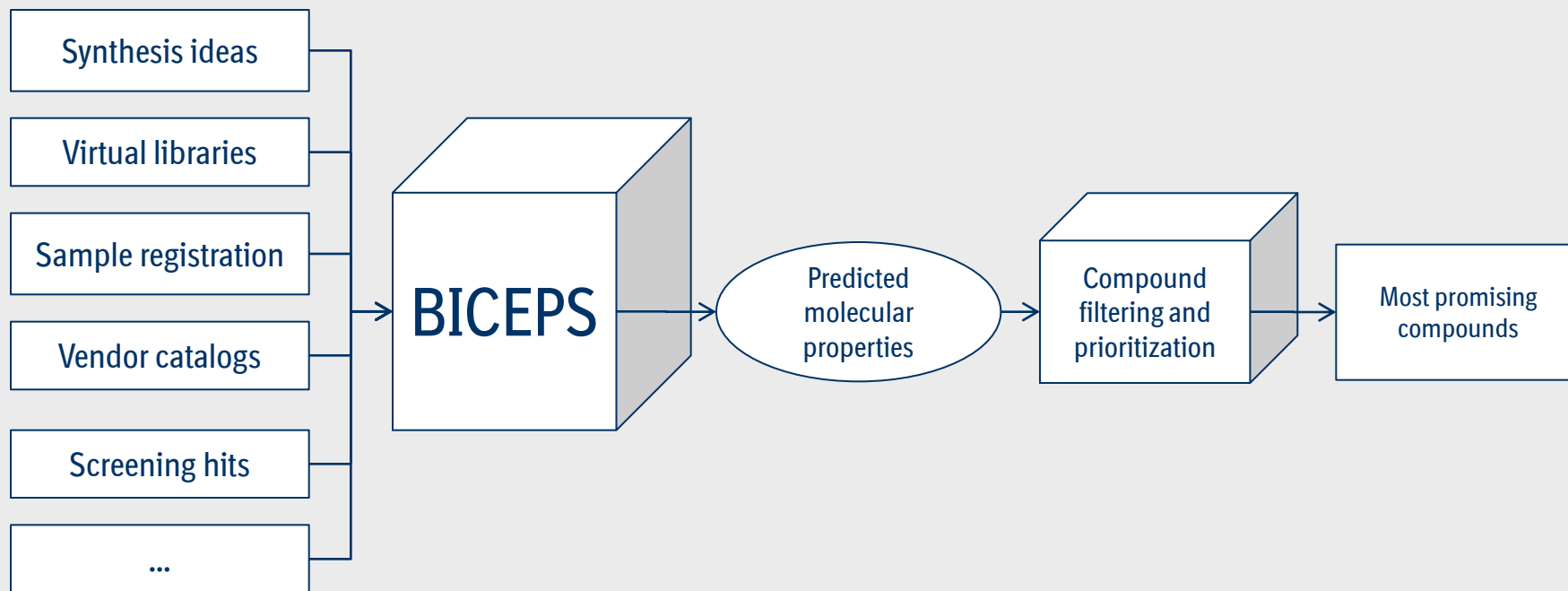


# KNIME as a platform for distributed molecular property predictions

Johannes Koppe, Andreas Teckentrup, [Nils Weskamp](#)



Boehringer  
Ingelheim



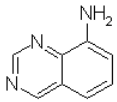
ISIS/Base - [prototyp.db(Read-Only)/main]

File Edit Options Object Database Search List Window Help

Get single Compound Code Get list of Compound Codes Delete list of Compound Codes Help Descriptors Doku

Forms Query Browse Update <MOL> 58 of 67 Search Domain: All

Structure



18989

Sample Identifier	Library ID	Chemist	Class	Subclass
AT-20090114				
MolWeight	Molformula	Purpose	Successor of	
145.16	C <sub>8</sub> H <sub>7</sub> N <sub>3</sub>			
Num. Comment 1	Num. Comment 2	Num. Comment 3	Num. Comment 4	Num. Comment 5
Text Comment 1	Text Comment 2	Text Comment 3	Text Comment 4	Text Comment 5

Main	Cyp
PhysChem	MetStab
Prop	Caco2
Daysi	hERG
	in-vivo
	Solubility

Version: 3.2

Release Date: 13.09.2010 09:18:56

CLogP	CLogP message	MolWeight	# Acceptors	# Donors	TPSA	# Rotatable Bonds
0.76	All fragments measured	145.16	2	1	51.80	0

# Lipinski violations	Veber Message	HIA Message	Andrews binding energy	# pos. N-atoms	Abbott Bioavailability	Net charge @ pH 6
0	BA ok	good	-5.50	0	0.55	0.0

*inhouse ADMET-PC predictions ( confident = 'yes' )*

**Cyp Inhibition**

3A4 [uM]			2D6 [uM]			2C9 [uM]			2C19 [uM]		
Ver.	from	to	Ver.	from	to	Ver.	from	to	Ver.	from	to
V07A			V07A			V06A			V06A		
V07			V07			V06			V06	30.0	50.0

**Caco-2 Permeability**

ab [1E-6 cm/s]			ba [1E-6 cm/s]			Efflux			Intr. [1E-6 cm/s]		
Ver.	from	to	Ver.	from	to	Ver.	from	to	Ver.	from	to
V10	40.0	80.0	V05	50.0	65.0	V10	0.0	1.5	V05	45.0	60.0

**Metabolic stability**

HLM [%Qh]			RLM [%Qh]			MLM [%Qh]		
Ver.	from	to	Ver.	from	to	Ver.	from	to
V06			V06			V02	79.2	88.4

**hERG Inh. IC50 [uM]**

Ver.	from	to
V07A	10.0	50.0

**in-vivo**

CL - rat [ml/min/kg]			VSS - rat [l/kg]		
Ver.	from	to	Ver.	from	to
V05	90.0	120.0	V07	2.0	5.0

**Phospholipidosis**

Risk
low

**Solubility**

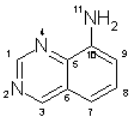
thermodynamic [g/l]							
pH	Ver.	from	to	pH	Ver.	from	to
1.0	V01	0.500	1.000	1.0	V01	0.500	1.000
3.0	V03	0.500	1.000	3.0	V03	0.500	1.000
5.0	V01			5.0	V01		
7.4	V04	0.100	0.500	7.4	V04	0.100	0.500

**nephelometric [mM]**

pH	Ver.	from	to
4.0	V2.0		
7.4	V2.0		

ACDlabs

Structure



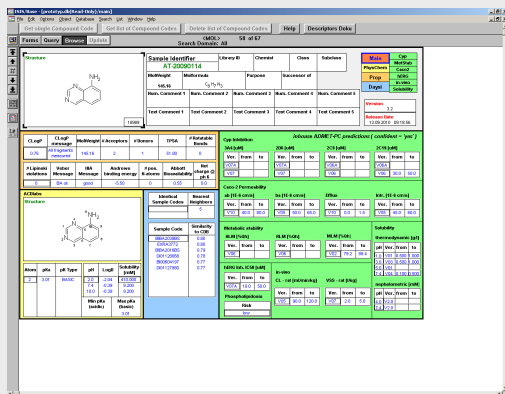
Atom	pKa	pK Type	pH	LogD	Solubility [mM]
2	3.01	BASIC	2.0	-2.04	410.000
			7.4	-0.39	9.200
			10.0	-0.39	9.200

Min pKa (acidic)	Max pKa (basic)
	3.01

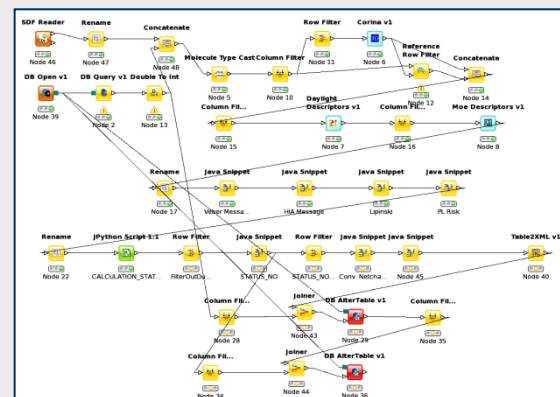
Identical Sample Codes	Nearest Neighbors
	5

Sample Code	Similarity to CDB
BIBA2036BS	0.80
EXRA3772	0.80
BIBA2016BS	0.79
DI01128068	0.78
BI00604197	0.77
DI01127980	0.77

## ISIS/Base Frontend

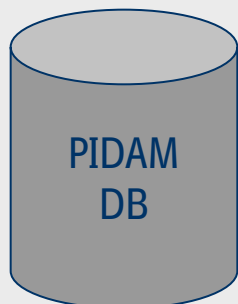


The screenshot shows the ISIS/Base Frontend interface. At the top, there's a search bar and navigation tabs. Below, a chemical structure is displayed on the left. The main area contains several data tables, including a table with columns for 'Mol Weight', 'Mol Weight', 'Mol Weight', etc., and another table with columns for 'Mol Weight', 'Mol Weight', 'Mol Weight', etc.

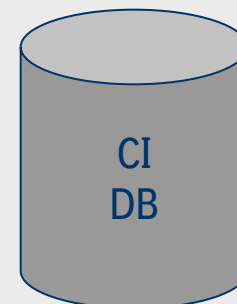


SDF  
(Conductor)

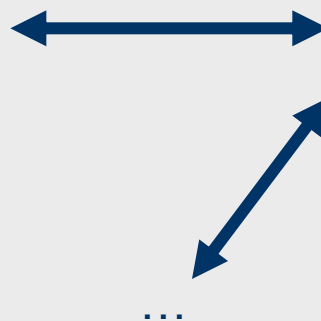
XML



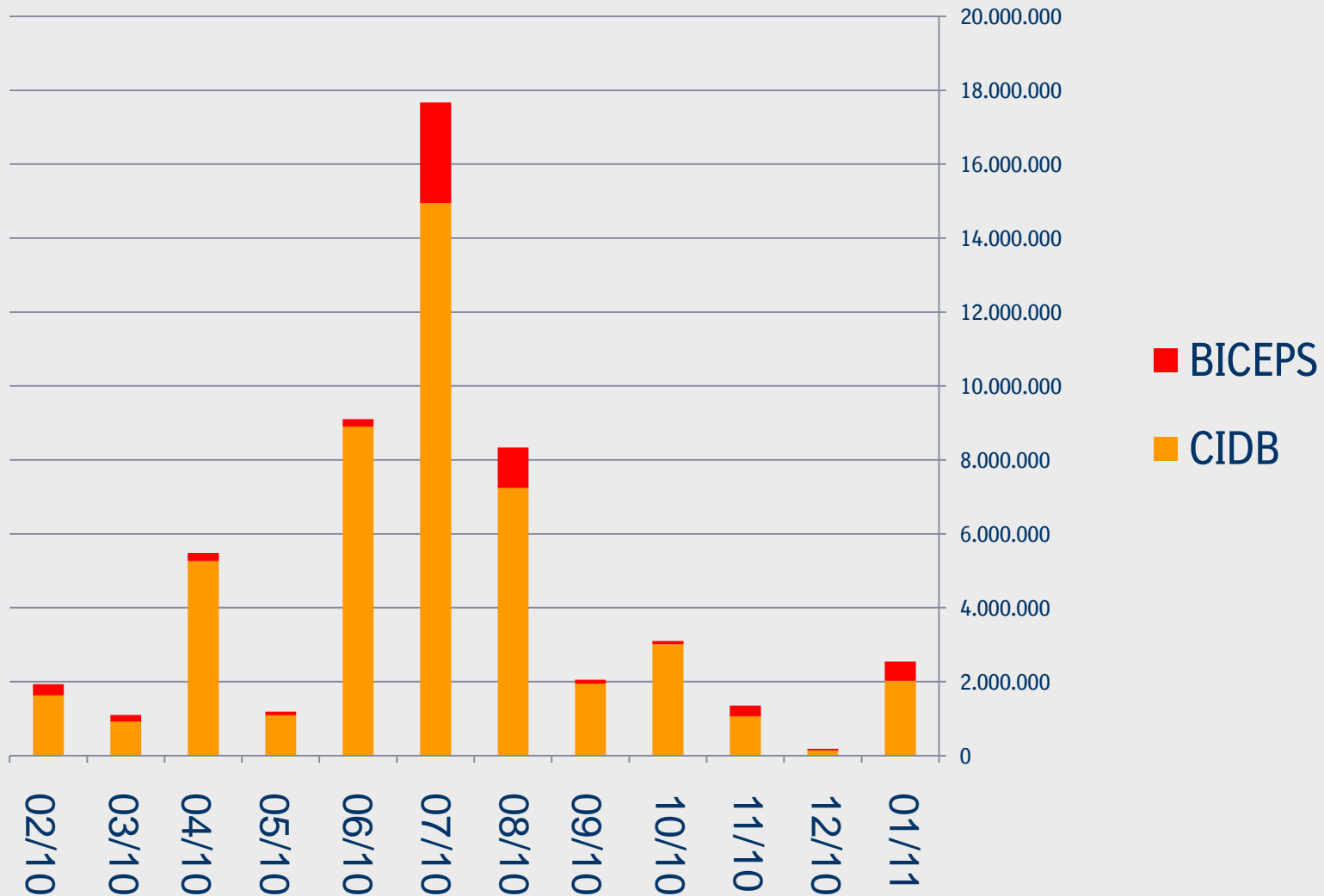
Oracle



Oracle

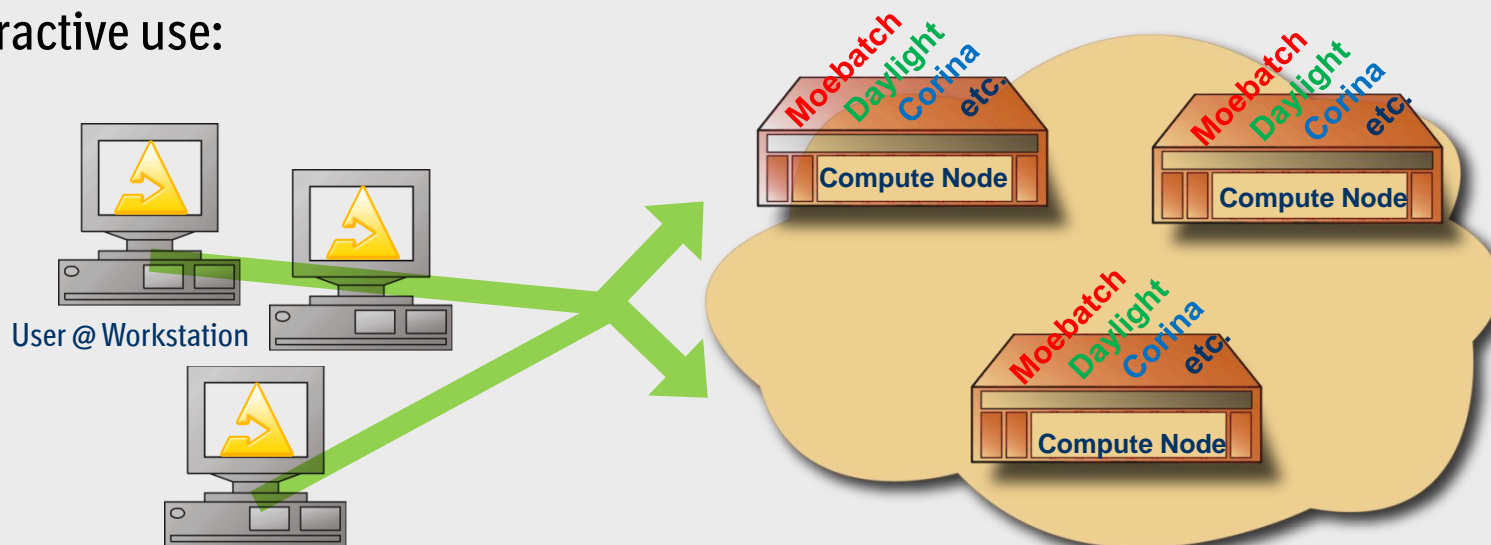


# BICEPS: number of generated data points over 12 months

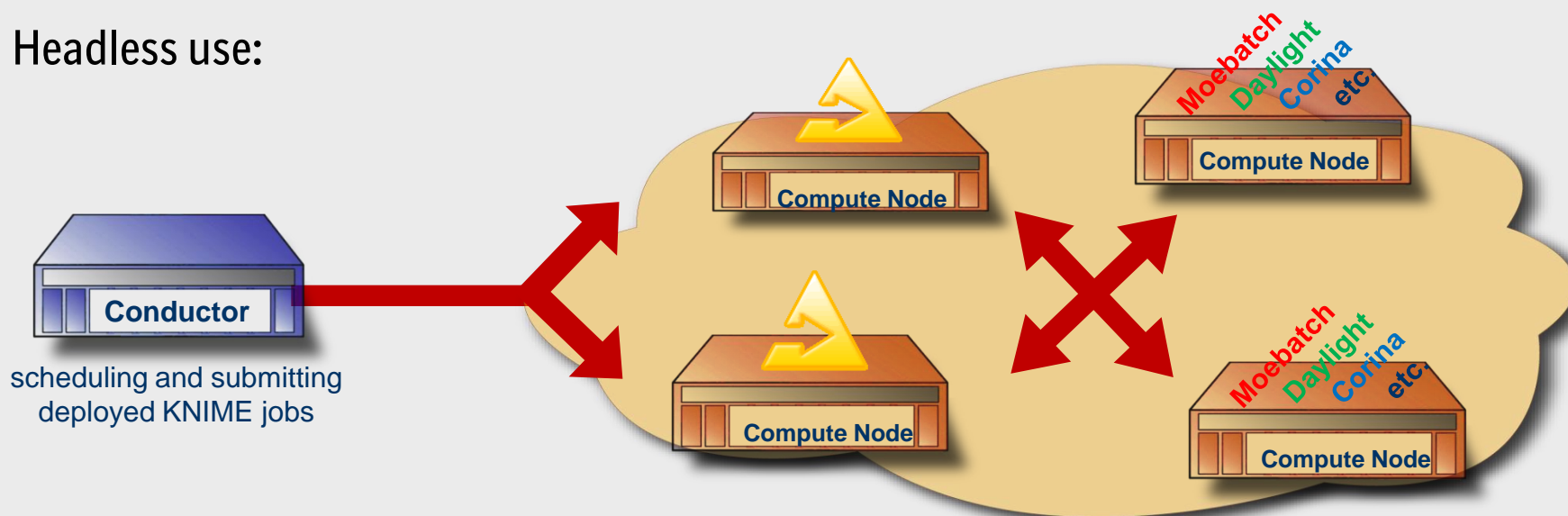


- Altogether, nearly 55 million results generated over the last 12 months
- High variability in monthly system load
  - Peaks in system load often due to model updates and new compound collections
- Issues of scalability, load balancing and reliability under high load conditions
  - Need for manual interventions and significant maintenance efforts
  - Client-Server architecture of many workflow tools leads to bottlenecks and single points of failure
  - Scale-up significantly increases license costs
- Decision to evaluate fully-distributed system architecture based on KNIME

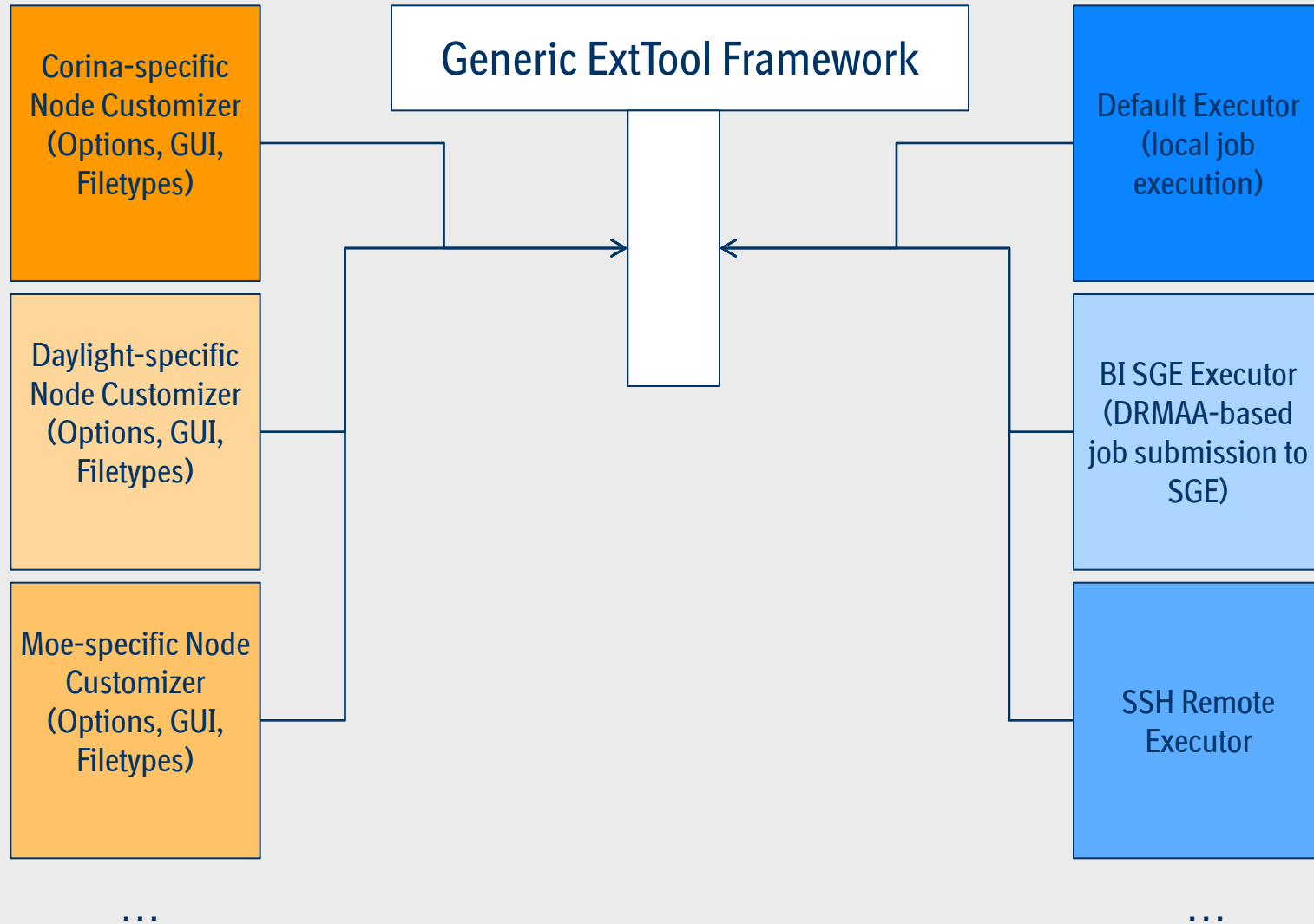
## Interactive use:



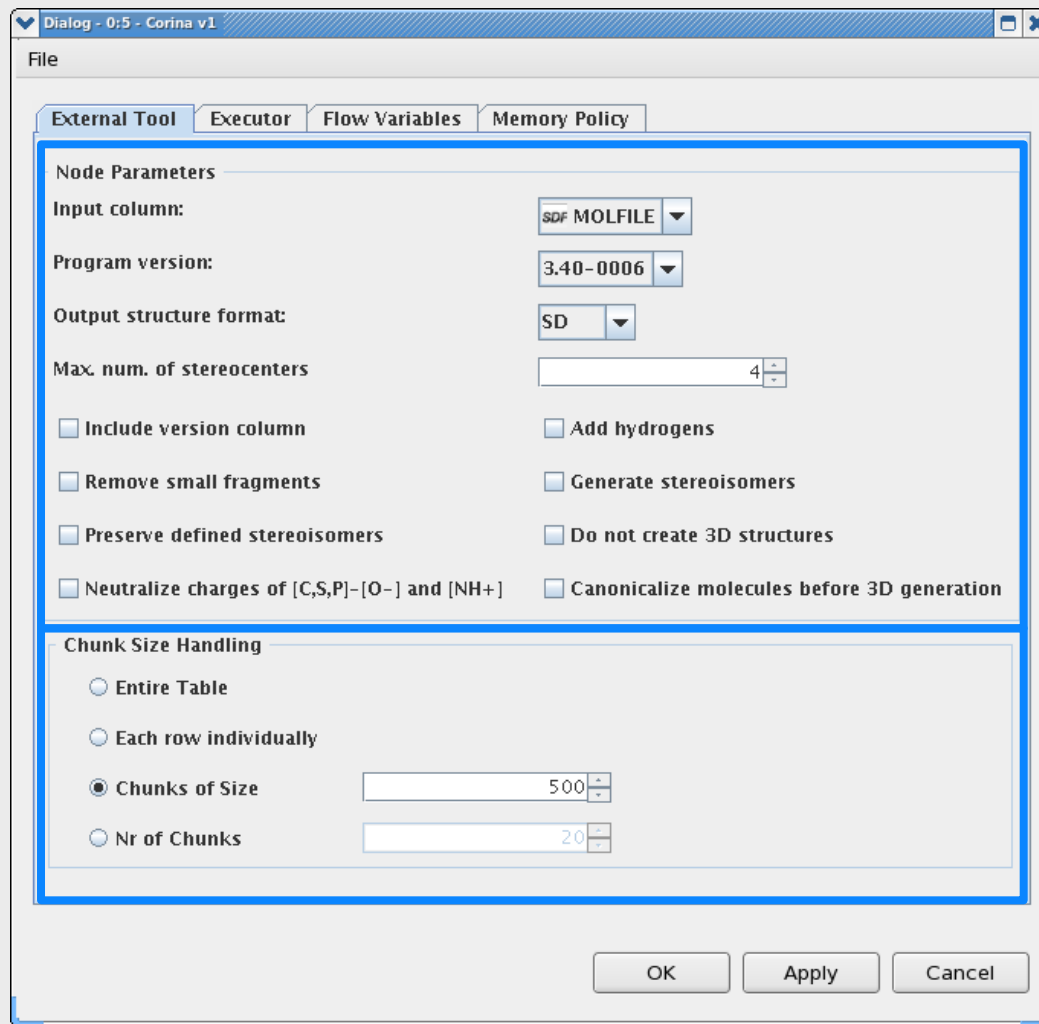
## Headless use:



- Existing workflows rely heavily on various external tools for structure manipulation and descriptor calculation
- Generic integration of external tools into KNIME necessary (incl. optional parallel cluster execution)
  - Workflow migration to KNIME without switching the underlying chemistry tools
- Sponsoring of a generic “External Tool” node/framework
  - Specifications by BI in collaboration with KNIME.com
  - Implementation by KNIME.com
  - Code part of the KNIME open source release
- Adaptation of the generic framework for specific tools and setup @ BI



## Corina options



Dialog - 0:5 - Corina v1

File

External Tool | Executor | Flow Variables | Memory Policy

**Node Parameters**

Input column: SDF MOLFILE

Program version: 3.40-0006

Output structure format: SD

Max. num. of stereocenters: 4

Include version column

Add hydrogens

Remove small fragments

Generate stereoisomers

Preserve defined stereoisomers

Do not create 3D structures

Neutralize charges of [C,S,P]-[O-] and [NH+]

Canonicalize molecules before 3D generation

**Chunk Size Handling**

Entire Table

Each row individually

Chunks of Size: 500

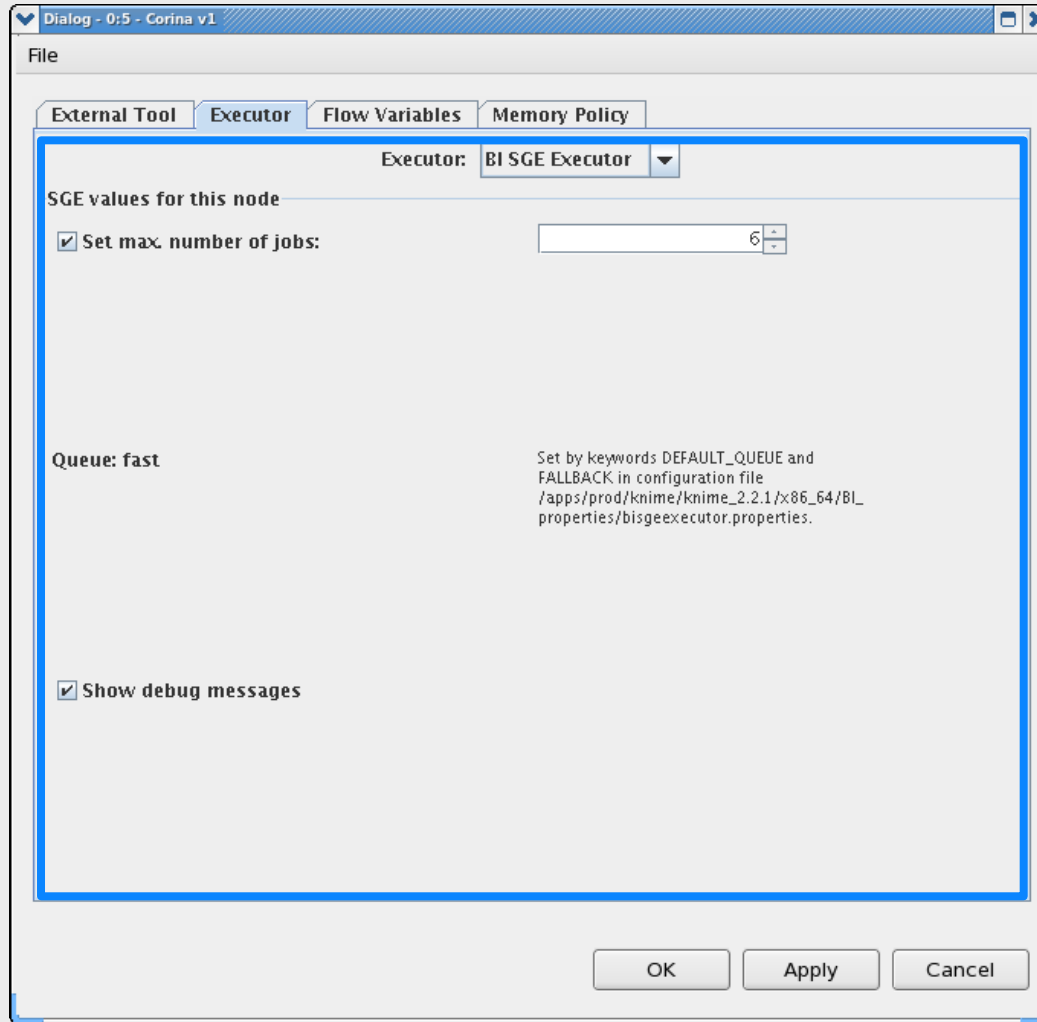
Nr of Chunks: 20

OK Apply Cancel

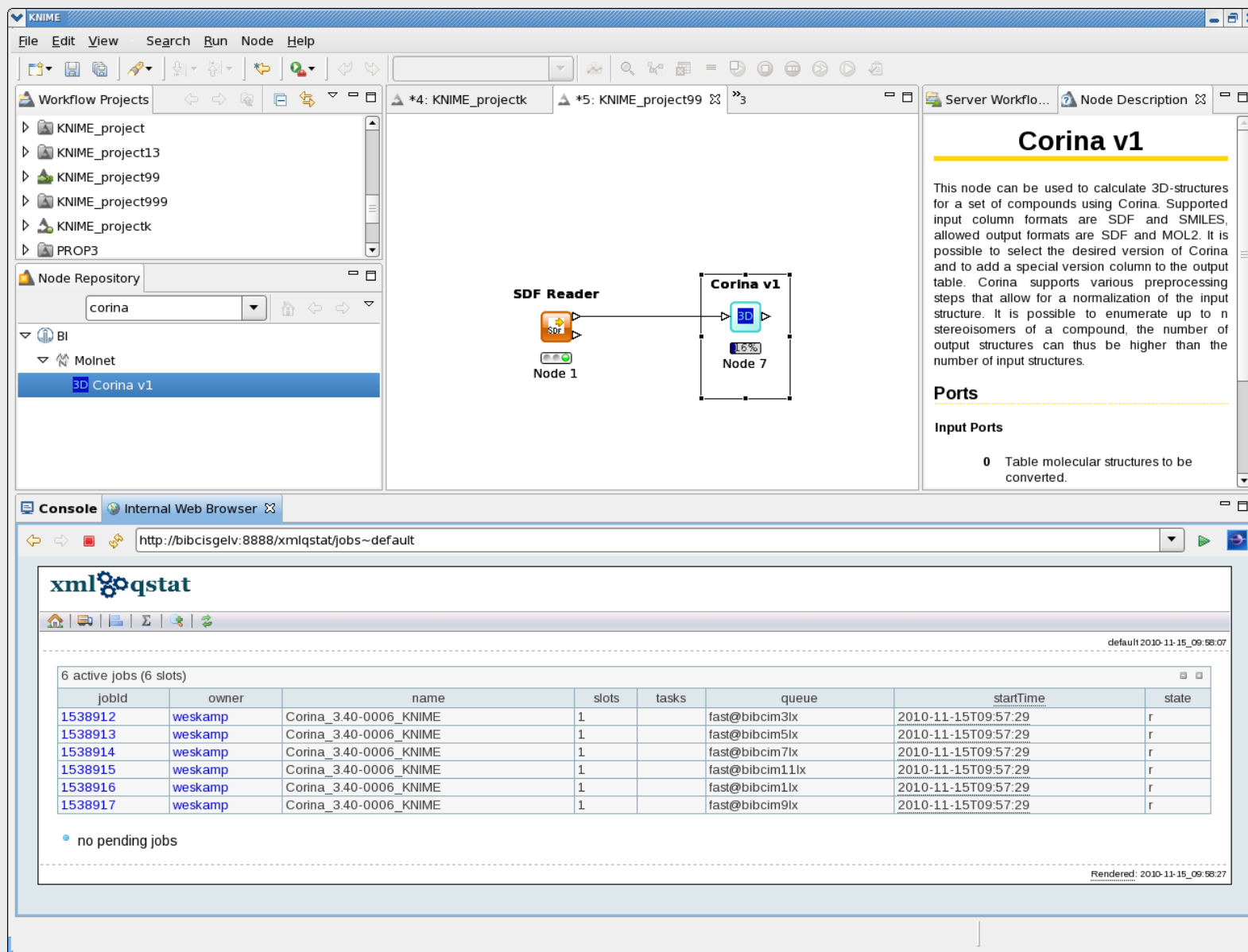
Batch creation  
for parallel  
execution

Local execution  
vs. grid engine  
submission

Error handling  
and automatic  
re-submission of  
failed jobs



Automatic queue  
selection based  
on external tool  
characteristics



The screenshot displays the KNIME software interface. On the left, the 'Node Repository' shows the '3D Corina v1' node selected. The main workspace contains a workflow with an 'SDF Reader' node (Node 1) connected to a 'Corina v1' node (Node 7). The 'Corina v1' node description is shown on the right, detailing its capabilities for calculating 3D-structures and listing its input ports. The bottom console shows an 'Internal Web Browser' displaying the 'xmlqstat' web interface, which provides a table of active jobs and their status.

**Corina v1**

This node can be used to calculate 3D-structures for a set of compounds using Corina. Supported input column formats are SDF and SMILES, allowed output formats are SDF and MOL2. It is possible to select the desired version of Corina and to add a special version column to the output table. Corina supports various preprocessing steps that allow for a normalization of the input structure. It is possible to enumerate up to n stereoisomers of a compound, the number of output structures can thus be higher than the number of input structures.

**Ports**

**Input Ports**

0 Table molecular structures to be converted.

**xmlqstat**

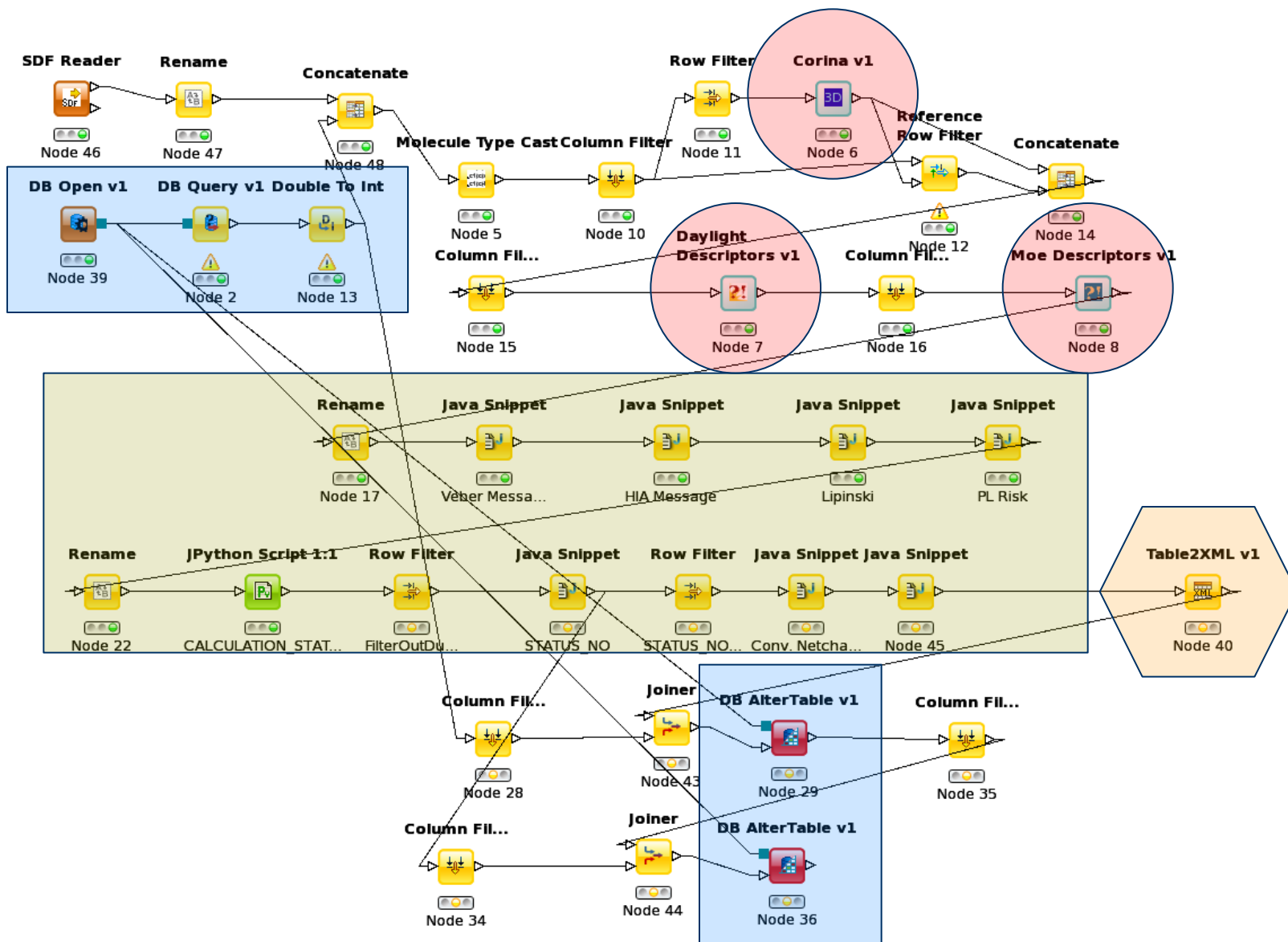
default 2010-11-15\_09:58:07

6 active jobs (6 slots)							
jobId	owner	name	slots	tasks	queue	startTime	state
1538912	weskamp	Corina_3.40-0006_KNIME	1		fast@bibcim3lx	2010-11-15T09:57:29	r
1538913	weskamp	Corina_3.40-0006_KNIME	1		fast@bibcim5lx	2010-11-15T09:57:29	r
1538914	weskamp	Corina_3.40-0006_KNIME	1		fast@bibcim7lx	2010-11-15T09:57:29	r
1538915	weskamp	Corina_3.40-0006_KNIME	1		fast@bibcim11lx	2010-11-15T09:57:29	r
1538916	weskamp	Corina_3.40-0006_KNIME	1		fast@bibcim1lx	2010-11-15T09:57:29	r
1538917	weskamp	Corina_3.40-0006_KNIME	1		fast@bibcim9lx	2010-11-15T09:57:29	r

no pending jobs

Rendered: 2010-11-15\_09:58:27

# Example workflow – property calculation (PROP4)



- Most of the necessary external chemistry tools available within KNIME
- Large number of workflows for property predictions ported to KNIME
  - Testing under realistic conditions yielded promising results
  - Identical results from old and new workflows
  - Switch to productive use of KNIME expected soon
- Rollout of KNIME within the computational chemistry group for interactive usage
  - First end-user training completed
  - Currently extensive evaluation within the group
  - Received a lot of feedback, mainly on usability issues
  - Decision on productive use of KNIME in this application expected in 2011

- Oliver Wissdorf
- Bernd Wiswedel ([KNIME.com](http://KNIME.com))