WHITEPAPER

ACCELRYS' ENHANCED STEREOCHEMICAL REPRESENTATION

THE CHALLENGE

Applied synthetic chemistry has been placing increasing emphasis in recent years on stereochemistry. In the pharmaceutical industry, for example, there is a marked trend toward creating molecules with specifically known stereochemistry, while working with molecules with larger numbers of stereogenic centers. These trends demand a corresponding increase in the capability of informatics tools used to store, search, and retrieve molecules and reactions. During the development of a drug candidate many samples will be produced with increasing knowledge about the configuration of the stereogenic centers within the common structure. There is a need to register the structural information associated with each sample, accurately and precisely, and to be able to retrieve related structures with their related property data. Accelrys is pleased to announce a significant upgrade in our stereochemistry representation to address customer requirements. This white paper provides a detailed explanation of the new capabilities and how they can be used to best advantage.

EXAMPLE SYNTHESES

Both the examples that follow require a representation for the relative stereoconfiguration of stereogenic centers. They also require the ability to assign the stereoconfiguration of individual stereogenic centers.

Example: non-stereoselective reduction

As an example, consider the reaction shown in Figure 1. The reactant contains four stereogenic centers of known absolute configuration. The keto group is reduced to the alcohol using a non-stereoselective procedure. This introduces a new stereogenic center, which can adopt either the R or S configuration. The sample, therefore, is a mixture of the two diastereomers (epimers). This transformation does not affect the stereochemical configuration of the stereogenic centers on the cyclopentane ring.

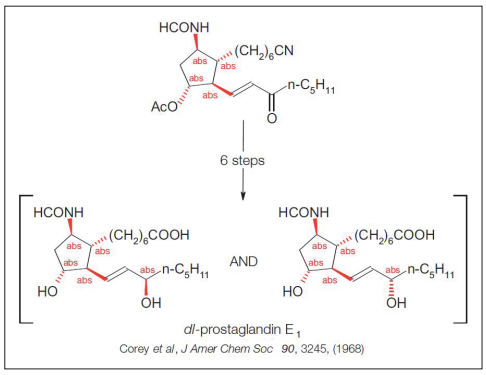


Figure 1. *Non-stereoselective reduction produces a sample that is a mixture of the two possible diastereomers.*

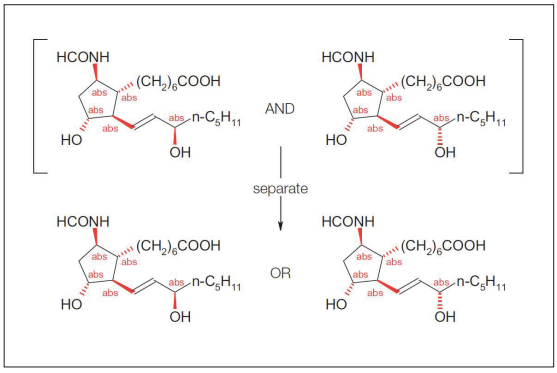


Figure 2. *Separation of a mixture of the two possible diastereomers produces two samples with a "relative" relationship.*

The sample is screened and found to be of interest. The two diastereomers are isolated in pure form, so that we now have two samples, each of which is a single diastereomer. The two samples are related by their relative configuration at the new stereogenic center. One sample has the R configuration and the other has the S configuration, but we do not know which is which. We cannot assign the configuration of either sample from the information that is currently available. This is shown in Figure 2.

After the samples are fully characterized, we obtain two samples, each with known absolute configuration at all stereogenic centers. Figure 3 shows this.

Example: syn-hydroxylation of an alkene

In this example, a benzodiazepine is hydroxylated using osmium tetroxide. This reagent introduces the hydroxy groups cis to each other, but we do not know which face of the cyclohexene is preferred. Thus, the reaction produces a pure product in which the hydroxy groups are cis, but we do not know whether the product has the (S, R) or the (R, S) configuration. That is, we know the relative stereoconfiguration of the two new stereogenic centers, but we do not know their absolute configuration. Figure 4 shows this.

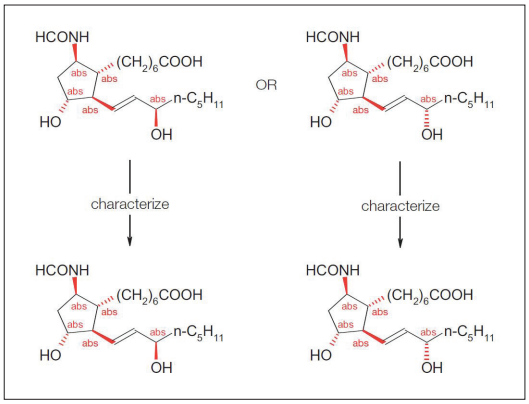


Figure 3. *Characterization produces two samples with absolute configurations.*

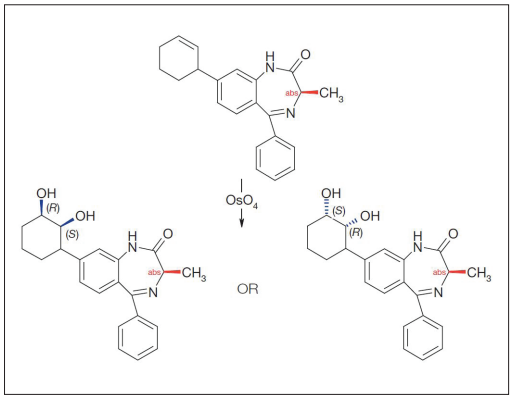


Figure 4. *Stereoselective syn-hydroxylation of an alkene.*

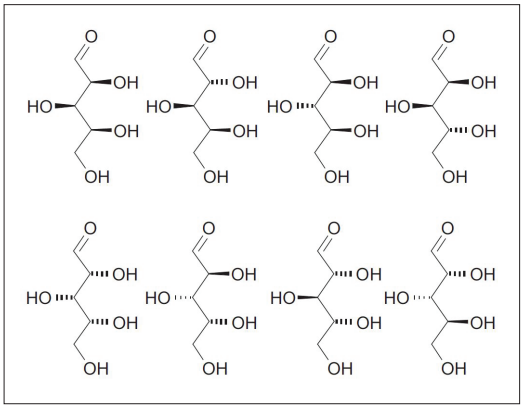


Figure 5. *The eight possible configurations for (-)-arabinose (a pentose).*

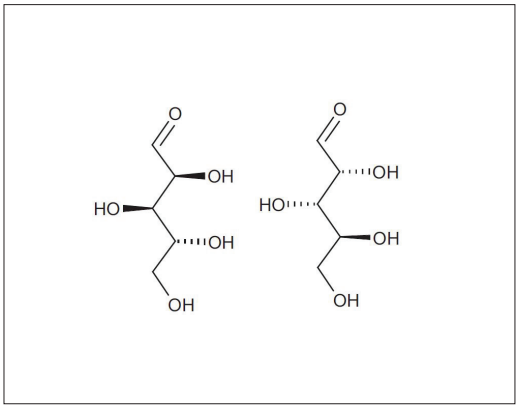


Figure 6. *(-)-arabinose determined to be one of two enantiomers.*

EMIL FISCHER'S ELUCIDATION OF THE STRUCTURE OF GLUCOSE

In 1891, Emil Fischer set out to elucidate 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 it must be one of the eight stereoisomers that are shown in Figure 5.

Fischer started by assuming the absolute stereoconfiguration of the simple molecule, glyceraldehyde. He was able to show that one of the carbon atoms in (-)-arabinose had the same configuration. Eventually, Fischer was able to show that (-)-arabinose was one of the two enantiomers that are shown in Figure 6. That is, Fischer was able to find the relative stereoconfigurations of all three stereogenic centers, but not their absolute configurations.

In 1951, it was determined that (-)-arabinose has the absolute configuration that is shown in Figure 7.

In this structure elucidation, Figures 5, 6, and 7 represent structural assignments for three physical samples. Only the last sample can be represented uniquely by a structure that uses existing stereochemical conventions, because it has a fully characterized structure.

ACCELRYS' ORIGINAL STEREOCHEMISTRY REPRESENTATION CONVENTIONS

The existing representation conventions use the following marks on the structure to represent stereochemical configuration:

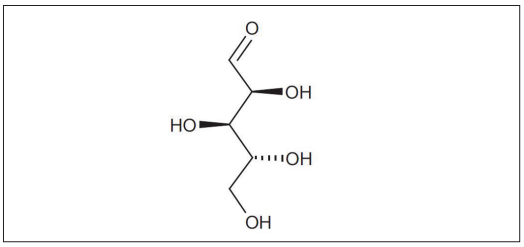


Figure 7. *Absolute configuration of (-)-arabinose determined by X-ray crystallography.*

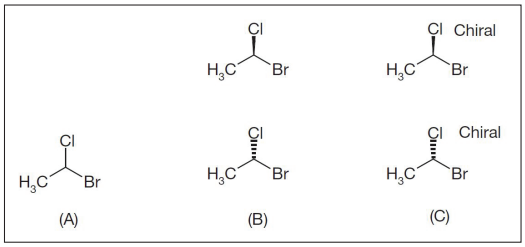


Figure 8. *Current representation.*

- Non-stereo bond to atom at stereogenic centers: Implies that no information is known about the configuration of a stereogenic center. It could be either of two stereoisomers, or a mixture of the two. See Figure 8A.
- Up (solid) wedge bond or down (hatched) wedge bond to atom at chiral center, with the narrow end of the bond pointing toward the stereogenic center: Implies that something is known about the configuration at that center. The structure implies that the sample is a mixture of the stereoisomer as drawn and its enantiomer. See Figure 8B.
- Wedge bonds to atoms at stereogenic centers with the chiral flag on the entire structure: When present, the chiral flag implies that a single isomer is present, and that the absolute configuration (R or S) is known for all stereogenic centers that are marked with wedge bonds. See Figure 8C.

These representations do not cover the situations discussed in the previous examples, where we know that the sample is pure, but we do not know the absolute configuration of all stereogenic centers.

COMPATIBILITY WITH A CHEMICAL DATABASE SYSTEM

A language is only useful if others can understand it. The chemical structure diagram is the universal language of chemistry, and well-understood conventions exist for most representation needs. Stereochemical representation, however, is one area where conventions are incomplete, particularly when it comes to the representation of structures with incomplete stereochemical information. A number of conventions are used. One of the most common is to differentiate the structures based on an assigned name. However, there are no established standards of nomenclature for all the possibilities that are discussed in these examples.

Accelrys' chemical representation needs to be understandable by two constituencies. One is the chemist who enters queries and retrieves the expected hits. The other constituency is the database search engine, which has to be able to translate queries unambiguously and retrieve the expected hits. Accuracy is vital, but in many cases the chemist wants only to retrieve a subset of all the possible hits. In this case, the chemist needs to use query features to narrow the search criteria and thereby reduce the number of hits.

The need for a system of chemical representation to encode structural information accurately to provide the desired precision when used as a query, and to be understandable by both chemists and a computer system, is a tremendous challenge. MDL's enhanced stereochemical representation meets this challenge.

ACCELRYS' ENHANCED STEREOCHEMICAL REPRESENTATION

The enhanced representation retains the currently supported representations:

- No stereochemical information at the stereogenic centers (no stereo bonds)
- A mixture of the stereoisomer as drawn and its enantiomer (stereo bonds marked and without the chiral flag attached to the structure)

• A pure sample in which the absolute configuration of the stereogenic centers is known (stereo bonds marked and with the chiral flag attached to the structure)

The enhanced representation is, therefore, fully backward-compatible with existing representations. Very importantly, the enhanced stereochemistry retains the current search and retrieval characteristics of the current representation.

The current representation has two deficiencies: it cannot represent a pure sample with known relative configuration of the stereogenic centers, and it cannot represent structures where some centers are known to have an absolute configuration but other stereogenic centers are known to have a relative configuration. The enhanced stereochemical representation addresses these issues.

GROUPS OF STEREOGENIC CENTERS

Accelrys' enhanced stereochemical representation introduces three types of identifier that can be attached to a stereogenic center. Each stereogenic center that is marked with wedge bonds belongs to a stereochemical group. Grouping allows us to specify stereogenic centers with a relative relationship. We can also use grouping to specify more than one relative grouping of stereogenic centers.

The only restriction on the number of stereogenic center collections is that a stereogenic center can be a member of solely one stereochemical group.

The three types of stereochemical groups are called ABSOLUTE, OR, and AND (see Figure 9 and Table 1).

The ABSOLUTE group replaces the chiral flag, which is no longer used in the enhanced system of representation. Instead of applying the chiral flag to the structure, you specify that all stereogenic centers that are marked with wedge bonds are members of the ABSOLUTE group. Stereogenic centers that are not part of the ABSOLUTE group must belong to an OR or an AND group.

The OR group is equivalent to a relative representation and the AND group is equivalent to a racemic representation. Either of these representations can be applied to a single stereogenic center.

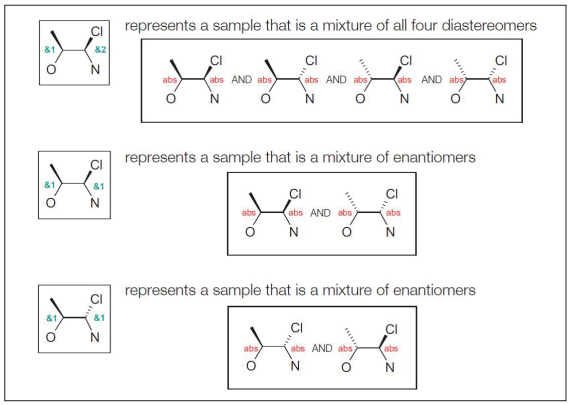


Figure 9. *Stereochemistry types in MDL's enhanced representation.*

Stereochemical Identifier	Represents
ABSOLUTE	Stereogenic center where the absolute configuration is known
OR	Stereogenic center where the relative configuration is known, but the absolute configuration is not known
AND	Mixture of stereoisomers; can be a pair of enantiomers or all the diastereomers

Table 1. *Characteristics of the different types of stereochemical groups.*

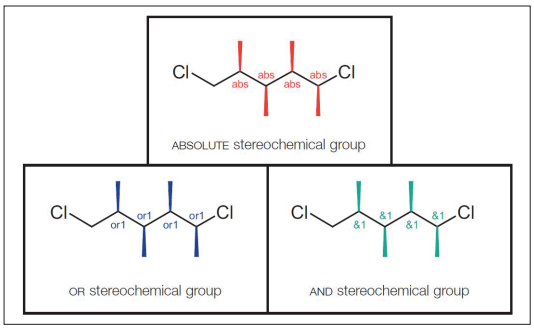


Figure 10. *Interpretation of AND stereochemical groups.*

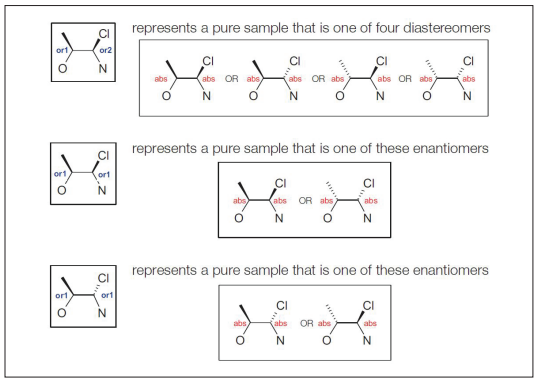


Figure 11. *Interpretation of OR stereochemical groups.*

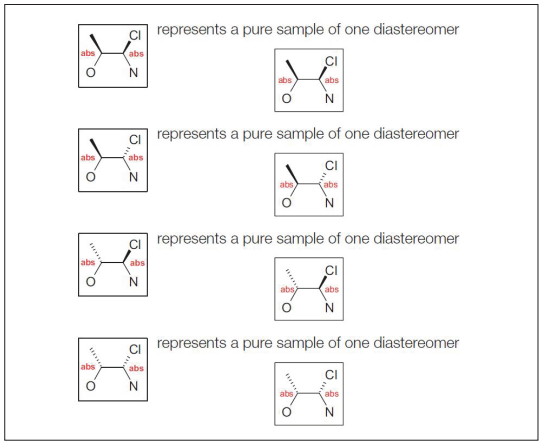


Figure 12. *Interpretation of ABSOLUTE stereochemical group.*

In the case of the OR group, the meaning is a structure that represents one stereoisomer that is either the structure as drawn (R,S) OR the epimer in which the stereogenic centers have the opposite configuration (S,R). The relationship can be between samples rather than within the structure. For example, Figure 4 shows two structures that are related by their relative configuration about a pair of stereogenic centers. This representation is an extension to the normally understood meaning of relative stereochemical configuration. To avoid confusion, the enhanced stereochemical representation uses the term OR, rather than relative. In the case of the AND group, the meaning is a mixture of two enantiomers: The structure as drawn AND its enantiomer. The term racemic, however, is normally applied to a sample that contains a 1:1 mixture of enantiomers. The term is so specific that it is better to use it to define an actual sample, rather than to name a structural representation. Consequently, the enhanced stereochemical representation uses the term AND, rather than racemic.

Examples of the three types of stereochemical groups are given in Figures 10, 11, and 12. In the figures, the label abs, or1, or2, &1, or &2 at each stereogenic center specifies the stereochemical group to which the center belongs. For more about these labels, see "Accelrys' nomenclature for the enhanced representation" on page 7.

If stereogenic centers are not marked with wedge bonds, then these unmarked centers are interpreted as having unknown stereochemistry.

A sample can contain all three types of stereochemical groups. The sample can contain solely one group of absolute stereocenters, but can contain multiple OR and AND groups. In addition, the structure can contain stereogenic centers that are not marked with wedge bonds. The situation can, therefore,

Type of Stereochemical Group	Value in Molfile	Display Label
ABSOLUTE	STEABS	abs
OR	STERELn	orn
AND	STERACn	&n

Table 2. *Stereochemical groups in the V3000 molfile.*

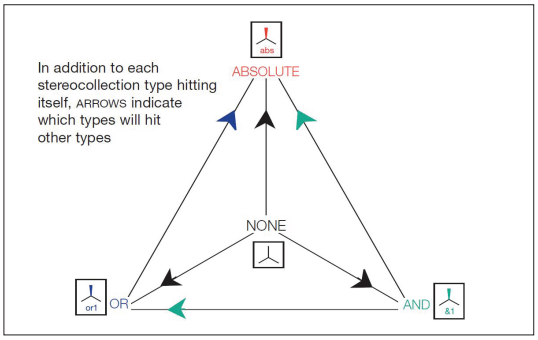


Figure 13*. Searching hierarchy for enhanced representation.*

become very complex. Representation of all this information is a major challenge. For more information, see "Use the enhanced representation sparingly" on page 9.

No standards have been adopted to name or represent structures with all this information. IUPAC suggests that R* and S*, or u (unlike) and l (like), be used to describe relative stereocenters (IUPAC Compendium of Chemical Terminology, 2nd edition (1997)). However, IUPAC does not appear to have considered cases where the sample contains more than one group of relative stereocenters. Once a standard is established, Accelrys will adopt it.

ACCELRYS' NOMENCLATURE FOR THE ENHANCED REPRESENTATION

The stereogenic centers and their associated stereochemical groups are identified within a V3000 molfile, the format of which has been extended to accommodate this information. The groups are identified using persistent collections within the molfile. For more detail, see Appendix A: Example V3000 Molfile.

A stereogenic center that is marked with a wedge bond can have one of three values in the molfile. The internal names of these collections are: STEABS, STERELn, and STERACn, where n is an index number that identifies a particular set (collection) of stereogenic centers of a given types. Table 2 shows the relationship of these internal names to the three types of stereogenic centers and their default display labels. The types of stereochemical groups and the values in the molfile are fixed, but the administrator can change the display labels.

The stereogenic center is also identified by one wedge bond of either the solid or hashed style, with the narrow end of the bond pointing toward the stereogenic center. You can also use color within the rendering technology to identify stereogenic centers of the same type. For example, you can specify red for the abs label, blue for the orn label, and green for the &n label. You can also apply the color of the label to the wedge stereo bonds. Both Accelrys Draw and Accelrys ISIS/Draw provide the option to use color to identify and classify stereogenic centers. The label names, their colors, and the color of the stereo bonds are all under the control of the system administrator.

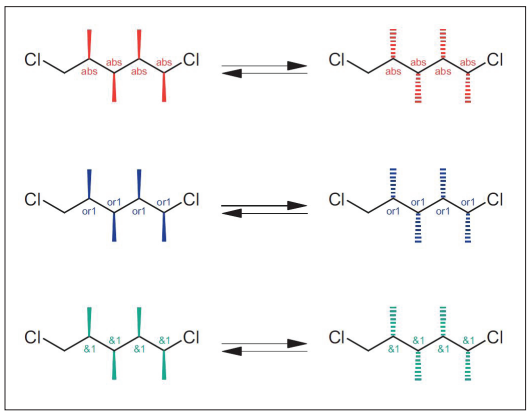


Figure 14. *SSS matching (1).*

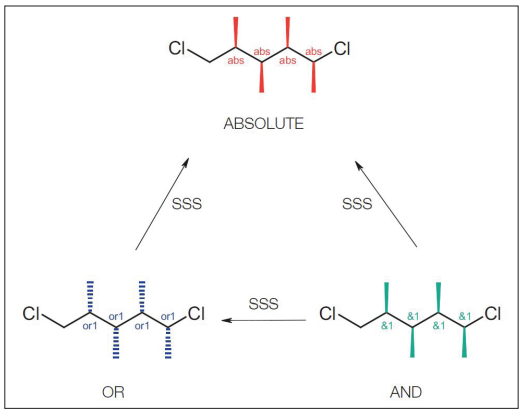


Figure 15. *SSS matching (2).*

Accelrys decided not to invent new bond styles to represent this information; although some new bond styles have been proposed (Maehr, J Chem Inf Comput Sci. 2002;42:894–902), they have not been formally adopted by IUPAC. When a standard is adopted it will be straightforward to change how the information is rendered. Accelrys is committed to conforming to a future standard.

SEARCH CHARACTERISTICS OF THE ENHANCED REPRESENTATION

Substructure searches follow the hierarchy shown in Figure 13. For SSS matching, ABSOLUTE stereochemical groups match only ABSOLUTE target groups; OR stereochemical groups match OR and ABSOLUTE target groups; and AND groups match any type of group (see Figures 14 and 15).

For atoms in ABSOLUTE stereo collections to match, all stereogenic centers must have the same stereoconfiguration. For atoms in OR or AND stereo collections to match, the stereogenic centers must match the structure as drawn or its enantiomer.

Stereogenic groups in a query can also match stereogenic groups that represent a larger set of stereoisomers, provided that the target structure entirely contains the set of stereoisomers that the query represents. For a detailed explanation of this behavior, see Appendix A of the ISIS/Host Relational Chemical Database Administration Guide, Version 5.0. Figure 16 shows all possible substructure search (SSS) matches (see also Appendix B: SSS Matching).

ENUMERATING GENERIC STRUCTURES

The enumeration of generic structures that contain stereogenic centers within the Rgroups and the scaffold can present difficulties, for example when one Rgroup contains one or more stereogenic centers with absolute configuration and other Rgroups and the scaffold contain stereogenic centers with relative configuration.

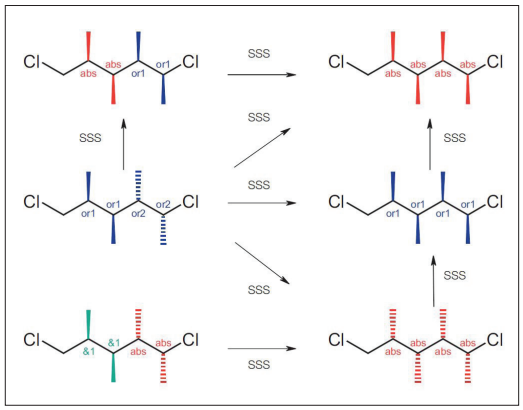


Figure 16. *SSS matching (3).*

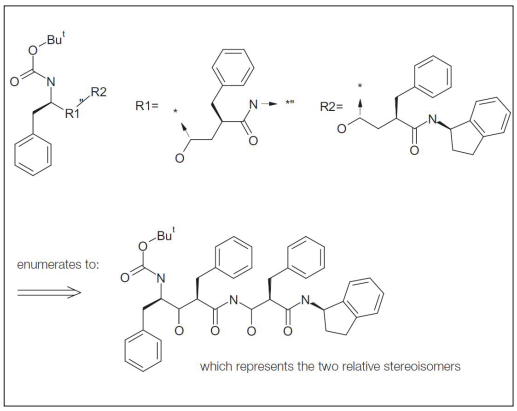


Figure 18. *Removal of the chiral flag leads to enumeration to incorrect representation.*

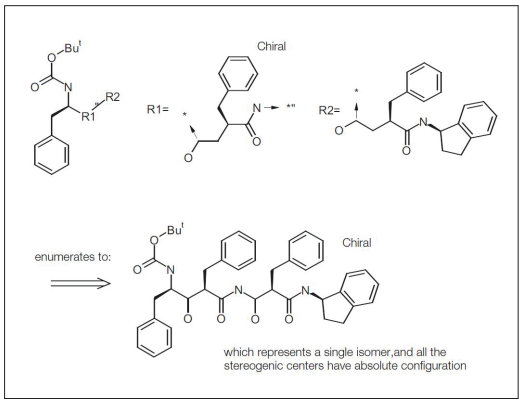


Figure 17. *Presence of chiral flag on one Rgroup leads to enumeration to incorrect representation.*

Consider the partial generic structure shown in Figure 17. It is represented using the original representation. The scaffold and R2 contain stereogenic centers but no chiral flags are attached. They represent, therefore, a mixture of the two relative isomers; R1 has a chiral flag attached and has, therefore, absolute configuration. Reacting these three fragments produces a sample that contains four stereoisomers (see Figure 20). Enumeration of this generic will produce a structure with an attached chiral flag, which applies to the whole structures. The product of the enumeration represents a single stereoisomer. This is not correct, but the structure is contained within the product of the chemical reaction.

Removal of the chiral flag prior to enumeration leads to a representation that represents two relative stereoisomers. See Figure 18. This too is incorrect, and the relative isomer that is implicit in this representation is not present in the product.

The use of separate and stereogenic center groups in the generic structures corrects this defect. Enumeration of the generic shown in Figure 19 produces a representation that correctly represents all of the four isomers that would be produced in the chemical reaction.

USE THE ENHANCED REPRESENTATION SPARINGLY

The enhanced representation is very rich and allows much detail to be assigned to stereogenic centers. For example, a structure with one stereogenic center can be represented in four ways: no information (no stereo bonds), AND, OR, or ABSOLUTE. This is manageable and understandable. A structure with two centers has 65 registerable representations, and this reduces to 29 non-degenerate, registerable representations. Interpreting these structures and their relationships is a challenge and limits need to be

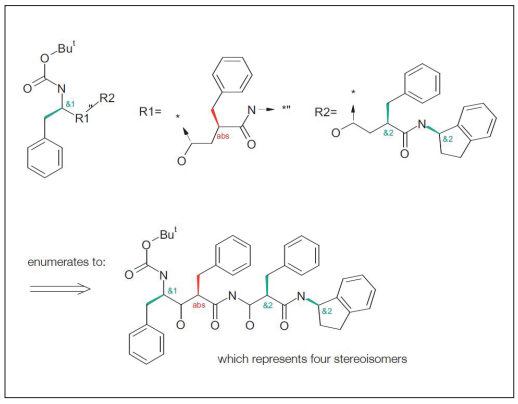


Figure 19. *MDL's enhanced stereochemical representation produces the correct result.*

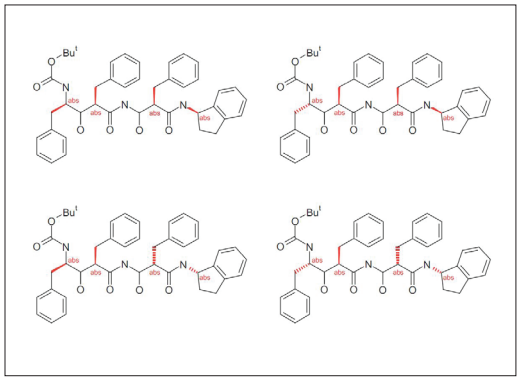


Figure 20. *The four structures implied in the generic structure shown in Figures 17, 18, and 19.*

applied to the amount of detail that is acceptable in a valid structure.

The limits need to be developed as a corporate business rule, but Accelrys suggests that a preferred representation should contain no more than four different stereochemical groups.

For examples see Appendix C: Registerable Representations.

SUMMARY

- Accelrys' enhanced stereochemical representation satisfies a need of scientists and curators of chemical structure databases.
- The enhanced representation introduces a relative stereochemical representation that is structuredefining and searchable.
- A structure can contain more than one type of stereochemical group.
- The stereochemical groups define a structure's uniqueness.
- In the absence of a standard, Accelrys proposes the names ABSOLUTE, OR, and AND for the types of stereochemical groups.
- Accelrys will conform to a standard when it is adopted by IUPAC.
- The enhanced stereochemical representation is supported by all of Accelrys' chemistry products.
- The enhanced stereochemical representations allows considerable detail in describing information about the stereogenic centers within a structure.
- The enhanced stereochemical representation needs to be incorporated in a company's chemical structure business rules.
- It is recommended that the enhanced stereochemical representation be used sparingly.

APPENDIX A: EXAMPLE V3000 MOLFILE

```
MDL-Draw12200218542D
0 0 0 0 0 999 V3000
M V30 BEGIN CTAB
M V30 COUNTS 29 32 0 0 0
M V30 BEGIN ATOM
M V30 1 C -0.0544 -0.428 0 0
M V30 2 C -0.0544 -1.278 0 0 CFG=2
M V30 3 C -0.7962 -1.6974 0 0
M V30 4 C -0.7962 -2.544 0 0
M V30 5 C -0.0703 -2.9817 0 0
M V30 6 C 0.6716 -2.5622 0 0 CFG=1
M V30 7 C 0.6716 -1.7051 0 0 CFG=2
M V30 8 C 1.4208 -1.2852 0 0
M V30 9 C 1.4093 -2.9942 0 0
M V30 10 C -1.2907 1.3481 0 0
M V30 11 C -2.0326 0.9232 0 0
M V30 12 C -2.0326 0.0734 0 0
M V30 13 C -1.293 -0.3579 0 0
M V30 14 C -0.5534 0.9274 0 0
M V30 15 C -0.5534 0.074 0 0
M V30 16 N 0.77 -0.3022 0 0
M V30 17 C 1.1707 0.466 0 0 CFG=1
M V30 18 C 0.8061 1.2704 0 0
M V30 19 N -0.0142 1.4289 0 0
M V30 20 O 1.3356 1.938 0 0
M V30 21 C -2.7699 1.3485 0 0 CFG=2
M V30 22 C -2.7699 2.2058 0 0 CFG=2
M V30 23 C -3.4976 2.6309 0 0 CFG=2
M V30 24 C -4.2412 2.2104 0 0
M V30 25 C -4.2412 1.355 0 0
M V30 26 O -3.4976 3.4805 0 0
M V30 27 O -2.0178 2.624 0 0
M V30 28 C 2.0249 0.466 0 0
M V30 29 C -3.4995 0.9299 0 0
M V30 END ATOM
M V30 BEGIN BOND
M V30 1 1 2 1 CFG=1
M V30 2 1 7 8 CFG=1
M V30 3 1 6 9 CFG=1
M V30 4 1 2 3
M V30 5 1 2 7
M V30 6 1 3 4
M V30 7 1 4 5
M V30 8 1 5 6
M V30 9 1 6 7
M V30 10 2 1 16
M V30 11 1 16 17
M V30 12 1 17 18
M V30 13 1 19 14
M V30 14 1 18 19
M V30 15 2 18 20
M V30 16 1 21 11 CFG=1
M V30 17 1 21 22
M V30 18 1 22 23
M V30 19 1 23 24
M V30 20 1 24 25
M V30 21 1 23 26 CFG=3
M V30 22 1 22 27 CFG=3
M V30 23 1 17 28 CFG=1
M V30 24 1 10 11
M V30 25 2 10 14
M V30 26 2 11 12
M V30 27 1 12 13
M V30 28 2 13 15
M V30 29 1 14 15
M V30 30 1 15 1
M V30 31 1 25 29
M V30 32 1 29 21
M V30 END BOND
M V30 BEGIN COLLECTION
M V30 MDLV30/STEABS ATOMS=(1 17)
M V30 MDLV30/STEREL2 ATOMS=(2 6 7)
M V30 MDLV30/STEREL1 ATOMS=(2 22 23)
M V30 MDLV30/STERAC2 ATOMS=(1 2)
M V30 MDLV30/STERAC1 ATOMS=(1 21)
M V30 END COLLECTION
M V30 END CTAB
M END
```

Figure 21. *Example V3000 molfile.*

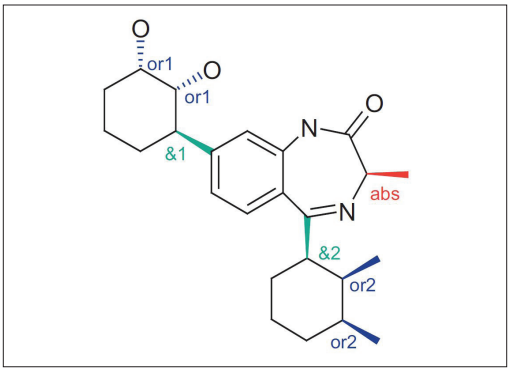


Figure 22. *Structure from Figure 21 rendered by MDL® ISIS/Draw 2.5.*

APPENDIX B: SSS MATCHING

To understand the SSS matches shown in Figure 16, you can enumerate the individual stereoisomers that are implied by the representation. The diagram below shows the stereoisomers that are contained within the representations in Figure 16. This information shows the overlap between the representations and, therefore, the valid SSS matches.

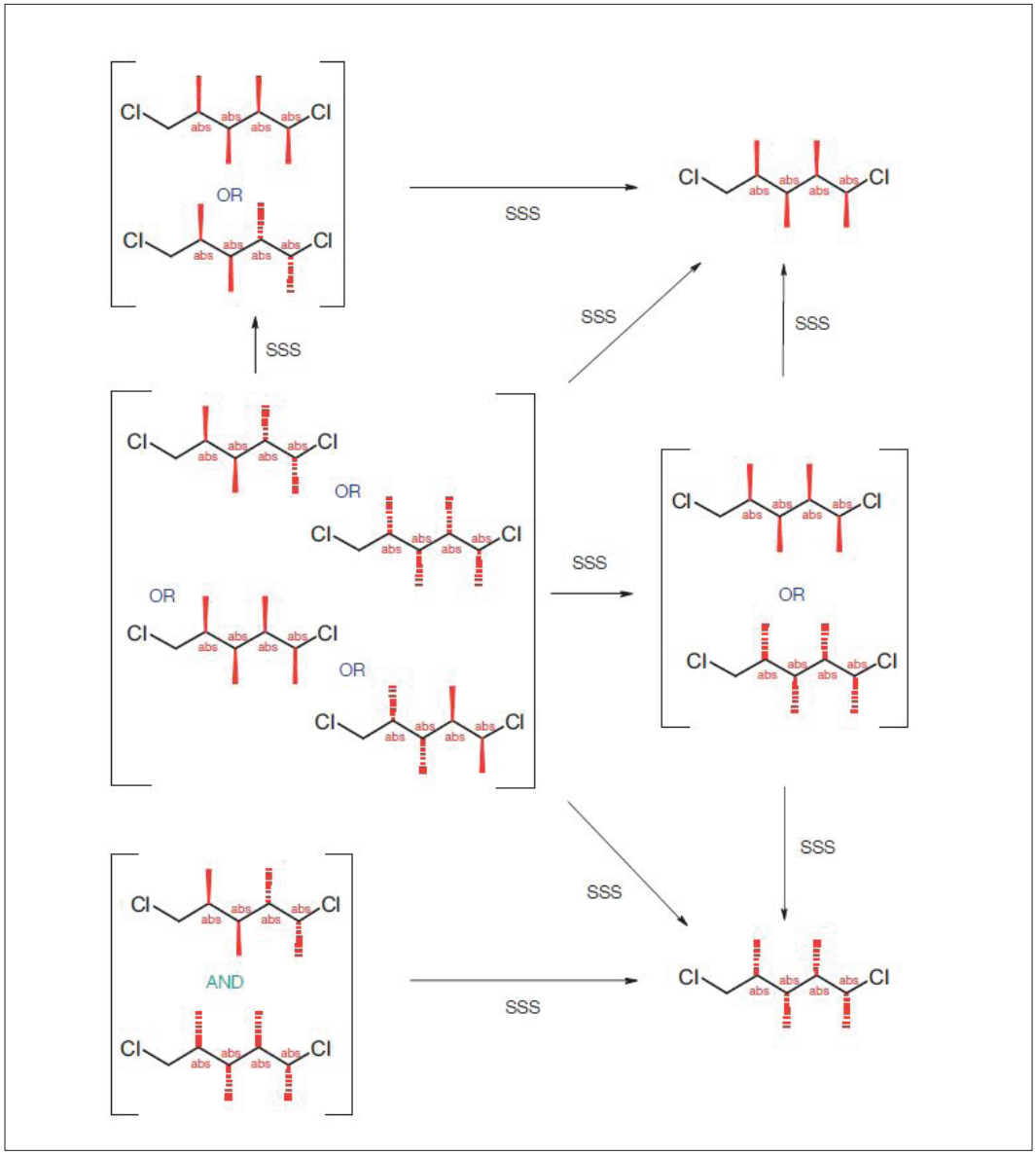
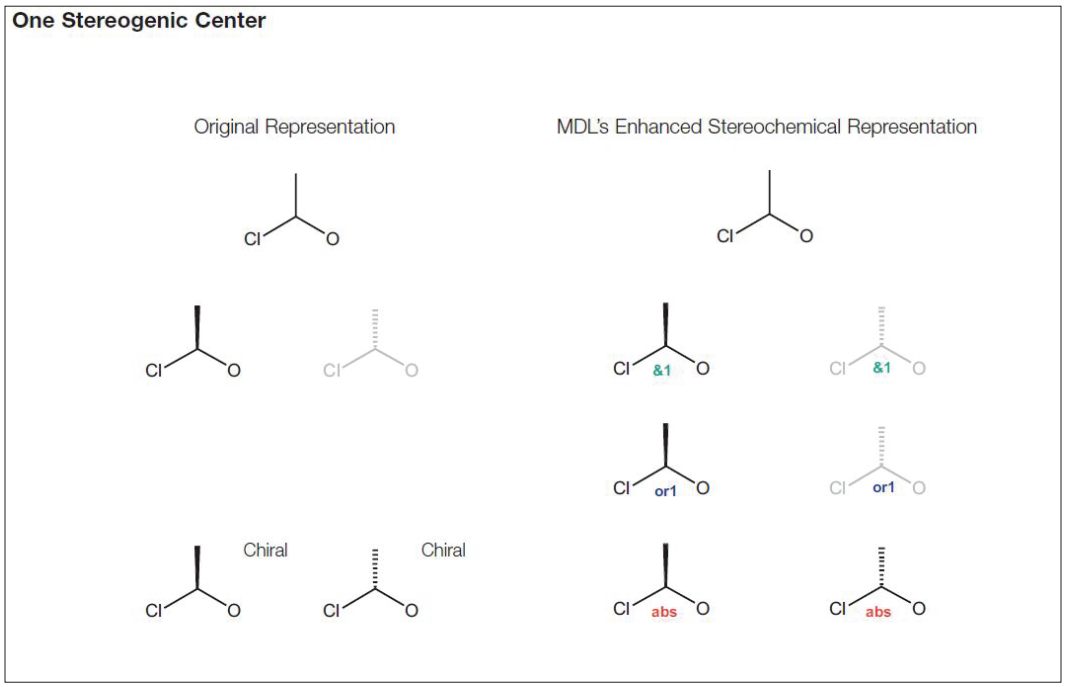


Figure 23. *SSS matching (4): enumerating the representations clarifies matches.*

APPENDIX C: REGISTERABLE REPRESENTATIONS

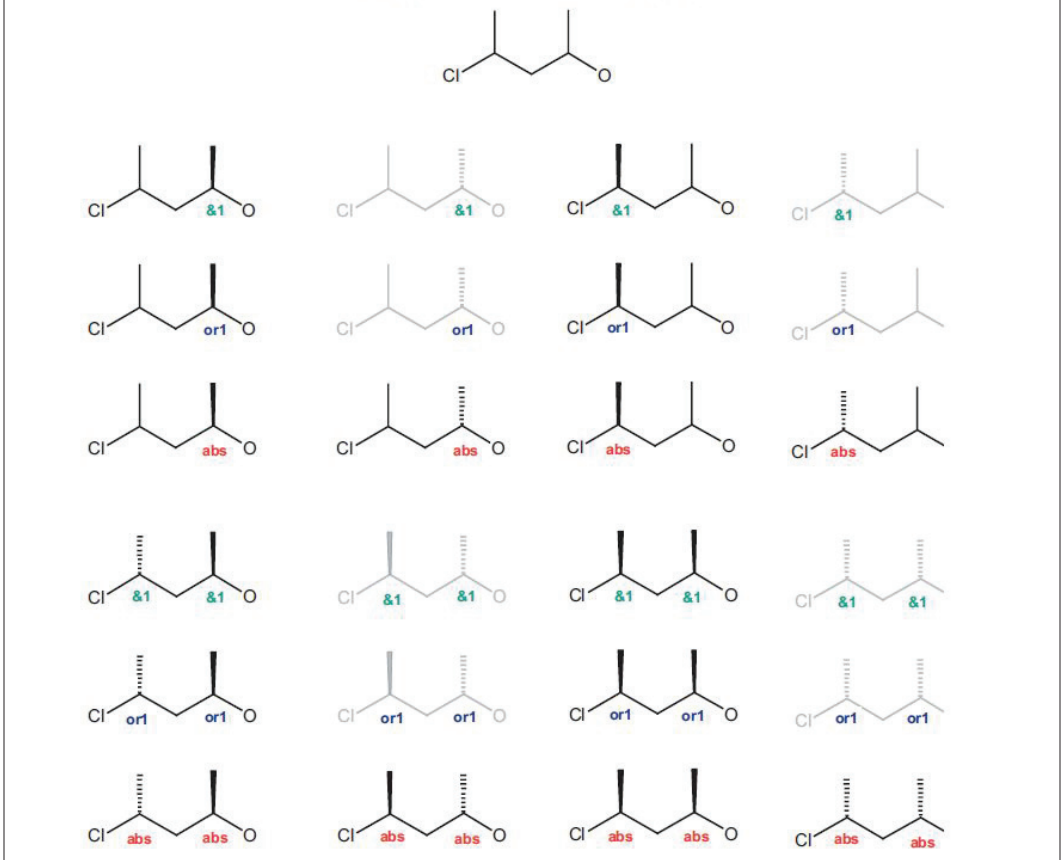
In the representations that follow all the representations shown in gray are degenerate with one of the structures shown in black. All the representations are valid and are registerable but degenerate representations are treated as duplicates.



APPENDIX C (CONTINUED)

Two Stereogenic Centers

Enhanced Stereochemical Representation



APPENDIX C (CONTINUED)

Two Stereogenic Centers (continued)

Enhanced Stereochemical Representation (continued)