سری کارگاه‌های یادگیری عمیق

نمایش ساختار مولکولی داروها در شبکه های عصبی گرافی

افروز راشدی

دانشجوی ارشد هوش مصنوعی

Deep Learning Workshops

Representing Molecular Structure of Drugs in Graph Neural Networks

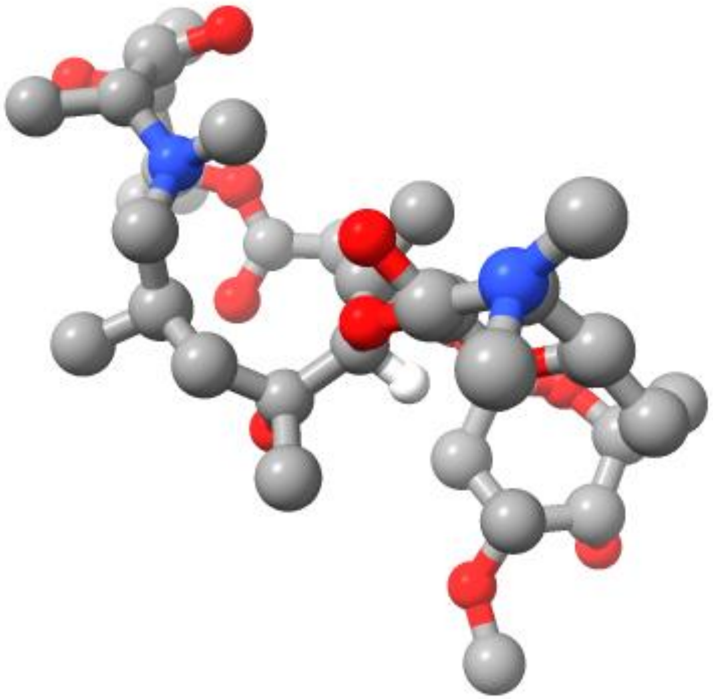
Afrooz Rashedi

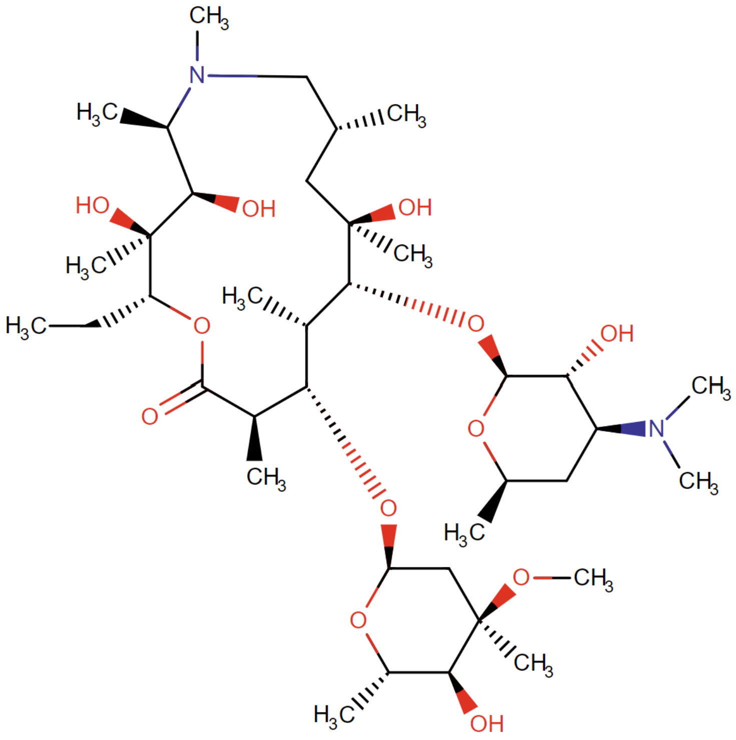
MSc. AI Student

Table of Contents

1. Drug Structure
2. Introduction to SMILES
3. SMILES Notation Rules
4. Canonical SMILES
5. Canonical SMILES Algorithm
6. Cangen Algorithm
7. Cangen Example
8. Related Contents

Drug Structure





What is SMILES?

Simplified Molecular Input Line Entry System

- Translate a chemical’s 3-dimensional structure into a string of symbols
- Simplicity and Readability to the human eye

SMILES Notation

Non-Hydrogen atoms represented by their atomic symbols

- Any unfulfilled valency of an atom is assumed to be Hydrogen

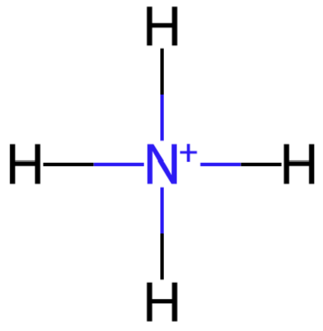
Example:

- Writing a simple C means a CH_4 (Methane)
  - A simple N is NH_3 (Ammonia)
  - A simple O is H_2O (Water)
- Representing elemental atoms by []
  - [C] is elemental Carbon

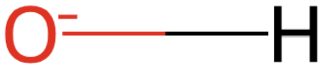
SMILES Notation

- Representing Charged Molecules
  - Using []
  - Positive charge represented by + sign
  - Negative charge represented by – sign

➢ Ammonium Cation
[NH4+]



➢ Hydroxyl Anion
[OH-]



SMILES Notation

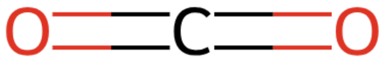
- Representing Bonds
  - Single -
  - Double =
  - Triple #
  - Aromatic :

➢ Single and Aromatic bonds are often omitted for simplicity

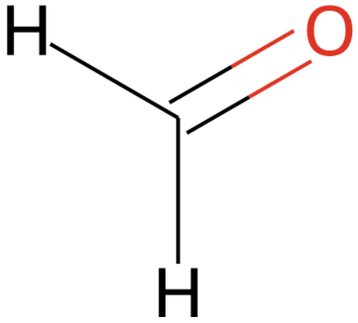
➢ Adjacent atoms assumed to be connected by a single or aromatic bond

SMILES Notation

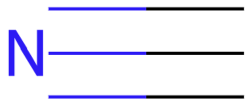
- Representing Bonds (Examples)
- Carbon dioxide (CO_2) O=C=O



- Formaldehyde (CH_2O) C=O



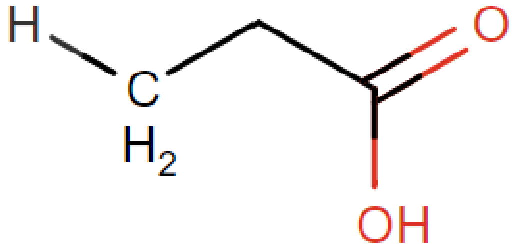
- hydrogen cyanide (HCN) C#N



SMILES Notation

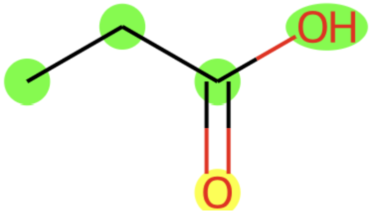
Representing Branches

- specified by enclosures in parentheses
- Propanoic acid $C_3H_6O_2$

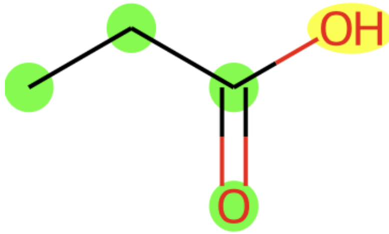


- Green highlighted atoms are taken as *base chain*
- Yellow highlighted atoms are *branch*

CCC(=O)O



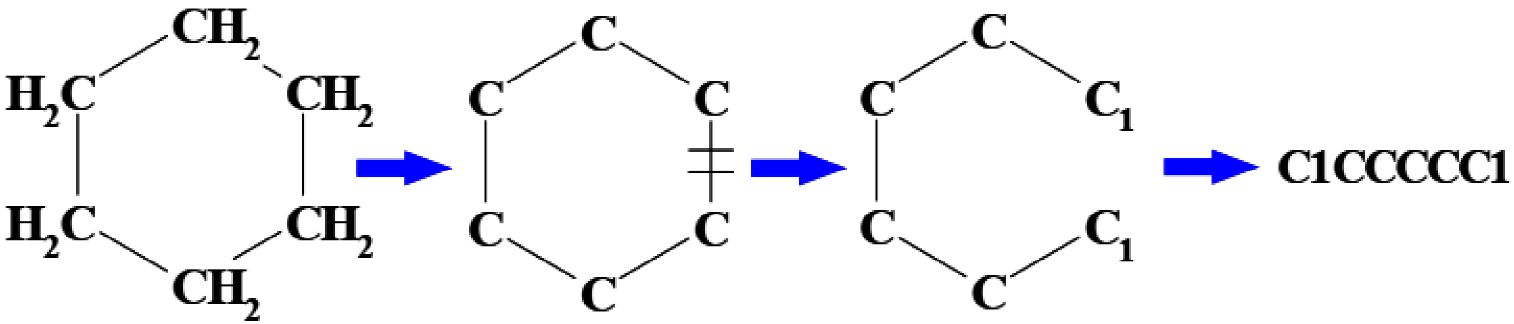
CCC(O)=O



SMILES Notation

Representing Cyclic Structures

- Break one *single* or *double (aromatic) bond* in each ring
- Bonds are numbered in any order, designating ring-opening/closure bonds by a digit immediately following the atomic symbol at each ring closure

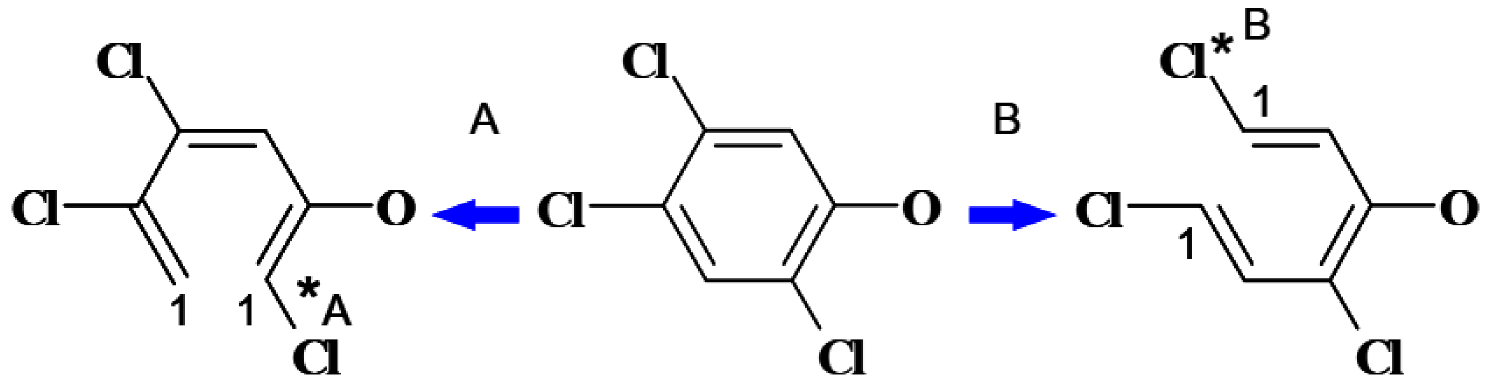


Cyclohexane 110-82-7

SMILES Notation

Representing Cyclic Structures

- Different SMILES notations for the *same structure*
- Breaking a ring in different places



SMILES A
c1(Cl)c(O)cc(Cl)c(Cl)c1

2,4,5-Trichlorophenol
CAS RN 95-95-4

SMILES B
Clc1cc(O)c(Cl)cc1Cl

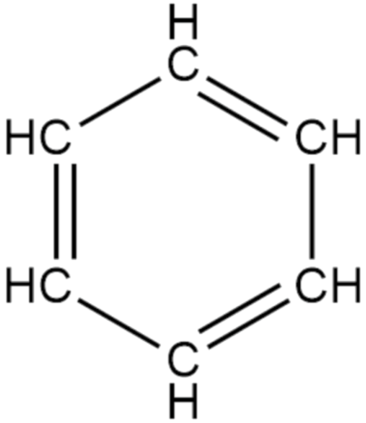
SMILES Notation

Aromatic Rings

- Lowercase letters tells us this is an aromatic ring signifying alternate single and double bonds

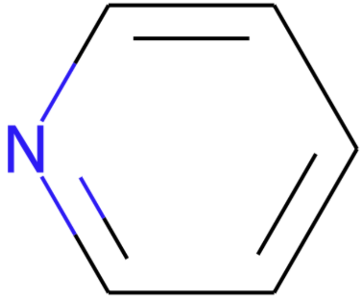
Benzene (C_6H_6)

c1ccccc1



Pyridine

c1ccncc1

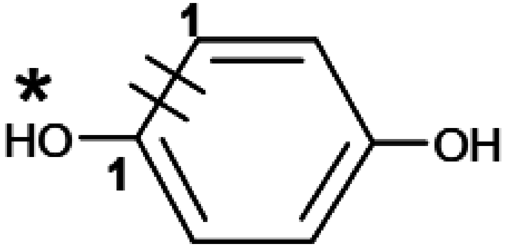


SMILES Notation

Nonaromatic Rings

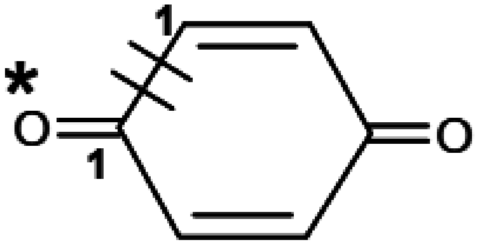
Non-aromatic: indicated by UPPER CASE

- Hydroquinone



aromatic O=c1ccc(O)cc1

- Quinone



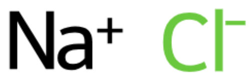
non-aromatic O=C1C=CC(=O)C=C1

SMILES Notation

- Disconnected Structures
  - Ions are not connected by a covalent bond
  - Disconnected compounds are separated by a “.”

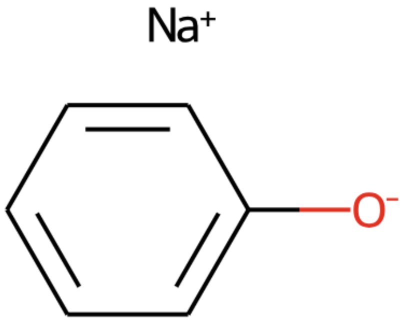
Sodium Chloride

[Na+].[Cl-]



Sodium Phenoxide

[Na+].[O-]c1ccccc1

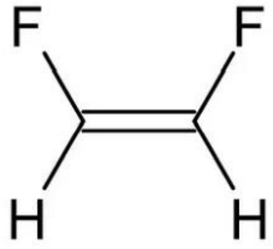


SMILES Notation

Isotopic Specification

- The *double bond* configuration is indicated by placing ‘/’ or ‘\’ between the atom constituting the double bond and their subsequent bonding partners.

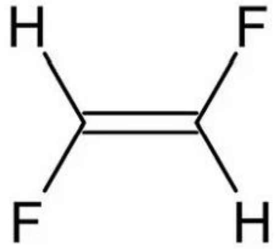
(Z)-1,2-Difluoroethylen



F\C=C/F

F/C=C\F

(E)-1,2-Difluoroethylen



F\C=C\F

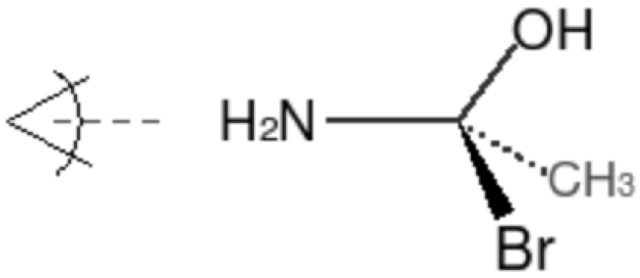
F/C=C/F

SMILES Notation

Tetrahedral Centers

- *anticlockwise*, an '@' is inserted after the central C-atom in []
- For a *clockwise* order, '@@' is inserted after the central C-atom in []

look from N towards C (chiral center)



list the neighbors anticlockwise

N[C@](Br)(O)C

...or clockwise

N[C@@](Br)(C)O

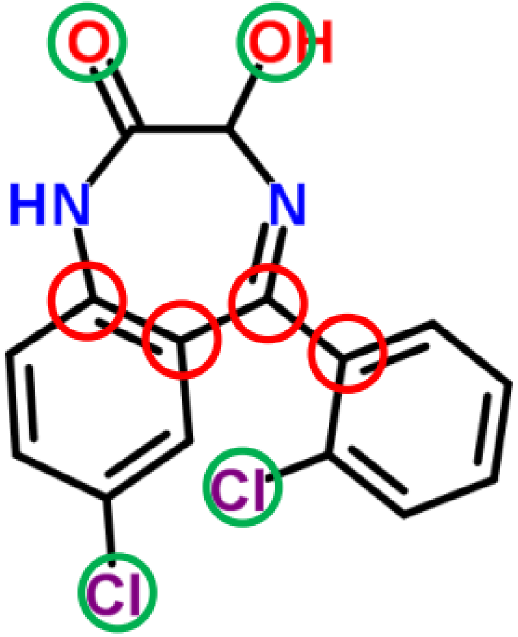
Disadvantages of SMILES string

- SMILES is not unique
- Doesn’t provide information about positions of each atom in space

Canonical SMILES

- ➢ Absolute or Unique SMILES
- ➢ Easy identification of identical molecules
- ➢ Solves "graph isomorphism" problem

- Cangen Algorithm
- A practical, efficient algorithm for generating canonical SMILES



Canonical SMILES Algorithm

Calculating the initial ranks of atoms:

1. Number of connections
2. Number of non-H bond orders
3. Atomic number
4. Sign of charge: 0 for non-negative, 1 for negative charge
5. Formal charge
6. Number of attached hydrogens
7. Isotope mass number

Cangen Algorithm

- Improved invariants
- “Stable” prioritization
- Avoids ambiguities from combining invariants
- Resolves “symmetric” atoms

Cangen Initial Invariants

❖ Initial invariants encode atom type information
  - Number of neighbors
  - Sum of bond orders
  - Charge
  - Number of attached hydrogens

❖ Initial invariants are transformed to ranks

Cangen Update Rule for Invariants

- ❖ Rank is mapped to corresponding prime (Carbon)
- ❖ New invariant:
  - primes of neighbors are multiplied

New ranks are determined based on:

- Old ranks
- New invariants

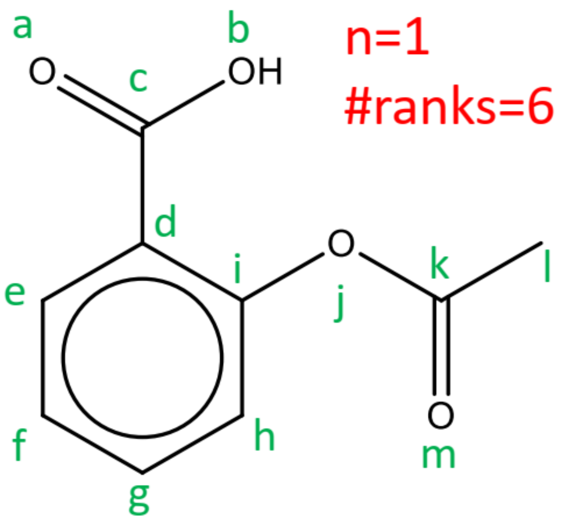
Cangen Iteration

- Repeat until ranking is stable
  - Calculate new invariants
  - Re-rank atoms

- Final ranking yields priorities
- Generate Smiles

Cangen Example

Initial invariants encode atom type information



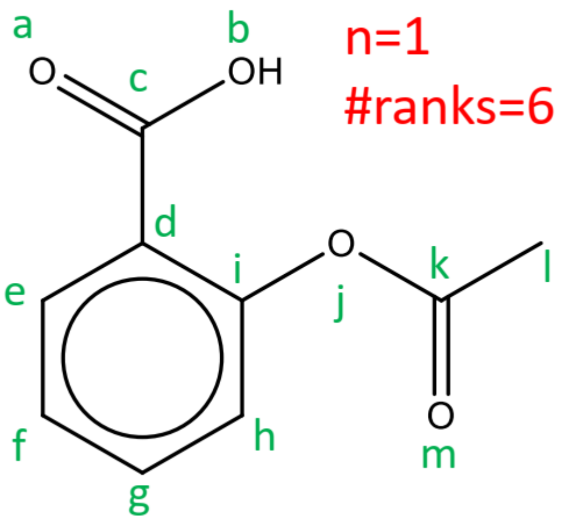
n=1

#ranks=6

Atom	#bds	Σ bds	At.Nr	Chg.	#H	rank
a	1	2	08	0	0	**3**
b	1	1	08	0	1	**2**
c	3	4	06	0	0	**6**
d	3	4	06	0	0	**6**
e	2	3	06	0	1	**5**
f	2	3	06	0	1	**5**
g	2	3	06	0	1	**5**
h	2	3	06	0	1	**5**
i	3	4	06	0	0	**6**
j	2	2	06	0	0	**4**
k	3	4	06	0	0	**6**
l	1	1	06	0	3	**1**
m	1	2	08	0	0	**3**

Cangen Example

Primes of neighbors are multiplied



n=1

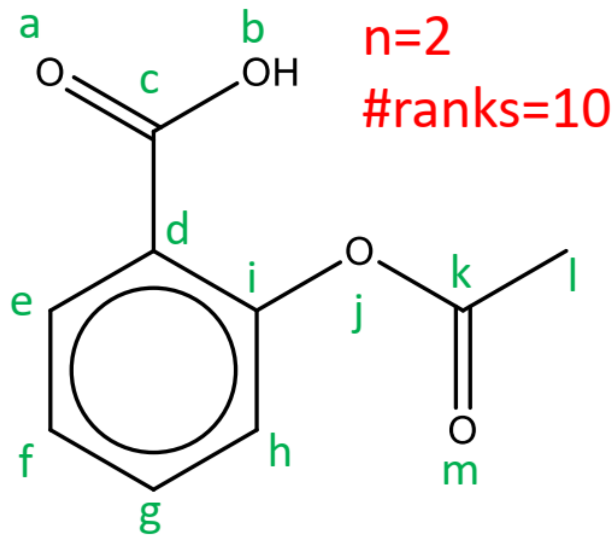
#ranks=6

Atom	rank	prime	Nbors	New Inv.
a	3	5	c	13
b	2	3	c	13
c	6	13	a,b,d	195 = 5*3*13
d	6	13	c,e,i	1889 = 13*11*13
e	5	11	d,f	143 = 13*11
f	5	11	e,g	121 = 11*11
g	5	11	f,h	121 = 11*11
h	5	11	g,i	143 = 11*13
i	6	13	d,h,j	1001 = 13*11*7
j	4	7	i,k	169 = 13*13
k	6	13	j,l,m	70 = 7*2*5
l	1	2	k	13
m	3	5	k	13

Cangen Example

Repeat until ranking is stable

Iteration 2



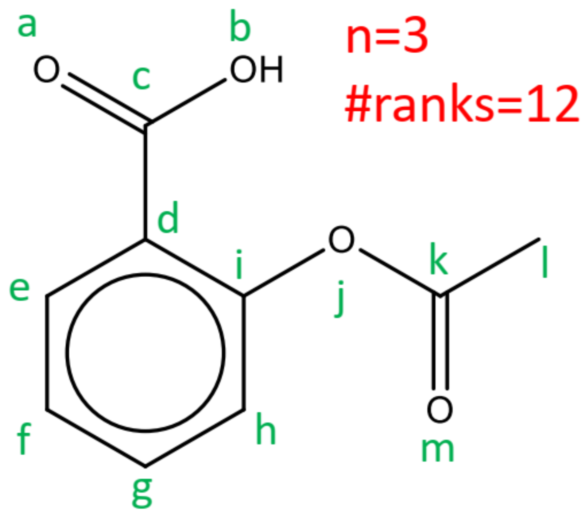
n=2

#ranks=10

Atom	rank	New Inv.	(rk.,inv.)	New rank
a	3	13	(3,13)	**3**
b	2	13	(2,13)	**2**
c	6	195	(6,195)	**8**
d	6	1889	(6,1889)	**10**
e	5	143	(5,143)	**6**
f	5	121	(5,121)	**5**
g	5	121	(5,121)	**5**
h	5	143	(5,143)	**6**
i	6	1001	(6,1001)	**9**
j	4	169	(4,169)	**4**
k	6	70	(6,70)	**7**
l	1	13	(1,13)	**1**
m	3	13	(3,13)	**3**

Cangen Example

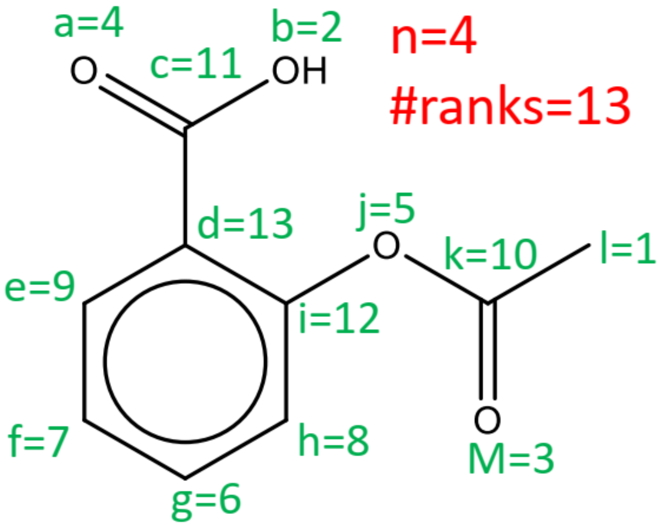
Iteration 3



Atom	rank	New Inv.	(rk.,inv.)	New rank
a	3	19	(3,19)	4
b	2	-	(2,-)	2
c	8	-	(8,-)	10
d	10	-	(10,-)	12
e	6	319	(6,319)	8
f	5	143	(5,143)	6
g	5	143	(5,143)	6
h	6	243	(6,243)	7
i	9	-	(9,-)	11
j	4	-	(4,-)	5
k	7		(7,-)	9
l	1	-	(1,-)	1
m	3	17	(3,17)	3

Cangen Example

Iteration 4 (Final ranking)



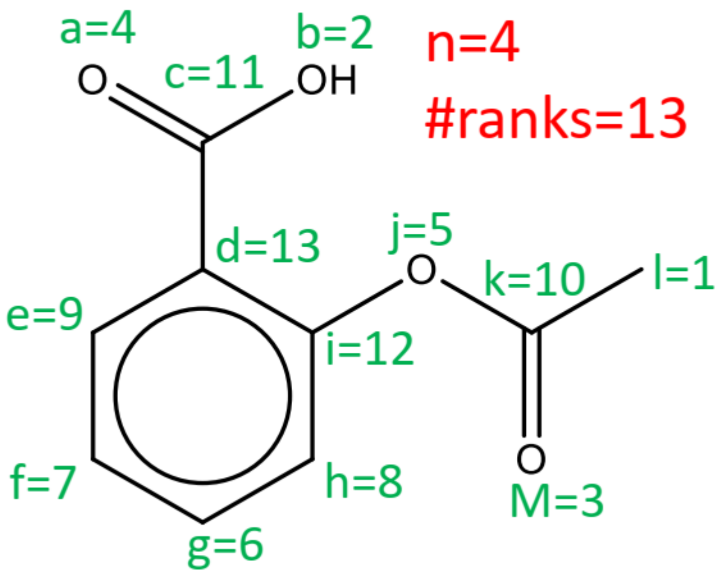
n=4

#ranks=13

Atom	rank	New Inv.	(rk.,inv.)	New rank
a	4	-	(4,-)	4
b	2	-	(2,-)	2
c	10	-	(10,-)	11
d	12	-	(12,-)	13
e	8	-	(8,-)	9
f	6	249	(6,249)	7
g	6	221	(6,221)	6
h	7	-	(7,-)	8
i	11	-	(11,-)	12
j	5	-	(5,-)	5
k	9		(9,-)	10
l	1	-	(1,-)	1
m	3	-	(3,-)	3

Cangen Example

Generating SMILES



n=4

#ranks=13

```
lk  m ji hgfed c a  b
CC(=O)Oc1ccccc1C(O)=O
```

Canonical Limitations

- Depends on the aromaticity model
- Different variations of the algorithms exist

Related Python Libraries

✓ RDKit

✓ pubchempy

✓ py3Dmol

Related Websites

- ✓ https://go.drugbank.com/
- ✓ https://pubchem.ncbi.nlm.nih.gov/
- ✓ https://chembl.gitbook.io/chembl-interface-documentation/downloads
- ✓ https://drugs.ncats.io/

➢ Online Tools:

Canonical SMILES generator:

https://www.antvaset.com/canonical-smiles-generator

SMILES checker:

https://www.cheminfo.org/flavor/malaria/Utilities/SMILES_generator__checker/index.html

Related Websites

Represented By:

Afrooz Rashedi

MSc student of AI

Feel Free To Contact Me:

afrooz.rashedi@gmail.com