

## InChI Isotopologue and Isotopomer Development Team

### Purpose:

Develop enhanced specifications within the regular InChI standard for representing isotopologues and isotopomers. More specifically, augment the isotopic layer specifications of the regular InChI standard so that specific isotopologues, isotopomers, partial isotopomers, and isotopologue fragments can be represented by a single InChI string and used to identify isotope-informative analytical features.

### Current Team Members:

Hunter Moseley - [hunter.moseley@gmail.com](mailto:hunter.moseley@gmail.com)

Philippe Rocca-Serra - [proccaserra@gmail.com](mailto:proccaserra@gmail.com)

Reza Salek - [r7salek@gmail.com](mailto:r7salek@gmail.com)

Masanori Arita - [masanori.arita@gmail.com](mailto:masanori.arita@gmail.com)

Emma Schymanski - [schymane@gmail.com](mailto:schymane@gmail.com)

### Problems Being Solved:

While the InChI standard has an isotopic layer for representing exact isotopomers, there is not a specification for representing a range (set) of isotopomers. The fundamental issue is that an ambiguous location of specific isotopes of certain atoms cannot be represented in the current standard. This is needed to represent a set of isotopomers that correspond to a specific isotopologue. Currently, an InChI string can represent a specific isotopomer, but not a set of mass-equivalent isotopomers. For example, here is the InChI string for alpha-D-glucopyranose:

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1

[C\(\[C@@H\]1\[C@H\]\(\[C@@H\]\(\[C@H\]\(\[C@@H\]\(O1\)O\)O\)O\)O](#)

The following InChI string represents the explicit <sup>13</sup>C isotopomer with the 4th carbon labeled:

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/*i1+0,2+0,3+0,4+1,5+0,6+0*

[\[12CH2\]\(\[12C@@H\]1\[12C@H\]\(\[13C@@H\]\(\[12C@H\]\(\[12C@@H\]\(O1\)O\)O\)O\)O](#)

The isotopic layer is highlighted in red. However, there is no way to create a single valid InChI string that represents the isotopologue of alpha-D-glucopyranose containing one <sup>13</sup>C atom at an undefined atomic location, which is the signal that would be observed in mass spectrometry experiments. The use of multiple InChI strings to represent this is impractical for many isotopologues.

### Tasks:

- 1) Develop a more complete InChI isotopic layer specification and examples for isotopologues, isotopomers, partial isotopomers, and isotopologue fragments for naming isotopically-resolved molecular features detectable by analytical instrumentation.
- 2) Develop InChI isotopic layer specification and examples for non-isotopically-resolved isotopologues detected by less resolved/accurate mass analytical instrumentation.
- 3) Review and refine developed specifications.
- 4) Present proposed specifications to the IUPAC InChI subcommittee and the InChI Trust.
- 5) Present proposed specifications to the relevant community and solicit feedback.

### Timeline:

- Develop a specification the proposal within 4-8 weeks (February to March 2018).
- Present and refine the proposal with the IUPAC InChI subcommittee (April to May 2018).
- Present and refine the proposal with the relevant community (June to August 2018).

## Definitions and Background Information

### Definitions:

*Structural Isomer* - or constitutional isomerism - is one of a set of isomeric molecules with the same molecular formula but with different bonding patterns and atomic organization (i.e. line formulae - see <https://goldbook.iupac.org/html/C/C01285.html>).

*Stereoisomer* - is one of a set of isomeric molecules that have the same molecular formula and sequence of bonded atoms (constitution), but differ in the configuration of their atoms in three-dimensional space (see <https://goldbook.iupac.org/html/S/S05984.html>).

*(Extended) Isotopomer* - also known as isotopic isomers, these are isomers (<https://goldbook.iupac.org/html/I/I03289.html>) with the same number of each isotope of each element but differing in their positions within the chemical structure. For example,  $\text{CH}_3\text{CHDCH}_3$  ([CC\(\[2H\]\)C](#)) and  $\text{CH}_3\text{CH}_2\text{CH}_2\text{D}$  ([CCC\(\[2H\]\)](#)) are a pair of constitutional isotopomers of [propane](#). Isotopomers can also be specific to stereochemistry, where they are known as isotopic stereoisomers (e.g. (R)- and (S)- $\text{CH}_3\text{CHDOH}$  - see <https://goldbook.iupac.org/html/I/I03352.html>). By definition, isotopomers have the same mass, since they have the same isotopic composition. However, the set of all isotopomers which span all possible isotopic compositions are often collectively referred to as isotopomers. While technically inaccurate, it is a pragmatic use of the term to refer to all possible isotopomers spanning all isotopic compositions for a given chemical entity. Therefore, we will refer to isotopomer as any isotope-position-distinct isomer of a given chemical entity. The term "mass-equivalent isotopomer" will be used to refer to the technical definition of isotopomer.

*Mass-equivalent Isotopomer* - see above; we will use this term to refer to the technical definition of an isotopomer, as one in a collection of isomers with the same number of each isotope of each element but differing in their positions within the chemical structure.

*Partial Isotopomer* - part of an isotopomer where the isotopic content of specific atoms is known. In nuclear magnetic resonance spectroscopy (NMR) and in rare circumstances in tandem mass spectrometry, individual spectral features may directly indicate that certain atoms of a particular molecule have certain isotopes. As an example from NMR, certain analytical features may indicate that carbon 1 of alpha-D-glucopyranose is  $^{13}\text{C}$  labeled without knowing the isotopic status of the other carbon atoms.

*Isotopologue* - representing molecules of the same chemical entity and same isotopic composition. Related *isotopologues* represent the same chemical entity but differ in the isotopic composition (number of isotopic substitutions - see <https://goldbook.iupac.org/html/I/I03351.html>). The isotopologues of water that span the stable isotope compositions of hydrogen and oxygen are as follows:  $^1\text{H}_2^{16}\text{O}$ ,  $^1\text{H}^2\text{H}^{16}\text{O}$ ,  $^2\text{H}_2^{16}\text{O}$ ,  $^1\text{H}_2^{17}\text{O}$ ,  $^1\text{H}^2\text{H}^{17}\text{O}$ ,  $^2\text{H}_2^{17}\text{O}$ ,  $^1\text{H}_2^{18}\text{O}$ ,  $^1\text{H}^2\text{H}^{18}\text{O}$ , and  $^2\text{H}_2^{18}\text{O}$ . A single *isotopologue* represents a set of mass-equivalent isotopomers (e.g.  $^{12}\text{C}_2^1\text{H}_6$  has one isotopomer;  $^{12}\text{C}_2^1\text{H}_5^2\text{H}$  has two isotopomers).

*Nominal-Mass Isotopologue* - a set of isotopologues with the same elemental and nucleon compositions. In many mass spectra, the detected spectral features are not isotopically-resolved, but are elementally and nucleon resolved, thus representing what we will term a nominal-mass isotopologue.

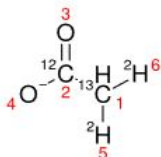
*Isotopologue Fragment* - a refined set of isotopomers where the ambiguity of isotope location is limited to a subset of the atoms. This is a subset of the mass-equivalent isotopomers representing the full isotopologue. Tandem mass spectrometry may provide spectral features that indicate where (possibly mixed) isotopes are localized within a chemical structure. For example,  $^{13}\text{C}_1$  may be limited to the glycerol part of a glycerophospholipid and not the fatty acid chains.

*Disclosure: The IUPAC Gold Book was a major source for crafting these definitions (<https://goldbook.iupac.org>). Wikipedia was another significant source for crafting some of these definitions, especially for [structural isomer](#), [stereoisomer](#), [isotopomer](#) and [isotopologue](#).*

### Background Information:

The current InChI specification contains an isotopic layer that can specify exact isotopomers. There is also some ability to specify limited atomic location ambiguity for hydrogen isotopes. For example, acetate containing one  $^{13}\text{C}$  and two  $^2\text{H}$  isotopes in the methyl group has the following InChI string:

InChI=1S/C2H4O2/c1-2(3)4/h1H3,(H,3,4)/p-1/i1+1D2,2+0  
[\[13CH\]\(\[12C\]\(=O\)\[O-\]\)\(\[2H\]\)\[2H\]](#)



Here are example molecules we will use for demonstration in the following sections:

alpha-D-glucopyranose (left)

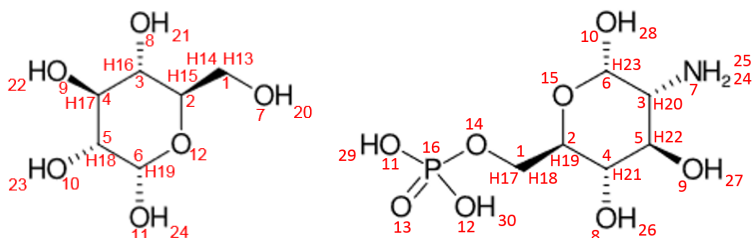
InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1

[C\(\[C@@H\]1\[C@H\]\(\[C@@H\]\(\[C@H\]\(\[C@@H\]\(O1\)O\)O\)O\)O\)O](#)

2-Amino-2-deoxy-6-O-phosphono-alpha-D-glucopyranose (right)

[C\(\[C@@H\]1\[C@H\]\(\[C@@H\]\(\[C@H\]\(\[C@@H\]\(O1\)N\)O\)OP\(=O\)\(O\)O\)O\)O](#)

InChI=1S/C6H14NO8P/c7-3-5(9)4(8)2(15-6(3)10)1-14-16(11,12)13/h2-6,8-10H,1,7H2,(H2,11,12,13)/t2-,3-,4-,5-,6+/m1/s1



## **Current InChI Isotopomer Specification:**

**Simple Definition:** /i#<+|->#[l#],...

### **Complete Definition:**

/i<atom\_number><isotope\_designation>[<hydrogen\_isotope><hydrogen\_isotope\_count>]

/i<atom\_number>[<hydrogen\_isotope><hydrogen\_isotope\_count>]

<atom\_number> - specific atom that has the specified isotope.

<isotope\_designation> - isotope designation indicated by a sign (+ or -) and number indicating the unit mass difference from the rounded average atomic mass of the element. For example, the average atomic mass of Sn (118.710) is rounded to 119 and the most abundant isotope <sup>120</sup>Sn is specified by "+1".

<hydrogen\_isotope> - D for deuterium. etc.

<hydrogen\_isotope\_count> - number of the specified hydrogen isotope bonded to the designated atom (<atom\_number>); 1 if omitted.

### **Examples:**

1. <sup>13</sup>C at carbons 1 and 2 in alpha-D-glucopyranose:

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1+1,2+1

2. <sup>13</sup>C at carbon 1 with <sup>2</sup>H<sub>2</sub> attached in alpha-D-glucopyranose:

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1+1D2

3. Carbon 1 with <sup>2</sup>H<sub>2</sub> attached in alpha-D-glucopyranose:

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1D2

4. <sup>13</sup>C at carbon 1 with <sup>2</sup>H<sub>1</sub> attached in alpha-D-glucopyranose:

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1D/t1?,2-,3-,4+,5-,6+  
Technically, this is an isotopologue fragment and not an isotopomer, meaning there is limited ambiguity in the location of the deuterium isotope.

5. <sup>2</sup>H bonded to carbon 1 in a stereospecific configuration in alpha-D-glucopyranose:

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1D/t1-,2-,3-,4+,5-,6+  
This is a stereospecific isotopomer.

6. <sup>2</sup>H bonded to all heteroatoms of alpha-D-glucopyranose (as e.g. in exchange experiments):

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i7D,8D,9D,10D, 11D

## Proposal for New Isotopologue and Isotopomer Specifications

In the following proposed specifications, isotopologue designations are indicated by parentheses with the *element* identified first, while isotopomer designations lack parentheses with the *atom position* identified first: e.g. /a(C2+1) for isotopologue and /i7+1 for isotopomer. The lack of parentheses for isotopomers ensures backwards compatibility with previous InChI specifications. See individual specifications below for details.

### Isotopically-Resolved Isotopologue Specification:

**Rationale:** to enable the use of InChI to specify an isotopologue or isotopologue fragment.

**Simple Definition:** /a(Ee#<+|->#[,#]...)

**Complete Definition:**

/a(<element><isotope\_count><isotope\_designation>[,<atom\_number>])

<element> - one or two letter Element code (Ee).

<isotope\_count> - number of atoms with the designated isotope (#).

<isotope\_designation> - isotope designation indicated by a sign (+ or -) and number indicating the unit mass difference from the rounded average atomic mass of the element. For example, the average atomic mass of Sn (118.710) is rounded to 119. We specify two <sup>118</sup>Sn atoms as "/a(Sn2-1)".

<atom\_number> - specific atom(s) that may have the specified isotope. If empty, then all atoms of the given element may have the specified isotope (# in [,#]).

If multiple atoms need to be specified, these are comma separated.

**Regular Expression Definition:** [/a([()([A-Z][a-z]?[d+][+-]\d+([,]\d+)\*[)])]

### **Examples:**

1. <sup>13</sup>C<sub>2</sub> isotopologue of alpha-D-glucopyranose:

InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(C2+1)

2. <sup>13</sup>C<sub>2</sub><sup>2</sup>H<sub>3</sub> isotopologue of alpha-D-glucopyranose:

InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(C2+1),(H3+1)

3. <sup>17</sup>O<sub>2</sub><sup>18</sup>O<sub>1</sub> isotopologue of alpha-D-glucopyranose:

InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(O2+1),(O1+2)

4. <sup>13</sup>C<sub>2</sub> limited to atoms 4,5,6 isotopologue fragment of alpha-D-glucopyranose:

InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(C2+1,4,5,6)

5. <sup>13</sup>C<sub>1</sub> at atom 1, <sup>2</sup>H<sub>1</sub> limited to atoms 13,14 isotopologue fragment of alpha-D-glucopyranose:

InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1+1/a(H1+1,13,14)

6. <sup>13</sup>C<sub>2</sub><sup>15</sup>N<sub>1</sub> isotopologue of 2-Amino-2-deoxy-6-O-phosphono-alpha-D-glucopyranose

InChI=1/C6H14NO8P/c7-3-5(9)4(8)2(15-6(3)10)1-14-16(11,12)13/h2-6,8-10H,1,7H2,(H2,11,12,13)/t2-,3-,4-,5-,6+/m1/s1/i7+1/a(C2+1)

Notice the combination of isotopologue and isotopomer specification, since there is only 1 nitrogen atom in the molecule. Therefore when ambiguity of isotope location is present, it is clearly indicated in the isotopic layer.

### **Augmented Range Specification:**

**Rationale:** to enable a more concise string representation of isotopomer and isotopologues. Note in the isotopologue definition, the first sign and following number indicates the isotope, whereas in the isotopomer definition the last sign and following number indicates the isotope.

**Simple Isotopologue Definition:** /a(Ee#<+|->#[,#-#[,#]...)

**Complete Isotopologue Definition:**

/a(<element><isotope\_count><isotope\_designation>[,<start\_atom\_number>-<end\_atom\_number>])

**Simple Isotopomer Definition:** /i#-#<+|->#

The augmented isotopomer range specification is limited to a single atom\_range and isotope\_designation, due to parsing ambiguities that would arise if mixed atom\_range and single atom designations were allowed.

**Complete Isotopomer Definition:**

/i<start\_atom\_number>-<end\_atom\_number><isotope\_designation>

### **Examples:**

1.  $^{13}\text{C}_2$  limited to atoms 4,5,6 isotopologue fragment of alpha-D-glucopyranose:  
InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(C2+1,4-6)
2.  $^{13}\text{C}_6$  isotopomer of alpha-D-glucopyranose:  
InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1-6+1

### **Nominal-Mass Isotopologue Specification:**

**Rationale:** to enable an InChI representation for mass spectral features that are not isotope-resolved.

**Simple Nominal-Mass Isotopologue Definition:** /a(#n[,#]...)

**Complete Nominal-Mass Isotopologue Definition:**

/a(<neutron\_count>n[,<atom\_number>])

<neutron\_count> - number of neutrons of extra mass.

Notice the reversal of neutron\_count relative to the “n” representing neutron. This is done so that no ambiguity will exist with the isotopically-resolved isotopologue specification, where the element is indicated first and the isotope\_count second.

### **Examples:**

1. M+3 isotopologue of alpha-D-glucopyranose:  
InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(3n)  
The lack of a specified list of atoms indicates the whole molecule.
2. M+4 isotopologue limited to carbon atoms of alpha-D-glucopyranose:  
InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(4n,1,2,3,4,5,6)  
With augmented range specifications:  
InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i(4n,1-6)  
This situation could arise from use of tandem mass spectrometry features that localize the extra mass to certain elements.

### **Partial Isotopomer Specification:**

**Rationale:** to enable an InChI representation of a partial isotopomer, which is directly useful for describing most individual nuclear magnetic resonance spectroscopy spectral features.

**Simple Definition:** same as the full isotopomer specification but with the unknown isotope specifications removed (Example 2). The designation of the isotope for all atoms of a given element will represent a full isotopomer (Example 1).

### **Examples:**

- <sup>13</sup>C at 4th carbon for full isotopomer (with respect to carbon) of alpha-D-glucopyranose:  
InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1+0,2+0,3+0,4+1,5+0,6+0  
With augmented range specification:  
InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1-3+0,4+1,5-6+0
- <sup>13</sup>C at 4th carbon for partial isotopomer of alpha-D-glucopyranose:  
InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i4+1  
Only the atom with the known isotopic composition is indicated.

### **Cross-Constitutional Isomer Isotopologue Specification:**

**Rationale:** to enable a nonstandard InChI representation of isotopologues that span multiple constitutional isomers, i.e. either a fully or partially isotopically-resolved molecular formula. This specification is directly useful for describing individual features in mass spectrometry when the specific constitutional isomer is unknown.

**Simple Definition:** 1/Ee#[Ee#].../a(Ee#<+|->#) OR 1/Ee#[Ee#].../a(#n)

### **Examples:**

- <sup>13</sup>C<sub>2</sub> isotopologue of molecules with molecular formula C<sub>6</sub>H<sub>12</sub>O<sub>6</sub>:  
InChI=1/C6H12O6//a(C2+1)
- <sup>13</sup>C<sub>1</sub><sup>18</sup>O<sub>1</sub> isotopologue of molecules with molecular formula C<sub>6</sub>H<sub>12</sub>O<sub>6</sub>:  
InChI=1/C6H12O6//a(C2+1),(O1+2)
- M+3 isotopologue of molecules with molecular formula C<sub>6</sub>H<sub>12</sub>O<sub>6</sub>:  
InChI=1/C6H12O6//a(3n)

This last example enables a major use-case in mass spectrometry spectral feature assignment.



## Questions and Issues Posed by the IUPAC InChI Subcommittee

### Question 1:

*How will be stereochemistry covered, if it changes with hopping the isotopes (e.g. 2 x  $^{13}\text{C}$  in inositol). In regards to stereochemistry, non-standard InChi can distinguish between 'undefined' and 'unknown', will the isotopic enhancement cover similar cases?*

### Answer:

The answer is 'undefined' for stereochemistry created by the presence of isotope. Since the isotopologue represents a set of isotopomers, each isotopomer may have a specific stereochemistry based on specific isotope location. It is not that they are necessarily 'unknown', but the set can create a situation where there is not a single completely designated stereochemistry. For the stereochemistry not specified by isotope, this would naturally be included with the isotopologue designation, as illustrated in the following examples:

<https://comptox.epa.gov/dashboard/dsstoxdb/results?search=inositol>

InChI=1S/C6H12O6/c7-1-2(8)4(10)6(12)5(11)3(1)9/h1-12H

<https://comptox.epa.gov/dashboard/dsstoxdb/results?search=myo-Inositol>

InChI=1S/C6H12O6/c7-1-2(8)4(10)6(12)5(11)3(1)9/h1-12H/t1-,2-,3-,4+,5-,6-

The InChI strings for  $^{13}\text{C}_2$  isotopologues of inositol and myo-inositol would be:

InChI=1/C6H12O6/c7-1-2(8)4(10)6(12)5(11)3(1)9/h1-12H/a(C2+1)

InChI=1/C6H12O6/c7-1-2(8)4(10)6(12)5(11)3(1)9/h1-12H/t1-,2-,3-,4+,5-,6-/a(C2+1)

### Question 2:

*Is canonicalization depending on the mass (as in the case of inositol)?*

### Answer:

For the most part, the current InChI canonicalization algorithm creates a unique numbering that is isotope agnostic. However, problems arise when dealing with stereospecific isotopomers involving hydrogen. The canonicalization algorithm will generate the same atom numbering for two different stereospecific hydrogen isotopomers. The following is an example:

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1D/t1-,2-,3-,4+,5-,6+

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1D/t1-,2+,3+,4-,5+,6-/m0

The canonicalization algorithm in the InChI Trust software will generate identical MOL/SDF files with the deuteration indicated on atom number 13. Open Babel's "--gen3D" option has a segmentation fault, so the unique stereochemistry cannot be retrieved via that software. The Biological Magnetic Resonance Bank has been working on an approach called ALATIS to generate unique hydrogen atom numbering, but their solution is not practical for everyday use of InChI. It would be best that the canonicalization algorithm in the InChI Trust software at least generated unique atom numbering for stereospecific hydrogen isotopomers. An issue ([#1794](#)) has been posted on the Open Babel GitHub repository, so maybe a stereospecific MOL/SDF file could be generated from their software if they fix their bug.



**Question 3:**

*It is still about only 1 molecule, not a mixture? One molecule (partially defined) is acceptable; mixtures are more problematic and are being covered in a different project and code base*

**Answer:**

The current proposed specifications are mainly for one molecule; however, the cross-constitutional isomer isotopologue specification could represent a mixture of structural isomers with the same mass, as only the formula is defined.

**Question 4:**

*There is a potential overlap with variable structure proposal; the isotopologue proposal is asking a different and more specific question. Revision on the variable structure proposal could, potentially, allow for unspecified positions to include isotopologues.*

**Answer:**

We suspect that “dividing and conquering” (treating both cases in the separate sub-groups then ensuring compatibility of the developed approaches) may be a better strategy, since we have specifications to handle nominal mass and across (all) constitutional isomers. The nominal mass situation will likely be hard to handle in the variable structure proposal. The cross-constitutional isomer situation is probably handling a much wider range of variable structures than what the variable structure proposal is focused on handling. It would be good for some of our team to directly interact with the team developing the variable structure proposal to discuss this.

**Question 5:**

*What input format was suggested to feed InChI with an isotopologue, which has uncertain positions of isotopic atoms?*

*For instance, consider two Deuterium atoms located in any two out of 5 hydroxyl groups in a molecule. Of course, the examples may be much more complicated.*

*As far as I know, molfile formats, which are currently used to feed InChI with chemical structures, cannot express such an isotopologue (I think that query formats should not be used)*

*IMHO, no software development is possible without a suitable well-defined input format, a representative test set of isotopologues expressed in the suggested format, and their suggested InChI.*

**Answer:**

The ideal solution is to extend the MOL file format to include options for specifying ambiguous isotope location. However, such development and standardization is probably beyond scope of the current proposal development for InChI isotopologue specifications. This is due to the history of the CT File Format, the parent MOL file format which was released by MDL Information Systems and through a series of acquisitions is now part of the BioVia unit of

Dassault Systèmes. Is there even a formal process to extend the MOL file format and how would such an extension be disseminated? If a formal extension process for CT/MOL files can be found or established, we would be willing to initiate the development of extensions to the MOL file format that facilitate direct representation of isotopologues.

As an alternative, an SDF file could be used to provide a complete description of an isotopologue. This approach is functional and would be completely descriptive of an isotopologue by providing an enumeration of all isotopomers represented by a given isotopologue. It also holds to the concept that a single MOL file can represent a single isotopomer, while an isotopologue represents a compatible set of isotopomers that must be represented as a set. Here are two possible SDF representations that would be adequate:

- A) Including MOL files for all specific isotopomers represented by the isotopologue.
- B) Including masking MOL files for each isotope present that indicate the ambiguous location of isotope along with additional information in the SDF file to indicate the number of each isotope present for a given masking MOL file. The SDF file would also include partial isotopomer MOL files for isotope that is unambiguous in its location.

Such SDF file representations could be used both as an input format and for visualization purposes. Hunter Moseley's lab has an open codebase available on GitHub that can generate type-A SDF files: <https://github.com/MoseleyBioinformaticsLab/isoenum>. His lab is willing to extend the codebase to generate type-B SDF files as a demonstration of these isotopologue representing SDF files, which would be better for visualization purposes.

#### **Question 6:**

*To make the isotopologue InChI usable and adopted by masses, in addition to the well-defined input format an isotopologue chemical structure editor/viewer is a must. This requires a simple and easy to understand concept of visualization of any complexity isotopologues. Of course the editing is out of InChI project scope. However, the visualization concept may be required for extending winchi to isotopologues.*

#### **Answer:**

See answer to Question 5.

#### **Question 7:**

*The most serious concern I have is a mixture of "known" and "unknown" things appearing in the same /i layer, as the proposal presumes.*

*It is based on idea that a single /i layer is covering both isotopomer (precisely known) and isotopologue (unknown or unprecise) information.*

*InChI typically tries to employ somewhat different general design, which has some added benefits.*

*Namely, it first includes some layer capturing an unprecise information, then optionally includes its extension/analog, which accounts for more precise info.*

*This layered approach allows one to represent molecular information with a different degree of preciseness.*

*The example is tautomeric data [e.g., <https://www.inchi-trust.org/technical-faq/#6.1> ]*

*InChI string contains /h sub-layer in main layer, which lists all possible locations of mobile H's -- but not the precise locations themselves. However, the exact locations may be further specified by /h sub-layer in FixedH layer, following /f.*

*Most notably, you may cut longer ("precise" mobile H) InChI at /f and retain only the beginning: this would be valid ("unprecise" mobile H) parent of all possible longer strings representing tautomers.*

*Ideally, it would be most convenient to have such a layout at least partially preserved for the hierarchy comprising isotopologues, mass-equivalent isotopomers, nominal-mass isotopologues, and isotope-position-distinct isotopomers, etc. Current design in the proposal does not reflect such "parent"/"child", or "generic"/"specific" relations, at least represented in easily deducible manner.*

*I am not sure if it is really/easily achievable, but I would strongly suggest that the authors consider a possibility of introducing the layered hierarchical isotopologue/isomer layout, at least partially.*

**Answer:**

We will separate precise isotopomer specification from isotopologue specification using a separate layer "/a" for ambiguous location of isotope. We have updated the specification accordingly - assuming that "/a" is not being used by current or other proposed InChI specifications.

**Question 8:**

*Another thing I am not happy with is that combined isotopologue-isotopomer layer may eventually break the hierarchy already implemented in InChI.*

*Consider an interplay of isotopes and stereo. In general, "unprecise" indication of isotope position [isotopologue] may either (a) result in (change of) stereochemistry or (b) does not touch any stereochemical feature. However, it may be very difficult to distinguish between (a) and (b), as this would require full enumeration and analysis of the whole set of available isotopomers.*

*So the only reasonable option is to consider all stereo "appearing after" /i as unreliable/undefined (as is correctly specified in answers of the authors to some previous question, end of the proposal).*

*What is the worse, sometimes (when /i does not cover isotopologues but only precise isotopomers) old logics would still work and stereo "appearing after" /i would be absolutely legitimate.*

*Moreover, any other stereo features "appearing further at right", in InChI string (e.g., due to tautomerism with /f exact H positions, which are normally accounted for), would also be effectively invalidated by isotopologue-enabled /i layer...*

**Answer:**

The separation of isotopomer and isotopologue specifications into separate layers (see above) should help make it more obvious when the isotope touches stereochemical features. Also, a full enumeration of isotopomers is not needed to distinguish between changed (undefined) stereochemistry or untouched stereochemistry. Stereochemistry created by precise isotopomer designation will be undefined if another atom bonded to the created stereocenter has ambiguous isotope designation involving the same element of the isotopomer designation.

**Question 9:**

*The proposal suggests that "isotopologue designations are indicated by parentheses with the element identified first, while isotopomer designations lack parentheses with the atom position identified first: e.g. /i( C2 +1) for isotopologue and /i 7 +1 for isotopomer."*

*I would strongly object to the usage of element symbols here (and it's also the opinion of DT with whom we briefly chatted).*

*Currently, InChI contains element symbols only in the very first, formula layer. In all other contexts, atoms are referred to by their numbers.*

*It may seem convenient for specific purpose of the moment, but preserving some design principles is more important in a long run...*

**Answer:**

We understand the need for adhering to design principles; however the current design principles were primarily based on designation of characteristics to specific atoms. By extending the InChI to cover ambiguous designation, the design principles may need to be revisited to achieve a balance between the original concept and a compact, understandable representation. The specification of the element prevents the need for specifying a collection of atoms. For instance, we currently propose to represent two <sup>13</sup>C located somewhere across all carbon atoms as:

InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(C2+1)

Here are some alternative specifications that are not as concise nor as easy to interpret:

InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(2+1,1,2,3,4,5,6)

OR

InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(2+1,1-6)

We could still represent nominal-mass isotopologues of 3 neutrons spread across the molecule as:

InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(3n)

OR (if we change the original specification to be more consistent with this new alternative)

InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(3+n)

OR

InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(3+n,1-24)

Here is an example of 3 neutrons spread across the non-hydrogen atoms:

InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/a(3+n,1-12)

But cross-constitutional isomer isotopologues cannot be represented unless we assume some ordering of atoms like the typical C/ascending atomic number/H ordering seen generated by the InChI Foundation software:

InChI=1/C6H12O6/a(2+1,1,2,3,4,5,6)

InChI=1/C6H12O6/a(2+1,1-6)

### Question 10:

*Minor issues.*

(a)

*"The following InChI string represents the explicit 13 C isotopomer with the 4th carbon labeled:*

*InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1  
/i1+0,2+0,3+0,4+1,5+0,6+0"*

*Actually, this represents the explicit 13 C isotopomer with the 4th carbon labeled and the explicit 12C "labeling" for all the other carbons.*

*Construct "+0" always means that the atom is of a specific isotope whose mass number is the same as the rounded average atomic mass.*

*[<https://www.inchi-trust.org/technical-faq/#9.3>]*

### Answer:

We are proposing that "/i4+1" refers to the partial isotopomer designation and that "/i1+0,2+0,3+0,4+1,5+0,6+0" refers to the fully specified isotopomer designation. This was shown in examples 1 and 2 under the Partial Isotopomer specification above.

### Question 11:

(b)

*Text "Current InChI Isotopomer Specification ... <hydrogen\_isotope\_count> - number of the specified hydrogen isotope bonded to the designated atom (<atom\_number>)" should be corrected as: "... <hydrogen\_isotope\_count> - number of the specified hydrogen isotope bonded to the designated atom (<atom\_number>), 1 if omitted"*

*Consequently, all the example InChI strings containing "...D1" or so should be corrected.*

*InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1D1*

*should be*

*InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1D*

*InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i7D1,8D1,9D1,10D1,11D1*

*should be*

*InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i7D,8D,9D,10D,11D*

*etc.*

**Answer:**

Thanks for bringing this to our attention. We have fixed this error in the proposed isotopologue specifications above.

**Question 12:**

It looks like my previous explanations have not reached their goal. So, below are more of them.

(1) On #4:

Complaining that it is not easy to understand InChI string or manually convert InChI into a structure should be irrelevant to the InChI-related project: InChI is supposed to be understood either by software, or by experts, who feel comfortable with it. (this refers to words “this is not easy to tell from the example without knowing the atom numbering and structure [...]”)

**Answer:**

We agree. The partial isotopologue can be determined from the InChI itself. Therefore, we have removed the commentary about not being easy to discern.

**Question 13:**

(2) I disagree with #5:

The canonicalization algorithm generates InChI, not molfile (this refers to words “the canonicalization algorithm will generate two identical MOL/SDF files for the two possible stereospecific configurations”)

To generate two stereospecific molfiles out of two stereospecific InChIs one needs to have these InChI. These two InChI are:

*InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1D/t1-,2+,3+,4-,5+,6-/m0*

*InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1D/t1-,2-,3-,4+,5-,6+*

Since these InChIs have not been mentioned in the write-up, it is not clear how the molfiles were supposed to be generated (this refers to words “One problem is that it is not easy to generate a stereospecific MOL/SDF representation” and “the canonicalization algorithm will

generate two identical MOL/SDF files for the two possible stereospecific configurations” )

A stereospecific molfile may easily be created by any comprehensive chemical structure editor: ACD/ChemSketch, ChemBioDraw, etc. It may be drawn manually or created out of InChI since these editors use libinchi.dll/libinchi.so for InChI to structure conversion. (this paragraph refers to words “One problem is that it is not easy to generate a stereospecific MOL/SDF representation”)

InChI Trust software (libinchi.dll/libinchi.so) generates chemical structure connectivity with stereo parities out of InChI. This information is similar to stereo-SMILES and is sufficient for producing a stereospecific molfile by chemical structure editors supporting InChI to structure conversion. I guess, this fact is not known to the authors of the write-up, who used a buggy version of OpenBabel software for accessing this functionality of libinchi.dll/libinchi.so.

Producing molfiles with coordinates out of InChI was deliberately left out of the InChI Trust software functionality as well as, for instance, evaluation of E-Z or S-R transition barriers or pKa values.

Instead of complaining about this deliberately missing functionality one might have used the working software, not the buggy one.

(This section refers to words “The InChI Trust software does not create a stereospecific MOL/SDF file”).

Using OpenBabel for InChI to structure conversion was a wrong choice of software by the authors of the write-up. I do not think sharing this experience in the text is a good idea for at least two reasons: (i) The text may be still available after OpenBabel has fixed the bug, thus making the text factually incorrect; (ii) The preferred way to encourage bug fixing is reporting the bug to the OpenBabel team instead of using project documentation as a platform for publicly announcing a bug in free and open-source software. (this refers to words “openbabel’s “--gen3D” option generates a segmentation fault when trying to convert from InChI to MOL/SDF format”)

**Answer:**

Based on these suggestions, we have revised the description of example #5. With the questions and issues being posed and more extensive discussions with the developers of the InChI Trust software, we are gaining a better understanding of the methods and use-cases driving the software implementation of the InChI standard. Also, we have posted issues on the OpenBabel github repository for the bugs we have discovered so far in that software. We were just trying to provide a concrete example demonstrating some difficulty in generating stereospecific MOL/SDF representations. But it is a good suggestion not to use momentary software bugs to make this point. The main point we are trying to highlight is that there is a growing use-case for the stereospecific interconversion between InChI and MOL/SDF.



However, the isotopologue specification proposal is probably not the correct venue for this discussion. Therefore, we have revised the following commentary, but removed it from example #5:

“

One issue is that it is not easy to generate a stereospecific MOL/SDF representation. The InChI Trust software does not create a stereospecific MOL/SDF file and its MOL/SDF conversion algorithm will generate two identical MOL/SDF files for the following two possible stereospecific configurations:

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1D/t1-,2-,3-,4+,5-,6+

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6+/m1/s1/i1D/t1-,2+,3+,4-,5+,6-/m0

In all fairness, generation of stereospecific MOL/SDF is not a use-case for the InChI Trust software. Therefore, this use-case is de facto delegated to third-party software.

“

The growing importance of this stereospecific conversion use-case on use of molecular analytical data in a range of next-generation systems-level mathematical modeling of chemical and biochemical systems cannot be overstated.

## July 15, 2018 Progress Report

In late February 2018, the following team was assembled to develop a proposal for extensions to the InChI standard that enable the concise representation of isotopologues and isotopomers:

- Hunter Moseley - [hunter.moseley@uky.edu](mailto:hunter.moseley@uky.edu) - University of Kentucky, Lexington, KY USA
- Philippe Rocca-Serra - [proccaserra@gmail.com](mailto:proccaserra@gmail.com) -
- Reza Salek - [salekr@iarc.fr](mailto:salekr@iarc.fr) - International Agency for Research on Cancer, World Health Organization, 150 cours Albert Thomas 69372 Lyon CEDEX 08, France
- Masanori Arita - [arita@nig.ac.jp](mailto:arita@nig.ac.jp) - National Institute of Genetics, Yata 1111, Mishima, 411-8540 Shizuoka, Japan
- Emma Schymanski - [emma.schymanski@uni.lu](mailto:emma.schymanski@uni.lu) - Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg.

One purpose of these extensions is to enable unambiguous and concise annotation of nuclear magnetic resonance and mass spectral features with respect to isotope content, enabling a new level of data reusability for growing metabolomics repositories. By the end of March 2018, the team had developed an initial draft of the proposal, which was presented to various members of the InChI IUPAC subcommittee and the InChI Trust for feedback and refinement. Over April and May of 2018, the proposal was refined with multiple cycles of feedback. To elicit feedback from the broader scientific community that would benefit from the proposed extensions, the refined proposal was presented during the Stable Isotope Approaches to Metabolomics workshop at the 14th Annual Conference of the Metabolomics Society in Seattle, Washington in late June 2018. During this presentation, the isotope-specific spectral feature annotation problem was presented along with how the proposed extensions to InChI would solve this problem. Basic definitions and examples of the extensions were also presented and questions from the audience were answered during a panel discussion.