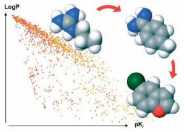


KNIME
Konstanz Information Miner

Data KNIME-ing Discovery: Workflows in Medicinal Chemistry

Mike Bodkin

What a Chemist needs to know about
Chemoinformatics and SAR
Wednesday 11 June 2008
SCI, London, UK

Lilly
Answers That Matter.

Data Pipelining

Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry
ELSEVIER
Computational Biology and Chemistry 31 (2007) 305–319
www.elsevier.com/locate/compbiolchem

Review Article

Workflow based framework for life scie

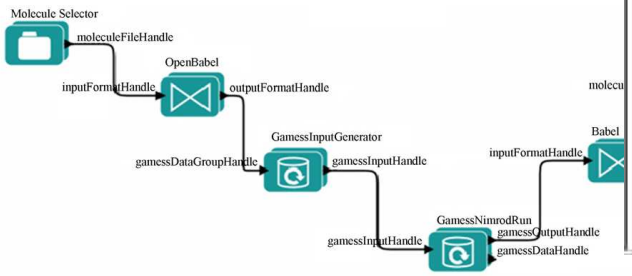
Abhishek Tiwari [a,*], Arvind K.T. Sek
[a] Informatics Division, GVK Biosciences, Hyderabad 5000
[b] School of Informatics, Indiana University (IUPUI), Indianap

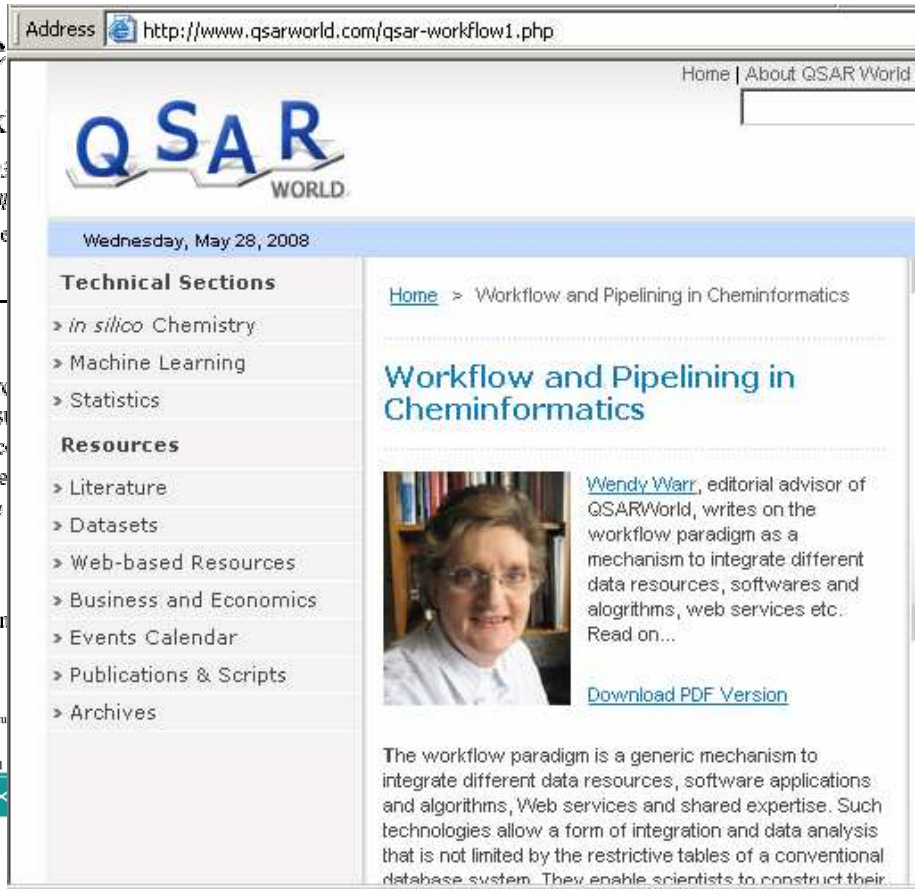
Received 19 February 2007; received in revised form 15 June 2007; acce

Abstract

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

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





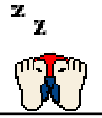
Lilly
Answers That Matter.

Pipeline Pilot

M.Bodkin@lilly.com

Lilly
Answers That Matter.

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

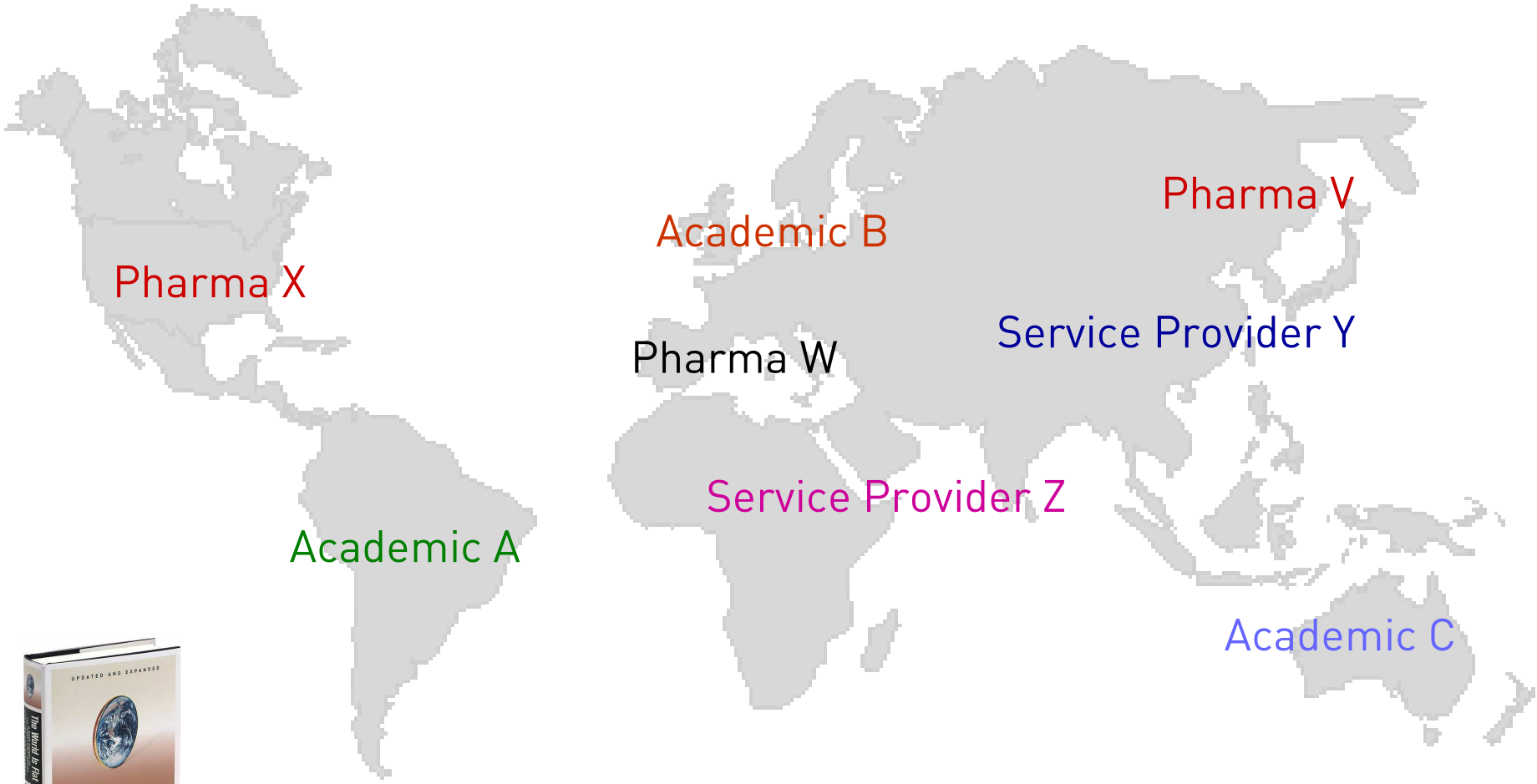
Cross discipline workflows (e.g., cheminformatics, data mining, statistics, imaging, bioinformatics, etc.)

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

M.Bodkin@lilly.com

A Common Collaborative Framework



'The world is flat'

M.Bodkin@lilly.com

Lilly
Answers That Matter.

What's KNIME?

KNIME
Konstanz Information Miner

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

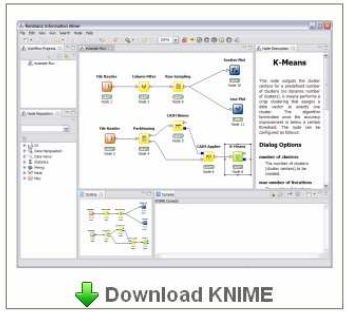
http://www.knime.org/index.html

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.

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. Customs 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.

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

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

KNIME
Konstanz Information Miner

home | download | documentation | community

wishlist contributors events

FAQ 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	61	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 Iqbal
Third Party			
Tripos Extensions Extensions implemented and distributed by Tripos	2	10	Mon Aug 20, 2007 5:00 pm daike
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

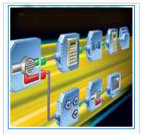
This data is based on users active over the past five minutes

Log in

Username: Password: Log me on automatically each visit Log in

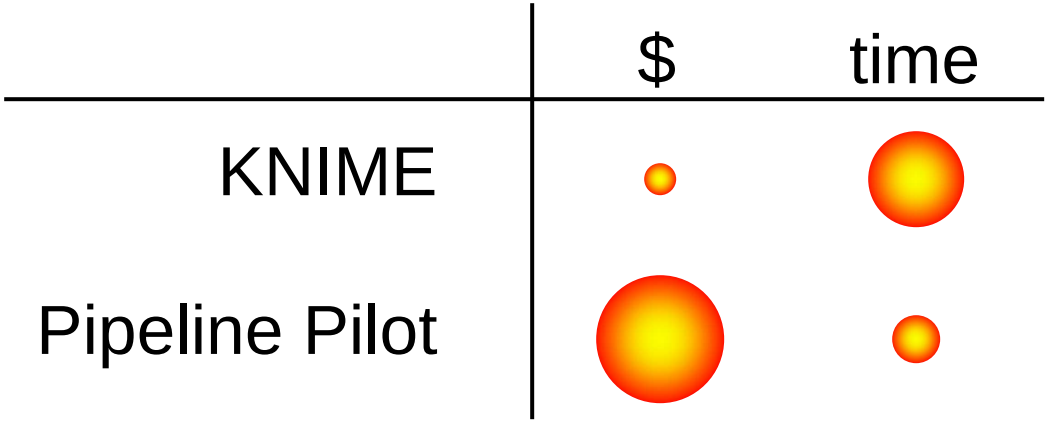
M.Bodkin@lilly.com

Lilly
Answers That Matter.



Pipeline Pilot v KNIME

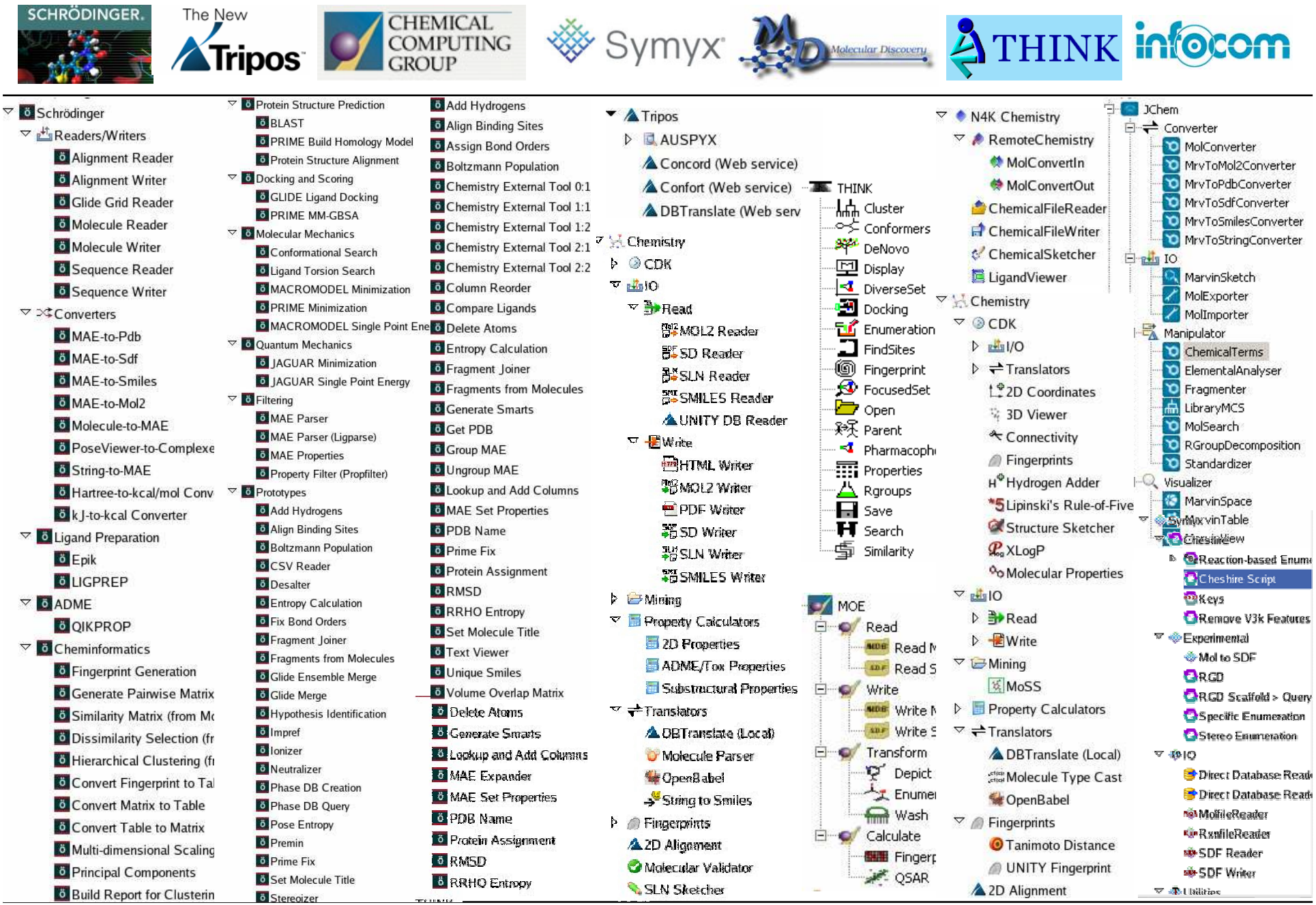
KNIME
Konstanz Information Miner



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

M.Bodkin@lilly.com

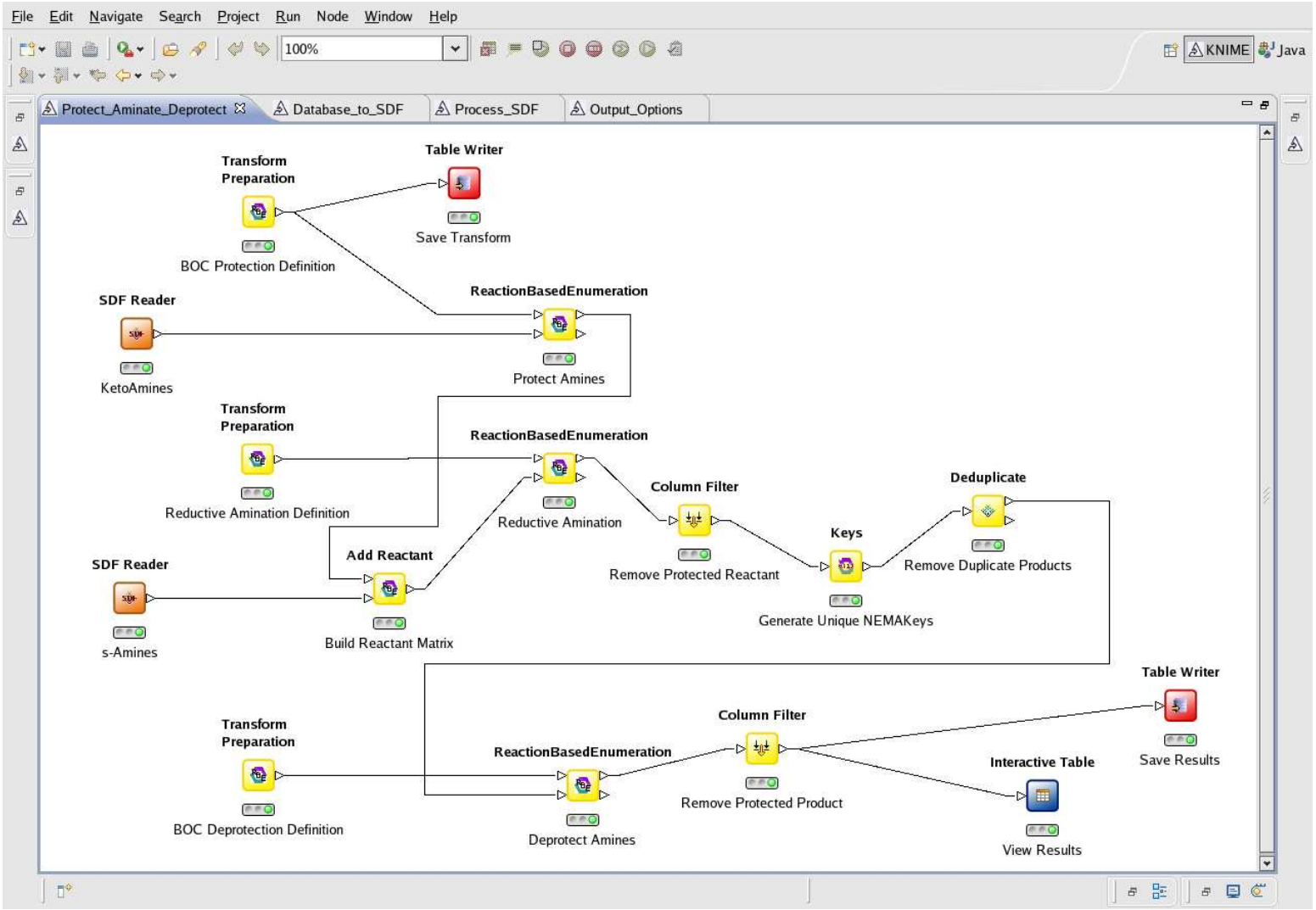
Lilly
Answers That Matter.



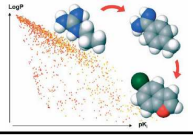
M.Bodkin@lilly.com

Answers That Matter.

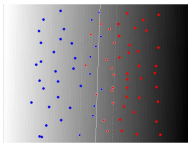
Reductive Amination Workflow*

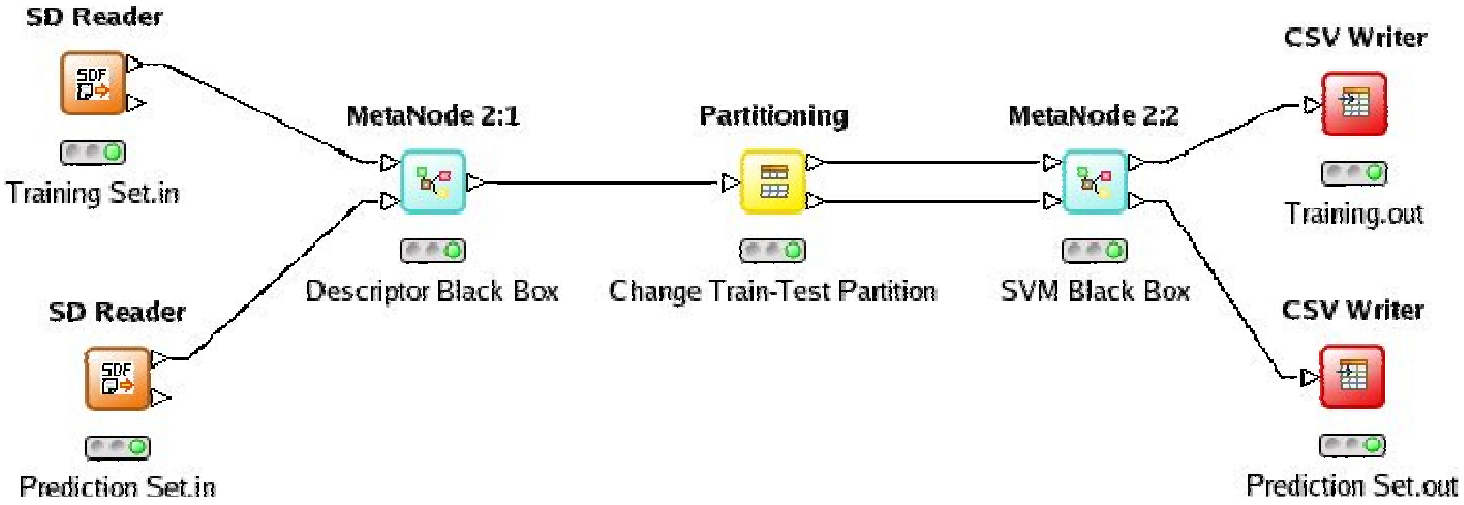


* With kind permission of Keith Taylor, Joe Durant Symyx

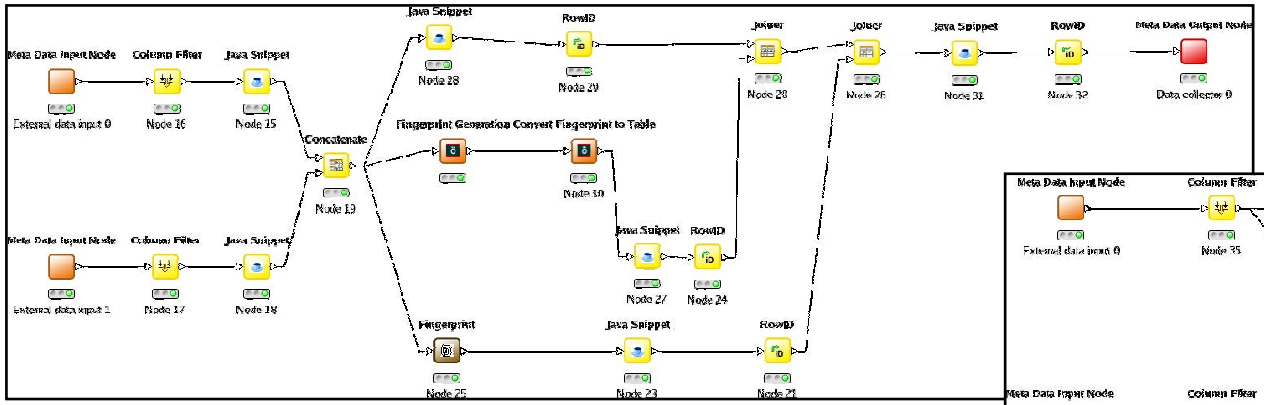


QSAR modelling

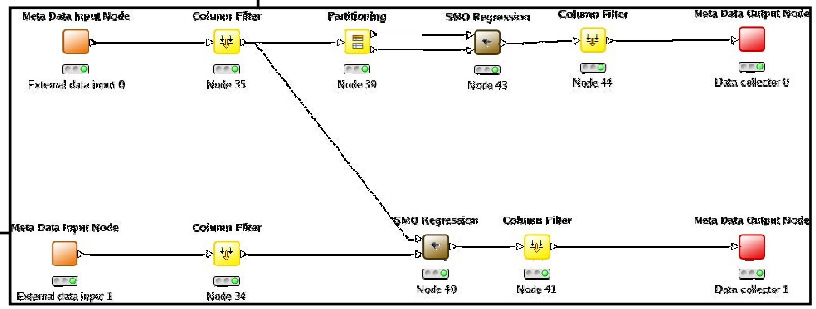




Descriptor Calculation Metanode2:1



SVM Model Building Metanode2:2



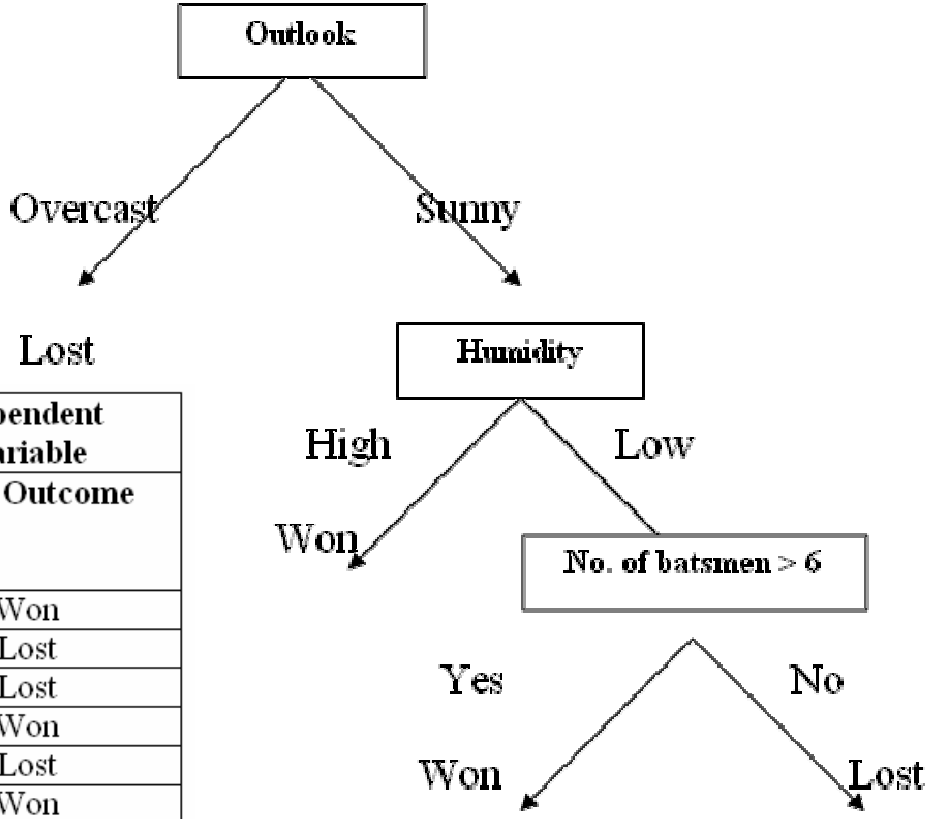
M.Bodkin@lilly.com

Lilly
Answers That Matter.

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

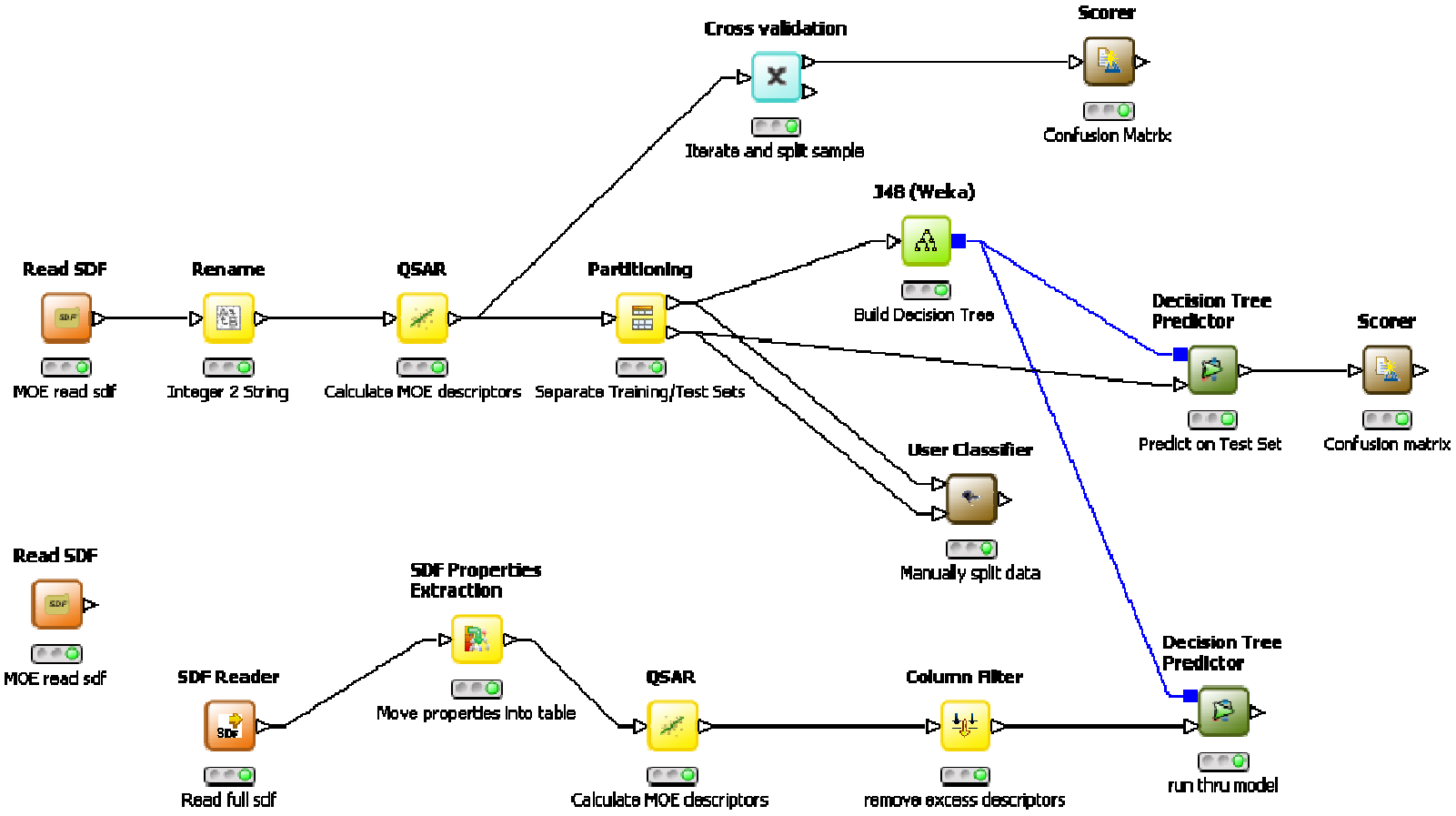


Splitting on highest information gain

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

M.Bodkin@lilly.com

Blood Brain Barrier Penetration

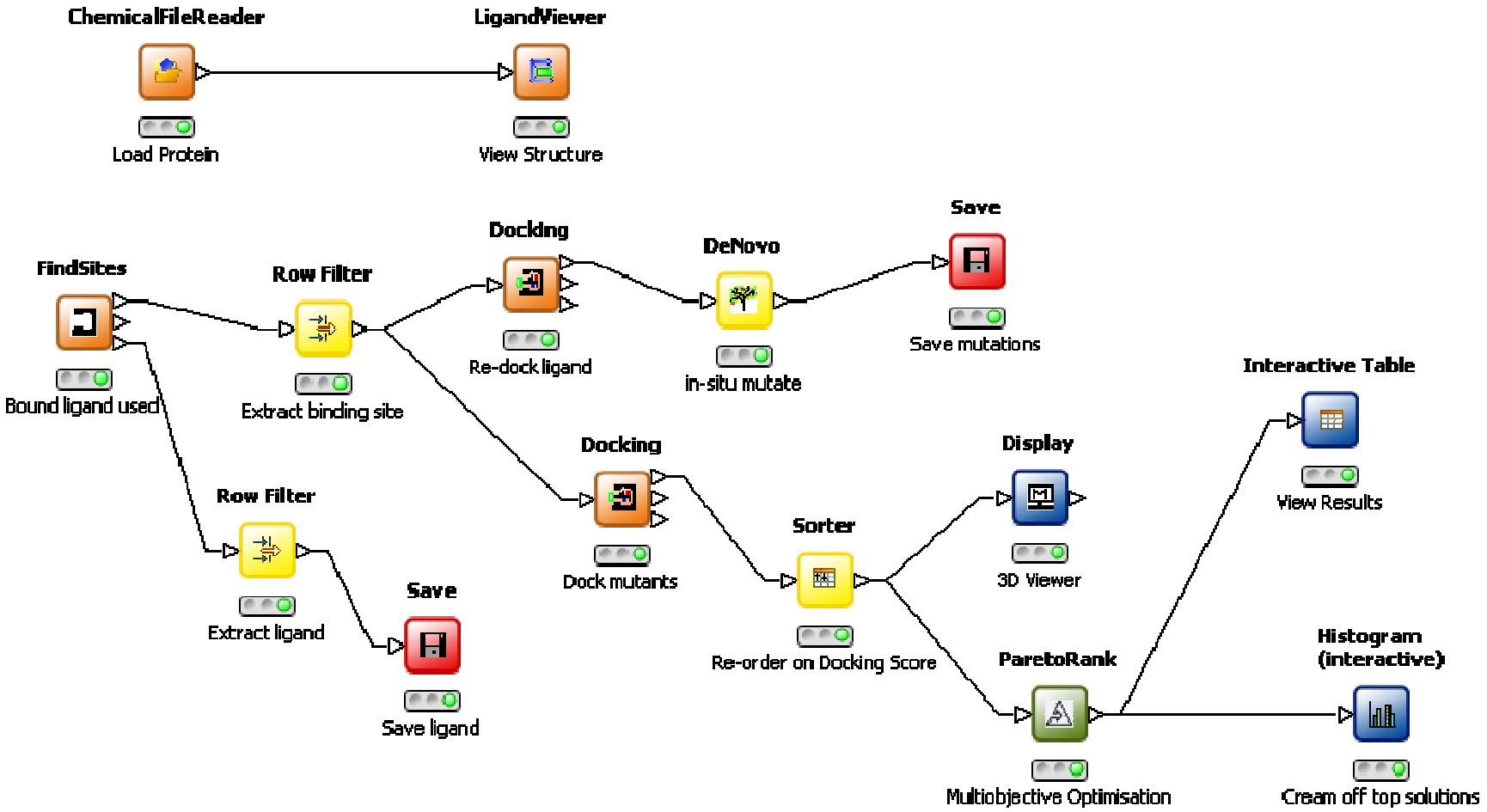


Data taken from J Chem Inf Model 2007, 47,170-175

M.Bodkin@lilly.com

Lilly
Answers That Matter.

Structure Based De Novo Design



http://www.treweren.com/

M.Bodkin@lilly.com

Answers That Matter.

Reaction Vectors

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

Electronic LabNotebooks allows us to store that data

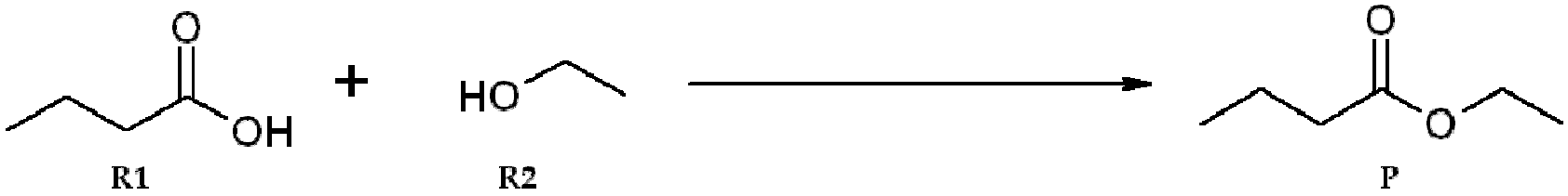
But what are we doing with that data?

And what could we do with the data?

M.Bodkin@lilly.com

Lilly
Answers That Matter.

Reaction Vectors in De Novo Design



R1 R2 P

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)

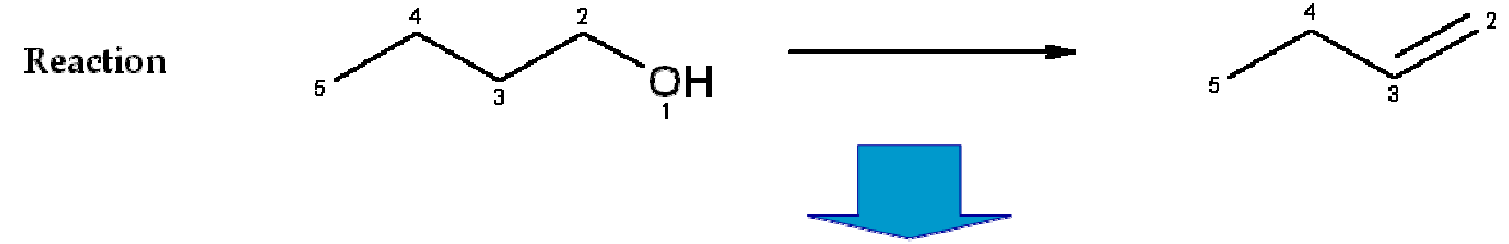
Therefore P = D + R or P= D + (R1 + R2)

M.Bodkin@lilly.com

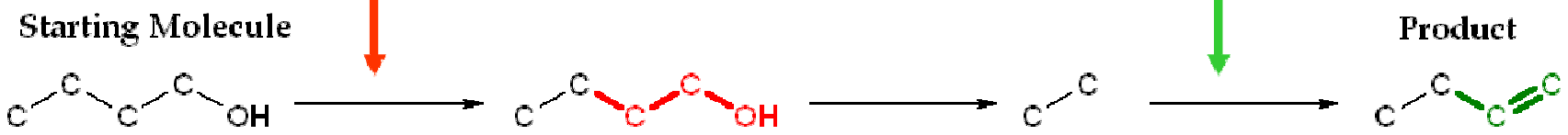
Answers That Matter.

Applying Reaction Vectors

The University Of Sheffield.



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



Atoms/bonds selected for removal using APs lost

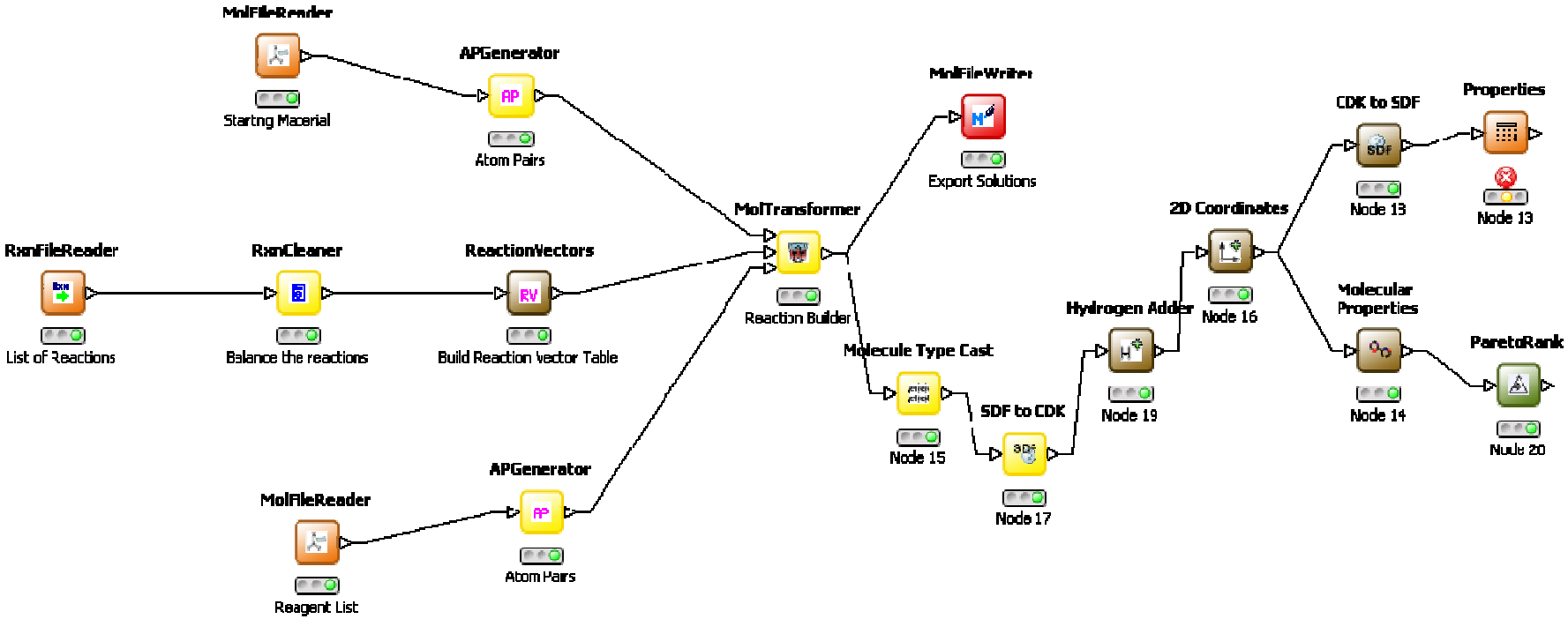
New atoms/bonds added using APs gained

M.Bodkin@lilly.com

Lilly Answers That Matter.

Knowledge Based De Novo Design

The University Of Sheffield.



M.Bodkin@lilly.com

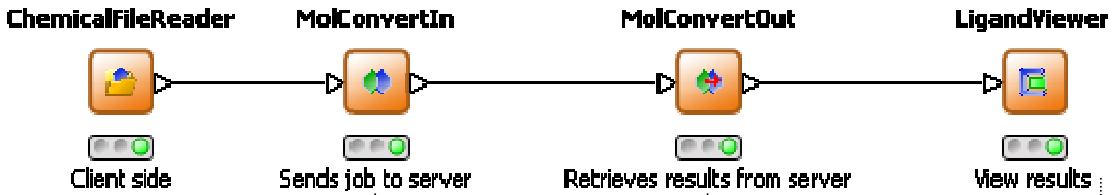
Lilly
Answers That Matter.

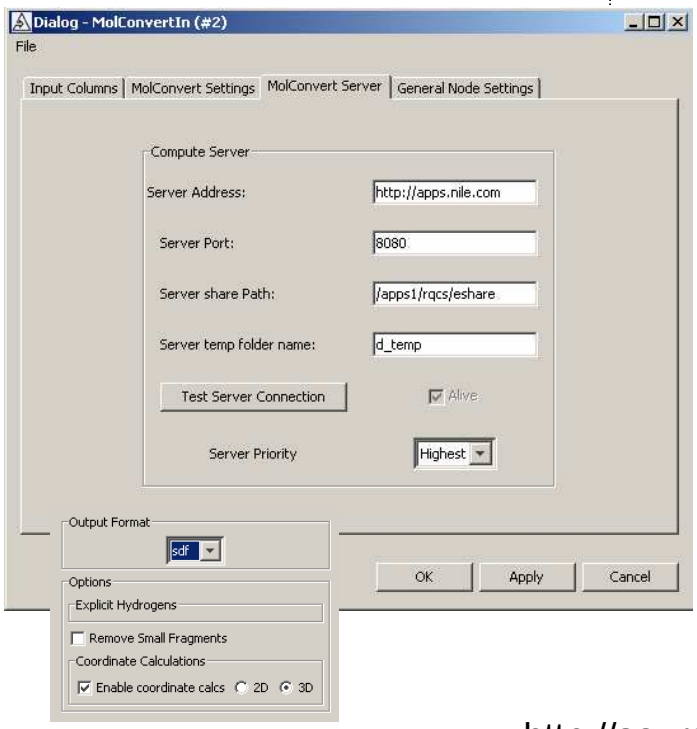
M.Bodkin@lilly.com

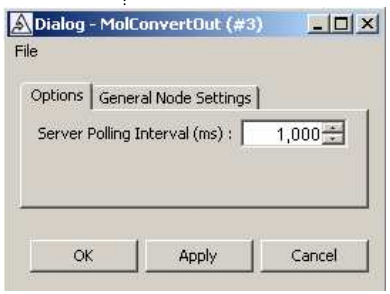
Answers That Matter.

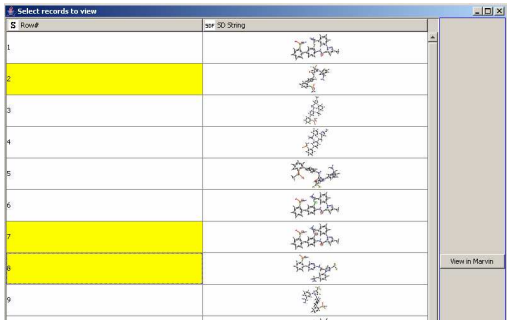
Nodes4knime: Web Services

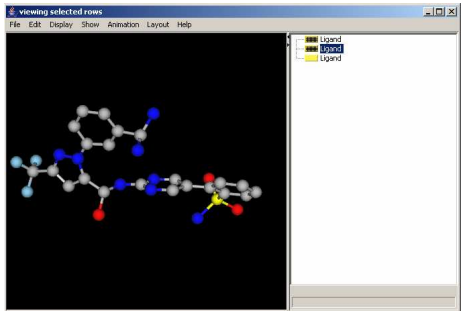
Simple example: 2d←→3D Conversion











http://sourceforge.net/projects/nodes4knime

M.Bodkin@lilly.com

Answers That Matter.

Conclusions

Introduction to workflow tools

Application in medicinal chemistry

Community of users

M.Bodkin@lilly.com

Acknowledgements

Linda Hirons

David Evans

David Thorner

Dirk Tomandl

Hina Patel

Val Gillet

Keith Davies

The KNIME guys

M.Bodkin@lilly.com

Answers That Matter.

KNIME Node Development

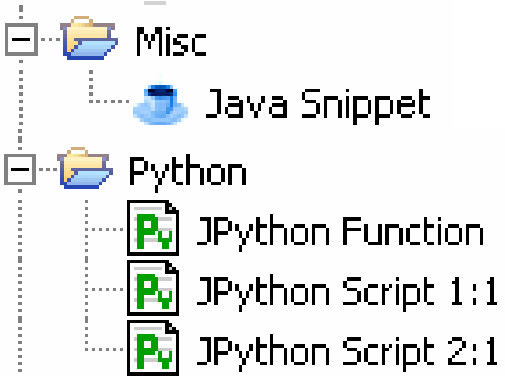
Three options to incorporate new methods or external programs:

External Tool

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

Java or Python Snippet

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

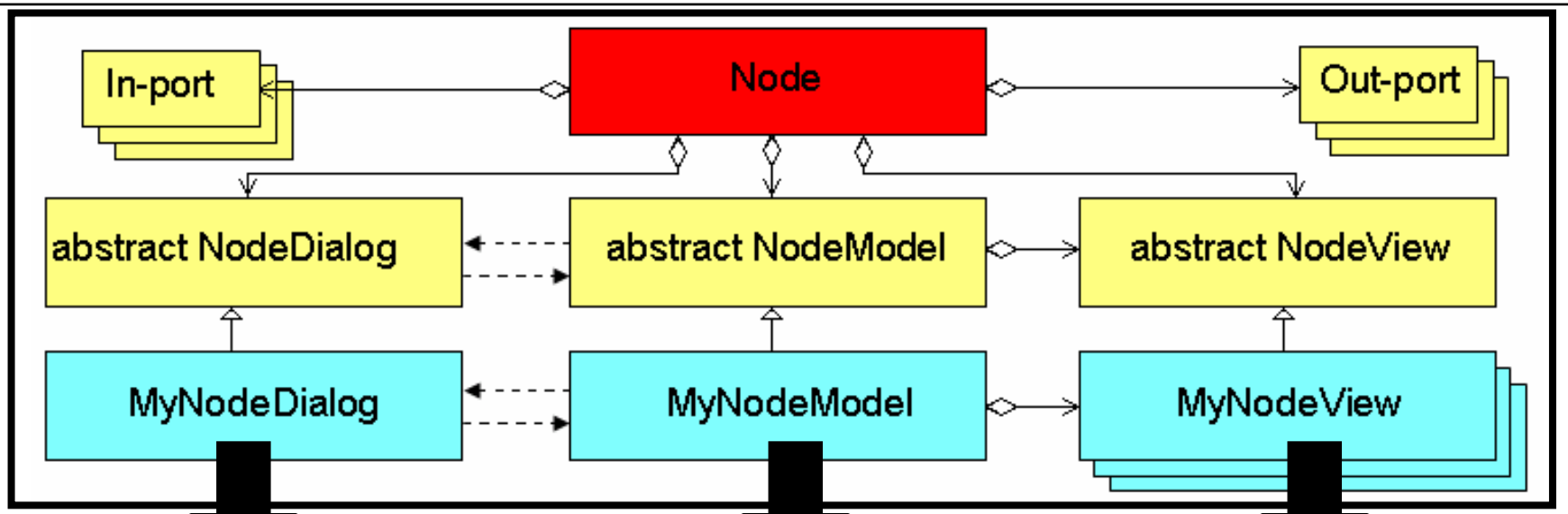


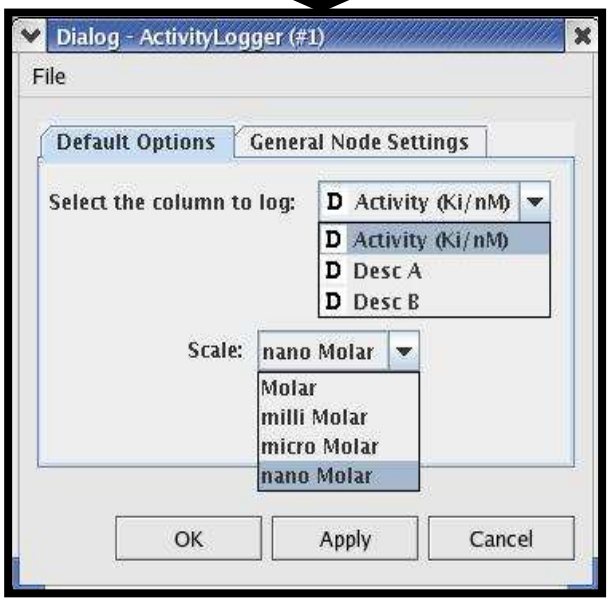
Full Node

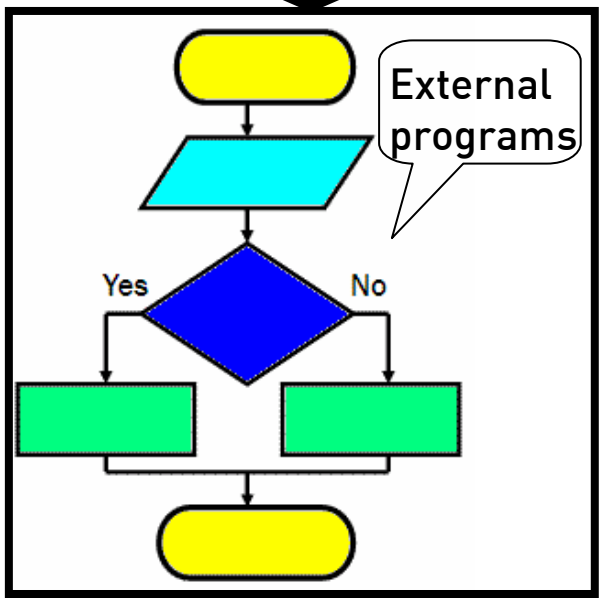
- Acts on KNIME data table
- Can create GUI configuration dialog
- Can implement interactive data views
- Full API
- Call external executable, or write whole procedure within node method
- Java required

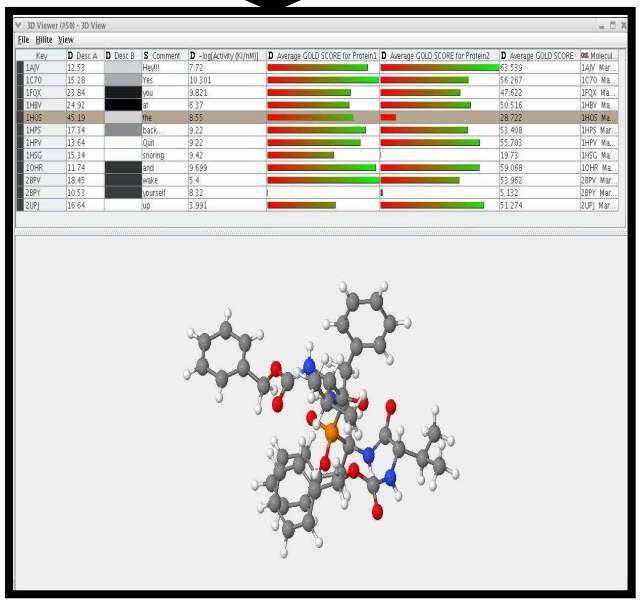
M.Bodkin@lilly.com

Node Architecture









M.Bodkin@lilly.com

Lilly
Answers That Matter.