AI/ML for prediction of biological properties of molecules

Module 1. Using AI models for drug discovery

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

Ersilia

An applied example

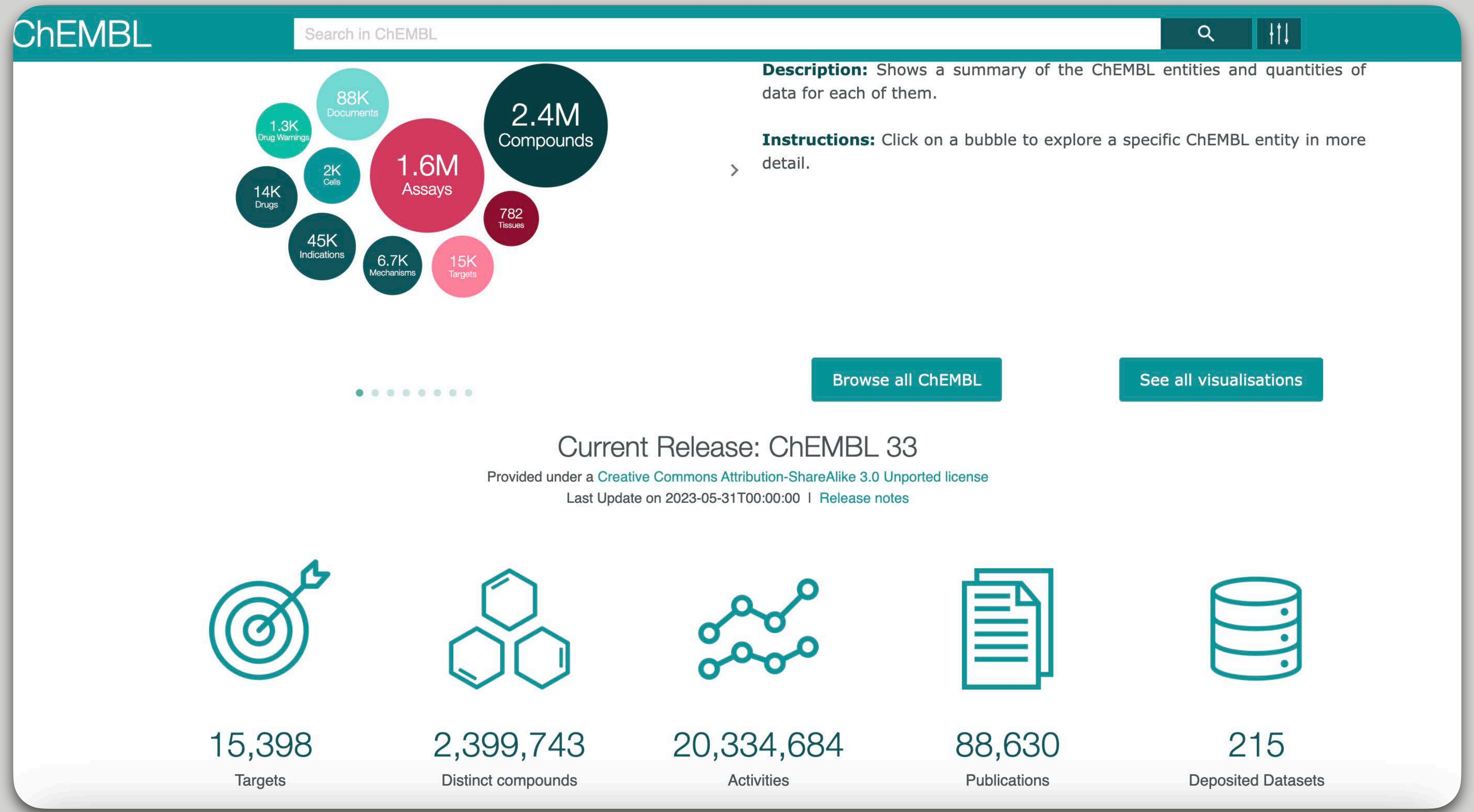
Go to menti.com and introduce 1655 9573

The problem: introductory case

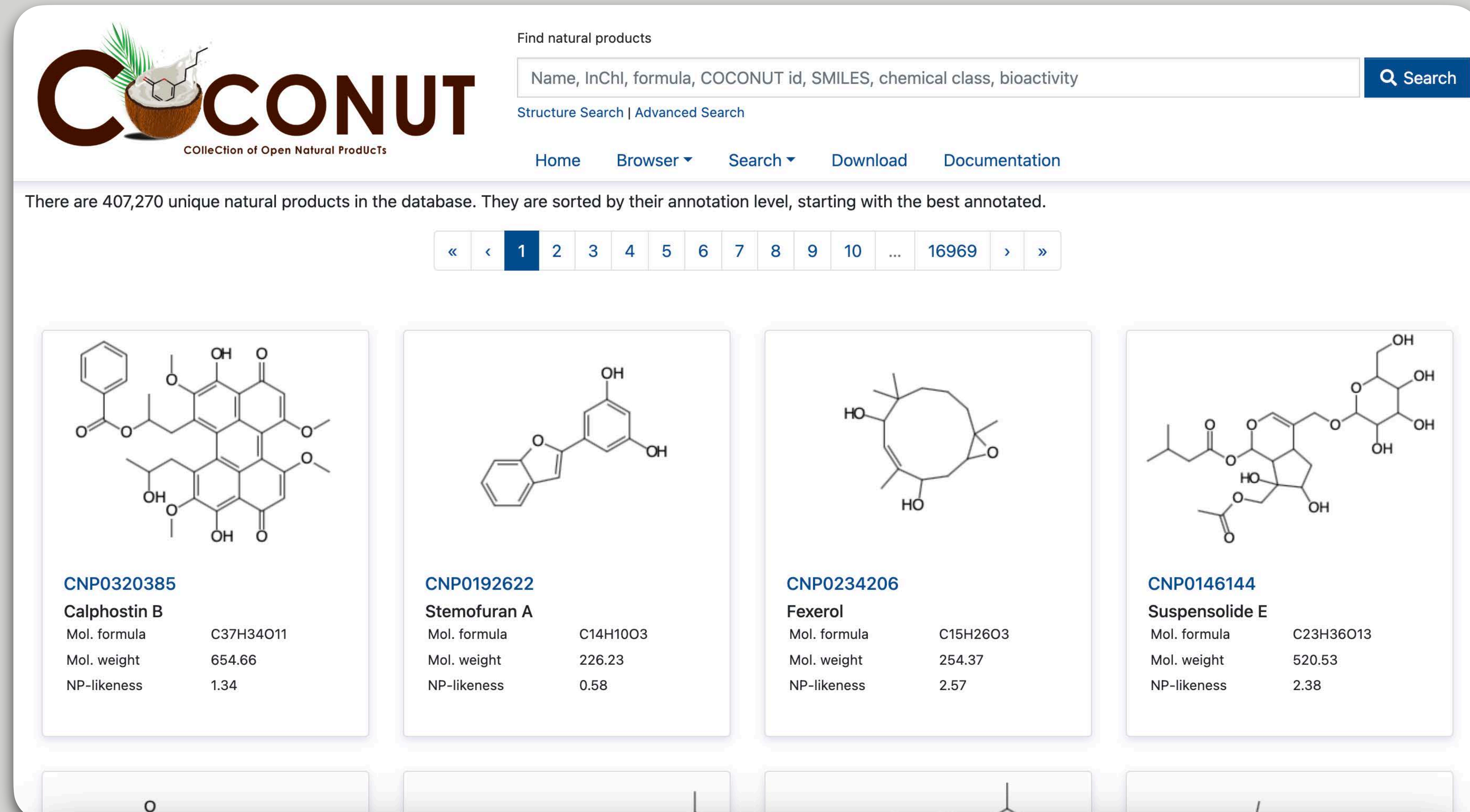
A scientist is working with two collections of ~400 compounds:

— Capacity to test 20 molecules
— Molecules must be easy to synthesize
— Nature-inspired chemistry is a plus
— Maximise chances of success in advanced stages

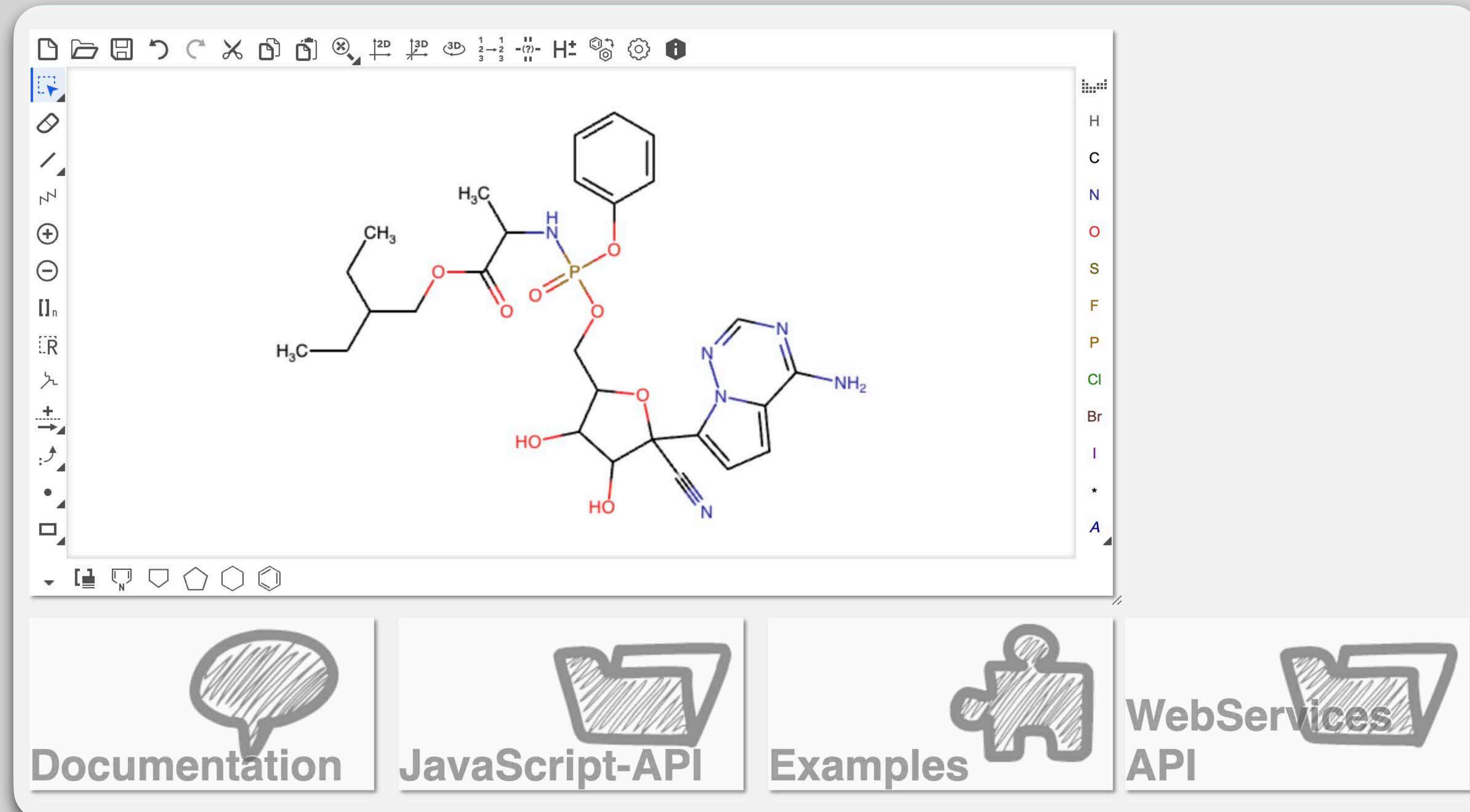
The ChEMBL database
https://ebi.ac.uk/chembl



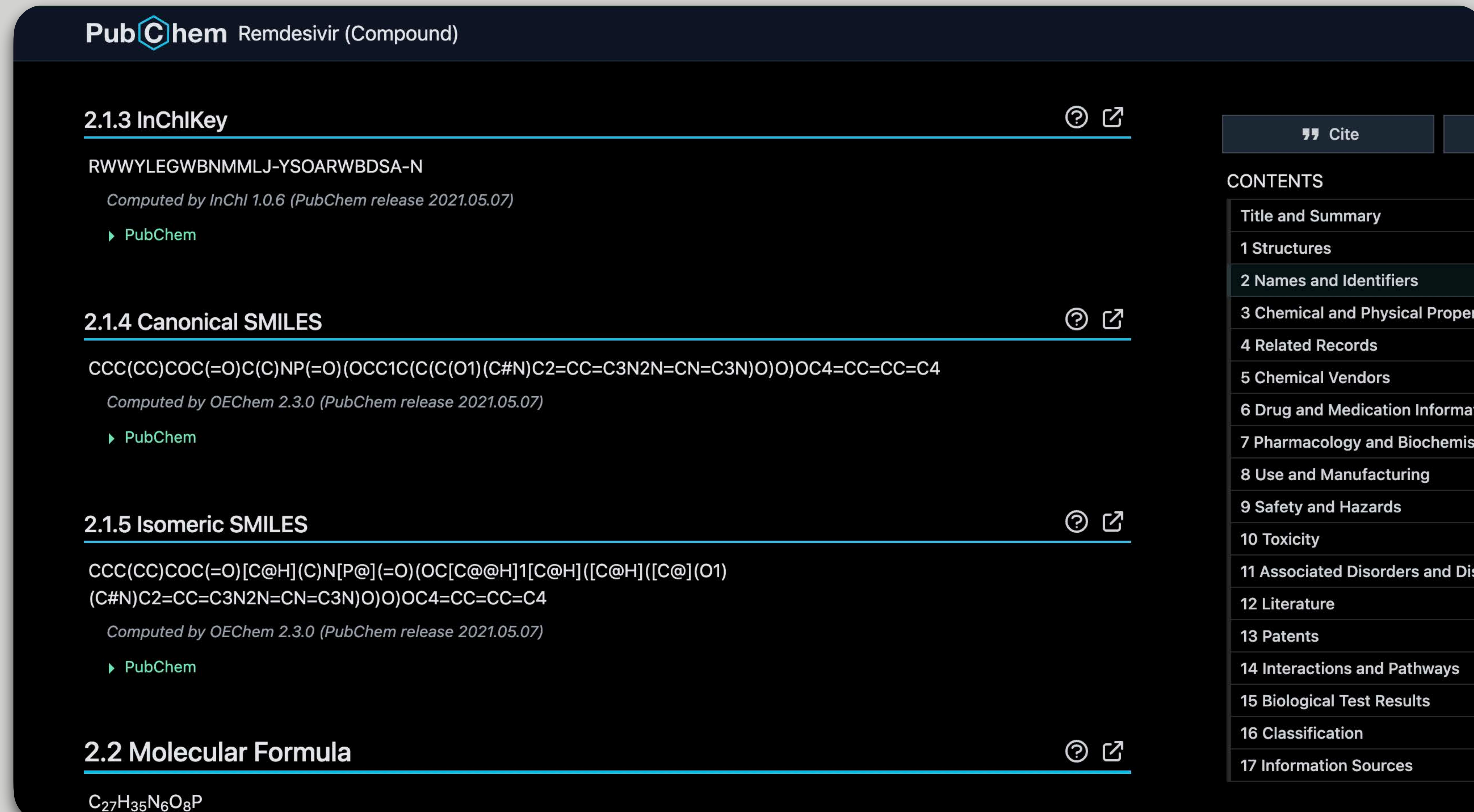
The COCONUT database
https://coconut.naturalproducts.net



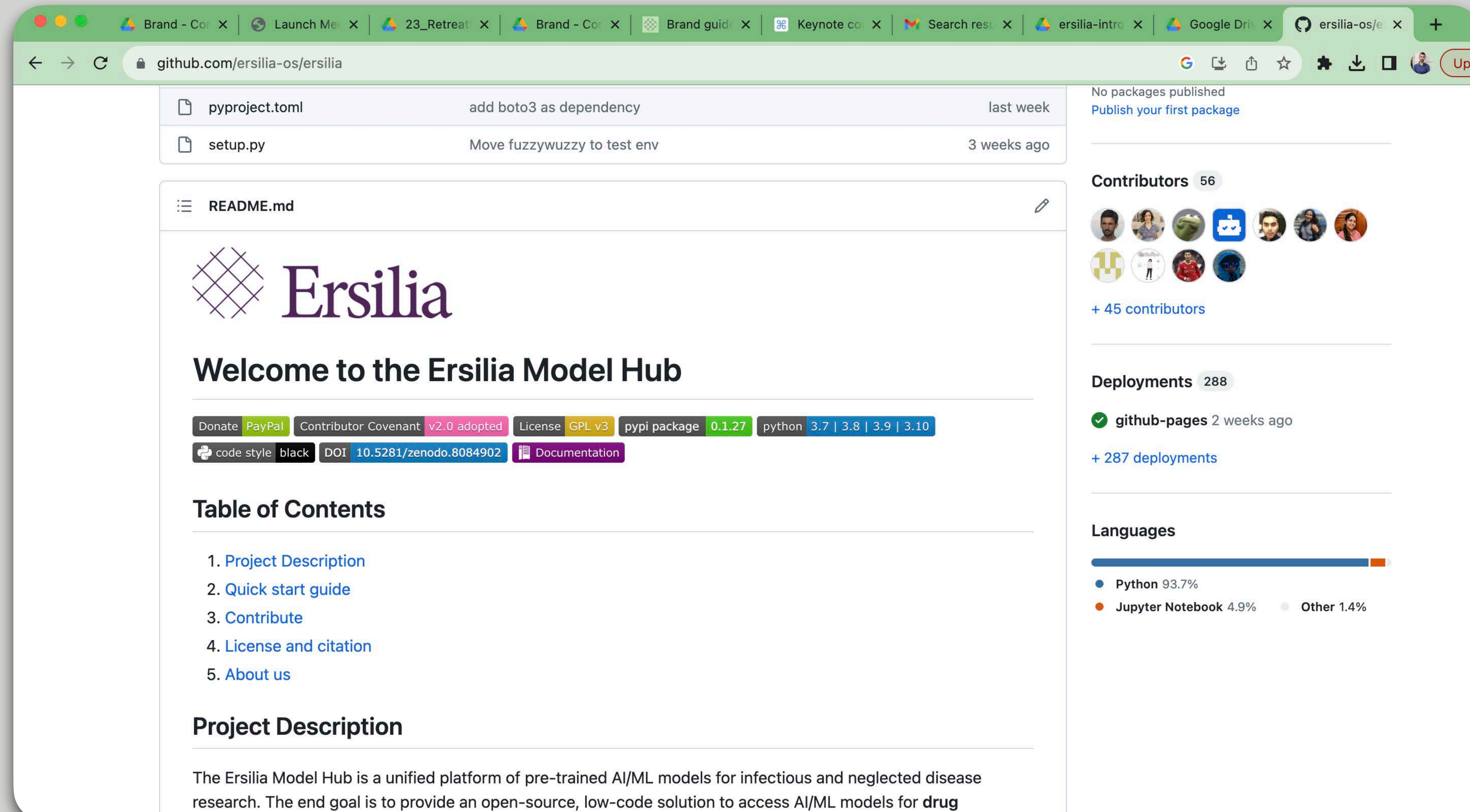
Marvin-js
https://marvinjs-demo.chemaxon.com



PubChem
https://pubchem.ncbi.nlm.nih.gov



This is a conventional screenshot slide



🧑‍💻 Exercise 1. Download the datasets of the case-study

— Go to your email inbox
— Download the chembl_selected.csv and coconut_selected.csv files
— Open and explore these files
— You can draw examples of the molecules in MarvinJs

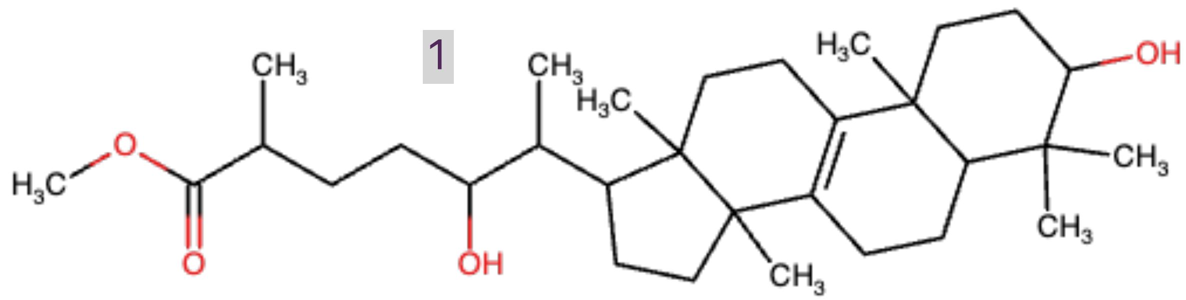
💡 Exercise 1

— What can we do to explore the data?

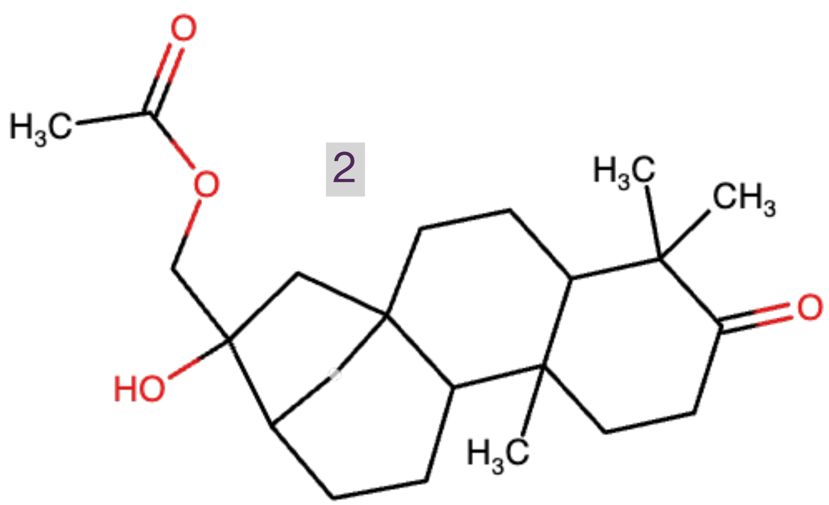
ChEMBL molecules

Coconut molecules

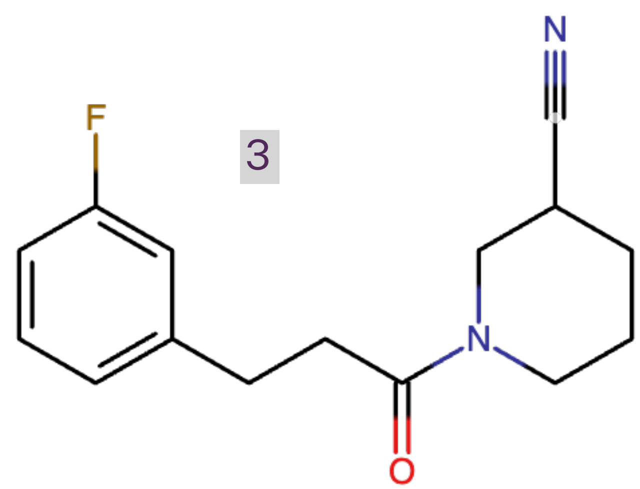
1



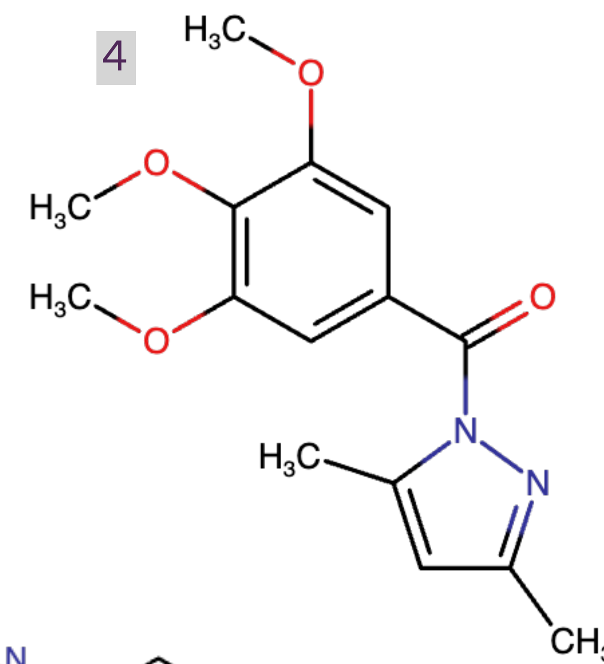
2



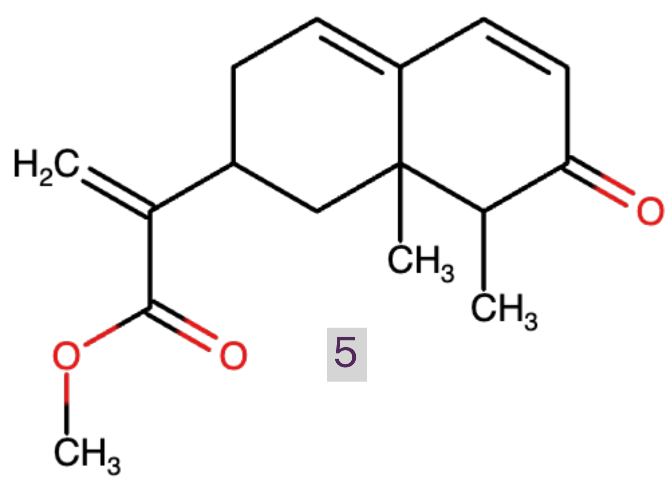
3



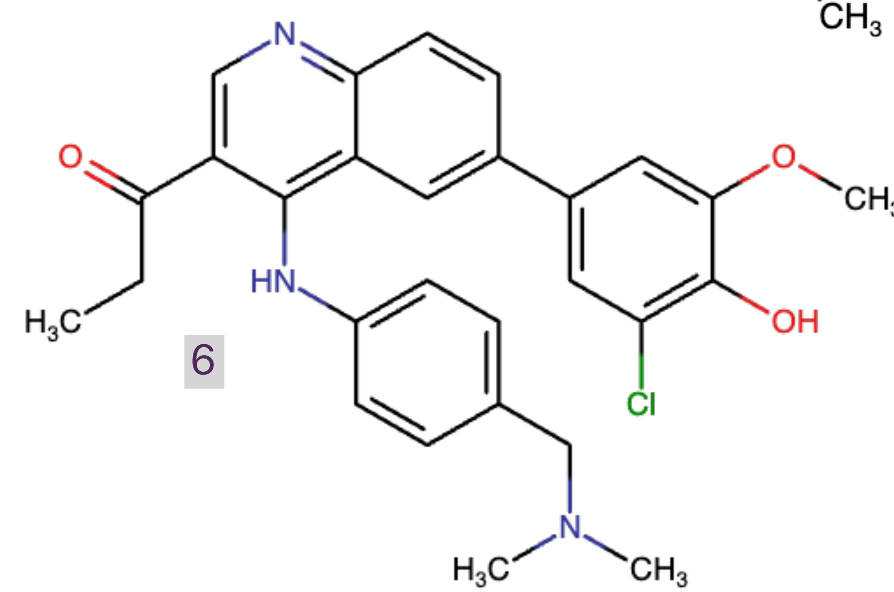
4



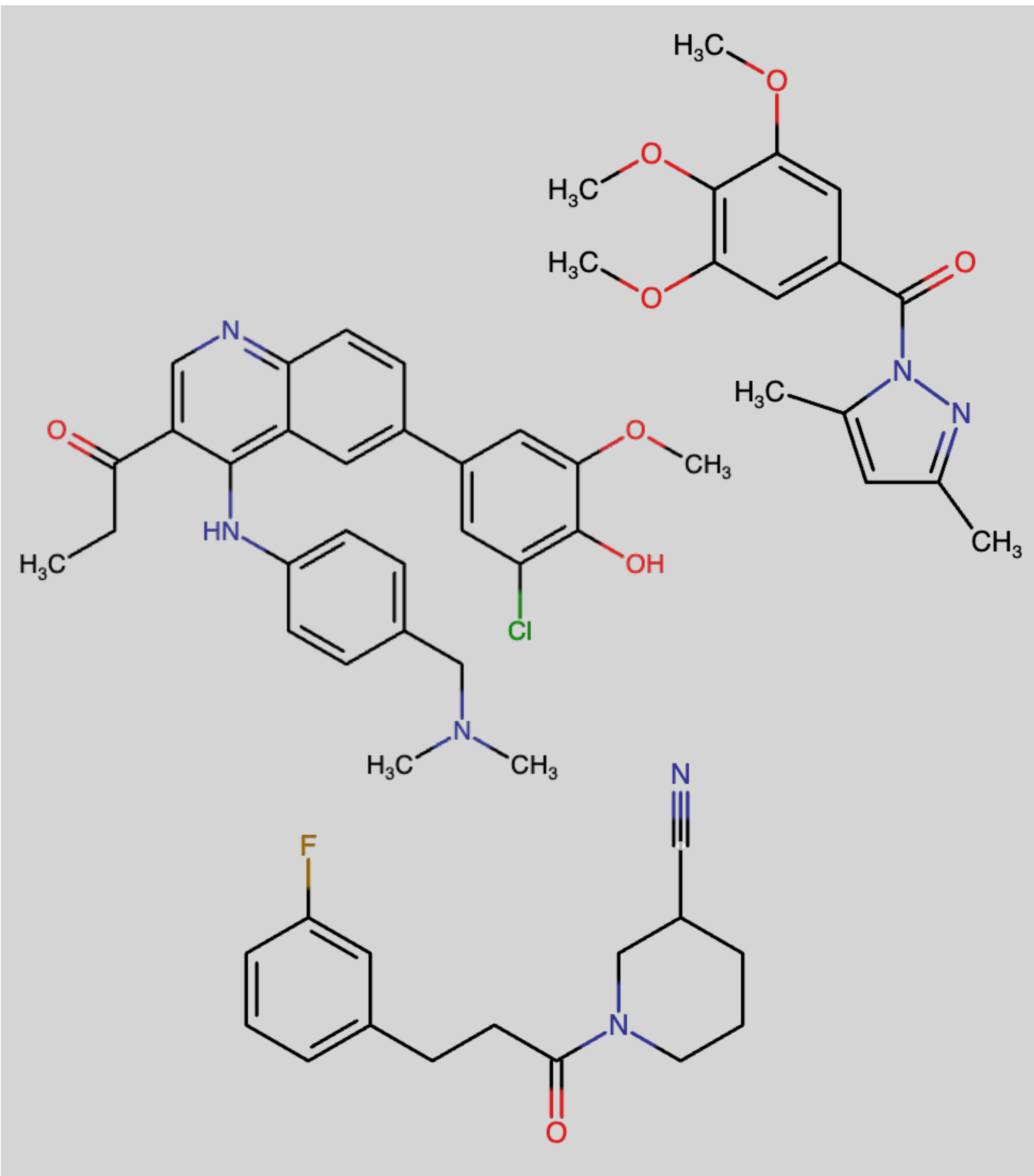
5



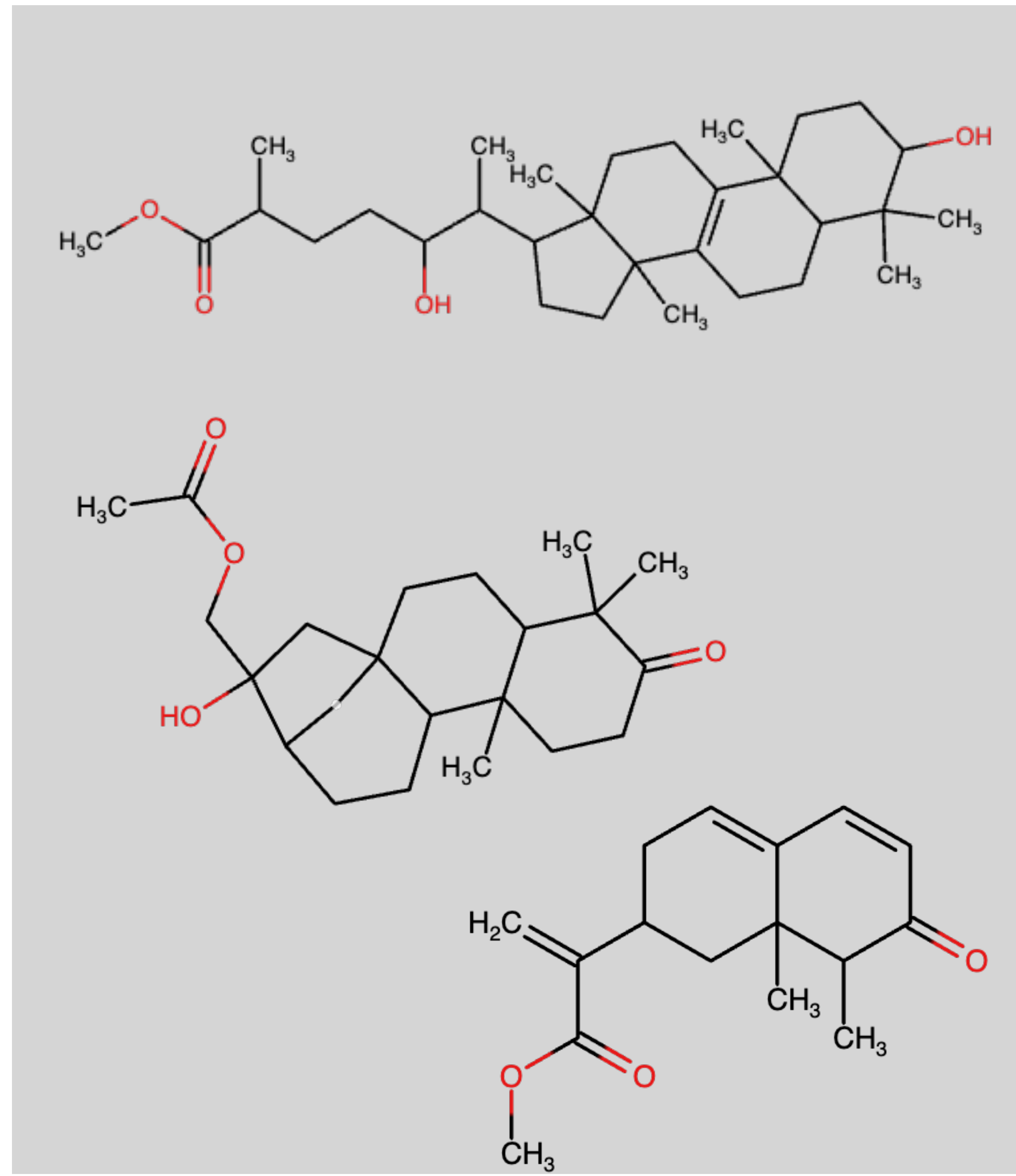
6



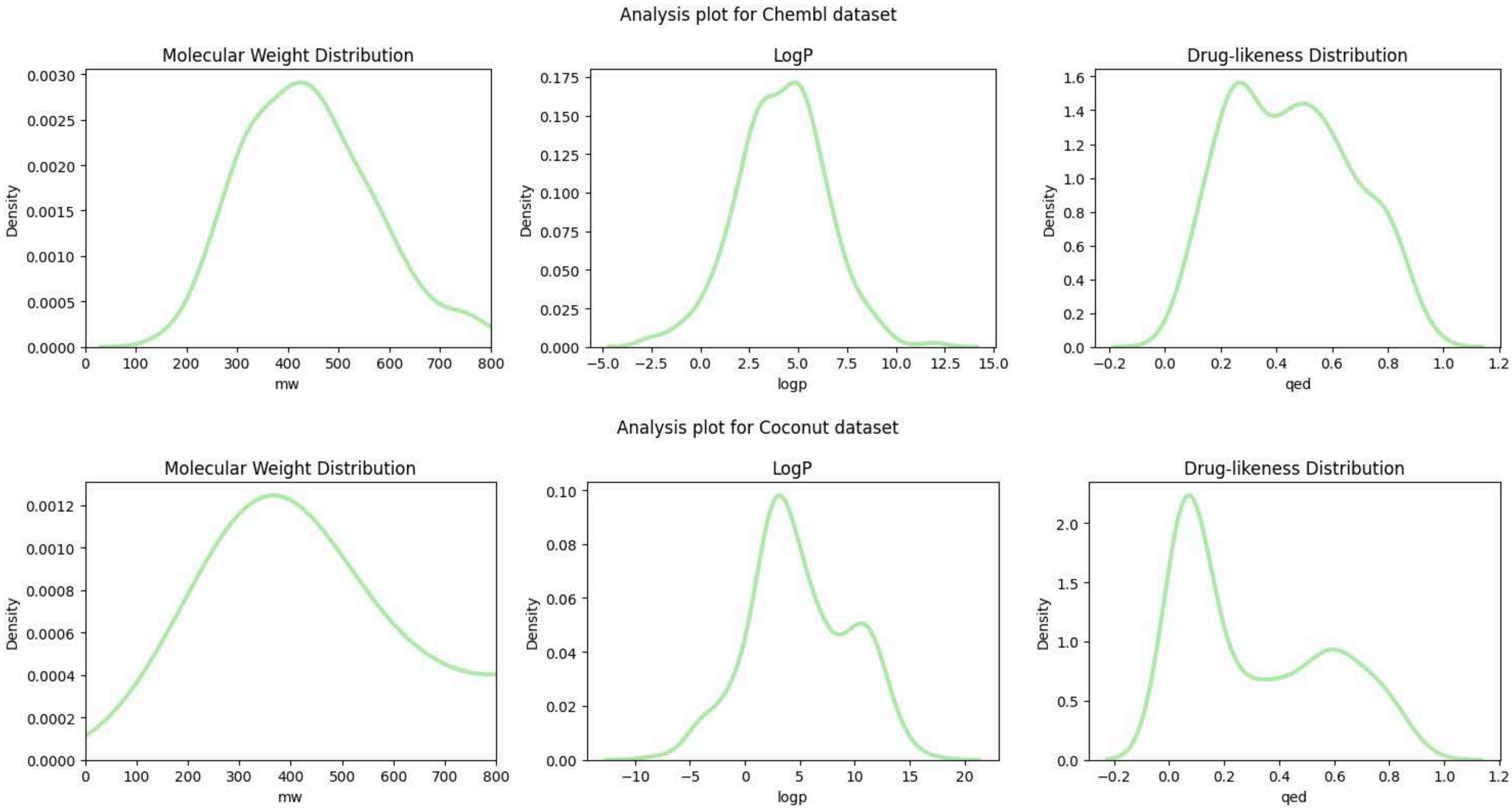
ChEMBL molecules



Coconut molecules



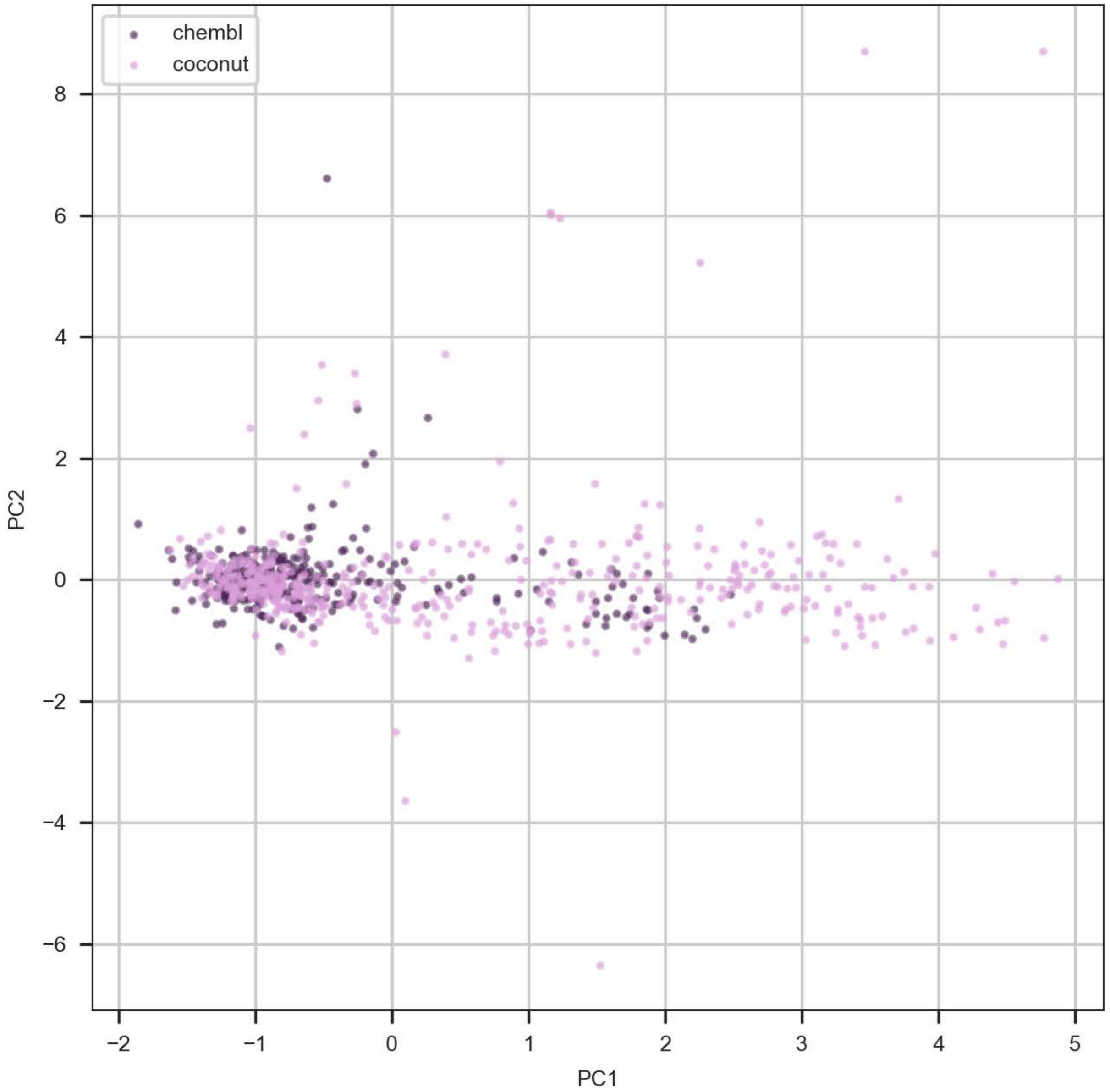
Physicochemical properties



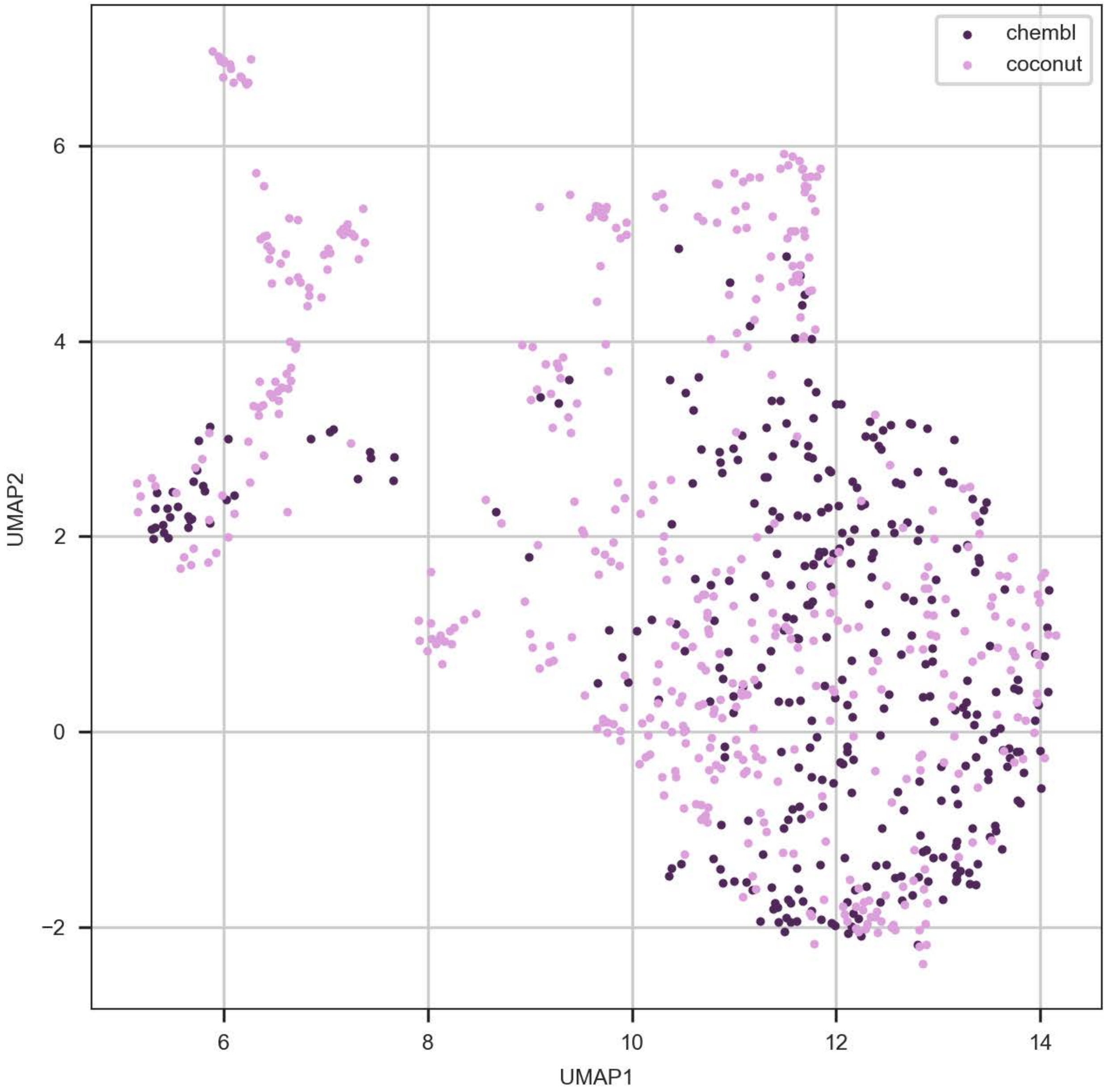
Calculated with RdKIT

Analyse the chemical space

PCA



UMAP



🧑‍💻 Exercise 1. Download the datasets of the case-study

— Go to your email inbox
— Download the chembl.csv and coconut.csv files
— Open and explore these files
— You can draw examples of the molecules in MarvinJs

💡 Discussion 1

— What is the difference between both datasets?
— Which one do you think would be easier to work with?
— What are the advantages of using one or the other?

AI tools that could aid us

🧑‍💻 Exercise 2. Look for suitable models in the Ersilia Model Hub

- Go to: https://ersilia.io/model-hub
- Note down the Ersilia code (eos0abc) for AI models that could help in our task
  - Antimalarial activity
  - Broad spectrum antibiotic activity
  - Antihelminthic activity
  - Cytotoxicity
  - hERG cardiotoxicity
  - Solubility
  - Synthetic accessibility
  - Natural product score

🧑‍💻 Exercise 2. Look for suitable models in the Ersilia Model Hub

— Go to: https://ersilia.io/model-hub
— Note down the Ersilia code (eos0abc) for AI models that could help in our task

💡 Discussion 2

— Go to menti.com and add the code: 1655 9573

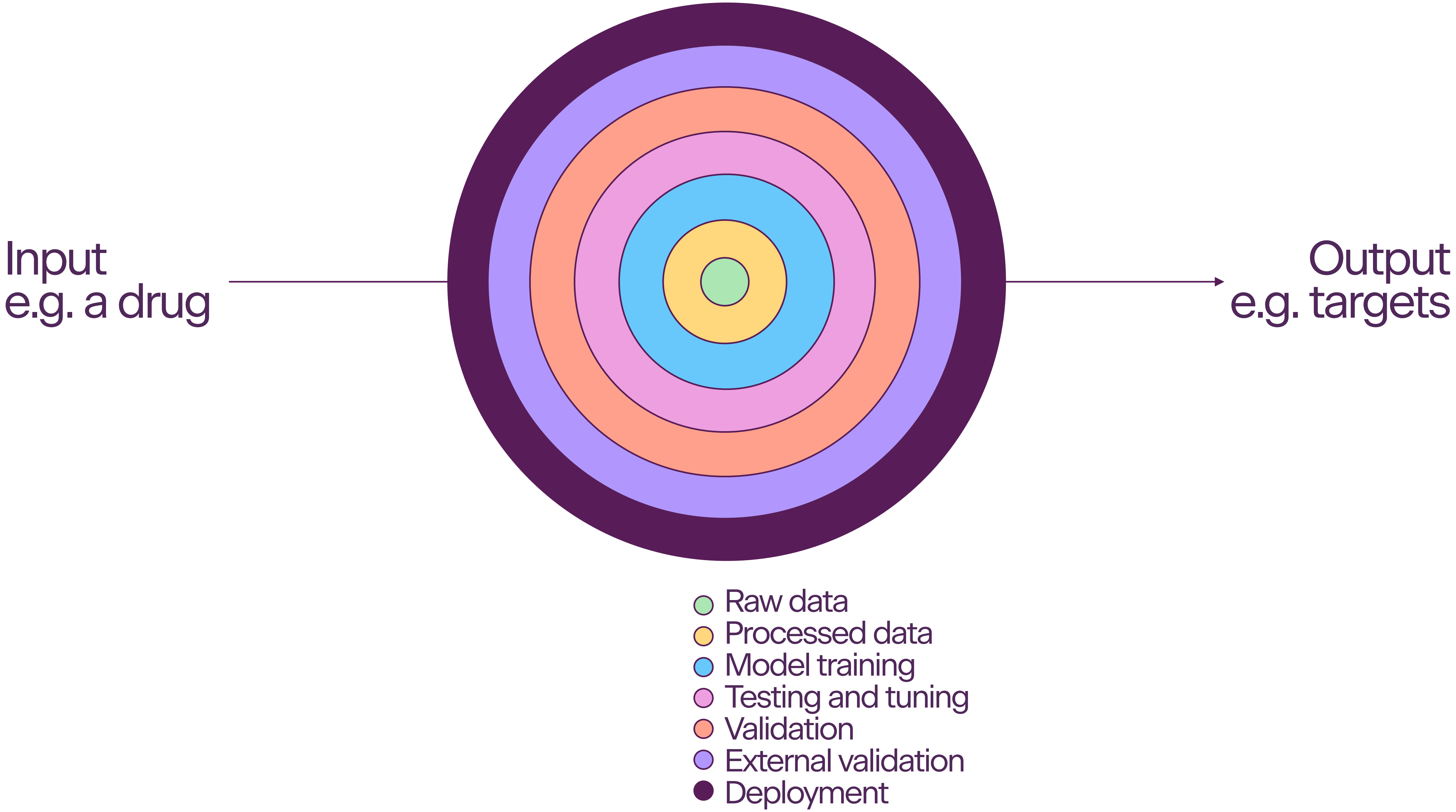
Selected models

— Antimalarial activity by MMV: eos4rta
— Antimalarial activity by Open Source Malaria: eos7yti
— ChemProp antibiotic: eos4e41
— Antischistosomiasis activity by SwissTPH: eos2l0q
— Cardiotoxicity: eos4tcc
— Cytotoxicity in HepG2 cells: eos3le9
— Solubility: eos6oli
— Synthetic accessibility: eos9ei3
— Natural product score: eos9yui

👩‍💻 Exercise 3

— Go to the online inference available through the Ersilia Model Hub for the selected models: https://bit.ly/eos4rta
— Run the predictions for both datasets for the model eos4rta
— Let’s analyse the results together
— What relevant questions could we ask ourselves?
— What information can we gather about the model?

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

Antimalarial prediction with MMV data (eos4rta)

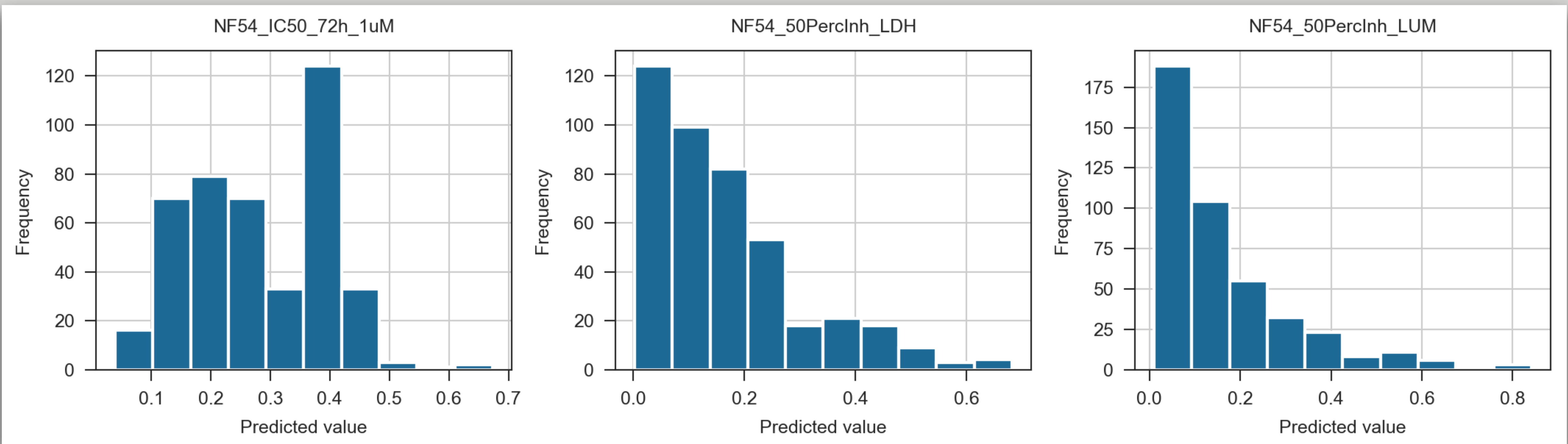
— Task: Classification
— Output: Probability of inhibiting the malaria parasite (strain NF54) in IC50 (threshold 1uM) and percentage of inhibition (50%, measured by LDH and Lum)
— Training set: MMV dataset

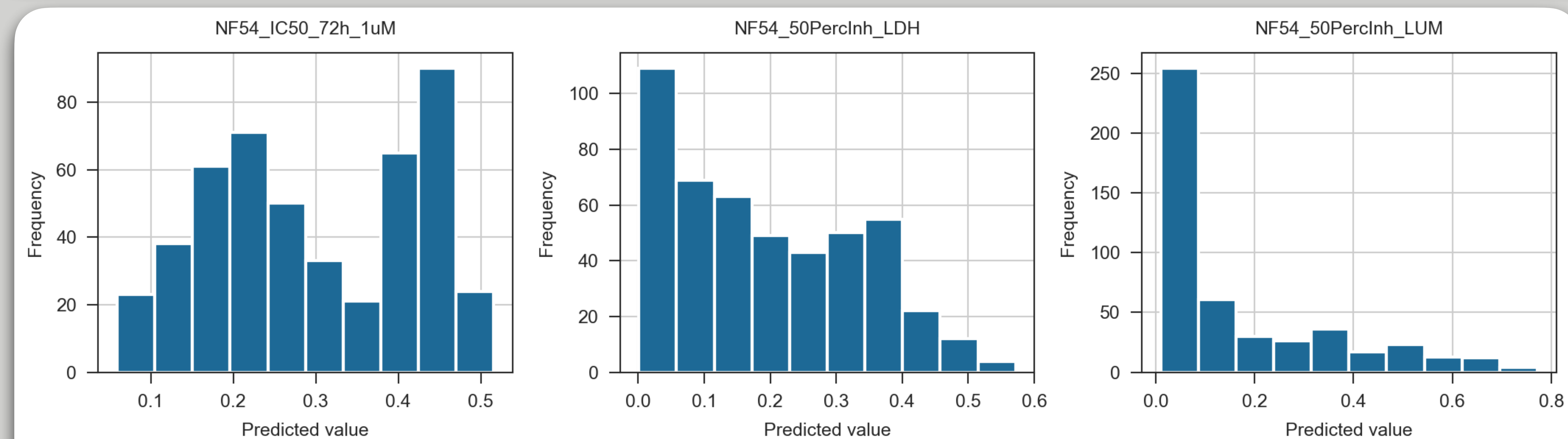
— Relevance to our problem?
— What value do we want to optimise?
— Can we make any assumptions about the applicability domain of the model?

Antimalarial prediction with MMV data (eos4rta)

1	key	input	NF54_IC50_72h_1uM	NF54_50PercInh_LDH	NF54_50PercInh_LUM
2	HWGPBEQLDAATTP-UHFFFAOYSA-N	N#CC1CCCN(C(=O)CCC2=CC=CC(F)=C2)C1	0.19404025869214928	0.007060029655472325	0.02215968
3	VZEQMVMGOXXSDA-UHFFFAOYSA-N	CCC(=O)C1=CN=C2C=CC(C3=CC(Cl)=C(O)C(OC)=C3)=CC2=C1NC1=CC=C(CN(C)C)C=C1	0.37599067020623106	0.2687784437758614	0.56457806
4	XPDWCQMOAYLTHH-CCVNUDIWSA-N	C/C(=N\NC(=O)C1=NC2=C(C(=O)N1)C1CCCN1C(=O)N2C1=CC=CC=C1)C1=CC=C(Cl)C=C1	0.23471794699679766	0.15549179065885121	0.161725
5	WWAFZFZKTQQHTL-ILRYNQFESA-N	CC(=O)N[C@H]1[C@H](SCCCN2C=C(CN3C(=O)C4=CC=CC5=CC=CC(=C45)C3=O)N=N2)O[C@H](CO)[C@@H](O)[C@@H]1O	0.21402212125712888	0.13970746426092914	0.031084577
6	BSKQAAYIGGYUAZ-VGOFMYFVSA-N	OC1=C(/C=N/C2=NC=CS2)C2=CC=CC=C2N1	0.2693705165689197	0.2214959085993597	0.11503028
7	QSNBHLXYLHVCLT-QPPBQGQZSA-N	CC(=O)O[C@@H]1C(C)(C)OC(=O)[C@]12COC1=CC=C3C(=O)C=C(C4=CC=CC=C4)OC3=C21	0.21265540763595348	0.10763128810323211	0.06936009
8	NHCSOGQGYIMJJG-UHFFFAOYSA-N	COC1=CC(NC(=O)CN2N=C3C(SC4CCCCC4)=NC=CN3C2=O)=CC(OC)=C1	0.11331766446698903	0.07980804957251031	0.04113348
9	BUMLEIQSRWKWTF-UHFFFAOYSA-N	COC1=CC=CC=C1N1CCN(CCCCC(=O)NC2CCCC3=CC=CC(OC)=C23)CC1	0.16842902769078913	0.05639335168658563	0.10543706
10	YXYPAHMTJNXFTE-ZVHZXABRSA-N	COC1=CC(/C=C2/SC(N3N=C(C4=CC=CC=C4)CC3C3=CC=CC=C3O)=NC2=O)=CC(OC)=C1O	0.23464995544273506	0.1532669401132952	0.2107249
11	CMXZAXQUIAMXTH-UHFFFAOYSA-N	CC1=CC=C(C2=NOC(CNC(=O)N3CCC(CO)CC3)=C2)C=C1	0.224875905759012	0.01609998922995054	0.029294686
12	OOKWFQQDDHURHZ-UHFFFAOYSA-N	CC(=O)N1C(C2=CC=C(C)C=C2)SC(C)(C)C1C(=O)O	0.1421700090791356	0.006912590037419948	0.023652889
13	RPFGULHOFYSDAK-UHFFFAOYSA-N	CN(C)C(=O)CNCCC1=CC=CC(OCCCCC(F)(F)F)=C1	0.1361932078310655	0.0028507629099750495	0.04296503
14	MEEWKYUFROITOK-UHFFFAOYSA-N	CCC1=CC=CC2=C1C=CC1=C2OC(=O)C2=C1OC=C2C	0.2507538100341272	0.06487105965670041	0.053030923
15	KUFQYQWVZDXHJN-UHFFFAOYSA-N	COC1=CC(C(=O)N2N=C(C)C=C2C)=CC(OC)=C1OC	0.1846704496223233	0.03033886434156533	0.01774607
16	VCVQSRCYSKKPBA-UHFFFAOYSA-N	CC(C)(C)NCC(O)COC1=CC=CC=C1C#N	0.1266497485804591	0.08719589427013674	0.020389792
17	TXOGMSNEULUYAF-UHFFFAOYSA-N	Br.CC1N(C)C2CCCC1(C1=CC=CC(OC(=O)C3=CC=CN=C3)=C1)C2	0.17037676675418562	0.01929204116227353	0.037059054
18	LLQHRNDLBMDQHR-UHFFFAOYSA-N	CS(=O)(=O)N(CC(O)CN1C2=CC(F)=CC=C2C2=CC=C(F)C=C12)C1CC1	0.21495711025965167	0.18772608817362973	0.12503982
19	AEQYZGAQFDSQIR-UHFFFAOYSA-N	CN(CCCOCCOCCC1=CC=CC=C1)CCC1=CC=C(O)C2=C1SC(O)=N2	0.220578258598462	0.21976090369785006	0.098191075
20	ULZOVHDYBVKSJL-HYARGMPZSA-N	COC1=CC(OC)=C(OC)C=C1/C=N/NC(=O)C1=CC=CC(S(=O)(=O)N2CCOCC2)=C1	0.1007243304226347	0.07429913961349913	0.049029544
21	WGHIRYPZICNFTM-UHFFFAOYSA-N	O=C(CC(CC1=CNC2=CC=CC=C12)(NC(=O)OC1C2CC3CC(C2)CC1C3)C(=O)NCCC1=CC=CC=C1)OCC1=CC=CC=C1	0.2563680639893735	0.09914915282881873	0.13676733
22	PBJIOVQCYBRCRK-UHFFFAOYSA-N	CN(CC1=CC=CO1)C1=NC=NC2=CC=C(C3=CC=C4C(=C3)OCO4)C=C12	0.47896541603871234	0.28667421280238103	0.093372725
23	XKIZIFRQOMJEGT-UHFFFAOYSA-N	CC1=CC(C)=C(CNC(=O)C2=CC(C3=CN(C)N=C3)=CC(N(C)C3CCCCC3)=C2C)C(=O)N1	0.2823441970109414	0.1934216569207105	0.18941434
24	UYNMHCOEYXEJMK-UHFFFAOYSA-N	CC1=CC(C)=C(NC2=NC(N)=NC(NC3=CC=C(C#N)N=C3)=N2)C(C)=C1.Cl	0.2635951400723174	0.3172307211761615	0.15022285
25	KCFIWGIFJLSTCC-UHFFFAOYSA-N	NCCCCCNC1=CC=C(NCCCCCN)C2=C1C(=O)C1=CC=NC=C1C2=O	0.14930099137258168	0.22184310934281415	0.121442325
26	VUYHQRNOICWQLK-RGCMKSIDSA-N	CC1=CN([C@@H]2O[C@H](COP(=O)(O)OP(=O)(O)OP(=O)(O)O)[C@@H](O)[C@H]2O)C(=O)C2=CC=CC=C12	0.20440955032358654	0.059101234839148746	0.029773388
27	YHPYKUDCFAWLRI-OAQYLSRUSA-N	COC1=CC=C(S(=O)(=O)NC2=CC=C(N[C@H](C(=O)O)C(C)(C)C)C3=CC=CC=C23)C=C1	0.09641623152644103	0.10141103840031902	0.16634862
28	UEGYOFFNGADXPX-UHFFFAOYSA-N	CC(C)(/N=C(\S)NC1=CC=C(NC(=O)C2=CC=CC=C2F)C=C1)C1=CC=CC=C1	0.15241632634371952	0.015270571433384577	0.26353943
29	OCHZNYFFBSVCIJ-VXKWHMMOSA-N	O=C(C1=CC=CC=N1)[C@@H]1CCCN1C(=O)[C@@H]1CCCN1C(=O)CCCC1=CC=CC=C1	0.24688721782694656	0.033307025324251185	0.03346359
30	XHRBNXOROUSHMG-UHFFFAOYSA-N	C=C(C(=O)C1=CC=C(C)C=C1)N1C=NC=N1	0.11227243173090821	0.0121530577075702	0.067444526
31	XPHIFDXVOGTFLO-UHFFFAOYSA-N	COC1=CC=C(CCNC(=O)CC2=CC=C(Cl)C=C2)C=C1OC	0.06525940013158374	0.021986501972883946	0.067917645
32	KOGBUXRCFBDPLI-FLPBZWPXSA-N	O=C1NC(=O)N([C@H]2CO[C@H](CO)O2)C=C1/C=C/I	0.19131867013179316	0.025953917598287843	0.010016828
33	HJWGSRNLLRXEPX-UHFFFAOYSA-N	CC1=CC=CC=C1N1C(=O)C2=CC=CC=C2N2C(N3CCOCC3)=NN=C12	0.21108852014817692	0.031247588504776925	0.048215564
34	LDAQXINHLRVUBF-UHFFFAOYSA-N	COC1=CC(NC(C)CCCN(CC2=CC=C(Cl)C=C2)C(=O)NC2=CC=CC=C2F)=C2N=CC=CC2=C1	0.186221323861707	0.1196811102470078	0.19354615
35	DCAQIXBZZGJKTP-UHFFFAOYSA-N	O=[N+]([O-])C1=CC=C(C2=CN3C=C(F)SC3=N2)C=C1	0.1384919249078545	0.038394872347370844	0.08510643
36	ABSMFJVWTLRCJI-UHFFFAOYSA-N	CC1=NN(CC(=O)NCCCN2CCC(N3CCCCC3)CC2)C(=O)C2=CC(C3=CC=CC=C3)=NN12	0.367079751755813	0.16346366039185015	0.27601156

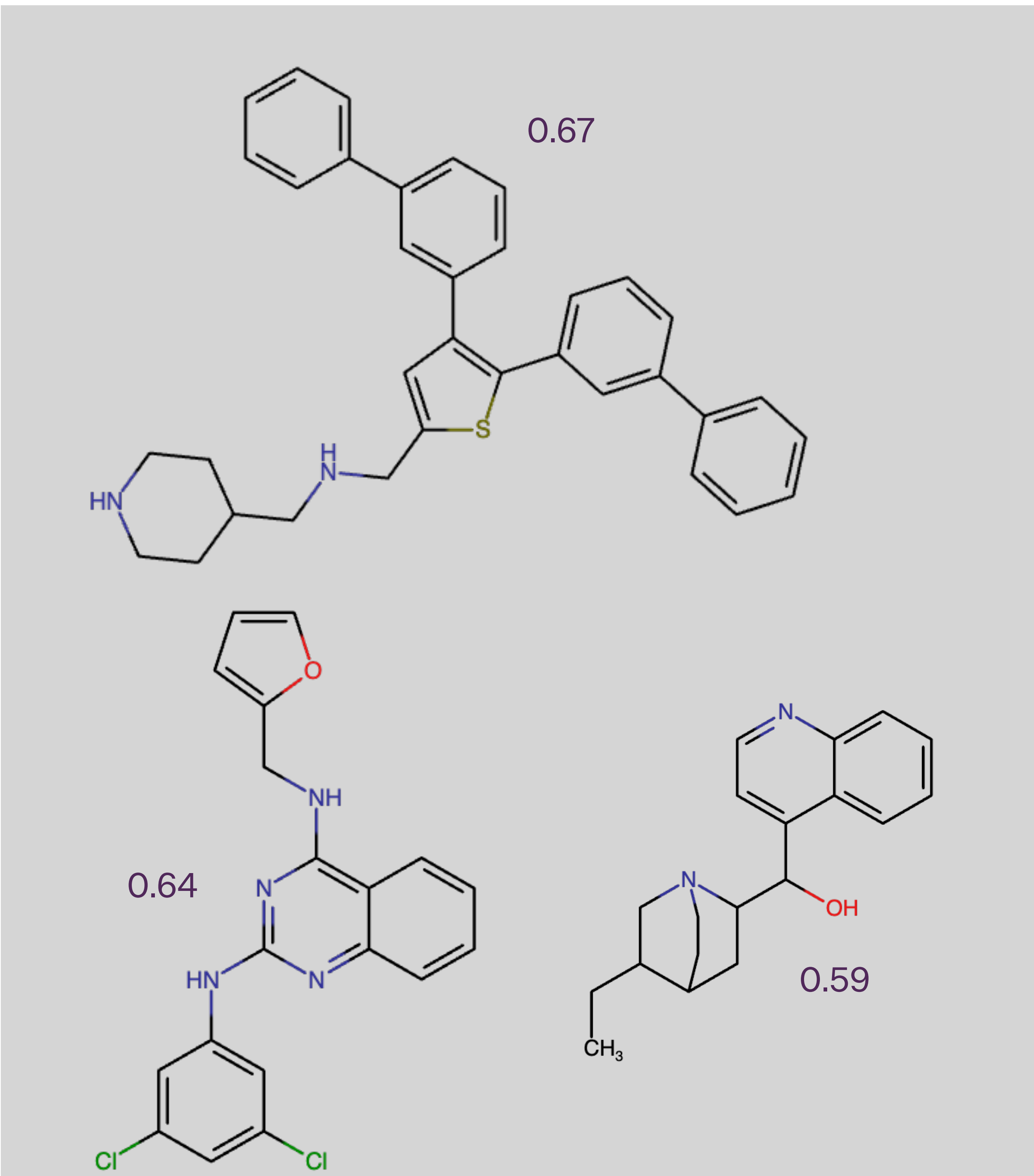
Antimalarial prediction with MMV data (eos4rta)



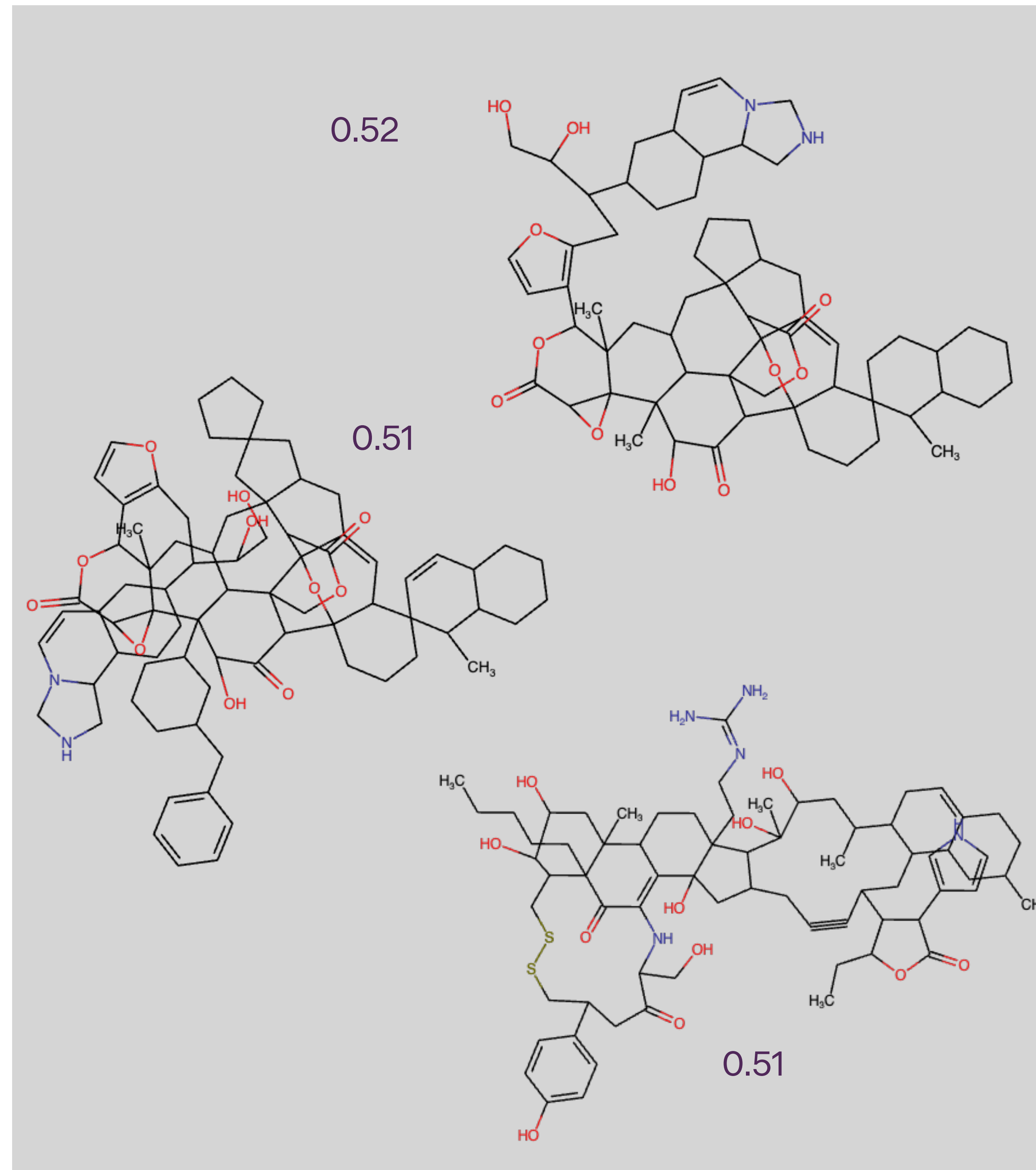


Top molecules

ChEMBL



Coconut



Antimalarial prediction with OSM data (eos7yti)

— Task: Classification
— Output: Probability of killing P.falciparum in vitro (IC50 < 1uM and 2.5uM, respectively)
— Training set: Open Source Malaria

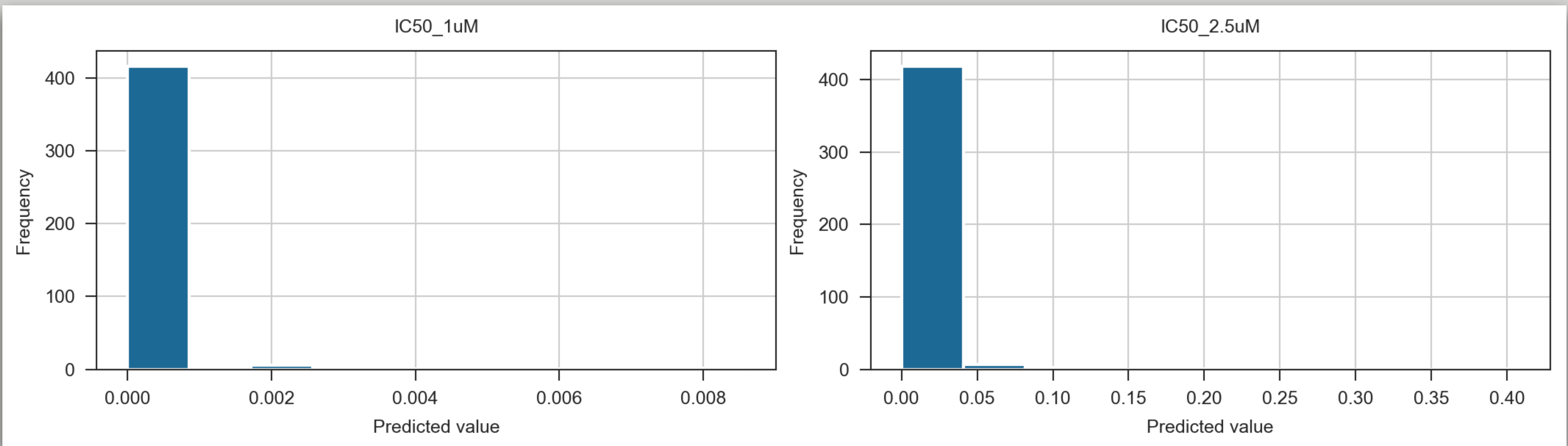
— Relevance to our problem?
— What value do we want to optimise?
— Can we make any assumptions about the applicability domain of the model?

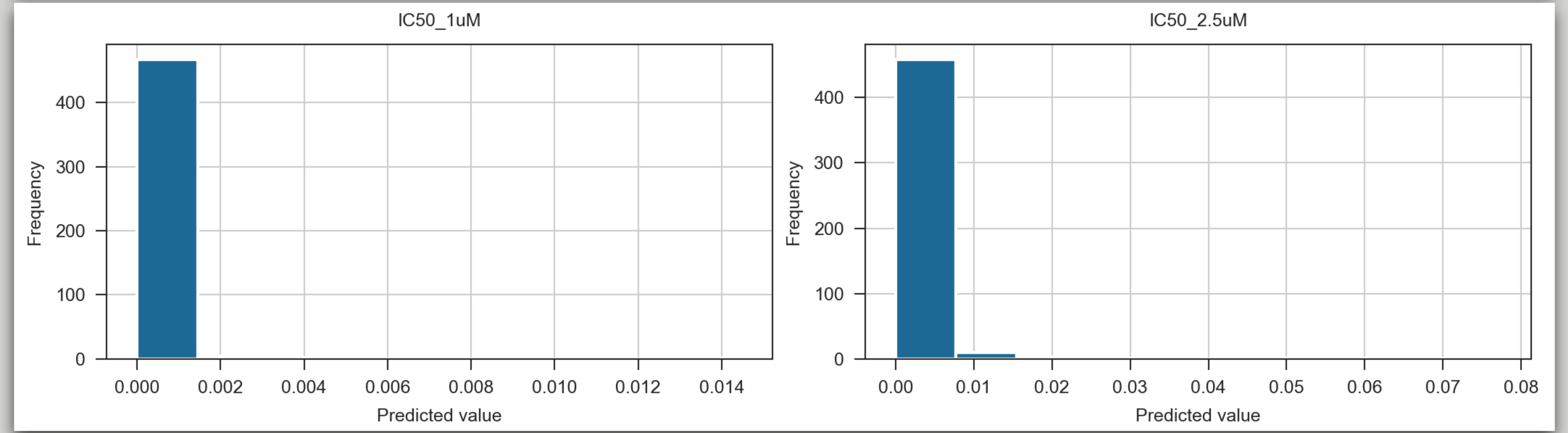
Antimalarial prediction with OSM data (eos7yti)

chembl_selected_eos7yti_predictions

key	input	IC50_1uM	IC50_2.5uM
OLGGEMHVMHDMBQ-CKJJVQESSA-N	C[C@H](C1=CC=CC=C1)N1CC2=C(OC(N)=C(C3=NC(C4=CC=C(Cl)C=C4)=NO3)C2C2=CC=CC3=CC=CC=C23)/C(=C/C2=CC=CC3=CC=CC=C23)C1	0.0004378277290802	0.2171111233365295
JDXKRKYSUPFHSI-UHFFFAOYSA-N	C=C(C1=CC=C(C2=CC=CC=C2)C=C1)C1CCOC2(CCC(NC3=CC=C(C(F)(F)F)C=C3)CC2)OO1	2.214106131100807e-06	0.0029736285656708
ITZCQKLMGVPQPT-ZLYZMSFYSA-N	CC(C)[C@@H]1CC[C@]2(CO)CC[C@]3(C)[C@H](CC[C@@H]4[C@@]5(C)CC6=C(ON=C6)C(C)(C)[C@@H]5CC[C@@]34C)[C@@H]12	7.752534626013493e-07	0.0006278432551692
QVHOWIYXXSOEIV-GHXNOFRVSA-N	CC(NC1=NC2=C(/C=C3\NC(=O)NC3=O)C=NN2C(NC2CC2)=N1)C1=CC=CO1	6.140151853332196e-06	0.0004765144158255
AJOIPROOJINBPD-UHFFFAOYSA-N	NCCCC(=O)NCCC(=O)NC1=NN=C(S(N)(=O)=O)S1	2.0708282359644947e-07	4.1798490204619994e-06
GLNKQEUBUVUTNJ-UHFFFAOYSA-N	O=C1CC(C2=CC=CC=C2)CC2=C1C1(CCCCC1)N=C(NC1=NC3=CC=CC=C3O1)N2	6.079850143390151e-07	0.0039926084967902
WXXKIULLBFLFPW-DVWZZLGYSA-N	COC(=O)CC[C@@H](C)[C@H]1CC[C@H]2[C@@H]3C/C(=N/NC(=S)NC4=CC=C(C)C=C4)[C@@H]4C/C(=N/NC(=S)NC5=CC=C(C)C=C5)CC[C@]4(C)[C@H]3CC[C@]12C	9.185641243027404e-06	3.56490302632571e-05
DVPYLGDJYYVKAJ-CWCCKSQSSA-N	CC(=O)O[C@H]1CC[C@@]2(C)[C@@H](CC[C@]3(C)[C@@H]2CC=C2[C@@H]4[C@@H](C)[C@H](C)CC[C@]4(C(=O)N(C)O)CC[C@@]32C)C1(C)C	5.055493847346875e-07	0.000144560547262
SSZMRTAPOVWYHW-LKYMCVAFSA-N	C[C@H]1C[C@@H](N2C=NC3=C(N)N=CC(F)=C23)[C@H](O)[C@@H]1O	5.852887621547444e-07	0.0004649515449903
NTCUGFMIELJXCS-UHFFFAOYSA-N	CCN1CCN(CCCNC2=C3C(=NC4=CC=CC=C24)C=CC=C3)CC1	1.4935725365451757e-06	0.0015950373564847
ZLXMEKUSNLBNSL-SVEHJYQDSA-N	CC1(C)CC2=NC(C3CCN(C4=NC=C(O)C=N4)CC3)=C([C@@H](F)C3=CC=C(C(F)(F)F)C=C3)C(C3CCC(F)(F)CC3)=C2[C@@H](O)C1	8.045653787310577e-05	0.0456071722231443
BZMUHPHCPVKBAC-FNORWQNLSA-N	O=C(/C=C/C1=CC=CC2=C1N(CC1=CC=C(Cl)C=C1Cl)C(=O)C2)NS(=O)(=O)C1=CC(Cl)=C(Cl)S1	0.0001132368960172	0.0008238488948243
BYCWLOQDZYMODR-DWXRJYCRSA-N	C[C@@H]1CCC[C@H](N2CCC(C3=C(C#N)C=CC(Cl)=C3F)=CC2=O)C2=NC=CC(=C2)C2=CC=CC=C2NC1=O	2.154915507657875e-06	0.0197872593891658
FUYDACWLMIPEJT-UHFFFAOYSA-N	COC1=CC=CC(C2=NC(CN3C=CN=C3C=O)=CO2)=C1	1.4274704509198098e-07	0.0008600318333556
SRNYIEUASJEHNI-UHFFFAOYSA-N	CCCN1C(=O)C2=C(N=C(CCCC3=CC=CC=C3)N2)N(CCCOC)C1=O	8.679527861206315e-07	0.0004936922822348
JHYNXXDQQHTCHJ-UHFFFAOYSA-M	CC[P+](C1=CC=CC=C1)(C1=CC=CC=C1)C1=CC=CC=C1.[Br-]	1.4971119654221377e-07	0.0060150355548496
NQNMTRSTEYEOGI-UHFFFAOYSA-N	CCCCN1N=C(C(=O)C2=CC=CC=C2N)CC1C(=O)OCC	7.567228028676208e-07	2.3651568063984105e-05
XEAKAWDCKUCKBY-AMWOSJAMSA-N	CC(C)C[C@H](NC(=O)OC(C1=CC=CC=C1)C1CCNCC1)C(=O)N[C@@H](CCCNC(=N)N)C(=O)C1=NC2=CC=CC=C2S1	9.66383499794039e-07	0.0012739788702609
WFMZEOAASVEOPI-SPSPGWCGSA-N	CC1CCC2C(=O)N3C(CCC(C)[C@@H]3C3=CC=C(Br)C=C3)C(=O)N2C1C1=CC=C(Br)C=C1	1.8748294560469473e-08	0.0002468533268354
YQGXBSXHFMPWQM-KLCAMILTSA-N	CCC(=O)O[C@H]1CC[C@@]2(C)[C@@H](CC[C@]3(C)[C@@H]2CC=C2[C@@H]4CC(C)(C)CC[C@]4(C(=O)NCCCC(=O)N[C@H](C(=O)O)C(C)C)CC[C@@]32C)C1(C)C	5.643365281574339e-07	2.0450477123839355e-05
QDUPLBCEMNFINR-SOFGYWHQSA-N	O=C(O)/C=C/C1=CC=C2C(=C1)CC1(CC3=CC=CC=C3C1)C2	1.1599784665818626e-07	0.0003513803311255
XWMBNHQWCOHJQH-UHFFFAOYSA-N	FC1=CC=C(C2=C(C3=CC=NC(NCCN4CCSCC4)=N3)SC(C3CCNCC3)=N2)C=C1	1.5966185490990446e-06	0.0006289843175537
DEZJLIXJXQEJDP-WEVVVXLNSA-N	O=C(O)CN1C(=O)S/C(=C/C2=CC=CC([N+](=O)[O-])=C2)C1=O	2.392214453133201e-07	5.505205797345498e-05
YJMQHDDLZWBZTR-UHFFFAOYSA-N	O=C1N=C2C=CC=CN2C=C1CC1=CC=CC(OC2=CC=CC=C2)=C1	4.157655309997209e-06	0.0055286656907683
LSTDAQHMTQYRJL-UHFFFAOYSA-N	Cl.O=C1CCN(CC2=CC=CC=C2F)CC1C(C1=CC=C(F)C=C1)C1=CC=C(F)C=C1	0.000208501589857	0.0081064388920074
GZBJYWLKANIMIX-CDUMDVBJSA-N	C=C(C)[C@@H]1CC=C(CNC2=NC=NC3=C2N=CN3[C@@H]2O[C@H](CO)[C@@H](O)[C@H]2O)CC1	1.0027610436647369e-07	9.45650463370943e-05
JYGCJRGDFNDTFE-UHFFFAOYSA-N	O=C(NC1=CC=CC=C1)N1C2CCC1CC(O)(C1=CC=CN=C1)C2	1.13581563132442e-06	0.0001498427109909
SHLVDRLISHVGSO-ZSQFBXSQSA-N	O=C(NCC1=CC=CN=C1)NC[C@H]1CCC[C@H](OCC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)[C@@H]1C1=CC=CC=C1	2.9846781027030165e-05	0.0042985627781289
LJXCMWIXKXYBMH-IUHHBDENSA-N	CCCCCCCC(=O)N(C)[C@@H](CC(C)C)C(=O)N[C@H](C(=O)N(C)[C@H](C(=O)N1C[C@@H](O)C[C@H]1C(=O)N1C(=O)C=C[C@@H]1C)C(C)C)C(C)OC(C)=O	1.444113463165923e-06	3.486047505518835e-06
AHQFRLVRGMRTIK-RPWUZVMVSA-N	COC1=CC=CC([C@@H]2OC3=CC=C(OC)C=C3C[C@H]2OC(=O)NS(=O)(=O)C2=CC=C(C)C=C2)=C1	5.41299661432453e-06	0.0009151132511432
YBUPKAFNJNXZCU-UHFFFAOYSA-N	CC(C)CSC1=CC2=C(C=C1Cl)C=C(C(=O)O)C(C(F)(F)F)O2	1.4410015854032548e-06	6.016413145911683e-05
OWNKDAHEOJCUPR-UHFFFAOYSA-N	O=S(=O)(CCC1=CC=CC=C1)C(F)F	1.1079218037000825e-06	0.0002231968955587

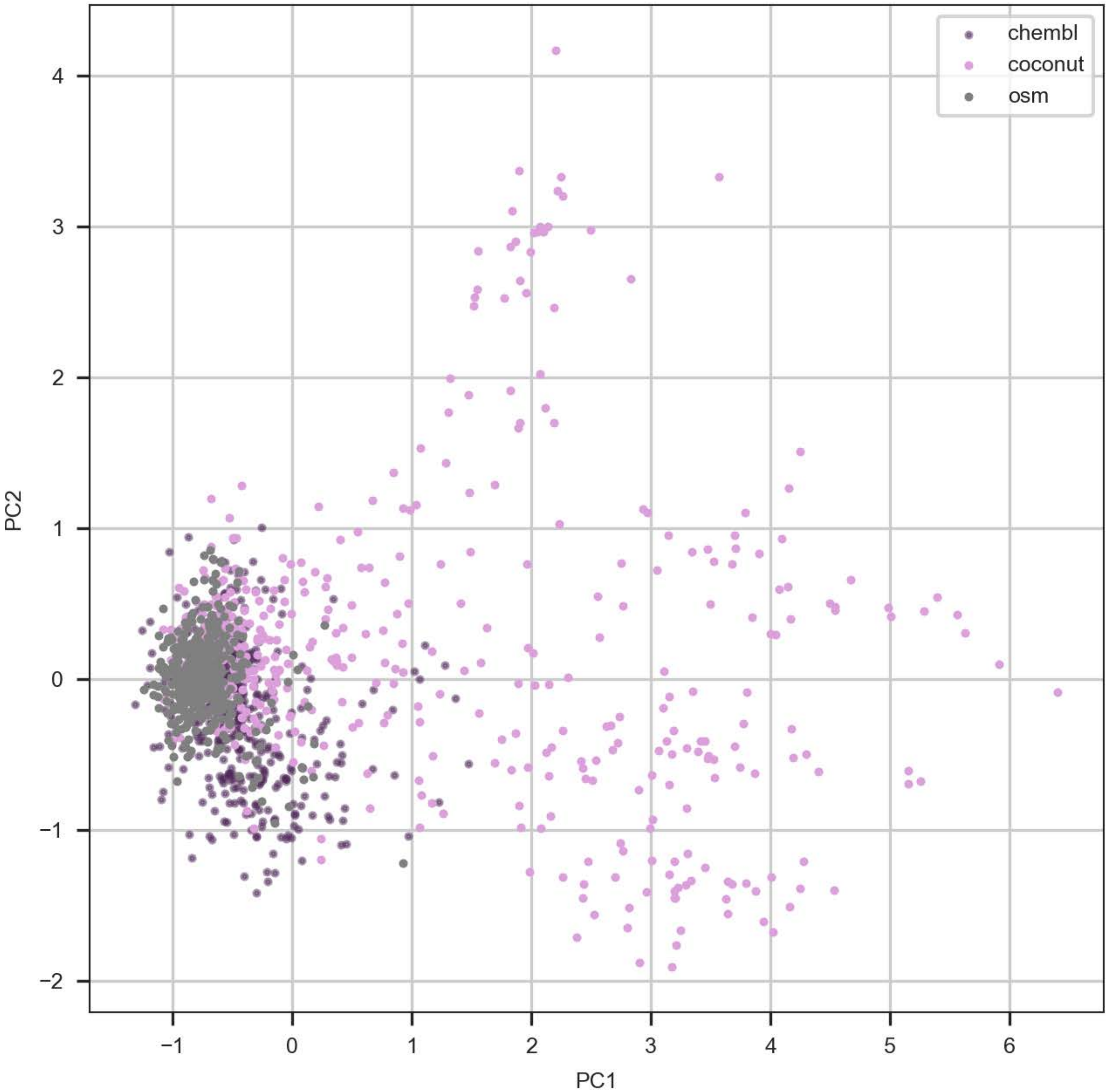
Antimalarial prediction with OSM data (eos7yti)



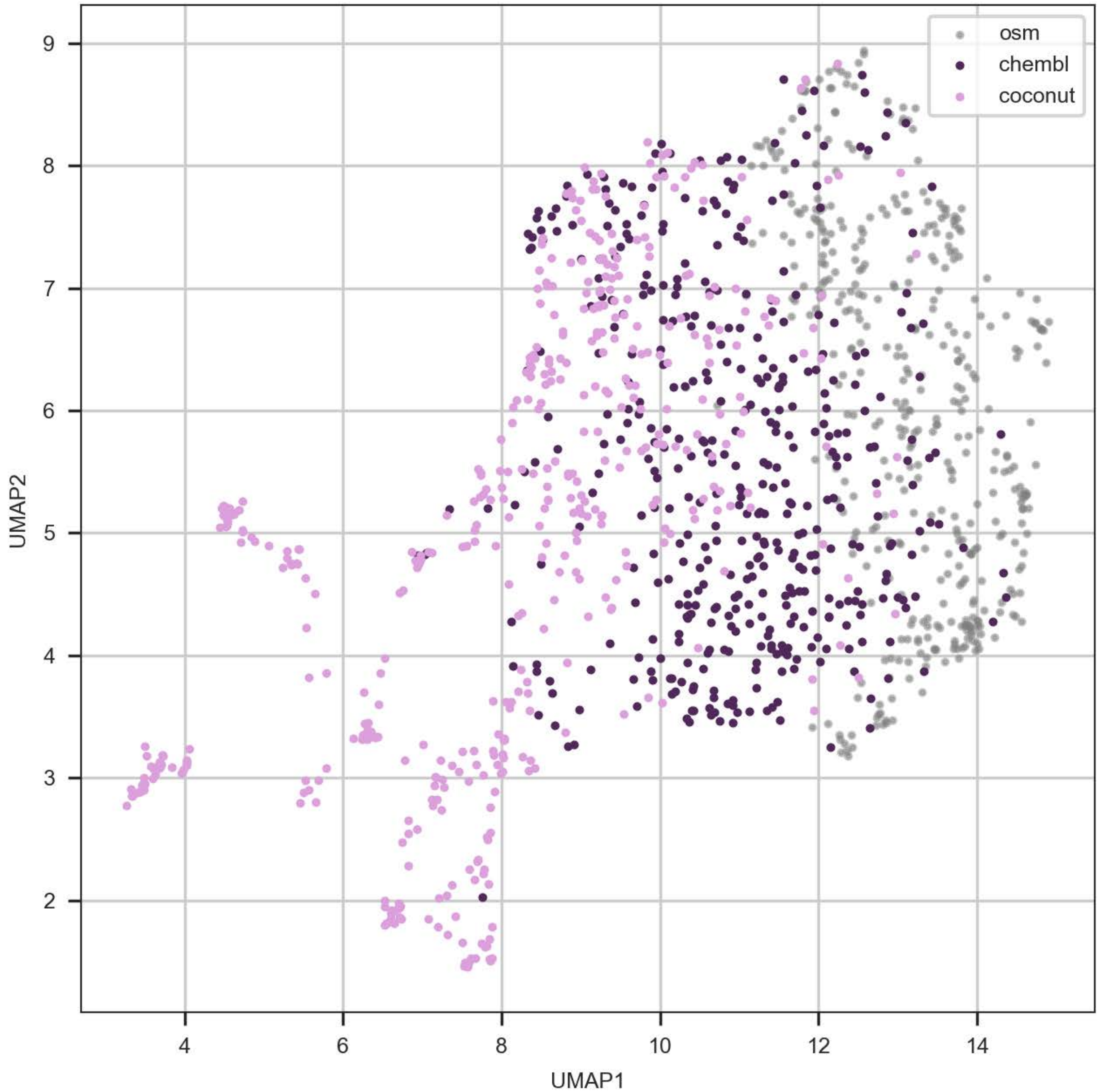


OSM original data - comparison

PCA



UMAP



Selected models

- Antimalarial activity by MMV: eos4rta
- Antimalarial activity by Open Source Malaria: eos7yti
- ChemProp antibiotic: eos4e41
- Antischistosomiasis activity by SwissTPH: eos2l0q
- Cardiotoxicity: eos4tcc
- Cytotoxicity in HepG2 cells: eos3le9
- Solubility: eos6oli
- Synthetic accessibility: eos9ei3
- Natural product score: eos9yui

👩‍💻 Exercise 4

- Let's split up in pairs
- Take up a model and download the predictions for that model
- Look up the information about the model
- Look at the distribution of the activities for your model
- Select three molecules and explain why to the rest

This is an exercise, there is no right or wrong answer

🧑‍💻 Exercise 4 guidance

Step 1: Model prediction & interpretation

For each model, think about the following questions:
— What type of model is it (classification or regression)
— What is the training dataset? (refer to the original publication if possible)
— What is the interpretation of the model outcome?
— What cut-off, if any, we should use for that particular model?
In addition, think about the following concepts:
— Does the outcome of the model make sense? If it does not make sense, perhaps we have the wrong interpretation of the model output
— Is the cut-off I have selected too stringent (i.e, I am losing too many molecules and I should be more permissive?)
— Is this model very relevant for the current dataset (i.e, is malaria activity equally important as natural product likeness?)

Step 2: molecule selection

Use the predicted values to select the 20 molecules that you would take for experimental testing if you had to choose. To that end, you can think of:
— What are the most important activities you want to optimize
— What are strict no-go points
— What are activities that are easiest to optimize at lead stage

Step 3: prepare the presentation

Prepare a short presentation for the other group. This should cover:
— Which models did you choose and why
— What selection strategy did you decide
— Which were your selected molecules

Discussion

An applied example - final exercise

Go to menti.com and introduce 1655 9573

Virtual screening cascade

	Activity	Result	Hit values	Relevance
eos4rta	Malaria	Probability	High	High
eos7yti	Malaria	Probability	High	Low
eos4tcc	Cardiotox	Probability	Low	High
eos3le9	Cytotox	Probability	Low	High
eos6oli	Solubility	LogS	Average	Medium
eos9ei3	Synth.Acc	Score	High	Medium
eos9yui	NP-like	Score	Average	Low

👩‍💻 Exercise 4

- Go to your email inbox
- Download the master file with all the predictions
- Let's prioritise some compounds:
  - Go back to small groups
  - Select 3 compounds that look good according to the predicted activities
  - Present them to the rest of the group, showing their predicted activities, structure...
  - Be critical! No compound will be perfect!

* This exercise is intended solely for training exercises, it is not a real case-study. The molecules have been selected to facilitate discussion

Any questions?

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