SFI Centre for Research Training in GENOMICS DATA SCIENCE

IBM

Centre for Research Training

Science Foundation Ireland sfi For what's next

UCD DUBLIN University College Dublin University for All

Partitioning, representation, and automation in canonical and non-canonical peptide modelling

Speaker: **Raúl Fernández-Díaz (PhD Candidate UCD – IBM Research)**

UCD: R. Cossio-Pérez, C. Agoni, D.C. Shields

IBM Research: T.L. Hoang, V. Lopez

Novo Nordisk: R. Ochoa

Part 0 - Introduction

More information and contact info

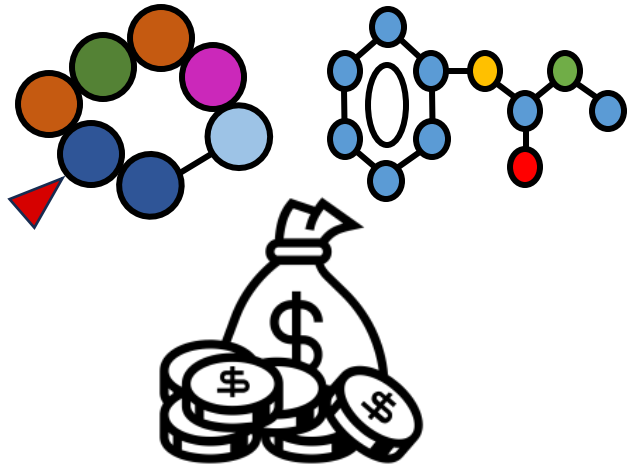
A tale of two peptides

Natural (or canonical) peptides

- Protein sequences (20 aa)
- Cheap
- Not great drugs

Synthetic peptides (or non-canonical) and peptidomimetics

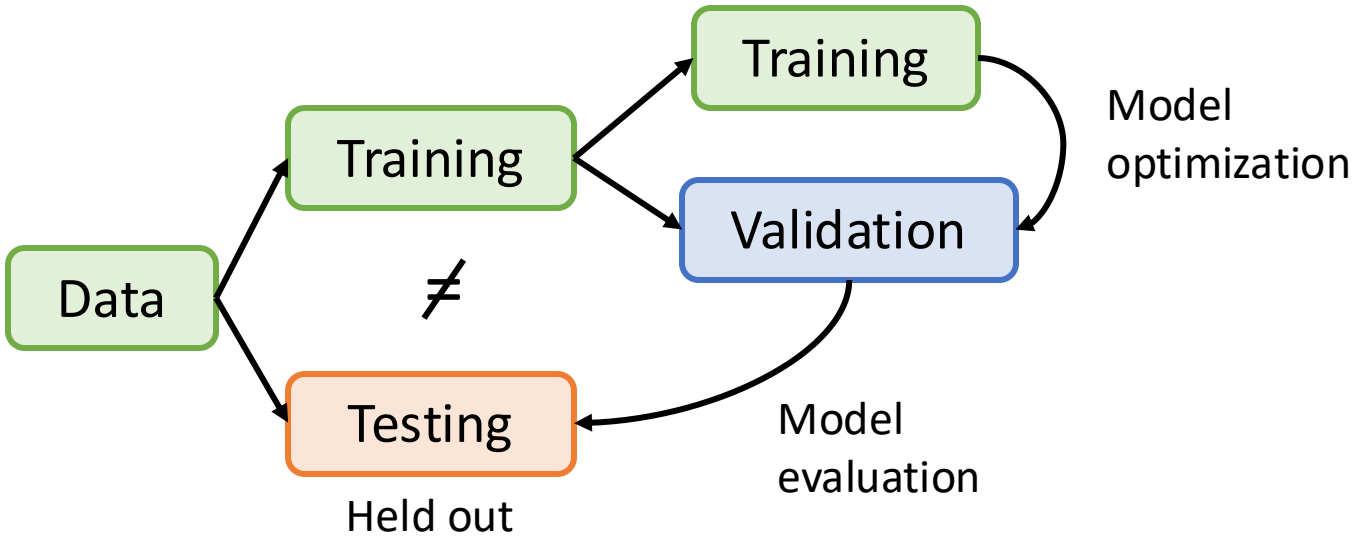
- Modified peptides + small molecules
- Expensive
- Better drug candidates

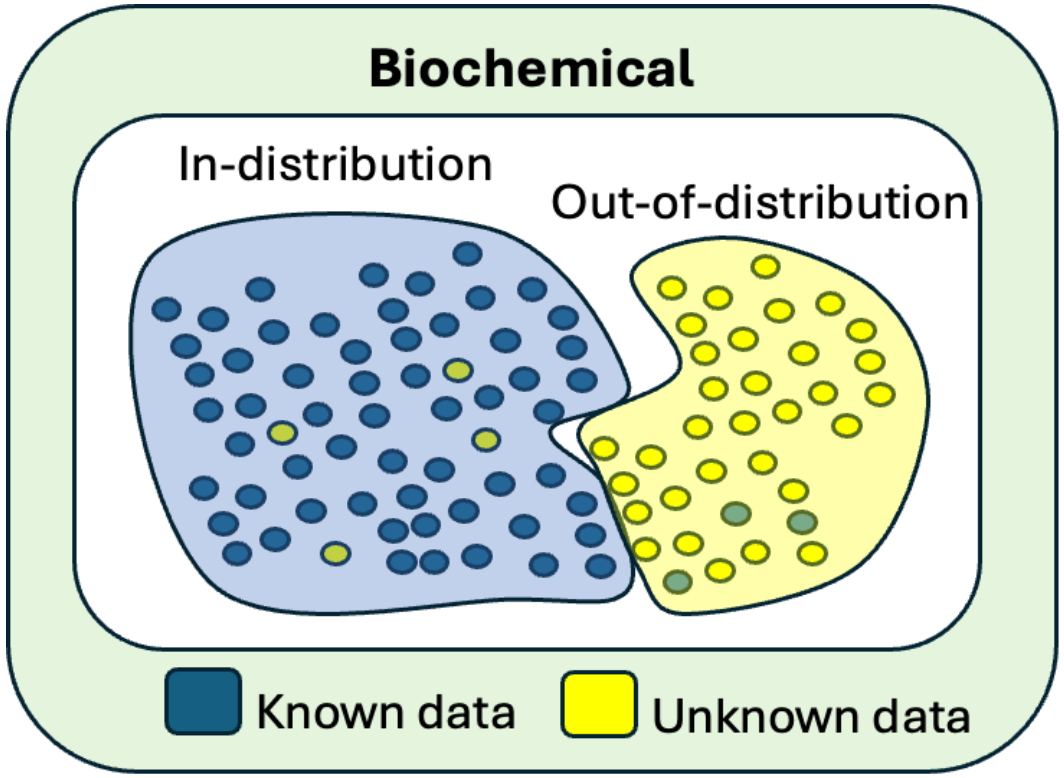


More information and contact info

Dataset partitioning

Central Assumption of ML: “Training data is representative of prediction data”





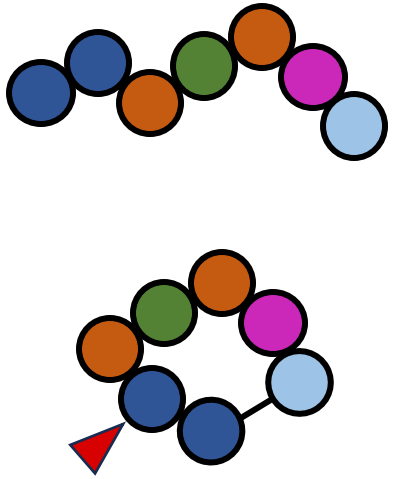
Training and testing need to have different molecules, otherwise we cannot evaluate generalization/extrapolation

Objective: Estimate model performance in unseen data

More information and contact info

Peptide representation

We need to transform molecules into numerical vectors



Methods

[1, 0.3, 0.4, 2.1, …]

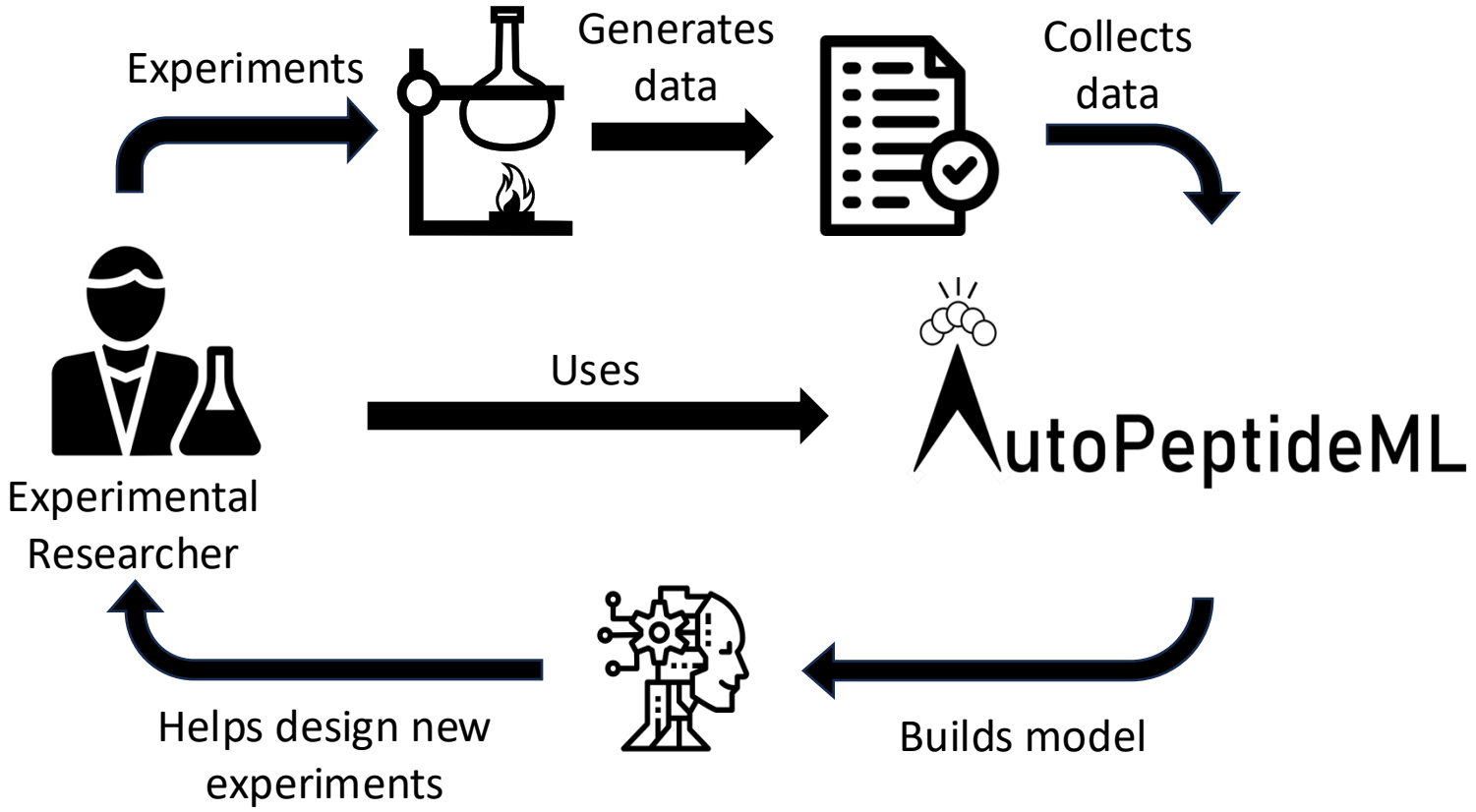
- **Heuristic**: Molecular FPs
  - Substructure-based: ECFP/Avalon
  - Monomer sequence: PepFuNN
- **Data-driven**: Pre-trained representation models
  - [Protein/Chemical/Peptide] Language Models
  - GNNs

More information and contact info

Automation

AutoPeptideML

R. Fernández-Díaz et al., AutoPeptideML: a study on how to build more trustworthy peptide bioactivity predictors, *Bioinformatics*, Volume 40, Issue 9, September 2024, btae555



Design Requirements

1. Robust evaluation
2. Reproducibility
3. Easier to integrate with experimental workflow

More information and contact info

Objectives

1. How to automatically build peptide property prediction models (and how to evaluate them)
2. How to extrapolate from natural to synthetic peptides or peptidomimetics

SFI Centre for Research Training in GENOMICS DATA SCIENCE

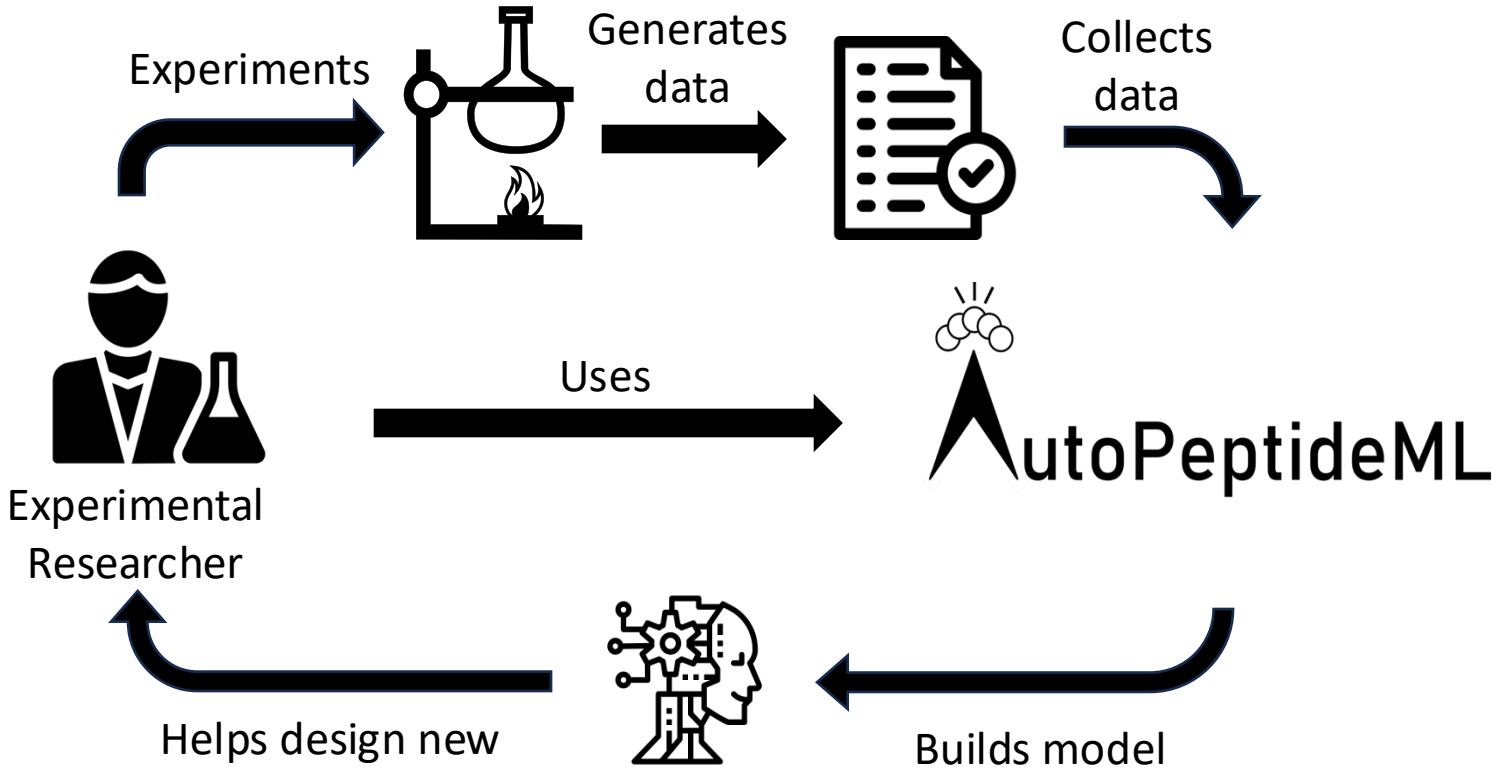
Part 1 - AutoPeptideML

More information and contact info

Automation

AutoPeptideML

R. Fernández-Díaz et al., AutoPeptideML: a study on how to build more trustworthy peptide bioactivity predictors, *Bioinformatics*, Volume 40, Issue 9, September 2024, btae555



Design Requirements

1. Competitive performance
2. Reliable evaluation so that experimental scientist can trust the models
3. Easy to use

SFI Centre for Research Training in GENOMICS DATA SCIENCE

More information and contact info

How good can automation really be?

Collected **18** datasets used for building different peptide bioactivity predictors

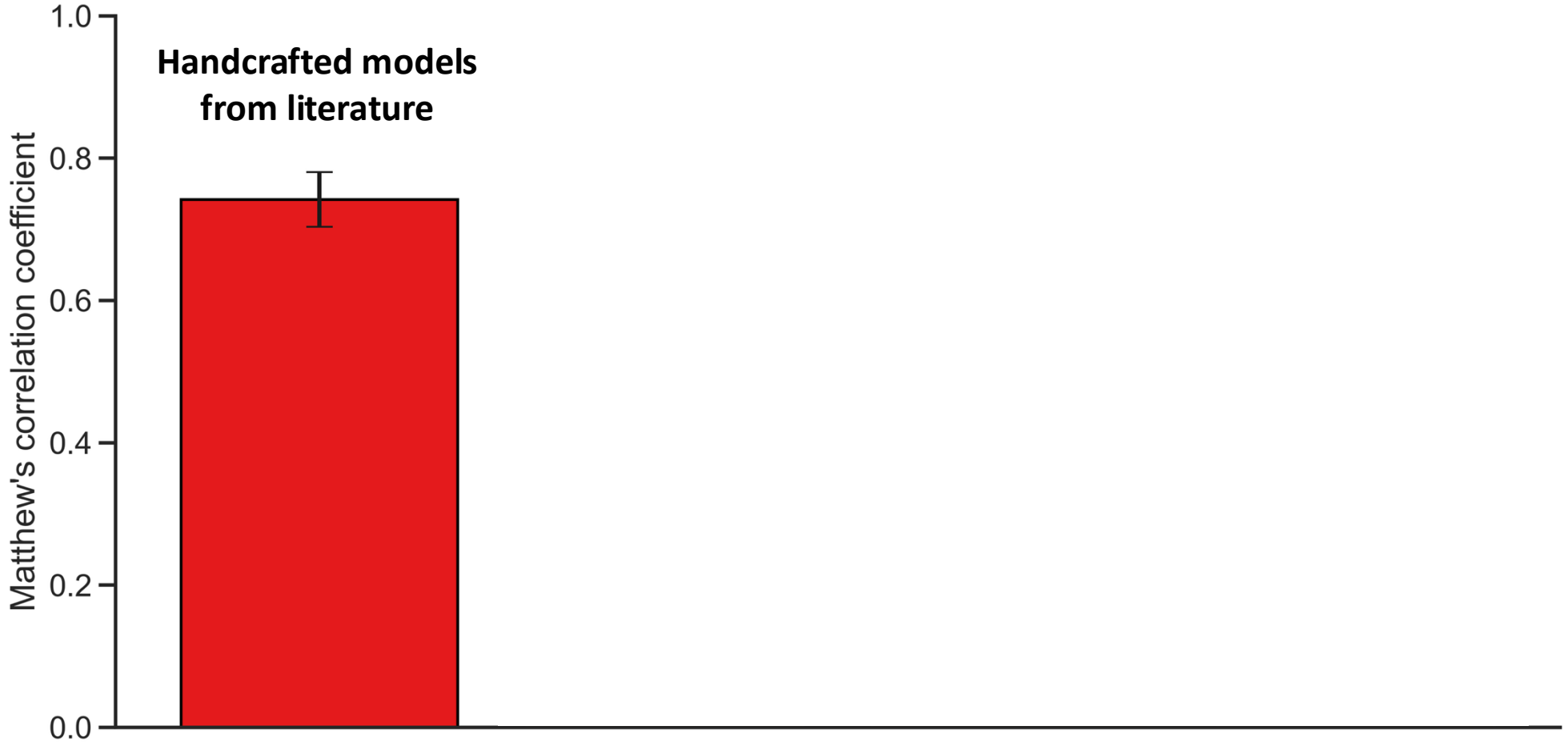
AutoPeptideML

R. Fernández-Díaz et al., AutoPeptideML: a study on how to build more trustworthy peptide bioactivity predictors, *Bioinformatics*, Volume 40, Issue 9, September 2024, btae555

More information and contact info

How good can automation really be?

Collected **18** datasets used for building different peptide bioactivity predictors



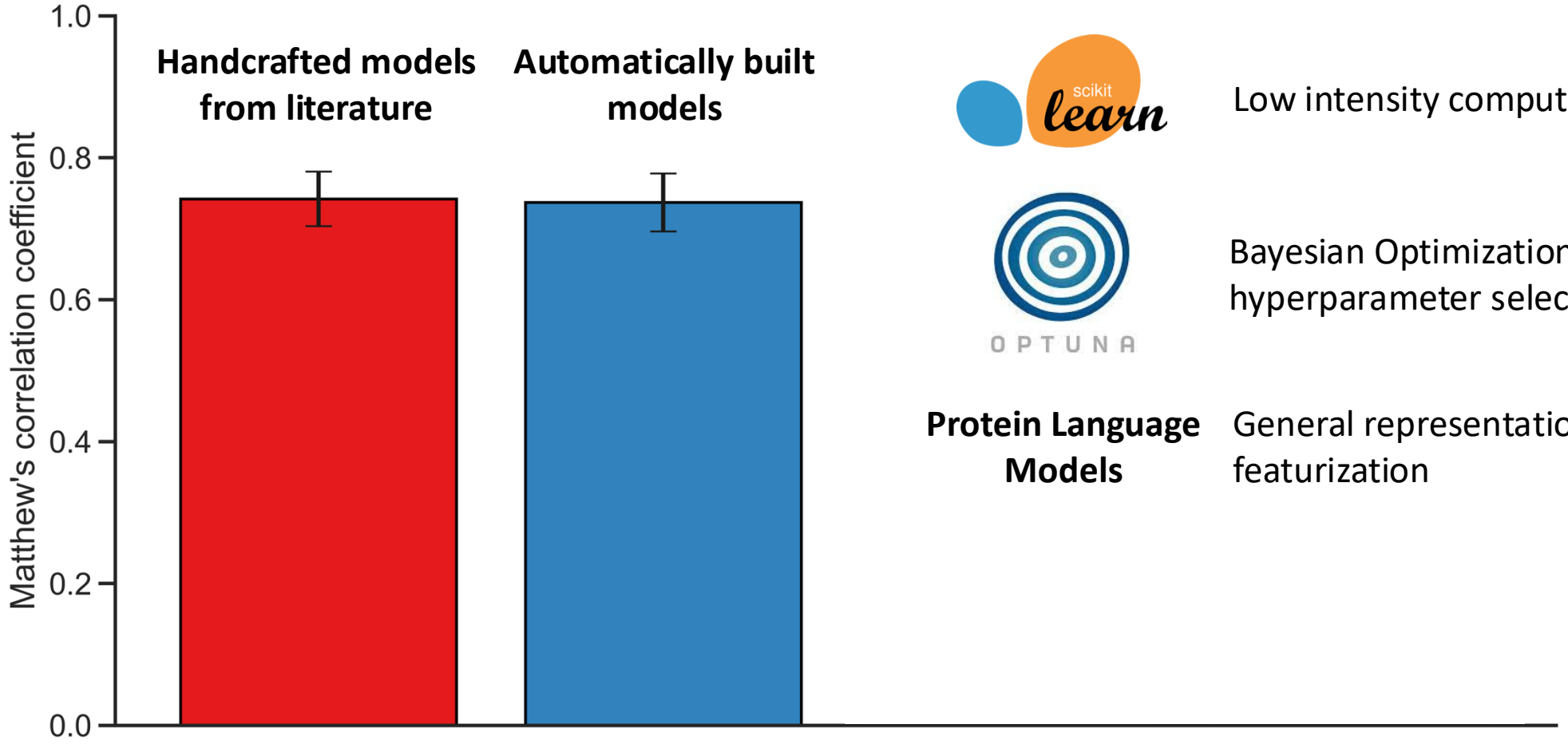
AutoPeptideML

R. Fernández-Díaz et al., AutoPeptideML: a study on how to build more trustworthy peptide bioactivity predictors, *Bioinformatics*, Volume 40, Issue 9, September 2024, btae555

More information and contact info

Automation is as good as human developers

Collected **18** datasets used for building different peptide bioactivity predictors



scikit learn — Low intensity computing

OPTUNA — Bayesian Optimization for hyperparameter selection

Protein Language Models — General representation/ featurization

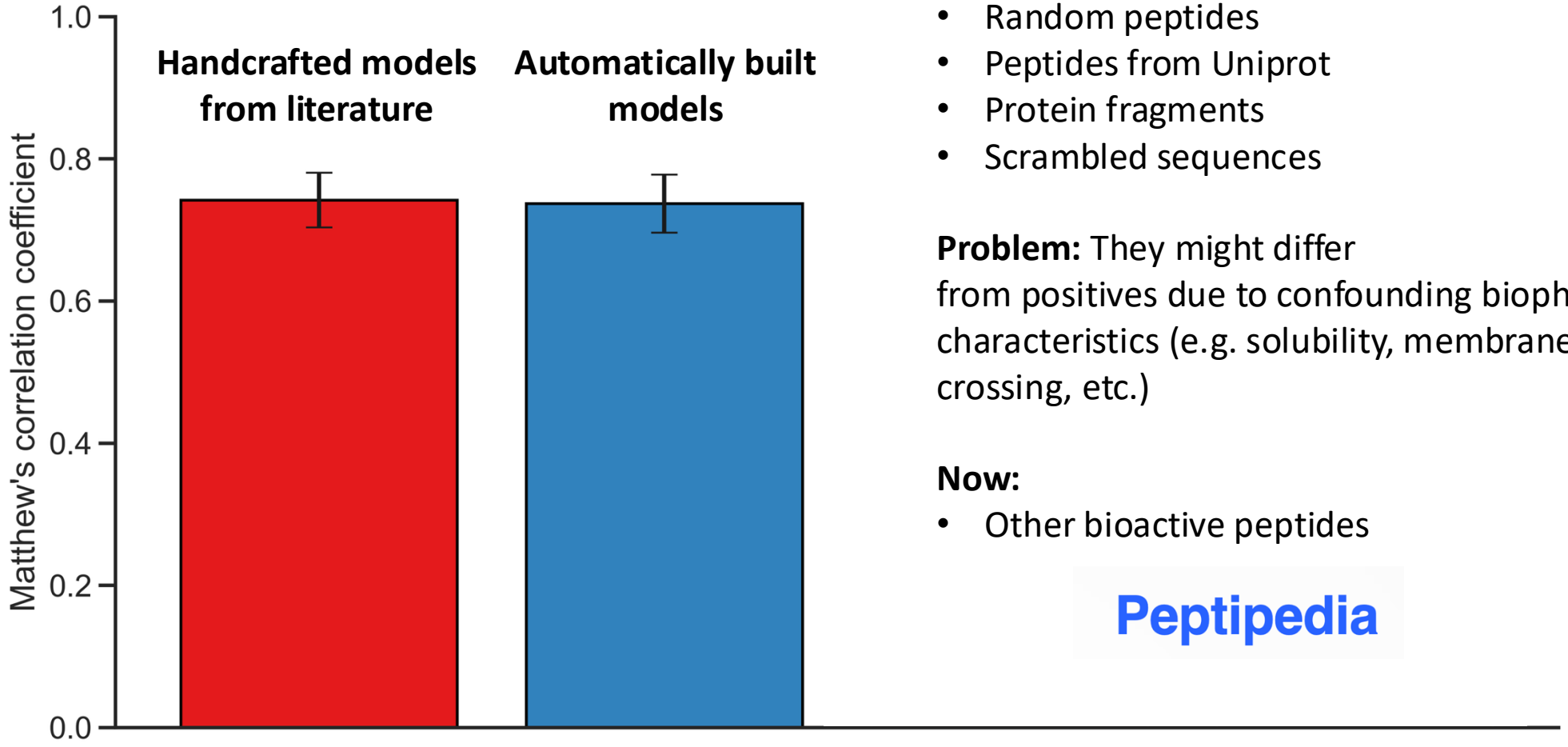
AutoPeptideML

R. Fernández-Díaz et al., AutoPeptideML: a study on how to build more trustworthy peptide bioactivity predictors, *Bioinformatics*, Volume 40, Issue 9, September 2024, btae555

More information and contact info

How important is the choice of negative samples?

Collected **18** datasets used for building different peptide bioactivity predictors



Before:

- Random peptides
- Peptides from Uniprot
- Protein fragments
- Scrambled sequences

Problem: They might differ from positives due to confounding biophysical characteristics (e.g. solubility, membrane crossing, etc.)

Now:

- Other bioactive peptides

Peptipedia

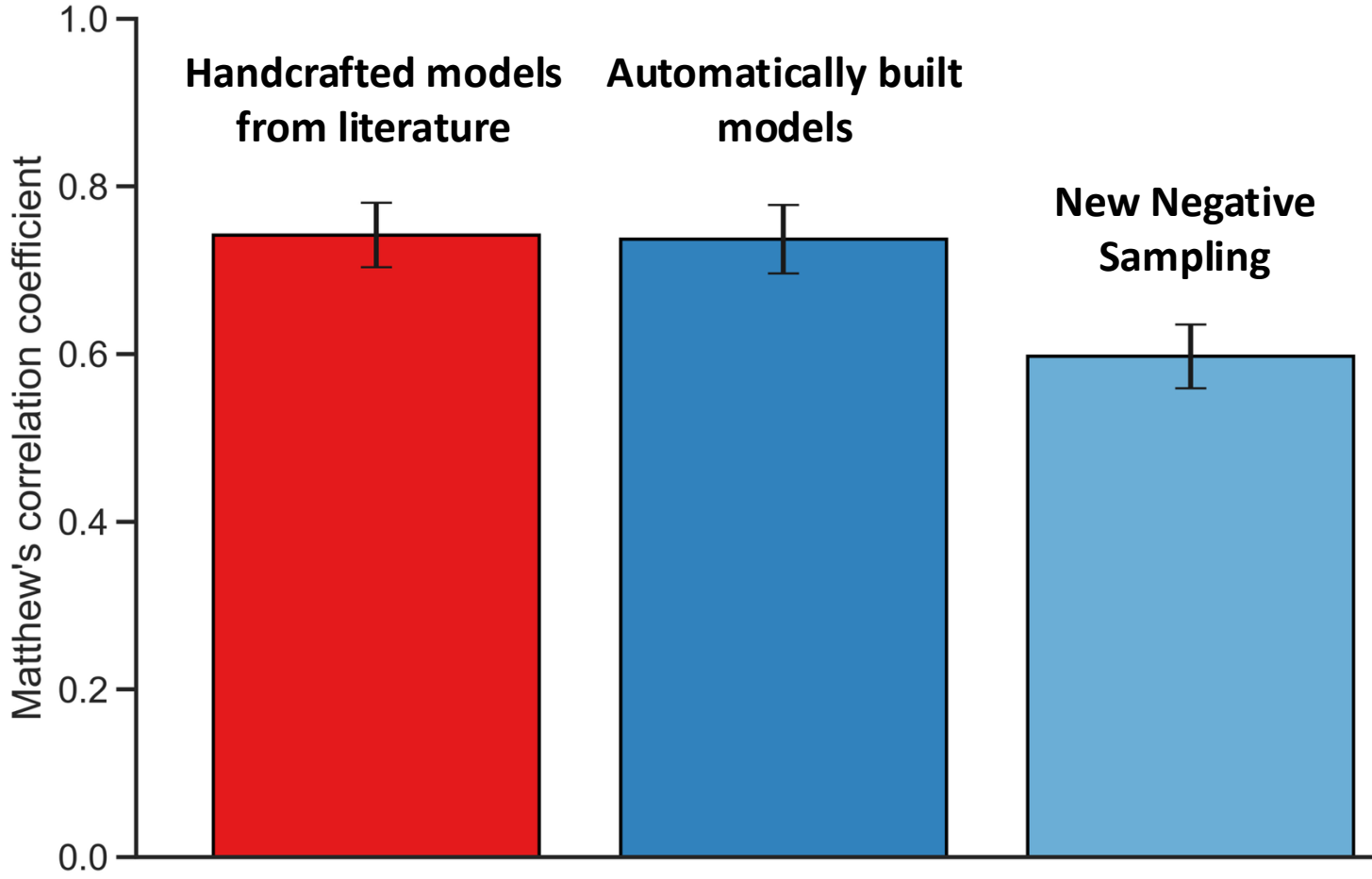
AutoPeptideML

R. Fernández-Díaz et al., AutoPeptideML: a study on how to build more trustworthy peptide bioactivity predictors, *Bioinformatics*, Volume 40, Issue 9, September 2024, btae555

More information and contact info

Proper choice of negatives is quite important

Collected **18** datasets used for building different peptide bioactivity predictors



Before:

- Random peptides
- Peptides from Uniprot
- Protein fragments
- Scrambled sequences

Problem: They might differ from positives due to confounding biophysical characteristics (e.g. solubility, membrane crossing, etc.)

Now:

- Other bioactive peptides

AutoPeptideML

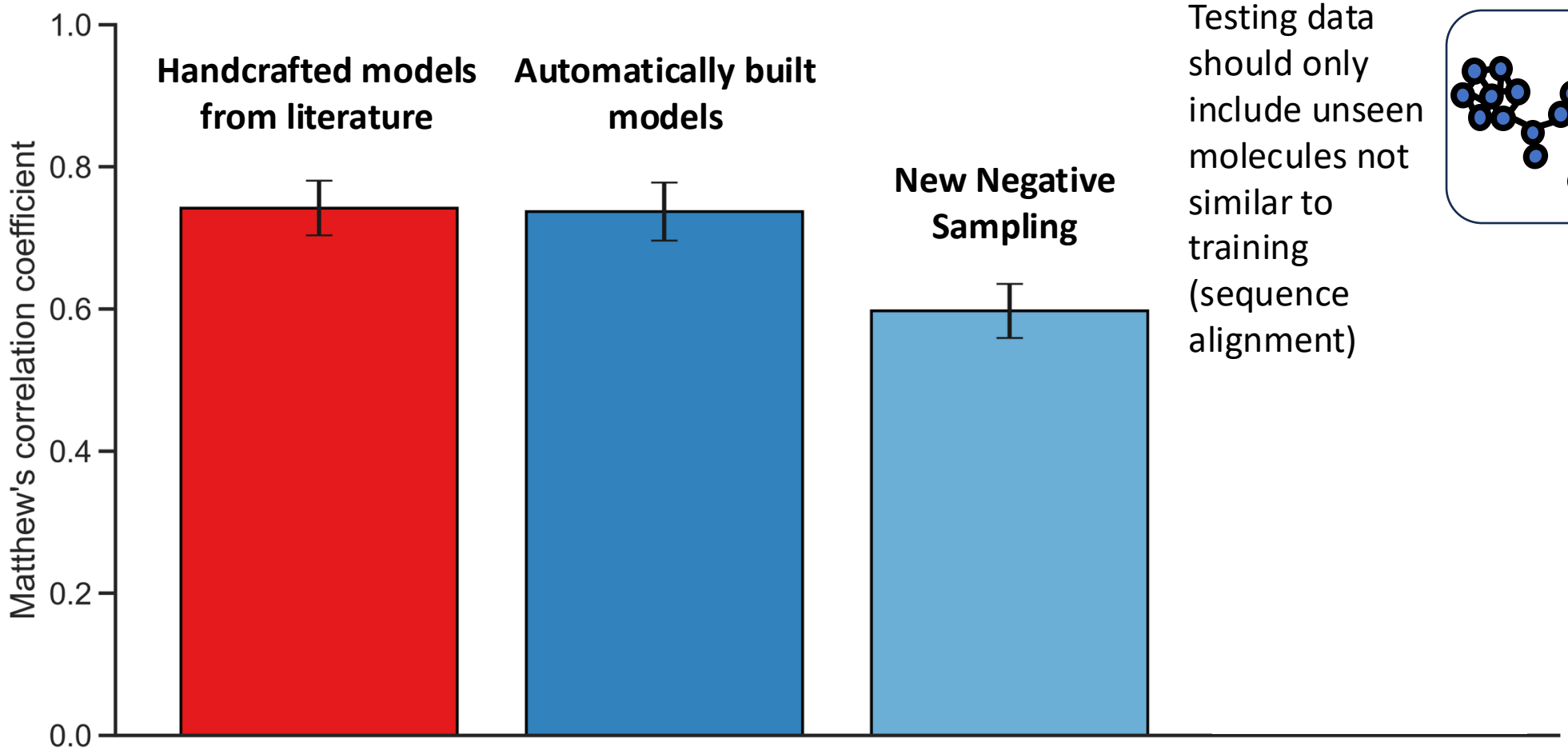
R. Fernández-Díaz et al., AutoPeptideML: a study on how to build more trustworthy peptide bioactivity predictors, *Bioinformatics*, Volume 40, Issue 9, September 2024, btae555

Peptipedia

More information and contact info

How important is to have independent test sets?

Collected **18** datasets used for building different peptide bioactivity predictors



Testing data should only include unseen molecules not similar to training (sequence alignment)

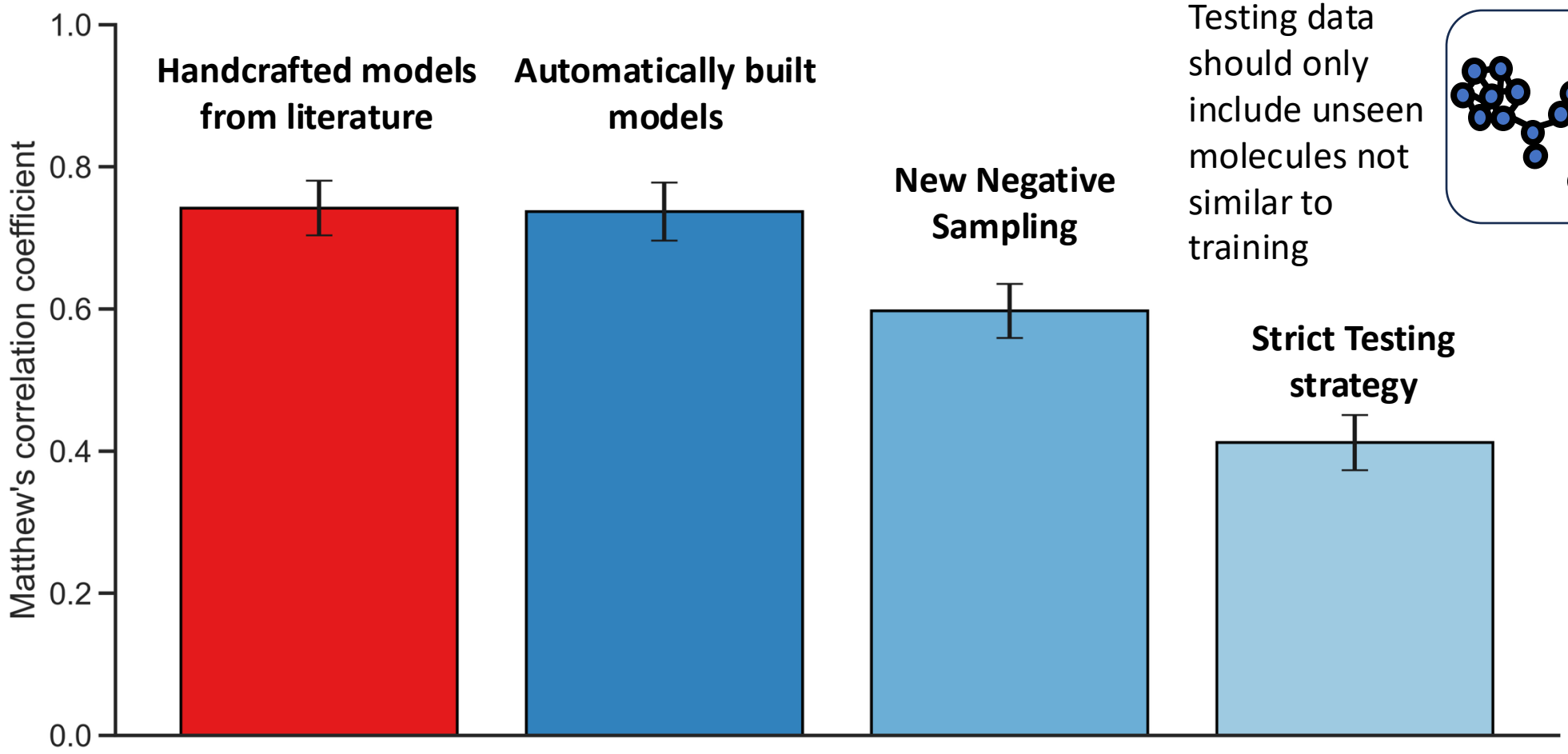
AutoPeptideML

R. Fernández-Díaz et al., AutoPeptideML: a study on how to build more trustworthy peptide bioactivity predictors, *Bioinformatics*, Volume 40, Issue 9, September 2024, btae555

More information and contact info

Independence in test sets is crucial

Collected **18** datasets used for building different peptide bioactivity predictors



Testing data should only include unseen molecules not similar to training

AutoPeptideML

R. Fernández-Díaz et al., AutoPeptideML: a study on how to build more trustworthy peptide bioactivity predictors, *Bioinformatics*, Volume 40, Issue 9, September 2024, btae555

More information and contact info

Automating ML for natural peptides

AutoPeptideML

R. Fernández-Díaz et al., AutoPeptideML: a study on how to build more trustworthy peptide bioactivity predictors, *Bioinformatics*, Volume 40, Issue 9, September 2024, btae555

Webserver - GUI

AutoPeptideML

BIOINFORMATICS PUBLICATION CHEMRXIV PREPRINT GITHUB REPOSITORY DOCUMENTATION

Welcome to the AutoPeptideML webserver.

The next steps will help you build your own model.

0. Modelling task

First, start by defining the prediction task.

What is the prediction problem you are facing?

Classification (categorical values)

1. Inputs

In this section you will define the data from which you want the model to learn.

Download sample dataset

Please upload dataset with your peptides and their labels if available

Drag and drop file here
Limit 200MB per file
Browse files

CLI tool

```
--------------------------
| AutoPeptideML v.2.0.3  |
| By Raul Fernandez-Diaz |
--------------------------

Model builder


Part 1 - Define the data and preprocessing steps
[?] What is the modelling problem you're facing?:
  > Classification (returning categorical value)
    Regression(returnin continuous value)
```

Python Package

```python
df = pd.read_csv(osp.join(PATH, 'original_data', f'c-{dataset}.csv'))
apml = AutoPeptideML(
    data=df,
    outputdir=f'apml-{dataset}',
    sequence_field='SMILES',
    label_field='labels'
)
apml.build_models(
    task='class',
    reps=['esm2-8m', 'peptideclm', 'chemberta-2', 'ecfp-16'],
    models=['svm', 'knn', 'rf', 'lightgbm', 'xgboost'],
    device='mps',
    n_trials=10
)
apml.create_report()
return apml
```

Part 2 – Peptide representation

More information and contact info

Can we leverage data on cheaper experiments to prioritise more expensive experiments?

ChemRxiv™

Fernández-Díaz R, et al. How to build machine learning models able to extrapolate from standard to modified peptides. ChemRxiv. 2025; doi:10.26434/chemrxiv-2025-ggp8n-v3

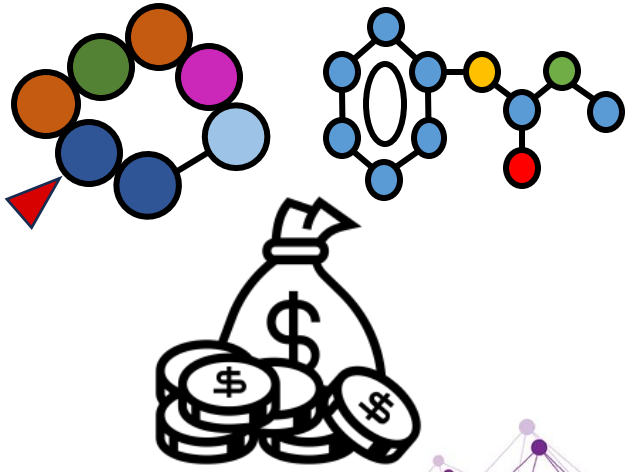
Natural peptides
(cheap)

Train

AutoPeptideML

Predict

Synthetic peptides and peptidomimetics
(expensive but better drugs)



SFI Centre for Research Training in GENOMICS DATA SCIENCE

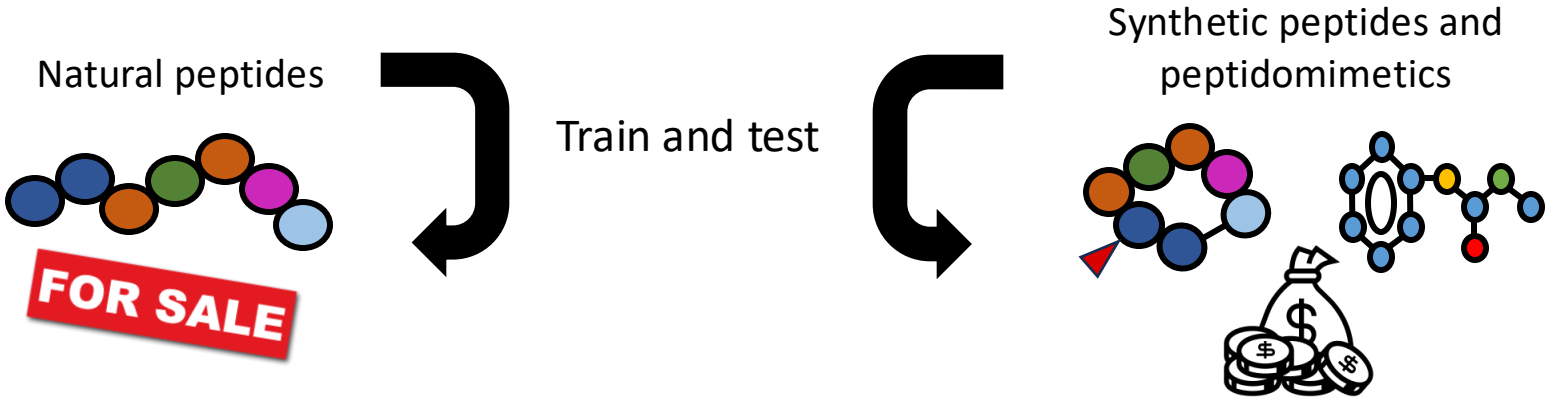
More information and contact info

Computational experiments

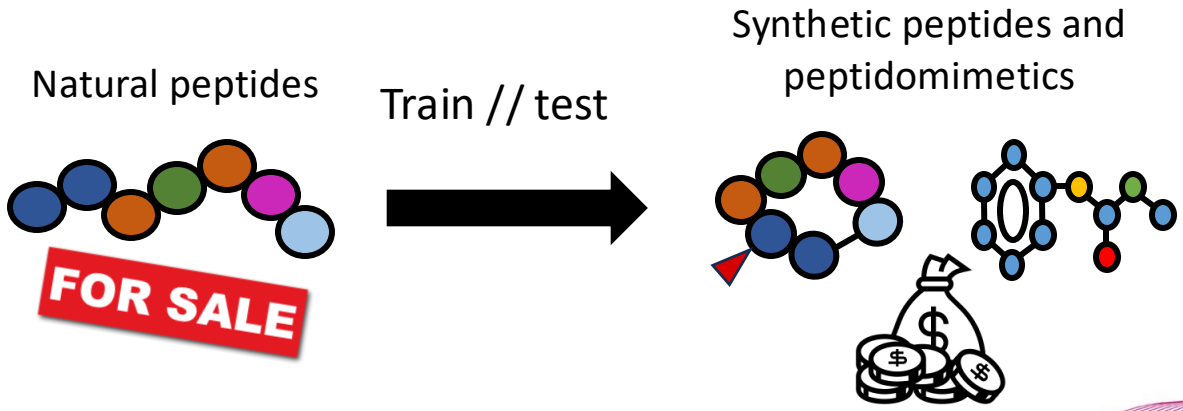
ChemRxiv™

Fernández-Díaz R, et al. How to build machine learning models able to extrapolate from standard to modified peptides. ChemRxiv. 2025; doi:10.26434/chemrxiv-2025-ggp8n-v3

Interpolation



Extrapolation



SFI Centre for Research Training in GENOMICS DATA SCIENCE

More information and contact info

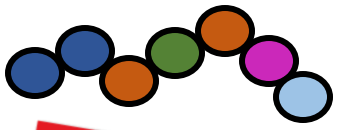
Computational experiments

ChemRxiv™

Fernández-Díaz R, et al. How to build machine learning models able to extrapolate from standard to modified peptides. ChemRxiv. 2025; doi:10.26434/chemrxiv-2025-ggp8n-v3

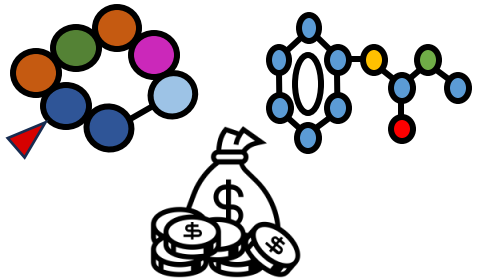
Interpolation

Natural peptides



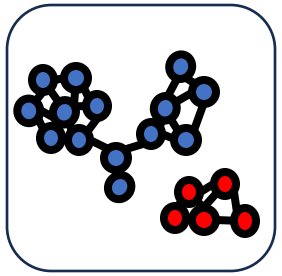
Train and test

Synthetic peptides and peptidomimetics



How to build train/test splits: Should we use sequence alignment for measuring peptide similarity?

Similarity-informed train/test split



SFI Centre for Research Training in GENOMICS DATA SCIENCE

More information and contact info

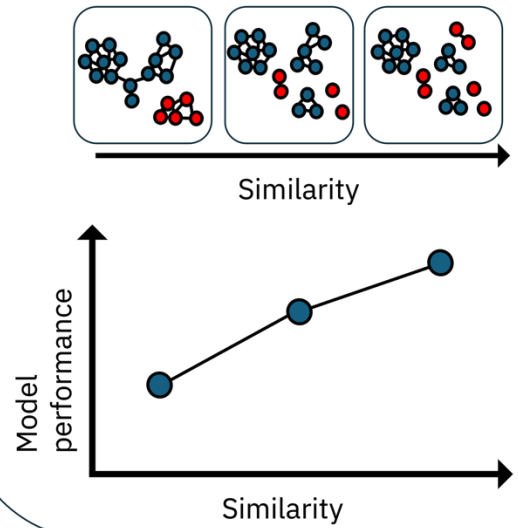
Finding the best similarity metric

Fernandez-Diaz R, et al. A new framework for evaluating model out-of-distribution generalisation for the biochemical domain. InThe Thirteenth International Conference on Learning Representations 2025.

Hestia-GOOD framework

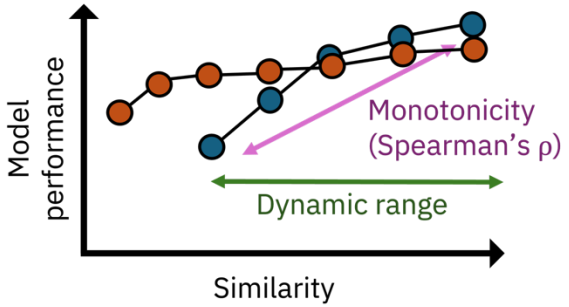
GOOD curve

Model performance as a function of train-test similarity



Similarity metric selection

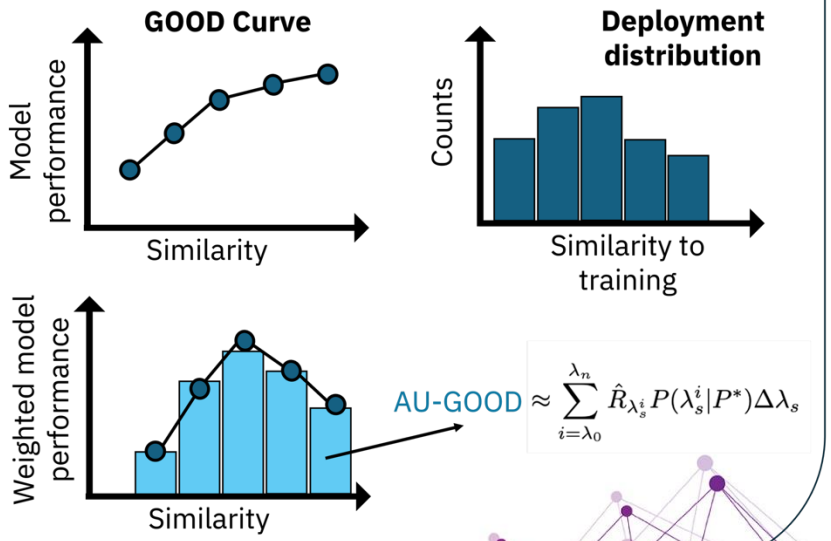
Quantitative analysis of best similarity function for a given task/dataset



- **Monotonicity:** Is model performance a function of train-test similarity?
- **Dynamic range:** What is the resolution of the similarity metric?

AU-GOOD metric

Estimation of model performance conditioned on a deployment distribution



SFI Centre for Research Training in GENOMICS DATA SCIENCE

More information and contact info

Best metrics for each dataset

ICLR 2025

Fernandez-Diaz R, et al. A new framework for evaluating model out-of-distribution generalisation for the biochemical domain. InThe Thirteenth International Conference on Learning Representations 2025.

Metrics explored:

- ECFP various radii
- MAPc various radii
- MMSeqs2 (alignment)
- Needleman-Wunsch (alignment)
- ESM2-8M embedding distance
- Molformer-XL embedding distance

- Analysis of **8** datasets:
  - 4 natural
  - 4 synthetic
- Chemical FPs are better than sequence alignment for natural peptides

SFI Centre for Research Training in GENOMICS DATA SCIENCE

More information and contact info

Best metrics for each dataset

ICLR 2025

Fernandez-Diaz R, et al. A new framework for evaluating model out-of-distribution generalisation for the biochemical domain. InThe Thirteenth International Conference on Learning Representations 2025.

Metrics explored:

- ECFP various radii
- MAPc various radii
- MMSeqs2 (alignment)
- Needleman-Wunsch (alignment)
- ESM2-8M embedding distance
- Molformer-XL embedding distance

- Analysis of **8** datasets:
  - 4 natural
  - 4 synthetic
- **Chemical FPs are better than sequence alignment for natural peptides**

Dataset	Peptide type	Task	Similarity Type	Similarity	Dynamic range (↑) [a]	Monotonicity (↑) [b]
Protein-peptide binding affinity	Standard	Regression	Chemical FP	MAPc-8	70 %	0.8 ± 0.1
Protein-peptide binding affinity	Modified	Regression	Chemical FP	MAPc-20	80 %	0.95 ± 0.03
Cell penetration	Standard	Classification	Chemical FP	MAPc-8	60 %	0.98 ± 0.04
Cell penetration	Modified	Classification	Chemical FP	MAPc-12	60 %	0.5 ± 0.2
Antibacterial	Standard	Classification	Chemical FP	MAPc-8	60 %	0.97 ± 0.02
Antibacterial	Modified	Classification	Chemical FP	ECFP-12	50 %	0.9 ± 0.1
Antiviral	Standard	Classification	Sequence Alignment	MMSeqs2	80 %	0.96 ± 0.05
Antiviral	Modified	Classification	Chemical FP	MAPc-12	70 %	0.6 ± 0.2

SFI Centre for Research Training in GENOMICS DATA SCIENCE

More information and contact info

Best metrics for each dataset

ICLR 2025

Fernandez-Diaz R, et al. A new framework for evaluating model out-of-distribution generalisation for the biochemical domain. InThe Thirteenth International Conference on Learning Representations 2025.

Metrics explored:

- ECFP various radii
- MAPc various radii
- MMSeqs2 (alignment)
- Needleman-Wunsch (alignment)
- ESM2-8M embedding distance
- Molformer-XL embedding distance

- Analysis of **8** datasets:
  - 4 natural
  - 4 synthetic
- **Chemical FPs are better than sequence alignment for natural peptides**

Dataset	Peptide type	Task	Similarity Type	Similarity	Dynamic range (↑) [a]	Monotonicity (↑) [b]
Protein-peptide binding affinity	Standard	Regression	Chemical FP	MAPc-8	70 %	0.8 ± 0.1
Protein-peptide binding affinity	Modified	Regression	Chemical FP	MAPc-20	80 %	0.95 ± 0.03
Cell penetration	Standard	Classification	Chemical FP	MAPc-8	60 %	0.98 ± 0.04
Cell penetration	Modified	Classification	Chemical FP	MAPc-12	60 %	0.5 ± 0.2
Antibacterial	Standard	Classification	Chemical FP	MAPc-8	60 %	0.97 ± 0.02
Antibacterial	Modified	Classification	Chemical FP	ECFP-12	50 %	0.9 ± 0.1
Antiviral	Standard	Classification	Sequence Alignment	MMSeqs2	80 %	0.96 ± 0.05
Antiviral	Modified	Classification	Chemical FP	MAPc-12	70 %	0.6 ± 0.2

SFI Centre for Research Training in GENOMICS DATA SCIENCE

More information and contact info

Best metrics for each dataset

ICLR 2025

Fernandez-Diaz R, et al. A new framework for evaluating model out-of-distribution generalisation for the biochemical domain. InThe Thirteenth International Conference on Learning Representations 2025.

Metrics explored:

- ECFP various radii
- MAPc various radii
- MMSeqs2 (alignment)
- Needleman-Wunsch (alignment)
- ESM2-8M embedding distance
- Molformer-XL embedding distance

- Analysis of **8** datasets:
  - 4 natural
  - 4 synthetic
- **Chemical FPs are better than sequence alignment for natural peptides**

Dataset	Peptide type	Task	Similarity Type	Similarity	Dynamic range (↑) [a]	Monotonicity (↑) [b]
Protein-peptide binding affinity	Standard	Regression	Chemical FP	MAPc-8	70 %	0.8 ± 0.1
Protein-peptide binding affinity	Modified	Regression	Chemical FP	MAPc-20	80 %	0.95 ± 0.03
Cell penetration	Standard	Classification	Chemical FP	MAPc-8	60 %	0.98 ± 0.04
Cell penetration	Modified	Classification	Chemical FP	MAPc-12	60 %	0.5 ± 0.2
Antibacterial	Standard	Classification	Chemical FP	MAPc-8	60 %	0.97 ± 0.02
Antibacterial	Modified	Classification	Chemical FP	ECFP-12	50 %	0.9 ± 0.1
Antiviral	Standard	Classification	Sequence Alignment	MMSeqs2	80 %	0.96 ± 0.05
Antiviral	Modified	Classification	Chemical FP	MAPc-12	70 %	0.6 ± 0.2

SFI Centre for Research Training in GENOMICS DATA SCIENCE

More information and contact info

Interpolation experiments

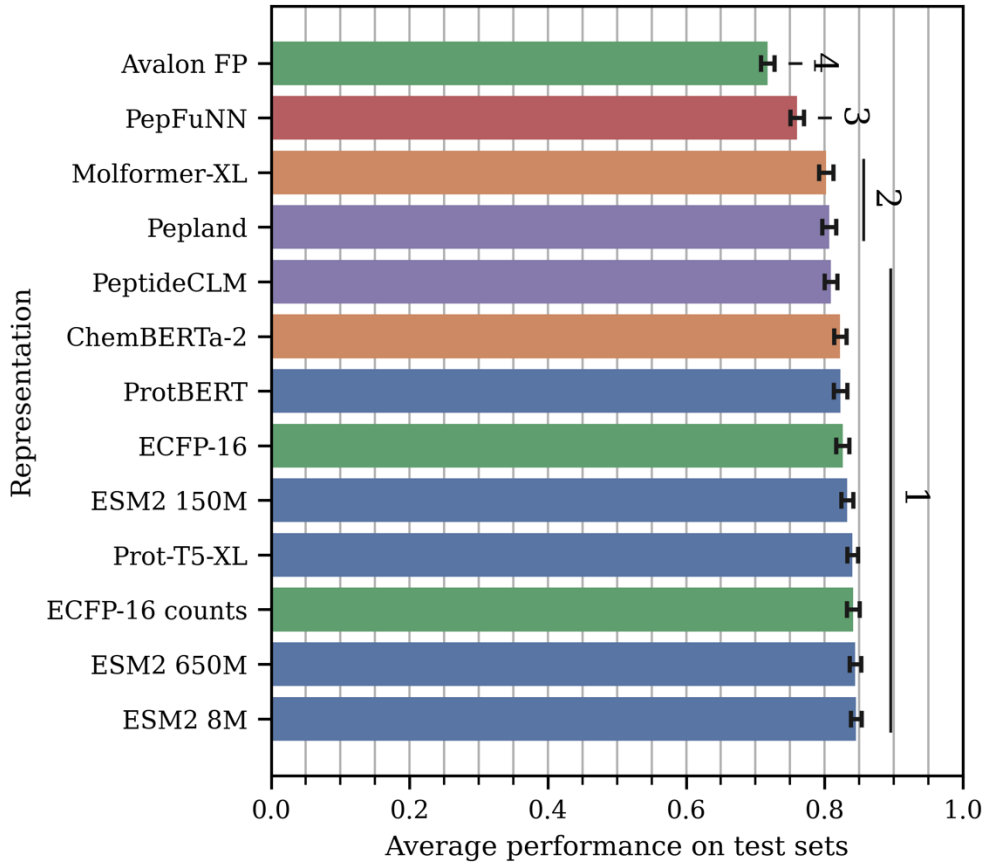
ChemRxiv™

Fernández-Díaz R, et al. How to build machine learning models able to extrapolate from standard to modified . ChemRxiv. 2025; 134/chemrxiv-2025- gp8n-v3

- Average on **4** datasets.
- Statistical analysis are Kruskal-Wallis with *post-hoc* Wilcoxon test.
- Significance defined with Bonferroni correction.

Natural to natural

Synthetic to synthetic



Representation Family

PLM | CLM | Chemical FP | Peptide FP | Peptide LM/GNN

SFI Centre for Research Training in GENOMICS DATA SCIENCE

More information and contact info

Interpolation experiments

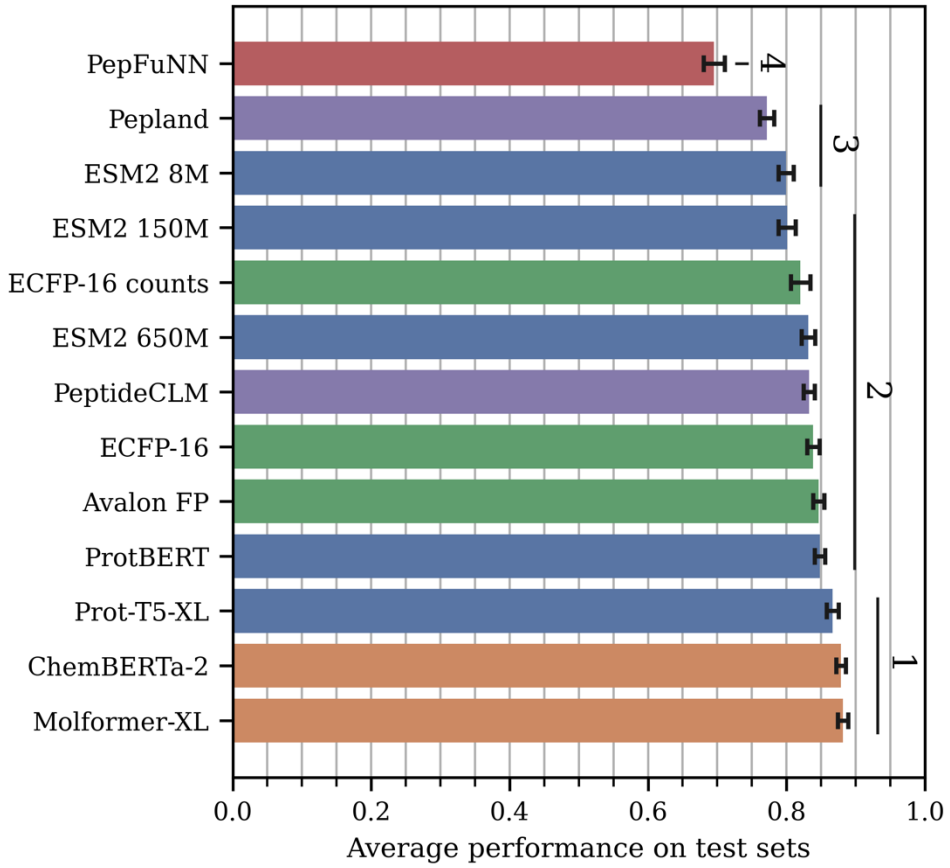
ChemRxiv™

Fernández-Díaz R, et al. How to build machine learning models able to extrapolate from standard to modified . ChemRxiv. 2025; 134/chemrxiv-2025-gp8n-v3

- Average on **4** datasets.
- Statistical analysis are Kruskal-Wallis with *post-hoc* Wilcoxon test.
- Significance defined with Bonferroni correction.

Natural to natural

Synthetic to synthetic



SFI Centre for Research Training in GENOMICS DATA SCIENCE

Representation Family

PLM CLM Chemical FP Peptide FP Peptide LM/GNN

More information and contact info

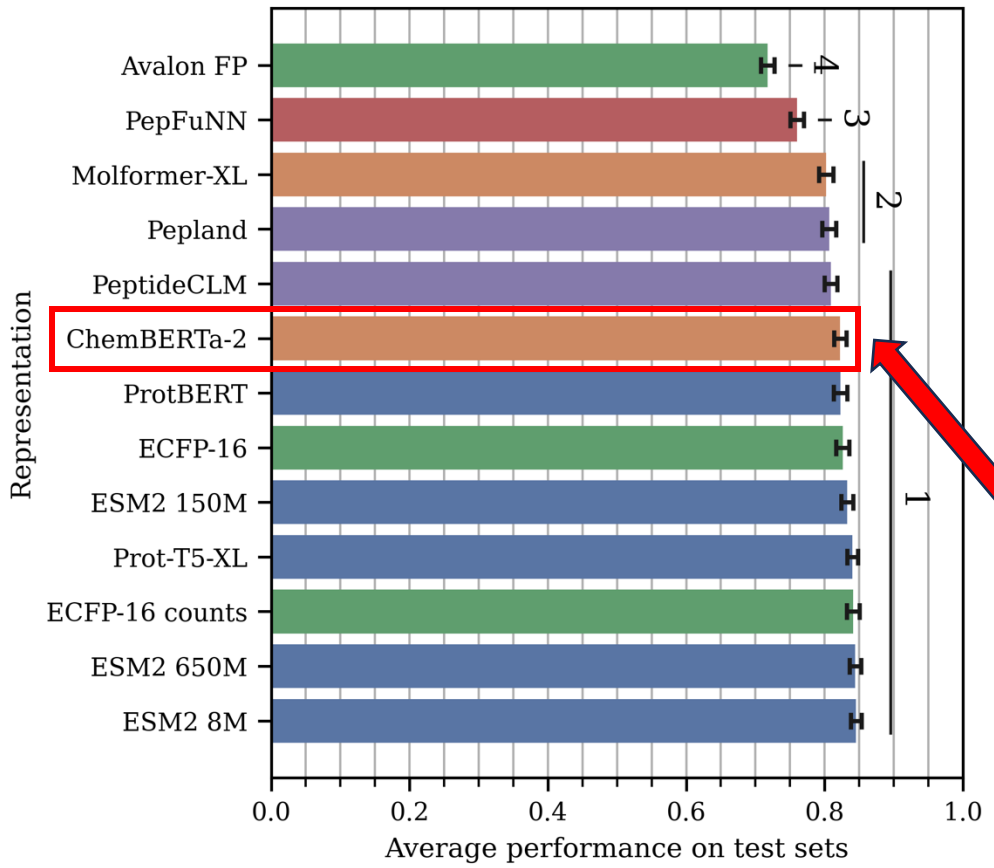
Interpolation experiments

ChemRxiv™

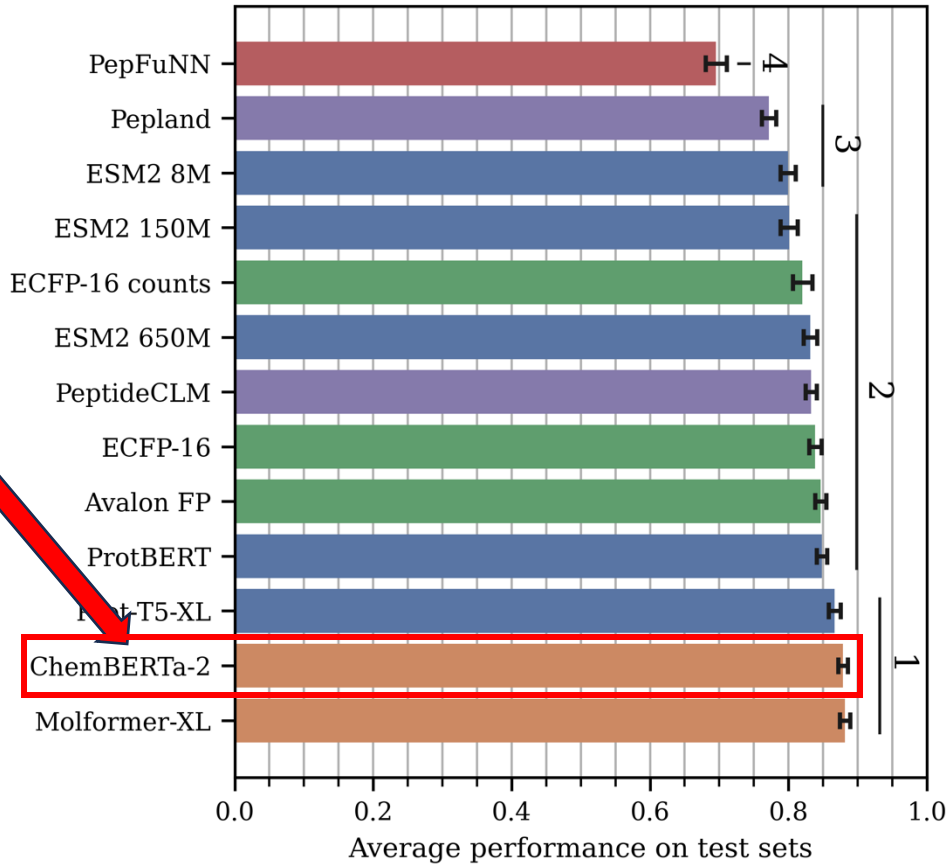
Fernández-Díaz R, et al. How to build machine learning models able to extrapolate from standard to modified. ChemRxiv. 2025; 134/chemrxiv-2025-gp8n-v3

- Average on **4** datasets.
- Statistical analysis are Kruskal-Wallis with *post-hoc* Wilcoxon test.
- Significance defined with Bonferroni correction.

Natural to natural



Synthetic to synthetic



Representation Family

PLM | CLM | Chemical FP | Peptide FP | Peptide LM/GNN

SFI Centre for Research Training in GENOMICS DATA SCIENCE

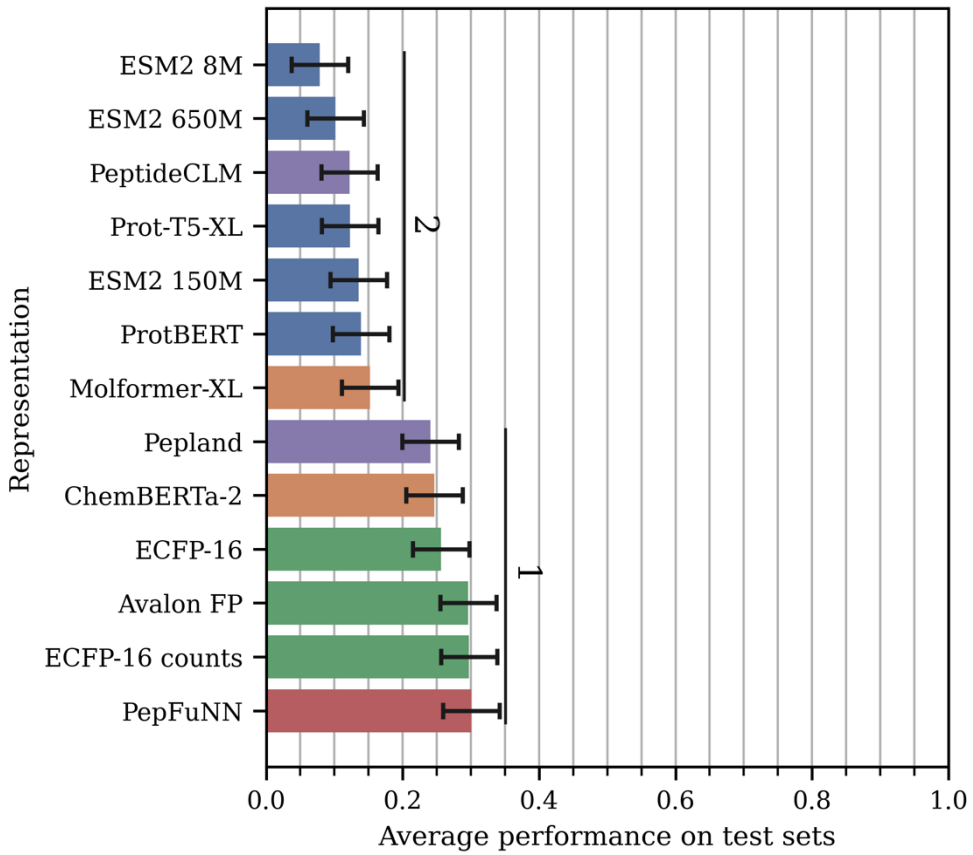
More information and contact info

Natural to synthetic extrapolation

ChemRxiv™

Fernández-Díaz R, et al. How to build machine learning models able to extrapolate from standard to modified peptides. ChemRxiv. 2025; doi:10.26434/chemrxiv-2025-ggp8n-v3

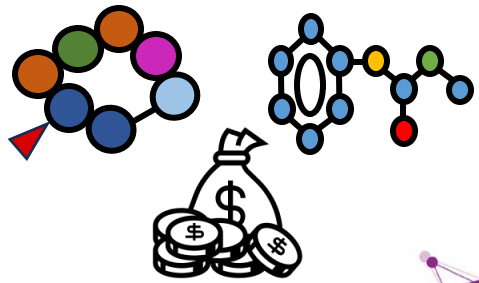
- Average on **4** datasets.
- Statistical analysis are Kruskal-Wallis with *post-hoc* Wilcoxon test.
- Significance defined with Bonferroni correction.



Natural peptides

Train // test

Synthetic peptides and peptidomimetics



Representation Family

PLM CLM Chemical FP Peptide FP Peptide LM/GNN

SFI Centre for Research Training in GENOMICS DATA SCIENCE

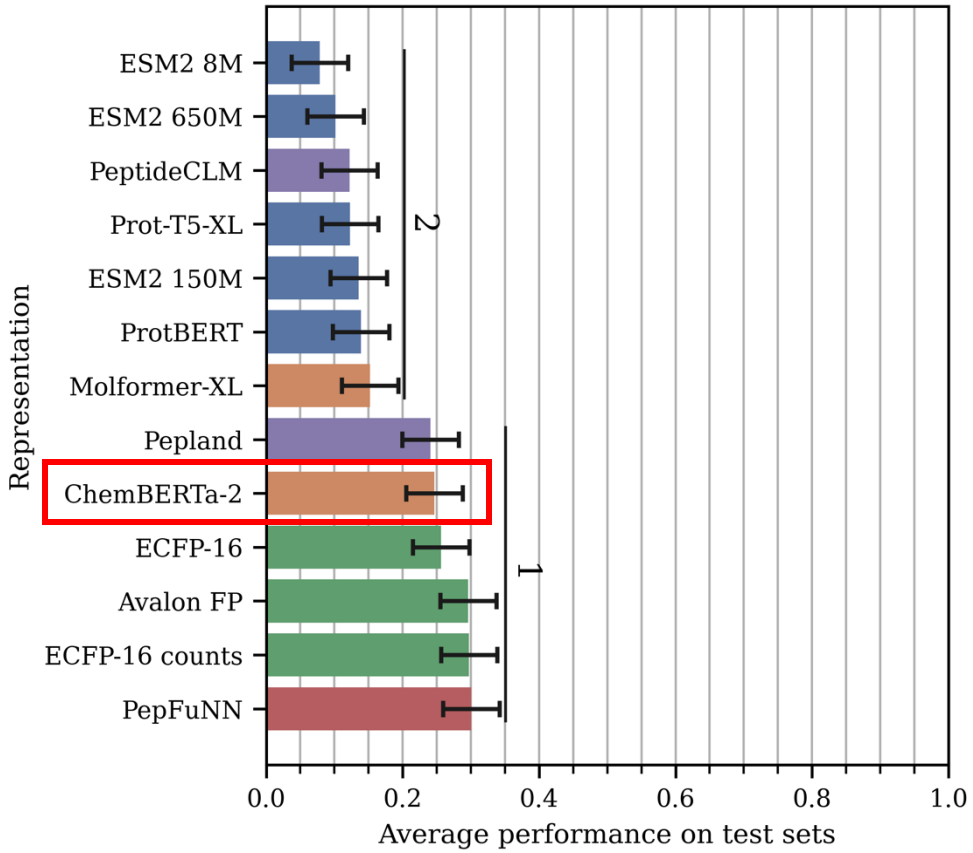
More information and contact info

Natural to synthetic extrapolation

ChemRxiv™

Fernández-Díaz R, et al. How to build machine learning models able to extrapolate from standard to modified peptides. ChemRxiv. 2025; doi:10.26434/chemrxiv-2025-ggp8n-v3

- Average on **4** datasets.
- Statistical analysis is ANOVA with *post-hoc* Tukey's HSD



Representation Family

PLM CLM Chemical FP Peptide FP Peptide LM/GNN

Conclusions

1. Natural to synthetic extrapolation is possible, but models are less reliable
2. ChemBERTa-2 appears to be the most versatile tool to work with peptides
3. Chemical and Peptide Fingerprints are robust options as well

SFI Centre for Research Training in GENOMICS DATA SCIENCE

More information and contact info

Conclusions

1. Dataset building (negative definition) and partitioning (train/test split) are key for proper model evaluation
2. Chemical fingerprints are better for partitioning natural and synthetic datasets than sequence alignment, contrary to standard practice
3. Natural to synthetic extrapolation is possible, but there is room for improvement. We release the benchmark datasets to make it easier for the community to improve
4. ChemBERTa-2 appears to be the most versatile tool, closely followed by chemical fingerprints

SFI Centre for Research Training in GENOMICS DATA SCIENCE

Contact info, papers, and slides of the presentation

More information and contact info

SFI Centre for Research Training in GENOMICS DATA SCIENCE

SFI Centre for Research Training in GENOMICS DATA SCIENCE

IBM

Centre for Research Training

Science Foundation Ireland sfi For what's next

UCD DUBLIN University College Dublin University for All

Partitioning, representation, and automation in canonical and non-canonical peptide modelling

Speaker: **Raúl Fernández-Díaz (PhD Candidate UCD – IBM Research)**

UCD: R. Cossio-Pérez, C. Agoni, D.C. Shields

IBM Research: T.L. Hoang, V. Lopez

Novo Nordisk: R. Ochoa

More information and contact info