



Organizing and Clustering Molecule Libraries Using SARvision

A product of



www.ChemApps.com

support@chemapps.com

phone: 858-259-8161

Welcome to SARvision

SARvision2.2 has a number of features that allow for facile organization of large groups of molecules. Currently **SARvision2.2** will easily organize up to 10,000 molecules and **SARvision2.2Plus** will organize up to 100,000 molecules. **SARvision** groups molecules based on structural motifs in the molecules. These structural motifs, or *Scaffolds*, are identified using a unique ruled based method that approximates the decision making process of the medicinal chemist as closely as is possible for a nonhuman algorithm. These scaffolds are then organized hierarchically into a tree structure to allow the user to easily navigate from trivial scaffolds at the base of the tree (*root scaffolds*) extending out to complex scaffolds located at the end of each tree (*leaf scaffolds*). We have found that this is the most intuitive way for the user to organize molecular data that allows easy identification of interesting motifs in the tree and to group molecules into relevant families. The user can easily navigate the tree, edit the tree and/or prune the tree to create the most useful scaffold dataset possible. Note that a molecule can belong to more than one scaffold. Each scaffold thus represents a unique view of the data.

This is a product of several years of development and research. The SARvision project has benefited from much feedback from the user. Please contact us (support@chemapps.com) to add any recommendations that you may have for future releases or to report any bugs that you may encounter. We want to hear from you.

1. Load a dataset into SARvision

This guide will outline features available for organizing compounds. CombiBlock's (www.bioblocks.com) database is used as an example and is available in the install directory ([bioblocks.sdf](#)) but any compound dataset can be used.

To get started, import an sd file into **SARvision** under **File:Import: Molecule Dataset(sdf)**. This file format is the most common format employed for sharing molecule datasets between programs and users. It should be available for download from your database. If not contact your informatics group. For this example, we use the commercially available chemical library provided by BioBlocks ([BioBlocks.sdf](#)).

1. Load Molecule Dataset as an SD file.

All Data Fields In File Imported as Column Data.

ID	compound_id	quantity_gram	price	mol
1	AA001	2.500000	240.00	cis-
2	AA002	2.500000	190.00	
3	AA003	2.500000	270.00	tran
4	AA004	2.500000	190.00	

Once loaded, molecules and associated data are shown in the table on the right. Under the menu **Table: Display Options**, the color of the displayed molecules can be changed. By right clicking on individual columns, the data can be formatted, sorted or deleted. The columns can be 'drag and dropped' in any order you desire and by double clicking the column header, locked to the left of the table.

2. Build a Scaffold Tree

SARvision has a number of options to control the scaffold tree building algorithm. To access these options, go under **Tree: Scaffold Id Options** and select the desired options for identifying scaffolds. These are options to control the speed and the accuracy of the scaffold perception algorithm.

Select **Build all Layers** so that Scaffold Identification will exhaustively calculate all layers of the scaffold tree from the top to the bottom. When unselected, only parts of the scaffold tree are calculated at a given time.

Select **Build in New Folder** to build the scaffold tree in a new folder separate from any existing folders or scaffolds.

Unselect **Fast Build** option. Selection of this option makes several approximations to accelerate scaffold identification. These assumptions include ignoring trivial ring systems at all levels of the tree (such as cyclohexane and benzene) and motifs found in chains (such as an amide).

Select **Delete Trivial Scaffolds** to prevent the generation of scaffolds beginning with trivial functionalities such as benzene or cyclopropyl.

Discard Scaffolds without Rings. Selection of this option causes the tree builder algorithm to ignore scaffolds that are composed only of atom chains. Unless the user is specifically interested in this type of non-cyclic scaffolds, it is recommended that this option not be selected.

Set **Discard Families < 3 Molecules.** This is the size of the Scaffold groupings at which point **SARvision** stops calculating addition children. For instance, 'expanding' a scaffold further to produce more complex children scaffolds will not occur unless the Family size is greater than that set for this option.

Set **Sampling Size** to 100 %. Since **SARvisionPlus** can analyze up to 100,000 molecules, this option allows the user to work with subsets of the database when identifying scaffolds to speed up the tree building procedure.

Unselect **Consider Only Scaffolds in Active Property Range.** Selecting this option would only construct a scaffold tree utilizing molecules that belong to the Active Property ranges set by the user (**Data: Set Active Properties**).

Finally, when the desired options have been selected, build the scaffold under **Tree: Identify Scaffolds.**

The screenshot displays the SARvision 2.2 interface with the 'Identify Scaffolds...' dialog box open. The 'Tree Building Options' tab is selected, showing the following settings:

- Build All Layers
- Fast Build
- Build Tree In New Folder
- Delete Trivial Scaffolds (Recommended)
- Discard Scaffolds without Rings (Recommended)
- Discard Families < 3 Molecules
- Molecule Subset: 100 %
- Consider Only Scaffolds in Active Property Range

Annotations on the dialog box include:

- 2a. Set Scaffold Identification Options
- Build Complete Tree.
- Ignore Scaffolds Like benzene and cyclohexane
- Ignore Scaffolds that have no rings.
- Build Tree using a Subset of Molecules.
- 2b. Build Hierarchical Scaffold Tree.
- Do not include scaffolds with <3 molecules. Do not build children scaffolds unless moleculeset >= 4.

The Scaffold Tree structure is shown on the right, with a 'Data Set' table listing the scaffolds:

Structure	ID	compound_id
<chem>NC(=O)Cc1ccccc1</chem>	21	AA022
<chem>NC(=O)Cc1ccc(Cl)cc1</chem>	22	AA023
<chem>NC(=O)Cc1ccc(Cl)cc1</chem>	23	AA024
<chem>NC(=O)Cc1ccc(F)cc1</chem>	24	AA025
<chem>NC(=O)Cc1ccccc1</chem>	25	AA026

Construction of a Scaffold Tree can take from a few seconds up to hours depending on the size and complexity of the dataset. The Scaffolds are arranged hierarchically

into the tree such that each parent scaffold is a substructure of all of its children scaffolds, the children substructure of the grand children and so forth.

3. Identify Scaffolds with names and Correlate these to each Molecule in the set.

Select [Tree: AutoName Tree Nodes](#) to name the scaffolds in the tree based on the position in the tree. Each child scaffold takes on the name of the parent with the addition of ".1", ".2"... for each. If the parent is "1", then the children are named in order "1.1", "1.2", "1.3"...

To correlate molecules in the table back to scaffolds in the tree, select [Tree: Correlate Molecules to Scaffolds](#). This creates a column named 'Scaffolds' in the table and adds the names of the Scaffolds to which it belongs for each molecule in the set.

The screenshot displays the SARvision 2.2 interface with the 'BioBlocks.sdf' dataset. On the left, a scaffold tree is shown with three nodes: ScaffoldTree1.1 (45), ScaffoldTree1.2 (40), and ScaffoldTree1.3 (22). A callout box labeled '3a. AutoName Scaffolds by Tree Position.' points to the 'Auto-name Tree Nodes' menu option. In the center, a 'Data' menu is open, highlighting 'Correlate Molecules to Scaffolds'. A callout box labeled '3b. Correlate Molecules Back to Scaffold. Add Scaffold Column.' points to this menu option. On the right, a data table is shown with columns for 'Molecules' (ID#), 'Scaffolds', and 'ID#'. The table contains four rows of chemical structures, each associated with the scaffold name 'ScaffoldTree1 ScaffoldTree1.1' and an ID number (21, 22, 23, 24).

Molecules	Scaffolds	ID#
<chem>NC(Cc1ccccc1)C(=O)O</chem>	ScaffoldTree1 ScaffoldTree1.1	21
<chem>NC(Cc1ccc(Cl)cc1)C(=O)O</chem>	ScaffoldTree1 ScaffoldTree1.1	22
<chem>NC(Cc1ccc(Cl)cc1)C(=O)O</chem>	ScaffoldTree1 ScaffoldTree1.1	23
<chem>NC(Cc1ccc(F)cc1)C(=O)O</chem>	ScaffoldTree1 ScaffoldTree1.1	24

4. Build Scaffold Tree from Scaffold root

Often times, the medicinal chemist knows the type of motifs that are relevant to a given project. In these cases, a tree can be build using a scaffold of interest as the root of the tree to build children scaffolds. For this example, click on the draw

The image displays two screenshots of the SARvision 2.2 software interface, illustrating the process of building a scaffold tree. The left screenshot shows a scaffold tree with a root node 'ScaffoldTree 362' containing a drawn scaffold (NH-OH). A context menu is open over this scaffold, with 'Add Scaffold' selected. The right screenshot shows the same scaffold tree with a child scaffold (a benzene ring) added under node 83. A table on the right lists molecules that contain both scaffolds as substructures.

5a. Draw a Scaffold.

5b. Add A Child Scaffold.

An Example of a NonHierarchical Scaffold Tree.

Molecules shown are those that contain both Scaffolds as Substructures.

Structure	Sc
<chem>CC(C)C(N)C(O)C(=O)O</chem>	1
<chem>CC(C)C(N)C(O)C(=O)O</chem>	2
<chem>C1CCCCC1C(N)C(O)C(=O)O</chem>	3