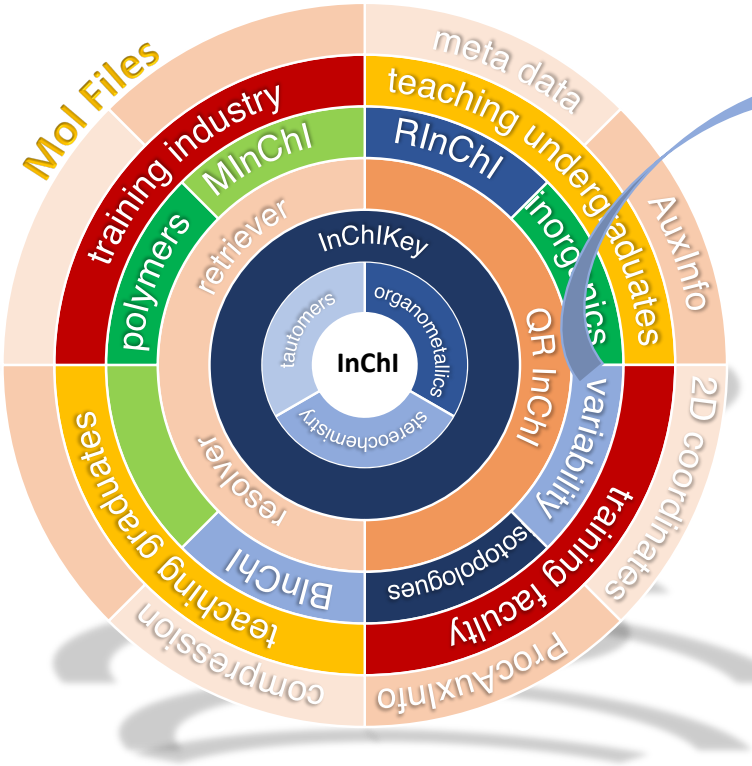
# InChI

## Markush and Variability

Jonathan M Goodman

*Working group:*
*Gerd Blanke, István Öri, Anthony Baston*

# Variable Structures in the InChI Ecosystem




***Canonical Identifier for Variability***

*(Markush)*

**VInChI**

*Compact identifier for multiple structures*

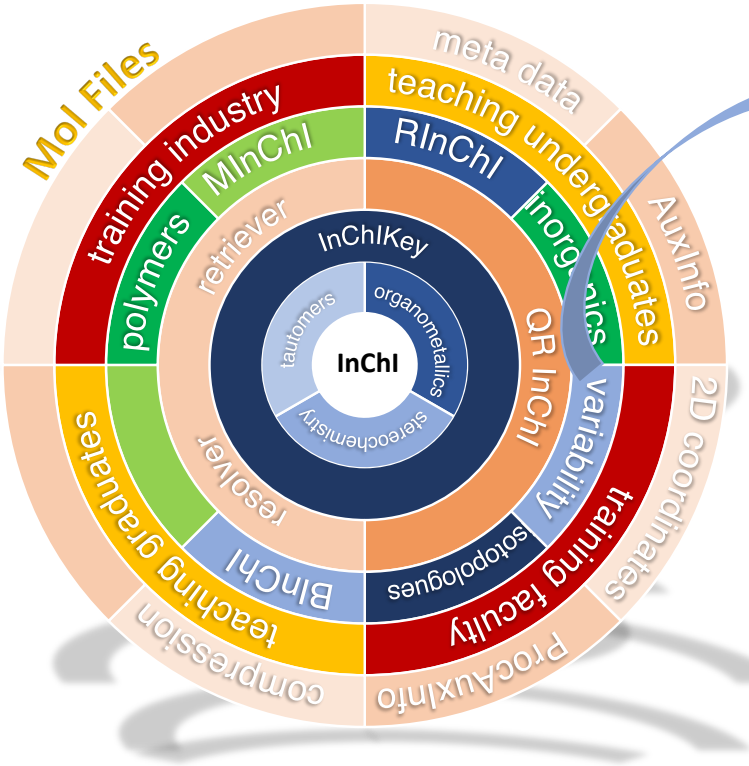
**Generic Reactions**

*Rapid manipulation of large groups of molecules*

*Extended Standard InChI?*

*Encoding of non-molecular Markush information*

# Variable Structure Working Group



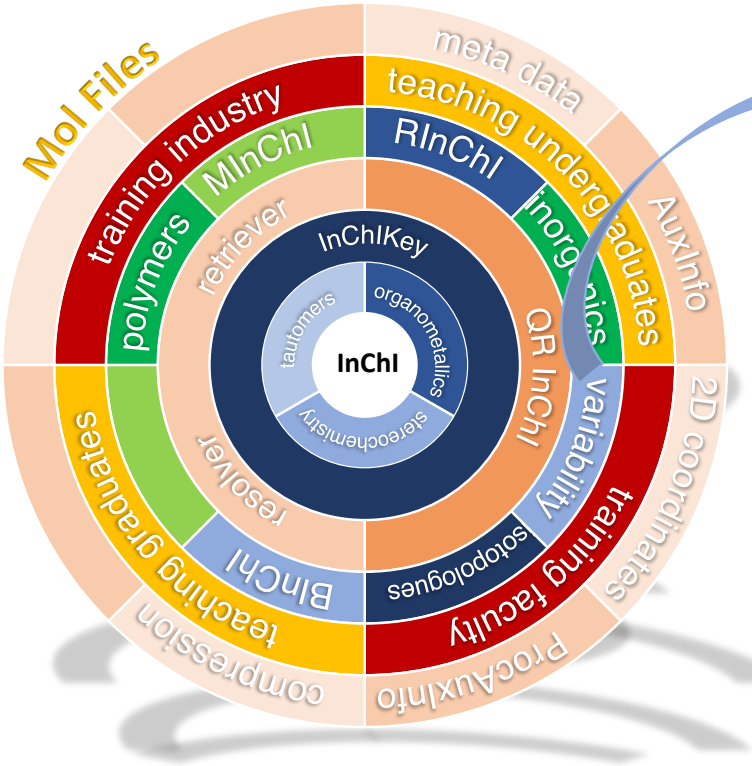

## *Open Meeting of the Working Group*

**VInChI**

***5 pm (UK)***
***April 12th***

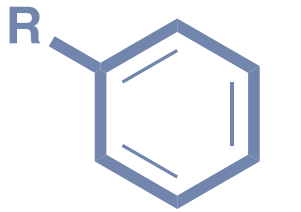
*Contact:*
***jonathan@inchi-trust.org***
***www-jmg.ch.cam.ac.uk/inchi***

# Two ***Canonical*** Variable Structure Challenges:



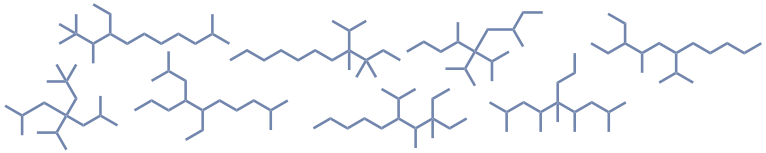

***Can we encode Markush-like structures as an InChI?***

**VInChI**

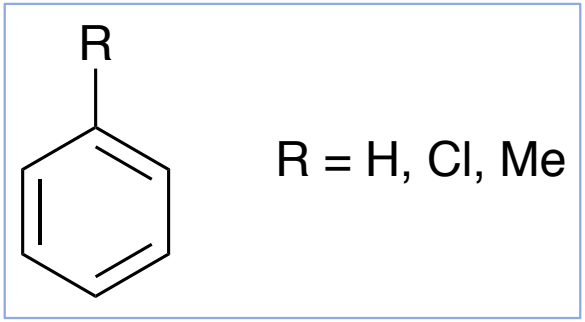



R= H, Me, Cl

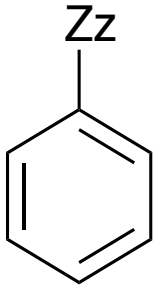
***Can we encode a group of molecules as an InChI?***



# Simple Variability: *MarkInChI*




Pseudo-element introduced in InChI v1.06

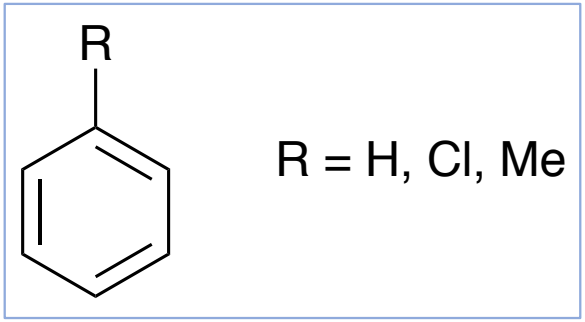



InChI=1B/C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H

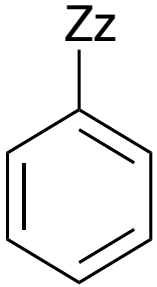
A new section lists options for the pseudo-element, in alphabetical order, separated by "!"

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

# Simple Variability: *MarkInChI*




Pseudo-element introduced in InChI v1.06

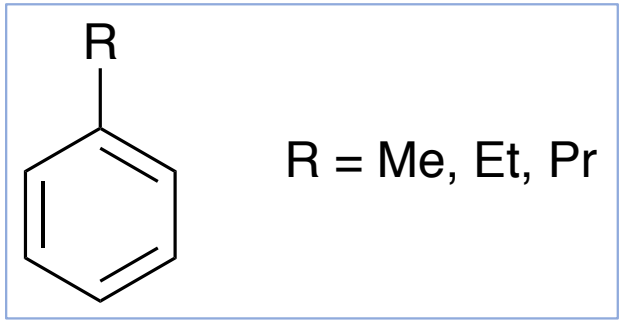



InChI=1B/C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H

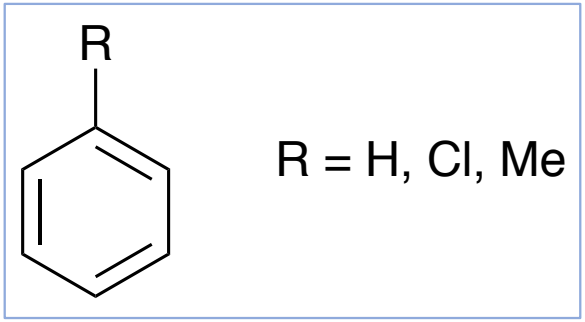
A new section lists options for the pseudo-element, in alphabetical order, separated by "!"

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

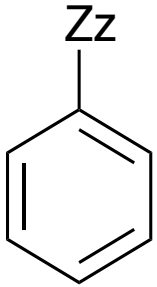
MarkInChI=1B/C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H
<>C!C2H5Zz/c1-2-3/h2H2,1H3!C3H7Zz/c1-2-3-4/h2-3H2,1H3



# Simple Variability: *MarkInChI*




Pseudo-element introduced in InChI v1.06

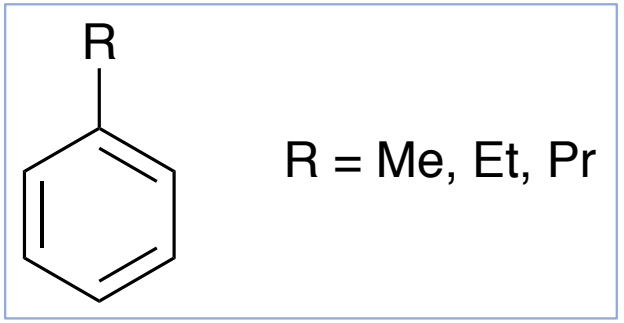


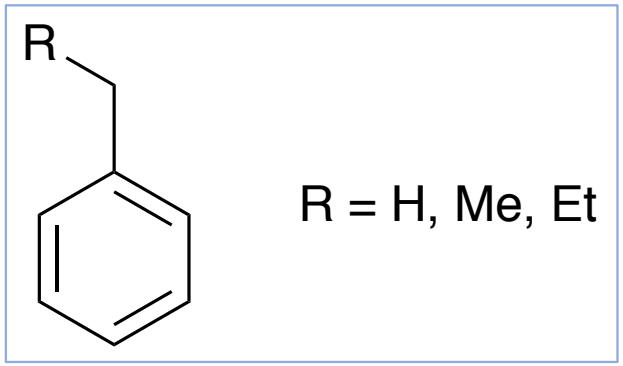

InChI=1B/C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H

A new section lists options for the pseudo-element, in alphabetical order, separated by "!"

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

MarkInChI=1B/C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H
<>C!C2H5Zz/c1-2-3/h2H2,1H3!C3H7Zz/c1-2-3-4/h2-3H2,1H3



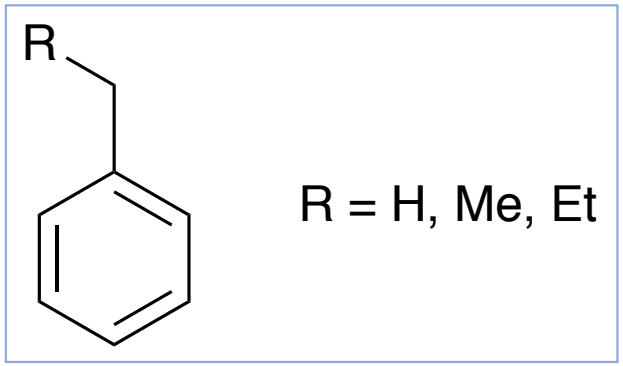



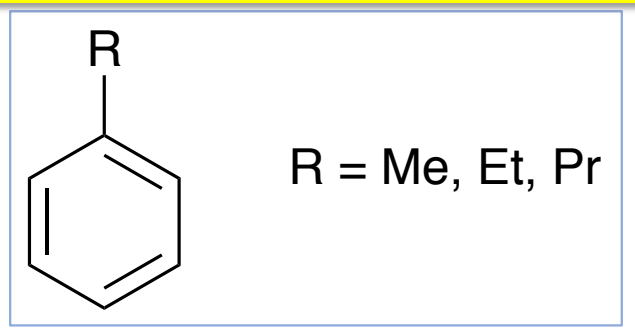

MarkInChI=1B/C7H7Zz/c8-6-7-4-2-1-3-5-7/h1-5H,6H2
<>H!C!C2H5Zz/c1-2-3/h2H2,1H3

# Simple Variability: *MarkInChI*

***Which is the canonical MarkInChI?***

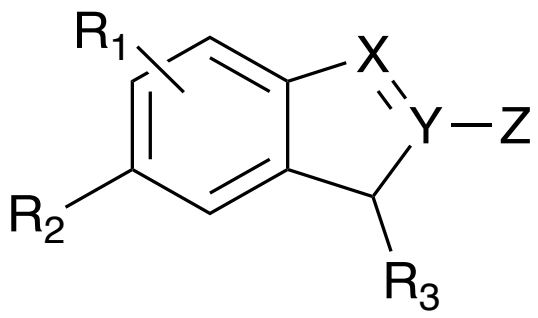
MarkInChI=1B/C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H
<>C!C2H5Zz/c1-2-3/h2H2,1H3!C3H7Zz/c1-2-3-4/h2-3H2,1H3




MarkInChI=1B/C7H7Zz/c8-6-7-4-2-1-3-5-7/h1-5H,6H2
<>H!C!C2H5Zz/c1-2-3/h2H2,1H3

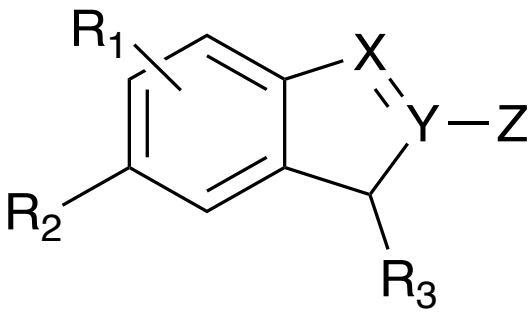
# More Complex Variability




$R_1$ = H, Me
$R_2$ = Me, Et, Pr, Bu
$R_3$ = Ph, tolyl
X = N, CH
Y = C, N+
Z = Cl, $CH_2Cl$

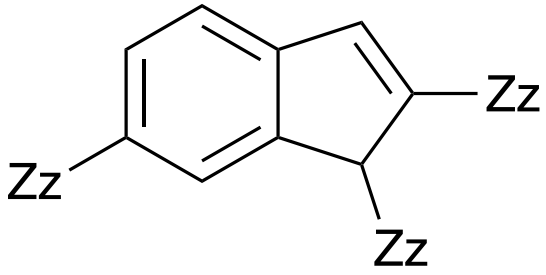
128 distinct structures

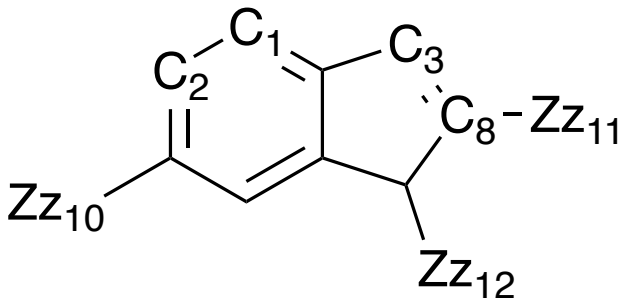
# More Complex Variability



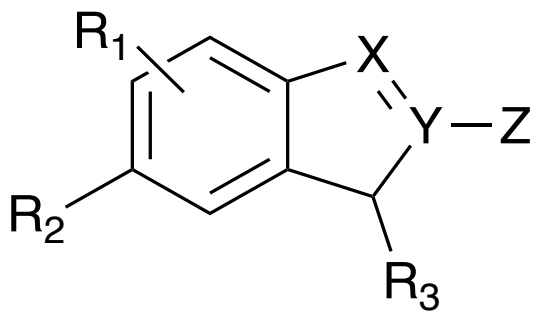

$R_1$ = H, Me
$R_2$ = Me, Et, Pr, Bu
$R_3$ = Ph, tolyl
X = N, CH
Y = C, N+
Z = Cl, $CH_2Cl$

128 distinct structures



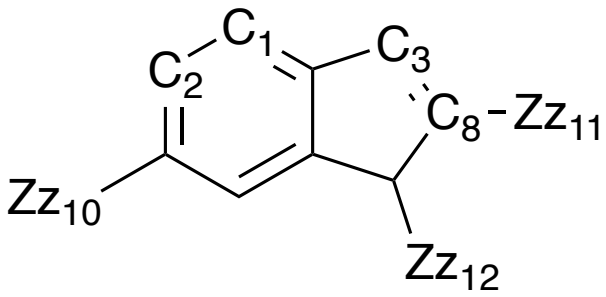



# More Complex Variability



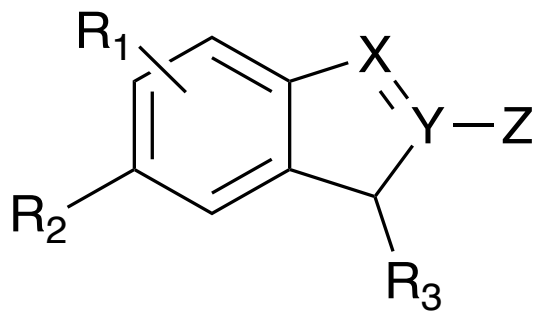

$R_1$ = H, Me
$R_2$ = Me, Et, Pr, Bu
$R_3$ = Ph, tolyl
X = N, CH
Y = C, N+
Z = Cl, $CH_2Cl$

128 distinct structures



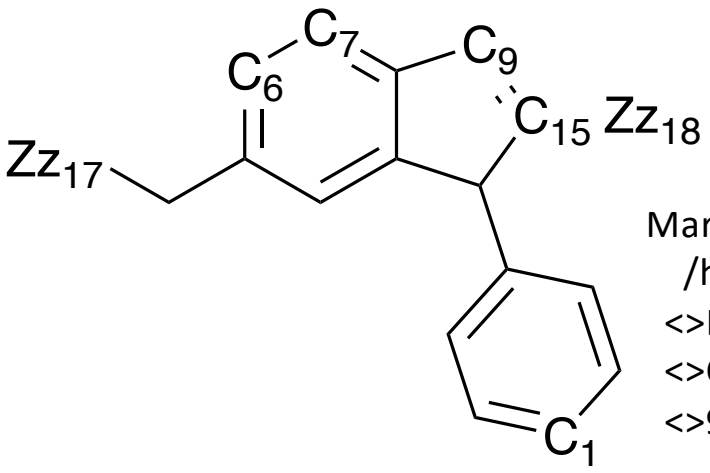

MarkInChI=1B/C9H5Zz3/c10-6-2-1-5-3-8(11)9(12)7(5)4-6/h1-4,9H
<>C!C2H5Zz/c1-2-3/h2H2,1H3!C3H7Zz/c1-2-3-4/h2-3H2,1H3
!C4H9Zz/c1-2-3-4-5/h2-4H2,1H3
<>CH2ClZz/c2-1-3/h1H2!Cl
<>C6H5Zz/c7-6-4-2-1-3-5-6/h1-5H!C7H7Zz/c1-6-2-4-7(8)5-3-6/h2-5H,1H3
<>3-C!N<>8-C!N+<>1H,2H-C!H

# More Complex Variability



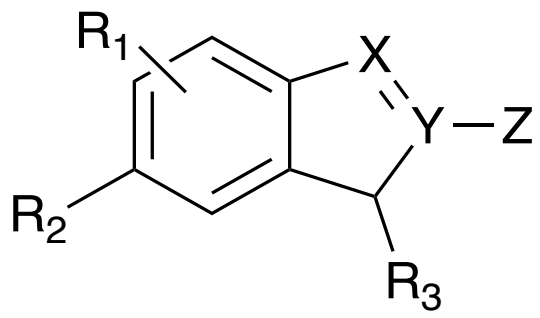

$R_1$ = H, Me
$R_2$ = Me, Et, Pr, Bu
$R_3$ = Ph, tolyl
X = N, CH
Y = C, N+
Z = Cl, $CH_2Cl$

128 distinct structures




MarkInChI=1B/C16H12Zz2/c17-10-11-6-7-13-9-15(18)16(14(13)8-11)12-4-2-1-3-5-12
/h1-9,16H,10H2
<>H!C!C2H5Zz/c1-2-3/h2H2,1H3!C3H7Zz/c1-2-3-4/h2-3H2,1H3
<>CH2ClZz/c2-1-3/h1H2!Cl
<>9-C!N<>15-C!N+<>1H-C!H<>6H,7H-C!H

# More Complex Variability

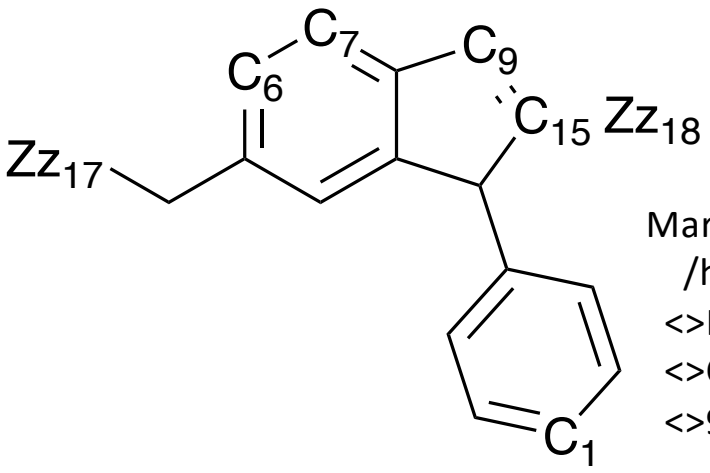



$R_1$ = H, Me
$R_2$ = Me, Et, Pr, Bu
$R_3$ = Ph, tolyl
X = N, CH
Y = C, N+
Z = Cl, $CH_2Cl$

128 distinct structures

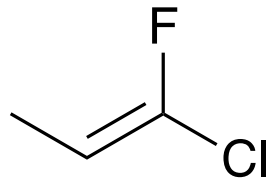
Do not need to use Zz – can just use the existing atom numbers

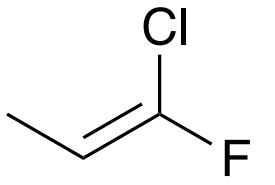
But the use of Zz makes the intentions of the writer clearer




MarkInChI=1B/C16H12Zz2/c17-10-11-6-7-13-9-15(18)16(14(13)8-11)12-4-2-1-3-5-12/h1-9,16H,10H2
<>H!C!C2H5Zz/c1-2-3/h2H2,1H3!C3H7Zz/c1-2-3-4/h2-3H2,1H3
<>CH2ClZz/c2-1-3/h1H2!Cl
<>9-C!N<>15-C!N+<>1H-C!H<>6H,7H-C!H

# Problem: different Zz create stereochemistry

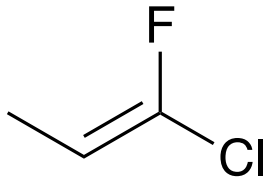


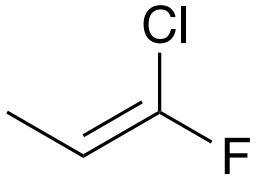

InChI=1S/C3H4ClF/c1-2-3(4)5/h2H,1H3/b3-2-

InChI=1S/C3H4ClF/c1-2-3(4)5/h2H,1H3/b3-2+

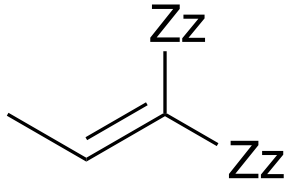
# Problem: different Zz create stereochemistry



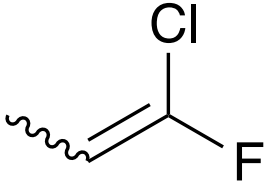

InChI=1S/C3H4ClF/c1-2-3(4)5/h2H,1H3/b3-2-




InChI=1S/C3H4ClF/c1-2-3(4)5/h2H,1H3/b3-2+




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

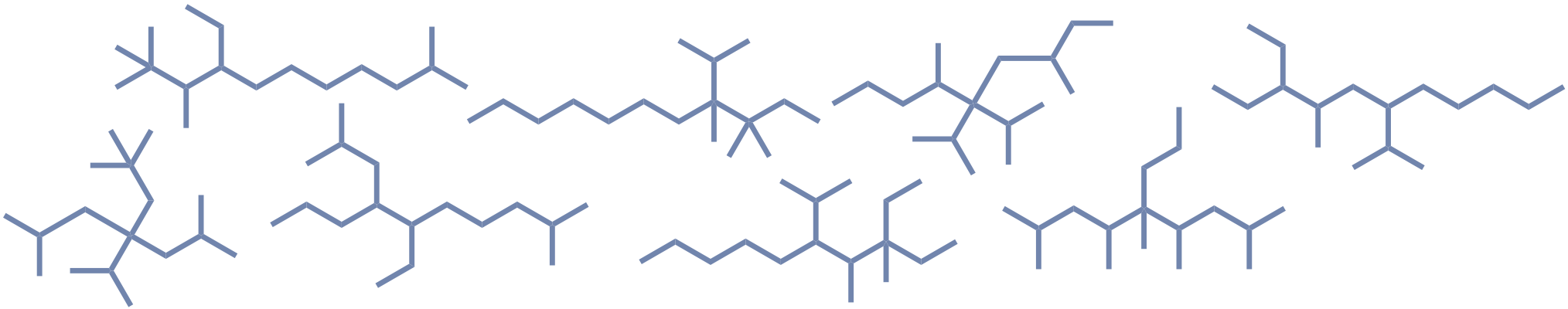



InChI=1B/C3H4Zz2/c1-2-3(4)5/h2H,1H3<>Cl!F<>Cl!F

# Variable InChI for Isomeric Alkanes (Anthony Baston)

- Input InChI:

```
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
```



# Variable InChI for Isomeric Alkanes (Anthony Baston)

- Input InChI:

```
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:

```
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

# Markush InChI

- Things to do:
  - Library of standard abbreviations: R, Ar, X, *etc*
  - Create a tool to visualize a Markush InChI
  - Create a tool to generate Markush InChI
    - By hand ; from molfile ; from a list of structures ; from other Markush InChI
  - Tools to manipulate Markush InChI (ideally without enumeration)
    - Structure searching within a Markush InChI
    - What do two Markush InChI have in common?
    - Search a database for members of a Markush InChI
  - Generic reactions (RInChI)
  - Compare two Markush InChI representations of the same system and choose the better one
  - Canonicalisation of Markush InChI

# Summary: Markush and Variable InChI

- A framework is been defined, based on InChI v1.06
- Canonicalisation and stereodifferentiation under examination
- Creating VInChI: provide a list of InChI and VInChI
- Creating MarkInChI: *currently hand-crafted by artisans*

- Optimisation needed: what best fits use-cases?
- Working group: 5 pm (UK), April 12th

# InChI

## Markush and Variability

**Working group: 5 pm (UK), April 12th**

*Jonathan M Goodman, Gerd Blanke,*
*István Öri, Anthony Baston*

*jonathan@inchi-trust.org*
*www-jmg.ch.cam.ac.uk/inchi*