SCIENCE MEETS LIFE

PEPTIDE AND PROTEIN IDENTIFICATION

lennart martens

lennart.martens@vib-ugent.be
@compomics
computational omics and systems biology group
Ghent University and VIB, Ghent, Belgium

VIB

GHENT UNIVERSITY

Comp omics

MS/MS spectra and identification

Database search algorithms in three phases

Sequencial search algorithms

Decoys and false discovery rate calculation

The future: machine learning

Protein inference: bad, ugly, and not so good

MS/MS spectra and identification

Database search algorithms in three phases

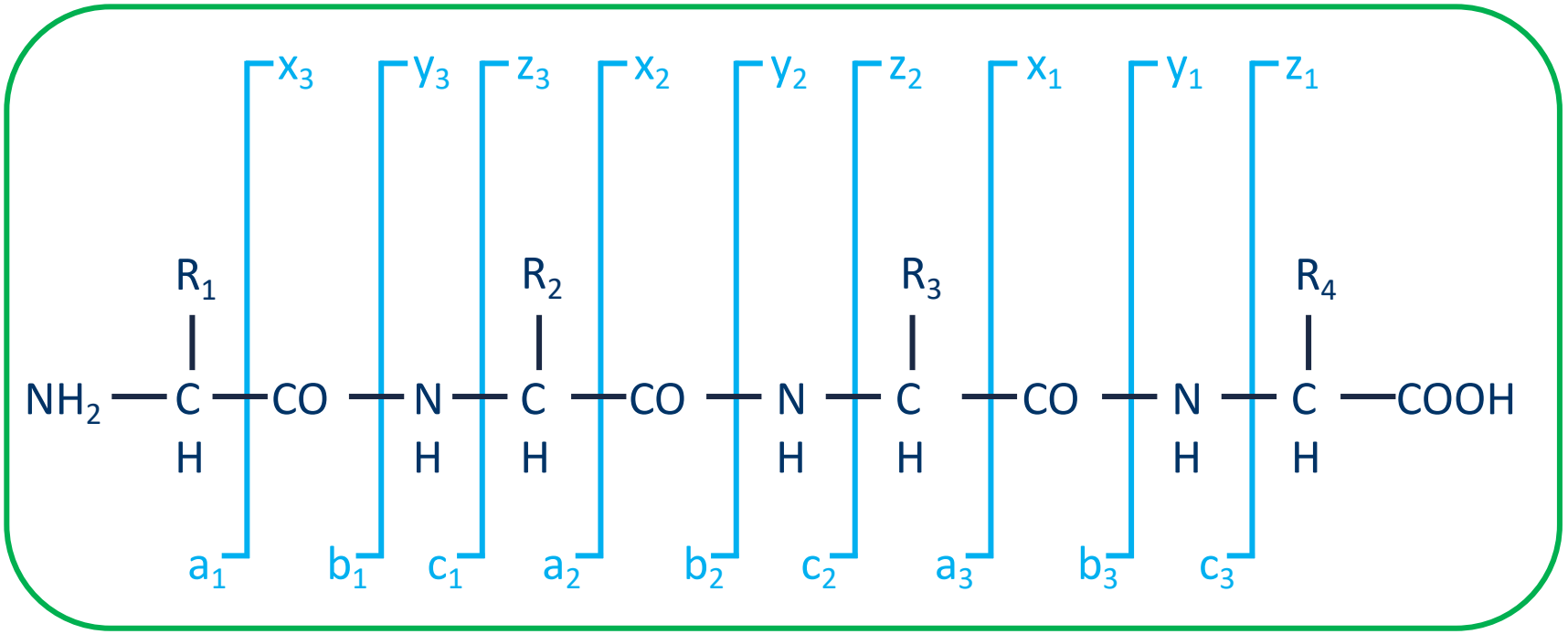
Sequencial search algorithms

Decoys and false discovery rate calculation

The future: machine learning

Protein inference: bad, ugly, and not so good

Peptides subjected to fragmentation analysis can yield several types of fragment ions

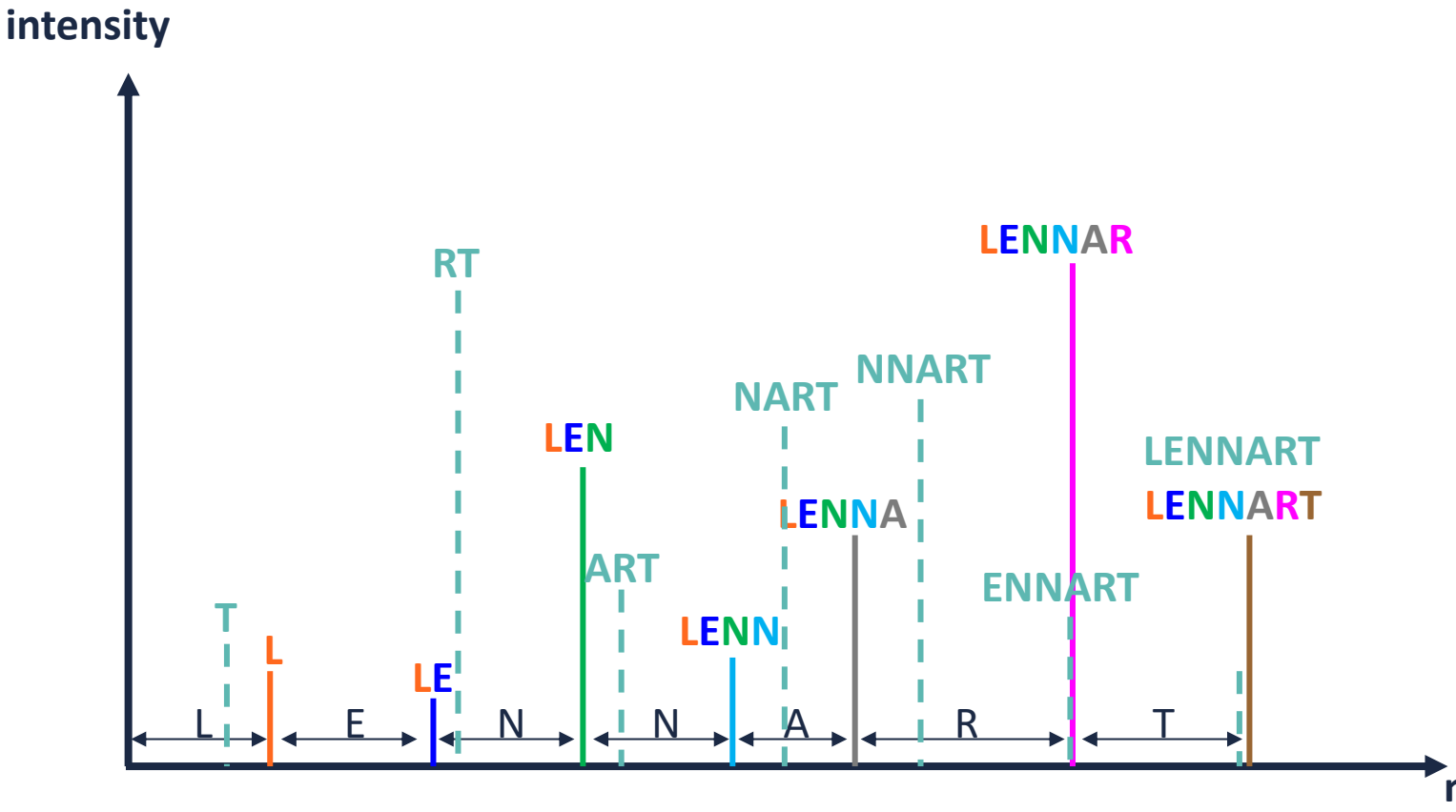


There are several other ion types that can be annotated, as well as 'internal fragments'. The latter are fragments that no longer contain an intact terminus. These are harder to use for 'ladder sequencing', but can still be interpreted.

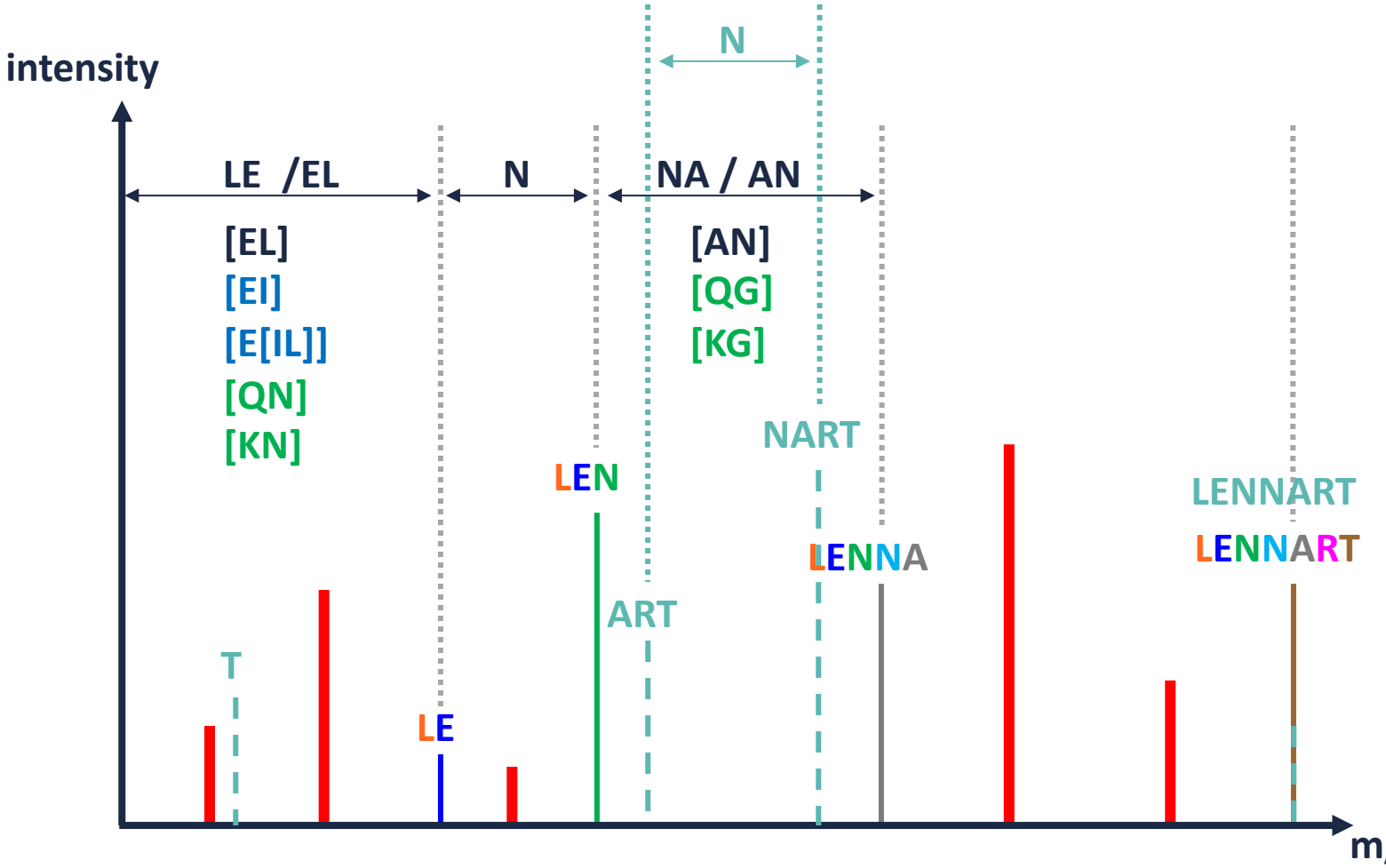
This nomenclature was coined by **Roepstorff and Fohlmann** *(Biomed. Mass Spec.*, 1984) and **Klaus Biemann** *(Biomed. Environ. Mass Spec.*, 1988) and is commonly referred to as 'Biemann nomenclature'. Note the link with the Roman alphabet.

In an ideal world, the peptide sequence will produce directly interpretable ion ladders

L E N N A R T



Real spectra usually look quite a bit worse, which introduces ambiguity in interpretation



MS/MS spectra and identification

Database search algorithms in three phases

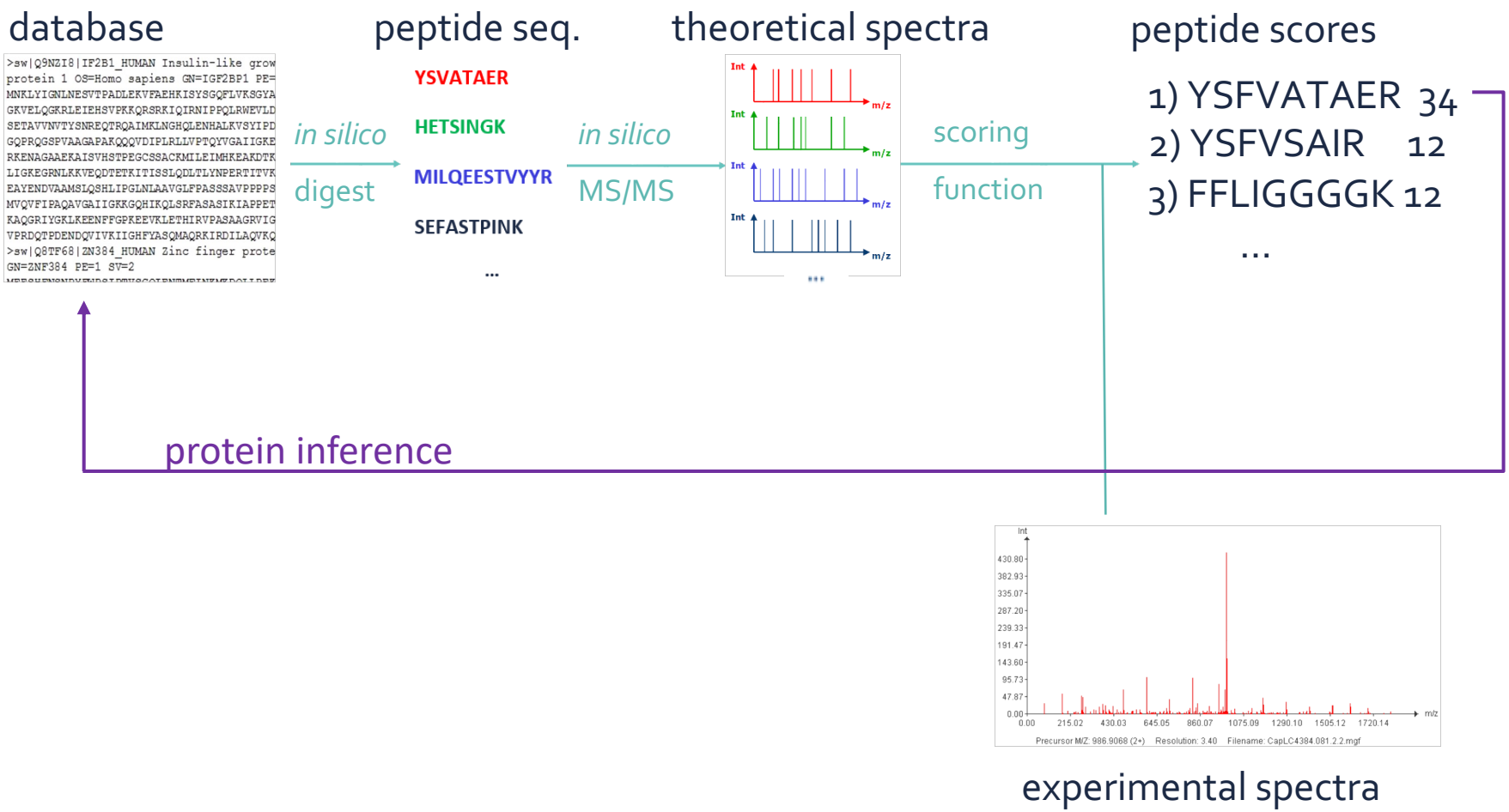
Sequencial search algorithms

Decoys and false discovery rate calculation

The future: machine learning

Protein inference: bad, ugly, and not so good

Database search engines match experimental spectra to known peptide sequences



Three popular algorithms illustrate the three types of scoring systems

SEQUEST (UWashington, Thermo Fisher Scientific)

Intensity-based scoring system

MASCOT (Matrix Science) / Andromeda (Jürgen Cox)

Peak counting-based scoring system

X!Tandem (The Global Proteome Machine Organization)

Hybrid scoring system

SEQUEST is the original search engine, and is based on ion intensity matching

Can be used for MS/MS (PFF) identifications

Based on a cross-correlation score (includes peak height)

Published core algorithm (patented, licensed to Thermo), Eng, *JASMS* 1994

Provides preliminary (Sp) score, rank, cross-correlation score (XCorr),

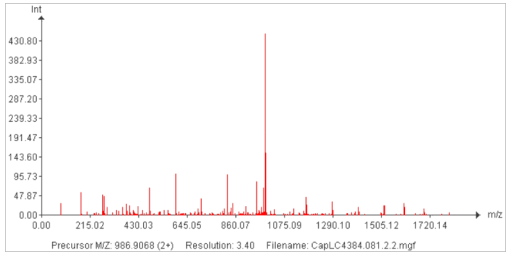
and score difference between the top tow ranks (deltaCn, ΔCn)

Thresholding is up to the user, and is commonly done *per* charge state

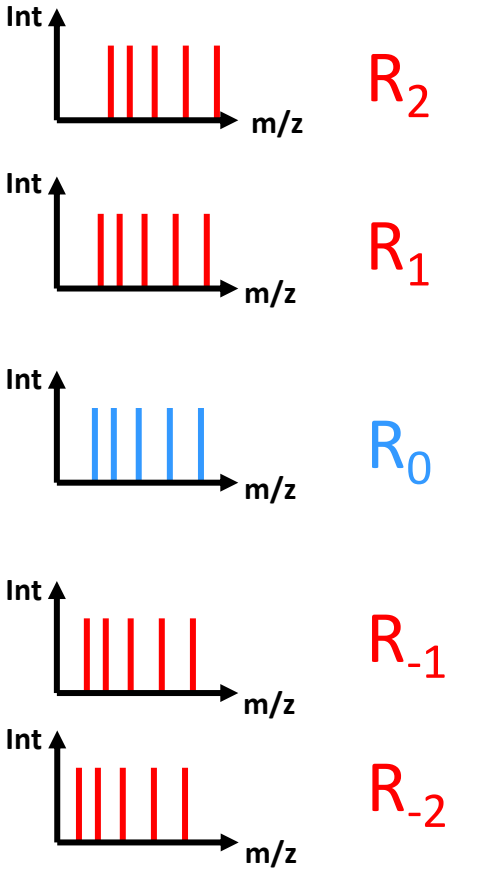
Many extensions exist to perform a more automatic validation of results

The correlation score (R_i) is calculated as the matched ion intensity

$$\sum_{i=-75}^{+75} Ri$$



$$R_i = \sum_{j=1}^{n} x_j \cdot y_{(j+i)}$$

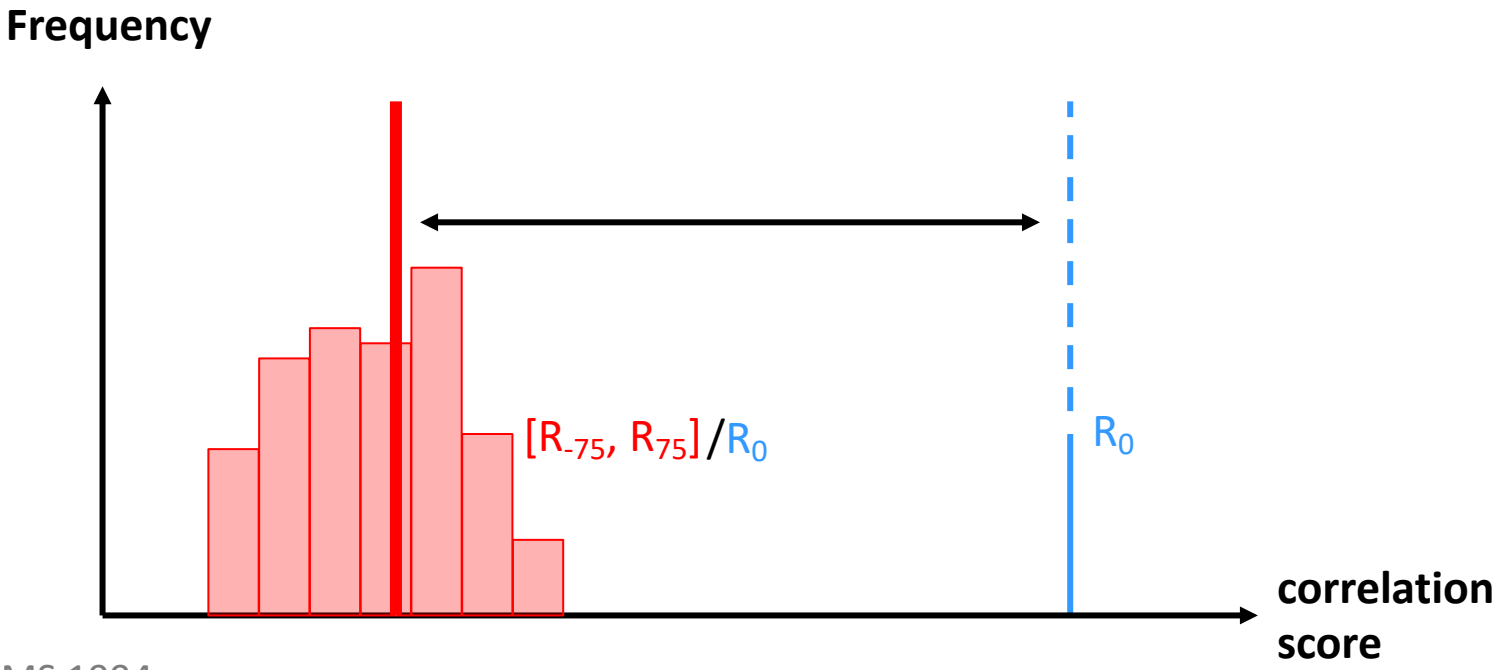


Eng, JASMS 1994

Yılmaz, Proteome Bioinformatics (MMB), Springer, 2017

The cross-correlation score (*Xcorr*) is R_0 calibrated by the average random correlation

$$\text{XCorr} = R_0 - \frac{1}{150}\left(\sum_{i=-75/R_0}^{+75} Ri\right)$$

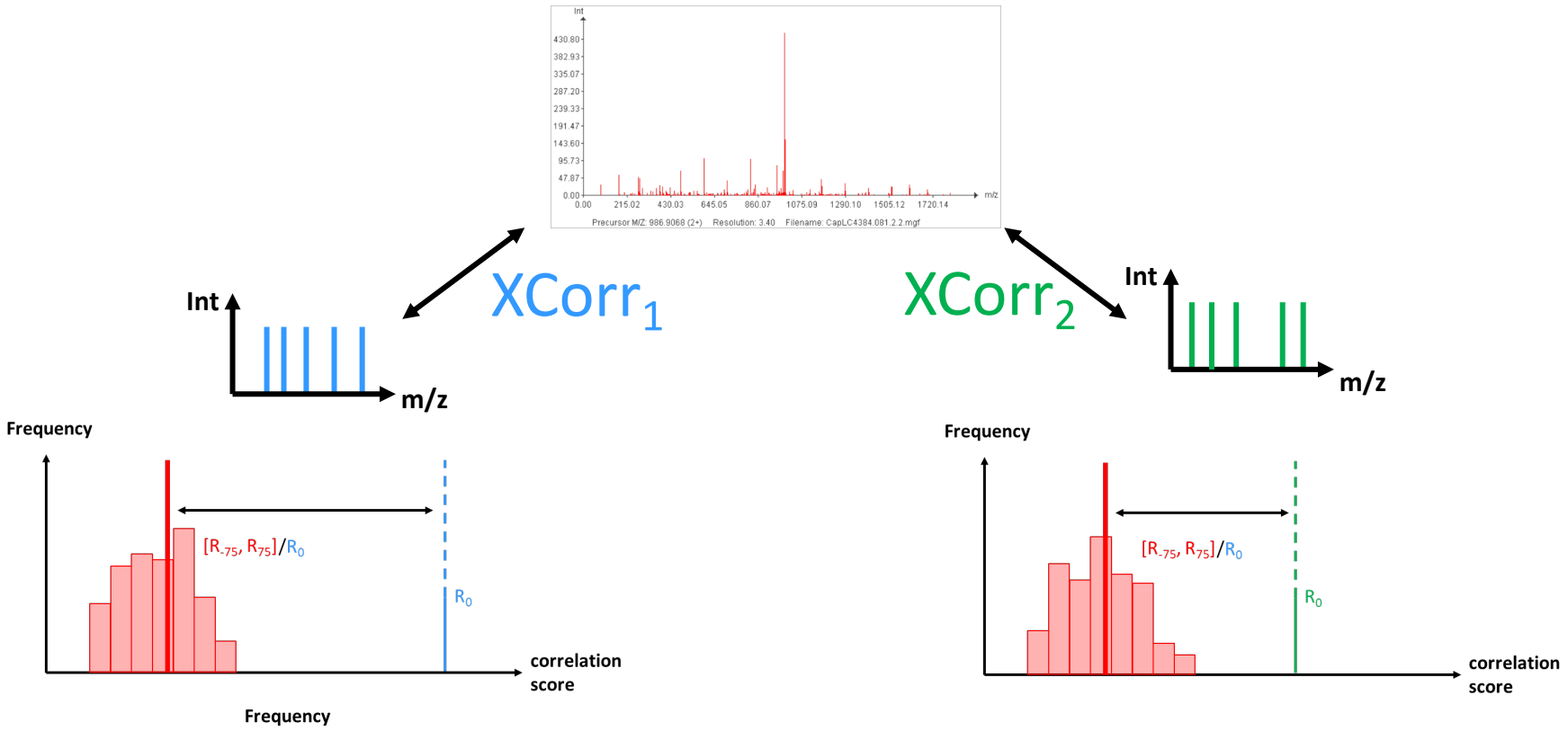


Eng, JASMS 1994

Yılmaz, Proteome Bioinformatics (MMB), Springer, 2017

The best theoretical match is then compared to the second-best theoretical match

$$\text{deltaCn} = \frac{XCorr_1 - XCorr_2}{XCorr_1}$$



Eng, JASMS 1994

Yılmaz, Proteome Bioinformatics (MMB), Springer, 2017

Mascot is an equally recognized search engine, but is based on peak counting

Very well established search engine, Perkins, *Electrophoresis* 1999

Can do MS (PMF) and MS/MS (PFF) identifications

Based on the MOWSE score,

Unpublished core algorithm (trade secret)

Predicts an *a priori* threshold score that identifications need to pass

From version 2.2, Mascot allows integrated decoy searches

Provides rank, score, threshold and expectation value per identification

Customizable confidence level for the threshold score

Through Andromeda, we understand MASCOT

$$s = -10 \times \log_{10} \sum_{j=k}^{n} \left[\binom{n}{j} (p)^{j} (1-p)^{n-j} \right]$$

n = number of theoretical peaks
k = number of matched peaks (within a given fragment tolerance)
p = probability of finding a single, matched peak by chance
p is calculated by dividing the number of highest intensity peaks (q) by a mass-window size (100 Da)
q is limited by a maximum value, and is optimized for maximum *s*

based on **peak counting** instead of intensity sums

Cox, J Prot Res, 2011
Yılmaz, Proteome Bioinformatics (MMB), Springer, 2017

X!Tandem introduces a hybrid score, based on both peak counting and ion intensity

A successful open source search engine, Craig and Beavis, *RCMS* 2003

Can be used for MS/MS (PFF) identifications

Based on a hyperscore (*Pi* is either 0 or 1):

$$HyperScore = \left(\sum_{i=0}^{n} I_i * P_i \right) * N_b! * N_y!$$

Relies on a hypergeometric distribution (hence hyperscore)

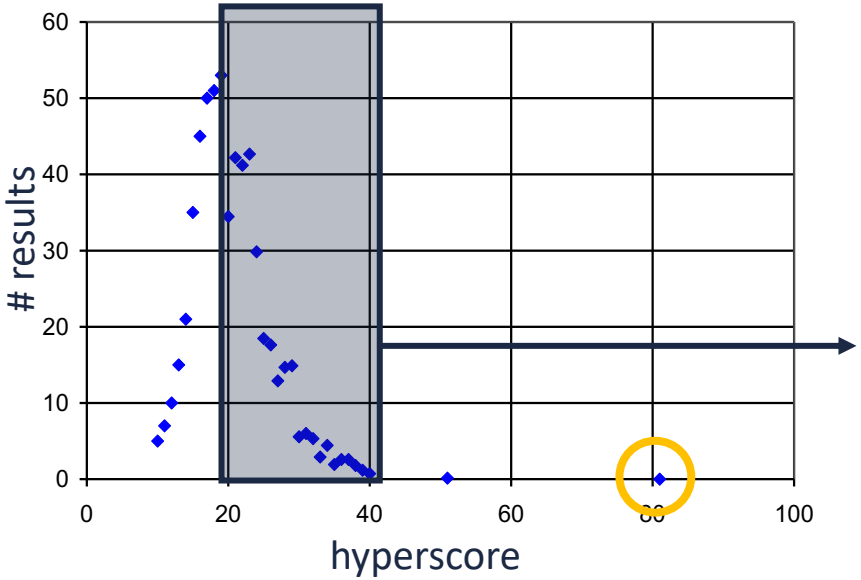
Published core algorithm, and is freely available

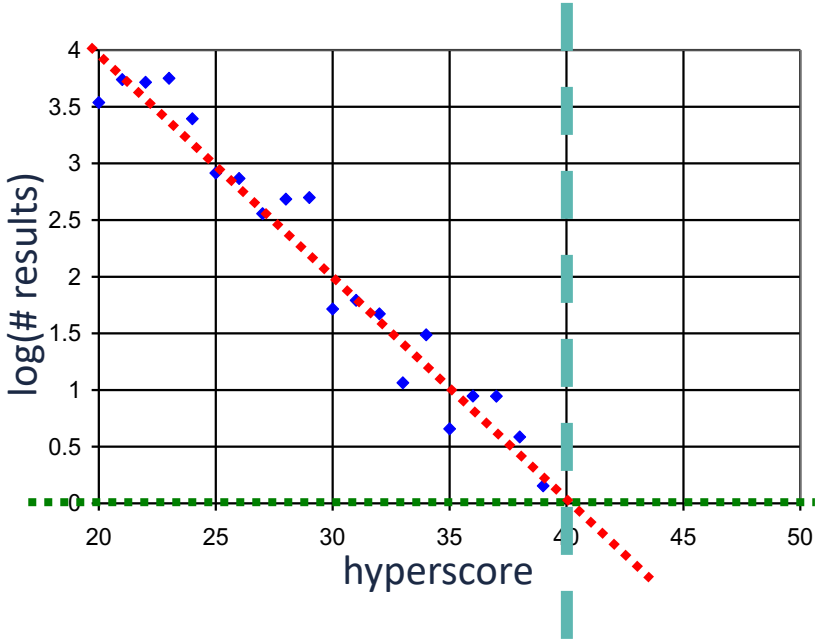
Provides hyperscore and expectancy score (the discriminating one)

X!Tandem is fast and can handle modifications in an iterative fashion

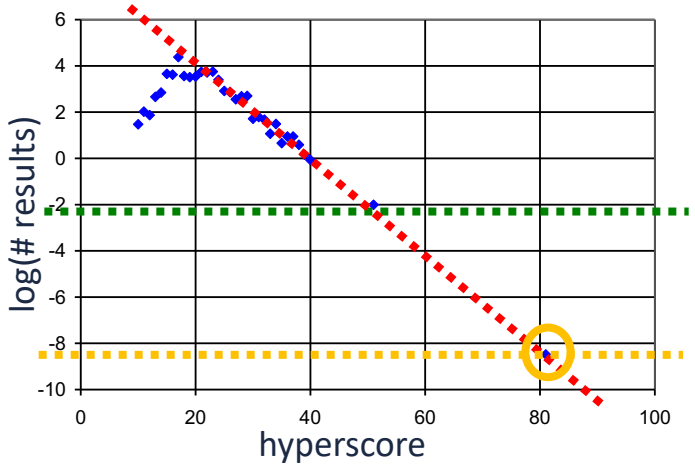
Has rapidly gained popularity as (auxiliary) search engine

X!Tandem's significance calculation for scores can be seen as a general template





significance threshold



E-value=$e^{-8.2}$

Adapted from: Brian Searle, ProteomeSoftware, http://www.proteomesoftware.com/XTandem_edited.pdf

MS/MS spectra and identification

Database search algorithms in three phases

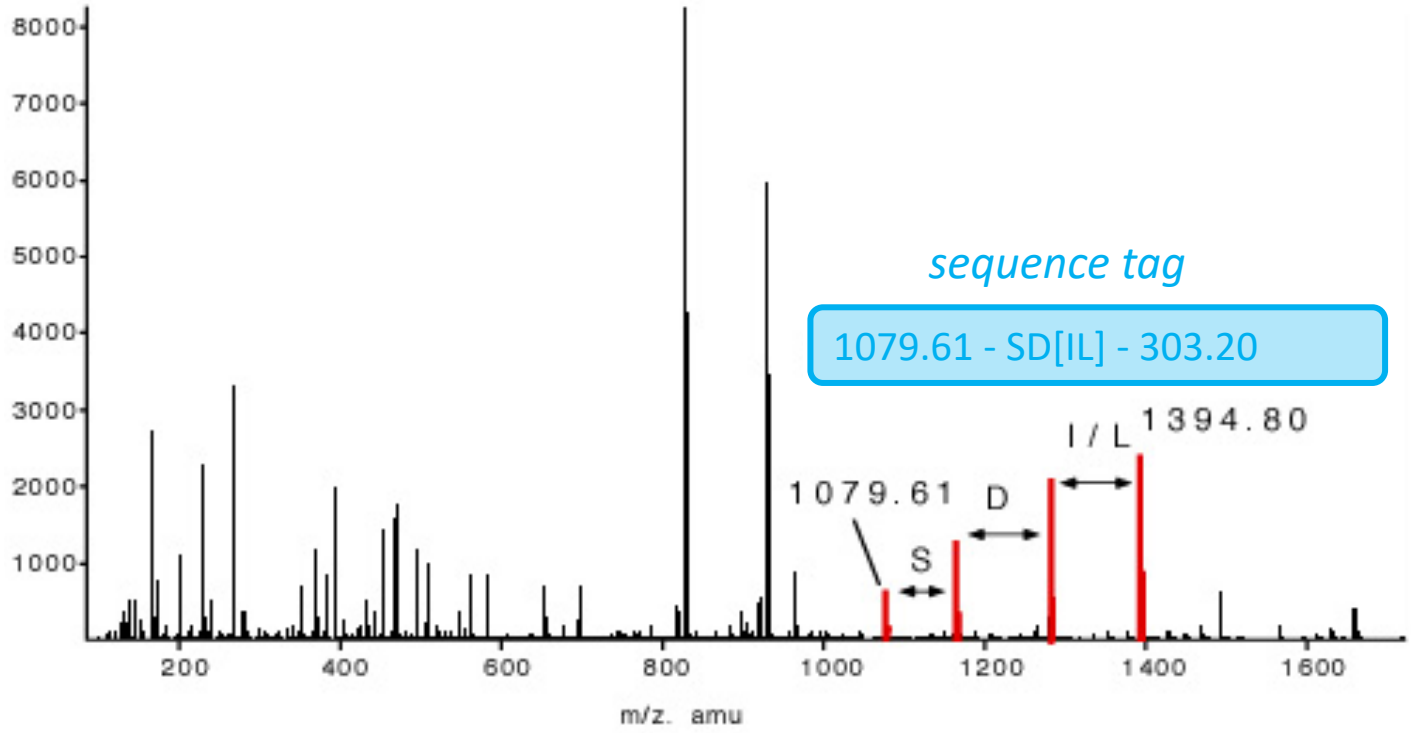
Sequencial search algorithms

Decoys and false discovery rate calculation

The future: machine learning

Protein inference: bad, ugly, and not so good

Sequence tags are as old as SEQUEST, and still have a role to play today



The concept of sequence tags was introduced by Mann and Wilm

Mann, Analytical Chemistry, 1994

GutenTag, DirecTag, TagRecon

Tabb, *Anal. Chem.* 2003, Tabb, *JPR* 2008, Dasari, *JPR* 2010

Recent implementations of the sequence tag approach

Refine hits by peak mapping in a second stage to resolve ambiguities

Rely on a empirical fragmentation model

Published core algorithms, DirecTag and TagRecon freely available

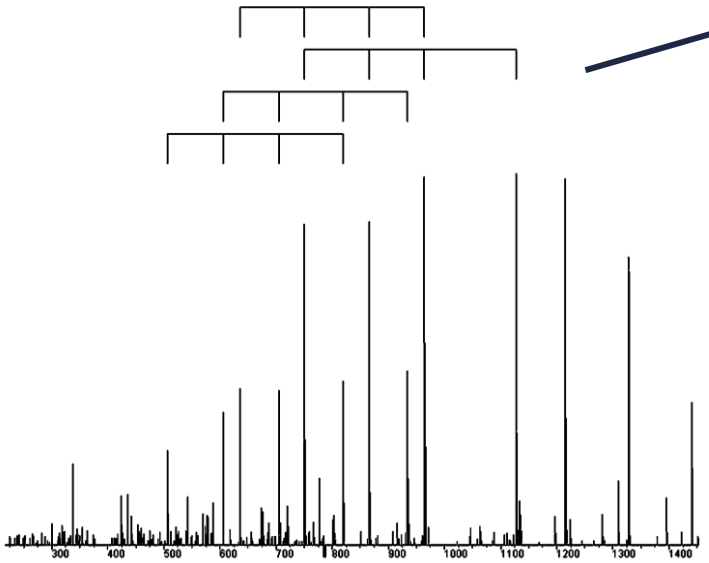
GutenTag/DirecTag extracts tags, TagRecon matches tags to database

Very useful to retrieve unexpected peptides (modifications, variations)

Entire workflows exist (e.g., combination with IDPicker)

GutenTag: two stage, hybrid tag searching

1. Generate sequence tags



2. Search DB for matches

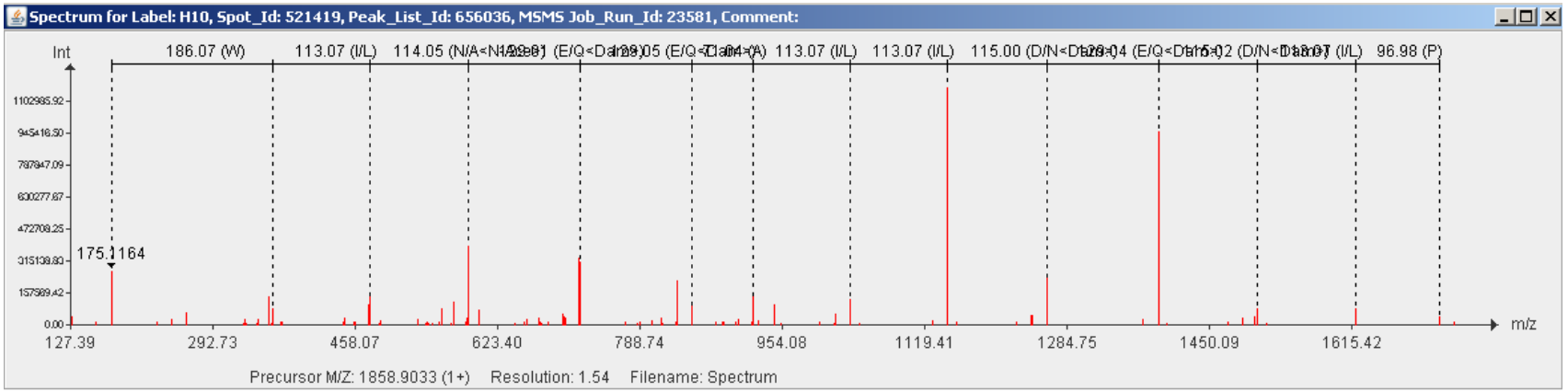
DDG → -DDGNSDRS
YVD → -YVDVNKFKD
VDD → KLLSYVDDEAFIR
DDE → EGDEANSDDEEEDL
DDV → -DDVDIDEN
VVD → SSCTAVVD-
DVY → AFQYLKDVY-

3. Score DB Sequences

KLLSYVDDEAFIR	19.36
-DDVDIDEN	8.56
-DDGNSDRS	6.94
-YVDVNKFKD	6.25
SSCTAVVD-	5.74
EGDEANSDDEEEDL	5.64
AFQYLKDVY-	5.61

Tabb, Analytical Chemistry, 2003

De novo sequencing tries to read the entire peptide sequence from the spectrum



Example of a manual de novo of an MS/MS spectrum

No more database necessary to extract a sequence!

Algorithms	***References***
Lutefisk	Dancik 1999, Taylor 2000
Sherenga	Fernandez-de-Cossio 2000
PEAKS	Ma 2003, Zhang 2004
PepNovo	Frank 2005, Grossmann 2005
RapidNovor	Ma 2015
…	…

MS/MS spectra and identification

Database search algorithms in three phases

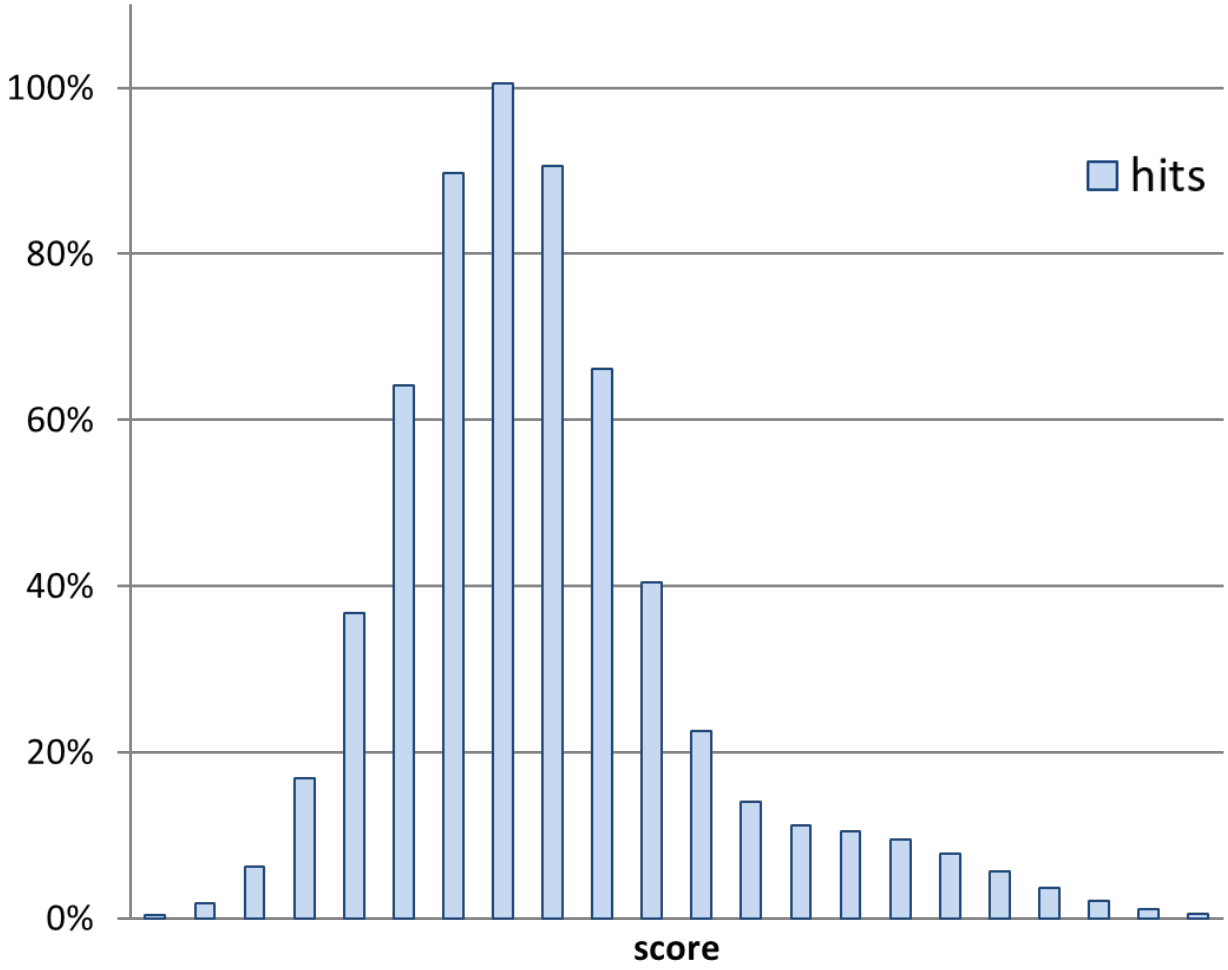
Sequencial search algorithms

Decoys and false discovery rate calculation

The future: machine learning

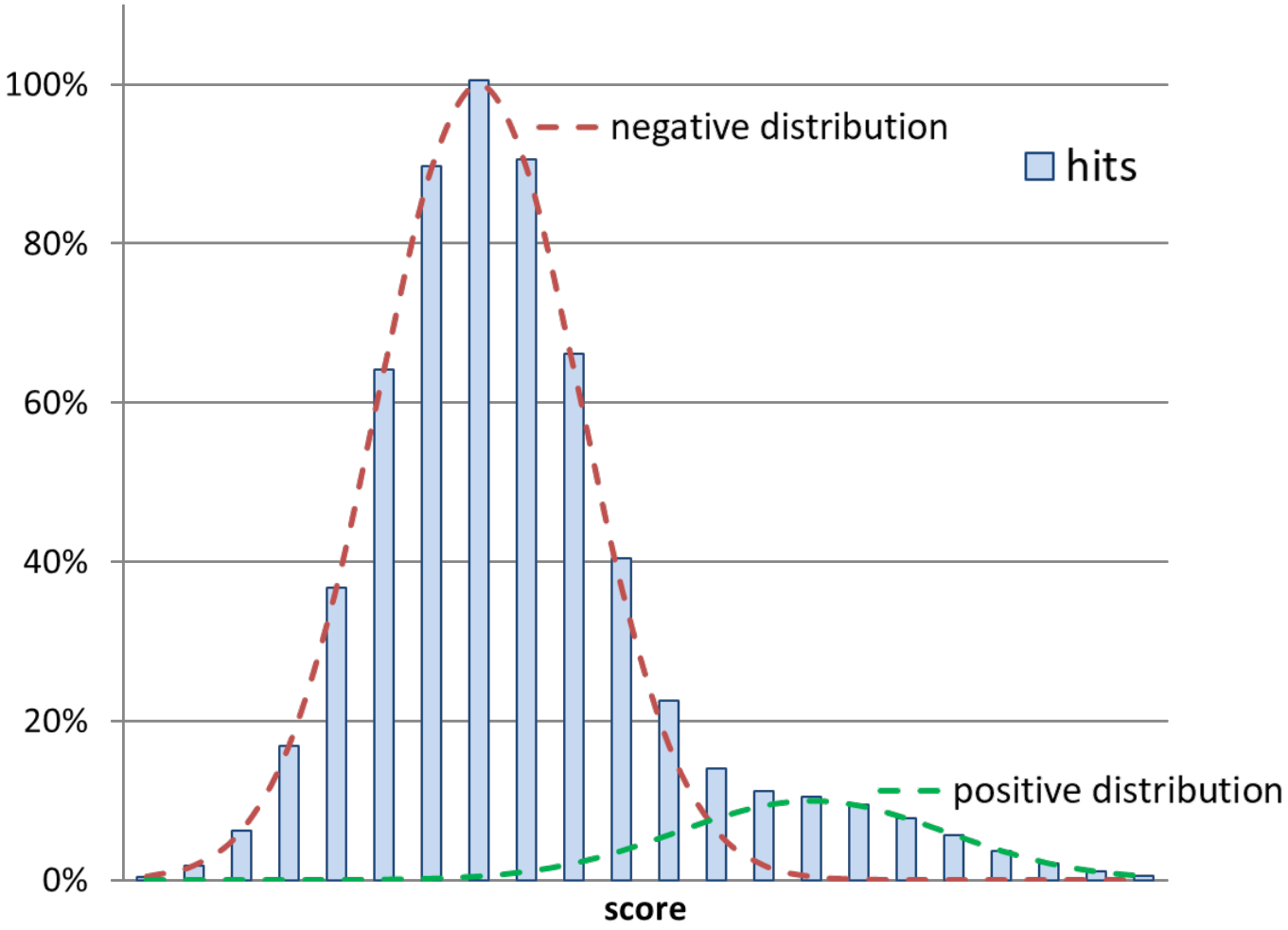
Protein inference: bad, ugly, and not so good

All hits, good and bad together, form a distribution of scores

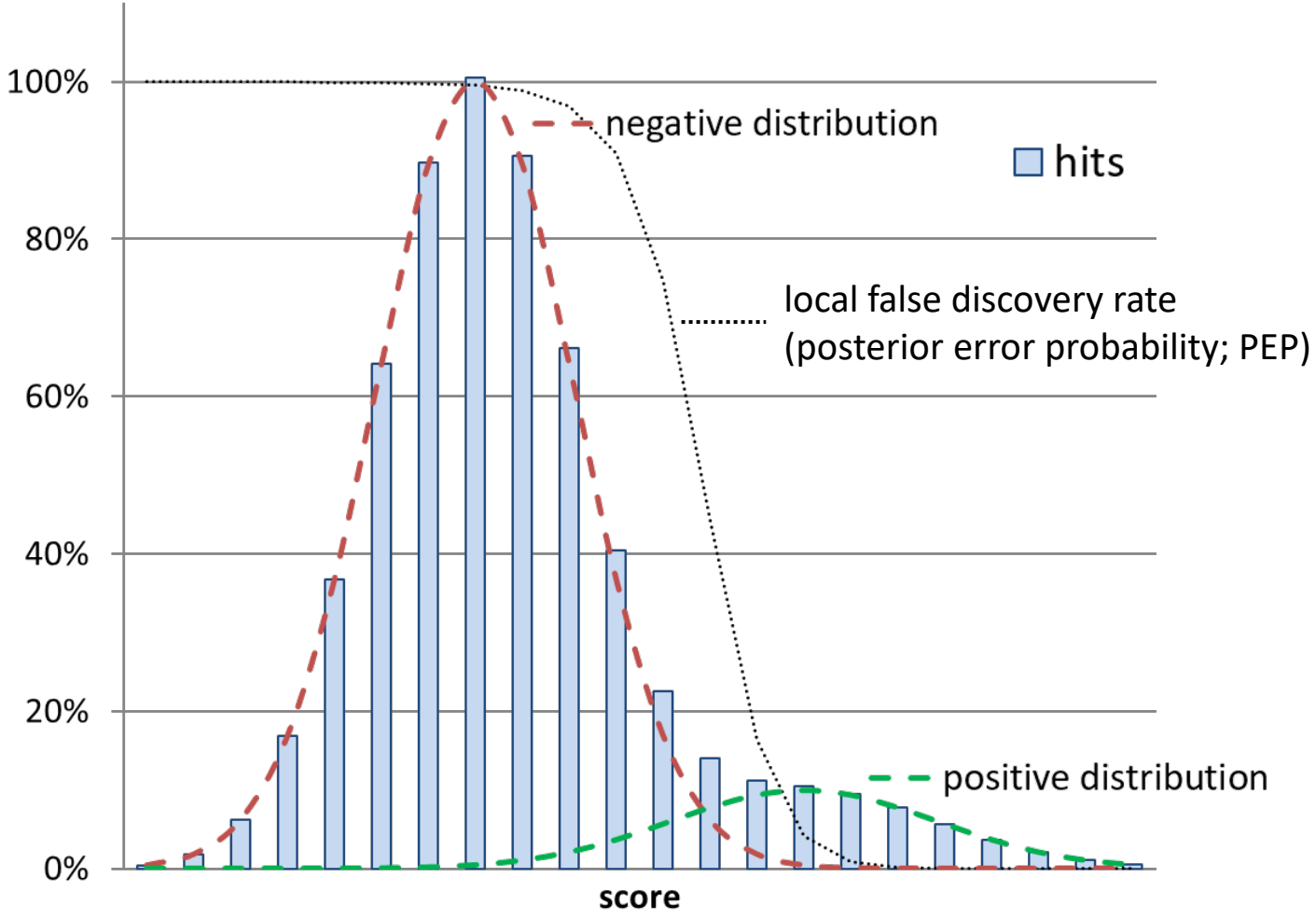


Nesvizhskii, J Proteomics, 2010

If we know how scores for bad hits distribute, we can distinguish good from bad by score



The separation is not perfect, which leads to the calculation of a local false discovery rate



Decoy databases are false positive factories, assumed to deliver representative bad hits

Three main types of decoy DB's are used:

- Reversed databases (*easy*)

  LENNARTMARTENS → SNETRAMTRANNEL

- Shuffled databases (*slightly more difficult*)

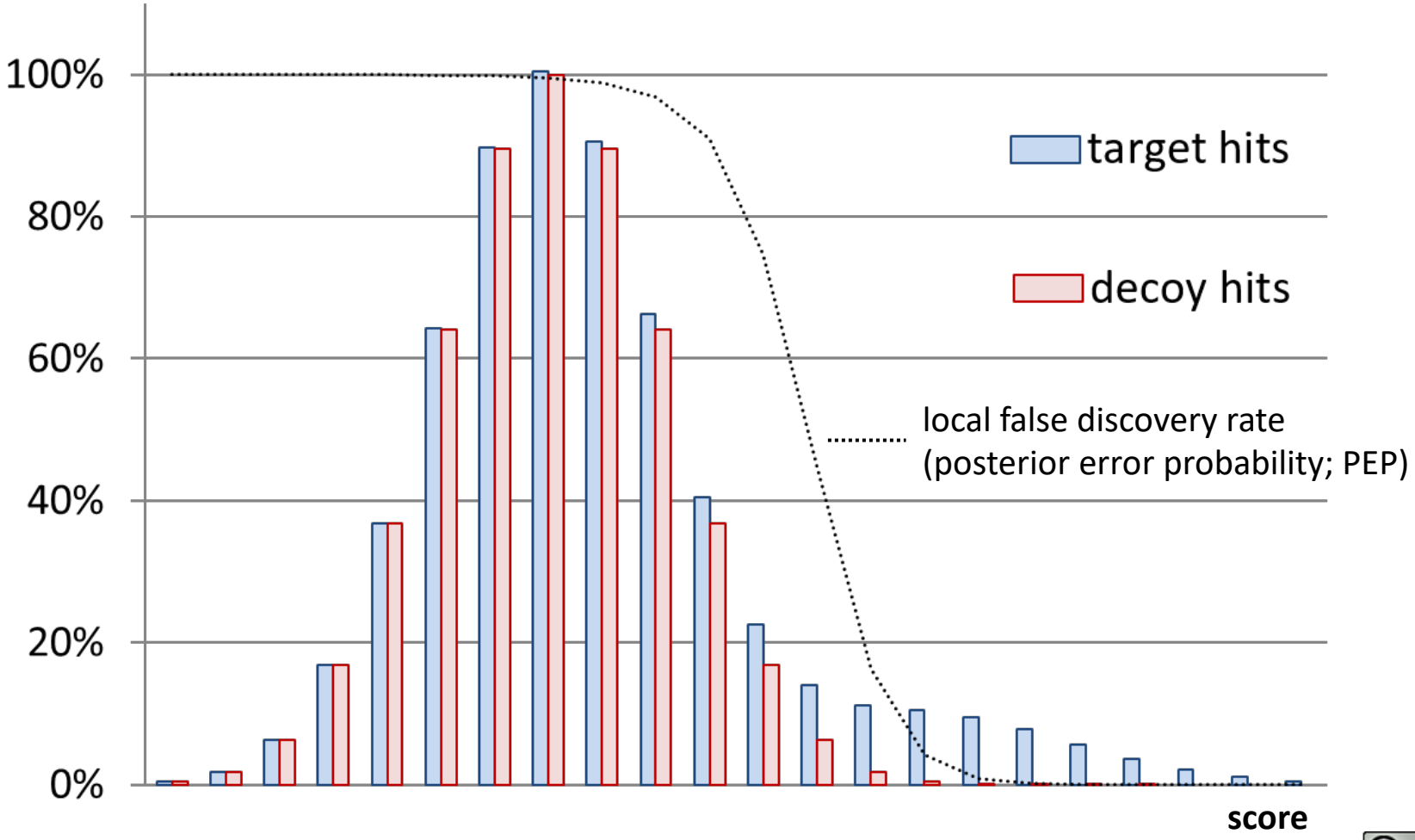
  LENNARTMARTENS → NMERLANATERTTN *(for instance)*

- Randomized databases (*as difficult as you want it to be*)

  LENNARTMARTENS → ~~GFVLAEPHSEAITK~~ *(for instance)*

The concept is that each peptide identified from the decoy database is an incorrect identification. By counting the number of decoy hits, we can estimate the number of false positives in the original database, **provided that the decoys have similar properties as the forward sequences.**

With the help of the scores of decoy hits, we can assess the score distribution of bad hits



Käll, Journal of Proteome Research, 2008

MS/MS spectra and identification

Database search algorithms in three phases

Sequencial search algorithms

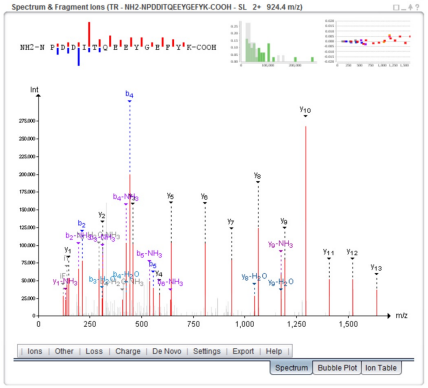
Decoys and false discovery rate calculation

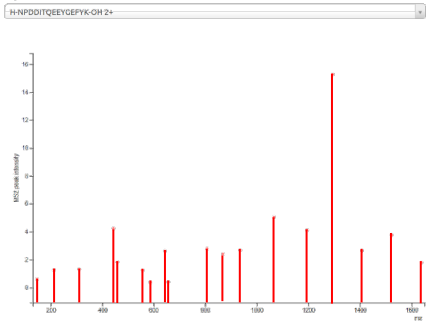
The future: machine learning

Protein inference: bad, ugly, and not so good

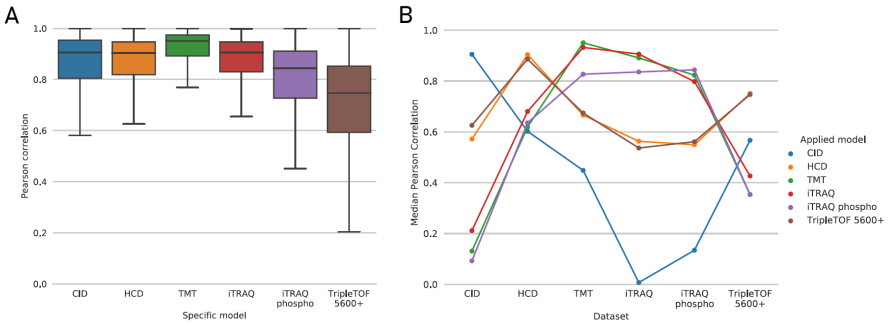
Our MS2PIP fragmentation model accurately predicts peptide fragmentation behaviour

Vaudel, Nat. Biotech., 2015
PeptideShaker



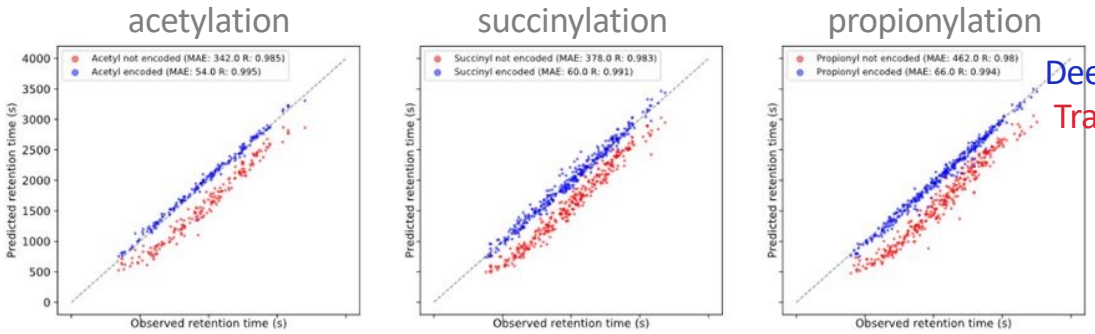


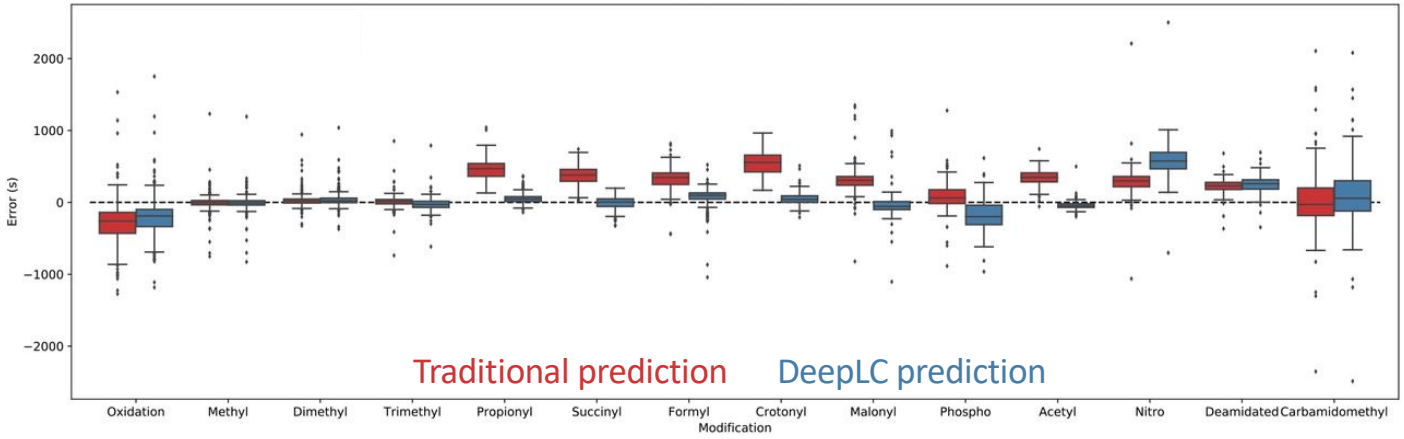
https://iomics.ugent.be/ms2pip
Degroeve, Bioinformatics, 2013
Degroeve, Nucleic Acids Research, 2015



https://iomics.ugent.be/ms2pip/
Gabriels, Nucleic Acids Research, 2019

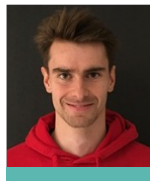
Our DeepLC model accurately predicts retention times of peptides with unseen modifications

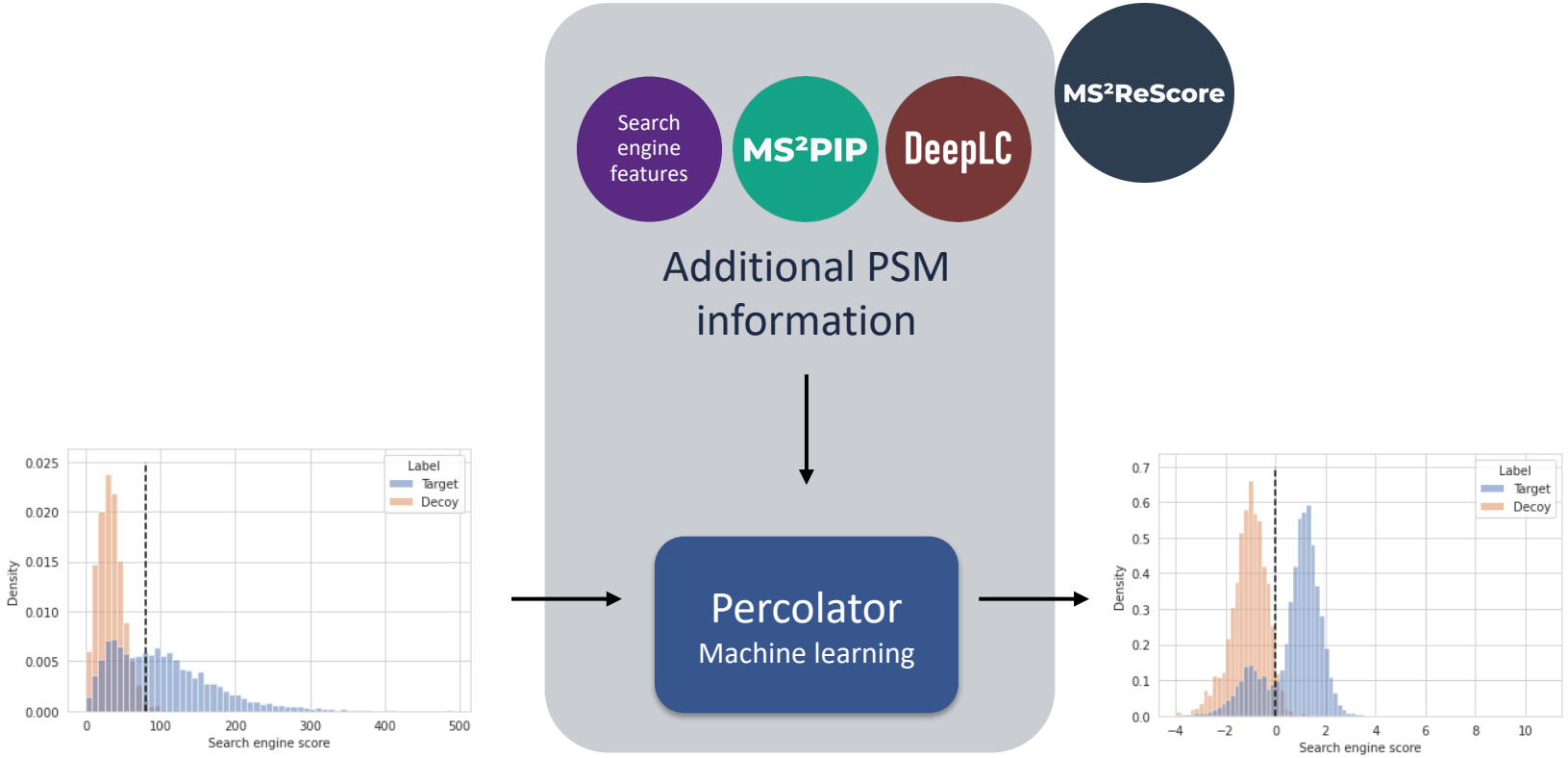




Dataset from Zolg, MCP, 2018
Bouwmeester, Nature Methods, 2021

MS²Rescores uses machine learning predictions to boost identification sensitivity and specificity

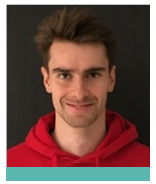


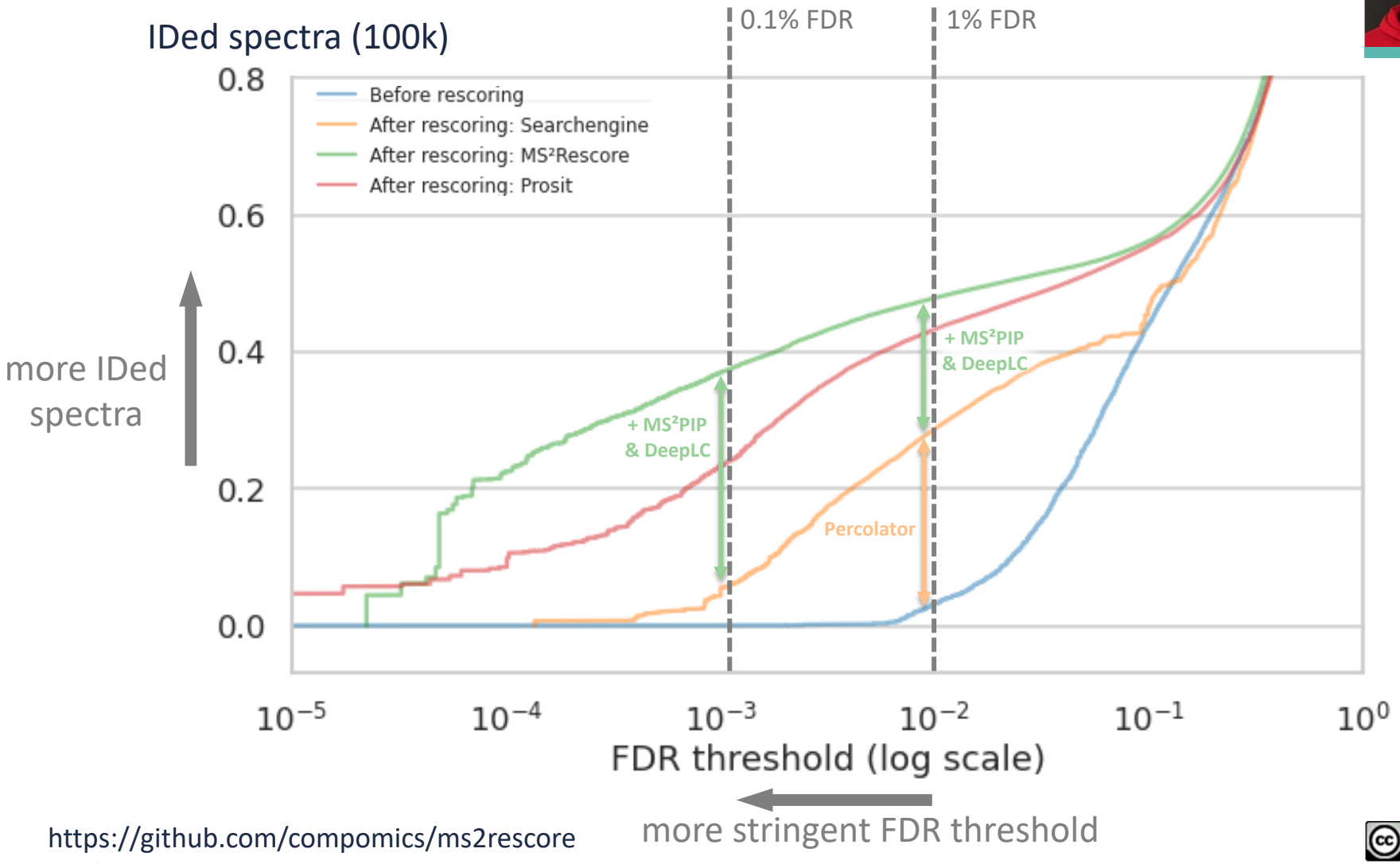


MS²Rescore: C. Silva, Bioinformatics (2019) & A. Declerq, MCP, 2022
MS²PIP: R. Gabriels, Nucleic Acids Research (2019)

DeepLC: R. Bouwmeester, Nature Methods (2021);
Percolator: L. Käll, Journal of Proteome Research (2009)

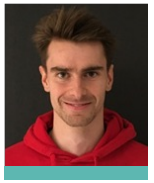
MS2PIP and DeepLC in MS²Rescore dramatically boost identification in immunopeptidomics

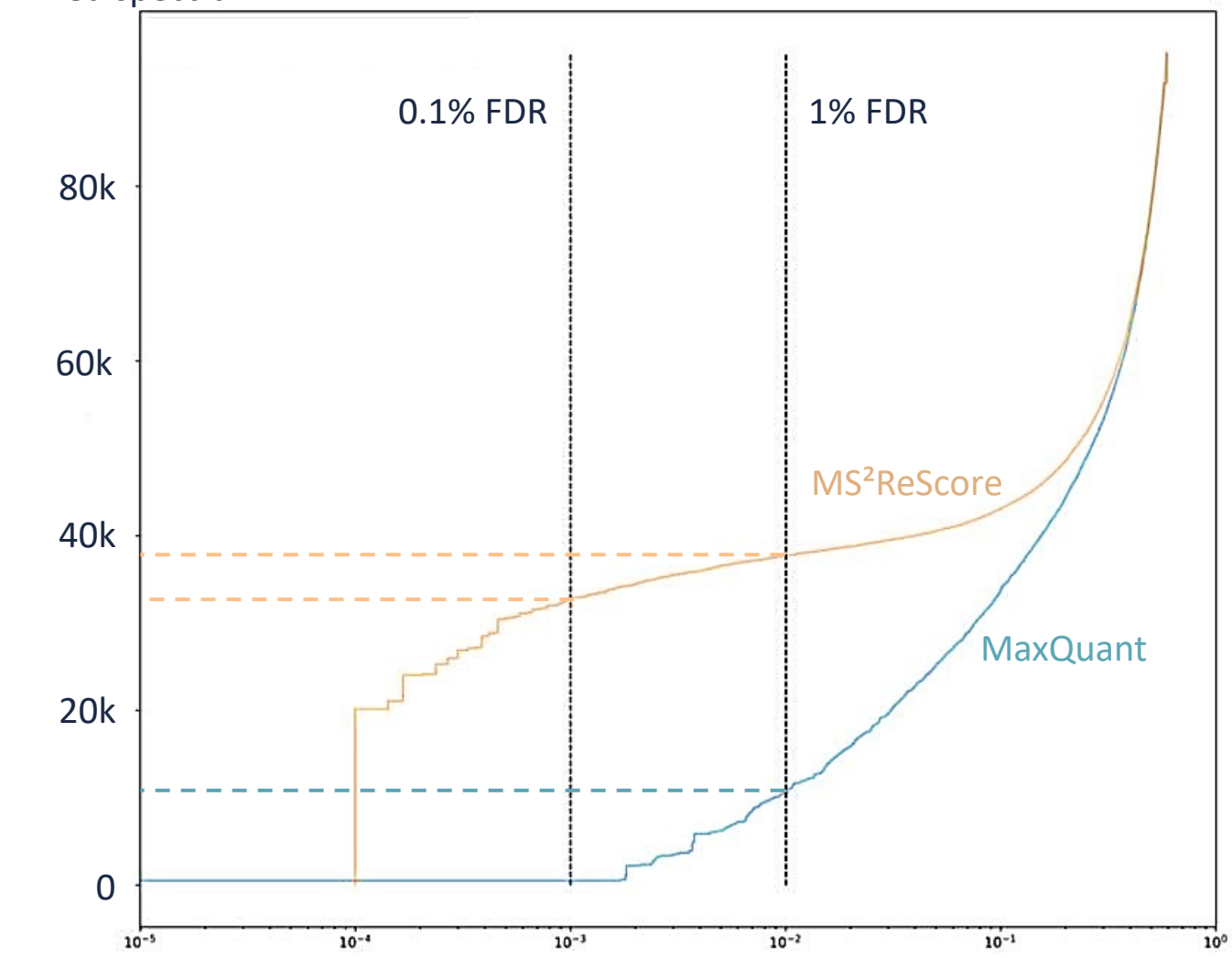




https://github.com/compomics/ms2rescore
Declercq, MCP, 2022

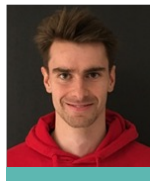
MS²Rescore can also be applied to generic peptidomics data

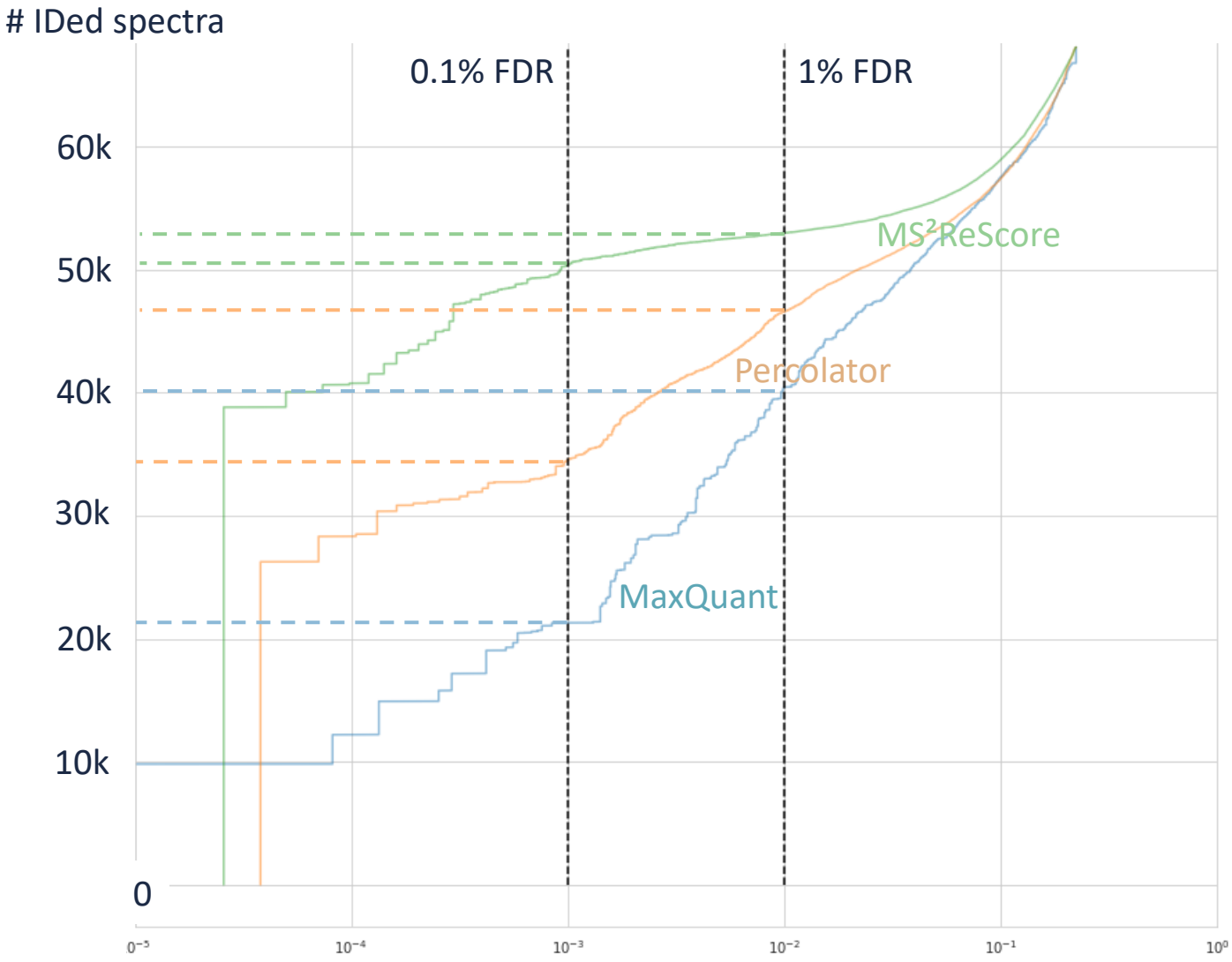




https://github.com/compomics/ms2rescore

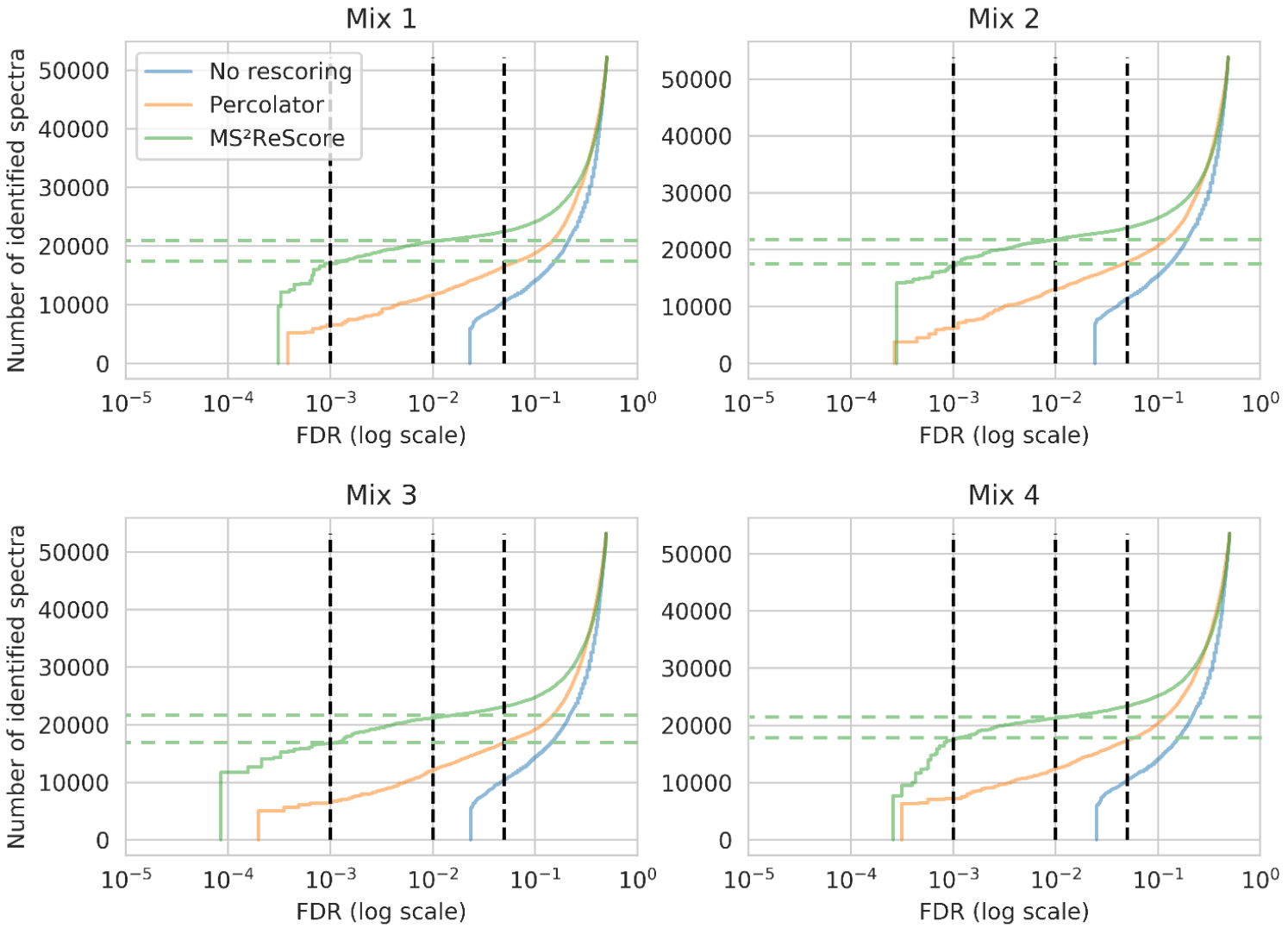
MS²Rescore can also be applied and to single cell data

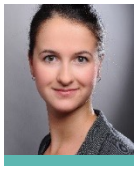


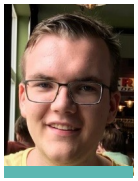


https://github.com/compomics/ms2rescore

MS²Rescore also boosts metaproteomics, opening up the prospect of meta-immunopeptidomics

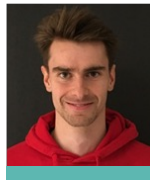


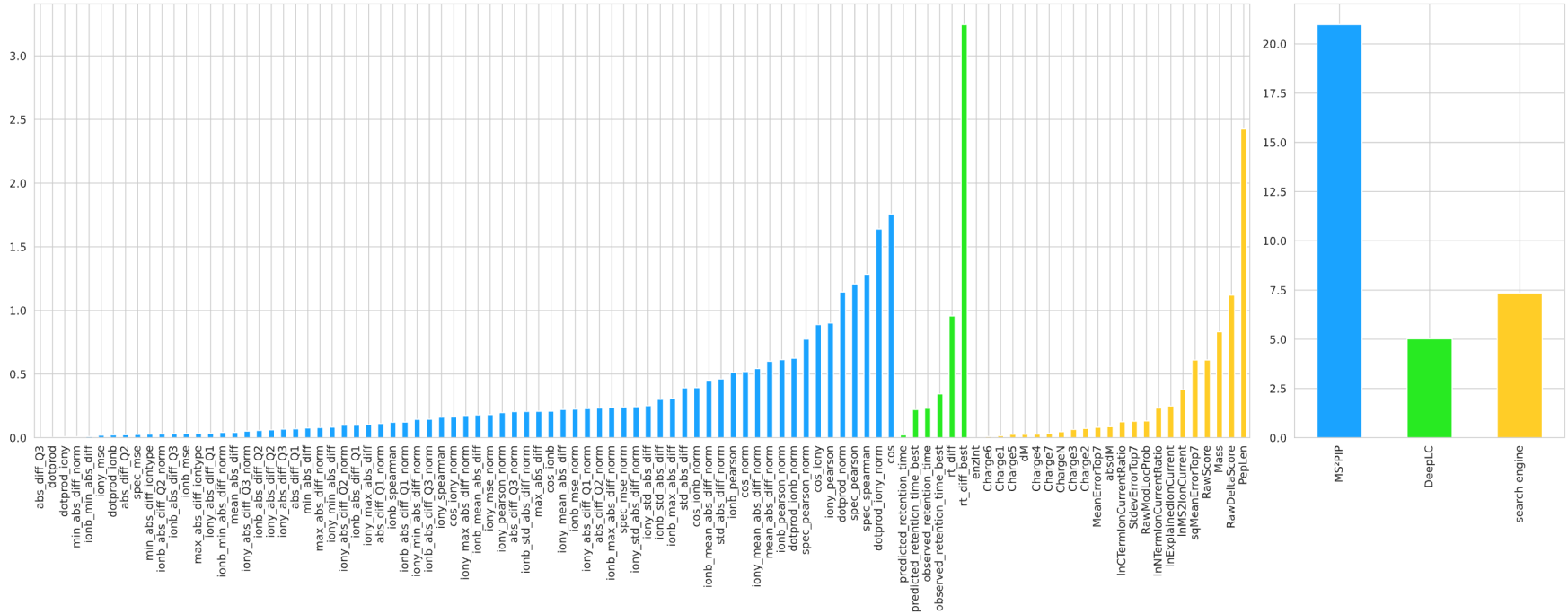




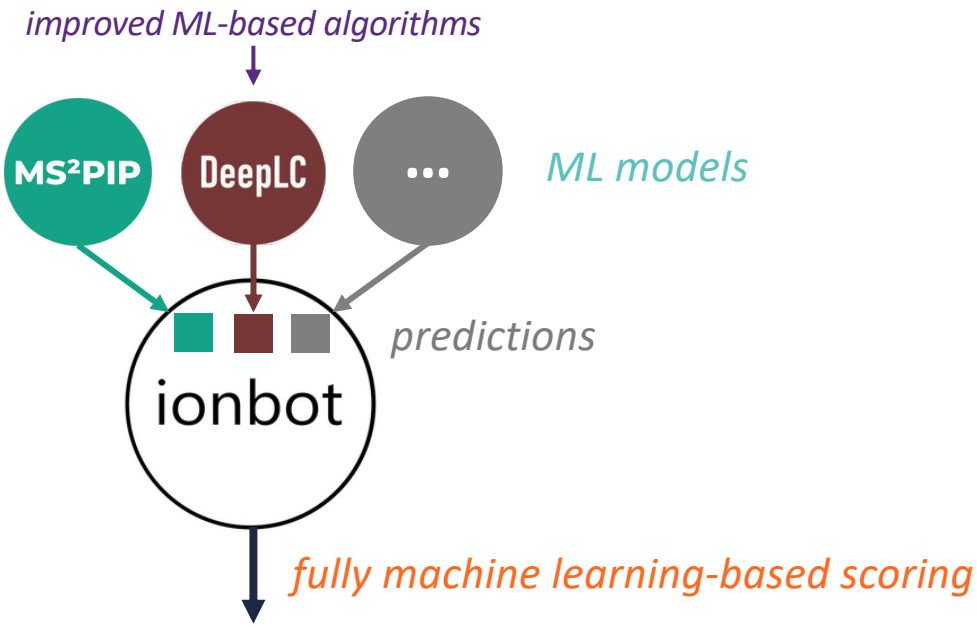
Van Den Bossche *et al.*, *in preparation*

The feature weights in MS²Rescore show that predicted features matter – a lot.





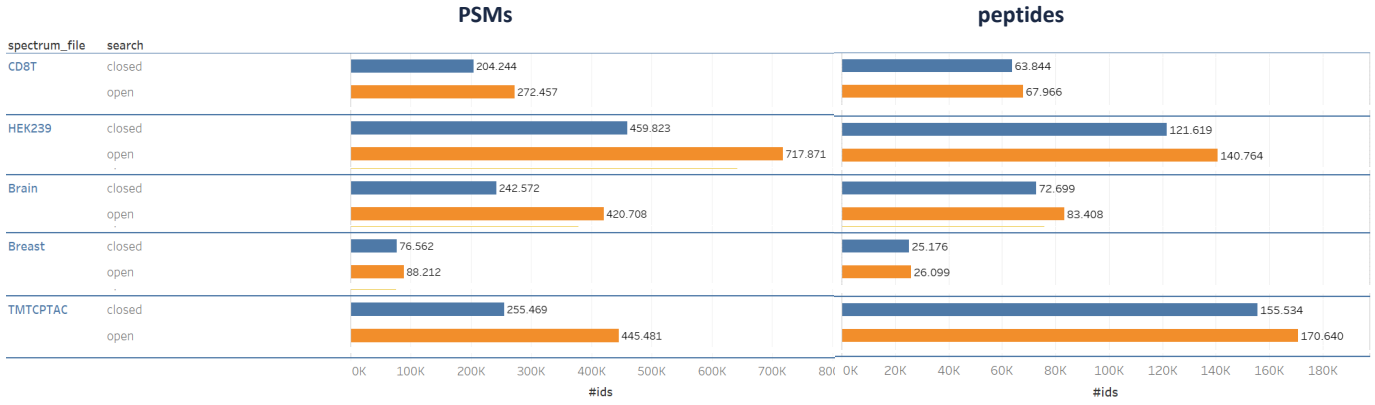
MS2PIP and DeepLC power ionbot, a novel open modification search engine with high reliability

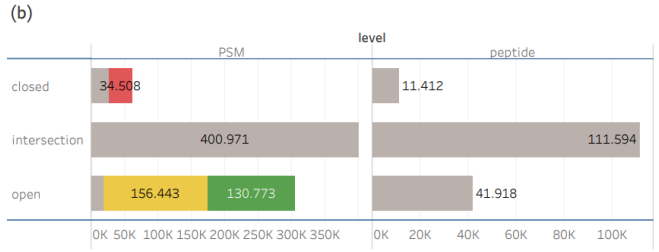


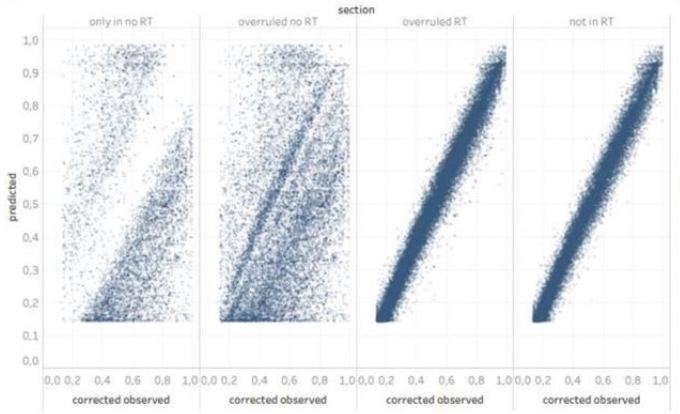
sensitive and highly accurate
identification of (modified) peptides

https://ionbot.cloud
Degroeve, https://www.biorxiv.org/content/10.1101/2021.07.02.450686v2

ionbot shows the value of open modification searches, and of accurate prediction models

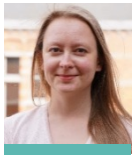




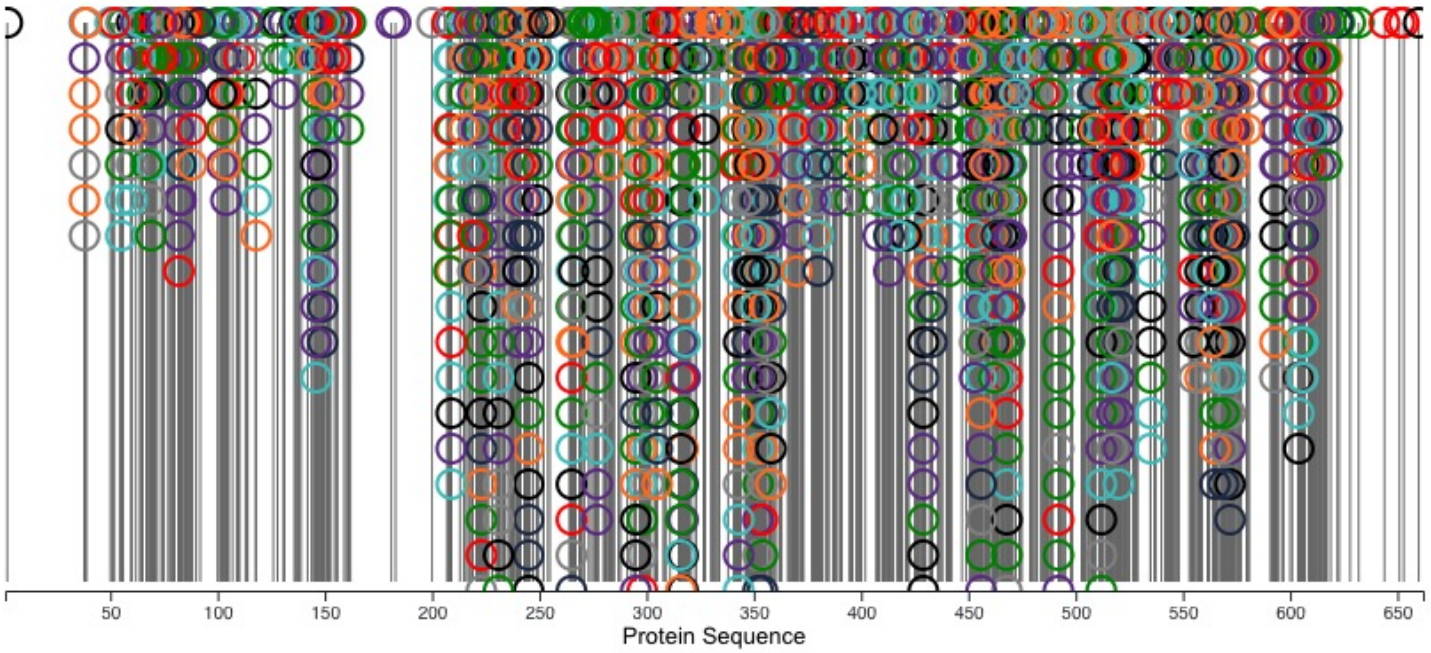


https://ionbot.cloud
Degroeve, https://www.biorxiv.org/content/10.1101/2021.07.02.450686v2

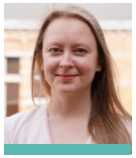
When all PTMs are considered, our view of proteins is changed

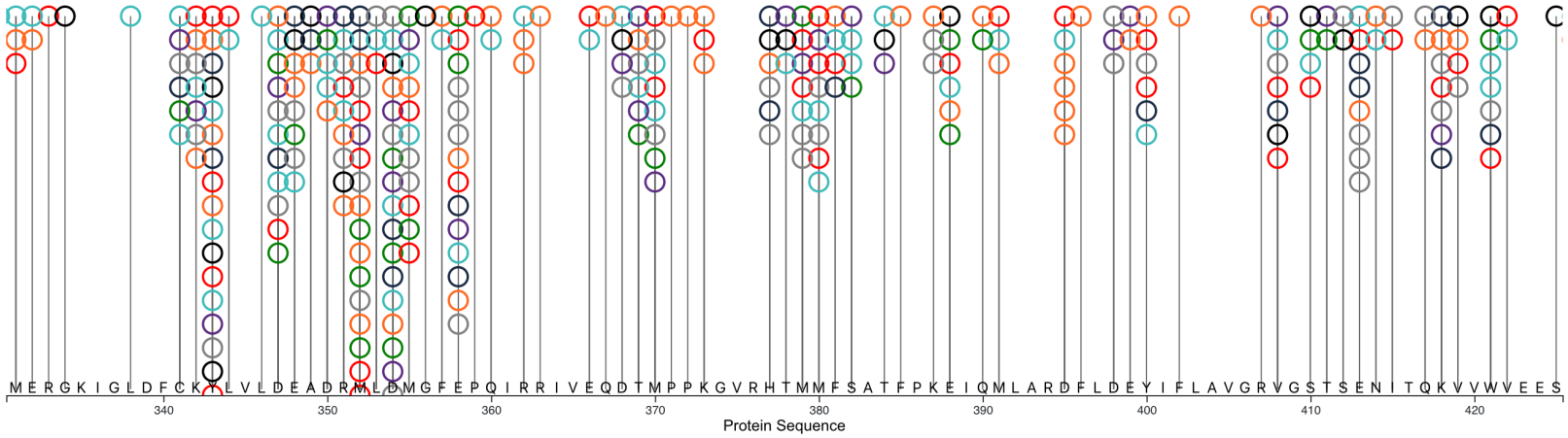


p00571 Summary Peptides Structures Mutations

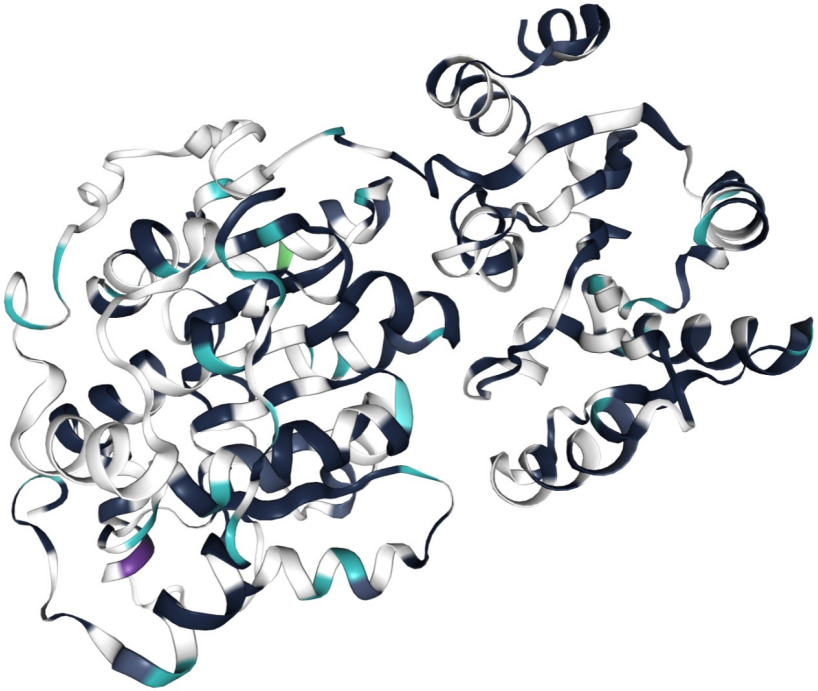


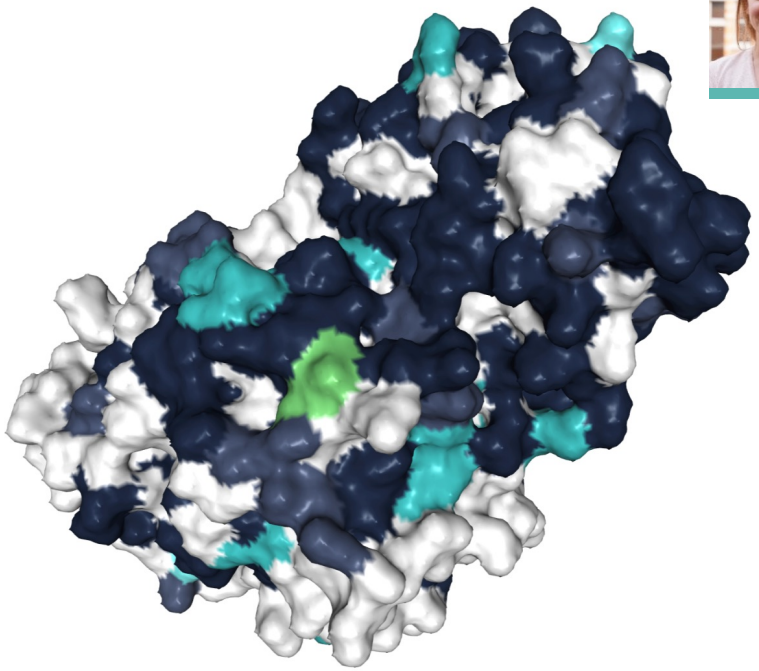
Zooming in shows that not all residues are created equal

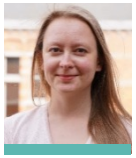




The 3D structure view also becomes rather crowded







MS/MS spectra and identification

Database search algorithms in three phases

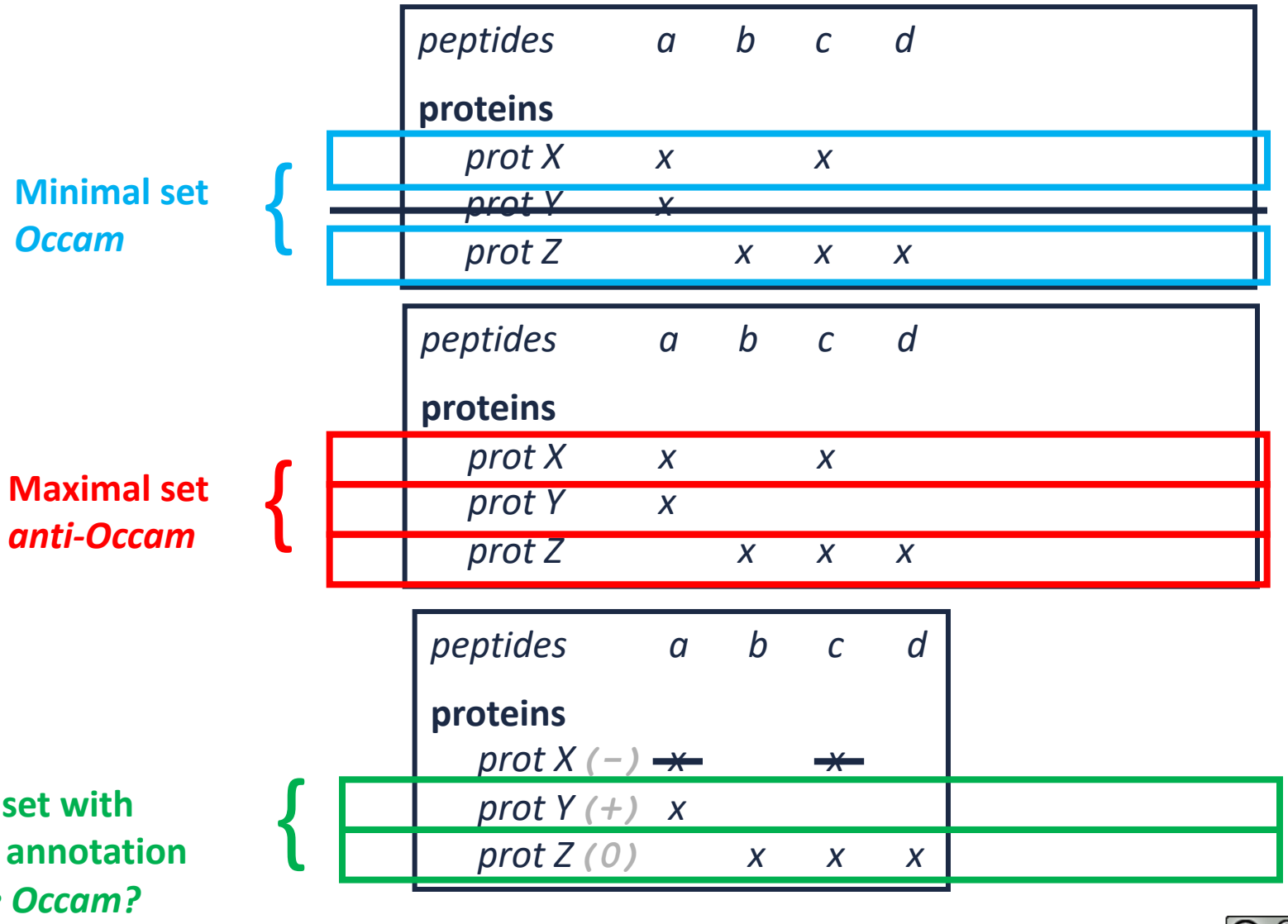
Sequencial search algorithms

Decoys and false discovery rate calculation

The future: machine learning

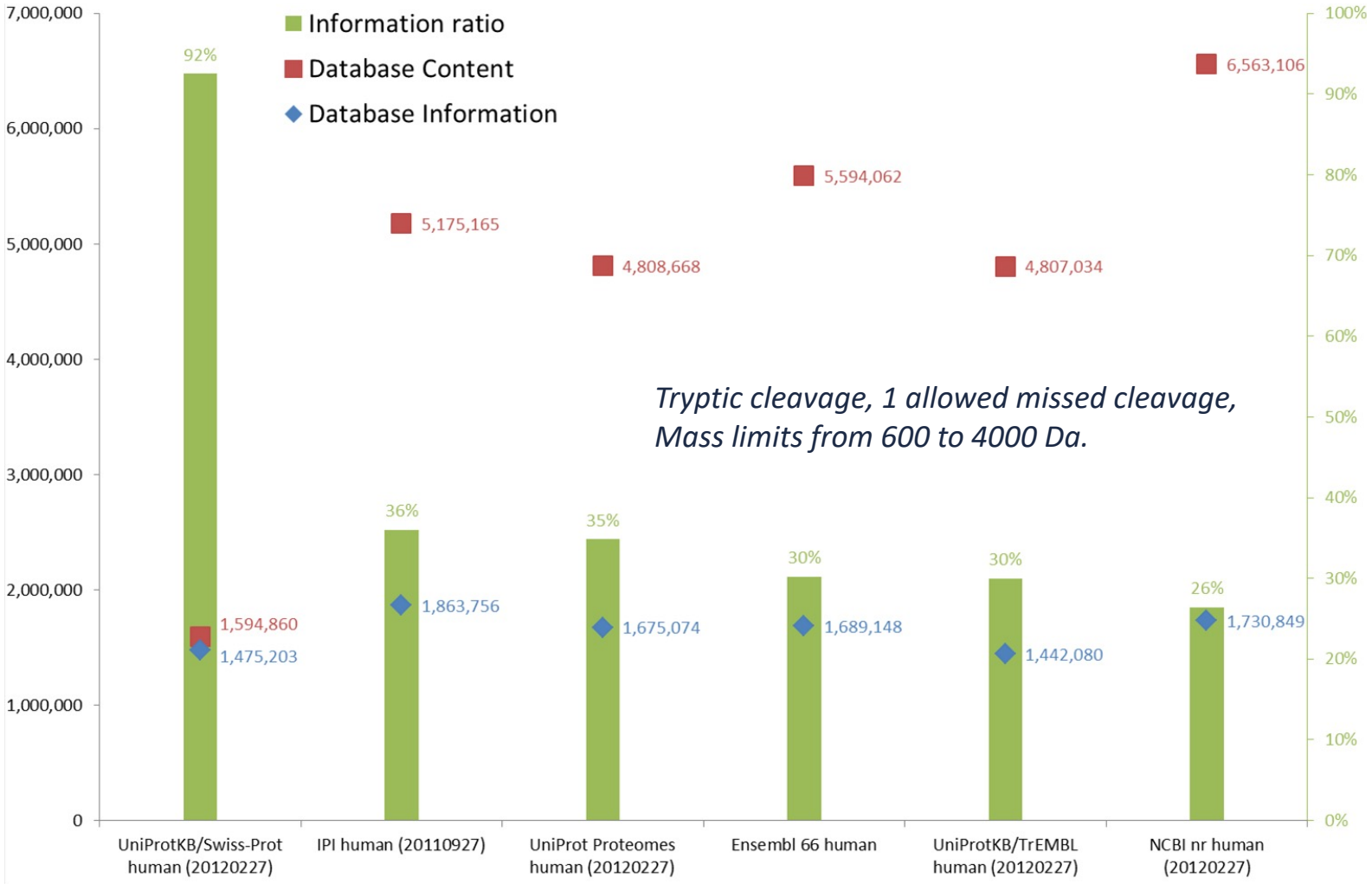
Protein inference: bad, ugly, and not so good

Protein inference is a question of conviction



Martens, Molecular Biosystems, 2007

The complexity of protein inference is linked to the information ratio of a database



Barsnes, Amino Acids, 2013

In real life, protein inference issues will be mainly bad, often ugly, and occasionally good