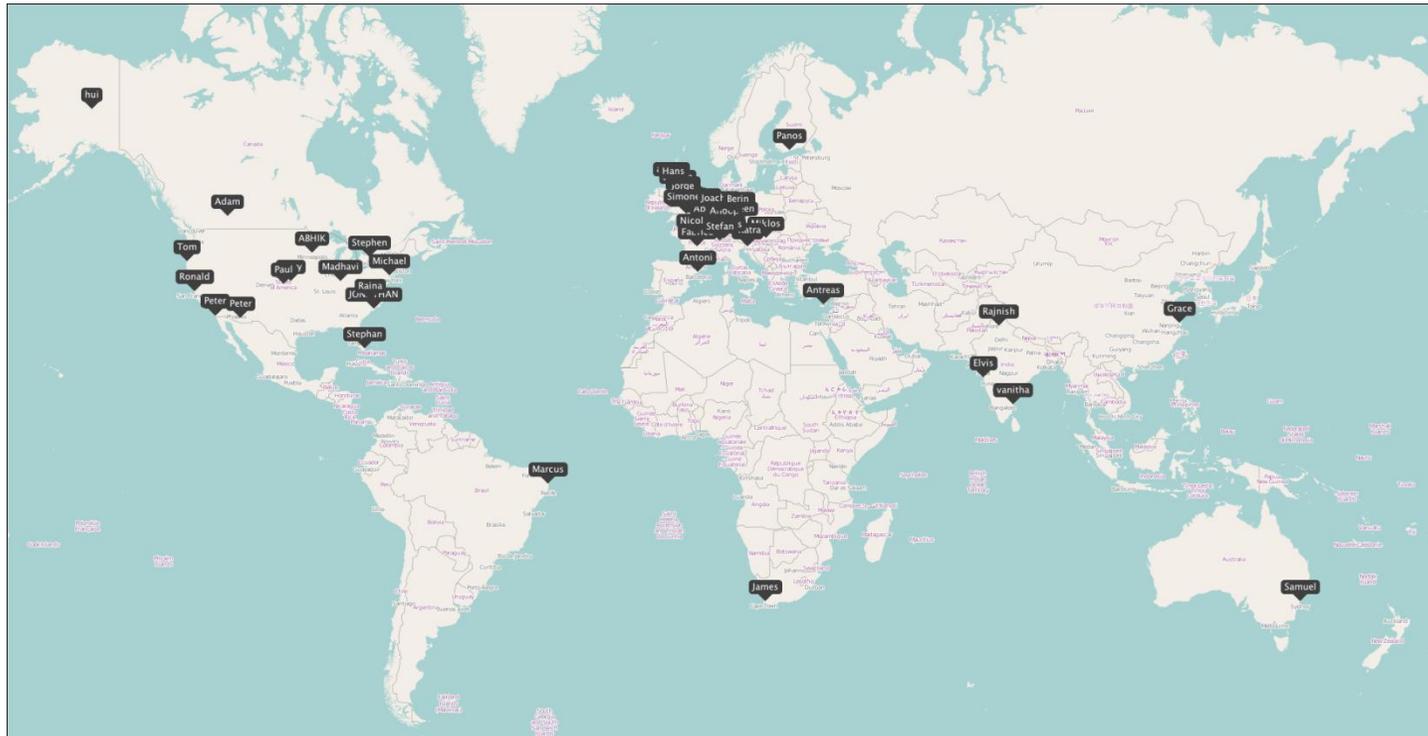




Chemistry in KNIME

For questions and suggestions please contact education@knime.com

Welcome!



- Aaron Hart will be presenting Rosaria Silipo will be moderating.
- Please mute your mic and ask questions by chat.
- This meeting is being recorded, and the video will be posted soon.

Goal of the Course

- Survey KNIME's basic concepts
- Sketch a chemical structure into a KNIME table
- Convert between chemistry types
- Read/write and manipulate in SDF
- Calculate molecular descriptors with RDKit
- Access ChEMBL via webservices (Optional)

Course Material

Prerequisite -> www.knime.org

- KNIME 2.7 or later
- KNIME Extensions (Chemistry, including Marvin + Reporting)
- KNIME Community extensions (including EBI and RDKit nodes)

Supplementary -> www.knime.org/files/Intro_to_KNIME_Chem.zip

- Data
- Example workflows
- These slides

Additional Resources

- **KNIME** pages (www.knime.org)
 - **APPLICATIONS** for example workflows
 - **RESOURCES** with links to a number of resource pages, like downloads, updates, documentation, ...
- **KNIME Tech** pages (<http://tech.knime.org/>)
 - **FORUM** for questions and answers
 - **DOCUMENTATION** for documentation, FAQ, changelogs, ...
 - **LABS** where to find new experimental nodes
 - **COMMUNITY CONTRIBUTIONS** for development instructions and third party nodes



The KNIME Workbench

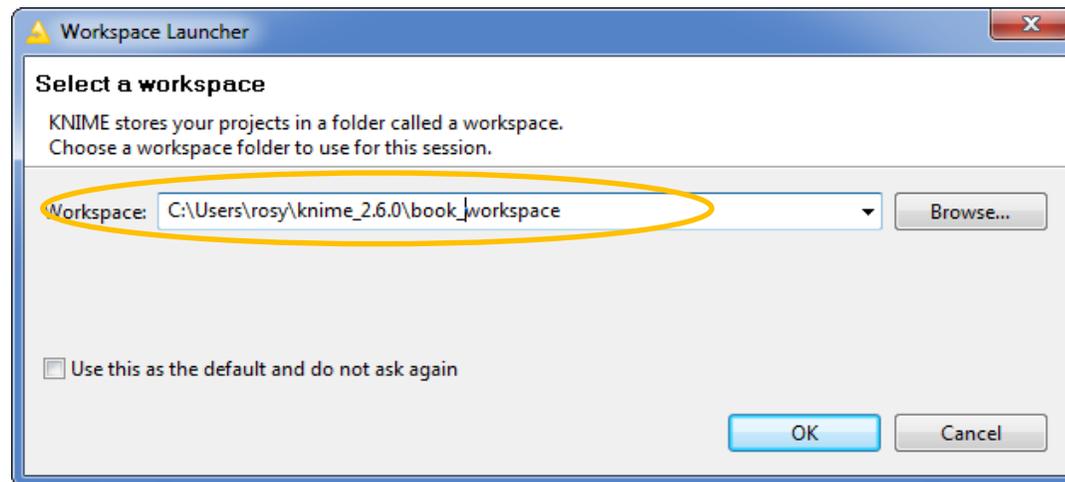
Overview of the Platform

Overview

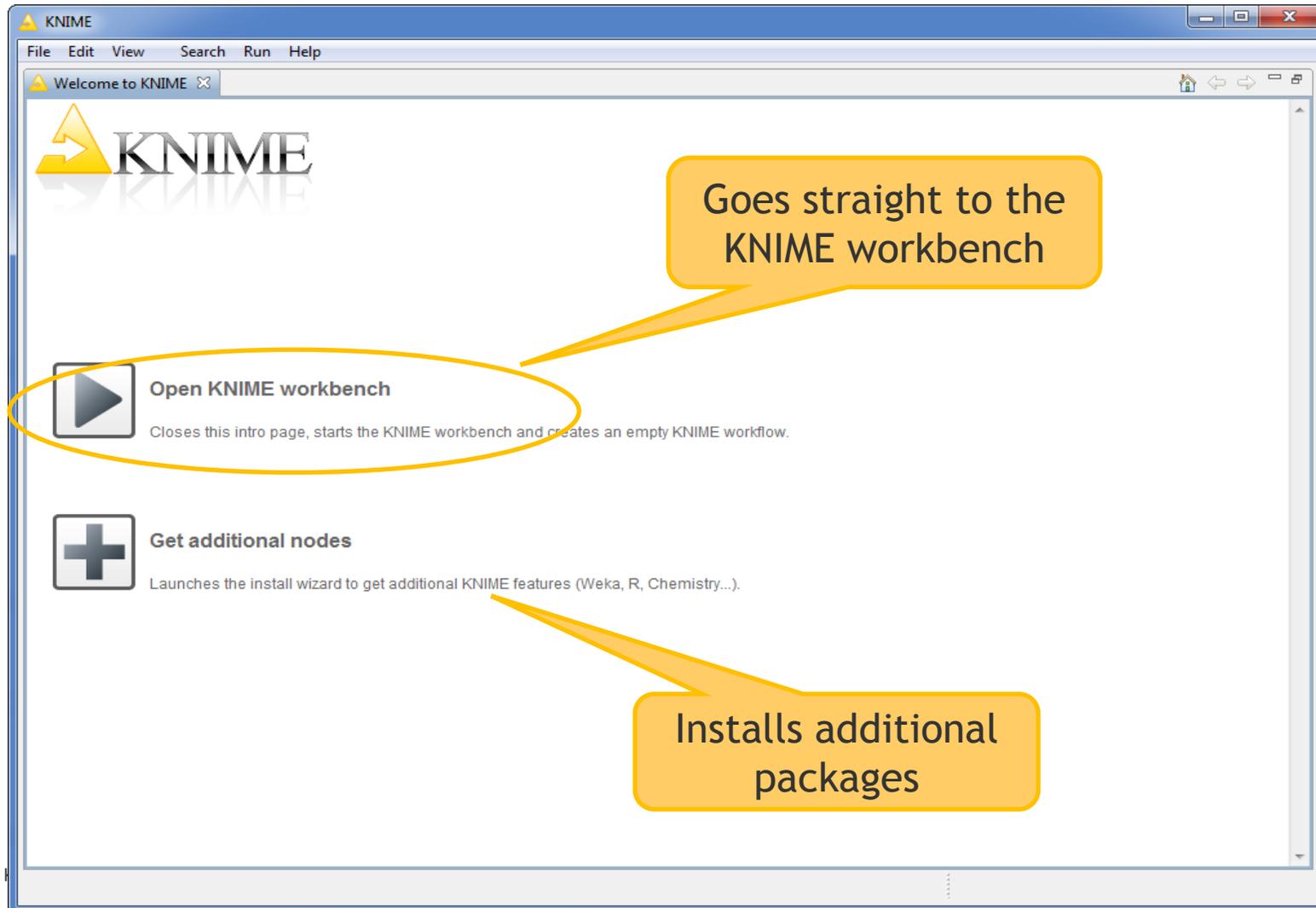
- The KNIME Workspace
- Installation of KNIME Extensions
- The KNIME Workbench
 - Help
 - Node Description

The Workspace

- The workspace is the **folder** in which work for the current KNIME session is stored.
- Using different workspaces for different projects keeps your KNIME work neat and clean.



First Time Start Screen



The Workbench



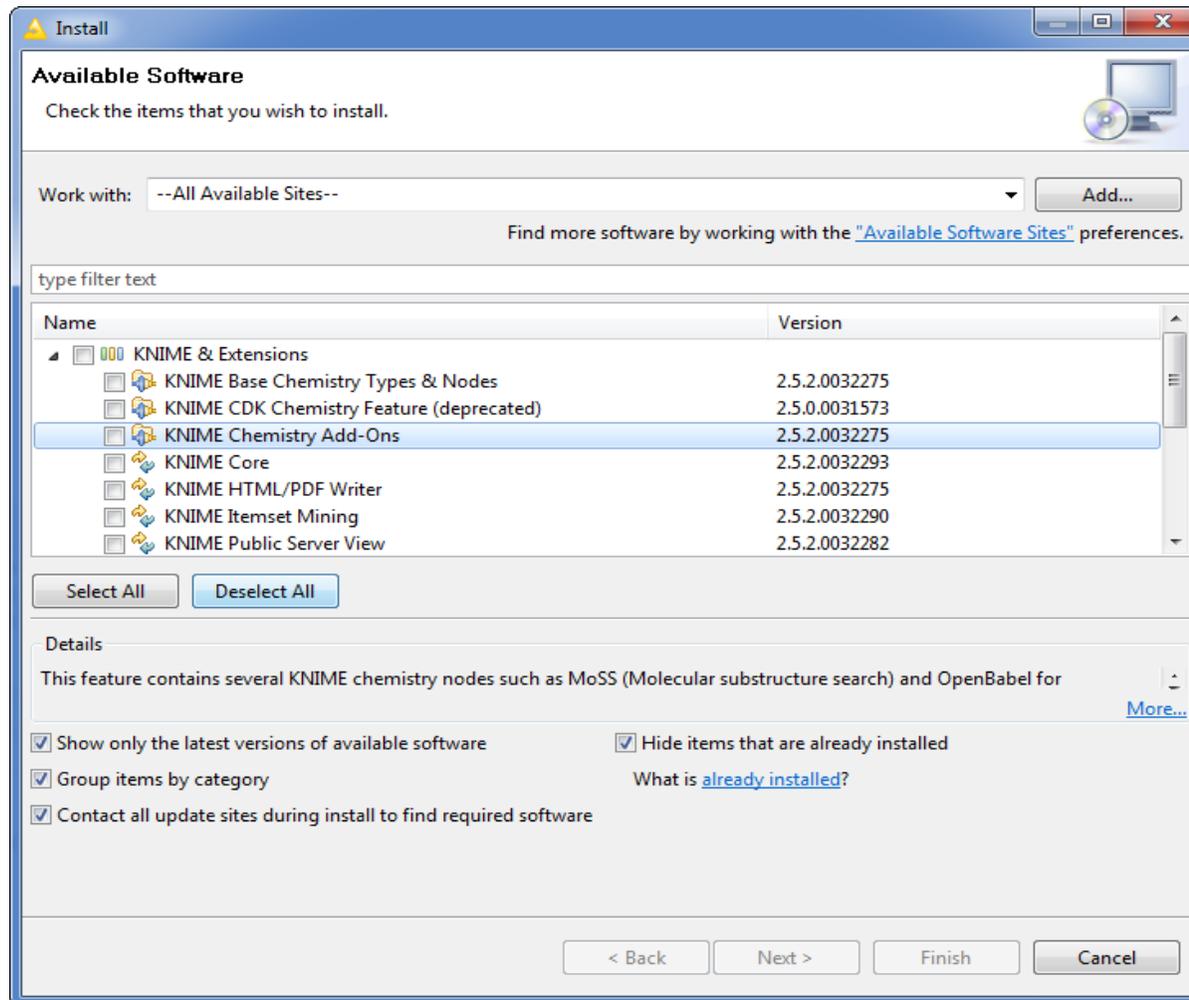
The screenshot shows the KNIME Workbench interface with several components annotated with red boxes and labels:

- Servers and Workflows:** Located in the KNIME Explorer on the left, showing a tree view of local and server-side projects.
- Workflow Editor:** The central workspace showing a workflow with nodes: File Reader (adult.data set), Column Filter (rm "final weight"), Row Filter (born outside US), and CSV Writer (write new file). A yellow text box below the workflow reads: "This workflow is my first KNIME workflow. It reads data, removes uninteresting columns and rows from the data table and writes the result to a CSV file."
- Node Description:** A panel on the right showing the "File Reader" node's description and "Dialog Options". A red arrow points from this panel to the workflow editor.
- Node Repository:** A panel on the left showing a categorized list of nodes such as IO, Database, Data Manipulation, etc.
- Outline:** A small thumbnail of the current workflow located at the bottom left.
- Console:** A panel at the bottom right showing the KNIME Console output, including a welcome message and a warning: "WARN CSV Writer Selected output file exists and will be overwritten!".

Install KNIME Extensions

- Click Help -> Install New Software
- Work with --All Available Sites--
- To add new sites, click Add
- Expand node **KNIME & Extensions**
- Select your extensions.
- Click Next and follow the Wizard instructions

Install KNIME Extensions

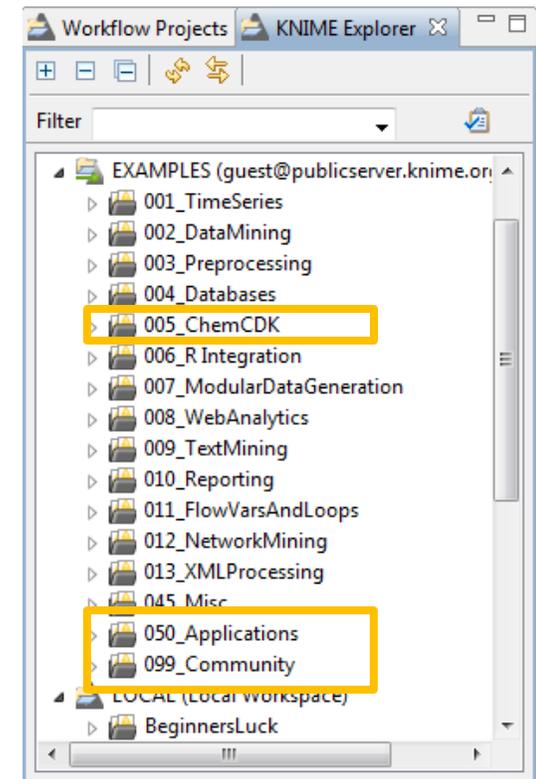


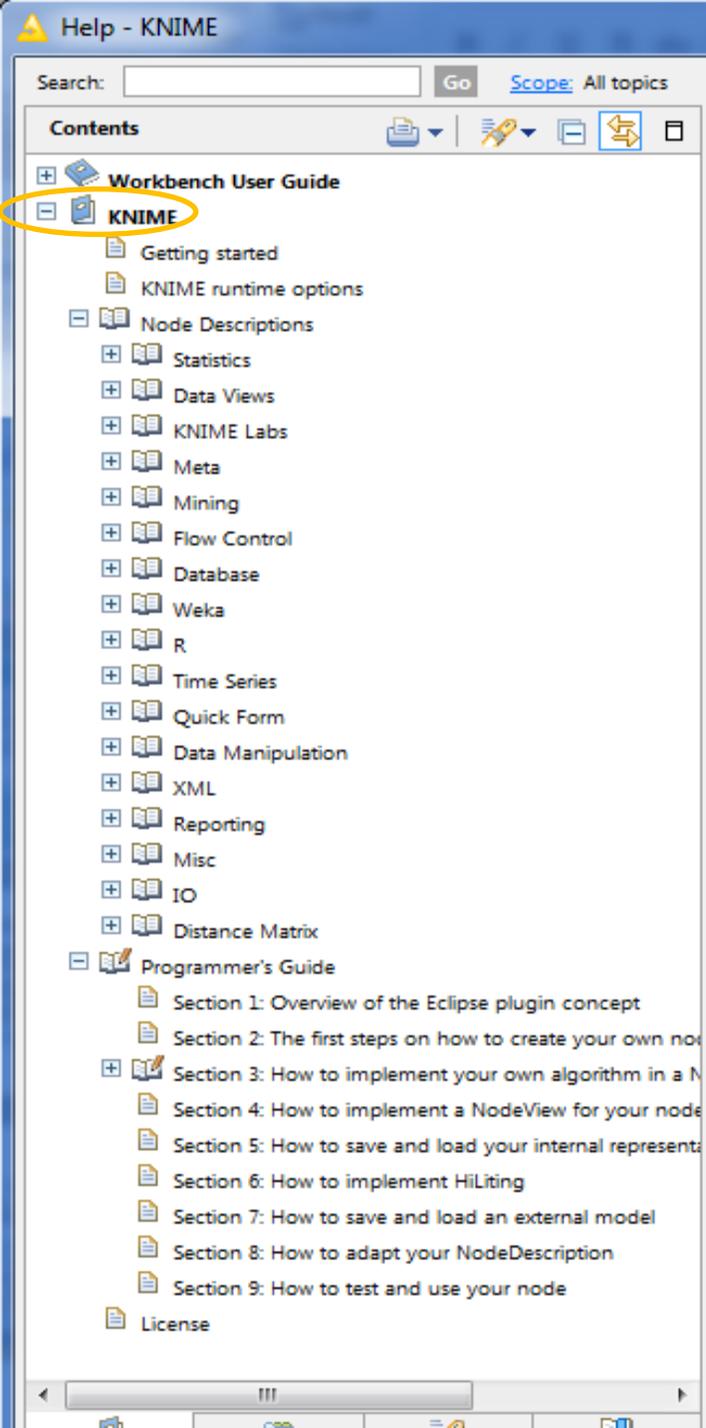
The KNIME Public Server

On the KNIME Public Server you can find a number of useful examples on how to implement specific tasks with KNIME.

In the KNIME Explorer panel:

- right click the public server
- select “Login”
- No login credentials required





Help

Go to the Help menu, select Help Contents.

The KNIME section contains:

- A quick guide on getting started;
- The Node Descriptions chapter. This lists all the node descriptions that appear in KNIME when you select a node or category. You can use the search bar for a specific node description.
- The Programmer's Guide chapter gives details on how to create a new node for KNIME.

What is a Node?

A node is the **single processing unit** of a workflow.

A node takes a data set as input, processes it, and makes it available on its output port.

The “processing” action of a node can be:

- Reading a file
- Data manipulation
- Generating a plot
- Calculating properties or statistics
- .. and more

Nodes containing other nodes are called **meta nodes**.

More on Nodes

A node has a one of three states:

File Reader



Idle:

The node is not yet configured and can not be executed with it's current settings.

File Reader



Configured:

The node has been set up correctly, and may be executed at any time

File Reader



Executed:

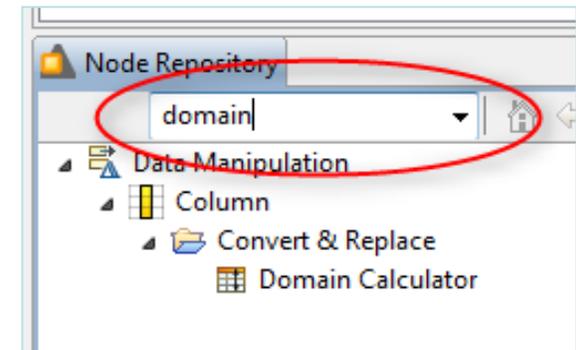
The node has been successfully executed. Results may be viewed and used in downstream nodes.

Node Repository Search Box

The Node Repository contains a **search box**.

If you type a keyword in the search box and press Enter, you obtain the list of nodes with that keyword in the name.

Press the Esc key to see all nodes again.

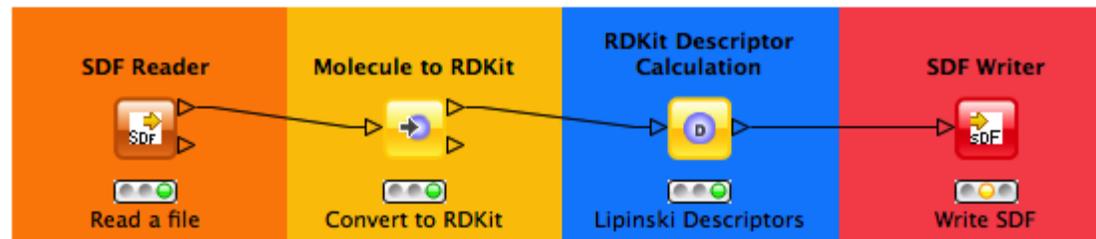


What is a Workflow?

A workflow is an **analysis flow**,
i.e. a sequence of analysis steps
necessary to reach a given
result.

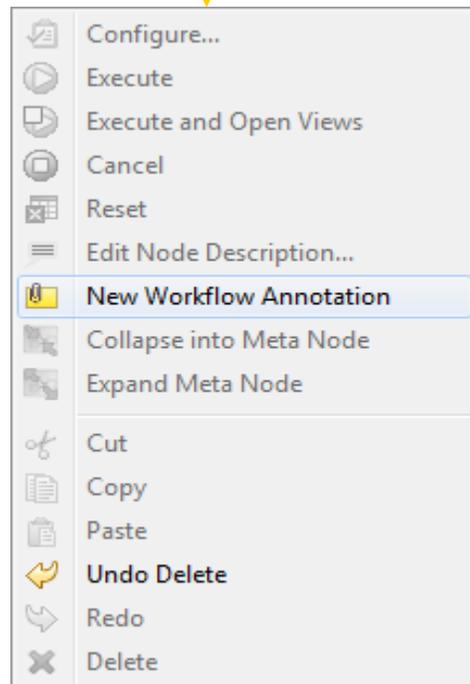
Example:

- Step 1: Read data file
- Step 2: Manipulate types
- Step 3: Analyze
- Step 4: Export Results

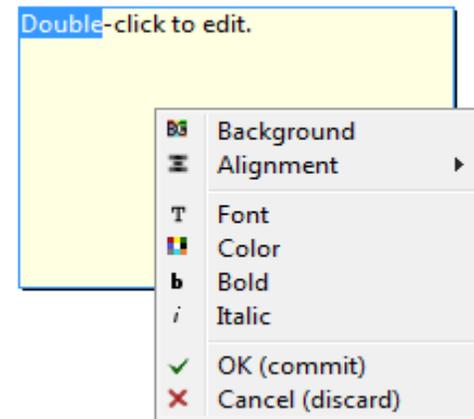


Annotations

Right-click anywhere in the Workflow Editor panel.



Double-click the annotation to write text in it.



Context-menu of the annotation for the annotation editor

Demo

- Survey KNIME Workbench
- Import the demo workflows

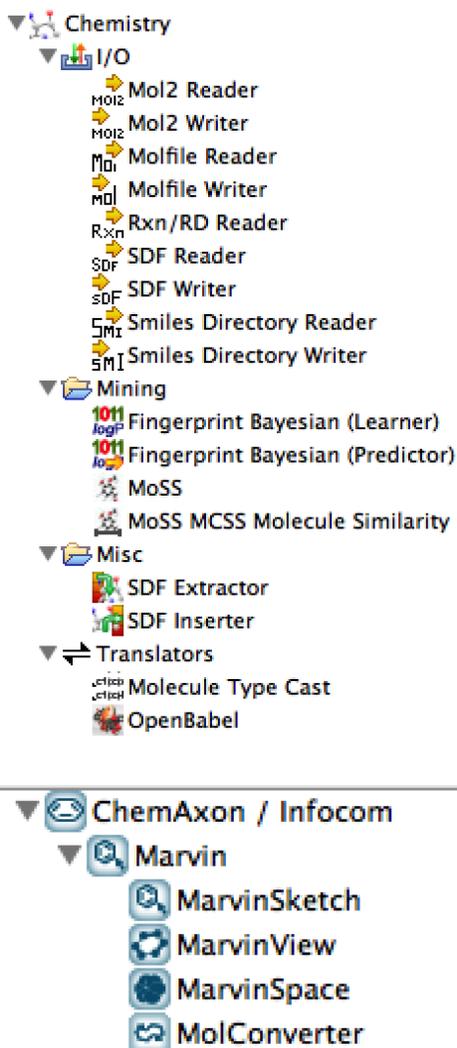


The KNIME Chemistry Extensions

Overview

- KNIME Core chemistry
 - Types, IO, translation, similarity, modeling
- Community Contributions
 - RDKit, CDK, Indigo, Erl Wood etc.
- Commercial Extensions
 - CCG MOE, Chemaxon, Schrodinger etc.

KNIME Core Chemistry



- IO
 - Readers and writers for Mol2, Mol, Rxn, SDF and Smiles.
- Mining
 - Fingerprint learner/predictor
 - MoSS & Moss MCSS similarity
- Misc
 - SDF Property extraction and Insertion
- Translators
 - Molecule Type Cast converts strings to native KNIME chemistry types
 - OpenBabel translates between chemistry types
- Marvin
 - Sketch, Convert and View (2d and 3d)

Chemistry Extensions from the Community

(a few tools...)

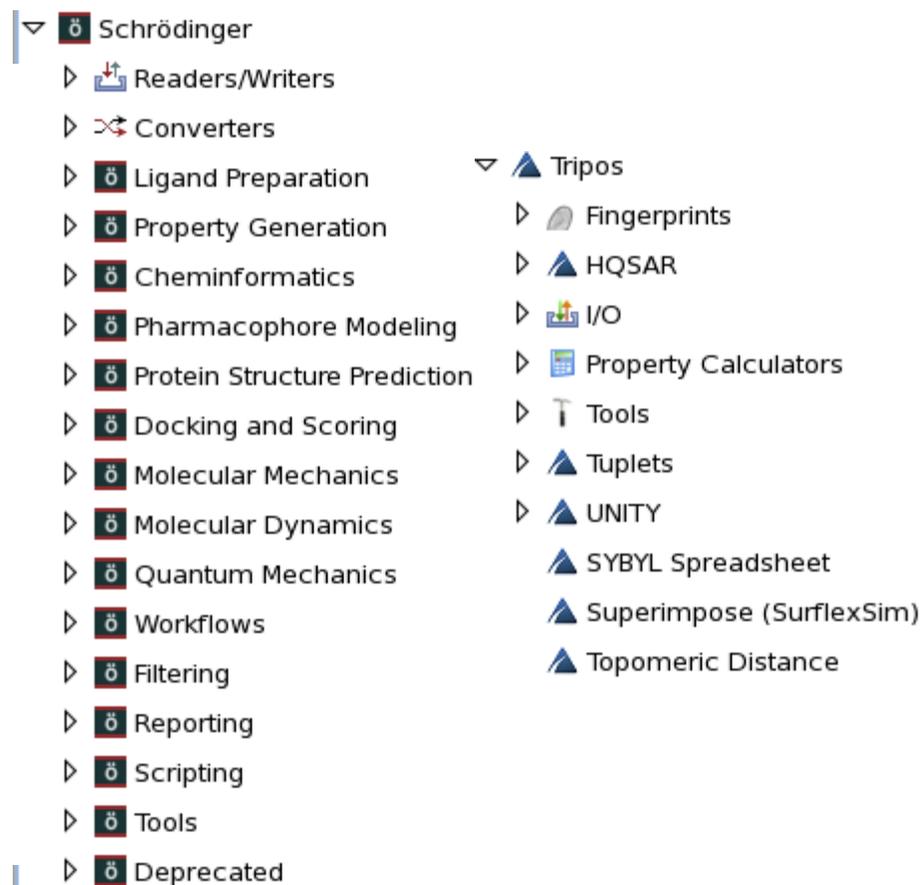
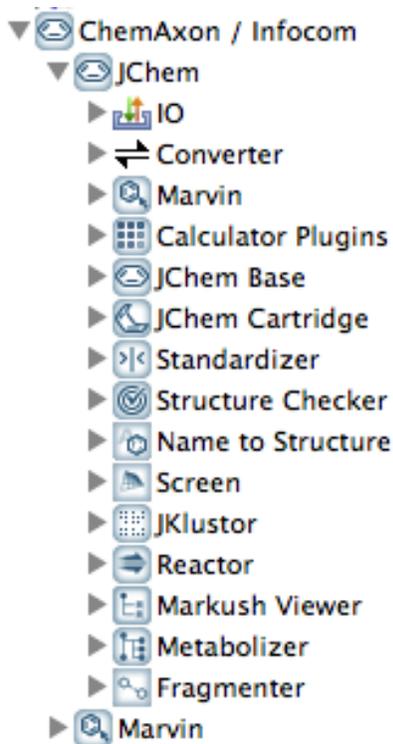
- RDKit
 - Experimental
 - RDKit Diversity Picker
 - RDKit Find Murcko Scaffolds
 - RDKit Molecule Fragmenter
 - RDKit R Group Decomposition
 - Molecule to RDKit
 - RDKit To Molecule
 - InChI to RDKit
 - RDKit To InChI
 - IUPAC to RDKit
 - RDKit Canon SMILES
 - RDKit Fingerprint
 - RDKit Substructure Filter
 - RDKit Dictionary Substructure Filter
 - RDKit One Component Reaction
 - RDKit Highlighting Atoms
 - RDKit Interactive Table
 - RDKit SMILES Headers
 - RDKit Descriptor Calculation
 - RDKit Fingerprint Reader
 - RDKit Fingerprint Writer
 - RDKit Functional Group Filter
 - RDKit Generate Coords
 - RDKit Molecule Substructure Filter
 - RDKit Salt Stripper
 - RDKit Substructure Counter
 - RDKit Two Component Reaction

- CDK
 - 3D
 - 3D Coordinates
 - 3D D-Moments
 - 3D D-Similarity
 - 3D RMSD
 - 3D Viewer
 - S_{i,k} 3D WHIM
 - I/O
 - CDK Mol CDK to Molecule
 - Mol CDK Molecule to CDK
 - 2D Coordinates
 - SgnAtom Signatures
 - ChemSpider
 - Connectivity
 - Element Filter
 - Fingerprint Similarity
 - Fingerprints
 - Hydrogen Manipulator
 - Lipinski's Rule-of-Five
 - Mass Calculator
 - Molecular Properties
 - OPSPIN
 - Structure Sketcher
 - Substructure Search
 - Sugar Remover
 - CH₂ Sum Formula
 - Symmetry
 - XLogP

- Erl Wood Cheminformatics
 - Activity Cliffs
 - Activity Cliffs Viewer
 - Similarity network viewer
 - Calculators
 - Column Merger
 - Fingerprint Similarity
 - Virtual Screening Metrics
 - Convertors
 - Fingerprints Expander
 - Old Bit Vector To New Bit Vector
 - Docking
 - Docking Job Lister
 - Docking Job Retriever
 - Docking Job Submitter
 - IO
 - Chemical Reactions File Reader
 - Text Input
 - Multi-objective
 - Desirability
 - Multi-Objective Loop End
 - Multi-Objective Loop Start
 - Pareto Ranking
 - RGroup Analysis
 - MCS Distance
 - MCS Matrix
 - Matched Pairs Detector
 - Matched Pairs Finder
 - RGroup Efficiency
 - Reaction Generation
 - Reaction Generator
 - Reaction Vectors Database Reader
 - Reaction Vectors Database Writer
 - Viewers
 - 2D/3D Scatterplot
 - Jmol Docking Pose Viewer
 - Jmol Viewer
 - Similarity Viewer
 - Vida Viewer

- Indigo
 - Molecule Translators
 - Molecule to Indigo
 - Query Molecule to Indigo
 - Indigo to Molecule
 - Indigo to Query Molecule
 - Reaction Translators
 - Reaction to Indigo
 - Query Reaction to Indigo
 - Indigo to Reaction
 - Indigo to Query Reaction
 - Molecule Nodes
 - Component Combiner
 - Component Separator
 - Highlighter
 - Isomer Enumerator
 - MCS Scaffold Finder
 - Molecule Transformation (beta)
 - Murcko Scaffold
 - R-Group Decomposer
 - Substructure Match Counter
 - Substructure Matcher
 - Reaction Nodes
 - Reaction Automapper
 - Reaction Builder
 - Reaction Splitter
 - Substructure Matcher
 - Combinatorial Chemistry
 - Combinatorial Reaction Enumeration (beta)
 - Manipulators
 - Aromatizer
 - Atom Replacer
 - Bond Replacer
 - Dearomatizer
 - Feature Remover
 - Generate 2D Coordinates
 - Hydrogen Adder
 - Hydrogen Remover
 - Properties
 - Fingerprint Similarity
 - Indigo Fingerprint
 - Molecule Properties
 - Valence Checker

Commercial Chemistry Extensions (and even more...)



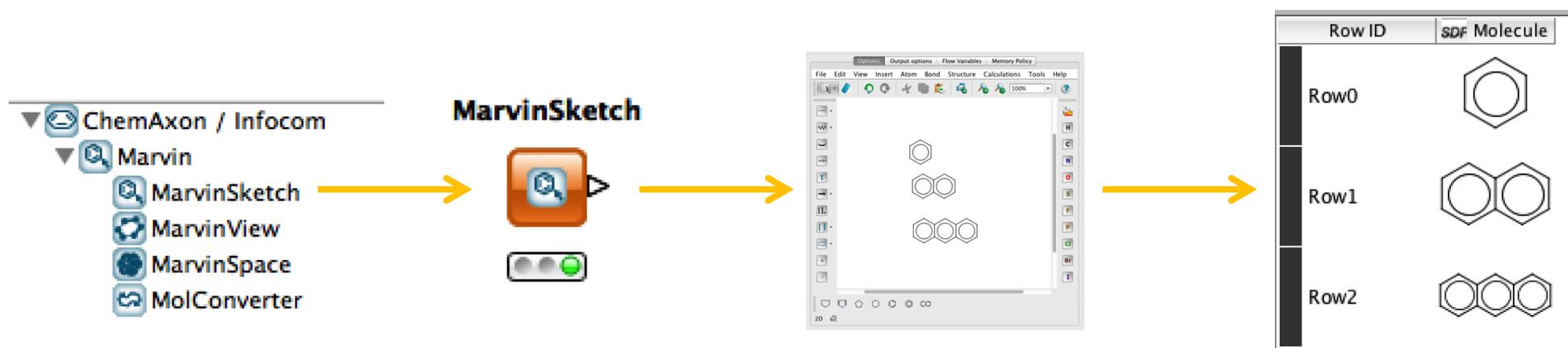


My First Molecule

(Sketching in KNIME)

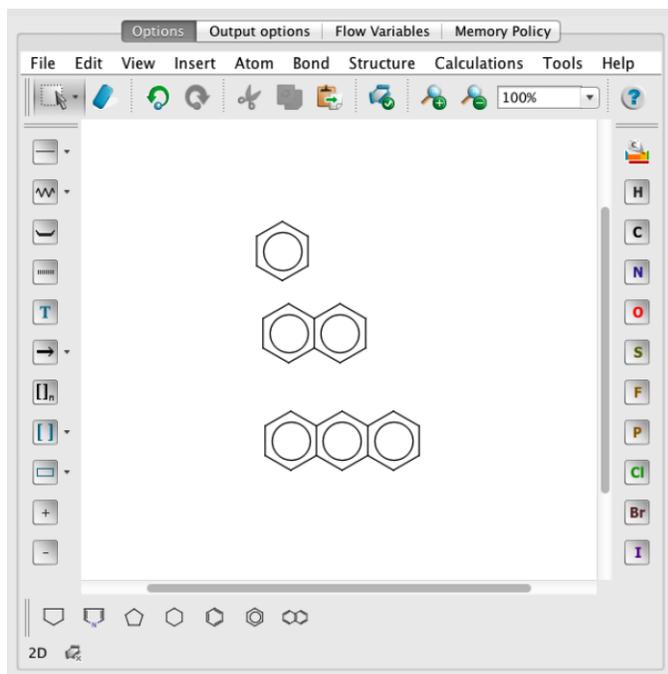
The Marvin Sketcher

- Available from the KNIME Update Site
- Donated by Chemaxon/Infocom
- Integrates with Webportal to sketch in a browser

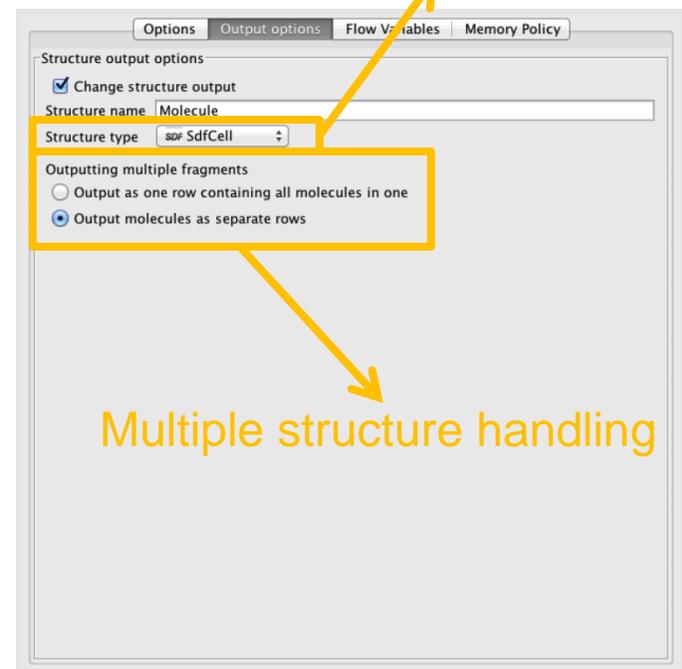


The Marvin Sketcher II

- Full featured Sketcher
- Advanced options shown on in additional tab



Output data type



Structure output options

- Change structure output
- Structure name: Molecule
- Structure type: sdf SdfCell
- Outputting multiple fragments
 - Output as one row containing all molecules in one
 - Output molecules as separate rows

Multiple structure handling

The screenshot shows the 'Structure output options' dialog box. A yellow box highlights the 'Structure type' dropdown menu and the 'Outputting multiple fragments' section. A yellow arrow points from the text 'Output data type' to the 'Structure type' dropdown, and another yellow arrow points from the text 'Multiple structure handling' to the 'Output molecules as separate rows' radio button.

DEMO

- Sketch some molecules, export as an SDF cell, generate a new data row for each molecule



Chemistry Data Basics

Types and IO

Chemistry Types Overview

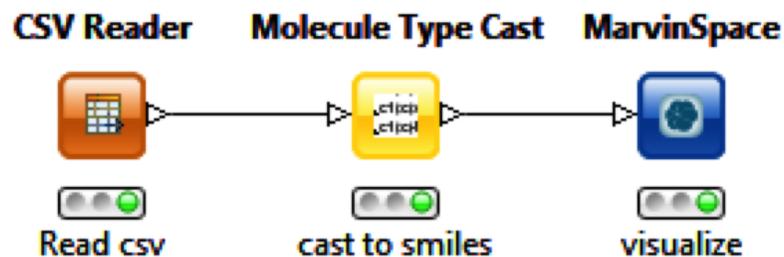
- KNIME Basic types
 - String, Int, Double

- KNIME Core Chemistry Types
 - SDF, MOL, MOL2, Smiles
 - These types can be rendered in KNIME Tables

- Special nodes convert between types.

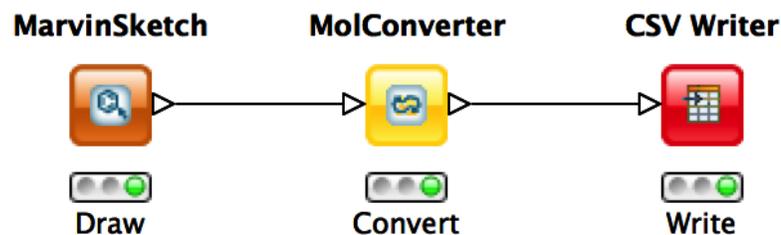
Node: Molecule Type Cast

- KNIME Core Chemistry
- Convert string cells to a renderable smiles type
- Use to convert structures from string text files or databases to smiles cells.



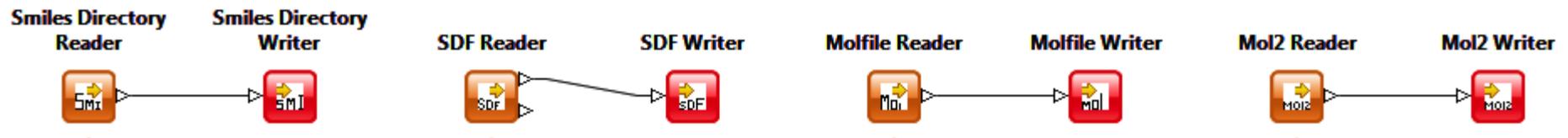
Nodes: MolConverter & OpenBabel

- KNIME Core Chemistry
- Convert to/from virtually any format
- MolConverter provided by Chemaxon/Infocom
- www.openbabel.org



Chemistry IO

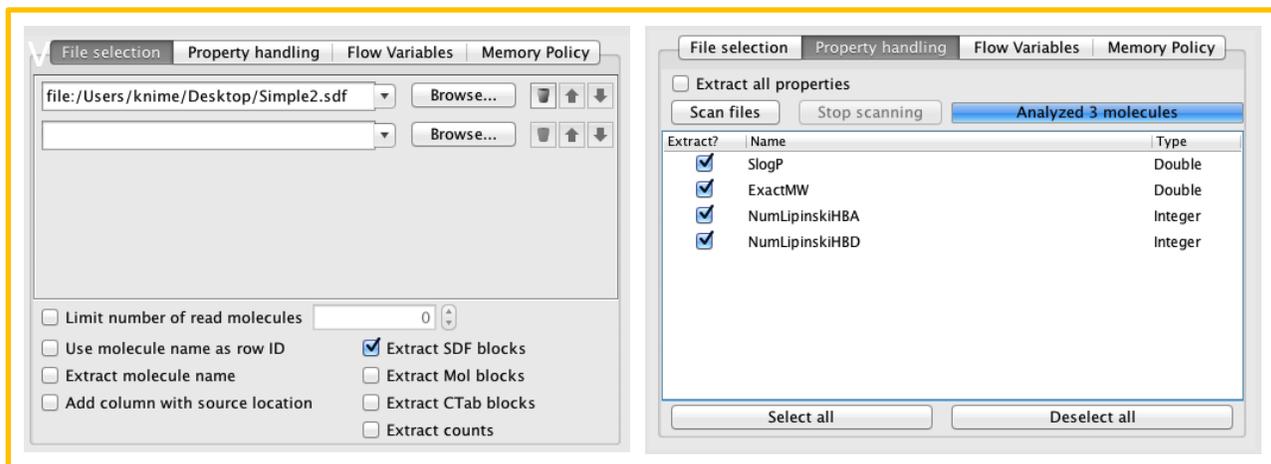
- File Readers and writers available for:
 - SDF, MOL, MOL2, Smiles



Chemistry IO II: SDF

- SDF Reader can import properties
- SDF Writer can write columns as properties
- Dedicated nodes also exist for this (SDF Inserter and SDF Extractor)

SDF Reader

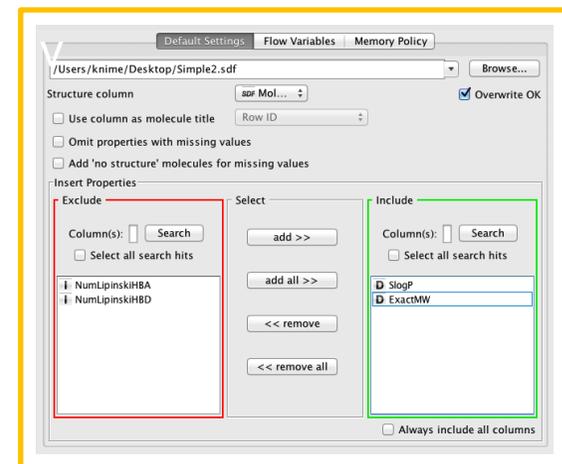


The SDF Reader node configuration window is shown with the following settings:

- File selection:** file:/Users/knime/Desktop/Simple2.sdf
- Property handling:**
 - Extract all properties
 - Buttons: Scan files, Stop scanning, Analyzed 3 molecules
 - Table of properties to extract:

Extract?	Name	Type
<input checked="" type="checkbox"/>	SlogP	Double
<input checked="" type="checkbox"/>	ExactMW	Double
<input checked="" type="checkbox"/>	NumLipinskiHBA	Integer
<input checked="" type="checkbox"/>	NumLipinskiHBD	Integer
- Flow Variables:**
 - Limit number of read molecules: 0
 - Use molecule name as row ID
 - Extract molecule name
 - Add column with source location
 - Extract SDF blocks
 - Extract Mol blocks
 - Extract CTab blocks
 - Extract counts

SDF Writer



The SDF Writer node configuration window is shown with the following settings:

- File selection:** /Users/knime/Desktop/Simple2.sdf
- Structure column:** sor Mol...
- Property handling:**
 - Use column as molecule title
 - Omit properties with missing values
 - Add 'no structure' molecules for missing values
- Insert Properties:**
 - Exclude:** NumLipinskiHBA, NumLipinskiHBD
 - Select:** (empty)
 - Include:** SlogP, ExactMW
- Always include all columns

Demo

- Create new workflow named Exercise 3
- Read structures from 124mols.sdf
- Write smiles to csv
- Read resulting csv
- Cast molecule strings to smiles and translate to sdf
- Visualize in report with BIRT



Molecular Descriptors

An RDKit example

Overview

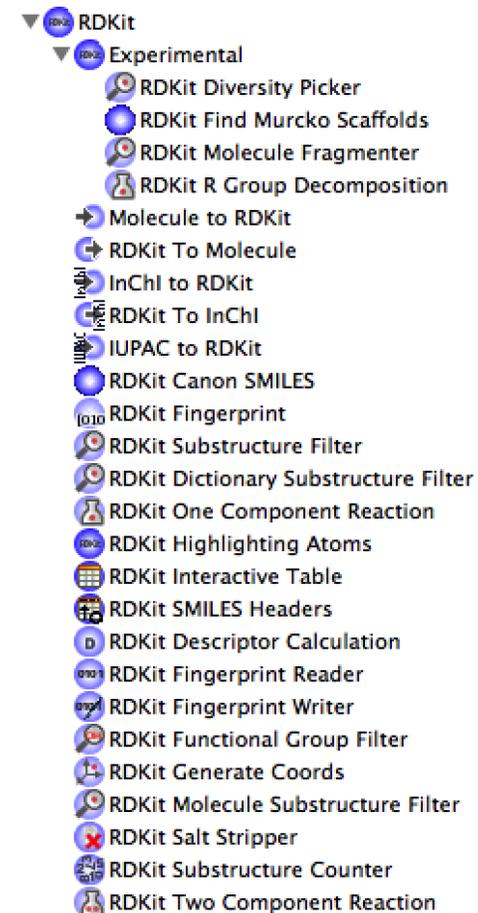
- Intro to RDKit
- Special Type Conversion
- Molecular properties
- InChI keys
- Smiles



Open-Source Cheminformatics
and Machine Learning

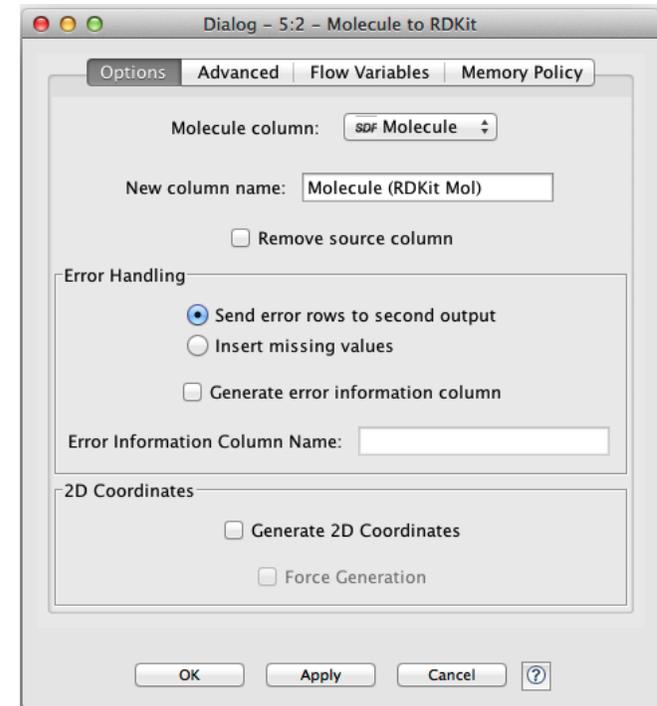
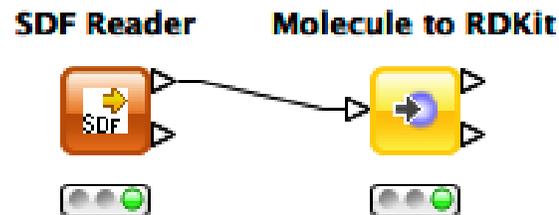
RDKit in KNIME

- Cheminformatics and Machine Learning
- Written in C++ and python.
- www.rdkit.org for more information



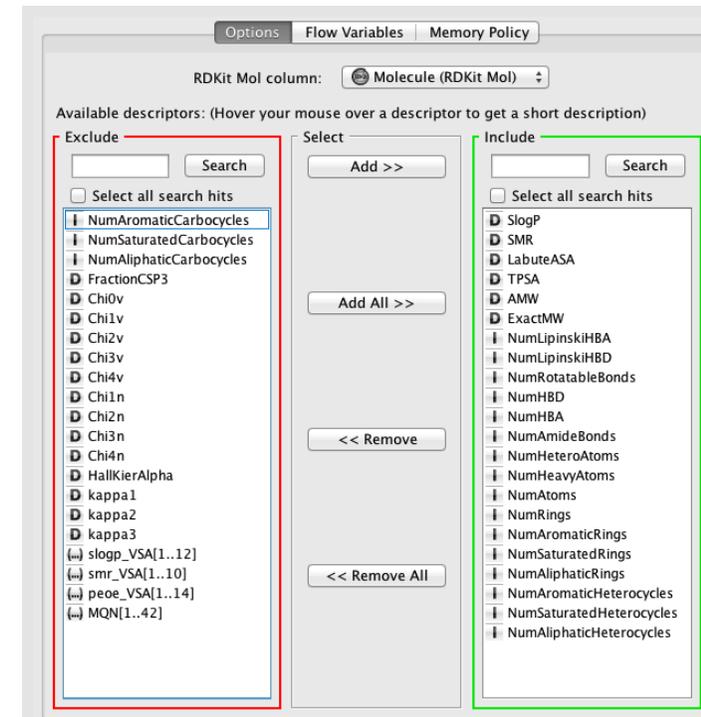
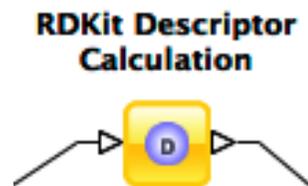
Node: Molecule to RDKit

- Input smiles or SDF
- Creates RDKit data column
- Required before most RDKit nodes



Node: RDKit Descriptor Calculator

- Input RDKit Molecules
- Calculates many descriptors
 - MW, logp, h-bond donors etc.



Node: InChI Keys

- Input RDKit Molecules
- Generates InChI Keys



Options | Advanced | Flow Variables | Memory Policy

RDKit Mol column:

Remove source column

InChI Code Generation

New column name for InChI codes:

InChI Key Generation

Generate also InChI keys

New column name for InChI keys:

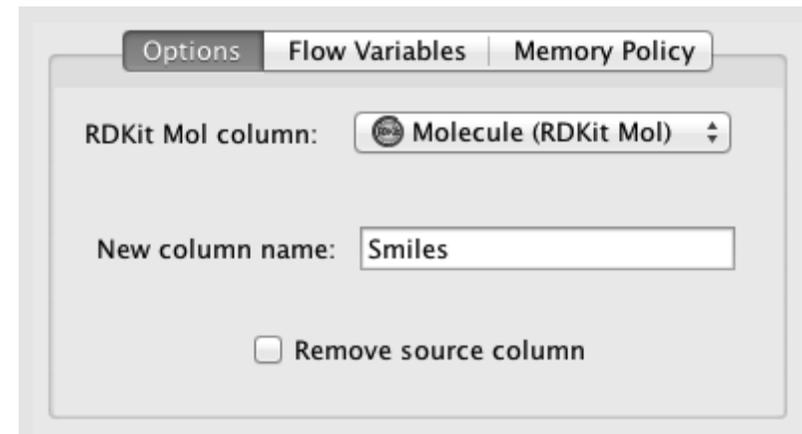
Extra InChI Generation Information

New column name prefix for extra information:

Return Code Column Aux Info Column Message Column Log Column

NodeRDKit Cannon Smiles

- Input RDKit Molecule
- Generates canonical smiles



Demo

- Read structures from 124mols.sdf
- Convert to RDKit
- Calculate Properties
- Generate InChI keys
- Generate smiles
- Write to SDF



KNIME Chemistry

Similarity and Bioactivity from ChEMBL

Workflow Background

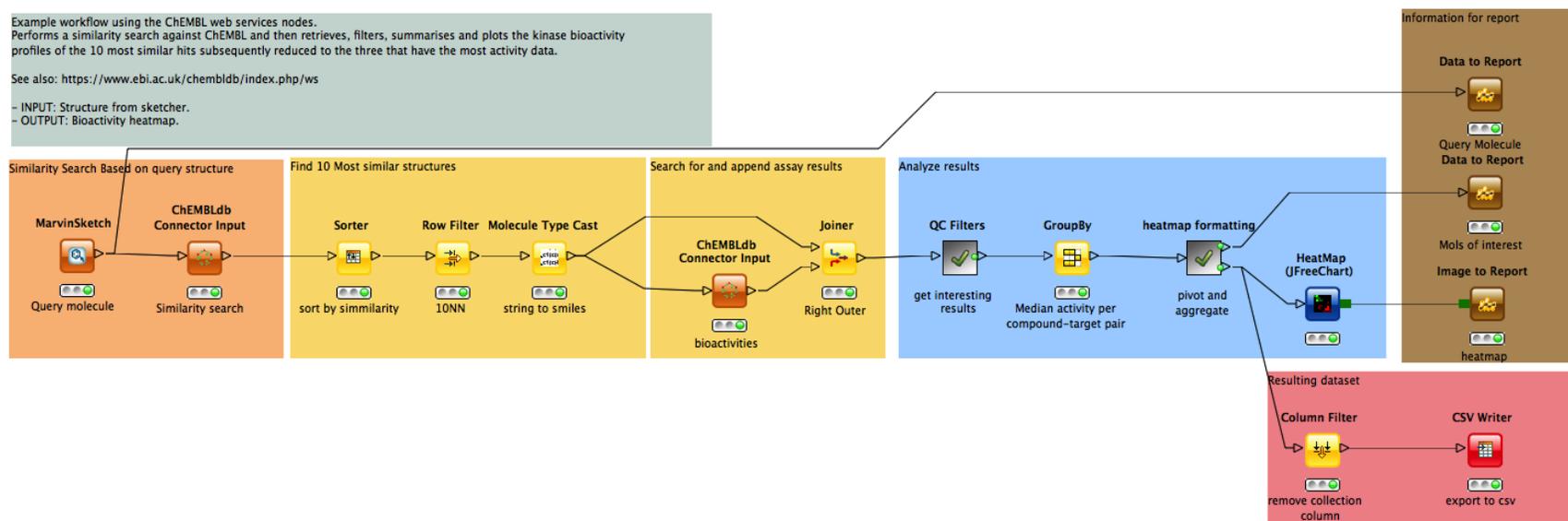
- Donated by EBI (Along with the nodes!)
- Similarity search and bioactivity query with ChEMBL Webservices nodes
- Available on the public examples server.

Example workflow using the ChEMBL web services nodes.
Performs a similarity search against ChEMBL and then retrieves, filters, summarises and plots the kinase bioactivity profiles of the 10 most similar hits subsequently reduced to the three that have the most activity data.

See also: <https://www.ebi.ac.uk/chembl/index.php/ws>

- INPUT: Structure from sketcher.

- OUTPUT: Bioactivity heatmap.



ChEMBL Webservices Demo

- Modify the workflow to include assay data from only the 3 molecules with the most activity data.

Thank you

education@knime.com