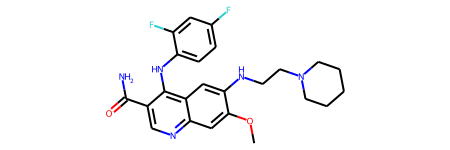
**TECS PhD Machine Learning Rotation: What causes a drug/molecule to be lipophilic and can we predict it?**



**Lipophilicity of 0.295**

**ML**

Lipophilicity is an important feature of drug molecules that affects both membrane permeability and solubility. The lipophilicity dataset, curated from ChEMBL database, provides experimental results of octanol/water distribution coefficient (logD at pH 7.4) of 4200 compounds.

See:  
<https://deepchem.readthedocs.io/en/latest/api_reference/moleculenet.html#lipo-datasets>  
And references in there.

**Aims:**

1. Have a look at the dataset, are there any patterns to it?
2. What sort of chemical features are associated with lipophilicity?
3. Train a series of models to predict toxicity, can you beat the state of the art?

This problem is a regression problem: the output will be a number.

You will need to play around with a few ML models, and change hyperparameters to get the best model. Don’t forget to do more than one run for each model so you can display the results with error bars. The state of the art for this dataset is given in DeepChem and the MoleculeNet paper, and possibly in the references for the dataset. Chemical focussed ML algorithms are in deepchem, or sklearn contains general machine learning algorithms. It is worth trying out a few different featurizations as well.

You will need to use rdkit to get measures of chemical features.

See   
<https://moleculenet.org/>

And   
<https://deepchem.readthedocs.io/en/latest/index.html>

There’s a considerable amount of freedom as to where you take this, but you will want to show off trained predictive models, have a poke around the literature, be prepared for questions from non-machine learning chemists about the chemistry and questions from ML chemists about how you trained the models.