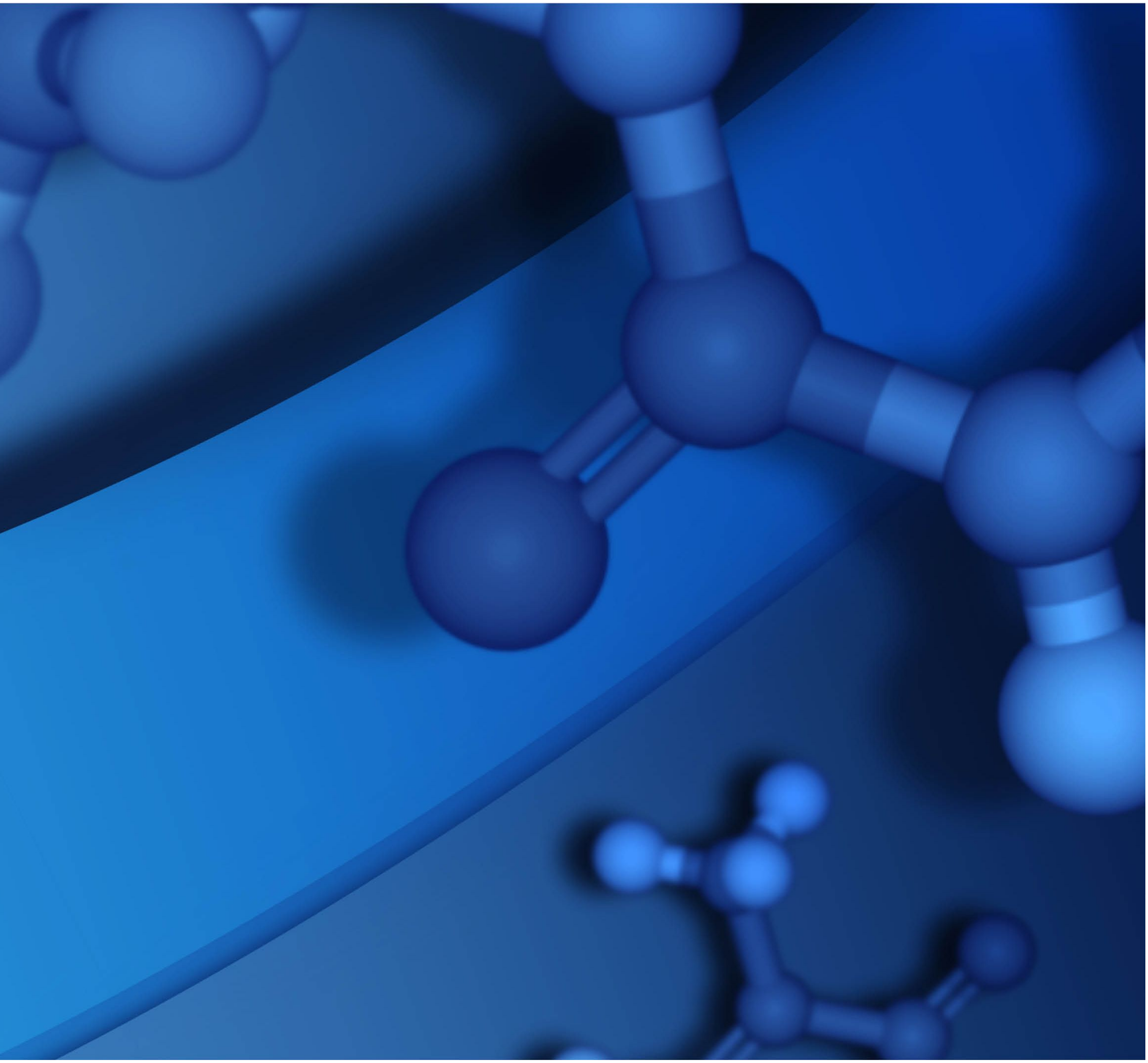
# CHEMICAL REPRESENTATION

2021



## Copyright Notice

## Acknowledgments and References

To print photographs or files of computational results (figures and/or data) obtained by using Dassault Systèmes software, acknowledge the source in an appropriate format. For example:

> "Computational results were obtained by using Dassault Systèmes BIOVIA software programs. BIOVIA Direct was used to perform the calculations and to generate the graphical results."

Dassault Systèmes may grant permission to republish or reprint its copyrighted materials. Requests should be submitted to Dassault Systèmes Customer Support, either by visiting https://www.3ds.com/support/ and clicking **Call us** or **Submit a request**, or by writing to:

Dassault Systèmes Customer Support
10, Rue Marcel Dassault
78140 Vélizy-Villacoublay
FRANCE

# Contents

# Chapter 1:
# Overview

This guide provides the information on BIOVIA chemical representation that you need for chemical structure and reaction registration to chemical databases.

## Audience for this Guide

This guide is for people who:

- Administer BIOVIA chemical databases
- Create and maintain business rules for registration of chemical structures
- Create applications that allow scientist end-users to search BIOVIA chemical databases
- Provide technical support to scientists who search BIOVIA chemical databases

## Prerequisite Knowledge

This guide assumes that you are familiar with:

- Introductory Organic Chemistry
- Your operating system
    - Microsoft Windows Server
    - Linux Server
    - Sun Solaris Server

Experience with BIOVIA Draw will also be helpful to you.

Knowledge of the Oracle RDBMS, especially Oracle SQL*Plus and Oracle database administration, is helpful but is not required.

## Related BIOVIA Documentation

The following table lists related, useful documentation.

**Table 1** Related BIOVIA Documentation

| Document | Purpose |
|---|---|
| *BIOVIA Draw Help* | Drawing chemical structures and reactions |
| *BIOVIA Draw Configuration Guide* | Developing applications that use BIOVIA Draw |
| *BIOVIA Draw Developers Guide* | Developing applications that use BIOVIA Draw |
| *BIOVIA Draw API Reference* | Developing customized interfaces for BIOVIA Draw |
| *BIOVIA CTFile Formats* | Describes the file formats used to represent structures.<br>This file is included in the *BIOVIADirect_2021_Documentation.zip* file. |

| Document | Purpose |
|---|---|
| *BIOVIA Direct Administration Guide* | Configuring BIOVIA Direct for your site |
| *BIOVIA Direct Developers Guide* | Developing applications using BIOVIA Direct databases |
| *BIOVIA Direct Reference Guide* | Developing applications using BIOVIA Direct |
| *BIOVIA Direct Linux Installation and Configuration Guide*<br>*BIOVIA Direct Windows Installation and Configuration Guide* | Installing and configuring BIOVIA Direct for your platform |
| *BIOVIA Direct Secure Extproc Listener Guide* | Setting up and configuring a secure extproc listener |

# Chapter 2: Molecule Representation

## Substances, Structures and Fragments

In describing the representation of chemical entities, it is important to distinguish the following terms:

- A chemical structure (or molecule) is the graphical depiction of a chemical compound that can be drawn and displayed in BIOVIA Draw.
- A substance is matter of particular composition that is reasonably homogeneous. A substance can have associated sets of intrinsic properties. A substance can consist of one or more chemical structures.
- A fragment is a single, disconnected atom; or a set of connected atoms and bonds. A structure can consist of zero, one, or more fragments.

**Note:** A chemical structure that consists of zero fragments (that is, zero atoms and bonds) is called a no-structure. For more information on no-structures, see Structural Uncertainty on page 31.

## The BIOVIA Periodic Table

The BIOVIA periodic table or Ptable defines the atom symbols that you can register to your database and use in graphical search queries. The Ptable also provides a way to specify custom atomic weights for the standard elements, hydrogen through lawrencium. The Ptable consists of three sections:

- Entries that provide the custom atomic weight for elements H through Lr.
- Elements that specify additional man-made chemical elements, Rf through Uuo.
- Entries that specify pseudoatoms, which are atom symbols that do not correspond to any of the chemical elements.

You can add your own pseudoatoms and man-made elements to the Ptable in your database and you can customize atomic weights for the standard elements. For detailed information on the BIOVIA Ptable, see Customizing the BIOVIA Ptable on page 206.

**Note:** Do not confuse pseudoatoms with abbreviated structures. An abbreviated structure contains within it all the atoms and bonds of the underlying structure, that is, it contains *the complete connection table (CTAB)*. In contrast, a pseudoatom is a single atom with no underlying connection table. For information on abbreviated structures, see Abbreviated Structures on page 30.

## Atom Properties

### Charges, Radicals, and Isotopes

Use BIOVIA Draw to specify atomic charges, radicals, and isotopes. You can specify a charge at an atom ranging from -15 to +15.

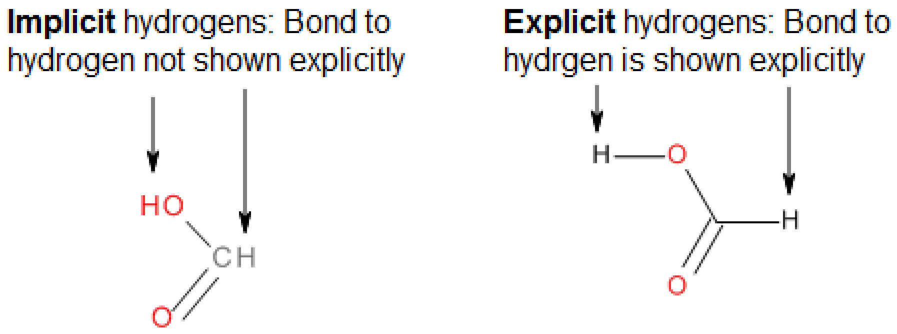
Specify radicals at an atom to indicate a number of unpaired electrons and electronic spin states:

- use a period (.) for unpaired electrons.
- use a colon (:) for two unpaired electrons, singlet state.
- use two carets (^^) for two unpaired electrons, triplet state.

You can add an isotope label to an atom ranging from 99 below to 99 above the atomic weight in the Ptable. Atoms without isotopic labels represent the natural mixture of isotopes.

## Valences and Implicit Hydrogens

The valence property of an atom is the number of covalent bonds that can attach. If you draw a structure with fewer non-hydrogen attachments than the valence, hydrogens are assumed to be present at unfilled valences. These are implicit hydrogens. An implicit hydrogen is a hydrogen that is either assumed to be present (invisible) or attached to an atom by an invisible bond. An explicit hydrogen is a hydrogen that is attached to an atom by a visible bond.


**Implicit** hydrogens: Bond to hydrogen not shown explicitly
HO
CH
O
**Explicit** hydrogens: Bond to hydrgen is shown explicitly
H
O
H
O

You can use a BIOVIA Draw setting to hide or display implicit hydrogens.

## Default Valences

Each atom has one or more default valences. The number of implicit hydrogens at an atom is equal to the allowed valence minus the number of bonds to non-hydrogen atoms, up to the next allowed valence. For example, a sulfur atom with one bond to a non-hydrogen atom has one implicit hydrogen, and a sulfur atom with two bonds has zero implicit hydrogens, because the next highest valence is 2. A sulfur atom with three bonds has one implicit hydrogen, because the next highest valence is 4.

The table that follows shows allowed default valences for neutral main group elements:

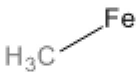
| 1a | | 2a | 3a | | 4a | | 5a | | 6a | | 7a | | 8 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | (1) | | | | | | | | | | | | He | (0) |
| | | | B | (3) | C | (4) | N | (3) | O | (2) | F | (1) | Ne | (0) |
| | | | | | Si | (4) | P | (3,5) | S | (2,4,6) | Cl | (1,3,5,7) | Ar | (0) |
| | | | | | | | As | (3,5) | Se | (2,4,6) | Br | (1,3,5,7) | Kr | (0) |
| | | | | | | | | | Te | (2,4,6) | I | (1,3,5,7) | Xe | (0) |
| | | | | | | | | | | | At | (1,3,5,7) | Rn | (0) |
| | | | | | | | | | | | | | | |

Implicit hydrogens are never added to metal atoms or ions (implied valence is zero), with the exception of Al(-1) which has a default valence of 4.

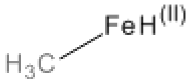
## Explicit Valence

You can use explicit valence to assign a non-default valence to an atom, and thereby control the number of implicit hydrogens at an atom. For example:

Use explicit valence to specify implicit hydrogens on-metal hydrides. By default, the valence for a metal is set to the number of bonds that are attached to non-hydrogen atoms. If you assign an explicit valence, however, the atom has implicit hydrogens at unfilled valences. The following example uses explicit valence to specify the correct number of hydrogens for an organometallic compound:



Default valence=1
1 non-hydrogen attachment
No implicit hydrogen



Explicit Valence=2
1 non-hydrogen attachment
1 implicit hydrogen

## Rules for Calculation of Valence and Implicit Hydrogen

A radical subtracts 1 from the valence. For example, C^ has valence 3, Na^ has valence 0, and Mg^ has valence 1.

Charge has the following effects on valence:

- A charged atom has the same valence as its isoelectronic neutral atom.

  For example, a single carbon atom with a charge of +1 is isoelectronic with boron, and therefore has three implicit hydrogens. A nitrogen atom with a charge of +1 is isoelectronic with carbon and therefore has four implicit hydrogens.
- Charged atoms in group 3a-7a elements with atomic number >20 are an exception because of the presence of the transition elements. If the charge shifts the element into or through the transition region of the table, then the atom will have the behavior of a metal and it will have no implicit hydrogens.

# Bond Types

The following types of bonds are represented in BIOVIA databases: single bonds (including bonds to salts and bonds between mixtures of two neutral compounds), double bonds, and triple bonds.

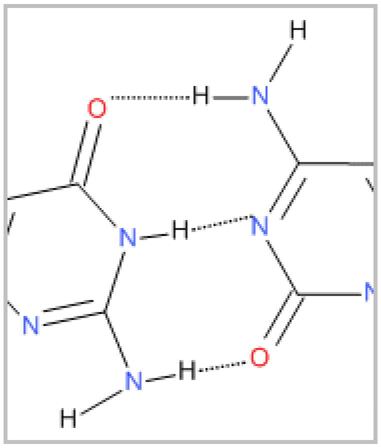
The following bond types denote stereochemistry:

- A Double Either bond represents a double bond that might be either cis or trans. See Cis and Trans Stereochemistry on page 133.
- Up and Down stereo bonds can be used to specify tetrahedral stereochemistry and/or stereochemistry of allenes and biaryls. See Tetrahedral Stereochemistry on page 9 and Stereochemistry of Allenes and Biaryls on page 24.

Additional bond types are available for special purposes:

- Either stereo bonds: This bond type is for display only and is equivalent to no stereo bond. For more information on usage of stereo bonds in representation of stereochemistry, see Tetrahedral Stereochemistry on page 9 and Stereochemistry of Allenes and Biaryls on page 24.

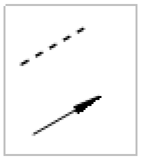
- Query bonds (Any, Aromatic, Single/Double, Single/Aromatic) are used in substructure search queries, but cannot be registered to a database. For information on query bond types, see Query Features on Atoms and Bonds on page 138.
- Zero-order bonds, which are bonds that do not affect valence. Zero-order bonds can be of two types:
  - hydrogen bond, such as this partial depiction of Guanine and Cytosine



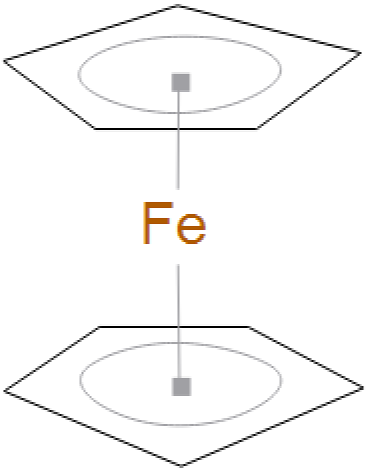
  - coordination bond. A coordination bond can be displayed in the coordination style or the dative style. If a coordination bond is not set to the coordination display style or the dative display style, it displays like a normal single bond.

  Structures with these bond types can be registered.

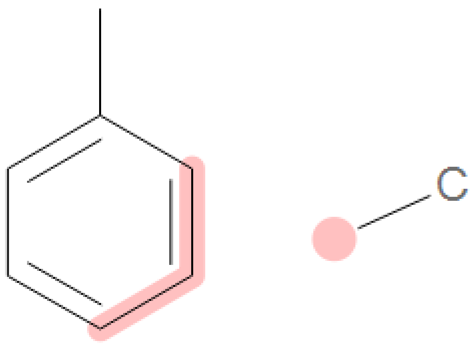
  This graphic represents a hydrogen bond above a coordination bond (dative).



- Multi-endpoint bonds can be of any bond type, are registerable, and support two kinds of one-to-many association:
  - a haptic bond to all atoms in a collection, such as this metallocene.



- a variable attachment bond to one of any atom in a collection. For example, this can be used to represent structures with an unspecified location for a substituent, such as this indeterminate isomer of xylene.



## Aromaticity

Aromaticity is perceived by counting the number of pi electrons in the ring and then applying the Hückel Rule (4N+2). Each atom in the ring is typed for hybridization and pi-electron count.

Check for atoms which cannot be aromatic. These are checks for hybridization and invalid valence. The presence of one or more rejected atoms will prevent the ring from being aromatic.

The following atoms contribute one pi electron:

- A double bond is two pi electrons, so for each double bond in a ring, the atom at each endpoint contributes one electron.
- The same applies for triple bonds because the additional electrons are in a different plane.
- Additionally, bonds that have been previously marked as aromatic from the perception of a neighboring fused ring will have one electron contributed by each endpoint atom.
- An atom with a radical will contribute one pi electron.

Remaining atoms should contain only single bonds and are checked for the presence of a lone pair. The following atoms (and their isoelectronic equivalents) can contribute a lone pair: N, O ,P, As, S, Se, Te. If an atom contributes a lone pair, that adds 2 pi electrons to the ring.

Any atoms that remain are checked to see if they contain an empty p-orbital which can act as a pi-electron acceptor. Only positively charged non-metals are allowed here. Examples include C+, Si+ and Ge+. These atoms contribute 0 pi electrons to the ring.

If an atom connect be classified above as contributing 0, 1, or 2 electrons, then the atom is rejected and the ring cannot be aromatic.

If a ring contains no rejected atoms, then electron contributions are summed and the Hückel Rule is applied. Electron counts of 2, 6, 10, 14, 18, and so on, indicate an aromatic ring.

Pairs of fused rings are considered for azulene-type cases. When the outer-ring is perceived as aromatic then both inner rings (including the fusion bond) will be marked aromatic. Iteration is used for larger fused systems where marking one ring aromatic can trigger the aromaticity in a neighboring ring with shares a bond.

## Tautomers

Tautomers are sets of structural isomers which can readily interconvert. This interconversion usually occurs by the movement of an explicit hydrogen atom or a formal charge from one heavy atom to another and the accompanying changes in bond orders to compensate.:

In Pipeline Pilot, a number of components are available for users to identify and enumerate tautomers. See the Pipeline Pilot *Advanced Chemistry Guide* for more details.

In Direct, one specific set of Tautomer options is used for determining which structures are to be determined equivalent when the tautomeric flexmatch switch (TAU) is enabled.

- ConsiderCarbonAsDonor: BondedToAcceptor(1_3) - Carbon atom needs to be attached to at least one atom which is attached to an acceptor
- MakeAllSp2AtomsAcceptors - True - Marks all $sp^2$-hybridized atoms as acceptors
- Amides Tautomerization - Tautomerize all amide groups
- Perceive Charge Tautomerization - False - Mobile charges are not considered when identifying tautomer information

A tautomeric fragment is defined as a set of contiguous atoms which can act as tautomeric donors or acceptors.Tautomer groups must contain the same number of hydrogens in the tautomeric region, with a tolerance limit. The tolerance limit is the sum of the absolute values of charges plus the sum of the number of radicals plus the number of metal bonds. Diradicals count as two radicals. It is possible have multiple tautomeric fragments in a single molecule. The *Identify Tautomeric Fragments* component in Pipeline Pilot can be used to provide the information the tautomeric fragments present in a structure.

> **Note:** It is possible for one structure to see itself as a tautomer of another but not vice versa. This asymmetry in tautomeric perception is caused by restrictions in the donor and acceptor definitions above. The post-tautomerization state may not be re-perceived as being in a tautomeric group. When this occurs users may see asymmetry in the use of the tautomeric flexmatch switch (TAU). Structure A may map Structure B, but not vice versa.

## Salts

Salts can be represented in different ways. For example, you can represent a salt as a single chemical structure that comprises multiple fragments (parent structure, counterions, and water of hydration). Alternatively, you might want to store only the parent compound as a chemical structure and store information on counterions and fragments in a separate database field.

If you choose to represent salts as a single chemical structure, the following fragments will be perceived as counterions and hydrates in searching and registration:

- Alkali metals: Li, Na, K
- Halogens: F, Cl, Br
- Carbonates, nitrates, nitrites, perchlorates, and sulfates
- Water (hydrates)

To be perceived as the counterion or hydrate of a salt, an atom or fragment must be listed in the BIOVIA Salts definition. You can add your own definitions of counterions to your company's Salts definition. For detailed information on the BIOVIA Salts definition, see Introduction on page 215.

> **Note:** Exact (flexmatch) searching uses the BIOVIA Salts definition. For more information, see Exact Search (Flexmatch) on page 107.

## Tetrahedral Stereochemistry

You can specify the stereoconfiguration at asymmetric tetrahedral centers (chiral centers) at the following atom types (and their isoelectronic equivalents): C, N, Si, P, S, As, Se, or Te.

Trivalent Group V (nitrogen, phosphorus, or isoelectronic equivalents) atoms are not considered to be valid tetrahedral stereocenters, with two exceptions – trivalent nitrogen and trivalent phosphorous.

Trivalent nitrogen must have three single bonds each in a ring, and no more than two of those can be connected to sp2 atoms. Trivalent phosphorus must have three single bonds none of which are connected to hydrogen.

These nitrogen and phosphorus atoms are considered to be valid tetrahedral centers. Marked atoms which are not considered to be valid stereocenters will have bond marks removed on registration, and bond marks are ignored for stereospecific search queries.

To summarize stereochemical representation:

- *No stereo bond pointing to the atom at a stereogenic center:* To be perceived as a *defined* stereogenic center, the atom must be at the narrow end of the Up or Down stereo bond.

  The lack of an Up or Down stereo bond to a stereogenic atom, or the presence of the wide end of an Up or Down bond, implies that no information is known about the configuration of that stereogenic center. It could be either of two stereoisomers or a mixture of the two. A stereogenic center without an Up or Down bond is referred to as an undefined stereogenic center.

  **Note:** For more information on stereochemical perception, see Perception of Stereoconfiguration at Tetrahedral Centers on page 222.

- *Either stereo bond pointing to atom at a stereogenic center:* Has the same meaning as no stereo bond. That is, the configuration at the stereogenic center is *undefined.*
- *The narrow end of an Up or Down stereo bond at a stereogenic center:* Implies that something is known about the configuration at that center; either its absolute configuration or its configuration relative to other stereogenic centers that are marked with stereo bonds. A stereogenic center that is marked with an Up or Down stereo bond is a *defined* stereogenic center.
- *Stereogroup label:* Provides information on the stereoconfiguration of a *defined* stereogenic center (that is, a stereogenic center at the narrow end of an Up or Down stereo bond). For more information on stereogroup labels, see Grouping Related Stereogenic Centers on page 9.
- *Chiral label:* Summarizes stereogroup information for an entire structure by indicating whether more than one stereogroup is present. For more information, see Chiral Labels on page 13.

### Grouping Related Stereogenic Centers

In many cases, you might not know the absolute configuration of two stereogenic centers, but might have some information about their relative configuration. For example, you might know the relative configuration of the two centers. Alternatively, you might know that you have a mixture of stereoisomers. You can define stereochemical groups to show that the configurations of the stereogenic centers are related. To show this, you put the related stereogenic centers into the same OR or AND group:

- An *OR group* is a set of stereogenic centers in which you know the relative configuration of all the stereogenic centers in that group. Only one enantiomer is present, but you do not know which one.
- An *AND group* is a set of stereogenic centers that represents a mixture of two possible stereoisomers in which the relative configuration of all the centers in the group is known.

The index numbers on the group labels (`or1, or2, or3..., &1, &2, &3...`) show which stereogenic centers of a given type (OR or AND) are related, and have no other significance. That is, assigning identical index numbers to OR and AND stereogroup labels (for example, `or1` and &1) does not imply a stereochemical relationship between the groups. The stereochemical relationship between groups of the same type with different index numbers (for example, `or1, or2, or3`...,) is also undefined.

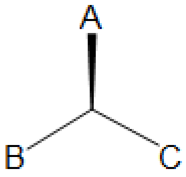
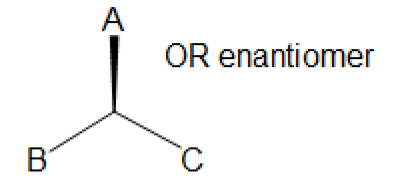
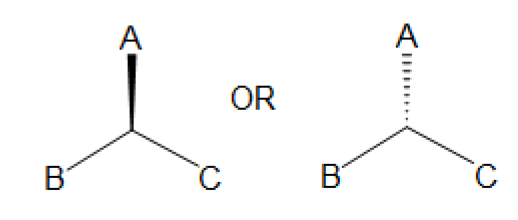
The *ABS stereogroup* contains all stereogenic centers for which the absolute configuration is known.

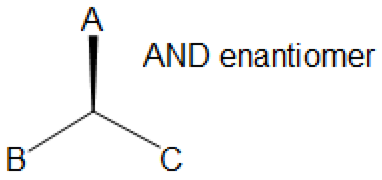
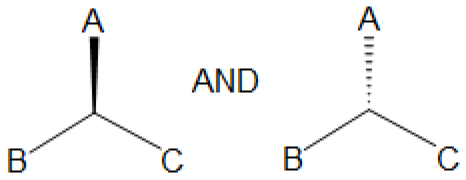
Where possible Pipeline Pilot uses the original representation because most structures are V2000 with Chiral flag on or off. Using V2000 with the Chiral Flag on is equivalent to having all centers in a single ABS group and using V2000 with the Chiral Flag off is equivalent to having all centers in a single AND group. For more information on the original Accelrys representation, see Original Representation of Tetrahedral Stereochemistry on page 256.

A given structure can have any combination of OR, AND, and ABS stereogroups. For examples, see Examples of Stereogroups on page 17.

## Structures with One Stereogenic Center

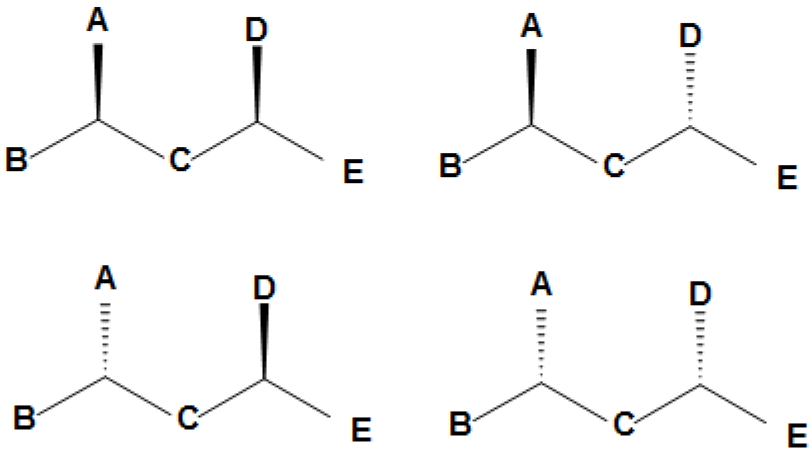
The table that follows shows how you can use stereogroups to show how much information you have about the configuration of a structure that contains one stereogenic center. The chiral label (blank, OR enantiomer, or AND enantiomer) indicates whether the stereogenic center is in the ABS, or1, or &1 stereogroup.

| Structure | Interpretation |
|---|---|
|  | A single stereoisomer whose absolute configuration is known (ABS stereogroup). |
|  | A single stereoisomer whose configuration is unknown (OR stereogroup)<br> <br>The chiral label "OR enantiomer" on the structure shows that all stereogenic centers on the structure are members of a single OR stereogroup, `or1`. See Chiral Labels on page 13. |

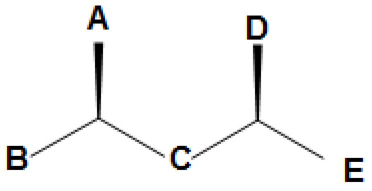
| Structure | Interpretation |
|---|---|
| <br> | A mixture of enantiomers (AND stereogroup)<br><br> <br>The chiral structure label "AND enantiomer" on the structure shows all stereogenic centers on the structure are members of a single AND stereogroup &1. |

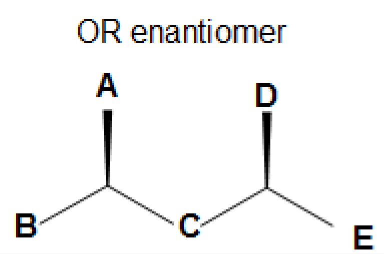
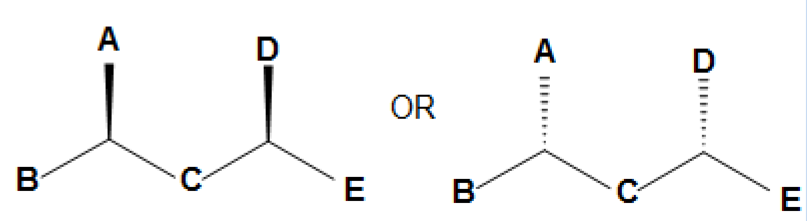
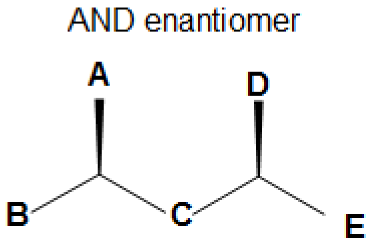
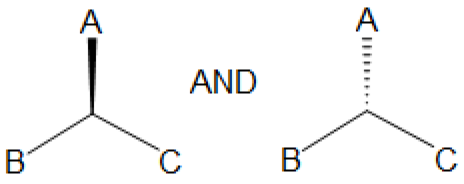
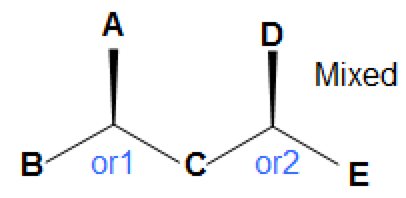
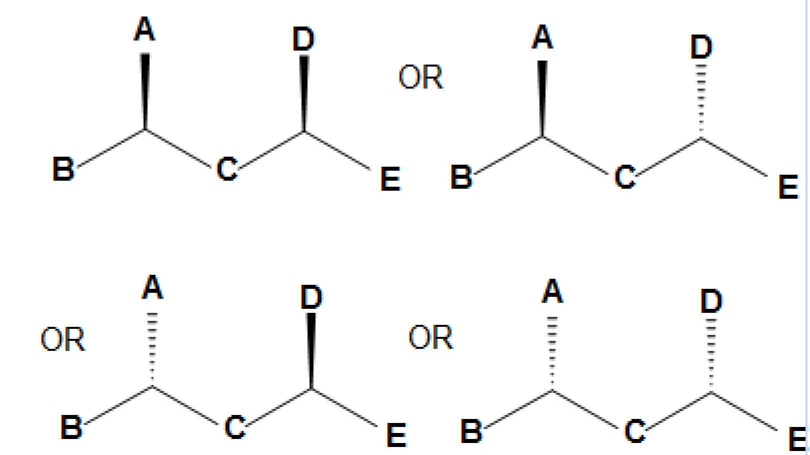
## Structures with Multiple Stereogenic Centers

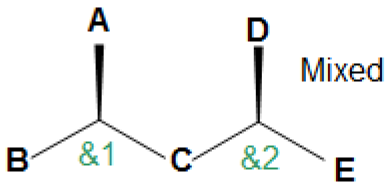
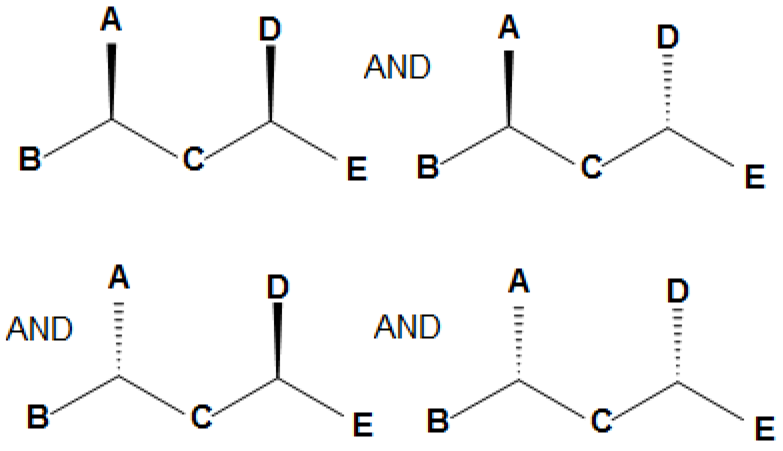
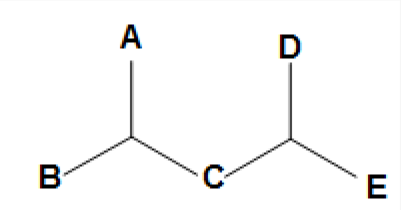
Structures with multiple stereogenic centers introduce the possibility of incomplete information on the relationships between the configurations of the various stereogenic centers. For example, a structure with two stereogenic centers can have four possible stereoisomers:



The table that follows shows how you can use stereogroups to illustrate how much information you have about the configuration of a structure that contains more than one stereogenic center. The *chiral label* (blank, `OR enantiomer`, `AND enantiomer`, or `Mixed`) indicates whether the stereogenic centers in the structure are in more than one stereogroup.

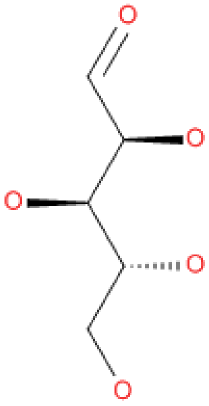
| Structure | Interpretation |
|---|---|
| <br> | A single stereoisomer For which you know the absolute configuration at both stereogenic centers. |

| Structure | Interpretation |
|---|---|
| <br> | A single unknown enantiomer in which the relative configuration of the stereogenic centers is known. That is, you know you have either of the two stereoisomers that follow:<br><br> <br>The chiral label "OR enantiomer" on the structure shows that all stereogenic centers on the structure are members of a single OR subgroup, or1. See Chiral Labels on page 13. |
| <br> | A mixture of two stereoisomers with the same relative configuration. For example:<br><br> <br>The chiral label "AND enantiomer" on the structure shows that all stereogenic centers on the structure are members of a single AND stereogroup, &1. For more information, see Chiral Labels on page 13. |
| <br> | Any one of the four stereoisomers that are possible for structures with two stereogenic centers:<br><br> A<br>D<br>B<br>C<br>E<br>OR<br>A<br>D<br>B<br>C<br>E<br>OR<br>A<br>D<br>B<br>C<br>E<br>OR<br>A<br>D<br>B<br>C<br>E<br>The chiral label "Mixed" on the structure shows that the structure has more than one stereogroup, in this case, or1 and or2. |

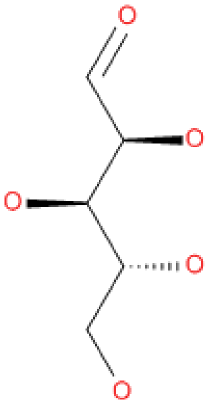
| Structure | Interpretation |
|---|---|
|  | A mixture of all four stereoisomers:<br> A<br>D<br>B<br>C<br>E<br>AND<br>A<br>D<br>B<br>C<br>E<br>AND<br>A<br>D<br>B<br>C<br>E<br>AND<br>A<br>D<br>B<br>C<br>E<br>The chiral label "Mixed" on the structure shows that the structure has more than one stereogroup, in this case, &1 and &2. |
|  | Nothing is known about the configuration of the two stereogenic centers. The sample might be a single stereoisomer, a racemic mixture, or a mixture of any of the four possible stereoisomers in any proportion. |

## Chiral Labels

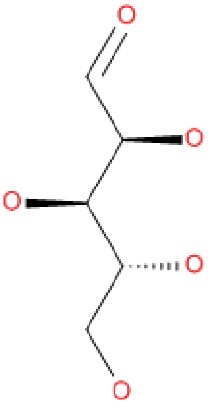
A *chiral label* is a text label that BIOVIA Draw displays adjacent to a structure that summarizes the stereogroup information on the structure. For example:



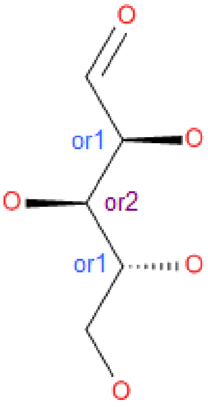
A



B



C



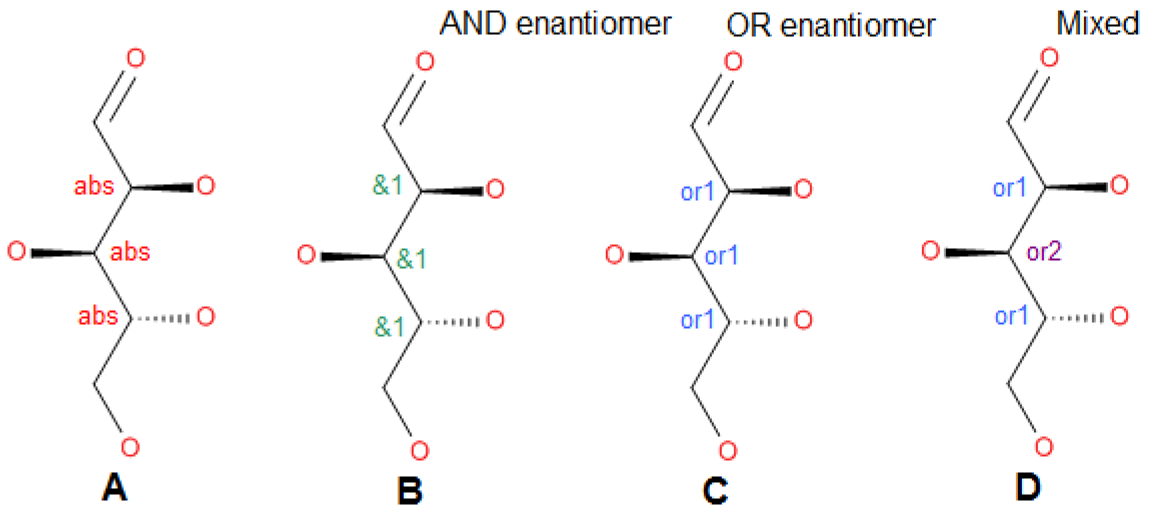
D

BIOVIA Draw provides four chiral labels:

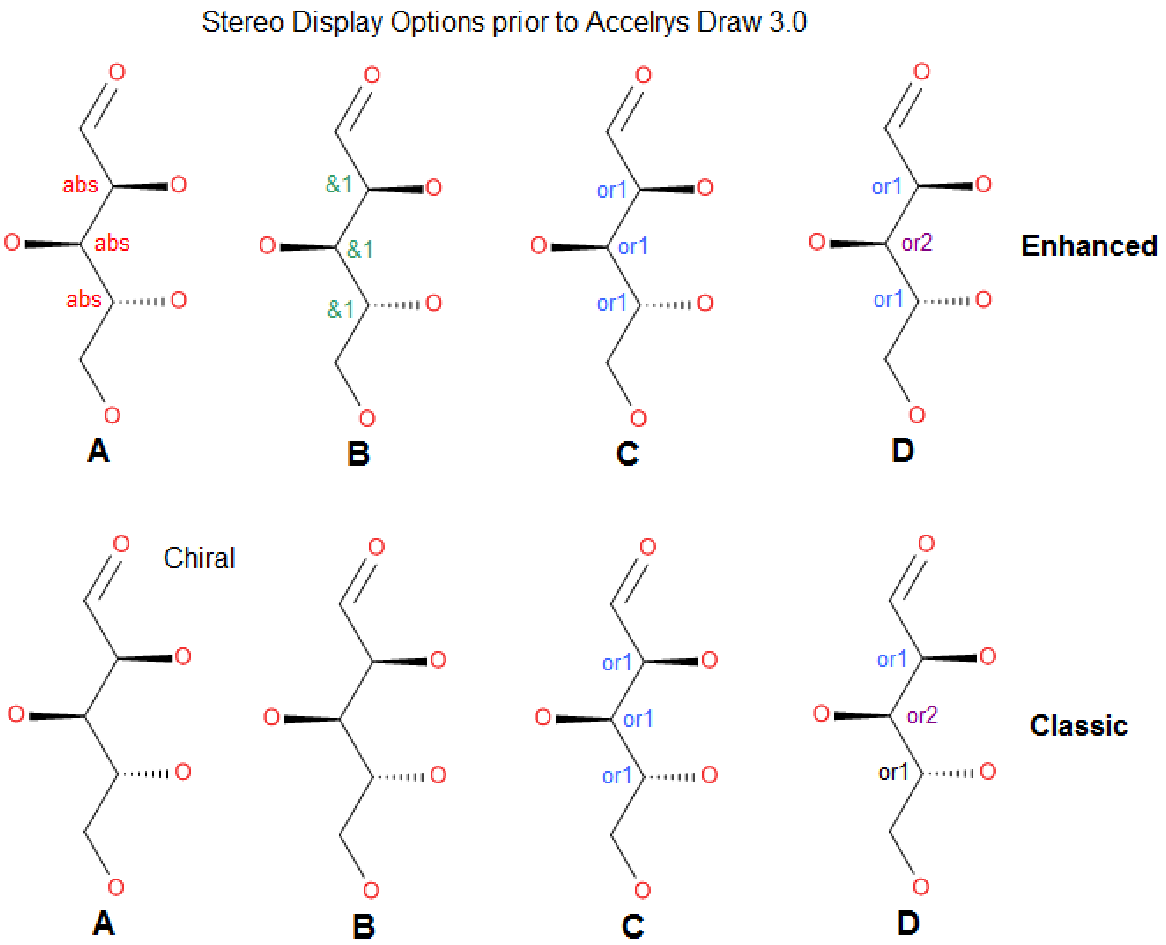
- Absolute chiral label: Displays when all defined stereogenic centers are in the ABS stereogroup and/or that the structure contains no defined stereogenic centers (Structure A). Default value: `(blank)`.
- AND enantiomer chiral label: Displays when all defined stereogenic centers are in the &1 stereogroup (Structure B). Default value: `AND enantiomer`.
- OR enantiomer chiral label: Displays when all defined stereogenic centers are in the `or1` stereogroup (Structure C). Default value: `OR enantiomer`.
- Mixed chiral label: Displays when multiple stereogroups are present (Structure D). Default value: `Mixed`.

**IMPORTANT!** Do not confuse chiral labels with the chiral flag. Chiral labels are solely for display: BIOVIA Draw displays the appropriate chiral label based on information on stereogroups present on the structure. Chiral labels are not explicitly saved to the molfile, and Accelrys Cheshire has no corresponding property. In contrast, The chiral flag is explicitly saved to the molfile and corresponds to the Accelrys Cheshire property M_CHIRAL. For more information on the chiral flag, see Original Accelrys Representation of Tetrahedral Stereochemistry on page 16.

A display setting in BIOVIA Draw allows you to display both the chiral label and the stereogroup labels at individual stereogenic center. The abs stereogroup label displays at each stereogenic center in the ABS stereogroup. For example:



**Note:** Chiral labels were not available prior to Symyx Draw 3.0. Instead, BIOVIA Draw provided only one setting, Display stereochemistry as, with two options: Enhanced and Classic. The following figure shows these options for structures A, B, C, and D:



BIOVIA Draw provides additional display settings for stereochemistry. For more information, see the following topic in the index for the *BIOVIA Draw Online Help*: "settings".

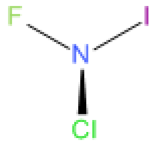
## Rules for Unambiguous Representation of Tetrahedral Stereochemistry

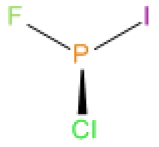
Since molecular entities are commonly represented as two-dimensional drawings in journal publications, chemists have devised a number of conventions for representing three-dimensional information in two-dimensional media, such as perspective drawings, Haworth projections, and Fischer projections. However, some of these conventions can produce drawings that might have more than one interpretation. To ensure that the stereochemistry of structure drawings that you register to your corporate database unambiguously represent the configuration that you intend, the drawings must follow recommended conventions for representing three-dimensional structures in a two dimensional drawing. For more information, see Representation of Stereochemistry in BIOVIA Databases on page 222.

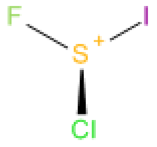
**Note:** Standards for representing the stereochemistry of three-dimensional structures in two-dimensional drawings are given in the following report from the International Union of Pure and Applied Chemistry (IUPAC): J. Brecher, "Graphical Representation of Configuration (IUPAC recommendations 2006)" *Pure Appl. Chem.* (2000), **78**(10), 1897-1970. This report is available as a PDF online at: http://www.iupac.org/publications/pac/2006/pdf/7810x1897.pdf.

## Stereochemistry of Sulfur

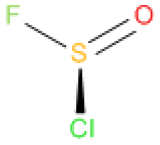
Trigonal pyramidal geometry, or tetrahedral with a lone pair is not considered stereo in Pipeline Pilot Client chemistry.



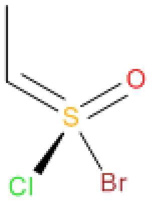




The exception is sulfonyl groups which are puckered. These are tetrahedral with lone pair and considered stereo.

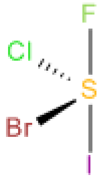


Full four-coordination-number tetrahedral sulfur is also considered stereo.

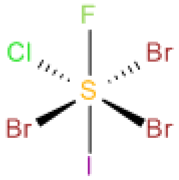


Non-tetrahedral sulfur cases:

- The SeeSaw - Trigonal bipyramidal with lone pair on equator.



- Octahedral geometries.



## Original Accelrys Representation of Tetrahedral Stereochemistry

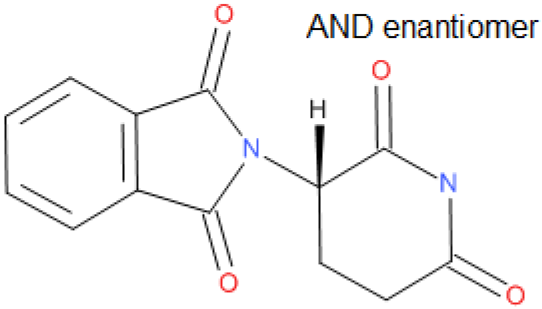
Prior to BIOVIA ISIS/Host and BIOVIA Direct, BIOVIA software used a more limited representation of stereochemistry. The original representation did not support stereogroups on individual stereogenic centers, but used the chiral flag to specify stereochemistry for an entire molecule. For information on the original stereochemical representation, its limitations, and its relationship to the current representation, see Original Representation of Tetrahedral Stereochemistry on page 256.

## Examples of Stereogroups

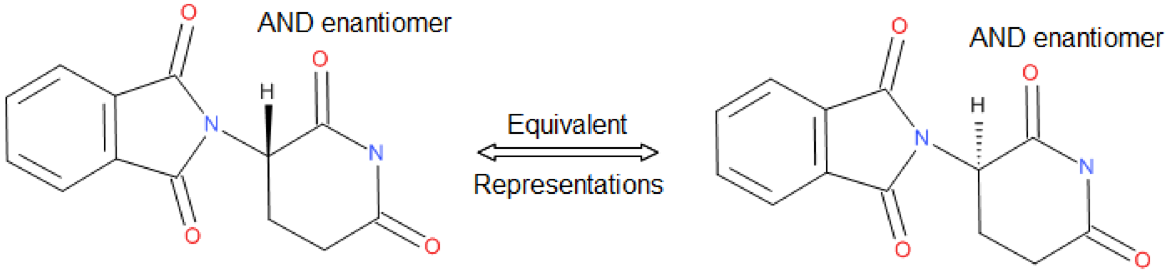
This section contains examples of structures with tetrahedral stereochemistry.

### Example 1: Racemic Mixture

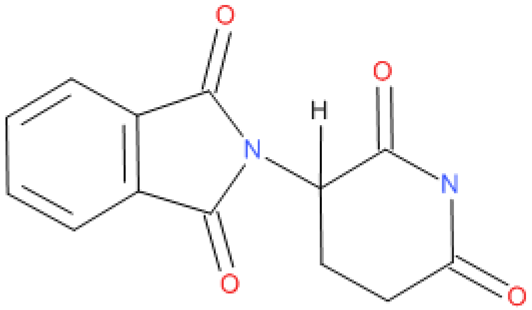
Thalidomide is a racemic mixture of two stereoisomers, only one of which is teratogenic. You can represent the racemic mixture by adding the AND enantiomer chiral label to the stereogenic center:



It is unimportant whether you use an Up or Down bond at the stereogenic center, because both drawings represent exactly the same pair of structures:



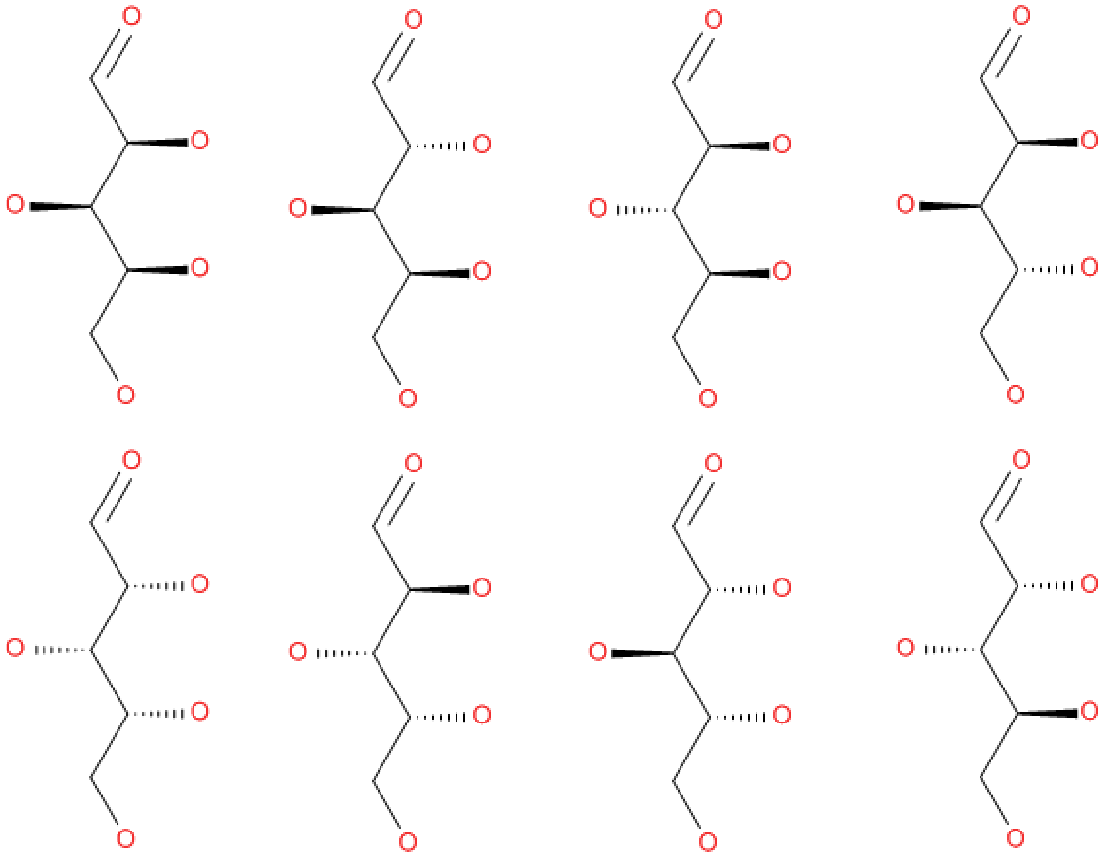
**Note:** Many chemists might want to represent the structure of the racemic mixture by drawing the structure without a stereo bond:



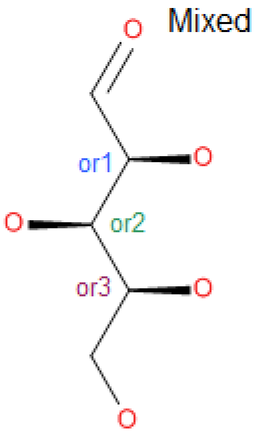
Drawing the structure without the stereo bond, however, implies that you know nothing about the stereoconfiguration of the substance. In this case, you do know something about the substance: you know that it is a racemic mixture.

## Example 2: Acquiring Increasing Amounts of Information on the Stereochemistry of a Sample

You can use OR and AND groups to track a compound as you learn more about the configuration of its stereogenic centers. For example, consider Emil Fischer's 1891 proof of the relative configurations of the stereogenic centers of (-)-arabinose, (+)-glucose, and other sugars. In the case of (-)-arabinose, Fischer began his proof by knowing that the structure must be one of eight stereoisomers:

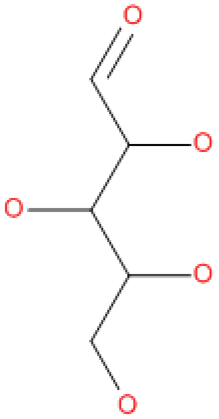


You can represent these possibilities by drawing one of the stereoisomers with each stereogenic center in a different OR group:



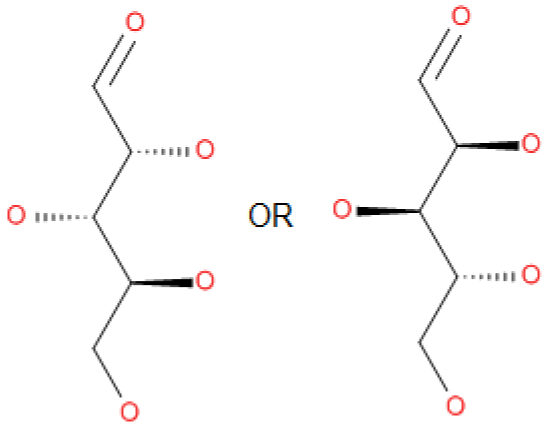
By using OR groups, you show that the substance is a single stereoisomer. By assigning each stereogenic center to a different OR group, you show that you know nothing about the relative configurations of those centers. The number of possible stereoisomers is equal to 2x, where x= the number of OR groups. Thus, 23 = 8.

Many chemists might represent a stereogenic center of unknown configuration by drawing the structure without stereo bonds:

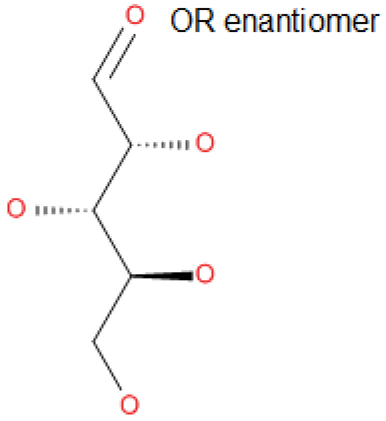


However, this representation implies that you know nothing at all about the structure. It implies that you do not know whether the structure is a single stereoisomer of unknown configuration or a mixture of some or all of the eight possible stereoisomers. That is not true in this case, because you know that the structure is a single stereoisomer and not a mixture.

Eventually, Fischer was able to show that (-)-arabinose was one of two enantiomers:

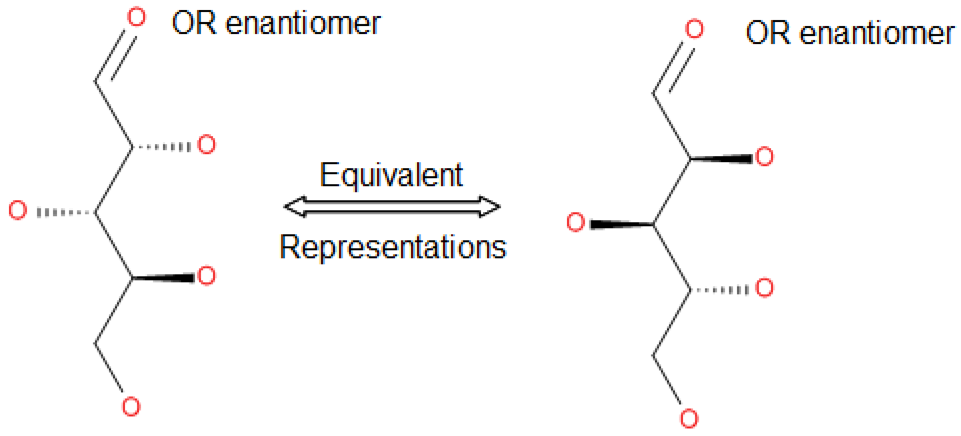


That is, Fischer was able to find the relative configurations of all three stereogenic centers, but not their absolute configurations. You can represent this level of information by drawing one of the enantiomers and using the OR enantiomer chiral label to assign all three stereogenic centers to the same OR group:

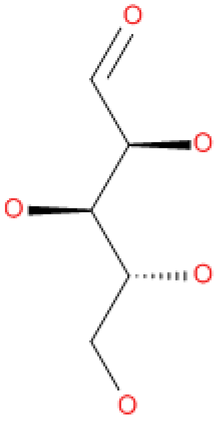


By using orN labels, you show that the substance is a single stereoisomer. By assigning each stereogenic center to a single OR group, you show that you know the relative configurations of all three centers. The number of possible stereoisomers is equal to 21 = 2.

For structures that contain OR and/or AND groups, it does not matter which of the possible stereoisomers that you draw. The or1 labels make the two structures equivalent:

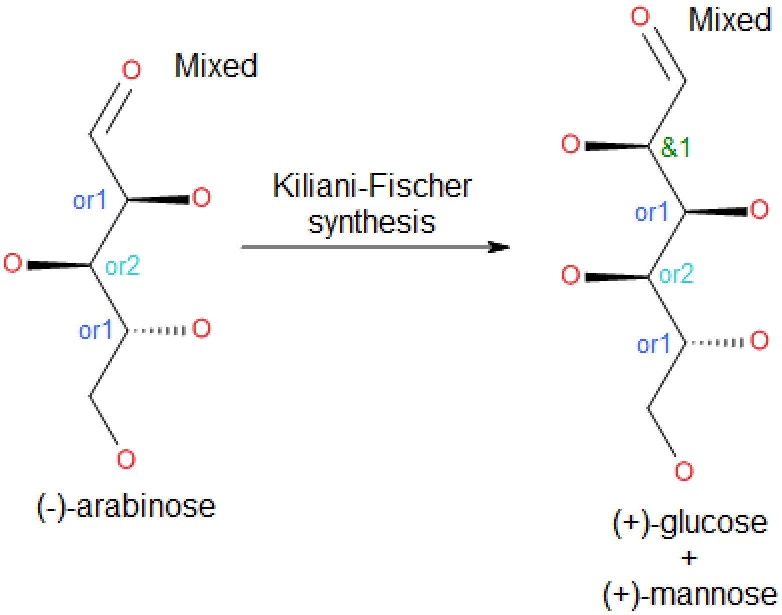


In 1951, it was determined that the absolute stereoconfiguration of (-)-arabinose is:

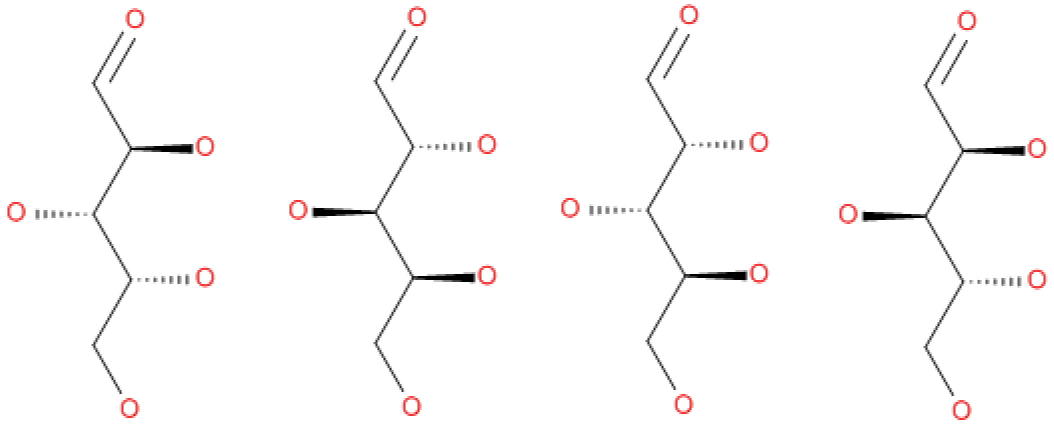


## Example 3: A Mixture of Epimers in a Reaction Product

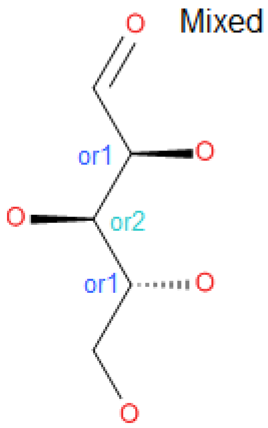
In one stage of his proof of the relative configuration of (+)-glucose, Fischer used the Kiliani-Fischer synthesis to make a mixture of (+)-glucose and (+)-mannose from (-)-arabinose:



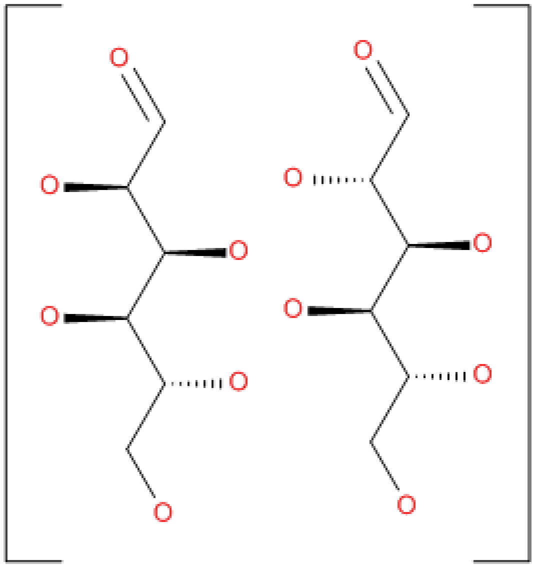
At this point in his proof, Fischer knew the relative configurations of the stereogenic centers at the 2- and 4-carbons of the reactant. That is, Fischer knew that (-)-arabinose was one of four stereoisomers:

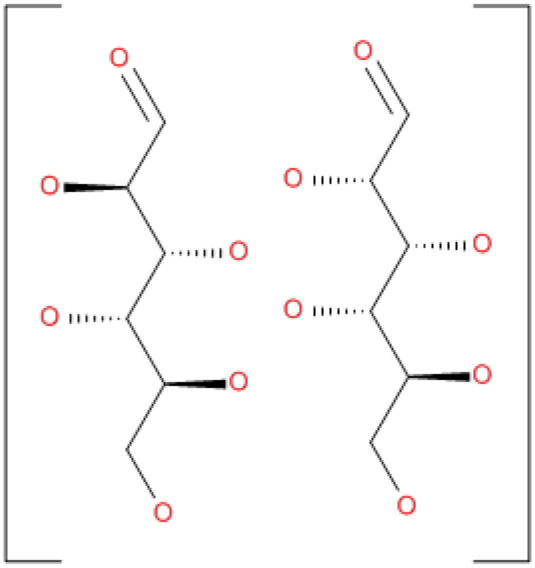


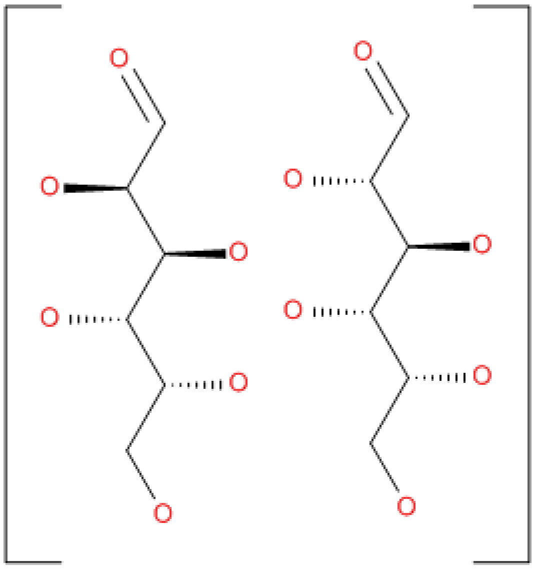
You can represent these possibilities by assigning the stereogenic centers at the 2- and 4-carbon atoms to the same OR group. The stereogenic center at the 3-carbon is in a separate OR group, because at this point Fischer had no information on how its configuration was related to the configuration of the other two centers:

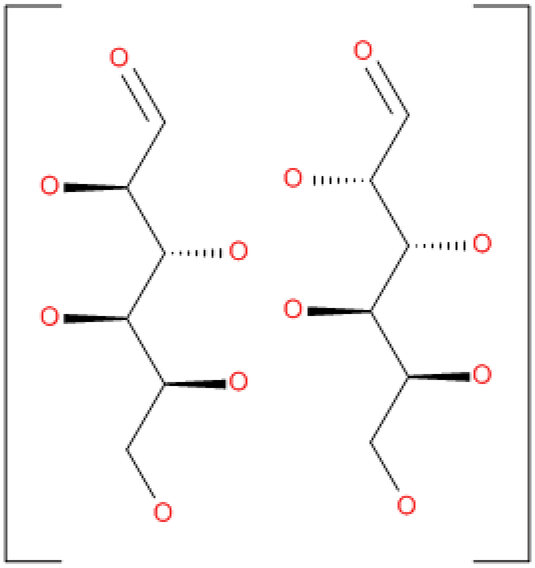


Fischer knew that the relationships between the configurations at the 3-, 4-, and 5-carbons of (+)-glucose and (+)-mannose must be the same as those at the 2-, 3-, and 4-carbons of (-)-arabinose, because the synthesis does not break any bonds to stereogenic centers. He also knew that the synthesis was not stereoselective, so the product would be a mixture of both epimers. Therefore, Fischer knew that the reaction product must be one of four pairs of epimers:

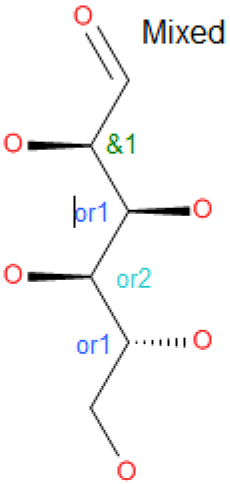








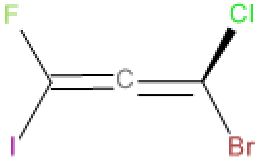
This reaction product can be represented as follows:

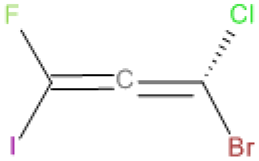


The orN labels at the 3-, 4-, and 5-carbons are the same as those for the 2-, 3-, and 4-carbons of (-)-arabinose. The AND group at the stereogenic center at the 2-carbon (&1) shows that the latter is a mixture of stereoisomers that differ only in their configuration at that stereogenic center (epimers).

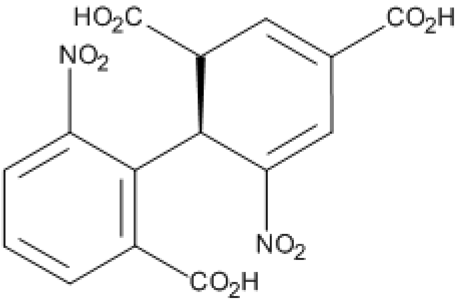
## Stereochemistry of Allenes and Biaryls

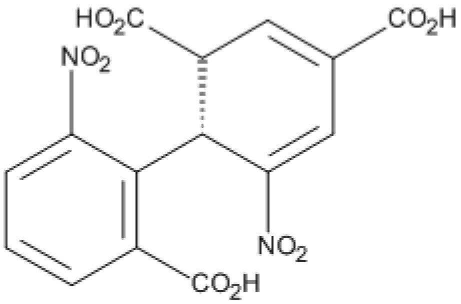
Structures such as allenes and biaryls possess an *axis of chirality*, rather than one or more *chirality centers* like structures with tetrahedral stereochemistry. For example, the following allenes are non-equivalent and are perceived as stereoisomers:





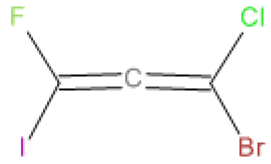
*Ortho*-substituted biaryls can possess an axis of chirality because of hindered rotation about the single bond. For example, the following structures are perceived as stereoisomers:

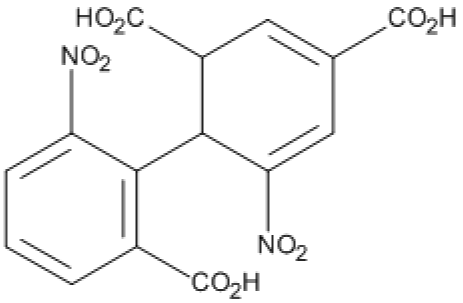




Use the following guidelines for representing these structures in your database:

- If you know the absolute configuration of the structure, use an Up or Down bond to indicate stereochemistry, as in the examples shown previously.
- Do not use Up or Down stereo bonds if the absolute configuration of the allene or biphenyl is unknown. For example:



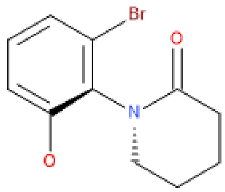


ABS, OR, and AND stereogroups can be applied to allene and biaryl chirality centers. The same stereogroup can contain tetrahedral, allene, and biaryl centers.

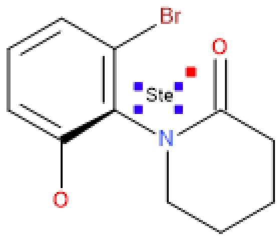
## Perception of Atropisomer Centers

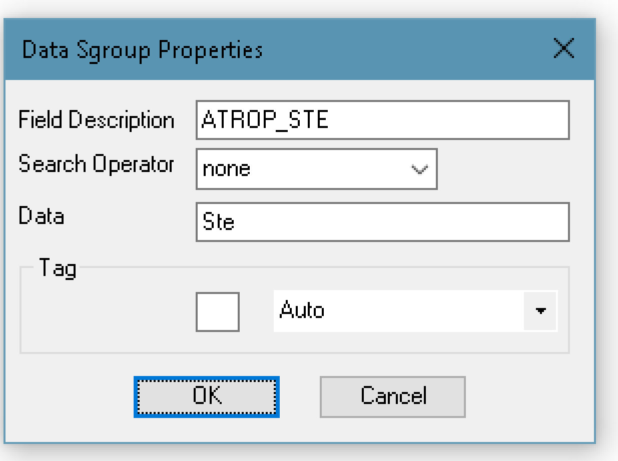
Atropisomer centers are perceived when:

- Wedges are added to both sides of the atropisomers center. The wedges must be on one of the single bonds attached to the atoms that bound the atropisomers center.



- The bond connecting the rings the contains a DAT Sgroup named ATROP_STE with a value of *On*, `True`, or `Ste`. For example, in Draw:





Data Sgroup Properties

Field Description: ATROP_STE

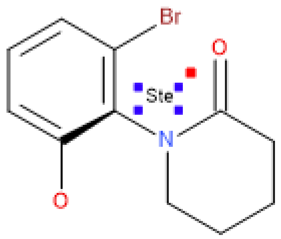
Search Operator: none

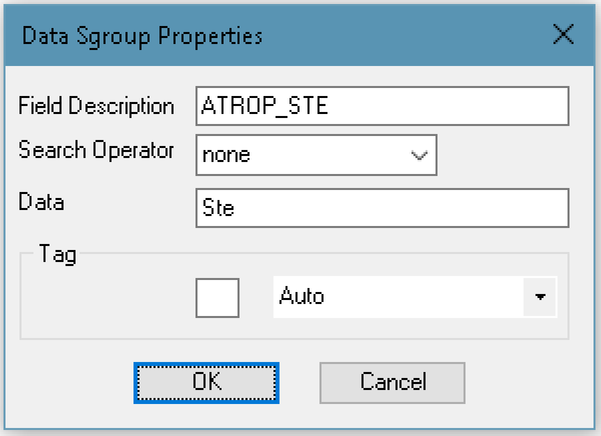
Data: Ste

Tag: □ Auto

- The following conditions are met:
  - A single, non-aromatic bond must connect two rings.
  - The atoms connecting the rings have a hybridization of sp2.
  - The four atoms attached to the atoms connecting the rings have a hybridization of sp2.
  - 3 or 4 alpha attachments restricting rotation between the rings.

  These rules are overridden if the bond connecting the rings the contains a DAT Sgroup named ATROP_STE with a value of *Off*, `False`, or `NoSte`. For example, in Draw:





Field Description: ATROP_STE

Search Operator: none

Data: Ste

Tag: □ Auto

## Additional Information

For additional information and rules for unambiguous representation of cis-trans stereochemistry in structures that you register to databases, see Stereochemistry of Asymmetric Double Bonds on page 236.

# Higher Order Stereochemistry

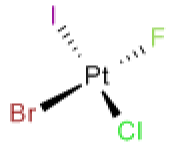
BIOVIA supports square planar, trigonal bipyramidal, and octahedral structures.

See the following examples for recommended representations.

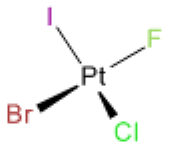
## Square Planar

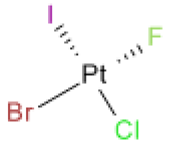
Square planar geometry results from $sp^3d^2$ hybridization with two lone pairs. It is perceived for metals with four attachments and for group VIII elements and their isoelectronic equivalents in period 3 or higher. Users should be aware that metals with four bonds can also exhibit tetrahedral stereo. To mark a metal with tetrahedral stereo, use a single wedge bond. If present, hydrogen atoms must be drawn explicitly. Square planar structures have a coordination number of 4, with 3 isomers.

Recommended:



Also supported:

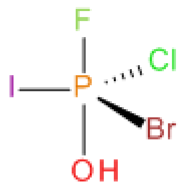


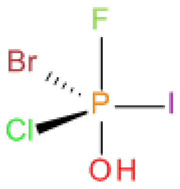


## Trigonal Bipyramidal

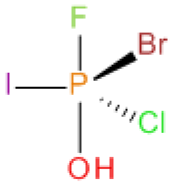
Trigonal bipyramidal geometry results from $sp^3d$ hybridization. It is perceived for metals with five attachments and for group V elements and their isoelectronic equivalents in period 3 or higher. If present, hydrogen atoms must be drawn explicitly. Trigonal bipyramidal structures have a coordination number of 5, with 20 isomers.

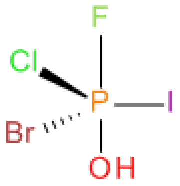
Recommended:





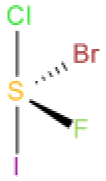
Also supported:





Special Case:

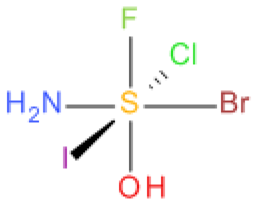
Lone Pair at one Vertex, *See Saw*, Group VI elements and their isoelectronic equivalents in period 3 or higher.

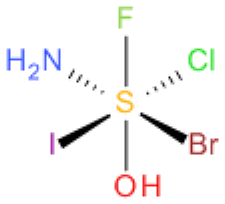


## Octahedral

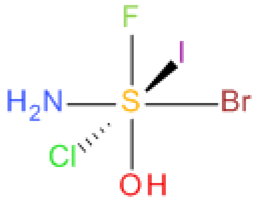
Octahedral geometry results from $sp^3d^2$ hybridization. It is perceived for metals with six attachments and for group VI elements and their isoelectronic equivalents in period 3 or higher. If present, hydrogen atoms must be drawn explicitly. Octahedral structures have a coordination number of 6, with 30 isomers.

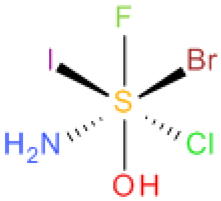
Recommended:





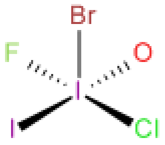
Also supported:





Special Case:

Lone Pair at one Vertex, *Square Pyramidal*, Group VII elements and their isoelectronic equivalents in period 3 or higher.



### Limitations and Restrictions

- Metals with Coordination Number (CN) = 4 can be tetrahedral as well as square planar. Use a single wedge for tetrahedral and either 2 or 4 wedges for square planar.
- Bonds must be covalent. Stereo involving dative or haptic bonds is not supported.
- If present, you must draw hydrogen atoms explicitly.
- 2D coordinates are required.

### Use in Queries

Higher order stereochemistry is normally considered as structure-differentiating for substructure (`SSS`, `RSS`) searches and for exact match (`FLEXMATCH`) searches that include the `/STE` switch. You can ignore higher order stereochemistry during matching by setting the option `IgnoreHigherOrderStereo` in the Substructure Map component or by adding the option to the options argument in a BIOVIA Direct `SSS`, `RSS`, or `FLEXMATCH` search.

## Cis and Trans Stereochemistry

Represent asymmetric double bonds as follows:

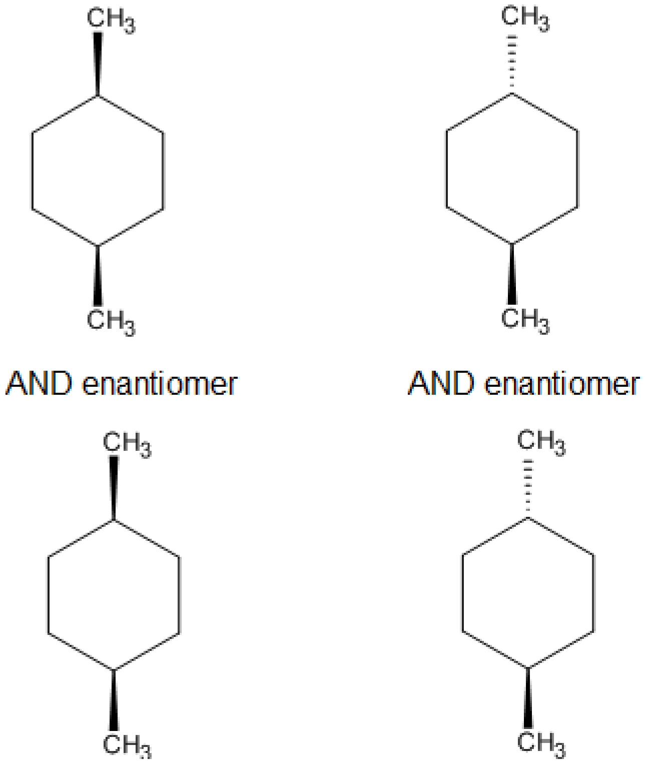
- If you know the exact stereoconfiguration of the double bond, draw the structure in the correct configuration (cis or trans).
- If you do not know the stereoconfiguration of the double bond, draw the structure in either the cis or trans configuration and use a Double Either bond rather than a double bond.

**IMPORTANT!** Do *not* use the Either stereo bond or a colinear arrangement of bonds to represent an unknown configuration. For more information, see Stereochemistry of Asymmetric Double Bonds on page 236.

## Meso Compounds

A meso compound is an achiral compound with an axis of symmetry. It is optically inactive although it contains two or more identical stereocenters. The stereocenters lie on opposite side of the plane of symmetry rendering the compound optically inactive.

An example is the stereochemical representation of di-substituted cyclohexane:



Due to the presence of an axis of symmetry, these structures are superimposable on their mirror images. If you flip all of the wedges they are the same compound, so the all-ABS configuration (chiral flag on) and the all-&1 configuration (chiral-flag-off) are equivalent. Because of this, BIOVIA Direct will ignore any chiral flag present on meso structures during FLEXMATCH matching.

## Spiro Compounds

Spiro compounds are bicyclic compounds with rings connected at a single atom. Spiroatoms may exhibit chirality even if they do not have four different substituents at the chiral center.

## Chemical and Data Substance Groups (Sgroups)

A *substance group (Sgroup)* is a collection of atoms, bonds, or fragments that form a related unit within a structure. Sgroups allow you to represent chemical information that might not be structurally explicit. There are two types of Sgroups: chemical Sgroups and data Sgroups.

A *chemical Sgroup* is an Sgroup that is represented as a structure enclosed by brackets. A chemical Sgroup can have crossing bonds, which connect it to other structures. The crossing bonds and the atoms and bonds within the brackets comprise the chemical definition of the Sgroup (also called the Sgroup basis). Chemical Sgroups can be nested, but cannot overlap one another.

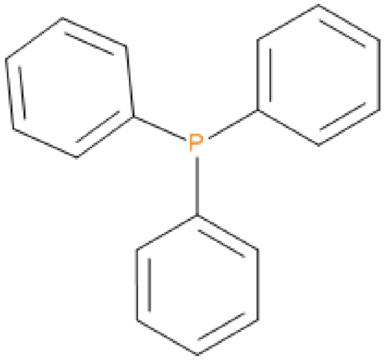
A *data Sgroup* contains chemically-significant text or numeric data that is linked to an atom, a bond, a group of atoms and bonds, or to a chemical Sgroup. The atoms, bonds, and brackets comprise the chemical definition (Sgroup basis) of the data Sgroup. Data Sgroups allow you to attach structure-differentiating data (called attached data or Sgroup data) directly to a structure or a portion of a structure, rather than store the information in separate fields. Data Sgroups are not enclosed in brackets, but are identified by the associated attached data. Attached data acts as a handle for identifying and manipulating data Sgroups in much the same way as the brackets of a chemical Sgroup. Unlike chemical Sgroups, data Sgroups are not hierarchical and can overlap. Each data Sgroup on a structure must be associated with an Sgroup field, which defines the data type as text or numeric.

## Data Sgroups

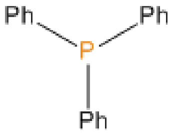
For information on using attached data (Sgroup data) for structure representation, see Attached Data on page 154.

## Abbreviated Structures

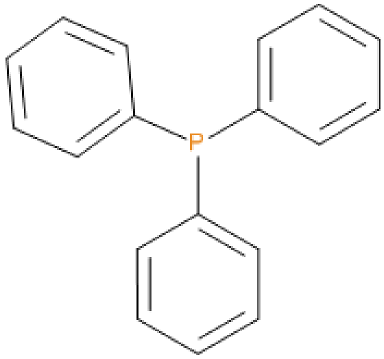
An abbreviated structure (also called an abbreviation or abbreviation Sgroup) is a type of chemical Sgroup that displays a text label to represent all or part of a molecule. The abbreviated structure can be expanded to display the underlying structure. Abbreviated structures are equivalent to the same structures without the abbreviations. For example, the following structures are all equivalent:



No Abbreviation Sgroups



With Abbreviation Sgroups, Contracted Display



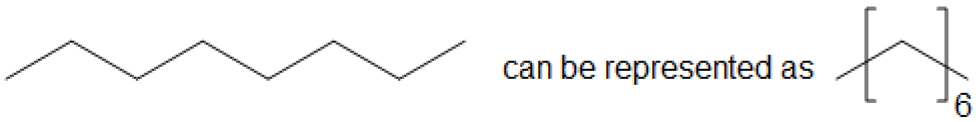
Abbreviation Sgroups, Expanded Display

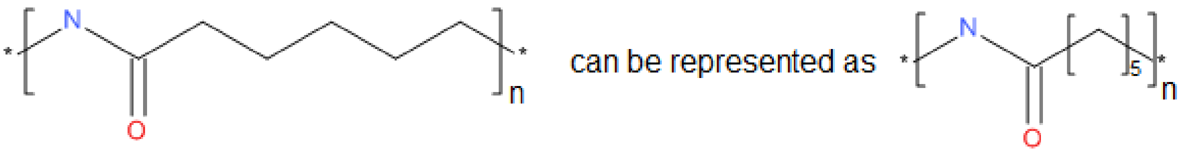
Biopolymers such as protein, RNA, or DNA sequences use abbreviated structures to represent the residues. For more information, see Biopolymer Representation on page 53.

**Note:** Abbreviated structures were formerly called superatoms. The term superatom was dropped to avoid confusion with pseudoatom.

## Multiple Groups

A multiple group Sgroup indicates a fixed number of replications of a fragment or a part of a structure in contracted form. For example:





Like abbreviated structure Sgroups, multiple group Sgroups affect the display only. A multiple group can contain abbreviated structures or other multiple groups.

See also, Variable Repeat Group on page 148 and Allowing Additional Atoms in a Chain or Ring (Link Node) on page 147.

### Other Chemical Sgroups

Other categories of chemical Sgroup include:

- *Mixture* and *component* brackets (mixture and component Sgroups), which can be used to represent ordered and unordered mixtures. See Mixture Representation on page 74.
- *Polymer* brackets (polymer Sgroups), which can be used to represent non-biological polymers. See Polymer Representation on page 78.
- *Generic* brackets (generic Sgroups), which can be used to define a set of atoms and bonds with which you can associate attached data. See Attached Data on page 154.

## Structural Uncertainty

Information on chemical structure is often not completely defined. For example, if you lack information on the configuration of a tetrahedral stereogenic center, you represent it as an undefined stereogenic center, that is a center without Up or Down stereo bonds.

In other cases, you might have no information on an entire structure or a portion of a structure. In these cases, you use *atoms (star atoms) to represent a portion of a structure, or a no-structure to represent an entire structure.

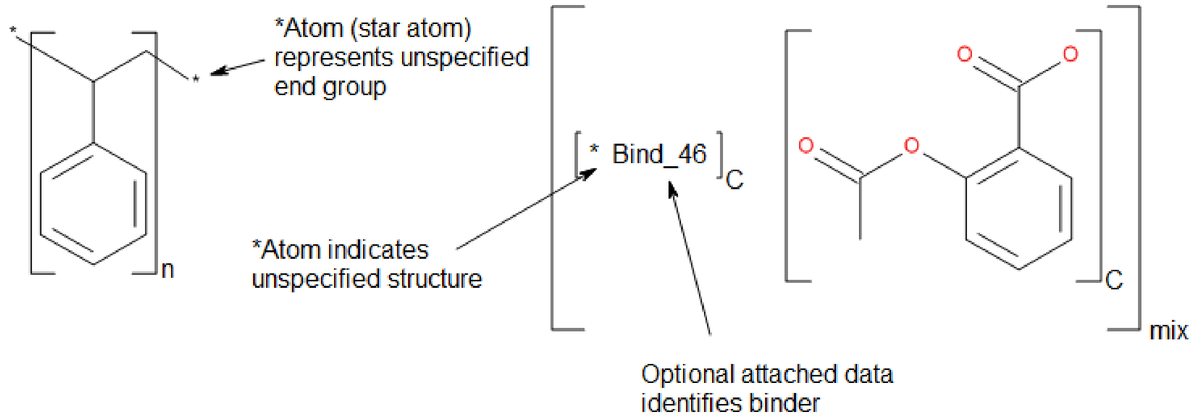
Starting with Direct 2017 R2 it is possible to store certain query atom types in the database, which can also be used to represent structural uncertainty.

### Star Atoms (*)

A *atom (star atom) generally represents a portion of a structure that is unknown or undefined. Common applications of *atoms include representation of:

- Polymer end groups with undefined structure. For information on polymer representation, see Polymer Representation on page 78.
- Components of a mixtures with an undefined structure. For information on mixture representation, see Mixture Representation on page 74.
- Biopolymer residues in condensed form. For information on condensed representation of biopolymers, see Condensed Representation of Biopolymers on page 54.

The following figure shows examples of *atoms in polymers and mixtures:



A *atom has an atomic weight of zero, and cannot be distinguished from other *atoms in searching. To distinguish structures that are chemically distinct but whose chemical structure is unknown, use attached data (Sgroup data). In the previous example, the data that is attached to the *atom distinguishes chemically different binders (Bind_46, Bind_45, Bind_01) in searching. In *atom representation of biopolymers, each residue consists of a single *atom with attached data that distinguishes chemically different residues.

When star atoms are used in a molecule which is not a polymer, a *flexmatch* search only maps a star atom in the query to a star atom in the target. However, a *substructure* search maps a star atom in the query to any non-hydrogen atom in the target, not just to star atoms. A star atom in a substructure query acts the same as the Atom Query Feature: Any Atom (A) on page 139. Attached data restricts the search results as described above.

For more information on attached data, see Attached Data on page 154.

## No-Structures

A no-structure is a chemical structure that consists of zero fragments, that is, zero atoms and zero bonds. Use no-structures to represent chemical structures that are unknown. For example:

- In a *molecule* database, use a no-structure when you have data to register but the structure is unknown or not yet defined.
- In a *reaction* database, you register a no-structure for a reactant or product whose structure is unknown or not yet defined.

Do not confuse no-structures with *atoms. Use *atoms to represent a *portion* of a chemical structure that is unknown or undefined. Use a no-structure solely if the *entire chemical structure* is unknown or undefined.

## Registered Atom Query Types

Molecules and reactions can be registered even when they contain the following substructure query atom types or features. These may be used to represent structural uncertainty:

- A - Represents any atom except R and hydrogen. Matched by substructure query features:
  - Atoms *, A, R, Z
- Q - Represents any atom except hydrogen or carbon. Matched by substructure query features:
  - Atoms *, A, Q, R, Z
  - NOT atom list containing only carbon

- X - Represents any halogen atom. Matched by substructure query features:
  - Atoms *, A, X, R, Z
  - Atom list containing halogen elements F, Cl, Br, I (As is not required, it may also contain additional elements)
  - NOT atom list which does not contain any halogen element
- M - Represents any metal atom. Matched by substructure query features:
  - Atoms *, A, M, R, Z
  - NOT atom list which does not contain any metallic element
- R - Represents any atom including hydrogen. Matched by substructure query features:
  - Atoms *, R
- Z - Represents any atom except hydrogen. Matched by substructure query features:
  - Atoms *, A, R, Z
- Atom list - Represents the elements contained within the list. Matched by substructure query features:
  - Atoms *, A, R, Z, and by Q if the registered atom list does not contain carbon
  - Atom list containing all of the same elements as in the registered atom list, though it can also contain additional elements
  - NOT atom list which does not contain any of the elements in the registered atom list
- NOT atom list - Represents all of the elements except for those within the list. Matched by substructure query features:
  - Atoms *, A, R, Z, and by Q if the registered atom list does not contain carbon
  - Atom list containing all of the same elements as in the registered atom list, it may also contain additional elements
  - NOT atom list - Represents some or all of the elements within the registered atom list (but no other elements)

# Chapter 3:
# Reaction Representation

## Introduction

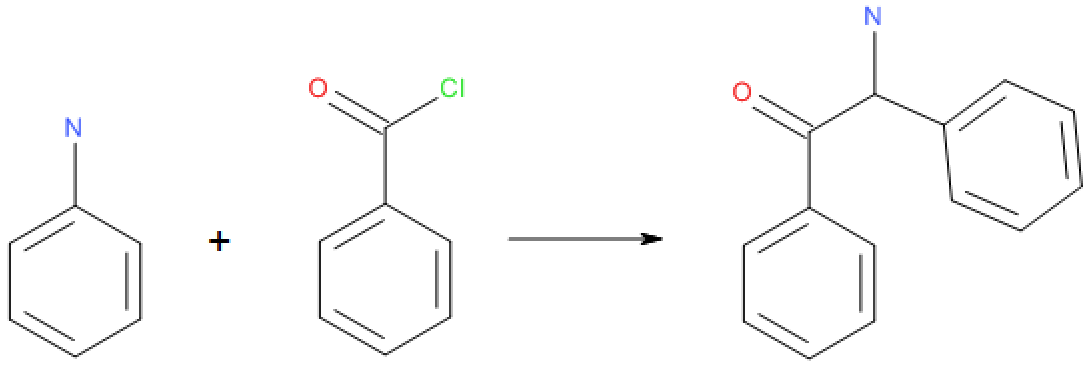
A chemical reaction is a process that results in the transformation of a set of substances into another set of substances. A chemical reaction can consist of one or more single-step reactions.

In an BIOVIA reaction database, a database field contains the diagram of a single-step reaction, which consists of one or more reactant molecules and one or more product molecules. Other fields in the database specify additional information, such as how single-step reactions combine into a multi-step reaction, reagents such as solvents and catalysts, and other data such as yield.

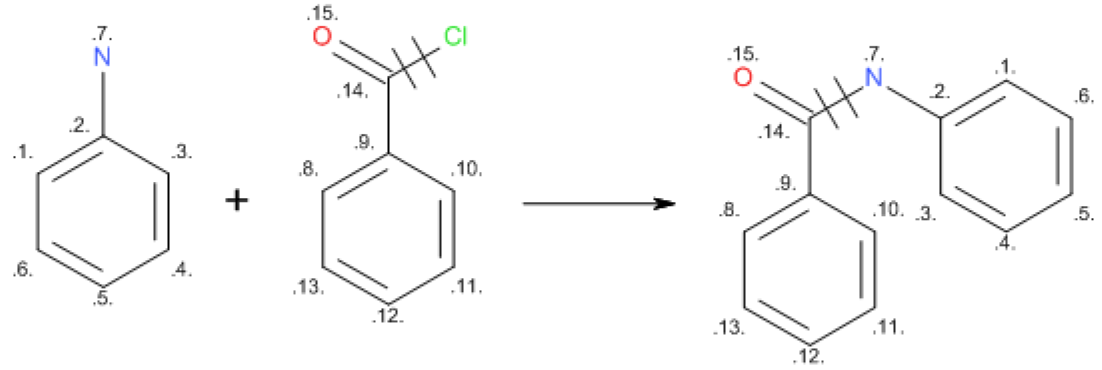
This chapter explains the characteristics of reaction diagrams in single-step reactions. For more information on reaction databases, see BIOVIA Direct databases: BIOVIA DirectAdministration.

## Reaction Mapping

Reaction mapping specifies the correspondence between the atoms and bonds in reactants and the atoms and bonds in products. The figure that follows shows an unmapped reaction:

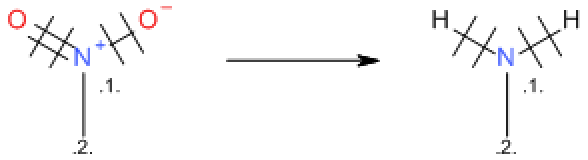


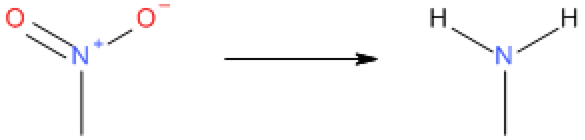
The mapped reaction is:



The numbers on the atoms are atom-atom map numbers that define exactly the atoms in the reactant (s) that correspond to atoms in the product(s) of a reaction. Reacting center marks on the bonds are properties that define what happens to the bond in the reaction, such a s whether the bond is broken or formed, changes order, or does not change. Some of the atoms also have properties that specify whether stereochemistry is inverted or retained in the reaction. For more information on atom and bond properties in mapped reactions, see Stereoconfiguration Atom Properties in Mapped Reactions on page 37 and Properties of Bonds in Mapped Reactions on page 37.

Reaction mapping enables more precise specification of exactly which portions of a structure undergo transformation. This precision is especially important in queries for reaction substructure (RSS) search. For example, both the mapped and unmapped versions of the following query retrieve reactions in which a nitro group is reduced to a primary amine:





The unmapped query, however, also retrieves additional reactions that do not undergo the desired transformation. Because the unmapped query lacks information on atom and bond transformation, it retrieves any reaction that contains a nitro group in the reactant and an amino group in the product, regardless of whether the reaction transforms the nitro group into an amino group.

For more information on reaction substructure search, see Reaction Substructure Search (RSS) on page 191.

Unmapped queries not only retrieve unwanted reactions, but execute much more slowly. For more information on performance of reaction queries, see Performance of RSS Queries on page 195.

## Mapping Reactions Automatically

You can automatically map or unmap reactions in BIOVIA Draw and BIOVIA Direct by using:

- The `AutoMap Reactions` and `UnMap Reactions` commands in the Chemistry menu of BIOVIA Draw. For more information, see the following topic in the index for the *BIOVIA Draw Online Help:* "reactions, automatic mapping".
- The following functions and operators in BIOVIA Direct: `rxnautomap`, `rxnautomapchange`, `rxnautomapstatus`, `mdlaux.automap`, and `mdlaux.regenaamaps`. For information on these operators, see the *BIOVIA Direct Reference Guide*.

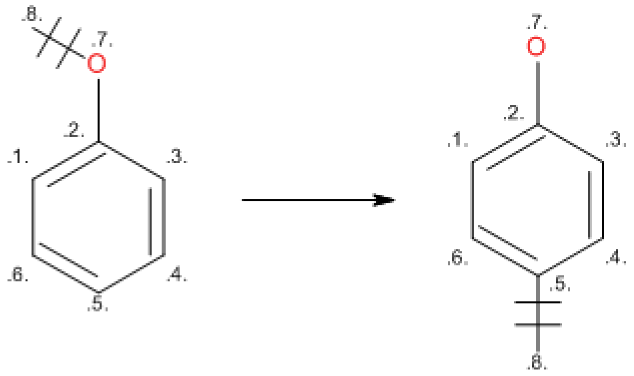
**IMPORTANT!** Always inspect the atom-atom maps on a reaction before you register the reaction or use it in a search. If you notice that a map is incorrect, unmap the reaction and use the procedure in the following section to map the reaction manually.

## Mapping Reactions Manually

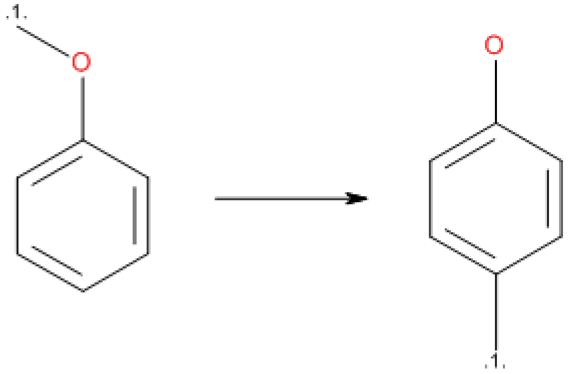
To map a reaction manually, use the Atom-Atom Mapping Tool in BIOVIA Draw to map a few atoms manually, and then use the **`AutoMap Reactions`** command in the BIOVIA Draw Chemistry menu to automatically map the remainder of the reaction correctly. You might need to repeat this process several times.

If you notice that a reaction map is incorrect, unmap the reaction, map a few atoms manually, and then map the rest of the reaction automatically. Repeat this procedure as needed until you have mapped enough of the atoms manually to allow the automap feature to map the remainder of the reaction correctly.

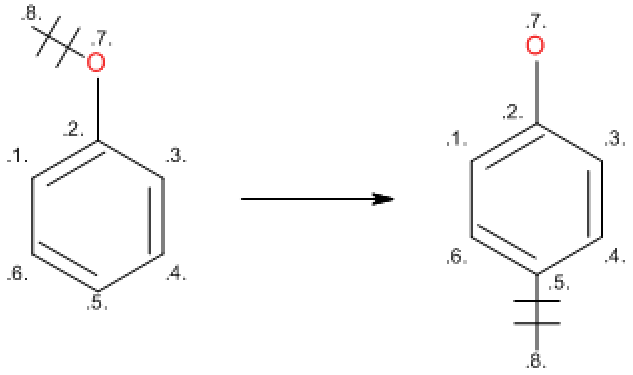
For example, the automap feature failed to detect the migration of the methyl group in the rearrangement of a phenolic ether, as shown by the lack of atom-atom mapping numbers on the methyl carbon:



After unmapping the reaction, map the methyl carbons manually:



When you use the automap feature on this reaction, you obtain a reaction with correct mappings:



When you map reactions manually, set the BIOVIA Draw reaction settings to display all atom and bond properties. The correct settings are:

- Display atom-atom mapping: `On`
- Display reacting centers: `All marks`
- Display stereomarkers: `On`

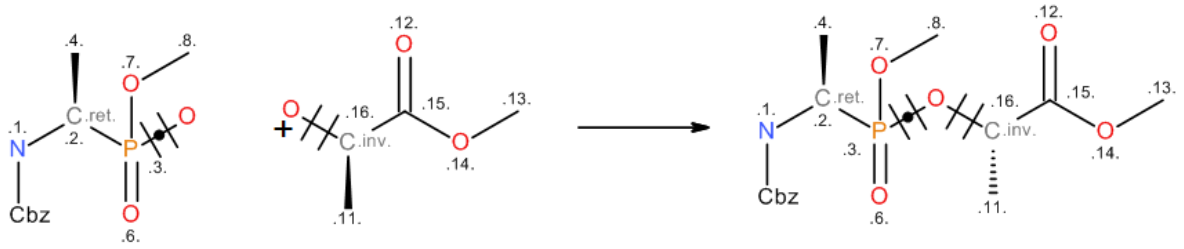
For information on reaction display settings, see the following index entry in the *BIOVIA Draw Online Help:* "settings".

## Stereoconfiguration Atom Properties in Mapped Reactions

In addition to mapping numbers, atoms in mapped reactions have properties that specify changes in stereoconfiguration in a reaction:

| Atom Property | Meaning |
|---|---|
| .ret. | Stereogenic center retains its stereoconfiguration during the reaction. |
| .inv. | Stereogenic center inverts its stereoconfiguration during the reaction. |

For example, the stereoconfiguration of the alcohol in the following Mitsunobu reaction is inverted, but the configuration of the other stereogenic center is retained:



## Properties of Bonds in Mapped Reactions

### Simple Bond Properties

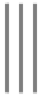
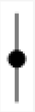
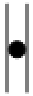
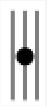
The table that follows shows bond properties for mapped reactions:

| Name | Symbol | Meaning in Reactant | Meaning in Product |
|---|---|---|---|
| No Change | ● | Bond must not change in the reaction. | Bond must not change in the reaction. |
| Change | ∣ | Bond changes type in the reaction. | Bond changes type in the reaction. |
| Make/Break | ∥ | Bond breaks in the reaction. | Bond forms in the reaction. |

Conversion of any of the following bond types is considered a change in bond type: Single, Double, Triple, Aromatic. Query bonds are not included because query bonds cannot be registered to a database. For information on bond types and bond perception, see Bond Types on page 5 and Aromaticity on page 7.

## Combined Bond Properties

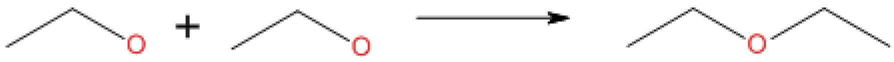
In some reactions, the fate of a particular atom or bond in a reactant can be ambiguous or even unknown. In these cases, the bond property is a combination of the simple bond properties that represent the different fates of the bond. The table that follows shows the possible combinations of simple bond properties:

| Name | Symbol | Meaning in Reactant | Meaning in Product |
|---|---|---|---|
| Change/Make or Break |  | Bond changes bond type OR breaks | Bond changes bond type OR forms |
| Change/No Change |  | Bond changes type OR must not change | Bond changes type OR must not change |
| Make or Break/No Change |  | Bond breaks OR does not change | Bond forms OR does not change |
| Change/Make or Break/No Change |  | Bond breaks OR changes bond type OR must not change | Bond forms OR changes bond type OR must not change |

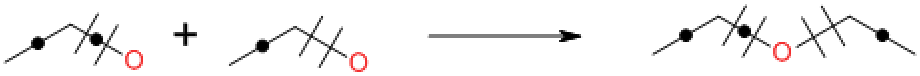
For examples of combined bond properties, see the sections that follow.

### Example: Ambiguity in the Fate of Reacting Atoms and Bonds

Consider the industrial synthesis of diethyl ether through the dehydration of ethyl alcohol:



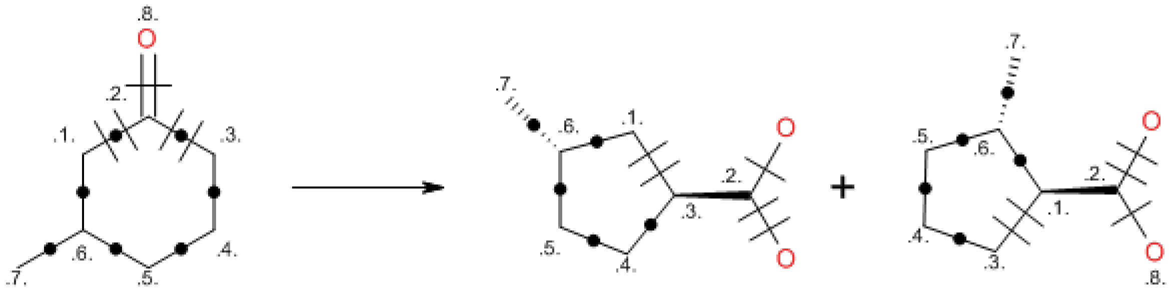
In this reaction, it is impossible to know which of the two oxygen atoms in the reactants is present in the product. In fact, either oxygen atom can be present in the product, and therefore either carbon-oxygen bond can be broken. The following figure shows how to use atom-atom mapping numbers and a combined bond property to indicate the ambiguity:



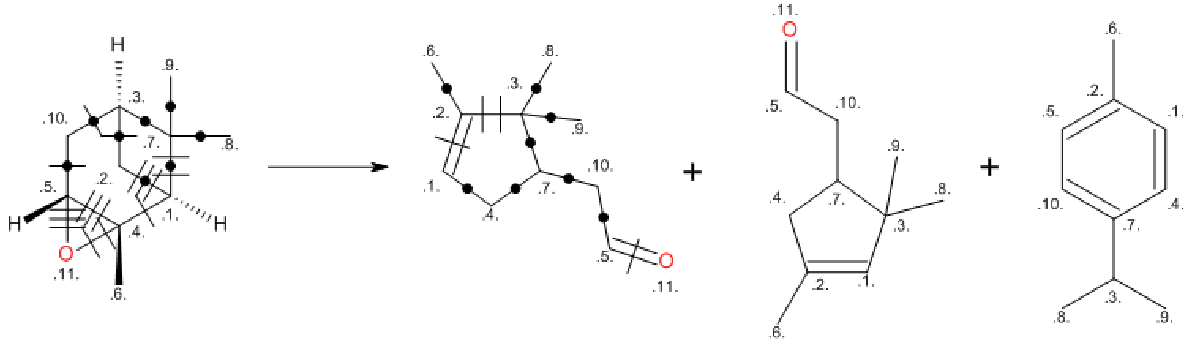
The oxygen atoms lack mapping numbers, because you do not know which oxygen atom in the reactants is incorporated into the product. The combined bond property Breaks/No Change in the carbon-oxygen bond shows the two possible fates of the bond in the reaction.

### Example: Multiple Fates of Bonds in Products

In the following reaction, the ring bond between mapped atoms 1 and 2 and between mapped atoms 1 and 3 are unchanged in the first product and broken in the second product:



In the following reaction, the bond between mapped atoms 1 and 4 is unchanged in product A, is broken in product B, and changes bond type from single to double in product C. The bond between mapped atoms 4 and 8 is unchanged in products A and B and is broken in product C:



## Stereogroup Information in Reactants and Products

Each defined tetrahedral stereogenic center (that is, each center that is marked with an Up or Down stereo bond) must belong to a stereochemical group (stereogroup), as described in Tetrahedral Stereochemistry on page 9. The following rules apply to stereogroups in reactions:

- Unless you have specific information about relationships between configurations of AND stereogenic centers in different reactants, you should create distinct AND stereogroups for each reactant and product molecule. The same applies to OR stereogroups in reactants and products.
- AND stereogenic centers in reactants *must* belong to different stereogroups than the corresponding stereogenic centers in the products. The same is true for OR stereogroups.
- A structure or reaction can have only one ABS stereogroup. Therefore, ABS stereogenic centers in each product or reactant structure all belong to the same stereogroup: the ABS stereogroup.

For example, if each reactant and product in a reaction of the form A+B→C+D contains a set of stereogenic centers that are part of an AND stereogroup, then AND centers in reactant A should be assigned to stereogroup &1; centers in reactant B should be assigned to stereogroup &2; centers in product C should be assigned to stereogroup &3; and centers in product D should be assigned to stereogroup &4.

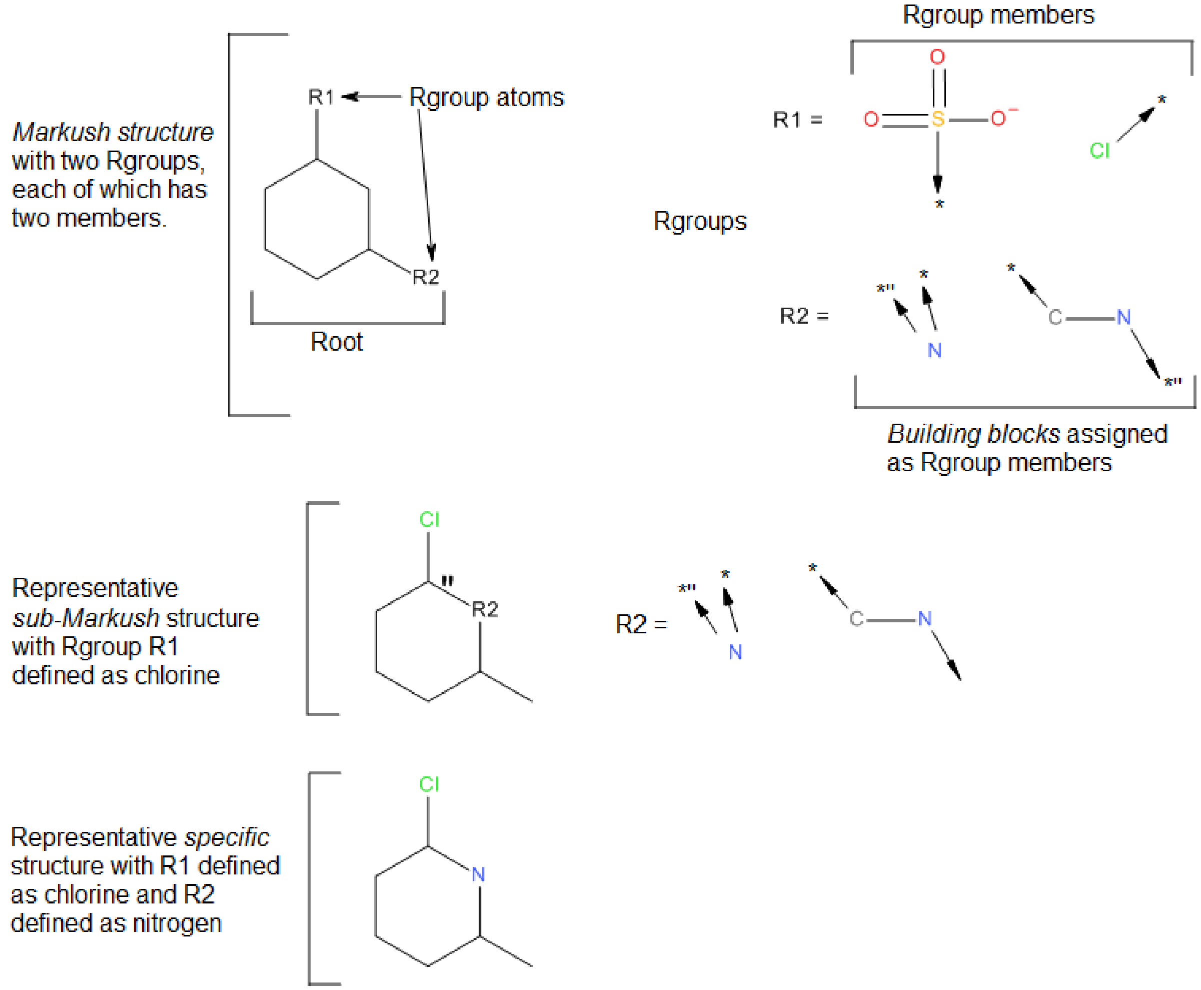
The numeric values of the stereogroup index numbers 1, 2, 3, 4... have no significance other than to indicate that groups with different numbers are unrelated.

# Chapter 4:
# Markush (Rgroup) Structures

## Library Representation

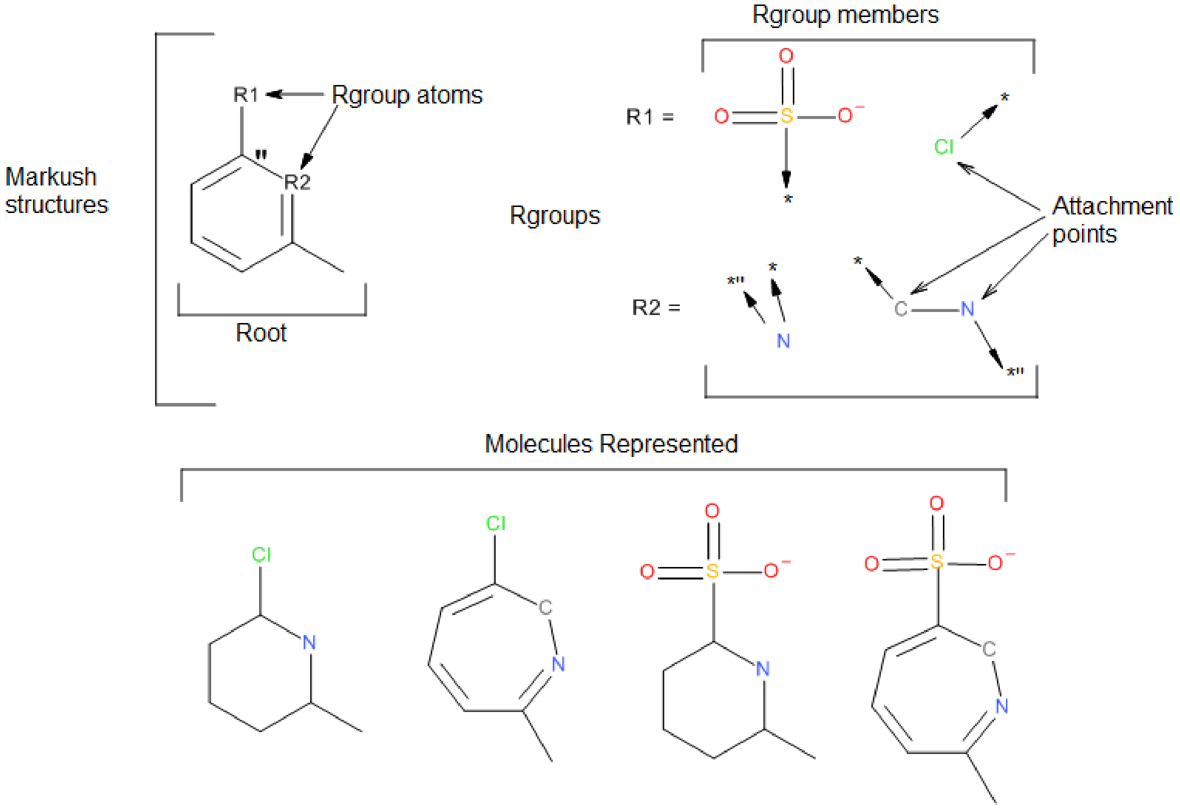
*Markush structures* are used to represent combinatorial *libraries* of related compounds. The figure that follows summarizes the terminology of combinatorial chemistry.

A combinatorial library includes a top-level Markush structure and specific structures it represents. In diagram 3.1, only one representative sub-Markush and specific structure are shown.


Markush structure with two Rgroups, each of which has two members.
R1
Rgroup atoms
R2
Root
Rgroup members
Rgroups
R1 =
O
O
S
O−
*
Cl
*
R2 =
*''
*
N
*
C
N
*''
Building blocks assigned as Rgroup members
Representative sub-Markush structure with Rgroup R1 defined as chlorine
Cl
''
R2
R2 =
*''
*
N
*
C
N
Representative specific structure with R1 defined as chlorine and R2 defined as nitrogen
Cl
N

A *library* is a set of chemical structures that are related through a common *Markush structure*. The set consists of a top-level Markush structure, plus all of the *sub-Markush* and *specific* structures that it represents.

A *Markush (Rgroup) structure* (also called a "generic" structure) represents one or more actual structures. A Markush structure consists of a specific *root structure*, or *scaffold*, with specific Rgroup atoms in specific positions on the root, and an explicit set of fragments and identifying information comprising the members for each Rgroup atom:



A *specific structure* is a fully defined chemical structure (with no Markush features). For example, all of the structures that are described in Molecule Representation on page 3, are specific structures. The process of *enumeration* creates specific structures from a Markush structure and its Rgroups.

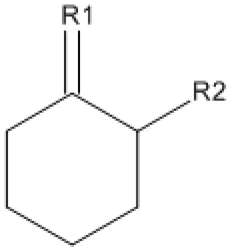
The sections that follow describe the major components of a Markush structure.

## The Root Structure

The *root structure* consists of a chemical structure with attached Rgroup atoms. Each atom marks a position of variability on the structure where any one of a set of substituents (known as Rgroup *members* or *building blocks*) can attach. A root structure must be present in all Markush structures.

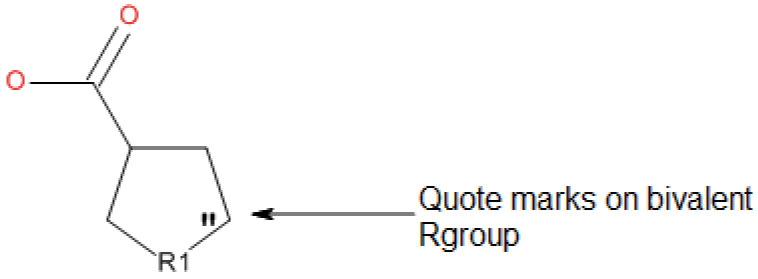
The following guidelines apply to Rgroup atoms:

- Rgroup atoms can be attached to the root with the same or different bond types:

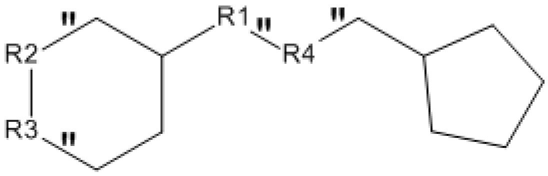
  

  The bond type must satisfy the valence requirements of each Rgroup member. A double or triple bond represents a *single* point of attachment.

- An Rgroup can have up to two attachments. An Rgroup with two attachments is referred to as a bivalent or doubly connected Rgroup. When you draw a bivalent Rgroup structure, BIOVIA Draw automatically marks one of the bonds on the root structure with a double-quote to distinguish between the two bonds.



- Rgroup atoms can exist on adjacent bond junctions:



- An Rgroup atom need not be connected to any atoms (zero attachments).

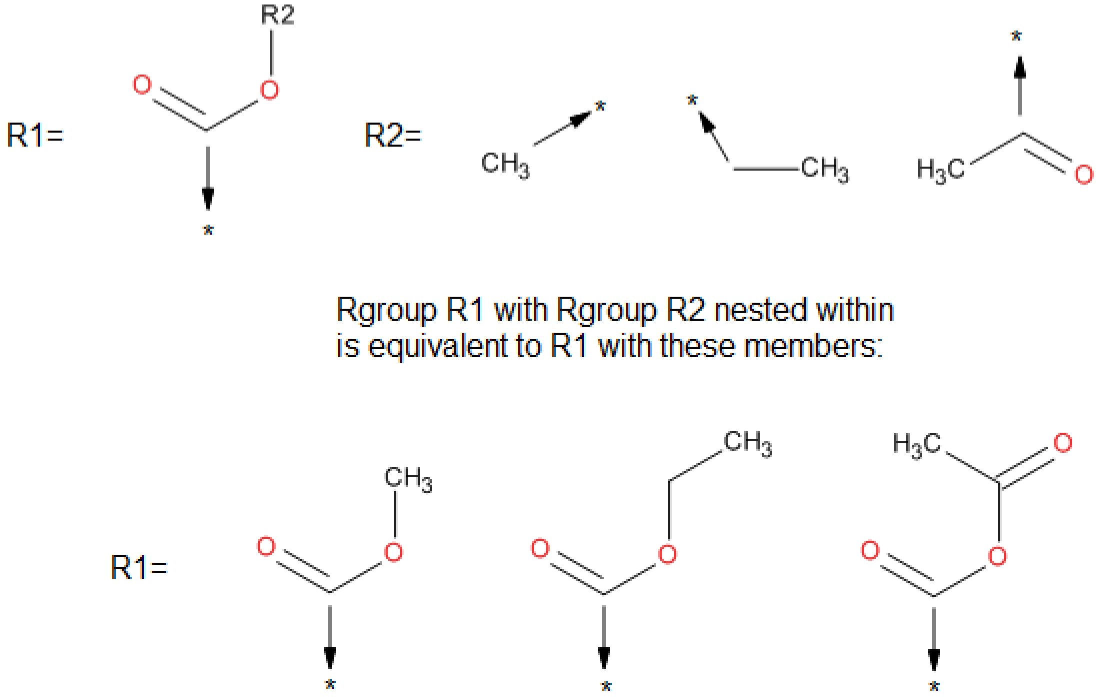
## Rgroups and Rgroup Members

An *Rgroup* consists of a set of structural fragments called *Rgroup members* or building blocks, that specify the structures that can be present at the Rgroup atom site. Each Rgroup member has zero, one, or two *attachment points*, marked by an arrow and an asterisk. An attachment point specifies the *atom* on the Rgroup member that attaches to the Rgroup atom site on the root (represented by the asterisk). This atom *replaces* the Rgroup atom on the root. The arrow does *not* represent a bond.

> **Note:** Direct does not allow a substructure query to contain an atom list with hydrogen inside of an Rgroup member. Direct 9.0 removed the hydrogen from the atom list and then executed the substructure search. Because this led to misleading results, Direct 9.1 and later report an error and not execute the search.

### Nested Rgroups

You can nest Rgroups by defining Rgroup members that contain other Rgroup atoms. In the example below, the R2 members are nested within the R1 member:



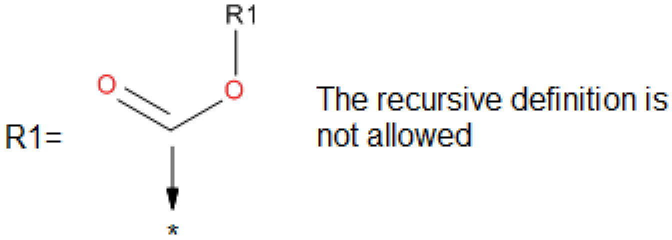
Rgroup R1 with Rgroup R2 nested within
is equivalent to R1 with these members:

You can nest:

- More than one Rgroup within a member
- Rgroups with either one or two attachment points

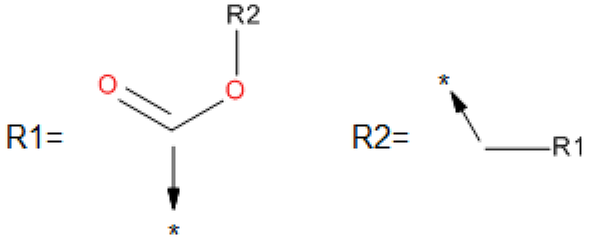
You cannot nest:

- An Rgroup within one of its own members:



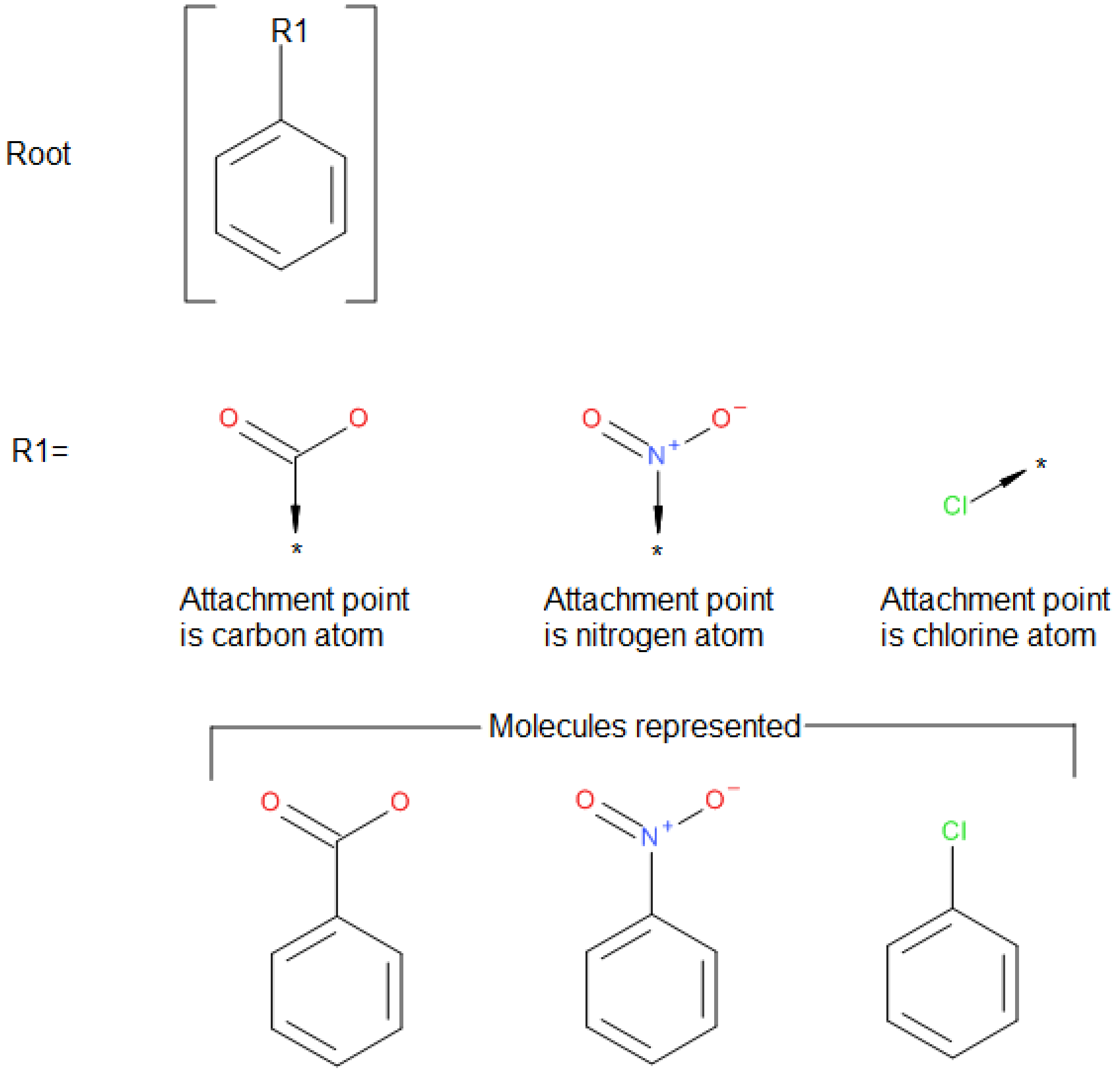
The recursive definition is
not allowed

- The same Rgroup more than once within a member
- A member of R1 containing R2 if R2 contains a member with a nested R1. Thus, the following recursive definition is invalid:



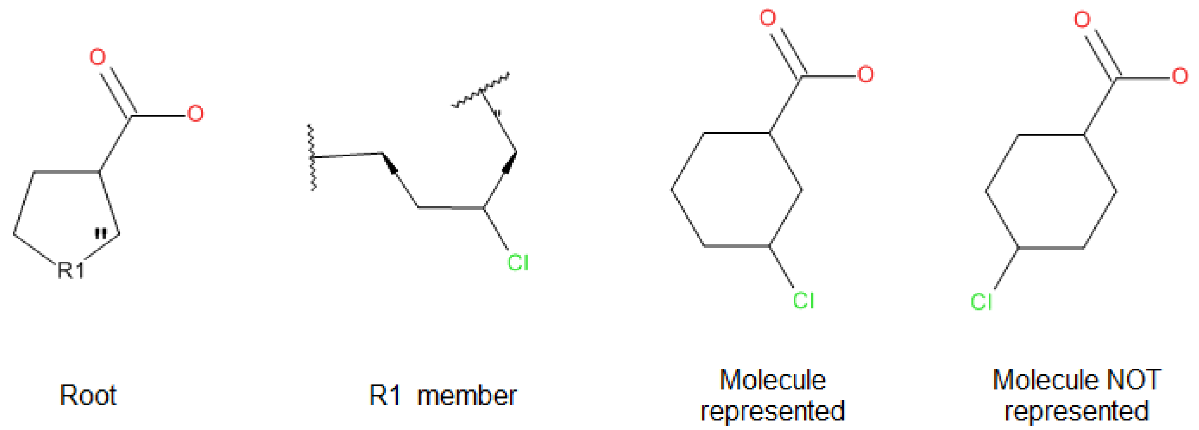
## One Attachment Point

An Rgroup atom at the terminus of a bond requires a single attachment point.

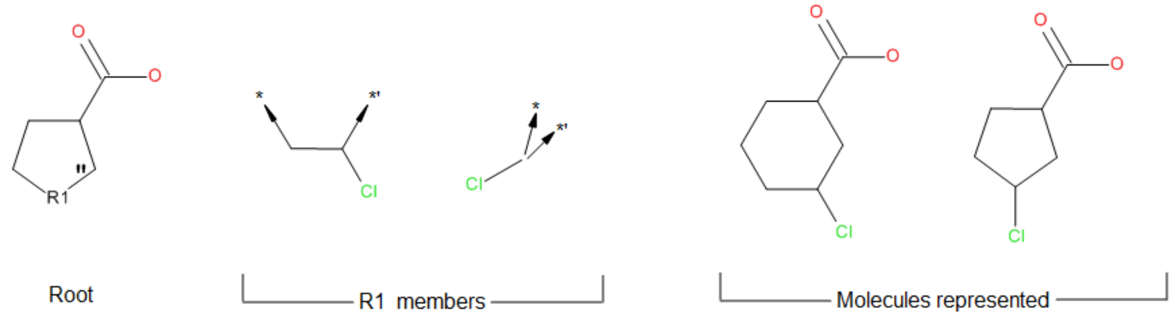

Root
R1
R1=
O
O
N+
O−
Cl
*
Attachment point is carbon atom
Attachment point is nitrogen atom
Attachment point is chlorine atom
Molecules represented

## Two Attachment Points

For an Rgroup atom at the junction of two bonds, each bond must attach to an atom on the Rgroup member. In this case, you must specify *two* attachment points on each Rgroup member. The second attachment point is marked with a double quote to correspond to the bond in the root marked with a double quote:

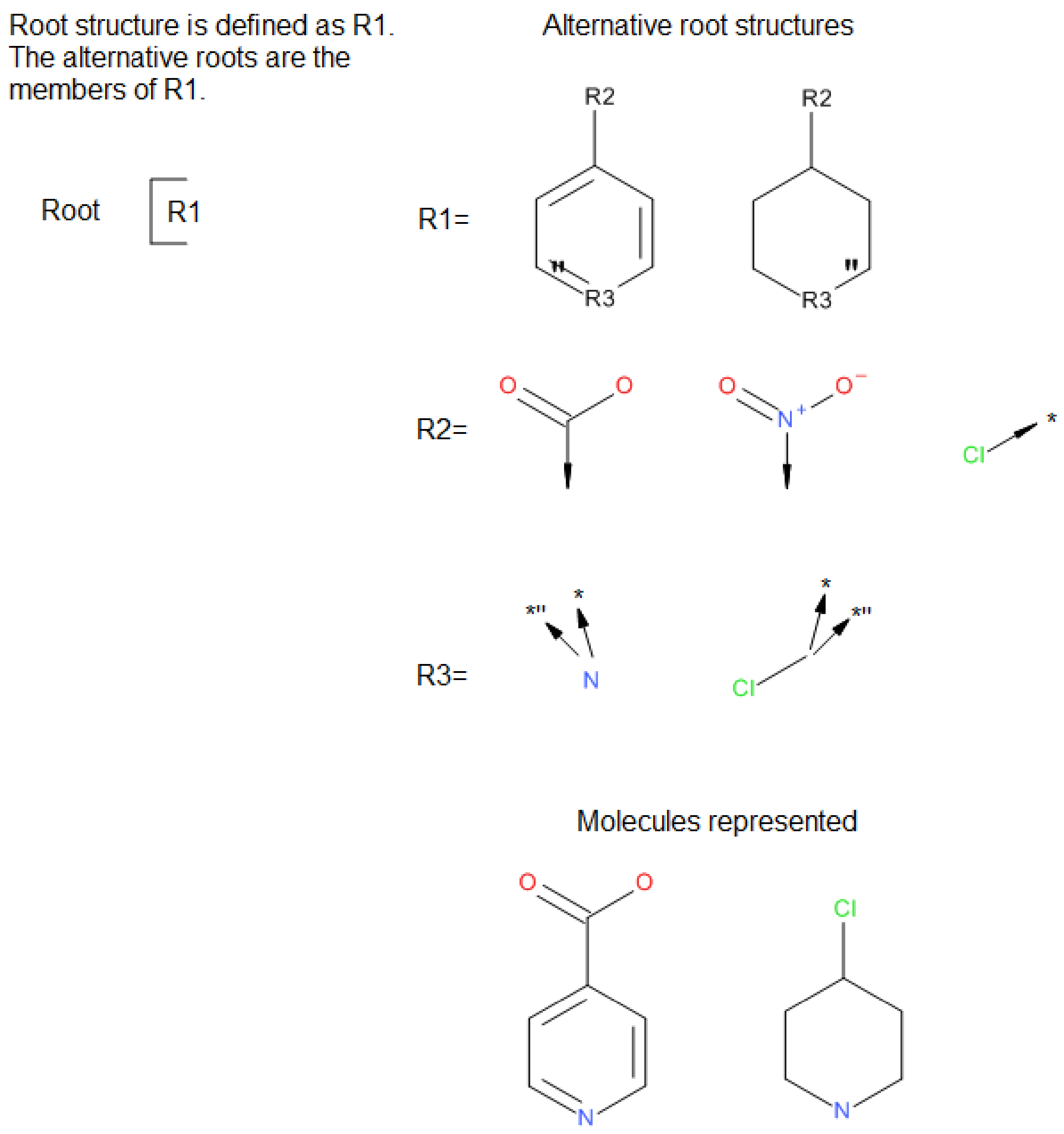


The two attachment points can be different atoms on the molecule fragment, or they can be the same atom:



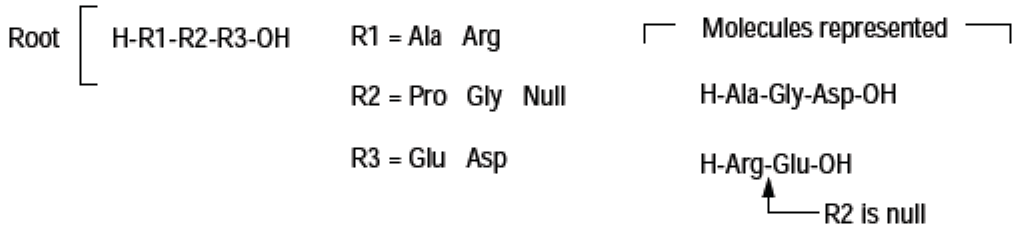
## Unconnected Rgroup Atom

An unconnected Rgroup atom in a root lets you define two or more *alternate root structures*. This technique is very useful for defining libraries which cannot be represented by a single Markush structure. The structure of each alternative root can differ in any way except that the same Rgroup atoms must be present in each. For example, you can generate a library if a building block is the main part of the scaffold.


Root structure is defined as R1.
The alternative roots are the
members of R1.
Alternative root structures
Root
R1
R1=
R2
R2
R3
R3
R2=
O
O
O
N+
O−
Cl
*
R3=
*"
*
N
*
*"
Cl
Molecules represented
O
O
N
Cl
N

## Null Members

You can add a *null Rgroup* member to a singly or doubly connected group. The null Rgroup member allows a direct bonding between its neighbor Rgroups or atoms. The following example includes a null member as part of R2 and represents a peptide sequence in which amino acid residues from the R1 and R3 members are directly attached. In this case, the null Rgroup member represents an experiment in which an Rgroup is explicitly omitted during a synthesis.



## Markush Structures Differ From Markush Queries

A *Markush query* finds structures that contain your query as a substructure wholly within a larger structure. Both Markush structures and Markush queries contain a common structure with points of variation represented by numbered Rgroup atoms (R1, R2, R3, and so on). However, these similarities are superficial. The two types of "Rgroup" structures are used for quite different tasks and differ from one another in many structural details.

For information on substructure search, see Substructure Search (SSS). For information on the special features of Markush queries, see Markush Search Queries on page 160.

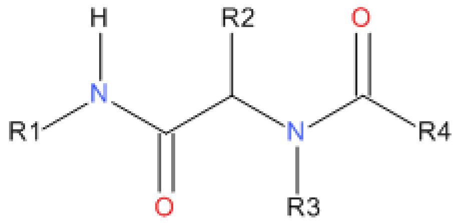
# Enumeration of Markush Structures

Enumeration is the process that generates the set of fully defined structures (called specific structures or specifics) that a Markush structure represents. BIOVIA provides two mechanisms for enumerating structures:

- Scaffold-based Enumeration, which creates specific product structures from a Markush product structure.
- Reaction-based Enumeration,which uses a Markush reaction and a set of specific reactants or precursors to generate the specific structures that are products of the reaction.

## Scaffold-based Enumeration

*Scaffold-based enumeration* begins with a Markush structure with defined Rgroup members, and generates a set of specific structures that correspond to all possible combinations of Rgroup members. For example, the root structure of the product of an Ugi reaction is:



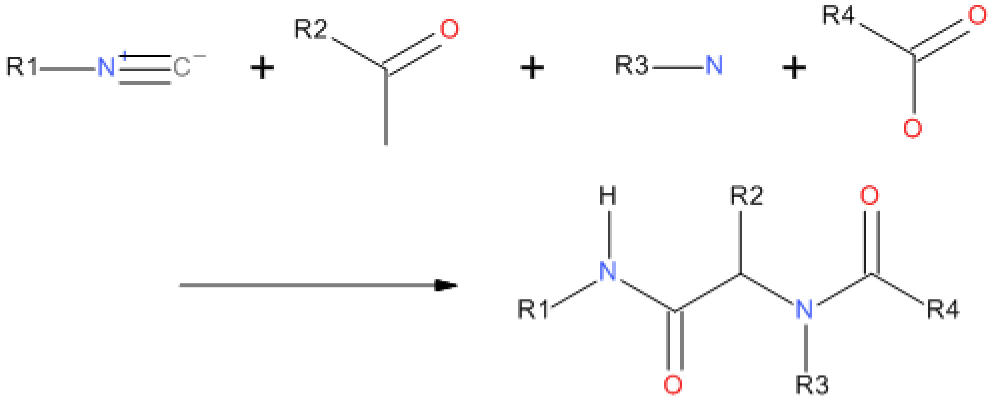
If each of the four Rgroups has four members, then enumeration generates a total of 4x4x4x4 = 256 specifics.

To perform this enumeration, you would use the Rgroup enumeration components in Pipeline Pilot. See the "Enumeration by Rgroups" section of the *Advanced Chemistry Guide* in the *Pipeline Pilot Chemistry Collection*.

**Note:** Pipeline Pilot Rgroup enumerators expect the core and members to be on separate data records. The *Extract Rgroup Fragments* component can be used to separate a single Markush structure into multiple core and members records.

### Reaction-based Enumeration

In reaction-based enumeration, the reaction chemistry is specified by a reaction in which the components are structures with Rgroup atoms that show the points of variability: that is, a reaction whose components are the roots of Markush structures. The following example shows a generic Ugi reaction:

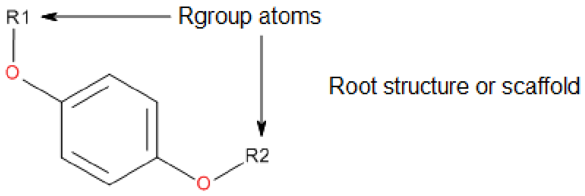


To specify the Rgroup members of the Markush structures, you must specify a set of specific reactant structures for each Markush reactant. For the Ugi reaction, you need to specify four sets of specific reactants: isocyanides (R1), aldehydes (R2), primary amines (R3), and carboxylic acids (R4). For each combination of specific reactants, the enumeration algorithm discerns which portions of those reactants need to be combined to create the product structure.

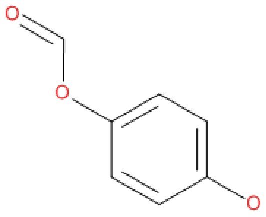
Use the Reaction Enumeration components in Pipeline Pilot to perform this type of enumeration. See the "Enumeration by Reactions" section of the *Basic Chemistry Guide* in the *Pipeline PilotChemistry Collection*. For best results, use the reactions should contain atom-to-atom mapping (aamap) information. Use the Add Reaction Mapping component to add these if they are not already present.

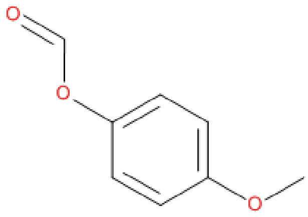
## Rgroup Decomposition

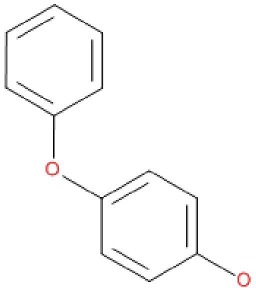
*Rgroup decomposition* is the process of "decomposing" a set of fully defined (specific) structures using the most common structural feature: the *root structure* or *scaffold*. In effect, Rgroup decomposition is the reverse of scaffold-based enumeration. For example:

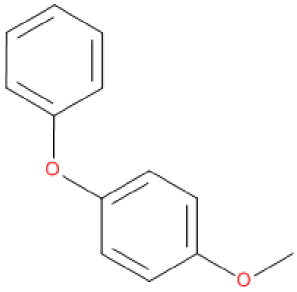


Structures to be decomposed:

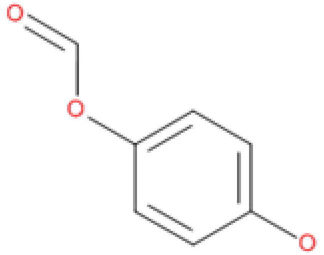
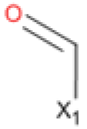
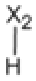
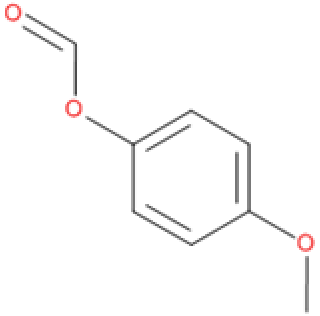
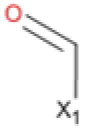
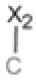
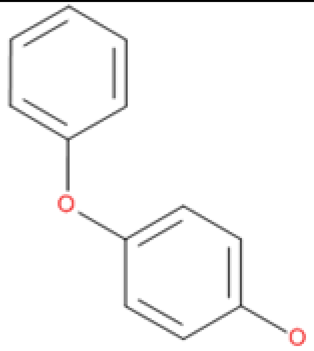
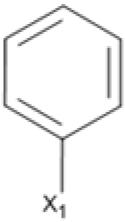
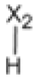
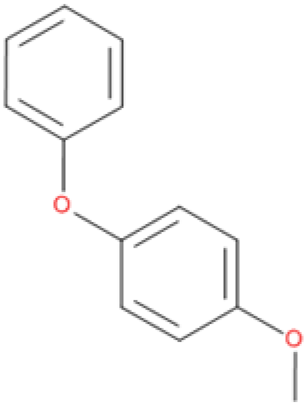
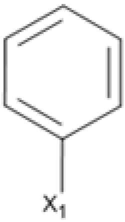
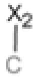








Insight for Excel displays the results of the Rgroup decomposition in a table of Rgroup substituents:

| | R1 | R2 | EC50 |
|---|---|---|---|
|  |  |  | 300 |
|  |  |  | 5 |
|  |  |  | 30 |
|  |  |  | 500 |

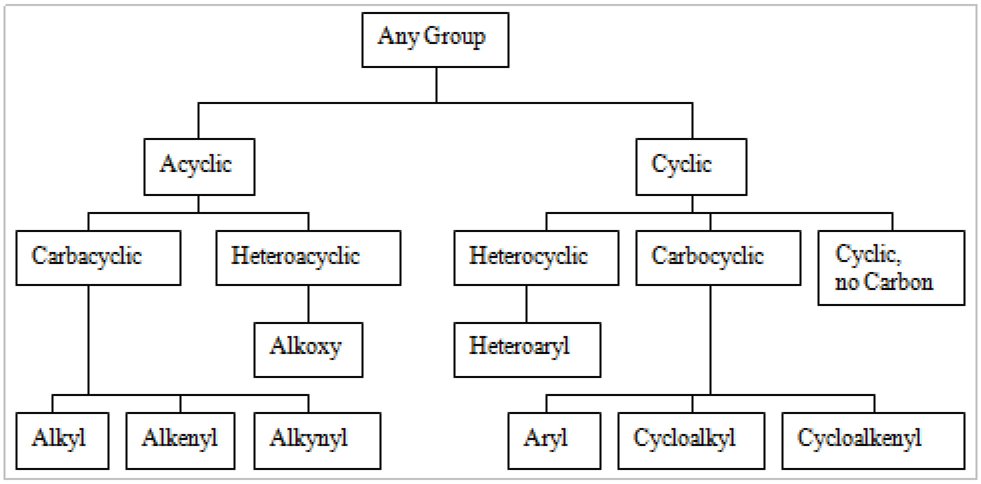
You can transfer the table to Microsoft Excel, where you can manipulate chemical structures and data in an Excel spreadsheet.

BIOVIA provides the capability for Rgroup decomposition in Pipeline Pilot with the *Generate RGroups* and *Generate SAR Information* components. The *Generate RGroups* component outputs the fragments as separate data records while *Generate SAR Information* annotates each input record with fragment SMILES. See the "Structure-Activity Relationship Tables" chapter of the *Advanced Chemistry Guide* in the *Pipeline Pilot Chemistry Collection* for more information.

# Chapter 5:
# Homology Groups

## About Homology Groups

A Homology Group is a single star atom with an abbreviation Sgroup which represents a structural feature, such as "alkyl group" or "acyclic group", that might be ambiguous. Homology group star atoms may be registered and used in substructure search queries. When used as a query, the homology group will match the group of atoms and bonds which it represents. For example an Alkenyl will match a singly connected group of carbons with no rings and with only single and double bonds between atoms in the group. The homology group will also match registered star atom homology groups. Using the same example, Alkenyl will match a registered Alkenyl, a registered Carbacyclic, a registered Acyclic or a registered *Any Group*.The hierarchy for homology groups is shown below:



For details on Homology Groups see *BIOVIA Direct Developers Guide* " Using BIOVIA Direct Homology Groups Searching and Registration".

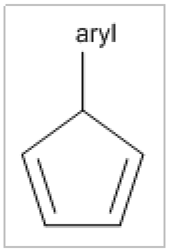
## Creating Substructure Search (SSS) Queries

You create a homology group query in BIOVIA Draw as follows:

1. Add a star atom, which has the atom symbol * or Zigzag line, at the point where you want the homology group. The star atom can stand alone or it can be bonded to one other atom. The star atom cannot have more than one bond connected to it.
2. Highlight the star atom and right click, then select `Create Template/Abbreviation`.
3. The `Create Chemical Template/Abbreviation` dialog box appears on the screen. In the field labeled Display abbreviation as:, specify the name of the homology group. Use one of the names shown in the hierarchical table above.

**IMPORTANT!** Although the name is not case-sensitive, you must spell the characters of the name exactly as shown in the hierarchical table above, including any spaces and commas. For example, `Any Group` has a space between the two words, and `Cyclic, no Carbon` has spaces and a comma. If the name does not match exactly (except for case), no error message appears, but BIOVIA Direct homology group search and registration operations will not work for that name.

For example:



You can use a name from any level in the hierarchy. For example, carbocyclic finds all types of rings and ring assemblies containing only carbon.

When used in a substructure search (SSS) the query returns all molecules containing:

- the specified homology group
- any specific homology groups that are subclasses of the specified homology group.

## Limitations

Molecules containing star atoms with homology group abbreviations can be registered. Similarity search is available for Homology Groups, but the results might not be intuitive when the database contains registered structures containing homology group abbreviations. BIOVIA therefore recommends using SSS instead of similarity search.

For registered structures, any star atom with homology group abbreviation must be connected to exactly one other atom. The bond type can only be single, double, or triple. If a homology group star atom stands alone or is connected to more than one other atom, an error is generated.

For an SSS query, the star atom with homology group abbreviation can be unconnected or connected to one other atom, however it cannot be connected to more than one atom.

# Chapter 6: Biopolymer Representation

## Configuring Databases for Biopolymer Registration and Searching

Over the years, chemical structure databases have used many formats for storing biopolymer molecules. The two most important formats in use today are full CTAB and SCSR:

- Full CTAB stores every atom and bond of the biopolymer molecule in the database. A protein with 100 amino acids will have around 900 atoms and 900 bonds which are stored in the molecule connection table (CTAB) inside the database. Sequence monomer information is encoded using abbreviation Sgroups (superatoms) which wrap each monomer and its leaving groups.
- SCSR (Self Contained Sequence Representation) stores one atom for each monomer in the biopolymer, plus the atoms and bonds for each distinct monomer that is present in the biopolymer. A protein with 100 amino acids, of which 20 are distinct, will have 100 SCSR template atoms and about 100 bonds between those atoms, plus about 180 additional atoms and bonds defining the monomers.

The HELM format is becoming popular as a way of entering biopolymers into the database, but it is not suitable for searching and must be converted to a standard molecule connection table during structure registration.

The SCSR format is generally preferred because it allows for storage of larger sequences. It is also easier to interconvert SCSR with the other formats, including HELM.

While these molecule connection table formats are interconvertible, generally speaking a structure in one format will not match the logically identical structure in a different format in an exact match or substructure search. Chemists creating databases using a modern version of BIOVIA Draw will generally use SCSR format. If their database also contains legacy structures in full CTAB format, a search using one format will usually not match the same biopolymer stored in the other format.

Starting with the BIOVIA Direct 2018 release, "small" sequences entered in SCSR format are stored as both SCSR structures and as full CTAB structures, which provides a partial solution to this searching problem. As long as all of the sequences stored in the database are below the expansion cutoff (the value of "small"), structure searches will work regardless of the format used to store the structures.

The definition of "small" is provided by the user when the BIOVIA Direct domain index is created, either as a maximum monomer count or a maximum molecular weight. In practice, we have found that performance becomes unacceptable when it exceeds about 100 monomers.

Storing both formats requires more storage space and processing time but allows queries and targets that do not exceed the threshold (the value of "small") to match each other regardless of the original structure format.

When an input structure is stored in SCSR format, all nucleic acid monomers and any amino acid monomers that have been modified (not present in the standard list of global SCSR templates) or are cross-linked are expanded on the fly into full CTAB units during indexing and searching. On-the-fly expansion has been available since BIOVIA Direct version 8. Starting with BIOVIA Direct 2018, all SCSR input structures below a certain size cutoff have all monomers expanded into full CTAB format, and the expanded full CTAB structure is stored in the index so that on-the-fly expansion is not required during matching.

Starting with the BIOVIA 2020 release, options are available to ignore the 3' and 5' terminal phosphates in substructure and exact matches, allowing nucleic acids represented with traditional SCSR template classes to match structures created from HELM format as long as the sugars, bases, and connecting phosphates are the same.

To see how queries match or do not match targets, see the Flexmatch Searching tables in the *BIOVIA Direct Administration Guide*. For more information, see the *BIOVIA Direct Administration Guide > Biopolymer Searching in Direct*.

## Condensed Representation of Biopolymers

Biopolymer residues can be represented abbreviated structures with the full structure (connection table or CTAB) underneath. The full structure convention preserves all chemical information on the sequence.

The full structure convention works well for oligopeptides and small proteins of less than 5,000 molecular weight. Chemists routinely work with polypeptide chains of as many as 3,000 amino acids. Because the average amino acid contains about 8 atoms, this implies a connection table of about 25,000 atoms. Searching and registering structures of this size can be extremely slow.

To address this need, BIOVIA provides additional conventions for representing biopolymers.

### Hybrid Representation

Biopolymers are potentially very large molecules with thousands of monomers. However, biopolymers are composed of a relatively small number of distinct monomer types. The hybrid representation uses templates to represent the residues, which allows for significant compression of the structure stored in the database and transmitted in molfile format between different programs. The hybrid representation retains the ability to elaborate the atoms and bonds in the full structure. Therefore, the hybrid representation is formally known as Self-Contained Sequence Representation (SCSR).

For example, the double repetition of alanine and glycine as AGAG takes only a few more lines than representing a single pair of AG. The entire alanine template can be represented as a single template label, A. Similarly, glycine can be compressed to G. Using template labels, such as A and G, is the preferred means of representing biopolymers.

A mechanism for converting legacy structures to the preferred format will be offered in future releases. In the meantime, BIOVIA supports both the hybrid representation and legacy structures, just as BIOVIA supports both the preferred V3000 format and the legacy V2000 format.

**See also**

*BIOVIA Direct Developers Guide* > Using BIOVIA Direct > Biopolymer Searching and Registration

*BIOVIA Direct Administration Guide*, in the chapter titled Creating and Managing Molecule Tables and Indexes, the topic "Biopolymer (sequence) molecule index".

BIOVIA Draw Help, **Administration of Biopolymer Structure Conventions: Concepts and Procedures**

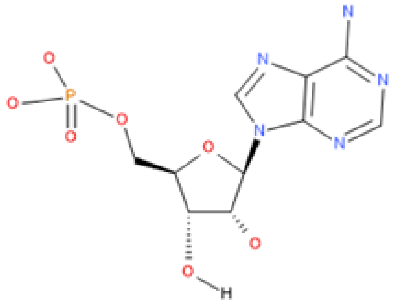
*BIOVIA CTFile Formats* - see Connection Table (CTAB) chapter's section on Template block. This document is available for download and is distributed with BIOVIA Direct, BIOVIA Draw, and BIOVIA Isentris.

### Representation of Nucleic Acids

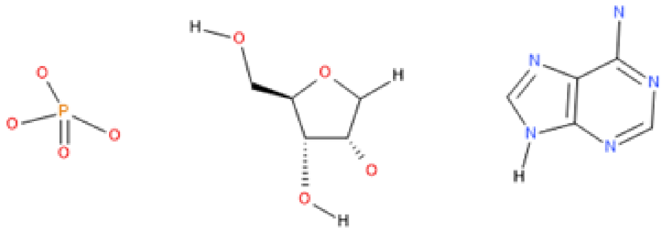
Nucleic acid sequences (RNA and DNA) are represented in the same way as other biopolymers, by using either SCSR templates or full CTAB format. The individual nucleotides in nucleic acid sequences can be represented in either of the following ways.

- **Traditional representation**: Each nucleotide is represented by a single SCSR template or SUP Sgroup that includes all three base, sugar, and 5’ phosphate moieties.

- **Granular representation**: Each sugar, base, and phosphate group is represented by a separate SCSR template that defines how they connect to each other. This granular representation is available starting with the BIOVIA 2020 release.



Traditional representation
(one template for entire nucleotide)



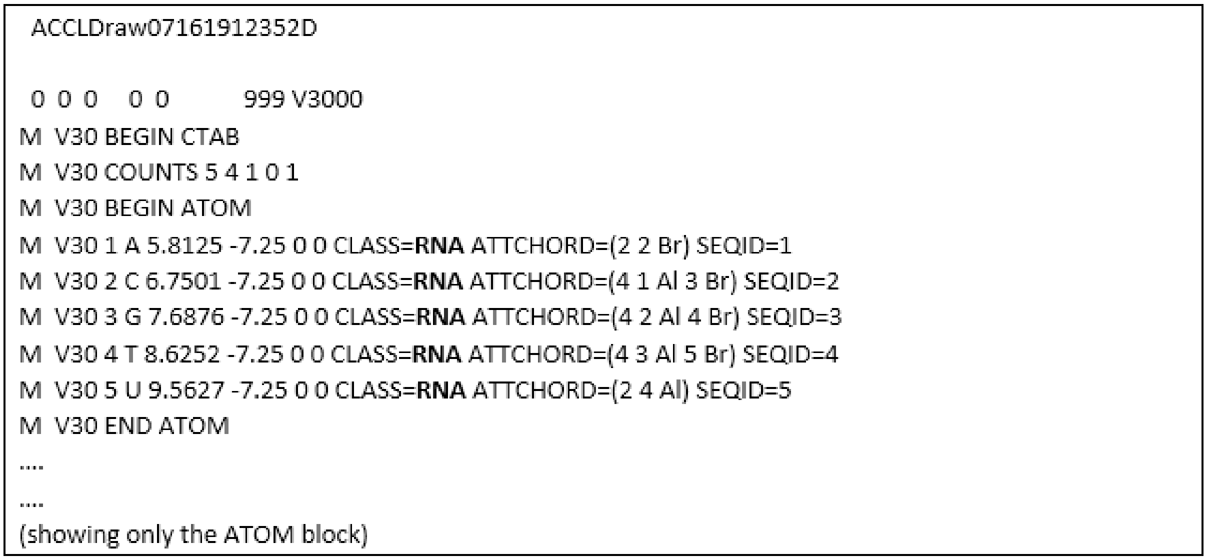
Granular representation
(separate templates for sugar, base and phosphate)

The different representations use different template classes to represent the nucleotide. In some cases, using a different representation can affect searches and matches.

### Traditional Representation

Only five nucleotides are needed to represent naturally occurring RNA nucleic acids – AMP (A), CMP (C), GMP (G), TMP (T), and UMP (U). The corresponding nucleotides can also represent DNA, with deoxyribose instead of ribose as the sugar. Therefore, the traditional representation uses the RNA and DNA abbreviation classes to specify the chemistry of the nucleotides in a condensed format.

For example, the following figure shows the traditional SCSR representation of the ACGTU sequence, with one sequence atom representing each nucleotide.



```
  ACCLDraw07161912352D

  0  0  0     0  0           999 V3000
M  V30 BEGIN CTAB
M  V30 COUNTS 5 4 1 0 1
M  V30 BEGIN ATOM
M  V30 1 A 5.8125 -7.25 0 0 CLASS=RNA ATTCHORD=(2 2 Br) SEQID=1
M  V30 2 C 6.7501 -7.25 0 0 CLASS=RNA ATTCHORD=(4 1 Al 3 Br) SEQID=2
M  V30 3 G 7.6876 -7.25 0 0 CLASS=RNA ATTCHORD=(4 2 Al 4 Br) SEQID=3
M  V30 4 T 8.6252 -7.25 0 0 CLASS=RNA ATTCHORD=(4 3 Al 5 Br) SEQID=4
M  V30 5 U 9.5627 -7.25 0 0 CLASS=RNA ATTCHORD=(2 4 Al) SEQID=5
M  V30 END ATOM
....
....
(showing only the ATOM block)
```

BIOVIA Draw exports templates in traditional representation, by default.

### Granular Representation

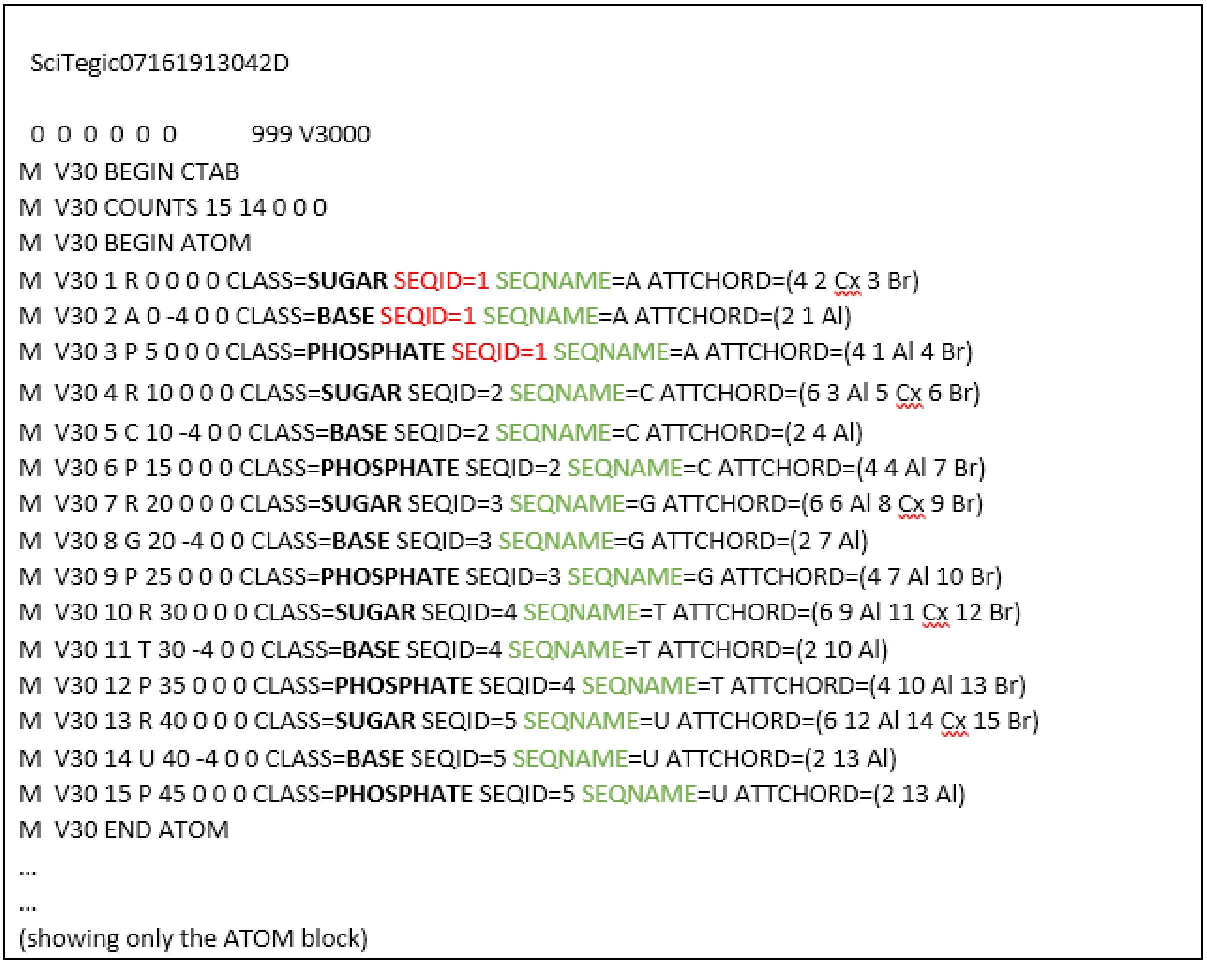
Starting with the BIOVIA 2020 release, the SCSR and full CTAB biopolymer formats can accommodate a more granular representation of nucleotides as separate sugar, base, and phosphate groups. The granular representation works well for representing oligonucleotides with modified bases, sugars, or phosphate groups, which are becoming widely used in drug design programs. It also allows more control over the position of the terminal phosphate groups (at the 3' or 5' end of the sequence).

The granular representation results in more consistent interconversion with the HELM format, which also represents nucleic acids with separate monomers for the sugar, base, and phosphate groups. For example, the HELM string for the ACGTU sequence in the previous example is:

`RNA1{R(A)P.R(C)P.R(G)P.R(T)P.R(U)P}$$$$`

where `R` represents Ribose, `P` represents Phosphate, and `A`, `C`, `G`, `T`, `U` represent the different Nitrogen bases.

The following figure shows the granular SCSR representation of an ATP nucleotide.



```
  SciTegic07161913042D

  0  0  0  0  0  0         999 V3000
M  V30 BEGIN CTAB
M  V30 COUNTS 15 14 0 0 0
M  V30 BEGIN ATOM
M  V30 1 R 0 0 0 0 CLASS=SUGAR SEQID=1 SEQNAME=A ATTCHORD=(4 2 Cx 3 Br)
M  V30 2 A 0 -4 0 0 CLASS=BASE SEQID=1 SEQNAME=A ATTCHORD=(2 1 Al)
M  V30 3 P 5 0 0 0 CLASS=PHOSPHATE SEQID=1 SEQNAME=A ATTCHORD=(4 1 Al 4 Br)
M  V30 4 R 10 0 0 0 CLASS=SUGAR SEQID=2 SEQNAME=C ATTCHORD=(6 3 Al 5 Cx 6 Br)
M  V30 5 C 10 -4 0 0 CLASS=BASE SEQID=2 SEQNAME=C ATTCHORD=(2 4 Al)
M  V30 6 P 15 0 0 0 CLASS=PHOSPHATE SEQID=2 SEQNAME=C ATTCHORD=(4 4 Al 7 Br)
M  V30 7 R 20 0 0 0 CLASS=SUGAR SEQID=3 SEQNAME=G ATTCHORD=(6 6 Al 8 Cx 9 Br)
M  V30 8 G 20 -4 0 0 CLASS=BASE SEQID=3 SEQNAME=G ATTCHORD=(2 7 Al)
M  V30 9 P 25 0 0 0 CLASS=PHOSPHATE SEQID=3 SEQNAME=G ATTCHORD=(4 7 Al 10 Br)
M  V30 10 R 30 0 0 0 CLASS=SUGAR SEQID=4 SEQNAME=T ATTCHORD=(6 9 Al 11 Cx 12 Br)
M  V30 11 T 30 -4 0 0 CLASS=BASE SEQID=4 SEQNAME=T ATTCHORD=(2 10 Al)
M  V30 12 P 35 0 0 0 CLASS=PHOSPHATE SEQID=4 SEQNAME=T ATTCHORD=(4 10 Al 13 Br)
M  V30 13 R 40 0 0 0 CLASS=SUGAR SEQID=5 SEQNAME=U ATTCHORD=(6 12 Al 14 Cx 15 Br)
M  V30 14 U 40 -4 0 0 CLASS=BASE SEQID=5 SEQNAME=U ATTCHORD=(2 13 Al)
M  V30 15 P 45 0 0 0 CLASS=PHOSPHATE SEQID=5 SEQNAME=U ATTCHORD=(2 13 Al)
M  V30 END ATOM
…
…
(showing only the ATOM block)
```

The SCSR template classes, `SUGAR`, `BASE`, and `PHOSPHATE`, represent the separate groups in a nucleotide. The `ATTCHORD` field defines how the groups connect to each other. The three groups in each nucleotide have the same value for the `SEQID` field, which corresponds to the position of the nucleotide in the sequence. An optional field, `SEQNAME`, can specify a name to represent the nucleotide when it is displayed in sequences.

Similar template classes are used when the nucleotides are represented by expanded SUP Sgroups. The `SUGAR`, `BASE`, and `PHOSPHATE` classes represent the separate groups of a nucleotide, and the `SEQID` and `SEQNAME` fields define the sequence properties.

Pipeline Pilot converts HELM strings to granular SCSR representation. BIOVIA Draw also provides options to export template libraries in granular representation. Users can sketch nucleic acid sequences with modified sugars, bases, or phosphates, and share libraries of modified templates with the Pipeline Pilot Chemistry Collection, Direct, and other BIOVIA applications.

**Note:** For information on managing shared libraries, see *Synchronizing Chemistry Configuration Files* on the Pipeline Pilot Help Center.

### Searching and Matching

Nucleotide sequences that have a 5’ terminal phosphate but do not have a 3’ terminal phosphate can be represented either with the traditional or with the granular SCSR template classes. When searching for or matching these types of nucleotide sequences, the two representations will match each other.

Nucleotide sequences that have a 3’ terminal phosphate and do not have a 5’ terminal phosphate, such as those typically created from HELM strings, cannot be represented with the traditional SCSR template classes. Therefore, these sequences will not match a sequence constructed from the traditional SCSR template classes, by default.

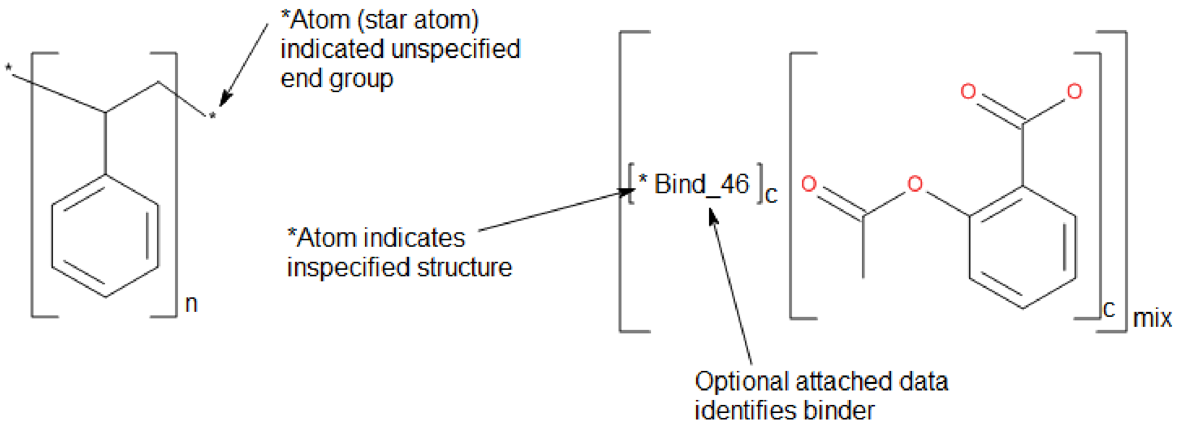
However, you can set options for substructure and exact matches to allow structures to match if they differ only in the presence or absence of 3' and 5' terminal phosphates. For example:

- In Pipeline Pilot, for the *Substructure Map* and *Exact Structure Map* components, select the *Ignore Terminal Phosphates* option.
- In BIOVIA Direct, use the *IgnoreTerminalPhosphates* flag.

These options are available starting with the BIOVIA 2020 release. Adding these options to your protocols or searches causes both the 3' and 5' terminal phosphates to be ignored in both queries and targets, allowing structures represented with traditional SCSR template classes to match structures created from HELM format as long as the sugars, bases, and connecting phosphates are the same.

## *Atoms Alone

A *atom generally represents a portion of a structure that is unknown or undefined. For example:


*Atom (star atom) indicated unspecified end group
n
*Atom indicates inspecified structure
* Bind_46
c
O
O
O
O
c
mix
Optional attached data identifies binder

A *atom has an atomic weight of zero. In a search query, a *atom hits any atom or group of atoms at that position. In the previous example, the data that is attached to the *atom (called attached data) distinguishes chemically different binders (Bind_46, Bind_45, Bind_01) in searching.

In *atom representation of biopolymers, each residue consists of a single *atom with attached data that distinguishes chemically different residues.

Consider the following information when you decide whether you want to use *atoms for condensed representation of biopolymers:

- The routine that is used to calculate molecular weight for registration and searching of BIOVIA databases cannot calculate accurate molecular weights for structures that use the *atom convention. To allow your users to search by molecular weight, you need to create a separate column in the molecule table of the database that contains the correct molecular weight. For more information on searching by molecular weight, see Molecular Weight of Standard Structures on page 205.

- Regarding the BIOVIA Draw implementation:
    - You must add the Sgroup field `MDL_RESIDUE_ATTACHMENT_ORDER` to your database to preserve information on the connectivity of atoms within the residue to structures outside it, such as protecting groups. For information on why this field is required, see Sgroup Field for Identifying Attachment Atoms on page 61.
    - You *must* also define the Sgroup field `MDL_STARATOM_NAME` in your database. This Sgroup field defines the attached data that distinguishes chemically different residues. For information on why this field is required, see Sgroup Field for *Atom Representation on page 62.

## Pseudoatoms Alone

The BIOVIA extended periodic table (extended Ptable) contains pseudoatoms, atom symbols that do not correspond to any of the chemical elements. The extended Ptable contains pseudoatoms for the 20 canonical natural amino acids, where the pseudoatom symbol corresponds to the amino acid three-letter abbreviation.

You can customize the default extended Ptable to include more pseudoatom symbols. For more information on Ptable customization, see Customizing the BIOVIA Ptable on page 206 .

> **Note:** Do not confuse pseudoatoms with abbreviated structures. Abbreviated structures can be expanded to reveal the underlying structure. By contrast, pseudoatoms consist of a single atom that cannot be expanded.

When you decide whether you want to use pseudoatoms for condensed representation of biopolymers, consider:

- Regarding the BIOVIA Draw implementation: You must add the Sgroup field `MDL_RESIDUE_ATTACHMENT_ORDER` to your database to preserve information on the connectivity of atoms within the residue to structures outside it, such as protecting groups. For information on why this field is required, see Sgroup Field for Identifying Attachment Atoms on page 61.
- The default extended Ptable does not specify the masses of the 20 canonical amino acid residues. If you want to provide your users with the ability to search biopolymers by molecular weight, your Ptable needs to specify weights for all pseudoatom symbols that you use in biopolymers. BIOVIA Draw provides an example Ptable that contains suitable weights. This example Ptable is located at `[Draw_home]\Examples\SequenceTool\ptable_sequencetool.dat`

For more information on molecular weight calculation, see Molecular Weight of Standard Structures on page 205.

- The BIOVIA Ptable is limited to a total of 200 entries, including the 103 natural elements, the 20 canonical amino acids, and a few specialized pseudoatoms provided by BIOVIA. If you need more symbols than the Ptable can provide, you must use *atoms to represent the structures.
- Searching performance for structures that are represented as pseudoatoms is slightly better than for structures that are represented as *atoms.
- The template atom convention is preferred if you want to represent biopolymer sequences in condensed form. If you use the pseudoatom convention, you will need to perform additional procedures to migrate your data to the BIOVIA preferred form, the template atom.

## *Atoms and Pseudoatoms

If you are already using pseudoatoms for biopolymer representation in your database, you might want to continue using this convention. To provide condensed representations for additional structures, however, BIOVIA recommends that you use *atoms rather than add entries to your customized Ptable.

## Summary of Biopolymer Structure Conventions

The table that follows summarizes the advantages and disadvantages of the conventions.

| Structure Convention | Advantages | Disadvantages |
|---|---|---|
| Hybrid representation (template representation of residues) (condensed) | Compact representation. | An expanded template definition does not match against the collapsed template atoms. |
| Full | Complete information on the biopolymer structure is stored in the database. The only convention that supports subsequence searching in Isentris. See Subsequence Searching in Isentris Applications on page 72. | Users may experience slow performance when searching and registering large biopolymer sequences (for example, peptides with molecular weight > 5,000). |
| *Atoms alone (condensed) | Better performance in searching and registration of large biopolymers. No need to customize your Ptable. | Storing solely the condensed representation in the database creates a slight risk of information loss. You must create the following Sgroup fields in your database: MDL_STARATOM_NAME and MDL_RESIDUE_ATTACHMENT_ORD ER. Requires special procedures for accurate searching by molecular weight. |
| Pseudoatoms alone (condensed) | Better performance in searching and registration of large biopolymers. No need to change existing customized periodic table (Ptable). | Storing solely the condensed representation in the database creates a slight risk of information loss. You must create the following Sgroup field in your database: MDL_RESIDUE_ATTACHMENT_ORD ER. Pseudoatoms provide slightly faster performance in searching and registration than *atoms. Space for pseudoatoms in the BIOVIA Ptable is limited. |

| Structure Convention | Advantages | Disadvantages |
|---|---|---|
| *Atoms and pseudoatoms (condensed) | Improved performance in searching and registration. No need to change existing customized periodic table (Ptable). If your customized Ptable is full, you can use *atoms to represent additional residues. | Storing the condensed representation in the database creates a slight risk of information loss. You must create the following Sgroup fields in your database: MDL_STARATOM_NAME and MDL_RESIDUE_ATTACHMENT_ORD ER. Requires special procedures for accurate searching by molecular weight. |

## Compatibility of Sequences Created in ISIS/Draw

This section describes compatibility issues between the default configuration of the ISIS/Draw Sequence Tool and the Sequence Tool in BIOVIA Draw. If you have customized the ISIS/Draw Sequence Tool, these differences might not apply.

### ISIS/Draw Sequences Use the Full Structure Convention

Sequences that are created in ISIS/Draw always use the full-structure convention for representation of biopolymers.

If you want to use a condensed representation of biopolymer residues, you need to convert existing structures to the condensed representation. BIOVIA Cheshire can assist with this conversion.

### ISIS/Draw Residue Templates Lack Explicit Attachment Atoms

Except for cysteine, peptides that are created in ISIS/Draw lack a third attachment atom and explicit hydrogen leaving group on amino acids with reactive side chains, such as serine. These explicit attachment atoms enable users of BIOVIA Draw to easily attach protecting groups to contracted residues.

These differences do not affect structural equivalence for registration and searching.

For more information, see Explicit Hydrogen Leaving Groups on Attachment Atoms of Reactive Chains on page 67 and Explicit Hydrogen Leaving Groups on Histidine on page 69.

### Stereochemistry of ISIS/Draw Residue Template

The stereochemistry of chiral amino acid residues in ISIS/Draw is incompletely defined, while the stereochemistry of these residues in BIOVIA Draw is fully defined. That is, the BIOVIA Draw residues specify that the absolute configuration of each stereogenic center is known, while stereogenic centers in ISIS/Draw lack a specification.

The structures of DNA and RNA residues in ISIS/Draw lack stereo bonds, while the corresponding residues in BIOVIA Draw have the correct stereo bonds and specify that the absolute configuration is known.

Consequently, a biopolymer search query that you create in BIOVIA Draw cannot retrieve a sequence that was created in ISIS/Draw in a substructure search. To address this issue, the best long-term solution is to convert existing structures in your database to the full stereochemistry. In the short term, you can edit the residue templates for BIOVIA Draw to match the stereochemistry of the corresponding structures in ISIS/Draw.

For more information on BIOVIA representation of tetrahedral stereochemistry, see Tetrahedral Stereochemistry on page 9. For information on how to specify configuration at stereogenic centers, see the following index entry in the *BIOVIA Draw Online Help*: Enhanced stereochemistry, concepts and examples.

## Compatibility of Condensed and Full Structure Conventions

Full and condensed representations of biopolymers are not equivalent in searching. For example, a sequence that uses the template atom convention is not equivalent to one that uses the full structure convention. If your database primarily consists of small peptides, and you now want to store larger proteins, you need to consider whether to convert existing structures from full to a condensed representation:

- If it is important that scientists who work on small peptides and on proteins be able to retrieve both types of structures with a single chemical structure query, then you should convert existing structures to a condensed representation.
- If scientists who work with small peptides rarely need to retrieve larger structures, however, then you might consider using the full representation for small peptides and a condensed representation solely for large proteins.

## Required Sgroup Fields for Biopolymer Representation

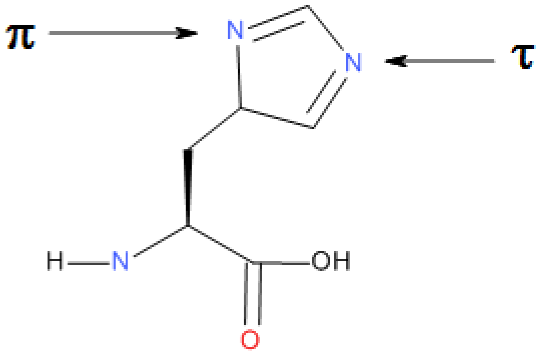
A substance group (Sgroup) field is a special kind of database field that assigns meaning to a specific type of attached data that is stored with the structure. This section explains the function of the attached data in condensed representation of biopolymers. Both types of attached data are invisible to end-users of BIOVIA Draw.

For detailed information on the characteristics of these fields, and for procedures for creating these fields in your database, see Sgroup Fields on page 154.

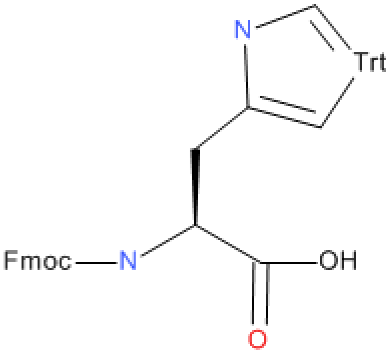
### Sgroup Field for Identifying Attachment Atoms

Add the Sgroup field `MDL_RESIDUE_ATTACHMENT_ORDER` to your database if you want to represent biopolymer residues in condensed legacy forms: *atom, pseudoatom, or a combination of the two.

Each biopolymer residue has at least two predefined attachment atoms that the BIOVIA Draw Sequence Tool uses to connect the abbreviated residues within a sequence. Amino acid templates with reactive side chains have a third attachment atom so that users can add or remove protecting groups without expanding the structure of the residue. The histidine side chain has two reactive sites (the π and τ nitrogens), so the histidine template has two additional attachment atoms.



When you use a condensed representation of biopolymer residues, you might lose information on which of the defined attachment atoms are bonded to particular protecting groups. In the following example, an amino terminal histidine residue has an Fmoc protecting group on the N-terminal amino group and a trityl protecting group on the τ nitrogen:



Fmoc and Trt are abbreviations for the Fmoc and trityl protecting groups, respectively. When you contract the structure to a condensed representation of the residue (either a *atom or a pseudoatom), the information on whether the Fmoc is attached to the amino group, the τ nitrogen, or the π nitrogen is lost.

To preserve information on group attachments, BIOVIA Draw attaches Sgroup data to each group that identifies the atom on the residue to which the group is attached. This Sgroup data is associated with the Sgroup field `MDL_RESIDUE_ATTACHMENT_ORDER`.

### Sgroup Field for *Atom Representation

Add the Sgroup field `MDL_STARATOM_NAME` to your database if you want to represent any biopolymer residues as *atoms, or to use *atoms to represent polyethylene glycol (PEG) groups. The data that is attached to each *atom distinguishes chemically different residues. For example the attached data for a *atom that represents a serine residue is "AA-Ser", while the attached data for a *atom that represents a threonine residue is "AA-Thr". This Sgroup data is associated with the Sgroup field `MDL_STARATOM_NAME`.

## Structures Used in Biopolymer Representation

This section describes the two types of structures that are used in biopolymer representation in BIOVIA Draw: biopolymer residue templates and templates for single-attachment groups. These structures are available as templates on the top toolbar of the BIOVIA Draw Sequence Tools.

### Biopolymer Residues

The BIOVIA Draw Sequence Tools enable you to draw any of the following biopolymer sequences by typing the following symbols:

- 1-letter amino acid sequences and 3-letter amino acid sequences.

  **Note:** The 1-letter and 3-letter options are purely for display. A search query that uses 3-letter abbreviations matches sequences that use the 1-letter abbreviations.

- DNA sequences
- RNA sequences

You can customize the BIOVIA Draw Sequence Tools to add your own symbols for non-natural amino acids and/or non-natural nucleotides. See the *BIOVIA Draw Configuration Guide* topic on Customizing BIOVIA Draw for Registration and Searching of Biopolymers.

## Single-attachment Groups

Chemists often derivatize amino acid side chains and terminal groups with protecting groups or polyethylene glycol (PEG) molecules. These groups are called single-attachment groups because they contain a single attachment atom for bonding to a biopolymer residue. This section contains guidelines for representing these groups in biopolymer sequences.

You can customize the BIOVIA Draw Sequence Tools to add symbols for additional single-attachment groups.

### Protecting Groups

Protecting groups are small molecules of defined structure, for example t-butyl or benzyl. Protecting groups might be added and removed several times in the course of a peptide synthesis.

The example protecting groups in BIOVIA Draw use the full-structure convention. If your database contains many structures with protected amino acids, however, you might want to use *atoms to represent protecting groups.

### PEG Molecules

PEG molecules consist of a variable number of repeating units of polyethylene glycol and are characterized by average molecular weight, typically 750 to 20,000 Daltons. Therapeutic proteins can be conjugated with polyethylene glycol (PEG) molecules to improve bioavailability.

BIOVIA recommends that you represent the PEG moiety as a *atom with associated Sgroup data, even if you choose to represent biopolymer residues as the complete structure. Chemists generally do not need to change the chemistry of the PEG moiety, so there is no need to expand the abbreviation to see the underlying structure.

# Special Features of Abbreviated Structures

This section provides reference information on structural features that are unique to templates for biopolymer residues.

## Template Format for Biopolymer Residues

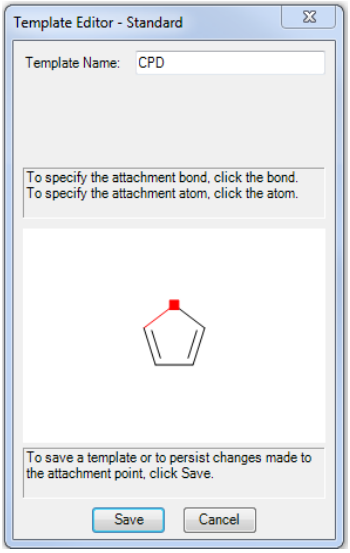
Chemical structure templates in earlier versions of BIOVIA Draw contain a single attachment atom by which you can attach an abbreviated template structure to an atom in another structure.

- You can draw bonds between two abbreviated template structures without expanding the abbreviations. For example, you can draw a bond between an abbreviated protecting-group template and an abbreviated amino acid residue in a sequence.
- You can specify multiple attachment atoms on a template. These attachment atoms have special attributes that enable you to define the ways that enhanced templates attach to each other.

In BIOVIA Draw, the toolbars that display in the Sequence Tools all contain enhanced templates. Other template toolbars and directories in BIOVIA Draw contain standard templates.

> **Note:** Template format does not affect structural equivalence in searching or registration. For example, a query that is constructed using single-attachment templates in standard format can find structures that contain these templates in enhanced format.

Use the Standard Template Editor dialog in BIOVIA Draw to create and edit standard templates:



Template Name: CPD

To specify the attachment bond, click the bond.
To specify the attachment atom, click the atom.

To save a template or to persist changes made to the attachment point, click Save.

For information on creating and editing standard templates, see the *BIOVIA Draw Online Help*.

For information on the special attributes of enhanced templates, see the remaining topics in this section.

## Abbreviation Class

The abbreviation class specifies the chemistry of the abbreviated structures in biopolymer residues. Allowed values for abbreviated structures of residues are shown in the table that follows:

| Residue Type | Meaning |
|---|---|
| AA | Amino acid. Identifies the abbreviated structure as an amino acid residue. |
| DNA | DNA. Identifies the abbreviated structure as a deoxyribonucleotide residue. |
| RNA | RNA. Identifies the abbreviated structure as a ribonucleotide residue. |

In the molfile, the abbreviation class is specified in the SCL tag. For more information on molfile tags, see Molfile Features in Biopolymer Templates on page 70.

For more information on abbreviated structures, see Abbreviated Structures on page 30.

## Terminal Leaving Groups

The Sequence Tool allows you to create biopolymer sequences by typing standard residue symbols at the keyboard. As you type each letter of the sequence, BIOVIA Draw automatically adds the residue template that corresponds to that structure. For example, if you want to enter the amino-acid sequence "ASER" using 1-letter abbreviations, you choose the 1-letter Amino Acid Sequence tool from the palette and type "ASER" in the drawing area. The following figure shows what you see as you type:

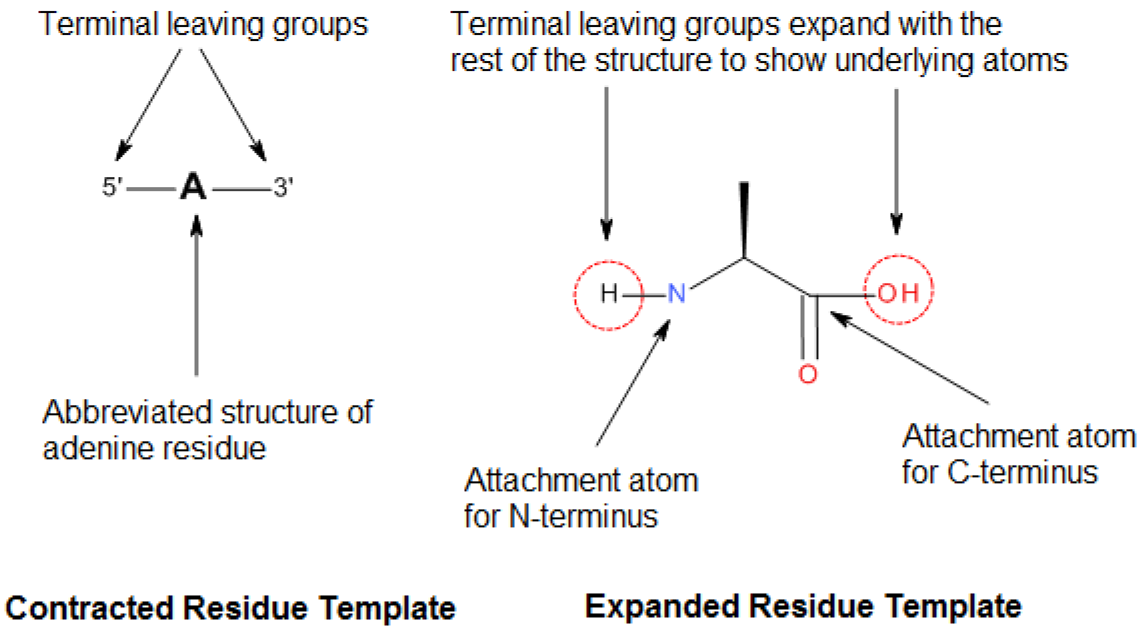
H-A-OH

H-A S-OH

H-A S E-OH

H-A S E R-OH

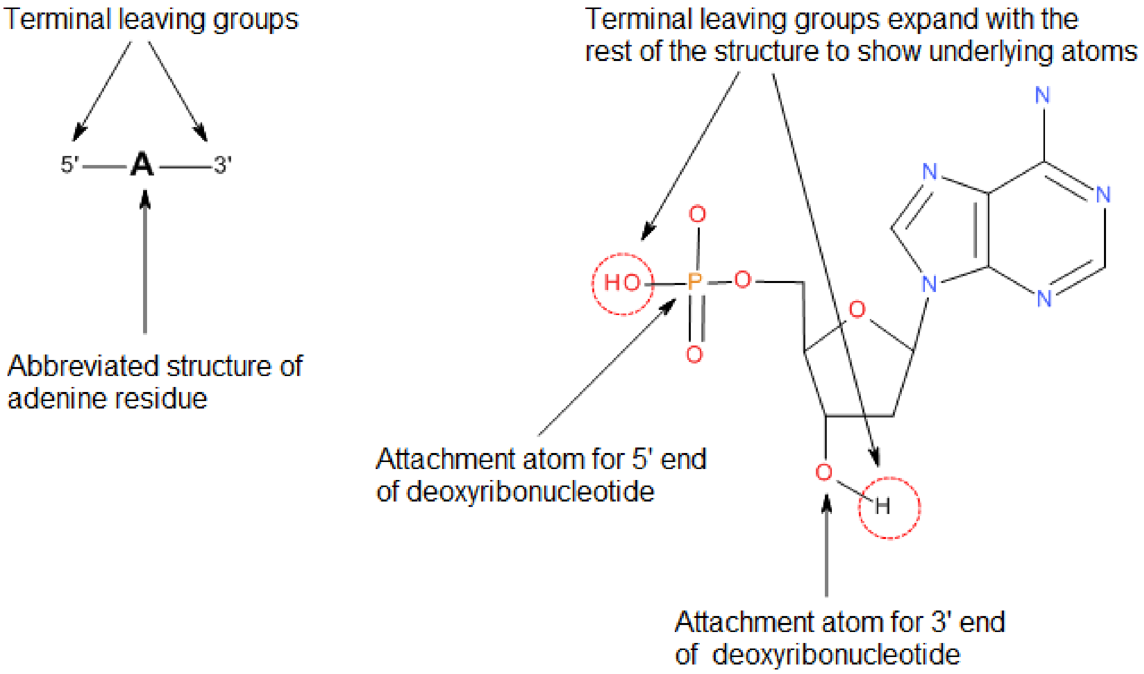
Each template for a biopolymer residue consists of an abbreviated structure (abbreviation Sgroup) with at least two attachment atoms. Each of these attachment atoms has an associated terminal leaving group (or terminal group), so called because the groups leave the structure when you join residues together. The following figure shows the residue template for L-alanine with the residue structure and leaving groups:


Terminal leaving groups
5'
A
3'
Abbreviated structure of adenine residue
Terminal leaving groups expand with the rest of the structure to show underlying atoms
H
N
OH
O
Attachment atom for N-terminus
Attachment atom for C-terminus
Contracted Residue Template
Expanded Residue Template

The H and OH are abbreviations for the atoms that are lost when a peptide bond is formed. The loss of these atoms when you type the residue symbols ensures that the chemistry of the sequence is correct.

Thus, every amino acid residue template contains at least three abbreviated structures: one for the residue structure, and one each for the N-terminal and C-terminal leaving groups:

- The abbreviated structure for the residue, which has an abbreviation class of AA and a text label (SMT tag) that represents the residue abbreviation.
- The abbreviated structure for the N-terminal leaving group, which has an abbreviation class of LGRP and a text label (SMT tag) of H.
- The abbreviated structure for the C-terminal leaving group, which has an abbreviation class of LGRP and a text label (SMT tag) of OH


Terminal leaving groups
5'
A
3'
Abbreviated structure of adenine residue
Terminal leaving groups expand with the rest of the structure to show underlying atoms
N
N
N
N
N
O
HO
P
O
O
O
O
H
Attachment atom for 5' end of deoxyribonucleotide
Attachment atom for 3' end of deoxyribonucleotide

**Contracted Residue Template**

**Expanded Residue Template**

Residue templates for DNA and RNA are constructed using similar principles. For example:

For nucleotide residues, the abbreviations for the leaving groups have names that show the polarity of the sequence: 5' is the abbreviation for the hydroxyl group that is lost in the ligation reaction, and 3' is the abbreviation for the hydrogen atom that is lost.

## Order and Bond Matching Attributes of Attachment Atoms

The order attribute of an attachment atom controls the order in which other templates can attach to the residue. The bond matching attribute ensures that the correct atoms in the abbreviated templates are linked together when you type the residue symbols, or when you draw bonds between residues. An attachment atom with bond matching=left joins to an attachment atom with bond matching=right. For peptides:

- The attachment atom at the N-terminus has order=1 and bond matching=left.
- The attachment atom at the C-terminus has order=2 and bond matching=right.

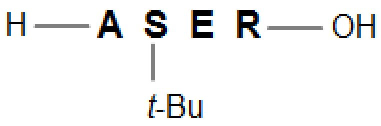
Amino acid residues with reactive side chains can have additional attachment atoms, as described in the following sections.

### Attachment Atoms on Reactive Chains

If you want to add the t-butyl protecting group to the hydroxyl group of the serine residue in the following amino acid sequence.

H—**A S E R**—OH

You can attach the protecting group directly to the serine side chain:



All amino acids with reactive side chains have the reactive atom defined as a third attachment atom. These attachment atoms must have the following properties:

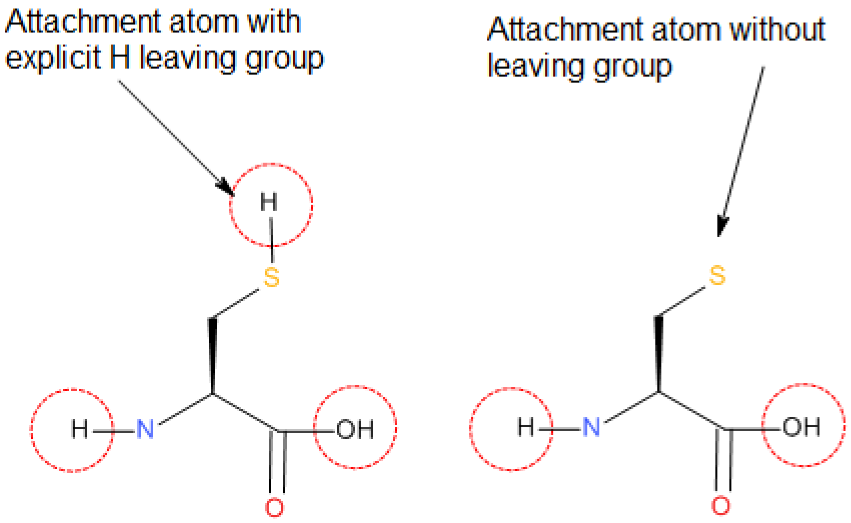
- Bond matching = cross (for crosslink)
- Order=3 or higher.

> **Note:** For cysteine, defining the sulfur as an attachment atom also allows you to draw disulfide bonds between abbreviated cysteine residues.

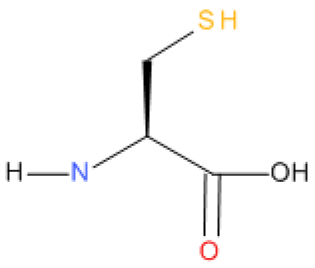
The histidine side chain has two possible bonding sites for protecting groups, the π nitrogen and the τ nitrogen, and therefore the side chain has two attachment atoms.

### Explicit Hydrogen Leaving Groups on Attachment Atoms of Reactive Chains

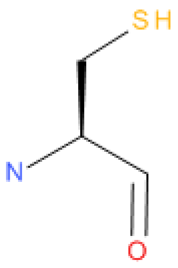
The templates for amino acid residues in BIOVIA Draw all have explicit hydrogen leaving groups attached to the attachment atom on reactive side chains. For example:


Attachment atom with explicit H leaving group
Attachment atom without leaving group
H
S
H
N
OH
O
S
H
N
OH
O

The explicit hydrogen leaving group is necessary to support the use of molecular formula and molecular weight calculations for biopolymer residues that use the pseudoatom representation. The structure of the cysteine residue shows this:

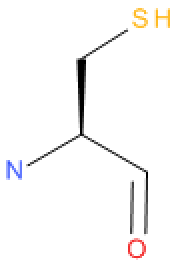


Amino acid template with leaving groups:
Molformula = C3 H7 N O2 S
Formula Wt = 121.1581

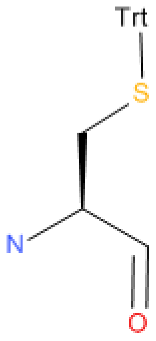


Amino acid residue within a sequence:
Molformula = C3 H5 N O S
Formula Wt = 103.1428

The hydrogen atom that is attached to the sulfur is an implicit hydrogen. This implicit hydrogen is not actually present on the structure and is not registered to the database. When you register a structure to a database, the routine that calculates formula weight recognizes that it needs to add an H to obtain a correct formula weight. On the other hand, if a protecting group is attached to the cysteine sulfydryl group, the calculation routine does not add a hydrogen when it calculates the formula weight:



Unprotected amino acid residue:
Molformula = C3 H5 N O S
Formula Wt = 103.1428

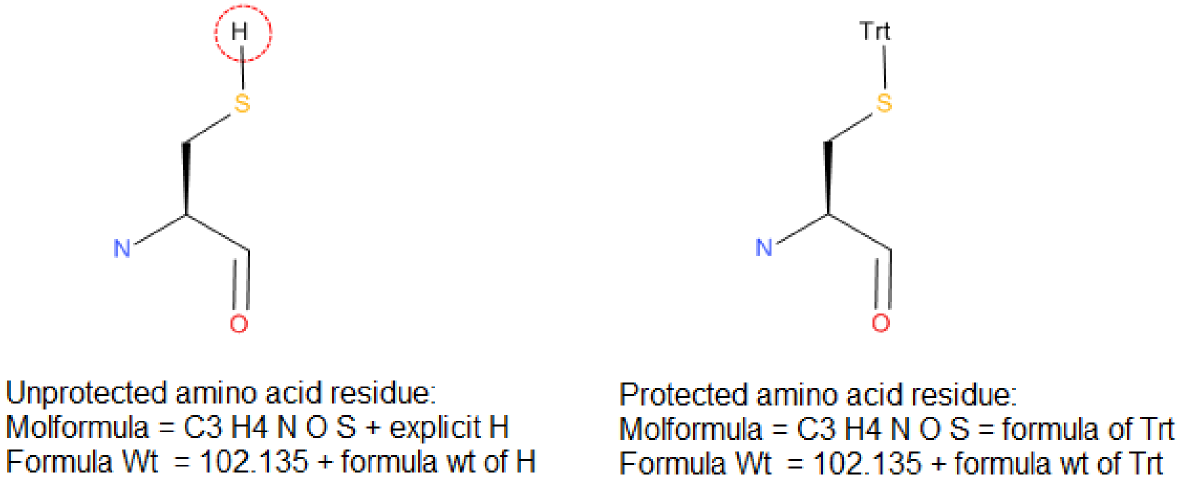


Protected amino acid residue:
Molformula = C3 H4 N O S = formula of Trt
Formula Wt = 102.135 + formula wt of Trt

The calculation routine can do this because the sulfur atom has an assigned valence of 2. However, pseudoatoms do not possess an assigned valence. Consequently, if your database represents amino acid residues as pseudoatoms, the algorithm can no longer count the atoms within it, and must obtain the formula weight from the atomic weight of the pseudoatom definition in the Ptable:

- If you assign the pseudoatom an atomic weight that excludes the fifth hydrogen (102.135), then molecular weights for protected cysteines are correct, but molecular weights for cysteine residues with free sulfhydryl groups will be too low by one hydrogen atom.
- If you assign the pseudoatom an atomic weight that includes the fifth hydrogen (103.1428), then molecular weights for unprotected cysteines will be correct, but residues with protected cysteines will be too high by one hydrogen atom.

The solution to this problem is to attach an explicit hydrogen leaving group to the internal attachment atom, and to assign the pseudoatom an atomic weight that excludes the fifth hydrogen (102.135):



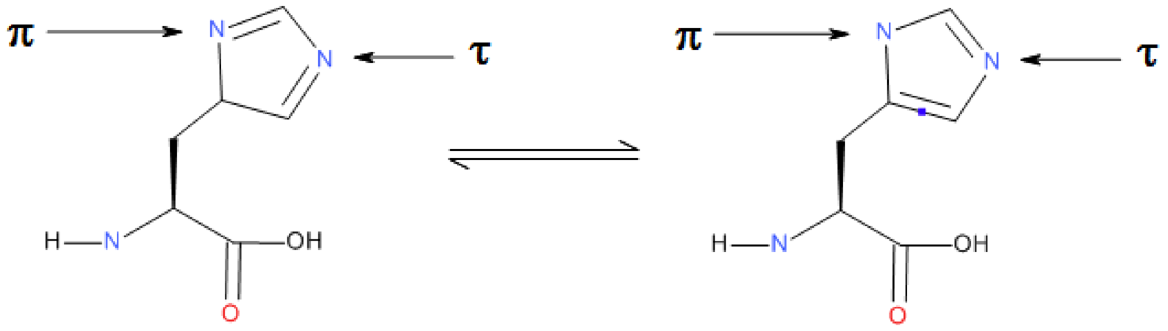
Unprotected amino acid residue:
Molformula = C3 H4 N O S + explicit H
Formula Wt = 102.135 + formula wt of H

Protected amino acid residue:
Molformula = C3 H4 N O S = formula of Trt
Formula Wt = 102.135 + formula wt of Trt

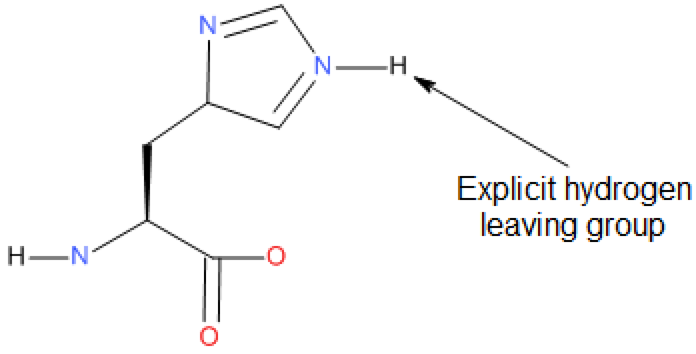
**Note:** The routine that BIOVIA supplies for calculating molecular weights for registration to databases uses an algorithm that is different from that of the Calculator command in the BIOVIA Draw Chemistry menu. The Calculator command in BIOVIA Draw always calculates molecular weight and molecular formula correctly, regardless of whether you use pseudoatoms, *atoms, or full structures to represent biopolymer residues. For information on calculation of molecular weights for registration, see Molecular Weight of Standard Structures on page 205.

### Explicit Hydrogen Leaving Groups on Histidine

The histidine side chain has two reactive nitrogens that are part of a tautomeric ring:



Consequently, the template for histidine contains one explicit hydrogen leaving group attached to the τ nitrogen, which is the most commonly protected ring nitrogen:



The attachment atoms on the histidine side chain are defined as follows:

- τ nitrogen: order=3, bond matching=cross, explicit hydrogen leaving group
- π nitrogen: order=4, bond matching=cross, no leaving group

## Molfile Features in Biopolymer Templates

This section provides additional details on the special features of biopolymer templates that can be seen by viewing the template molfile in a text editor. For complete information on all features of the molfile format, see the document, *CTFile Formats*, which is available by free download from the BIOVIA web site.

### Abbreviation Class (SCL)

The abbreviation class (SCL) molfile tag is present solely in template molfiles in enhanced format. The abbreviated structure for the residue must have one of the abbreviation classes AA, DNA, or RNA, as described in Abbreviation Class on page 64.

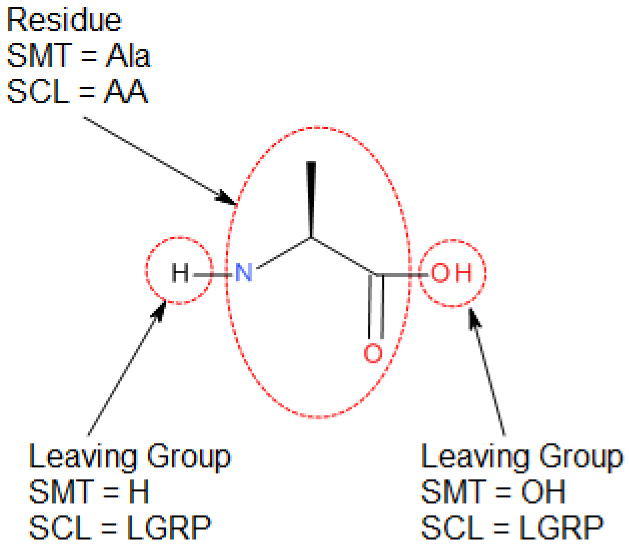
Another abbreviation class called leaving group (LGRP) identifies an abbreviated structure as a leaving group on a biopolymer residue. Each template for a biopolymer residue contains at least two LGRP abbreviations.

**Note:** Abbreviation class was originally referred to as superatom class, hence the tag name SCL. The term superatom was dropped to avoid confusion with pseudoatom.

### Sgroup Subscript (SMT)

The Sgroup subscript (SMT tag) is the text label that displays when an abbreviated structure (abbreviation Sgroup) is contracted. Abbreviated structures in both standard and enhanced templates have SMT tags.

The figure that follows shows the SMT tags on the three abbreviated structures that comprise a biopolymer residue template:



### Abbreviation Attachment Atom (SAP)

The abbreviation attachment atom (SAP) molfile tag is present solely in template molfiles in enhanced format.

> **Note:** The abbreviation attachment atom was originally referred to as the superatom attachment point, hence the tag name SAP.

Templates in standard format contain a single attachment atom that is defined as the first atom in the connection table (CTAB). Templates in enhanced format use the SAP tag to specify the attachment atom and its attributes. The following example shows the SAP entries for the four attachment atoms in the template for L-histidine:

```
M  SAP  4  4  1  0  Dx  6  2  Al  5  3  Br  13  11  Cx
```

For each attachment atom, the letter specifies the order (A=1, B=2, and so on), and the character that follows specifies bond matching (l=left, r=right, and x=cross).

In this example:

- Cx identifies the τ nitrogen, which is the third attachment point and has a bond matching attribute of cross.
- The numbers that precede the letter C identify the attachment atom and the atom in the leaving group to which it is attached. The τ nitrogen is identified as the thirteenth atom in the atom table in the molfile, and is attached to the eleventh atom, which is the explicit hydrogen leaving group.
- Dx identifies the π nitrogen, which is the fourth attachment atom and also has a bond matching attribute of cross. The π nitrogen is identified as the first atom in the molfile. The π nitrogen lacks an explicit hydrogen leaving group, so the number of the atom in the leaving group is zero.

## Creating and Enforcing Conventions for Biopolymer Representation

The structure convention that you choose for biopolymer representation becomes part of your company's business rules for chemical structure registration. This section describes how to use *standard abbreviations* to enforce these business rules.

### Choose One Convention for Each Chemical Entity

Structures that use different conventions for representing biopolymer residues are not equivalent in searching. For example, none of the following representations of L-alanine retrieves the full representation of L-alanine in a search:

- template atom
- *atom
- pseudoatom

Also, the template atom and *atom representations of L-alanine do not retrieve the pseudoatom representation.

Therefore, if you want your users to be able to find any biopolymer sequence in the database, be sure to use the same convention for each type of biopolymer residue and single-attachment group.

### Use Standard Abbreviations

The BIOVIA Draw Sequence Tools allow you to create sequences either by typing symbols at the keyboard or by clicking template tools on the toolbar for a sequence tool. As you draw the sequence, BIOVIA Draw converts the symbols or template tools that you choose into abbreviated structures (abbreviations) in the sequence.

To enforce your business rules for biopolymers, you must ensure that everyone at your company uses the same template for each chemical entity. These abbreviation templates should be write-protected so that end users cannot accidentally alter them.

Unlike other toolbars in BIOVIA Draw, structures on the toolbar in the sequence tool are defined in a special file called a standard abbreviation definitions file. This file consists of a set of entries for each Sequence Tool, plus a set of single-attachment groups. Each entry within a set consists of, at minimum:

- The `displayName` attribute, which is the symbol for the biopolymer residue that the user types and that displays on the toolbar.
- The `molfile` attribute, which specifies the path to the molfile that contains the full structure of the residue or single-attachment group as an abbreviated structure in the contracted state.

In addition, each entry in the standard abbreviation definitions file can contain additional attributes that are specific to the chemistry of the corresponding chemical entity and the convention used to represent it (full, *atom, or pseudoatom).

Expanding a residue or single-attachment group that is defined in the standard abbreviation definitions file causes BIOVIA Draw to display the expanded structure that is specified by the molfile attribute in the standard abbreviations file. When you contract the abbreviation, BIOVIA Draw replaces the expanded structure with an abbreviated structure that contains either the full structure, a *atom, or a pseudoatom, as specified by the corresponding entry in the standard abbreviations file.

Thus, the differences between full and condensed representations are transparent to users of BIOVIA Draw. Biopolymer structures look and behave the same, regardless of whether the contracted abbreviations contain *atoms, pseudoatoms, or complete structures.

# Subsequence Searching in Isentris Applications

## Subsequence Search Differs from Substructure Search

The substructure search (SSS) that is provided in Isentris applications is not necessarily useful for sequence searching. For example, the query sequence "GGG" as a full structure representation retrieves structures similar to the following:

GGG AAF FFR NLP RFL

That is, the search retrieves all tripeptides that contain glycylglycylglycine as a substructure. Because all 20 canonical natural amino acids contain glycine as a substructure, the query retrieves all combinations of the 20, or up to 203 different tripeptides.

This searching behavior also applies to amino acids whose structures are embedded within other amino acids. For example, alanine is a substructure of 18 of the 20 amino acids, and valine is a substructure of isoleucine.

The template, *atom, and pseudoatom representations only hit individual amino acids if they have the same representation in the database, such as identical collapsed templates or identical template atoms.

In most cases, chemists who use a sequence as a query want results that are similar to those of the BLAST program. That is, they would expect the query "GGG" to retrieve biopolymer sequences that contain the query sequence embedded within them, as shown in the following example (the query sequence is in boldface):

GGG

GFGGG GGGDGT

ALLGGGDGT AMFGGGDLG

This type of search is a subsequence search.

## BIOVIA Draw Programming Interface (API) for subsequence search

Manually applying these changes to even the smallest sequences is extremely cumbersome. For this reason, BIOVIA Draw API provides options that automatically apply these changes to sequences that are used in search queries. For details, see the BIOVIA Draw API Reference for information on:

- Renderer.MolfileStringForSubsequenceQuery
- Renderer.ChimeStringForSubsequenceQuery
- StructureConverter.ConvertMolfileStringToSubsequenceQuery

> **Note:** The *BIOVIA Draw API Reference* is available from `Start > Programs > BIOVIA > BIOVIA Draw [version] > BIOVIA Draw Documentation`.

## Implementing Subsequence Search for Full Structure Representations

Provided the molecules have full structure representation, you can use substructure search for subsequence search by performing the following edits on your substructure search query in BIOVIA Draw:

1. Delete the terminal groups (N-terminal hydrogen and C-terminal hydroxyl). The presence of the terminal groups prevents a match except with a sequence of the same length.
2. Delete non-terminal leaving groups (for example, the explicit hydrogen on cysteine). The presence of these groups can prevent a query from finding sequences with substituents on side chains. For example, a serine in the query does not find sequences in which serine has a protecting group on its side chain. Similarly, a query that contains cysteine with a free sulfhydryl group cannot find sequences that contain a cysteine in which the sulfur atom is part of a disulfide bond.
3. Apply the Substitution as Drawn (s*) query feature to all non-hydrogen atoms within each residue. The Substitution as Drawn query feature is necessary solely if you use the full-structure convention for sequence representation. Applying Substitution as Drawn ensures that glycine retrieves solely glycine; alanine retrieves solely alanine, and so on.

# Related Documentation for BIOVIA Draw

See the *BIOVIA Draw Configuration Guide* topics on:

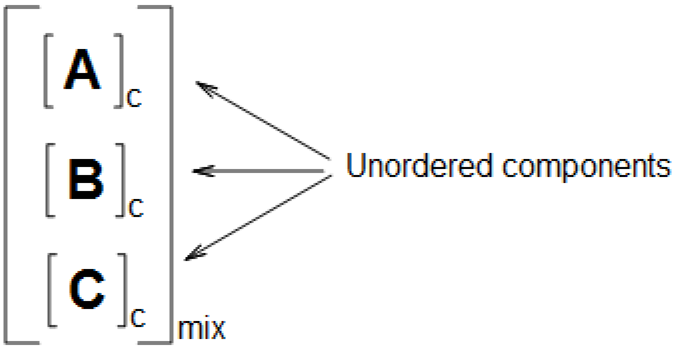
- Enabling the Data Sgroup Tool to Control Sgroup Data
- Customizing BIOVIA Draw for Registration and Searching of Biopolymers

# Chapter 7: Mixture Representation

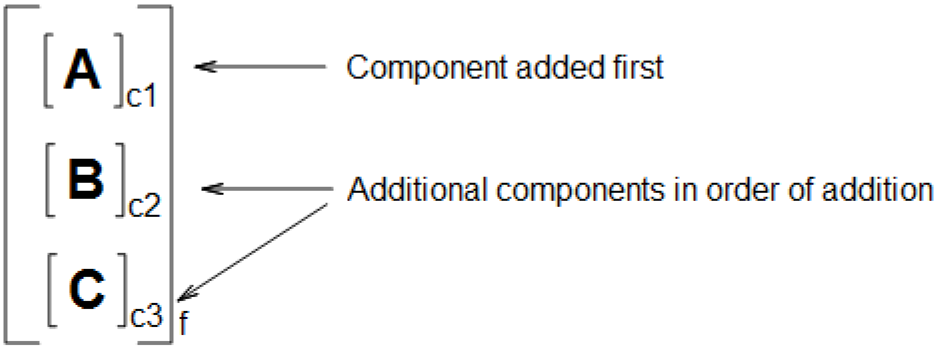
## Ordered and Unordered Mixtures

To represent mixtures, enclose each structure within *component* Sgroups (component brackets), and then enclose the components within mixture Sgroups (mixture brackets). Mixtures can be *unordered* or *ordered*:

- If the order of addition of components is not important, specify an unordered mixture (mix). Unordered mixtures contain components (c) that are not numbered:



- If the order of addition of components of the formulation is important, use an ordered mixture (f). Ordered mixtures contain components that are numbered (c1, c2, c3, and so forth):



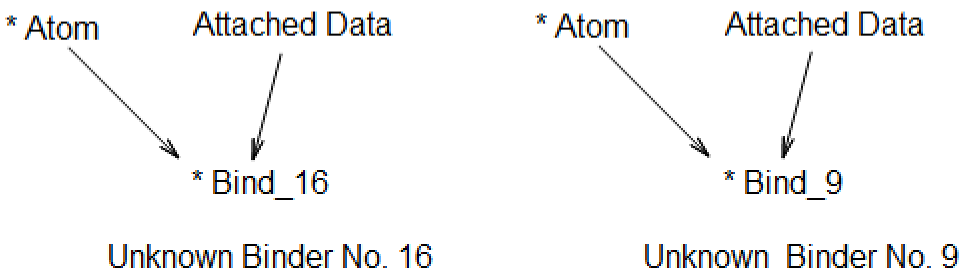
Ordered Mixture

For more information, see Examples of Mixtures on page 75.

**Note:** In earlier versions of BIOVIA software, ordered mixtures were called *formulations*.

## Using * atoms for Unspecified Structures in Mixtures

You can use *atoms with attached data to represent substances where the structure is unknown, or when you do not want to specify the structure. For example, the following *atoms represent two distinct binders of unspecified structure:



In each *atom representation, the attached data is associated with the Sgroup field `BINDER_DESCRIPTION`.

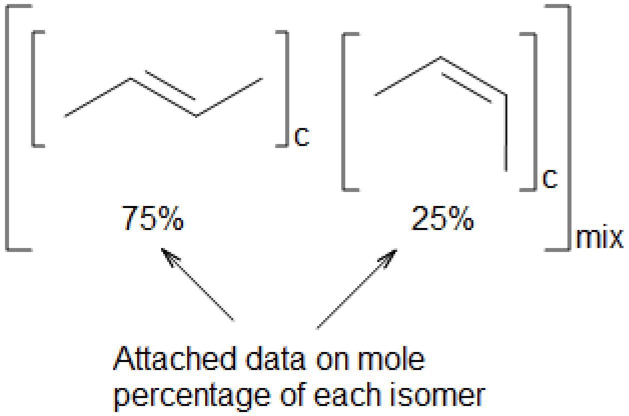
## Examples of Mixtures

This section contains examples of non-polymer mixtures only. For an example of a polymer mixture, see Guidelines for Defining Additional Attached Data on page 88.

### Unordered Mixtures

#### Reaction Product

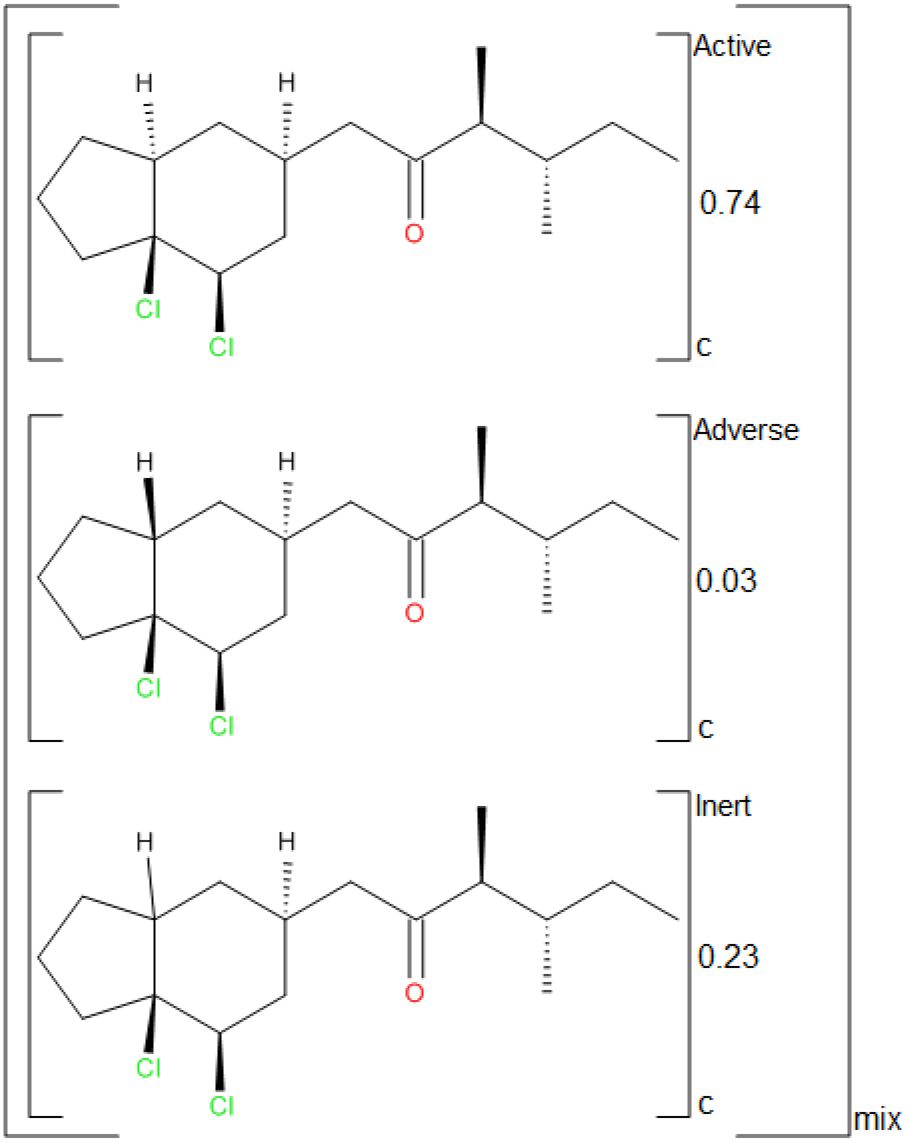
To draw a mixture of compounds such as a reaction product, draw a mixture with unordered components (unordered mixture). You use an unordered mixture because the order of addition is not important. The following example shows a reaction product that consists of a mixture of geometric stereoisomers:



Each component of the mixture is enclosed within component brackets (c). The numeric data that is attached to each set of component brackets specifies the mole percentage of that component of the mixture. Unordered mixture brackets (mix) enclose the component brackets

#### Mixture of Stereoisomers

The following example is a mixture of optical stereoisomers:



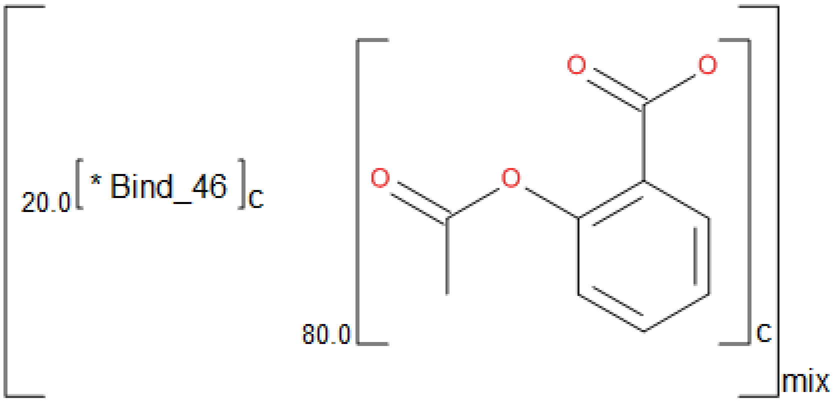
Each component bracket (c) within the mixture has two pieces of attached data associated with it:

Numeric data (0.74, 0.03, or 0.23) that represents the fraction of each component in the mixture. The numeric data is associated with the Sgroup field COMPONENT_FRACTION.

Non-numeric data (Active, Adverse, or Inert) that represents the activity of each component. This non-numeric data is associated with the Sgroup field ACTIVITY_TYPE.

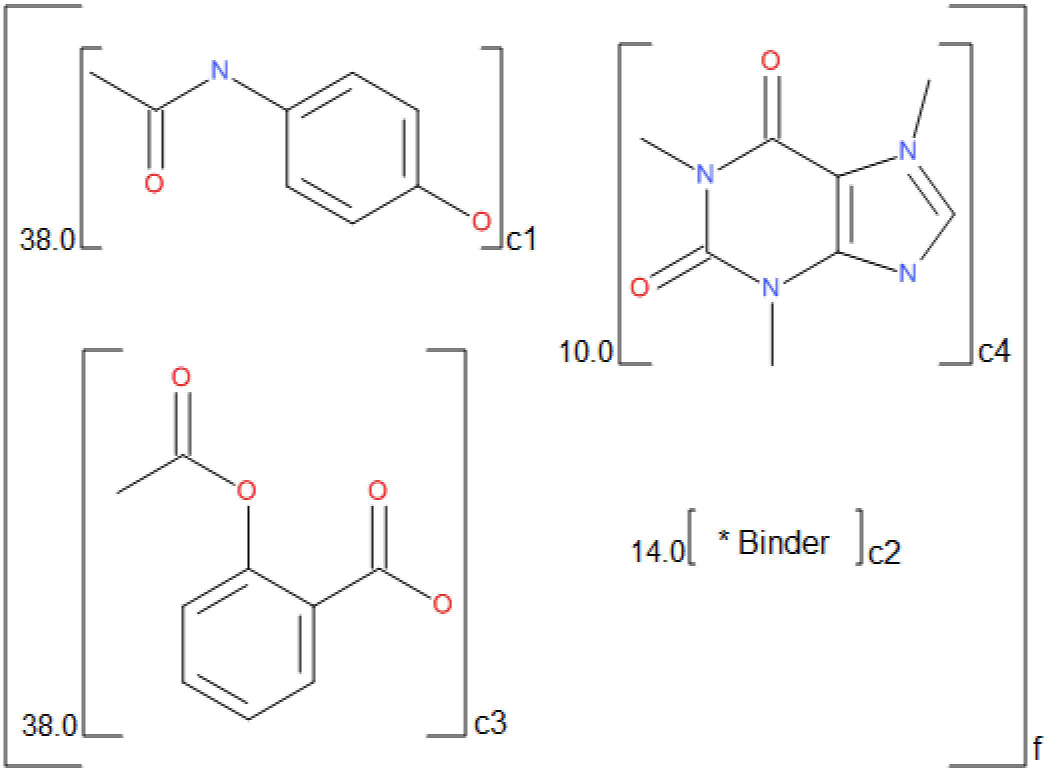
## Aspirin Tablet

The following mixture is the formulation of an aspirin tablet. The representation uses mixture brackets (mix), because the components (c) can be combined in any order. Each component bracket has attached data that specifies the weight percentage of each component (the associated Sgroup field is WEIGHT_PERCENT). One component (*atom with attached data Bind_46) is a binder (the associated Sgroup field is BINDER_DESCRIPTION):



## Ordered Mixture

In the following formulation, the acetaminophen (c1) must be added before the binder (c2) and other components (c3 and c4).



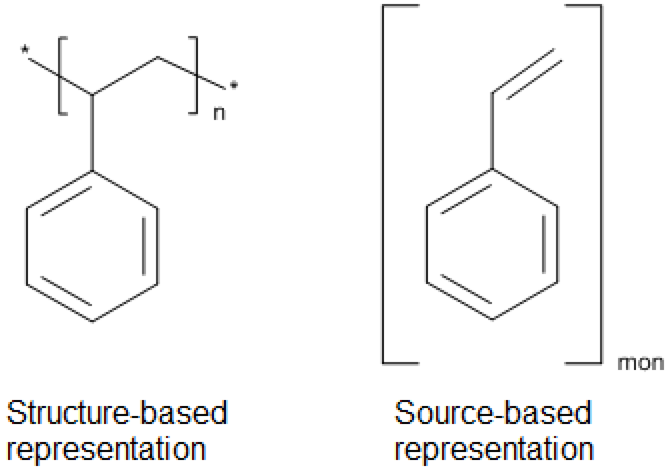
The attached data specifies the weight percentage of each component of the mixture (the associated Sgroup field is `WEIGHT_PERCENT`).

# Chapter 8:
# Polymer Representation

## Introduction to Polymer Representation

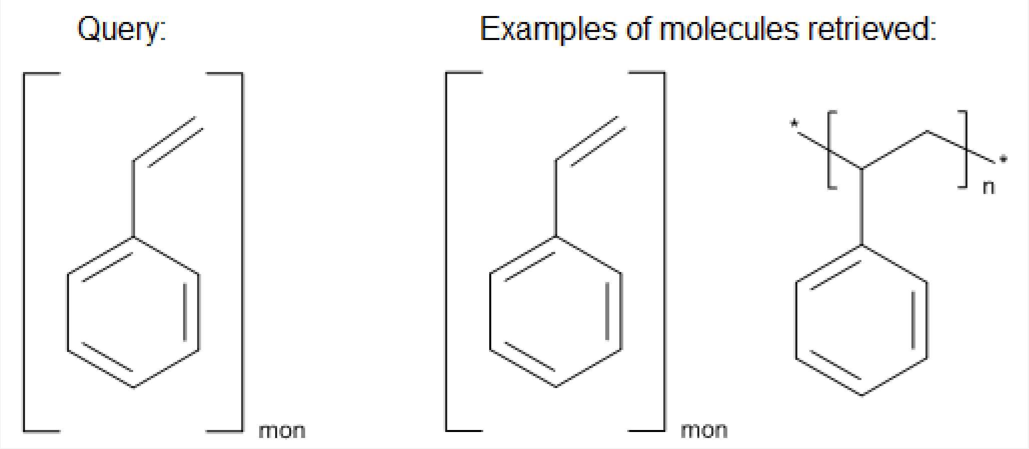
### Structure-based and Source-based Representation

Polymers can be drawn using either the structure-based or source-based representation. The following example shows both representations of polystyrene, a simple homopolymer:



For information on conventions for drawing structure-based representations of polymers and copolymers, see Examples of Structure-Based Representation on page 89. For information on conventions for drawing source-based representations of polymers and copolymers, see Examples of Source-based Representation on page 101.

**Note:** For certain types of polymerization reactions, the source-based and structure-based representations of a polymer can find each other. For example:



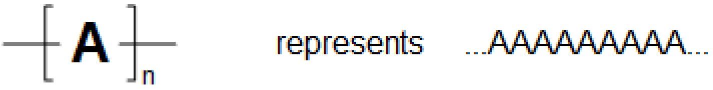
For more information on this type of polymer search, see Sourced-based and Structure-based Polymer Search on page 135.

## Polymer Bracket Types

A polymer structure consists of a collection of structural fragments that are enclosed by polymer brackets (polymer Sgroups) This section describes the types of polymer brackets that you can use to build polymer structures.

### Structural Repeating Unit (SRU) Brackets

Use Structural Repeating Unit (SRU) brackets to enclose the structural repeating unit (constitutional repeating unit) of the structure-based representation of a homopolymer:



SRU brackets imply that the structure within the brackets can repeat with itself. Consequently, you should enclose the structure "A" in SRU brackets solely if the sequence ...AAAAA... can occur.

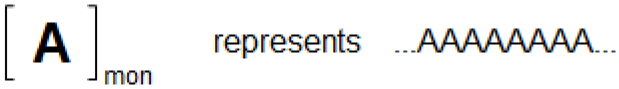
Each set of SRU brackets has two or more bonds, called crossing bonds, that overlay the brackets. The crossing bonds of an SRU show how the repeating units connect to each other within the polymer (connectivity).

> **Note:** For SRUs that have 2 or 4 crossing bonds, the repeat pattern of the SRU specifies additional information on connectivity. For more detailed information on repeat pattern, see Polymer Repeat Pattern on page 81.

For examples of polymer structures that use SRU brackets, see Examples of Structure-Based Representation on page 89.

### Monomer Brackets (mon)

Use Monomer (mon) brackets to draw the repeating unit(s) for the source-based representation of a homopolymer:



For examples of source-based representations of polymers, see Examples of Source-based Representation on page 101.

### Mer Brackets (mer)

For source-based representations of copolymers, use mer brackets (mer) to enclose a monomer that cannot repeat with itself. For example, if you know that a diol monomer for a polyester does not homopolymerize to form polyperoxides within the copolymer, enclose the diol in mer brackets. For examples of copolymers that use mer brackets, see Copolymers from Monomers that Do Not Homopolymerize on page 103.

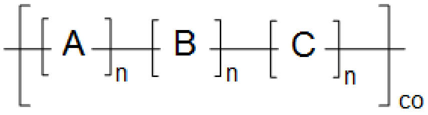
### Copolymer Brackets (co)

If a structure contains more than one set of SRU, monomer, and/or mer brackets, enclose the entire structure within copolymer brackets. Always use copolymer brackets when you require more than one set of polymer brackets to represent the structure. For example, polymers that are chemically modified after polymerization are represented as copolymers. For examples of chemically modified polymers, see the examples that begin at Alternating and Other Periodic Polymers on page 103.

If the order of the repeating units within the copolymer is known (for example, a regular block copolymer), draw the structure-based representation with connecting bonds between the SRUs. If the sequence repeats, draw the structure so that bonds from the SRUs cross the copolymer brackets:

poly(polyA-*block*-polyB-*block*-polyC

(segmented regular block)

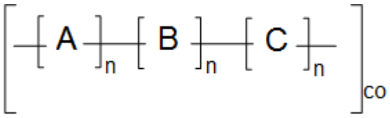


represents ...AAA...BBB...CCC...
AAA...BBB...CCC...

Only one sequence of blocks is possible; three-block sequence repeats with itself, giving a segmented block structure.

polyA-*block*-polyB-*block*-polyC

(regular block copolymer)

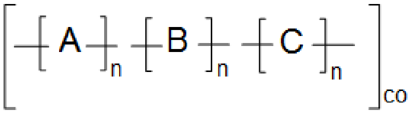


represents ...AAA...BBB...CCC...

Only one sequence of blocks is possible; three-block sequence repeats (regular block copolymer)does **not necessarily** repeat with itself.

polyA-*block*-polyB-*block*-polyC

(irregular block copolymer); also statistical, randon, unspecified, and all other copolymers.

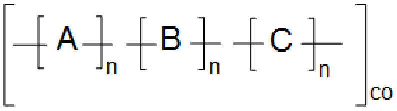


represents ...AAA...BBB...CCC...
**or**
...BBB...AAA...CCC...

for a total of six possible sequences of repeating units; repeating units are not necessarily grouped together.

If the sequence of the repeating units is unknown or unspecified (for example, an irregular block copolymer, or a random copolymer), draw the structure-based representation with no connecting bonds between the SRUs:

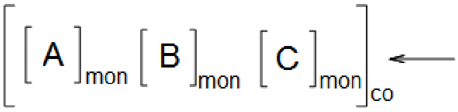
Statistical, random, unspecified and other copolymers



represents ...AAA...BBB...CCC...
**or**
...BBB...AAA...CCC...

for a total of six possible sequences of blocks.

Source-based representation



Connectivity and sequence unspecified

When you draw copolymer brackets with no crossing bonds, specify the either/unknown repeat pattern. The head-to-tail and head-to-head repeat patterns are meaningful solely when the copolymer brackets have crossing bonds. For more information on repeat pattern, see Polymer Repeat Pattern on page 81.

For examples of polymer structures that use copolymer brackets, see Examples of Structure-Based Representation on page 89 and Examples of Source-based Representation on page 101.

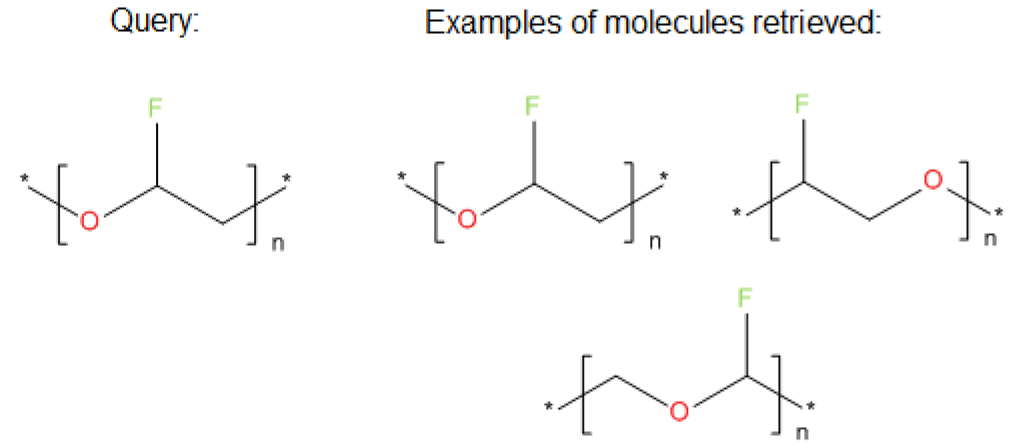
## Additional Polymer Brackets

Additional polymer and copolymer bracket types can be used to represent specific types of copolymers.

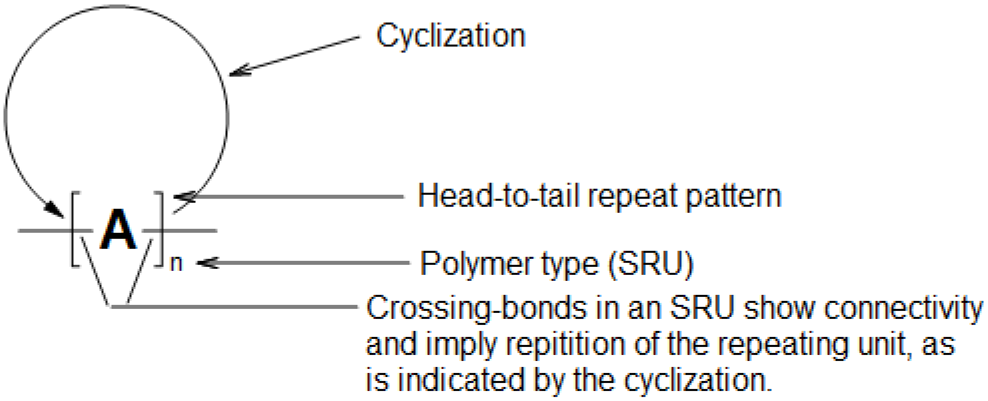
- copolymer bracket types: alternating (alt), block (blk), and random (ran)
- polymer bracket types: cross-link (xl), modified (mod), and graft (grf)

## Cyclization and Phase Shifting

For SRUs that contain a pair of brackets and that have the head-to-tail repeat pattern, you can define the structural repeating unit anywhere along the polymer backbone (that is, you can "phase-shift" the brackets along the polymer backbone). For example:



You can visualize the connectivity between SRUs as a cyclization of the crossing bonds:



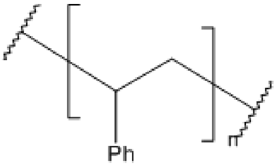
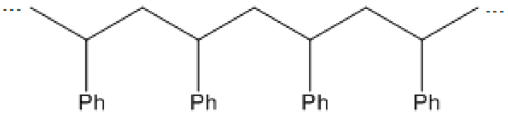
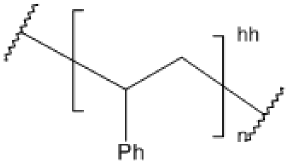
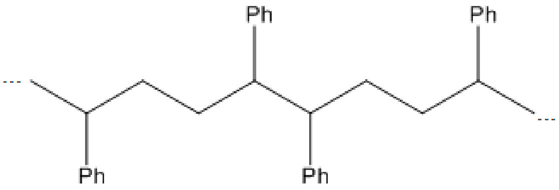
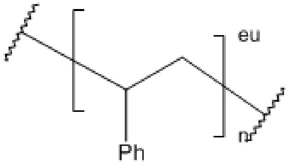
Crossing-bonds in an SRU show connectivity and imply repitition of the repeating unit, as is indicated by the cyclization.

## Polymer Repeat Pattern

For polymer brackets that contain crossing bonds (SRUs and some copolymers), the polymer repeat pattern shows how the repeating units are joined together. The repeat patterns that you can specify depend on the number of crossing bonds in the polymer.

### Polymers with Two Crossing Bonds

If the polymer bracket has one crossing bond on each bracket of the SRU or copolymer, you have three possible repeat patterns: head-to-tail (the default), head-to-head, and either/unknown. The following figure shows examples of the three repeat patterns in a homopolymer:

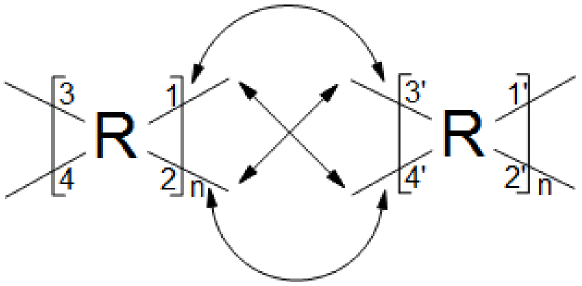
| Polymer Configuration | Polymer Drawing | Arrangement of Units within Polymer Structure |
|---|---|---|
| Head-to-tail |  |  |
| Head-to-head |  |  |
| Either/Unknown |  | Undefined.<br>Can represent a mixture of head-to-tail and head-to-head (**either**), or an **unknown** repeat pattern |

For an example of a homopolymer that uses either/unknown repeat pattern, see Irregular Homopolymers on page 90.

## Ladder-type Polymers

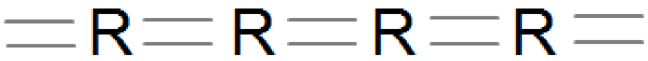
For paired brackets with two crossing bonds on each bracket (ladder-type polymers), you must also specify how the two crossing bonds on each bracket connect to the corresponding bonds of the adjacent SRUs:

- If the repeat pattern is head-to-tail, you must specify whether the SRU flips around the polymer backbone when it attaches to the adjacent SRU. Consequently, you have two options for connecting the corresponding bonds:

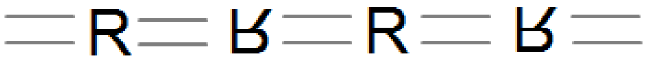


Possible combinations for head-to-tail repeat pattern:

1 & 3' = (2 & 4') head-to-tail with no flip
1 & 4' = (2 & 3') head-to-tail with flip

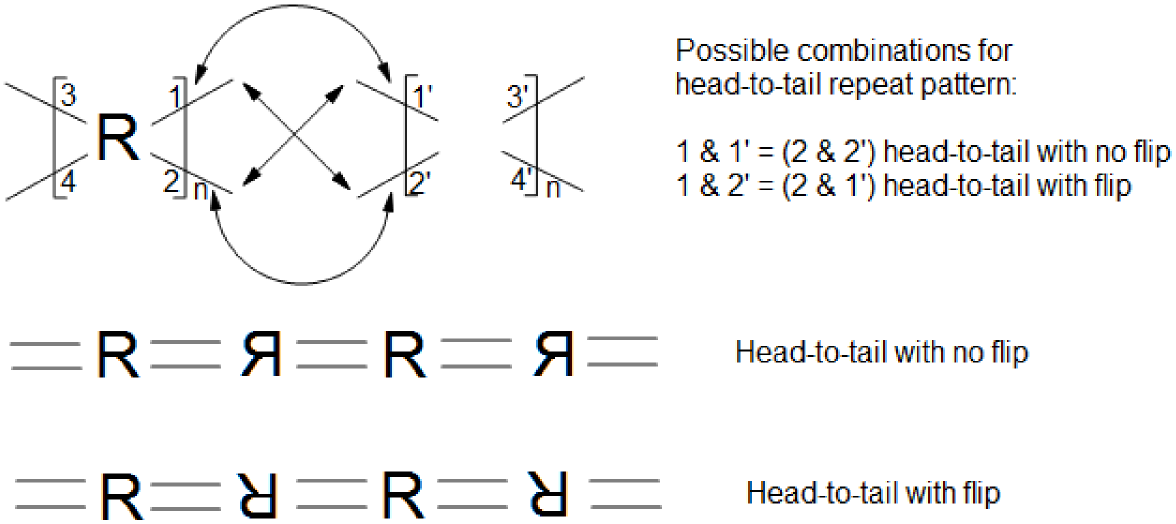


Head-to-tail with no flip



Head-to-tail with flip

- If the repeat pattern is head-to-head, the options for connecting the corresponding bonds are as follows:

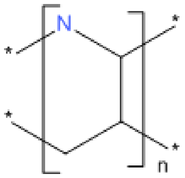
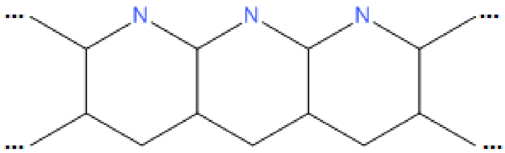
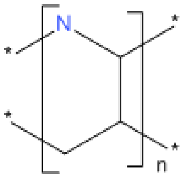
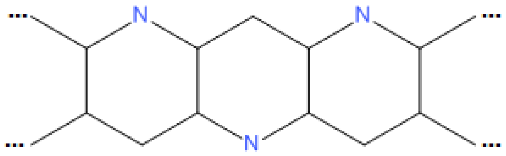
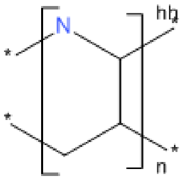
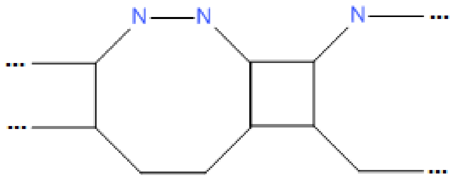
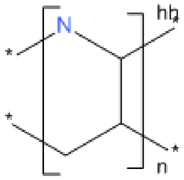
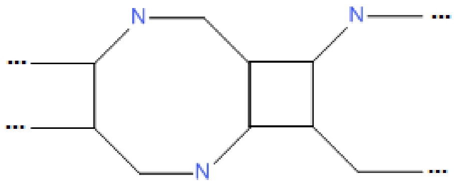
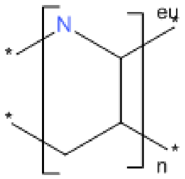


Possible combinations for
head-to-tail repeat pattern:

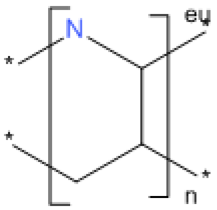
1 & 1' = (2 & 2') head-to-tail with no flip
1 & 2' = (2 & 1') head-to-tail with flip

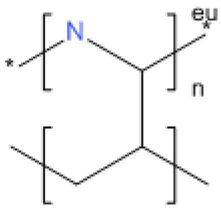
- If the repeat pattern is either/unknown, the connections between corresponding bonds might be unknown. Alternatively, the polymer might contain a mixture of units that are joined head-to-head, and other units that are joined head-to-tail (either).

The following figure shows examples of the five repeat patterns that are possible for ladder-type polymers:

| Polymer Repreat Pattern | Polymer Drawing | Arrangement of Units Within Polymer Structure |
|---|---|---|
| Head-to-tail no flip |  |  |
| Head-to-tail flipped |  |  |
| Head-to-head no flip |  |  |
| Head-to-head flipped |  |  |
| Either/Unknown |  | Undefined. Can represent a mixture of repeat patterns (**either**) or an **unknown** repeat pattern. |

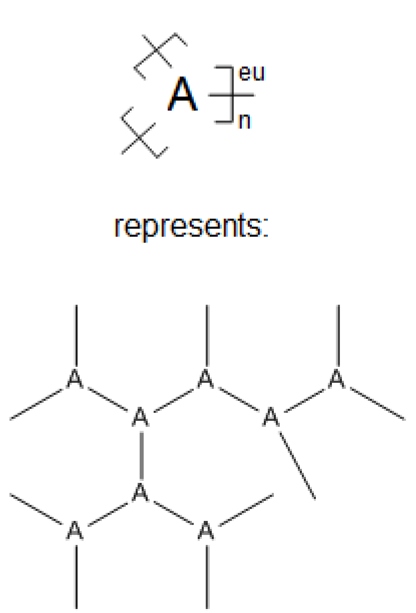
For the either/unknown repeat pattern, a structure that contains two brackets with four crossing bonds is equivalent to a structure that contains four brackets with one crossing bond on each bracket. For example, the following structures are equivalent:

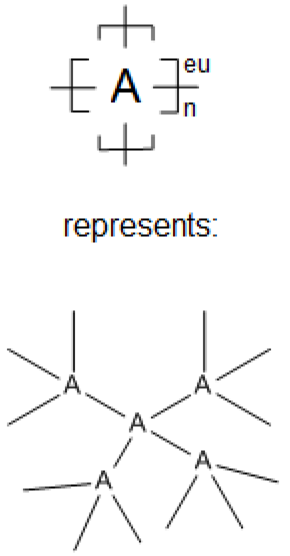




## Polymers with Three or More Brackets

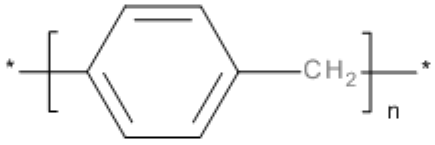
Polymers with three or more brackets always have the either/unknown (eu) repeat pattern. For example:



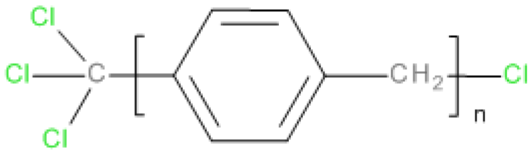


## Polymer End Groups

The end groups of polymers are generally unknown. Use *atoms (star atoms) to represent unknown or unspecified end groups. For example:



Polymer with unknown or unspecified end groups



Polymer with known end groups

**Note:** When you draw the brackets of a polymer structure, the drawing automatically changes terminal carbon atoms to *atoms.

## Why Are Representation Conventions Important?

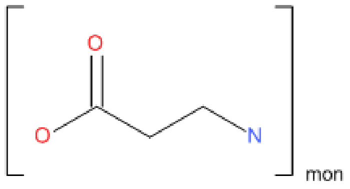
A consistent system of polymer representation for your corporate database ensures that your users can create graphical search queries that retrieve the structures that they want. Following BIOVIA' guidelines for polymer representation also ensures that your users are able to retrieve both source-based and structure-based representations of a polymer. To enable your users to create effective graphical search queries for polymers and mixtures, you need to follow BIOVIA conventions for graphical representation of polymers.

## Guidelines for Graphical Representation of Polymers

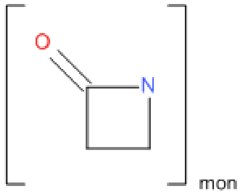
Use the information in this section to decide how you want to represent polymer structures that are stored in your corporate database.

## Structure-based and Source-based Representation

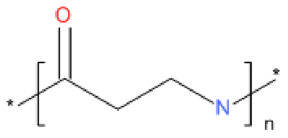
BIOVIA recommends that you use structure-based representation for structures that you register to your corporate database. The main advantage of structure-based representation is that the polymer structure is always the same, regardless of how the polymer was prepared. For example, Nylon 3 can be prepared from two different source materials: 2-azetidinone or beta-alanine. The following figure shows the corresponding source-based and structure-based representations of Nylon 3:



Source-Based Representation,
Beta-Alanine monomer



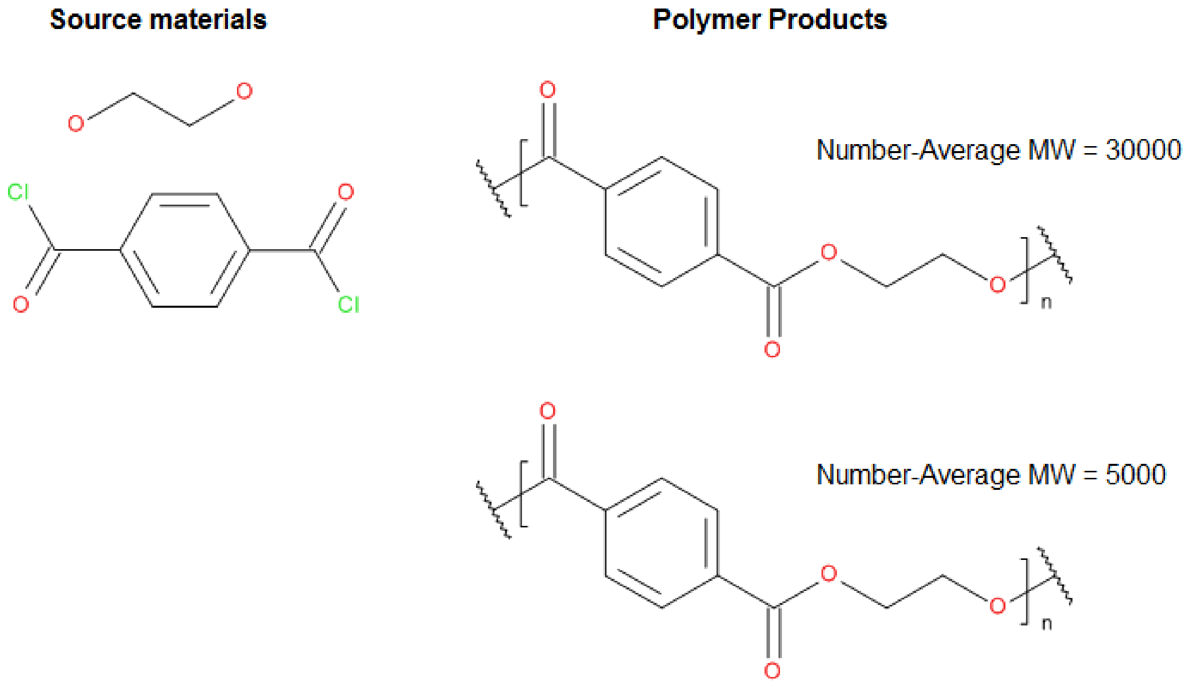
Source-Based Representation,
2-Azetidinone monomer



Structure-Based Representation,
Either monomer

Storing the structure-based representation in the database ensures that users can use either a source-based or structure-based query to retrieve the polymer structure.

In many cases, however, you might need to differentiate polymers that have the same graphical representation, but have different properties and/or were produced by different processes. For example, poly(ethylene glycol terephthalate) can be prepared from either the dimethyl ester or the diacid chloride of terephthalic acid. The two methods of preparation produce polymers with very different molecular-weight distributions, and hence different properties:



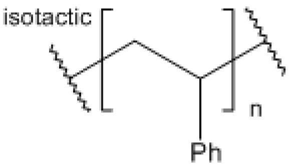
For polymers like these, BIOVIA recommends that you store the structure-based representation in your molecule database. You can store information on process and properties in relational tables, as attached data, or through a combination of relational tables and attached data.

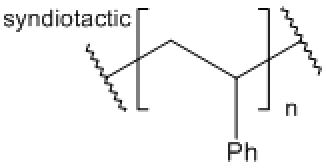
For examples of different types of structure-based polymers, see Examples of Structure-Based Representation on page 89.

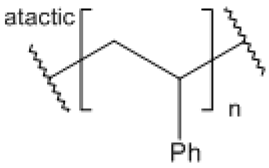
**Note:** In some circumstances, you might want to use solely source-based representation in your corporate database. For information on the reasons for choosing source-based representation, and for examples of source-based polymers, see Examples of Source-based Representation on page 101.

### Stereoregularity in Polymers

Avoid using Up and Down stereo bonds to show stereoregularity in polymer structures. A source-based graphical query cannot find a structure-based representation that contains Up and Down stereo bonds. Therefore, use attached data to show stereoregularity. In the following example, attached data shows tacticity in polystyrene:







## Using Attached Data in Polymer Representation

The BIOVIA administrator is responsible for creating the Sgroup fields in your corporate database and specifying the types of data that can be attached to structures. When you design your corporate database, you need to decide which data to store in database fields (including relational tables) and which data to store as attached data (that is, as structure annotations that are associated with Sgroup fields). This section explains how attached data is used in polymer representation.

For specific information on creating Sgroup fields for attached data in your database, see Sgroup Fields on page 154.

### Required Attached Data for Polymer Representation

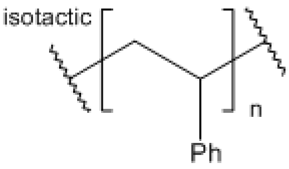
The following types attached data are required for BIOVIA' recommended structure conventions for polymer representation.

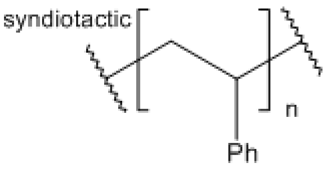
#### Polymer or Copolymer Type

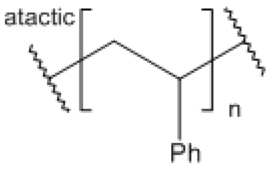
Use attached data that is associated with the Sgroup field POLYMER_TYPE to specify the type of polymer or copolymer in structure-based representation of polymers. For more examples of polymer type, see Examples of Structure-Based Representation on page 89.

#### Stereoregularity

Use attached data to show stereoregularity, because a source-based query cannot find a structure-based polymer that contains Up and Down stereo bonds. The following example shows tacticity in polystyrene:



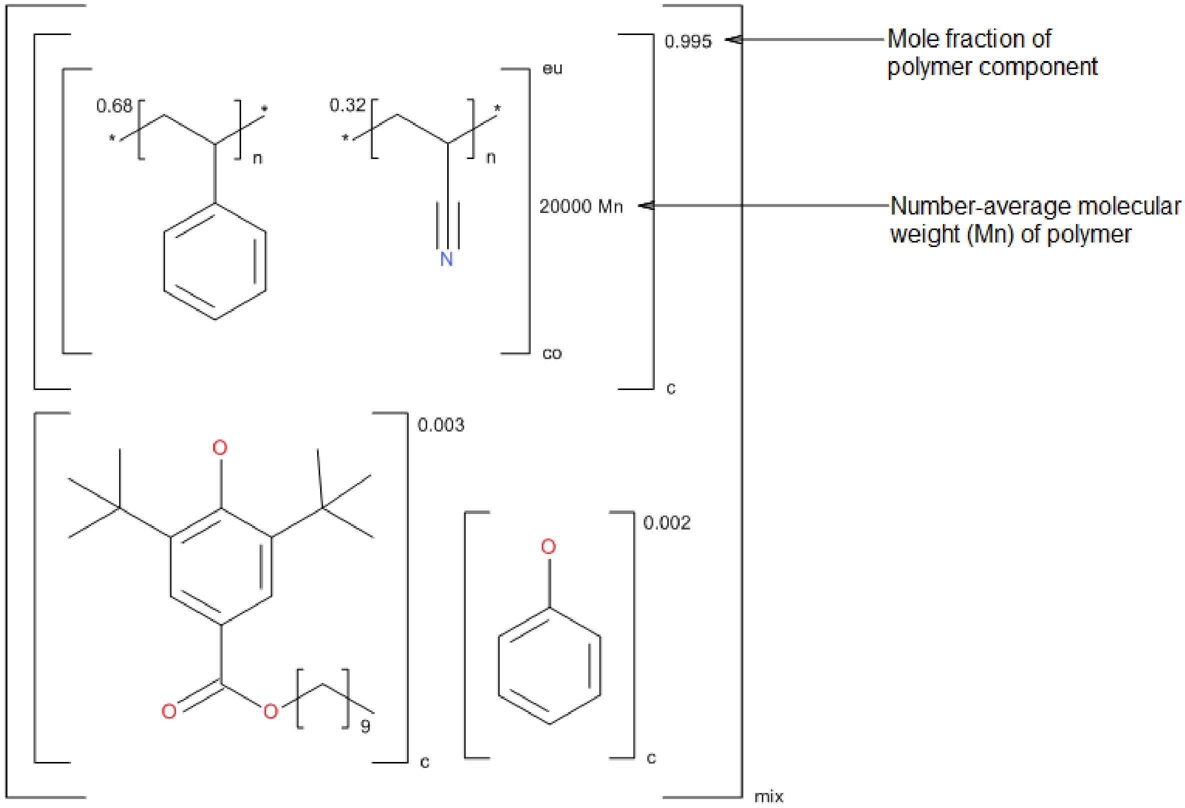




## Guidelines for Defining Additional Attached Data

In addition to the required Sgroup fields, you might want to define additional types of attached data (with associated Sgroup fields) to identify polymers.

For example, the following structure of a mixture contains attached data on the number-average molecular weight (Mn) of the copolymer (20,000 Mn), and on the mass fraction of each structural repeating unit within the copolymer (0.68 and 0.32), and on the mole fraction of each component of the mixture (0.995, 0.003, and 0.002):


0.68
n
0.32
n
eu
co
0.995
20000 Mn
c
0.003
O
O
O
9
c
0.002
O
c
mix
N
Mole fraction of polymer component
Number-average molecular weight (Mn) of polymer

Use the following guidelines to decide whether to define additional types of attached data for polymer representation:

1. If the information applies to a complete structure record, store the information either in a field in the molecule database, or in a relational-database table. Examples of this type of data include:
   - The corporate ID number of a polymer or other compound
   - The commercial name of a polymer or polymer formulation
   - Physical properties that are associated with a polymer or polymer formulation, such as elasticity
2. Use attached data for any other information (in addition to polymer type and stereoregularity) that applies to a portion of a structure. For example:
   - Use attached data for the weight fraction (or weight percentage) of individual monomers that comprise a copolymer. In the previous figure, the numeric data attached to the SRU brackets is associated with the Sgroup field POLYMER_MASS_FRACTION.

- Use attached data for the mole fraction of individual components of a mixture, such as a composite of several polymers. In the figure, the numbers that are attached to the component brackets are associated with the Sgroup field `MOLE_FRACTION`.
- Use attached data to differentiate polymers with the same structural repeating unit (SRU) but different properties. In the figure, the number-average molecular weight (20000 Mn) is associated with the Sgroup field `MOLWT_NUMBER_AVERAGE`.

For more examples of attached data, see the sections Examples of Structure-Based Representation on page 89 and Examples of Source-based Representation on page 101.

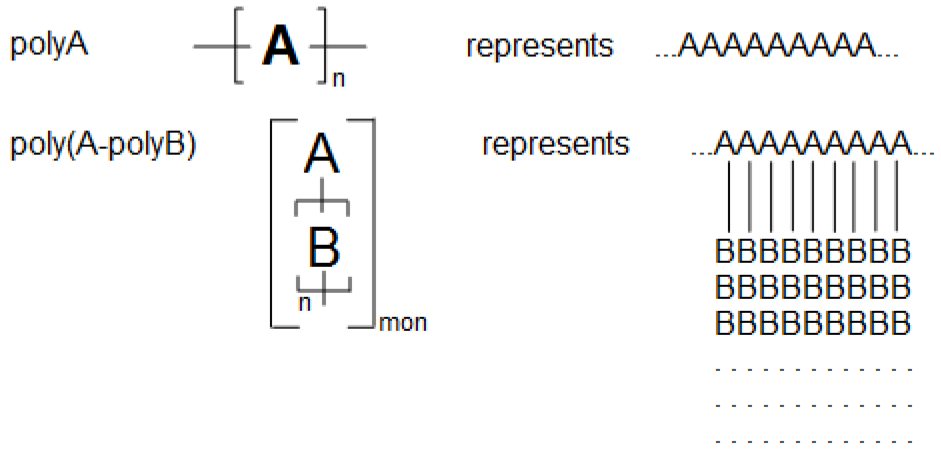
## Examples of Structure-Based Representation

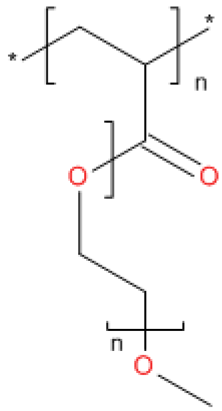
The examples in this section show how to draw structure-based representations of different kinds of polymers and copolymers. The conventions in this section are based on standards for polymer representation that have been established by the International Union of Pure and Applied Chemistry (IUPAC). For more information on IUPAC standards, see Bareiss, R. E.; Kahovec, J.; and Cratochvil, P.; "Graphical Representations (Chemical Formulae) of Macromolecules (IUPAC Recommendations 1994)" *Pure Appl. Chem.* 1994, **66**(12), 2469-2482. This report is available as a PDF online at:

http://www.iupac.org/publications/pac/1994/pdf/6612x2469.pdf

### Regular Homopolymers

#### Simple Homopolymers





homopolymer of a macromer

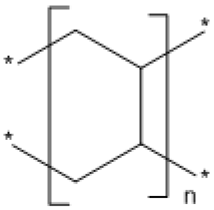
### Stereoregularity

Use attached data that is associated with the Sgroup field POLYMER_STEREO to specify to specify information on stereoregularity in polymers. For examples, see Stereoregularity on page 87.

### Ladder-type Polymers

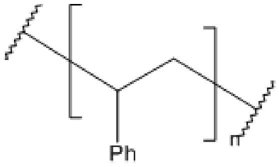
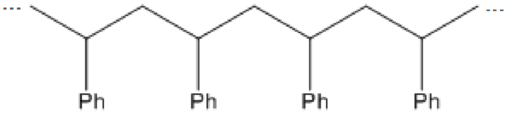
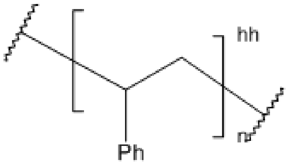
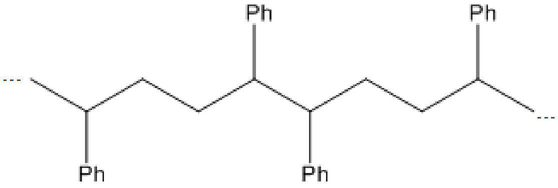
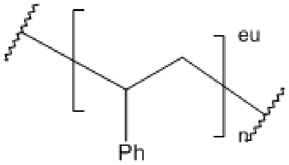
To draw ladder-type polymers, use paired brackets with two bonds over each bracket. For example:

poly(but-1-ene-1,4:3,2-tetrayl)



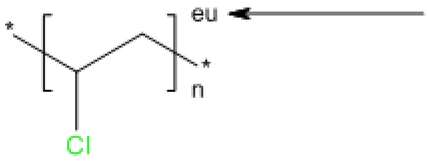
## Irregular Homopolymers

The repeat pattern of a polymer specifies whether the SRUs are connected head-to-head or head-to tail. The either/unknown repeat pattern specifies a polymer in which the repeat pattern is unknown, or in which the repeat pattern can be either head-to-head or head-to tail. The following example shows possible repeat patterns in polystyrene:

| **Polymer Configuration** | **Polymer Drawing** | **Arrangement of Units within Polymer Structure** |
|---|---|---|
| Head-to-tail |  |  |
| Head-to-head |  |  |
| Either/Unknown |  | Undefined.<br>Can represent a mixture of head-to-tail and head-to-head (**either**), or an **unknown** repeat pattern |

In the following example, the structural repeating units of poly(chloroethylene) can be joined head-to-head (1-chloroethylene) or head-to-tail (2-chloroethylene). The either/unknown repeat pattern shows this irregularity:

poly(1-chloroethylene/ 2-chloroethylene)

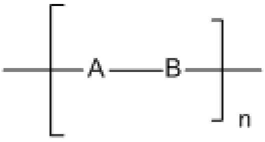


Polymer repeat pattern **either/unknown**

## Alternating and Periodic Polymers

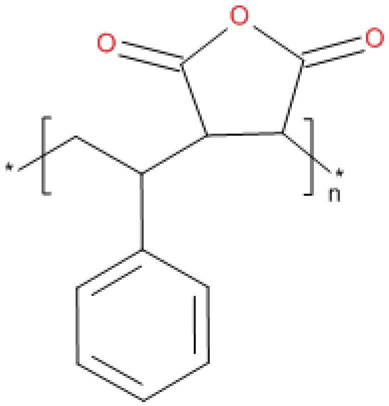
Represent alternating and other periodic polymers as homopolymers. That is, enclose the structural unit within a single SRU. For example:

poly(A-*alt*-B)

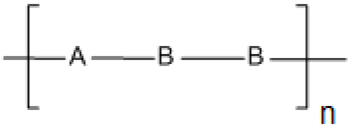


represents ...ABABABAB...

Poly[(2,5-dioxotetrahydrofuran-3,4-diyl)(1-phenylethylene)]



poly(a-per-B-per-B)



represents ...ABBABBABBABB...

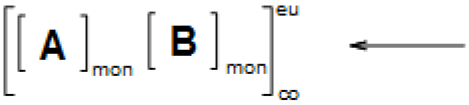
## Statistical, Random, and Unspecified Copolymers

In statistical, random, and unspecified copolymers, the sequence of repeating units is irregular or unspecified. Therefore, you need to draw the copolymer with no connecting bonds between the SRUs.

Always use the either/unknown repeat pattern for copolymer brackets that do not have bonds that overlay the copolymer brackets. These bonds are called crossing bonds.

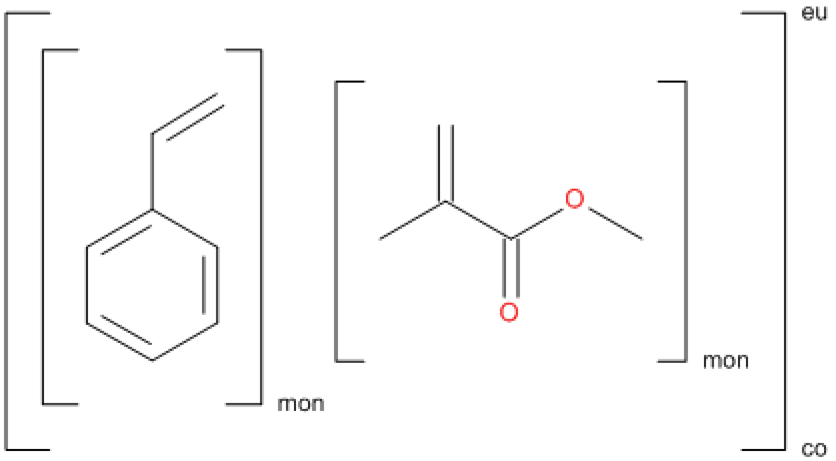
### Unspecified Copolymers

poly(A-*co*-B)



Sequence of repeating units is unspecified

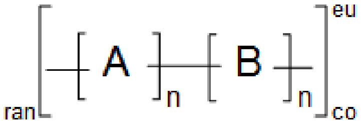
Poly[styrene-*co*-(methyl methacrylate)]



## Random Copolymers

Represent random copolymers as a combination of SRUs without connecting bonds. Attach data (ran) to the brackets to show the copolymer type. For example:

poly(A-*ran*-B)

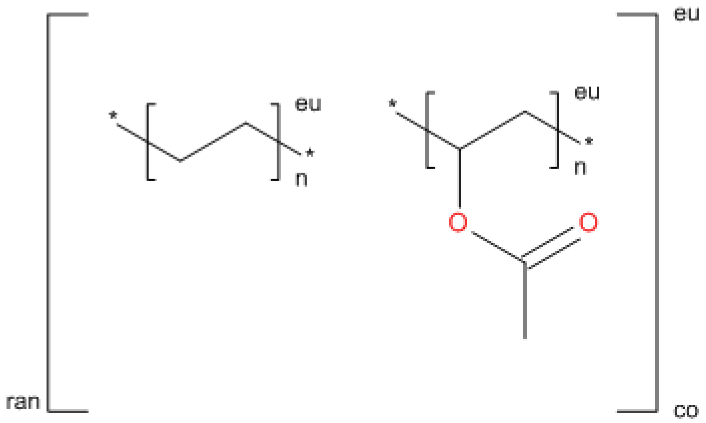


represents ...ABAABBBABAB...

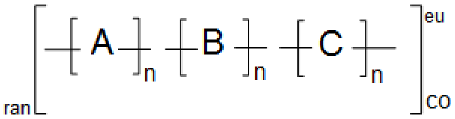
ran = attached data on polymer type

poly[ethylene-*ran*-(vinyl acetate)]

ran = attached data on polymer type



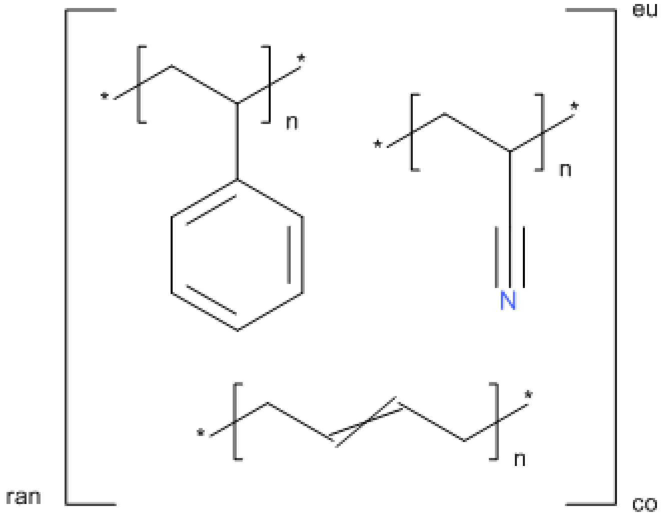
ploy(A-*ran*-B-*ran*-C)



represents ...AABBCBAACABC...

ran = attached data on polymer type

poly(styrene-*ran*-acrylonitrile-*ran*-butadiene)

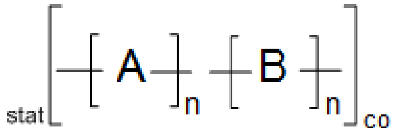


ran = attached data on polymer type

## Statistical Copolymers

Represent "statistical" copolymers as random copolymers. Use attached data (stat) to specify the polymer type, and/or to specify probabilities of repeating units. For example:

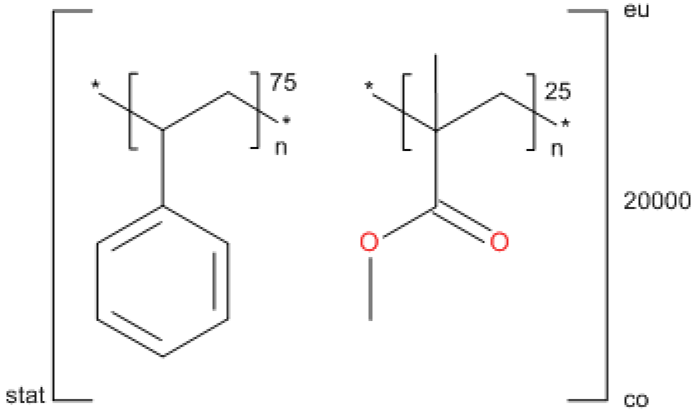
poly(A-*stat*-B)



represents ...ABBABAABABBB...

stat = attached data on polymer type

poly[styrene-*stat*-(methyl methacrylate)]



**Attached Data**

| | |
|---|---|
| stat | Polymer Type |
| 75<br>25 | } Mass percentage of each repeating unit |
| 20000 | Number-average molecular weight |

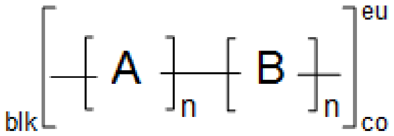
## Regular Block Copolymers

In a regular block copolymer, the sequence of repeating units is known. Therefore, you need to draw the copolymer in sequence, with connecting bonds between the SRUs.

If the copolymer brackets have crossing bonds, specify the head-to-tail or head-to-head repeat pattern as necessary. Otherwise, specify either/unknown.

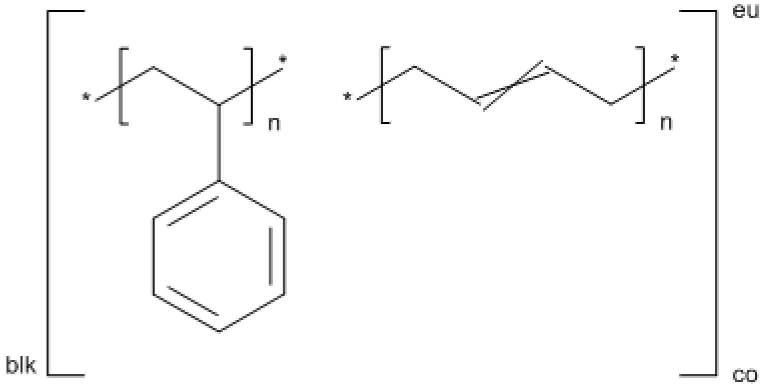
### Ordered Diblock

polyA-block-polyB



represents ...AAABBB...

polystyrene-*block*-polybutadiene



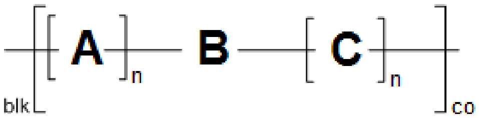
blk = attached data on polymer type

= Double Either bond (can be cis or trans)

## Block Copolymer with Junction Unit

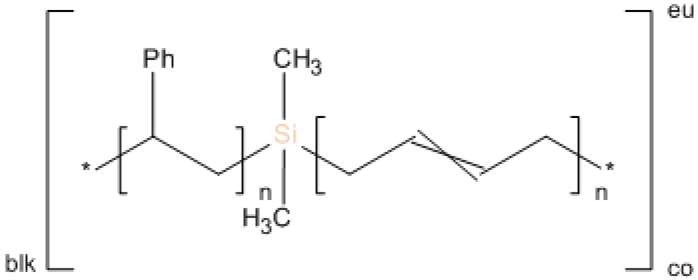
A *junction unit* is a non-repeating structural unit that joins two sequences within a polymer. Do not enclose a junction unit in SRU brackets, because SRU brackets imply that the unit can repeat with itself. For example:

polyA-*block*-B-*block*-polyC



represents ...AAA...B...CCC...
AAA...B...CCC...

polystyrene-*block*-dimethylsilanedyl-*block*-1,4-polybutadiene

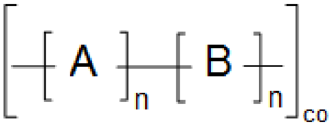


blk = attached data on polymer type

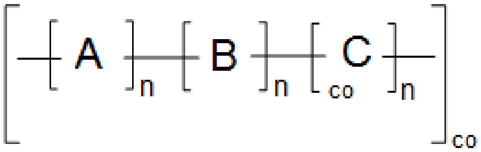
= Double Either bond

## Segmented Block Copolymers

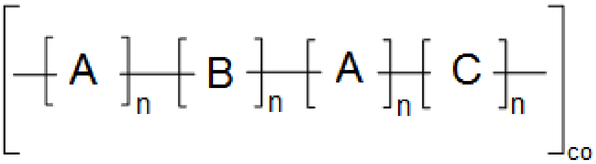
If bonds cross the copolymer brackets, then the entire sequence of blocks repeats, giving a segmented block structure. For example:



represents ...AAA...BBB...AAA..
.BBB...AAA...BBB...

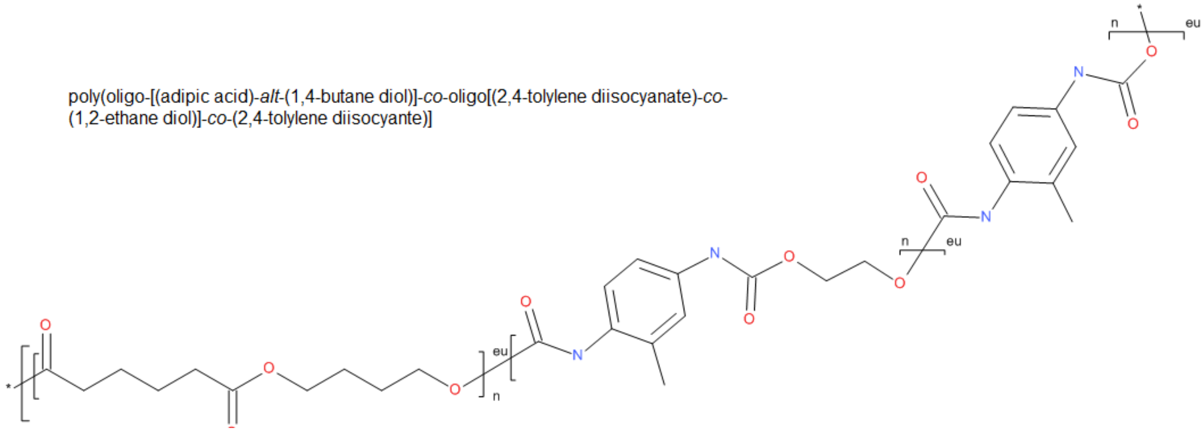


represents ...AAA...BBB...CCC..
.AAA...BBB...CCC...

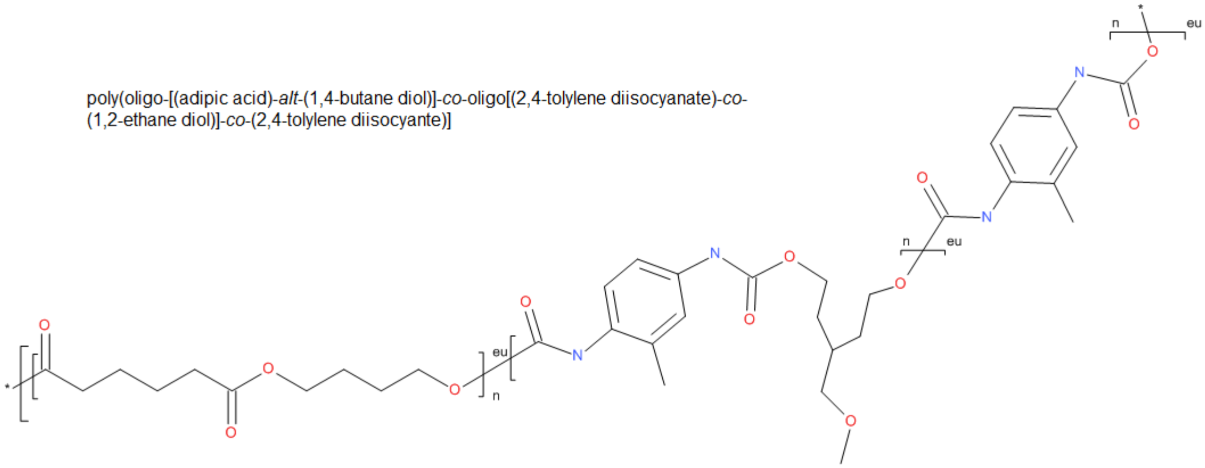


represents ...AAA...BBB...AAA..
.CCC...AAA...BBB...
.BBB...AAA...CCC...

poly(oligo-[(adipic acid)-*alt*-(1,4-butane diol)]-*co*-oligo[(2,4-tolylene diisocyanate)-*co*-(1,2-ethane diol)]-*co*-(2,4-tolylene diisocyante)]

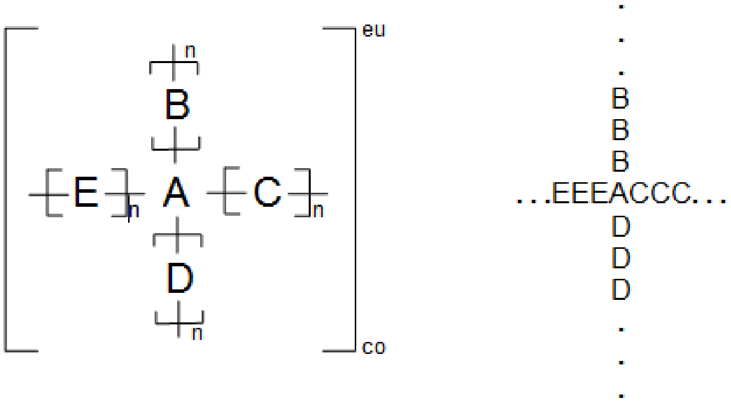


poly(oligo-[(adipic acid)-*alt*-(1,4-butane diol)]-*co*-oligo[(2,4-tolylene diisocyanate)-*co*-(1,2-ethane diol)]-*co*-(2,4-tolylene diisocyante)]

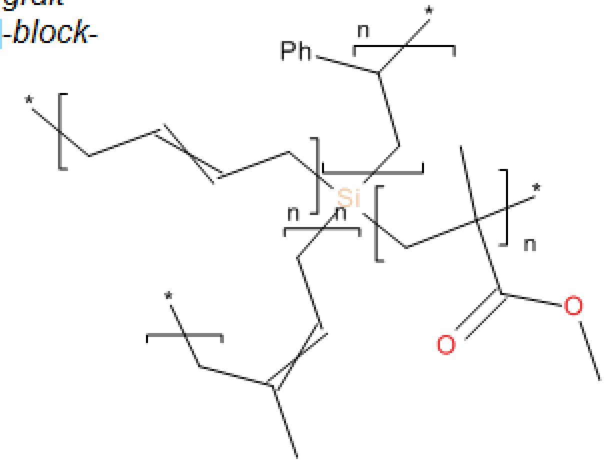


## Star Block Copolymers

polyE-*block*-[A-*graft*-(polyB;polyB)]
-*block*-polyC



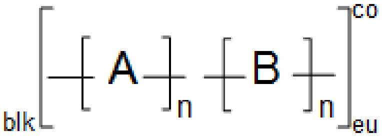
polystyrene-*block*-[silanetetrayl-*graft*-(polybutadiene;poly isoperene)]-*block*-poly(methyl methacrylate)



## Irregular Block Copolymers

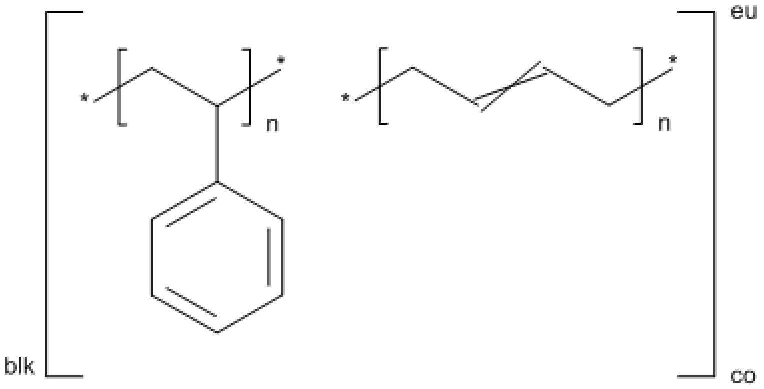
In a block copolymer that consists of irregular blocks, the sequence of repeating units is unknown. Therefore, draw the copolymer with no connecting bonds between the SRUs, and specify the either/unknown repeat pattern on the copolymer brackets:

polyA-block-polyB



represents ...AAA...BBB...
**or**
...BBB...AAA....

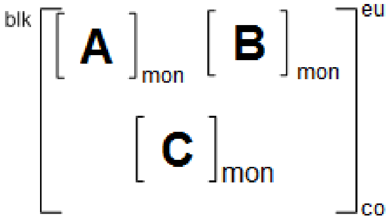
polystyrene-*block*-polybutadiene



blk = attached data on polymer type

= Double Either bond (can be cis or trans)

polyA-*block*-polyB-*block*-polyC



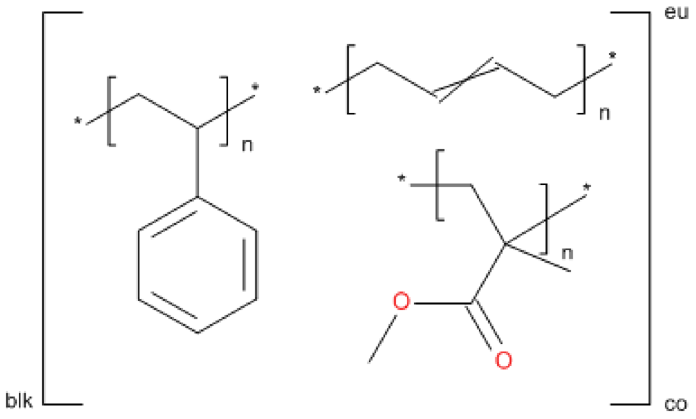
represents AAA...BBB...CCC
**or**
BBB...AAA...CCC
for a total of six possible block sequences

blk = attached data on polymer type

polystyrene-block-polybutadiene-block-poly(methyl methacrylate)

blk = attached data on polymer type

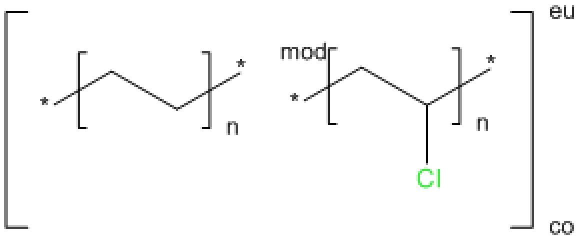
= Double Either bond



## Chemically Modified Polymers

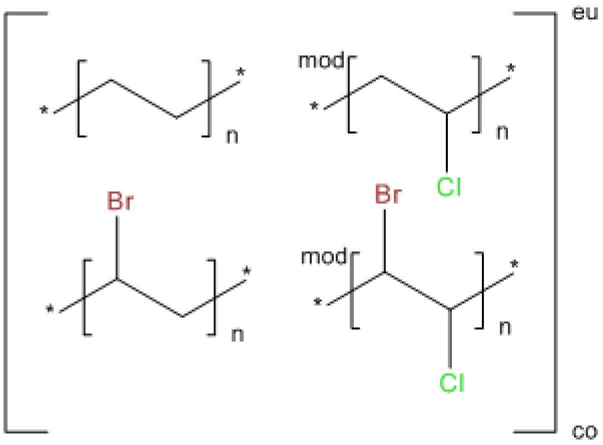
Represent a chemically modified polymer as a copolymer that contains the original SRU and the chemically modified SRU.

Chlorinated polyethylene



mod = optional attached data on polymer type

brominated (chlorinated polyethylene)



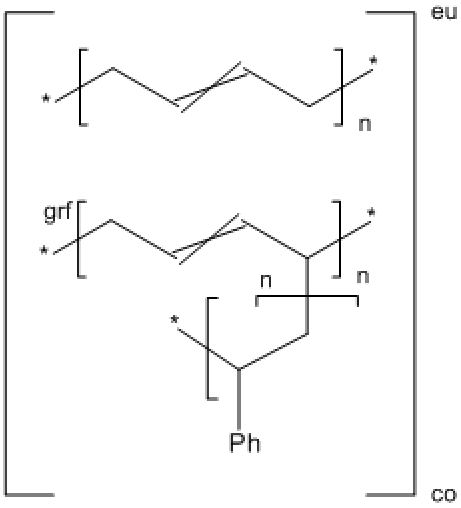
mod = optional attached data on polymer type

## Graft Polymers and Copolymers

Represent a graft polymer as a copolymer that contains the original SRU and the grafted SRU.

### Single Graft at a Known Site

polybutadiene-*graft*-polystyrene

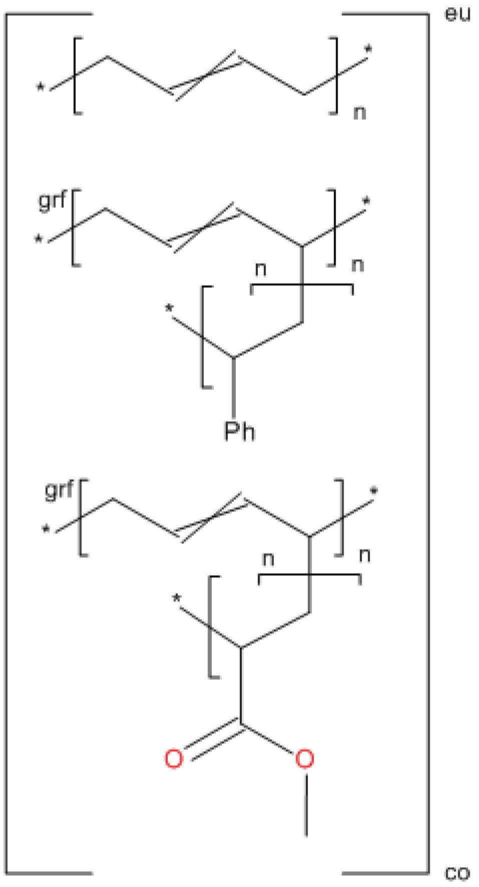


grf = attached data on polymer type

### Mixed Graft at a Known Site

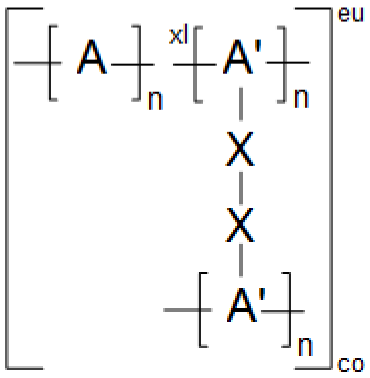
polybutadiene-*graft*-[polystyrene; poly-(methyl methacrylate)]

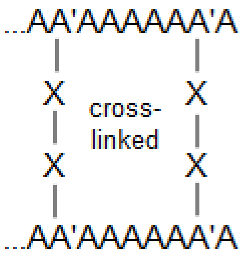
grf = attached data on polymer type



### Cross-linked Polymers

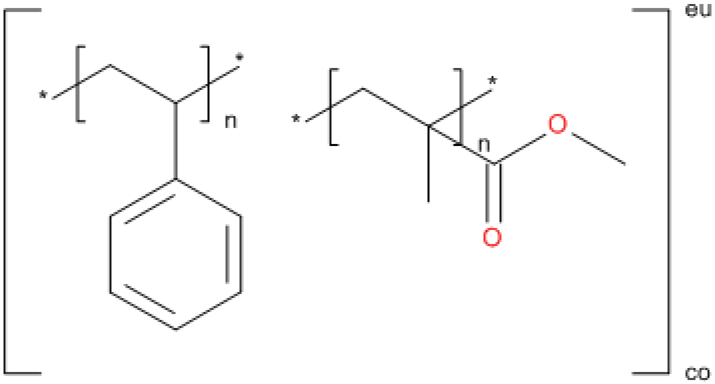
Represent cross-linked polymers as copolymers. That is, represent the polymer as a copolymer that contains both the original and cross-linked SRU.



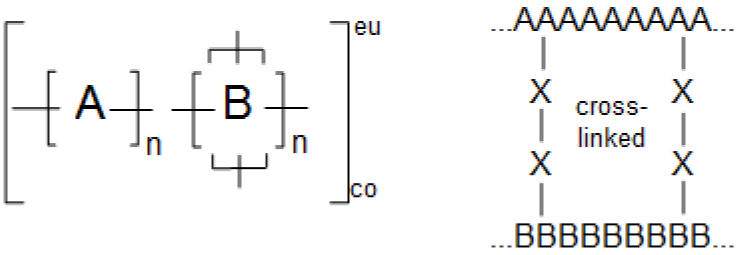


xl = optional attached data on polymer type

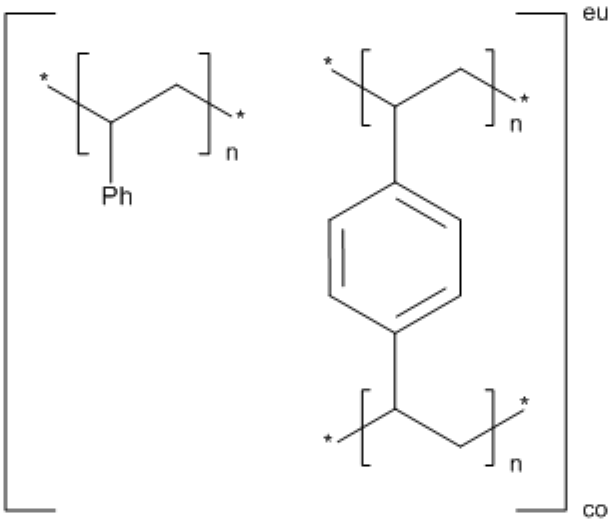
poly[styrene-*co*-(methyl methacrylate)]



polyA-*co*-polyB



poly[styrene-*ran*-(divinyl benzene)]



## Examples of Source-based Representation

You might decide to use solely source-based representation for your corporate databases. You might prefer source-based representation for the following reasons:

- Your company creates many polymers in which the structures of the repeating units are uncertain, but whose starting materials are known.
- Your users are more interested the monomer from which a polymer is prepared than in the structure of the resulting polymer. That is, your users commonly ask questions such as, "What polymers in the database were made from this monomer?"

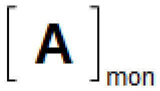
Use the guidelines in this section to learn how to create source-based representations of polymers and copolymers.

## Guidelines for Source-based Representation

The source-based representation of a polymer is based on the starting materials of the polymer. Use the following guidelines when you draw source-based representations of polymers:

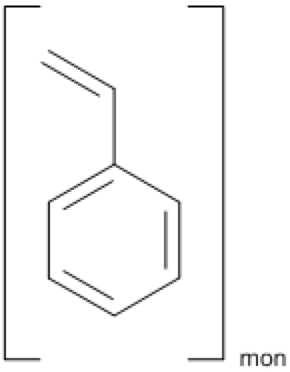
- Use monomer (mon) brackets to enclose all the source materials from which a single structural repeating unit of the polymer is made. In drawing copolymers, enclose a structure within monomer brackets solely if you know that the monomer can homopolymerize.
- If you know that a monomer does not homopolymerize, enclose the monomer within mer brackets. For example, the diol and diacid monomers of a polyester generally do not react to form polyperoxides or polyethers and polyanhydrides. For more information, see Copolymers from Monomers that Do Not Homopolymerize on page 103.
- If the structure contains more than one set of monomer and/or mer brackets, enclose the structure within copolymer brackets (co).
- Specify the either/unknown repeat pattern on copolymer brackets. Copolymer brackets in source-based polymers do not have bonds that overlay the brackets (crossing bonds), and therefore should not have the ht or hh repeat pattern.

### Homopolymers

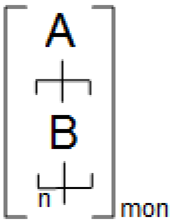


represents ...AAAAAAA...

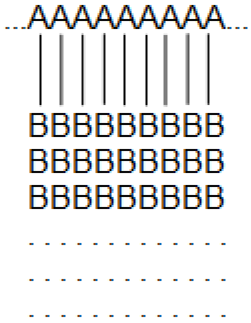
polystyrene



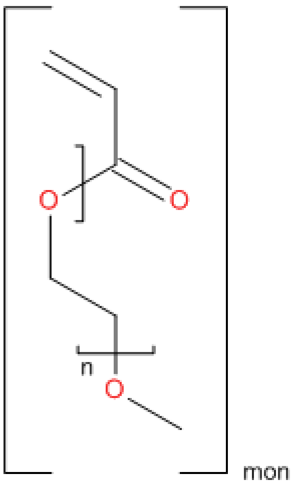
poly(A-polyB)



represents



homopolymer of a macromer

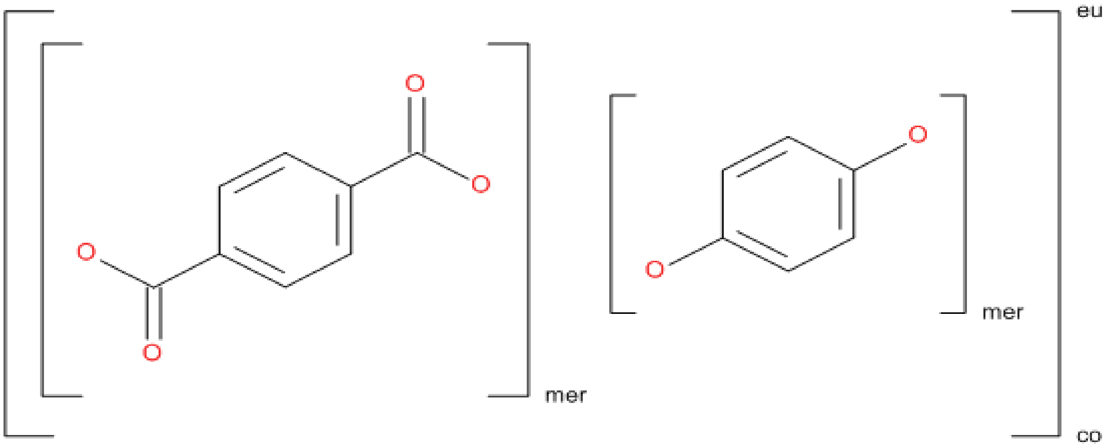


## Alternating and Other Periodic Polymers

For alternating and other periodic polymers, you generally know enough about the structure to draw the structure-based representation. Therefore, BIOVIA strongly recommends that you do not register source-based alternating and periodic copolymers.

## Copolymers from Monomers that Do Not Homopolymerize

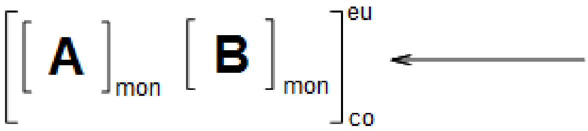
Use mer brackets to enclose monomers that you know do not homopolymerize. For example, when you combine a diol monomer and a diacid monomer to create a polyester, you might know that the two monomers do not homopolymerize to form polyperoxides, polyethers, or polyanhydrides. To show this, you enclose the diol and diacid monomers in mer brackets:



## Statistical, Random, and Unspecified Copolymers

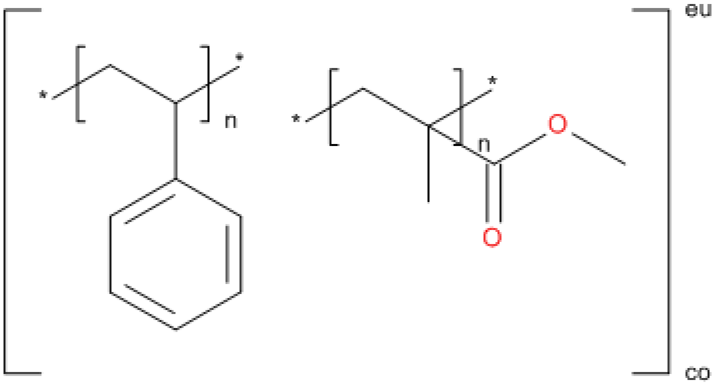
### Unspecified Copolymers

poly(A-*co*-B)



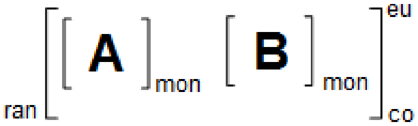
Sequence of repeating units is unspecified

poly[styrene-*co*-(methyl methacrylate)]



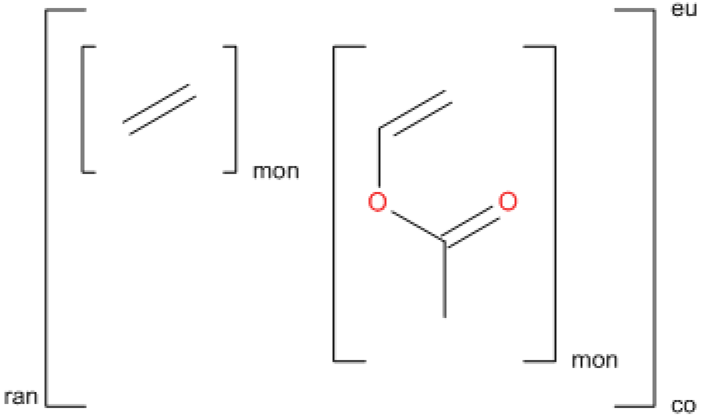
## Random Copolymers

poly(A-*ran*-B)



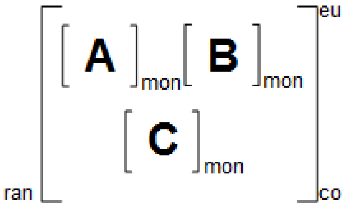
represents ...ABBABAABABBB...

poly[ethylene-*ran*-(vinyl acetate)]



ran = attached data on polymer type

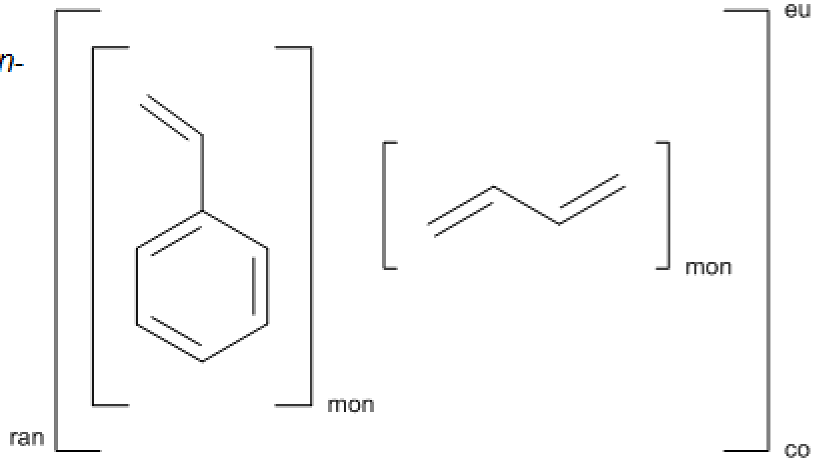
poly(A-*ran*-B-*ran*-C)



represents ...ACBBACBAACCBABBCB...

ran = attached data on polymer type

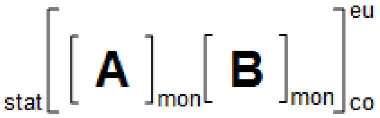
poly(styrene-*ran*-butadiene)



ran = attached data on polymer type

## Statistical Copolymers

poly(A-*stat*-B)

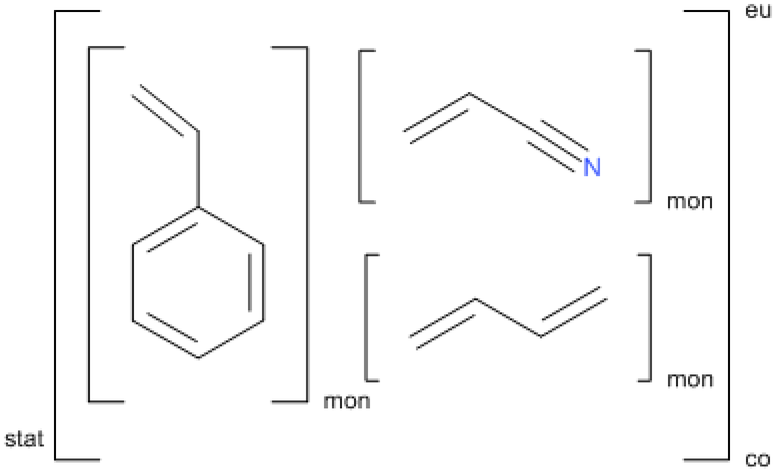


represents ...ABBABAABABBB...

stat = attached data on polymer type

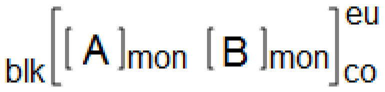
Poly[styrene-*stat*-acrylonitrile-*stat*-butadiene]

stat = attached data on polymer type



## Block Copolymer

polyA-*block*-polyB

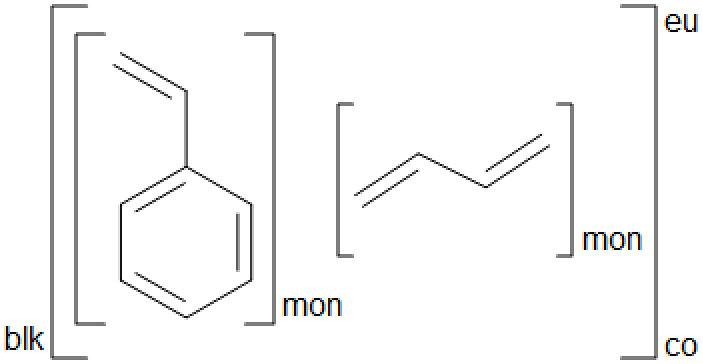


represents

AAA...BBB
or
BBB...AAA

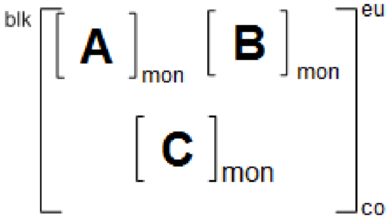
blk = attached data on polynmer type

polystyrene-*block*-polybutadiene



blk = attached data on polynmer type

polyA-*block*-polyB-*block*-polyC

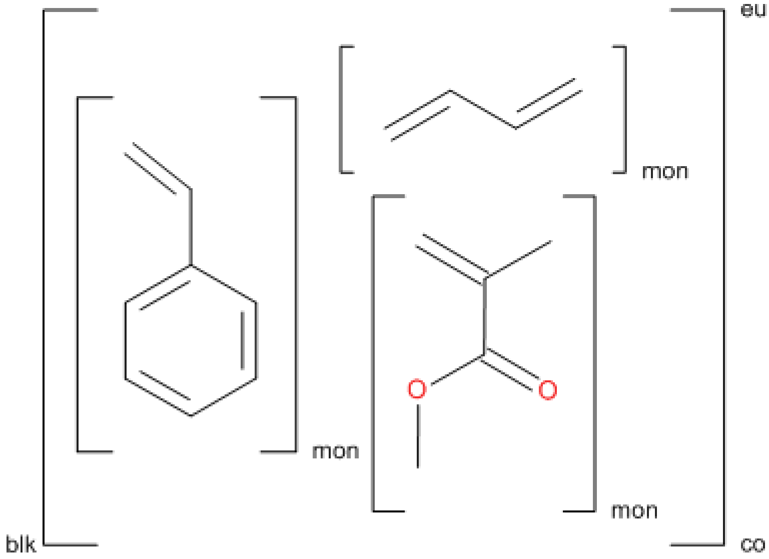


represents AAA...BBB...CCC
**or**
BBB...AAA...CCC
for a total of six possible block sequences

blk = attached data on polymer type

Polystyrene-*block*-polybutadiene-*block*-poly(methyl methacrylate)

blk = attached data on ploymer type



# Chapter 9: Exact Search (Flexmatch)

## Flexmatch Switches

The flexmatch search operator finds structures that match your query molecule exactly, except in ways that you specify with any of the flexmatch switches. The flexmatch switches allow you to selectively restrict or relax the criteria that are used to determine whether a structure is an exact match.

The information in this chapter describes how to use the flexmatch switches to obtain the results that you want. For example, you can use flexmatch switches to find:

- Structures that match your query, including higher order stereochemistry such as octahedral, trigonal pyramidal and square planar stereochemistry at certain atoms.
- Structures that match your query, excluding higher order stereochemistry.
- Structures that match your query exactly, excluding tautomers
- Structures that match your query exactly, including tautomers
- Stereoisomers of your query, excluding tautomers
- Stereoisomers of your query, including tautomers
- Structures that are salts of your query structure, or that are parent compounds of the salts in your query.

You can also use flexmatch switches to specify whether isotopes and attached data must match. The flexmatch switches are used as:

- A criterion for determining whether a duplicate of the structure being registered already exists in the database. See Definition of Duplicate Structure on page 203.
- Arguments for the `flexmatch` search operator in BIOVIA Direct. For more information, see the *BIOVIA Direct  Reference Guide*.
- Arguments for the `rxnflexmatch` search operator in BIOVIA Direct.
- Parameters in the Exact Structure Map component in Pipeline Pilot.

**Note:** Both the target and the query must contain only structural features that can be registered to a molecule database. For example, neither target nor query can contain atom or bond query features. If either does, the flexmatch operation will fail. For information on features that cause flexmatch to fail, see Structural Features and Registration on page 202.

## How to Specify Switches

You have the following options available for use with flexmatch searches:

- Three-letter switches
- Long-name switches

To set or enable switches, list them in the argument, for example, `FRA,TAU,SAL` provides a tautomer match. `CHA,RAD IgnoreChargesinPISystems` provides a metal search that ignores matching charge and radical counts.

## Three-letter Switches

If you specify more than one flexmatch switch, use a comma (,) or a forward slash (/) as a delimiter, for example FRA,TAU,SAL or FRA/TAU/SAL.

Each flexmatch switch that you set adds a condition to the matching criteria, which generally permits fewer hits. Switches that are not included in the list are Off (not set). For example, if you specify solely FRA, all other switches, such as SAL, are turned off.

Use MATCH to set switches, for example, MATCH=FRA,TAU,SAL is equivalent to FRA,TAU,SAL. Use ALL to set the most restrictive set of compatible switches. Use NONE to set all switches Off.

Use IGNORE to turn off the specified switches, for example, IGNORE=DAT ignores attached data, but sets all other switches.

**IMPORTANT!** Be careful when using IGNORE, since any switches that you do not explicitly include in the list will be turned on. The results might surprise you.

## Long-name Switches

Long-name switches (such as IgnoreHigherOrderStereo) are qualifiers that can increase the number of search results.

You cannot include long-name switches within MATCH or IGNORE lists. You must separate long-name switches with a space rather than a comma. The following examples demonstrate the proper syntax:

- Using a long-name switch within a list:

```
FRA,TAU,SAL,STE IgnoreHigherOrderStereo
```

- Using a long-name switch in conjunction with a MATCH list:

```
MATCH=FRA,TAU,SAL,STE IgnoreHigherOrderStereo
```

- Using a long-name switch in conjunction with a MATCH list:

```
IgnoreHigherOrderStereo MATCH=FRA,TAU,SAL,STE
```

For more information, see Dependencies Between Switches on page 114.

# Description of Switches

This section contains descriptions of individual switches. For examples that show how switches are used, see Examples of the Interactivity of Flexmatch Switches on page 115 and Useful Combinations of Flexmatch Switches on page 130.

| Name | Description |
|---|---|
| **Three-letter Switches** | |
| BON | Bond types must match. |
| CHA | Atom charges must match |
| DAT | Data sgroups must match |
| END | Polymer end groups must match |
| FRA | No additional fragments |

| Name | Description |
|---|---|
| HYD | Atom hydrogen counts must match |
| ION | Fragment charges must match |
| MAS | Atom isotope values must match |
| MET | Bonds to metals must match |
| MIX | Mixture types and components must match |
| MSU | Monomers do not match SRUs |
| POL | Polymer definitions must match |
| RAD | Atom radicals must match |
| SAL | Salt fragments must match |
| STE | Atom and bond stereochemistry must match |
| TAU | Tautomeric bond types must match |
| TYP | Polymer types must match |
| VAL | User-defined (abnormal) atom valence must match |
| **Long-name Switches** | |
| IgnoreChargesinPiSystems | Ignore total charge and radical counts in pi systems |
| IgnoreHigherOrderStereo | Ignore higher order (e.g. octahedral) atom stereochemistry |
| IgnoreTerminalPhosphates | Ignore terminal phosphates on RNA/DNA sequences |

## Isotopes (MAS)

When *On*, the MAS switch specifies that all isotopic labels must match those on your query. When *Off*, structures with different isotopic labels are perceived as equivalent.

## Bonds (BON, STE, TAU)

Use these switches to specify how closely bonds in the target match bonds in the query. For examples that use these switches, see Exact Match/As Drawn plus Tautomers on page 131 and Exact Match/As Drawn plus Stereoisomers on page 132.

### Bond (BON)

If you turn on BON, all bond types must match, including aromatics and perceived topology (ring or chain).

Turning on BON automatically disables TAU, the tautomer switch.

### Stereochemistry (STE,IgnoreHigherOrderStereo)

When ***ON***, STE specifies that stereoisomers of your query must match. You can retrieve stereoisomers that contain:

- Geometric stereochemistry for atoms separated by a double-bond and with each endpoint having two additional single bonds, including implicit hydrogens. Trivalent N is allowed if the additional attachment is not a hydrogen.
- Tetrahedral stereochemistry for the following atoms: C, Si, P, As, S, Se, Te, and their isoelectronic equivalents. Only nitrogen has tetrahedral stereo in charged quaternary form. Neutral P and As are tetrahedral in the pentavalent form (three single bonds and one double bond). Neutral S, Se, Te are tetrahedral in one of the tetravalent forms (2 single bonds, one double bond and a chiral lone pair) and one of the hexavalent forms (2 single bonds and two double bonds).
- Allenes.
- Biaryls with hindered rotation – heavy attachments at least 3 of the 4 ortho locations.
- Higher order stereochemistry: square planar, trigonal bipyramidal and octahedral.

The following types of stereoisomers are perceived as equivalent:

- Metal complexes or other higher order stereochemistry which is not square planar, trigonal bipyramidal or octahedral.
- Complex structural stereochemistry, such as spiral polyphenanthrenes.

For E/Z geometric stereochemistry, you must use a double bond to define the stereochemistry. A query with a double bond retrieves solely structures with the same stereochemistry, plus structures with the Double Either bond. Query structures that use Double Either bond have undefined stereochemistry and therefore can retrieve only structures with the Double Either bond.

For tetrahedral, allene, and biaryl and higher order stereochemistry, your query must use Up and Down stereo bonds to define the stereochemistry. A query that lacks stereo bonds, or that contains Either bonds, retrieves solely structures with undefined stereochemistry.

For tetrahedral stereochemistry, STE requires that the stereogroup labels also match. For more information and examples of matching stereogroup labels, see Flexmatch Search of Structures with Tetrahedral Stereochemistry on page 118.

If STE is present and IGNOREHIGHERORDERSTEREO is also present, then tetrahedral, allene and biaryl stereo is validated but higher-order stereo is ignored.

### Tautomer Bonds (TAU)

When *On*, the tautomer groups in the query structure and the target structure must contain the same number of hydrogens in the tautomeric region, within a tolerance limit. The tolerance limit is the sum of the absolute values of charges plus the sum of the number of radicals plus the number of metal bonds. Diradicals count as two radicals.

When BON is On and TAU is *Off*, the impact of TAU is to apply hydrogen matching to each tautomeric region, rather than to the structure as a whole.

If you turn on both HYD and TAU, the tolerance limit is zero.

A structure can contain multiple tautomeric groups, and these groups can overlap. A tautomeric group might include part of an aromatic ring.

## Salts and Parent Compounds (CHA, ION, FRA, SAL)

The switches in this section are used to search structures that are perceived as salts and parent compounds according to the BIOVIA Salt definition.

For information on the BIOVIA Salt definition, see Customizing the BIOVIA Salts Definition on page 215.

For examples that use these switches, see Exact Match/As Drawn plus Salts on page 133.

## Charge (CHA and ION)

When *On*, the CHA switch specifies that the charge on each atom must match those in your query. In contrast, the ION switch specifies that the total charge in each fragment must match.

CHA is a specific case of the more general ION switch. If you want only the sum of the charges to match, without matching charges atom-by-atom, turn *off* CHA and turn *on* ION.

Setting CHA *On* disables the ION switch.

## Fragments (FRA)

When *On*, FRA specifies that each fragment in the target structure must match each fragment in you query, with no additional fragments allowed. Setting FRA to *Off* allows additional fragments.

For example, a query that contains fragments A and B only matches structures that contain both A and B. The query does not match structures with fragments A, B, and C, where C is a fragment that is not in your query.

For polymer structures, the settings of POL and TYP influence how structures are perceived. For details, see Copolymer Search on page 136.

## Hydrogen Count (HYD)

When *Off*, HYD specifies that the number of hydrogen atoms on each fragment must match your query, within a tolerance limit. The tolerance limit for each fragment is the sum of the absolute value of charges plus the number of radicals plus the number of metal bonds. Diradicals count as two radicals.

When HYD is *On*, the tolerance limit is set to zero, so the number of hydrogen atoms on each fragment must match your query exactly. Additionally, when BON, CHA, and RAD are also *On*, the number of hydrogen atoms on matching query and target atoms must be the same.

## Salts (SAL)

When *On*, SAL specifies that the salt counterions in the structures retrieved must match the counterion in your query. A counterion is any fragment or single atom in the Salts definition. You can customize the Salts definition to add your own counterions. For more information, see Customizing the BIOVIA Salts Definition on page 215.

When SAL is *Off*, you retrieve molecules with different counterions than those in your salt query. When SAL is *Off* and FRA is *On*, you retrieve only the parent structures of salts because counterions are removed from the query before matching.

**Note:** The SAL switch affects search results solely if your business rules for salts use a single chemical structure to represent the salt. Alternative data models might represent salts differently. For example, you might want to store only the parent compound as a chemical structure and store information on counterions and hydrates in a separate database field. You might also define a salt as a substance with the chemical structures that comprise the salt, the parent compound, counterion, water of hydration, in different rows within an Oracle database table.

# Alternative Structure Representations (MET, RAD, VAL, IgnoreChargesInPiSystems)

Use the switches in this section to loosen the criteria for structures such as metallocenes and organometallic compounds to retrieve alternative representations of these structures. For examples that use these switches, see Thiocarboxylic Acid Salts on page 115 and Organometallic Complexes on page 116.

## Metal Bonds (MET)

When *On*, `MET` specifies that the connectivity of all metal bonds must match your query. A metal is any atom other than H, D, T, He, B through Ne, Si through Ar, As through Kr, Te, I, Xe, At, and Rn.

If `MET` is *Off* and `SAL` is also *Off*, any metals that are listed in the Salts file are removed before matching.

If both `MET` and `BON` are *Off*, bond types are ignored for bonds connected to metals. For example, you could match all representations of ferrocene regardless of how the pi-metal bonds are represented.

> **Note:** The metal bonds are still included in the hydrogen tolerance (see `TAU`), so you might also want to turn off `HYD`.

### Radicals (RAD)

When *On*, `RAD` specifies that the radical value on each atom must match those in your query.

### Valence (VAL)

You can specify that the valence on each atom must match those in your query.

If no valence is specified, an implicit valence is calculated from the structure. For more information, see Valences and Implicit Hydrogens on page 4.

A query with an unspecified valence, a calculated implicit valence, can retrieve a structure that contains the matching explicit valence. Alternatively, a query with an explicit valence can retrieve a structure that contains a calculated implicit valence of equal value.

Some valences are illegal. For transition metals, a valence is illegal if it is smaller than the number of bonds to the atom. For other real atom types, an illegal valence is ignored, and the query retrieves any valence.

### IgnoreChargesInPiSystems

By default, if CHA and RAD are turned on, a query containing a pi-system (haptic bonding) matches only a target that has the same total charge and radical count within its pi-system. Add the `IgnoreChargesInPiSystems` flag to the front of your flexmatch switches to allow pi-systems that have different total charges and radical counts to match. For example, a cyclopentadienyl anion in the query and a cyclopentadienyl radical in the target.

For an example that uses this switch, see Organometallic Complexes on page 116.

## DNA and RNA Sequences (IgnoreTerminalPhosphates)

Add `IgnoreTerminalPhosphates` to your Flexmatch switches to allow an RNA or DNA sequence query containing a 3' or 5' phosphate to match a target that does not have a 3' or 5' phosphate, and vice versa. The switch causes both 3' and 5' phosphates to be ignored in query and target.

For example, an RNA sequence, ACG, created with BIOVIA Draw will contain a 5' phosphate but not a 3' phosphate. It is equivalent to the HELM string `RNA1{P.R(A)P.R(C)P.R(G)}$$$$`. Drawing ACG in a HELM editor will typically create the HELM string `RNA1{R(A)P.R(C)P.R(G)P}$$$$`, which does not have a 5' phosphate but does contain a 3' phosphate. These two sequences will match one another only if you include the `IgnoreTerminalPhosphates` switch. For example `'IgnoreTerminalPhosphates MATCH=ALL'`.

For more information, see Representation of Nucleic Acids on page 54.

## Substance Groups (Sgroups)

Use these switches to specify features on structures that contain polymer Sgroups, mixture Sgroups, or attached data (data Sgroups). The On or Off state of these switches does not affect the results that you obtain from structures that do not contain Sgroups.

For more information on Sgroups, see Chemical and Data Substance Groups (Sgroups) on page 29.

For examples that use these switches, see Flexmatch Search of Polymers on page 128.

### Attached Data (DAT)

When *On*, specifies that attached data (Sgroup data) must match the attached data in your query.

> **Note:** BIOVIA recommends that you use attached data on polymer brackets to different types of polymers and copolymers. For more information, see Attached Data on page 154.

### Polymer End Groups (end)

When *On*, `END` specifies that polymer end groups must match those in your query. A polymer end group is a connected set of atom and bonds that:

- Is connected to the rest of the structure by just one crossing bond (a crossing bond is a bond that crosses a bracket).
- Contains no brackets.

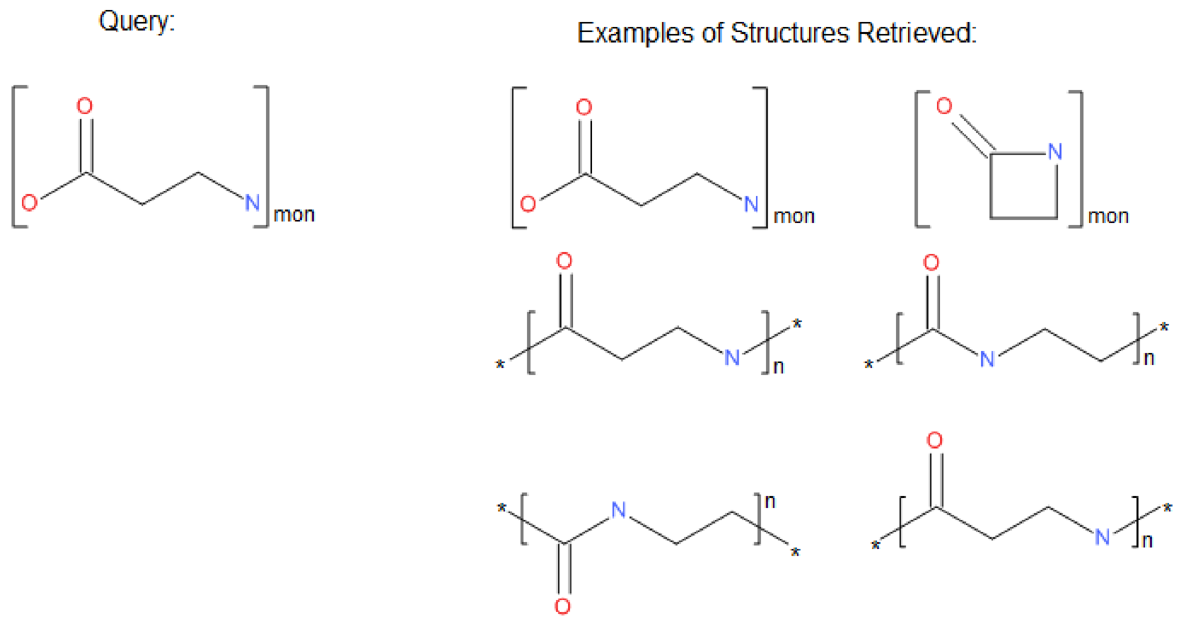
`END` depends on `POL`. `END` can be on only when `POL` is also on.

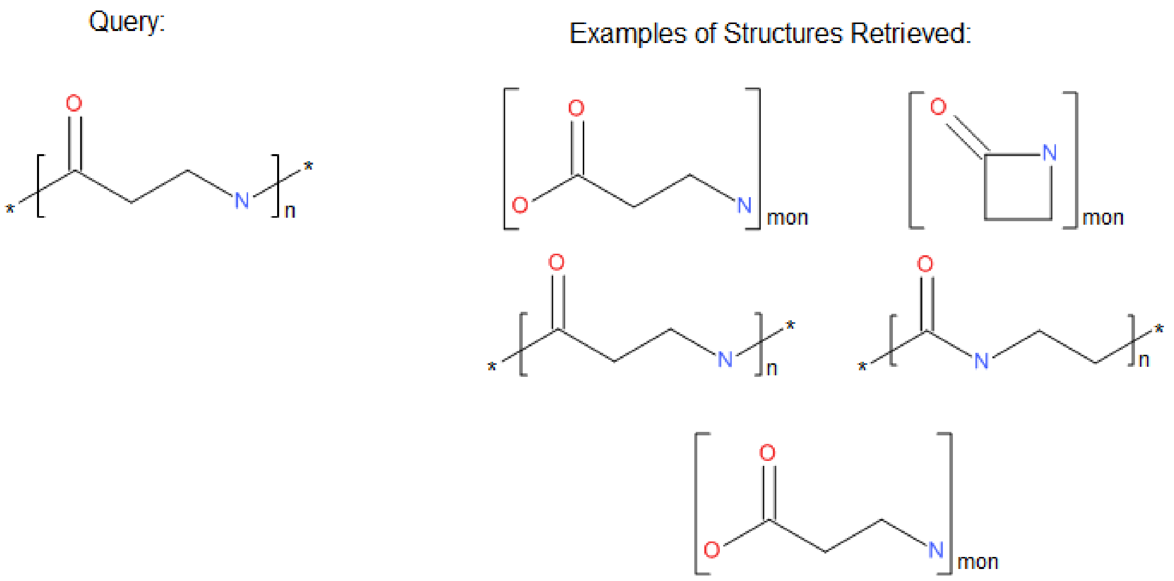
### Mixtures (MIX)

When *On*, `MIX` specifies that the type of mixture, ordered or unordered, and the order of components in an ordered mixture, must match the corresponding structures in your query.

### Monomer/SRU Uniqueness (MSU)

`MSU` specifies that the source-based and structure-based representations of the same polymer are considered distinct structures. When `MSU` is *Off*, the source-based and structure-based representations of a polymer find each other. For example:





MSU depends on both POL and TYP. POL and TYP must be on for MSU to function.

When MSU is *Off*, the settings of BON and HYD for structures that are within monomer and SRU brackets are ignored.

## Polymers (POL)

When *On*, POL specifies that all polymer definitions must match. The polymer definition is the set of atoms and bonds that is associated with a polymer Sgroup.

The TYP, END, and MSU switches depend on POL. That is, turning off POL automatically turns off TYP, END, and MSU. For monomer and SRU brackets, the setting of POL also determines how fragments within the polymer are perceived. For more information, see Fragments (FRA) on page 111 and Copolymer Search on page 136.

## Polymer Type (TYP)

When *On*, TYP specifies that the polymer Sgroup or bracket type and polymer connectivity must match your query. Polymer Sgroup types include: monomer (mon), SRU (n), mer, crosslink (xl), modification (mod), unspecified copolymer (co), alternating copolymer (alt), graft copolymer (grf), block copolymer (blk), and random copolymer (ran).

> **Note:** Modification (mod), alternating copolymer (alt), graft copolymer (grf), block copolymer (blk), and random copolymer (ran) were used in earlier, obsolete conventions for representation of polymers.

TYP depends upon the setting of POL. That is, POL must be on for TYP to be on. MSU also depends on TYP; TYP must be on for MSU to be on. For monomer and SRU brackets, the setting of TYP also influences how fragments in the polymer are perceived. For more information, see Copolymer Search on page 136.

# Dependencies Between Switches

Certain flexmatch switches depend on the settings of other switches. Some switches disable other switches while other require other switches to be turned on. For the following two cases, turning on a switch for a specific property disables" a switch for a more general property.

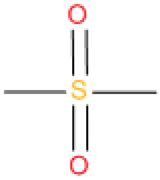
| Switch | Setting | Dependent Switch | Setting |
|---|---|---|---|
| CHA | On | ION | On |
| BON | On | TAU | Off |
| POL | Off | TYP | Off |
| POL | Off | END | Off |
| TYP | Off | MSU | Off |

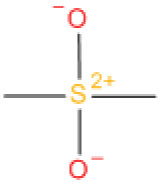
## Examples of the Interactivity of Flexmatch Switches

The flexmatch switches are typically interactive. To obtain the correct results, you need to know which flexmatch switches require other flexmatch switches to function correctly. With the correct interactive settings, you can retrieve closely related structures that are represented in a variety of ways in your database.

### Sulfones

Sulfones can be represented as:





If your database contains both representations, you can retrieve them both with the same query by setting TAU and ION On with BON and CHA Off:

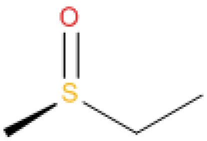
'TAU,ION'

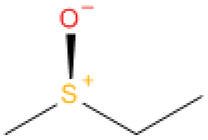
Alternatively, with additional switches that do not affect tautomer or ion matching:

'TAU,ION,FRA,HYD,MAS,MET,RAD,SAL,STE,VAL,DAT'

### Sulfoxides

Sulfoxides can be represented as:



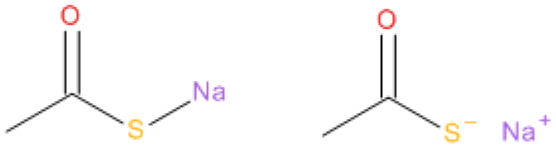


If your database contains both representations, you can retrieve them both with the same query by using the same settings as for sulfoxides, plus the STE switch to ensure matching stereo bonds:

'TAU,ION,STE'

### Thiocarboxylic Acid Salts

Thiocarboxylic acid salts can be represented as:



If your database contains both representations, you can retrieve them both with the same query by setting TAU and SAL to On with BON, ION, MET, and CHA Off:

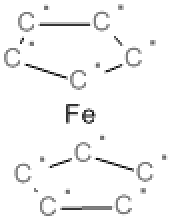
```
'SAL,TAU'
```

Alternatively, with additional switches that do not affect matching of tautomers, salts, or metal bonds:

```
'FRA,HYD,MAS,MET,RAD,SAL,STE,VAL,DAT'
```

## Organometallic Complexes

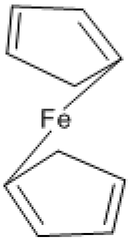
BIOVIA databases depict ligands in organometallic complexes with radicals. For example, ferrocene is represented as:



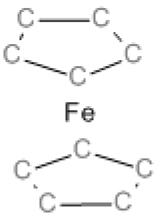
The following combination of flexmatch switches:

```
'ION,TAU'
```

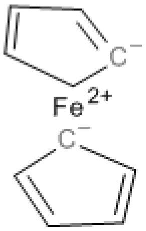
matches these representations of ferrocene:



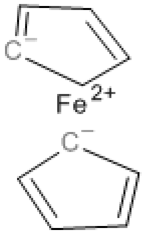
A



B



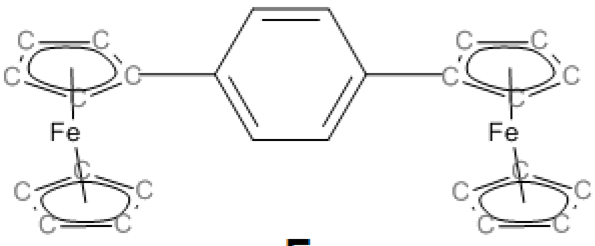
C



D

A, C, and D match B because RAD and FRA are *Off*. In addition:

- A matches B because MET is *Off*.
- C and D match B because ION is On and CHA is *Off*. The following structure also matches, because FRA is *off*:

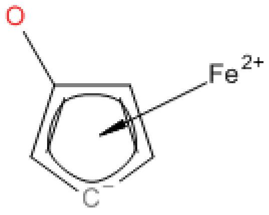


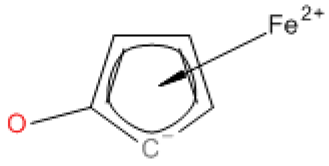
E

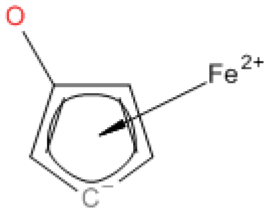
The table that follows summarizes the effect of individual switches on the structures that match the query:

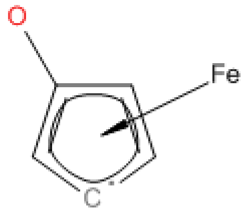
| | Matching Structures | | | | |
|---|---|---|---|---|---|
| **Switch Set On** | **A** | **B** | **C** | **D** | **E** |
| (none) | X | X | X | X | X |
| BON | | X | | X | |
| CHA | X | X | | X | |
| FRA | X | X | X | | |
| ION | X | X | | X | |
| MAS | X | X | X | X | |
| MET | | | | | |
| RAD | X | | X | | |

Newer BIOVIA databases represent pi-bonded ligands in organometallics using the haptic bond type to indicate the bond between the pi system and the metal. When matching two organometallics which both use haptic bonds, flexmatch searches automatically disregard the bond types within the pi system and the position of charge and radical within the pi system. Even with flexmatch switches set to `MATCH=ALL` the following structures will be seen as identical even though the negative charge is in a different position in the ring:





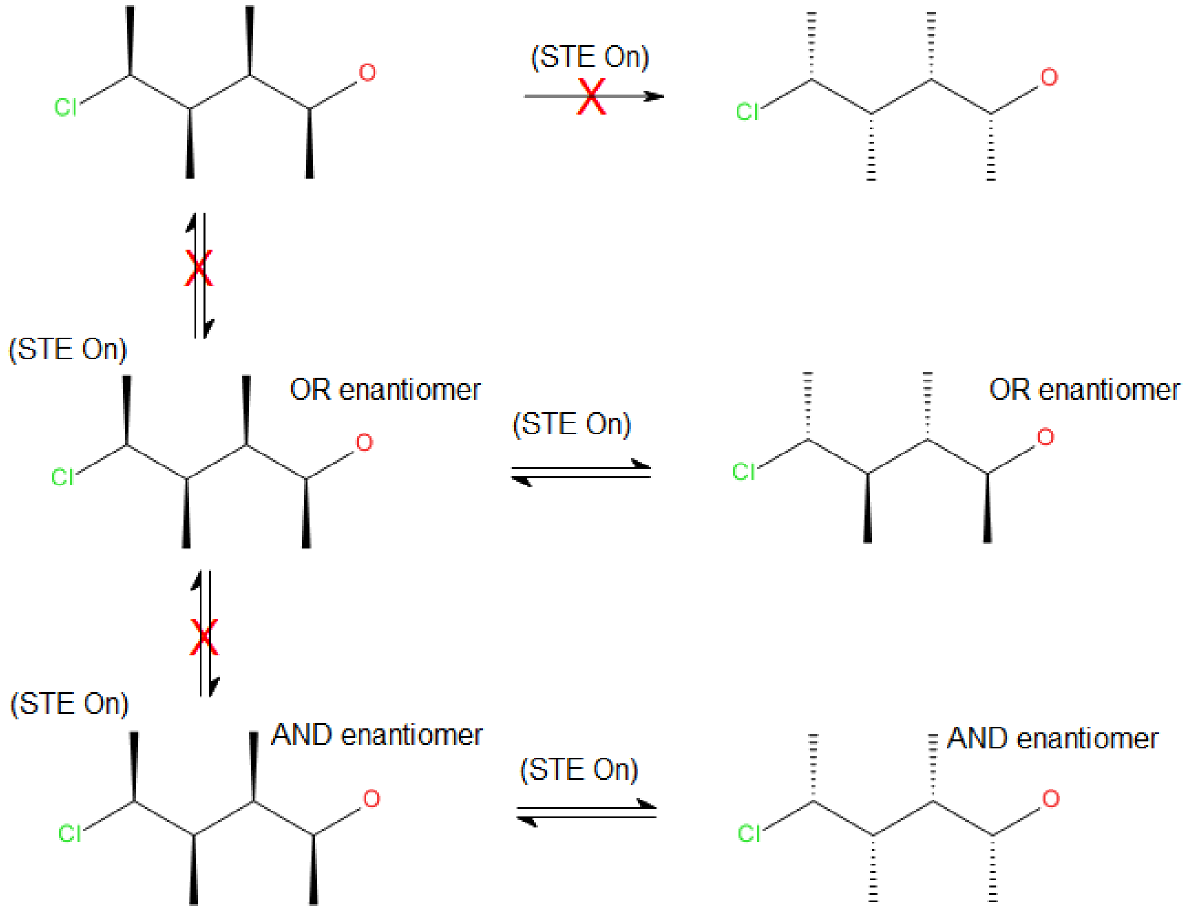




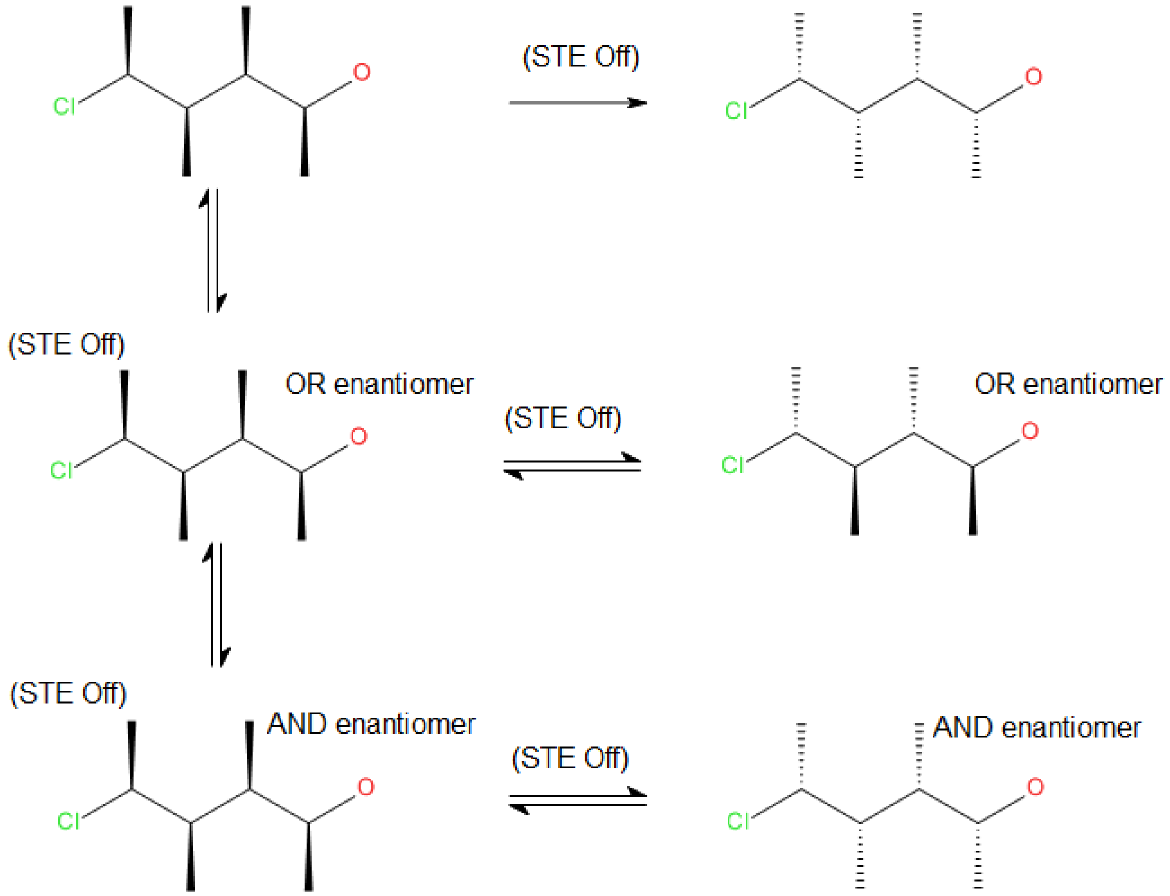
Total charge and radical counts within the two pi systems and the attached metals must match, assuming CHA and RAD switches are set. You may add the flag `IgnoreChargesInPiSystems` at the front of your flexmatch query switches to relax this restriction and allow for example the following two structures to match even with flexmatch switches set to ALL. The flexmatch switches in this case would be `IgnoreChargesInPiSystems MATCH=ALL`.

## Flexmatch Search of Structures with Tetrahedral Stereochemistry

Exact search uses the flexmatch parameters that provide the most restrictive match of structural features. The flexmatch parameter STE controls the matching of stereochemical groups and stereo bonds. When STE is on, as is the case in Exact search, the groups of stereogenic centers on the structures retrieved must match those of the query exactly. Thus, a structure with all stereogenic centers in the ABS stereogroup does not retrieve a structure with stereogenic centers in OR and AND stereogroups, and so on. For example:



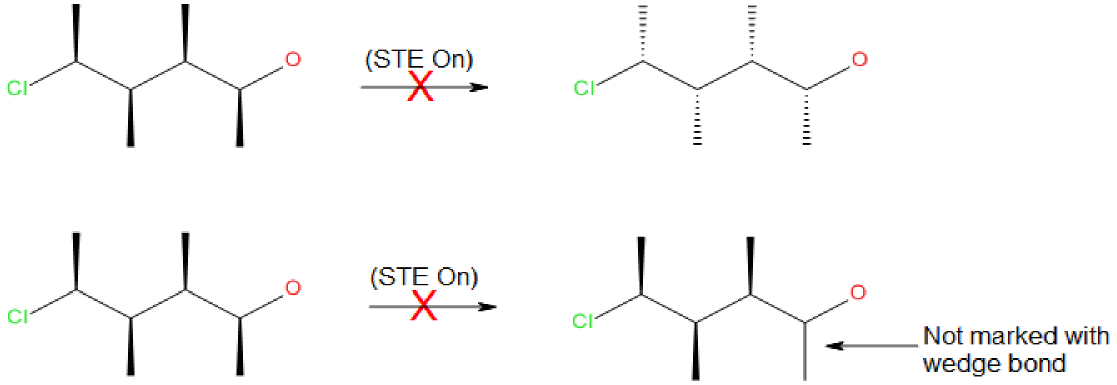
In an Isomer search, the flexmatch switch STE is set *Off*, which causes stereo bonds and the stereo labels on atoms to be ignored. Consequently, all the structures shown in the previous figure for Exact match can retrieve each other in an Isomer search.



The sections that follow provide more details and examples of Exact stereo searching STE On.

## Absolute Configuration

In exact search, structures with all stereogenic centers in the ABS stereogroup match only if all stereogenic centers have exactly the same stereoconfiguration, as implied by Up or Down stereo bonds and stereogroups:

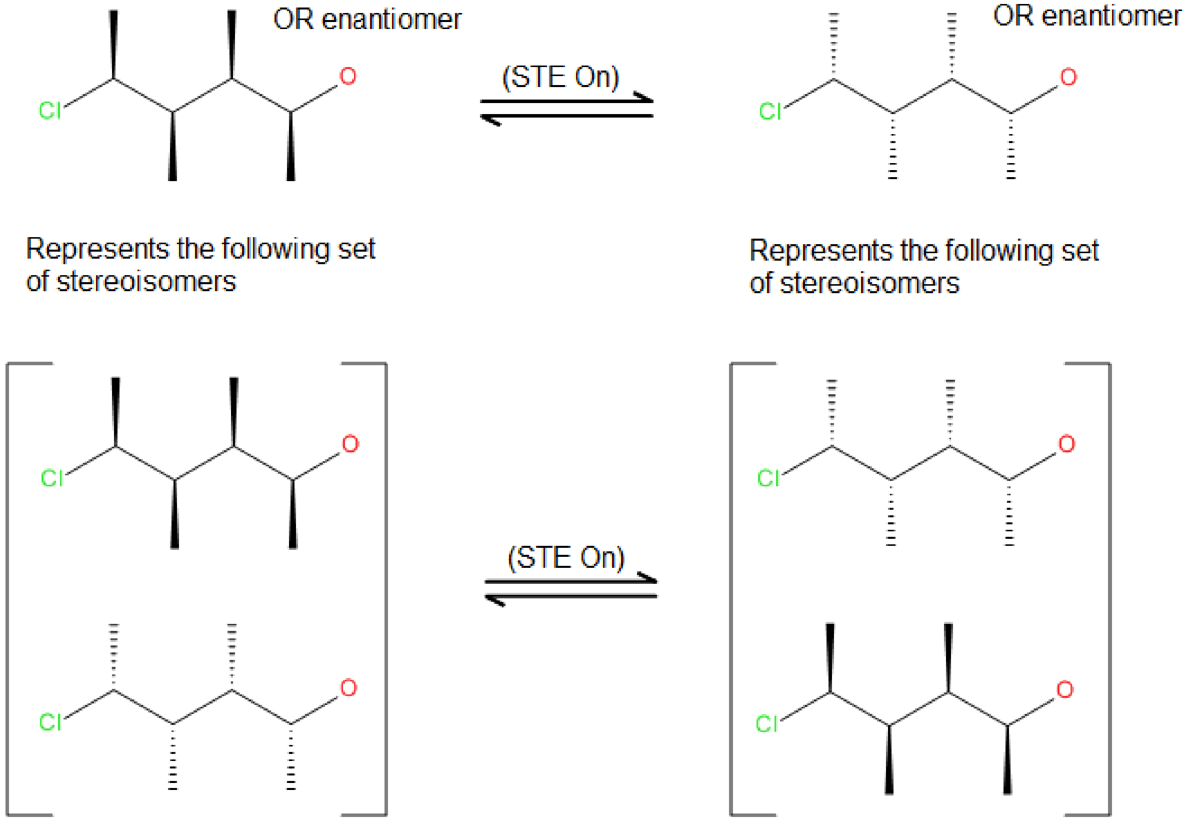


## Relative Stereoconfigurations (OR groups)

See the following examples.

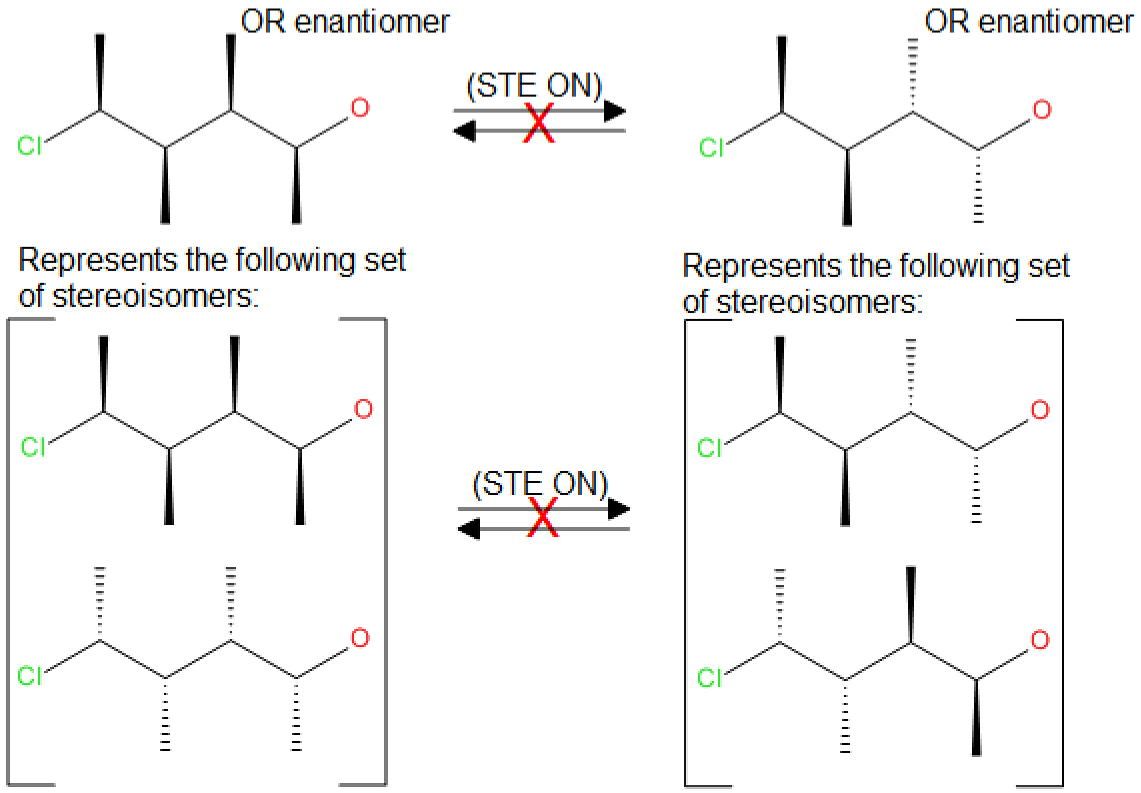
### Example 1

For stereogenic centers in OR groups to match, the structure as drawn, with Up or Down stereo bonds, must represent one of the alternative stereoisomers. The reasons for this become clear when the stereoisomers that each structure represents are compared. In the example that follows, both structures as drawn represent the same stereoisomer because the Up and Down stereo bonds indicate the same relative configuration, and the OR enantiomer label indicates that all stereogenic centers are in the same OR group:


OR enantiomer
Cl
O
(STE On)
OR enantiomer
Cl
O
Represents the following set of stereoisomers
Represents the following set of stereoisomers
Cl
O
Cl
O
(STE On)
Cl
O
Cl
O

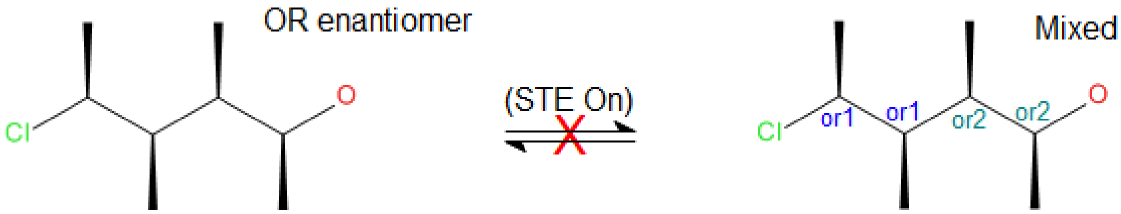
### Example 2

In this example, the Up and Down stereo bonds on the structures denote that the relative configurations are different. Consequently, the two structures do not represent the same stereoisomer:


Cl
O
OR enantiomer
(STE ON)
Cl
O
OR enantiomer
Represents the following set of stereoisomers:
Cl
O
Cl
O
(STE ON)
Represents the following set of stereoisomers:
Cl
O
Cl
O

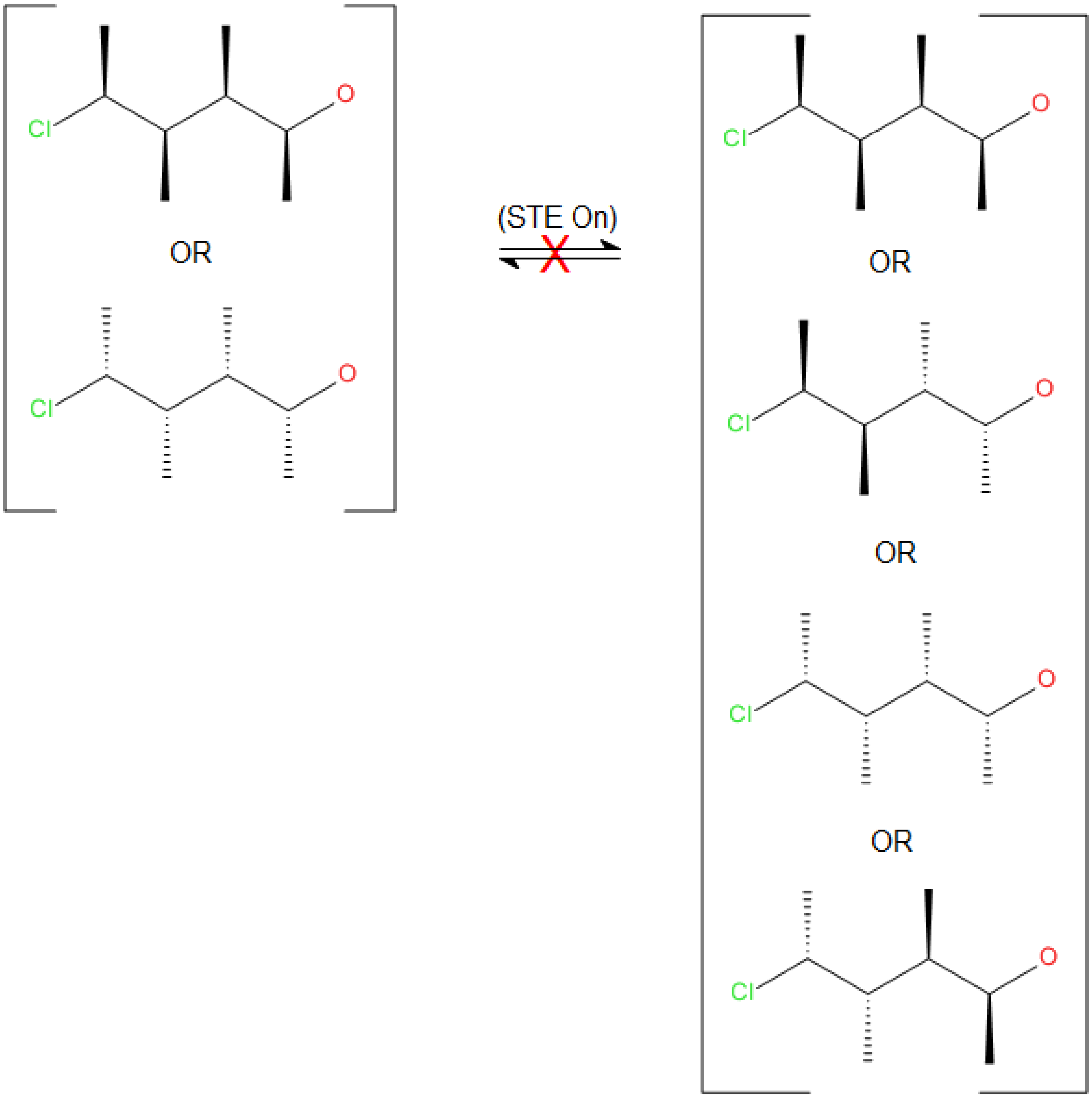
### Example 3

In this example, the stereo bonds in the structures are drawn exactly the same. The structure on the left, however, represents an alternative between only two stereoisomers, while the structure on the right represents one alternative among four possible stereoisomers:



Represents the following set of stereoisomers

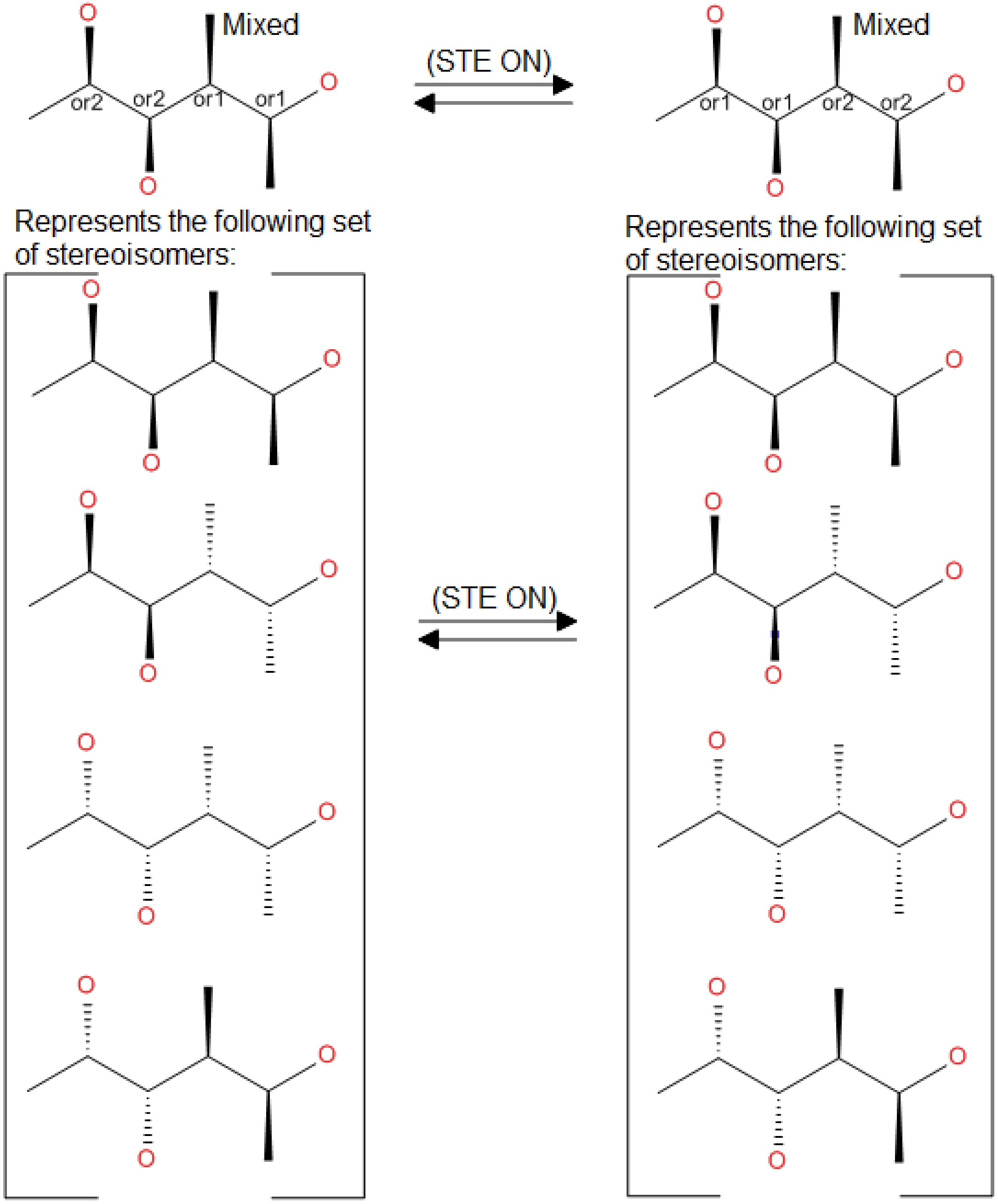
Represents the following set of stereoisomers



**Note:** In a substructure search, the structure on the right does match the structure on the left, because the stereoisomers that the left-hand structure represents are entirely contained within the stereoisomers that the right-hand structure represents. For more information on substructure search of stereoisomers, see Substructure Search of Structures with Tetrahedral Stereochemistry on page 171.

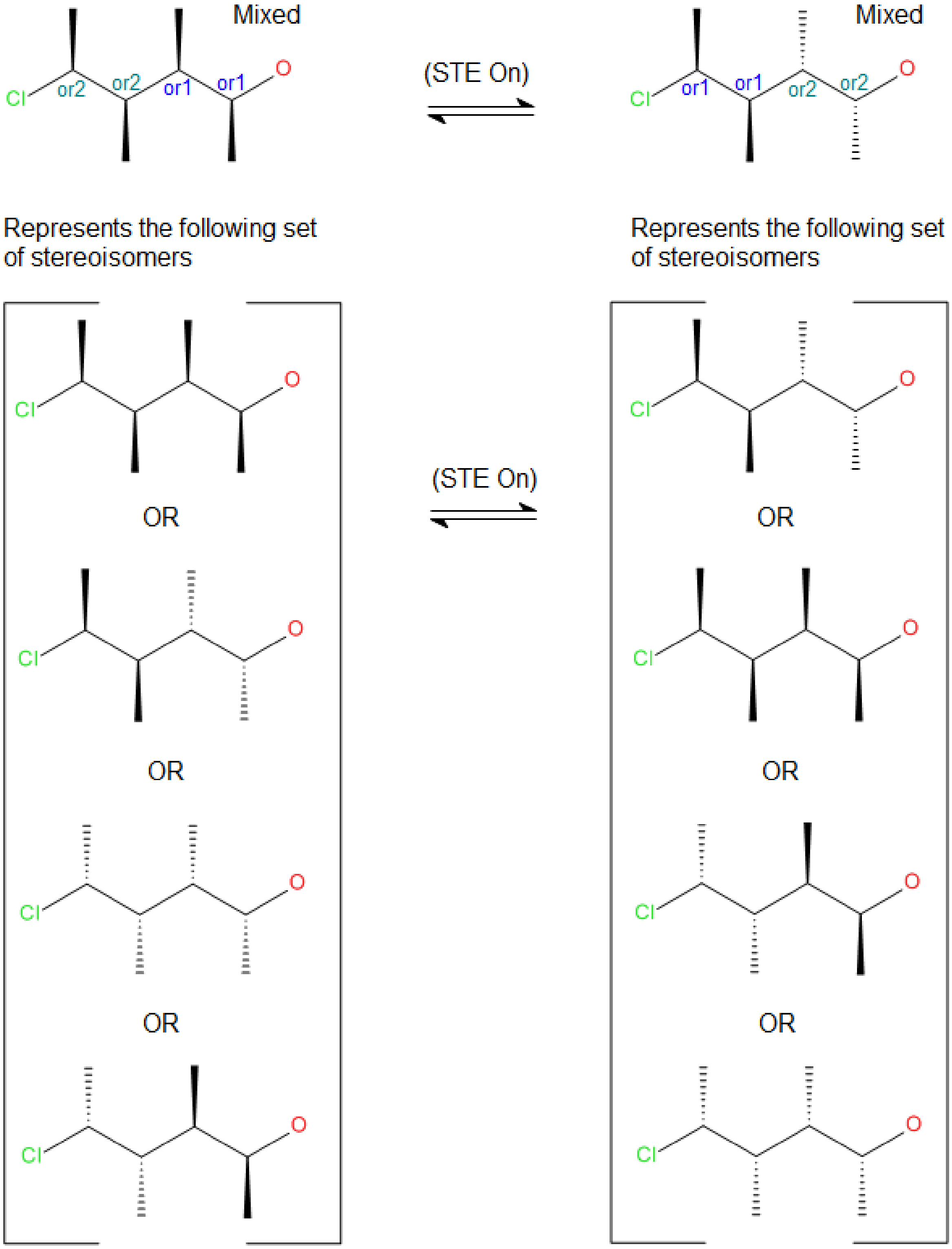
## Example 4

In this example, the structures match because each structure represents the same set of stereoisomers. Even though the index numbers on the OR groups are reversed, the two structures represent exactly the same set of alternative stereoisomers. The actual values of the index numbers on OR groups are unimportant. Only the group definitions must be the same:


Mixed
or2
or2
or1
or1
(STE ON)
Mixed
or1
or1
or2
or2
Represents the following set of stereoisomers:
Represents the following set of stereoisomers:
(STE ON)

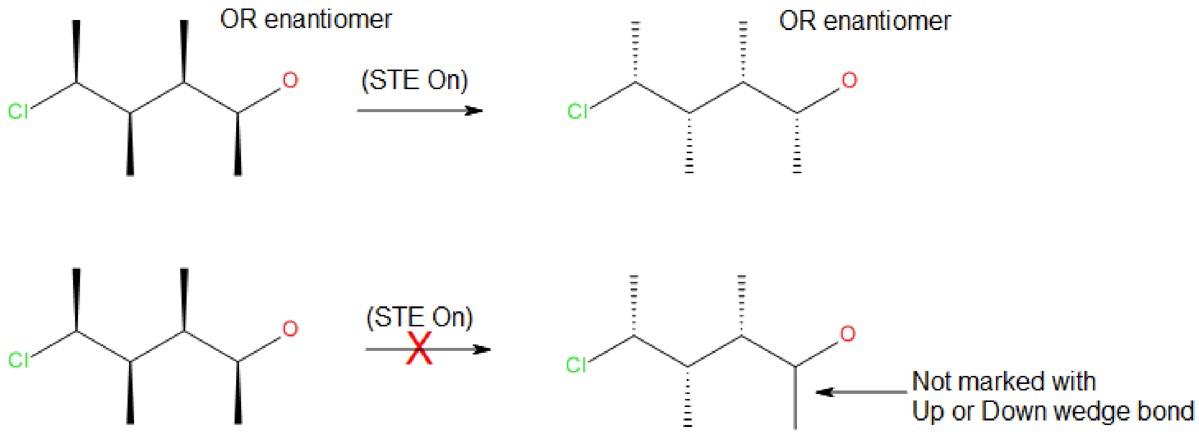
## Example 5

In this example, the stereo bonds in one structure represent a different member of the set of alternative stereoisomers. However, the two structures still represent exactly the same set of stereoisomers and are equivalent:



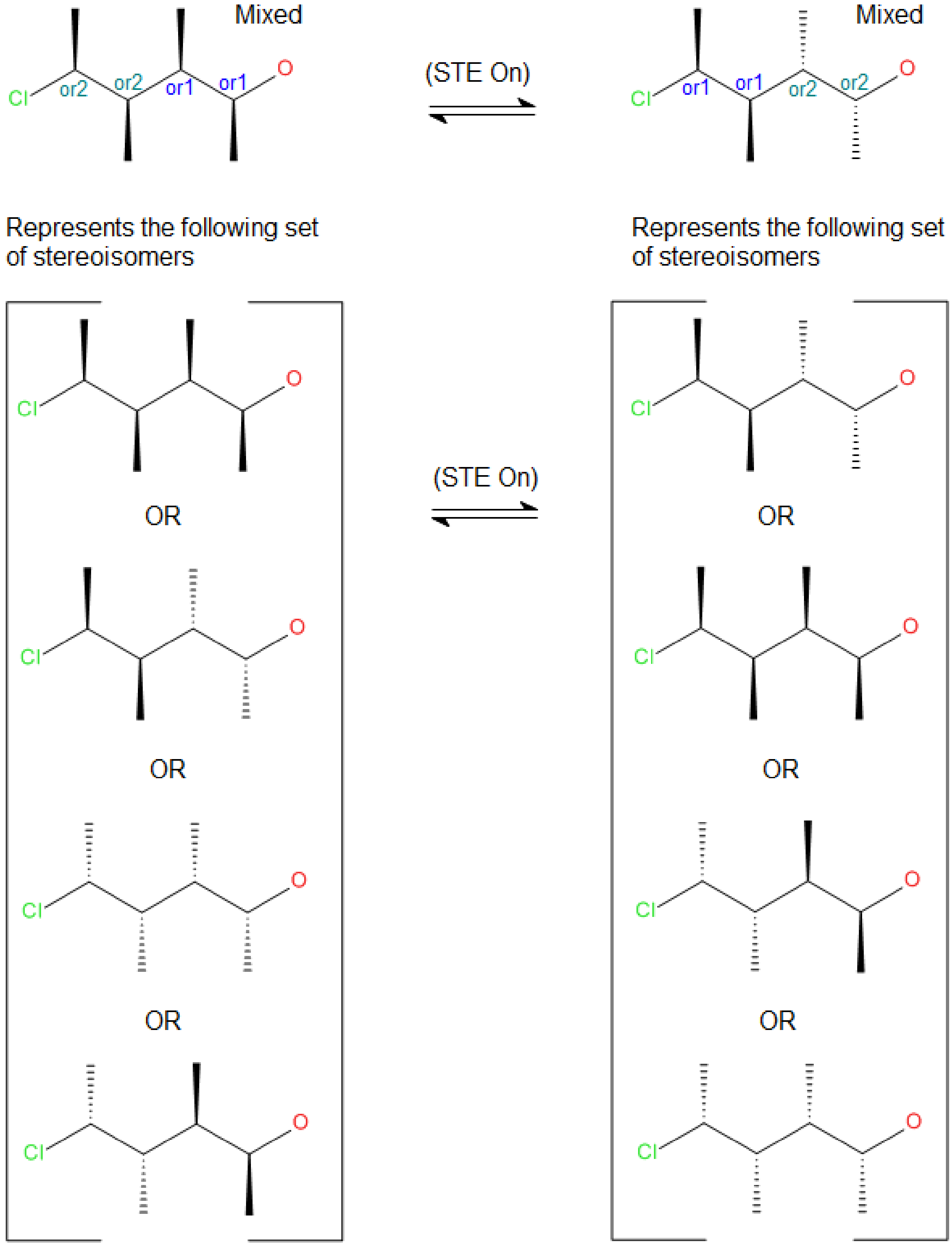
## Example 6

Structures with OR groups match solely when all the stereogenic centers are defined, that is, are marked with Up or Down stereo bonds. For example:



## Mixtures of Relative Stereoconfigurations (AND groups)

Everything that applies to flexmatch matching of OR groups applies also to flexmatch matching of AND groups. Specifically, the examples in the preceding section behave exactly the same for AND groups as for OR groups. To get the rules for AND groups, substitute *AND* for *OR* and *&* for *or* in the drawings in the previous section. The figure below shows this for Example 5:


Mixed
Cl
or2
or2
or1
or1
O
(STE On)
Mixed
Cl
or1
or1
or2
or2
O
Represents the following set of stereoisomers
Represents the following set of stereoisomers
Cl
O
OR
Cl
O
OR
Cl
O
OR
Cl
O
(STE On)
Cl
O
OR
Cl
O
OR
Cl
O
OR
Cl
O

## Flexmatch Search of Polymers

To retrieve the polymer representation that matches your query exactly, use the following flexmatch switches:

'POL,TYP,MSU'

That is, use a source-based query to retrieve the source-based representation, and a structure-based query to retrieve the structure-based representation. To retrieve polymer structures in which the source-based representation of a polymer must match the corresponding structure-based representation, specify the following switches (MSU Off):

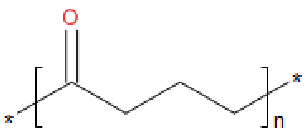
'POL,TYP'

To retrieve polymer structures in which the source-based representation of a polymer must match the corresponding structure-based representation, use the following switches:

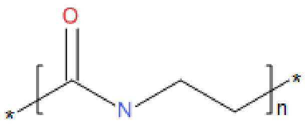
'POL,TYP,MSU' or
'FRA,HYD,MAS,MET,RAD,SAL,STE,VAL,BON,TAU,ION,CHA,MIX,POL,TYP,MSU'

Setting MSU On also ignores cyclization. For more information see Cyclization and Phase Shifting on page 80. For example:

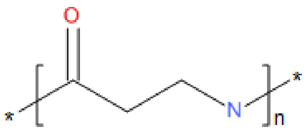
Query:



Example of
Record Retrieved:



Example of
Record NOT Retrieved:



To retrieve all three representations of the polymer, the query and phase-shifted version, set MSU *Off*.

If you also want polymer end groups to match exactly, add the END switch:

'POL,TYP,MSU,END'

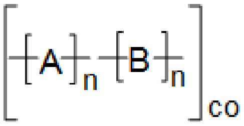
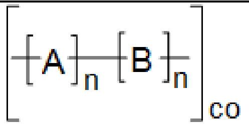
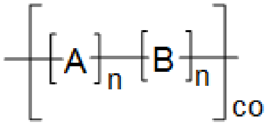
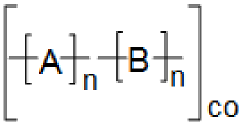
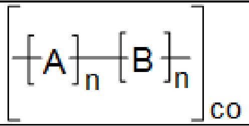
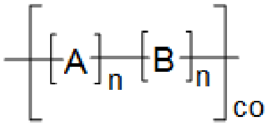
or

'POL,TYP,END'

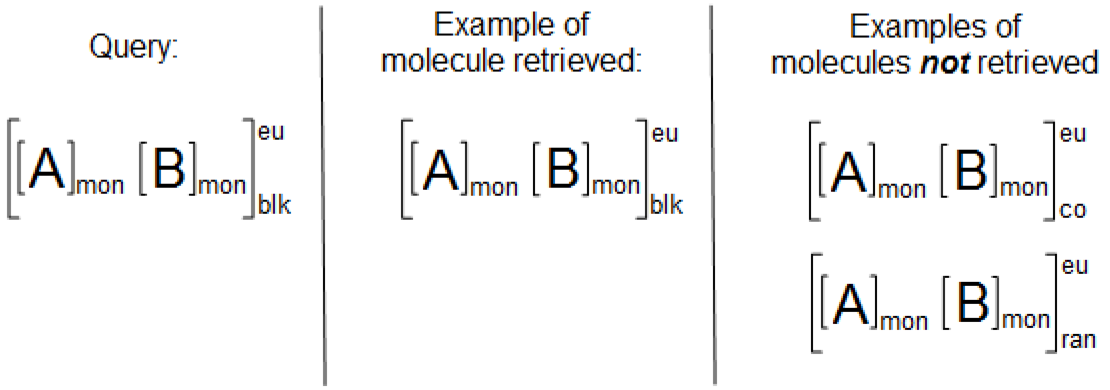
A polymer with the either/unknown repeat pattern finds all other repeat patterns. Other polymer repeat patterns must match exactly. The following table summarizes the search characteristics of repeat patterns for ladder-type polymers:

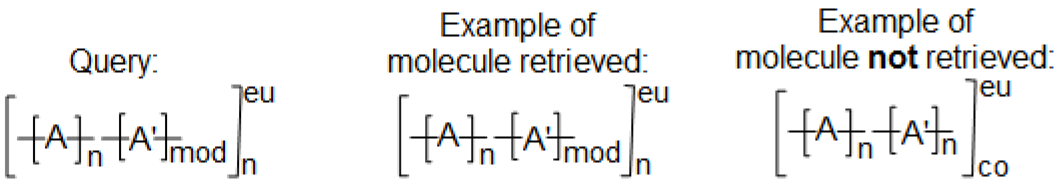
| | | Database Structures | | | | |
|---|---|---|---|---|---|---|
| | | Head-to-Tail With Flip | Head-to-tail With No Flip | Head-to-Head With No Flip | Head-to-Head With Flip | Either/ Unknown |
| Queries | Head-to-Tail With Flip | ✓ | | | | |
| | Head-to-tail With No Flip | | ✓ | | | |
| | Head-to-Head With No Flip | | | ✓ | | |
| | Head-to-Head With Flip | | | | ✓ | |
| | Either/ Unknown | ✓ | ✓ | ✓ | ✓ | ✓ |

Copolymers in which the sequence of SRUs is unspecified match structures in which the sequence is specified. The following table summarizes the search characteristics of copolymer structures that contain SRUs when the polymer switches are set to '`POL,TYP,MSU,END`' or '`POL,TYP,MSU`':

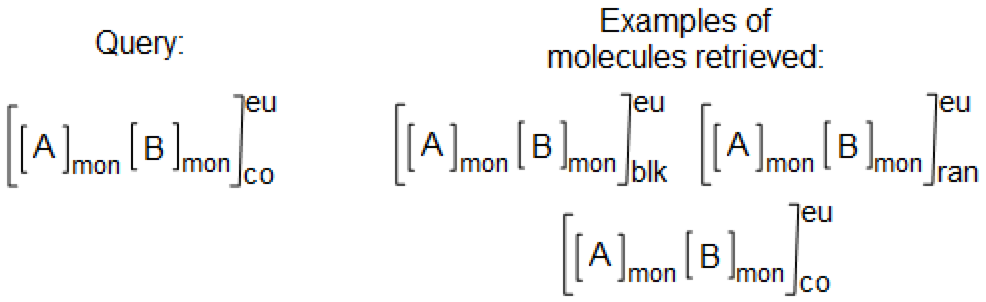
| | | Database Structures | | |
|---|---|---|---|---|
| | |  |  |  |
| Queries |  | ✓ | ✓ | ✓ |
| |  | | ✓ | ✓ |
| |  | | | ✓ |

Polymer and copolymer types must match the query exactly. For example:




Query:
Example of molecule retrieved:
Example of molecule **not** retrieved:
A n A' mod eu n
A n A' mod eu n
A n A' n eu co

Unspecified copolymer brackets (co) match any type of copolymer (co, blk, ran, alt).



## Useful Combinations of Flexmatch Switches

The flexmatch switches are provided as tools for developers to create custom searches.

**IMPORTANT!** Because the flexmatch switches often interact in complex and unexpected ways, it is strongly recommended that controls for flexmatch switches not be exposed to end users. Instead,BIOVIA recommends that you provide the searching capabilities of the switches through useful combinations, such as those that follow.

### Exact Match/As Drawn

To obtain the most restricted match to your query, specify all flexmatch switches. For example:

'FRA,HYD,MAS,MET,RAD,SAL,STE,VAL,BON,TAU,ION,CHA,DAT'

or because the ION switch is ignored when CHA is on, and BON disables TAU:

'FRA,HYD,MAS,MET,RAD,SAL,STE,VAL,BON,CHA,DAT'

If your query contains polymers and/or mixtures, add the MIX, POL, TYP, and MSU switches:

'FRA,HYD,MAS,MET,RAD,SAL,STE,VAL,BON,TAU,ION,CHA,DAT,MIX,POL,TYP,MSU'

or

'ALL' 'MATCH=ALL'

To retrieve both source-based and structure-based representations of a polymer query, remove the MSU switch:

'FRA,HYD,MAS,MET,RAD,SAL,STE,VAL,BON,TAU,ION,CHA,DAT,MIX,POL,TYP'

or

'IGNORE=MSU'

## Exact Match/As Drawn plus Tautomers

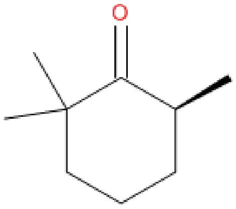
The most restricted match does not retrieve tautomers. To retrieve your query and its tautomers, but not stereoisomers, remove the BON switch and add the TAU switch:

'FRA,HYD,MAS,MET,RAD,SAL,STE,VAL,TAU,CHA,DAT'

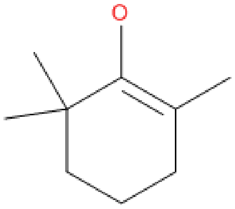
### Tetrahedral Stereochemistry

Structures that contain stereochemistry in tautomeric groups can retrieve tautomers of the structure that do not contain stereochemistry. Conversely, query structure with a tautomeric region that contains no stereochemistry matches tautomers of the structure that contain stereochemistry.

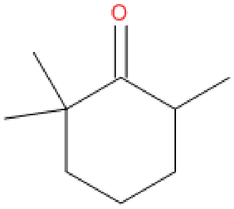
For example, consider the following four structures:



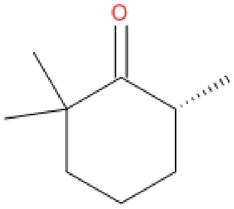
A



B



C



D

In a Tautomer search (TAU,STE):

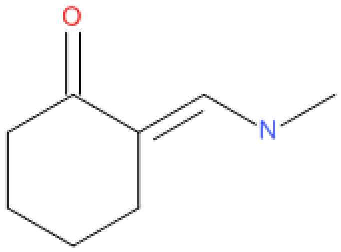
- Query structure A matches structure B but not C or D. C is the identical tautomeric form, but contains undefined stereochemistry. D does not match because D is a stereoisomer of A.
- Query structure B matches structures A, C, and D.
- Query structure C matches structure B but not A or D. B and C always match because they are tautomers. C contains undefined stereochemistry and therefore cannot match A and D.

A query structure with defined stereochemistry does not match the identical structure with undefined stereochemistry. For more information see Stereochemistry (STE,IgnoreHigherOrderStereo) on page 109.

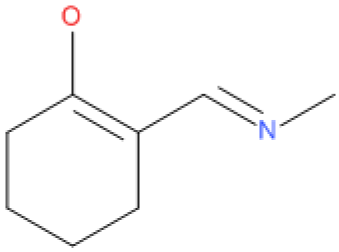
For examples of searching tetrahedral stereochemistry in structures that do not contain tautomer groups, see Flexmatch Search of Structures with Tetrahedral Stereochemistry on page 118.

### Cis/Trans Geometric Stereochemistry

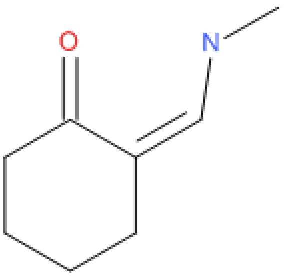
If a double bond in a tautomeric group in the query matches a single bond in the target, the configuration must match, even if the double bonds are part of a tautomer. For example:



A



B



C

In a Tautomer search FRA,HYD,MAS,MET,RAD,SAL,STE,VAL,TAU,ION,CHA,DAT or TAU,STE:

- Query structure A matches structure B but not C. C is the identical tautomeric form, but is a stereoisomer.
- Query structure B matches structures A and C.
- Query structure C matches structure B but not A.

## Exact Match/As Drawn plus Stereoisomers

To retrieve your query as drawn, that is, excluding tautomers, and stereoisomers of your query, use the most restricted switch settings with BON On, TAU Off, and STE Off:

'BON,FRA,HYD,MAS,MET,RAD,SAL,VAL,ION,BON,CHA,DAT'

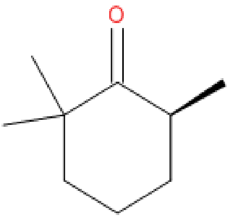
or because the ION switch is ignored when CHA is on:

'FRA,HYD,MAS,MET,RAD,SAL,VAL,BON,CHA,DAT'

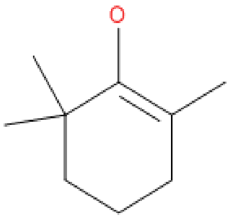
### Tetrahedral Stereochemistry

For examples of searching tetrahedral stereochemistry in structures that do not contain tautomer groups, see Flexmatch Search of Structures with Tetrahedral Stereochemistry on page 118. Structures that contain stereochemistry in tautomeric groups can retrieve stereoisomers of the structure with the same tautomer.

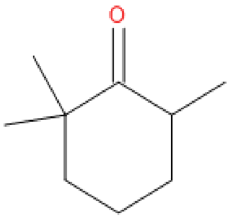
For example, consider the following four structures:



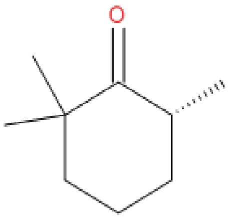
A



B



C



D

In an Isomer search, BON *On* and STE *Off*:

- Query structure A matches structure C and D, but not structure B. Structures C and D have the same tautomer but different stereochemistry. Structure B is a different tautomer.
- Query structure B does not match structure A, C, or D.
- Query structure C matches structure A and D, but not structure B.

### Cis and Trans Stereochemistry

Represent asymmetric double bonds as follows:

- If you know the exact stereoconfiguration of the double bond, draw the structure in the correct configuration (cis or trans).
- If you do not know the stereoconfiguration of the double bond, draw the structure in either the cis or trans configuration and use a Double Either bond rather than a double bond.

**IMPORTANT!** Do *not* use the Either stereo bond or a colinear arrangement of bonds to represent an unknown configuration. For more information, see Stereochemistry of Asymmetric Double Bonds on page 236.

## Original Tautomer Search

Earlier versions of BIOVIA software did not allow you to search tautomers and stereoisomers independently as shown in the previous sections, because setting the TAU switch disabled the STE switch. Consequently, you could not retrieve tautomers without also retrieving stereoisomers. This type of searching applies to:

- BIOVIA ISIS/Host, Version 5.0 and earlier
- BIOVIA Direct, Version 5.0 and earlier

If you want tautomer search to behave the same as in these earlier releases, use the following switch settings with TAU On and with BON and STE both Off:

FRA,HYD,MAS,MET,RAD,SAL,VAL,TAU,ION,CHA,DAT or because the ION switch is ignored when CHA is on: FRA,HYD,MAS,MET,RAD,SAL,VAL,TAU,CHA,DAT

## Exact Match/As Drawn plus Salts

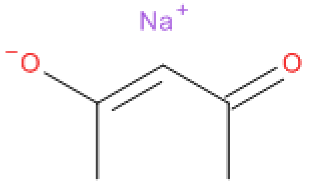
To retrieve all structures that are salts of your query, or that contain your query in combinations, such as hydrates and compounds, set the SAL switch to Off.

'BON,RAD,VAL,MET,MAS,STE,DAT'

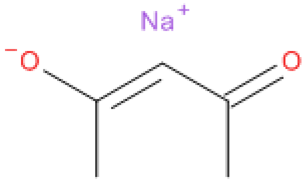
For a query that represents a salt, this combination of switches matches structures that contain:

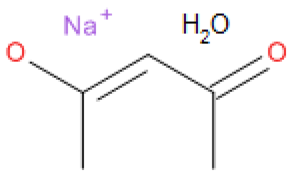
- Your query stripped of its counterions (called the parent compound)
- The parent compound with different counterions and/or additional fragments such as hydrates

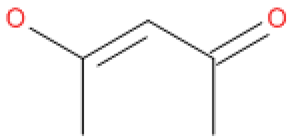
For example, a salt search with the compound sodium acetylacetonate:



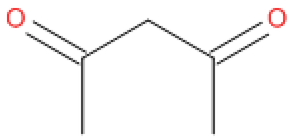
retrieves the query structure, its monohydrate, and the parent compound in its enol form:



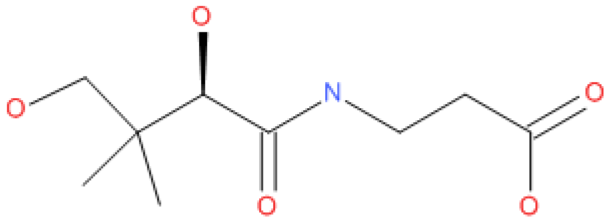




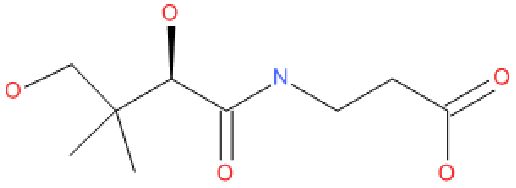
The keto form of the parent compound is not retrieved:

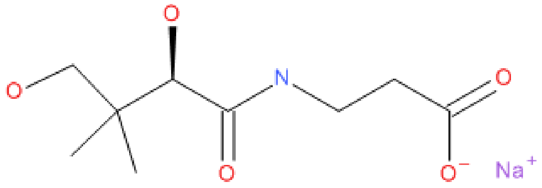


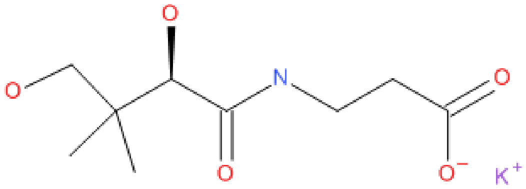
You can also use a parent compound as your query. For example, if your query is D-pantothenic acid:

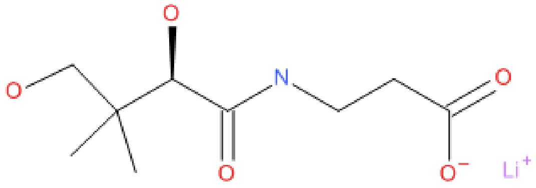


A salt search retrieves the query structure and its sodium, potassium, and lithium salts:

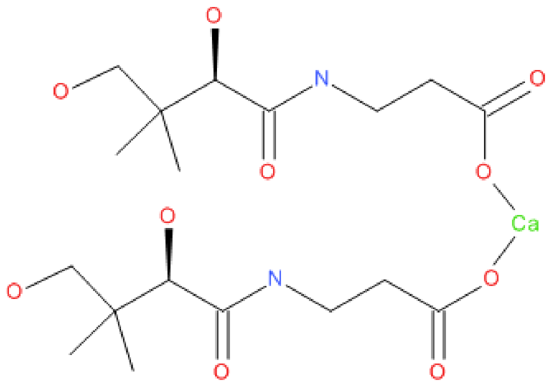




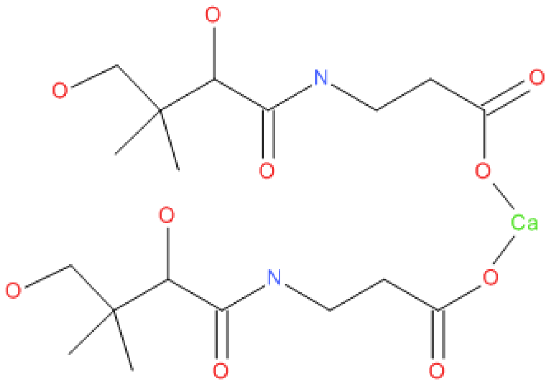




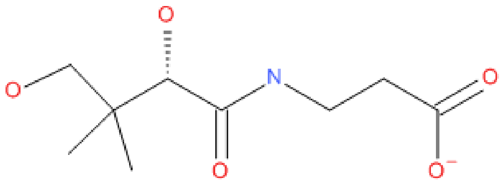
The following structures are not retrieved:



A



B



C

Structure A is not retrieved because the salt is represented with covalent bonds between the metal and other atoms, MET is *On*. Structure B is not retrieved because the stereochemistry is undefined, and Structure C is not retrieved because it is a stereoisomer, STE is *On*.

To retrieve solely the parent compound of a salt query, use the salt search, but set FRA On:

'BON,RAD,VAL,MET,MAS,STE,DAT,FRA'

## The Least Restrictive Flexmatch Switches

To retrieve the least restricted exact match of your query, set all switches *Off*. For example:

'MATCH=NONE' or 'IGNORE=ALL'

## Exact Match/As Drawn Polymer Search

To retrieve polymer structures that match your query exactly, set the MSU (monomer/SRU uniqueness) switch *O*n. For example:

'HYD,MAS,MET,RAD,SAL,STE,VAL,BON,TAU,ION,CHA,FRA,POL,TYP,MSU'

To specify that polymer end groups, mixtures, and attached data must match, add the switches END,MIX,DAT, respectively.

## Sourced-based and Structure-based Polymer Search

To retrieve polymer structures in which the source-based representation of a polymer must match the corresponding structure-based representation, use the same switches as for exact polymer search, but omit the MSU switch:

'HYD,MAS,MET,RAD,SAL,STE,VAL,BON,TAU,ION,CHA,FRA,POL,TYP'

For examples of structure-based and source-based polymers, see Polymer Representation on page 78.

## Copolymer Search

A copolymer search retrieves structures that are copolymers of your query.

To specify a copolymer search, use the same switches as for search-based and structure-based polymer search, but omit the FRA switch:

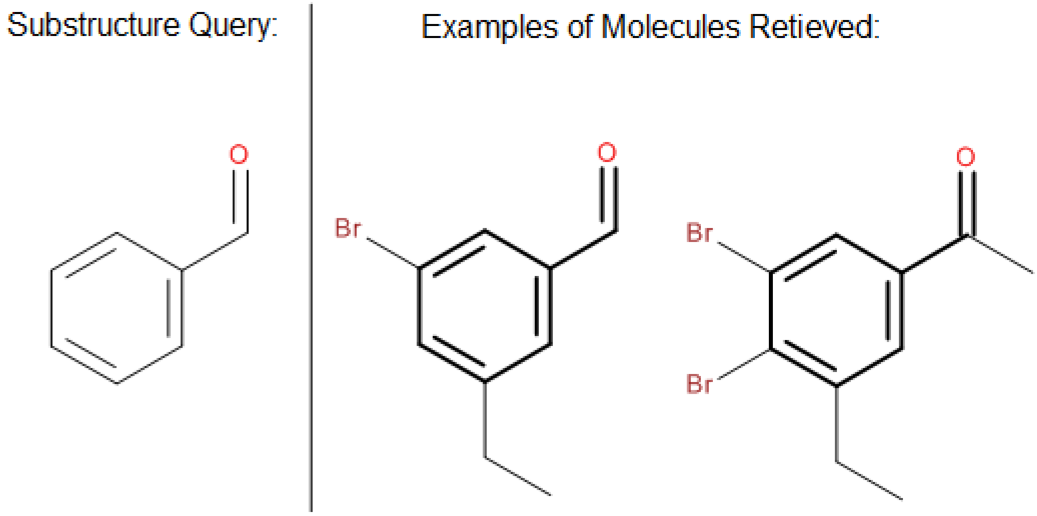
'HYD,MAS,MET,RAD,SAL,STE,VAL,BON,TAU,ION,CHA,POL,TYP'

For examples of copolymers, see Polymer Representation on page 78.

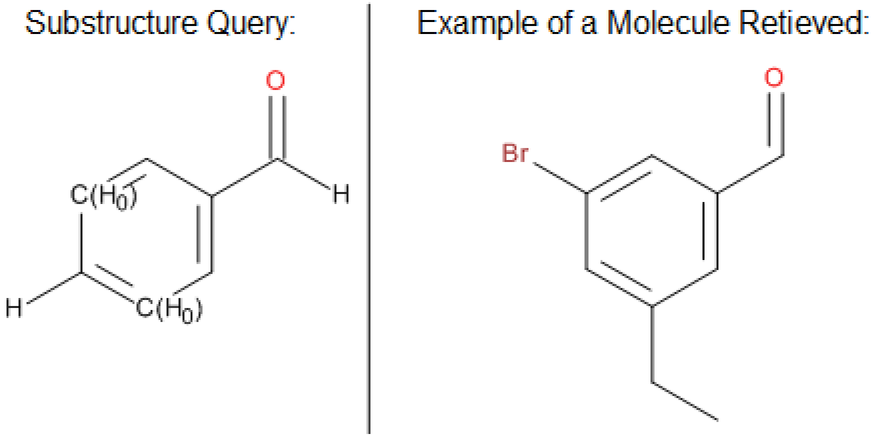
# Chapter 10:
# Substructure Search

## Definition of Substructure Search

A substructure search (SSS) finds structures that contain your query as a substructure wholly within a larger structure. Your substructure query is a two-dimensional representation of a portion of a molecule (a 2D substructure). For example:



You can increase the power of an SSS search by adding specific properties (called query features) to atoms and/or bonds. In the following example, the atom query feature (H0) prohibits the attachment of hydrogens on a specific atom, and the atom query feature Substitution as drawn (s*) prohibits the attachment of non-hydrogen atoms:



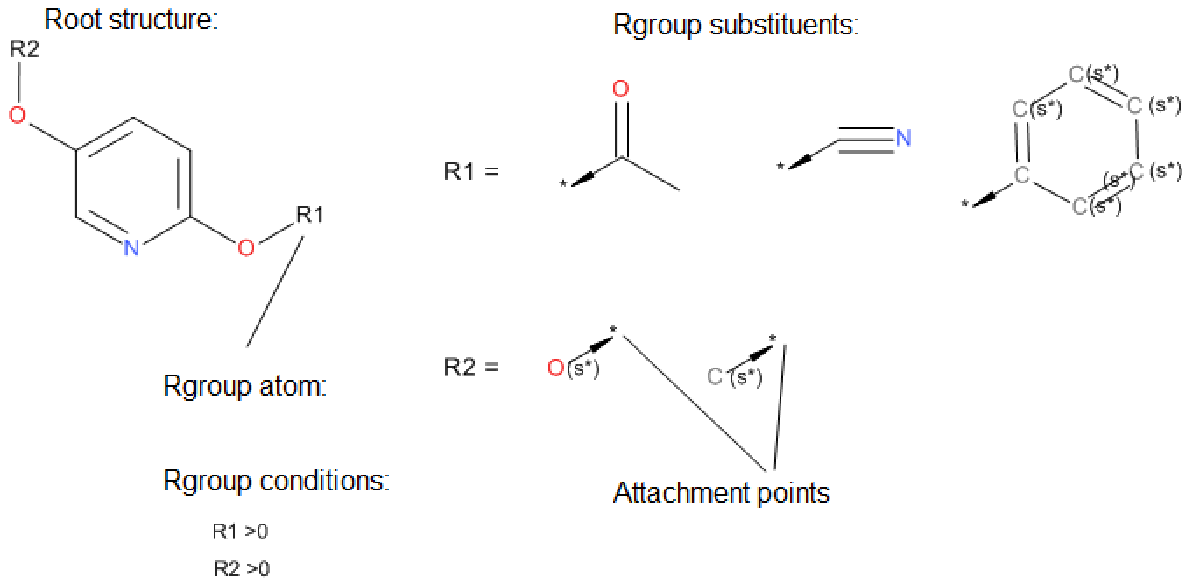
For information on query features, see Query Features on Atoms and Bonds on page 138.

You can use a Markush query to specify additional restrictions on your SSS query. For example, you can:

- Specify the structural fragments (such as functional groups or atoms) of your choice at specific sites within a molecule.
- Exclude structural fragments of your choice at specific sites within a molecule.

- Specify that the presence of one structural fragment requires the presence of another (Rgroup conditions).

An example of a Markush query is as follows:

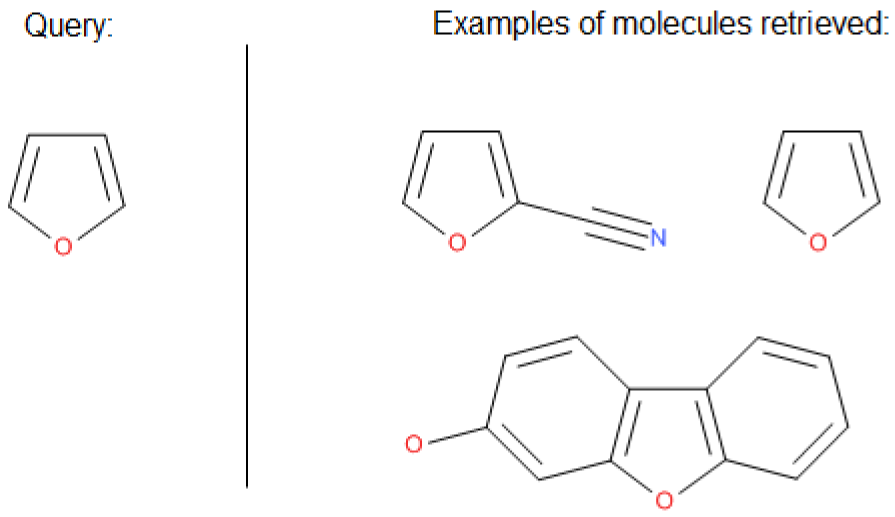


**IMPORTANT!** Do not confuse *Markush queries* with *Markush structures*. Although superficially similar to Markush queries, Markush structures have unique characteristics that do not apply to Markush queries. For example, Markush structures can be registered to databases, but Markush queries cannot. For information on these differences, see Differences Between Markush Queries and Markush Structures on page 161.

For general information on Markush queries, see Markush Search Queries on page 160.

## Query Features on Atoms and Bonds

A query feature on an atom and/or a bond in a molecule is a property that specifies the retrieval of certain types of molecule records. If you do not use query features, the molecules retrieved contain solely your exact substructure query embedded wholly within them. For example:



## Allowing or Excluding Specific Atoms

Use atom query features to allow or exclude the atoms of your choice.

### Atom Query Feature: Any Atom (A)

Specifies any atom except R, hydrogen, and hydrogen isotopes:

To disable A as a query feature use the flag `InterpretQueryAtomsLiterally`, A will then only specify A in the target.

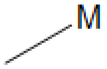
### Atom Query Feature: Heteroatoms (Q)

Specifies any atom except A, R, carbon, an atom list containing carbon, such as [C,N,O], hydrogen and hydrogen isotopes:

To disable Q as a query feature use the flag `InterpretQueryAtomsLiterally`, Q will then only specify Q in the target.

### Atom Query Feature: Metal Atom (M)

Specifies any metal atom, M, or an atom list containing only metal atoms, such as [Fe,Co]:



To disable M as a query feature use the flag `InterpretQueryAtomsLiterally`, M will then only specify M in the target.

### Atom Query Feature: Any Atom Including Hydrogen (R)

Specifies any atom including hydrogen:

To disable R as a query feature use the flag `InterpretRAtomsLiterally`, R will then only specify R in the target.

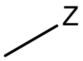
## Atom Query Feature: Halogen Atom (X)

Specifies any halogen atom (F, Cl, Br, I, At), X, or an atom list containing only halogens, such as [Cl,Br]:

To disable X as a query feature use the flag `InterpretXAtomsLiterally` or `InterpretQueryAtomsLiterally`, X will then only specify X in the target.

## Atom Query Feature: Any Atom (Z)

Specifies any atom except hydrogen, and hydrogen isotopes:



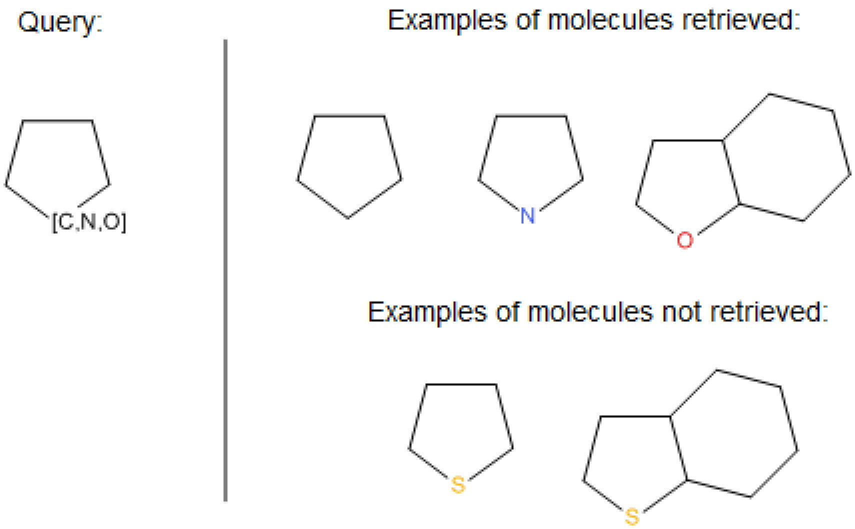
To disable Z as a query feature, use the flag `InterpretZAtomsLiterally`, Z will then only specify Z in the target.

Z is used in Pipeline Pilot to represent attachment points in Markush members. It is present in Direct for compatibility with Pipeline Pilot.

## Atom Query Feature: List

Specifies any atom on a list of your choice, or an atom list which is a subset of your atom list.

In the following example, only the atoms C, N, and O are allowed at the position specified within the list brackets:

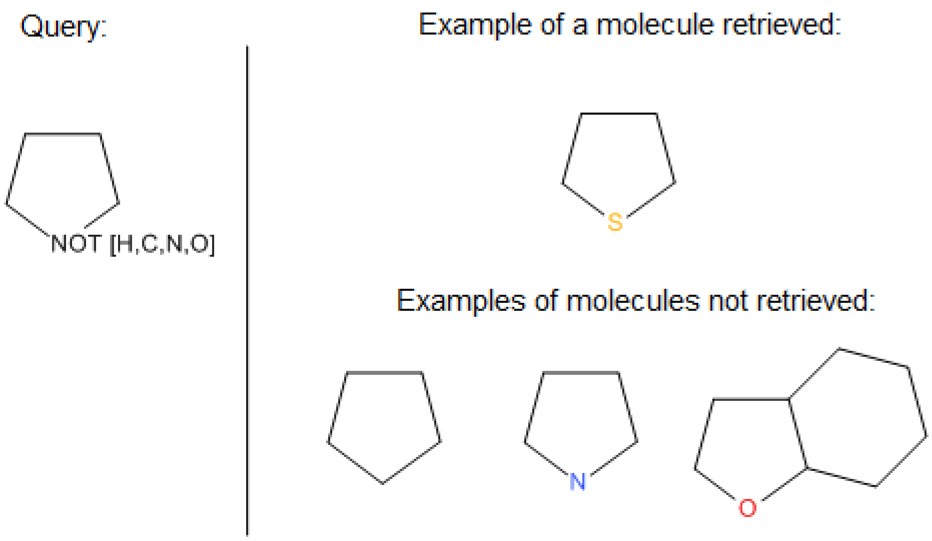

Query:
[C,N,O]
Examples of molecules retrieved:
N
O
Examples of molecules not retrieved:
S
S

## Atom Query Feature: Not List

Specifies any atom except those on a list of your choice, an atom list containing any of the atoms on your list, and hydrogen atoms.

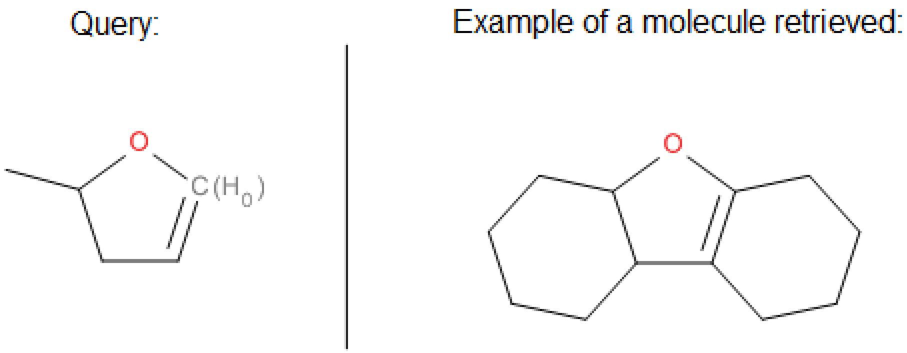
**Note:** The query NOT [C,N,O] is equivalent to NOT [H,C,N,O]; H is always excluded.

In the following example, only the atoms C, N, O, and H are not allowed at the position specified within the list brackets:


Query:
NOT [H,C,N,O]
Example of a molecule retrieved:
S
Examples of molecules not retrieved:
N
O

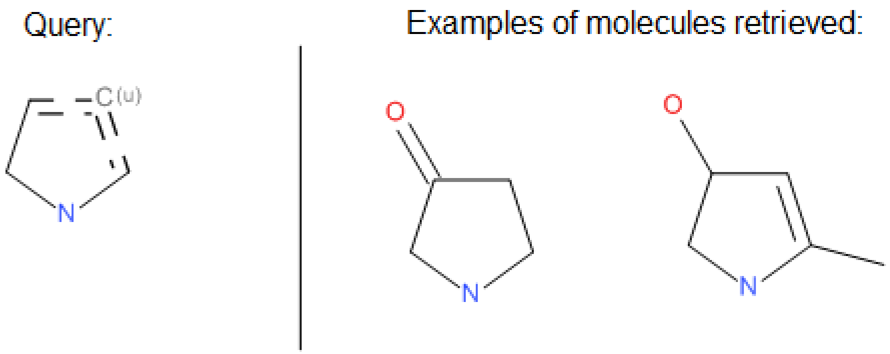
## Atom Query Feature: H0

Prohibits the attachment of hydrogens. For example:



## Atom Query Feature: Unsaturated Atom (u)

Specifies that the atom of your choice is attached to at least one multiple bond (double, triple, or aromatic). For example, the following query specifies both an unsaturated atom (u) and single or double bonds (the double dashed lines).



## Allowing a Specific Number of Attachments

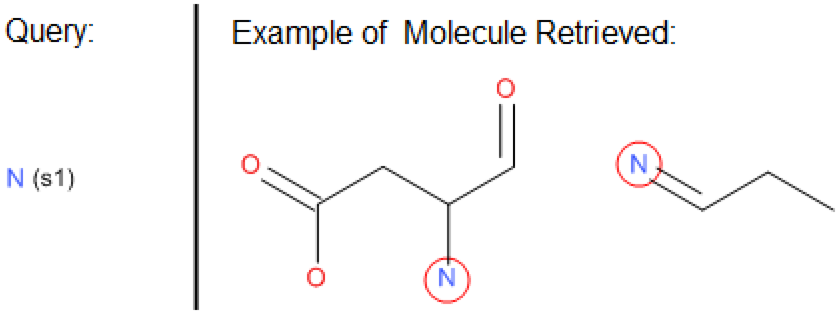
A non-hydrogen attachment is called a substituent.

### Atom Query Feature: Substitution Count of Zero (s0)

Specifies that the records retrieved have no non-hydrogen attachments at the specified position. All open valences at that position have solely implicit hydrogen atoms. This query feature implies the retrieval of a single atom with attached hydrogens (if any).

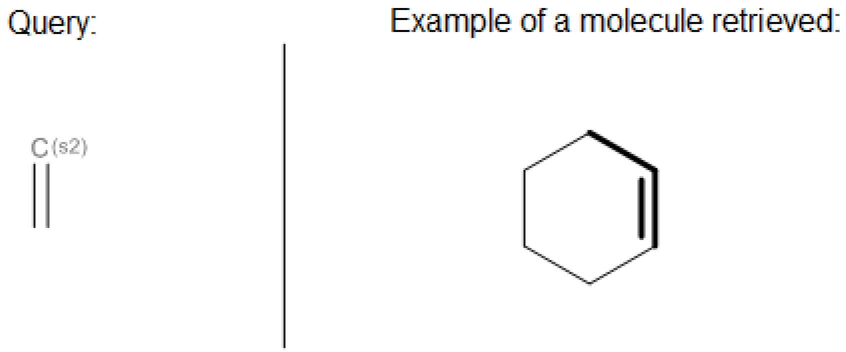
### Atom Query Feature: Substitution Count of One (s1)

Specifies that the records retrieved have solely one non-hydrogen attachment of any bond type at the specified position. For example:



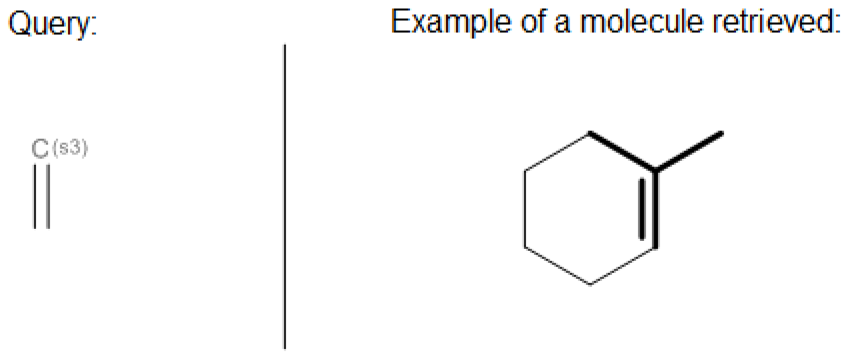
### Atom Query Feature: Substitution Count of Two (s2)

Specifies that the records retrieved have solely two non-hydrogen attachments of any bond type at the specified position. For example:



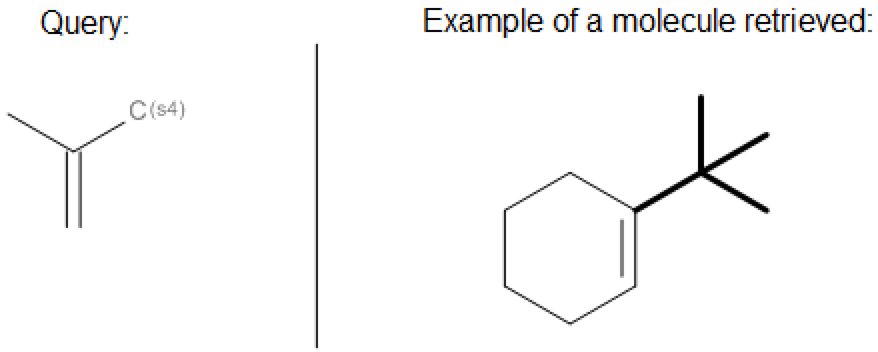
### Atom Query Feature: Substitution Count of Three (s3)

Specifies that the records retrieved have solely three non-hydrogen attachments of any bond type at the specified position. For example:



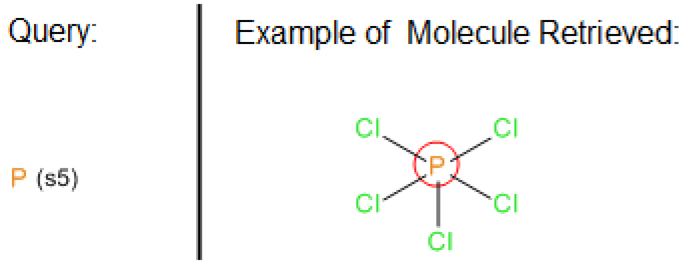
### Atom Query Feature: Substitution Count of Four (s4)

Specifies that the records retrieved have solely four non-hydrogen attachments of any bond type at the specified position. For example:



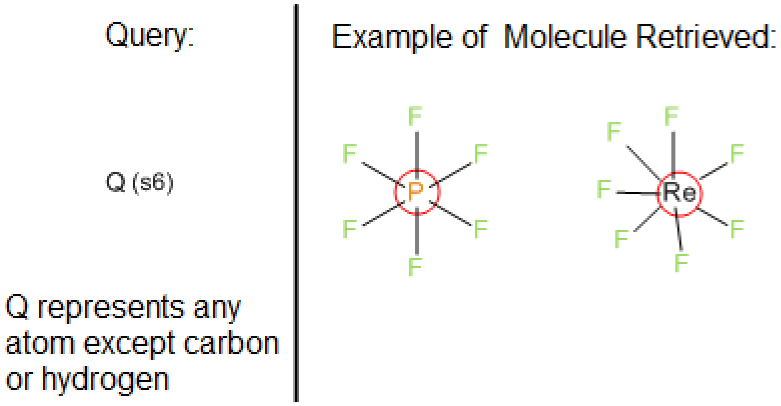
## Atom Query Feature: Substitution Count of Five (s5)

Specifies that the records retrieved have solely five non-hydrogen attachments of any bond type at the specified position. For example:



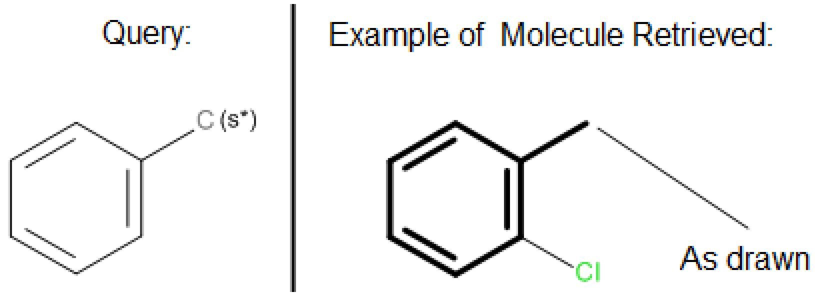
## Atom Query Feature: Substitution Count of at Least Six (s6)

Specifies that the records retrieved have at least six non-hydrogen attachments of any bond type at the specified position. For example:


Query:
Q (s6)
Q represents any atom except carbon or hydrogen
Example of Molecule Retrieved:
F
F
F
P
F
F
F
F
F
F
F
Re
F
F
F

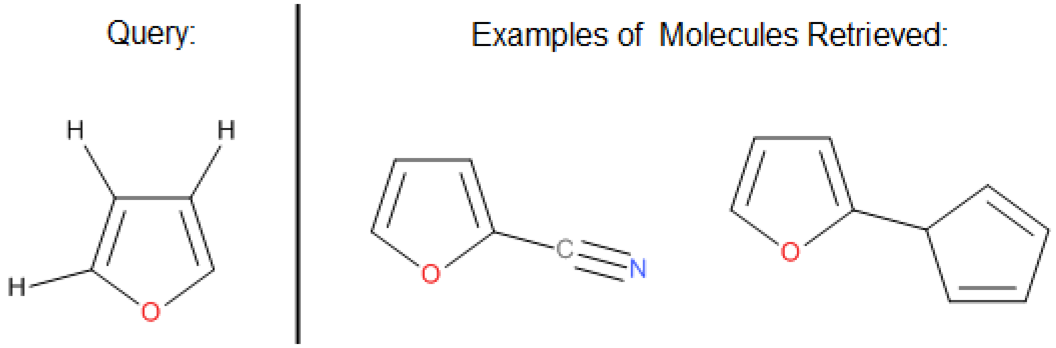
## Atom Query Feature: Substitution Count as Drawn (s*)

Specifies that the records retrieved have solely those non-hydrogen attachments that you see at the specified position. For example:



## Explicit Hydrogens

You can also use explicit hydrogens to block substitution. For example:



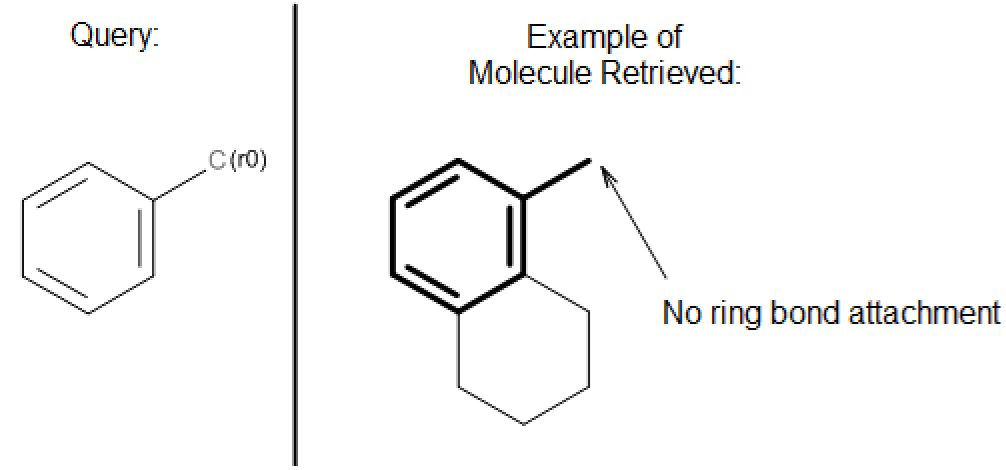
However, queries with explicit hydrogens are not as precise as queries that use substitution count. For an example of the differences, see Aliphatic Alcohols on page 184.

## Allowing a Specific Number of Ring Bond Attachments

Use atom query features to specify the number of ring bond attachments.

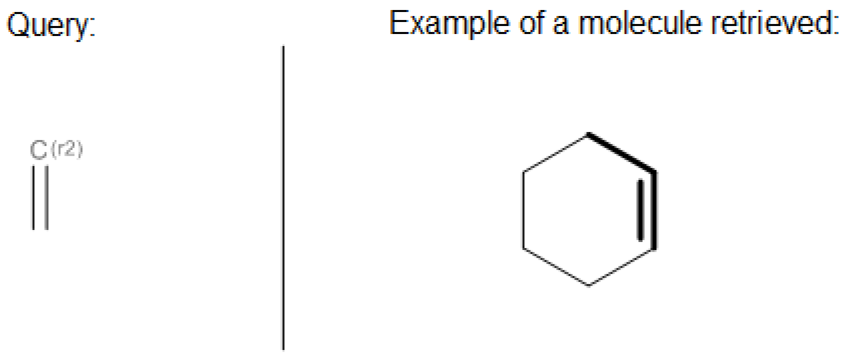
### Atom Query Feature: Ring Bond Count of Zero (r0)

Specifies that the records retrieved have no ring bond attachments at the specified position. All atoms with this mark are not part of a ring. For example:



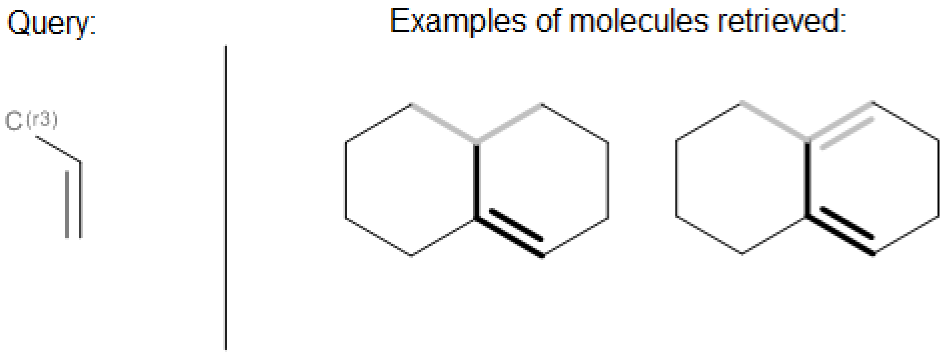
### Atom Query Feature: Ring Bond Count of Two (r2)

Specifies that the records retrieved have solely two ring bond attachments at the specified position. For example:



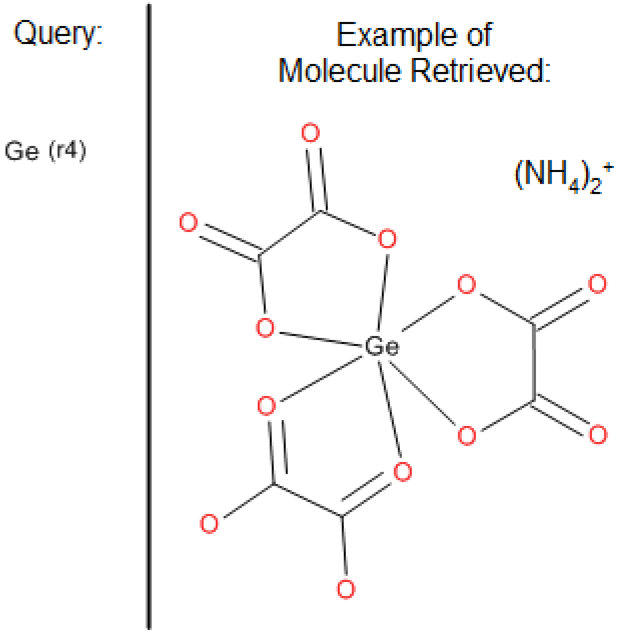
## Atom Query Feature: Ring Bond Count of Three (r3)

Specifies that the records retrieved have solely three ring bond attachments at the specified position. For example (bonds that are not drawn explicitly in the query are in gray):



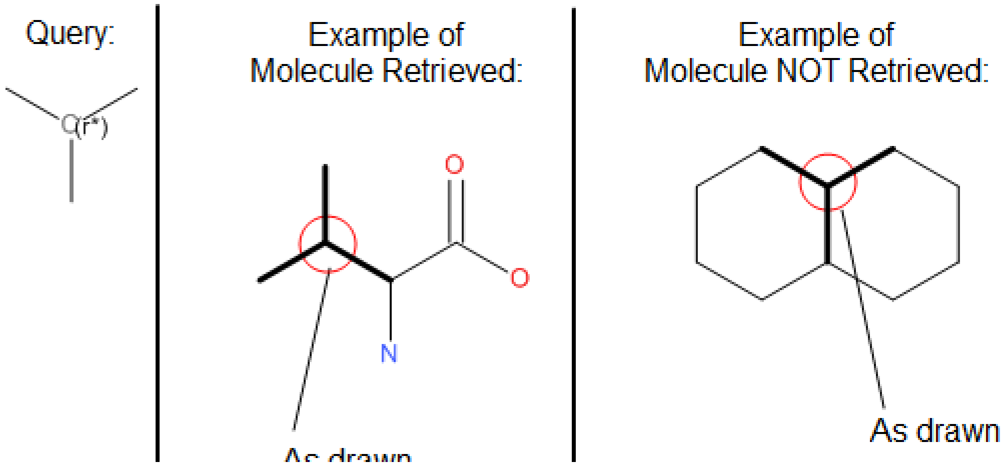
## Atom Query Feature: Ring Bond Count of at Least Four (r4)

Specifies that the records retrieved have at least four ring bond attachments at the specified position. For example:



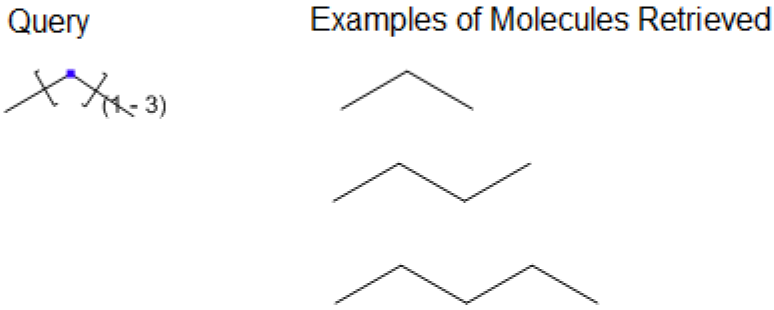
## Atom Query Feature: Ring Bond Count as Drawn (r*)

Specifies that the records retrieved have solely those ring bond attachments that you see at the specified position. For example:

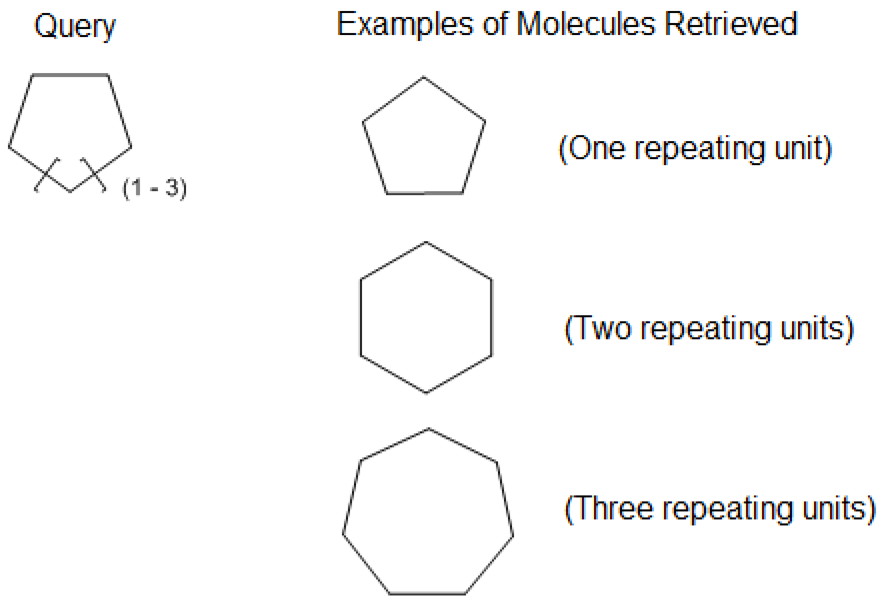

Query:
Q(r*)
Example of Molecule Retrieved:
O
O
N
As drawn
Example of Molecule NOT Retrieved:
As drawn

## Allowing Additional Atoms in a Chain or Ring (Link Node)

Use a link node to specify the addition of variable numbers repeating units (atoms) to a ring or chain. A repeating unit must be attached to two adjacent atoms. For example, you can specify between one and three repeating carbon atoms in an existing 5-membered ring with the link node [L1-3]:



Alternatively, you can specify between one and three repeating carbon atoms in an existing 3-carbon chain with the link node [L1-3]:



A link node is a special type of chemical Sgroup.

To summarize, a link node is characterized as:

- A single atom, connected to exactly two non-repeating atoms.
- Optional side chains, that repeat together with the link atom.
- A repeat count from 1 to a fixed upper limit.

See also Variable Repeat Group on page 148 and Multiple Groups on page 30.

## Variable Repeat Group

A variable repeat group is a query feature (non-registerable) that:

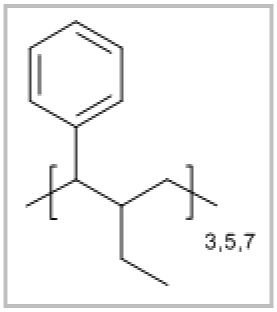
- Is a group of one or more atoms connected to exactly two non-repeating atoms.
- Has a repeat count that is controlled by a specified set of occurrences, including optional ranges, such as:

  1-3, 7, 11, 13-15

### Comparison with link node and multiple group

A variable repeat group is:

- Similar to a link node but with the capability to support a *group* of atoms
- Also a superset of a multiple group because it adds the option of having a basis that repeats in a flexible manner. The repeat count can be a single, fixed repeat like a multiple group, or a set of repetitions.



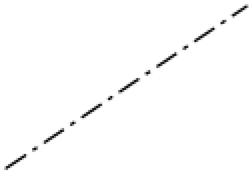
For more information, see Multiple Groups on page 30 and Allowing Additional Atoms in a Chain or Ring (Link Node) on page 147.

## Allowing Multiple Bond Types

Use bond query features to specify more than one bond type.

### Bond Query Feature: Any bond

Specifies any bond type: single, double, triple, or aromatic, at that position:

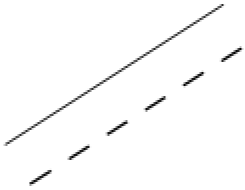


### Hydrogen Query is a Special Case

Explicit hydrogens that have an 'any' query bond type are interpreted as a search for special hydrogen environments, such as an explicit hydrogen bonded to a non-single bond type, or having more than one bonded neighbor of any registerable bond type. Such queries do not match single-bonded hydrogens. To get hits on single-bonded hydrogens, either remove explicit hydrogens from the queries or ensure that such hydrogens retain their original bond type.

### Bond Query Feature: Aromatic Bond

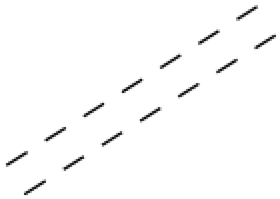
Specifies solely an aromatic bond at that position:



An aromatic query bond hits any bond in all types of aromatic rings: 6-membered ring bonds, 5-membered ring bonds, and so on. For more information on the structures that are interpreted as aromatic, see How Compounds are Perceived as Aromatic on page 152.

### Bond Query Feature: Single/Double Bond

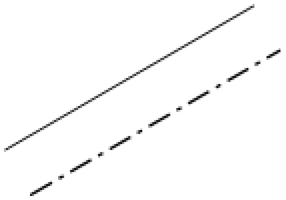
Specifies either a single or a double bond at that position:



A Single/Double query bond does not hit bonds in 6-membered aromatic rings, or 5-membered rings. For more information, see How Compounds are Perceived as Aromatic on page 152.

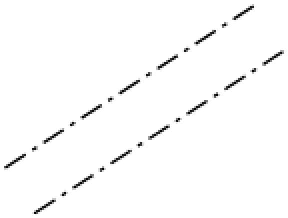
### Bond Query Feature: Single/Aromatic Bond

Specifies either a single or an aromatic bond at that position:



### Bond Query Feature: Double/Aromatic Bond

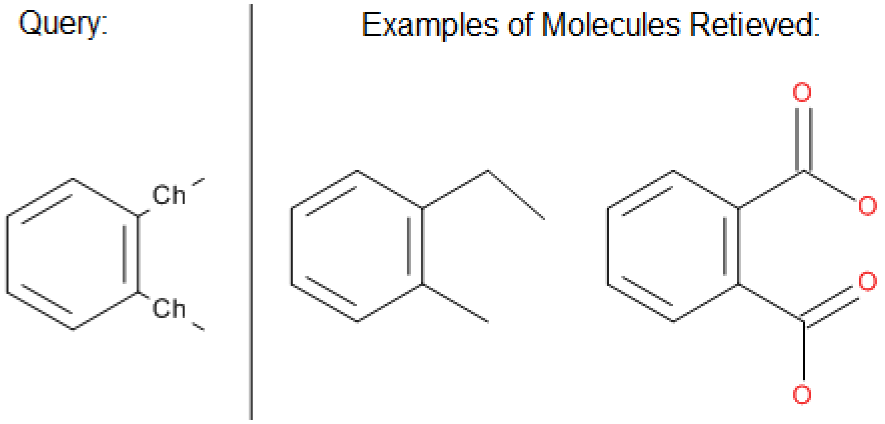
Specifies either a double or an aromatic bond at that position:



## Bond Topology

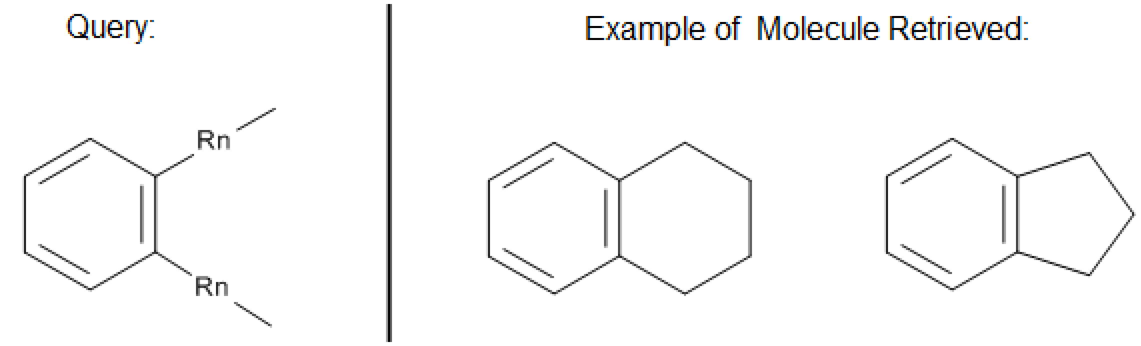
### Bond Query Feature: Ch

Specifies the bond is part of an acyclic structure. For example:



## Bond Query Feature: Rn

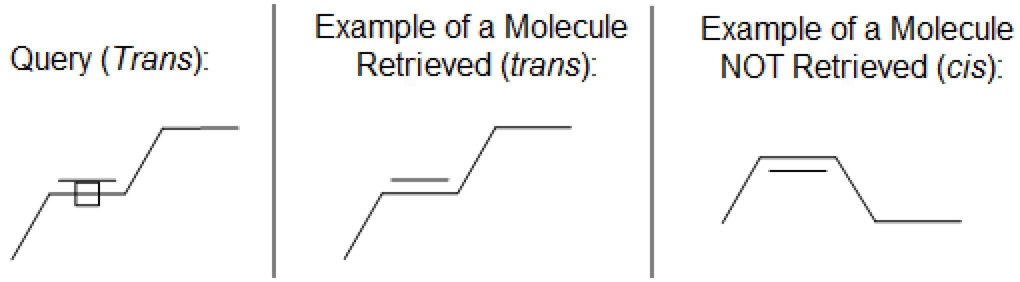
Specifies that the bond is part of a cyclic structure. For example:



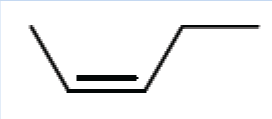
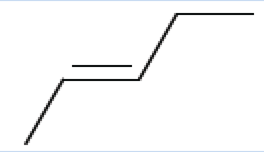
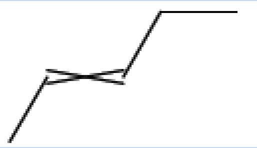
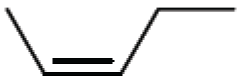
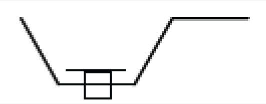
## Bond Query Feature: Cis/Trans Geometric Double Bonds as Drawn

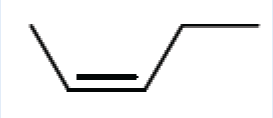
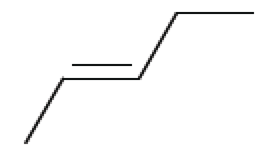
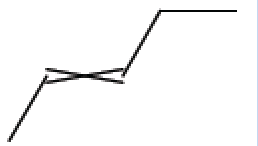
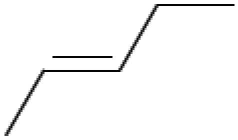
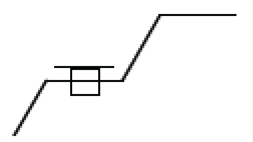
**Note:** ISIS/Draw uses the term "Stereo Box" for this query feature.

Adding this query feature to asymmetric double bonds allows you to retrieve molecules with matching cis or trans stereochemistry. For example:


Query (*Trans*):
Example of a Molecule Retrieved (*trans*):
Example of a Molecule NOT Retrieved (*cis*):

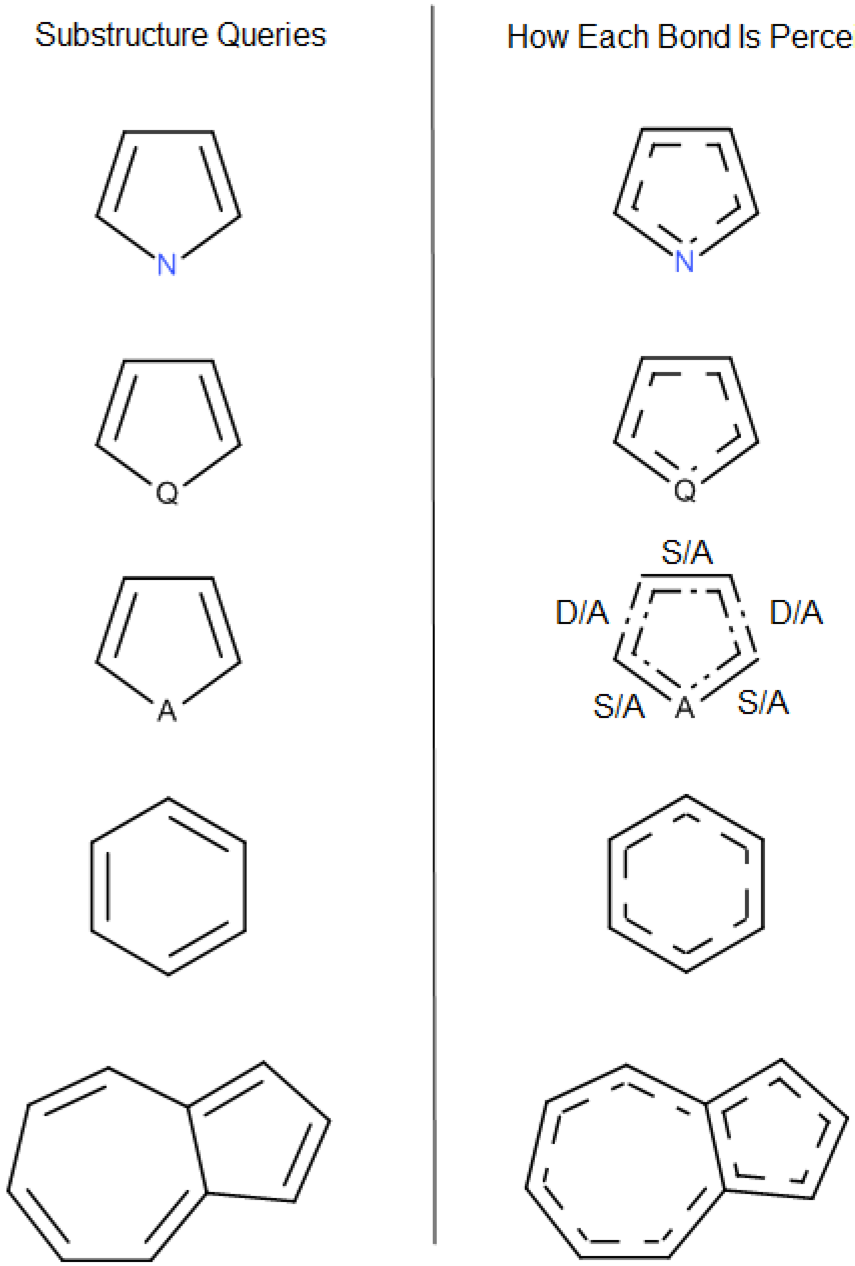
The table that follows summarizes the effect of the stereo box in a substructure search:

| | Structures in Database | | |
|---|---|---|---|
| Queries | Cis Isomer | Trans Isomer | Either Isomer |
| |  |  |  |
|  | ✓ | ✓ | ✓ |
|  | ✓ | | |

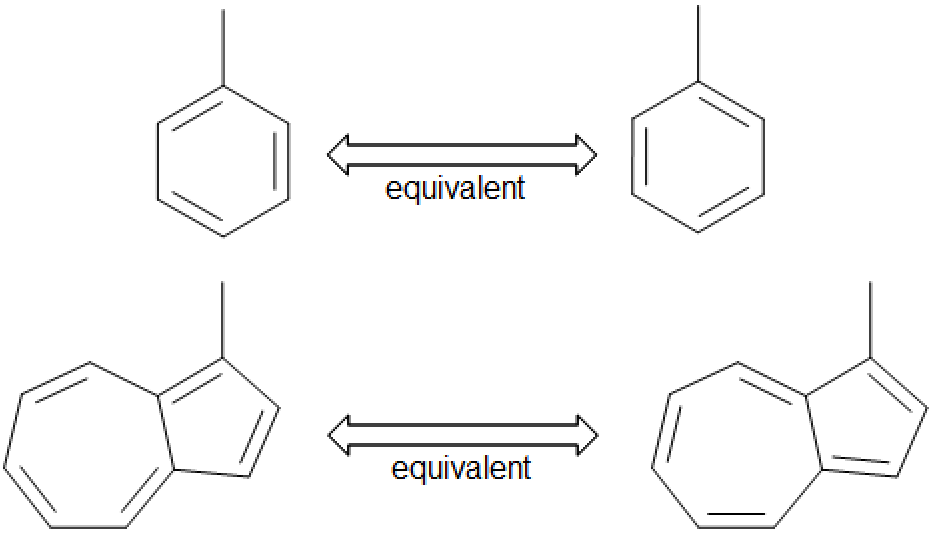
| | Structures in Database | | |
|---|---|---|---|
| Queries | Cis Isomer<br> | Trans Isomer<br> | Either Isomer<br> |
|  | ✓ | ✓ | ✓ |
|  | | ✓ | |

## How Compounds are Perceived as Aromatic

The following aromatic substructure queries specify the retrieval of aromatic molecules. Aromaticity is determined using the 4N+2 electron rule, with the query atom Q (any atom except hydrogen or carbon) treated as a heteroatom with a lone pair. All of these queries except for the query with A (any atom except hydrogen) contain only aromatic bonds. The query with A is treated as if it contains single or aromatic (S/A) and double or aromatic (D/A) query bond types so that it will match a cyclopentadiene target:



The bonds of aromatic ring systems are stored in the database with additional information that allows alternative representations of the same structure to be perceived as equivalent, for example:



# Attached Data

## Introduction

Attached data (also called Sgroup data) is numeric or text data that you can associate with all or part of a structure. Attached data is normally chemically significant, because it is stored directly with the structure, not in a separate database field. Consequently, attached data can be stored in the structure field of a database and/or used in a graphical search query. Using an Sgroup field name which starts with NOSEARCH: will prevent the attached data from being chemically significant, see Reserved Names for Sgroup Fields on page 156.

You can attach data to an atom, a bond, a fragment, or to any collection of atoms and bonds.

## Sgroup Fields

Each piece of attached data is associated with a special kind of database field called an Sgroup field. The Sgroup field defines what the data means. For example, text data that is associated with a particular Sgroup field is not equivalent to identical text data that is associated with a different Sgroup field.

You cannot associate a form box with an Sgroup field. To search for data in an Sgroup field, you must use a graphical search query. For more information on creating search queries for attached data, see Substructure Search of Attached Data on page 166.

### Required Sgroup Fields

Attached data must be used to represent the following structure types:

- Biopolymers which are represented using a condensed form with * atoms or pseudoatoms
- Non-biological polymers

### Sgroup Fields for Condensed Representation of Biopolymers

BIOVIA uses the following reserved Sgroup field names for representation of biopolymer residues in condensed form (*atoms and/or pseudoatoms).

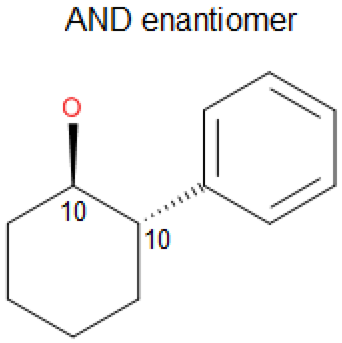
**Note:** These Sgroup field names are not required or used when representing biopolymers using the SCSR (Self Contained Sequence Representation) representation and global templates. The fields are only used with legacy representations using * atoms or pseudoatoms.

| Sgroup Field Name | Purpose of Attached Data | Characteristics of Data |
|---|---|---|
| MDL_STARATOM_NAME | Distinguishes different types of residues or single-attachment groups that use the *atom representation.<br>See Sgroup Field for *Atom Representation on page 62. | Variable text |
| MDL_RESIDUE_ ATTACHMENT_ORDER | Stores information on non-terminal attachment atoms for residues that are represented in condensed form (*atom or pseudoatom).<br>See Sgroup Field for Identifying Attachment Atoms on page 61. | Fixed text: 2 characters |

### Sgroup Field for Stereochemical Purity at Stereogenic Centers

Marking a stereogenic center with the AND enantiomer chiral label does not necessarily imply that the two stereoisomers are present in an exact 50:50 mixture.

If you want to display quantitative information on the amounts of stereoisomers, you might want to specify it as attached data at the stereogenic atom. On the following structure, the AND enantiomer chiral label indicates a mixture of two stereoisomers, and the attached data represents a 10% enantiomeric excess of the structure as drawn:



The attached data differentiates this structure from others of identical chemical structure but different stereochemical purity.

The Stereochemistry dialog in BIOVIA Draw has a control for stereochemical purity that is hidden by default. The attached data is associated with the reserved Sgroup field `MDL_PURITY`.

This example uses enantiomeric excess as the measurement of stereochemical purity. BIOVIA has no specific recommendations for indicating stereochemical purity.

### Sgroup Fields for Polymer Representation

BIOVIA uses the following reserved Sgroup field names for representation of polymers:

| Sgroup Field Name | Meaning of Attached Data | Characteristics of Data |
|---|---|---|
| POLYMER_TYPE | Stores information on different types of structural repeating units.<br>See Polymer or Copolymer Type on page 87. | Variable text Allowed values: stat ran blk xl ran grf mod |
| POLYMER_STEREO | Stores information on stereoregularity in a polymer repeating unit.<br>See Stereoregularity in Polymers on page 87. | Fixed text: 2 characters |

### Reserved Names for Sgroup Fields

Certain Sgroup field names are reserved for use by BIOVIA. If you want to define additional categories of attached data for structure representation, avoid the following names for Sgroup fields:

- All Sgroup field names that begin with the characters "MDL_"
- All Sgroup field names that begin with the characters "SMMX:"
- The names MEMB_NAME and MEMB_TAG1
- The names POLYMER_TYPE and POLYMER_STEREO
- The name ATROP_STE

Use the prefix "NOSEARCH:" in Data Sgroup field names to ignore in BIOVIA Direct substructure or exact-match searches. The prefix is not case sensitive. For example, a data Sgroup field named NMRShift would have its values compared between query and target when searching, while a field named nosearch:NMRShift would be ignored during searching.

## Guidelines for Defining Your Own Attached Data

In addition to the types of attached data that are described in the previous section, you might want to incorporate additional categories of attached data into your business rules for chemical substance representation. Use the following guidelines to decide whether to use attached data.

- If the information that you want to store applies to an entire structure or substance, store the information in a separate database field.
- Use attached data for any information that applies to a portion of a structure, and that you want to use to differentiate structures. For more information, see Examples of Attached Data on page 156.

## Examples of Attached Data

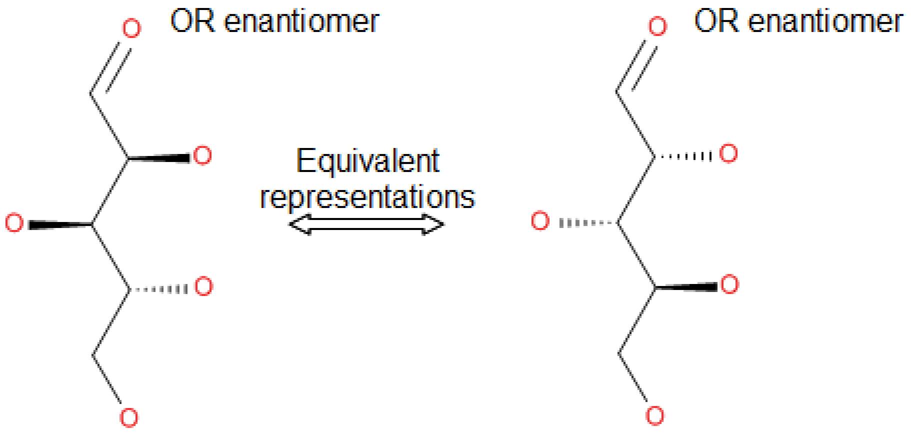
This section contains examples that illustrate how attached data can be used to differentiate chemical structures.

For details on the searching behavior of attached data, see:

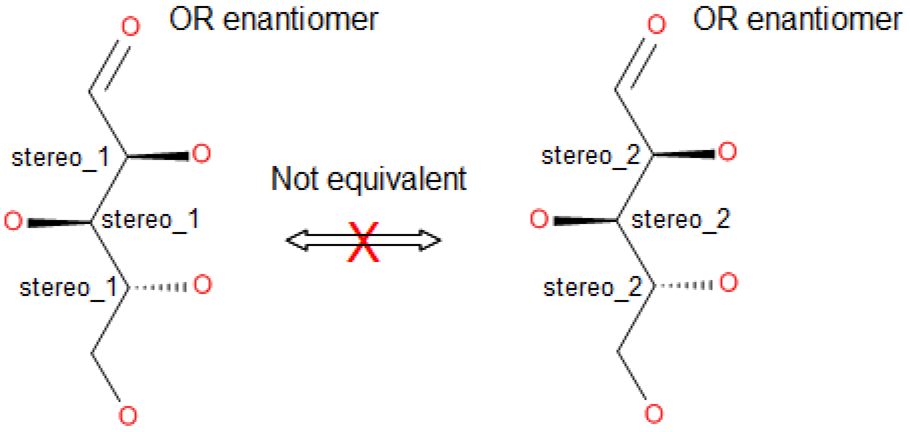
- Exact Search (Flexmatch) on page 107
- Substructure Search on page 137

### Distinguishing Relative Stereoisomers

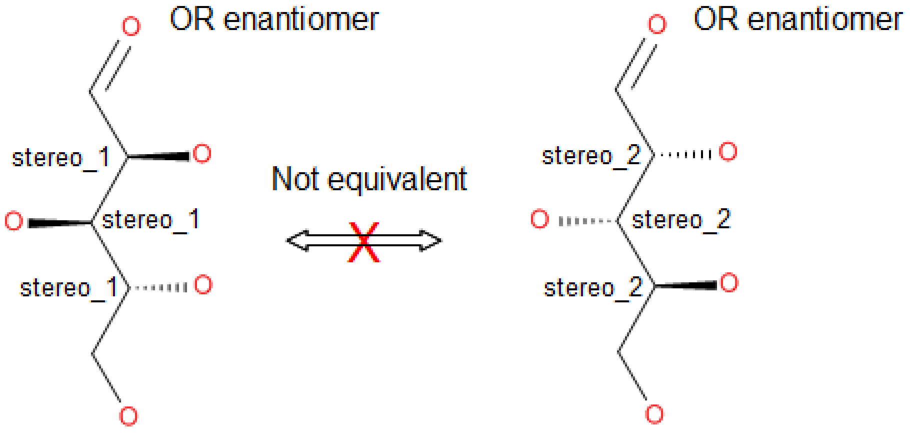
An OR stereogroup is a set of stereogenic centers in which you know the relative configuration of all the stereogenic centers in that group. Only one enantiomer is present, but you do not know which one. When you draw the structure, it does not matter which of the two enantiomers you draw, because the chiral label OR enantiomer makes the two structures equivalent:

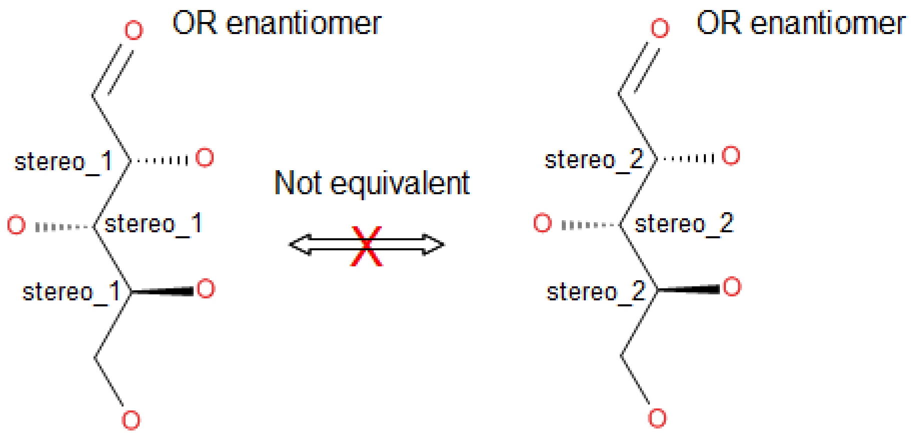


The two structures are equivalent, so you cannot register them as distinct structures. You might want to register them as distinct structures, however, if you have isolated both enantiomers, but do not know the absolute configuration of either one. To register the enantiomers as distinct structures, use attached data to mark the structures. In the above example, you could use the text "`Stereo_1`" to mark one enantiomer, and the text "`Stereo_2`" to mark the second enantiomer:



The attached data ensure that the structures are perceived as distinct, although the stereo bonds and stereogroups are identical. The figures that follow show additional examples of structures that are equivalent if the attached data is removed, but not equivalent otherwise:



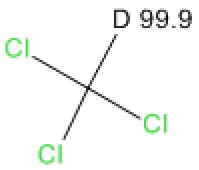


For more information on stereochemical representation, see Tetrahedral Stereochemistry on page 9.

**IMPORTANT!** Prior to the introduction of enhanced stereochemical representation, the recommended convention for specifying related groups of stereogenic centers was to mark the centers with attached data. This convention is no longer recommended, however, because of the advantages of the enhanced stereochemical representation. For more information on the advantages of stereogroups over attached data, see Attached Data Marks Groups of Related Stereogenic Centers on page 259.

### Isotopic Purity

In the following example, the data that is attached to the deuterium atom specifies the percentage of deuteration at that site:



The attached data differentiates this structure of deuterochloroform from others with the same chemical structure but different amounts of deuteration.

## Creating Sgroup Fields in Your Database

Use the following steps to define Sgroup fields in your database:

1. Use the guidelines in the preceding sections to decide the categories of attached data that you want in your database. Each category of attached data should have an associated Sgroup field.
2. Use Oracle SQL*Plus to load an Sgroup fields definition file into the global chemical environment. New databases will use these Sgroup fields, to install updated Sgroup fields into existing databases use the SQL*Plus command `ALTER INDEX` *`indexname`* `REBUILD PARAMETERS ('ENVIRONMENT')`. For more information, see the chapter, "Setting the Chemical Environment" in *BIOVIA Direct Administration Guide*.
3. If you have defined your own categories of attached data according to your company's business rules for chemical substance representation, you must customize BIOVIA Draw to enable your users to create and edit attached data. See the section that follows.

## Data Sgroup Queries

Data Sgroup queries are created in BIOVIA Draw using the *Attach Data* option.

To create a data Sgroup query:

1. Right-click on the atom. Select *Attach Data...*.
2. Fill in the *Field Description* with the name of the Data Sgroup Field.

   **Note:** The name is not case sensitive.

3. Use the *Search Operator* menu to select one of the following query types:
   - >
   - >=
   - <
   - <=
   - <>
   - between
   - contains
   - like
4. Fill in the *Data* field with the query. The query may be:
   - A single number
   - A range of numbers separated by a dash
   - Text, including % and _ wildcard characters when used with the LIKE operator

NOTES:

- If the query value begins with a single (') or double (") quote character, enclose the entire value with the other type of quote. For example the value 'x23 must be typed in as "'x23".
- Numeric queries are used with the relational operators > (greater than), >= (greater than or equal to), < (less than), <= (less than or equal to), <> (not equal to) and between (between two values inclusive).

  For example, a query with operator > and Data value 1.0 will match targets which contain data Sgroup values greater than 1.0.

  To allow for round-off, numeric comparisons allow plus or minus 5.0 in the seventh digit depending on the operator. In the example above this means that the data Sgroup value must be greater than 1.000005 in order to match the query.
- Text queries may be used with any of the relational operators and with LIKE (wild card search) and CONTAINS (text is contained in). Text queries are normally case-dependent. A query Abs will only match Abs, not ABS or abs - but this may be changed in the Sgroup definition file in BIOVIA Direct.
- The LIKE operator is the same as in Oracle. Include the % character in your query to match zero or more characters, include the _ character in your query to match any single character. For example, a query with operator LIKE and Data value %abs will match both "abs" and "relabs", but will not match "absrel".
- The CONTAINS operator will match data Sgroup vales which contain the query text. It is the same as using LIKE with a Data value which has a % character at the front and back of the text.
- Text queries using the relational operators compare characters based on their ASCII values, thus "A" is lower than "a".

BIOVIA Direct decides whether to perform a numeric search or a text search by performing the following checks. The first check that succeeds determines the type:

1. If the query operator is LIKE or CONTAINS a text search is done.
2. If the query Data value is enclosed in double (") or single (') quote characters the quotes are removed and a text match is done.
3. If there is an SGROUPFIELDS definition file stored in the domain index's environment, the specific field type (text or numeric) contained in that definition is used.
4. If the molfile or rxnfile containing the data Sgroup field has an 'N' or 'C' FIELDTYPE flag a numeric search is done. If the file has a 'T' or 'F' flag a text match is done.
5. If the query can be parsed into either a single floating point number or a range of two floating point number separated by a dash (-) character a numeric match is done. (If a text match is desired you must enclose the value in single or double quotes.)
6. A text match is done.

The SGROUPFIELDS definition file stored in the environment may be empty, in which case any data Sgroup field name is allowed and the numeric/search processing ignores step 3, or it may contain a list of allowed field names, types and whether or not a text field search is case-dependent or not.

BIOVIA recommends using an SGROUPFIELDS definition file to control the allowed data Sgroup fields, to ensure that searches use the appropriate numeric or text matching, and to allow for case-independent text searching.

See the *BIOVIA Direct Administration Guide* section on Setting the Direct C$DIRECT2021 Environment for more information.

### Related Documentation for BIOVIA Draw

See the BIOVIA Draw Configuration Guide topics on:

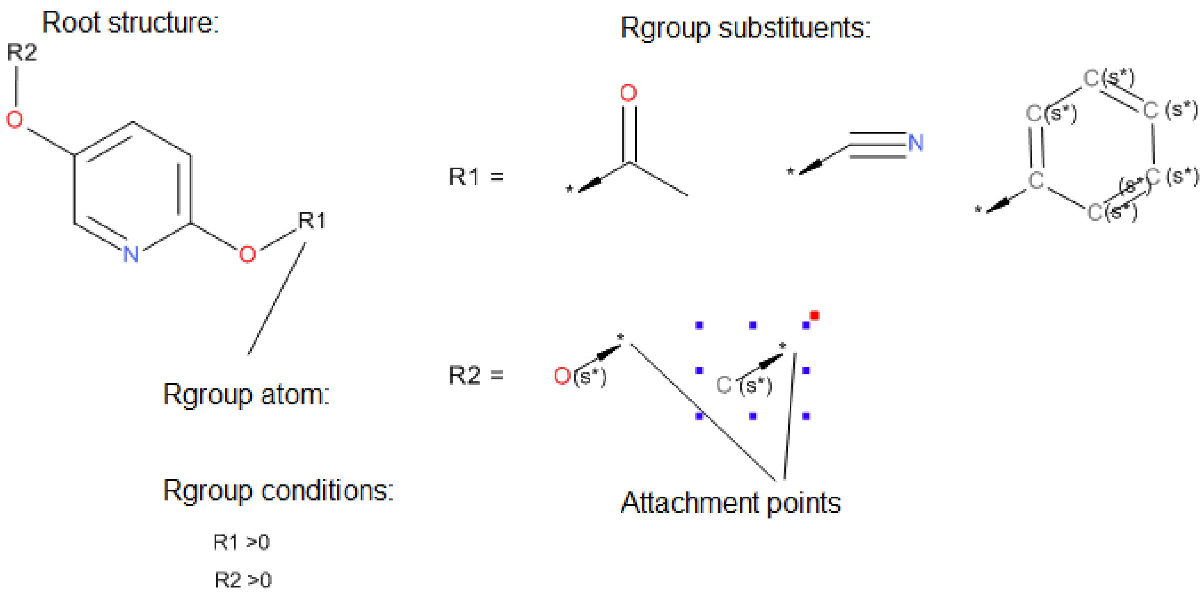
- Enabling the Data Sgroup Tool to Control Sgroup Data
- Customizing BIOVIA Draw for Registration and Searching of Biopolymers

## Markush Search Queries

A *Markush query* finds structures that contain your query as a substructure wholly within a larger structure with additional restrictions. For example, you can:

- Specify the structural fragments (such as functional groups or atoms) of your choice at specific sites within a molecule.
- Exclude structural fragments of your choice at specific sites within a molecule.
- Specify that the presence of one structural fragment requires the presence of another (Rgroup conditions).
- You can also increase the power of an Markush query by adding specific restrictions (called query features) on atoms and/or bonds.

The following figure shows an example of a Markush query that matches 2,5-disubstituted pyridines in which the substituent at position 2 can be COOH, CN, or a phenyl group (Ph), and the substituent at position 5 can be OH or CH3:



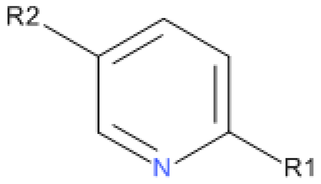
## Differences Between Markush Queries and Markush Structures

Both Markush structures and Markush queries contain a common structure with points of variation that are represented by numbered Rgroup atoms (R1, R2, R3, and so on). However, these similarities are superficial. The two types of Rgroup structures are used for quite different tasks and differ from one another in many structural details. To summarize the differences:

- Markush structures can be registered to databases, but Markush queries cannot. A Markush query can be used solely in a substructure search.
- A Markush structure can contain Rgroup atoms on adjacent bond junctions (for example: R1-R2-R3), but a Markush query cannot. See The Root Structure on page 41.
- The root of a Markush structure can contain unconnected Rgroup atoms, but a Markush query cannot. See Unconnected Rgroup Atom on page 45.
- A Markush structure can contain Rgroups that are nested to any number of levels, but a Markush query can contain only one level of nesting. For examples of Markush structures, see Nested Rgroups on page 42. For examples of Markush queries, see Nested Rgroup Substituents on page 164.
- A Markush structure can contain only one instance of each Rgroup atom, but a Markush query can contain any number of instances of an Rgroup atom. See One Rgroup at Multiple Positions on page 162.
- A Markush structure can contain only one Rgroup atom at each atom site, but a Markush query can contain multiple Rgroup atoms at a single atom site. See Multiple Rgroup Atoms at One Position on page 162.
- A Markush structure can contain null Rgroup members, but a Markush query cannot. See Null Members on page 46.
- A Markush query can contain Rgroup conditions (Occurrence, RestH, and/or If-Then), but a Markush structure cannot. See Rgroup Conditions on page 165.

## The Root Structure

The root structure contains the portion of the structure that does not vary among the structures retrieved. In the example in the previous section, the root structure is pyridine with Rgroups R1 and R2.

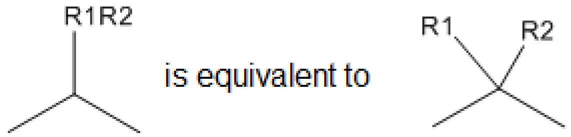


## Rgroup Atoms

On the root structure, Rgroup atoms (R1 through R32) are placeholders that designate where the lists of functional groups and atoms that can attach to the root structure. In the example above, the Rgroup R1 designates that the members of the list R1 (the Rgroup substituents) are allowed at position 5 on the root structure. The Rgroup R2 designates that the members of the list R2 (the Rgroup substituents) are allowed at position 2 on the root structure.

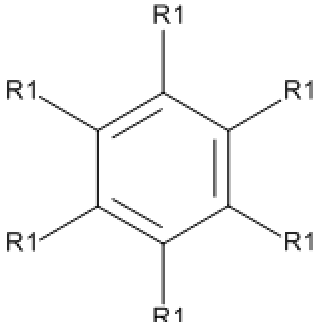
## Multiple Rgroup Atoms at One Position

You can specify multiple Rgroups at a single site on the root structure. For example:



## One Rgroup at Multiple Positions

You can specify a single Rgroup (such as R1) at multiple positions on the root structure. For example:



To restrict the number of Rgroups that must be occupied, see Rgroup Conditions on page 165. A molecule might not need to satisfy all the Rgroup sites on the root structure to be retrieved. For example, the Rgroup condition R1 > 0 specifies that any molecule in which at least one of the R1 sites is filled by an R1 member will be retrieved. In addition, the Rgroup condition RestH determines whether the unsatisfied sites must be filled with hydrogen atoms or if other molecule fragments are allowed to be attached if they are not on the list.

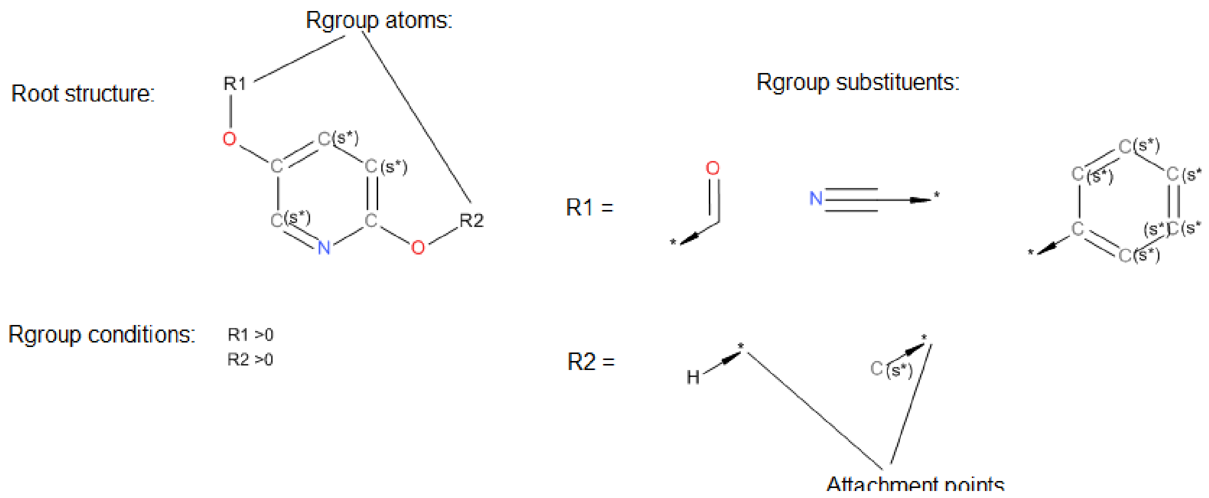
## Multiple Fragments in the Root Structure of a Markush Query

If the root structure of a Markush query contains multiple fragments, you must group the root structures before you do a search. To group multiple fragments in BIOVIA Draw, select them, and then choose **Object > Group**.

## Rgroup Substituents

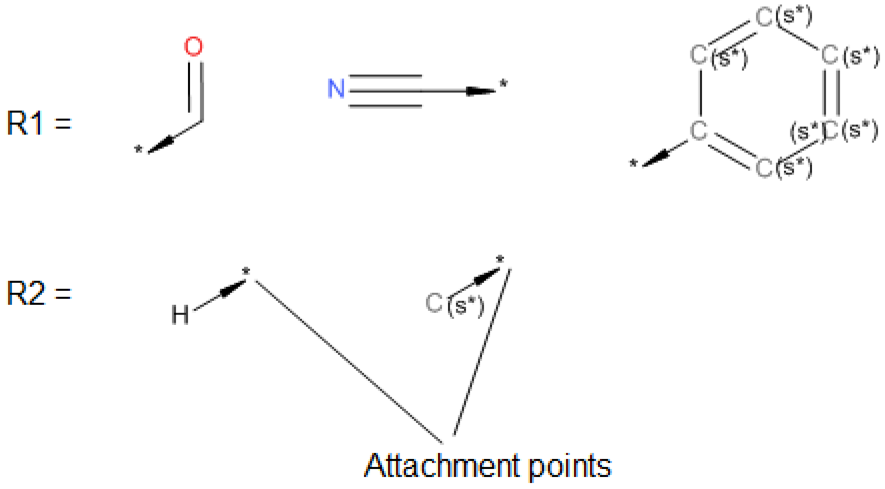
Rgroup substituents are the members of the list of Rgroups. You define the Rgroup substituents that you will allow to be found at specified positions in the molecules retrieved. Rgroup substituents can be any structural fragment, including functional groups and single atoms.

In the example below, the Rgroup substituents COOH, CN, or Ph are allowed at position 5 (R1) on the root structure and the Rgroup substituents OH or CH3 are allowed at position 2 (R2) on the root structure:

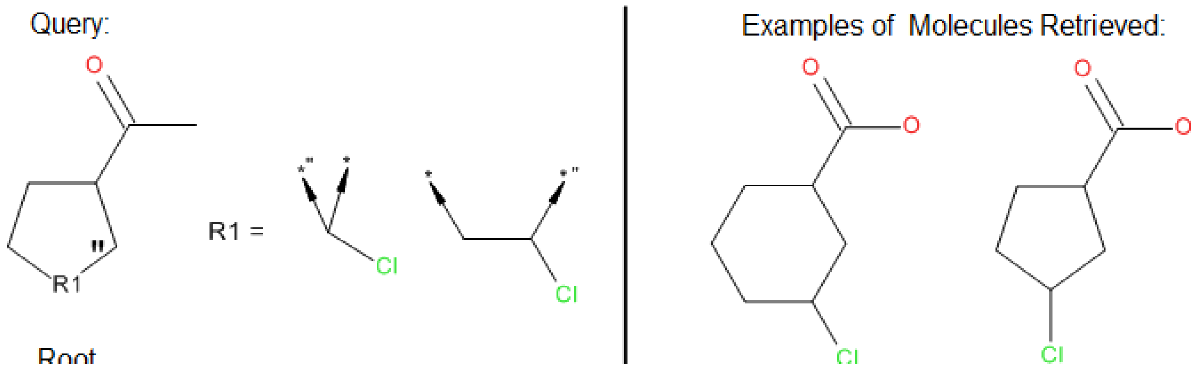


## Attachment Points

An attachment point (marked with an arrow and an asterisk) designates the atom on the Rgroup substituent that is attached to the root structure:



You must designate an attachment point on an Rgroup substituent for each bond that is attached to the substituent. For example, if an Rgroup is attached to two bonds on the root structure, you must designate two attachment points:

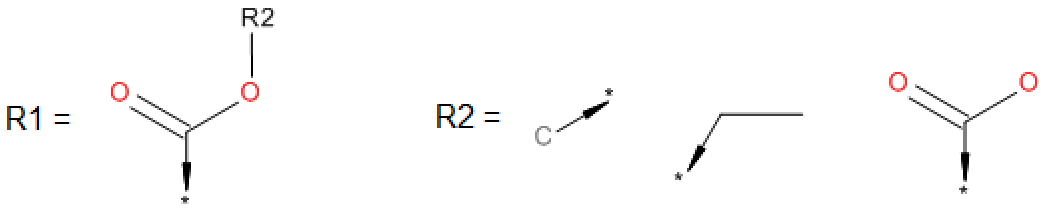


In this example, one bond on the root structure is marked with a single quote that corresponds to the attachment point with a single quote on the Rgroup substituent. The other bond on the root structure is marked with a double quote that corresponds to the attachment point with a double quote on the Rgroup substituent.

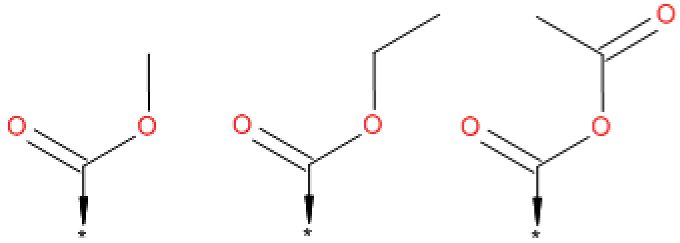
To obtain the correct search results, be certain to assign the attachment point with a single quote on the Rgroup substituent to the corresponding atom on the root structure.

## Nested Rgroup Substituents

A nested Rgroup is an Rgroup that is a placeholder within another Rgroup rather than within the root structure. (You cannot nest an Rgroup any lower than this level.) In the following example, Rgroup R2 is nested within the Rgroup R1:



These nested Rgroups are equivalent to the following structures:



You cannot specify more than one nested occurrence of a single Rgroup substituent (such as R2) within another (unnested) Rgroup substituent (such as R1). For example, you cannot specify that the nested Rgroup R2 be retrieved at both positions 1 and 2 within the R1 Rgroup substituent benzene. Instead, create a query in which a set of nested Rgroups (such as R2 and R3) contains identical substituents. Then specify the occurrence of each nested Rgroup only one time. For example, specify that the nested Rgroup R2 be retrieved at position 1 and that the nested Rgroup R3 be retrieved at position 2 within the R1 Rgroup substituent benzene.

## Rgroup Conditions

### RestH

With the Rgroup condition RestH, you can specify whether you will allow non-hydrogen substituents other than your defined Rgroup substituents to be attached to an unsatisfied Rgroup site. An unsatisfied Rgroup site is an Rgroup site on the root structure that does not have an Rgroup member attached. RestH can be On or Off:

- If RestH is Off, any functional group or atom may be attached at unsatisfied Rgroup sites on the root structure.
- If RestH is On, only hydrogens can be attached at unsatisfied Rgroup sites on the root structure.

### Occurrence

With the Rgroup condition Occurrence, you can specify the number of Rgroup sites that must be occupied in the records retrieved. For example, the occurrence designation R1 > 0 specifies that an Rgroup substituent from the R1 list occupies at least one R1 site on the root structure in the records retrieved. Type one of the following values in the Occurrence range text box:

| Value | Description |
|---|---|
| 0 | All members of the Rgroup are excluded from the Rgroup site (such as R1). To allow any other Rgroup substituent at the Rgroup site, set RestH to off. |
| >0 | At least one Rgroup must be satisfied. If RestH is off, any other substituent can attach to unsatisfied sites. If RestH is on, unsatisfied sites must be filled solely by hydrogens. |
| n | Exactly n number of sites must be satisfied. RestH controls substitution at unsatisfied sites. |
| <n | Less than n number of sites (including zero) must be satisfied. RestH controls substitution at unsatisfied sites. |
| m-n | Any number of sites falling within the range of m to n number of sites must be satisfied. RestH controls substitution at unsatisfied sites. |
| m,n | Any combination of the above expressions, separated by commas. For example: 1,3-5,9,>12. |

For example, the Rgroup condition R1 > 0 specifies that an Rgroup substituent from the R1 list occupies at least one R1 site on the root structure in the records retrieved.

In another example, the Rgroup condition R1 > 0 and R2 > 0 specifies that an Rgroup substituent from the R1 list AND an Rgroup substituent from the R2 list must occupy at least one site on the root structure in the records retrieved.

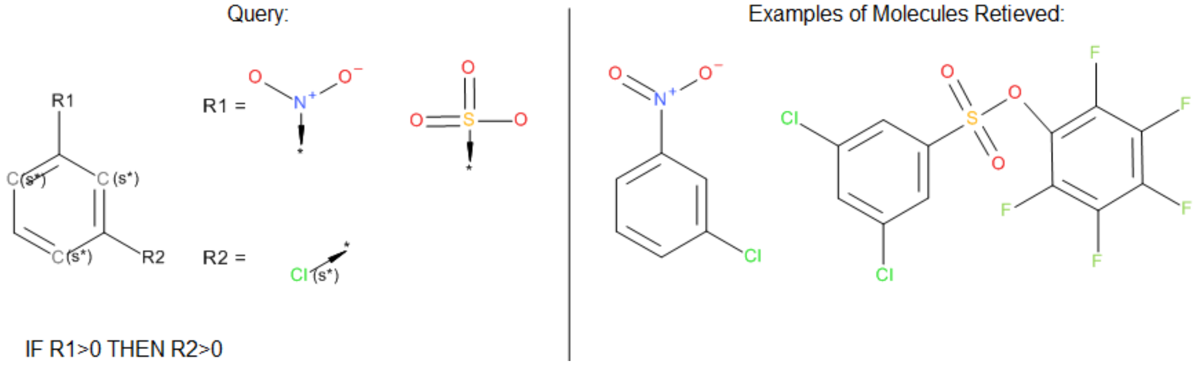
Finally, the Rgroup conditions R1 = 1, 3-5, 9, >12 specify that either 1, 3, 4, 5, 9, or more than 12 sites must be satisfied.

### If/Then

With the Rgroup condition If/then, you can specify whether the presence of one Rgroup is dependent on the presence of another Rgroup.

For example, If R1 then R2 specifies that if the conditions for R1 are satisfied, then the conditions for R2 must also be satisfied. (If the conditions for R1 are not satisfied, then the conditions for R2 are ignored.) The If/then condition implies that a molecule may be retrieved even though R1 is not satisfied.

In the following query, R1 contains a nitro group and a sulfonic acid group. R2 contains a chlorine atom. The condition If R1 then R2 specifies that if a nitro group or a sulfonic acid group is attached at R1 on the root structure, then a chlorine must also be attached at R2. However, if the R1 site on the root structure does not have either the nitro or the sulfonic acid group (R1 is unsatisfied), then R2 can have any substituent (R2 is ignored):



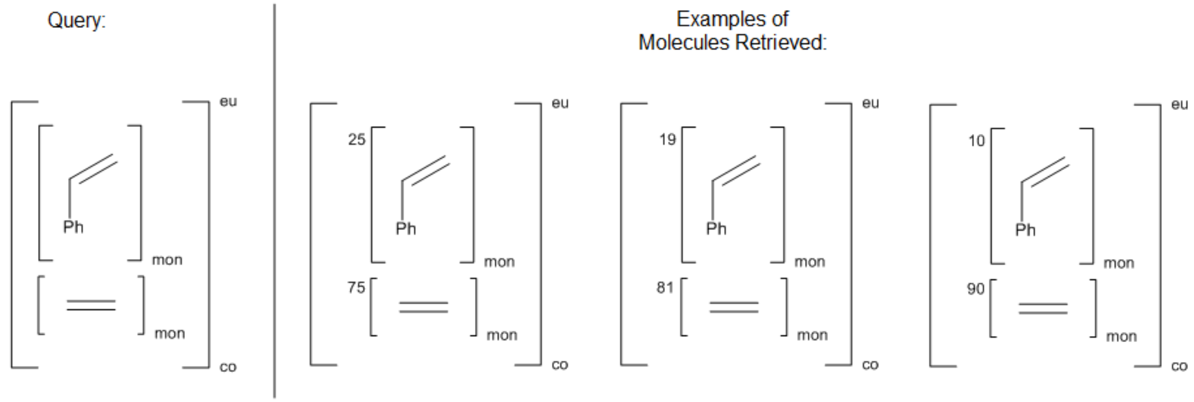
## Substructure Search of Attached Data

You can specify a text/numeric query (search operator with conditions) for attached data, a target value (specific data) of attached data, or no attached data. The following sections show the search results that you obtain for each type of query.

**Note:** In these examples, Ph is an abbreviated structure that represents a phenyl group.

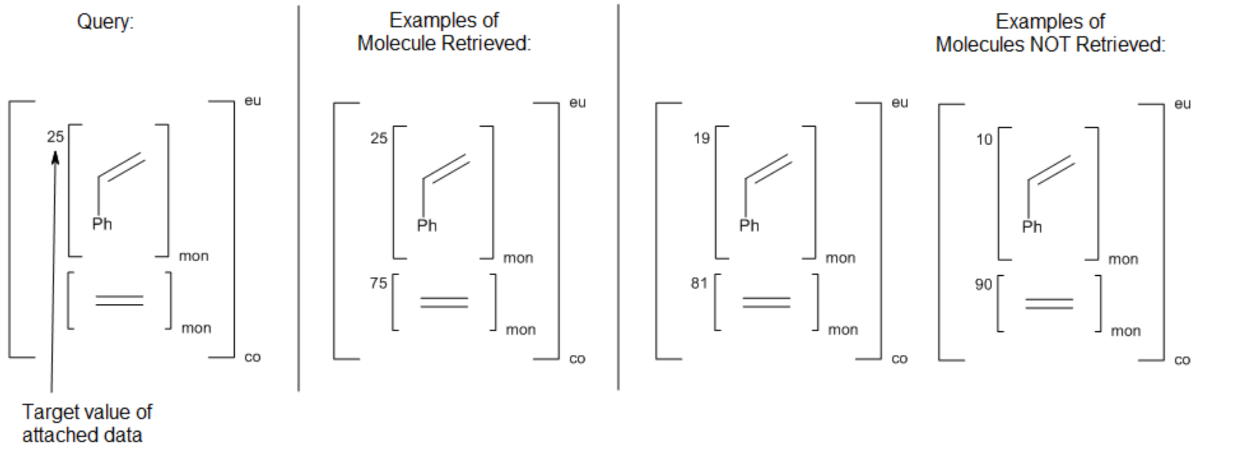
### No Attached Data

A query that contains no attached data matches structures that contain any attached data. In the following example, the numbers on the brackets are attached data:



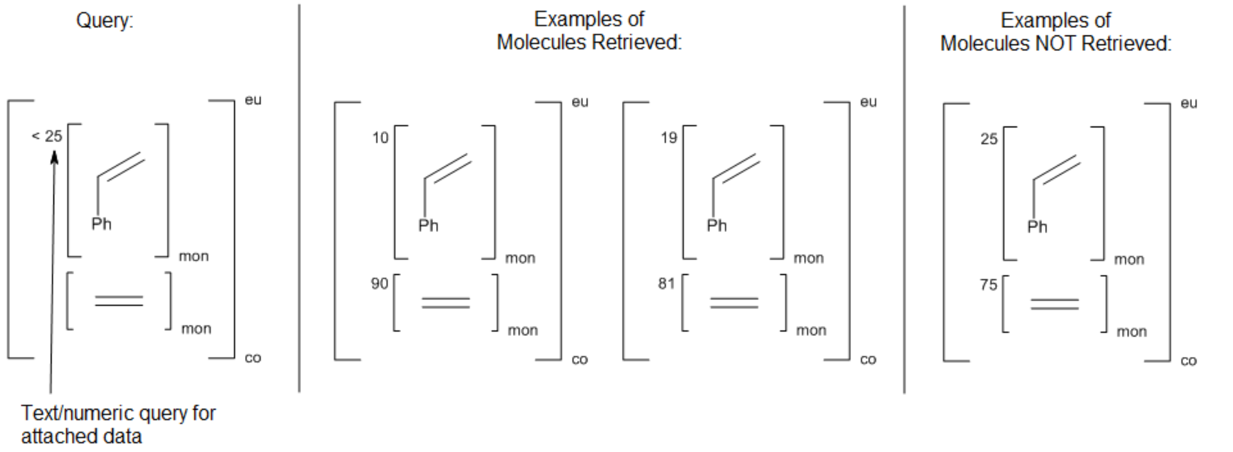
### Target Value (Specific Data)

A query that contains a target value (specific data) matches structures in which the attached data matches exactly. For example:



## Search Operator and Target Value

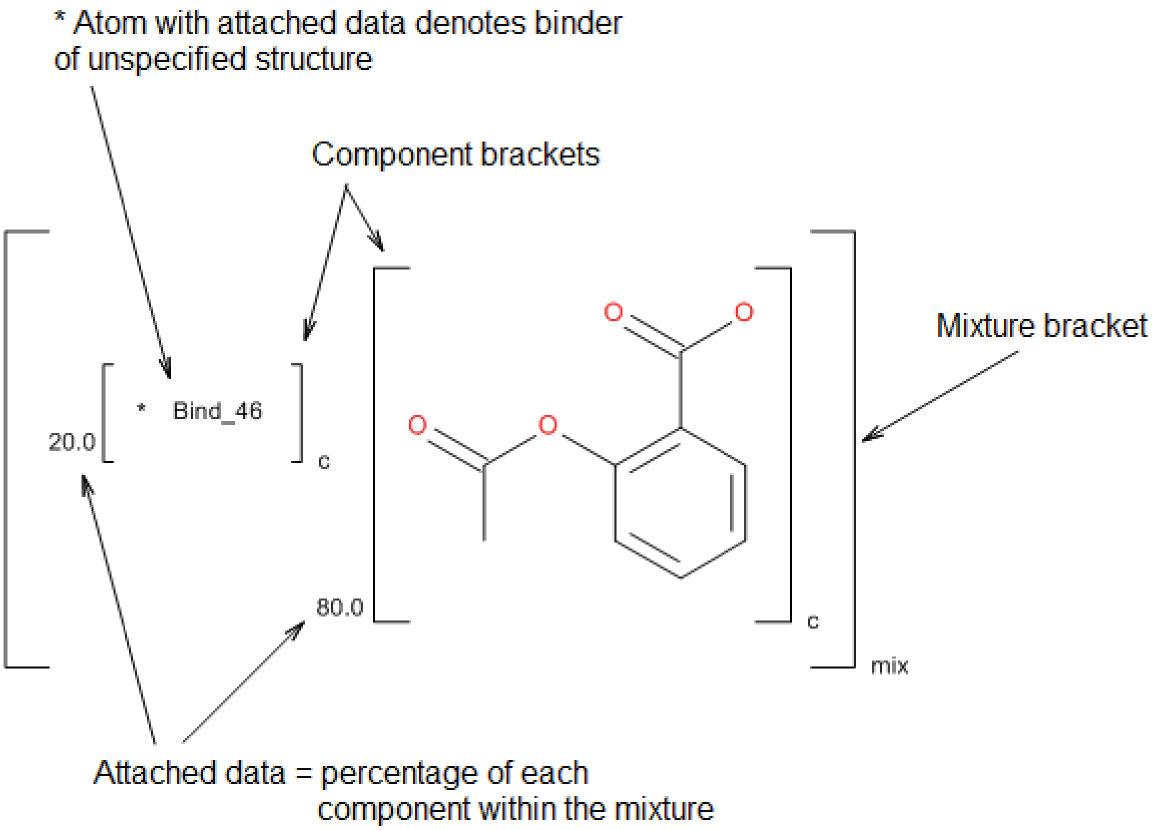
A query that contains a search operator and target value matches structures in which the attached data meets the conditions that are specified. For example, to find copolymers that contain less than 25% polystyrene, attach the text/numeric query < 25 to the monomer brackets of polystyrene:



In BIOVIA Draw, you use the Attached Data dialog to specify search operators. The Attached Data dialog is hidden by default. For information on customizing BIOVIA Draw to enable this dialog for the attached data that your business rules require, see Related Documentation for BIOVIA Draw on page 73.

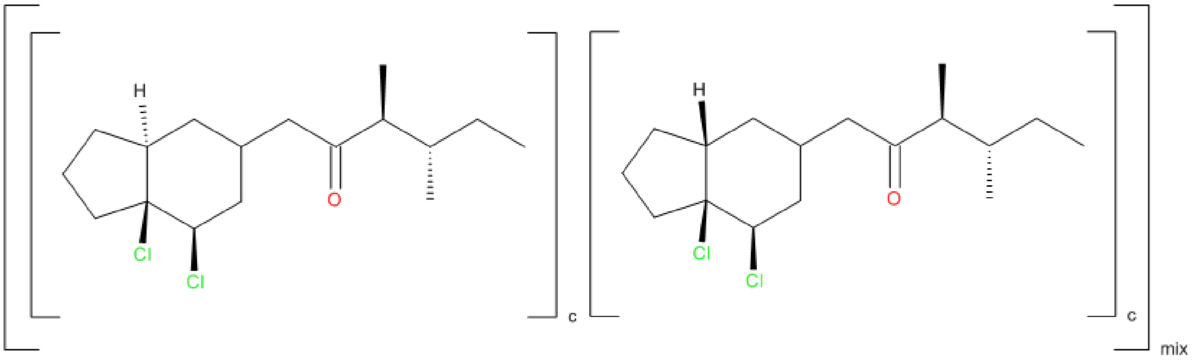
## Substructure Search of Mixtures

You can use mixture brackets and component brackets in substructure search queries. For example, the following structure represents a mixture of two components: aspirin (80% by weight) and a binder of unspecified structure called Bind_46 (20% by weight).

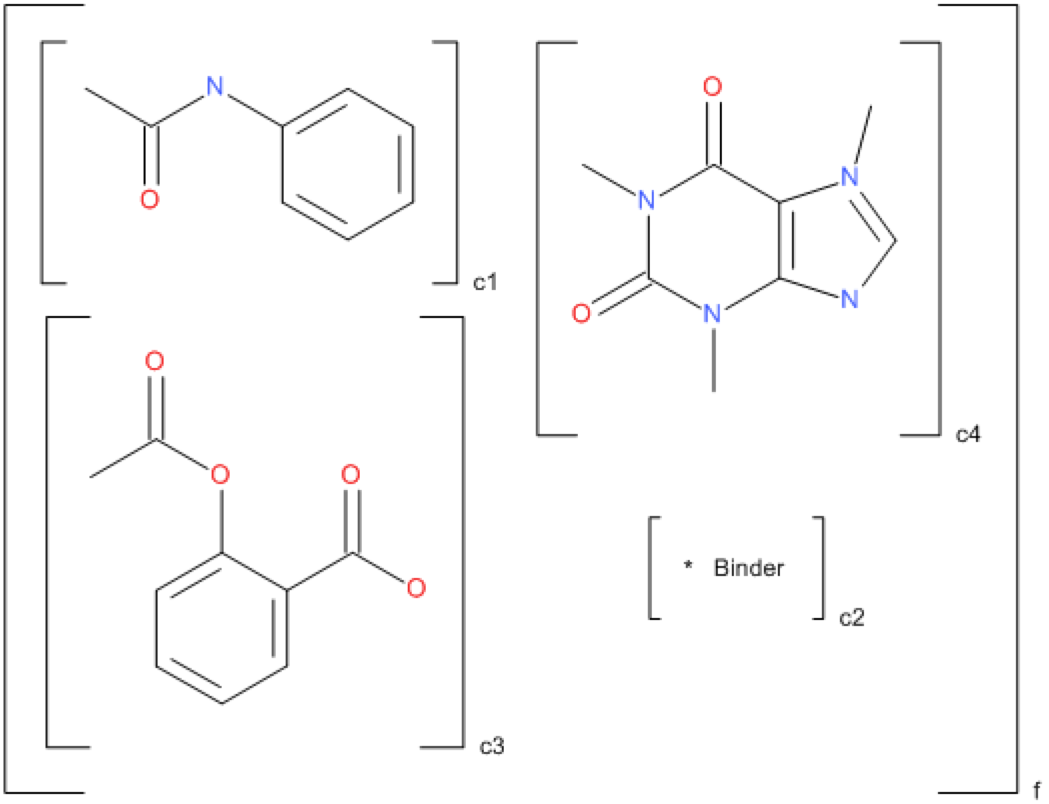

* Atom with attached data denotes binder of unspecified structure
Component brackets
Mixture bracket
20.0
*
Bind_46
c
80.0
O
O
O
O
c
mix
Attached data = percentage of each component within the mixture

If your database contains structures with mixture brackets and component brackets, study the following sections to learn how to create search queries for these structures.

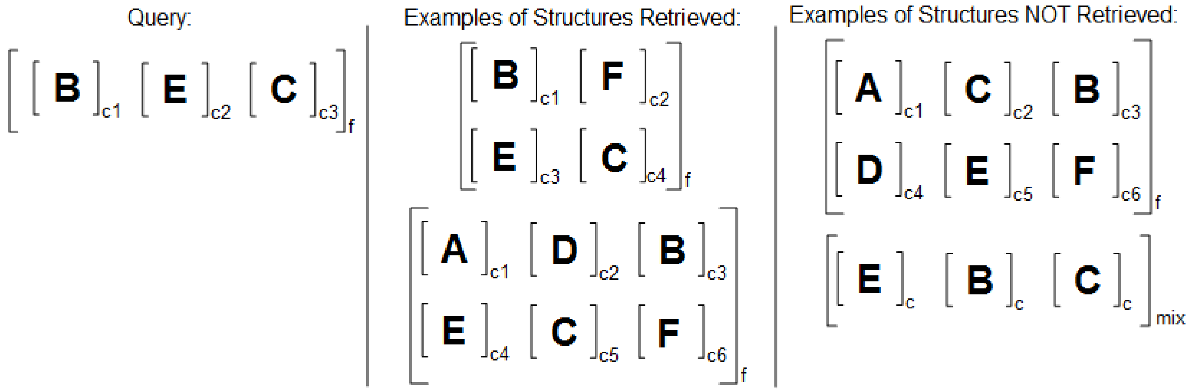
If the order of addition of the components of a mixture is not important, specify an unordered mixture (mix). Unordered mixtures contain components (c) that are not numbered. For example, the following is a mixture of stereoisomers from a reaction product:



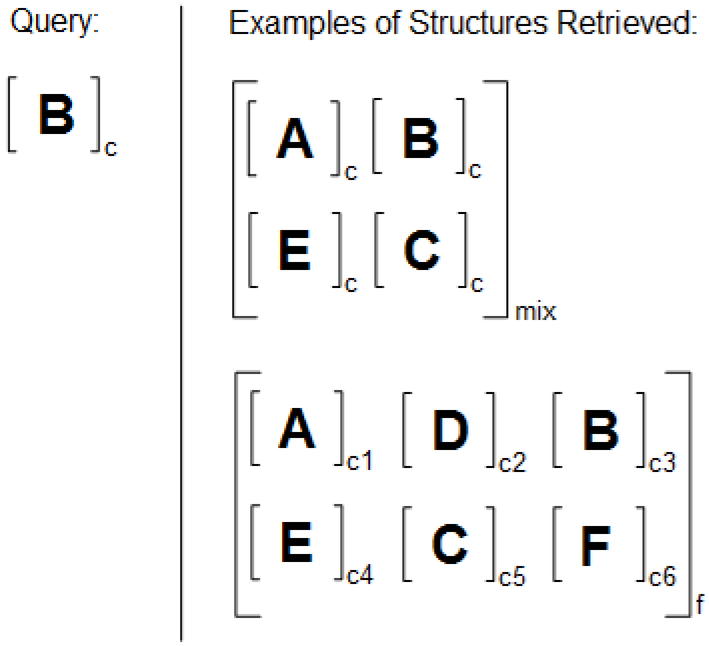
If the order of addition of components of the mixture is important, specify an ordered mixture (f). Ordered mixtures contain components that are numbered (c1, c2, c3, and so forth). In the following example, the acetaminophen (c1) must be added before the binder (c2) and other components (c3 and c4):



A query that contains ordered components finds structures that contain ordered components with the same relative (but not necessarily sequential) order. For example:



To find both unordered and ordered mixtures that contain specified components, use an unordered component without brackets. For example:



For more information and examples of structures that use mixtures, see Mixture Representation on page 74.

## Substructure Search of Polymers

Substructure search of polymers complies with the following rules:

- The following polymer bracket types must match exactly: SRU(n), mon, mer, blk, ran, alt
- Unspecified copolymer brackets (co) match specific copolymer bracket types: blk, ran, alt
- The number of crossing bonds must match exactly
- A repeat pattern of eu can hit any repeat pattern. All other repeat patterns must match exactly
- For polymers with the head-to-tail (ht) repeat pattern, you can define the structural repeating unit anywhere along the polymer backbone (that is, you can "phase-shift" the brackets along the polymer backbone)

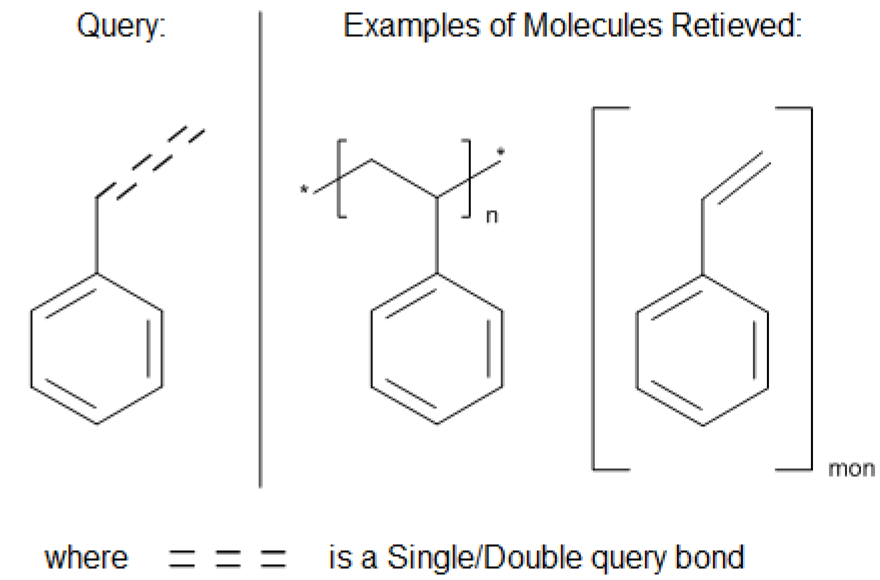
For information on polymers, see Polymer Representation on page 78.

For information on flexmatch search of polymers, see Substance Groups (Sgroups) on page 113.

Exceptions to this behavior are:

- A polymer with *atom end groups hits structures with defined end groups
- Monomer/SRU matching is not performed
- You can use the Anypolymer (anyp) bracket type in the query. The anyp bracket matches all other polymer bracket types

You can use the Anypolymer bracket type to create a query that matches both source-based and structure based representations of a polymer. For example, you might use atom and bond query features to create a Markush polymer query that matches both the structure-based and source-based representations of an addition polymer:

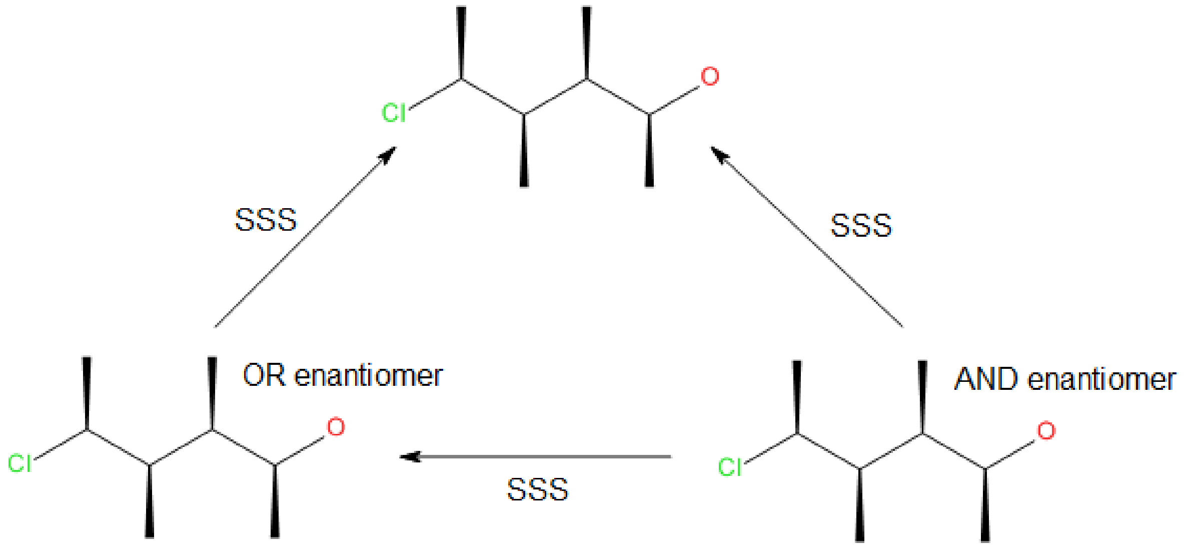

Query:
Examples of Molecules Retieved:
*
*
n
mon
where
is a Single/Double query bond

In this example, the bond query feature Single/Double finds both single and double bonds. The anyp brackets match structures with either monomer (mon) or SRU (n) brackets.

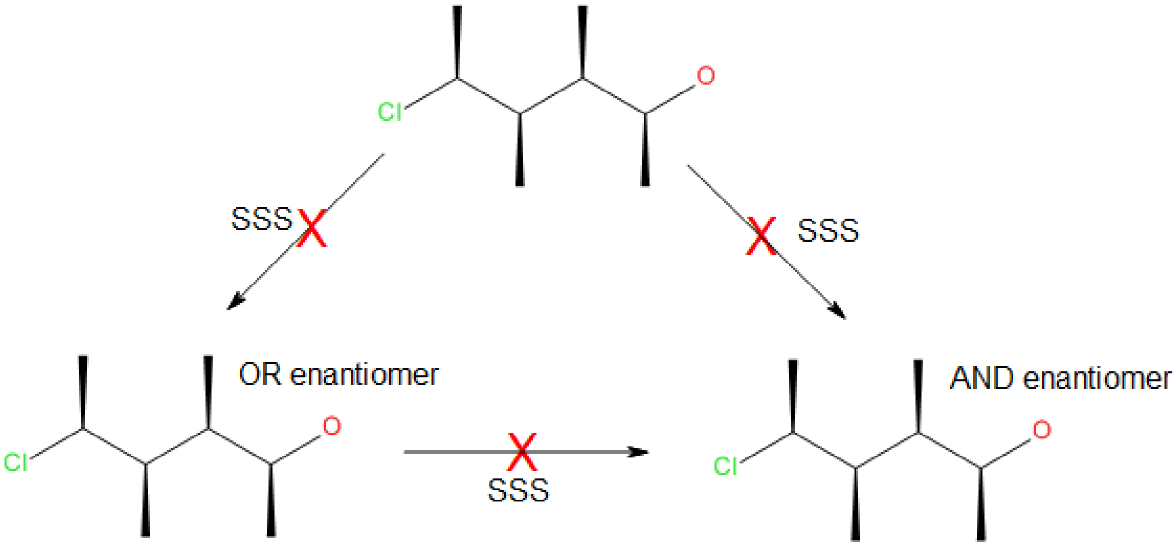
For more information on query features, see Query Features on Atoms and Bonds on page 138.

## Substructure Search of Structures with Tetrahedral Stereochemistry

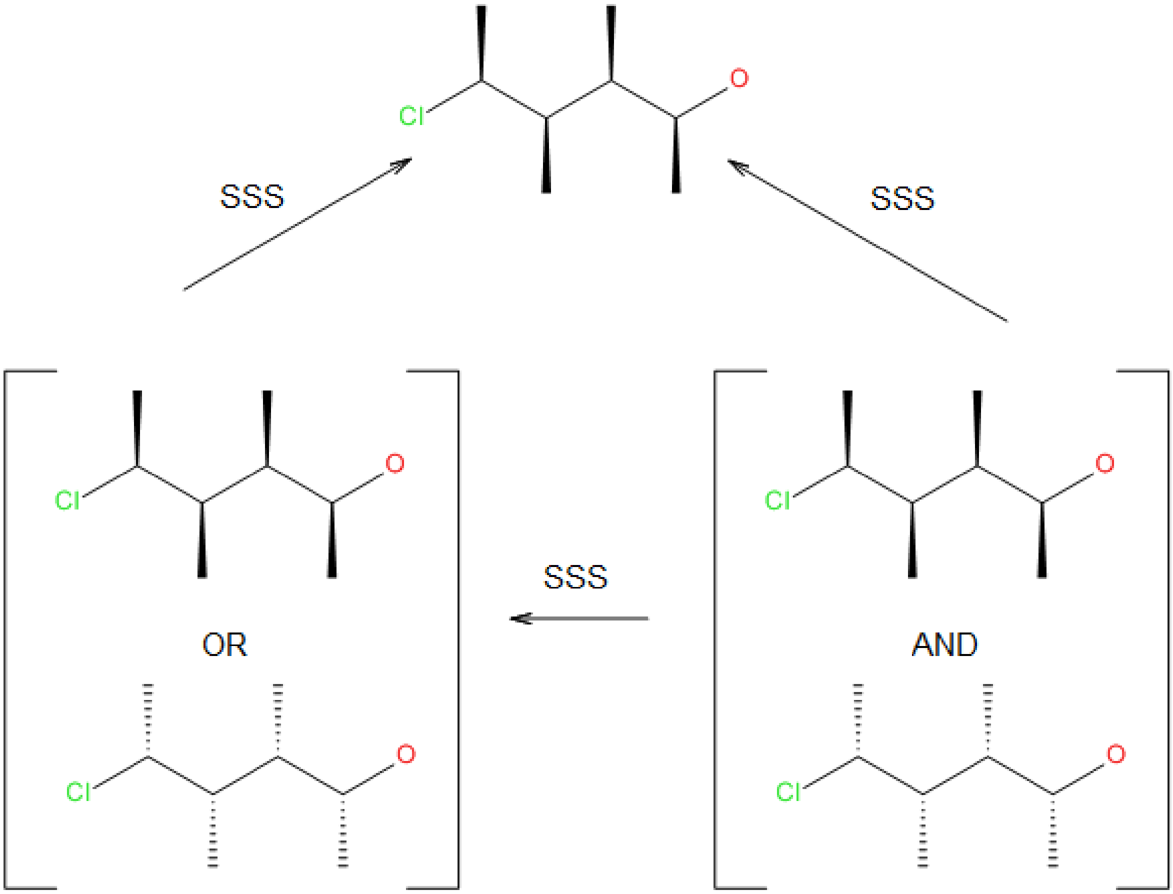
In substructure searching (SSS), stereogenic centers in the ABS stereogroup match only structures in the ABS stereogroup. Centers in OR stereogroups match centers with either the ABS or OR stereogroups. Centers in AND stereogroups match centers with ABS, OR, or AND stereogroups:



The reverse is not true, however, as shown in the figure that follows



The reasons for these results become clear when the OR and AND structures are enumerated to show the structures that they actually represent:

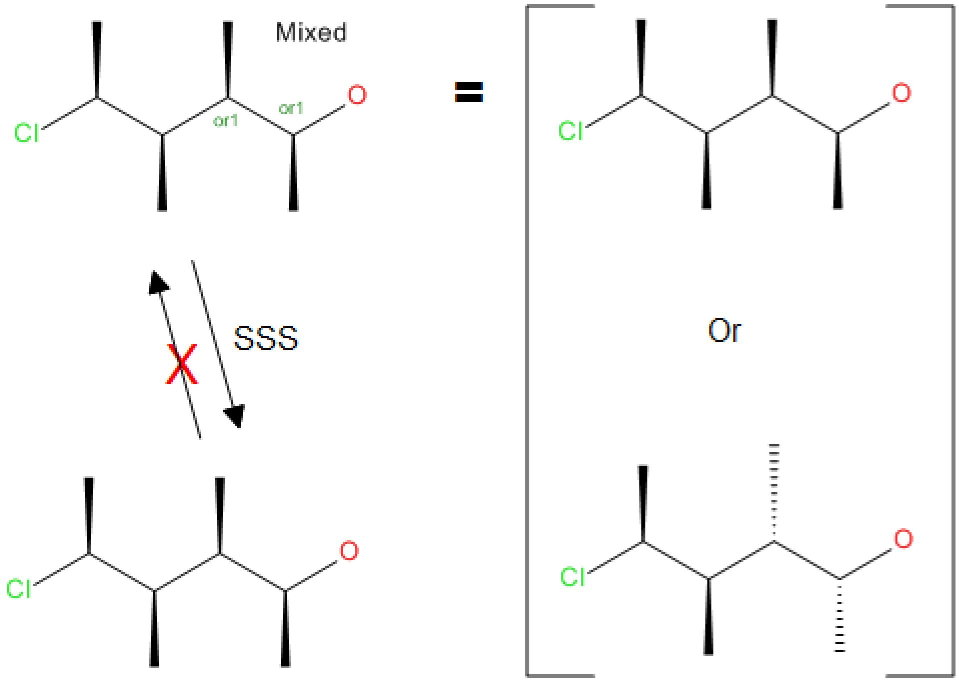


The OR and AND structures match the absolute configuration because both OR and AND structures contain that absolute configuration. AND matches OR, but OR does not match AND. The OR structure means, "This is a single stereoisomer, but I do not know which one it is", while the AND structure means, "Both of these stereoisomers are present." Clearly, the AND structure contains both alternative OR structures, while the OR structure contains only one of the two structures in the AND structure.

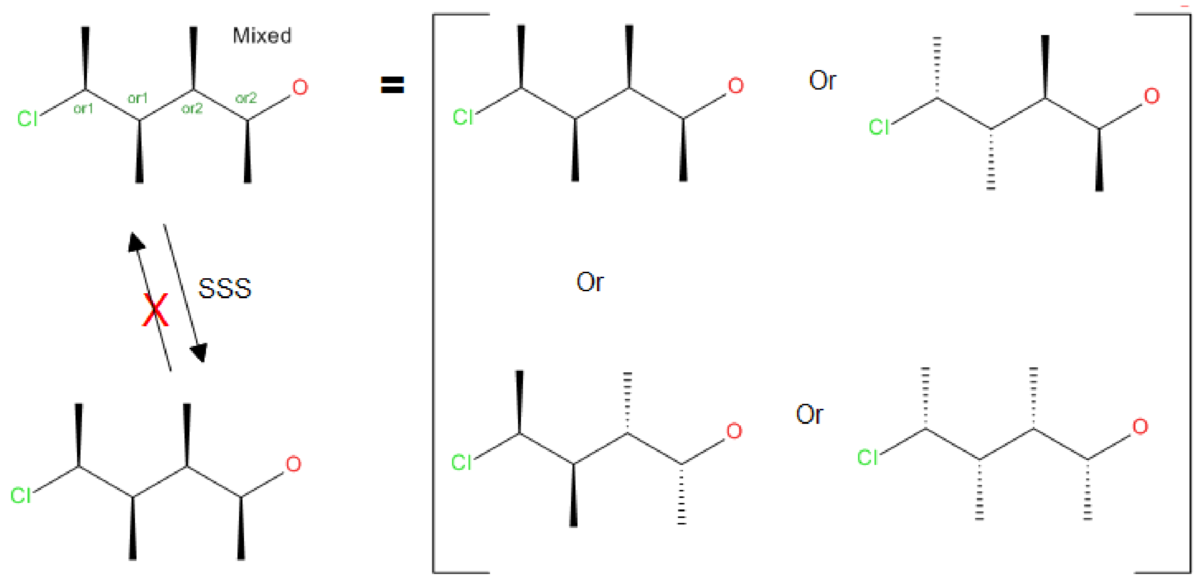
In summary, the structures that the query matches must represent stereoisomers that are entirely contained within the stereoisomers that are specified by the query. The sections that follow contain more examples that illustrate this.

## Example 1

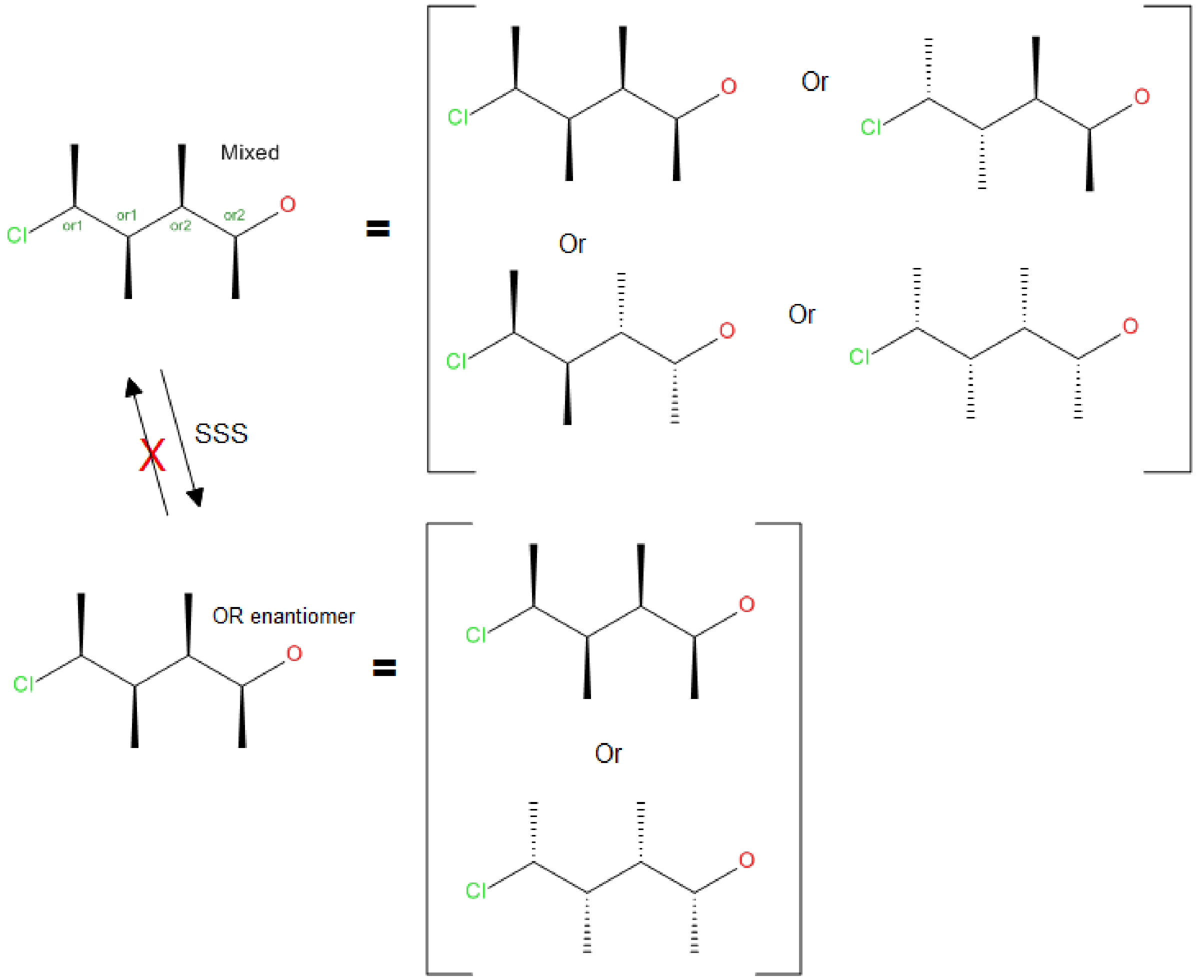
The structure with two stereochemical groups (the ABS group plus one OR group) matches the absolute configuration, because that configuration is entirely contained within the pair of stereoisomers that the query represents.



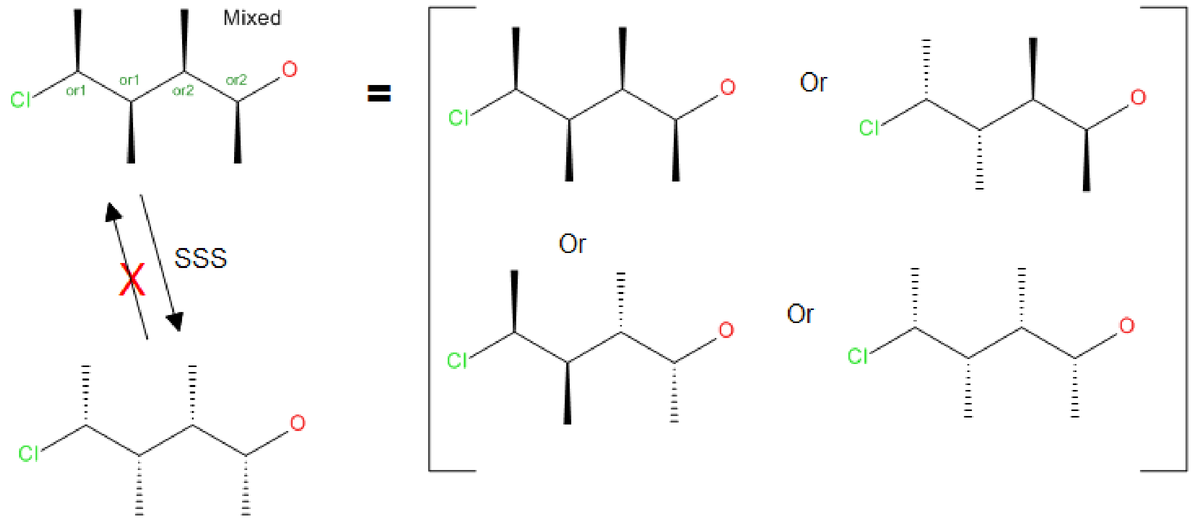
## Example 2



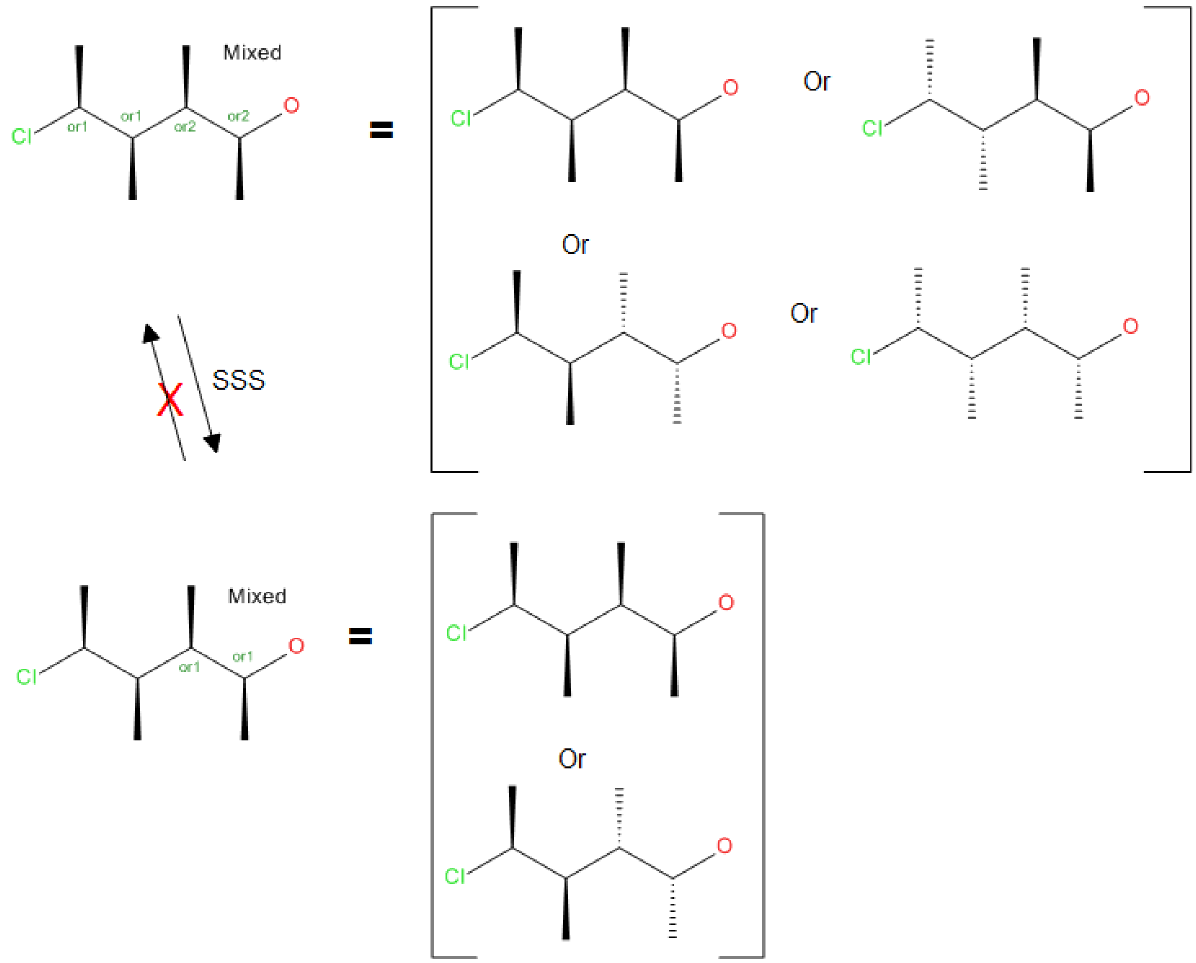
## Example 3



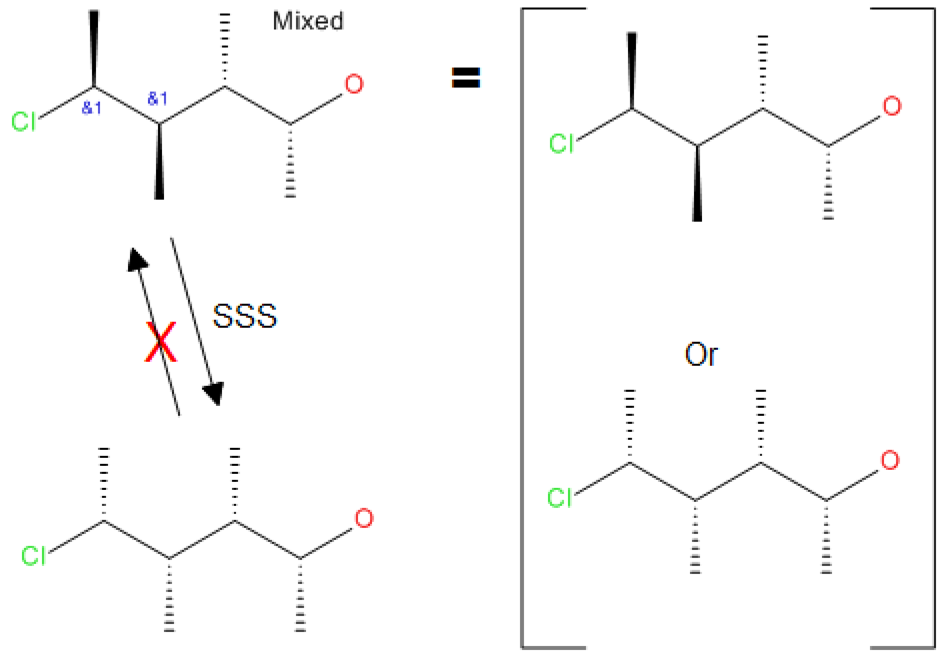
## Example 4



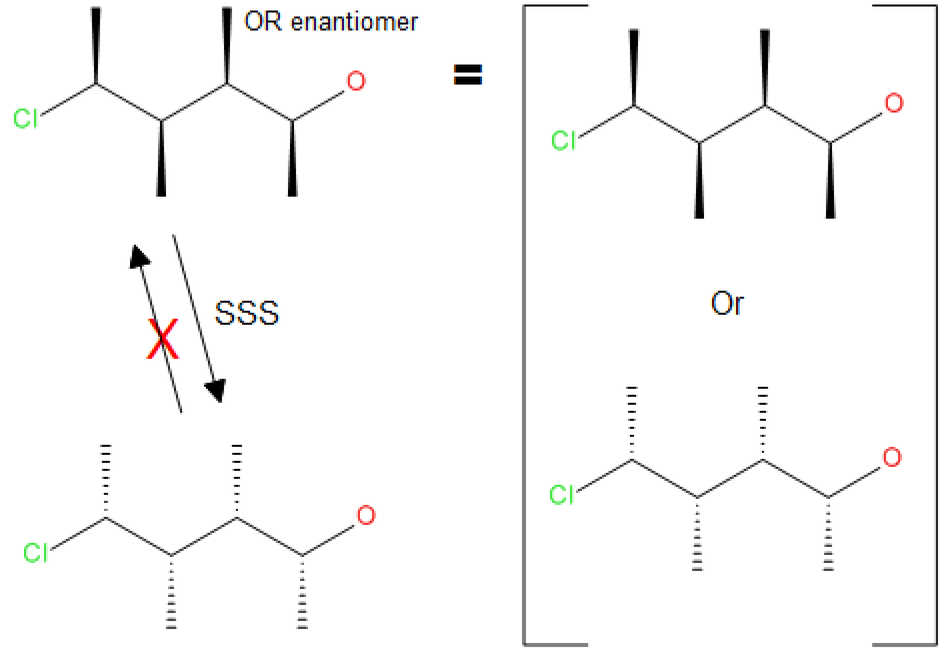
## Example 5



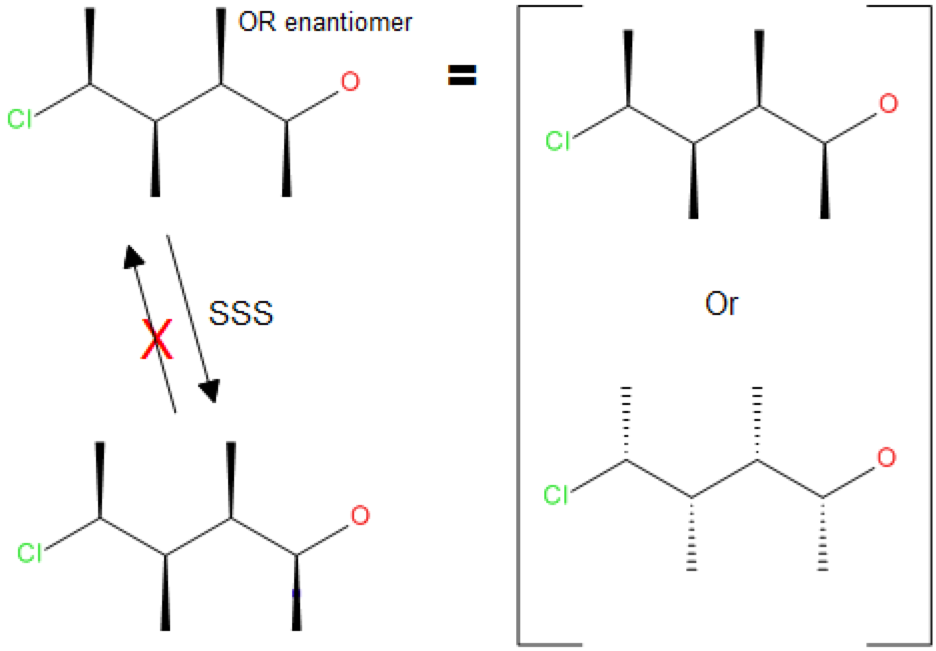
## Example 6



## Example 7

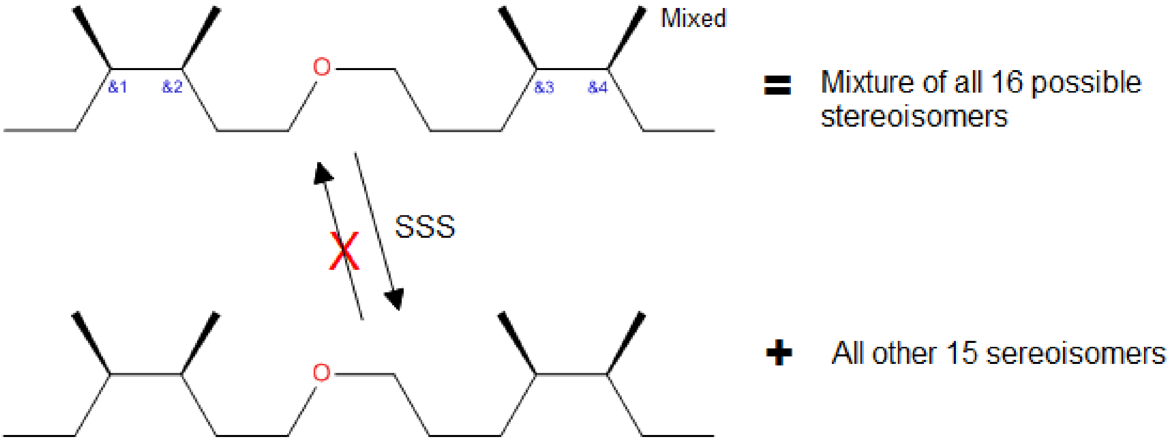


## Example 8



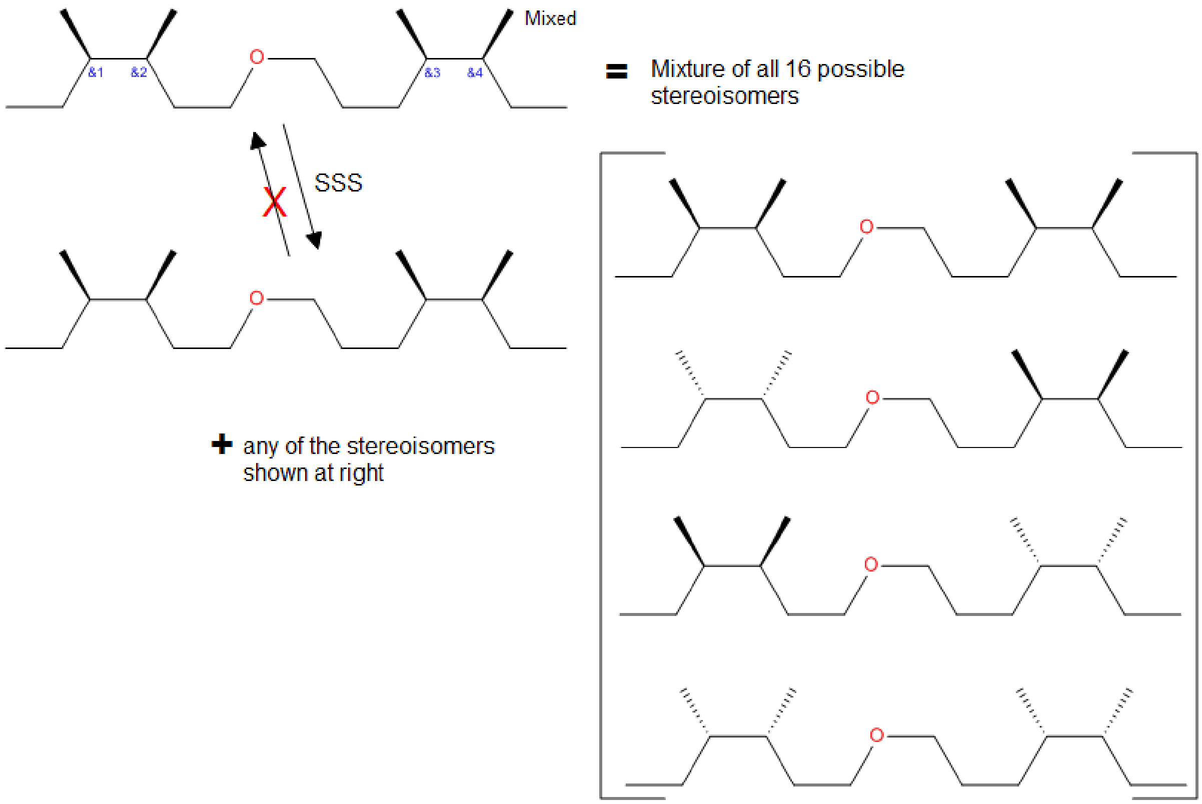
## Example 9

Each of the four stereogenic centers is in a different AND group. Therefore, the number of stereoisomers is 24 = 16. The structure with the AND groups matches all 16 possible stereoisomers, because all 16 stereoisomers are contained within the set of 16 that the structure represents.


Mixed
&1
&2
O
&3
&4
=
Mixture of all 16 possible stereoisomers
SSS
O
+
All other 15 sereoisomers

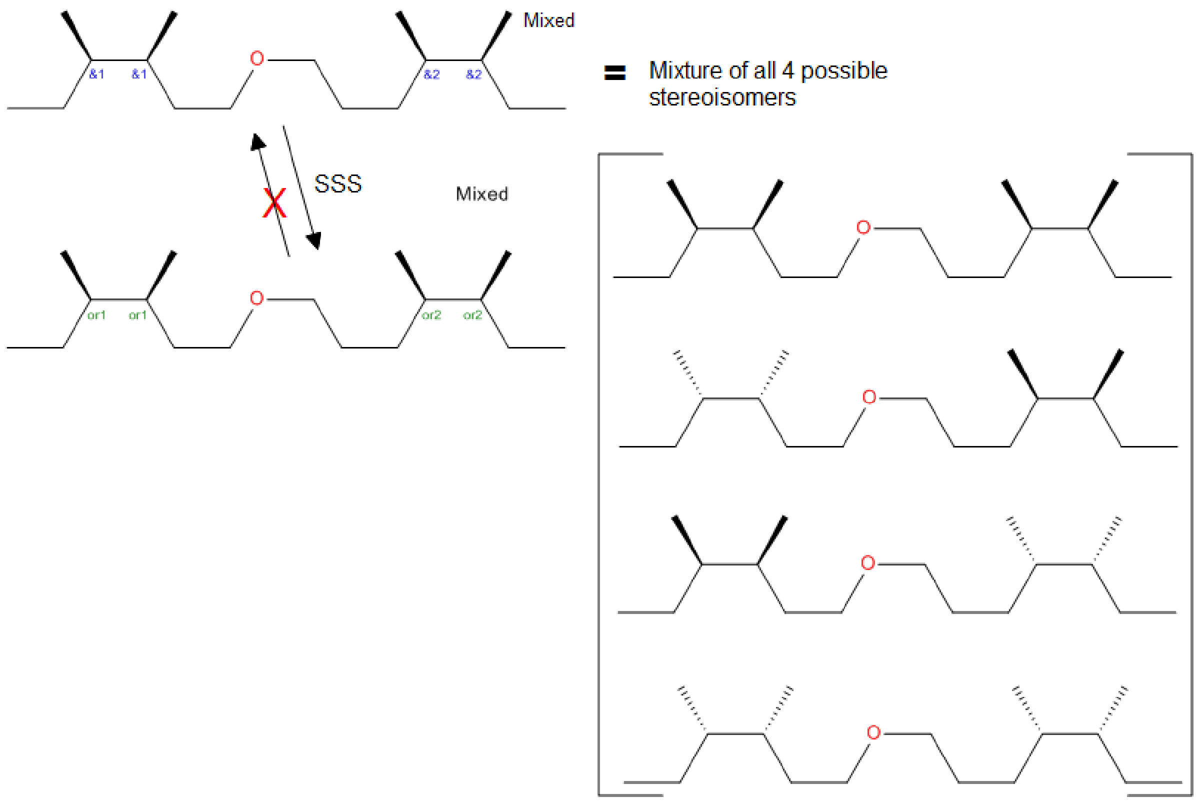
## Example 10

The structure with four AND groups matches the structure with two AND groups, because the latter is entirely contained within the set that is specified by the former.


&1
&2
&3
&4
Mixed
SSS
=
Mixture of all 16 possible stereoisomers
+ any of the stereoisomers shown at right

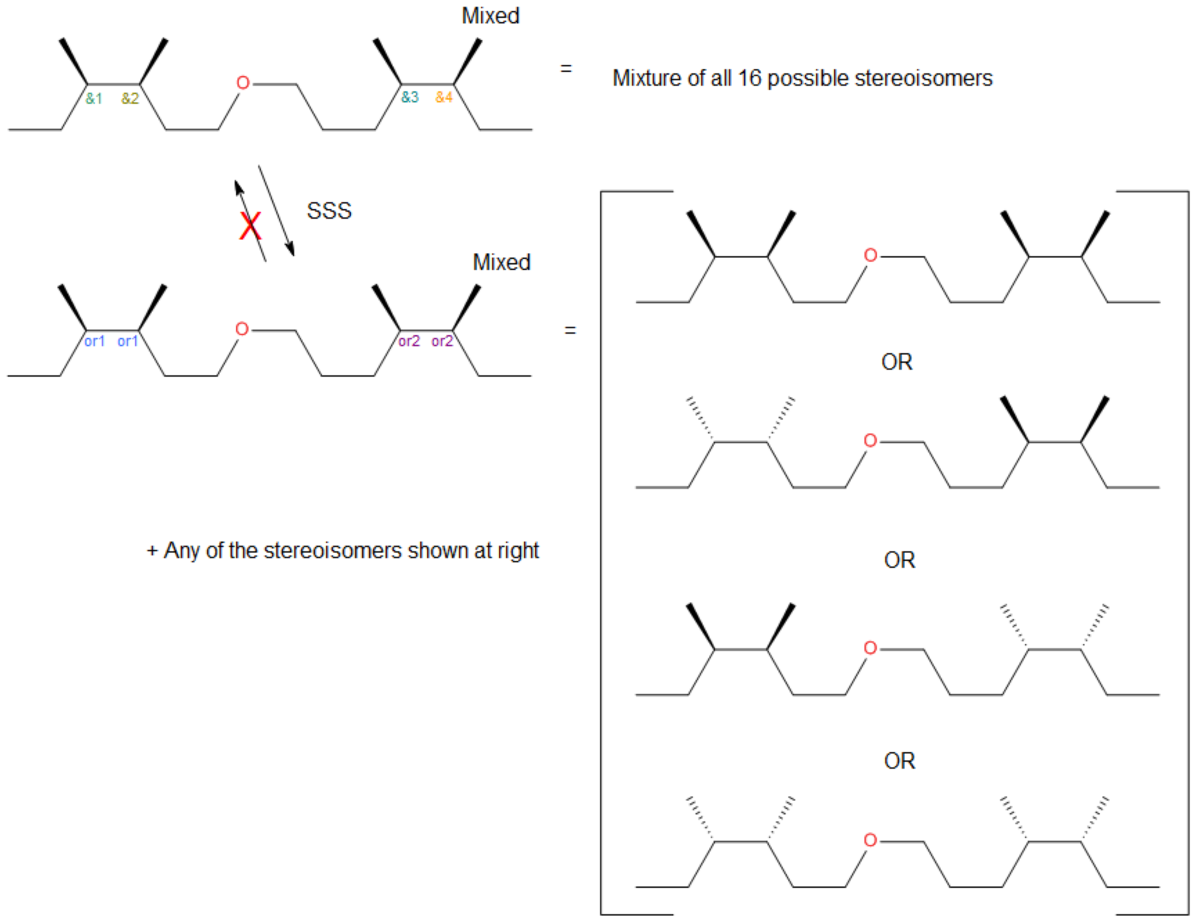
## Example 11

The structure with the two OR groups represents any one of 4 possible stereoisomers. This set of stereoisomers is entirely contained within the structures specified by the structure with two AND groups.



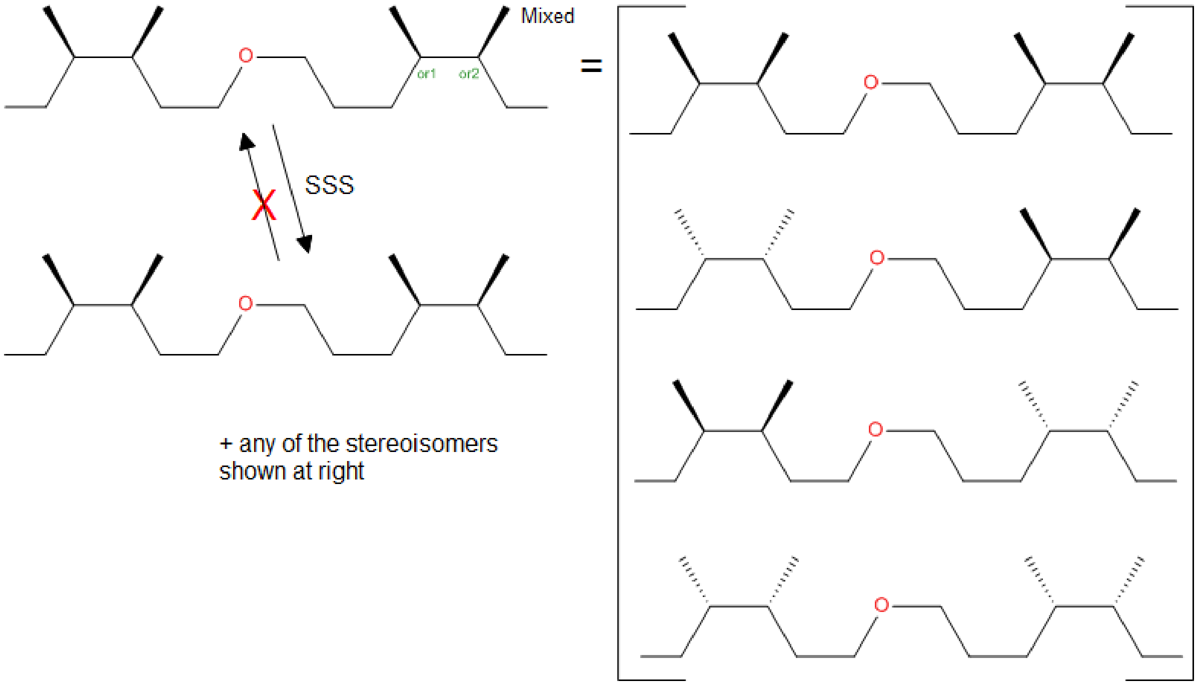
## Example 12

The query matches solely stereoisomers with the same relative configuration of the two stereogenic centers on the right, because these two centers belong to the same OR group.


Mixed
&1
&2
&3
&4
=
Mixture of all 16 possible stereoisomers
SSS
Mixed
or1
or1
or2
or2
=
OR
OR
OR
+ Any of the stereoisomers shown at right

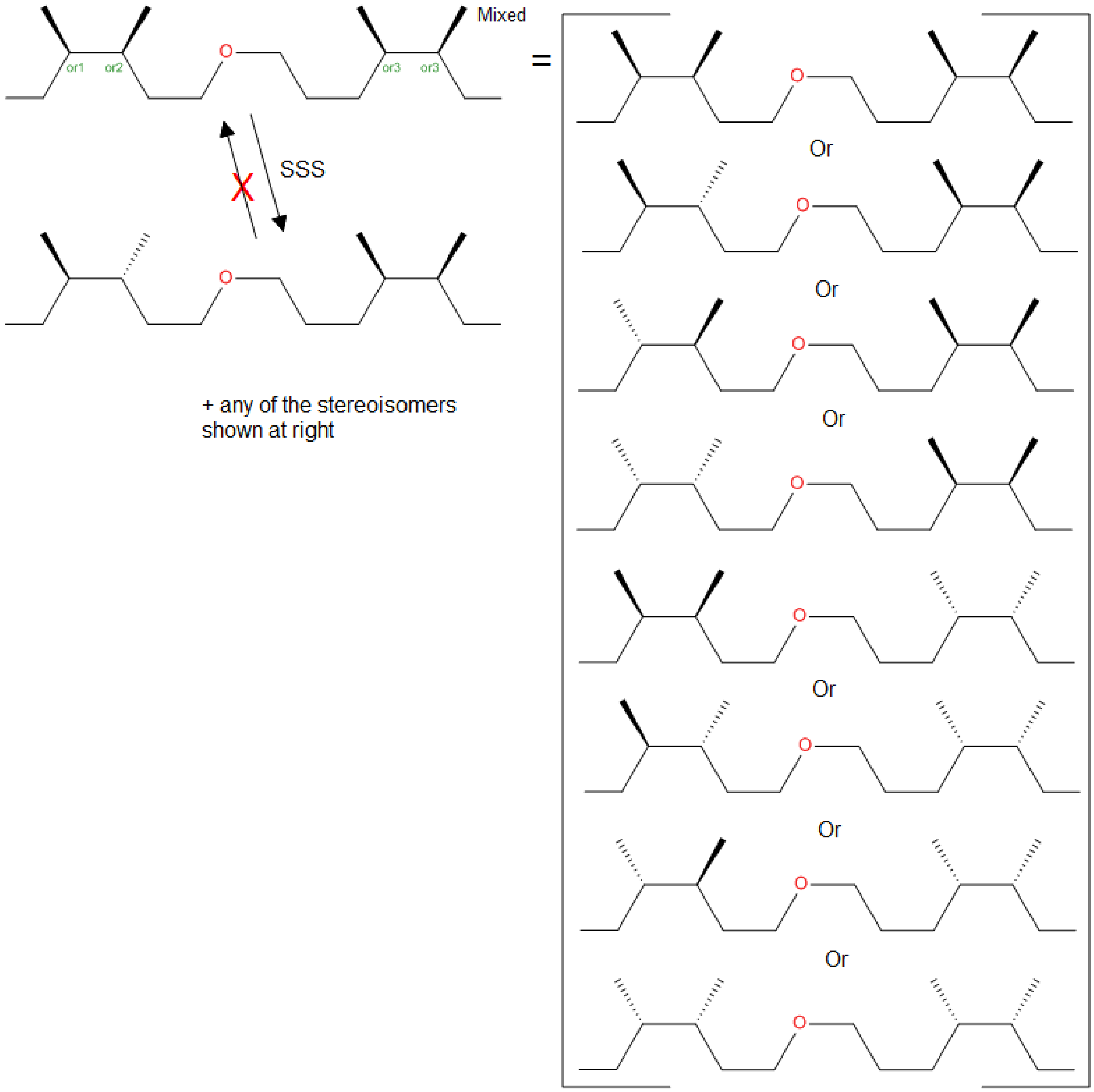
## Example 13

The query matches stereoisomers with any combination of configurations of the two stereogenic centers on the right of the structure, because these two centers are in different OR groups.



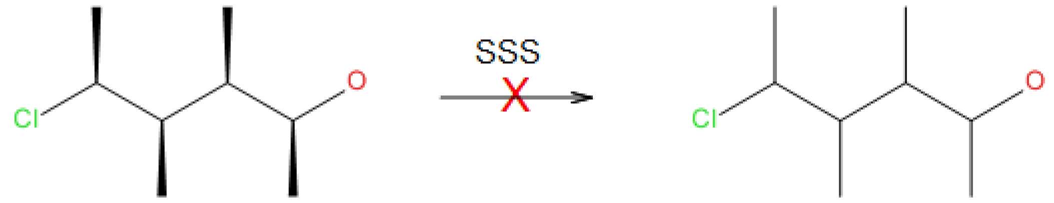
## Example 14

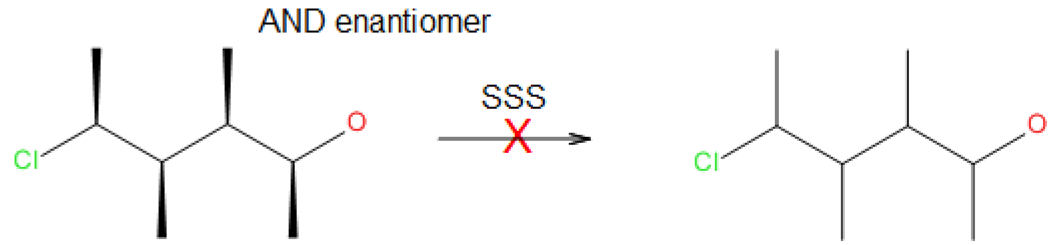
The query matches only 8 of the 16 stereoisomers that are possible for structures with 4 stereogenic centers. The query specifies 8 different stereoisomers because its 4 centers belong to only 3 different OR groups. The two stereogenic centers on the right have the same relative configuration because they belong to the same OR group.



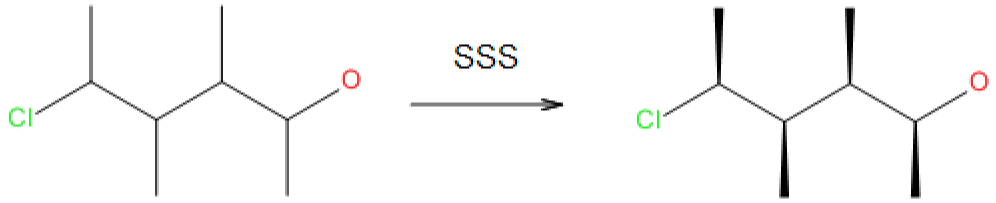
## Example 15

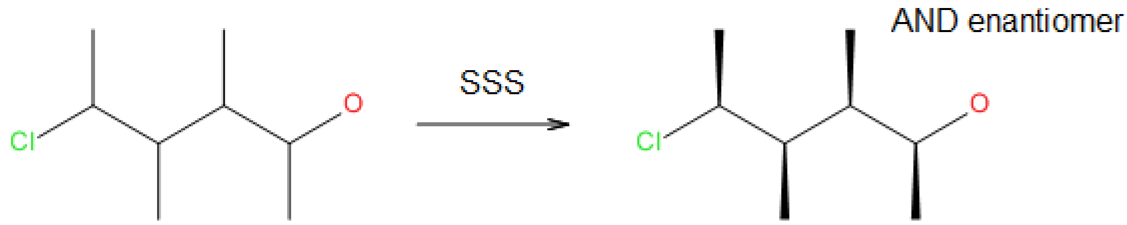
None of the queries in the preceding examples matches a structure that contains unspecified stereogenic centers, for example:

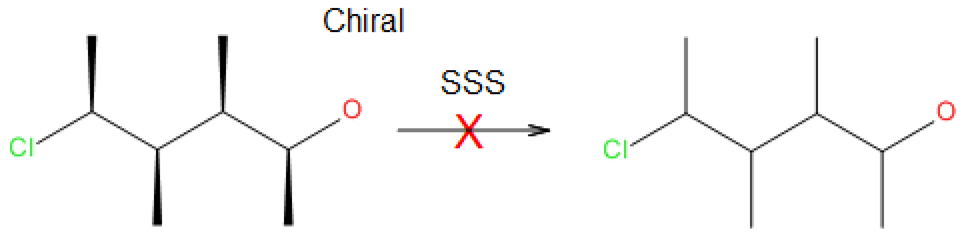


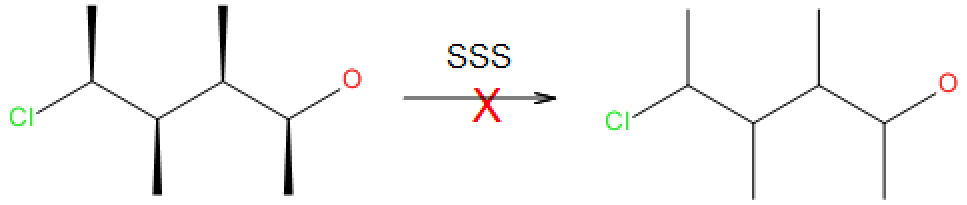


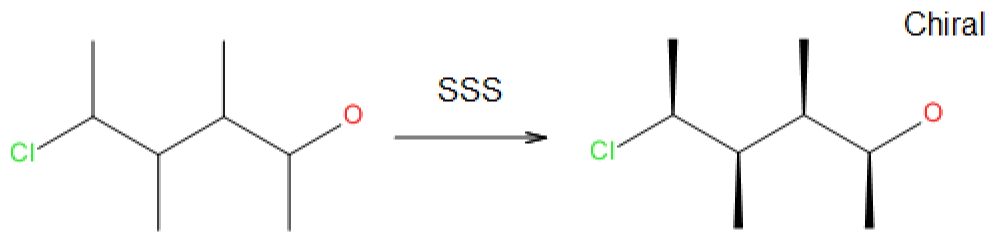
However, a structure with undefined stereogenic centers matches any structure that contains it as a substructure, whether the centers are marked or not:

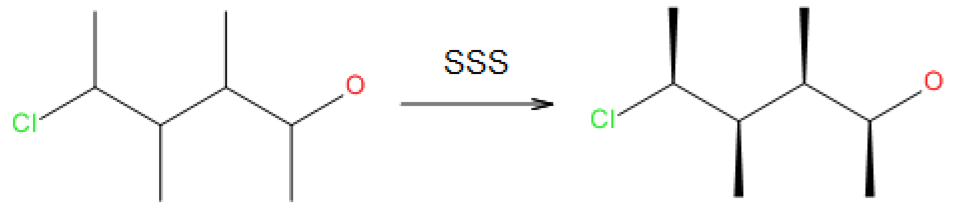










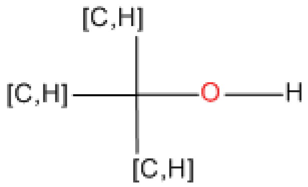


## Queries for Common Functional Groups

### Alcohols and Phenols

#### Aliphatic Alcohols

The following query matches all alcohols (primary, secondary, and tertiary):

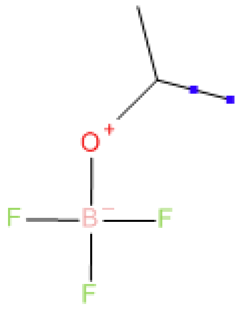


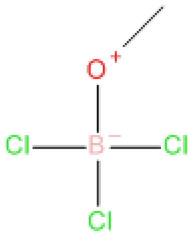
The four single bonds to the alcohol carbon ensure that the carbon will be saturated, as required by the IUPAC definition of an alcohol.

The hydrogen atom on the oxygen is drawn explicitly to prevent substitution of other atoms.

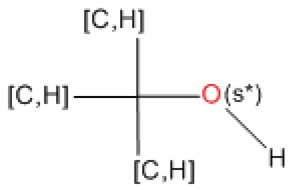
The atom list query feature [C,H] specifies that only carbon or hydrogen can be attached to the alcohol carbon.

This query also matches a small number of structures like the following examples:



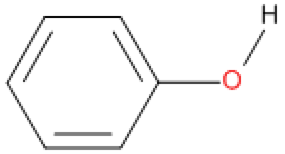


To avoid retrieving these structures, add the Substitution Count as Drawn (s*) query feature to the oxygen atom:

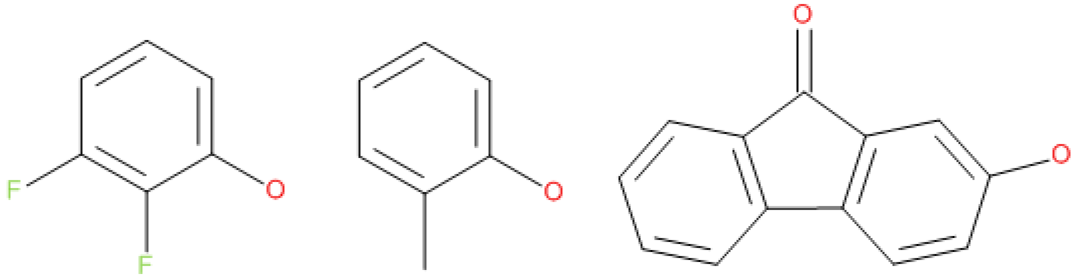


## Phenols

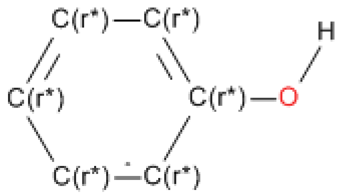
The following query matches phenols:



The following are examples of the structures that you retrieve:



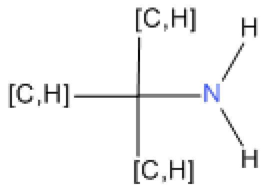
To avoid retrieving phenols that contain fused rings, apply the Ring Bond Count as Drawn (r*) query feature to all the atoms of the ring:



# Amines

## Primary Aliphatic Amines

The following query matches primary aliphatic amines, butamides, thioamides, or enamines:

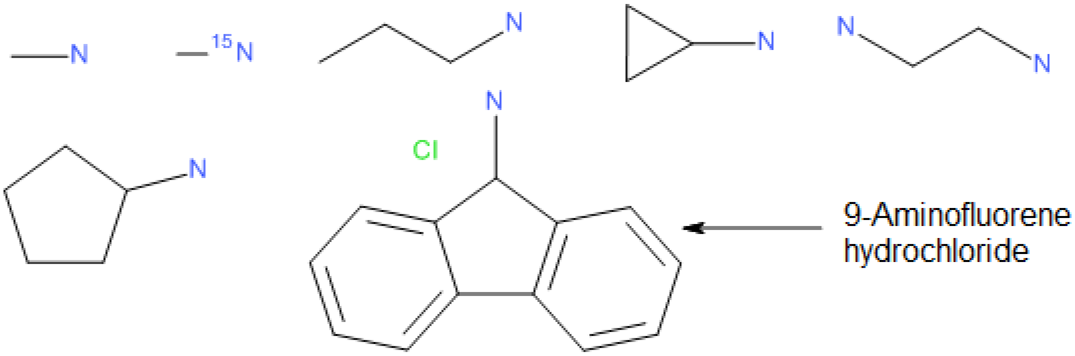


The four single bonds to the carbon that is attached to the amine nitrogen ensure that the carbon is saturated. This avoids the retrieval of compounds such as amides and enamines.

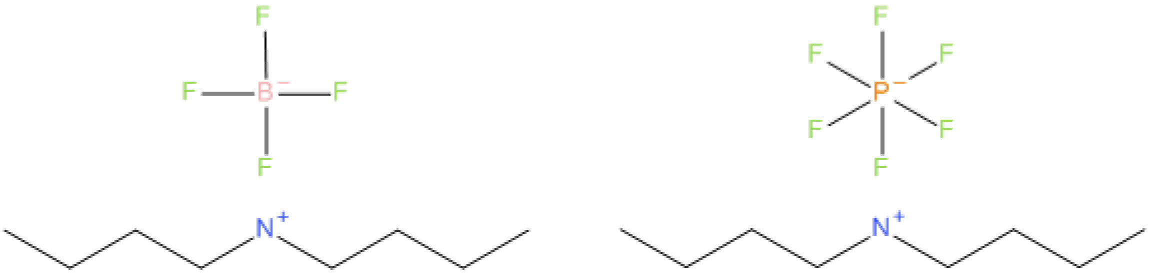
The hydrogen atoms on the nitrogen are drawn explicitly to prevent substitution of other atoms.

The Atom list query feature [C,H] specifies that only carbon or hydrogen can be attached to the carbon at the amine.

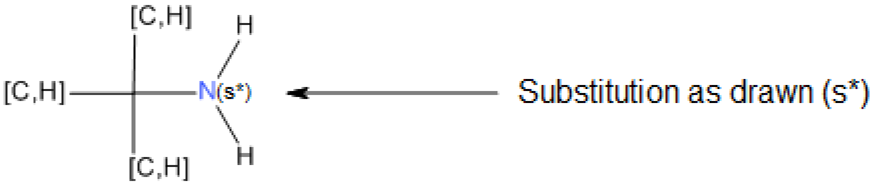
The following are examples of the structures that you retrieve:



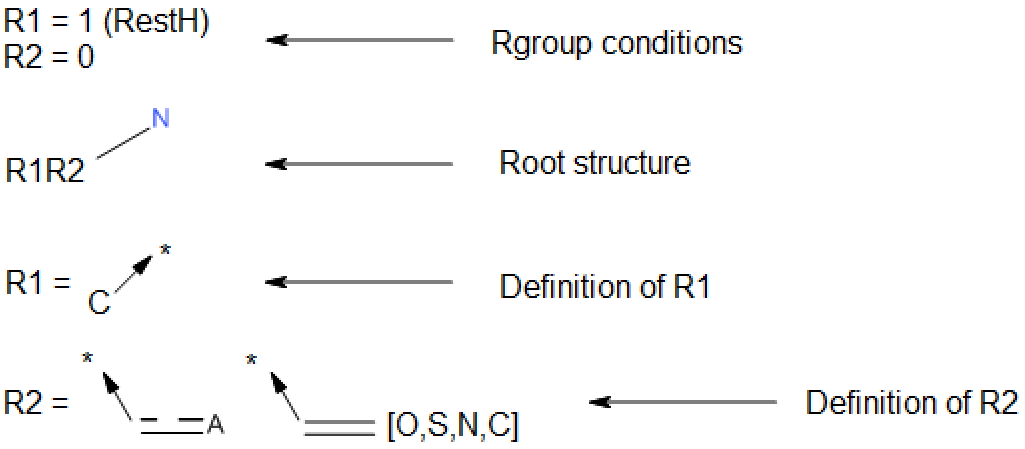
This query also matches a small number of secondary ammonium salts like the following examples:



To avoid retrieving these structures, add the Substitution Count as Drawn (s*) query feature to the nitrogen atom:



The following example uses a Markush query to specify a primary aliphatic amine:



The query matches primary aliphatic amines. It does not retrieve aromatic amines, carboxamides, thioamides, enamines, or nitrogen compounds with imino (-C=N) substituents, such as guanidines.

The root structure of the Rgroup specifies a single nitrogen atom with the Rgroup labels R1 and R2 attached by a single bond. Together, the two Rgroups specify the substituents that can attach to the nitrogen.

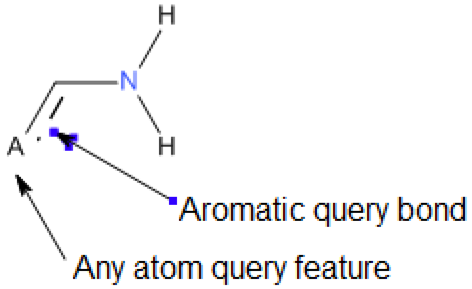
The Rgroup condition R1=1 specifies that the nitrogen atom must have exactly one attachment (Rgroup member) from R1. The (Rest H) Rgroup condition specifies that all other substituents of the nitrogen must be hydrogen atoms. R1 contains one member, a single carbon atom. Thus, R1 specifies primary amines, but not compounds such as hydrazines or hydroxylamines.

The Rgroup condition R2=0 specifies that none of the Rgroup members in the definition of R2 can attach to the nitrogen atom. Thus, R2 specifies the groups that are not retrieved by the query, as follows:

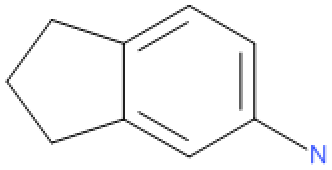
- The first Rgroup member in R2 contains a carbon atom attached to an Any atom (A) query feature through an Aromatic query bond. This specifies aromatic amines, so the query does not retrieve aromatic amines.
- The second Rgroup member in R2 contains a carbon atom attached to the Atom list query feature [O,S,N,C] through a double bond. This specifies carboxamides, thioamides, structures that contains imino groups and enamines, so the query does not retrieve these structures.

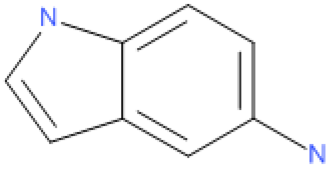
## Aromatic and Heteroaromatic Primary Amines

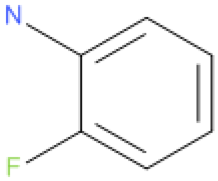
To retrieve all aromatic primary amines, including those with heteroatoms in the ring, use the Any atom (A) query feature and the Aromatic query bond. For example:

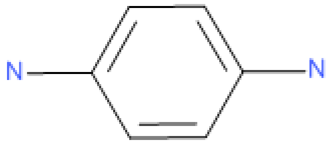


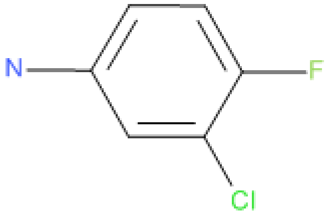
The following are examples of the structures that you retrieve:

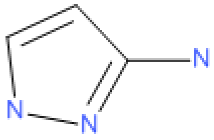


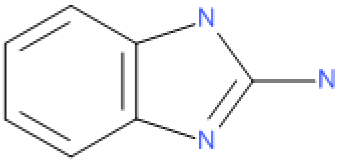




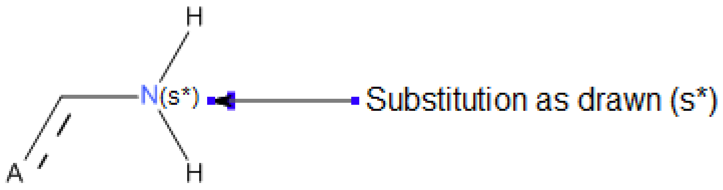






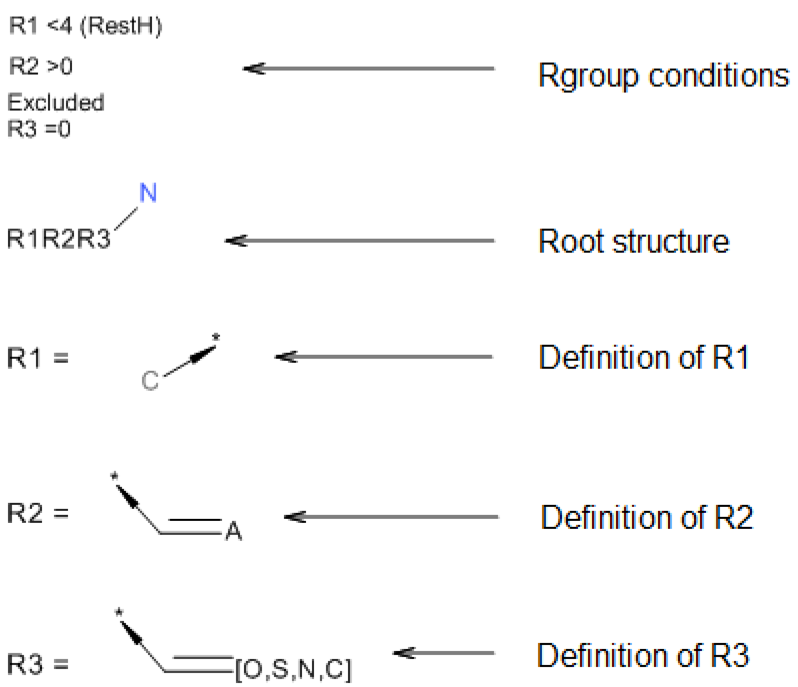


To avoid retrieving secondary ammonium salts, add the Substitution Count as Drawn (s*) query feature to the nitrogen atom:



## Aromatic Primary, Secondary and Tertiary Amines

The following query matches primary, secondary, and tertiary amines that contain at least one aromatic substituent. The query does not retrieve carboxamides, thioamides, enamines, or nitrogen compounds with imino (-C=N) substituents, such as guanidines.



The root structure of the Rgroup specifies a single nitrogen atom with the Rgroup labels R1, R2, and R3 attached by a single bond. Together, the three Rgroups specify the substituents that can attach to the nitrogen.

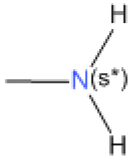
The Rgroup condition R1<4 specifies that the nitrogen atom can have one, two, or three attachments (Rgroup members) from R1. The (Rest H) Rgroup condition specifies that all other substituents of the nitrogen must be hydrogen atoms. R1 contains one Rgroup member, a single carbon atom. Thus, R1 specifies primary, secondary, and tertiary amines, but not compounds such as hydrazines or hydroxylamines.

The Rgroup condition R2>0 specifies that at least one of the attachments specified by R2 must be attached to the nitrogen atom. R2 contains one Rgroup member, a carbon atom attached to an Any atom (A) query feature through an Aromatic query bond. This combination of query features specifies any aromatic group. Thus, R2 specifies that at least one attachment of the nitrogen must be aromatic.

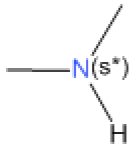
The Rgroup condition R3=0 specifies that none of the attachments specified by R3 can attach to the nitrogen atom. Thus, R3 specifies the groups that the query does not retrieve. R3 contains one Rgroup member, a carbon atom attached to the Atom list query feature [O,S,N,C] through a double bond. This member specifies carboxamides, thioamides, structures that contains imino groups, and enamines. Therefore, the query does not (R3=0) retrieve these structures.

## Amino Groups

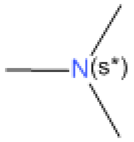
The following queries retrieve all compounds that contain amino groups, including aliphatic amines, aromatic amines, enamines, amides, and so on:



Unsubstituted amino group



N-substituted amino group



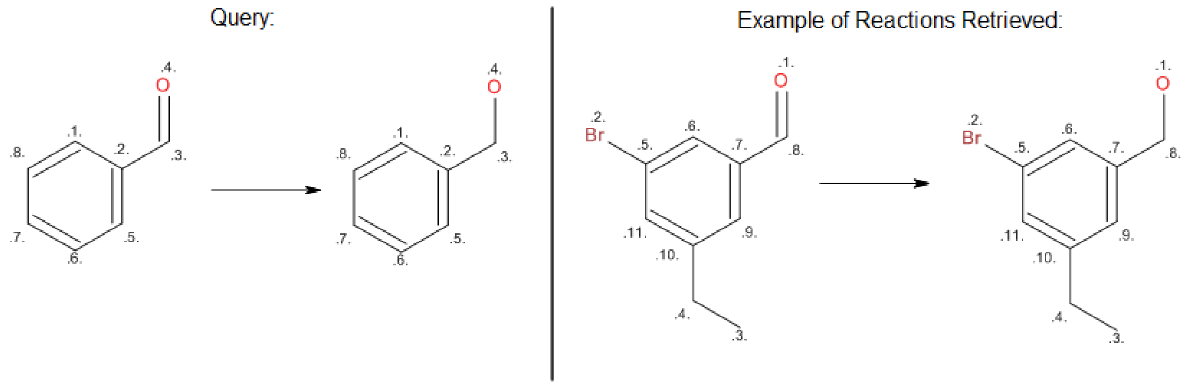
N,N-disubstituted amino group

For all three queries, the Substitution Count as Drawn (s*) query feature and explicitly drawn hydrogen atoms on the nitrogen atom ensure that no additional substitution can occur at the nitrogen.

# Chapter 11: Reaction Substructure Search (RSS)

## Definition of Reaction Structure Search

A reaction substructure (RSS) search finds reaction records in your database that contain your query as a reaction substructure wholly within a larger reaction. Your reaction substructure query is a two-dimensional representation of a portion of a reaction (a 2D reaction substructure) with mapped atoms and your choice of restrictions on the reacting centers. For example:



An atom-atom map on the reaction components in a query specifies exactly which atoms in the reactants correspond to the atoms in the products. For more information on atom-atom maps, see Reaction Mapping on page 34.

Your query can also contain restrictions on atoms and bonds. Allowed restrictions include:

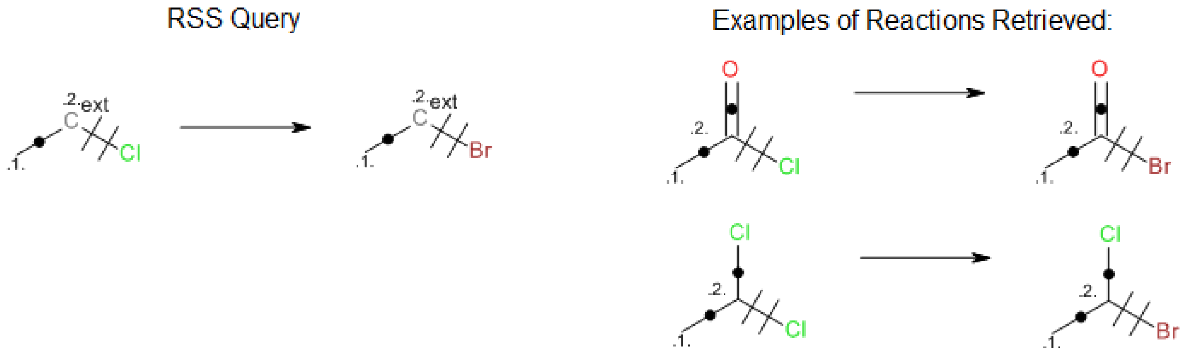
- Molecule atom and bond SSS query features. For information on molecule atom and bond query features, see Query Features on Atoms and Bonds on page 138.
- Atom-atom map numbers. For information on atom-atom mapping, see Reaction Mapping on page 34.
- Reaction atom and bond properties. For information on reaction atom and bond properties, see Stereoconfiguration Atom Properties in Mapped Reactions on page 37 and Properties of Bonds in Mapped Reactions on page 37.
- Reaction atom and bond query features. Like molecule query features, these properties can only be used in a substructure search and cannot be registered to a database. For more information on these query features, see Reaction Bond Query Features on page 192.

In this example, the reaction bond property Change (/) specifies that the bond type in the reactant must change in the product at the specified position.

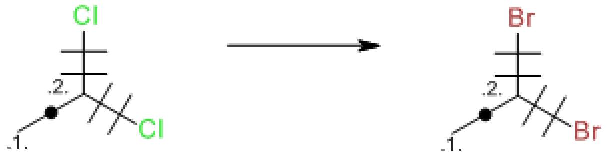
## Reaction Atom Query Feature

### Exact Change Atom Query Feature: .Ext.

Specifies that an atom changes exactly as you specify it in the reactants and products. You must add the **.Ext.** feature to each atom in the reaction that has the same atom-atom map number. For example, if the reaction contains one reactant and one product, then you must mark the corresponding atoms in both the reactant and product. The following example shows a query and the reactions that the query retrieves:

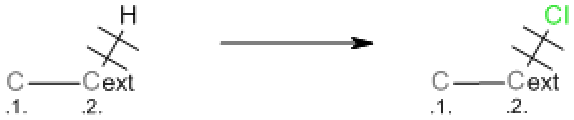


The following reaction is not retrieved, because another bond to the atom has been broken, in addition to the bond that you specified in your query.



NOT Retrieved

You must specify any changes in the number of hydrogens that are attached to this atom. For example, to show the chlorination of an ethyl group, draw a chlorine in the product that replaces an explicit hydrogen in the reactant. Both bonds must have the Make/Break property (//). For example:

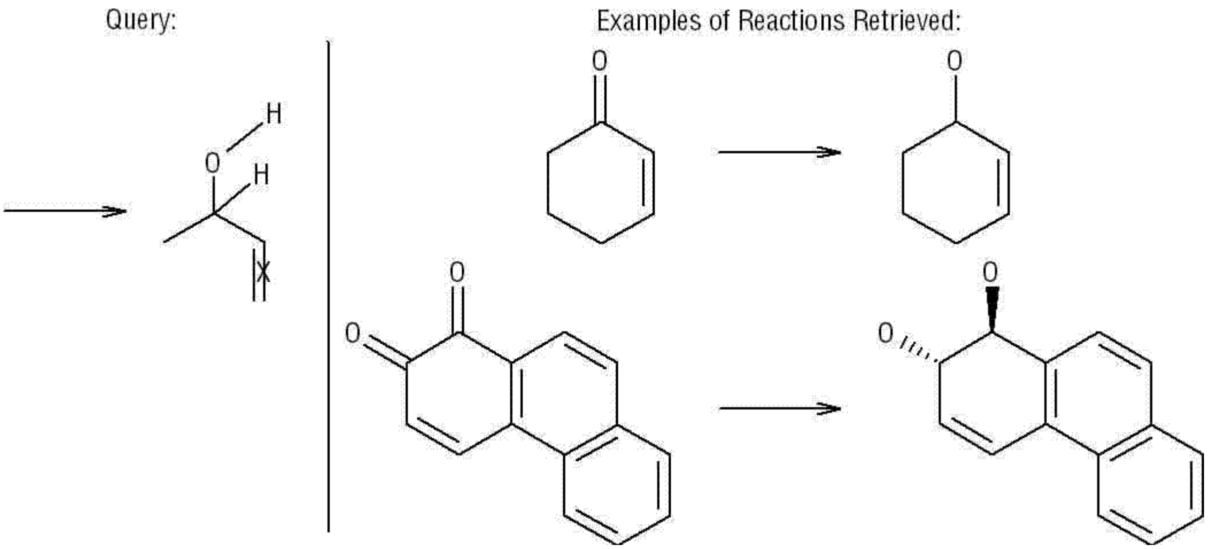


You must apply atom-atom maps to the reaction before you use this query feature. If you do not apply atom-atom maps to your reaction, the **.Ext.** query feature is ignored during the search

## Reaction Bond Query Features

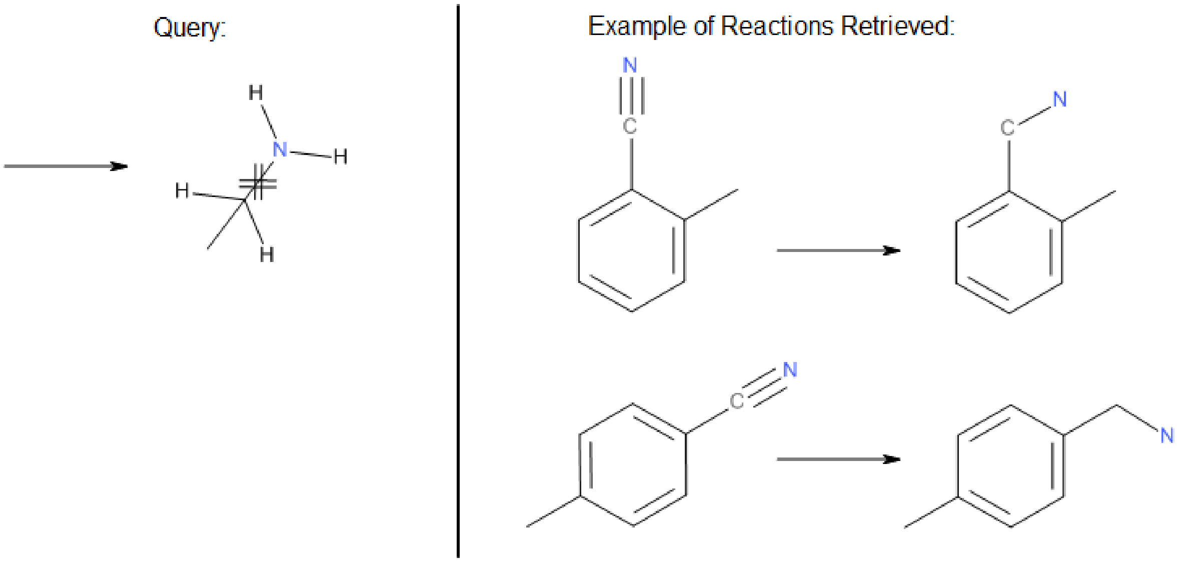
### Reaction Bond Query Feature: Not Center (X)

Specifies that a bond must not change in the records retrieved. For example, to specify the reduction of a carbonyl to an alcohol, the adjacent double bonds must not change as a result of the transformation:



## Reaction Bond Query Feature: Center (#)

Specifies that a bond changes in an unspecified way (can form, break, or change its bond type) in the records retrieved. For example, to specify the reduction of a nitrile to a primary amine:



# Using Reaction Atom and Bond Properties in RSS Searching

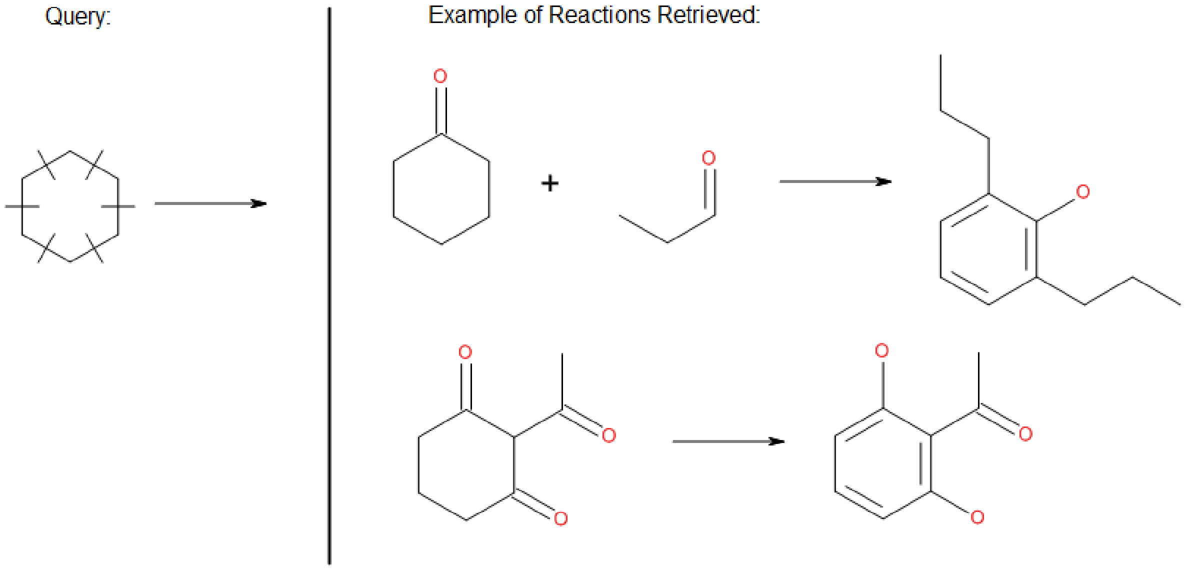
Like reaction query features, you can use reaction atom and bond properties to put restrictions on your reaction query. These properties are particularly useful if your query is a half-reaction, that is, solely reactants or solely products.

For more information on reaction atom properties, see Stereoconfiguration Atom Properties in Mapped Reactions on page 37.

For more information on reaction bond properties, see Properties of Bonds in Mapped Reactions on page 37.

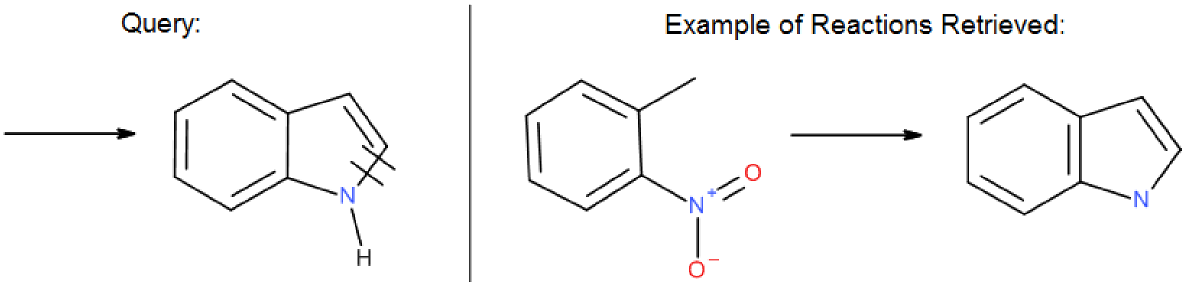
## Reaction Bond Property: Change (/)

Specifies that a bond changes solely its bond type in the records retrieved. For example, to change the bond type of all bonds in a cyclohexane reactant:



## Reaction Bond Property: Make/Break (//)

Specifies that a bond is either formed or broken in the records retrieved. For example, to specify the formation of indoles:

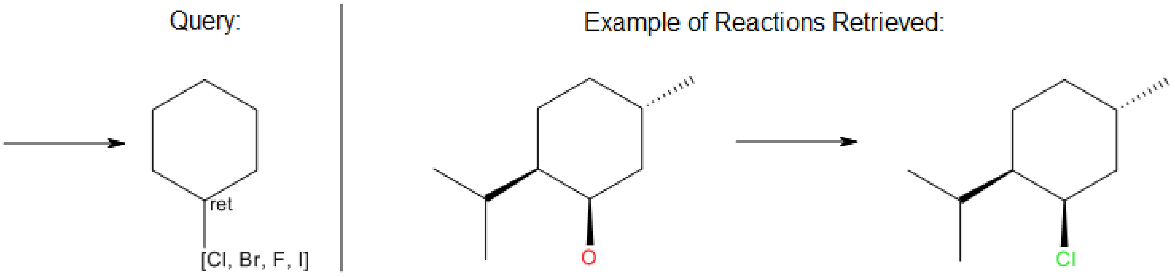


## Reaction Bond Property: No Change(•)

The No Change (•) bond property can be used in the same way as the Not Center (X) bond query property.

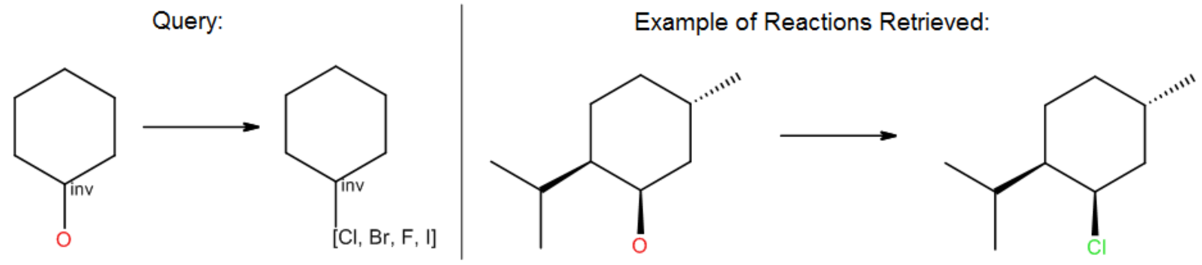
## Stereo Center Atom Property: .Inv.

Specifies that the stereo bond is inverted in the records retrieved. For example:



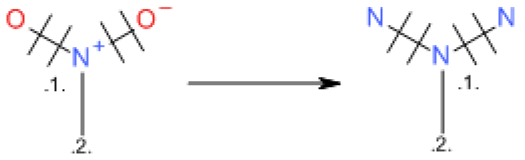
### Stereo Center Atom Property: .Ret.

Specifies that the stereo bond is retained (does not invert) in the records retrieved. For example:

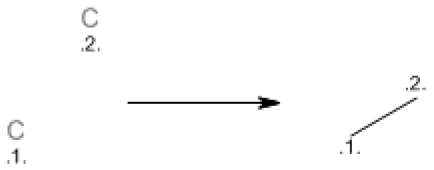


## Performance of RSS Queries

Chemists often want to find reactions in which a single functional group is transformed into another, such as the reduction of a nitro group to a primary amine:



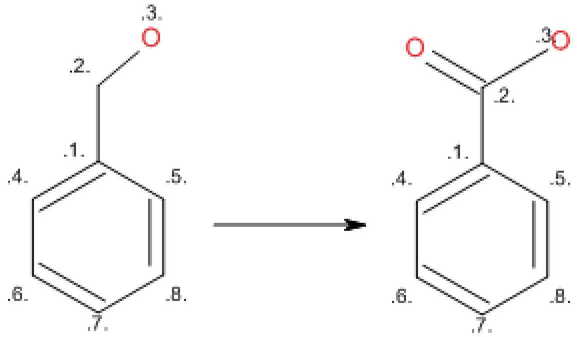
Queries like this often execute slowly, partly because of the small number of atoms and bonds that they contain. As an extreme case, a query that finds reactions in which a carbon-carbon single bond is formed can require several minutes to execute:



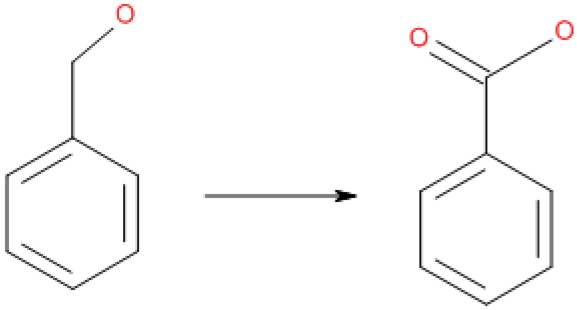
You can improve the usability of your searching application by providing a warning to the user if they draw a query that is likely to cause performance problems. The remainder of this section describes characteristics of RSS queries that can cause slow performance, with suggestions for detecting these conditions in a reaction query.

### Unmapped Reaction

Neglecting to map reactants to products in a reaction query can slow search performance considerably. In the following example, the mapped query executes almost 100 times faster than the unmapped query:



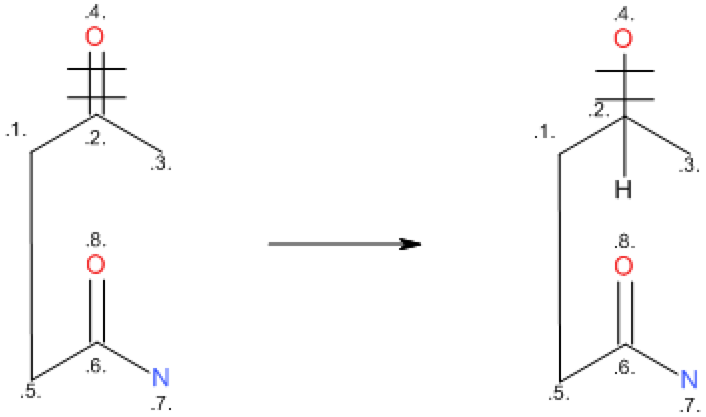
Mapped Query:
10x faster



Unmapped Query

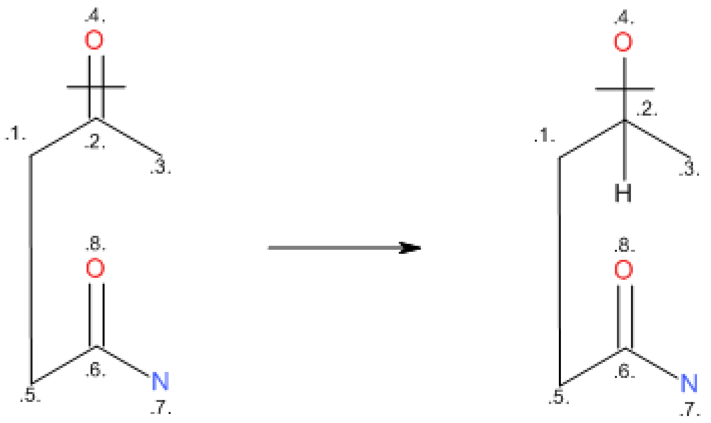
## Multifragment Reactants and Products

A reactant that contains multiple fragments can execute very slowly when each fragment maps to a separate fragment in a multifragment product A typical example is a query for the reduction of a ketone in the presence of an amide:



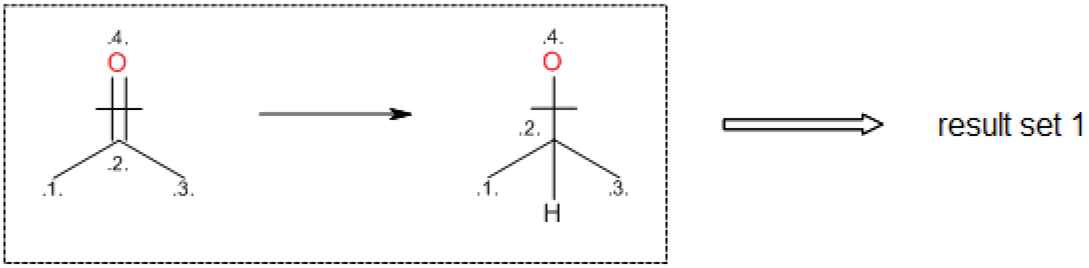
In this example, the atoms in the ketone fragment map solely to atoms in the alcohol fragment, while the amide reactant maps solely to the amide product.

Drawing bonds to create a single fragment in the reactant and product increases search speed by as much as five-fold:

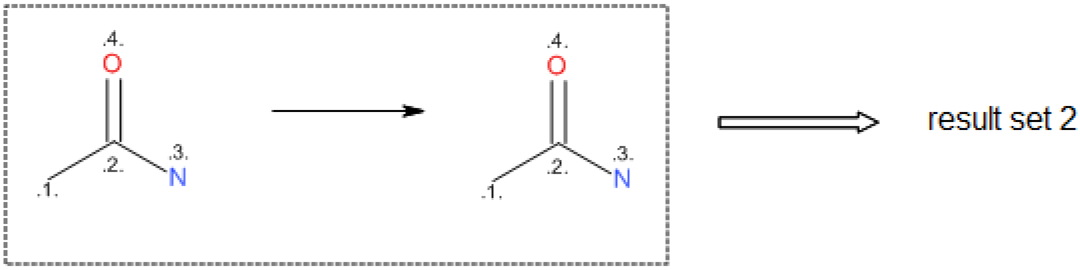


If you do not want to combine the fragments in the query, you can split the query into two sub-reactions, run a separate search for each query:

Query 1



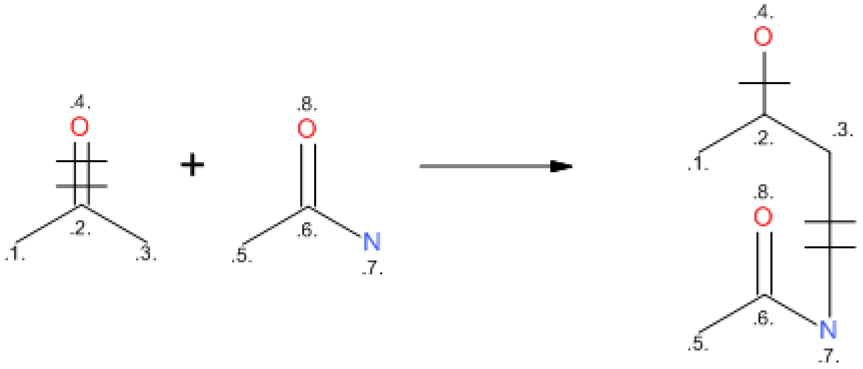
Query 2



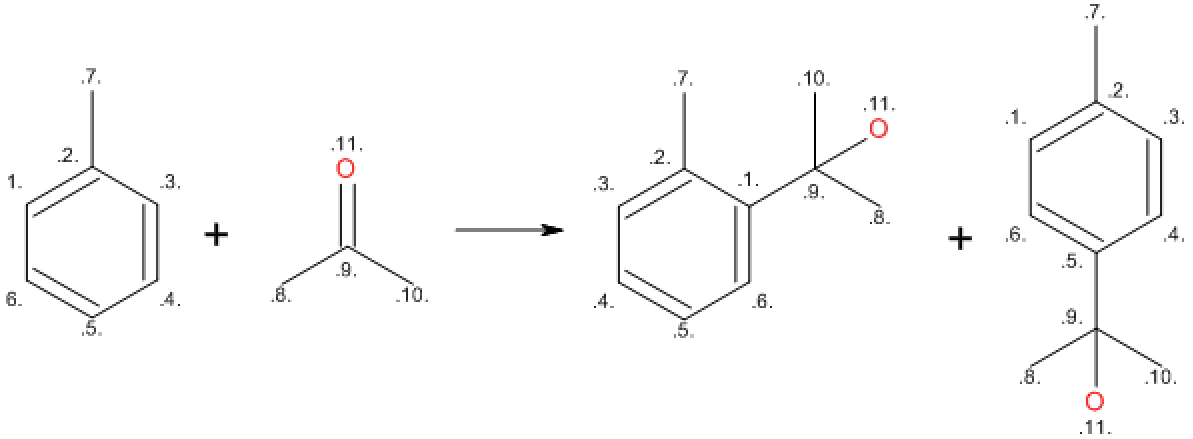
Then, use AND logic to combine the two result sets:

result set 1 AND result set 2 ⟹ result set 3

Reactions of the form A+B->C do not cause this problem, because the product contains atoms from both reactants:



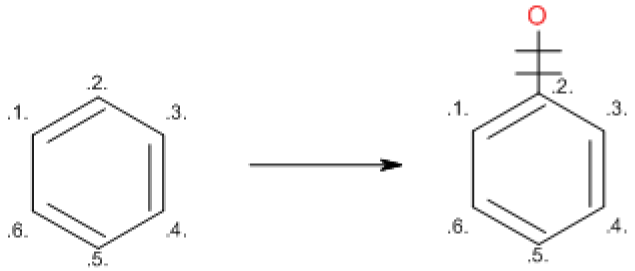
A variation of this form is A+B→C+C', where C and C' are compositional isomers or stereoisomers.

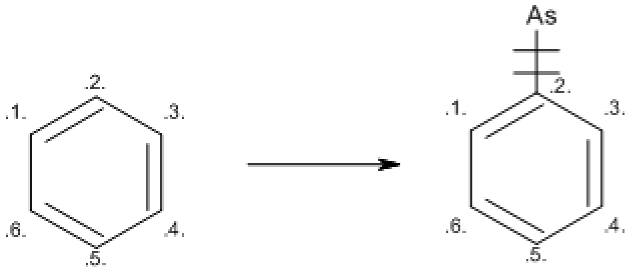


## Nonspecific Queries

Nonspecific queries not only retrieve an unmanageably large number of hits, but might also run very slowly. The following characteristics can make a query nonspecific:

- Elements solely from the first two rows of the periodic table. In the following example, the second query executes over 10 times faster than the first:





- A small number of atoms and bonds. Example: A reactant that contains only two carbon atoms, as shown in Performance of RSS Queries on page 195.

### Atom and Bond Query Features

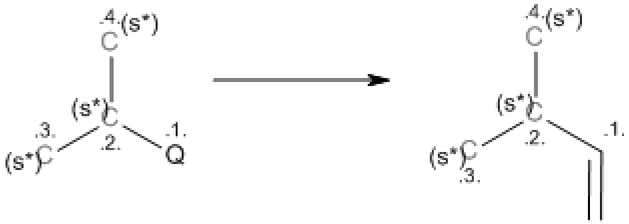
Some query feature make a query less specific, while others make it more specific. Query features that make a query less specific and hence might cause slow search performance include:

- Any atom except H (A)
- Any atom except H or C (Q)
- Atom list
- Any bond
- Single/double bond, Single/aromatic bond, or Double/aromatic bond

Query features that make a query more specific and therefore enhance search performance include:

- Substitution count: s*, s0, s1, s2, s3, s4, s6, or adding a bond with an explicit hydrogen atom
- Ring bond count: r*, r0, r2, r3, r4
- Bond topology: Ring bond only (Rn) or Chain bond only (Ch)
- Unsaturated atom (u)
- NOT atom list

In the following example, the Substitution as Drawn query feature (s*) on the carbon atoms causes the query to execute almost 100 times faster than the same query without the s* query feature:



## Enhanced RSS Search Algorithm

An enhanced RSS search algorithm was introduced in BIOVIA Direct 5.1 and BIOVIA ISIS/Host 5.1. The enhanced algorithm corrects the following limitations of earlier algorithms:

- The Exact Change (.Ext.) query feature must be present in at least one mapped atom in the reactant and in the corresponding mapped atom(s) in the product. The previous RSS algorithm failed to detect this error, and could therefore produce false hits. The new algorithm can detect such errors so that the query produces no false hits.

- The previous RSS algorithm often handled atoms with the .Ext. query feature incorrectly if the number of attached hydrogens changed between reactant and product.
- The previous RSS algorithm sometimes retrieved reactions in which the atom-atom mapping in the target did not match atom-atom mapping in the query, thereby producing false hits.
- The enhanced RSS algorithm performs more extensive checking for errors in reacting bond properties (reacting center errors) than the previous RSS algorithm, based on atom-atom mapping relationships. Consequently, the enhanced RSS algorithm might not match a reaction to itself if the reaction (either query or target, or both) contains reacting center errors. If you find that the enhanced RSS algorithm cannot match a reaction to itself, examine the reaction and correct the errors. For information on reaction mapping and reaction bond properties, seeReaction Representation on page 34.

A detailed description of the new RSS algorithm (RSS3) and its comparison with previous RSS methods can be found in the following paper:

Lingran Chen, James G. Nourse, Bradley D. Christie, Burton A. Leland, David L. Grier. "Over 20 Years of Reaction Access Systems from MDL: A Novel Reaction Substructure Search Algorithm," *J. Chem. Inf. Comput. Sci.* **(2002)**, 42(6), 1296-1310.

# Chapter 12:
# NEMA Key Search

Reliability and speed of searching is greatly simplified if a canonical name can be generated for a structure, and the process is further simplified if that can be machine generated. NEMA was developed to meet these needs.

NEMA (New Extended Morgan Algorithm) is a method of generating a unique numeric "name" for a molecule. The NEMA key (or NEMAKEY) is a compressed version of the complete NEMA "name". The key is 30 characters long, and is composed of numbers and letters.

BIOVIA Draw can generate a NEMA key for structure it renders. For example, benzene has the key:

```
MMS8JV3UB9DBYU1ZVXWPAT1DPT4H74
```

- NEMA key values for any given structure are subject to change between product releases. Compare the NEMA keys only if they were generated by the same version of NEMA.
- NEMA keys do not contain information about enhanced stereochemistry, haptic (pi) bonds, Markush bonds, polymer Sgroups, data Sgroups, or Rgroups. Do not use NEMA keys to compare two molecules containing these features.
- For structures with template atoms (such as some biologics), the same key is generated regardless of the expanded or contracted state of the template atoms.

**See also**

*NEMA Key Based Exact Match Searching*, a free download available at download at http://accelrys.com/products/pdf/exact-match-searching.pdf

*BIOVIA Direct Developers Guide* > Using BIOVIA Direct > NEMAKEY searching and key generation

*BIOVIA Direct Reference Guide* > Molecule-Specific Operators and Functions > Molecule-Specific Functions:

- mdlaux.molnemakey
- molnemakey
- mdlaux.rownemakey

*BIOVIA Draw Help* - search for *NEMA*.

# Chapter 13:
# Registration to Chemical Databases

## Registration of Molecules

### Structural Features and Registration

The structural features that are described in Molecule Representation on page 3 can be registered to a molecule database. Structural features that cannot be registered are:

- Attached data that contains a search operator. For more information, see Substructure Search of Attached Data on page 166.
- Mixture brackets that do not contain at least one component bracket. Unordered mixtures can contain only unordered components, and ordered mixtures can contain only numbered components. Other structures can be used in substructure-search queries, but cannot be registered. For details, see Substructure Search of Mixtures on page 167.
- Polymers: Polymer structures that contain the Anypolymer (anyp) bracket. For more information, see Substructure Search of Polymers on page 170.

In a flexmatch search, the presence of these features in either the query, the target, or both cause flexmatch search to fail.

Beginning with Version 2021, Pipeline Pilot Client Chemistry and BIOVIA Direct support registration and searching of the following atom query features (see Registration of Query Features on page 202) which, in the past, were restricted to use only in substructure queries:

- A, Q, M, X, R, Z, atom lists such as [C,N,O] or NOT[C,N,O]

Positive logic atom lists containing hydrogen are not registerable.

### Registration of Query Features

Pipeline Pilot Chemistry and BIOVIA Direct support registration and searching of the following atom query features:

- A, Q, M, X, R, Z, atom lists such as [C,N,O] or NOT[C,N,O]

Positive logic atom lists containing hydrogen are not registerable.

When a molecule contains one of these features and it is used in a FLEXMATCH search, the atom feature only matches itself. Thus Q matches Q and [C] matches [C] but not C.

When a molecule contains one of these features and it is used in a substructure search, the query feature matches an atom in the target that is logically a subset of the query.

The R atom is special, it is effectively removed from the query leaving an open valence; it matches any atom including hydrogen.

#### Query Target

| Query Atom | Property |
|---|---|
| R | Any atom including H, also including any of the newly registerable features |
| Z | Any atom except H, including any of the other newly registerable features |

| Query Atom | Property |
|---|---|
| A | Any atom except R and H, including any of the other newly registerable features |
| Q | Any atom except R, A, C and H, including M, X and atom lists which do not contain C |
| M | Any metal atom, M, and positive logic atom lists containing only metals |
| X | Any halogen atom, X, and positive logic atom lists containing only halogens |
| [N,O,S] | N, O, S, [N,O], [N,S], [O,S] and [N,O,S] |
| NOT [N,O,S] | Any atom except N, O, S, any positive logic atom list which does not contain N, O or S, and any negative logic atom lists which includes N, O, and S |

BIOVIA Direct now respects several substructure search flags. Add these flags to the third argument of a substructure query, or set them globally for the cartridge installation.

| Flag | Use |
|---|---|
| `InterpretRAtomsLiterally` | Causes R atoms in the query to only match R atoms in the target. |
| `InterpretZAtomsLiterally` | Causes Z atoms in the query to only match Z atoms in the target. |
| `InterpretXAtomsLiterally` | Causes X atoms in the query to only match X atoms in the target. |
| `InterpretQueryAtomsLiterally` | Causes A, Q, M, or X atoms in the query to only match the corresponding A, Q, M, or X atoms in the target. |

## Definition of Duplicate Structure

You can configure an BIOVIA chemical database to either allow or prohibit the registration of duplicate chemical structures. For example:

- Your data model might require that you register different batches or samples of the same compound, each of which has different data associated with it. In this case, you need to allow registration of duplicate structures.
- Your data model might require that each chemical structure that is stored in the database be unique, so that you have solely one record of each compound with its associated data. Different batches or samples of a compound and their associated data are stored in subtables. In this case, you need to prohibit registration of duplicate structures.

If your data model requires unique structures, you also need to specify the definition of a duplicate structure, that is, you need to specify which structural features should be used to determine whether a duplicate of a structure that is being registered already exists in the database. For example, you might want to register the tautomers of a compound as distinct structures, or you might want to consider tautomers of a structure to be equivalent.

The procedure for allowing duplicate structures depends upon the type of database, see BIOVIA Direct Databases on page 204.

### BIOVIA Direct Databases

By default, BIOVIA Direct molecule and reaction databases allow registration of duplicate structures.

To prohibit registration of duplicate structures, use Oracle SQL*Plus command `ALTER INDEX … REBUILD PARAMETERS('UNIQUE=options')`, where the options are flexmatch switches. For example:

```
SQL> ALTER INDEX CORPDB_MOL_MDLIX REBUILD PARAMETERS
('UNIQUE="match=all/allow_timeout"');
```

The definition `match=all` corresponds to the same switches as the most restricted flexmatch search:

```
"FRA,HYD,MAS,MET,RAD,SAL,STE,VAL,BON,ION,CHA,DAT,MIX,POL,TYP,MSU,END"
```

or

```
"IGNORE=NONE"
```

For more information, see How to Specify Switches on page 107.

To change the definition of a duplicate structure so that tautomers are perceived as equivalent structures, remove the BON switch from the definition of duplicate structure and add the TAU switch:

```
"FRA,HYD,MAS,MET,RAD,SAL,STE,VAL,TAU,ION,CHA,DAT,MIX,POL,TYP,END,MSU"
```

or

```
"IGNORE=BON"
```

This definition of a duplicate structure prohibits registration of tautomers of structures that already exist in the database, because the tautomers of a structure are equivalent.

**IMPORTANT!** Be careful when you decide whether to change the definition of a duplicate structure in the database. When used to define duplicate structures for registration, the flexmatch switches have the opposite result as in searching. For example, if you set the STE switch Off in a search, you retrieve your query structure and its stereoisomers. If you set the STE switch Off in the duplicate structure definition, then you cannot register stereoisomers separately, because all stereoisomers are perceived as the same structure. Thus, to register stereoisomers as separate structures, do not set STE Off in the duplicate definition.

For more information and detailed procedures, see "Altering Duplicate Checking Behavior" in the chapter, "Maintaining Molecule Tables and Indexes" in the *BIOVIA Direct Administration Guide*.

## Registration of Reactions

Not all structural features of a reaction can be registered to a reaction database. The structural features that are described in Molecule Representation on page 3 and Reaction Representation on page 34 can be registered to a reaction database. Structural features that cannot be registered are:

- Molecule features that cannot be registered. For details, see Structural Features and Registration on page 202.
- The Exact Change (.Ext.) reaction atom query feature. For more information, see Exact Change Atom Query Feature: .Ext. on page 191.
- The bond query features Not Center (X) and Center (#). For more information, see Reaction Bond Query Features on page 192.

## Molecular Weight of Standard Structures

You can provide your users with the ability to search by molecular weight in both BIOVIA Direct databases and relational-chemical databases. In both cases, however, the routine that these databases use to calculate molecular weight does not work correctly for certain types of structures, for example:

***Biopolymers*** that use *atoms for condensed representation of biopolymer residues. Since a *atom has an atomic weight of zero, the molecular weights of biopolymer residues cannot be calculated. For more information on the *atom representation, see Condensed Representation of Biopolymers on page 54.

Unlike the routine that BIOVIA supplies for calculating molecular weights in databases, the Calculator command in the BIOVIA Draw Chemistry menu can correctly calculate molecular weights for biopolymer structures that use the *atom representation. The Calculator command uses information in the standard abbreviation definitions file to calculate molecular weight, but the routine in the database uses solely the information in the Ptable. For more information on the *standard abbreviation definitions file*, see Creating and Enforcing Conventions for Biopolymer Representation on page 71.

***Artificial polymers.*** The molecular weights of these substances are generally specified as a range, or as a number-average or weight-average. Because the representation of a polymer contains only one instance of the structural repeating unit, the routine cannot use it to calculate an accurate molecular weight.

For structures like these, create a separate column in the molecule table for the molecular weight and use a suitable method to generate the correct molecular weight. For more information on searching by molecular weight, see the BIOVIA Direct Administration Guide.

# Chapter 14:
# Customizing the BIOVIA Ptable

## Introduction

The BIOVIA periodic table, or Ptable, defines the atom symbols that you can register to your database and provides for optional overrides to the atomic weights defined in BIOVIA Foundation.

The Ptable is standardized and synchronized across all BIOVIA products. For example, the Ptable is part of local BIOVIA Draw installations (for example: C:\Program Files (x86)\BIOVIA\BIOVIA Draw 2017 R2\PTable\PeriodicTable.txt).

You can customize the default BIOVIA Ptable to add your own pseudoatom symbols, or to change the atomic weights that are assigned an atom or pseudoatom.

## Changing the Current Ptable

Perform the following steps to replace your current Ptable with a revised Ptable.

1. Make a copy of your current Ptable for future reference from the installed location.
2. Follow the guidelines in the section Guidelines for Customizing the Ptable File on page 206, to customize your Ptable.
3. Load the revised Ptable as instructed for your application. For example, in BIOVIA Direct:
   - Use Oracle SQL*Plus to load the revised Ptable to a text file. For more information, see the chapter, "Setting the Chemical Environment" in the *BIOVIA Direct Administration Guide*.

## Guidelines for Customizing the Ptable File

The BIOVIA periodic table, or Ptable, defines the atom symbols that you can register to your database and provides for optional overrides to the atomic weights defined in BIOVIA Direct.

### Format of the Ptable

Comments are allowed in the Ptable. They must have a '#' as the first character of the line. Any text following the '#' is ignored.

Each element entry in the Ptable includes 11 items of information on one line. Each item is separated by space from the previous item.

- Atom symbol: must start with a letter and can be followed by up to two more letters or numbers.
- Atom name (not used by BIOVIA Direct)
- Van der Waals radius: in angstroms (not used by BIOVIA Direct)
- Covalent radius: in angstroms (not used by BIOVIA Direct)
- First legal oxidation state (1..n, or 0 if none)
- Second legal oxidation state (1..n, or 0 if none)
- Third legal oxidation state (1..n, or 0 if none)
- Fourth legal oxidation state (1..n, or 0 if none)
- Fifth legal oxidation state (1..n, or 0 if none)

- Atomic weight
- Exact mass of the most common isotope

If the atom symbol is specified as one of the elements H through Lr, only the Van der Waals radius, covalent radius, atomic weight and exact mass of the most common isotope are used to override the built-in AEP values. The atom name and oxidation states are not overridden.

When specifying pseudoatom symbols, if a symbol might stand alone, set all oxidation states to zero. Otherwise, set them to the values 1 through 5. This prevents implicit hydrogens from being added.

## Rules for Adding or Modifying Entries

**IMPORTANT!** When you change a periodic table definition entry, the new table can have only new atom symbols. You cannot remove or change existing atom symbols.

Observe these rules when adding or modifying symbols:

- You can only alter the atomic weight and exact mass of most common isotope for atoms H through Lr. To do this, add a line with the appropriate atomic symbol and all of the required fields, replacing the atomic weight and exact mass of most common isotope fields with the desired numbers. For example to change the value for carbon from its default of 12.0107 to 12.011 you would add the following line:

  C Carbon 1.7 0.77 0 0 0 0 0 12.011 12

- Be careful when you add pseudoatoms to the Ptable.

  **Note:** Do not confuse pseudoatoms with abbreviated structures. An abbreviated structure contains within it all the atoms and bonds of the underlying structure, that is, it contains the complete connection table, or CTAB. In contrast, a pseudoatom is a single atom with no underlying connection table. For information on abbreviated structures, see Abbreviated Structures on page 30.

- Restrict atomic symbols and pseudoatoms to three characters or less. The first character must be alphabetic (A-Z, a-z). Other characters can be alphanumeric (A-Z, a-z, 0-9). Alphabetic characters are not case sensitive. For example, T and t are treated as the same character.
- If you customize your Ptable so that it contains atom symbols with numbers in them, you cannot retrieve the symbols unless you enclose the string with the non-alphabetic characters in double quotes. Thus, for formula search queries such as C5 H13 Bg1, the Bg1 must be double-quoted:

  -
    ```
    select molregno,molformula from 0 where molformula = 'C5 H10 "Bg1"';
    Found 1 MOL records MOL1(MOLREGNO): 15 MOL1(MOLFORMULA): "C5 H10 "Bg1""
    ```
  - To avoid these complications, BIOVIA recommends that you avoid using numbers in atom symbols.
- Except for biopolymer residues that use the pseudoatom representation (see Biopolymer Representation on page 53), a pseudoatom cannot be expanded to the full group it represents or any part of it. For example, if you define the symbol Ts to represent a tosylate, a substructure-search query that contains an explicitly drawn tosylate group, or any fragment of the tosylate group will not hit a molecule registered in the database containing the Ts symbol.
- Do not use symbols that represent reserved atom types. Symbols that represent reserved types include elements, and the symbols A, L, LP, M, Q, R, R1 through R32, R#, X, Z, and *.
- Do not use symbols that are identical to template names (for example, Et or Ph).
- Do not remove or reorder existing entries in the file.

## File Listing for the Default Ptable

The following is the 2021 version of `PeriodicTable.txt`.

```
#
# Each line has the following fields, for standard elements
# only the two radii and two weights can be modified:
#
# symbol name VDW_radius covalent_radius 5X_legal_oxidation_states atomic_
weight exact_mass_most_common_isotope
#
# Standard elements
# Lines are comments unless radii or weights are modified
  H   Hydrogen      1.09 0.32 1 0 0 0 0   1.007940000   1.007825032
  He  Helium        1.40 1.60 0 0 0 0 0   4.002602000   4.002603250
  Li  Lithium       1.82 1.31 1 0 0 0 0   6.941000000   7.016004000
  Be  Beryllium     1.53 0.91 2 0 0 0 0   9.012182000   9.012182100
  B   Boron         1.92 0.82 3 0 0 0 0  10.811000000  11.009305500
  C   Carbon        1.70 0.77 4 2 0 0 0  12.010700000  12.000000000
  N   Nitrogen      1.55 0.75 3 0 0 0 0  14.006700000  14.003074010
  O   Oxygen        1.52 0.73 2 0 0 0 0  15.999400000  15.994914620
  F   Fluorine      1.47 0.72 1 0 0 0 0  18.998403200  18.998403200
  Ne  Neon          1.54 1.12 0 0 0 0 0  20.179700000  19.992440180
  Na  Sodium        2.27 1.66 1 0 0 0 0  22.989770000  22.989769670
  Mg  Magnesium     1.73 1.36 2 0 0 0 0  24.305000000  23.985041900
  Al  Aluminum      1.84 1.18 3 0 0 0 0  26.981538000  26.981538440
  Si  Silicon       2.10 1.10 4 0 0 0 0  28.085500000  27.976926530
  P   Phosphorus    1.80 1.06 3 5 0 0 0  30.973761000  30.973761510
  S   Sulfur        1.80 1.02 2 4 6 0 0  32.065000000  31.972070690
  Cl  Chlorine      1.75 0.99 1 3 5 7 0  35.453000000  34.968852710
  Ar  Argon         1.88 1.54 0 0 0 0 0  39.948000000  39.962383120
  K   Potassium     2.75 2.06 1 0 0 0 0  39.098300000  38.963706900
  Ca  Calcium       2.31 1.74 2 0 0 0 0  40.078000000  39.962591200
  Sc  Scandium      2.11 1.44 3 0 0 0 0  44.955910000  44.955910200
  Ti  Titanium      2.00 1.32 3 4 0 0 0  47.867000000  47.947947100
  V   Vanadium      2.00 1.21 5 4 3 2 0  50.941500000  50.943963700
  Cr  Chromium      2.00 1.26 6 3 2 0 0  51.996100000  51.940511900
  Mn  Manganese     2.00 1.26 7 6 4 2 3  54.938049000  54.938049600
  Fe  Iron          2.00 1.26 2 3 0 0 0  55.845000000  55.934942100
  Co  Cobalt        2.00 1.21 2 3 0 0 0  58.933200000  58.933200200
  Ni  Nickel        1.63 1.15 2 3 0 0 0  58.693400000  57.935347900
  Cu  Copper        1.40 1.17 2 1 0 0 0  63.546000000  62.929601100
  Zn  Zinc          1.39 1.25 2 0 0 0 0  65.409000000  63.929146600
  Ga  Gallium       1.87 1.26 3 2 0 0 0  69.723000000  68.925581000
  Ge  Germanium     2.11 1.27 4 0 0 0 0  72.640000000  73.921178200
  As  Arsenic       1.85 1.20 3 5 0 0 0  74.921600000  74.921596400
  Se  Selenium      1.90 1.17 2 4 6 0 0  78.960000000  79.916521800
  Br  Bromine       1.85 1.14 1 3 5 7 0  79.904000000  78.918337600
  Kr  Krypton       2.02 1.60 0 0 0 0 0  83.798000000  83.911507000
  Rb  Rubidium      3.03 2.21 1 0 0 0 0  85.467800000  84.911789300
  Sr  Strontium     2.49 1.86 2 0 0 0 0  87.620000000  87.905614300
  Y   Yttrium       2.00 1.66 3 0 0 0 0  88.905850000  88.905847900
  Zr  Zirconium     2.00 1.41 4 0 0 0 0  91.224000000  89.904703700
  Nb  Niobium       2.00 1.31 3 5 0 0 0  92.906380000  92.906377500
```

```
Mo  Molybdenum    2.00 1.30 2 3 4 5 6  95.940000000  97.905407800
Tc  Technetium    2.00 1.27 7 0 0 0 0  98.000000000  97.907216000
Ru  Ruthenium     2.00 1.16 2 3 4 6 8 101.070000000 101.904349500
Rh  Rhodium       2.00 1.25 2 3 4 0 0 102.905500000 102.905504000
Pd  Palladium     1.63 1.26 2 4 0 0 0 106.420000000 105.903483000
Ag  Silver        1.72 1.34 1 0 0 0 0 107.868200000 106.905093000
Cd  Cadmium       1.58 1.48 2 0 0 0 0 112.411000000 113.903358100
In  Indium        1.93 1.44 3 0 0 0 0 114.818000000 114.903878000
Sn  Tin           2.17 1.40 2 4 0 0 0 118.710000000 119.902196600
Sb  Antimony      2.06 1.41 3 5 0 0 0 121.760000000 120.903818000
Te  Tellurium     2.06 1.37 2 4 6 0 0 127.600000000 129.906222800
I   Iodine        1.98 1.33 1 3 5 7 0 126.904470000 126.904468000
Xe  Xenon         2.16 1.31 0 0 0 0 0 131.293000000 131.904154500
Cs  Cesium        3.43 2.46 1 0 0 0 0 132.905450000 132.905447000
Ba  Barium        2.68 2.01 2 0 0 0 0 137.327000000 137.905241000
La  Lanthanum     2.00 1.81 3 0 0 0 0 138.905500000 138.906348000
Ce  Cerium        2.00 1.65 3 4 0 0 0 140.116000000 139.905434000
Pr  Praseodymium  2.00 1.71 3 4 0 0 0 140.907650000 140.907648000
Nd  Neodymium     2.00 1.64 3 0 0 0 0 144.240000000 143.910083000
Pm  Promethium    2.00 1.63 3 0 0 0 0 145.000000000 144.912744000
Sm  Samarium      2.00 1.62 2 3 0 0 0 150.360000000 151.919728000
Eu  Europium      2.00 1.85 3 3 0 0 0 151.964000000 152.921226000
Gd  Gadolinium    2.00 1.61 3 0 0 0 0 157.250000000 157.924101000
Tb  Terbium       2.00 1.59 3 4 0 0 0 158.925340000 158.925343000
Dy  Dysprosium    2.00 1.59 3 0 0 0 0 162.500000000 163.929171000
Ho  Holmium       2.00 1.58 3 0 0 0 0 164.930320000 164.930319000
Er  Erbium        2.00 1.57 3 0 0 0 0 167.259000000 165.930290000
Tm  Thulium       2.00 1.56 2 3 0 0 0 168.934210000 168.934211000
Yb  Ytterbium     2.00 1.56 2 3 0 0 0 173.040000000 173.938858100
Lu  Lutetium      2.00 1.56 3 0 0 0 0 174.967000000 174.940767900
Hf  Hafnium       2.00 1.41 4 0 0 0 0 178.490000000 179.946548800
Ta  Tantalum      2.00 1.31 5 0 0 0 0 180.947900000 180.947996000
W   Tungsten      2.00 1.30 2 3 4 5 6 183.840000000 183.950932600
Re  Rhenium       2.00 1.28 1 2 4 6 7 186.207000000 186.955750800
Os  Osmium        2.00 1.26 2 3 4 6 8 190.230000000 191.961479000
Ir  Iridium       2.00 1.27 2 3 4 6 0 192.217000000 192.962924000
Pt  Platinum      1.72 1.30 2 4 0 0 0 195.078000000 194.964774000
Au  Gold          1.66 1.34 1 3 0 0 0 196.966550000 196.966552000
Hg  Mercury       1.55 1.49 1 2 0 0 0 200.590000000 201.970626000
Tl  Thallium      1.96 1.48 1 3 0 0 0 204.383300000 204.974412000
Pb  Lead          2.02 1.47 2 4 0 0 0 207.200000000 207.976636000
Bi  Bismuth       2.07 1.46 3 5 0 0 0 208.980380000 208.980383000
Po  Polonium      1.97 1.46 2 4 0 0 0 209.000000000 208.982416000
At  Astatine      2.02 1.45 1 3 5 7 0 210.000000000 209.987131000
Rn  Radon         2.20 1.90 0 0 0 0 0 222.000000000 222.017570500
Fr  Francium      3.48 1.80 1 0 0 0 0 223.000000000 223.019730700
Ra  Radium        2.83 2.00 2 0 0 0 0 226.000000000 226.025402600
Ac  Actinium      2.00 1.81 3 0 0 0 0 227.000000000 227.027747000
Th  Thorium       2.00 1.65 4 0 0 0 0 232.038100000 232.038050400
Pa  Protactinium  2.00 1.66 4 5 0 0 0 231.035880000 231.035878900
U   Uranium       1.86 1.42 3 4 5 6 0 238.028910000 238.050782600
Np  Neptunium     2.00 1.61 3 4 5 6 0 237.000000000 237.048167300
Pu  Plutonium     2.00 1.61 3 4 5 6 0 244.000000000 244.064198000
```

```
  Am  Americium     2.00 0.92 3 4 5 6 0 243.000000000 243.061372700
  Cm  Curium        2.00 0.91 3 0 0 0 0 247.000000000 247.070347000
  Bk  Berkelium     2.00 0.90 3 4 0 0 0 247.000000000 247.070299000
  Cf  Californium   2.00 0.89 3 0 0 0 0 251.000000000 251.079580000
  Es  Einsteinium   2.00 0.88 3 0 0 0 0 252.000000000 252.082970000
  Fm  Fermium       2.00 0.87 3 0 0 0 0 257.000000000 257.095099000
  Md  Mendelevium   2.00 0.86 2 3 0 0 0 258.000000000 258.098425000
  No  Nobelium      2.00 0.85 2 3 0 0 0 259.000000000 259.101020000
  Lr  Lawrencium    2.00 0.84 3 0 0 0 0 262.000000000 262.109690000
#
# Additional elements and pseudoatoms
  Rf  Rutherfordium 1.58 0.84 1 2 3 4 5 261.000000000   261.0000000
  Db  Dubnium       1.58 0.84 1 2 3 4 5 262.000000000   262.0000000
  Sg  Seaborgium    1.58 0.84 1 2 3 4 5 266.000000000   266.0000000
  Bh  Bohrium       1.58 0.84 1 2 3 4 5 264.000000000   264.0000000
  Hs  Hassium       1.58 0.84 1 2 3 4 5 277.000000000   277.0000000
  Mt  Meitnerium    1.58 0.84 1 2 3 4 5 268.000000000   268.0000000
# Ds  Darmstadtium  1.58 0.84 1 2 3 4 5 280.000000000   280.0000000
# Rg  Roentgenium   1.58 0.84 1 2 3 4 5 280.000000000   280.0000000
# Cn  Copernicium   1.58 0.84 1 2 3 4 5 285.000000000   285.0000000
# Uut Ununtrium     1.58 0.84 1 2 3 4 5 284.000000000   284.0000000
# Fl  Flerovium     1.58 0.84 1 2 3 4 5 289.000000000   289.0000000
# Uup Ununpentium   1.58 0.84 1 2 3 4 5 288.000000000   288.0000000
# Lv  Livermorium   1.58 0.84 1 2 3 4 5 293.000000000   293.0000000
# Uus Ununseptium   1.58 0.84 1 2 3 4 5 294.000000000   294.0000000
# Uuo Ununoctium    1.58 0.84 1 2 3 4 5 294.000000000   294.0000000
  Pol PolymerBead   1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
  Mod Modification  1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Rgp R             1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
  H2  Hydrogen      2.18 0.64 0 0 0 0 0   2.015880000   2.015650064
  H+  Hydrogen      1.09 0.32 0 0 0 0 0   1.007280000   0.000000000
# Ala Alanine       1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Arg Arginine      1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Asn Asparagine    1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Asp Asparagine    1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Cys Cystine       1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Gln Glutamine     1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Glu GlutamicAcid  1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Gly Glycine       1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# His Histidine     1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Ile Isoleucine    1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Leu Leucine       1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Lys Lysine        1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Met Methionine    1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Phe Phenylalanine 1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Pro Proline       1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Ser Serine        1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Thr Threonine     1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Trp Tryptophan    1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Tyr Tyrosine      1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
# Val Valine        1.80 0.79 1 2 3 4 5   0.000000000   0.000000000
```

## Default Ptable

The BIOVIA Ptable contains the following natural elements and values:

| Symbol | Name | VDW Radius | Covalent Radius | Legal Oxidation States | Atomic Weight | Exact Mass |
|---|---|---|---|---|---|---|
| H | Hydrogen | 1.09 | 0.32 | 1 0 0 0 0 | 1.00794 | 1.007825032 |
| He | Helium | 1.4 | 1.60 | 0 0 0 0 0 | 4.002602 | 4.00260325 |
| Li | Lithium | 1.82 | 1.31 | 1 0 0 0 0 | 6.941 | 7.016004 |
| Be | Beryllium | 1.53 | 0.91 | 2 0 0 0 0 | 9.012182 | 9.0121821 |
| B | Boron | 1.92 | 0.82 | 3 0 0 0 0 | 10.811 | 11.0093055 |
| C | Carbon | 1.7 | 0.77 | 4 2 0 0 0 | 12.0107 | 12 |
| N | Nitrogen | 1.55 | 0.75 | 3 0 0 0 0 | 14.0067 | 14.000307401 |
| O | Oxygen | 1.52 | 0.73 | 2 0 0 0 0 | 15.9994 | 15.99491462 |
| F | Fluorine | 1.47 | 0.72 | 1 0 0 0 0 | 18.9984032 | 18.9984032 |
| Ne | Neon | 1.54 | 1.12 | 0 0 0 0 0 | 20.1797 | 19.99244018 |
| Na | Sodium | 2.27 | 1.66 | 1 0 0 0 0 | 22.98977 | 22.98976967 |
| Mg | Magnesium | 1.73 | 1.36 | 2 0 0 0 0 | 24.305 | 23.985419 |
| Al | Aluminum | 1.84 | 1.18 | 3 0 0 0 0 | 26.981538 | 26.98153844 |
| Si | Silicon | 2.1 | 1.10 | 4 0 0 0 0 | 28.0855 | 27.98153844 |
| P | Phosphorus | 1.8 | 1.06 | 3 5 0 0 0 | 30.973761 | 30.9736151 |
| S | Sulfur | 1.8 | 1.02 | 2 4 6 0 0 | 32.065 | 31.97207069 |
| Cl | Chlorine | 1.75 | 0.99 | 1 3 5 7 0 | 35.453 | 34.96885271 |
| Ar | Argon | 1.88 | 1.54 | 0 0 0 0 0 | 39.948 | 39.96238312 |
| K | Potassium | 2.75 | 2.06 | 1 0 0 0 0 | 39.0983 | 38.963769 |
| Ca | Calcium | 2.31 | 1.74 | 2 0 0 0 0 | 40.078 | 39.9625912 |
| Sc | Scandium | 2.11 | 1.44 | 3 0 0 0 0 | 44.95591 | 44.9559102 |
| Ti | Titanium | 2 | 1.32 | 3 4 0 0 0 | 47.867 | 47.9479471 |
| V | Vanadium | 2 | 1.21 | 5 4 3 2 0 | 50.9415 | 50.9405119 |
| Cr | Chromium | 2 | 1.26 | 6 3 2 0 0 | 51.9961 | 51.9405119 |
| Mn | Manganese | 2 | 1.26 | 7 6 4 2 3 | 54.938049 | 54.9380496 |
| Fe | Iron | 2 | 1.26 | 2 3 0 0 0 | 550845 | 55.9349421 |
| Co | Cobalt | 2 | 1.21 | 2 3 0 0 0 | 58.9332 | 58.9332002 |

| Symbol | Name | VDW Radius | Covalent Radius | Legal Oxidation States | Atomic Weight | Exact Mass |
|---|---|---|---|---|---|---|
| Ni | Nickel | 1.63 | 1.15 | 2 3 0 0 0 | 58.6934 | 57.9353479 |
| Cu | Copper | 1.4 | 1.17 | 2 1 0 0 0 | 63.546 | 62.9296011 |
| Zn | Zinc | 1.39 | 1.25 | 2 0 0 0 0 | 65.409 | 63.9291466 |
| Ga | Gallium | 1.87 | 1.26 | 3 2 0 0 0 | 69.723 | 68.925581 |
| Ge | Germanium | 2.11 | 1.27 | 4 0 0 0 0 | 72.64 | 73.9211782 |
| As | Arsenic | 1.85 | 1.20 | 3 5 0 0 0 | 74.9216 | 74.9215964 |
| Se | Selenium | 1.9 | 1.17 | 2 4 6 0 0 | 78.96 | 79.9165218 |
| Br | Bromine | 1.85 | 1.14 | 1 3 5 7 0 | 79.904 | 78.9183376 |
| Kr | Krypton | 2.02 | 1.60 | 0 0 0 0 0 | 83.798 | 83.911507 |
| Rb | Rubidium | 3.03 | 2.21 | 1 0 0 0 0 | 85.4678 | 84.9117893 |
| Sr | Strontium | 2.49 | 1.86 | 2 0 0 0 0 | 87.62 | 87.9056143 |
| Y | Yttrium | 2 | 1.66 | 3 0 0 0 0 | 88.90585 | 88.9058479 |
| Zr | Zirconium | 2 | 1.41 | 4 0 0 0 0 | 91.224 | 89.9047037 |
| Nb | Niobium | 2 | 1.31 | 3 5 0 0 0 | 92.90638 | 92.9063775 |
| Mo | Molybdenum | 2 | 1.30 | 2 3 4 5 6 | 95.94 | 97.9054078 |
| Tc | Technetium | 2 | 1.27 | 7 0 0 0 0 | 98 | 97.907216 |
| Ru | Ruthenium | 2 | 1.16 | 2 3 4 6 8 | 101.07 | 101.9043495 |
| Rh | Rhodium | 2 | 1.25 | 2 3 4 0 0 | 102.9055 | 102.905504 |
| Pd | Palladium | 1.63 | 1.26 | 2 4 0 0 0 | 106.42 | 105.903483 |
| Ag | Silver | 1.72 | 1.34 | 1 0 0 0 0 | 107.8682 | 106.905093 |
| Cd | Cadmium | 1.58 | 1.48 | 2 0 0 0 0 | 112.411 | 113.9033581 |
| In | Indium | 1.93 | 1.44 | 3 0 0 0 0 | 114.818 | 114.903878 |
| Sn | Tin | 2.17 | 1.40 | 2 4 0 0 0 | 118.71 | 119.9021966 |
| Sb | Antimony | 2.06 | 1.41 | 3 5 0 0 0 | 121.76 | 120.903818 |
| Te | Tellurium | 2.06 | 1.37 | 2 4 6 0 0 | 127.6 | 129.9062228 |
| I | Iodine | 1.98 | 1.33 | 1 3 5 7 0 | 126.90447 | 126.904468 |
| Xe | Xenon | 2.16 | 1.31 | 0 0 0 0 0 | 131.293 | 131.9041545 |
| Cs | Cesium | 3.43 | 2.46 | 1 0 0 0 0 | 132.90545 | 132.905447 |
| Ba | Barium | 2.68 | 2.01 | 2 0 0 0 0 | 137.327 | 137.905241 |

| Symbol | Name | VDW Radius | Covalent Radius | Legal Oxidation States | Atomic Weight | Exact Mass |
|---|---|---|---|---|---|---|
| La | Lanthanum | 2 | 1.81 | 3 0 0 0 0 | 138.9055 | 138.906348 |
| Ce | Cerium | 2 | 1.65 | 3 4 0 0 0 | 140.116 | 139.905434 |
| Pr | Praseodymium | 2 | 1.71 | 3 4 0 0 0 | 140.90765 | 140.907648 |
| Nd | Neodymium | 2 | 1.64 | 3 0 0 0 0 | 144.24 | 143.910083 |
| Pm | Promethium | 2 | 1.63 | 3 0 0 0 0 | 145 145 | 144.912744 |
| Sm | Samarium | 2 | 1.62 | 2 3 0 0 0 | 150.36 | 151.919728 |
| Eu | Europium | 2 | 1.85 | 3 3 0 0 0 | 151.964 | 152.921226 |
| Gd | Gadolinium | 2 | 1.61 | 3 0 0 0 0 | 157.25 | 157.924101 |
| Tb | Terbium | 2 | 1.59 | 3 4 0 0 0 | 158.92534 | 158.925343 |
| Dy | Dysprosium | 2 | 1.59 | 3 0 0 0 0 | 162.5 | 163.929171 |
| Ho | Holmium | 2 | 1.58 | 3 0 0 0 0 | 164.93032 | 164.930319 |
| Er | Erbium | 2 | 1.57 | 3 0 0 0 0 | 167.259 | 165.93029 |
| Tm | Thulium | 2 | 1.56 | 2 3 0 0 0 | 168.93421 | 168.934211 |
| Yb | Ytterbium | 2 | 1.56 | 2 3 0 0 0 | 173.04 | 173.9388581 |
| Lu | Lutetium | 2 | 1.56 | 3 0 0 0 0 | 174.967 | 174.9407679 |
| Hf | Hafnium | 2 | 1.41 | 4 0 0 0 0 | 178.49 | 179.9465488 |
| Ta | Tantalum | 2 | 1.31 | 5 0 0 0 0 | 180.9479 | 180.947996 |
| W | Tungsten | 2 | 1.30 | 2 3 4 5 6 | 183.84 | 183.9509326 |
| Re | Rhenium |  | 1.28 | 1 2 4 6 7 | 186.207 | 186.9557508 |
| Os | Osmium | 2 | 1.26 | 2 3 4 6 8 | 190.23 | 191.961479 |
| Ir | Iridium | 2 | 1.27 | 2 3 4 6 8 | 192.217 | 192.962924 |
| Pt | Platinum | 1.72 | 1.30 | 2 4 0 0 0 | 195.078 | 194.964774 |
| Au | Gold | 1.66 | 1.34 | 1 3 0 0 0 | 196.96655 | 196.966552 |
| Hg | Mercury | 1.55 | 1.49 | 1 2 0 0 0 | 200.59 | 201.970626 |
| Tl | Thallium | 1.96 | 1.48 | 1 3 0 0 0 | 204.3833 | 204.974412 |
| Pb | Lead | 2.02 | 1.47 | 2 4 0 0 0 | 207.2 | 207.976636 |
| Bi | Bismuth | 2.07 | 1.46 | 3 5 0 0 0 | 208.98038 | 208.980383 |
| Po | Polonium | 1.97 | 1.46 | 2 4 0 0 0 | 209 | 208.982416 |
| At | Astatine | 2.02 | 1.45 | 1 3 5 7 0 | 210 | 209.987131 |

| Symbol | Name | VDW Radius | Covalent Radius | Legal Oxidation States | Atomic Weight | Exact Mass |
|---|---|---|---|---|---|---|
| Rn | Radon | 2.2 | 1.90 | 0 0 0 0 0 | 222 | 222.0175705 |
| Fr | Francium | 3.48 | 1.80 | 1 0 0 0 0 | 223 | 223.0197307 |
| Ra | Radium | 2.83 | 2.00 | 2 0 0 0 0 | 226 | 226.0254026 |
| Ac | Actinium | 2 | 1.81 | 3 0 0 0 0 | 227 | 227.027747 |
| Th | Thorium | 2 | 1.65 | 4 0 0 0 0 | 232.0381 | 232.0380504 |
| Pa | Protactinium | 2 | 1.66 | 4 5 0 0 0 | 231.03588 | 231.0358789 |
| U | Uranium | 1.86 | 1.42 | 3 4 5 6 0 | 238.02891 | 238.0507826 |
| Np | Neptunium | 2 | 1.61 | 3 4 5 6 0 | 237 | 237.0481673 |
| Pu | Plutonium | 2 | 1.61 | 3 4 5 6 0 | 244 | 244.064198 |
| Am | Americium | 2 | 0.92 | 3 4 5 6 0 | 243 | 243.0613727 |
| Cm | Curium | 2 | 0.91 | 3 0 0 0 0 | 247 | 247.070347 |
| Bk | Berkelium | 2 | 0.90 | 3 4 0 0 0 | 247 | 247.070299 |
| Cf | Californium | 2 | 0.89 | 3 0 0 0 0 | 251 | 251.07958 |
| Es | Einsteinium | 2 | 0.88 | 3 0 0 0 0 | 252 | 252.08297 |
| Fm | Fermium | 2 | 0.87 | 3 0 0 0 0 | 257 | 257.095099 |
| Md | Mendelevium | 2 | 0.86 | 2 3 0 0 0 | 258 | 258.098425 |
| No | Nobelium | 2 | 0.85 | 2 3 0 0 0 | 259 | 259.10102 |
| Lr | Lawrencium | 2 | 0.84 | 3 0 0 0 0 | 262 | 262.10969 |

# Chapter 15: Customizing the BIOVIA Salts Definition

## Introduction

The BIOVIA Salts definition specifies fragments within a structure that flexmatch recognizes as salt counterions or hydrates. When on, the flexmatch switch SAL strips any recognized counterions from the query before conducting the search. For more information on flexmatch search, see Exact Search (Flexmatch) on page 107.

The default Salts definition recognizes the following as salt counterions:

- Alkali metals: Li, Na, K
- Halogens: F, Cl, Br
- Carbonates, nitrates, nitrites, perchlorates and sulfates
- Water (hydrates)

You can add your own definitions of counterions to your company's Salts definition.

You need to customize the Salts definition solely if your business rules use a single chemical structure to represent salts. Other data models might represent salts differently. For example, you might want to store only the parent compound as a chemical structure and store information on counterions and hydrates in a separate database field.

## Changing the Current Salts Definition

Perform the steps that follow to replace your current Salts definition with a revised Salts definition:

1. Make a copy of your current Salts definition for future reference.
   - Using Oracle SQL*Plus to write the current Salts definition to a text file. For more information, see the chapter, "Setting the Chemical Environment" in the *BIOVIA Direct Administration Guide*. For example:

   ```
   SQL> connect c$direct<version>/password
   SQL> select mdlaux.getenvfile('SALTS', 'c:\current_salts.txt') from
   dual;
   ```

2. Follow the guidelines in Format of the Salts Definition on page 220, to edit the Salts definition.
3. Load the revised Salts definition.
   - Use Oracle SQL*Plus to read the revised Salts definition file. For more information, see the chapter, "Setting the Chemical Environment" in the *BIOVIA Direct Administration Guide*. For example:

   ```
   SQL> connect c$direct<version>/password
   SQL> select mdlaux.setenvfile('SALTS', 'c:\new_salts.txt') from dual;
   ```

## Default Salts Definition

The salts definition is an SDFile containing counterion definitions and an optional counterion removal code. Each counterion is a separate record in the SDFile. The counterion removal code is an optional data field entry following the molfile counterion definition. For a sdfile definition, see the CT Files document.

Use BIOVIA Draw to define each counterion and save as a molfile. Then use an editor to add the counterion molfile to the SDFile. Terminate each record in the SDFile with four dollar sign characters on a line by themselves, for example:

```
Dichloride
  ACCLDraw04091209072D

  2  0  0  0  0  0  0  0  0  0999 V2000
    5.0000   -4.0625    0.0000 Cl  0  5  0  0  0  0  0  0  0  0  0  0
    4.9688   -5.2500    0.0000 Cl  0  5  0  0  0  0  0  0  0  0  0  0
M  CHG  2   1  -1   2  -1
M  END
$$$$
```

If no counterion removal code is specified, the counterion is removed from the query only if it is a completely separate fragment. The file that BIOVIA ships with Direct uses only this default.

You can also specify that the counterion should be removed if it is a completely separate fragment, or if it is attached to a fragment apart from the main structure. In this case the entire fragment including the counterion is removed. To specify this option, include a data field named `FragmentFlag` with a value of **F** in the SDFile, for example:

```
Nitrate-1
  ACCLDraw04091209252D

  4  3  0  0  0  0  0  0  0  0999 V2000
    5.1250   -5.0313    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    5.1250   -3.9439    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    6.1479   -5.6218    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
    4.1021   -5.6218    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  1  3  1  0  0  0  0
  1  4  2  0  0  0  0
M  CHG  1   3  -1
M  END
>  <FragmentFlag>
F

$$$$
```

where:

`<FragmentFlag>` specifies the field name. **F** is the value for the field. The blank line following the value is required.

## The BIOVIA Direct default salts definition file

```
Dichloride
  ACCLDraw04091209072D

  2  0  0  0  0  0  0  0  0  0999 V2000
    5.0000   -4.0625    0.0000 Cl  0  5  0  0  0  0  0  0  0  0  0  0
    4.9688   -5.2500    0.0000 Cl  0  5  0  0  0  0  0  0  0  0  0  0
M  CHG  2   1  -1   2  -1
M  END
$$$$
Chloride
```

```
  ACCLDraw04091209072D


  1  0  0  0  0  0  0  0  0  0999 V2000
    5.0000   -4.0625    0.0000 Cl  0  5  0  0  0  0  0  0  0  0  0  0
M  CHG  1   1  -1
M  END
$$$$
Bromide
  ACCLDraw04091209192D


  1  0  0  0  0  0  0  0  0  0999 V2000
    3.0313   -2.6875    0.0000 Br  0  5  0  0  0  0  0  0  0  0  0  0
M  CHG  1   1  -1
M  END
$$$$
Fluoride
  ACCLDraw04091209202D

  1  0  0  0  0  0  0  0  0  0999 V2000
    3.3438   -2.5000    0.0000 F   0  5  0  0  0  0  0  0  0  0  0  0
M  CHG  1   1  -1
M  END
$$$$
Lithium
  ACCLDraw04091209202D

  1  0  0  0  0  0  0  0  0  0999 V2000
    3.0938   -3.5000    0.0000 Li  0  3  0  0  0  0  0  0  0  0  0  0
M  CHG  1   1   1
M  END
$$$$
Sodium
  ACCLDraw04091209202D

  1  0  0  0  0  0  0  0  0  0999 V2000
    2.5313   -2.9688    0.0000 Na  0  3  0  0  0  0  0  0  0  0  0  0
M  CHG  1   1   1
M  END
$$$$
Potassium
  ACCLDraw04091209212D

  1  0  0  0  0  0  0  0  0  0999 V2000
    2.4375   -2.7500    0.0000 K   0  3  0  0  0  0  0  0  0  0  0  0
M  CHG  1   1   1
M  END
$$$$
Dicarbonate
  ACCLDraw04091209222D

  8  6  0  0  0  0  0  0  0  0999 V2000
```

```
    5.1250   -5.0313    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.1250   -3.8501    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    6.1479   -5.6218    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
    4.1021   -5.6218    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
    4.0709   -9.4656    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
    6.1166   -9.4656    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
    5.0938   -7.6939    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    5.0938   -8.8750    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  1  3  1  0  0  0  0
  1  4  1  0  0  0  0
  8  5  1  0  0  0  0
  8  6  1  0  0  0  0
  8  7  2  0  0  0  0
M  CHG  4   3  -1   4  -1   5  -1   6  -1
M  END
$$$$
Carbonate
  ACCLDraw04091209222D

  4  3  0  0  0  0  0  0  0  0999 V2000
    5.1250   -5.0313    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.1250   -3.8501    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    6.1479   -5.6218    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
    4.1021   -5.6218    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  1  3  1  0  0  0  0
  1  4  1  0  0  0  0
M  CHG  2   3  -1   4  -1
M  END
$$$$
Sulfate
  ACCLDraw04091209232D

  5  4  0  0  0  0  0  0  0  0999 V2000
    5.1250   -5.0313    0.0000 S   0  0  3  0  0  0  0  0  0  0  0  0
    4.5625   -3.9751    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    6.1479   -5.6218    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
    4.1021   -5.6218    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
    5.8054   -3.9779    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  1  3  1  0  0  0  0
  1  4  1  0  0  0  0
  1  5  2  0  0  0  0
M  CHG  2   3  -1   4  -1
M  END
$$$$
Nitrate-1
  ACCLDraw04091209252D

  4  3  0  0  0  0  0  0  0  0999 V2000
    5.1250   -5.0313    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    5.1250   -3.9439    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    6.1479   -5.6218    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
    4.1021   -5.6218    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  1  3  1  0  0  0  0
  1  4  2  0  0  0  0
M  CHG  1   3  -1
M  END
$$$$
Nitrate-2
  ACCLDraw04091209252D

  4  3  0  0  0  0  0  0  0  0999 V2000
    5.1250   -5.0313    0.0000 N   0  3  0  0  0  0  0  0  0  0  0  0
    5.1250   -3.9439    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    6.1479   -5.6218    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
    4.1021   -5.6218    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  1  3  1  0  0  0  0
  1  4  1  0  0  0  0
M  CHG  3   1   1   3  -1   4  -1
M  END
$$$$
Nitrite
  ACCLDraw04091209262D

  3  2  0  0  0  0  0  0  0  0999 V2000
    5.1250   -5.0313    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    5.1250   -3.9439    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    5.1021   -6.0593    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  1  3  1  0  0  0  0
M  CHG  1   3  -1
M  END
$$$$
Perchlorate
  ACCLDraw04091209322D

  5  4  0  0  0  0  0  0  0  0999 V2000
    3.7188   -3.6875    0.0000 Cl  0  0  0  0  0  0  0  0  0  0  0  0
    5.0229   -3.6907    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    3.7188   -4.8686    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    3.7271   -2.6907    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    2.6021   -3.7156    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  1  3  2  0  0  0  0
  1  4  2  0  0  0  0
  1  5  1  0  0  0  0
M  CHG  1   5  -1
M  END
$$$$
Hydrochloride
  ACCLDraw04091209342D
  2  1  0  0  0  0  0  0  0  0999 V2000
    3.2813   -2.3125    0.0000 Cl   0  0  0  0  0  0  0  0  0  0  0  0
    4.4624   -2.3125    0.0000 H    0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
M  END
$$$$
Hydrate
  ACCLDraw04091209342D

  3  2  0  0  0  0  0  0  0  0999 V2000
    3.2813   -2.3125    0.0000 O    0  0  0  0  0  0  0  0  0  0  0  0
    4.4624   -2.3125    0.0000 H    0  0  0  0  0  0  0  0  0  0  0  0
    2.6907   -3.3354    0.0000 H    0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  1  0  0  0  0
M  END
$$$$
```

## Format of the Salts Definition

The Salts definition contains two types of information: comments and counterion definitions.

### Comments

Comment lines must begin with an asterisk.

### Counterion definitions

Each counterion definition includes:

- A counterion removal code (which can be a blank)
- A counterion name
- A counterion definition

The counterion removal code occupies character position 1. It specifies how a counterion should be stripped from a parent-search query. The table that follows explains the functions of the codes:

| Code | Function |
|---|---|
| A Blank | Specifies counterion removal only if it is a completely separate fragment. The file that BIOVIA ships uses this code only. |
| The letter E | Specifies counterion removal if it is a completely separate fragment, or if it is attached to the main molecular structure. |
| The letter F | Specifies counterion removal if it is a completely separate fragment, or if it is attached to a fragment apart from the main structure. In the latter case, the entire fragment including the counterion is removed. |

The counterion name begins in character position 5. It can be a name of any length and must be followed by at least one space. The counterion name is a group you want removed from a query structure.

The counterion definition can begin in any character position following the counterion name and at least one space.

## Rules for modifying the Salts definition

Observe these rules when you modify the Salts definition file:

- BIOVIA recommends that the line length limit in the Salts definition file be 255 characters.
- To remove a specific number of instances of a given ion from a query, specify multiple copies of that ion in the counterion definition, as shown in the dichloride and dicarbonate entries.
- Order larger counterions and salts before smaller counterions and salts.
- The order of entries using code E or F is very important. Always list the more exclusive counterions and salt fragments at the beginning of the file, for example, nitrates before nitrites and dianions before anions.

The entries that follow are an example of entries in the incorrect order. In this example, you list nitrite before nitrate using the code F:

F Nitrite O=N*(-O")

F Nitrate O=N*(-O")-O"

Using a Salts definition with these incorrect entries results in the following behavior when BIOVIA Direct encounters a nitrate group in a parent-search query:

a. It views the group as a nitrite fragment attached to an oxygen anion.
b. It removes the nitrogen, the double-bonded oxygen, and one of the oxygen anions according to the nitrite definition.
c. It continues the parent search for an exact match of the stripped query. The results of the search incorrectly list the query structure as having included a nitrite group instead of a nitrate group.
d. Because of the F code, the second oxygen anion attached to the nitrite group is also removed.

# Appendix A: Representation of Stereochemistry in BIOVIA Databases

## Perception of Stereoconfiguration at Tetrahedral Centers

Stereochemical perception is the process by which BIOVIA software determines stereoconfiguration from a two-dimensional structure drawing. In effect, stereochemical perception is the reverse of projecting a three-dimensional structure onto a plane. The two-dimensional representation of the structure necessarily involves a loss of information because the drawing lacks information on z-coordinates. By convention, chemists use Up and Down stereo bonds to supply the missing information.

To reconstruct the intended three-dimensional structure, the software must make certain assumptions about how to interpret the information from the x-y atom coordinates and the Up/Down stereo bonds. For applications of these assumptions, see Rules for Drawing Structures with Tetrahedral Stereochemistry on page 227.

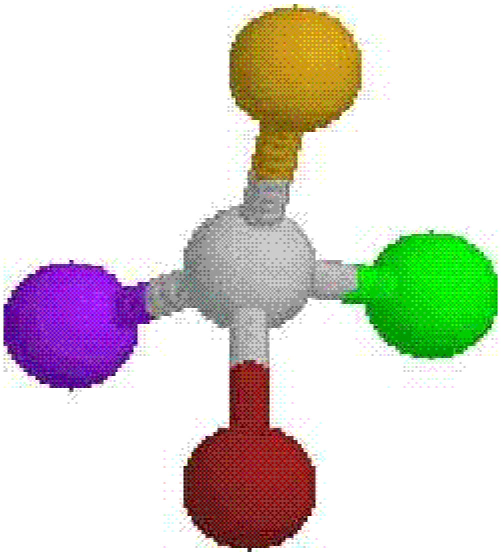
### Axioms

Axioms are assumptions that are self-evident from the geometry that you obtain when you project a three-dimensional structure with tetrahedral geometry onto a plane. In the following axioms, the drawing is assumed to be projected onto the x-y plane.

### Axiom 1

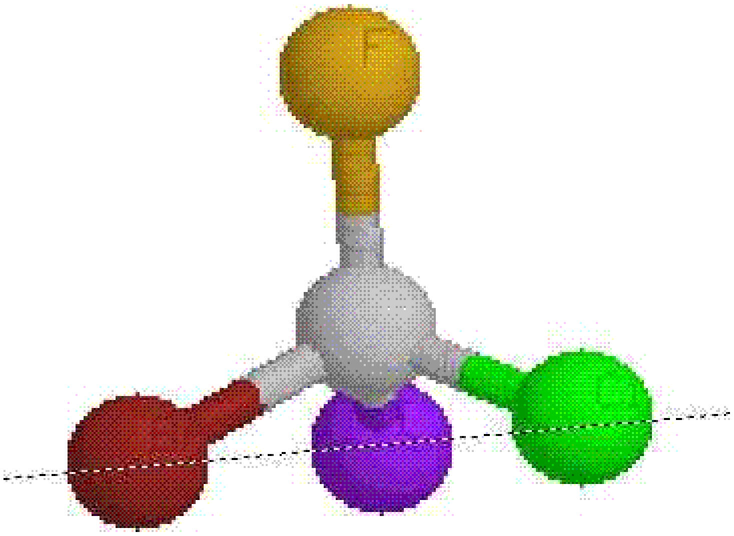
If the structure is oriented so that the central atom is in the x-y plane and its attachments are out of the plane, then:

- Atoms attached to bonds that lie opposite each other must lie on the same side of the x-y plane (either both above or both below).
- Atoms attached to adjacent bonds must lie on opposite sides of the x-y plane (one above and one below). For example:



### Axiom 2 (Colinearity Axiom)

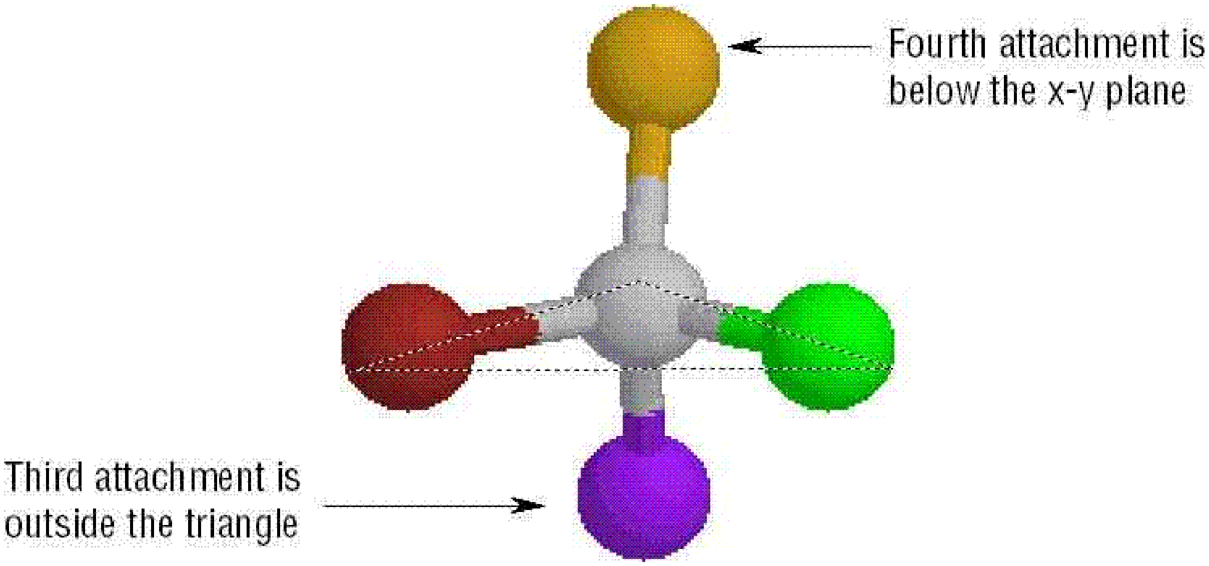
When the central atom and one of its attached atoms both lie in the x-y plane, the atoms in the remaining attachments must be colinear:



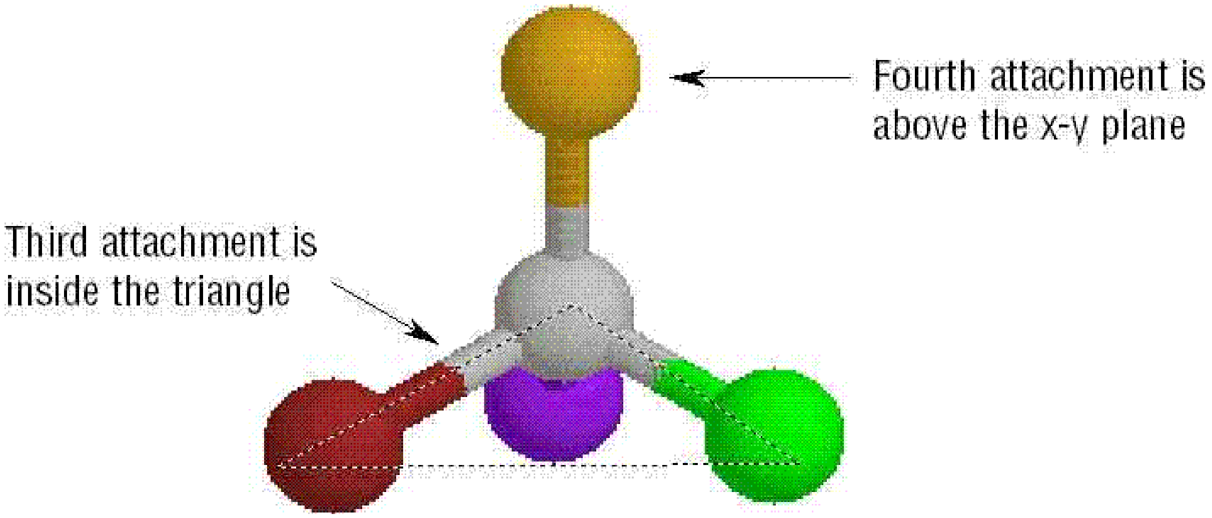
### Axiom 3 (Triangle Axiom)

When the structure is oriented so that the central atom and two attachments that lie opposite one another are above the x-y plane, and the third attachment lies below the x-y plane, then the following are true:

- If the third attachment lies outside the triangle that is formed by the central atom and the first two attachments, then the fourth attachment must also lie below the x-y plane:

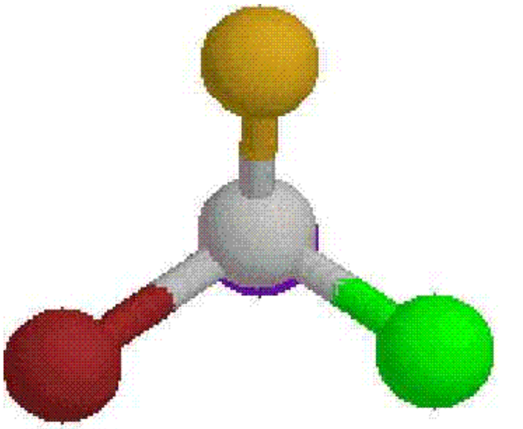

Fourth attachment is below the x-y plane
Third attachment is outside the triangle

- If the third attachment lies inside the triangle that is formed by the central atom and the first two attachments, then the fourth attachment must also lie above the x-y plane:

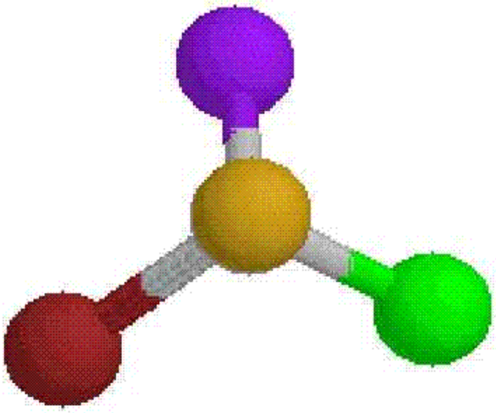

Fourth attachment is above the x-y plane
Third attachment is inside the triangle

### Axiom 4

When the structure is oriented so that one of the attached atoms is eclipsed (completely hidden) by the central atom, then the other three attachments must lie above the x-y plane. Conversely, if the central atom is eclipsed by one of its attachments, then the other three attachments must lie below the x-y plane:



Central atom eclipses one atom
Other three atoms lie above x-y plane



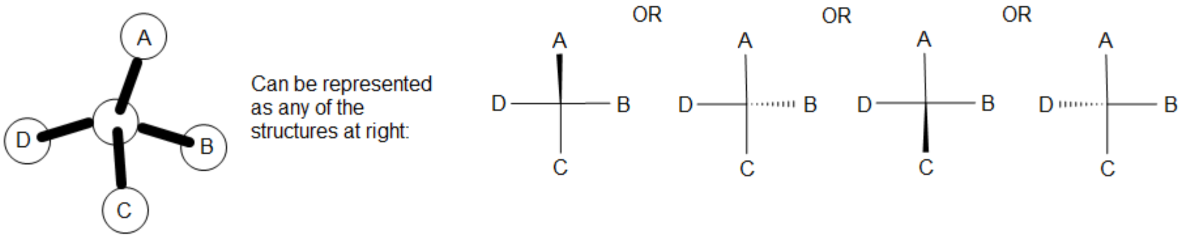
One atom eclipses central atom
Other three atoms lie below x-y plane

## Postulates

Postulates are assumptions that the stereoperception algorithm must make to calculate the stereoconfiguration (the implied three-dimensional arrangement of the atoms) from the information in the drawing. These postulates are necessary to resolve contradictions that might arise between the information implied by the x-y coordinates of the atoms and the information implied by Up/Down stereo bonds.

## Postulate A

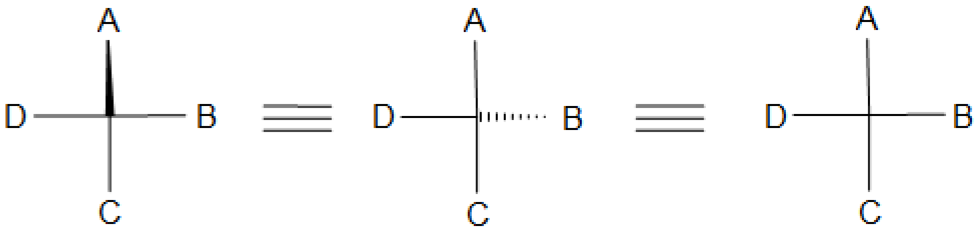
At least one (and only one) Up or Down stereo bond is necessary to specify the stereoconfiguration of a tetrahedral stereogenic center. A Down stereo bond indicates that the atom at the wide end of the stereo bond lies below the x-y plane, while an Up bond indicates that the atom at the wide end lies above the x-y plane. If none of the attached bonds is a stereo bond, the algorithm cannot calculate a stereoconfiguration, so the stereoconfiguration is undefined:



## Postulate B

Stereoconfiguration is perceived solely at the narrow end of the Up or Down stereo bond. Thus, the narrow end of the stereo bond must attach to the stereogenic center whose configuration it specifies.

For example, the central atom in all three structures below is perceived as an undefined stereogenic center; hence the structures are equivalent:

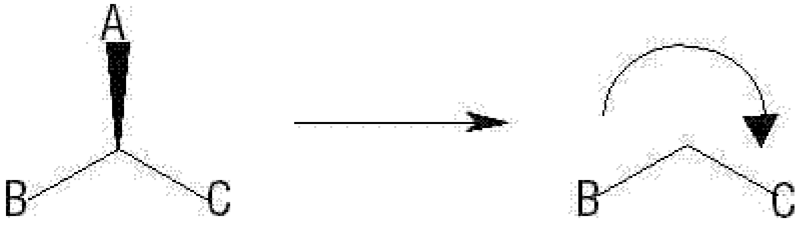


## Postulate C

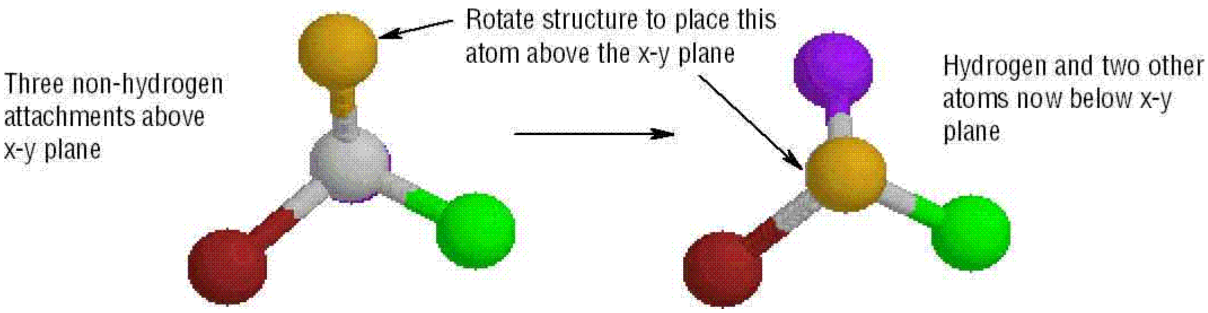
### BIOVIA Direct 8

For a stereogenic center with an implicit hydrogen attachment, the stereo-perception algorithm must calculate the position of the implicit hydrogen and determine the stereoconfiguration from the resulting structure. The software calculates the position of the implicit hydrogen from the two-dimensional drawing using a procedure that can be visualized as follows:

- Move the atom that is attached to the Up stereo bond so that the atom is above the plane of the drawing, thereby eclipsing the stereogenic atom, see Axiom 4 on page 224:

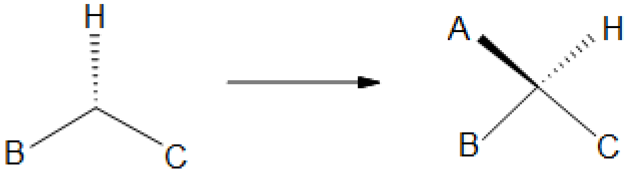


In three dimensions, the process can be visualized as:

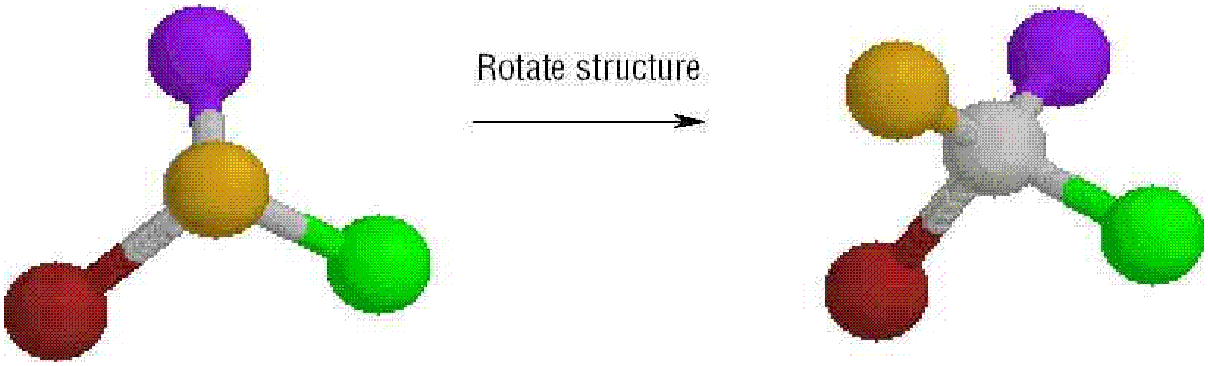

Three non-hydrogen attachments above x-y plane
Rotate structure to place this atom above the x-y plane
Hydrogen and two other atoms now below x-y plane

When you rotate the structure so that atom A eclipses the central atom, then by Axiom 4 the bond to the hydrogen atom and the two attachments with non-stereo bonds are now below the plane of the paper.

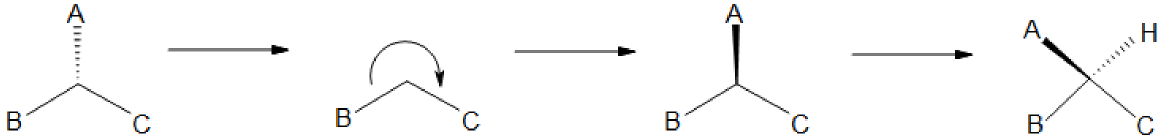
- Place the bond to the hydrogen atom within the larger of the two angles that are formed by the two non-stereo bonds, and mark it as a Down stereo bond. Move the atom attached to the Up stereo bond so that it is adjacent to the explicit hydrogen atom, see Axiom 1 on page 222:



In three dimensions, the process can be visualized as:



For a drawing with three explicit attachments and a Down wedge bond to a non-hydrogen attachment, the procedure is:



## BIOVIA Direct 9

For a stereogenic center with an implicit hydrogen attachment, the stereo-perception algorithm must calculate the position of the implicit hydrogen and determine the stereoconfiguration from the resulting structure. The software calculates the position of the implicit hydrogen from the two-dimensional drawing using a procedure that can be visualized as follows:

1. Initially place all atoms in the x-y plane with z-coordinates of 0.
2. Apply a z-adjustment for the atom on the other side of the wedge. Set z = -1 for an atom on the other side of a Down wedge bond and z=+1 for an atom on the other side of an Up wedge bond.
3. Initially the implicit hydrogen is assumed to be directly opposite the other three attachments: p[H] = -(p[1] + p[2] + p[3]) and attempt to determine the stereo parity for the center.
4. If the parity calculation in #3 fails, make a second attempt to place the implicit hydrogen. This time place the implicit hydrogen either directly above or below the central atom so that it is opposite the wedge that is present

### Postulate D

The stereo-perception algorithm always calculates a stereoconfiguration for any stereogenic center with at least one Up or Down bond, even if information on stereoconfiguration from the x-y atom coordinates conflicts with the information from the stereo bonds. In this case, the calculated stereoconfiguration might be different from that intended.

The algorithm does not return a warning when the stereoconfiguration is calculated from conflicting information. Therefore, it is important to follow stereochemical drawing rules to avoid errors and ensure that the structures that you draw represent the stereoconfiguration that you intend. For more information and examples of the errors that can occur, see Rules for Drawing Structures with Tetrahedral Stereochemistry on page 227.

### Postulate E

You can specify the stereoconfiguration at asymmetric tetrahedral centers (chirality centers) solely if the atom is one of the following: C, N, Si, P, As, S, Se or Te or any of their isoelectric equivalents (for example, B-).

## Stereochemical Groups

Each defined tetrahedral stereogenic center on a structure (that is, each center that is marked with an Up or Down stereo bond) must belong to a stereochemical group (stereogroup). There are three categories of stereogroups: ABS, OR, and AND. If no groups are explicitly specified for the structure, then all centers are placed in a single stereo group. If the chiral flag is set, then all centers are placed in the ABS stereogroup, otherwise all centers are placed in the same AND stereogroup. For more information on the different types of stereogroup and their meanings, see Tetrahedral Stereochemistry on page 9.

# Rules for Drawing Structures with Tetrahedral Stereochemistry

## One Stereo Bond per Stereogenic Center

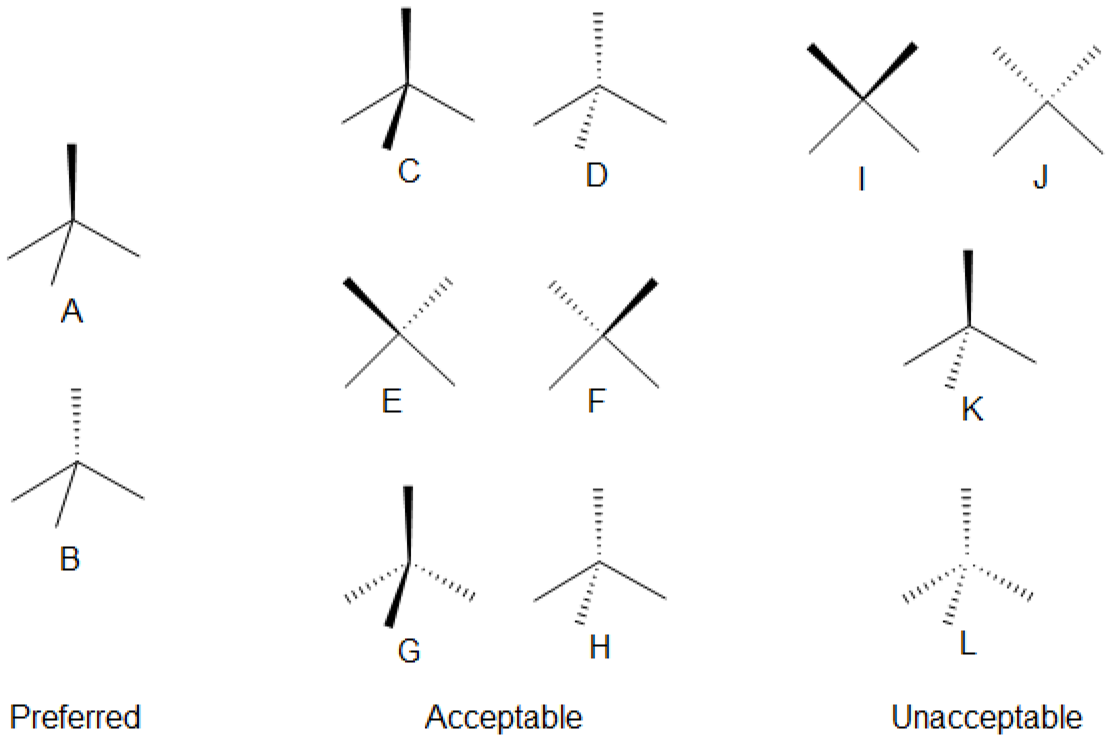
**Note:** The more stereo bonds that you draw to a particular atom, the greater the possibility of introducing errors.

One stereo bond per stereogenic center can unambiguously represent the stereoconfiguration (see Axiom 1 on page 222), so you should avoid drawing more than one stereo bond to a stereogenic center. If it is necessary to draw more than one stereo bond to a stereogenic center, follow these guidelines:

1. Never draw more than two stereo bonds to a stereogenic center with four explicit attachments.
2. Never draw more than one stereo bond to a stereogenic center that has three explicit attachments and an implicit hydrogen.

3. If you cannot avoid drawing two stereo bonds to a stereogenic center with three explicit attachments, then draw the hydrogen explicitly and place one of the stereo bonds to the explicit hydrogen.
4. Never draw more than one stereo bond of the same type (Up or Down) adjacent to each other.
5. Never draw stereo bonds of opposite type across from each other.

The following figure shows preferred, acceptable, and unacceptable drawings (terminal atoms have been omitted for clarity, assume that the four substituents are different):

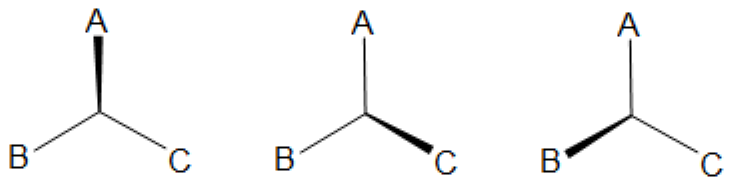


These rules result from Axiom 1, Postulate A, and Postulate D, as follows:

- Structures A and B are preferred.
- Structures A through H are acceptable because they are consistent with the geometry defined in Axiom 1.
- Structures I through L illustrate the errors that can occur when a stereogenic center has more than one stereo bond. These structures are unacceptable because they contradict the geometry specified by Axiom 1. The algorithm calculates a stereoconfiguration for these structures (Postulate D), but the configuration might be different from what you intend.

## Equivalence Rule

For a tetrahedral stereogenic center with three explicit attachments, it does not matter which of the three bonds is marked with a stereo bond. That is, the following structures are equivalent:



## Wedge Bond Rule

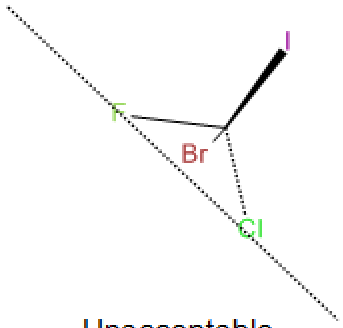
In addition to the Up Wedge and Down Wedge bond tools, BIOVIA Draw provides tools for bold bonds and dashed bonds. Do not use bold bonds or dashed bonds to draw structures that you use as search queries or register to databases, because these bonds do not distinguish the narrow from the wide end of the bond.

## Triangle Rule

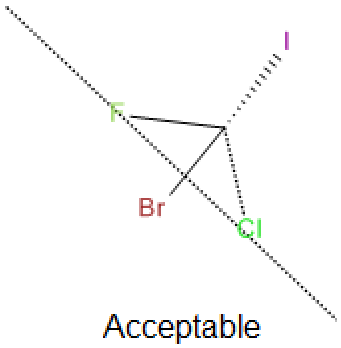
For a tetrahedral stereogenic center with four explicit attachments, the atom at the end of a non-stereo bond should not fall within the triangle defined by the stereogenic center and the two remaining attachments. This rule is necessary because of the way that the stereoconfiguration is calculated:

- If the atom at the end of the non-stereo bond that is opposite the stereo bond falls outside the triangle defined by the stereogenic center and the other non-stereo bonds, then the calculated stereoconfiguration is the same as that implied by the stereo bond.
- If the atom at the end of the non-stereo bond that is opposite the stereo bond falls inside the triangle defined by the stereogenic center and the other non-stereo bonds, then the calculated stereoconfiguration is the opposite of that implied by the Up/Down stereo bond.

For example:



Unacceptable

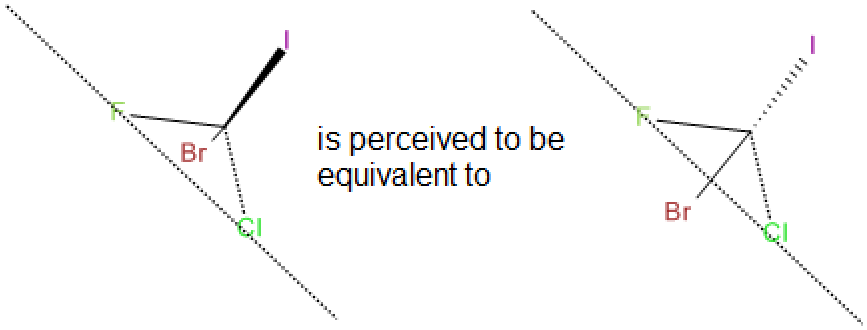


Acceptable

By Postulate A the Br and I atoms must both lie below the x-y plane. Axiom 3 (Triangle Axiom) (triangle axiom) requires that the Br and I atoms in the structure on the left must both lie below the x-y plane, while the structure on the right must have the Br atom below the x-y plane and the I atom above it.

The structure on the right is acceptable because the presence of the Down stereo bond is consistent with the information from Postulate A and Axiom 3 (triangle axiom). The structure on the left is not acceptable, because the presence of the Down stereo bond conflicts with the requirement of Axiom 3 (triangle axiom) that the I atom must lie above the x-y plane.

To resolve the conflict between the information in the x-y coordinates and the information in the stereo bond, the software interprets the drawing as having the opposite stereoconfiguration to that implied by the stereo bond:

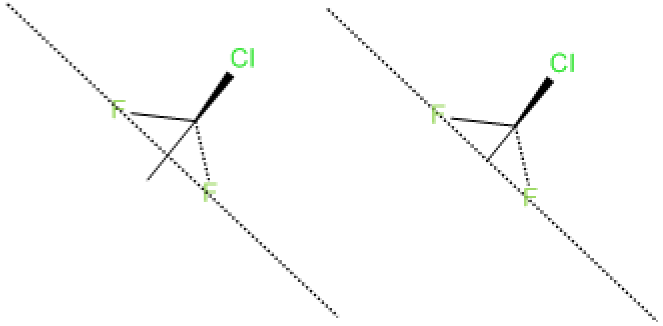


Following the triangle rule avoids this error.

## Colinearity Rule

Certain actions (for example, when the structure is saved to a molfile, used as a search query, or registered to a database) cause the stereoconfiguration of a structure to be re-perceived, that is, recalculated. The colinearity rule ensures that the calculated stereoconfiguration is the same each time the structure is re-perceived. The colinearity rule applies to stereogenic centers with three or four explicit attachments:

- For a tetrahedral stereogenic center with four explicit attachments and one stereo bond, the terminal atoms of the non-stereo bonds cannot be colinear.



Acceptable
Unacceptable

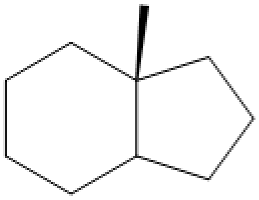
The structure on the left is unacceptable because, by Axiom 2 (Colinearity Axiom) (colinearity axiom), the bond to atom A must lie entirely in the x-y plane, so the presence of an Up stereo bond is a contradiction.

For structures that violate the colinearity rule, small differences in calculated results can invert the stereoconfiguration:

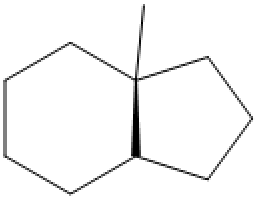
These discrepancies can result from the rounding-off errors inherent in arithmetic operations that involve floating-point numbers. Following the colinearity rule avoids these discrepancies.

## Rule for Adjacent Stereogenic Centers

Avoid drawing an Up or Down stereo bond between two adjacent stereogenic centers. Use an explicit hydrogen to indicate the stereoconfiguration. For example:



Preferred

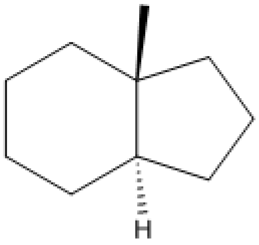


Less desirable

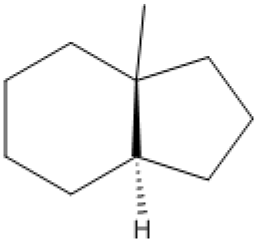
By Postulate B, stereoconfiguration is perceived solely at the narrow end of a stereo bond. Applying this rule helps prevent errors when marking adjacent stereogenic centers.

## Rules for Stereogenic Centers in Rings

Avoid drawing a stereo bond in a ring. For example:

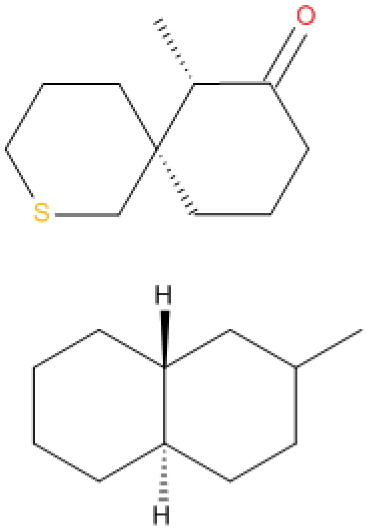


Preferred

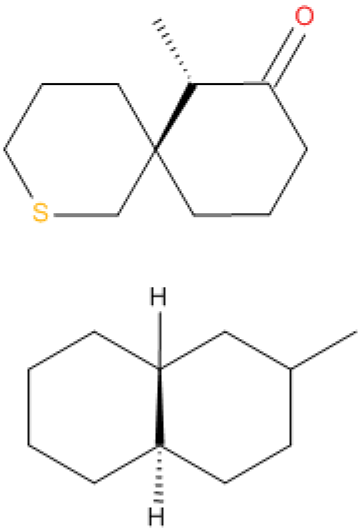


Less desirable

To avoid stereo bonds in rings, you might need to draw explicit hydrogen atoms to carry the stereo bond. For example:



Preferred



Less desireable

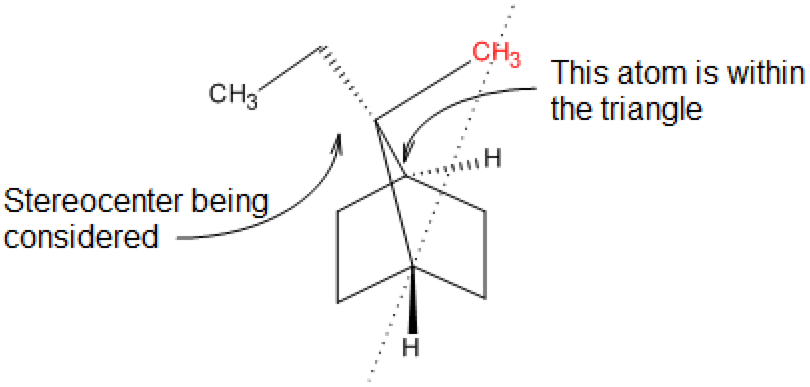
Like the rule for adjacent stereogenic centers, this rule helps prevent errors in marking stereogenic centers.

## Rule for Explicit Hydrogen

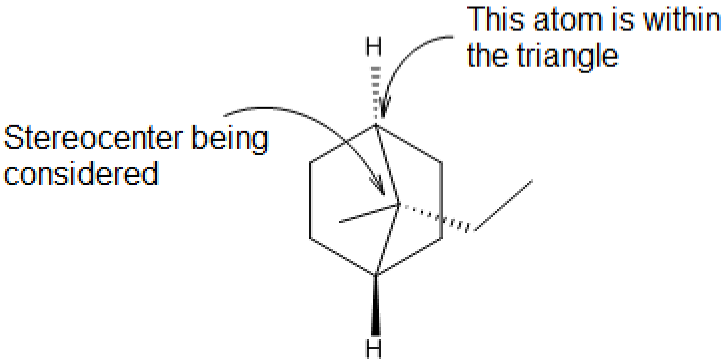
Avoid drawing explicit hydrogen except when necessary to specify stereochemistry. If the stereogenic center has an explicit hydrogen atom attached, always place the stereo bond on that hydrogen atom.

## Non-perspective Drawings are Preferred

Non-perspective drawings, such as Mills depictions, are preferred to perspective drawings, because mistakes are easier to avoid in non-perspective drawings. For example, the triangle rule is often broken in perspective drawings, though it is not easy to detect when you violate this rule:


CH3
CH3
This atom is within the triangle
H
Stereocenter being considered
H

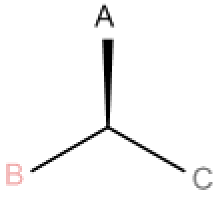
Violation of the triangle rule is easier to detect in the equivalent non-perspective drawing, for example:

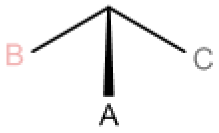

This atom is within the triangle
H
Stereocenter being considered
H

## Rules for Stereogenic Centers with Implicit Hydrogen and Three Explicit Attachments

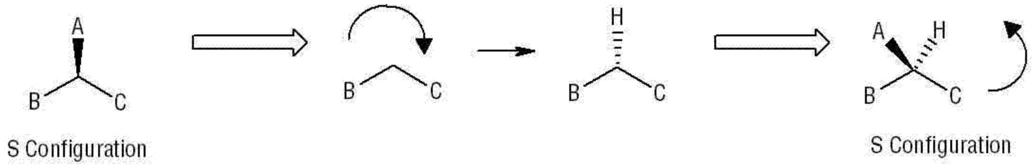
Follow these rules when you want to change the arrangement of bonds at a stereogenic center (that is, the x-y coordinates of the attached atoms) without changing the stereoconfiguration. For example, you might need to change the layout of a complex structure.

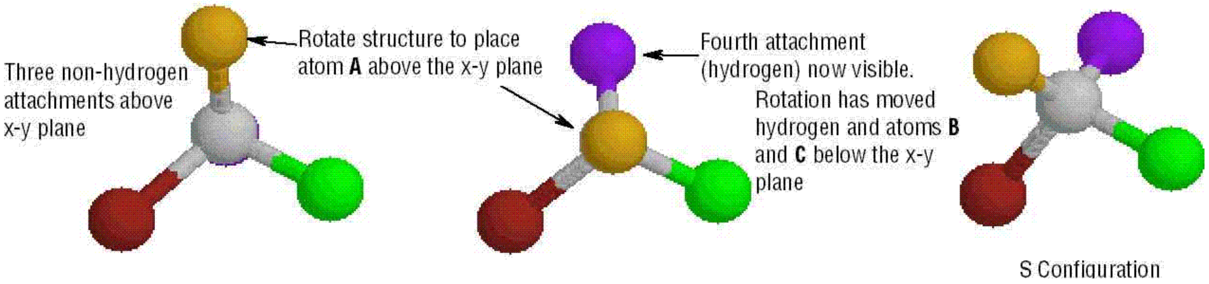
- The position of the Up or Down stereo bond relative to the two non-stereo bonds can be freely changed without affecting the stereoconfiguration.



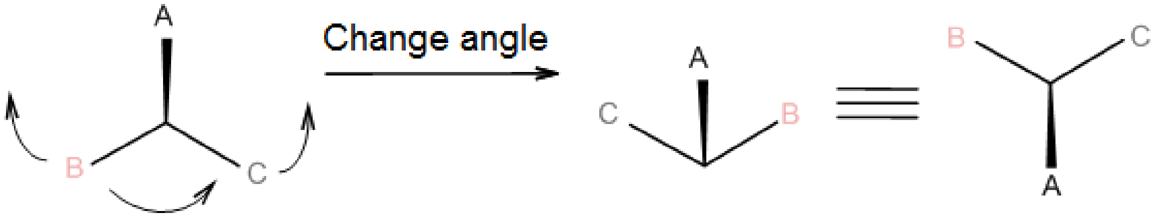


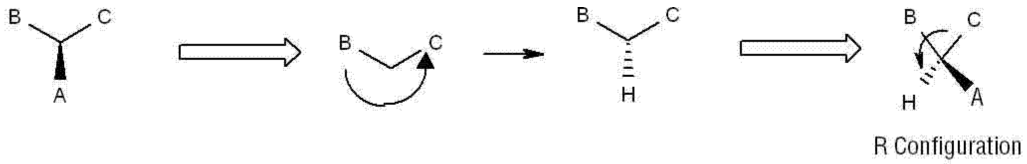
- Do not change the angle between the two non-stereo bonds from less than 180 degrees to more than 180 degrees.

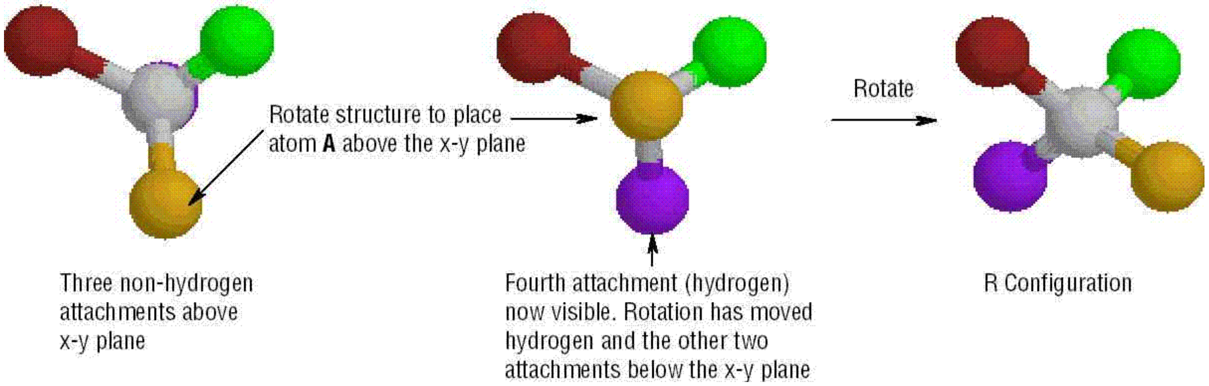



Three non-hydrogen attachments above x-y plane
Rotate structure to place atom **A** above the x-y plane
Fourth attachment (hydrogen) now visible.
Rotation has moved hydrogen and atoms **B** and **C** below the x-y plane
S Configuration

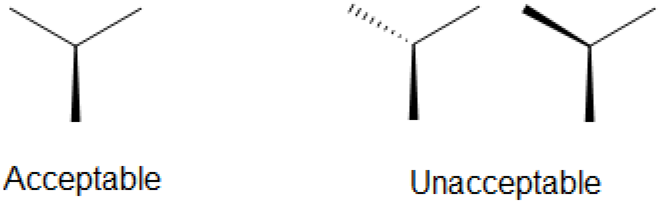
When you change the angle between the non-stereo bonds, you obtain:






Rotate structure to place atom **A** above the x-y plane
Rotate
Three non-hydrogen attachments above x-y plane
Fourth attachment (hydrogen) now visible. Rotation has moved hydrogen and the other two attachments below the x-y plane
R Configuration

- Never draw more than one stereo bond to a stereogenic center with three explicit attachments:

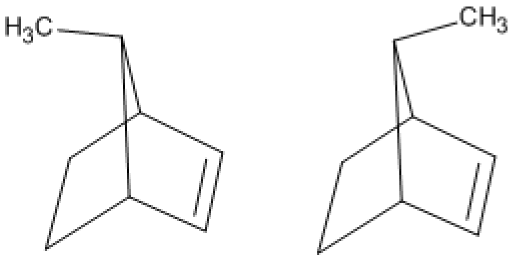


For structures that violate this rule, the algorithm might calculate a different stereoconfiguration each time the structure is re-perceived (for example, when the structure is written to a molfile, used as a search query, or registered to a database).

### Avoid Fischer Projections, Haworth Projections and Perspective Drawings

By Postulate A and Postulate B, stereoconfiguration cannot be determined from x-y atom coordinates alone. Many conventions for drawing stereochemistry, such as Haworth and Fischer projections, lack stereo bonds and are therefore perceived as having undefined stereochemistry.

Like Haworth and Fischer projections, perspective drawings that lack stereo bonds are also perceived as having undefined stereochemistry. For example, a chemist might want to represent two stereoisomers of a bicyclic compound like this:



However, the lack of stereo bonds means that all the stereogenic centers have undefined stereochemistry, hence the two structures are perceived as equivalent.

This rule complies with the recommendations for the unambiguous representation of stereochemistry in two-dimensional structures from the International Union of Pure and Applied Chemistry (IUPAC). These guidelines are given in the report: J. Brecher, "Graphical Representation of Configuration (IUPAC recommendations 2006)" *Pure Appl. Chem.* (2000), **78**(10), 1897-1970.

## Stereochemistry of Allenes and Biaryls

Allenes of the type abC=C=Ccd (or abC=C=Cab) and ortho-substituted biaryls with hindered rotation possess an *axis of chirality*. This section describes how to indicate the three-dimensional stereochemistry of these structures in a two-dimensional drawing.

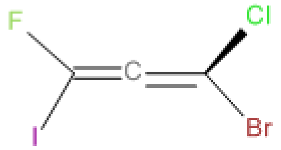
### Perception of Stereoconfiguration

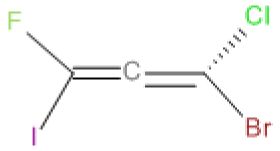
As with tetrahedral stereochemistry, the stereochemical perception algorithm uses information on the x-y coordinates of the atoms in the structure. However, the algorithm does not use information from any Up or Down stereo bonds that are attached to the asymmetric double bond.

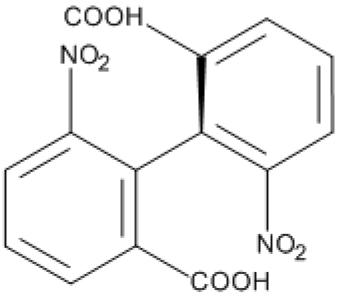
### Rules for Drawing Allenes and Biaryls

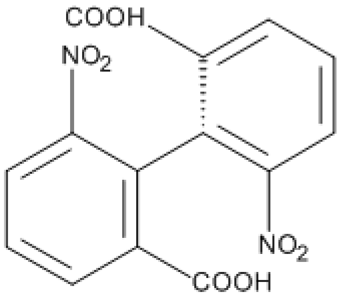
Use the following guidelines for drawing structures that you register to your database:

- If you know the absolute configuration of the structure, use an Up or Down bond to indicate stereochemistry. For example:

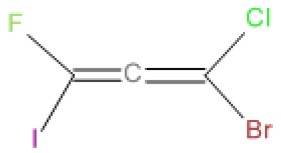


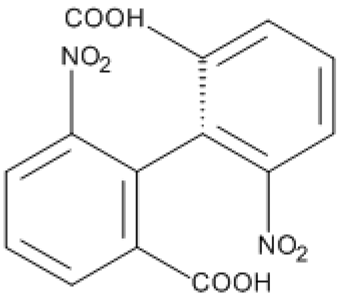






- Do not use stereo bonds if the absolute configuration of the allene or biphenyl is unknown. For example:

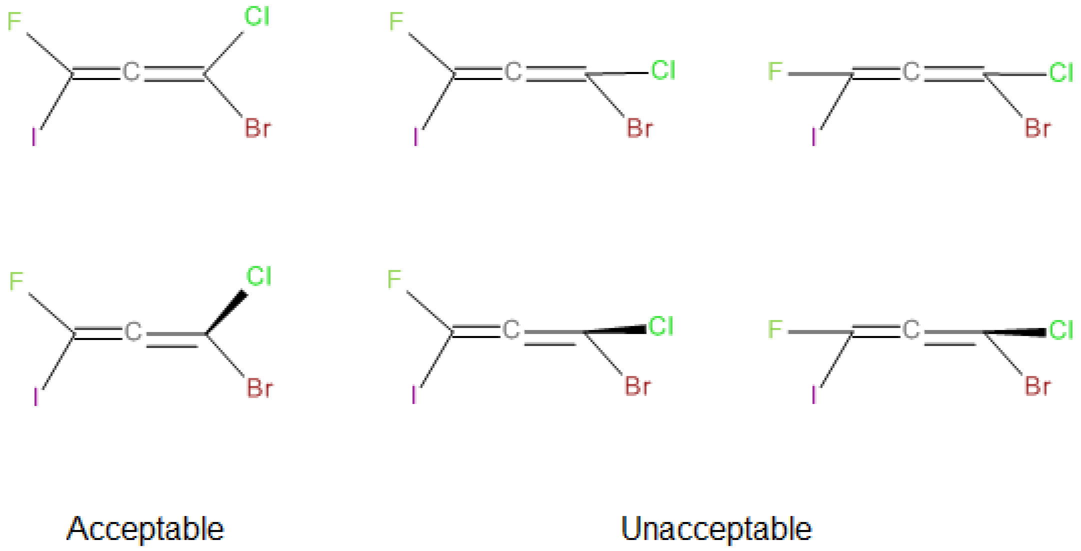




The following guidelines for tetrahedral stereochemistry and asymmetric double bonds also apply to these structures:

- One Stereo Bond per Stereogenic Center on page 227
- Wedge Bond Rule on page 229

- Colinearity Rule on page 230. For example, bonds to substituents of allenes cannot be colinear:



## Stereochemistry of Asymmetric Double Bonds

Double bonds of the type abC=Ccd (or abC=Cab) have cis-trans stereoisomerism. This section describes how to indicate the three-dimensional stereochemistry of these structures in a two-dimensional drawing.

### Perception of Stereoconfiguration

As with tetrahedral stereochemistry, the stereochemical perception algorithm uses information on the x-y coordinates of the atoms in the structure. However, the algorithm does not use information from any Up or Down stereo bonds that are attached to the asymmetric double bond. Instead, the algorithm calculates stereoconfiguration as follows:

- The E stereoconfiguration on the basis of CIP rules 0-2.
- The Z stereoconfiguration on the basis of CIP rules 0-2.
- Unknown: The E or Z stereoconfiguration cannot be determined on the basis of CIP rules 0-2.
- Potential: The bond is a Double Either stereo bond. Hence the stereoconfiguration might be either E or Z.
- None: The bond is not a geometric asymmetric double bond or Double Either bond.

The following conditions can cause in an unknown stereoconfiguration:

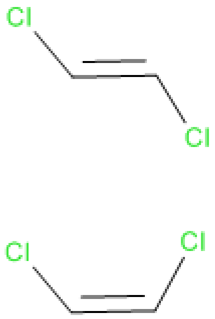
- The algorithm cannot order substituents using the atomic number and mass rules (CIP 0-2).
- The algorithm encounters an "indeterminate atom" as it is ordering ligands. Such "indeterminate atoms" include query atoms (A, Q), the Atom List query feature, Rgroup attachment points, and Rgroup atoms.

### Rules for Drawing Structures with Cis/Trans Geometric Stereochemistry

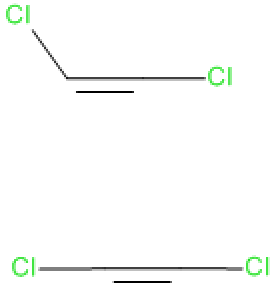
#### Known Configurations

If the configuration is known and specified in the original literature, draw the structure as specified.

The colinear rule applies to cis/trans geometric double bonds. That is, the double bond must not be colinear with one or both attached single bonds. For example:

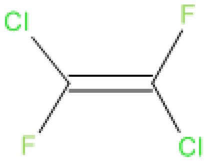


Acceptable

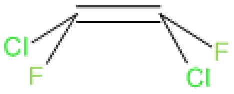


Unacceptable

The two attachments of each atom of a double bond must not be on the same side of the double bond. For example:



Acceptable



Unacceptable

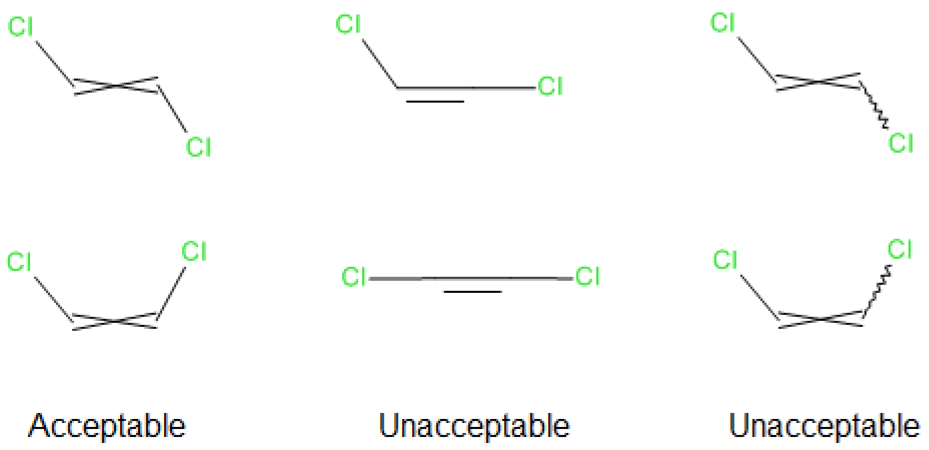
## Unknown Configurations and Mixtures

If the configuration is not specified in the original literature or if the author specified a mixture of *cis* and *trans*, draw the structure using the Double Either bond. Either bonds are intended to mark unknown tetrahedral centers.

Do *not* use an Either (squiggle) stereo bond to represent unknown stereoconfiguration of an asymmetric double bond.

Do *not* use a colinear bond arrangement to indicate an unknown configuration.

Examples:



## BIOVIA Stereo Parity

Stereo parity comparisons are made in any operation in which one structure (conventionally referred to as a *query structure*) is *mapped* to another structure, referred to as the target structure. For example:

- In product-based enumeration, a Markush reactant is mapped to each specific reactant to generate the fragments that are assembled into the products. See Reaction-based Enumeration on page 48.
- In scaffold-based enumeration, building blocks are created by mapping a clipping rule to a reactant structure. See Scaffold-based Enumeration on page 47.

Different types of stereo-isomerism use different parity calculations:

- For tetrahedral stereogenic centers, the algorithm calculates the parity of the atom at the stereogenic center (atom stereo parity).
- For allenes and biaryls, the algorithm calculates the parity of a collection of bonds (stereo parity of a collection).
- For asymmetric geometric double bonds, the algorithm calculates the parity of the double bond (bond stereo parity).

BIOVIA software uses parity calculations to compare stereo-configurations of mapped structures because parity calculations always provide a value that can be compared. In contrast, it is not always possible to calculate a stereoconfiguration from CIP rules. Moreover, parity calculations are much faster than calculations that use CIP rules.

The sections that follow describe these parity calculations.

### Stereo Parity of an Atom

The calculation of atom parity uses the following information:

- Whether the atom is stereogenic, that is, whether the atom is one of the following types: C, N, Si, P, As, S, Se, Te or any isoelectronic equivalent (for example, B-).
- The number of explicit and implicit hydrogen atoms attached to the atom.
- The order of the atom's attachments in the bond section of the molfile connection table (CTAB).
- Whether the atom has attached Up, Down, Either, or single bonds.

The parity calculation cannot determine whether the atom is a true stereogenic center, because it examines solely the atoms that are directly attached to the stereogenic atom.

Known tetrahedral stereocenters are marked with an "Even" or "Odd" AtomStereo. Even and odd are calculated in a similar manner to R and S chirality. However, for increased speed and simplicity, they are calculated using the order of the bonds on the central atom as the priority order instead of the rigorous CIP priority rules:

1. Number the atoms surrounding the stereo center with 1, 2, 3, and 4 in the order that they appear in the central atom's list of attachments. Implicit hydrogens are always numbered 4 and given coordinates opposite the other three attachments (that is, the negative average of the other three vectors).
2. The center is viewed from a position such that the bond connecting the highest numbered atom (4)projects behind the plane formed by atoms 1, 2, and 3.
3. The remaining three atoms (1, 2, and 3) are arranged in a clockwise or counterclockwise direction in ascending numerical order. The application assigns Even for a clockwise arrangement and Odd for a counterclockwise arrangement.

## Relationship to CTAB's Stereo Parity

This definition is analogous to the BIOVIA CTAB (or MOL-file) stereo parity definition with the exception that MOL files use the atom's index in the molecule as the priority order and assigns "1" for a clockwise and "2" for a counterclockwise arrangement. AtomStereo can be related to MOL- file StereoParity using the Atom::HasEvenParity() function, which analyzes the relationship between the atom order and the bond order. If an atom has PP AtomStereo and PP AtomParity that are both even or both odd, the MOL-file Parity is "1". Otherwise, the MOL-file Parity is "2".

## Relationship to Up and Down Wedge Bonds

In many formats, including BIOVIA CTAB, tetrahedral stereo information is persisted in the form of wedge bonds emanating from the stereocenter. The readers automatically perceive Even, Odd, or Unknown stereo for atoms that are marked with wedge bond. No perception is performed for possible stereo centers that are unmarked (that is, no attached wedges). To increase speed, these marked centers are not validated. The parity is determined from the order of the attachments as described above. Then the atom stereo marking is the canonical location for the stereocenter configuration. The wedge bond is kept as a display feature only. When modifying the stereocenter with the molecular toolkit, it is the atom stereo that should be modified. To synchronize the wedge bonds with the atom stereo, use the RepositionStereoBonds option in the *Standardize Molecule* component.

## Stereocenters in 3D CTAB Files

In 3D CTAB files, there are no wedge bonds, so Even and Odd parities are not perceived on import. In most cases, the parity markings are not needed as information on the configuration is included in the atomic coordinates. It is useful to assign parities on stereo atoms before converting the 3D molecule into 2D representation either with the 2D Coords component or by exporting to a 2D format such as SMILES. To perform this task, use the SetStereoFromCoordinates option in the Standardize Molecule component.

## Limitations of Parity Calculation

The parity calculation uses only the atoms that are directly attached to the stereogenic center, and prioritizes these atoms solely by their order in the molfile connection table (CTAB). The parity calculation does *not* check the atom attachments to discern whether the attachments are unlike, nor does it use the Cahn-Ingold-Prelog (CIP) priority rules to establish the R or S stereoconfiguration.

## Stereo Parity of a Bond

A parity value can be calculated for an asymmetric geometric double bond. Unlike stereo parity at tetrahedral stereogenic centers, the parity is calculated with reference to a bond.

The calculation of parity on a bond uses the following information:

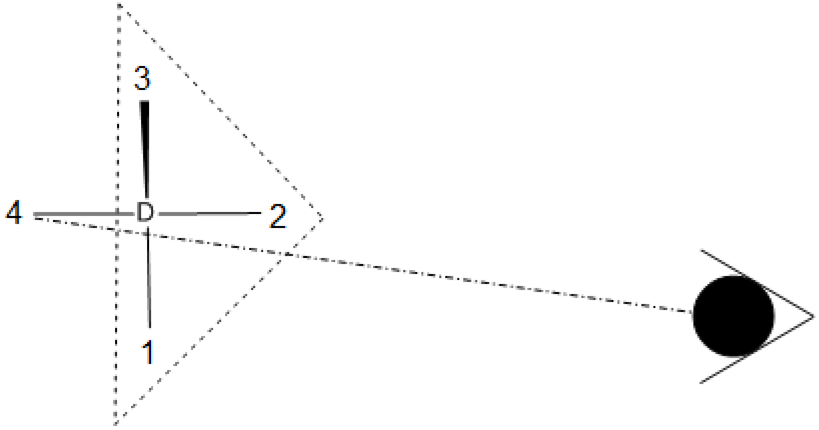
Bond type, for example, Double, Double Either, or another bond type. Bond parity can have one of four values, as shown in the following table:

| Parity Description | Meaning |
|---|---|
| cis | Indicates that the bond is a possible asymmetric geometric double bond with the cis stereoconfiguration. |
| trans | Indicates that the bond is a possible asymmetric geometric double bond with the trans stereoconfiguration. |
| unknown | Indicates that the bond is a possible asymmetric geometric double bond with a configuration of either Z or E because the bond type is Double Either. |
| none | Indicates that the bond cannot be a possible asymmetric geometric double bond because the bond type is neither Double nor Double Either. For example, a bond that is perceived as aromatic has no parity. For information on the type of bonds that are perceived as aromatic, see Aromaticity on page 7. |

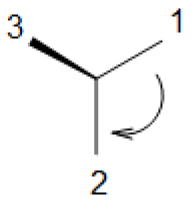
### Parity Calculation

Stereo parity of atoms with at least three non-hydrogen atom attachments is determined using a geometric calculation that can be visualized as follows:

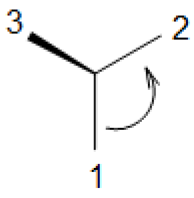
1. Mark a bond attached at a stereogenic center Up or Down to define the configuration.
2. Number the atoms surrounding the stereogenic center with 1, 2, 3, and 4 as follows: Assign the highest numbers to hydrogen atoms, with other atoms ordered by their position in the atom block of the molfile connection table (CTAB).
3. View the center from a position such that the bond connecting the highest-numbered atom (4) projects behind the plane formed by atoms 1, 2, and 3. In the figure, atoms 1, 2, and 4 are all in the plane of the paper, and atom 3 is above the plane.



4. Sighting towards atom number 4 through the plane (123), you see that the three remaining atoms can be arranged in either a clockwise or counterclockwise direction in ascending numerical order.



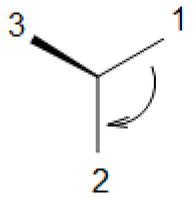
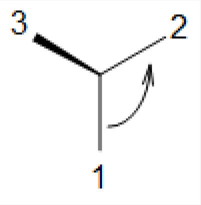
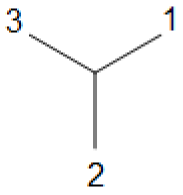
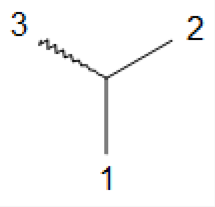
odd parity

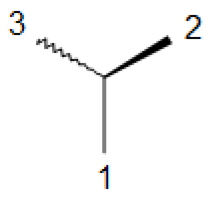
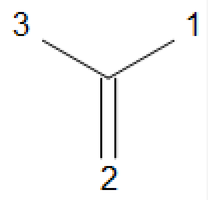
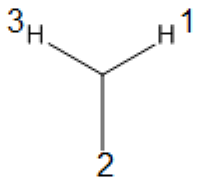


even parity

In this example, a parity value can be calculated for the atom even though it is not a stereogenic center (three of the atom's attachments are identical).

The following table shows examples of parity values:

| Atom with Attachment | Parity | Meaning |
|---|---|---|
|  | odd | Central atom is a possible defined stereogenic center with odd parity. |
|  | even | Central atom is a possible defined stereogenic center with even parity. |
|  | potential | Central atom is a possible undefined stereogenic center, because it lacks an Up or Down stereo bond to define the parity value and stereoconfiguration. |
|  | potential | Central atom is a possible undefined stereogenic center, because it lacks an Up or Down stereo bond to define the parity value and stereoconfiguration. An Either bond is equivalent to a single bond. |

| Atom with Attachment | Parity | Meaning |
|---|---|---|
|  | none | An Either bond negates any other Up or Down bonds that are attached to the stereogenic center. |
|  | none | Central atom cannot be a stereogenic center because of the double bond to carbon. |
|  | none | Central atom cannot be a stereogenic center because the atom has more than one hydrogen attachment. |
|  | none | Central atom cannot be a stereogenic center because stereochemistry is not perceived at Ge atoms. |

### Limitations of Parity Calculation

The parity calculation uses only the atoms that are directly attached to the double bond, and prioritizes the attached atoms solely by their order in the molfile connection table (CTAB). Thus, the same limitations apply as for atom parity calculation.

## Axial Parity of (Allenes and Biaryls)

A parity value can be calculated for allenes and biaryls. Unlike stereo parity at tetrahedral stereogenic centers, the parity is calculated with reference to a collection of bonds:

- For allenes, the collection contains the two double bonds.
- For biaryls, the collection contains the bond between the rings.

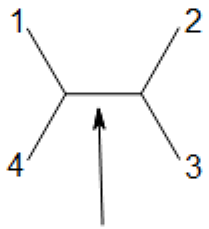
The calculation of parity on the collection of atoms and bonds uses the following information:

- The order of the atoms attached to the bond in the atom block of the molfile connection table (CTAB).
- (Allenes only) Whether the atoms in the collection contain any atoms other than C or Si.
- (Biaryls only) Whether at least three non-hydrogen atoms are attached at the ortho-positions on the aryl rings. The presence of two or more hydrogen atoms at the *ortho* positions allows free rotation about the bond between the aryl rings. Biraryls with free rotation about the bond are not stereogenic.

The parity calculation is similar to that of tetrahedral centers. The four attachments are in tetrahedral-like positions but instead of the center being a single atom, the center is a bond or group of bonds which have restricted rotation.

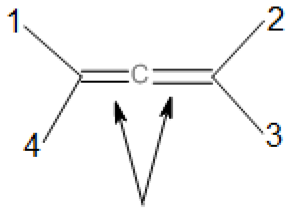
## Parity Calculation

The calculation of axial parity prioritizes attachments based on the order they occur in the bond block of the molfile connection table. No special consideration is made for explicit hydrogen attachments. The figure that follows show the bonds and the attached atoms that are used in the parity calculation:



Bond collection used in parity calculation for biaryls (bond between aryl rings)

1, 2, 3, 4 = Atoms used in parity calculation for biaryls (ring atoms *ortho* to bond)



Bond collection used in parity calculation for allenes

1, 2, 3, 4 = Atoms used in parity calculation for allenes

### Limitations of Parity Calculation

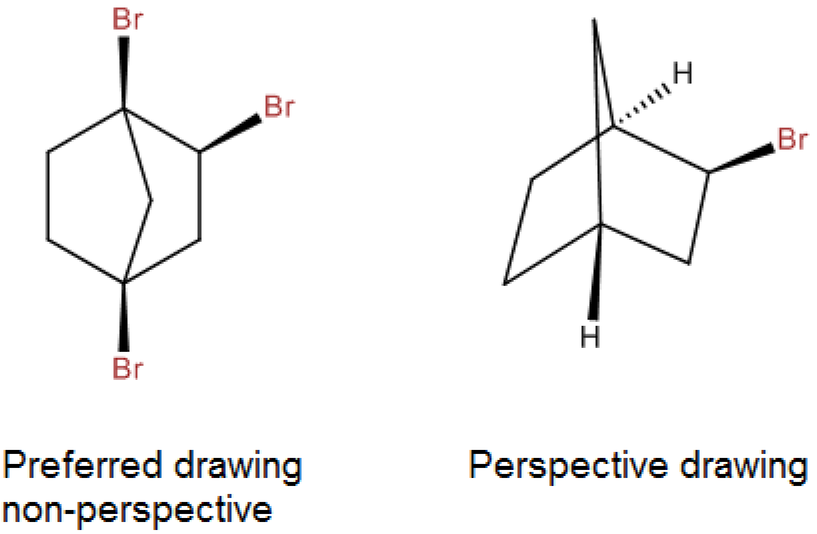
The parity calculation uses only the atoms that are directly attached to the collection, and prioritizes the attached atoms solely by their order in the molfile connection table (CTAB). Thus, the same limitations apply as for atom parity calculation.

## Converting Conventional Depictions to Acceptable Drawings

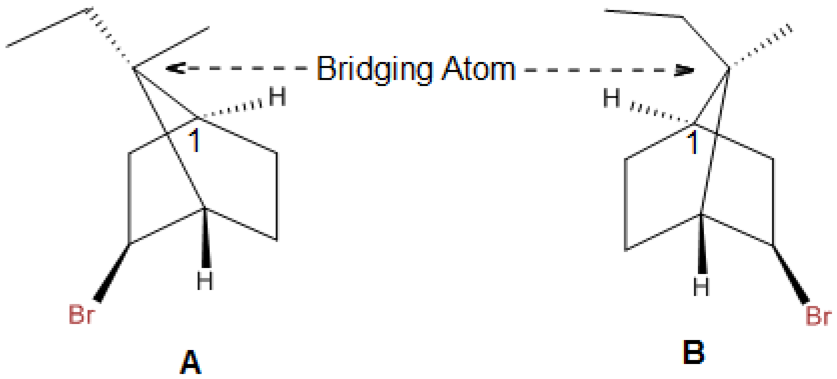
### Perspective Drawings

The following rules apply when drawing polycyclic molecules with bridging atoms:

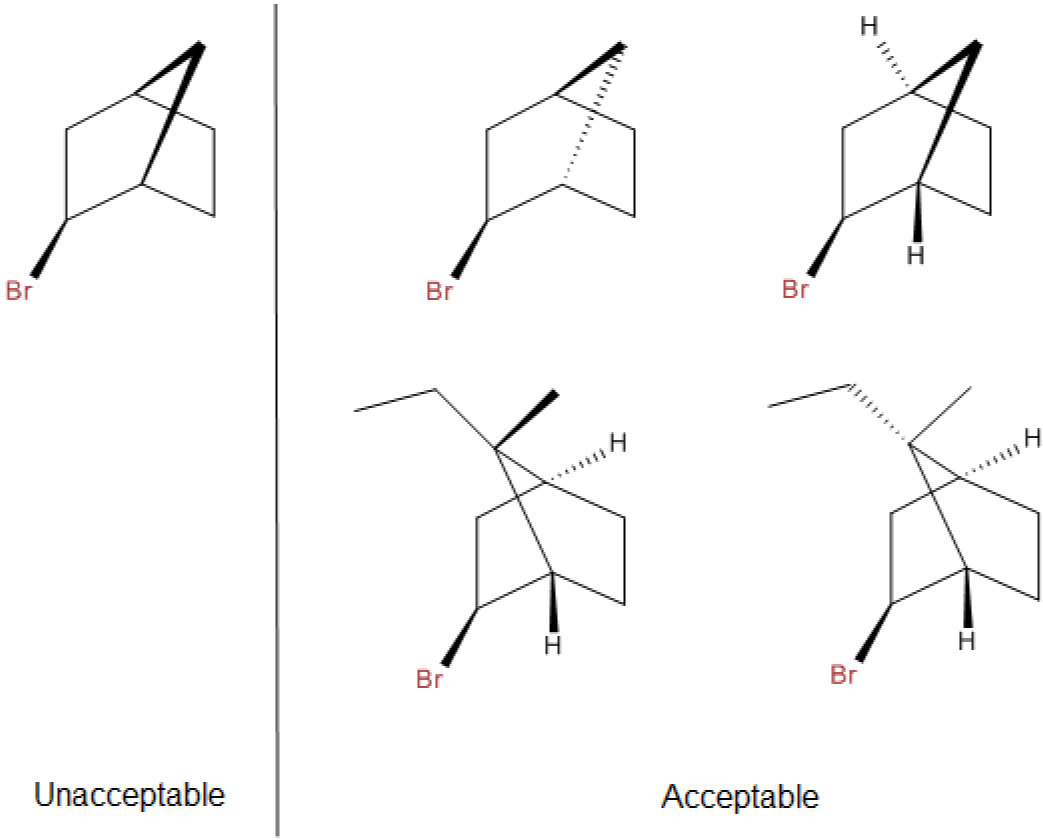
- Whenever possible, make non-perspective drawings.



- Whenever possible, do not mark the bonds that are attached to the bridging atoms as stereo bonds. Use explicit hydrogens instead.
- When marking a stereo bond to an asymmetric bridgehead atom that is bonded to hydrogen, you should:
  - Draw the hydrogen atom explicitly.
  - Mark the bond to hydrogen as the stereo bond.
- When marking stereo bonds for substituents attached to an asymmetric bridging atom, the following applies:
  - When the bridging atom is drawn to the left of atom 1 (the rear bridgehead), the left substituent is marked Down (see drawing A below).
  - When the bridging atom is to the right of atom 1, the right substituent is marked Down (see drawing below).



The following example shows acceptable and unacceptable drawings.

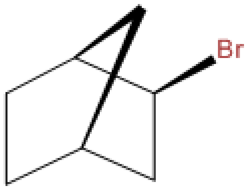


**Tip:** To understand the rules in this section, it can be helpful to refer to a three-dimensional model of the bicyclic structure shown here. For example, the changes that are described in Rule 4 correspond to what you see when you rotate the three-dimensional structure from the orientation in drawing A to that in drawing B. In the examples that follow, changes in stereo bonds (for example, changing Up to Down or changing which bond is marked as Up or Down) result when you view the structure from the side rather than from above.

## Drawing Stereo Bonds to the Non-Stereogenic Bridging Atom

Stereo bonds should not be drawn from the bridgehead to the bridging atom unless unavoidable. Instead, use explicit hydrogens to indicate stereochemistry as in the example above (drawing A).

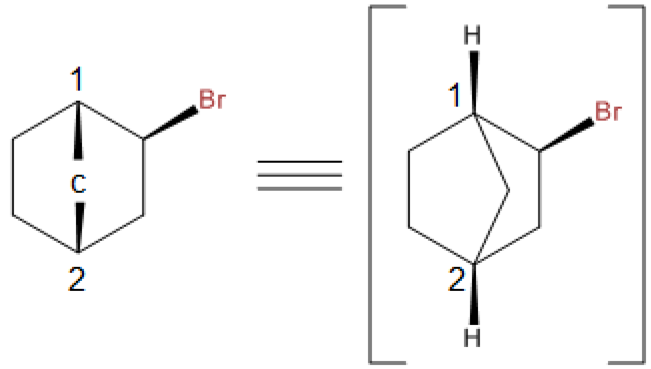
When you encounter a molecule such as molecule C, it might seem at first glance that you should mark both bonds leading to the bridging atom as Up; however, that is not the case.



**C**

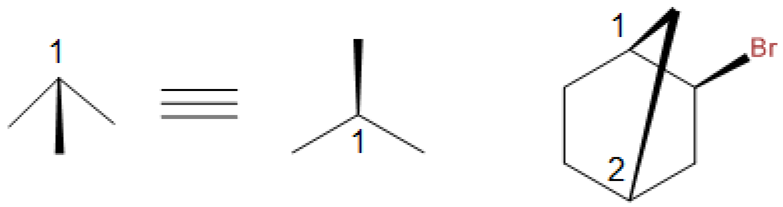
Unacceptable

To illustrate this, you can derive how to mark the two bonds by beginning with a non-perspective drawing, molecule D. When you look down on the bridging atom, it is coming out toward you, and both bonds should be marked Up as shown.



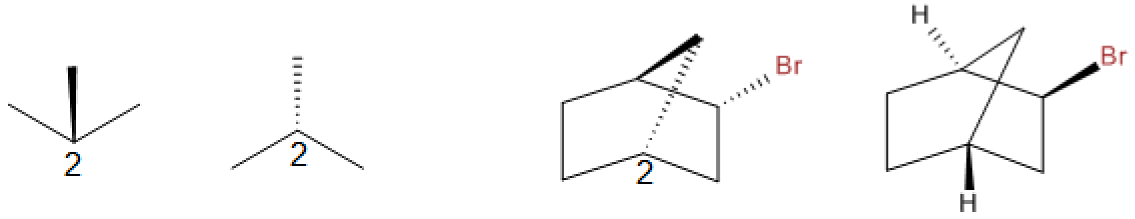
D

Moving the bridging atom Up as shown in molecule E does not change the stereochemistry about atom 1. For more information see Rules for Stereogenic Centers with Implicit Hydrogen and Three Explicit Attachments on page 232.



E

Now move atom 2 to form a re-entrant angle (a polygonal interior angle greater than 180 degrees). Changing the angle as shown reverses the configuration about atom 2. Therefore, the bond to the bridging atom must be marked Down in order to maintain the stereochemistry of atom 2.



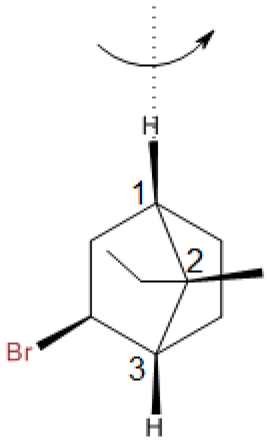
The correct way to mark stereo bonds to the bridging atom is shown in molecule F. This is true as long as both the bridging atom and one of the bridgehead atoms 1 or 2 are not stereogenic; there should be no stereo bonds marked to the bridging atom in that case. In general, marking stereo bonds to the bridging atom is discouraged.

## Marking Bonds to Stereogenic Bridging Atoms

Be careful when marking bonds in perspective drawings that involve the bridging atom. To illustrate how to mark such bonds, a non-perspective drawing is used as a starting point and two examples are discussed.

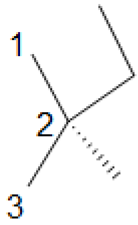
### Example 1

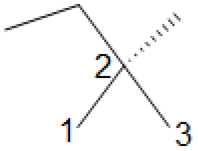
When looking down upon the bridging atom as in molecule A, the two substituents attached to the bridging atom are both directed out of the plane toward the observer and stereoconfiguration at the bridging atom is indicated by marking one of the groups Up.

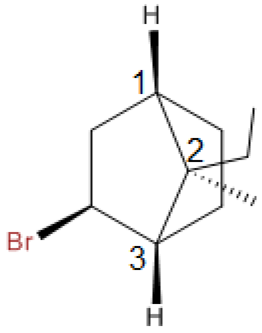


**A**

Rotating the bridging atom about the y axis, as shown in drawing A, gives you molecule B. The methyl group that was the substituent to the right in molecule A is now Down.

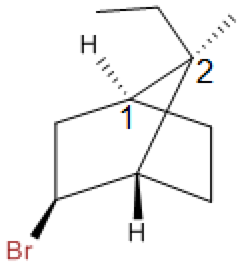




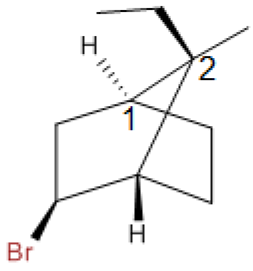


**B**

Moving the bridging atom 2 up and to the right of bridgehead 1 as shown in molecule C does not change the stereochemistry about the bridging atom and the methyl group is still Down. (However, the relative positions of substituents on atom 1 have changed.) Drawings C and D are acceptable.



C

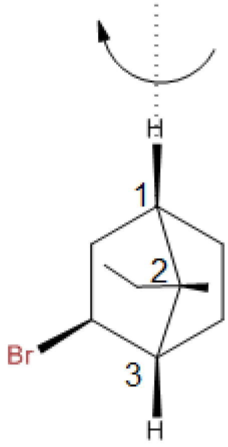


D

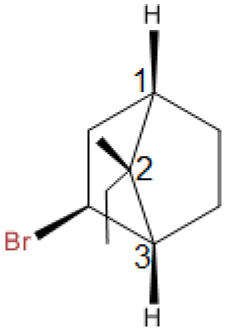
**Note:** To preserve the stereoconfiguration at atom 1, the stereo bond to the explicit hydrogen must change from Up in drawings A and B to Down in drawings C and D. This difference corresponds to what you see when you change from viewing the three-dimensional structure from above (A and B) to viewing it from the side (C and D).

## Example 2

Beginning again with drawing A, rotate the bridging atom about the y axis in the other direction. The methyl group is Up.

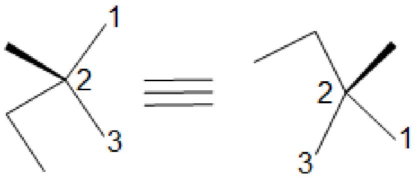


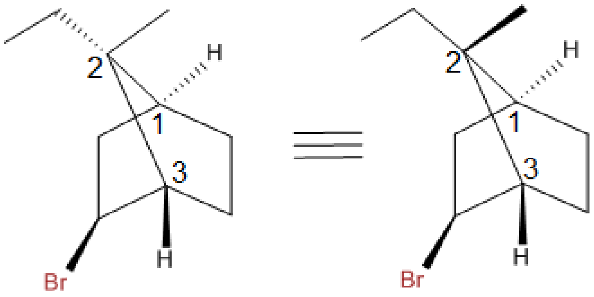
A



E

Moving the bridging atom 2 up and to the left of atom 1 as shown in molecule F does not change the stereochemistry at the bridging atom and the methyl group remains Up. (Again, however, the stereochemistry was reversed at atom 1.) This is also an acceptable drawing. From Postulate A, molecule G is an equivalent drawing and is also acceptable.



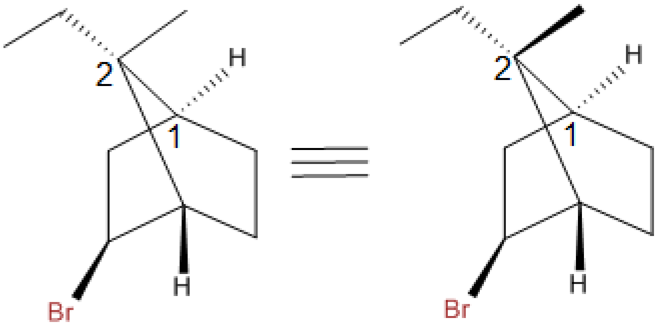


Although both molecules are equivalent, molecule F is the preferred representation. In molecule G, tilting the non-stereo bond at atom 2 by only a small angle would cause the structure to violate the Triangle Rule. In that case, the stereoconfiguration would be inverted, and the two structures would no longer be equivalent. Placing the stereo bond on the right eliminates the problem.

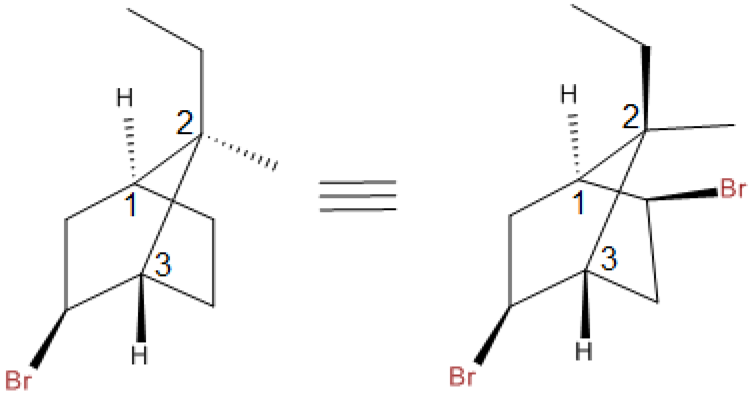
## Marking Bonds to Substituents

You need to consider the location of the bridging atom with respect to the bridgehead atom 1 in marking the bonds to its substituents. For example, consider the stereogenic bridging atom in the perspective drawings that follow:

- When the bridging atom 2 is to the left of atom 1, mark the left substituent with a Down stereo bond.

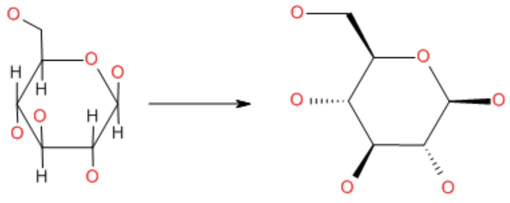


- Alternatively, mark the right substituent with an Up stereo bond. When the bridging atom 2 is to the right of atom 1, mark the left substituent with an Up stereo bond. Alternatively, mark the right substituent with a Down stereo bond.

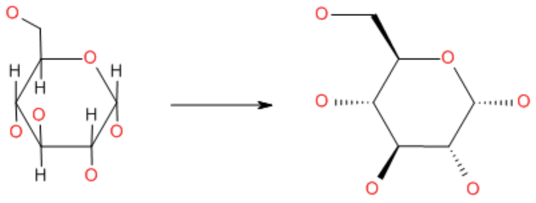


## Haworth Projections

Haworth projections are easy to translate into acceptable drawings. Bonds above the plane of the carbon ring are marked Up, and bonds beneath the plane of the carbon ring are marked Down. Hydrogens, which are explicit in Haworth projections, are implicit in structures that you draw for registration:



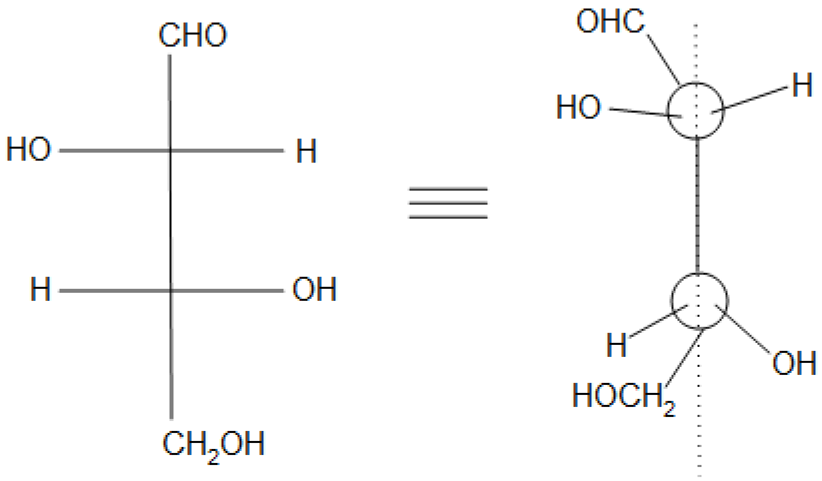
β-D-Glucose



α-D-Glucose

## Fischer Projections

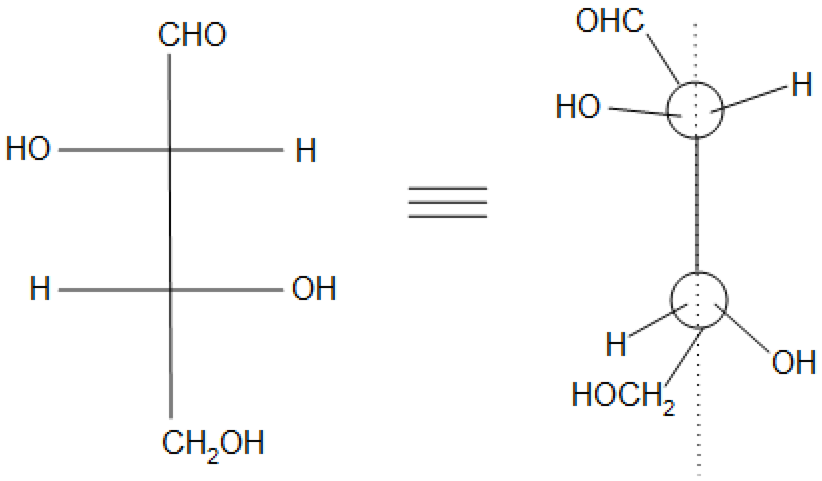
In a Fischer projection, the horizontal lines represent bonds coming out towards the observer, and vertical lines represent bonds going away from the observer behind the plane of the paper.



Two examples will be used to show how to translate a Fischer projection into an acceptable drawing for registration (use of a molecular model while going through each step can be helpful). While following the examples below, keep in mind the following: Flipping about an asymmetric center inverts its stereochemistry. Therefore, the stereo bond connected to a "flipped" atom must be reversed.

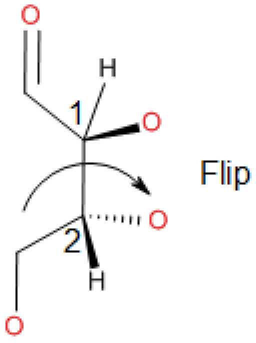
### Example 1: Two Stereogenic Centers

The Fischer projection for D-threose can be visualized as shown in drawing A below:



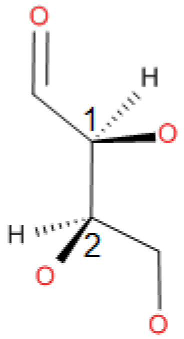
A

Now rotate the molecule 90 degrees about a vertical axis so that the bonds that were projecting into the plane of the paper are now lying in the plane of the paper. As you can see, the horizontal bonds that were to the left of the vertical bonds (in the projection) are coming out of the plane, and those that were to the right are now directed into the plane of the paper.

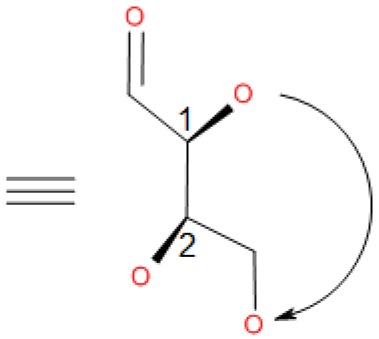


B

Flipping about atom 2 gives you molecule C. The -OH group that was a Down bond is now an Up bond. Drawing C is equivalent to drawing D. The explicit hydrogens have been removed.

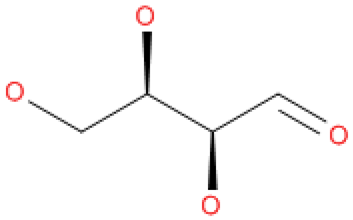


C

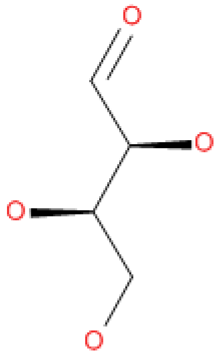


D

Rotating about an axis perpendicular to the plane of the paper gives rise to drawing E. Molecules E or F are acceptable drawings.

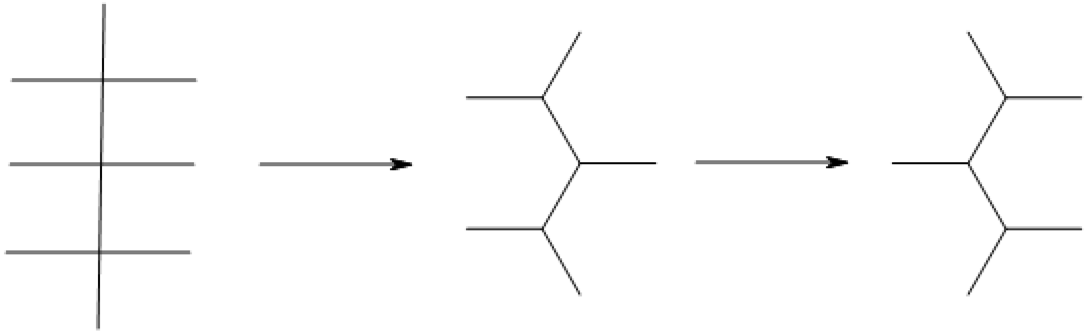


E



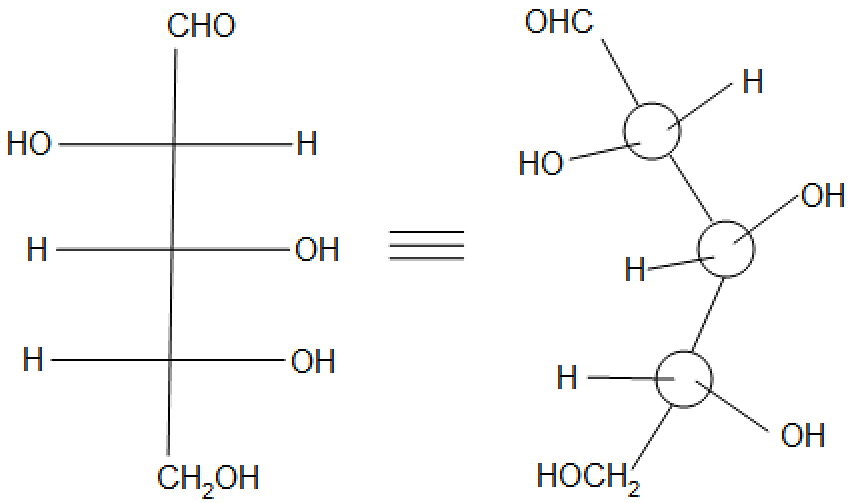
F

As you can see, to translate the Fischer projection into an acceptable drawing, the molecule was rotated so that the vertical bonds of the Fischer projection are coplanar with the plane of the paper. In addition molecules such as sugars are customarily drawn so that the molecule is in an extended form and the vertical bonds form a zig zag pattern:



## Example 2: Three Stereogenic Centers

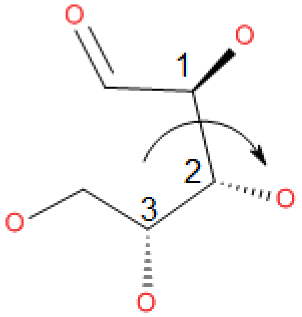
You follow the same procedure as previously for a molecule with three stereogenic centers. The Fischer projection can be visualized as in drawing F.



**F**

Rotating the molecule gives rise to molecule G.

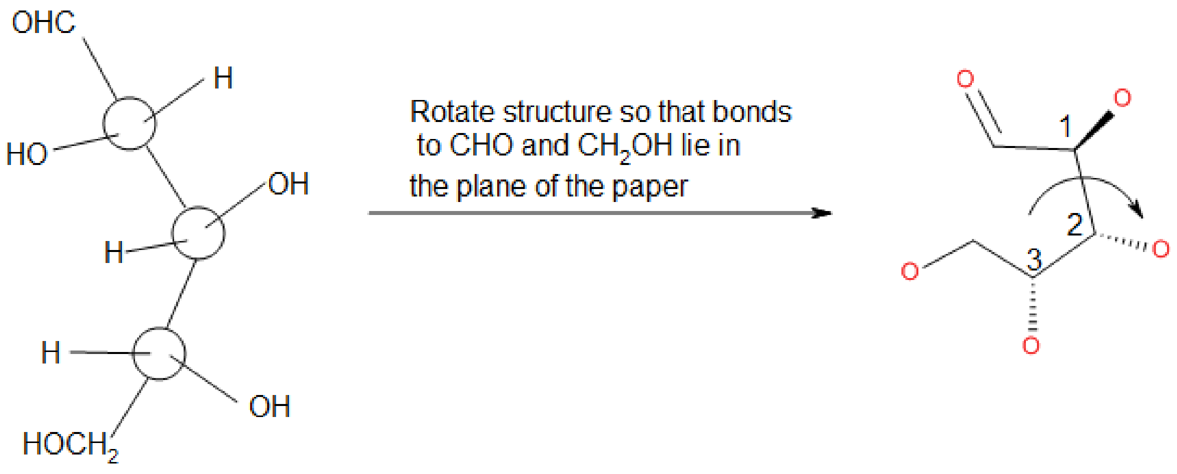
> **Note:** The bonds that were vertical in the projection form an arc, and the molecule is not in its most extended form.



**G**

Now flip molecule G about atom 2 using the bond between atoms 1 and 2 as the axis of rotation.

> **Note:** Atom 3 has been flipped also. (Remember, flipping inverts stereochemistry.)

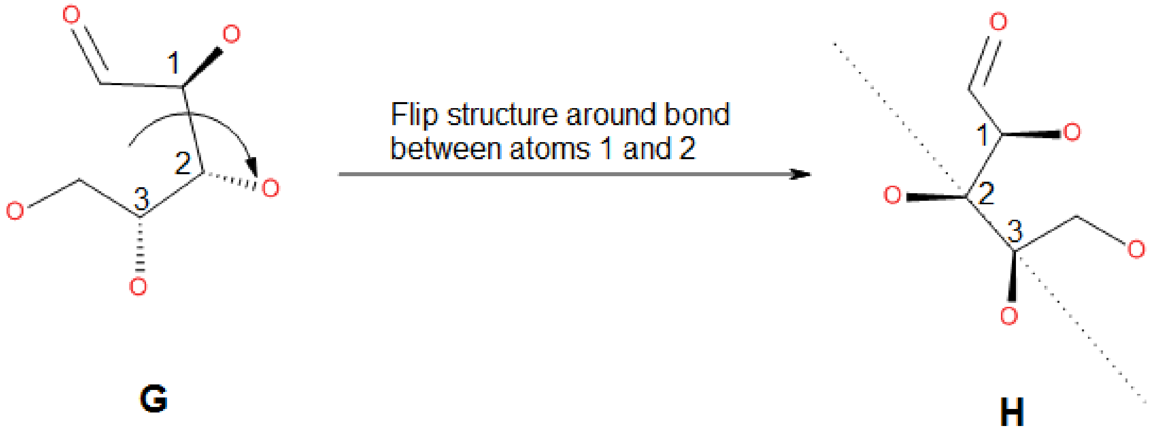


Rotate structure so that bonds to CHO and $CH_2OH$ lie in the plane of the paper

**F**

**G**

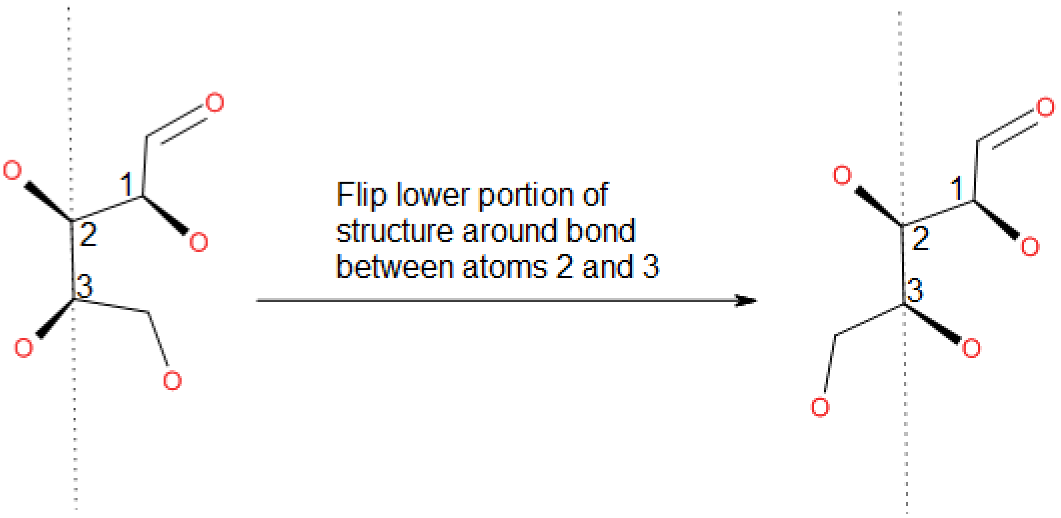
Flipping once again about atom 3 gives you molecule I.



Flip structure around bond between atoms 1 and 2

**G**

**H**



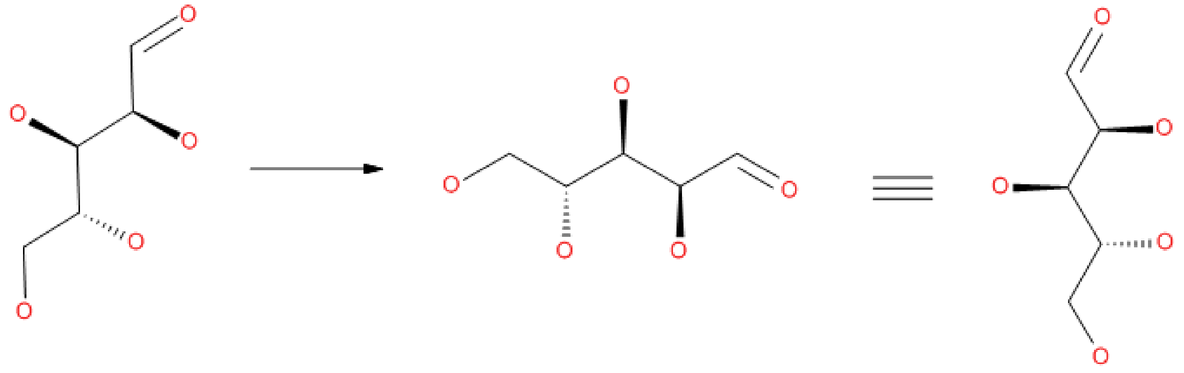
Flip lower portion of structure around bond between atoms 2 and 3

**H**

**I**

Rotating the molecule gives rise to molecule G.

**Note:** The bonds that were vertical in the projection form an arc, and the molecule is not in its most extended form.



I J K

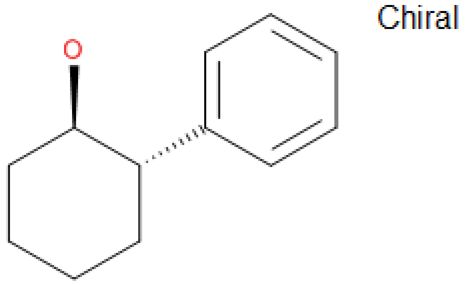
# Appendix B: Original Representation of Tetrahedral Stereochemistry

## Chiral Flag for Absolute Stereochemistry

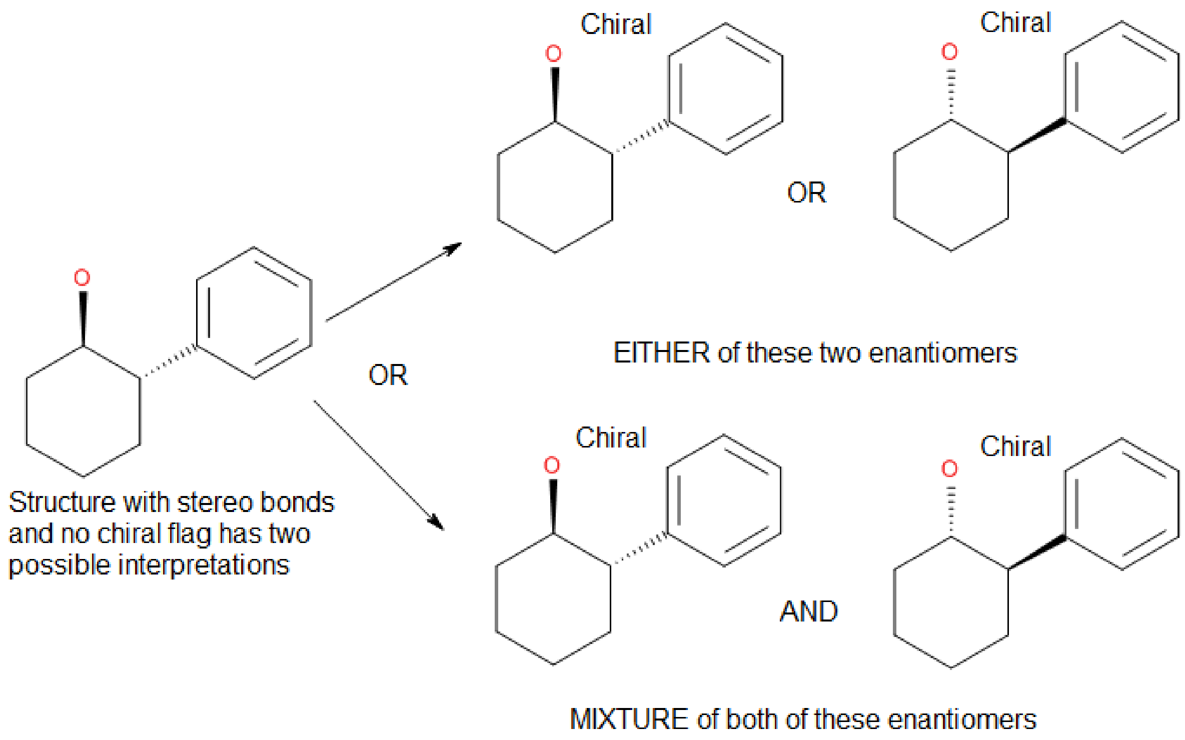
The original representation of tetrahedral stereochemistry uses a structural property called the chiral flag to indicate stereochemistry of an entire structure:

- Setting the chiral flag to On indicates that the structure as drawn represents a single stereoisomer whose absolute stereochemistry is known.
- Setting the chiral flag to Off indicates that the absolute configuration of the structure is unknown, but that the relative configurations of the stereogenic centers are known. The structure might represent either a single diastereomer or a mixture of the two stereoisomers.

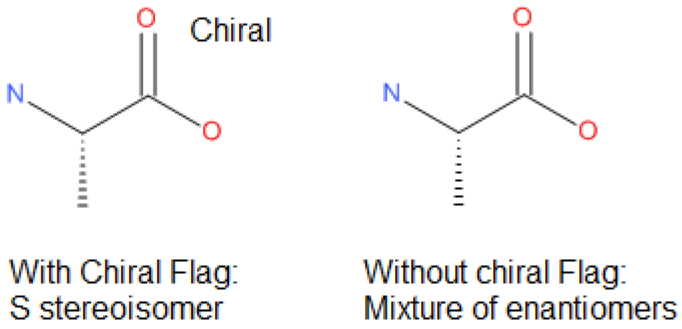
For example, the text label `Chiral` on the following structure indicates that the chiral flag is set to On. Therefore, the structure represents a single stereoisomer, (1R,2S)-(-)-trans-2-phenyl-1-cyclohexanol:



In the following figure, the absence of the text label `Chiral` on the structure indicates that the chiral flag is set to Off. The structure might represent either a single diastereomer or a mixture of the two stereoisomers:


Chiral
Chiral
OR
OR
EITHER of these two enantiomers
Structure with stereo bonds and no chiral flag has two possible interpretations
Chiral
Chiral
AND
MIXTURE of both of these enantiomers

For structures with only one stereogenic center, a structure with stereo bonds but without the chiral flag represents a mixture of enantiomers:



The original Accelrys representation has these limitations:

- The chiral flag setting cannot distinguish a mixture of stereoisomers from a single enantiomer in which you know the relative configuration of all the stereogenic centers.
- The chiral flag setting applies to all fragments in the structure. You cannot specify that one fragment represents a mixture of enantiomers and another represents the absolute configuration.
- There is no way to specify different levels of information for individual stereogenic centers. For example, in a multi-step synthesis in which one or more stereogenic centers are created at each step, you might know the absolute configuration at some centers, but only the relative configuration at others. There is no consistent method for representing different levels of information on the structure.

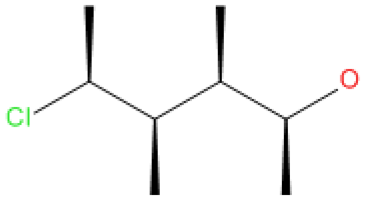
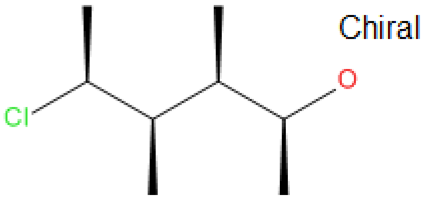
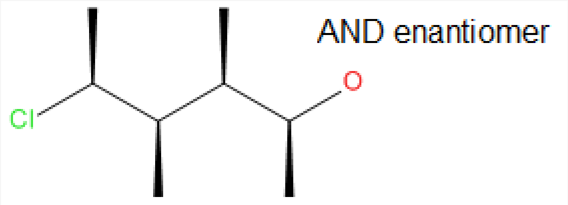
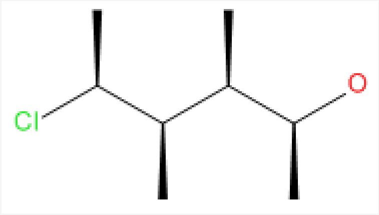
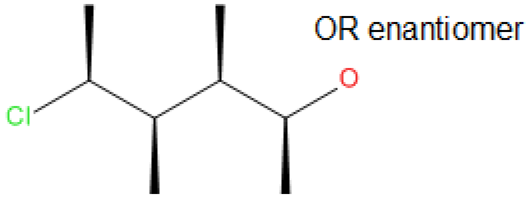
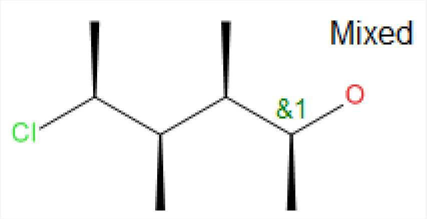
The enhanced stereochemical representation removes these limitations, as described in Tetrahedral Stereochemistry on page 9.

## Compatibility of Original and Enhanced Stereochemical Representations

The original stereochemical representation is a subset of the enhanced stereochemical representation. Consequently, users can continue to register structures that use your existing business rules until you adapt your business rules to the enhanced representation. Searching and registration of structures that use the original Accelrys representation are not affected.

### Structure Representation

The table that follows shows the equivalence of the original and enhanced stereochemical representation:

| Enhanced Representation | Original Representation |
|---|---|
|  |  |
|  |  |
|  | No equivalent |
|  | No equivalent |

The term "equivalent" means that structures with the enhanced stereochemical representation behave exactly the same as structures with the original stereochemical representation in searching and registration.

### Reaction Representation

Reaction components with marked stereogenic centers that lack the chiral flag are automatically assigned to distinct AND stereogroups when you read the reaction from a rxnfile or retrieve it from a database. For example, if each reactant and product in a reaction of the form A+B->C+D contains defined stereogenic centers that are not marked with the chiral flag (that is, each component contains a set of stereogenic centers that are part of a single AND stereogroup), then the AND centers in reactant A might be assigned to stereogroup &1; centers in reactant B might be assigned to stereogroup &2; centers in product C might be assigned to stereogroup &3; and centers in product D might be assigned to stereogroup &4.

The numeric values of the stereogroup index numbers 1, 2, 3, 4 .. have no significance other than to indicate that groups with different numbers are unrelated.

## Adapting Business Rules to Use the Enhanced Stereochemical Representation

After you adapt your business rules to use the enhanced stereochemical representation, you need to decide whether to revise the existing structures in your database to take advantage of the enhanced stereochemical representation. If you decide to convert existing structures to the enhanced representation, see Converting Existing Structures to Enhanced Representation on page 259 for a description of common workarounds for the limitations of the original stereochemical representation.

**IMPORTANT!**
If you decide not to convert existing structures to the enhanced stereochemical representation, your database will contain a mixture of structures that use two different sets of business rules: Your original business rules, based on the original stereochemical representation, and your revised business rules, based on the enhanced representation.

A search that retrieves structures from one set of business rules might not retrieve structures from the other. You can broaden your search criteria to ensure that you retrieve all structures: For flexmatch: set STE Off; for SSS, replace all Up and Down stereo bonds with single bonds. However, this strategy retrieves many unwanted structures.

## Converting Existing Structures to Enhanced Representation

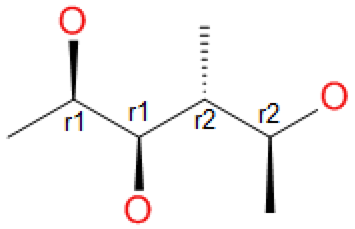
When you adapt your business rules to use the enhanced stereochemical representation, you need to decide whether to revise the existing structures in your database to take advantage of the enhanced stereochemical capabilities. This section contains examples of common workarounds for the limitations of the original stereochemical representation, and provides suggestions for structure conversion.

### Attached Data Marks Groups of Related Stereogenic Centers

The examples in this section use (STE On) and (DAT On) to indicate that the corresponding flexmatch switches are set On. For more information on the STE flexmatch switch in stereochemical searching, see Flexmatch Search of Structures with Tetrahedral Stereochemistry on page 118. For information on the DAT flexmatch switch, see Attached Data (DAT) on page 113.

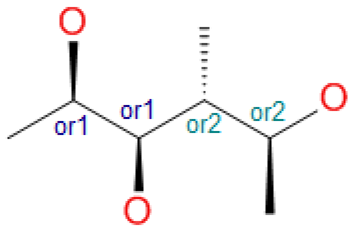
Attached data (Sgroup data) can be used to mark related stereogenic centers. In the following example, the structure on the left uses text labels (r1, r2, and so on) to mark groups of stereogenic centers with the same relative configuration. The structure on the right uses OR stereogroups:

Unrelated centers identified
by different Attachment Data



Using Attached Data to mark groups
of stereogenic centers with the
same relative configuration

Unrelated centers identified
by different Stereochemical Labels



Using Stereochemical labels to mark
groups of stereogenic centers with the
same relative configuration

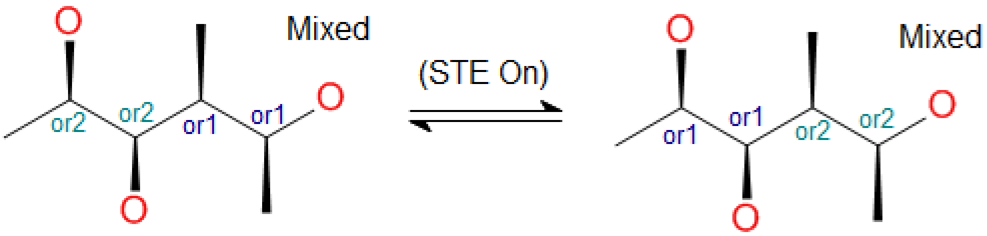
Both attached data and stereogroups can store precise information on the stereochemistry of individual stereogenic centers. However, converting the attached data representation to the enhanced stereochemical representation provides two advantages:

- No customization of BIOVIA Draw is required to use the enhanced representation.

  The Stereochemistry dialog, which applies stereogroup labels to structures, is enabled by default in BIOVIA Draw. If you want to keep your existing representation, you need to hide the Stereochemistry dialog and enable the Attached Data dialog, which is hidden by default. For information on customizing BIOVIA Draw, see the BIOVIA Draw Configuration Guide. This document is available from Start > Programs > BIOVIA > BIOVIA Draw [version] > BIOVIA Draw Documentation.

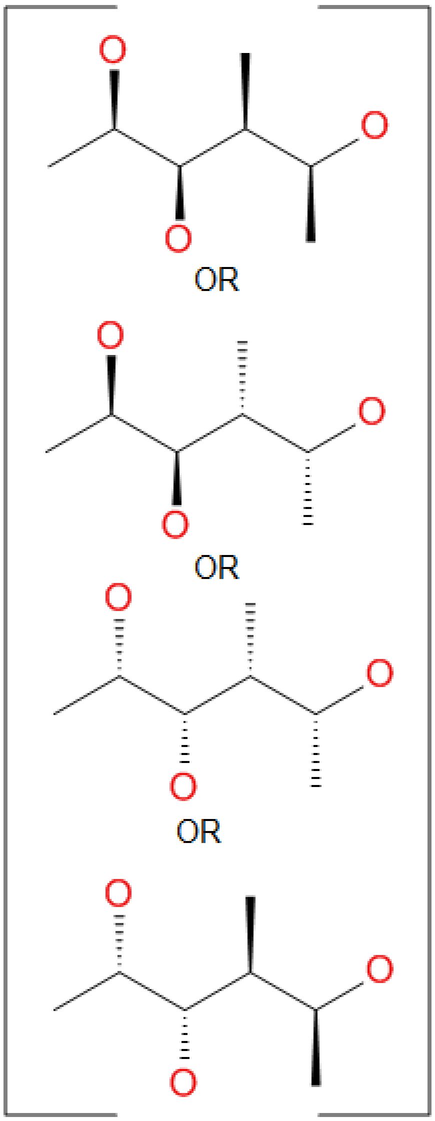
- Using attached data to identify groups of stereogenic centers can cause structures that should be perceived as duplicates to be perceived as different.

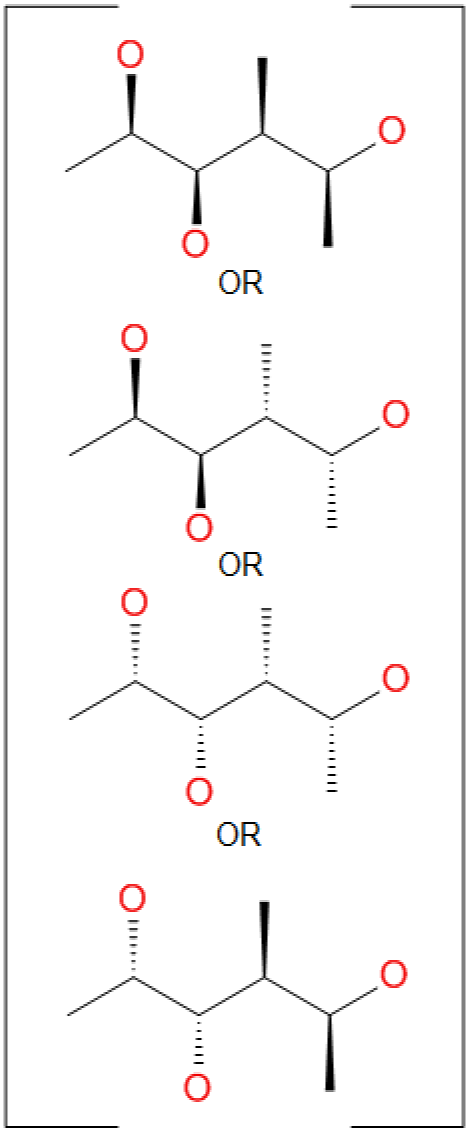
  To illustrate this, consider the structure given earlier in this section, which contains two sets of stereogroup centers whose relative configurations are known. The figure below shows two possible ways the structure can be drawn using the enhanced stereochemical representation:



Represents the following set of stereoisomers:

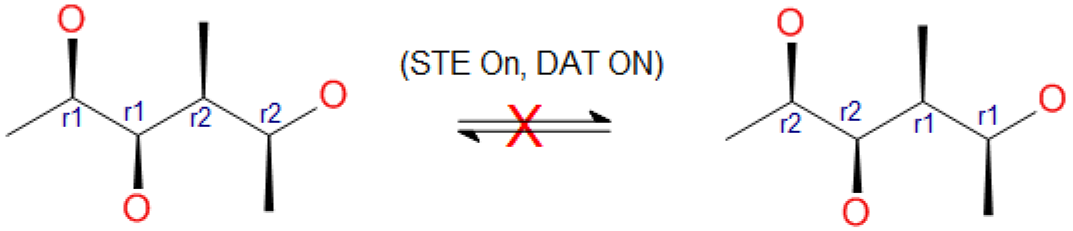
Represents the following set of stereoisomers:



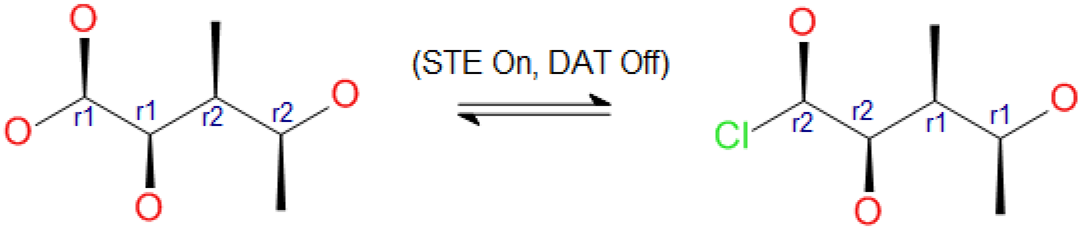


Although the index numbers on the stereogroup labels are reversed, the structures are equivalent because each one represents exactly the same set of stereoisomers. The structures find one another in a flexmatch search (with STE On), and are perceived as duplicates when you register them to the database.

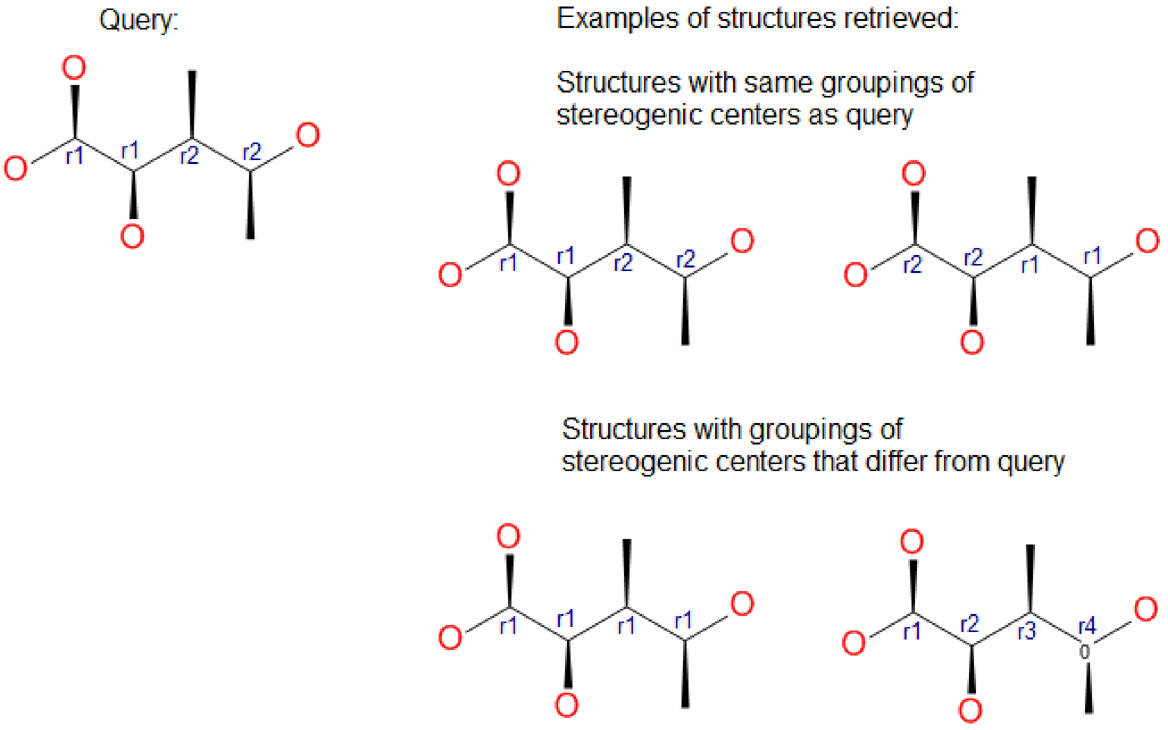
In contrast, consider what happens if you use attached data rather than stereogroup labels to mark the stereogenic centers. The figure below shows the same structures as the previous figure, but uses attached data to group stereogenic centers. Reversing the values of attached data results in two structures that are not equivalent, even though they define exactly the same relationships between the two sets of stereogenic centers:



You can make the structures equivalent by setting the flexmatch DAT switch to Off:

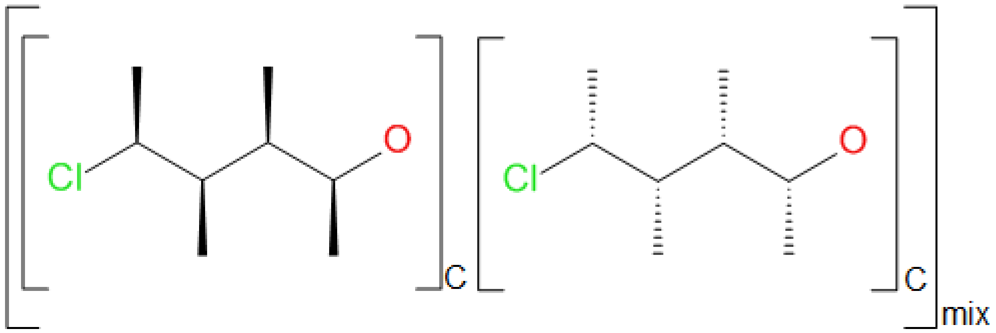


However, setting DAT to Off also retrieves a large number of structures in which the stereogenic centers are grouped differently. In the following example, the query matches not only structures with the equivalent stereogroup assignments, but also a structure with only one stereogroup and another structure with four stereogroups:

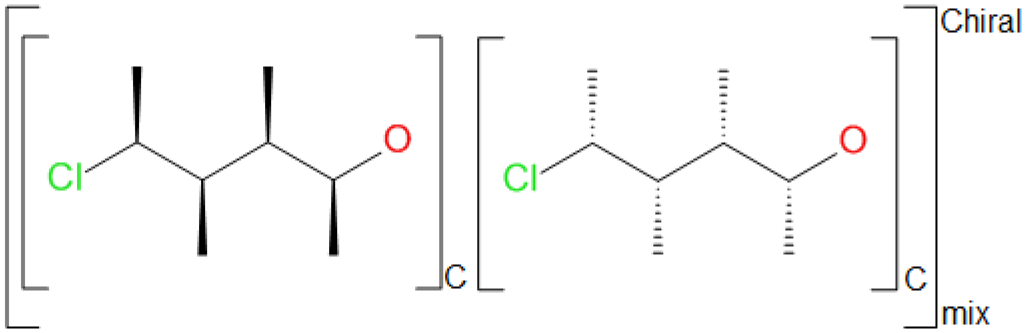

Query:
O
O
r1
r1
r2
r2
O
O
Examples of structures retrieved:
Structures with same groupings of stereogenic centers as query
O
O
r1
r1
r2
r2
O
O
O
O
r2
r2
r1
r1
O
O
Structures with groupings of stereogenic centers that differ from query
O
O
r1
r1
r1
r1
O
O
O
O
r1
r2
r3
r4
0
O
O

## Mixture and Component Brackets Represent Mixtures of Stereoisomers

You can represent a mixture of enantiomers as follows:



The original representation displays the mixture as:



In this case, you do not need to convert structures in your database, because all stereogenic centers belong to the ABS stereogroup. In the original representation, the chiral flag setting applies to both components of the mixture.

## No Distinction Between a Mixture of Enantiomers and a Single Stereoisomer

The original Accelrys representation did not distinguish a mixture of stereoisomers (AND) from a single stereoisomer whose relative configuration is known (OR). Consequently, your database might contain structures that represent a single stereoisomer of known relative configuration (OR enantiomer), but which are currently represented as mixtures (AND enantiomer). If you need to distinguish single stereoisomers from mixtures, then you need to convert the stereogenic centers on these structures from AND to OR.

## Undefined Stereogenic Centers Indicate Multiple Stereogroups

The original Accelrys representation did not allow you to define groups of related stereogenic centers on the same molecule. One strategy for dealing with this was to:

- Represent structures with multiple stereogenic centers without Up and Down stereo bonds, that is, with undefined stereogenic centers.
- Store information on stereoconfiguration separately from the structures, in a database field.

If you want to change these structures to take advantage of the enhanced stereochemistry, you need to add stereo bonds and stereogroup information.