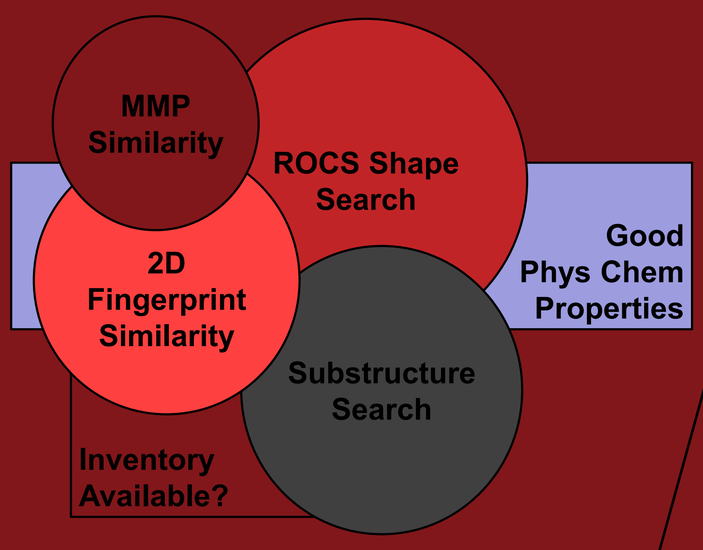
You can't always get what you want, but KNIME gets what you need!

Luke Bullard [1], James A Lumley [1], Thomas Wilkin [1]
Dana Benesh[2], Sonia Gutierrez [4], Jason Ochoada[2], David Thorner [3], Jibo Wang[2]

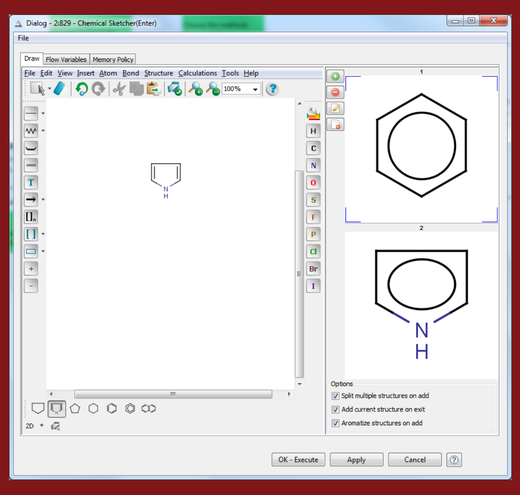
[1] LRL Research IT & [3] LRL Computational Chemistry, Lilly Research Centre, Erl Wood Manor, Windlesham, Surrey, UK
[2] LRL Computational Chemistry, Lilly Corporate Center, Indianapolis, IN 46285 USA
[4] LRL Computational Chemistry, Avenida de la Industria, 30 – 28108 Alcobendas, Madrid, Spain

The Problem: Too many compounds to test. Many compound searching and scoring tools running on different IT systems. Extremely difficult to standardise the search process. No easy interface for a non-expert user

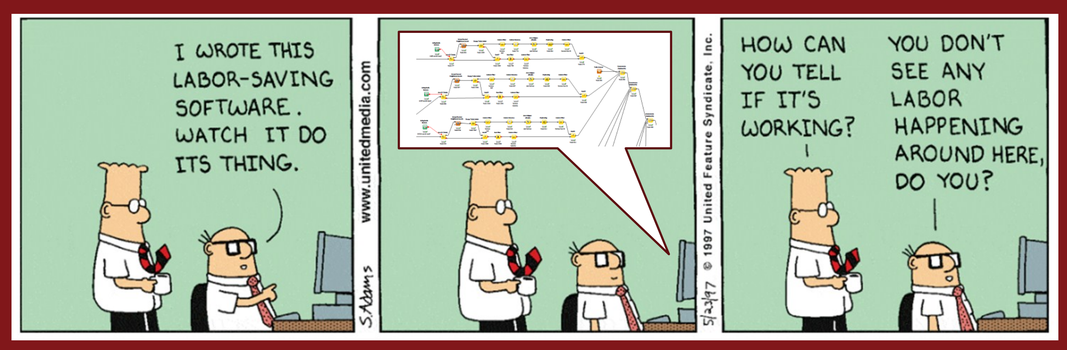
The Workflow: A flexible framework for abstracting out the most interesting compounds from Internal, External, and Virtual sources given a set of input starting points. Fast selection using consensus from all search methods. Also allows facile swapping of search methods in/out.

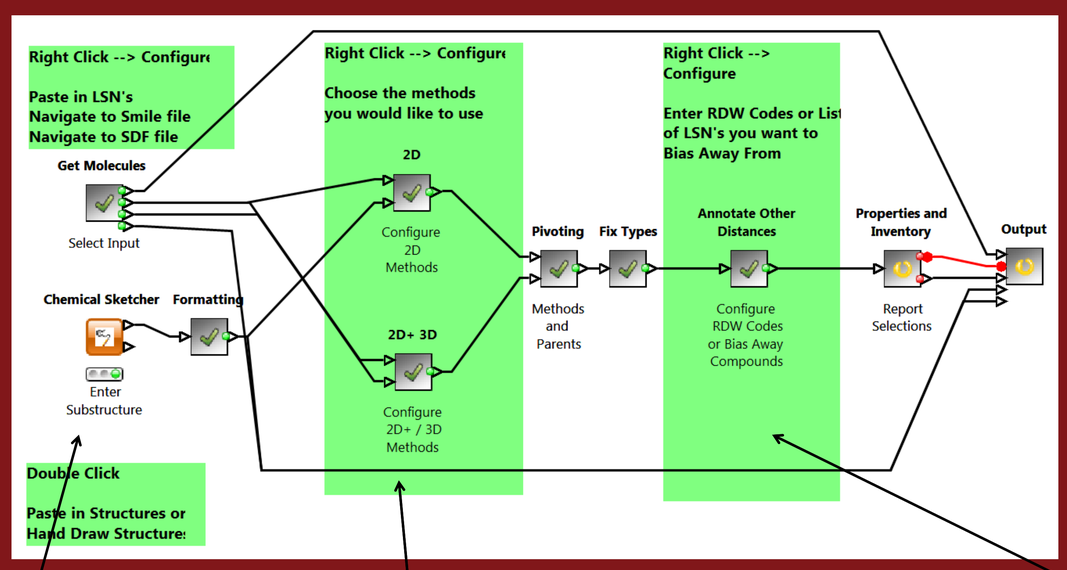


Input: A user sketches a number of chemical substructures and/or specifies a set of Lilly compound Identifiers from flat file or free text input (quickform)

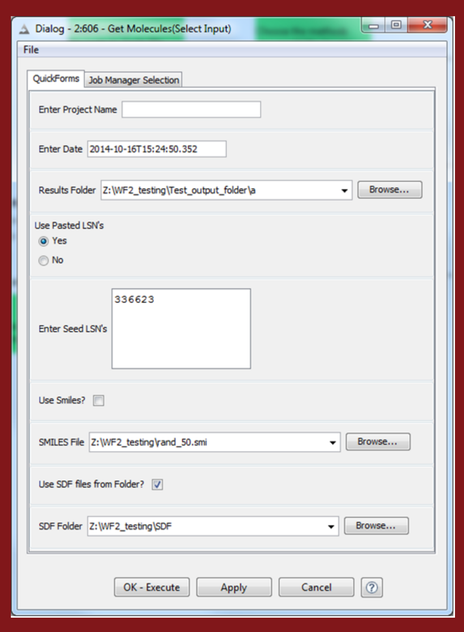


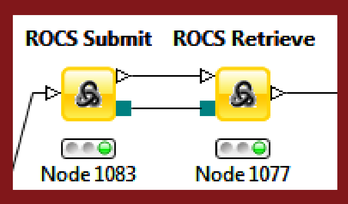
Opensource Contributions: Lilly develops and supports the 'Erl Wood' (EW) extensions that include the above multi molecule sketcher node, utilities for accessing authenticated data sources, a framework for handling molecule conversions in the node dialogue (based on chemaxon) and other nodes focused on cheminformatics tasks: http://tech.knime.org/community/erlwood

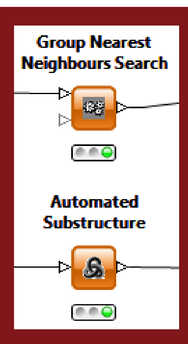




Methods: Input molecules are used to search for 'Near Neighbours' in 2D and 3D space using fingerprint similarity and ROCS shape based searching

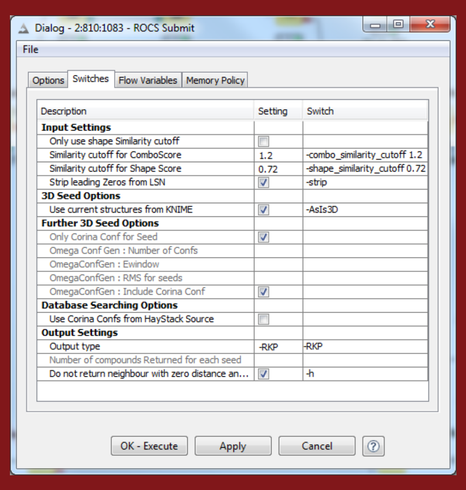




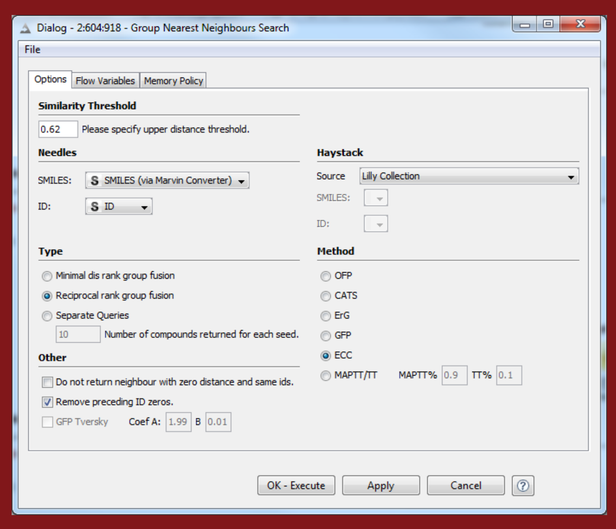


Meta nodes allow workflow simplification and easy workflow configuration for non-advanced users

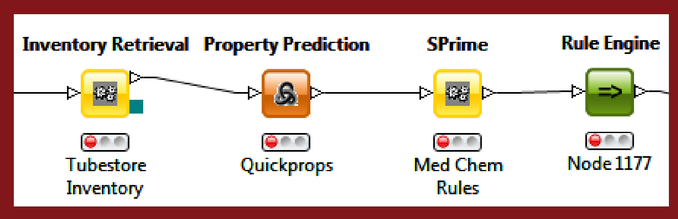
XML Based Node Config & Cluster Submission: Internal nodes and matching SOA layer provide cluster job submission e.g.: "ROCS Submit". Node UI development for complex command lines is simplified via XML config for easy switch definition before submission of command line to cluster



Molecule Conversion Framework: GroupNN (and other internal nodes) use the molecule conversion framework' to take any input molecule format, convert to smiles then submit to the GroupNN node backend server

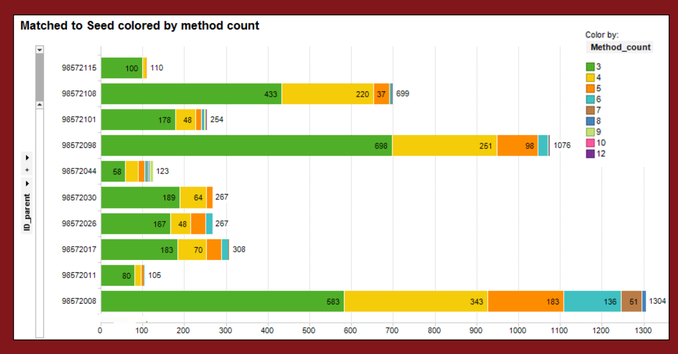


Bias Away: Annotate fingerprint distances to compounds already tested. Specify Lilly Assay code or a input a list of Lilly Identifiers. Remove if tested already



Selected compounds are checked for (a) internal inventory via SOA access to current sample management data and (b) filtered by calculated properties via SOA access to our proprietary property prediction engine SPrime.

Output: Spotfire view of results to show total number of compounds found for each input 'seed' ID. The colours indicate the number of compound found by 3,4,5, etc. methods



The Future State:
We're extending the KNIME cluster execution code to allow us to submit workflow chunks to our internal HPC environment via qsub. This strategy will allow searching multi Millions and Billions of compounds. Webportal deployment is also being considered