NovaMechanics | Chemoinformatics & Bioinformatics Solutions

KNIME Enalos+ Modelling nodes

A Brief Tutorial

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

Table of Contents

Introduction

Rapid development of information and communication technologies during the last few decades has dramatically changed our capabilities of collecting, analyzing, storing and disseminating all types of data. This process has had a profound influence on the scientific research in many disciplines, including the development of new generations of effective and selective medicines. Large databases containing millions of chemical compounds tested in various biological assays such as PubChem are increasingly available as online collections. In order to find new drug leads, there is a need for efficient and robust procedures that can be used to screen chemical databases and virtual libraries against molecules with known activities or properties. To this end, Quantitative Structure-Activity Relationships (QSAR) modelling provides an effective means for both exploring and exploiting the relationship between chemical structure and its biological action towards the development of novel drug candidates.

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 the Modelling Enalos+ nodes of Novamechanics Ltd. This case study deals with a Linear-Quantitative-Structure-Activity-Relationship (QSAR) model, presented for modelling and predicting the inhibition of CXCR3 receptor.

Step 1-Workbench overview

The KNIME workbench is organized as follows:

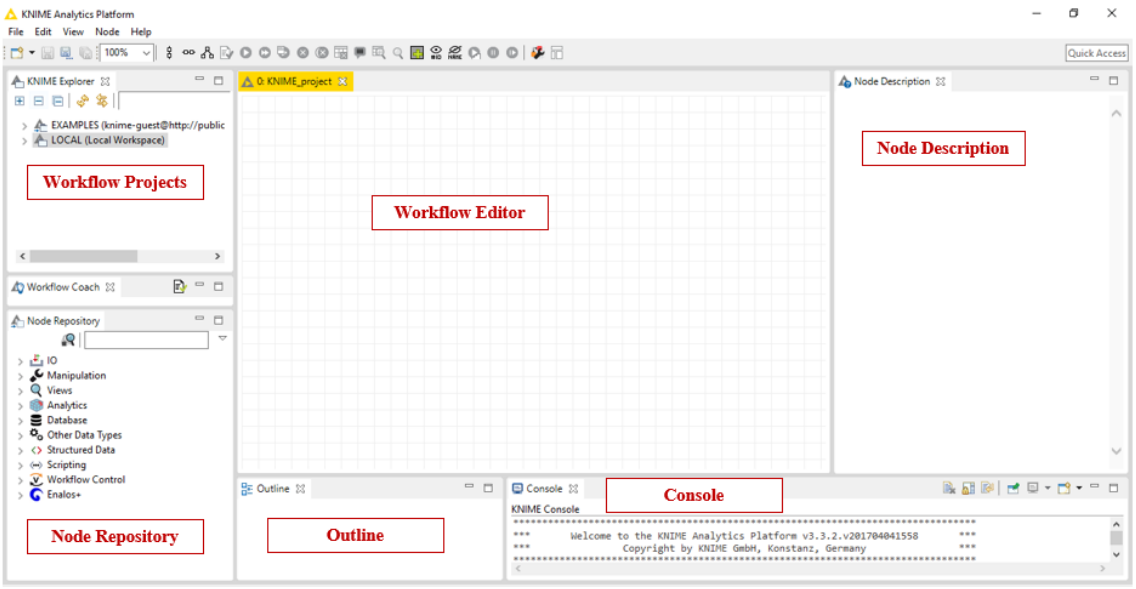


Fig. 1: KNIME workbench

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.

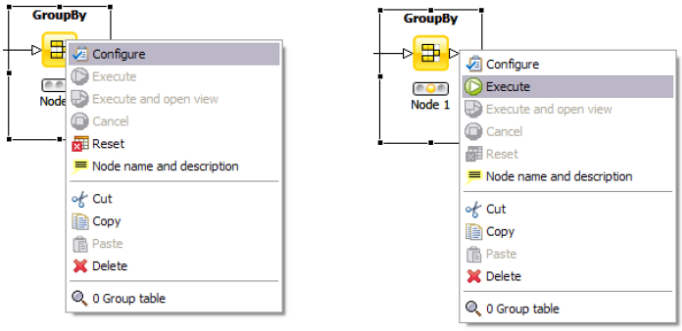


Fig. 2: Configuring and executing nodes

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 Depository, expand the *IO* and the contained *Read* category and choose *Excel Reader (XLS)* node (Fig. 3). Then, drag & drop the *Excel Reader (XLS)* icon into the Workflow Editor window. Do it twice, in order to have 2 *Excel Reader (XLS)* in the Workflow editor.

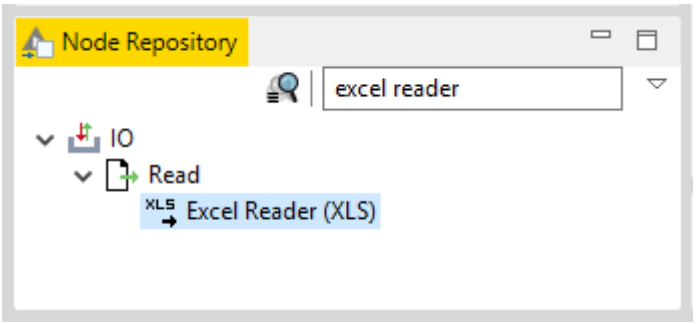


Fig. 3: Node Depository interface

Then, expand the *Enalos+* category followed by the *Modelling* category and drag into the Workflow Editor *MLR* and *Model Acceptability Criteria* as shown below (Fig. 4).

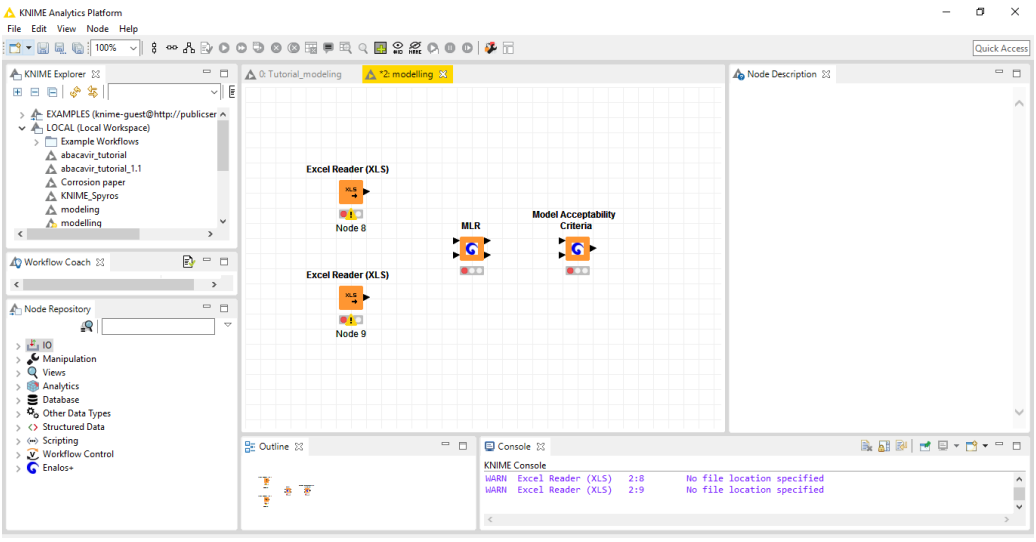


Fig. 4: Workflow editor

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). *MLR* node takes in the 1st input port the "training set" data and in the 2nd input port the "test set" data. *Model Acceptability Criteria* node takes as 1st input a table containing values for the dependent variable, predicted by the model (ypred) and the dependent variable for the test set (yexp), and as 2nd input values for the dependent variable for the training set (ytr). The nodes will not show a green status as long as they are not yet configured and executed.

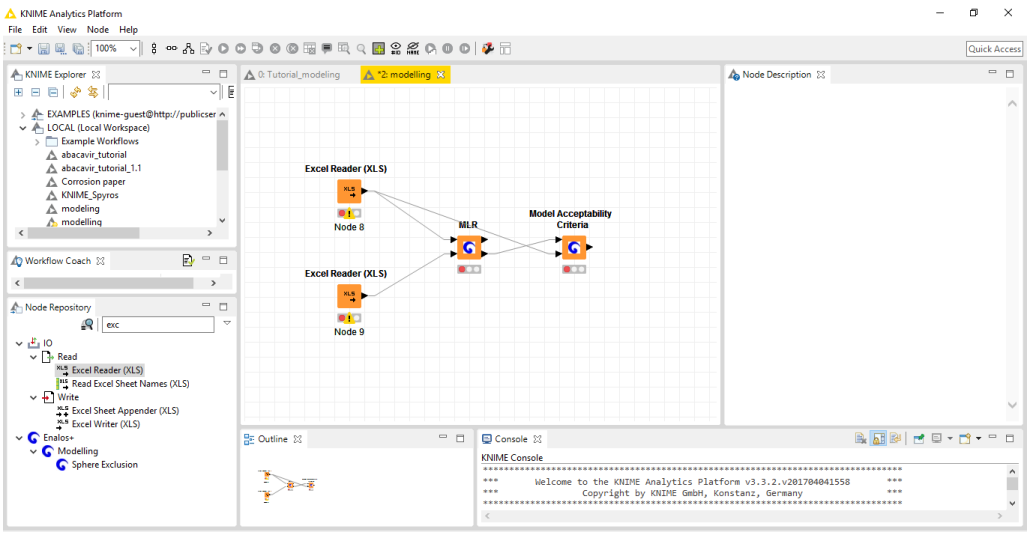


Fig. 5: Connecting nodes

Step 7-Configuring nodes

Fully connected nodes showing a red status icon need to be configured. Start with the *Excel Reader (XLS)*, right click it and select "Configure" from the menu. Press "Browse" button and Select an .xls file to read (Fig. 6). Follow the same steps for the 2nd *Excel Reader (XLS)* node. The 1st node refers to the "training set" and the 2nd node to the "test set". You can also rename these two nodes by "training set" and "test set". Press "Apply" and "OK" to close the dialog of the *Excel Reader (XLS)* nodes. Once the node has been configured correctly, it switches to yellow (meaning ready for execution).

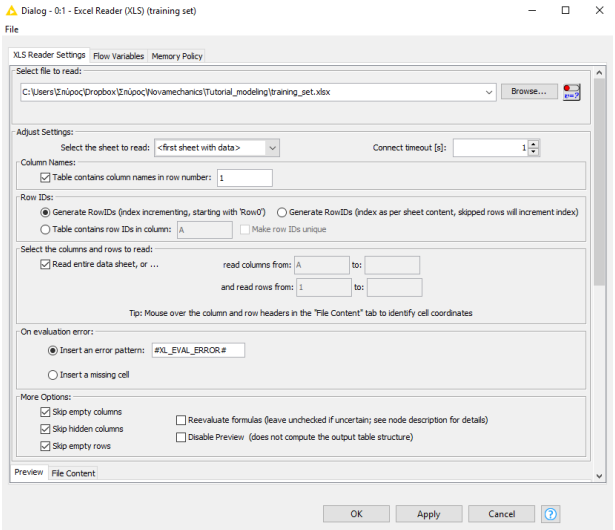


Fig. 6: Configuring the Excel Reader (XLS) node

Then, configure *MLR* node by choosing a "Y train column" and a "Y test column" (Fig. 7). In Fig. 7 " dep_var" is the depended variable (Y). Press "Apply" and "OK" to exit the dialog of the *MLR* node.

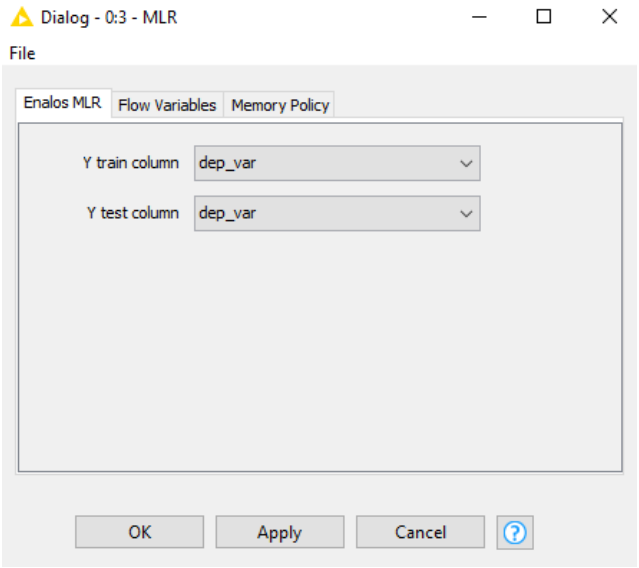


Fig. 7: Configuring MLR node

Model Acceptability Criteria node is configured as shown in Fig. 8. Complete the depended variable's Predicted, Experimental and Training values.

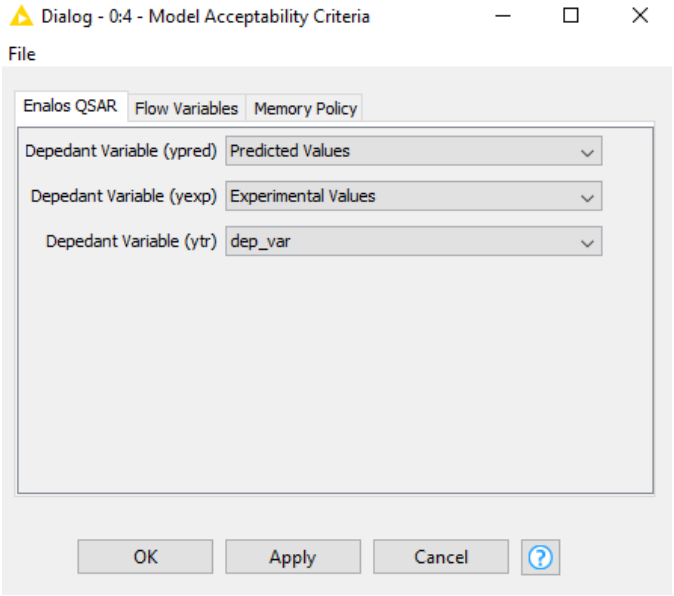


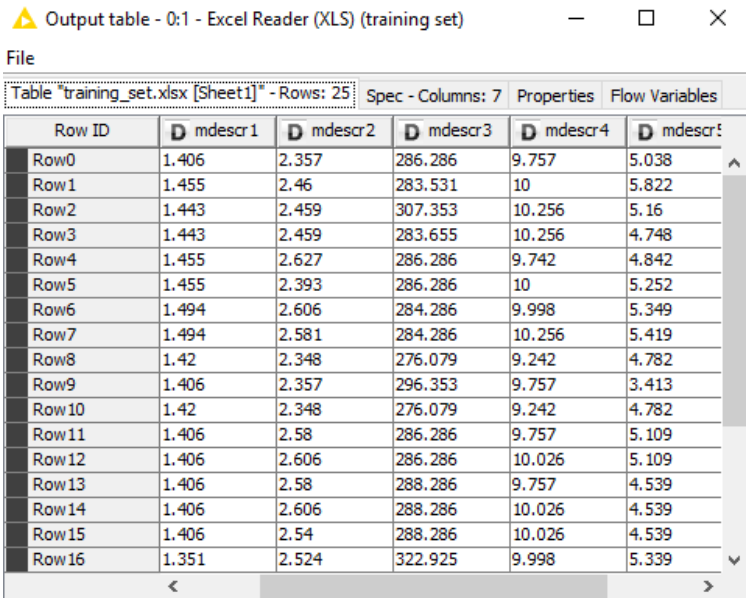
Fig. 8: Configuring Model Acceptability Criteria

Step 8-Executing nodes

Now, right click on the *Model Acceptability Criteria* node and execute it. The workbench will execute all predecessor nodes for you. 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 *Excel Reader (XLS)* output port the table read in is extracted (Fig. 9).


Output table - 0:1 - Excel Reader (XLS) (training set)

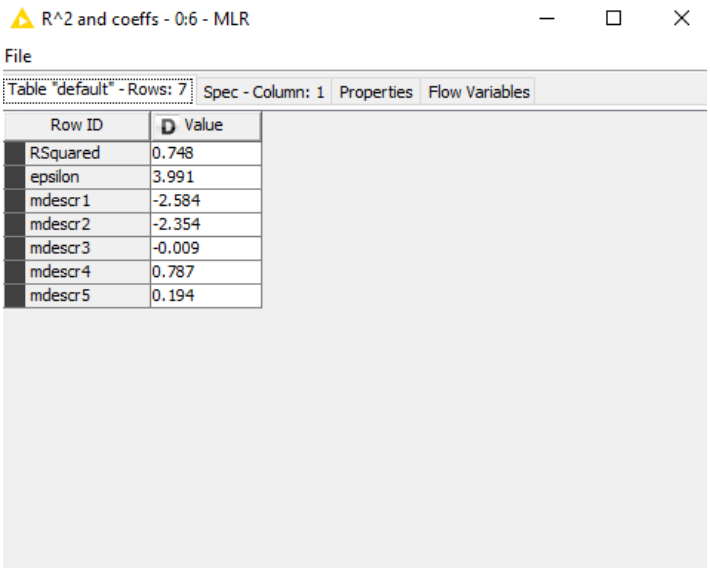
File

Table "training_set.xlsx [Sheet1]" - Rows: 25 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	mdescr1	mdescr2	mdescr3	mdescr4	mdescr5
Row0	1.406	2.357	286.286	9.757	5.038
Row1	1.455	2.46	283.531	10	5.822
Row2	1.443	2.459	307.353	10.256	5.16
Row3	1.443	2.459	283.655	10.256	4.748
Row4	1.455	2.627	286.286	9.742	4.842
Row5	1.455	2.393	286.286	10	5.252
Row6	1.494	2.606	284.286	9.998	5.349
Row7	1.494	2.581	284.286	10.256	5.419
Row8	1.42	2.348	276.079	9.242	4.782
Row9	1.406	2.357	296.353	9.757	3.413
Row10	1.42	2.348	276.079	9.242	4.782
Row11	1.406	2.58	286.286	9.757	5.109
Row12	1.406	2.606	286.286	10.026	5.109
Row13	1.406	2.58	288.286	9.757	4.539
Row14	1.406	2.606	288.286	10.026	4.539
Row15	1.406	2.54	288.286	10.026	4.539
Row16	1.351	2.524	322.925	9.998	5.339

Fig. 9: Excel Reader (XLS) results

MLR node has 2 output ports. The first one extracts a table containing the coefficient of determination (R squared), constant (y-intercept) and the regression coefficients for each independent variable of the multiple linear model (Fig. 10). The second one exports the experimental and the predicted values of the dependent variable y (Fig. 11).

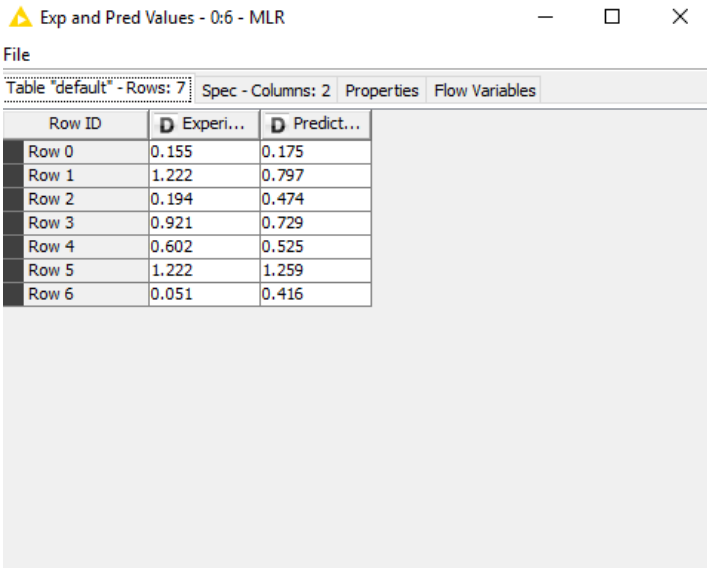

R^2 and coeffs - 0:6 - MLR

File

Table "default" - Rows: 7 | Spec - Column: 1 | Properties | Flow Variables

Row ID	Value
RSquared	0.748
epsilon	3.991
mdescr1	-2.584
mdescr2	-2.354
mdescr3	-0.009
mdescr4	0.787
mdescr5	0.194

Fig. 10: MLR results (1)


Exp and Pred Values - 0:6 - MLR

File

Table "default" - Rows: 7 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	Experi...	Predict...
Row 0	0.155	0.175
Row 1	1.222	0.797
Row 2	0.194	0.474
Row 3	0.921	0.729
Row 4	0.602	0.525
Row 5	1.222	1.259
Row 6	0.051	0.416

Fig. 11: MLR results (2)

Model Acceptability Criteria node exports the Quality of Fit and Predictive Ability Statistics of a continuous QSAR Model (Fig. 12).

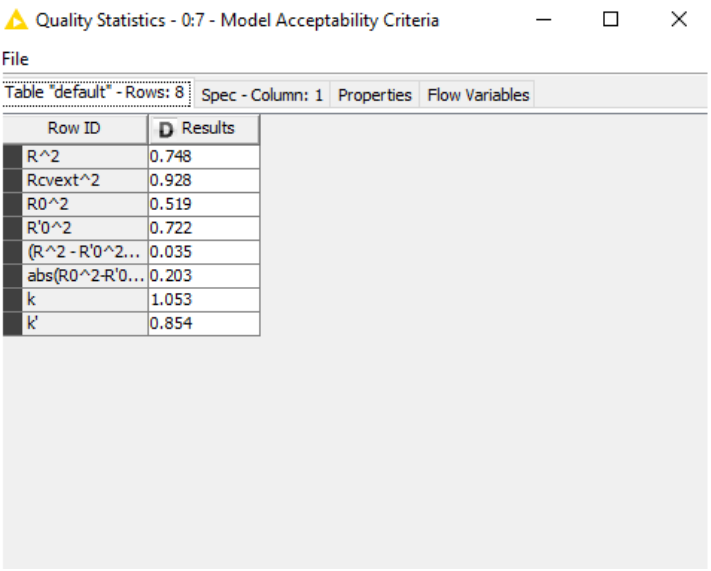


Fig. 12: Model Acceptability Criteria results (2)

The executed workflow is depicted in Fig. 13:

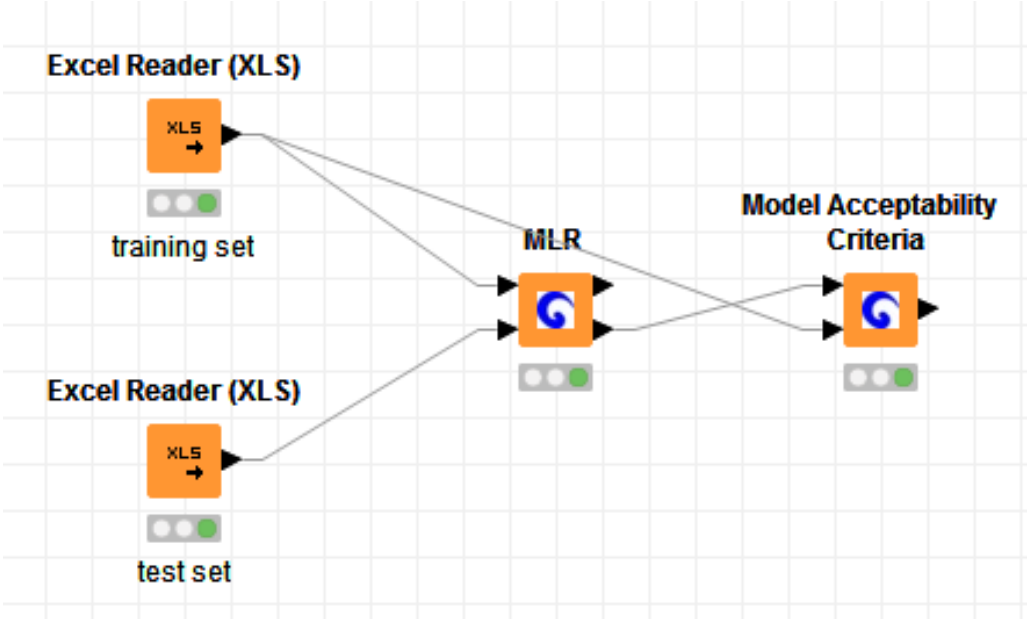


Fig. 13: Workflow using Enalos+ Modelling nodes

Step 10-Extending the main Workflow

Now you can extend the previous workflow by adding other Modelling nodes:

In order to define the model applicability domains, add *Domain-APD* and *Domain-Leverage* nodes. These two nodes, take as input the training test (1st input) and the test set (2nd input), except from the depended variable. To do so, you can add *Column Splitters* in order to remove the depended variable y.

You can also add *Y-Randomization* node. Y-randomization (or Y-scrambling) is a technique, applied to ensure a QSAR model's robustness. This test consists of repeating all the calculations with scrambled values of the response variable of the training set. In this case you will need to

connect a *Column Splitter* to the output of *Y-Randomization* node in order to include only one randomization per time.

Configure these nodes as shown below:

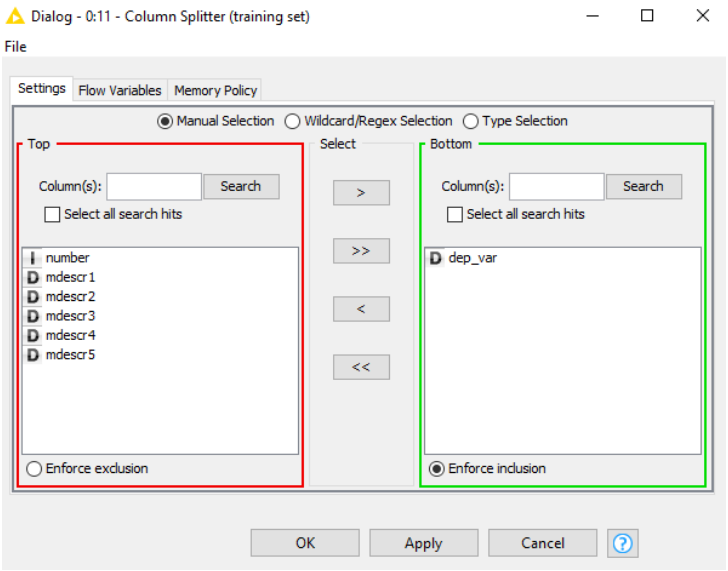


Fig. 14: Configuring Column Splitter node

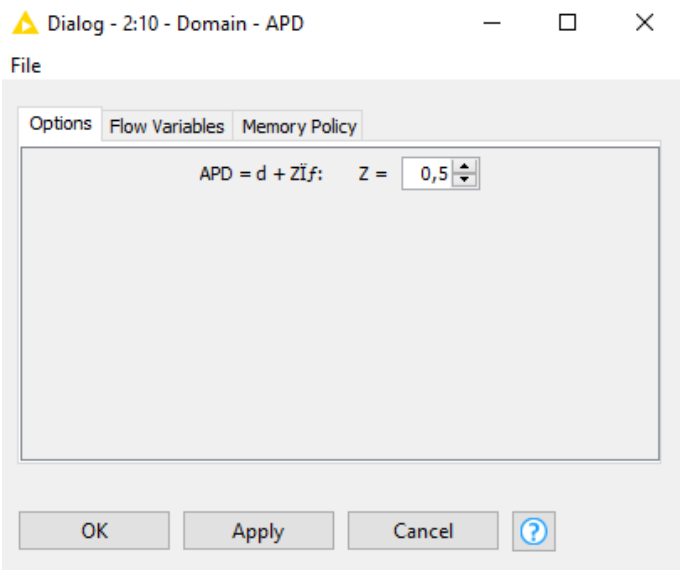


Fig. 15: Configuring Domain-APD node

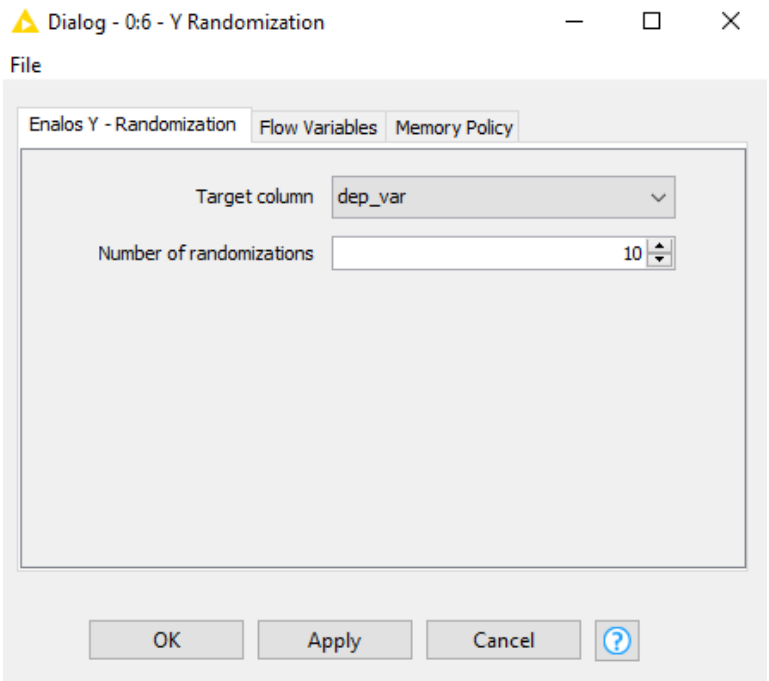


Fig. 16: Configuring Y-Randomization node

The updated workflow is depicted in Fig. 17.

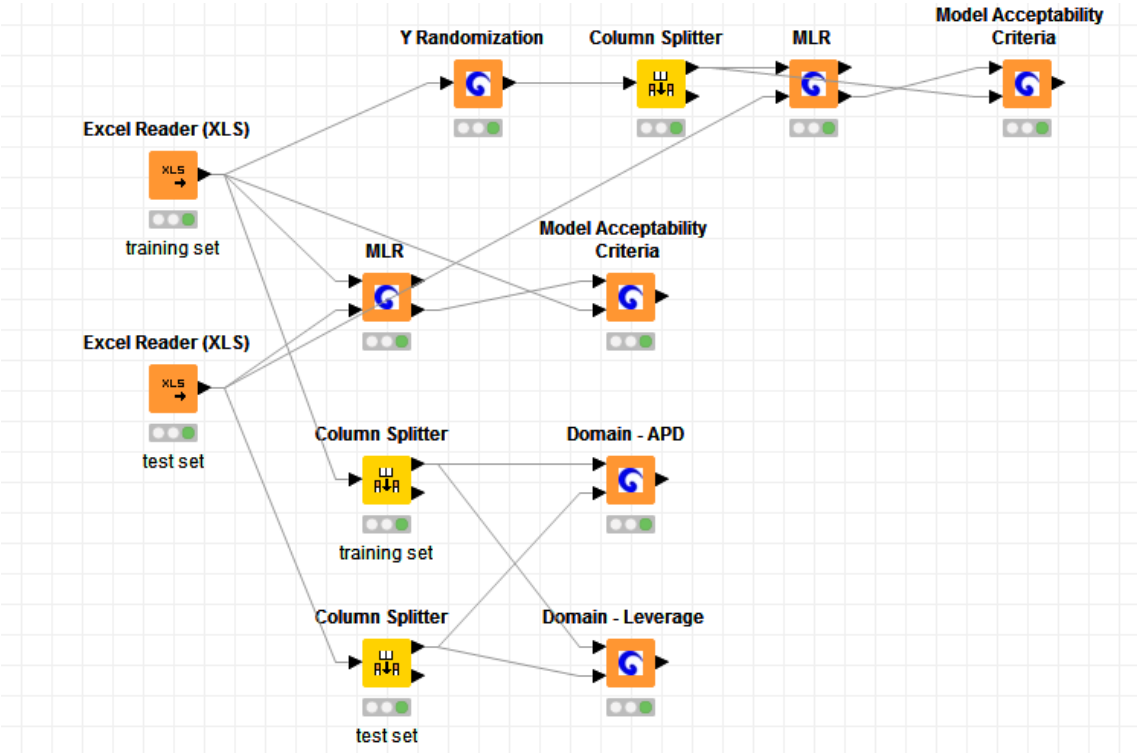


Fig. 17: Updated Workflow

Step-11 Converting the main Workflow

Now, you can convert the main workflow. The workflow was built from the beginning using a training test for the model prediction and a test set for the model validation. Generally, the experimental data are split in training and test sets using the appropriate algorithms. Such algorithms are Kennard-Stone and Sphere Exclusion:

- The Kennard and Stone method allows the selection of two representative subsets (as training and test sets) with a uniform distribution over an initial dataset.
- The Sphere Exclusion method 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.

Drag & drop the *Excel Reader (XLS)* icon into the Workflow Editor window. This node takes all experimental data from an .xls file. Connect the output to *Kennard and Stone* node. Configure *Kennard and Stone* by defining the "Target column" which refers to the depended variable and the "Model percentage". In general, the test set should be about 15-20% of the entire dataset. So, you can select, say, 80% as a "Model percentage" (Fig. 18)

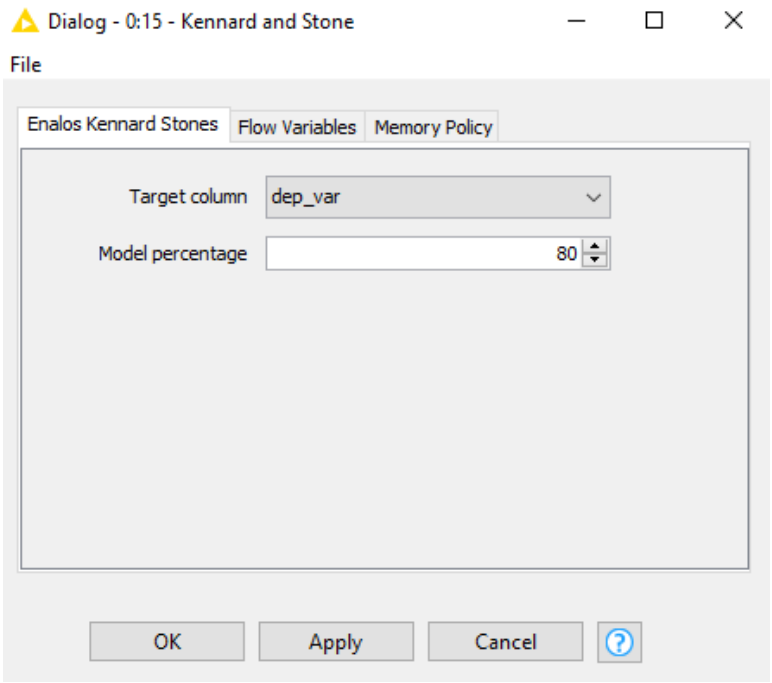


Fig. 18: Configuring Kennard and Stone node

Then, add, connect, configure and execute *MLR* and *Model Acceptability Criteria* nodes in the same way as before (see Fig. 4, Fig. 5, Fig. 7, Fig. 8). The new, converted results of the *MLR* and *Model Acceptability Criteria* nodes are depicted in Fig. 19, Fig. 20, Fig. 21, Fig. 22.

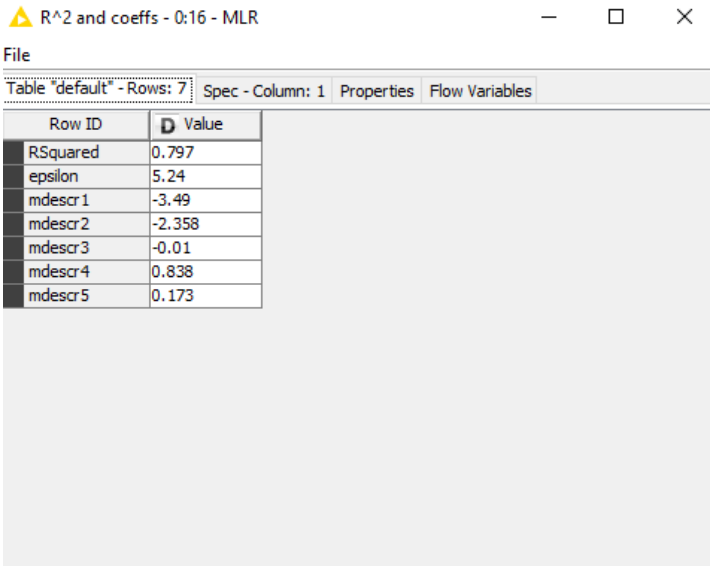


Fig. 19: MLR results (1)

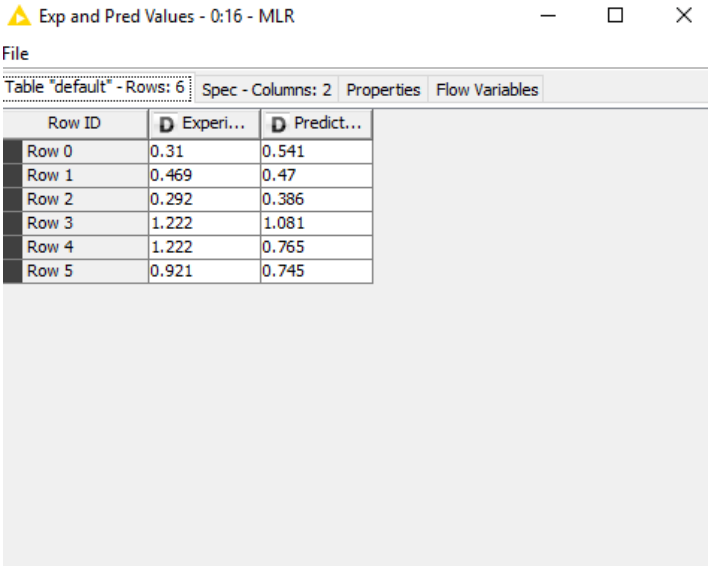


Fig. 20: MLR results (2)

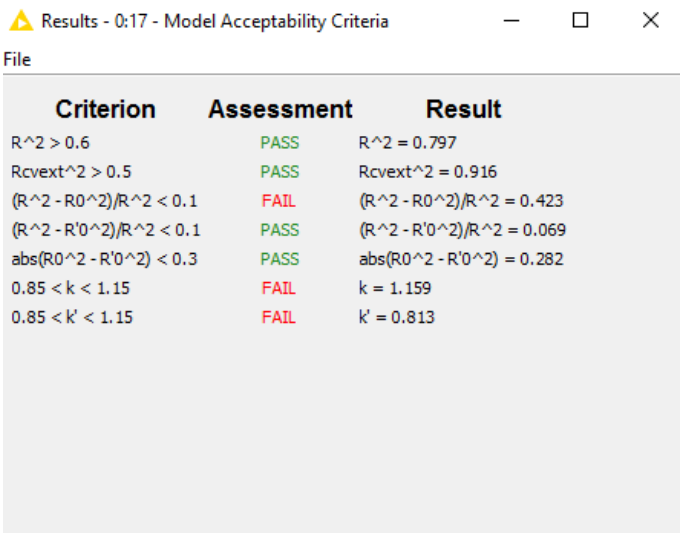


Fig. 21: Model Acceptability Criteria results (1)

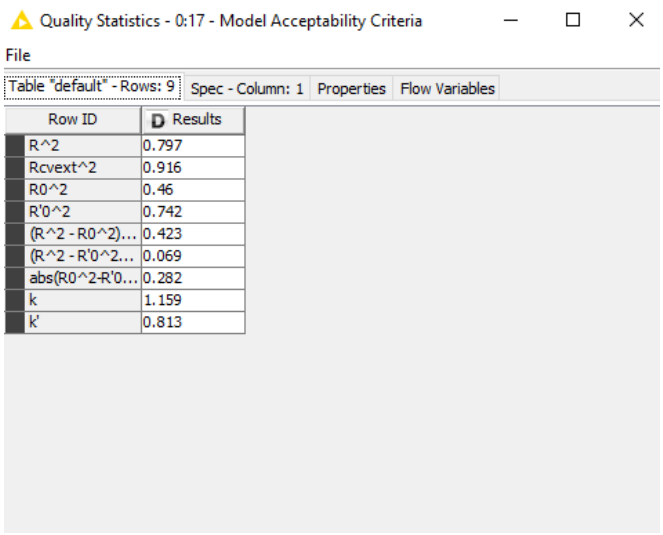


Fig. 22: Model Acceptability Criteria results (2)

Now, you can construct another branch. Connect the *Excel Reader (XLS)* output with *Sphere Exclusion* node instead of *Kennard and Stone* node. Then, add, connect, configure and execute *MLR* and *Model Acceptability Criteria* nodes in the same way as before. Inspect the results of the new pipeline (Fig. 23, Fig. 24).

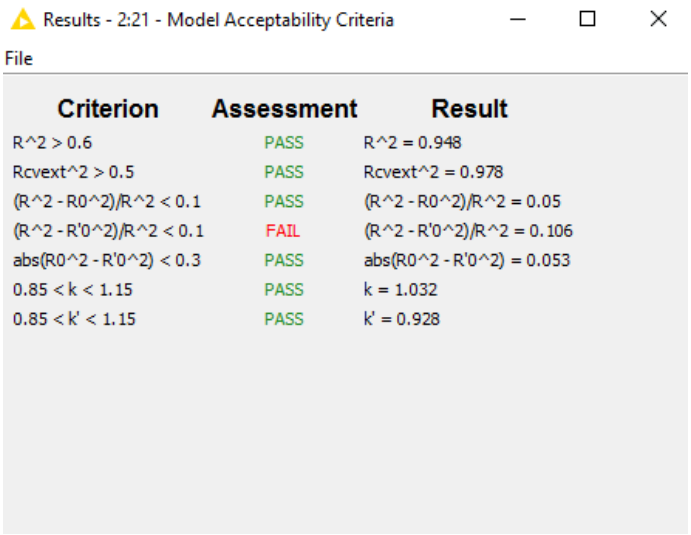


Fig. 23: Model Acceptability Criteria results (1)

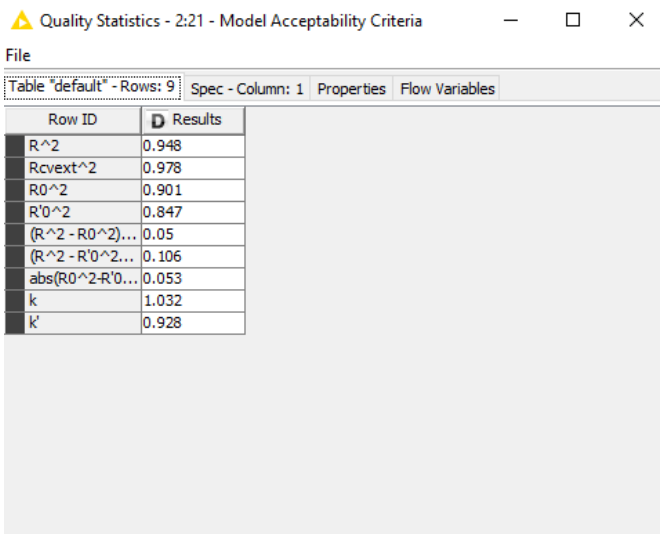


Fig. 24: Model Acceptability Criteria results (2)

The new, converted workflow is depicted in Fig. 25:

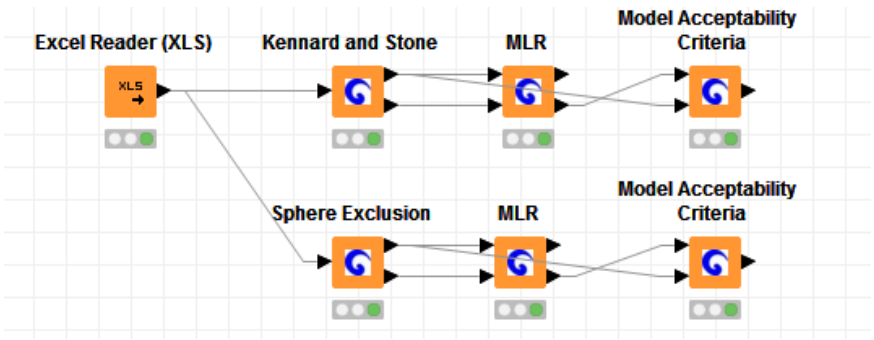


Fig. 25: Converted Workflow

You can also *Domain-APD* and *Domain-Leverage* nodes in order to define the model applicability domains and *Y-randomization* node to ensure a QSAR model's robustness (see Step 10-Extending the main Workflow). Obviously the extension possibilities of the workflow can be endless. The user can add other KNIME nodes and combine them with Enalos+ nodes in order to construct the appropriate model.

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.