Ersilia

H3D FOUNDATION

H3D Foundation and Ersilia present

Bringing data science and AI/ML tools to infectious disease research

Event Sponsors

wellcome

CS&S

Code for Science & Society

Session 2: virtual screening cascade

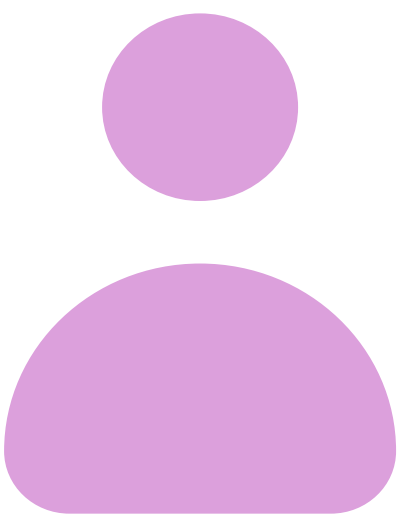
Breakout session

Gemma Turon, @TuronGemma, gemma@ersilia.io
Miquel Duran-Frigola, @mduranfrigola, miquel@ersilia.io
Ersilia Open Source Initiative, @ersiliaio, https://ersilia.io
28th September 2022

Ersilia

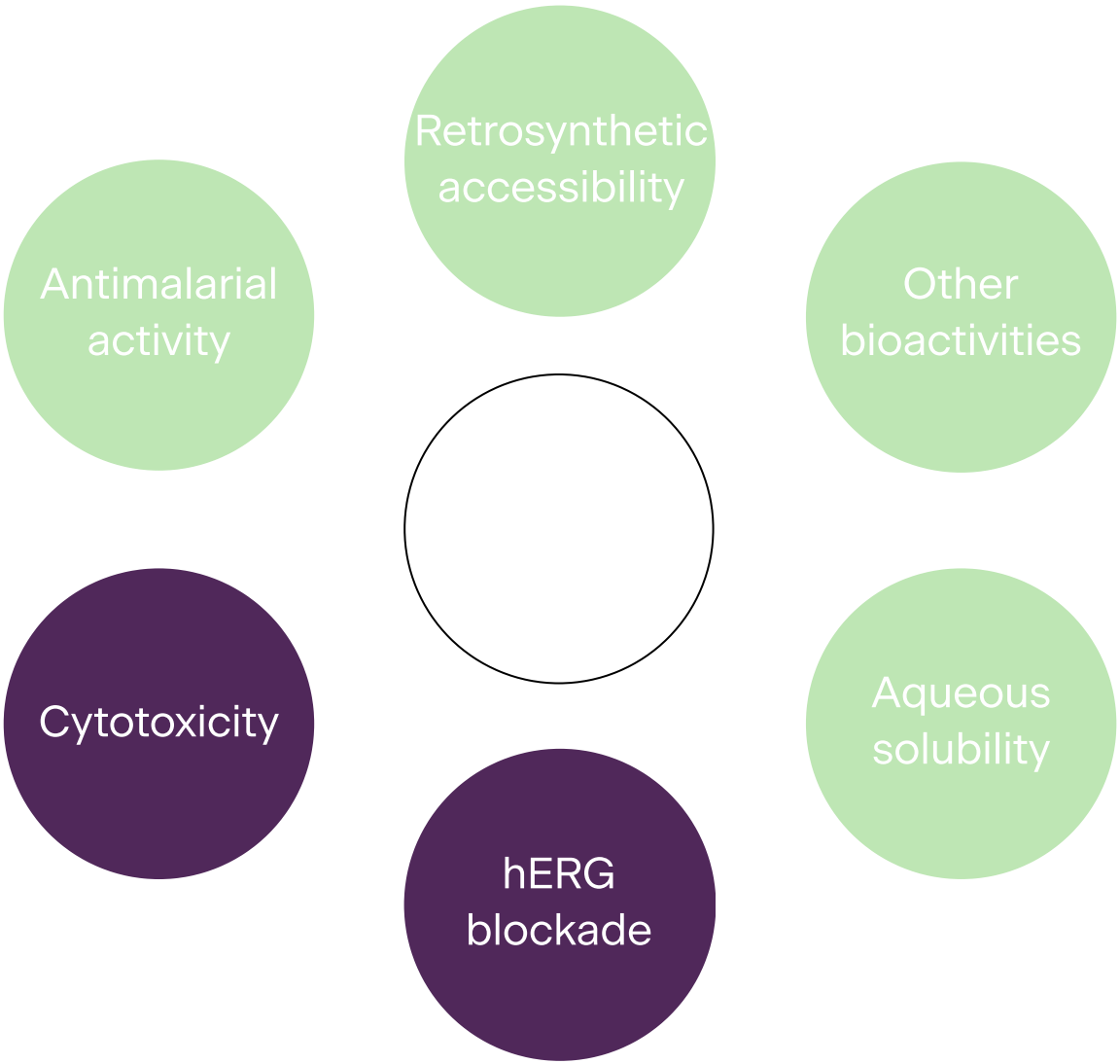
H3D FOUNDATION

I have access to a library of compounds for only their structure (SMILES) is available. I want to identify a few hits for an anti-infective drug, but only have the capacity to test 10 molecules in the first round. What can I do?

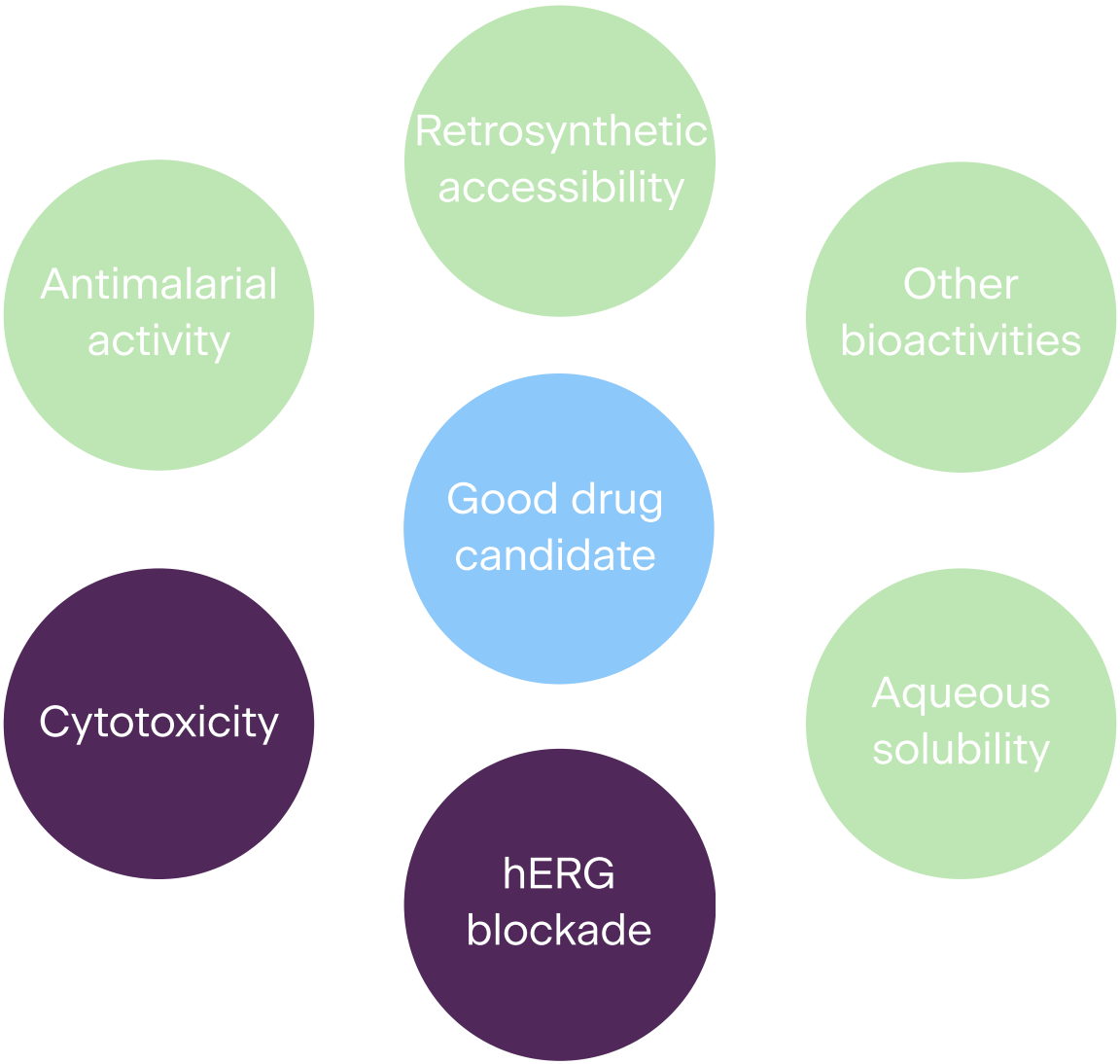


Ersilia

Virtual Screening Cascade



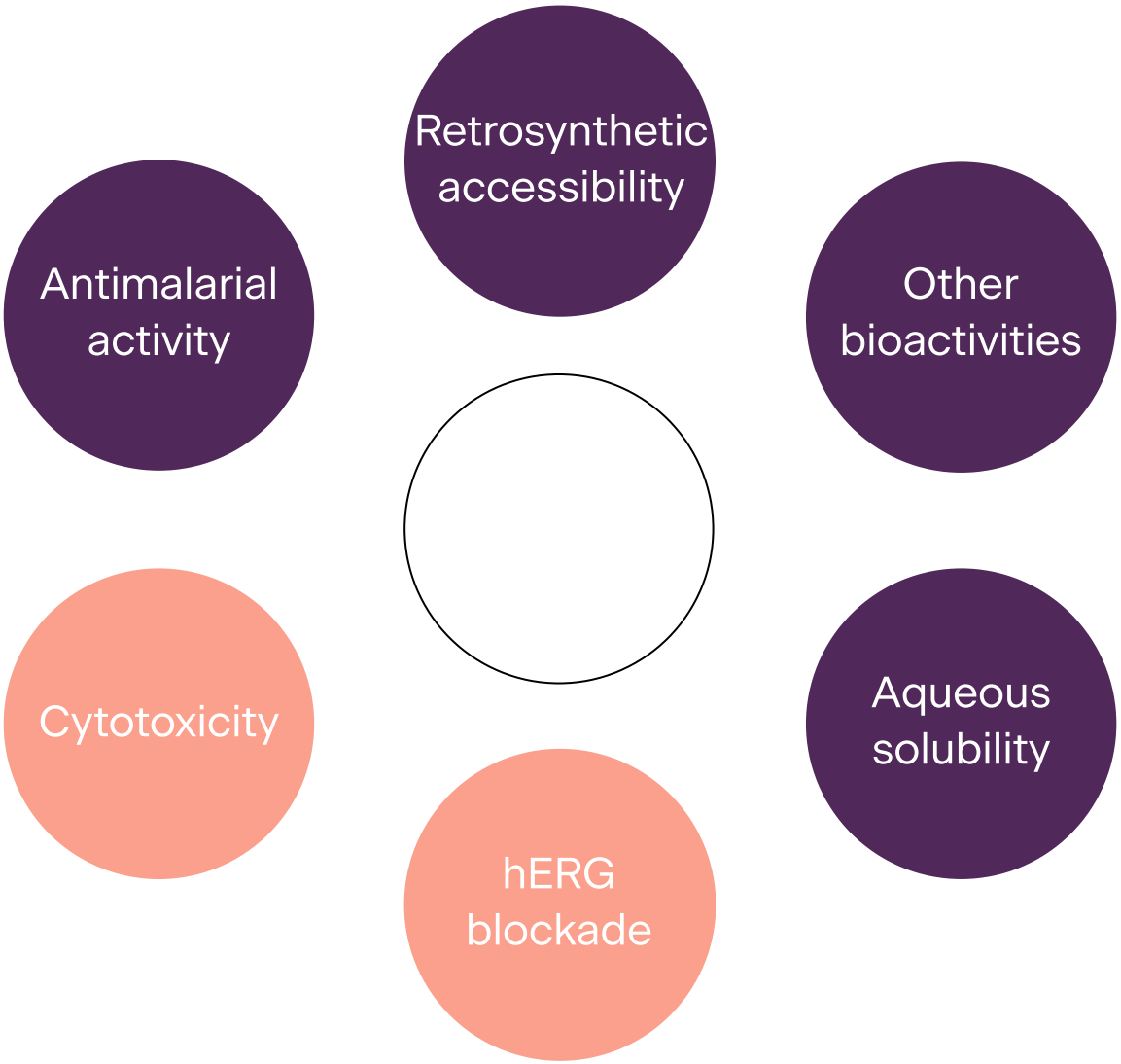
Virtual Screening Cascade



Ersilia

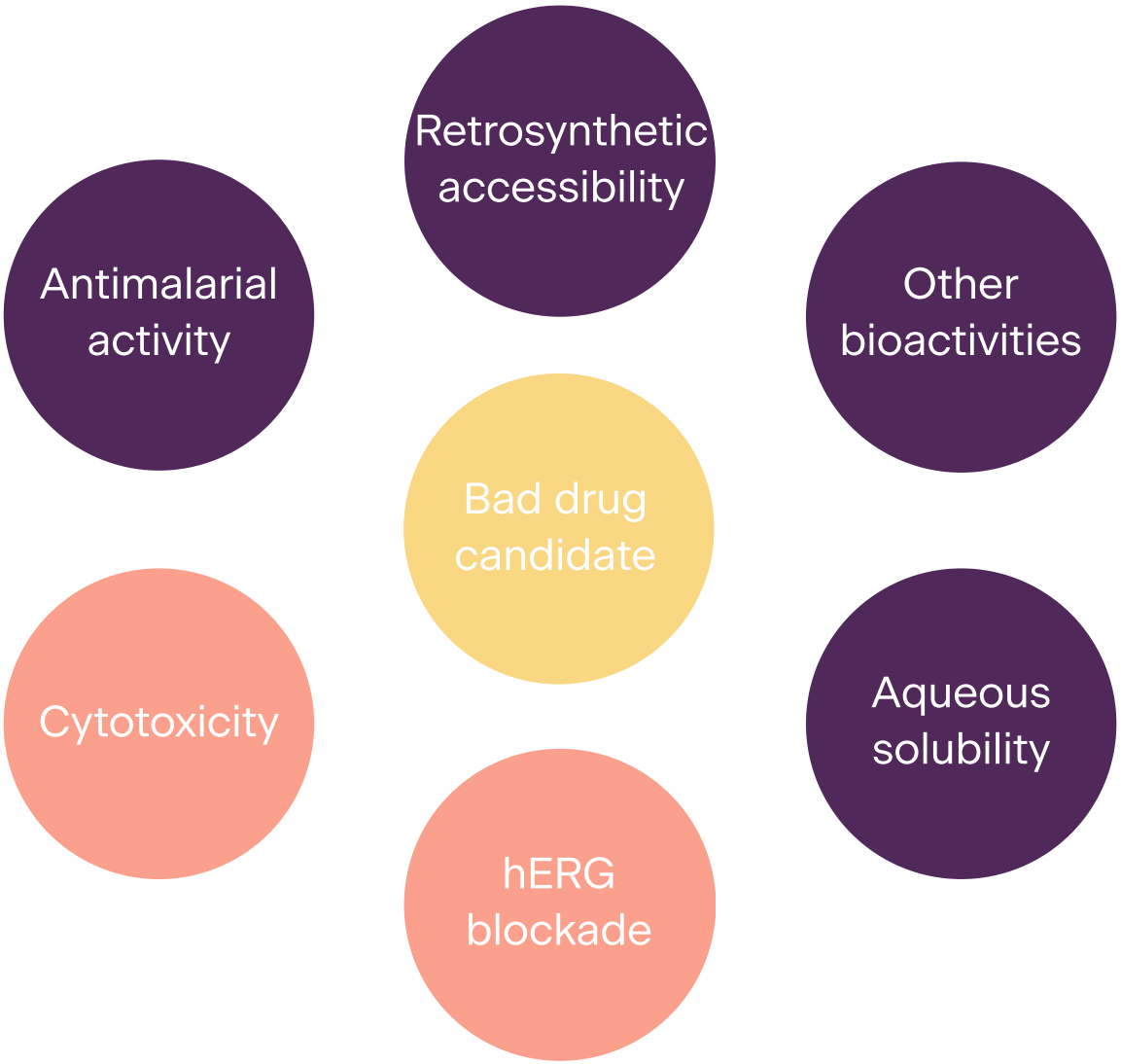
H3D FOUNDATION

Virtual Screening Cascade



Ersilia

Virtual Screening Cascade



Ersilia Model Hub

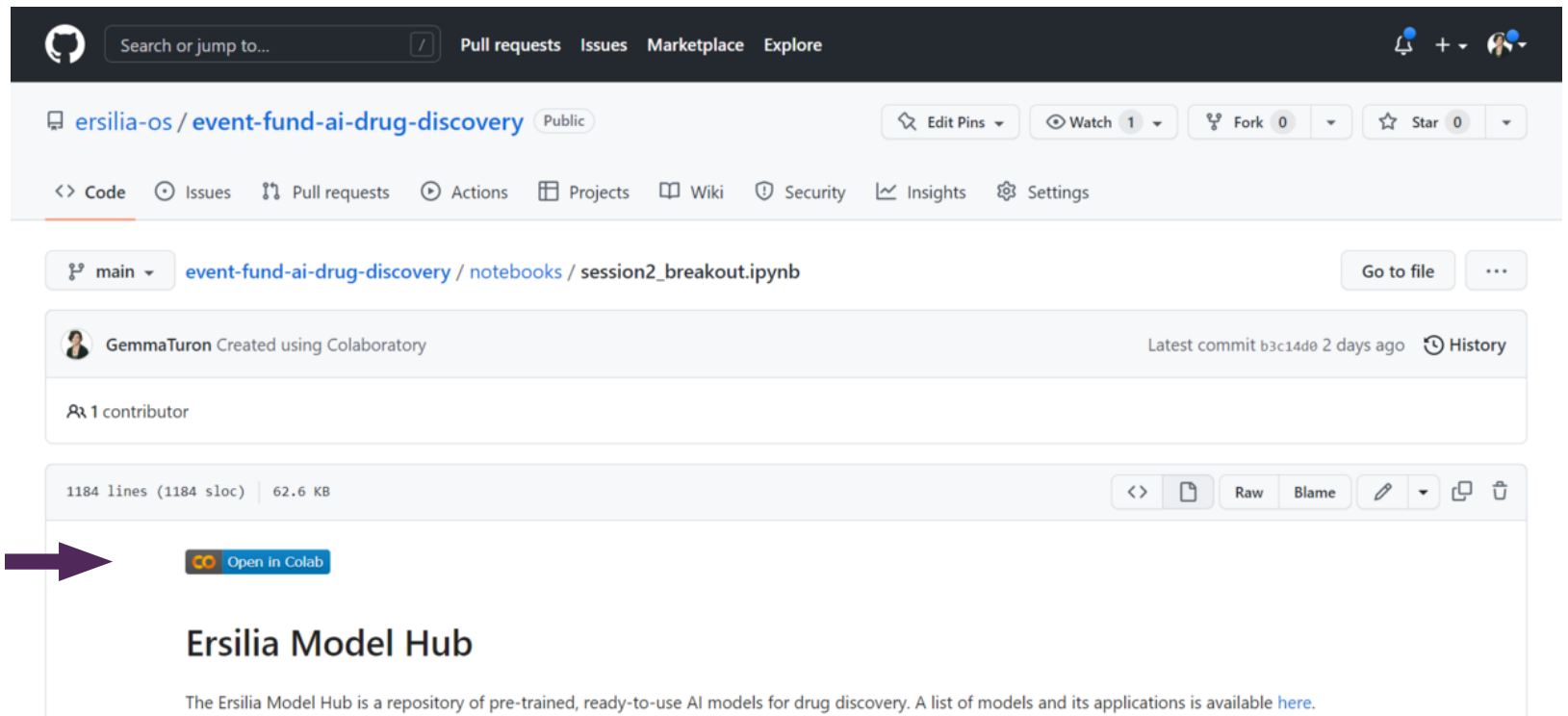
Repository of pre-trained, ready to use AI/ML models for drug discovery

- Available models can be browsed at https://ersilia.io/model-hub
- The code is available at https://github.com/ersilia-os/ersilia (GPLv3 License)
- Accessibility:
  - Command Line Interface
  - Google Colab implementation

Turon, G., & Duran-Frigola, M. (2021). Ersilia Model Hub (Version 1.0.0) [Computer software]

Getting Started

https://github.com/ersilia-os/event-fund-ai-drug-discovery



Ersilia

Ersilia Model Hub in Colab

Click on the play button to install Ersilia in this Colab notebook.

Show code

```
#@title The Ersilia Model Hub
#@markdown Click on the play button to install Ersilia in this Colab notebook.

%%capture
%env MINICONDA_INSTALLER_SCRIPT=Miniconda3-py37_4.12.0-Linux-x86_64.sh
%env MINICONDA_PREFIX=/usr/local
%env PYTHONPATH={PYTHONPATH}:/usr/local/lib/python3.7/site-packages
%env CONDA_PREFIX=/usr/local
%env CONDA_PREFIX_1=/usr/local
%env CONDA_DIR=/usr/local
%env CONDA_DEFAULT_ENV=base
!wget https://repo.anaconda.com/miniconda/$MINICONDA_INSTALLER_SCRIPT
!chmod +x $MINICONDA_INSTALLER_SCRIPT
!./$MINICONDA_INSTALLER_SCRIPT -b -f -p $MINICONDA_PREFIX
!python -m pip install git+https://github.com/ersilia-os/ersilia.git
!python -m pip install requests --upgrade
import sys
_ = (sys.path.append("/usr/local/lib/python3.7/site-packages"))
```

Turon, G., & Duran-Frigola, M. (2021). Ersilia Model Hub (Version 1.0.0) [Computer software]

Ersilia Model Hub

Each model is identified by a code (eosxxxx) and a slug (1-2 word reference).

There are five basic commands on Ersilia:

1. Fetch the model from its online storage
2. Serve the model on your system
3. Run the desired API (predict)
4. Close the model (stop serving)
5. Delete the model from the system

Turon, G., & Duran-Frigola, M. (2021). Ersilia Model Hub (Version 1.0.0) [Computer software]

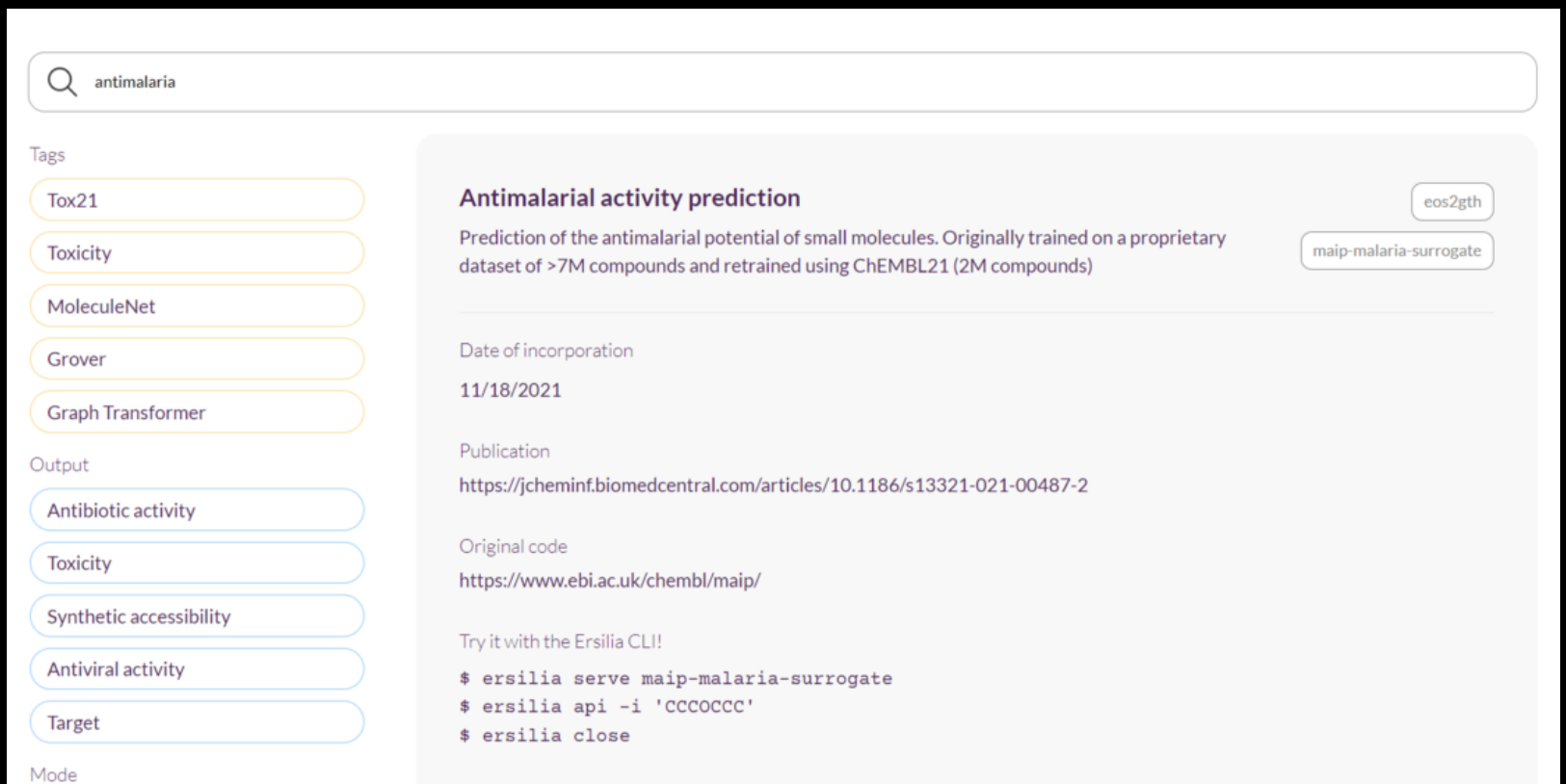
Breakout Session Exercise

We will use the 400 compounds from the MMV Malaria Box to run a series of predictions using models available in the Ersilia Model Hub and select the molecules with higher interest for experimental testing.

Steps:

1. Install the Ersilia Model Hub in Colab (together)
2. Run a model for antimalarial activity prediction and analyse the results (together)
3. In small groups, run and analyse the outcomes of the different suggested models

Antimalarial Activity



Bosc et al, Journal of Cheminformatics, 2021

Mounting Google Drive

We will first mount google drive on Colab to store the model predictions and import some basic Python Packages

```
#mount your own GDrive in Colab
from google.colab import drive
drive.mount('/content/drive')

import matplotlib.pyplot as plt
import pandas as pd
```

MMV Malaria Box

The dataset is prepared and stored as a csv file in the /data folder of the h3d_ersilia_ai_workshop in your personal Drive

```
#we can open it as a pandas dataframe
smiles = "drive/MyDrive/h3d_ersilia_ai_workshop/data/session2/mmv_malariabox.csv"
df=pd.read_csv(smiles)
df.head()
```

	CAN_SMILES
0	COc1ccccc1CNC(=O)CCn1c(=O)[nH]c2ccsc2c1=O
1	CN(C)c1ccc(C(O)(c2ccc(N(C)C)cc2)c2ccc(N(C)C)cc...
2	Cc1ccc(-c2cc3c(SCC(=O)Nc4cc(C(F)(F)F)ccc4Cl)nc...
3	CCOC(=O)C1=C(c2ccccc2)N=c2sc(=Cc3cc(C)n(-c4ccc...
4	CCOC(=O)c1cnc2c(C)cc(C)cc2c1Nc1ccc(OC)c(OC)c1

Antimalarial Activity: eos2gth

We use the ! notation to access the CLI commands for Ersilia from the notebook

Data is prepared in the folder /data already in your google drive, and predictions will also be stored there.

```
#Step One: Retrieve the model from the internet
!ersilia fetch eos2gth
```

```
Fetching model eos2gth: maip-malaria-surrogate
Model eos2gth fetched successfully!
```

Antimalarial Activity: eos2gth

```python
#we must import the ErsiliaModel Function for Python
from ersilia import ErsiliaModel

#Step 2: load the model in the notebook
model = ErsiliaModel("eos2gth")
#Step 3: bring the model alive
model.serve()
#Step 4: run predictions for the input smiles
output = model.predict(input=smiles, output="pandas")
#Step 5: close the model
model.close()
```

We have obtained the predictions as "output" in pandas format, which we can directly save to Drive as .csv

```python
output.to_csv("drive/MyDrive/DataScience_Workshop/data/day2/eos2gth.csv",
index=False)
```

Analysing predictions

To analyse the predictions, we can read them from the Drive

```
df = pd.read_csv("drive/MyDrive/DataScience_Workshop/data/day2/eos2gth.csv")
df.head()
```

	key	input	score
0	ALGPHOUNWIZIOQ-UHFFFAOYSA-N	COc1ccccc1CNC(=O)CCn1c(=O)[nH]c2ccsc2c1=O	6.886159
1	QFVDKARCPMTZCS-UHFFFAOYSA-N	CN(C)c1ccc(C(O)(c2ccc(N(C)C)cc2)c2ccc(N(C)C)cc...	15.483176
2	HKNNPGWJKJDXCN-UHFFFAOYSA-N	Cc1ccc(-c2cc3c(SCC(=O)Nc4cc(C(F)(F)F)ccc4Cl)nc...	27.288107
3	QQNUVMRXVVLAAU-UHFFFAOYSA-N	CCOC(=O)C1=C(c2ccccc2)N=c2sc(=Cc3cc(C)n(-c4ccc...	12.532308
4	MSGARPVVCYTSLR-UHFFFAOYSA-N	CCOC(=O)c1cnc2c(C)cc(C)cc2c1Nc1ccc(OC)c(OC)c1	37.554440

InChiKey SMILES Prediction

Analysing predictions

We can sort the molecules from higher to lower score

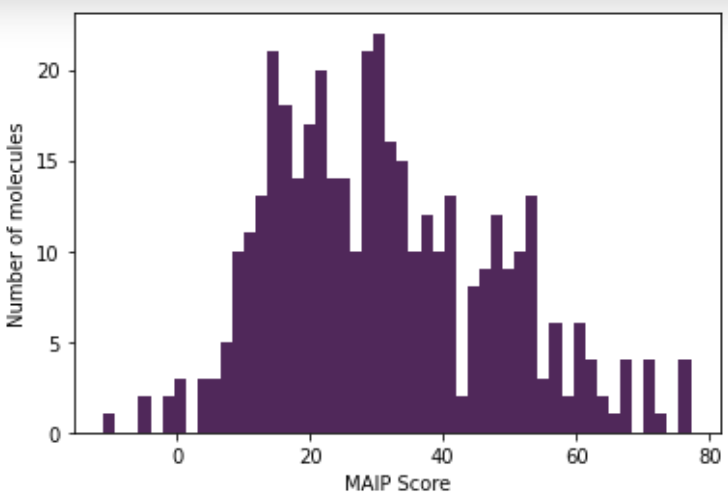
```
1 output.sort_values("score", ascending=False).head()
```

	key	input	score
136	HPFVQAYLQOSFOL-UHFFFAOYSA-N	COc1cccc(Nc2nc(NCCO)c3ccccc3n2)c1	77.199020
399	ZGMMVVYGDFQTBB-UHFFFAOYSA-N	OCCNc1nc(Nc2ccc(Cl)c(Cl)c2)nc2ccccc12	76.798370
69	RHZLKBRFIAZMTN-UHFFFAOYSA-N	Cc1ccc(Nc2nc(NCCO)c3ccccc3n2)cc1C	76.052498
339	NPWXHTXMBIOHKI-UHFFFAOYSA-N	Cn1c(=O)n(C)c2cc(CNCCNc3ccnc4cc(Cl)ccc34)ccc21	76.012589
55	YZZGEZJZTYYZGG-UHFFFAOYSA-N	CCN(CC)CCNc1ncnc2c1[nH]c1ccc(Cl)cc12	71.951157

Analysing predictions

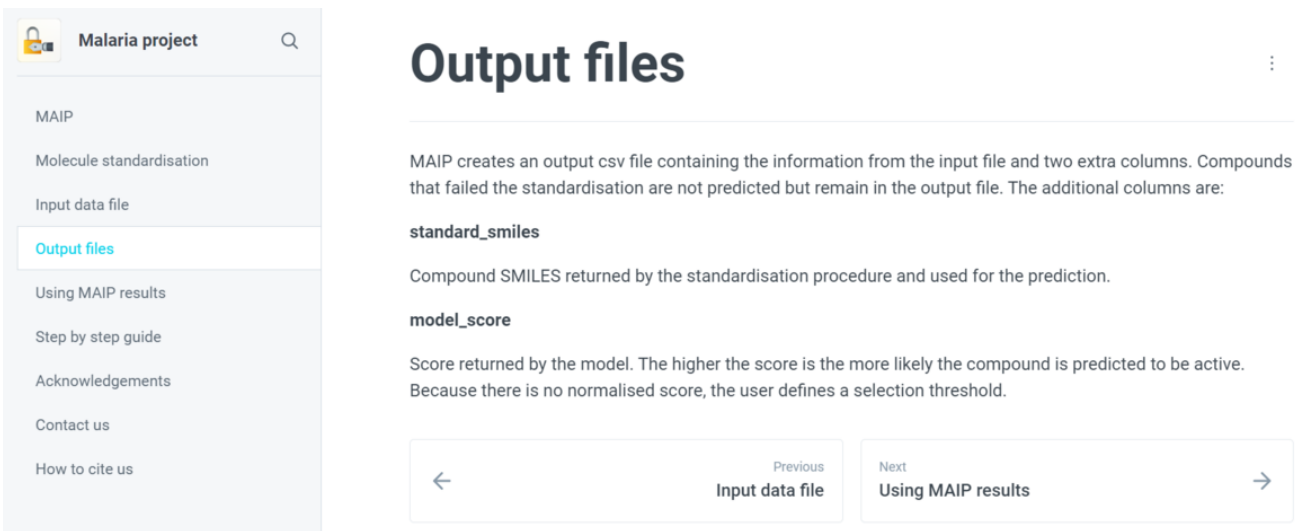
Or plot the distribution of all scores in a histogram

```
plt.hist(output["score"], bins=50, color="#50285a")
plt.xlabel("MAIP Score")
plt.ylabel("Number of molecules")
plt.show()
```



Analysing predictions

- Is this model a regression or a classification?
- How can we interpret the model "score" as antimalarial potential?

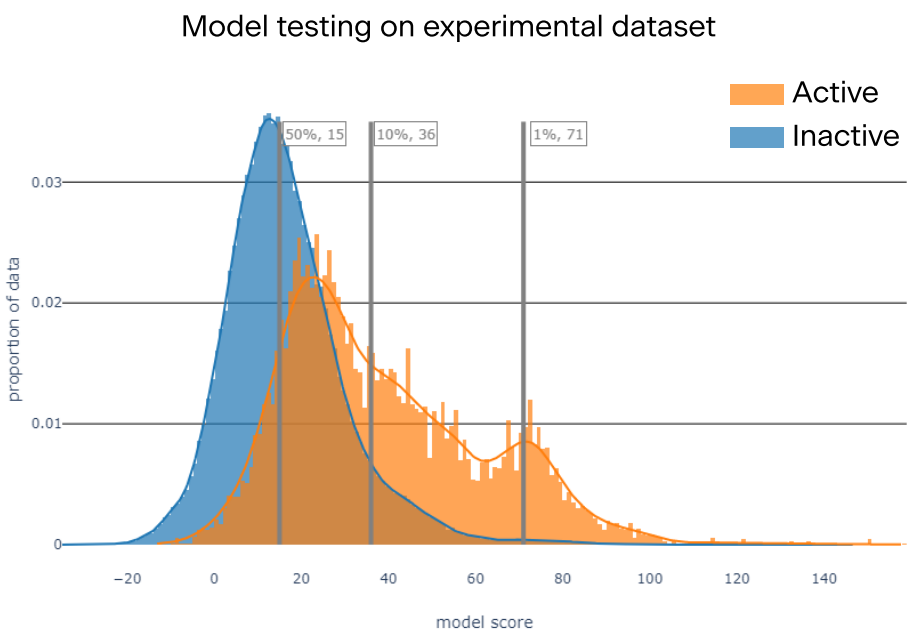


https://chembl.gitbook.io/malaria-project/output-file

Plotting the results

To better understand the model outputs, we need to go back to the original model publication where the training data is explained

Given this test set, what could we choose as a good activity threshold for our dataset?



https://chembl.gitbook.io/malaria-project/output-file

Next steps

In groups, run additional predictions for the MMV Dataset and select the molecules that you would move on to experimental testing.

We provide a list of suggested models to be used, you do not need to run predictions for all:

- eos46ev (anti tuberculosis)
- eos4e41 (antibiotic activity)
- eos2ta5 (hERG blockade)
- eos2r5a (retrosynthetic availability)
- eos6oli (aqueous solubility)
- eos9yui (natural product likeness)

Next steps

Some pointers:

- Run predictions for relevant models, storing the results in Google Drive
- Analyse the results on the Google Colab Notebook or directly using Excel / Google Sheets
- Discuss the results and rank/select best molecules
- Prepare a short (10 min presentation) of your findings

There is no right or wrong answer, it's just a practise exercise

https://ersilia.gitbook.io/event-fund/