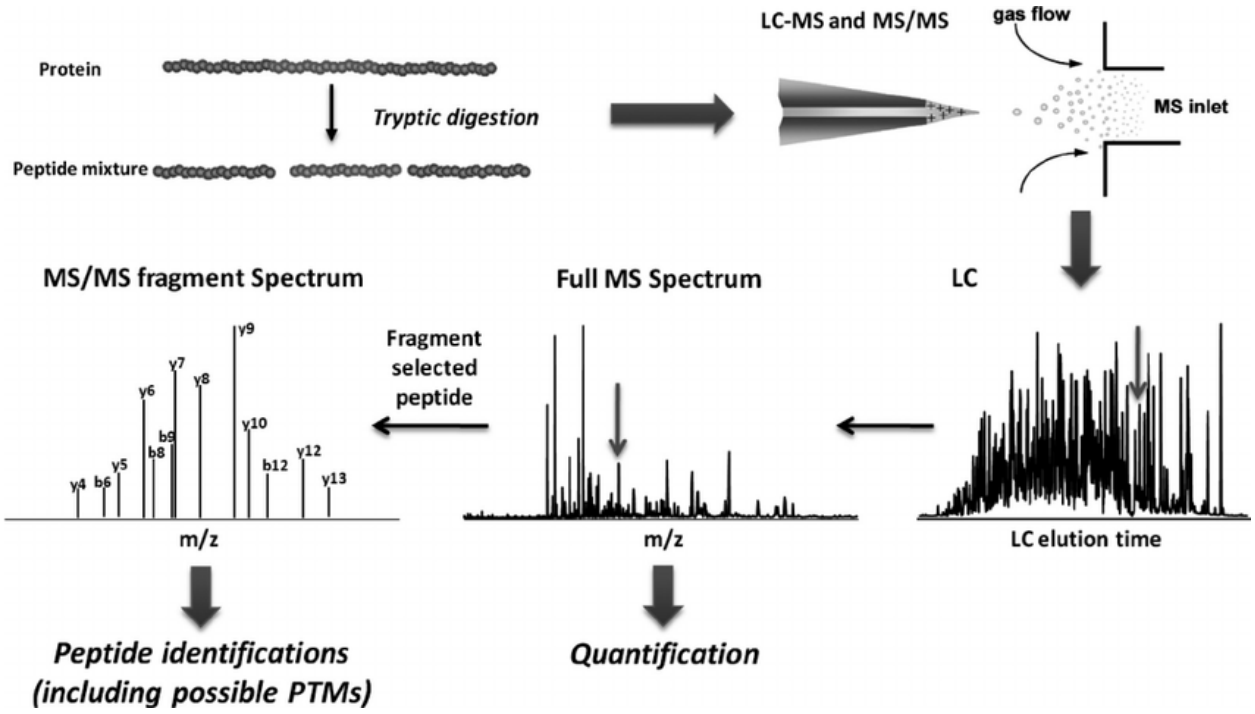
Technische Universität Braunschweig

Functional Genomics Practical

Mohammad Rezaei| November 20245

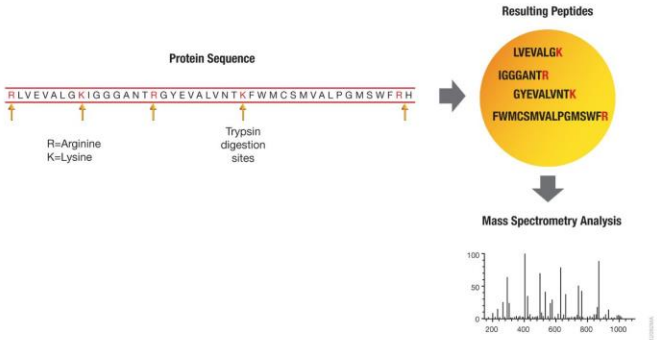
Proteomics Work Flow



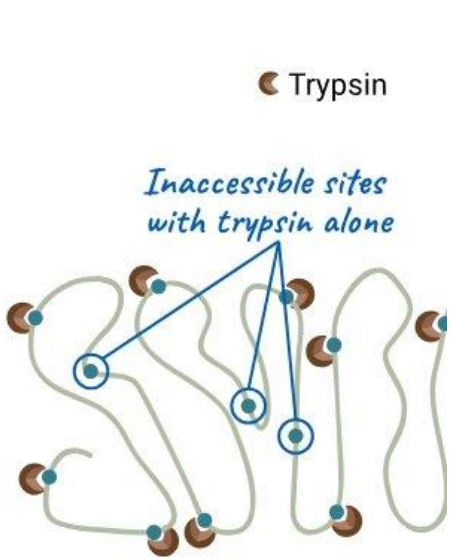
DOI: 10.1074/jbc.R110.199703

From theoretical peptides to real sample digest

The ***"theoretical digest"*** means: for a given protein sequence, apply enzyme rules
(e.g. trypsin always cuts after K/R) → list of all possible peptides.



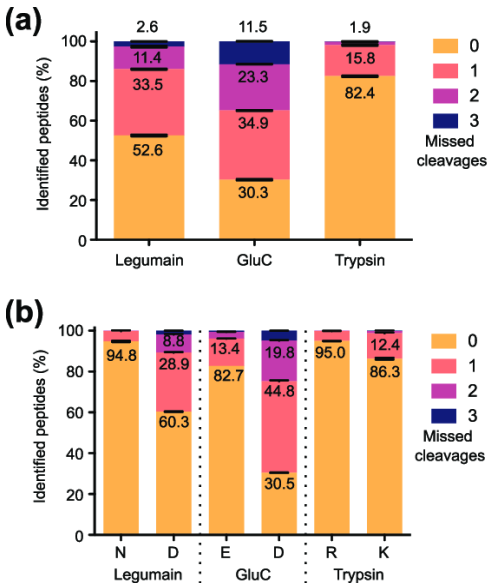
https://www.promega.de/en/resources/guides/protein-analysis/protease-digestion-for-mass-spec/



DOI: 10.1074/jbc.R110.199703

From theoretical peptides to real sample digest

- In reality: protein extraction → denaturation → digestion → cleanup → LC-MS/MS. At each step peptides may be lost or transformed.
- So we should expect:
- ***Maximum theoretical peptides*** ≥ ***actual peptides detected*.**



DOI: 10.1021/acs.analchem.9b03604

Why many peptides don’t show up: key limiting factors

Digestion inefficiencies: e.g. missed cleavages, blocked/modified sites, protein structure impeding enzyme access → fewer “ideal” peptides.

Peptides with unsuitable properties: too small/too large, extreme hydrophobicity, weak ionisation → poor LC/MS detection.

Sample & LC losses: adsorption to tubes or tips, co-elution suppression, low abundance overshadowed by high abundance peptides.

Acquisition & selection issues: in DDA (data dependent acquisition) the instrument selects only top N-ions → many lower-abundance peptides never get fragmented.

Data analysis filters: some peptides are present but the MS/MS spectra are of poor quality, or search parameters too strict → they go undetected or unassigned. Systematic errors arise from e.g. modifications, shared peptides, etc.

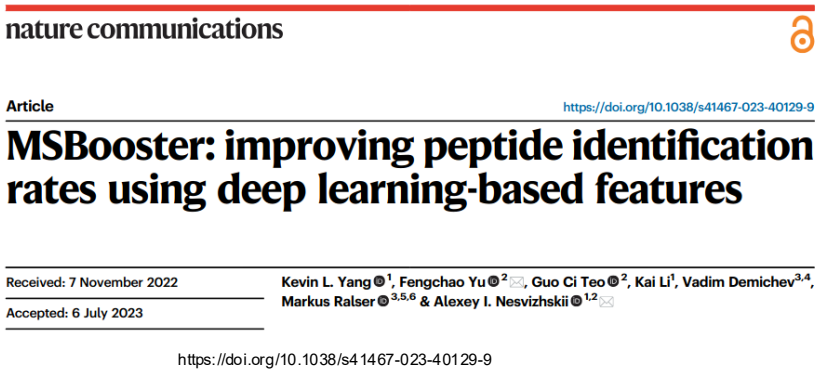
https://egor-pro.medium.com/missing-with-out-a-trace-da186405e02b

Consequences & mitigation strategies

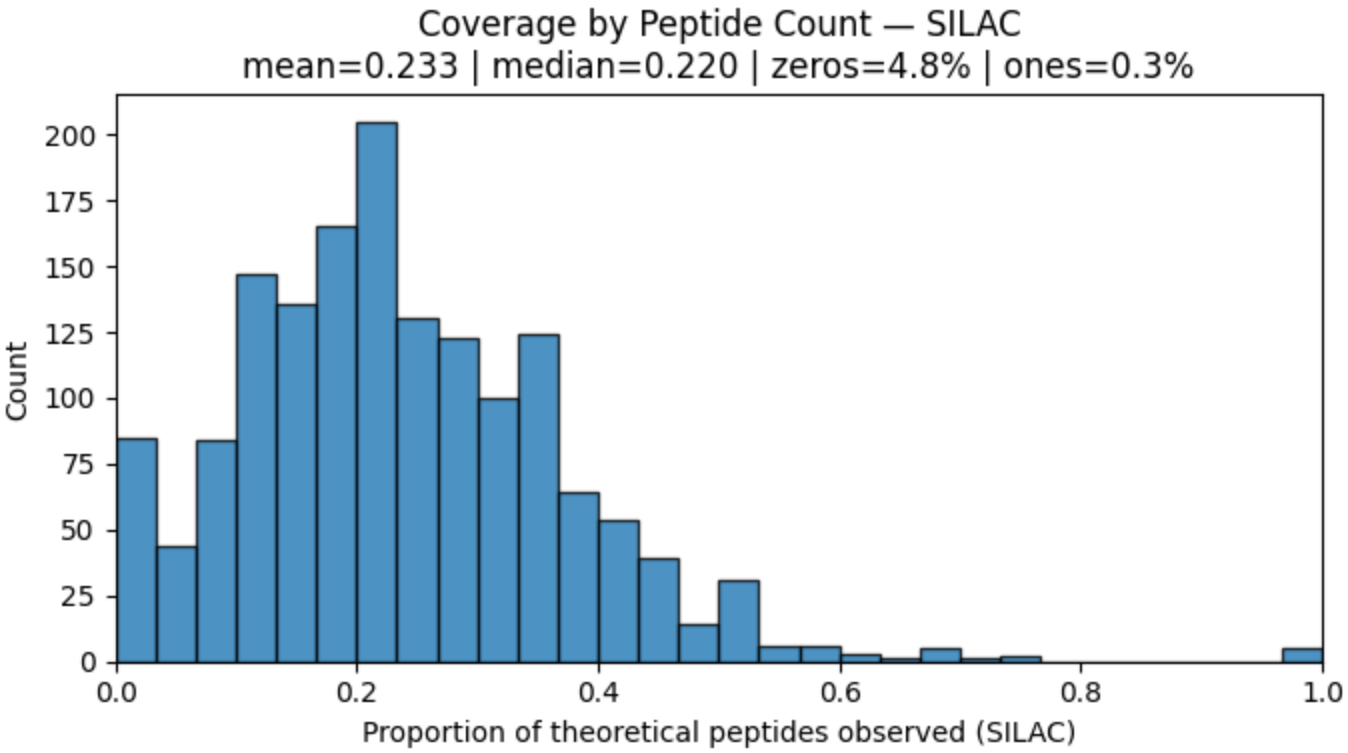
Consequences: lower proteome/peptide coverage than predicted; gaps (“missing values”) in quantitative datasets; bias toward high-abundance peptides; under-representation of low abundance or difficult peptides (e.g., from membrane proteins, PTMs)

Mitigation strategies:

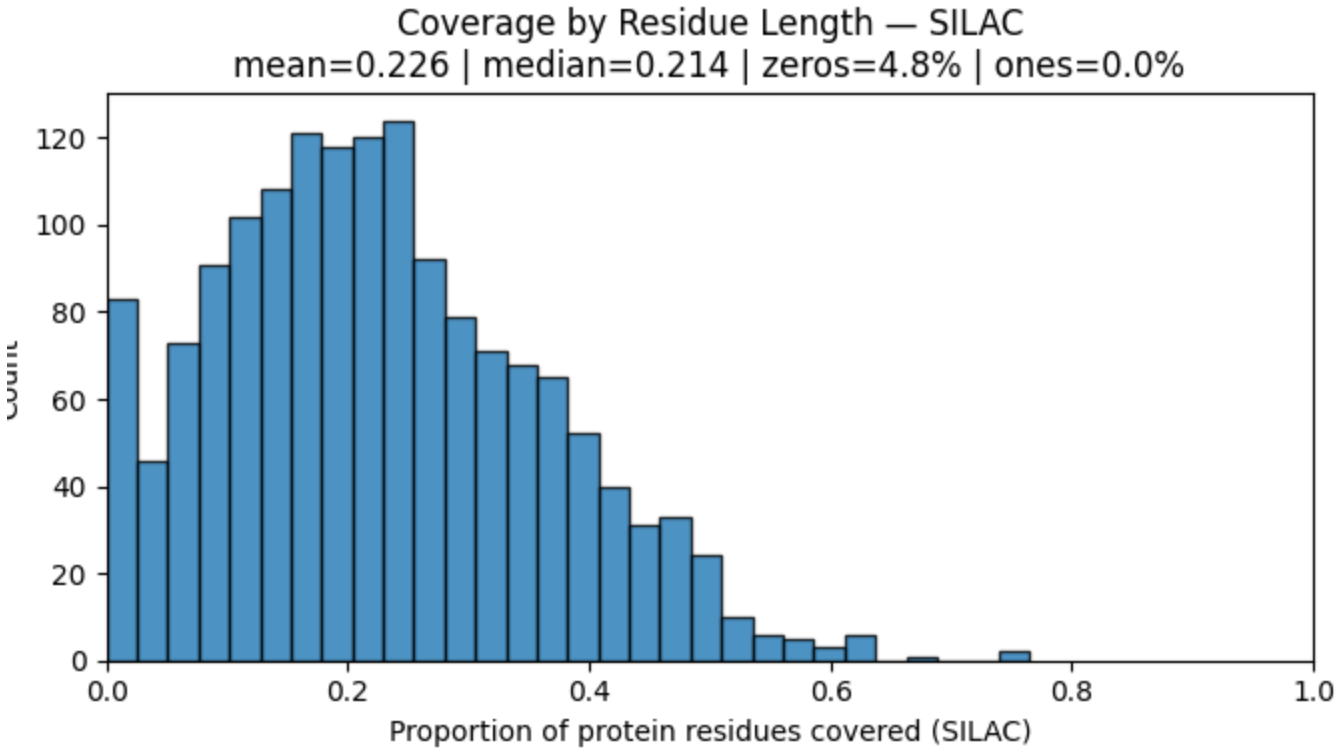
- Improve digestion efficiency (optimize enzyme:substrate ratio, denaturation, time, etc).
- Use alternative proteases or multi-enzyme digestions to cover “hard” regions.
- Choose acquisition mode carefully: consider DIA (data independent acquisition) to reduce stochastic missing values compared to DDA.



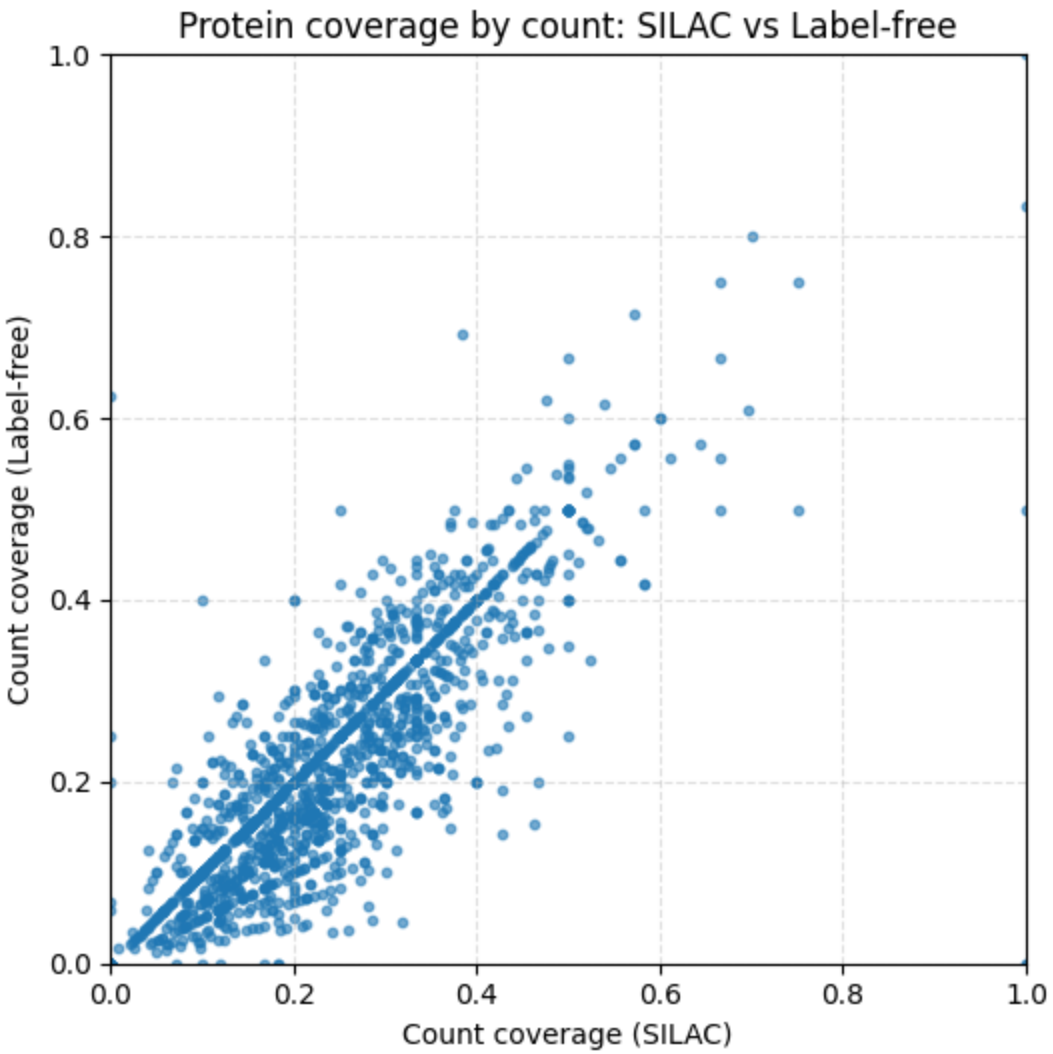
Sample Statistics



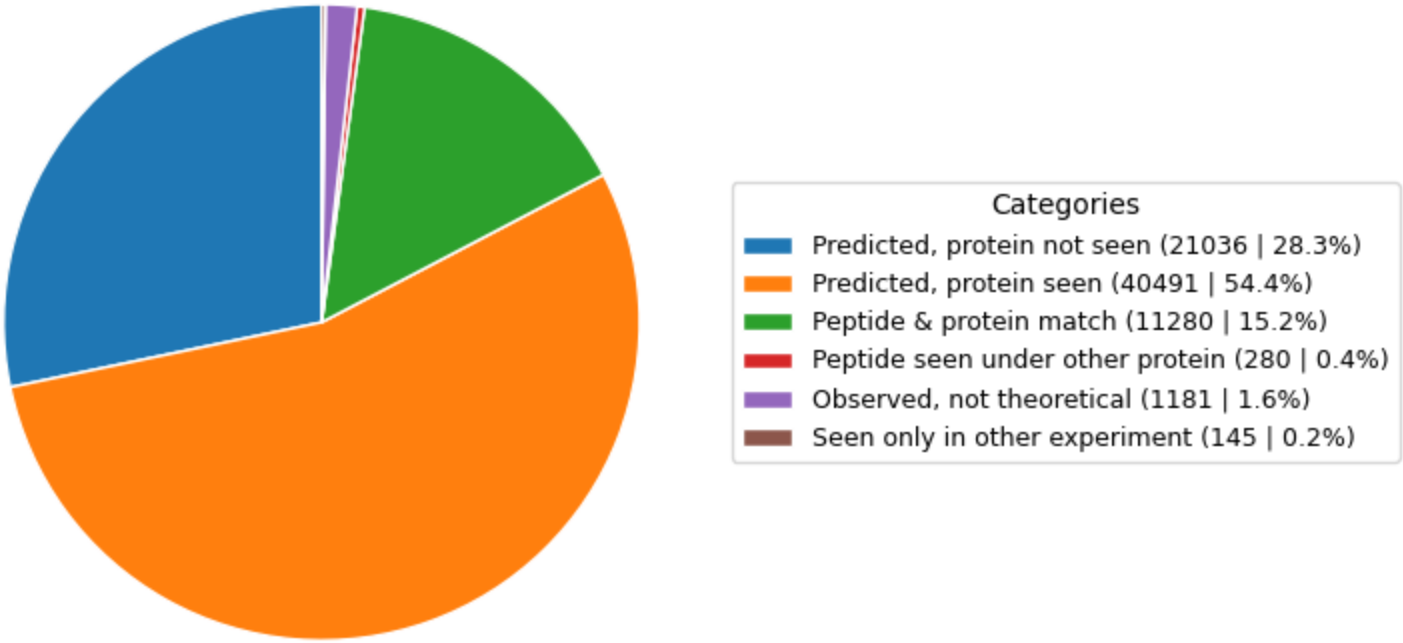
Sample Statistics



Sample Statistics



Sample Statistics

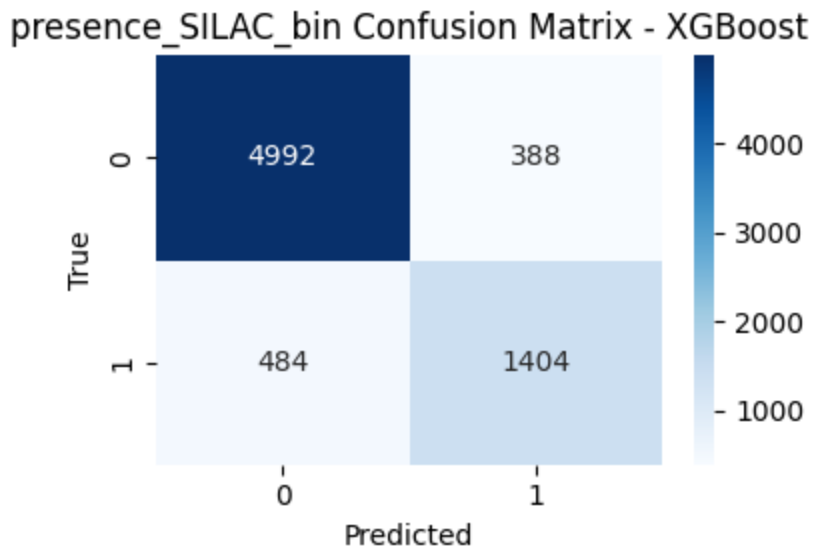


Training ML Models

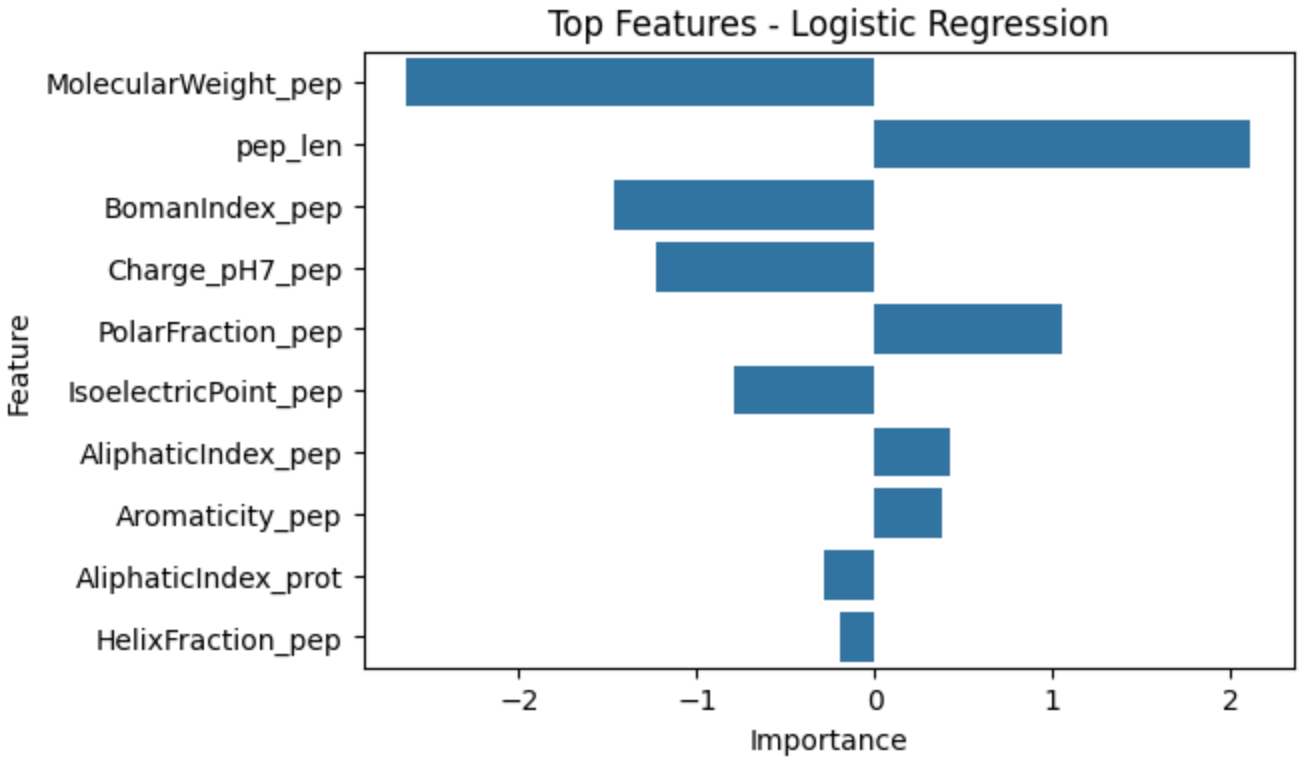
```
LogisticRegression | Accuracy: 0.820 | F1: 0.617
      RandomForest | Accuracy: 0.869 | F1: 0.731
           XGBoost | Accuracy: 0.881 | F1: 0.764
               SVM | Accuracy: 0.859 | F1: 0.710
               MLP | Accuracy: 0.847 | F1: 0.700
ESM2 Fine tuning | Accuracy: 0.887 | F1: 0.888
```

Training ML Models

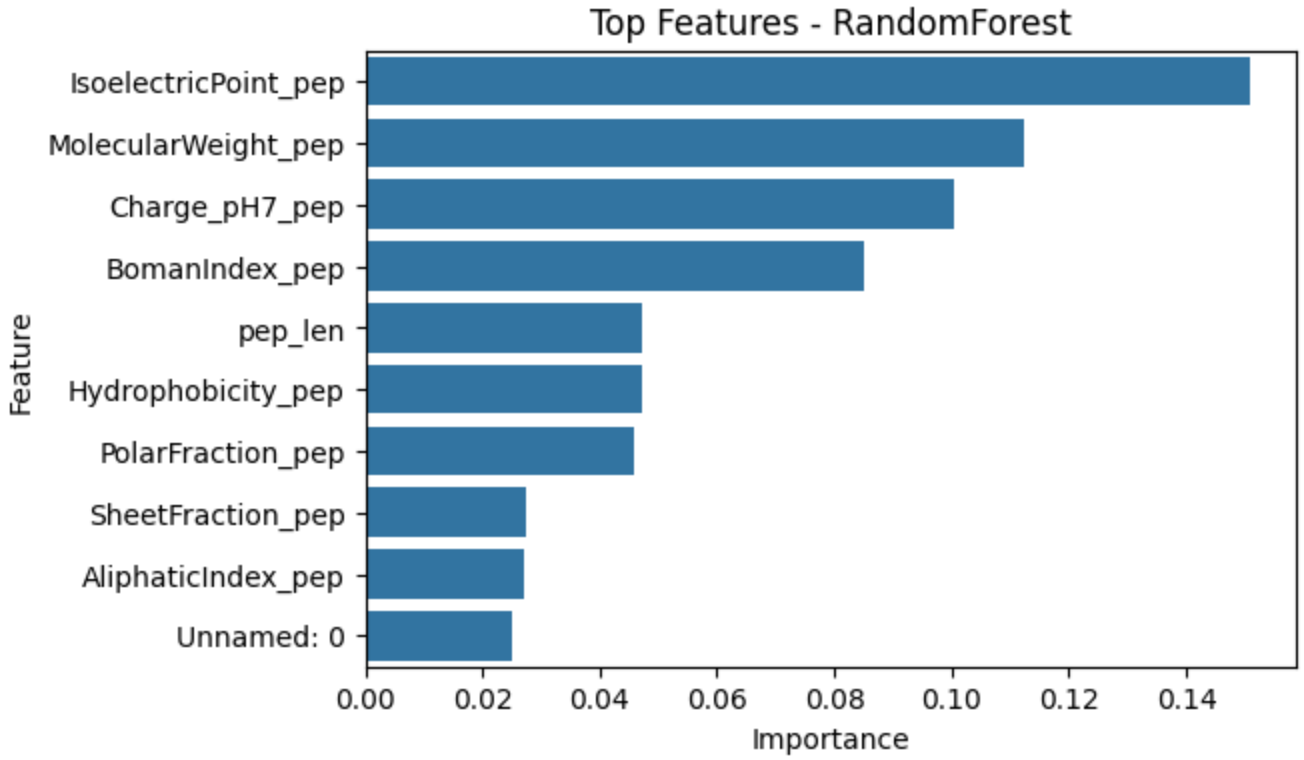
```
=== presence_SILAC_bin - XGBoost ===
Accuracy: 0.880, F1-score: 0.763
Classification report:
              precision    recall  f1-score   support

           0       0.91      0.93      0.92      5380
           1       0.78      0.74      0.76      1888

    accuracy                           0.88      7268
   macro avg       0.85      0.84      0.84      7268
weighted avg       0.88      0.88      0.88      7268
```



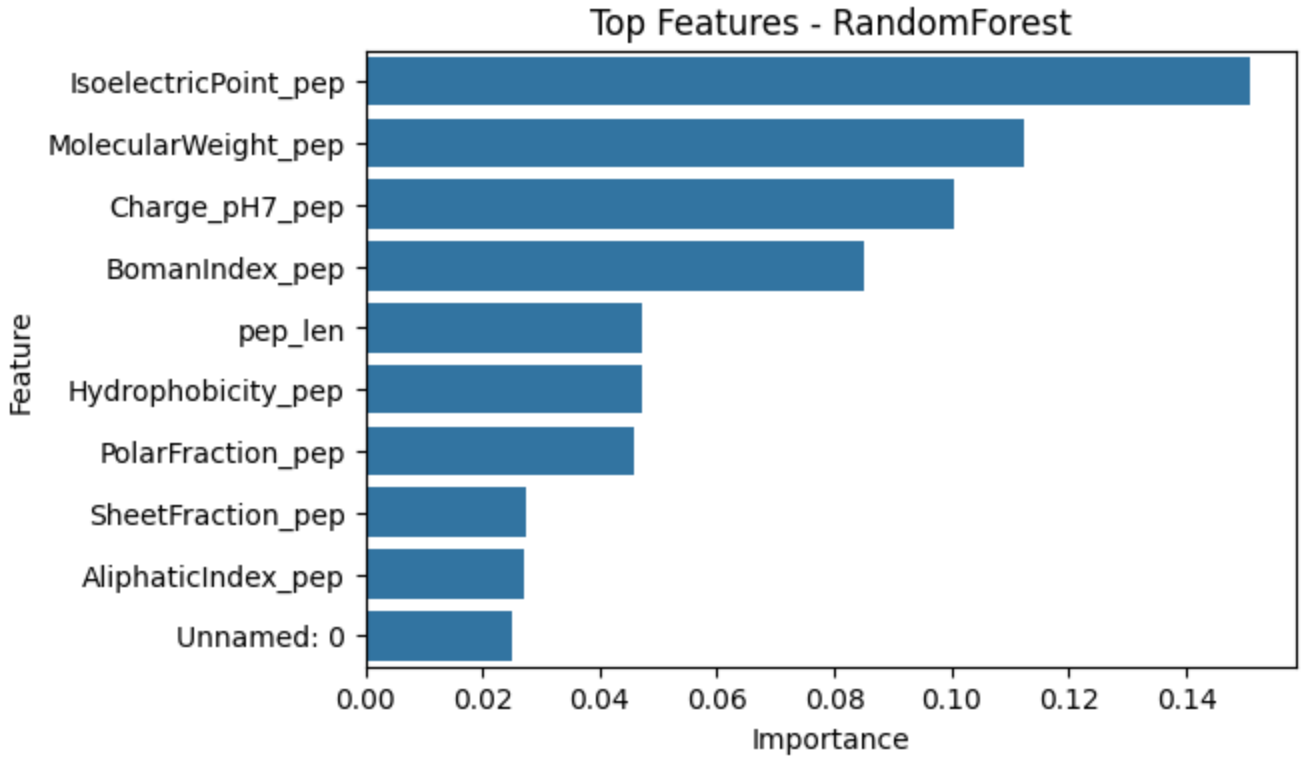
ML Interpretability



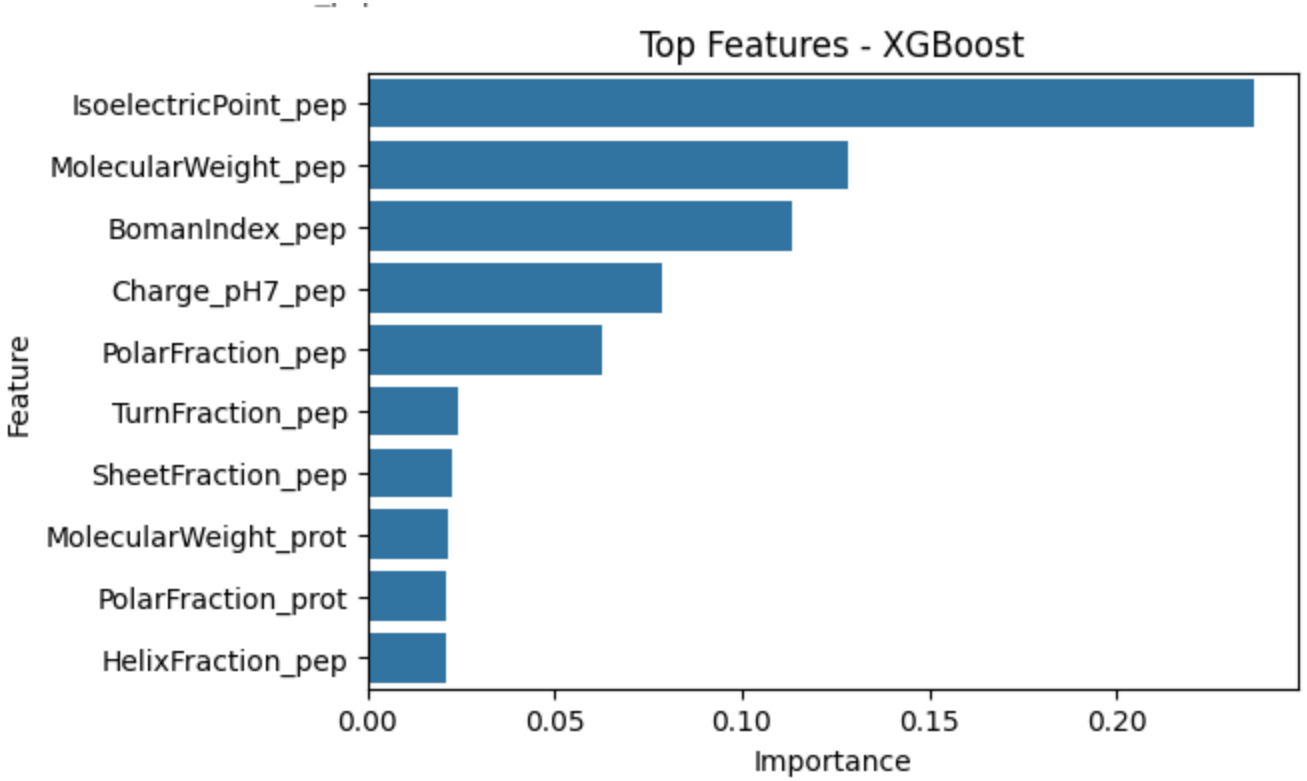
ML Interpretability



ML Interpretability



ML Interpretability

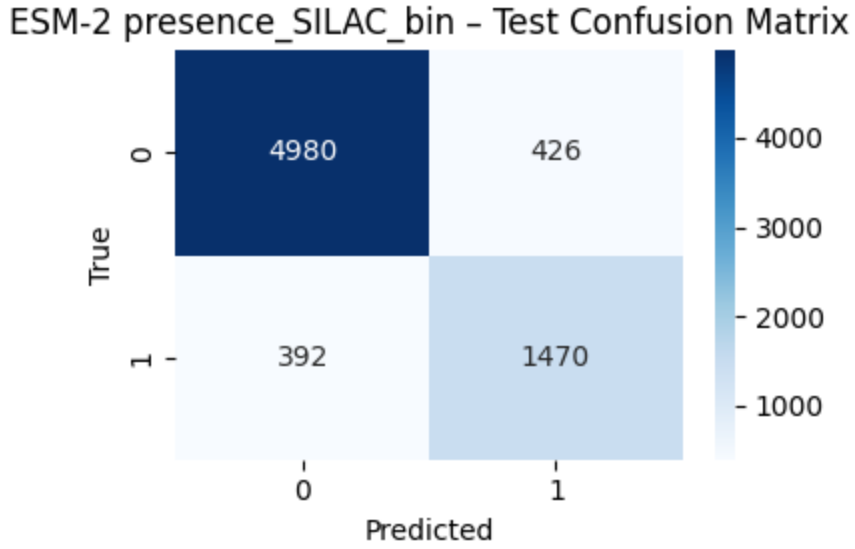


Deep Learning Models : Experiment with ESM 2

Deep Learning Models : Experiment with ESM 2

Classification report (test):

	precision	recall	f1-score	support
0	0.927	0.921	0.924	5406
1	0.775	0.789	0.782	1862
accuracy			0.887	7268
macro avg	0.851	0.855	0.853	7268
weighted avg	0.888	0.887	0.888	7268



Deep Learning Models : Experiment with ESM 2

```
Example predictions on test set:
00 | peptide: EVTFKDEPGVTYVVQPISTNK | true: 1 | pred: 1
01 | peptide: KPGLYK               | true: 0 | pred: 0
02 | peptide: LDPNIR               | true: 0 | pred: 0
03 | peptide: ALMGANMQR            | true: 1 | pred: 1
04 | peptide: GISVTSSVMQFDYDDYK    | true: 1 | pred: 1
05 | peptide: ATNEESYLMQK          | true: 1 | pred: 1
06 | peptide: NKQVDGFTTNPSLMAK     | true: 0 | pred: 0
07 | peptide: DIVAESPDLVIVGGGIANADDPVEAAK | true: 1 | pred: 1
08 | peptide: ETTAIDIPFAAR         | true: 1 | pred: 0
09 | peptide: IRETAR               | true: 0 | pred: 0
```

Practical Outline

1. First iteration: EDA + initial ML training ✅ *(completed)*

What was done

- Performed **exploratory data analysis (EDA)** on the available samples.
- Trained **several baseline ML models** (e.g. regression/classification models) to predict the target(s).

Main outcomes

- Identified **which models perform reasonably well**.
- Identified **key challenges**:
  - Limited sample size → risk of overfitting
  - Reproducibility of results

Practical Outline

2. Experiment with digestion metrics & integrate into pipeline
⏳ **~2 days**

Goal

- Design and test **digestion-related parameters**
- **Integrated End-To-End pipline**

Planned work

- Integrate these metrics into preprocessing & feature engineering steps.
- See how digestion parameters affects the predictions

Practical Outline

2. Experiment with digestion metrics & integrate into pipeline
⏳ **~2 days**

Goal

- Design and test **digestion-related parameters**
- **Integrated End-To-End pipline**

Planned work

- Integrate these metrics into preprocessing & feature engineering steps.
- See how digestion parameters affects the predictions

Practical Outline

3. Comprehensive data analysis + full training on all samples ⏳ ~1 week (approx.)

Goal

- Move from “prototype” to **complete analysis** using **all available samples** and the improved feature set.

Planned work

- Full, cleaned EDA on the **entire dataset**
- Evaluate performance using solid metrics (AUC, F1).
- Perform **error analysis**: which types of peptides/samples are systematically mis-predicted?

Expected result

- A **robust, documented pipeline** from raw features → digestion metrics → model → performance.
- Clear understanding of **limitations and strengths** of the approach.

Practical Outline

4. Final report, findings, and literature-based suggestions ⏳ ~1 week

Goal

- Summarise all practical work and link it to a report.
- Provide **recommendations** for improving peptide detection and experimental design.

Planned work

- Write **final report** including:
  Introduction & background, Methods, Results, and Discussion.