

Data KNIME-ing Discovery: Workflows in Medicinal Chemistry

Mike Bodkin

What a Chemist needs to know about
Cheminformatics and SAR

Wednesday 11 June 2008
SCI, London, UK

Answers That Matter.

Data Pipelining



Available online at www.sciencedirect.com



Computational
Biology and
Chemistry

Computational Biology and Chemistry 31 (2007) 305–319

www.elsevier.com/locate/complchem

Review Article

Workflow based framework for life science

Abhishek Tiwari^{a,*}, Arvind K.T. Sek

^a Informatics Division, GVK Biosciences, Hyderabad 50003

^b School of Informatics, Indiana University (IUPUI), Indiana

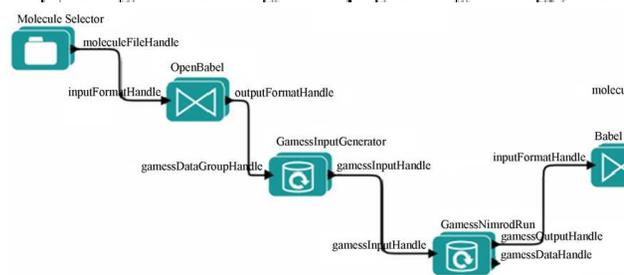
Received 19 February 2007; received in revised form 15 June 2007; accepted

Abstract

Workflow technology is a generic mechanism to integrate diverse types of available resources (different services) which facilitate knowledge exchange within traditionally divergent fields such as computational science, physics, chemistry and statistics. Researchers can easily incorporate and adapt their own research protocols for scientific analysis. Application of workflow technology has been used in large-scale gene expression analysis, proteomics, and system biology. In this article, we have discussed the trends in applications of workflow based systems.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: Workflow technology; Data pipelining; Data streaming; Visual programming; Ontology; Semantic



Address <http://www.qsarworld.com/qsar-workflow1.php>

Home | About QSAR World

QSAR WORLD

Wednesday, May 28, 2008

- Technical Sections
 - > *in silico* Chemistry
 - > Machine Learning
 - > Statistics
- Resources
 - > Literature
 - > Datasets
 - > Web-based Resources
 - > Business and Economics
 - > Events Calendar
 - > Publications & Scripts
 - > Archives

[Home](#) > Workflow and Pipelining in Cheminformatics

Workflow and Pipelining in Cheminformatics

[Wendy Warr](#), editorial advisor of QSARWorld, writes on the workflow paradigm as a mechanism to integrate different data resources, softwares and algorithms, web services etc. Read on...

[Download PDF Version](#)

The workflow paradigm is a generic mechanism to integrate different data resources, software applications and algorithms, Web services and shared expertise. Such technologies allow a form of integration and data analysis that is not limited by the restrictive tables of a conventional database system. They enable scientists to construct their



Pipeline Pilot



Pipeline Pilot Professional Client - [New Protocol1\Calculate ADMET Properties*]

File Edit View Tools Window Help

75

Protocols Favorites

Search Demo (MDL Direct)
Data Modeling
Analysis
Learning and Clustering
Decision Trees
Drug Like
R Statistics
Reporting

1000 1000 674 326 674 494 180 180 180 180 180 180

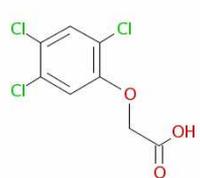
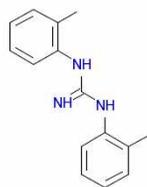
NCI Drugs Reader ADMET Aqueous Solubility Keep only molecules with goo... ADMET Blood Brain Barrier Keep only molecules with goo... ADMET Intestinal Absorption Keep only molecules with goo... Remove Hydrogens 2D Coords HTML Molecular Table Vi...

Molecules outside the confidence area of the BBB model are sent to the Fail port. The BBB value for these molecules could not be calculated.

Display molecules with good Solubility, Absorption and BBB.

Data Record Tree View
Data record # 1
Molecule
430
Name : 430
NSC : 430
CAS_RN : 93-76-5
ADMET_Solubility : -3.627
ADMET_Solubility_Level : 3
ADMET_Unknown_AlogP98 : 0
ADMET_BBB : 0.118
ADMET_BBB_Level : 1
ADMET_Absorption_Level : 0

Display records 1 to 25 of 180 Update <<First <Previous Next> Last>>

Molecule	Name	NSC	CAS_RN	ADMET_Solubility	ADMET_Solubility_Level	ADMET_Unknown_AlogP98	ADMET_BBB	ADMET_BBB_Level	ADMET_Absorption_Level
	430	430	93-76-5	-3.6270	3	0	0.11800	1	
	473	473	97-39-2	-3.8660	3	0	0.17200	1	0

Finished
YE75924 Protocols Components
New Protocol1*

Help

Reads a file of 10,000 known drugs collected from the NCI data
An SD file reader of a file of known drugs collected from the NCI compo
See Also:
■ This component is a specific instance of the [SD Reader](#).

Additional Options

Why use a workflow?



Facilitate project support:

- rapid development of workflows for one-time usage

Synergies between developers

Less redundancies between developers because of single, central location of components/nodes

Cleaner software architecture:

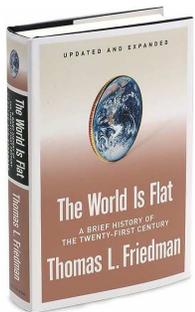
- separation between node (algorithm) development and workflows
- workflows can quickly be changed with little if any programming
- easier new tool development and rapid prototyping

Cross discipline workflows (e.g., cheminformatics, data mining, statistics, imaging, bioinformatics, etc.)

Facilitate scientific software development by separating algorithm development from user interfaces

- develop algorithmic workflows => web services
- write tailored GUI in, say, C# that uses the web services for the computational parts

A Common Collaborative Framework



'The world is flat'

What's KNIME?



KNIME - Konstanz Information Miner - Microsoft Internet Explorer provided by Eli Lilly and Company

Address: http://www.knime.org/index.html

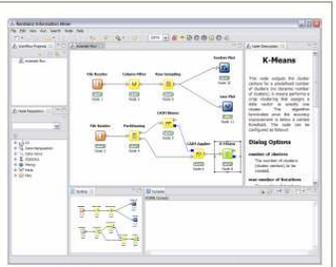
KNIME
Konstanz Information Miner

home download documentation community

news screenshots applications examples service partners about

Welcome!

KNIME, pronounced [naim], is a modular data exploration platform that enables the user to visually create data flows (often referred to as pipelines), selectively execute some or all analysis steps, and later investigate the results through interactive views on data and models.



[Download KNIME](#)

KNIME was developed (and will continue to be expanded) by the [Chair for Bioinformatics and Information Mining](#) at the [University of Konstanz](#), Germany. The group headed by Michael Berthold is also using KNIME for teaching and research at the University. Quite a number of new data analysis methods developed at the chair are integrated in KNIME. Let us know if you are looking for something in particular, not all of those modules are part of the standard KNIME release just yet...

KNIME base version already incorporates over 100 processing nodes for data I/O, preprocessing and cleansing, modelling, analysis and data mining as well as various interactive views, such as scatter plots, parallel coordinates and others. It includes all analysis modules of the well known Weka data mining environment (<http://www.cs.waikato.ac.nz/ml/weka>) and additional plugins allow R-scripts (www.r-project.org) to be run, offering access to a vast library of statistical routines.

KNIME is based on the [Eclipse](#) platform and, through its modular API, easily extensible. Custom nodes and types can be integrated within hours enabling KNIME to be used not only in production environments but also for teaching and research prototyping. If you would like to read a more detailed description of the software, please download our White Paper as a PDF file [here](#).

KNIME is available through a dual licensing scheme. A non-profit open source license allows KNIME to be downloaded, distributed, and used freely as long as the software or its use is not distributed per profit. See [license arrangements](#) for details.

Address: http://www.knime.org/downloads.html

KNIME: Download: Extensions - Microsoft Internet Explorer provided by Eli Lilly and Company

KNIME: Forum: Index - Microsoft Internet Explorer provided by Eli Lilly and Company

Address: http://www.knime.org/forum/index.php

KNIME
Konstanz Information Miner

home download documentation community

knime wishlist contributors events

FAO Search Memberlist Usergroups Register
Profile Log in to check your private messages Log in

The time now is Wed May 28, 2008 2:43 pm

KNIME Forum Index View unanswered posts

Forum	Topics	Posts	Last Post
KNIME			
KNIME Users For users of the KNIME Data Analysis/Workbench	90	439	Wed May 28, 2008 7:37 am wiswedel ↗
KNIME Developers Developers extending the functionality of KNIME with their own Nodes	81	256	Wed May 28, 2008 7:25 am thor ↗
KNIME General General Questions regarding KNIME	26	103	Wed May 07, 2008 10:04 am wiswedel ↗
R Statistics Nodes and Integration	6	27	Thu May 08, 2008 12:07 am lobal ↗
Third Party			
Tripos Extensions Extensions implemented and distributed by Tripos	2	10	Mon Aug 20, 2007 5:00 pm dalko ↗
Schrödinger Extensions Extensions implemented and distributed by Schrödinger	0	0	No Posts
JChem Extensions Extensions implemented and distributed by Infocom	3	4	Mon Mar 31, 2008 8:56 am tohshima ↗

All times are GMT + 1 Hour

Who is Online

Our users have posted a total of 839 articles
We have 157 registered users
The newest registered user is [Ender](#)

In total there are 2 users online :: 0 Registered, 0 Hidden and 2 Guests [Administrator] [Moderator]
Most users ever online was 69 on Wed Nov 07, 2007 7:38 pm
Registered Users: None

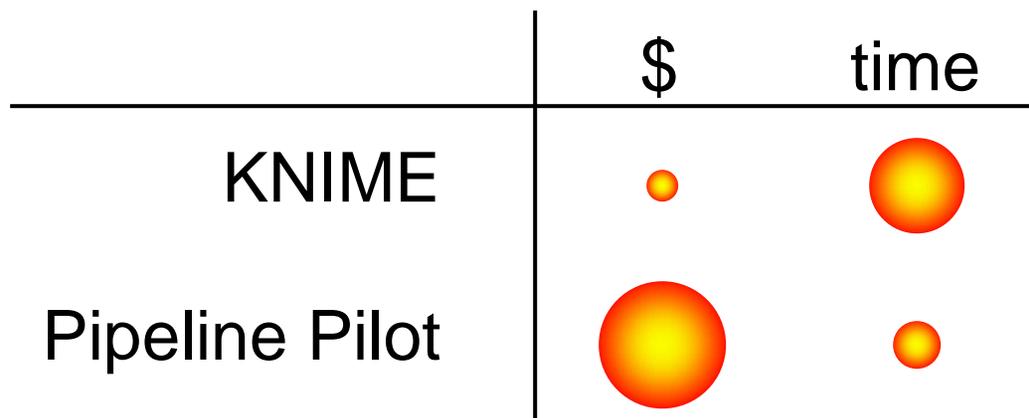
This data is based on users active over the past five minutes.

Log in

Username: Password: Log me on automatically each visit



Pipeline Pilot v KNIME



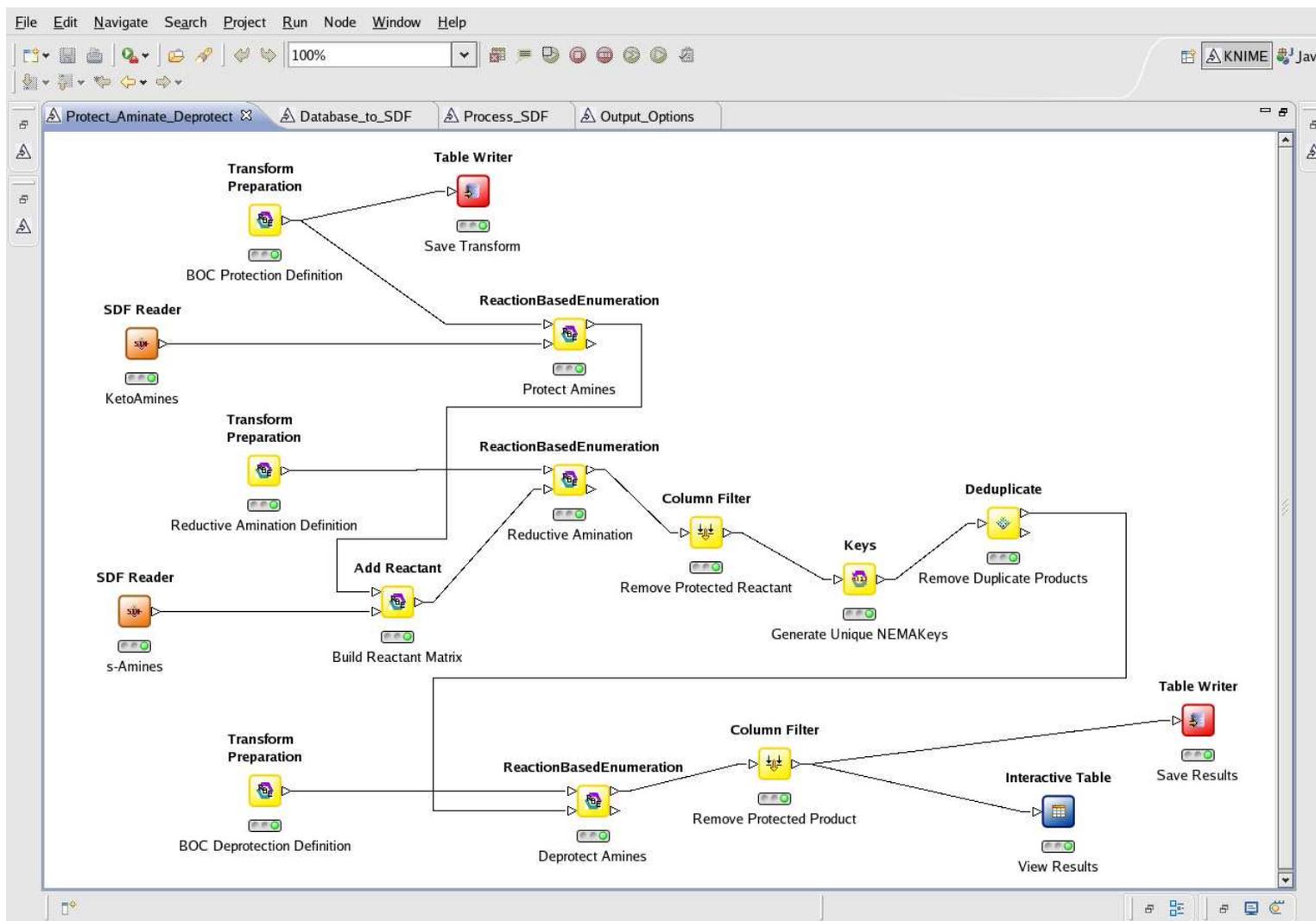
Pipeline Pilot	KNIME
commercial product - expensive	open source - free
mature	immature
industry standard	picking up,
comprehensive compound/node collections	compchem focussed node collections
well scalable with increasing number of developers and users	not well scalable with increasing number of developers and users
client/server architecture	thick client
supports all prominent programming languages	Java; Python
pipelining	workflow
fast	unclear how scalable with large datasets
provides API for internal molecular object	optional CDK molecular object



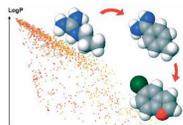
- Alignment Reader
 - Alignment Writer
 - Glide Grid Reader
 - Molecule Reader
 - Molecule Writer
 - Sequence Reader
 - Sequence Writer
 - Converters
 - MAE-to-Pdb
 - MAE-to-Sdf
 - MAE-to-Smiles
 - MAE-to-Mol2
 - Molecule-to-MAE
 - PoseViewer-to-Complex
 - String-to-MAE
 - Hartree-to-kcal/mol Conv
 - kJ-to-kcal Converter
 - Ligand Preparation
 - Epik
 - LIGPREP
 - ADME
 - QIKPROP
 - Cheminformatics
 - Fingerprint Generation
 - Generate Pairwise Matrix
 - Similarity Matrix (from M
 - Dissimilarity Selection (fr
 - Hierarchical Clustering (fr
 - Convert Fingerprint to Tal
 - Convert Matrix to Table
 - Convert Table to Matrix
 - Multi-dimensional Scaling
 - Principal Components
 - Build Report for Clusterin
- Protein Structure Prediction
 - BLAST
 - PRIME Build Homology Model
 - Protein Structure Alignment
- Docking and Scoring
 - GLIDE Ligand Docking
 - PRIME MM-GBSA
- Molecular Mechanics
 - Conformational Search
 - Ligand Torsion Search
 - MACROMODEL Minimization
 - PRIME Minimization
 - MACROMODEL Single Point Ene
- Quantum Mechanics
 - JAGUAR Minimization
 - JAGUAR Single Point Energy
- Filtering
 - MAE Parser
 - MAE Parser (Ligparse)
 - MAE Properties
 - Property Filter (Propfilter)
- Prototypes
 - Add Hydrogens
 - Align Binding Sites
 - Boltzmann Population
 - CSV Reader
 - Desalter
 - Entropy Calculation
 - Fix Bond Orders
 - Fragment Joiner
 - Fragments from Molecules
 - Glide Ensemble Merge
 - Glide Merge
 - Hypothesis Identification
 - Impref
 - Ionizer
 - Neutralizer
 - Phase DB Creation
 - Phase DB Query
 - Pose Entropy
 - Premin
 - Prime Fix
 - Set Molecule Title
 - Stereoizer
- Add Hydrogens
- Align Binding Sites
- Assign Bond Orders
- Boltzmann Population
- Chemistry External Tool 0:1
- Chemistry External Tool 1:1
- Chemistry External Tool 1:2
- Chemistry External Tool 2:1
- Chemistry External Tool 2:2
- Column Reorder
- Compare Ligands
- Delete Atoms
- Entropy Calculation
- Fragment Joiner
- Fragments from Molecules
- Generate Smarts
- Get PDB
- Group MAE
- Ungroup MAE
- Lookup and Add Columns
- MAE Set Properties
- PDB Name
- Prime Fix
- Protein Assignment
- RMSD
- RRHO Entropy
- Set Molecule Title
- Text Viewer
- Unique Smiles
- Volume Overlap Matrix
- Delete Atoms
- Generate Smarts
- Lookup and Add Columns
- MAE Expander
- MAE Set Properties
- PDB Name
- Protein Assignment
- RMSD
- RRHO Entropy

- Tripos
- AUSPYX
 - Concord (Web service)
 - Confort (Web service)
 - DBTranslate (Web serv
- Chemistry
 - CDK
 - Read
 - MOLZ Reader
 - SD Reader
 - SLN Reader
 - SMILES Reader
 - UNITY DB Reader
 - Write
 - HTML Writer
 - MOLZ Writer
 - PDF Writer
 - SD Writer
 - SLN Writer
 - SMILES Writer
 - Mining
 - Property Calculators
 - 2D Properties
 - ADME/Tox Properties
 - Substructural Properties
 - Translators
 - DBTranslate (Local)
 - Molecule Parser
 - OpenBabel
 - String to Smiles
 - Fingerprints
 - 2D Alignment
 - Molecular Validator
 - SLN Sketcher
- N4K Chemistry
 - RemoteChemistry
 - MolConvertIn
 - MolConvertOut
 - ChemicalFileReader
 - ChemicalFileWriter
 - ChemicalSketcher
 - LigandViewer
 - Chemistry
 - CDK
 - I/O
 - Translators
 - 2D Coordinates
 - 3D Viewer
 - Connectivity
 - Fingerprints
 - Hydrogen Adder
 - Lipinski's Rule-of-Five
 - Structure Sketcher
 - XLogP
 - Molecular Properties
 - IO
 - Read
 - Read M
 - Read S
 - Write
 - Write M
 - Write S
 - Transform
 - Depict
 - Enumer
 - Wash
 - Calculate
 - Fingerp
 - QSAR
 - JChem
 - Converter
 - MolConverter
 - MrvToMol2Converter
 - MrvToPdbConverter
 - MrvToSdfConverter
 - MrvToSmilesConverter
 - MrvToStringConverter
 - IO
 - MarvinSketch
 - MolExporter
 - MolImporter
 - Manipulator
 - ChemicalTerms
 - ElementalAnalyser
 - Fragmenter
 - LibraryMCS
 - MolSearch
 - RGroupDecomposition
 - Standardizer
 - Visualizer
 - MarvinSpace
 - MarvinTable
 - Cheshire
 - Reaction-based Enum
 - Cheshire Script
 - Keys
 - Remove V3k Features
 - Experimental
 - Mol to SDF
 - RGD
 - RGD Scaffold > Query
 - Specific Enumeration
 - Stereo Enumeration
 - IO
 - Direct Database Read
 - Direct Database Read
 - MolfileReader
 - RxnfileReader
 - SDF Reader
 - SDF Writer
 - Utilities

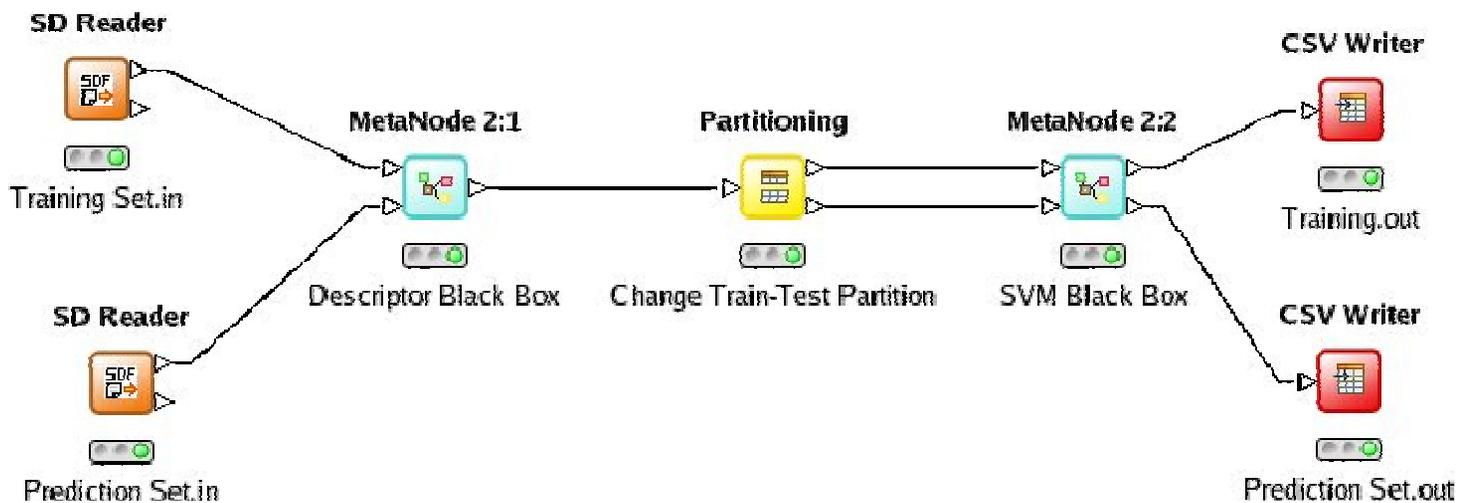
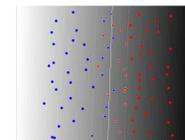
Reductive Amination Workflow*



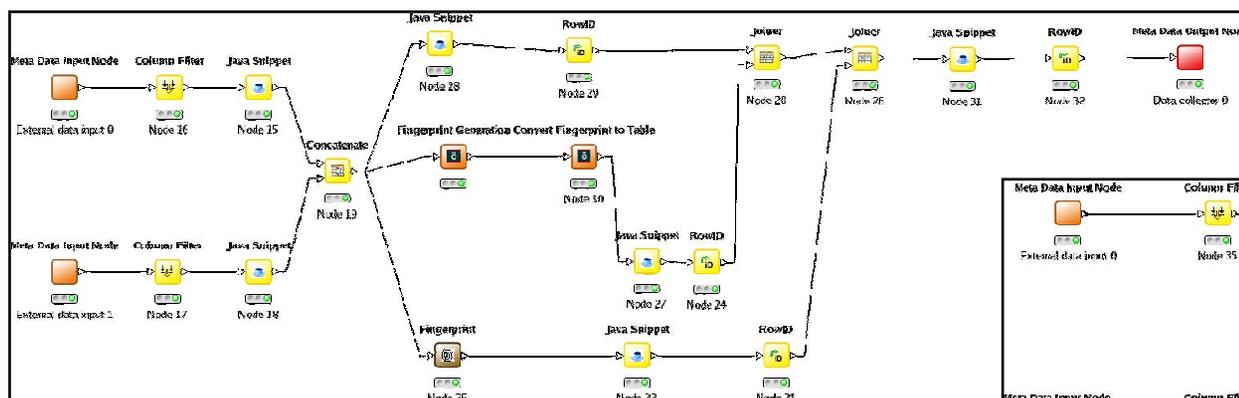
* With kind permission of Keith Taylor, Joe Durant Symyx



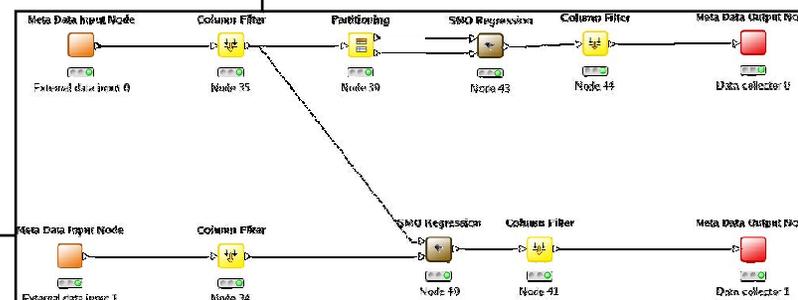
QSAR modelling



Descriptor Calculation *Metanode2:1*



SVM Model Building *Metanode2:2*

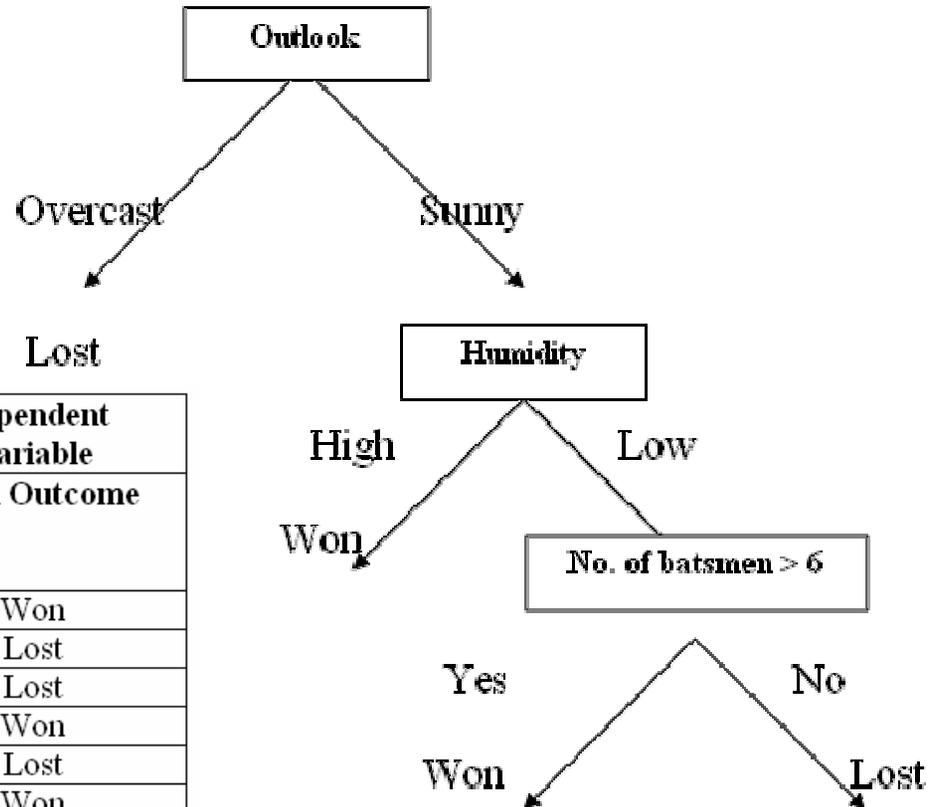


J48 Decision Trees



Lilly Cricket Team
Last 10 matches

Independent Variables			Dependent Variable
Outlook	Humidity	Number of batsmen in team > 6	Final Outcome
Sunny	High	Yes	Won
Overcast	High	No	Lost
Sunny	Low	No	Lost
Sunny	High	No	Won
Overcast	Low	Yes	Lost
Sunny	Low	Yes	Won
Sunny	Low	No	Lost
Sunny	High	No	Won
Sunny	Low	Yes	Won
Sunny	Low	Yes	Won

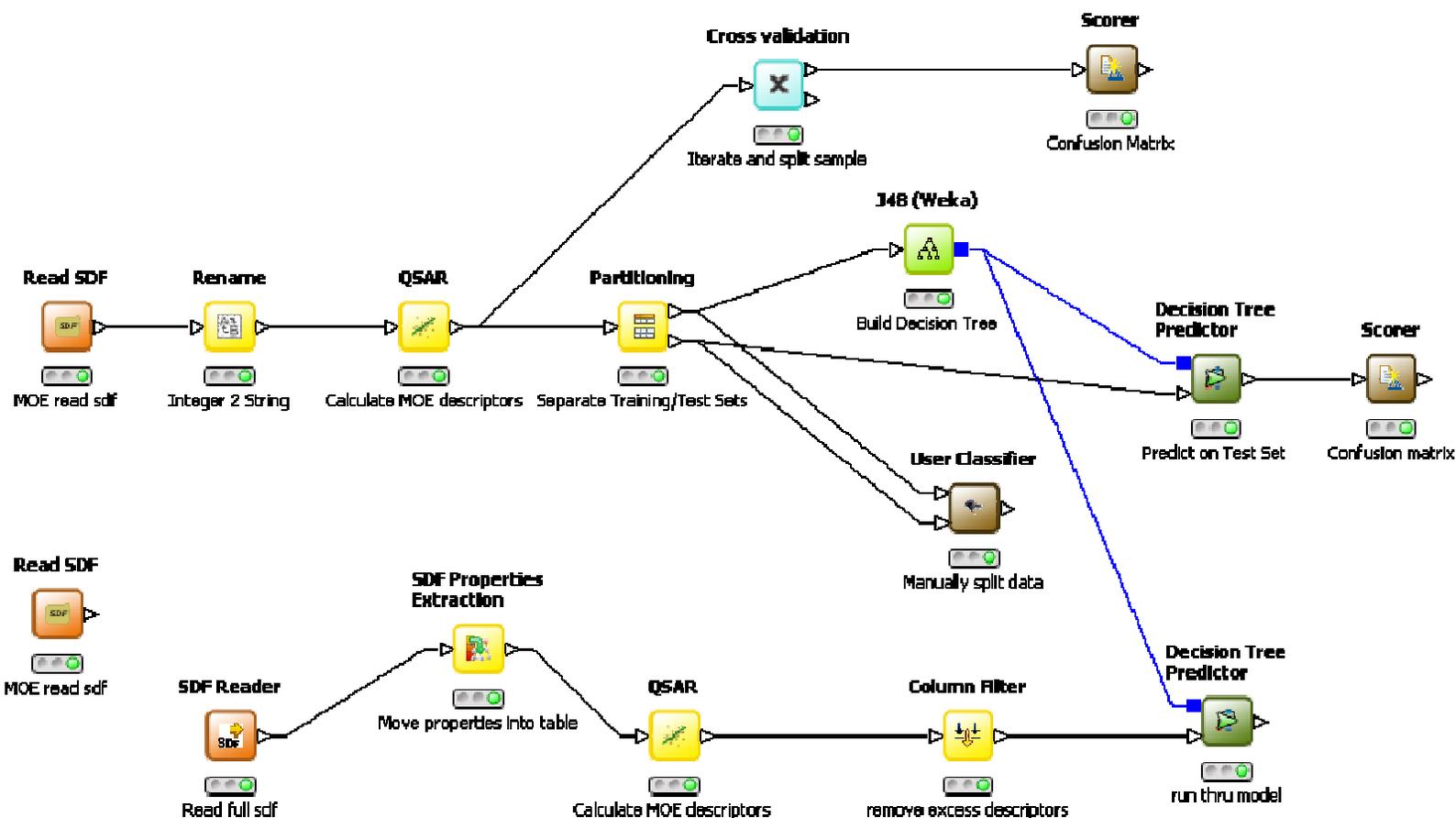


Splitting on highest information gain

<http://www.d.umn.edu/~padhy005/Chapter5.html>

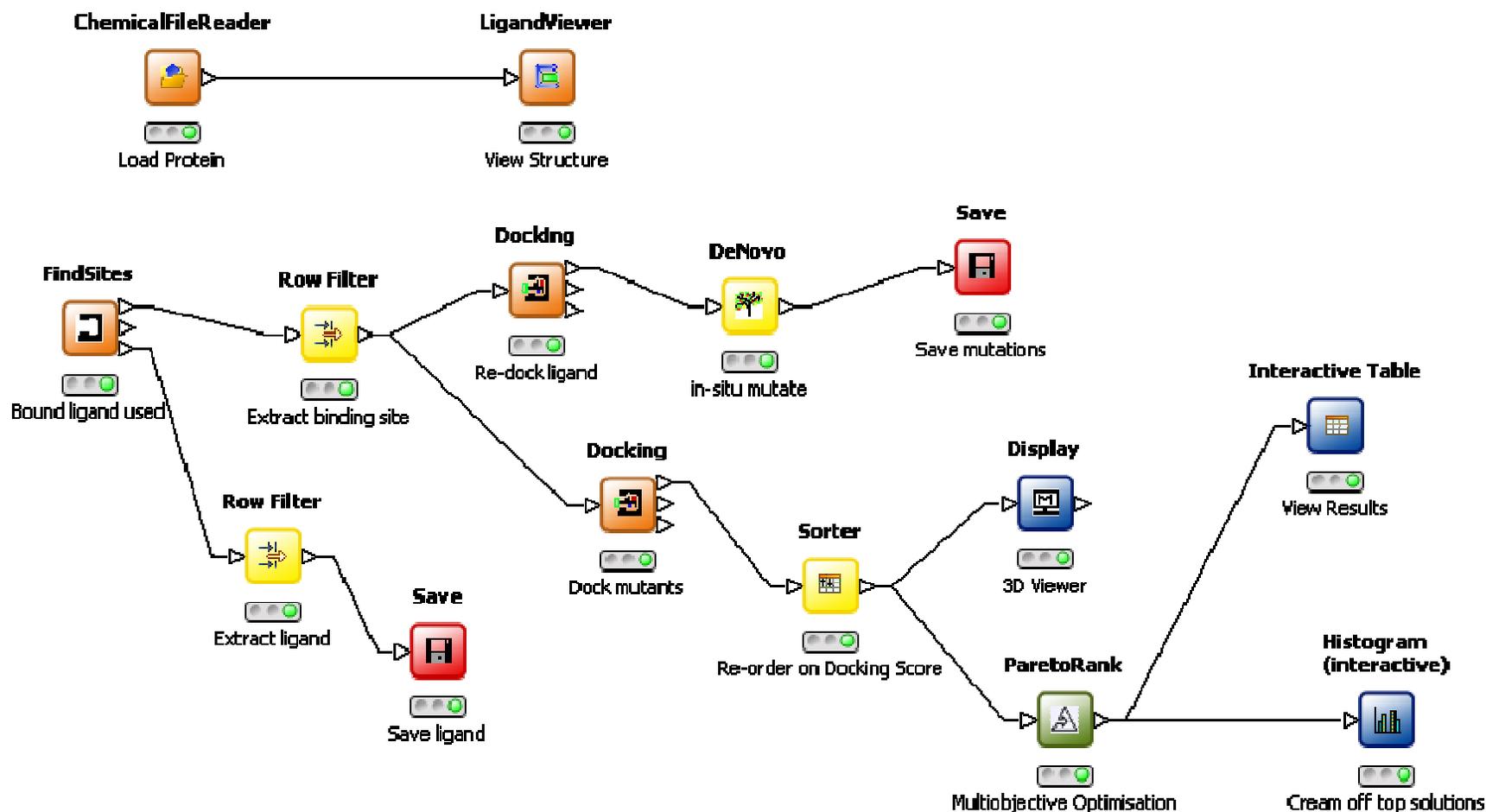


Blood Brain Barrier Penetration



Data taken from J Chem Inf Model 2007, 47,170-175

Structure Based De Novo Design



<http://www.treweren.com/>

Reaction Vectors



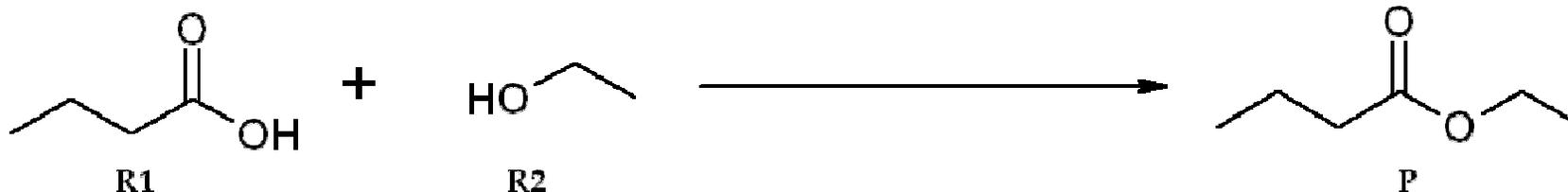
500 Chemists – 5 reactions per wk – 125000 Reactions per yr

Electronic LabNotebooks allows us to store that data

But what are we doing with that data?

And what could we do with the data?

Reaction Vectors in De Novo Design



I	1	2	3	4
Bond	C-C	C-O	C-OH	C-OR
#	4	1	2	0

REACTANT VECTOR
(R1 + R2)

I	1	2	3	4
Bond	C-C	C-O	C-OH	C-OR
#	4	1	0	2

PRODUCT VECTOR (P)

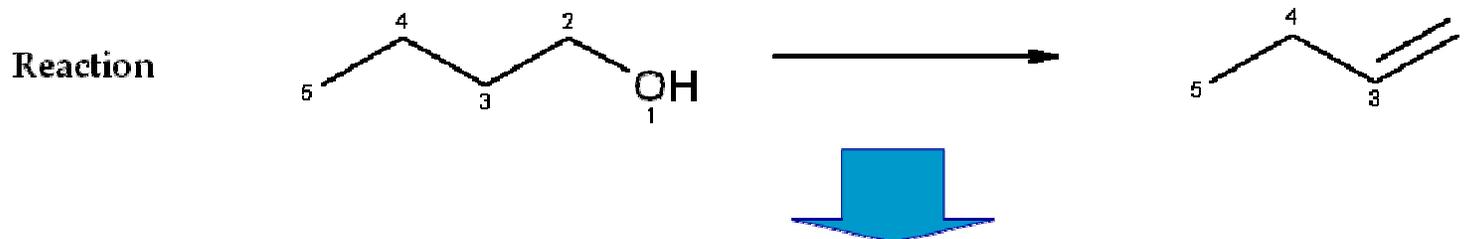
I	1	2	3	4
Bond	C-C	C-O	C-OH	C-OR
#	0	0	-2	2

REACTION VECTOR (D)

$$D = P - (R1 + R2)$$

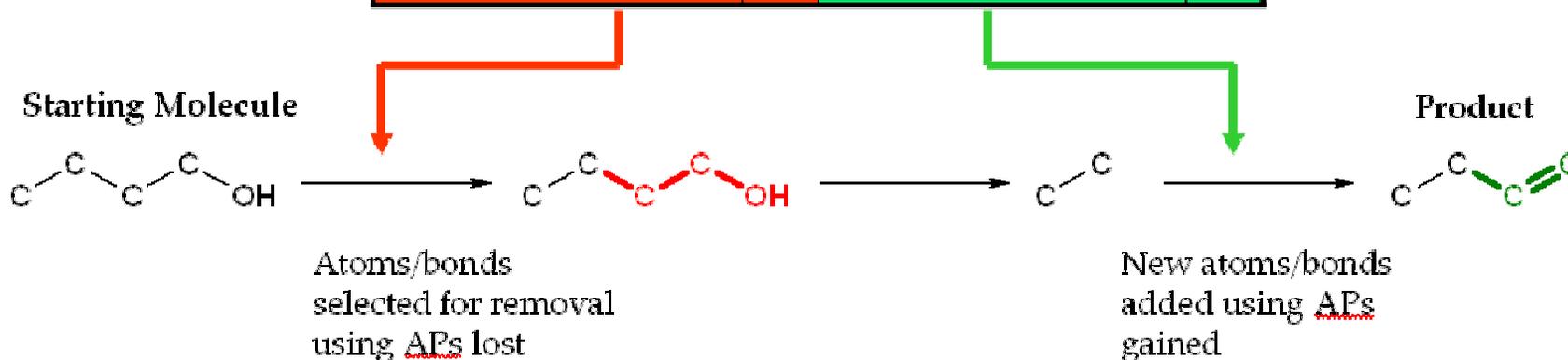
Therefore $P = D + R$ or $P = D + (R1 + R2)$

Applying Reaction Vectors

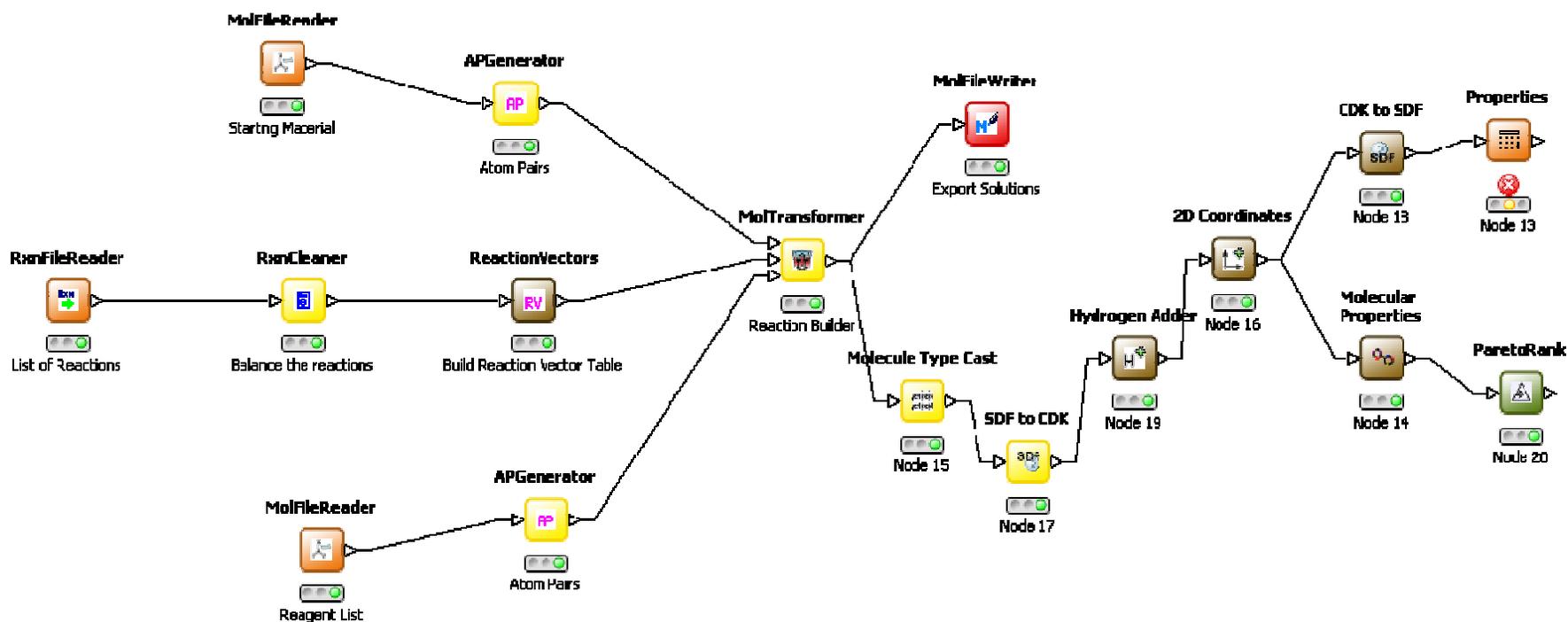


Reaction Vector

APs Lost		APs Gained	
C(2,0,0)-2(1)-O(1,0,0)	-1	C(2,1,0)-2(1)-C(2,0,0)	+1
C(2,0,0)-2(1)-C(2,0,0)	-2	C(2,1,0)-2(2)-C(1,1,0)	+1



Knowledge Based De Novo Design

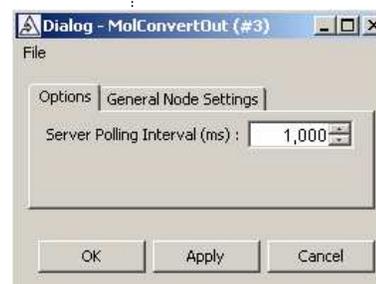
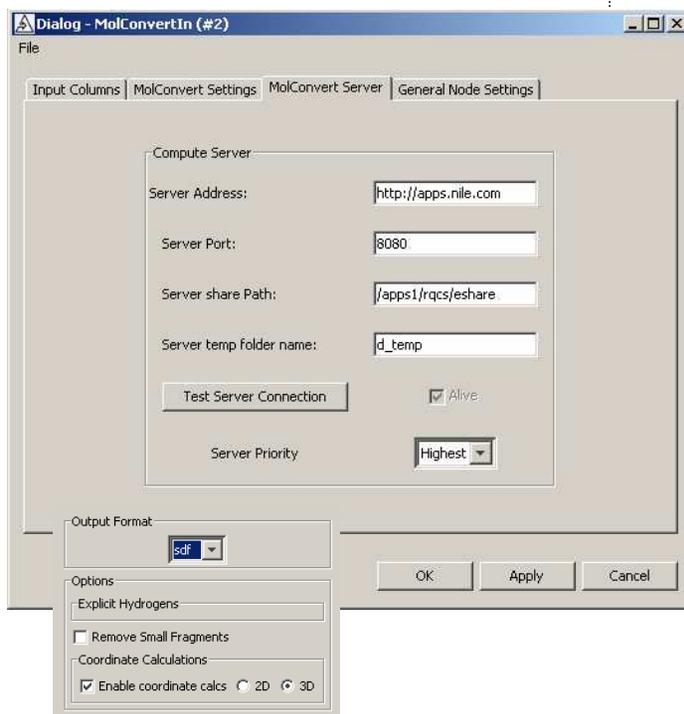
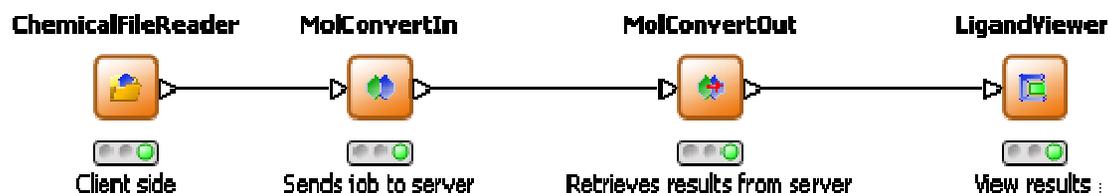




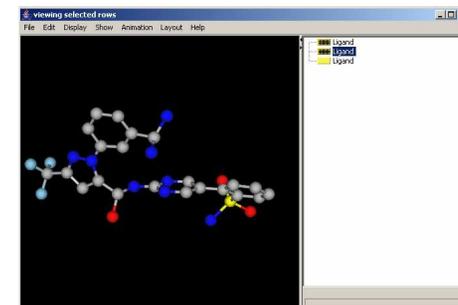
Nodes4knime: Web Services



Simple example: 2d \leftrightarrow 3D Conversion



Row#	2D String	3D String
1		<chem>C1=CC=CC=C1</chem>
2		<chem>C1=CC=CC=C1</chem>
3		<chem>C1=CC=CC=C1</chem>
4		<chem>C1=CC=CC=C1</chem>
5		<chem>C1=CC=CC=C1</chem>
6		<chem>C1=CC=CC=C1</chem>
7		<chem>C1=CC=CC=C1</chem>
8		<chem>C1=CC=CC=C1</chem>
9		<chem>C1=CC=CC=C1</chem>
10		<chem>C1=CC=CC=C1</chem>



<http://sourceforge.net/projects/nodes4knime>

Conclusions

Introduction to workflow tools

Application in medicinal chemistry

Community of users

Acknowledgements



Linda Hiron

David Evans

David Thorner

Dirk Tomandl

Hina Patel

Val Gillet

Keith Davies

The KNIME guys



KNIME Node Development

Three options to incorporate new methods or external programs:

External Tool

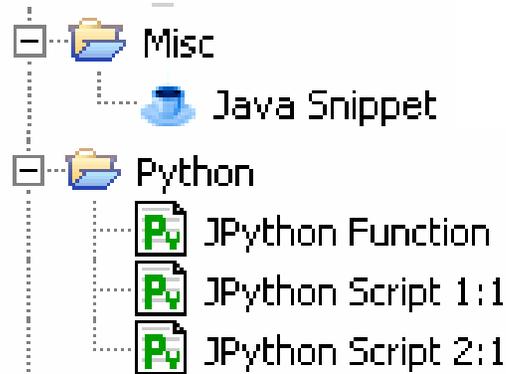
- CSV File → CSV File
- External script with command-line options
- Any language

External Tool



Java or Python Snippet

- Acts on KNIME data table
- Manipulates single columns
- Java/Python required



Full Node

- Acts on KNIME data table
- Can create GUI configuration dialog
- Can implement interactive data views
- Full API
- Call external executable, or write whole procedure within node method
- Java required

Node Architecture

