IUPAC International Chemical Identifier (InChI) Programs

InChI version 1, software version 1.04 (September 2011)

User's Guide

Last revision date: September 12, 2011

This document is a part of the release of the IUPAC International Chemical Identifier with InChIKey, version 1, software version 1.04.

CONTENTS

I. OVERVIEW

About InChI

The IUPAC International Chemical Identifier (InChI) provides unique labels for well-defined chemical substances. These labels are generated by converting an input chemical structure, in the form of a 'connection table', to a unique and predictable series of ASCII characters. They offer a means for representing chemical compounds in a manner that does not depend on how they were drawn. Note that they are re-expressions of chemical structures, they are not registry or registration numbers and do not require access to a database. They were developed primarily as a means of 'naming' a compound in digital media although they are expressed as simple text that may be manually interpreted. This document describes the operation and output of the present version of the program that generates this Identifier.

The Identifier is designed to process single, well-defined chemical compounds (which may be composed of multiple components).

Technical details are given in a separate document, the InChI Technical Manual. The basic algorithms were taken from the literature, with selection, testing and implementation done primarily at NIST; with modifications and additions by IUPAC and the InChI Trust.

In the several years of its development, many individuals contributed to the development of the InChI at meetings and through correspondence. The chemical rules employed are intended to represent a consensus view of the concept of chemical identity. The computer program described in this document applies these algorithms to input structures and generates both the Identifier and an annotated depiction of the structure.

Derivation of the InChI from an input chemical structure proceeds through three steps: 1) normalization – all input information not needed for structure identification is discarded and structure information is divided into 'layers'; 2) canonicalization – each atom is given a label that depends only on its position in the structure; 3) serialization – a

string of characters, the Identifier, is generated from the canonical labels. All 'chemical' rules are applied in the first step.

The current version of the Identifier is 1; the current version of the InChI software is 1.04 (September 2011). Previously released versions 1.01 (2006), 1.02-beta (2007), 1.02-standard (2009), and 1.03 (June 2010) as well as all earlier versions, are now considered obsolete.

Standard and non-standard InChI

InChI has a layered structure which allows one to represent molecular structure with a desired level of detail. Accordingly, the InChI software may generate different InChI strings for the same molecule, depending on the choice of a multitude of options (e.g., distinguishing or not distinguishing tautomers). This flexibility, however, may be considered a drawback with respect to standardization/interoperability. The standard InChI which is always produced with fixed options was defined by the IUPAC InChI Subcommittee in response to these concerns.

The standard InChI was defined to ensure interoperability/compatibility between large databases/web searching and information exchange. As related to its internal layered structure, standard InChI, introduced in v.1.02-standard (2009) release of InChI software, is a subset of IUPAC International Chemical Identifier v.1. The layered structure of the standard InChI conforms to the following requirements.

- Standard InChI organometallic representation does not include bonds to metal for the time being.
- Standard InChI distinguishes between chemical substances at the level of 'connectivity', 'stereochemistry', and 'isotopic composition', where:
    - connectivity means tautomer-invariant valence-bond connectivity (different tautomers have the same connectivity/hydrogen layer);
    - stereochemistry means configuration of stereogenic atoms and bonds; unknown stereo designations are treated as undefined;

- isotopic composition means mass numbers of isotopic atoms (when specified)

Standard InChI v.1 was introduced in v. 1.02-standard release of the InChI software in 2009 (this software version was able of generating only standard InChIs).

The present release of InChI software, v. 1.04, has merged functionality. It allows one to produce both standard and non-standard InChI strings, as well as their hashed representation (InChIKey).

By default, InChI software v. 1.04 produces standard InChI (for brevity, stdInChI below). In particular, the standard identifier is generated when the software is used without any specifically added options. If some options are specified, and at least one of them qualifies as related to non-standard InChI (see section 'InChI Software Options' below), the program produces non-stdInChI/InChIKey.

The standard InChI is designated by the prefix: "InChI=1S/…….. " (that is, letter 'S' immediately follows the Identifier version number, '1'; Identifier version numbers should always be whole numbers).

Non-standard InChI is designated by the prefix: "InChI=1/…….. " (that is, letter 'S' is omitted).

About InChIKey

The InChIKey is a character signature based on a hash code of the InChI string. A hash code is a fixed length condensed digital representation of a variable length character string. Providing a hash derived from an InChI string should be helpful in search applications, including Web searching and chemical structure database indexing; also, this hash may serve as a checksum for verifying InChI, for example, after transmission over a network.

The InChIKey consists of two blocks. The first block is always the same for the same molecular skeleton. All isotopic substitutions, changes in stereoconfiguration, tautomerism and protonation are reflected in the second block.

A standard InChIKey, which is a key produced from a standard InChI, does not account for tautomerism and may indicate only absolute stereo (or completely ignore stereo). It also does not account for the original structure's bonds to metal.

The two hash blocks of InChIKey are based on a truncated SHA-256 cryptographic hash function (http://en.wikipedia.org/wiki/SHA_hash_functions#SHA-2).

Note that due to the very essence of hash functions, collisions (the same InChIKey for different InChIs/structures) are unavoidable in very large collections. A theoretical – optimistic – estimate of collision resistance (i.e., the minimal size of a database at which a single collision is expected, that is, an event of the two hashes of two different InChI strings being the same) is 6.1×10^{9} molecular skeletons $\times$ 3.7×10^{5} stereo/ isotopomers per skeleton $\approx$ 2.2×10^{15}. To exemplify: the probability of a single first block collision in a database of 1 billion compounds is 1.3%. In other words, a single first block collision is expected in 1 out of 100/1.3 = 75 databases of 10^{9} compounds each. For 10^{8} (100 million) compounds in a database this probability is 0.014%.

A beta-version of the InChIKey was introduced in software v. 1.02-beta (2007). The standard InChIKey was introduced in v. 1.02-standard release (2009) as an InChIKey computed from the standard InChI and intended for the principal purpose of a search-engine-style lookup of chemical information.

The present release of the InChI software, v. 1.04 of 2011, has merged functionality. It allows one to produce both standard and non-standard InChIKey.

Note that the current format of InChIKey is different from that of the beta version (2007); the format of the standard InChIKey is the same as that of v. 1.02-standard (2009).

II. ABOUT InChI PROGRAMS

This document is accompanied by version 1.04 of the InChI generator. This program runs under 32-bit Microsoft Windows Operating Systems. The main program, `winchi-1.exe`, is a conventional Windows graphical-interface application.

The 'command line' version `inchi-1.exe` is also included. A version recompiled under Linux (32 and 64 bit executables, `inchi-1`) without any changes is also supplied.

The program `winchi-1` takes an input structure and generates both graphical and text output in a form designed to allow critical examination of the InChI. The Identifier and associated text output may be parsed and annotated in either a simple plain text or XML (eXtensible Markup Language) format.

As structure input, the program currently accepts standard SDfiles, Molfiles [see "Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited" by Arthur Dalby, James G. Nourse, W. Douglas Hounshell, Ann K. I. Gushurst, David L. Grier, Burton A. Leland, and John Laufer, Journal of Chemical Information and Computer Sciences, 1992; 32(3); pp. 244-255; a more recent description of V2000 format may be downloaded from http://www.mdli.com/downloads/public/ctfile/ctfile.jsp], or its own output produced when the "Full auxiliary information" option is selected. Input may originate from individual disk files or through the Windows clipboard.

InChI may be also generated by using Software Library/application programming interface (API). This is described later.

III. RUNNING InChI PROGRAMS

Graphical Interface Program (winchi-1)

Introduction

The InChI generation program is provided along with sample chemical structures in a 'zip' file INCHI-1-BIN.zip. To use this program, first extract the contents of the file to a directory of your choice. To start the program, run the file `winchi-1.exe` that was extracted from the zip file. Figure 1 then appears on your monitor.

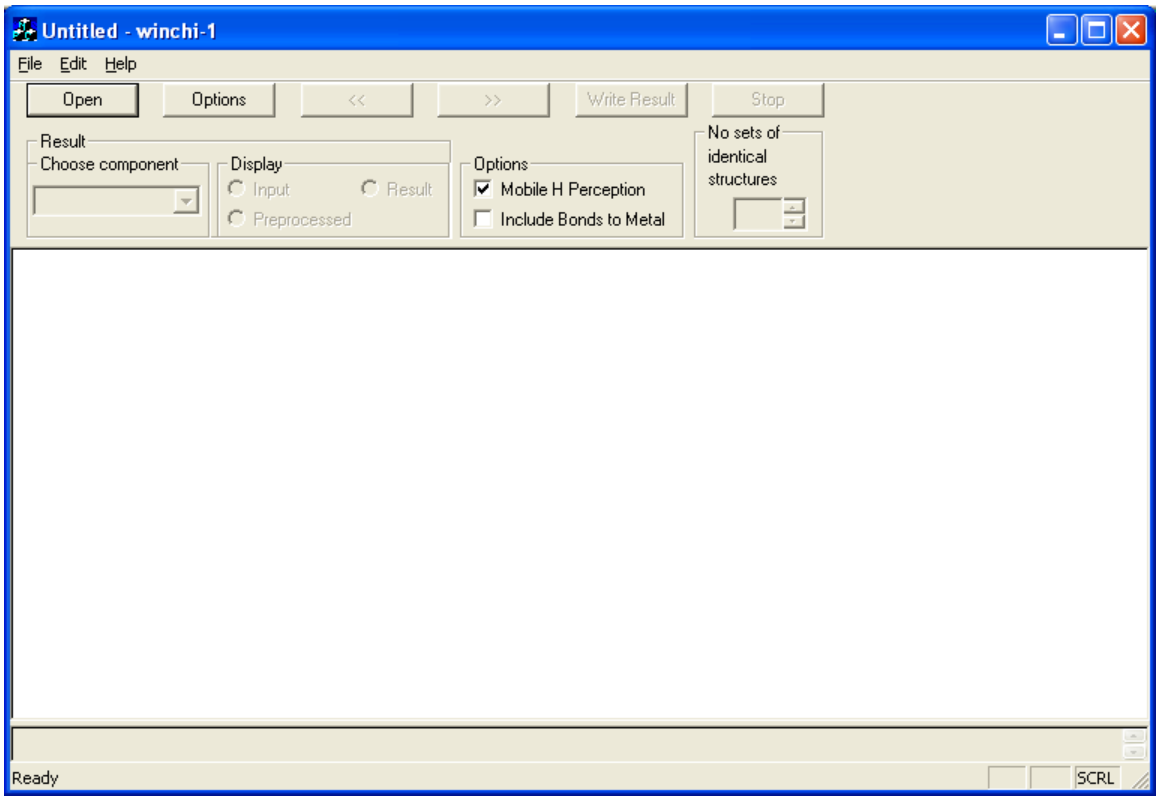


Figure 1

Generating an InChI begins with the selection of an input structure file. The simplest way is to drag the input structure file from Windows Explorer directory list into the InChI window. Structures also may be copied from certain chemical structure editors (ISIS/Draw with "Copy Mol/Rxnfile to the Clipboard" option or from ACD/ChemSketch) and pasted into the InChI window (Select Edit → Paste from InChI menu). The input structure file pathname may be provided as a command line option when you start `winchi-1`. Selection of the input structure file may also be done by first clicking on the 'Open' button (top left corner of Figure 1) and then, in the dialog box that appears (as shown in Figure 2),

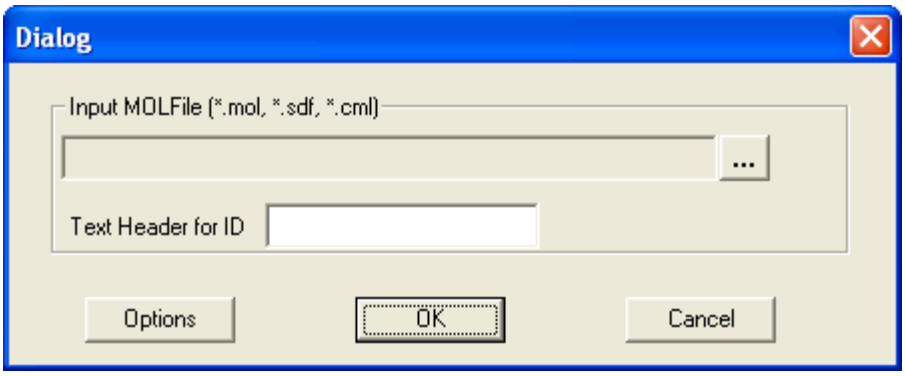


Figure 2

selecting a structure file using the '…' button on the right of the 'Input Structure File' field. You may select any of the sample .mol or .sdf files for initial testing. In this dialog you may also enter "Text Header for ID"; this will simply add to the InChI header a structure ID if it is present in an input SDfile (from other input formats the header and ID are extracted automatically). Ignore this box for now.

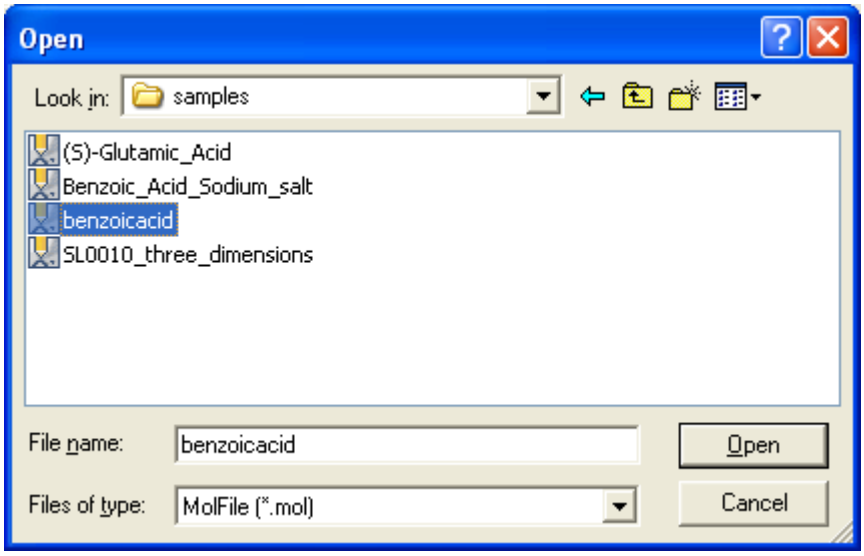


Figure 3

Figure 3 shows the selection of a structure file. In this case it is entitled benzoicacid.mol, which was prepared by a separate structure-drawing program. Clicking the file name copies it into "File name:" line. After that click "Open" to close the dialog.

At this point you may also change InChI processing options. (The choices for the options that can be changed are shown in Figure 4, but no changes are made in this example.)

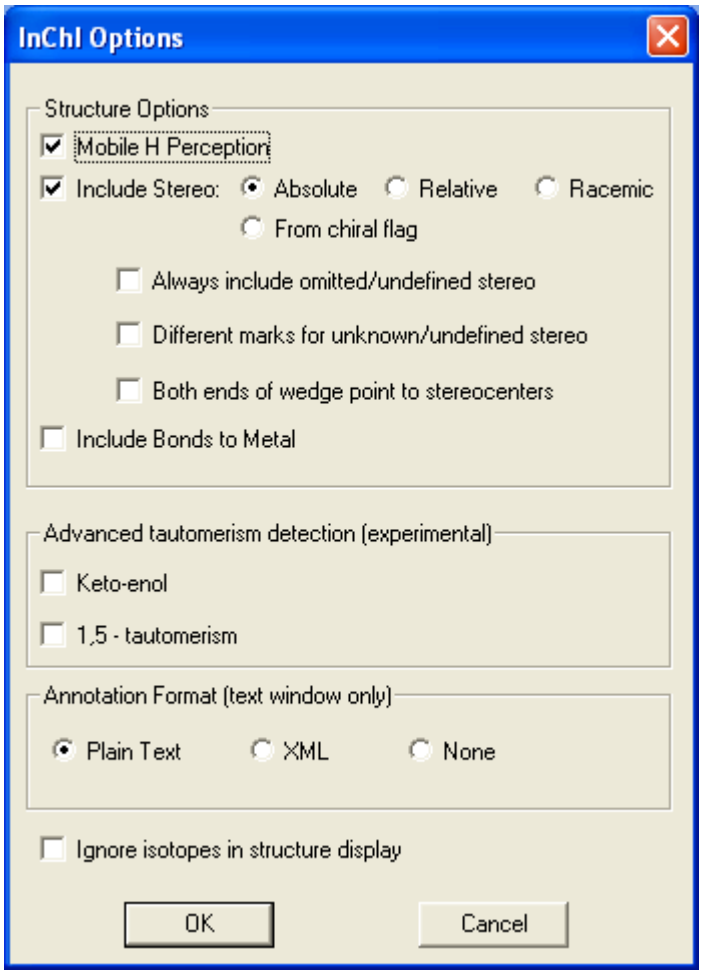


Figure 4

Close InChI Options dialog if you opened it and select OK in the dialog (Fig. 2) when done; the result is Figure 5.

The main output window is composed of two sections: the upper section (shown in white in Figure 5) shows structural information graphically and the lower section (shown in gray in Figure 5) shows text output.

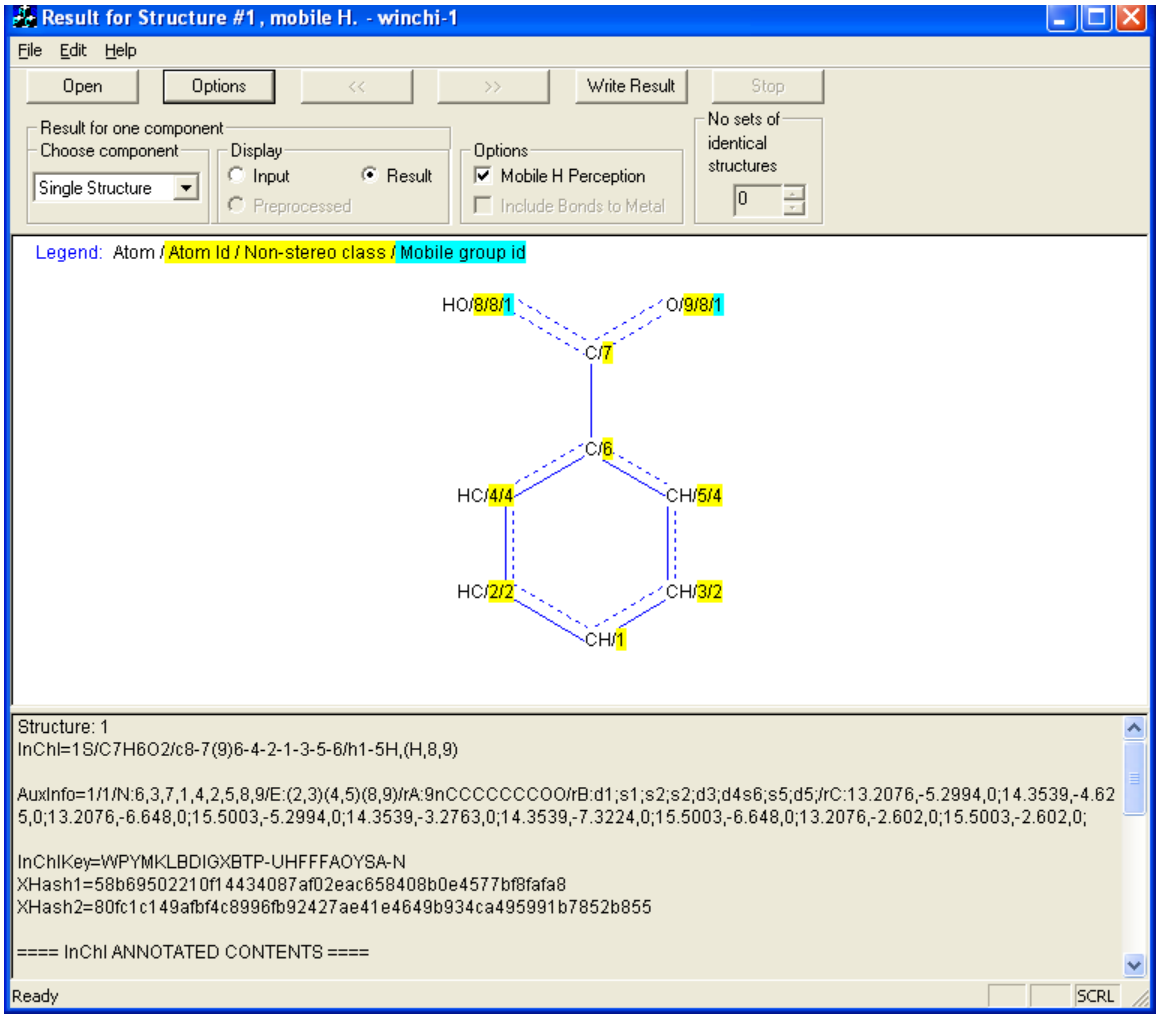


Figure 5

Upper section

The structure is displayed along with labels generated by InChI algorithms. In cases where an SDF file is input, the first structure shown is the first entry in the input file. The example shown in Figure 5 is a single component example. If more than one component (independent structure) is found in the first structure file (such as benzoic acid, sodium salt shown in Figure 6), each may be separately examined using the "Choose component" 'combo box' on the upper left of the screen, although they are treated as part of a single compound by InChI (Figures 7 and 8).

The buttons under “Display” permit viewing of the input structure and the preprocessed structure if it differs from the input structure. . The buttons under “Options” are the same as in the “Options” dialog box. “Mobile H Perception” removes the “fixed-H” part of the identifier. Figure 9 shows the same structure with the option “Mobile H Perception” off.

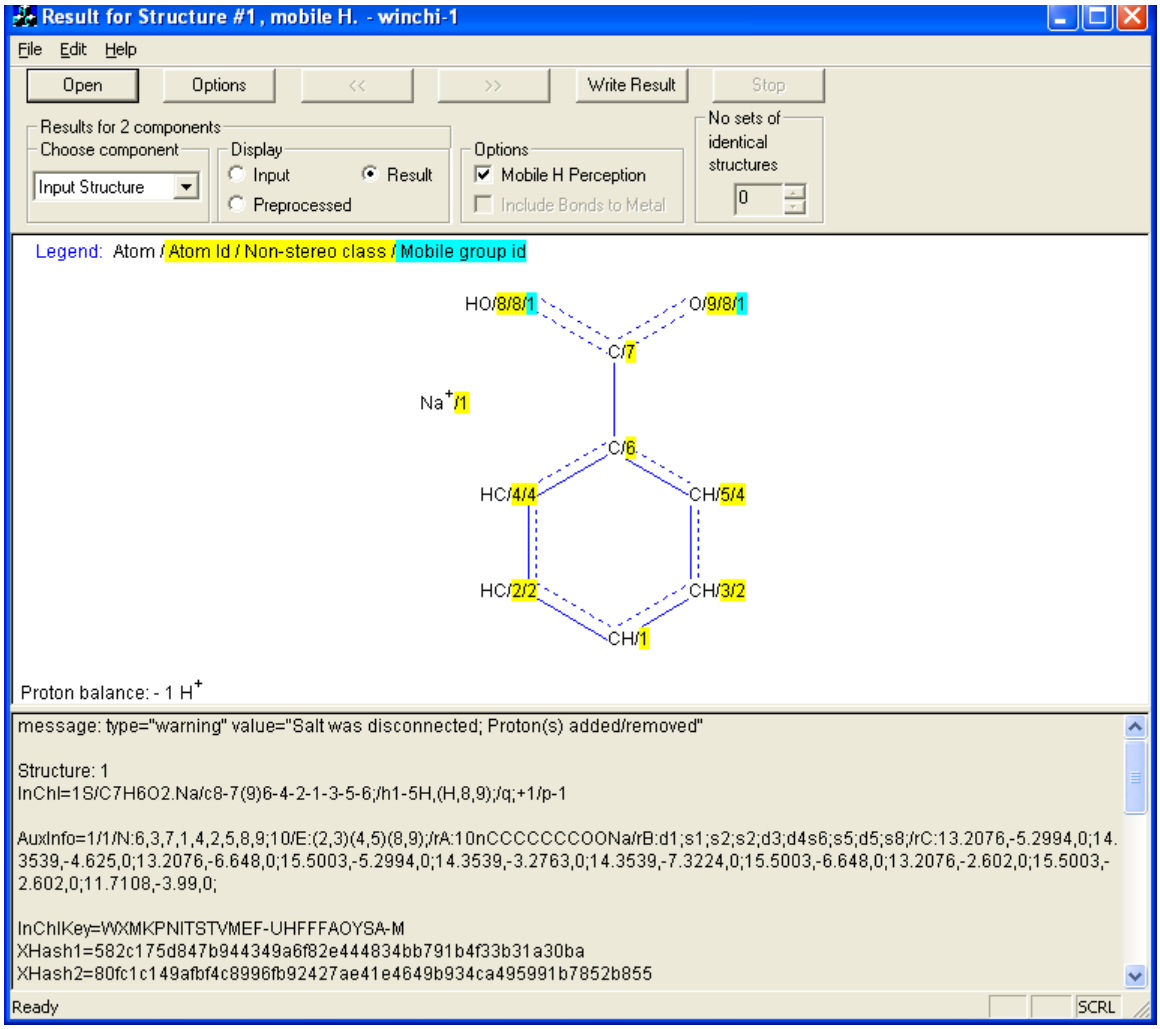


Figure 6

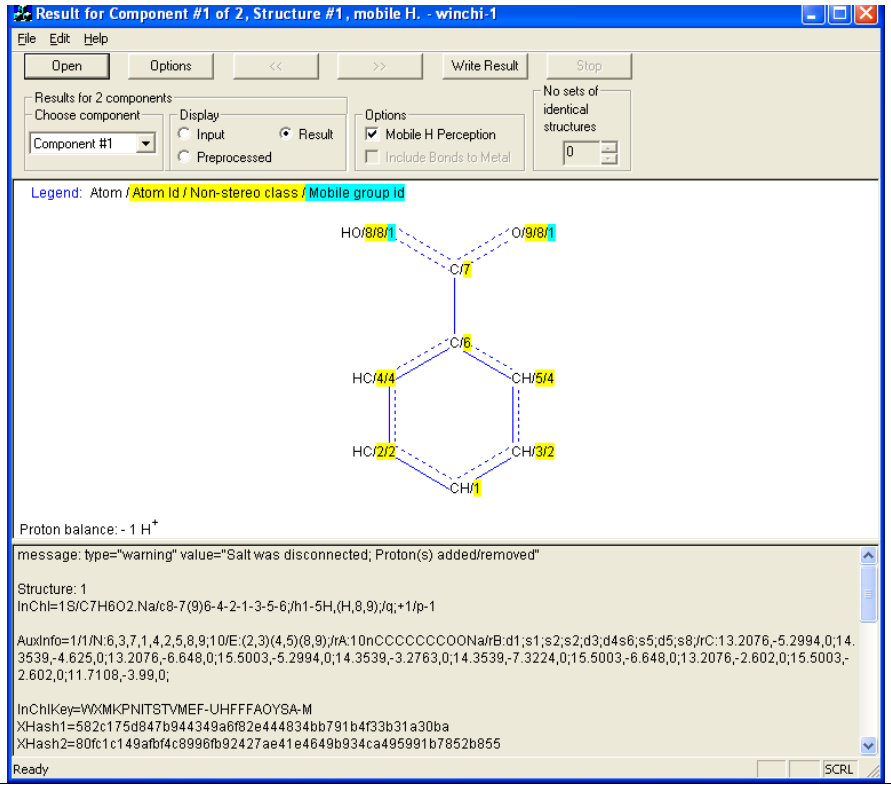


Figure 7

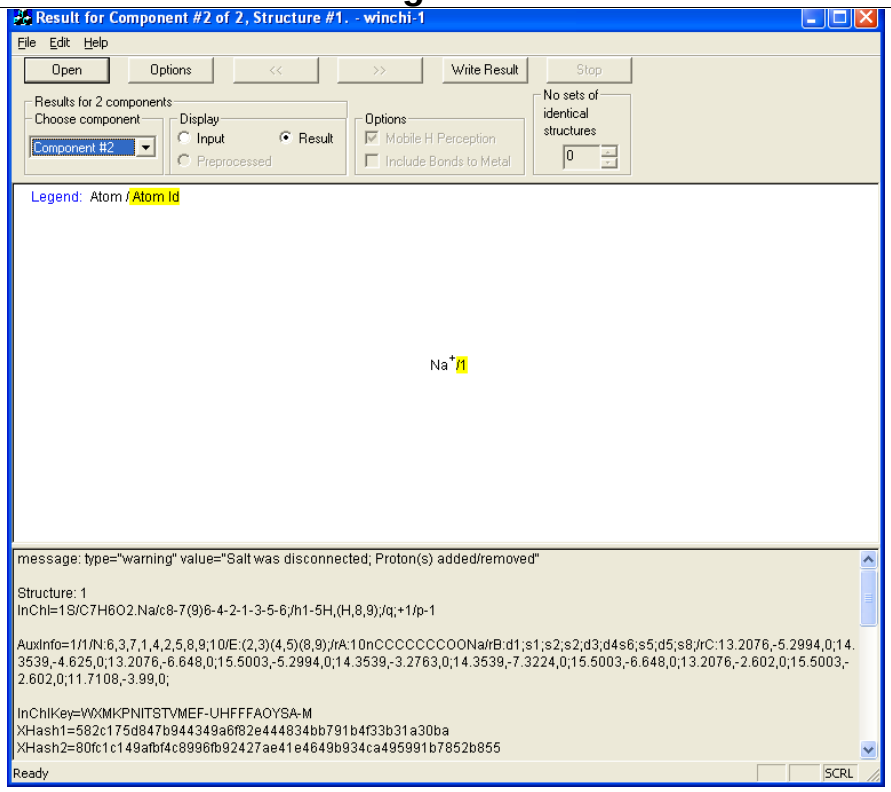


Figure 8

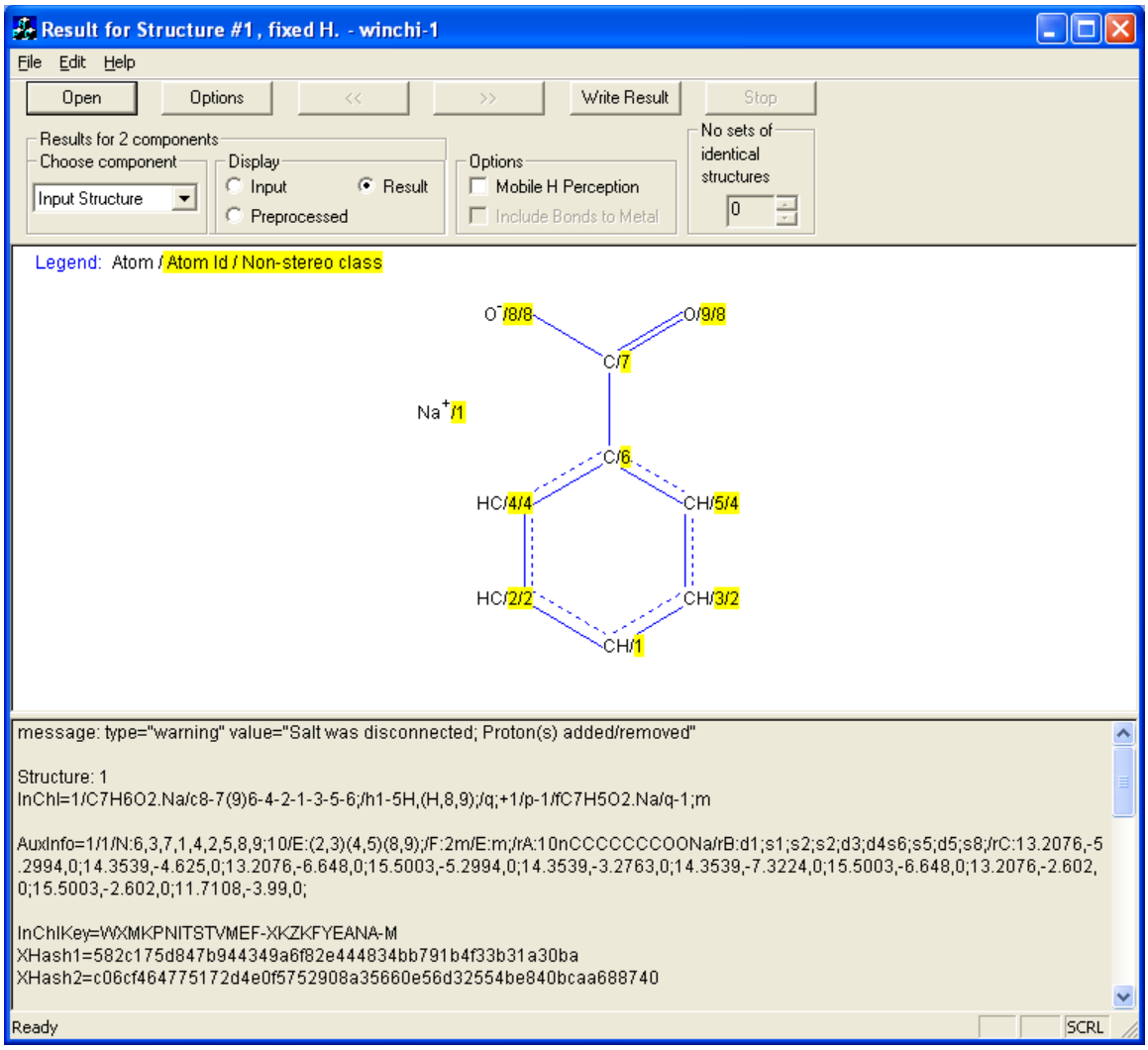


Figure 9.

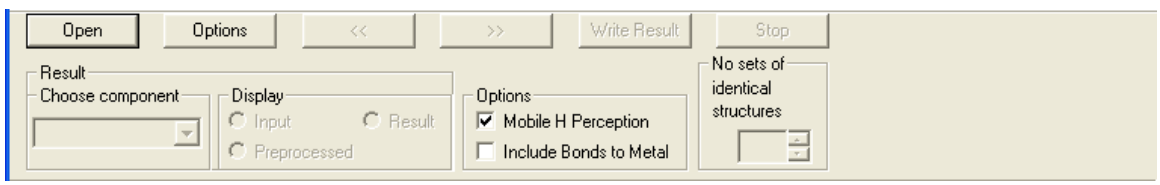


Figure 10. InChI Toolbar

On the InChI Toolbar the rightmost box displays the number of sets of equivalent components. When equivalent components are found, they may be highlighted by making a selection in the box. This provides a quick way to determine if two depictions of the same compound are considered to be the same by InChI algorithms, although the actual InChI generated will represent the collection of structures as a single compound.

The structure display shows the canonical identification number of each atom along with the non-stereo equivalence class number assigned to that atom. The canonical number is the unique number given to an atom and used for 'serialization' (creation of the actual InChI). The non-stereo equivalence class number is a number assigned to each set of equivalent atoms (all atoms having the same equivalence class number are indistinguishable, *ignoring stereochemistry*; the equivalence class number is the smallest canonical identification number in the class of equivalent atoms). This information is only intended to assist in the understanding of results of InChI processing and is not directly used in InChI generation except in the processing of stereochemistry.

Stereochemical parities of bonds and atoms are also displayed. A question mark symbol indicates that stereoisomerism is possible, but the configuration has not been specified. Bonds that have been found to be variable by alternation or movement of mobile H-atoms or charges are shown by dotted lines. This information is used only for deciding which bonds may exhibit double bond (*Z/E*) isomerism. By design, the Identifier does not explicitly represent bond types.

Lower Section

The InChI along with auxiliary data and explanatory information is shown in the lower section of the output window, such as seen in Figure 6 (see Section VI). Unlike the graphical display, even if more than one disconnected component is found, all textual results for a single input structure file are shown together. This reflects the important point that all components of a submitted structure are considered by InChI to be part of a single compound. Results for different (disconnected) components of a single substance are separated by semicolons, except for chemical formulas, which, in keeping with common conventions, are separated by dots.

Options

Pressing the Options Button opens the InChI Options Dialog Box. The following options are then available (as seen in Figure 4):

- Mobile H Perception – turning Off will fix all H-atoms (disallow H-migration), this allows the generation of a fixed-H section of the Identifier (and makes the resulting InChI non-standard).
- Include Stereo (Absolute, Relative, Racemic, From chiral flag) – include stereo layer and choose its type or exclude all stereo information from the identifier. If the last option is selected then in presence of a chiral flag stereochemistry is considered absolute, otherwise relative.
  For standard InChI the only allowed choice is absolute stereochemistry or omission of all stereo; other choices make InChI non-standard.
- Always include omitted/undefined stereo – by default, InChI does not include unknown/undefined stereo unless at least one defined stereo is present in the input structure. Turning this option On results in inclusion of unknown/undefined stereo in all cases.
- Different marks for unknown/undefined stereo – turning this option On will result in usage of the two different signs, 'u' and '?', for "unknown" and "undefined" stereo. Briefly: "undefined" means not given while "unknown" means explicitly marked as unknown, e.g., with "wavy" bonds. By default, this option is turned off and the twoh signs are merged to '?' (that is, "unknown" stereo treated as "undefined").
- Both ends of wedge point to stereocenters – by default, this option is turned Off. This means that that a stereo bond depicted by a wedge affects the stereochemistry of only the atom 'pointed to' by the narrow end of that wedge. However, it may be turned On if the user is completely sure that a stereobond affects both atoms it connects (that is, for 2D structures complying to the legacy "perspective" stereochemistry drawing style).

- Include Bonds to Metal - turning On will add a layer that includes specific bonding to metals (in case of salts the bonds between a metal and an acid cannot be reconnected – as seen in Figures 6-9 where that choice is "grayed out" and cannot be ticked or checked).
- Annotation Format (Plain Text; XML, None) – choose appropriate format for explanatory information.
- Ignore Isotopes in Structure Display – this does not change the identifier, it only affects the structure appearance and the display of sets of equivalent components.

Note that the above options form a subset of a full options set available in the command-line executable `inchi-1` (see section 'InChI Software Options' below).

Text File Output

At any time you may select 'Write Result' to analyze the input file and write all textual results to an output file located in the same directory as the program. The name of this file is derived from the name of the input structure file and is displayed when it is created (the name has extension .txt). Figure 11 is an example of this for benzoic acid. It shows the directory/location on the computer as well as the file names given to the three (3) output files. Two other files to assist in diagnosing problems, should they occur, are created and their names displayed. One of them is a log file; it contains names of input and output files, a list of selected options, warning and error messages, number of processed structures, processing time, etc. The name of this file has extension .log. Another file – a problem file -- contains input structure file records that caused errors. This file (its name has extension .prb) may be important to determine reasons for the errors. A listing of errors and warnings is given in the Appendix 1.

Figures 12-14 show the content of the three output files. The .prb file is, of course, empty, since there were no problems encountered in generating the InChI for benzoic acid.

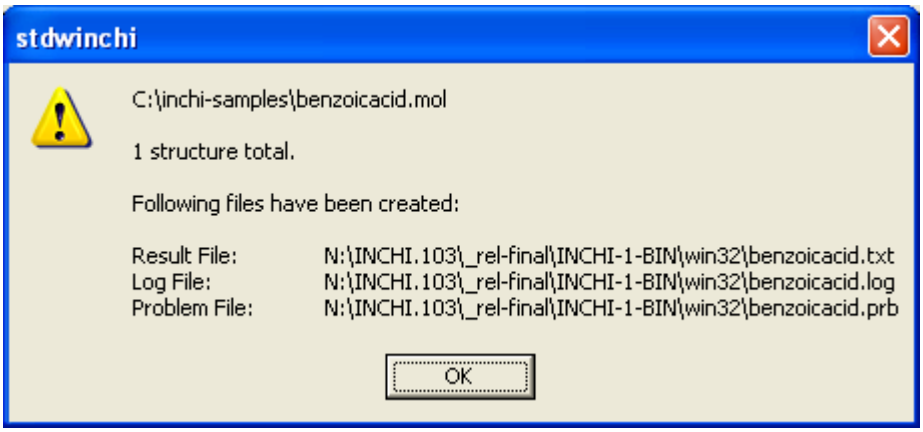


Figure 11

```
* Input_File: "C:\inchi-samples\benzoicacid.mol"
Structure: 1
InChI=1S/C7H6O2/c8-7(9)6-4-2-1-3-5-6/h1-5H,(H,8,9)
AuxInfo=1/1/N:6,3,7,1,4,2,5,8,9/E:(2,3)(4,5)(8,9)/rA:9nCCCCCCCOO/rB:d1;s1;s2;s2;d3;d4s6;s5;d5;/rC:13.2076,-5.2994,0;14.3539,-4.625,0;13.2076,-6.648,0;15.5003,-5.2994,0;14.3539,-3.2763,0;14.3539,-7.3224,0;15.5003,-6.648,0;13.2076,-2.602,0;15.5003,-2.602,0;

InChIKey=WPYMKLBDIGXBTP-UHFFFAOYSA-N
XHash1=58b69502210f14434087af02eac658408b0e4577bf8fafa8
XHash2=80fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855
```

Figure 12. File benzoicacid.txt

```
InChI version 1, Software version 1.04 Build of September 9, 2011
Opened log file 'benzoicacid.log'
Opened input file 'benzoicacid.mol'
Opened output file 'benzoicacid.txt'
Opened problem file 'benzoicacid.prb'
The command line used:
"C:\inchi-samples\inchi-1.exe benzoicacid.mol benzoicacid.txt benzoicacid.log benzoicacid.prb"
Generating standard InChI
Input format: MOLfile
Output format: Plain text
Full Aux. info
Timeout per structure: 60.000 sec
Up to 1024 atoms per structure

End of file detected after structure #1.
Finished processing 1 structure: 0 errors, processing time 0:00:00.00
```

Figure 13. File benzoicacid.log

Figure 14. File benzoicacid.prb

Command Line Program (inchi-1)

For those familiar with the Windows 'Command Prompt', an executable program is also provided – `inchi-1.exe`. This program uses 'command line' arguments that are shown by invoking the program without any arguments. The principal use of the program is to allow batch processing within other programs for the processing of multiple structure files. At present, this program is intended primarily for processing SDF files. This program may be recompiled and used under Linux without any changes. The Linux version does not display chemical structures.

Standard redirection may be used to suppress `inchi-1` console output.

Under Windows:

```
inchi-1 /AuxNone input.sdf output.txt logfile.log NUL 2>NUL
```

Under Linux:

```
inchi-1 -AuxNone input.sdf output.txt logfile.log NUL
2>/dev/null
```

">" or "1>" redirects standard output, "2>" redirects standard error output.

To process files greater than 2 GB with `inchi-1`, the output of a problem file should be suppressed. To do that, the output and log file names should be included in the command line; the name of the problem file should be NUL, for example:

```
inchi-1 input.sdf  output.txt logfile.log NUL
```

Note that the graphical program `winchi-1.exe` cannot process files greater than 2 GB.

AMI (Allow Multiple Inputs) mode

Since InChI software v. 1.04, the possibility of processing multiple input files at a single run was added to the inchi-1 executable (both Windows and Linux versions).

This mode is activated by the inchi-1 command line option "/AMI" (Windows) or "–AMI" (Linux; AMI stands for "Allow Multiple Inputs"). In this mode, all the file names supplied in the command line are considered as the names of separate input files.

For further convenience, the common file name wildcards ("*" and "?") are supported.

For example, issuing a command

```
inchi-1 *.mol /AMI (Windows)
```

```
inchi-1 *.mol -AMI (Linux)
```

will instruct the executable to process all the mol-files in the current directory.

Note, that omitting the switch "AMI" assumes working in a conventional single-input mode which may result in undesired treatment of wildcards[1].

In AMI mode, the names of output, log and problem files could not be individually specified. Instead, they are formed, for each of multiple inputs, by appending the file name with suffixes ".txt", ".log" and ".prb". However, to partially mimic the behavior of inchi-1 in conventional single-input mode, three additional command line options are introduced (see section "Availability of InChI software options", Table 6). They allow one to redirect the output to stdout, log to stderr, as well as to suppress creation of problem files.

Examples *(Windows, Linux)*:

```
inchi-1 nci*.mol /AMI /AMIOutStd /AMIPrbNone /AuxNone /Key
```

```
./inchi-1 /home/me/mol/nci/*.mol -AMI -AMILogStd -
AMIPrbNone
-RecMet -FixedH
```

[1] There is an important difference in wildcard expansion under Windows and Linux.
Under Windows, inchi-1 executable makes an expansion itself (if "AMI" switch is specified). That is, if "AMI" is omitted, no expansion occurs and "*.mol" is just considered as an invalid file name.
Under Linux, wildcards are always expanded by shell. That is, if "AMI" is omitted, "*.mol" will be expanded to the list of file names; the first four of them will be treated by inchi-1, according to single-input rules, as input, output, log and problem file names (which means that the last three files will be over-written).

As indicated by tests, processing of multiple MOL files in AMI mode may be several times faster (the exact speed-up depends on many details; anyway the processing time is still significantly longer than that for a single SDF file containing the same data).

InChI Software Library (libinchi)

For advanced users who may want to create the Identifier in their own software the InChI Software Library (InChI API) is provided in a separate package. The package contains 'C' source code for `inchi-1.exe`, 'C' source code for the InChI Library that may be compiled into a Dynamic Link Library (DLL) `libinchi.dll` under Windows or Shared Object (SO) `libinchi.so` under Linux; also, there are 'C' and Python examples of simple applications that read input Molfile and use the InChI Library to produce Identifiers.

The InChI Library does not display structures and is not able to read chemical structural data from the input file. It uses specially formatted input binary data and produces three strings: InChI, the Auxiliary Information, and, if necessary, an error or warning message. Also, there are procedure to calculate InChIKey and other service routines. The source code is accompanied with makefiles tested with gcc under Windows and Linux.

The InChI Library allows one to generate both standard and non-standard InChIs/InChIKeys. For example, an API function `GetINCHI()` produces standard InChI by default and non-standard InChI if some "InChI creation option" is specified in input parameters. However, for compatibility with the previous v. 1.02-standard (2009) release, the procedures which deal only with stdInChI – for example, `GetStdINCHI()` - are retained.

The InChI API calls are documented in the separate "InChI API Reference Sheet" document and source code header file "inchi_api.h" included in the package.

InChI Software Options

The exact set of InChI software options has been changing from release to release. The description below refers to the current, v. 1.4 (2011) release.

The options are available in graphical program `winchi-1`, command line executable `inchi-1` and through InChI API. Not all the options are available for all the parts of software; the maximal set of options is available for the `inchi-1` program.

Structure perception and InChI creation options

Options affecting generation of InChI are divided on "structure perception" options and "InChI creation" options.

The perception options are considered drawing style/edit flags which affect the input structure interpretation and are not memorized. It is assumed that the user may deliberately use these options to account for the specific features of structure collections. Whence, perception options may be used while generating standard InChI without loss of its "standardness".

Perception options are listed in the following table. Presented here are command line switches available (they should be used with the appropriate prefix - i.e., 'NEWPSOFF' should be entered as '/NEWPSOFF' under Windows and '-NEWPSOFF' under Linux).

Table 1. Structure perception options.

Structure perception option	Meaning	Default behavior (standard; if no option supplied)
`NEWPSOFF`	Both ends of a wedge (which indicates stereochemistry) point to stereocenters	Only the narrow end of a wedge points to a stereocenter
`DoNotAddH`	All hydrogens in input structure are explicit	Add H according to usual valences
`SNon`	Ignore stereo	Use absolute stereo

There are several options (Table 2) which modify the interpretation of input stereochemical data. In principle, they also may be considered "structure perception"

options. However, as the standard InChI, by definition, requires the use of absolute stereo (or no stereo at all), these "perception" options assume generation of non-standard InChI.

Table 2. Stereo interpretation options (lead to generation of non-standard InChI).

Stereo option	Meaning	Default behavior (standard; if no option supplied)
SRel	Use relative stereo	Use absolute stereo
SRac	Use racemic stereo	Use absolute stereo
SUCF	Use Chiral Flag in MOL/SD file record: if On – use absolute stereo, Off – relative	Use absolute stereo (or another option if requested by SRel /SRac/SNon switches)

The creation options affects the InChI algorithm, not structure perception. They modify the defaults which are specified for standard InChI and significantly affect the final appearance (e.g., additional InChI layers may appear). Whence, using any of the creation options qualifies the resulting identifier as non-standard.

Creation options used for generation of a particular non-standard InChI may be appended to the created identifier, see below.

InChI creation options are listed in the following table.

Table 3. InChI creation options.

InChI creation option	Meaning	Default behavior (if no option supplied)
SUU	Always indicate unknown/undefined stereo	Does not indicate unknown/undefined stereo unless at least one defined stereocenter is present
SLUUD	Stereo labels for "unknown" and "undefined" are different, 'u' and '?', resp. (new option; see explanation)	Stereo labels for "unknown" and "undefined" are the same ('?')
RecMet	Include reconnected metals results	Do not include
FixedH	Include Fixed H layer	Do not include
KET	Account for keto-enol tautomerism (experimental extension to InChI v. 1)	Ignore keto-enol tautomerism
15T	Account for 1,5-tautomerism (experimental extension to InChI v. 1)	Ignore 1,5-tautomerism

The standard InChI is always generated if no InChI creation/stereo modification options are specified. This means:

- include tautomerism (i.e., turn mobile H perception on, exclude "fixed hydrogen atoms" layer) except for keto-enol and 1,5-tautomerism;
- omit reconnection of bonds to metal atoms;
- only the narrow end of a wedge points to a stereocenter;
- exclude unknown/undefined stereo if no other stereo is present;
- treat stereochemistry as absolute (not relative or racemic).

Inversely, if any of SUU/SLUUD/RecMet/FixedH/Ket/15T/SRel/SRac/SUCF options are specified in the command line, the generated InChI will be non-standard.

Saving InChI creation options

Since the software v. 1.03, the command-line option "/SaveOpt" ("-SaveOpt" under Linux) was introduced. It allows one to append saved InChI creation options to a non-standard InChI string.

The "SaveOpt appendix" currently consists of the two capital Latin letters which are separated from the InChI string by a backslash '\'. Note that this appendix is not considered as an integral part (layer) of InChI itself; rather, it is an optional complement. It may or may not be present after the end of an InChI string (by default – no "SaveOpt" option – it is absent). To signify this, the appendix is separated from the previous sequence of symbols by a character which may not appear in any other place, a backslash.

Note also that the InChI generation option "/SaveOpt" (and the saved-options appendix) is not available for standard InChI as the latter is always created with the same options.

As for the encoding of saved options, the first SaveOpt letter encodes whether RecMet/FixedH/SUU/SLUUD switches were activated. Each of them is a binary switch ON/OFF, giving a total of 2*2*2*2=16 values which are encoded by capital letters 'A' through 'P'.

The second letter encodes experimental (InChI 1 extension) options KET and 15T. Each of these options is a binary switch ON/OFF, so there are 2*2=4 combinations, encoded by ‘A’ through ‘D’. Note that anything but 'A' here would indicate "extended" InChI 1. Note that here is some reservation for future needs: the 2nd memorization character may accommodate two more ON/OFF binary options (at 26-base encoding).

The exact encoding scheme is specified in the tables below.

Table 4. Meaning of the 1st SaveOpt letter.

Letter	RecMet	FixedH	SUU	SLUUD
A	OFF	OFF	OFF	OFF
B	OFF	OFF	OFF	ON
C	OFF	OFF	ON	OFF
D	OFF	OFF	ON	ON
E	OFF	ON	OFF	OFF
F	OFF	ON	OFF	ON
G	OFF	ON	ON	OFF
H	OFF	ON	ON	ON
I	ON	OFF	OFF	OFF
J	ON	OFF	OFF	ON
K	ON	OFF	ON	OFF
L	ON	OFF	ON	ON
M	ON	ON	OFF	OFF
N	ON	ON	OFF	ON
O	ON	ON	ON	OFF
P	ON	ON	ON	ON

Table 5. Meaning of the 2nd SaveOpt letter.

Letter	Ket	15T
A	OFF	OFF
B	OFF	ON
C	ON	OFF
D	ON	ON

Examples:

```
InChI=1/C9H11NO2.Na/c1-3-5(7(3)9(10)12)6-4(2)8(6)11;/h5-
6,11H,1-2H3,(H2,10,12);/q;+1/p-
1/t5?,6?;/i/hD/fC9H10NO2.Na/h11h,10H2;/q-1;m/i10D;\OA
```

(this identifier was created with options /RecMet /FixedH /SUU and /SaveOpt)

```
InChI=1/C9H11NO2.Na/c1-3-5(7(3)9(10)12)6-4(2)8(6)11;/h5-
6,11H,1-2H3,(H2,10,12);/q;+1/p-1/t5?,6?;/i/hD\KA
```

(this identifier was created for the same input structure with options /RecMet /SUU and /SaveOpt)

```
InChI=1S/C9H11NO2.Na/c1-3-5(7(3)9(10)12)6-4(2)8(6)11;/h5-
6,11H,1-2H3,(H2,10,12);/q;+1/p-1/i/hD
```

(this identifier was created for the same input structure with no InChI creation options)

The next table summarizes the availability of various options in the various parts of the InChI software.

Table 6. Availability of InChI software options.

Options availability			Command line option (without / or – prefix)	Explanation
winchi	inchi	API		
Input				
–	Yes	–	STDIO	Use standard input/output streams
–	Yes	–	InpAux	Input structures in InChI default aux. info format (for use with STDIO)
Yes	Yes	Yes	SDF:*name*	Read from the input SDfile the ID under the named data header
–	Yes	–	AMI	Allow multiple input files
Output				
–	Yes	Yes	AuxNone	Do not produce Auxiliary Information
–	Yes	–	NoLabels	Omit structure number, DataHeader and ID from InChI output
–	Yes	Yes	SaveOpt	Save custom InChI creation options
–	Yes	–	Tabbed	Separate structure number, InChI, and AuxIndo with tabs
Always	Yes	–	D	Display the structure
Yes	Yes	–	Equ	Display sets of identical components

-	Yes	-	F*number*	Set display font size (points)
-	Yes	Yes	OutputSDF	Convert InChI created with default auxiliary info to a SDfile
-	Yes	Yes	SdfAtomsDT	Output Hydrogen Isotopes to SDfile as Atoms D and T
-	Yes	-	AMIOutStd	Write output to stdout (in AMI mode only)
-	Yes	-	AMILogStd	Write log messages to stderr (in AMI mode only)
-	Yes	-	AMIPrbNone	Suppress creation of problem files (in AMI mode only)
Structure perception				
Yes	Yes	Yes	NEWPSOFF	Both ends of wedge point to stereocenters
-	Yes	Yes	DoNotAddH	Do not add H according to usual valences
Yes	Yes	Yes	SNon	Ignore stereo information in input structures
Stereo perception modifiers (non-standard InChI)				
Yes	Yes	Yes	SRel	Relative stereo
Yes	Yes	Yes	SRac	Racemic stereo
Yes	Yes	Yes	SUCF	Use Chiral Flag: On means Absolute stereo, Off - Relative
Customizing InChI creation (non-standard InChI)				
Yes	Yes	Yes	SUU	Always include omitted unknown/undefined stereo
Yes	Yes	Yes	SLUUD	Make labels for unknown and undefined stereo different
Yes	Yes	Yes	RecMet	Include reconnected metals results
Yes	Yes	Yes	FixedH	Include Fixed H layer
Yes	Yes	Yes	KET	Account for keto-enol tautomerism (experimental)
Yes	Yes	Yes	15T	Account for 1,5-tautomerism (experimental)
Generation				
60 sec	Yes *)	Yes*)	Wnumber	Set time-out per structure in seconds
-	Yes	Yes	WarnOnEmptyStructure	Warn and produce empty

				InChI for empty structure
Always	Yes	- **)	Key	Generate InChIKey
Always	Yes	- **)	XHash1	Generate hash extension (to 256 bits) for 1st block of InChIKey
Always	Yes	- **)	XHash2	Generate hash extension (to 256 bits) for 2nd block of InChIKey
Conversion				
-	Yes	-	InChI2Struct	Convert standard InChI string(s) into structure(s)
-	Yes	-	InChI2InChI	Convert InChI string(s) into InChI string(s)

*) W0 means unlimited time. In InChI Library the default is W0, in inchi the default is 60 seconds (W60).
**) In InChI Library, generation of InChIKey/hash extensions is performed via a separate API call.

Test Files

A number of Molfiles (*.mol) and two SDfiles (*.sdf) are included with the program for illustrative purposes. Some Molfiles contain more than one fragment – each may be viewed separately using the 'combo-box' on the upper left of the screen. Multiple structures are given in the SDfiles, which may be viewed in order by pressing the 'Next Structure' (">>") and 'Previous Structure' ("<<") buttons. File Samples.sdf contains all of the individual Molfiles from Samples.zip. These SDfiles contain names of the structures. To display them enter word "name" (without quotes) in "Structure ID Header" field (Fig. 2).

IV. CHEMICAL STRUCTURE INPUT

Molfiles or the program output produced with the "Full auxiliary information" option may be used for input. Molfile structures may be submitted either as a single Molfile or as a series of concatenated Molfiles (an SDfile). A number of programs, some of them freely available, may be used to create these Molfiles. Information on how to produce and

convert If an input structure contains more than one independent structure, each component is individually shown in the graphical output section of the program, though this has no effect on the InChI. Text results are given for all layers and all components (different components of a single substance are separated by semicolons in each layer, except for chemical formulas, which, by convention, are separated by dots.).

While structure normalization methods built into the program perceive a range of different structure drawing conventions, it is possible that other conventions may not be properly recognized. Examination of the graphical results of InChI processing, especially for equivalent atom classes and stereo labeling, should reveal such problems.

If an SDfile is 'labeled', the program can supply these labels in its output. If the tag name is 'Name' and the data field is '2-methylanthracene', this information would appear in the SDfile as 3 lines (the last line is blank):

```
>    <Name>
 2-methylanthracene
```

In this case, if the tag 'Name' is entered in the 'Structure ID Header' field in the input dialog box, '2-methylanthracene' will appear in the output text.

A variety of structure files are provided for testing. Individual Molfiles have extension .MOL, concatenated Molfiles have extension .SDF.

V. InChI AND InChIKey BY EXAMPLE

The InChI graphical program parses and annotates the InChI and associated auxiliary information and displays it in the textual output region. An understanding of this information requires an understanding of InChI layering, which is described in detail in the Technical Manual. A summary is presented here for understanding program output.

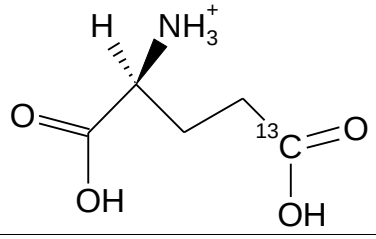


Figure 15. (*S*)-Glutamic acid

To provide an example of some of the InChI layers for a "real" molecule, we have chosen the structure of isotopically substituted (*S*)-glutamic acid in Figure 15 above for illustrative purposes.

Figure 16 shows InChI, AuxInfo and InChIKey. As no specific generation options have been activated, standard InChI and InChIKey were generated by default.

Structure: 1
InChI=1S/C5H9NO4/c6-3(5(9)10)1-2-4(7)8/h3H,1-2,6H2,(H,7,8)(H,9,10)/p+1/t3-/m0/s1/i4+1

AuxInfo=1/1/N:5,6,2,7,1,4,8,9,10,11/E:(7,8)(9,10)/it:im/I:/E:m/rA:11nCCHN+CCC.i13OOOO/rB:s1;N2;P2;s2;s5;s6;s7;d7;d1;s1;/rC:6.1671,-19.3365,0;7.0125,-18.4864,0;6.4113,-17.4485,0;7.6089,-17.4485,0;7.8578,-19.3318,0;8.891,-18.7306,0;9.7363,-19.576,0;9.7316,-20.7735,0;10.8916,-19.266,0;5.0071,-19.0265,0;6.1624,-20.534,0;

InChIKey=WHUUTDBJXJRKMK-MYXYCAHRSA-O
XHash1=96b009f08de5ce91a1c59d8a61fce52decfaadc669291edb
XHash2=80cd24563259e8848f0436b53b223970c42fcc8326e9858ecf148d78

Figure 16. Standard InChI, AuxInfo and InChIKey/hash extensions for (*S*)-glutamic acid.

Figure 17 shows the input structure display. Figure 18 – "Preprocessed" – shows the result of the preprocessing – an attempt to eliminate charges with purpose to reduce different protonation forms to one. Figure 19 shows the result of the structure analysis by the InChI algorithm.

Figure 20 shows the results with "Mobile H Perception" turned off (this `winchi-1` option corresponds to the command line option `FixedH`). Notice that the content of the text window has changed: a string that starts with "/f" has been appended to InChI. Also, notice that both InChI and InChIKey became non-standard (no 'S' flag in InChI, 'N' flag instead of 'S' in InChIKey) as a non-standard InChI creation option has been used.

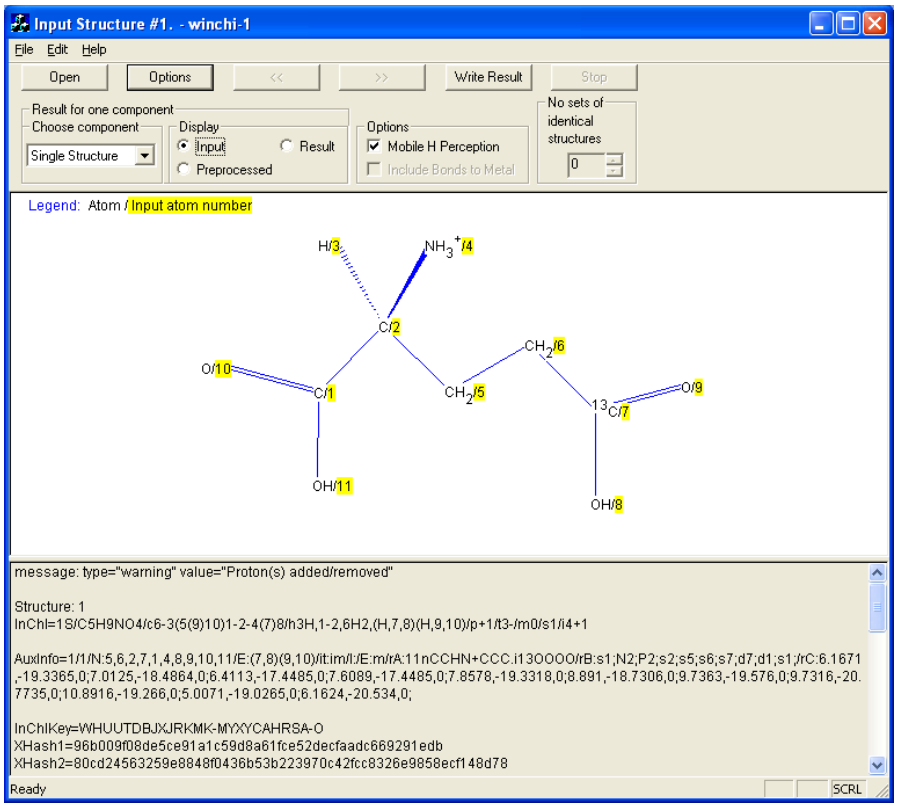


Figure 17.

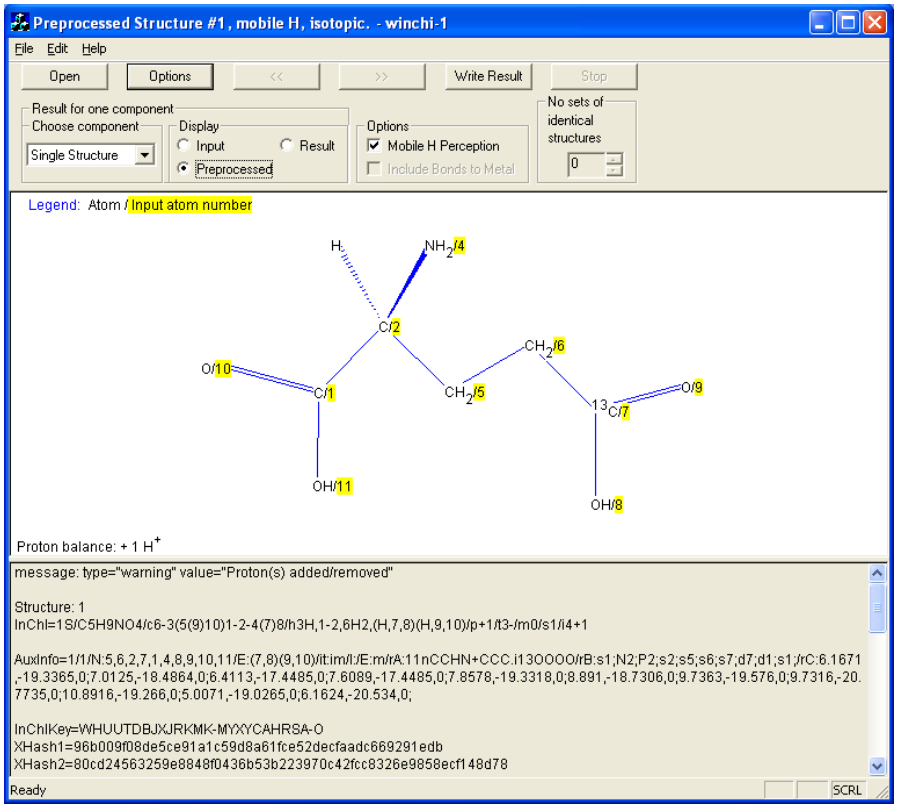


Figure 18.

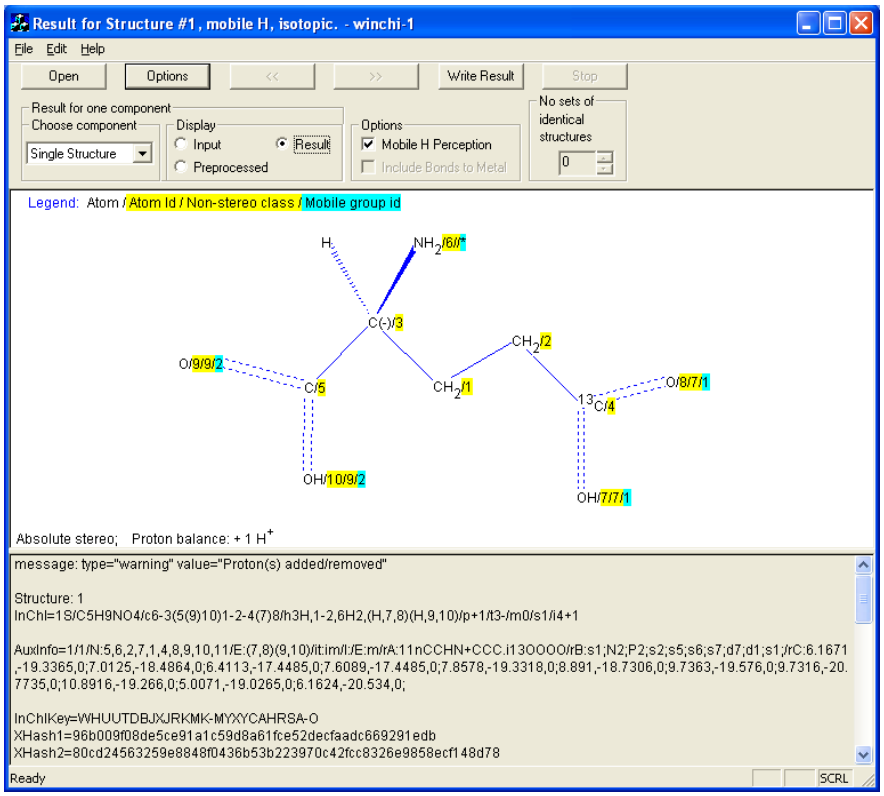


Figure 19.

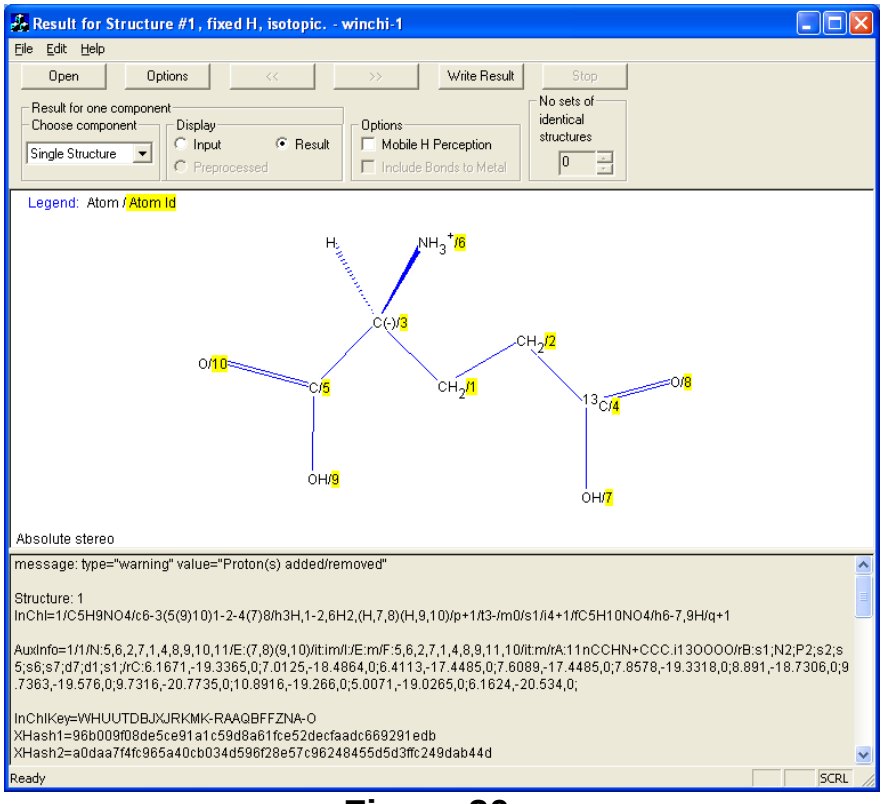


Figure 20.

Figure 21 shows the full contents of the text output window in the case of "Mobile H Perception" turned off.

The "InChI ANNOTATED CONTENTS" provides annotations to each item of the Identifier and Auxiliary information. Note that the Auxiliary information is not a part of the Identifier.

```
message: type="warning" value="Proton(s) added/removed"

Structure: 1
InChI=1/C5H9NO4/c6-3(5(9)10)1-2-4(7)8/h3H,1-2,6H2,(H,7,8)(H,9,10)/p+1/t3-/m0/s1/i4+1/fC5H10NO4/h6-7,9H/q+1

AuxInfo=1/1/N:5,6,2,7,1,4,8,9,10,11/E:(7,8)(9,10)/it:im/l:/E:m/F:5,6,2,7,1,4,8,9,11,10/it:m/rA:11nCCHN+CCC.i13OOOO/rB:s1;N2;P2;s2;s5;s6;s7;d7;
d1;s1;/rC:6.1671,-19.3365,0;7.0125,-18.4864,0;6.4113,-17.4485,0;7.6089,-17.4485,0;7.8578,-19.3318,0;8.891,-18.7306,0;9.7363,-19.576,0;9.7316,-
20.7735,0;10.8916,-19.266,0;5.0071,-19.0265,0;6.1624,-20.534,0;

InChIKey=WHUUTDBJXJRKMK-RAAQBFFZNA-O
XHash1=96b009f08de5ce91a1c59d8a61fce52decfaadc669291edb
XHash2=a0daa7f4fc965a40cb034d596f28e57c96248455d5d3ffc249dab44d

==== InChI ANNOTATED CONTENTS ====

Structure: 1

InChI=
{version}1
/{formula}C5H9NO4
/c{connections}6-3(5(9)10)1-2-4(7)8
/h{H_atoms}3H,1-2,6H2,(H,7,8)(H,9,10)
/p{protons}+1
/t{stereo:sp3}3-
/m{stereo:sp3:inverted}0
/s{stereo:type (1=abs, 2=rel, 3=rac)}1
/i{isotopic:atoms}4+1
/f{fixed_H:formula}C5H10NO4
/h{fixed_H:H_fixed}6-7,9H
/q{fixed_H:charge}+1

AuxInfo=
{version}1
/{normalization_type}1
/N:{original_atom_numbers}5,6,2,7,1,4,8,9,10,11
/E:{atom_equivalence}(7,8)(9,10)
/it:{abs_stereo_inverted:sp3}im
/l:{isotopic:original_atom_numbers}
/E:{isotopic:atom_equivalence}m
/F:{fixed_H:original_atom_numbers}5,6,2,7,1,4,8,9,11,10
/it:{fixed_H:abs_stereo_inverted:sp3}m
/rA:{reversibility:atoms}11nCCHN+CCC.i13OOOO
/rB:{reversibility:bonds}s1;N2;P2;s2;s5;s6;s7;d7;d1;s1;
/rC:{reversibility:xyz}6.1671,-19.3365,0;7.0125,-18.4864,0;6.4113,-17.4485,0;7.6089,-17.4485,0;7.8578,-19.3318,0;8.891,-18.7306,0;9.7363,
-19.576,0;9.7316,-20.7735,0;10.8916,-19.266,0;5.0071,-19.0265,0;6.1624,-20.534,0;
```

Figure 21. The Identifier, InChIKey, Auxiliary information, Annotated Identifier and Auxiliary information for (*S*)-glutamic acid with "Mobile H Perception" turned off (non-standard).

Notes:

- Since the displayed InChI is non-standard, it has the "InChI=1/" prefix.
- The Auxiliary information is not a part of the Identifier.
- Since the displayed InChIKey is non-standard, it contains the flag letter 'N':
  `InChIKey=WHUUTDBJXJRKMK-RAAQBFFZ`**`N`**`A-O`
- The InChIKey contains the protonation flag not equal to 'N' (the last character of the InChIKey is 'O') which indicates that the "p" segment of InChI string is not empty.
- InChIKey is followed by the two hash extensions, XHash1 and XHash2.
  They represent the rest of 256-bit SHA-2 signature for the 1st and 2nd blocks, resp.

The InChI represents the structure of a covalently bonded compound in four distinct 'layers' described below.

1. Main Layer

1.1 Chemical Formula

This is a conventional Hill-sorted formula with components separated by periods (dots). In the example in Figure 21, the formula is:

```
/{formula}C5H9NO4
```

1.2 Connections

Defines the covalent bonds between atoms in the structure. It is partitioned into as many as three sublayers: H-atoms omitted, immobile H-atoms included and, mobile H-atoms included.

In the example in Figure 21, the connections are:

```
/c{connections}6-3(5(9)10)1-2-4(7)8
/h{H_atoms}1-2H2,3H,6H2,(H,7,8)(H,9,10)
```

where part `(H,7,8)(H,9,10)` is responsible for mobile H

2. Charge Layer

This simply represents net charge, and may appear in two sublayers. Unlike other layers, this layer is independent of all others and when omitted indicates that the charge is not specified.

`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 by the formal positions of their electrons.

2.2 Protons

Number of protons removed from or added (if the number is negative) to the substance to make same components with variable protonation (e.g. amino acids) identical.

In the example in Figure 21 the proton(s) are:

```
/p{protons}+1
```

Note that InChIKey indicates this as protonation: the last character is 'O' which corresponds to "`p+1`".

3. Stereochemical Layer

This layer is composed of two sublayers. The first accounts for double bond, sp^2, and the second for sp^3 tetrahedral stereochemistry and allenes. The latter stereo descriptions are first given for relative stereochemistry only, followed by a designation of whether the absolute stereochemistry is required (and if this requires inversion of the relative stereochemistry).

In the example in Figure 21 the stereo layer is:

```
/t{stereo:sp3}3-
/m{stereo:sp3:inverted}0
/s{stereo:type (1=abs, 2=rel, 3=rac)}1
```

4. Isotopic Layer

This is a layer in which different isotopically labeled atoms are distinguished from each other. Mobile isotopic hydrogen atoms are listed separately. The layer also holds any changes in stereochemistry created by the presence of isotopic atoms.

In the example in Figure 21 the isotopic layer is:

```
/i{isotopic:atoms}4+1
```

5. Fixed-H Layer

This layer (which may be present only in non-standard InChI), provides the location of H-atoms considered mobile in earlier layers along with any needed changes to earlier layers.

In the example in Figure 21, the Fixed-H layer is:

```
/f{fixed_H:formula}C5H10NO4
/h{fixed_H:H_fixed}6-7H,9H
/q{fixed_H:charge}+1
/t{fixed_H:stereo:sp3}m
/m{fixed_H:stereo:sp3:inverted}0
/s{fixed_H:stereo:type (1=abs, 2=rel, 3=rac)}1
/i{fixed_H:isotopic:atoms}m
```

Note that these names of layers are used in the annotated InChI output. In the identifier itself the layers are preceded by two characters, a '/' followed by a letter.

For any input structure, the first layer will always be generated. Other layers will appear only when the input structure contains the associated information. For instance, if *Z/E* stereochemistry, but not sp^3 stereochemistry is entered, only the *Z/E* (sp^2) stereochemistry sublayer will be represented.

The contents of a layer may depend on prior layers. For instance, the stereochemical layer uses identification numbers of atoms defined in the formula layer.

The Charge layer is simply the overall charge of the component, hence is independent of the other layers. It is possible to extend this layer by adding other 'whole molecule' attributes, such as electronically excited state, vibrational/rotational state and state of aggregation (phase).

The Protons layer refers to the entire structure. The specific state of protonation (or deprotonation) may be ignored by omitting this layer.

VI. OUTPUT TEXT FORMAT

The text output from the InChI program is written in plain text format as described below. This text is visible in the lower region of the main window and in the text file generated by selecting 'Write Result' in the main window.

Note that InChI for a substance is strictly defined as a string of characters composed of a series of text fields. The specific text format described here is meant only for those interested in the details of the representation and is not required for effective use of the InChI (for more details consult the Technical Manual).

The actual fields present in a given representation will depend on the information present in the input structure and the intent of the structure author. If, for instance, it is desired to represent a structure with mobile H-atoms, a fixed H-atom layer is not generated. If a structure cannot have stereoisomers, no stereo layers will be present.

All text output originating from a single chemical substance input (structure file) is provided in up to four lines; all lines except the second one may be suppressed:

```
Structure NUMBER. STRUCTURE_ID_HEADER =VALUE
InChI=1/… or InChI=1S/…
AuxInfo=1/…
InChIKey=…
```

where NUMBER is the sequence number of the structure in the input file. When Molfiles or SDfiles are used and a "STRUCTURE_ID_HEADER" has been entered in the

"Structure ID Header" field in the input dialog box (see Chemical Structure Input section above), VALUE represents the contents of that field. If the field was left blank then STRUCTURE_ID_HEADER=VALUE is omitted. The output AuxInfo line is optional. The Tabbed option produces output of the same items merged in one line with tab characters as separators.

InChI string

Following the /? InChI delimited tags are individual layer values. Curly braces contain annotations (the values for each layer follow the closing curly brace).

```
Main Layer (immediately follows the InChI version)
/{formula}
/c{connections}
/h{H_atoms}
Charge layer
/q{charge}
/p{protons}
Stereo layer
/b{stereo:dbond}
/t{stereo:sp3}
/m{stereo:sp3:inverted}
/s{stereo:type (1=abs, 2=rel, 3=rac)}
Isotopic Layer
/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)}
```

The Main Layer is divided into three layers: formula layer, connections layer, and H atoms layer. The Stereo layer is divided into four layers: stereo double bond, sp^3 stereochemistry, sp^3 inversion flags, and type of sp^3 layer.

A description of the contents of each of these layers follows.

Main Layer

This provides the elemental composition and connectivity of the structure. This layer, which is always present, is subdivided into several segments. The first segment is a conventional chemical formula, which also provides the InChI identification numbers used for each atom. These numbers are determined by the sequence numbers of elements

in the chemical formula (excluding H). In the formula each element is represented by the form 'El#', where 'El' is the element symbol and '#' is the number of atoms. For example, in case of C2H6O two atoms C have identification numbers 1 and 2 and atom O (3rd non-H atom) has number 3. Atoms H are not given any identification numbers, except for bridging atoms H which, when present, are given the highest identification numbers. When a given component is present multiple times, this formula may be preceded by this number of occurrences. The case of H+ is special – it is represented simply as '1' in the Protons layer (the formula and connections segments are empty).

As noted above, the position of each element in the first (formula) segment is used as its identification number. These numbers are used in the second segment of the InChI, connections (/c), to indicate bonding partners. To illustrate, in this segment isobutane (C4H10) is represented as "1-4(2)3", which means that the 1st atom listed is bonded to the 4th, the 4th is bonded to the 2nd and the 3rd atoms. If the connections segment is empty (for example, in case of methane) it is omitted entirely.

The 3rd segment, the hydrogen layer (/h), describes positions of hydrogen atoms attached to the molecular skeleton described by formula and connections layers. For isobutene, "1-3H3,4H" means that each of atoms 1 to 3 has 3 H, and atom 4 has one H. A mobile H may migrate between different atoms. For example, acetic acid (C2H4O2) has connections "1-2(3)4" and a hydrogen layer "1H3,(H,3,4)". Parentheses contain the number of mobile H (one in this case) and the identification numbers of the atoms that share these mobile H atoms (3 and 4).

Isotopic layer

Isotopic layers consist of a series of isotopic atoms with their identification numbers. Specific isotopes are represented by integers giving their atomic mass relative to the rounded average atomic mass of the element. For example, if atom number 6 is ^{37}Cl, it is represented as 6+2 (average atomic mass of Cl is 35.453, rounding to the nearest integer gives 35, 37 – 35 = +2).

Hydrogen isotopes are exceptions to this labeling rule. They are explicitly denoted as *a*D*n* (deuterium) and *a*T*n* (tritium), where *a* is the identification number of the atom to which they are attached and *n* is the number of these atoms attached to the a^{th} atom; *n* = 1 is omitted. Isotopic hydrogen atoms that are mobile or belong to atoms that are recognized as proton donors or acceptors are considered to belong to the whole substance and shown in the /h (exchangeable_H) segment of the isotopic layer.

Stereo layer

The stereo layer expresses the 'parity' of the atoms and bonds that define the stereochemistry. This layer is divided into four sub-layers, the first, dbond (/b), provides double bond (*Z/E*) stereochemistry, the second, sp3 (/t), represents sp3 tetrahedral stereochemistry and allenes, the third, sp3:inverted (/m), is present only in case of absolute configuration, the fourth (/s) describes whether the stereo representation is absolute, relative or racemic. Future versions of the InChI may add other forms of stereochemistry. Note that standard InChI supports only absolute stereochemistry or the absence of it in case of non-chiral chemical structures.

The stereo label of a double bond is represented in the format *a*-*b*X, where *a* and *b* are the identification numbers of the bonded atoms (*a*>*b*) and X is a parity label, with the possible values: +, -, or ?. The + and - labels indicate that the stereochemical configuration has been defined, however these values only have meaning relative to the atom identification numbers assigned in the labeling process. The numbers do not coincide with CIP priorities. If, for example, atoms in a similar structure were given different labels, the parity might change even if a chemist might consider the stereochemistry to be the equivalent. Additional rule-based processing would be needed to label a bond as '*Z*' or '*E*', for example. By default, a question mark ('?') indicates that stereochemistry either has not been specified or explicitly entered as 'unknown'. However, in a non-standard InChI generated with option 'SLUUD' turned On, the symbol 'u' is used to indicate explicitly entered 'unknown' stereo.

Labels for sp^3 stereochemistry are expressed in the format *n*X where *n* is the identification number of the atom and X is the parity, as computed by InChI. The parity is

allowed the same values as discussed for double bond stereochemistry. Also, as for double bond stereochemistry, parity values themselves depend on the particular labeling of the structure and are not readily converted to standard CIP notation. Currently the user may request absolute, relative, and racemic sp3 stereo (the last two choices are not available in standard InChI). In case of absolute stereo the algorithm processes both the input structure and inverted structure; after that "the smallest" sp3 layer is chosen. Therefore enantiomers have an identical sp3 section. The fact of choosing the inverted configuration is shown as 1 in sp3:inverted (/m) segment, otherwise there is 0 or period if inversion does not bring a change. The type of stereo is shown in /s segment as /s1 (absolute); non-standard InChI also uses /s2 (relative) and /s3 (racemic).

Fixed-H layer

The fixed-H (/f) layer follows (it may appear only in non-standard InChI). This layer adds information required to fix the positions of all mobile H atoms. It is structured the same as the main layer, except for absent connections and H_fixed replacing H_atoms layer. For example, H_fixed for acetic acid is 3H: a single mobile H position is fixed at atom 3. Negative values should be subtracted from corresponding H_atoms; in this case lowercase h is used.

Layer transposition

The order of the components in the main section of the identifier may differ from the order in fixed-H section. This is shown in the /o (transposition) segment. The transposition usually occurs because in the sorted order of the components constitution has higher priority than stereo and isotopic layers.

Mobile-H Limitations

Not all possible forms of tautomerism are represented. Specifically, when there are no charged heteroatoms (normalization_type, the first segment in the Auxiliary Info, is 1) this version perceives 1,3 (and limited 1,4 and 1,5)-H-atom transfer as well as 1,2-H-atom migration in 5-membered rings. In the case of charged heteroatoms the detection of mobile atoms H and removal of protons may be 'aggressive' (see InChI Technical

Manual). Note that the mobility of H-atoms depends on the environment of a substance as well as its structure, hence can at best be a useful approximation. The definitions used by InChI were intended to represent, to the degree possible, a common current practice.

InChIKey string

InChIKey has five distinct components.

(1) 14-character hash of the basic (Mobile-H) InChI layer. It encodes molecular skeleton (connectivity);

(2) 8-character hash of the remaining layers (except for the "/p" segment, which accounts for added or removed protons: it is not hashed at all; the number of protons is encoded at the end of the InChIKey). It encodes stereochemistry and isotopic substitution information, associated with molecular connectivity expressed by the first block. In case of non-standard InChI, it also encodes information on the exact position of tautomeric hydrogens (if any), as well as on the related stereo/isotopic data.

(3) 1 flag character,

(4) 1 version character

(5) the last character is a [de]protonation indicator.

All symbols of InChIKey except the delimiter (a dash, that is, a minus) are uppercase English letters representing a "base-26" encoding.

The overall length of InChIKey is fixed at 27 characters, including separators (dashes):
`AAAAAAAAAAAAAA-BBBBBBBBFV-P`

Here

(1) `AAAAAAAAAAAAAA` is a 14-character first hash block.

(2) `BBBBBBBB` is an 8-character second hash block.

(3) `F` is a flag indicating kind of InChIKey: it has either the value 'S' for standard InChIKey or the value 'N' for non-standard.

(4) `V` is a character indicating InChI version: 'A' for version 1 (current), 'B' for version 2, etc.

(5) `P` is an indicator for the number of protons; this number is not encoded in the hash but is indicated as a separate 2-character block at the end, where one character is a hyphen, as -N for neutral, -M for -1 hydrogen, -O for +1 hydrogen, etc.

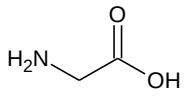
The exact layout is presented in the following table.

Char	Protons	Char	Protons
N	0		
M	-1	O	+1
L	-2	P	+2
K	-3	Q	+3
J	-4	R	+4
I	-5	S	+5
H	-6	T	+6
G	-7	U	+7
F	-8	V	+8
E	-9	W	+9
D	-10	X	+10
C	-11	Y	+11
B	-12	Z	+12
A	< -12 or > +12		

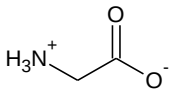
InChIKey inherits some layered structure from InChI. The first block is always the same for the same molecular skeletons. All isotopic substitutions, changes in stereoconfiguration, tautomerism and protonization are reflected in the second block. Note that, by definition, standard InChIKey, like standard InChI, does not account for tautomerism and may indicate only absolute stereo. It also does not account for the original structure's bonds to metal, if they were present and disconnected on standard InChI generation.

Note also that different protonation states of the same compound will have InChIKeys which differ only by the last character, the protonation flag (unless the both states have number of inserted/removed protons > 12; in this case the protonation flag will also be the same, 'A'). Additionally, by design of standard InChI/InChIKey, different tautomers of the same compound (as far as their particular tautomerism is perceived by InChI) will have the same standard identifiers.

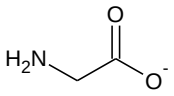
As an example, shown below are standard InChIKeys as well as standard InChI strings for neutral, zwitterionic, anionic and cationic states of glycine (its neutral and zwitterionic states do not differ in total number of protons so they have the same InChI/InChIKey):



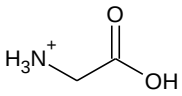
InChI=1S/C2H5NO2/c3-1-2(4)5/h1,3H2,(H,4,5)
InChIKey=DHMQDGOQFOQNFH-UHFFFAOYSA-N



InChI=1S/C2H5NO2/c3-1-2(4)5/h1,3H2,(H,4,5)
InChIKey=DHMQDGOQFOQNFH-UHFFFAOYSA-N



InChI=1S/C2H5NO2/c3-1-2(4)5/h1,3H2,(H,4,5)/p-1
InChIKey=DHMQDGOQFOQNFH-UHFFFAOYSA-M



InChI=1S/C2H5NO2/c3-1-2(4)5/h1,3H2,(H,4,5)/p+1
InChIKey=DHMQDGOQFOQNFH-UHFFFAOYSA-O

Auxiliary Information Output

A variety of additional information is optionally provided along with the Identifier: a mapping of canonical identification atom numbers on original atom numbers, constitutional equivalence, inverted sp3 stereo and its numbering, isotopic and fixed-H layer information, and 'reversibility' information which allows the redrawing of the original structure and recalculation of the identifier. This additional analysis information is shown in the line that starts with `AuxInfo=`

```
AuxInfo=
{version}1
/{normalization_type}
```
Main part
```
/N:{original_atom_numbers}
/E:{atom_equivalence}
/gE:{group_equivalence}
/it:{abs_stereo_inverted:sp3}
/iN:{abs_stereo_inverted:original_atom_numbers}
```
Isotopic part
```
/I:{isotopic:original_atom_numbers}*
/E:{isotopic:atom_equivalence}
/gE{isotopic:group_equivalence}
```

```
/it:{isotopic:abs_stereo_inverted:sp3}
/iN:{isotopic:abs_stereo_inverted:original_atom_numbers}
```
Reversibility part
```
/CRV:{charge_radical_valence}
/rA:{reversibility:atoms}
/rB:{reversibility:bonds>
/rC:{reversibility:xyz}
```

The original number of an atom with identification number *n* is given as the n^{th} member of this list for a component; the lists are separated with ";".

Classes of equivalent atoms or groups are given as lists of identification numbers within parentheses.

Inverted absolute sp3 stereo provides the stereo layer of the inverted (reflected in a mirror) substance.

Unusual valences, atomic charges, and radical locations in the input data are shown in the charge-radical-valence (/CRV:) section. Together with the identifier this information allows to reconstruct a representative of a set of structures each of which produces the same identifier. The examples are: 22+1, 22.3, 22+1.3, 22d, 22d3, where 22 is the atom identification number, +1 is charge, 3 is valence, d is radical-doublet (t = triplet, s = singlet).

Upon requesting "Full auxiliary information" (always ON in the `winchi` program) the reversibility section is added; it includes all input information that allows the display of the input structure and the calculation of the identifier. This reversibility information is not used in InChI2Struct conversion.

The Identifier of a reconnected structure (bonds to metals are disconnected by default), if requested (non-standard InChI only), is separated by /r and may contain all layers describer earlier.

Error/Warning Output

If problems are encountered during the processing of a structure, they are shown in the first line of the `winchi-1` text window or in the `inchi-1` log file.

In the structure display, stereogenic atoms that caused warnings "Ambiguous stereo: center(s)" are displayed in red as well as atoms that caused warnings "Ambiguous stereo: bond(s)" concerning stereogenic bonds. Parities of these stereogenic elements are also displayed in red.

VII. PRINTING

The upper or lower sections of the output display may be printed by pressing the RIGHT mouse button with the cursor over the section and then selecting the print option. Text in the lower section may be copied using standard Windows controls.

VIII. OTHER OUTPUT FILES

In addition to the standard InChI output file discussed above (extension .txt), selection of the 'Write Results' option generates two other files that use the same base name as the input structure file. One is a .log file that records the progress of the program. The other is a .prb file that records processing problems. We would appreciate being sent a copy of these files if problems are encountered with program operation.

IX. SOURCE CODE

The basic InChI generation code is written in the 'C' language and the user interface code of `winchi-1` is written in C++ using Microsoft Foundation Classes. All 'C' language source code, including Microsoft Visual C++ project files and gcc makefiles, is available through the links at http://www.iupac.org/inchi and http://www.inchi-trust.org.

X. CONTACT INFORMATION

International Union of Pure and Applied Chemistry (IUPAC)
InChI Subcommittee

Steve Heller (Subcommittee Chair)
steve@hellers.com

Igor V. Pletnev
igor.pletnev@gmail.com

Appendix. InChI Software Warning and Error Messages

Two varieties of problems detected during processing are reported. Warnings provide processing information that show any ambiguities in the input structure or special actions taken during processing. An InChI will be produced. When an error is generated, a valid InChI cannot be produced due to invalid input. It is expected that additional errors and warnings will reported in the final version.

Notes:

1. Messages ending with ":…" are followed by additional information
2. Symbol # represents an integer

Types of Warnings/Errors

- Input structure warnings
- Input structure errors
- InChI calculation errors
- Reading Molfile warning messages
- Reading Molfile error messages
- Reading pre-existing InChI output errors
- Internal errors (possible software error)

List of InChI warning and error messages

Input structure warnings

```
"Proton(s) added/removed"
"Charges neutralized"
"Omitted undefined stereo"
"Ambiguous stereo: [center(s)][bond(s)]"
"Unusual valence(s):…"
"Charges were rearranged"
"Salt was disconnected"
"Metal was disconnected"
"Not chiral"
```

Input structure errors

```
"Unknown element(s):…"
"Bond to nonexistent atom"
"Multiple bonds between two atoms"
"Atom has more than 3 aromatic bonds"
"Too many atoms"
"Empty structure"
"Atom ‘X’ has more than 20 bonds" (X is the chemical
element symbol)
```

InChI calculation errors

```
"Output buffer overflow"
"Cannot process free radical center"
"Time limit exceeded"
"User requested termination"
```

Reading Molfile warnings

```
"Too long counts line"
"Too long atom block line"
"Too long properties block line"
"Charge not recognized:…"
"Radical not recognized:…"
"Isotopic data not recognized:"
"Too long SData line truncated" (SData line was truncated
to 200 characters)
```

Reading Molfile errors

```
"Unrecognized bond type:#"
"Unrecognized bond stereo"
"Program error interpreting MOLfile"
"Unknown error"
"Cannot read counts line"
"Cannot interpret counts line:…"
"Cannot read atom block line"
"Cannot interpret atom block line:…"
"Cannot read bond block line"
"Cannot interpret bond block line:…"
"Cannot read STEXT block line"
"Cannot read properties block line"
"Unexpected SData header line
"Bypassing to next structure"
```

Reading pre-existing InChI output errors

```
"Missing atom data"
"Wrong atoms data"
"Wrong number of atoms"
"Wrong bonds data"
"Wrong bond type"
"Wrong number of bonds"
"Missing atom coordinates data"
"Wrong atom coordinates data"
"Wrong number of coordinates"
"Wrong version of auxiliary information"
"Cannot interpret reversibility information"
"Program error interpreting InChI aux"
"Unknown error"
```

Internal errors (possible software error)

```
"Out of RAM"
"Cannot disconnect metal error"
"Fatal undetermined program error"
"Cannot allocate output data. Terminating"
"Cannot distinguish components”
"Cannot extract Component"
"ARRAY OVERFLOW"
"LENGTH_MISMATCH"
```

```
"OUT_OF_RAM"
"RANKING_ERR"
"ISOCOUNT_ERR"
"TAUCOUNT_ERR"
"ISOTAUCOUNT_ERR"
"MAPCOUNT_ERR"
"ISO_H_ERR"
"STEREOCOUNT_ERR"
"ATOMCOUNT_ERR"
"STEREOBOND_ERR"
"REMOVE_STEREO_ERR"
"CALC_STEREO_ERR"
"STEREO_CANON_ERR"
"CANON_ERR"
"UNKNOWN_ERR(#)
"No description(#)"
```