SCIENCE MEETS LIFE

Generating high-quality libraries for DIA datasets

Ralf Gabriels

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

CompOmics

16/11/2022

EMBO Course Targeted Proteomics

Overview

Why use predicted libraries for DIA

Practical considerations

Benchmarking results

Future development

Take home messages

Overview

Why use predicted libraries for DIA

Practical considerations

Benchmarking results

Future development

Take home messages

Mass spectrometry data usually contains three dimensions: RT, m/z and intensity

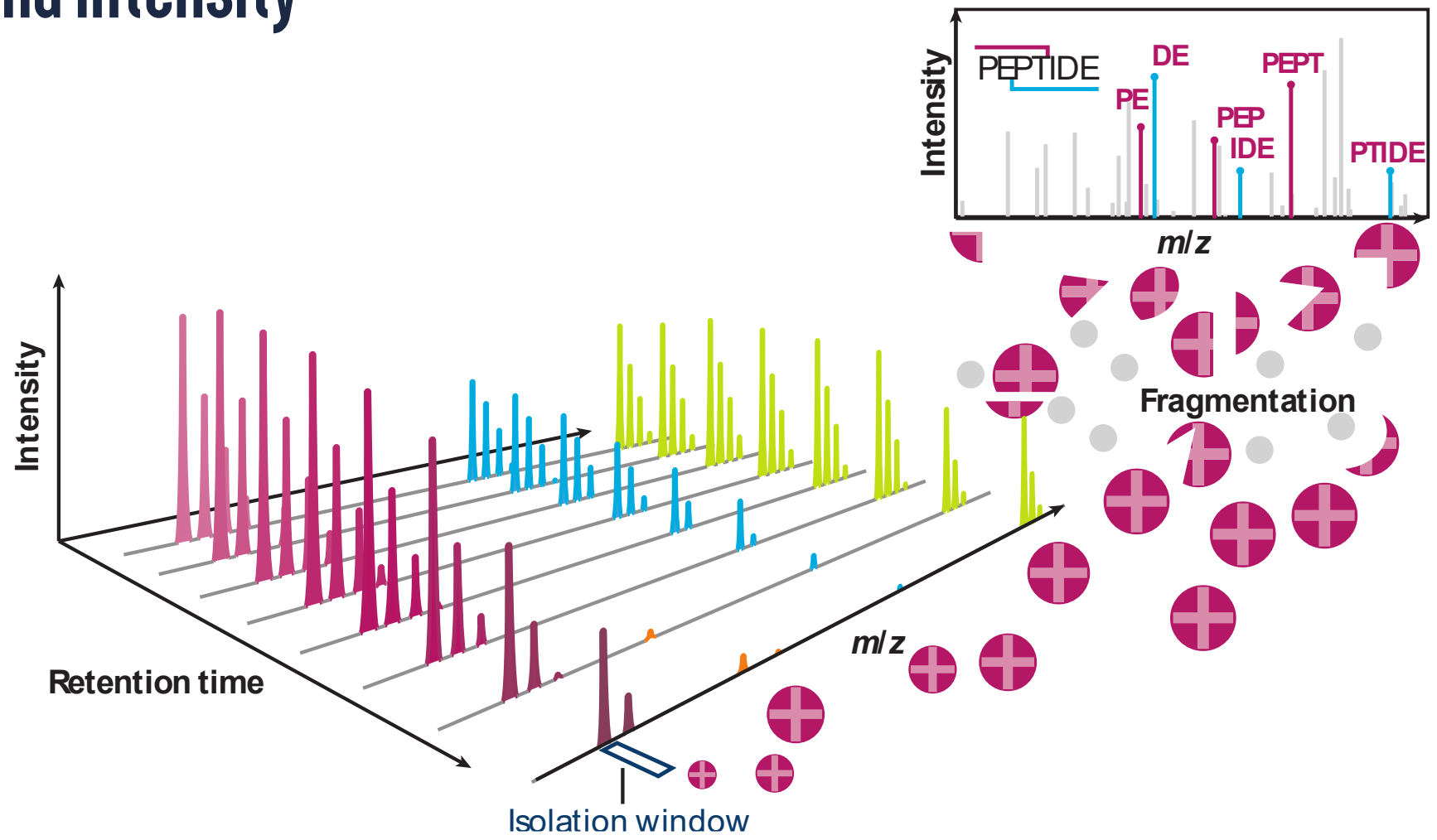
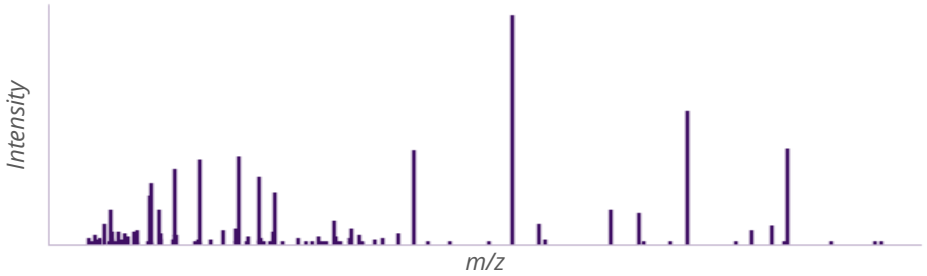


Figure adapted from Pavel Sinitcyn et al. Annu. Rev. Biomed. Data Sci. (2018)

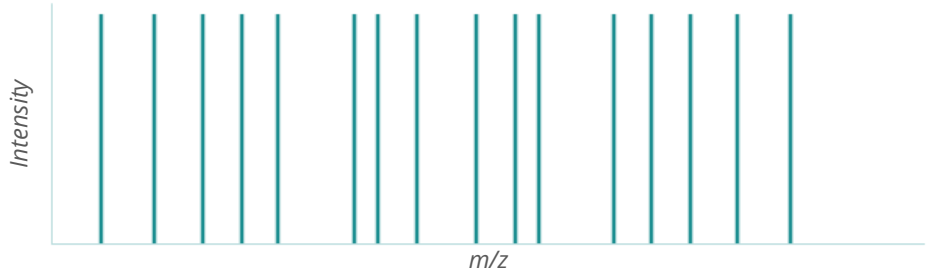
DDA proteomics data is traditionally identified using only the m/z dimension

Empirical spectrum



?

Theoretical spectrum



H-PEPTIDEK-OH

While DIA removes the stochasticity of DDA, the resulting spectra are much more complex to identify

+ Easy to identify
- Not all precursors fragmented

- Hard to identify
+ All precursors fragmented

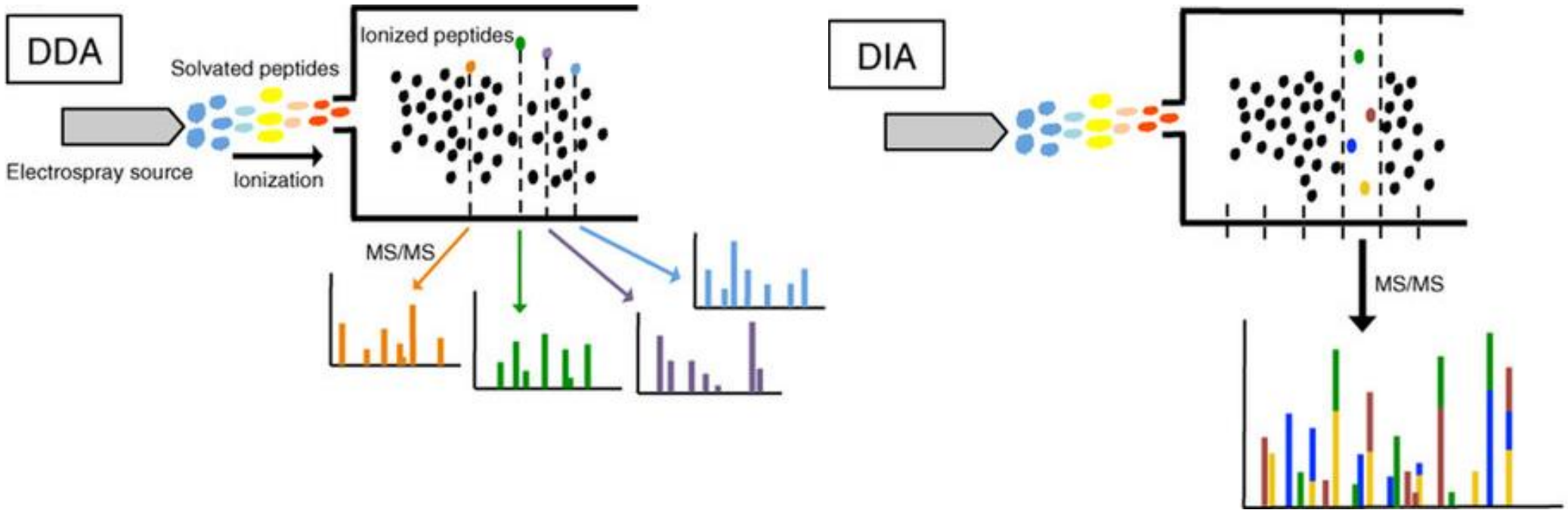
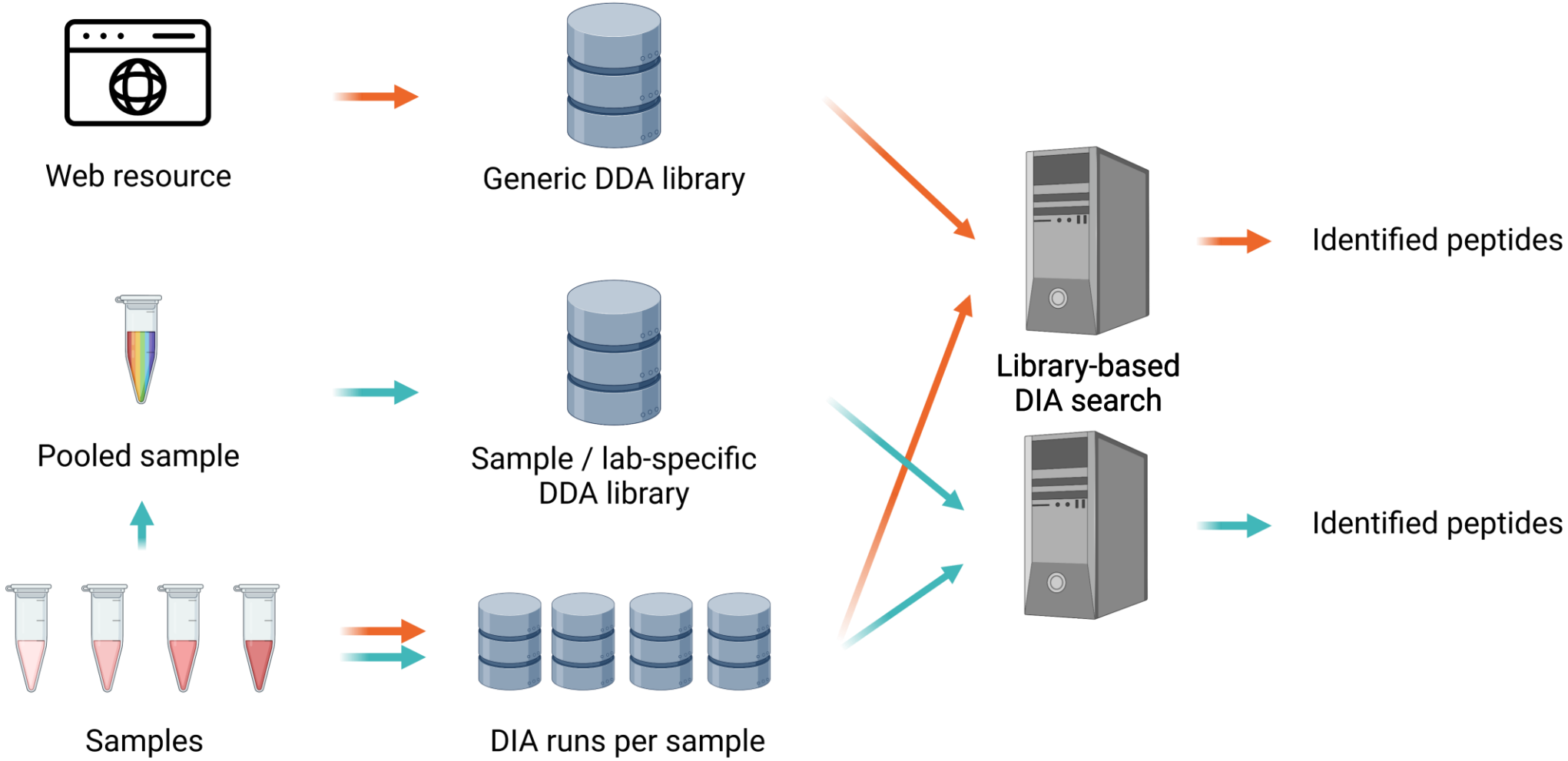


Figure adapted from R Peckner et al., Nat. Methods 2018, doi:10.1038/nmeth.4643

To cope with the increase in identification ambiguity, the additional RT and intensity dimensions are usually sourced from spectral libraries



Unfortunately, library-based DIA carries over DDA's drawback of stochasticity: Only the most intense precursors are acquired

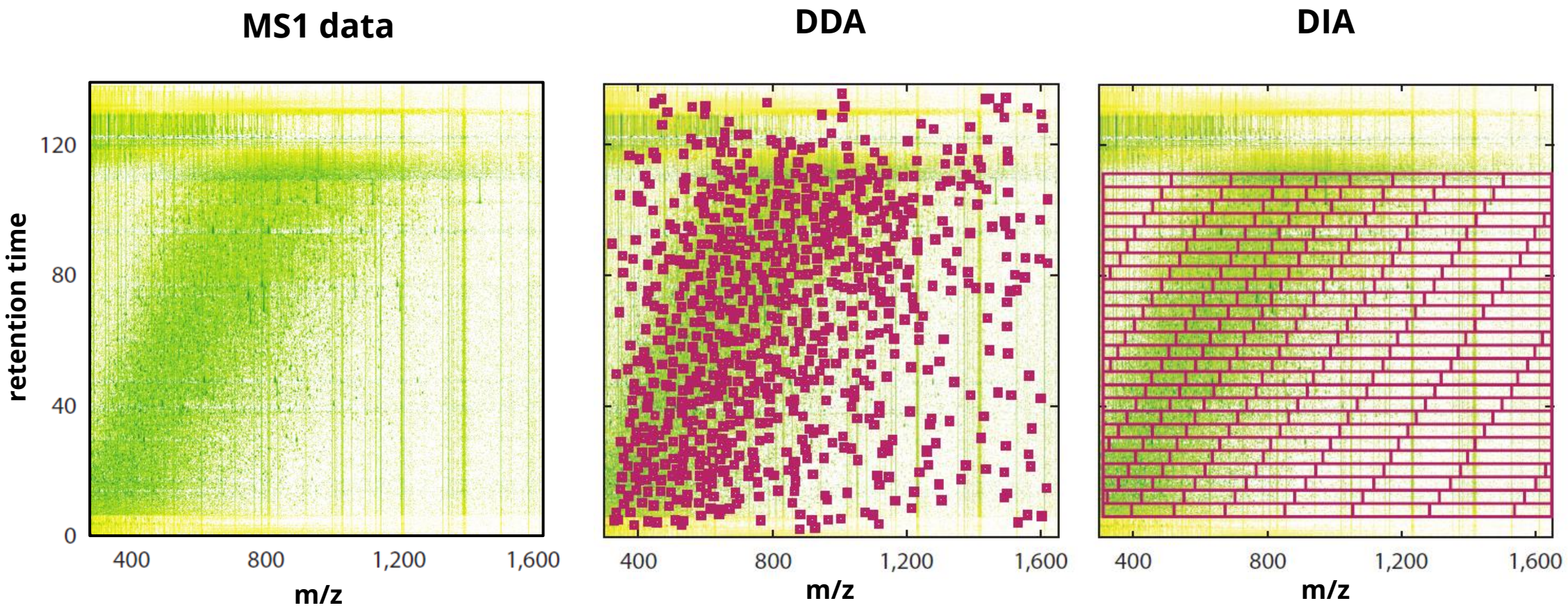
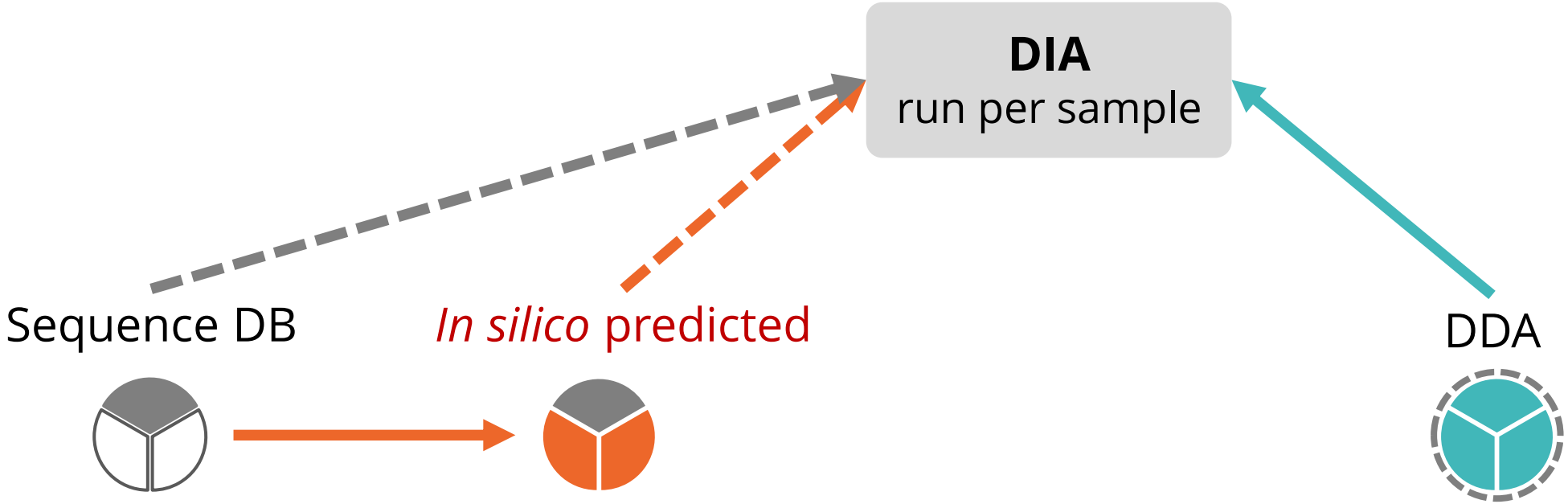
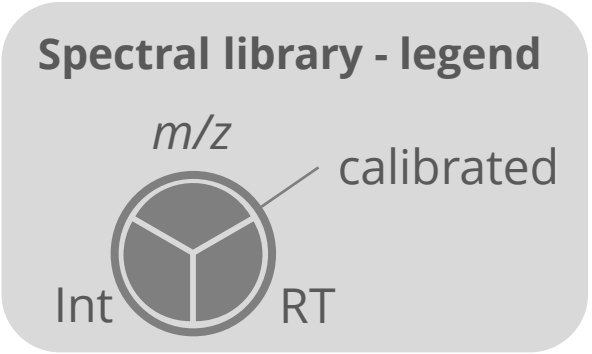


Figure adapted from Pavel Sinitcyn et al. Annu. Rev. Biomed. Data Sci. (2018)

In silico predicted libraries can combine the best of both worlds

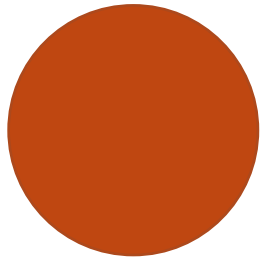




The smaller search space is a major, but often overlooked advantage of spectral libraries

DDA Pan-Human library

211 000 precursors



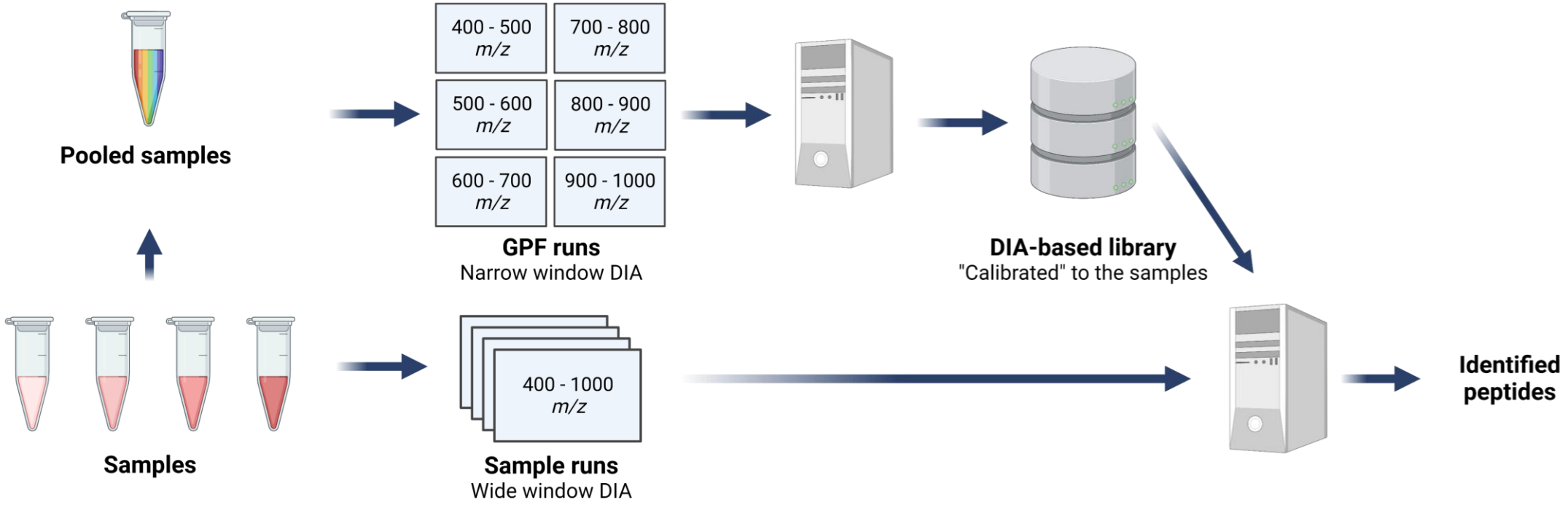
* to scale

Full proteome digest

3 000 000 precursors

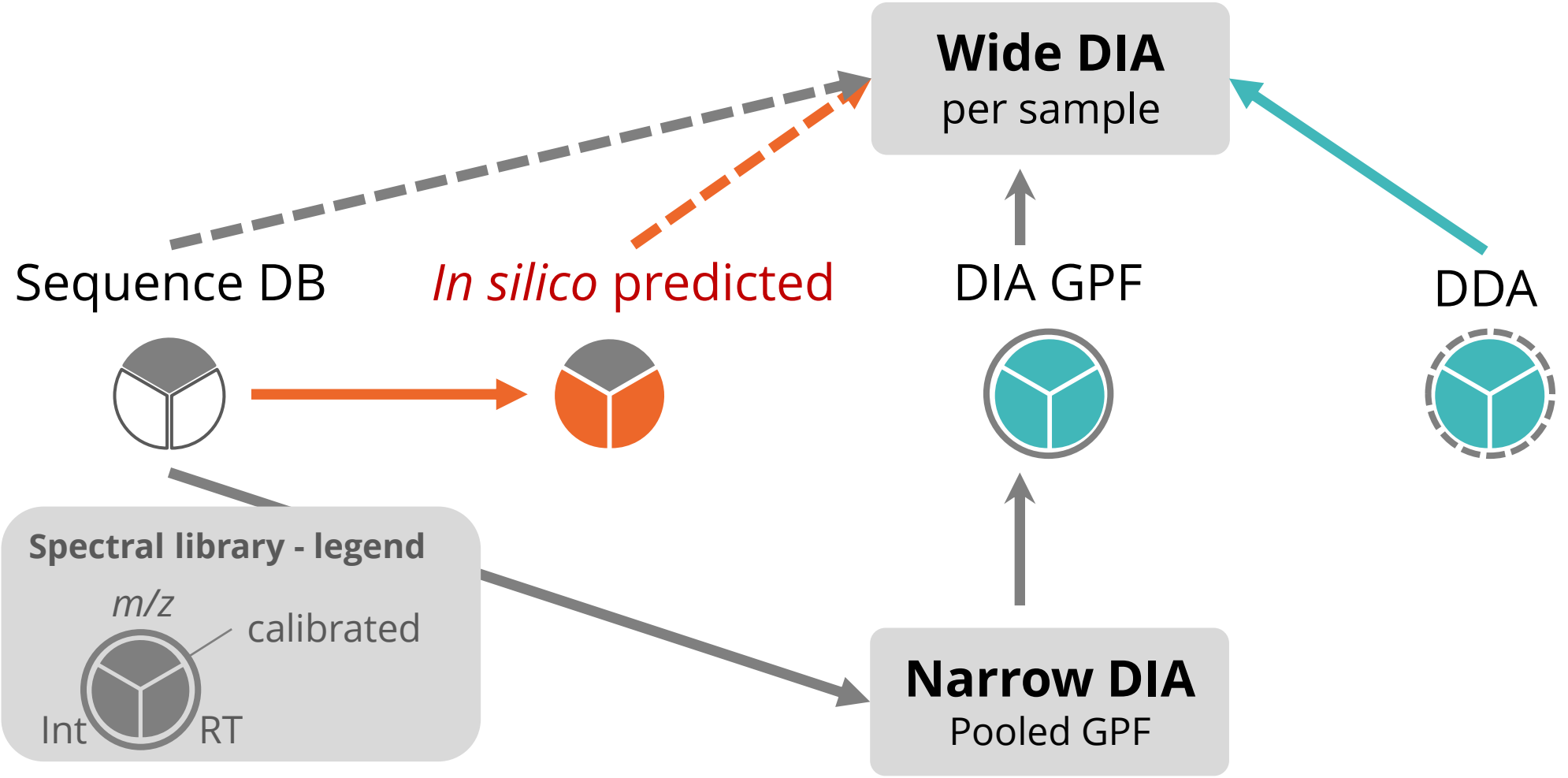
2+ / 3+
cmm@C, ox@M
2 missed cleavages
minimum 7 amino acids
maximum 5000 Da
m/z 400 – 1000 Da

An intermediate "chromatogram library" from a pooled, gas phase-fractionated DIA run helps to overcome the search space problem



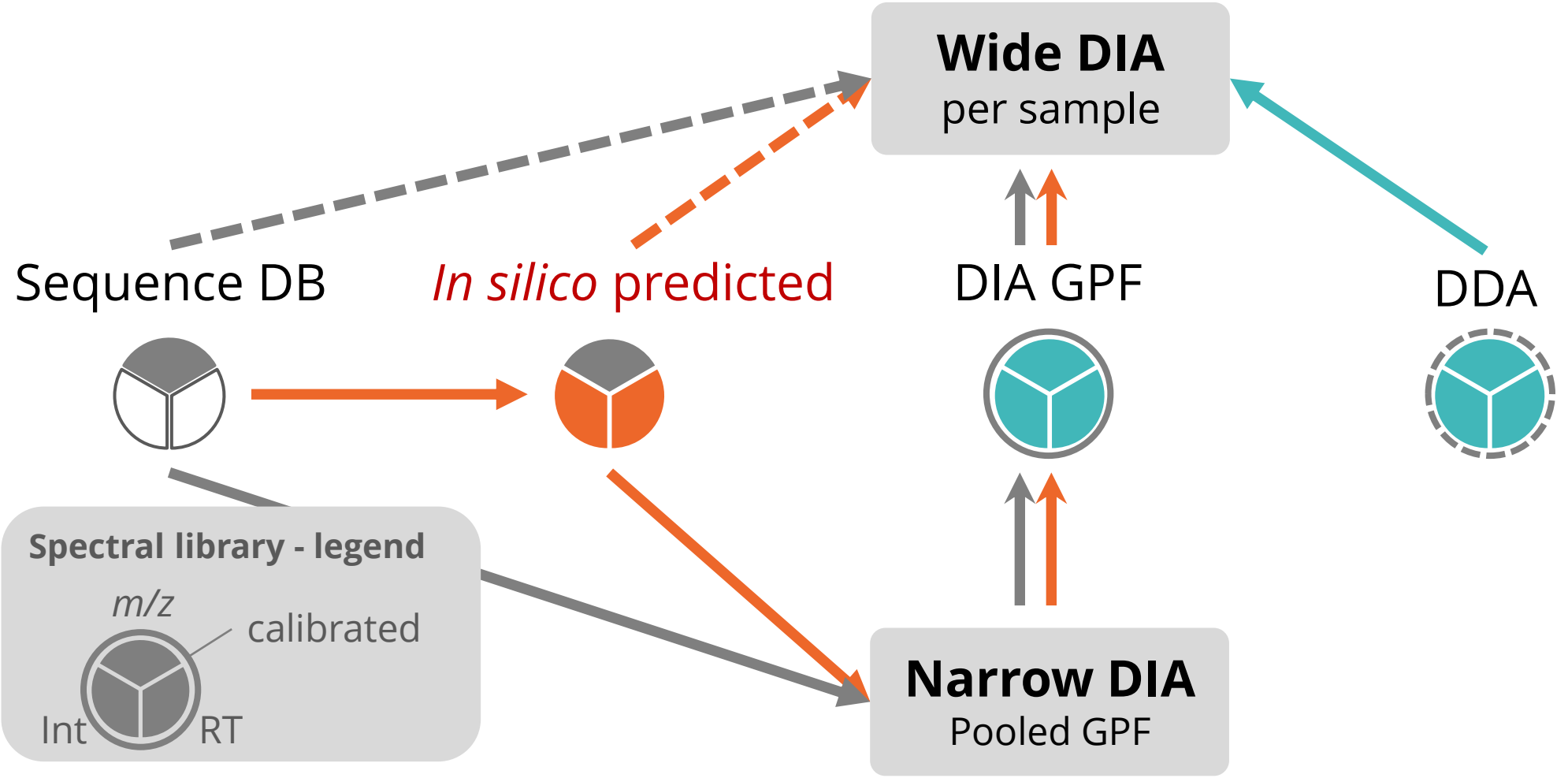
Brian C Searle et al. Nat. Commun. (2018). doi:10.1038/s41467-018-07454-w

An intermediate "chromatogram library" from a pooled, gas phase-fractionated DIA run helps to overcome the search space problem

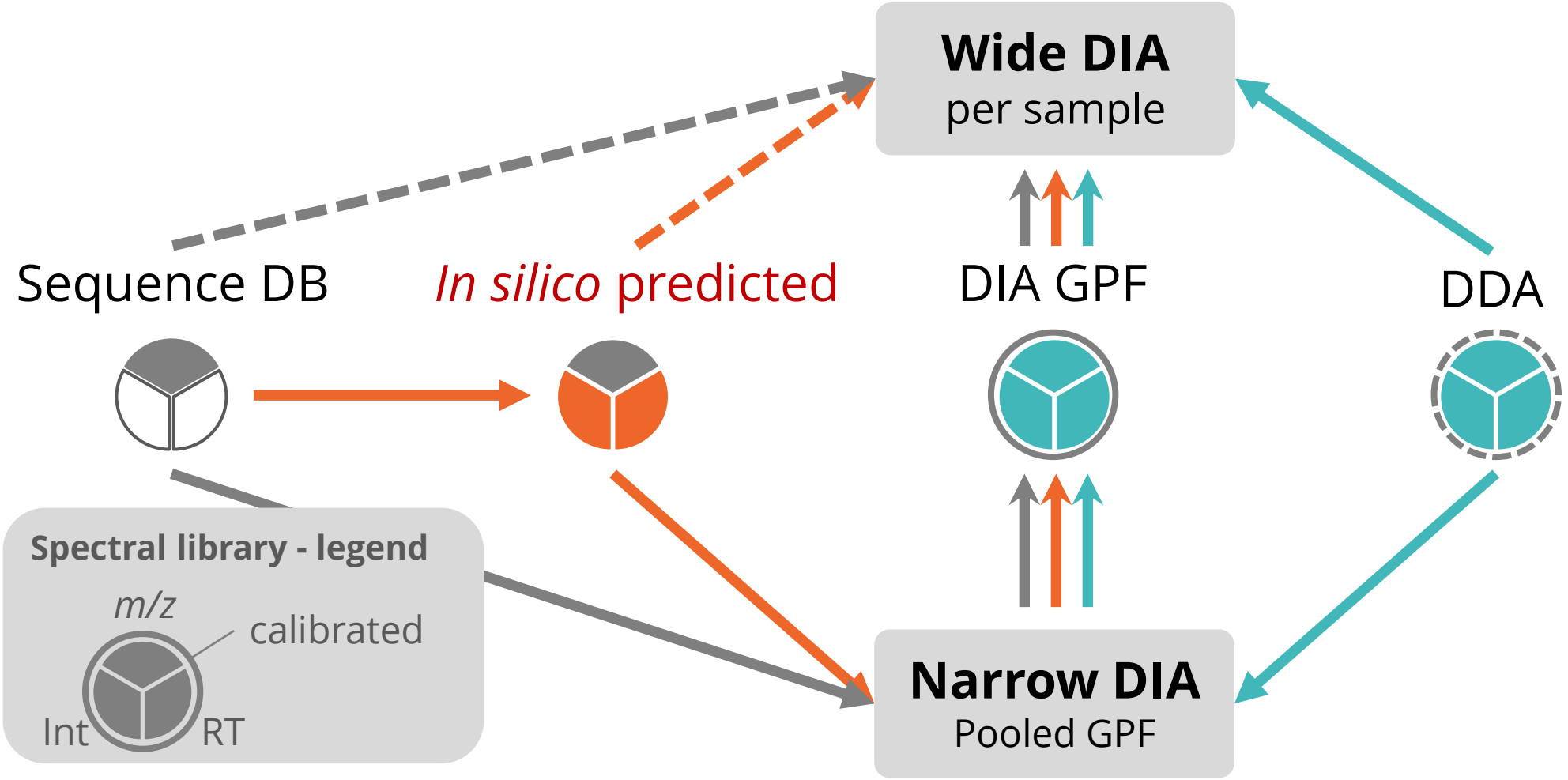


Bart Van Puyvelde*, Sander Willems*, Ralf Gabriels* et al. Proteomics. (2020) DOI:10.1002/pmic.201900306

An intermediate "chromatogram library" from a pooled, gas phase-fractionated DIA run helps to overcome the search space problem



An intermediate "chromatogram library" from a pooled, gas phase-fractionated DIA run helps to overcome the search space problem



Bart Van Puyvelde*, Sander Willems*, Ralf Gabriels* et al. Proteomics. (2020) DOI:10.1002/pmic.201900306

Overview

Why use predicted libraries for DIA

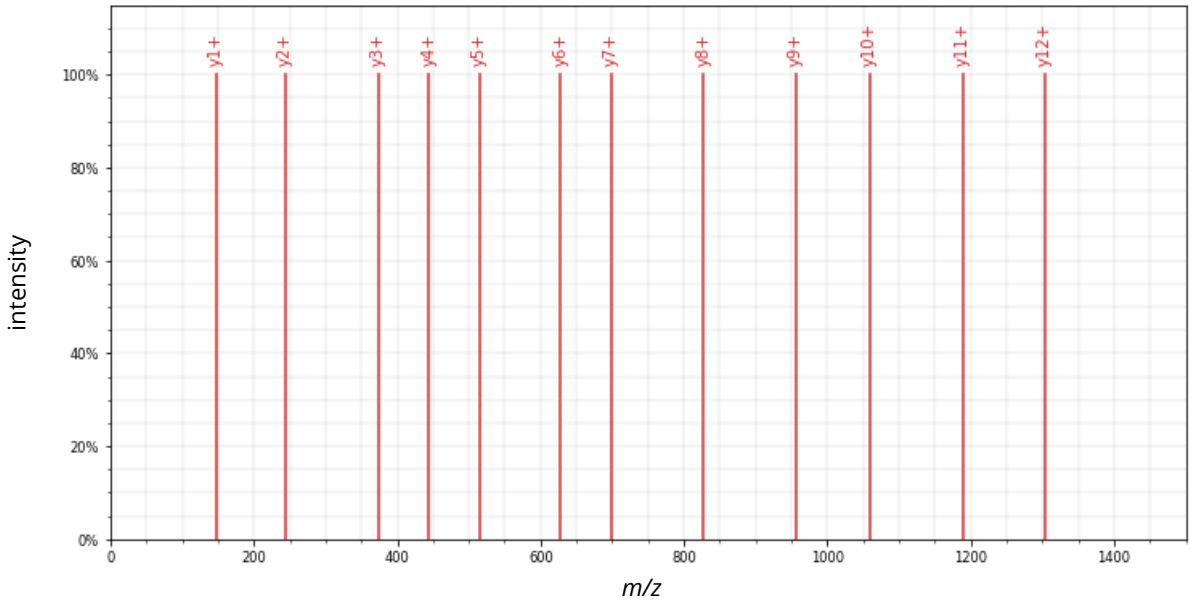
Practical considerations

Benchmarking results

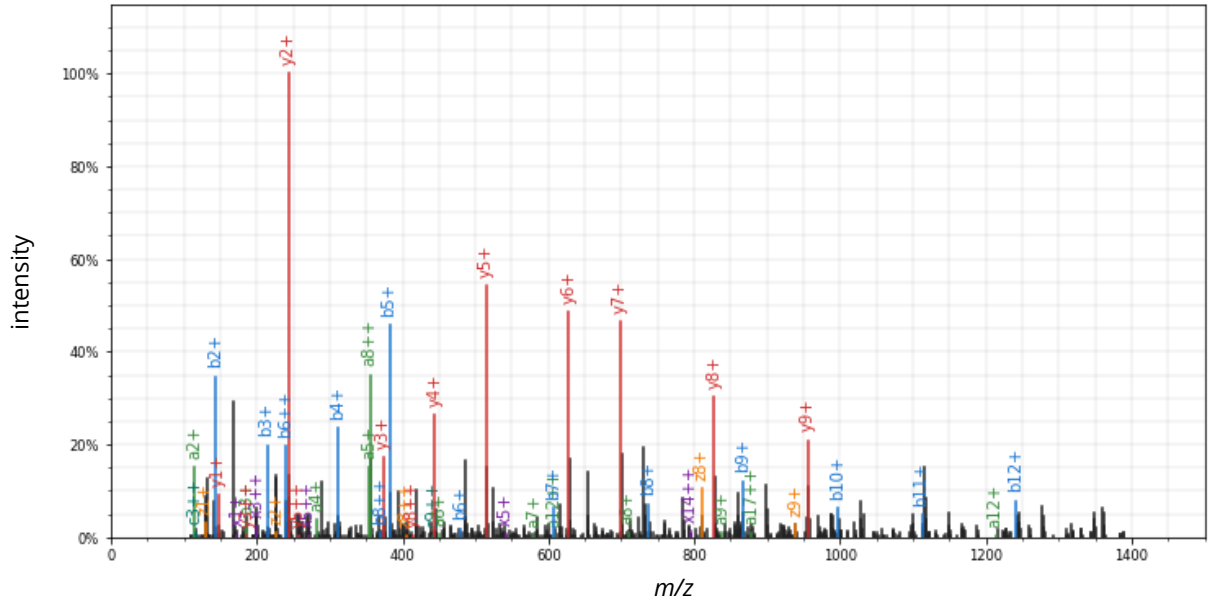
Future development

Take home messages

Fragment intensities are the most helpful, but arguably most complex, dimension to predict

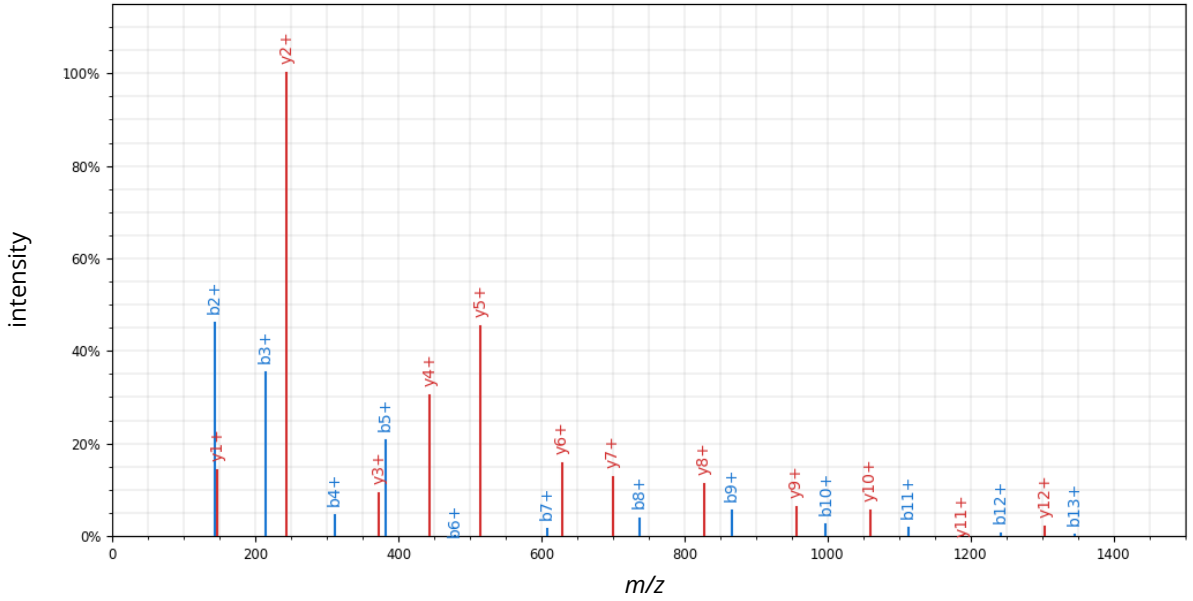


theoretical spectrum

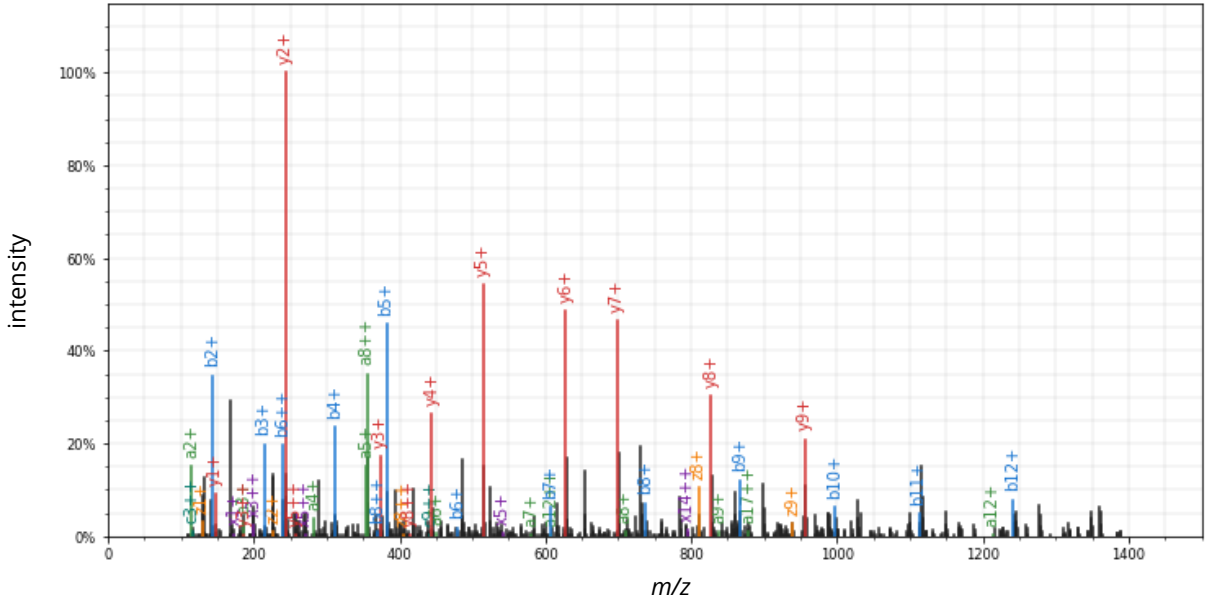


observed spectrum

Fragment intensities are the most helpful, but arguably most complex, dimension to predict

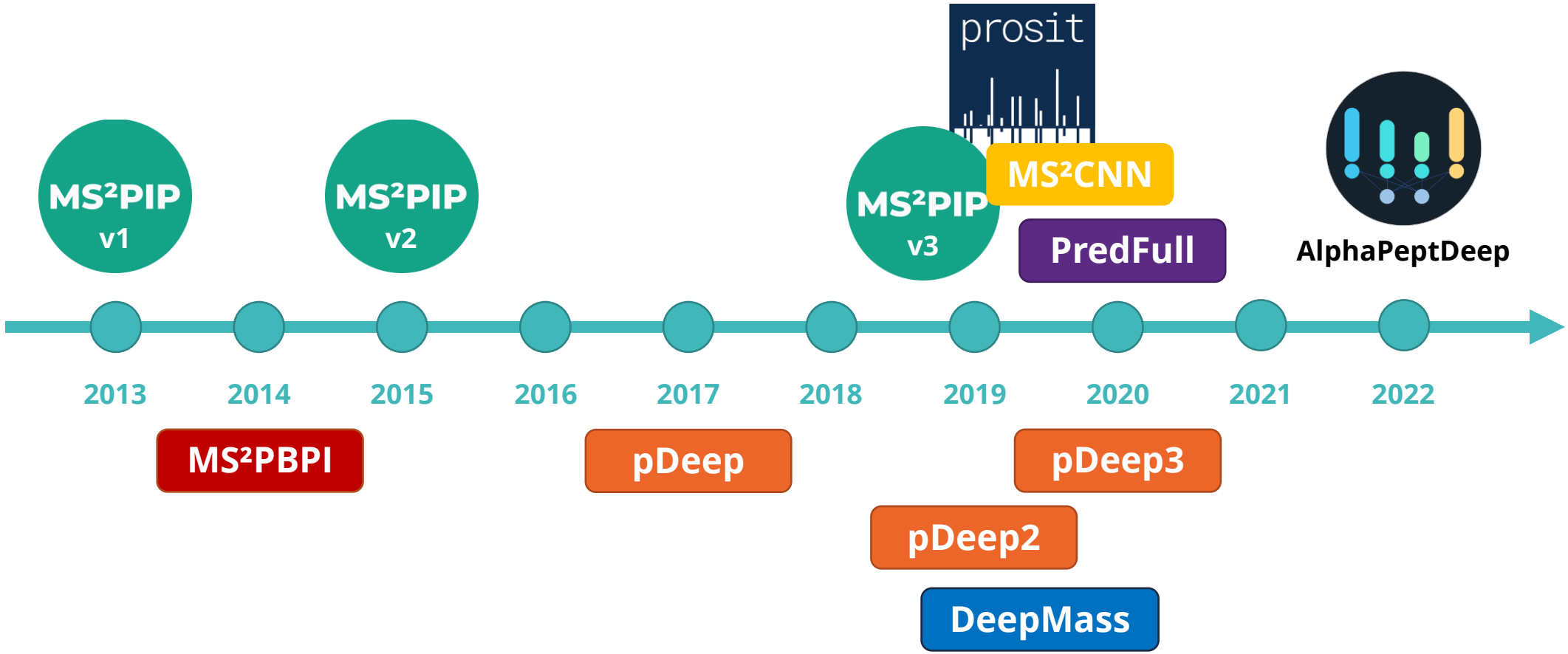


predicted spectrum

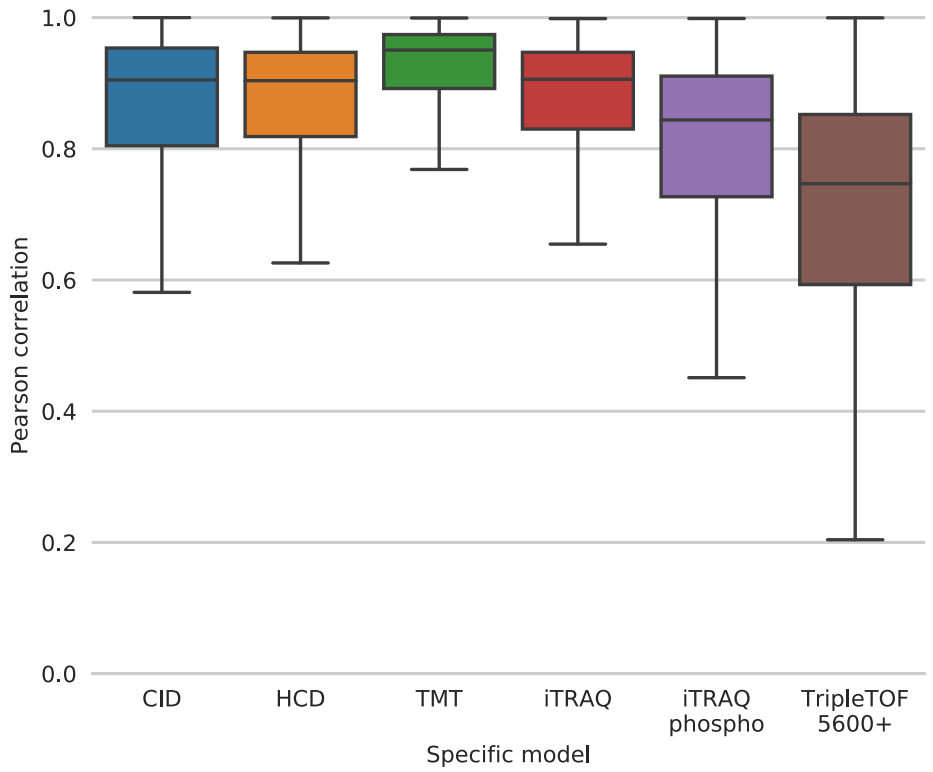


observed spectrum

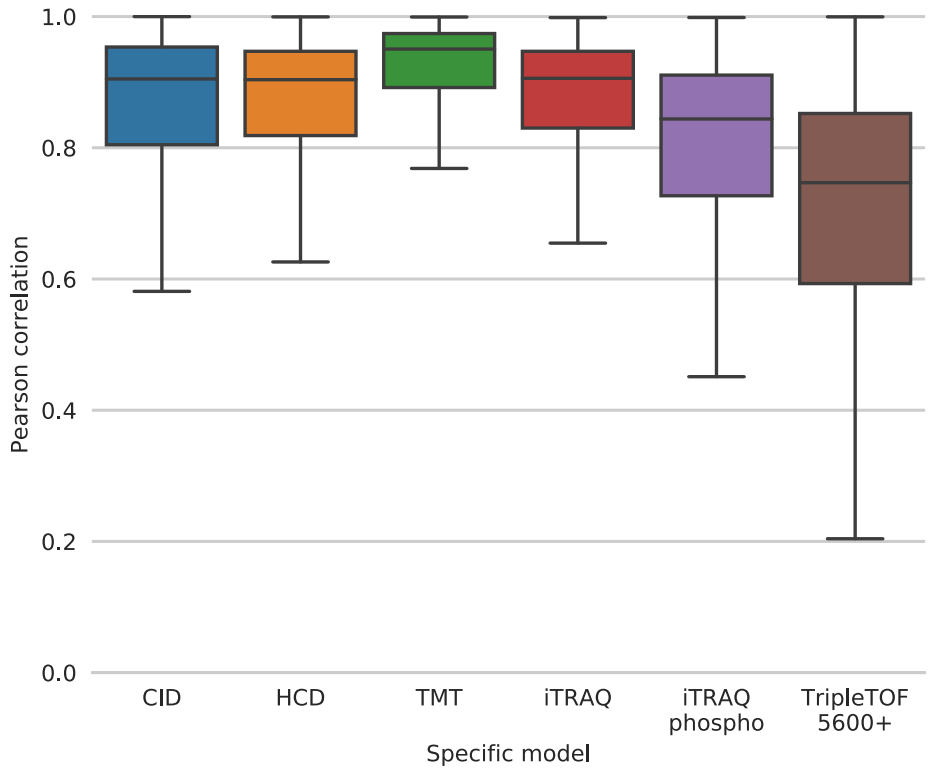
Many predictors have been published in the last decade, each with their own specific advantages and disadvantages

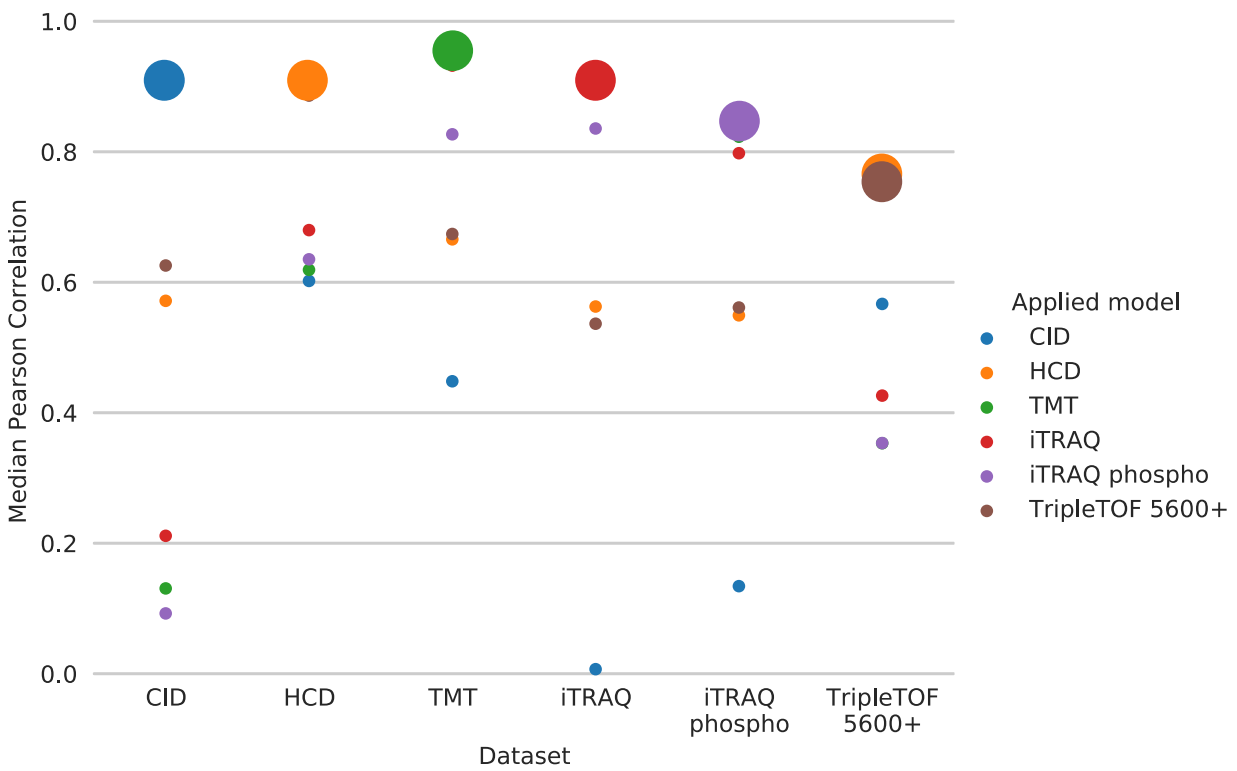


Intensities follow different patterns, depending on peptide type, instrument, instrument settings, and modifications

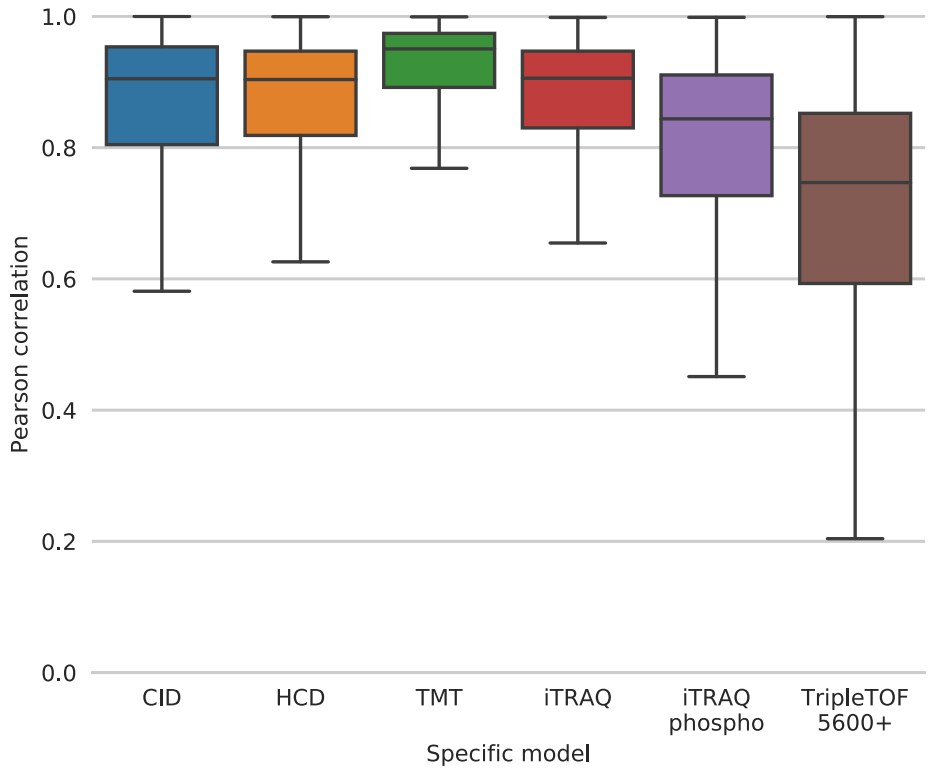


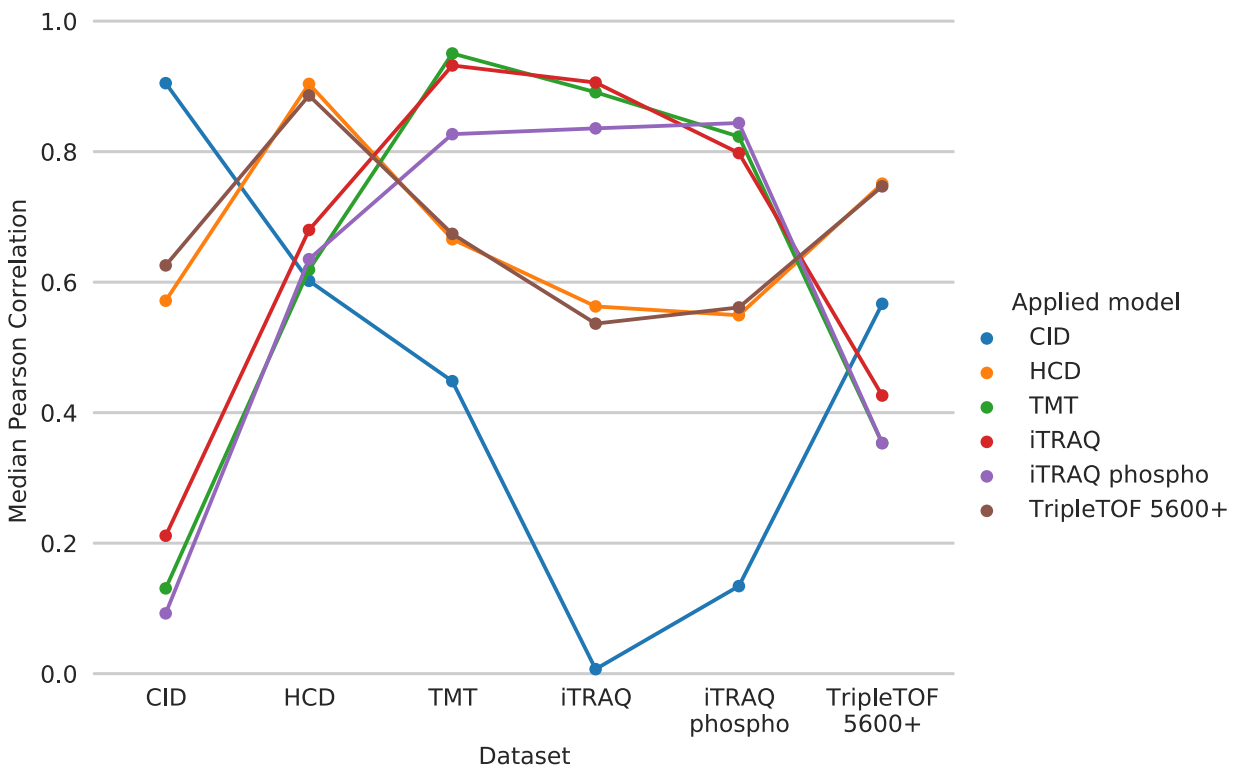
Intensities follow different patterns, depending on peptide type, instrument, instrument settings, and modifications





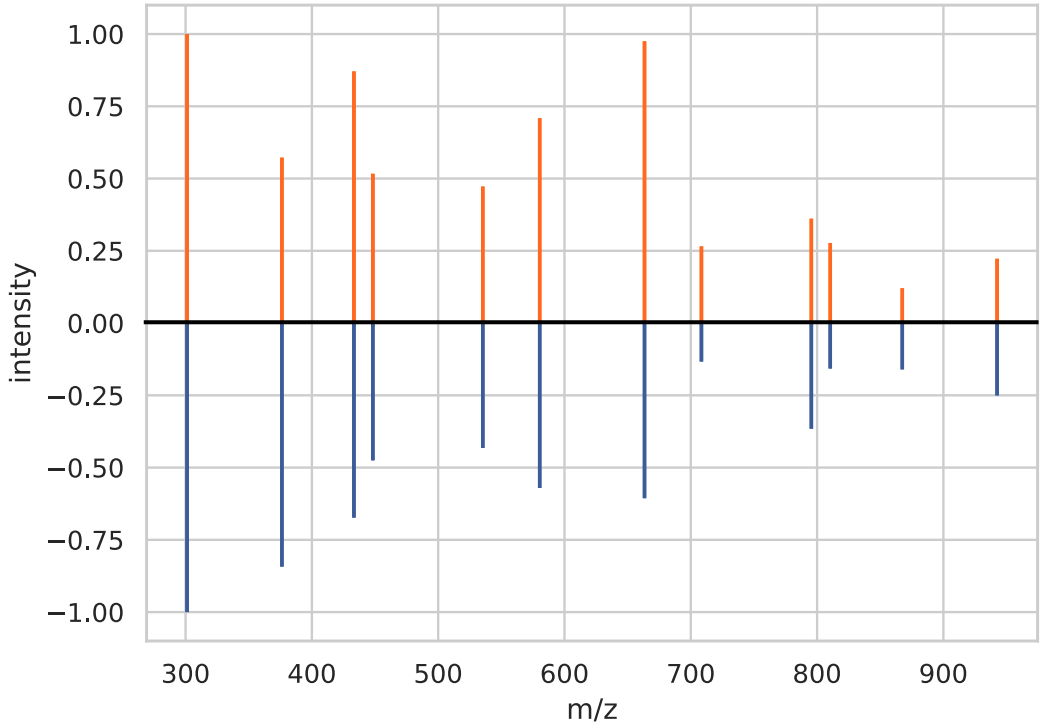
Intensities follow different patterns, depending on peptide type, instrument, instrument settings, and modifications





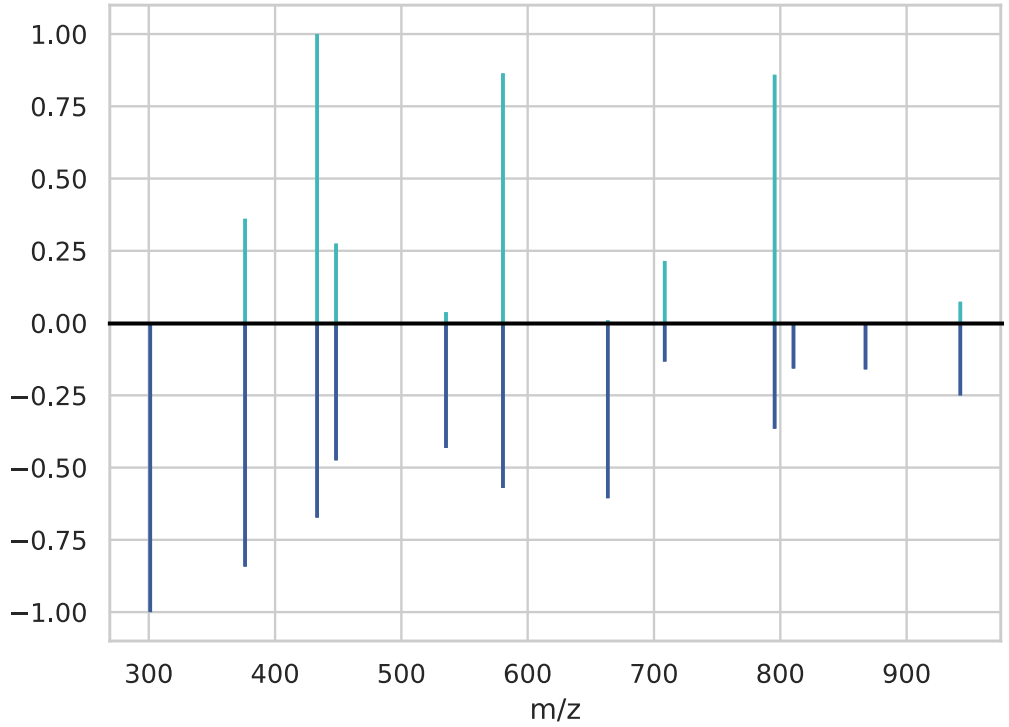
Intensities follow different patterns, depending on peptide type, instrument, instrument settings, and modifications

Prediction TMT model



Empirical TMT spectrum

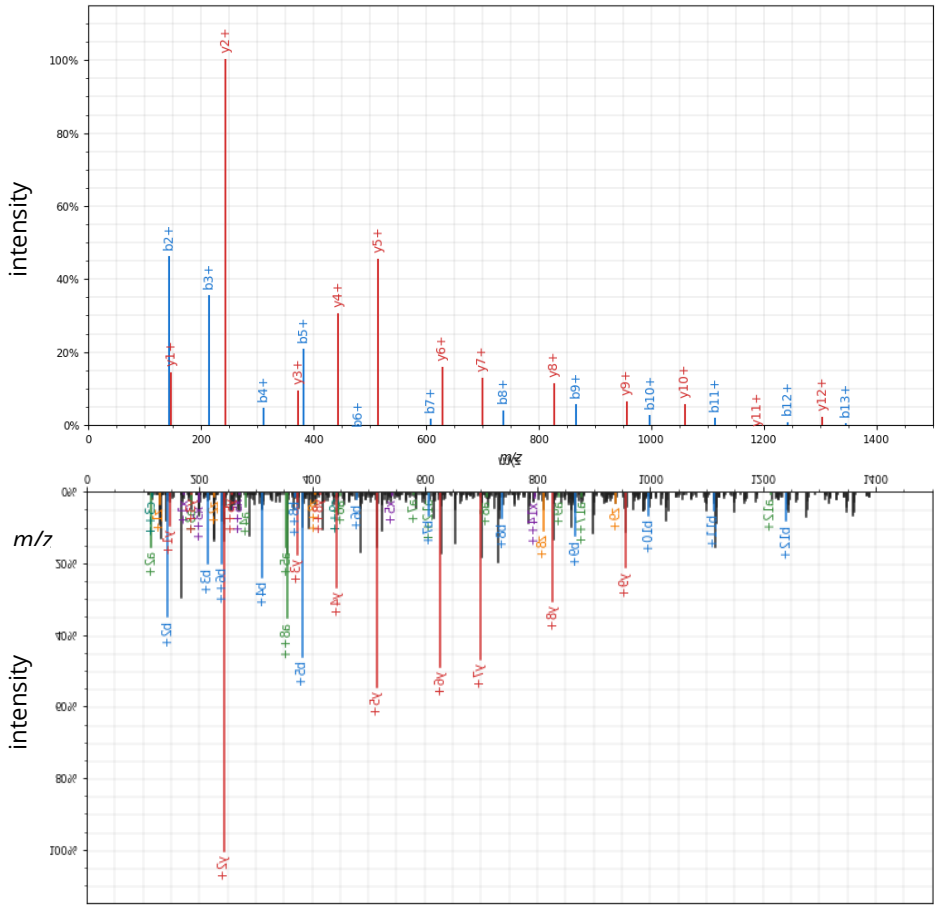
Prediction HCD model



Empirical TMT spectrum

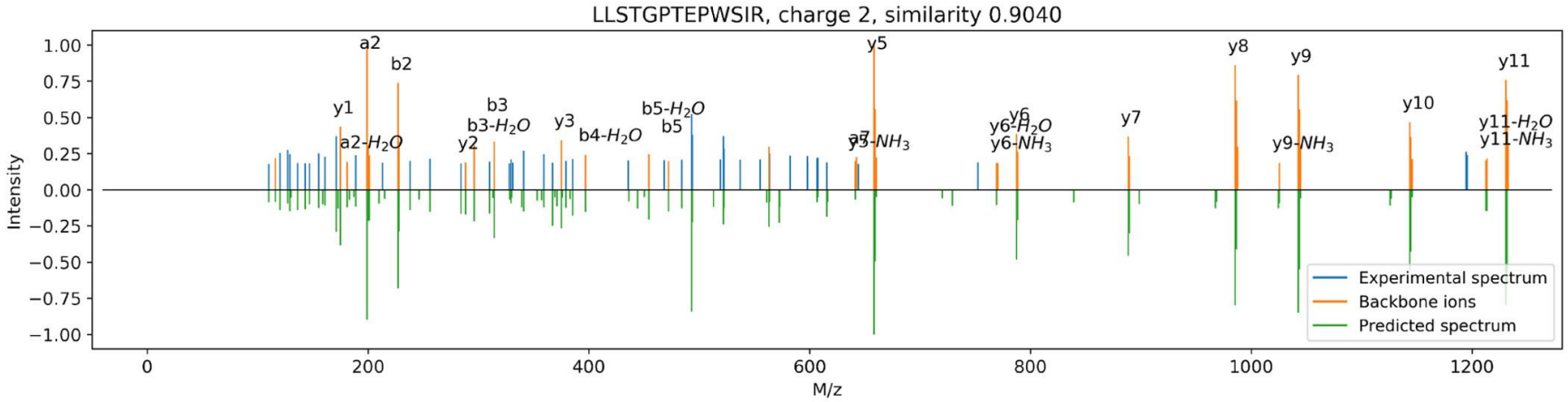
The scoring function should fit the intensities that are predicted

predicted

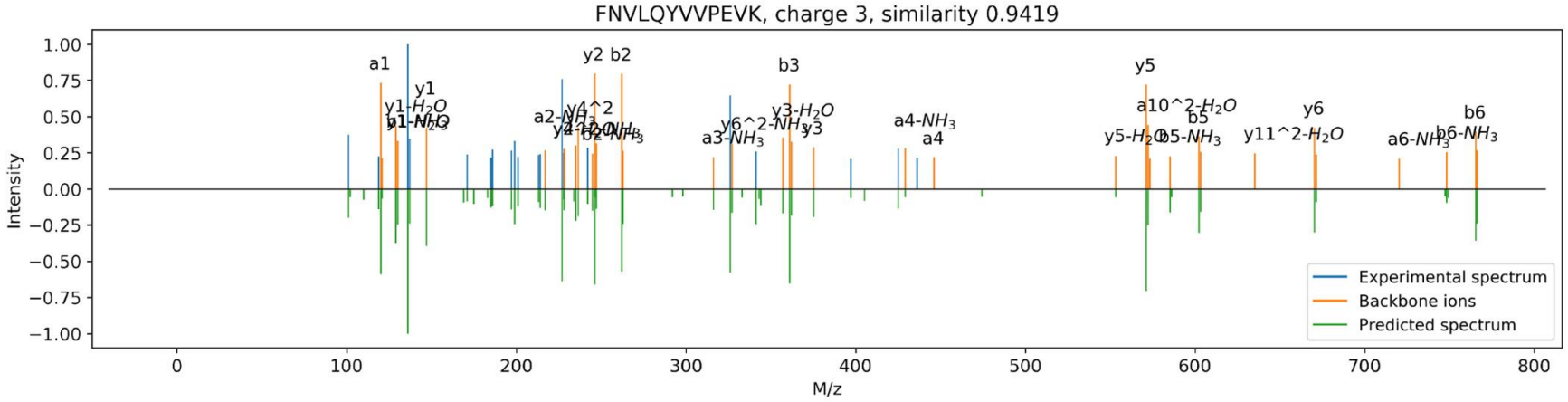


observed

The scoring function should fit the intensities that are predicted



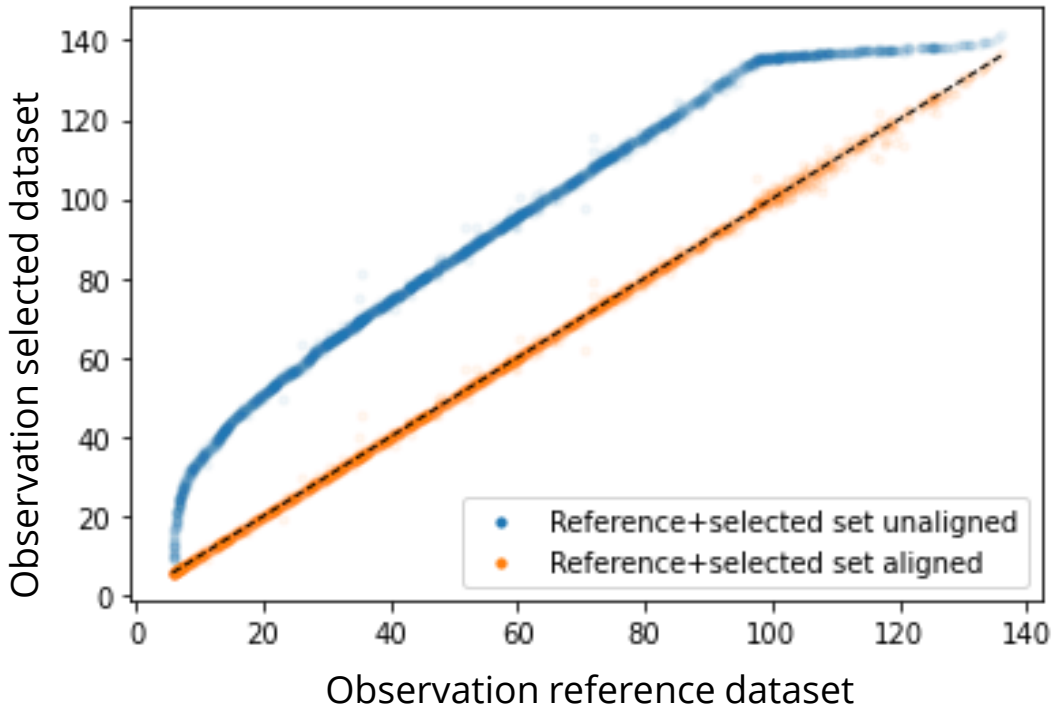
(a)



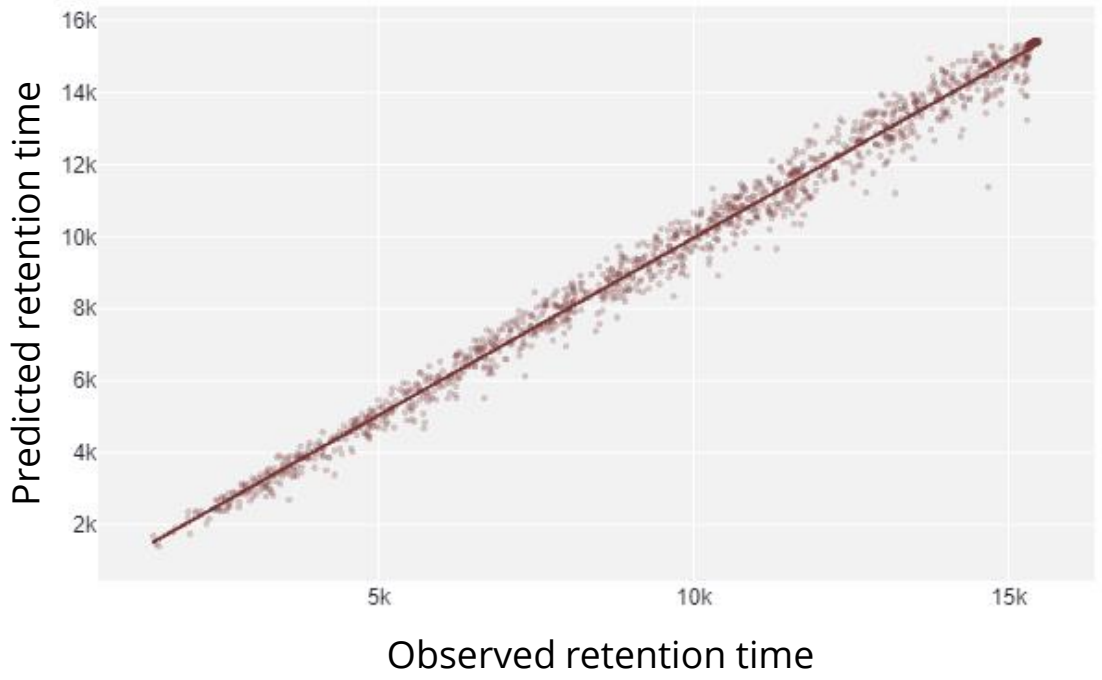
(b)

Figure adapted from Kaiyuan Liu et al. Analytical Chemistry. (2020) doi:10.1021/acs.analchem.9b04867

Retention time is straightforward to predict, but must be calibrated to each LC setup



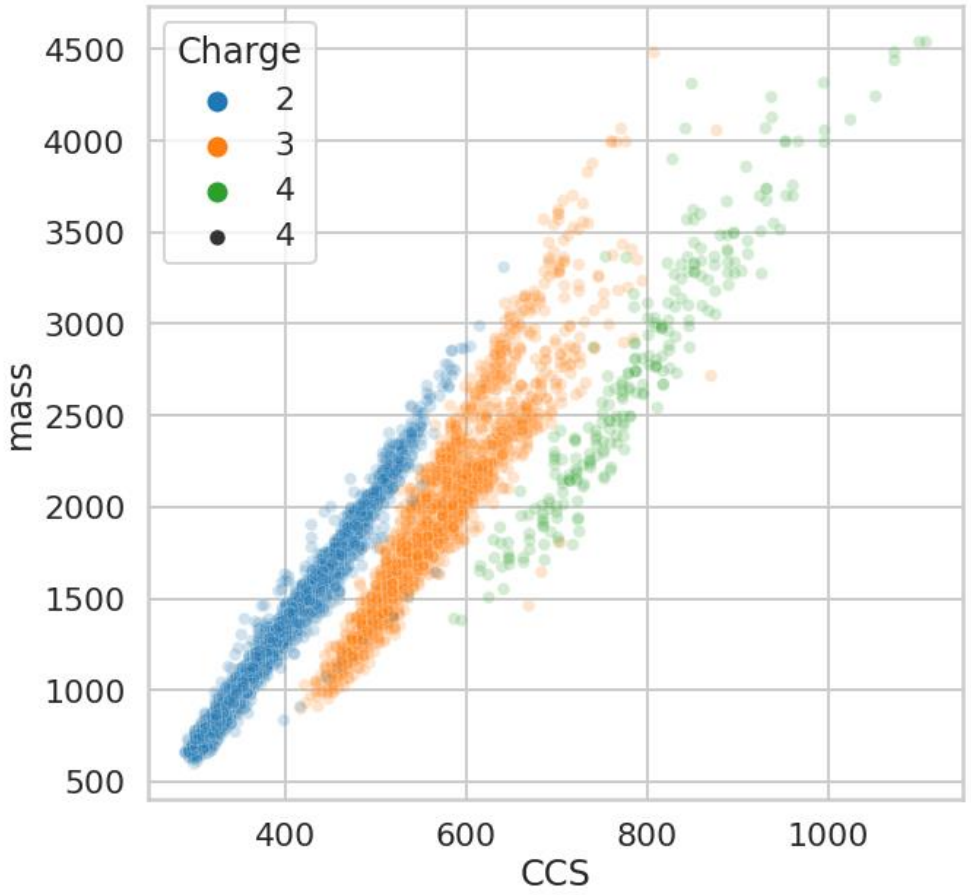
Calibration of different datasets

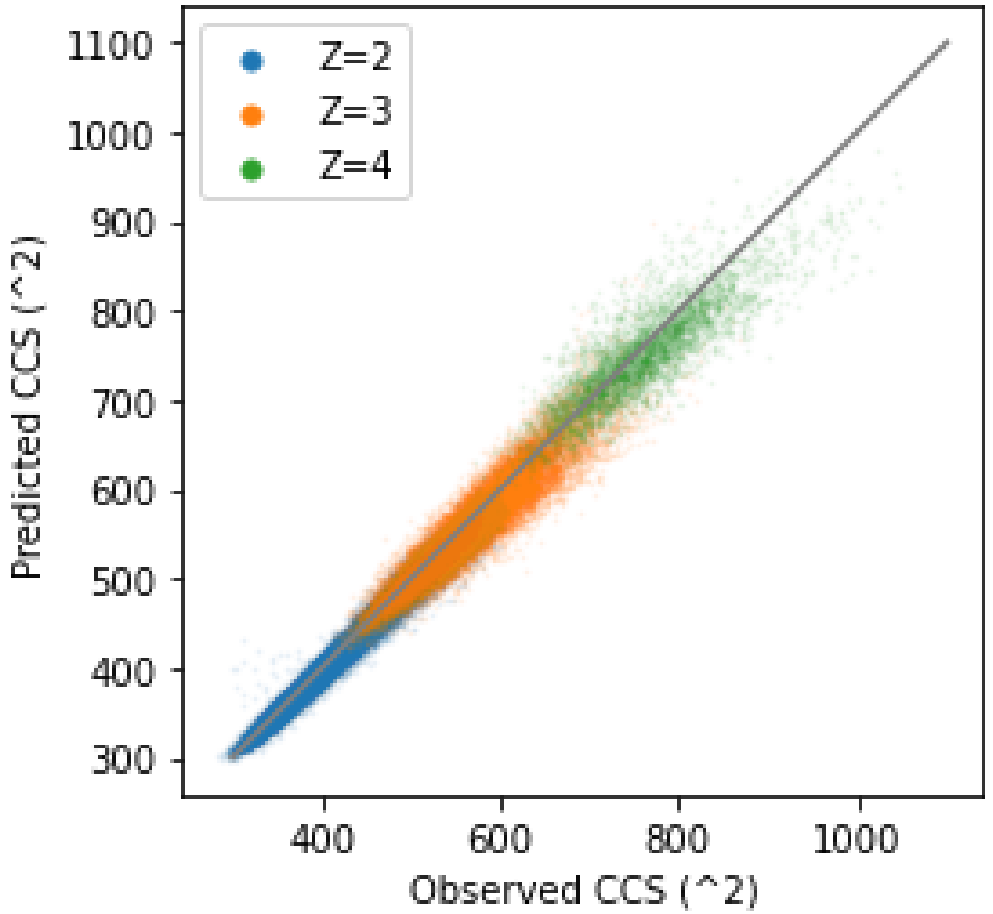


DeepLC prediction + calibration

Figures adapted from: Robbin Bouwmeester et al. Nature Methods (2021) doi:10.1038/s41592-021-01301-5; ProteomicsML.org

Ion mobility is the new kid on the block, but might not bring a lot of value to the identification step itself

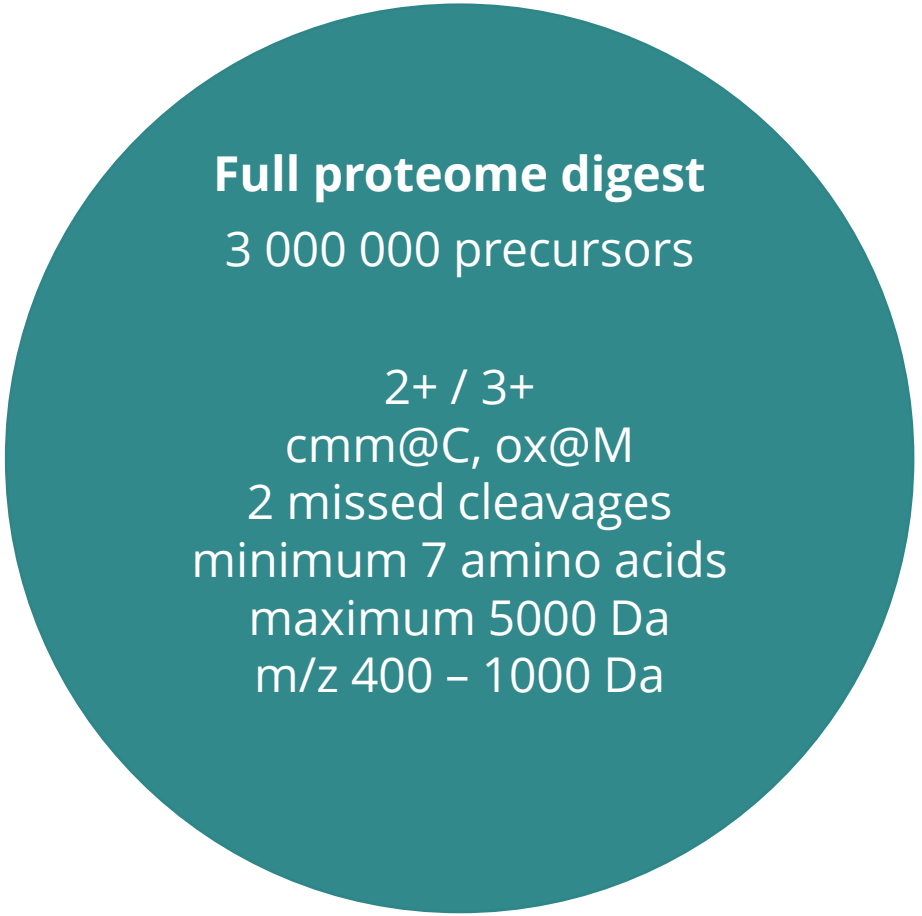




Figures adapted from ProteomicsML.org

Predicted libraries allow for customizability of the search space with plenty of parameters

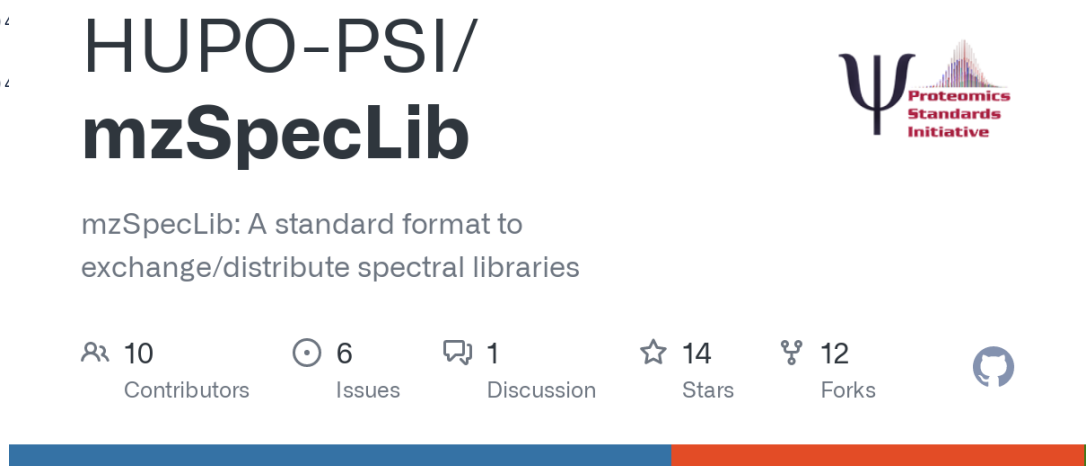
- Protein database
- Digestion rules
- Missed cleavages
- Charge states
- Modifications
- Length
- Peptide mass
- Precursor m/z
- …



Many spectral library formats exist, unfortunately without much consensus

```
Name: Vecuronium cation
Comment: NIST Mass Spectrometry Data Center
Num peaks: 22
81.0699 9.29 "C6H9=p-C28H48N2O4/0.2ppm;C12H18^2=p-C22H40N2O4^2/-3.1ppm 14/14"
86.0962 56.04 "C5H12N=p-C29H45NO4/-2.6ppm;C10H24N2^2=p-C24H34(
98.0964 3.20 "C6H12N=p-C28H45NO4/-0.2ppm;C12H24N2^2=p-C22H34O4
99.1042 2.10 "C6H13N=p-C28H44NO4/-0.5ppm;C12H26N2^2=p-C22H32O4
```

NIST MSP
SpectraST sptxt
BiblioSpec MS2/SSL
EncyclopeDIA DLib
Spectronaut CSV
…



Overview

Why use predicted libraries for DIA

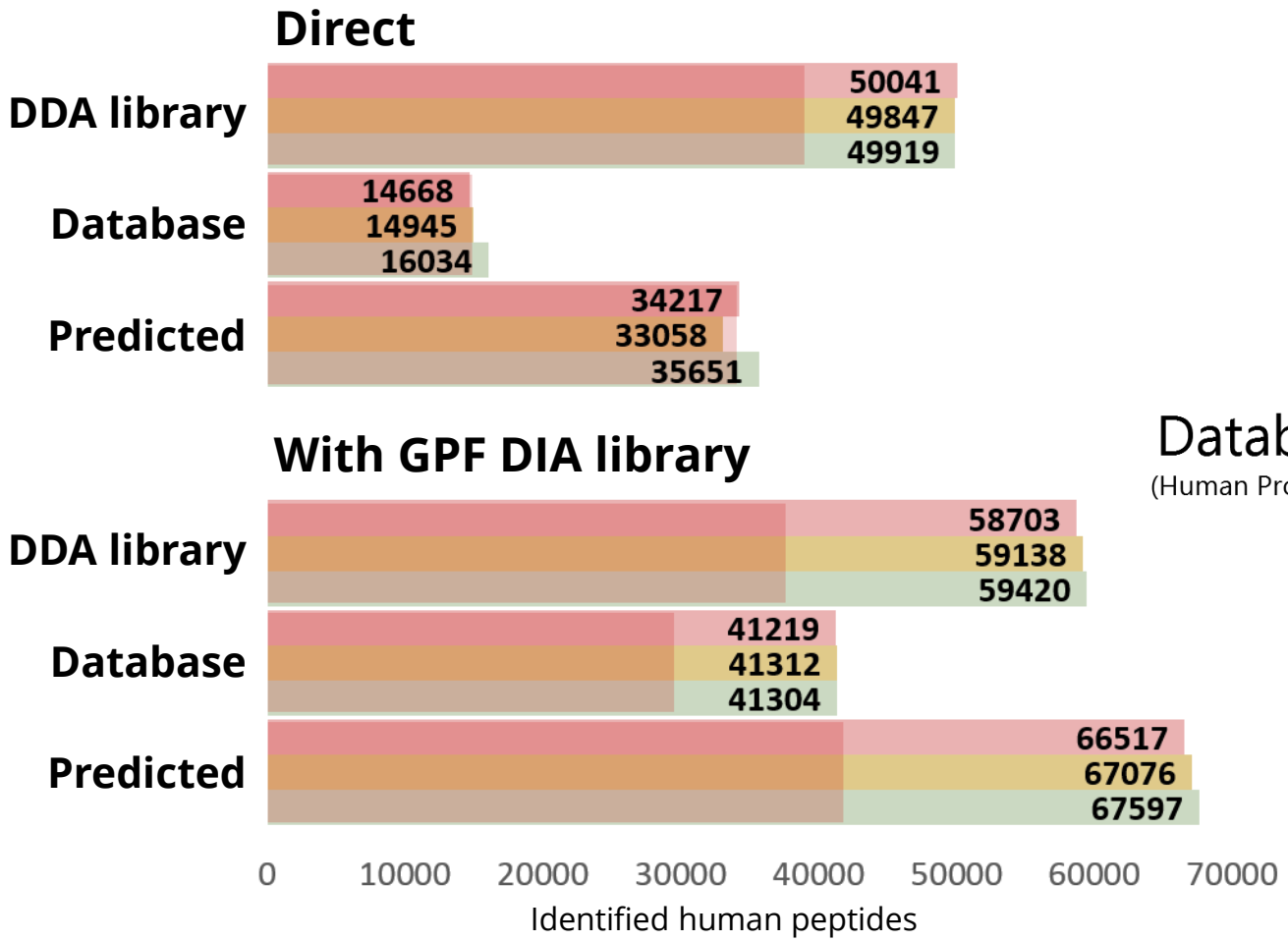
Practical considerations

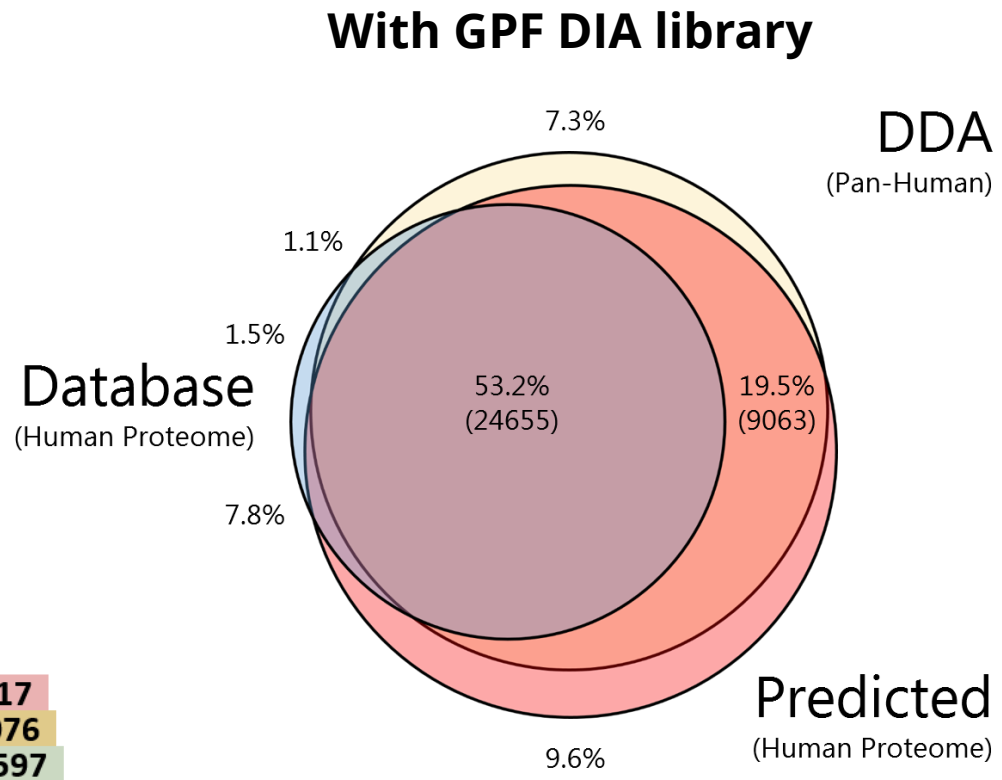
Benchmarking results

Future development

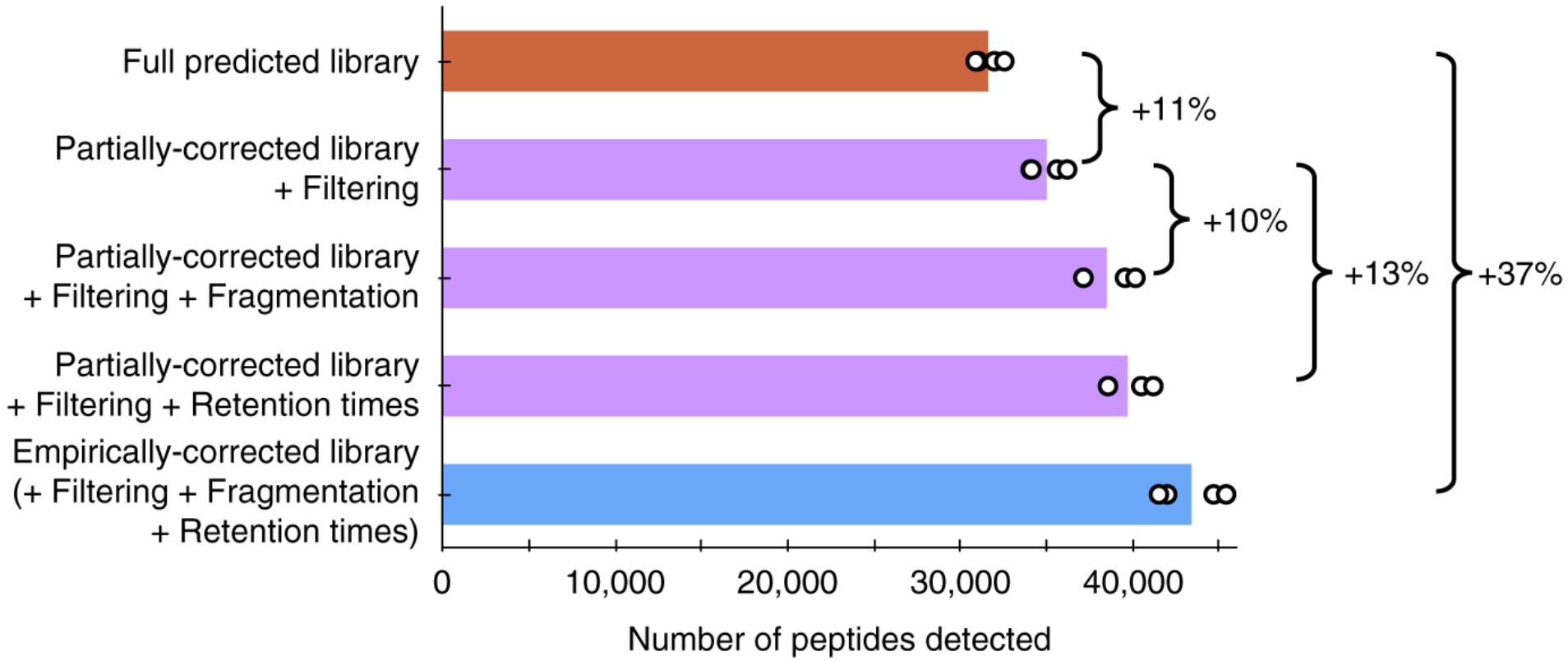
Take home messages

Predicted libraries in combination with an intermediate pooled GPF DIA library outperforms other methods



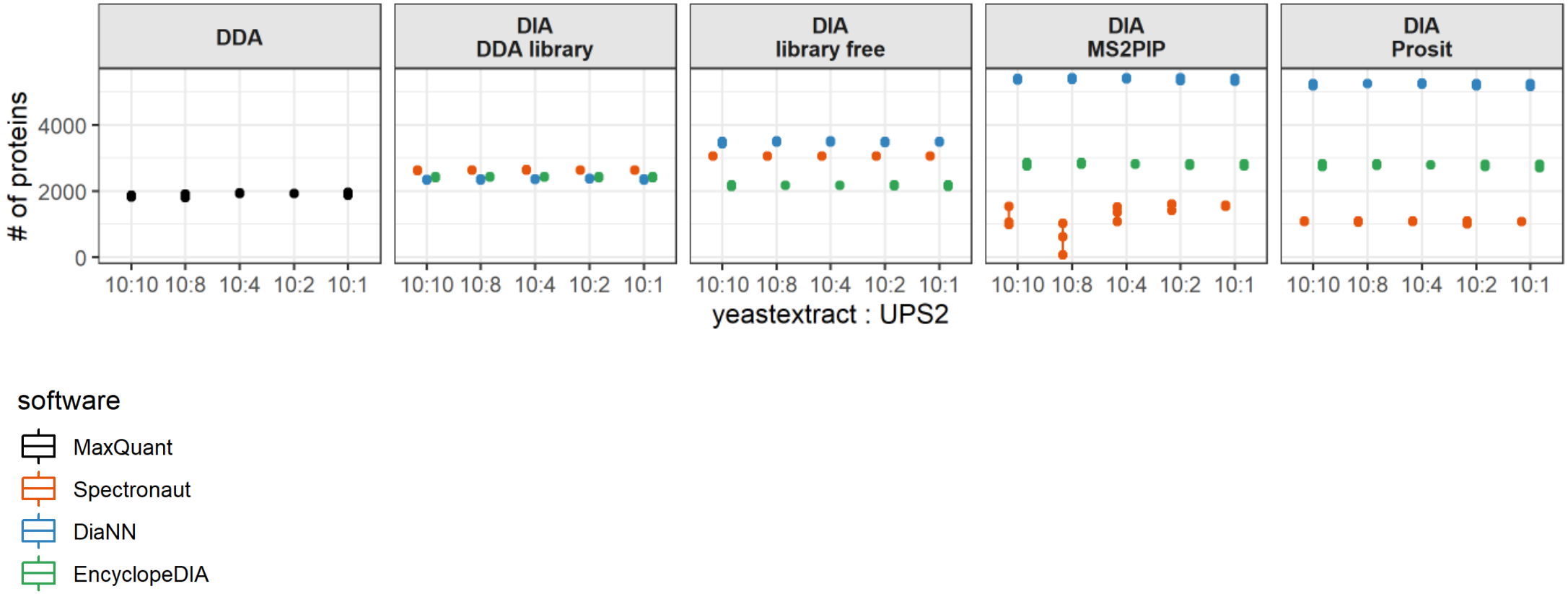


Similar results were obtained simultaneously by another group



Brian C. Searle et al. Nat. Communications (2020) doi:10.1038/s41467-020-15346-1

A recent large-scale methodological comparison shows that the combination of library type and software greatly influences the results



An Staes et al. (manuscript in preparation)

Overview

Why use predicted libraries for DIA

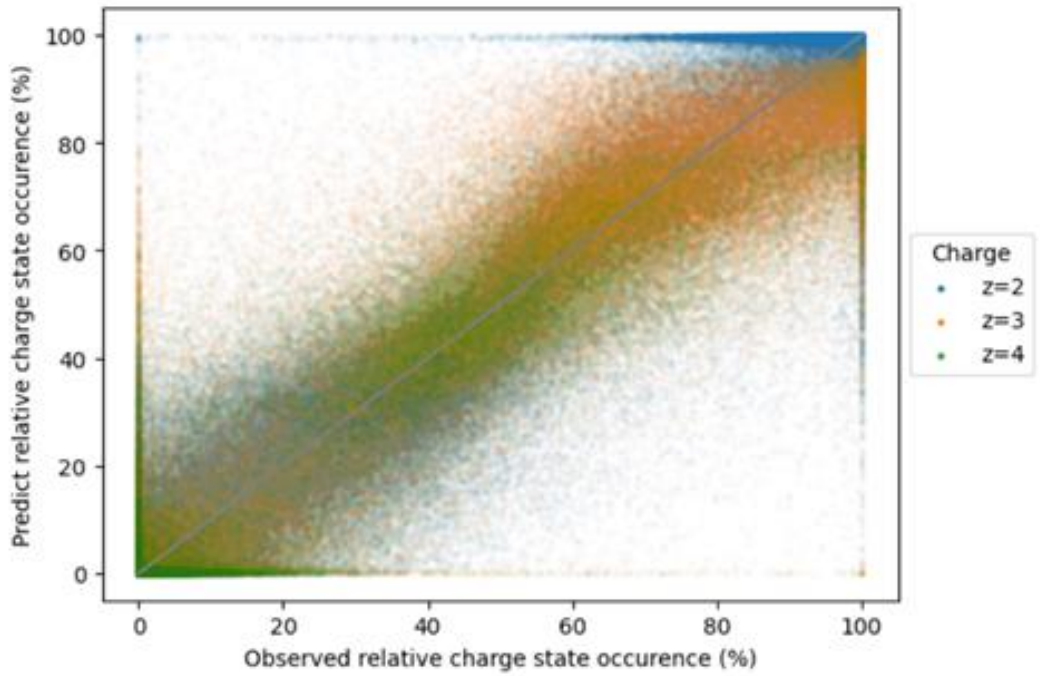
Practical considerations

Benchmarking results

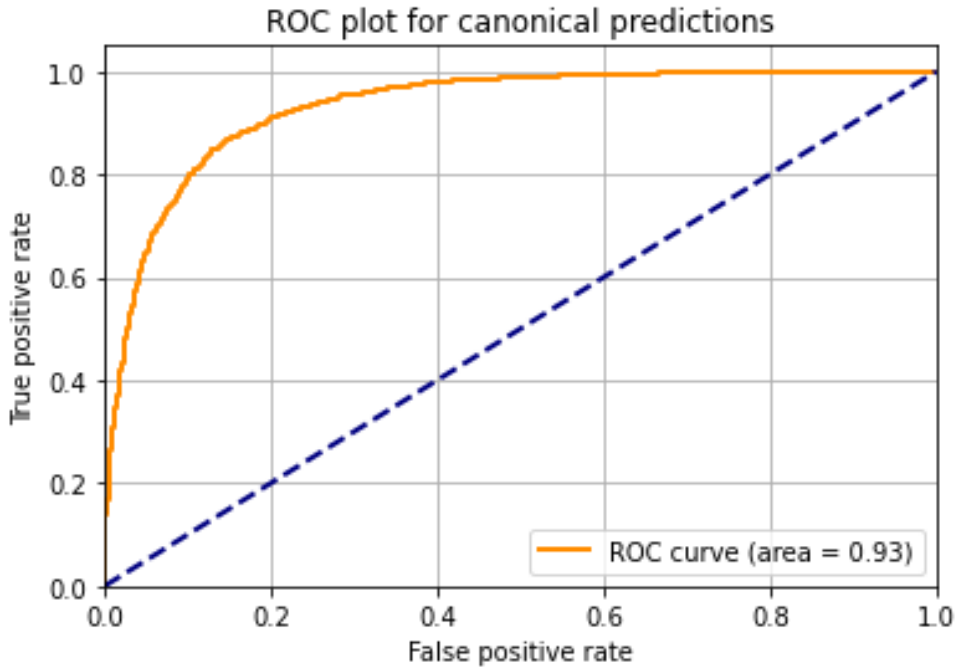
Future development

Take home messages

More aspects of LC-IM-MS/MS can be (and will be) predicted



Charge state occurrence



Peptide detectability

Left: Unpublished work; Right: ProteomicsML.org

Using ML predictions on-the-fly for both DIA and DDA will become standard practice

Already happening for DDA

ionbot

DIA and DDA seem to come closer

Wide window DDA

Narrow window DIA

Overview

Why use predicted libraries for DIA

Practical considerations

Benchmarking results

Future development

Take home messages

Take home messages

Use predicted libraries whenever possible

Use the correct models for your setup

No *one-fits-all*: Evaluate multiple strategies on your use-case

ProteomicsML

datasets, tutorials, and open science

www.proteomicsml.org

Lennart Martens

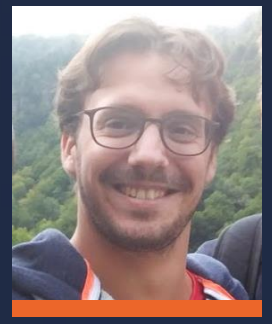
Sven Degroeve

Bart Van Puyvelde

Maarten Dhaenens

Robbin Bouwmeester

Arthur Declercq



Sander Willems

Comp omics

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

UNIVERSITEIT GENT

Ralf Gabriels

@RalfGabriels

https://ralf.gabriels.dev

https://compomics.com