

Canvas

Customizable Cheminformatics Platform

**Schrödinger
2009 Overview**

Canvas - Customizable Cheminformatics Platform

- Chemical spreadsheet
- Fast, interactive charting and plotting
- Many features and options

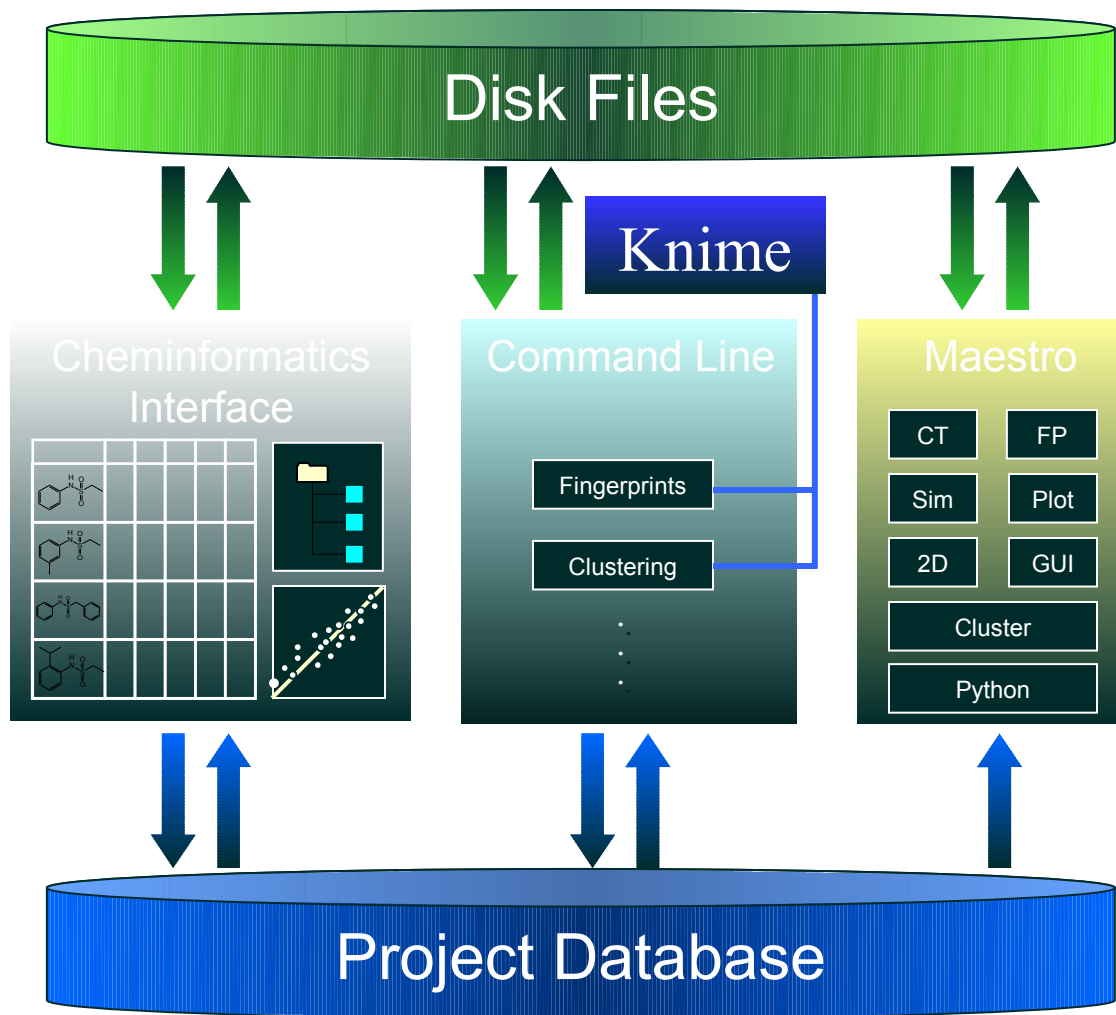
2D Molecular Properties
2D Fingerprints
Similarity/Distance Matrix
Similarity/Distance Screen

Diversity Analysis
Hierarchical Clustering
Self-Organizing Map
K-means
Leader-Follower

Principal Components Analysis
Bayes Classification
Multiple Linear Regression
Partial Least-Squares Regression
Principal Components Regression
Neural Networks
Maximum Common Substructure



Multiple Access Points to Canvas

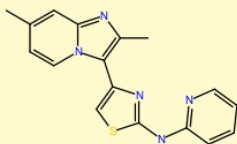


High Quality 2D Images in the Project Table

Project Table --- project_2009

Table Select Entry Property Group ePlayer

2D

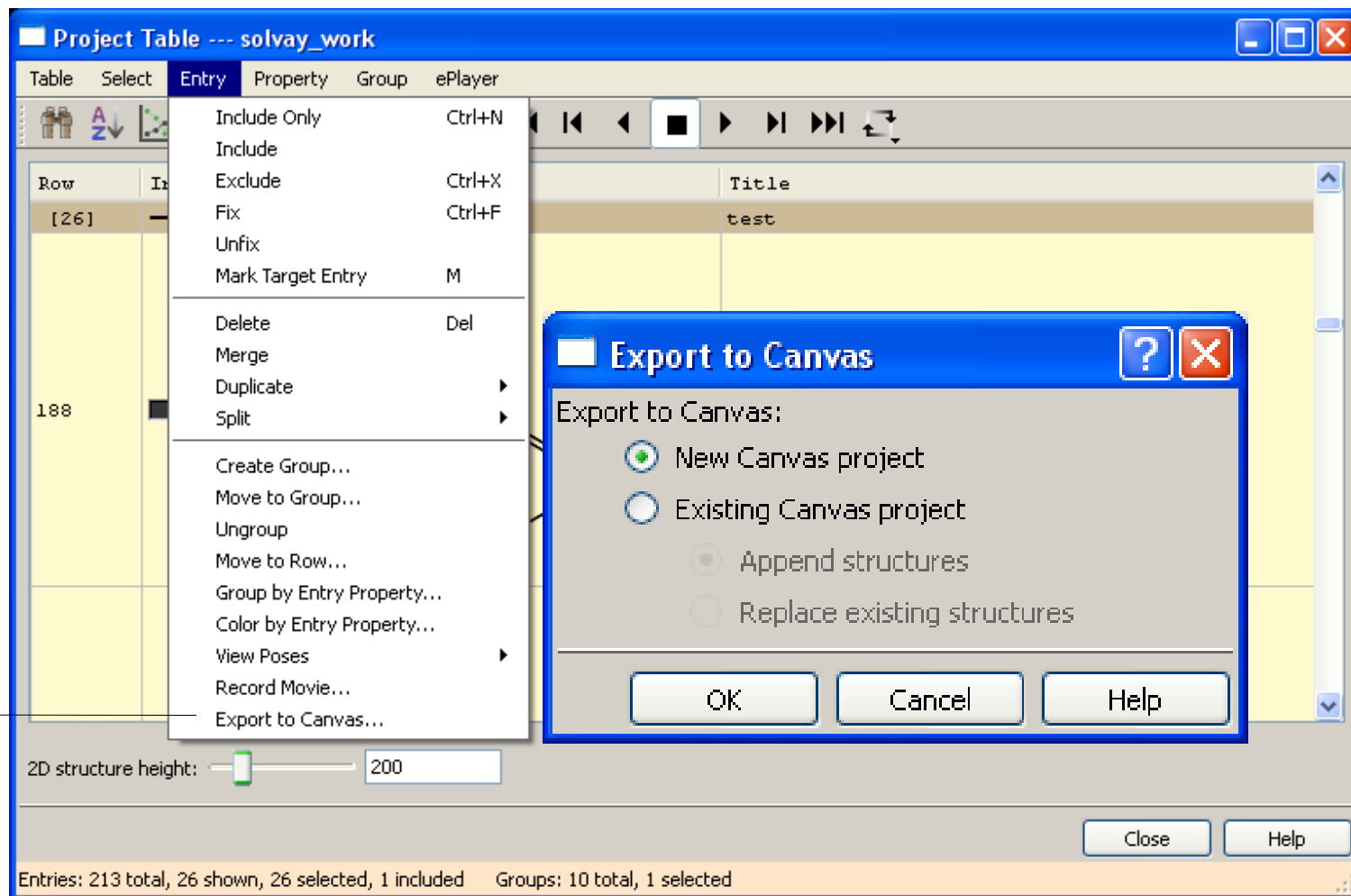
Row	In	2D Structure	Title	glide lnum	glide gscore	glide lipo	glide hbond	glide metal	glide rewards	glide evdw	glide ecoul	glide erotb	glide esite	glide emodel	glide energy
166	<input type="checkbox"/>		419851	12	-7.30	-2.51	-0.25	0.00e+00	-1.75	-34.24	-7.94	0.18	-0.06	-66.07	-42.18
167	<input type="checkbox"/>		620317	28	-6.94	-3.11	-0.19	0.00e+00	-1.65	-40.86	-2.34	0.40	0.00e+00	-61.59	-43.19
168	<input type="checkbox"/>		151943	5	-6.56	-2.18	-0.31	0.00e+00	-2.06	-40.99	-3.33	0.54	0.00e+00	-60.38	-44.32
169	<input type="checkbox"/>		451167	15	-6.55	-1.77	-0.44	0.00e+00	-2.45	-32.67	-8.22	0.99	-0.02	-62.09	-40.89

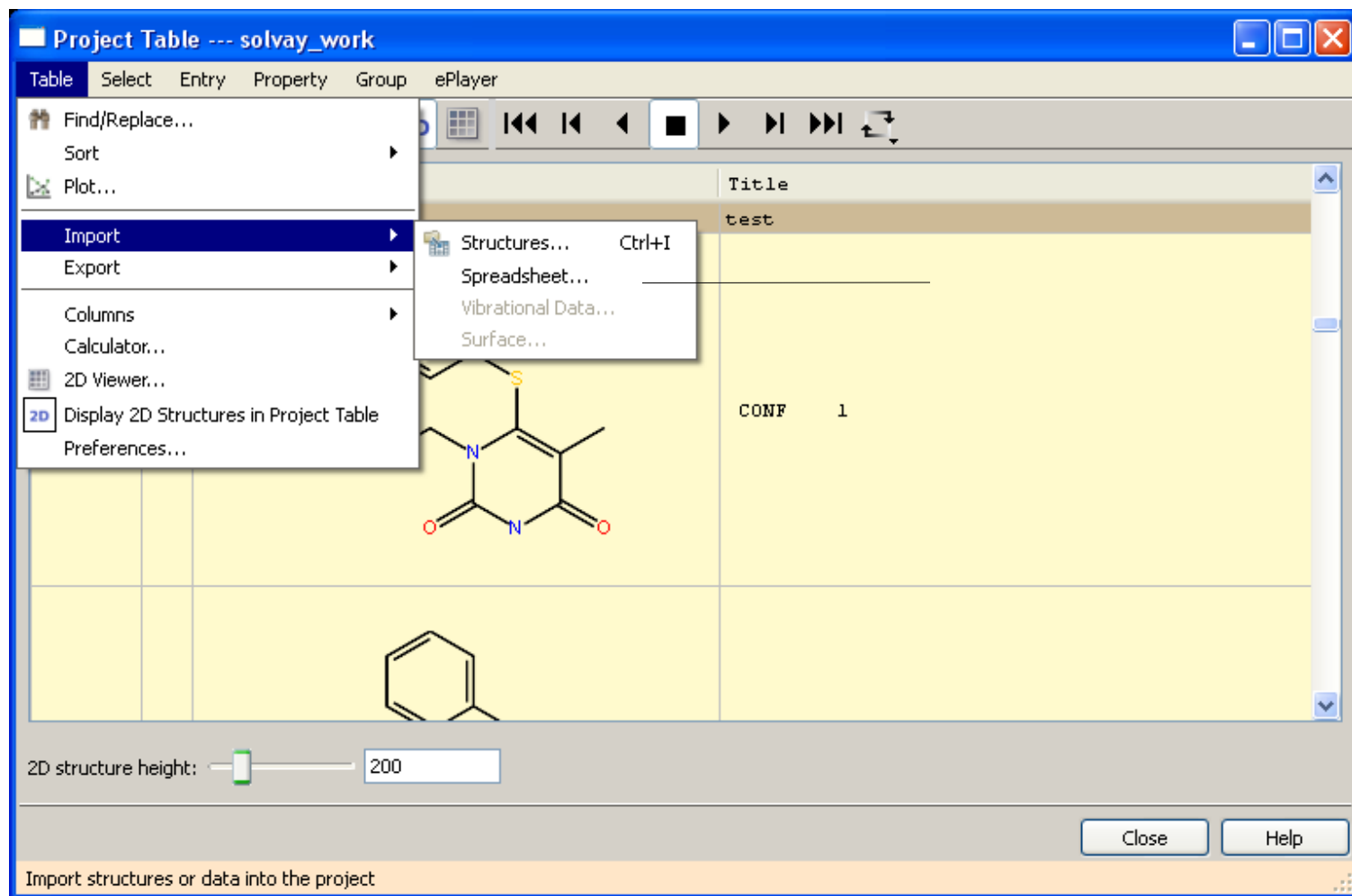
2D structure height:

Close Help

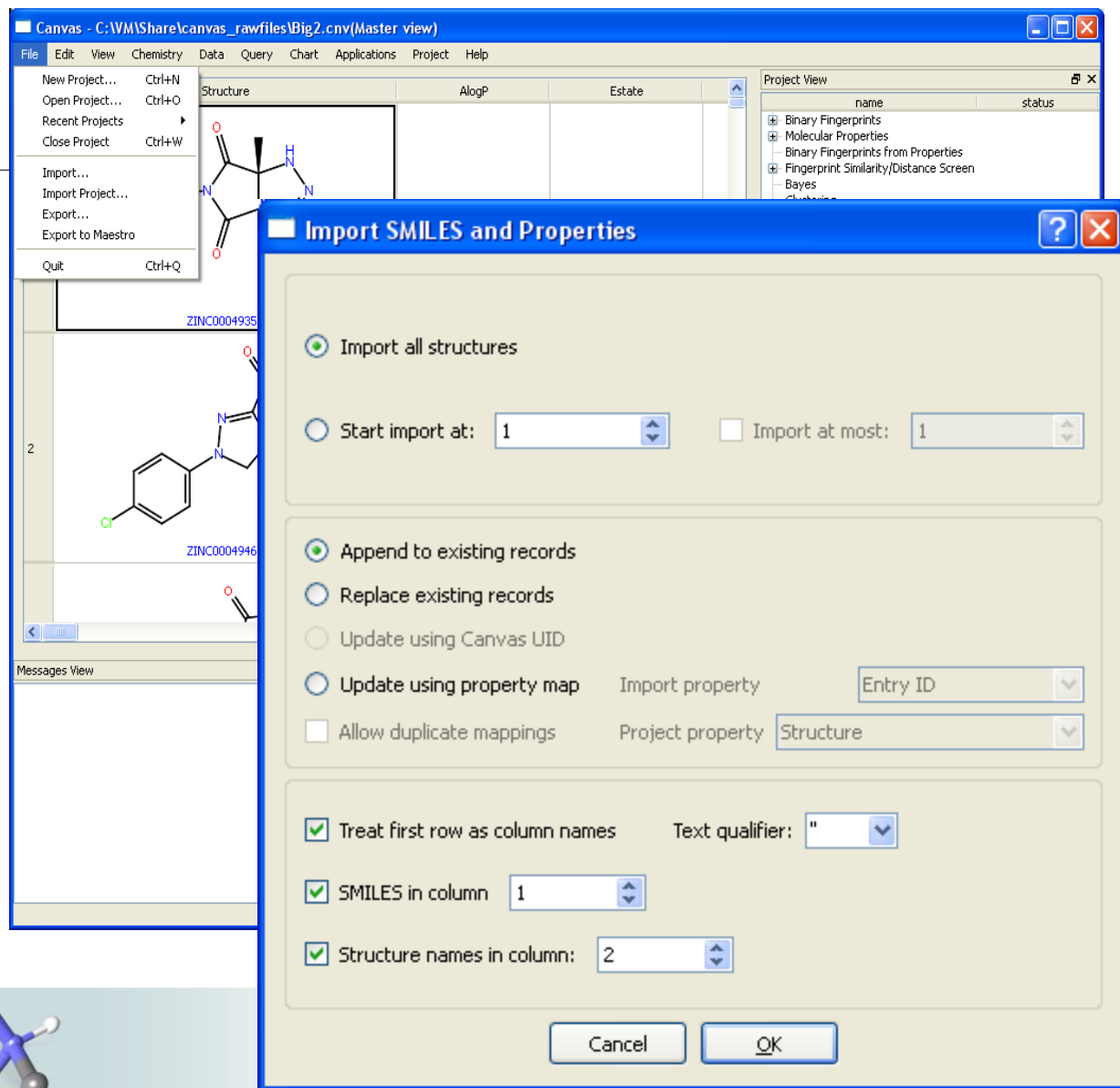


SCHRÖDINGER.





Csv
Mae
Sd
Smiles



GUI Features

- Input/output of standard file types
- 2D chemical spreadsheet
- Custom views
- Substructure searching
- Sorting
- Chemical & property filtering
- Statistics
- Charting
- Heat maps
- Molecular property calculations
- Binary fingerprints
- Similarity searching
- Clustering
- Diversity
- Self-organizing maps
- Maximum common substructure
- Naïve Bayes classification
- Multiple linear regression
- Partial least-squares regression
- Principal components analysis & regression
- Neural networks

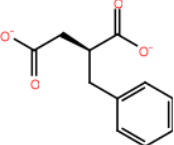
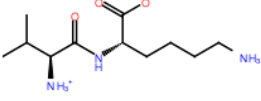
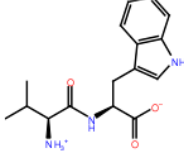


Overview

- Import
 - Maestro
 - SD
 - CSV
 - SMILES
- Open Canvas project
- Create custom views
- Sort/filter/query
- Statistics
- Heat maps
- Charts
- Applications
- View results
- Scroll smoothly over millions of rows

Canvas - (Master view)

File Edit View Chemistry Data Query Chart Applications Maestro Project Help

	Structure	FP_Linear	MW	AlogP	HBA	HBD
1	 1hyt_1	bit count= 56	206.195	0.1853	0	0
2	 1lna_1	bit count= 62	246.327	-3.0869	1	7
3	 3tmn	bit count= 150	303.356	-0.6693	1	5

Project View

name	status
Binary Fingerprints	
FP_Linear	Incorporated
Binary Fingerprints from Properties	
Fingerprint Similarity/Distance Screen	
Sim	Incorporated
Bayes	
Clustering	
Diversity	
Sphere_Excl	Finished
MCS	
MLR	
NNET	
PCA	
PCAReg	
PLS	
SOM	
SOM_Props	Finished
SOMBits	
SOM_FP_Linear	Finished
Views	
Passed Rule-of-5	
Sim_To_3tmn	
Chemistry Filters	
Sulfonamides	
Property Filters	
Rule-of-5	

Messages View

Total/Selected Rows: 402/0, Columns: 13/0



Unique Features...

FP types + Any combinations

ashing

- Storage 2^{32} -bit or 2^{64} -bit
- Cater to 4 billion distinct features in sparse space
- Advantage is to 'avoid' overlap or collisions
- Each feature can be 'picked up' and directly related back to activity
- Not lost amongst other features that are encoded in the same bit space
- SciTegic were first to use 2^{32} space ...non trivial to extend to 2^{64}
 - Hashing algorithm different so collision rate different
 - 2^{32} FPs is 0.1% Canvas (lower than SciTegic)



chemically understand what's driving a particular match

SCHRÖDINGER.

Unique Features...

atom typing schemes

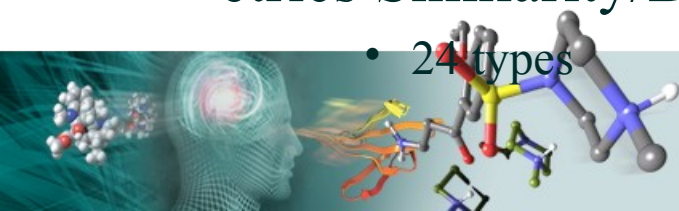
- ‘Treatment of Atoms and Bonds’
- 11 built-in, Custom (Phase atom-types, and OPLS atom-types)
- Typically only coded 1 or 2 atom typing schemes which can be applied to minimal FPs

calculating options (Count)

- Derived from topological descriptors
- Counts frequency of the features to give an idea of ‘feature density’
Better idea of functionality of the whole molecule
- 11 +Custom, Available in ALL FPs (Typically only in MACCS)

metrics Similarity/Diversity

- 24 types



Atom Typing Schemes

- 1 - All atoms equivalent; all bonds equivalent.
- 2 - Atoms distinguished by HB acceptor/donor; all bonds equivalent.
- 3 - Atoms distinguished by hybridization state; all bonds equivalent.
- 4 - Atoms distinguished by functional type: {H}, {C}, {F,Cl}, {Br,I}, {N,O}, {S}, {other}; bonds by hybridization. DEFAULT WITH RADIAL
- 5 - Mol2 atom types; all bonds equivalent.
- 6 - Atoms distinguished by whether terminal, halogen, HB acceptor/donor; bonds distinguished by bond order.
- 7 - Atomic number and bond order. DEFAULT WITH LINEAR
- 8 - Atoms distinguished by ring size, aromaticity, HB acceptor/donor, ionization potential, whether terminal, whether halogen; bonds distinguished by bond order.
- 9 - Carhart atom types (atom-pairs approach);
- 10 - Daylight invariant atom types; bonds distinguished by bond order.

C - Custom.
E - Estate atom types.

FP Scaling options

- 0 - no scaling (default)
- 1 - scale counts by feature size to unity
- 2 - scale counts by feature size to feature size
- 3 - scale counts by feature size to molecule size
- 4 - scale squares of counts by feature size to unity
- 5 - scale squares of counts by feature size to feature size
- 6 - scale squares of counts by feature size to molecule size
- 7 - scale sqrt of counts by feature size to unity
- 8 - scale sqrt of counts by feature size to feature size
- 9 - scale sqrt of counts by feature size to molecule size
- 10 - use raw feature counts
- 11 - use raw feature counts squared
- 12 - use sqrt of raw feature counts

Fingerprints

Metrics

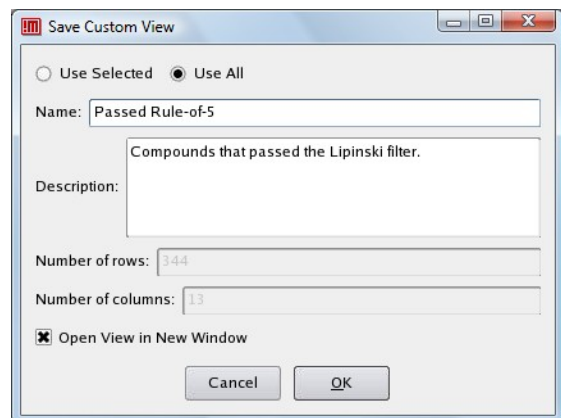
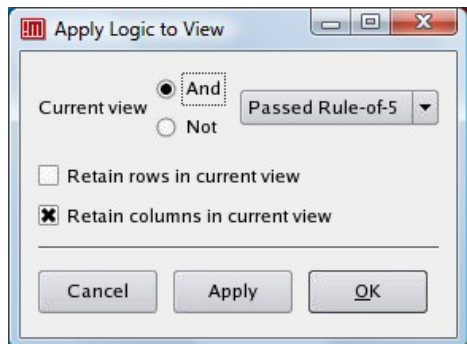
Buser Cosine Dice Hamann Hamming Dixon
Euclidean Kulczynski Matching
PatternDifference **Tanimoto** modifiedTan Shape
Size Simpson Petke Tversky Yuel Variance
Soergel rogersT Pearson Minmax

25,344 combinations!



Custom Views

- Create from selected/all
- Perform logic
- Export row IDs
- Undo operations



	Structure	FP_Linear	MW	AlogP
1	 1hyt_1	bit count= 56	206.195	0.1853
2	 3tmn	bit count= 150	303.356	-0.6693
3	 4tln_1	bit count= 18	146.188	-1.7731



Sorting

Canvas_2009.ppt [Compatibility Mode] - Microsoft PowerPoint

stry Data Query Chart Applications Maestro Project Help

Sort...
Property Filter...
Chemistry Filter...
Heat Map...
Statistics...

Sort Ascending
Sort Descending

Sort

☒ New sort ☐ Reverse current order

Sort by: AlogP ☐ Ascending ☒ Descending

Add
Remove

MW ↑
AlogP ↓

OK Cancel

- Single-property sort (right-click on column header)
- Multi-property sort



Filters and Queries

- Property filters with AND/OR logic
- REOS (Vertex)
- Custom chemistry filters
- Substructure queries with preview for literal substructures

Property Filter

Use existing filter: (none selected)

Show rows where:

MW <= 500

Filter String:

☒ And ☐ Or (AlogP <= 5.0) and (MW <= 500)

Chemistry Filter

Use existing filter: (none selected)

	SMARTS	Min	Max	Comments
1	[Cl,Br,I]	0	2	2 non-fluorine halogens
2	[N;+0,+1;\$(N(=O)~[O;H0;-0,-1])]	0	0	No nitro groups
3				
4				
5				
6				

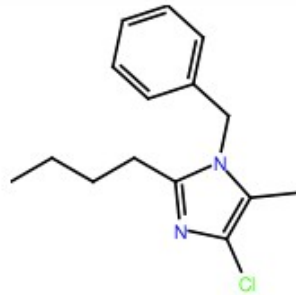
Substructure Query

Screen: ☐ Selected rows ☒ All rows

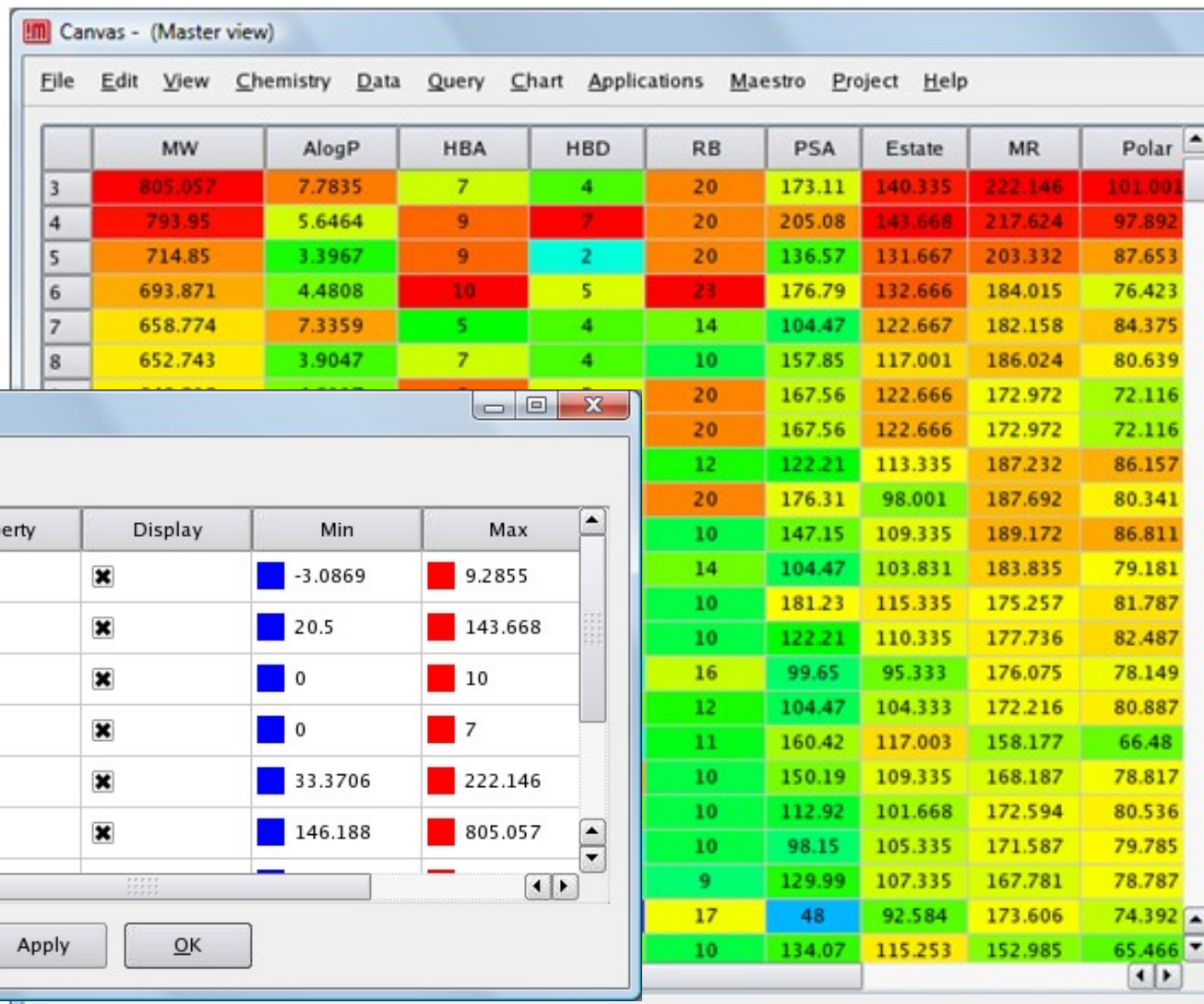
☐ Require exact match

SMARTS

Enter SMARTS: CCCCc(nc(Cl)c1C)n1Cc2cccc2



Heat Maps



Statistics

Statistics

Correlation	
Covariance	
Mean	
Median	
Mode	
Root-Mean-Square Deviation	
Standard Deviation	
Variance	

	Mean
AlogP	2.89681
Estate	62.3065
HBA	3.20398
HBD	1.55473
MR	100.08
MW	359.528
PSA	76.1972
Polar	44.0785
PSA	5.02786

Compute Export... Cancel

- Univariate → column
- Bivariate → matrix

Statistics

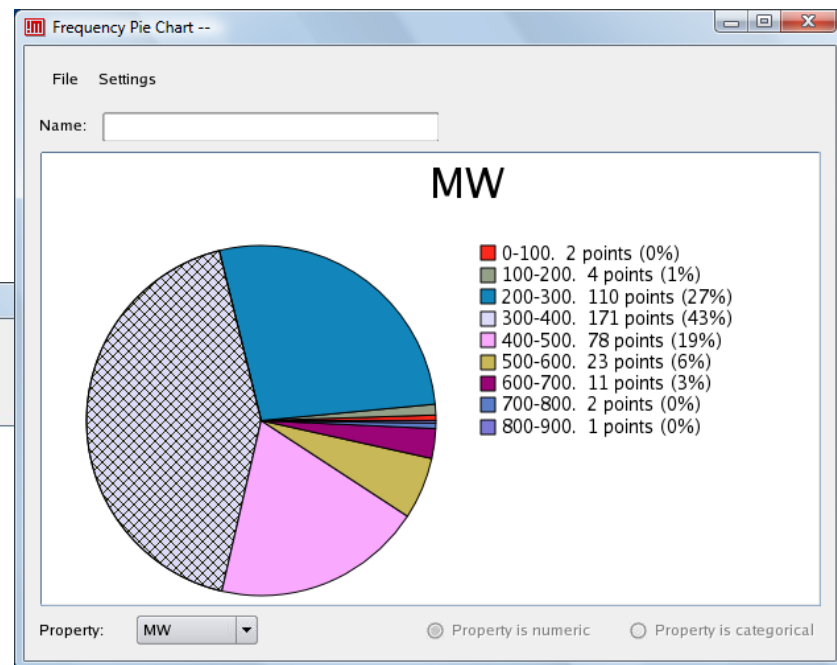
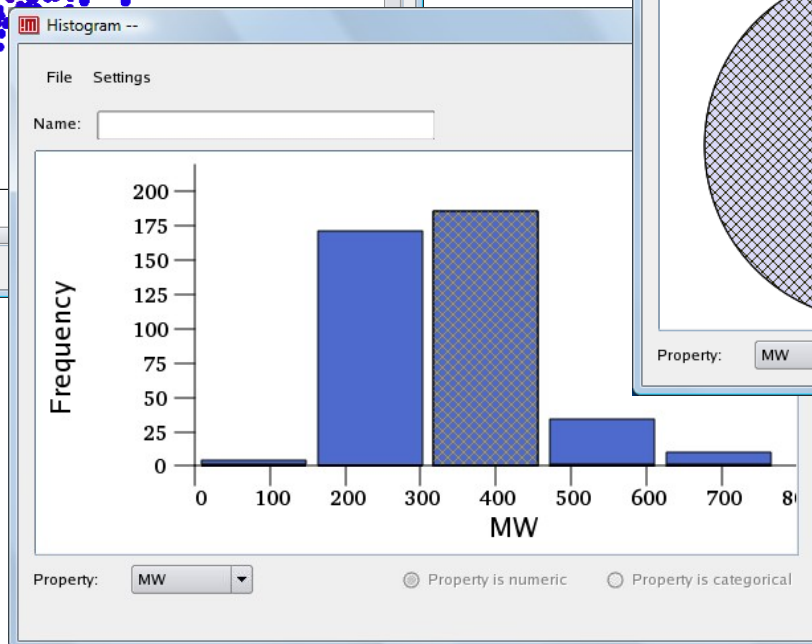
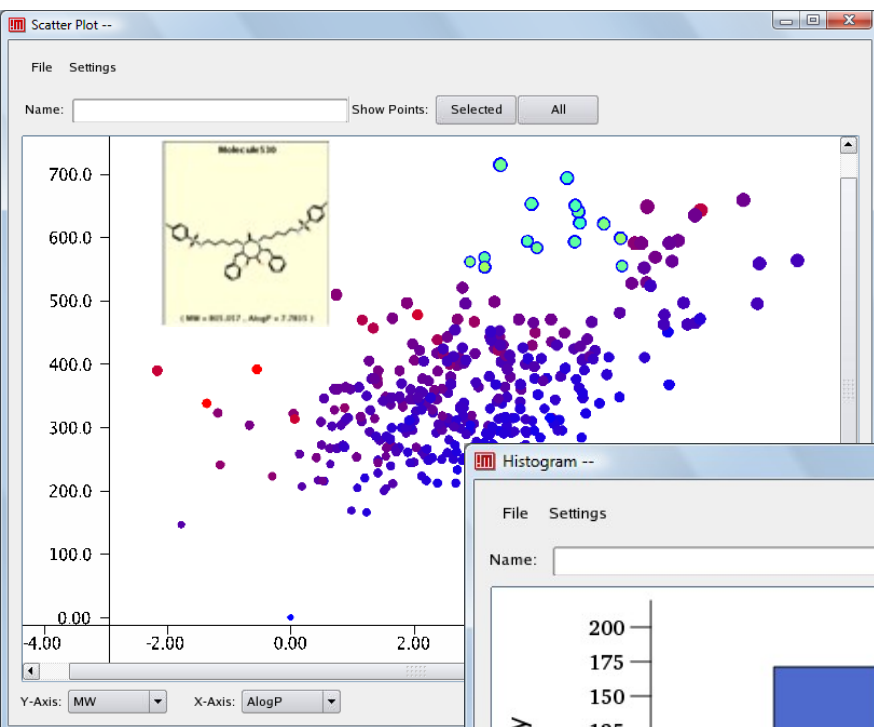
Correlation	
Covariance	
Mean	
Median	
Mode	
Root-Mean-Square Deviation	
Standard Deviation	
Variance	

	AlogP	Estate	HBA	HBD	MR	MW
AlogP	1	0.442202	-0.00975017	-0.0801911	0.683916	0.591047
Estate	0.442202	1	0.708151	0.379977	0.905857	0.959044
HBA	-0.0097...	0.708151	1	0.448408	0.52161	0.608988
HBD	-0.0801...	0.379977	0.448408	1	0.315646	0.34871
MR	0.683916	0.905857	0.52161	0.315646	1	0.979057
MW	0.591047	0.959044	0.608988	0.34871	0.979057	1
PSA	-0.172206	0.638568	0.707554	0.519652	0.411934	0.541107
Polar	0.706285	0.89838	0.502091	0.316499	0.987785	0.967318

Compute Export... Cancel

Charts

- Visualize up to four dimension in scatter plots (color/size of points)
- Selections are mirrored in spreadsheet
- Mouse-over points to see structures



Diversity Analysis

- Sphere exclusion
- Directed sphere exclusion
- MAXSUM
- MAXMIN
- View diverse structures
- Save as custom view

Diversity Analysis

Dataset: ☐ Selected rows ☒ All rows

Save results as:

Similarity/distance matrix source:

☒ From fingerprints ☐ From file

Fingerprint column:

File name:

Metric:

Description:

Diversity selection method:

Description:

☒ Diverse subset size:

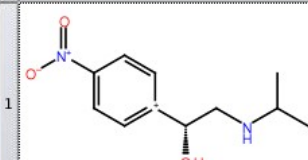
Exclusion sphere size:

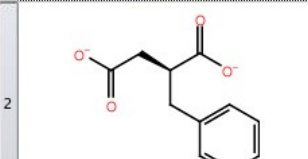
Initialization method: Seed:

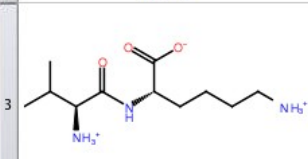
Diverse Compounds (5 t...

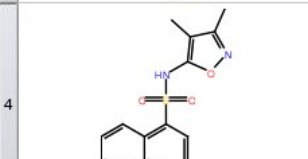
View Chemistry

Structure

1 
Molecule426

2 
1hyt_1

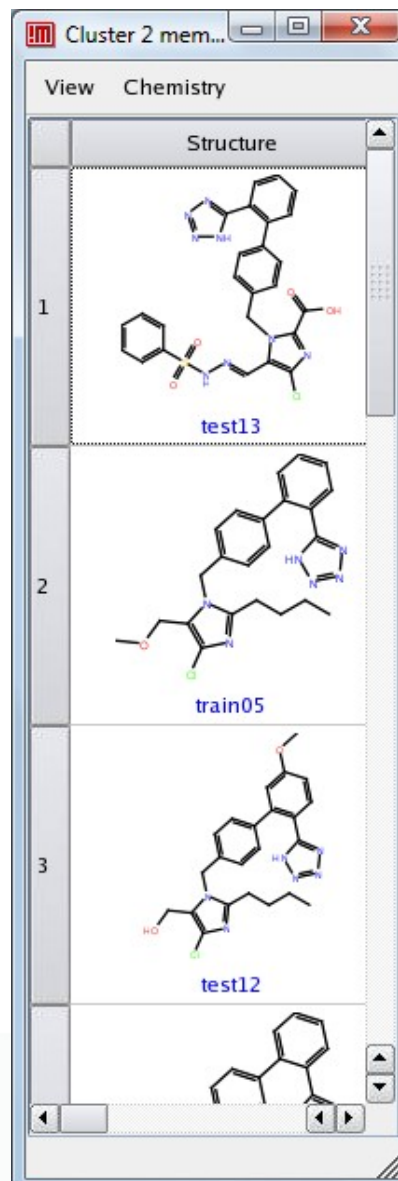
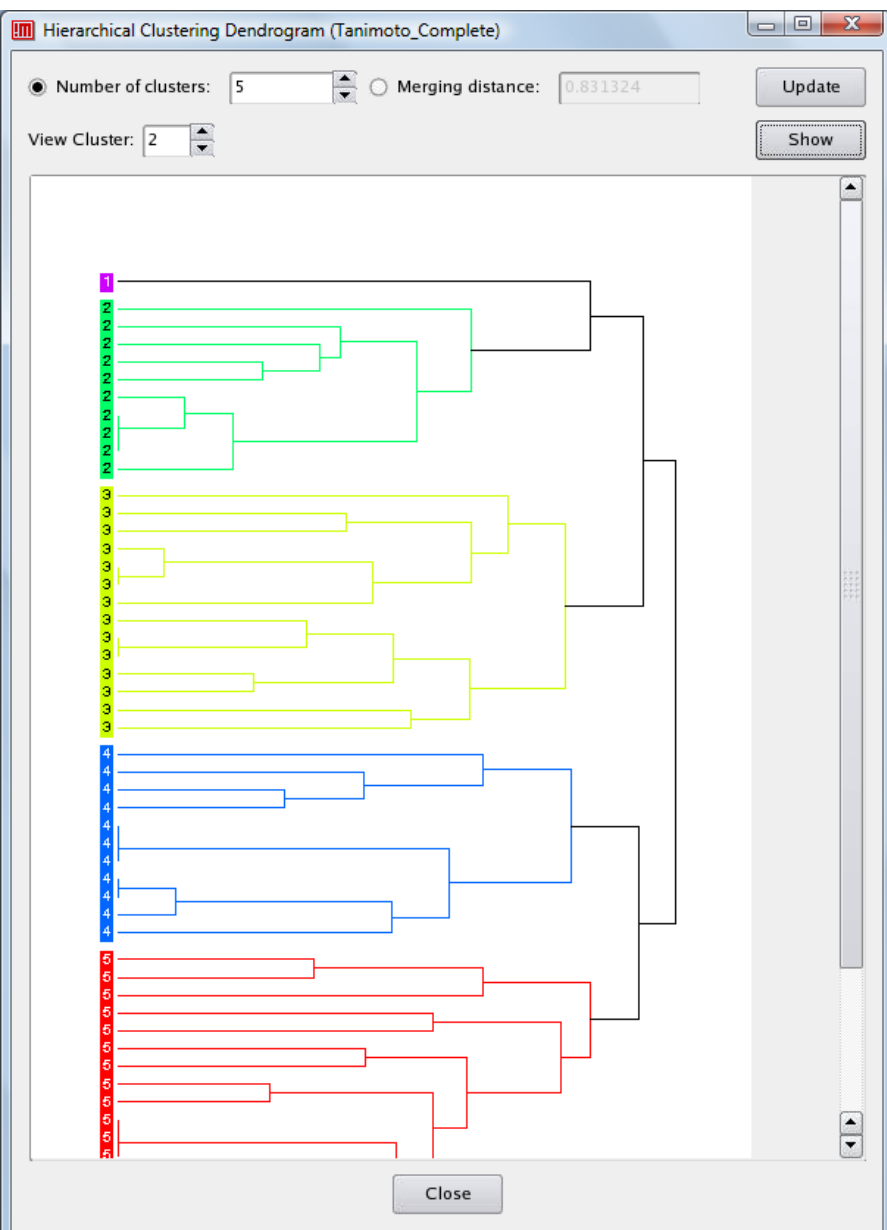
3 
1lna_1

4 
endo-1



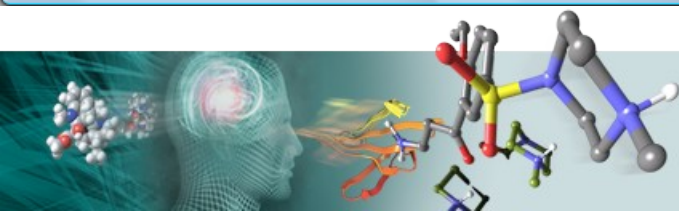
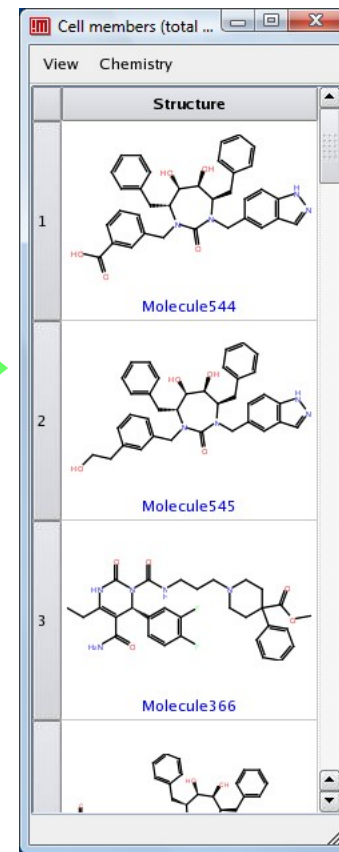
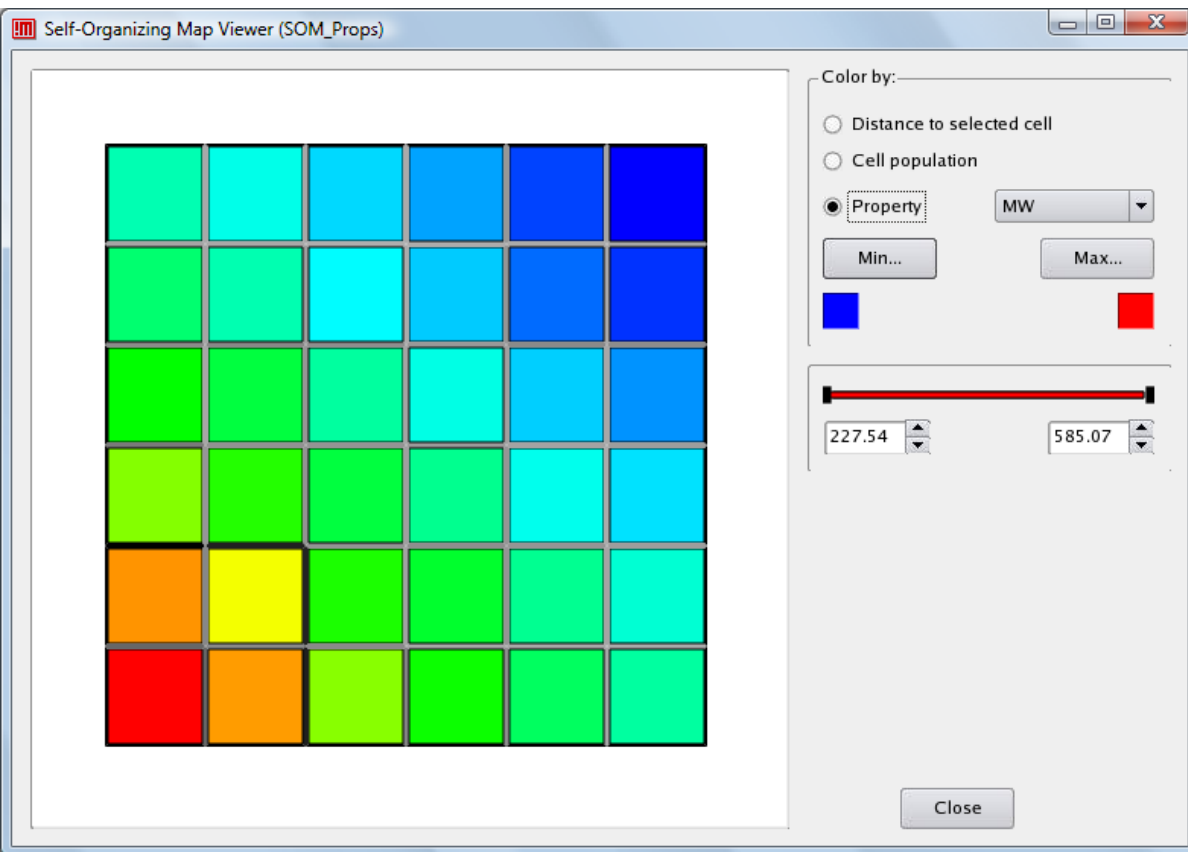
Hierarchical Clustering

- View different clustering levels
- Display cluster members
- Save as custom view

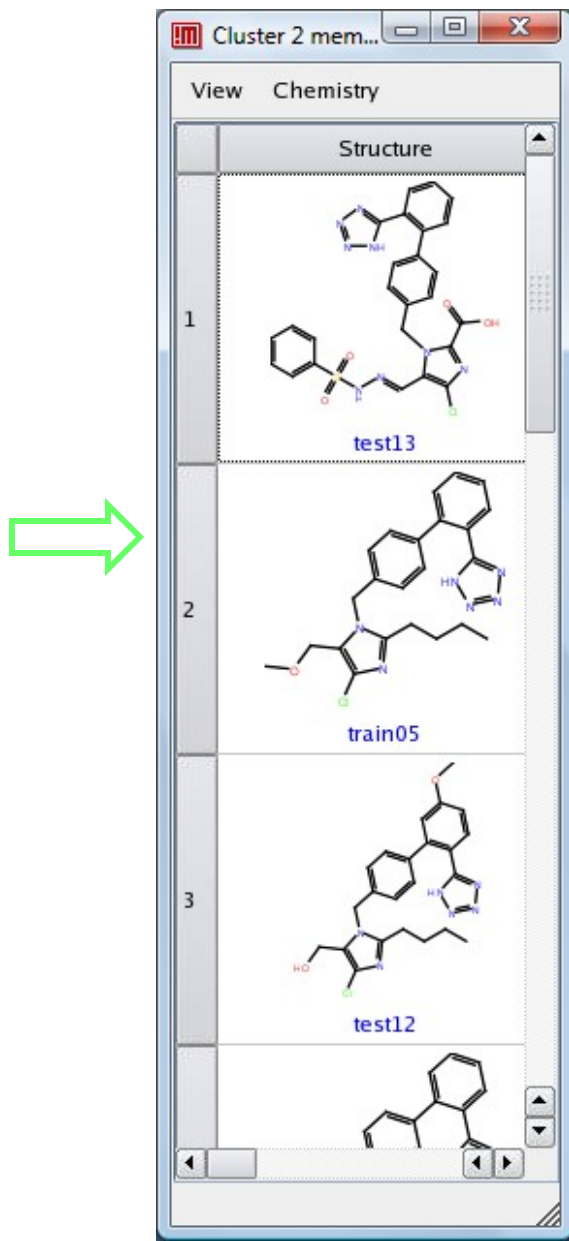


Self-Organizing Maps

- Create from properties or fingerprint
- Color cells by distance, population, property
- Right-click cell to see members



Place holder for Knime workflow slide



`$SCHRODINGER/utilities/canvasDBCS...`

Define # clusters, and, choice of how to pick from clusters

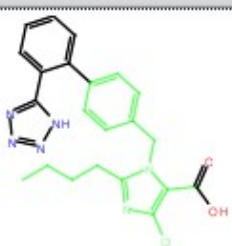
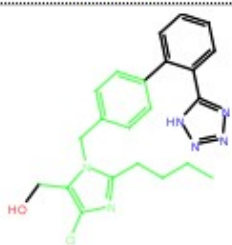
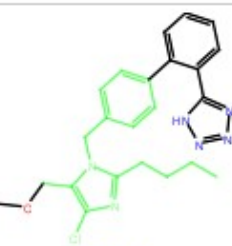
- n 4 -init representative
pick out centroid 'most representative'
- n 4 -init random
pick out random 'default'
- n 1 -init dissimilar
pick out edge 'dissimilar/diverse'

Maximum Common Substructure

MCS Matches (MCS_40-50)

View substructures that match exactly 40 compounds

Substructure size: 37 (18 atoms + 19 bonds) < Prev Group 1 of 1 Next >

	Structure	Substructure SMARTS
1	 train01	<chem>CCCCc(nc(Cl)c1C)n1Cc2ccccc2</chem>
2	 train02	<chem>CCCCc(nc(Cl)c1C)n1Cc2ccccc2</chem>
3	 train05	<chem>CCCCc(nc(Cl)c1C)n1Cc2ccccc2</chem>

Export... Close

- Match between 2 and n compounds
- View different MCS sizes simultaneously
- Paste SMARTS to Substructure Query tool
- 100 structures in ~1 second

Bayes Classification

- Use numeric properties and/or fingerprints
- Choose class divisions with visual feedback

Bayes Classification

Dataset: ☐ Selected rows ☒ All rows Save results as:

☒ Build new model ☐ Apply saved model ☐ Apply imported model

X variables Original X variables

Y variable Original Y variable

☒ Y is numeric ☐ Y is categorical

Assign Random Training Set Percentage: % Seed:

Advanced Options...

	Structure	Training Set	pIC50-Exp	Observed	Predicted
1	train01	<input checked="" type="checkbox"/>	8.886	Class 2	Class 2
2	train02	<input checked="" type="checkbox"/>	7.721	Class 2	Class 2
3	train03	<input checked="" type="checkbox"/>	7.699	Class 2	Class 2
4	train04	<input checked="" type="checkbox"/>	7.538	Class 2	Class 2
5	train05	<input checked="" type="checkbox"/>	7.377	Class 2	Class 2
6	train06	<input checked="" type="checkbox"/>	7.097	Class 2	Class 2

Training Test

	Observed	Correct	Incorrect
Class 1	10	3	
Class 2	10	2	
Totals:	20	5	

Bayes Classification - Bins

Y variable:

Number of bins: ☒ Equal widths ☐ Equal populations

3.82 6.35 8.89

Class 1 (27) Class 2 (23)

Partial Least-Squares Regression

Partial Least-Squares Regression

Dataset: ☐ Selected rows ☒ All rows Save results as: PLS_QSAR

☒ Build new model ☐ Apply saved model ☐ Apply imported model Import...

X variables Original X variables

Activity
Bit1
Bit10
Bit100
Bit101
Bit102
Bit103
Bit104
Bit105
Bit106
Bit107
Bit108
Bit109
Bit11

Y variable Original Y variable

Activity

Maximum number of PLS factors: 3 ☐ Stop adding PLS factors when SD drops to: -1.0

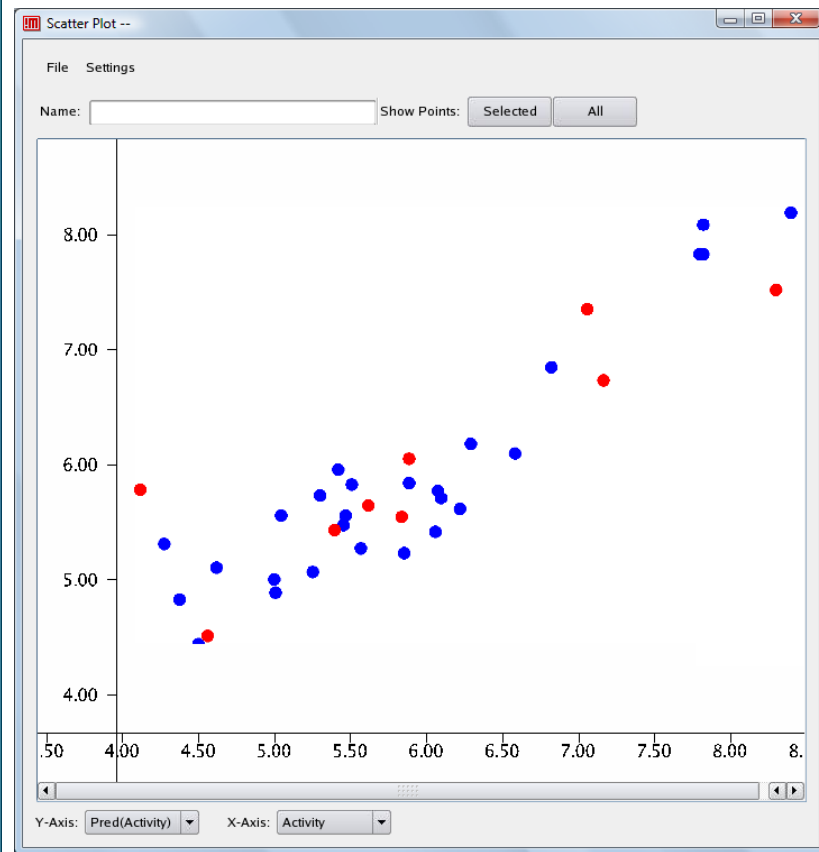
☐ Autoscale X variables ☐ Eliminate X variables with [t-value] <

Assign Random Training Set Percentage: 75 % Seed: 12345

Structure	Training Set	Activity	Pred(Activity)
1 endo-1	<input checked="" type="checkbox"/>	5.509	5.82248
2 endo-3	<input checked="" type="checkbox"/>	5.469	5.54966
3 endo-5	<input checked="" type="checkbox"/>	7.824	7.82072
4 endo-6	<input type="checkbox"/>	7.06	7.3482
5 endo-8	<input checked="" type="checkbox"/>	8.398	8.23882
6 endo-9	<input type="checkbox"/>	8.301	7.5154
7 endo-13	<input type="checkbox"/>	5.839	5.5412
8 endo-14	<input checked="" type="checkbox"/>	5.252	5.05955
9 endo-15	<input checked="" type="checkbox"/>	6.06	5.41133

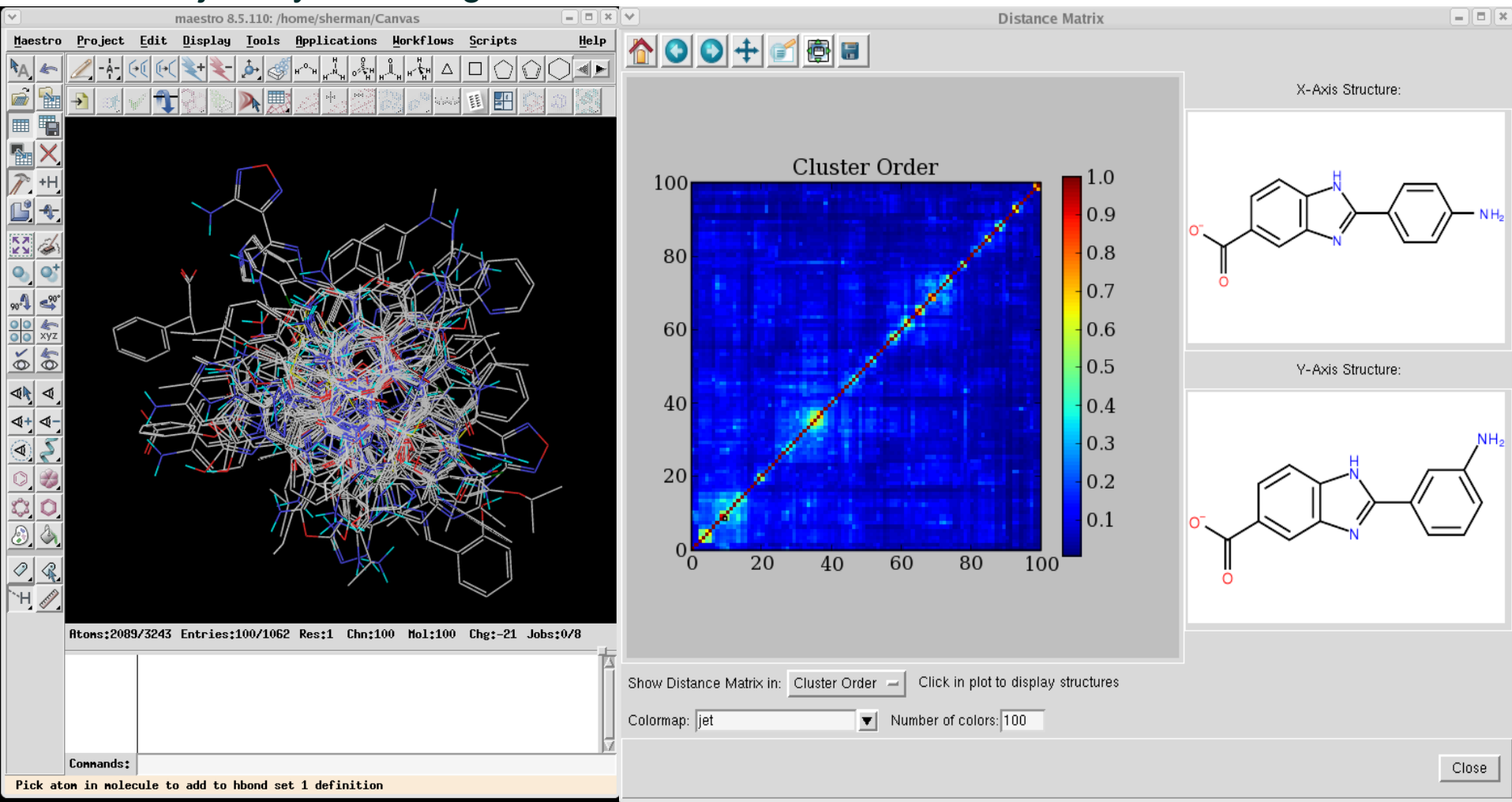
#PLS Factors	SD	R^2	R^2-CV	Stability
1	0.669898	0.643075	0.507797	0.929578
2	0.505899	0.804585	0.483726	0.853199
3	0.429946	0.864738	0.458017	0.751337

Build Model Apply Model Export Model... Scatter Plot... Cancel



Canvas in Maestro

- Interactive gui for FP generation, similarity calculation, and clustering
- Structural interaction fingerprints
- Volume clustering
- Trajectory clustering



Hole Filling Approach with canvasDBCS

- WKS-SCHROD-JAS# ls
50DivZincdb.mae 71_knowncdk2inhibs.mae
50zinc.fp 71inhibs.fp

```
WKS-SCHROD-JAS# $SCHRODINGER/utilities/canvasDBCS -ifp 50zinc.fp -ifp2  
71inhibs.fp -n 10
```

```
"subset index","representative","distance"
```

```
1,ZINC00150686,0.96827  
2,ZINC00066585,0.95916  
3,ZINC00174484,0.948655  
4,ZINC00406225,0.948505  
5,ZINC00341309,0.943493  
6,ZINC04334275,0.941558  
7,ZINC01559545,0.934981  
8,ZINC04156836,0.933824  
9,ZINC02482648,0.929924  
10,ZINC00160320,0.928775
```

```
CPU time = 0.22 sec
```

```
canvasDBCS successfully completed.
```

Methods for Diversity

Sphere exclusion (Sphere) - Hudson, B. D.; Hyde, R. M.; Rahr, E.; Wood, J. Parameter Based Methods for Compound Selection from Chemical Databases.

Quant. Struct.-Act. Relat. 1996, 15, 285-289.

Directed Sphere Exclusion (DISE) - Gobbi, A. G.; Lee, M.-L. DISE: Directed Sphere Exclusion. J. Chem. Inf. Comput. Sci. 2003, 43, 317-323.

MAXSUM - Maximum sum of pairwise distances. Each iteration adds the compound with the largest sum of distances to the compounds already in the set.

MAXMIN - Largest minimum distance. Each iteration adds the compound with the largest minimum distance to all compounds already in the set.



The Canvas Interface

Model Building
Handling Large Datasets
Command Line



SCHRÖDINGER