



Solving big problems on small computers



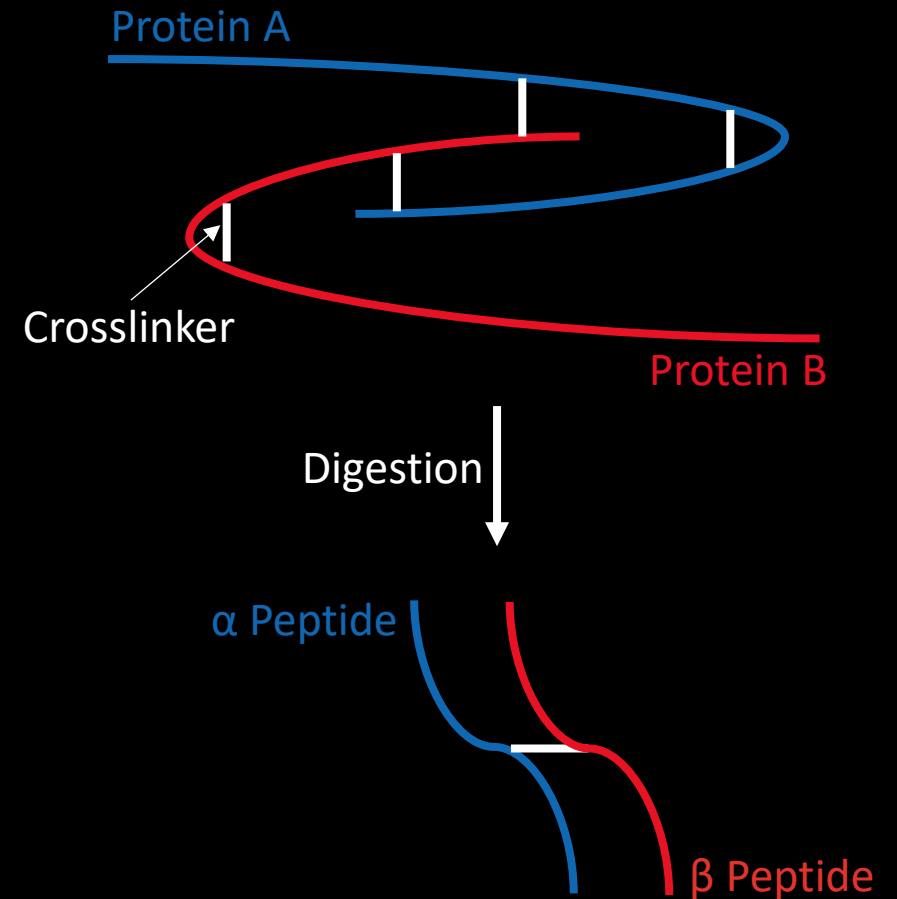
alternative title:

Proteome-wide Non-Cleavable Crosslink Identification Using
Sparse Matrix Multiplication with MS Annika 3.0

Micha Birklbauer, September 2024

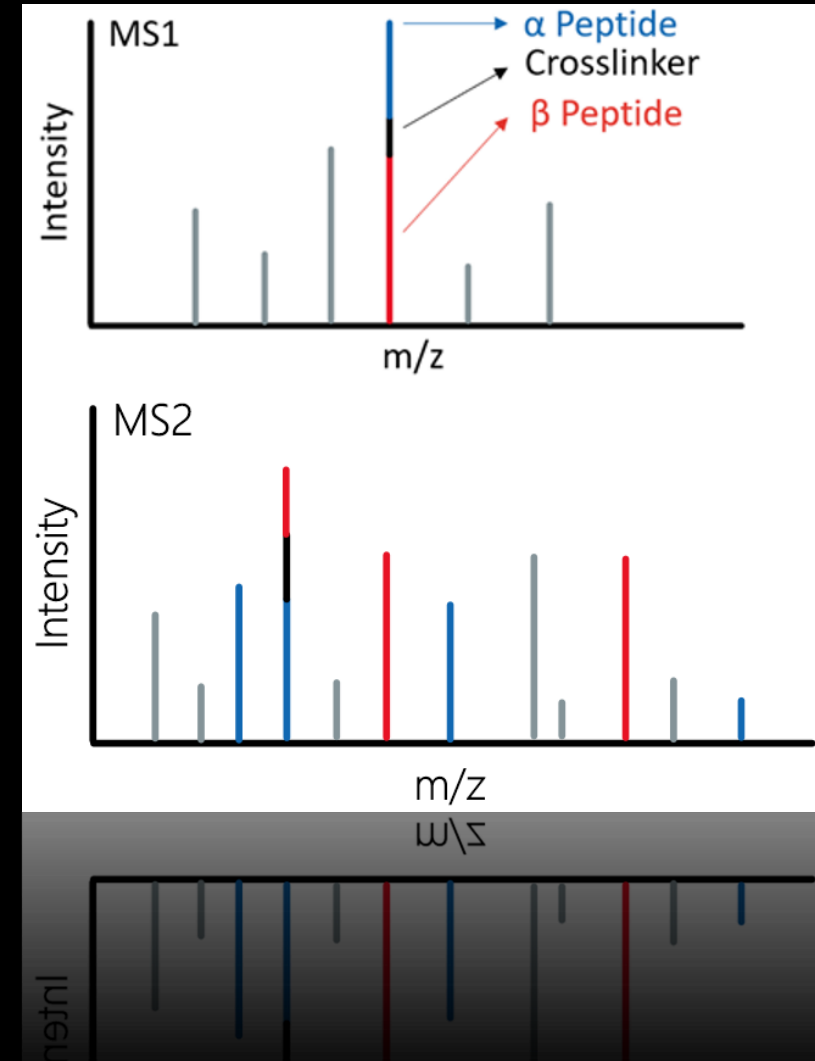
Introduction: Cross-linking

- **Crosslinker:** Small molecule that covalently links amino acids in or between proteins
- Digestion
- Smaller connected fragments
→ α Peptide and β Peptide
- Analyzed with mass spectrometry
- Used for structure analyses and to study protein-interaction networks



The Problem

- In non-cleavable cross-linking experiments we cannot infer the masses of the two individual peptides
- We know the mass of the complete cross-linked entity (precursor)
- For identification we need to consider any combination of two peptides that make up the precursor mass

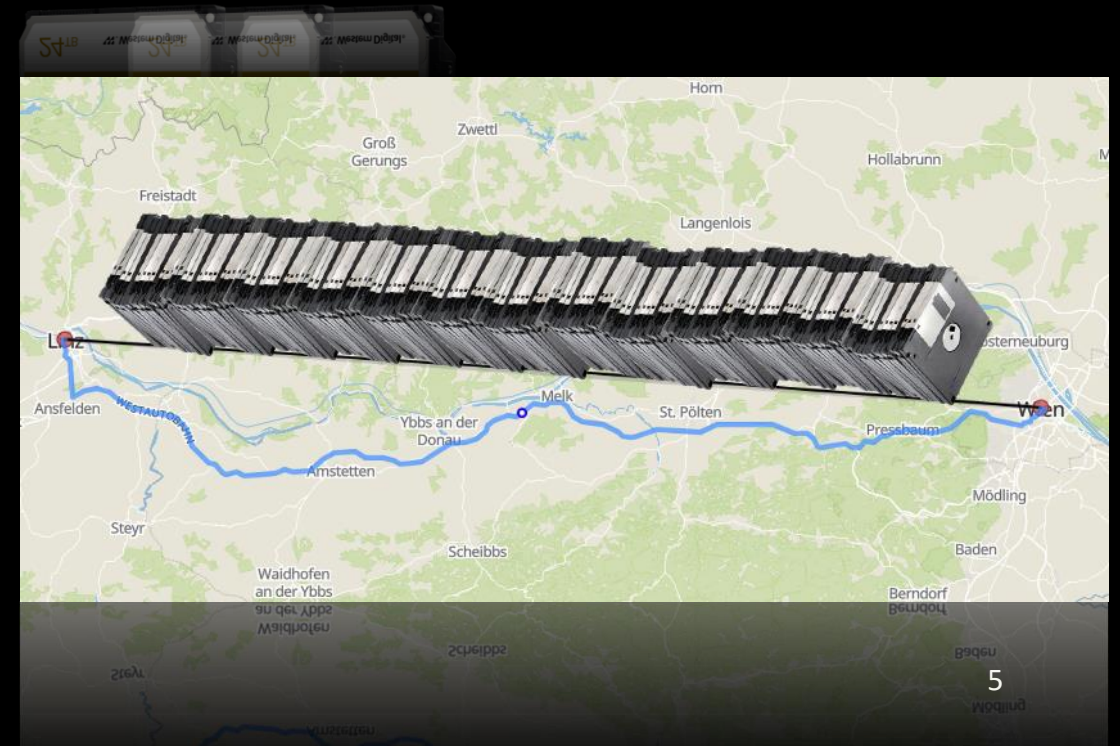


The n^2 Problem

- Considering every combination of peptides leads to very large search spaces
- Consider a database of n peptides, the number of possible combinations c would be:
 - $c = \binom{n}{2} + n = (n + 1) * \frac{n}{2}$
 - For big n this can be considered $O(n^2)$ complexity
- For the human proteome (SwissProt, 20 337 p), trypsin digest, maximum 3 missed cleavages, 5 – 30 peptide length:
 - $n = 2\,749\,058$
 - $c = 3\,778\,661\,318\,211$
 - one combination = 16 bytes
 - Total = ?

54.98676 TB

- This would roughly fit on three enterprise-level hard disks
- Or on enough 3.5" floppy disks to cover the distance from Linz to Vienna



Tackling the n^2 Problem

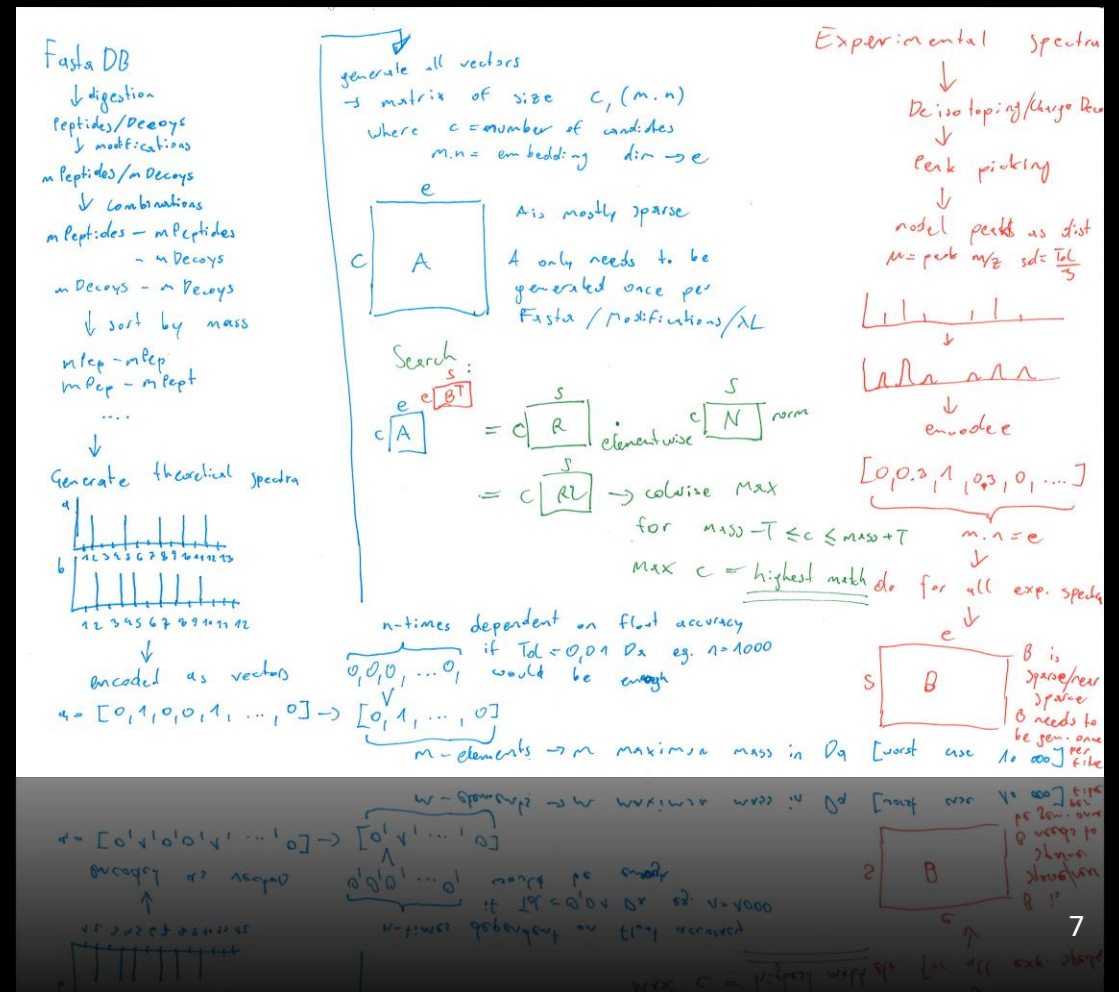
- Three different options:
 - Cry
 - Reducing the search space
 - Speeding up the search process so more combinations can be considered

source: giphy.com



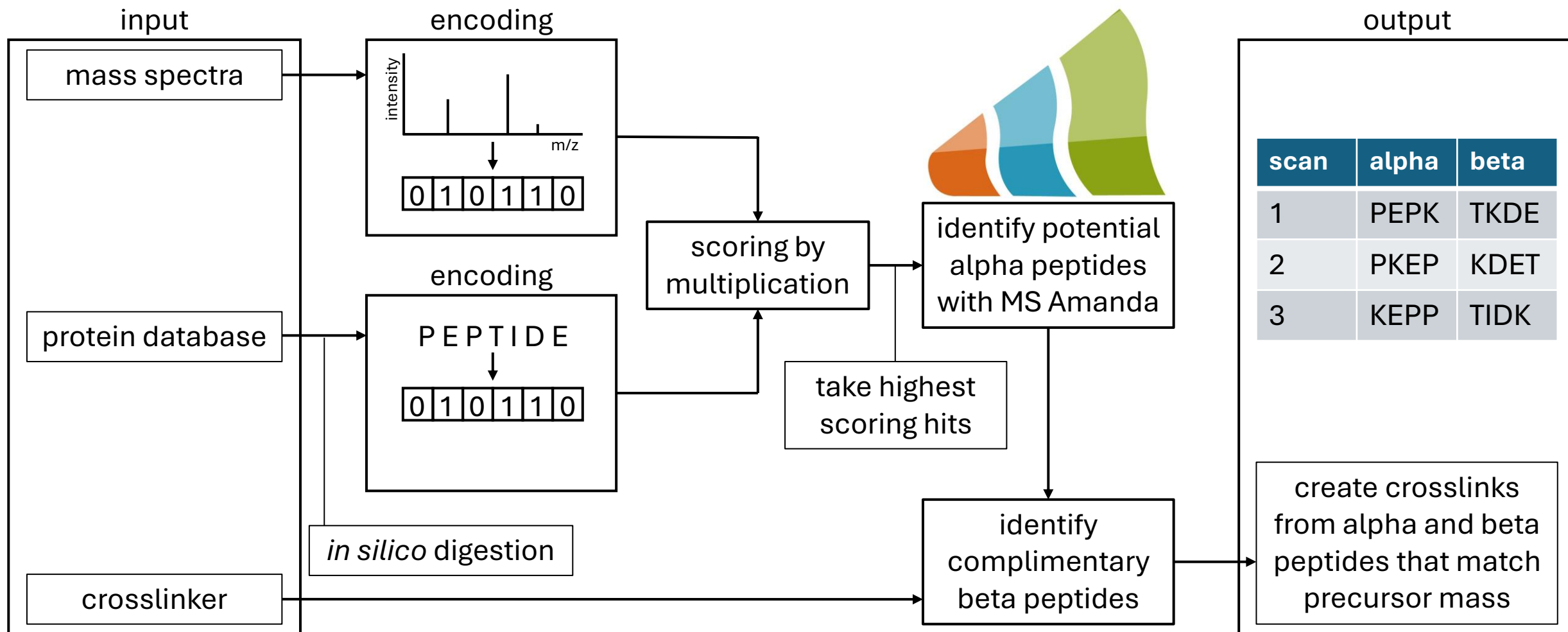
The Idea

- Reduction of search space:
 - Identify one of two peptides first
 - → Limits the number of combinations to consider
- Identification of the peptide by a fast approximate search:
 - Encoding of peptides and spectra as sparse vectors
 - Scoring purely based on addition and multiplication
 - Adhering to the SIMD paradigm



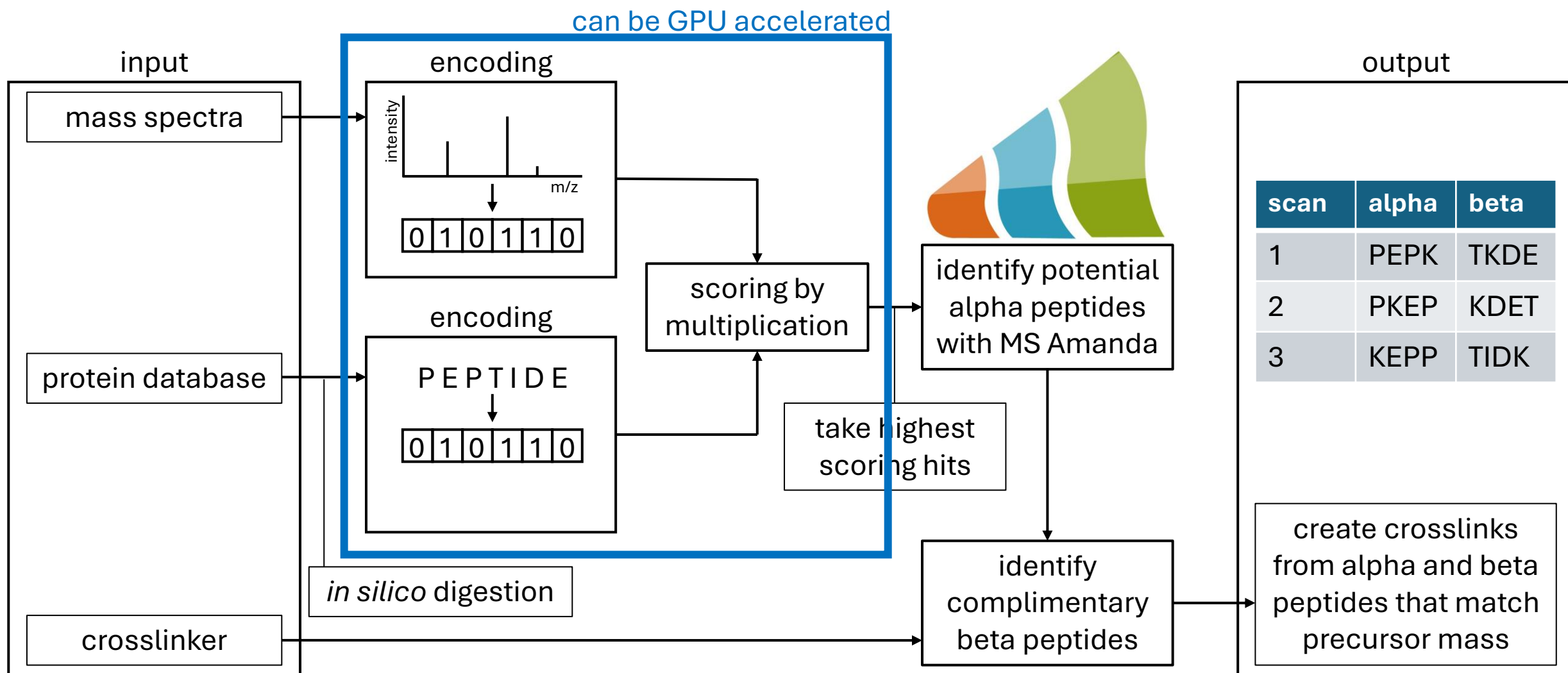


MS Annika 3.0





MS Annika 3.0



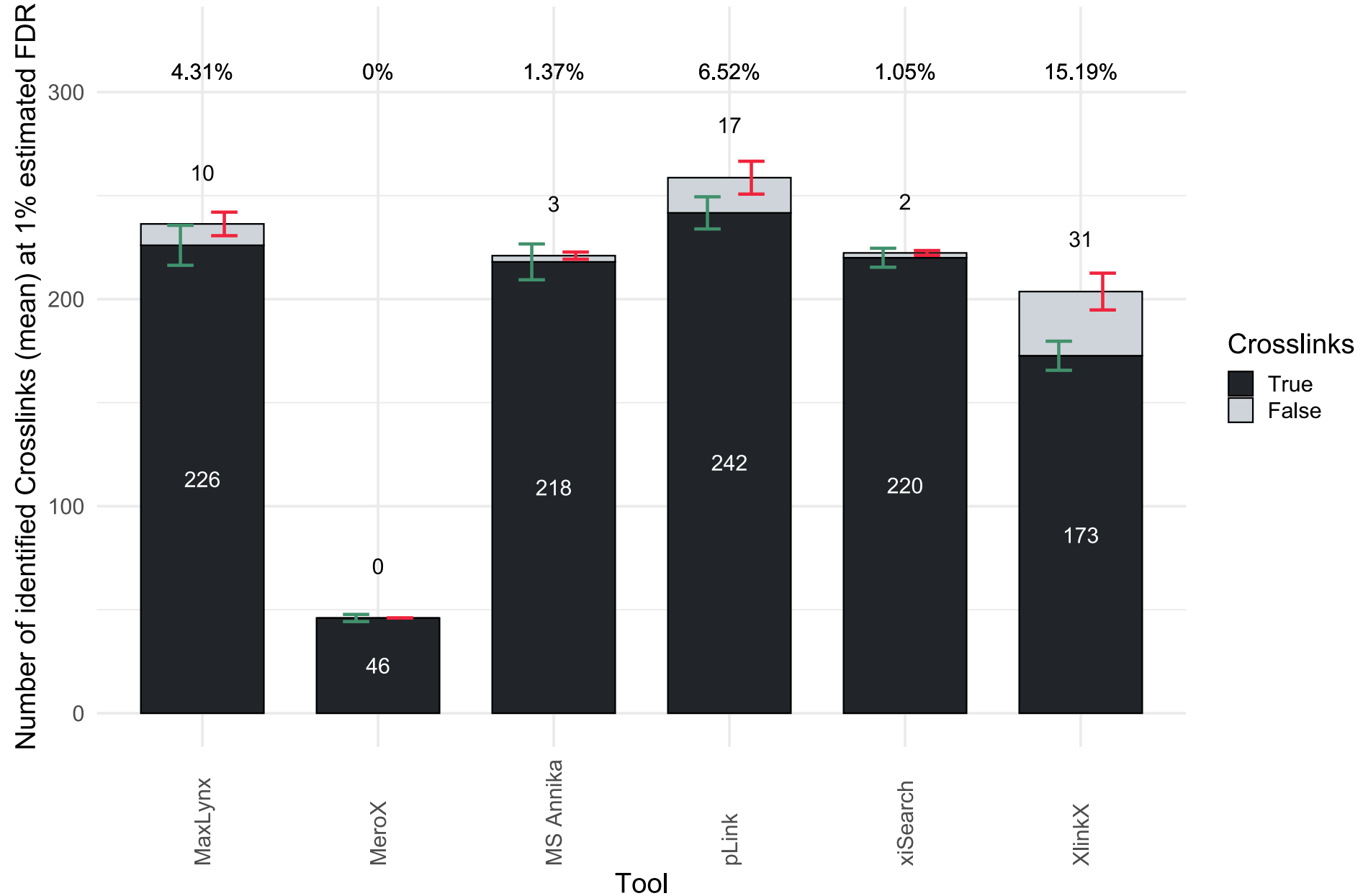
The End Product

Quite literally...

Comparison of MS Annika 3.0 to Other Search Engines

- Dataset by Beveridge *et al.* → PXD014337
- Synthetic peptides cross-linked with DSS
 - allows calculation of “true” FDR
 - allows assessment of which crosslinks are true positive identifications and which are false positive identifications
- Comparison against MaxLynx (MaxQuant), MeroX, pLink, xiSearch and XlinkX
- **Goal:** Identify as many true positive crosslinks as possible while staying close to 1% FDR

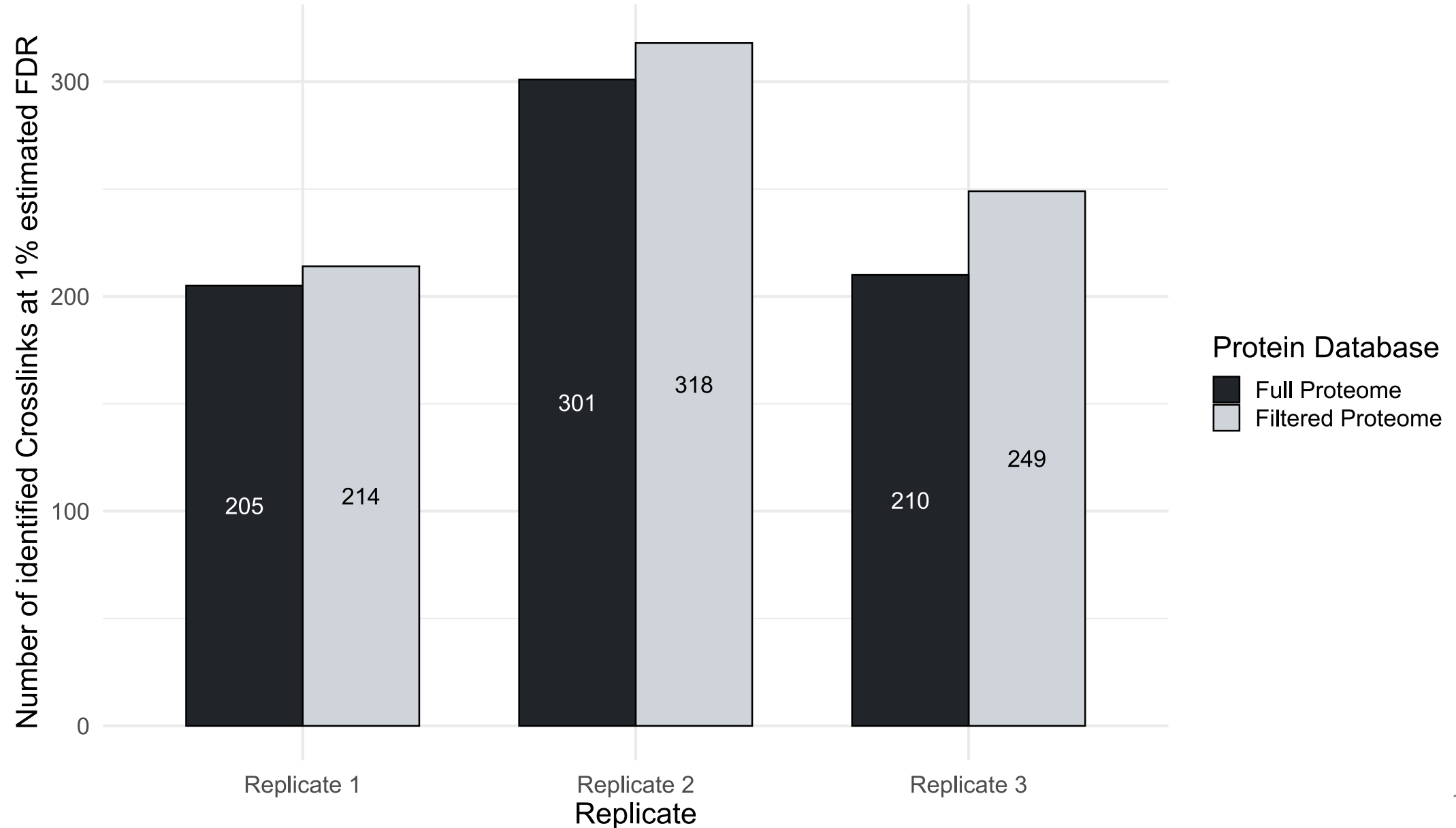
Dataset of synthetic peptides by Beveridge et al., 2020:
 Number of identified crosslinks per tool at 1% estimated FDR
 (3 replicates, crosslinker: DSS)

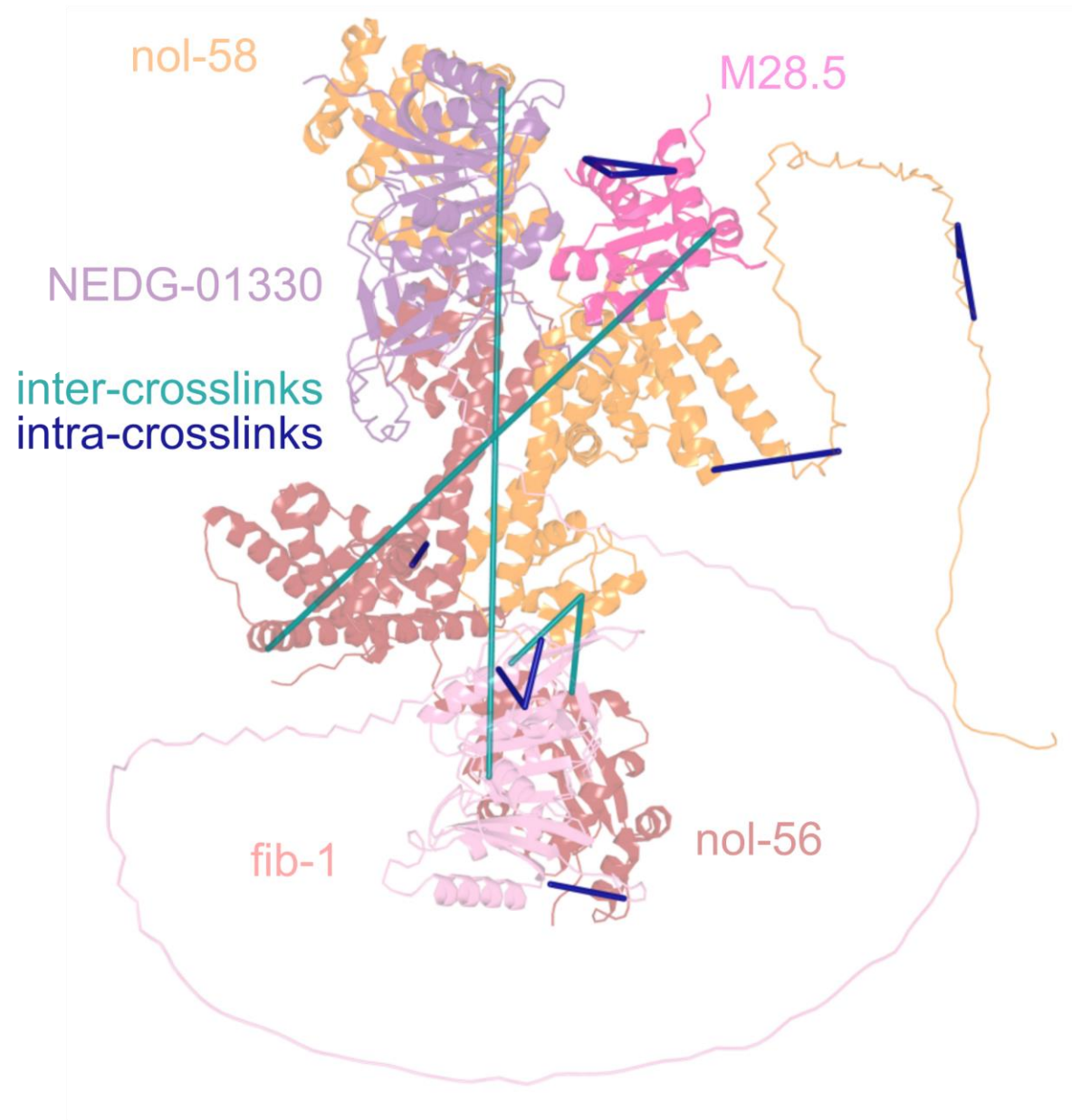


The Real Test: Large Scale Proteome-Wide Crosslink Identification

- Mass spectra from *C. elegans* nuclei samples
- Samples cross-linked with DSG
- Comparison of two searches:
 - Searched against database of most abundant proteins ($n \approx 3000$)
 - Searched against the full *C. elegans* proteome using all sequences available in UniProt ($n \approx 26\,000$)
- Validated for 1% FDR with xiFDR

Dataset of *C. elegans* nuclei by Müller et al., 2024:
Number of identified crosslinks per replicate at 1% estimated FDR
(3 replicates, crosslinker: DSG)





AlphaLink2 ipTM score: 0.721

Refined structure of
the Box C/D RNP
complex in *C. elegans*

Conclusions

- MS Annika 3.0 can successfully tackle proteome-wide non-cleavable crosslink studies
- The implemented algorithm is super efficient, allowing proteome-wide searches on commodity hardware
- This will allow researchers to:
 - Perform non-cleavable crosslink studies of complex samples that were previously unfeasible
 - Re-analyze published crosslink data with bigger protein databases, potentially uncovering new biological insights
- The algorithm presents a transferable solution for big search space problems

Acknowledgements

- Fränze Müller, Sowmya Geetha, Manuel Matzinger & Karl Mechtler
- Viktoria Dorfer
- FWF
- The Eigen developers & community
- My research group 😊



Discussion

Pre-Print:



MS Annika:

github.com/hgb-bin-proteomics/MSAnnika

Contact:

micha.birklbauer@fh-hagenberg.at
github.com/michabirklbauer
michabirklbauer.me

source: giphy.com



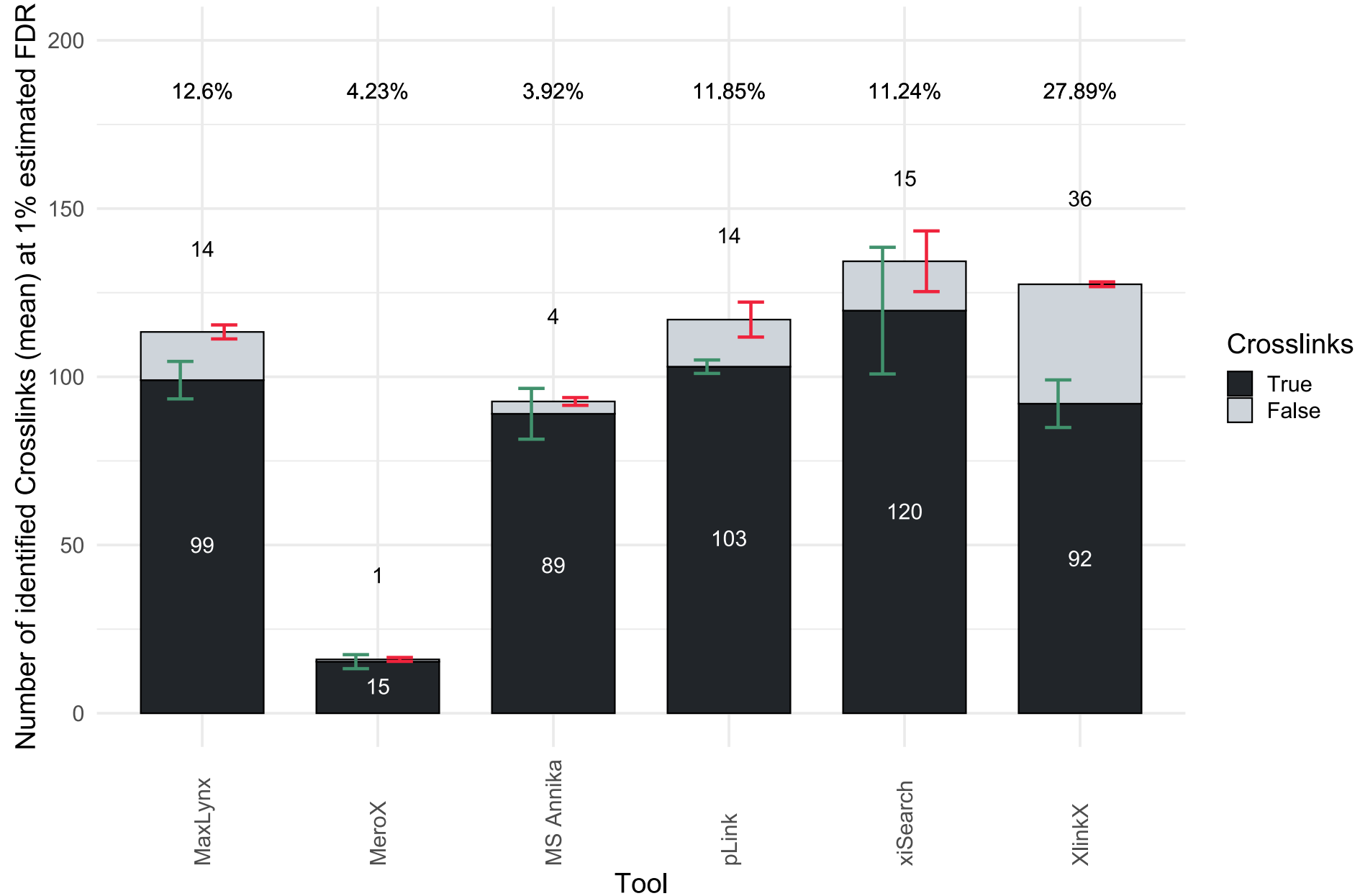
Appendix A

More Comparisons

Comparison of MS Annika 3.0 to Other Search Engines

- Dataset by Matzinger *et al.* → PXD029252
- Synthetic peptides cross-linked with ADH
 - allows calculation of “true” FDR
 - allows assessment of which crosslinks are true positive identifications and which are false positive identifications
- Comparison against MaxLynx (MaxQuant), MeroX, pLink, xiSearch and XlinkX
- **Goal:** Identify as many true positive crosslinks as possible while staying close to 1% FDR

Dataset of synthetic peptides by Matzinger et al., 2022:
 Number of identified crosslinks per tool at 1% estimated FDR
 (3 replicates, crosslinker: ADH)

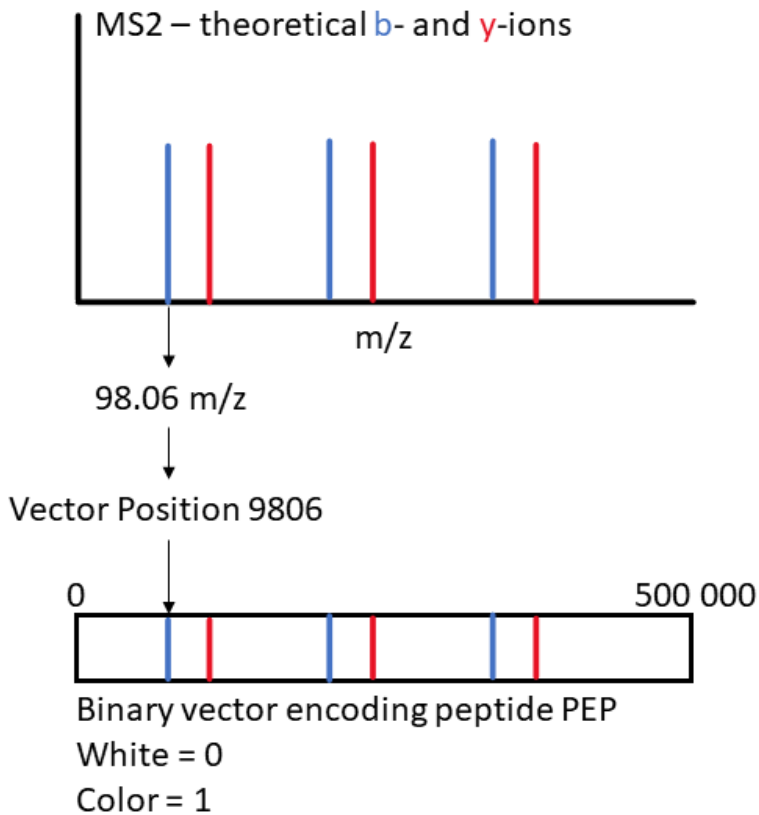


Appendix B

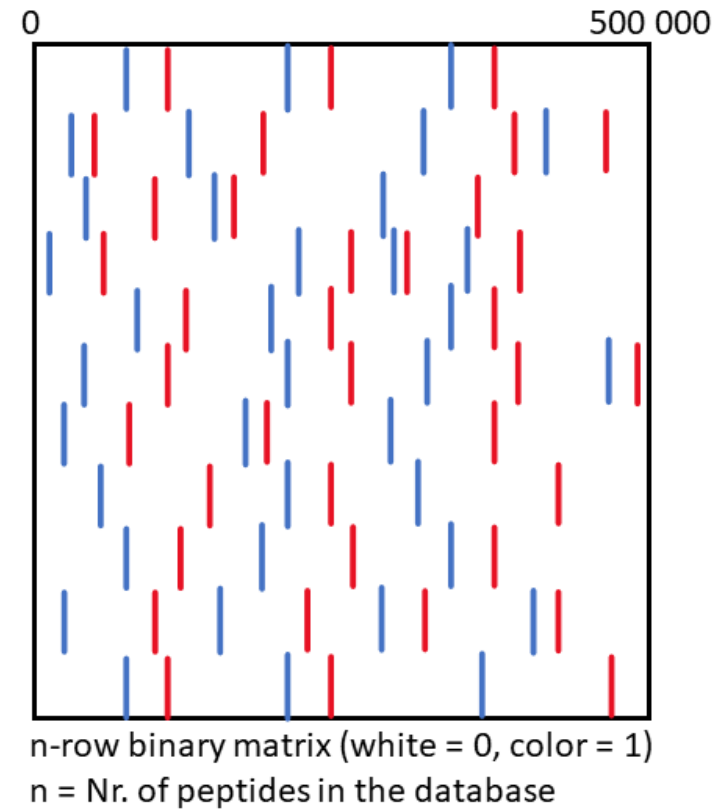
Matrix Encoding

Vectorizing example

Peptide: **PEP**

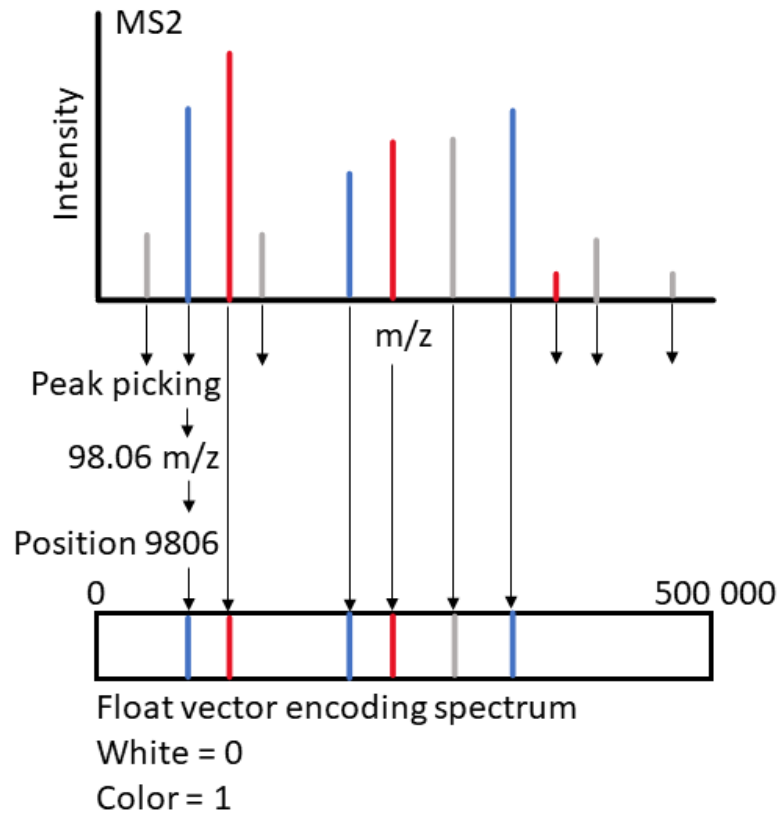


Do for every peptide in database:

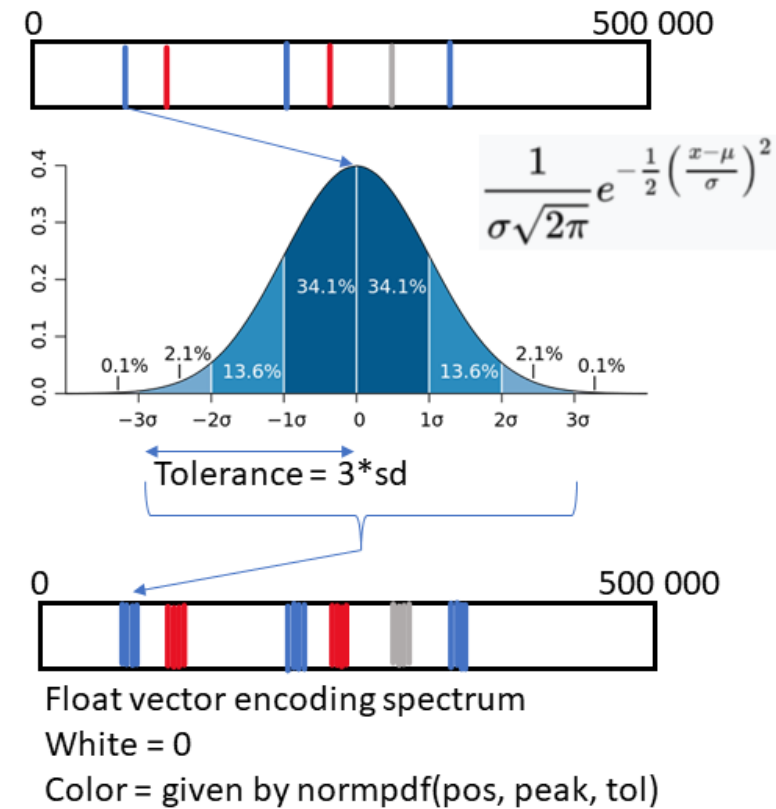


Vectorizing example

Experimental spectrum:



Considering instrument tolerance:



Vectorizing example - search

Peptide: **PEP**



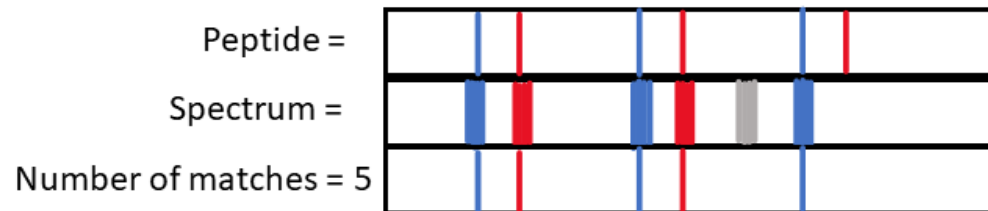
Binary vector encoding peptide PEP

Spectrum:



Float vector encoding spectrum

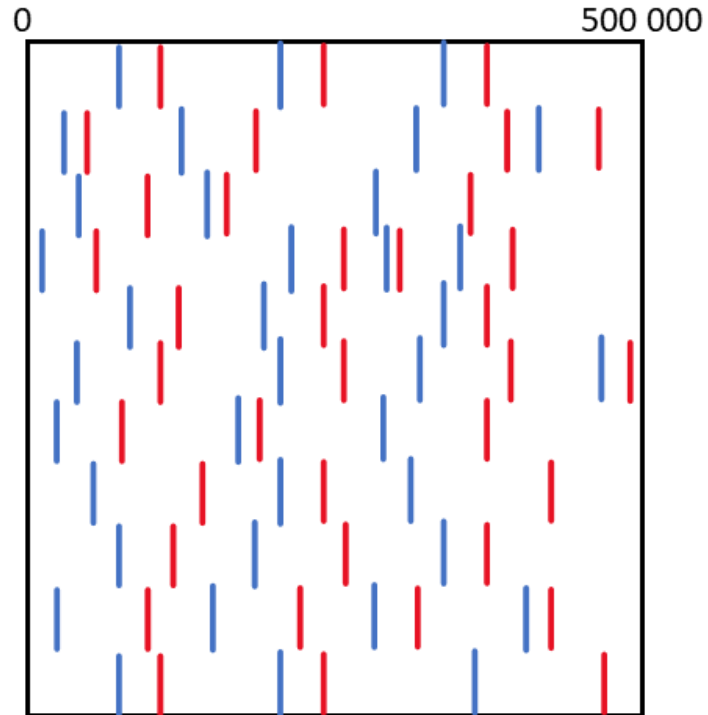
Similarity is given by the dot product of the two vectors
= the sum of overlapping positions
= the number of peaks that match



(Normalization = Number of matches / Number of ions of PEP = $5/6 = 0.8$)

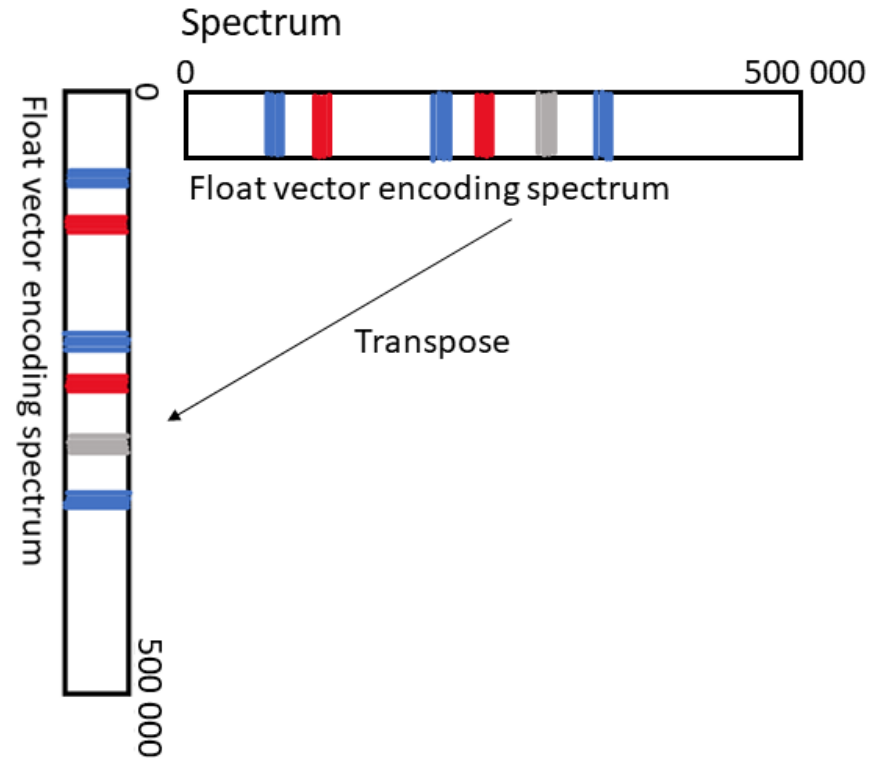
Vectorizing example - search

All peptides:



n-row binary matrix (white = 0, color = 1)
n = Nr. of peptides in the database

X

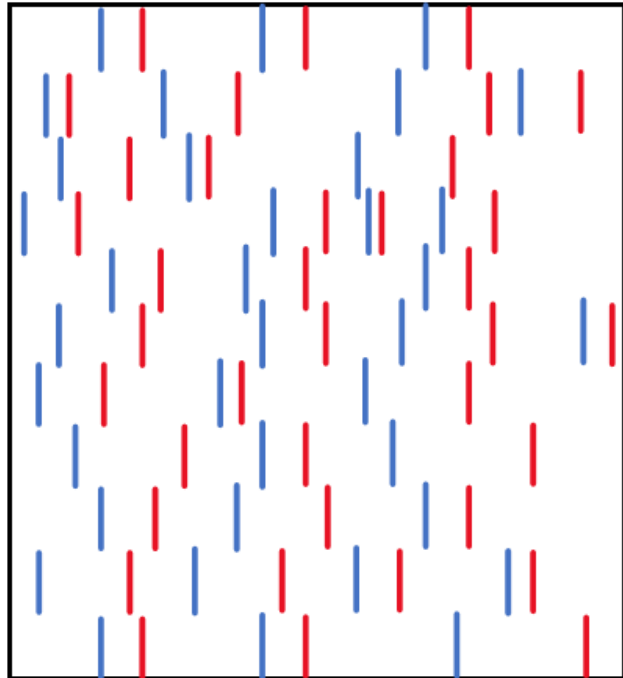


Similarity is given by the matrix product

Vectorizing example - search

All peptides:

0 500 000



n-row binary matrix (white = 0, color = 1)
n = Nr. of peptides in the database

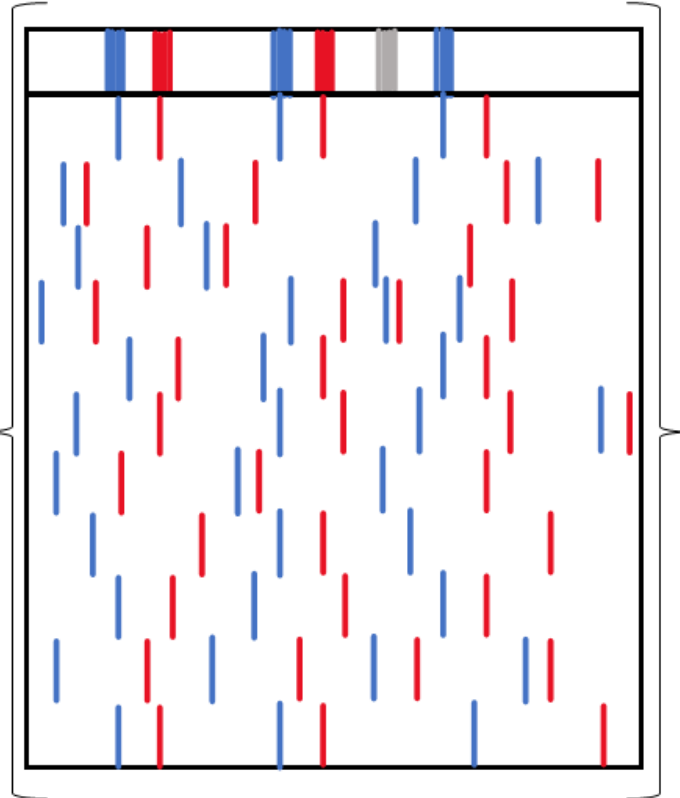
\times

Float vector encoding spectrum

0

500 000

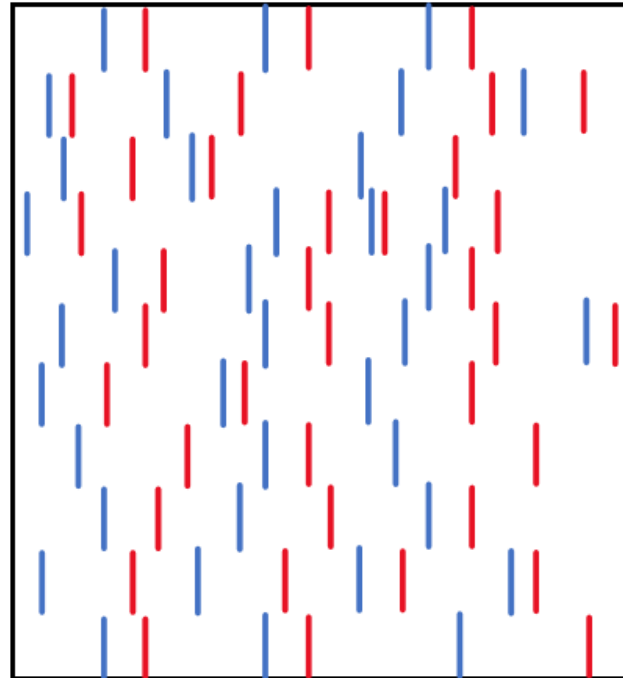
$=$



Vectorizing example - search

All peptides:

0 500 000



n-row binary matrix (white = 0, color = 1)
n = Nr. of peptides in the database

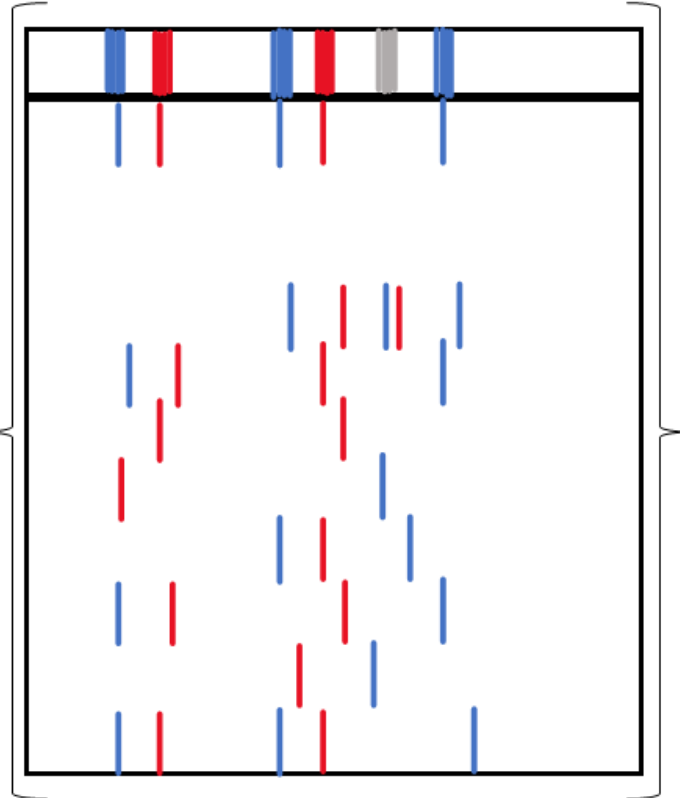
X

Float vector encoding spectrum

0

500 000

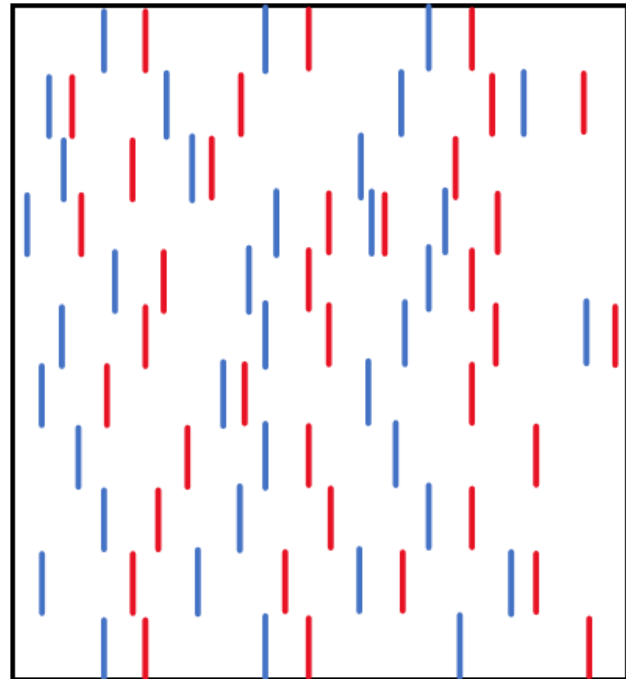
=



Vectorizing example - search

All peptides:

0 500 000



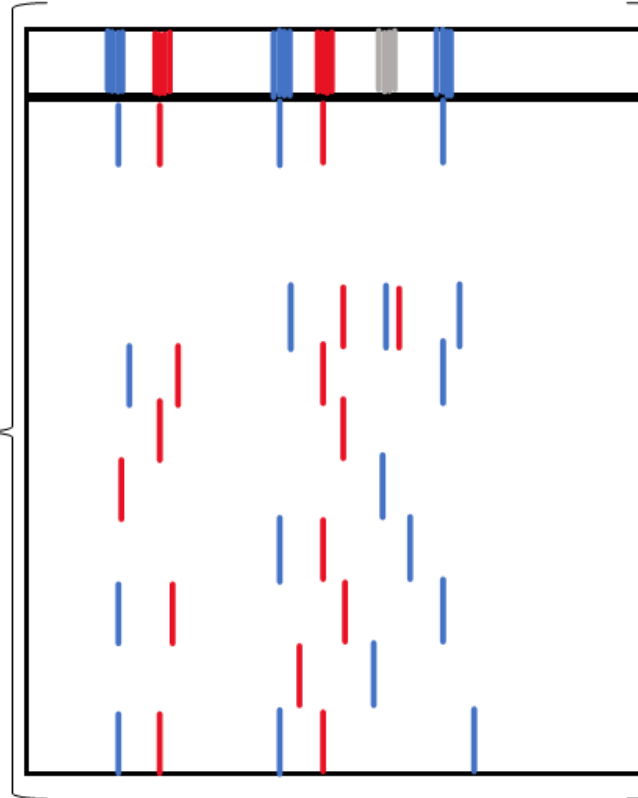
n-row binary matrix (white = 0, color = 1)
n = Nr. of peptides in the database

\times

Float vector encoding spectrum

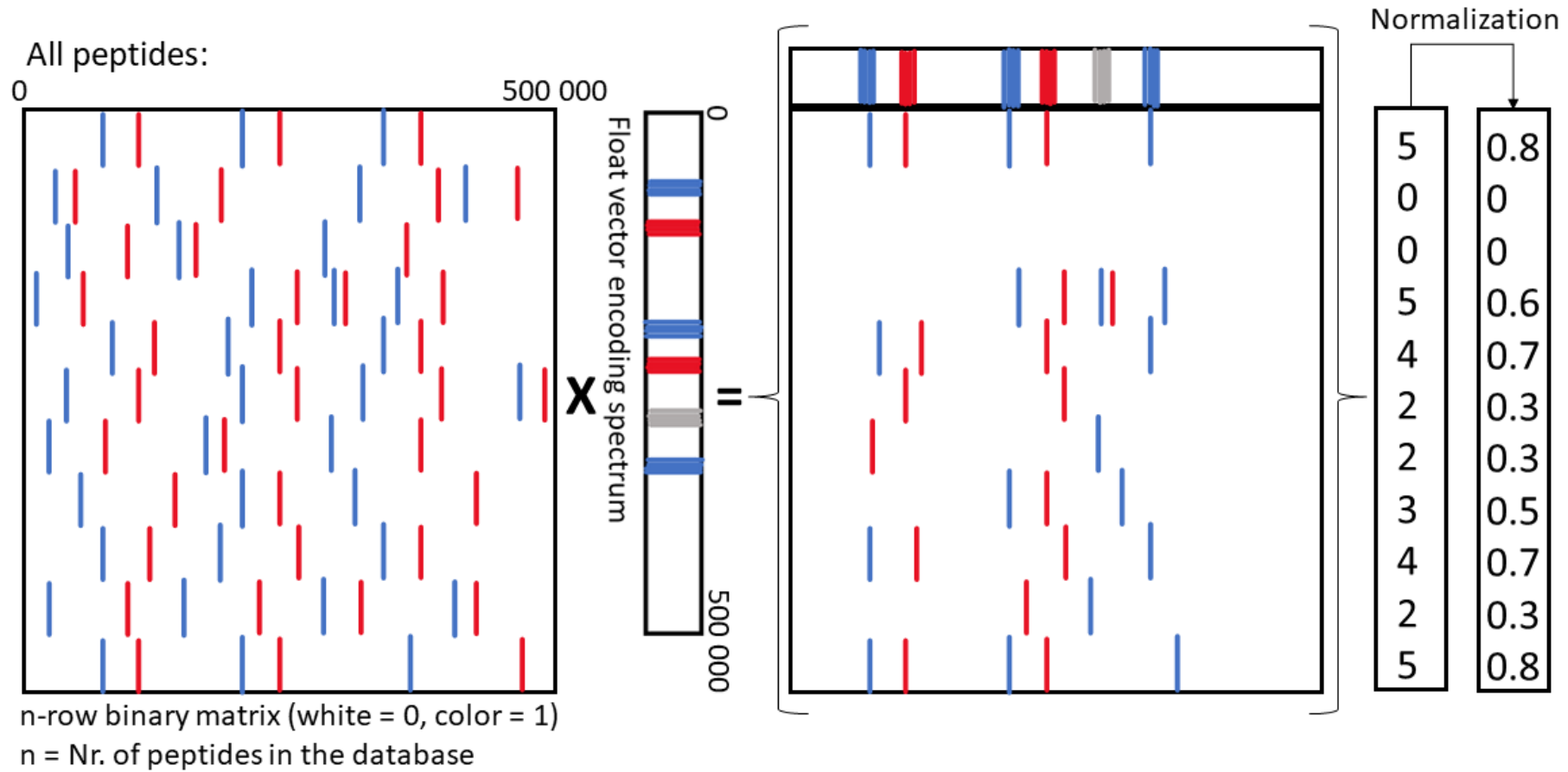


$=$

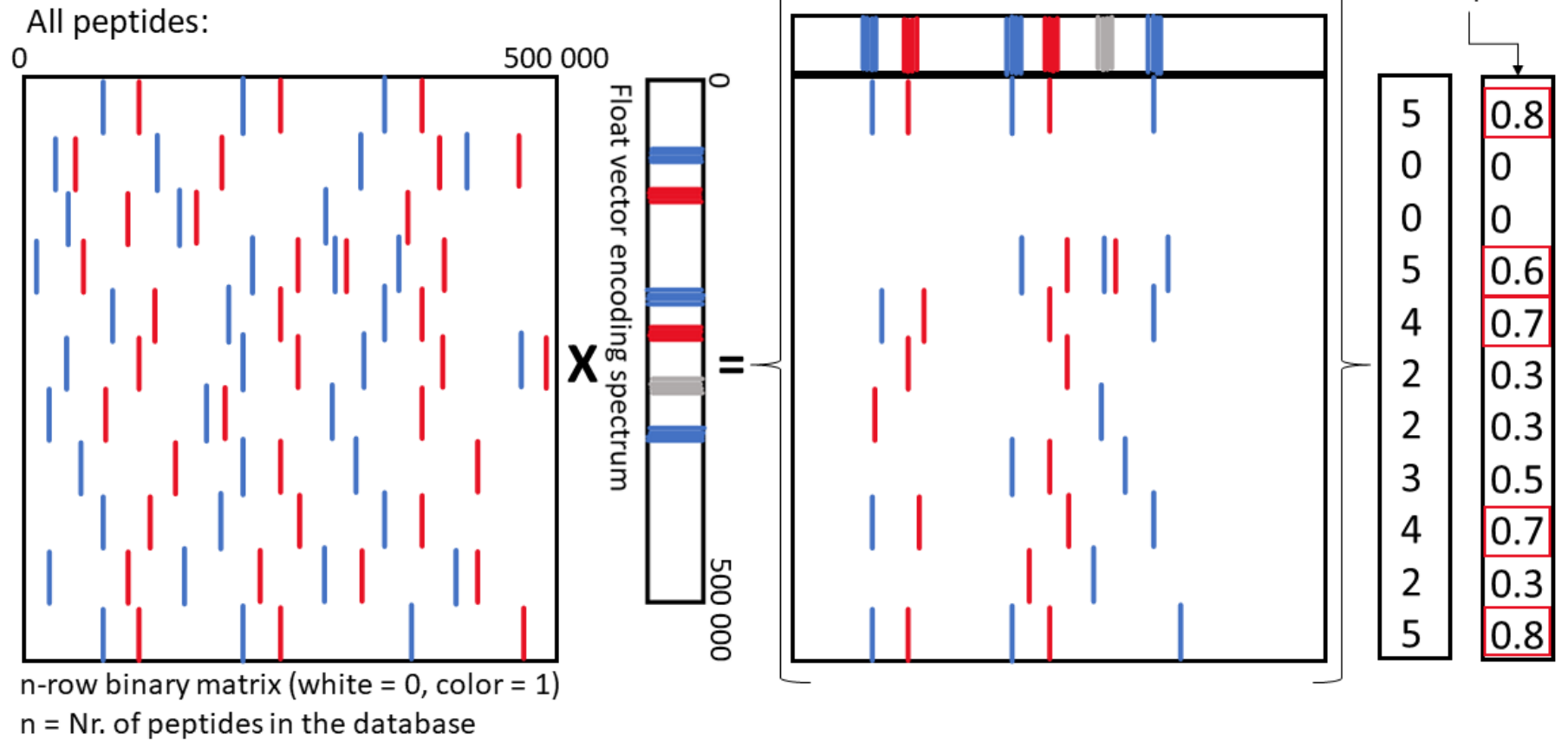


5
0
0
5
4
2
2
3
4
2
5

Vectorizing example - search



Vectorizing example - search



Vectorizing example - search

Take top n = 5



Generate all combinations of the peptide with peptides that make up the precursor mass

Pool all candidate combinations



Exact search with MS Amanda