

Building a Drug Discovery Workflow in 8+1 Steps with KNIME

Dóra Barna, Norbert Sas
ChemAxon

Extended KNIME Spring Summit 2020
Webinar, May 7, 2020





Before we start

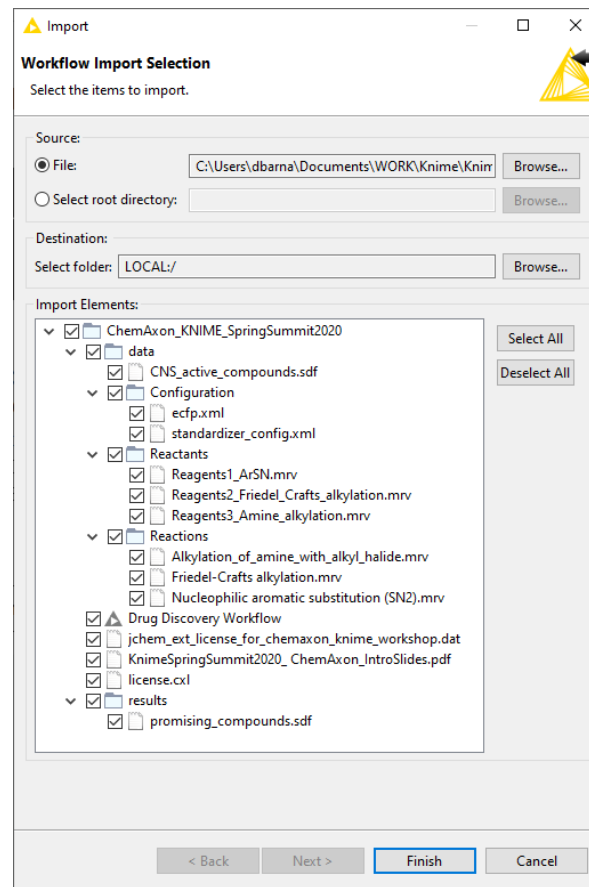
- Install the KNIME Analytics Platform
- Download the workflow group
- Open the workflow
- Install the ChemAxon/Infocom nodes
- Install the ChemAxon and the JChem Extensions licenses





Download and open the workflow

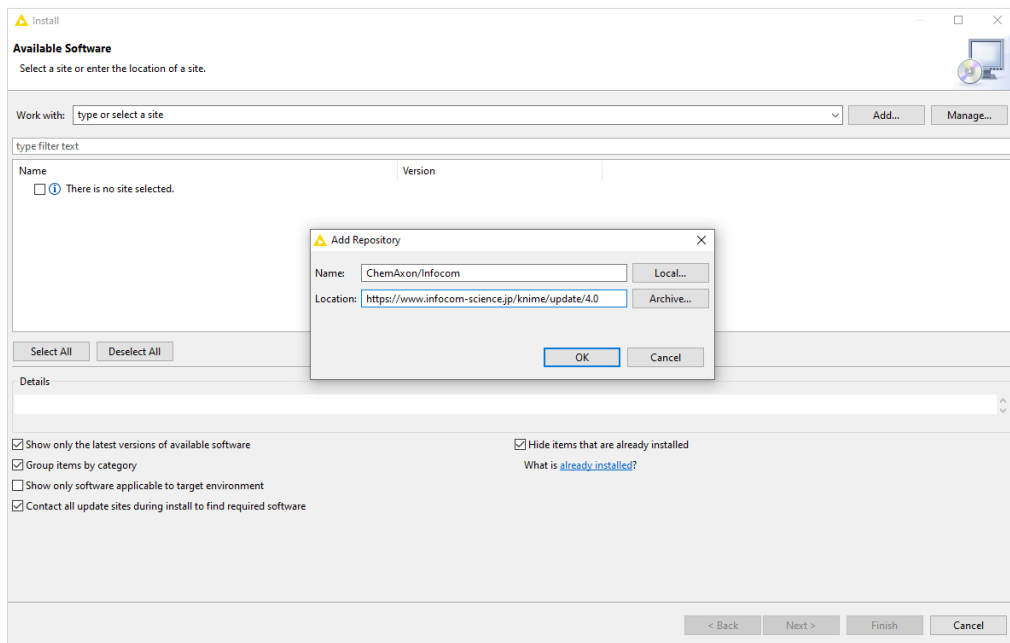
- Workflow group to download:
ChemAxon_KNIME_SpringSummit2020.knar
- Double-click on the downloaded workflow group
- Import all files to the KNIME Analytics Platform
- Open the **Drug_Discovery_Workflow** file





Installing the ChemAxon/Infocom nodes

URL with the latest version: <https://www.infocom-science.jp/knime/update/4.0>

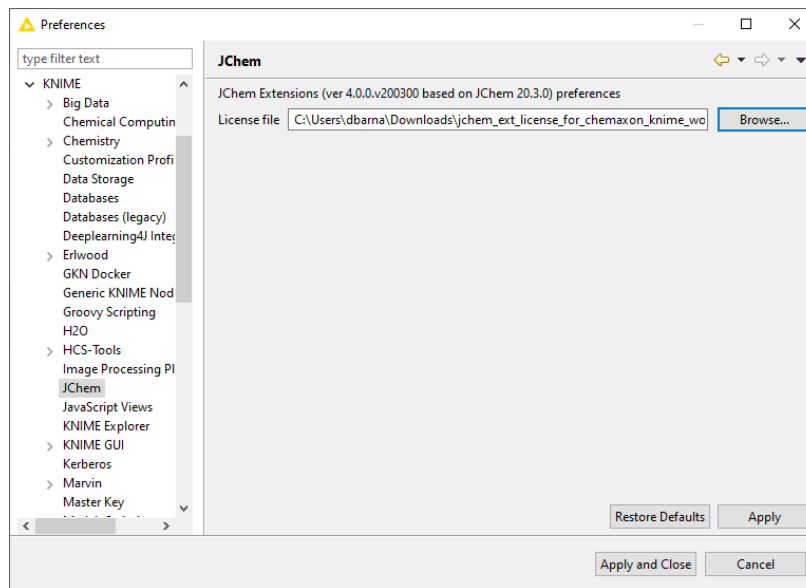
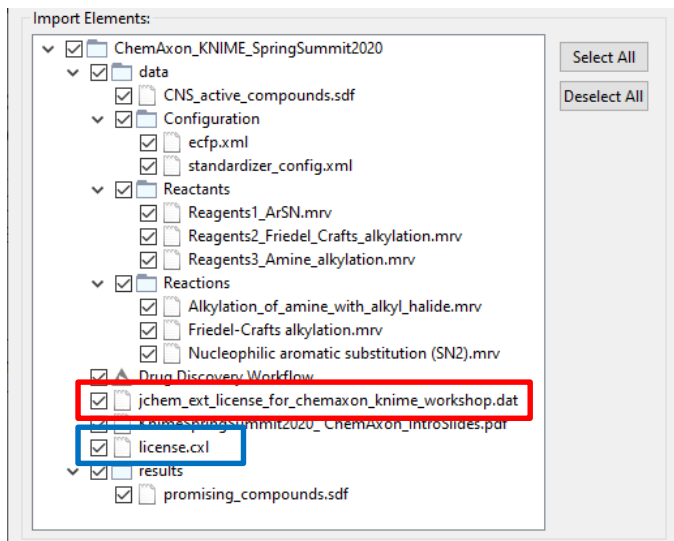


- Help >> Install Software... >> Add...
- Use the URL for „Location”
- OR: download the zipped version and use „Local” installation
- Install both:
 - JChem Extensions,
 - Marvin Extensions
- Restart KNIME when the installation is ready
- [Detailed install guide](#)
- [ChemAxon/Infocom nodes](#)



Installing the ChemAxon and JChem Extensions licenses

- JChem Extensions license: **File >> Preferences >> KNIME >> JChem**
- ChemAxon license: **\$USER_HOME/(.)chemaxon/licenses/license.cxl**



Installing the license files: https://docs.chemaxon.com/KNIME_Nodes_Licensing.html



ChemAxon – Who we are

We provide **software solutions** and **services** to enhance drug discovery and other chemistry related fields with **chemical** & **biological** intelligence.



ChemAxon – What we do





ChemAxon and KNIME

ChemAxon & Infocom

- Chemical drawing and molecule visualization
- Chemoinformatics toolkits
- Marvin Extensions
- JChem Extensions

Server-based ChemAxon software

- *via* the REST Web Services nodes of KNIME
- Structure search, calculations, working with macromolecules...





ChemAxon in KNIME – The Marvin nodes

Node Repository

- IO
- Manipulation
- Views
- Analytics
- DB
- Other Data Types
- Structured Data
- Scripting
- Tools & Services
- Community Nodes
- KNIME Labs
- Workflow Control
- Workflow Abstraction
- Social Media
- Reporting
- Chemistry
 - ChemAxon / Infocom
 - Marvin**
 - MarvinSketch
 - MarvinView
 - MarvinSpace
 - MolConverter
 - Jchem
- Testing
- LigandScout
- Scientific Strategy
- TIBCO
- VEGA
- MOE

Multistep reaction enumeration
Uses chemaxon_reactor_collection.mrx that contains 242 generic reaction equations with rules based on literature.

1. Reaction - Niementowski (ring closure)

2. reaction - Acylation

3. Analyzing products

MarvinTable

List of reactants and products for all individual reaction

MarvinView

JCX4L Writer

Node 128

Dialog - 4118 - MarvinSketch (Draw reaction)

Options: Output options | Flow Variables | Job Manager Selection | Memory Policy

File Edit View Insert Atom Bond Structure Calculations Services Help

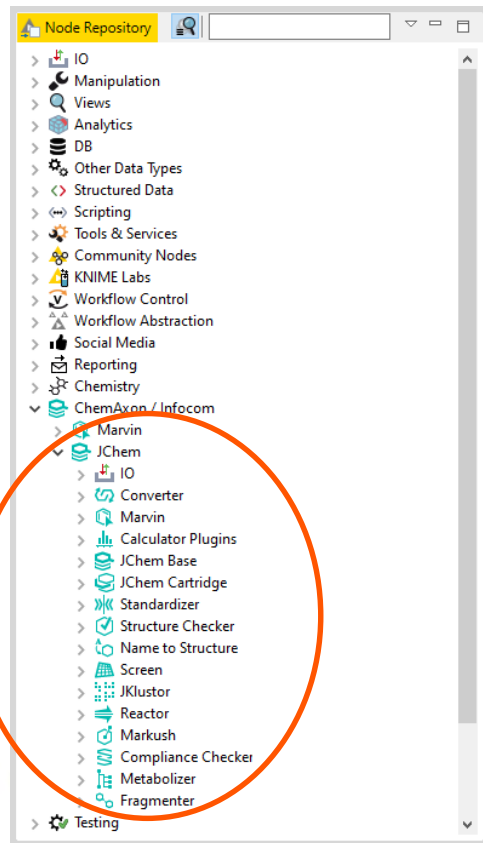
20

OK Apply Cancel

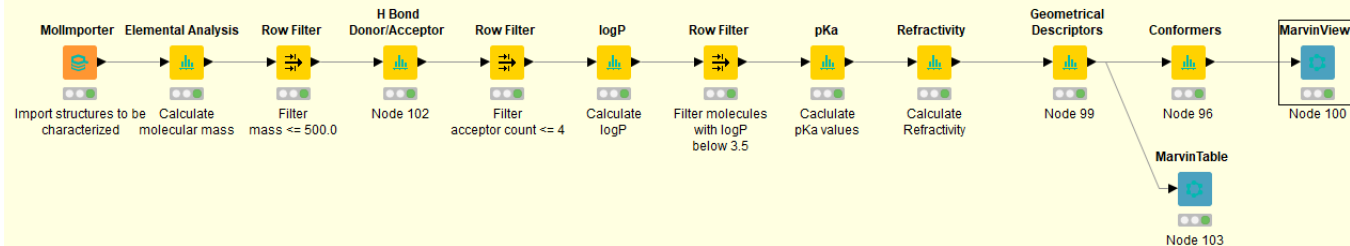
MarvinTable - 4127 - MarvinTable (List of reactants and products)

Row ID	Molecule	COMPO...	Reactant ID	Molecule	StandardizerResult	COMPO...
Row7		CPD_00115	REACTANT_3_1		Remove Explicit Hydrogens (ID: Remove Explicit Hydrogens)	CPD_00020
Row8		CPD_00115	REACTANT_3_1		Remove Explicit Hydrogens (ID: Remove Explicit Hydrogens)	CPD_00020
Row9		CPD_00115	REACTANT_3_1		Remove Explicit Hydrogens (ID: Remove Explicit Hydrogens)	CPD_00020
Row10		CPD_00115	REACTANT_3_1		Remove Explicit Hydrogens (ID: Remove Explicit Hydrogens)	CPD_00020
Row11		CPD_00115	REACTANT_3_1		Remove Explicit Hydrogens (ID: Remove Explicit Hydrogens)	CPD_00020
Row12		CPD_00115	REACTANT_3_1		Remove Explicit Hydrogens (ID: Remove Explicit Hydrogens)	CPD_00020
Row13		CPD_00115	REACTANT_3_1		Remove Explicit Hydrogens (ID: Remove Explicit Hydrogens)	CPD_00020
Row14		CPD_00115	REACTANT_3_1		Remove Explicit Hydrogens (ID: Remove Explicit Hydrogens)	CPD_00020

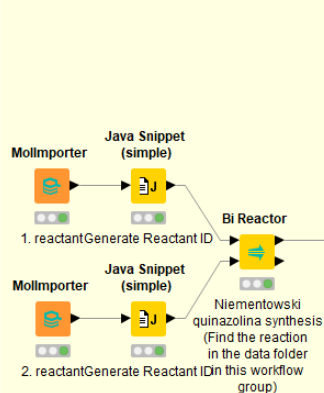
ChemAxon in KNIME – The JChem nodes



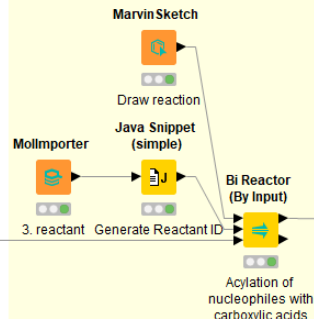
Filter structures based on calculated and predicted properties



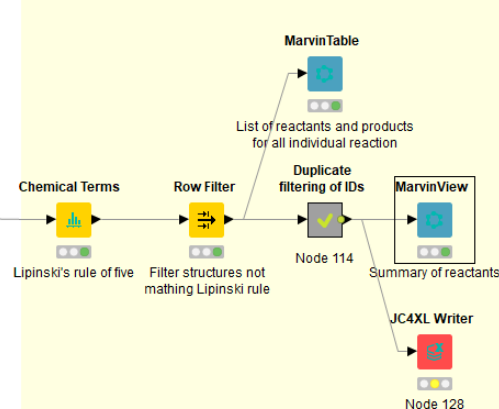
1. Reaction - Niementowski (ring closure)



2. reaction - Acylation

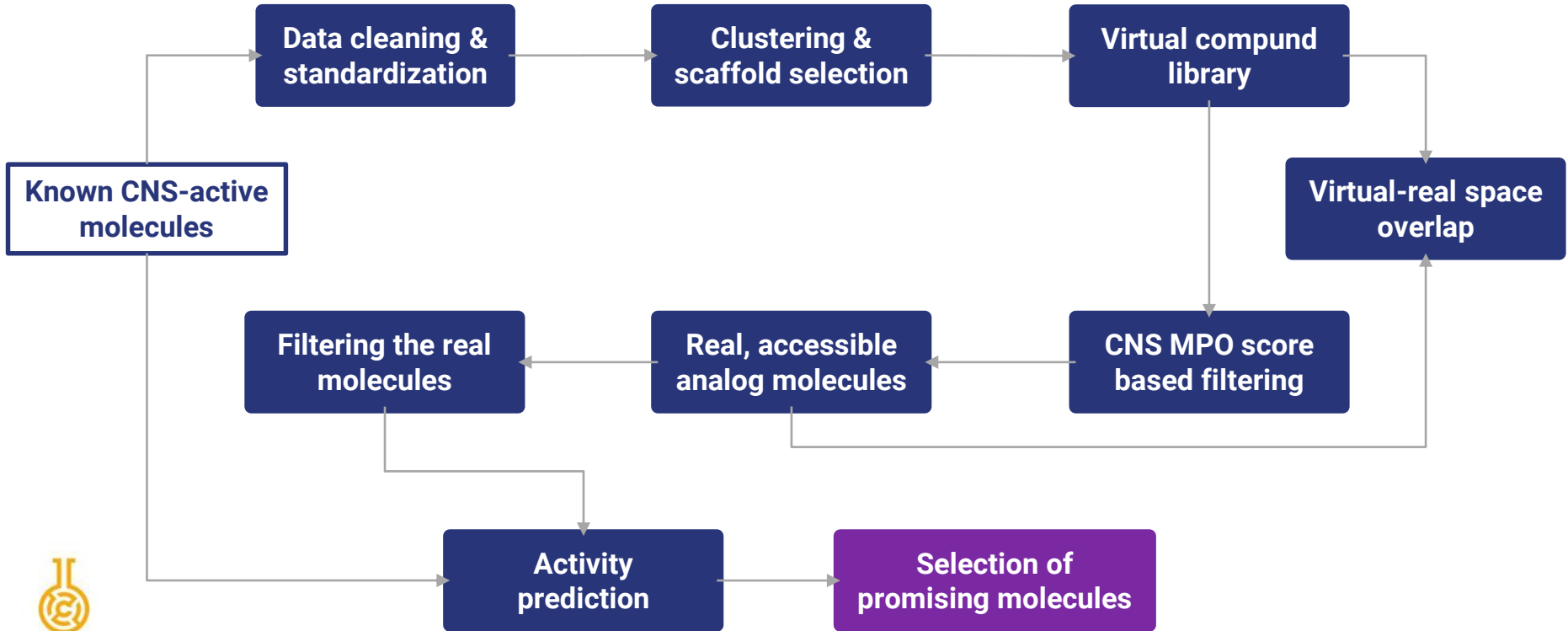


3. Analyzing products



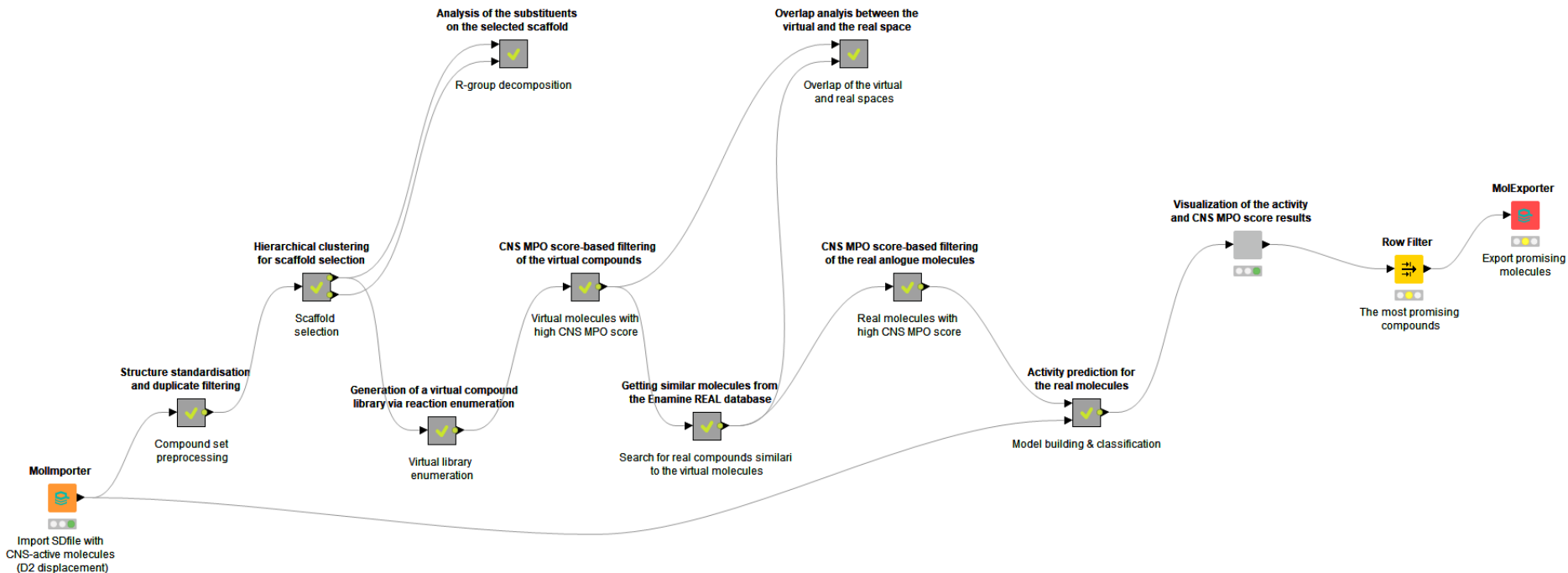


A Drug Discovery Workflow in 8+1 Steps

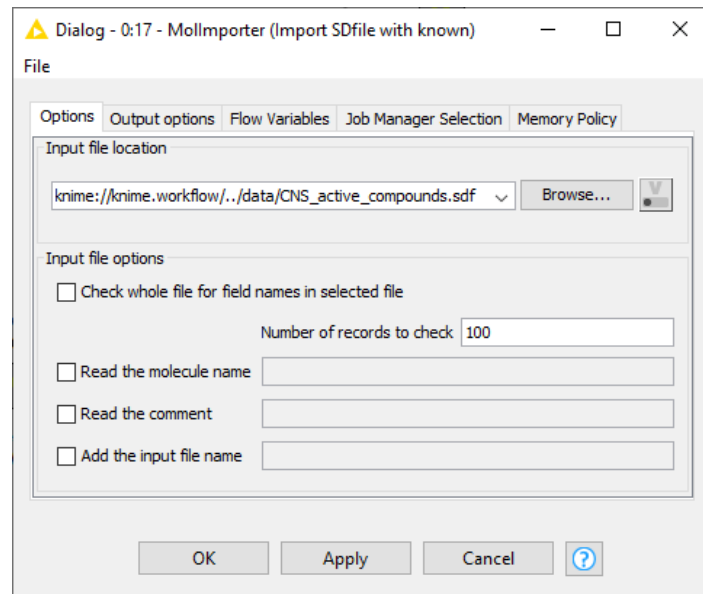
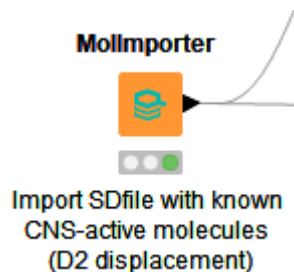




A Drug Discovery Workflow in 8+1 Steps



Step 1 – The input data

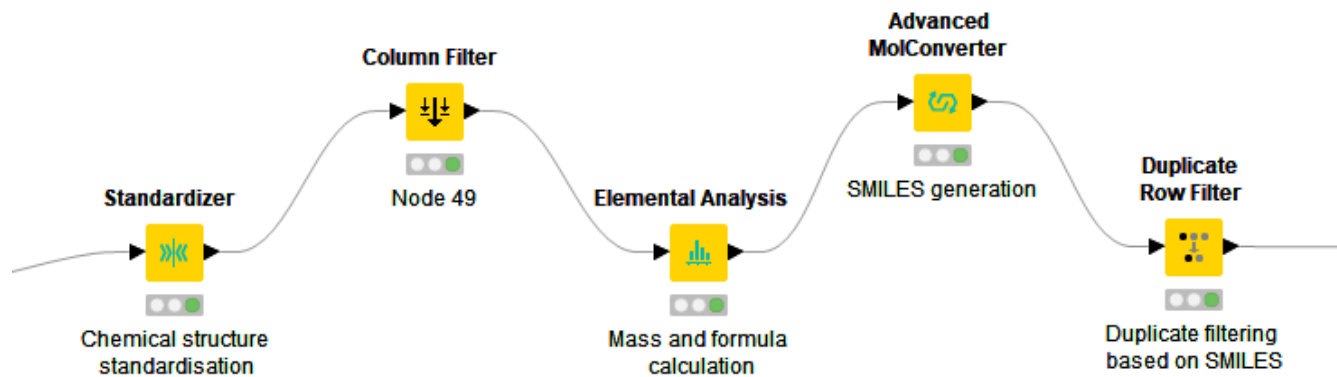


3257 compounds from ChEMBL that were tested against
the [human dopamine D2 receptor](#) (have a pChEMBL value)





Step 2 – Structure pre-processing

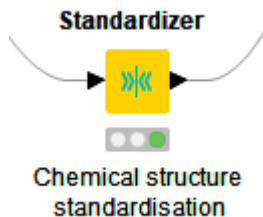


- Structure standardisation
- Molecular weight and SMILES generation
- Duplicate filter on the standardized structures using SMILES

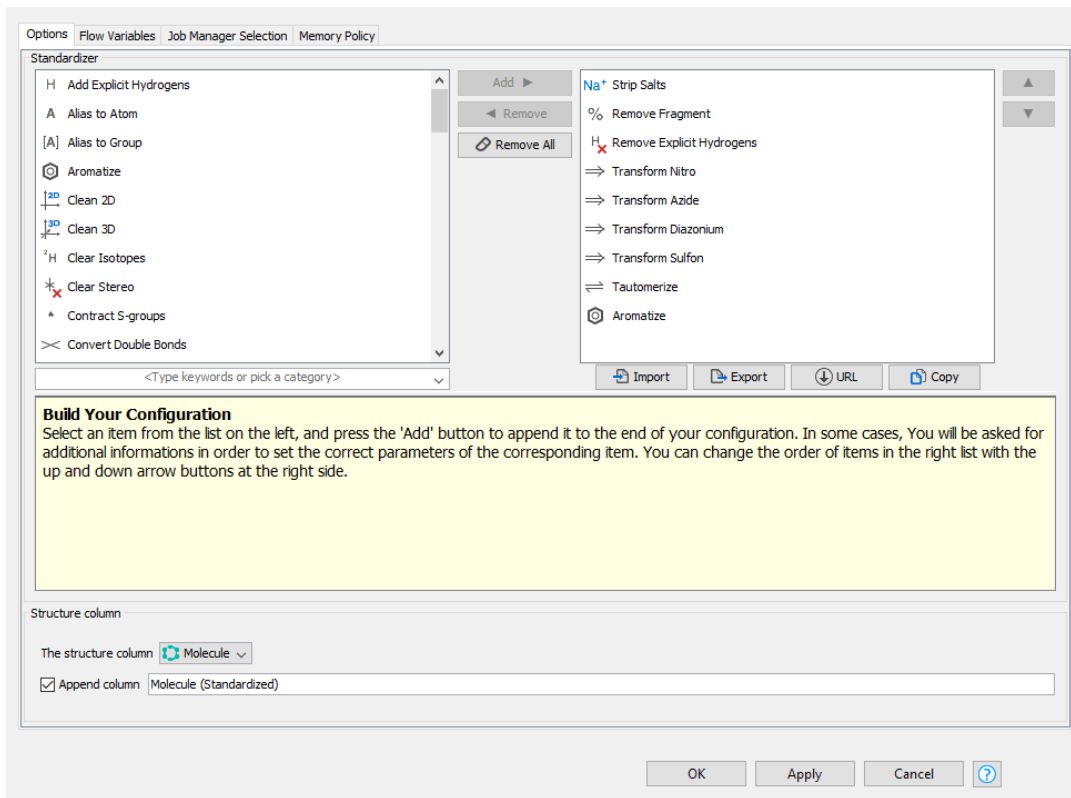




Step 2 – Structure standardisation

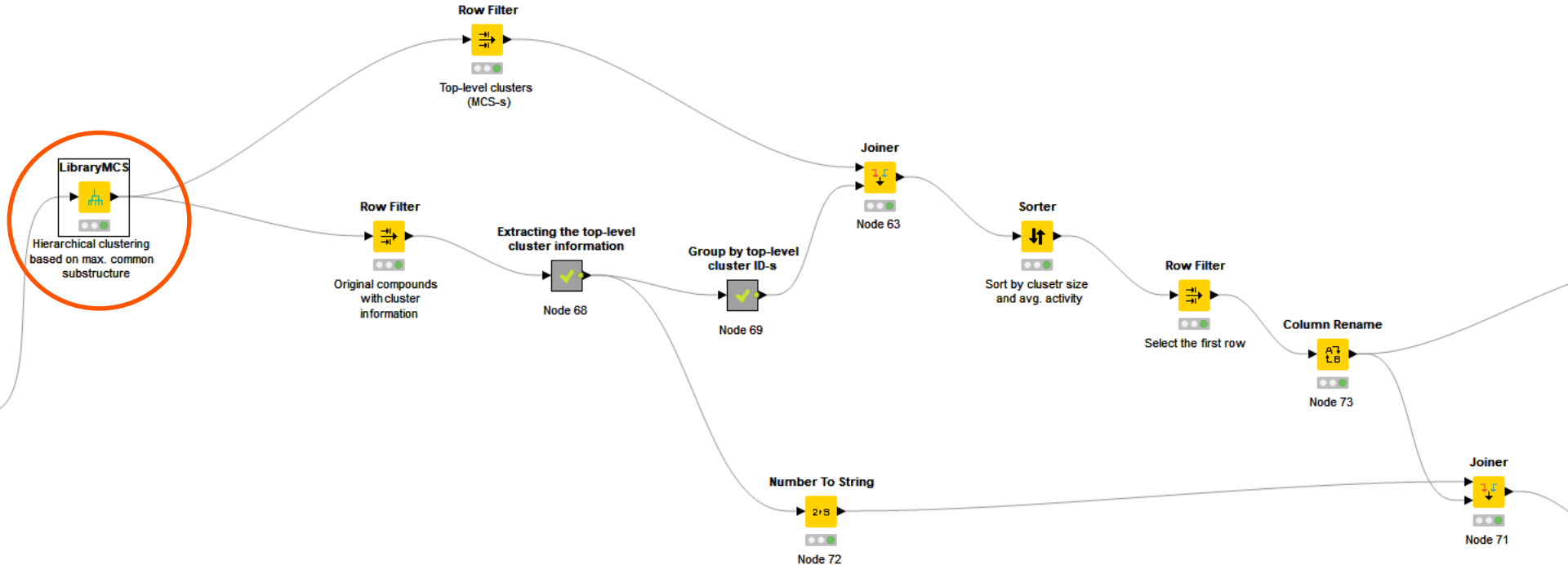


- Aromatize, strip salts, add/remove expl. H-s, tautomerize, neutralize...
- Transform typical functional groups or structural patterns to make them consistent





Step 3 – Clustering and scaffold selection





Step 3 – MCS-based hierarchical clustering

LibraryMCS



Hierarchical clustering
based on max. common
substructure

Flow Variables Options Job Manager Selection Clustering Memory Policy Output options

Minimal MCS Size

Minimal MCS Size: 12

MCS Mode

☒ Normal

☐ Fast

Matching Parameter

☒ Atom type

☒ Bond type

☒ Charge

☐ Radicals

☐ Isotopes

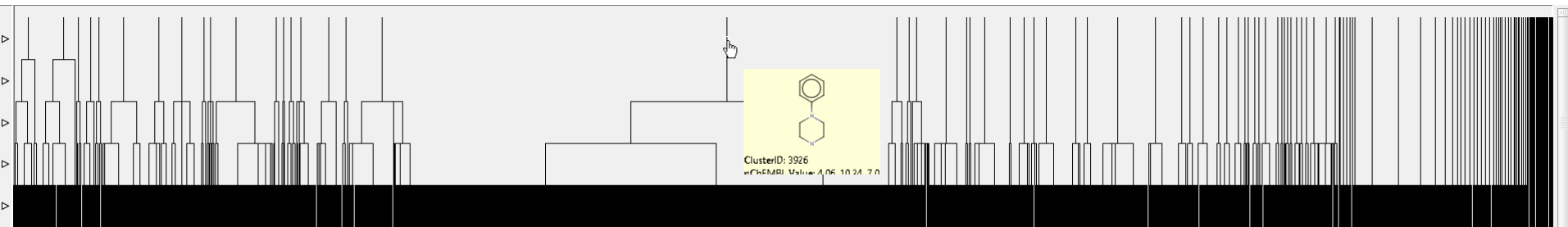
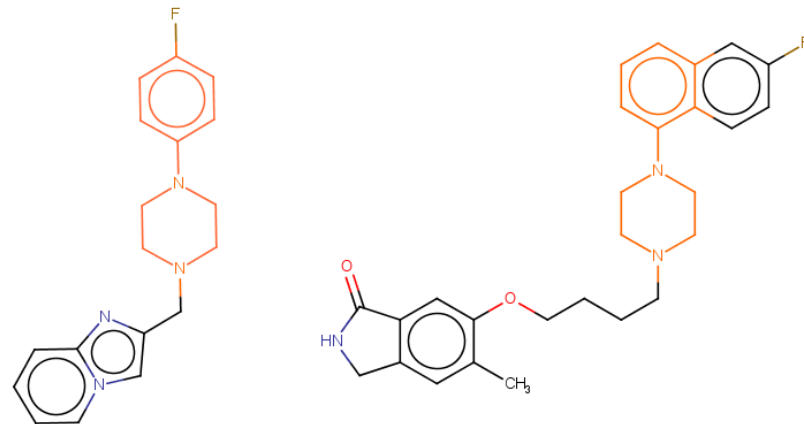
Structure column

The structure column: Molecule (Standardized) ▼

☒ Keep properties of structure

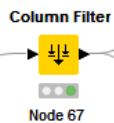
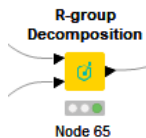
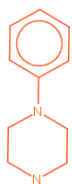
OK Apply Cancel ?

MCS (max. common substructure) of 2 molecules





Step 3 – R-group decomposition



Ortho substituents
on the aromatic ring



Node 91

Substituents on the
piperazine nitrogen

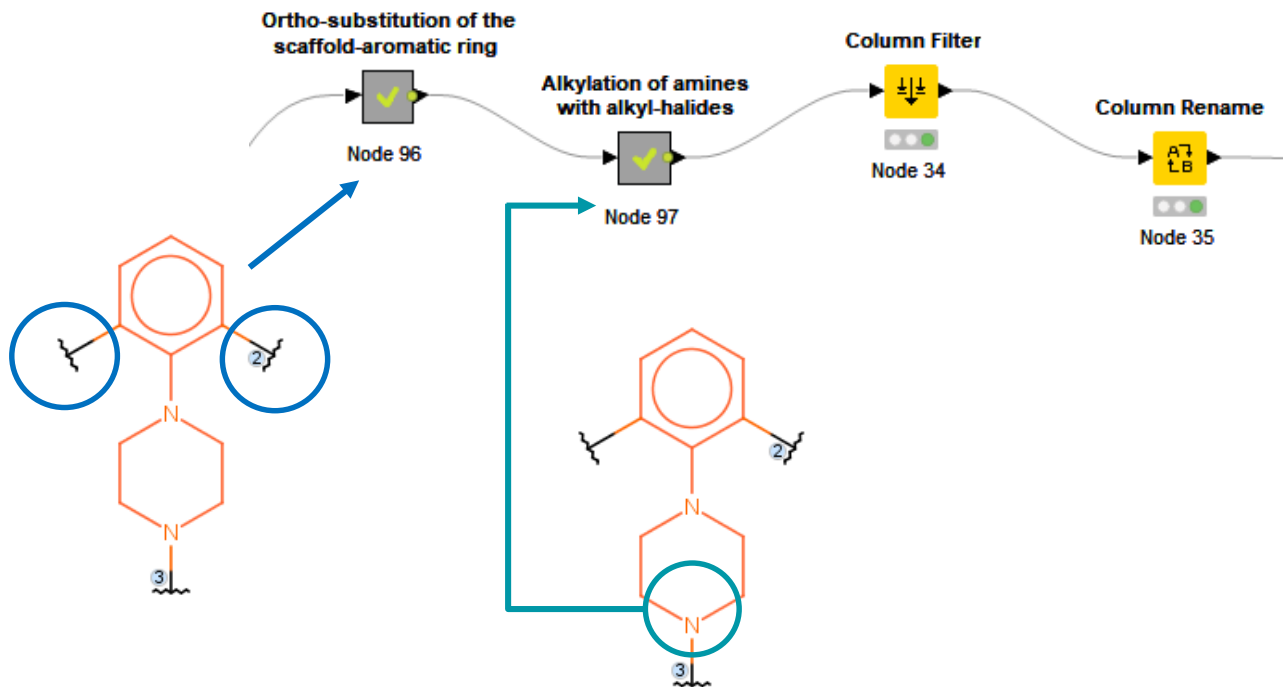


Node 92

Row ID	Molecule (cluster)	Target molecule	Ligand 1	Ligand 2	Ligand 3	Ligand 4	Ligand 5	Ligand 12	Ligand 14	Molecule Ch...
Row1153_Ro...			Cl ¹	Cl ²	3 H	4 H	5 H	12 H		CHEMBL244774
Row1056_Ro...			1 H	2 H	3 H	4 H	5 O	12 H		CHEMBL272602
Row1885_Ro...			1 H	2 H	3 H	4 H	5 O	12 H		CHEMBL3758330
Row1363_Ro...			1 H	2 H	3 H	4 H	5 O	12 H		CHEMBL225512
Row1394_Ro...			1 H	O ²	3 H	4 H	5 H	12 H		CHEMBL3759491

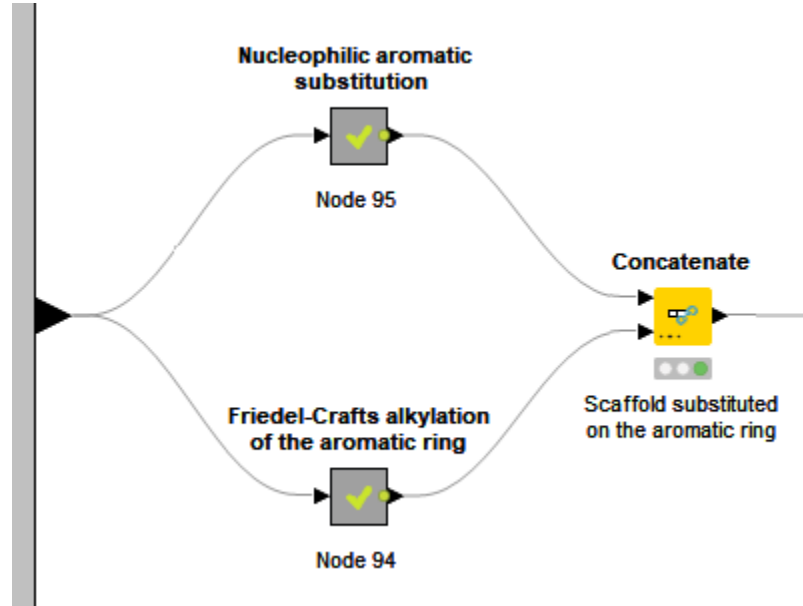
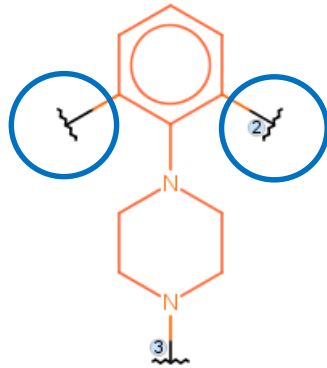


Step 4 – Virtual library enumeration

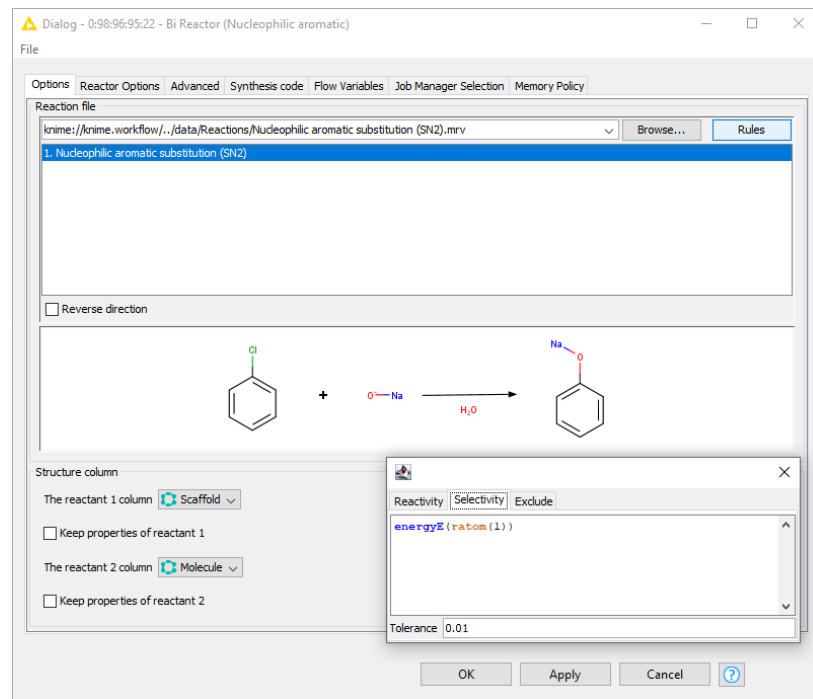
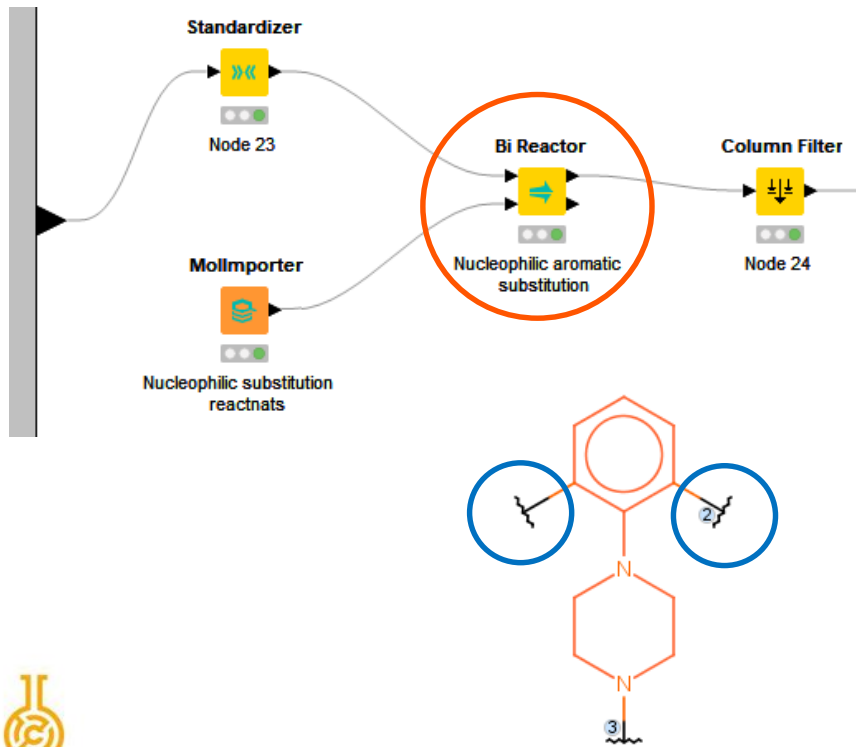




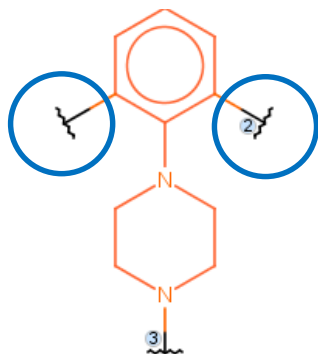
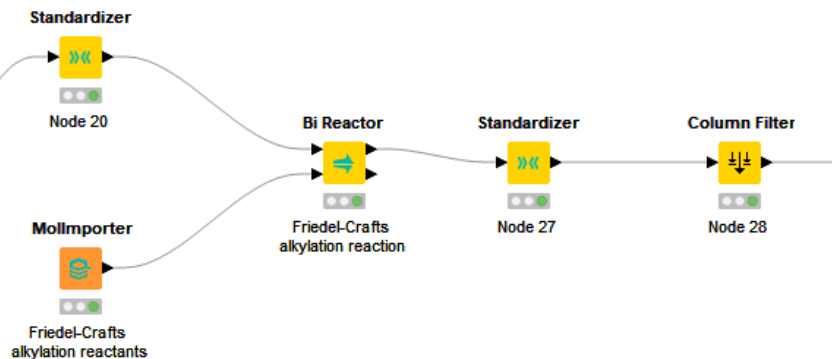
Step 4 – Substitution of the aromatic ring



Step 4 – Nucleophilic aromatic substitution



Step 4 – Friedel-Crafts alkylation



Dialog - 0:98:96:94:18 - Bi Reactor (Friedel-Crafts)

File

Options Reactor Options Advanced Synthesis code Flow Variables Job Manager Selection Memory Policy

Reaction file

krime://krime.workflow/./data/Reactions/Friedel-Crafts alkylation.mrv Browse... Rules

1. Friedel-Crafts alkylation

☐ Reverse direction

Structure column

The reactant 1 column Scaffold ☐ Keep properties of reactant 1

The reactant 2 column Molecule ☐ Keep properties of reactant 2

Reactivity Selectivity Exclude

```
charge(ratom(1), "aromaticsystem") <= -0.2
```

OK Apply Cancel ?



Step 4 – Alkylation of the piperazine nitrogen

Alkylation of amines
with alkyl-halides

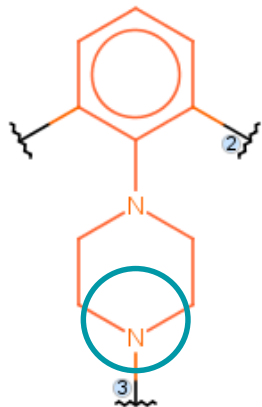


Node 97

Bi Reactor



Alkylation of the
piperazine nitrogen



Reactivity	Selectivity	Exlude
<pre>match(ratom(1), "[#6:1] [NX3H2;!\$(NC~[!#1!#6]):2]") or match(ratom(1), "[#6:1] [NX3H1;!\$(NC~[!#1!#6]):2]") and match(ratom(2), alkyl_halide) and !match(ratom(2), aryl_halide)</pre>		

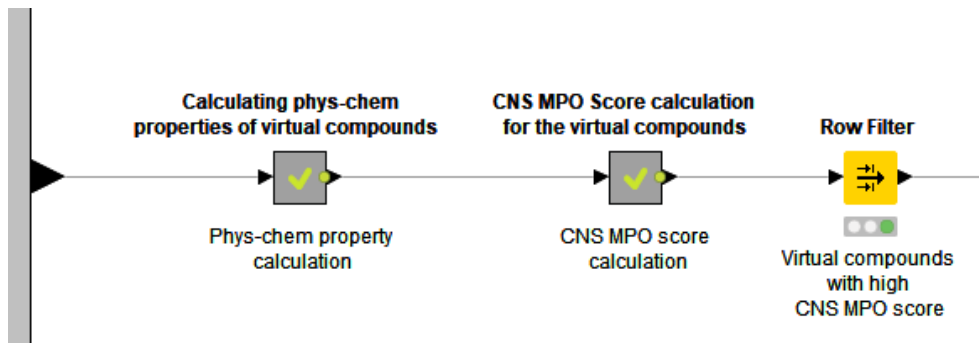


Table "default" - Rows: 45416 Spec - Column: 1 Properties Flow Variables

Row ID	Virtual compounds
Row0	
Row1	
Row2	
Row3	
Row4	



Step 5 – Property prediction and library filtering



- Molecules with a high probability to cross the blood-brain barrier
- BBB penetration influenced by key phys-chem parameters
- CNS MPO score^{*}, ^{**}: a weighted scoring function assessing 6 phys-chem properties

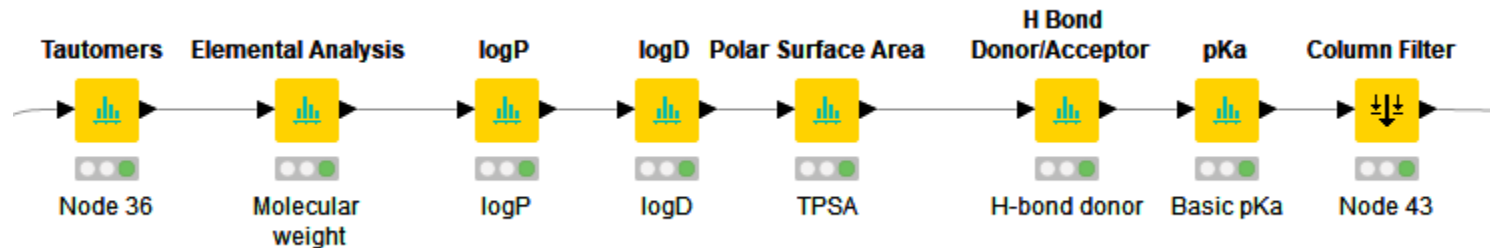


(^{*}): Wager TT et al. Moving beyond rules: The development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties. ACS Chem. Neurosci. 2010 1, 435-449.

(^{**}): Wager TT et al. Central nervous system multiparameter optimization desirability: Application in drug discovery. ACS Chem. Neurosci. 2016 7, 767-7



Step 5 – Physicochemical property prediction

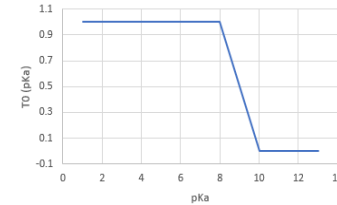
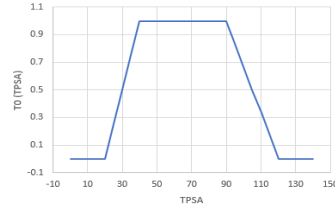
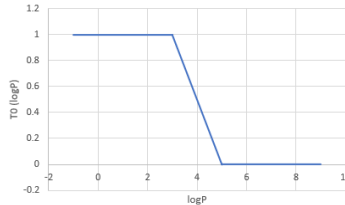
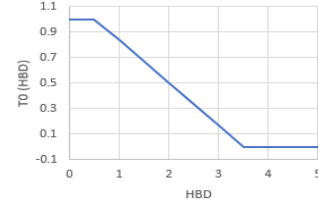
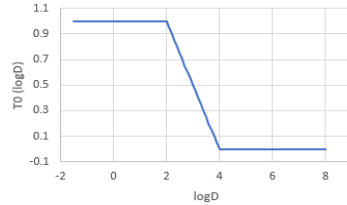
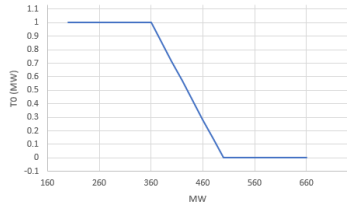
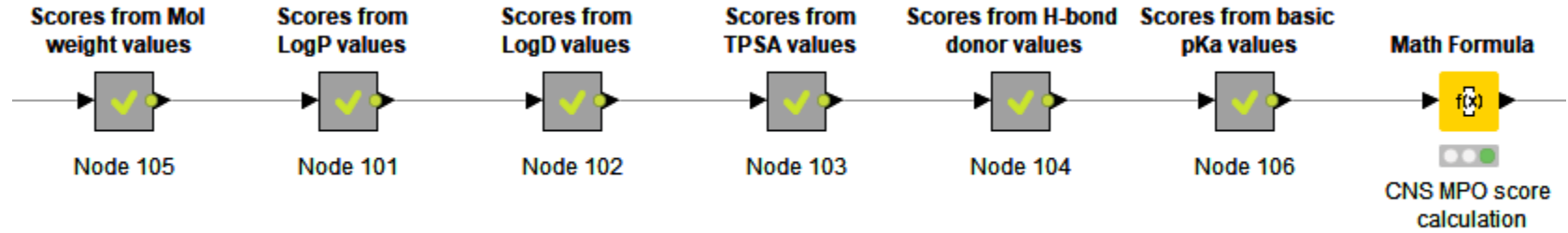


- Molecular weight
- logP
- logD at pH = 7.4
- Polar surface area
- Hydrogen bond donor atoms
- Strongest basic pKa





Step 5 – Weighted scoring function and filtering

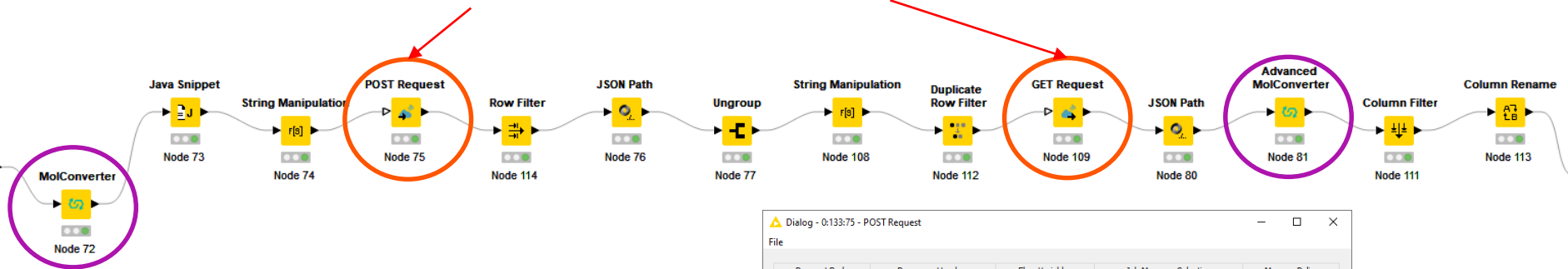




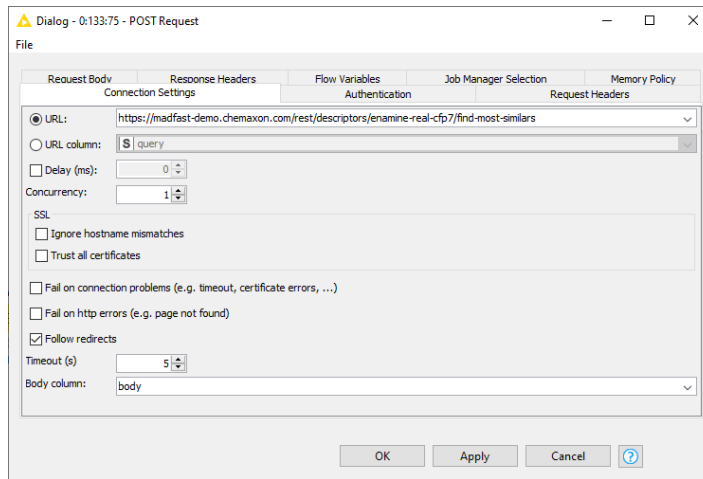
Step 6 – Looking for accessible analogue molecules

Web service calls to find the
20 most similar analogues

















Web service calls to get the
SMILES of the Enamine cpds.



- Similarity search in a subset of the Enamine REAL DB (~170M cpds.)
- In-memory search with Madfast Sim.
- Search via web service calls


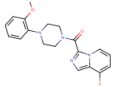

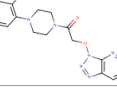

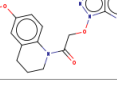

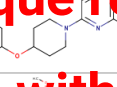

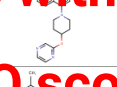

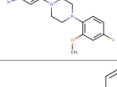

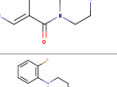

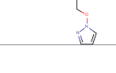


Step 7 – CNS MPO score based filtering of the analogues

Table "default" - Rows: 5598 Spec - Columns: 4 Properties Flow Variables				
Row ID	Virtual molecule	D CNS MPO Score (virtual molecule)	D Dissimilarity	Real analogue molecule
Row43_1		5.74	0.464	
Row43_3		5.74	0.47	
Row43_4		5.74	0.473	
Row43_5		5.74	0.476	
Row43_6		5.74	0.478	
Row43_7		5.74	0.478	
Row43_8		5.74	0.479	
Row43_10		5.74	0.483	

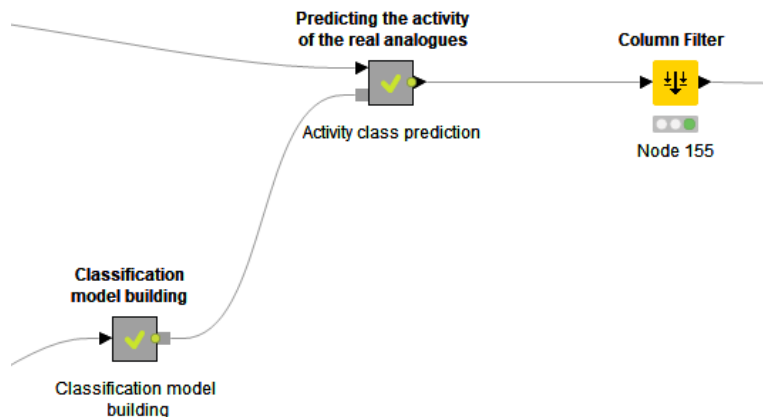
**5598 unique
real compounds**



Table "default" - Rows: 3493 Spec - Columns: 6 Properties Flow Variables					
Row ID	Virtual ...	D CNS MP...	D Dissimilarity	Real analogue molecule	D CNS MP... D ID
Row344_6		5	0.502		6 1
Row345_1		5	0.542		6 2
Row535_11		5	0.546		6 3
Row723_4		5	0.492		6 4
Row723_9		5	0.502		6 5
Row1135_13		5	0.562		6 6
Row2196_17		5	0.545		6 7
Row6833_4		5	0.566		6 8

**3493 unique real
compounds with high
CNS MPO score**

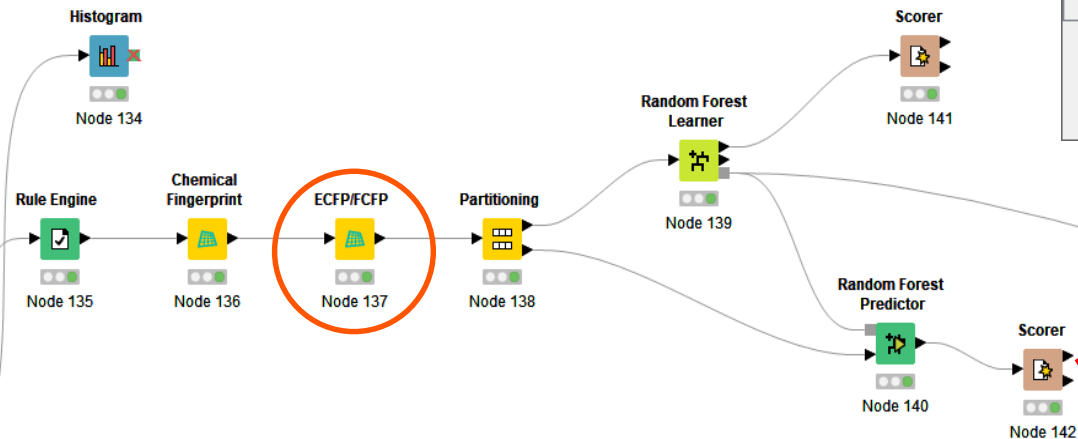
Step 8 – Activity-based classification of the analogues



- Categorizing the Enamine molecules to „active” and „non-active” groups
- Binary classification model with the Random Forest algorithm



Step 8 – Model building



Confusion Matrix - 0:161:147:142 - Scorer

File Hilite

Prediction (Activity) \ Activity	Active	Non-active
Active	206	52
Non-active	58	336

Correct classified: 542	Wrong classified: 110
Accuracy: 83.129 %	Error: 16.871 %
Cohen's kappa (κ) 0.649	





Step 8 – Predicting the activity of the molecules

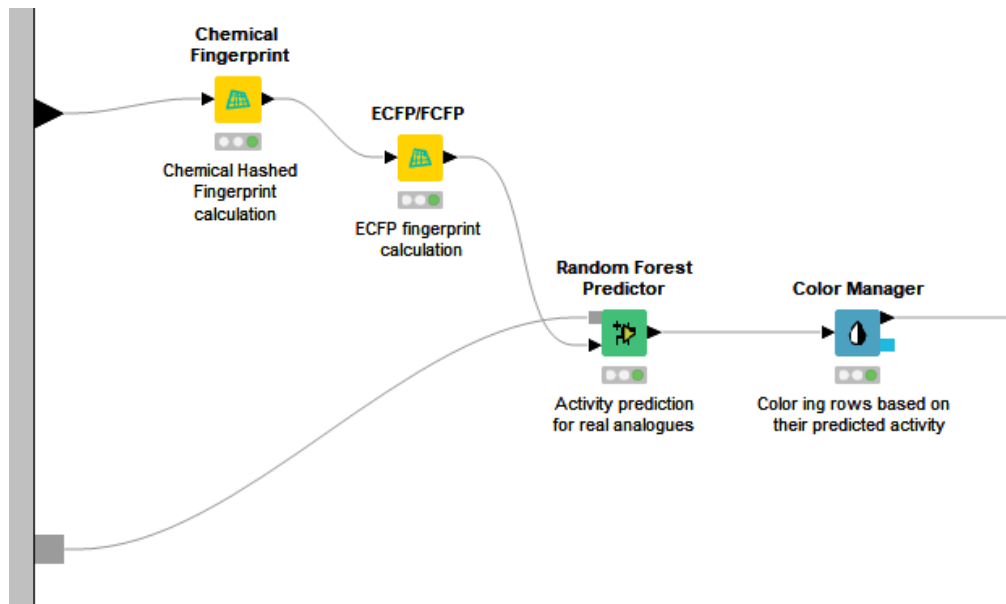
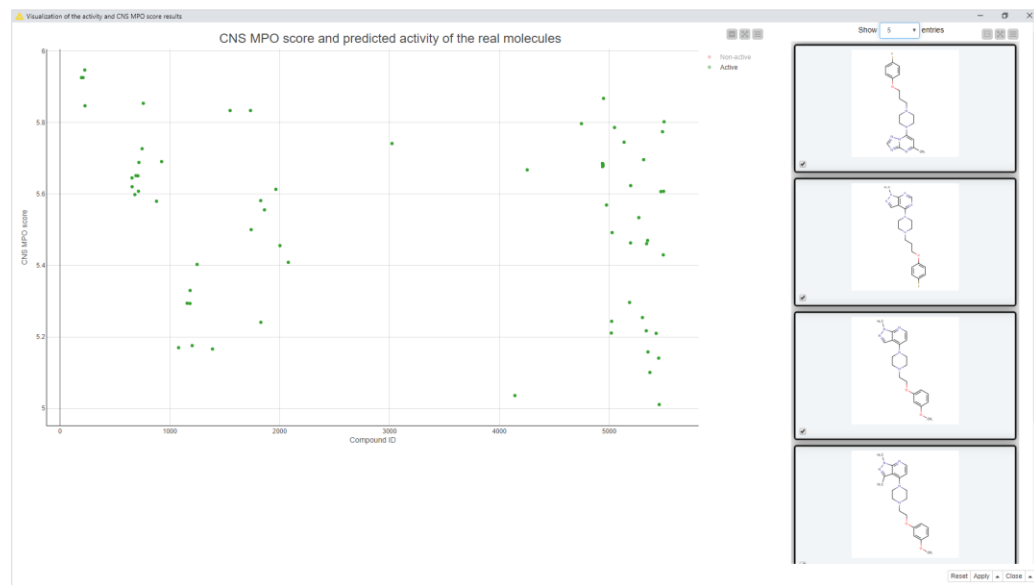
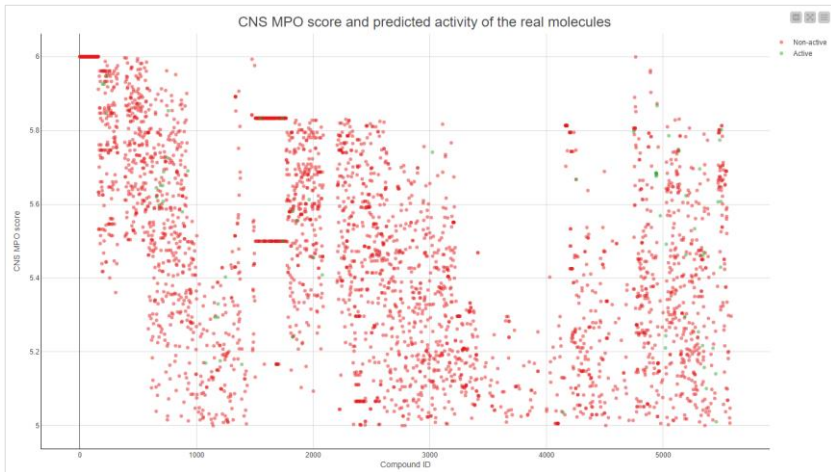
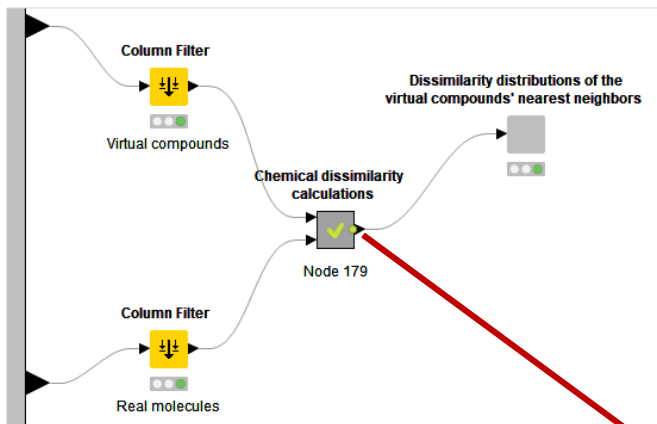


Table "default" - Rows: 3493 Spec - Columns: 5 Properties Flow Variables					
Row ID	Real analogue molecule	ChS MPO Score - Real Analogue	D ID	S Prediction (Activity)	D Prediction (Activity) (Confidence)
Row12085_18		5.294	1,184	Active	0.5
Row12085_19		5.33	1,185	Active	0.5
Row14014_4		5.15	1,189	Non-active	0.75
Row31944_11		5.218	1,200	Non-active	0.68
Row31944_18		5.218	1,201	Non-active	0.67
Row34420_7		5.159	1,202	Non-active	0.57
Row34793_18		5.176	1,204	Active	0.51
Row7555_9		5.38	1,208	Non-active	0.65
Row13498_3		5.1	1,209	Non-active	0.75

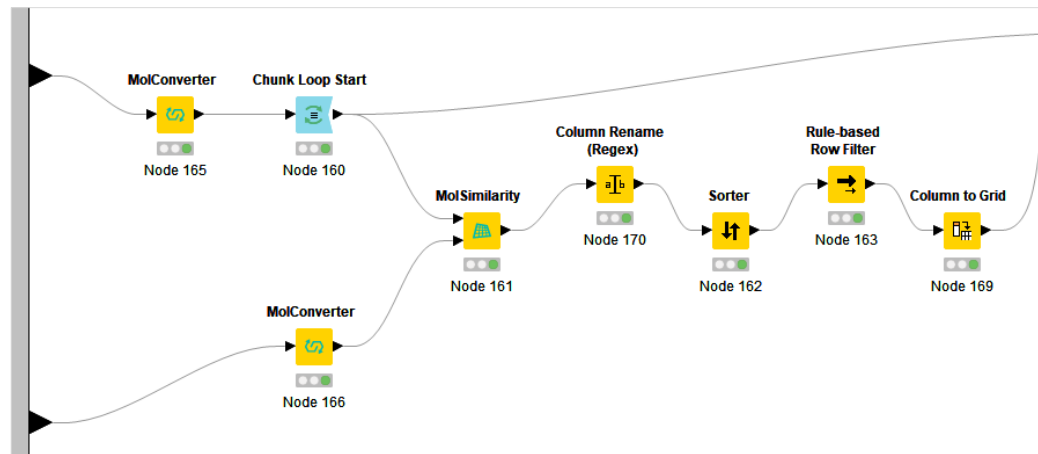
"What to synthesize next?"



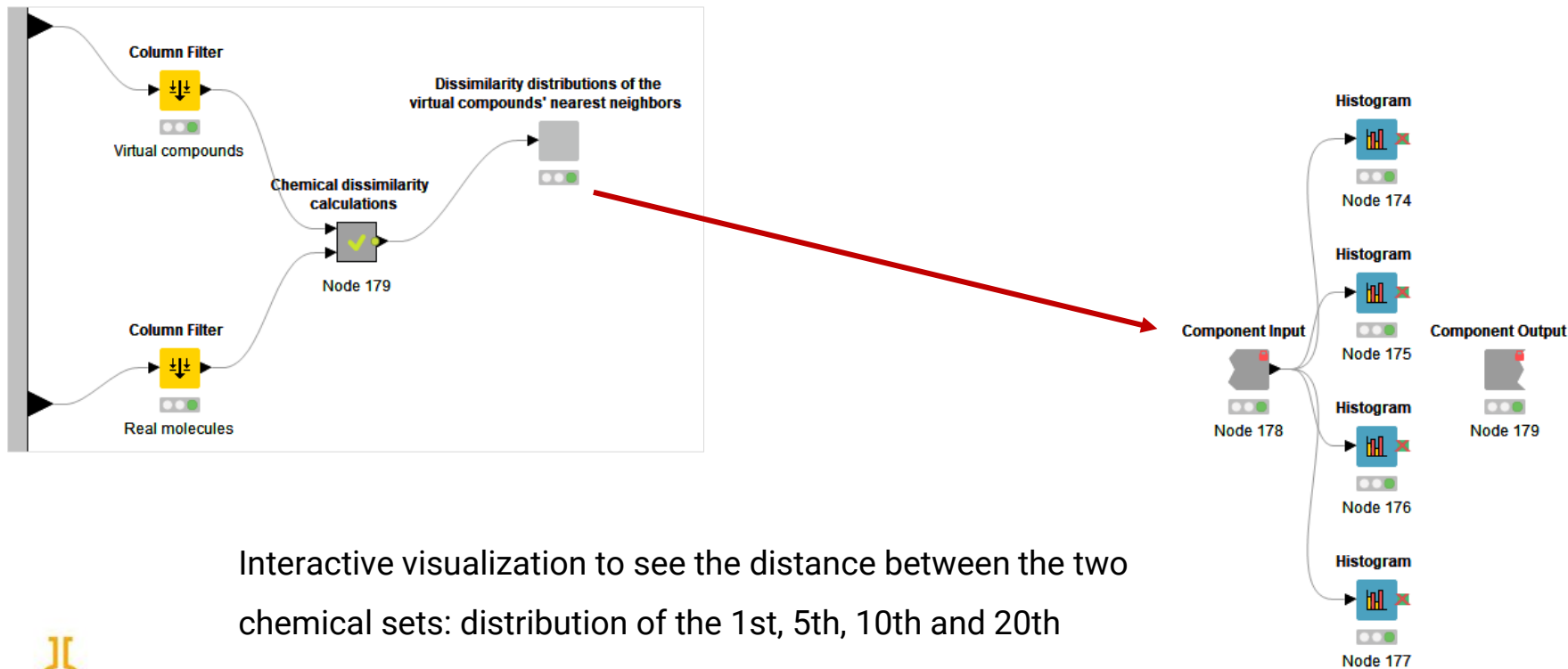
Step +1 – Overlap between the virtual and the real space



- How large is the dissimilarity between the virtual and the „real” molecules?
- Could we explore a new part of the chemical space?



Step +1 – Overlap between the virtual and the real space



Interactive visualization to see the distance between the two chemical sets: distribution of the 1st, 5th, 10th and 20th nearest neighbor dissimilarities



Thank you.

Dóra Barna, Norbert Sas

dbarna@chemaxon.com

nsas@chemaxon.com

