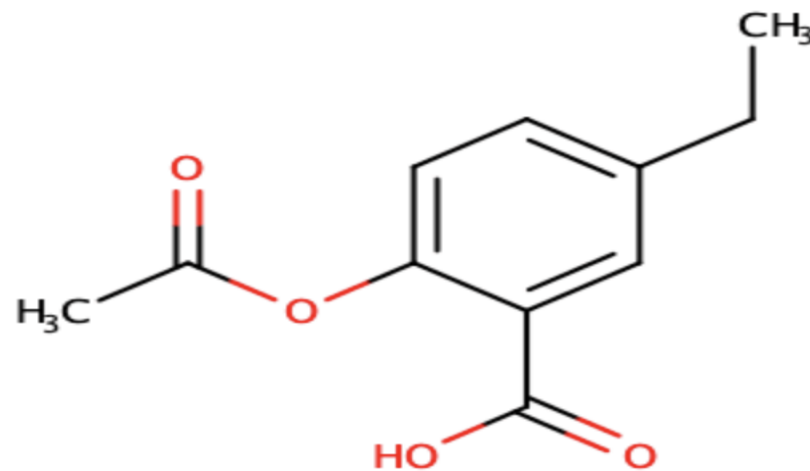
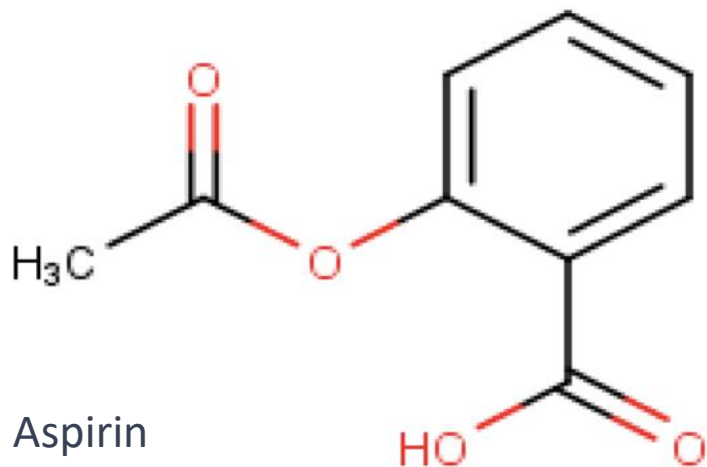


- AI 신약개발 경진대회 2nd

## **smiles code information**

**2024년 08월 09일**

# Smiles 표기법이란?



- Simplified Molecular Input Line Entry System의 약어로 molecule structure를 문자열 형태로 표기하여 머신러닝/딥러닝에 활용할 수 있게 표현하는 표기법
- 왼쪽 molecule는 CC(=O)OC1=CC=CC=C1C(=O)O 로 오른쪽 molecule 는 CCC1=CC=C(OC(C)=O)C(=C1)C(=O)O 와 같이 표현

# Smiles structure 를 구성하는 5가지 요소

1) Atom(원자)

2) 결합 (bond)

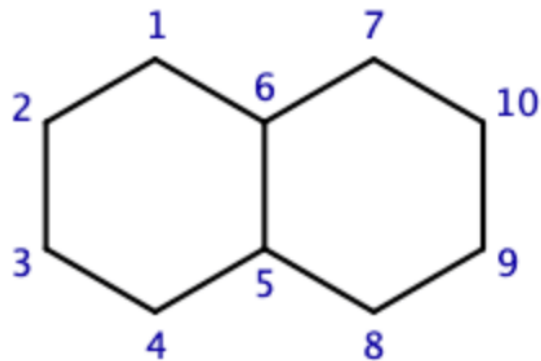
.	결합되지 않음	\$	사중결합
-	단일결합	:	방향족결합
=	이중결합	/	입체화학 표시
#	삼중결합	\	입체화학 표시

# Smiles structure 를 구성하는 5가지 요소

## 3) 고리 (ring)

고리가 시작되는 원자로부터 반시계방향으로 돌아가면서 표기하고, 고리의 시작 부분과 마지막 부분 원자 두 개에 번호를 표시합니다.

ex) C1CCCC2C1CCCC2



# Smiles structure 를 구성하는 5가지 요소

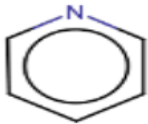

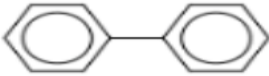
## 4) 방향족 (aromaticity)

벤젠과 같은 방향족 고리들을 아래와 같이 표현할 수 있습니다.

i) C1=CC=CC=C1 → 단일결합과 이중결합이 교대로 일어나는 Kekule form

ii) C:1:C:C:C:C:1 → 방향족결합 기호(:)를 사용

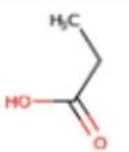

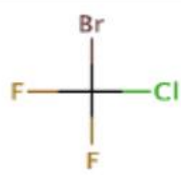
iii) c1ccccc1 → 방향족 고리에 포함된 B, C, N, O, P, S 원자를 소문자로 표시합니다.

분자	SMILES
	<chem>n1ccccc1</chem> (pyridine)
	<chem>o1ccccc1</chem> (furan)
	<chem>c1ccccc1-c2ccccc2</chem> (biphenyl)

# Smiles structure 를 구성하는 5가지 요소

## 5) 가지 (branch)

분자의 가지는 괄호()로 표현합니다. 가지는 다양한 순서로 나타낼 수 있으며 결합도 괄호 안에 포함되어 있어야 합니다.

분자	SMILES
	<chem>CCC(=O)O</chem> (propionic acid)
	<chem>FC(F)F</chem> (fluoroform)
	<chem>FC(Br)(Cl)F,</chem> <chem>BrC(F)(F)Cl,</chem> <chem>C(F)(Cl)(F)Br, ...</chem>

# IC50 을 구하기 위한 Smiles 코드로 얻을 수 있는 정보

## 1. Molecular structure information

- 약물의 표적이 되는 세포나 효소와의 결합력이 중요하기 때문에 이를 고려하여 약물의 분자 구조를 잘 설계해야 함
- 약물의 분자가 생체내에서 잘 녹는지 흡수가 잘되는지도 분자 구조로 결정됨

# IC50 을 구하기 위한 Smiles 코드로 얻을 수 있는 정보

## 2. Functional Groups(작용기)

### - Determining reactivity

작용기에 따라 분자가 물에 잘 녹을지 기름에 잘 녹을지 결정됨

➔ 약물이 몸속에서 어떻게 이동하고 어디로 도달할지를 결정

### - Binding affinity

작용기는 약물이 생물학적 표적(예: 단백질, 효소)과 얼마나 잘 결합할지를 크게 영향을 미치는데 알코올기(-OH)나 아민기(-NH<sub>2</sub>) 같은 작용기가 있으면, 약물이 표적과 강하게 결합할 수 있음

➔ 결합이 강할수록 약물의 결합 효과가 커지고, IC50 값이 낮아짐

### - 특정 반응 유도

- 산성 작용기와 pH에 따른 반응성

- 산성 작용기(-COOH, -SO<sub>3</sub>H 등)를 가진 약물은 산성 환경에서 더 잘 이온화(양성자 H<sup>+</sup>를 방출)되는데, 약물의 생물학적 표적(예: 단백질의 특정 부위)과 결합하는 데 중요한 역할을 함

->위는 산성 환경을 가지고 있어서, 산성 작용기를 가진 약물이 위에서 더 잘 작용할 수 있음



# IC50 을 구하기 위한 Smiles 코드로 얻을 수 있는 정보

- 3. 물리화학적 성질의 추출:
- 분자량: SMILES로부터 원자들의 수와 종류를 통해 분자량을 계산할 수 있음.
- LogP (분배 계수): 지용성과 수용성의 비율을 나타내며, 약물의 흡수, 분포, 대사 및 배설에 중요한 역할을 함.
- 극성 표면적 (PSA): 분자의 극성 부분의 표면적을 계산할 수 있으며, 이는 분자의 투과성에 영향을 미칠 수 있음.
- 수소 결합 공여자/수용자 수: 분자의 극성과 결합 가능성을 나타내는 중요한 요소로, 생물학적 활성을 결정하는 데 중요한 역할을 함.

# 정보 추출을 위한 라이브러리

- **RDKit**: 오픈소스 화학 정보 라이브러리로, SMILES 코드를 분자 구조로 변환하고 다양한 화학적 특성을 계산할 수 있음
- **Open Babel**: 다양한 화학 파일 형식을 변환하고, SMILES로부터 물리화학적 특성을 추출할 수 있는 소프트웨어
- **ChemAxon**: SMILES 코드를 사용하여 분자의 다양한 특성을 분석할 수 있는 상용 소프트웨어

# 데이터 수집 및 전처리 방법

1. SMILES 코드로부터 분자 특성을 추출.

여기에는 분자량, LogP, PSA, 수소 결합 공여자/수용자 수 등이 포함되어야함 .

2. 실험적으로 얻은 IC50 값을 타겟 변수로 사용하여 데이터셋 구축

-> 결측값 처리, 데이터 스케일링 등 전처리 과정을 수행하여 모델 학습에 적합한 형태로 데이터 준비