Vernalis
RESEARCH
part of Vernalis plc

What's Cooking for KNIME at Vernalis?

Steve Roughley

KNIME Cheminformatics SIG

London, UK

18th October 2016

THE QUEEN'S AWARDS
FOR ENTERPRISE:
INTERNATIONAL TRADE
2014

Outline

Vernalis RESEARCH part of Vernalis plc

- Introduction to Vernalis
- Changes since last SIG
  - Matched Molecular Pairs
  - PDB Connector
  - PMI
- Forthcoming Releases
  - SpeedySMILES
  - File Readers
  - Benchmarking
- Future directions
  - PDBe Webservices
  - UI Enhancements

Introduction to Vernalis

Vernalis RESEARCH part of Vernalis plc

- ~ 60 staff in research
  - Based in Cambridge, UK (*Granta Park*)
  - Structure-based drug discovery since 1997
  - Fragment-based lead discovery since 2001
  - Biophysics & Structural Biology expertise
    - X-ray, NMR, ITC and SPR
- Portfolio of discovery projects
  - Seven development candidates generated in the past seven years
  - Structure, fragments and modelling integrated with medicinal chemistry
  - Collaborations with large and small pharmaceutical companies
- Trusted community contributor since June 2013
  - 2 KNIME-trained developers

Current release status

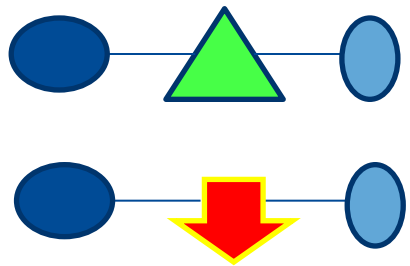
Vernalis RESEARCH part of Vernalis plc

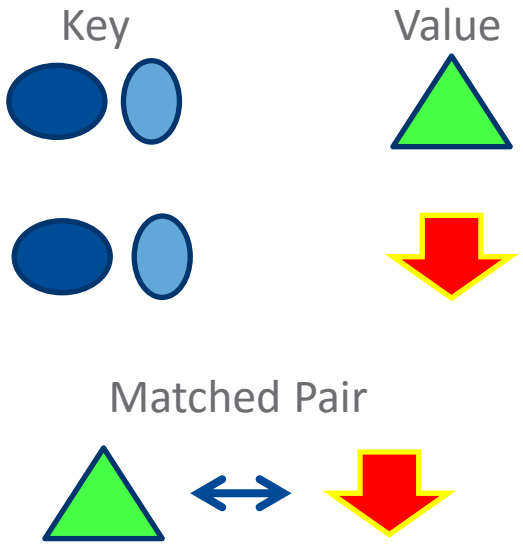
- Version 1.7.0
- New releases since last meeting:
  - Improvements to Matched Molecular Pairs
  - Major Re-write of PDB Connector
  - Added additional RCSB PDB Webservice nodes
  - Added Principal Moments of Inertia nodes (Community request)
- Now stands at 65 nodes in 6 plugins
  - Several nodes now deprecated following transfer to KNIME core
- Many ~275 released internally (and more ongoing)
  - Permission to release a significant number of new nodes in coming weeks

Matched Molecular Pairs (MMPs)

Background

- Process involves 2 steps
- Fragment molecules
  - Break molecules along matching bonds
  - Gives ‘Key’-’Value’ Pairs
- Create matched Pairs
  - Two identical Keys with different values





Matched Molecular Pairs (MMPs)

Improvements made in recent releases

Vernalis RESEARCH part of Vernalis plc

- Moved filtering of fragmentations forwards in process
  - Significant performance improvements for filtered fragmentations (Unchanging HAC and Changing/Unchanging HAC ratio)
- Added check for presence of unassigned double bonds before trying to do any assignment
  - Upto 25% performance improvement
- Filter/Splitter nodes no longer pass multicomponent structures
- Improved garbage collector class to remove possible references to deleted native objects

Matched Molecular Pairs (MMPs)

Bug fixes in recent releases

Vernalis RESEARCH part of Vernalis plc

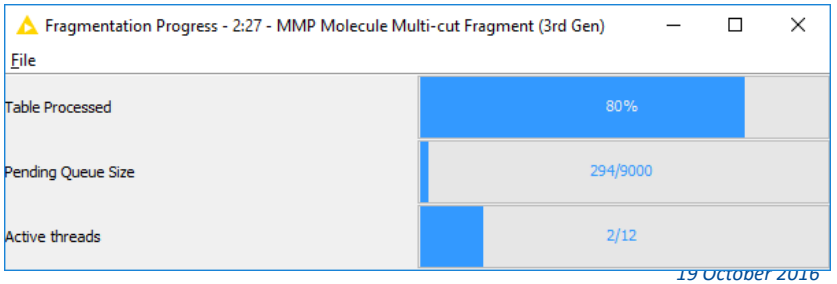
- Fragmentor nodes no longer fail with multicomponent structures
- Fixed bug in which certain double bond geometries to near-symmetrical aromatic rings became corrupted causing crashes
- Fixed bug causing multi-fragment node to only perform two cuts to a bond ([2*]-[1*] value) when the maximum number of cuts was 2
- Fixed loss of stereointegrity issue in MMP fragmentation
- Fixed canonicalisation issue resulting from above fix (NB 'Values' will not be differently canonicalised to previously)
- Fixed graceless failures for some salted forms and apparent successful failures of other salted forms
- Fixed loss of double bond geometry (Change in RDKit toolkit behaviour)
- Fixed scrambling of attachment point labels and node failures in certain circumstances
- MMP canonicalisation is modified to deal with duplicate components in the 'Key'

Matched Molecular Pairs (MMPs)

Forthcoming update (est. October 2016)

Vernalis RESEARCH part of Vernalis plc

- Despite significant effort, MMP nodes were still leaking memory
  - Now identified as issue in Java SWIG wrappers of RDKit
    - Fixed – Release imminent (Thanks Greg!)
- H-Removal has been reverted to RDKit removeHs method
- Additional options to skip high-complexity molecules (based on number of combinations of matching bonds)
- Added a progress view showing threads/completed rows queue status
  - Incoming rows are processed by passing to an available thread
  - Output is held in queue until it's place in the output table is reached
  - A complex molecule can effectively block other threads if queue is filled
    - Bigger queue (see preferences), which takes more memory but keeps more threads active
    - Sort input table by complexity – similarly complex molecules are processed at same time
    - Further node re-engineering…
      - Add option to maintain/ignore incoming row order
      - Process table row by row, distributing fragmentations across threads
- NodeModel class / Internal worker classes now generic
  - Allows alternative toolkit implementations
  - CDK Implementation in-house
    - Requires modifications to CDK
    - Significantly slower than latest RDKit version

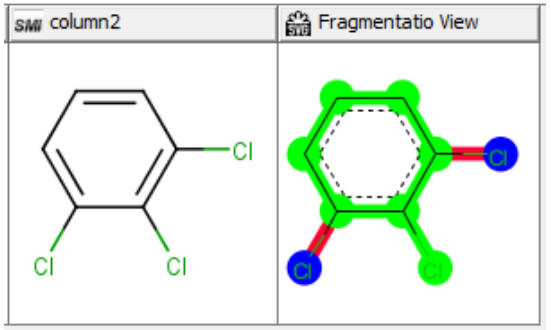


Matched Molecular Pairs (MMPs)

Possible Future directions…

Vernalis RESEARCH part of Vernalis plc

- Add column chooser to allow user to keep incoming table columns associated with fragmented molecules
- Add Table to Progress view showing which input rows being processed
- Fragmentation Visualisation
  - E.g. The mock-up using RDKit Molecule Highlighting Node
- Re-working parallelisation
  - Add option to preserve/ignore incoming row order
    - Would dispense with queue blocking as thread releases output directly to output tables rather than queue
  - Process table row-by-row, distributing fragmentations across threads
    - Complex molecules use all threads for longer



PDB Connector Re-write

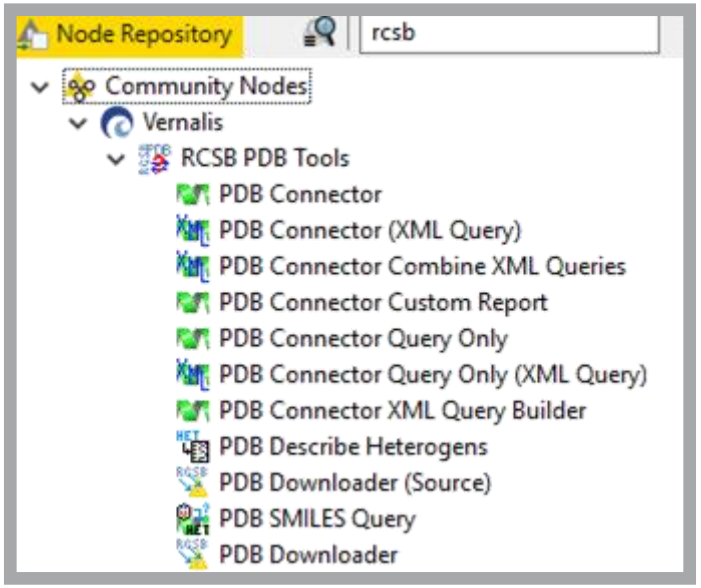
Vernalis RESEARCH part of Vernalis plc

- PDB Connector was original Vernalis node
  - Originally written by Dave Morley (Enspiral Discovery, Ariana Pharma, formerly at Ribotargets, where wrote rDock)
  - 3 functions in 1 node…
    - Query building
    - Query execution (webservice call)
    - Report generation (different webservice call)
  - XML Query variant subsequently added (lacks the query building)
    - Allowed databasing of regular queries or (inefficient!) generation of 2nd report
- Increasingly difficult to maintain as RCSB change webservice behaviour
  - Various “sticking-plaster” fixes applied
  - Recent changes led to near-complete rewrite
    - Many new features added
    - In addition to keeping the popular existing nodes, single nodes added with each feature alone, and each logical combination

PDB Connector Re-write

Vernalis RESEARCH part of Vernalis plc

- New nodes
  - PDB Connector (XML Query) **(Query Only)**
  - Query Builder
    - Returns the XML query as a flow variable
  - Custom Reporter
    - Runs custom report
  - Query Combiner node
    - Combines two XML advanced queries AND or OR logic



- All Query generator or executor nodes now have 2 views:
  - 'Logical' query view
  - XML query view
- Some column types are now returned as collection cells
  - Fixes parsing errors of some numeric types which are occasionally deliminated values
  - This was the start of the re-write!
- Additionally, “new” nodes have been released to access the PDB 'Describe Heterogens' webservice and PDB SMILES Query webservice

PDB Connector Re-write

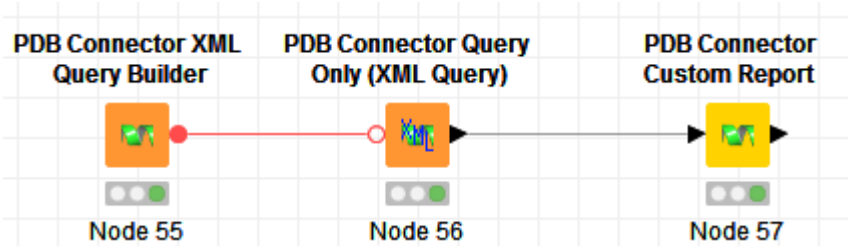
Vernalis RESEARCH part of Vernalis plc

- POST and GET methods now allow chunking of report table
  - Resolving strange behaviour for large result sets due to change in behaviour of webservice
- Execution can now be cancelled at any point during the service request
  - No longer stuck waiting for an incorrect query to complete before cancellation comes into effect
- All dialogs (where relevant) have options to
  - Change variable name of XML Query
  - Copy query to clipboard (can paste into other nodes or RCSB website)
- Data parsing problems are now added to a column at the end of the report table in addition to the console
  - Previously only reported in console
- When both a query and a report are run, the execution balance has been adjusted from 30/70 to 10/90
- Query results are parsed directly to a BufferedDataContainer, reducing memory overhead for large result sets
  - Previously, all hits were stored in memory
  - Allowed separation of all 3 functions

New PDB Connector nodes

More 'KNIME-like'

Vernalis RESEARCH part of Vernalis plc



Collection columns

Custom report table - 2:57 - PDB Connector Custom Report

File

Table "default" - Rows: 32 | Spec - Columns: 12 | Properties | Flow Variables

Row ID	Structu...	(...) Macrom...	(...) Structure Author	(...) !
Row0	358.61	[Protein]	[Dymock, B.W.,Barril, X.,Brough, P.A.,...	?
Row1	646.35	[Protein]	[Dymock, B.W.,Barril, X.,Brough, P.A.,...	?
Row2	067.53	[Protein]	[Brough, P.A.,Aherne, W.,Barril, X.,...]	?
Row3	059.89	[Protein]	[Brough, P.A.,Aherne, W.,Barril, X.,...]	?
Row4	809.49	[Protein]	[Brough, P.A.,Barril, X.,Borgognoni, J.,...	?
Row5	694.49	[Protein]	[Brough, P.A.,Barril, X.,Borgognoni, J.,...	?
Row6	782.49	[Protein]	[Brough, P.A.,Barril, X.,Borgognoni, J.,...	?
Row7	946.18	[Protein]	[Brough, P.A.,Barril, X.,Borgognoni, J.,...	?
Row8	955.38	[Protein]	[Brough, P.A.,Barril, X.,Borgognoni, J.,...	?
Row9	969.23	[Protein]	[Brough, P.A.,Barril, X.,Borgognoni, J.,...	?
Row10	082.39	[Protein]	[Brough, P.A.,Barril, X.,Borgognoni, J.,...	?
Row11	802.31	[Protein]	[Roughley, S.D.,Hubbard, R.E.]	?
Row12	706.25	[Protein]	[Roughley, S.D.,Hubbard, R.E.]	?
Row13	690.25	[Protein]	[Roughley, S.D.,Hubbard, R.E.]	?
Row14	664.21	[Protein]	[Roughley, S.D.,Hubbard, R.E.]	?
Row15	660.22	[Protein]	[Roughley, S.D.,Hubbard, R.E.]	?
Row16	791.48	[Protein]	[Roughley, S.D.,Hubbard, R.E.]	?

XML Query - 2:54 - PDB Connector

File

```
<orgPdbCompositeQuery version="1.0">
    <queryRefinement>
        <queryRefinementLevel>0</queryRefinementLevel>
        <orgPdbQuery>
            <queryType>org.pdb.query.simple.AdvancedKeywordQuery</queryType>
            <keywords>HSP90</keywords>
        </orgPdbQuery>
    </queryRefinement>
    <queryRefinement>
        <queryRefinementLevel>1</queryRefinementLevel>
        <conjunctionType>and</conjunctionType>
        <orgPdbQuery>
            <queryType>org.pdb.query.simple.AdvancedAuthorQuery</queryType>
            <searchType>0</searchType>
            <audit_author.name>Roughley,SD</audit_author.name>
            <exactMatch>false</exactMatch>
        </orgPdbQuery>
    </queryRefinement>
</orgPdbCompositeQuery>
```

Logical Query - 2:54 - PDB Connector

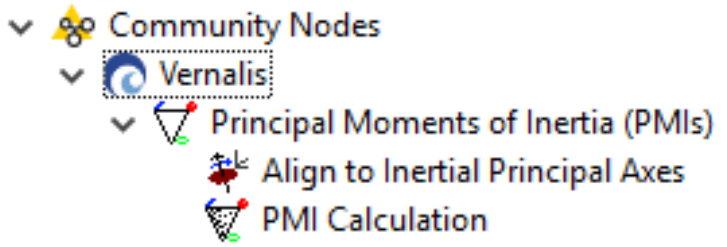
File

```
Text Search:
    Text: "HSP90"

AND

Author Name:
    Search Type: "All Authors"
    Author: "Roughley,SD"
    Exact Match: "false"
```

Principal Moments of Inertia (PMI, nPMI)

Vernalis RESEARCH part of Vernalis plc

- Release following community request
  - Already written and released in-house



- PMI is a measure of (molecular) shape
  - For near-spherical shapes, the 3 PMI are approximately equal
  - Other combinations indicate rod-like or disk-like shapes
  - The 'normalised' PMI are more commonly used
    - The smaller two values are divided by the largest to give the nPMIs 'npr1' and 'npr2'
    - Comparable between molecules as always in range 0 - 1 and 0.5 – 1
    - Node will calculate either/both
- Also released an "Align to Principal Axes" node at same time
- Many further nodes for molecular shape descriptors and PMI Plots are available in-house
  - Release pending ongoing publication

WHAT'S COOKING?

Nodes with release-permission

Nodes in development

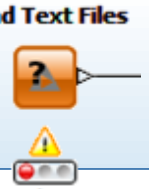
"Speedy SMILES"

Est. release Oct. 2016

Vernalis RESEARCH part of Vernalis plc

- NOT Another chemical toolkit!
  - Based entirely on String processing
    - Small memory/time footprint
    - Lightweight pre-processing before taking into a toolkit for 'heavy lifting'
  - Nodes for fast processing of SMILES
    - Filters/Splitters
    - Manipulators (e.g. invert stereochem, desalt)
    - Property calculators (e.g. charges, HAC, chiral centres etc)
  - FAST: 1.4m molecules (ChEMBL 19)
    - Desalted - ~14 s
    - Split chiral/non-chiral molecules - ~11 s
    - Heavy Atom Count ~6 seconds
  - First Vernalis nodes to be Streaming API-compliant
    - ~35% faster when multiple nodes are streamed sequentially

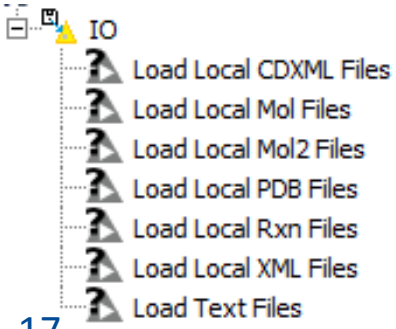
Text file loading source nodes

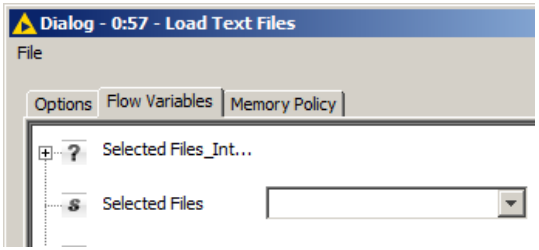


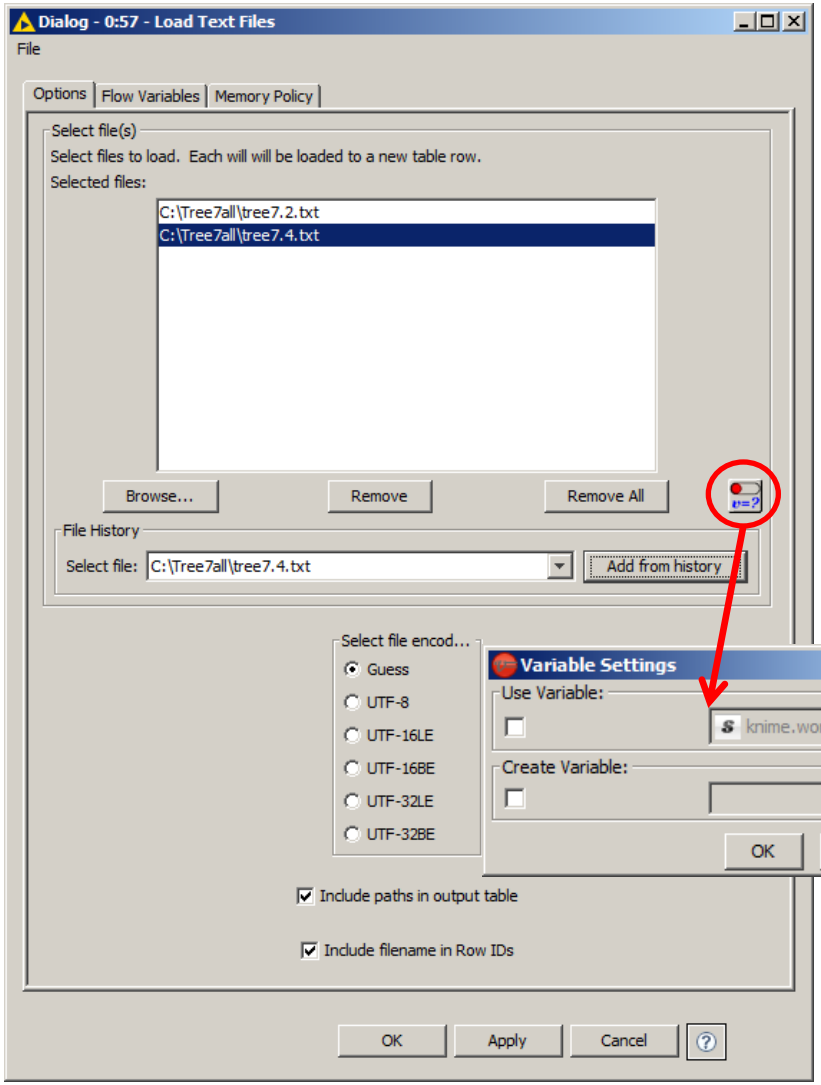
Est. Release Oct 2016

Vernalis RESEARCH part of Vernalis plc

- Complements released Load text-based files manipulator node
- StringArray, but with *single* flow variable replacement
  - List of files can be passed via flow variable without the flow variables tab 'Array problem'
  - New DefaultDialog component and SettingsModel
    - Available for use in other plugins from our core plugin
- Also available in other 'flavours':





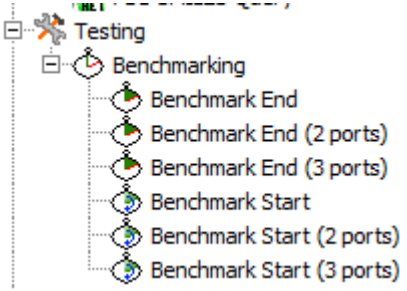


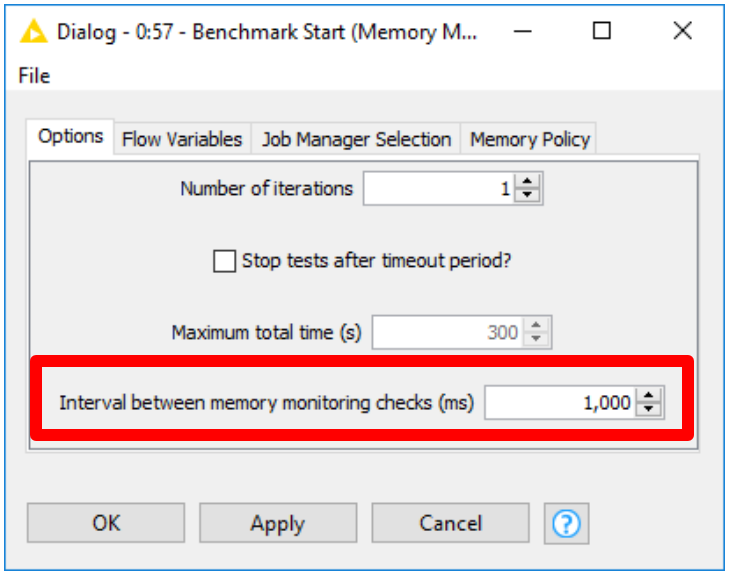
Benchmarking Nodes

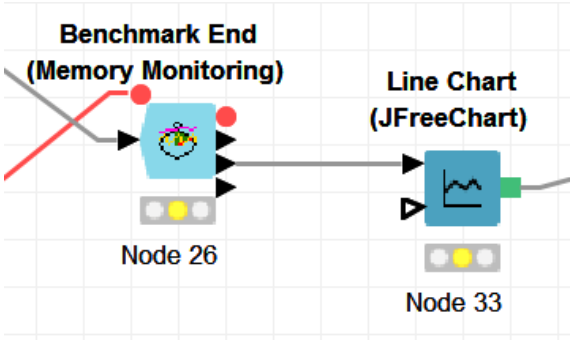
Est. Release Oct. 2016

Vernalis RESEARCH part of Vernalis plc

- Time execution of intervening node(s)
  - Multiple iterations
  - Optional time limit (next iteration doesn't start if already elapsed)
  - Min/Max/Mean iteration time
  - Individual iteration
  - Last iteration table(s) passed through
  - New Nodes added with memory monitoring
    - Output to Loop End extra table
    - Easily fed into e.g. LineChart (JFreeChart) node for visualisation
  - New 'Memory Use' node







PDBe Webservices

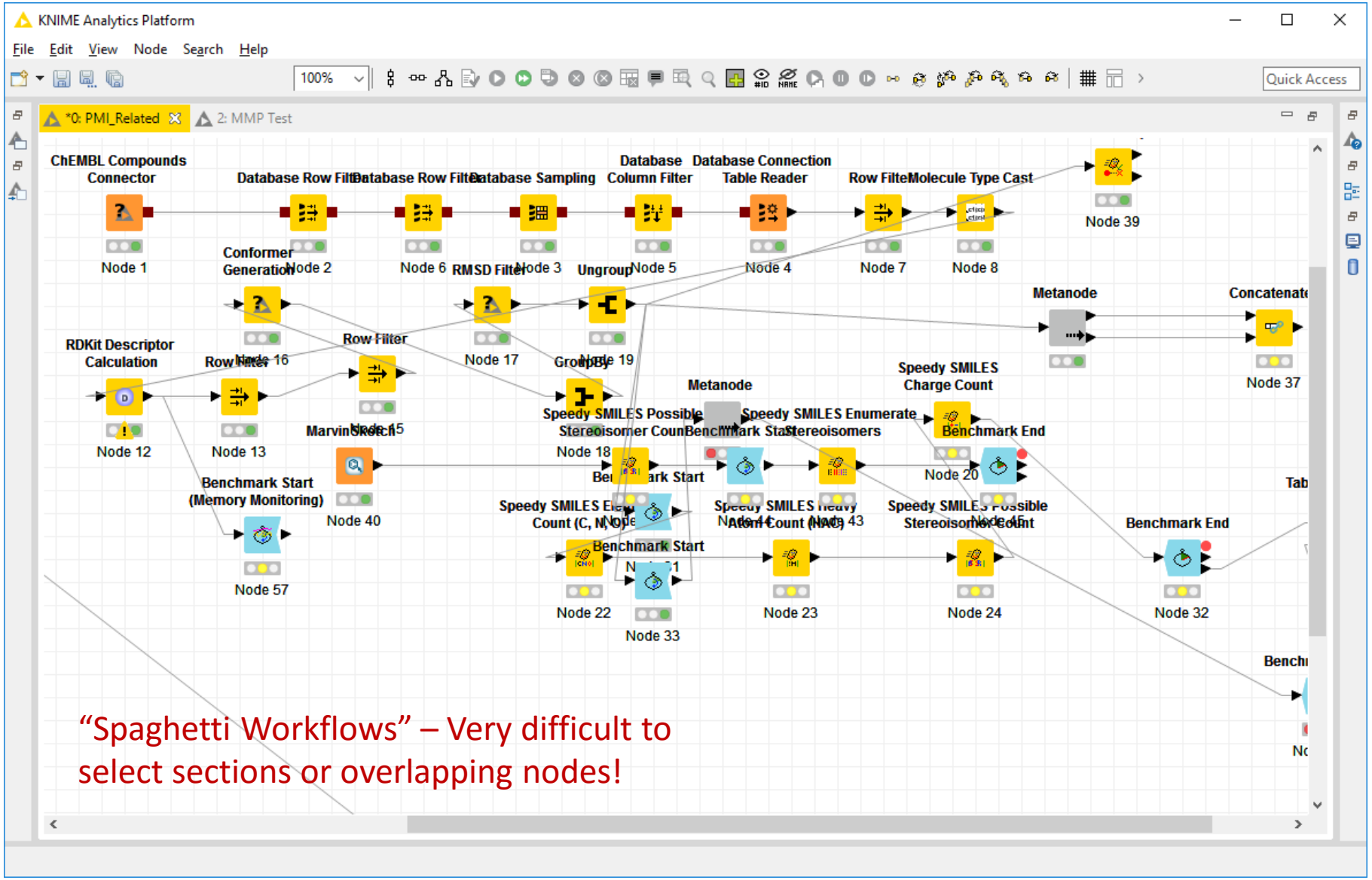
In internal development

Vernalis RESEARCH part of Vernalis plc

- The 'PDB in Europe' (PDBe) has numerous webservices
  - http://www.ebi.ac.uk/pdbe/pdbe-rest-api
  - Responsive support team based at Hinxton, UK
- Architecture built to define nodes based on service URL and column outputs
  - Services all return JSON, which sometimes requires re-shaping/reformatting in addition to result extraction/conversion to KNIME DataCells
  - JSON manipulators might be amenable to implementation as stand-alone nodes too
- Most services now implemented
  - Working on testing framework
    - Data updates happen with sufficient regularity as to make test workflows near-impossible to design otherwise!
- Longer term possibility of user-defined webservice nodes via XML config file

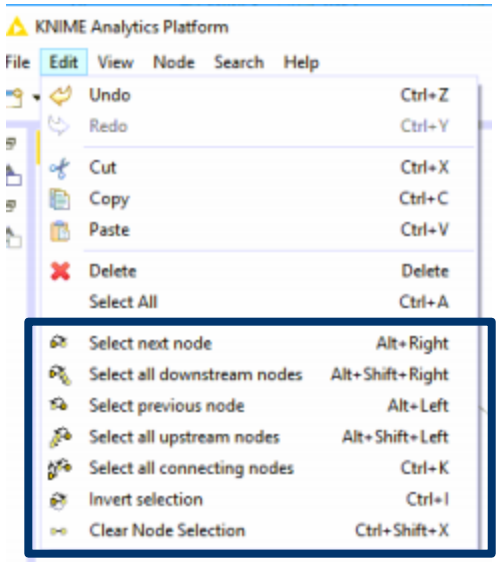
I'm sure we've all ended up with this…

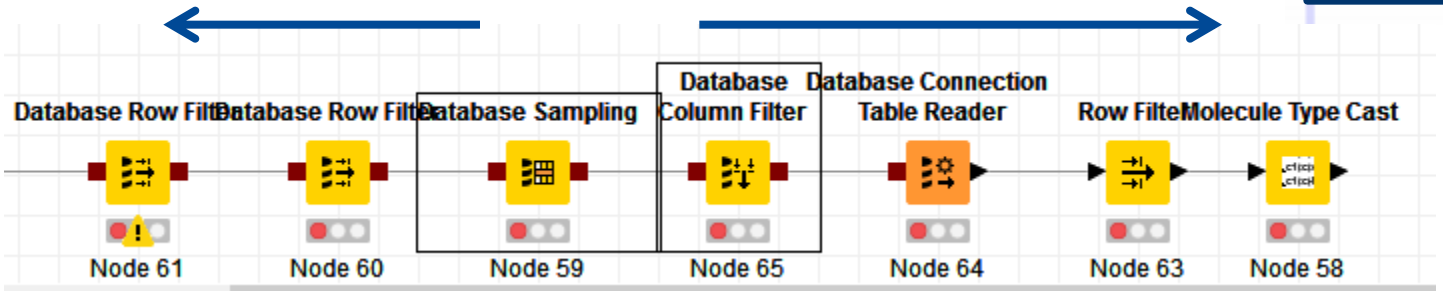
Vernalis RESEARCH part of Vernalis plc

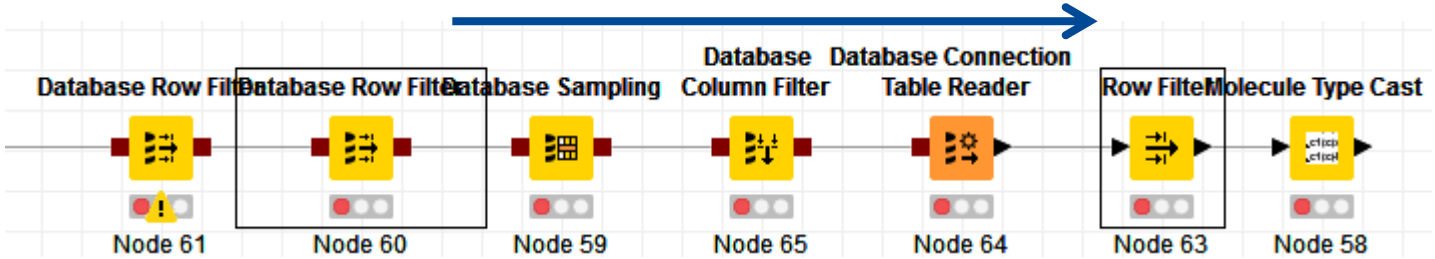


UI Enhancements

- Selection modifiers
  - Select next/all upstream/downstream node(s)
  - Select all connecting nodes
  - Invert selection
  - Clear Selection
  - Still a little buggy – particular with metanodes







Summary

Vernalis RESEARCH part of Vernalis plc

- Release of new nodes to nightly build expected by end of this week
  - 'Speedy SMILES' (14 nodes)
  - File Readers (7 nodes)
  - Benchmarking nodes / with memory monitoring (8 nodes)
  - Side effect is all file nodes now accept '`knime:/`' URL protocols
- Matched Pairs update within next 2-3 weeks
  - Complete testing first
- Questions?

- knime@vernalis.com
- S.roughley@vernalis.com
- KNIME Forum

Fingerprint nodes

PENDING RELEASE

Vernalis RESEARCH part of Vernalis plc

- BitVector and ByteVector
- Type /format conversions
- Logic operations
- Properties

- Also aggregators (GroupBy, Column combiner nodes)

- Some functionality released by KNIME at around time permission for release obtained
  - "De-duplicate"