Peptide detectability prediction to improve protein identification in mass spectrometry using machine learning

By Anima Sutradhar

MSc Bioinformatics, QMUL SBCS
Supervisor: Professor Conrad Bessant
Co-supervisors: Esteban Gea & Hajar Saihi

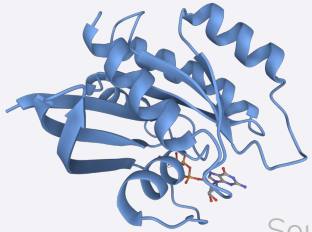
What is proteomics?

The promise of proteomics

Proteomics is the study of the complete set of proteins produced by the genome at any one time.

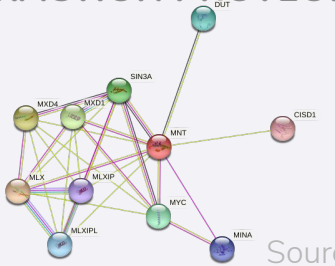
PROTEOMICS GOAL: identify and quantify all different proteins in a sample.

STRUCTURAL PROTEOMICS



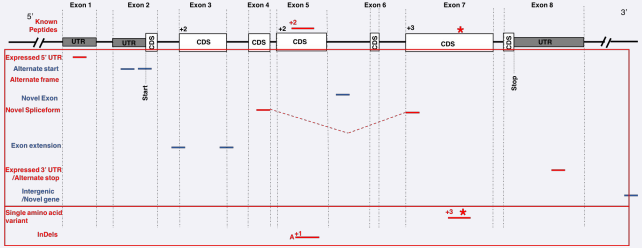
Source: PDB ID 5p21

INTERACTION PROTEOMICS



Source: string-db.org

PROTEOGENOMICS

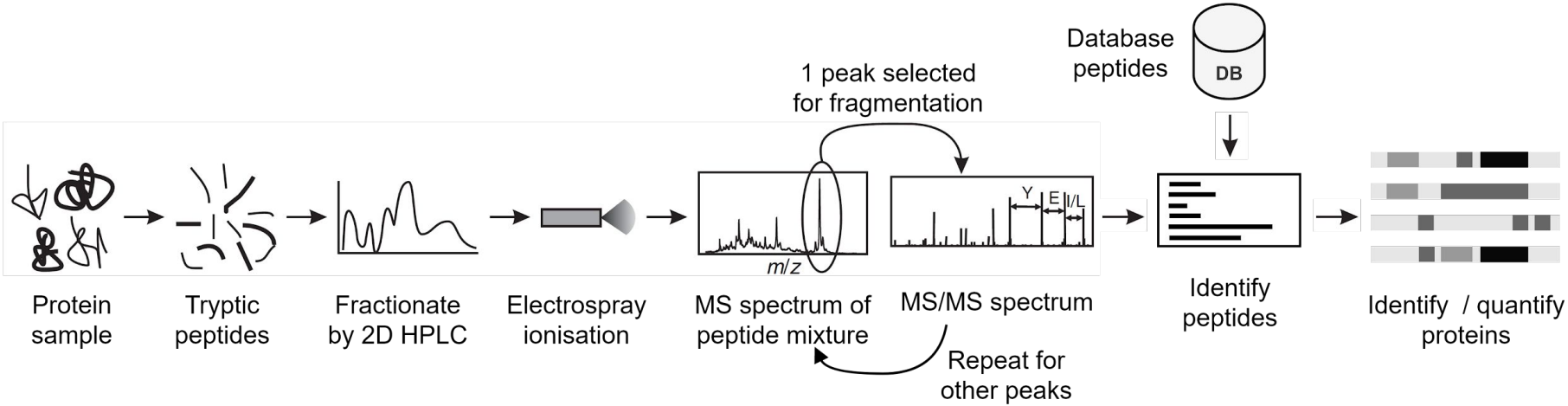


Source: galaxyproject.github.io

How can proteins be identified?

Standard shotgun proteomics workflow

PROTEOMICS GOAL: identify and quantify all different proteins in a sample.

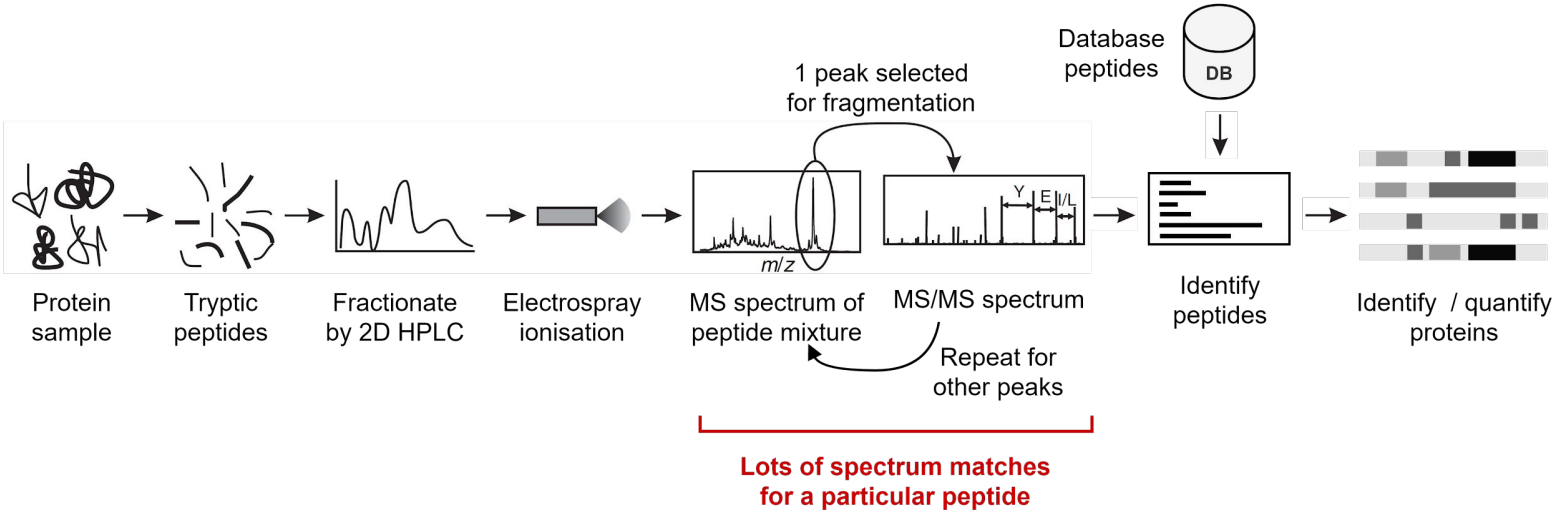


Adapted from Lu et al., Nature (2006).

Where does peptide detectability come into this?

The concept of peptide detectability

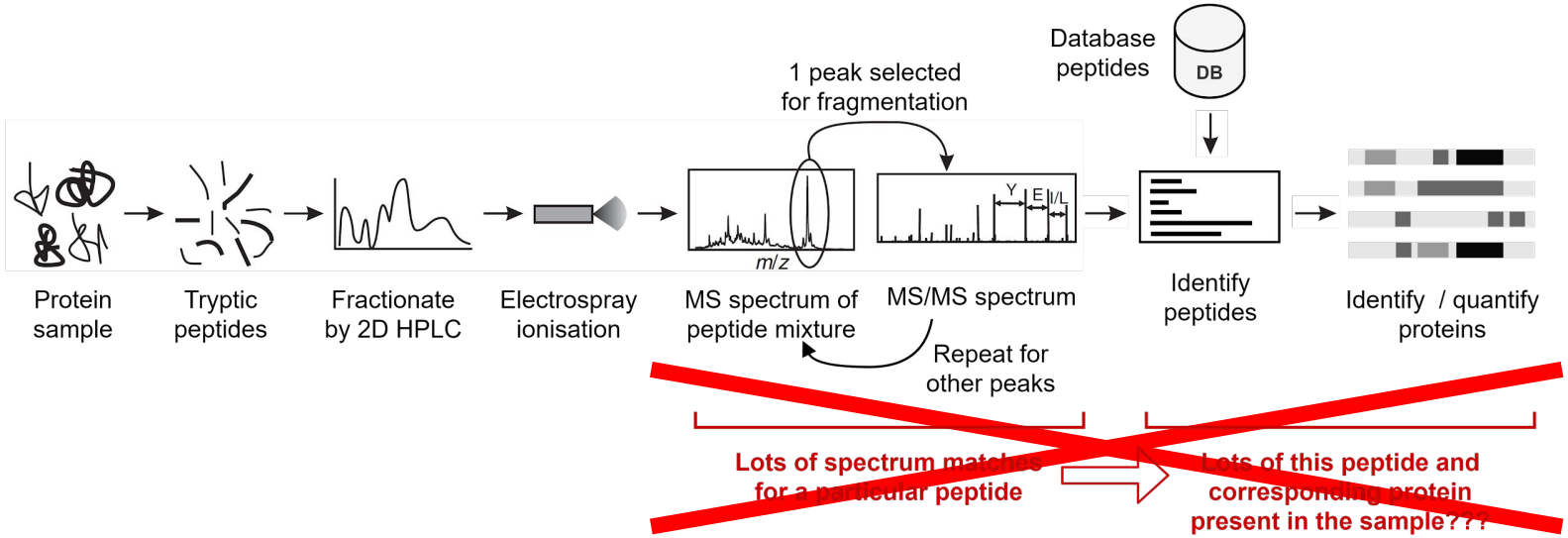
PROTEOMICS GOAL: identify and quantify all different proteins in a sample.



Adapted from Lu et al., Nature (2006).

The concept of peptide detectability

PROTEOMICS GOAL: identify and quantify all different proteins in a sample.



Adapted from Lu et al., Nature (2006).

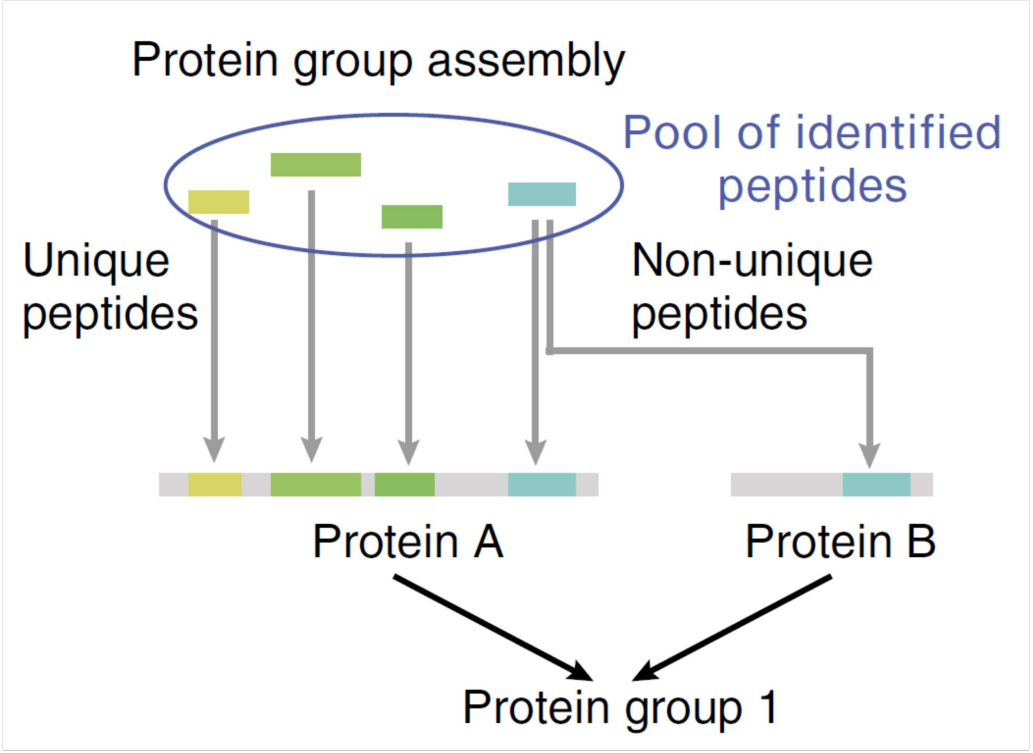
The concept of peptide detectability

PROTEOMICS GOAL: identify and quantify all different proteins in a sample.

TWO MAJOR ISSUES:

1) **Variability in sample runs**: you can run the exact same experiments, with the exact same samples and conditions → and still identify different peptides and their corresponding proteins in different quantities.
2) **Majority peptides go undetected** (50-90% spectra unassigned).

The concept of peptide detectability

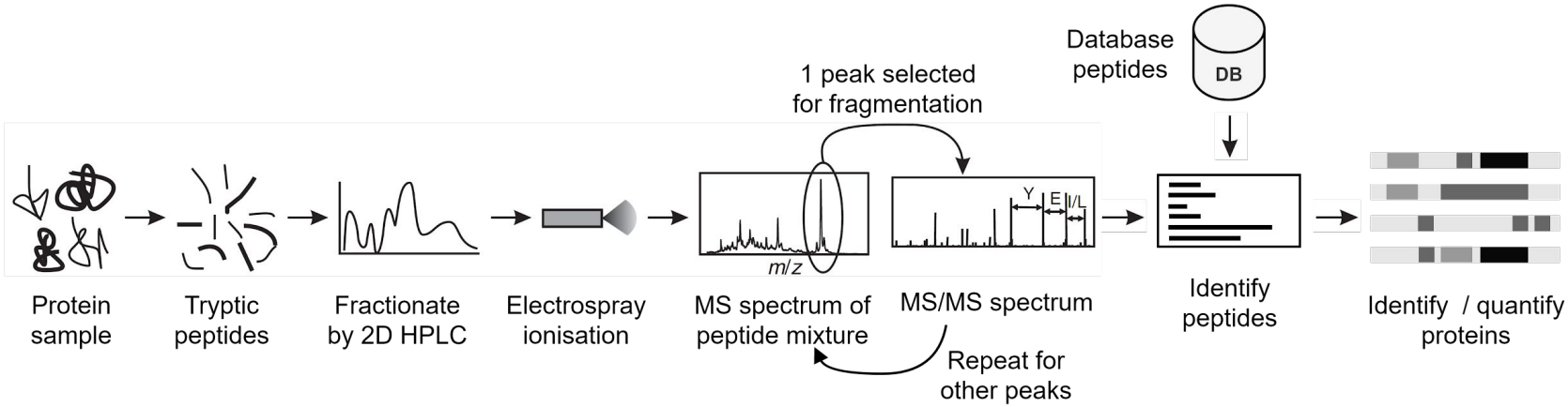


We can't assume the number of identified peptides correlates to the number of corresponding proteins in that sample.

Tyanova, Temu & Cox, Nature Protocols (2016).

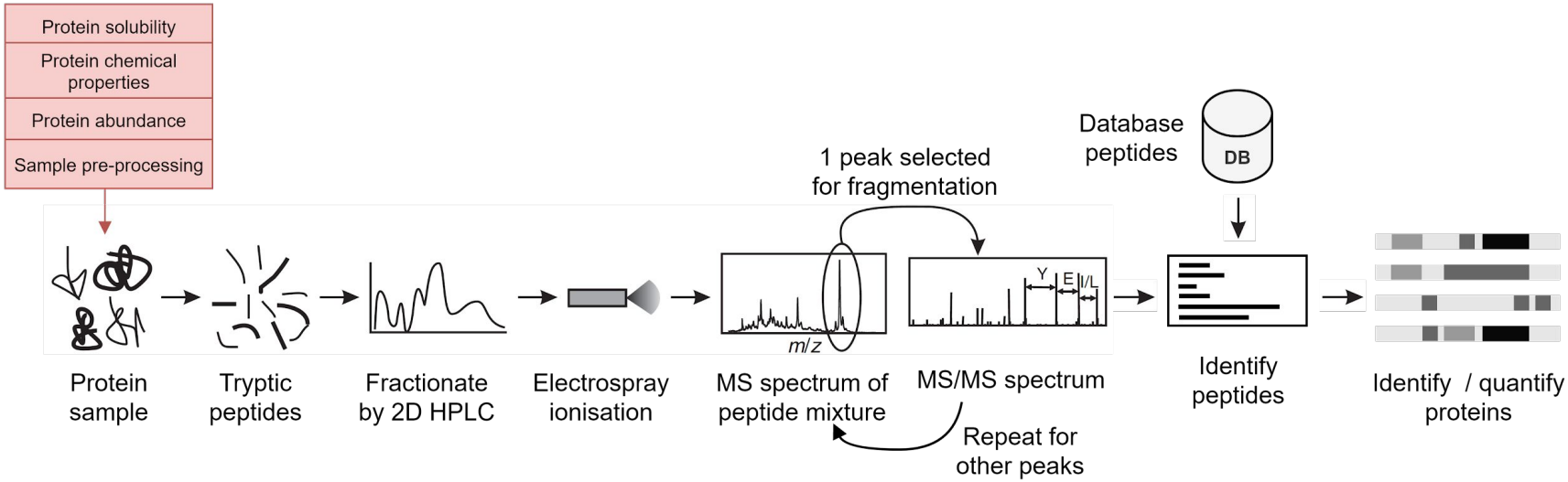
So what's causing this variability and so many peptides to go undetected?

Factors influencing peptide detectability



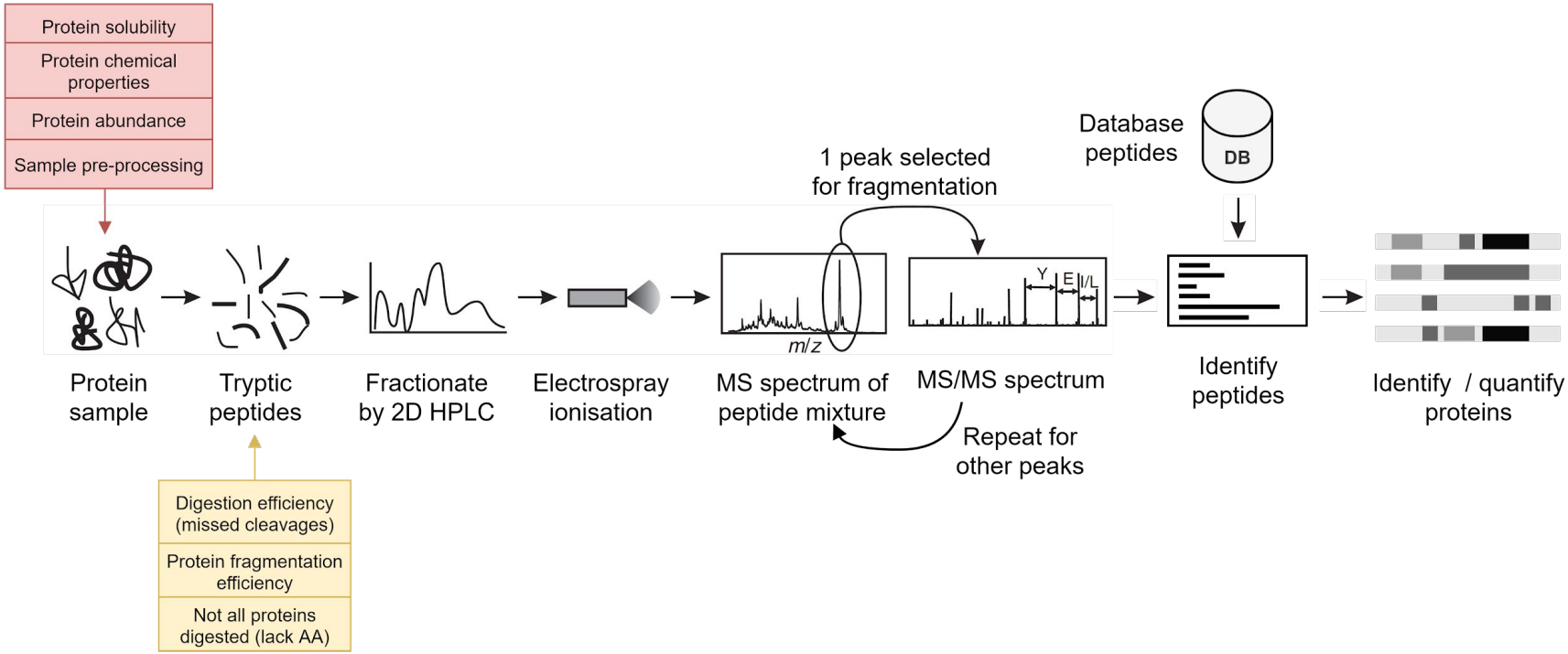
Adapted from Lu et al., Nature (2006).

Factors influencing peptide detectability



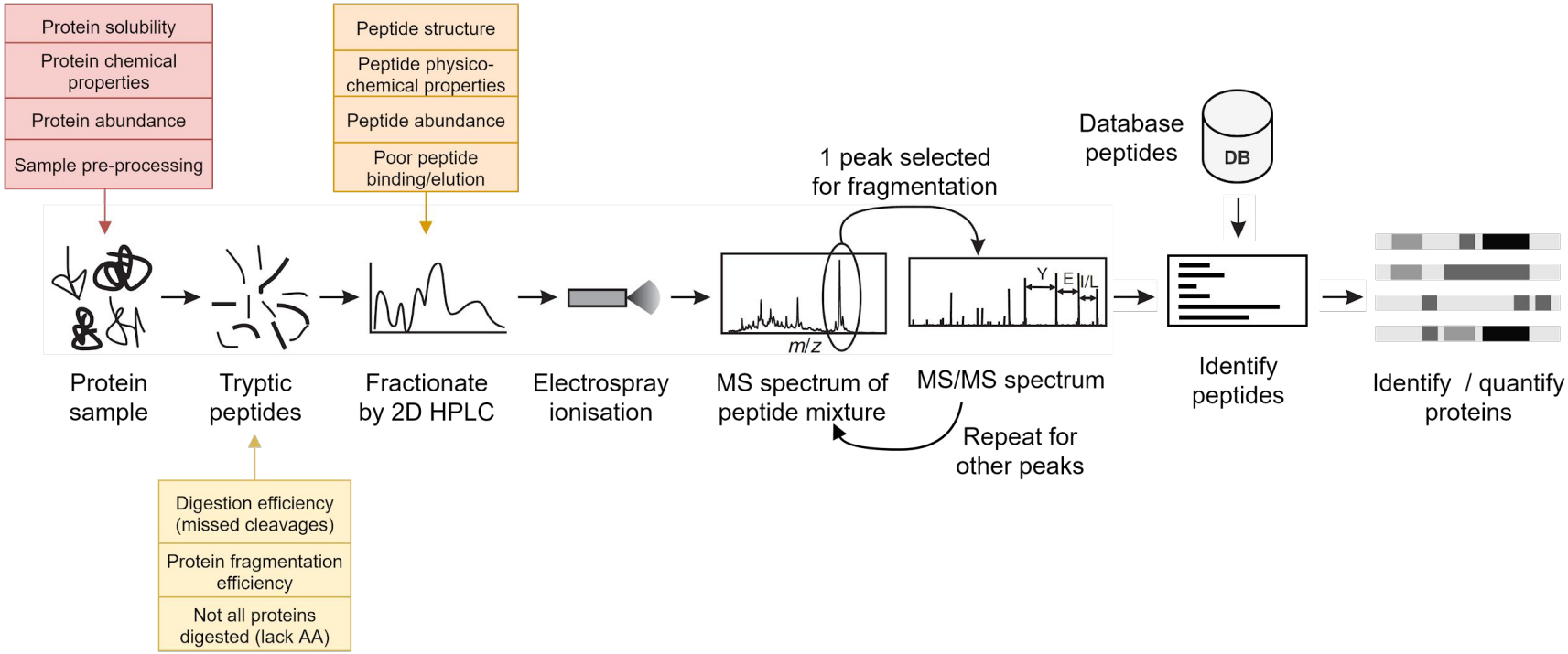
Adapted from Lu et al., Nature (2006).

Factors influencing peptide detectability



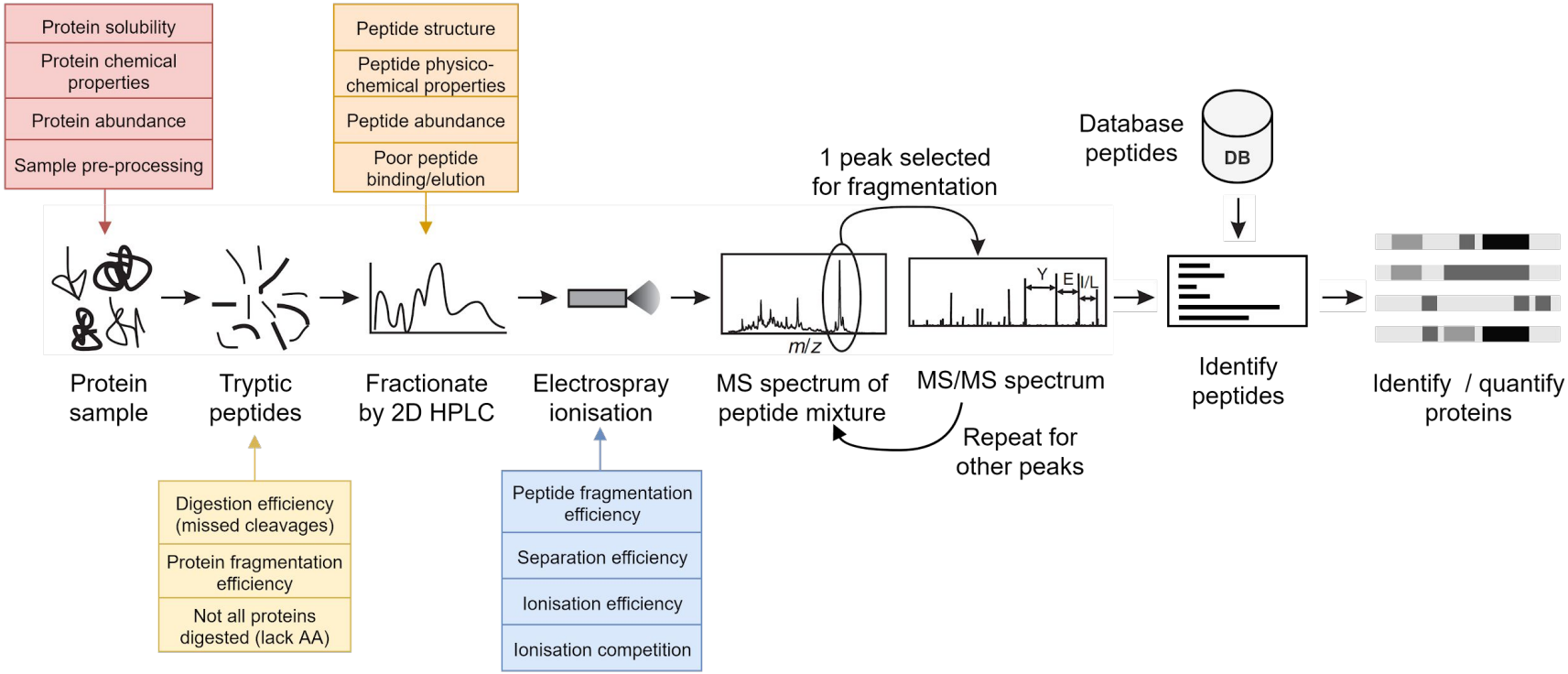
Adapted from Lu et al., Nature (2006).

Factors influencing peptide detectability



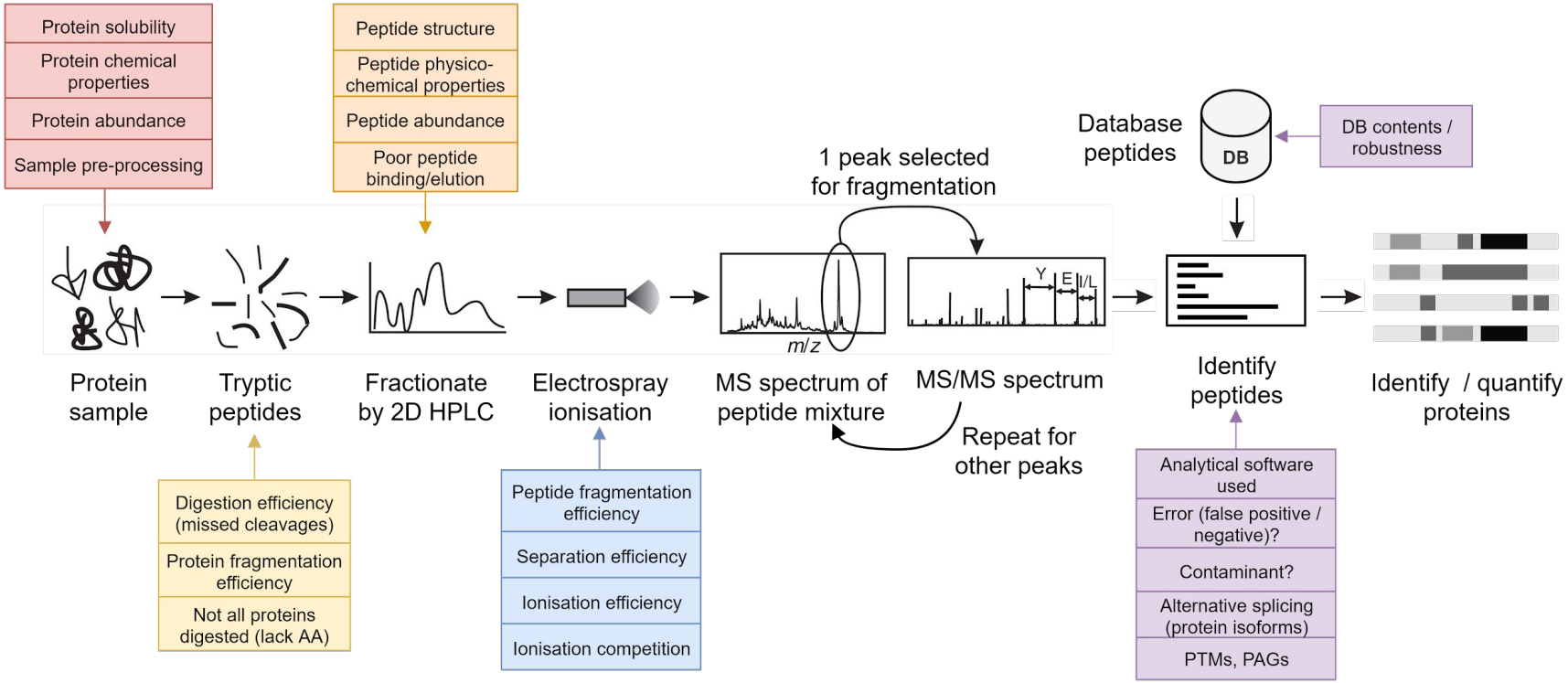
Adapted from Lu et al., Nature (2006).

Factors influencing peptide detectability



Adapted from Lu et al., Nature (2006).

Factors influencing peptide detectability

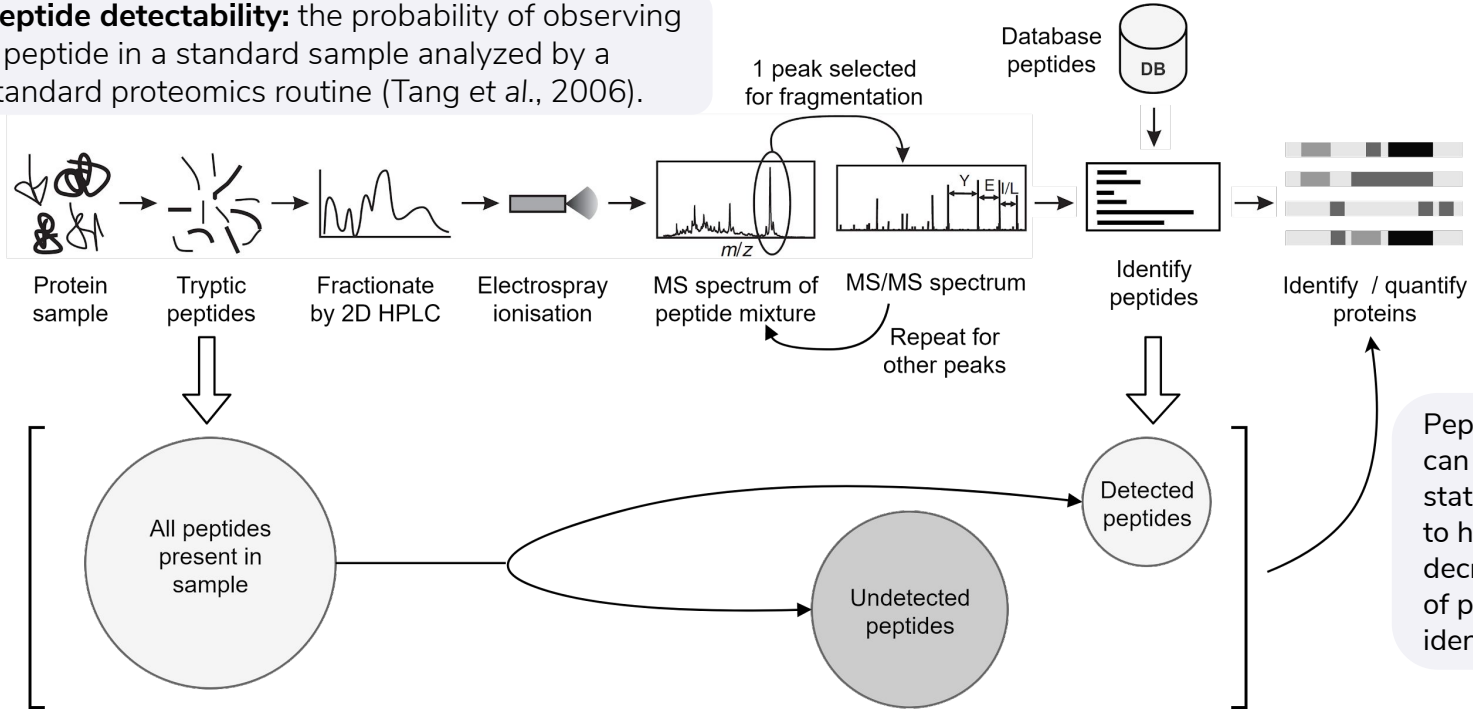


Adapted from Lu et al., Nature (2006).

How can peptide detectability help in protein identification and quantitation?

Integrating peptide detectability in protein identification

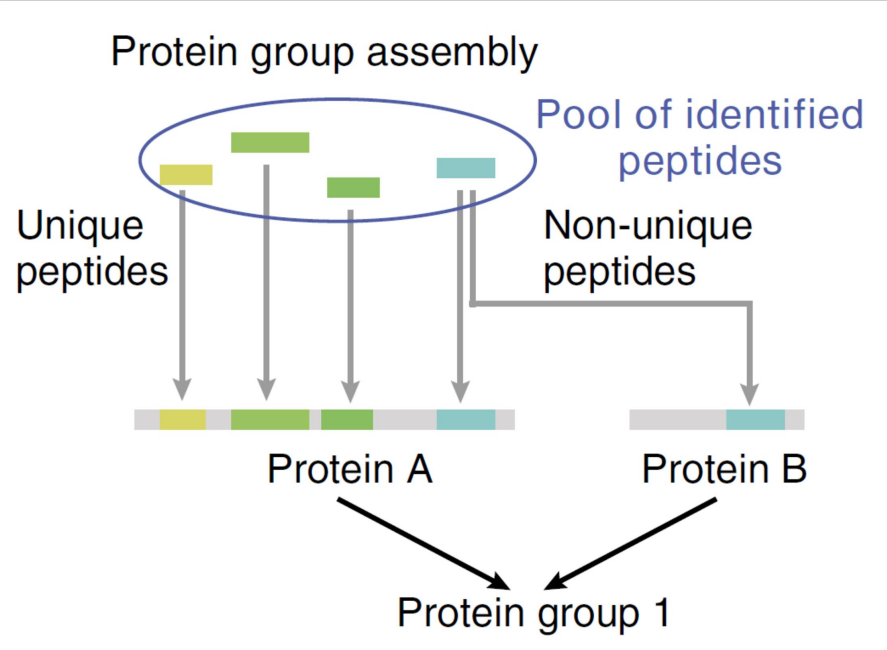
Peptide detectability: the probability of observing a peptide in a standard sample analyzed by a standard proteomics routine (Tang et *al.*, 2006).



Peptide detectability can be used as a statistical measure to help increase / decrease credibility of protein identification results.

Adapted from Lu et al., Nature (2006).

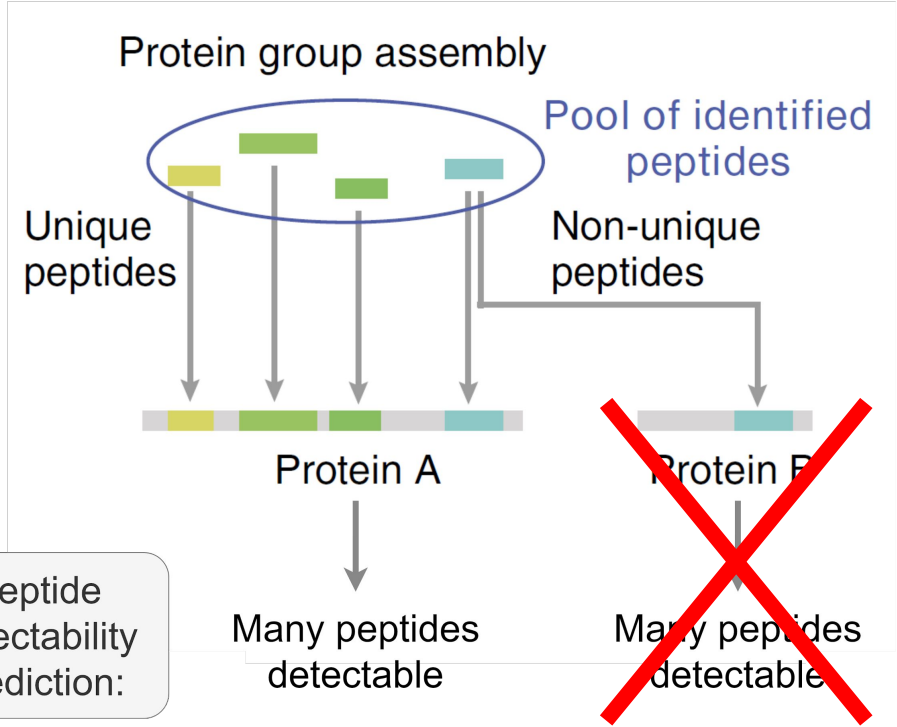
Integrating peptide detectability in protein identification



Peptide detectability can be used as a statistical measure to help increase / decrease credibility of protein identification results.

Adapted from Tyanova, Temu & Cox, Nature Protocols (2016).

Integrating peptide detectability in protein identification



Peptide detectability can be used as a statistical measure to help increase / decrease credibility of protein identification results.

Adapted from Tyanova, Temu & Cox, Nature Protocols (2016).

INTRODUCING MY PROJECT

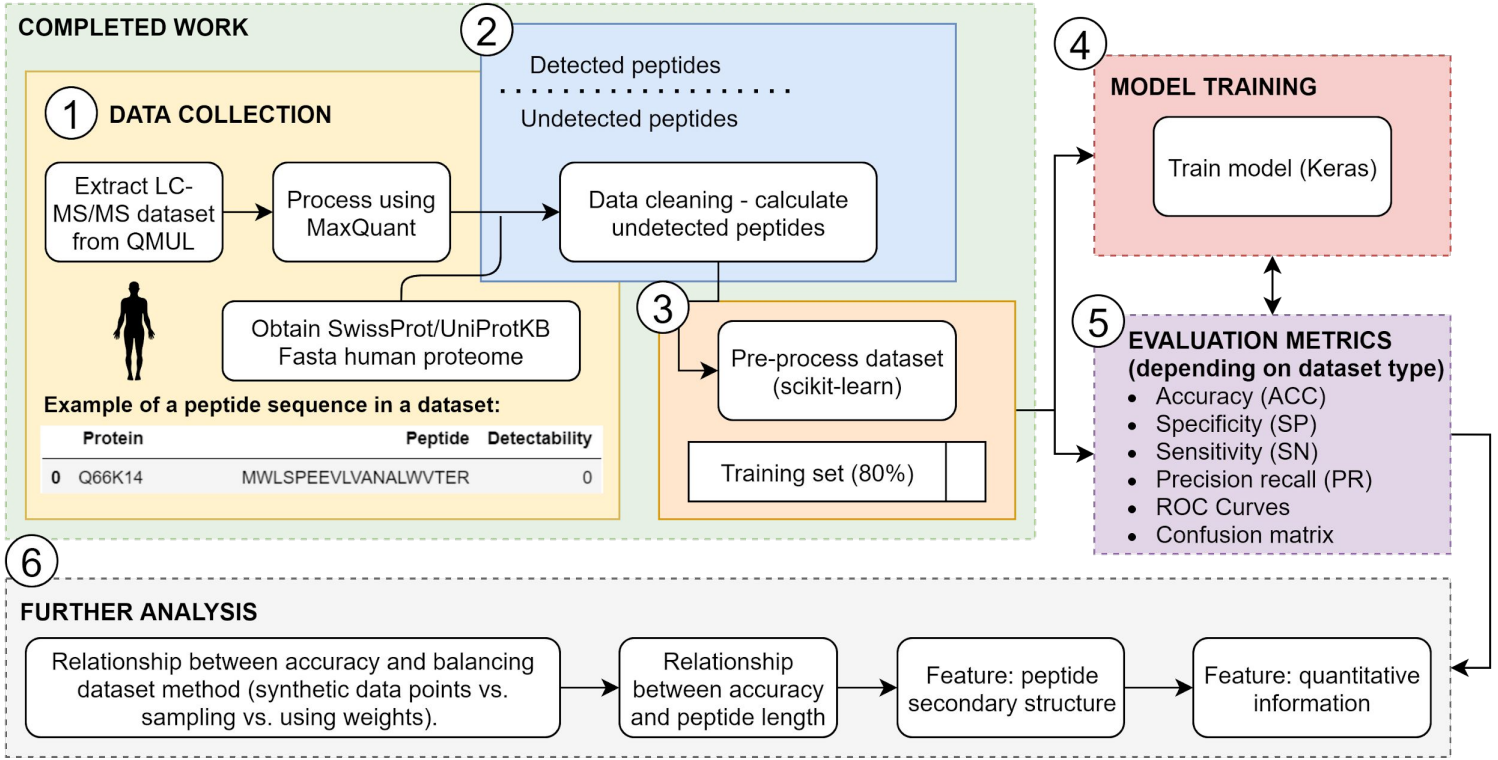
Project aims

BIG PICTURE GOAL: using peptide detectability to improve the accuracy and confidence of protein identification in mass spectrometry.

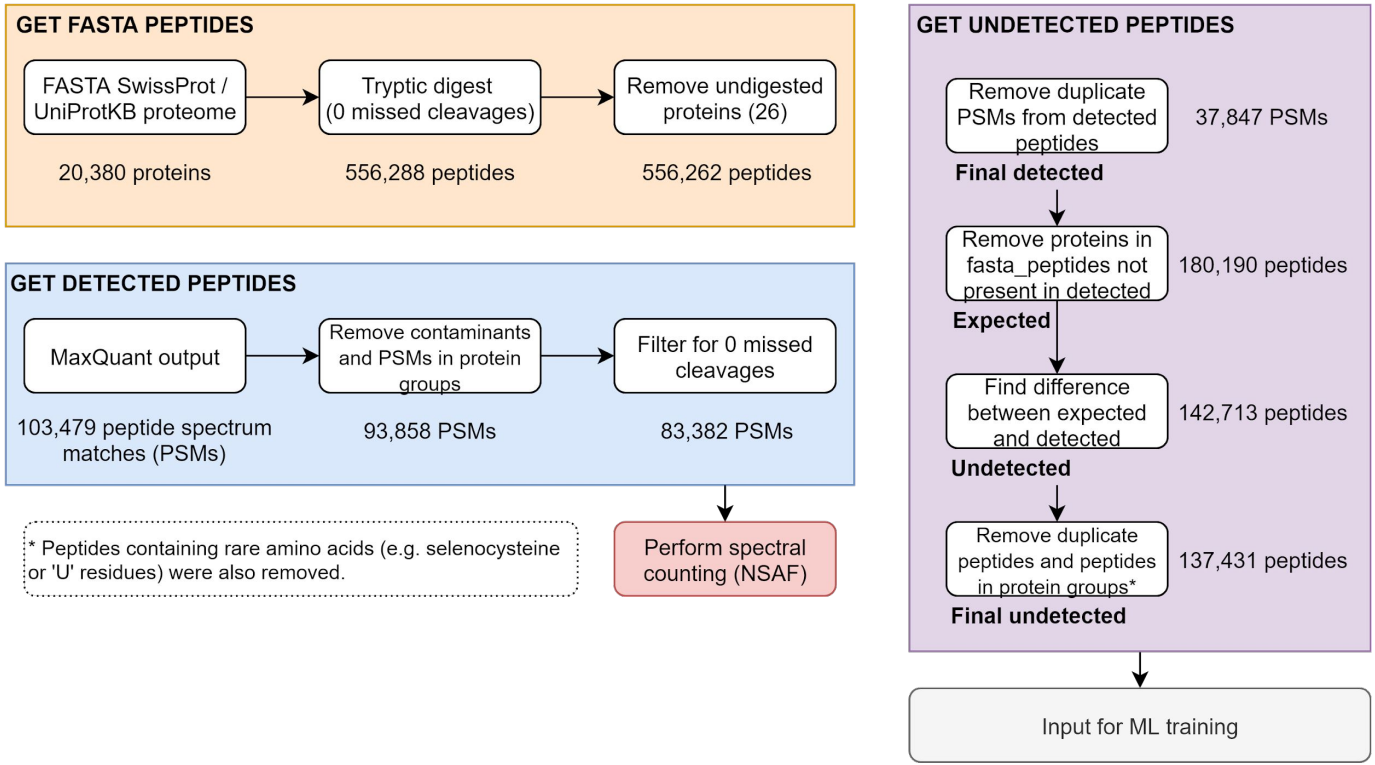
Creating a model that can successfully predict the detectability of peptides of a protein, through:

1. Data processing using MaxQuant + data cleaning and pre-processing for ML.
2. Optimisation of an already built transformer network (a type of neural network).
3. Incorporating features (such as protein abundance) and observing effect on model.

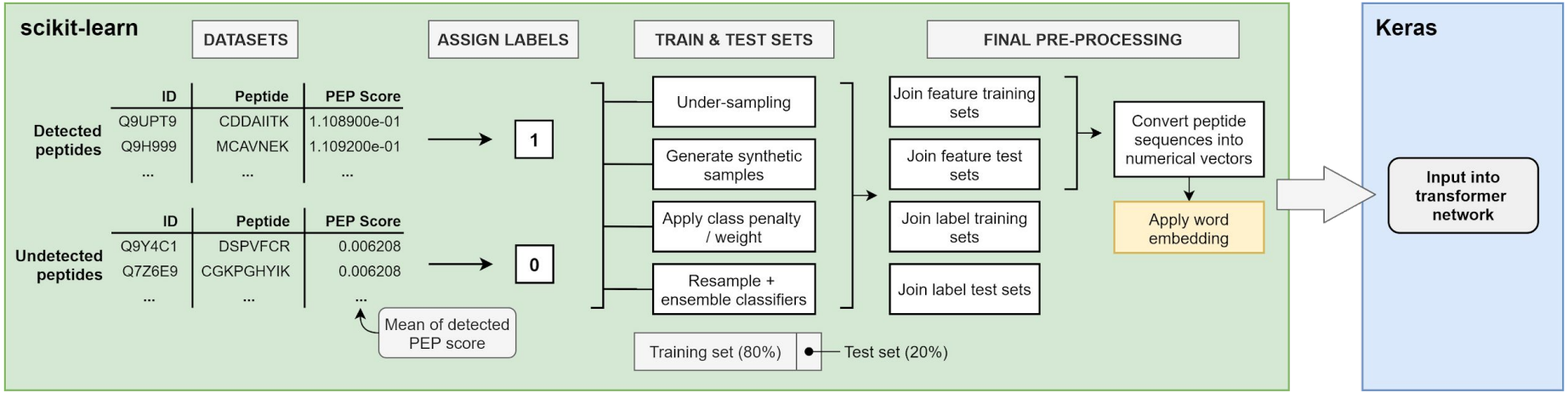
Methodology



Data cleaning & quality control



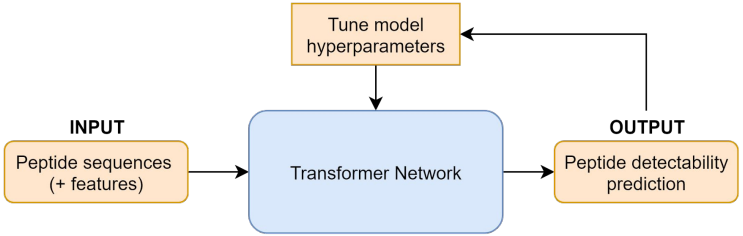
Data preprocessing for model training

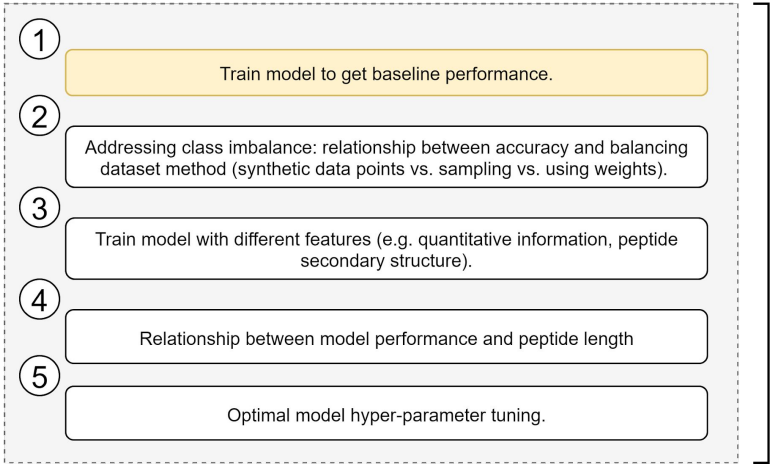


Future work

Research questions:

1) Which features best improve model performance, and why?
2) How does model performance vary with peptide length?
3) How does model performance vary with dataset size?
4) What are the optimal hyperparameter tunings for the model?





Apply evaluation metrics:

- Accuracy (ACC)
- Specificity (SP)
- Sensitivity (SN)
- Precision recall (PR)
- ROC Curves
- Confusion matrix

Acknowledgements

Big thank you to...

Conrad (supervisor) for your patience and support as I got my head around this project.

Esteban (co-supervisor) for nicely answering all of my questions (repeatedly).

Hajar (co-supervisor) for your invaluable project tips when I felt lost and confused.

The **Bessant lab** for the warm welcome, encouragement and banter.

Jessica & Siv for your feedback on my presentation, and **Alej** for moral support / supplying memes.

And last but certainly not least, my **family and friends** for helping me maintain my sanity!

Thank you! Questions?