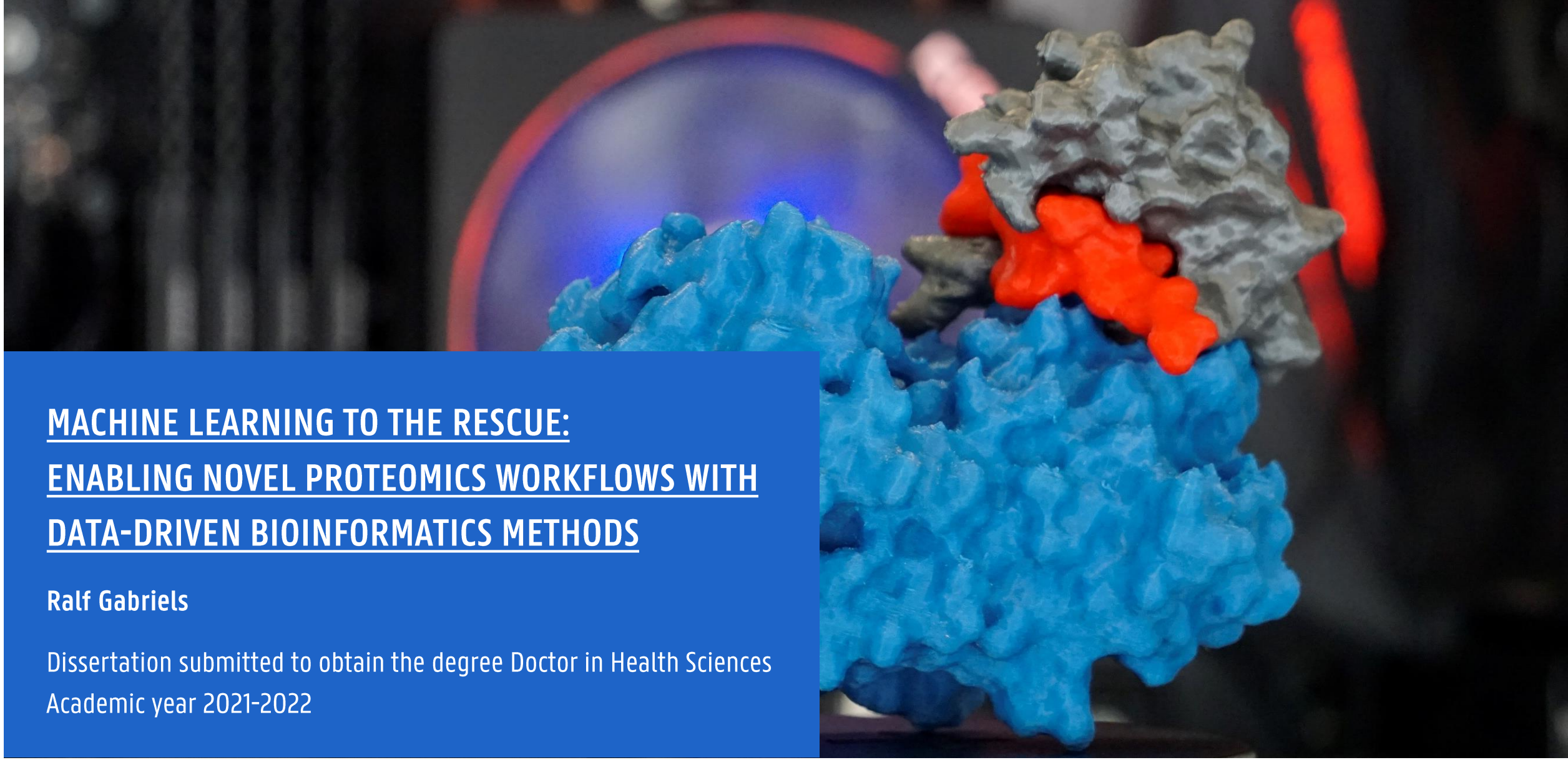


MACHINE LEARNING TO THE RESCUE: ENABLING NOVEL PROTEOMICS WORKFLOWS WITH DATA-DRIVEN BIOINFORMATICS METHODS

Ralf Gabriels

Dissertation submitted to obtain the degree Doctor in Health Sciences

Academic year 2021-2022

GHENT UNIVERSITY

FACULTY OF MEDICINE AND HEALTH SCIENCES

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

CompOmics

fwo

Introduction (Dutch)

proteins and proteomics

mass spectrometry

machine learning

Results (English)

MS²PIP: Peptide spectrum prediction for multiple fragmentation methods, instruments, and labeling techniques

Removing the hidden data dependency of DIA with predicted spectral libraries

MS²Rescore: Leveraging spectrum predictions to enable novel proteomics workflows

MS²DIP: spectrum prediction for modified peptides

Discussion and future perspectives (English)

proteins and proteomics

Image: sciencenewsforstudents.org

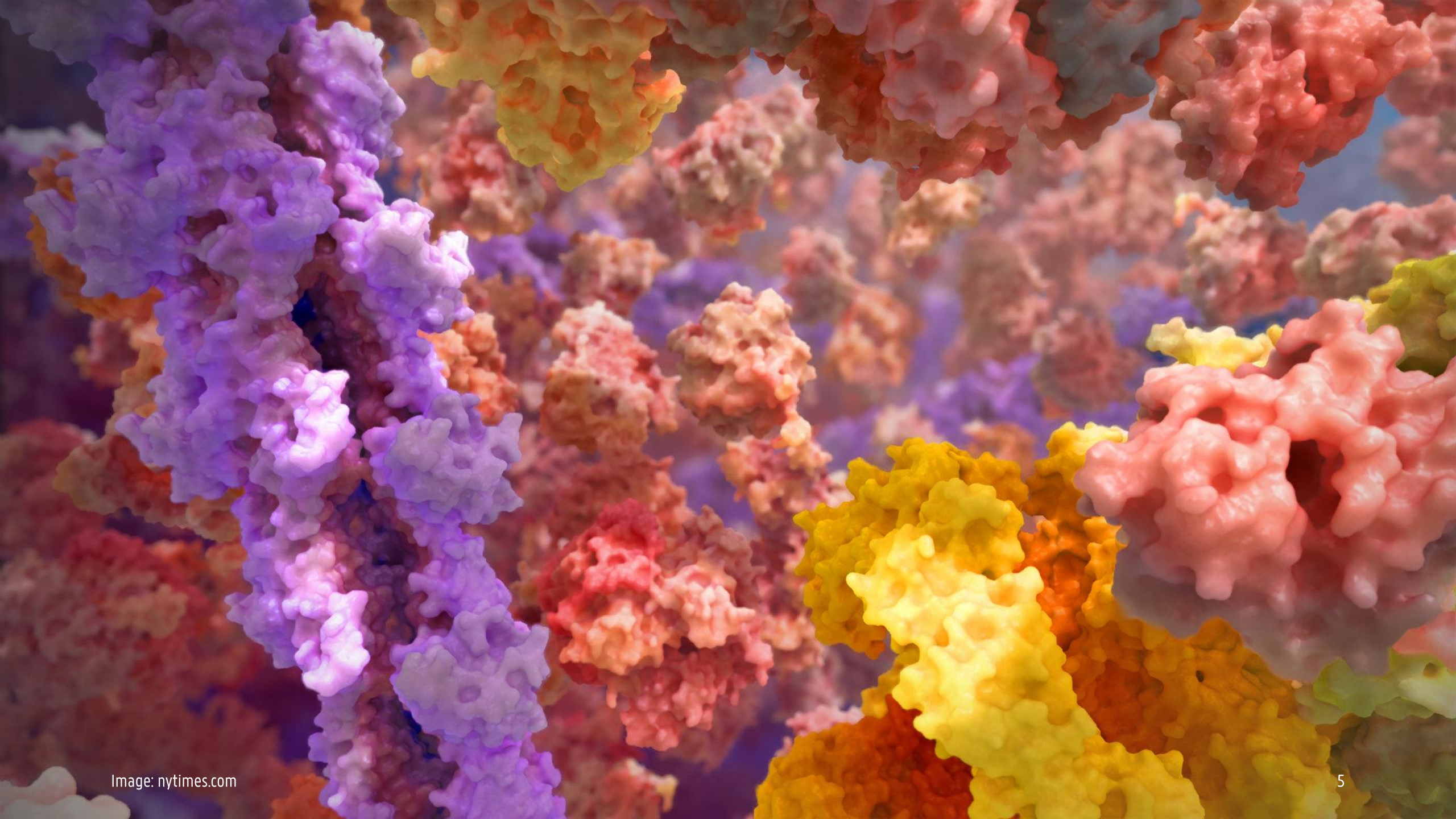


Image: nytimes.com

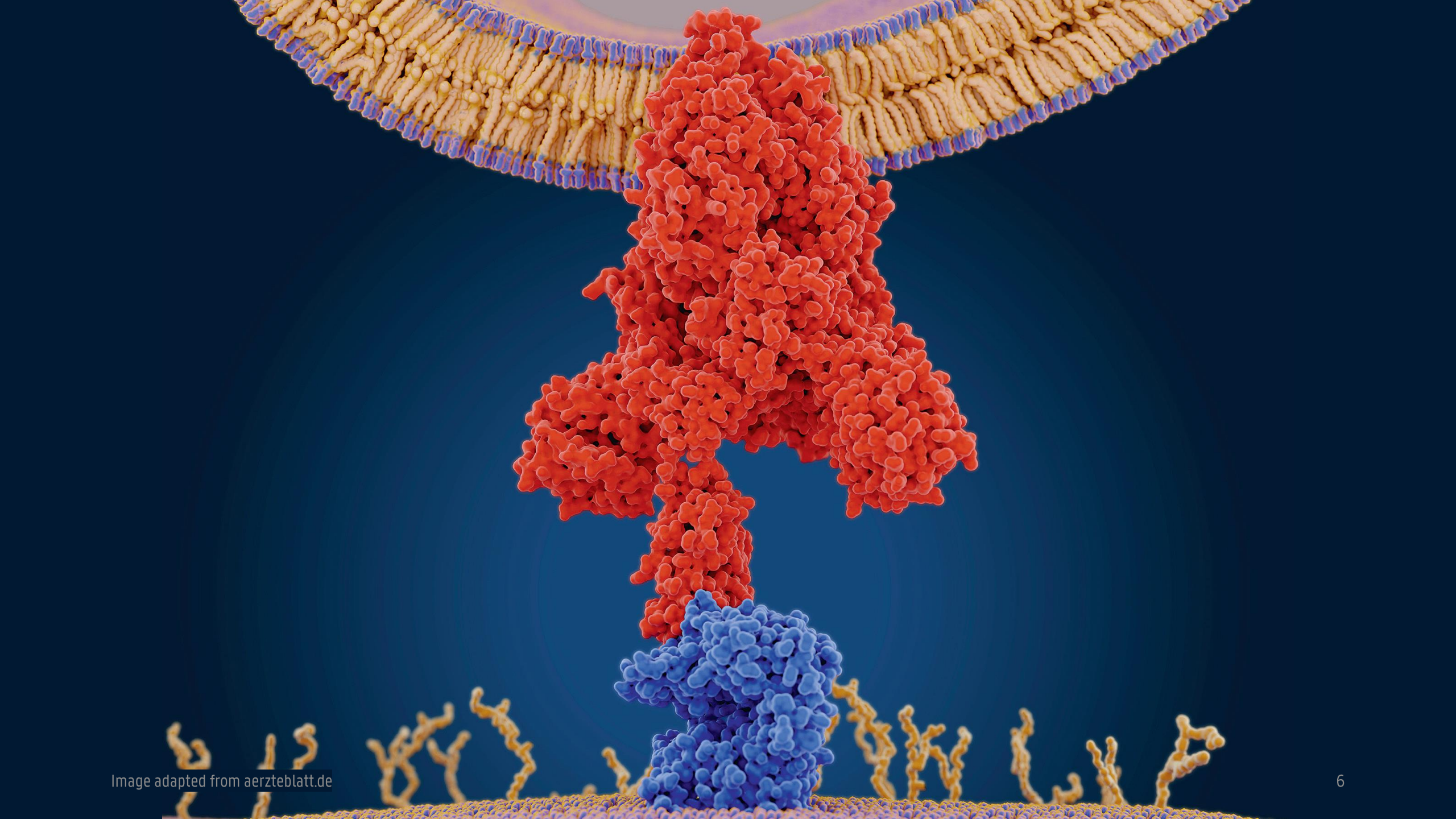
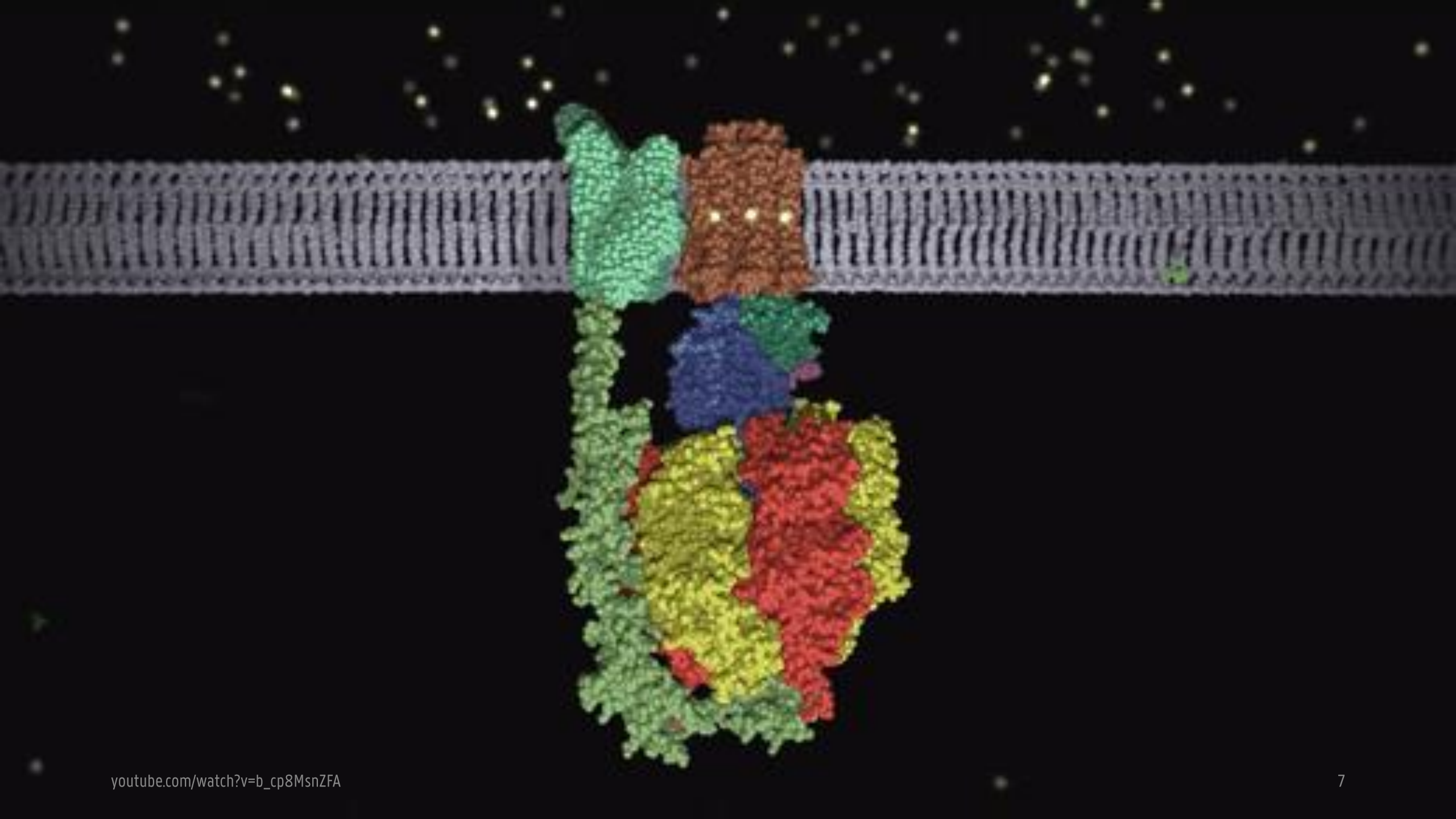
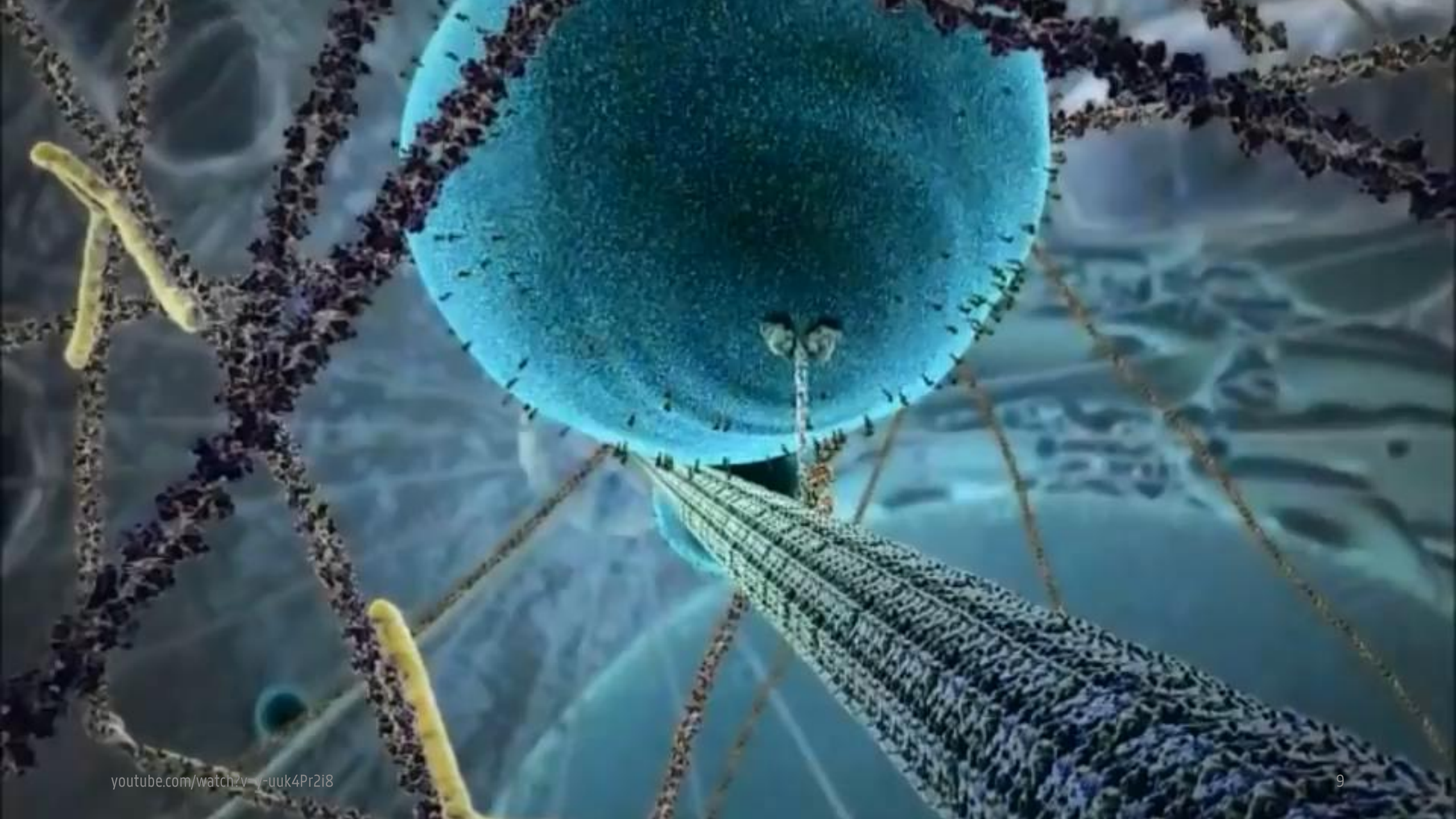


Image adapted from aerzteblatt.de



youtube.com/watch?v=b_cp8MsnZFA

youtube.com/watch?v=XI8m6oOgXDY



youtube.com/watch?v=y-uuk4Pr2i8

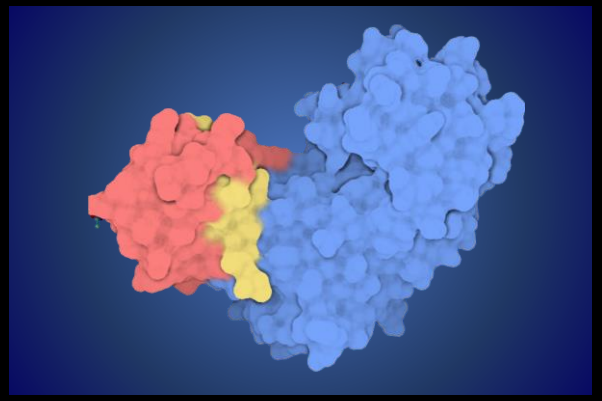
Central dogma of molecular biology

DNA / GENE

ATGGCGGCGGCCGGAGATAG
TGGGGCCTACGATTGGG[…]

RNA / TRANSCRIPT

AUGGCGGCGGCCGGAGAUAG
UGGGGCCUACGAUUGGG[…]



PROTEIN

MGVPSGLILCVLIGAFFISM
AAAGDSGAYDWVMPARS[…]

source

recipe

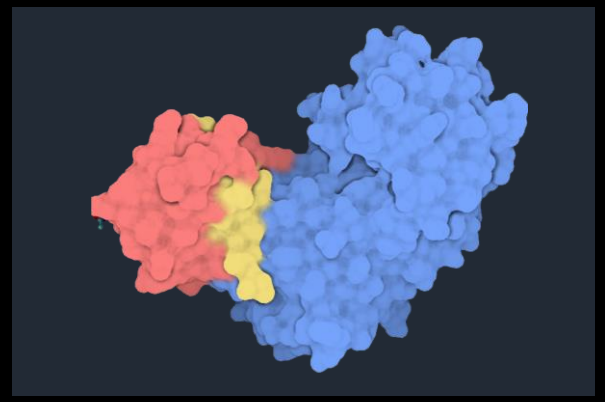
delicious waffles

DNA / GENE

TTGCCCATTAGCTGTTGTCT
TAGTTCAAGATTTTGTT[…]

RNA / TRANSCRIPT

AUGGCGGCGGCCGGAGAUAG
UGGGGCCUACGAUUGGG[…]



PROTEIN

MGVPSGLILCVLIGAFFISM
AAAGDSGAYDWVMPARS[…]

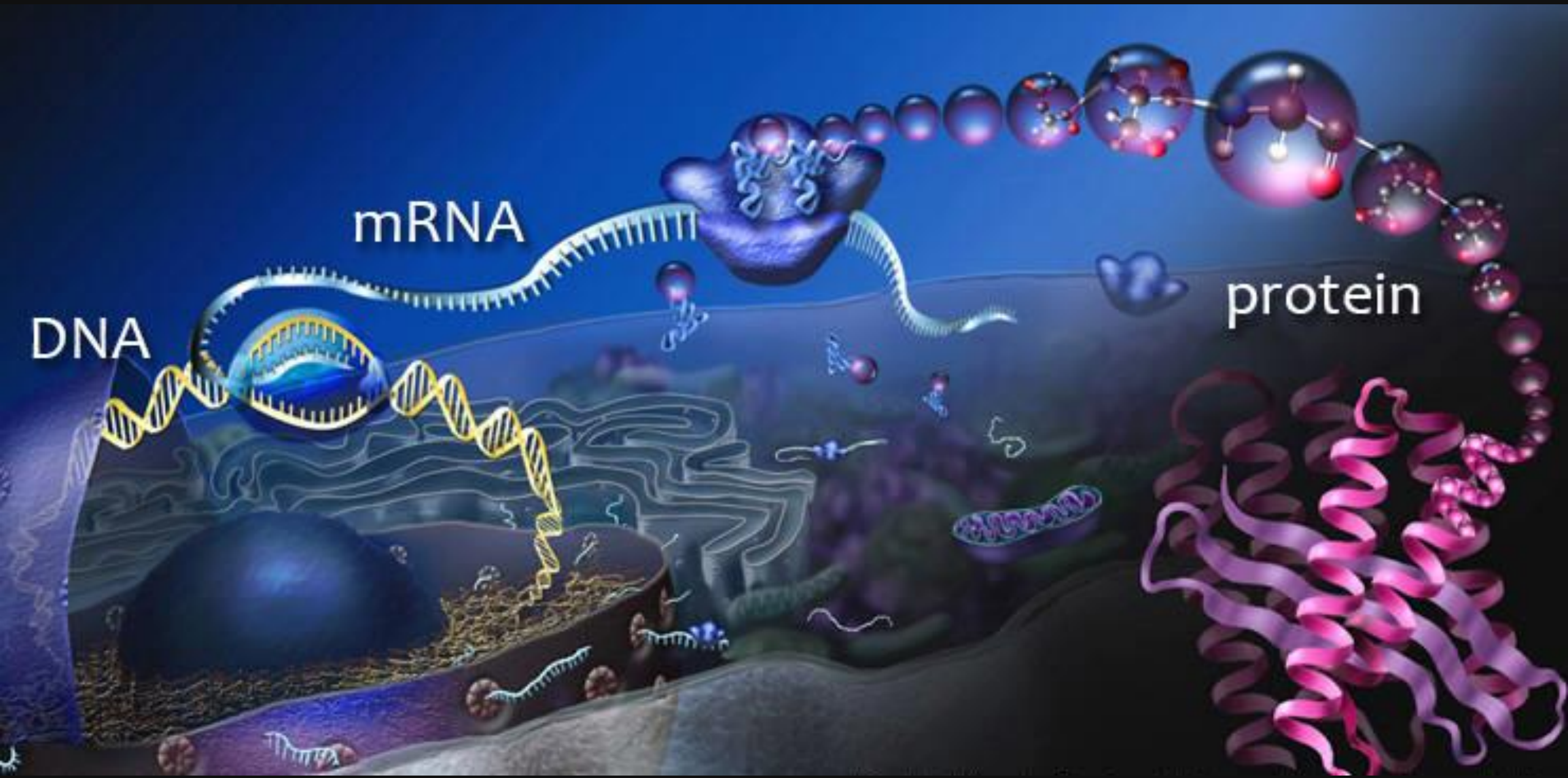
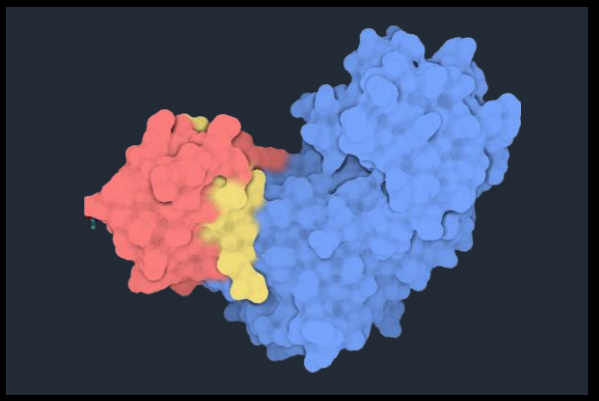


Image: Nicolle Rager, National Science Foundation

DNA / GENE

TTGCCCATTAGCTGTTGTCT
TAGTTCAAGATTTTGTT[…]

4 characters
linear
1 or 2 copies per cell



PROTEIN

MGVPSGLILCVLIGAFFISM
AAAGDSGAYDWVMPARS[…]

20 characters
linear folded into 3D structure
1 – 10^7 copies per cell
post-translational modifications

-omics

= study of biomolecules in full context (holistic)

molecule
DNA
RNA
protein

-omics

= study of biomolecules in full context (holistic)

molecule	entity name
DNA	gene
RNA	transcript
protein	protein

-omics

= study of biomolecules in full context (holistic)

molecule	entity name	collection
DNA	gene	genome
RNA	transcript	transcriptome
protein	protein	proteome

-omics

= study of biomolecules in full context (holistic)

molecule	entity name	collection	field of study
DNA	gene	genome	gen**omics**
RNA	transcript	transcriptome	transcript**omics**
protein	protein	proteome	prote**omics**

proteomics

= study of proteins in full context (holistic)

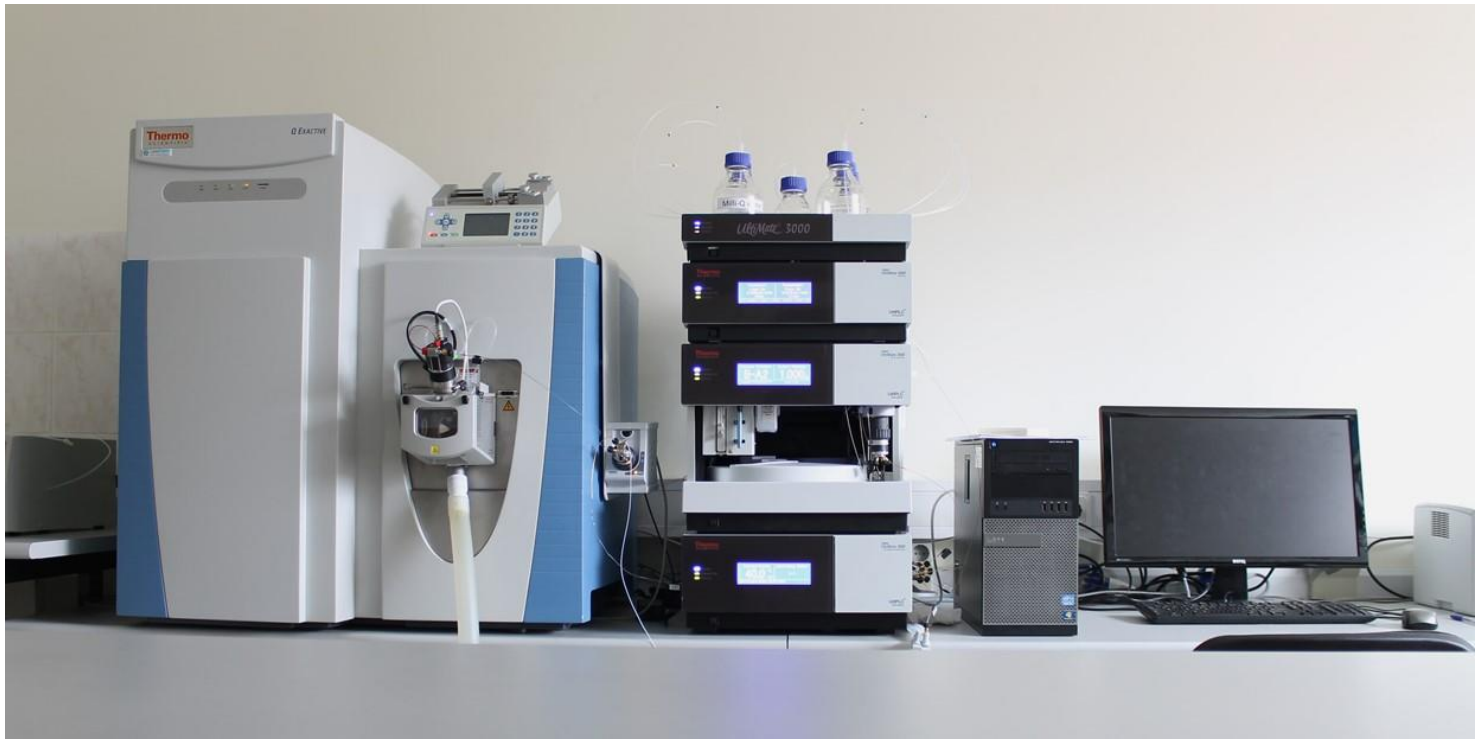


Image: scientificservices.eu

mass spectrometry

mass spectrometer

very sensitive molecular “scale”

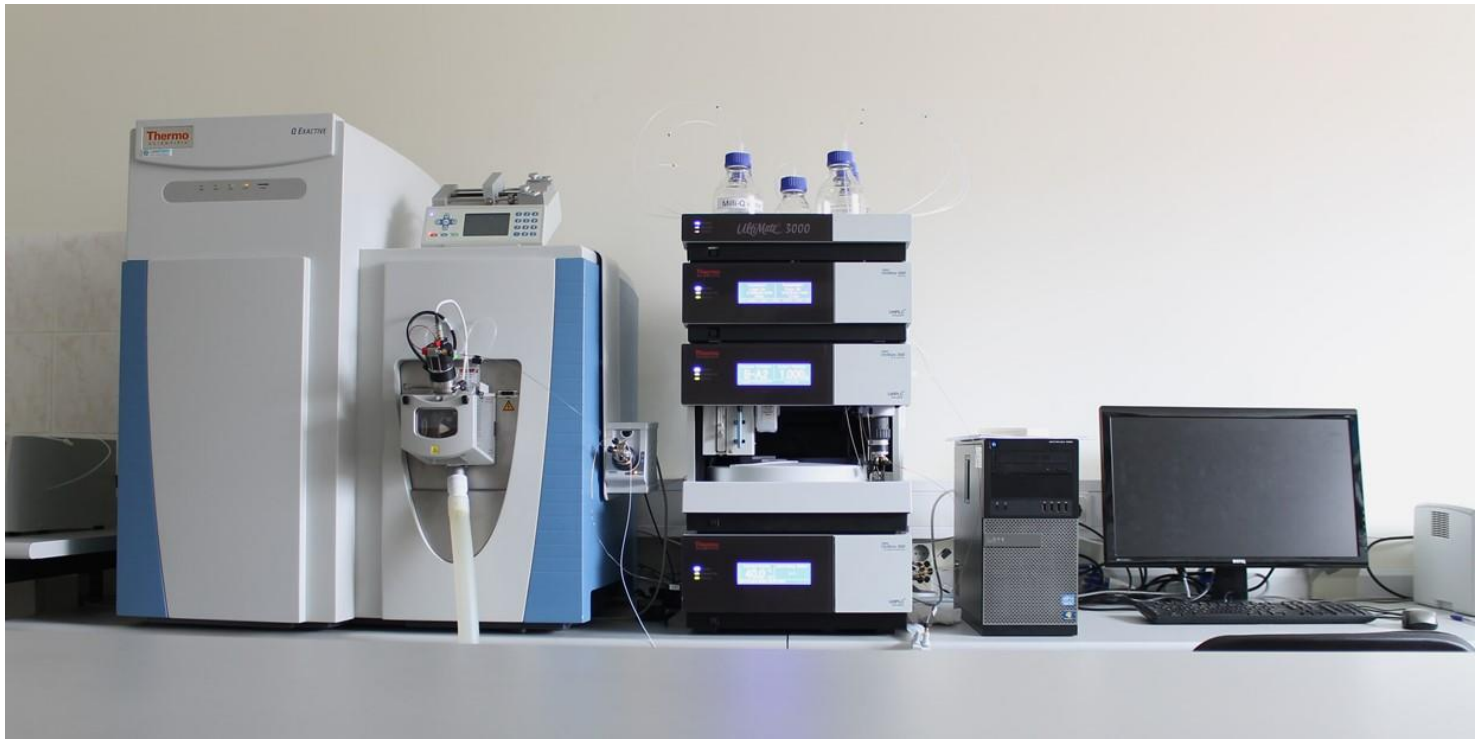


Image: scientificservices.eu

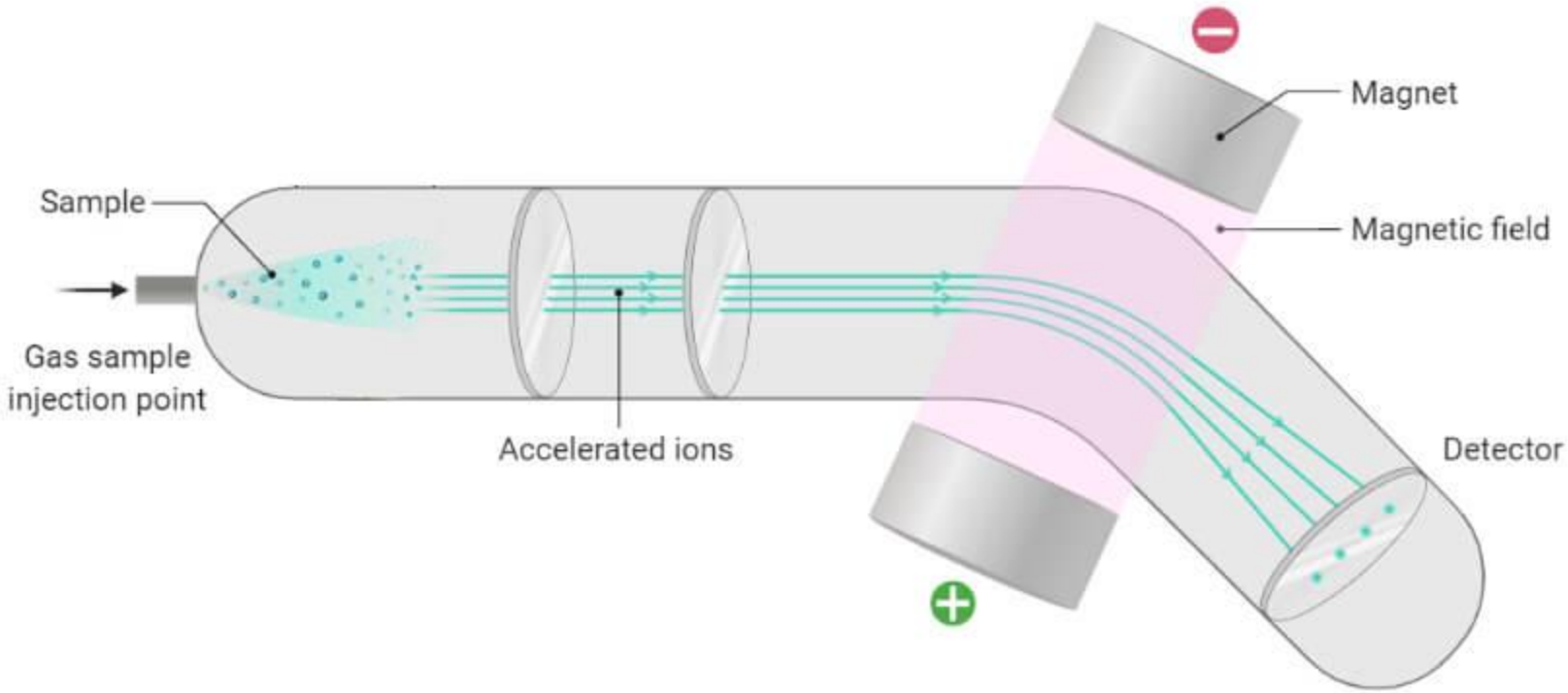


Image: microbenotes.com

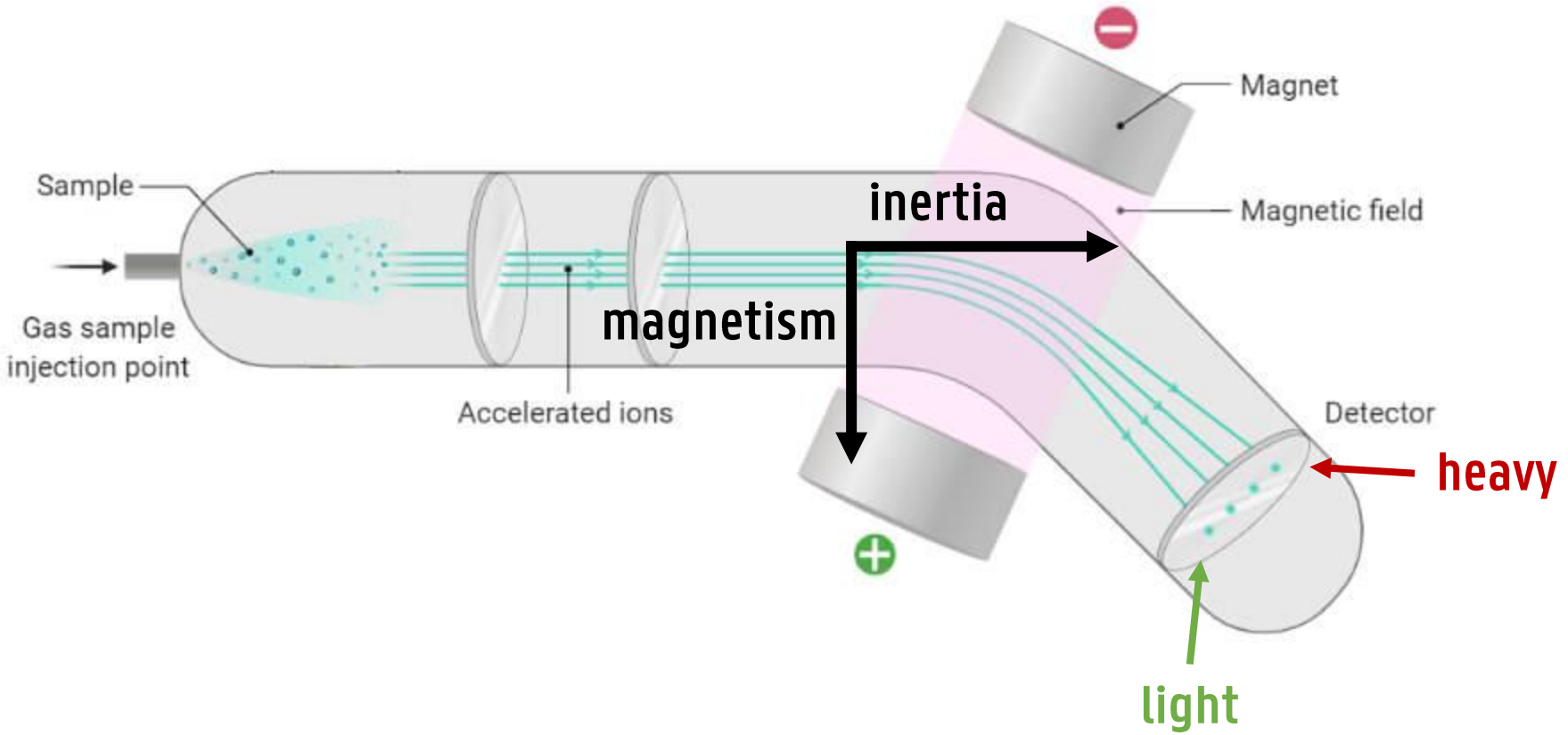


Image: microbenotes.com

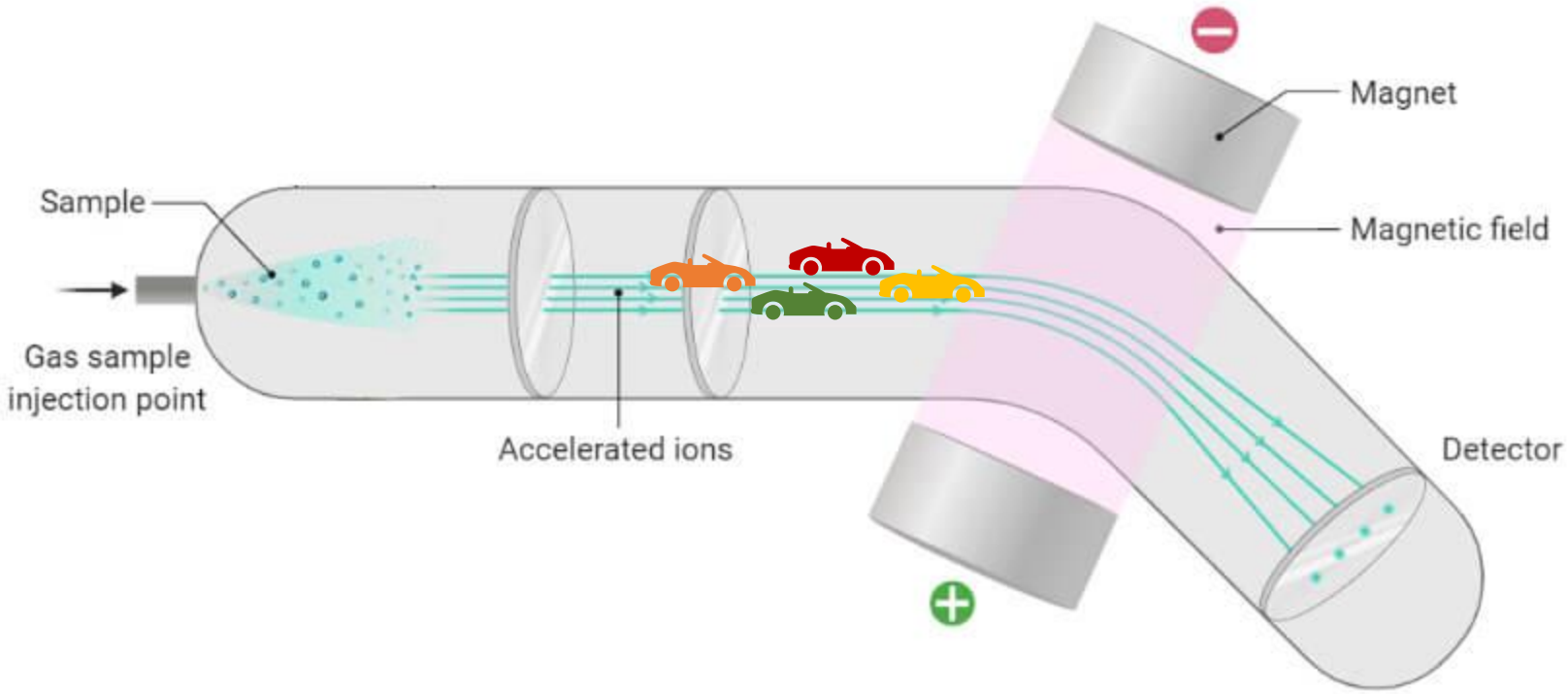


Image: microbenotes.com

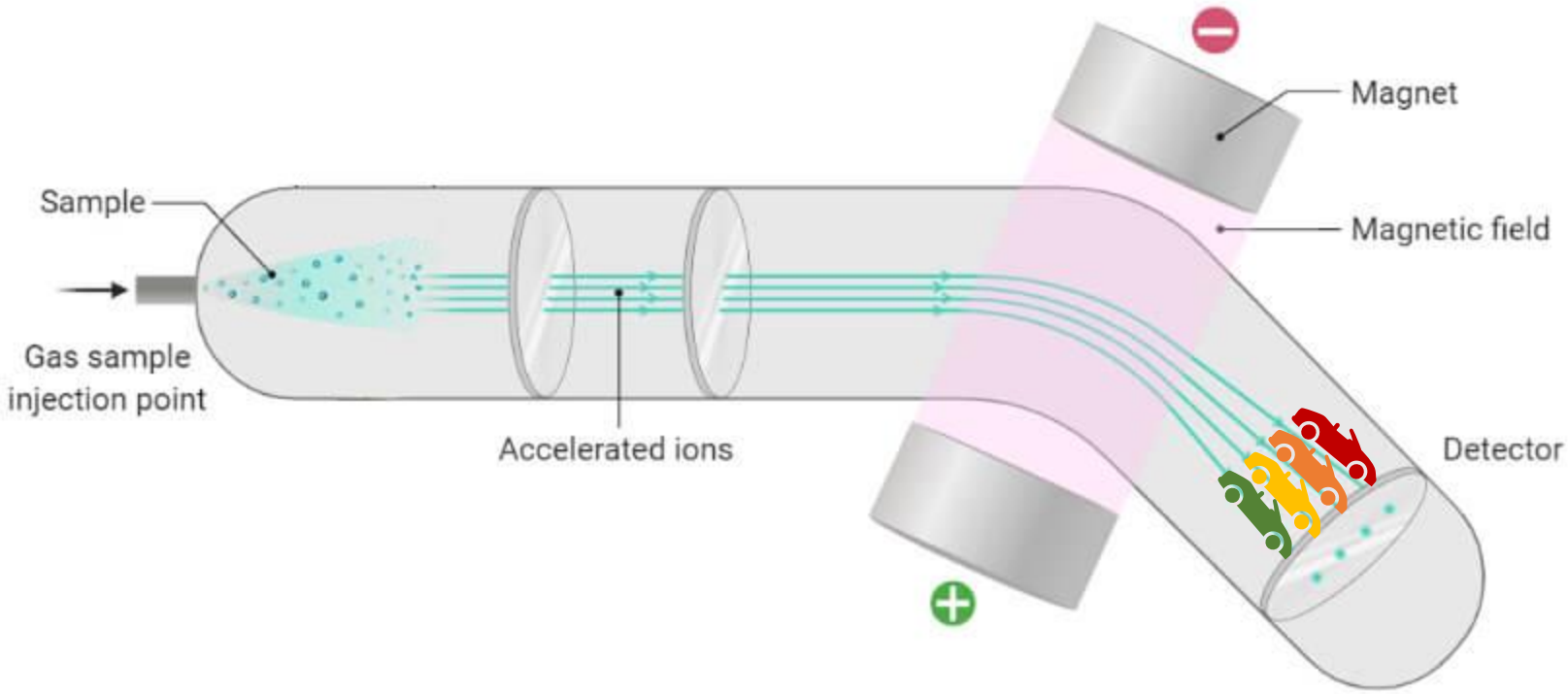
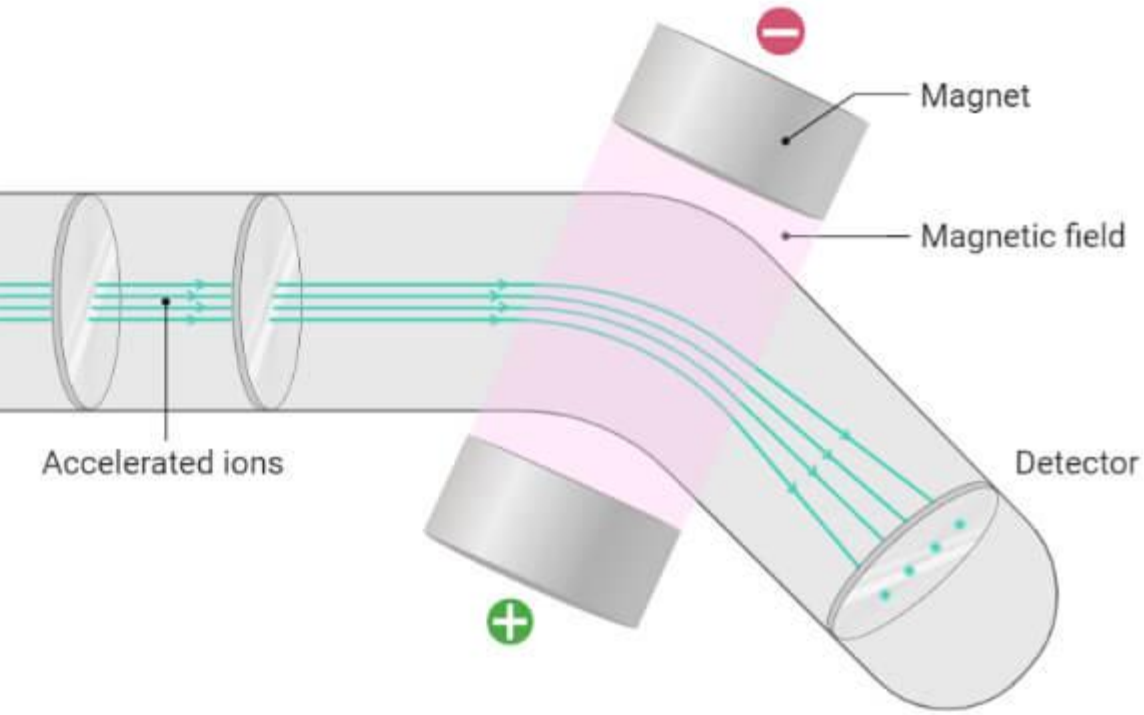


Image: microbenotes.com



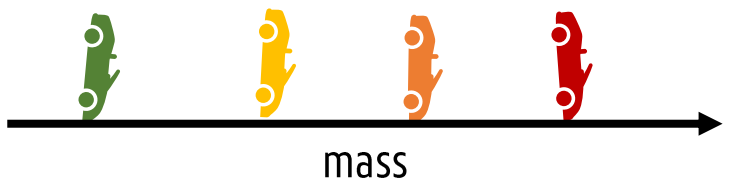
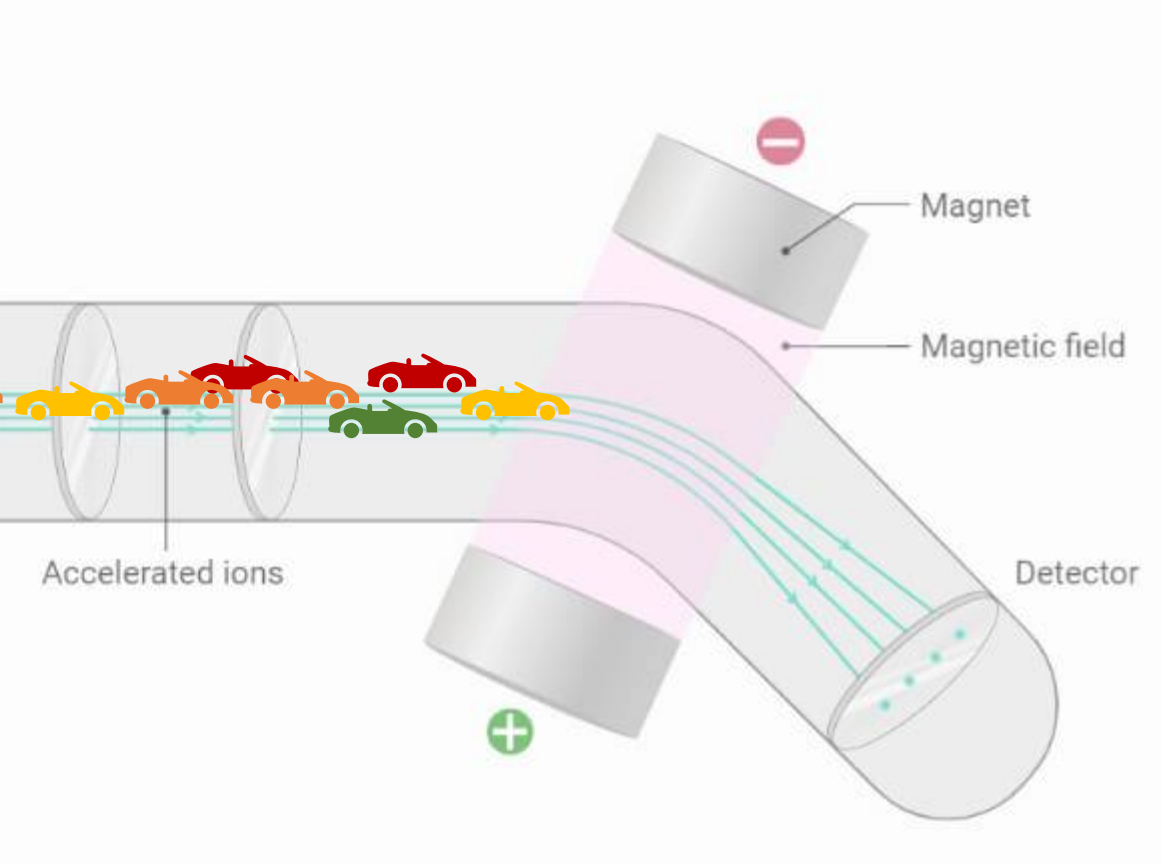


Image: microbenotes.com



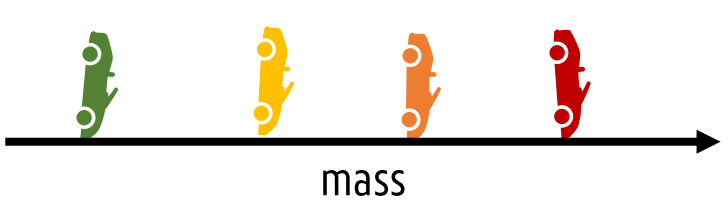
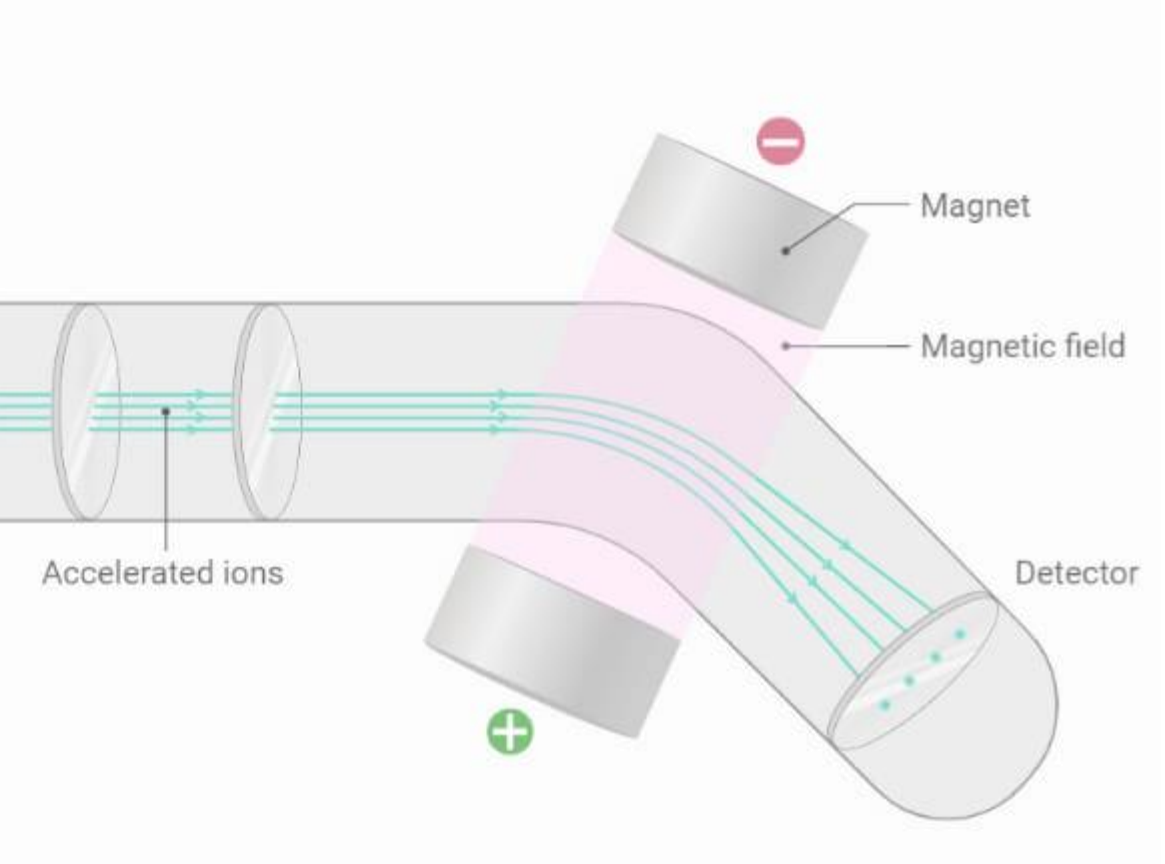


Image: microbenotes.com



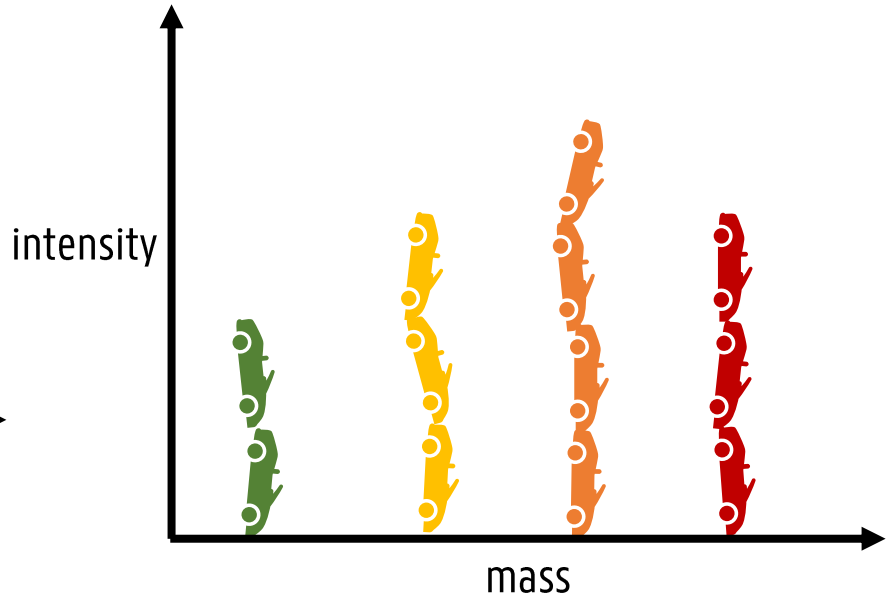


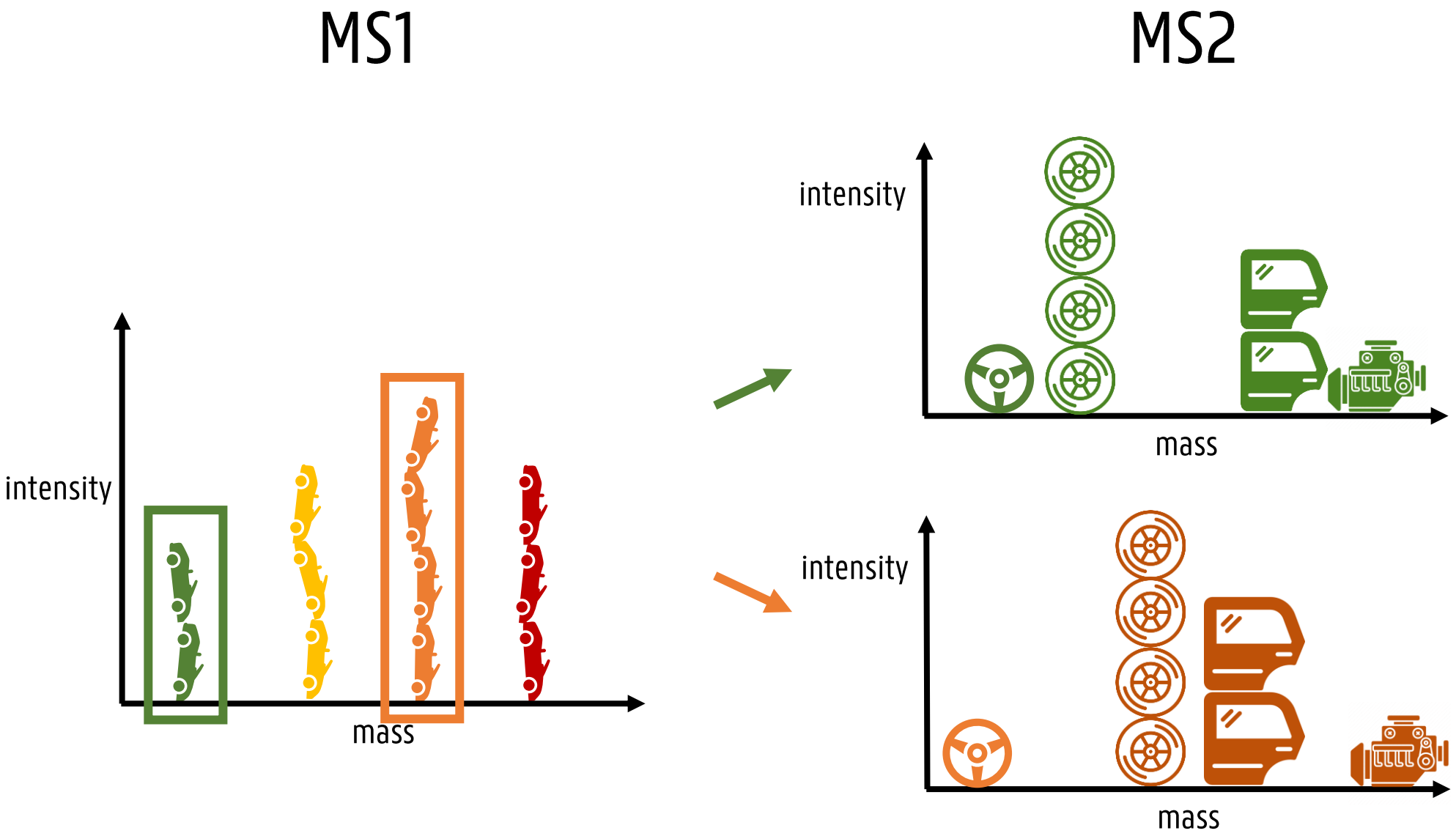
Image: microbenotes.com

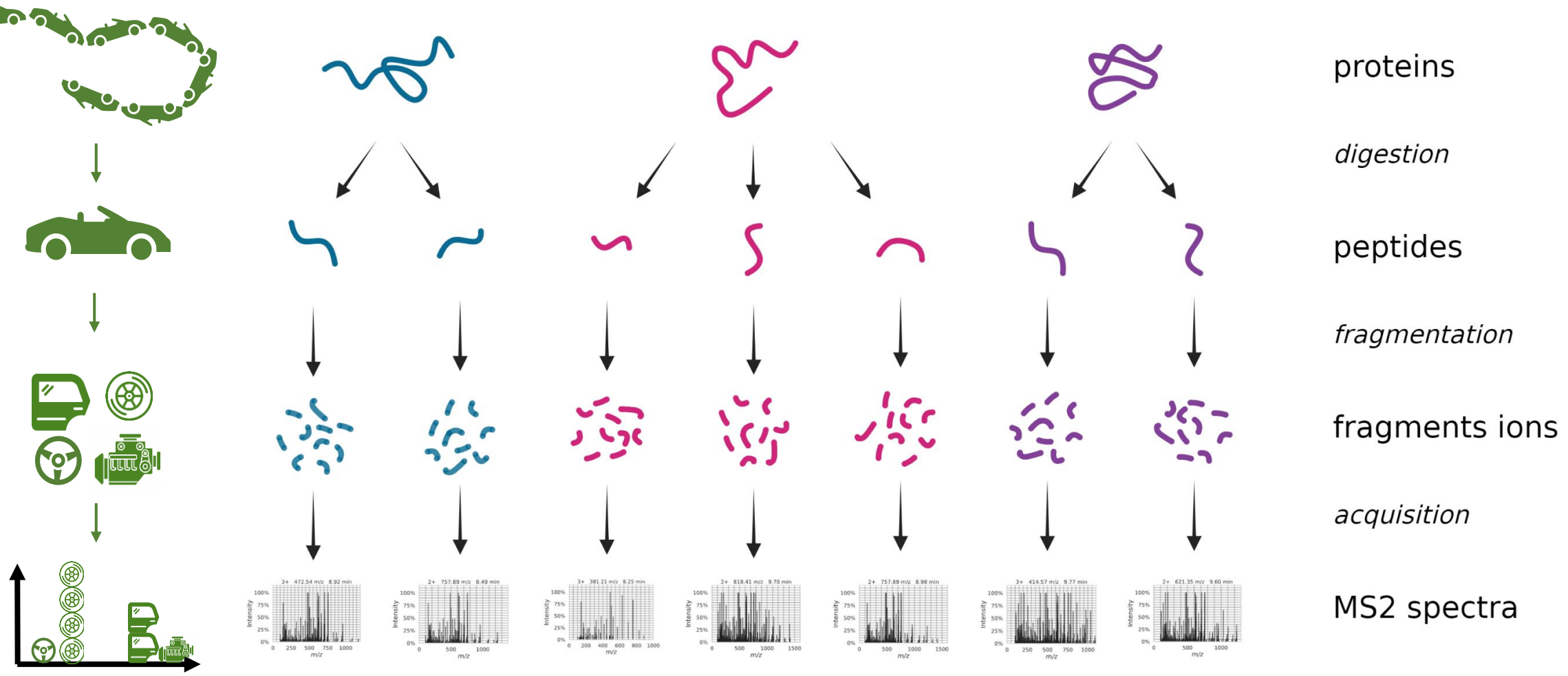

M FU 7055

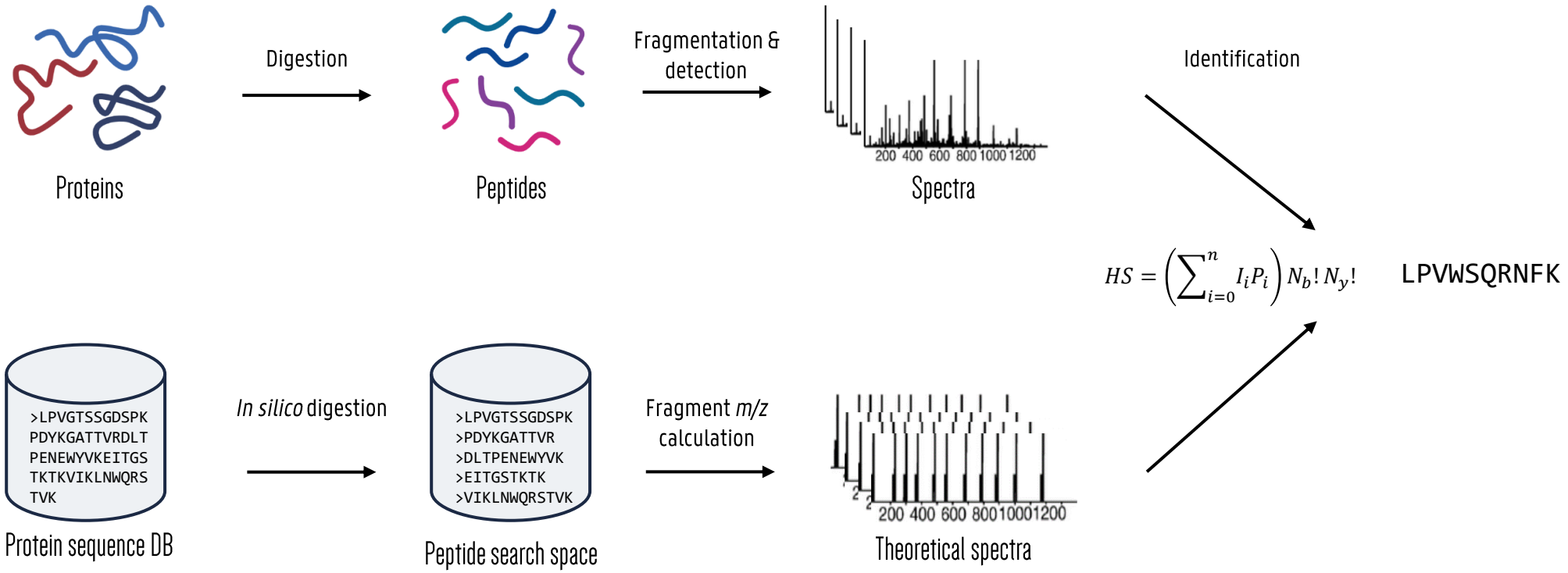

NEW YORK
MTR 75A9

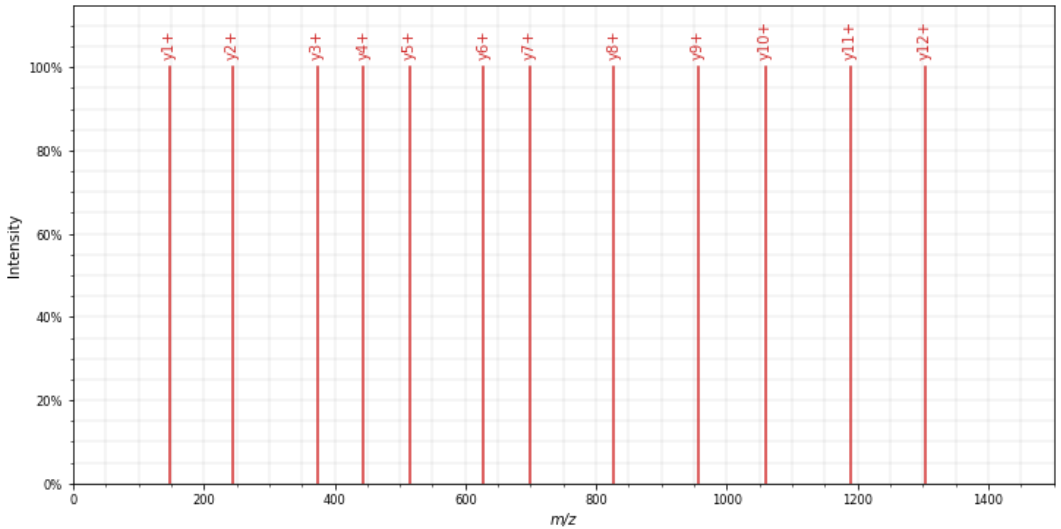




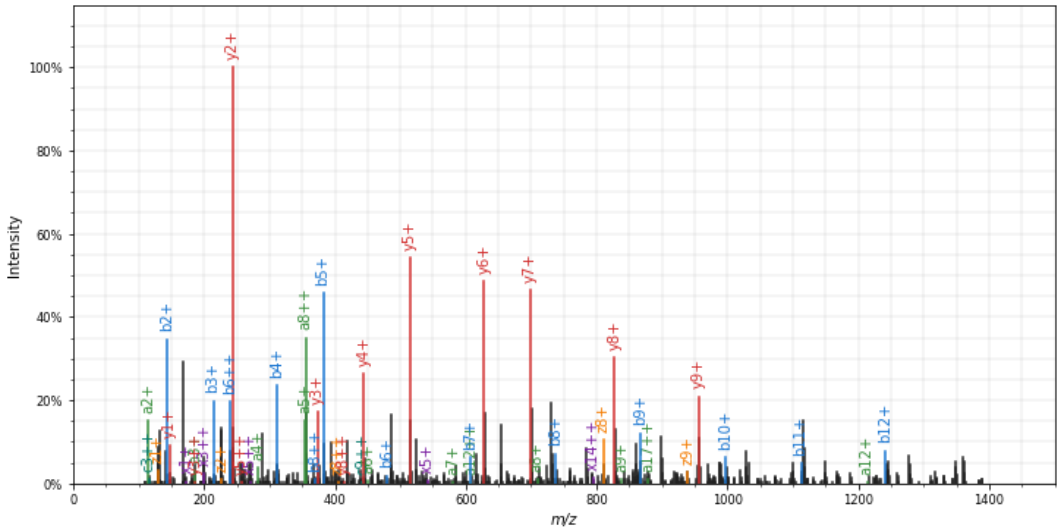

MS1
MS2
intensity
mass
intensity
mass
intensity
mass


proteins
digestion
peptides
fragmentation
fragments ions
acquisition
MS2 spectra


Proteins
Digestion
Peptides
Fragmentation & detection
200 400 600 800 1000 1200
Spectra
Identification
$HS = \left(\sum_{i=0}^{n} I_i P_i\right) N_b! \, N_y!$
LPVWSQRNFK
>LPVGTSSGDSPK
PDYKGATTVRDLT
PENEWYVKEITGS
TKTKVIKLNWQRS
TVK
Protein sequence DB
In silico digestion
>LPVGTSSGDSPK
>PDYKGATTVR
>DLTPENEWYVK
>EITGSTKTK
>VIKLNWQRSTVK
Peptide search space
Fragment *m/z* calculation
200 400 600 800 1000 1200
Theoretical spectra

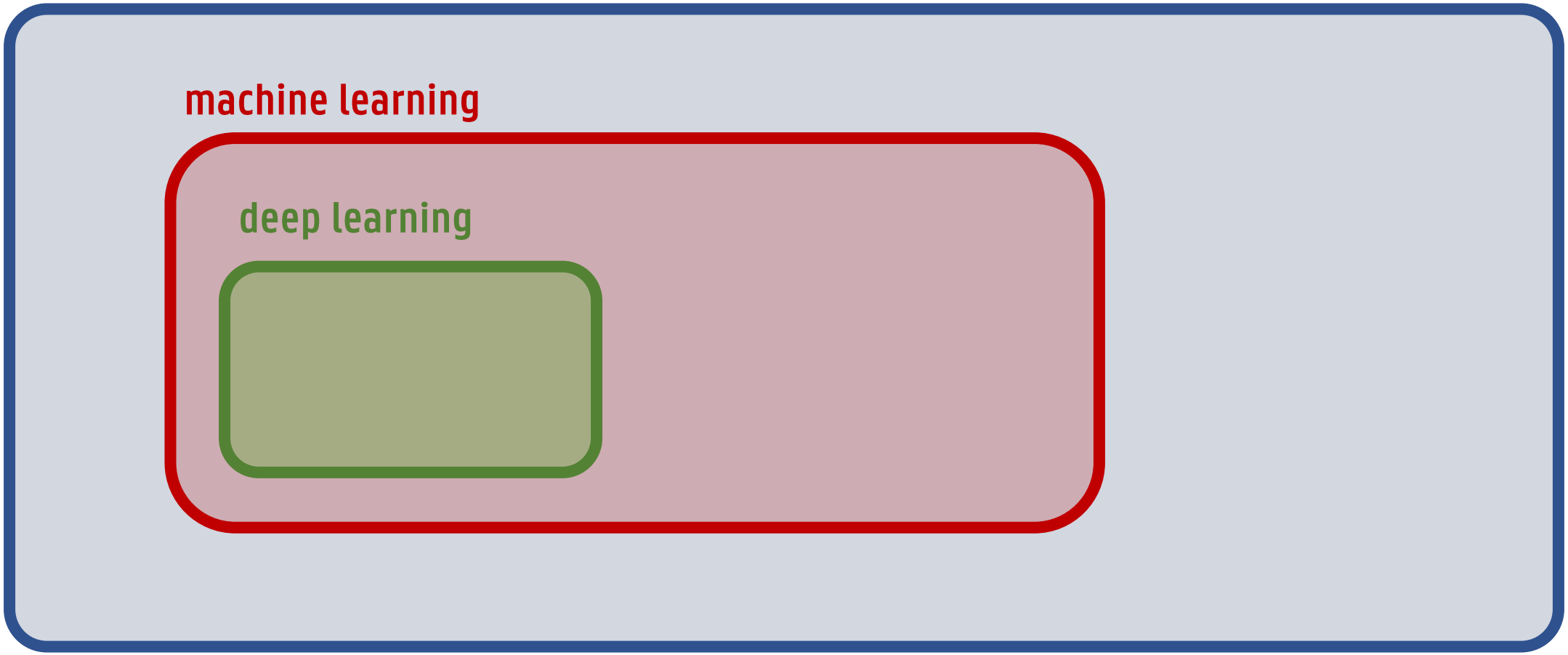


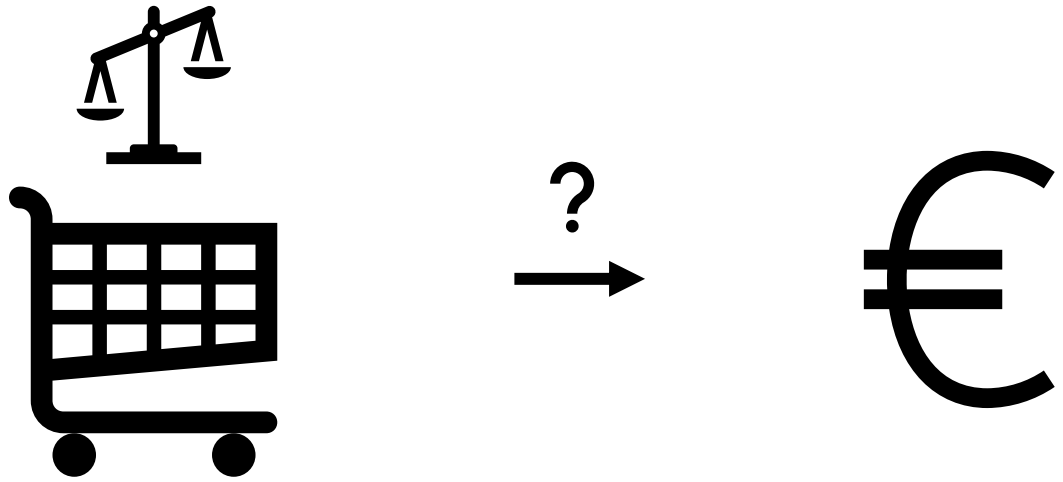
theory

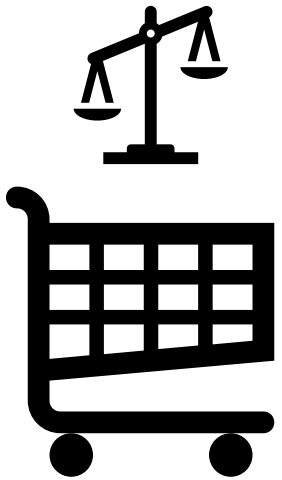


reality

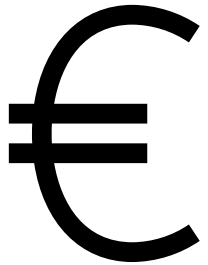
machine learning


artificial intelligence
machine learning
deep learning

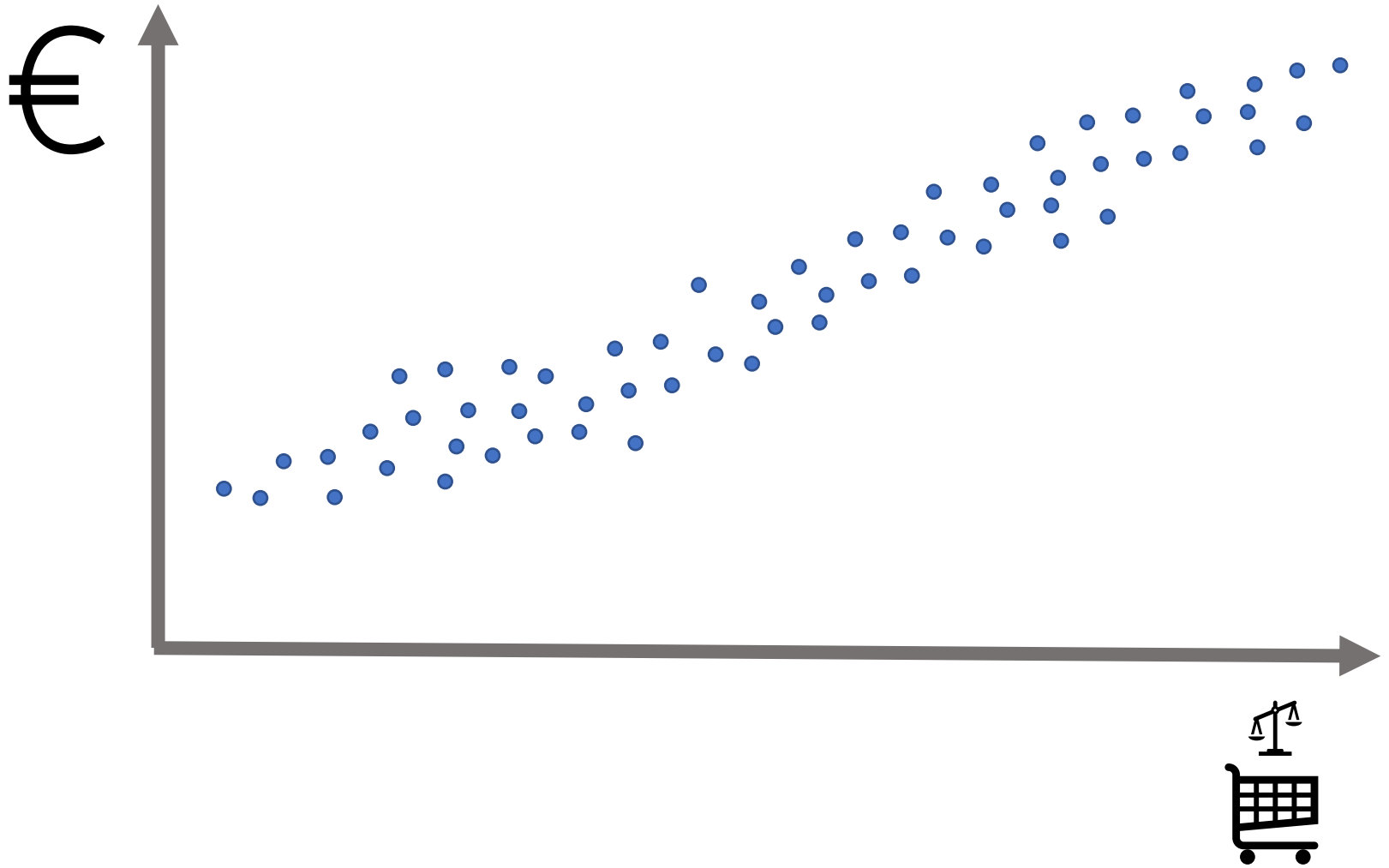


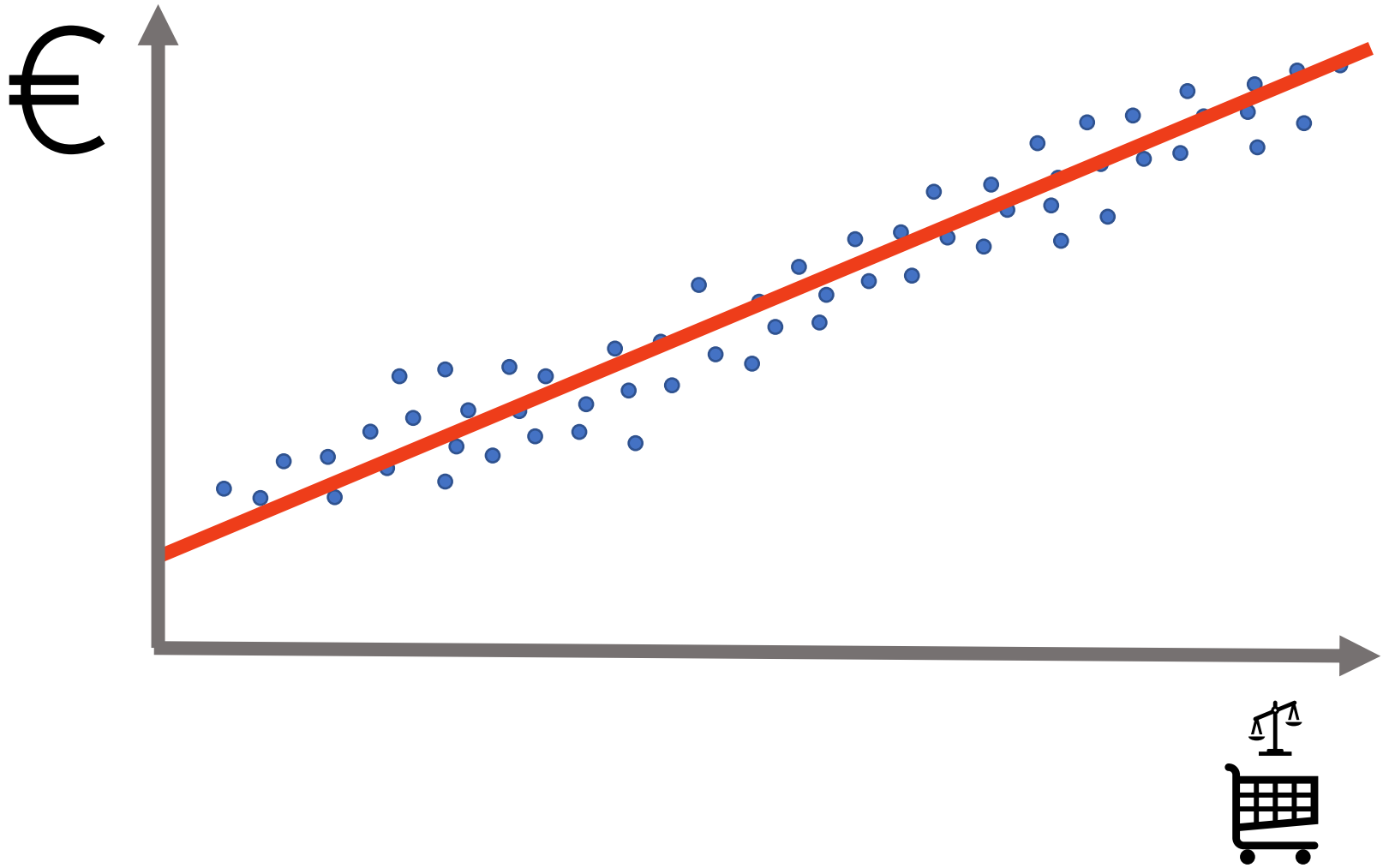


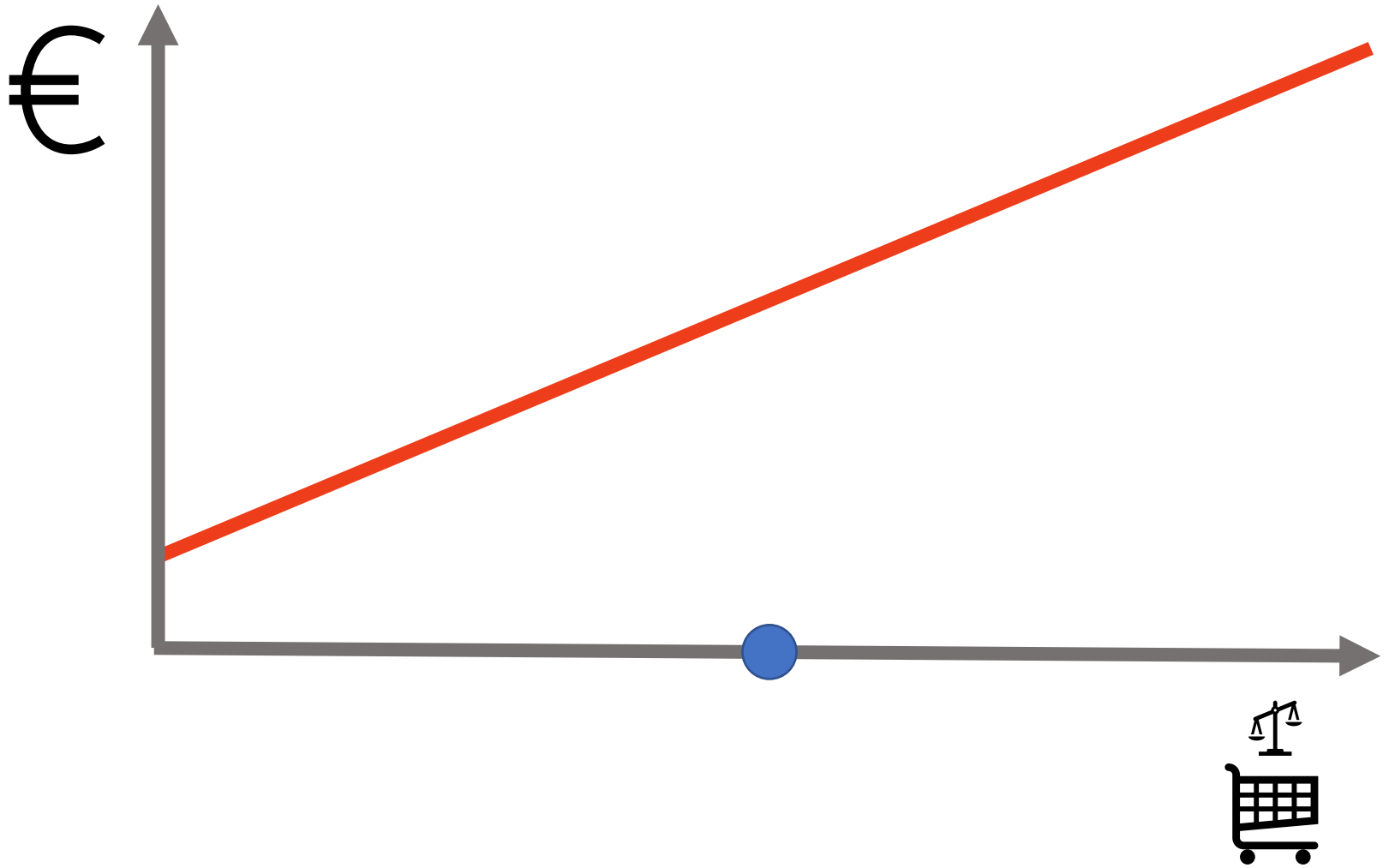
?

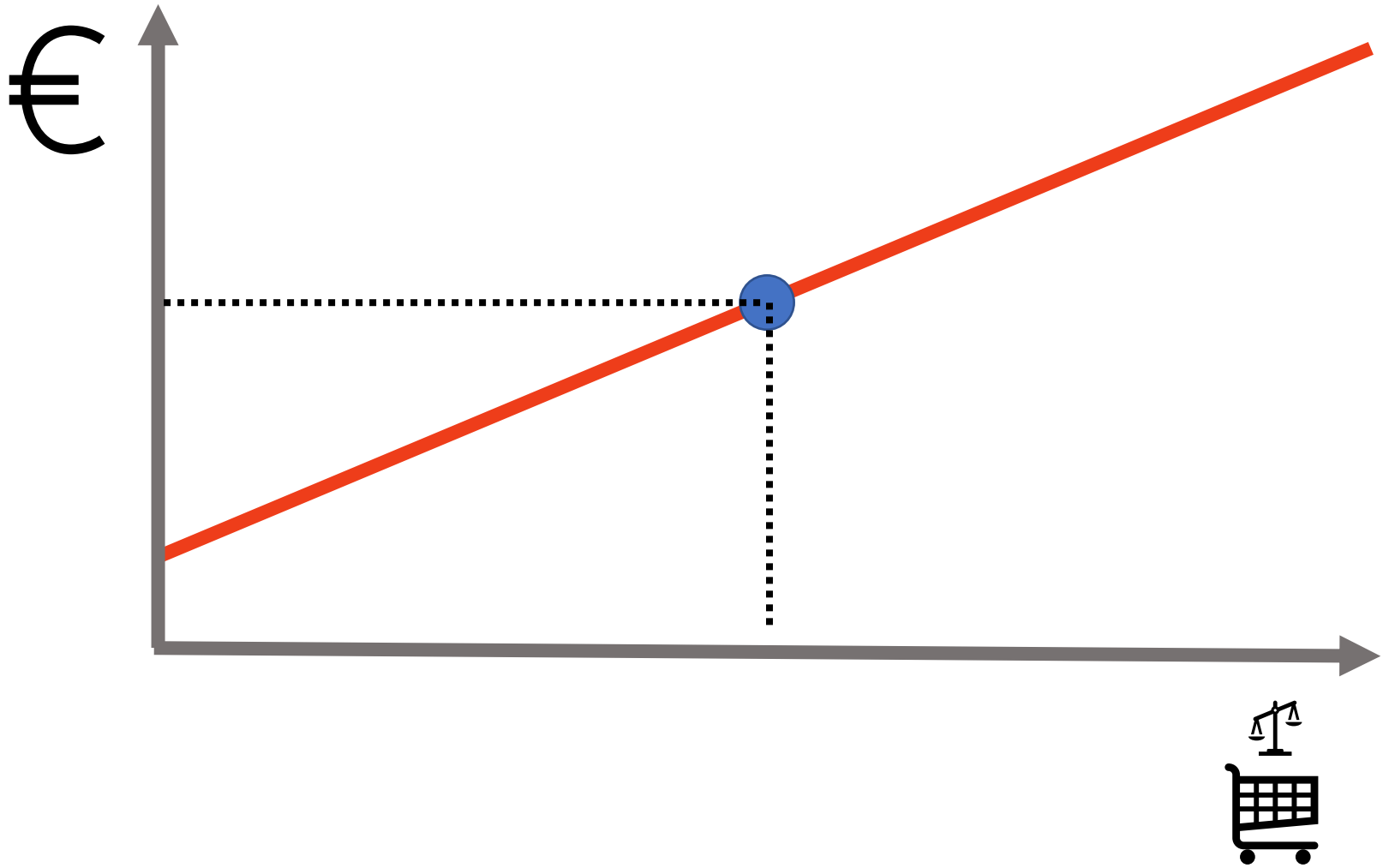


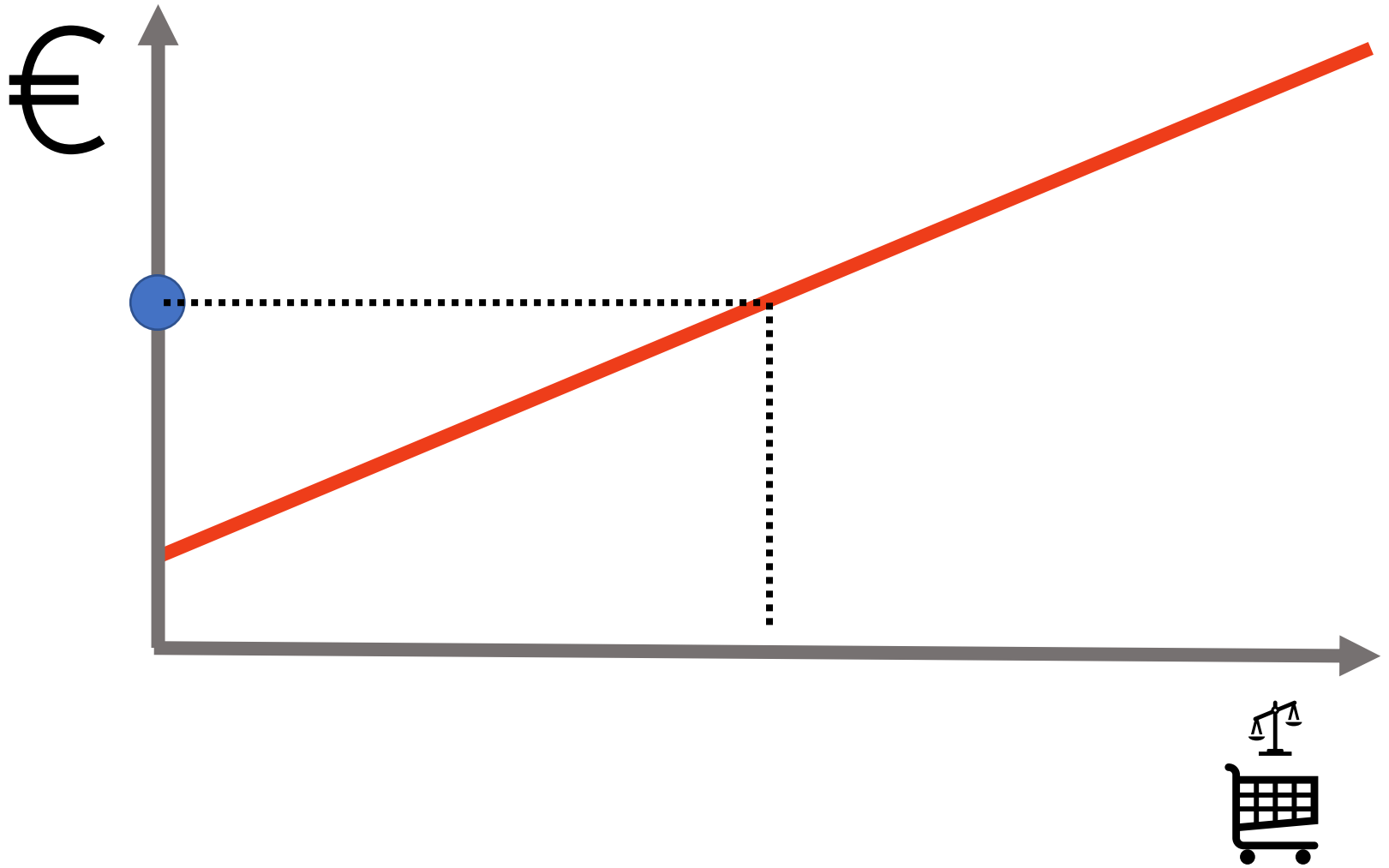
weight	cost
34 kg	€54
12 kg	€21
64 kg	€108
...	...

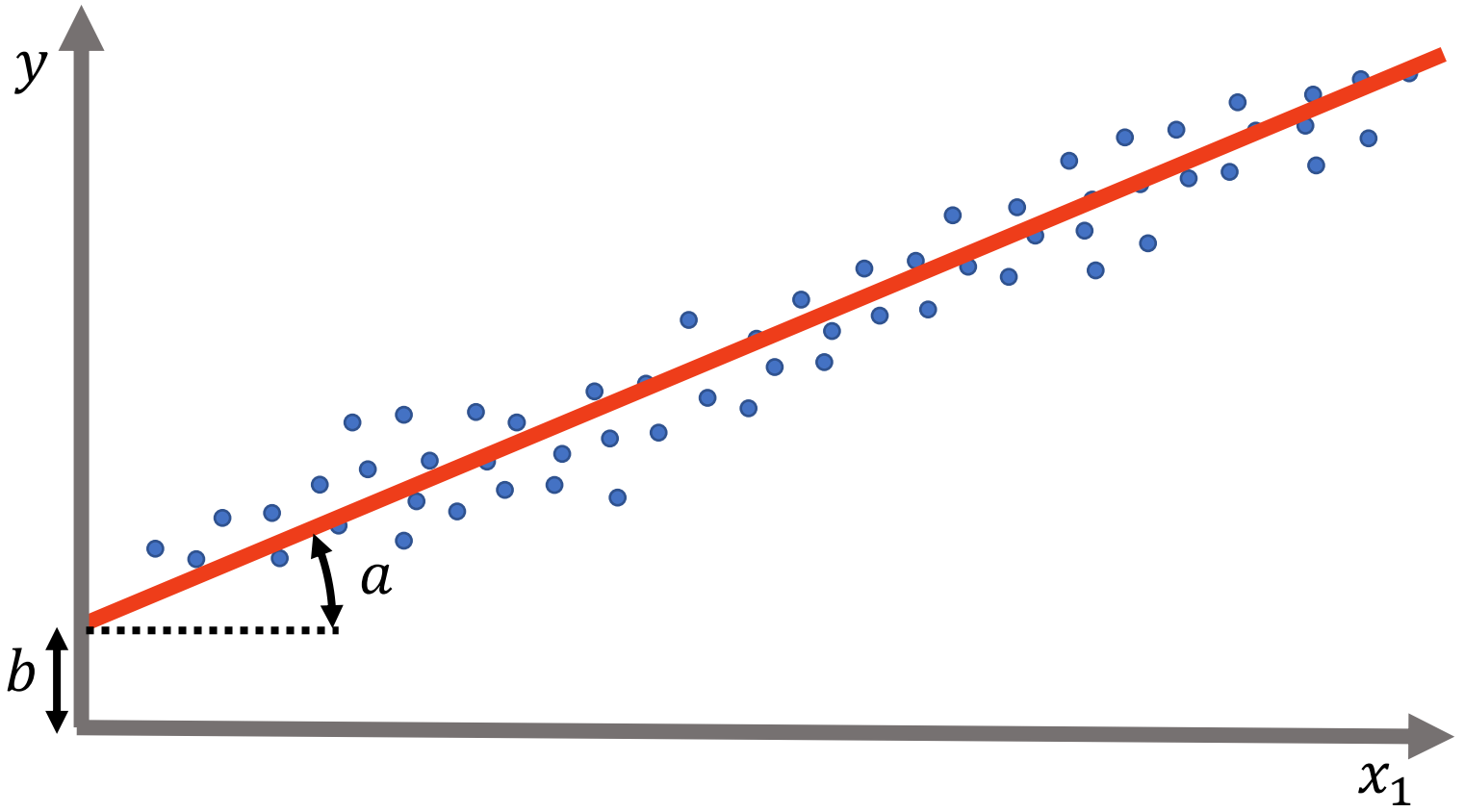












$$y = a \cdot x_1 + b$$

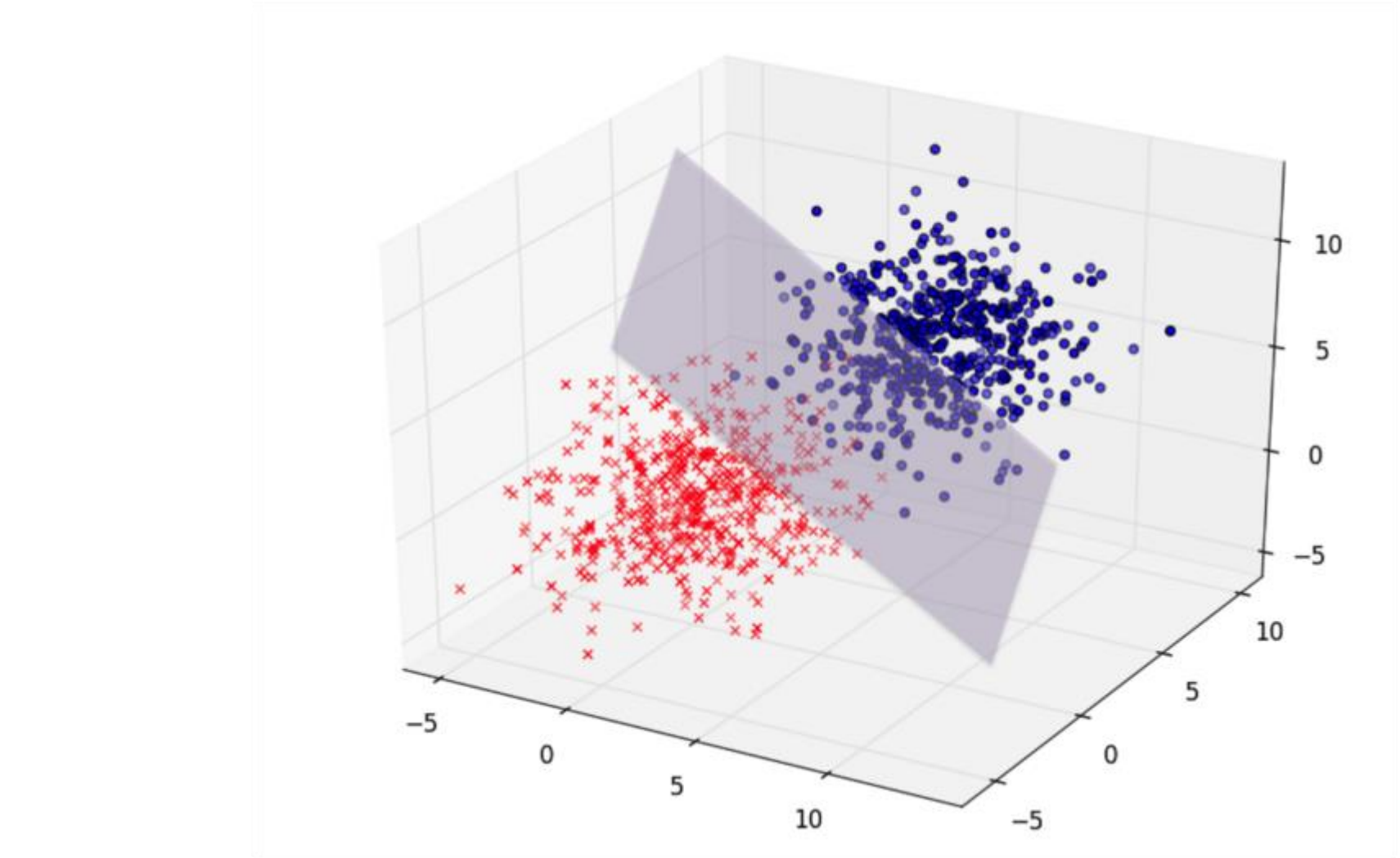
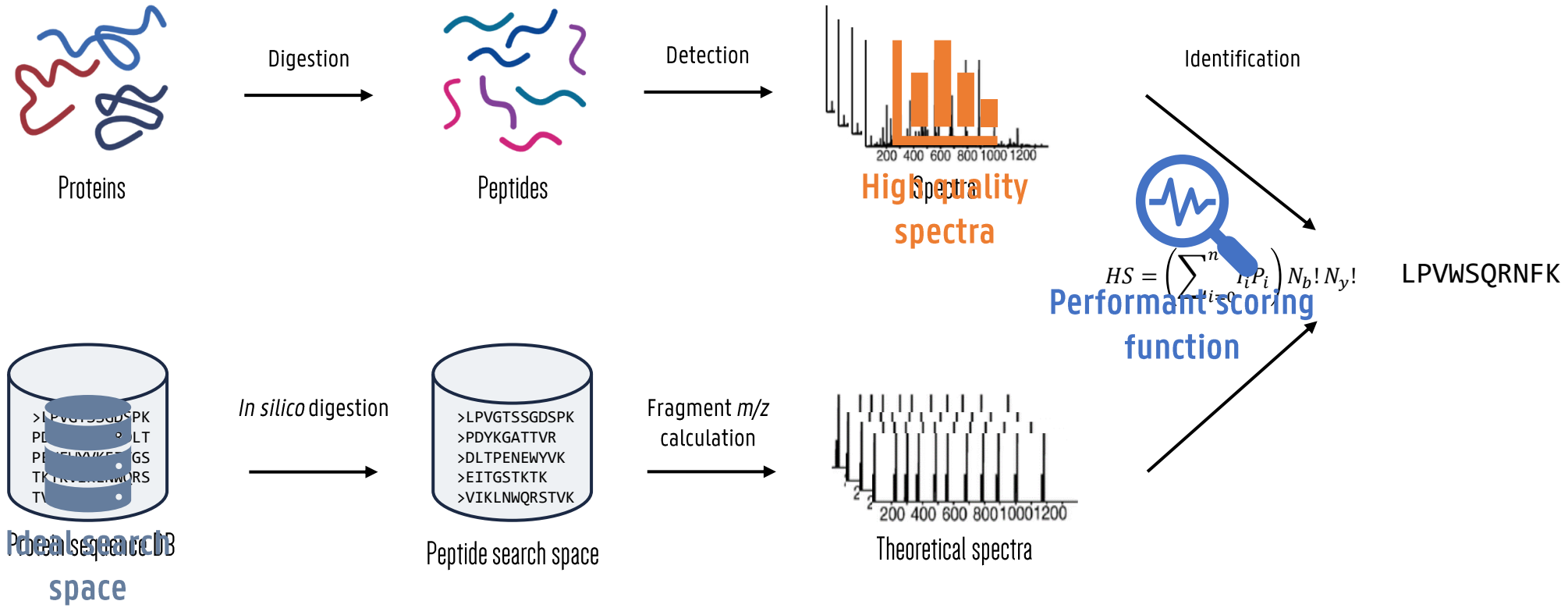


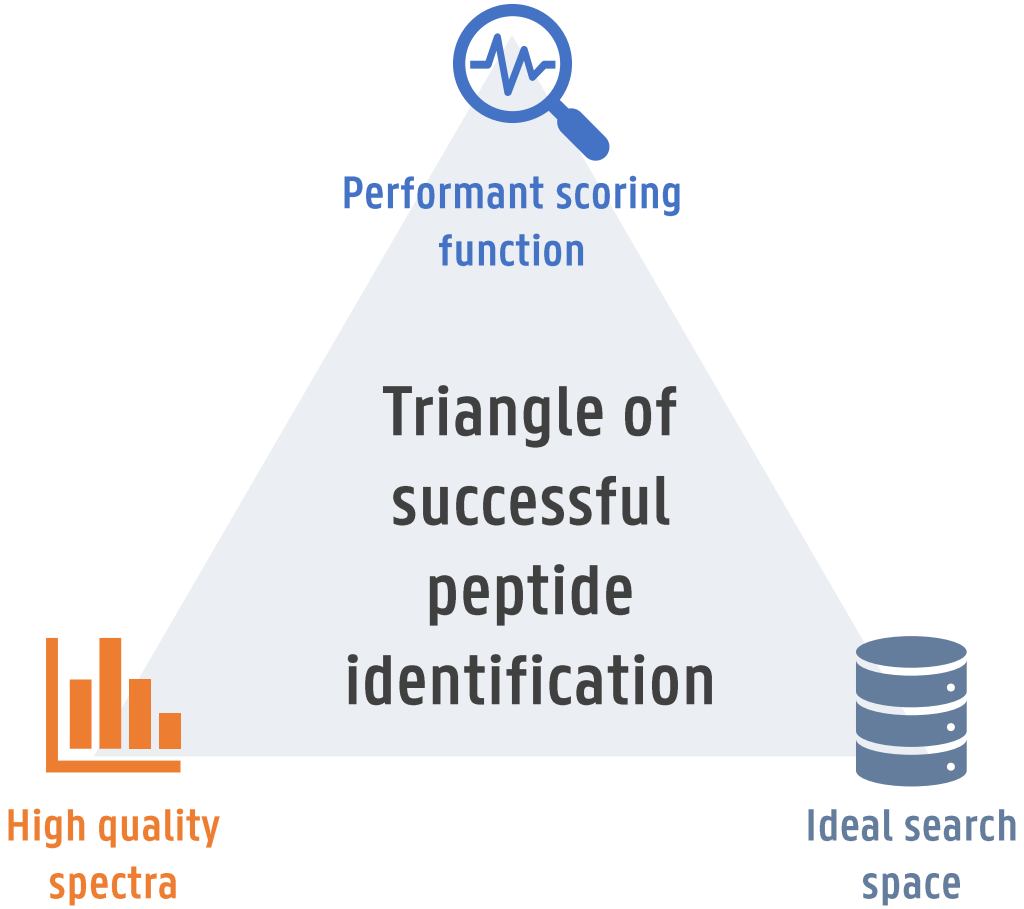
Image: projectrhea.org

proteomics
+ mass spectrometry
+ machine learning
= ?

Key elements to successful peptide spectrum identification?



Key elements to successful peptide spectrum identification?



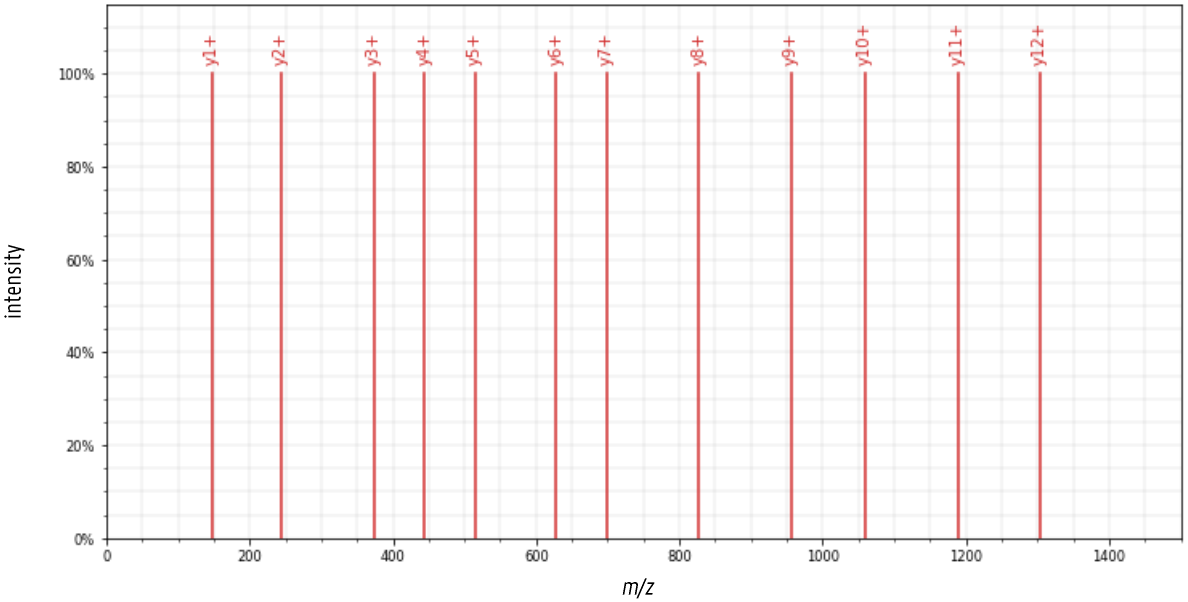
Key elements to successful peptide spectrum identification?

Triangle of successful peptide identification

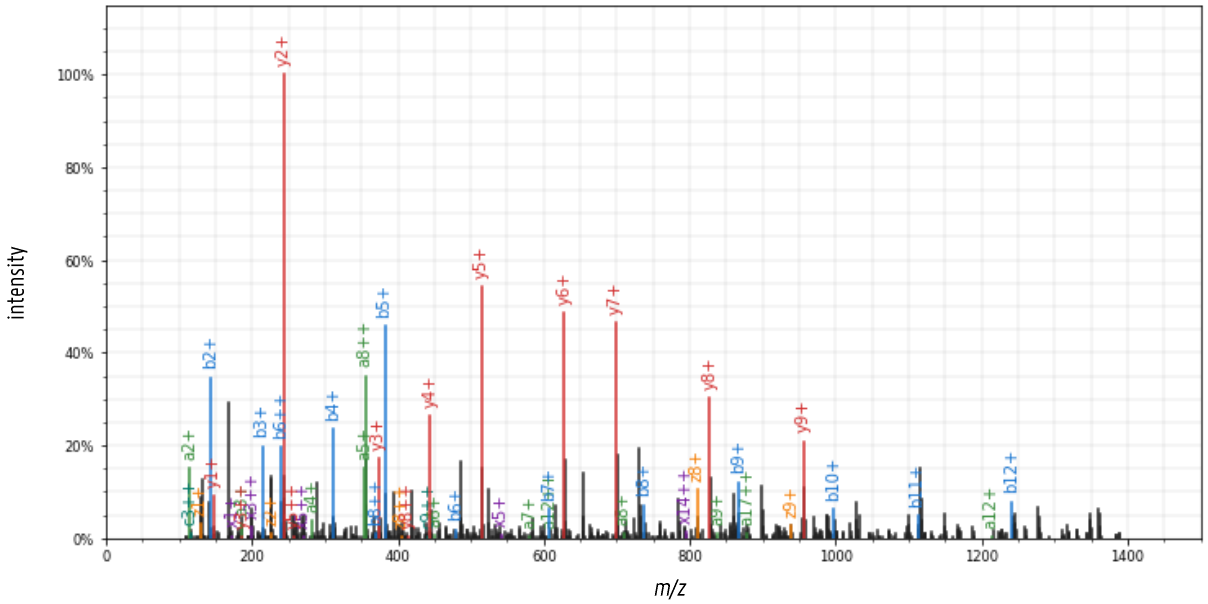
High quality spectra

Ideal search space

Machine learning to the rescue!

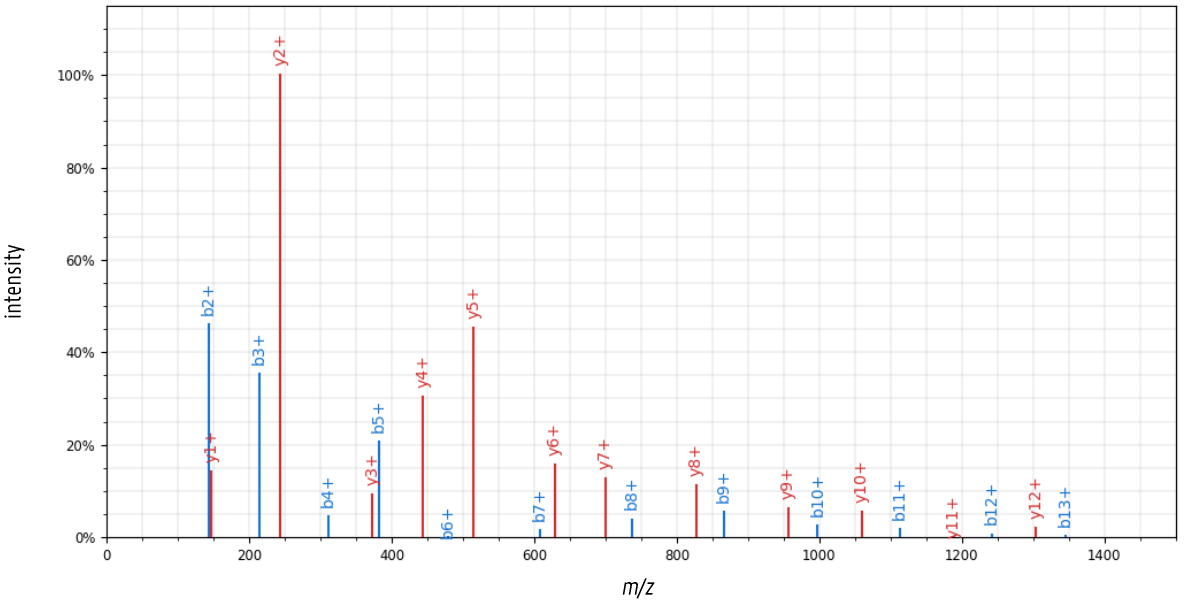


theoretical spectrum

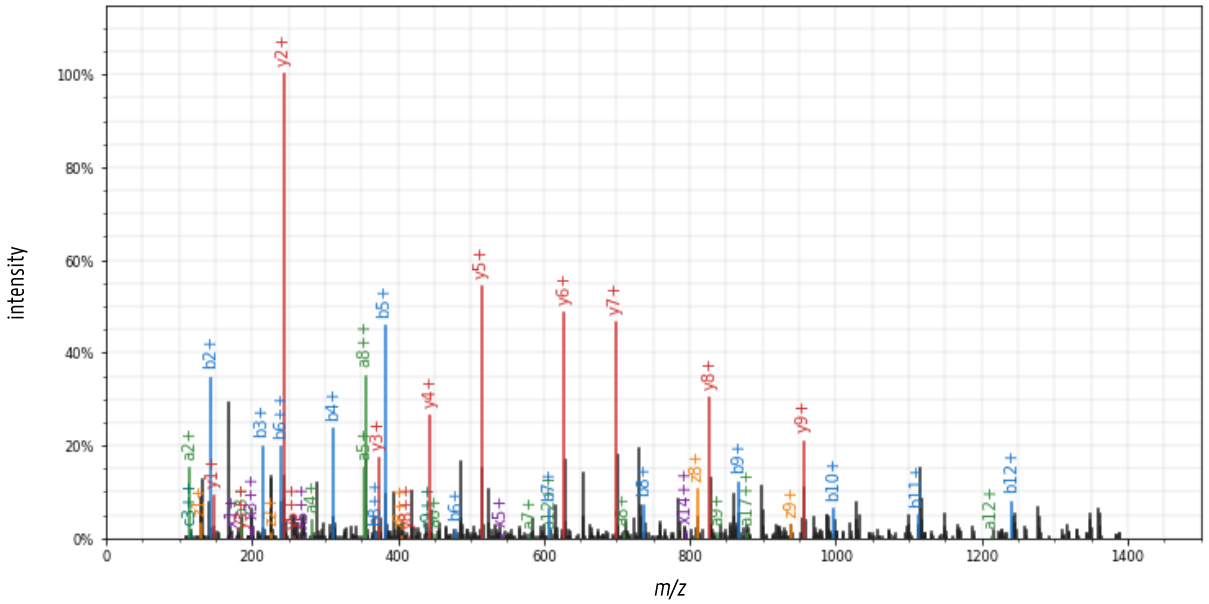


observed spectrum

Machine learning to the rescue!

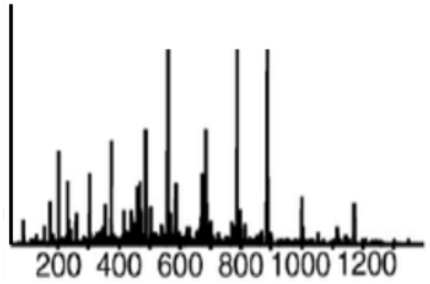


predicted spectrum

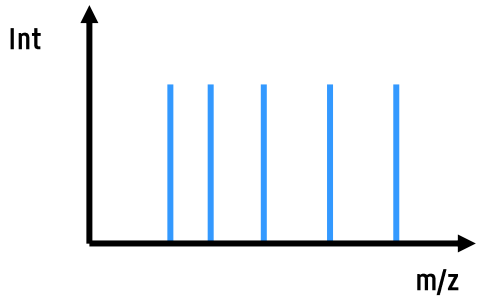


observed spectrum

Machine learning to the rescue!



observed spectrum



theoretical spectrum

Introduction (Dutch)

proteins and proteomics

mass spectrometry

machine learning

Results (English)

MS²PIP: Peptide spectrum prediction for multiple fragmentation methods, instruments, and labeling techniques

Removing the hidden data dependency of DIA with predicted spectral libraries

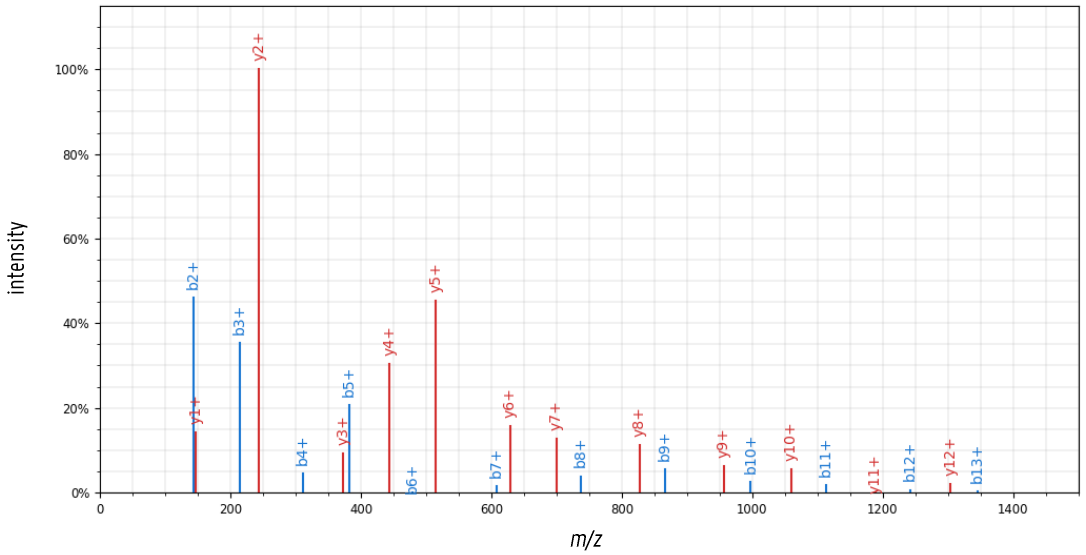
MS²Rescore: Leveraging spectrum predictions to enable novel proteomics workflows

MS²DIP: spectrum prediction for modified peptides

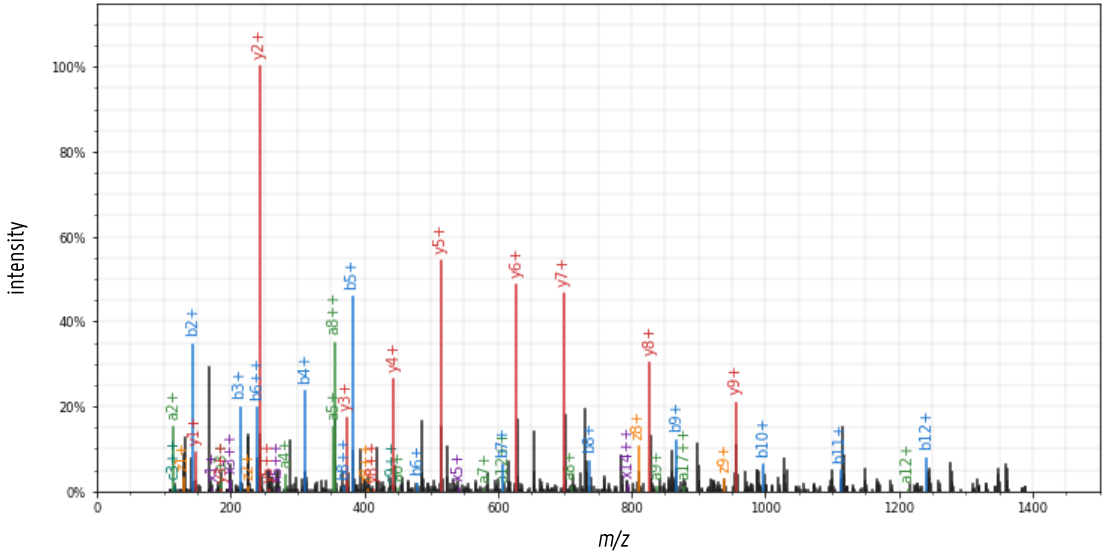
Discussion and future perspectives (English)

MS²PIP: Peptide spectrum prediction for multiple fragmentation methods, instruments, and labeling techniques

MS²PIP can accurately predict peptide fragmentation spectra

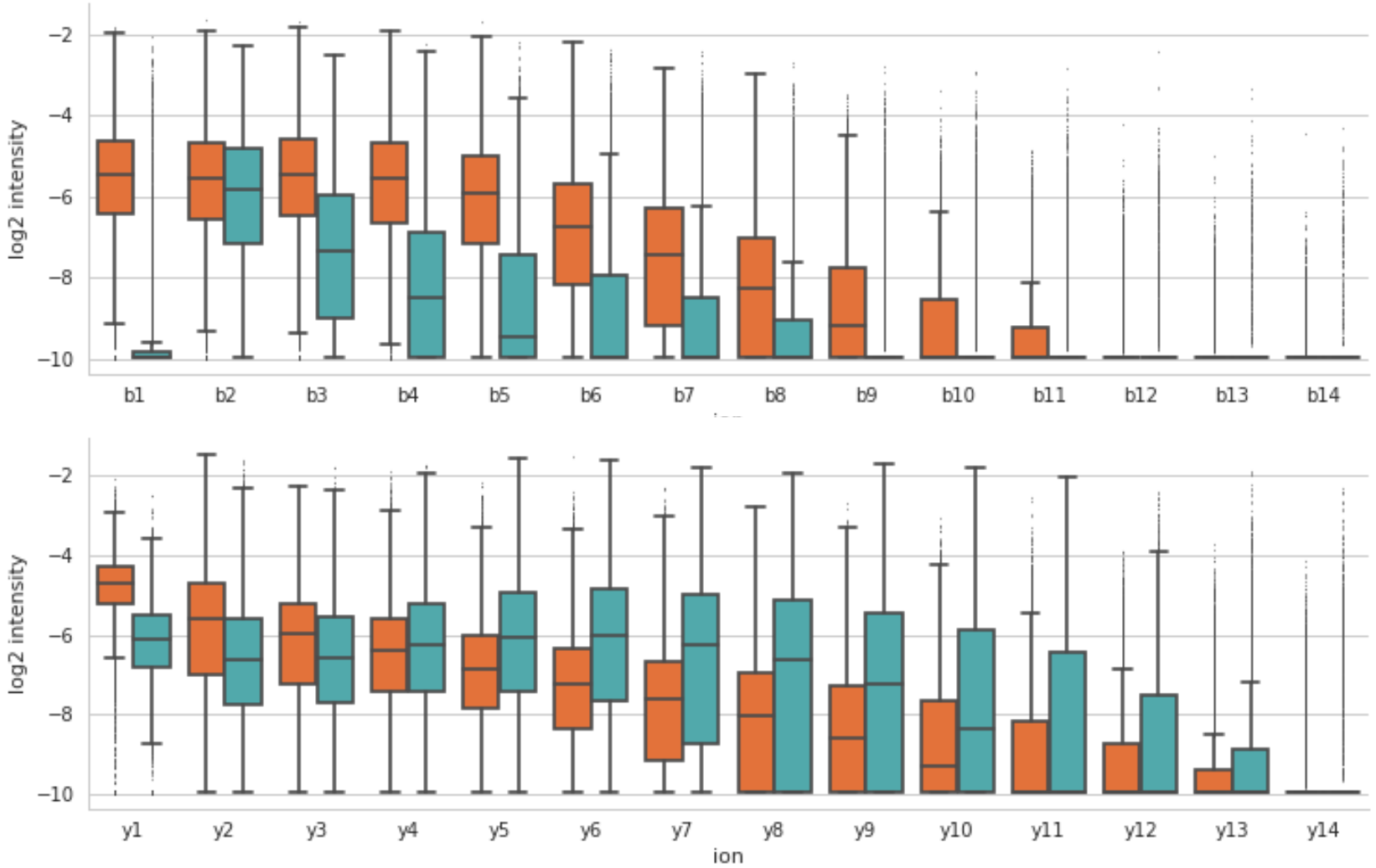


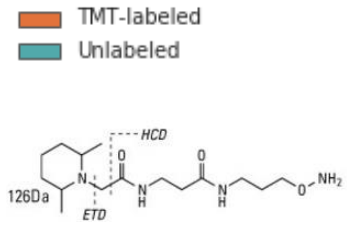
predicted spectrum



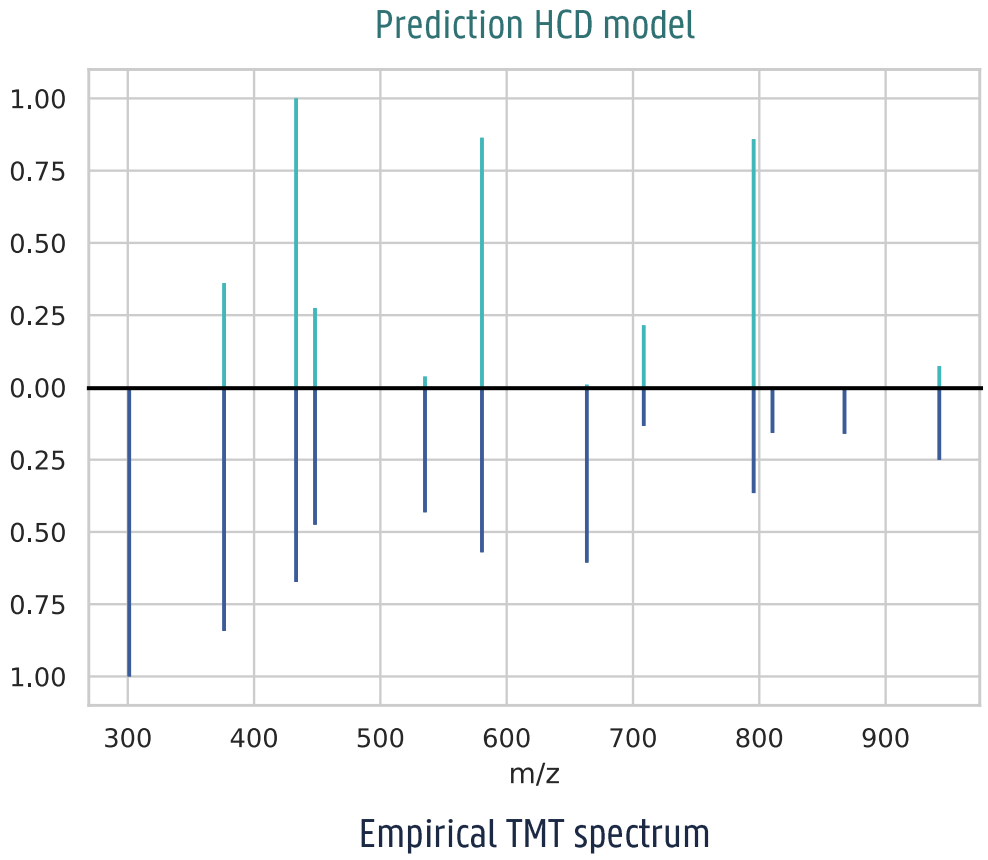
observed spectrum

However, peptide fragmentation can drastically differ between experimental setups



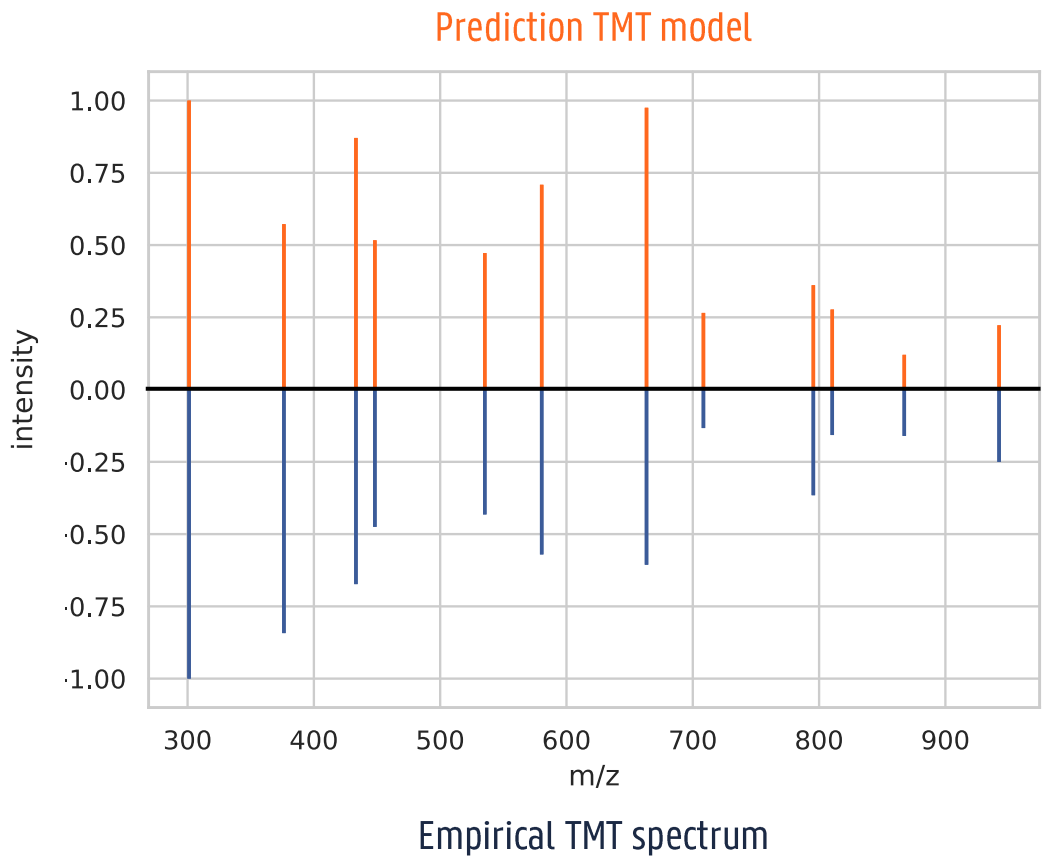


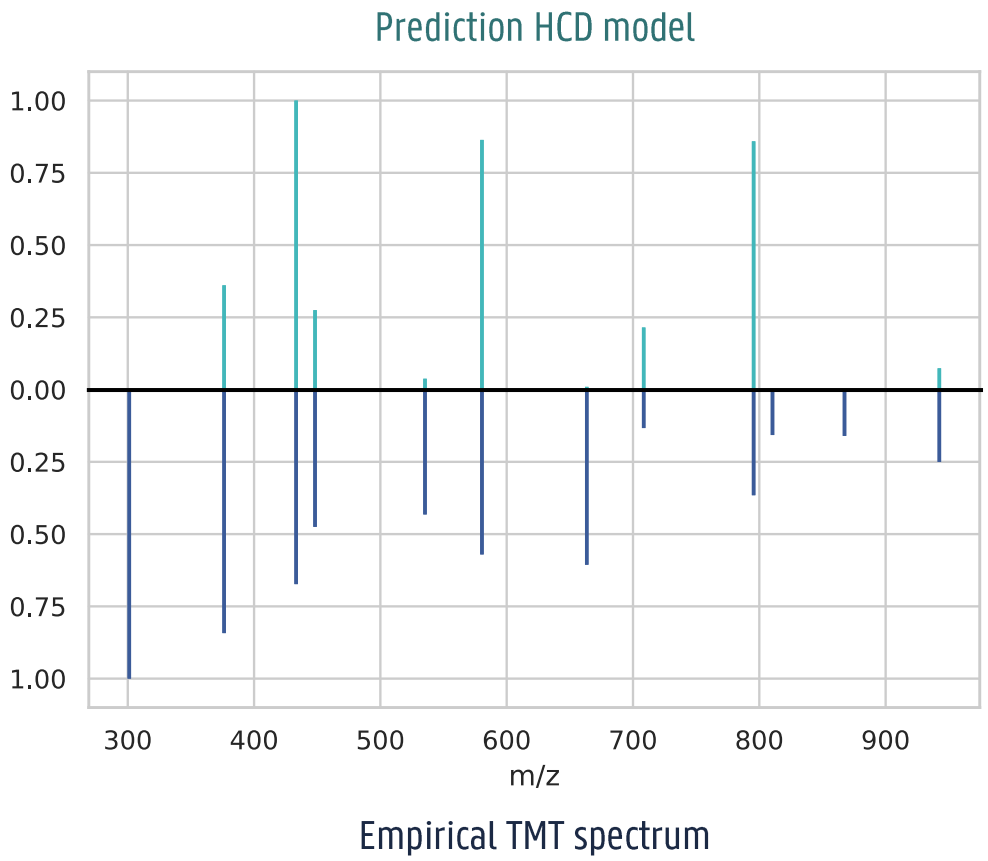
Specialized setups require specialized MS²PIP models



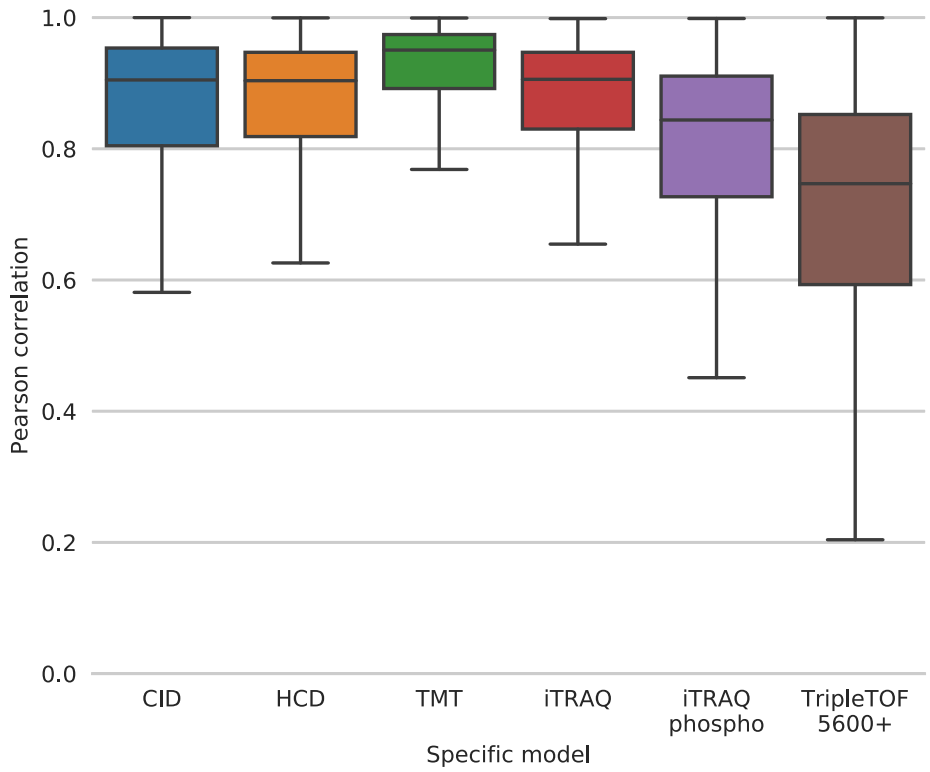
→ Pearson correlation = 0.20

Specialized setups require specialized MS²PIP models

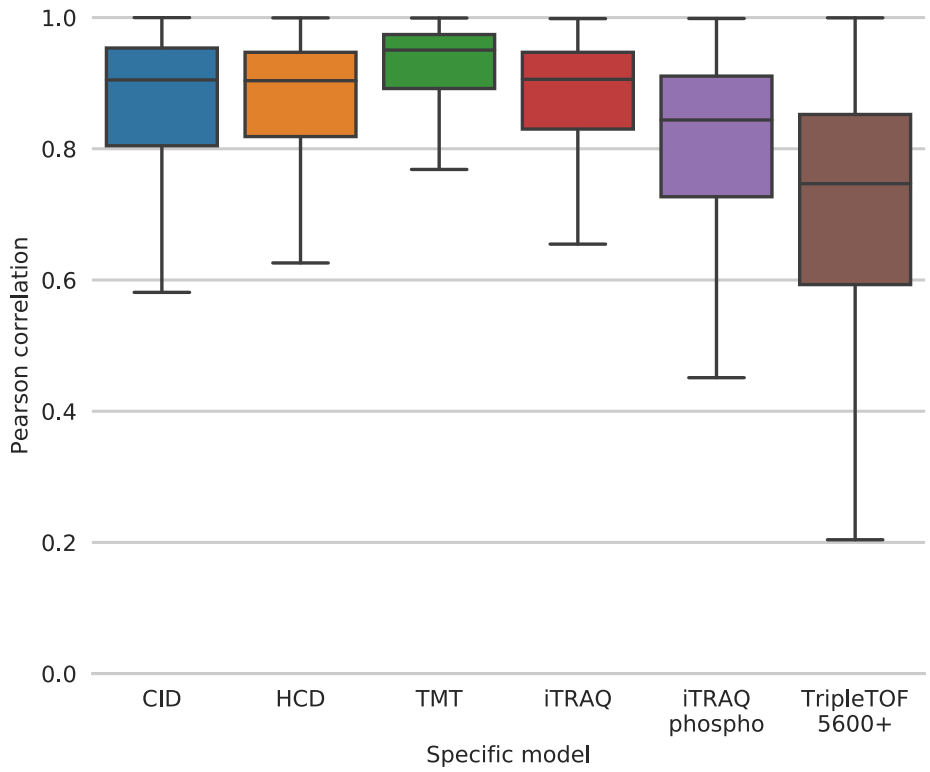


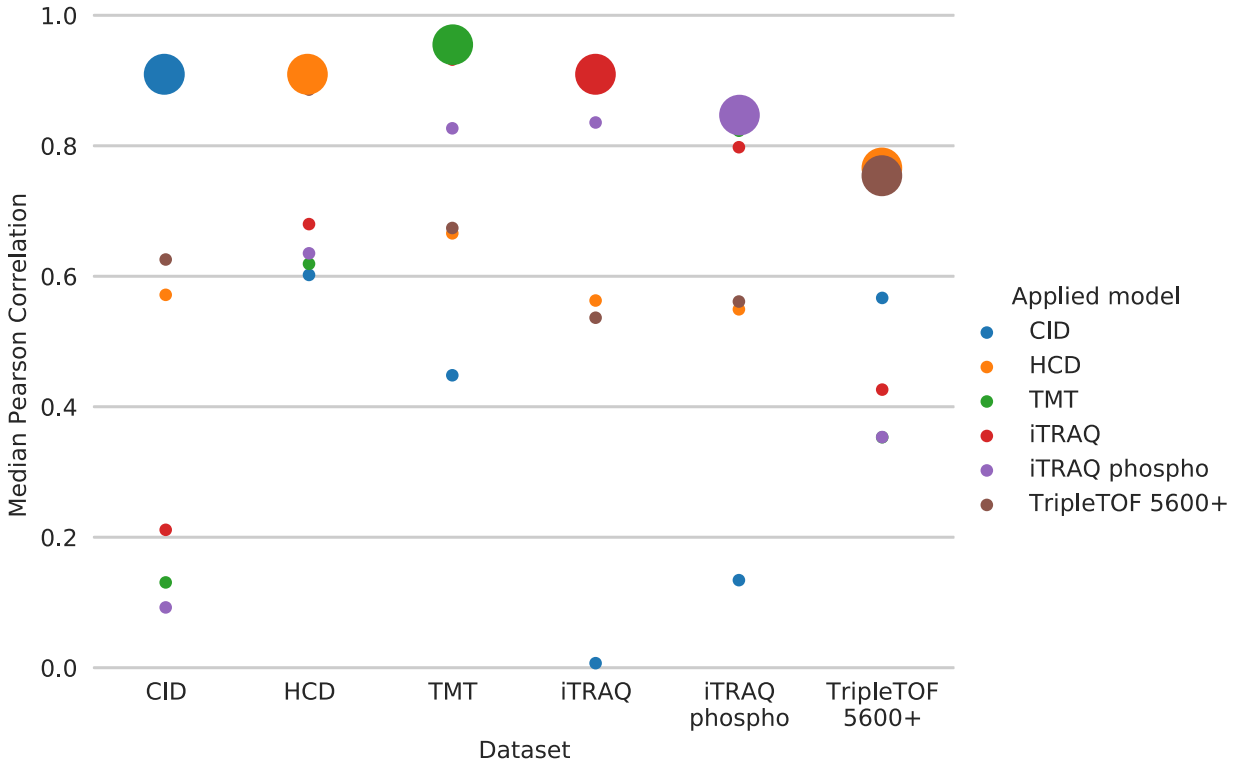


I trained new models for multiple fragmentation methods, instruments and labelling techniques

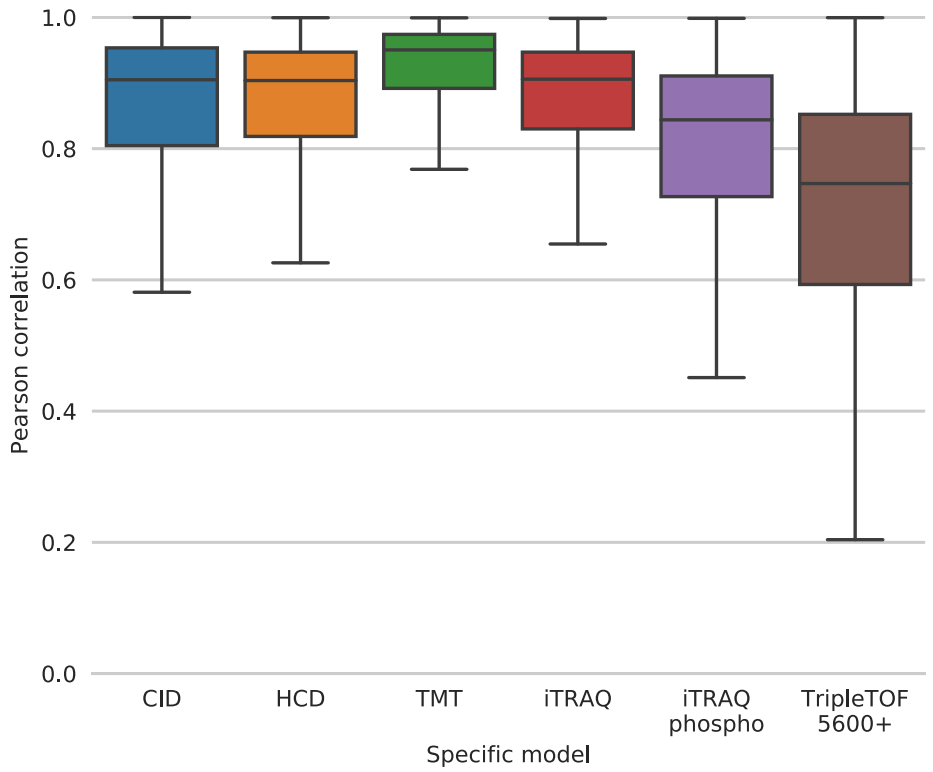


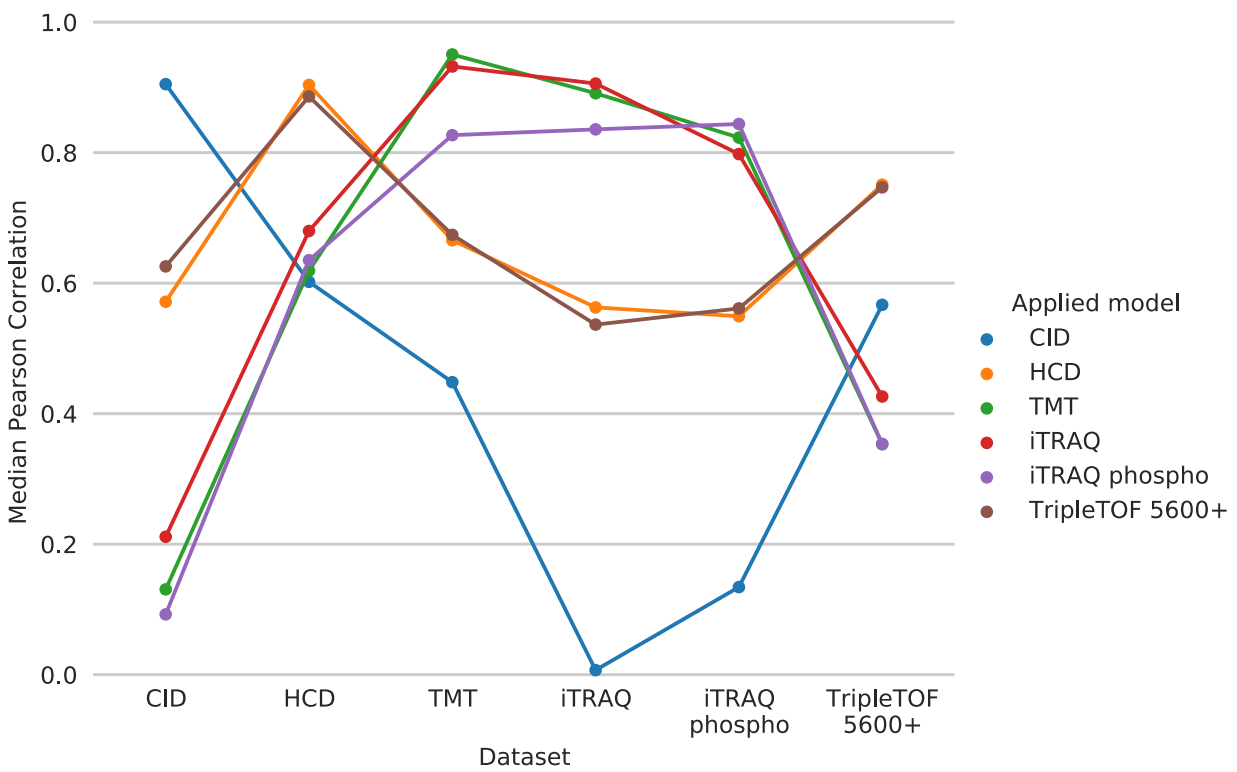
As expected, training data-specific models substantially improves the predictions





Interestingly, similar fragmentation effects across setups are highlighted by the model performances





Removing the hidden data dependency of DIA with predicted spectral libraries

DIA uses wide isolation windows, resulting in complex chimeric spectra

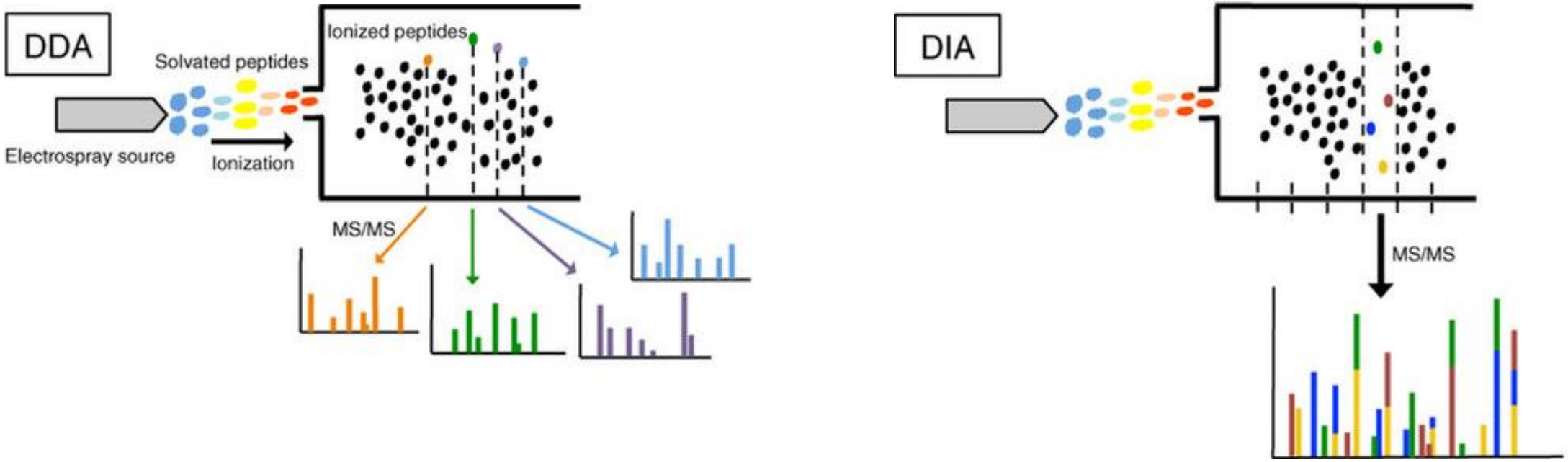
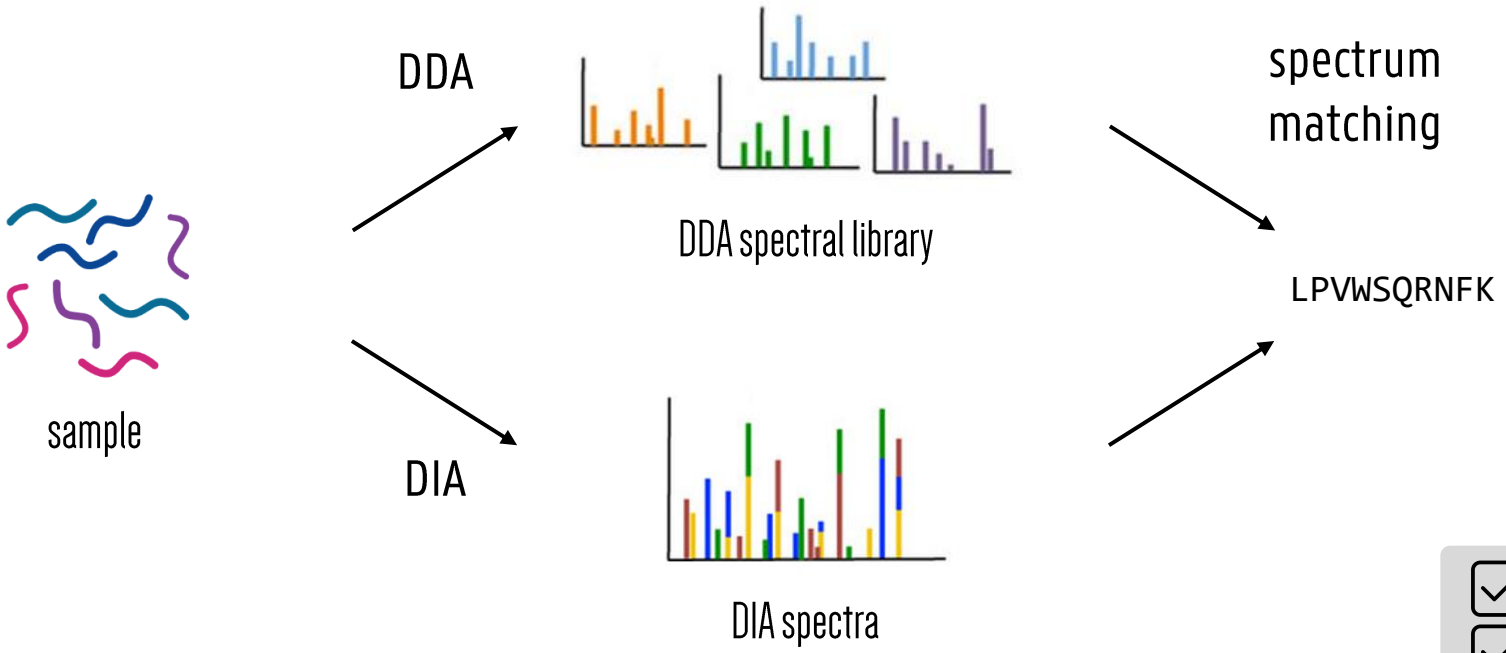


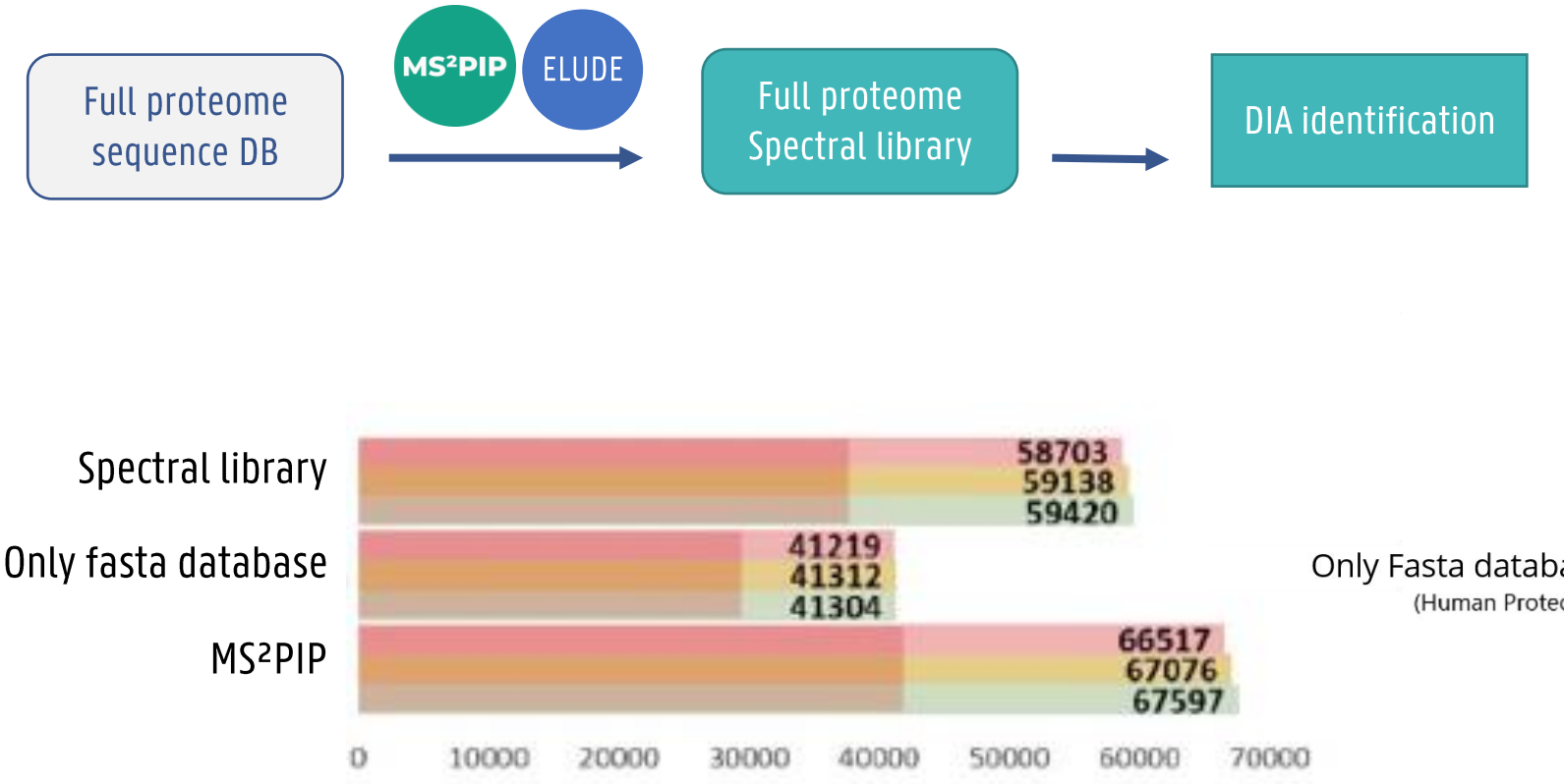
Figure adapted from R Peckner et al., Nat. Methods 2018, doi:10.1038/nmeth.4643

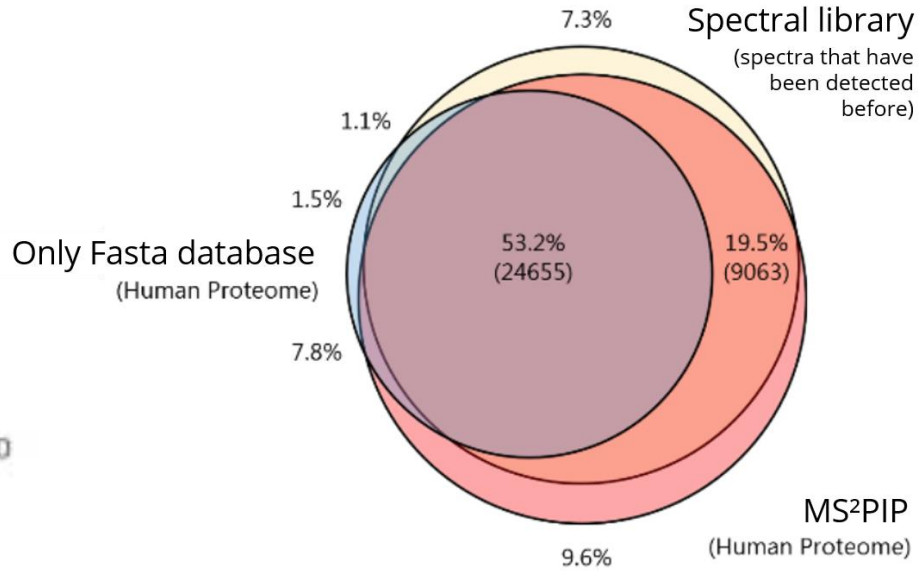
Therefore, DDA spectral libraries are often used for spectrum identification



- ☑ Full LC-MS/MS information
- ☑ Manageable search space
- ✗ Limited to what DDA can identify

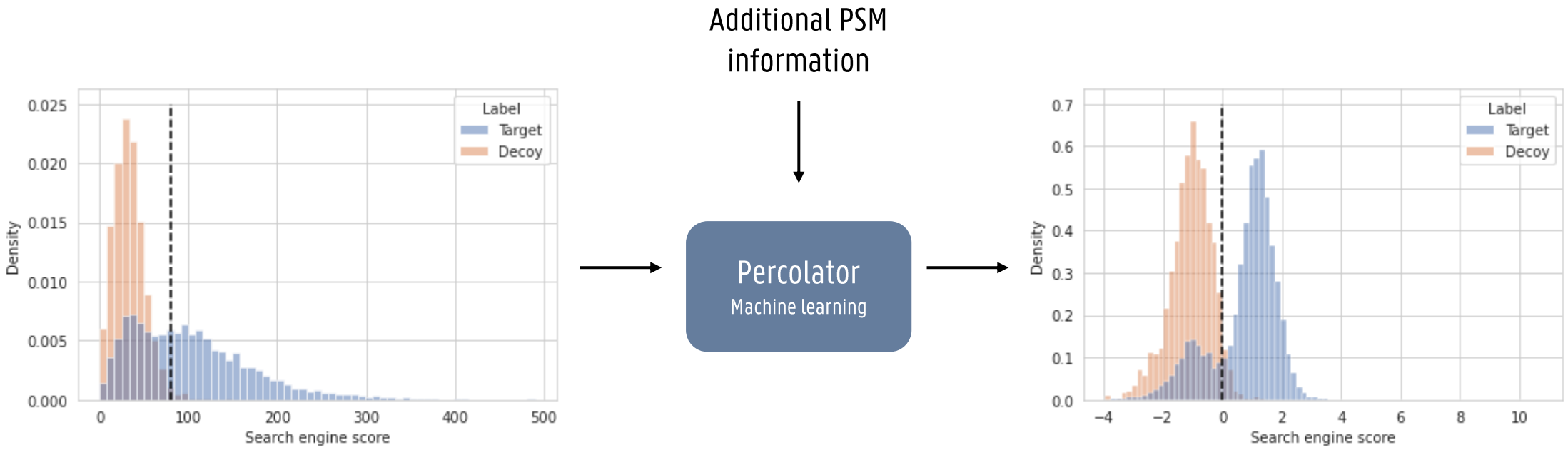
In silico predicted spectral libraries can free DIA from this DDA data dependency



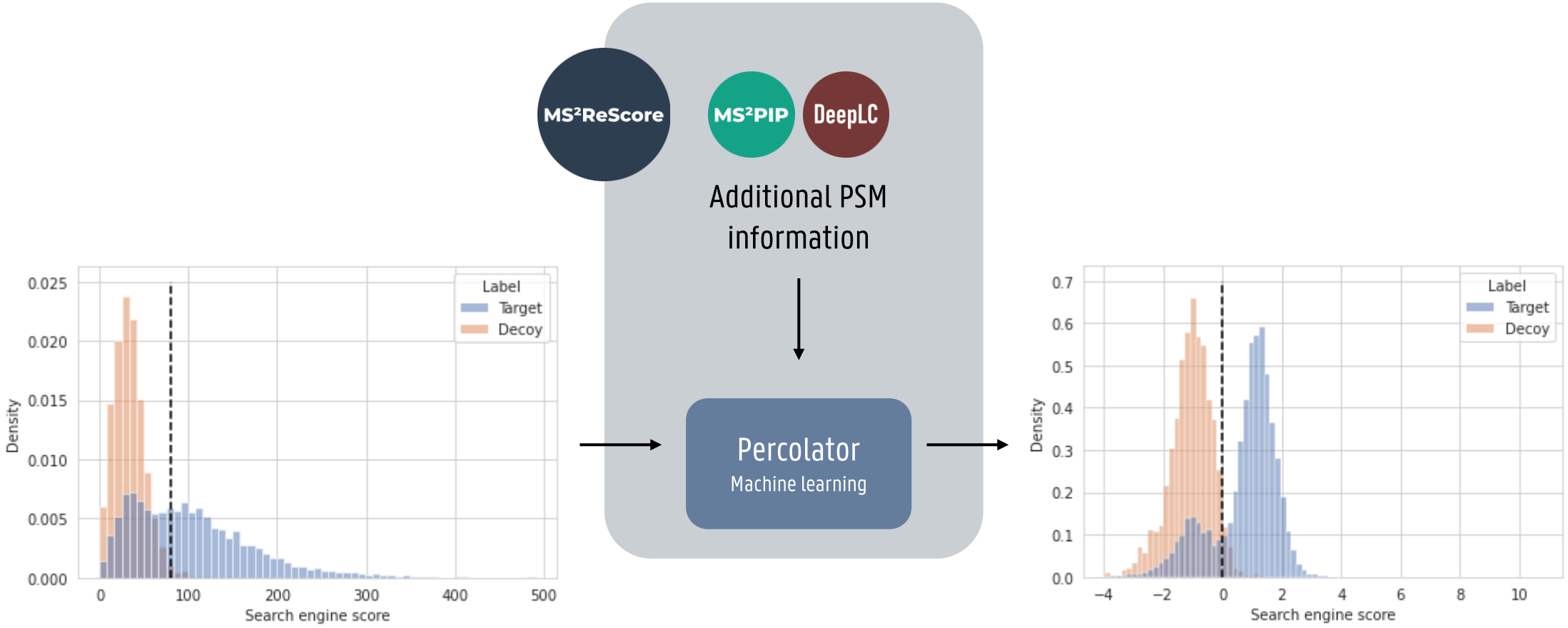


MS²Rescore: Leveraging spectrum predictions to enable novel proteomics workflows

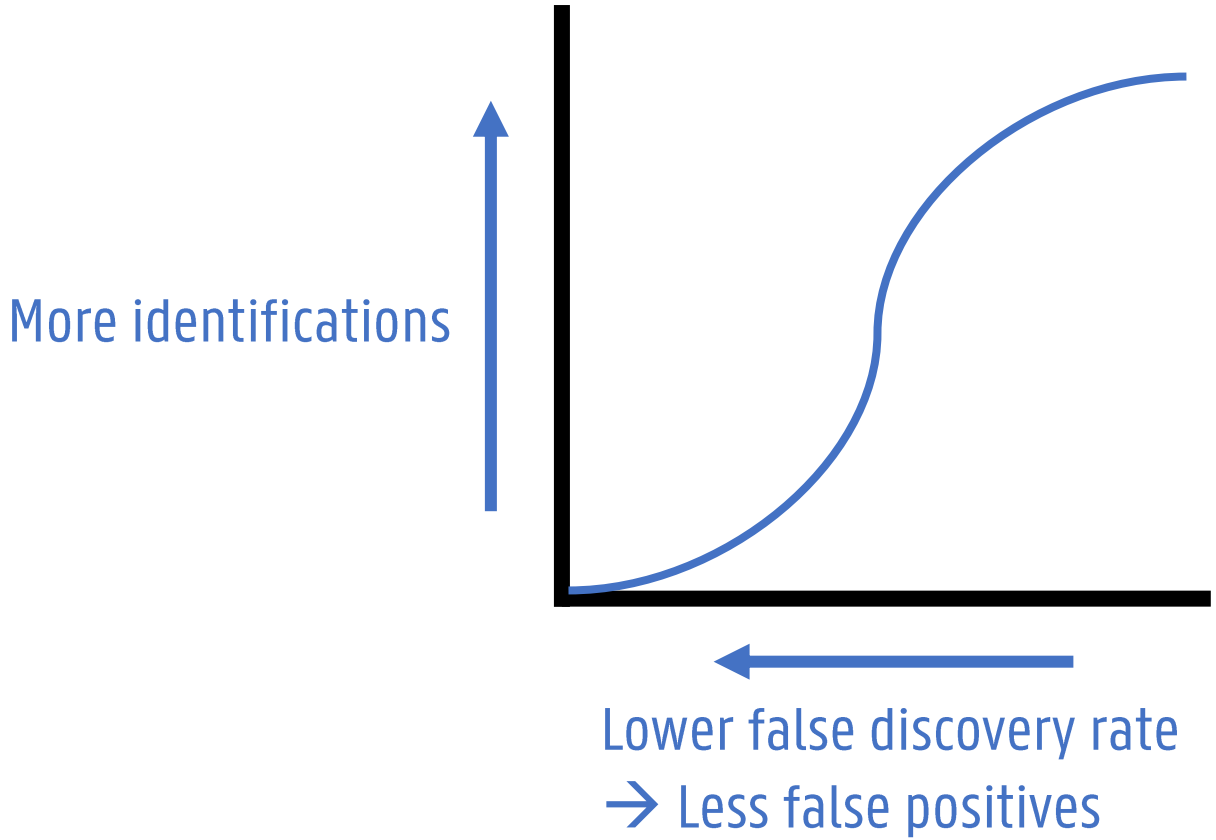
Percolator combines different peptide match information sources and learns how to separate true from false identifications



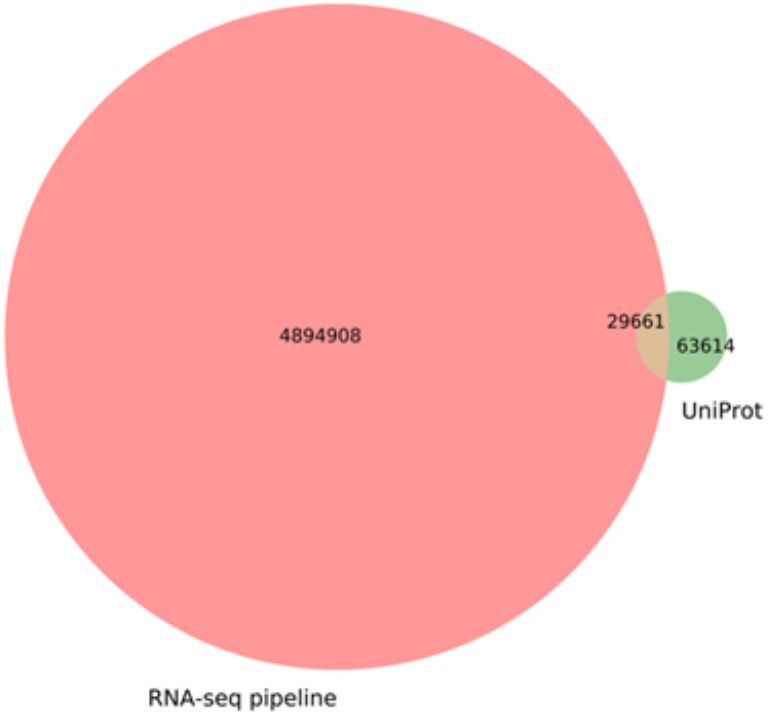
Machine learning-based information can be passed to Percolator for improved rescoring



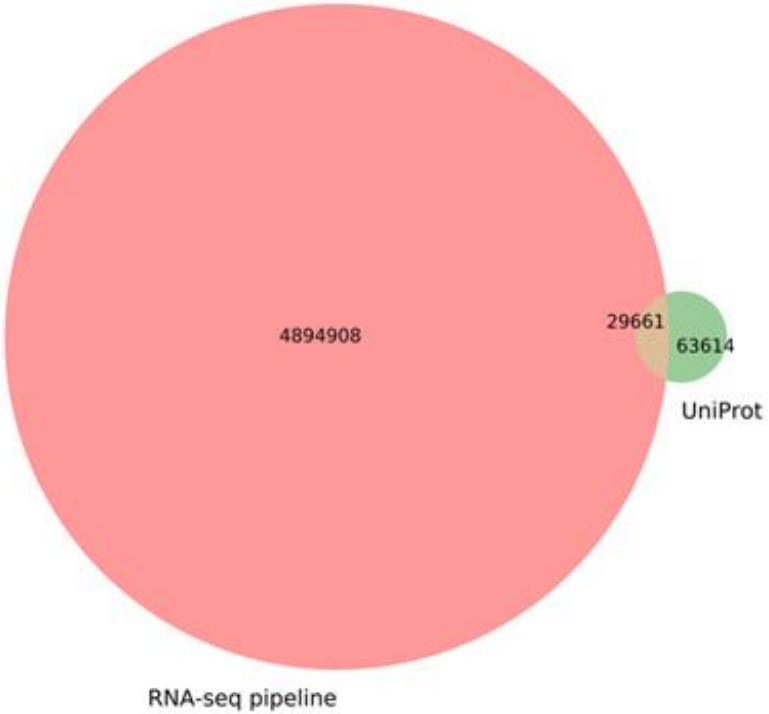
Explaining FDR/identification rate plots

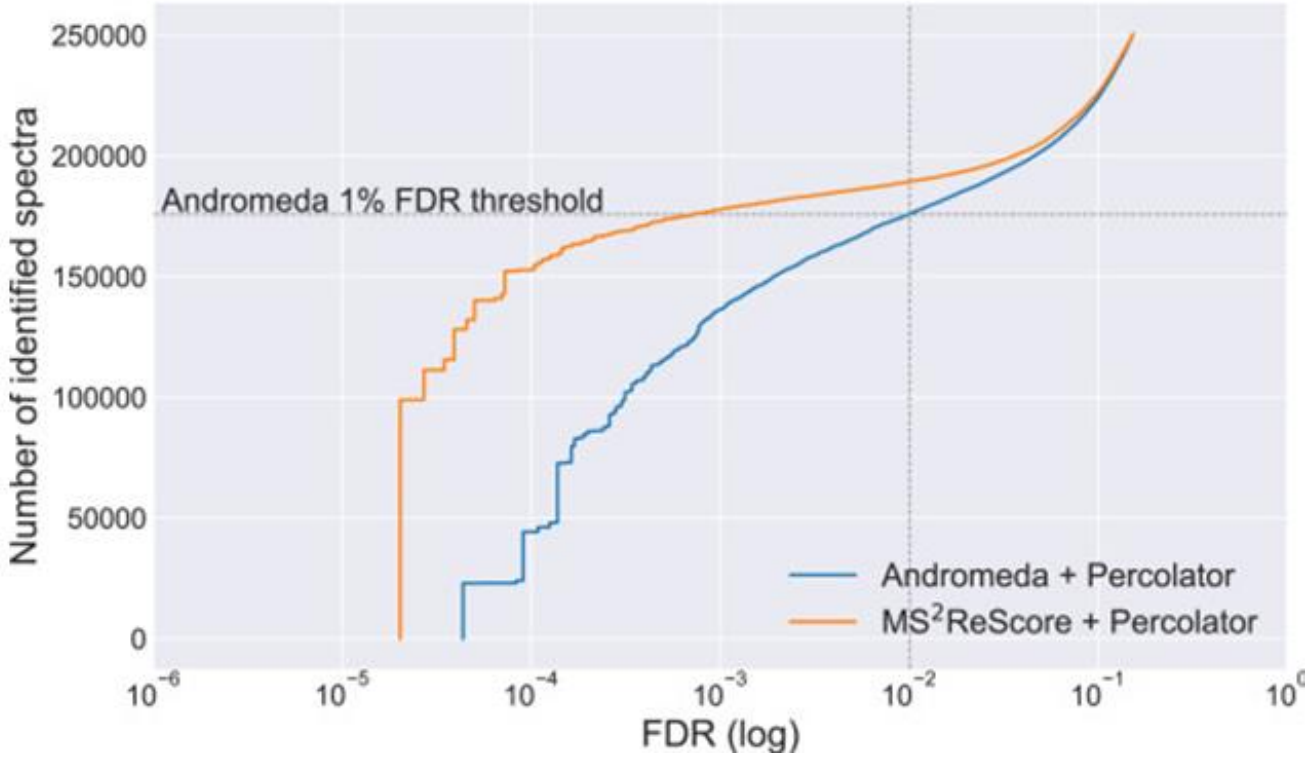


MS²Rescore in proteogenomics: Higher ID rate, at a 10-fold lower FDR threshold



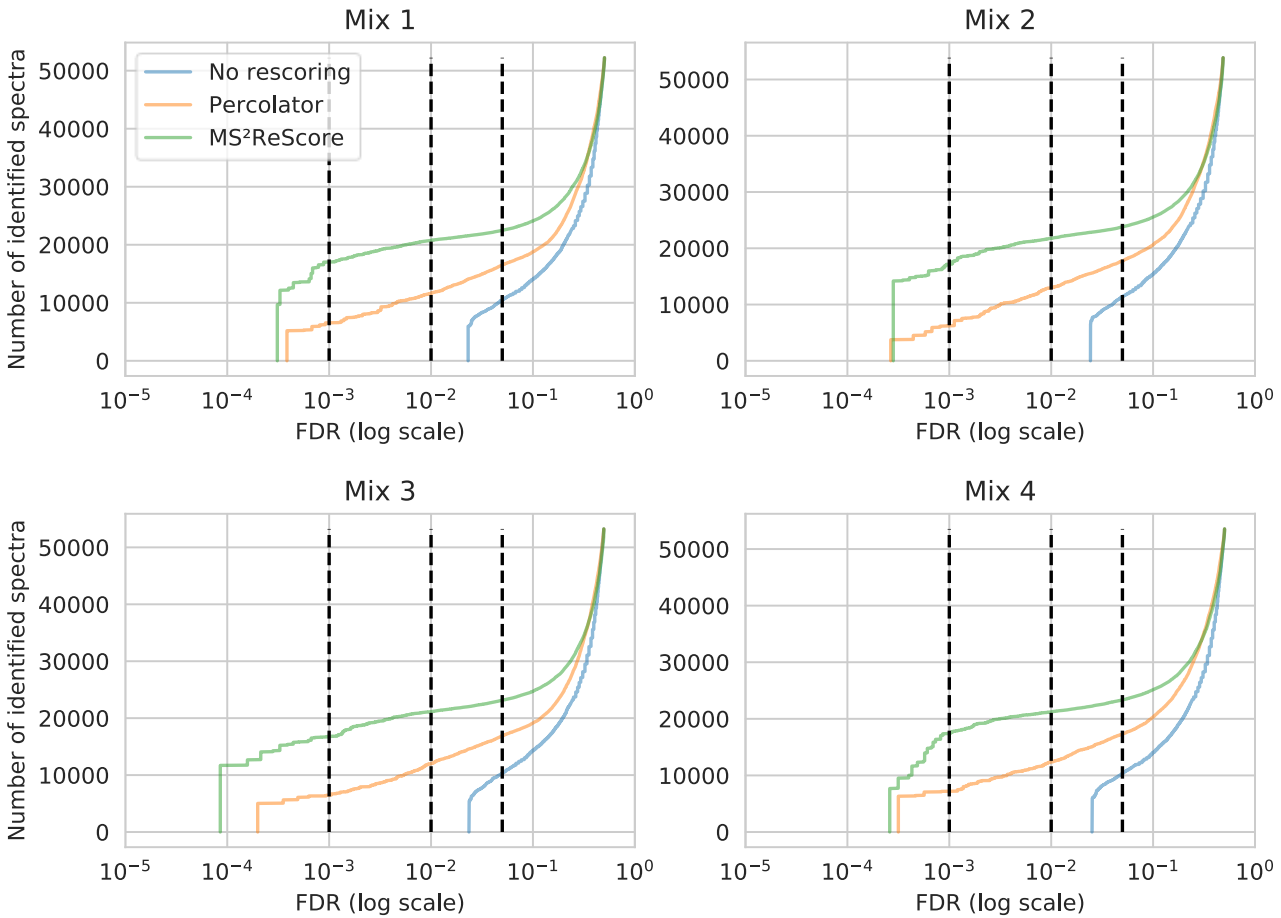
MS²Rescore in proteogenomics: Higher ID rate, at a 10-fold lower FDR threshold





Steven Verbruggen (2021) MCP. doi:10.1016/j.mcpro.2021.100076

MS²Rescore in metaproteomics: From 0 to 20 000 identified spectra

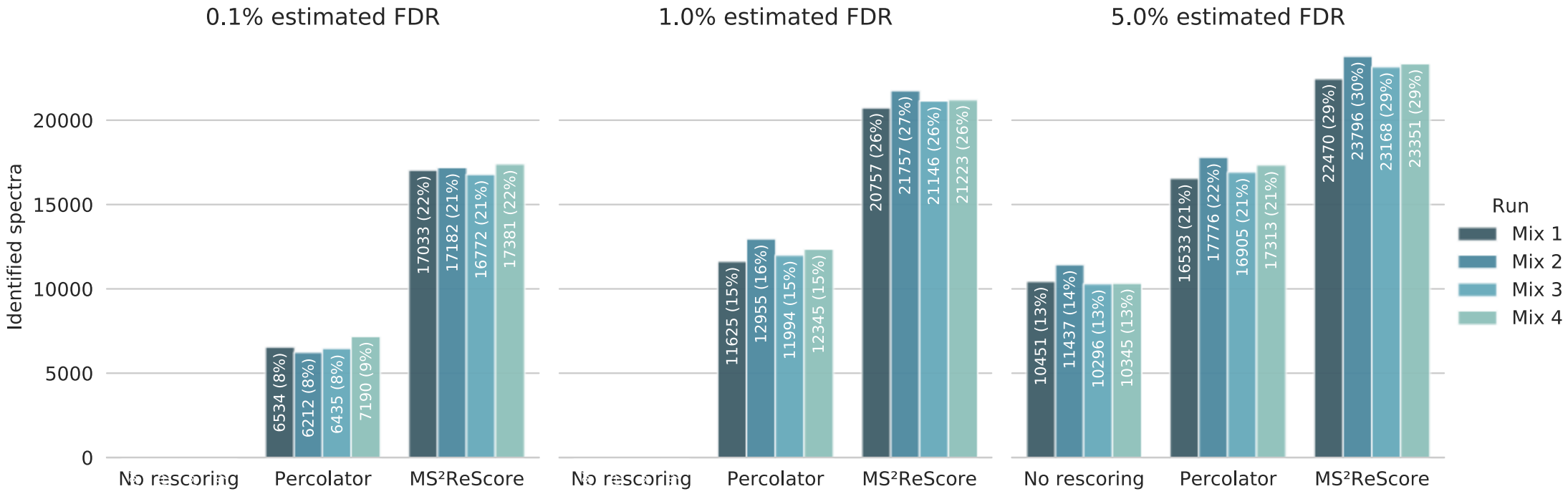


iPRG 2020 study:

- 4 microbial mixes, with unknown sequences
- Extremely large search space
- X!Tandem
- CID ion trap acquisition

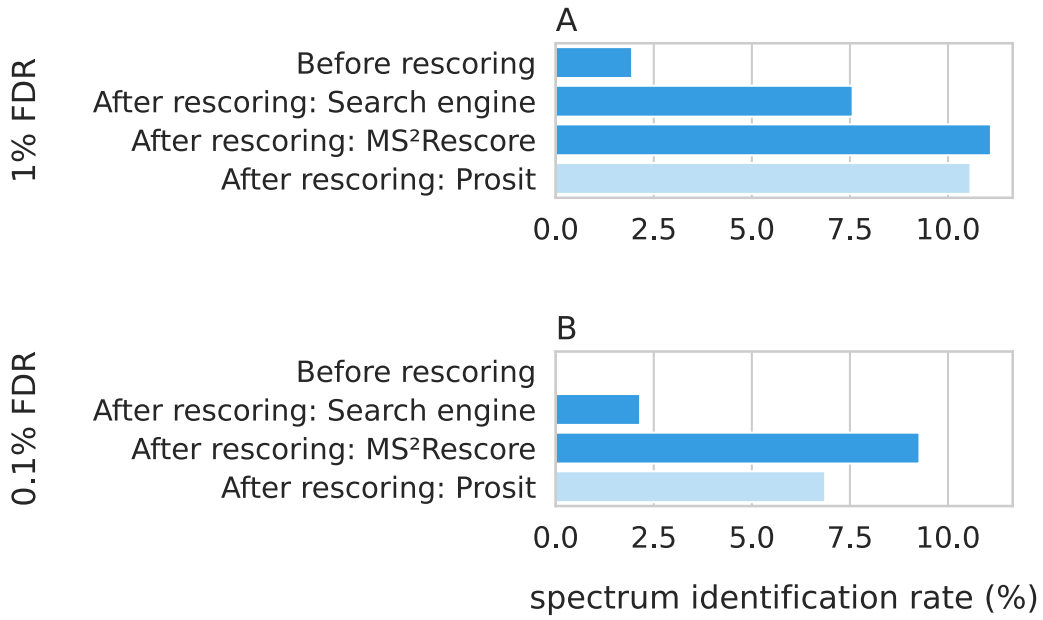
Manuscript in preparation

MS²Rescore in metaproteomics: From 0 to 20 000 identified spectra

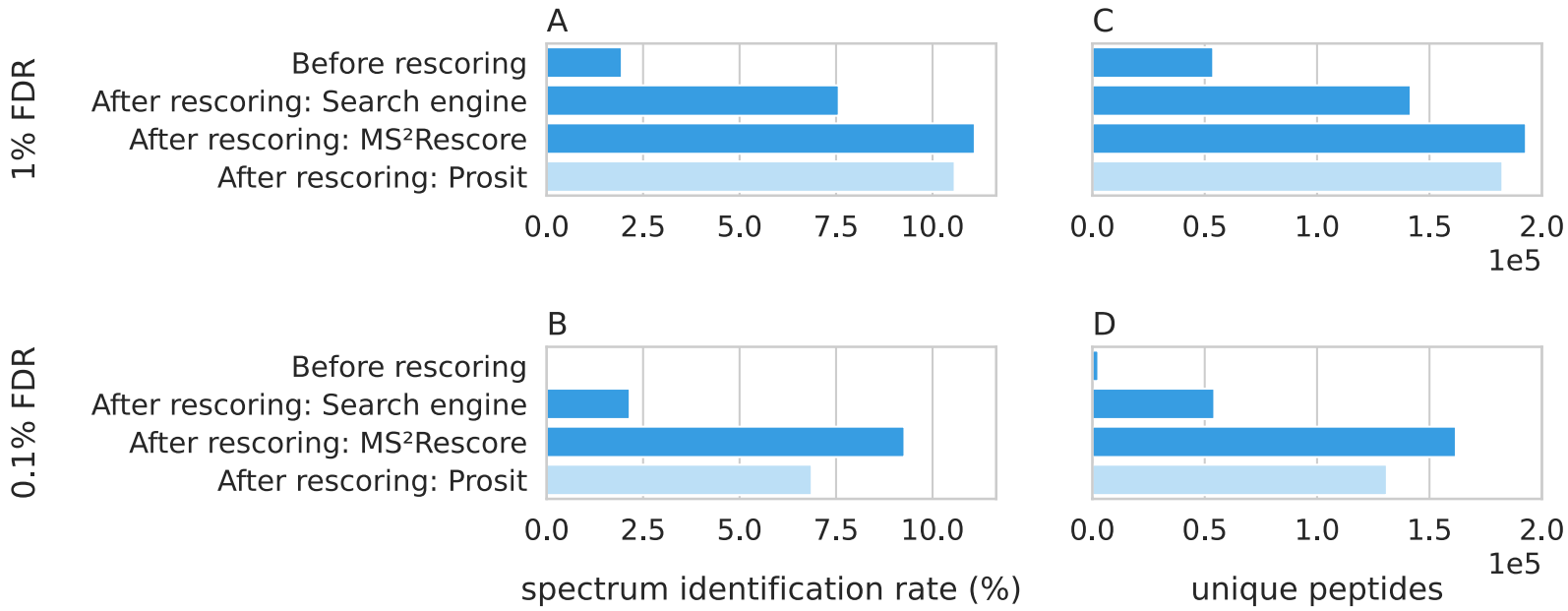


Manuscript in preparation

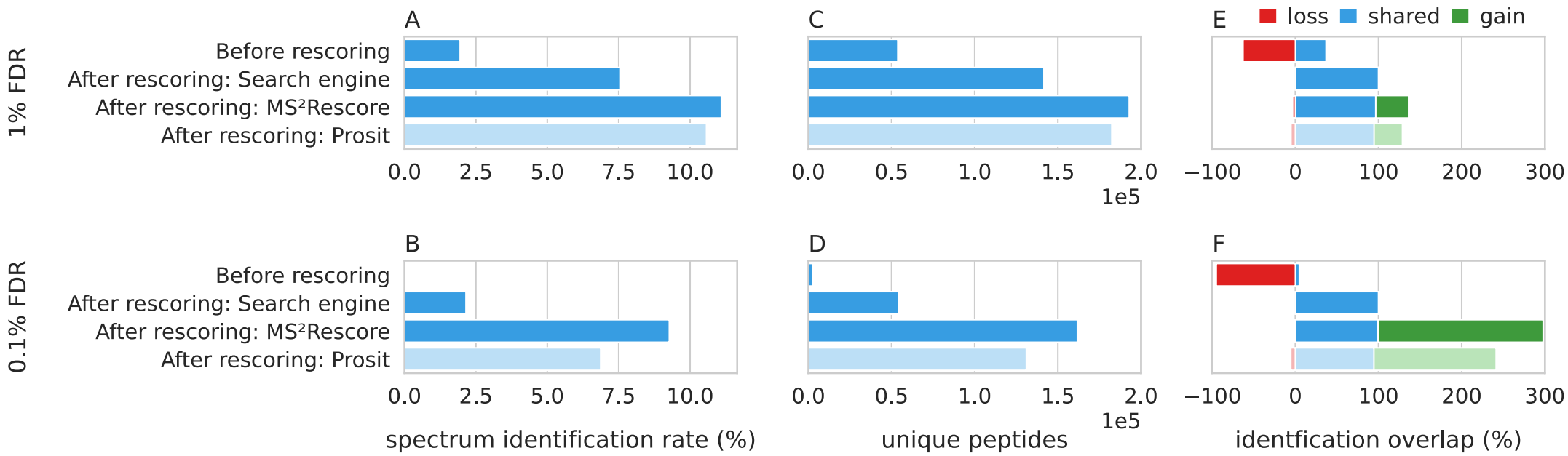
MS²Rescore in immunopeptidomics: +46% spectrum identification rate



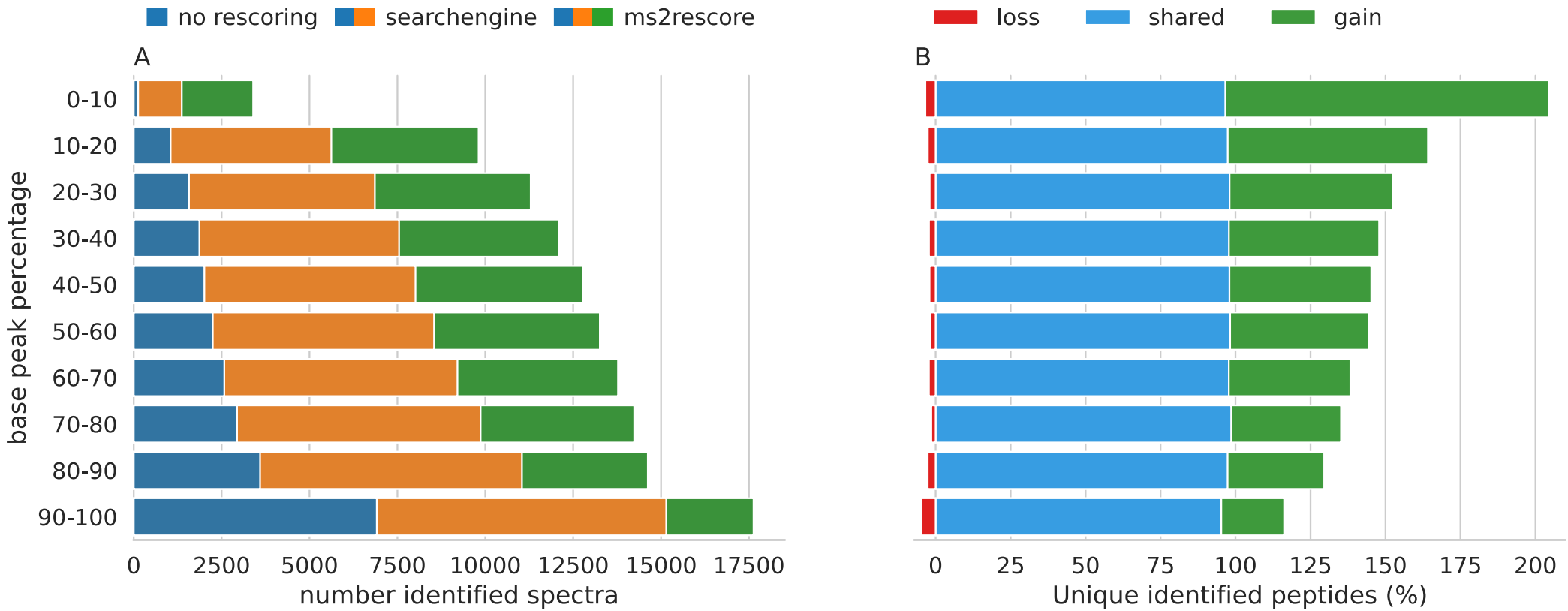
MS²Rescore in immunopeptidomics: +36% unique identified peptides



MS²Rescore in immunopeptidomics: +36% unique identified peptides



MS²Rescore in immunopeptidomics: +46% identification rate



MS²DIP:
spectrum prediction for modified peptides

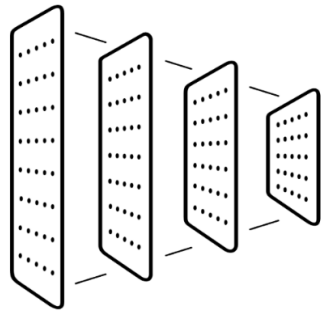
Deep learning allows a generalization across unmodified and modified peptides

large-scale reprocessing of public data on PRIDE Archive

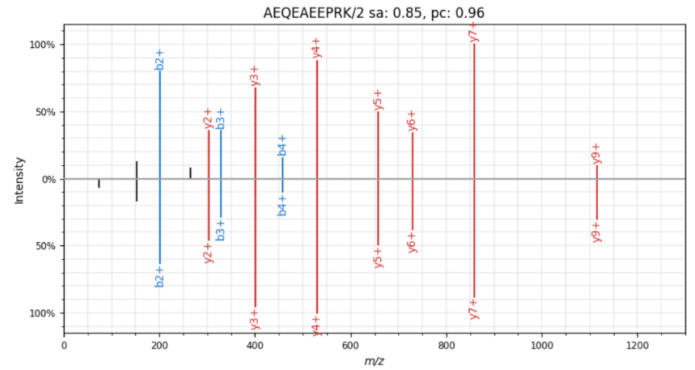
ionbot

open modification search with novel search engine ionbot

spectral library of 11 million unique PSMs

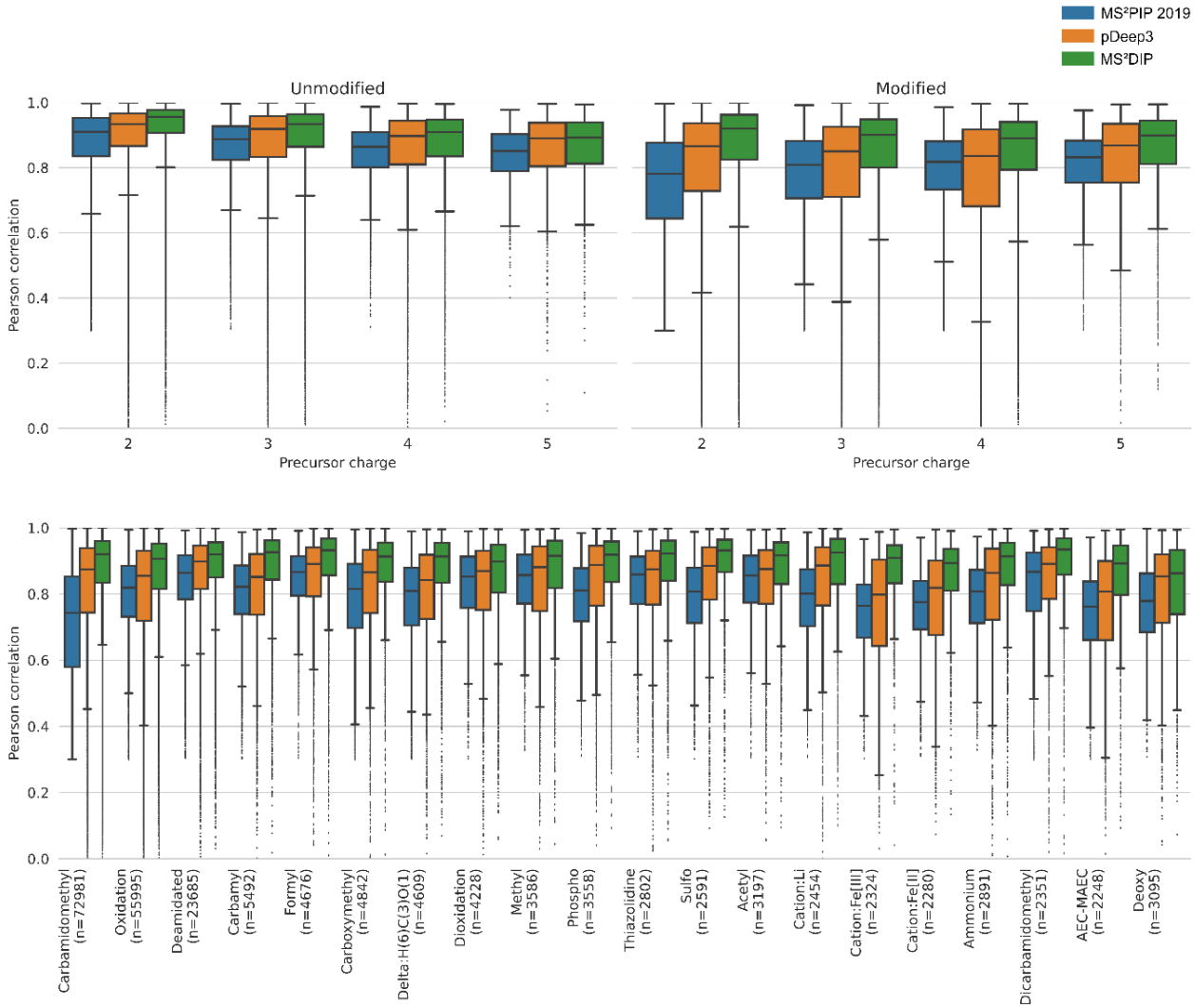


Convolutional neural network



Spectrum predictions for (un)modified peptides

Preliminary results already outperform MS²PIP and out-of-the-box pDeep3



Introduction (Dutch)

proteins and proteomics

mass spectrometry

machine learning

Results (English)

MS²PIP: Peptide spectrum prediction for multiple fragmentation methods, instruments, and labeling techniques

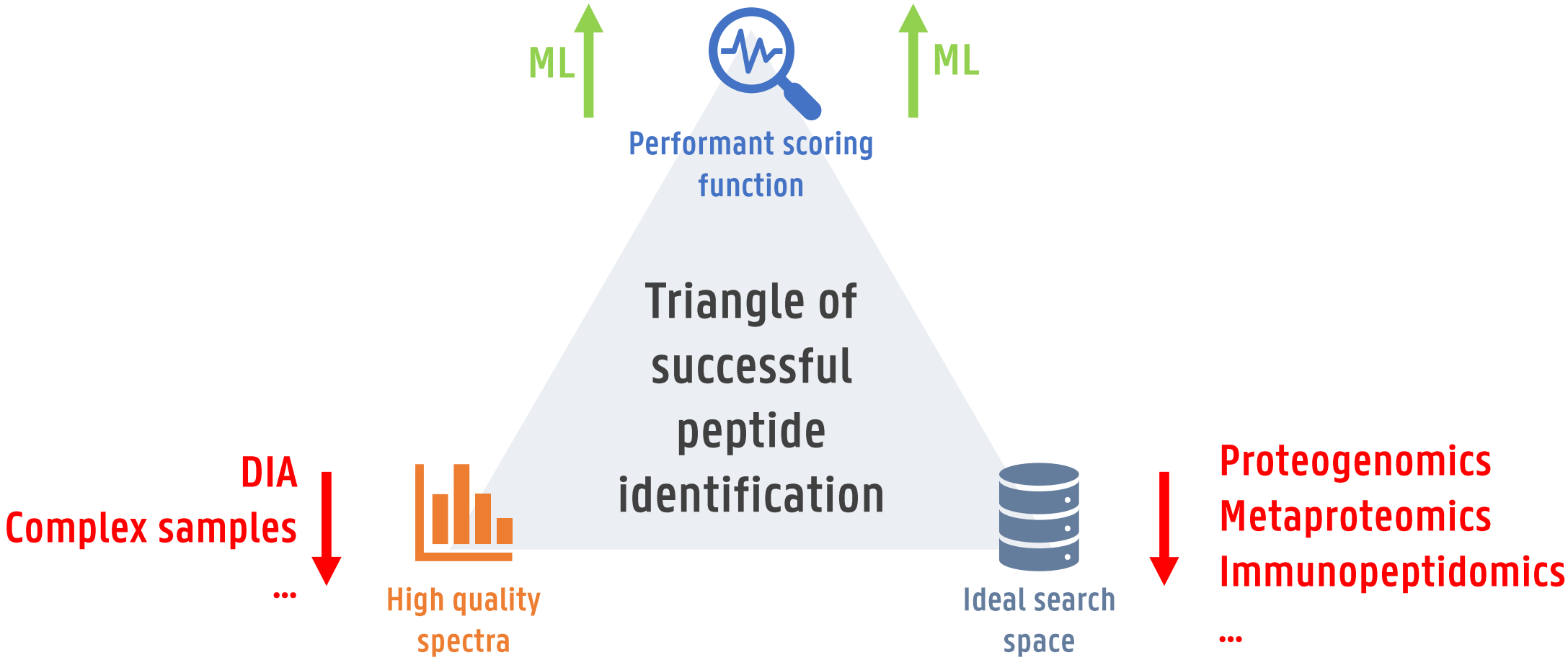
Removing the hidden data dependency of DIA with predicted spectral libraries

MS²Rescore: Leveraging spectrum predictions to enable novel proteomics workflows

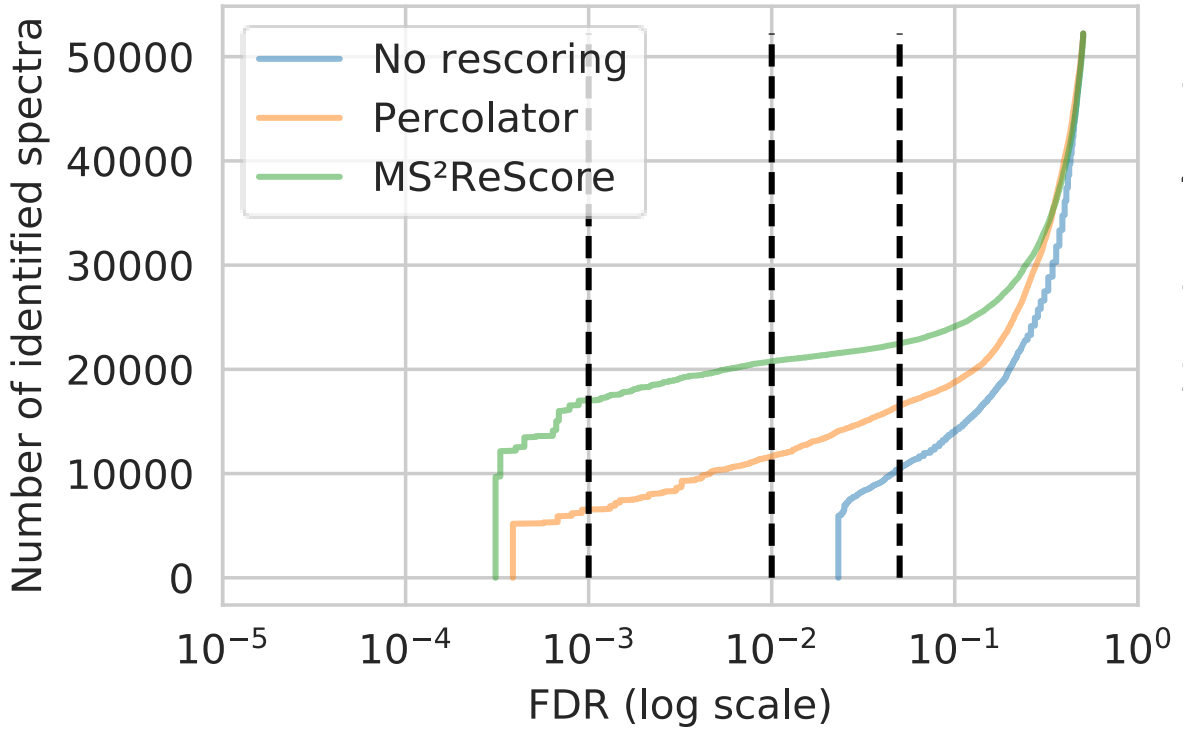
MS²DIP: spectrum prediction for modified peptides

Discussion and future perspectives (English)

In challenging experimental workflows, machine learning can rescue peptide identification rates



More sensitive scoring methods also allow for more specificity: 0.1% FDR to replace 1% FDR?



My “ML in proteomics wish list” for the (near) future

Better metadata in public data

Use of standardized file formats

More modularity and interoperability in bioinformatics software

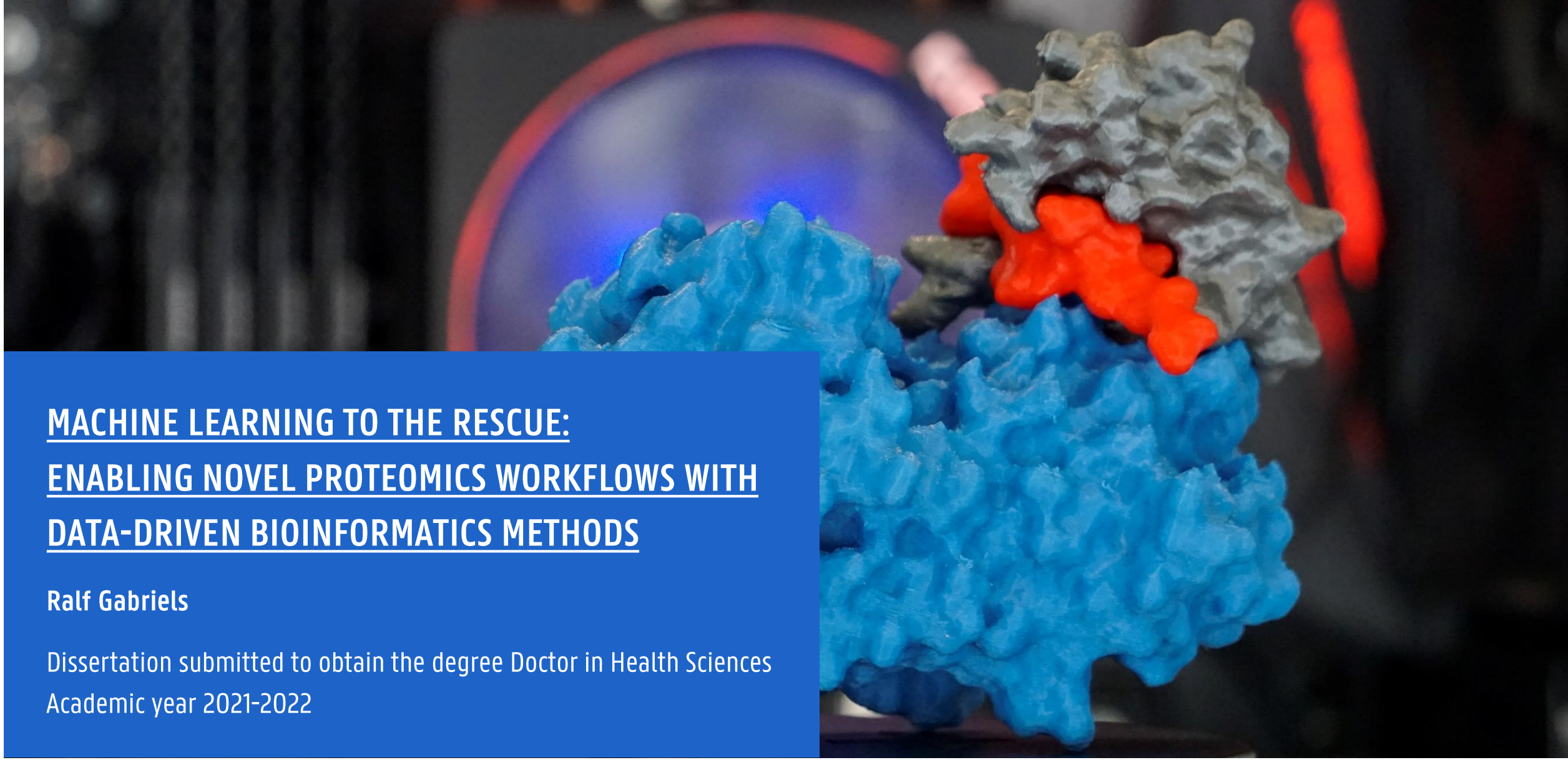
Modularity and sharing of deep learning networks

High quality, community-curated training data

The future of MS-based proteomics lies in the discovery of the unexpected

requiring more sensitive methods

where machine learning will take a central role



MACHINE LEARNING TO THE RESCUE: ENABLING NOVEL PROTEOMICS WORKFLOWS WITH DATA-DRIVEN BIOINFORMATICS METHODS

Ralf Gabriels

Dissertation submitted to obtain the degree Doctor in Health Sciences

Academic year 2021-2022

GHENT UNIVERSITY

FACULTY OF MEDICINE AND HEALTH SCIENCES

VIB-UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

fwo

CompOmics