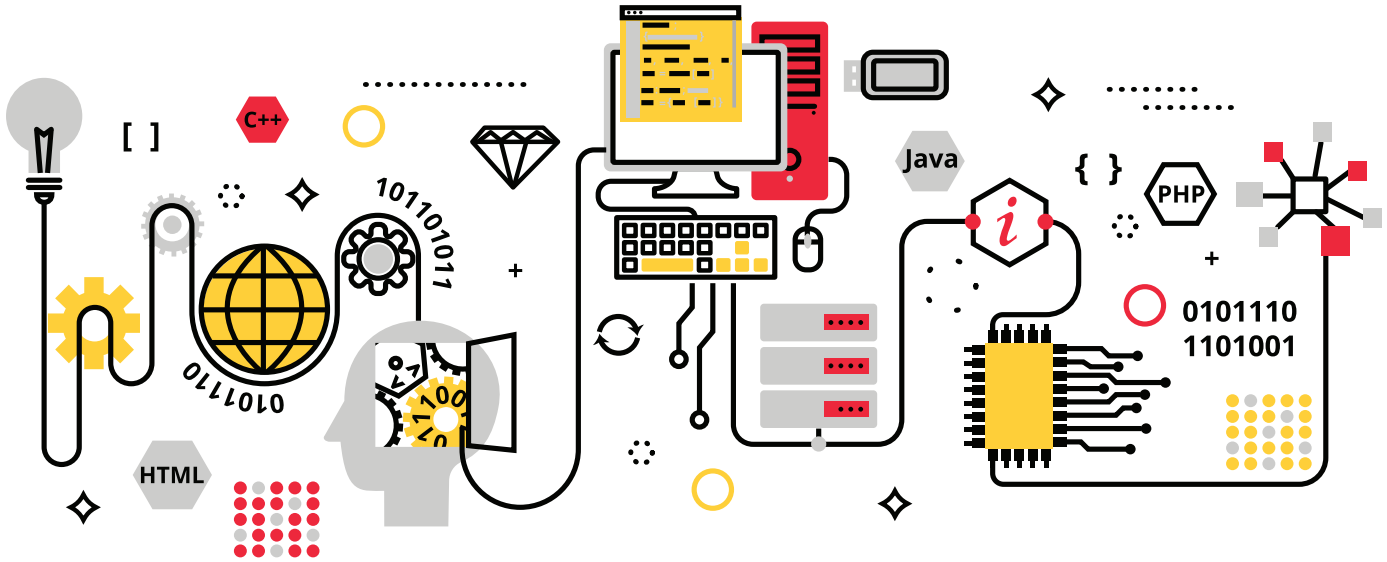
wellcome centre anti-infectives research



KNIME: Reaction Enumeration

Wellcome Centre for Anti-Infectives Research (*WCAIR*)

TRAINING

contents

1.0 About this exercise

In this exercise we will make use of a data analytics software known as KNIME to perform a 'Reaction Enumeration'. We will virtually create a collection of compounds – amides in this case – by reacting primary amines and carboxylic acids from a commercial catalogue. With this collection in place, we will analyse the results, filter them to remove molecules with unwanted features and then de-risk them using the ChEMBL database to check if there are any known compounds like the ones we generated in the exercise.

Installing KNIME (Konstanz Information Miner): **www.knime.com/downloads**

Create a new folder in your "Documents" called "knime-workspace".
Ensure you download the following into this folder:

- A KNIME workflow (.knwf) file, **"WCAIR_01_Reaction_Enumeration.knwf"**
- An SDF file, **"sigma-aldrich-zinc.sdf"** that contains a prepared version of the Sigma-Aldrich commercial building blocks catalogue obtained from the ZINC database

1.1 KNIME Interface

The software has its screen divided in panes (**Figure 1**).

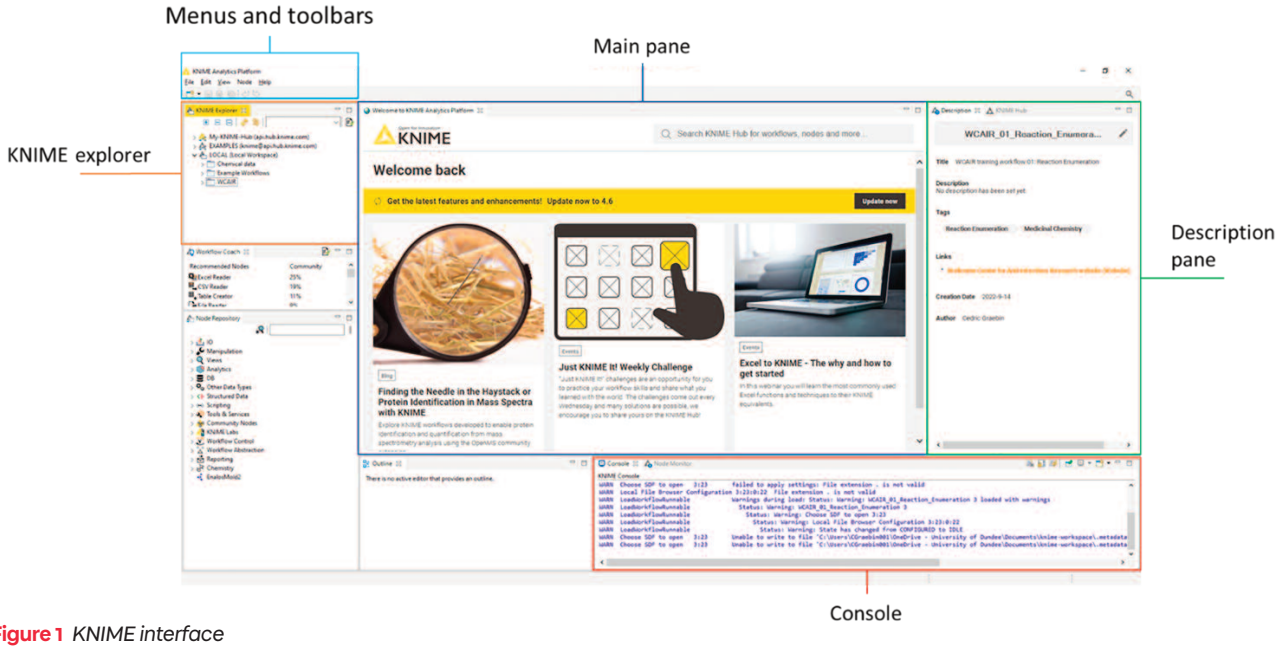


Figure 1 *KNIME interface*

The **KNIME explorer** is where the workflow files will appear when imported. The workflow itself will appear on the **Main pane**. The **Description pane** will show the description of the tasks being executed and the **console** will output any error and warning messages that may appear.

2.0 Download and set-up KNIME

Download KNIME at **www.knime.com/downloads**

Select the "KNIME Analytics Platform" and use the "self-extracting archive" option if you don't have administrator privileges on your machine. This option only copies the necessary files onto your computer without changing configurations in your system.

Note: If you're using the self-extracting setup, you must run KNIME manually by finding "knime.exe" in the folder in which you installed the software.

2.1 For self-extracting or zip setup only (important):

When you run KNIME for the first time, the software will configure itself to save all workflows from a folder in which you may not have full control, meaning that KNIME may fail if you need to save or write temporary files. As a workaround, we will change KNIME's standard saving point to one in which there will be no error messages.

With KNIME open, click on "File" > "Switch Workspace" > "Other..." (**Figure 2**).

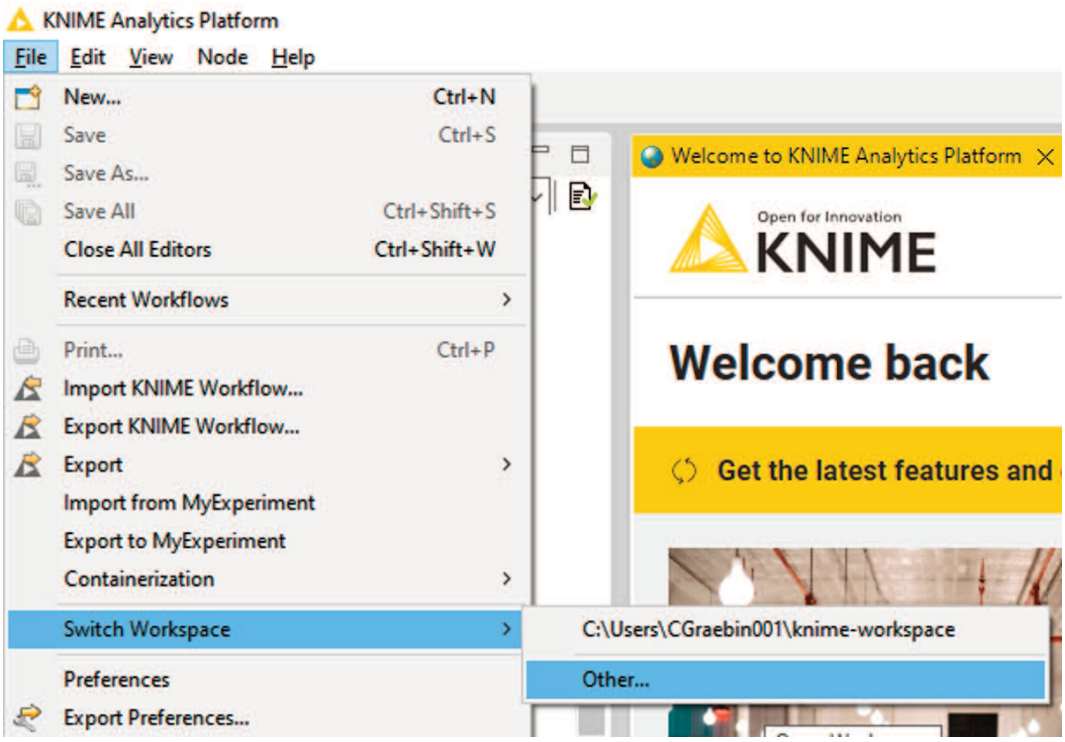


Figure 2 *KNIME menu and workspace switching*

A dialog box will open (**Figure 3**).

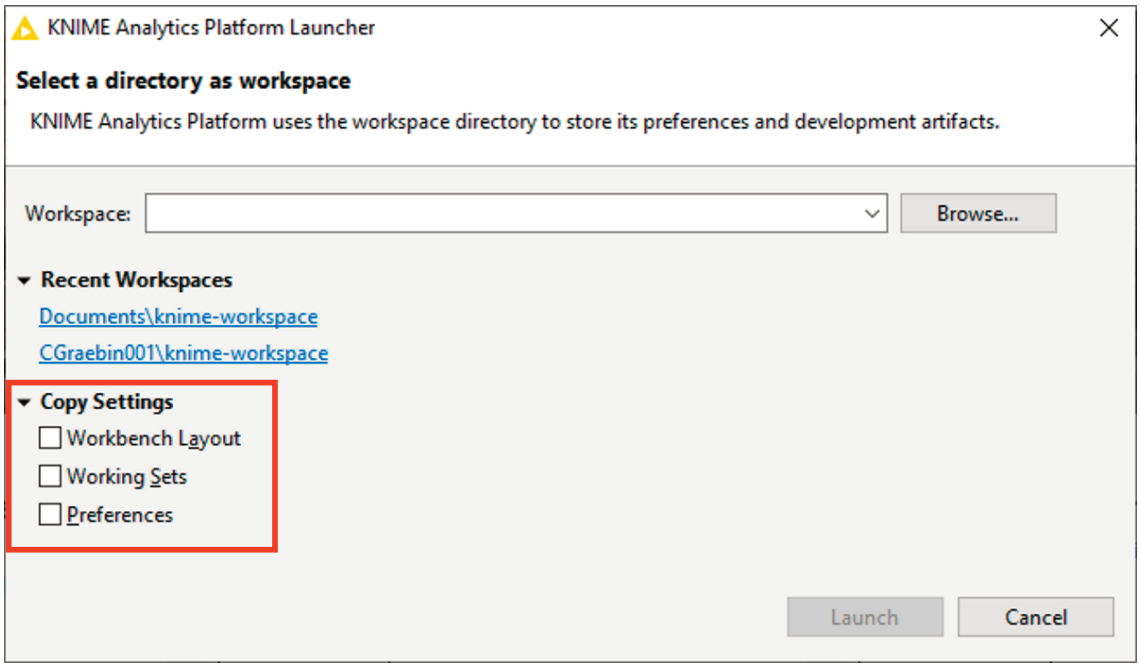


Figure 3 *KNIME Analytics Platform Launcher. Use "Browse" to locate your workspace.*

2.2 To choose a new workspace folder

On the "Select a directory as workspace" dialog box (**Figure 3**), click "Browse > Locate "knime-workspces" folder > "Select Folder".

On the KNIME Analytics Platform Launcher window, make sure to check all the options in the "Copy Settings" before pressing "Launch". After some moments, the data will be copied to the new location and KNIME will ask you to restart the software for the changes to take effect. If you check or have checked before the "don't ask me again" option that appears, KNIME will restart automatically from now on. Accept the restart and wait for KNIME to launch again. Now we are ready to continue to configure KNIME.

2.3 Opening the Workflow

Click on “File” > “Import KNIME Workflow...” (**Figure 4**)”.

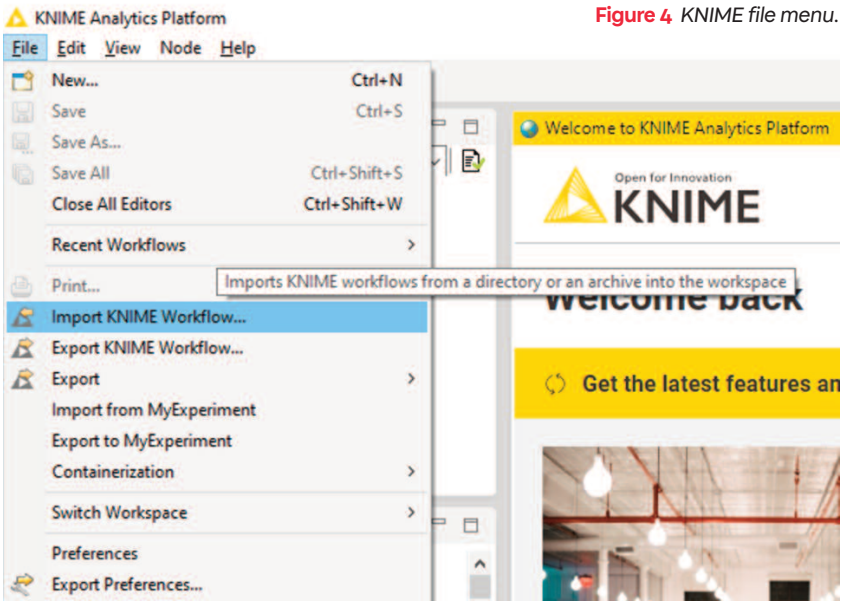


Figure 4 *KNIME file menu.*

A dialog box will open.

In “Source:”, click “Select File”, then press the “Browse” button to choose the workflow file WCAIR_01_Reaction_Enumeration.knwf (**Figure 5**).

Make sure that the workflow is selected in the “Import Elements” list by pressing the “Select All” button of the right.

Press “Finish” to import the workflow.

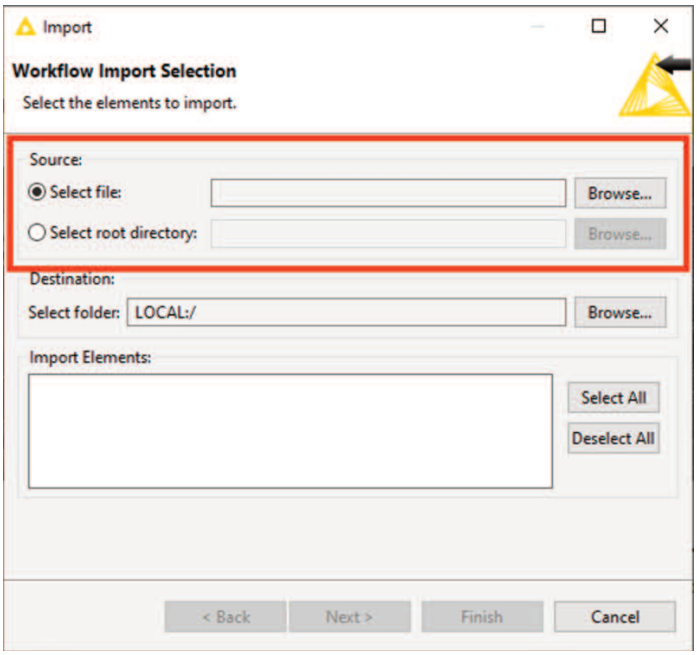


Figure 5 *Import window for KNIME*

A new file will appear on the **KNIME Explorer tab**.

If the workflow does not open by itself, double click on “WCAIR_01_Reaction_Enumeration” to open the workflow (Figure 6).

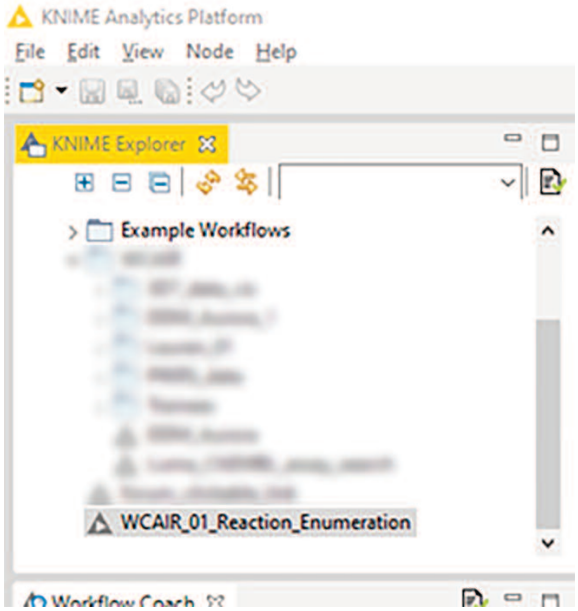


Figure 6 *KNIME explorer pane showing the imported workflow.*

If this is the first time you have used KNIME, KNIME will ask you to install required extensions (Figure 7).

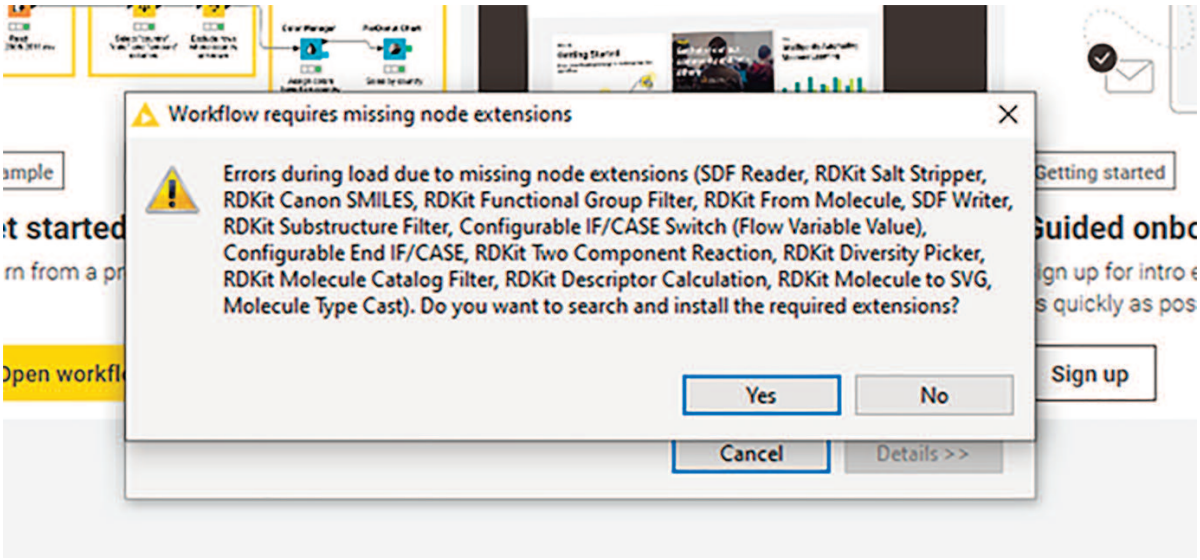


Figure 7 *Error message indicating missing nodes that need to be installed to open the workspace.*

Click “Yes”.

In the new dialogue box, click “Next”.

Ensure that "I accept the terms of the license agreements" is selected prior to clicking "Finish". The necessary nodes have now been installed.

After installing the nodes, KNIME will ask to be restarted. Accept the restart and open the workflow again.

When you load the workflow for the first time, a Warning message may appear indicated that the workflow has been loaded with warnings (**Figure 8**). This can be ignored. If this message appears just press OK and keep on.

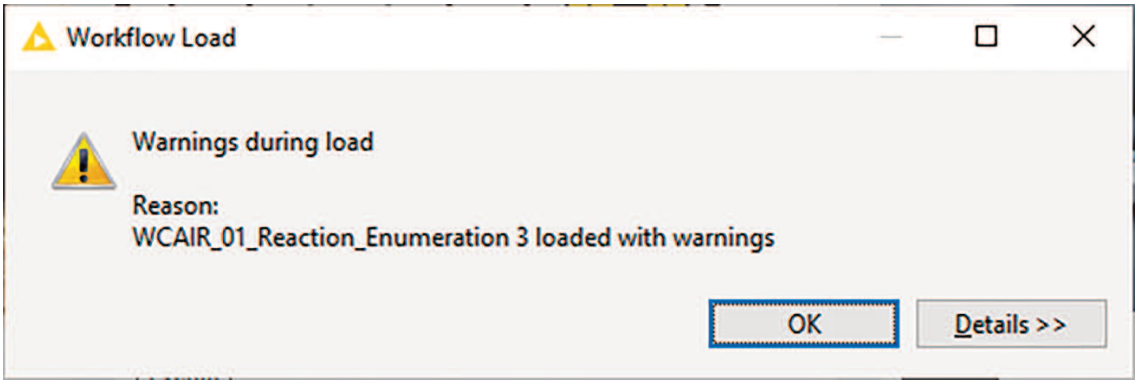


Figure 8 *Warnings during load message box. This message may not appear when loading the workflow.*

2.4 Adjusting renderer preferences

To view the chemical structures, we need to set the proper renderer for KNIME.

Click "File" > "Preferences"

A dialog box will open (**Figure 9**).

On the left pane, expand KNIME and look for "Preferred Renderers".

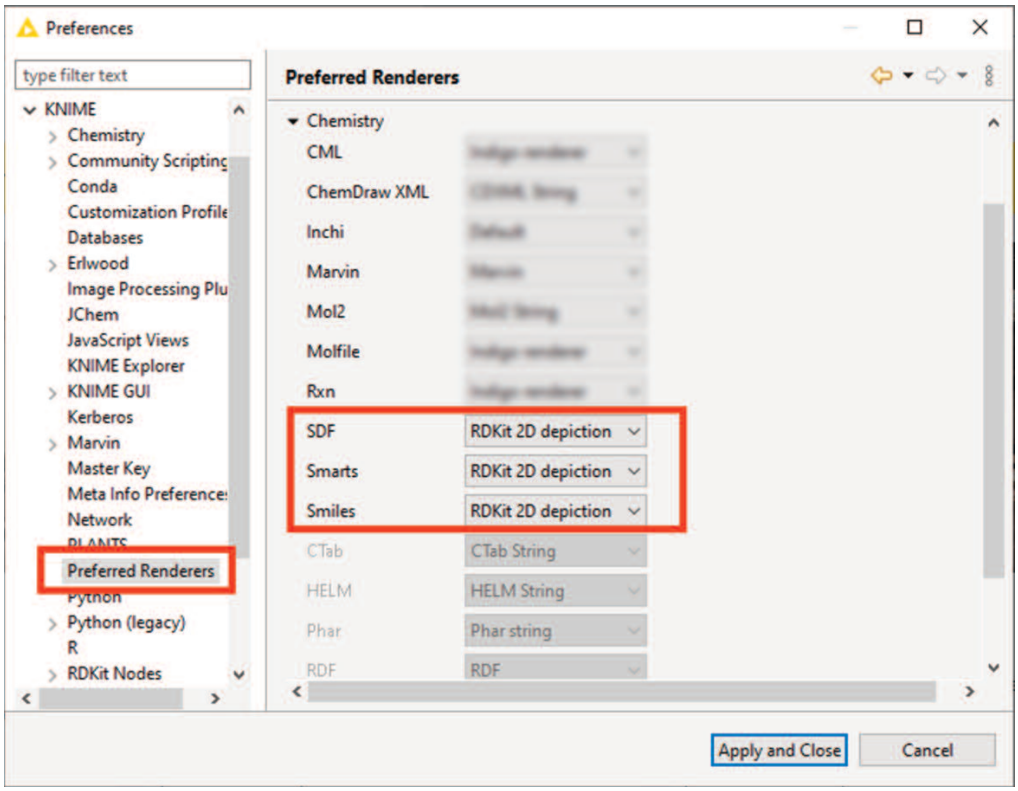


Figure 9 Knime Preferences dialog box with the 'Preferred Renderers' pane open.

Make sure that the renderer for "SDF", "SMARTS" and "SMILES" are the "RDKit 2D depiction"*

* RDKit was from one of the extensions you installed in the last step and this will tell KNIME to use the RDKit to recognize and render (i.e.: draw) the SDF, SMILES and SMARTS fields as chemical structures.

Press "Apply and Close" when done.

3.0 Nodes

With the workflow open, KNIME interface will show the file as a graphical workflow (**Figure 10**).

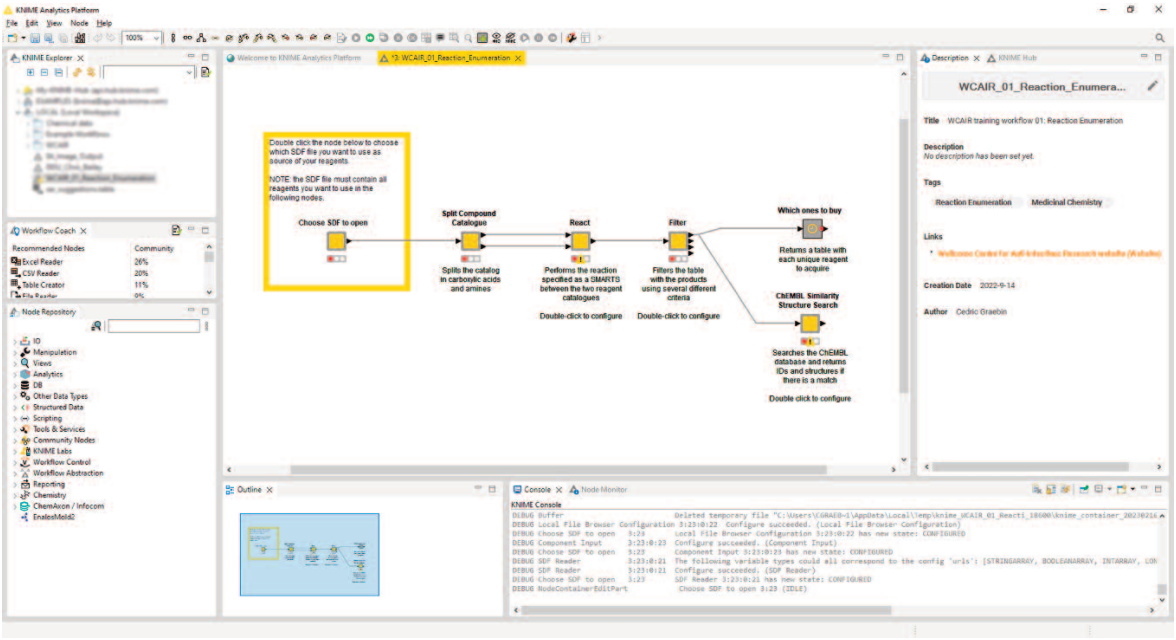


Figure 10 *KNIME with the workflow loaded.*

You'll see six squares connected by arrows.

Each square, connected by arrows, is called a "node" (**Figure 11**). Nodes indicate a data transformation. KNIME compartmentalises the transformation and allows the user to see what is happening after each node execution.

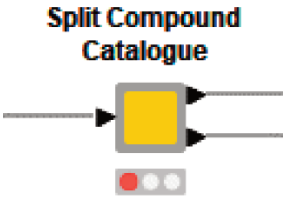


Figure 11 *A KNIME node.*

An arrow to the left of the node is the "input port", i.e: the data table that will be transformed by the node. One or more arrows to the right are the "output ports", or the data tables that are the result of the transformation done by the node.

The traffic lights at the bottom of each node denotes the node state.

- A **red** light means that the node has not been executed yet or is not ready to be run (requiring configuration)
- A **yellow** light means that the node is configured and ready to be executed
- A **green** light means that the node was successfully executed.

You will notice that all the six nodes are marked as red because they are not ready to be executed. You need to configure the nodes so they can be properly executed.

3.1 Processing the molecule catalogue file

Begin with 'Choose SDF to open' node. Double click it to open its configuration dialog (**Figure 12**).

Use the "Browse" button to look for the **"sigma-aldrich-zinc.sdf"** file provided with this exercise. Click "OK" to close this dialog box.

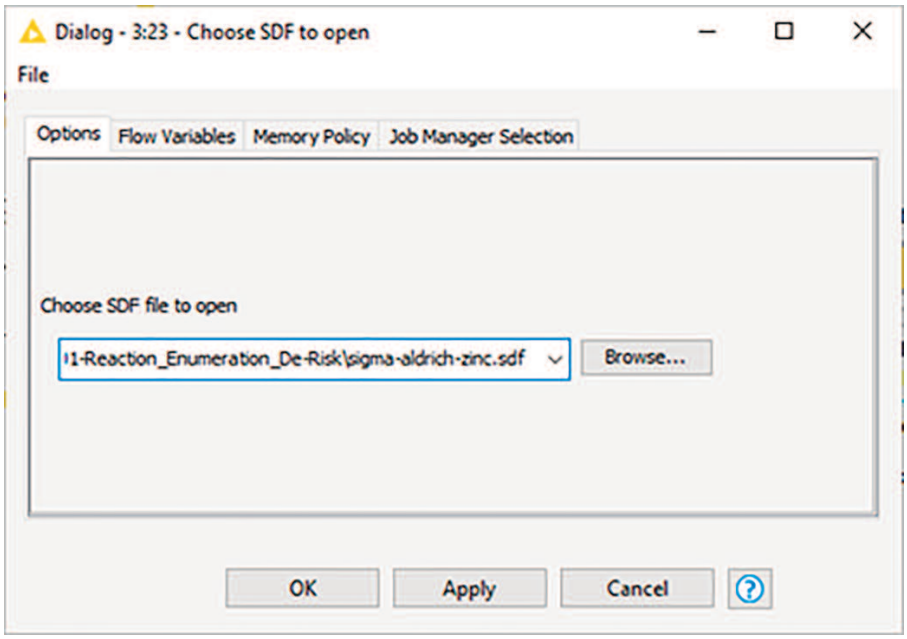


Figure 12 *First node configuration dialog box.*

Right-click the first node and choose the option "Execute" (**Figure 13**). Wait for the node execution to be complete (the traffic light below the node will turn green).

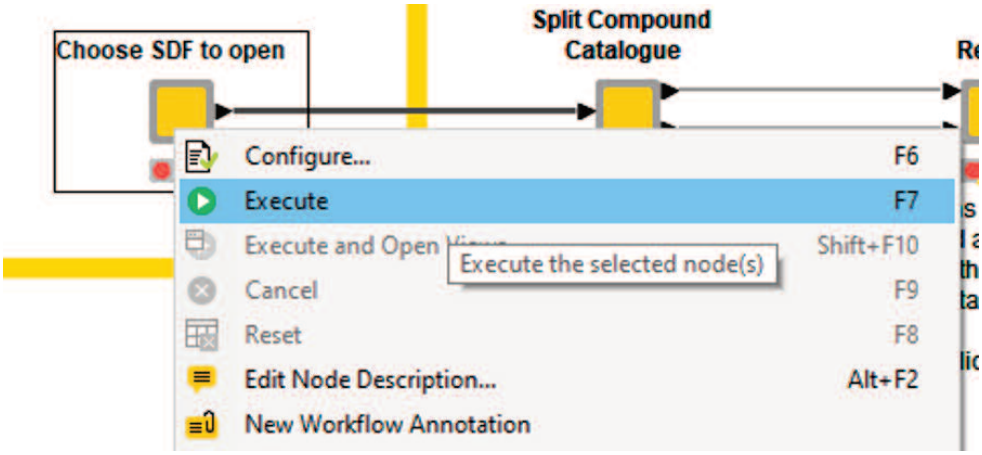


Figure 13 *Node context menu highlighting the 'Execute' option.*

3.1 Processing the molecule catalogue file

If the node now has a green light (**Figure 14**), this means that it was executed properly.

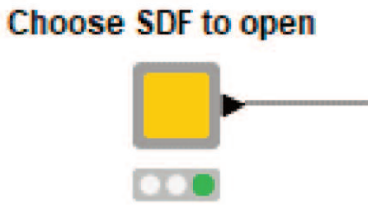


Figure 14 *Green light indicates the node executed properly*

Right click the node and choose the last option, named “Molecule catalog” (**Figure 15**).

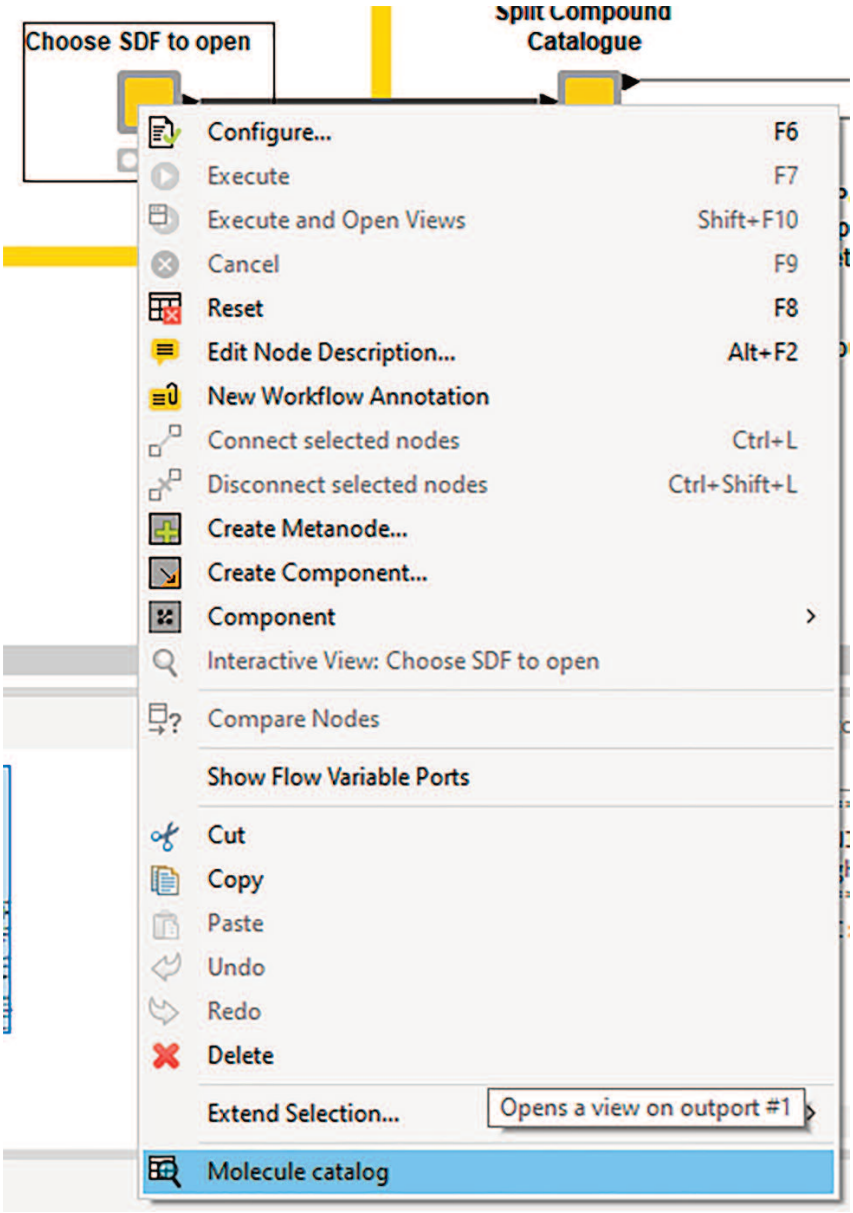


Figure 15 *Node context menu highlighting the 'Molecule catalog' option.*

A table will now appear. Study the table and look to see what has been included.
How many rows and columns does this table have?

Click "X" to close the window.

3.2 Split Compound Catalogue

The first table showed several chemical structures and their catalogue code. The aim of this exercise is to make amide couplings with these reagents. Will all these reagents react in an amide coupling? Probably not!

The second node, "Split Compound Catalogue", will extract all the primary amines and carboxylic acids from the catalogue and give them as two separate outputs.

Right click over the node and click on "Execute".

After the execution, explore the two output tables (right-click, click on the last two options: "Amines", and then "Carboxylic acids").
How many compounds of each functional group are present in the catalogue?

3.3 Perform reactions

The next node "React" will perform the reaction between the amines and carboxylic acids from the node "Split Compound Catalogue". It has two input ports – one for each compound table.

Double click the third node to configure its execution (**Figure 16**).

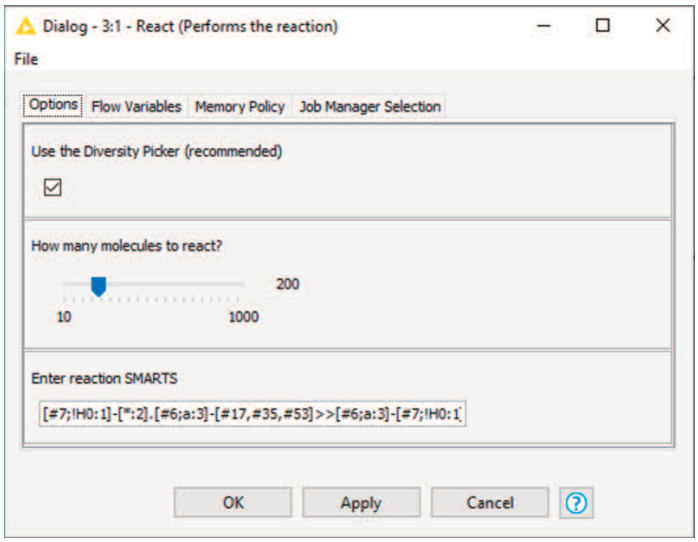


Figure 16 *Configuration window for the React node.*

It is strongly recommended to use the Diversity Picker. This picker will select a diverse subset of both catalogue tables. If this is run multiple times, the diversity picker will pick different compounds each time.

You can change how many molecules to generate using the "How many molecules to react?" slider. If you choose not to use the picker, the node will try to react all the carboxylic acids with all the amines, and this may take a long time to process. Ensure "200" is selected.

The reaction SMARTS is a text strings that describe an amide coupling using the SMARTS chemical language. This exercise will not cover SMARTS. Further information on this topic can be found **here**.

Press "OK", then right-click the node again and choose "Execute".

After the execution, right-click the node again and inspect the "Products" table (last option on the drop-down menu).
How many products are there? Does the table contain the reagents used for each coupling?

3.4 Filtering the results

The next step is to filter these products (unless you're willing to analyse 40,000 compounds) into something more manageable.

Double click the "Filter" node to open its configuration dialog box (**Figure 17**).

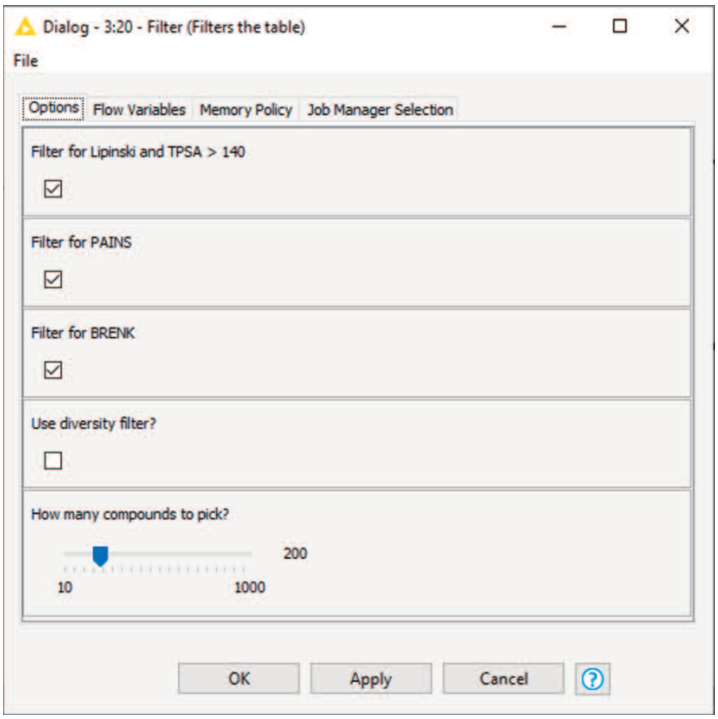


Figure 17 *Node context menu highlighting the 'Molecule catalog' option.*

For this exercise, keep the default options checked (i.e.: "Filter for Lipinski and TPSA > 140", "Filter for PAINS", "Filter for BRENK", "Use diversity filter?", and set the slider to 200 compounds to pick).

Press "OK". Right click the node > "Execute".

Wait for the node to complete its tasks (this may take a few minutes).

Explore the four outputs (**Figure 18**) using your mouse and right clicking the node.
How many molecules did not pass the filter criteria?

The tables with BRENK and PAINS filters also have a detailed explanation of what made the molecule to be classified as PAINS or as a compound presenting undesirable features.
Why were there only 200 "Filtered compounds"?

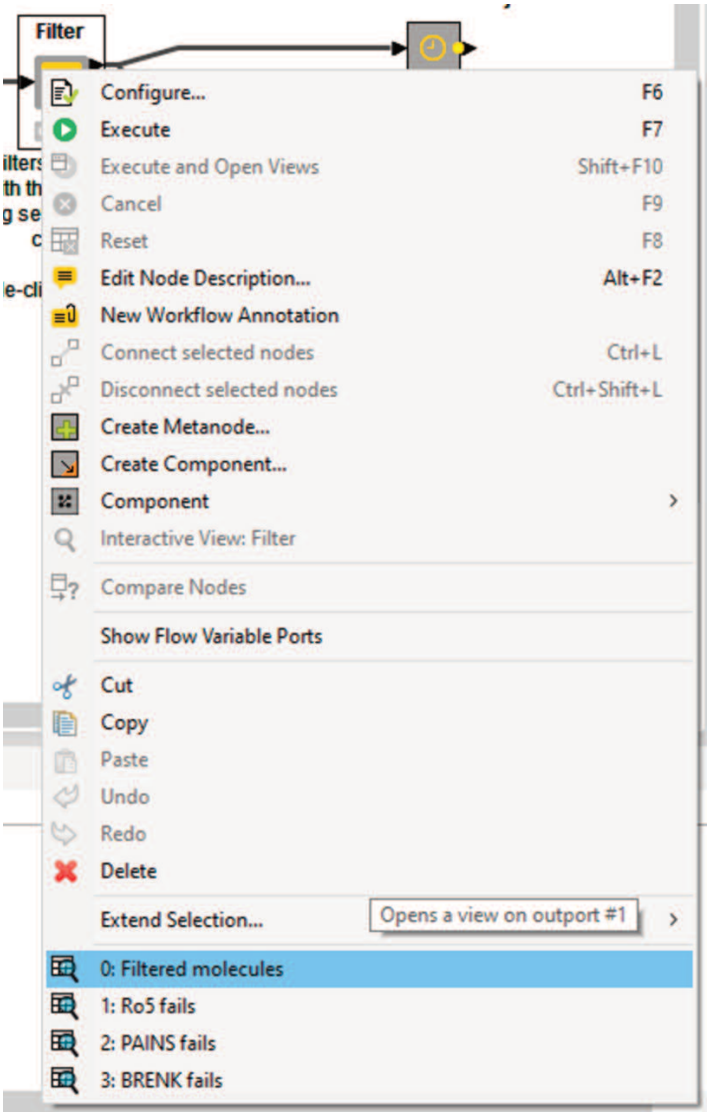


Figure 18 *Context menu for the Filter node.*

3.5 How many reagents shall I purchase?

Now that you have the filtered list of products, what are the reagents that are needed to be acquired? Maybe one or more of them are being used twice and it would be good to avoid double purchases.

To check this, use the "Which ones to buy" node. Right click on it > "Execute".

After the execution, the node symbol will change to a green checkmark. Right click on it again and choose the last option "Connected to: Filtered/Labelled Data" to see the suggested reagents for procurement.
How many reagents are there?

3.6 What if my compounds were already synthesized and tested?

KNIME has many good features. One of them is that you can utilise external data services, make queries, and fetch the results. We will do this by making use of the freely available chemical database ChEMBL (from the European Molecular Bioinformatics Laboratory) similarity search service.

Double click on the "ChEMBL Similarity Structure Search" node to show its configuration dialog box (**Figure 19**).

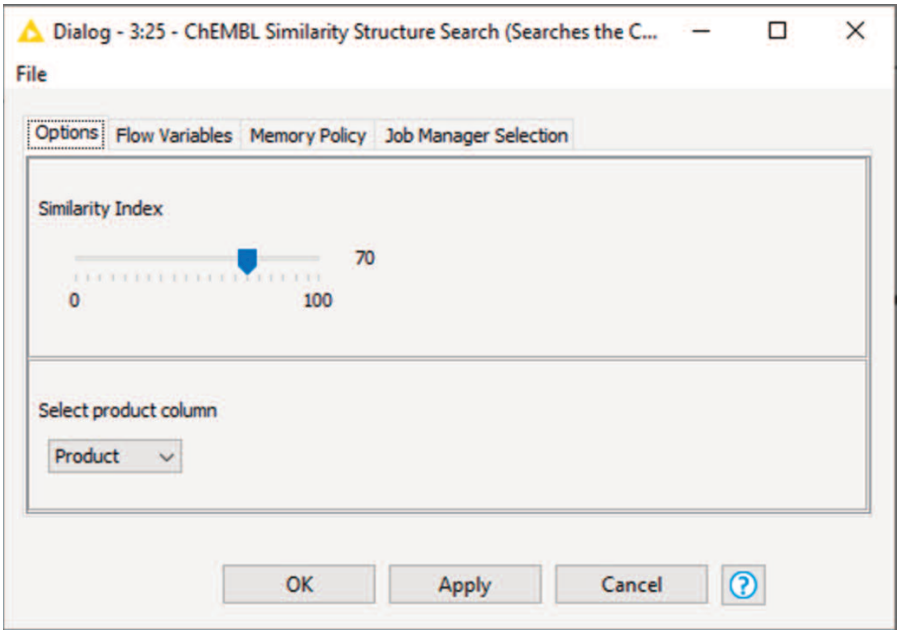


Figure 19 *ChEMBL similarity search configuration window*

The similarity index tells ChEMBL how similar the molecule the user is looking for must be from the presented results. Normally 95% or greater similarity returns only the chosen molecule (if it exists), and if you want to see close analogues a value between 60-70% will do the trick (although it will also increase the processing time in the requests).

Choose **60** for this exercise.

This node requires a working internet connection to execute this node.

Press "OK" to close the dialog box.

Right click > 'Execute and Open Views'.

This is also one of the great features of KNIME. It allows the user to create interactive views, in which the data can be manipulated and transformed in real time. We will make use of this in this view.

After processing, a new window (**Figure 20**) with a table will appear on the screen (if there are any results to report, of course).

Depending on how the diversity filter was set up, the compounds you will get might not be the same as the ones in the Figure below.

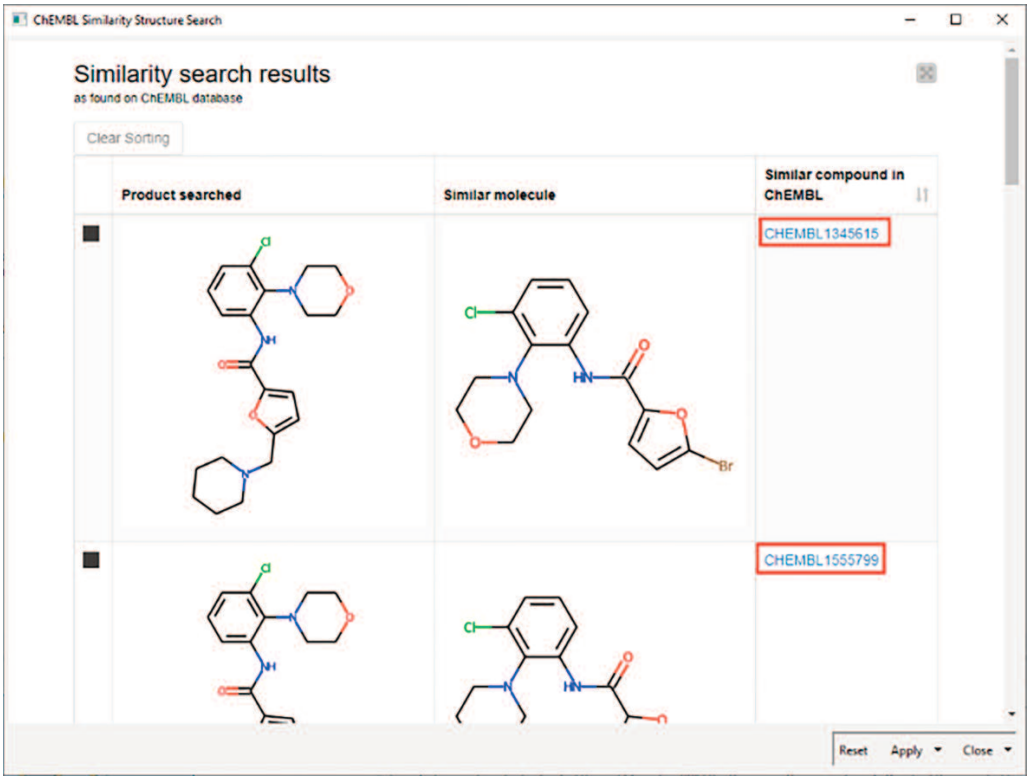


Figure 20 *"Similarity search results" interactive view.*

The interactive view will show you the product you searched, the similar molecules that were found on the ChEMBL database (one per row), and its ChEMBL ID number.

If you click this CHEMBL identifier on the right column, a new window will open redirecting you to the ChEMBL website (see **Figure 21**) with all the details for this molecule.

There may be cases in which the "Product searched" will be repeated, because more than one similar product was found in the database.

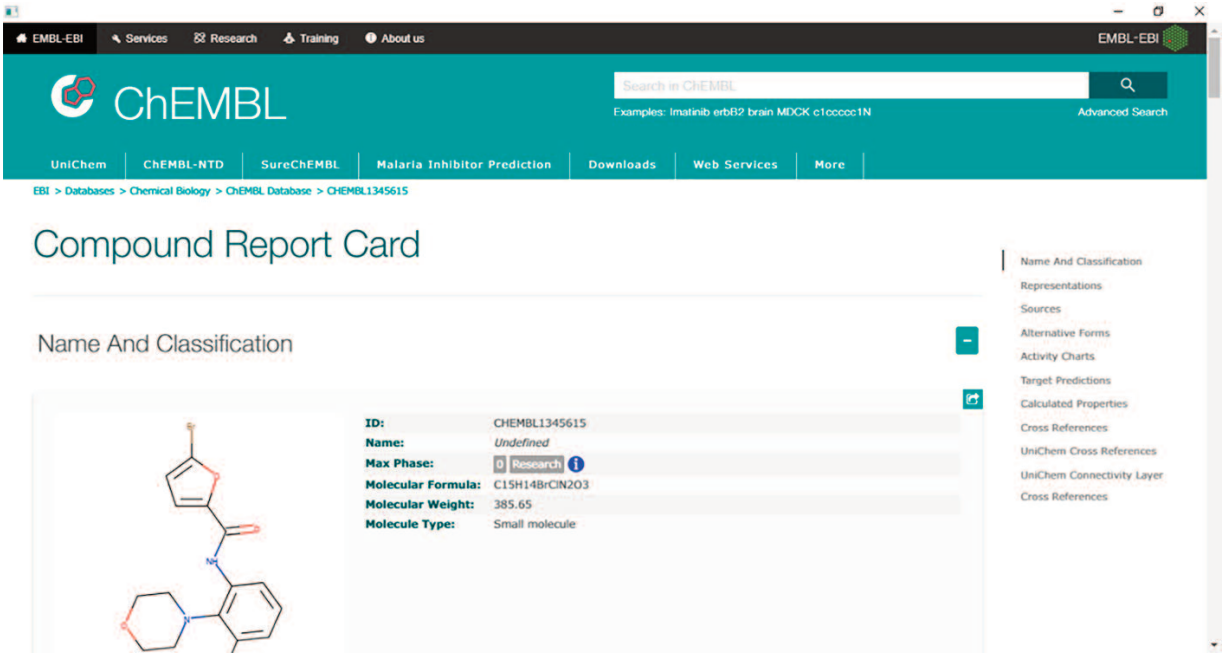


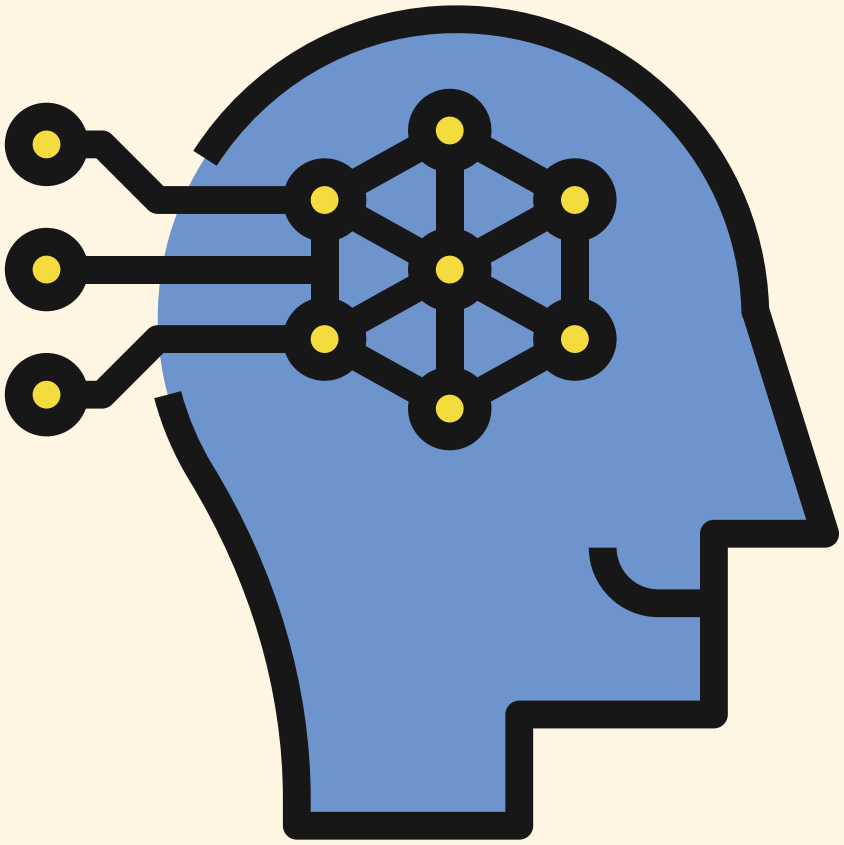
Figure 21 *ChEMBL compound report card window.*

How many similar products have you found in the database search?

4.0 And now...

As we finally approach the end of this exercise, you now have given your first steps on how to use KNIME and its chemistry-related features. You were able to explore a catalogue of chemical products, extract from them two tables containing two different families of organic compounds, perform a reaction using them and analyse the results, including searching for them in an open-access chemical compound database. And this is just the tip of the iceberg! KNIME can do a lot more for chemistry and life sciences-based applications. You can read more on the **KNIME website**.

5.0 Answers for in-text questions



5.0 Answers for in-text questions

5.1 Processing the molecule catalogue file

How many rows and columns does this table have?

28355 rows and 2 columns.

5.2 Split Compound Catalogue

How many compounds of each functional group are present in the catalogue?

4366 amines and 2819 carboxylic acids.

5.3 Perform reactions

How many products are there? Does the table contain the reagents used for each coupling?

40,000 if the diversity picker was set to 200.

5.4 Filtering the results

How many molecules did not pass the filter criteria?

In our test runs, 3,470 molecules failed the Rule-of-five filter, 713 failed the PAINS filter, and 16,685 failed the BRENK filter.

Why were there only 200 "Filtered compounds"?

This was the number of compounds to pick in the filter node (see Figure 17).

5.5 How many reagents shall I purchase?

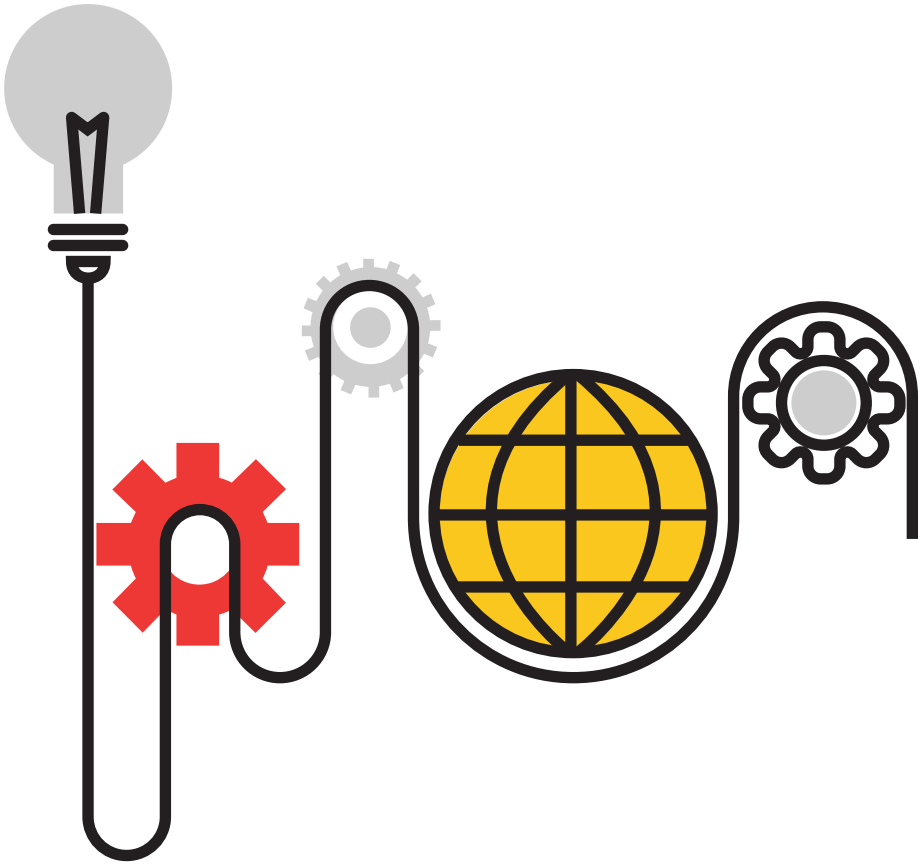
How many reagents are there?

This depends largely on how many compounds will be filtered off from the results (see last question). In our test runs, there were 290-310 compounds on the table. This number may vary if more compounds are chosen in the Filter node (again, see last question).

5.6 What if my compounds were already synthesized and tested?

How many similar products have you found in the database search?

This may also vary, since it depends on how many compounds were picked in the Filter node and the similarity search value. In our test runs these numbers were varying from 30 – 60 when using a 60% as similarity value cut-off.



Wellcome Centre for Anti-Infectives Research
School of Life Sciences
University of Dundee
Dow Street
Dundee DD1 5EH

WCAIR@dundee.ac.uk

www.wcair.dundee.ac.uk

@WCAIRDundee

University of Dundee

wellcome centre anti-infectives research