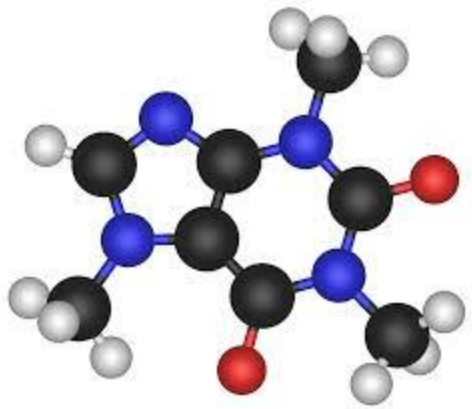
Project Assignment I

ML on Covid bioactivity data

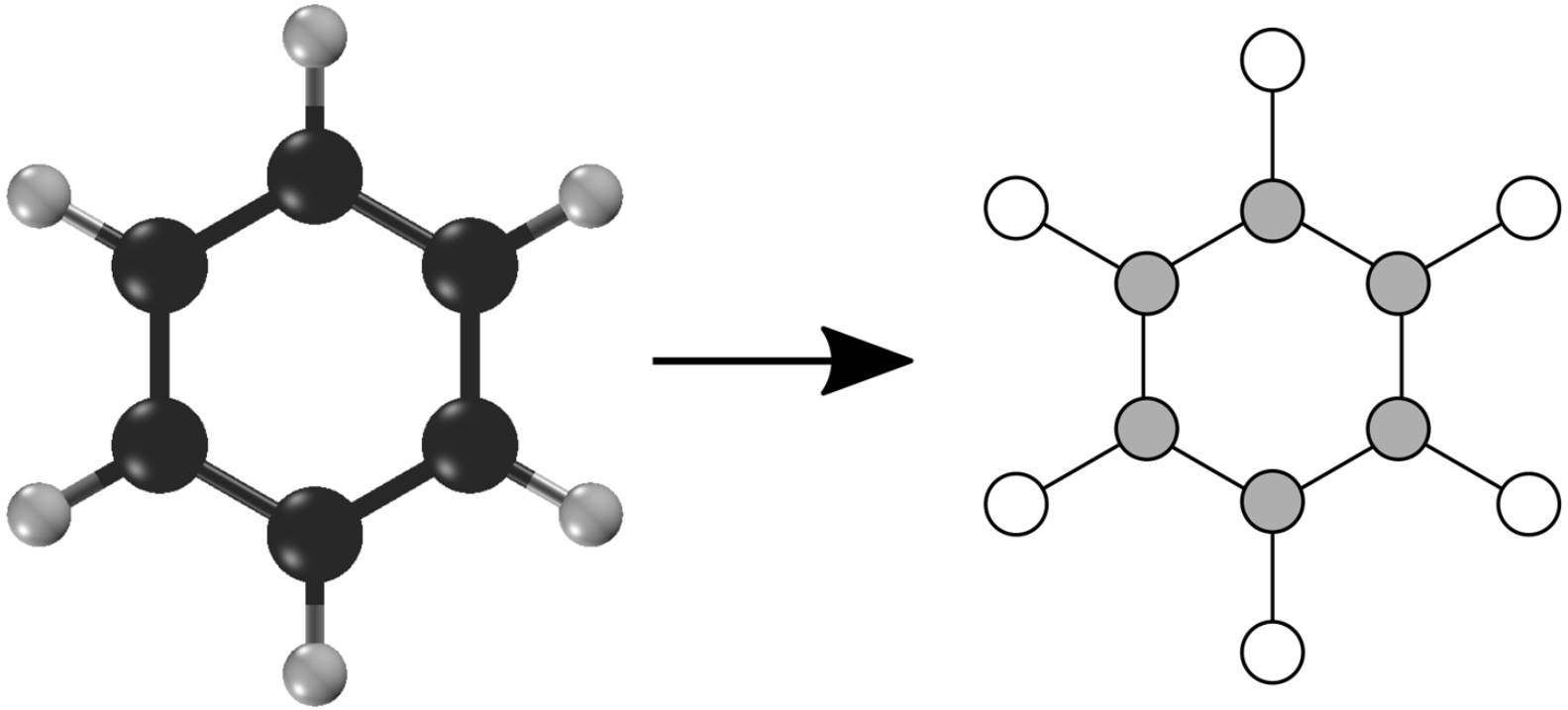
The objective is to build conventional machine learning models such as SVM, Random Forest, Linear Regression etc (and NOT neural networks) to predict Bioactivity values (y=bioactivity).

1. Do the data cleaning, if required
2. Remove the records, if there are missing values
3. Molecules are in SMILES format. Please use the descriptors such as ALogP, PSA etc to build the model
4. Split the data into training data and test data, build the models and estimate the Accuracy
5. If needed goto pubmed and retrieve some papers with keywords Machine learning and Chembl and study how they have used the machine learning and descriptors for building the model.

A simple representation of a caffeine molecule as a "ball-and-stick" diagram.



An example of converting a benzene molecule into a molecular graph. Note that atoms are converted into nodes and chemical bonds into edges.



SMILES Strings

SMILES is a popular method for specifying molecules with text strings. The acronym stands for “Simplified Molecular-Input Line-Entry System”, which is sufficiently awkward-sounding that someone must have worked hard to come up with it. A SMILES string describes the atoms and bonds of a molecule in a way that is both concise and reasonably intuitive to chemists. To nonchemists, these strings tend to look like meaningless patterns of random characters. For example,

“OCCc1c(C)[n+](cs1)Cc2cnc(C)nc2N”

describes the important nutrient thiamine, also known as vitamin B1.

RDkit

Molecular ML methods use SMILES strings as its format for representing molecules inside datasets. There are some deep learning models that directly accept SMILES strings as their inputs, attempting to learn to identify meaningful features in the text representation. But much more often, we first convert the string into a different representation (or featurize it) better suited to the problem at hand.

Molecular ML programs depends on another open source chemoinformatics package, RDKit, to facilitate its handling of molecules. RDKit provides lots of features for working with SMILES strings. It plays a central role in converting the strings in datasets to molecular graphs

Molecular Descriptors

It’s useful to describe molecules with a set of physiochemical descriptors. These usually correspond to various computed quantities that describe the molecule’s structure. These quantities, such as the log partition coefficient or the polar surface area, are often derived from classical physics or chemistry. The RDKit package computes many such physical descriptors on molecules.

https://www.ebi.ac.uk/chembl/

ebi.ac.uk/chembl/

Apps films Sep2019 IISERB Alcourse Week 1 – Lecture: H... Haxel2020 Python ML Haxel2020a Python1 DL ML1 Songs DL1

EMBL-EBI Services Research Training About us EMBL-EBI

ChEMBL

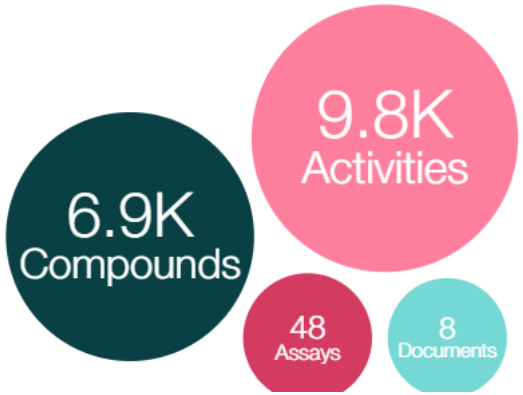
Search in ChEMBL

Examples: Imatinib erbB2 brain MDCK c1ccccc1N

Draw a Structure | Enter a Sequence

UniChem ChEMBL-NTD SureChEMBL Malaria Inhibitor Prediction Downloads Web Services More

ChEMBL is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs.

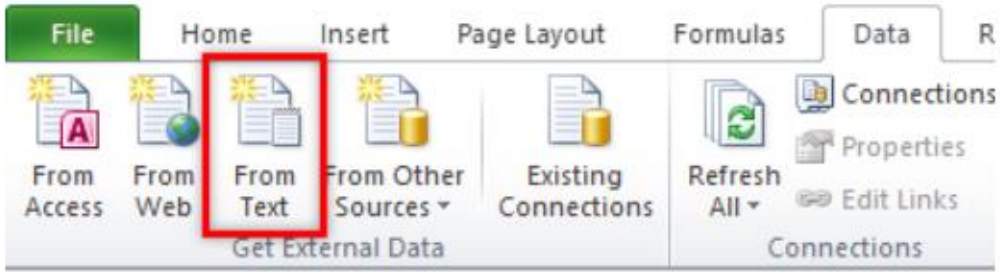


Explore SARS-CoV-2 data

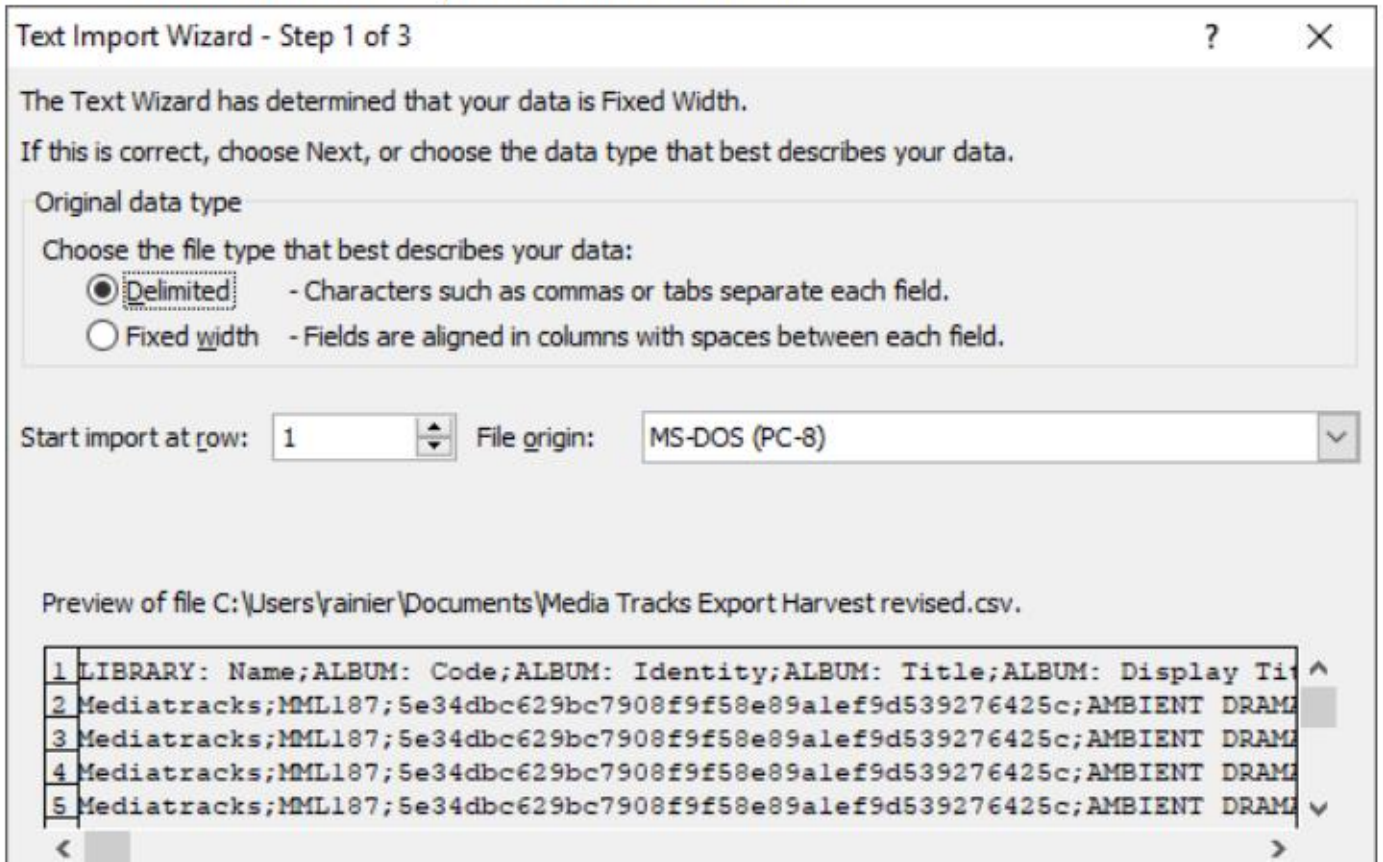
Description: Shows a summary of SARS-CoV-2 related ChEMBL entities and quantities of data for each item.

Instructions: Click on a bubble to explore a specific ChEMBL entity in more detail.

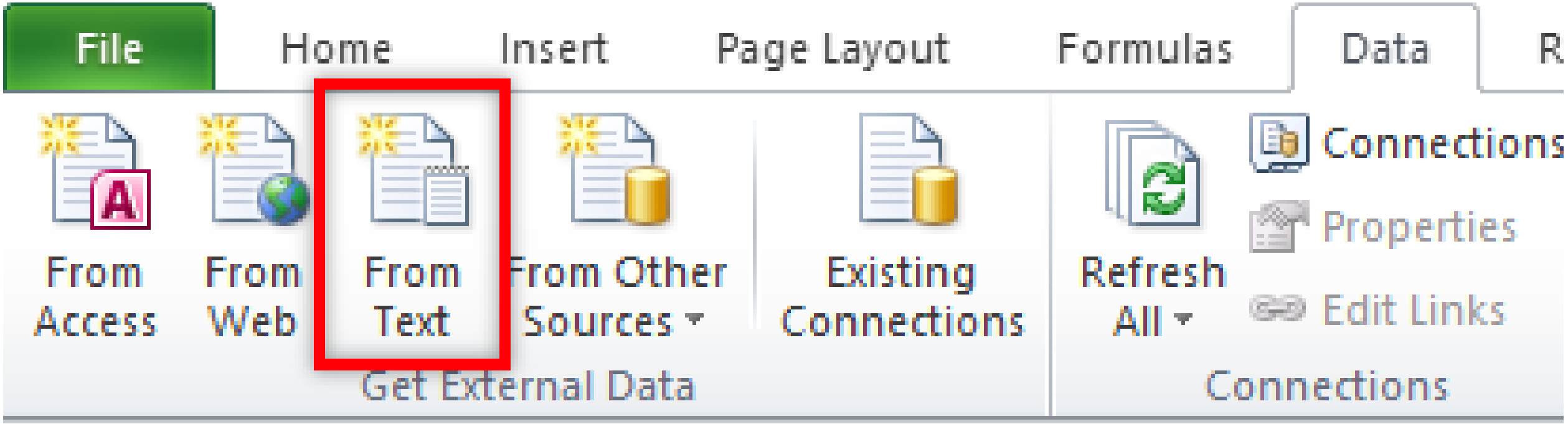
1. Open a new empty spreadsheet in Excel
2. Go to the Data tab and select 'From Text'



3. Select the file you want to open
4. Choose the 'Delimited' option and click Next



https://harvestmedia.zendesk.com/hc/en-us/articles/360023978031-Opening-Excel-files-with-the-correct-CSV-list-separator



Pubmed

Key words: Machine learning, Chembl

https://pubmed.ncbi.nlm.nih.gov/31703452/

ChEMBL ID	Name	Synonyms	Type	Max Phase	Molecular Weight	Targets	Bioactivities	AlogP	PSA	HBA	HBD	#RO5 Violations
CHEMBL3039407			Small molecule	0	454.97	1	1	3.89	72.83	5	1	0
CHEMBL192254			Small molecule	0	257.25	1	1	-2.17	130.83	8	4	0
CHEMBL3545175	PF-03882845	PF-03882845|PF-3	Small molecule	1		1	1					
CHEMBL507327			Small molecule	0	295.73	1	1	3.47	50.68	3	1	0
CHEMBL1187085			Small molecule	0	138.15	1	1	0.21	41.18	1	1	0
CHEMBL603764			Small molecule	0	347.22	1	1	-1.86	186.07	10	5	0
CHEMBL549485			Small molecule	0	424.4	1	1	-1.37	148.82	10	3	0
CHEMBL1236027			Small molecule	0	465.64	1	1	4.77	102.01	5	3	0
CHEMBL471792			Small molecule	0	366.84	1	1	5.51	54.37	3	1	1
CHEMBL4303667			Small molecule	0	308.12	1	1	0.57	33.42	2	0	0
CHEMBL4303310			Small molecule	0	368.39	1	1	3.50	93.06	6	2	0
CHEMBL301922			Small molecule	0	284.36	1	1	-1.13	113.32	4	4	0
CHEMBL423894			Small molecule	0	353.6	1	1	1.65	60.15	5	5	0
CHEMBL4303206			Small molecule	0	1317.46	1	1	None	None	None	None	None
CHEMBL4303150			Small molecule	0	562.51	1	1	5.30	90.46	4	4	2
CHEMBL4303162			Unknown	0		1	1					
CHEMBL4303167			Small molecule	0	1093.33	1	1	None	None	None	None	None
CHEMBL4303326			Small molecule	0	1541.55	1	1	None	None	None	None	None
CHEMBL4303221			Small molecule	0	429.55	1	1	5.65	75.86	7	2	1
CHEMBL4303231			Small molecule	0	276.26	1	1	-0.83	100.74	9	0	0
CHEMBL4303152			Small molecule	0	283.24	1	1	-2.69	159.51	9	5	0
CHEMBL4303340			Small molecule	0	505.02	1	1	4.26	86.42	7	1	1
CHEMBL4303349			Small molecule	0	294.31	1	1	3.38	90.30	6	4	0
CHEMBL4303351			Small molecule	0	311.35	1	1	1.57	76.30	7	1	0
CHEMBL4303226			Small molecule	0	465.48	1	1	4.11	122.91	6	2	0
CHEMBL4303644			Small molecule	0	344.45	1	1	3.63	60.44	4	0	0
CHEMBL3039094			Small molecule	0	330.47	1	1	4.45	43.37	3	0	0
CHEMBL4303508			Small molecule	0	417.51	1	1	3.53	78.43	7	1	0

AB	AC	AD	AE
ht (Monoisotopic)	Molecular Species	Molecular Formula	Smiles
454.1922	NEUTRAL	C24H32ClFO5	CC1(C)O[C@@H]2CC3C4CCC5=CC(=O)CCC5(C)C4(F)C(O)CC3(C)[C@]2(C(=O)CCl)O1
257.1012	NEUTRAL	C10H15N3O5	CC1(O)C(O)C(CO)OC1n1ccc(N)nc1=O
295.0512	NEUTRAL	C16H10ClN3O	O=c1c2c[nH]c3ccccc3c-2nn1-c1ccc(Cl)cc1
138.055	ACID	C7H8NO2+	C[n+]1cccc(C(=O)O)c1
347.0631	ACID	C10H14N5O7P	Nc1ncnc2c1ncn2C1O[C@H](COP(=O)(O)O)[C@@H](O)[C@H]1O
424.1369	NEUTRAL	C20H24O10	C[C@@H]1C(=O)O[C@H]2[C@H](O)C34[C@H]5C[C@@H](C(C)(C)C)[C@]36[C@@H](OC(=O)[C@@H]6
465.3452	ZWITTERION	C23H50N2O5P+	CCCCCCCCCCCCC/C=C/[C@@H](O)[C@@H](N)CO[P@](=O)(O)OCC[N+](C)(C)C
366.1023	ACID	C22H19ClO3	O=C1C(O)=C(C2CCC(c3ccc(Cl)cc3)CC2)C(=O)c2ccccc21
181.0972	NEUTRAL	C9H13IN2O2	CN(C)C(=O)Oc1ccc[n+](C)c1.[I-]
368.126	NEUTRAL	C21H20O6	COc1ccc(C2Oc3c(CC=C(C)C)c(O)cc(O)c3C(=O)C2=O)cc1
284.1848	BASE	C13H24N4O3	CC(C)CC(NC(=O)C1CCCN1)C(=O)NCC(N)=O
353.3518	BASE	C20H43N5	CCNC/C=C\CNCCCCNCCCCNC/C=C\CNCC
1166.5589	None	C70H80N10O16	CN1C[C@H](C(=O)N[C@]2(C)O[C@@]3(O)[C@@H]4CCCN4C(=O)[C@H](Cc4ccccc4)N3C2=O)CC2c3ccc
561.1961	BASE	C29H34Cl2FN3O3	CC(C)(C)CC1NC(C(=O)NC2CCC(O)CC2)C(c2cccc(Cl)c2F)C12C(=O)Nc1cc(Cl)ccc12
1092.6431	None	C52H88N10O15	CCC(C)CC(C)CCCCCCCCC(=O)N[C@H]1C[C@@H](O)[C@@H](NCCN)NC(=O)[C@@H]2[C@@H](O)CCN2
1540.6628	None	C63H112O42	CC(O)COC[C@H]1OC2O[C@@H]3[C@@H](COCC(C)O)OC(O[C@@H]4[C@@H](COCC(C)O)OC(O[C@@
429.1623	NEUTRAL	C24H23N5OS	CC(C)n1cnc2c(NCCc3ccc(O)cc3)nc(-c3csc4ccccc34)nc21
276.0971	NEUTRAL	C11H12N6O3	Cc1nc(Cn2cnc3c2c(=O)n(C)c(=O)n3C)no1
283.0917	NEUTRAL	C10H13N5O5	Nc1nc2c(ncn2[C@@H]2O[C@@H](CO)[C@@H](O)[C@H]2O)c(=O)[nH]1
504.2041	NEUTRAL	C27H29ClN6O2	Cc1c(C(=O)Nc2ccc(N3CCC(N4CCOCC4)CC3)c(C#N)c2)cnn1-c1ccc(Cl)cc1
294.1117	NEUTRAL	C16H14N4O2	Oc1cccc(Nc2ccc(Nc3cccc(O)c3)nn2)c1
311.1382	NEUTRAL	C16H17N5O2	Cn1ncc2c(N3CCOCC3)nc(-c3cccc(O)c3)nc21
465.0882	ACID	C24H19NO7S	CS(=O)(=O)Nc1ccc(-c2coc3cc(OCc4cccc(C(=O)O)c4)ccc3c2=O)cc1
344.1988	NEUTRAL	C21H28O4	CC(=O)O[C@H]1CC[C@@]2(C)C(=CC(=O)[C@H]3[C@@H]4CCC(=O)[C@@]4(C)CC[C@@H]32)C1
330.2195	NEUTRAL	C21H30O3	CC(=O)O[C@H]1CC[C@@]2(C)C(=CCC3C4CCC(=O)[C@@]4(C)CCC32)C1
417.2165	BASE	C24H27N5O2	COCCOc1ccc2c(c1)ncn2-c1ccc2cccc(N3CCC(N)CC3)c2n1
366.1368	NEUTRAL	C24H18N2O2	O=C1c2ccccc2-c2c(NCCc3ccccc3)c(=O)[nH]c3cccc1c23