AUTOMATED INTEGRATION OF STRUCTURAL, BIOLOGICAL AND METABOLIC SIMILARITIES TO SUSTAIN READ-ACROSS

An automated procedure for the selection of analogues for data gap-filling is made available here. Given a target compound, analogues are identified with a decision algorithm that integrates three orthogonal similarity metrics, considering toxicologically relevant aspects:

1. **Structural similarity** based on structural fingerprints
2. **Biological similarity** based on biological fingerprints codifying the outcomes of high throughtput screening (HTS) assays from PubChem
3. **Metabolic similarity** based on the presence of common metabolic pathways between the target and the analogue as simulated by SyGMa library (https://github.com/3D-e-Chem/sygma) Structural filters based on the presence of maximum common substructures (MCS) and common functional groups can be also applied to narrow the chemical space for the analog(s) search. Read-across prediction is made for the target based on the activity of analogue(s) identified with multiple similarity methods.

INSTALLATION

HOW TO INSTALL THE WORKFLOW

1. Install the last version of KNIME. It can be downloaded at https://www.knime.com/knime-analytics-platform (Windows version, 64bits).
2. Open KNIME.
3. Go to *"File -> Import KNIME Workflow"*.
4. Tick *"Select File:"* and go to *"Browse..."*. Select the *.knwf file of the workflow.
5. Click to *"Finish"*.
6. The workflow now is in your "KNIME Explorer" menu on the left of the screen. Double click on the workflow to open it.
7. If some of the plugins used for the workflow are missing, a message will appear asking you to install the missing extensions. Click on *"Ok"*. The procedure will guide you in the installation of the missing extensions. Restart KNIME to make the new plugins working.
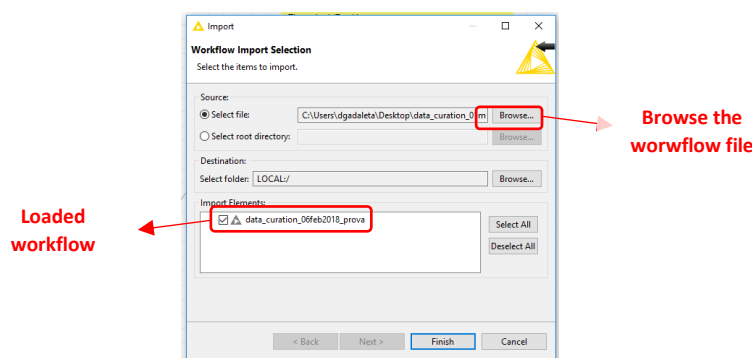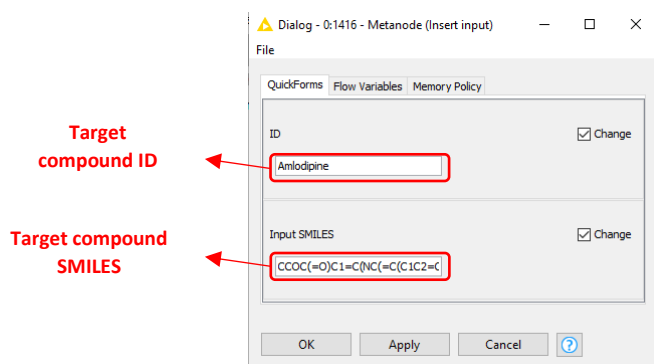
USE THE WORKFLOW

HOW TO USE THE WORKFLOW

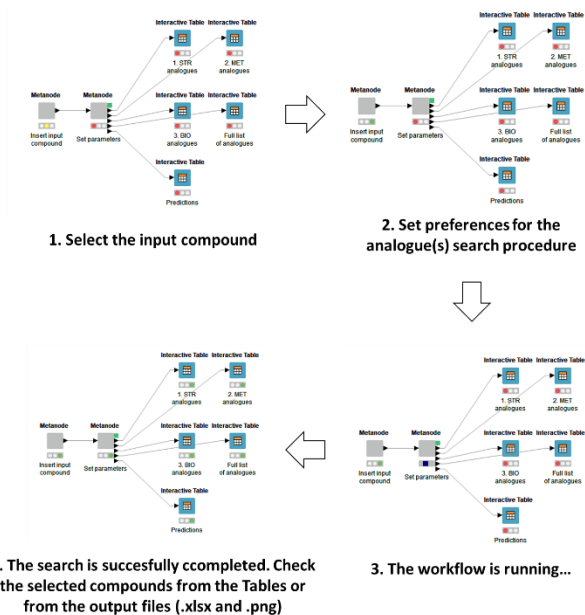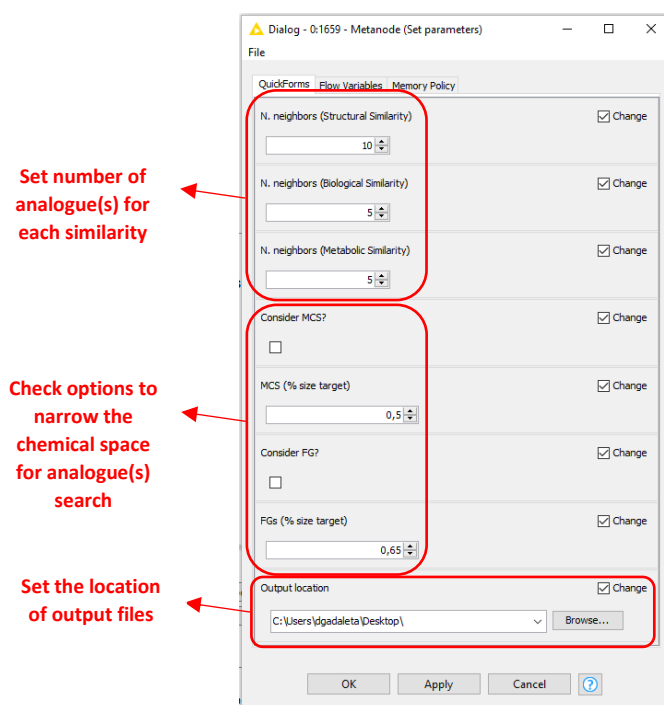1. Load the input file in the *"Insert input compound"* wrapped metanode:

- Double click on the metanode.
- Type the ID and the input SMILES.
- Clink on *"Ok"*.

2. Execute the workflow by clicking on the ⏩ button on the top of the window or press Shift+F7.
3. Modify the settings in the second wrapped metanode, "*Set parameters*".

- Double click on the metanode.
- Select the number of analogue(s) to consider for each of the three similarity. By default, ten analogue(s) are retrieved based on structural similarity and five based on metabolic and biological similarity.
- If appropriate, check the *"Consider MCS?"* and/or *"Consider FG?"* options to narrow the chemical space to search analogue(s). If the options are checked, it is possible to set the thresholds ("*% size target*") to include a chemicals in the analogue(s) selection.
- Select the path of the output file. By default, a new directory named *"RAX_results"* is created on the Desktop, that includes output files.
- Clink on *"Ok"*.
- Execute the workflow by clicking on the ⏩ button on the top of the window or press Shift+F7. The "Set parameters" node will change from the "Paused" state (🕐) to the "Running" state (⏩). If the first part of the workflow is executed successfully, the node will change to the "Executed" state (✅). If the node return to the "Paused" state (🕐), some problems occurred during the procedure.
- Press 💾 on the top-left of the window to save the workflow.

Set number of analogue(s) for each similarity

Check options to narrow the chemical space for analogue(s) search

Set the location of output files

1. Select the input compound

2. Set preferences for the analogue(s) search procedure

4. The search is succesfully ccompleted. Check the selected compounds from the Tables or from the output files (.xlsx and .png)
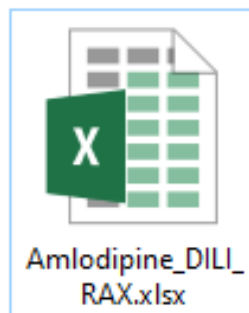
3. The workflow is running...

HOW TO READ THE OUTPUT

1. To inspect the list of selected analogue(s) within KNIME, select the *"Interactive Table"* nodes, then click on the 🔍 button on the top of the page or press F10. To inspect the RAX plot, right click on the "Set parameters" metanode, then select the 🔍 *"Port 1"* option from the drop-down menu.

2. To inspect the results reported in the output files, go to the *"RAX_results"* directory that is present in the output location specified before. By default, the *"RAX_directory"* is created on the Desktop. The directory include two files:

Amlodipine_DILI_RAX.png   Amlodipine_DILI_RAX.xlsx

.png file that shows the list of analogue(s) as a pie chart. as a pie chart. Each slice of the chart represent a different analog, with the background color indicative of the analog's activity and histogram bars that give estimation of the four similarities with respect of the target (i.e., blue for structural, pink for biological and green for metabolic similairty). Outlined slices indicate analogues included in multiple similarity lists (i.e., dark blue for three lists and light blue for two lists).

.xlsx file including four different sheets:

- **Structural analogues:** list of closest neighbor compounds based on structural similarity. For each analogue, the structural similarity with respect of the target is reported. By default, ten analogues are selected. If checked, the presence of MCS and common FGs are verified here (yes/no), and a list of functional groups identified for each analogue is reported.

- **Biological analogues:** list of closest neighbor compounds based on biological similarity. For each analogue the following information are reported: the number of biological assays that are negative (Cn) or positive (Cp) for both the target and the analogue; the number of biological assays that are positive for the target but negative for the analogue (Tp); the number of biological assays that are negative for the target but positive for the analogue (Sp); the biological similarity ($BioSim$); the reliability associated to the BioSim (i.e., based on the number of assays in common between the target and the analogue); the final BioSim value weighted on the reliability.

- **Metabolic analogues:** list of closest neighbor compounds based on metabolic similarity. For each analogue the following information are reported: the metabolic similarity value; the number and the list of metabolic pathways that are present only in the target, only in the analogue or in both compounds; the number and the list of common metabolites between the target and the analogue.

- **Full_list_analogues:** complete list of all the previous analogues that includes all the information above. For each analogue, the list(s) in which it appears and the relative ranks are indicated (e.g. STR1, BIO3 is the top-ranked structural analogue and the third-ranked biological analogue). Analogue(s) included in multiple similarity lists are considered more suitable with respect of those included in a single list.

- **Predictions:** reads-across predictions. Prediction of activity for the target is made by averaging activities of analog(s) included in at least one, two or three similarity lists (*"group_threshold"*). For each prediction, the number of analogues used to sustain the prediction and their activity is indicated. If structural filters based on MCS and common functional groups are used, this is specified in the *"Source_filter"* column ("*none*" is displayed when no filters are used).

REFERENCE

Further details on the algorithms used to calculate similarities can be found in the reference pubblication:

CONTACT

Domenico Gadaleta, PhD
Computational Toxicology Unit
Laboratory of Environmental Chemistry and Toxicology
Department of Environmental Health Sciences
Istituto di Ricerche Farmacologiche Mario Negri IRCCS
Via Mario Negri 2, 20156 Milano, Italy
Tel. +39 02 3901 4396 e-mail: domenico.gadaleta@marionegri.it