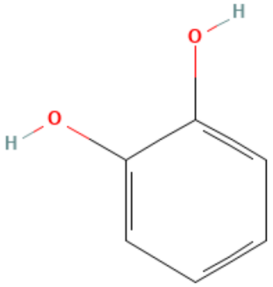
Heinrich Heine
Universität Düsseldorf

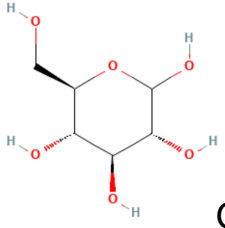
Representations of small molecules

Molecular Graphs

- A molecular graph represents molecules, with atoms represented as nodes and bonds represented as edges between nodes.



Catechol (C6H6O2)



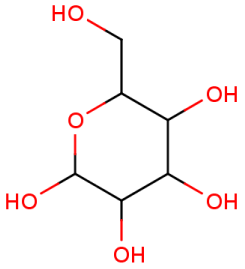
Glucose (C6H12O6)

- Rules for Drawing Molecular Graphs:
  - **Nodes**: Each node represents an atom (typically labeled with the element symbol)
    Exception: sometimes carbon atoms do not have carbon symbol “C”
  - **Bonds:** Each edge represents a bond between atoms
    - Number of parallel lines indicate type of bonds
  - **Hydrogen on Carbon:** Hydrogens attached to carbon atoms are generally not shown.
  - **Stereochemistry (**spatial arrangements of atoms):
    - Solid wedges: bonds coming out of the plane
      Hashed wedges (hashed triangles): bonds going behind the plane.
    - Plain Lines: bonds in the plane of the screen

SMILES

- SMILES - simplified molecular-input line-entry system
- A SMILES string represents the structure of a small molecule using a single string
  - Can consist of letters, numbers and special characters
- SMILES string for Glucose: C(C1C(C(C(C(O1)O)O)O)O)O

encodes this molecule



- Does not store stereochemical information yet, but we can add it

Image generated with: https://pubchem.ncbi.nlm.nih.gov

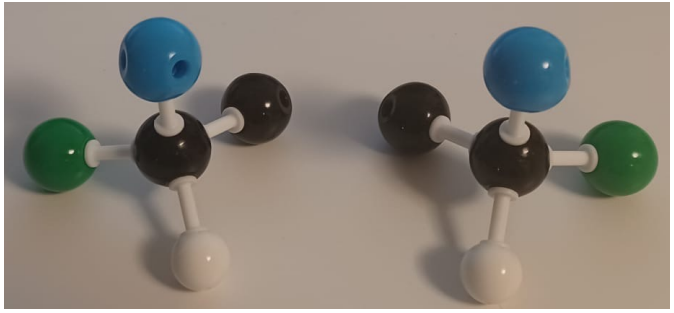
SMILES Specification Rules (1)

- Atoms are represented by their chemical symbols in square brackets, e.g. [Fe]
- Hydrogen atoms are usually implied and not explicitly written
- Brackets can be omitted if all of the following rules apply:
  - The atom is part of the following atoms: C, N, O, P, S, F, Cl, Br, B, or I
  - Has no formal charge
  - Number of attached hydrogen atoms can be implied
  - They are the normal isotopes
  - They are not chiral centers
    - Definition: A specific atom in a molecule that has four different groups attached to it, leading to molecules that are non-superimposable mirror images of each other.

[N-] ← Nitrogen with add. electron

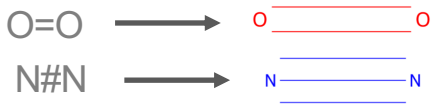
[OH-]

[13C]

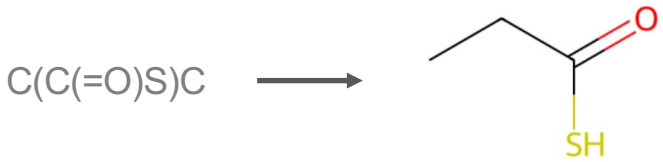


SMILES Specification Rules (2)

- Bonds:
  - Single bonds are not explicitly represented unless necessary (‘-’)
  - Double and triple bonds are represented by ‘=‘ and ‘#’

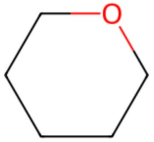


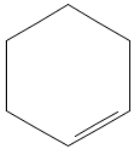
- Branches
  - described by using parentheses “(“ and “)”

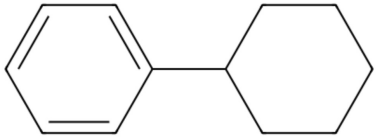


SMILES Specification Rules (3)

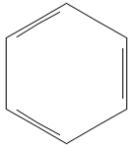
- Rings / cyclic structures are indicated by using numbers to represent the connection points. The same number is used for both ends of the ring.

C1CCCCO1 → 

C1=CCCCC1 → 

C1(CCCCC1)C2=CC=CC=C2 → 

- Atoms that are part of an aromatic bond are denoted by lower case letters

c1ccccc1 → 

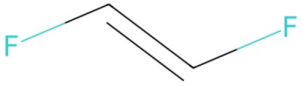
SMILES Specification Rules (4)

- Stereochemistry:
  - Chiral center: mirror images are not identical
  - @ and @@ indicate the arrangement of the four bonds attached to the chiral center
    - Consider the four bonds in the order in which they appear, left to right, in the SMILES string
    - Looking toward the central carbon from the perspective of the first bond
    - The following three bonds are counter-clockwise (@) or clockwise (@@)
  - Example:

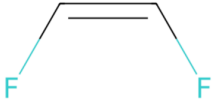
SMILES Specification Rules (5)

- Stereochemistry: trans and cis isomers
  - Arrangements of bonds typically around double bonds
  - Latin prefixes: “cis”=the side of; “trans”=the other side of
  - Can be specified by “/” and “\”
  - Two slashes are used to specify the stereochemistry:
    - Using the same slash twice indicates cis-molecule
    - Two different slashes indicate trans-molecule
  - Example:

F/C=C/F

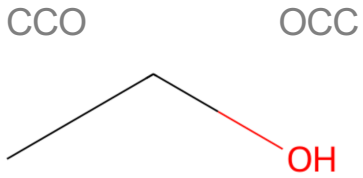


F\C=C/F



Canonical SMILES

- SMILES strings are not unique
  - For example, they differ depending on with which molecule you start



- Canonical SMILES: generating SMILES strings such that the same molecule always produces the same SMILES
  - involves applying a standardized set of rules and algorithms

Alternative small molecule representations

- InChI (International Chemical Identifier) strings:
  - A textual identifier that provides a unique description of a chemical substance
  - 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 is the InChI string for glucose

- Mol file format:
  - Describing chemical molecules.
  - Detailed information about the positions of atoms and the bonds connecting them:

```
  3  2  0  0  0  0  0  0  0  0999 V2000
   21.8400  -11.9918    0.0000 C   0  0  0  0  0  0
   20.6288  -12.6940    0.0000 O   0  0  0  0  0  0
   23.0512  -12.6940    0.0000 O   0  0  0  0  0  0
  1  2  2  0     0  0
  1  3  2  0     0  0
M  END
```


MOL file: https://www.genome.jp/entry/C00011