



Open-Source Cheminformatics



Open source cheminformatics in KNIME with the RDKit: update

Gregory Landrum

NIBR IT

Novartis Institutes for BioMedical Research, Basel

5th KNIME Users Group Meeting

Zurich, 2 February 2012

Acknowledgements

■ Novartis:

- Tom Digby (Legal)
- John Davies (CPC)
- Steve Litster (NIBR IT)
- Richard Lewis (GDC)
- Patrick Warren (NIBR IT)
- Andy Palmer (NIBR IT)
- Manuel Schwarze (NIBR IT)
- Dillip Kumar Mohanty (NIBR IT)
- Sayantan Sengupta (NIBR IT)

■ Rational Discovery:

- Santosh Putta (currently at Nodality)
- Julie Penzotti

- RDKit open-source community

- KNIME forum members

- knime.com

- Michael Berthold
- Thorsten Meinl
- Bernd Wiswedel



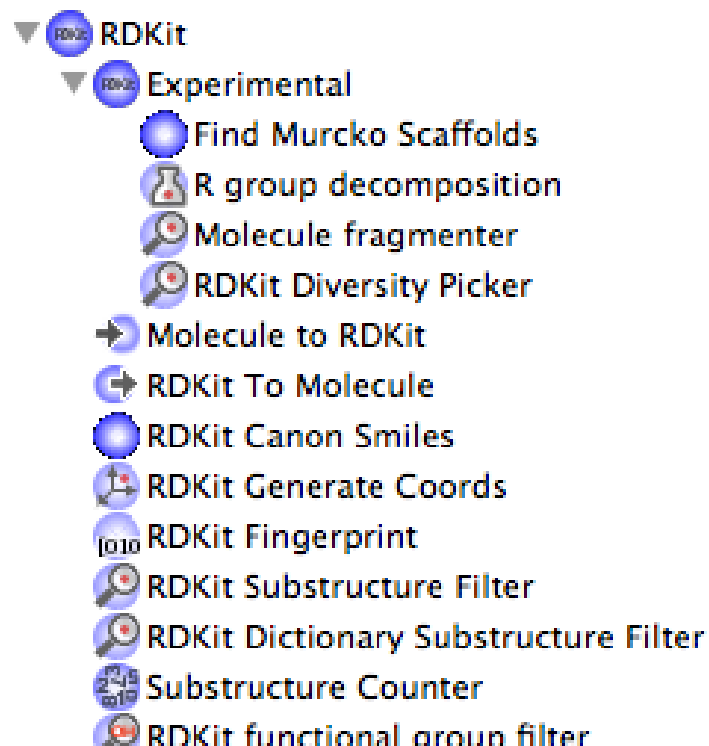
RDKit: What is it?

- Python (2.x), Java, C++ toolkit for cheminformatics
 - Core data structures and algorithms in C++
 - Heavy use of Boost libraries
 - Python wrapper generated using Boost.Python
- Functionality:
 - 2D and 3D molecular operations
 - Descriptor generation for machine learning
 - Database cartridge for substructure and similarity searching
 - Supports Mac/Windows/Linux
- History:
 - 2000-2006: Developed and used at Rational Discovery for building predictive models for ADME, Tox, biological activity
 - June 2006: Open-source (BSD license) release of software, Rational Discovery shuts down
 - to present: Open-source development continues, use within Novartis, contributions from Novartis back to open-source version

Knime integration¹

- Out of the box Knime is strong on data processing and mining, weak on chemistry.
- Goal: develop a set of *open-source* RDKit-based nodes for Knime that provide basic cheminformatics functionality

- Distributed from knime community site
- Binaries available as an update site (no RDKit build/installation required)
- Work in progress: more nodes being added (new wizard makes it easy)



¹ Work done together with knime.com

What's there?

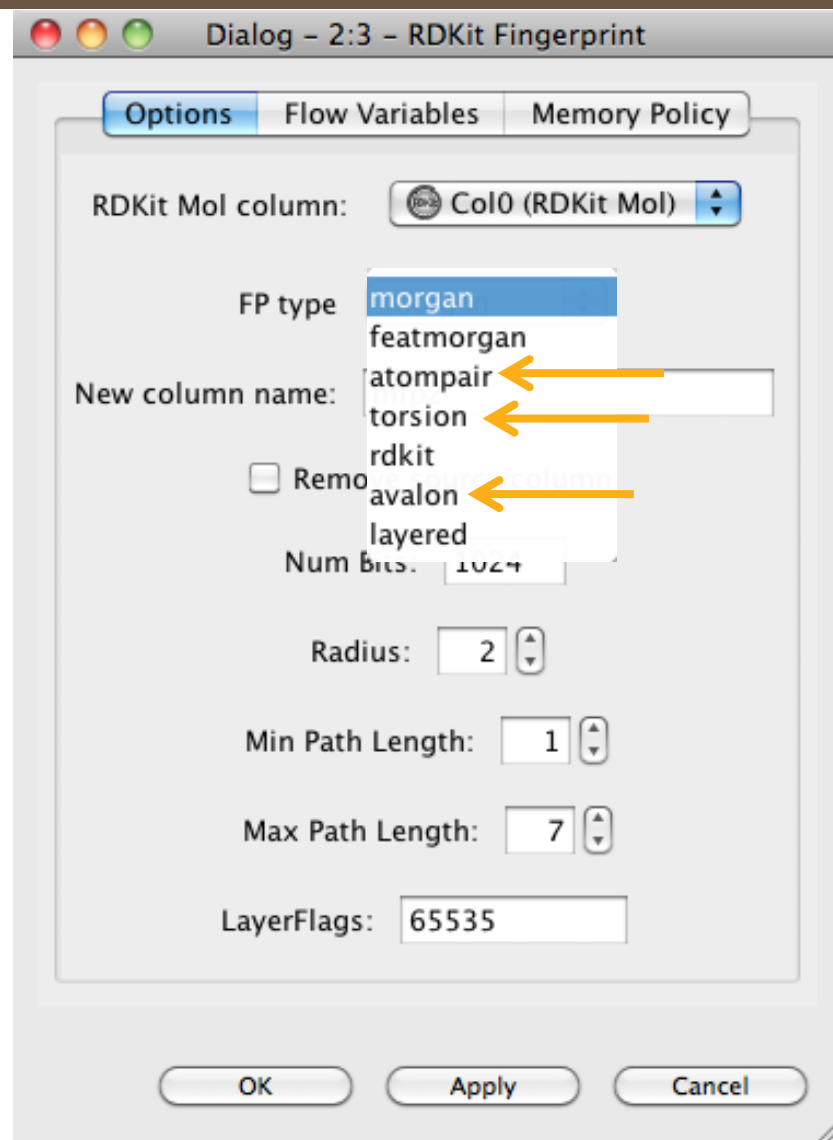
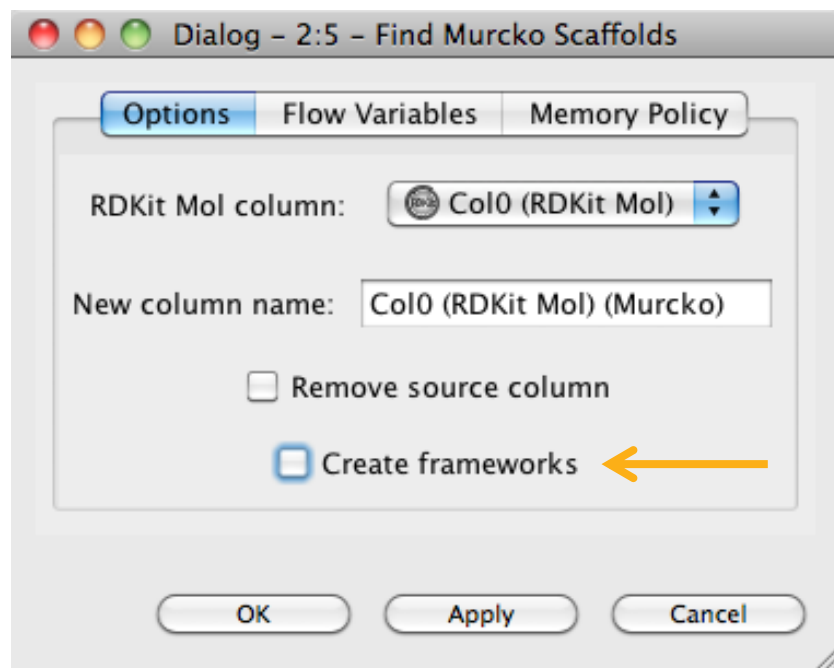
- RDKit
 - Experimental
 - Find Murcko Scaffolds ←
 - R group decomposition
 - Molecule fragmenter
 - RDKit Diversity Picker ←
 - Molecule to RDKit
 - RDKit To Molecule ←
 - RDKit Canon Smiles
 - RDKit Generate Coords
 - RDKit Fingerprint ←
 - RDKit Substructure Filter ←
 - RDKit Dictionary Substructure Filter
 - Substructure Counter ←
 - RDKit functional group filter ←
 - RDKit One Component Reaction
 - RDKit Two Component Reaction
 - RDKit Salt Stripper ←
 - Descriptor Calculation ←
 - Fingerprint Writer ←
 - Fingerprint Reader ←

Another new piece: 64bit Windows support

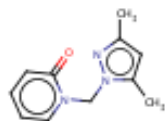
←
New

←
Updated

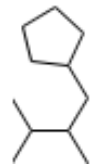
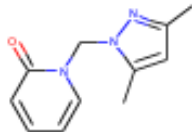
Modified nodes



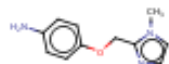
Row3693



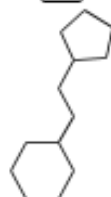
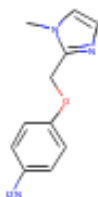
ZINC053793...



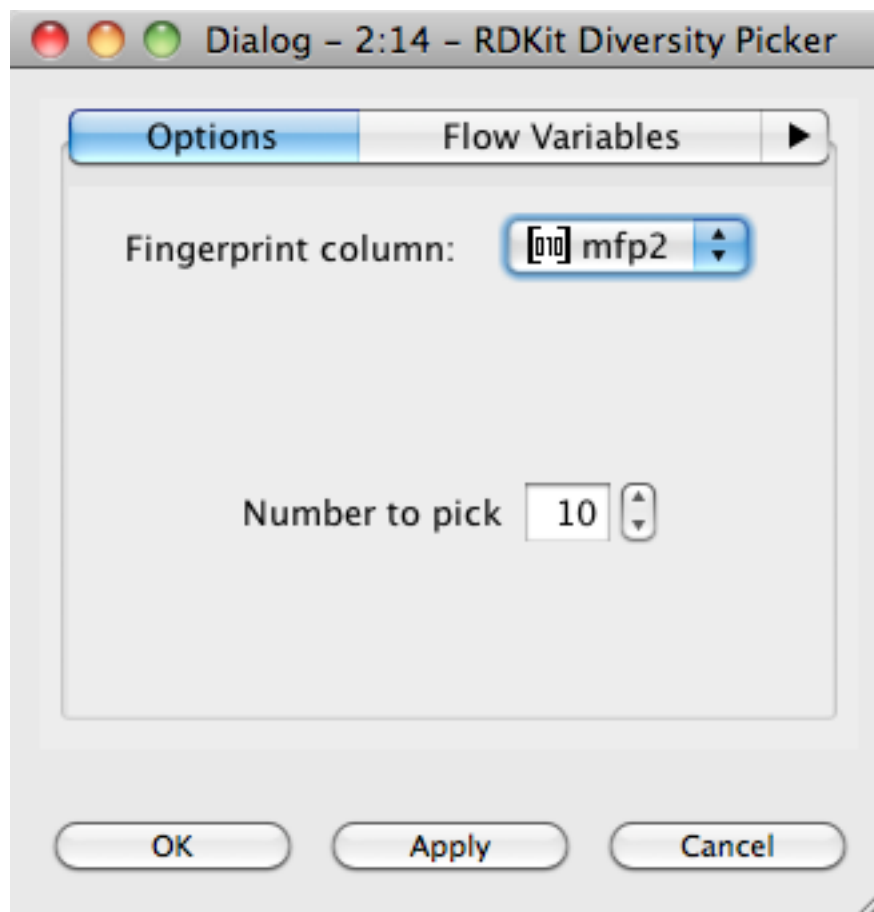
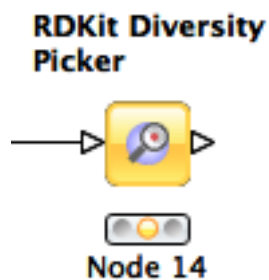
Row3694



ZINC319775...

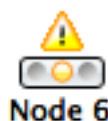


Diversity Picker



Functional group filter

RDKit functional group filter



Dialog - 2:6 - RDKit functional group filter

Settings | Flow Variables | Memory Policy

Select molecule column

Column name : Col0 (RDKit Mol)

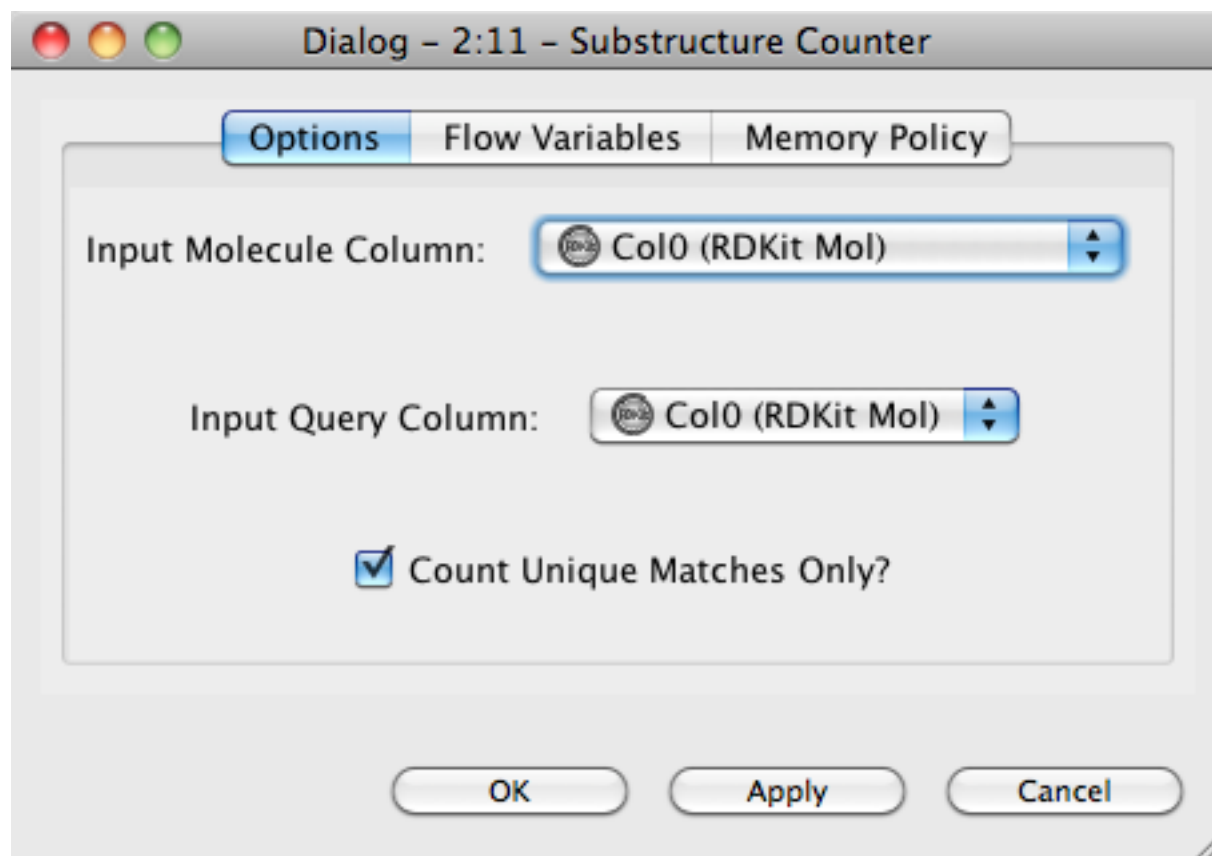
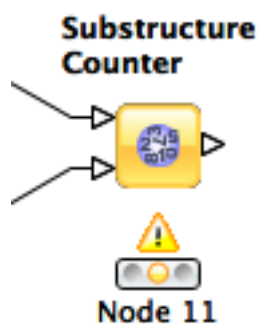
Select functional group definition file (Optional)

File path : Browse... Load default

List of available functional group filters

Select	Functional Group Name	Qualifier	Count
<input type="checkbox"/>	AcidChloride	Exactly(=)	0
<input type="checkbox"/>	Aromatic AcidChloride	Exactly(=)	0
<input type="checkbox"/>	Aliphatic AcidChloride	Exactly(=)	0
<input type="checkbox"/>	CarboxylicAcid	Exactly(=)	0
<input type="checkbox"/>	Aromatic CarboxylicAcid	Exactly(=)	0
<input type="checkbox"/>	Aliphatic CarboxylicAcid	Exactly(=)	0
<input type="checkbox"/>	AlphaAmino CarboxylicAcid	Exactly(=)	0
<input type="checkbox"/>	SulfonylChloride	Exactly(=)	0
<input type="checkbox"/>	Aromatic SulfonylChloride	Exactly(=)	0
<input type="checkbox"/>	Aliphatic SulfonylChloride	Exactly(=)	0
<input checked="" type="checkbox"/>	Amine	Exactly(=)	1
<input type="checkbox"/>	Primary Amine	Exactly(=)	0

Substructure Counter

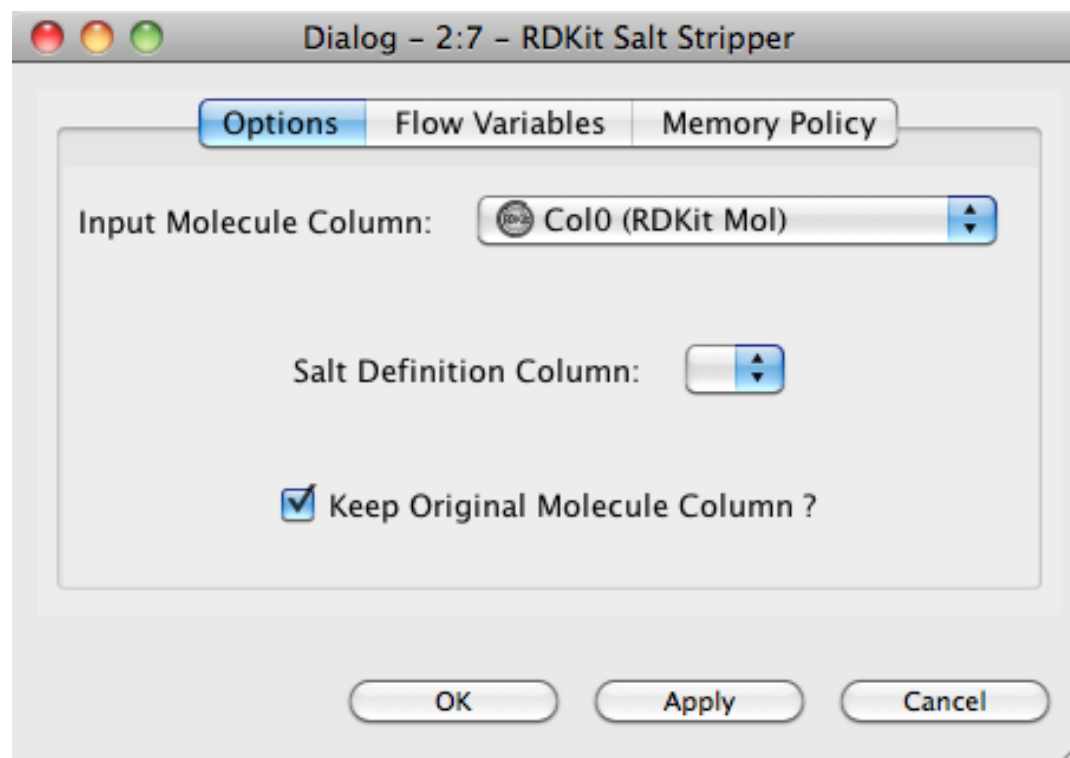


Salt stripper

RDKit Salt Stripper

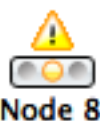


Node 7



Descriptor calculator

Descriptor Calculation



The screenshot shows a software dialog box titled "Dialog - 2:8 - Descriptor Calculation". It has three tabs: "Settings" (selected), "Flow Variables", and "Memory Policy".

At the top, there is a dropdown menu for "Molecule column:" set to "Col0 (RDKit Mol)".

Below this, the "Available descriptors:" section is divided into three main areas:

- Exclude:** A red-bordered box containing a "Descriptor(s):" input field, a "Search" button, and a checkbox labeled "Select all search hits". Below this is an empty list box.
- Select:** A central area with five buttons: "add >>", "add all >>", "<< remove", and "<< remove all".
- Include:** A green-bordered box containing a "Descriptor(s):" input field, a "Search" button, and a checkbox labeled "Select all search hits". Below this is a list box containing the following descriptors: "slogp", "smr", "LabuteASA", "TPSA", "AMW", "ExactMW", "NumLipinskiHBA", "NumLipinskiHBD", and "NumRotatableBond".

At the bottom of the dialog are three buttons: "OK", "Apply", and "Cancel".

Fingerprint reader/writer

Fingerprint Writer

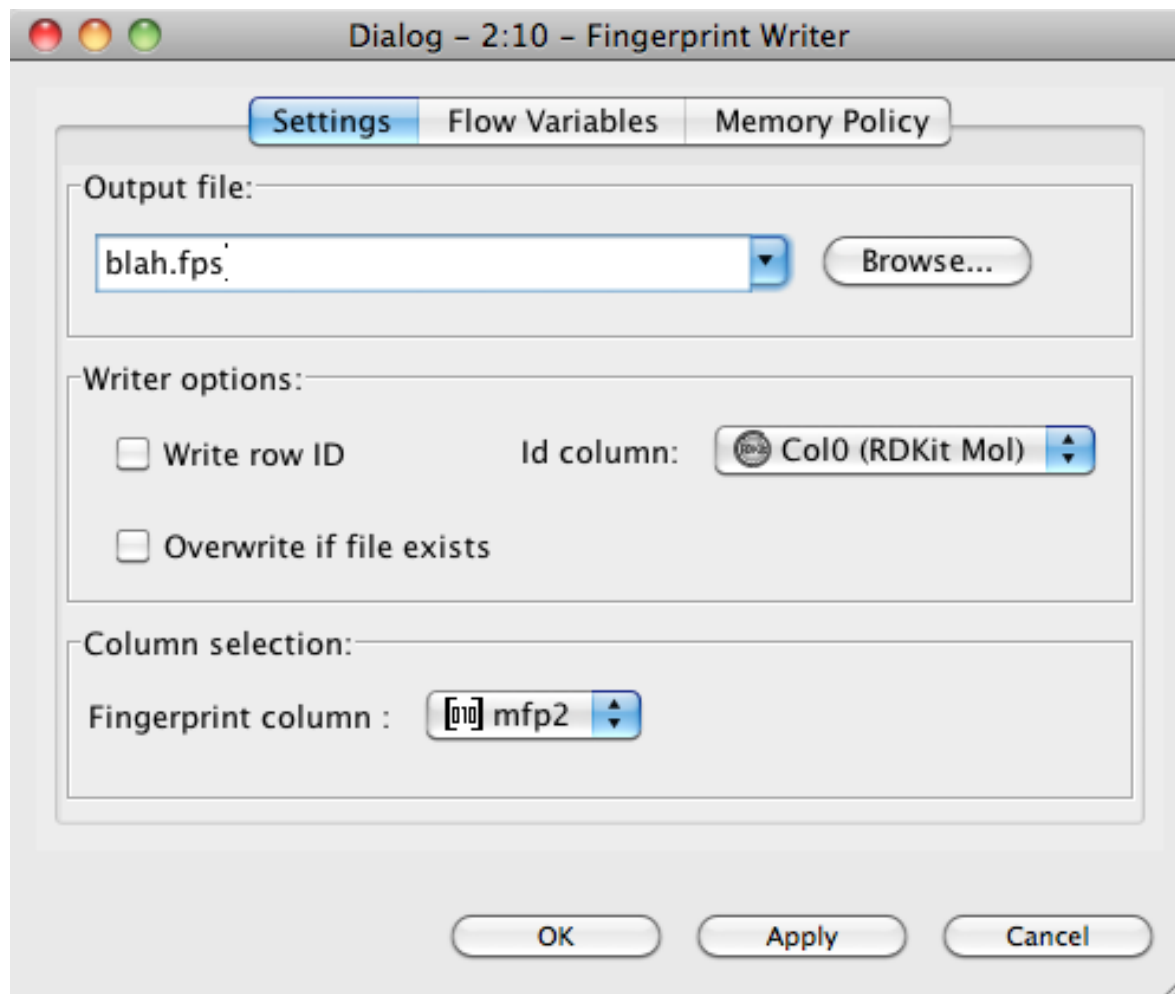


Node 10

Fingerprint Reader



Node 9



<http://code.google.com/p/chem-fingerprints/>