# Project plan

You have received a QSAR dataset extracted from ChEMBL. It consists of two columns, SMILES strings and binding affinities. You are allowed to adapt code from the internet, e.g. ChatGPT. Create a well documented Jupyter notebook in which you perform the following tasks:

- Investigate the dataset
    - Make a plot of the distribution of the binding affinities in your dataset
    - Calculate and plot the distributions of molecular weight and log P for your dataset
    - Make a plot of the chemical space spanned up by your dataset. Choose a molecular descriptor and make a, *e.g.*, a PCA or UMAP plot.
- Data preprocessing
    - Choose and calculate features for one or more types of molecular descriptors.
    - Scale features to unit variance.
    - Discretize your dataset to a binary classification problem, i.e. create a label for each object to be 0 or 1. You can for example use a binding affinity of 10 μM as threshold.
- Supervised machine learning modelling
    - Split the dataset into 10% for hyperparameter tuning and 90% for modeling.
    - Train supervised models for KNN and Random Forest and perform hyperparameter tuning on the 10% dataset.
    - Perform k-fold crossvalidation on the 90% dataset using the tuned hyperparameters.
    - Make a table with accuracy, AUROC and F1 score for each modeling method.
    - Plot AUROC and make a confusion matrix of F1 scores.

**Deliverables**

- Hand in your well-documented jupyter notebook
- Create a poster for the poster presentation