

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

AIDD  
April 19<sup>th</sup>, 2022

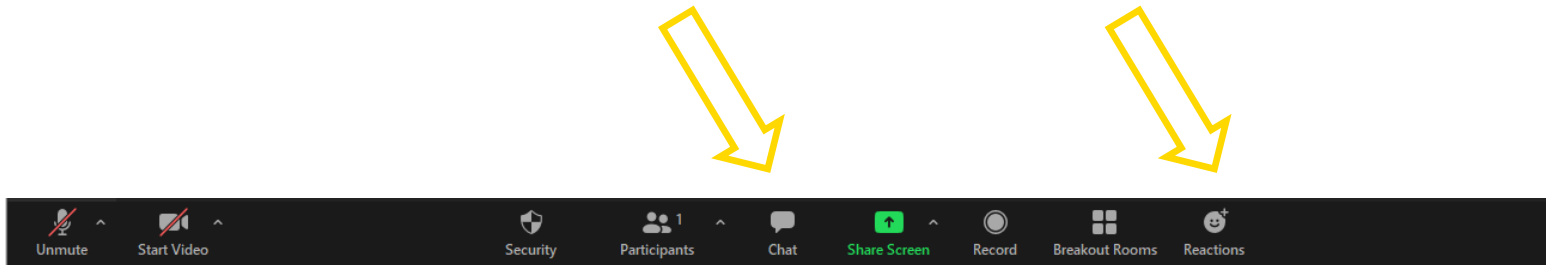
Daria Goldmann  
**KNIME AG**



# 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

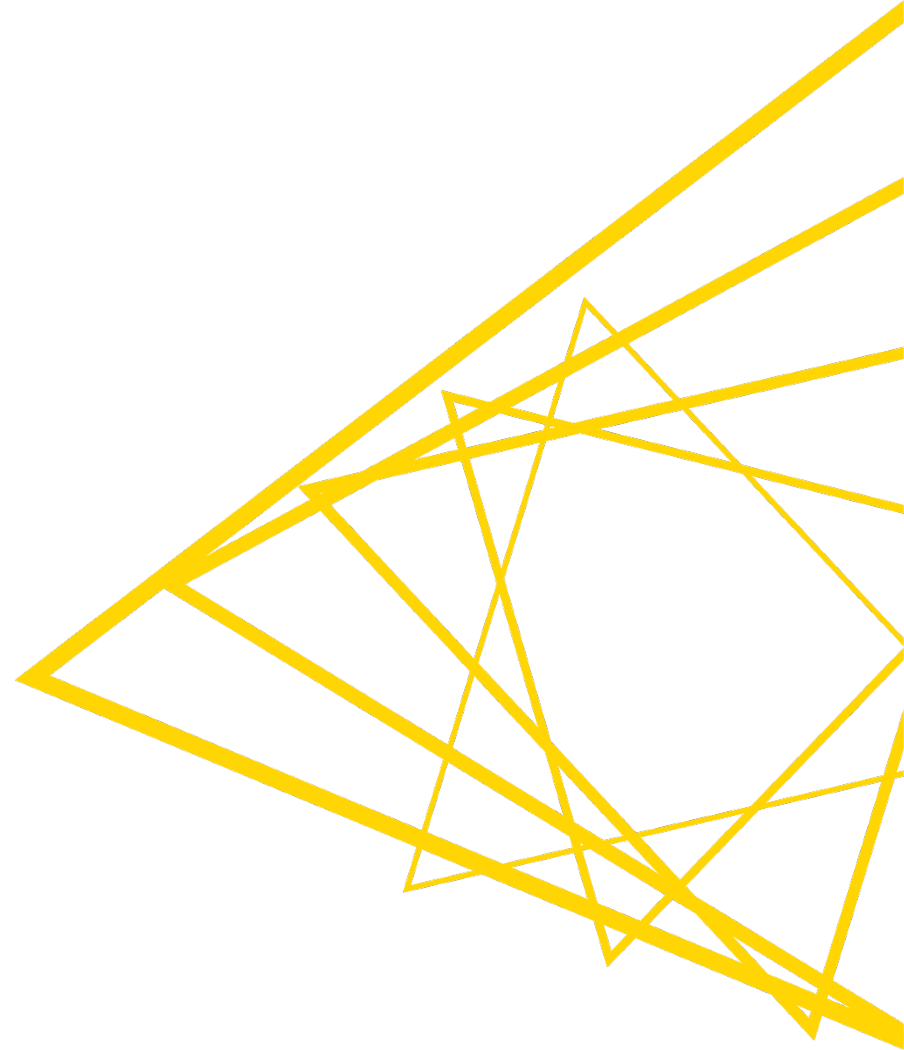


# Scenario

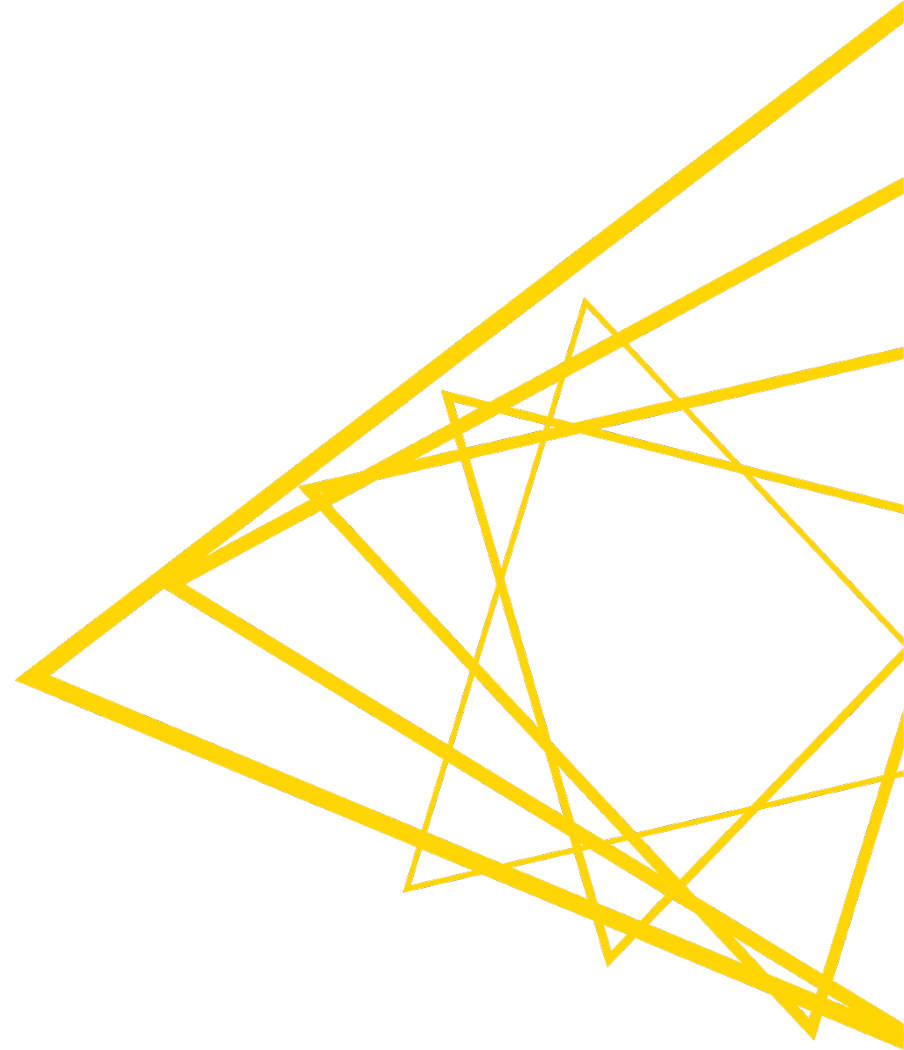
---

- 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 colleagues

**Teaser of the view**

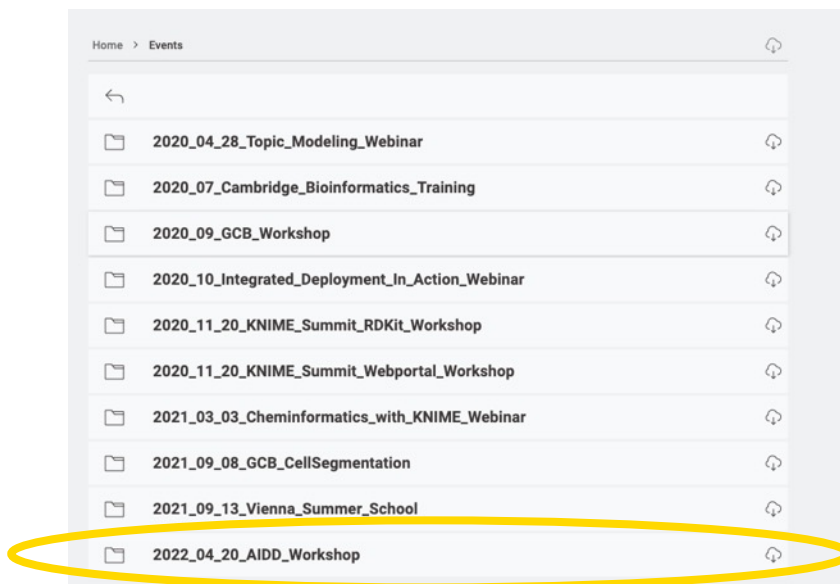


# Setup



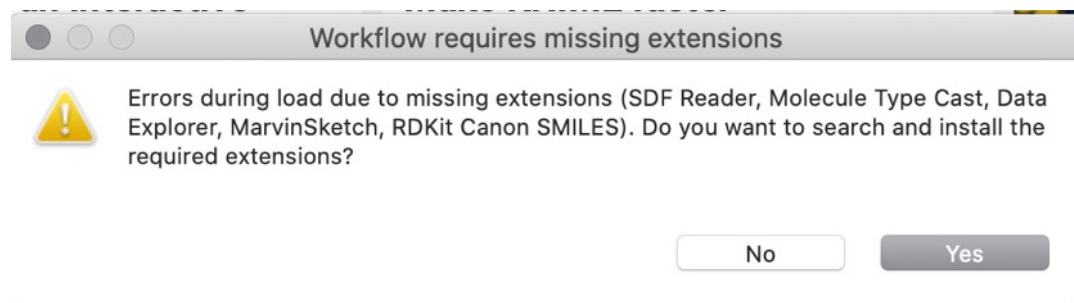
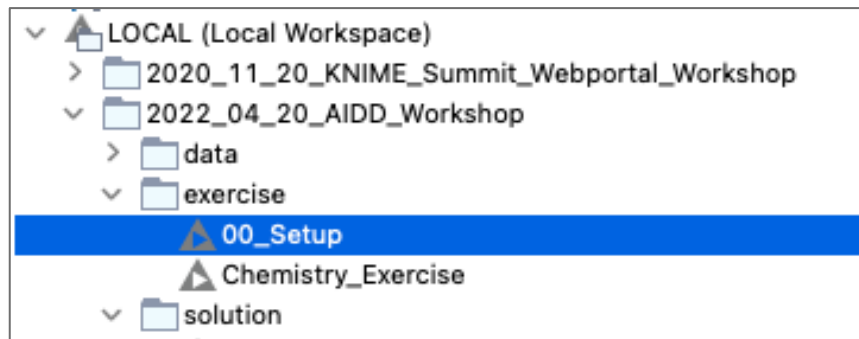
# Set up KNIME Analytics Platform

- Download exercises (2022\_04\_20\_AIDD\_Workshop) from [https://hub.knime.com/knime/spaces/Life%20Sciences/latest/Events~ekRSJneVS0b0RD\\_k/](https://hub.knime.com/knime/spaces/Life%20Sciences/latest/Events~ekRSJneVS0b0RD_k/)

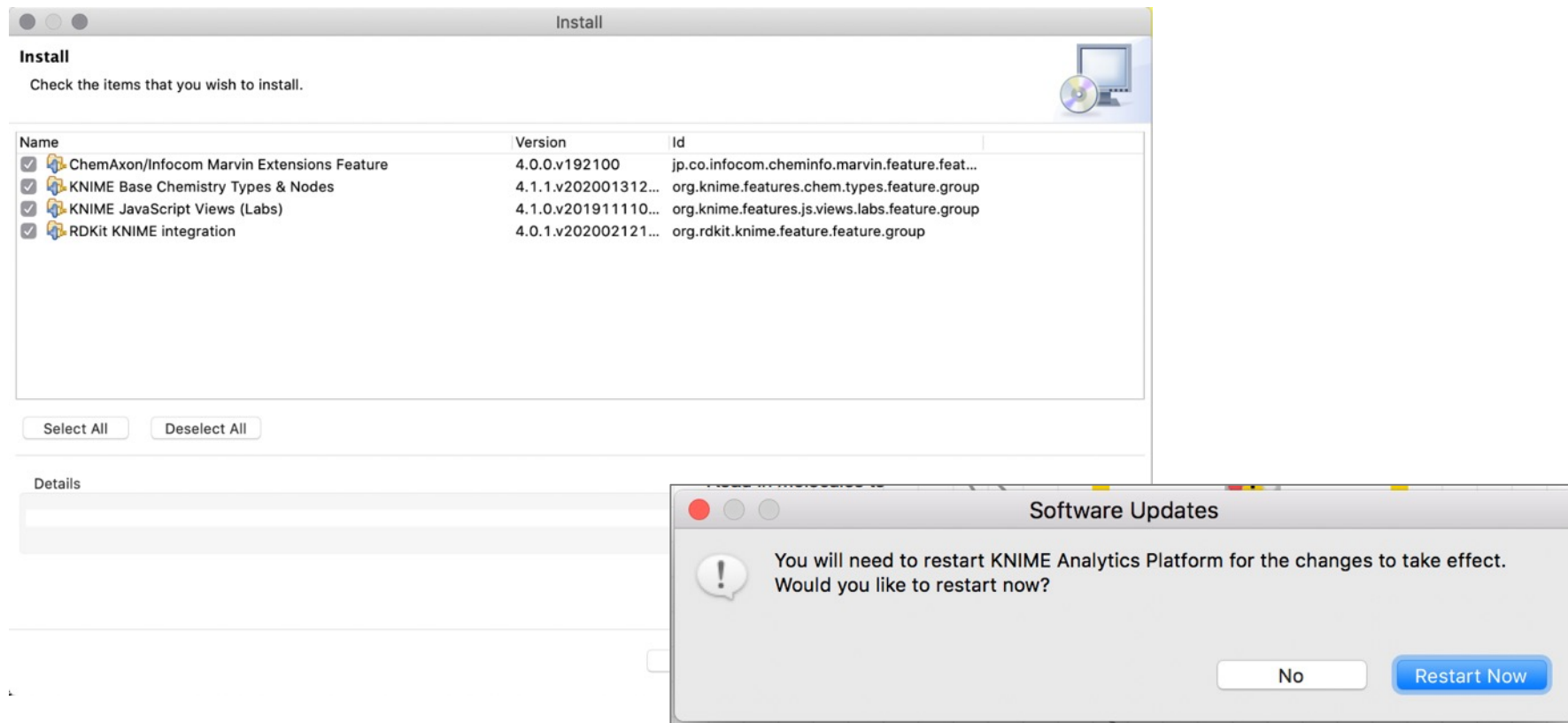


- Import the file into your local workspace

# Open the 00\_Setup workflow

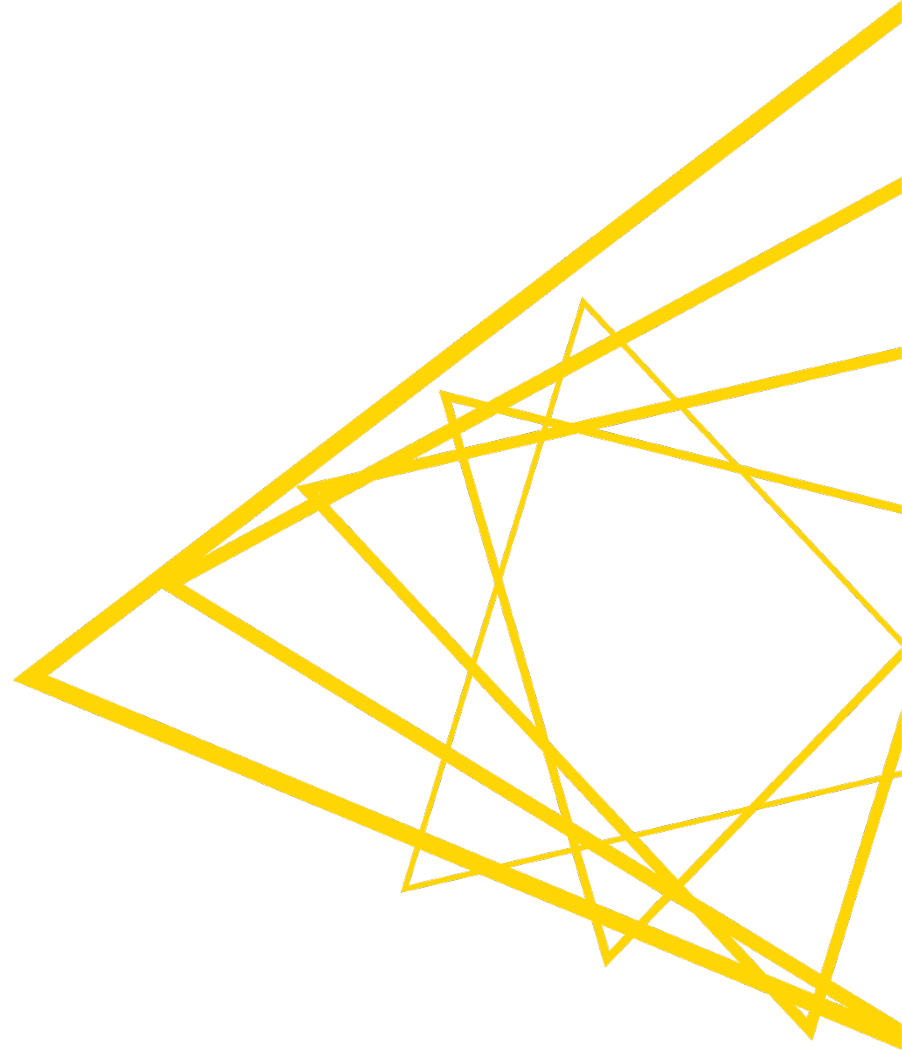


# Install the Extensions and Restart





**You are now ready to go**

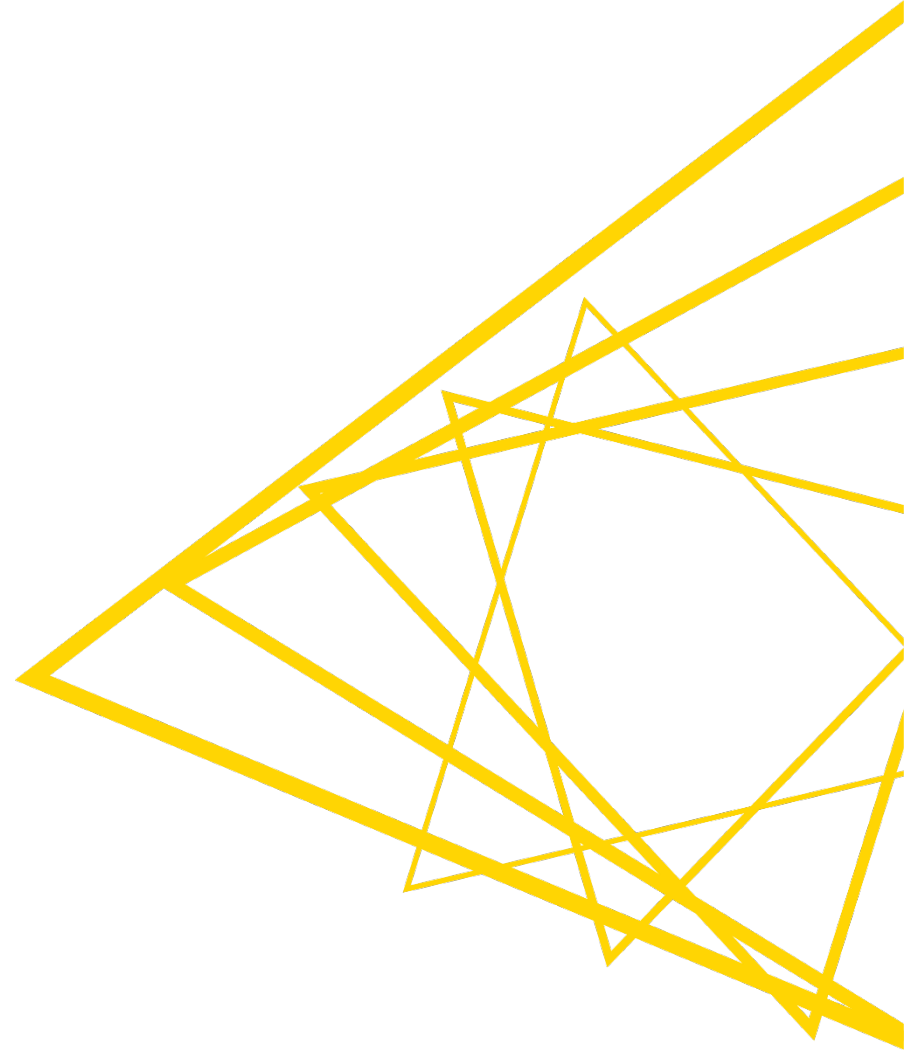


# Agenda for Today

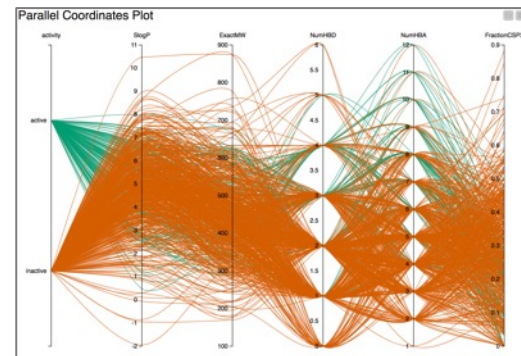
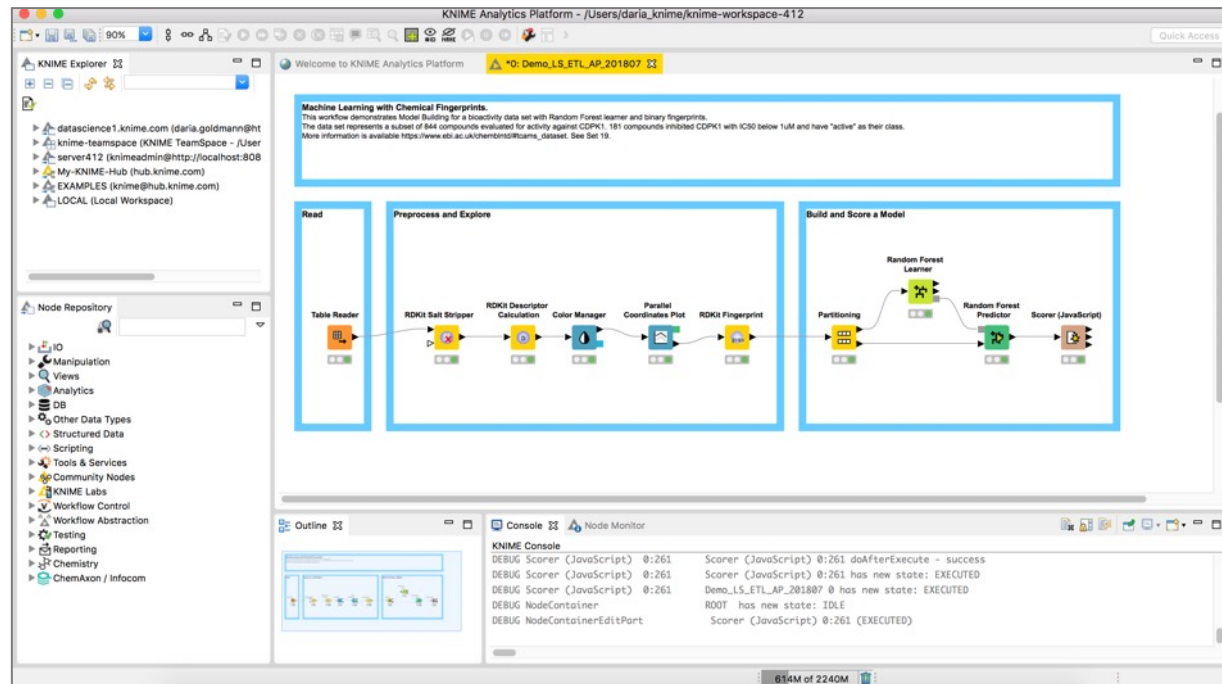
---

- 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

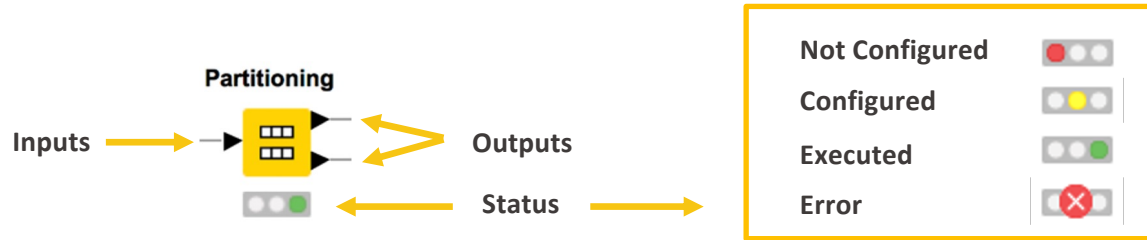
# The KNIME Workbench

The screenshot displays the KNIME Analytics Platform interface with the following components labeled:

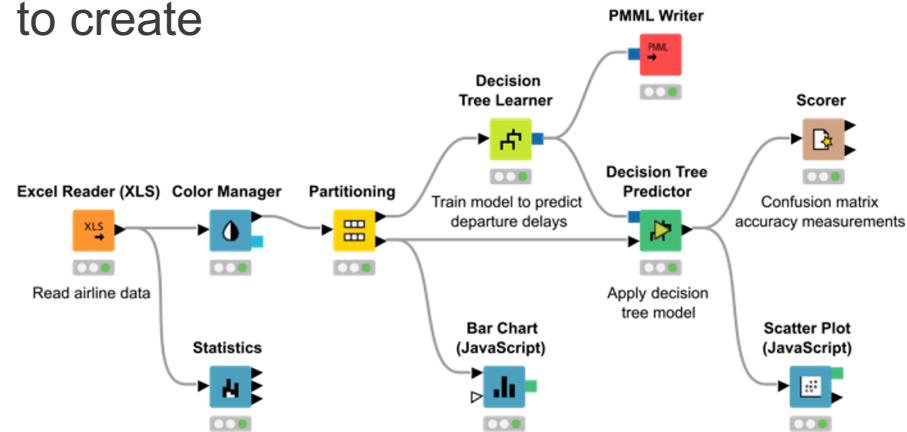
- Servers and Workflows:** Located in the top-left pane (KNIME Explorer), showing a project tree with folders like 'data', 'exercises', and 'solutions', and files like '01\_Chemistry\_Basics.knimeproject'.
- Node Recommendations:** Located in the middle-left pane (Workflow Coach), showing a list of recommended nodes such as 'File Reader', 'CSV Reader', and 'Table Creator'.
- Node Repository:** Located in the bottom-left pane, showing a hierarchical tree of node categories including 'IO', 'Manipulation', 'Views', 'DB', 'Analytics', 'Other Data Types', 'Structured Data', 'Scripting', 'Tools & Services', 'Community Nodes', 'KNIME Labs', 'Workflow Control', 'Workflow Abstraction', 'Reporting', 'Chemistry', and 'ChemAxon / Infocom'.
- Workflow Editor:** The central workspace showing a workflow diagram with nodes like 'File Reader', 'SDF Reader', 'Concatenate', 'Molecule Type Cast', 'RDKit Canon SMILES', and 'Duplicate Row Filter'. The workflow is divided into two steps: 'Step 1: Read data from different sources' and 'Step 2: Remove duplicates'.
- Node Description:** A yellow callout pointing to the 'Node Description' pane on the right, which provides details about the selected node.
- KNIME Hub:** A yellow callout pointing to the 'KNIME Hub Search' pane on the right, which allows searching for workflows, nodes, and more.
- Console:** A yellow callout pointing to the 'Console' pane at the bottom right, which displays log messages and warnings.
- Outline:** A yellow callout pointing to the 'Outline' pane at the bottom left, which shows a hierarchical view of the workflow nodes.

# Visual KNIME Workflows

**NODES** perform tasks on data



Nodes are combined to create  
**WORKFLOWS**

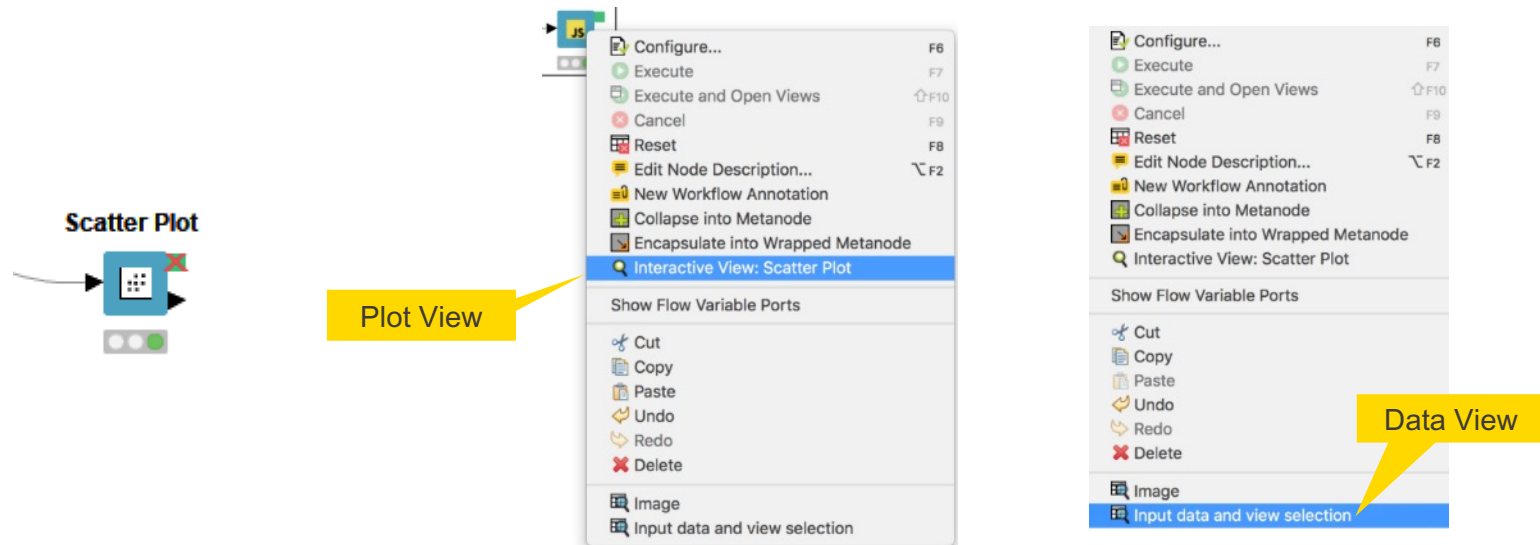


# Node Outputs and Views

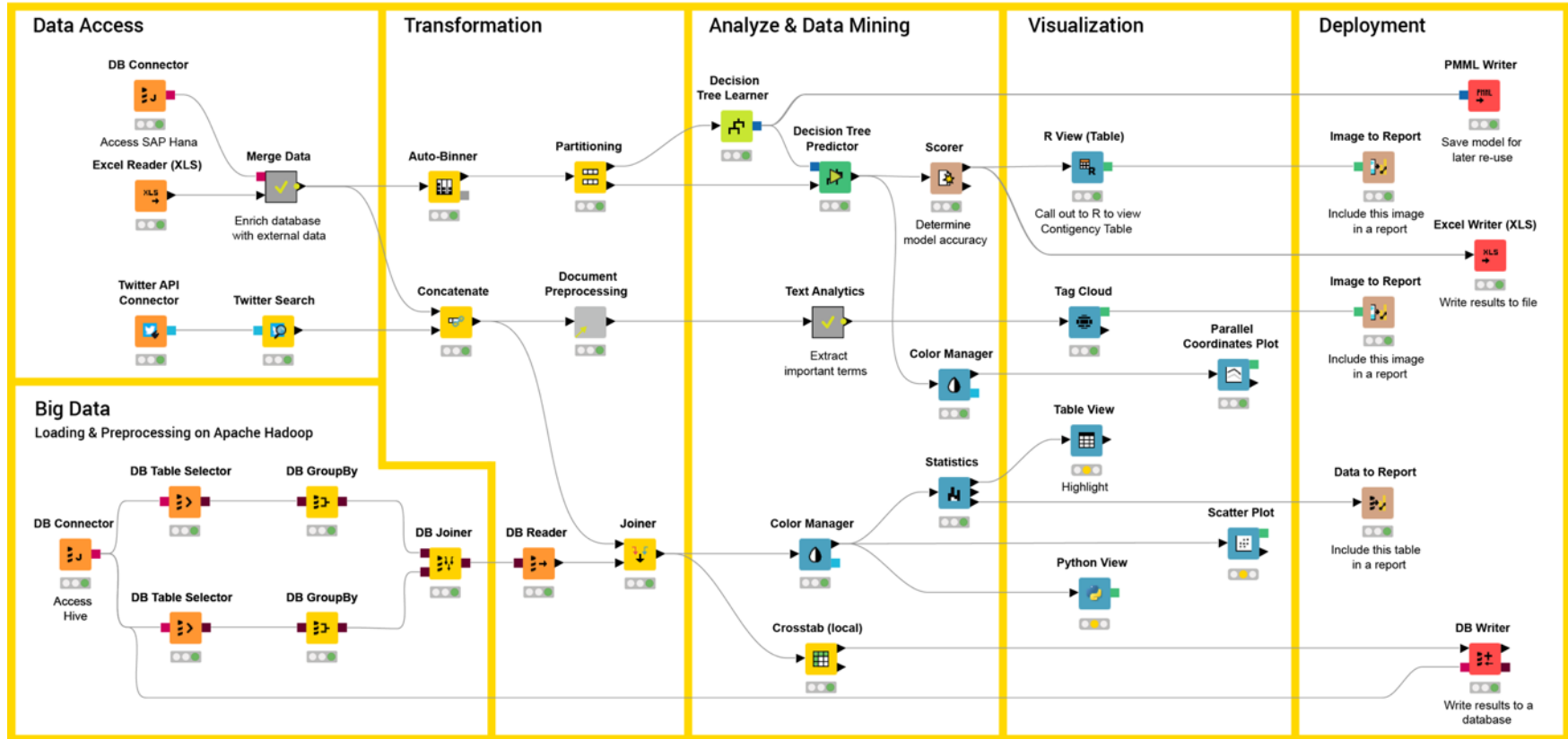
- Right-click executed node
- Select View option in context menu

OR

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

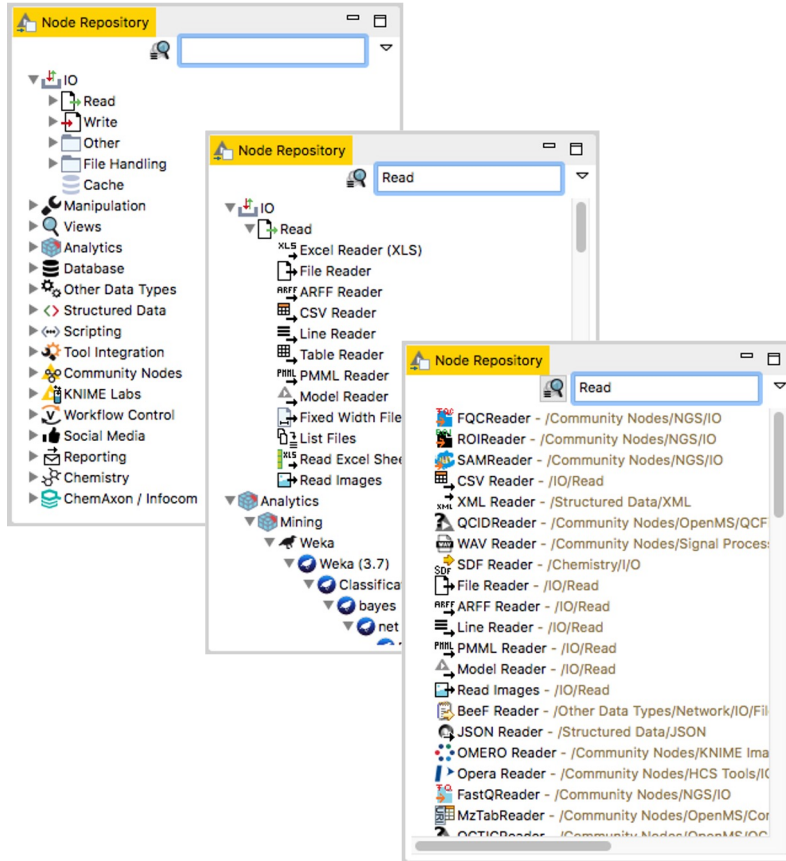




# 4000+ Nodes for all Steps of End-To-End Data Science





# 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

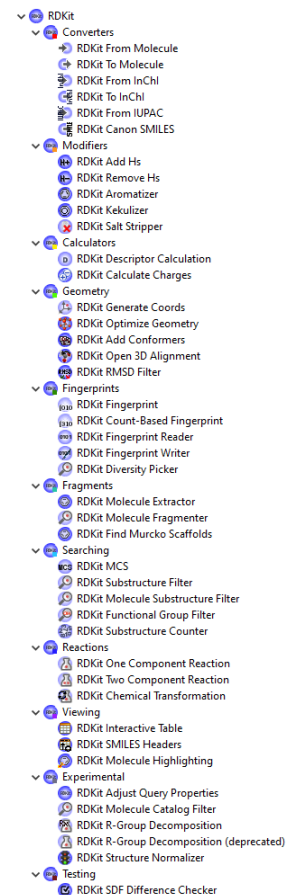
The image displays four panels of open-source extensions for cheminformatics, each with a tree view structure. The extensions are categorized as follows:

- 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



- ▼ BioSolveIT Nodes
  - ▶ CoLibri (Chemistry Spaces)
  - ▶ IO
    - ▶ Assess Affinity with Hyde in SeeSAR
    - ▶ Compute FTrees Similarity
    - ▶ Compute FlexS Alignments
    - ▶ Compute LeadIT Docking
    - ▶ Convert Molecules with Naomi
    - ▶ FTrees Query Generator
    - ▶ Filter Molecules with Naomi
    - ▶ FlexX Docking
    - ▶ 3D Generate 3D Coordinates
    - ▶ Generate Protomers / Tautomers with Naomi
    - ▶ Interactive BioSolveIT Table
    - ▶ Interactive SeeSAR Viewer
    - ▶ Prepare Receptor with LeadIT
    - ▶ Run ReCore Interactively
    - ▶ Search FTrees Fragment Space
    - ▶ SeeSAR Project Generator



- ▼ ChemAxon / Infocom
  - ▼ JChem
    - ▶ IO
      - ▶ Converter
      - ▶ Marvin
      - ▶ Calculator Plugins
      - ▶ JChem Base
      - ▶ JChem Cartridge
      - ▶ Standardizer
      - ▶ Structure Checker
      - ▶ Name to Structure
      - ▶ Screen
      - ▶ JKlustor
      - ▶ Reactor
      - ▶ Markush Viewer
      - ▶ Metabolizer
      - ▶ Fragmenter
    - ▶ Marvin



- ▼ Cresset
  - ▼ Forge
    - ▶ Models
    - ▶ Project
    - ▶ Forge Align
    - ▶ Activity Miner
    - ▶ FieldTemplater
  - ▼ Spark
    - ▶ Spark Fragment Selector
    - ▶ Generate Spark Database
    - ▶ Spark Database Search
  - ▼ XedTools
    - ▶ XedMin
    - ▶ XedeX
    - ▶ Torch/Forge Molecule Viewer



- ▼ MOE
  - ▶ Input
  - ▶ Output
  - ▶ Convert
  - ▶ Transform
  - ▶ Process
  - ▶ Calculate
  - ▶ QuaSAR
  - ▶ Fingerprints
  - ▶ Simulations
  - ▶ Bioinformatics
  - ▶ Fragment Based Design
  - ▶ CombiChem
  - ▶ Miscellaneous
  - ▶ Pharmacophore
  - ▶ Materials



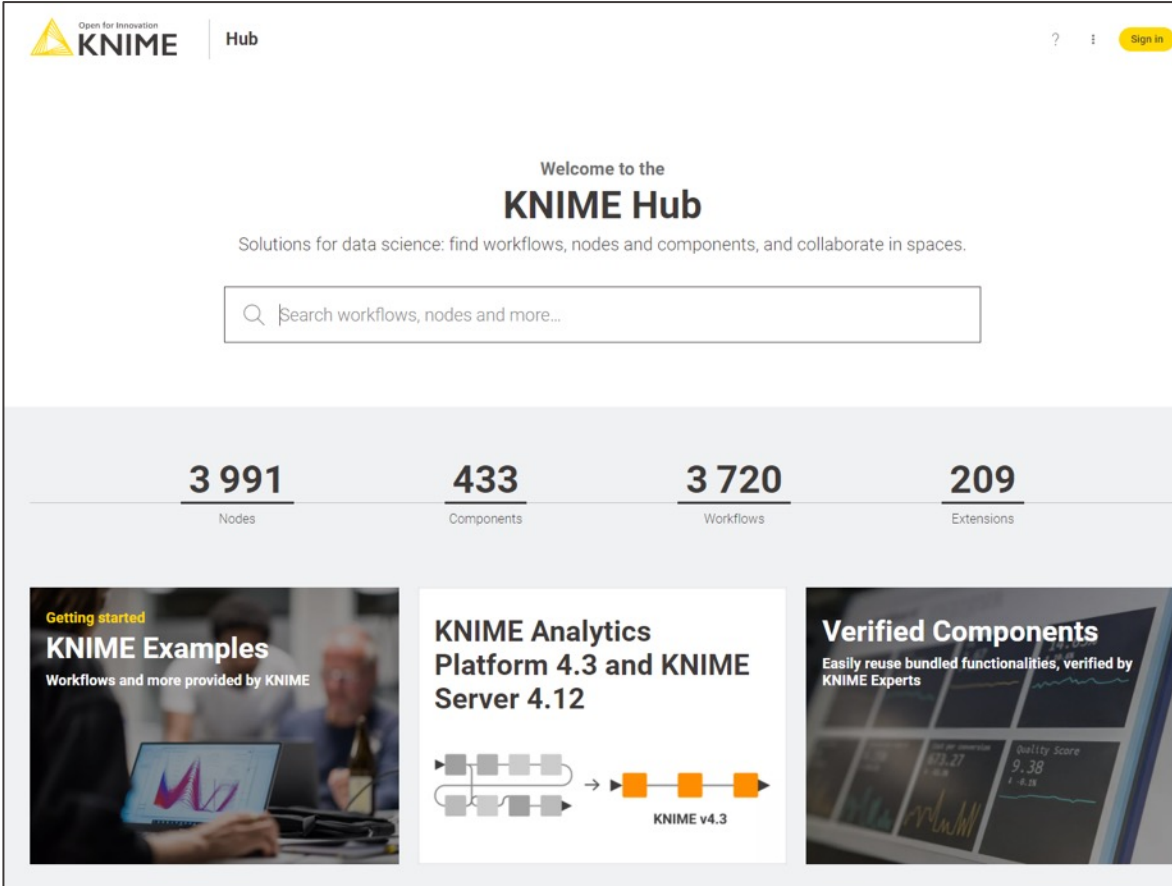
- ▶ Schrödinger
  - ▶ Readers/Writers
  - ▶ Converters
  - ▶ Ligand Preparation
  - ▶ Property Generation
  - ▶ Cheminformatics
  - ▶ Pharmacophore Modeling
  - ▶ Protein Structure Prediction
  - ▶ Docking and Scoring
  - ▶ Molecular Mechanics
  - ▶ Molecular Dynamics
  - ▶ Quantum Mechanics
  - ▶ Workflows
  - ▶ Filtering
  - ▶ Reporting
  - ▶ Scripting
  - ▶ Tools

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



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

# KNIME Hub: Searching, Sharing, and Collaborating



The screenshot shows the KNIME Hub homepage. At the top left is the KNIME logo with the tagline 'Open for Innovation'. To its right is the word 'Hub'. On the top right, there are links for help (question mark) and a 'Sign in' button. The main heading reads 'Welcome to the KNIME Hub' followed by the subtitle 'Solutions for data science: find workflows, nodes and components, and collaborate in spaces.' Below this is a search bar with the placeholder text 'Search workflows, nodes and more...'. A statistics bar displays four categories: 'Nodes' with 3,991 items, 'Components' with 433 items, 'Workflows' with 3,720 items, and 'Extensions' with 209 items. The bottom section features three promotional tiles. The first tile, 'Getting started KNIME Examples', shows a person using a laptop with a data visualization. The second tile, 'KNIME Analytics Platform 4.3 and KNIME Server 4.12', includes a diagram of a workflow with nodes and arrows, labeled 'KNIME v4.3'. The third tile, 'Verified Components', shows a dashboard with various charts and a 'Quality Score' of 9.38.

Open for Innovation  
**KNIME** Hub

Welcome to the  
**KNIME Hub**

Solutions for data science: find workflows, nodes and components, and collaborate in spaces.

Search workflows, nodes and more...

**3 991**  
Nodes

**433**  
Components

**3 720**  
Workflows

**209**  
Extensions

**Getting started**  
**KNIME Examples**  
Workflows and more provided by KNIME

**KNIME Analytics Platform 4.3 and KNIME Server 4.12**

**Verified Components**  
Easily reuse bundled functionalities, verified by KNIME Experts

Quality Score  
9.38

KNIME v4.3

# 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>

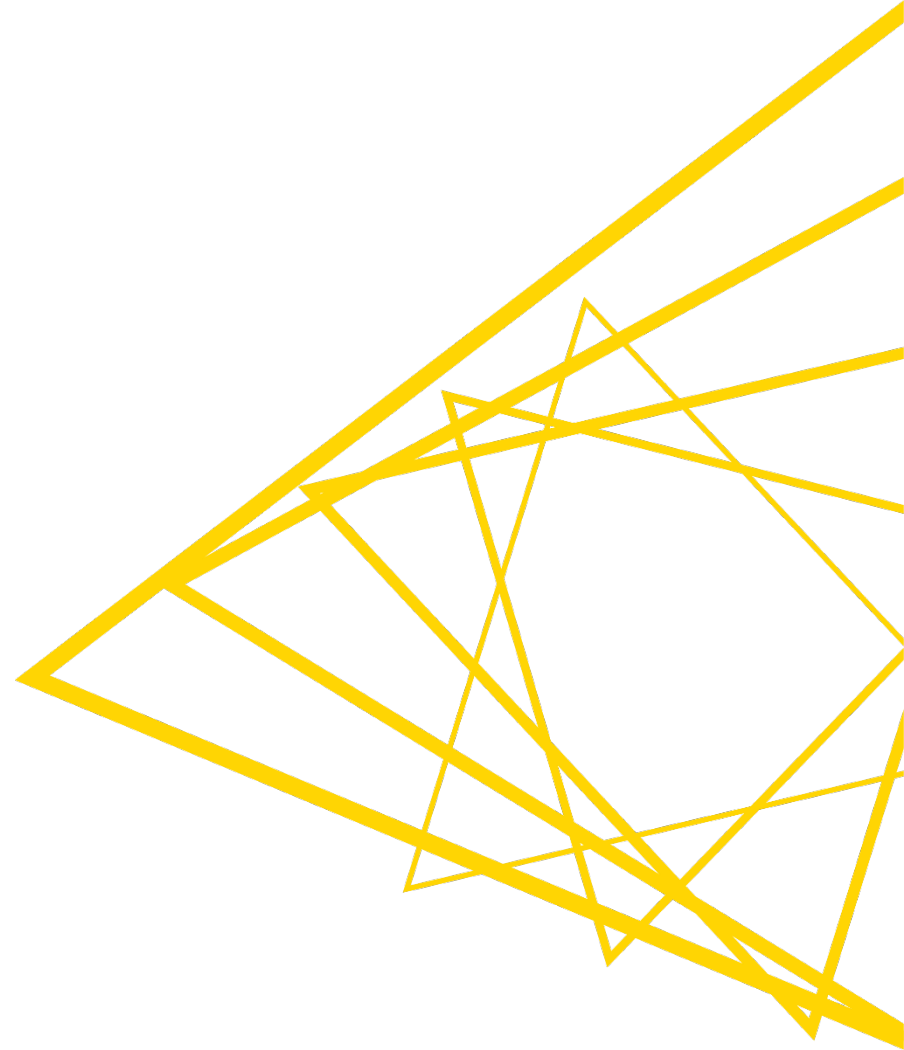
**KNIME Tech** pages

**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 coefficient 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:

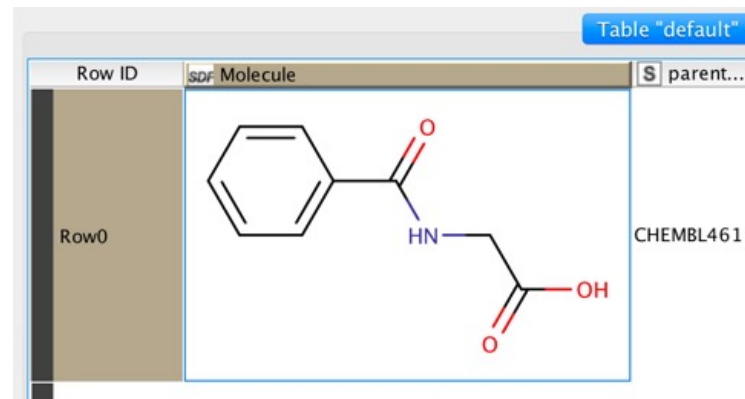
- [PubChem AID 686912](#): GSK\_TB: GSK in-house hydrophobicity assay at pH7.4 (logD)
- 148 molecules
- CSV file with experimental data
- SD file with structural data

## ■ 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)



# File Reader (Complex Format) Configuration

File Reader  
(Complex Format)



File system

File path

Advanced  
settings

Column type

Dialog - 0:562 - File Reader (Complex Format) (in-house)

Settings | Flow Variables | Job Manager Selection | Memory Policy

Enter ASCII data file location: (press 'Enter' to update preview)

Read from: Relative to | Current workflow

File:  Browse...

☐ Preserve user settings for new location Rescan

Basic Settings

☐ read row IDs Column delimiter: ; Advanced...

☒ read column headers ☐ ignore spaces and tabs

☐ Java-style comments Single line comment:

Preview

Click column header to change column properties (\* = name/type user settings)

Row ID	S	Molecule Ch...	S	Mole...	I	Mo...	D	Mol...	I	#ROS...	D	AlogP	S	Comp...	use *Smiles
Row0		CHEMBL2178940		0		391.48	0			2.89		4721			
Row1		CHEMBL168899		0		238.06	?			?		5805			
Row2		CHEMBL573214 NORBINA...		0		661.8	1			3.82		4168			

OK Apply Cancel ?

# Common Settings: Four Default File Systems

## ■ Local File System

Input location

Read from:

Mode: ☒ File ☐ Files in folder

File:

## ■ Relative to ...

Read from: 

- Current mountpoint
- Current workflow data area
- Current workflow

File:

## ■ Mountpoint

Read from:

File:

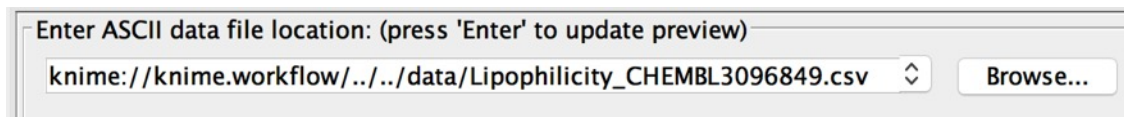
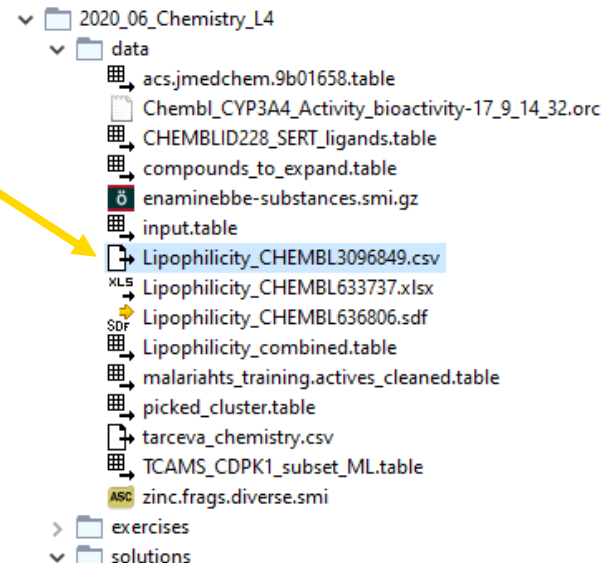
## ■ Custom URL

Read from:

URL:

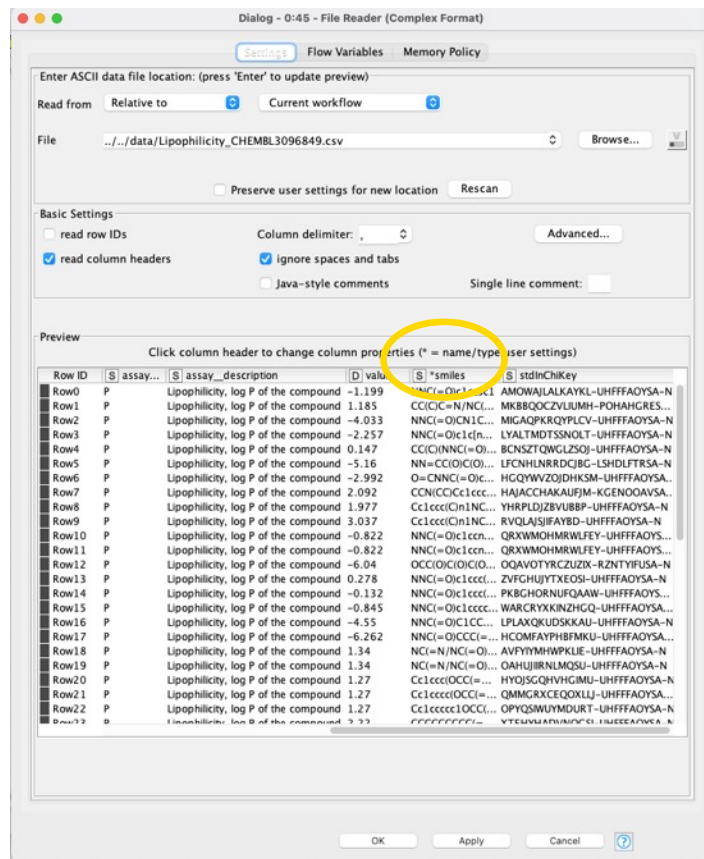
# Workflow-Relative File Paths

- 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

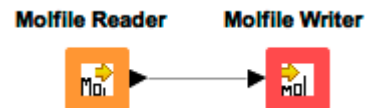
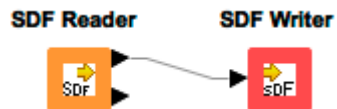
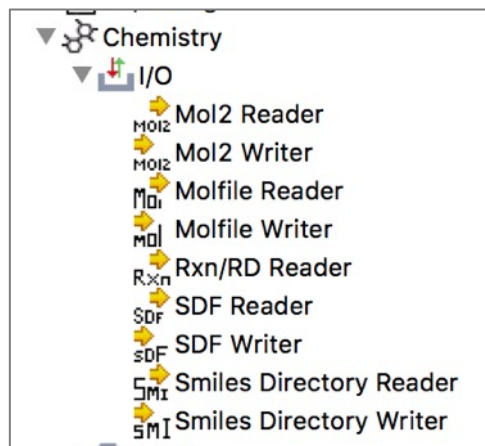


D value	smiles	stdInChiKey
-1.199		AMOWAJLALKAYKL-UHFFFAOYSA-N
1.185		MKBBQOCZVLUMH-POHAHGRE...
-4.033		MIGAPKRQYPLCV-UHFFFAOYSA-N
-2.257		LYALTMDSSTNOLT-UHFFFAOYSA-N
0.147		BCNSZTQWGLZSOJ-UHFFFAOYSA-N

# Nodes for Reading and Writing files

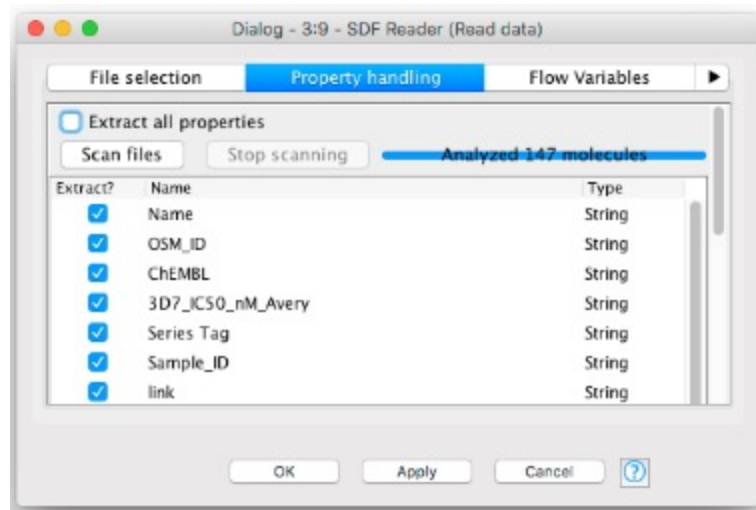
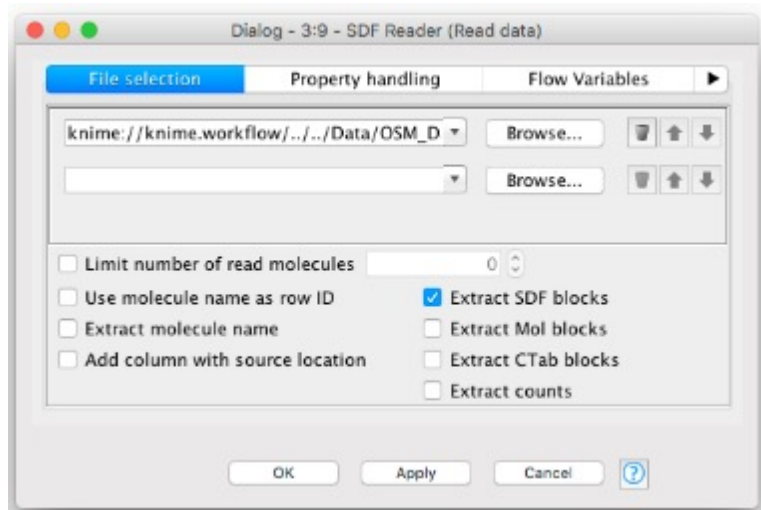
Reader and writers provided for:

- SDF, SMILES, mol, mol2

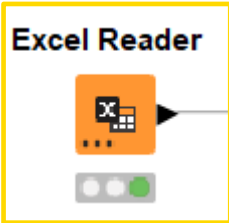




# A bit more about reading SD files



# Excel Reader



Dialog - 0:44 - Excel Reader

Settings Transformation Advanced Settings Encryption Flow Variables Memory Policy

Input location

Read from Relative to Current workflow

Mode ☒ File ☐ Files in folder

File ../data/Lipophilicity\_CHEMBL633737.xlsx Browse...

Sheet selection

☒ Select first sheet with data (default)

☐ Select sheet with name default

☐ Select sheet at index 0 (Sheet indexes start with 0.)

Column header

☒ Use Excel column name e.g. A, B, C ☐ Use column index e.g. Col1, Col2

☒ Table contains column names in row number 1 (Row numbers start with 1. See "File Content" tab to identify row numbers.)

Empty column name prefix: empty\_

Row ID

☒ Generate row IDs ☐ Table contains row IDs in column A

Sheet area

☒ Read entire data of the sheet ☐ Read only data in columns from A to , rows from 1 to . (See "File Content" tab to identify columns and rows.)

Preview with current settings

The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	S parent_c...	S bioact...	S opera...	S units	S assay_che...	S assay...	S assay_description	D value	S smiles
Row0	CHEMBL422385	LogP	=	Unspecified	CHEMBL633737 A		Partition coefficient (logP)	4.99	[O-][N+](=O)
Row1	CHEMBL164282	LogP	=	Unspecified	CHEMBL633737 A		Partition coefficient (logP)	0.96	[O-][N+](=O)
Row2	CHEMBL167688	LogP	=	Unspecified	CHEMBL633737 A		Partition coefficient (logP)	5.61	[O-][N+](=O)
Row3	CHEMBL353064	LogP	=	Unspecified	CHEMBL633737 A		Partition coefficient (logP)	3.24	[O-][N+](=O)
Row4	CHEMBL73297	LogP	=	Unspecified	CHEMBL633737 A		Partition coefficient (logP)	3.52	[O-][N+](=O)
Row5	CHEMBL167121	LogP	=	Unspecified	CHEMBL633737 A		Partition coefficient (logP)	3.43	Cc1cc2cccc
Row6	CHEMBL56970	LogP	=	Unspecified	CHEMBL633737 A		Partition coefficient (logP)	2.39	[O-][N+](=O)
Row7	CHEMBL16374	LogP	=	Unspecified	CHEMBL633737 A		Partition coefficient (logP)	5.87	[O-][N+](=O)
Row8	CHEMBL167986	LogP	=	Unspecified	CHEMBL633737 A		Partition coefficient (logP)	1.92	[O-][N+](=O)
Row9	CHEMBL144962	LogP	=	Unspecified	CHEMBL633737 A		Partition coefficient (logP)	2.73	[O-][N+](=O)
Row10	CHEMBL165395	LogP	=	Unspecified	CHEMBL633737 A		Partition coefficient (logP)	3.26	[O-][N+](=O)
Row11	CHEMBL355315	LogP	=	Unspecified	CHEMBL633737 A		Partition coefficient (logP)	4.69	[O-][N+](=O)
Row12	CHEMBL350310	LogP	=	Unspecified	CHEMBL633737 A		Partition coefficient (logP)	1.80	[O-][N+](=O)

OK Apply Cancel ?

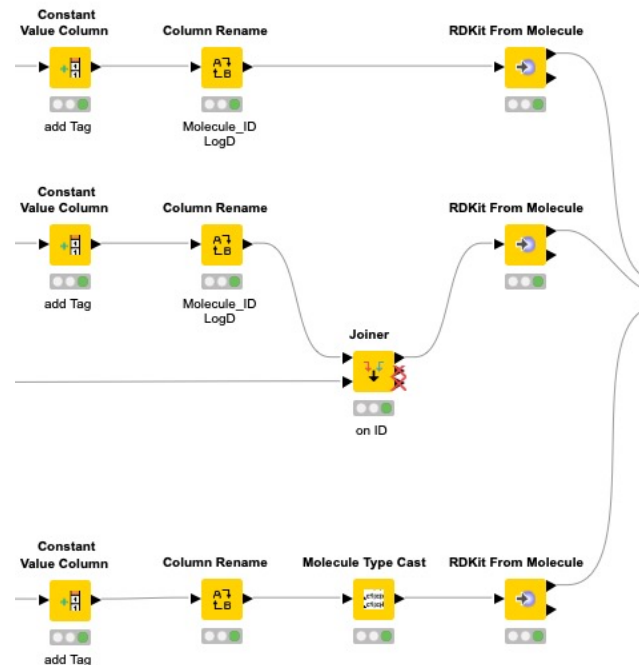
# Reminder of the task

---

- 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 colleagues

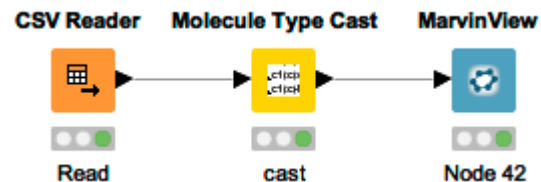
# Tedious repetitive preprocessing tasks

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



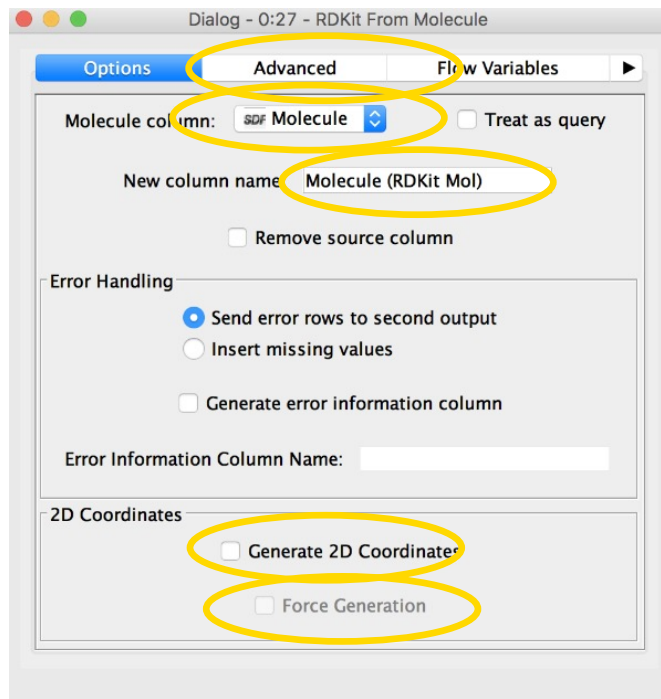
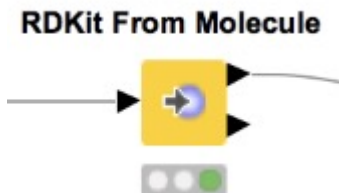
# Nodes for Type Manipulation

- **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.



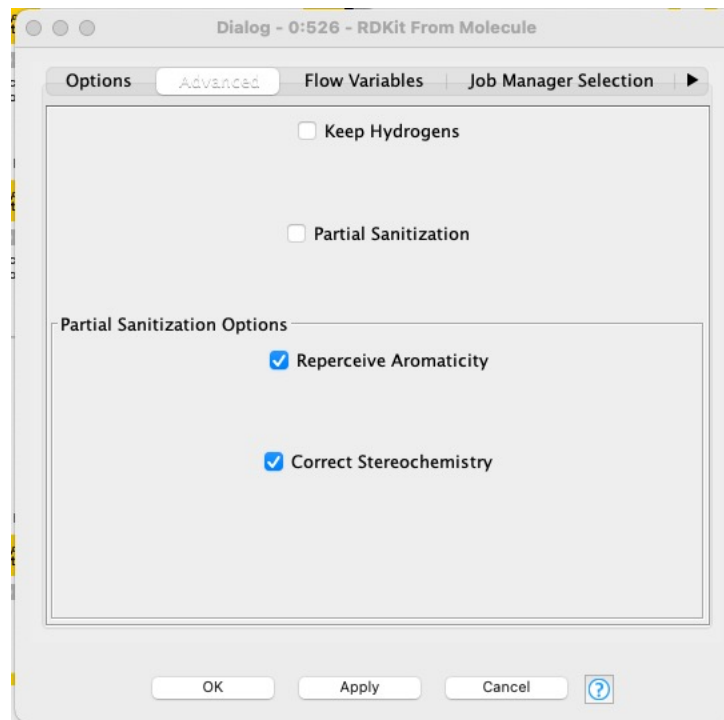
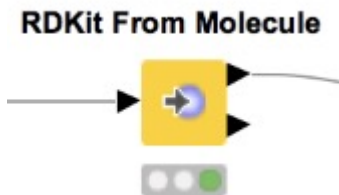
# RDKit From Molecule

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

Inner 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

Left Outer 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

Right 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

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)

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

Join by Mol Reg No

Full Outer Join

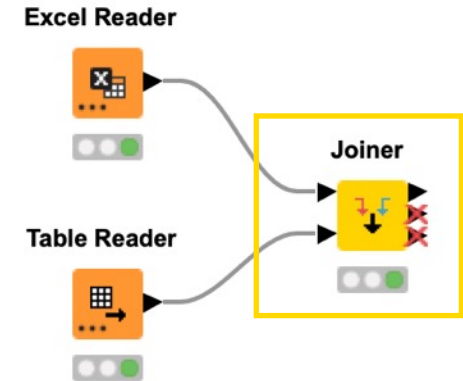
Mol Reg No	Chembl ID	SMILES	Ki value	Ki relation	Ki unit
17	?	?	76.0	=	nM
65	?	?	6.56	=	nM
35	?	?	100	>	nM
15	CHEMBL1794855	CCCN(CCC)	8	=	nM
10	CHEMBL278751	CCN(C)	95.8	=	nM
22	CHEMBL103772	CCCN1CC	?	?	?
24	CHEMBL328107	C1CN(CCN1)	?	?	?

Missing values in the left table

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

The screenshot shows the 'Joiner Settings' dialog box in KNIME. The 'Join columns' section is highlighted with a yellow box and contains the following elements:

- Match:** ☒ all of the following ☐ any of the following
- Top Input ('left' table):** A dropdown menu showing 'StoreID'.
- Bottom Input ('right' table):** A dropdown menu showing 'StoreID'.
- Buttons:** '+', '-', and a second '+' button.

Below the 'Join columns' section, the 'Compare values in join columns by' section has three radio buttons: ☒ value and type, ☐ string representation, and ☐ making integer types compatible.

The 'Include in output' section has three checked checkboxes: ☒ Matching rows, ☒ Left unmatched rows, and ☒ Right unmatched rows. To the right of these checkboxes is a Venn diagram labeled 'Full outer join' showing two overlapping circles.

The 'Output options' section has three checkboxes: ☒ Route unmatched rows to separate ports, ☐ Merge join columns, and ☐ Hiliting enabled.

The 'Row Keys' section has two radio buttons: ☒ Concatenate original row keys with separator \_ and ☐ Assign new row keys sequentially.

At the bottom of the dialog are buttons for 'OK', 'Apply', 'Cancel', and a help icon.

# Joiner Configuration – Column Selection

Dialog - 0:303 - Joiner

Joiner Settings **Column Selection** Performance Flow Variables Memory Policy

Top Input (left table)

☒ Manual Selection ☐ Wildcard/Regex Selection ☐ Type Selection

Exclude

Filter

No columns in this list

☒ Enforce exclusion

Include

Filter

☒ Enforce inclusion

Bottom Input (right table)

☒ Manual Selection ☐ Wildcard/Regex Selection ☐ Type Selection

Exclude

Filter

☒ Enforce exclusion

Include

Filter

☐ Enforce inclusion

Duplicate column names

☐ Do not execute

☒ Append custom suffix (right)

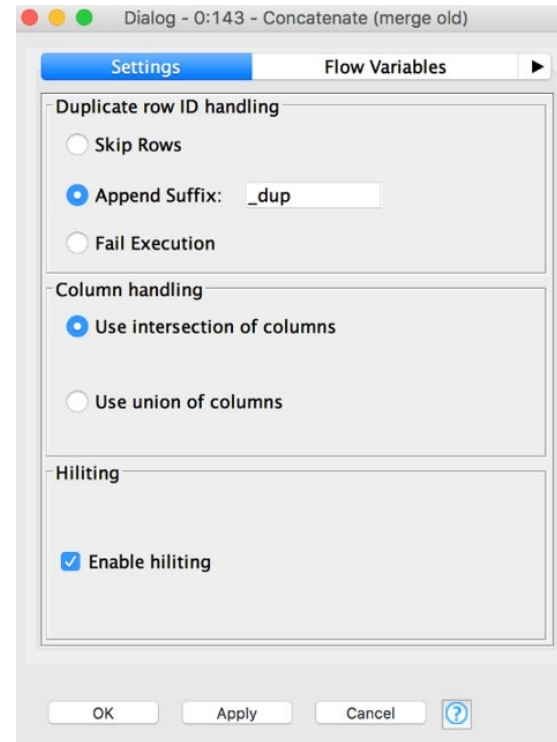
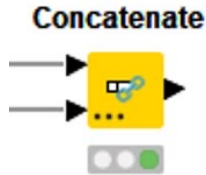
OK Apply Cancel ?

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

File Reader



Excel Reader



Add input port  
Remove input port

File Reader



Excel Reader



Concatenate



SDF Reader



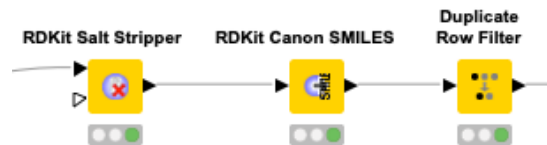
SDF Reader



# [Some] standardization

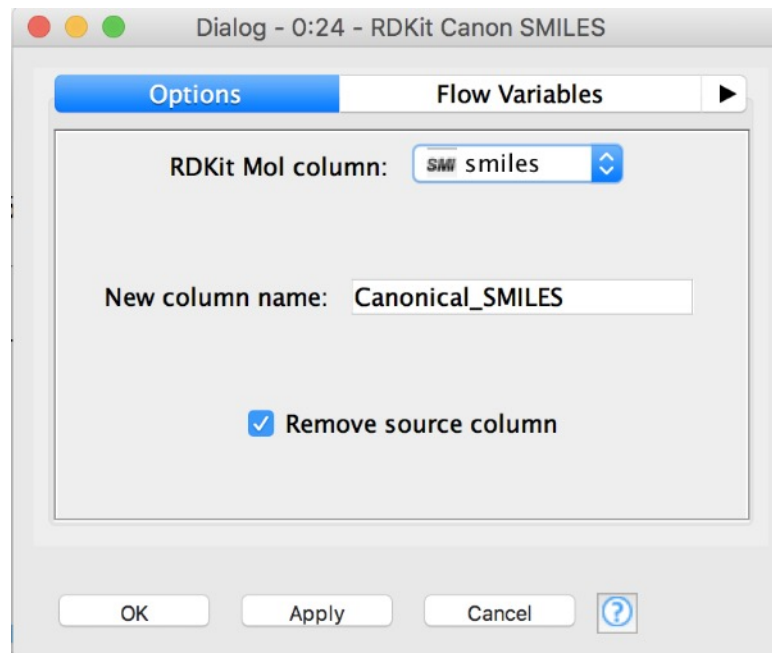
---

- Removing salts
- Generating canonical SMILES
- Removing duplicates



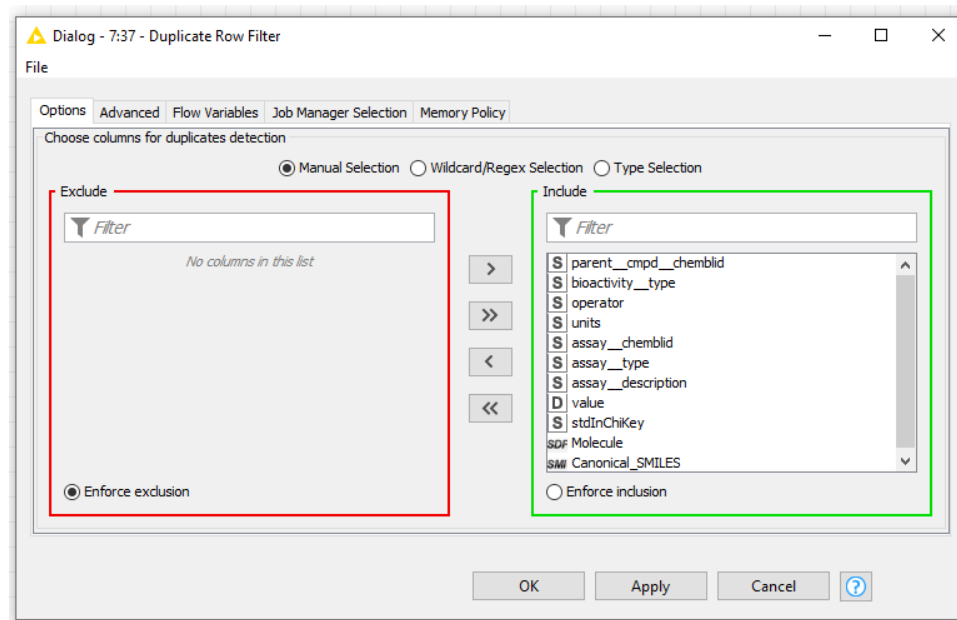
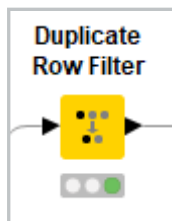
# Standardization

## Generate canonical SMILES





# Remove Duplicates

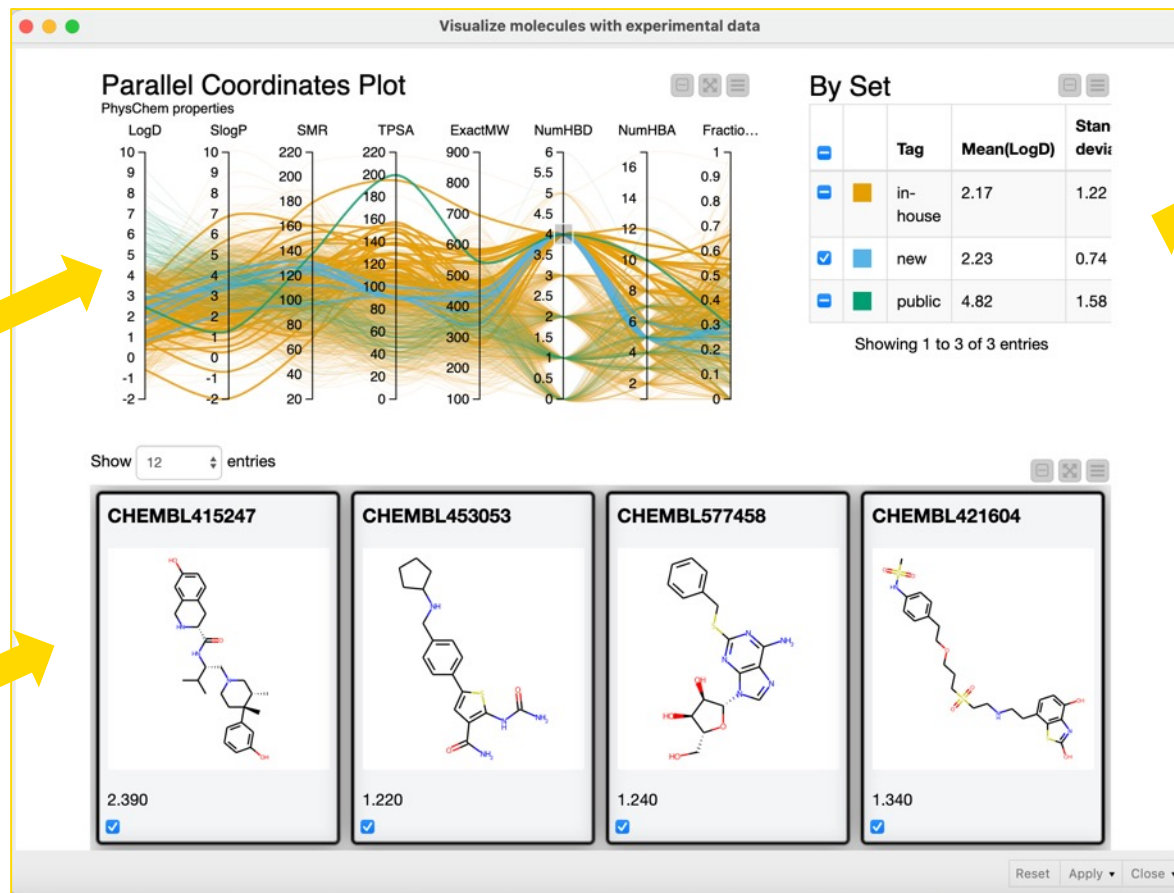


# Reminder of the task

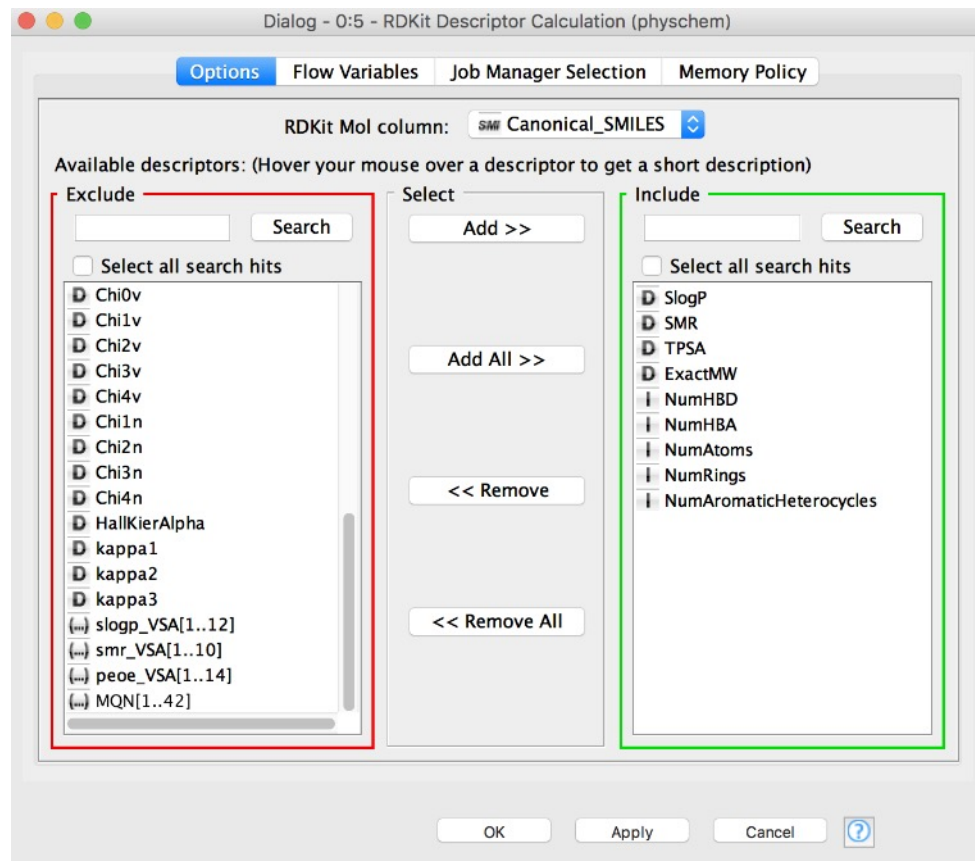
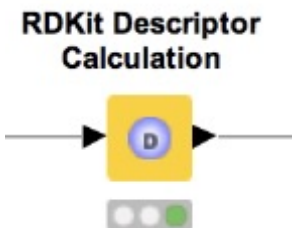
---

- 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 colleagues

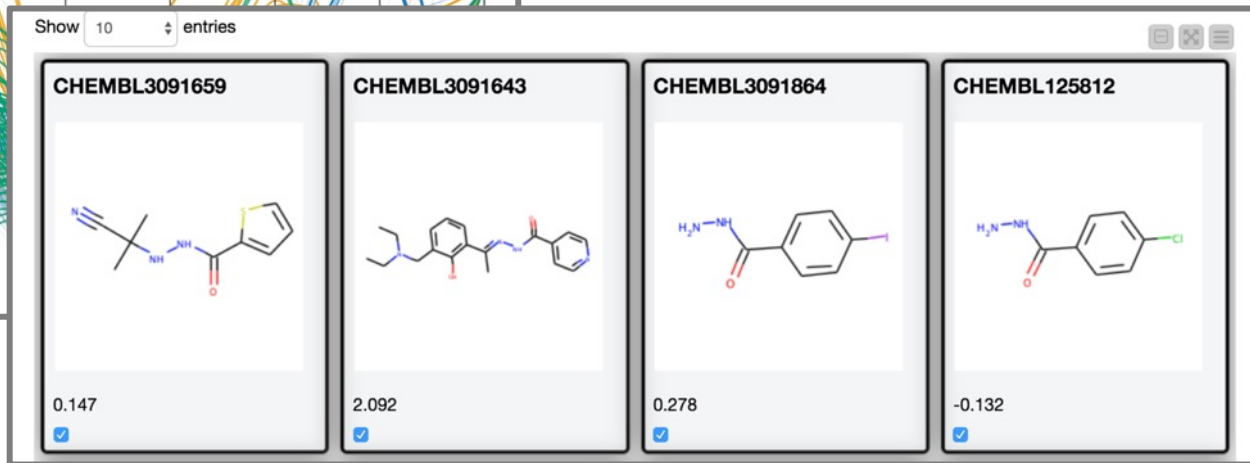
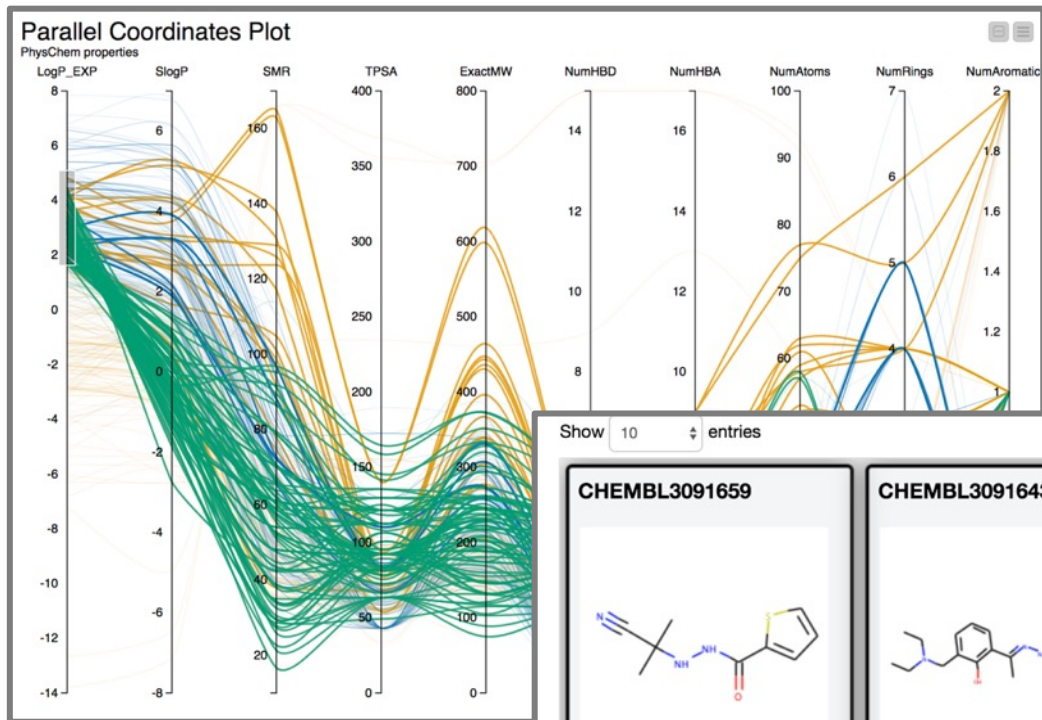
# Reminder of a possible interactive view



# Compute Descriptors

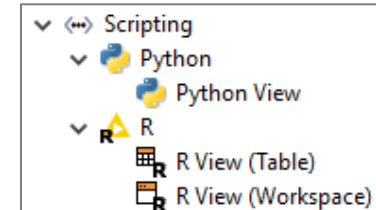
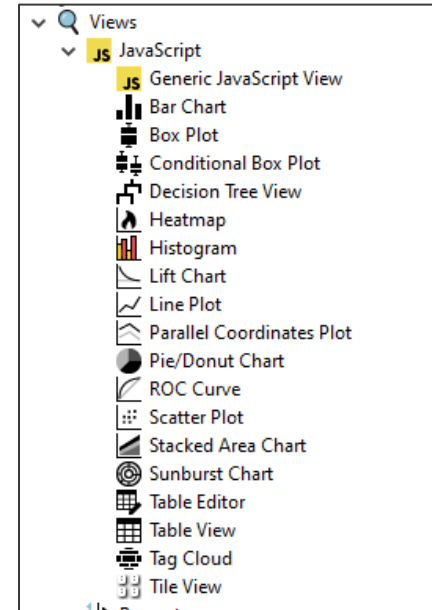


# Visualize Chemical Structures Together with other Views

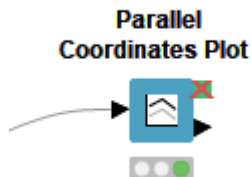


# Data Visualization

- 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



Dialog - 0:25 - Parallel Coordinates Plot (JavaScript)

Options   General Plot Options   Control Options   Selection and Filter   Flow Variables

General Settings

☐ Generate image

Maximum number of rows 2,500

☒ Manual Selection   ☐ Wildcard/Regex Selection

Column(s):  Search

☐ Select all search hits

mol\_chemblid  
assay\_id

add >>

add all >>

<< remove

<< remove all

☒ Enforce exclusion

Column(s):  Search

☐ Select all search hits

LogP\_EXP  
SlogP  
SMR  
TPSA  
ExactMW  
NumHBD  
NumHBA  
NumAtoms  
NumRings  
NumAromaticHeterocycles

☐ Enforce inclusion

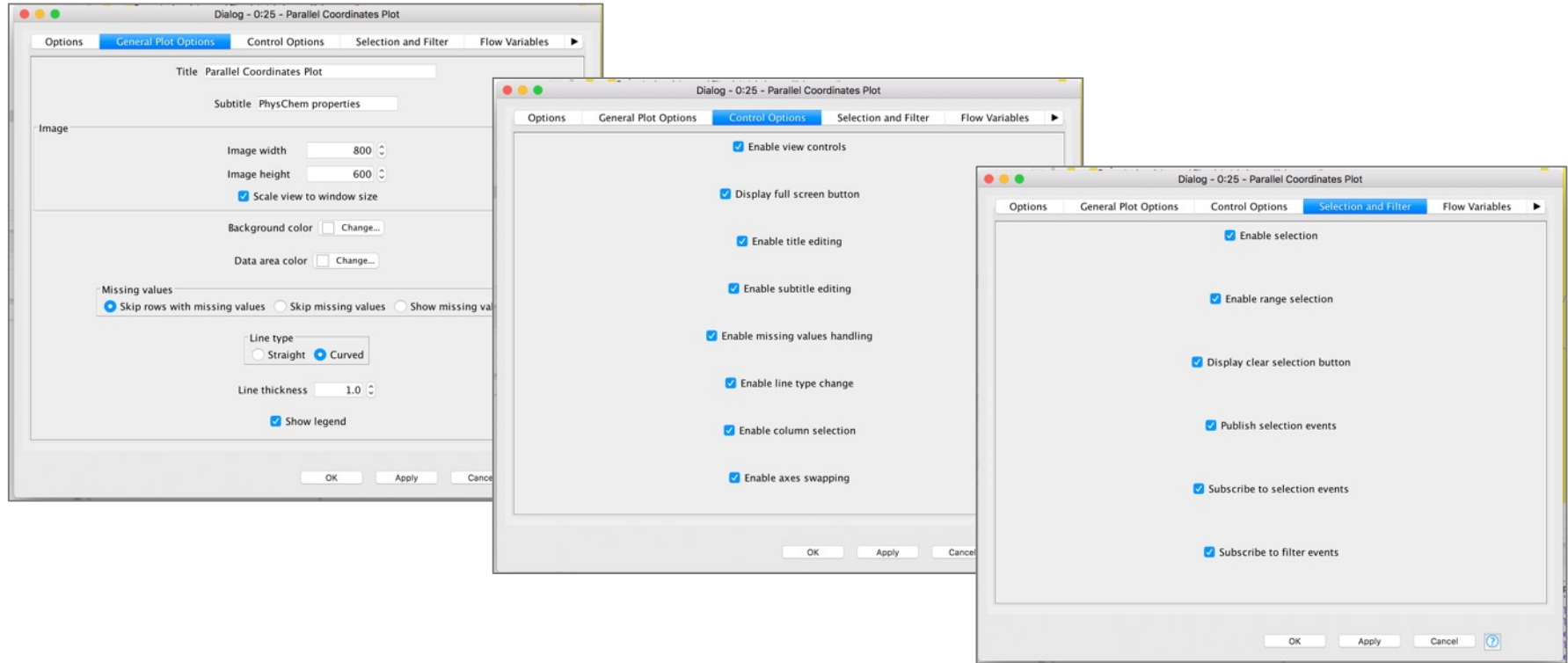
☒ Use colors from spec

Color Column ? <none>

OK   Apply   Cancel   ?

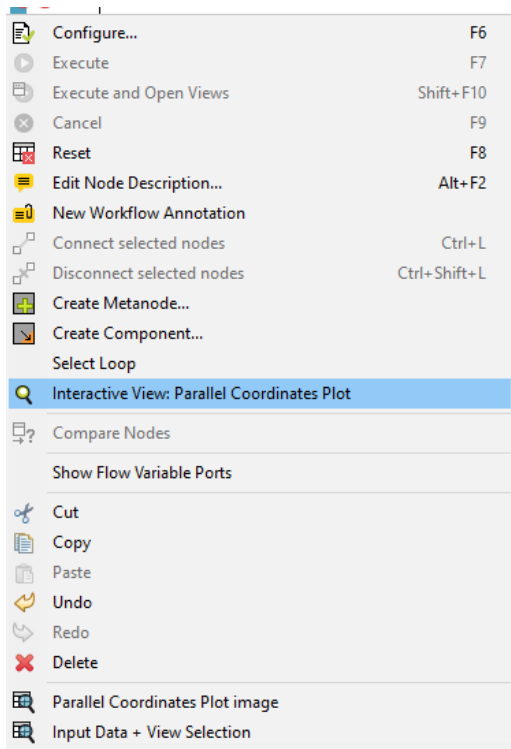
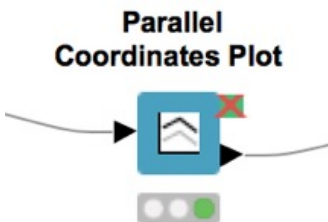
# Parallel Coordinates Plot

- Additional configuration tabs

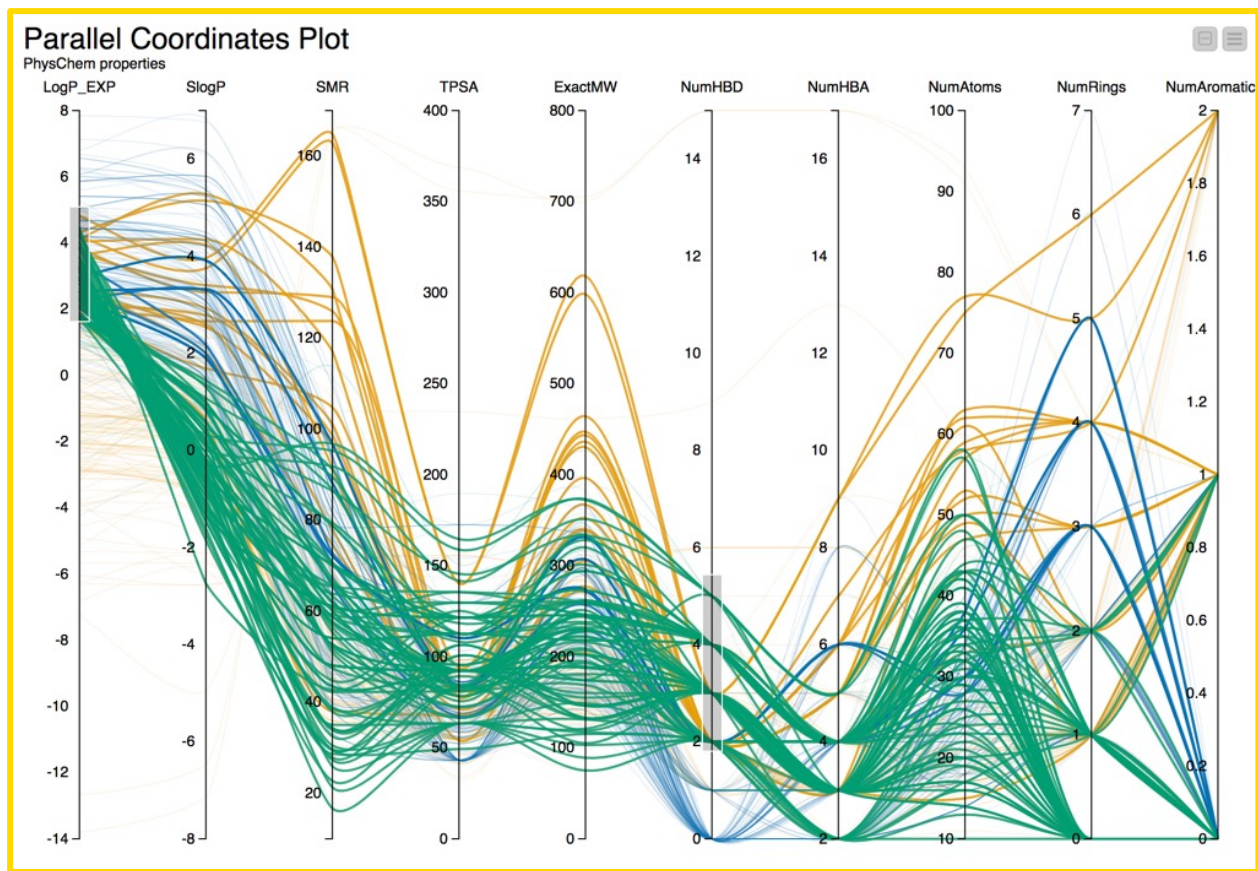
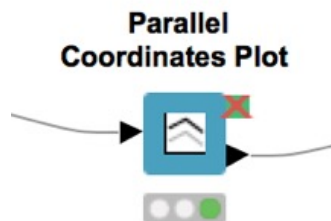




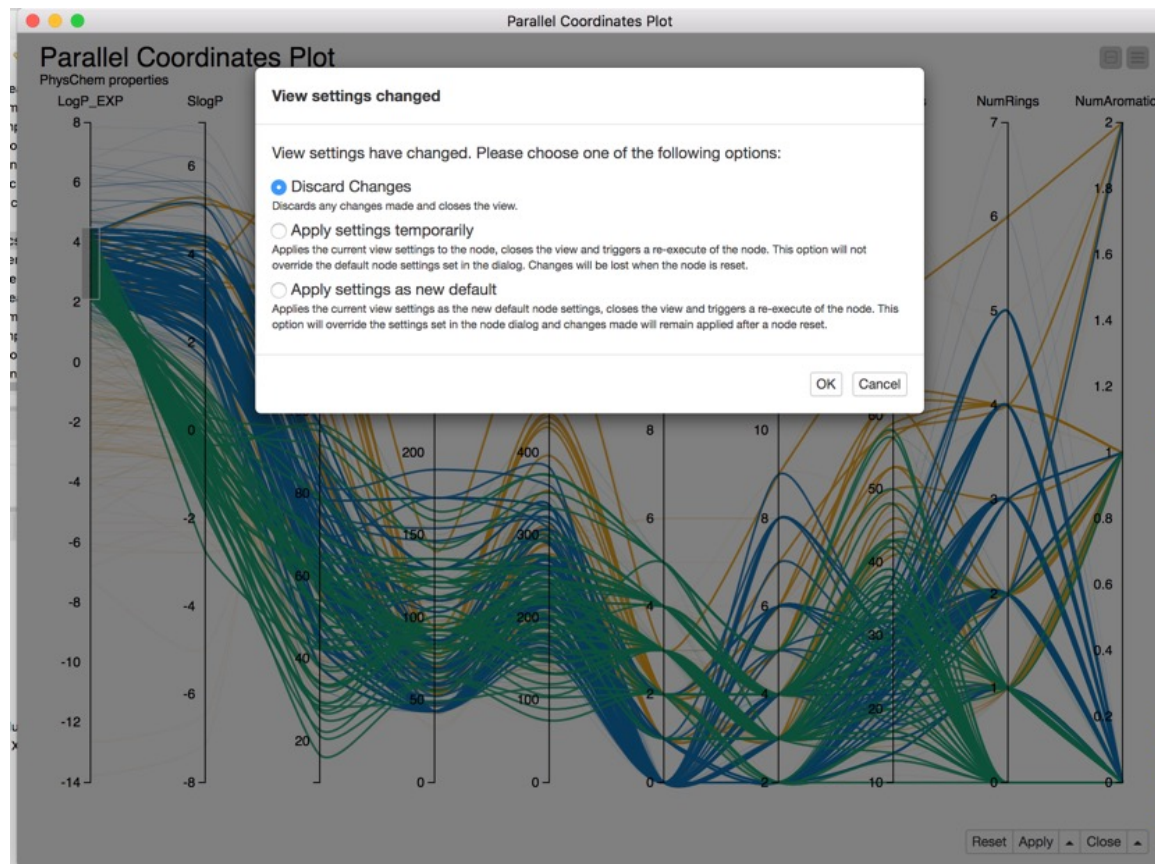
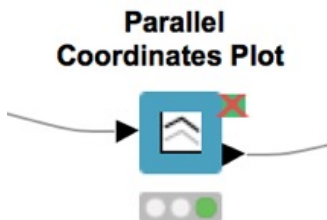
# Parallel Coordinates Plot



# Interactively Filter on Multiple Properties



# Parallel Coordinates Plot

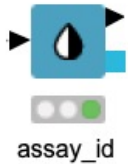


# Color Manager

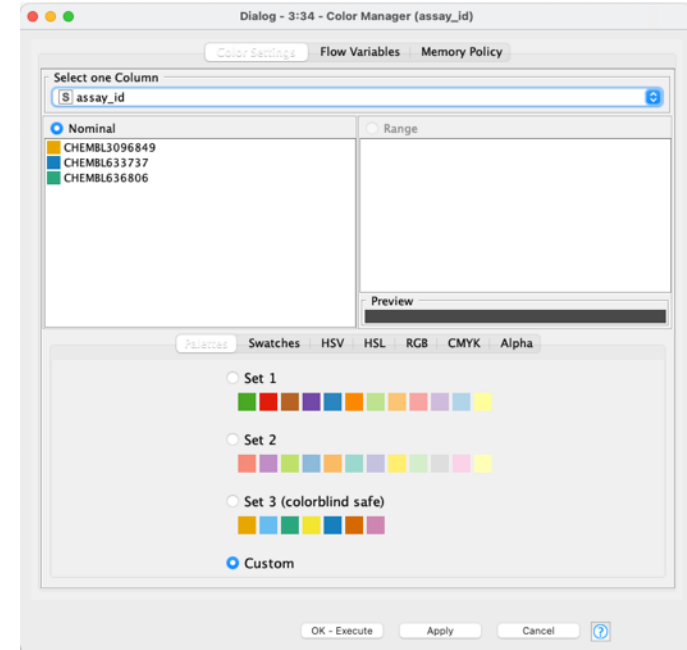
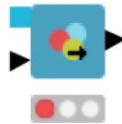
One of several visual property managers (e.g. size, shape)

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

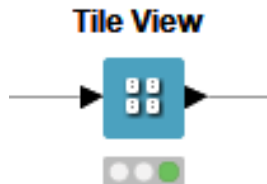
Color Manager



Color Appender



# Tile View



Dialog - 0:85:0:74 - Tile View (mol)

Options Interactivity Formatters Flow Variables Memory Policy

**General Options**

No. of rows to display: 100,000

Title:

Subtitle:

**Display Options**

☐ Display row colors ☒ Display column headers ☒ Display fullscreen button

☒ Fixed number of tiles per row (1 - 100) 1

☐ Fixed tile width (30 - 5000px) 180

Select text alignment: ☒ Left ☐ Center ☐ Right

Choose a title column:  
S compound\_chembl\_id

Columns to display:

☒ Manual Selection ☐ Wildcard/Regex Selection ☐ Type Selection

**Exclude**

Filter

- canonical\_smiles (RDKit Mol)
- S compound\_chembl\_id
- L molregno
- S standard\_relation
- S assay\_chembl\_id
- S description
- R1

☐ Enforce exclusion

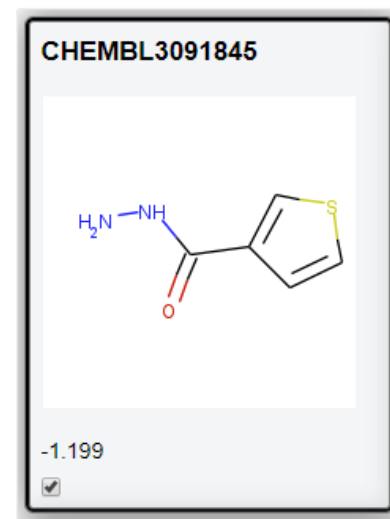
**Include**

Filter

- D pIC50
- Mol\_svg

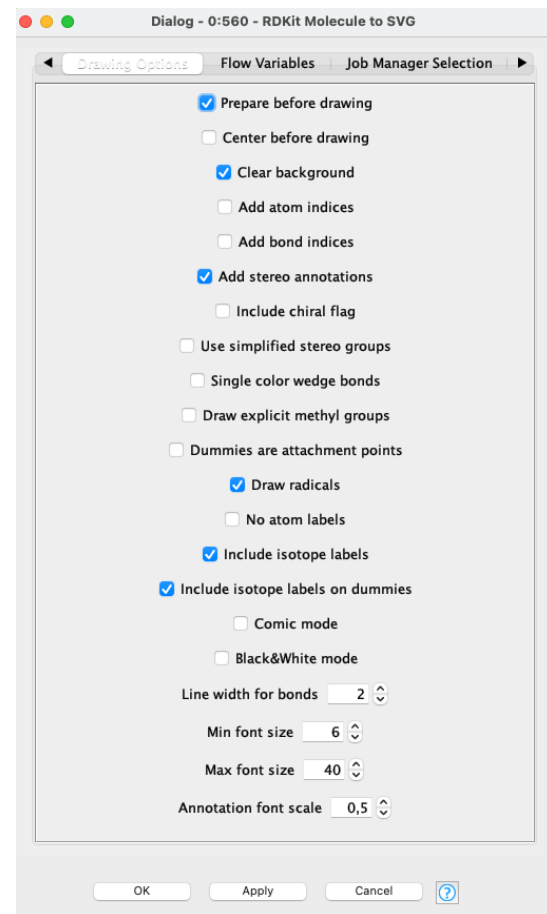
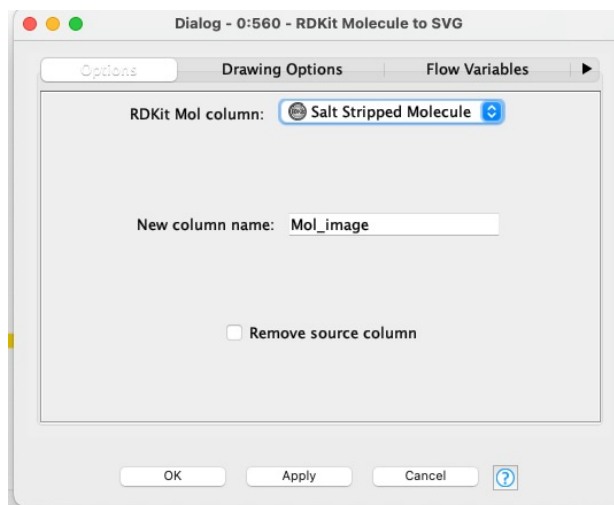
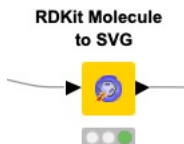
☒ Enforce inclusion

OK Apply Cancel ?

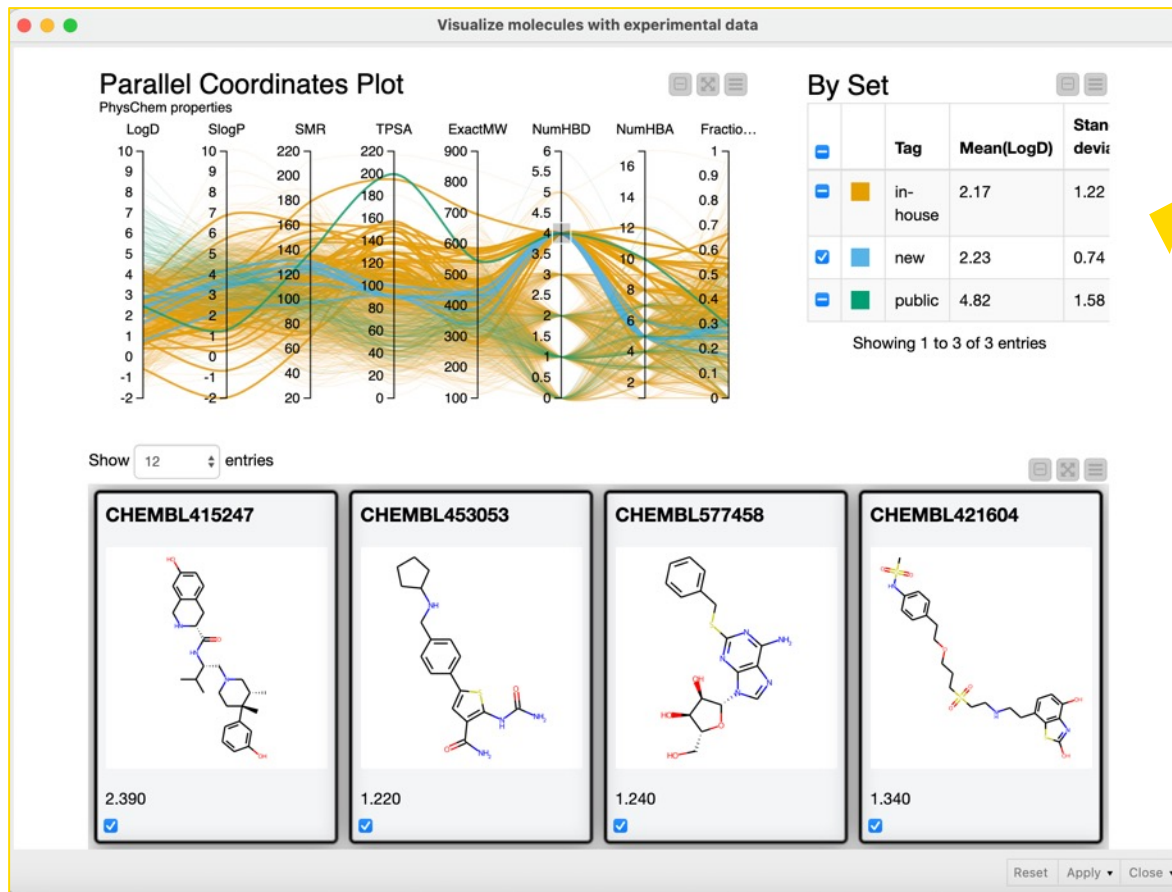


# RDKit molecule to SVG

- 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



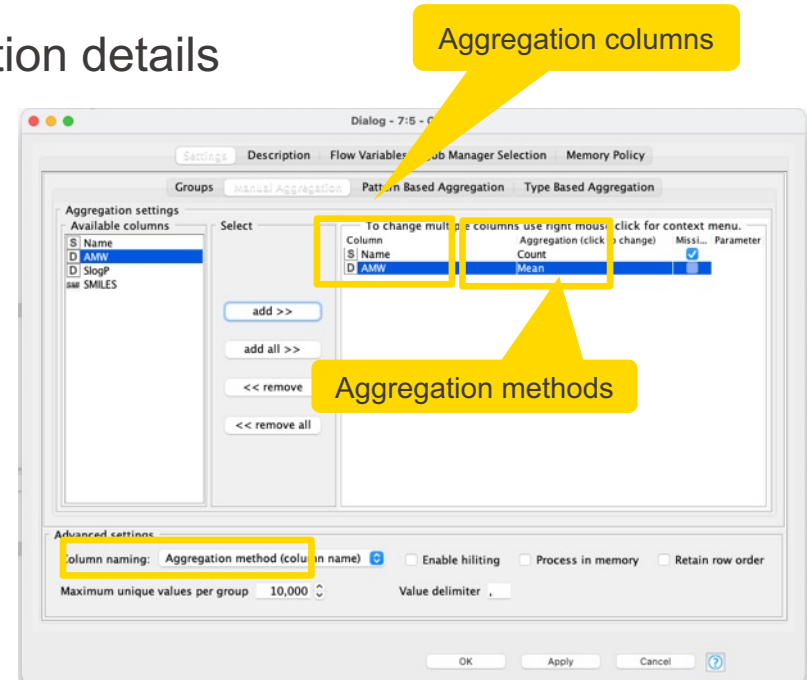
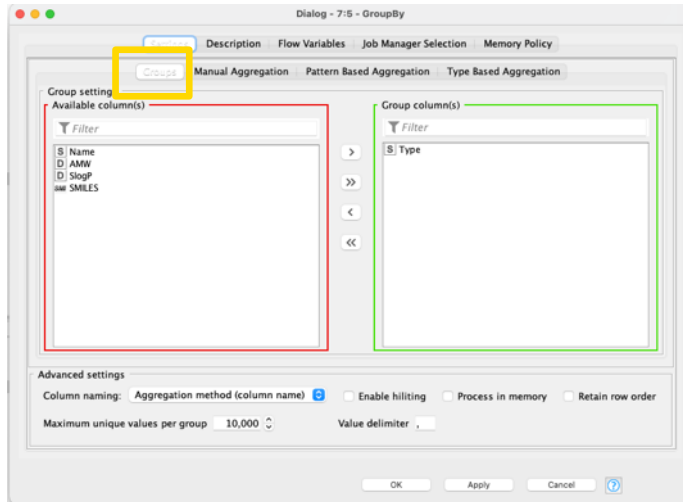
Type	Count(Name)	Mean(Weight)
NSAID	4	208.44
PPI	2	364.40
SSRI	4	326.45



# GroupBy

Aggregate to summarize data

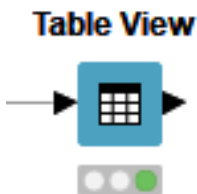
- First tab provides grouping options
- Second tab provides control over aggregation details



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



JavaScript Table View

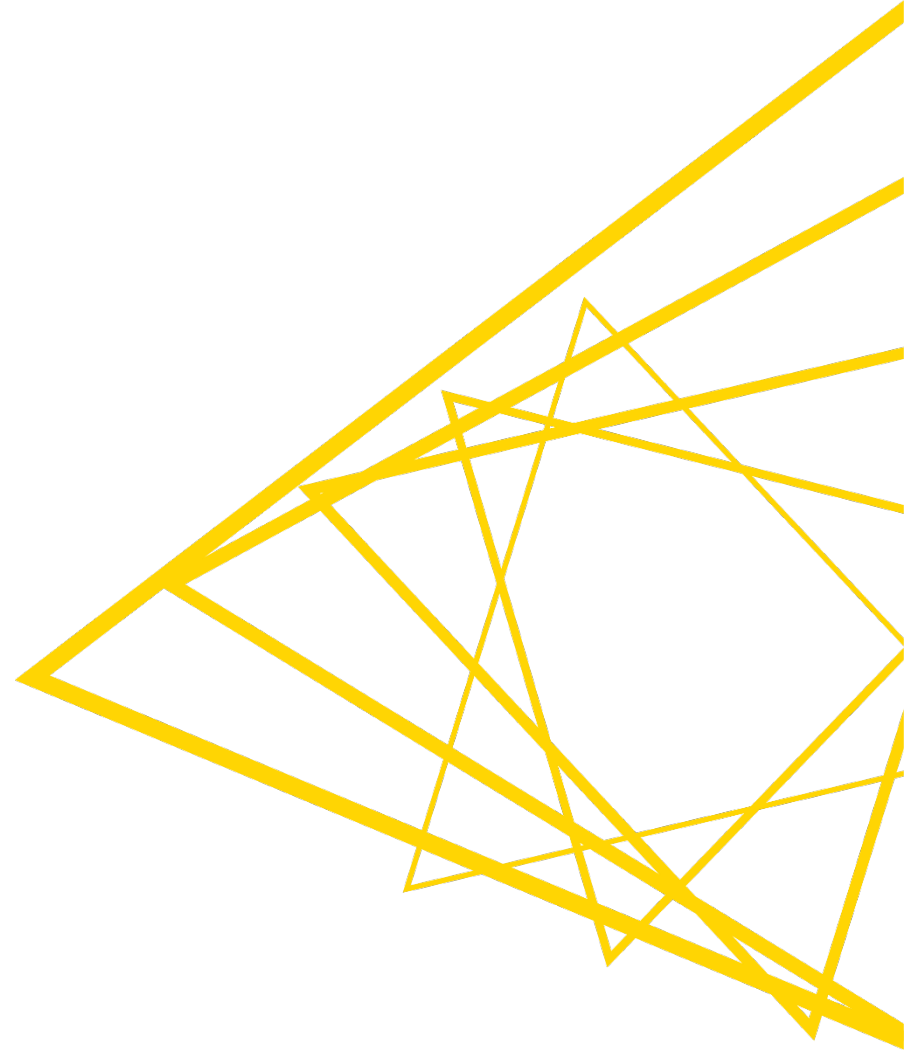
Show 10 entries

Search:

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

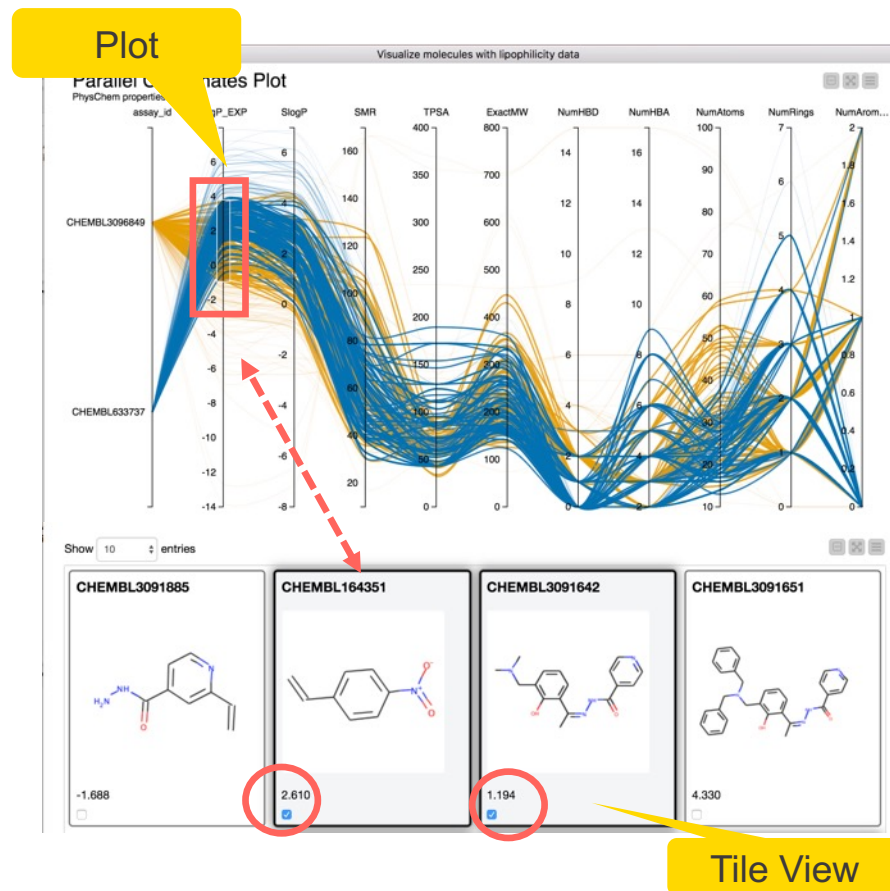
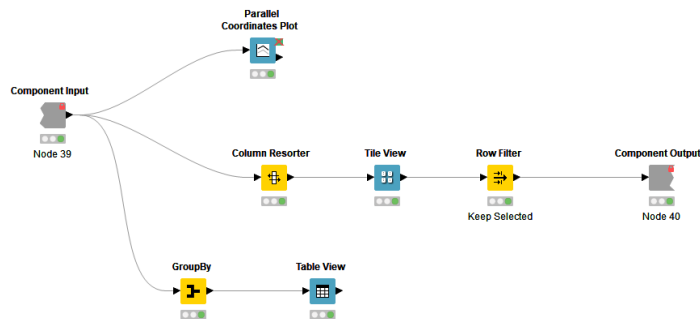
Loading data (28710 of 29170 records) - Displaying 1 to 10 of 29170 entries.

# Composite views

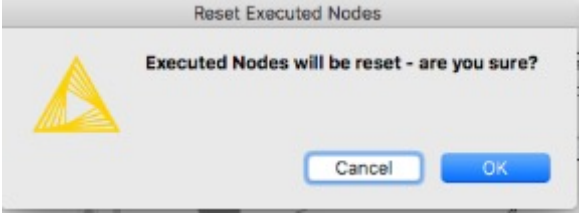
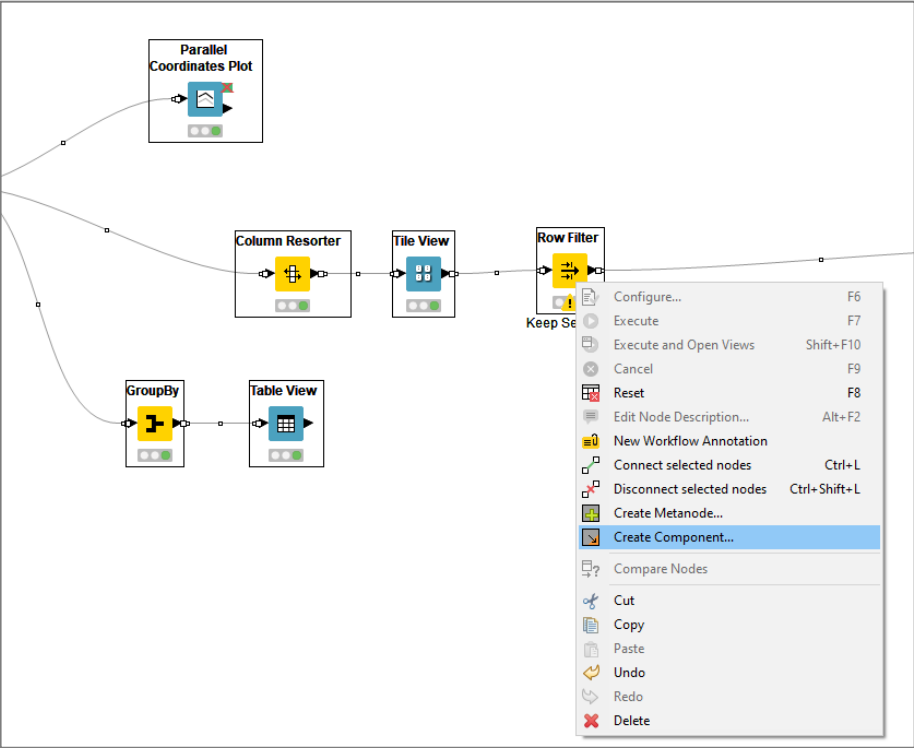


# Components – Combined Views

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



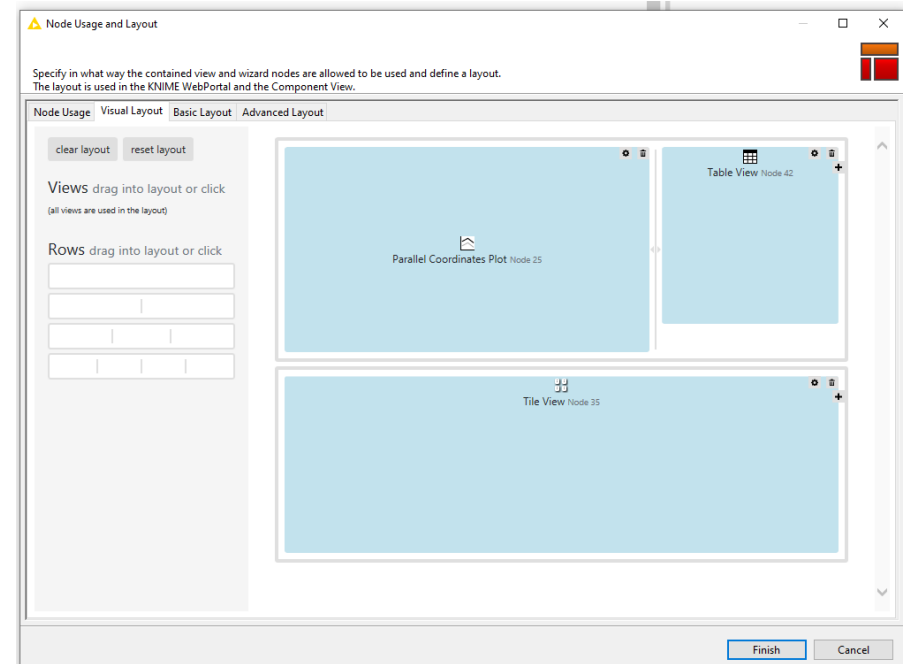
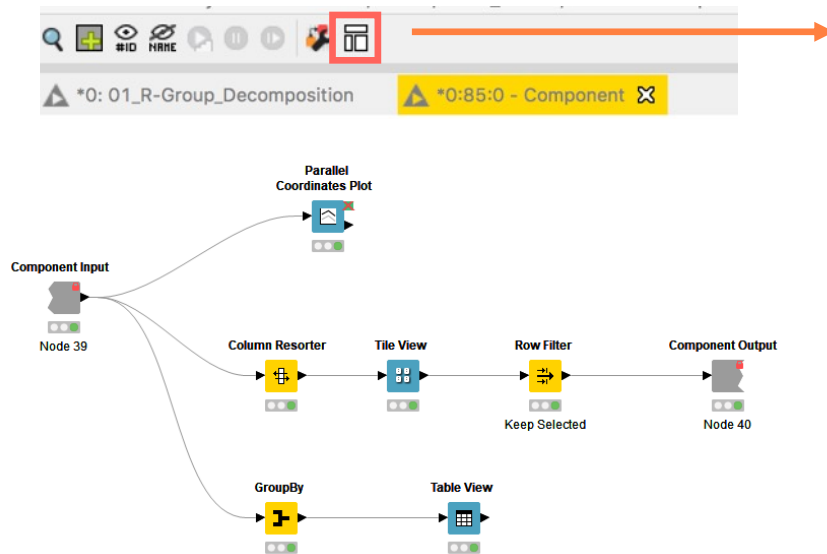
# Create a Component



# Configure Content and Views Layout

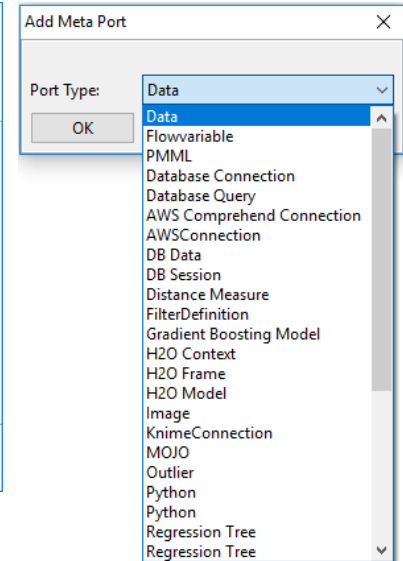
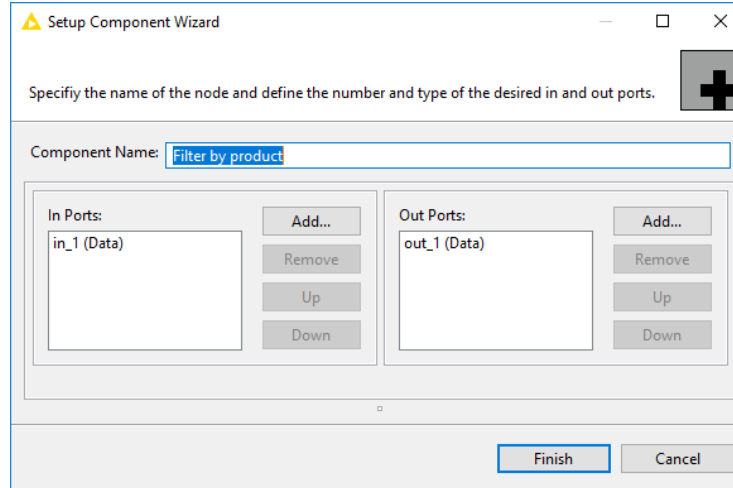
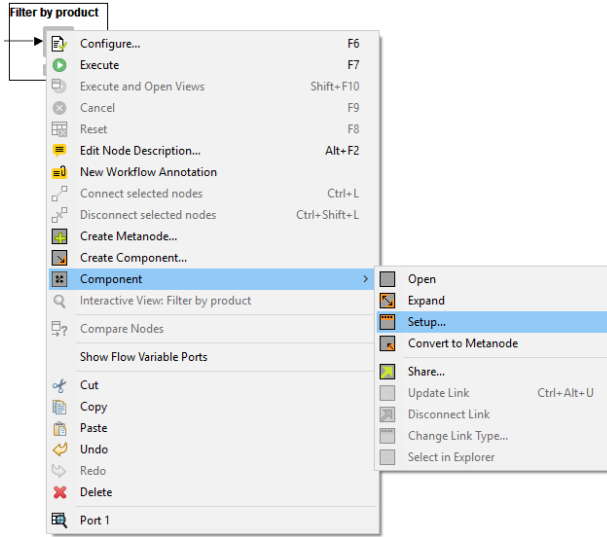
- Click layout button when inside component to assign views to rows and columns

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



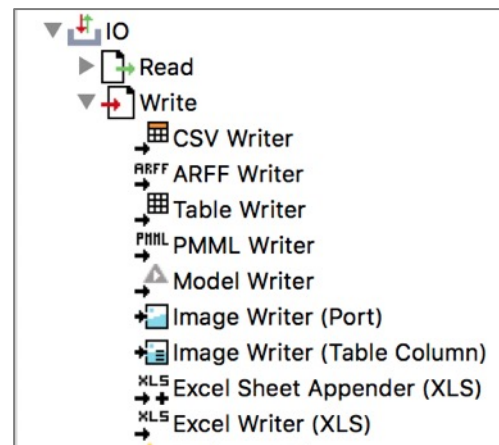
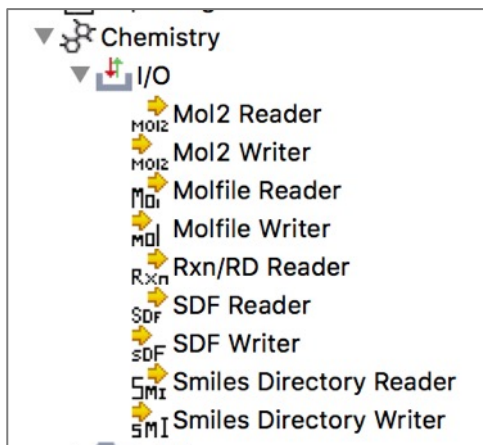
# Configure Component Ports

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



# Saving Files with Writer Nodes

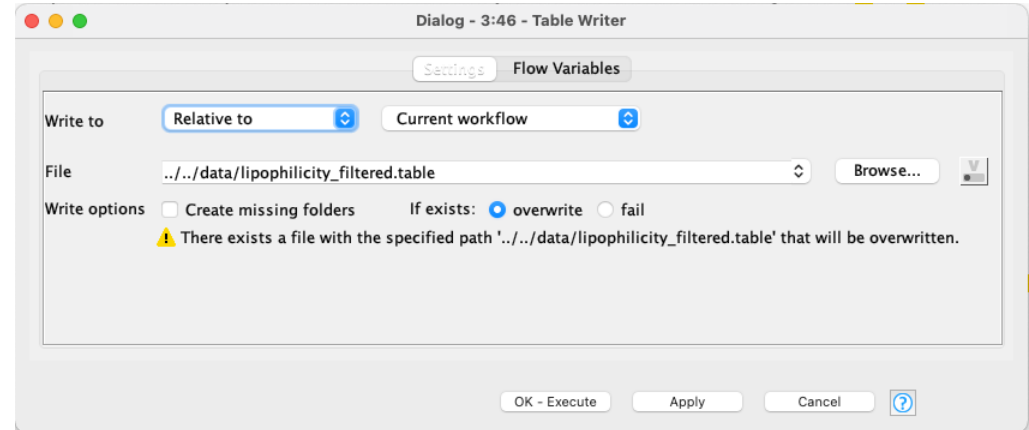
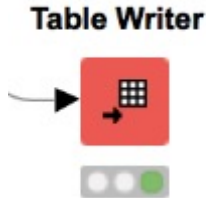
- Writers for chemical formats
- Writers for traditional data types





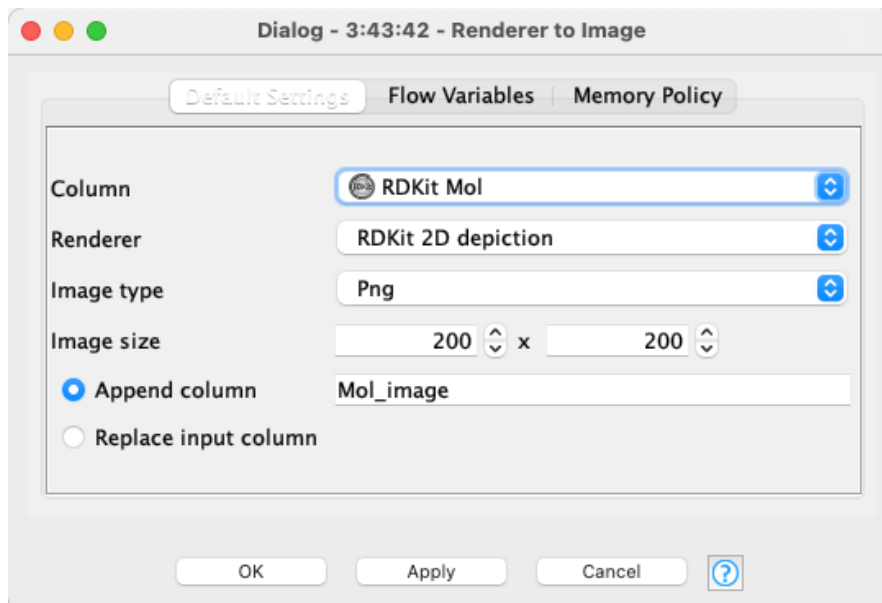
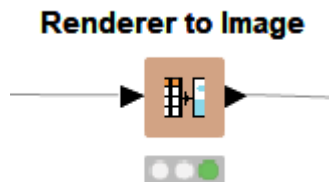
# Table Writer

- 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](https://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 & Selenium

Vernalis

Sequan

**Partner Extensions**

For discussions related to extensions developed by partners

JChem Extensions

Schrödinger

**KNIME Server**

For discussions related to KNIME Server

**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 files

M

M

Marten\_Pfennenschmidt

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 Document Data Extractor, or use the Text Processing nodes which can work with Document Columns.

[hub.knime.com](https://hub.knime.com)

KNIME

data

KNIME Hub • Search

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. Available pre-processing steps: - Autom...

knime > Examples > 02\_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

Source

**02\_Hive\_WritingtoDB\_Exercise**

KNIME

Download workflow

By downloading the workflow, you agree to our terms and conditions.

© 2020 KNIME AG

Short link

https://www.knime.com/workflows/02\_Hive\_WritingtoDB\_Exercise

Big Data Course Homework Exercise #2

Make sure you have executed the 02\_Hive\_WritingtoDB\_Exercise/02\_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!**

