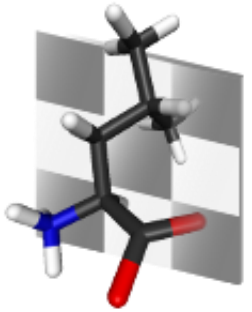
CheS-Mapper: New Developments

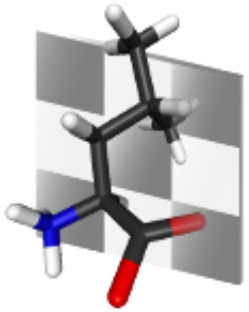


Martin Gütlein
mguetlein@fdm.uni-freiburg.de

Andreas Karwath & Stefan Kramer

OpenTox Euro 2013, October 2

CheS-Mapper - Chemical Space Mapping and Visualization in 3D



- 3D viewer for small molecule datasets
- Published in Journal of Cheminformatics, March 2012, >6000 accesses in 18 months
- Project homepage: http://ches-mapper.org
- Open-source java software
- Uses: Jmol, CDK, WEKA, OpenBabel, R
- Compatible to OpenTox dataset services

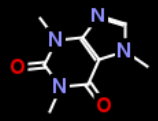
demo.smi - CheS-Mapper (v1.12.1)

File Edit View Highlighting Help

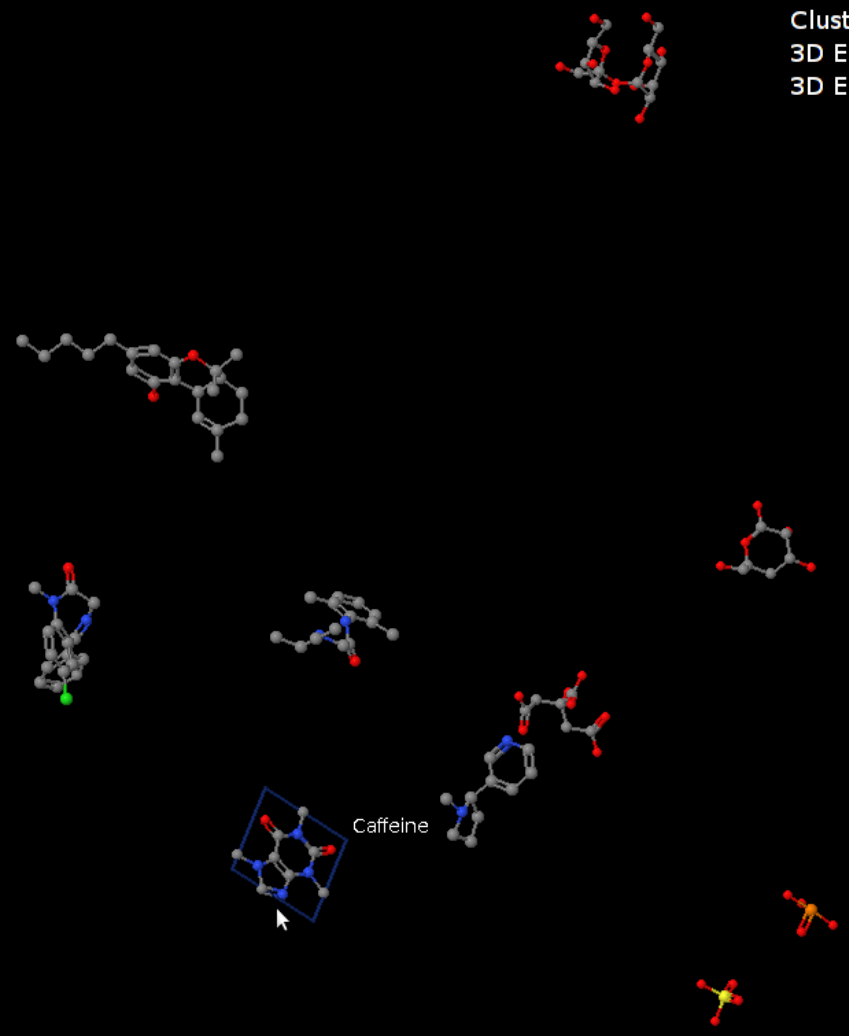
- Caffeine
- Citric acid
- Diazepam
- Glucose
- Lidocaine
- Nicotine
- Phosphoric acid
- Sucrose
- Sulfuric acid
- THC

☐ Superimpose

Dataset:	demo.smi
Num compounds:	10
Cluster algorithm:	No Dataset Clustering
3D Embedding:	Sammon 3D Embedder (R)
3D Embedding Quality:	excellent (CCC: 1, r^2: 0.99)



Compound	Caffeine
Smiles	Cn1cnc2c1c...
abonds	*10*
dbonds	*2*
sbonds	*3*
HBA2	*6*
HBD	*0*
SMIdbNAME	*Caffeine*
logP	*-1.03*
bonds	*15*
MR	*52.04*
atoms	*14*
HBA1	*3*
MW	*194.19*
nF	*0*
tbonds	*0*
TPSA	*61.82*



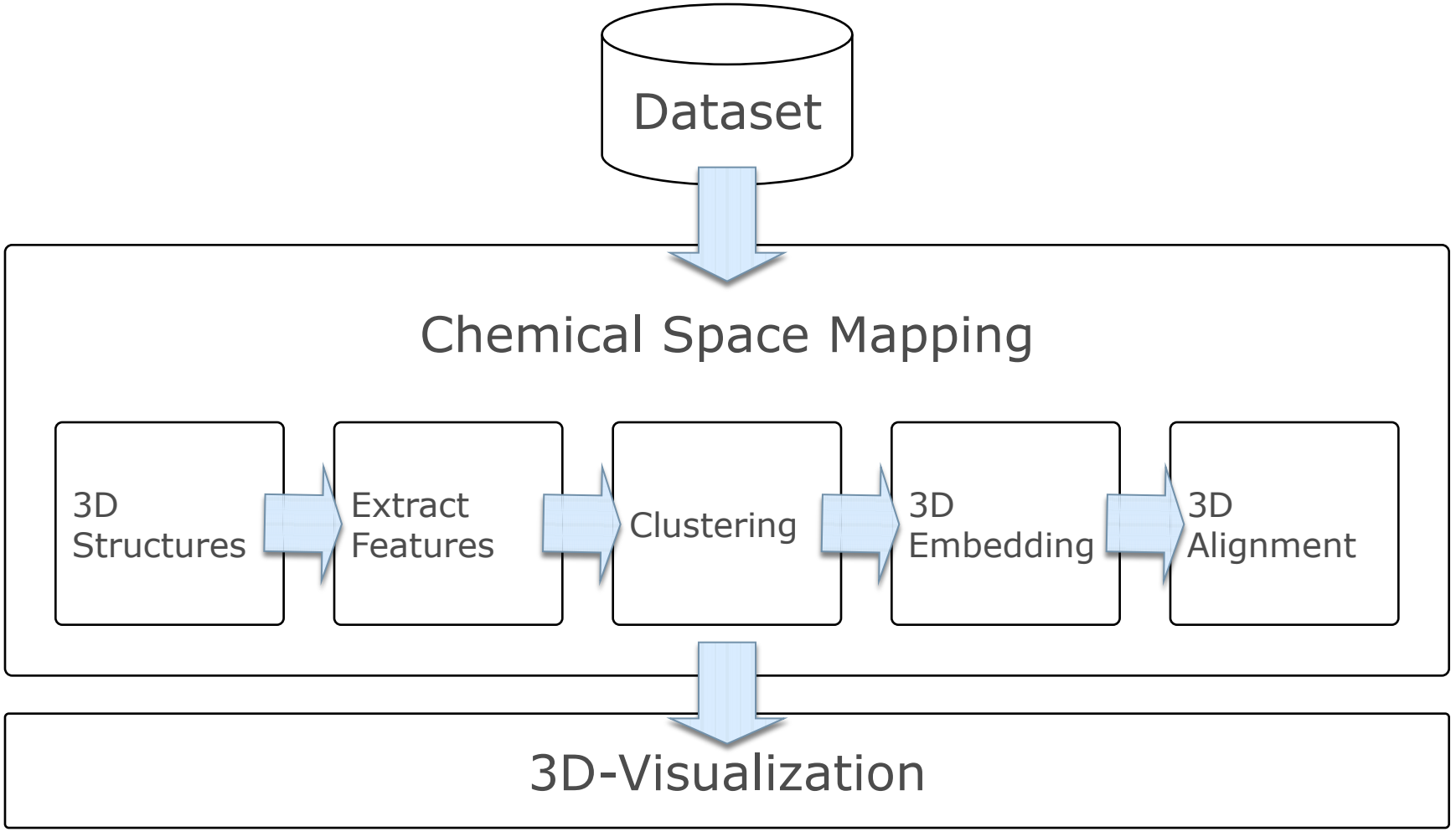
- + ○ Wireframe ● Balls & Sticks ● Dots

Feature: None (show atom types)

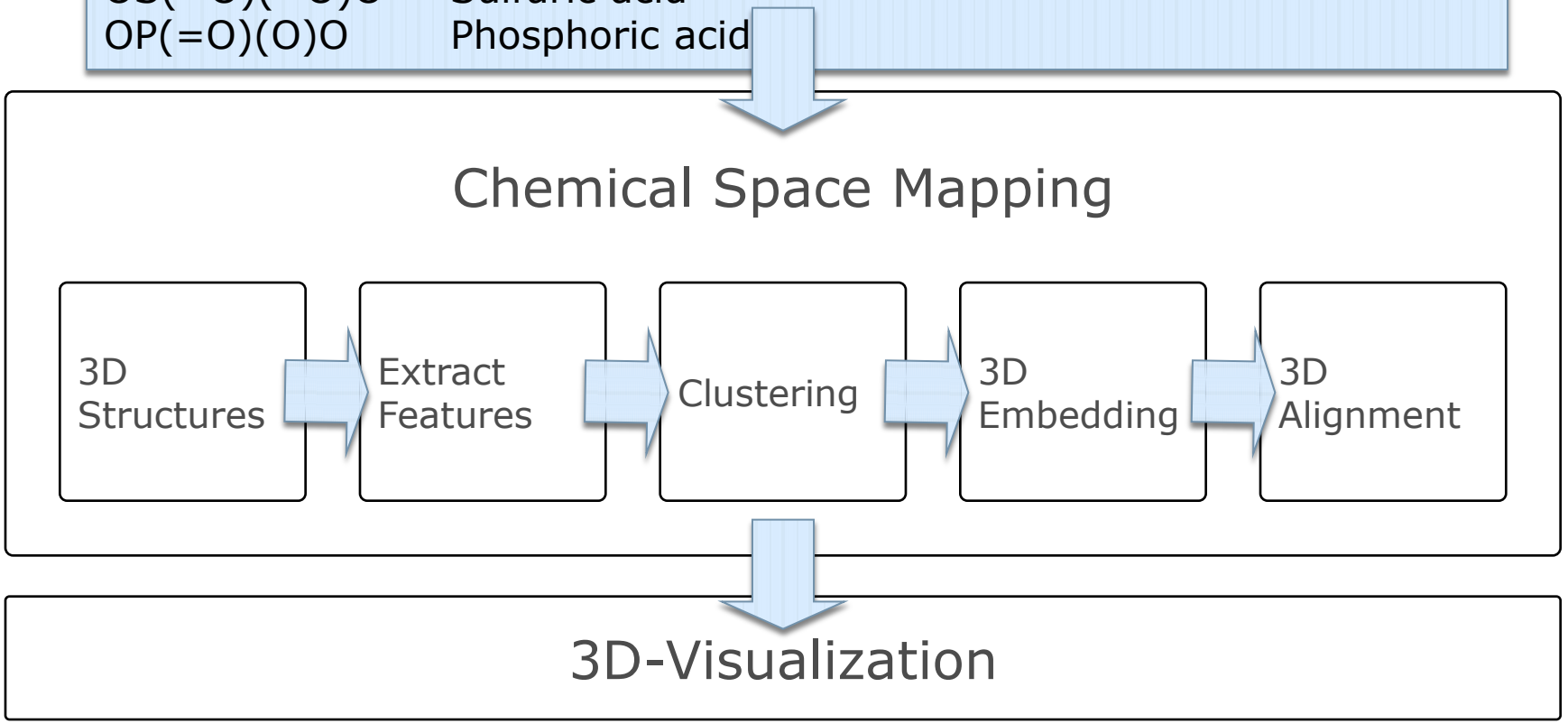
Artifical test dataset with 10 compounds

Jmol

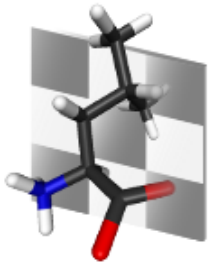
CheS-Mapper Workflow



Cn1cnc2c1c(=O)n(c(=O)n2C)C Caffeine
CCCCCc1cc2c(C3C=C(CCC3C(O2)(C)C)C)c(c1)O THC
O=C(Nc1c(cccc1C)C)CN(CC)CC Lidocaine
CN1CCCC1c1cccnc1 Nicotine
CN1c2c(C(=NCC1=O)c1ccccc1)cc(Cl)cc2 Diazepam
O1C(CO)C(O)C(O)C(O)C1O[C@@]1(OC(C(O)C1O)CO)CO Sucrose
OCC1OC(O)C(O)C(O)C1O Glucose
OC(=O)CC(O)(C(=O)O)CC(=O)O Citric acid
OS(=O)(=O)O Sulfuric acid
OP(=O)(O)O Phosphoric acid



CheS-Mapper Wizard (v1.12.1)



1. Load Dataset
2. Create 3D Structures
3. Extract Features
4. Cluster Dataset
5. Embed into 3D Space
6. Align Compounds

Load Dataset (step 1 of 6)

Select a dataset from your file system for clustering, embedding and visualization.

Select dataset file (Copy a http link into the textfield to load a dataset from the internet):

/home/martin/data/presentation/demo.smi | Open file... | Search ChEMBL

Recently used datasets

Load Dataset

Dataset Properties:

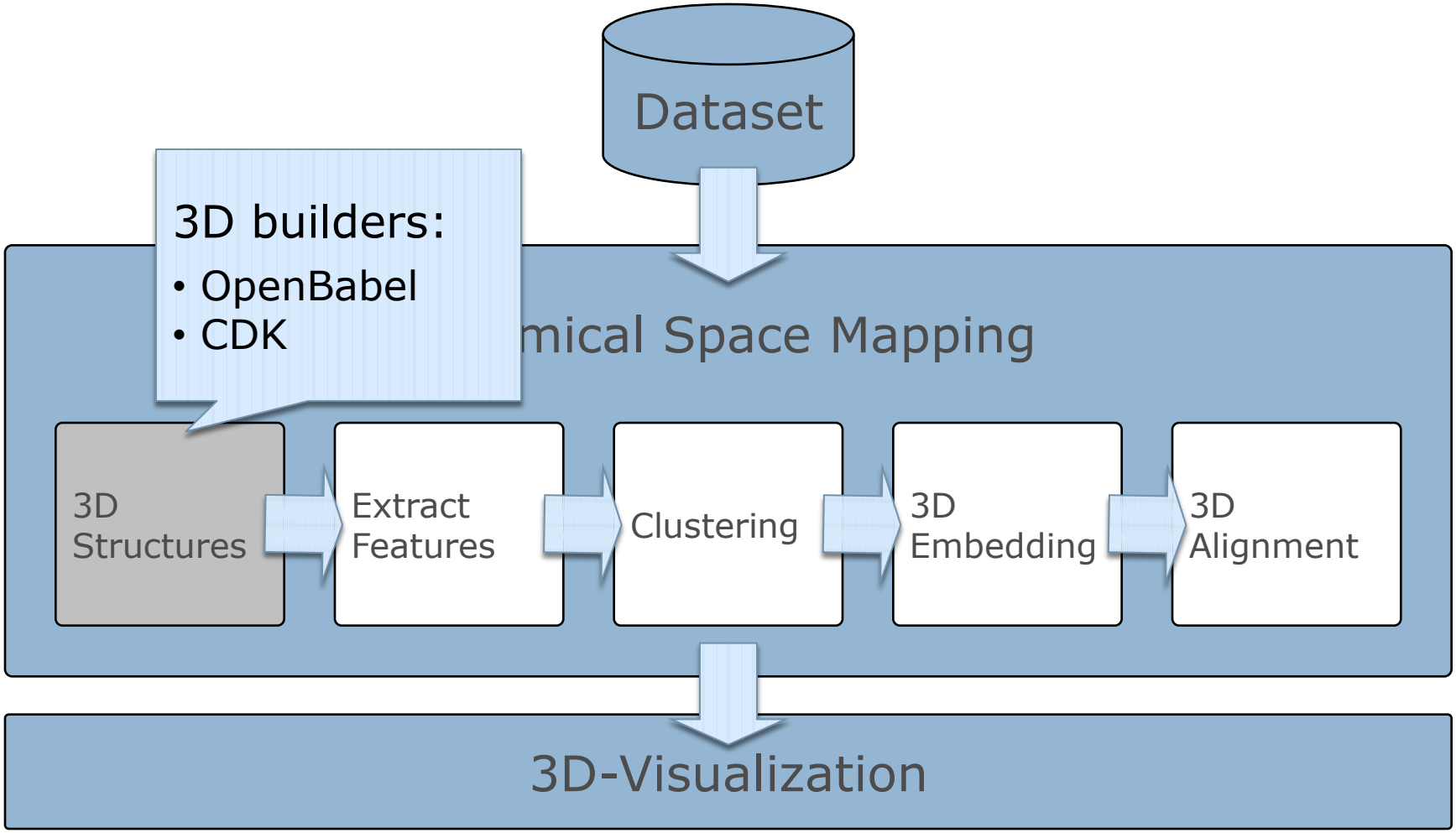
File:	demo.smi
Num compounds:	10
Num properties per compound:	2
3D available:	true

Chemical space mapping can be configured with wizard

OpenTox

Help | Import | Cancel | Previous | Next | Start mapping

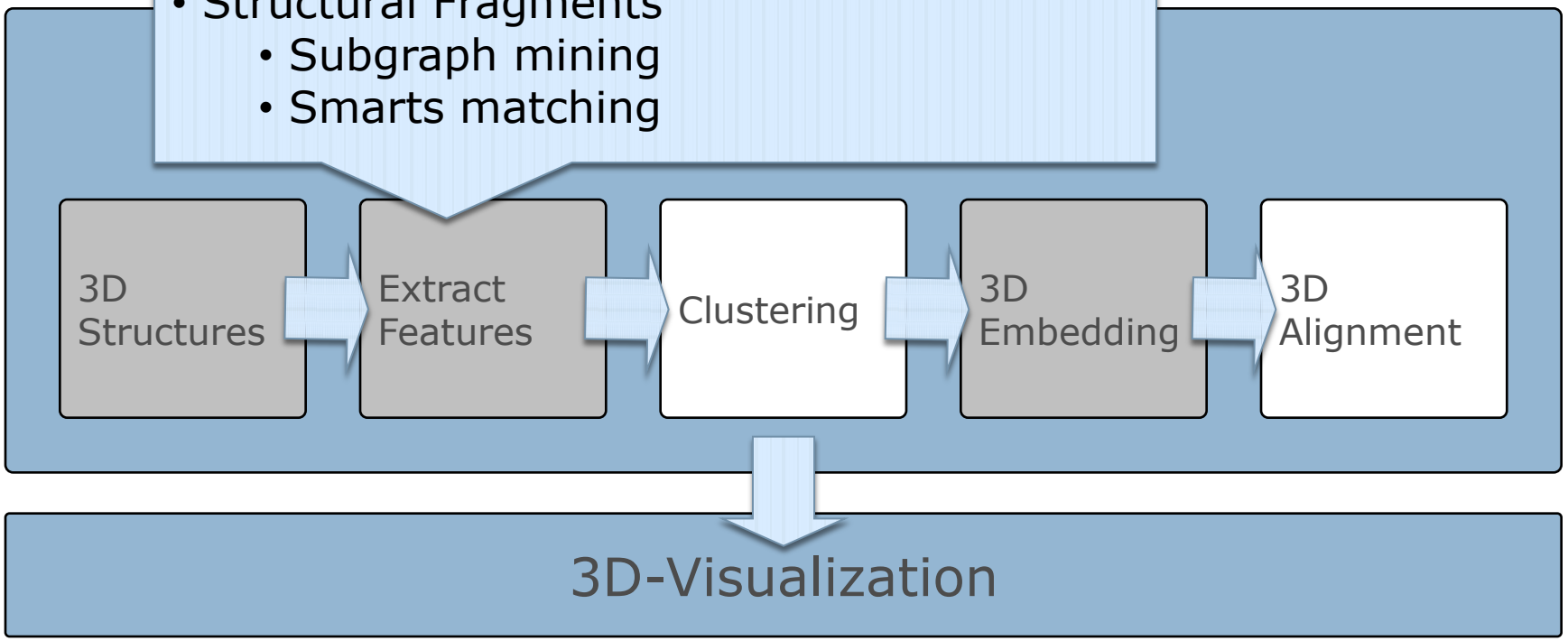
CheS-Mapper Workflow



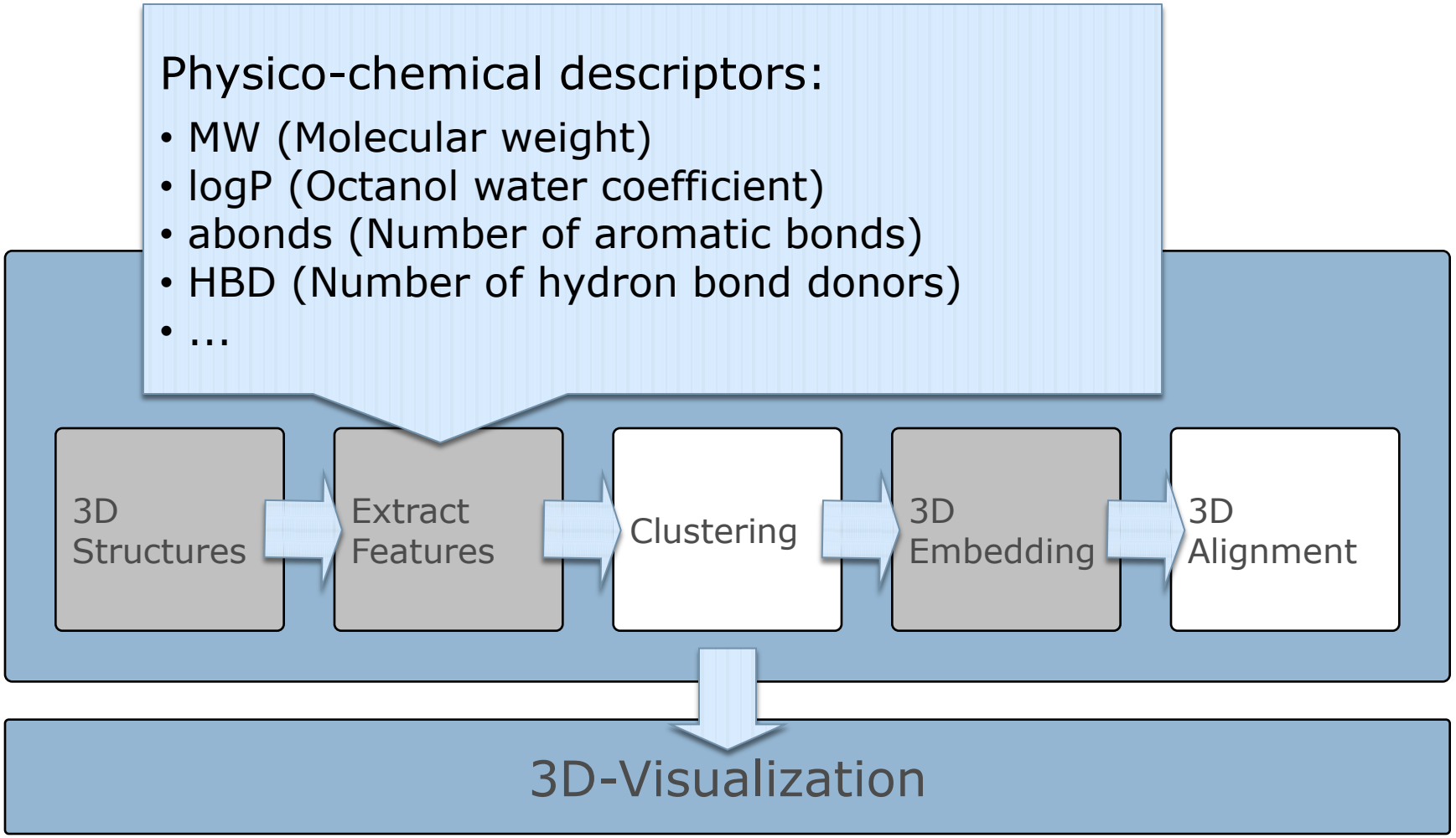
CheS-Mapper Workflow

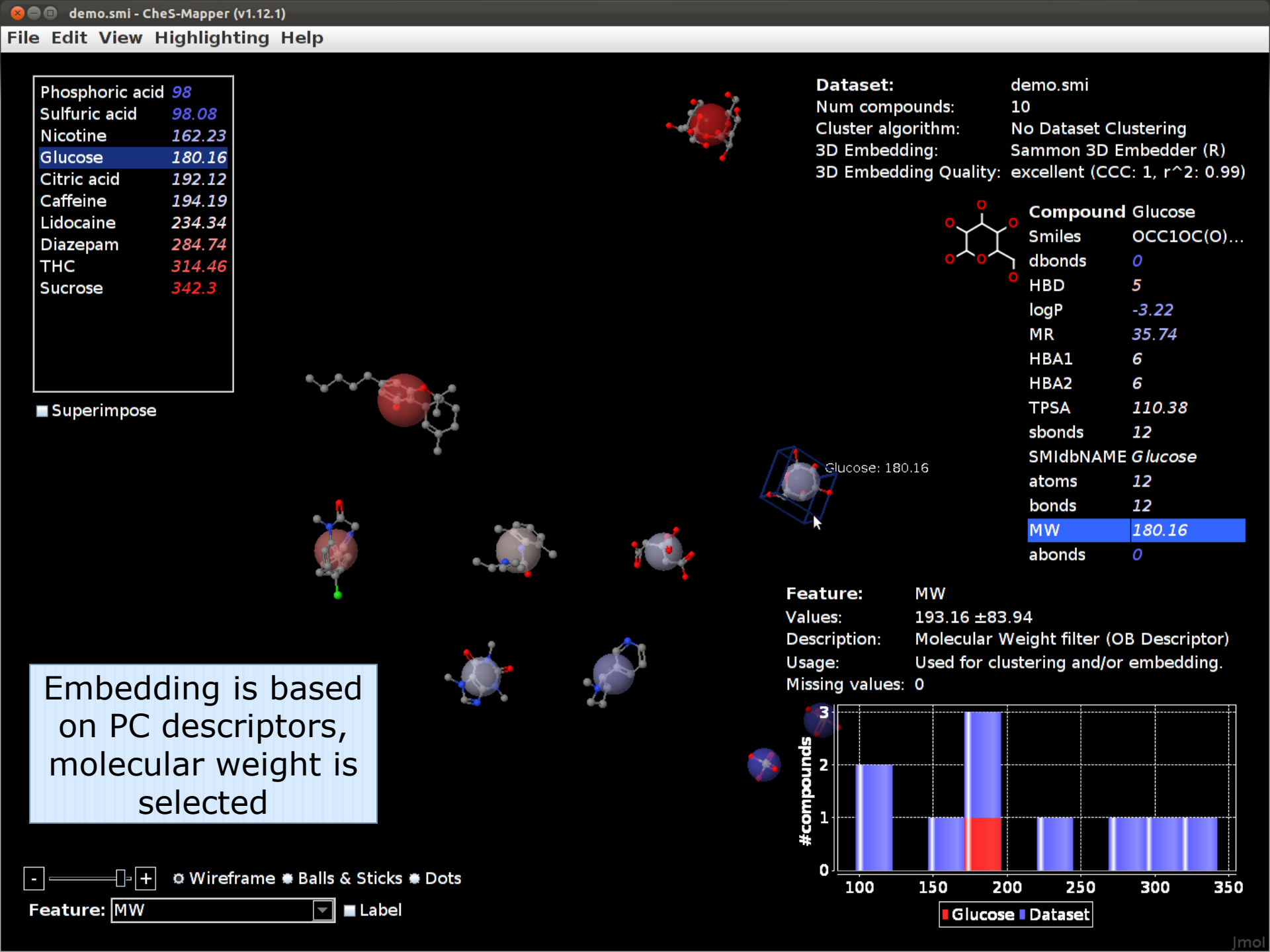
Feature Extraction:

- Included in the dataset
- PC-Descriptors (computed with CDK or OpenBabel)
- Structural Fragments
  - Subgraph mining
  - Smarts matching



CheS-Mapper Workflow

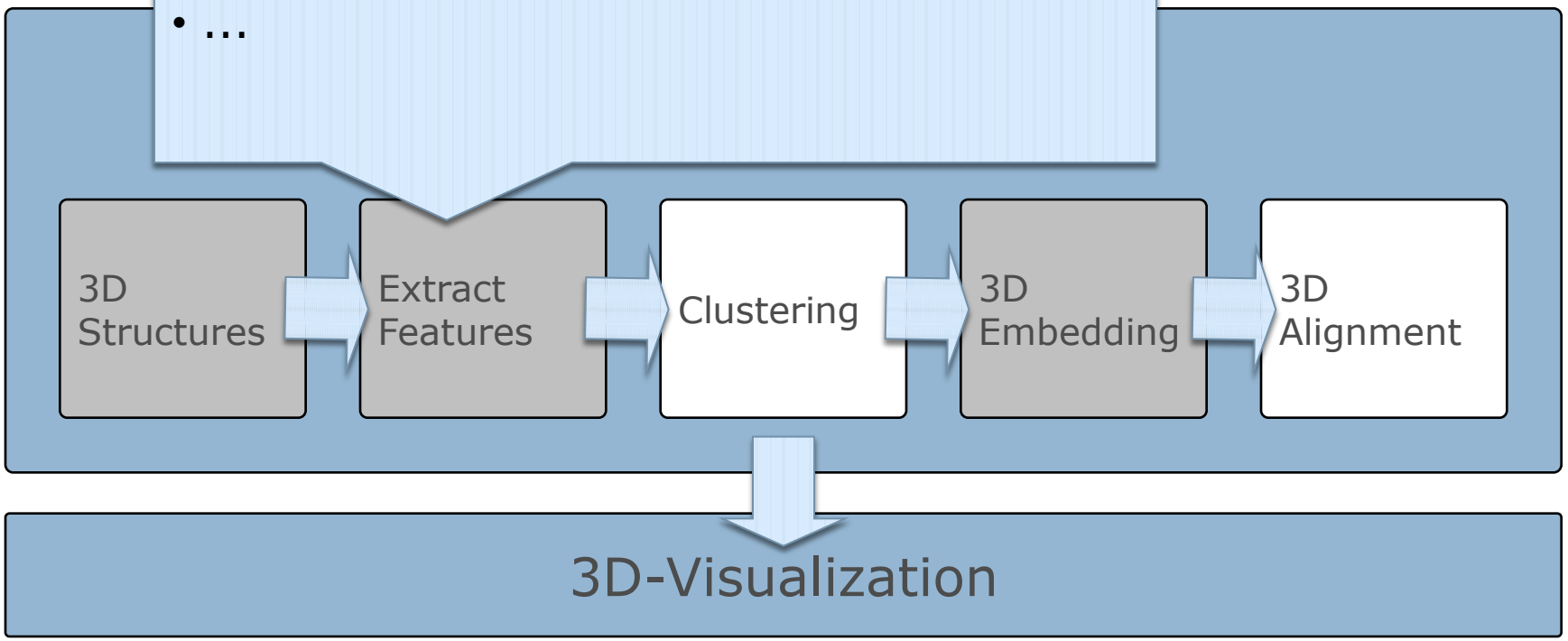


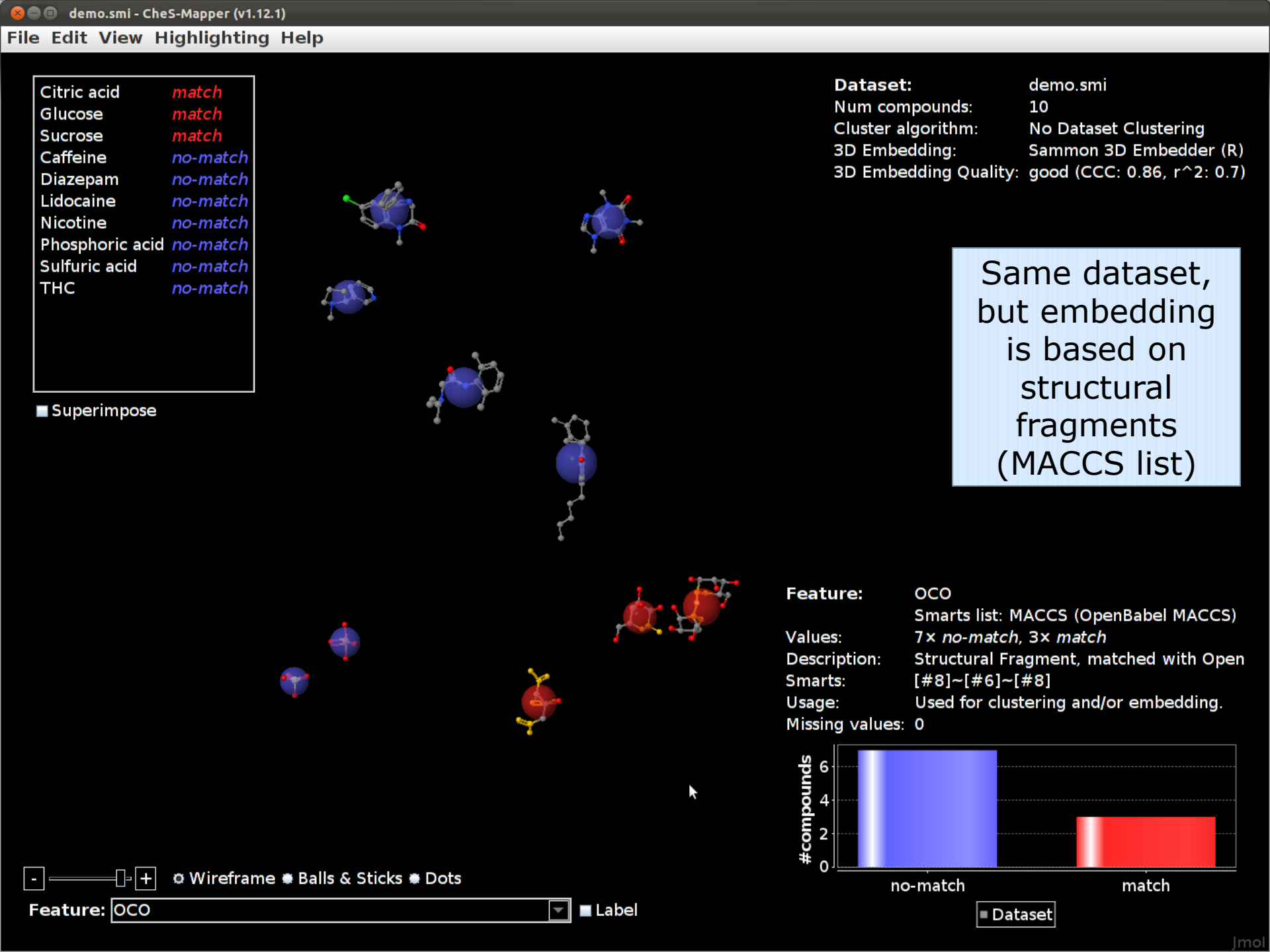

demo.smi - CheS-Mapper (v1.12.1)
File Edit View Highlighting Help
Phosphoric acid 98
Sulfuric acid 98.08
Nicotine 162.23
Glucose 180.16
Citric acid 192.12
Caffeine 194.19
Lidocaine 234.34
Diazepam 284.74
THC 314.46
Sucrose 342.3
Superimpose
Dataset: demo.smi
Num compounds: 10
Cluster algorithm: No Dataset Clustering
3D Embedding: Sammon 3D Embedder (R)
3D Embedding Quality: excellent (CCC: 1, r^2: 0.99)
Compound Glucose
Smiles OCC1OC(O)...
dbonds 0
HBD 5
logP -3.22
MR 35.74
HBA1 6
HBA2 6
TPSA 110.38
sbonds 12
SMIdbNAME Glucose
atoms 12
bonds 12
MW 180.16
abonds 0
Glucose: 180.16
Feature: MW
Values: 193.16 ±83.94
Description: Molecular Weight filter (OB Descriptor)
Usage: Used for clustering and/or embedding.
Missing values: 0
Embedding is based on PC descriptors, molecular weight is selected
#compounds
0
1
2
3
100
150
200
250
300
350
Glucose Dataset
Wireframe Balls & Sticks Dots
Feature: MW
Label
Jmol

CheS-Mapper Workflow

MACCS list (166 SMARTS fragments):

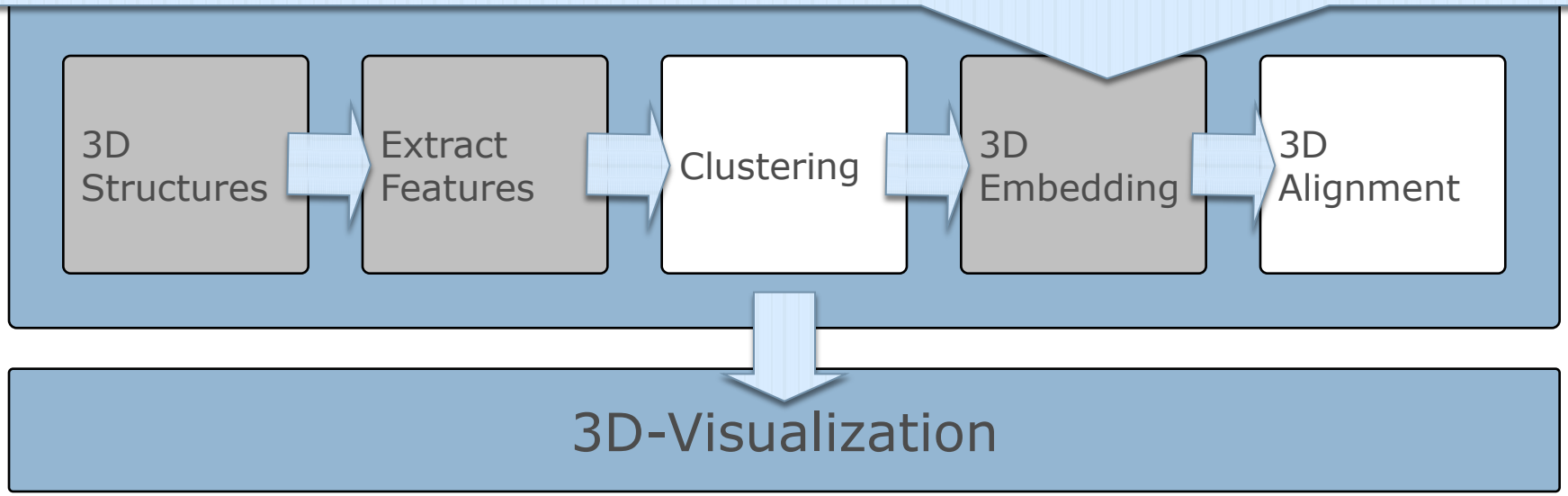
- P
- [#8]~[#6]~[#8]
- *~[CH2]~[#8]
- ...



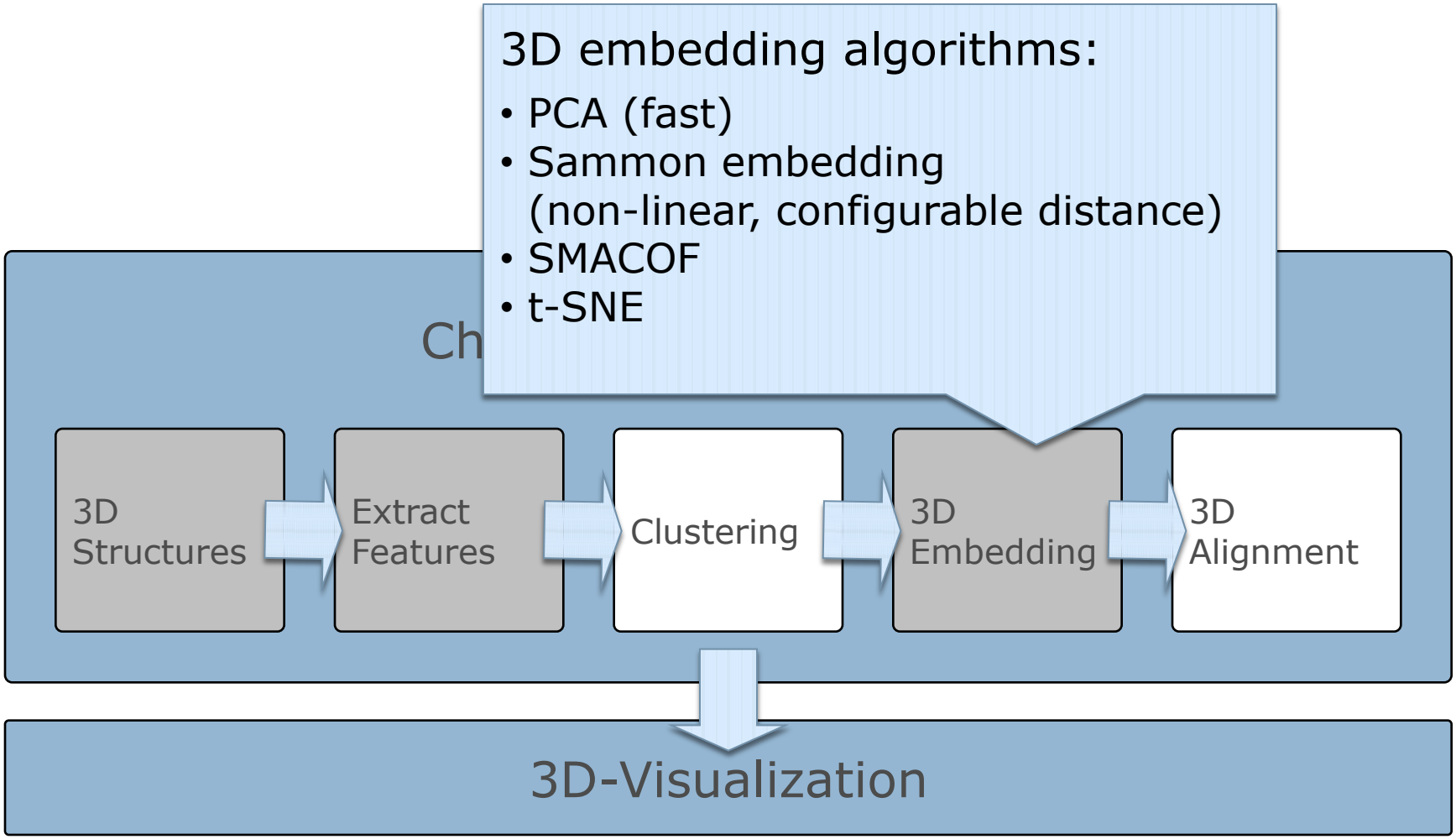

demo.smi - CheS-Mapper (v1.12.1)
File Edit View Highlighting Help
Citric acid match
Glucose match
Sucrose match
Caffeine no-match
Diazepam no-match
Lidocaine no-match
Nicotine no-match
Phosphoric acid no-match
Sulfuric acid no-match
THC no-match
Superimpose
Dataset: demo.smi
Num compounds: 10
Cluster algorithm: No Dataset Clustering
3D Embedding: Sammon 3D Embedder (R)
3D Embedding Quality: good (CCC: 0.86, r^2: 0.7)
Same dataset, but embedding is based on structural fragments (MACCS list)
Feature: OCO
Smarts list: MACCS (OpenBabel MACCS)
Values: 7× no-match, 3× match
Description: Structural Fragment, matched with Open
Smarts: [#8]~[#6]~[#8]
Usage: Used for clustering and/or embedding.
Missing values: 0
#compounds
6
4
2
0
no-match
match
Dataset
- + Wireframe Balls & Sticks Dots
Feature: OCO
Label
Jmol

	MW	LogP	abonds	...	HBD
Caffeine					
THC					
Glucose					
Sulfuric acid					
...					

	X	Y	Z
Caffeine			
THC			
Glucose			
Sulfuric acid			
...			



CheS-Mapper Workflow



Quality of the 3D Embedding

	MW	LogP	abonds	...	HBD
Caffeine					
THC					
Glucose					
Sulfuric acid					
...					

	X	Y	Z
Caffeine			
THC			
Glucose			
Sulfuric acid			
...			

How well do the 3D positiones reflect the feature values?

Compound distance matrix based on feature values

Compound distance matrix based on 3D positions

Global embedding quality:
Correlation between distance matrixes (CCC, R^2)

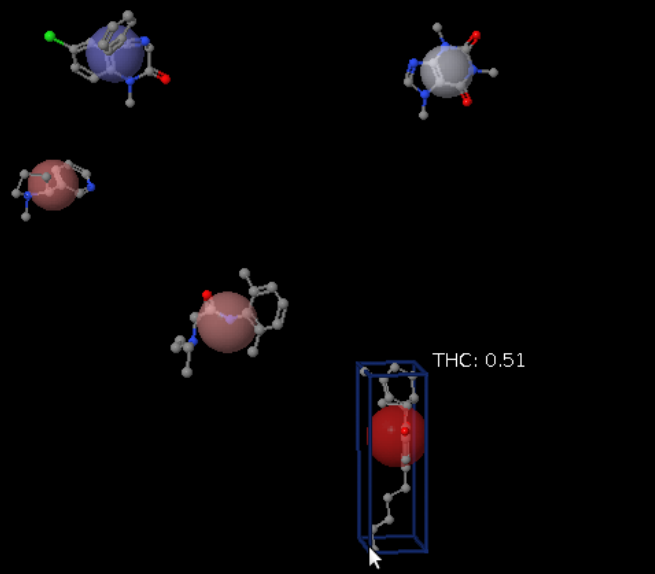
Embedding stress for each compound:
1 – correlation of the correspoding matrix rows

demo.smi - CheS-Mapper (v1.12.1)

File Edit View Highlighting Help

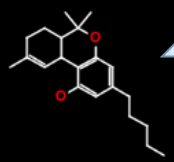
Citric acid	0.05
Glucose	0.07
Sucrose	0.08
Sulfuric acid	0.1
Phosphoric acid	0.11
Diazepam	0.12
Caffeine	0.26
Lidocaine	0.38
Nicotine	0.4
THC	0.51

☐ Superimpose



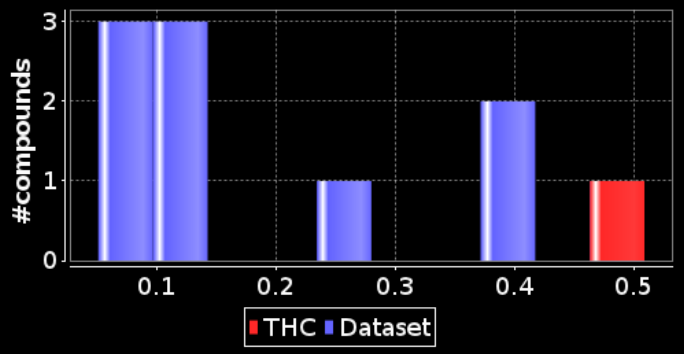
THC is the compound with the highest embedding stress

Dataset: demo.smi
Num compounds: 10
Cluster algorithm: No Dataset Clustering
3D Embedding: Sammon 3D Embedder (R)
3D Embedding Quality: good (CCC: 0.86, r^2: 0.7)



Compound	THC
Smiles	CCCCCc1cc...
C(A)(A)C	match
C(A)(A)CH3	match
C=C(A)A	match
C=C	match
CC(C)(C)A	match
CH3AAACH2A	match
Onot%A%A	match
CH3CH2A	match
CH3ACH2A	match
ACH2CH2A ...	match
ACH2AACH2A	match
6M ring > 1	match
ACH2CH2A	match

Feature: Embedding stress
Values: 0.11 ±0.17
Description: 0 := perfectly embedded (no stress), 1: h
Usage: NOT used for clustering and/or embeddin
Missing values: 0

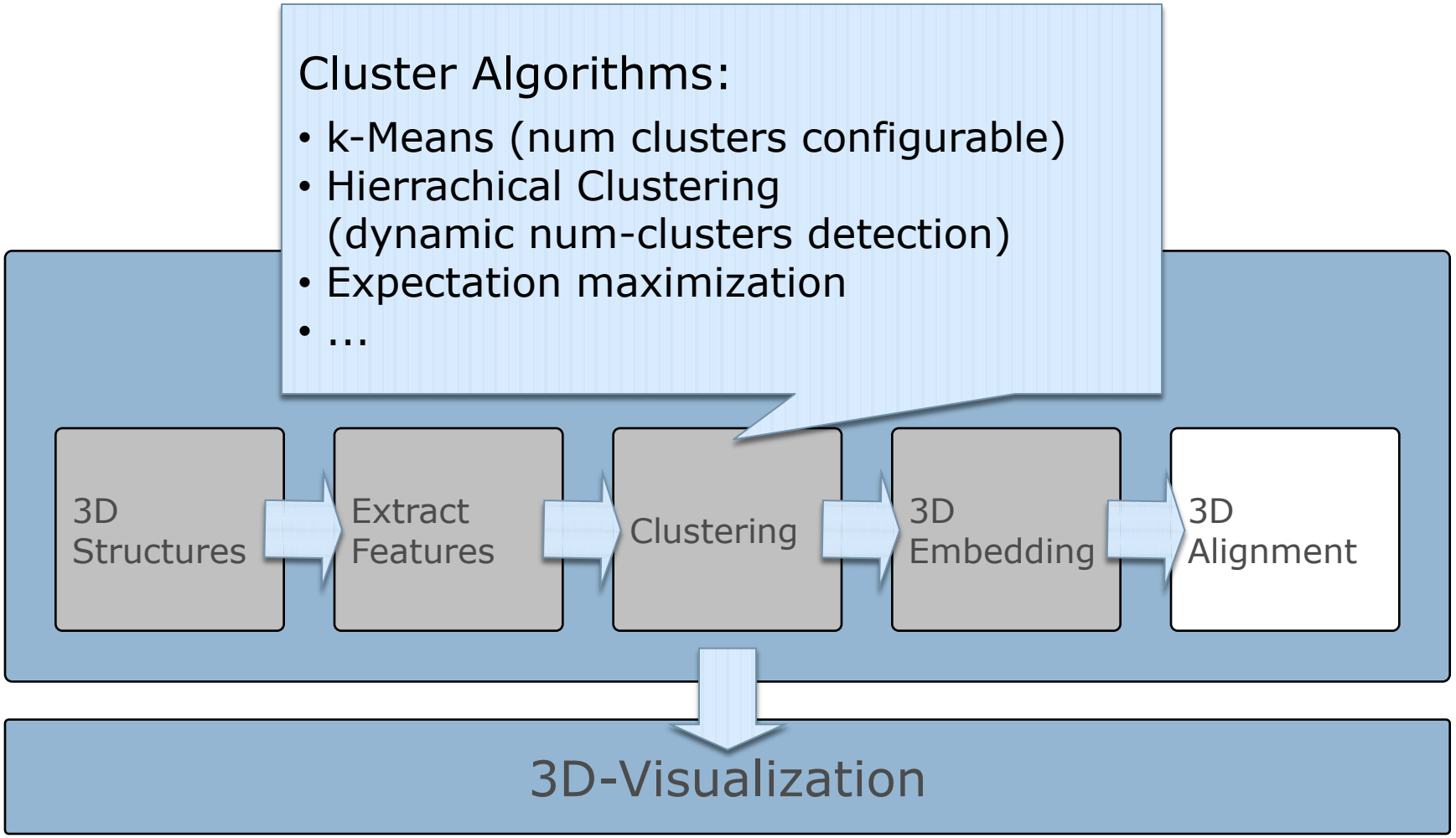


- + ○ Wireframe ● Balls & Sticks ● Dots

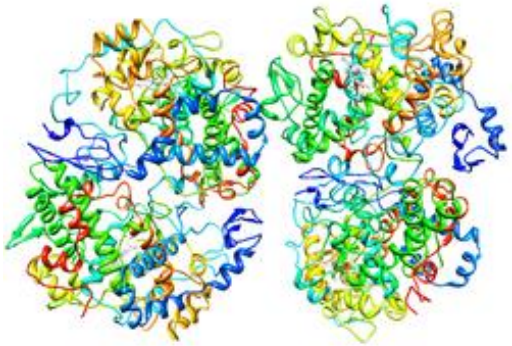
Feature: Embedding stress ☐ Label

Jmol

CheS-Mapper Workflow

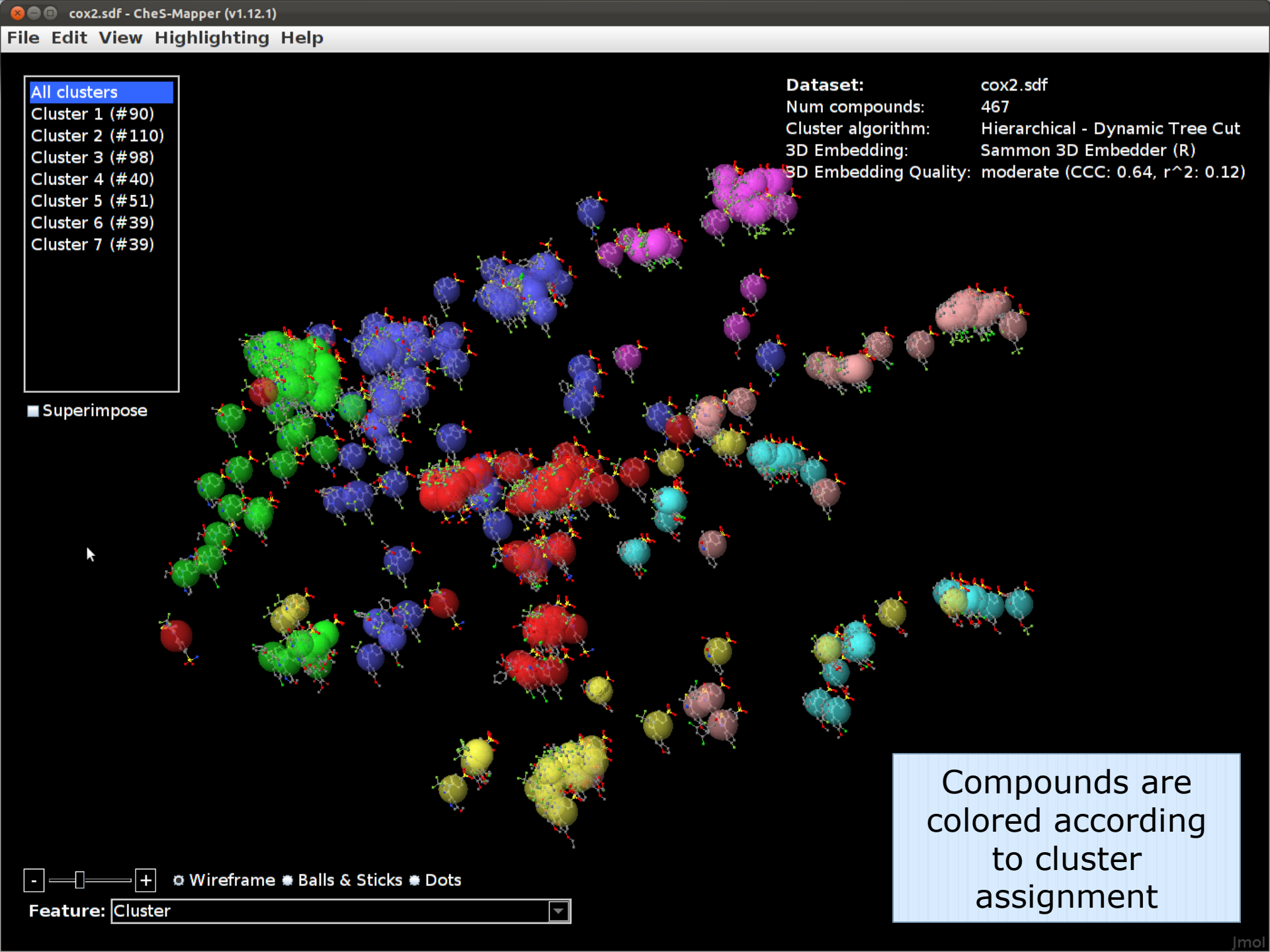


COX-2 dataset

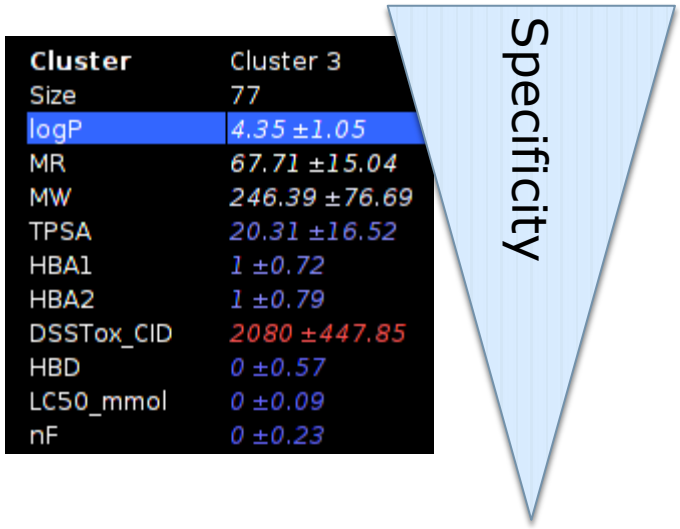


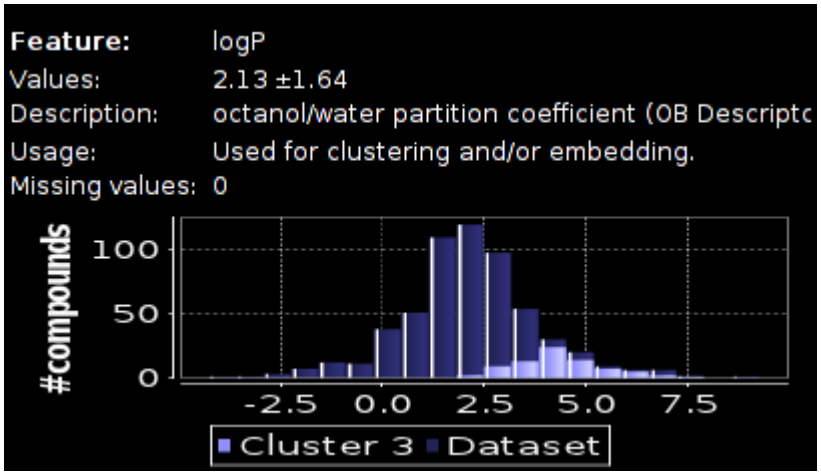
- 467 COX-2 Inhibitors
- Inhibitation of the enzyme Cyclooxygenase-2 (COX-2) is investigated in cancer studies
- Compounds are structurally very similar (docking dataset)
- Endpoint: IC_{50} µMol (half maximal inhibitory concentration)

Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure−Activity Relationships
Jeffrey J. Sutherland, Lee A. O'Brien, and, and Donald F. Weaver
Journal of Chemical Information and Computer Sciences 2003 43 (6), 1906-1915

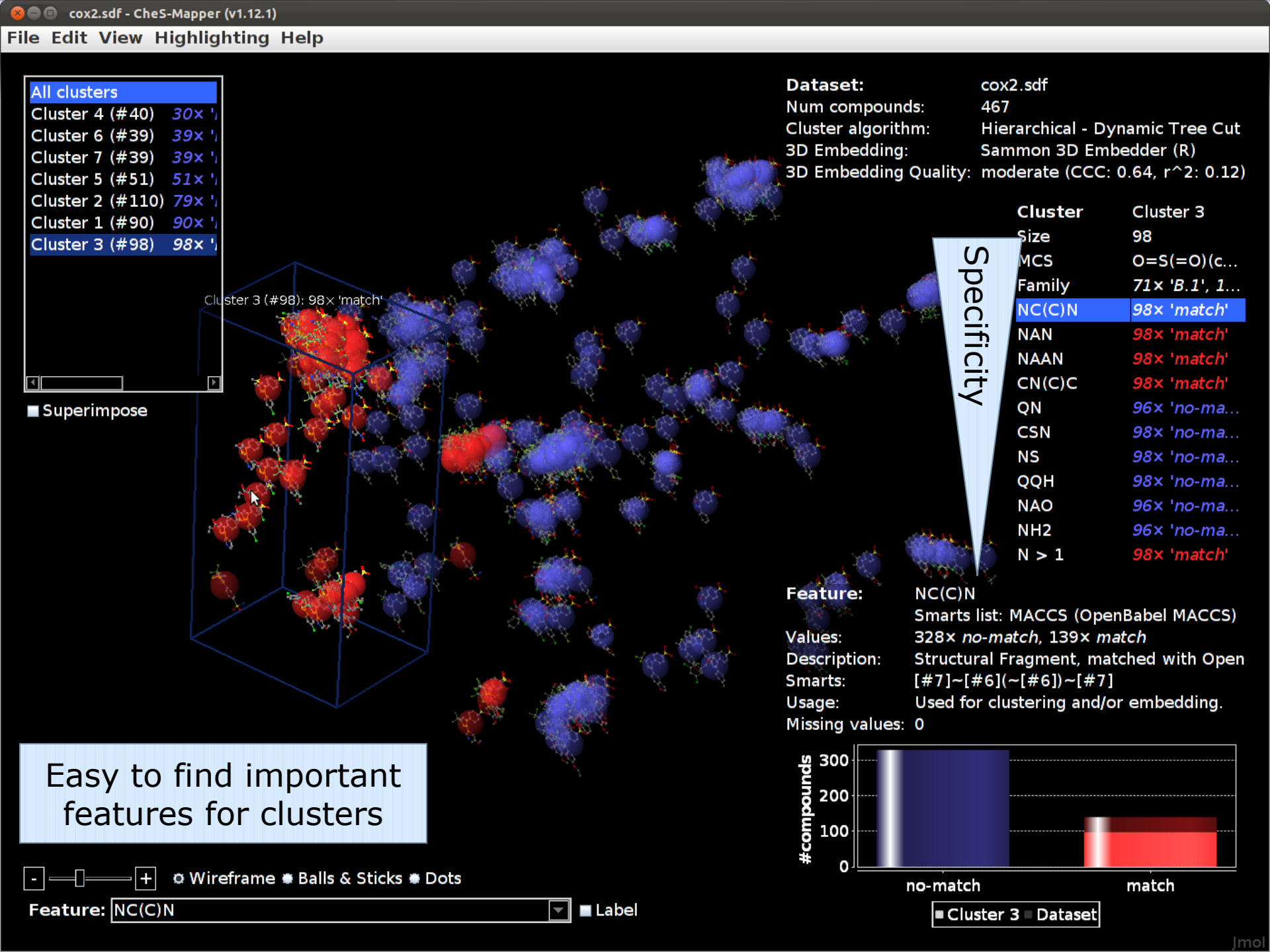

cox2.sdf - CheS-Mapper (v1.12.1)
File Edit View Highlighting Help
All clusters
Cluster 1 (#90)
Cluster 2 (#110)
Cluster 3 (#98)
Cluster 4 (#40)
Cluster 5 (#51)
Cluster 6 (#39)
Cluster 7 (#39)
Superimpose
Dataset: cox2.sdf
Num compounds: 467
Cluster algorithm: Hierarchical - Dynamic Tree Cut
3D Embedding: Sammon 3D Embedder (R)
3D Embedding Quality: moderate (CCC: 0.64, r^2: 0.12)
Wireframe Balls & Sticks Dots
Feature: Cluster
Compounds are colored according to cluster assignment
Jmol

Sorting of features



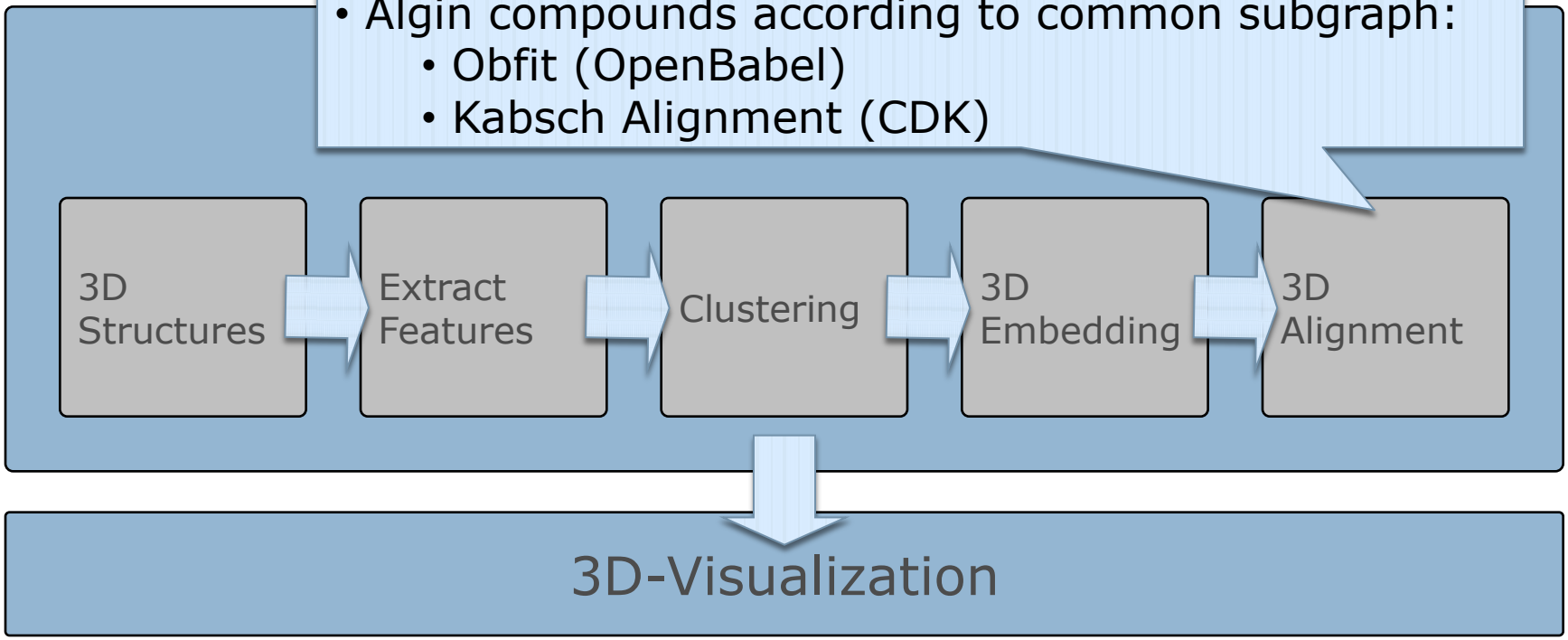


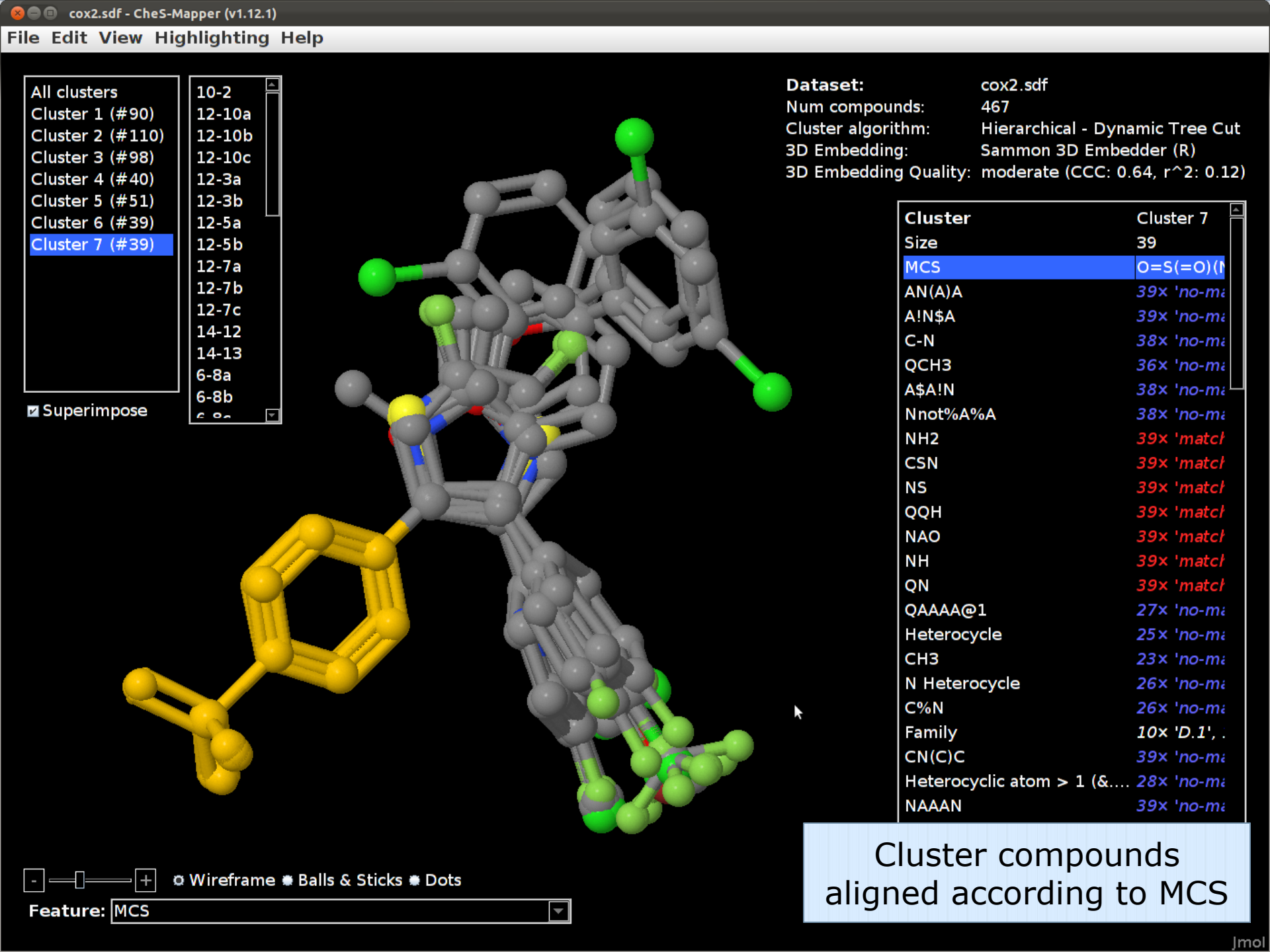
- Feature values are shown for the selected compound or cluster
- Features are sorted according to the p-value of
  - X^2 test for nominal features
  - ANOVA test for numerical features
- The most ‚important' features are listed first:
  the feature values of this compound/cluster differ the most from the complete dataset

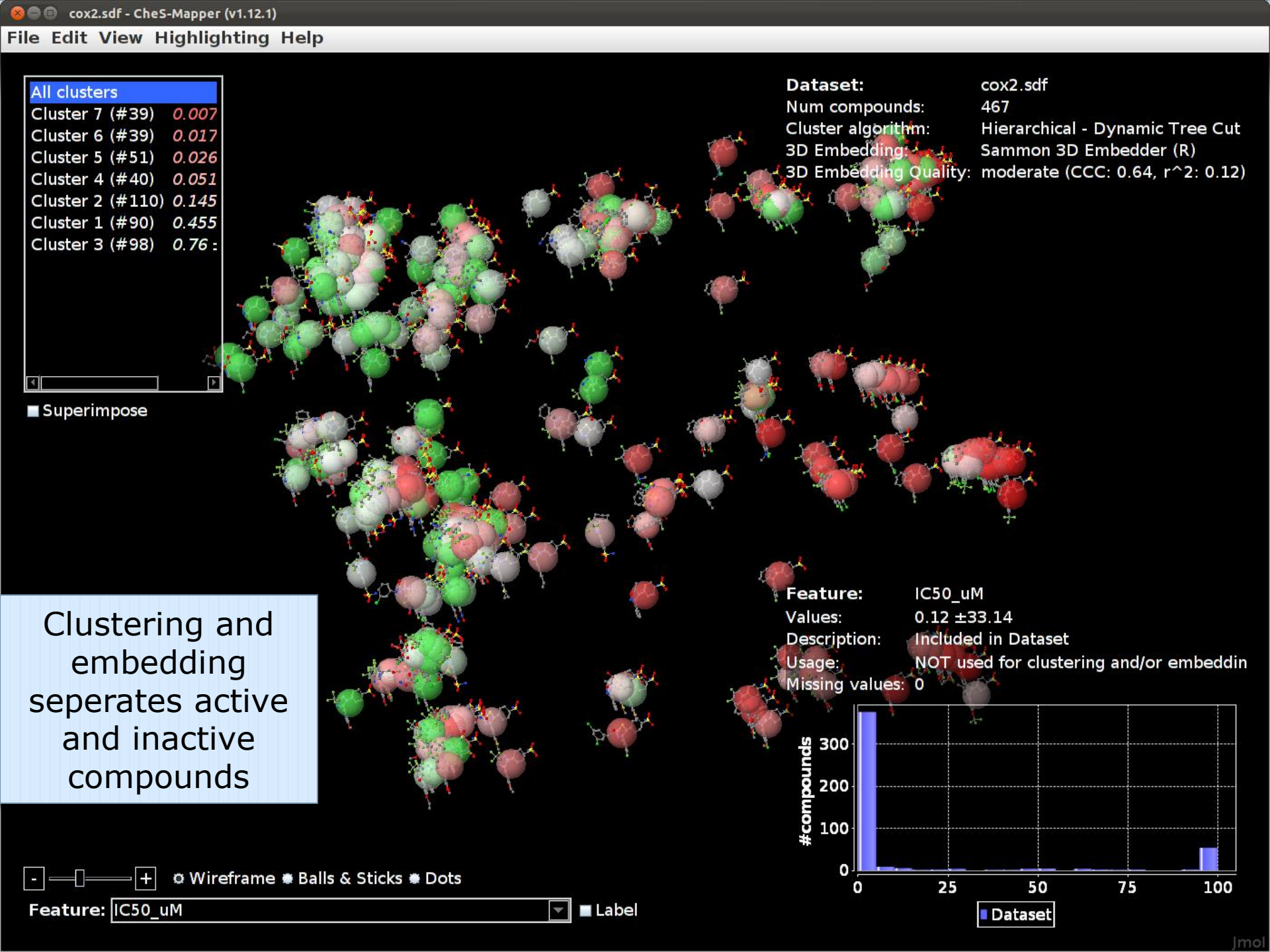

cox2.sdf - CheS-Mapper (v1.12.1)
File Edit View Highlighting Help
All clusters
Cluster 4 (#40) 30× '
Cluster 6 (#39) 39× '
Cluster 7 (#39) 39× '
Cluster 5 (#51) 51× '
Cluster 2 (#110) 79× '
Cluster 1 (#90) 90× '
Cluster 3 (#98) 98× '
Superimpose
Cluster 3 (#98): 98× 'match'
Dataset: cox2.sdf
Num compounds: 467
Cluster algorithm: Hierarchical - Dynamic Tree Cut
3D Embedding: Sammon 3D Embedder (R)
3D Embedding Quality: moderate (CCC: 0.64, r^2: 0.12)
Specificity
Cluster Cluster 3
Size 98
MCS O=S(=O)(c...
Family 71× 'B.1', 1...
NC(C)N 98× 'match'
NAN 98× 'match'
NAAN 98× 'match'
CN(C)C 98× 'match'
QN 96× 'no-ma...
CSN 98× 'no-ma...
NS 98× 'no-ma...
QQH 98× 'no-ma...
NAO 96× 'no-ma...
NH2 96× 'no-ma...
N > 1 98× 'match'
Feature: NC(C)N
Smarts list: MACCS (OpenBabel MACCS)
Values: 328× no-match, 139× match
Description: Structural Fragment, matched with Open
Smarts: [#7]~[#6](~[#6])~[#7]
Usage: Used for clustering and/or embedding.
Missing values: 0
#compounds
300
200
100
0
no-match
match
Cluster 3 Dataset
Easy to find important features for clusters
Wireframe Balls & Sticks Dots
Feature: NC(C)N
Label
Jmol

CheS-Mapper Workflow

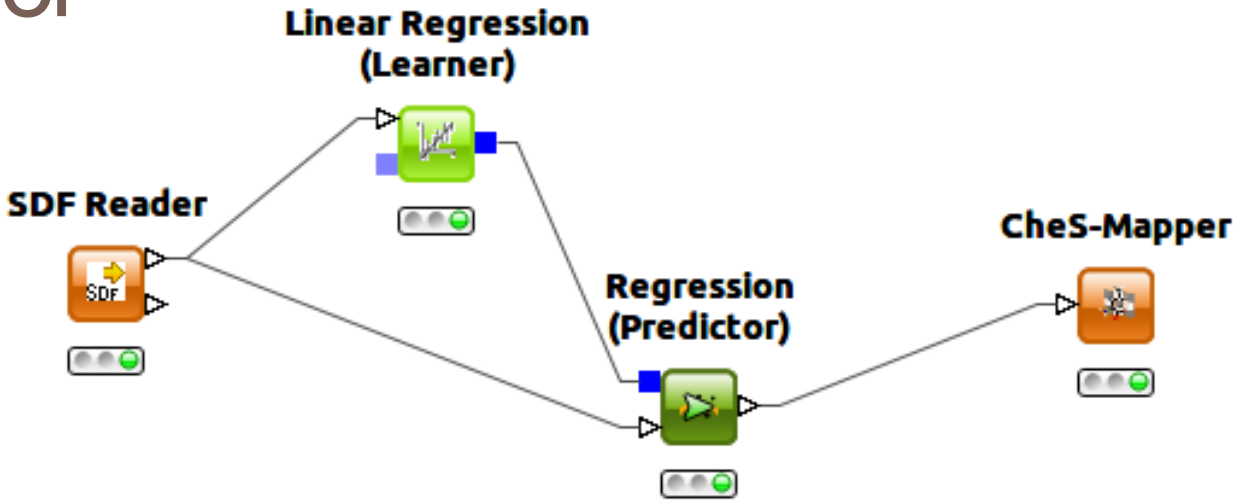
3D Alignment:
- Detect common subgraph of each cluster:
  - Compute MCS (maximum common subgraph)
  - Use largest fragment
  - Manually specifiy subgraph
- Algin compounds according to common subgraph:
  - Obfit (OpenBabel)
  - Kabsch Alignment (CDK)




cox2.sdf - CheS-Mapper (v1.12.1)
File Edit View Highlighting Help
All clusters
Cluster 1 (#90)
Cluster 2 (#110)
Cluster 3 (#98)
Cluster 4 (#40)
Cluster 5 (#51)
Cluster 6 (#39)
Cluster 7 (#39)
10-2
12-10a
12-10b
12-10c
12-3a
12-3b
12-5a
12-5b
12-7a
12-7b
12-7c
14-12
14-13
6-8a
6-8b
Superimpose
Dataset: cox2.sdf
Num compounds: 467
Cluster algorithm: Hierarchical - Dynamic Tree Cut
3D Embedding: Sammon 3D Embedder (R)
3D Embedding Quality: moderate (CCC: 0.64, r^2: 0.12)
Cluster Cluster 7
Size 39
MCS O=S(=O)(
AN(A)A 39× 'no-ma
A!N$A 39× 'no-ma
C-N 38× 'no-ma
QCH3 36× 'no-ma
A$A!N 38× 'no-ma
Nnot%A%A 38× 'no-ma
NH2 39× 'match
CSN 39× 'match
NS 39× 'match
QQH 39× 'match
NAO 39× 'match
NH 39× 'match
QN 39× 'match
QAAAA@1 27× 'no-ma
Heterocycle 25× 'no-ma
CH3 23× 'no-ma
N Heterocycle 26× 'no-ma
C%N 26× 'no-ma
Family 10× 'D.1', .
CN(C)C 39× 'no-ma
Heterocyclic atom > 1 (&.... 28× 'no-ma
NAAAN 39× 'no-ma
Wireframe Balls & Sticks Dots
Feature: MCS
Cluster compounds aligned according to MCS
Jmol


cox2.sdf - CheS-Mapper (v1.12.1)
File Edit View Highlighting Help
All clusters
Cluster 7 (#39) 0.007
Cluster 6 (#39) 0.017
Cluster 5 (#51) 0.026
Cluster 4 (#40) 0.051
Cluster 2 (#110) 0.145
Cluster 1 (#90) 0.455
Cluster 3 (#98) 0.76 :
Superimpose
Dataset: cox2.sdf
Num compounds: 467
Cluster algorithm: Hierarchical - Dynamic Tree Cut
3D Embedding: Sammon 3D Embedder (R)
3D Embedding Quality: moderate (CCC: 0.64, r^2: 0.12)
Clustering and embedding seperates active and inactive compounds
Feature: IC50_uM
Values: 0.12 ±33.14
Description: Included in Dataset
Usage: NOT used for clustering and/or embeddin
Missing values: 0
#compounds
300
200
100
0
0
25
50
75
100
Dataset
Wireframe Balls & Sticks Dots
Feature: IC50_uM
Label
Jmol

The CheS-Mapper extension for KNIME

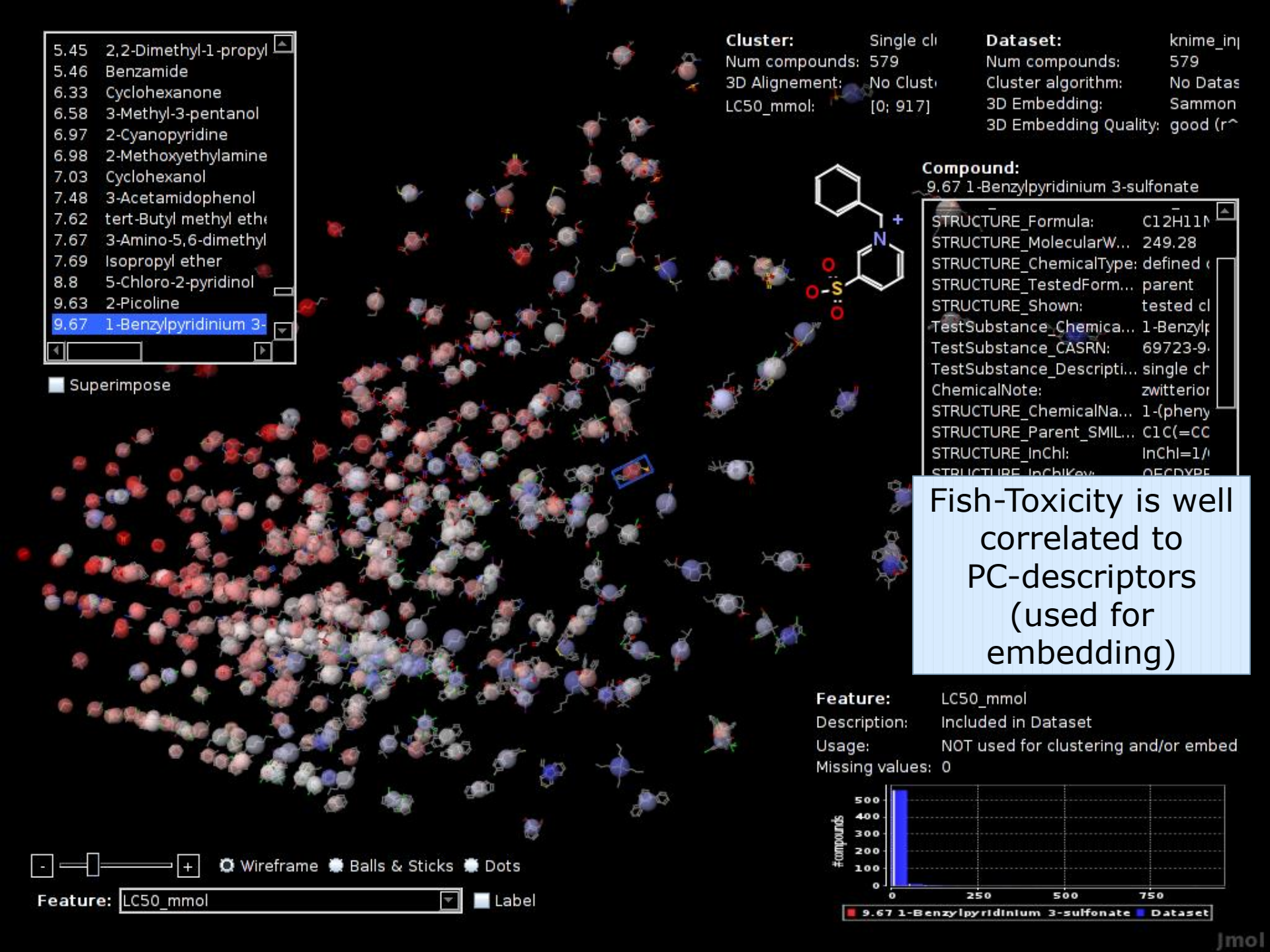


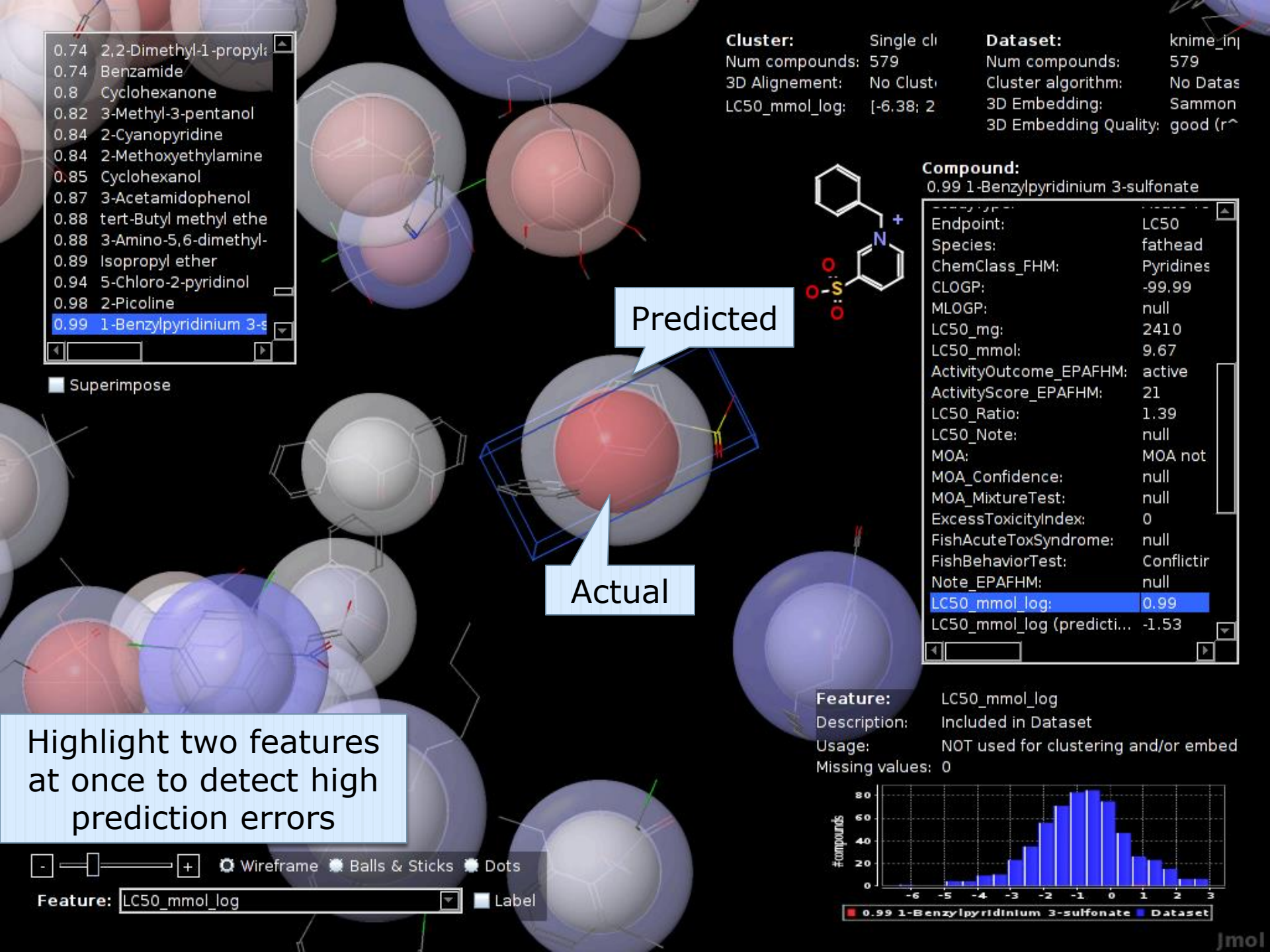
- KNIME:
  - graphical workbench for data access, investigation and predictive analysis
  - various extensions to process chemical data
- CheS-Mapper integration as visualization node

EPAFHM dataset

- EPAFHM:
  US Envrionmental Protection Agency (EPA)
  Fathead Minnow Acute Toxicity Database File
- 617 industrial organic chemicals
- Endpoint: LC50 mMol ((lethal) concentration that kills 50%)

Predicting modes of action from chemical structure: Acute toxicity in the fathed minnow (Pimephales promelas).
Russom, C.L., S.P. Bradbury, S.J. Broderius, D.E. Hammermeister, and R.A. Drummond (1997)
Environmental Toxicology and Chemistry, 16(5): 948-967.


5.45 2,2-Dimethyl-1-propyl
5.46 Benzamide
6.33 Cyclohexanone
6.58 3-Methyl-3-pentanol
6.97 2-Cyanopyridine
6.98 2-Methoxyethylamine
7.03 Cyclohexanol
7.48 3-Acetamidophenol
7.62 tert-Butyl methyl eth
7.67 3-Amino-5,6-dimethyl
7.69 Isopropyl ether
8.8 5-Chloro-2-pyridinol
9.63 2-Picoline
9.67 1-Benzylpyridinium 3-
Superimpose
Cluster: Single cl
Num compounds: 579
3D Alignement: No Clust
LC50_mmol: [0; 917]
Dataset: knime_in
Num compounds: 579
Cluster algorithm: No Datas
3D Embedding: Sammon
3D Embedding Quality: good (r^
Compound:
9.67 1-Benzylpyridinium 3-sulfonate
STRUCTURE_Formula: C12H11
STRUCTURE_MolecularW... 249.28
STRUCTURE_ChemicalType: defined
STRUCTURE_TestedForm... parent
STRUCTURE_Shown: tested cl
TestSubstance_Chemica... 1-Benzyl
TestSubstance_CASRN: 69723-9
TestSubstance_Descripti... single ch
ChemicalNote: zwitterio
STRUCTURE_ChemicalNa... 1-(pheny
STRUCTURE_Parent_SMIL... C1C(=CC
STRUCTURE_InChI: InChI=1/
Fish-Toxicity is well correlated to PC-descriptors (used for embedding)
Feature: LC50_mmol
Description: Included in Dataset
Usage: NOT used for clustering and/or embed
Missing values: 0
#compounds
500
400
300
200
100
0
0
250
500
750
9.67 1-Benzylpyridinium 3-sulfonate
Dataset
Wireframe
Balls & Sticks
Dots
Feature: LC50_mmol
Label
Jmol


0.74 2,2-Dimethyl-1-propyl
0.74 Benzamide
0.8 Cyclohexanone
0.82 3-Methyl-3-pentanol
0.84 2-Cyanopyridine
0.84 2-Methoxyethylamine
0.85 Cyclohexanol
0.87 3-Acetamidophenol
0.88 tert-Butyl methyl ethe
0.88 3-Amino-5,6-dimethyl-
0.89 Isopropyl ether
0.94 5-Chloro-2-pyridinol
0.98 2-Picoline
0.99 1-Benzylpyridinium 3-s
Superimpose
Cluster: Single cl
Num compounds: 579
3D Alignement: No Clust
LC50_mmol_log: [-6.38; 2
Dataset: knime_in
Num compounds: 579
Cluster algorithm: No Datas
3D Embedding: Sammon
3D Embedding Quality: good (r^
Compound:
0.99 1-Benzylpyridinium 3-sulfonate
Endpoint: LC50
Species: fathead
ChemClass_FHM: Pyridines
CLOGP: -99.99
MLOGP: null
LC50_mg: 2410
LC50_mmol: 9.67
ActivityOutcome_EPAFHM: active
ActivityScore_EPAFHM: 21
LC50_Ratio: 1.39
LC50_Note: null
MOA: MOA not
MOA_Confidence: null
MOA_MixtureTest: null
ExcessToxicityIndex: 0
FishAcuteToxSyndrome: null
FishBehaviorTest: Conflictir
Note_EPAFHM: null
LC50_mmol_log: 0.99
LC50_mmol_log (predicti... -1.53
Predicted
Actual
Feature: LC50_mmol_log
Description: Included in Dataset
Usage: NOT used for clustering and/or embed
Missing values: 0
#compounds
0.99 1-Benzylpyridinium 3-sulfonate
Dataset
Highlight two features at once to detect high prediction errors
Wireframe
Balls & Sticks
Dots
Feature: LC50_mmol_log
Label
Jmol

More features

- Export clusters/compounds/features
- Export high-res images
- Access ChEMBL database
- Save and share embedding settings
- Data tables to browse through raw compound/feature/cluster data
- Command line interface
- Configurable highlight and view settings
  - Adjust highlight color gradient
  - Enable log highlighting
  - Switch between sphere and atom-color highlighting
  - …

Quotes about CheS-Mapper

… a very niece piece of software.

Basil Hartzoulakis, PhD, Xention Ltd, Cambridge

You help to make me look like a genius to my bosses ... this is the only open source tool I know of that is usable by a regular bench chemist.

Kerry W. Fowler, Ph.D., Senior Scientist
Kineta Inc, Seattle

CheS-Mapper has come in handy.

Kaushik Hatti, Vittal Mallya Scientific Research Foundation, Banagalore

I am very impressed by the ches-mapper software ... the tool is very effective to spot trends of a chemical group

Hiroshi Nara, Organic chemist, Japan

thanks for the gorgeous ches-mapper.

Santi Villalba, University College Dublin

Many compliments because we have found that it is highly useful and well working.

Prof. Paola Gramatica, University of Insubria