The IUPAC Chemical Identifier – Technical Manual

Stephen E. Stein, Stephen R. Heller, Dmitrii V. Tchekhovskoi
Physical and Chemical Properties Division
National Institute of Standards and Technology
Gaithersburg, Maryland, U.S. 20899-8380

CONTENTS

I. ABSTRACT

This document presents a technical description of the IUPAC-NIST Chemical Identifier (INChI). It describes the scope of application of the INChI and explains the methods used for the creation of its various output ‘layers’. Rules used to resolve ambiguities in input information are also described. Appendices list valences and abbreviations, discuss the processing of the Identifier and describe some general problems in the representation of chemical identity.

II. INTRODUCTION

a. The IUPAC-NIST Chemical Identifier (INChI)

The objective of the Identifier is to provide a string of characters capable of uniquely representing a chemical compound. This involves finding and implementing a set of rules that transform an input ‘connection table’ into an output sequence of characters. Since INChI is intended to serve as a precise digital signature of a compound, it must have two properties: 1) different compounds (as defined by their ‘connection tables’) must have different identifiers and 2) a single compound must have a single identifier, regardless how its structure is drawn. The first requires the inclusion of all of the chemical features that distinguish one compound from another. The second requires the elimination of input information that reflects only the conventions used for drawing the compound.

Since a given compound may be represented at different levels of detail, in order to create a robust expression of chemical identity it was decided to create a hierarchical ‘layered’ form of the Identifier, where each layer holds a distinct and separable class of structural information, with the layers ordered to provide successive structural refinement. In addition to basic ‘connectivity’ and overall charge, the principal varieties of layers are mobile/fixed H-atoms (expresses tautomerism), isotopic composition and stereochemistry. The Identifier is created from the input structure in three steps: normalization (removing information not needed for layer construction and separating information into layers); canonicalization (generating a set of atom labels that do not depend on how the structure was initially drawn); and serialization (converting the set of labels derived from canonicalization into a string of characters, the INChI). The chemical ideas employed for creating the INChI appear in the normalization step, where conventions are removed while maintaining a complete description of the compound.

This ‘layered’ model allows chemists to represent chemical substances at a level of detail of their choosing. Except for main layer (atoms and their bonds), the presence of a layer is not required and appears only when corresponding input information has been provided. Moreover, because of inherent difficulties in fully describing certain structural details of some chemical substances, the main layer of the INChI is expected to provide a stable and reliable means for identifying complex chemical substances.

Adequate perception of mobile H-atoms required rather complex rules. They were needed to deal with combined effects of the different conventions employed for drawing chemical structures and the fact that mobile H-atoms are defined by their rapid isomerization reactions, which can depend on structural details and chemical environment. Rules were based on earlier published work with extensions based on experience with test sets, also using input from interested parties. Other rules were developed largely to perceive and, when possible correct inadequacies in input stereochemical information.

b. **Objective of this Document**

The principal objective of this document is to describe the technical issues involved in structure ‘normalization’ steps in the current version of the INChI. This includes both the scope of chemical substances covered by the INChI as well as the means of dealing with a variety of common problems involved in the representation of chemical compounds. Mathematical details of the algorithms used will not be presented. They have been derived from methods reported in the literature (listed in the Bibliography section at the end of this document). They will be made available in the form of tested and documented source code along with the final version of the INChI.

III. DISCUSSION

a. The Scope of the INChI

It was agreed at IUPAC meetings prior to the start of this project that the first version of the INChI should cover well-defined, covalently-bonded organic molecules. It was also agreed to include substances with mobile hydrogen atoms (tautomers, for instance). In the course of this project, it was found that with a straightforward extension organometallic compounds could be represented. Methods were found to also include variable protonation. Not included are polymers, molecular class representations (Markush structures), and conformations. Also, the present version only considers traditional organic stereochemistry (double bond - sp^2 and tetrahedral - sp^3) and the most common forms of H-migration (tautomerism). However, the layered structure of the INChI allows future refinements with little or no change to the layers described here.

By design, the INChI represents only a single type of connectivity (it ignores bond orders except for analyzing stereochemistry and H-migration) and does not explicitly represent positions of electrons. While this is not the conventional method for representing compounds, it provides an effective means of representing their identity.

In summary, the INChI is a series of characters derived by applying a set of rules to a chemical structure to provide a unique digital ‘signature’ for a compound. It has been developed under IUPAC auspices to serve as a uniform, openly available digital ‘name’ for a compound. It has been designed for use as a ‘plug-in’ for other chemical structure-based software systems.

b. Construction of the INChI

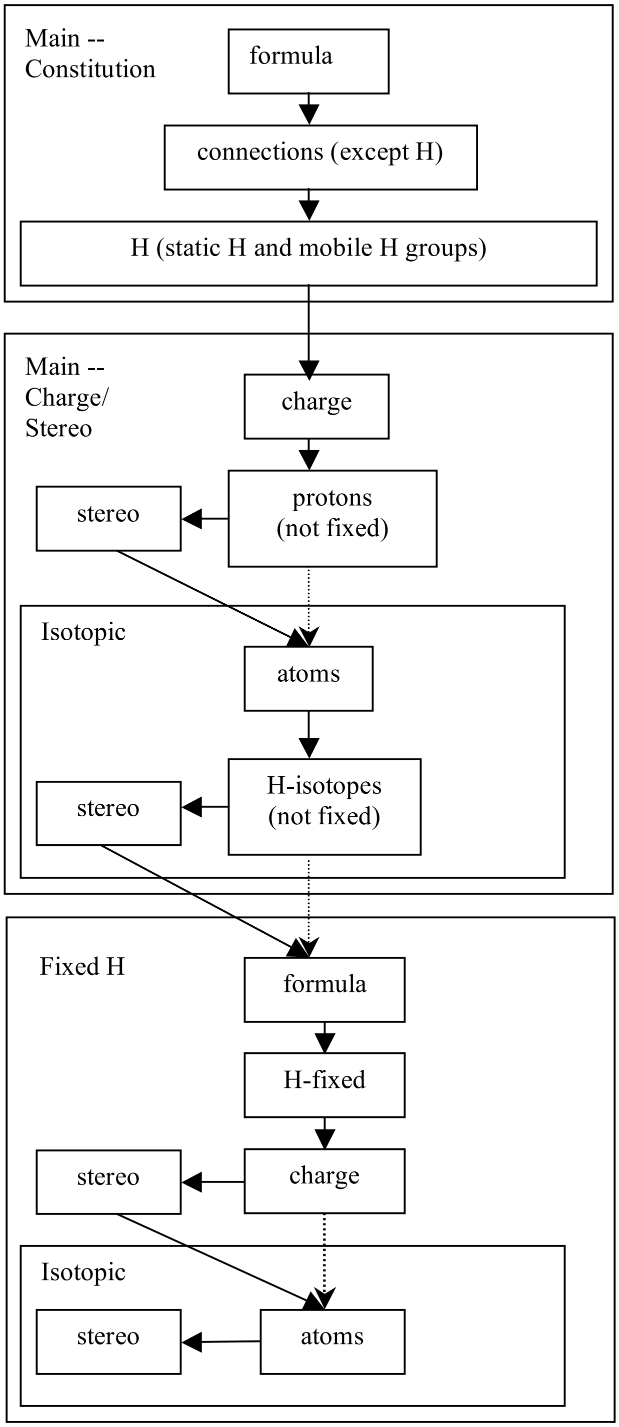
The INChI string is composed of one or more ‘layers’ that are successively built from information in an input ‘connection table’. Each layer is expressed as a string of characters. Layers are appended to one another in a strictly defined order: each layer except for the first layer has one and only one preceding (parent) layer. If the data necessary to create a layer is not available, that layer is omitted from the INChI representation. Values computed for each layer depend on prior layers. As a consequence, for example, two stereochemical layers for different compounds cannot be directly compared – comparisons must involve the complete set of preceding layers. Layers do not, however, depend on successive layers. Therefore, if two INChI strings are identical up to a layer, then the structural characteristics of the two represented structures are also identical up to that point. For brevity, if a requisite layer is identical to an earlier one in the same Identifier, abbreviations are used. For example, isotopic sublayers that are exactly equal to their non-isotopic counterparts are omitted when no ambiguity is created. Abbreviations are given in Appendix 2.

c. INChI Components

For structures that are composed of multiple interconnected (covalently bonded) components, a single INChI is generated, but each of the components retains its identity. Each layer contains information pertinent to all of the components (these are represented as conventional ‘dot-disconnected’ units in the formula layer, or with semicolons in other layers). In general, a valid INChI of an individual component of a more complex compound may be obtained by simply excising it from the INChI string. Details are given in Appendix 3.

d. The Five INChI ‘Layer’ Types

Depending on the information contained in the input structure, the INChI may be composed of up to five distinct varieties of ‘layers’, each representing a different class of structural information. These layers are discussed below and illustrated in Figure 1. Specific examples and more details are presented later.



Layer creation:
Main contains basic connectivity and is always present.
Charge/Stereo is added when needed.
When no isotopic atoms are present both 'isotopic' boxes are bypassed.
When mobile H-atoms are 'on', fixed H box is omitted.
When stereo is not present, all stereo boxes are bypassed.

Comparing INChI:
Before comparison the following items are optionally removable: protons, H-isotopes, entire fixed-H box, entire isotopic box, all stereo.
To compare absolute stereo to relative or racemic, "abs.inverted" and "type" stereo segments should be ignored.
For approximate comparison only formulas and connections may be used.

Order of Comparison for Sorting Components:
formula (without H)
connections (except terminal H)
total number of H
hydrogen atoms (except mobile)
(mobile-H only) H atoms (mobile groups)
(fixed H only) fixed-H chemical formula
(fixed H only) fixed H connections of H-fixed
The following is either for mobile-H or fixed-H
stereo bonds
stereo centers
stereo center inversions
isotopic atoms (isotopic H not included)
isotopic H (mobile and exchangeable H in mobile-H case are not included)
isotopic mobile groups (canonical IDs only)
isotopic stereo bonds
isotopic stereo centers
isotopic stereo center inversions
component charges

If fixed-H layers are same then main layers are compared
If main layers are same then fixed-H layers are compared if present.

Figure 1. INChI Layer Flowchart

The first two layers (chemical formula and connections) are derived solely from the simple connectivity information represented in the input structure. They entirely ignore pi-electrons and charge as well as stereochemical, tautomeric and isotopic information.

1. Main Layer

1.1. Chemical Formula

For a compounds composed of a single component, this is the conventional Hill-sorted elemental formula. For compounds containing multiple components, the Hill-sorted formulas of the individual components are sorted according to the guidelines in Figure 1 and separated by dots.

1.2. Connections

This lists the bonds between the atoms in the structure, partitioned into as many as three sublayers. The first represents all bonds other than those to non-bridging H-atoms, the second represents bonds of all immobile H-atoms, and the third provides locations of any mobile H-atoms. The last sublayer represents H-atoms that can be found at more than one location in a compound due to well-known varieties of isomerization. It identifies the groups of atoms that share one or more mobile hydrogen atoms.

2. Charge Layer

This represents net charge (surplus of protons over electrons) and does not depend on the contents of other layers. It may appear in as many as two sublayers:

2.1. Component Charge

The net charges of the components are represented in this layer as independent tags. By design, the INChI does not distinguish between structures that differ only in the formal positions of their electrons.

2.2. Protons

The number of protons removed from or added to the substance so that a given components may be represented without regard to their degree of protonation.

3. Stereochemical Layer

This is composed of two sublayers, the first accounts for double bond, sp^2 stereochemistry and the second for sp^3 tetrahedral stereochemistry and allenes. Note that the first sublayer is independent of the second, but not vice-versa.

3.1 Double Bond sp^2 (Z/E) Stereo

Expression of this stereo configuration is easily done in 2-dimensional drawings, and when double bonds are rigid, stereoisomerism is readily represented without ambiguity. However, in alternating bond systems, some non-rigid bonds may be formally drawn as double. Bonds in these systems, when discovered by INChI algorithms, are not assigned stereo labels. Also, to avoid needless stereodescriptors in aromatic and other small rings, no sp^2

stereoisomerism information is generated in rings containing 7 or fewer members.

3.2 Tetrahedral sp^3 Stereo

Tetrahedral sp^3 stereochemistry is readily represented using conventional wedge/hatch (out/in) bonds commonly employed in 2D drawings. Relative sp^3 stereochemistry is represented first, optionally followed by a tag to indicate absolute stereochemistry. When a stereocenter configuration is not known to the structure author, an ‘unknown’ descriptor may be specified, which will then appear in the stereo layer. If a possible stereocenter is found, but no stereo information is provided, when a stereolayer is requested this stereocenter will be represented by a not-given (‘undefined’) flag.

4. Isotopic Layer

This is a layer in which different isotopically labeled atoms are identified. Exchangeable isotopic hydrogen atoms (deuterium and tritium) are listed separately. The layer also holds any changes in stereochemistry caused by the presence of isotopes.

5. Fixed-H Layer

When potentially mobile H atoms are detected and the user specifies that they should be immobile (tautomerism not allowed), this layer binds these H atoms to the atoms specified in the input structure. When this, in effect, causes a change in earlier layers, appropriate changes are added to this layer (earlier layers 1-4 are not affected).

e. INChI Structure

Figure 2 below describes the ordering of all possible layers in an Identifier. Individual layers are preceded by /? where ? is a lowercase letter that distinguished that layer. In the Identifier itself, actual layer contents replace the annotations shown below in curly braces. Titles in *Italics* are only shown for clarity.

```
{INChI version}
```

1. Main Layer (M):

```
/{formula}
/c{connections}
/h{H_atoms}
```

2. Charge Layer

```
/q{charge}
/p{protons}
```

3. Stereo Layer

```
/b{stereo:dbond}
/t{stereo:sp3}
/m{stereo:sp3:inverted}
/s{stereo:type (1=abs, 2=rel, 3=rac)}
```

4. Isotopic Layer (MI):

```
/i{isotopic:atoms}*
/h{isotopic:exchangeable_H}
/b{isotopic:stereo:dbond}
/t{isotopic:stereo:sp3}
/m{isotopic:stereo:sp3:inverted}
/s{isotopic:stereo:type (1=abs, 2=rel, 3=rac)}
```

5. Fixed H Layer (F):

```
/f{fixed_H:formula}*
/h{fixed_H:H_fixed}
/q{fixed_H:charge}
/b{fixed_H:stereo:dbond}
/t{fixed_H:stereo:sp3}
/m{fixed_H:stereo:sp3:inverted}
/s{fixed_H:stereo:type (1=abs, 2=rel, 3=rac)}
```

(6.) Fixed/Isotopic Combination (FI)

```
/i{fixed_H:isotopic:atoms}*
/b{fixed_H:isotopic:stereo:dbond}
/t{fixed_H:isotopic:stereo:sp3}
/m{fixed_H:isotopic:stereo:sp3:inverted}
/s{fixed_H:isotopic:stereo:type (1=abs, 2=rel, 3=rac)}
/o{transposition}
```

Figure 2. Layers of the identifier

Construction of the Identifier involves, in effect, the in-memory labeling of the atoms in the input connection table, with equivalent atoms assigned identical labels by the canonicalization procedure along with rules for outputting (serializing) that information. The isotopic layers (one derived from the main layer and a second derived, when needed, from the fixed H layer) simply provide the labels of the isotopic atoms. Stereo layers contain the parity assignments for specific atoms and bonds, and are separately derived as required for up to four layer sequences, M, MI, F, FI . A ‘serialization’ routine produces output character strings for each of the layers. Note that if no information is available for a layer, that layer is omitted. For example, when there is no isotopic labeling, the layer MI and FI part of Fixed H layer are not present.

f. Implementation

Algorithms described here have been implemented in Microsoft Windows-based programs described in an accompanying ‘User’s Guide’ (wINChI11b.exe and cINChI11b.exe programs). Test programs are available at http://chemdata.nist.gov/IChI/INChIv11b.zip. An API (application program interface) will also be provided as well as full source code (in the ‘C’ programming language). Structures shown in this document are provided along with these programs.

IV. DETAILS AND EXAMPLES

a. Main Layer (M)

The identity of each atom and its covalently-bonded partners provide all of the information necessary to construct this layer. All information regarding π-bonds, charge, isotopic composition, tautomerism and stereochemistry is ignored. This ‘normalization’ process avoids many of the complexities commonly encountered in structure representation. For example, nitro groups can be input using any of the common representations and problems associated with representing zwitterions and special valences are avoided as are issues concerning alternating bonds and aromaticity.

This form of representation is unusual for most chemists, since conventional structure representations generally include bonding details that are not needed for chemical identification and often omit some or all hydrogen atoms. In effect, the present representation describes the single-bond network of a molecule and avoids any description of the ‘positions’ of all pi-bonds bonds and electrons. Any excess or deficit of electrons (overall charges) is represented in a separate layer.

Figures 3-6 are examples of input structures, their normalized structures and results of canonicalization for this basic layer. In the canonical numbering structure, the large numbers designate classes of equivalent atoms (an ‘in-memory’ representation), small numbers (canonical identification numbers) are used in the actual INChI generation process (serialization).

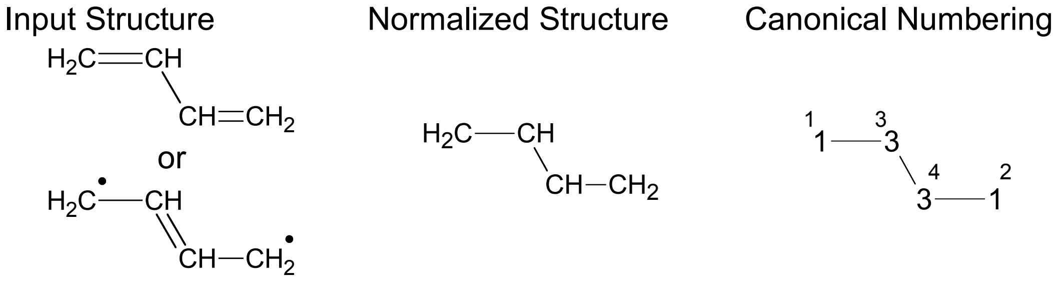
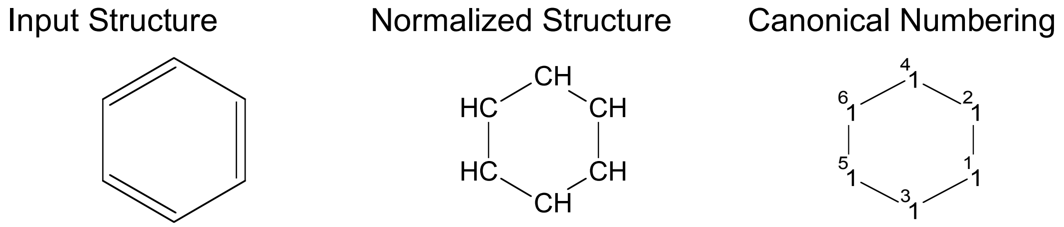


Figure 3

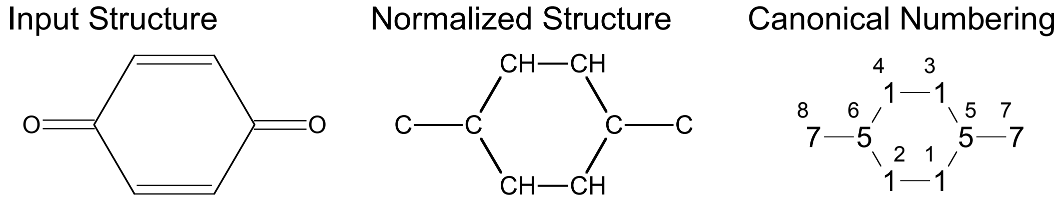


Input Structure

Normalized Structure

Canonical Numbering

Figure 4

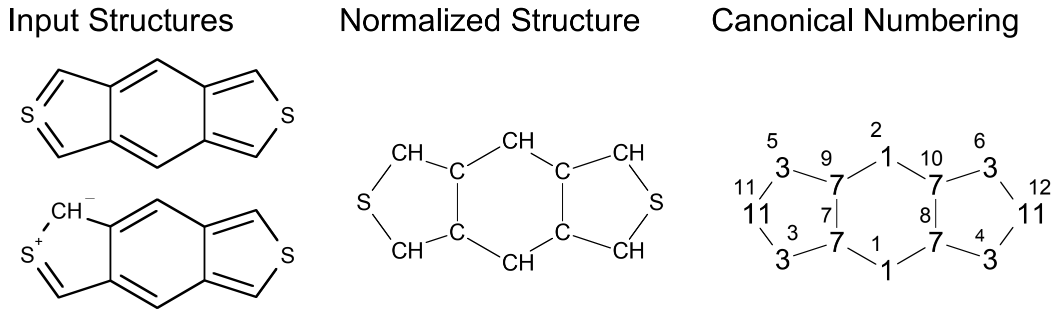


Input Structure

Normalized Structure

Canonical Numbering

Figure 5



Input Structures

Normalized Structure

Canonical Numbering

Figure 6

‘Munchnones’ serve to illustrate the many different ways that certain structures may be represented, the last being the normalized form used for the INChI

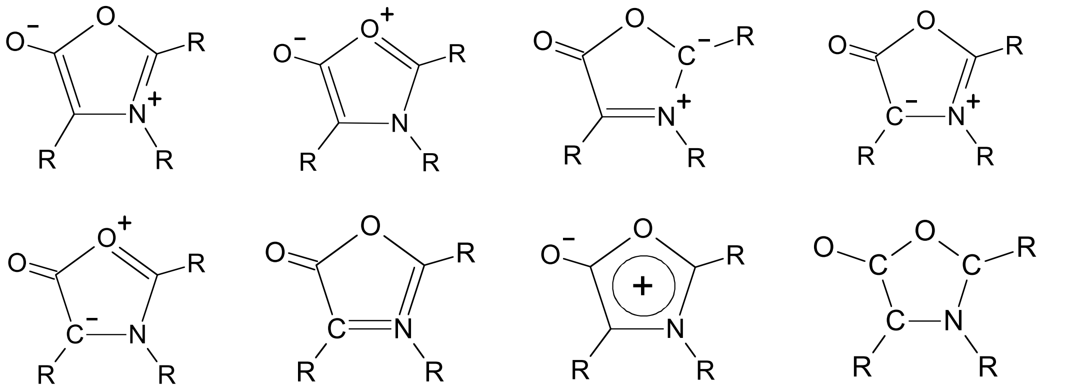


Figure 7

While bond orders are not used in the representation, hydrogen atoms are required. If there is ambiguity concerning the number of H-atoms in a structure (i.e., its chemical formula is not clear), a reliable INChI cannot be created. The INChI generator uses accepted valence rules to detect such ambiguity and issues warnings when detected.

b. Normalization

The structures in the above examples did not need normalization steps except for ignoring bond types and charges. However, the following additional normalization steps are sometimes needed to, in effect, deal with ambiguities in structure representation, especially those involving mobile hydrogen atoms.

INChI applies as many as six varieties of normalization rules to a given structure. These are described briefly here and in detail later. Steps 1-4 are designed to eliminate a variety of structure drawing conventions that could interfere with later processing. Step 5 finds protons necessary to dealing with variable protonation. Step 6, the final normalization step, includes the discovery of conventional tautomeric patterns (depicted in Table 4) and ‘resonances’ that may occur due to bond alternation or positive charge migration along paths of alternating bonds. When certain negatively charged heteroatoms are present or additional work is required for complete ‘hard’ proton addition/removal, step 6 discovers additional patterns of exchanging hydrogen atoms and charges. In some compounds, resolving these ambiguities results in an increased ‘mobility’ of H-atoms relative to conventional tautomeric rules.

The normalization and the stereochemical part heavily relies on testing whether a bond order can be changed due to presence of an alternating bond circuit or possibility of a hydrogen atom, charge, or radical center to migrate along a path of alternating bond. This testing is based on matching algorithm described in details in [4].

The specific type of normalization performed in provided in the Auxiliary information section of INChI output. This includes: (1) conventional tautomerism, (2) additional exchange of H and negative charges typical for products of heterolytic dissociation, and (3) ‘hard’ removal/addition of protons that is accompanied by a wider exchange of H and negative charges. When a binary representation of the normalization type includes the bit corresponding to 2^n then the type number (n+1) was invoked. For example, normalization type = 6 = 2 + 4 = $2^{3-1} + 2^{2-1}$ means that (3), ‘hard’ proton addition/removal, and (2), additional exchange of H and possibly negative charges, were invoked.

For the fixed H layer, only moving positive charges along paths of alternating bonds is allowed.

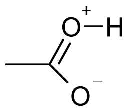
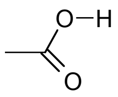
The normalization steps are:

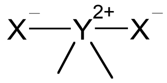
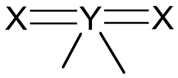
1. Alter the structure drawing
2. Disconnect “salts”
3. Disconnect metals
4. Eliminate radicals if possible
5. Process variable protonation
6. Process charges and mobile H

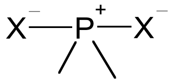
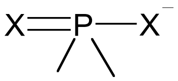
Note. Many examples of chemical structures below are hypothetical; they were selected to illustrate the concepts on small structures.

Step 1. Alter the structure drawing

Table 1. Altering the structure drawing

type	Input fragment		Fixed fragment	note
1	$X-H^{+}$	→	$\overset{+}{X}-H$	X is any atom except H
2	$X-H^{-}$	→	$\overset{-}{X}-H$	X is any atom except H
3	$\overset{-}{X}-Y=X^{+}$	→	$X=Y-X$	X=N, P, As, Sb, O, S, Se, Te
Example of 3		→		X = O, Y = C

4		→		X=O, S, Se, Te
				Y=S, Se, Te

5		→		X=O, S, Se, Te

Step 2. Disconnect “salts”

Some salts are commonly represented in either connected or disconnected forms. The approach used by INChI is to always disconnect salts. The base definition for recognition of connected salts is:

M-X or Y-M-X

where M is a metal atom and HX, HY are “acids”.

In connected “salts”, metals are connected by single bonds only and do not have H-atoms connected to them. Metal valences should be the lowest known to

INChI valence or, for some metals, the valence may also be the 2nd lowest valence. Positively charged metals should have the lowest known to INChI valence (See Appendix 1).

Metals are all elements except these:

Table 2. Non-metals

IIIA	IVA	VA	VIA	VIIA	VIIIA
13	14	15	16	17	18
				H	He
B	C	N	O	F	Ne
	Si	P	S	Cl	Ar
	Ge	As	Se	Br	Kr
			Te	I	Xe
				At	Rn

“Acid” is one of the following three:

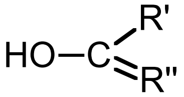
HX (X=F, Cl, Br, I)		HO−C≡R

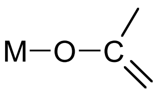
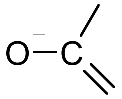
Figure 8. Acid definition

Upon disconnection atom X or O of the acid receives a single negative charge; the charge of the metal is incremented.

Substances drawn as H_4N-X are disconnected to NH_3 and HX.

Several examples are shown in the table 3:

Table 3. Examples of salt disconnection

	connected		disconnected		
1	NH_4—O—C	→	NH_3	+	HO—C
2	NH_4—X	→	NH_3	+	HX (X=F, Cl, Br, I)
	Below M is a metal				“acid” anion
3		→	M+	+	
4	M−O−C≡	→	M+	+	O^-−C≡
5	M—X	→	M+	+	X^- (X=F, Cl, Br, I)

Note that inorganic acids do not fit the salt definition. For example, sodium nitrate is treated as a coordination compound, so may be reconnected on user request.

Step 3. Disconnect metals

In an effort to deal with the various different conventions used for drawing organometallic compounds, all metal atoms are disconnected in the main layer. In the process, the charges for disconnected F, Cl, Br, I, At, O, S, Se, Te, N, P, As, B are adjusted if possible by transferring charge to the metal atom.

The user may request to add a “reconnected” layer that generates an INChI that contains all bonds given in the input structures. A disconnected “salt” (step 2 above) cannot be reconnected this way.

Step 4. Eliminate radicals if possible

This can be illustrated as follows:

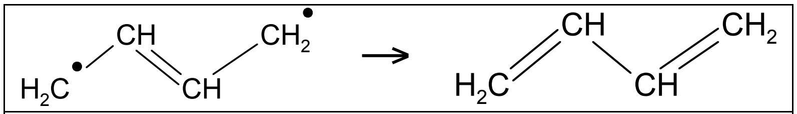


Figure 9. Radical cancellation

Step 5. Process variable protonation (charges and mobile H).

This step is needed to represent substances that have variable or unknown degrees of protonation. The necessary condition for this step is existence of charges +1 or -1 located on non-metal atoms that have standard valences (See Appendix 1). The total charge on these atoms is also counted and used later. Charges on atoms that are adjacent to other charged atoms are not counted. Non-ring bonds altered during variable protonation processing are marked as non-stereogenic. Described below aggressive (‘hard’) proton removal or addition may be turned off (command line option /NoADP or ‘Disable Aggressive (De)protonation).

Step 5.1. Remove protons from charged heteroatoms.

This step finds and disconnects protonated atoms and places them in a separate proton (charge) layer. If the structure contains atom $Y'H_m^+$ ($m \geq 1$, Y' is N, P, O, S, Se, or Te) then it is replaced with $Y'H_{m-1}$. This is a “simple removal” of a proton.

Since some protonated atoms are, in effect, concealed by alternating bond conventions, a separate effort is made to find and disconnect these protons. This “hard removal” involves changing bonds and removing H from formally uncharged atoms. It may be illustrated as follows. If there exist atoms $=N^+$ or $\equiv N^+$ **and** $-NH_m$ ($m \geq 1$, at least one neighbor of N must be Y) or =Y-$\underline{O}$H (Y= C, N, P,

As, Sb, S, Se, Te, Cl, Br; $\underline{O}$=O, S, Se, Te) then an attempt is made to find a fragment containing an alternating path (a, b,… are other atoms) and remove a proton:
$H_mN{-}b{=}c{-}d{=}N^+ \rightarrow H_mN^+{=}b{-}c{=}d{-}N \rightarrow H_{m-1}N{=}b{-}c{=}d{-}N + H^+$ or
$\underline{HO}{-}Y{=}a{-}b{=}c{-}d{=}N^+{<} \rightarrow H\underline{O}^+{=}Y{-}a{=}b{-}c{=}d{-}N{<} \rightarrow \underline{O}{=}Y{-}a{=}b{-}c{=}d{-}N{<} + H^+$

More aggressive transformations are also possible, for example

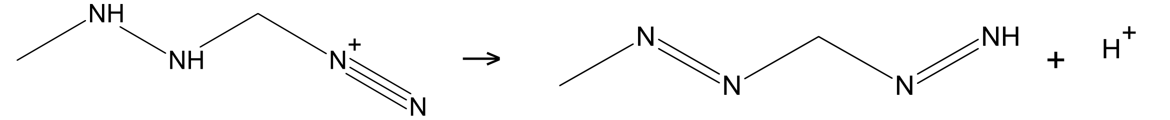


Figure 10. Example of ‘hard’ proton removal

During this process:

(a) positive charges may be moved between N$^+$, N$^-$ and N (except N in -N=$\underline{O}$);

(b) negative charges may be moved between N$^+$, N$^-$, N, and $\underline{O}$, $\underline{S}$ in -Y=$\underline{O}$, =Y=$\underline{O}$, =Y-$\underline{O}$X, ≡Y-$\underline{O}$X, -C-$\underline{S}$X, -$\underline{O}$'-$\underline{O}$X, ≡N$^+$-$\underline{O}$H, =N$^+$=$\underline{O}$, -N$^-$-$\underline{O}$H, where $\underline{O}$ is O, S, Se, or Te; $\underline{S}$ is S, Se, or Te; X is H or -; Y≠C≠N may carry ±1 charge; N in –N=$\underline{O}$ is excluded.

(c) atoms H may be moved between atoms described in (b)

The neutralization of positive and negative charges may occur. A simple exchange of atom H and a negative charge between two atoms without changing bonds is not allowed.

Examples of such normalization are given in sample structures provided with the test program and on Fig. 10 and 11(1a-c, 2a-b).

Step 5.2. Remove protons from neutral heteroatoms

If the total charge referred in Step 5 is positive and the structure has fragments =C-$\underline{O}$H, -$\underline{O}$-$\underline{O}$H, C-$\underline{S}$H, or =N-$\underline{O}$H, then hydrogen atoms are removed from the fragments and replaced with negative charges until either no more hydrogens are available or the charge has been reduced to zero. This is a “simple removal” of a proton. It is illustrated on Fig. 11(3a-c).

If the total charge is still positive then a “hard proton removal” procedure similar to the previously described one is executed.

During this process:

(d) positive charges may be moved between atoms described in 5.1 (a);

(e) negative charges may be moved between atoms described in 5.1 (b)

(f) atoms to receive H if the procedure succeeds: $\underline{O}$ in -C=$\underline{O}$, =C=$\underline{O}$, =N$^+$=$\underline{O}$, and -N=$\underline{O}$

(g) atoms H may be moved between atoms described in 5.1 (b) except atoms described in (f) above

If the procedure succeeds it moves H from atoms described in 5.2(g) to atom $\underline{O}$ described in 5.2(f). After that the H is removed from that $\underline{O}$ as a proton, leaving negatively charged $\underline{O}^{-}$ thus reducing the positive charge. An example is on Fig. 11(4a-c).

Step 5.3. Add protons to reduce negative charge

If the total charge referred in Step 5 is negative or has become negative due to positive charge removal and the structure has fragments $\text{-C=}\underline{O}^{-}$, $\text{-}\underline{O}\text{-}\underline{O}^{-}$, $\text{C-}\underline{S}^{-}$, or $\text{=N-}\underline{O}^{-}$, then protons are added to the fragments replacing negative charges with atoms H until the total charge is reduced to minimal or zero. This is a “simple addition” of a proton.

If the total charge is still negative then a “hard proton addition” procedure similar to the previously described one is executed.

During this process:

(h) positive charges may be moved between atoms described in 5.1 (a);
(i) atoms to receive negative charge if the procedure succeeds are atoms described in 5.2(f):
(j) negative charges may be moved between atoms described in 5.1 (b) except atoms described in (i) above
(k) atoms H may be moved between atoms described in 5.1 (b)

If the procedure succeeds it moves negative charge from atoms described in 5.3(j) to atom $\underline{O}$ described in 5.3(i).. After that this negative charge is replaced with atom H which is equivalent to a proton addition thus reducing the negative charge. An example is on Fig. 11(5a-c).

Step 6. Process charges and mobile H

For a structure that does not have charged atoms the tautomeric atoms and bonds are detected and marked. Atoms that may exchange hydrogen atoms are considered to belong to a “mobile H group”.

As evident from the foregoing discussion, the existence of a ‘protonated’ site is sometimes not readily apparent in a structural drawing. The normalization algorithm is designed to resolve complications that arise from ambiguities introduced at step 5 during “hard” or incomplete “simple” removal or addition of protons and in case of charged atoms resembling results of heterolytic dissociation. Below are examples of such ambiguities.

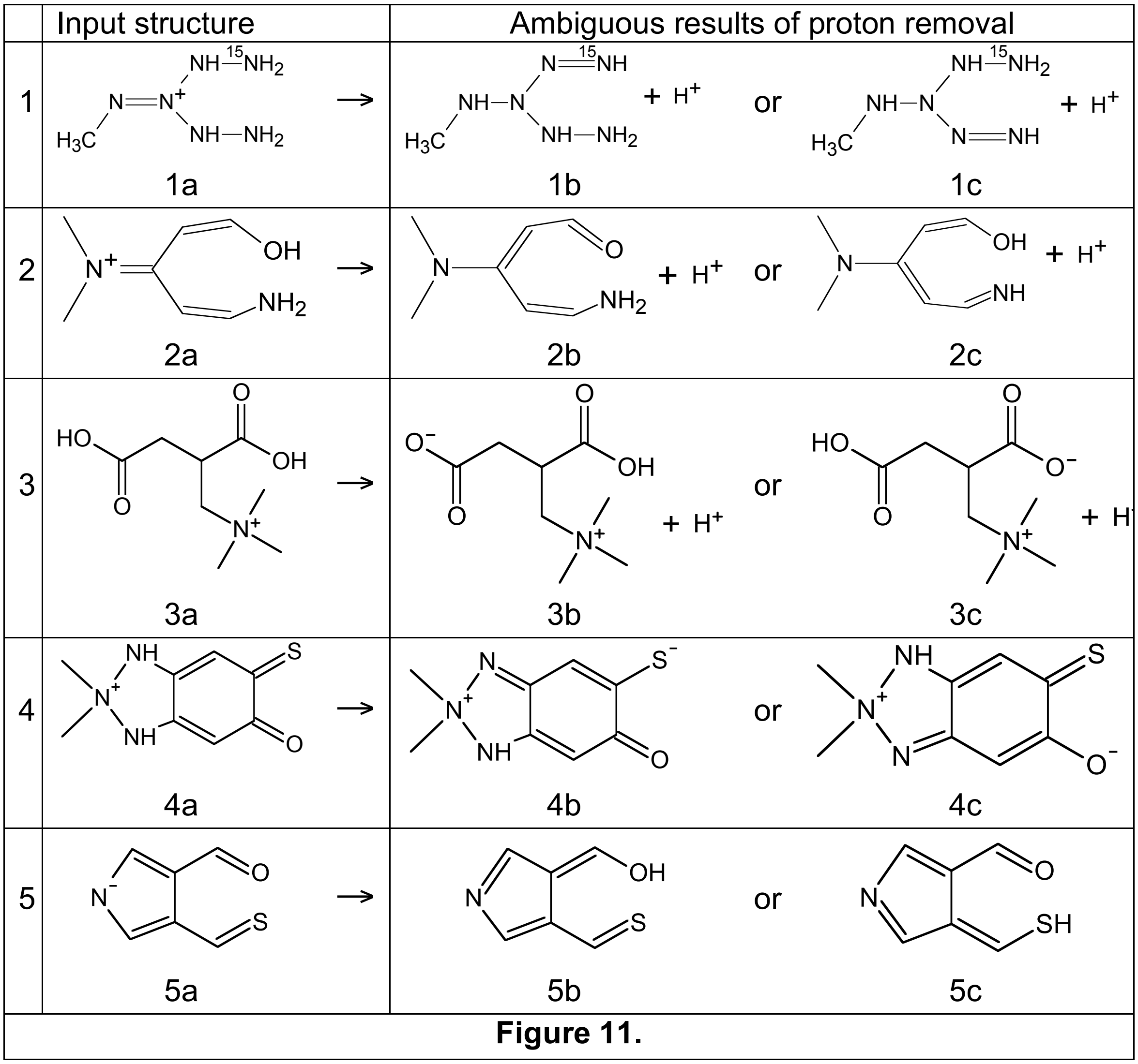


Figure 11.

Rows 1, 2, 4, and 5 illustrate “hard” proton removal ambiguities, row 3 illustrates incomplete “simple” removal of protons; structures 3b and 3c also illustrate ambiguous representation in case of charged atoms resembling results of heterolytic dissociation.

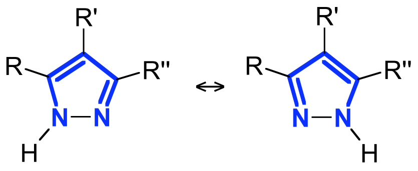
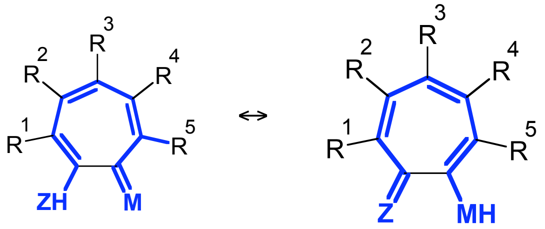
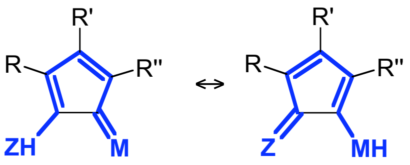
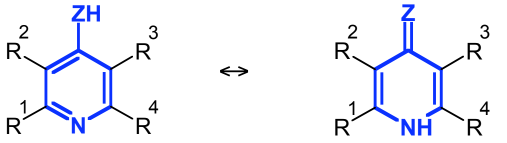
The information about the type of normalization invoked is in the first item in the INChI Auxiliary information. It is a number such that in its binary representation each bit manifests a specific invoked type of normalization. The bit corresponding to 2^2 means hard proton removal, bit 2^1 means treatment of negative charge position ambiguity similar to 3b and 3c on Fig. 11. The bit corresponding to 2^0 means “simple” tautomerism.

Step 6, procedure 1: Simple tautomerism detection

The Main layer must be the same for any arrangement of mobile hydrogen atoms. This is achieved by the logical removal of mobile H-atoms and the tagging of H-donor and H-receptor atoms. To identify these H-atoms we have adopted the straightforward varieties of H-transfer tautomerism listed in Table 4 (see also reference 1) and illustrated in Figures 12 and 13 using Guanine as an example.

Table 4	

M=Q–ZH ↔ MH–Q=Z,
or
M=Q–Z^- ↔ M^-–Q=Z
	M, Z = N^{III}, O^{II}, S^{II}, Se^{II}, Te^{II} (Roman superscripts designate chemical valence)

	Q = C, N, S, P, Sb, As, Se, Te, Br, Cl, I
	H = hydrogen, deuterium, or tritium
The “=” bond may be a double bond, a bond in the alternating single/double bond ring, or a “tautomeric” bond (shown in blue) Below H atom can be replaced with a negative charge	
	
	

Guanine example.

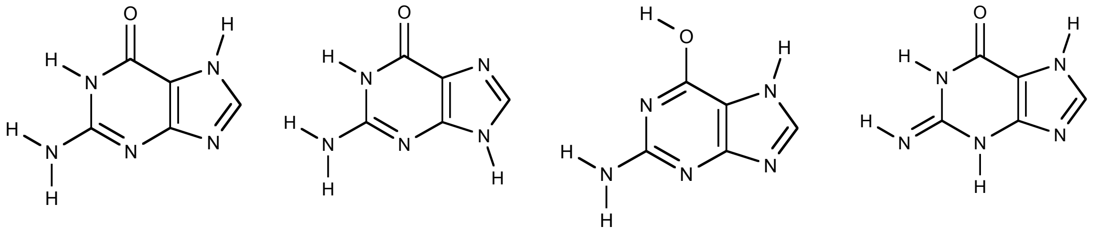


Figure 12. Tautomeric structures of Guanine (not all possible are shown)

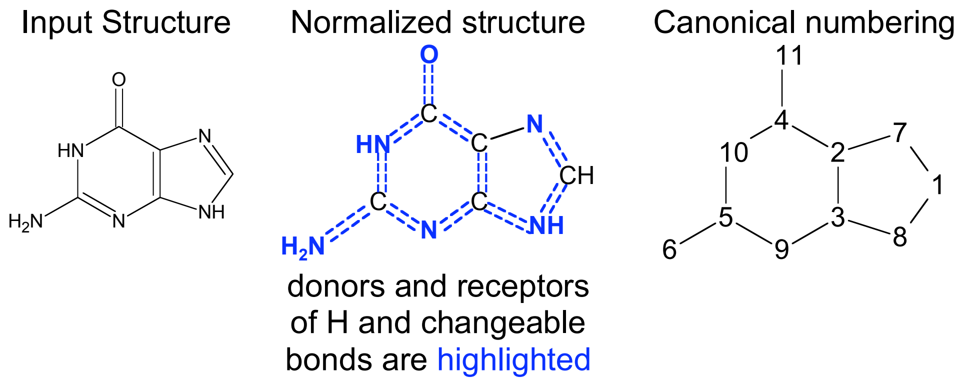


Figure 13. Guanine normalization and canonical numbering

INChI for Guanine (optional fixed H layer included) is

INChI=1.12Beta/C5H5N5O/c6-5-9-3-2(4(11)10-5)7-1-8-3/h1H,(H4,6,7,8,9,10,11)/f/h6H2,8H,10H

The same INChI of Guanine with added annotations {in curly braces} is

```
INChI=
{version}1.12Beta
/{formula}C5H5N5O
/c{connections}6-5-9-3-2(4(11)10-5)7-1-8-3
/h{H_atoms}1H,(H4,6,7,8,9,10,11)

/f{fixed_H:formula}
/h{fixed_H:H_fixed}6H2,8H,10H
```

Explanation of Guanine Identifier:

/h{H_atoms}**1H,(H4,6,7,8,9,10,11)**
atom number 1 has one H, 4 atoms H are shared by atoms 6,7,8,9,10, and 11

/h{fixed_H:H_fixed}**6H2,8H,10H**
atom 6 has 2H, atom 8 has 1H, atom 10 has 1H.

/f{fixed_H:formula}
is empty because the chemical formula for fixed H layer is same as in the Main layer.

This example illustrates important features of INChI:

- Ignoring the fixed H layer (beginning with /f in the box above) establishes the equivalence of different tautomeric forms of Guanine.
- Including the fixed H layer specifies a single tautomeric form of Guanine.

Step 6, procedure 2. Moveable positive charge detection

Positive charges located on N-atoms are considered moveable along alternating bonds between these atoms. This also applies to phosphorus atoms. Atoms that may exchange positive charges are assigned to a “mobile charge group”. The interference between mobile H and mobile charges may occur.

Hypothetical structures on Fig. 14a-14c serve as an illustration.

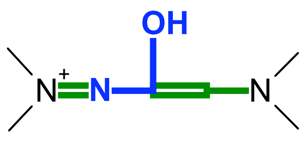
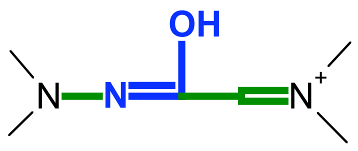
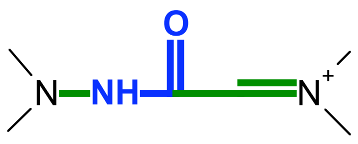
		

Figure 14a	**Figure 14b**	**Figure 14c**

Structure 14b was obtained from 14a by formally moving the positive charge from left to right along an alternating bond path. This allows the discovery (Fig 14b) of a tautomeric pattern (highlighted in **blue**). Bonds that may be changed by moving positive charges are highlighted in **green**. Fig. 14c shows another tautomeric form obtained from 14b. Note that Fig. 14c does not allow movement of a positive charge back from right to left. These three structures generate the same identifier.

For the purpose of detecting stereogenic bonds the algorithm must also provide a means for testing whether a bond order is changeable. INChI assumes that a changeable bond cannot support Z/E stereoisomerism. This is accomplished by introducing fictitious bonds and atoms (used only for internal processing) that represent a mobile H group (red H below) and charge group (red plus below). In the mobile H group fictitious double bonds (red) point to the atom-donors of H or negative charge; in mobile positive charge group fictitious single bonds point to positively charged atoms.

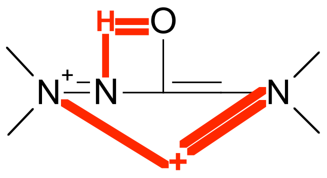
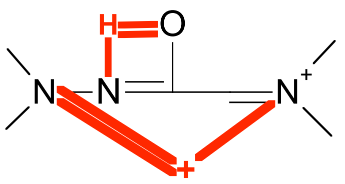
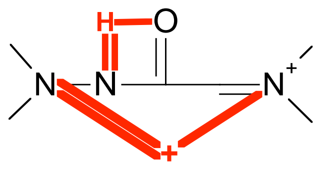
		

Figure 15a	**Figure 15b**	**Figure 15c**
Internal representation of structures from Figures 14a-14c.		

After the discovery of a new mobile group it is added to the structure. This results in the discovery of changeable bonds. In case of structure on Fig. 14a adding a charge group allows to discover changeable bond N-C (shown in blue) and, as a result, discover the mobile H group. These processing steps correct for common ambiguities in input information for conjugated systems where Z/E stereochemistry is implied by the drawing, but was not really intended.

Step 6, procedure 3. Additional normalization

As mentioned above, complications arise from ambiguities introduced at step 5 during “hard” or incomplete “simple” removal or addition of protons and in case of charged atoms resembling results of heterolytic dissociation. Since there could be more than one possible set of added/removed proton locations or more than one alternating path for “hard” addition or removal, ambiguity may be introduced. Another potential source of ambiguity (already mentioned in the introduction to Step 6 and illustrated on Fig, 11, structures 3b and 3c) can be found in a hypothetical zwitterionic structure that may be drawn in more than one way:

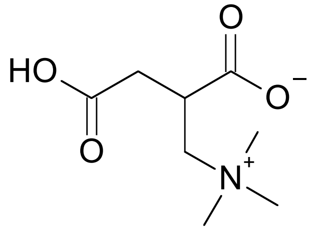
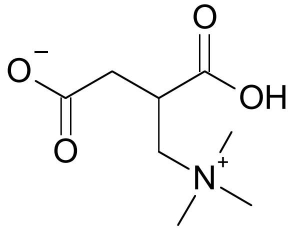
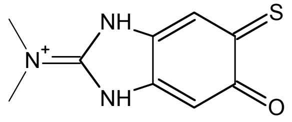
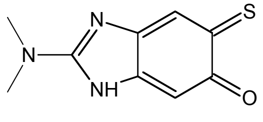
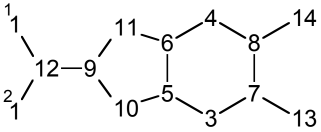
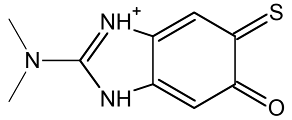
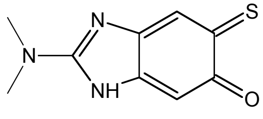
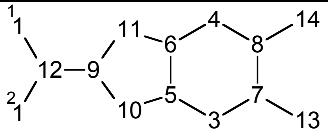
	

Figure 16a	**Figure 16b**

To avoid the ambiguity due to “hard” removal of protons or uncertain location of acidic hydrogen atoms the structure is tested for formal possibilities of:

1) moving positive charges between two atoms N or two atoms P;
2) discovery new tautomeric patterns described in Table 4;
3) moving H or negative charges between heteroatoms along paths of alternating bonds between atoms M and Z located in fragments MX-Q and Z=Q (M, Z, and Q are defined in Table 4)
4) removing a pair of H and/or negative charges from a pair of heteroatoms M connected by a path of alternating bonds and attaching this pair of H and/or negative charges to another pair of heteroatoms Z connected by a path of alternating bonds. If ‘hard’ proton addition or removal was done M and Z definition is relaxed to include donors and acceptors of H and negative charges defined in Step 5.1(b).
5) The final step is executed only in case of mobile negative charges. It puts all discovered atoms that possess previously discovered H and negative charges into a single mobile group. Atoms O and S located in fragments -Q-SX and -O'-OX (definition of Q is in Table 4, definitions of O, S and X are in the Step 5.1(b)), if present, are added to this group.

Normalization Limits. This approach avoids the possibility of representing different tautomeric representations of a given substance with different Identifiers in most, but not all cases. This imprecision results from a compromise between keeping in the identifier as many structural features as possible while avoiding ambiguities introduced by 'hard' (de)protonation, a process necessary for dealing with variable protonation. One of the underlying normalization conventions is a free migration of positive charges between atoms N along paths of alternating bonds. An example of a pair of structures that differ by the positive charge location is of Fig. 16c and 16f.

	Input Structure		Step 5 result	Canonical numbering
1	 **Figure 16c**	→ (hard)	 **Figure 16d**	 1 1 11 6 4 8 14 12 9 2 1 10 5 3 7 13 **Figure 16e**
2	 **Figure 16f**	→ (simple)	 **Figure 16g**	 1 1 11 6 4 8 14 12 9 2 1 10 5 3 7 13 **Figure 16h**

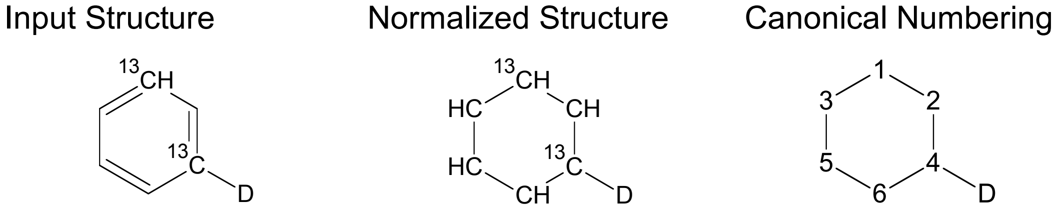
These structures should have same identifiers. Unfortunately, they don't. The Step 5 of the normalization produces seemingly identical structures (Fig. 16d and 16g). However, these structures are treated differently at Step 6. Since the 'hard' deprotonation triggers additional 'collectivization' of hydrogen atoms, Step 6 allows atom H to migrate between 4 heteroatoms of structure 16c (canonical numbers 10,11, 13, 14) and only between two atoms N (canonical numbers 10, 11) of structure 16f. The identifiers for 16c, 16f, and 16g are:

.

16c	**1.12Beta/C9H9N3OS/c1-12(2)9-10-5-3-7(13)8(14)4-6(5)11-9/h1-2H3,3-4H,(H,10,11,13,14)/p+1**
16f	**1.12Beta/C9H9N3OS/c1-12(2)9-10-5-3-7(13)8(14)4-6(5)11-9/h1-2H3,3-4H,(H,10,11)/p+1**
16g	**1.12Beta/C9H9N3OS/c1-12(2)9-10-5-3-7(13)8(14)4-6(5)11-9/h1-2H3,3-4H,(H,10,11)**

c. Isotopic Layer (I)

This is the most straightforward structural layer to compute.



INChI=1.12Beta/C6H6/c1-2-4-6-5-3-1/h1-6H/i1+1,4+1D

Figure 17

The isotopic layer is /i1+1,4+1D. It contains canonical atom number followed by the isotopic shift (13 – 12 = +1) followed by isotopic hydrogen (D) if present.

The only complexity arises for isotopically labeled hydrogen atoms that can undergo tautomerism. In the mobile H group these hydrogen atoms are treated as non-isotopic; the number of these mobile isotopic hydrogen atoms is appended to the ”exchangeable isotopic hydrogen atoms” part of isotopic layer. The same is done to isotopic hydrogen atoms that may be subject to heterolytic dissociation in aqueous solution (for example, D in R-SD)

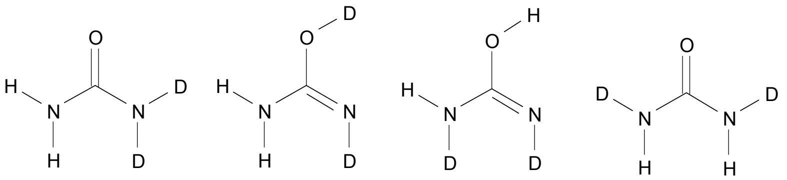
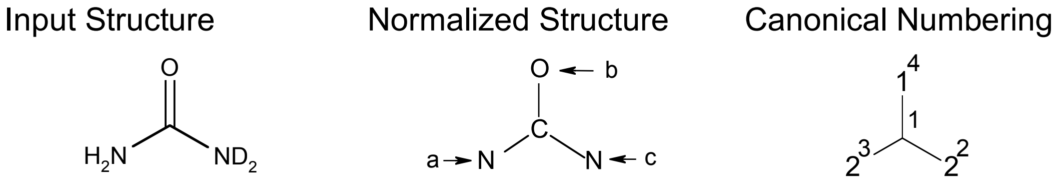


Figure 18. Tautomeric structures of isotopic urea



INChI = CH4N2O/c2-1(3)4/h(H4,2,3,4)/i/hD2

- Moblie H group (a,b,c) has 4 H located at atoms 2, 3, 4:: /h(H4,2,3,4)
- 2 isotopic hydrogen atoms D belong to the whole structure: /i/hD2

Figure 19

Also note that there are, in effect, two possible isotopic layer representations, one that is applied to the Main layer with mobile H and another to Main without mobile H or to fixed-H layer. The optional fixed H layer for the same urea structure is /f/h2-3H2/i2D2; the full identifier is

INChI=1.12Beta/CH4N2O/c2-1(3)4/h(H4,2,3,4)/i/hD2/f/h2-3H2/i2D2

d. Stereochemical Layer (S)

Because of common problems in perceiving and representing stereochemistry, the input information for this layer is the most likely to be incomplete or inaccurate. Further, it is not uncommon for structure collections to contain incomplete or no stereochemical information in their connection tables. Two-dimensional input structures will usually contain adequate information for Z/E stereo perception, though tetrahedral stereo information generally entered using wedge/hatch bonds may be absent or incomplete. An important advantage of the INChI layered format is the isolation of these potential problems and sources of variability in separate layers.

As noted earlier, layer values depend on contents of preceding layers. For example, the value produced for this layer will depend on whether it was derived from a Main layer or Fixed-H layer and whether it belongs to an isotopic layer. Therefore, this type of layer may be present at several locations in an Identifier (Figure 1 and 2).

Two distinct classes of stereochemistry are represented, sp^2 (double bond or Z/E) and sp^3 (tetrahedral). The sp^2 sublayer precedes sp^3 sublayer; as a result, properties of the tetrahedral atom neighbors do not affect sp^2 layer. This enables the proper representation of Z/E stereochemistry in conventional two-dimensional drawings even when stereo bond descriptions are incomplete or absent.

INChI algorithm allows two different systems of wedged and hatched bond interpretation in two-dimensional drawings. By default a “perspective” system is used, where a wedged or hatched bond affects stereochemistry of the two atoms it connects. Another system is invoked by selecting “Narrow end of wedge points to stereocenter” option. In this case the bond affects the stereochemistry of only one atom. Both systems assume that the narrow end of the bond is in the plane of the drawing. Figure 20 illustrates the difference.

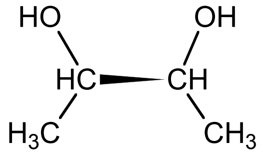
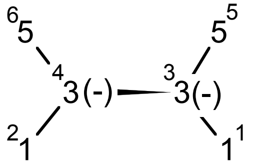
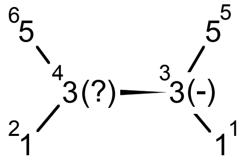
Input structure	Canonical numbering and sp^3 parities	
		
	(a) “perspective” system (default)	(b) Narrow end of wedge points to stereocenter
(a) INChI=1.12Beta/C4H10O2/c1-3(5)4(2)6/h1-2H3,3-6H/t3-,4-/m0/s1 (b) INChI=1.12Beta/C4H10O2/c1-3(5)4(2)6/h1-2H3,3-6H/t3-,4?/m0/s1		

Figure 20. Two systems of wedged/hatched bond interpretaion

On Fig. 20, ‘(?)’ means not-given (‘undefined’) stereo, (-) is a well-defined parity (see next paragraph) calculated by INChI.

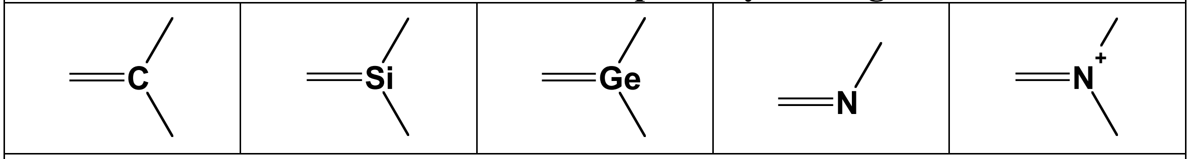
The calculation of stereodescriptors in cases where neighbors to a stereogenic element are not constitutionally identical is straightforward: the parities are calculated from canonical numbers and geometry. Sp^3 parity is ‘+’ if the canonical numbers of neighbors increase clockwise when observed from a hydrogen atom or an atom that has the smallest canonical number; parity of a double bond is '-' if neighbors with greater canonical numbers are located on the same side of the bond.

When constitutionally identical neighbors are present, several equivalent canonical numberings are possible. To resolve this ambiguity (break ties) the algorithm finds a numbering that minimizes a specific internal representation of the stereo layer. In this case it is desirable to determine whether a possibly stereogenic element is in fact stereogenic. To determine this, the following heuristic approach is used. A pair of constitutionally identical neighbors (we call them right and left neighbors) of a possibly stereogenic element is selected. These two neighbors and atoms around them are mapped on their constitutionally equivalent counterparts. After the mapping is complete the canonical numbers are switched between left and right (this leaves the non-stereochemical part of the identifier unchanged). Stereochemical layers corresponding to these two canonical numberings are compared. If the only change occurs to the stereogenic element in question and there are not more than two such constitutionally identical stereogenic elements then these elements are not marked as stereogenic. The origin of this rule is discussed later.

sp^2 stereochemistry

When using input originating from drawings, the perception of formal double bonds capable of supporting Z/E isomerism employs pi-electron information derived from the input connection table along with atom coordinates.

Double bonds treated as possibly stereogenic



Only one of two atoms connected by a possibly stereogenic double bond is shown

Figure 21

In alternating single/double bond cyclic systems, bond-finding algorithms determine whether a formal double bond can exist between each two attached atoms. If such a bond can be drawn between sp^2 hybridized atoms, and the remainder of the pi-electron structure can be completed with alternating bonds, that bond is presumed to be a double bond, hence stereogenic (can support Z/E isomerism).

In some structures, after fixing the location of a double bond, completion of alternating bonds in the remaining structure is not possible. In these cases, one or more 'free electrons' will remain. This commonly occurs for radicals and ions as well as for species with unconventional valences, especially those commonly represented using formal charge pairs (zwitterions). In such cases it INChI assumes that this bond cannot support Z/E stereoisomerism. INChI uses the convention that only formal double, localized bonds that produce a complete alternating pi-network can be stereogenic.

The 3-butene-1-yl radical illustrates an incomplete alternating system – the proposed simplification could not distinguish Z- from E- isomers. It, in effect, presumes that these species rapidly interconvert:

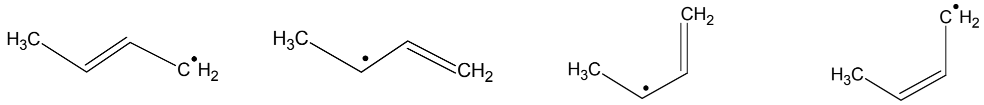


Fig. 22

This approximation allows the representation of stereoisomers that contain these uncertain stereo-bonds along with clearly-defined stereo features, such as:

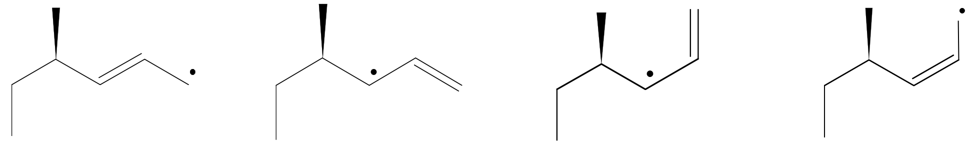


Fig. 23

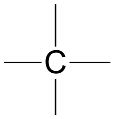
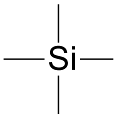
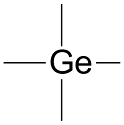
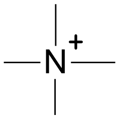
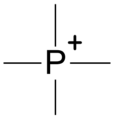
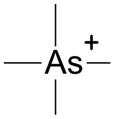
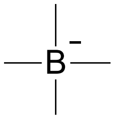
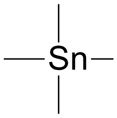
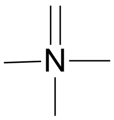
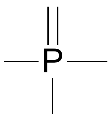
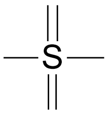
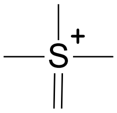
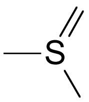
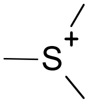
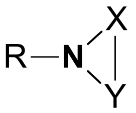
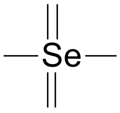
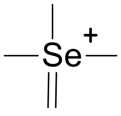
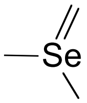
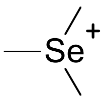
These species would generate the same INChI.

The INChI supports a ‘not-known’ descriptor for marking double bonds where the Z/E isomer is not certain. That is, the stereolayers would be different for Z-2-butene, E-2-butene and 2-butene.

sp^3 *stereochemistry*

Stereochemical descriptors will be processed for tetrahedral atoms such as C, Si and Ge. Currently INChI recognizes only the following atoms as capable of supporting *sp*^{3} stereochemistry:

Table 5. Atoms treated as possibly stereogenic

				
				
				
				
An atom or positive ion **N**, **P**, **As**, **S**, or **Se** is not treated as stereogenic if it has (a) Terminal **H** atom neighbor or (b) At least two terminal neighbors, –**XH**$_m$ and –**XH**$_n$, (n+m>0) connected by any kind of bond, where **X** is **O, S, Se, Te,** or **N**.				

The correctness of a drawing depicting stereogenic elements deserves special consideration. In INChI the following rules are used for two-dimensional drawings:

Table 6. Definition of 2D drawing correctness (4 ligands)

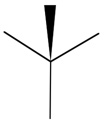
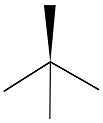
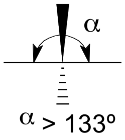
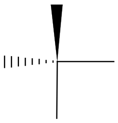
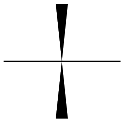
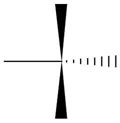
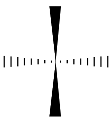
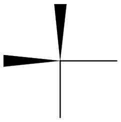
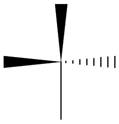
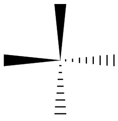
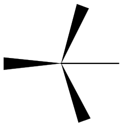
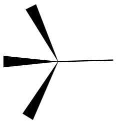
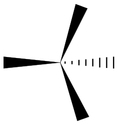
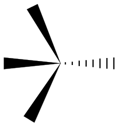
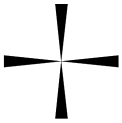
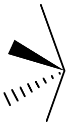
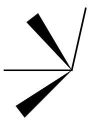
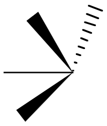
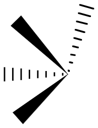
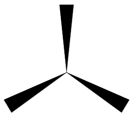
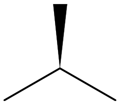
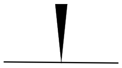
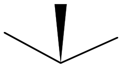
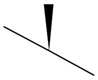
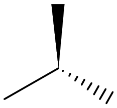
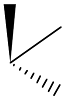
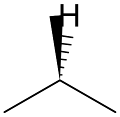
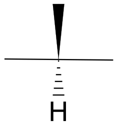
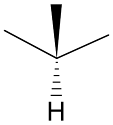
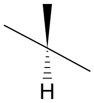
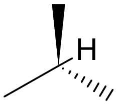
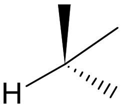
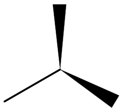
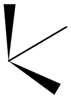
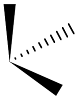
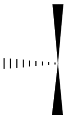
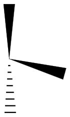
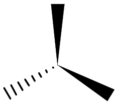
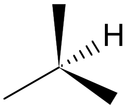
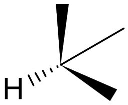
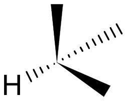
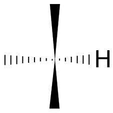
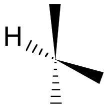
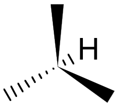
ok	warning	undefined	Ok	ok	ok	ok
						
undefined	undef	undefined	Ok	warning	ok	warning
						
undefined	warn: bonds inside 180° sector (examples)					ok
						

Table 7. Definition of 2D drawing correctness (3 ligands)

	ok	undefined	ok	undefined	Undefined	undefined
input						
interpreted as						
	ok	ok	ok	ok	Undefined	undefined
input						
interpreted as						

The parity of a stereogenic atom is calculated as a volume of an oriented tetrahedron. A wide end of a wedge bond is lifted at an angle of 45° to the plane; a wide end of a hatched bond is lowered at the 45° from the plane. Before the volume is calculated all bonds are reduced to the same length. In addition to the warnings described in the tables above, an additional warning is issued if the central atom is outside of the tetrahedron.

When a complete stereo-description is provided it is straightforward to derive the INChI for a stereoisomer. Problems arise for representation of structures that contain inexact stereochemical information. In these cases stereochemical layers of INChI for different input representations of the same substance will match only if they contain precisely the same sets of inexact information. Moreover, stereochemical layers for inexact structures will not match stereochemical layers for a fully described stereoisomer.

Nevertheless, significant interest was expressed for including partial stereochemical information in the INChI. For this purpose, absolute and unknown stereochemical descriptors can be employed (Figure 24 – left structure is absolute, the C-BrC2H stereocenter in the right structure is unknown):

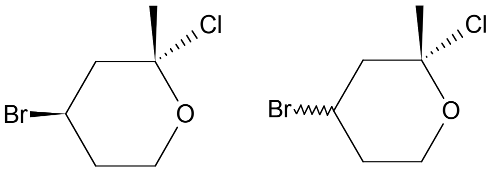


Fig. 24

Representing relative stereochemistry of the whole structure is illustrated for tartaric acid in Fig 25, where it is known that the structure is described by either structure 1 or 2.

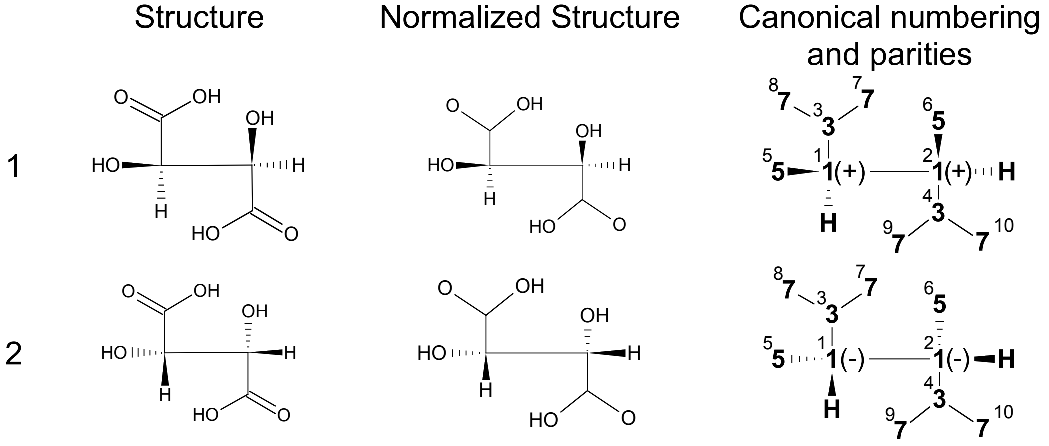


Fig. 25

The identifiers for these structures (case of absolute stereochemistry) are

1	**INChI=1.12Beta/C4H6O6/c5-1(3(7)8)2(6)4(9)10/h1-2H,5-6H,(H,7,8)(H,9,10)/t1-,2-/m1/s1**
2	**INChI=1.12Beta/C4H6O6/c5-1(3(7)8)2(6)4(9)10/h1-2H,5-6H,(H,7,8)(H,9,10)/t1-,2-/m0/s1**

INChI considers both enantiomers and selects the one that has “smaller” identifier. /m0 signifies that the selected one has the exact stereo arrrangement as the input structure; /m1 means that the selected one has inverse arrangement. /s1 means absolute stereochemistry was requested.

To identify relative stereochemistry the /m segment of the identifier is dropped. As the result the identifiers (case of relative stereochemistry) are the same:

1	**INChI=1.12Beta/C4H6O6/c5-1(3(7)8)2(6)4(9)10/h1-2H,5-6H,(H,7,8)(H,9,10)/t1-,2-/s2**
2	**INChI=1.12Beta/C4H6O6/c5-1(3(7)8)2(6)4(9)10/h1-2H,5-6H,(H,7,8)(H,9,10)/t1-,2-/s2**

/s2 means relative stereochemistry was requested.

Allenes belong to sp^3 layer. They precede other sp^3 stereogenic atoms. Cumulenes are treated as double bonds. The following rules are used to recognize allenes and cumulenes:

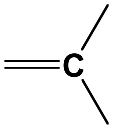
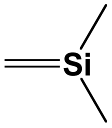
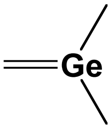
Cumulenes treated as possibly stereogenic		
Terminal atoms		
		
Middle atoms		
$=C=$	$=Si=$	$=Ge=$

Figure 26

Only cumulenes that have 3 double bonds and allenes that have 2 double bonds are treated as possibly stereogenic.

Examples and limitations of the "not more than two constitutionally identical stereogenic elements" rule

Fig. 27 shows an example illustrating three constitutionally identical stereogenic elements. Atom 4 on Fig. 27(b) is the atom in question; upon switching its neighbors only its parity changed, from ^{4}4(-) to ^{4}4(+) (Fig. 27(c)). Therefore this atom is considered stereogenic.

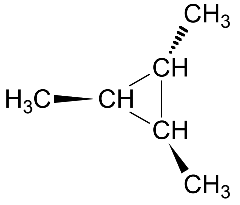
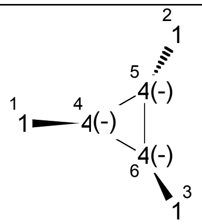
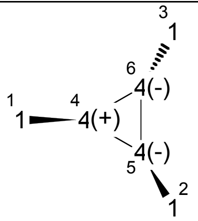
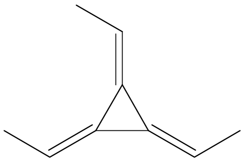
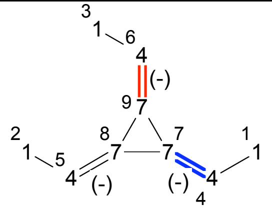
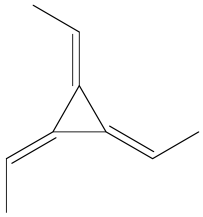
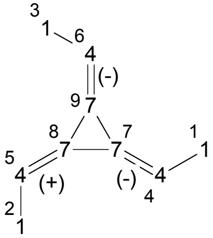
Input structure	Canonical numbering and sp^3 parities	Numbering switched between atoms 1 and 2 and corresponding parities
		
(a)	(b)	(c)
(b) INChI =1.12Beta/C6H12/c1-4-5(2)6(4)3/h1-3H3,4-6H/t4-,5-,6- (c) Switched =1.12Beta/C6H12/c1-4-5(2)6(4)3/h1-3H3,4-6H/t4+,5-,6-		

Figure 27. Switching neighbors of a possibly stereogenic atom

Another example of the same rule applied to stereogenic double bonds is on Fig. 28.

	Input structure	Canonical numbering and double bond parities
(a)		
(b)		

(a) INChI=1.12Beta/C9H12/c1-4-7-8(5-2)9(7)6-3/h1-3H3,4-6H/b7-4-,8-5-,9-6-
(b) INChI=1.12Beta/C9H12/c1-4-7-8(5-2)9(7)6-3/h1-3H3,4-6H/b7-4-,8-5+,9-6-

Figure 28

This (as well as Fig. 27) illustrates the limitation in using parities to mark individual stereogenic atoms or bonds and application of the “not more than two constitutionally identical stereogenic elements” rule.

Consider structure (a) on Fig. 28. Although the parity was assigned by the INChI algorithm to bond 6-9 of the structure (vertical red double bond), the bond definitely does not look stereogenic: the part of the molecule below the bond is symmetric with respect to the double bond axis. The same is true for the bond 7-4 of the same structure (blue double bond). Marking both colored bonds non-stereogenic makes the third double bond, 8-5, also non-stereogenic. As the result, the structure on Fig. 28(a) appears to have no stereochemistry at all and therefore indistinguishable from a structure that really has no data to determine its stereochemistry. The “not more than two” rule forces the retention of the parities of these three double bonds.

However, since the purpose of INChI is to provide an identifier, and not to reveal a true stereochemistry of the submitted structure, this is not a limitation of the INChI: the stereochemical layers of the two stereoisomers on Fig. 28 are different. This rule enables one stereoisomer to be distinguished from another.

e. Canonicalization

Canonicalization is a generating of a set of atom labels that do not depend on how the structure was initially drawn. This is a well-known mathematical problem. It has been discussed both in chemical and mathematical literature. For canonicalization that does not involve stereochemistry an algorithm as described in classic publication [5] was implemented and modified to accommodate for the layered structure of INChI.

The stereochemical canonicalization is based on an exhaustive mapping of non-stereochemical canonical numbering on the structure using previously found constitutional equivalence of the atoms with aim to find the smallest internal representation of the stereochemical layer while keeping other previously found layers unchanged. To avoid combinatorial explosion in case of highly symmetrical structures two approaches are used: (1) elimination of non-stereogenic elements and (2) backtrack search method that prunes the search tree [6].
The canonicalization is performed in stages; each stage adds one more layer to 'minimize' while keeping previously found layers unchanged. This makes splitting the identifier into layers meaningful. Fig. 29 shows the canonicalization flowchart. As it can be seen, the first layer of the Identifier is actually a hydrogenless chemical formula and connections (including bridging hydrogen atoms).

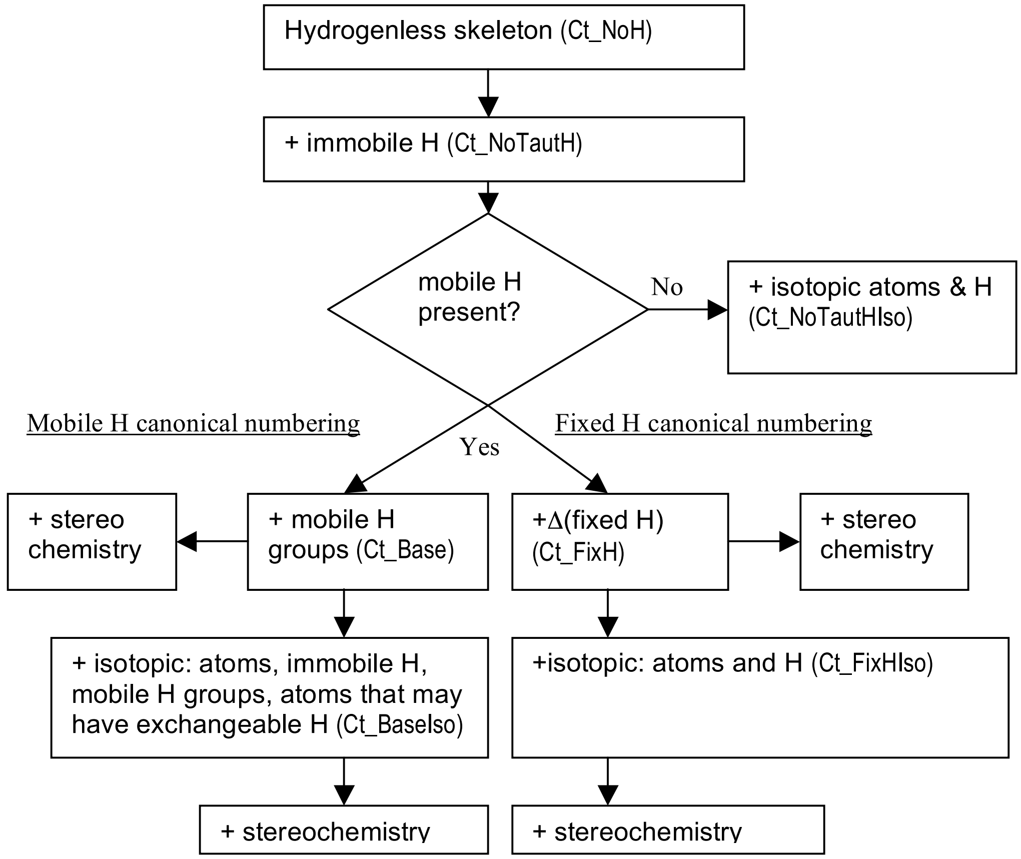

Hydrogenless skeleton (Ct_NoH)
+ immobile H (Ct_NoTautH)
mobile H present?
No
+ isotopic atoms & H (Ct_NoTautHIso)
Mobile H canonical numbering
Fixed H canonical numbering
Yes
+ stereo chemistry
+ mobile H groups (Ct_Base)
+Δ(fixed H) (Ct_FixH)
+ stereo chemistry
+ isotopic: atoms, immobile H, mobile H groups, atoms that may have exchangeable H (Ct_BaseIso)
+isotopic: atoms and H (Ct_FixHIso)
+ stereochemistry
+ stereochemistry

Figure 29. Canonicalization order flowchart (except stereochemistry)

Notes.

- Each set of canonical numberings is a subset of the previous one located up the tree.
- Δ(fixed H) = (number of fixed H on an atom) – (number of H in "mobile-H" structure on the same atom).
- Names in parentheses e.g. (Ct_NoH) are names of data structures in the code.

V. REFERENCES

1. "The Chemical Abstracts Service Chemical Registry System. VII. Tautomerism and Alternating Bonds", by Mockus, J. and Stobaugh, R. E.; J. Chem. Inf. Comput. Sci. 1980, 20, p. 18-22.

2. “Chemical Abstracts Service Chemical Registry System. 13. Enhanced Handling of Stereochemistry”, Blackwood, J. E., Blower, P. E., Jr., Layten, S. W., Lillie, D. H., Lipkus, A. H., Peer, J. P., Qian, C., Staggenborg, L. M., Watson, C. E., J. Chem. Inf. Comput. Sci., 1991, vol. 31, p. 204-212.

3. “Fun with Chirality” Dr. Ron Beavon, http://www.rod.beavon.clara.net/chiralit.htm, 2001.

4. W. Kocay, D. Stone, "An Algorithm for Balanced Flows", The Journal of Combinatorial Mathematics and Combinatorial Computing, vol. 19 (1995) pp. 3-31.

5. B. D. McKay, "Practical Graph Isomorphism", Congressus Numerantium, Vol. 30 (1981), pp. 45 – 87.

6. G. Butler, “Fundamental Algorithms for Permutational Groups”, Berlin ; New York: Springer-Verlag, 1991 (Series: Lecture Notes in Computer Science, 559), Chapter 11.

VI. BIBLIOGRAPHY

Canonical (Unique) Numbering Algorithms:

The Generation of a Unique Machine Description for Chemical Structures---A Technique Developed at Chemical Abstracts Service
Morgan, H.L.
Journal of Chemical Documentation
Vol. 5, pp. 107-113, **1965**

Stereochemically Unique Naming Algorithm
Wipke, W.T.; Dyott, T.M.
Journal of the American Chemical Society
Vol. 96, No. 15, pp. 4834-4842, **1974**

Canonical Numbering and Constitutional Symmetry
Jochum, C.; Gasteiger, J.
Journal of Chemical Information and Computer Sciences
Vol. 17, No. 2, pp. 113-117, **1977**

Computer Perception of Topological Symmetry
Shelley, C.A.; Munk, M.E.
Journal of Chemical Information and Computer Sciences
Vol. 17, No. 2, pp. 110-113, **1977**

Computer Perception of Topological Symmetry
Shelley, C.A.; Munk, M.E.
Journal of Chemical Information and Computer Sciences
Vol. 17, No, 2, pp. 110-113, **1977**

On Canonical Numbering of Atoms in a Molecule and Graph Isomorphism
Randic, M.
Journal of Chemical Information and Computer Sciences
Vol. 17, No. 3, pp. 171-180, **1977**

Algorithms for Unique and Unambiguous Coding and Symmetry Perception of Molecular Structure Diagram. I. Vector Functions for Automorphism Partitioning
Uchino, M.
Journal of Chemical Information and Computer Sciences
Vol. 20, pp. 116-120, **1980**

Computer Perception of Topological Symmetry via Canonical Numbering of Atoms
Randic, M.; Brissey, G.M.; Wilkins, C.L.
Journal of Chemical Information and Computer Sciences
Vol. 21, pp. 52-59, **1981**

Computer Perception of Topological Symmetry via Canonical Numbering of Atoms
Randic, M.; Brissey, G.M.; Wilkins, C.L.
Journal of Chemical Information and Computer Sciences
Vol. 21, pp. 52-59, **1981**

Unique Numbering and Cataloguing of Molecular Structures
Hendrickson, J.B.; Toczko, A.G.
Journal of Chemical Information and Computer Sciences
Vol. 23, pp. 171-177, **1983**

Canonical Numbering, Stereochemical Descriptors, and Unique Linear Notations for Polyhedral Clusters
Herndon, W.C.; Leonard, J.E.
Inorganic Chemistry
Vol. 22, pp. 554-557, **1983**

Unique Description of Chemical Structures Based on Hierarchically Ordered Extended Connectivities (HOC Procedures). I. Algorithms for Finding Graph Orbits and Canonical Numbering of Atoms
Balaban, A.T.; Mekenyan, O.; Bonchev, D.
Journal of Computational Chemistry
Vol. 6, No. 6, pp. 538-551, **1985**

Unique Description of Chemical Structures Based on Hierarchically Ordered Extended Connectivities (HOC Procedures). III. Topological, Chemical, and Stereochemical Coding of Molecular Structure
Balaban, A.T.; Mekenyan, O.; Bonchev, D.
Journal of Computational Chemistry
Vol. 6, No. 6, pp. 562-569, **1985**

SMILES. 2. Algorithm for Generation of Unique SMILES Notation
Weininger, D.; Weininger, A.; Weininger, J.L.
Journal of Chemical Information and Computer Sciences
Vol. 29, pp. 97-101, **1989**

Canonical Indexing and Constructive Enumeration of Molecular Graphs
Kvasnicka, V.; Pospichal, J.
Journal of Chemical Information and Computer Sciences
Vol. 30, pp. 99-105, **1990**

Computer Perception of Constitutional (Topological) Symmetry: TOPSYM, a Fast Algorithm for Partitioning Atoms and Pairwise Relations Among Atoms into Equivalence Classes
Rucker, G.; Rucker, C.

Journal of Chemical Information and Computer Sciences
Vol. 30, pp. 187-191, **1990**

On Using the Adjacency Matrix Power Method for Perception of Symmetry and for Isomorphism Testing of Highly Intricate Graphs
Rucker, G.; Rucker, C.
Journal of Chemical Information and Computer Sciences
Vol. 31, pp. 123-126, **1991**

ESSESA: An Expert System for Structure Elucidation from Spectra. 4. Canonical Representation of Structures
Huixiao, H.; Xinquan, X.
Journal of Chemical Information and Computer Sciences
Vol. 34, pp. 730-734, **1994**

A Computer-Oriented Linear Canonical Notational System for the Representation of Organic Structures with Stereochemistry
Agarwal, K.K.; Gelernter, H.L.
Journal of Chemical Information and Computer Sciences
Vol. 34, pp. 463-479, **1994**

Detection of Constitutionally Equivalent Sites from a Connection Table
Fan, B.T.; Barbu, A.; Panaye, A.; Doucet, J.-P.
Journal of Chemical Information and Computer Sciences
Vol. 36, pp. 654-659, **1996**

Isomorphism, Automorphism Partitioning, and Canonical Labeling Can Be Solved in Polynomial-Time for Molecular Graphs
Faulon, J.-L.
Journal of Chemical Information and Computer Sciences
Vol. 38, pp. 432-444, **1998**

Isomorphism, Automorphism Partitioning, and Canonical Labeling Can Be Solved in Polynomial-Time for Molecular Graphs
Faulon, J.-L.
Journal of Chemical Information and Computer Sciences
Vol. 38, pp. 432-444, **1998**

Conversion of Unique Names to Fixed Length Name (Hash)

Hash Functions for Rapid Storage and Retrieval of Chemical Structures
Wipke, W.T.; Krishnan, S.; Ouchi, G.I.
Journal of Chemical Information and Computer Sciences
Vol. 18, No. 1, pp. 32-37, **1978**

Structure Searching in Chemical Databases by Direct Lookup Methods
Christie, B.D.; Leland, B.A.; Nourse, J.G.
Journal of Chemical Information and Computer Sciences
Vol. 33, pp. 545-547, **1993**

Hash Codes for the Identification and Classification of Molecular Structure Elements
Ihlenfeldt, W.D.; Gasteiger, J.
Journal of Computational Chemistry
Vol. 15, No. 8, pp. 793-813, **1994**

Representation of Chemical Structures (relevant to naming)

Simulation and Evaluation of Chemical Synthesis. Computer Representation and Manipulation of Stereochemistry
Wipke, W.T.; Dyott, T.M.
Journal of the American Chemical Society
Vol. 96, No. 15, pp. 4825-4834, **1974**

An Efficient Design for Chemical Structure Searching. III. The Coding of Resonating and Tautomeric Forms
Feldman, A.
Journal of Chemical Information and Computer Sciences
Vol. 17, No. 4, pp. 220-223, **1977**

A Representation of π Systems for Efficient Computer Manipulation
Gasteiger, J.
Journal of Chemical Information and Computer Sciences
Vol. 19, No. 2, pp. 111-115, **1979**

The Chemical Abstracts Service Chemical Registry System. VII. Tautomerism and Alternating Bonds
Mockus, J.; Stobauch, R.J.
Journal of Chemical Information and Computer Sciences
Vol. 20, No. 1, pp. 18-22, **1980**

Computer-Assisted Mechanistic Evaluation of Organic Reactions. 2. Perception of Rings, Aromaticity, and Tautomers
Roos-Kozel, B.L.; Jorgenson, W.L.
Journal of Chemical Information and Computer Sciences
Vol. 21, pp. 101-111, **1991**

Chemical Abstracts Service Chemical Registry System. 13. Enhanced Handling of Stereochemistry
Blackwood, J.E.; Blower, pp.E., Jr.; Layten, S.W.; Lillie, D.H.; Lipkus, A.H.; Peer, J.P.; Qian, C.; Staggenborg, L.M.; Watson, C.E.
Journal of Chemical Information and Computer Sciences
Vol. 21, pp. 204-212, **1991**

Stereochemistry and Sequence Rules A Proposal for Modification of Cahn-Ingold-Prelog System
Perdih, M.; Razinger, M.
Tetrahedreon: Asymmetry
Vol. 5, No. 5, pp. 835-861, **1994**

Computerized Stereochemistry: Coding and Naming Configurational Stereoisomers

Razinger, M.; Perdih, M.
Journal of Chemical Information and Computer Sciences
Vol. 34, pp. 290-296, **1994**

Implementation of the Cahn-Ingold-Prelog System for Stereochemical Perception in the LHASA Program
Mata, pp.; Lobo, A.M.; Marshall, C.; Johnson, A.P.
Journal of Chemical Information and Computer Sciences
Vol. 34, pp. 491-504, **1994**

Chemical eXchange format (CXF)
Chemical Abstracts Service
Version 1.0
September, **1994**

Overcoming the Limitations of a Connection Table Description: A Universal Representation of Chemical Species
Bauerschmidt, S.; Gasteiger, J.
Journal of Chemical Information and Computer Sciences
Vol. 37, pp. 705-714, **1997**

Fundamental Aspects of Unique Naming Methods

Erroneous Claims Concerning the Perception of Topological Symmetry
Carhart, R.E.
Journal of Chemical Information and Computer Sciences
Vol. 18, pp. 108-110, **1978**

Conformation Specification of Chemical Structures in Computer Programs
Fella, A.L.; Nourse, J.G.; Smith, D.H.
Journal of Chemical Information and Computer Sciences
Vol. 23, pp. 43-47, **1983**

Chemical Abstracts Service Chemical Registry System. 13. Enhanced Handling of Stereochemistry
Blackwood, J.E; Blower, Jr., pp.E.; Layten, S.W.; Lillie, D.H.; Lipkus, A.H.; Peer, J.P.; Qian, C.; Staggenborg, L. M.; Watson, C. E.
Journal of Chemical Information and Computer Sciences
Vol. 31, No. 2, pp. 204-212, **1991**

Counts of All Walks as Atomic and Molecular Descriptors
Rucker, G.; Rucker, C.
Journal of Chemical Information and Computer Sciences
Vol. 33, pp. 683-695, **1993**

Symmetry of Chemical Structures: A Novel Method of Graph Automorphism Group Determination
Bohanec, S.; Perdih, M.
Journal of Chemical Information and Computer Sciences
Vol. 33, pp. 719-726, **1993**

Mathematical Relation between Extended Connectivity and Eigenvector Coefficients
Rucker, G.; Rucker, C.
Journal of Chemical Information and Computer Sciences
Vol. 34, pp. 534-538, **1994**

Computational Techniques for the Automorphism Groups of Graphs
Balasubramanian, K.
Journal of Chemical Information and Computer Sciences
Vol. 34, pp. 621-626, **1994**

Computer Generation of Automorphism Groups of Weighted Graphs
Balasubramanian, K.
Journal of Chemical Information and Computer Sciences
Vol. 34, pp. 1146-1150, **1994**

Algorithm for Computer Perception of Topological Symmetry
Hu, C.-Y.; Xu, L.
Analytica Chimica Acta
Vol. 295, pp. 127-134, **1994**

Determination of Topological Equivalence Classes of Atoms and Bonds in C20-C68 Fullerenes Using a New Prolog Coding Program
Laidboeur, T.; Cabrol-Bass, D.; Ivanciuc, O.
Journal of Chemical Information and Computer Sciences
Vol. 36, pp. 811-821, **1996**

Appendix 1. INChI Standard Valences.

(from array ElData[ ] located in source code file util.c)

Elements, average atomic mass, metals, standard valences, and conditions for implicit H addition.

ele-	Ave.	metal**	Add	Standard valence of atoms for charges from -2 to +2				
ment	at.		H	-2	-1	0	+1	+2
	mass							
H	1	-	Yes	-	-	1	-	-
He	4	-	-	-	-	-	-	-
Li	7	M1	Yes	-	-	1	-	-
Be	9	M1	Yes	-	-	2	1	-
B	11	-	Yes	3	4	3	2	1
C	12	-	Yes	2	3	4	3	2
N	14	-	Yes*	1	2	3 5*	4	3
O	16	-	Yes	-	1	2	3 5	4
F	19	-	Yes	-	-	1	2	3 5
Ne	20	-	-	-	-	-	-	-
Na	23	M1	Yes	-	-	1	-	-
Mg	24	M1	Yes	-	-	2	1	-
Al	27	M1	Yes	3 5	4	3	2	1
Si	28	-	Yes	2	3 5	4	3	2
P	31	-	Yes	1 3 5 7	2 4 6	3 5	4	3
S	32	-	Yes*	-	1 3 5 7	2 4* 6	3 5	4
Cl	35	-	Yes	-	-	1 3 5 7	2 4 6	3 5
Ar	40	-	-	-	-	-	-	-
K	39	M1	Yes	-	-	1	-	-
Ca	40	M1	Yes	-	-	2	1	-
Sc	45	M1	-	-	-	3	-	-
Ti	48	M1	-	-	-	3 4	-	-
V	51	M1	-	-	-	2 3 4 5	-	-
Cr	52	M1	-	-	-	2 3 6	-	-
Mn	55	M2	-	-	-	2 3 4 6	-	-
Fe	56	M2	-	-	-	2 3 4 6	-	-
Co	59	M2	-	-	-	2 3	-	-
Ni	59	M2	-	-	-	2 3	-	-
Cu	64	M1	-	-	-	1 2	-	-
Zn	65	M1	-	-	-	2	-	-
Ga	70	M1	Yes	3 5	4	3	-	1
Ge	73	-	Yes	2 4 6	3 5	4	3	-
As	75	-	Yes	1 3 5 7	2 4 6	3 5	4	3
Se	79	-	Yes	-	1 3 5 7	2 4 6	3 5	4
Br	80	-	Yes	-	-	1 3 5 7	2 4 6	3 5
Kr	84	-	-	-	-	-	-	-
Rb	85	M1	Yes	-	-	1	-	-
Sr	88	M1	Yes	-	-	2	1	-
Y	89	M1	-	-	-	3	-	-
Zr	91	M1	-	-	-	4	-	-
Nb	93	M1	-	-	-	3 5	-	-
Mo	96	M1	-	-	-	3 4 5 6	-	-
Tc	98	M1	-	-	-	7	-	-
Ru	101	M1	-	-	-	2 3 4 6	-	-
Rh	103	M1	-	-	-	2 3 4	-	-

Pd	106	M1	-	-	-	2 4	-	-
Ag	108	M1	-	-	-	1	-	-
Cd	112	M1	-	-	-	2	-	-
In	115	M1	Yes	3 5	2 4	3	-	1
Sn	119	M2	Yes	2 4 6	3 5	2 4	3	-
Sb	122	M1	Yes	1 3 5 7	2 4 6	3 5	2 4	3
Te	128	-	Yes	-	1 3 5 7	2 4 6	3 5	2 4
I	127	-	Yes	-	-	1 3 5 7	2 4 6	3 5
Xe	131	-	-	-	-	-	-	-
Cs	133	M1	Yes	-	-	1	-	-
Ba	137	M1	Yes	-	-	2	1	-
La	139	M1	-	-	-	3	-	-
Ce	140	M2	-	-	-	3 4	-	-
Pr	141	M2	-	-	-	3 4	-	-
Nd	144	M1	-	-	-	3	-	-
Pm	145	M1	-	-	-	3	-	-
Sm	150	M2	-	-	-	2 3	-	-
Eu	152	M2	-	-	-	2 3	-	-
Gd	157	M1	-	-	-	3	-	-
Tb	159	M2	-	-	-	3 4	-	-
Dy	163	M1	-	-	-	3	-	-
Ho	165	M1	-	-	-	3	-	-
Er	167	M1	-	-	-	3	-	-
Tm	169	M2	-	-	-	2 3	-	-
Yb	173	M2	-	-	-	2 3	-	-
Lu	175	M1	-	-	-	3	-	-
Hf	178	M1	-	-	-	4	-	-
Ta	181	M1	-	-	-	5	-	-
W	184	M2	-	-	-	3 4 5 6	-	-
Re	186	M2	-	-	-	2 4 6 7	-	-
Os	190	M2	-	-	-	2 3 4 6	-	-
Ir	192	M2	-	-	-	2 3 4 6	-	-
Pt	195	M2	-	-	-	2 4	-	-
Au	197	M1	-	-	-	1 3	-	-
Hg	201	M2	-	-	-	1 2	-	-
Tl	204	M2	Yes	3 5	2 4	1 3	-	-
Pb	207	M2	Yes	2 4 6	3 5	2 4	3	-
Bi	209	M1	Yes	1 3 5 7	2 4 6	3 5	2 4	3
Po	209	M2	Yes	-	1 3 5 7	2 4 6	3 5	2 4
At	210	-	Yes	-	-	1 3 5 7	2 4 6	3 5
Rn	222	-	-	-	-	-	-	-
Fr	223	M1	Yes	-	-	1	-	-
Ra	226	M1	Yes	-	-	2	1	-
Ac	227	M1	-	-	-	3	-	-
Th	232	M2	-	-	-	3 4	-	-
Pa	231	M2	-	-	-	3 4 5	-	-
U	238	M2	-	-	-	3 4 5 6	-	-
Np	237	M2	-	-	-	3 4 5 6	-	-
Pu	244	M2	-	-	-	3 4 5 6	-	-
Am	243	M2	-	-	-	3 4 5 6	-	-
Cm	247	M1	-	-	-	3	-	-
Bk	247	M1	-	-	-	3 4	-	-
Cf	251	M1	-	-	-	3	-	-
Es	252	M1	-	-	-	3	-	-
Fm	257	M1	-	-	-	3	-	-
Md	258	M1	-	-	-	3	-	-

No	259	M1	-	-	-	1	-	-
Lr	260	M1	-	-	-	1	-	-
Rf	261	M1	-	-	-	1	-	-

* Do not add H to reach valences marked with *

** M1 – use only lowest valence for salt disconnection;
M2 – use also the 2nd lowes valence.

Appendix 2. Abbreviations

In case of similar or identical components of a multicomponent compound the segments of a layer related to different components may be identical. In such cases the segment is not repeated in the identifier; instead it is preceded by a multiplier in form NUMBER in a chemical formula (for example, 2H2O for two molecules H_2O) and NUMBER* in the rest of the identifier (for example /h2*1H2, where /h1H2 is a hydrogen layer for H_2O, /h2*1H2 is a hydrogen layer for $2H_2O$).

In some cases a layer can appear at more than one place in the INChI output. For example, stereochemical layer in the Main and Fixed-H layers may be identical. When the contents of a layer for a component have appeared in an earlier layer, an abbreviation is used instead of the second instance. All possible abbreviations are given in this Appendix.

Different letters are used to refer to different locations of the first instance of the same layer information:

m – item in the first section of the Identifier, but not in the isotopic segment

M – item in the isotopic part of the first section

n – item in the fixed-H section, but not in the isotopic segment

N – item in the isotopic part of fixed-H section

i – prefix to **m, M, n**, or **N** – indicates that sp^3-stereo has been inverted.

Repetitions of these abbreviations further abbreviated with multipliers, for example “m;m;m” is replaced with “3m”,

Abbreviations used in the Identifier

Abbreviated item	Is identical to	Abbreviation
	stereo:dbond and **stereo:sp3**	
Isotopic:stereo	stereo	m*
fixed-H:stereo	stereo	m
fixed-H:isotopic:stereo	stereo	m
fixed-H:isotopic:stereo	isotopic:stereo	M
fixed-H:isotopic:stereo	fixed-H:stereo	n*
	isotopic:atoms	
fixed-H:isotopic:atoms	isotopic:atoms	m
	charge	
fixed-H:charge	charge	m

*the isotopic stereo is omitted if it is exactly same as non-isotopic stereo for all components.

Abbreviations used in the auxiliary information section

In the Fixed-H section of the Auxiliary Information the original_atom_numbers for the components are in the same order as in the Main section of the Identifier even if a transposition (/o) is present. Other Fixed-H items are subject to the transposition.

Word “orig_at_nums” is used instead of “original_atom_numbers”

Prefix “Aux” distinguishes items that belong to AuxInfo from those belonging to the Identifier

Inv(…) means formal replacing sp^3 parity “+” with “-“ and vice versa

Abbreviated item	Is identical to	Abbre-viation
Aux:original_atom_numbers		
Aux:isotopic:orig_at_nums	Aux:orig_at_nums	m
Aux:fixed-H:orig_at_nums	Aux:orig_at_nums	m
Aux:fixed-H:isotopic:orig_at_nums	Aux:orig_at_nums	m
Aux:fixed-H:isotopic:orig_at_nums	Aux:fixed-H:orig_at_nums	n
Aux:fixed-H:isotopic:orig_at_nums	Aux:isotopic:orig_at_nums	M
Aux:atom_equivalence or Aux:group_equivalence		
Aux:isotopic:atom_equivalence or Aux:isotopic:group_equivalence	Aux:atom_equivalence or Aux:group_equivalence	m
Aux:fixed-H:atom_equivalence or Aux:fixed-H:group_equivalence	Aux:atom_equivalence or Aux:group_equivalence	m
Aux:fixed-H:isotopic:atom_equivalence or Aux:fixed-H:isotopic:group_equivalence	Aux:atom_equivalence or Aux:group_equivalence	m
Aux:fixed-H:isotopic:atom_equivalence or Aux:fixed-H:isotopic:group_equivalence	Aux:fixed-H:atom_equivalence or Aux:fixed-H:group_equivalence	n
Aux:fixed-H:isotopic:atom_equivalence or Aux:fixed-H:isotopic:group_equivalence	Aux:isotopic:atom_equivalence or Aux:isotopic:group_equivalence	M
Aux:abs_stereo_inverted:sp3*		
Aux:abs_stereo_inverted:sp3	Inv.(stereo:sp3)	im
Aux:isotopic:abs_stereo_inverted:sp3	Aux:abs_stereo_inverted:sp3	m
Aux:isotopic:abs_stereo_inverted:sp3	Inv.(stereo:sp3)	im
Aux:isotopic:abs_stereo_inverted:sp3	Inv.(isotopic:stereo:sp3)	iM
Aux:fixed-H:abs_stereo_inverted:sp3	Aux:abs_stereo_inverted:sp3	m
Aux:fixed-H:abs_stereo_inverted:sp3	Inv.(stereo:sp3)	im
Aux:fixed-H:abs_stereo_inverted:sp3	Inv.(fixed-H:stereo:sp3)	in
Aux:isotopic:fixed-H:abs_stereo_inverted:sp3	Aux:abs_stereo_inverted:sp3	m
Aux:isotopic:fixed-H:abs_stereo_inverted:sp3	Aux:fixed-H:abs_stereo_inverted:sp3	n
Aux:isotopic:fixed-H:abs_stereo_inverted:sp3	Aux:isotopic:abs_stereo_inverted:sp3	M
Aux:isotopic:fixed-H:abs_stereo_inverted:sp3	Inv.(stereo:sp3)	im
Aux:isotopic:fixed-H:abs_stereo_inverted:sp3	Inv.(isotopic:stereo:sp3)	iM
Aux:isotopic:fixed-H:abs_stereo_inverted:sp3	Inv.(fixed-H:stereo:sp3)	in
Aux:isotopic:fixed-H:abs_stereo_inverted:sp3	Inv.(fixed-H:isotopic:stereo:sp3)	iN
Aux:abs_stereo_inverted:original_atom_numbers		
Aux:abs_stereo_inverted:orig_at_nums	Aux:orig_at_nums	m
Aux:isotopic:abs_stereo_inverted:orig_at_nums	Aux:orig_at_nums	m
Aux:isotopic:abs_stereo_inverted:orig_at_nums	Aux:isotopic:atom.orig	M
Aux:isotopic:abs_stereo_inverted:orig_at_nums	Aux:abs_stereo_inverted:orig_at_nums	im
Aux:fixed-H:abs_stereo_inverted:orig_at_nums	Aux:orig_at_nums	m
Aux:fixed-H:abs_stereo_inverted:orig_at_nums	Aux:fixed-H:orig_at_nums	n
Aux:fixed-H:abs_stereo_inverted:orig_at_nums	Aux:abs_stereo_inverted:orig_at_nums	im
Aux:fixed-H:isotopic:abs_stereo_inverted:orig_at_nums	Aux:orig_at_nums	m
Aux:fixed-H:isotopic:abs_stereo_inverted:orig_at_nums	Aux:fixed-H:orig_at_nums	n
Aux:fixed-H:isotopic:abs_stereo_inverted:orig_at_nums	Aux:isotopic:atom.orig	M
Aux:fixed-H:isotopic:abs_stereo_inverted:orig_at_nums	Aux:fixed-H:isotopic:atom.orig	N
Aux:fixed-H:isotopic:abs_stereo_inverted:orig_at_nums	Aux:abs_stereo_inverted:orig_at_nums	im
Aux:fixed-H:isotopic:abs_stereo_inverted:orig_at_nums	Aux:fixed-H:abs_stereo_inverted:orig_at_nums	in
Aux:fixed-H:isotopic:abs_stereo_inverted:orig_at_nums	Aux:isotopic:abs_stereo_inverted:orig_at_nums	iM

Notes:
*Stereo-Inv. identical to non-inverted or in case of relative or racemic is omitted

Appendix 3. Extracting Layers from INChI

The Identifier is a string separated into parts by two-character delimiters ‘/?’ where ‘?’ is a lowercase letter.

Fig. A3-1 represents a hierarchical structure of the Identifier:

```
    1------------------------------------1  1------------------------------------1
    |          2-----2  2---------------2|  |          2-----2  2---------------2|
    |          |     |  |        3----3 ||  |          |     |  |        3----3 ||
    |          |     |  |        |    | ||  |          |     |  |        |    | ||
VER/(chqpbtms/i(hbtms)/f(hqbtms/i(btms)o))/r(chqpbtms/i(hbtms)/f(hqbtms/i(btms)o))
   |         |        |        |      | |||          |        |        |      | ||
   | non-iso-| iso-   |non-iso-| iso- | |||  non-iso-| iso-   |non-iso-| iso- | ||
   | topic   | topic  |topic   | topic| |||  topic   | topic  |topic   | topic| ||
   |         |        |        |      | |||          |        |        |      | ||
   +---------+--------+--------+------+-+|+----------+--------+--------+------+-+|
   |                  |                 |||                   |                 ||
   | main: may have   | fixed-H: may be |||  main: may have   | fixed-H: may be ||
   | mobile H         | present only if |||  mobile H         | present only if ||
   |                  | main has mobile |||                   | main has mobile ||
   |                  | H               |||                   | H               ||
   +------------------+-----------------+|+-------------------+-----------------+|
   |                                     ||  identifier of the reconnected       |
   | may be disconnected if metal        ||  structure; may be present only if   |
   | atom(s) present                     ||  metal atom(s) present               |
   |                                     ||                                      |
   +-------------------------------------++--------------------------------------+
```

Figure A3-1. Hierarchical structure of the Identifier

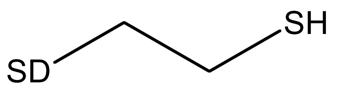
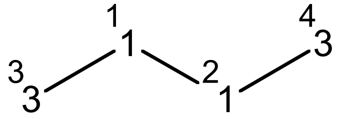
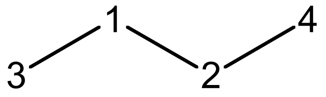
The Identifier starts with the version string (VER) followed by a slash ‘/’. The added for the sake of the explanation parentheses logically separate sections of the Identifier. Each section starts with two or three character combination, the last character being opening parenthesis:

Combination:	**/(**	**/i(**	**/f(**	**/r**
Starts layer:	Main	Isotopic	Fixed-H	Reconnected main

The matching closing parenthesis ends the section. Matching pairs of parentheses are shown by the lines above the Identifier string. The contents of the sections are explained below the Identifier string. Characters in the Identifier string that are not immediately preceded by a slash represent other possibly present items inside the section (see Fig. 2, Layers of the identifier). Slashes before them were omitted to avoid making the picture more obscure.

The serialization algorithm outputs the Identifier in such a way that if a section is not empty then its starting combination is always present. This makes parentheses unneeded in the output therefore they are not present in the Identifier.

As an example consider the identifier of a structure on Fig. A3-2 that includes Fixed-H layer.

Input structure	Canonical numbering (mobile H)	Canonical numbering (fixed H)
		
INChI=1.12Beta/C2H6S2/c3-1-2-4/h1-2H2,3-4H/i/hD/f/i3D		
Figure A3-2		

The deuterium atom in the Main isotopic layer is represented as
/i/hD
which means it is considered exchangeable; therefore its position in the Main layer is not defined. Even though the string that should immediately follow /i (isotopic:atoms, see Fig. 2) is not present, the "/i" itself is present in the Identifier to signify the isotopic layer.
In the Fixed-H layer
/f/i3D
its non-isotopic part that immediately follows /f is not present. However, "/f" is present to signify the fixed-H layer.

An algorithm to parse the Identifier may be described in the following way:

1) Find the first slash. The slash is preceded by the version and followed by a string (call it S) that contains all other layers of the identifier.
2) Search for "/r" in S. If "/r" is found then copy preceding "/r" substring to P[1] and the following "/r" string to P[2] else copy S to P[1]
(P[1] represents the whole identifier or an identifier of a disconnected structure; P[2] if not empty represents an identifier of a "reconnected" structure".)
3) Search for "/f" in each non-empty P[i]. If "/f" was found then copy the preceding string to Q[i][1] and the following string to Q[i][2] else copy P[i] to Q[i][1]
(Q[i][1] represents the Main layer; Q[i][2] represents fixed-H layer)
4) Search for "/i" in each non-empty Q[i][j]. If "/i" was found then copy the preceding string into R[i][j][1] and the following string into R[i][j][2] else copy Q[i][j] to R[i][j][1]
(R[i][j][1] represents the non-isotopic part of the layer; R[i][j][2] represents the isotopic layer)

At the end, non-empty strings R[i][j][k] (i, j, k = 1 or 2) contain:
i = 1: The identifier or the identifier of a disconnected structure
i = 2: The identifier of the "reconnected" structure
j = 1: The main layer
j = 2: The fixed-H layer

k = 1: The non-isotopic part
k = 2: The isotopic part of the layer

In case of multicomponent compound the parts of the identifier related to components are separated by semicolons “;” except for the chemical formula which is dot-disconnected. The order of the components within the segments of the Main layer is same; the fixed-H layer may have a different order of the components. In this case a transposition segment (/o) is present. For example, transposition (1,2,3) means that component #1 in the Main layer is component #2 in Fixed-H layer, component #2 in the Main layer is component #3 in Fixed-H layer and component #3 in the Main layer is component #1 in the Fixed-H layer. A simple example of a compound that exhibits a transposition is on Fig. A3-3 later.

To extract identifiers for individual components

- parse the identifier and obtain array of strings R as explained above
- split each of those strings into segments using “/?” as separators and identify the separators (see Fig. 2)
- split each segment by locating dots or semicolons into parts related to individual components and expanded them in case of multipliers and/or abbreviations describer in Appendix 2;
- transpose components in fixed-H part according to transposition (/o) if it is present
- pick the first entries (corresponding to the first component) and merge them together using previously found “/?” separators; add “version/” to the beginning of the string. The string is an identifier for the first component
- repeat for all other components

It should be noted that the number of components in fixed-H layer may be greater than the number of components in the Main layer. The difference is the number of free protons (H+) in the input structure.

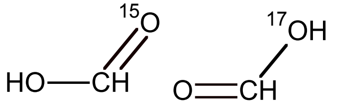
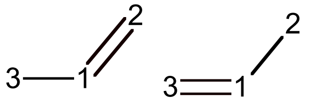
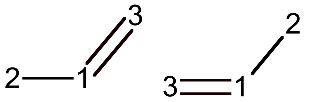
<table>
<tr><td>Input structure</td><td>Canonical numbering (mobile H, isotopic)</td><td>Canonical numbering (fixed-H, isotopic)</td></tr>
<tr><td>
</td><td>
</td><td>
</td></tr>
<tr><td>(a)</td><td>(b)</td><td>(c)</td></tr>
<tr><td colspan="3">INChI=1.12Beta/2CH2O2/c2*2-1-3/h2*1H,(H,2,3)/i2+1;2-1/f/h2*2H/i3-1;2+1/o(1,2)</td></tr>
<tr><td colspan="3"><b>Figure A3-3.</b> Example of a component transposition</td></tr>
</table>

Fig. A3-3 also shows use of multipliers (2 in the chemical formula and 2* in the rest of the identifier) that are employed to avoid repetitions of identical strings representing different components. The reason for the transposition is that in (b) the canonical numbering and the order are effectively determined by the isotopic composition (the mobile H is equally distributed between atoms O) while in (c) they are determined by the location of the fixed mobile H.

Appendix 4. Comparing INChI Representations For Finding Identical Compounds

If two INChIs are the same, then it is safe to assume that the compounds (structures) that they represent are the same. For many structures and for data collections where structures are entered using a uniform procedure, this should be sufficient for identification. However, the layered structured of INChI permits the representation of some compounds at different levels of detail or completeness. If, for example, one INChI is completely contained in another, then the second may be viewed as a more detailed representation of the first (for example, Z-but-2-ene may be viewed as a more detailed representation than but-2-ene). Or, for example, if one set of INChIs were derived from a collection with no stereo information and another contains complete stereo information, comparisons should be made with stereo information removed. Of course, manual confirmation may be necessary using chemical names if stereo distinctions are important.

Comparing INChI strings without regard to certain layers is, in effect, equivalent to removing the ignored layers and all information that logically follows it. This often amounts simply to truncating the INChI. If a shorter INChI matches the corresponding characters in a longer INChI, the longer is a more specific representation of the substance in the shorter. In other cases, this requires excising layers.

Stereochemistry:
Perhaps the most common problem in identifying two matching compounds is dealing with the absence of complete stereochemical information in one of them. Structure records in many large data collections may contain little or no such information. To find matching records, all stereo layers much be removed from the INChI.

Alternatively, if the structures of interest were derived from 2-D drawings, they may contain Z/E (sp^2) stereo information, but no tetrahedral (sp^3) stereo information. In this case, only the sp^3 sublayer of the stereo layer need to be removed.

Mobile H-Atoms (Tautomers):
If representations of tautomers are to be compared to INChI representations of the same substances but with fixed (immobilized) H-atoms, the fixed H layer should be removed. For INChI, fixing these H-atoms is a refinement of a structure, it forms an added layer that may be removed without affecting the preceding tautomer representation. To ignore all forms of H-migration, included those not defined in INChI rules (keto-enol tautomerism, for instance), one may compare just the formula and first connectivity sublayer (no H-atoms). Any matches found mean that the original structures have identical skeletons and atomic composition.

Isotopes:
To ignore isotopic substitution one simply needs to exclude the isotopic layer.

Charges and Protons:
The charge and proton layer are independent of all others and may be simply removed to eliminate dependence on charge or degree of protonation or deprotonation. Therefore, in a comparison to find identical compounds, do not consider these layers in a comparison unless you wish to distinguish different charge and proton states.

Comparison of Connectivity Only:
By using only the chemical formula and connection sublayers, it is possible to identify matching basic structures without regard to precise hydrogen location, stereochemistry and isotopic substitution. This provides a quick way to find matching compounds that may differ due to forms of hydrogen transfer not accounted for by INChI (keto-enol tautomerism, for instance.)

Appendix 5. Structure Representation Ambiguities

While a structural drawing is the generally accepted means of representing chemical identity, this method can be imprecise or ambiguous for a number of reasons. INChI can expose, but not resolve these inherent problems. When possible, it identifies them by issuing warnings and perhaps isolating them in specific layers. In some cases these problems arise from uncertainties on the part of the chemist, in other cases the problem is caused by the lack of accepted standards for representation. The individual who draws the structure may not even be aware of these uncertainties. Some of the more commonly encountered sources of such ambiguity are discussed in this section.

Mobile Hydrogen:

Since positions or even the number of certain hydrogen atoms (or protons) in a compound are may be be fixed or not known due to ease of migration, a single substance 'in a bottle' may not be readily represented as a single discrete chemical entity. On the other hand, in some circumstances substances are best represented where all 'mobile' H-atom locations are fixed, as, for instance, is often the case in the gas phase. To accommodate both forms of representation, INChI adds a separate 'layer' to fix specific locations of the mobile-H atoms defined in a prior layer. Stripped of this added layer, the INChI reverts the mobile H-atom representation. Also, the specific degree of protonation may also be specified if desired.

Contributing to the problem is that the possible locations of all mobile H-atoms may not be reliably known and may depend on the chemical environment (solvent, pH, and temperature, for instance).

Fortunately, a large fraction of the most common H-migration possibilities may be expressed by a few rules, which are implemented in INChI. The case of guanine is illustrated below.

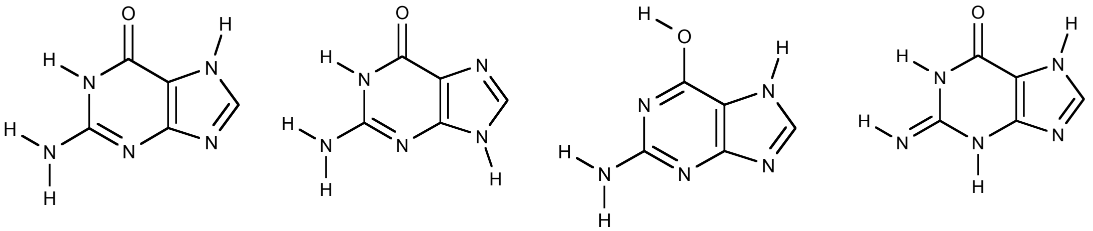


Figure A5-1. Tautomeric structures of Guanine

Another concern is the lack of a formal means of representing the types or even the presence of mobile hydrogen in a drawn structure. For INChI generation, the user must specify whether a fixed or mobile hydrogen representation is to be used. If no mobile hydrogen atoms are found to be present, this specification is simply ignored.

Stereochemistry:

Two varieties of stereochemistry are represented by INChI, double-bond (sp^2) and tetrahedral (sp^3). The former information may be extracted from atom coordinates given in conventional drawings. Uncertainties may arise, however, when a formal double bond is actually rotatable (facile rotation about the formal double bond), in which case Z/E stereoisomers are not distinguishable. Four examples of such ambiguity are listed below:

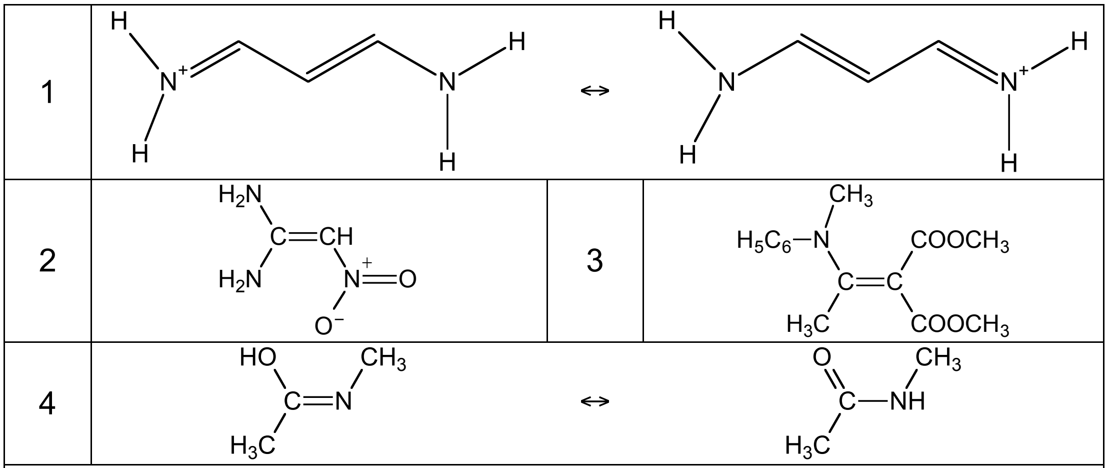


Figure A5-2. Examples of Rotatable Double Bonds

Among mobile hydrogen isomers, a bond may be formally rotatable when an H-atom is at one location, and not rotatable when in another (Example 4). In such cases, the bond is presumed by INChI to be rotatable. To further diminish possible ambiguities, Z/E stereochemistry is ignored when found in rings containing seven or fewer atoms. This, for example, eliminates that need for a stereolayer for benzene, which can formally exist in highly strained Z-forms.

While some drawing programs may allow users to express the lack of Z/E stereoisomerism in the examples above, unfortunately, users will often not use them.

If an input double bond suggests Z/E stereochemistry, but the INChI analysis indicates that it may also be represented as a single bond, a warning will be issued and the structure will ignore that Z/E stereochemistry *[the warning is not implemented yet].*

Regarding the representation of sp^3 stereochemistry, perhaps the most common problem is that the requisite stereo-information is partially or completely absent. Another problem is that errors are commonly made in complex drawings with multiple stereocenters. These difficulties, in fact, provided the principal initial motivation for creating a layered INChI, where these problems are held in a single layer that may be ignored if desired. It also allows structure collections without stereo descriptors to employ identifiers consistent with structure representations that have stereo-labeling. If structure representations are accurate and complete, full sp^3 layers will be the same for the same compound. In other cases, the sp3 layer may be ignored or processed further to confirm identity (by inspection of chemical name or use of third-party structure processing software).

Organometallic Compounds and Coordination Bonds:
No widely accepted means of representing organometallic substances exists. Ferrocene, for instance, may be drawn with the central iron atom connected to each of the two attached rings, to each of the atoms in the rings, to each of the bonds in the rings or not connected at all. The approach taken by INChI is to logically dissociate all atoms capable of forming coordination bonds (metals) and represent the structure as the individual, interconnected components along with the separated, unconnected metal atoms. For a large majority of organometallic compounds, this provides a unique INChI. If a bonded organometallic structure representation is desired, however, it may be specified by adding another series of layers to the INChI.

Multiple Components:
Many substances are best represented as multiple, independent structures. INChI will represent such substances by simply appending the individual layers for each component in each layer and sorting these components using a set of fixed rules. INChI creation assumes that if multiple structures are present in a single input connection table, they are components of a single compound. In most cases, it is possible to extract the INChI of each component from a composite INChI by excising the corresponding part of each layer. The order of the components in the layers is strictly defined.