Cheminformatics

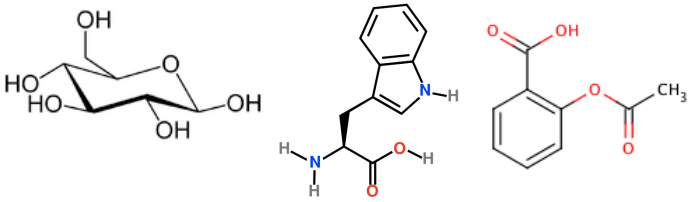
Practical work in class

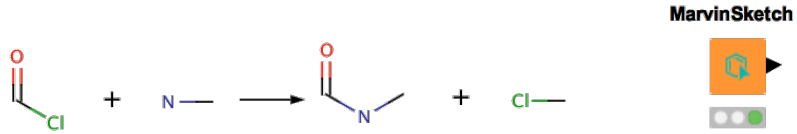
Jean-Loup Faulon
jean-loup.faulon@inra.fr

Playing with molecules and reactions

Tip: look at the workflow snapshots

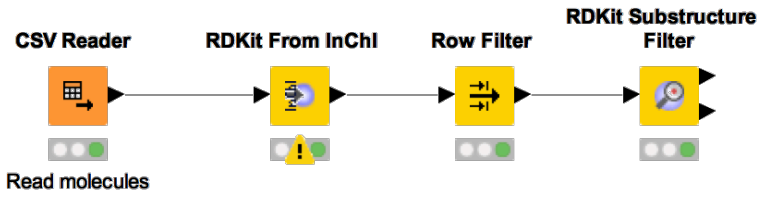
1. Give the SMILES of the following structures and reactions (stereochemistry included). The structures can be drawn using the node MarvinSketch. SMILES are found in Edit/Source of MarvinSketch, perform an atom-atom mapping for the reaction with MarvinSketch





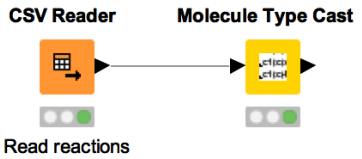
2. Reading and playing with molecules

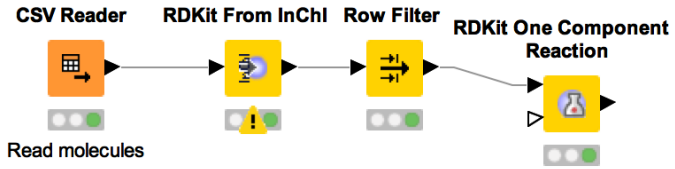
- Write a workflow visualizing the structures of the file Molecule-ecoli.cvs
- Filter out molecule having an InChi=None, how many molecules have been removed?
- Compute the SMILES for each remaining molecule
- How many molecules contain oxane (tetrahydropyran) as a substructure?



3. Reading and playing with reactions

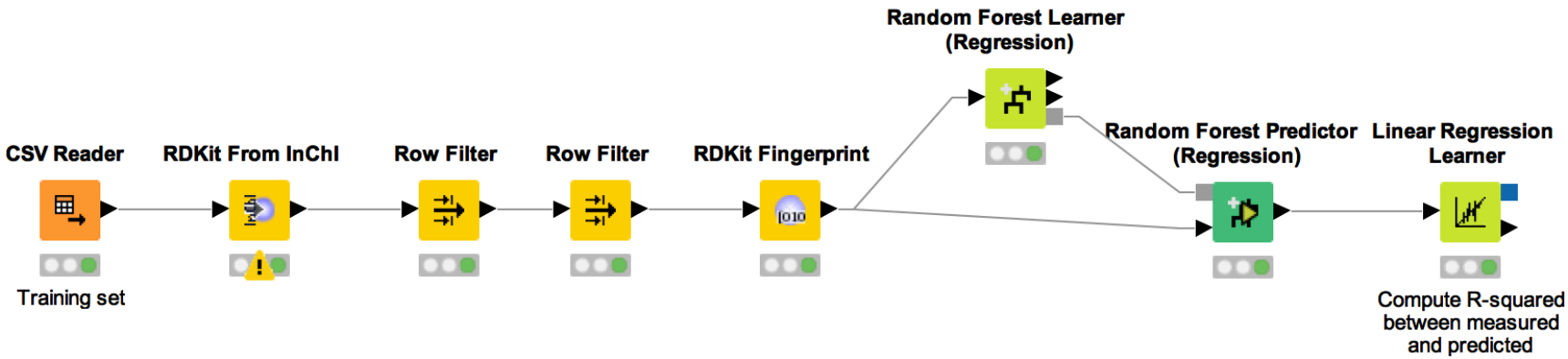
- Write a workflow visualizing the reactions of the file Reaction-rules.cvs
- Fire the first reaction of the EC class 2.3.1 on all valid molecules in Molecule-ecoli.cv. Tip: the reaction SMARTS is
  ([#16]-[#6:2](-[#6,#1:4])=[O:1])>>([H][#8]P(=O)([#8][H])[#8]-[#6:2](-[*:4])=[O:1])





Building a QSPR/QSAR

1. Write a Knime workflow that reads the file Tg-training.
2. Compute RDKit Fingerprint for each molecule
3. Perform a regression using random forest learner between the fingerprints and the Tg (glass transition temperature) values. What is the value for R2 ? (Tip: see snapshot of the workflow below)
4. Perform a LOO cross validation with the same dataset, using the X-partitioner and X-aggregator nodes. What is the Q2 value ?
5. Perform a 10-fold cross validation with the same dataset, using the X-partitioner and X-aggregator nodes. What is the Q2 value ?
6. Repeat the process using MACCS and Morgan fingerprints and answer questions 3-5 again



HomeWork assignment

Jean-Loup Faulon
jean-loup.faulon@inra.fr

Answer all questions in one single file saved in pdf and send it to jean-loup.faulon@inra.fr. The name of the file should be your-name.pdf. Please provide for each answer the question number. Please to not send zip or tar as I will not look at these.

HomeWork assignment (1)

Instructions to install:

1. Download and unzip workflow from https://www.myexperiment.org/workflows/5029.html
2. Download and unzip data from https://www.myexperiment.org/files/1927.html
3. Click on RetroPath2.0-Mods-iQSAR-v2.knwf to run Knime and once open save the workflow in your local workspace
4. Move the Aminoglycosides-Data in a folder you can easily access (desktop is fine)
5. Some of the questions given below can more easily be answered reading the comments and references provided with the workflow

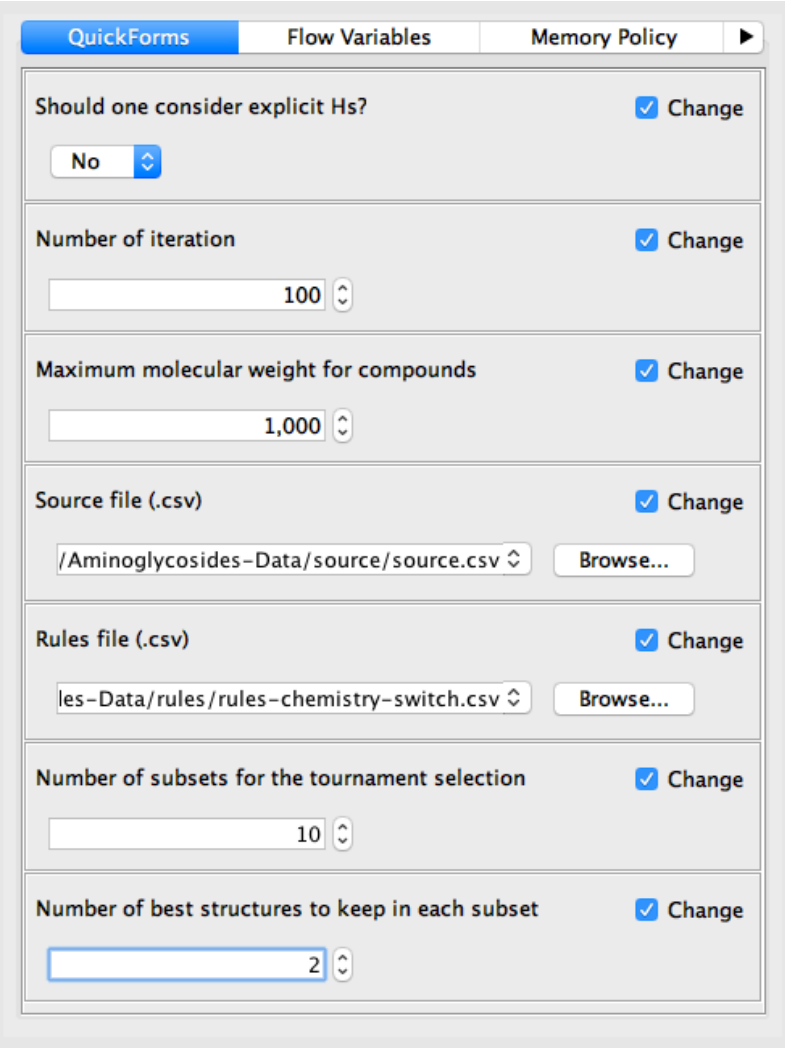
Questions:

1) (4 points) In the Aminoglycosides-Data/rule folder, draw the SMART reaction rule you find the in file. What is this rule suppose to do? Give an example of any product obtained applying the rule on Bisphenol A.

2) (4 points) Write a workflow to compute the SMILES strings of the first 10 molecules in the Aminoglycosides-Data/training-set folder. Report your results in a table giving the name, InChI and SMILES of each molecules.

3) (4 point) Open the QSAR node in the workflow. Which descriptors are used to represent molecules, which parameter is being predicted? Which machine learning method is used to perform the prediction?

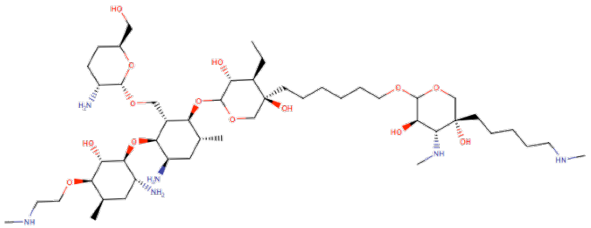
4) (2 point) Run the RetroPath2.0-Mods-iQSAR-v2 workflow with the parameters shown in the Figure (beware the process takes about 30 min). Briefly describe what is the workflow doing?

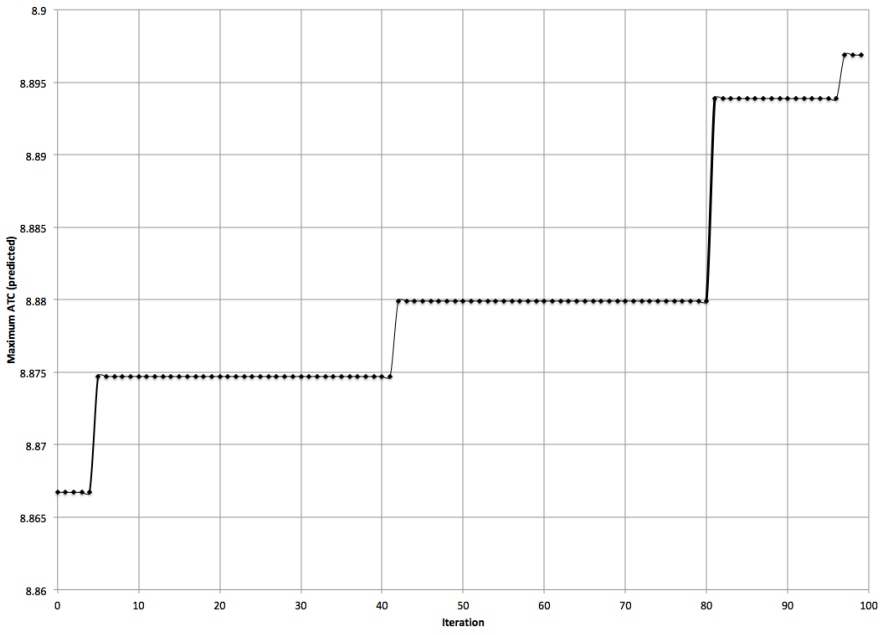


HomeWork assignment (2)

Questions:

5) (4 points) Draw the curve showing the maximum property predictions obtained vs. iteration number from the results obtained running the RetroPath2.0-Mods-iQSAR-v2 workflow. Draw the molecules having the highest predicted value at the first iteration and the last iteration (see examples in the Figure provided)





6) (2 points) Modify the RetroPath2.0-Mods-iQSAR-v2 such that molecules are represented by the 'Morgan' descriptor. Draw the curve showing the maximum property predictions obtained vs. iteration number. You might want to set up the maximum number of iteration to 50 as it takes time to run. Draw the molecules having the highest predicted value at the first iteration and the last iteration.