Open for Innovation
KNIME

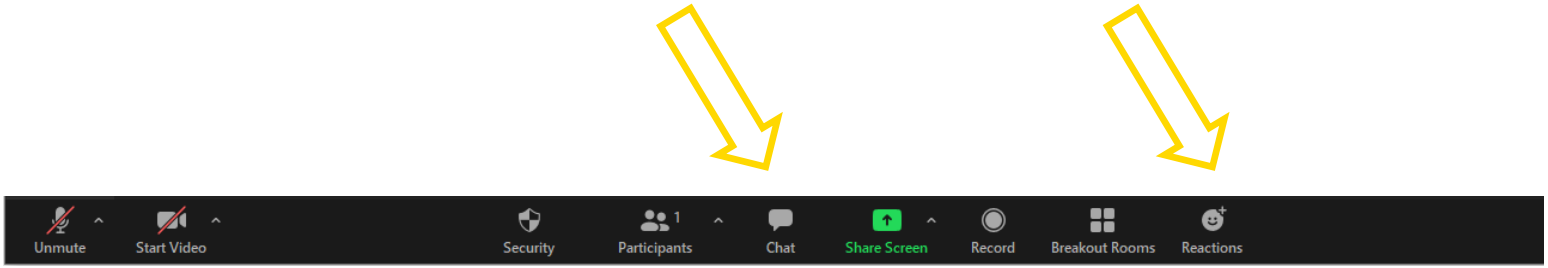
Hands-on: Data preparation and interactive visualization of chemical structures in KNIME Analytics Platform

AIDD
April 19th, 2022

Daria Goldmann
KNIME AG

Before we start…

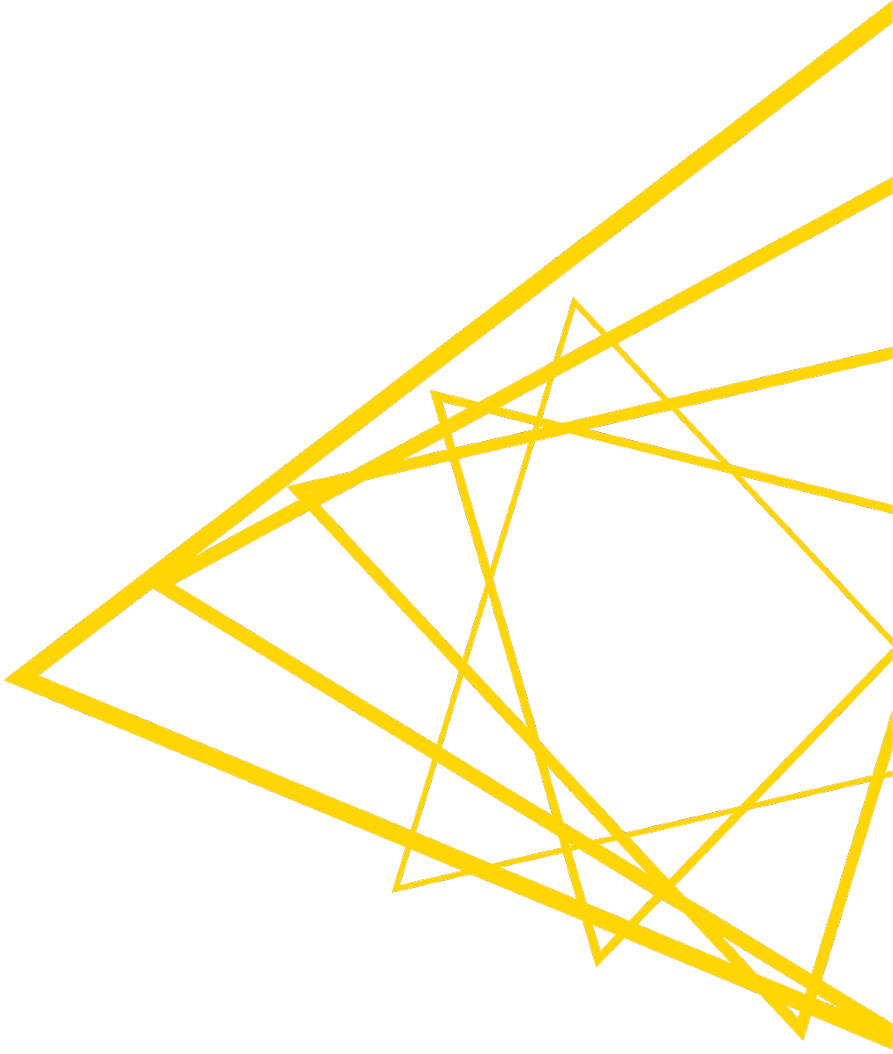
- The session is recorded
- Recording and slides are shared after the session with all participants
- Please use the chat to post your questions
- Please use the reactions to e.g. raise your hand and ask the questions live



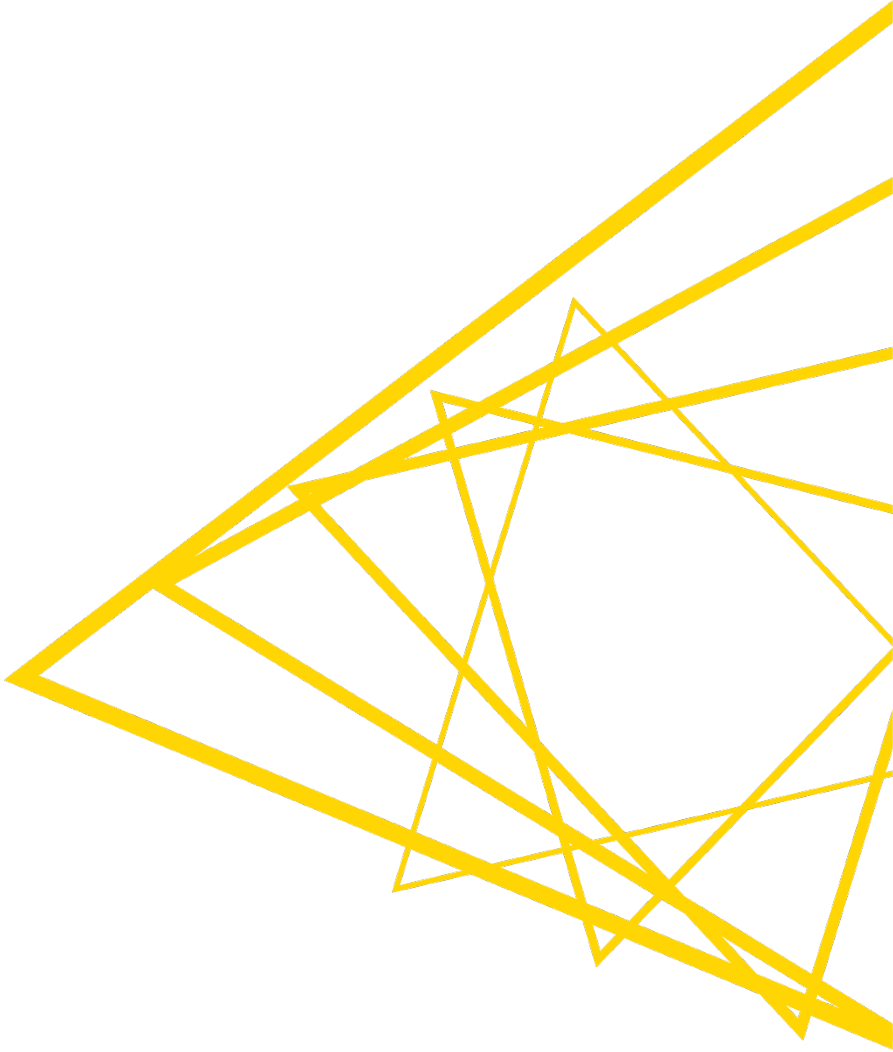
Scenario

- You are a computational chemist in a project
- Your task is to develop a model to estimate LogD
- How and where do you start?
- Previous model is built on the inhouse data (quite some time ago)
- You have collected some public data
- And there is a fresh set from the project
- You would like to harmonize the data: get rid of redunces, standardize chemical structures, remove duplicates
- You would like to develop an interactive visualization of the dataset to be able to explore the dataset, filter the data based on the insights, discuss the project data with the coleagues

Teaser of the view

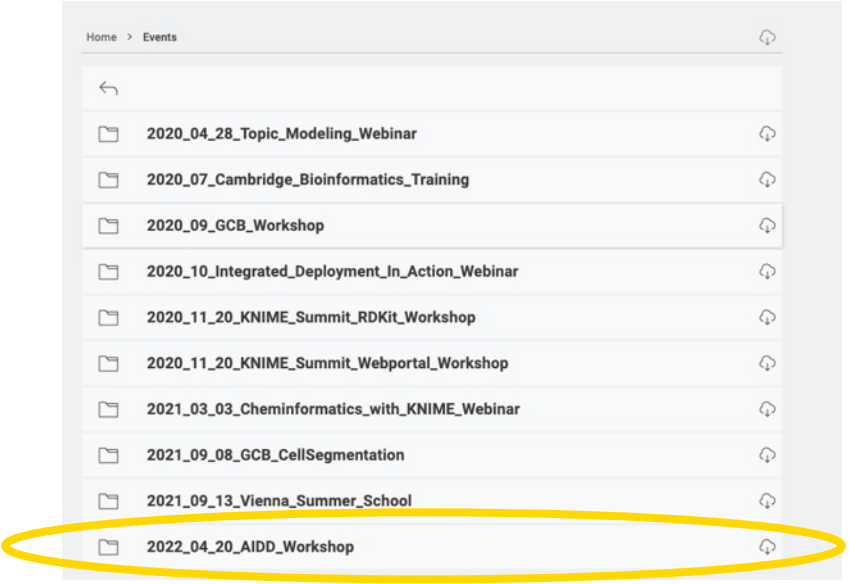


Setup



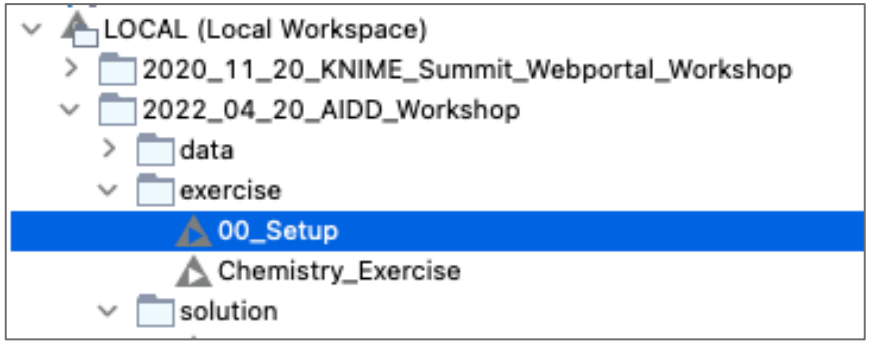
Set up KNIME Analytics Platform

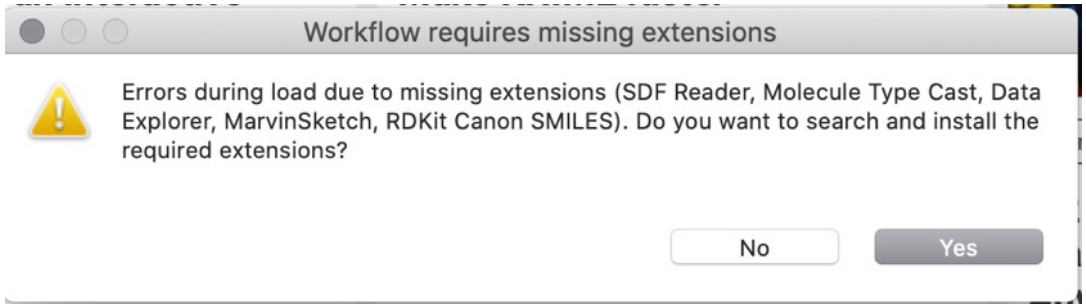
- Download exercises (2022_04_20_AIDD_Workshop) from https://hub.knime.com/knime/spaces/Life%20Sciences/latest/Events~ekRSJneVS0b0RD_k/



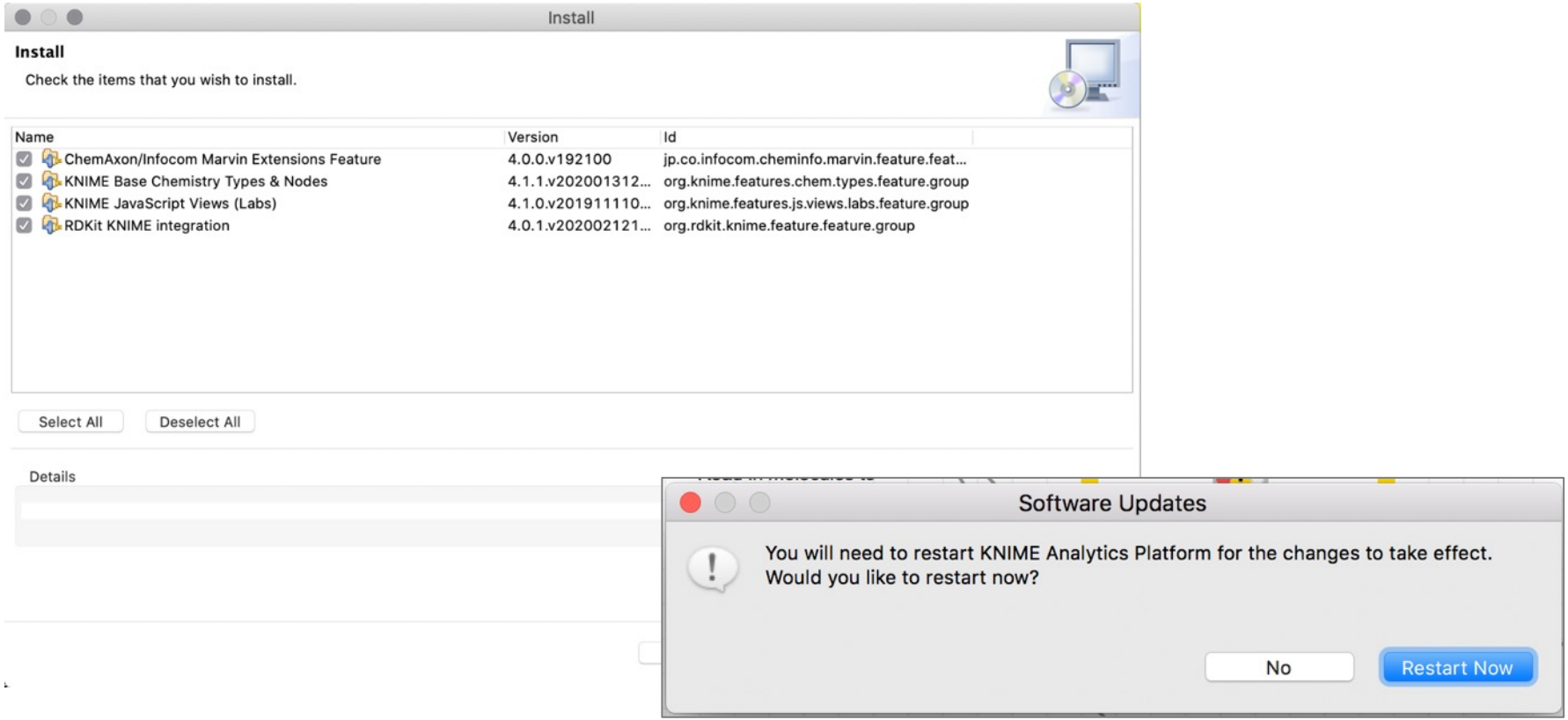
- Import the file into your local workspace

Open the 00_Setup workflow

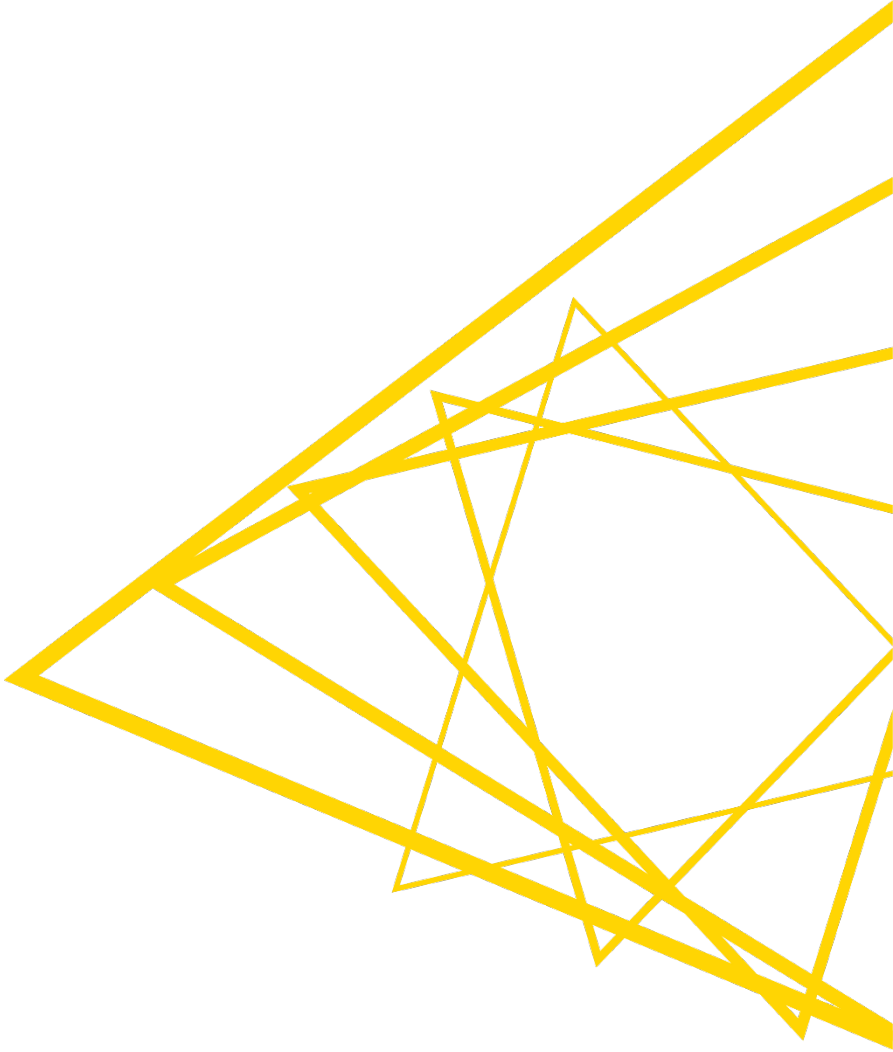




Install the Extensions and Restart



You are now ready to go

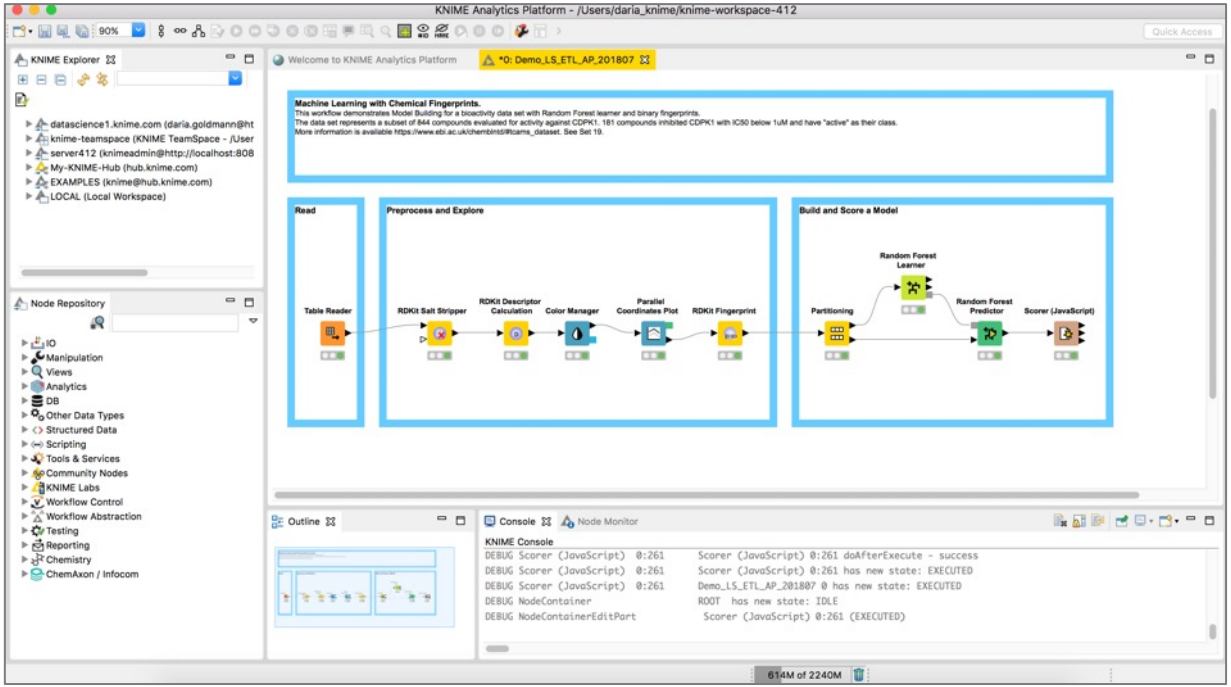


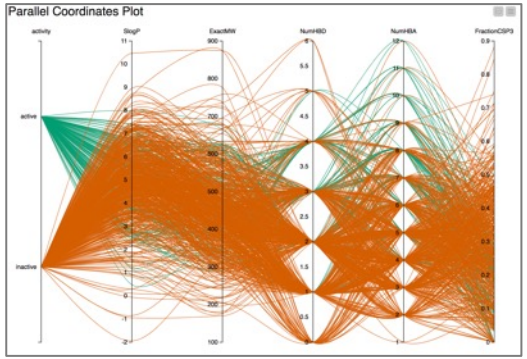
Agenda for Today

- Quick intro to KNIME
- Chemistry in KNIME
  - Chemistry formats
  - Standardization, duplicate removal
  - Filtering on multiple properties
  - Rendering chemical structures
  - Building a component
  - Saving files
- Work on the exercise and ask questions
- Explore possible solution

Brief intro to KNIME

The KNIME® Analytics Platform





Scorer View

Confusion Matrix

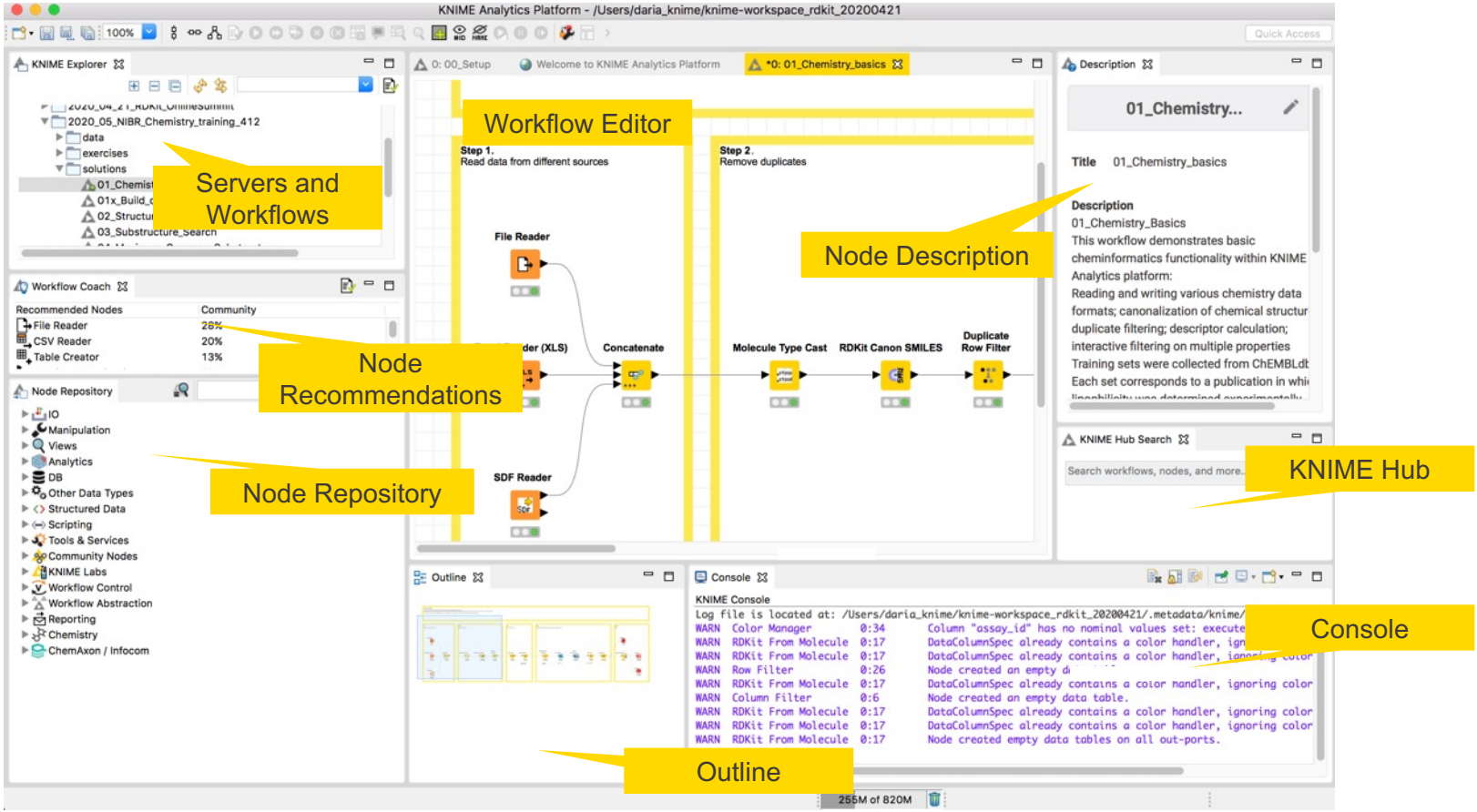
	active (Predicted)	inactive (Predicted)	
active (Actual)	27	9	75.00%
inactive (Actual)	4	129	96.99%
	87.10%	93.48%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
92.31%	7.69%	0.758	156	13

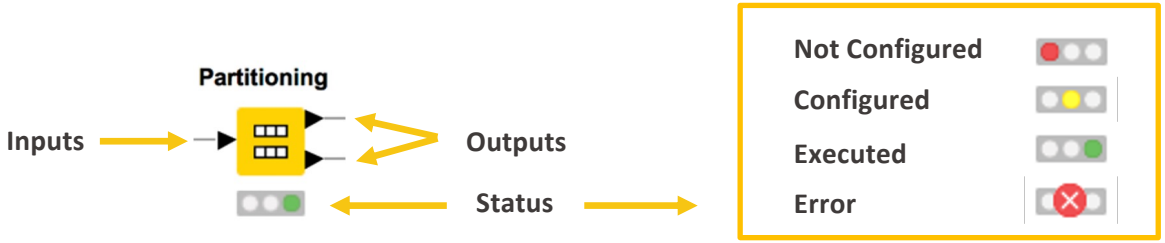
Open for Innovation KNIME

The KNIME Workbench

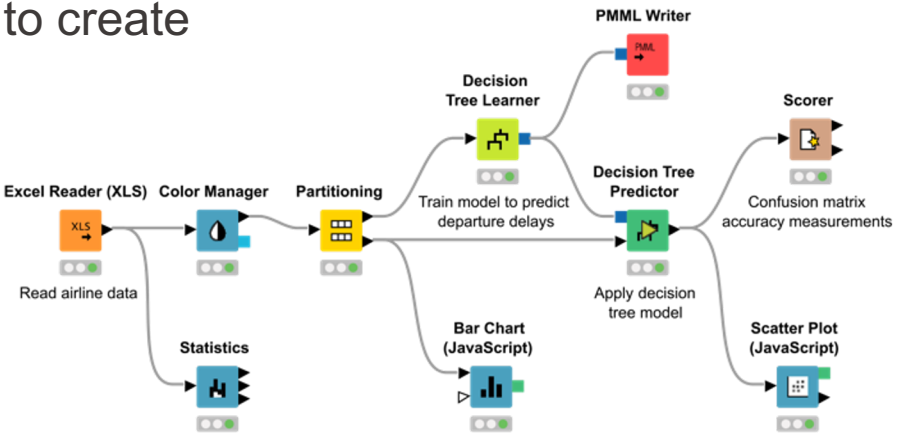


Visual KNIME Workflows

NODES perform tasks on data



Nodes are combined to create **WORKFLOWS**

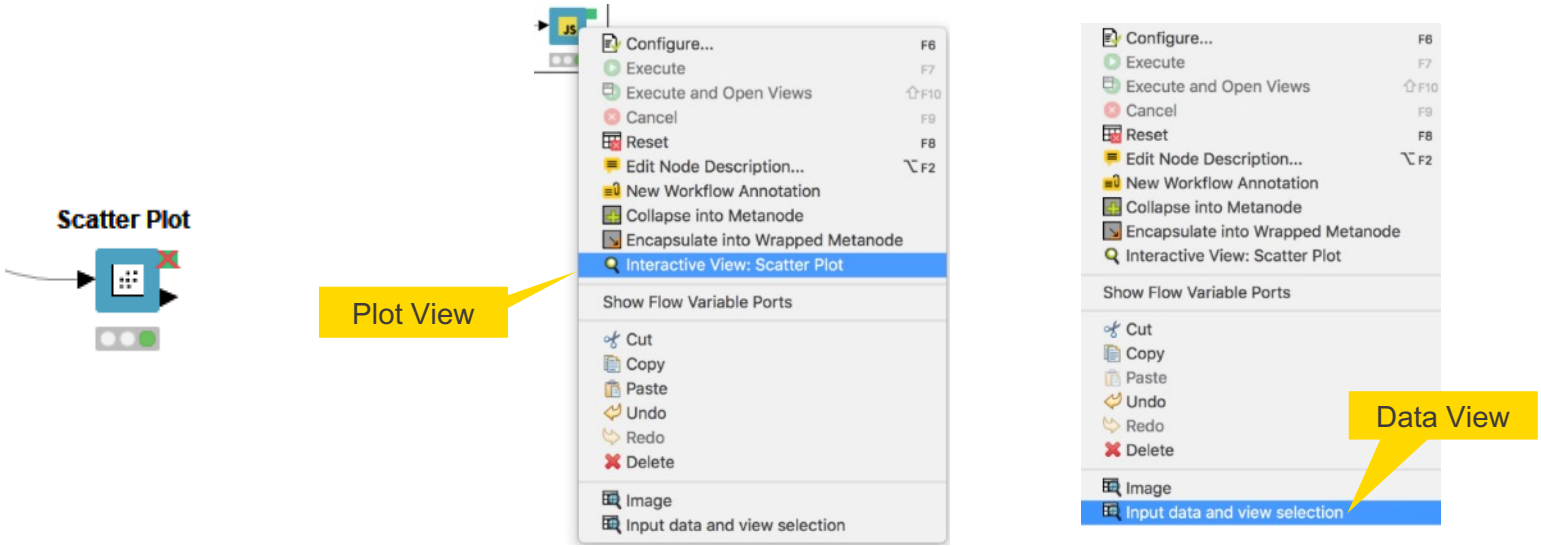


Node Outputs and Views

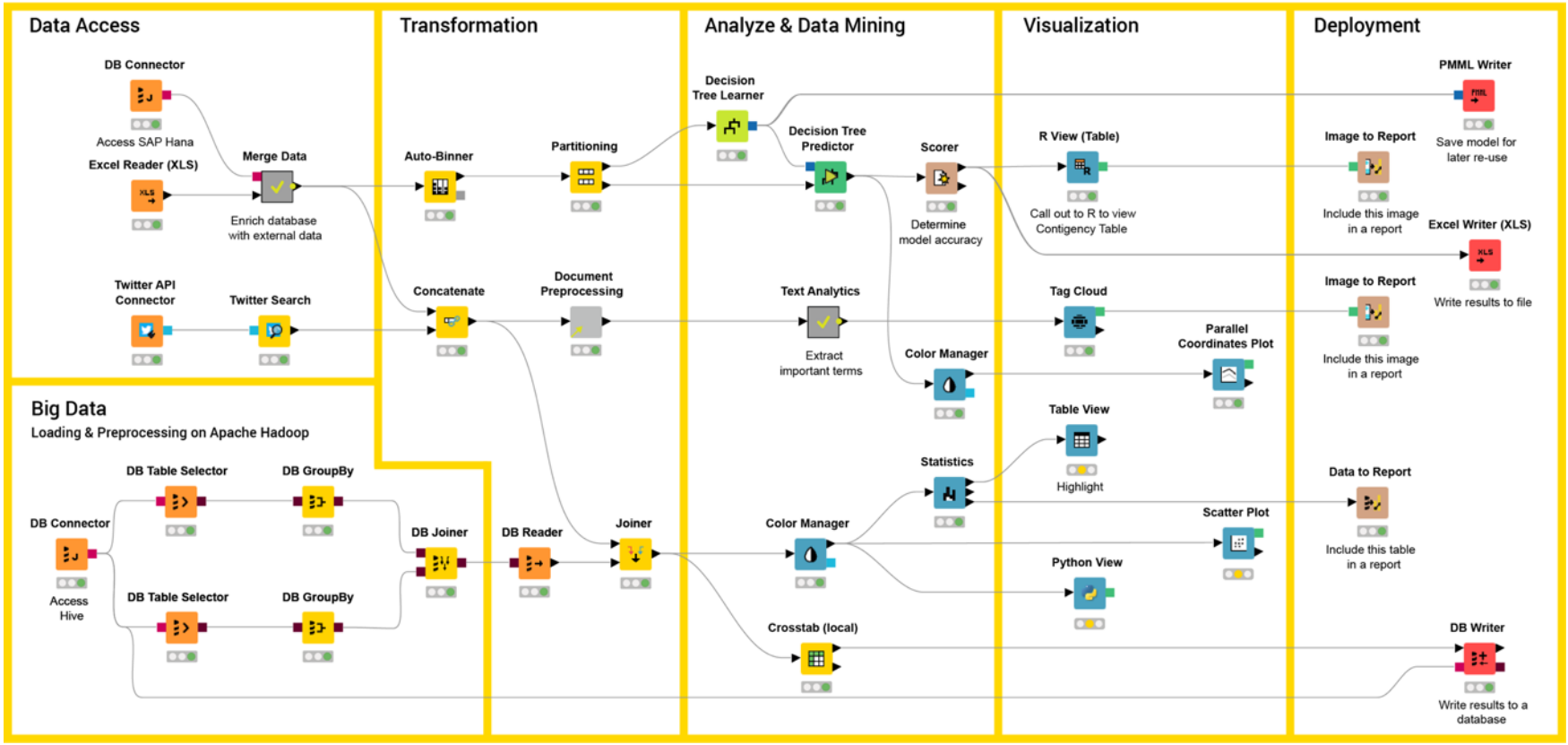
- Right-click executed node
- Select View option in context menu

OR

- Select output port (last item) to inspect execution results

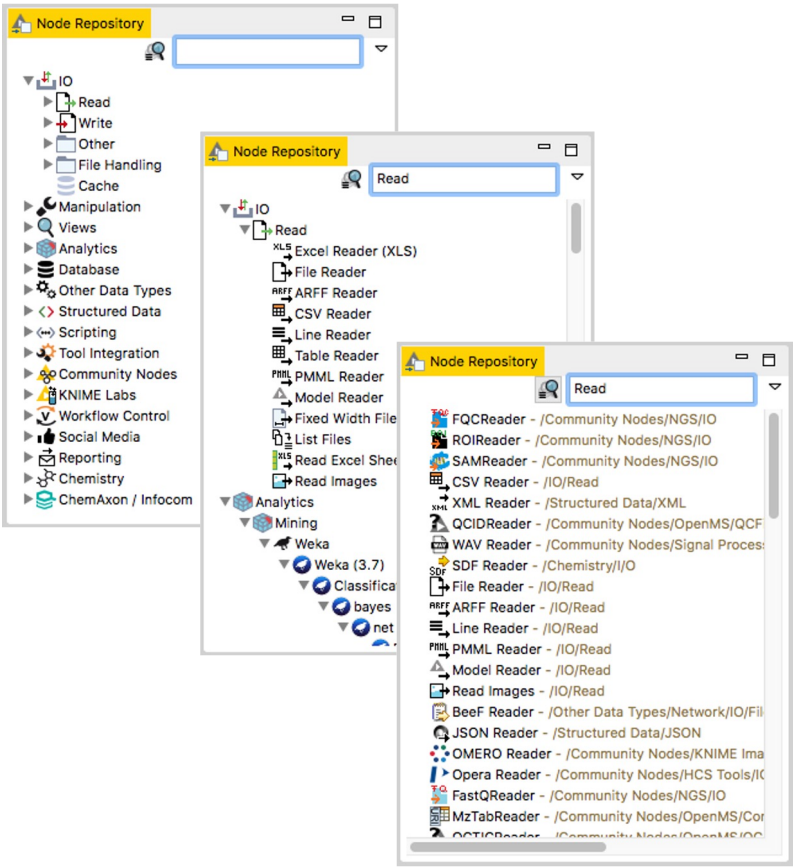


4000+ Nodes for all Steps of End-To-End Data Science



Open for Innovation KNIME

Node Repository



- The Node Repository lists all KNIME nodes
- The search box has 2 modes
  - Standard Search – exact match of node name
  - Fuzzy Search – finds the most similar node name
- Nodes can be added by drag and drop from the Node Repository to the Workflow Editor.

Selected Open Source Extensions for Cheminformatics

- RDKit
  - Converters
  - Modifiers
  - Calculators
    - RDKit Descriptor Calculation
    - RDKit Calculate Charges
  - Geometry
    - RDKit Generate Coords
    - RDKit Optimize Geometry
    - RDKit Add Conformers
    - RDKit Open 3D Alignment
    - RDKit RMSD Filter
  - Fingerprints
    - RDKit Fingerprint
    - RDKit Count-Based Fingerprint
    - RDKit Fingerprint Reader
    - RDKit Fingerprint Writer
    - RDKit Diversity Picker
  - Fragments
  - Searching
  - Reactions
  - Viewing
    - RDKit Interactive Table
    - RDKit SMILES Headers
    - RDKit Molecule Highlighting
  - Experimental

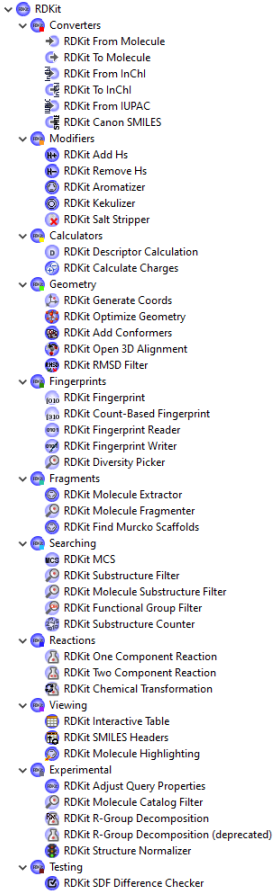
- CDK
  - 3D
    - 3D Coordinates
    - 3D D-Moments
    - 3D D-Similarity
    - 3D RMSD
    - 3D Viewer
    - 3D WHIM
  - AMBIT
  - I/O
  - 2D Coordinates
  - Atom Signatures
  - ChemSpider
  - Connectivity
  - Depiction
  - Element Filter
  - Fingerprint Similarity
  - Fingerprints
  - Hydrogen Manipulator
  - Lipinski's Rule-of-Five
  - Mass Calculator
  - Molecular Properties
  - OPSIN
  - SMARTS Query
  - Structure Sketcher
  - Substructure Search
  - Sugar Remover
  - Sum Formula
  - Symmetry
  - XLogP

- Erlwood Nodes
  - IO
  - Structure Data Format Converters
  - Structure Similarity
    - Fingerprint Similarity
  - Structure Properties
    - Plane of Best Fit Calculator
  - Virtual Screening
    - Virtual Screening Metrics
  - Evaluation and Ranking
    - Desirability Ranking
    - Pareto Ranking
  - SAR Analysis
    - Automated Matched Pairs
    - Free-Wilson Matched Pairs
  - Viewers
    - 2D/3D Scatterplot
    - Activity Cliffs Viewer
    - Similarity Viewer
  - Testing

- Vernalis
  - Databases
  - European PubMed Central
    - European PubMed Central Advanced Search
  - Fingerprints
  - Flow Control
  - IO
  - Matched Molecular Pairs (MMPs)
    - Filtering
    - Fragmentation
    - Pair Generation
    - Rendering
    - Transforms
    - MMP Calculate Maximum Cuts (RDKit)
    - MMP Fragmentation Type Loop Start
    - Uniquify IDs
  - Principal Moments of Inertia (PMIs)
  - RCSB PDB Tools
    - PDB Connector
    - PDB Connector (XML Query)
    - PDB Connector Combine XML Queries
    - PDB Connector Custom Report
    - PDB Connector Query Only
    - PDB Connector Query Only (XML Query)
    - PDB Connector XML Query Builder
    - PDB Describe Heterogens
    - PDB Downloader (Source)
    - PDB SMILES Query
    - PDB Downloader
  - Local PDB Tools
  - Sequence Tools
  - Speedy SMILES
  - Testing
  - Miscellaneous

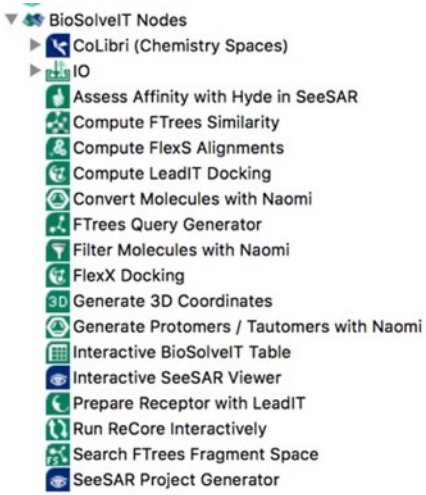
What is RDKit?

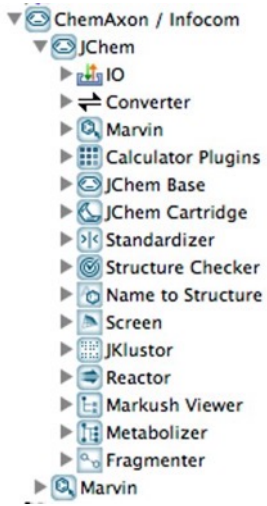
- Open source cheminformatics library in C++
- Wrappers for KNIME maintained by the open source community
- Useful for:
  - Descriptor calculation
  - Cleaning structures
  - InChI conversion
  - Canonical SMILES
  - Fingerprints
  - Scaffolds/substructures
  - Reaction simulation
  - and more…

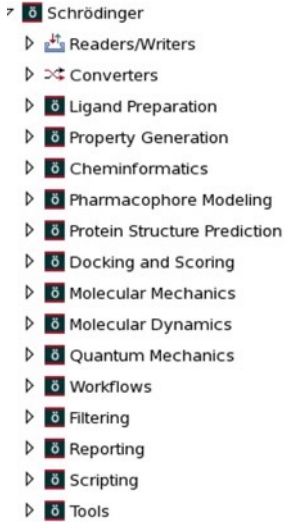


http://www.rdkit.org

Selected Commercial Life Science Extensions





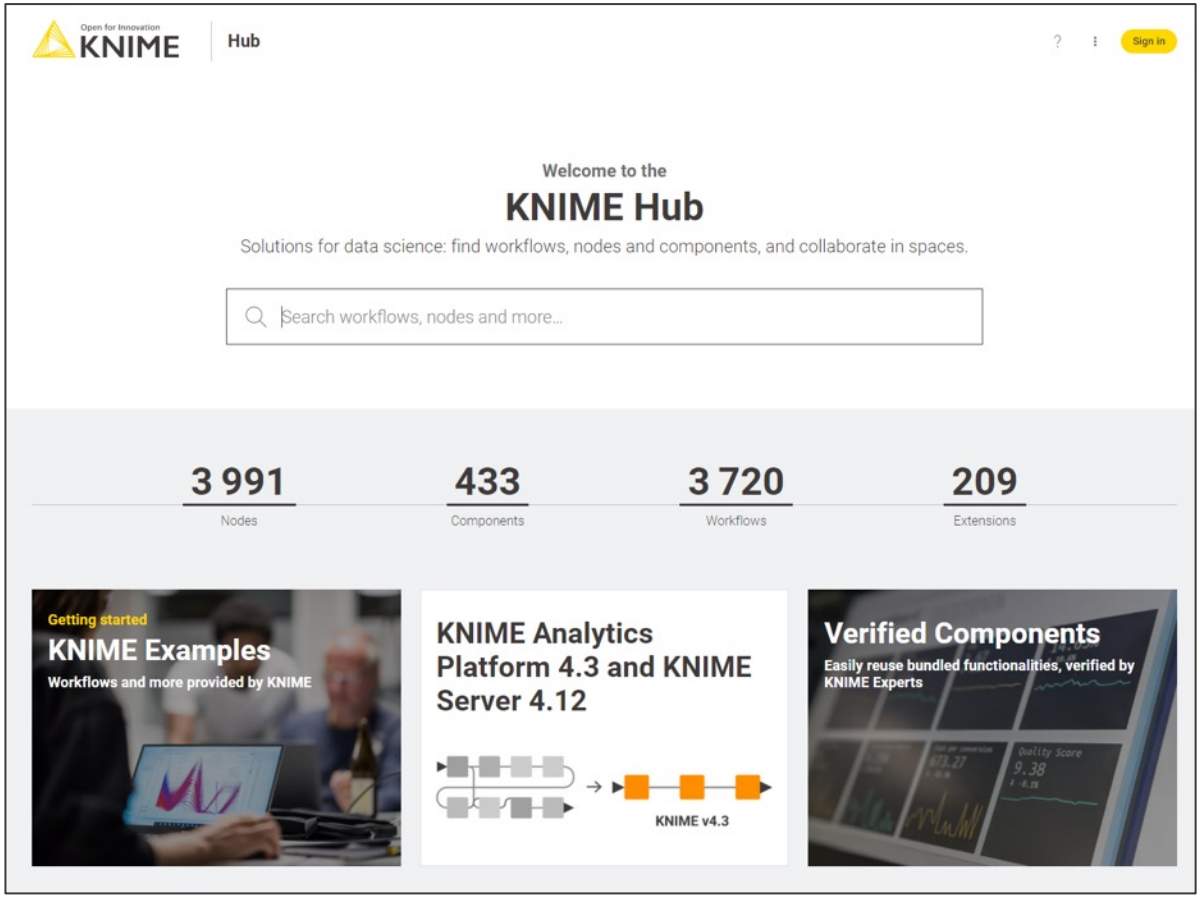


inte:ligand
Your partner for in-silico drug discovery.

Pharmacelera

Further extensions including detailed descriptions can be found at https://hub.knime.com

KNIME Hub: Searching, Sharing, and Collaborating



Additional Resources

KNIME pages (https://www.knime.com)

- Life Sciences landing page https://www.knime.com/why-knime-for-life-science
- RESOURCES **LEARNING HUB** https://www.knime.com/learning-hub
- RESOURCES **HUB** https://hub.knime.com/
- BOOK **WILL THEY BLEND** https://www.knime.com/knimepress/will-they-blend

KNIME Tech pages

FORUM for questions and answers https://forum.knime.com

- **DOCUMENTATION** for docs, FAQ, changelogs, ... https://docs.knime.com/
- **COMMUNITY CONTRIBUTIONS** for dev instructions and third party nodes https://www.knime.com/community

KNIME TV on **YouTube** https://www.youtube.com/user/KNIMETV

Chemistry in KNIME

Reminder of the task

- You are a computational chemist in a project
- Your task is to develop a model to estimate LogD
- How and where do you start?
- Previous model is build on the inhouse data (quite some time ago)
- You have collected some public data
- And there is a fresh set from the project
- You would like to harmonize the data: get rid of redunces, standardize chemical structures, remove duplicates
- You would like to develop an interactive visualization of the dataset to be able to explore the dataset, filter the data based on the insights, discuss the project data with the coleagues

Data sets

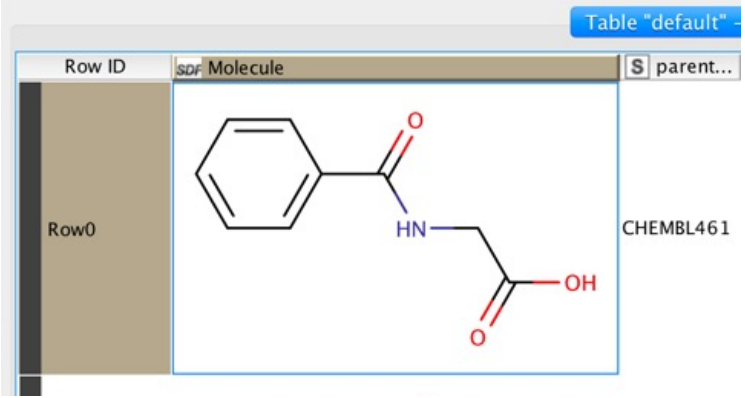
- Inhouse:
  - CHEMBL3301363: ASTRAZENECA: Octan-1-ol/water (pH7.4) distribution coefficent measured by a shake flask method described in J. Biomol. Screen. 2011, 16, 348-355. Experimental range -1.5 to 4.5
  - 4200 molecules
  - CSV file with experimental data and SMILES
- Public:
  - PubChem AID 686912: GSK_TB: GSK in-house hydrophobicity assay at pH7.4 (logD)
  - 148 molecules
  - CSV file with experimental data
  - SD file with structural data
- New inhouse:
  - CHEMBL851899: Partition coefficient (logD)
  - 11 molecules
  - XLS file with experimental data and SMILES

Overview of Types in KNIME

- Basic KNIME data types
  - String, integer, double

- KNIME core chemistry data types:
  - SMILES, SDF, mol, mol2
  - Structures in these formats can be rendered in KNIME tables

- Others
  - Image
  - Model
  - Connection
  - json/xml
  - …

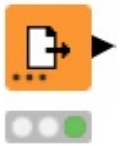


File Reader (Complex Format)

- Workhorse of KNIME Source nodes
- Reads text files
- Many advanced features allow it to read most ‘weird’ files
  - Short lines, inline comments, headers, and special encoding
  - **Distinguishes SMILES and SMARTS formats**

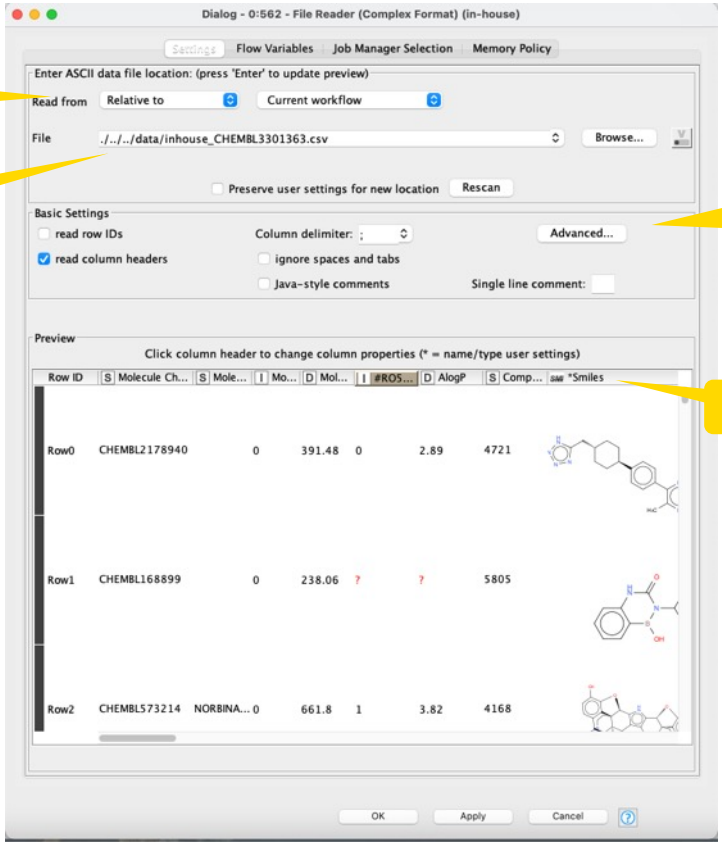
File Reader (Complex Format) Configuration

File Reader
(Complex Format)



File system

File path

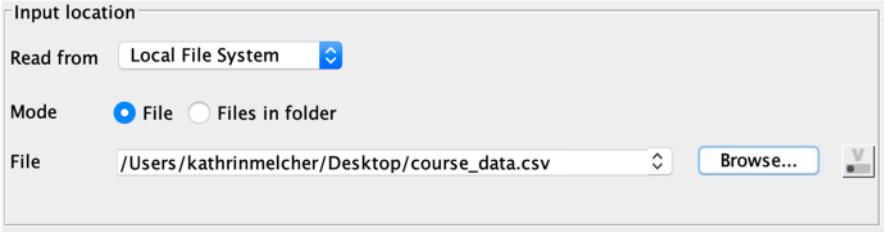


Advanced settings

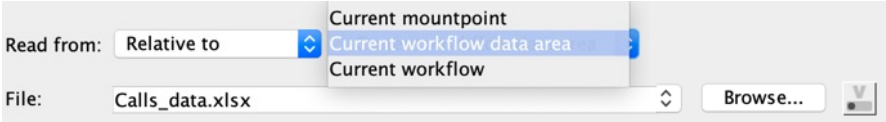
Column type

Common Settings: Four Default File Systems

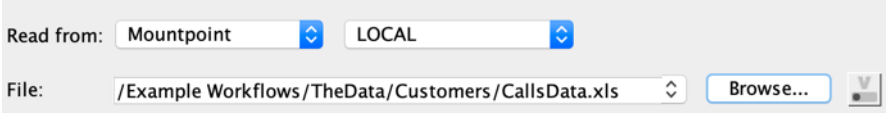
- ## Local File System



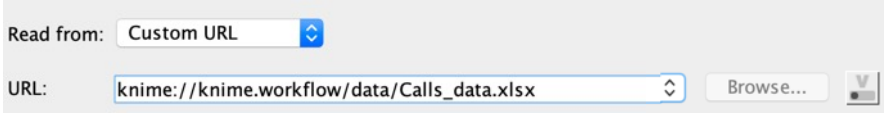
- ## Relative to …



- ## Mountpoint

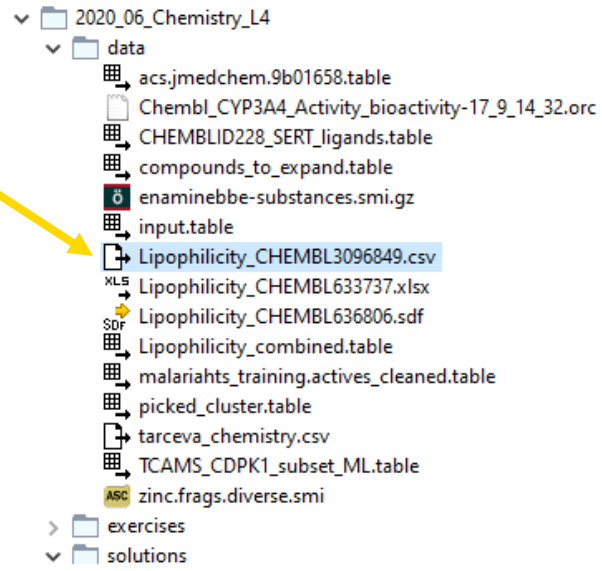


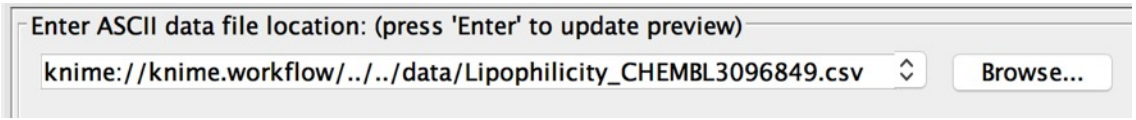
- ## Custom URL



Workflow-Relative File Paths

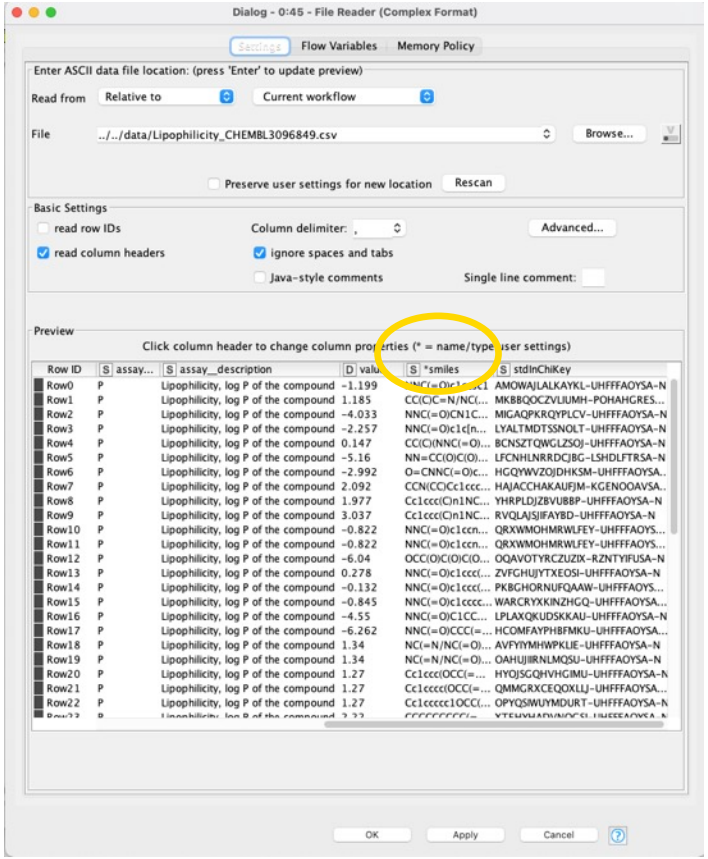
- Best choice if workflows are to be shared
- Requires matching folder structure within workflow group
  - Independent of environment outside of workflow group
- Example: Path to „Lipophilicity_CHEMBL3096849.csv“
  - Local path: /User/Name/knime-workspace/2020_06_Chemistry_L4/data/Lipophilicity_CHEMBL3096849.csv
  - Workflow relative:

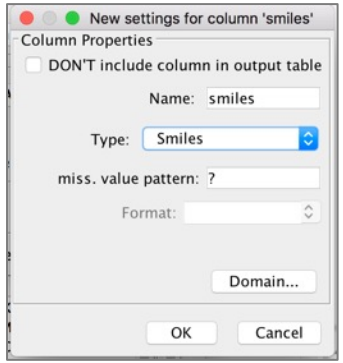


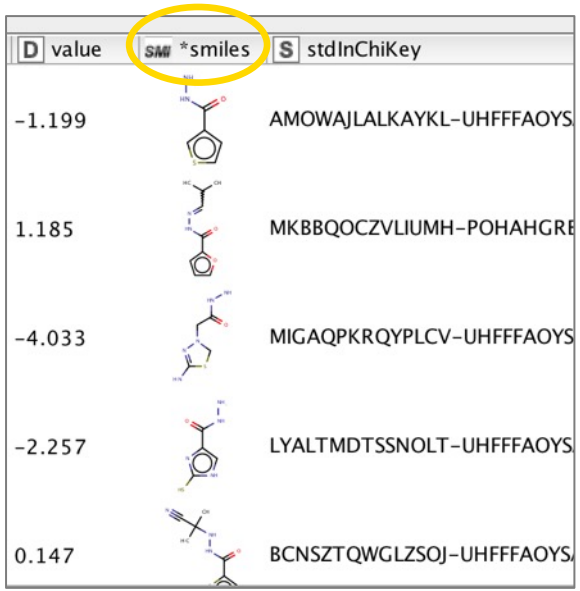


YouTube KNIME TV Channel: https://youtu.be/U9sP4g4yGwY

File Reader (Complex Format) Configuration



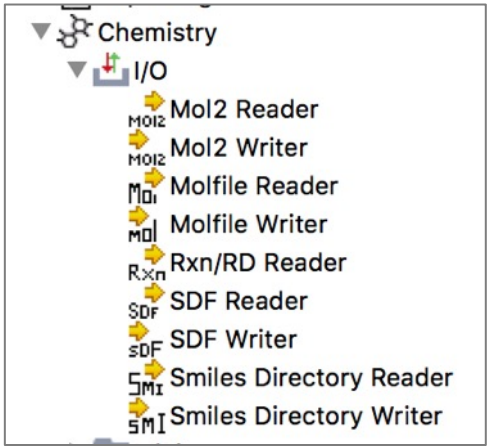


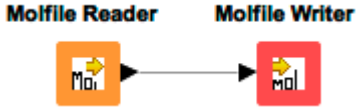


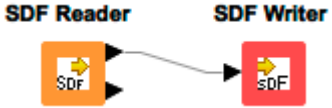
Nodes for Reading and Writing files

Reader and writers provided for:

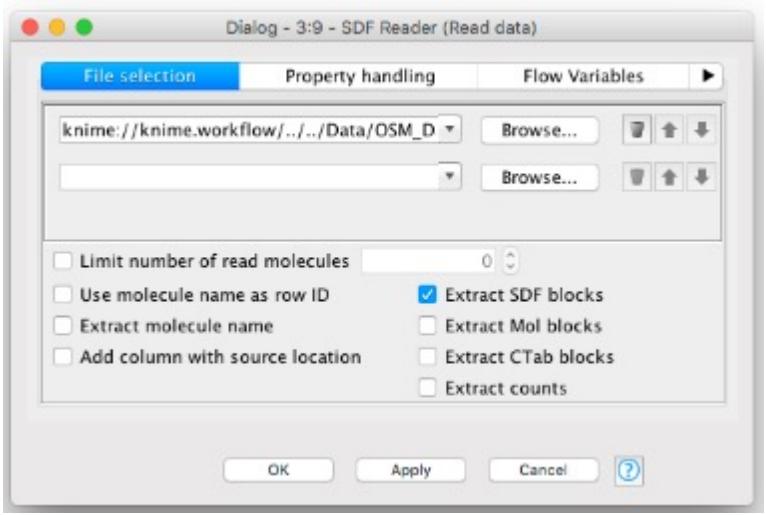
- SDF, SMILES, mol, mol2

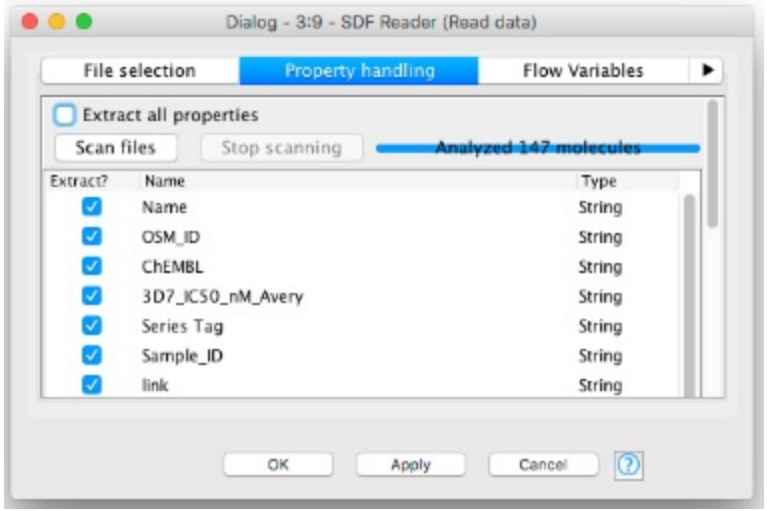




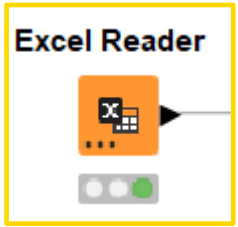


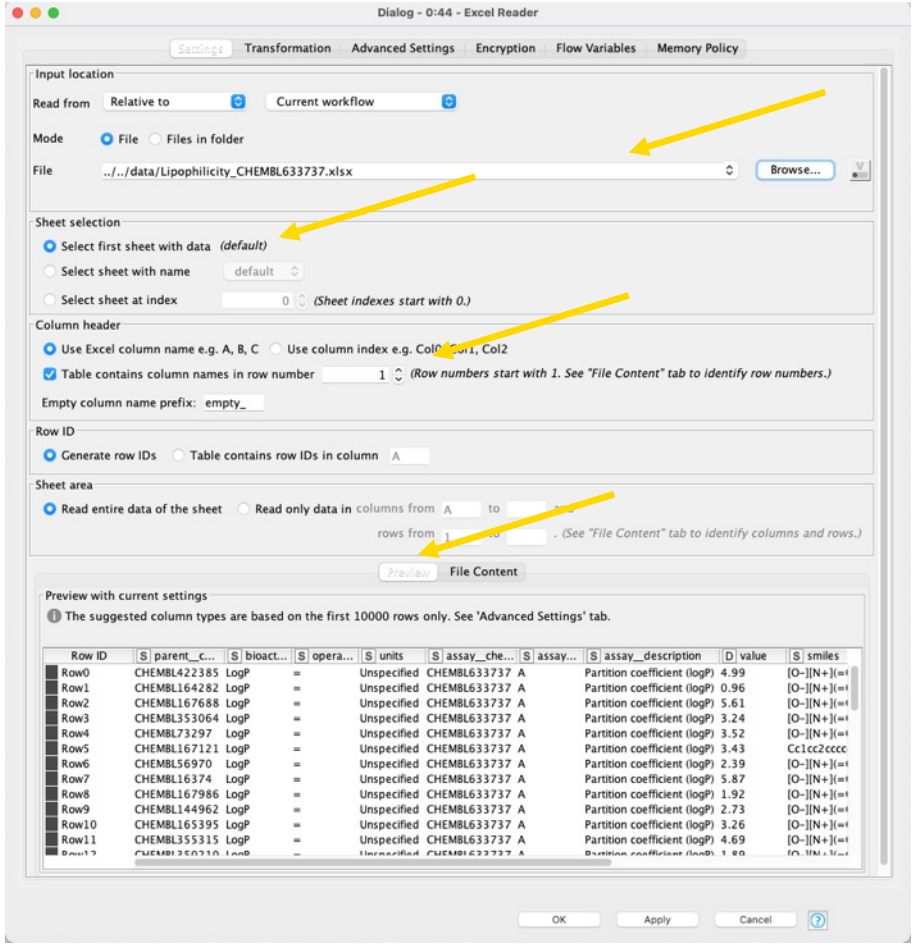
A bit more about reading SD files





Excel Reader





Reminder of the task

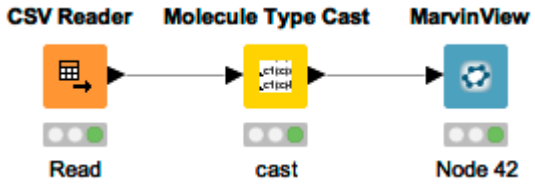
- You are a computational chemist in a project
- Your task is to develop a model to estimate LogD
- How and where do you start?
- Previous model is build on the inhouse data (quite some time ago)
- You have collected some public data
- And there is a fresh set from the project
- You would like to harmonize the data: get rid of redunces, standardize chemical structures, remove duplicates
- You would like to develop an interactive visualization of the dataset to be able to explore the dataset, filter the data based on the insights, discuss the project data with the coleagues

Tedious repetative preprocessing tasks

- Adding columns with constant values
- Renaming columns
- Casting strings to SMILES
- Joining data based on identifier
- Transforming SMILES/SDF to RDKit molecule

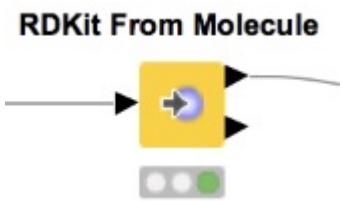
Nodes for Type Manipulation

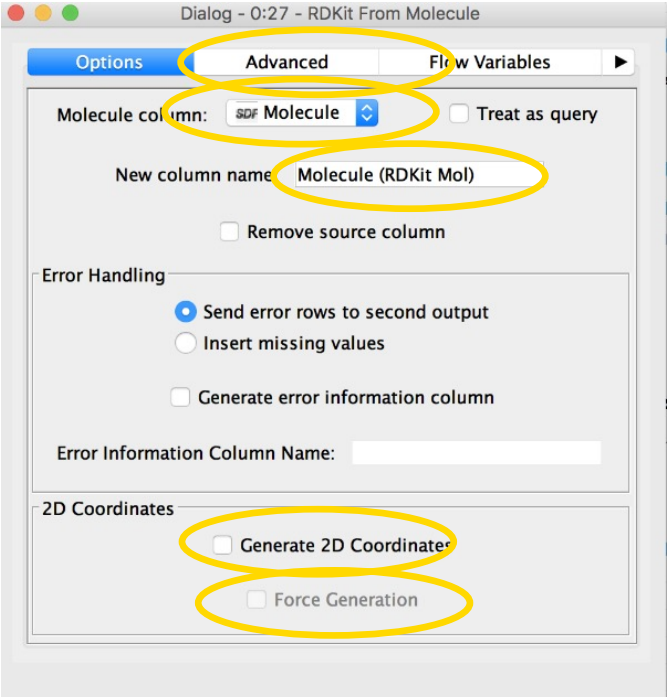
- Molecule Type Cast
  - Casts any string as a chemical type (i.e. It tells KNIME “This is a smiles string”)
  - Useful when reading data from a csv file or database.



RDKit From Molecule

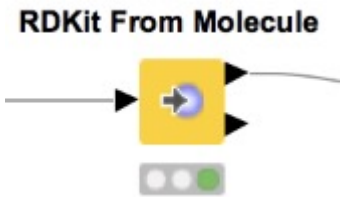
Generates RDKit molecule column from a molecule string representation (SMILES, SDF or SMARTS)

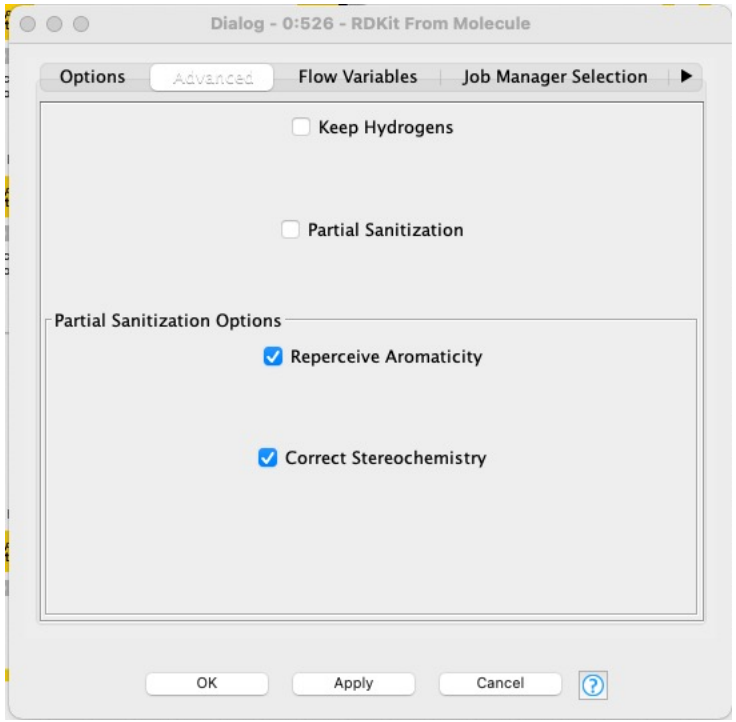




RDKit From Molecule

Generates RDKit molecule column from a molecule string representation (SMILES, SDF or SMARTS)





Joining Columns of Data

Left Table

Mol Reg No	Chembl ID	SMILES
22	CHEMBL1794855	CCCN(CCC)
24	CHEMBL278751	CCN(C)
15	CHEMBL103772	CCCN1CC
10	CHEMBL328107	C1CN(CCN1)

Join by Mol Reg No

Right Table

Mol Reg No	Ki value	Ki relation	Ki unit
17	76.0	=	nM
65	6.56	=	nM
35	100	>	nM
15	8	=	nM
10	95.8	=	nM

Inner Join

Mol Reg No	Chembl ID	SMILES	Ki value	Ki relation	Ki unit
15	CHEMBL103772	CCCN1CC	8	=	nM
10	CHEMBL328107	C1CN(CCN1)	95.8	=	nM

Left Outer Join

Mol Reg No	Chembl ID	SMILES	Ki value	Ki relation	Ki unit
22	CHEMBL1794855	CCCN(CCC)	?	?	?
24	CHEMBL278751	CCN(C)	?	?	?
15	CHEMBL103772	CCCN1CC	8	=	nM
10	CHEMBL328107	C1CN(CCN1)	95.8	=	nM

Right Outer Join

Mol Reg No	Chembl ID	SMILES	Ki value	Ki relation	Ki unit
17	?	?	76.0	=	nM
65	?	?	6.56	=	nM
35	?	?	100	>	nM
15	CHEMBL103772	CCCN1CC	8	=	nM
10	CHEMBL328107	C1CN(CCN1)	95.8	=	nM

Joining Columns of Data

Left Table

Mol Reg No	Chembl ID	SMILES
22	CHEMBL1794855	CCCN(CCC)
24	CHEMBL278751	CCN(C)
15	CHEMBL103772	CCCN1CC
10	CHEMBL328107	C1CN(CCN1)

Join by Mol Reg No

Full Outer Join

Right Table

Mol Reg No	Ki value	Ki relation	Ki unit
17	76.0	=	nM
65	6.56	=	nM
35	100	>	nM
15	8	=	nM
10	95.8	=	nM

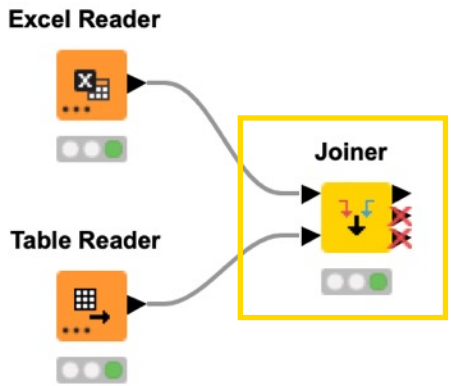
Missing values in the left table

Mol Reg No	Chembl ID	SMILES	Ki value	Ki relation	Ki unit
17	?	?	76.0	=	nM
65	?	?	6.56	=	nM
35	?	?	100	>	nM
15	CHEMBL1794855	CCCN(CCC)	8	=	nM
10	CHEMBL278751	CCN(C)	95.8	=	nM
22	CHEMBL103772	CCCN1CC	?	?	?
24	CHEMBL328107	C1CN(CCN1)	?	?	?

Missing values in the right table

Joiner

- Combines columns from two different tables
  - Top input port: “Left” data table
  - Bottom input port: “Right” data table



- Outputs:
  - Top port: Resulting joined table
  - Middle port: Unmatched rows from the left input table (top input port)
  - Bottom port: Unmatched rows from the right input table (bottom input port)

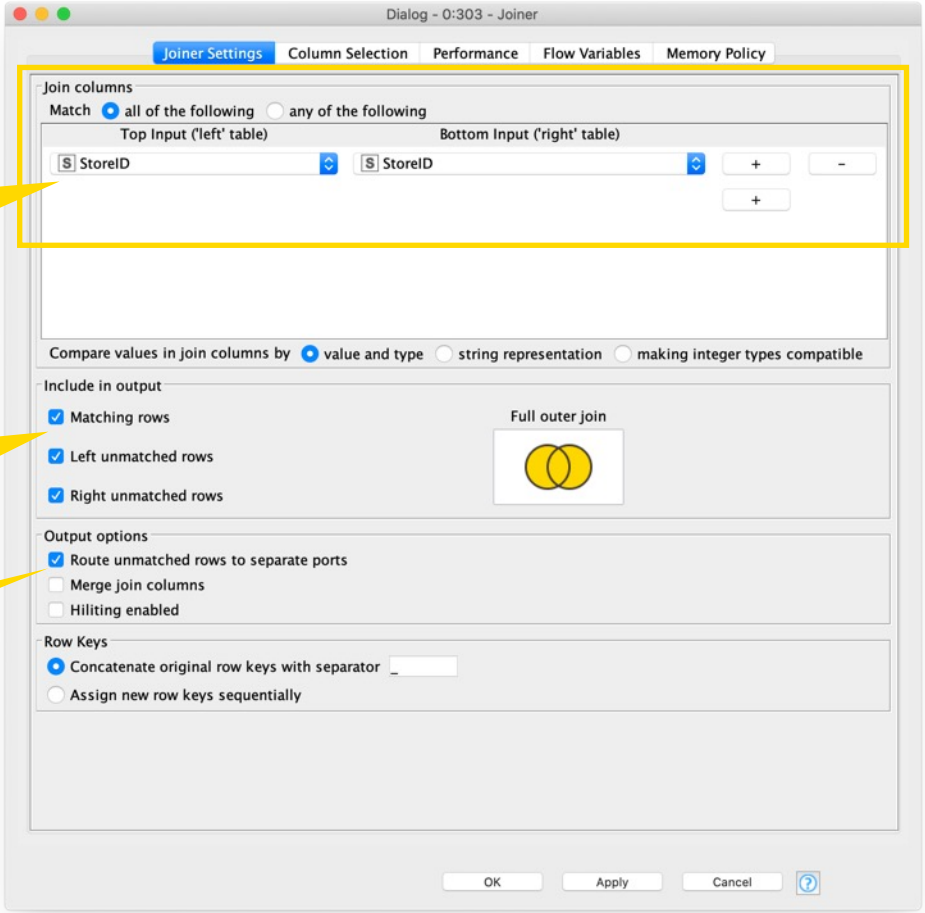
- By default the two bottom output ports are deactivated

Joiner Configuration – Linking Rows

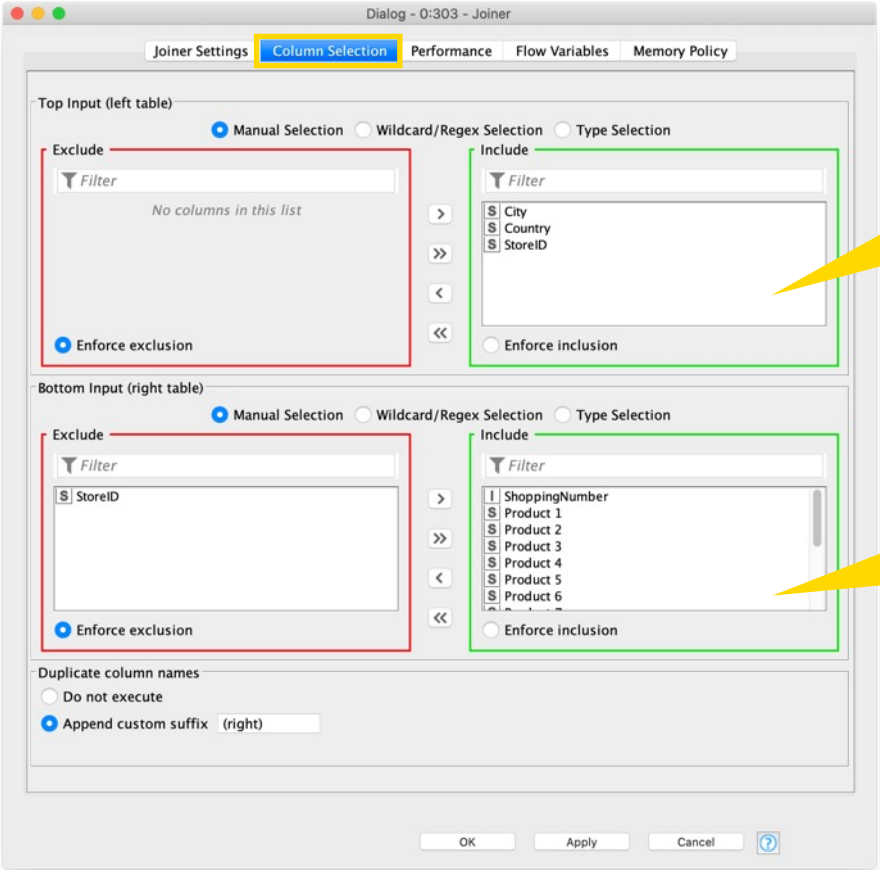
Values to join on. Multiple joining columns are allowed

Select the rows which should be included in the joined table

Activate this checkbox to activate the bottom output ports



Joiner Configuration – Column Selection

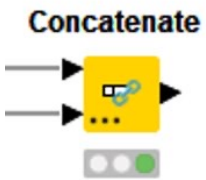


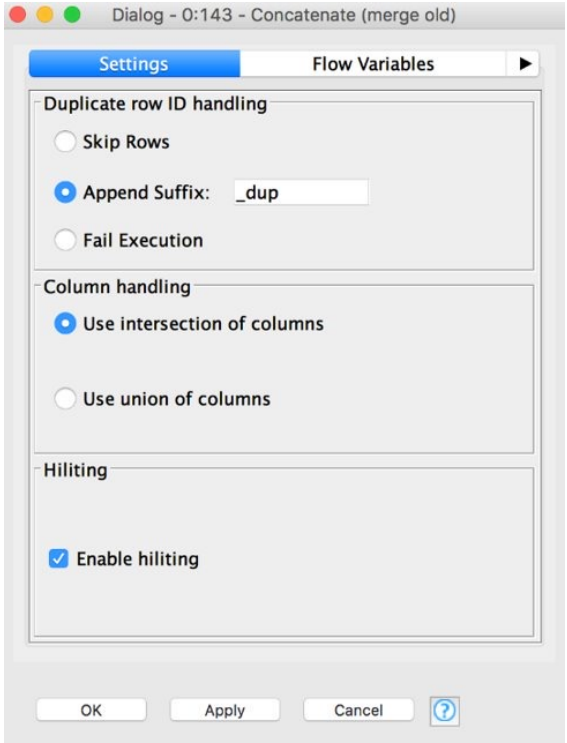
Columns from top table for joined table

Columns from lower table for joined table

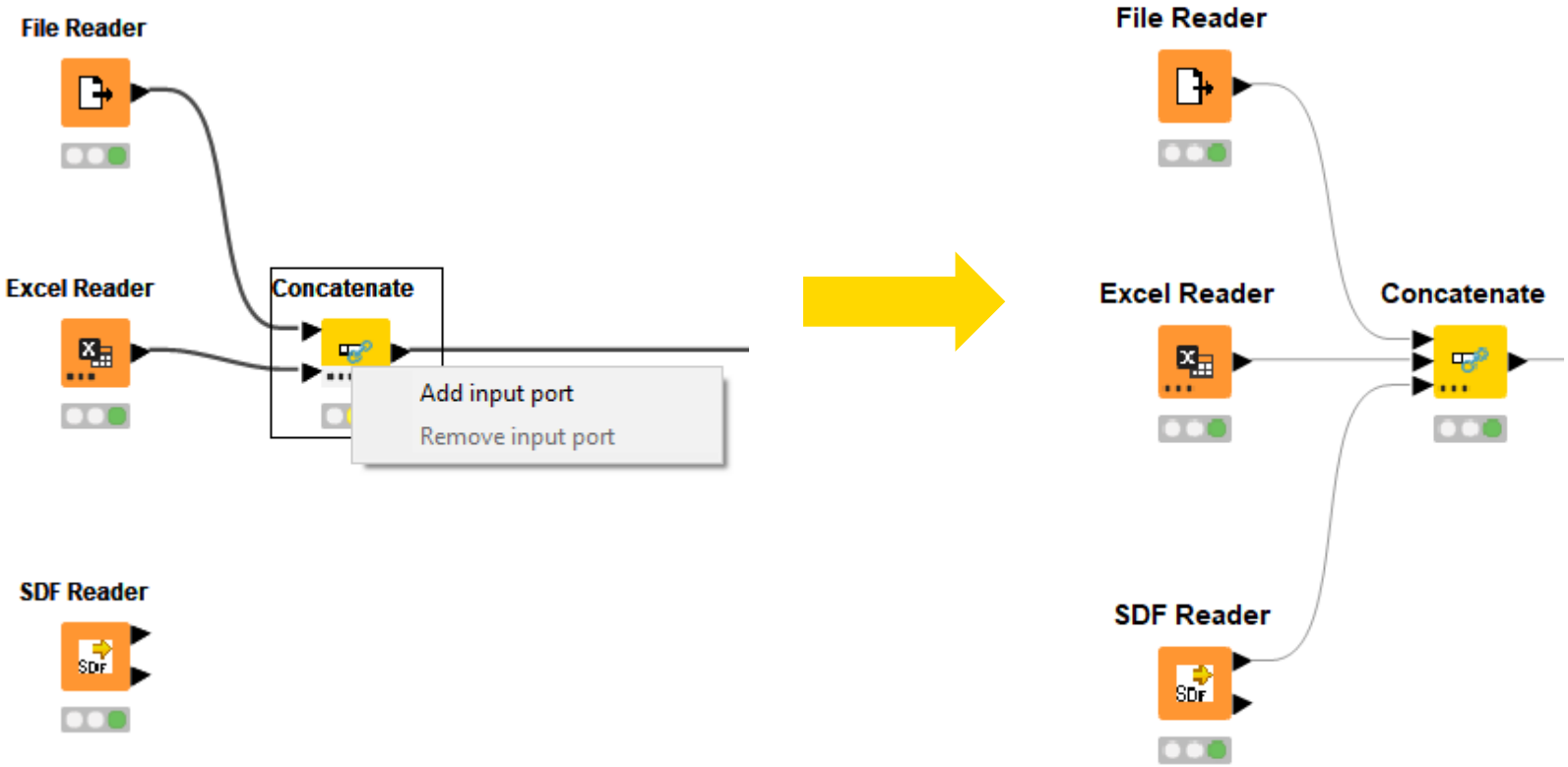
Concatenate

Combine rows from two or more tables with shared columns



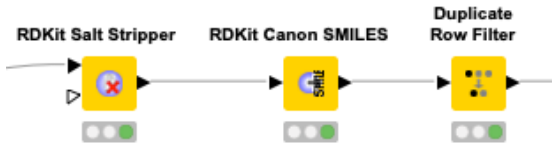


Dynamic Ports



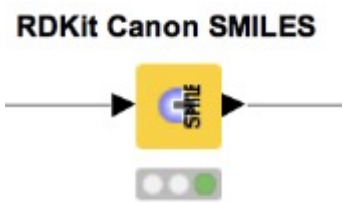
[Some] standardization

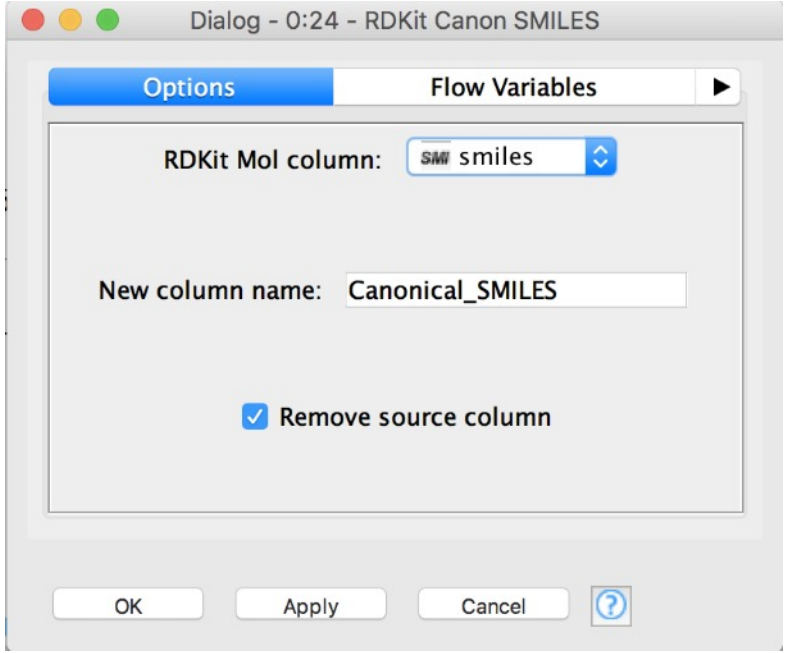
- Removing salts
- Generating canonical SMILES
- Removing duplicates



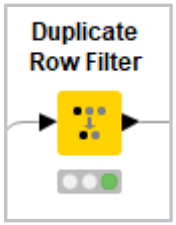
Standardization

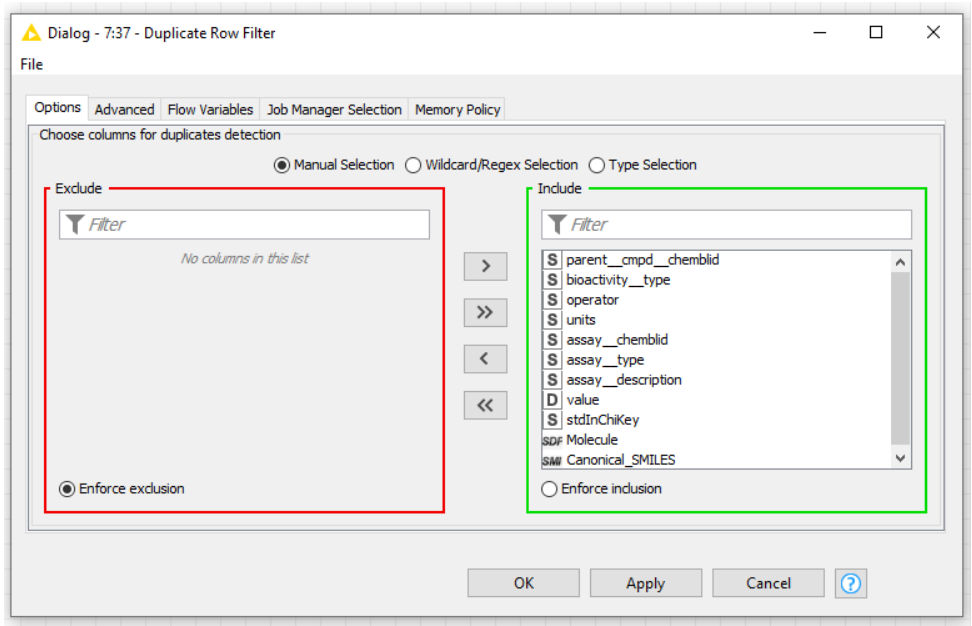
Generate canonical SMILES





Remove Duplicates

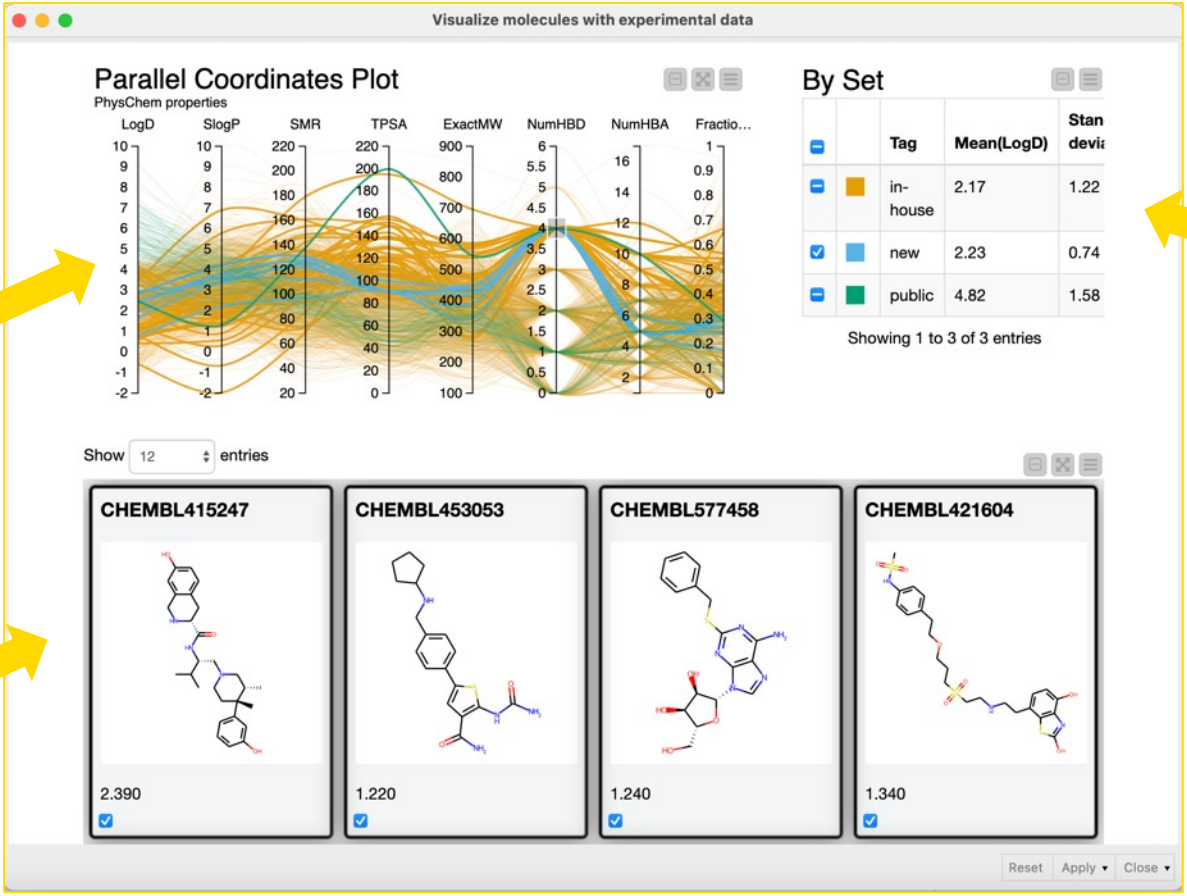




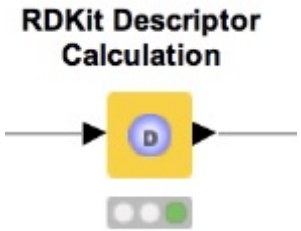
Reminder of the task

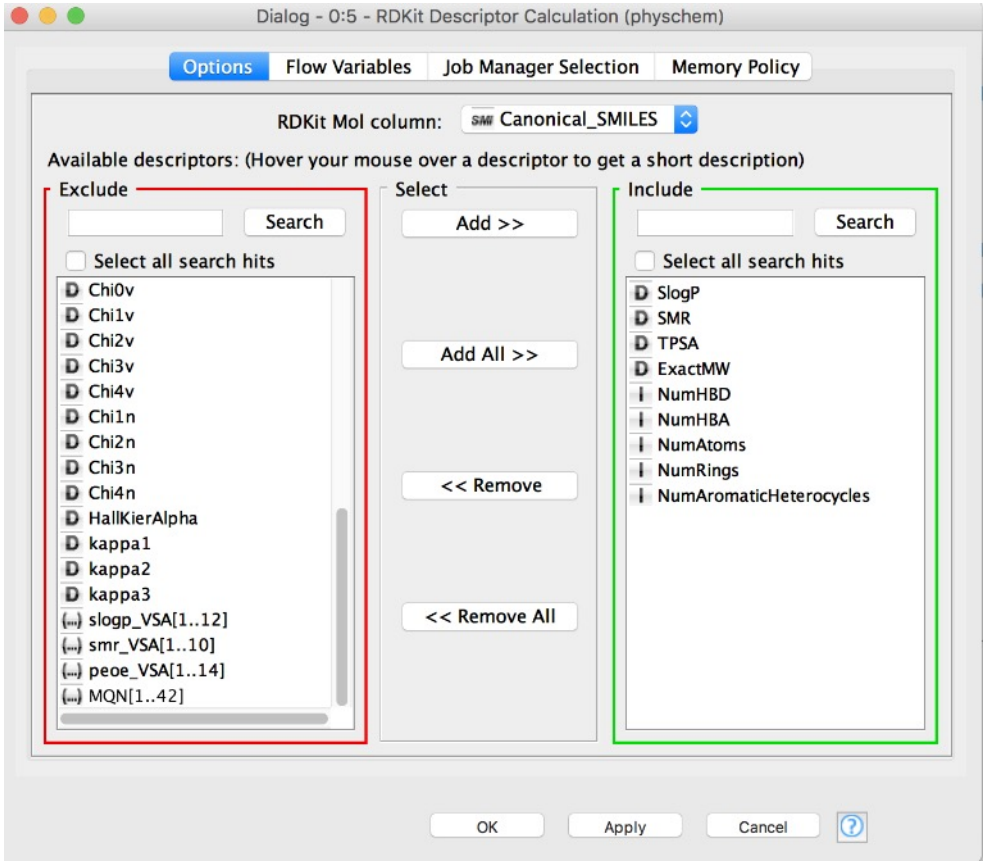
- You are a computational chemist in a project
- Your task is to develop a model to estimate LogD
- How and where do you start?
- Previous model is build on the inhouse data (quite some time ago)
- You have collected some public data
- And there is a fresh set from the project
- You would like to harmonize the data: get rid of redunces, standardize chemical structures, remove duplicates
- You would like to develop an interactive visualization of the dataset to be able to explore the dataset, filter the data based on the insights, discuss the project data with the coleagues

Reminder of a possible interactive view

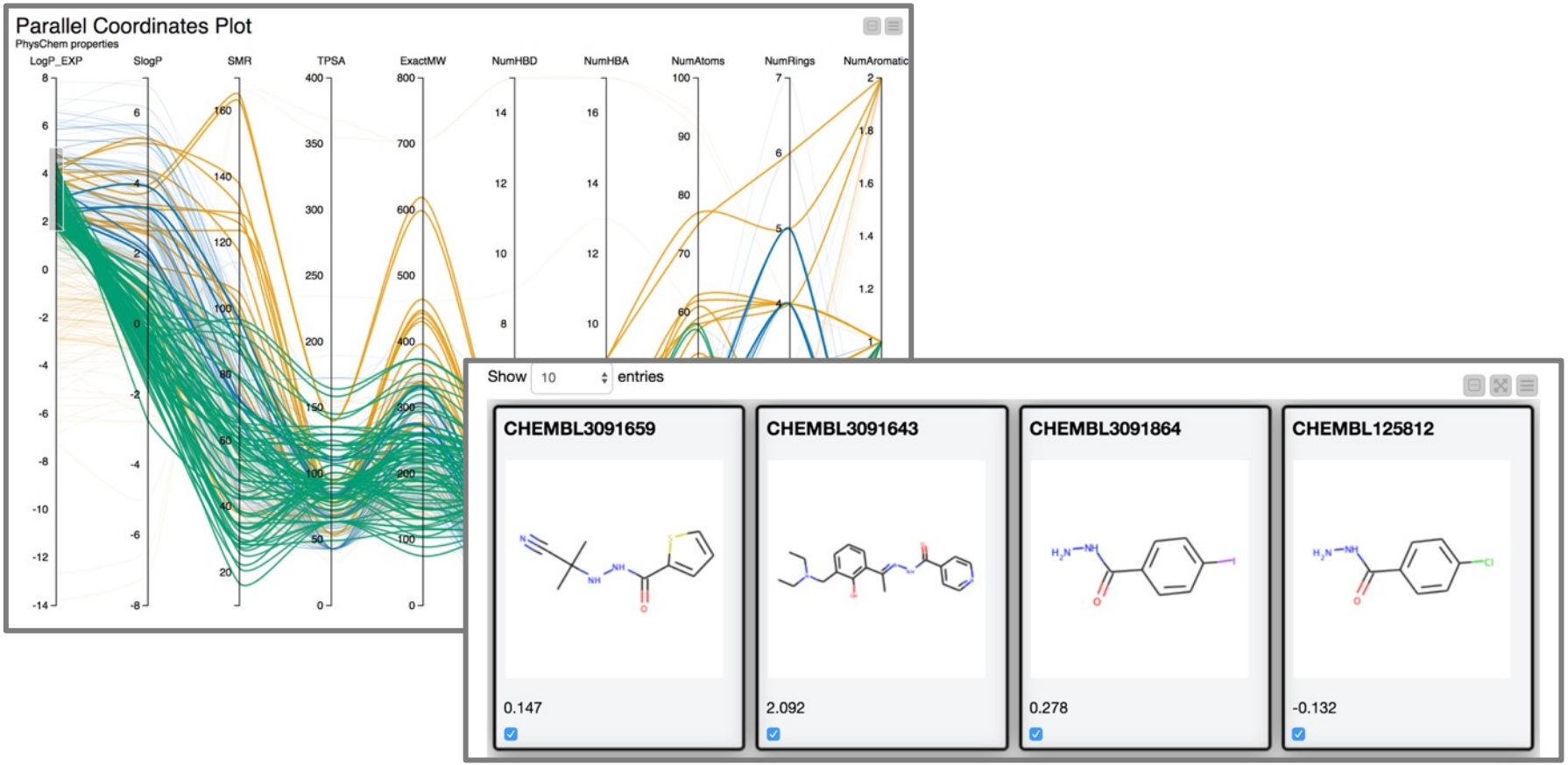


Compute Descriptors





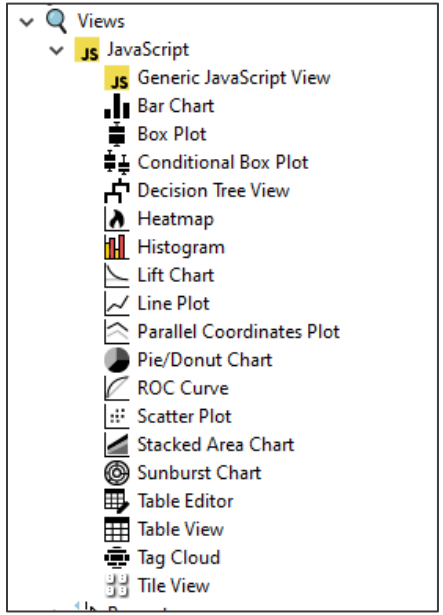
Visualize Chemical Structures Together with other Views

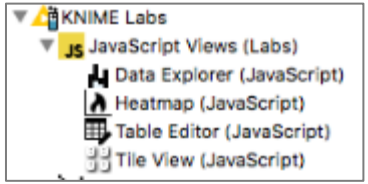


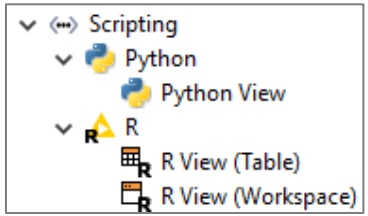
Data Visualization

- Large selection of easy to use visualization nodes
  - Web-based and interactive
  - Dedicated nodes, no scripting required

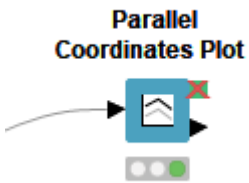
- R and Python View nodes for highly customizable graphics
  - Require scripting

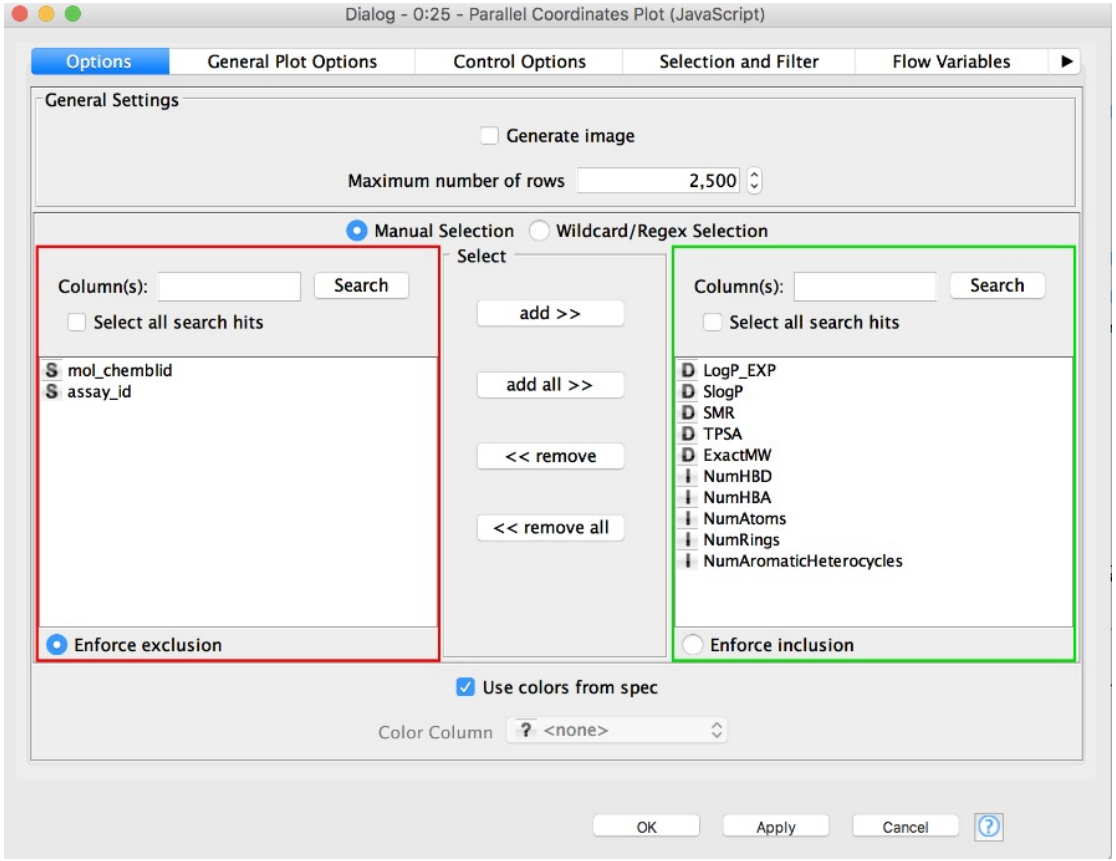






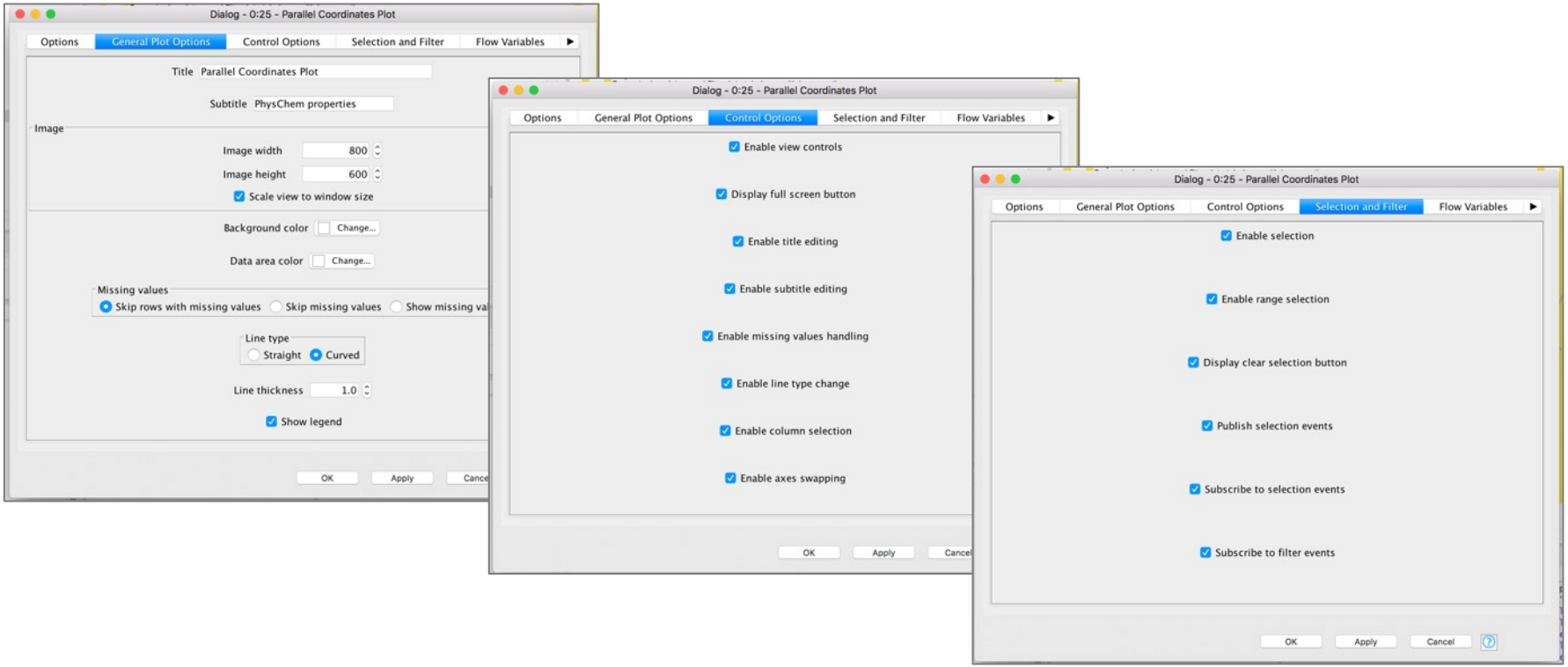
Parallel Coordinates Plot



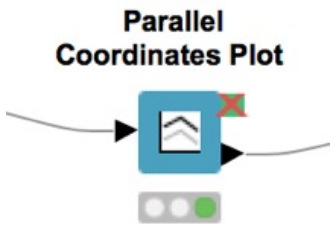


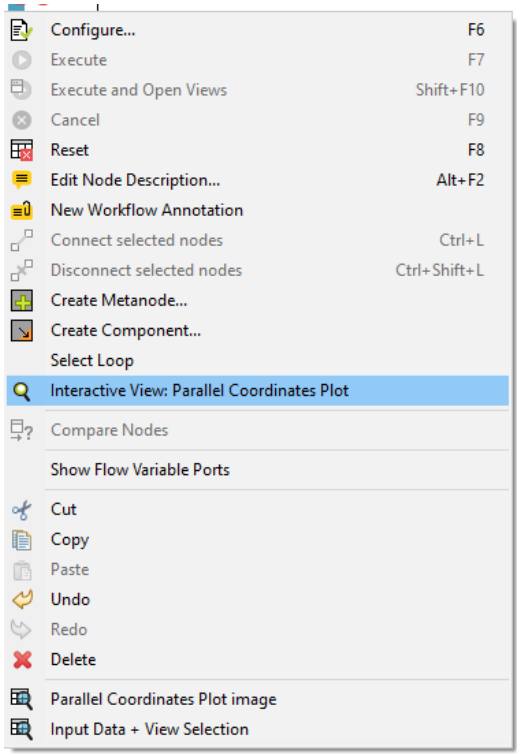
Parallel Coordinates Plot

- Additional configuration tabs

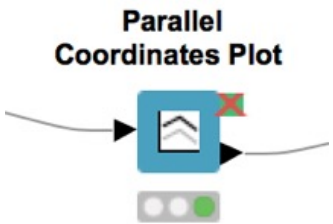


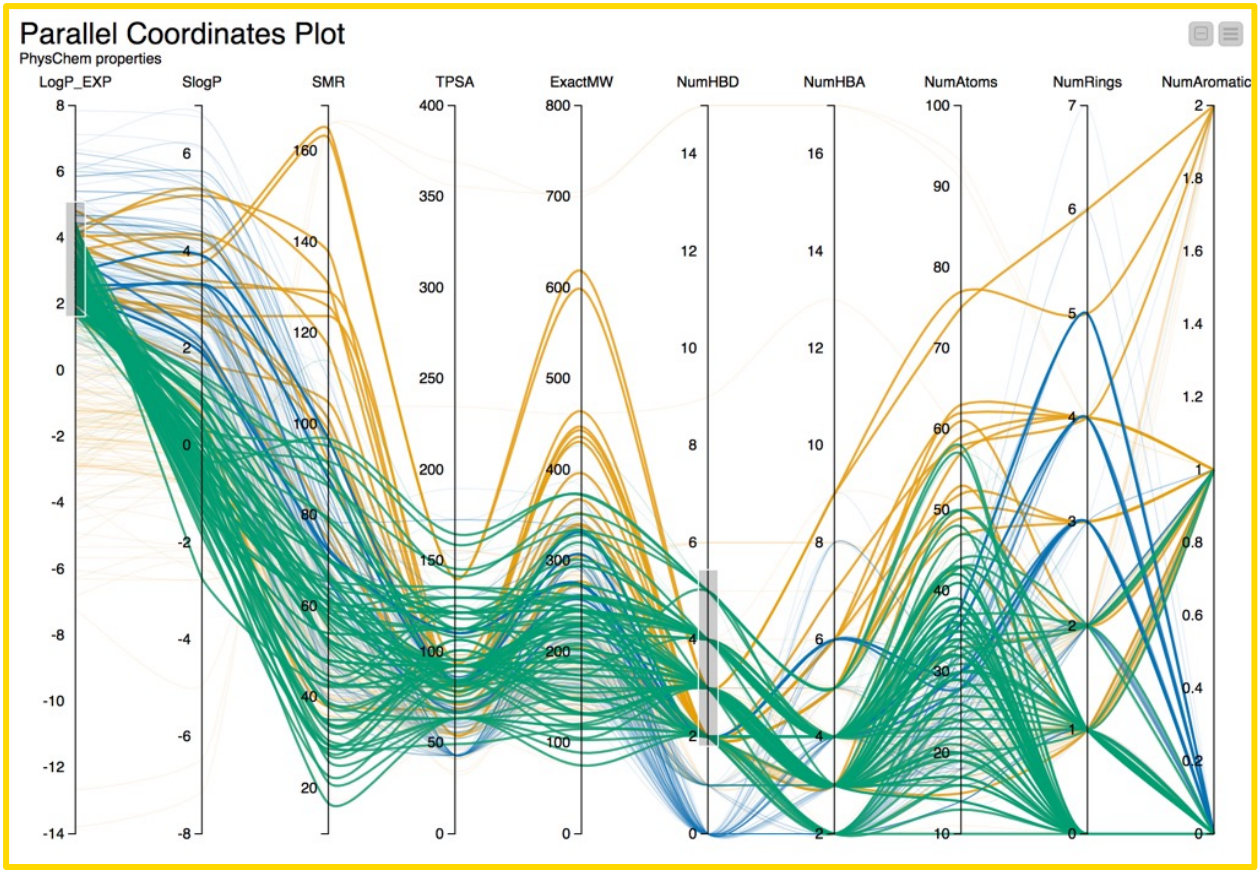
Parallel Coordinates Plot



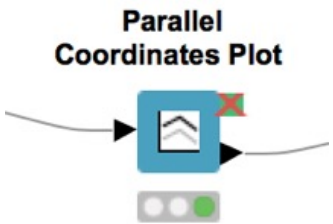


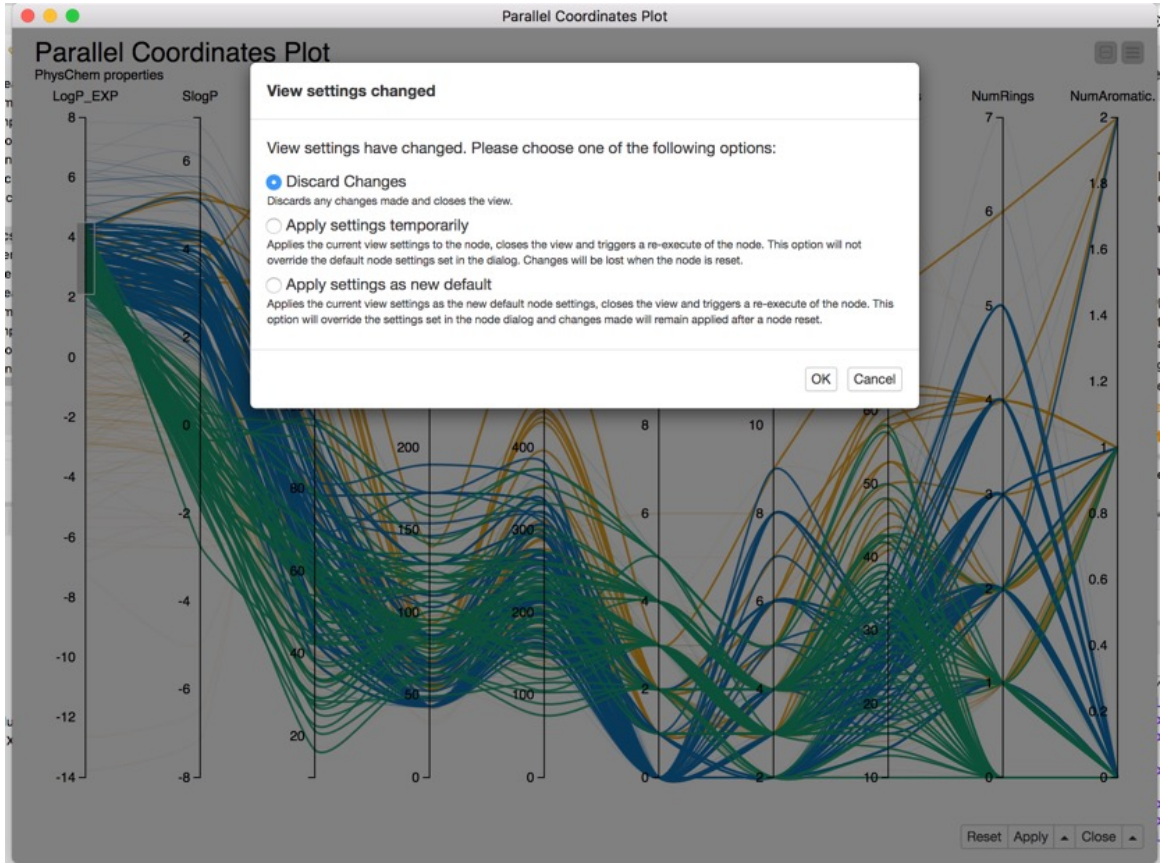
Interactively Filter on Multiple Properties





Parallel Coordinates Plot

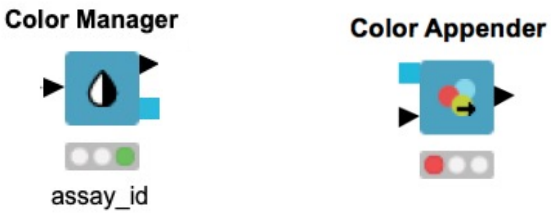


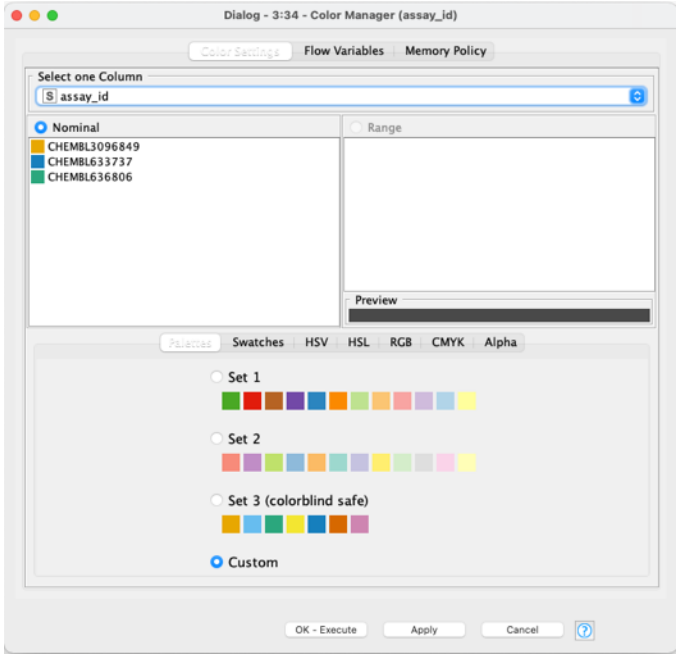


Color Manager

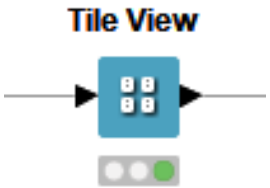
One of several visual property managers (e.g. size, shape)

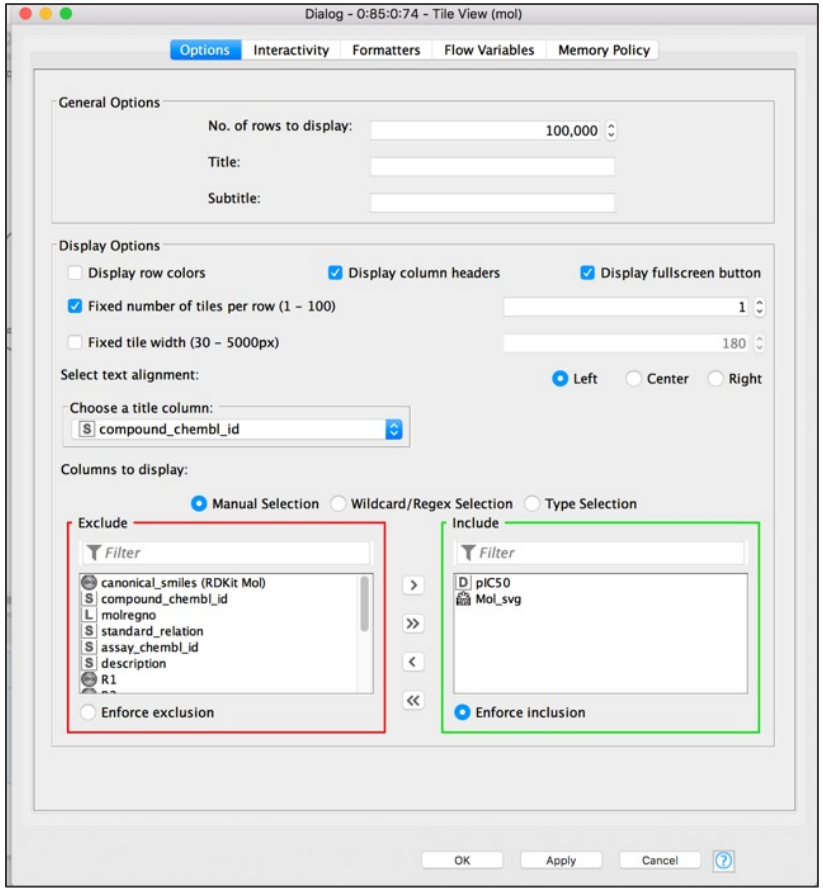
- Color by nominal or continuous values
- Sync colors between views using the color model port

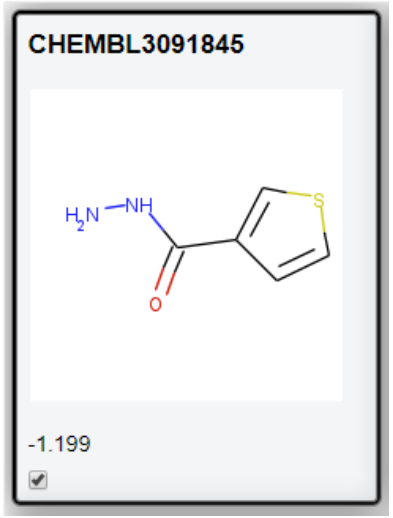




Tile View

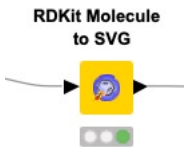


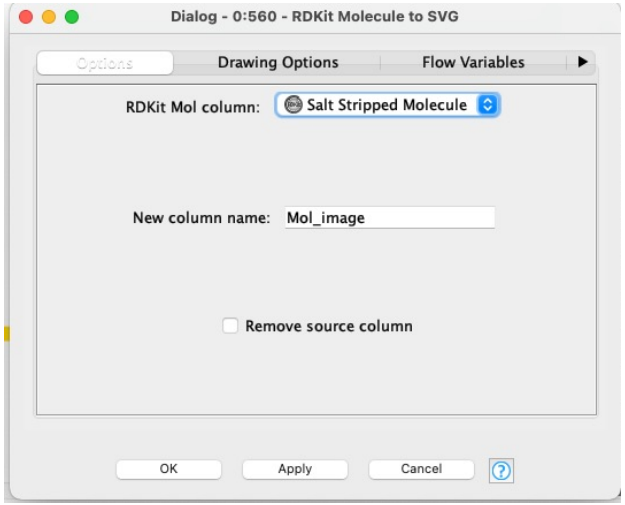


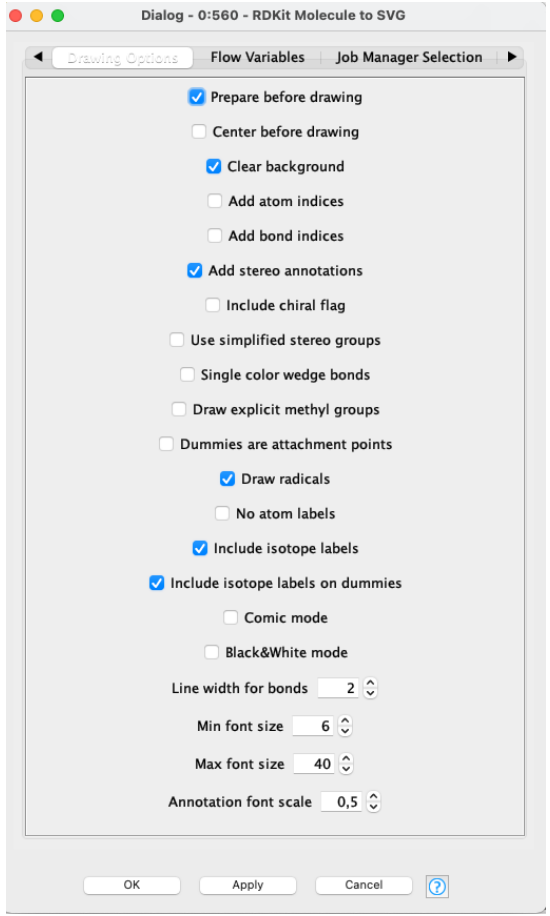


RDKit molecule to SVG

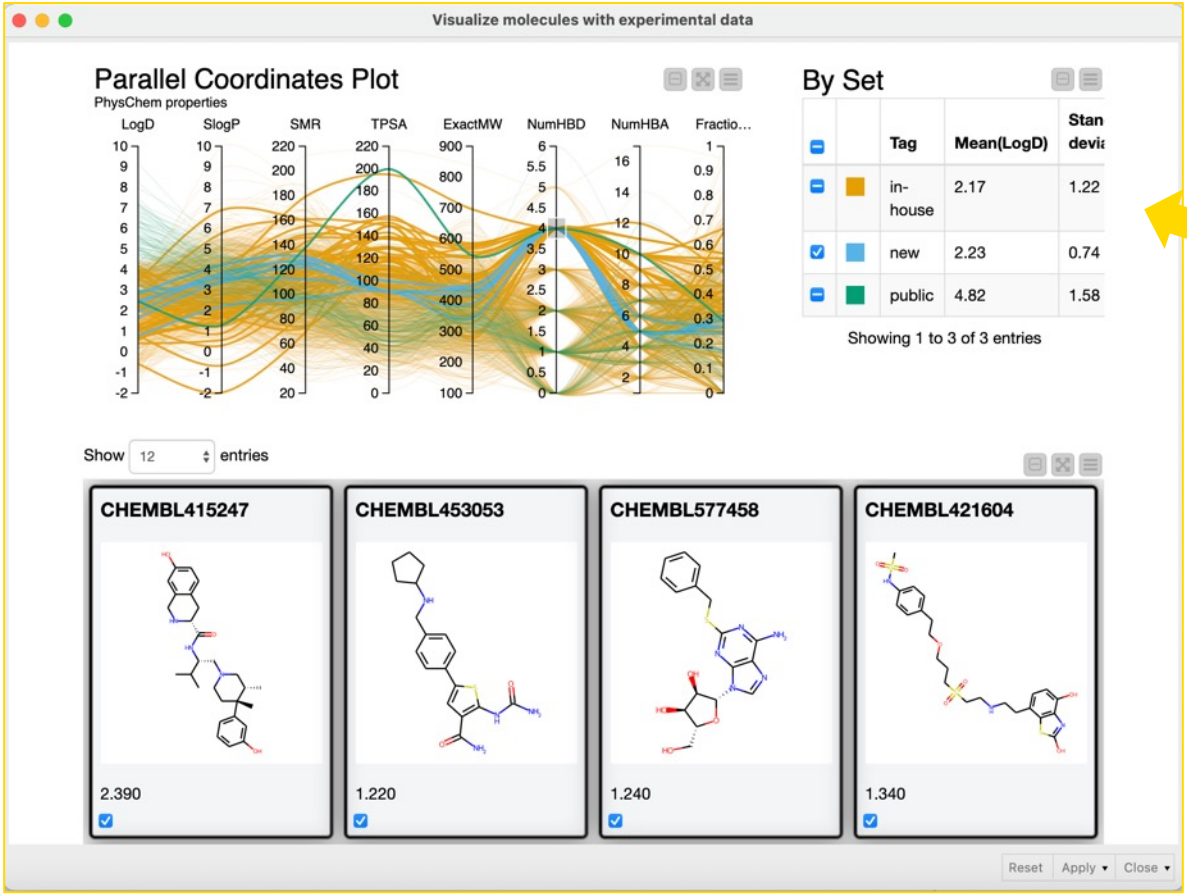
- Generate an SVG image for RDKit molecules







Reminder of a possible view



Data Aggregation (GroupBy)

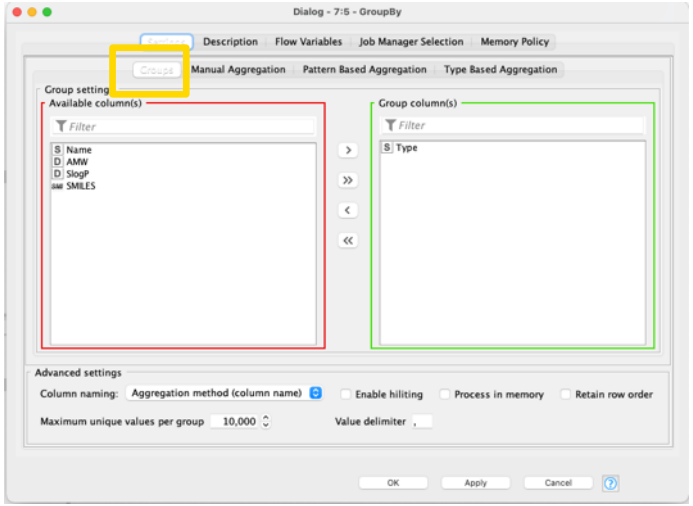
Type	Name	Weigt
NSAID	paracetamol	151.17
NSAID	aspirin	180.16
NSAID	ibuprofen	206.29
NSAID	diclofenac	296.15
PPI	omeprazole	345.42
PPI	pantoprazole	383.38
SSRI	fluoxetine	309.33
SSRI	paroxetine	329.37
SSRI	citalopram	324.40
SSRI	sertraline	342.70

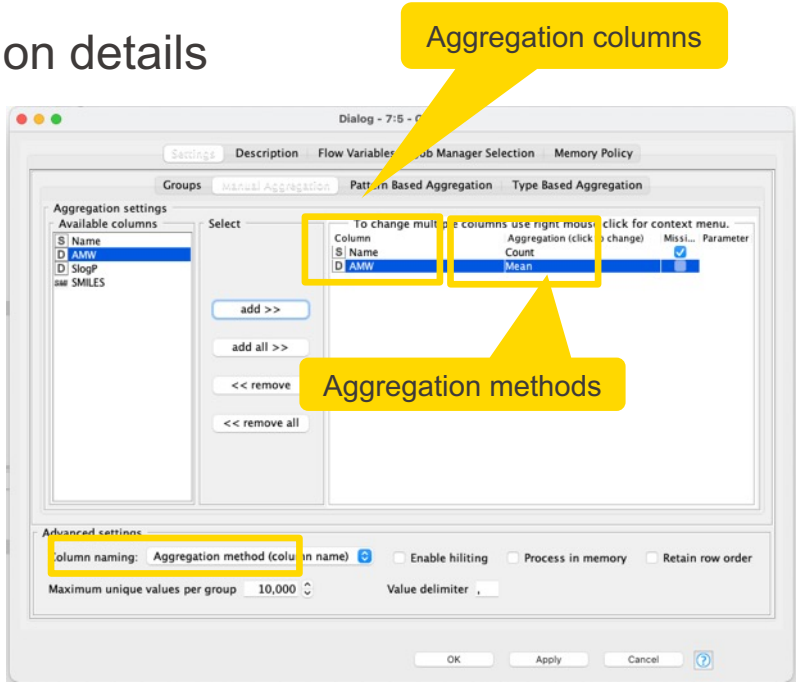
Type	Count(Name)	Mean(Weight)
NSAID	4	208.44
PPI	2	364.40
SSRI	4	326.45

GroupBy

Aggregate to summarize data

- First tab provides grouping options
- Second tab provides control over aggregation details

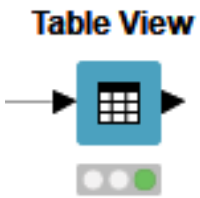


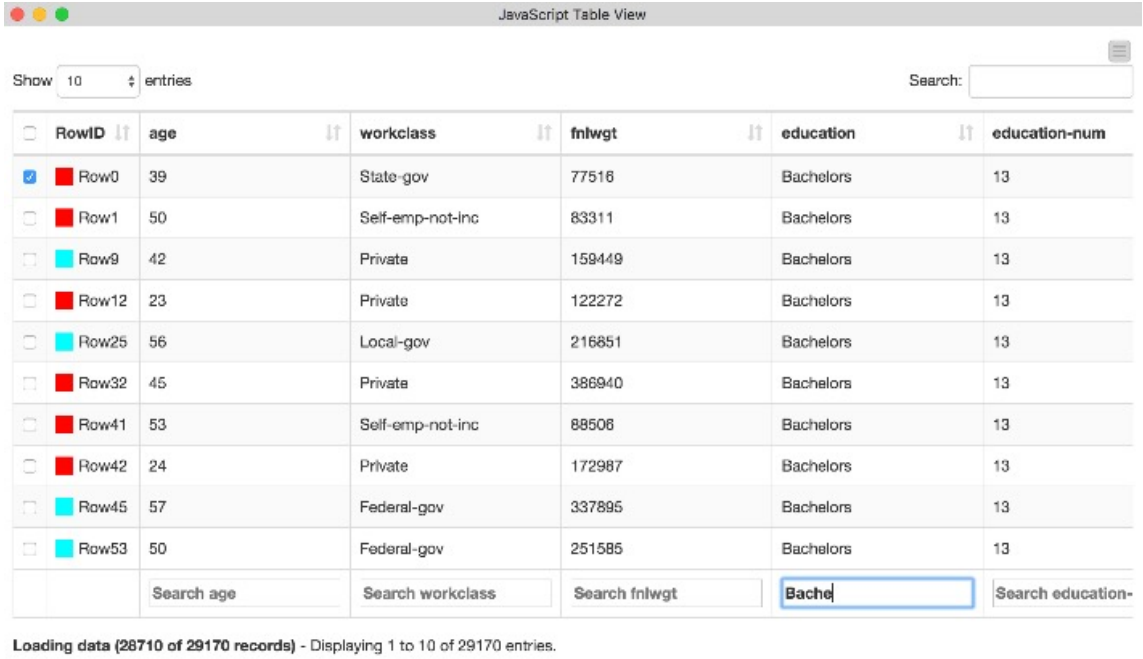


YouTube KNIME TV video: https://youtu.be/bDwF-TOMtWw

Table View

- Display data in an HTML table view.
- The view offers several interactive features, as well as the possibility to select rows



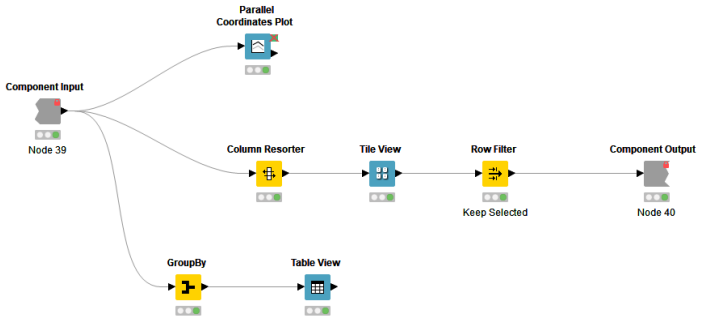


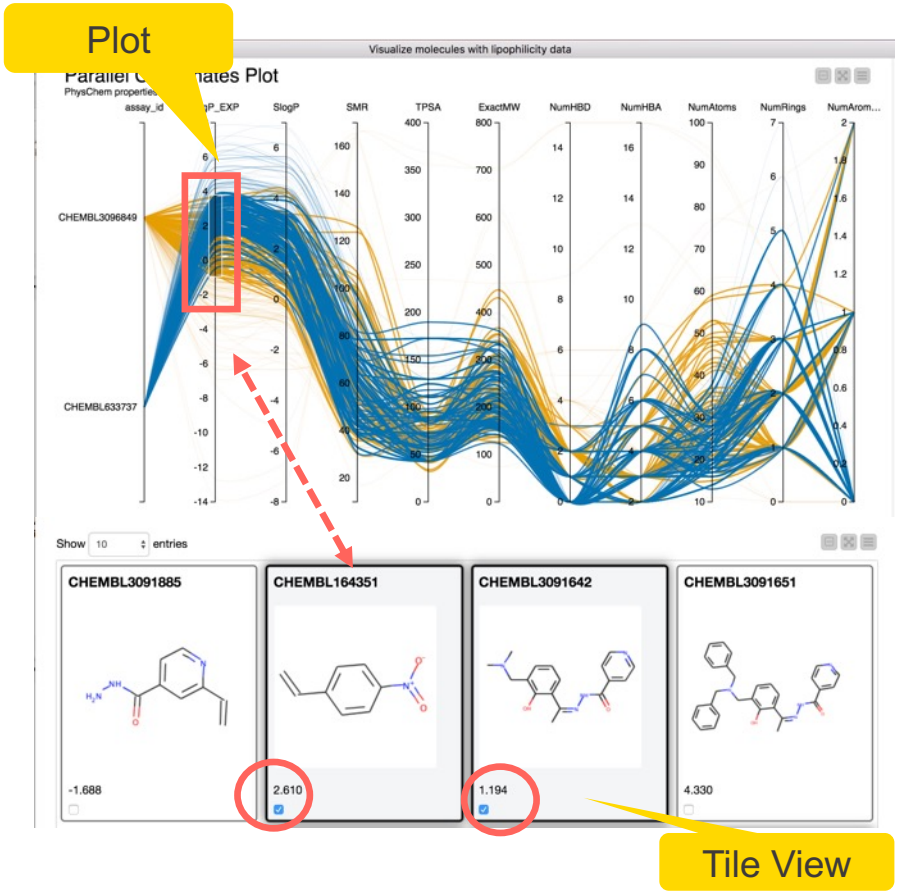
RowID	age	workclass	fnlwgt	education	education-num
Row0	39	State-gov	77516	Bachelors	13
Row1	50	Self-emp-not-inc	83311	Bachelors	13
Row9	42	Private	159449	Bachelors	13
Row12	23	Private	122272	Bachelors	13
Row25	56	Local-gov	216851	Bachelors	13
Row32	45	Private	386940	Bachelors	13
Row41	53	Self-emp-not-inc	88506	Bachelors	13
Row42	24	Private	172987	Bachelors	13
Row45	57	Federal-gov	337895	Bachelors	13
Row53	50	Federal-gov	251585	Bachelors	13
	Search age	Search workclass	Search fnlwgt	Bache	Search education-

Composite views

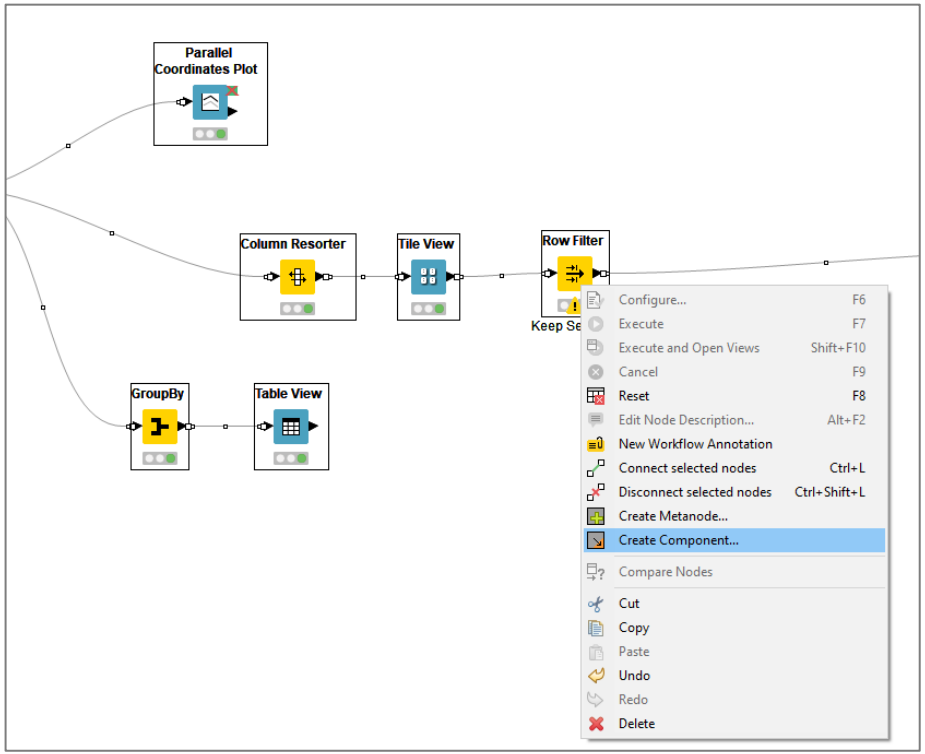
Components – Combined Views

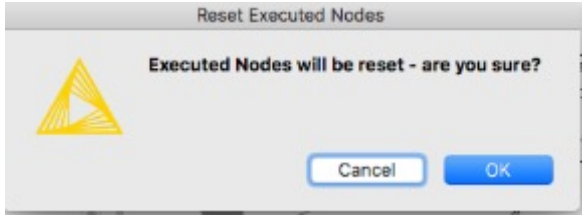
- Multiple JavaScript View nodes can be combined in a **component**
- Selections are transmitted to all other views
- Also for use on the KNIME WebPortal





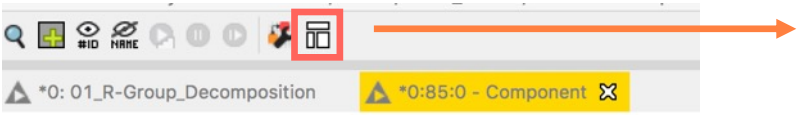
Create a Component

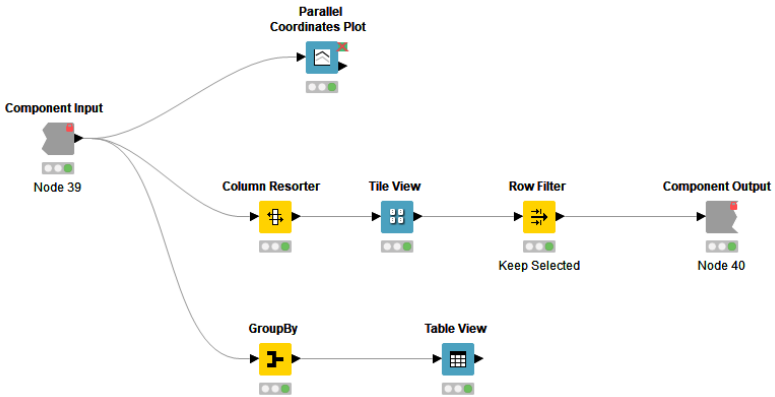




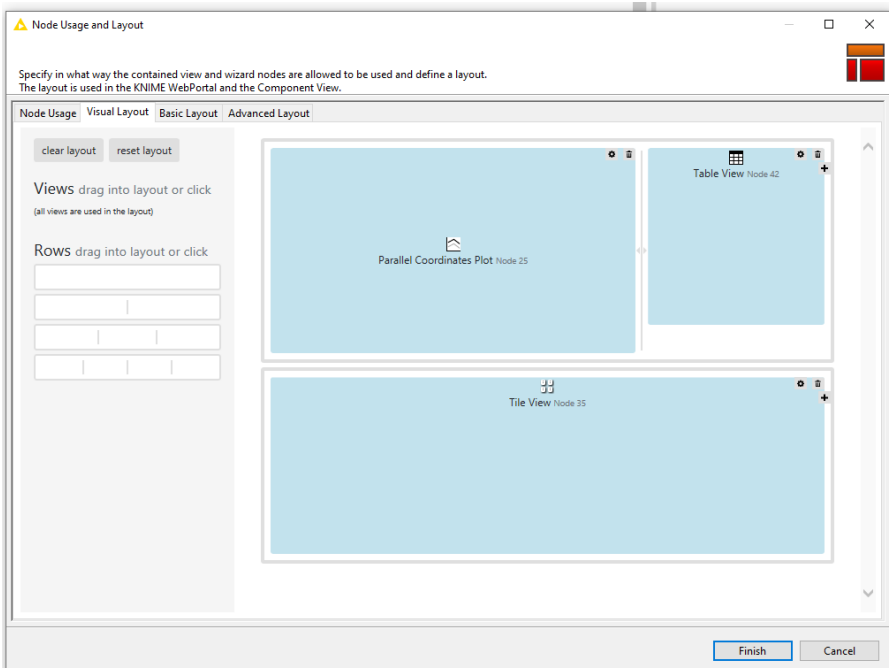
Configure Content and Views Layout

- Click layout button when inside component to assign views to rows and columns

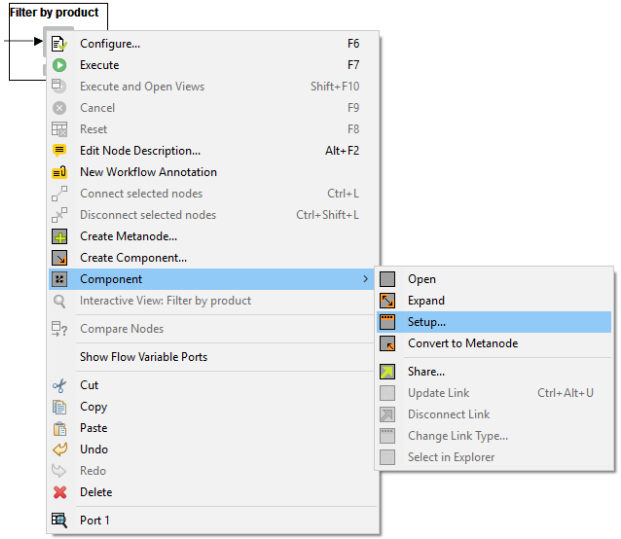




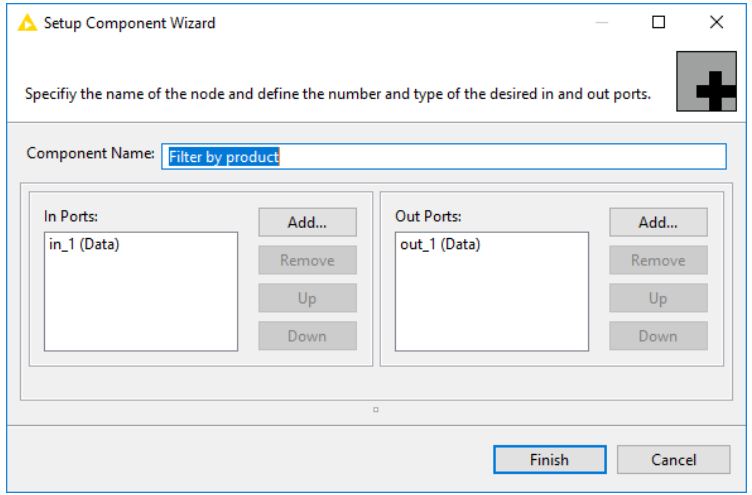
- Add views and rows via *drag&drop*
- Add columns using **+** buttons

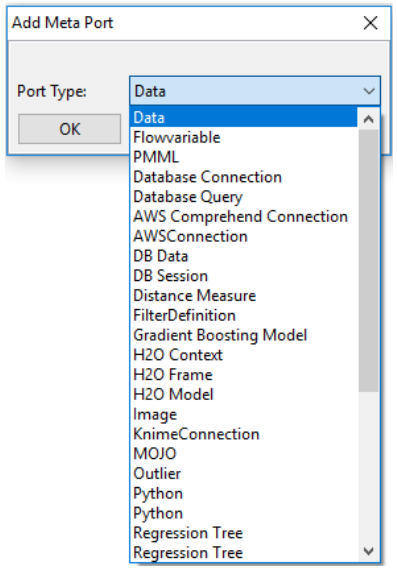


Configure Component Ports



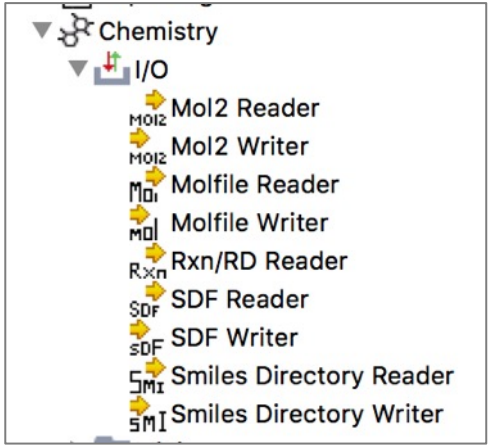
- Add input and output ports to Metanodes/Components
- Remove ports to adapt to changes after creation of Metanode/Component





Saving Files with Writer Nodes

- Writers for chemical formats
- Writers for traditional data types



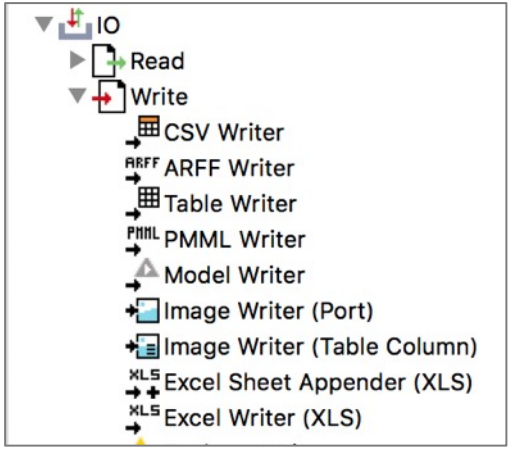
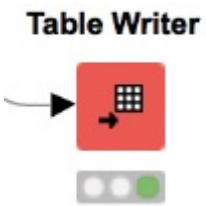
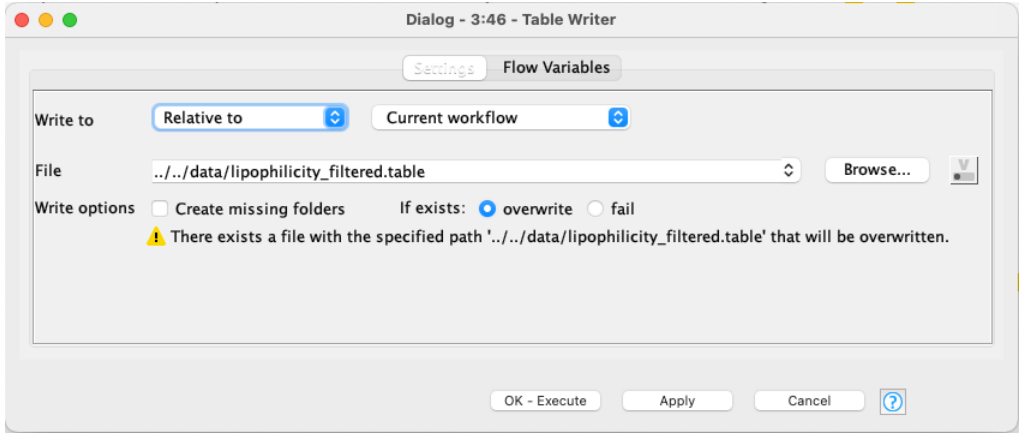


Table Writer

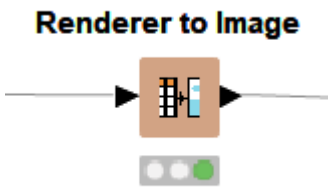
- Saves tables in native KNIME format
- Keeps the types

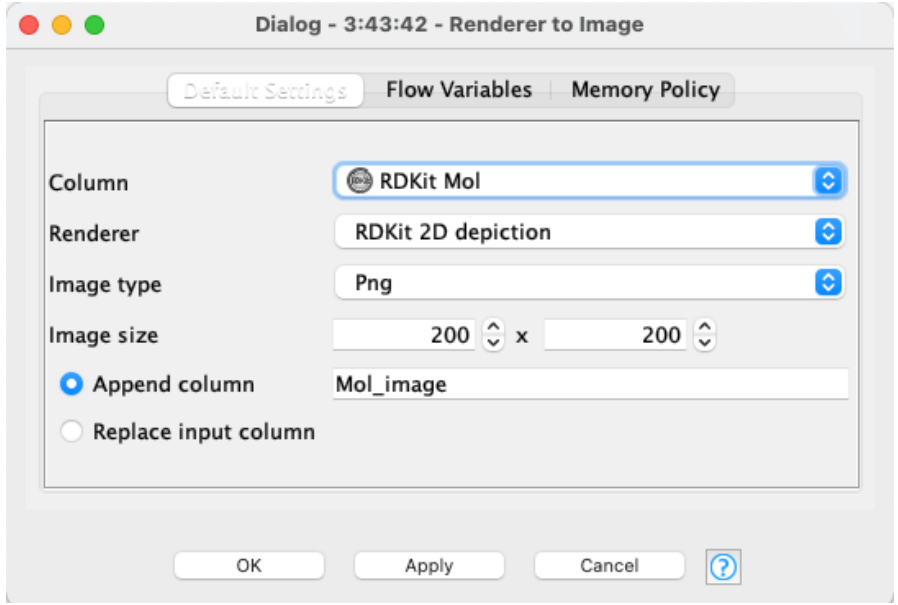




Renderer to Image

- Convert chemical structure to an image
- Various image types





Chemistry_Exercise

1. Read data from multiple files using corresponding Reader nodes. Find them in Node repository >> IO >> Read.
2. Customize the data by adding column names and removing redundant columns.
3. Generate canonical SMILES and remove duplicates.
4. Compute descriptors and use Parallel Coordinates Plot to filter data interactively on multiple properties. (Make sure to keep selection from the View)
5. Finally, save the data to TABLE, Excel, and SDF files.

The Community on the KNIME Hub & Forum

forum.knime.com

Category | Topics

KNIME Analytics Platform — 56 / week
For discussions related to KNIME Analytics Platform

KNIME Extensions — 18 / week
For discussions related to KNIME Extensions and Integrations
Text Processing · Scripting · Reporting · Image Processing · REST · Big Data · Deep Learning

Community Extensions — 3 / week
For discussions related to extensions developed by the KNIME community
RDKit · Scripting Extensions · HCS Tools · Palladian & Seleniu · Vernalis · Seqan

Partner Extensions
For discussions related to extensio partners
JChem Extensions · Schrödinge

KNIME Server
For discussions related to KNIME S

How to translate Chinese PDF documents to English ?
KNIME Extensions · Text Processing

michael19602016 — 5d

Dear All,

I would like to translate one or more chinese (Patent) document(s) to preferably english using for example the amazon (or google) translate node. Therefore, I tried to input some documents using the PDF parser. Settings were: Stanford NLP ChineseTokenizer and Charset = ISO-2022-CN.

However, the node created a table, displaying the path of the input documents in the "Document" column, but no text appeared.

My question is: What do I have to do to "feed" the translation node properly proceeding from some PDF documents in chinese language in a certain folder ?

Kind regards

Michael

created 5d · last reply 3h · 3 replies · 35 views · 2 users · 2 likes

Marten_Pfannenschmidt · KNIME Team Member — 1d

I assume you checked the "Use file path as title" box in configuration dialog of PDF Parser node. That is why the file path is included. If you want to make use of the text (which is stored within the document column) you can use the Doctument Data Extractor, or use the Text Processing nodes which can work with Document Columns.

hub.knime.com

KNIME Hub · Search

data

2 309 results

All · Nodes · Components · Workflows · Extensions

Interactive Data Cleaning
This KNIME component allows you to apply various data cleaning steps interactively. Default configuration will implement cleaning of missing values and outliers. Availble pre-processing steps: - Autom...
knime > Examples > 00_Components > Data Manipulation > Interactive Data Cleaning
Component

Reading and Pre-Processing Data
This component reads customer data from three different sources and pre-processes the data.
knime > Education > 03 KNIME Server Course > Components > Reading and Pre-Processing Data
Component

KNIME Big Data Connectors
KNIME nodes for accessing Big Data infrastructure such as Apache Hive, Apache Impala, or HDFS.
KNIME AG, Zurich, Switzerland

>Preferences->KimeHelpers->DataExamples
Source

02_Hive_WritingtoDB_Exercise

Download workflow
Short link

Big Data Course Hadoop Exercise #2

Make sure you have executed the /2_Hadoop/2_Exercises/00_Setup_Hive_Table workflow during your current KNIME session before running this workflow.

External resources
Slides for KNIME Analytics Platform Big Data Course

Add to KNIME Analytics Platform
Drag node into the workbench of KNIME Analytics Platform 4.x

Thank You!