AI/ML for prediction of biological properties of molecules

Module 4. The Ersilia Model Hub

Gemma Turon & Miquel Duran-Frigola
Ersilia Open Source Initiative (www.ersilia.io)
18th - 27th of September, 2023

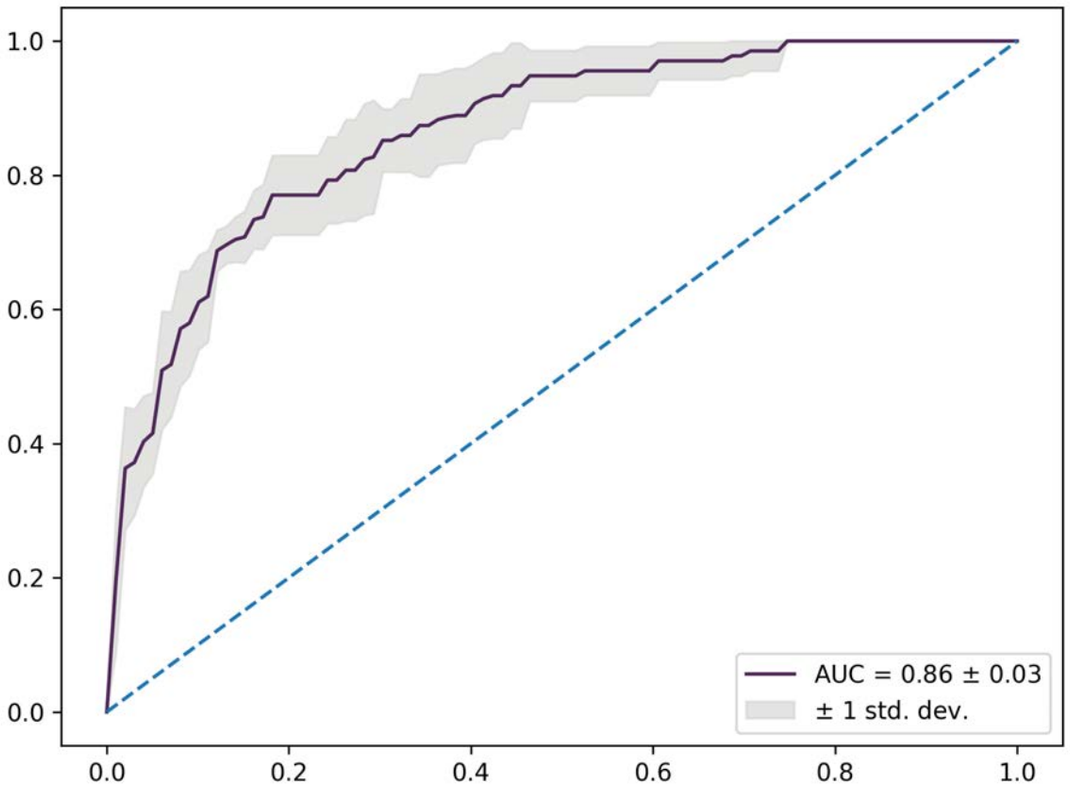
Ersilia

Our new models

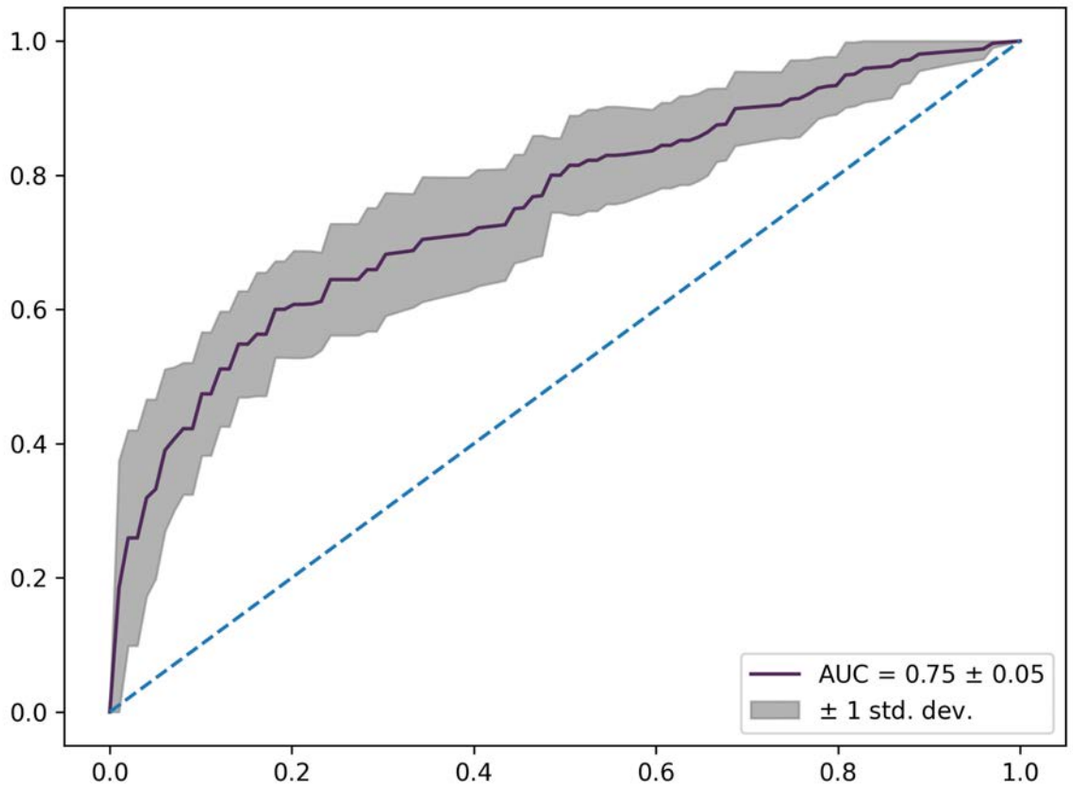
MycetOS

— Predicts the % of growth of mycetoma when incubated with 25 uM of the compound
— Cut-off: 20% growth

Morgan fingerprints
5-fold cross validation (train-test split 20%)



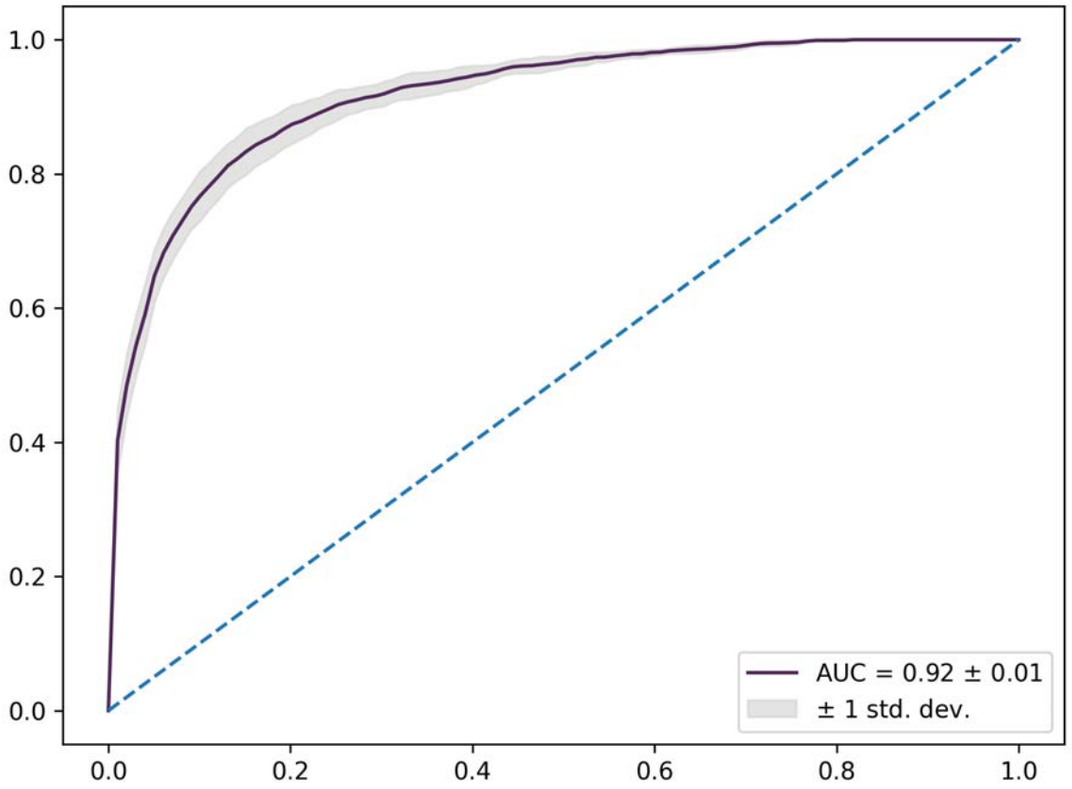
Ersilia embeddings
5-fold cross validation (train-test split 20%)



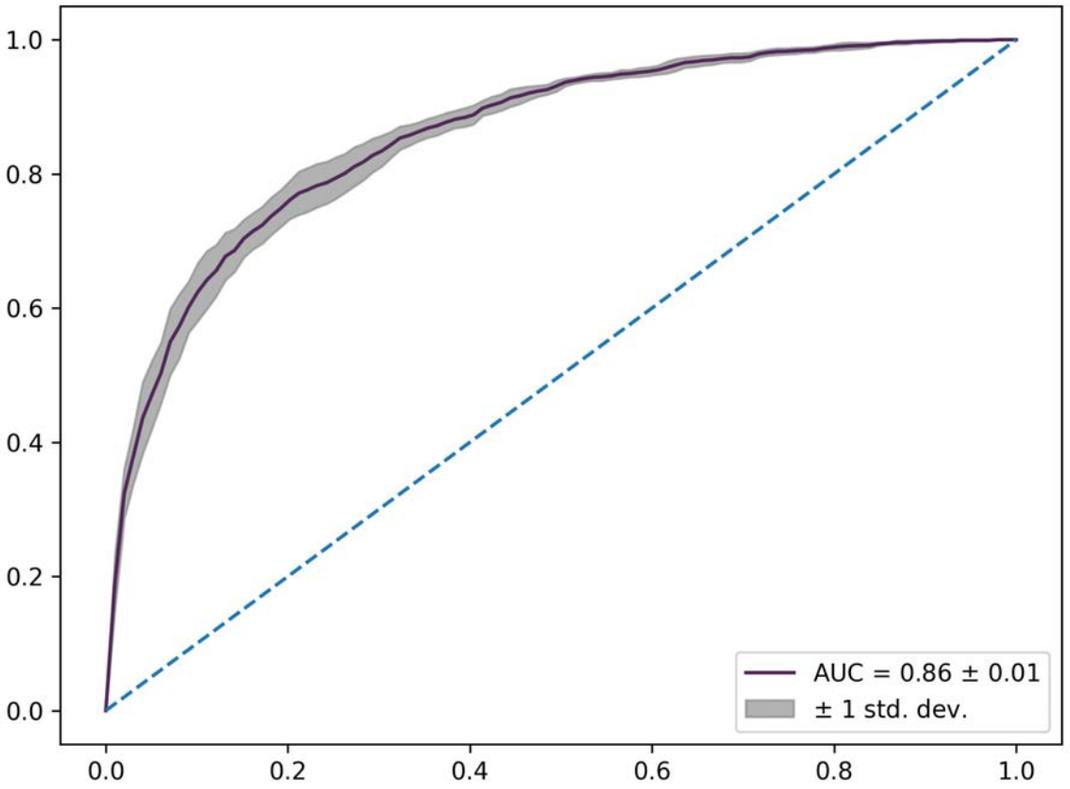
HDAC1

— Predicts the inhibition of HDAC1 (pChEMBL)
— Cut-off: 7 (higher compounds)

Morgan fingerprints
5-fold cross validation (train-test split 20%)

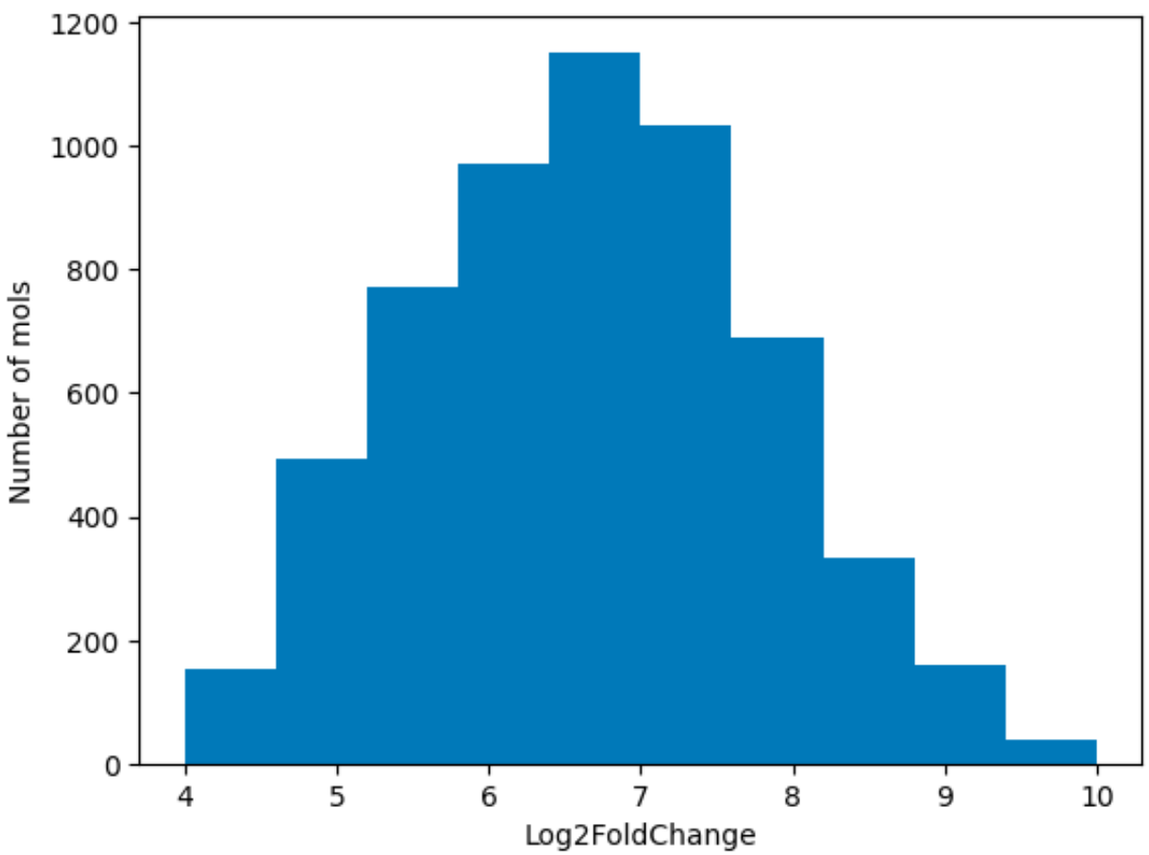


Ersilia embeddings
5-fold cross validation (train-test split 20%)

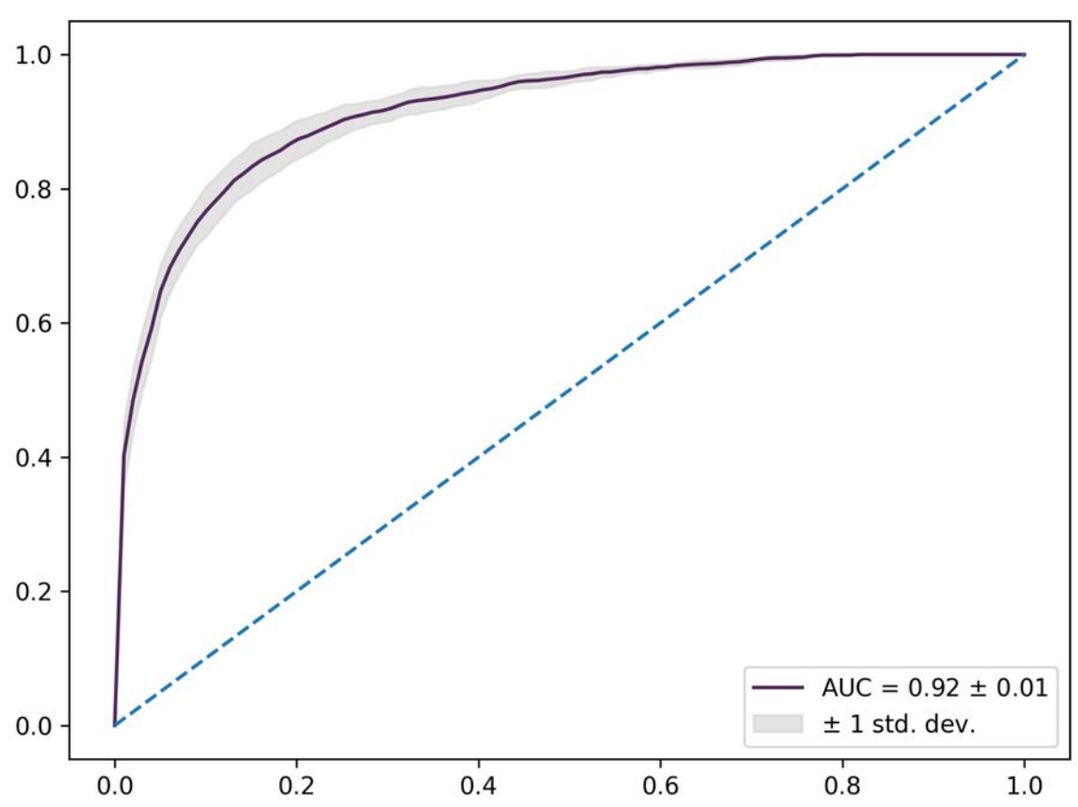


HDAC1

— Predicts the inhibition of HDAC1 (pChEMBL)
— Cut-off: 8 (higher compounds)



Morgan fingerprints
5-fold cross validation (train-test split 20%)



ACE2-Spike interaction

— Predicts the inhibition of the ACE2-Spike interaction in SARS-CoV-2
— Cut-off: -1

We need to work on the under sampling strategy to correct the class imbalance found in our dataset (with only 46 actives and over 2000 molecules in total)

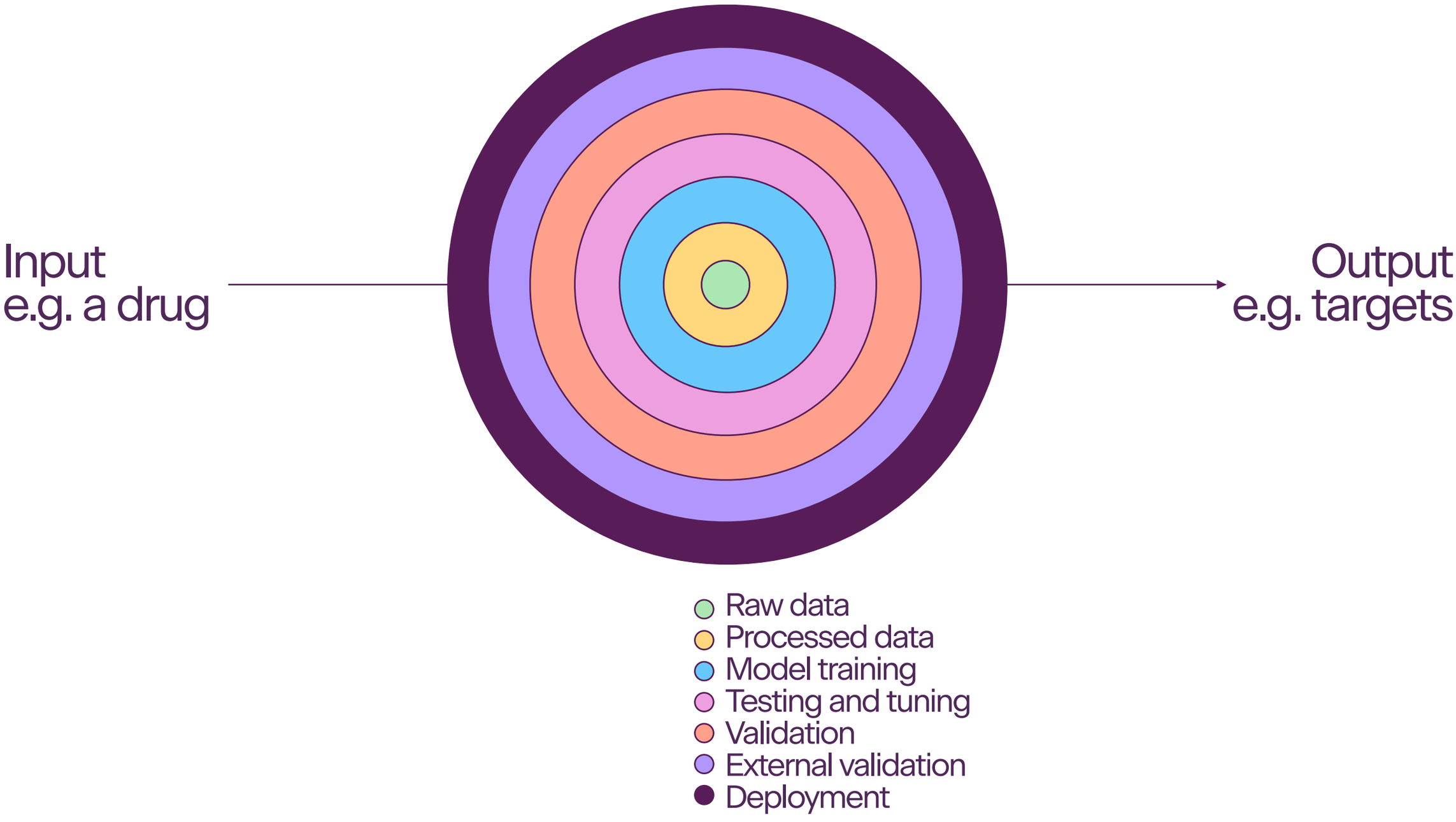
What to consider to continue using and developing these models

— Can we gather more data? Ideally, from our own or colleagues work?
— External validation of the models (labelled data from other sources, for example)
— Can we make subset models? (For example, for HDAC1)
— When making predictions, are our molecules already present in the training set?

How to run saved models locally

```
import pandas as pd
import joblib

# get your molecules of interest, for example from a .csv

df = pd.read_csv(“mynewdata.csv")

smiles = df["SMILES"]

# load the model using joblib

model = joblib.load(“mymodel.joblib")

my_preds = model.predict_proba(smiles)
```

Our goal: to provide ready-to-use AI models



Welcome to the Ersilia Model Hub!

https://ersilia.io/model-hub

Type to search model...

Tags

- Tox21
- Toxicity
- MoleculeNet
- Grover
- Graph Transformer

Output

- Antibiotic activity
- Toxicity
- Synthetic accessibility
- Antiviral activity
- Target

Mode

- Pretrained
- Retrained
- In-house
- Online

License

Carcinogenic potential of metabolites and small molecules

eos1579 metabokiller

Carcinogenicity is a result of several potential effects on cells. This model predicts the carcinogenic potential of a small molecule based on their potential to induce cellular proliferation, genomic instability, oxidative stress, anti-apoptotic responses and epigenetic alterations.
Metabokiller uses the Chemical Checker signaturizer to featurize the molecules, and the Lime package to provide interpretable results.
Using Metabokiller, the authors screened a panel of human metabolites and experimentally demonstrated two of the predicted carcinogenic metabolites induced carcinogenic transformations in yeast and human cells.

Molecular maps based on broadly learned knowledge-based representations

eos6m4j bidd-molmap

Descriptor-based or fingerprint-based molecular maps (images) are created. Typically, the goal is to use these images as inputs for an image-based deep learning model such as a convolutional neural network

SMILES transformer descriptor

eos2lm8 smiles-transformer

Molecular fingerprint based on natural language processing. It converts SMILES into fingerprints using an unsupervised model pre-trained on a very large SMILES dataset. The transformer is particularly well-suited for low-data drug discovery

Our models!

Ersilia

Model Information

Description

This model predicts the antimalarial potential of small molecules in vitro. We have collected the data available from the Open Source Malaria Series 4 molecules and used two cut-offs to define activity, 1 uM and 2.5 uM. The training has been done with the LazyQSAR package (Morgan Binary Classifier) and shows an AUROC >0.8 in a 5-fold cross-validation on 20% of the data held out as test. These models have been used to generate new series 4 candidates by Ersilia.

Identifiers

eos7yti | osm-series4

Results

Probability of killing P.falciparum in vitro (IC50 < 1uM and 2.5uM, respectively)

Antimalarial activity from OSM

Input molecules

Enter a list of molecules using SMILES notation and each molecule on a separate line

```
N#CC1=CC=C(C2=CC(O)=C3C(=N2)C=CC2=C3C=CN2)C=C1
CC1(C)CC(=O)C2=CN=C(NC3CCCC3)N=C2C1
CCC1=NC=NC(C2=CC(F)=C(C(=O)N3CCN4CCC[C@H]4C3)C(F)=C2)=C1C#CC1=CC=C(NC)N=C1
CCC(=O)N(C)C[C@H](O)C1=CC=CC(OCC2=NC3=CC=CC=C3N2C)=C1
```

Run

Or upload a CSV file with a single column named SMILES

Drag and drop file here
Limit 200MB per file • CSV

Browse files

Run

Course recap

In this course, we have...

— Explained how can AI help the Drug Discovery process
— Played with AI model interpretation
— Learnt the basic steps to train an AI model
— Introduced the Python programming language
— Tried cloud-base computing systems

What would you like to discuss?

— Local set up of workstations
— How to use the Ersilia Model Hub
— How to learn more Python
— How to learn more about AI
— How to better use existing databases
— ...?

Course evaluation

https://forms.gle/vfMivSPjb1nUrqro6

Thank You!

https://ersilia.io
hello@ersilia.io
@ersiliaio