
Responsible person: Henri Xhaard, Computational Drug Discovery

**Background:** The basis of QSAR/QSPR modelling is the assumption that similar molecules present similar activities or properties. “Similarity” between two molecules can be represented using a single numerical value, a “distance”. There are several ways to calculate a “distance” between two molecules, for example from a set of numerical “descriptors” themselves derived from the chemical structures, or by quantifying the amount of common chemical substructures. Among a dataset, molecules within the closest distance are called “neighbors”.

The \( k \)-nearest-neighbors (\( k \)-NN) method for QSAR/QSPR modelling is a novel method for the mathematical prediction of drug properties or activities (Zhen and Tropsha, 2000): for a given molecule, \( k \)-NN QSAR bases its predictions on only its \( k \) neighbors, not on the whole dataset as “classical” QSAR/QSPR do. \( k \)-NN QSAR/QSPR modelling have been shown to be a robust, conceptually simple method, and relatively easy to implement method.

**Project:** Using datasets present at the Centre for Drug Research (pharmacokinetic properties, transporter binding affinity, etc), the candidate will implement a \( k \)-NN method for QSAR/QSPR modelling using the software MATLAB. The idea is to make accessible the \( k \)-NN modelling in MATLAB to other groups at the CDR. In brief, implement mathematical functions and develop the code that:

- Calculate a descriptor-based distance between all pairs of molecules
- Select the \( k \) neighbors and compute a QSAR model based on this distance
- Measure the predictivity of the the model (\( r^2 \))
- Analyze the influence of diverse parameters (for example \( k \), or the distance selection methods, or the descriptors used) on the QSAR/QSPR model obtained (\( r^2 \))

In addition, if times permit,

- Use of \( k \)-NN methods for calculating an Applicability Domain, i.e. a mathematical function that rejects or accept molecules in the test set given the the molecules that were present in the training set. In other word, use the similarity principle to say whether a QSAR/QSPR model is applicable to a certain molecule.
- Experiment another type of QSAR where a distance threshold \( d \), not a discrete number of neighbour \( k \), serves as a basis for predicting properties

**Prerequisites:**
None. The mathematical functions necessary for the project have already been published and there is no need for the candidate to “invent” new ones. The candidate will learn how to use the software MATLAB as well as simple programming. The candidate will start with already computed molecular descriptors and during the project will learn how to compute new ones.
