

## Machine learning project for cheminformatics

This project will help you learn how to perform some basic machine learning and modelling techniques using python and packages such as pandas, SKLearn and RDKit. In this project, you will be attempting to replicate the multiple linear regression model presented in <https://pubs.acs.org/doi/abs/10.1021/ci9901338>, as well as attempting to improve it using new molecular descriptors and different modelling techniques. It is recommended that participants work in partners, particularly for the later notebooks.

### Outline:

1. Introductory notebooks and exercises:
  - i. Jupyter.ipynb
  - ii. Python.ipynb
  - iii. ModellingStatistics.ipynb
2. Loading and preprocessing data
  - i. LoadingData.ipynb
  - ii. RDKit.ipynb
3. Multiple Linear Regression main project
  - i. MultipleLinearRegression.ipynb
4. Additional (optional) extras:
  - i. MultipleLinearRegression.ipynb

### To begin

Simply cd into the directory containing the data files (.csv) (student\_ml\_project) and type the command:

```
jupyter notebook
```

This will allow you to open up a browser and direct to the Jupyter port where you will find a 'notebooks' folder. This will contain all the .ipynb files listed above.

Click on Jupyter.ipynb to begin.