NovaMechanics | Chemoinformatics & Bioinformatics Solutions

KNIME Enalos+ nodes

A Brief Tutorial

Novamechanics Ltd
Contact: info@novamechanics.com
Version 1, June 2017

Table of Contents

Introduction

Real world research seldom involves a single question being answered by means of a single operation, and the fields of molecular modelling, chemoinformatics and nanoinformatics are no exception. While researchers can create custom scripts to automate common procedures, this solution is less than ideal when projects demand rapid workflow prototyping, interactive data analysis, and robust, appropriately validated models. These are exactly the conditions for which Novamechanics Ltd Enalos+ nodes are best suited to open-source KNIME interface. Enalos+ nodes are designed to perform molecular modelling and help the user get straight access to multiple Chemical Databases for data mining and manipulation.

Enalos+ nodes built upon the existing KNIME infrastructure are divided in five main categories (Modelling, Molecular Descriptors, NCI, PubChem and UniChem) and significantly increase the number of the available nodes, the data handling tools and bridge different chemoinformatics and modelling tools upon the same interface.

The current tutorial is designed to help the user in going step-by-step through the process of building a KNIME workflow, using Enalos+ nodes of Novamechanics Ltd. This case study deals with the chemical compound abacavir, which is a medication used to prevent and treat HIV/AIDS.

Step 1-Workbench overview

The KNIME workbench is organized as follows:

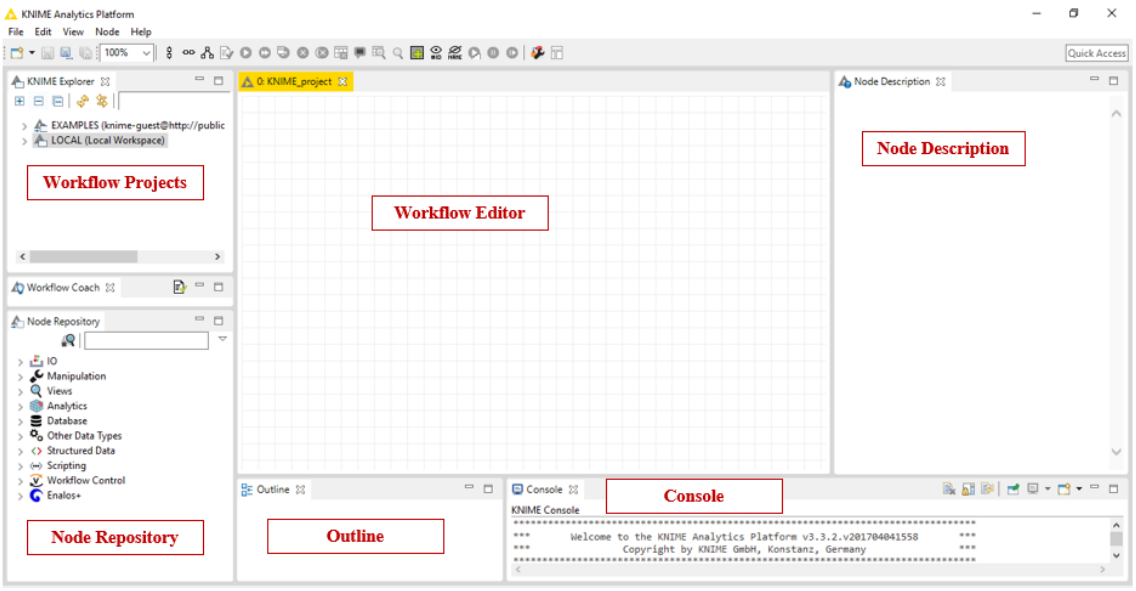


Fig. 1: KNIME workbench

It is composed of 6 main "windows": The Workflow Projects, the Workflow Editor, the Node Description, the Node Repository, the Outline and the Console. A short description of the KNIME's interface windows follows in Table 1:

Table 1: Description of KNIME interface

Workflow Projects	Workflow Editor	Node Description
Each workflow refers to a workflow project. All projects are displayed here. Import and export of workflows is supported. Status (closed, idle, executing and executed) is indicated by an icon.	Here the workflows are assembled by dragging nodes onto this editor, connecting, configuring and executing them.	Provides help about the selected node, its dialog options, views, expected input data and resulting output.
Node Repository	**Outline**	**Console**
Find all KNIME nodes here, ordered by categories. Help for selected nodes is displayed in the Node Description. Drag them onto the editor in order to add them to the workflow.	Overview over the workflow and navigation help for large workflows.	Status information, warnings and error messages are logged here. This information is also written to a log file.

Step 2-Building a workflow

The nodes are the basic processing units of a KNIME workflow. A workflow is built by dragging nodes from the Node Repository onto the Workflow Editor and connecting them, creating pipelines: Each node has a number of input-and/or output ports. Data (or a model according to each particular case) is transferred over a connection from an out-port to the in-port of another node.

1. Node status

When a node is dragged onto the workflow editor the status light shows red, which means that the node has to be configured in order to be able to be executed. A node is configured by right clicking it, choosing "Configure", and adjusting the necessary settings in the node's dialog. When the dialog is closed by pressing the "OK" button, the node is configured and the status light changes to yellow: the node is ready to be executed. Right-click on the node again shows an enabled "Execute" option; pressing it will execute the node and the result of this node will be available at the out-port (Fig. 2). After a successful execution the status light of the node is green. The result(s) can be inspected by exploring the out-port view(s): the last entries in the context menu open them. The above options "Configure", "Execute" and "View" are also available in the top ribbon of the KNIME interface window.

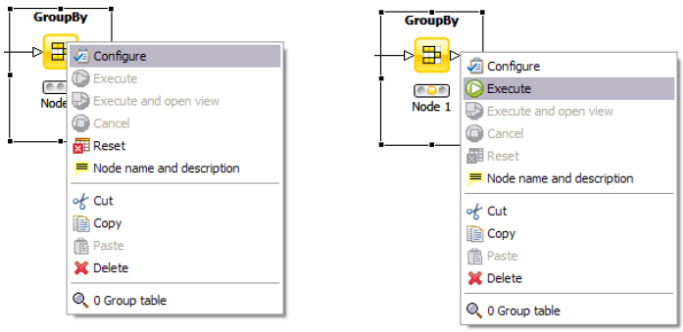


Fig. 2: Configuring and executing nodes

2. Ports

Ports on the left are input ports, where the data from the out-port of the predecessor node are provided. Ports on the right side of the node are called out-ports. The result of the node's operation on the data is provided at the out-port to successor nodes.

Step 3-Activate the Enalos+ nodes

In order to activate the Enalos+ nodes, the user has to copy the .jar file in the plugins folder and the .lic file in the license folder in the KNIME file location.

Step 4-A Brief Introduction

The Enalos+ nodes are divided into 5 main categories: Modelling, Molecular Descriptors, NCI, PubChem and UniChem.

1. Modelling

Modelling contains 11 nodes specified for data handling, preprocessing, testing modeling robustness and testing the accuracy of the predictions:

Table 2: Modelling nodes

Create New Molecules	**Domain APD**	**Domain Leverage**
Create New Molecules enables the user to create a list of molecules by combining a series of substituents with a core molecule.	*Domain APD* enables the user to define the domain of applicability of the model using a method based on the Euclidean distances.	*Domain Leverage* enables the user to define the domain of applicability of the model using a method based on the extent of extrapolation.
Int 2 Double	**Kennard and Stone**	**MLR**
Int 2 Double converts integer values of all columns to doubles.	*Kennard-Stone* node allows the selection of two representative subsets (as training and test sets) with a uniform distribution over an initial dataset.	*MLR* node performs Multiple Linear Regression in order to model the relationships between a scalar dependent variable y and two or more independent variables denoted as X.
Model Acceptability Criteria	**Remove Column**	**Remove Duplicates**
Model Acceptability Criteria gives information about the Quality of Fit and Predictive Ability of a continuous QSAR Model.	*Remove Column* node removes the selected input columns of the table that contain the same values at a percentage equal or higher than a specified cutoff limit.	*Remove Duplicates* enables the user to remove the rows of the input table that contain the same values in selected columns. The filtered table contains all rows that are unique and the first one of each repeated row.
Sphere Exclusion	**Y Randomization**	
Sphere Exclusion node allows the selection of two representative subsets (such as training and test sets). This method attempts to specify compounds which most effectively cover the available data space.	*Y Randomization* (or Y-scrambling) is a technique, applied to ensure a QSAR model's robustness.	

2. Molecular Descriptors

Molecular Descriptors contains *EnalosMold2* node.

3. NCI

NCI contains *CIR* node.

Table 3: Molecular Descriptors and NCI

EnalosMold2	**CIR**
Molecular Descriptors by *EnalosMold2* calculates a large and diverse set of molecular descriptors (777) encoding two-dimensional chemical structure information.	Enalos+ *CIR* node enables the user to get direct access to CIR (Chemical Identifier Resolver) through KNIME. The user has the option to select several output formats through a GUI menu.

4. PubChem

PubChem contains 8 nodes that give direct access to PubChem database through KNIME in order to extract useful information:

Table 4: PubChem nodes

Assay	**Assay Class**
Assay node gives the user access to PubChem database via substance or compound IDs (SID and CID), in order to find the Assays where a particular compound is tested. Using this node the user can download in KNIME information about the Assay and the Assay outcome.	*Assay Class* node searches in PubChem database according to one or more given AIDs (BioAssay identification numbers) and displays only the active or inactive compounds.
Main PubChem	**Patent**
Main PubChem node enables the user to search the PubChem database and obtain the following information for thousands of compounds with one request: PubChem CID (Compound ID), IUPAC Name, InChI, InChI-Key Molecular Formula, Molecular Weight, Canonical SMILES and the direct PubChem URL.	*Patent* node gives the user straight access to the PubChem database in order to obtain information about the patent coverage information for thousands of compounds with one request.
Patent to Sid	**Sid**
Patent to Sid node helps the user to search the PubChem database and obtain the SIDs (Substance IDs) of the compounds covered by the patents in request.	*Sid* node exports the CIDs (Compound IDs) of a given list of SIDs (Substance IDs), searching the PubChem database. The user can search the PubChem database and obtain information about the CIDs for thousands of compounds with one request.
Similarity	**Vendor**
Via *Similarity* node, the user can search the whole PubChem database for similar compounds (Tanimoto Similarity) and obtain the following information for thousands of compounds with one request: PubChem CID (Compound ID), Molecular Formula, Molecular Weight and Number of Rotatable Bonds.	*Vendor* node enables the user to search the PubChem database and obtain information about the commercial availability for thousands of compounds with one request.

5. UniChem

UniChem contains 2 nodes for accessing UniChem databases:

Table 5: UniChem nodes

UniChem	UniChem Connectivity
Enalos *UniChem* gives the user direct access to UniChem databases through KNIME. UniChem is a superset of all 27 available databases, separated in 5 friendly and easily recognizable categories.	*UniChem Connectivity* is an expanded version of the standard UniChem tool that allows you to find related molecules. Connectivity Search allows molecules to be first matched on the basis of complete identity between the connectivity layer of their corresponding Standard InChIs, and the remaining layers then compared to highlight stereo-chemical and isotopic differences.

Step 5-Adding Nodes

In the Node Repository, expand the *IO* and the contained *Other* category and choose *Table Creator* node (Fig. 3). Then, drag & drop the *Table Creator* icon into the Workflow Editor window.

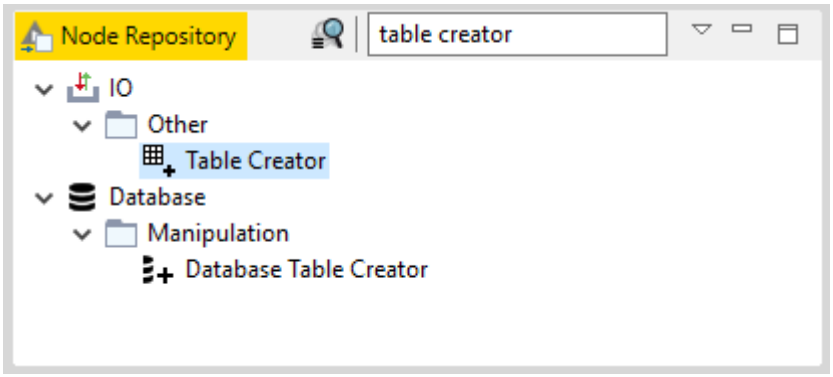


Fig. 3: Node Depository interface

Subsequently, expand the *Enalos+* category followed by the *PubChem* category and drag into the Workflow Editor *Main PubChem* and *Similarity*. The final nodes are *Assay*, *Patent* and *Vendor* and are positioned to the right of the *Similarity* node as shown below (Fig. 4).

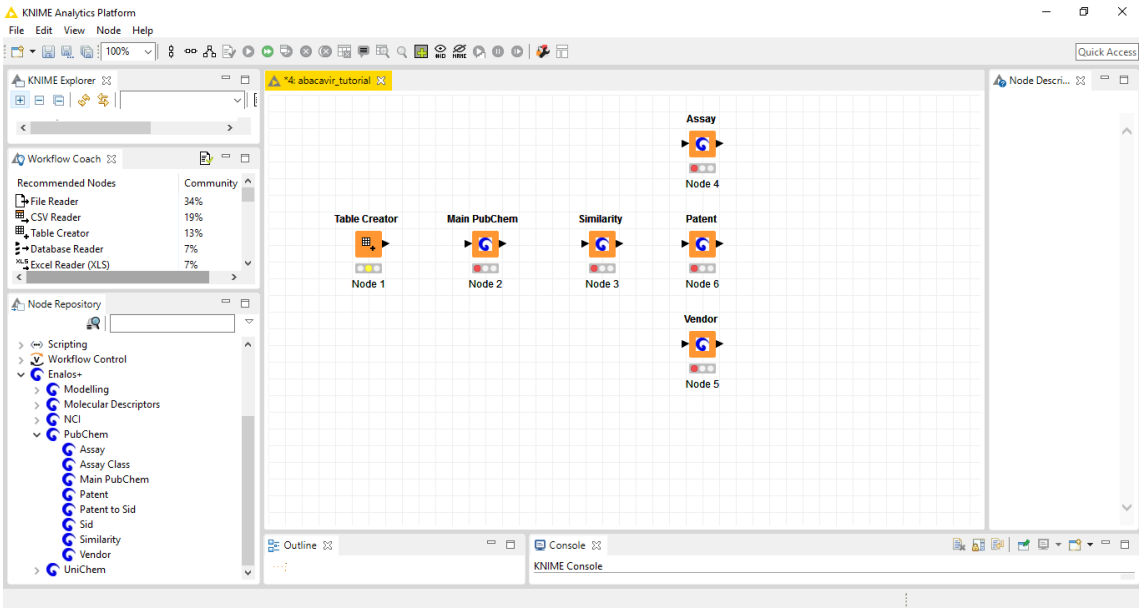


Fig. 4: Workflow editor

Step 6-Connecting Nodes

Now, you need to connect the nodes, in order to get the data flowing. Click an output port and drag the connection to an appropriate input port. Complete the flow as pictured below (Fig. 5). The nodes will not show a green status as long as they are not yet configured and executed.

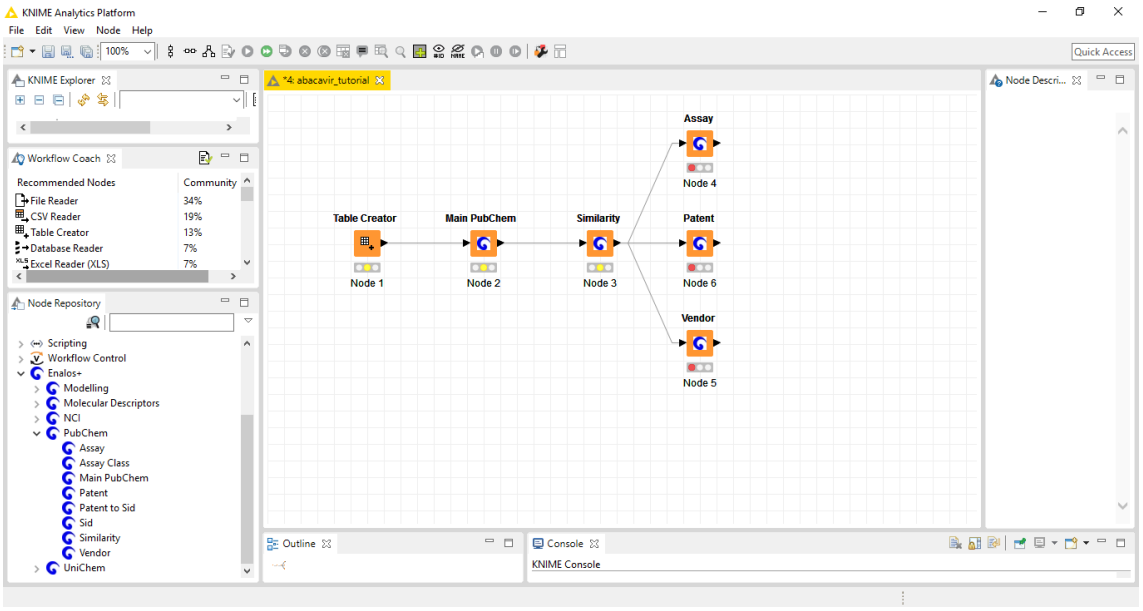


Fig. 5: Connecting nodes

Step 7-Configuring nodes

Fully connected nodes showing a red status icon need to be configured. Start with the *Table Creator*, right click it and select “Configure” from the menu. You can rename “Column1” by “Name” or etc. You can also right click a random row and select “Row ID properties”. In this menu you can set the prefix and the suffix of row IDs, as well as the number you want to start counting (usually: 0 (default) or 1). Then type in cell (1,1) the name of the chemical compound “abacavir”. Press “Apply” and “OK” to close the dialog of the *Table Creator* node (Fig. 6). Once the node has been configured correctly, it switches to yellow (meaning ready for execution).

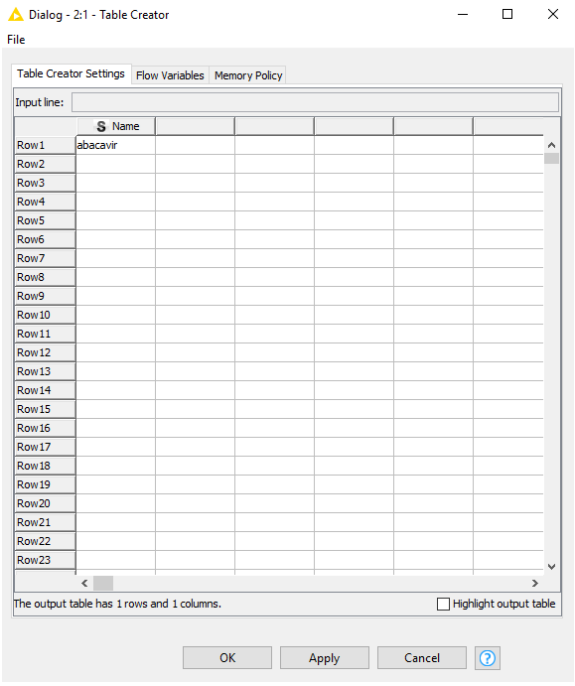


Fig. 6: Configuring Table Creator node

Then, configure *PubChem* node by choosing "Name" in the "Structure Identifier" and the "Input type" menus (Fig. 7). In this case, you have to choose "Name" because you have already given the name of the chemical compound (abacavir), while configuring the *Table Creator* node. Press "Apply" and "OK" to exit the dialog of the *Main PubChem*.

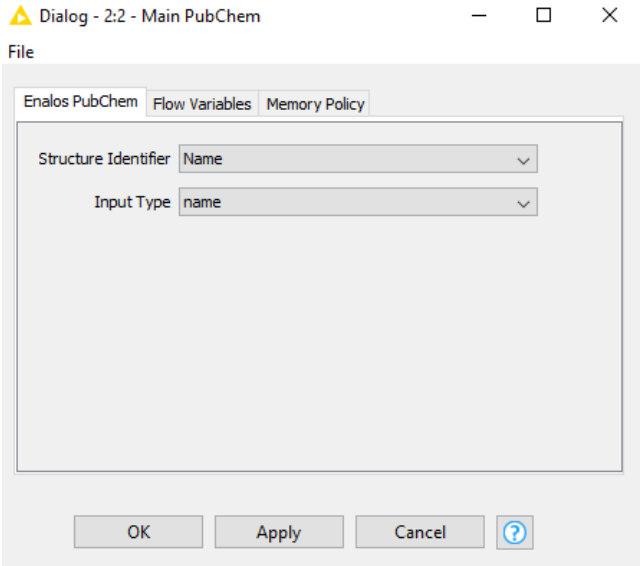


Fig. 7: Configuring Main PubChem node

Subsequently, configure *Similarity* node. Choose a "Structure Identifier", an "Input Type" (CID for example) and a "Tanimoto Similarity" value as depicted in Fig. 8.

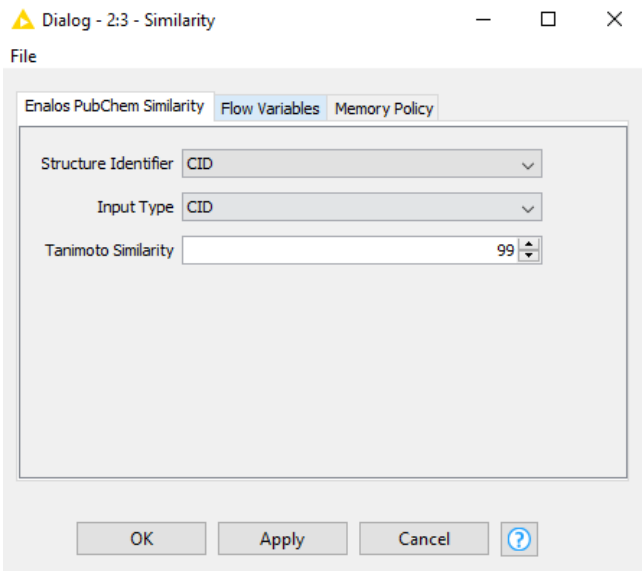


Fig. 8: Configuring Similarity node

Assay, *Patent* and *Vendor* nodes are configured by giving in the dialog window a "Structure Identifier" and an "Input Type". For more consistent results, when using PubChem Data Base, we usually choose CID for "Structure Identifier" and "Input Type" (Fig. 9).

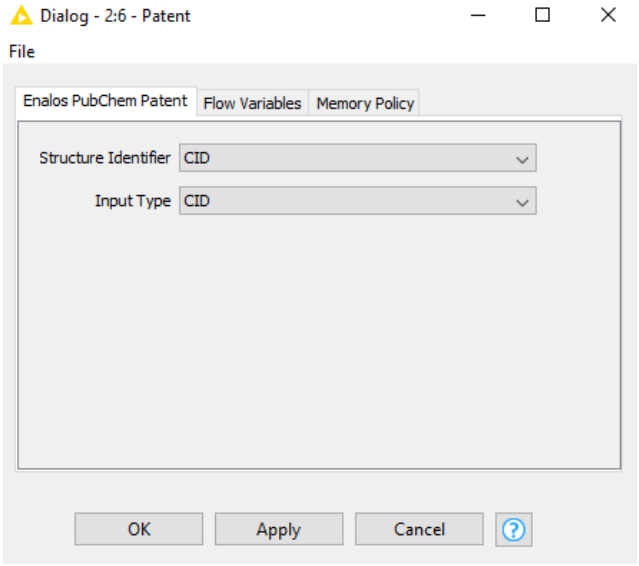


Fig. 9: Configuring Patent node

Step 8-Executing nodes

Now, right click on the *Assay* node and execute it. The workbench will execute all predecessor nodes for you. Then execute the *Patent* and the *Vendor* nodes. In a larger, more complex flow, you could select multiple nodes and trigger execution for all of them. The workflow manager will execute the nodes as needed, if possible in parallel. To execute all executable nodes press (Shift+F7).

Step 9-Inspecting the results

In order to examine the data and the results, open the nodes' views. From *PubChem* output port information is extracted about PubChem CID, IUPAC Name, InChI, InChI-Key Molecular Formula, Molecular Weight, Canonical SMILES and PubChem URL (Fig. 10).

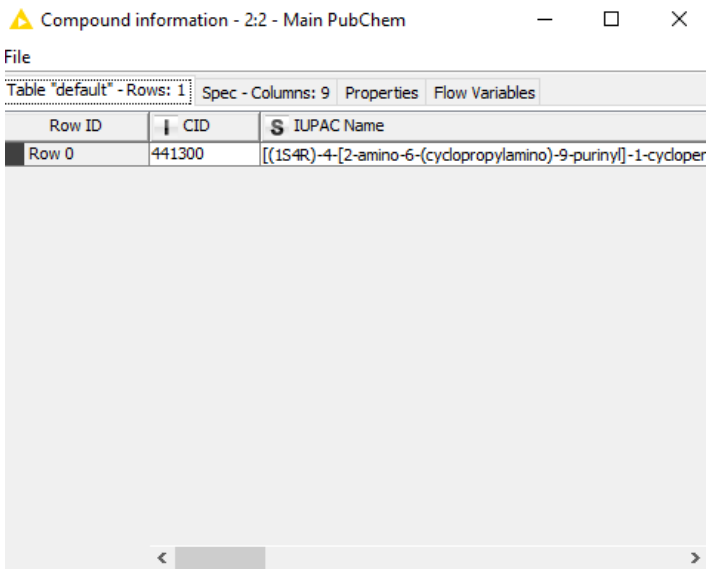
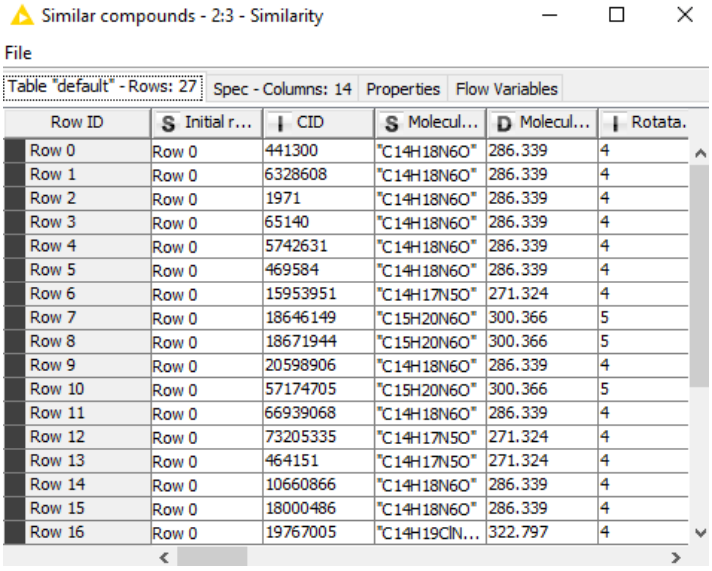


Fig. 10: Main PubChem results

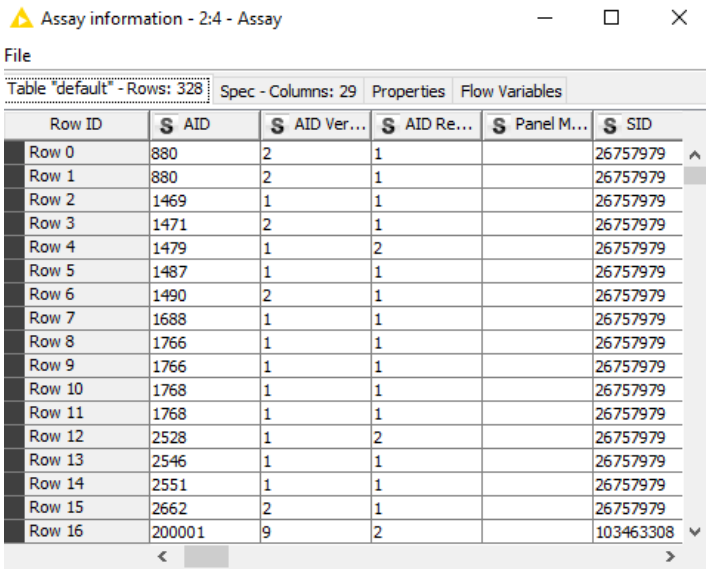
Similarity node extracts the CIDs of the similar compounds, along with their molecular formula, their molecular weight and the number of rotatable bonds (Fig. 11).



Row ID	Initial r...	CID	Molecul...	Molecul...	Rotata.
Row 0	Row 0	441300	"C14H18N6O"	286.339	4
Row 1	Row 0	6328608	"C14H18N6O"	286.339	4
Row 2	Row 0	1971	"C14H18N6O"	286.339	4
Row 3	Row 0	65140	"C14H18N6O"	286.339	4
Row 4	Row 0	5742631	"C14H18N6O"	286.339	4
Row 5	Row 0	469584	"C14H18N6O"	286.339	4
Row 6	Row 0	15953951	"C14H17N5O"	271.324	4
Row 7	Row 0	18646149	"C15H20N6O"	300.366	5
Row 8	Row 0	18671944	"C15H20N6O"	300.366	5
Row 9	Row 0	20598906	"C14H18N6O"	286.339	4
Row 10	Row 0	57174705	"C15H20N6O"	300.366	5
Row 11	Row 0	66939068	"C14H18N6O"	286.339	4
Row 12	Row 0	73205335	"C14H17N5O"	271.324	4
Row 13	Row 0	464151	"C14H17N5O"	271.324	4
Row 14	Row 0	10660866	"C14H18N6O"	286.339	4
Row 15	Row 0	18000486	"C14H18N6O"	286.339	4
Row 16	Row 0	19767005	"C14H19ClN...	322.797	4

Fig. 11: Similarity results

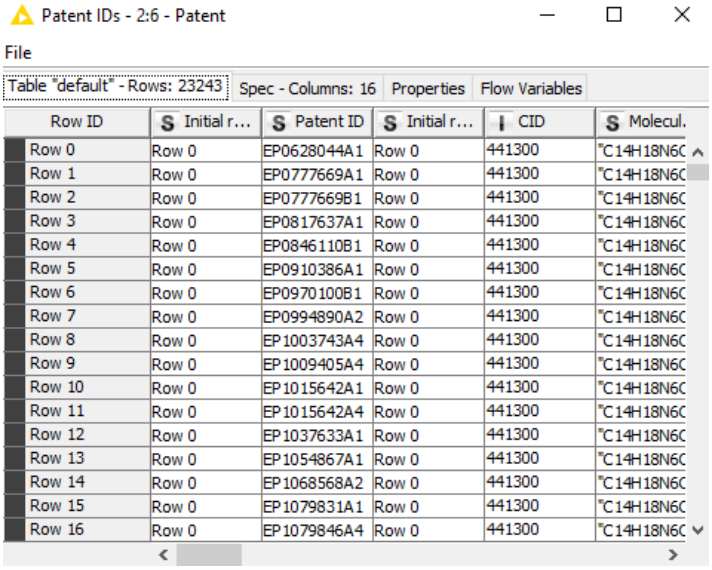
Assay node exports information from PubChem about the Assay name, AID (BioAssay identification number), Bioassay Type, Bioactivity Outcome, Target IG, Target GeneID, Activity Value in μm (whenever available) and the Activity Name (Fig. 12).



Row ID	AID	AID Ver...	AID Re...	Panel M...	SID
Row 0	880	2	1		26757979
Row 1	880	2	1		26757979
Row 2	1469	1	1		26757979
Row 3	1471	2	1		26757979
Row 4	1479	1	2		26757979
Row 5	1487	1	1		26757979
Row 6	1490	2	1		26757979
Row 7	1688	1	1		26757979
Row 8	1766	1	1		26757979
Row 9	1766	1	1		26757979
Row 10	1768	1	1		26757979
Row 11	1768	1	1		26757979
Row 12	2528	1	2		26757979
Row 13	2546	1	1		26757979
Row 14	2551	1	1		26757979
Row 15	2662	2	1		26757979
Row 16	200001	9	2		103463308

Fig. 12: Assay results

Patent node exports the Patent IDs, where every single compound is found. For the compounds that are not found in any patent of the PubChem, the node will export an appropriate message (Fig. 13).



Row ID	Initial r...	Patent ID	Initial r...	CID	Molecul.
Row 0	Row 0	EP0628044A1	Row 0	441300	"C14H18N6C
Row 1	Row 0	EP0777669A1	Row 0	441300	"C14H18N6C
Row 2	Row 0	EP0777669B1	Row 0	441300	"C14H18N6C
Row 3	Row 0	EP0817637A1	Row 0	441300	"C14H18N6C
Row 4	Row 0	EP0846110B1	Row 0	441300	"C14H18N6C
Row 5	Row 0	EP0910386A1	Row 0	441300	"C14H18N6C
Row 6	Row 0	EP0970100B1	Row 0	441300	"C14H18N6C
Row 7	Row 0	EP0994890A2	Row 0	441300	"C14H18N6C
Row 8	Row 0	EP1003743A4	Row 0	441300	"C14H18N6C
Row 9	Row 0	EP1009405A4	Row 0	441300	"C14H18N6C
Row 10	Row 0	EP1015642A1	Row 0	441300	"C14H18N6C
Row 11	Row 0	EP1015642A4	Row 0	441300	"C14H18N6C
Row 12	Row 0	EP1037633A1	Row 0	441300	"C14H18N6C
Row 13	Row 0	EP1054867A1	Row 0	441300	"C14H18N6C
Row 14	Row 0	EP1068568A2	Row 0	441300	"C14H18N6C
Row 15	Row 0	EP1079831A1	Row 0	441300	"C14H18N6C
Row 16	Row 0	EP1079846A4	Row 0	441300	"C14H18N6C

Fig. 13: Patent results

Vendor node extracts information from PubChem about: PubChem SID, Compound URL, Supplier and Category of the Vendor (Fig. 14).

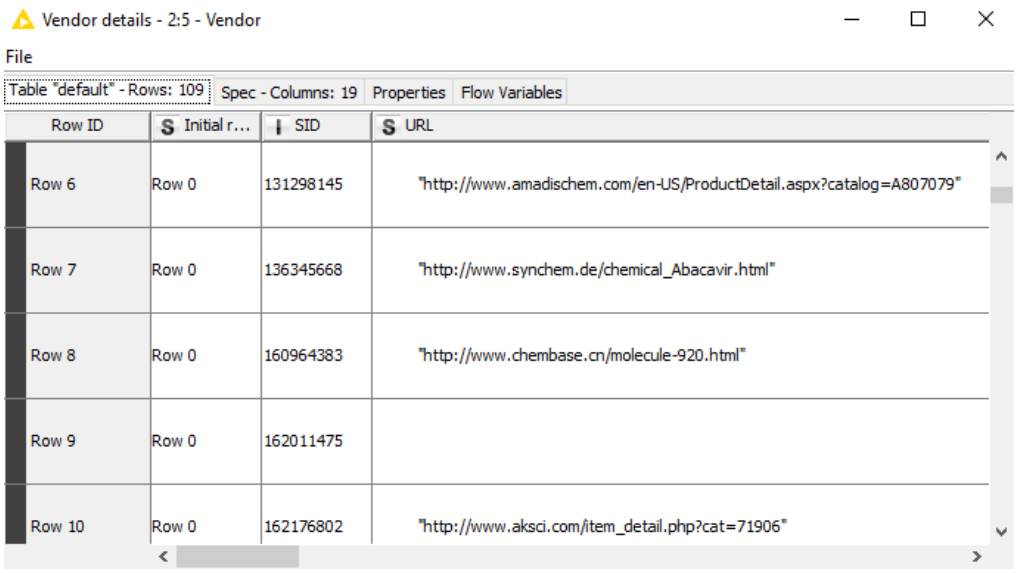


Fig. 14: Vendor results

The executed workflow is depicted in Fig. 15:

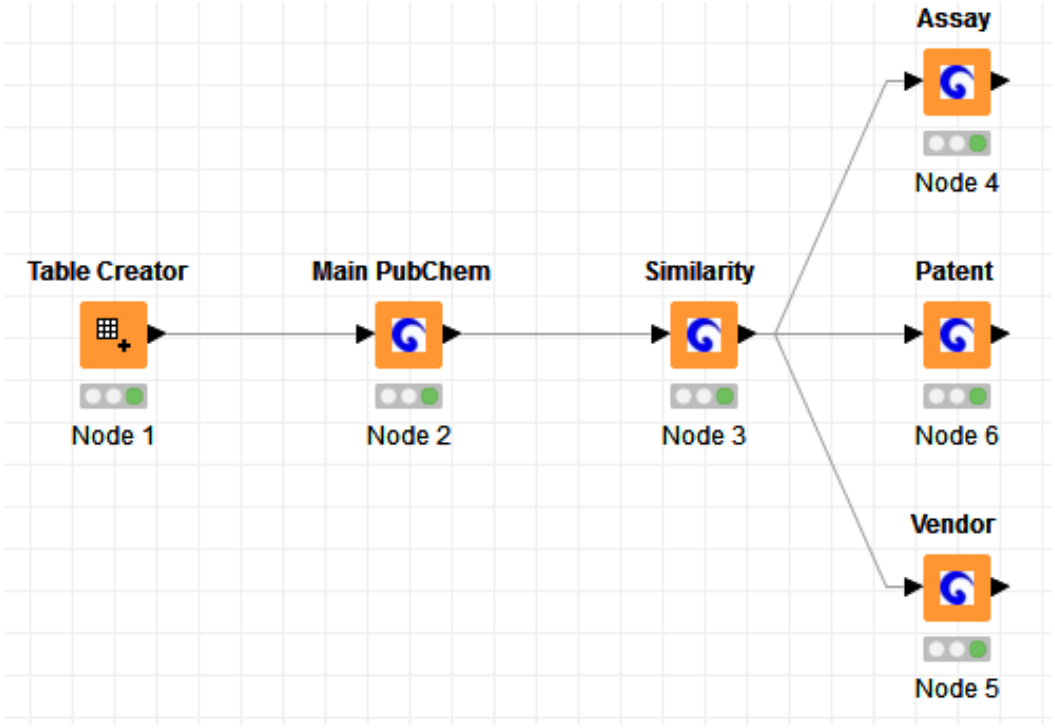


Fig. 15: Workflow using Enalos+ nodes

Step 10-Extending the main Workflow

Now you can extend the previous workflow by adding other NCI, PubChem and UniChem nodes. You can connect the *Main PubChem's* output with *CIR*, *UniChem* and *UniChem Connectivity*. To configure *CIR* node, you will need to specify the "Target Column" and the "Selected Representations" as shown in Fig. 16. In *UniChem* node you need to select InChI-Key as Input and consequently choose some of the available databases (Fig. 17). *UniChem Connectivity* node is configured in the same way as *UniChem* (Fig. 18).

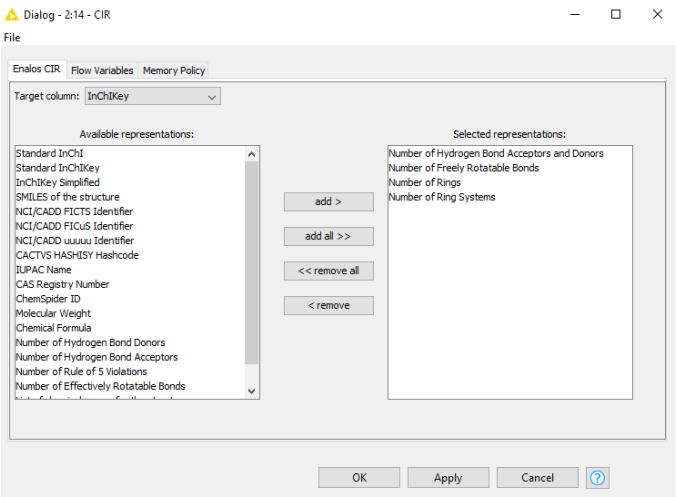


Fig. 16: Configuring CIR node

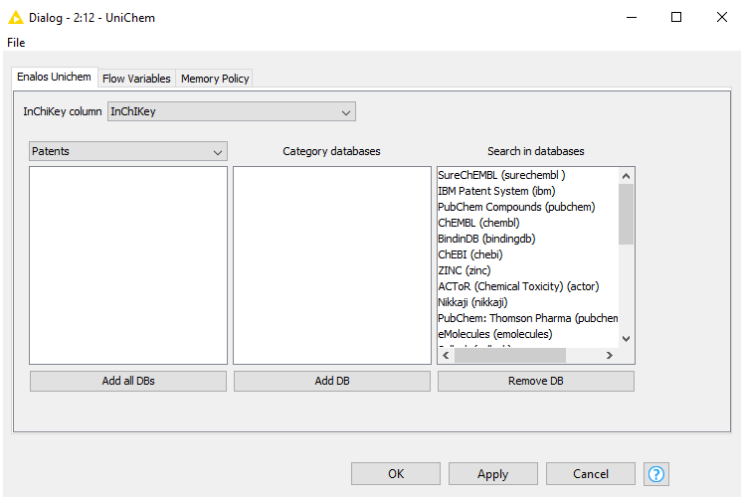


Fig. 17: Configuring UniChem node

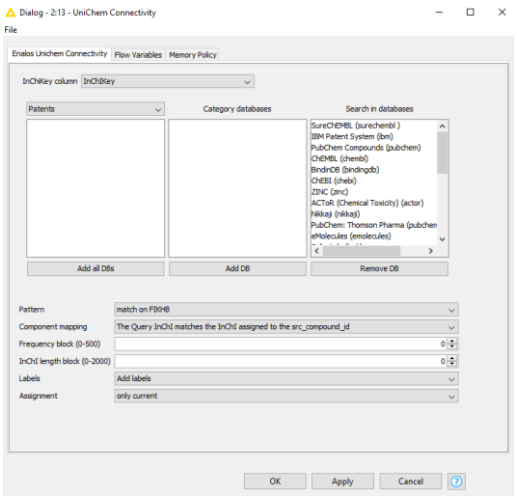


Fig. 18: Configuring UniChem Connectivity node

Then, try to add the following PubChem nodes as shown in Fig. 19.

- *Remove Duplicates* node removes the rows of the input table that contain the same values in selected columns. In this case you can select the CID column in the configuring menu (Fig. 23).
- Consequently, you can add 2 *Assay Class* nodes, in order to display both active and inactive compounds. In Fig. 24 you can see the "active" *Assay Class* configuring menu.
- In *Patent to Sid* node, give the Patent's ID as input, in order to obtain compounds' SIDs that are covered by the patents in request (Fig. 25).
- Finally, the *Sid* node, takes the compounds' SIDs as input and outputs the PubChem CIDs and synonyms (Fig. 26).

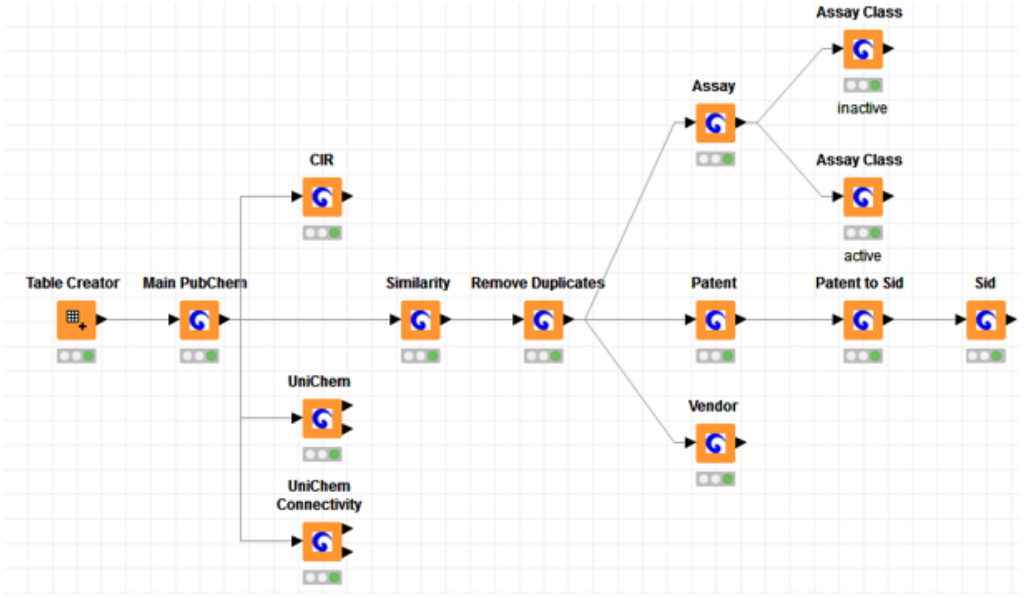


Fig. 19: Extended version of the main Workflow

- In order to control easily the data flow, you can use *Column Splitter* and *Row Splitter* nodes. Connect the *Column Splitter* to the output of *Remove Duplicates*, *Assay*, *Patent* and *Patent to Sid* nodes. You can select only one column for the Top partition and all the other columns for the Bottom partition (Fig. 21).[1] The Top partition Column will be used as input to the following nodes. For example, C.S.4 Top Partition ("SID") is used as *SID* node input. You can also connect *Row Splitter* to the output of *Assay* and *Patent* nodes, in order to export a data-table with less rows meeting the specified criteria (Fig. 22).
- The converted workflow, using the *Column Splitter* and *Row Splitter* nodes is depicted in Fig. 20.

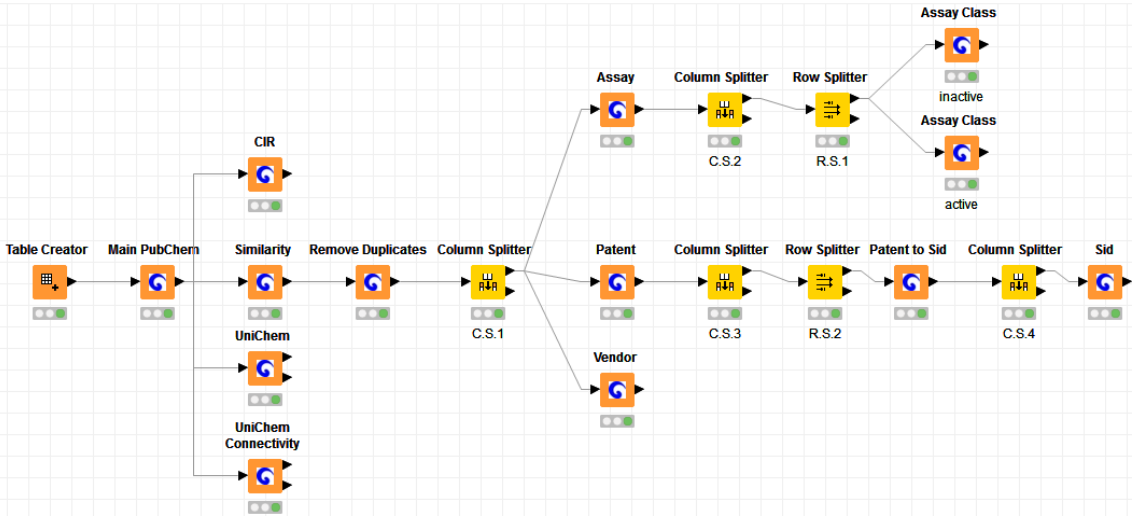


Fig. 20: Final version of the Workflow

[1] In the 1st *Column Splitter* select only CID for the Top Partition
In the 2nd *Column Splitter* select only AID for the Top Partition
In the 3rd *Column Splitter* select only Patent ID for the Top Partition
In the 4th *Column Splitter* select only SID for the Top Partition

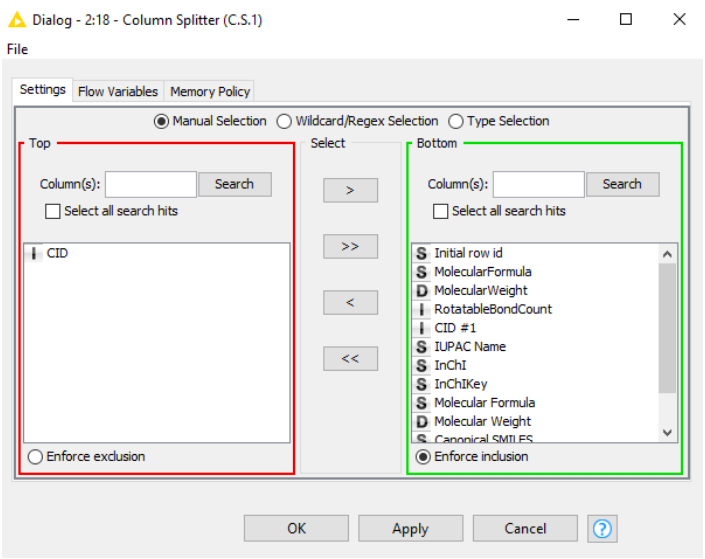


Fig. 21: Configuring Column Splitter

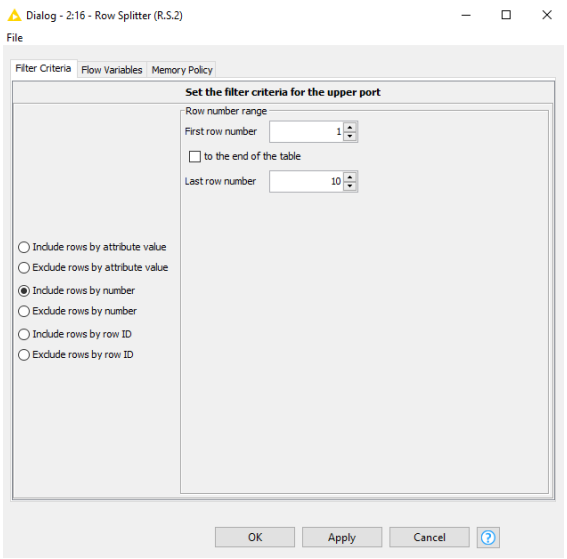


Fig. 22: Configuring Row Splitter

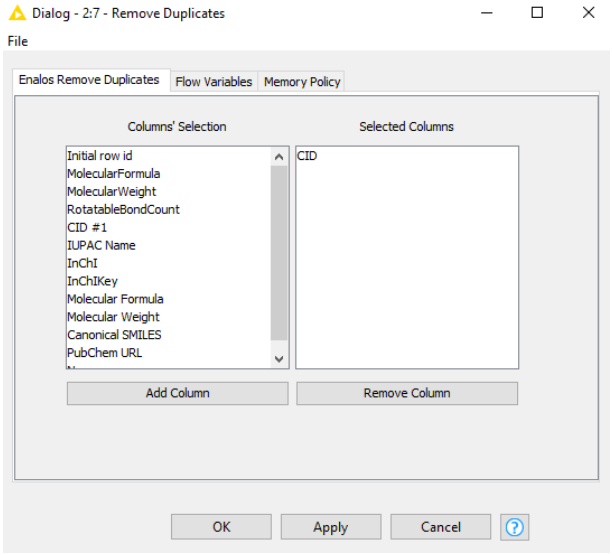


Fig. 23: Configuring Remove Duplicates node

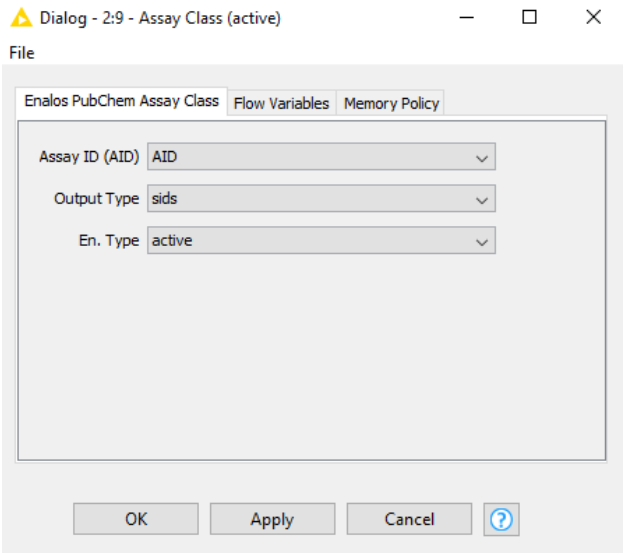


Fig. 24: Configuring Assay Class node

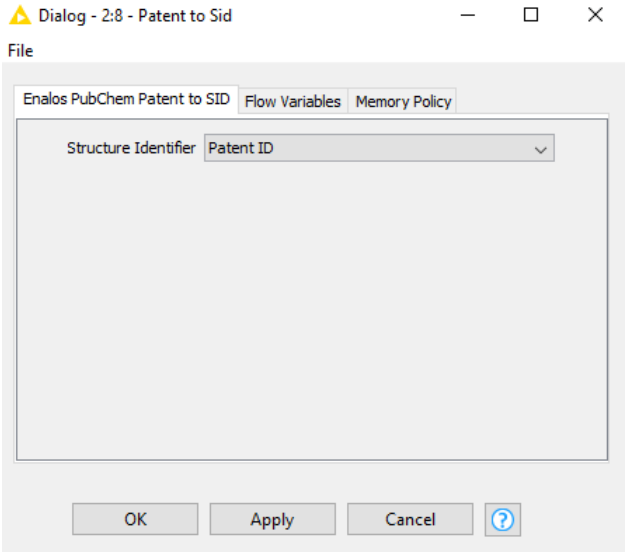


Fig. 25: Configuring Patent to Sid node

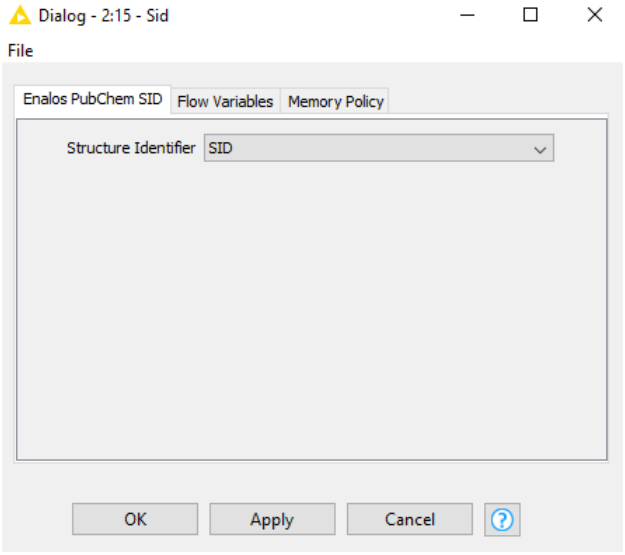


Fig. 26: Configuring Sid node

Now, inspect the results of the added nodes:

- *CIR* node exports a table with all pre-requested representations (Fig. 27).
- *UniChem* and *UniChem Connectivity* nodes output the InChiKeys that were found in the pre-selected UniChem databases and the InChiKeys that were not found in these databases (Fig. 28).
- *Assay Class* node exports a table with information from PubChem about PubChem SIDs (Substance IDs) of compounds that are active or inactive (Fig. 29, Fig. 30).
- *Patent to Sid* extracts a table with information from PubChem about PubChem SIDs of compounds that are covered by the patents in request. In case that no compounds are found under a patent, the number -1 is assigned as SID of this patent (Fig. 31).
- *Sid* outputs information from PubChem about CIDs and synonyms (Fig. 32).

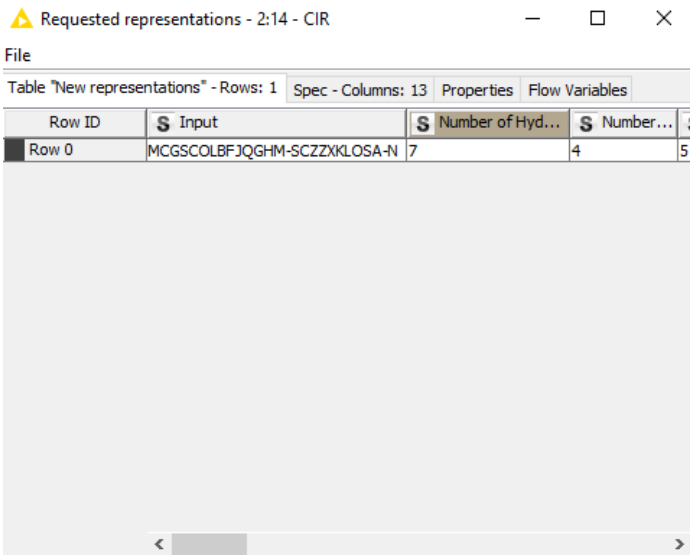


Fig. 27: CIR results

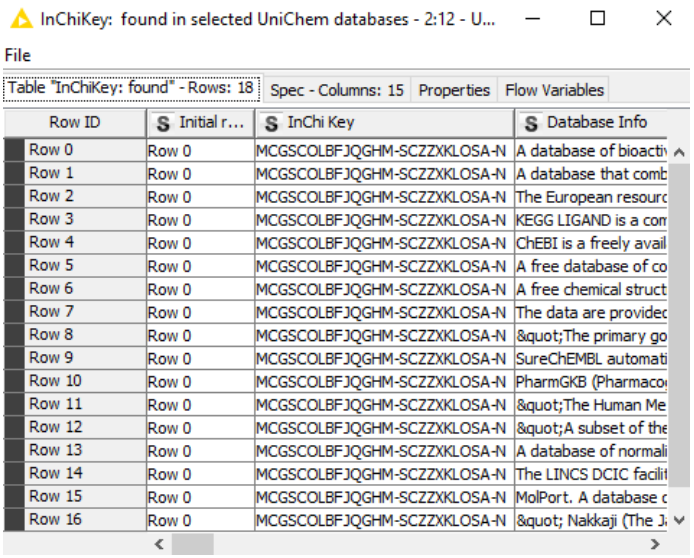


Fig. 28: UniChem results

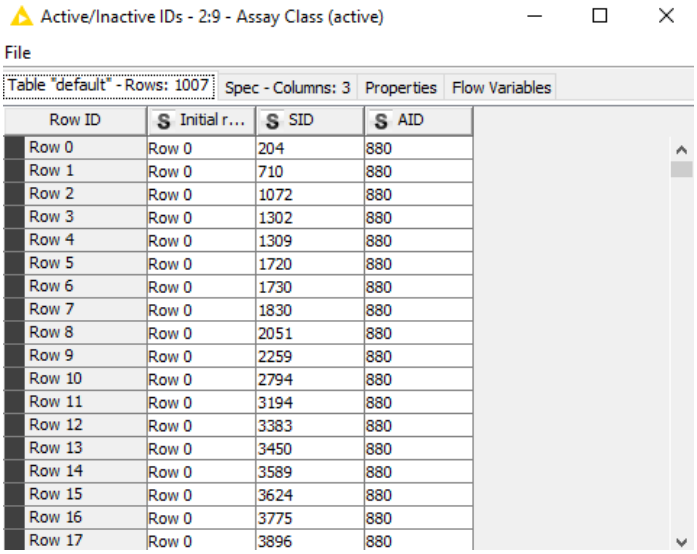


Fig. 29: Assay Class results (active)

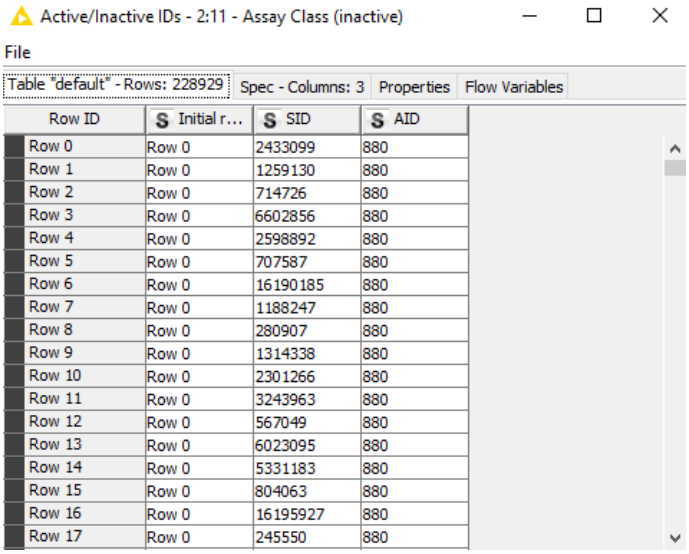


Fig. 30: Assay Class results (inactive)

SIDs under patent - 2:8 - Patent to Sid

File

Table "default" - Rows: 2035 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	Initial r...	SID	Patent ID
Row 0	Row 0	127415408	EP0628044B1
Row 1	Row 0	127427561	EP0628044B1
Row 2	Row 0	127564905	EP0628044B1
Row 3	Row 0	127602502	EP0628044B1
Row 4	Row 0	127720939	EP0628044B1
Row 5	Row 0	127804557	EP0628044B1
Row 6	Row 0	127807700	EP0628044B1
Row 7	Row 0	127965093	EP0628044B1
Row 8	Row 0	128037226	EP0628044B1
Row 9	Row 0	128090495	EP0628044B1
Row 10	Row 0	128110507	EP0628044B1
Row 11	Row 0	128165640	EP0628044B1
Row 12	Row 0	128167461	EP0628044B1
Row 13	Row 0	128211172	EP0628044B1
Row 14	Row 0	128246082	EP0628044B1
Row 15	Row 0	128310878	EP0628044B1
Row 16	Row 0	128413036	EP0628044B1
Row 17	Row 0	128428285	EP0628044B1

Fig. 31: Patent to Sid results

CIDs and synonims - 2:15 - Sid

File

Table "default" - Rows: 2035 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	Initial r...	SID	CID	synonyms	name
Row 51	Row 51	135764025	22721337		816
Row 52	Row 52	135855649	11108834		816
Row 53	Row 53	135855990	21310051		816
Row 54	Row 54	135933200	11459511		816
Row 55	Row 55	135996759	57133849		816
Row 56	Row 56	136042524	57155707		816
Row 57	Row 57	136045151	57157246		816
Row 58	Row 58	136049641	57159919		816
Row 59	Row 59	136178896	57236764		816
Row 60	Row 60	136209496	57255035		816
Row 61	Row 61	226393245	6101	SCHEMBL34	4488
Row 62	Row 62	226393282	6547	SCHEMBL122	4488
Row 63	Row 63	226393293	244	SCHEMBL147	4488
Row 64	Row 64	226393308	21226206	SCHEMBL172	4488
Row 65	Row 65	226393327	7509	SCHEMBL205	4488
Row 66	Row 66	226393487	7799	SCHEMBL454	4488
Row 67	Row 67	226393712	107428	SCHEMBL762	4488

Fig. 32: Sid results

Embark your own voyage of discovery!

Now, this was just a simple example to get you started. There is a lot more to discover. Try to explore it! We tried to keep it simple and intuitive. We would love to receive your feedback and find out what you liked and what you did not like; things you find not functional or things that did not seem to work.