Ersilia

H3D FOUNDATION

H3D Foundation and Ersilia present

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

Event Sponsors

wellcome

CS&S

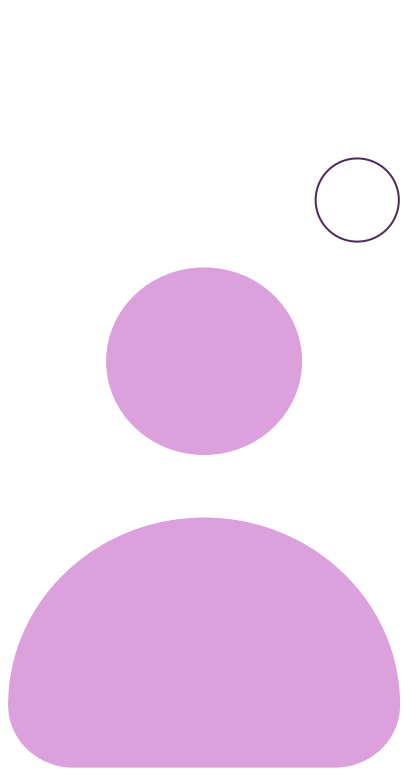
Code for Science & Society

Session 4: Generative models

Breakout Session 4

Gemma Turon, @TuronGemma, gemma@ersilia.io
Miquel Duran-Frigola, @mduranfrigola, miquel@ersilia.io
Ersilia Open Source Initiative, @ersiliaio, https://ersilia.io
30th September 2022

Ersilia



I have a list of selected hits, but I would like to improve some of the molecules with not so great activity and diversify the collection

Similarity search

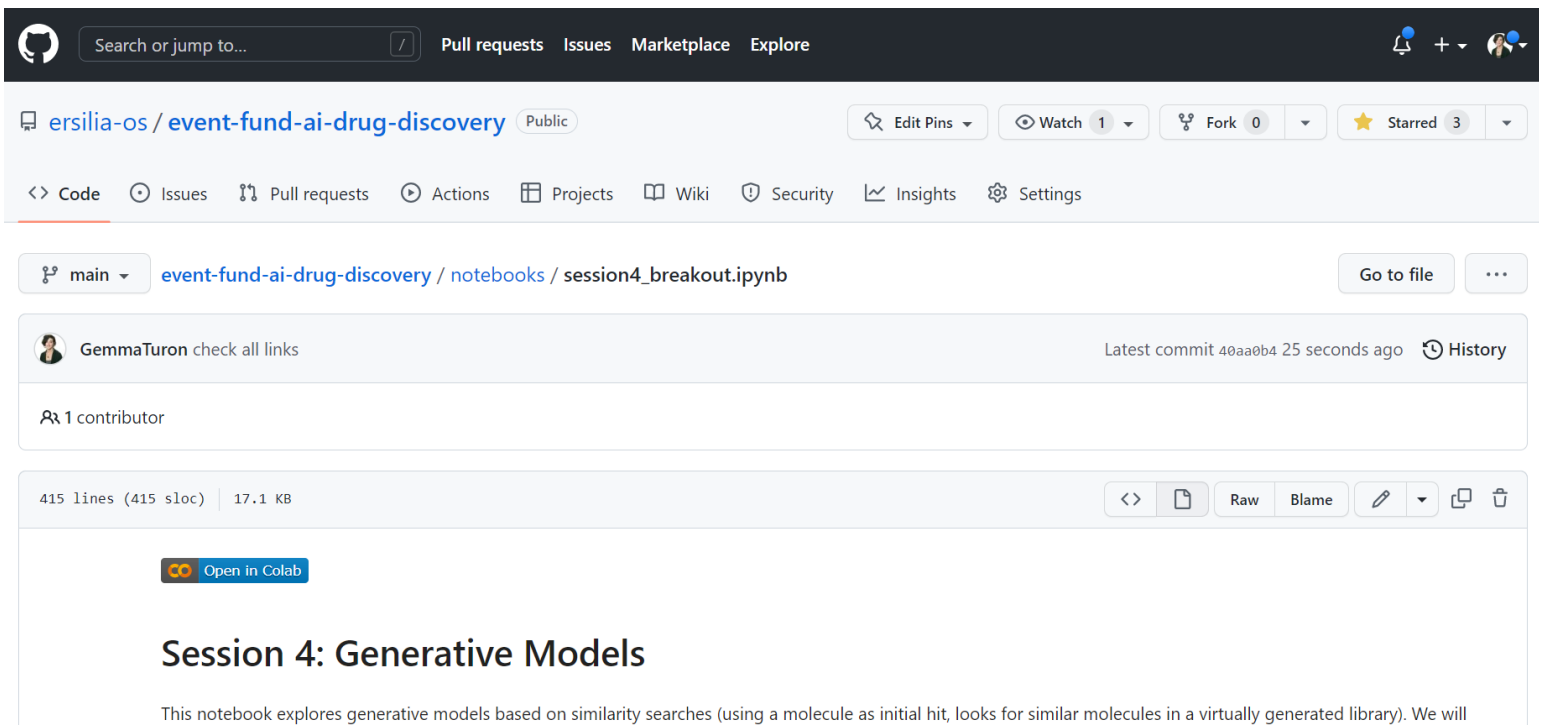
Similarity search is a type of generative model based on searching for similar molecules among an already generated virtual library.

In this case, the generative step has already been done and we only need to filter out molecules

- Much faster
- Potentially less diversity

Breakout Exercise

We will look for alternatives to one of the molecules from the MMV Malaria Box we used in session 2.



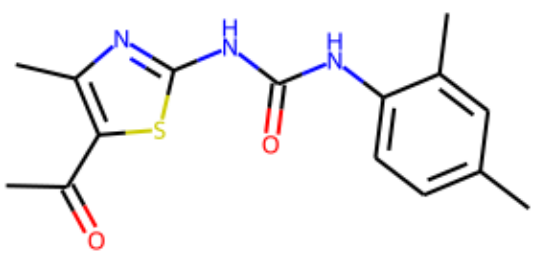
Ersilia

H3D FOUNDATION

Molecule Selection

Select a molecule of interest from the MMV Malaria Box Dataset (session2/data)

`CC(=O)c1sc(NC(=O)Nc2ccc(C)cc2C)nc1C`



Predicted activity against malaria:

Score = 4

Similarity Models in the EMH

We will work with two similarity models:

- eos4b8j: gdbchembl-similarity:
  - 166.4 billion possible molecules of up to 17 atoms
  - Interactive browsing at: http://faerun.gdb.tools/
- eos4b8j gdbmedchem-similarity
  - 10 million possible molecules curated from GDBChEMBL
  - Download and browse: http://gdb.unibe.ch

Both models provide the 100 closest molecules

Ersilia

Model Fetch and Predict

H3D FOUNDATION

```python
#running as python package
from ersilia import ErsiliaModel

model = ErsiliaModel("eos4b8j")
model.serve()
output_eos48bj = model.predict(input="Cc1ccc(Nc2nc(NCCO)c3ccccc3n2)cc1C", output="pandas")
model.close()

from ersilia import ErsiliaModel

model = ErsiliaModel("eos7jlv")
model.serve()
output_eos7jlv = model.predict(input="Cc1ccc(Nc2nc(NCCO)c3ccccc3n2)cc1C", output="pandas")
model.close()
```

Guidelines

Find the 100 compounds from each database that are most similar to your selected molecule:

- Are the hits obtained from each database different?
- Are hits from GDBMedChem synthetically more accessible than hits from GDBChEMBL?
- Do we have any molecule with predicted higher antimalarial potential than the original hit?
- Which molecules would you select for further screening?
- Is there any ADMET consideration you are taking into account for the selection, after what we reviewed on session 3?