A chemical notation system: SMILES

Goals:

(1) uniqueness of the description of the molecule graph

(2) user-friendly / human readability

(3) machine-friendly

Rules:

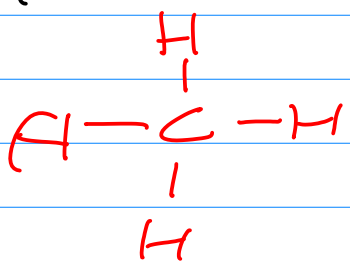
SMILES notation: a series of characters

(1) atoms represented by their atomic symbol, enclosed in square brackets [ ]

elements of the organic subset may be written without explicit hydrogen

B, C, N, O, P, S, F, Cl, Br

C

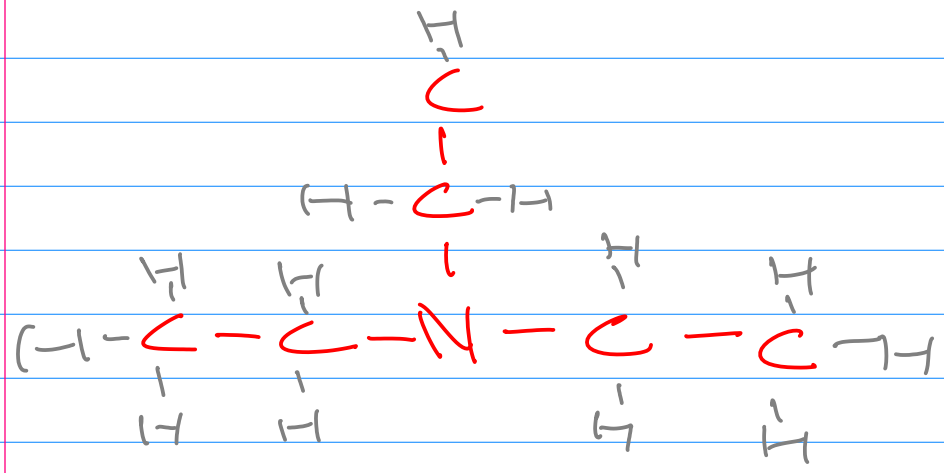


(2) bonds

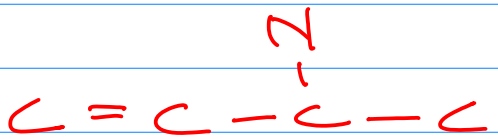
single bond	—	↓ may be omitted
double bond	=	
triple bond	#	
aromatic bond	:	

(3) Branches

encoded by enclosure parentheses



CCN(CC)CC



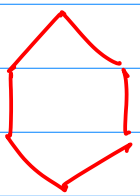
C=C-C(N)-C

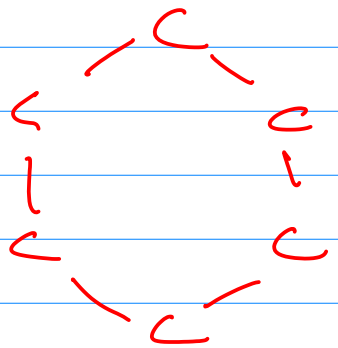
C=CC(N)C

CC(N)C=C

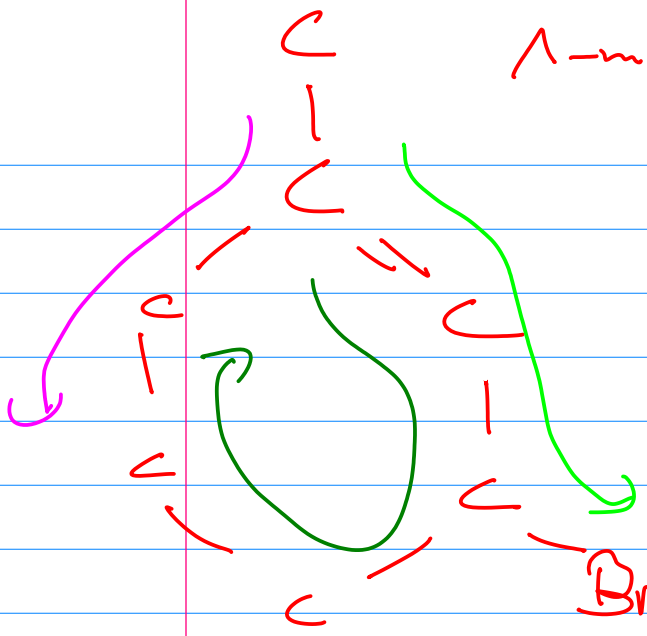
NC(C=C)-C

(4) cyclic structures





C1CCCCC1



1-methyl-3-bromo-cyclohexene

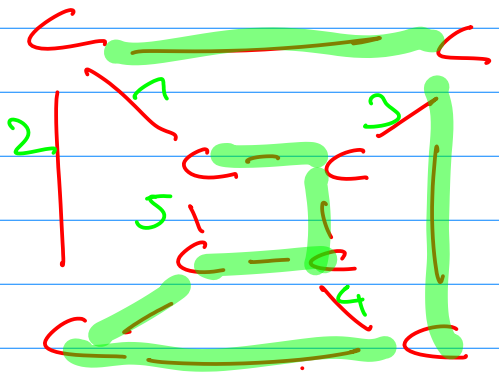
CC1=CC(Br)CCC1

CC1=CC(CCC1)Br

CC1CCCC(C

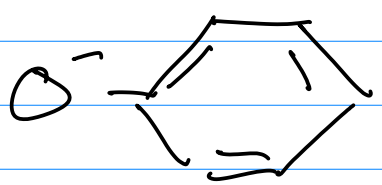
single atom might have many descriptors:

enclosure



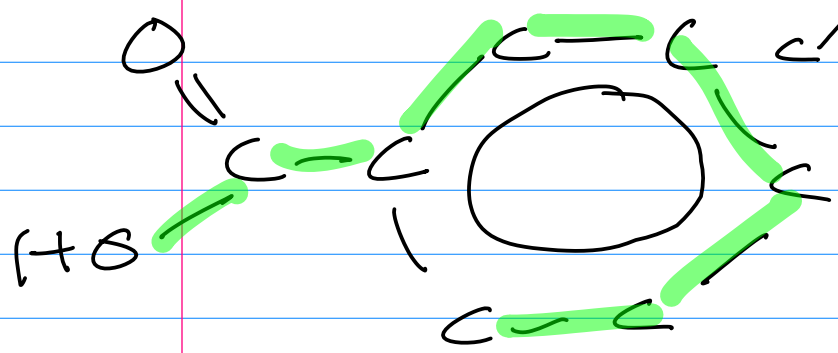
C12=C3C4C2C5C4C3C15

5) disconnected structures

Na+  [Na+].[O-]C1=CC=CC=C1

6) aromatic rings — small letters

benzoic acid



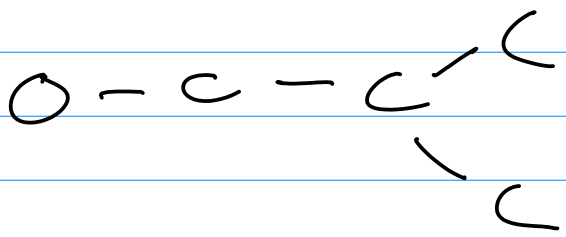
c1ccccc1C(=O)O

Generation of Unique SMILES Notation

method CANON and GENES

canonicalization — generation of unique SMILES

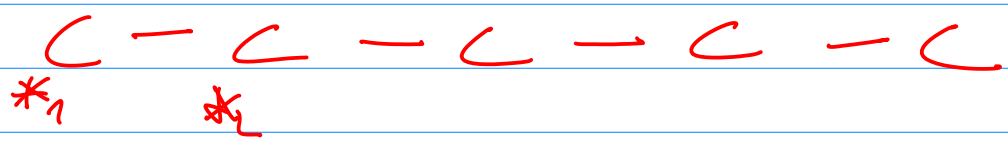
CANON must identify topologically equivalent in all respects molecules



CANGEN

(a) initial graph invariant

(1) # connections
(2) # non-H bonds
(3) atomic number
(4) sign of charge
(5) absolute charge
(6) # attached hydrogens



*1 (methyl carbon): 1,1,6,0,0,3
*2 (methylene carbon): 2,2,6,0,0,3

1160031 - 2260032 - 2260032 - ...

(b) rank equivalence

1 - 2 - 2 - 2 - 1

(c) simple extended connectivity

replace rank by (rank, $\sum$ rank neighbors)

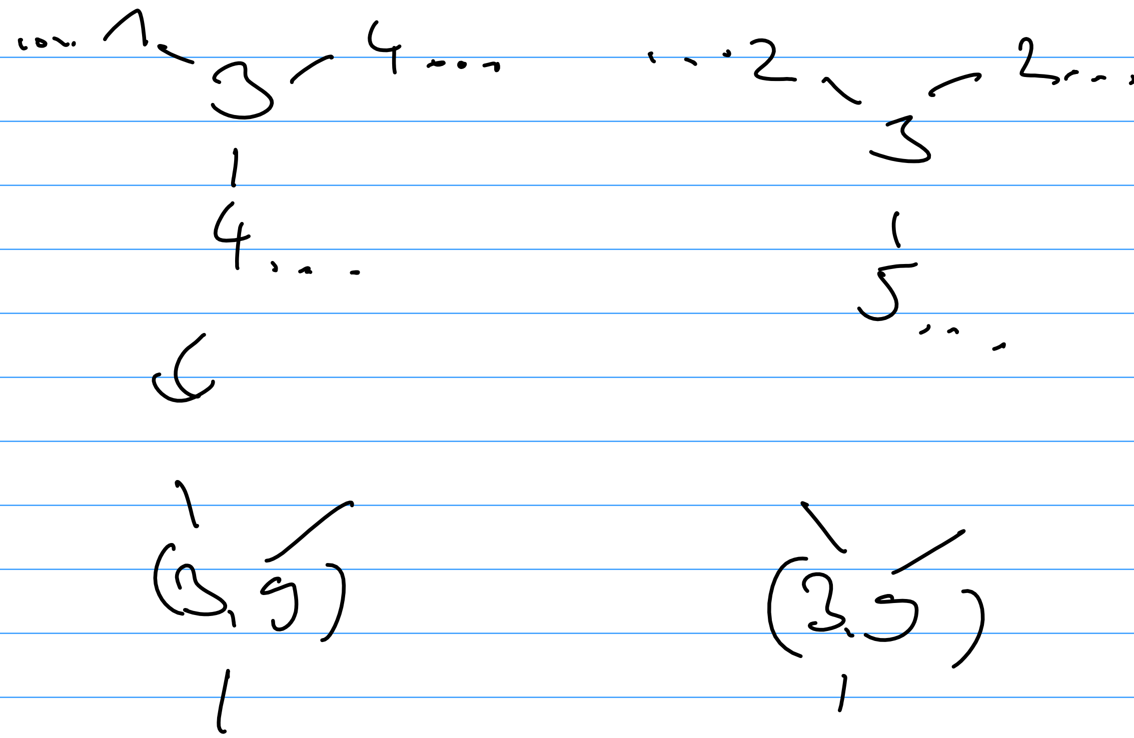
(1,2) - (2,3) - (2,4) - (2,3) - (1,2)

and replace:

1 - 2 - 3 - 2 - 1

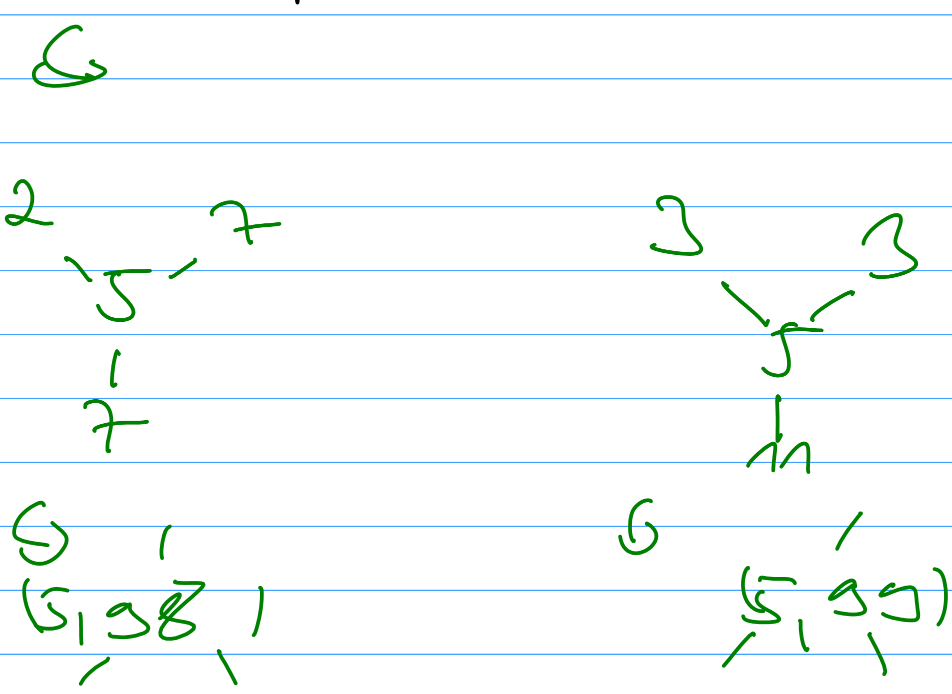
(d) extended connectivity using un-ambiguous functions

why:

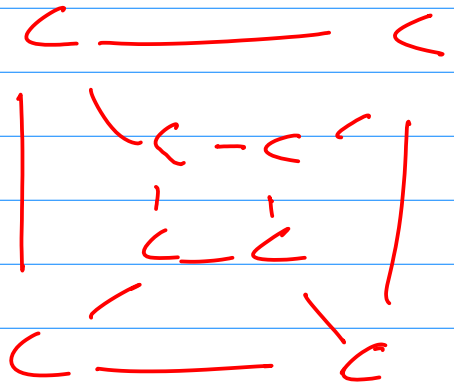


solution: - replace values by primes 2, 3, 5, 7, 11, ...

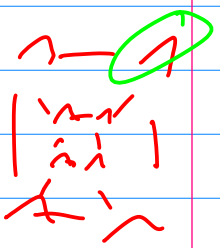
- use product instead of sums

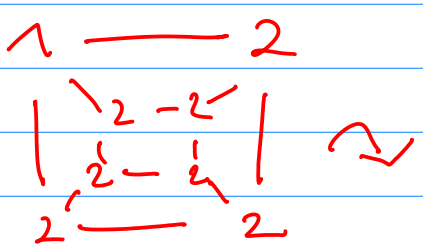


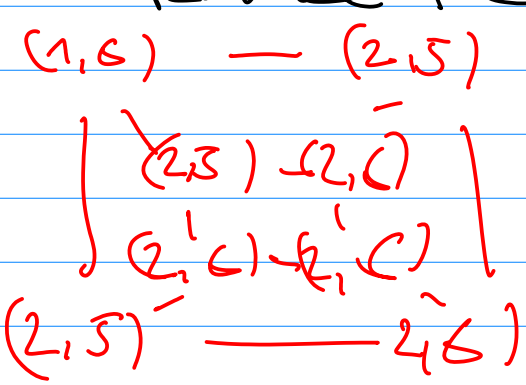
c) Scaling trees

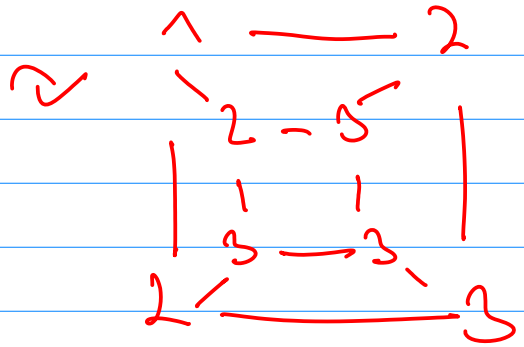


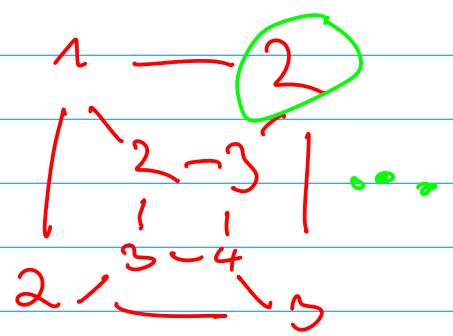
- choose an arbitrary scaling point (note, in culture now the remaining 7 C are no longer identical)
- double ranks and reduce the first by 1

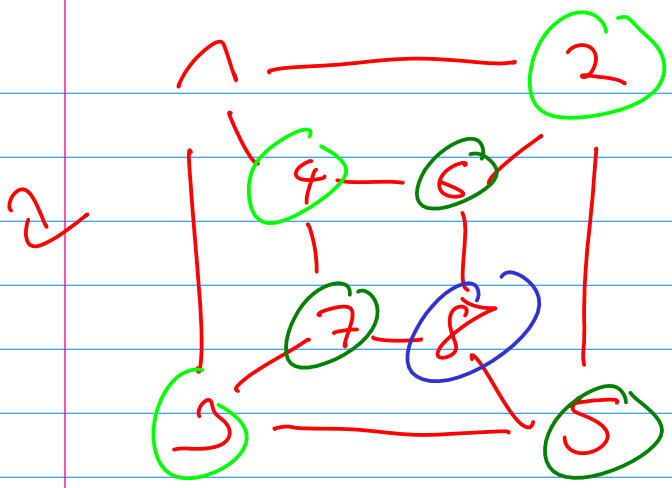












routine CANON N atoms

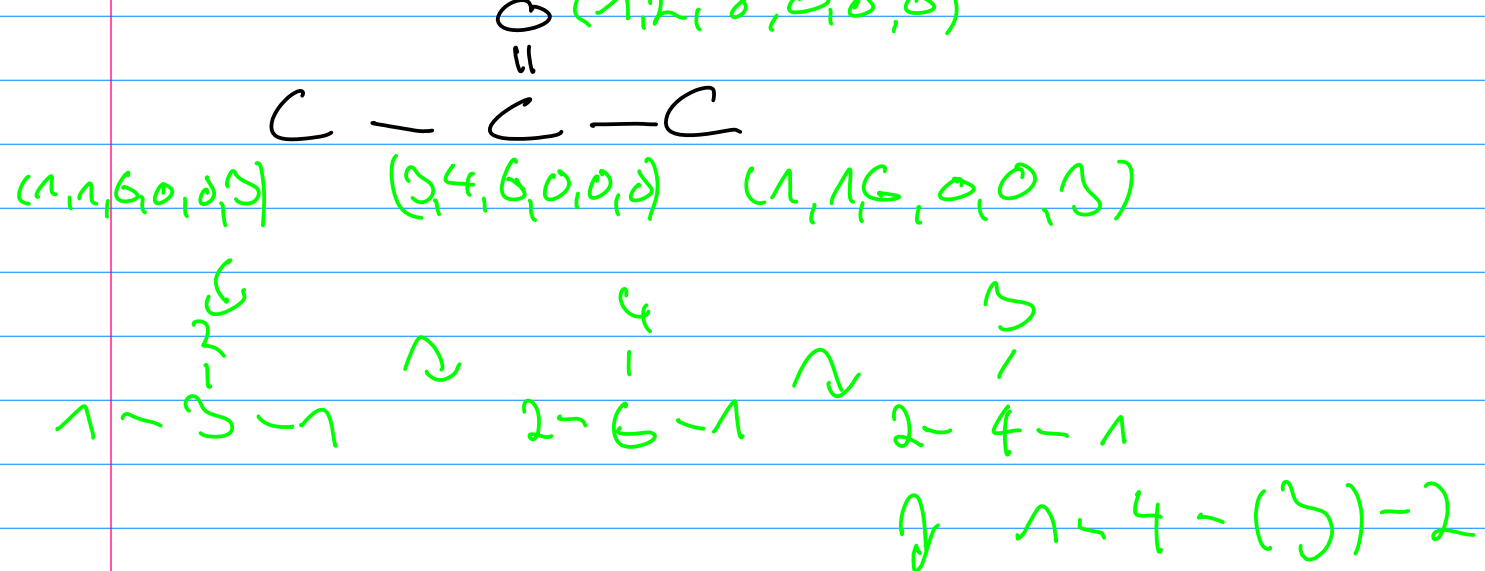
$O(N^2 \log N)$

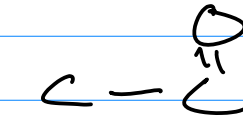
GENES : generation of unique SMILES

a) initial node ~ smallest canonical number

b) branching towards lower canonical number

Example: acetone



Therefore the canonical SMILES notation for  is CC(C)=O

and <u>not</u> CC(=O)C