Visual Analysis of Chemical Space with Scaffold Hunter

Nils Kriege

Dept. of Computer Science, TU Dortmund

OpenTox Euro, Mainz, 02. October 2013

Chemical Data in Drug Discovery

Chemical Space

- Theoretical chemical space: $\sim 10^{62}$ molecules
- *De-novo* libraries: several hundreds of millions
- Commercially-available: 21 million molecules (ZINC)

Trend

- Increasing amount of available data (public or in-house)
- Need to systematically explore and analyze data to speed up drug discovery process

Cyclic Knowledge Discovery by Visual Analysis

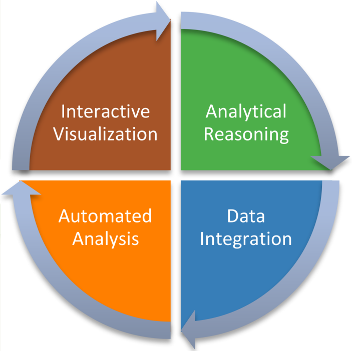
Classical Approach: Raw data → Analysis → Visualization

Visualization

- Analysis results
- Raw data
- Linked views

Analysis

- Clustering
- Classification
- SAR



Reasoning

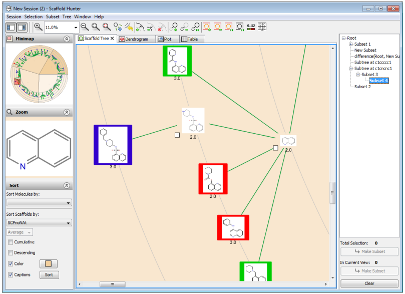
- New hypotheses
- Decision making
- Intuition of domain experts

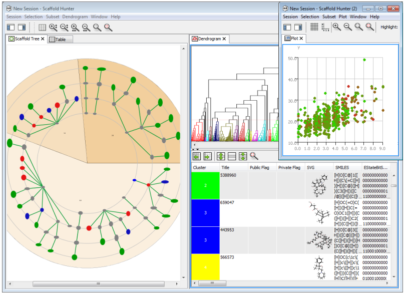
Data Integration

- Diverse data sources
- Experimental results

Scaffold Hunter

- Java-based Open Source tool
- Development started 2007, TU Dortmund

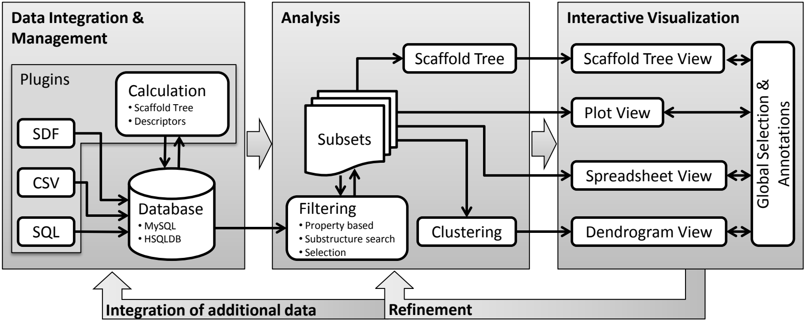



New Session - Scaffold Hunter
Session Selection Subset Dendrogram Window Help
Scaffold Tree
Table
Dendrogram
New Session - Scaffold Hunter (2)
Session Selection Subset Plot Window Help
Plot
Highlight:
Cluster
Title
Public Flag
Private Flag
SVG
SMILES
EStateBitS...

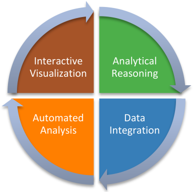
Goal:

- Import of data from a variety of sources
- Integrated visualization and analysis
- Interactive exploration in a systematic manner

Scaffold Hunter for Visual Analysis



- Facilitate cyclic knowledge discovery process
- Refinement of subsets, analysis parameters
- Integration of additional experimentally obtained data



Scaffold Tree: Concepts & Algorithms

- Hierarchical classification scheme based on core structures
- Rule-based parent scaffold selection
- Scaffolds as representatives for sets of similar molecules
- *Virtual* scaffolds without associated molecules

Algorithm

- For each molecule:
    1. Prune terminal side chains
       → scaffold
    2. Successively remove rings
       → unique parent scaffolds
- Merge multiple scaffolds
  → scaffold tree

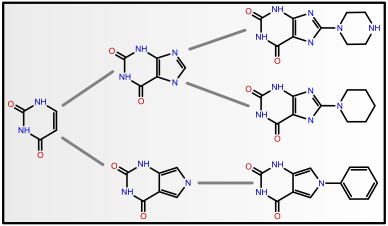
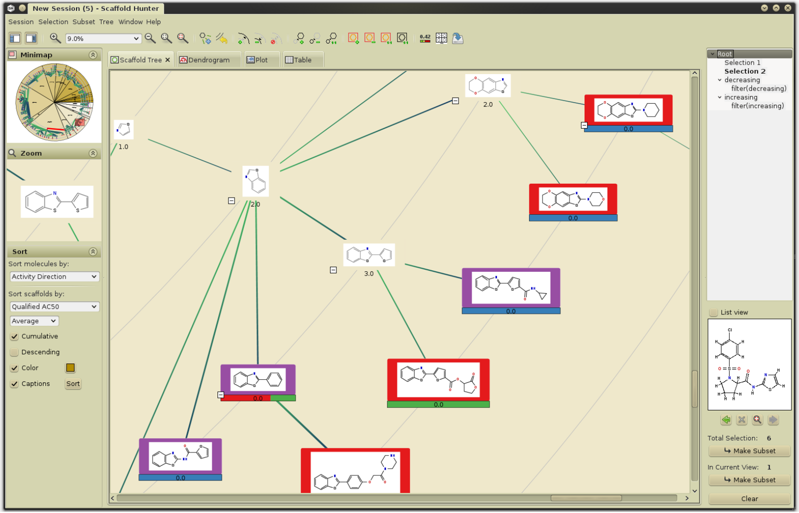


Figure: Branch of a scaffold tree

The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification
Schuffenhauer, Ertl, Roggo, Wetzel, Koch, Waldmann; J. Chem. Inf. Model., 2007, 47, 47-58

Scaffold Tree: Visualization



Sort molecules by: Activity Direction

Sort scaffolds by: Qualified AC50

Average

☑ Cumulative

□ Descending

☑ Color

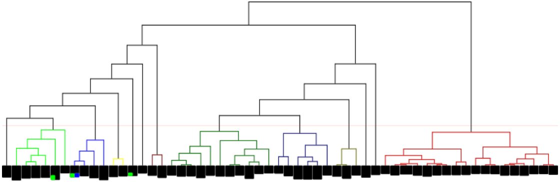
☑ Captions

Total Selection: 6

In Current View: 1

- Details-on-demand: Scaffold depiction adapts to zoom level
- Property Mapping: Representation by visual attributes

Hierarchical Clustering: Concepts & Algorithms



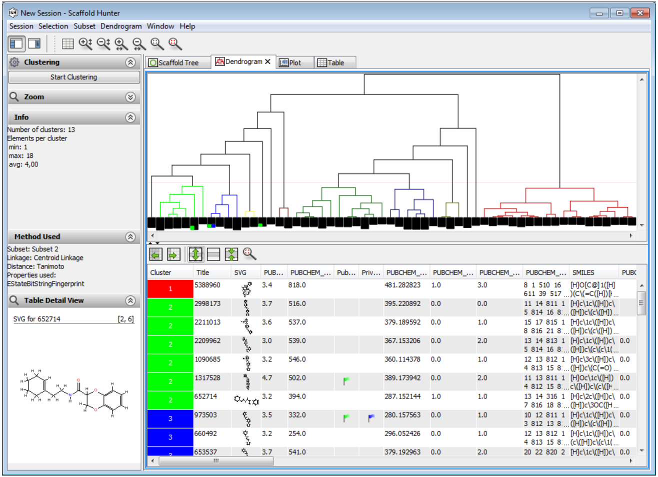
SAHN Clustering

- **Distance** between molecules, e.g., Tanimoto & fingerprints
- **Linkage:** Distance between clusters, e.g., Average or Ward
- **Algorithm:**
  1. Start with singleton clusters
  2. Merge pairs of clusters with minimum distance until a single cluster is obtained

Heuristic SAHN Clustering

- Subquadratic running time in practice, low memory footprint
- Support for arbitrary metric distance measures

Hierarchical Clustering: Visualization



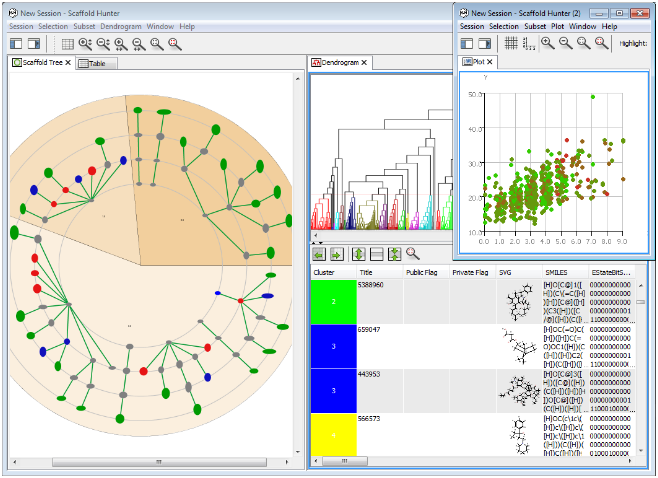
Number of clusters: 13
Elements per cluster
min: 1
max: 18
avg: 4,00

Subset: Subset 2
Linkage: Centroid Linkage
Distance: Tanimoto
Properties used:
EStateBitStringFingerprint

Cluster	Title	SVG	PUB...	PUBCHEM_...	Pub...	Priv...	PUBCHEM_...	PUBCHEM_...	PUBCHEM_...	PUBCHEM_...	SMILES	PUB(
1	5388960		3.4	818.0			481.282823	1.0	3.0	8 1 510 16 611 39 517 ...	[H]O[C@]1([H])(C\(=C([H])[H...	
2	2998173		3.7	516.0			395.220892	0.0	0.0	11 14 811 1 5 814 16 8...	[H]c\1c\([H])c\ ([H])c\(c\([H])...	
2	2211013		3.6	537.0			379.189592	0.0	1.0	15 17 815 1 8 816 21 8...	[H]c\1c\([H])c\ ([H])c\(c\([H])...	
2	2209962		3.0	539.0			367.153206	0.0	2.0	13 14 813 1 5 814 16 8...	[H]c\1c\([H])c\ ([H])c\(c\(c\1(...	0.0
2	1090685		3.2	546.0			360.114378	0.0	1.0	12 13 812 1 4 813 15 8...	[H]c\3c\([H])c\ ([H])c\(C(=O) ...	0.0
2	1317528		4.7	502.0	⚑		389.173942	0.0	2.0	11 13 811 1 4 812 15 8...	[H]Oc\1c\([H]) c\([H])c\(c\([H...	0.0
2	652714		3.2	394.0			287.152144	1.0	1.0	13 14 316 1 7 816 18 8...	[H]c\2c\([H])c\ ([H])c\3OC([H...	0.0
3	973503		3.5	332.0	⚑	⚑	280.157563	0.0	1.0	10 12 811 1 3 812 13 8...	[H]c\1c\([H])c\ ([H])c\(c\([H])...	0.0
3	660492		3.2	254.0			296.052426	0.0	1.0	12 13 812 1 4 813 15 8...	[H]c\1c\([H])c\ (c\([H])c\(c\1(...	0.0
3	653537		3.7	541.0			379.192963	0.0	2.0	20 22 820 2	[H]c\1c\([H])c\	0.0

- Zoomable user interface with details-on-demand
- **Cluster selection bar:** Interactive refinement of clustering
- **Table View:** Embeddable synchronized spreadsheet

Plot View

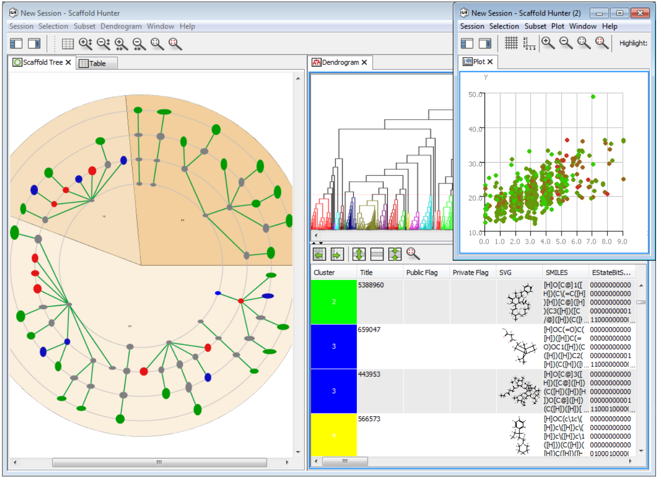


New Session - Scaffold Hunter
Session Selection Subset Dendrogram Window Help
Scaffold Tree
Table
Dendrogram
New Session - Scaffold Hunter (2)
Session Selection Subset Plot Window Help
Highlight:
Plot
y
50.0
40.0
30.0
20.0
10.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0

Cluster	Title	Public Flag	Private Flag	SVG	SMILES	EStateBitS...
2	5388960				[H]O[C@]1([H])(C\(=C([H])[H])[C@]([H])(C3([H])([C /@]([H])(C([I...	00000000000 00000000000 00000000000 00000000001 1100000000(...
3	659047				[H]OC(=O)C([H])([H])C(= O)OC1([H])(C ([H])([H])C2([H])(C([H])([I...	00000000000 00000000000 00000000000 00000000001 1100000000(...
3	443953				[H]O[C@]3([H])([C@]([H]) (C([H])([H])[H])O[C@]([H]) (C([H])([H])[...	00000000000 00000000000 00000000000 00000000001 1100010000(...
4	566573				[H]OC(c\1c\([H])c\([H])c\([H])c\([H])c\1 ([H]))(C([H])([H])C([H])([H	00000000000 00000000000 00000000000 00000000000 0100010000(

- 2D/3D scatter plot
- Mapping of attributes to axes, color, dot size etc.

Plot View

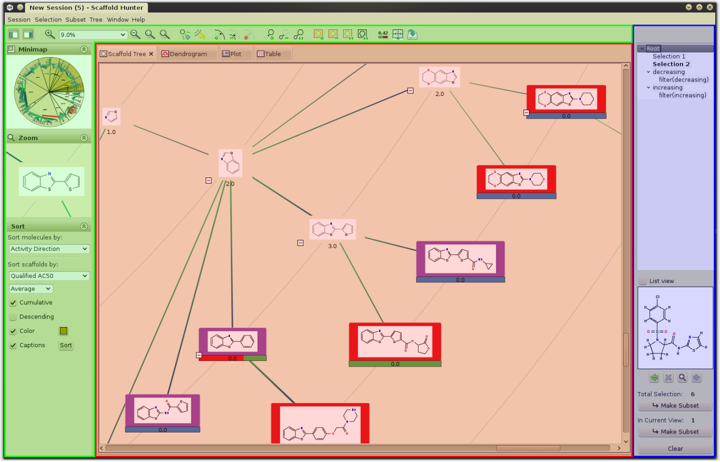

New Session - Scaffold Hunter
Session Selection Subset Dendrogram Window Help
Scaffold Tree
Table
Dendrogram
New Session - Scaffold Hunter (2)
Session Selection Subset Plot Window Help
Highlight:
Plot
y
50.0
40.0
30.0
20.0
10.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0

Cluster	Title	Public Flag	Private Flag	SVG	SMILES	EStateBitS...
2	5388960				[H]O[C@]1([H])(C\(=C([H])[H])[C@]([H])(C3([H])([C /@]([H])(C([I...	00000000000 00000000000 00000000000 00000000001 1100000000(...
3	659047				[H]OC(=O)C([H])([H])C(= O)OC1([H])(C ([H])([H])C2([H])(C([H])([I...	00000000000 00000000000 00000000000 00000000001 1100000000(...
3	443953				[H]O[C@]3([H])([C@]([H]) (C([H])([H])[H])O[C@]([H]) (C([H])([H])[...	00000000000 00000000000 00000000000 00000000001 1100010000(...
4	566573				[H]OC(c\1c\([H])c\([H])c\([H])c\([H])c\1 ([H]))(C([H])([H])C([H])([H	00000000000 00000000000 00000000000 00000000000 0100010000(

- 2D/3D scatter plot
- Mapping of attributes to axes, color, dot size etc.

→ **Effective use requires intuitive linking of views!**

Scaffold Hunter Main Window



Sort molecules by: Activity Direction

Sort scaffolds by: Qualified AC50

Average

☑ Cumulative

□ Descending

☑ Color

☑ Captions Sort

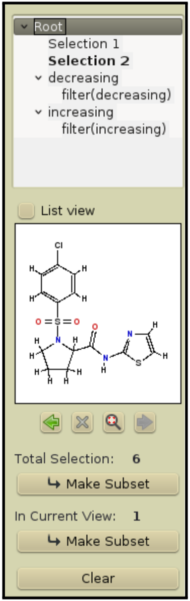
Total Selection: 6

In Current View: 1

- **Red**: Currently open views in tabs
- **Green**: View-specific tool- and sidebar
- **Blue**: Global subset and selection management

Coordination & Linkage of Views

- **Global selection:**
  - Synchronized selection over all views
  - Selection browser for quick access



Root
- Selection 1
- **Selection 2**
- decreasing
  - filter(decreasing)
- increasing
  - filter(increasing)

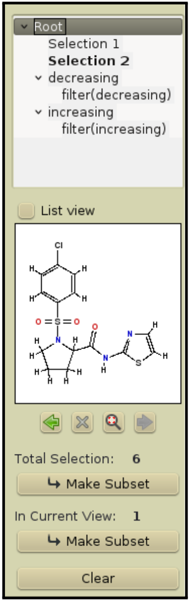
□ List view

Total Selection: **6**

In Current View: **1**

Coordination & Linkage of Views

- **Global selection:**
  - Synchronized selection over all views
  - Selection browser for quick access
- **Subset Management:**
  - Hierarchy of subsets
  - Change underlying subset of view
  - Multiple views on different subsets

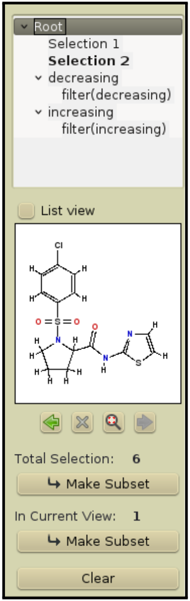


Total Selection: 6

In Current View: 1

Coordination & Linkage of Views

- **Global selection:**
  - Synchronized selection over all views
  - Selection browser for quick access
- **Subset Management:**
  - Hierarchy of subsets
  - Change underlying subset of view
  - Multiple views on different subsets
- **Filtering:** Selection, Property-based, SSS



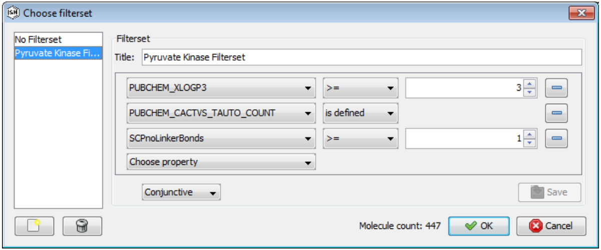
Root
Selection 1
Selection 2
decreasing
filter(decreasing)
increasing
filter(increasing)

☐ List view

Total Selection: 6
Make Subset

In Current View: 1
Make Subset

Clear



Choose filterset

No Filterset
Pyruvate Kinase Fi...

Filterset

Title: Pyruvate Kinase Filterset

PUBCHEM_XLOGP3 >= 3
PUBCHEM_CACTVS_TAUTO_COUNT is defined
SCPnoLinkerBonds >= 1
Choose property

Conjunctive

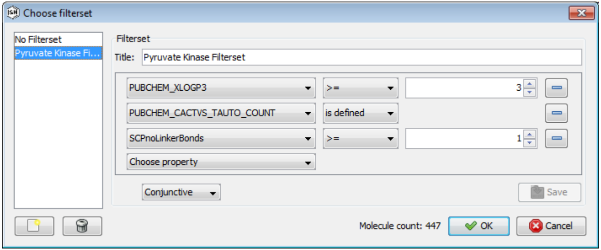
Save

Molecule count: 447

OK Cancel

Coordination & Linkage of Views

- **Global selection:**
  - Synchronized selection over all views
  - Selection browser for quick access
- **Subset Management:**
  - Hierarchy of subsets
  - Change underlying subset of view
  - Multiple views on different subsets
- **Filtering:** Selection, Property-based, SSS
- **Annotations:** Tooltip, comments, . . .



Choose filterset

No Filterset
Pyruvate Kinase Fi...

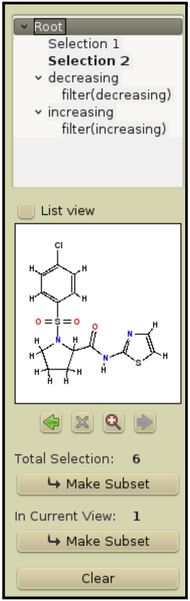
Filterset

Title: Pyruvate Kinase Filterset

PUBCHEM_XLOGP3	>=	3	−
PUBCHEM_CACTVS_TAUTO_COUNT	is defined		−
SCPnoLinkerBonds	>=	1	−
Choose property			

Conjunctive

Molecule count: 447



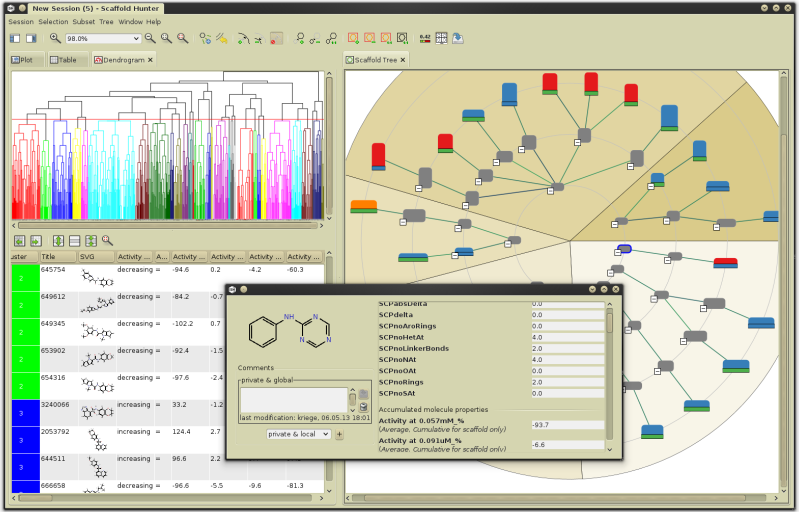
Root
- Selection 1
- **Selection 2**
- decreasing
  - filter(decreasing)
- increasing
  - filter(increasing)

□ List view

Total Selection: 6

In Current View: 1

Multiple Views & Tooltip


New Session (5) - Scaffold Hunter
Session Selection Subset Tree Window Help
98.0%
Plot
Table
Dendrogram
Scaffold Tree

uster	Title	SVG	Activity ...	A...	Activity ...	Activity ...	Activity ...	Activity ...
2	645754		decreasing	=	-94.6	0.2	-4.2	-60.3
2	649612		decreasing	=	-84.2	-0.7		
2	649345		decreasing	=	-102.2	0.7		
2	653902		decreasing	=	-92.4	-1.5		
2	654316		decreasing	=	-97.6	-2.4		
3	3240066		increasing	=	33.2	-1.2		
3	2053792		increasing	=	124.4	2.7		
3	644511		increasing	=	96.6	2.2		
	666658		decreasing	=	-96.6	-5.5	-9.6	-81.3

Comments

private & global

last modification: kriege, 06.05.13 18:01

private & local +

SCPabsDelta	0.0
SCPdelta	0.0
SCPnoAroRings	0.0
SCPnoHetAt	4.0
SCPnoLinkerBonds	2.0
SCPnoNAt	4.0
SCPnoOAt	0.0
SCPnoRings	2.0
SCPnoSAt	0.0

Accumulated molecule properties

Activity at 0.057mM_% *(Average, Cumulative for scaffold only)*	-93.7
Activity at 0.091uM_% *(Average, Cumulative for scaffold only)*	-6.6

Figure: Split View: Dendrogram, Scaffold Tree & Tooltip

Realization & Technical Details

- Freely available under GNU GPL v3
- Implemented in Java for platform independent use
- Modular software architecture:
  - Seamless integration of novel views and analysis features
  - Plugin system for data sources and property calculation

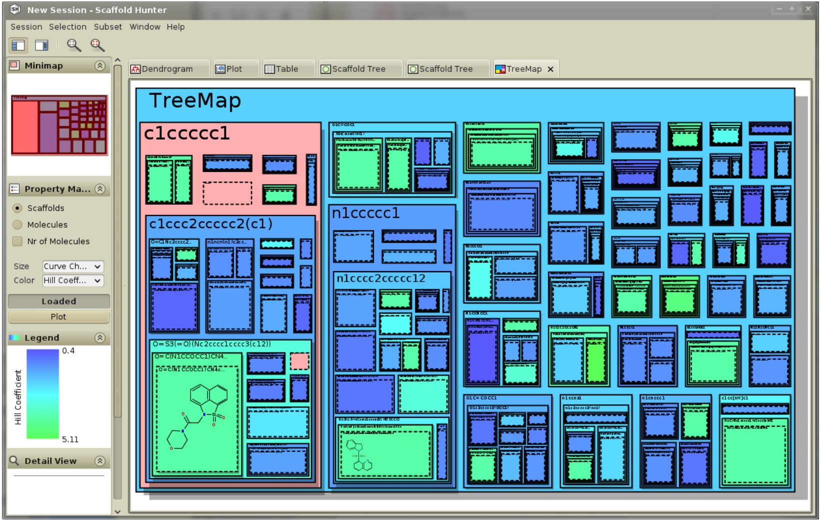
Realization & Technical Details

- Freely available under GNU GPL v3
- Implemented in Java for platform independent use
- Modular software architecture:
  - Seamless integration of novel views and analysis features
  - Plugin system for data sources and property calculation

Toolkits & Database Support

- **Chemistry Development Kit (CDK):** Various cheminformatics tasks
- **Piccolo2D:** Zoomable user interfaces
- **Batik:** SVG support
- **Hibernate:** Object-relational mapping
- **MySQL/HSQLDB:** Back-end databases

Future Work: Scaffold TreeMaps



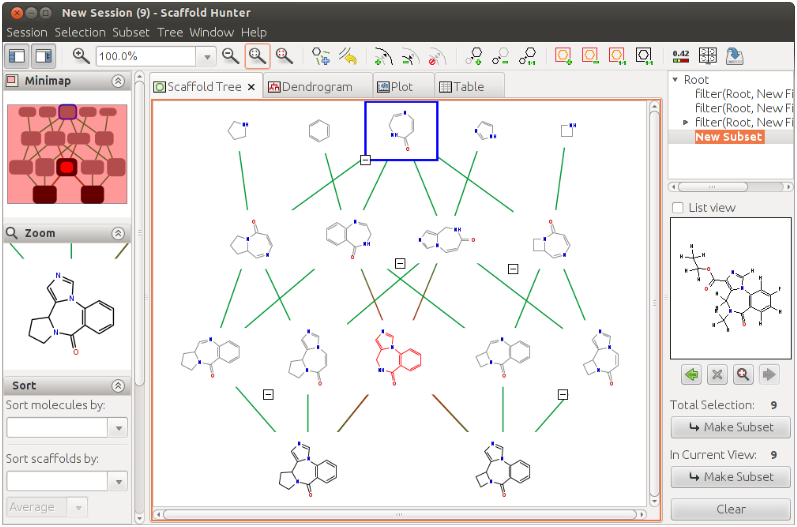
● Scaffolds
○ Molecules
□ Nr of Molecules

Size: Curve Ch...
Color: Hill Coeff...

Loaded
Plot

- Space-filling approach to visualize scaffold trees
- Google Summer of Code project 2013: Jeroen Lappenschaar

Future Work: Scaffold Networks



Total Selection: 9

In Current View: 9

- Visualization of multiple parent scaffolds (Sugiyama layout)
- Dynamic filtering of networks

Conclusion

- Exploratory visual analysis of chemical compound databases
- Clustering and classification of molecular datasets
- Multiple complementary interconnected views

Conclusion

- Exploratory visual analysis of chemical compound databases
- Clustering and classification of molecular datasets
- Multiple complementary interconnected views

Development & Acknowledgements

- **TU Dortmund, Prof. Mutzel:** Nils Kriege, Till Schäfer
- **University of Sydney:** Dr. Karsten Klein
- **GSoC 2013:** Jeroen Lappenschaar
- **Cooperation:**
  - MPI for Molecular Physiology Dortmund (Prof. Waldmann)
  - Dr. Koch, Computational Molecular Design, TU Dortmund