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

A molecule structure with red and blue lines

Description automatically generated

**ML**

**Not Toxic**

**😊**

The “Toxicology in the 21st Century” (Tox21) initiative created a public database measuring the toxicity of compounds, which has been used in the 2014 Tox21 Data Challenge. This dataset contains qualitative toxicity measurements for 8k compounds on 12 different targets, including nuclear receptors and stress response pathways. Each group will take 5 of the targets to look at.

Random splitting is recommended for this dataset.

The raw data csv file contains columns below:  
“smiles” - SMILES representation of the molecular structure  
“NR-XXX” - Nuclear receptor signaling bioassays results  
“SR-XXX” - Stress response bioassays results

please refer to <https://tripod.nih.gov/tox21/challenge/data.jsp> for details.

This problem is a classification problem: Toxic/non-toxic.

**Aims:**

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

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

There is 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.

**Assessment:** You will be asked to prepare an infographic of your findings as a group and present your infographic in a press conference style presentation where you will provide a short statement of your findings (approx. 5 minutes) then answer questions from the audience (approx. 20 minutes).