

Analysis of Mass Spectrometry Data with KNIME

Timo Sachsenberg, Julianus Pfeuffer

The Center for Integrative Bioinformatics (CIBI)



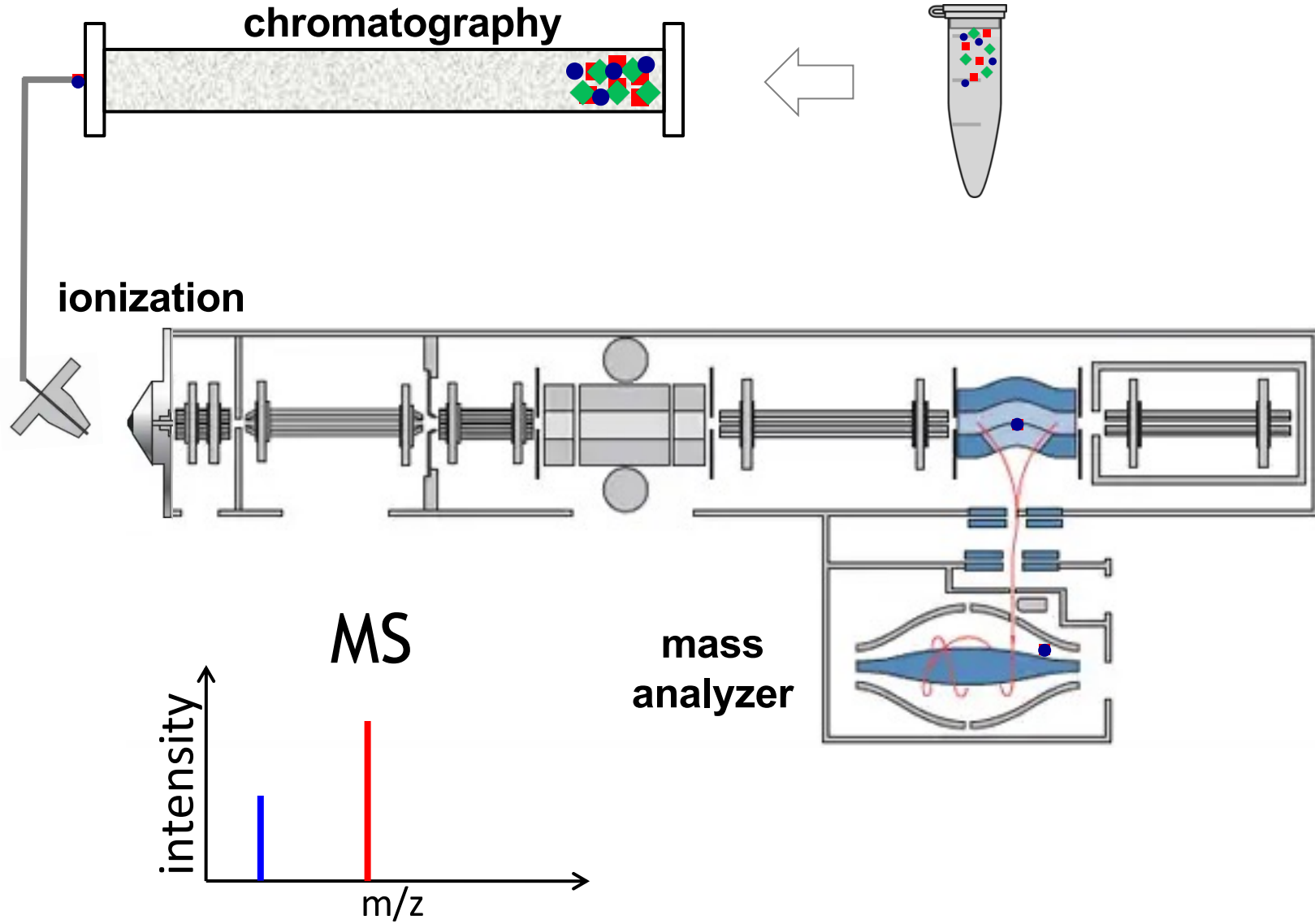
Mass Spectrometry (MS)

- Investigate samples at the molecular level
- Identify and quantify molecules like proteins, metabolites, chemicals (toxins, drugs)



Photo: BusinessWire

MS



Computational MS

- Large amount of data: >50k spectra, data >100gb
- Hundreds of *experimental methods* and protocols

Many *computational methods* for

- Identification
- Quantification
- Statistical analysis

often tailored to experimental method (e.g., **biomarker discovery**)

A single workflow or tool is not enough !

Solution: A set of tools that can be combined to highly flexible workflows



OpenMS

(py)OpenMS: an open-source C++ framework and python bindings for computational MS

Open source:

BSD 3-clause license, available on Windows, OSX, Linux

OpenMS tools - Building blocks:

One application for each analysis step

Integrated in many workflow systems:

KNIME, Galaxy, WS-PGRADE/gUSE, snakemake, nextflow

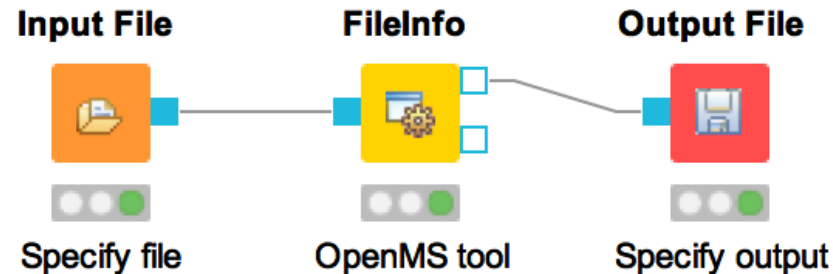


OpenMS in KNIME

Wrapping:

- OpenMS tool write configuration files: *Common Tool Description (CTD)*
- GenericKNIMENodes (GKN) generates Java source code (static) or an XML representation (dynamic) for nodes to show up in KNIME

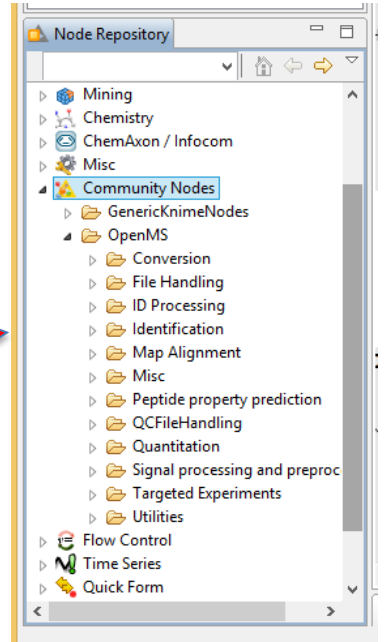
+ File handling nodes



The OpenMS KNIME plugin

- Community-contributions update site (stable & trunk)
- Provides ~180 MS related tools as Community nodes many of them recently added
- **Available soon: the OpenMS 3.0 KNIME plugins**

label-free SWATH
Protein-RNA XL
RNA PTM analysis
SILAC targeted extraction TMT
Protein-Protein XL
Protein-DNA XL
Epifany Protein-SIP
top-down
iTRAQ

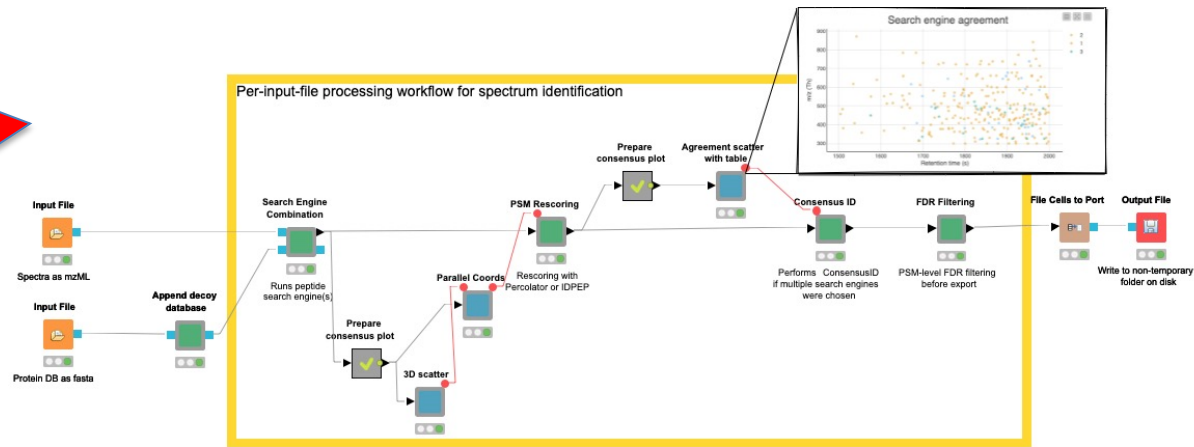


Workflow superpowers: Add pyOpenMS scripts!

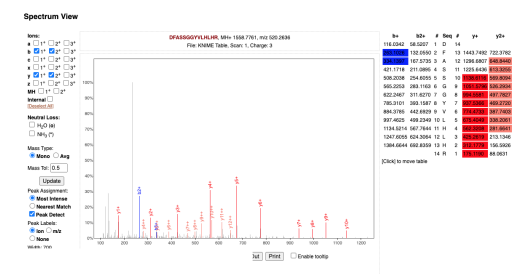
- Documentation: <https://pyopenms.readthedocs.io/en/latest/>



Example: Protein/peptide identification with visualization of spectra



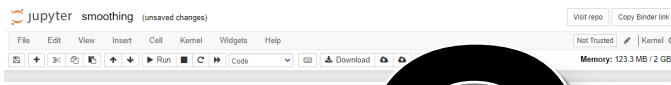
RowID	sequence	PSM_ID	accession	unique	database	database_version	search_engine	set
Row 1_spectrum-0	DFASGGVYLHLR	0	[3612_rev]	true	SimpleSearchEngine_1	null	[...]	16
Row 2_spectrum-1	VALSPRIVEALNDFPFDVWYMK	1	[35A1]	true	SimpleSearchEngine_1	null	[...]	42
Row 3_spectrum-2	RPSADESGHGGFGLAQAQGR	2	[35A1]	true	SimpleSearchEngine_1	null	[...]	34



Workflows on KNIME Hub: <https://hub.knime.com/openms-team/spaces/Blog%20workflows/latest/>

Blog post: <https://www.knime.com/blog/mass-spectrometry-protein-identification>

Develop tools with (py)OpenMS and create the next generation of workflows in KNIME!



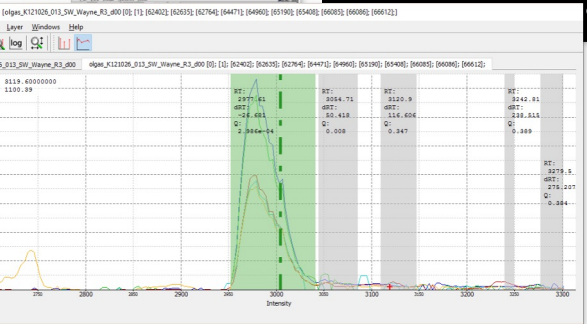
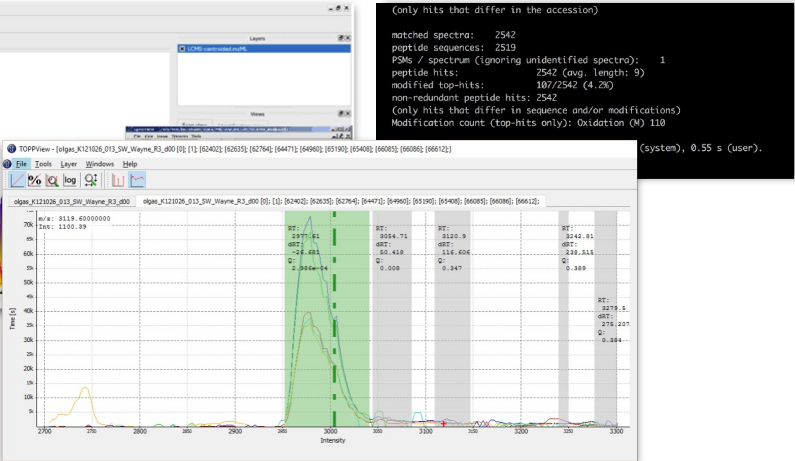
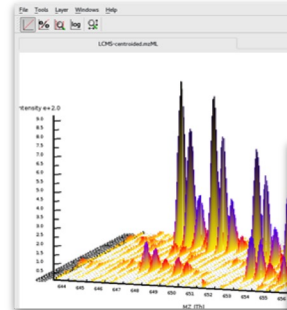
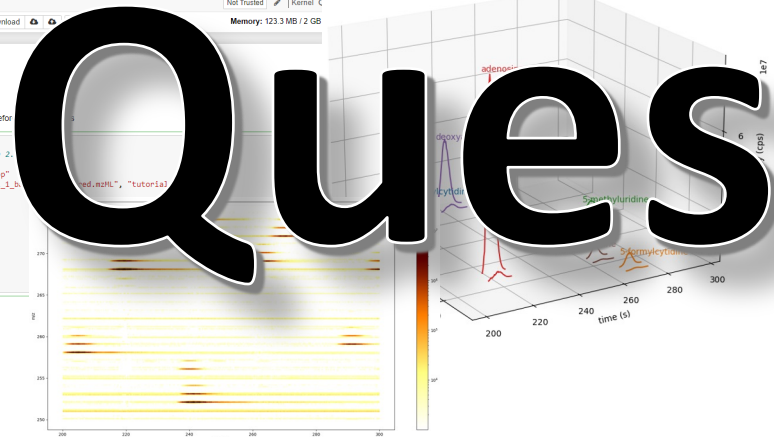
Smoothing

In many applications, mass spectrometry data should be smoothed first before

```
In [ ]: from urllib.request import urlopen
# from urllib import urlopen # use this code for Python 2
from pyopenms import *
gh = "https://raw.githubusercontent.com/OpenMS/OpenMS/develop"
urlopen(gh + "/share/OpenMS/examples/peakpicker_tutorial_1_1_1.mzML", "tutorial_1_1_1.mzML")

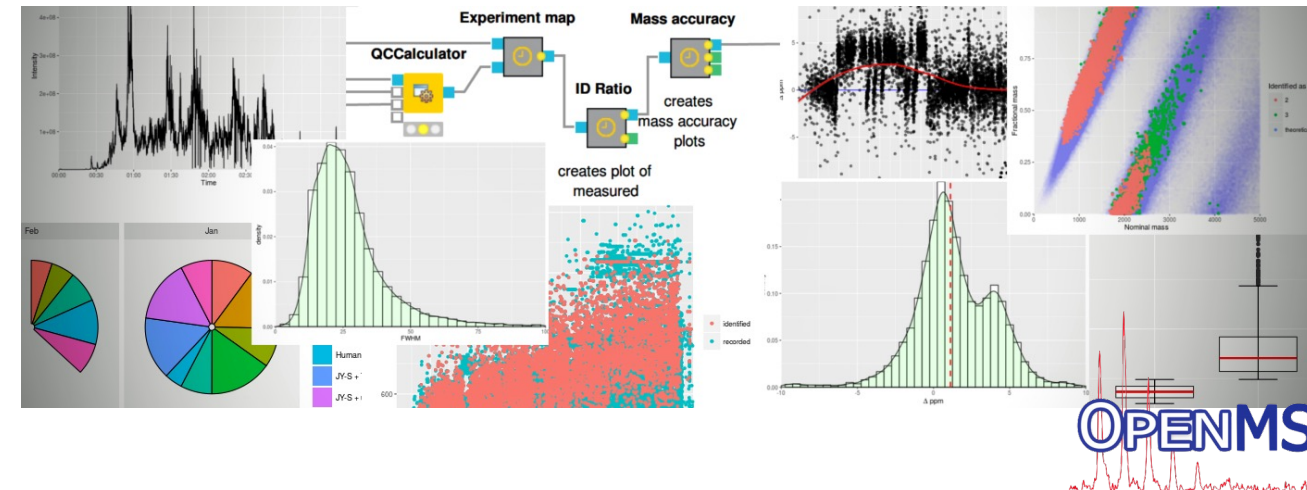
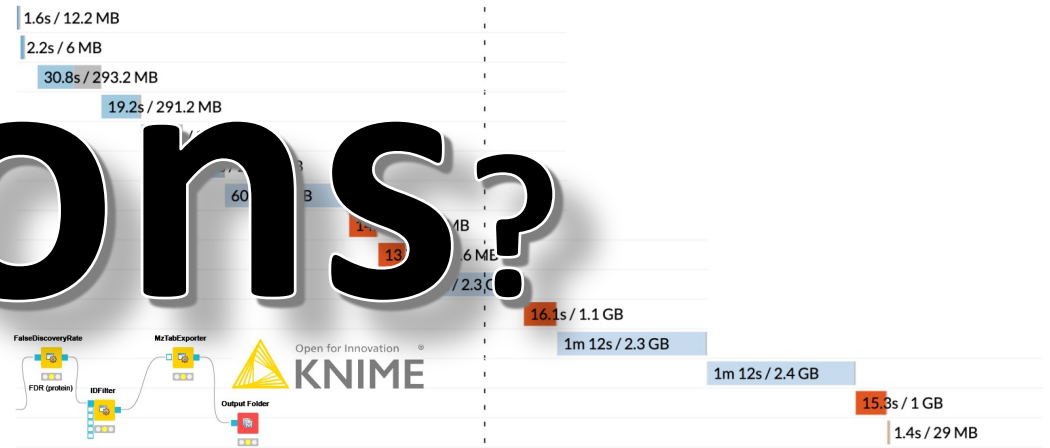
exp = MSExperiment()
gf = GausFilter()
param = gf.getParameters()
param.setValue("gaussian_width", 1.0) # needs wider width
gf.setParameters(param)

MSMFile().load("tutorial.mzML", exp)
gf.filterExperiment(exp)
MSMFile().store("tutorial.smoothed.mzML", exp)
```



Questions?

- sdrf_parsing (1)
- get_software_versions
- raw_file_conversion (1)
- raw_file_conversion (2)
- raw_file_conversion (3)
- raw_file_conversion (4)
- search_engine_msgf (1)
- search_engine_msgf (2)
- search_engine_comet (3)
- search_engine_msgf (3)
- search_engine_msgf (4)
- search_engine_comet (4)
- index_peptides (1)



What's new in the RDKit KNIME nodes

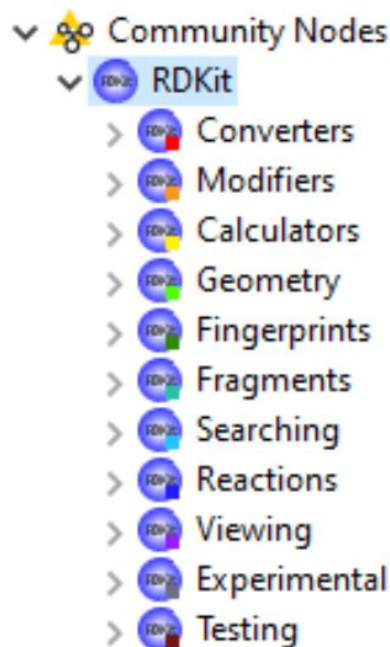
June 2022
Greg Landrum



Open-Source Cheminformatics
and Machine Learning

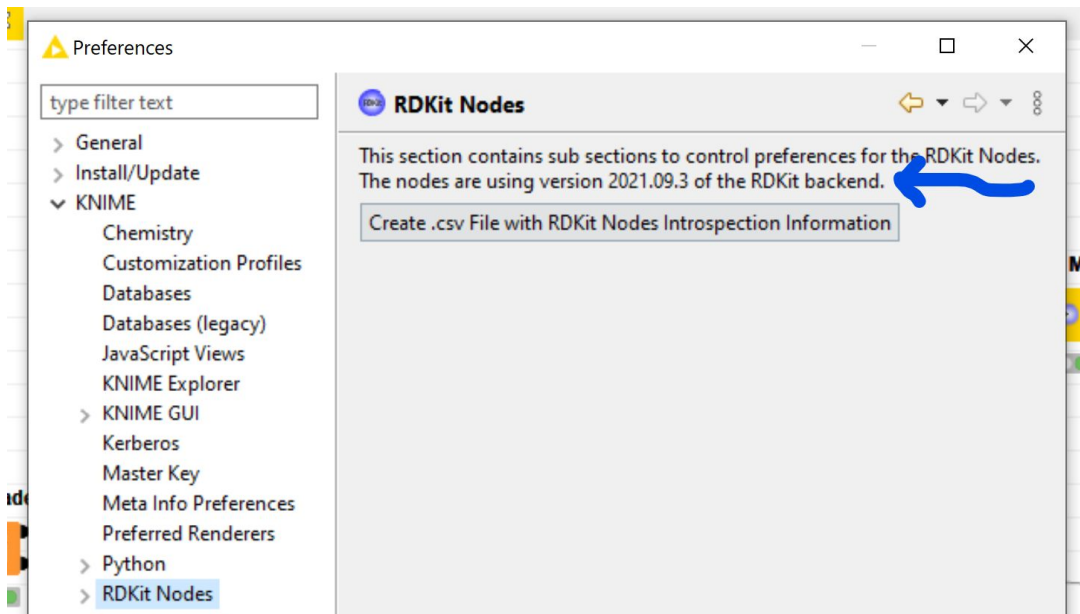
About the RDKit KNIME nodes

- KNIME trusted community nodes
- Open-source (<https://github.com/rdkit/knime-rdkit>)
- Primary developer: Manuel Schwarze (Novartis)
- Extensive cheminformatics functionality based on the RDKit cheminformatics toolkit (<https://www.rdkit.org>)



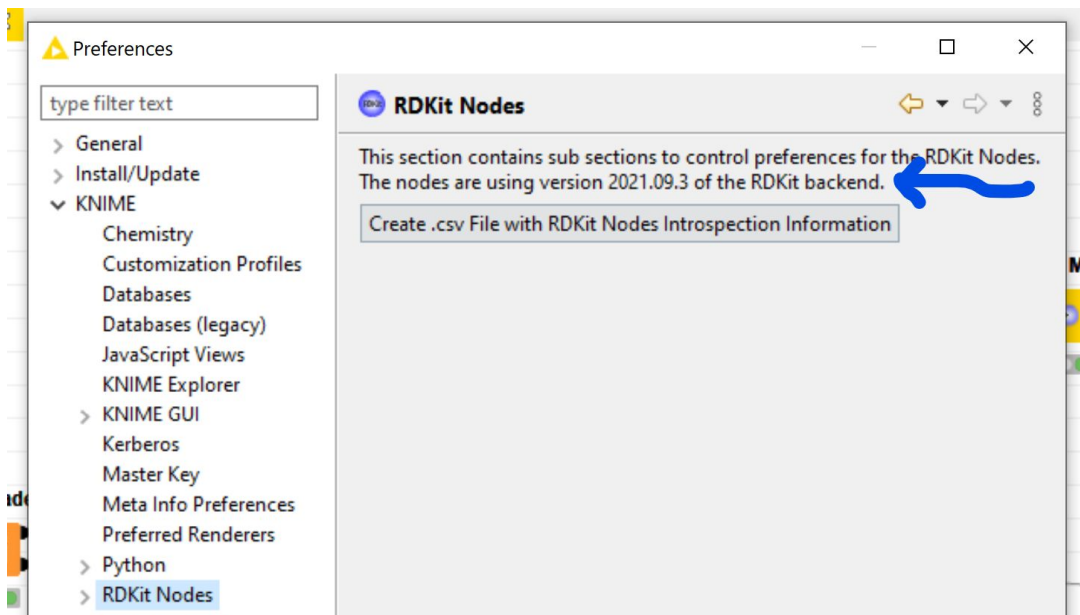
Improvements to existing nodes

Version info in preferences dialog



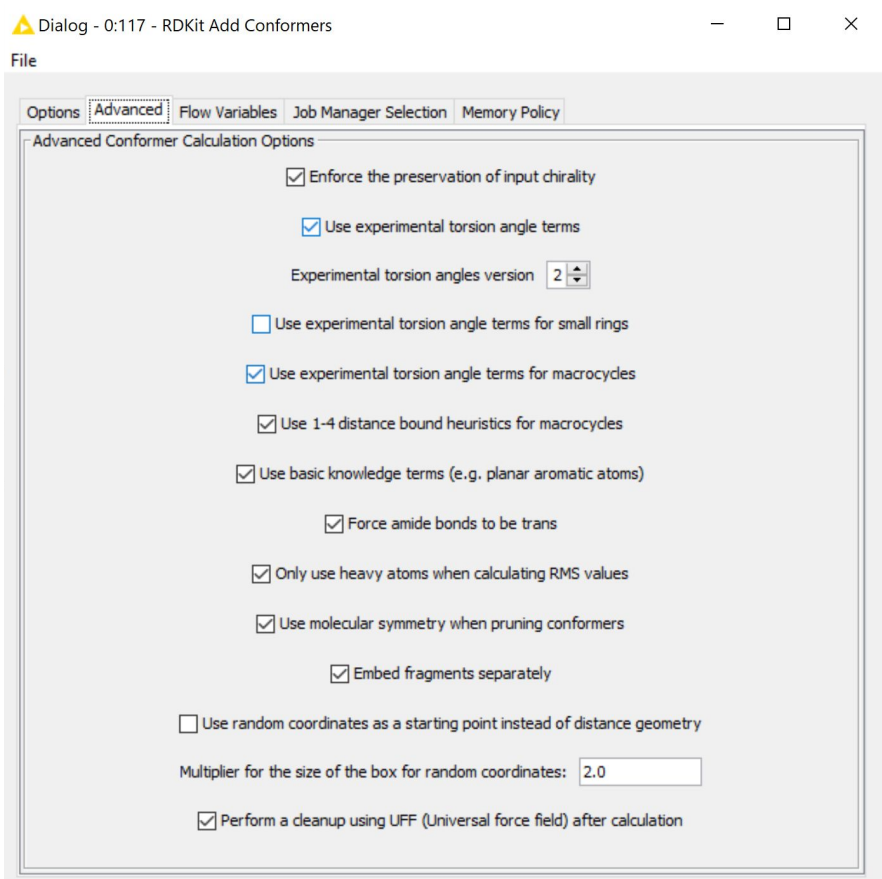
Improvements to existing nodes

Ongoing updates to current version of RDKit backend

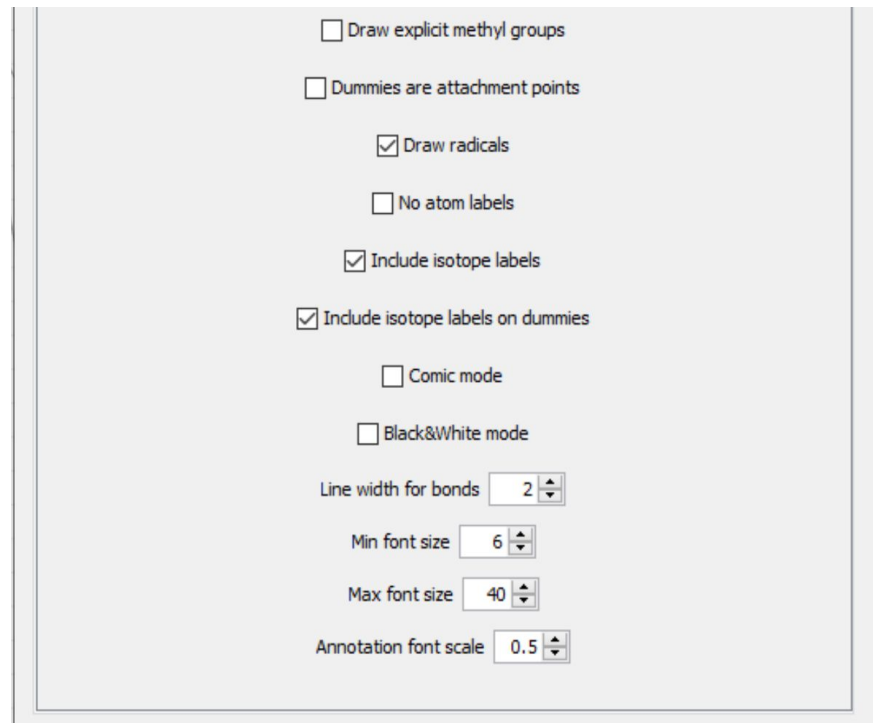
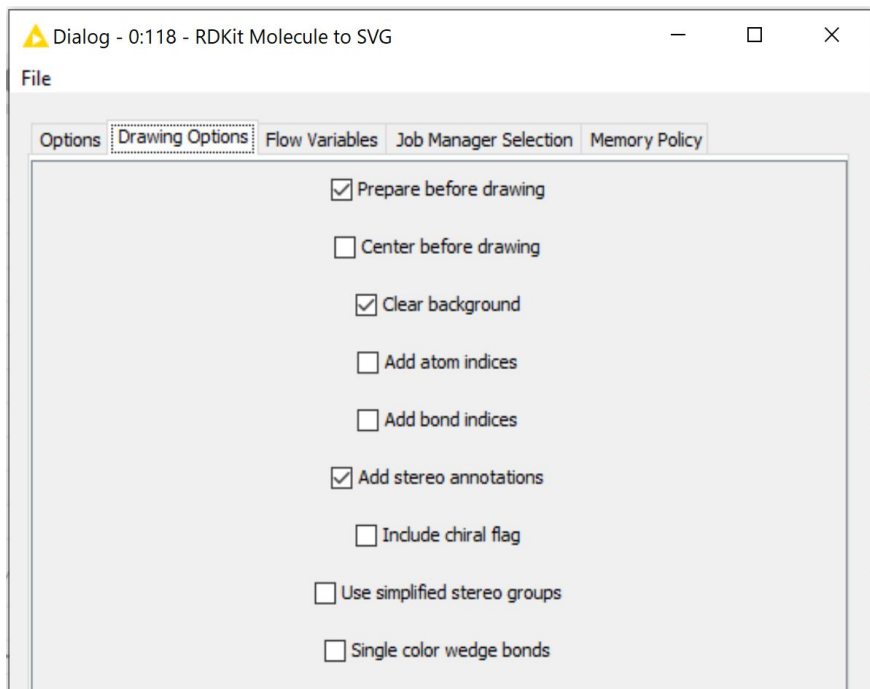
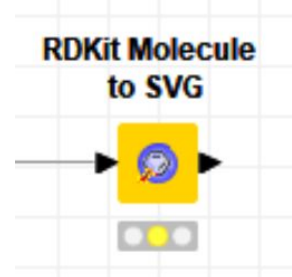


Improvements to existing nodes

RDKit Add Conformers: broad support of advanced options



New node: RDKit Molecule to SVG



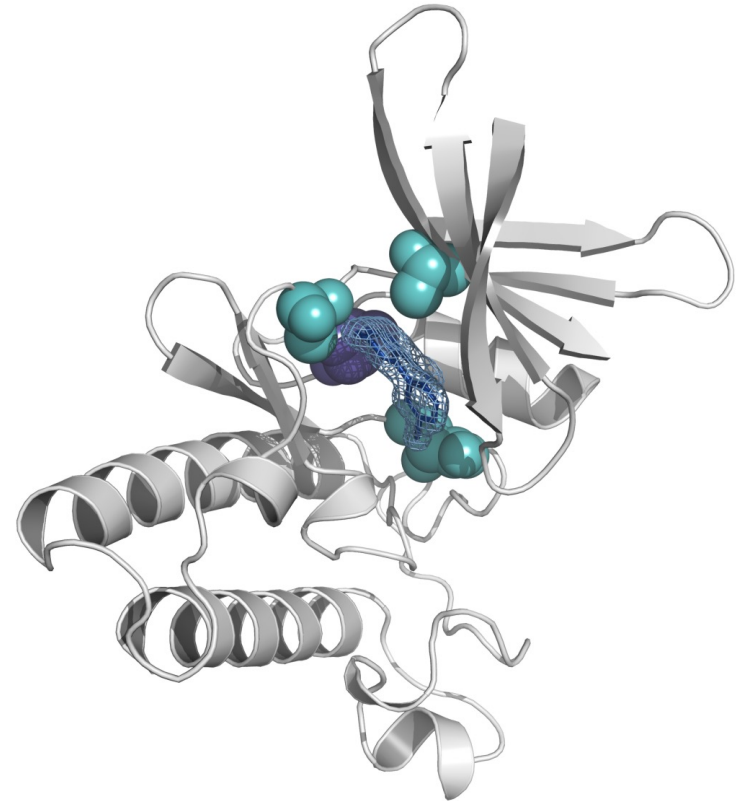
What's coming

Real soon now: updated RDKit backend version (2022.03.x)

Soon: new Python-based RDKit nodes (after KNIME v4.6 is released)

Vernalis KNIME Community Contribution

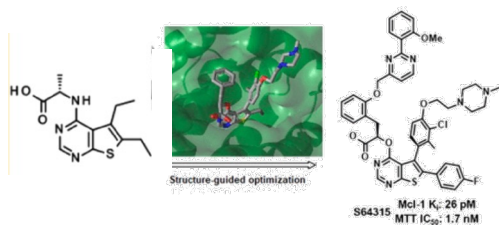
Steve Roughley
01/Jun/2022



>65 scientists based in Cambridge, UK, > 20 years in FBLD and SBDD

- A leader in fragment- and structure-based drug discovery
 - Closely integrated protein science, structural biology, biophysics, modelling and medicinal chemistry
 - A high level of innovation – developing, adapting, adopting methods to enable drug discovery
 - Proven success against highly challenging targets
 - Capabilities from target enablement to delivering clinical candidates
- Collaborations in all major therapeutic areas – joint projects from target to candidate
 - Recent funded collaborations include with Servier, Genentech, Lundbeck, AKP, Taisho, Daiichi Sankyo
 - 6 development candidates generated in the last 5 years

Challenging targets: NMR methods underpin from fragment to clinical candidate to inhibit the anti-apoptotic protein Mcl-1



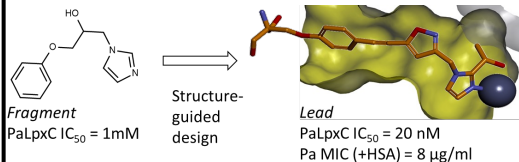
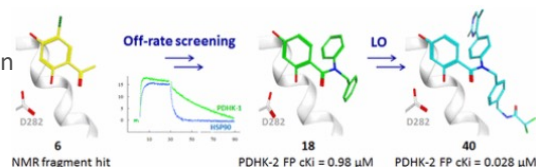
J Med Chem **62**:6913; **63**:13762

(also Nature **538**:477; ACS Omega **4**:8892)

Collaboration with Servier Oncology; clinical development partnered with Novartis

Recent publications from Vernalis Research

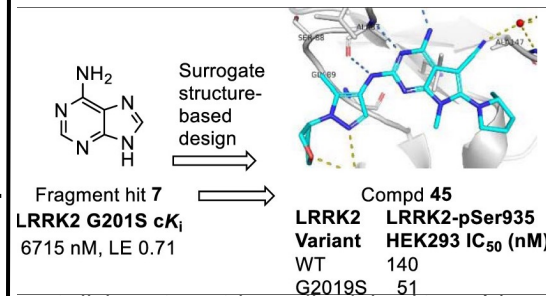
Innovation: rapid methods for hit to selective lead generation by screening of crude reaction mixtures
J Med Chem **60**: 2271;
Comm Chem **3**:122



Novel chemical entities:

discovery of non-hydroxamate inhibitors of Zn metalloprotease LpxC
J Med Chem **63**: 14805

Innovation: Design of a surrogate for LRRK2 kinase to enable structure-based discovery of brain-penetrant selective inhibitors



Collaboration with Lundbeck (Parkinson's)
J Med Chem **60**:8945; **64**:10312

Open Source at Vernalis

<https://github.com/vernalis>

- Heavy users of Open Source Software
- Contribute to open source projects
 - ‘Vernalis-owned’
 - KNIME community
 - 3D Printed flow chemistry / lab equipment
 - Digital lab
 - ‘External’ projects
 - Our developer team contributes to a range of open source project (RDKit, CDK, OpenBabel, BioJava...)
- Contributing publicly increases our code quality
- 1 KNIME-certified developer, 3 other developers, various KNIME desktop users

- Initial release 25/Jun/2012 – 1 node (PDB Connector)
- Now on v1.34.0 (released 24/May/2022)
 - 216 ‘active’ nodes (46 deprecated)
 - 9 Aggregation Operators (Bit and Byte vector operations)
 - 1 Port Type
 - 1 UI enhancement
 - 2 Extension points (for flow variable conditions and port type combiners)
- 4 Broad categories
 - Cheminformatics – RDKit-based (e.g Matched Molecular Pairs, Ertl Scaffold Trees)
 - Cheminformatics – Non-toolkit (‘Speedy SMILES’ nodes)
 - Accessing public data sources (PDB Connector nodes, EPMC Advanced Query)
 - General utility (loops, IF/Case Switches, End IF/Case, Collections, ...)

“Five Years of the KNIME Vernalis Cheminformatics Community Contribution”

Recent releases...

- UI 'Selection Modifiers'
- New Collection nodes
- New 'Binary Objects' nodes
- New IF/Case Switches / Ends
- Ertl 'Scaffold Keys' (www.peter-ertl.com)

Toolbar

Edit Menu →

- Select All (Ctrl+A)
- Select Next Node (Alt+Right)
- Move to Next Node (Ctrl+Alt+Right)
- Select All Downstream Nodes (Alt+Shift+Right)
- Select Previous Node (Alt+Left)
- Move to Previous Node (Ctrl+Alt+Left)
- Select All Upstream Nodes (Alt+Shift+Left)
- Select All Connecting Nodes (Ctrl+K)
- Invert Selection (Ctrl+I)
- Clear Node Selection (Ctrl+Shift+X)

Right-click

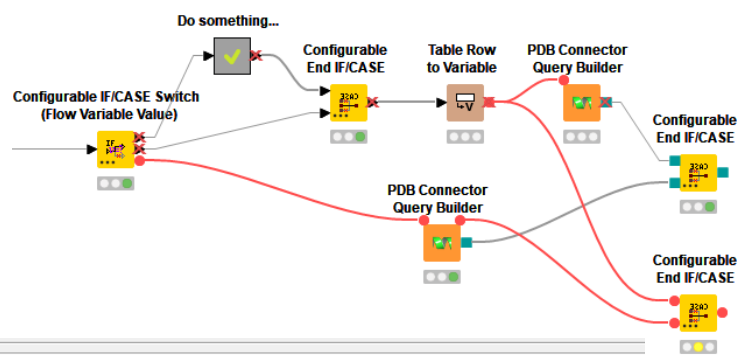
- Delete
- Extend Selection...
- Results
- Clear Node Selection (Ctrl+Shift+X)
- Select Connecting Nodes (Ctrl+K)
- Invert Node Selection (Ctrl+I)
- Select All Upstream Nodes (Alt+Shift+Left)
- Move to Previous Node (Ctrl+Alt+Left)
- Select Previous Node (Alt+Left)
- Select All Downstream Nodes (Alt+Shift+Right)
- Move to Next Node (Ctrl+Alt+Right)
- Select Next Node (Alt+Right)

Enhanced handling of updated node settings

Vernalis

- Binary Objects
 - GZip Compress Binary Object
 - GZip Decompress (Un-gzip) Binary Object
 - Zip Binary Object
 - UnZip Binary Object
- Chemistry
- Collections
 - Append to Collection
 - Collection Size
 - Collection to String
 - Column to Singleton Collection
 - Empty Collection to Missing
 - Insert into List
 - List To Set Converter
 - Mask Lists
 - Missing to Empty Collection
 - Set To List

SMILES	nAtoms	nar	nar (SVG)	near	near (SVG)	naal	naal (SVG)
<chem>c1ccccc1</chem>	9	9		6		3	
<chem>O=C1C=CC(=O)N1</chem>	6	4		0		6	
<chem>C1=CN=C1</chem>	9	9		5		4	
<chem>C1=CC=CC=C1</chem>	9	8		0		9	
<chem>C1=CC=CC=C1N</chem>	7	7		0		7	
<chem>C1=CC=C2C=CC=CC2=C1</chem>	14	14		10		1	



IF (Output port 0)

Variable Output Variable: NOT =

STRING Compare Options

Ignore Case Ignore leading / trailing whitespace

- Some possibilities...

- Binary Objects Tar Archive/Extract
- Heterocycle Regioisomer Enumerator (*c.f. J. Chem. Inf. Model., 2015, 55, 1130*)
- PDBe Webservices
- CIF Cell type / loader / typecast nodes
- 'Speedy Sequence' nodes for biological sequence manipulation
- Molecule Hash node (RDKit – NextMove hashes)
- More Screening Plate nodes
- Syrris Asia Flow Chemistry rack nodes

Tar Binary Object












UnTar Binary Object




HREMS Heterocycle
Regioisomer Enumerator



- ≡ Speedy Sequence
 -  Loop Start (Speedy Sequence Generator)
 - Speedy Sequence Alignment
 -  Speedy Sequence Alignment Visualisation
 -  Speedy Sequence Analysis
 -  Speedy Sequence Generator
 -  Speedy Sequence to SMILES

- ▼  Syrris Racks
 -  Append Syrris Rack IDs
 -  List Syrris Rack Sizes
 -  Syrris Rack IDs

- ▼  Plates
 -  Append Plate Well IDs
 -  List Plate Sizes
 -  Plate Well Expand Matrix
 -  Plate Well IDs
 -  Plot Plate Map
 -  Reformat Plate Well IDs

- Contact us...
 - KNIME Forum - <https://forum.knime.com/c/community-extensions/vernal/41>
 - GitHub - <https://github.com/vernal/vernal-knime-nodes/>
 - LinkedIn - <https://www.linkedin.com/in/stephen-roughley/>
 - E-Mail – address in the node description of every Vernalis node
- Acknowledgements
 - James Davidson (Vernalis)
 - Bernd Wiswedel, Thorsten Meinl, Gabriel Einsdorf, Tobias Koetter (KNIME)
 - Greg Landrum, Manuel Schwarze (RDKit)
 - KNIME users who give us feedback

Current Medicinal Chemistry, 2020, 27, 6495-6522

REVIEW ARTICLE



Five Years of the KNIME Vernalis Cheminformatics Community Contribution

