AutoPeptideML 2: An open source library for democratizing machine learning for peptide bioactivity prediction

Raúl Fernández-Díaz*, Thanh Lam Hoang, Vanessa Lopez, Denis Shields

University College Dublin

IBM Research

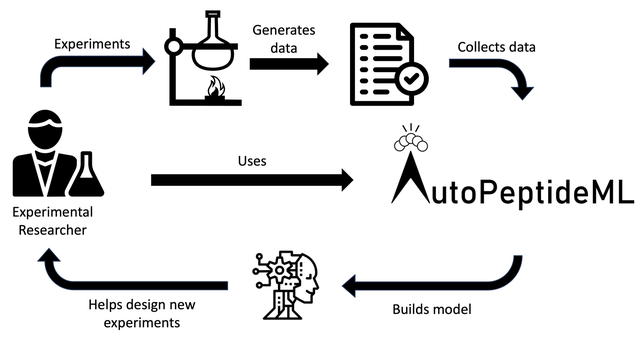
Introduction

Motivation

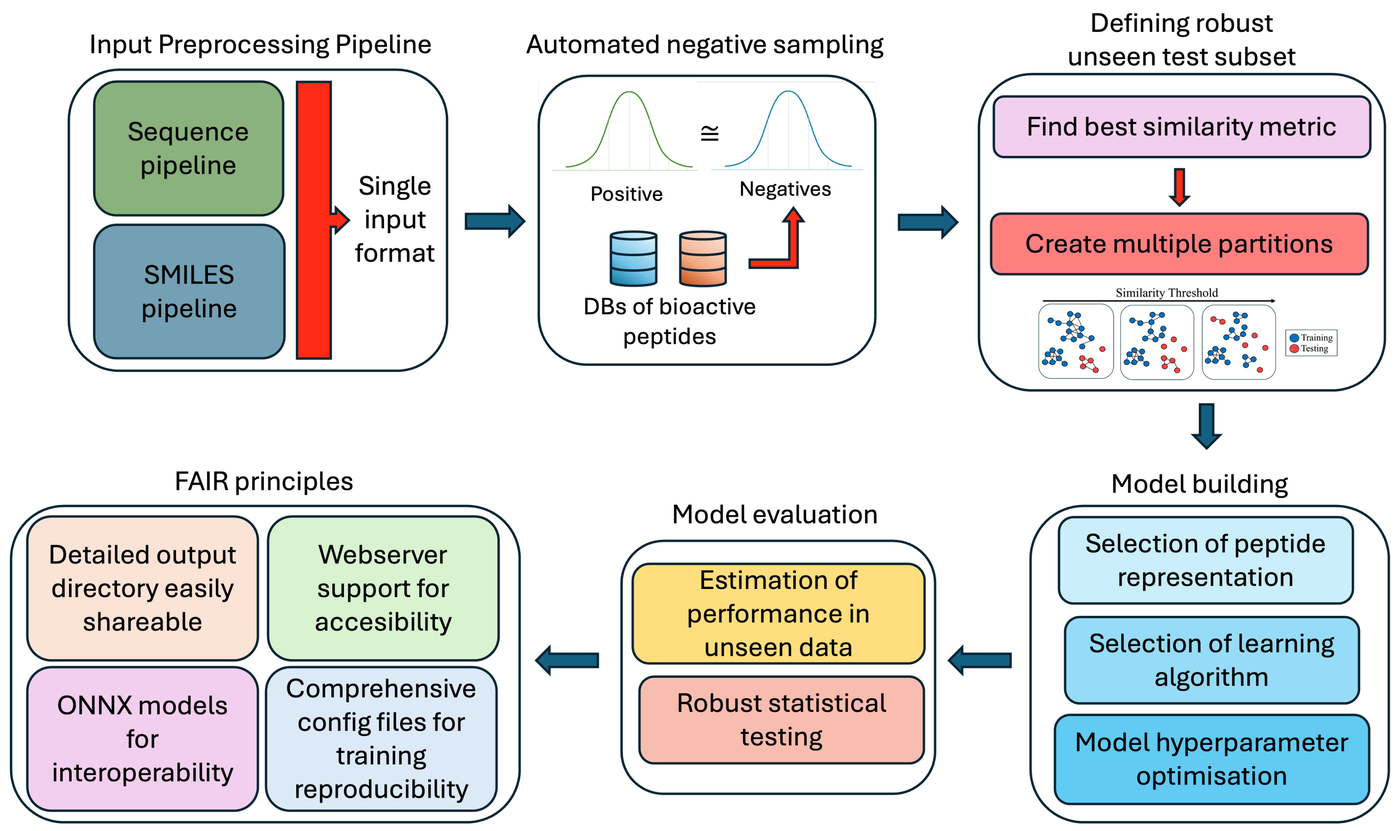
Integration of machine learning modelling to experimental workflows in peptide drug discovery/design can accelerate research.

Objectives

1. Empower experimental researchers to build custom models.
2. Generate output that follows FAIR principles.
3. Handle canonical, non-canonical peptides, and peptidomimetic molecules.

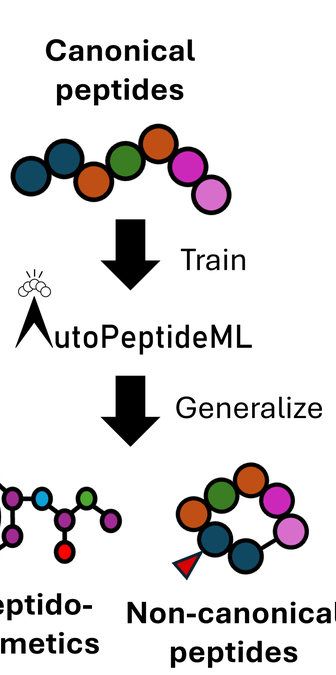


Modules

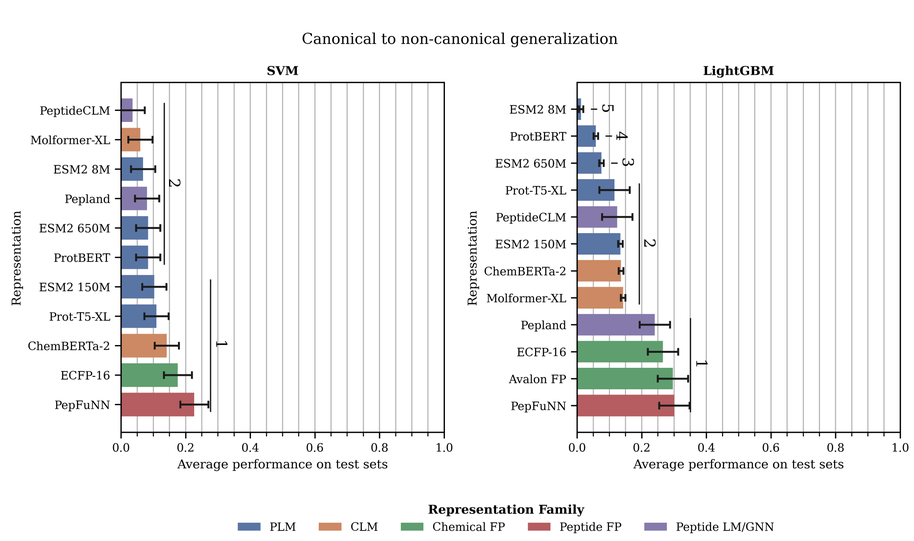


Canonical to non-canonical generalization

Overview



Results



Data corresponds to average model performance on 4 different datasets: antibacterial, antiviral, cell penetration, protein-peptide binding affinity

Conclusions

A. AutoPeptideML is a flexible tool that makes model-building easier and handles:
   1. Data preprocessing
   2. Negative sampling
   3. Dataset partitioning
   4. Model and representation selection
   5. Model hyperparameter optimization
   6. Model evaluation
   7. Training and evaluation PDF report

B. Its design facilitates compliance with FAIR principles and DOME guidelines.

C. Handle canonical, non-canonical peptides, and peptidomimetic small molecules.

D. Facilitates development of new methods of peptide representation by providing robust and reproducible workflows and evaluation

Dependencies

Data preparation

pandas RDKit

Model training

OPTUNA scikit learn

Evaluation + Report

Hestia GOOD quarto ONNX

Webserver

Streamlit FastAPI

Easy-to-use interfaces

Webserver

AutoPeptideML

BIOINFORMATICS PUBLICATION CHEMRXIV PREPRINT GITHUB REPOSITORY DOCUMENTATION

AutoPeptideML Model builder

Welcome to the AutoPeptideML webserver.

The next steps will help you build your own model.

0. Modelling task

First, start by defining the modelling task.

What is the modeling problem you are facing?

Classification (returning categorical values)

How do you want to define your peptides?

Macromolecules - allows for canonical, non-canonical, and peptidomimetics

Next step

http://peptide.ucd.ie/autopeptideml

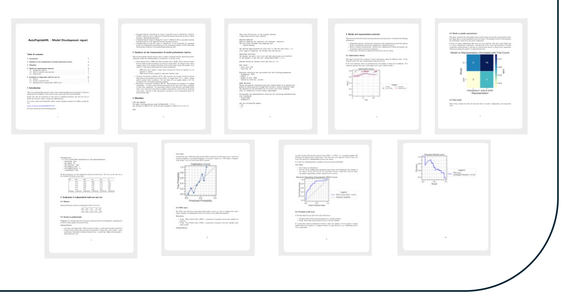
Python code

```python
import pandas as pd
from autopeptideml import AutoPeptideML

# Load data
df = pd.read_csv('antibacterial_data_canonical.csv')
df2 = pd.read_csv('antibacterial_data_noncanonical.csv')
all_inputs = df['sequence'].tolist() + df['SMILES'].tolist()

# Initialise AutoPeptideML
apml = AutoPeptideML(
    data=all_inputs,
    outputdir='demo'
)

# Preprocess
apml.preprocess_data(
    pipeline='to-smiles',
    n_jobs=5,
    verbose=True
)

# Build models
apml.build_models(split_strategy='min',
                  task='class',
                  reps=['chemberta-2', 'ecfp',
                        'esm2-8m',
                        'peptideclm'],
                  device='mps',
                  n_trials=200)
```

PDF report

AutoPeptideML - Model Development report

Table of contents



Contact info:

Raúl Fernández-Díaz*

ISMB/ECCB 2025 JULY 20–24 LIVERPOOL

33rd CONFERENCE ON Intelligent Systems for Molecular Biology

AND THE 24th European Conference on Computational Biology

DATES: July 20 – 24

Github Repository: