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

Evaluation of partitioning algorithms for trustworthy out-of-distribution evaluation of machine learning models in biochemistry

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

UCD: D.C. Shields

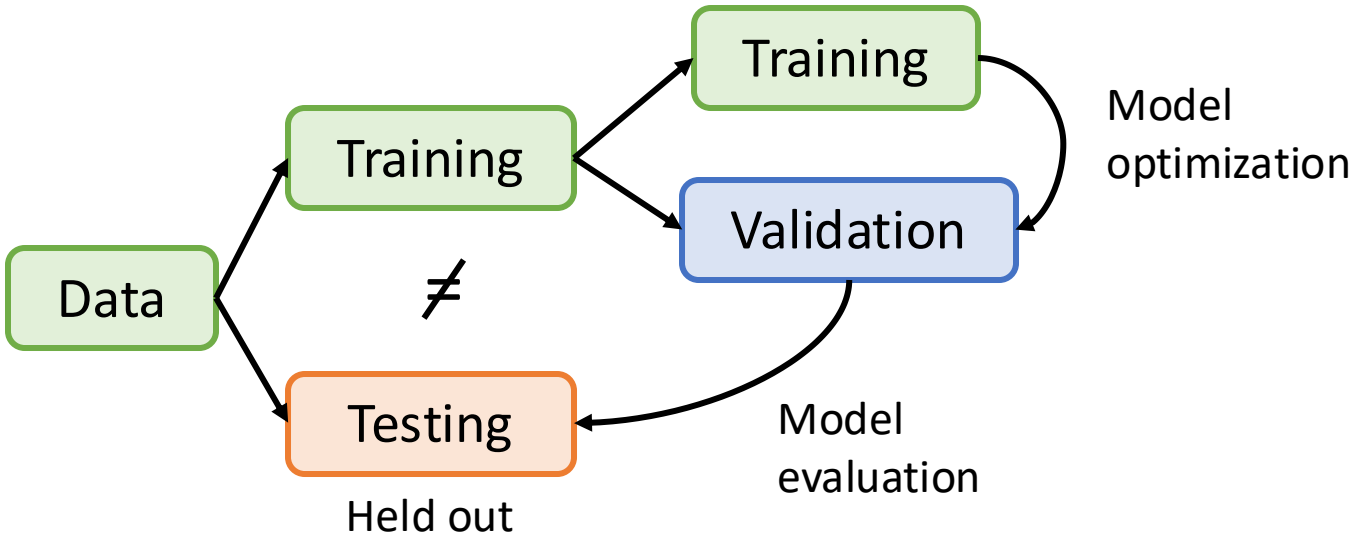
IBM Research: T.L. Hoang, V. Lopez

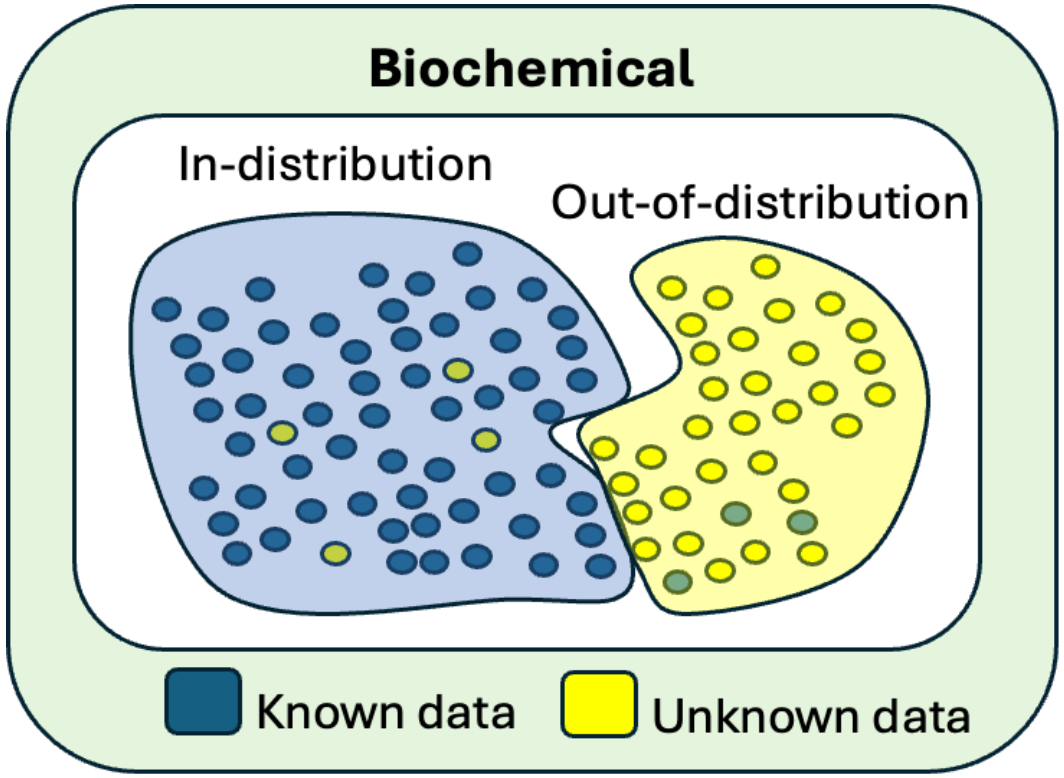
Part 0 - Introduction

More information and contact info

Dataset partitioning

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





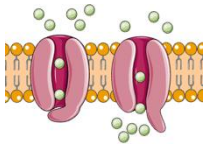
Dataset partitioning algorithms: Ensure that training and testing splits have different molecules, so that we can evaluate generalization/extrapolation

Objective: Estimate model performance in unseen data

More information and contact info

Use case: Pharmacological property prediction

Therapeutics Data Commons:
Collection of 22 datasets covering different pharmacological properties



Absorption

Distribution

Metabolism

Excretion

Toxicity

Research Question 1:

How important is the choice of **partitioning algorithm**?

We consider 8 partitioning algorithms:

1. Random (Baseline)
2. Scaffold (~ 2015)
3. CCPart (Ours)
4. Butina (1999)
5. UMAP (2024)
6. Sim-UMAP (Adapted UMAP)
7. CD-HIT-Part (Adapted CD-HIT)
8. BitBIRCH (2024)

Research Question 2:

How important is the definition of **molecular similarity**?

We consider 16 similarity metrics:

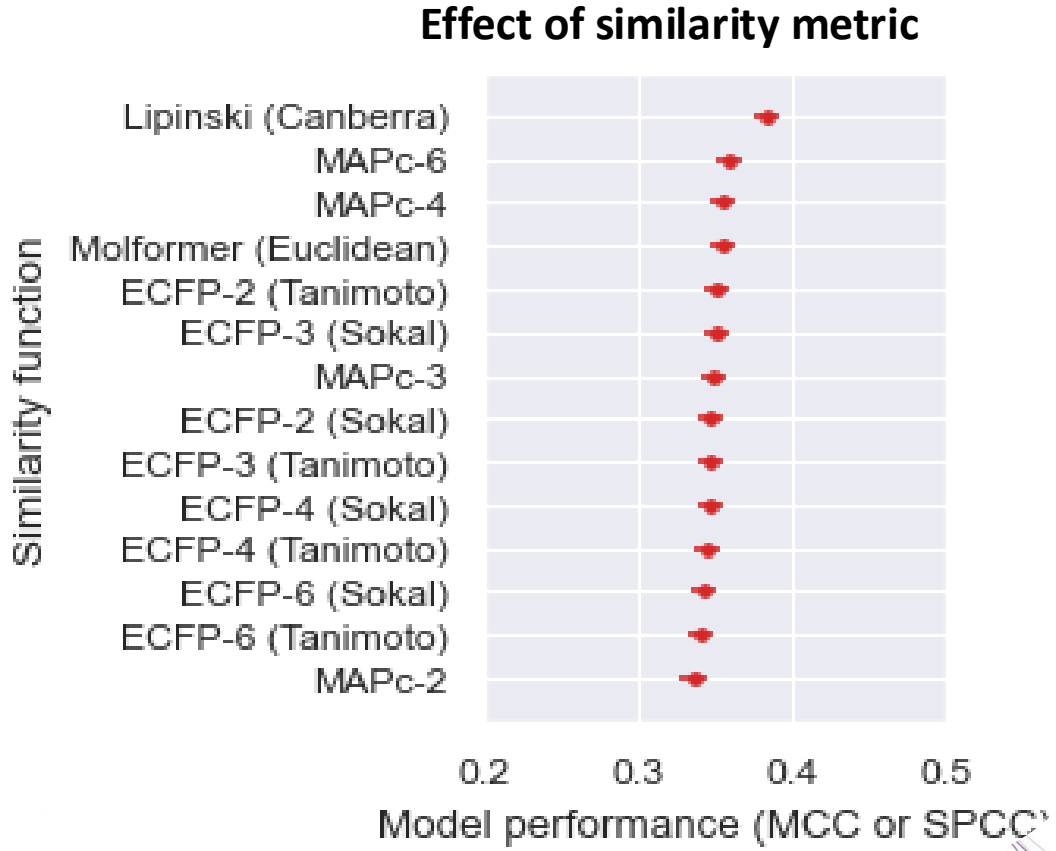
1. Tanimoto ECFP-2, 3, 4, and 6
2. Sokal ECFP-2, 3, 4, and 6 (199
3. Jaccard MAPc-2, 3, 4, and 6 (2024)
4. Canberra Lipinski vectors (Ours)
5. Euclidean Molformer-XL (Ours)

Results

More information and contact info

Effect of similarity metrics

1. Effect is statistically significant (Repeated measures ANOVA $p<0.05$
2. There is no significant difference between ECFP and MAPc fingerprints
3. Among the same fingerprint family
   a) In ECFP radius is not important
   b) In MAPc it is
4. Overall, magnitude of effect is small (around 3%)

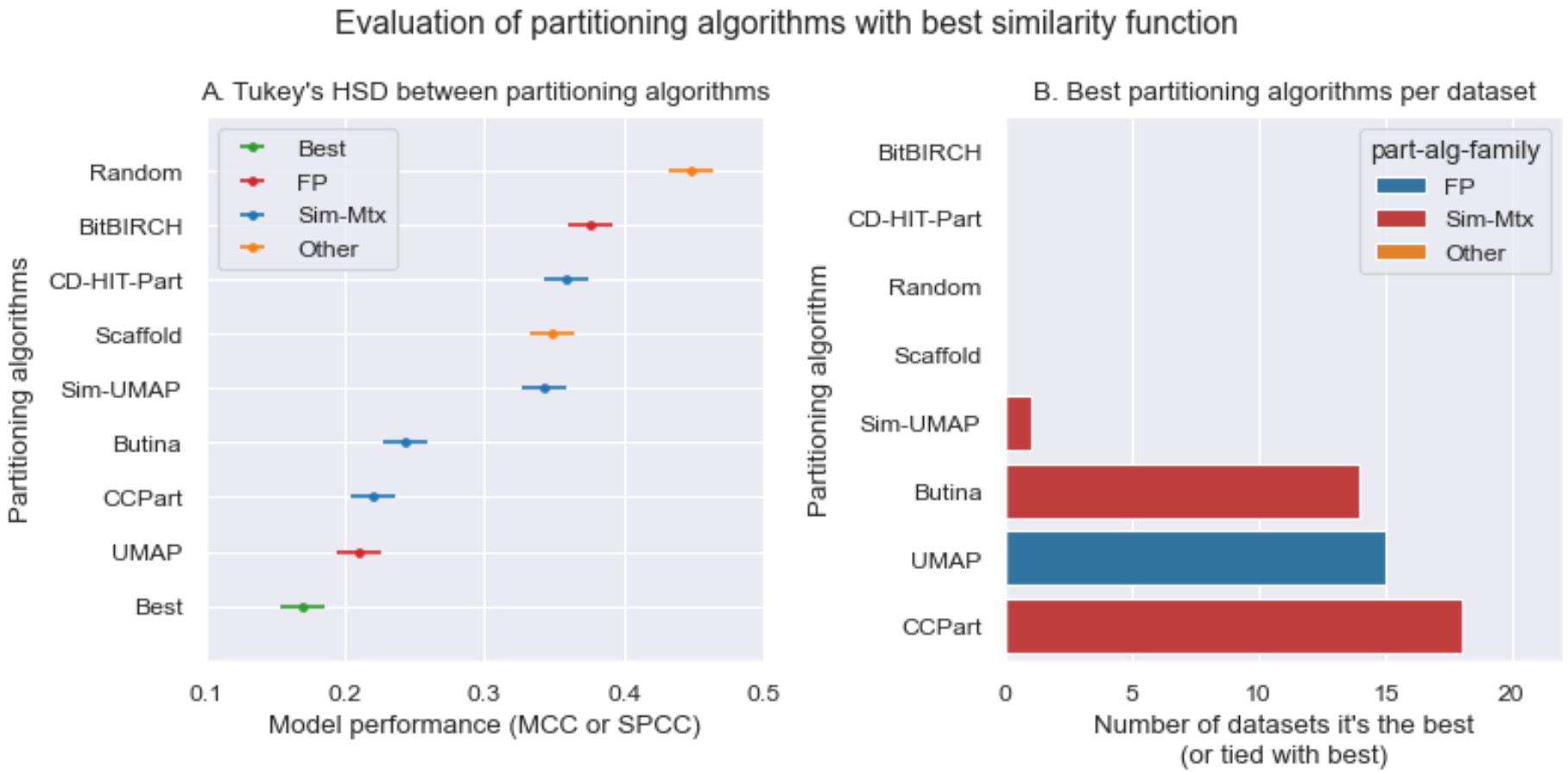


SFI Centre for Research Training in GENOMICS DATA SCIENCE

More information and contact info

Effect of partitioning algorithm

1. Random tends to overestimate model performance; so does scaffold (current standard)
2. Top 3 methods:
   a) CCPart
   b) Butina
   c) UMAP
3. Each dataset requires an optimal method (“Best”)



SFI Centre for Research Training in GENOMICS DATA SCIENCE

More information and contact info

Next steps: AutoHestia

Find automatically best way to partition a dataset

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
```

(Images are from another project, the idea is to design similar interfaces)

Conclusions

More information and contact info

Conclusions

1. Similarity partitioning is fundamental to avoid overestimating model performance
2. Each dataset requires a different combination of similarity metric and partitioning algorithm
3. We are building AutoHestia, a Python package and webserver, to automate the search for similarity metric and partitioning algorithm for any new dataset

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

Evaluation of partitioning algorithms for trustworthy out-of-distribution evaluation of machine learning models in biochemistry

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

UCD: D.C. Shields

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

More information and contact info