Large molecules, variability, and more…

Evan Bolton, Ph.D. – Program Head of Chemistry

(evan.bolton@nih.gov)

U.S. National Library of Medicine
National Center for Biotechnology Information

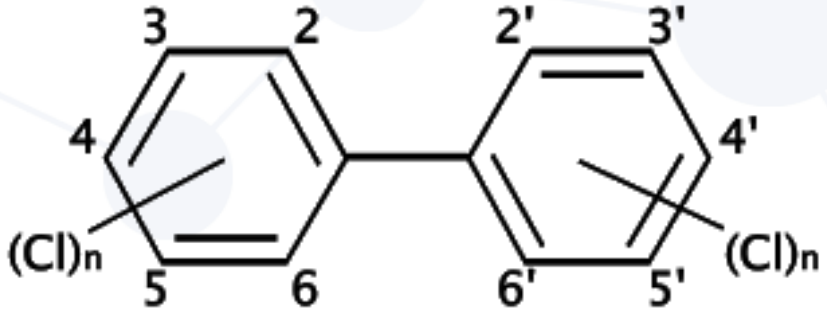
Many use cases of InChI

- Layered identifier
- Machine based
- Enhances FAIRness
- Open source
- Non-proprietary
- IUPAC Standard
- Hashed-form

- Discrete molecules
  - Molecules up to 32K atoms
  - Stereochemistry
  - Isotopic substitution
  - Mobile hydrogens
- Normalization
- Canonicalization
- Tautomers

Still much that InChI cannot handle

- Polymers (*fledgling support*)
- Mixtures (*proposal exists*)
- Molecular fragments (methyl)
- Variable attachment (PCBs)



- Small molecule space and large molecule space are similar yet different
  - Larger molecules tend to be harder to categorize
  - Know what is there but not always where or which
- …
- …

Image credit: https://sites.google.com/a/umich.edu/polychlorinated-biphenyls-pcbs/_/rsrc/1428007230337/characteristics-and-structure/Polychlorinated_biphenyl_structure.svg.png?height=144&width=320

Large molecules are very important

- Of the top-15 grossing drugs of 2018 .. 10 are biologics
- The biologics are chemically modified biopolymers and cannot be described by a sequence alone
- Active area of drug research
- There is a need to know when two are the ‘same’ (at some level)

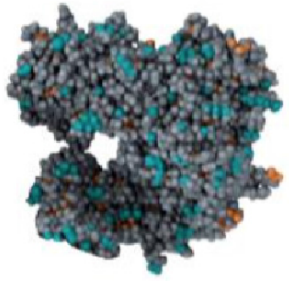
Small molecules vs. large (bio)molecules

Small molecule

Low molecular weight.
Chemically synthesised.
Well defined structure.

Biological molecule

High molecular weight.
Derived from living organisms.
Large and complex structure.



Monoclonal antibody

High molecular weight.
Derived from living organisms.
More complex structure.

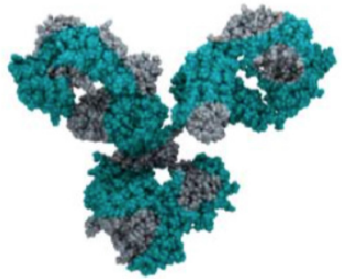
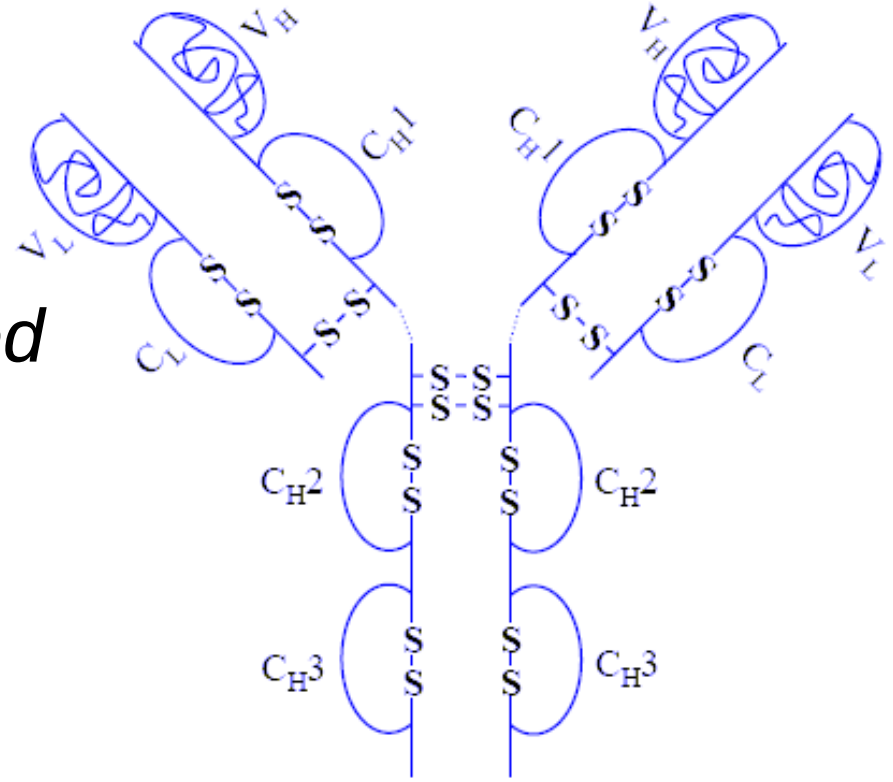


Image credit:
http://www.mabxience.com/wp-content/uploads/2017/09/170913_mAbxience_folleto-ext_gr%C3%A1ficos-web_eng_v2-01-1.png

Humira (adalimumab)

- *Recombinant monoclonal antibody directed to human TNF-α*
- *An IgG antibody composed of two kappa light chains* (each with a molecular weight of approximately 24 kDa) *and two IgG1z heavy chains* (each with a molecular weight of approximately 49 kDa)
- *Total molecular weight is 148 kDa.* Each light chain consists of 214 amino acid residues and each heavy chain consists of 451 amino acid residues



$C_{6428}H_{9912}N_{1694}O_{1987}S_{46}$

+20K atoms (within 32K InChI limit)

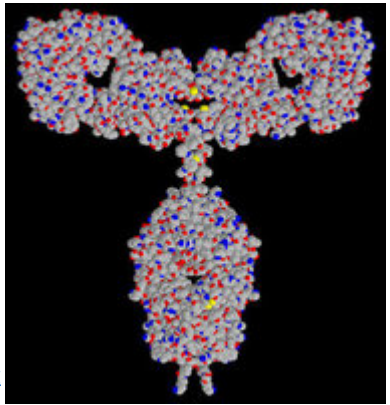
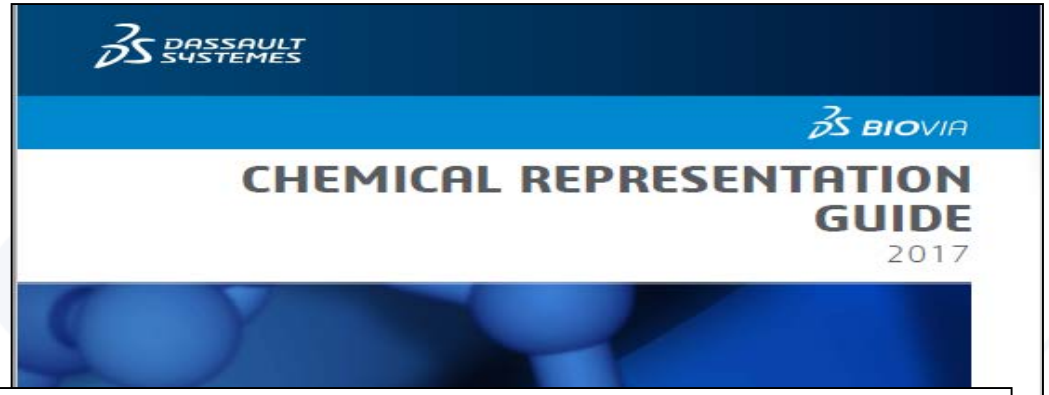


Image credit:
https://upload.wikimedia.org/wikipedia/commons/d/d2/Adalimumab_structure.png
http://www.gabionline.net/var/gabi/storage/images/media/images/adalimumab-2-v14a10/20732-1-eng-GB/Adalimumab-2-V14A10_medium.jpg

Large molecule formats



Extended SMILES, SMARTS

Codename: **cxsmiles,cxsmarts**

Contents:

Extended SMILES, SMARTS format

ChemAxon Extended SMILES/SMARTS is used for storing special features of the molecules after the SMILES string. Any information can be stored after the SMILES string if it is separated by space or tab characters as the SMILES parsers ignore them or use them as comment. The extended features are stored in the following format:

`SMILES_String |<feature1>,<feature2>,...|`

ChemAxon's extended SMILES/SMARTS does not contain non-ASCII characters, they are escaped in the usual form, "&#*n*;", with their character code, *n*. The ASCII characters ',', ';', '|', '{', '}' and ':' in Data Sgroup information are also escaped in this way. Moreover, the symbols '$', ';', '|', '{', '}' between dollar signs (see Atom labels / aliases / values.) are coded in the above mentioned way as well.

The extended feature description is economic. If some feature is missing in the molecule, then the corresponding special characters are not written. (Eg: If the atoms of the molecule has no alias strings at all, no "$" and ";" characters are written.) Moreover, if no feature of the molecule to be written, the extended feature field is omitted.

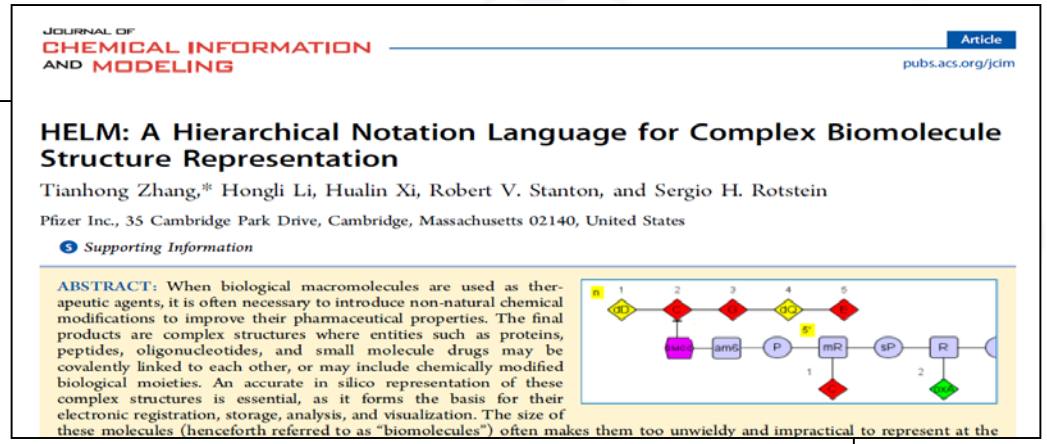
Please note that the SMILES string part generated in cxsmiles format is not always the same as the one generated by smiles output. Eg: In case of Ferrocene the coordinate bonds are not exported to plain SMILES ([Fe].c1cccc1.c1cccc1), but they appear in the cxsmiles (c12c3c4c5c1[Fe]23451234c5c1c2c3c45 |C:4.5,0.6,1.7,2.8,3.9,7.12,6.10,9.16,10.18,8.14|).

In extended smiles export the following additional features are exported:

- All aromatic atom are exported with lowercase letter in the SMILES string part.
  E.g. aromatic Boron is written with lowercase letter: b1ccccc1.
- Molecule absolute stereoconfiguration (For detailed description see the Stereochemistry section of the Query guide in JChem Base.)

  The relative stereoconfiguration is stored as "**r**". If a reaction contains components with absolute and relative stereo the indexes of the fragments with relative configuration is written. The absolute stereoconfiguration is the default, which is not marked. (Absolute stereoconfiguration known also as "Chiral flag" in MDL molfiles.)
  Example: "r:2,4,5"
- Enhanced stereochemical representation (For detailed description see the Stereochemistry section of the Query guide in JChem Base.)

  The following stereochemical group types are stored:



JOURNAL OF CHEMICAL INFORMATION AND MODELING — Article — pubs.acs.org/jcim

HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation

Tianhong Zhang,* Hongli Li, Hualin Xi, Robert V. Stanton, and Sergio H. Rotstein

Pfizer Inc., 35 Cambridge Park Drive, Cambridge, Massachusetts 02140, United States

Supporting Information

ABSTRACT: When biological macromolecules are used as therapeutic agents, it is often necessary to introduce non-natural chemical modifications to improve their pharmaceutical properties. The final products are complex structures where entities such as proteins, peptides, oligonucleotides, and small molecule drugs may be covalently linked to each other, or may include chemically modified biological moieties. An accurate in silico representation of these complex structures is essential, as it forms the basis for their electronic registration, storage, analysis, and visualization. The size of these molecules (henceforth referred to as "biomolecules") often makes them too unwieldy and impractical to represent at the

HELM Specification

"The limits of my language mean the limits of my world." Ludwig Wittgenstein

HELM Line Notation Extension - Ambiguity

Contents

Plus others .. E.g., FASTA

Large Molecules representation is (largely) a ‘solved’ problem

- Large molecule capabilities exist for in most commercial cheminformatic toolkits
  - Registration systems commonly available
  - Drawing packages abound
- Large molecule formats
  - MDL SDF
    - V2000/V3000 Sgroups / SCSR
  - ChemAxon Extended SMILES
  - HELM (V1/V2)
  - FASTA (w/ textual annotation)
  - PDB, mmCIF
  - IUPAC/IUBMB line notations, etc.

What can InChI offer large molecules?

- How can InChI offer value to this ‘crowded’ biopolymer field?
  - Identifier offers the same as a small molecule .. Normalization, canonicalization, identifier, hashed-key
- We would like to offer the same capabilities for large molecules as for small molecules
- Need extensions beyond discrete molecules as not all structural information is known (may be complex mixtures that can be readily described)

Problem space?

- What large molecule format (flavors) should InChI use as an input?
- To what degree can/should InChI handle the absence of (some) structural information?
- How should InChI canonicalize large molecule information content?
- What size molecule should InChI be able to handle?
  - (32K atoms → 32K residues/atoms?) (32K atoms → million atoms?)

Any we should add?

.Many. types of variability exist for large molecules

- Unknown biopolymer monomers or polymers
  - missing monomer, monomer mixture, or unknown
- Monomers/polymers in layers of AND or OR
  - with or without probabilities
  - mixtures
- Various levels of textual annotation
  - monomer, polymer, connection
- Numeric range of repeating units (A{23-25})

- Connection ambiguity
  - connected monomer type known, position is unknown
  - entities known but connection undefined
  - connection involves a polymer or group
  - connection involves a polymer and group
  - binding ratio between connected entities is unknown, decimal, or decimal range
  - connection between unknown polymers
  - polymer type known but sequence unknown
  - sequence partially known
  - component, connection and composition are variable

From Pistoia Alliance document: HELM Line Notation Extension - Ambiguity

Biopolymer handling

- Single attachment groups
  - Contains a single attachment atom for bonding to a biopolymer residue
  - Examples, PEG (polyethylene glycol), protecting groups (t-butyl, benzyl, FMOC, trityl)
- Abbreviated structures

SCSR – Hybrid representation

- Template representation of residues
- Small number of distinct monomer types but used many thousands of times
- Use of ‘*’ atoms (unknown or undefined)
- Use of psuedoatoms (dictionary of monomer representations, 1-3 letter codes, special connectivity semantics)

*atom approach

- Uses Sgroup fields MDL_RESIDUE_ATTACHMENT_ORDER and MDL_STARATOM_NAME
- Each residue consists of a single *atom with attached data that distinguishes chemically different residues
  - For example, ‘AA-THR’ to represent threonine

Pseudoatom approach

- Atom symbols that do not correspond to any of the chemical elements
- Uses Sgroup MDL_RESIDUE_ATTACHMENT_ORDER preserves information on the connectivity of atoms within the residue to structures outside it, such as protecting groups
- extended Ptable (size: ~200, can be modified)
  - three letter abbreviation (e.g., ALA)
  - condenses structure (no variability)
  - Ptable includes: 1-letter and 3-letter amino acid sequences, DNA, RNA
- Cannot be expanded as abbreviated structures can

Large molecule related proposal

‘Any’ atom

‘Monomer’ atom

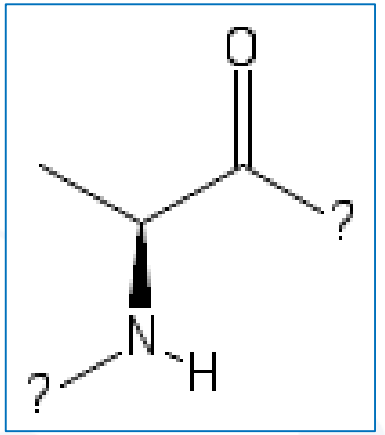
DyVinchi (Dynamic Variable InChI)

‘Any’ atom support

- **Methyl (‘*CH3’):** InChI=1B/CH3Zz/c2-1/h1H3
- **Methylene (‘*2CH2’):** InChI=1B/CH2Zz2/c3-1-2/h1H2
- **Ethyl (‘*CH2CH3’):** InChI=1B/C2H5Zz/c3-1-2/h1H2,2H3
- **Propyl (‘*CH2CH2CH3’):** InChI=1S/C3H7Zz/c4-1-2-3/h1-2H2,3H3

- Pseudo atom ‘any’ atom (‘Zz’)
- Treated like a defined atom
- Does not have stereo/isotope/hydrogen layer

‘Monomer’ atom support



- Three letter code (e.g., ‘Ala’)
  - First ‘A-Z’ else ‘a-z’
  - 17,576 possibilities
    - Can expand if more needed
- Internal knowledgebase
- Two connection points
  - Amino acid: N- and C-terminus
  - Nucleic acid: 3’- and 5’-terminus
- Non-standard InChI only
- Compact representation

- L-Alanine:

InChI=1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)/t2-/m0/s1 → InChI=1/Ala

- L-Alanyl-L-Alanine:

InChI=1S/C6H12N2O3/c1-3(7)5(9)8-4(2)6(10)11/h3-4H,7H2,1-2H3,(H,8,9)(H,10,11)/t3-,4-/m0/s1 → InChI=1/Ala2/c1-2

Dynamic Variability InChI (DyVinchi)

1. Define entities
   - in non-attached state
   - addressable by order
   - uses ‘any’ atom
2. Define attachments
   - heavy atom only
   - single valence only
3. Define variability
   - addressable by order
   - * spawns entity, 0=unkn

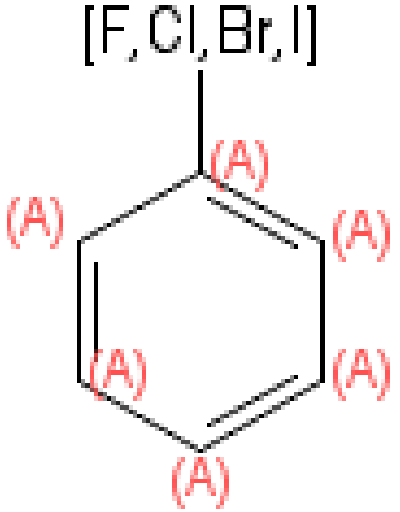
- Allows multiple attachments per entity
- Can nest variability
- Can provide counts
- Can provide ratios
- Handles Markush, variable connection, variable loading

The three parts of DyVinchi

- Define the unique set of molecular entities using an InChI surrounded by '[..]'. An 'any' atom ('Zz') defines each possible attachment.
- Define the unique set of attachment points
- Define the connectivity between attachments

'[#,#]' can be nested to any extent necessary, e.g., '[#,[#,#]', '[[#,#],#]', but always as a pair

Example: a **halogen** connected to **benzene**



[InChI=1B/C6H6Zz/c1-2-4-6-5-3-1-7/h1-6H] [InChI=1B/FZz][InChI=1B/ClZz] [InChI=1B/BrZz] [InChI=1B/IZz] |

[1,7][2,2][3,2][4,2][5,2] |

[1,[2,3,4,5]]

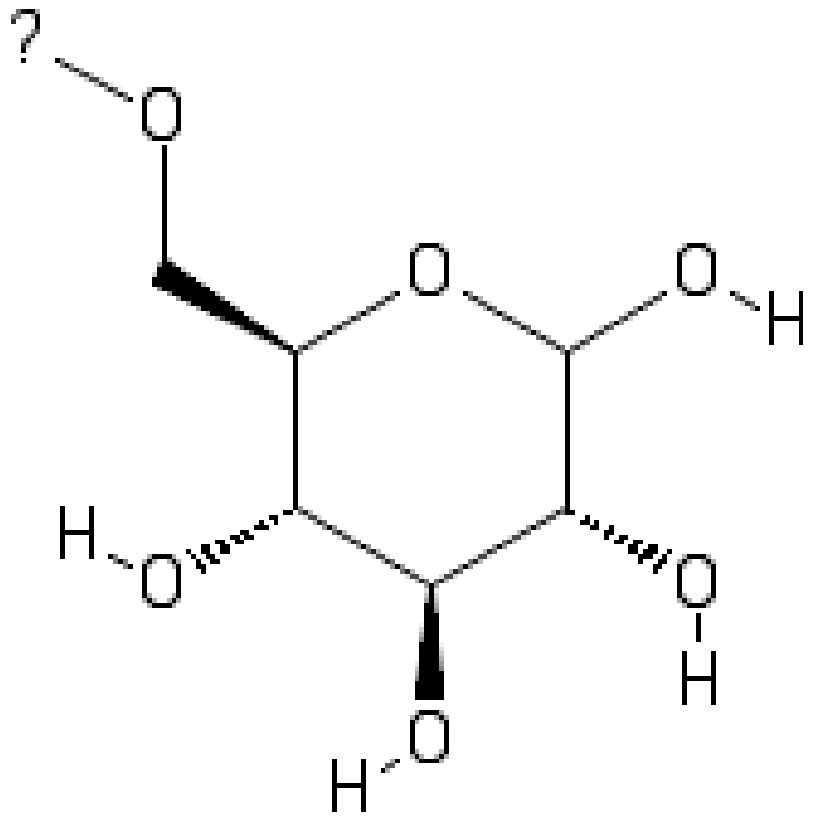
The DyVinchi above says: connect benzene to one of (fluorine, chlorine, bromine, and iodine).

Glucose bonded to undefined entity

[InChI=1B/C6H12O6**Zz**/c**13**-7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6?/m1/s1] | [1,**13**] | [1,0]

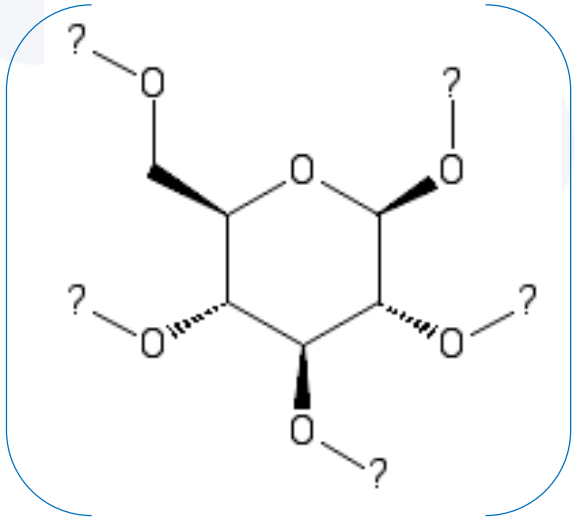
13 == Zz ‘any’ atom

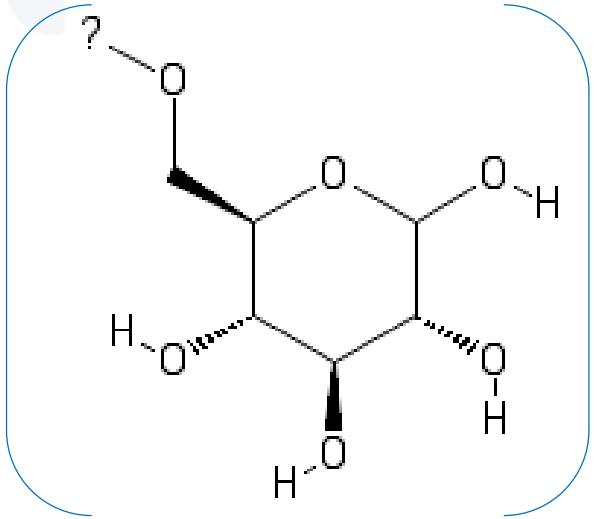
‘0’ is special molecular entity that means ‘unknown’ connection



Glucose (6 position) connected to any of the Glucose oxygens (1,2,3,4,6 position)

[InChI=1B/C6H12O6**Zz5**/c**13**-7-1-2-3(8-**14**)4(9-**15**)5(10-**16**)6(11-**17**)12-2/h2-11H,1H2/t2-,3-,4+,5-,6?/m1/s1]
[InChI=1B/C6H12O6**Zz**/c**13**-7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-,3-,4+,5-,6?/m1/s1]
| [1,[**13,14,15,16,17**]][2,**13**]
| [1,2]

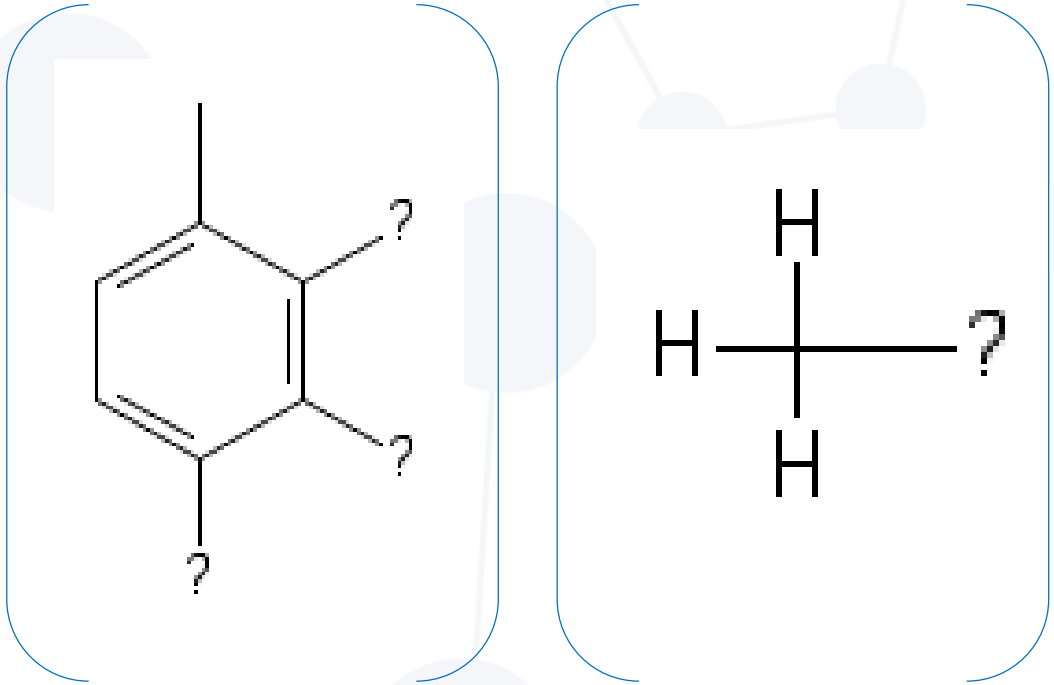




Xylenes

[InChI=1B/C6H6Zz3/c1-2(8)-4(9)-6(10)-5-3-1-7/h1-6H]
[InChI=1B/CH4Zz/c1-2/h1H4]
| [1,[8,9,10]][2,2] | [1,2]

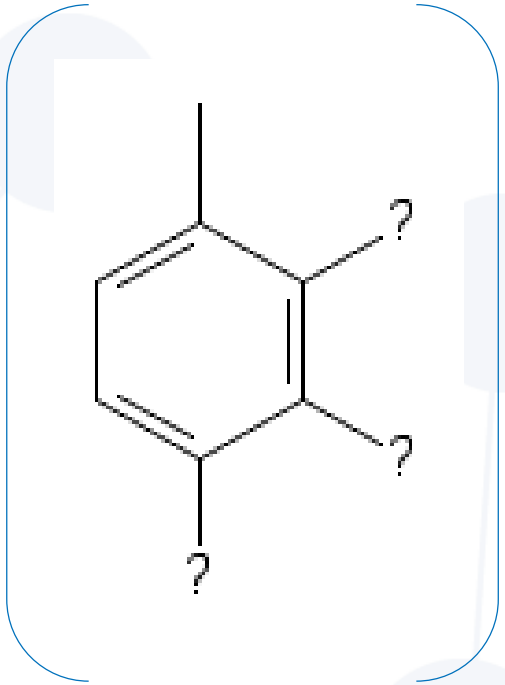
Toluene connected ortho, meta, or para to **methyl**

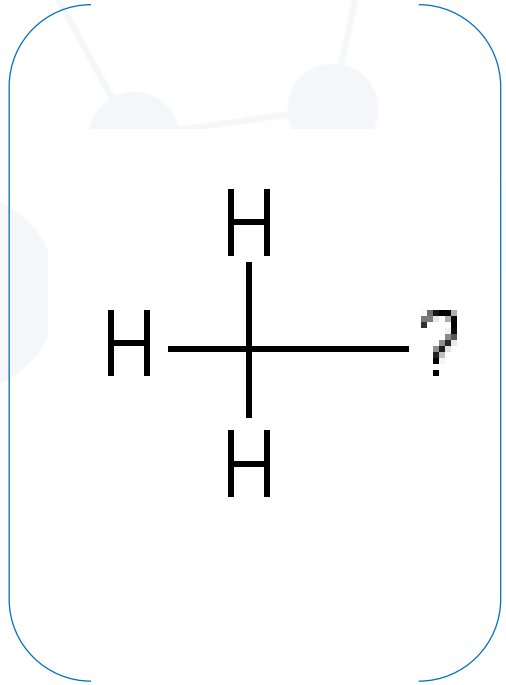


Xylenes w/ 20% ortho, 20% para, 60% meta

[InChI=1B/C6H6Zz3/c1-2(8)-4(9)-6(10)-5-3-1-7/h1-6H]
[InChI=1B/CH4Zz/c1-2/h1H4]
| [1,[8:20,9:20,10:60]][2,2]
| [1,2]

Optional ‘:’ indicates ratio





DyVinchi Specifics

- Define Entities
  - InChI with “[..]”
    [InChI=1S/C6H6/c1-2-4-6-5-3-1/h1-6H]
  - Order determines ordinal
- Layered separator “|”
- Define attachments
  - [#,[#,#]][#,#]
  - First # is entity
  - Second # is atom ID in InChI, can be a “[..]” list
  - Atom ID “0” means undefined
- Define variability
  - Two parts .. Attachment definition .. Attachment to what
  - Attachment definition can be nested with “[..]” list
  - Simplest is [#,#]
  - First number is entity
  - Second # is atom ID
  - Every # can be nested
  - Atom ID can include a ratio with “:” separator
    - E.g., “[2:1]” atom 2, 1 part

Dynamic Variability InChI – (Dy)Vinchi

1. Define entities
   - in non-attached state
   - addressable by order
2. Define attachments
   - heavy atom only
   - single valence only
3. Define variability
   - addressable by order
   - * spawns entity, 0=unkn

- Allows multiple attachments per entity
- Can nest variability
- Can provide counts
- Can provide ratios
- Handles Markush, variable connection, variable loading

DyVinchi Specifics

- Define Entities
  - InChI with “[..]”
    [InChI=1S/C6H6/c1-2-4-6-5-3-1/h1-6H]
  - Order determines ordinal
- Layered separator “|”
- Define attachments
  - [#,[#,#]][#,#]
  - First # is entity
  - Second # is atom ID in InChI, can be a “[..]” list
  - Atom ID “0” means undefined
- Define variability
  - Two parts .. Attachment definition .. Attachment to what
  - Attachment definition can be nested with “[..]” list
  - Simplest is [#,#]
  - First number is entity
  - Second # is atom ID
  - Every # can be nested
  - Atom ID can include a ratio with “:” separator
    - E.g., “[2:1]” atom 2, 1 part

Thank you.

This research was supported (in part) by the Intramural Research Program of the NIH, National Library of Medicine.

All collaborators and contributors.

Contact me if you would like more information:

evan.bolton@nih.gov