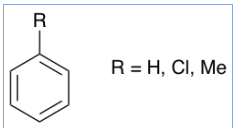
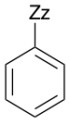
InChI: Variable Structures and Markush

Here are three approaches to describing variable structures. The names are intended to facilitate discussion and are not formal proposals

(i) MarkInChI: A central structure with a defined list of sidechains

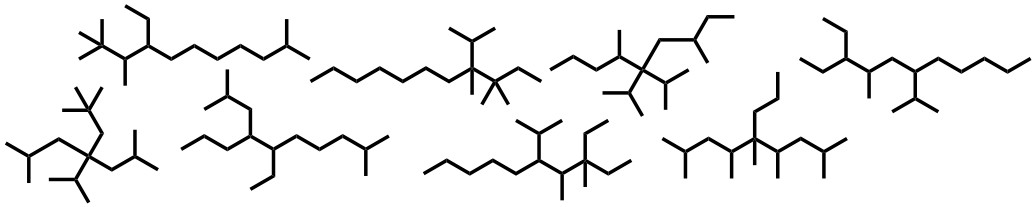




InChI=1B/C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H<>C!Cl!H

A pseudo-element (available in InChI v1.06) is used to mark the substitution point, and a list of possibilities for the pseudo-element is given after the <> separator. This corresponds quite closely to a traditional Markush structure, and may encode the intentions of the creator. More complex examples require precise specification of the syntax.

(ii) VInChI: A list of molecules which can be gathered into a single string



VInChI=1S/C17H36/c1-13(2)10-17(15(5)6,11-14(3)4)12-16(7,8)9/h13-15H,10-12H2,1-9H3
/pi-c(1+2+3+7;4+9+10+11-c(5+8+15;14+2+14-c(2+8+9+12;12+4+6+1)c(1+3+4+5;5+2+15+7)c(3+7+7+10;15+6+4+4)c(1+4+7;7+6+15)c(1+3+6;8+1+1)))

A list of InChI is compacted into a single string. The beginning of the VInChI string is the standard InChI of one of the list; the /pi layer describes how the others can be generated. It is easy to provide an input for this approach and it is straightforward to generate canonical VInChI. Ideally, the input could include VInChI as well as InChI.

(iii) PartInChI: A partial description of a structure with constraints

All saturated, acyclic alcohols with eight carbon atoms (89 isomers):
InChI=1B/C8H18O/h9H

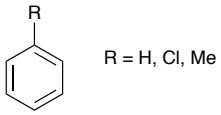
This approach makes it possible to generate descriptions of very complex and numerous structures without enumeration.

Jonathan Goodman, March 2021

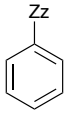
MarkInChI

The new release of InChI introduces pseudo-atoms which can be used to help describe Markush structures. This requires use of the non-standard -NPZz and -SAtZz flags, which allow non-polymer-related Zz atoms (pseudo element placeholders) and which allow stereochemistry at atoms connected to Zz (default: disabled). Although there is only one sort of pseudo-atom, Zz, the InChI algorithm produces a canonical numbering for all the pseudo-atoms in a molecule and so it is possible to generate a canonical (non-standard) InChI for a Markush structure with multiple R groups.

Example One:



This simple Markush structure could be represented by replacing the R group with a pseudo-atom:

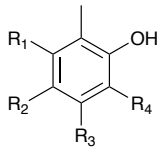


This structure has the non-standard InChI:
`InChI=1B/C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H`
The Markush structure can be represent by this InChI and a list of possible Zz groups (H, Cl, C in this case. The syntax could follow the form of the RInChI. The substitution groups are separated from the InChI of the core and from each other by "<>". The different options for the R group are separated by "!". This leads to a MarkInChI for this structure:
`MarkInChI=1B/C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H<>C!Cl!H`
The pseudo-atom, Zz, can be replaced by the atoms listed at the end of the string: carbon, chlorine, hydrogen. The options for R should be arranged in alphabetical order in a step towards a canonical identifier.

If the structure represented by R were more complex, for example, ethyl, propyl and iso-propyl, these could be represented by InChI with a Zz atom indicating the attachment point:
`InChI=1B/C2H5Zz/c1-2-3/h2H2,1H3`
`InChI=1B/C3H7Zz/c1-2-3-4/h2-3H2,1H3`
`InChI=1B/C3H7Zz/c1-3(2)4/h3H,1-2H3`

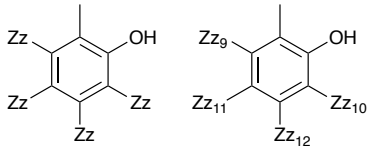
The InChI for RPh with R=H,Et,Pr,*i*Pr would be:
`MarkInChI=1B/C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H<>C2H5Zz/c1-2-3/h2H2,1H3!C3H7Zz/c1-2-3-4/h2-3H2,1H3!C3H7Zz/c1-3(2)4/h3H,1-2H3!H`

Example Two:



R_1 = H, Cl
R_2 = Me, Et
R_3 = OH, NH_2
R_4 = Me, Ph

In this example, it is necessary to distinguish between the different R groups, using the single pseudo-atom type. This is possible because the InChI algorithm numbers the atoms.



The representation on the left has the InChI:

`InChI=1B/C7H4OZz4/c1-2-3(8)5(10)7(12)6(11)4(2)9/h8H,1H3`

This numbers the pseudo-atoms as shown on the right, above. The R groups may be represented as:

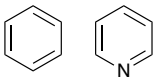
```
Zz9 = H, Cl
Zz10 = C, InChI=1B/C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H
Zz11 = C, InChI=1B/C2H5Zz/c1-2-3/h2H2,1H3
Zz12 = O, N
```

Using the RInChI-like symbols, the options for different pseudo-atoms would be listed in order, separated by <>:

`MarkInChI=1B/C7H4OZz4/c1-2-3(8)5(10)7(12)6(11)4(2)9/h8H,1H3<>Cl!H<>C!C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H<>C!C2H5Zz/c1-2-3/h2H2,1H3<>N!O`

MarkInChI=[*core InChI*]<>[*options for first Zz, sorted and separated by !*]<>[*options for second Zz, sorted and separated by !*]<>*etc*

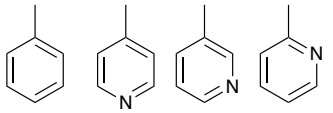
Example Three:



`InChI=1S/C6H6/c1-2-4-6-5-3-1/h1-6H`

In this example, a Zz atom is not used. Zz can only be monovalent. This makes it possible to use a standard InChI, but recommend consistent use of the -NPZz and -SAtZz flags. Since all the carbon atoms are equivalent, should use the one with the lowest index for the change:

`MarkInChI=1S/C6H6/c1-2-4-6-5-3-1/h1-6H<>1-C!N`

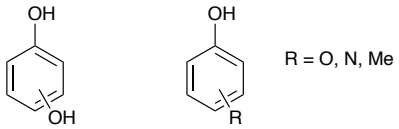


`InChI=1S/C7H8/c1-7-5-3-2-4-6-7/h2-6H,1H3`
`MarkInChI=1S/C7H8/c1-7-5-3-2-4-6-7/h2-6H,1H3<>2-C!N`
`MarkInChI=1S/C7H8/c1-7-5-3-2-4-6-7/h2-6H,1H3<>3-C!N`
`MarkInChI=1S/C7H8/c1-7-5-3-2-4-6-7/h2-6H,1H3<>5-C!N`

Always used the lowest index available. For all four structures, indicating just one carbon becomes a nitrogen:

`MarkInChI=1S/C7H8/c1-7-5-3-2-4-6-7/h2-6H,1H3<>2,3,5-C!N`

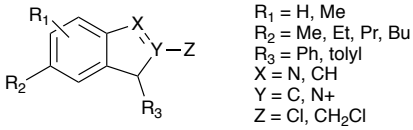
Example Four:



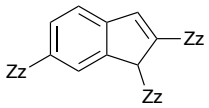
For structures with variable points of attachment, the hydrogens on atoms in the core are replaced, rather than the atoms themselves. This can be represented:

```
InChI=1S/C6H6O/c7-6-4-2-1-3-5-6/h1-5,7H
MarkInChI=1S/C6H6O/c7-6-4-2-1-3-5-6/h1-5,7H<>1H,2H,4H-O
MarkInChI=1S/C6H6O/c7-6-4-2-1-3-5-6/h1-5,7H<>1H,2H,4H-C!N!O
```

Example Five:

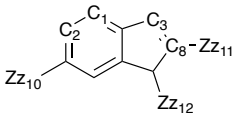


This more complex example would be represented by a core in which R_2, R_3 and Z are marked as pseudo-atoms, R_1 is omitted because it has a variable attachment point, and X and Y are represented by the first option in their respective lists, after sorting (X = CH; Y = C). The core structure, therefore, is:




```
InChI=1B/C9H5Zz3/c10-6-2-1-5-3-8(11)9(12)7(5)4-6/h1-4,9H
```

The InChI numbering includes:



Building up the MarkInChI:

R_1: This is not marked on the core-structure, so will come at the end of the list of substitutions in the MarkInChI. R_1 is C or H, attached to atoms 1 and 2, so: `<>1H,2H-C!H`

R_2: Me, Et, Pr, Bu replacing Zz_{10}. This is the first pseudo-atom, according to the InChI numbering, so these options come first in the list. Me can be represented as C, but the other options require InChI with pseudo-atoms indicating the point of attachment:

```
<>C!C2H5Zz/c1-2-3/h2H2,1H3
      !C3H7Zz/c1-2-3-4/h2-3H2,1H3!C4H9Zz/c1-2-3-4-5/h2-4H2,1H3
```

R_3: Ph and tolyl (assume *para*-tolyl) replace Zz_{12}:

```
<>C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H
      !C7H7Zz/c1-6-2-4-7(8)5-3-6/h2-5H,1H3
```

X: Carbon 3 (InChI numbering) can be CH or N: `<>3-C!N`

Y: Carbon 8 (InChI numbering) can be C or N+: `<>8-C!N+`

Z: Either Cl or CH_2Cl. This corresponds to Zz_{11} and so should come in the list between R_2 (Zz_{10}) and R_3 (Zz_{12}): <>CH2ClZz/c2-1-3/h1H2!Cl

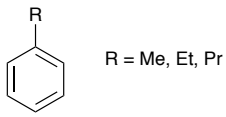
So the overall MarkInChI is:
MarkInChI=1B/C9H5Zz3/c10-6-2-1-5-3-8(11)9(12)7(5)4-6/h1-4,9H<>C!C2H5Zz/c1-2-3/h2H2,1H3!C3H7Zz/c1-2-3-4/h2-3H2,1H3!C4H9Zz/c1-2-3-4-5/h2-4H2,1H3<>CH2ClZz/c2-1-3/h1H2!Cl<>C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H!C7H7Zz/c1-6-2-4-7(8)5-3-6/h2-5H,1H3<>3-C!N<>8-C!N+<>1H,2H-C!H

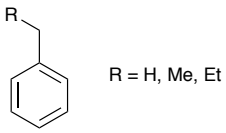
Illustrated by colour:
R_1: <>1H,2H-C!H
R_2: <>C!C2H5Zz/c1-2-3/h2H2,1H3
!C3H7Zz/c1-2-3-4/h2-3H2,1H3!C4H9Zz/c1-2-3-4-5/h2-4H2,1H3
R_3: <>C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H
!C7H7Zz/c1-6-2-4-7(8)5-3-6/h2-5H,1H3
X: <>3-C!N
Y: <>8-C!N+
Z: <>CH2ClZz/c2-1-3/h1H2!Cl

MarkInChI=1B/C9H5Zz3/c10-6-2-1-5-3-8(11)9(12)7(5)4-6/h1-4,9H<>C!C2H5Zz/c1-2-3/h2H2,1H3!C3H7Zz/c1-2-3-4/h2-3H2,1H3!C4H9Zz/c1-2-3-4-5/h2-4H2,1H3<>CH2ClZz/c2-1-3/h1H2!Cl<>C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H!C7H7Zz/c1-6-2-4-7(8)5-3-6/h2-5H,1H3<>3-C!N<>8-C!N+<>1H,2H-C!H

Current limitations

• Variable ring sizes are not currently covered by this proof-of-concept approach
• These MarkInChI need to be constructed by hand, which is reasonable for small numbers of molecules
• It will be possible to construct examples for which these rules are insufficient. For this proof-of-concept, we are content if this simple approach covers a useful subset of Markush structures.
• Canonical: there is a strong aspiration that these MarkInChI should be canonical
• Guidelines for substructure selection: Here are different Markush for the same molecules: which is better? Probably the latter as it has a larger core





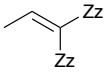
Jonathan Goodman, October 2020

InChI v1.06 and Markush Structures

February 2021 update

(i) The details of the syntax need to be precisely defined. For example, the separator "!" means "one from the list" so `1H,2H-C!H` means either 1H or 2H can be C, but not both (two possibilities). Alternatively, `1H-C!H<>2H-C!H` means either, or both, or neither (four possibilities). What about either or both but not neither? What about neither or one, but not both? How best to encode two methyls in any of five positions, or one methyl and one alcohol at any of four positions?

(ii) Although the InChI algorithm numbers pseudo-elements, making it possible to have an unlimited number of distinguishable substitution sites in a MarkInChI core, it does not allow for the creation of stereochemistry one the pseudo-elements represent different side chains. For example:



`InChI=1B/C3H4Zz2/c1-2-3(4)5/h2H,1H3`

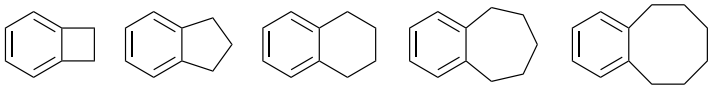
The two pseudo-elements are numbered four and five, but the double bond is not labelled as *E* or *Z* because the two pseudo-elements are the same. If they are substituted with different side-chains, the double bond geometry becomes significant. This can be addressed by a two stage process: first generate the stereochemistry-free InChI, and then substitute the pseudo-elements to generate all possible stereochemistry.

(iii) A controlled vocabulary for standard sidechains (R, Ar, X, *etc*) might be useful.

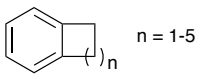
(iv) What are the best decisions to make on these and other issues? Someone with patent expertise is needed to help discover the design which will best cover likely use-cases.

Jonathan Goodman, March 2021

A similar approach can also be used to describe ring sizes.



This sequence may be represented by a single diagram



The InChI of the 6,4 ring system is:

`InChI=1S/C8H8/c1-2-4-8-6-5-7(8)3-1/h1-4H,5-6H2`

Atom 6 corresponds to the one with the brackets and n subscript in the diagram.

`InChI=1S/C8H8/c1-2-4-8-6-5-7(8)3-1/h1-4H,5-6H2<>6((1-5)C)`

This encodes the changing ring size. A fragment could be used in place of "*C*" for more complex examples. This takes the original atom, 6, and replaces it with a linear string of one, two, three, four or five methylenes. The number must be in brackets to allow more complex choices. If the structure could only have two, four or six methylenes, this could be written (2,4,6)C.

Jonathan Goodman, August 2017

VInChI

Anthony Baston, a Master's student at Cambridge, has developed a program to generate VInChI for a restricted group of molecules: acyclic alkanes. The program takes a list of InChI as its input and generates a single, canonical string to describe them all.

For example:

```
InChI=1S/C17H36/c1-13(2)10-17(15(5)6,11-14(3)4)12-
        16(7,8)9/h13-15H,10-12H2,1-9H3
InChI=1S/C17H36/c1-7-10-11-12-17(14(4)5)13-15(6)16(8-2)9-
        3/h14-17H,7-13H2,1-6H3
InChI=1S/C17H36/c1-7-10-17(13-15(5)6)16(8-2)12-9-11-
        14(3)4/h14-17H,7-13H2,1-6H3
InChI=1S/C17H36/c1-8-10-11-12-13-14-17(7,15(3)4)16(5,6)9-
        2/h15H,8-14H2,1-7H3
InChI=1S/C17H36/c1-8-11-12-13-16(14(4)5)15(6)17(7,9-2)10-
        3/h14-16H,8-13H2,1-7H3
InChI=1S/C17H36/c1-8-16(15(4)17(5,6)7)13-11-9-10-12-
        14(2)3/h14-16H,8-13H2,1-7H3
InChI=1S/C17H36/c1-9-10-17(8,15(6)11-13(2)3)16(7)12-
        14(4)5/h13-16H,9-12H2,1-8H3
InChI=1S/C17H36/c1-9-11-16(8)17(13(3)4,14(5)6)12-15(7)10-
        2/h13-16H,9-12H2,1-8H3

VInChI=1S/C17H36/c1-13(2)10-17(15(5)6,11-14(3)4)12-
16(7,8)9/h13-15H,10-12H2,1-9H3
/pi-c(1+2+3+7;4+9+10+11-c(5+8+15;14+2+14-
c(2+8+9+12;12+4+6+1)c(1+3+4+5;5+2+15+7)c(3+7+7+10;15+6+4+4)c(1
+4+7;7+6+15)c(1+3+6;8+1+1)))
```

A compact, canonical representation. The length of the VInChI is a measure of diversity

Jonathan Goodman, March 2021

PartInChI

The layered structure of the InChI makes it possible to give partial descriptions of molecules which might be interpreted as encoding everything that fits them. For example, octan-1-ol has the InChI:

`InChI=1S/C8H18O/c1-2-3-4-5-6-7-8-9/h9H,2-8H2,1H3`

An abbreviated form might reasonably be interpreted as including all structures with the same formula:

`InChI=1B/C8H18O`

The isomers of $C_8H_{18}O$ include ethers as well as alcohols. A PartInChI which listed just alcohols would need to have an additional section which specified either than an alcohol must be present, or else that an ether should not be present. This could be done in a variety of different ways. The simplest would be to include part of the /h layer and specify that atom nine, the oxygen, must have be attached to one hydrogen atom:

`InChI=1B/C8H18O/h9H`

More flexible structures are probably needed. A list of InChI for the fragments which are required or excluded would provide a much more flexible approach. The PartInChI could be combined with the MarkInChI approach to encode a very large number of sidechains around a fixed core.

Jonathan Goodman, March 2021