

Extensions to SYBYL[®] Line Notation for Representing Relative Stereochemistry and Isomeric Mixtures

By *Michael Mandl, Roman Portnoi, and Jon Swanson*
TRIPOS, Inc.
Olaf Hofmann and Rolf Jautelat
Schering A.G.
Brad Buckman
Berlex Biosciences

Table of Contents

EXECUTIVE SUMMARY	3
CHIRALITY IN DRUG DISCOVERY	3
RELATIVE STEREOCHEMISTRY AND MIXTURES.....	4
SKETCHING AND STEREOCHEMISTRY	9
RELATIVE STEREOCHEMISTRY AND 2D SEARCHING	10
OTHER VENDOR STEREOCHEMISTRY SUPPORT	13
SUMMARY	16

EXECUTIVE SUMMARY

This document describes an extension to the SYBYL[®] Line Notation (SLN) language that enables database storage and searching of chemical structures for which the exact stereochemistry is unresolved, and for those that exist as stereoisomeric mixtures. These extensions enable SLN representation of isomeric mixtures generated by reactions that produce chiral centers, and allow these mixtures to be registered and searched.

CHIRALITY IN DRUG DISCOVERY

The advent of sophisticated asymmetric synthesis methods and chiral separation technologies has meant that scientists can actively include stereochemistry as a design consideration in the development of biologically active compounds. Initially, these capabilities were exploited in chiral switching, which is the redevelopment as single enantiomers of drugs already approved as racemic or stereoisomeric mixtures. It quickly became evident that working with stereochemically pure compounds had pronounced benefits: no undetected false negatives from screening mixtures, more detailed information about ligand-receptor interactions, and the power to fine-tune pharmacological profile. As of 2002, chiral drugs accounted for more than \$150 billion in sales, more than 37% of total pharmaceutical sales.

The increased appreciation for the virtues of addressing stereochemistry in the early stages of drug discovery creates a concomitant need to store structures in a way that reflects the realities of chemical synthesis. In some cases there may not be a stereospecific route to a particular intermediate or end product. In other cases, reactions may be stereoselective rather than stereospecific. In either case, a mixture of products is generated, each having different stereochemistry. Pending a suitable separation to resolve the products and determine absolute configuration, the structures and associated information need to be stored in the corporate database.

SYBYL line notation (SLN)¹ is one of several widely used languages for storing molecular structures in databases and for creating search queries. These languages represent chemical structures as text strings, and specify atoms, bonds, and their attributes. One component of the grammar for systems describing chemical structures is a way to mark the stereochemistry at atoms and bonds. The stereochemistry is either marked in an absolute fashion (typically using Cahn-Ingold-Prelog (CIP)

¹ Ash, S.; Cline, M.A.; Homer, W.; Hurst, T.; Smith, G.B. "SYBYL Line Notation (SLN): A Versatile Language For Chemical Structure Representation" *J. Chem. Inf. Comput. Sci.* **1997**, 37, 71-79.

rules^{2,3}) or in a relative fashion, where atom weighting is based on atom ordering. Tripos' SYBYL Line Notation language and Conversational SMILES⁴ both allow marking atom chirality based on CIP rules. In the case of SLN, CIP nomenclature is translated to N/I nomenclature, where the atom weighting is based on the position of the atom in the SLN string. There are several other examples of this positional weighting scheme. Daylight Isomeric SMILES extends SMILES with @ and @@ atom flags, which define atom weighting based on atom order. MOL files⁵ have an atom parity flag that defines atom chirality as either even or odd. Except for provision of an unknown stereo atom attribute, there has been little provision for specifying structures where the absolute stereochemistry at each chiral center is not fully resolved.

RELATIVE STEREOCHEMISTRY AND MIXTURES

In order to provide a better syntax for registering stereoisomers into a database, it is necessary to do a better job of specifying unresolved stereoisomers than using a generic "unknown" atom stereo attribute. One should be able to identify whether the structure is a single pure stereoisomer for which absolute stereochemistry may or may not be known, a mixture of stereoisomers, or perhaps that nothing is known about the stereochemistry other than that the structure contains chiral centers. To provide this additional functionality requires an extension to the current atom stereo attribute in SLN.

The information to be encoded in this extension can be understood by looking at the scheme shown in Fig. 1. A reaction run under achiral conditions, which produces two chiral centers, will produce a mixture of the four stereoisomers (I). Typically, the first step in purifying this mixture is to resolve the mixture into two pairs of enantiomers (II). These enantiomers can be resolved into individual pure enantiomers (III), for example, by chiral HPLC. At this point, it may not be known which

² Cahn, V.R.S.; Ingold, C.; Prelog, V. "Spezifikation der Molekularen Chiralitat" *Angew. Chem.* **1966**, 78, 413-424.

³ Prelog, V.; Helmchen, G. "Basic Principles of the CIP-System and Proposals for a Revision" *Angew. Chemie. Int. Ed. Eng.* **1982**, 21, 567-583.

⁴ The original SMILES syntax did not include attributes from atom or bond chirality. (a) Weininger, D. "SMILES: A Chemical Language and Information System" *J. Chem. Inf. Comput. Sci.* **1988**, 28, 31-36. (b) Weininger, D.; Weininger, A.; Weininger, J. L. "SMILES II. Algorithm for Generation of Unique SMILES Notation" *J. Chem. Inf. Comput. Sci.* **1989**, 29, 97. Two separate extensions were proposed: Conversational SMILES by Robert Pearlman and coworkers at the University of Texas, Austin and Isomeric SMILES by Daylight Information Systems.

⁵ Dalby, A.; Nourse, J.G.; Hounshell, W.D.; Gushurst, A.K.I.; Grier, D.L.; Leland, B.A.; Laufer, J. "Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited" *J. Chem. Inf. Comput. Sci.* **1992**, 32, 244-255.

absolute stereochemistry corresponds to which isolated product. But before the final identification of the stereoisomers, it may be necessary to register these compounds and retain their relative relationship. A final stage, perhaps x-ray, will then determine the absolute stereochemistry of each isomer (IV).

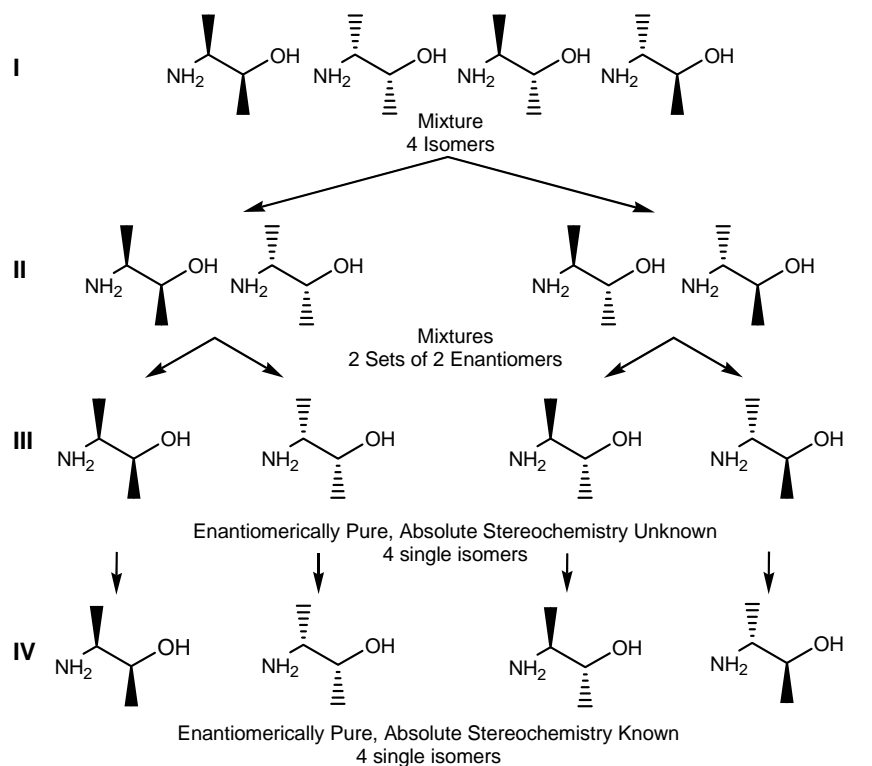


Figure 1. Steps in the resolution of stereoisomers.

Whether and to what extent the stereoisomers are resolved depends on the relevance of the compound. For example, if the mixture (I) shows no activity, the mixture will probably simply be registered as the mixture of 4 isomers and no further work will be done. If, on the other hand, the mixture is found to be active, the isomers will be partially or fully resolved. It is important to be able to register structures at each stage of the purification process.

Previous versions of SLN allowed marking chiral centers as N (normal or clockwise presentation) or I (inverted or counterclockwise presentation) if the absolute stereochemistry of the center was known (or alternatively, R or S). Otherwise, the only alternative was either to not

include a chiral attribute on the atom, or mark the atom as S=U (unknown or unspecified).

In order to be able to describe these mixtures and pure, but unresolved enantiomers, a modifier, the stereo mode, is added to the current N, I and U (or R, S, U) stereochemical attributes. R/S/U notation will be used for the rest of this paper, as it is more familiar to most chemists than N/I/U. These additional modifiers are:

- E** Explicit: exactly one stereoisomer (R/S) or completely unknown chirality (U).
- *** Relative: exactly one stereoisomer, but absolute stereochemistry not known.
- M** Mixture: more than one stereoisomer.

The combinations of the current stereochemical attributes and these additional modifiers produce a total of nine possible stereochemical values. These modifiers allow the differentiation between compounds registered as mixtures (RM, SM, and UM) and compounds that are single pure stereoisomers (R*, S*, U*, RE, and SE). The pure single compounds can be further categorized into the pure but unresolved isomers (R*, S*, and U*) and the pure resolved isomers (RE and SE).

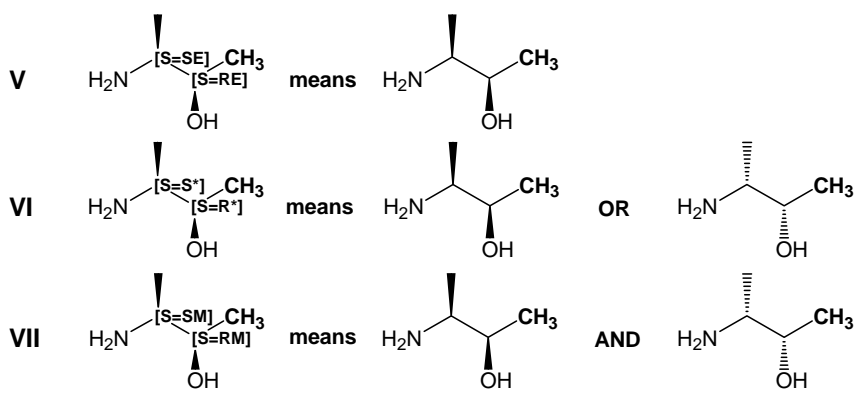


Figure 2. *Examples of Stereo Mode Attributes.*

The difference between E, *, and M is shown pictorially in Fig. 2. Structure V represents a single stereoisomer. The chiral centers will be marked with the E stereo mode modifier. To represent either this structure *or* its enantiomer, the * stereo mode modifier is used (VI). To represent both structure V *and* its enantiomer the M stereo mode modifier is used (VII). A structure with two chiral centers each marked R* represents either the stereoisomer with each chiral center marked R

or the stereoisomer with each chiral center marked S. A structure with two chiral centers marked RM represents both the stereoisomer with the chiral centers each marked R and the stereoisomer with each chiral center marked S. Note that in a structure with two chiral centers marked RM there is still a relationship between the individual chiral centers. The structure corresponds to just two isomers, not all possible combinations of each chiral center being set to R or S.

The UE attribute is a special case. The explicitly unknown case (UE attribute) is defined to represent the situation where the stereochemistry is totally unknown. The chiral center might either be part of a mixture or a pure isomer. In either case, the stereochemistry at the chiral center is unknown. This makes the UE attribute equivalent to the U attribute in earlier versions of SLN and provides backward compatibility. To further enhance backward compatibility, the E designator is optional. SLNs with atom stereo attributes written as S=RE, S=SE, or S=UE will be translated to the equivalent notation of S=R, S=S, or S=U.

An additional modifier is needed to handle groups of centers whose relative stereochemistry is known within a group but not between groups. For example, one might want to describe the product of a reaction between two pure but unresolved isomers. Numeric indices are allowed on the extended stereo attributes to designate multiple regions of relative stereochemistry within a single molecule. A molecule that has four atoms marked R*1, R*1, R*2, and S*2 has two sets of two chiral centers. The first set has two chiral centers such that if one is R, the other is R, or if one is S, the other is S. The second set has two chiral centers such that if one is R, the other is S, or vice versa. Uniqueing may change the value of the index, but not the groups to which the centers belong. If no index is supplied, the center is assumed to belong to group 0. This provides backward compatibility with earlier versions of SLN. The group modifier does not apply to the U chiral attribute. All U stereocenters belong to the same group. If an atom is marked, for example, S=U*5, it will be converted to S=U*. A group modifier cannot be applied to atoms with explicit stereochemistry, either. An atom marked S=RE5 will be converted to S=R (the E designator is not needed, but is not invalid).

An example is shown on the next page in Fig. 3. Two reactants, each a single, but unresolved isomer, combine to form a single isomer in the product. There are four possibilities for the absolute stereochemistry of this product.

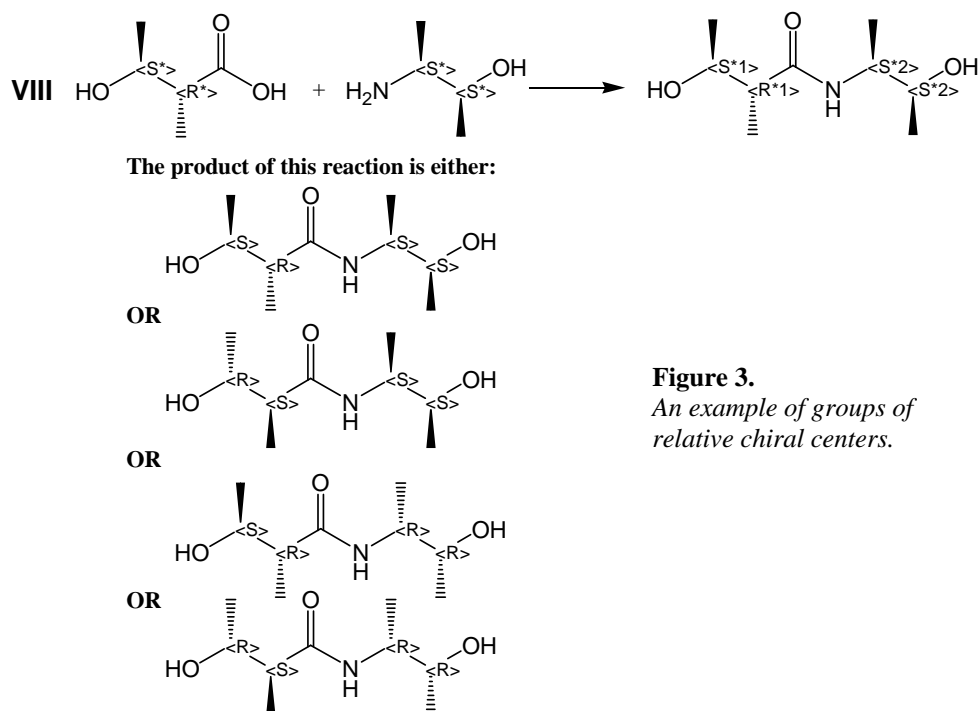


Figure 3.
An example of groups of relative chiral centers.

The relative stereo mode and group modifiers apply only within a molecule. They are not meant to have meaning across several molecules. This means that for a molecule with one chiral center, there is no functional difference between the attributes R*, S* and U*. Each represents a molecule that is one of the two possible pure isomers. One would normally describe the two pure isomers as U* with a label on the molecule to differentiate the two relative isomers. Since the R*, S*, RM and SM attributes are intended to be used within a single molecule to designate the relative orientation of chiral centers within a molecule, they should normally only be used when there are at least two such centers in the molecule. However, using the * or M modifier when there is only a single chiral center is not syntactically invalid.

If a molecule has two chiral centers, the first marked R* and the second marked S*, SLN treats this representation as exactly equivalent to a molecule with the same atoms and connections, but with the first marked S* and the second marked R*. Both represent molecules that contain two chiral centers such that either the first center is R and the other is S, or the first center is S and the other is R.

A few examples (Fig. 4, below) help illustrate how these attributes are used.

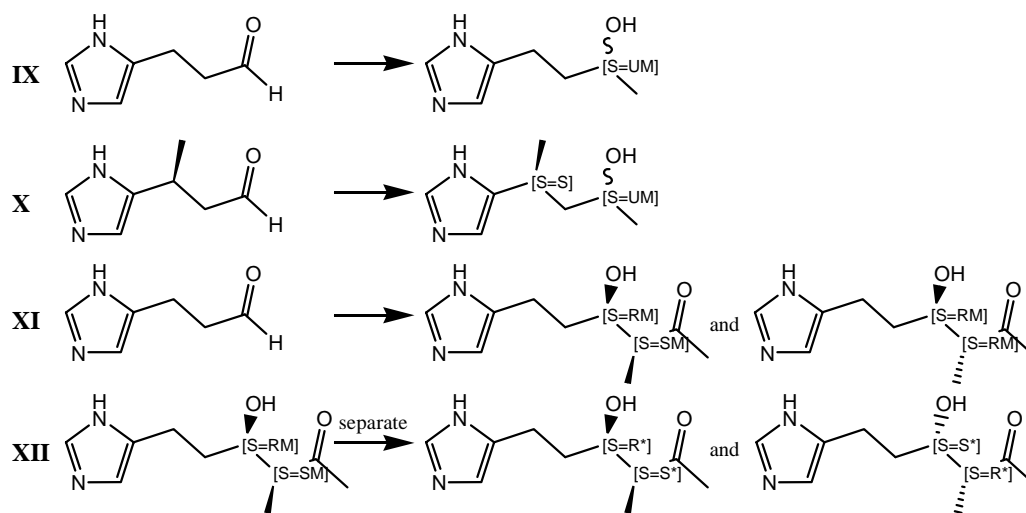


Figure 4. Examples of chiral molecules.

Structure IX shows the result of running a reaction that produces a chiral center, but the reaction produces a racemic mixture. Structure X shows the pair of diastereomers that would result from running this reaction with a chiral starting material. Structure XI represents the mixtures of enantiomers that will be produced from a reaction that produces two chiral centers. If one of the pairs of enantiomers were separated, but without elucidating the absolute stereochemistry, the individual isomers would be represented as in structure XII. Note that in this last example, there is no difference between what the two representations mean. Each represents a single stereoisomer that is R at one center and S at the other, or vice versa.

SKETCHING AND STEREOCHEMISTRY

The E, *, and M stereo mode modifiers allow representation of all stereo configurations without the need to consult wedge bond attributes, which, in SLN syntax, are attributes of the molecule as a whole. This provides a more compact representation, which is not dependent on retaining molecule attributes and simplifies the syntax of queries. However, in practice, it is generally more intuitive for chemists to sketch structures using wedge and dashed bonds. Wedge bond Ct (connection table) attributes of WEDGE_UP and WEDGE_DOWN are provided in SLN to retain the information sketched by the chemist.

It is possible when sketching to define the chirality by a mixture of wedge bonds and stereo mode qualifiers. This means it is important to define the priority of the atom and bond stereo designators, since both

might be present. For translation to SLN, the atom stereochemistry takes priority. Once converted to SLN, the fully qualified stereo attribute on the atom is used to determine the chirality of the atom; wedge bond information is not consulted. There are three major cases to consider when converting a sketched molecule to SLN:

1. Atom stereochemistry is not completely defined, i.e., the atom is labeled “*”, but wedge bond information is present to define the atom stereochemistry.

Convert the atom stereochemistry to a fully qualified attribute by determining R/S from the wedge bonds, i.e., mark the atom as “S=R*” and retain the wedge bond information.

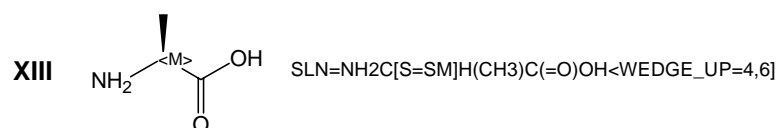
2. Atom stereochemistry is completely defined, i.e., the atom is labeled “R*,” and wedge bond information is present and consistent with the stereo label.

Mark the atom as “S=R*” and retain the wedge bond information.

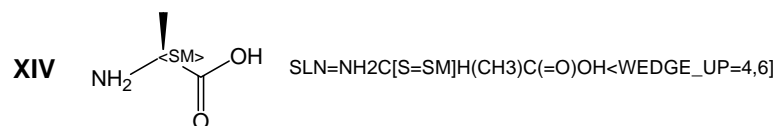
3. Atom stereochemistry is completely defined, i.e., the atom is labeled “R*,” and wedge bond information is present and inconsistent with the stereo label.

4. Mark the atom as “S=R*” and remove the wedge bond information.

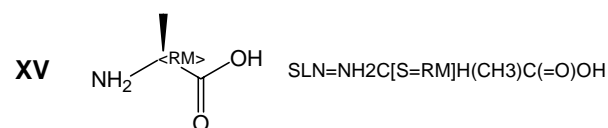
These cases are illustrated below in Fig. 5.



Case 1. Incompletely specified atom attribute and wedge bonds



Case 2. Completely specified atom attribute and consistent wedge bonds



Case 3. Completely specified atom attribute and inconsistent wedge bonds

Figure 5. Examples of how sketches will be interpreted as SLNs.

RELATIVE STEREOCHEMISTRY AND 2D SEARCHING

Having defined the attributes necessary to register compounds with unspecified or partially specified stereochemistry, it is necessary to define the behavior when using these attributes in 2D search queries. Two modes of operation have been defined for UNITY[®] 2D searching: explicit and hierarchical. The explicit mode will find exactly the chiral attribute specified in the query. For example, a query with an atom marked “S=U*” will only return structures with atoms marked as “U*” and not those marked with “R*” or “S*.” Any subset of registered compounds can be returned by using Boolean combinations of attributes in explicit mode. For example, to return any structures with relative stereochemistry, one can use “S=R*|S=S*|S=U*.” The hierarchical mode is supplied to make defining queries simpler. It follows the same general scheme as defined in Fig. 1. The first distinction made is between mixtures and pure compounds. Queries that include the “M” attribute return only mixtures. Queries that include the “*” attribute return single isomers of either known or relative stereochemistry. The second distinction is between known and unknown stereochemistry. Queries that use the E attribute (or no attribute) return only pure compounds of known stereochemistry. This scheme is illustrated in Fig. 6.

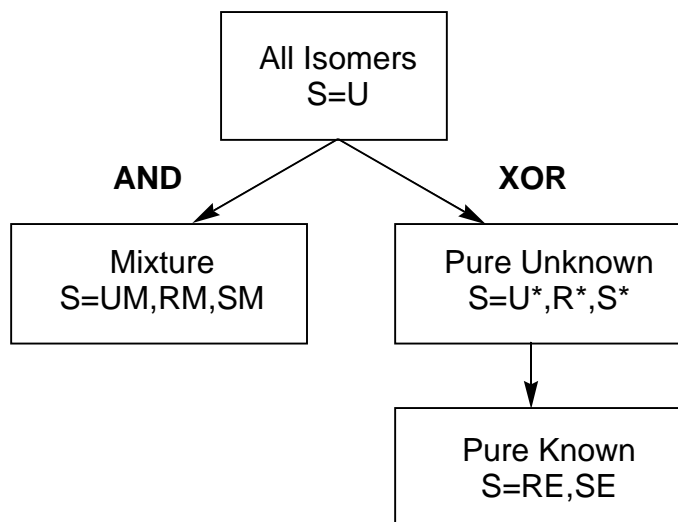


Figure 6. Hierarchical searching scheme for relative stereochemistry and mixtures.

The behavior of the R, S, and U attributes is similar. In explicit mode, R will only return chiral centers that are R. Likewise, U will only return chiral centers that are marked U (not those marked R or S). In hierarchical mode, U is defined to match U, R, or S.

Assuming one has a query with two chiral centers, Table 1 shows the hits that are returned in hierarchical mode. The attributes on the query atoms are shown in the columns. The attributes on structures in the database that might be returned as hits are shown in the rows. The cells indicate “yes” the structure is returned as a hit or “no” the structure is not returned as a hit. The notation RE/RE means the first chiral center is marked RE and the second chiral center is marked RE.

Table 1. Structures that will hit for various chiral queries in hierarchical mode.

Structures	Queries					
	RE/RE	R*/R*	RM/RM	UE/UE	U*/U*	UM/UM
RE/RE	yes	yes	no	yes	yes	no
SE/SE	no	yes	no	yes	yes	no
UE/UE	no	no	no	yes	no	no
R*/R*	no	yes	no	yes	yes	no
S*/S*	no	yes	no	yes	yes	no
U*/U*	no	no	no	yes	yes	no
RM/RM	no	no	yes	yes	no	yes
SM/SM	no	no	yes	yes	no	yes
UM/UM	no	no	no	yes	no	yes

Table 1 is somewhat simplified. For example, a query with S=UM at each chiral center would also return a structure marked S=RM at one center and S=UM at the other chiral center. These additional permutations were intentionally left out of the table because of the complication mentioned earlier regarding the relative stereo modifier. These should only be entered in the database in pairs or higher multiples. Instead of listing all possible permutations, only the simpler cases are listed in the table.

It is also important to discuss the relative stereochemistry attribute with respect to queries. The relative stereochemistry attribute describes the relationship of one chiral center in a molecule to another chiral center in the same molecule, so it does not make sense to submit a query with a single relative chiral center marked as other than U*. There is however, a difference in the search results when submitting a search with a single chiral center marked R* versus U*. The former query will hit structures with atoms marked RE, SE, R*, or S*. The latter query will hit structures with atoms marked RE, SE, R*, S*, or U*.

As with all chiral queries in UNITY, the structural part of the query must be chiral in order to do a search that involves stereochemistry. Chiral attributes on query atoms that are not chiral are removed. Note also that if an index is present on a molecule, it is only used to indicate which

chiral centers belong to the same group. The value of the index is not matched explicitly in either explicit or hierarchical mode.

Some examples (Fig. 7) help illustrate the results that will be returned from the searches in hierarchical mode. Query XVI shows a non-specific query. No chiral designators are present, so the chiral center will be treated as UE and all pure isomers and mixtures present in the database will be returned. Query XVII has one chiral center sketched without an explicit stereo mode that will be interpreted as explicit R (RE). The other two chiral centers are marked as relative with respect to each other. Based on the wedge bond information, these centers are R, so the SLN generated will have these centers marked as R*. Query XVIII has two chiral centers marked without an explicit stereo mode that will be interpreted as RE, and one center marked as relative. This latter center will be treated as a relative U and return the diastereomeric pair that is R at the first two centers and R or S at the third and, if present in the database, the structure that is R at the first two centers and U* at the third. Other hits are also possible, since structures in the database may have more than three chiral centers.

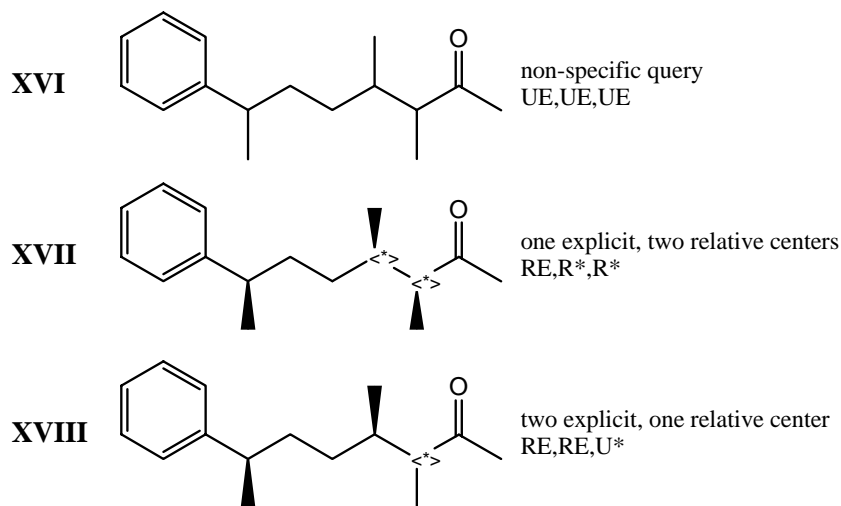


Figure 7. Example search queries involving chiral centers.

RELAXED MODE SEARCHING

Legacy data typically only distinguishes between molecules with absolute stereochemistry and mixtures. Some molecule sketchers can only mark molecules as being mixtures or absolute. The hierarchical mode when given a mixture in a query will not return molecules with

absolute stereo markings. In order to create more flexible searches a new stereochemistry search mode has been added. Relaxed mode searching treats relative and mixtures in the same fashion.

Table 2. Structures that will hit for various chiral queries in relaxed mode.

Structures	Queries					
	RE/RE	R*/R*	RM/RM	UE/UE	U*/U*	UM/UM
RE/RE	yes	yes	yes	yes	yes	yes
SE/SE	no	yes	yes	yes	yes	yes
UE/UE	no	no	no	yes	yes	yes
R*/R*	no	yes	yes	yes	yes	yes
S*/S*	no	yes	yes	yes	yes	yes
U*/U*	no	no	no	yes	yes	yes
RM/RM	no	yes	yes	yes	yes	yes
SM/SM	no	yes	yes	yes	yes	yes
UM/UM	no	no	no	yes	yes	yes

Relaxed mode differs from hierarchical mode in that when given a mixture or relative query it will return hits that match the query directly and the enantiomer of the query. Queries marked as absolute will only match structures marked as absolute.

Be aware that UE chiral center in a query would not match an achiral, symmetrical center in candidate for any stereo search mode. For example, in Fig. 8 query XIX will not hit symmetrical structure XX. To make this match occur you would have to remove S=UE attribute from the query XIX.

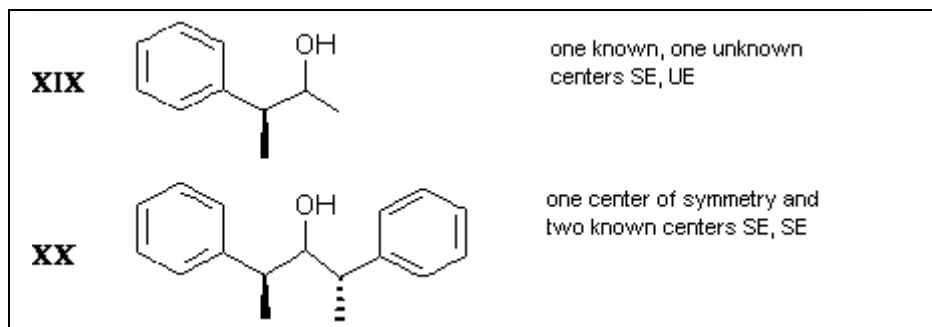


Figure 8. Example search a stereo query involving symmetrical structure.

An additional difference between hierarchical and relaxed mode is in respect to structures with multiple relative stereo groups. Such the hierarchical search mode matches stereo structure S with stereo query Q only when *each* isomer of the structure S matches *some* isomer of the query Q. The relaxed mode is more flexible: it matches structure S with

query Q when *some* isomer of the structure S matches *some* isomer of the query Q.

For example (see Table 3) the structure XXI matches the query XXII in the hierarchical mode because each isomer of this structure matches some isomer of the query: XXIa matches XXIIa and XXIb matches XXIIc. Contrariwise, the structure XXII does not match the query XXI in the hierarchical mode because there are some isomers of the structure (e.g. XXIIb and XXIIc) that does not match any query isomer. For the relaxed mode the both structures match each other: isomers XXIa and XXIIa are identical.

Table 3: Stereo Molecules and Isomers.

Structures	Isomers	
<p>XXI</p>	<p>XXIa</p>	<p>XXIb</p>
<p>XXII</p>	<p>XXIIa</p>	<p>XXIIb</p>
	<p>XXIIc</p>	<p>XXIIId</p>

OTHER VENDOR STEREOCHEMISTRY SUPPORT

In 2003, a supplier of chemical database software used extensively in the pharmaceutical industry announced that their product would support relative stereochemistry. Their approach is similar to that taken by Tripos, in that they have added support for marking a chiral center as being one of the following:

- **ABS** (explicit) means the center's chirality is completely determined, this is equivalent to SLN's E
- **OR** (relative), is equivalent to SLN's *.
- **AND** (mixture), is equivalent to SLN's M.

Their enhancements also make it possible to group stereocenters together by letting users add a numeric label to the center. Centers with the same label are grouped together. When a molecule has relative stereochemistry marked, the molecule's chiral flag is ignored.

This vendor's approach to dealing with relative stereochemistry is congruent with the approach taken by Tripos and the two are generally interchangeable.

SUMMARY

This article describes extensions to the SLN language to represent the isomeric mixtures that occur in reactions producing chiral centers. An additional modifier of E (explicit), * (relative), or M (mixture) is added to the standard stereo qualifiers of N, I, or U. In addition, an index is provided to group chiral centers together where the relative stereochemistry is known within the group, but not absolutely.

By adding the stereo mode modifier to the standard stereo attribute, enantiomeric mixtures and unresolved pure isomers can be described and searched. By also including a numeric index, several groups of chiral centers, whose chirality is known only relative to other atoms in the group, can be described. These extensions allow registration of any mixture that might be produced from these reactions into a UNITY or AUSPYX[®] database.

For more information

Please contact your Tripos Discovery Informatics consultant, call 1-800-323-2960 in the US, send email to contact_us@tripos.com, or visit our web site www.tripos.com.

USA
Tripos, Inc.
1699 South Hanley Road
St. Louis, MO 63144
1-314-647-1099

FRANCE
+33 1 69 59 29 49

GERMANY
+49 89 45 10 300

AUSTRALIA
+ 61 (7) 5439 9775

UNITED KINGDOM
+44 1908 650000

CANADA
+1 450 4334500

Tripos combines leading-edge technology and innovative science to deliver consistently superior chemistry-research products and services for the biotechnology, pharmaceutical and other life science industries. Within Tripos' Discovery Informatics business, the company provides software products and consulting services to develop, manage, analyze and share critical drug discovery information. Within Tripos' Discovery Research business, Tripos' medicinal chemists and research scientists partner directly with clients in their research initiatives, leveraging state-of-the-art information technologies and research facilities.

While the information presented in this white paper is believed to be current and accurate, it is presented as-is without guarantees or warranties of any kind. Tripos makes SYBYL, AUSPYX, and UNITY available subject to the terms of the Tripos Software License Agreement, and nothing in this white paper should be construed as a warranty concerning the performance of the product.

SYBYL, AUSPYX, UNITY, and the Tripos logo are trademarks or registered trademarks of Tripos, Inc. All other trademarks are the property of their respective owners.

©2006 Tripos, Inc. All rights reserved.

Printed in USA