**Chemoinformatics analysis for computer-aided drug design (CADD**

This workflow consists of 5 interconnected work blocks, each containing a central theme in computer-aided drug design (CADD).

The workflow is illustrated using as an example Acetylcholinesterase (AChE) and its respective active compounds, associated with IC50 values coming from the CHEMBL database, but can be easily extended for other targets of interest. The chemoinformatics analysis allows:

* List active and inactive compounds (according to the selected IC50 threshold).
* Evaluate the bioavailability of compounds and properties that adversely affect their absorption, distribution, metabolism, and excretion (ADME) based on chemical structure alone.
* Group compounds according to their structural similarity to evidence patterns of activity or construct new compounds through scaffold fusion. From this grouping, a diverse set of compounds can also be selected for further analysis.
* Calculate the maximum substructure that a set of molecules have in common.
* Find a central structure (scaffold) within all active compounds and identify its substituents at certain binding positions.

1. **Requirements**

* Knime version 4.3.1 or higher, a programming software through functional workflow Knime website
* Our Knime workflow to capture relevant information about an indication of interest using different databases PPI-network.
* Input files to run the workflow Inputs
* Output files obtained using Alzheimer's disease as an example for the user to compare results Outputs
* Here you can have a complete visualization of the workflow Workflow Visualization

1. **Data acquisition from ChEMBL**

Information on compound structure, bioactivity, and associated targets are organized in databases such as ChEMBL, PubChem, or DrugBank. Data were obtained from ChEMBL where all compounds associated with the target of interest (Acetylcholinesterase AChE) that had an IC50 value were filtered together with the respective Smiles of the molecules and the phase of the study in which they are found.

Active compounds were separated from inactive compounds using the following filter:

Active IC50 < 5 µM

Inactive IC50 >5 µM

1. **Molecular filtering: ADME and lead-likeness criteria**

Not all compounds are suitable starting points for drug development due to their undesirable pharmacokinetic properties, which for example adversely affect the absorption, distribution, metabolism, and excretion (ADME) of a drug. Therefore, these compounds are usually not included in data sets targeting promising drug candidates. Therefore, fewer drug-like molecules in the data set should be eliminated.

The bioavailability of a compound is an important property of ADME. The Lipinski rule of five was introduced to estimate the bioavailability of compounds based solely on their chemical structure. Ro5 states that malabsorption or permeation of a compound is more likely if the chemical structure violates more than one of the following rules:

* Molecular weight (MWT) <= 500 Da.
* Number of hydrogen bond acceptors (HBA) <= 10
* Number of hydrogen bond donors (HBD) <= 5
* Calculated LogP (octanol-water coefficient) <= 5
* Subsequently, those compounds complying with three or four of Lipinski's rules were filtered out for further chemometric analysis.

1. **Compound clustering**

Clustering can be used to identify groups of similar compounds, in order to pick a set of diverse compounds from these clusters for e.g. non-redundant experimental testing. The following steps show how to perform such a clustering based on a hierarchical clustering algorithm.

1. **Maximum common substructures**

To visualize the shared scaffolds and thus emphasize the extent and type of chemical similarities or differences of a cluster of compounds, the maximum common substructure (MCS) can be calculated and highlighted. In this workflow, the MCS was calculated for the four significant clusters obtained from the previous node using the FMCS algorithm.

1. **R-group Decomposition**

R-group decomposition is a special type of search that aims to find a central substructure (scaffold) and identify its substituents at certain binding positions. The query molecule consists of the scaffold and the binding sites represented by R-groups.

This block of the workflow shows how to perform R-group decomposition using the RDKit community extension. Its implementation consists of several steps.

I. Calculate MCS taking all active compounds as a starting point.

II. Perform the R-group decomposition

III. Find how many molecules with each combination of the two selected R-groups are in the data set.

IV. Visualize the results of the R-group decomposition.

1. Clustering can be used to identify groups of similar compounds, based on a hierarchical clustering algorithm.