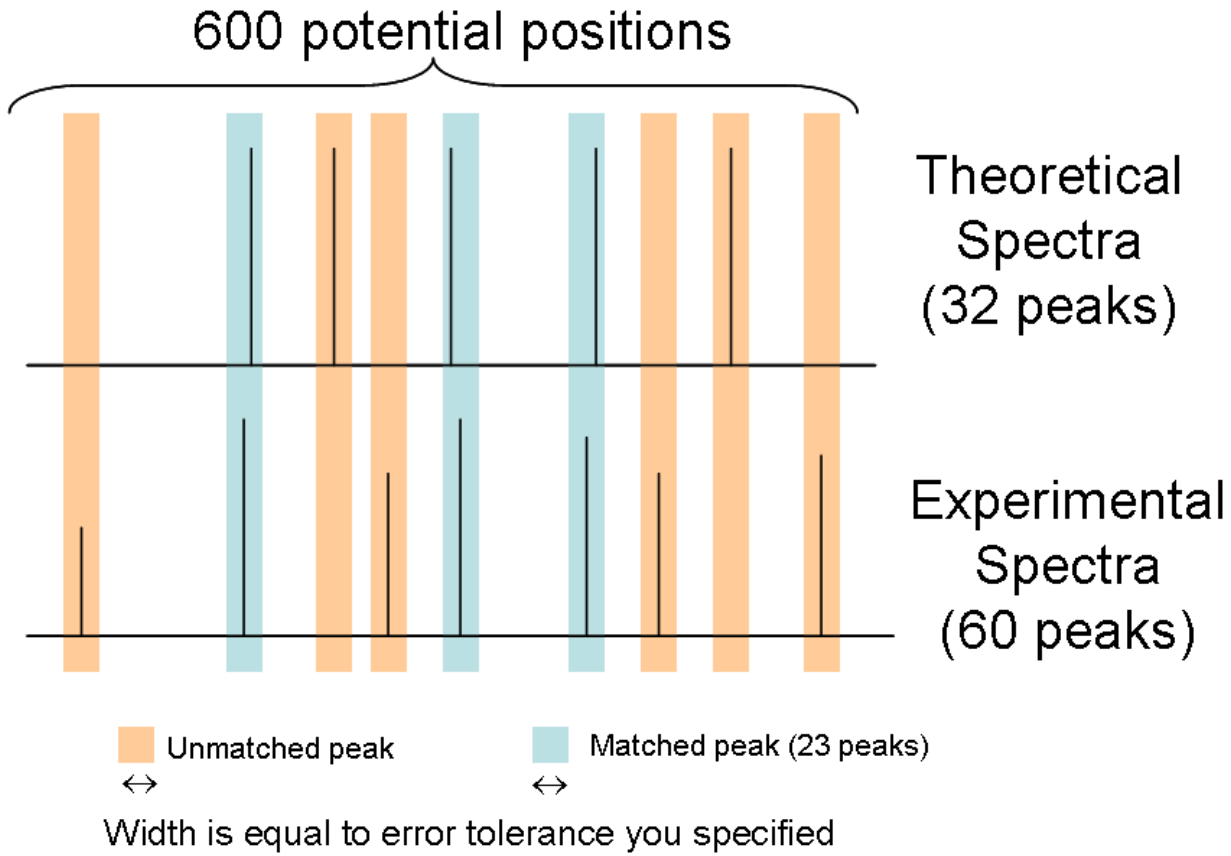
Vyhodnocení spekter a skórování

1. Na základě zadané databáze vygenerujeme všechny teoretické peptidy
2. Generování teoretických spekter
3. Hledáme počet shod mezi teoretickým a naměřeným spektrem
4. Jaká je pravděpodobnost, že nalezená shoda je náhodná?
5. Vyčíslíme pomocí **skóre** (obvykle –Log)



Vyhodnocení spekter a skórování

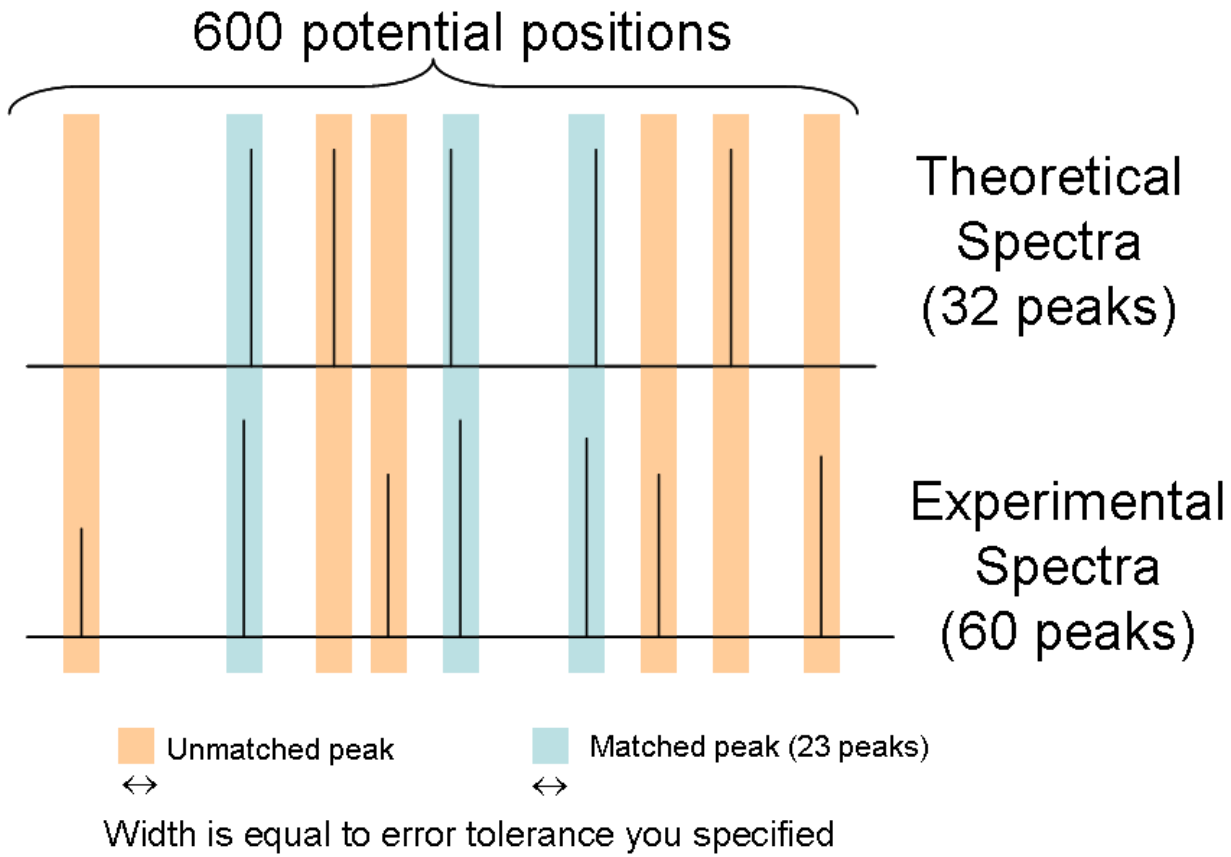
1. Na základě zadané databáze vygenerujeme všechny teoretické peptidy
2. Generování teoretických spekter
3. Hledáme počet shod mezi teoretickým a naměřeným spektrem
4. Jaká je pravděpodobnost, že nalezená shoda je náhodná?
5. Vyčíslíme pomocí **skóre** (obvykle –Log)



$$s(q,\text{loss}) = -10\log_{10}\sum_{j=k}^{n}\left[\binom{n}{j}\left(\frac{q}{100}\right)^{j}\left(1-\frac{q}{100}\right)^{n-j}\right]$$

Vyhodnocení spekter a skórování

Skóre = pravděpodobnost, že shoda naměřených a teoretických fragmentů je náhodná

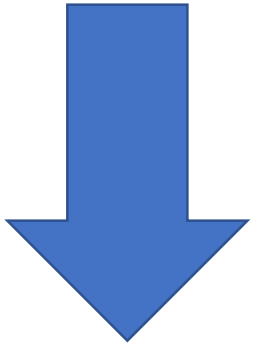


$$s(q,\text{loss}) = -10\log_{10}\sum_{j=k}^{n}\left[\binom{n}{j}\left(\frac{q}{100}\right)^{j}\left(1-\frac{q}{100}\right)^{n-j}\right]$$

False Discovery Rate

Samotné skóre nestačí pro zajištění dostatečné spolehlivosti identifikací.

Při porovnávání vysokého množství spekter v řádech statisíců dostáváme určitou frakci falešně pozitivních přiřazení.



Filtrování pomocí odhadu False Discovery Rate (FDR)

False Discovery Rate

$$\text{FDR} = \frac{D}{D+T}$$

D – počet hitů z decoy databáze nad tresholdem
T – počet hitů u target databáze

Pokud nastavíme treshold FDR na 1%, dostaneme dataset obsahující 99% správných identifikací a 1% falešně pozitivních.

FDR – target-decoy přístup

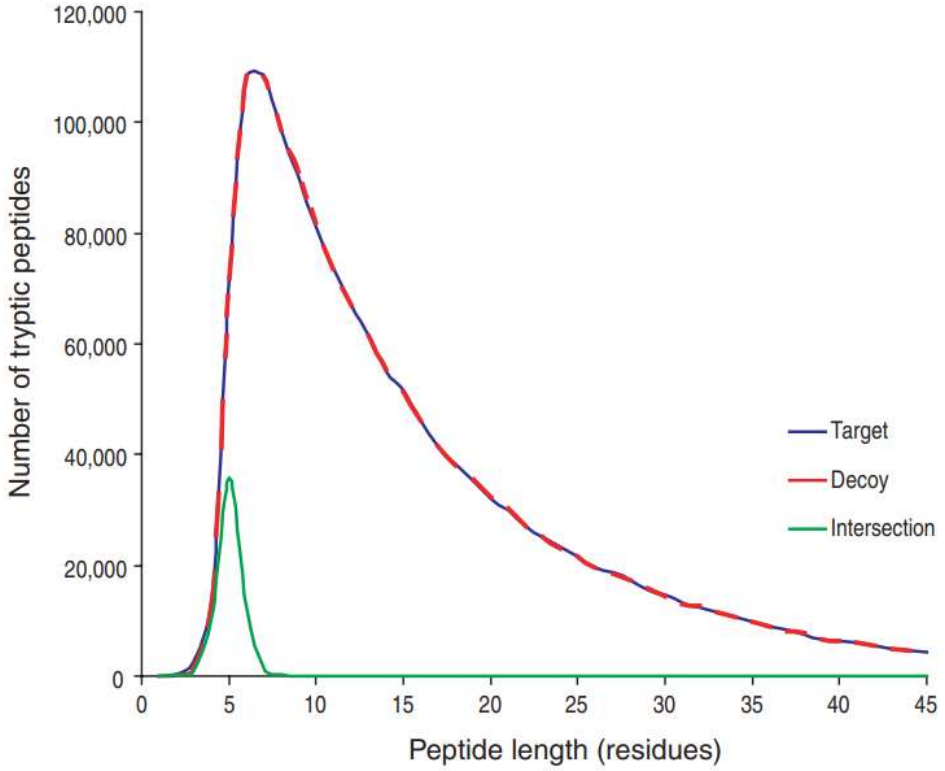
1. Vygenerujeme databázi neexistujících proteinů (návnad) otočením všech sekvencí v originální (target) databázi = **decoy databáze**

>sp|P01308|INS_HUMAN Insulin OS=Homo sapiens OX=9606 GN=INS PE=1 SV=1
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPL
ALEGSLQKRGIVEQCCTSICSLYQLENYCN

>REV:sp|P01308|INS_HUMAN Insulin OS=Homo sapiens OX=9606 GN=INS PE=1 SV=1
NCYNELQYLSCISTCCQEVIGRKQLSGELALPQLSGAGPGGGLEVQGVQLDEAERRTKPTYFFGREGCVLYLAEVLHSGCL
HQNVFAAAPDPGWLALLALLPLLRMWLAM

Předpoklady decoy databáze:

- **Decoy databáze obsahuje identické množství proteinů jako target databáze**
- **Distribuce délek peptidů je identická u obou databází**
- **Sekvence peptidů z target a decoy databáze se nepřekrývají (platí od zhruba 7 - 8 AMK)**



Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, Nature methods 2007

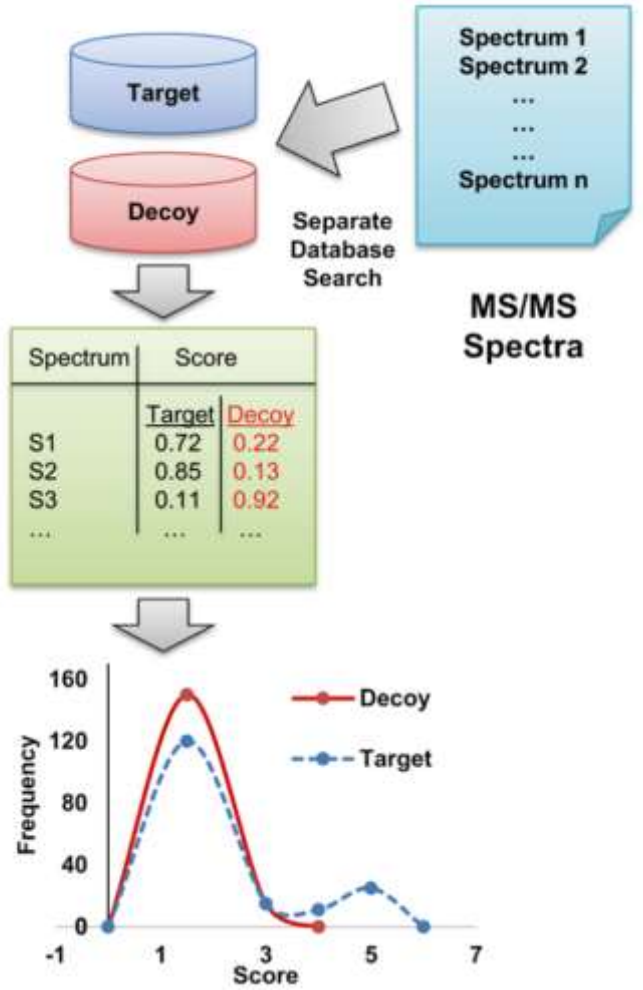
FDR – target-decoy přístup

1. Vygenerujeme databázi neexistujících proteinů (návnad) otočením všech sekvencí v databázi = **decoy databáze**

2. Naměřená spektra porovnáme jak proti smysluplné databázi (target), tak proti decoy databázi

3. Každé shodě je přiřazeno skóre.

4. V datasetu ponecháme jen identifikace odpovídající určité FDR (obvykle 1%)



$$\text{FDR} = \frac{D}{D+T}$$

D – počet hitů z decoy databáze nad tresholdem
T – počet hitů u target databáze

Klaus Jung (ed.), Statistical Analysis in Proteomics, Methods in Molecular Biology, vol. 1362, DOI 10.1007/978-1-4939-3106-4_7, © Springer Science+Business Media New York 2016

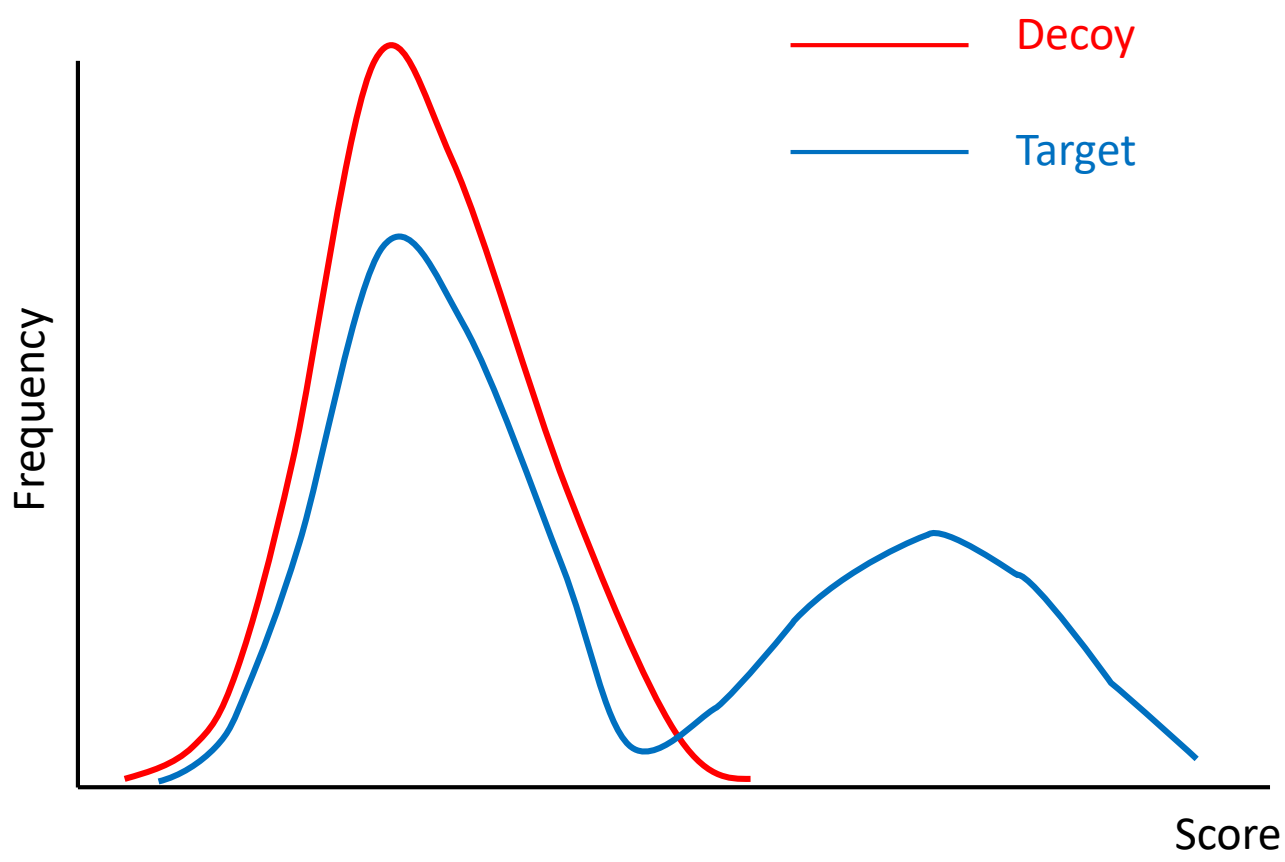
FDR – target-decoy přístup

1. Vygenerujeme databázi neexistujících proteinů (návnad) otočením všech sekvencí v databázi = **decoy databáze**

2. Naměřená spektra porovnáme jak proti smysluplné databázi (target), tak proti decoy databázi

3. Každé shodě je přiřazeno skóre.

4. V datasetu ponecháme jen identifikace odpovídající určité FDR (obvykle 1%)



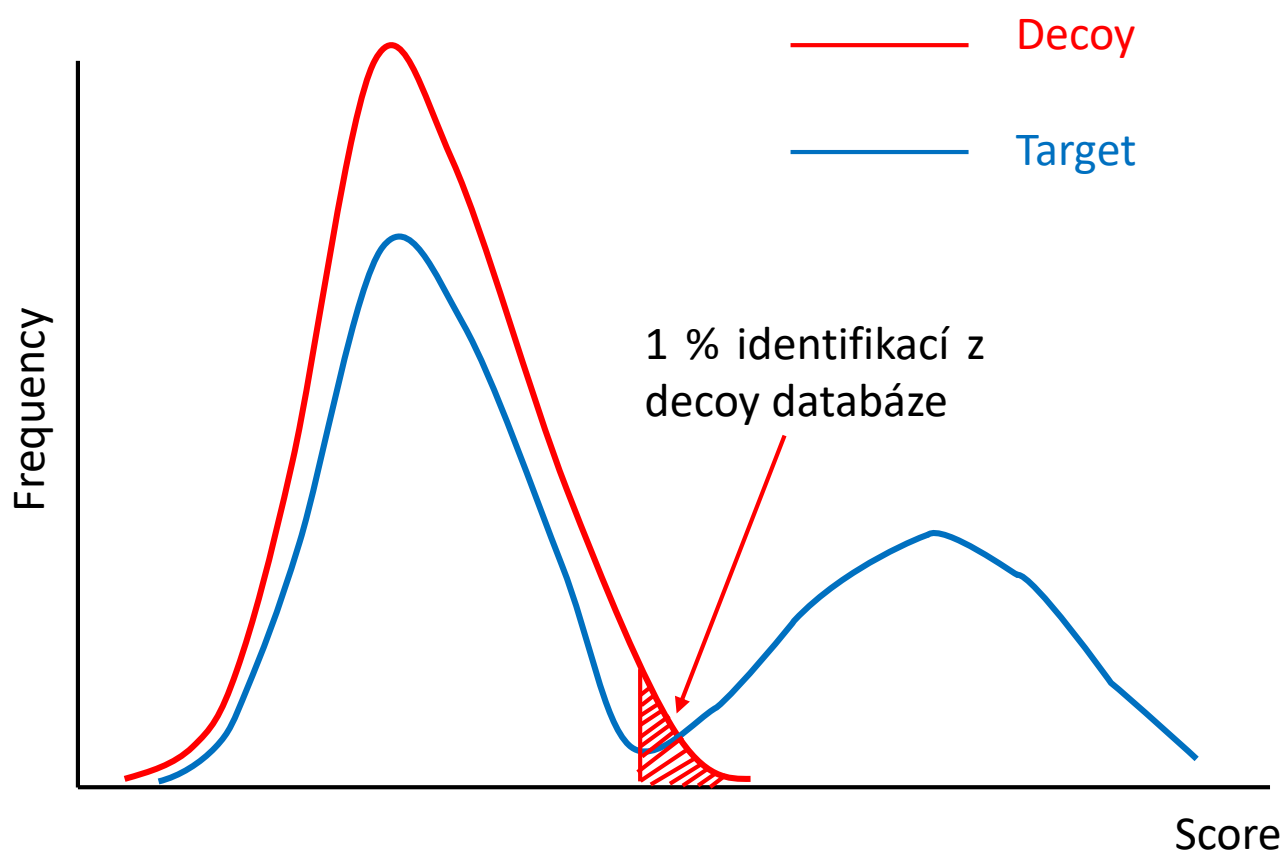
FDR – target-decoy přístup

1. Vygenerujeme databázi neexistujících proteinů (návnad) otočením všech sekvencí v databázi = **decoy databáze**

2. Naměřená spektra porovnáme jak proti smysluplné databázi (target), tak proti decoy databázi

3. Každé shodě je přiřazeno skóre.

4. V datasetu ponecháme jen identifikace odpovídající určité FDR (obvykle 1%)



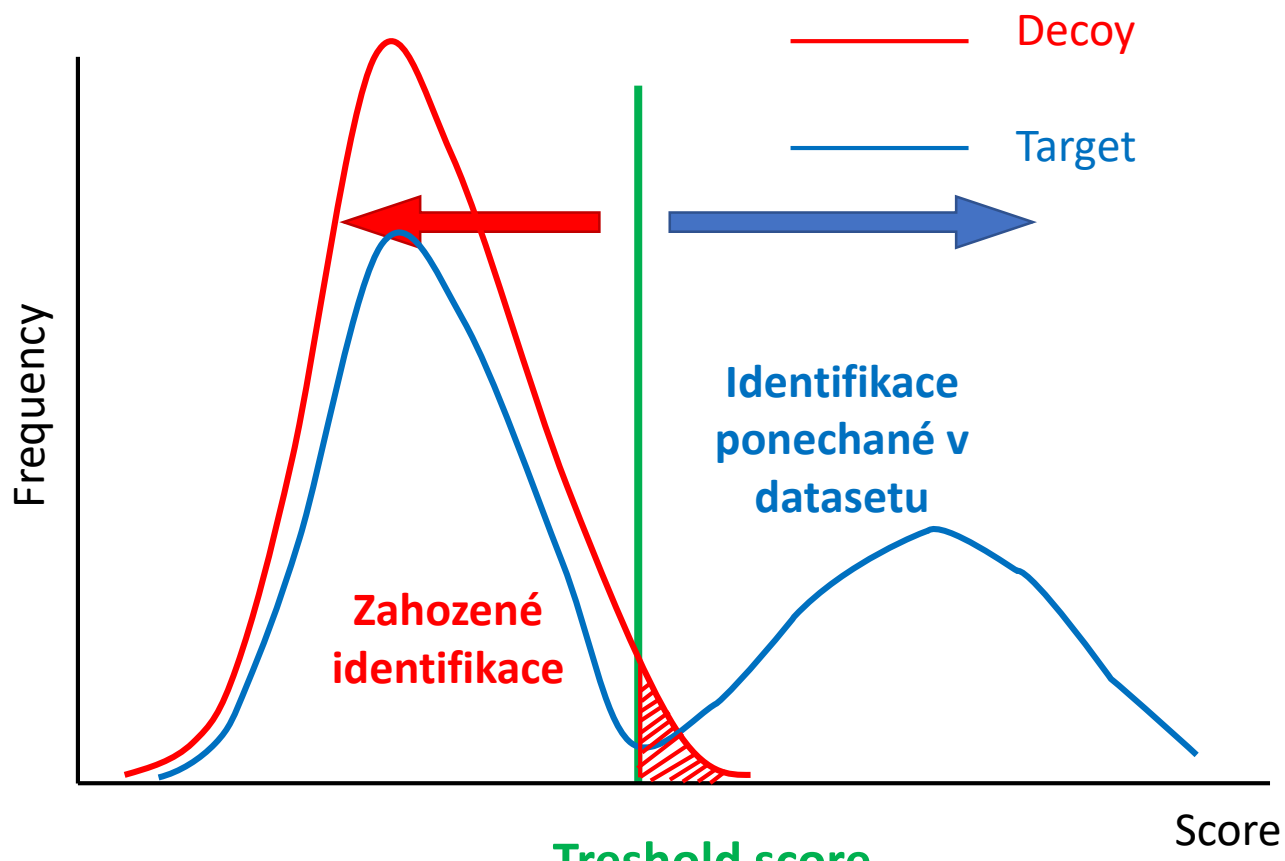
FDR – target-decoy přístup

1. Vygenerujeme databázi neexistujících proteinů (návnad) otočením všech sekvencí v databázi = **decoy databáze**

2. Naměřená spektra porovnáme jak proti smysluplné databázi (target), tak proti decoy databázi

3. Každé shodě je přiřazeno skóre.

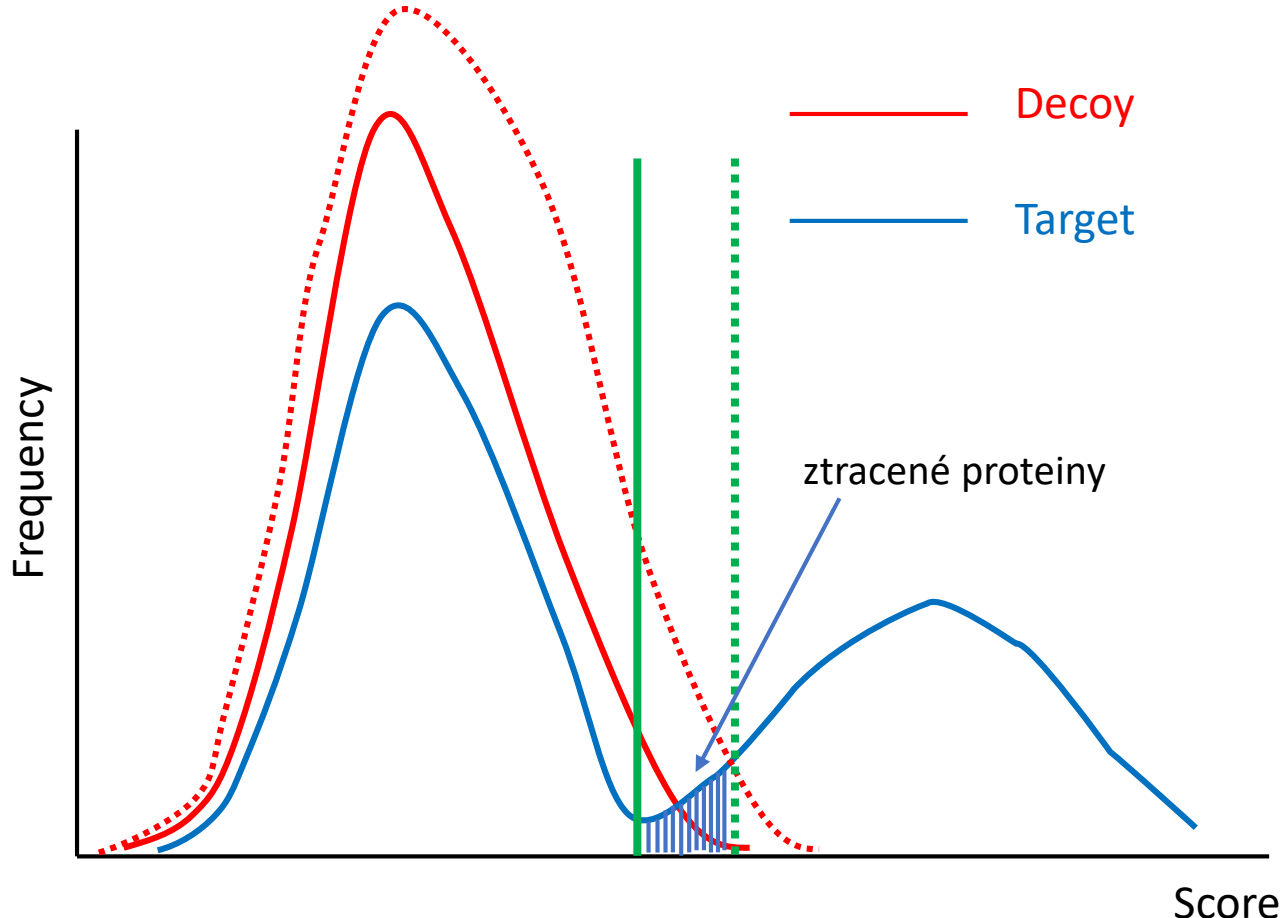
4. V datasetu ponecháme jen identifikace odpovídající určité FDR (obvykle 1%)



FDR – target-decoy přístup

Ovlivnění výsledků :

- použitím nespecifické nebo redundantní databáze
- definováním neúměrného počtu variabilních modifikací se zanedbatelným výskytem ve vzorku



Q-value

- asociovaná vždy s konkrétním PSM
- **minimální hodnota FDR, při které daný peptid projde FDR filtrováním**
- pokud má peptid *q*-value 0.01, znamená to, že filtrováním projde, pokud bude nastavena FDR alespoň na 1% (při FDR 0.5% by byl vyřazen)
- **vždy závisí na konkrétním vyhodnocení – stejný peptid bude mít různou *q*-value ve dvou různých experimentech**

Q-value

- **Příklad 1:**
  1. Data z lidského vzorku prohledáme proti **NESPECIFICKÉ DATABÁZI VŠECH EUKARYOT**
  2. Seřadíme identifikované peptidy podle skóre od nejvyššího
  3. Peptid EAMRPK je na pozici 100
  4. Mezi těmito 100 peptidy je 10 falešně pozitivních
  5. ***q*-value EAMRPK je 0.1**

- **Příklad 2:**
  1. Stejná data z lidského vzorku nyní prohledáme proti **SPECIFICKÉ LIDSKÉ DATABÁZI**
  2. Seřadíme identifikované peptidy podle skóre od nejvyššího
  3. Z původních 100 vypadlo 20 identifikací z důvodu menší databáze. Z toho 6 bylo falešně pozitivních.
  4. EAMRPK je na pozici 80.
  5. Mezi těmito 80 peptidy jsou 4 falešně pozitivní.
  6. ***q*-value EAMRPK je 4/80 = 0.05**

Q-value

Příklad 1 vs příklad 2:

stejné spektrum **dvě různé databáze**

stejný peptid **dvě různé q-value**

stejné skóre

Dva různé výsledky!

- **Zda peptid projde filtrováním je ovlivněno:**
  - použitou databází – čím specifičtější pro daný organismus, tím lépe
  - množstvím zadaných variabilních modifikací – expandují search space podobně jako nespecifická databáze

Příklady software pro necílenou proteomiku

DDA

MaxQuant – www.maxquant.org
FREE
www.youtube.com/@MaxQuantChannel

Proteome Discoverer
ThermoFisher Scientific

DIA

DiaNN - https://github.com/vdemichev/DiaNN
FREE

Spectronaut
Biognosys