AI 기반 신약 후보물질 탐색

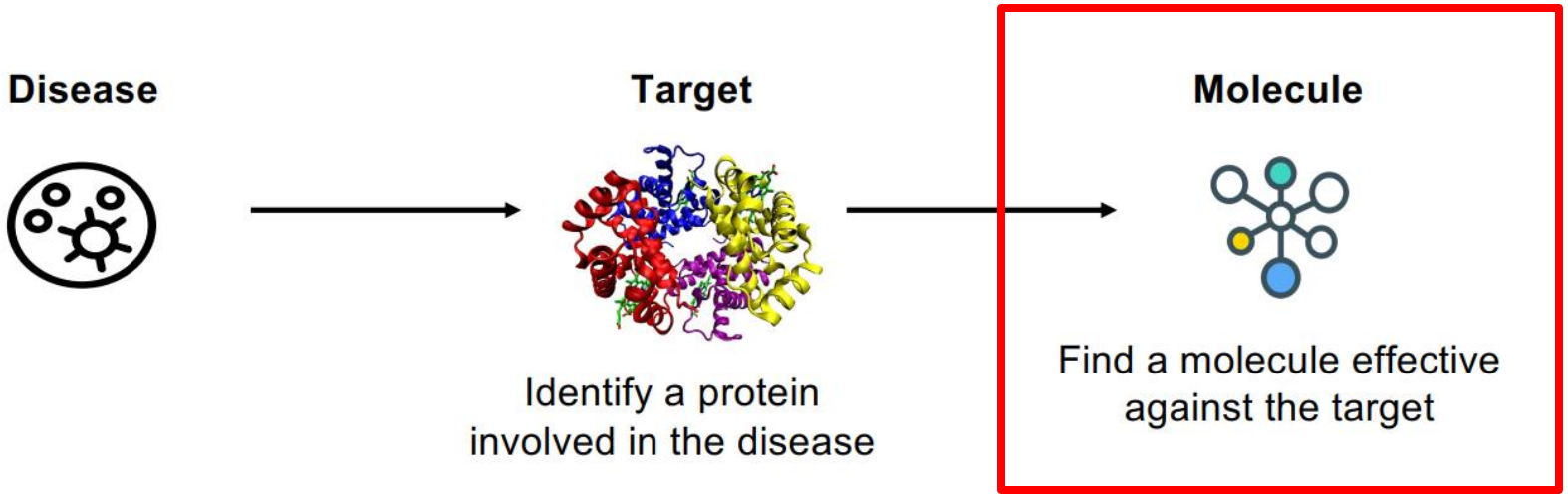
에이조스바이오

박혜진

2025.07.21

- Molecule

Molecule



Molecule

Representation	Name
Caffeine	Common Name
trimethylxanthine, theine, mateine, guaranine, methyltheobromine	Synonyms
$C_8H_{10}N_4O_2$	Empirical Formula
1,3,7-trimethylpurine-2,6-dione	IUPAC Name
58-08-2	CAS Registry Number
T56 BN DN FNVNVJ B F H	WLN
CN1C=NC2=C1C(=O)N(C(=O)N2C)C	SMILES
CN1C(=O)N(C)c2ncn(C)c2C1=O	SMILES (Aromatic)
1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3	InChI

- Molecule

위키백과
우리 모두의 백과사전

위키백과 검색

2022년 하반기 옛한글 문헌 전자화 프로젝트 봉사자 모집이 12월 1일부터 14일까지 진행됩니다.

아스피린

98개 언어

문서 토론 읽기 편집 역사 보기

위키백과, 우리 모두의 백과사전.

아스피린(Aspirin) 또는 **아세틸 살리실산**(acetylsalicylic acid, ASA)은 살리실산염 의약품이다. 진통제, 해열제로 쓰고, 혈중 농도를 낮추어 심혈관질환이나 심장마비 예방약으로 장기간 쓴다. 반감기는 300~650 mg 일 경우 3.1 - 3.2시간, 1g일 경우 6시간, 2g일 경우 9시간이다.

'아스피린'은 바이엘의 상표명이지만, 몇몇 나라에서는 아스피린을 아세틸살리실산이라는 물질명으로 부르기도 한다. 아스피린은 프로트롬빈 생성 억제를 통한 혈소판 제거 반응을 하기 때문에 항응고 작용을 하여 혈전을 예방할 수 있다. 이 때문에 급성 심장마비에 사용한다. 그러나 통증 억제를 위한 복용시 과다 출혈과 같은 부작용을 일으킬 수 있다.

역사 [편집]

버드나무 껍질에 함유된 살리실산이라는 물질에서 비롯된 아스피린은 기원전(BC) 1천500년쯤 고대 이집트에서 작성된 파피루스에서 언급된다. 아스피린하면 바이엘을 연상하지만 최초의 탄생역사는 기원전으로 올라가고 현대의약으로 특허를 최초로 등록한 독일의 바이엘AG사가 베를린의 제국 특허국에 자사의 상표를 등록한 날로부터 약 110년이 되었다. 그동안 아스피린은 세계에서 가장 생명이 길고 가장 놀라운 약임이 입증되었다. BC 400년쯤에는 '의학의 아버지'로 불리는 그리스의 히포크라테스가 사용했다는 기록이 있다. 그 후 2천여 년이 지나 영국에서 에드워드 스톤이라는 성직자가 백버드나무 껍질 즙을 열이 있는 사람 50명에게 먹여 해열작용을 확인했다. 그는 이 사실을 1763년 런던 왕립학회에서 발표했다. 약 60년 후에 이탈리아 화학자 피리아는 버드나무 껍질에서 약효의 주성분인 살리신을 분리했다. 그 뒤 몇 단계 화학 반응을 거쳐 아스피린의 모체인 살리을 얻었다.[4] 살리실산은 의학적인 효과가 있었지만 위벽을 자극하며 설사를 일으키고, 많이 먹을 경우 죽는 경우도 있었다. 1897년 독일 바이엘 사의 연구원 펠릭스 호프만[5]은 살리실산의 히드록시기를 아세틸기와 에스테르화 반응을 시켜 아스피린을 만들었는데 살리실산의 부작용이 크게 줄어들었다. 이는 최초의 합성 의약품이다. 바이엘 사는 1899년 3월 6일 "해열 진통제" 아스피린의 특허를 등록했다. 1914년, 초기의 가루 약을 알약 형태로 교체, 복용을 간편하게 하고 복용량을 표준화함으로써 더욱 대중화되었다.[6]

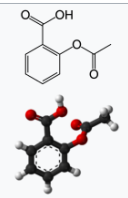
같이 보기 [편집]

- 이부프로펜
- 아세트아미노펜 - 타이레놀의 주성분

각주 [편집]

1. ↑ Haynes, William M., 편집. (2011). 《CRC Handbook of Chemistry and Physics》 92판. Boca Raton, FL: CRC Press. 3.8쪽. ISBN 1439855110.

아세틸 살리실산



체계적 명칭 (IUPAC 명명법)
2-Acetoxybenzoic acid
식별 정보
CAS 등록번호	50-78-2
ATC 코드	A01AD05 B01AC06, N02BA01
PubChem	2244
드러그뱅크	DB00945
ChemSpider	2157
화학적 성질
화학식	$C_9H_8O_4$
분자량	180g/mol
유의어	2-acetoxybenzoic acid acetylsalicylate acetylsalicylic acid *O*-acetylsalicylic acid
물리적 성질
밀도	1.40 g/cm³
녹는점	136 °C (277 °F) [1]
끓는점	140 °C (284 °F) (decomposes)

Aspirin

100 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

For the Persian-language TV series, see Aspirin (TV series).
Not to be confused with Robert Asprin.

Aspirin, also known as **acetylsalicylic acid** (**ASA**), is a nonsteroidal anti-inflammatory drug (NSAID) used to reduce pain, fever, and/or inflammation, and as an antithrombotic.[10] Specific inflammatory conditions which aspirin is used to treat include Kawasaki disease, pericarditis, and rheumatic fever.[10]

Aspirin is also used long-term to help prevent further heart attacks, ischaemic strokes, and blood clots in people at high risk.[10] For pain or fever, effects typically begin within 30 minutes.[10] Aspirin works similarly to other NSAIDs but also suppresses the normal functioning of platelets.[10]

One common adverse effect is an upset stomach.[10] More significant side effects include stomach ulcers, stomach bleeding, and worsening asthma.[10] Bleeding risk is greater among those who are older, drink alcohol, take other NSAIDs, or are on other blood thinners.[10] Aspirin is not recommended in the last part of pregnancy.[10] It is not generally recommended in children with infections because of the risk of Reye syndrome.[10] High doses may result in ringing in the ears.[10]

A precursor to aspirin found in the bark of the willow tree (genus *Salix*) has been used for its health effects for at least 2,400 years.[11][12] In 1853, chemist Charles Frédéric Gerhardt treated the medicine sodium salicylate with acetyl chloride to produce acetylsalicylic acid for the first time.[13] Over the next 50 years, other chemists, mostly of the German company Bayer, established the chemical structure and devised more efficient production methods.[13]:69–75

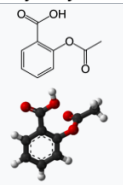
Aspirin is available without medical prescription as a proprietary or generic medication[10] in most jurisdictions. It is one of the most widely used medications globally, with an estimated 40,000 tonnes (44,000 tons) (50 to 120 billion pills) consumed each year,[11][14] and is on the World Health Organization's List of Essential Medicines.[15] In 2021, it was the 34th most commonly prescribed medication in the United States, with more than 17 million prescriptions.[16][17]

Brand vs. generic name [edit]

In 1897, scientists at the Bayer company began studying acetylsalicylic acid as a less-irritating replacement medication for common salicylate medicines.[13]:69–75[18] By 1899, Bayer had named it "Aspirin" and was selling it around the world.[19]

Aspirin's popularity grew over the first half of the 20th century, leading to competition between many brands and formulations.[20] The word *Aspirin* was Bayer's brand name; however, their rights to the trademark were lost or sold in many countries.[20] The name is ultimately a blend of the prefix **a**(cetyl) + **spir**. *Spiraea*, the

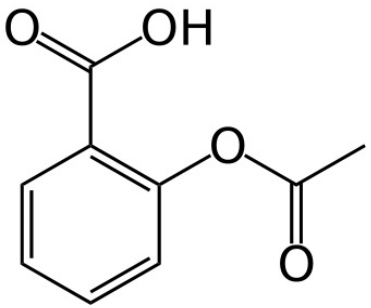
Acetylsalicylic acid



Clinical data	
Pronunciation	/əˌsiːtəl ˌsælɪˈsɪlɪk/
Trade names	Bayer Aspirin, others
Other names	2-acetoxybenzoic acid *o*-acetylsalicylic acid acetylsalicylic acid acetyl salicylate monoacetic acid ester of salicylic acid[1]
AHFS/Drugs.com	Monograph
MedlinePlus	a682878
License data	US DailyMed: Acetylsalicylic acid
Pregnancy category	AU: C[2]
Routes of administration	Oral, rectal
Drug class	Nonsteroidal anti-inflammatory drug (NSAID)
ATC code	A01AD05 (WHO) B01AC06 (WHO), N02BA01 (WHO)
Legal status	
Legal status	AU: OTC / Schedule 2, 4, 5, 6[3][4] CA: OTC[5] UK: General sales list (GSL, OTC) US: OTC / Rx-only
Pharmacokinetic data	
Bioavailability	80–100%[6]
Protein binding	80–90%[7]

• 분자표현 방법 정의

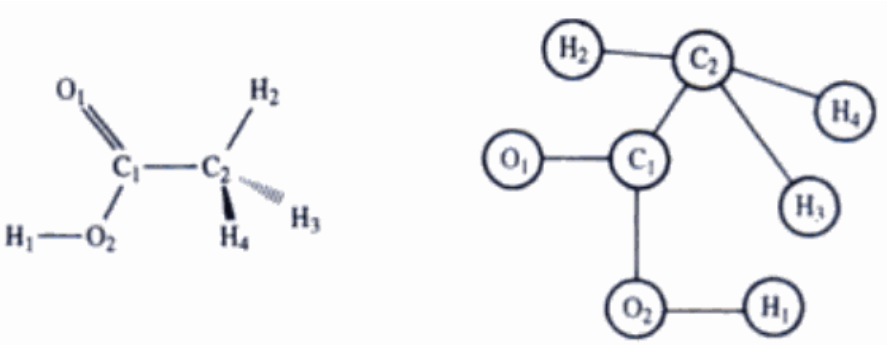
- 컴퓨터(DB)저장 방법
  - Common names: aspirin
  - IUPAC name: 2-acetoxybenzoic acid
  - Formula: $C_9H_8O_4$
  - As an image (PNG, GIF, etc.)
  - CAS number: 50-78-2
  - File format: ChemDraw file, MOL file, etc.
  - SMILES string: O=C(Oc1ccccc1C(=O)O)C
  - Binary Fingerprint: 10000100000001100000100100000001



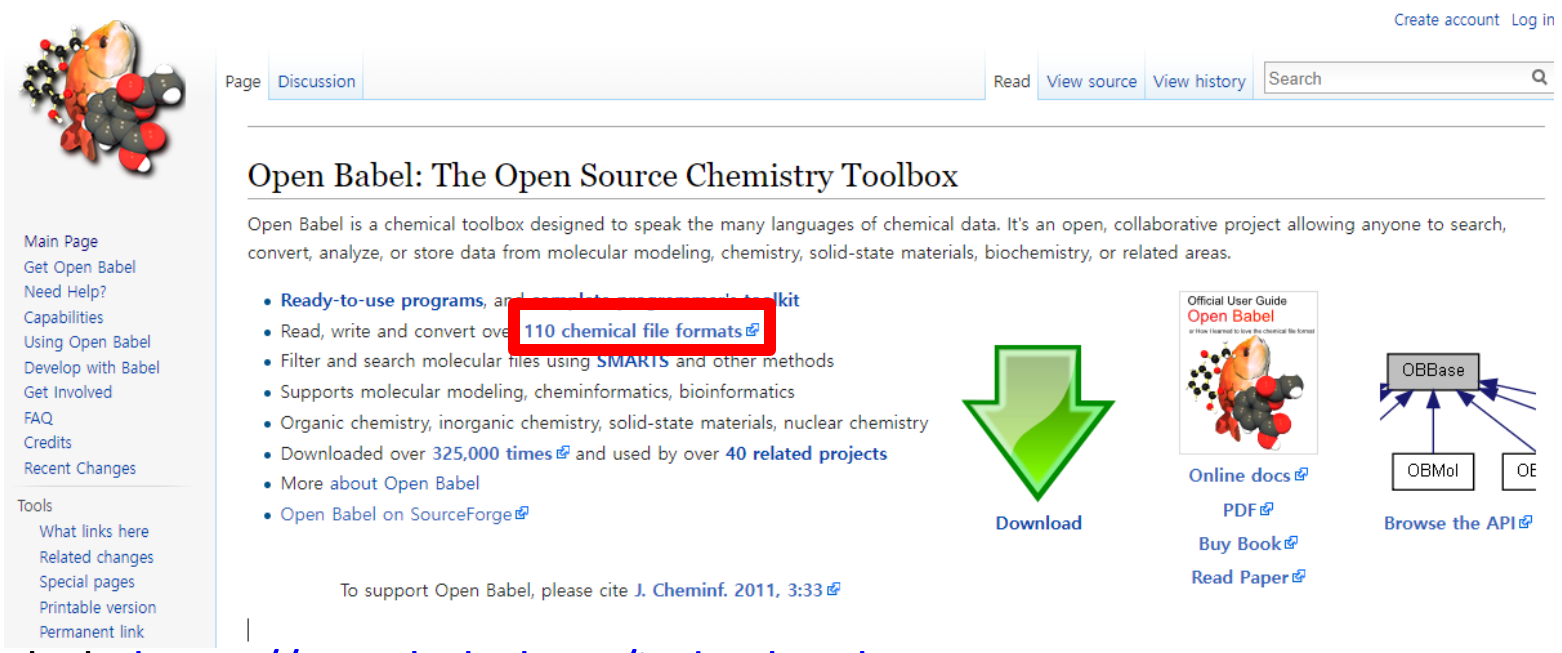
- 컴퓨터(DB) 저장시 고려사항
  - Benzene ring을 가진 구조를 찾아서 분석을 하고 싶다.
  - 각 구조들을 구분하여 설명할 수 있는 구분자를 가지고 싶다.

• 분자표현 방법 정의

- Molecule 구조를 Graph 형태로 표현 가능
  - Graph = collection of nodes and edges
  - Edge : Atom
  - Node : Bond
- Molecule Graph 표현형태
  - 1D : Line notation (SMILES)
  - 2/3D : Connection table (Atom들 간의 연결된 Bond 정보 표현) – mol, mol2, sdf, pdb, ...



- 분자 저장 방식의 다양성



- Open Babel : https://openbabel.org/index.html
- 110개의 Molecule 저장 format (여기에는 없는 저장 format 도 존재함)
  - MOL file for small-molecule structures (SDF)
  - PDB files for protein structures from crystallography
  - MOL2 files for protein structures from modelling software (e.g. after manipulation of the PDB file)

Molecule Graph 표현형태

-1D : Line notation

Chemical Line Notations

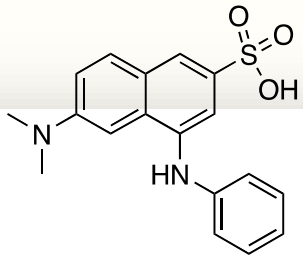
- Representations of molecules that fit on a single line. E.g. standard structural formulas. These work well for linear compounds, but less well for rings…

CH_3CH_2COOH

Line notations are:

- Compact
- Generally human readable/understandable

many line notations



6-Dimethylamino-4-phenylamino-naphthalene-2-sulfonic acid

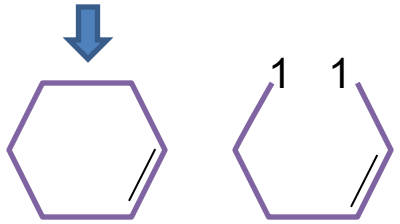
- Wiswesser line notation
  - 1L66J BMR& DSWQ IN1&1
  - An early line notation *(1949),* molecules as fragments
- Rosdal
  - 1=-5-=10=5,10-1,1-11N-12-=17=12,3-18S-19O,18=20O,18=21O,8-22N-23,22-24
  - A linear representation of a connection table developed by Beilstein
- SMILES
  - CN(C)C1=CC=C2C (C(NC3=CC=CC=C3)=CC(S(=O)(O)=O)=C2)=C1
  - Developed by Dave Weninger and Daylight Chemical Systems
- InChi
  - InChI=1S/C18H18N2O3S/c1-20(2)15-9-8-13-10-16(24(21,22)23)12-18(17(13)11-15)19-14-6-4-3-5-7-14/h3-12,19H,1-2H3,(H,21,22,23)
  - A compact chemical representation developed by IUPAC
- Sybyl Line Notation (SLN, Tripos)

A chemical file format: SMILES format

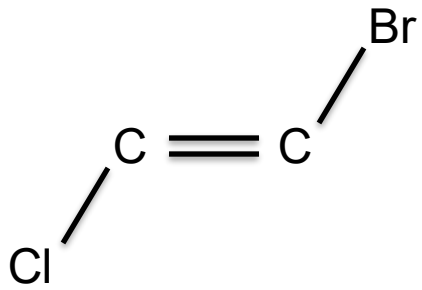
- Simplified Molecular Input Line Entry System
  - Weininger, J Chem Inf Comput Sci, 1988, 28, 31
  - 자세한 정의 참조 사이트 : http://opensmiles.org
  - Molecule의 연결 테이블과 Stereochemistry 를 표현하는 한줄 형태 포멧
  - 쿼리로 사용하거나, DB 에 저장 등 간단하게 표현이 가능하여 편하게 사용가능
- Examples:
  - CC : CH3CH3 (ethane)
  - CC(=O)O : CH3COOH (acetic acid)
- Basic guidelines:
  - Hydrogen들은 표현 안함 (특수한 경우 제외)
  - ( ) 는 또 다른 결합을 나타냄
  - 각 Atom 들은 왼쪽의 Atom에 연결됨을 표시
  - 단일 결합은 표시 안함, 이중결합은 =, 삼중결합은 # 으로 표시
- C(C)(C)(C)C ➔

A chemical file format: SMILES format

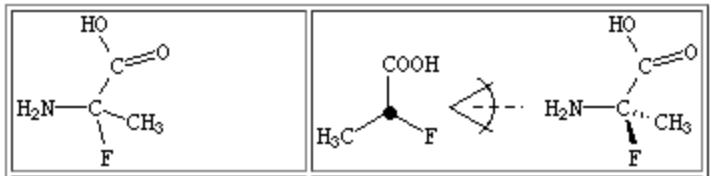
- Ring 표현



C1CCC=CC1



- double bond stereochemistry 표현
  - Cl/C=C/Br (trans), Cl/C=C\Br (cis)
- tetrahedral stereochemistry 표현
  - N[C@](C)(F)C(=O)O
    - N 에서 바라보면 CH3, F, COOH 가 반시계방향으로 정렬

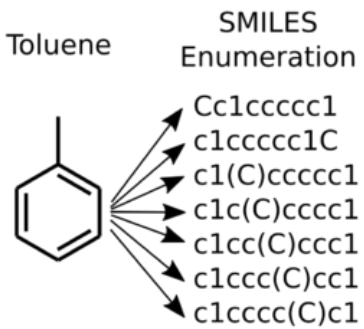


NC(C)(F)C(=O)O	N[C@](C)(F)C(=O)O
NC(F)(C)C(=O)O	N[C@@](F)(C)C(=O)O
(unspecified chirality)	(specified chirality)

- Aromaticity 표현은 소문자로 나타냄
  - C1CCCCC1 (cyclohexane)
  - c1ccccc1 (benzene)

A chemical file format: Canonical SMILES

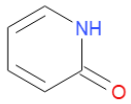
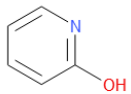
- 하나의 분자가 여러 가지 SMILES 형태로 표현이 가능
  - Not a unique identifier (one-to-many)



- "canonical SMILES" 형태로 표현
  - One molecule ⇔ One Canonical SMILES
  - 특정 구조에 대해 유니크한 SMILES 가짐
  - 다양한 Canonicalisation algorithm 이 존재하여 관련하여 주의가 필요함
- Database 에서 구조 중복 체크에 이용가능
  - 각 구조에 대해 Canonical SMILES를 이용하여 중복 구조를 체크
- Deep Learning 계산시 SMILES를 이용할 때 주의가 필요
  - SMILES가 아닌 Canonical SMILES 를 이용
  - 하나의 Algorithm을 이용하여 Canonical SMILES 를 생성

A chemical file format: InChI

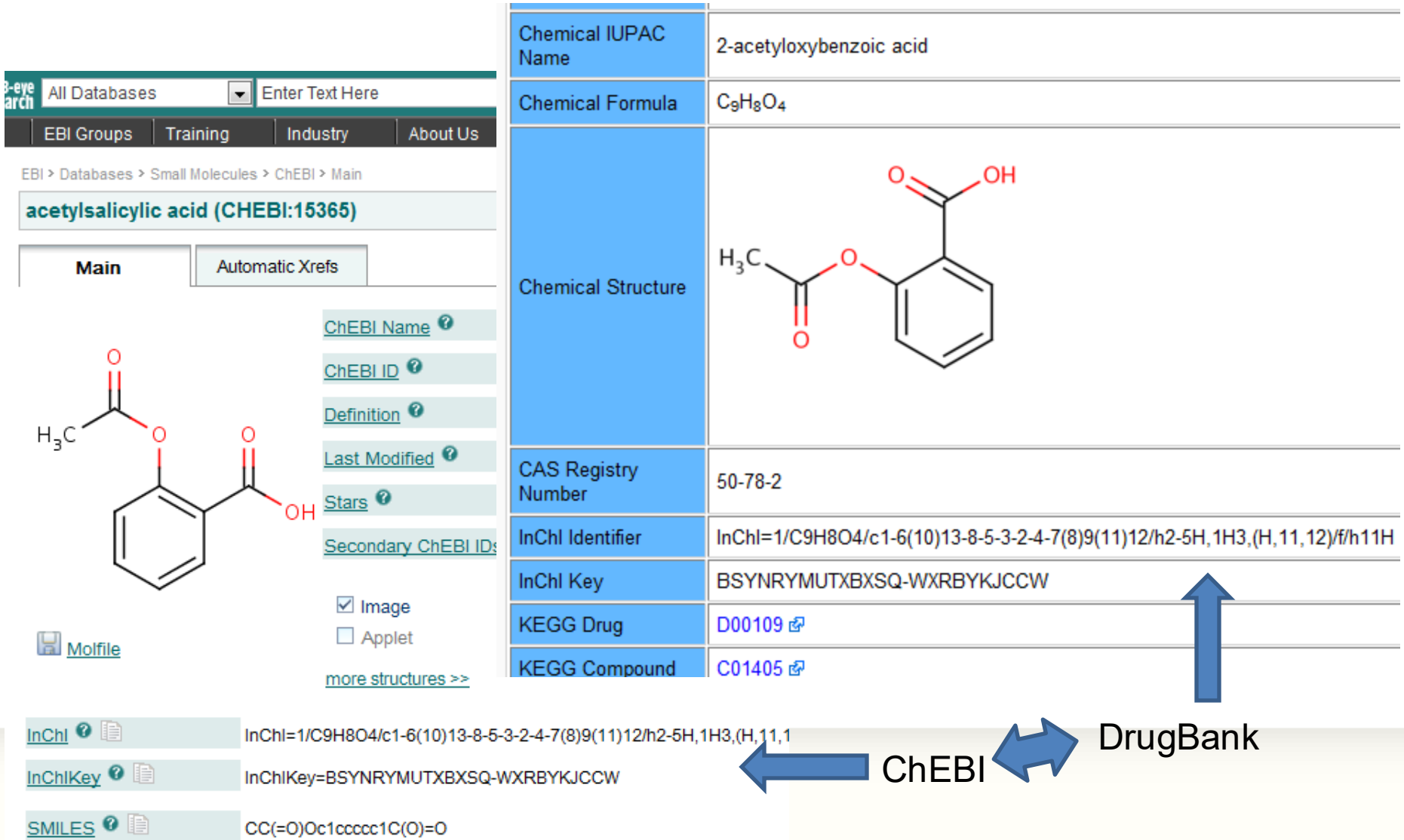
- International Chemical Identifier
  - NIST 와 IUPAC에서 개발
  - 목적 : 분자를 고유하게 식별하기 위한 인덱스
- Aspirin: InChI=1/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)/f/h11H
- 특징
  - 구조를 기반으로 생성 (unlike CAS number)
  - 구조와 1:1 매칭
  - 구조 이성질체(tautomer)가 같은 InChI 를 가짐 (SMILES는 아님)

Data Image	inpsmiles	InChI	InChI_AuxInfo
	O=c1[nH]cccc1	InChI=1S/C5H5NO/c7-5-3-1-2-4-6-5/h1-4H,(H,6,7)	AuxInfo=1/1/N:6,5,7,4,2,3,1/rA:7OCNCCCC/rB:d1;s2;s3;d4;s5;s2d6;/rC;;;;;;
	Oc1ncccc1	InChI=1S/C5H5NO/c7-5-3-1-2-4-6-5/h1-4H,(H,6,7)	AuxInfo=1/1/N:6,5,7,4,2,3,1/rA:7OCNCCCC/rB:s1;d2;s3;d4;s5;s2d6;/rC;;;;;;

- Notes
  - 컴퓨터가 생성하는 형태로 사람이 직접 읽고 쓰는 형태가 어려움
  - InChI Trust 가 제공하는 open source 코드를 이용하여 생성

* Source: http://inchi.info

A unique identifier makes it easy to link databases



Molecule Graph 표현형태

-2D/3D : Connection Table

molecular structure 파일 내용

- Whole molecule:
  - molecule name
  - journal article (for crystal structures)
  - creator or author(s)
- Atom:
  - atomic element (H, He, C, N, O, F, etc.)
  - atom name (E.g. in an amino acid N, CA, CB, CO O, etc.)
  - Cartesian coordinates (X, Y, Z) or Z-matrix atom number
  - Atom charge (formal and/or partial)
  - residue name (E.g. for a protein: Ala, Pro, etc.)
  - temperature factor and occupancy for crystal structures
- Bond
  - connection table : Atom 과 bond 정보
  - bond-orders (single, double, aromatic, etc.)

A chemical file format: PDB file

- The Protein Data Bank (PDB) file format
  - Brookhaven National Laboratory
  - to store protein crystal structure information
- Used by many molecular modelling programs
- The PDB format has limitations:
  - Columns are of fixed size
  - Does not contain information about bond orders (these are recorded in a separate database)
- The Databank has developed new formats to replace the PDB format. e.g. the mmCIF format (Macromolecular Crystallographic Information File)

- Ref: http://www.rcsb.org/ pdb/info.html.

A chemical file format: SD file format

- Developed by Molecular Design Limited (MDL).
- Can store 2D or 3D structures
- Can contain *query structures* which can contain variable atom and bond types. E.g an atom may be *either* nitrogen or carbon, or a bond could be *either* double or aromatic
- Can store additional information such as biological activity data associated with the molecule

```
-ISIS-  02991010 2D

 13 13 0 0 0 0 0 0 0 0999 V2000
  -0.0586   -1.1517    0.0000 C  0 0 1 0 0 0 0 0 0 0 0 0
  -1.7103   -0.5379    0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
    .
    .
    .
   0.6069    1.4103    0.0000 C  0 0 2 0 0 0 0 0 0 0 0 0
   2.8138    1.3828    0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
   3.9207   -0.5379    0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  -3.9207    0.7414    0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
   1.2724    2.1586    0.0000 O  0 0 0 0 0 0 0 0 0 0 0 0
   3.9207    0.7414    0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  1  2  1  0  0  0  0
  1  3  1  0  0  0  0
  1  4  1  1  0  0  0
     .
     .
     .
  1  5  1  6  0  0  0
  6  10 1  0  0  0  0
  10 13 1  0  0  0  0
  12 16 1  0  0  0  0
  15 18 2  0  0  0  0
M  END
>  <Isis_internal_number> (2)
2

>  <chemical_name> (2)
Minaprine dihydrochloride

>  <smiles_code> (2)
c1(c2ccccc2)(cc(c(NCCN3CCOCC3)nn1)C).Cl.Cl

>  <Plate position> (2)
66

$$$$
```

Atoms

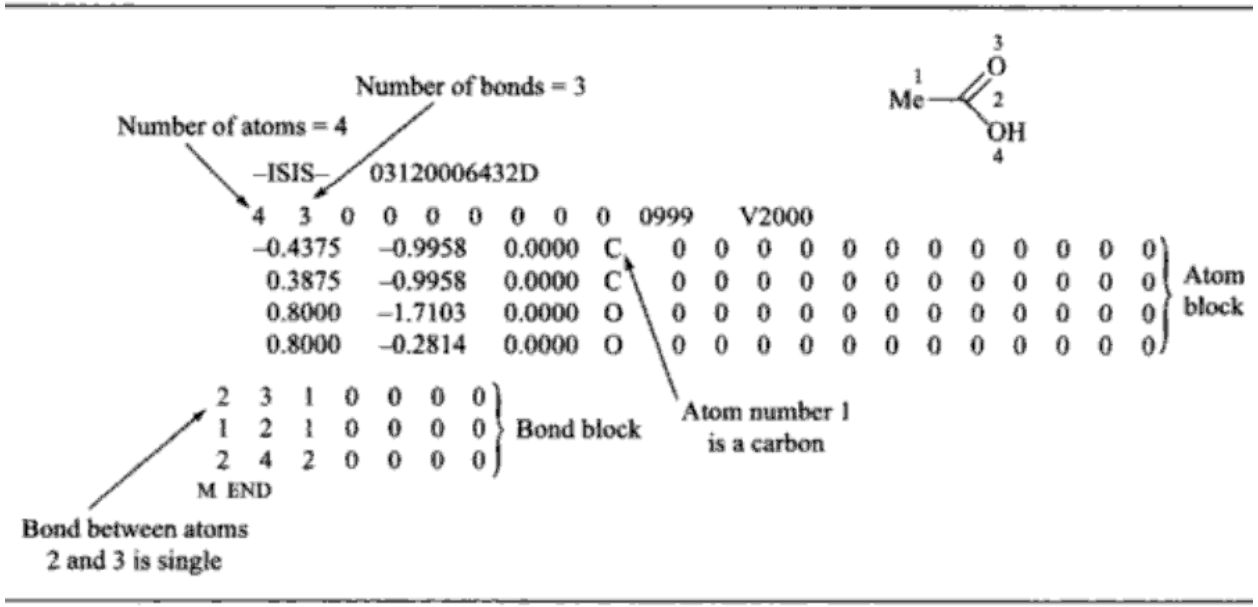
Bonds

Data

Record end

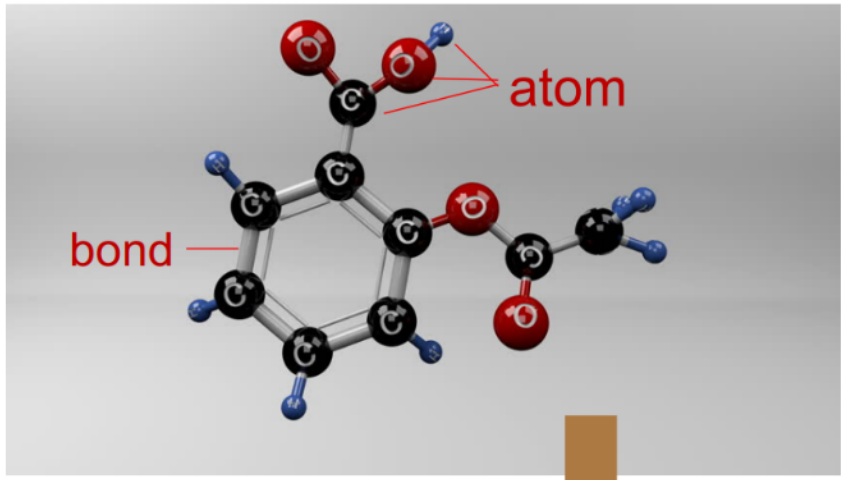
A chemical file format: MOL file

- 2D 및 3D 형태의 구조 표현



12.3: *MDL mol file for acetic acid, in the hydrogen-suppressed form.*

Molecule 표현 방법 (Descriptor)



Aspirin molecule.

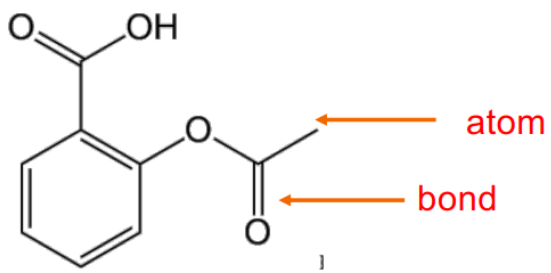
Formally, acetylsalicylic acid

1-D descriptors

Weight, solubility, charge, number of rotatable bonds, atom types, topological polar surface area

Molecule 표현 방법 (Descriptor)

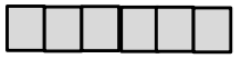
2D Graph Representation



Aspirin molecule.

fingerprinting

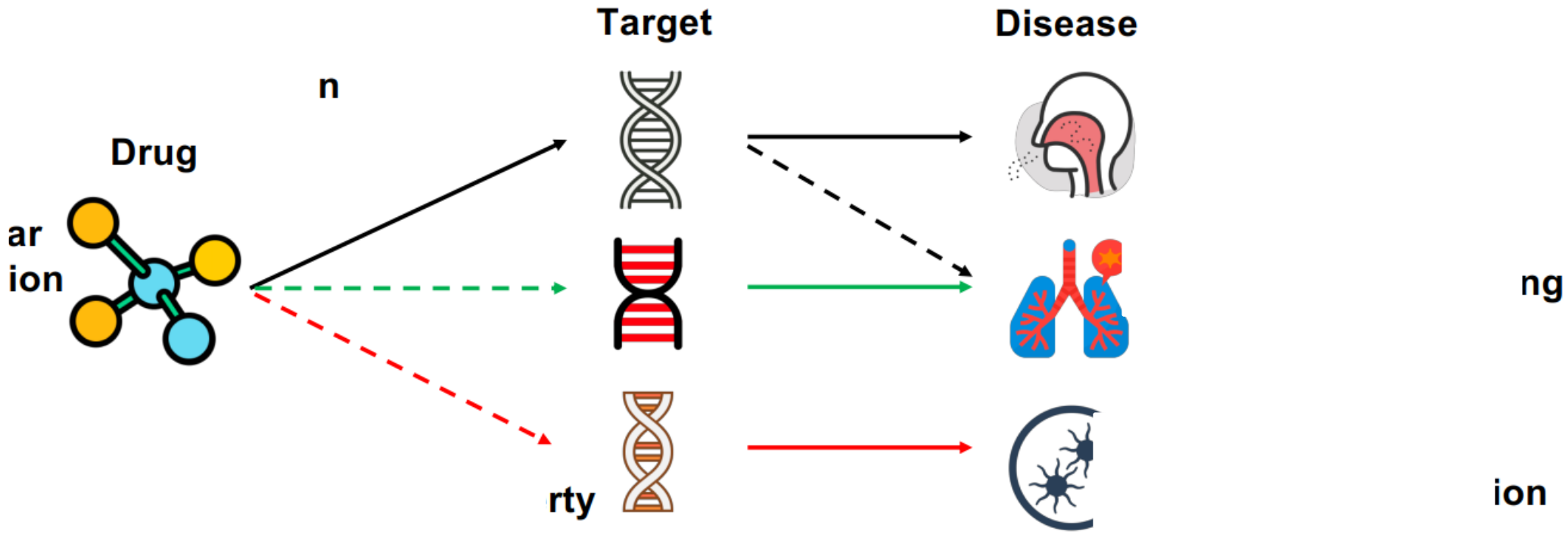
Circular fingerprinting



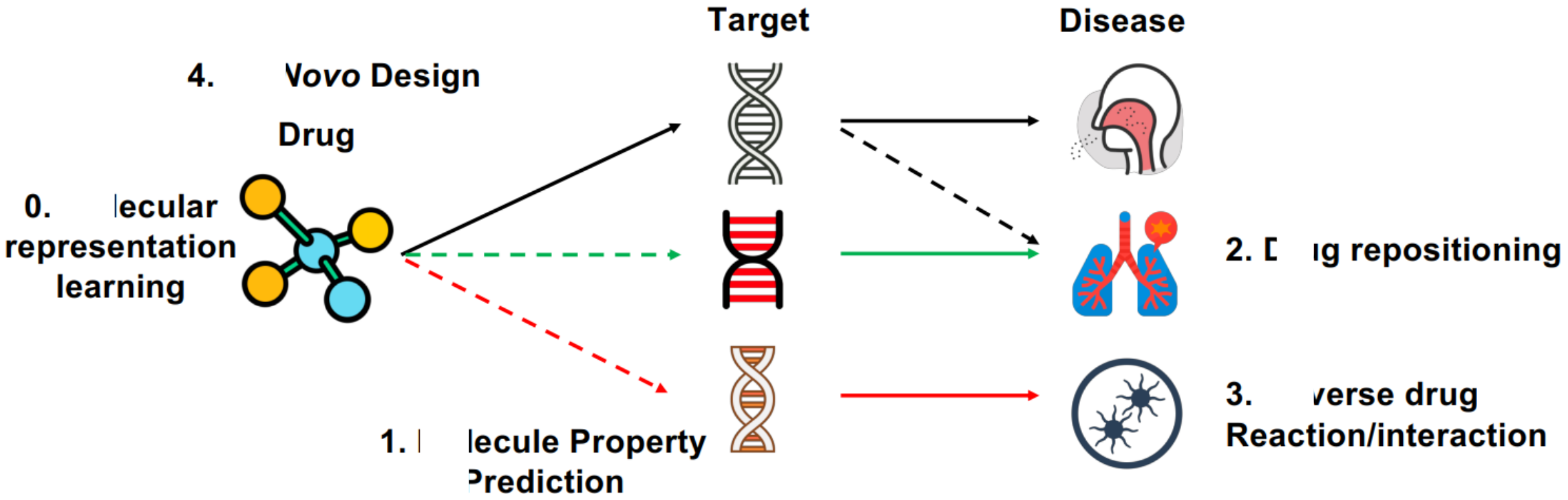
2-D Descriptor

taking into account the graph of covalent and aromatic bonds, but not spatial coordinates.

Molecule

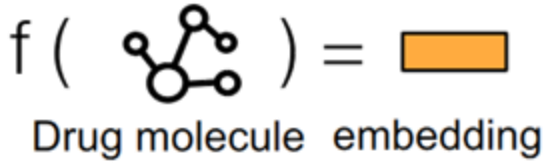


Molecule

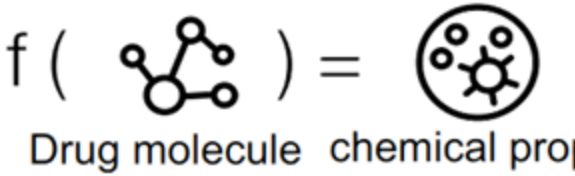

Target
Disease
4. Vovo Design
Drug
0. lecular
representation
learning
1. l lecule Property
Prediction
2. D ıg repositioning
3. verse drug
Reaction/interaction

Molecule

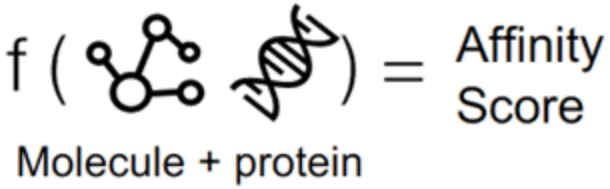
0. Molecular Representation Learning



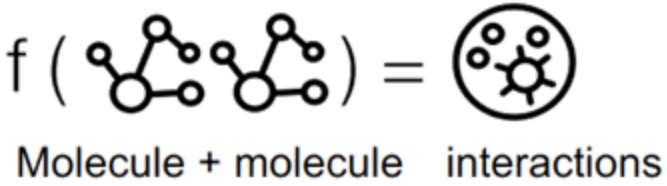
1. Molecule Property Prediction



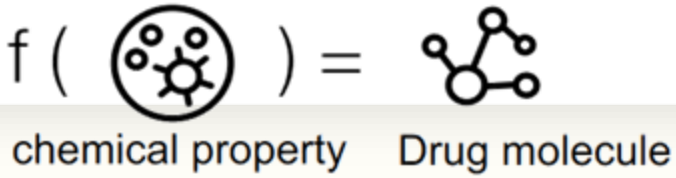
2. Drug repositioning



3. Adverse drug Reaction/interaction



4. De *Novo* Design



분자 구조 표현 방법

1D 표현 방법

- SMILES (Simplified Molecular Input Line Entry System)
  - 참조사이트
    - http://opensmiles.org/opensmiles.html
    - https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html
  - 실습사이트
    - http://cdb.ics.uci.edu/cgibin/Smi2DepictWeb.py : Sim2Depict
- Canonical SMILES
  - 실습 : openbabel
- InChI
  - 참조사이트 : http://inchi.info
  - 최신버전 다운로드 경로 : https://www.inchi-trust.org/download-latest-inchi-standard-software/
  - 명령어 예제
    - inchi-1.exe <<inputfile>> output.txt log.log NUL

SMILES

- Atom

Depiction	SMILES	Name	Note
	[Li]	Lithium	Square brackets ([ ]) are used to delimit individual atoms.
	O	Water	Elements in the "organic subset" may be written without brackets if the number of attached hydrogens conforms to the lowest normal valence consistent with explicit bonds: **B(3), C(4), N(3,5), O(2), P(3,5), S(2,4,6), F(1), Cl(1), Br(1), I(1)**
	*F	Unknown atom bonded to Fluorine	* is wildcard (any atom). The wildcard atom may also be written without brackets.
	C	Methane	
	[H]	Hydrogen Atom	Hydrogen is NOT part of the "organic subset" and therefore needs brackets.
	[C]	Elemental Carbon (Graphite)	If atoms are used within brackets, all Hydrogens must be specified, otherwise it is assumed that there are none.

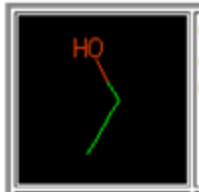
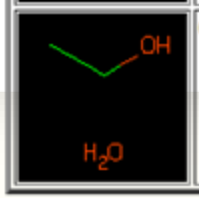
SMILES

- Properties of Atoms

	[Co+2] or [Co++]	Cobalt(II)	Brackets are required whenever atomic properties (including chirality) are specified. Charge is specified by sign and numerical value or by the quantity of signs.
	[NH4+]	Ammonium Ion	Hydrogen count is an atomic property and may be specified by including an H (and optionally an integer) after the Atomic Symbol. Hydrogens are not normally considered atoms. These Hydrogens are often referred to as "implicit" Hydrogens.
	[H+]	Proton	Hydrogen IS considered an atom when it is charged or has a specified mass (e.g. [2H], Deuterium). Hydrogen atoms are often referred to as "explicit" Hydrogens.
	[13C]	Carbon-13	Atomic mass is specified by including an integer before the Atomic Symbol. Default mass is "unspecified".
	[2H]O[2H]	Heavy Water	

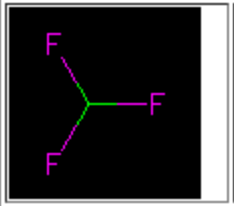
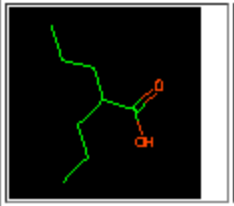
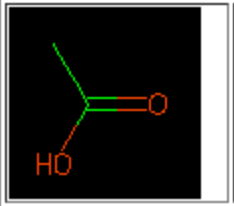
SMILES

- Bonds

	C-C-O or CCO	Ethanol	Single bonds are denoted by a dash, `-'. Single bond symbols and aromatic bond symbols (`:') may be omitted.
	O=C=O	Carbon Dioxide	Double bonds are denoted by an equals sign `='.
	C#N	Hydrogen Cyanide	Triple bonds are denoted by a pound sign `#'.
	[Na+]. [Cl-]	Sodium Chloride	The "dot" is a "non-bond", "disconnect", or "zero-order" bond.
	CCO.O	Ethanol and Water	Dot can be used to delineate mixtures. Each dot-disconnected SMILES is considered a "component" of the overall molecule or mixture.

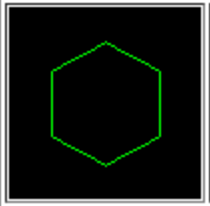
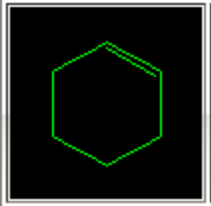
SMILES

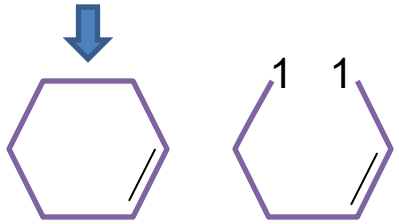
- Branching

	C(F)(F)F or FC(F)F	Fluoroform	Branches may be stacked.
	CCCC(C(=O)O)CCC	4-Heptanoic Acid	Branches may be nested.
	CC(=O)O	Acetic Acid	Bonds may be specified within branches.

SMILES

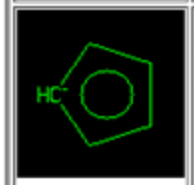
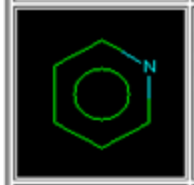
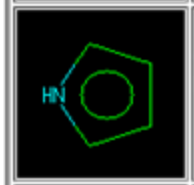
- Rings

	C1CCCCC1	Cyclohexane	Ring closure is specified by breaking bonds and numerically labeling (with the same number) the atoms that were connected to each other.
	C12CCCCC1CCCC2 or C1CC2CCCCC2CC1	Decalin	Atoms can have more than one ring closure.
	C1CCCCC1C1CCCCC1	Bicyclohexane	Closure numbers may be reused.
	C1=CCCCC1 or C=1CCCCC1 or C1CCCCC=1 or C=1CCCCC=1	Cyclohexene	The default bond order for the ring closure is single (or aromatic) but may be specified by including a bond symbol between the atom and the closure number.



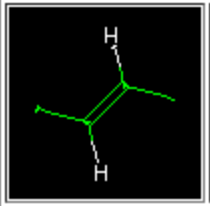
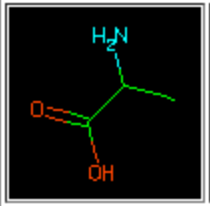
SMILES

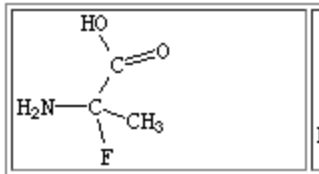
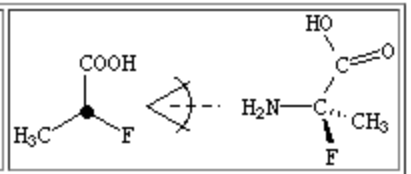
- Aromaticity

	c1ccccc1 or c:1:c:c:c:c:c1 or C1=CC=CC=C1	Benzene	Aromatic atoms may be specified by using lower case letters. Only the following atoms may be interpreted as aromatic: C, N, P, O, S, As, Se, and * (wildcard atom)
	[CH-]1C=CC=C1 or [cH-]1cccc1	Cyclopentadienyl Anion	Aromaticity detection is accomplished by using an extended version of Hueckel's rule. To qualify as aromatic, the number of available "excess" p-electrons in the ring (or ring system) must equal 4N+2. Here, the extra electron allows carbon to contribute 2 p electrons.
	n1ccccc1 or N1=CC=CC=C1	Pyridine	Pyridine nitrogen (5 valence electrons) has an unbound pair of electrons in an sp2 orbital and contributes 1 p electron
	[nH]1cccc1 or N1C=CC=C1	1-H-Pyrrole	Pyrrolyl nitrogen (5 valence electrons) contributes two p electrons.
	o1cccc1 or O1C=CC=C1	Furan	Oxygen (6 valence electrons) has an unbound pair of electrons in an sp2 orbital and contributes 2 p electrons.

SMILES

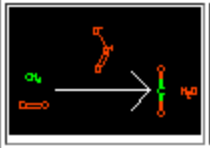
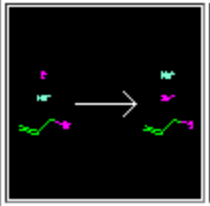
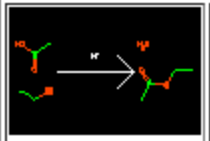
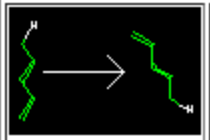
- Stereo Isomerism

	C/C=C/C	Trans-2-butene	E and Z type isomerism can be specified using the `/' and `\' characters. Double bond orientation may be unspecified. (e.g. CC=CC)
	N[C@@H](C)C(=O)O	L-alanine	Tetrahedral chirality can be specified using the "visual mnemonic" `@' character (anticlockwise) or two `@' characters (clockwise). Looking FROM the 1st neighbor listed in the SMILES TO the chiral atom, the other three neighbors appear anticlockwise or clockwise in the order listed.
	N[CH](C)C(=O)O	Alanine	Tetrahedral orientation may be unspecified.

	
NC(C)(F)C(=O)O	N[C@](C)(F)C(=O)O
NC(F)(C)C(=O)O	N[C@@](F)(C)C(=O)O
(unspecified chirality)	(specified chirality)

SMILES

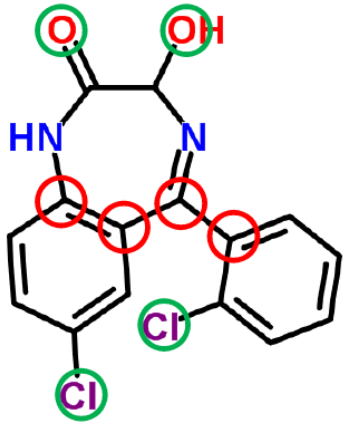
- Reactions

	C.O=O>O=[O+]-[O-]>O=C=O.O	Combustion of methane in the presence of ozone (non-stoichiometric)	Reactions are delineated by using the greater-than ('>') sign. The format is Reactants>Agents>Products. Atoms may be created or destroyed
	[I-].[Na+].C=CCBr>>[Na+].[Br-].C=CCI	Displacement Reaction	Agents are optional.
	CC(=[O:1])[OH:2].CC[OH:3]>[H+]>CC(=[O:1])[O:3]CC.[OH2:2]	Acid catalyzed esterification of acetic acid and ethanol	Atom maps ([<atom>:<map class>]) allow specification of the correspondence between reactant and product atoms. Map class is an atomic property.
	[CH2:1]=[CH:2][CH:3]=[CH:4][CH2:5][H:6]>>[H:6][CH2:1][CH:2]=[CH:3][CH:4]=[CH2:5]	A 1,5-hydride shift.	Atom-mapped hydrogens must be specified explicitly.

Canonical SMILES

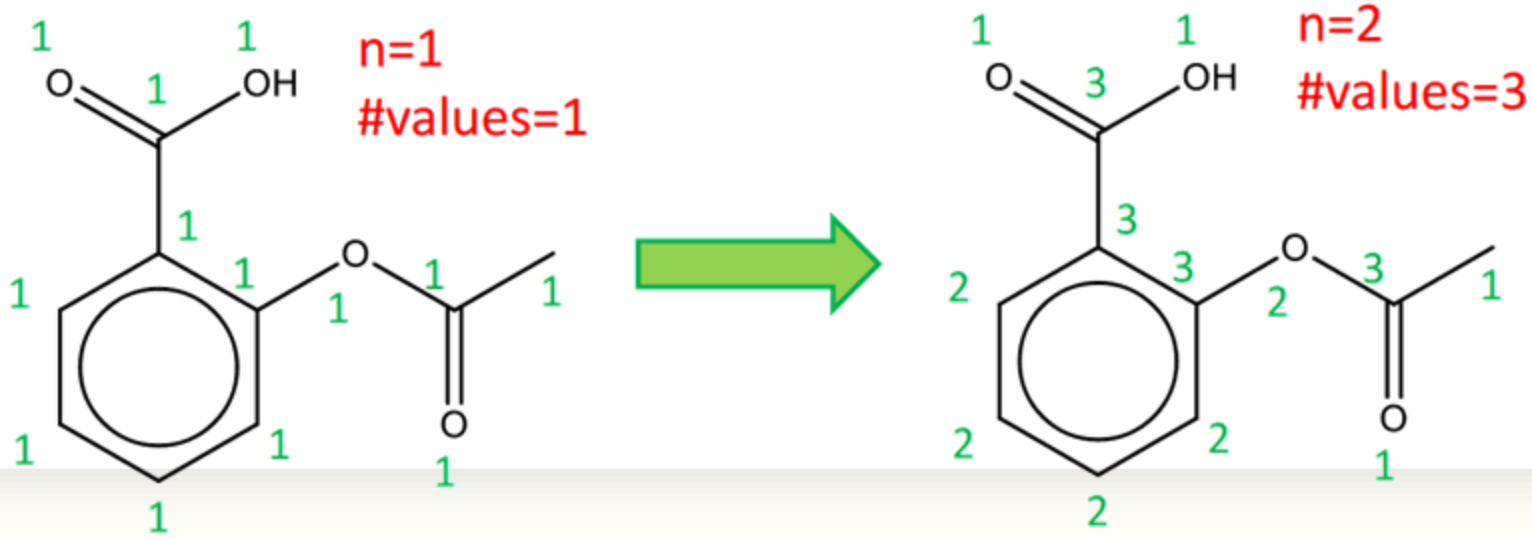
- Basic Idea
  - Topology 기반으로 atom의 우선 순위 부여

- Canonical Labeling을 위한 기본적인 아이디어
  - 아래의 정보의 조합으로 초기 값으로 Atom에 할당
    - Number of neighbors
    - Atom type
    - Number of hydrogens
    - …
  - 이웃에 부여된 값을 기준으로 Atom 값 변경
  - Atom들이 구분될 때까지 위의 과정을 반복
  - (또는 Atom의 값의 변화가 없을 때까지 반복)



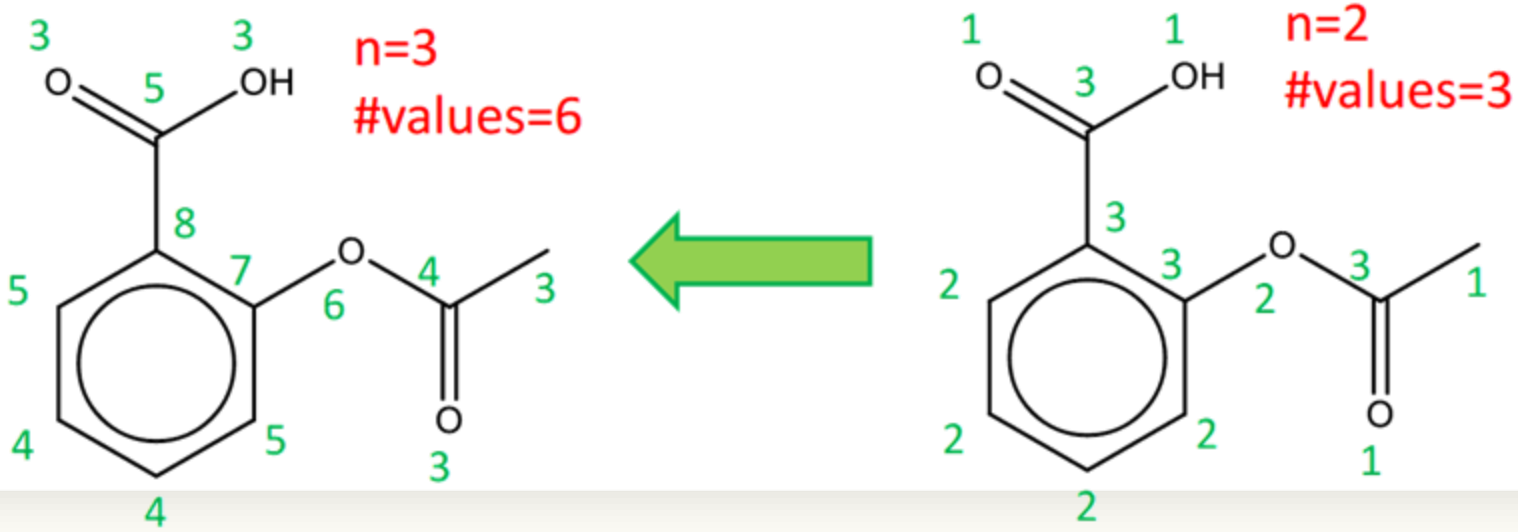
Canonical SMILES - Morgan Algorithm

1. Assign initial invariant of 1
2. New invariant: Sum of neighboring values
3. Determine number of values



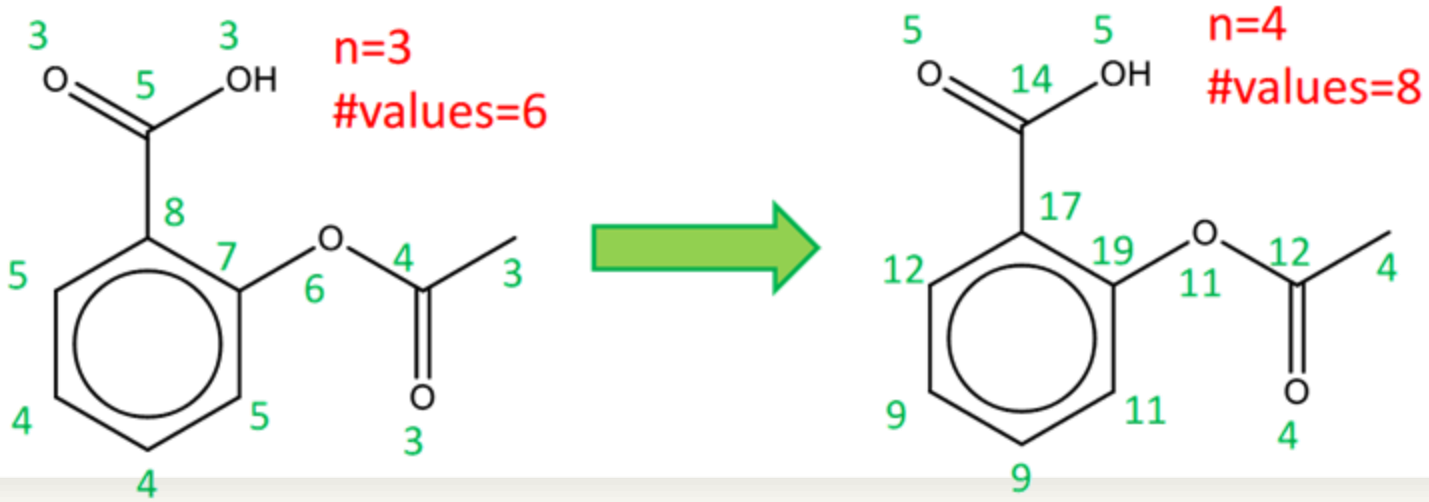
Canonical SMILES - Morgan Algorithm

- Repeat summing of neighboring values



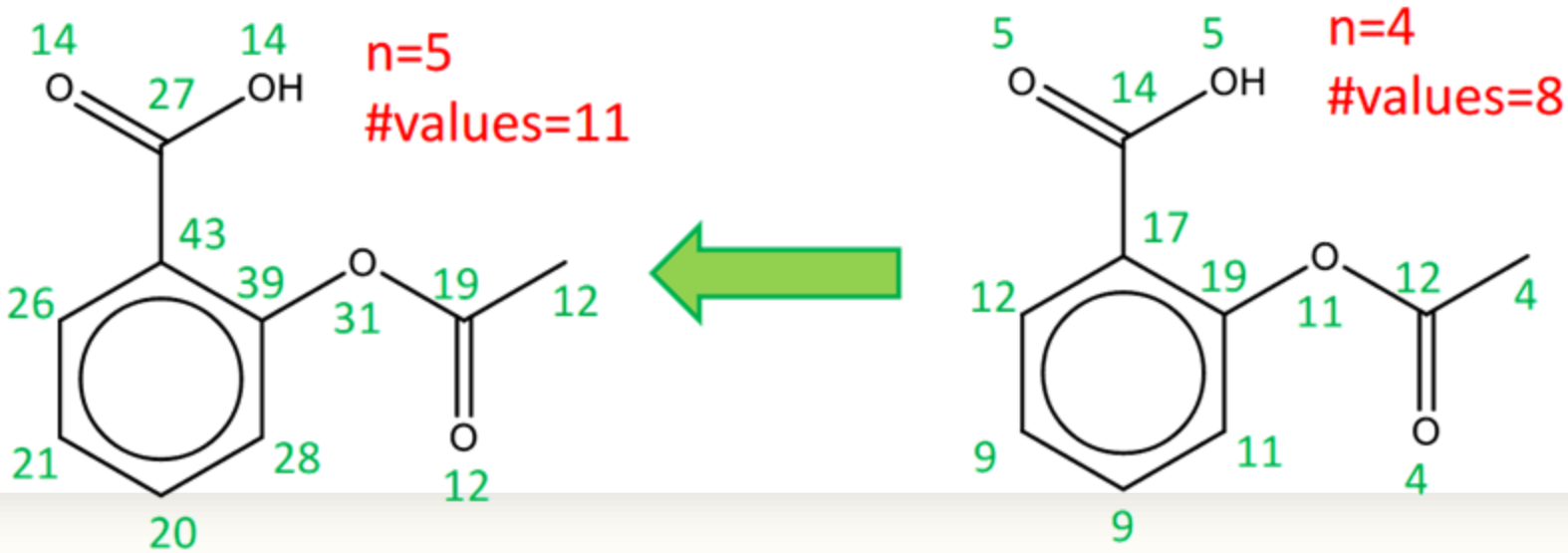
Canonical SMILES - Morgan Algorithm

- Repeat summing of neighboring values



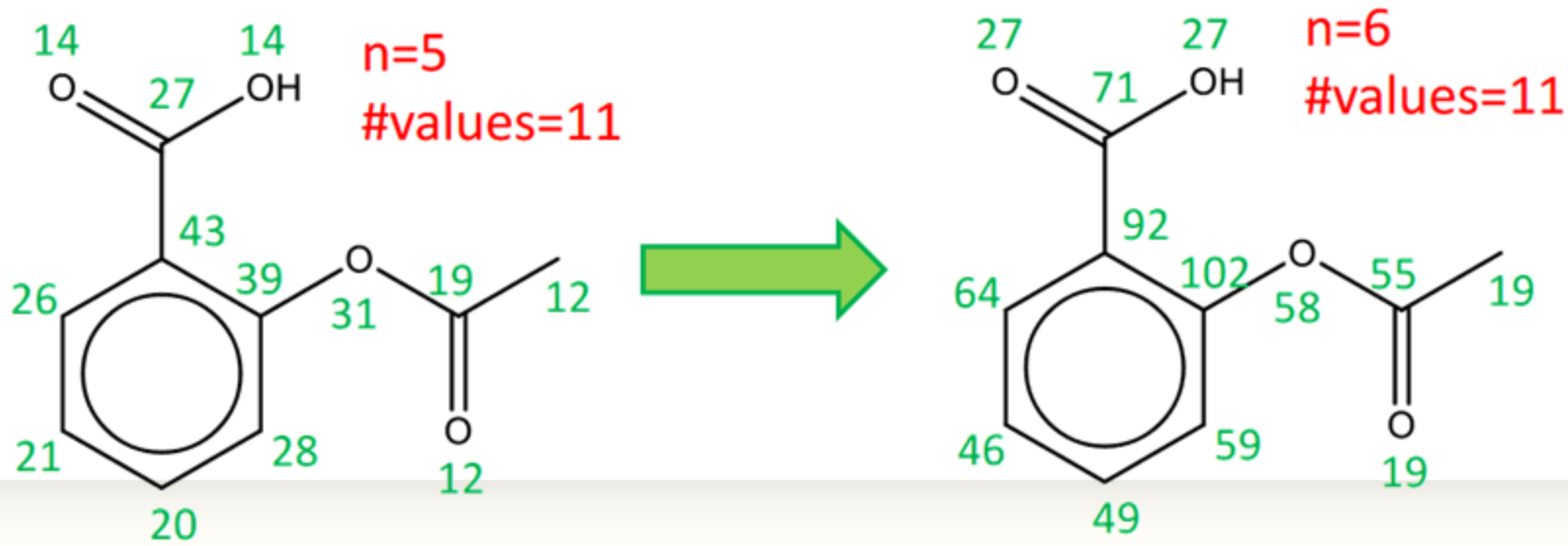
Canonical SMILES - Morgan Algorithm

- Repeat summing of neighboring values



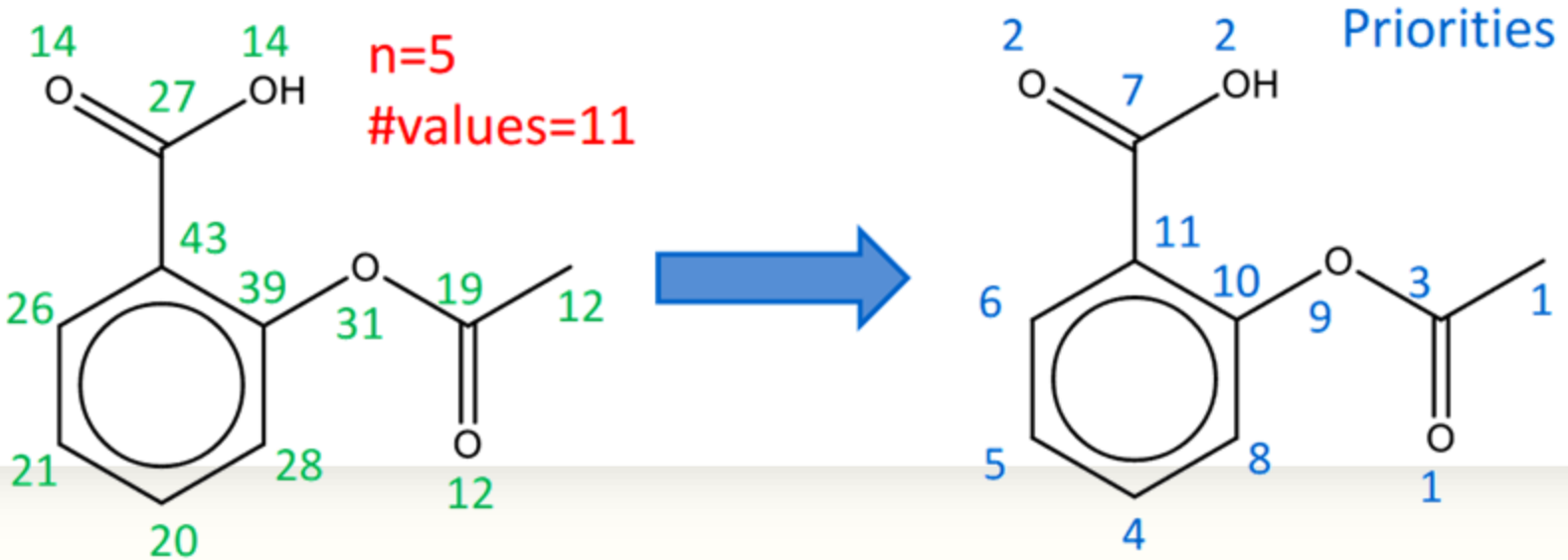
Canonical SMILES - Morgan Algorithm

- Repeat summing of neighboring values
- Until number of values does not increase anymore



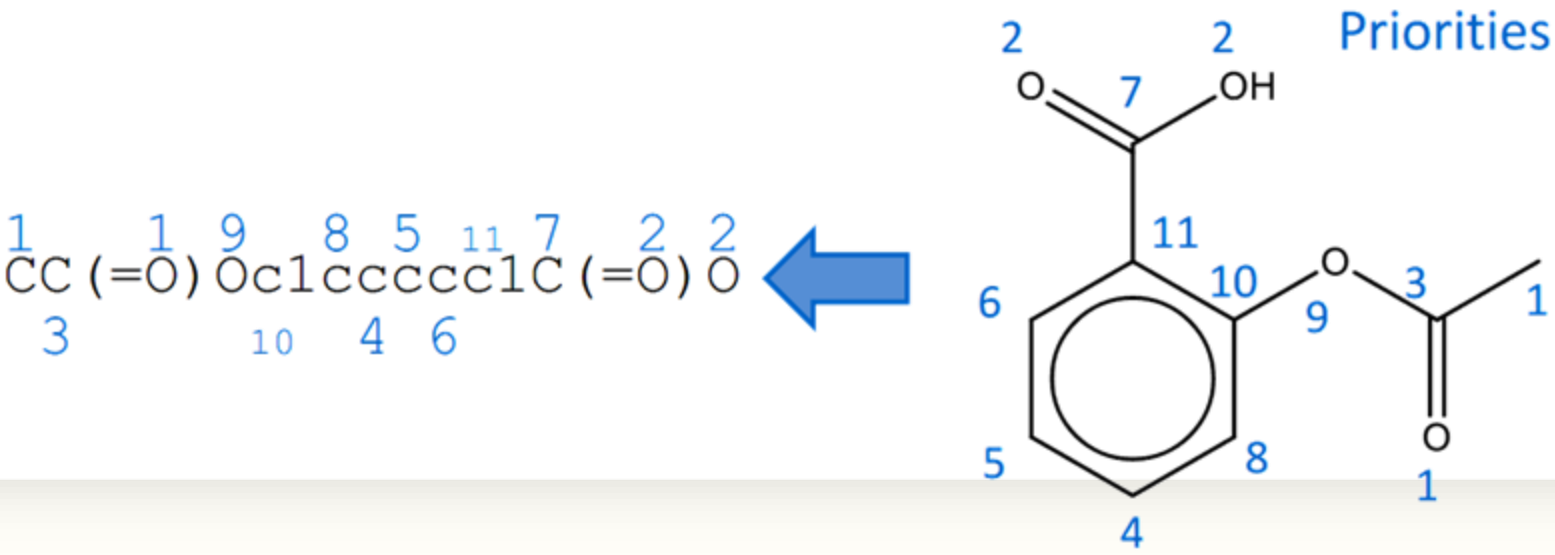
Canonical SMILES - Morgan Algorithm

- Assign priorities according to invariants



Canonical SMILES - Morgan Algorithm

- Disambiguate ties by
  - atom type
  - bond order
- Construct Smiles according to invariants



Canonical SMILES - Morgan Algorithm

- 특징
  - 초기값이 간단함
  - 초기에 서로 다른 원자 유형의 결합 순서를 구분할 수 없음
  - 수치 값이 폭발적으로 증가함
  - 모든 원자를 충분히 잘 구분할 수가 없음

Canonical SMILES - Cangen Algorithm

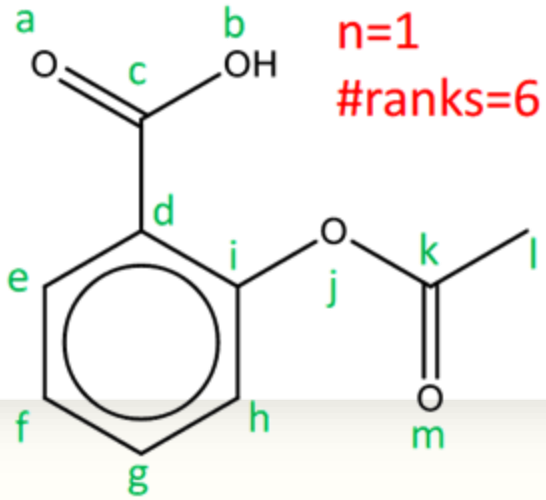
- Morgan Algorithm 의 단점을 보안하기 위해 Weininger 그룹에서 제안
  - 초기값 지정의 향상
  - 안정적인 우선순위 부여
  - 초기에 지정된 값의 결합으로 인한 모호성 방지
  - 대칭 Atom들의 문제 해결

Canonical SMILES - Cangen Algorithm

- Initial Invariants

Initial invariants encode atom type information

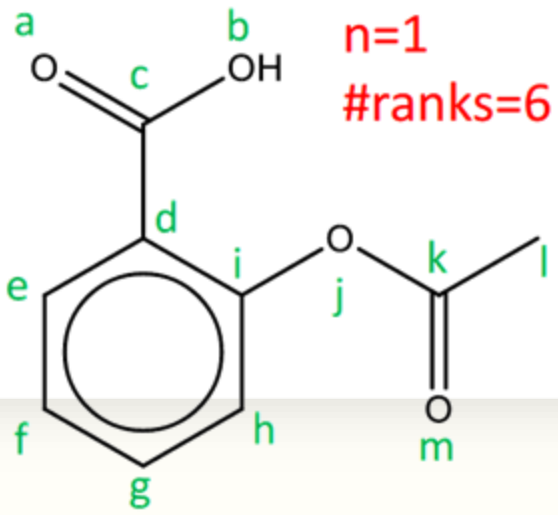
- Number of neighbors
- Sum of bond orders
- Atom type
- Charge
- Number of attached hydrogens



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

Canonical SMILES - Cangen Algioithm

- Initial Invariants
  - Initial invariants are transformed to ranks



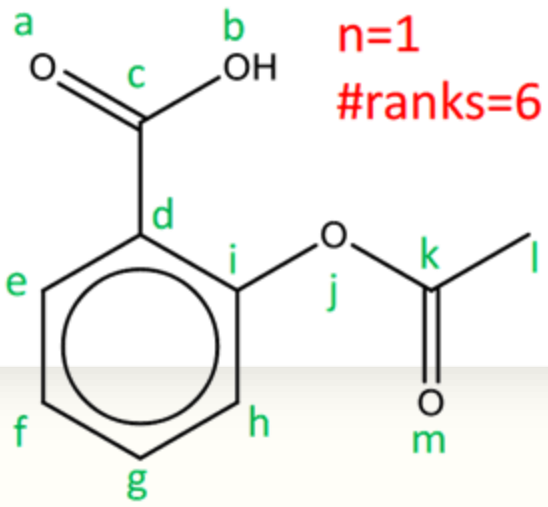
Atom	#bds	Σ bds	At.Nr	Chg.	#H	rank
a	1	2	08	0	0	3
b	1	1	08	0	1	2
c	3	4	06	0	0	6
d	3	4	06	0	0	6
e	2	3	06	0	1	5
f	2	3	06	0	1	5
g	2	3	06	0	1	5
h	2	3	06	0	1	5
i	3	4	06	0	0	6
j	2	2	06	0	0	4
k	3	4	06	0	0	6
l	1	1	06	0	3	1
m	1	2	08	0	0	3

Canonical SMILES - Cangen Algorithm

- Update Rule for Invariants
  - Rank is mapped to corresponding prime:

Rank	1	2	3	4	5	6	...
prime	2	3	5	7	11	13	...

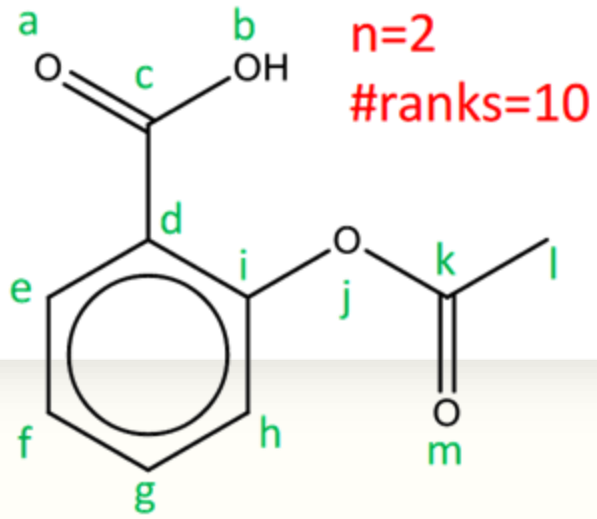
  - New invariant:
    - primes of neighbors are multiplied



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

Canonical SMILES - Cangen Algorithm

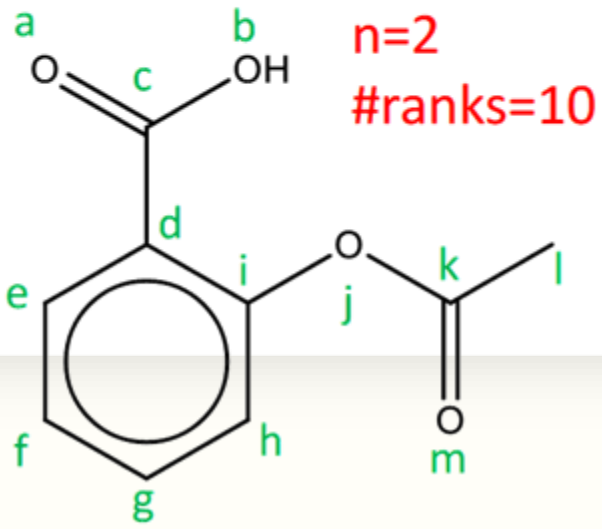
- Update Rule for Invariants
  - New ranks are determined on the basis of:
    - Old ranks
    - New invariants



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

Canonical SMILES - Cangen Algorithm

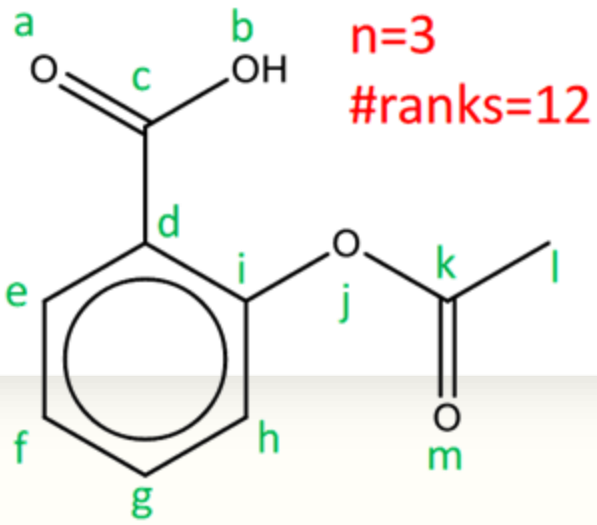
- Update Rule for Invariants
  - Repeat until ranking is stable:
    - Calculate new invariants
    - Re-rank atoms



Atom	rank	prime	Nbors	New Inv.
a	3	5	c	19
b	2	3	c	-
c	8	19	a,b,d	-
d	10	29	c,e,i	-
e	6	13	d,f	319 = 29*11
f	5	11	e,g	143 = 13*11
g	5	11	f,h	143 = 11*13
h	6	13	g,i	243 = 11*23
i	9	23	d,h,j	-
j	4	7	i,k	-
k	7	17	j,l,m	-
l	1	2	k	-
m	3	5	k	17

Canonical SMILES - Cangen Algorithm

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

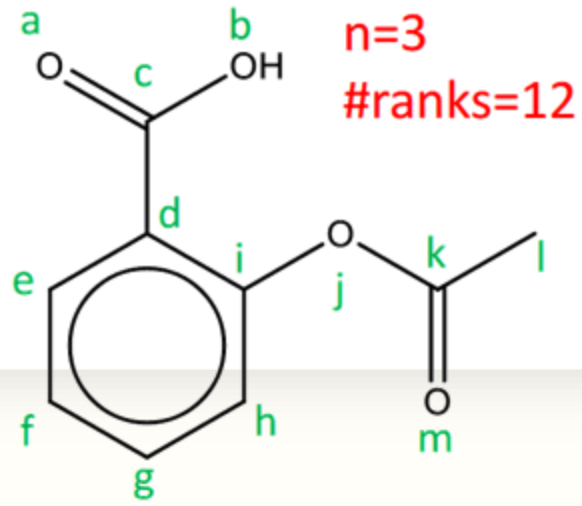


Atom	rank	New Inv.	(rk.,inv.)	New rank
a	3	19	(3,19)	4
b	2	-	(2,-)	2
c	8	-	(8,-)	10
d	10	-	(10,-)	12
e	6	319	(6,319)	8
f	5	143	(5,143)	6
g	5	143	(5,143)	6
h	6	243	(6,243)	7
i	9	-	(9,-)	11
j	4	-	(4,-)	5
k	7		(7,-)	9
l	1	-	(1,-)	1
m	3	17	(3,17)	3

Canonical SMILES - Cangen Algorithm

- Iteration

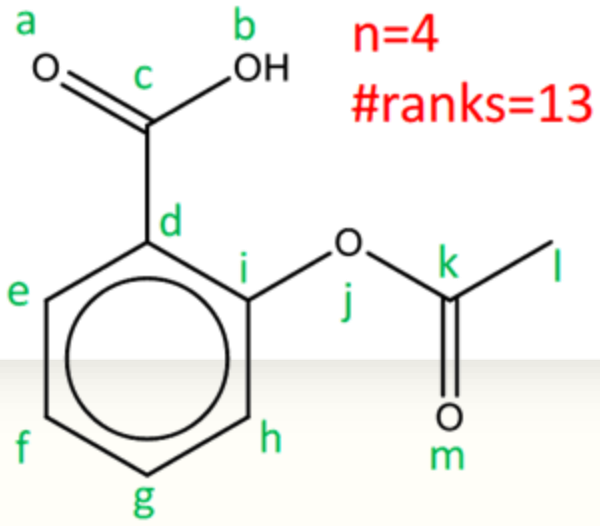
  - Repeat until ranking is stable:
    - Calculate new invariants
    - Re-rank atoms



Atom	rank	prime	Nbors	New Inv.
a	4	7	c	-
b	2	3	c	-
c	10	29	a,b,d	-
d	12	37	c,e,i	-
e	8	19	d,f	-
f	6	13	e,g	249 = 19*13
g	6	13	f,h	221 = 13*17
h	7	17	g,i	-
i	11	31	d,h,j	-
j	5	11	i,k	-
k	9	23	j,l,m	
l	1	2	k	-
m	3	5	k	-

Canonical SMILES - Cangen Algorithm

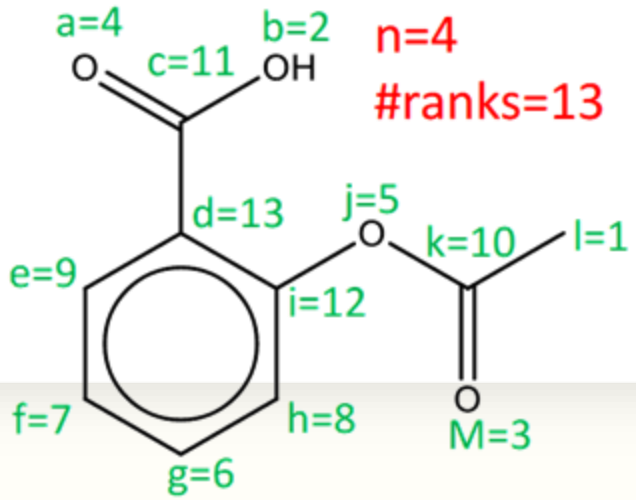
- Final Ranking
  - Repeat until ranking is stable:
    - Calculate new invariants
    - Re-rank atoms
  - Final ranking



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

Canonical SMILES - Cangen Algorithm

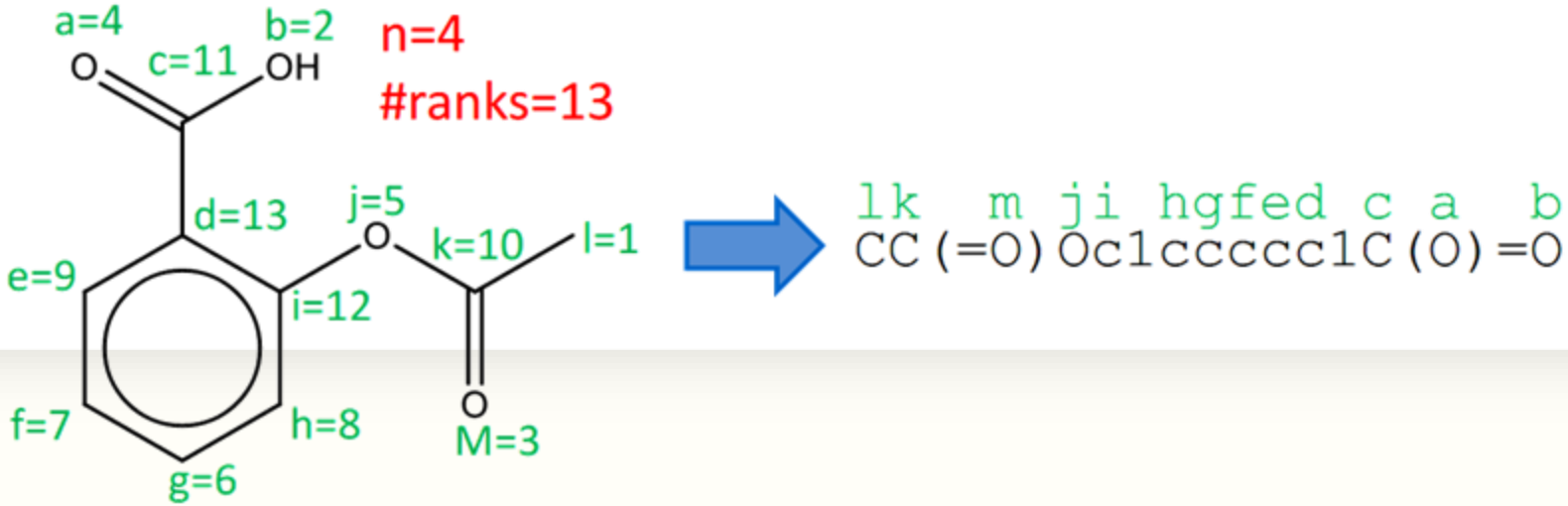
- Final Ranking
  - Repeat until ranking is stable:
    - Calculate new invariants
    - Re-rank atoms
  - Final ranking yields priorities



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

Canonical SMILES - Cangen Algorithm

- Generate Smiles
  - Repeat until ranking is stable:
    - Calculate new invariants
    - Re-rank atoms
  - Final ranking yields priorities
  - Generate Smiles

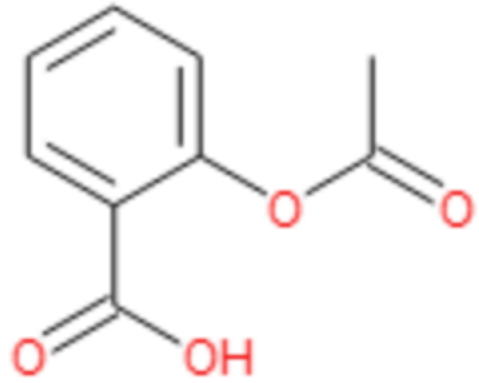


2D, 3D 표현 방법에 대한 여러 방법 소개

- SD file(mol)
  - 참조사이트 : https://www.daylight.com/meetings/mug05/Kappler/ctfile.pdf

- Mol2
  - 참조사이트 : https://chemicbook.com/2021/02/20/mol2-file-format-explained-for-beginners-part-2.html
- PDB
  - Small molecule도 표현은 하지만 주로 단백질 구조 표현에 사용
  - 참조사이트 : https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/pdbintro.html

SD file(mol file)

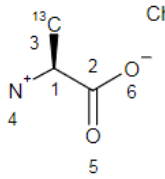



```
OpenBabel12032202322D

21 21  0  0  0  0  0  0  0  0999 V2000
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 H  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 H  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 H  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 H  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 H  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 H  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 H  0  0  0  0  0  0  0  0  0  0  0  0
 1  2  1  0  0  0  0
 1 14  1  0  0  0  0
 1 15  1  0  0  0  0
 1 16  1  0  0  0  0
 2  3  2  0  0  0  0
 2  4  1  0  0  0  0
 4  5  1  0  0  0  0
 5 10  1  0  0  0  0
 5  6  2  0  0  0  0
 6  7  1  0  0  0  0
 6 17  1  0  0  0  0
 7  8  2  0  0  0  0
 7 18  1  0  0  0  0
 8  9  1  0  0  0  0
 8 19  1  0  0  0  0
 9 10  2  0  0  0  0
 9 20  1  0  0  0  0
10 11  1  0  0  0  0
11 12  2  0  0  0  0
11 13  1  0  0  0  0
13 21  1  0  0  0  0
M  END
```

```
OpenBabel12032202372D

13 13  0  0  0  0  0  0  0  0999 V2000
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
   0.0000    0.0000    0.0000 O  0  0  0  0  0  0  0  0  0  0  0  0
 1  2  1  0  0  0  0
 2  3  2  0  0  0  0
 2  4  1  0  0  0  0
 4  5  1  0  0  0  0
 5 10  1  0  0  0  0
 5  6  2  0  0  0  0
 6  7  1  0  0  0  0
 7  8  2  0  0  0  0
 8  9  1  0  0  0  0
 9 10  2  0  0  0  0
10 11  1  0  0  0  0
11 12  2  0  0  0  0
11 13  1  0  0  0  0
M  END
```

SD file(mol) Overview –V2000 format

L-Alanine



Chiral

```
  6  5  0  0  1  0              3 V2000
   -0.6622    0.5342    0.0000 C   0  0  2  0  0  0
    0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
   -0.7207    2.0817    0.0000 C   1  0  0  0  0  0
   -1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
    0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
    1.9464    0.4244    0.0000 O   0  5  0  0  0  0
  1  2  1  0  0  0
  1  3  1  1  0  0
  1  4  1  0  0  0
  2  5  2  0  0  0
  2  6  1  0  0  0
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

Blocks not used in this Ctab: Atom List, Stext

Counts Line

Atom Block

Bond Block

Properties Block

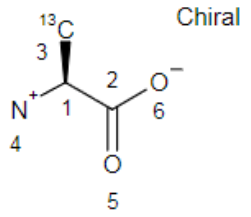
Connection Table (Ctab)

The format for a Ctab block is:

Counts line:	Important specifications here relate to the number of atoms, bonds, and atom lists, the chiral flag setting, and the Ctab version.
Atom block:	Specifies the atomic symbol and any mass difference, charge, stereochemistry, and associated hydrogens for each atom.
Bond block:	Specifies the two atoms connected by the bond, the bond type, and any bond stereochemistry and topology (chain or ring properties) for each bond.
Atom list block:	Identifies the atom (number) of the list and the atoms in the list.
Stext (structural text descriptor) block:	Used by ISIS/Desktop programs.
Properties block:	Provides for future expandability of Ctab features, while maintaining compatibility with earlier Ctab configurations.

SD file(mol) Overview –V3000 format

L-Alanine Chiral




```
L-Alanine
GSMACCS-II07189510252D 1   0.00366     0.00000     0
Figure 1, J. Chem. Inf. Comput. Sci., Vol 32, No. 3., 1992
  0  0  0     0  0            999 V3000
M  V30 BEGIN CTAB
M  V30 COUNTS 6 5 0 0 1
M  V30 BEGIN ATOM
M  V30 1 C -0.6622 0.5342 0 0 CFG=2
M  V30 2 C 0.6622 -0.3 0 0
M  V30 3 C -0.7207 2.0817 0 0 MASS=13
M  V30 4 N -1.8622 -0.3695 0 0 CHG=1
M  V30 5 O 0.622 -1.8037 0 0
M  V30 6 O 1.9464 0.4244 0 0 CHG=-1
M  V30 END ATOM
M  V30 BEGIN BOND
M  V30 1 1 1 2
M  V30 2 1 1 3 CFG=1
M  V30 3 1 1 4
M  V30 4 2 2 5
M  V30 5 1 2 6
M  V30 END BOND
M  V30 END CTAB
M  END
```

Header Block

Comment Line

Counts Line

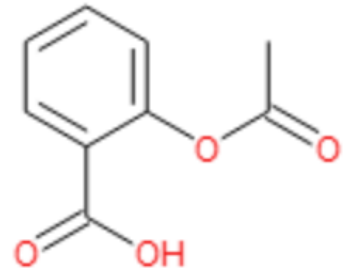
Atom Block

Bond Block

Connection Table (Ctab)

Blocks not used in this Ctab:
Sgroup block, Rgroup block, 3D block

Mol2, PDB




```
@<TRIPOS>MOLECULE
*****
 13 13 0 0 0
SMALL
GASTEIGER

@<TRIPOS>ATOM
      1 C          0.0000    0.0000    0.0000 C.3     1  UNL1        0.1216
      2 C          0.0000    0.0000    0.0000 C.2     1  UNL1        0.3220
      3 O          0.0000    0.0000    0.0000 O.2     1  UNL1       -0.2497
      4 O          0.0000    0.0000    0.0000 O.3     1  UNL1       -0.4239
      5 C          0.0000    0.0000    0.0000 C.ar    1  UNL1        0.1547
      6 C          0.0000    0.0000    0.0000 C.ar    1  UNL1        0.0378
      7 C          0.0000    0.0000    0.0000 C.ar    1  UNL1        0.0031
      8 C          0.0000    0.0000    0.0000 C.ar    1  UNL1        0.0004
      9 C          0.0000    0.0000    0.0000 C.ar    1  UNL1        0.0097
     10 C          0.0000    0.0000    0.0000 C.ar    1  UNL1        0.1143
     11 C          0.0000    0.0000    0.0000 C.2     1  UNL1        0.3907
     12 O          0.0000    0.0000    0.0000 O.co2   1  UNL1       -0.2404
     13 O          0.0000    0.0000    0.0000 O.co2   1  UNL1       -0.2404
@<TRIPOS>BOND
     1    1    2    1
     2    2    3    2
     3    2    4    1
     4    4    5    1
     5    5    6   ar
     6    6    7   ar
     7    7    8   ar
     8    8    9   ar
     9    9   10   ar
    10    5   10   ar
    11   10   11    1
    12   11   12   ar
    13   11   13   ar
```

```
COMPND    UNNAMED
AUTHOR    GENERATED BY OPEN BABEL 3.1.1
HETATM    1  C   UNL     1       0.000   0.000   0.000  1.00  0.00           C
HETATM    2  C   UNL     1       0.000   0.000   0.000  1.00  0.00           C
HETATM    3  O   UNL     1       0.000   0.000   0.000  1.00  0.00           O
HETATM    4  O   UNL     1       0.000   0.000   0.000  1.00  0.00           O
HETATM    5  C   UNL     1       0.000   0.000   0.000  1.00  0.00           C
HETATM    6  C   UNL     1       0.000   0.000   0.000  1.00  0.00           C
HETATM    7  C   UNL     1       0.000   0.000   0.000  1.00  0.00           C
HETATM    8  C   UNL     1       0.000   0.000   0.000  1.00  0.00           C
HETATM    9  C   UNL     1       0.000   0.000   0.000  1.00  0.00           C
HETATM   10  C   UNL     1       0.000   0.000   0.000  1.00  0.00           C
HETATM   11  C   UNL     1       0.000   0.000   0.000  1.00  0.00           C
HETATM   12  O   UNL     1       0.000   0.000   0.000  1.00  0.00           O
HETATM   13  O   UNL     1       0.000   0.000   0.000  1.00  0.00           O
CONECT    1    2
CONECT    2    1    3    3    4
CONECT    3    2    2
CONECT    4    2    5
CONECT    5    4   10    6    6
CONECT    6    5    5    7
CONECT    7    6    8    8
CONECT    8    7    7    9
CONECT    9    8   10   10
CONECT   10    9    9    5   11
CONECT   11   10   12   12   13
CONECT   12   11   11
CONECT   13   11
MASTER        0    0    0    0    0    0    0    0   13    0   13    0
END
```

신약개발의 데이터 정의(Molecule)

분자구조 표현 방법

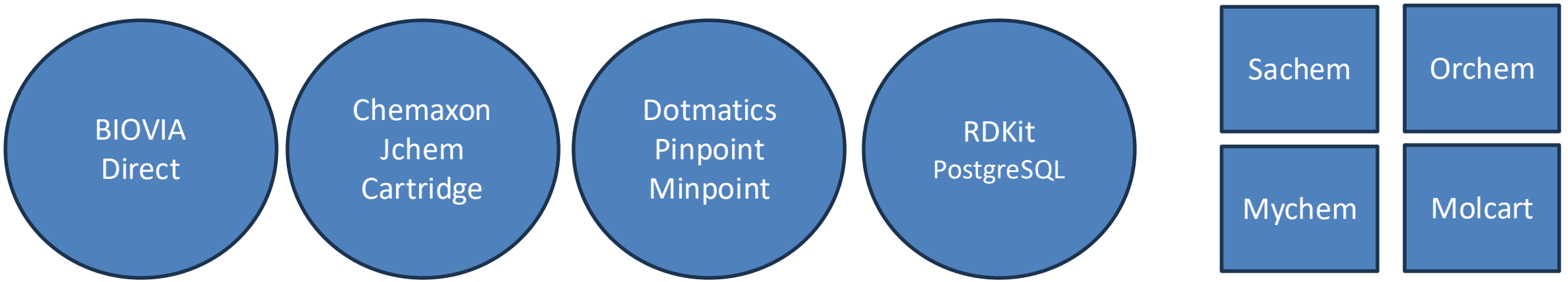
데이터 베이스

- Cartridge
- 구조 검색
- 공개 데이터베이스

QSAR 방법론

데이터 베이스

Molecular cartridge



- Cartridge 란 ?
  - 분자 구조 및 관련 데이터를 처리, 저장, 검색 및 조작하기 위해 화학 정보학 및 분자 모델링에 사용되는 소프트웨어 또는 도구를 의미
  - Oracle, PostgreSQL 또는 MySQL과 같은 관계형 데이터베이스 시스템에 사용되는 특수 확장 또는 플러그인 형식
  - 최근에는 DB 없이 검색이 가능한 형태도 개발이 되고 있음

Cartridge 작동 원리

관계형 데이터베이스와의 통합

- 화학 구조에 대한 특정한 추가 기능, 데이터 유형 및 인덱싱 메커니즘을 제공
- 특정 기능으로 확장된 SQL 쿼리를 사용하여 구조 기반 검색 및 분석을 수행

인덱싱

- Chemical fingerprint 나 Hash-based index과 같은 특수 색인 방법을 사용
- Molecular Fingerprints : 분자 특징(예: 특정 원자 유형의 존재, 결합 유형, 고리 시스템)을 나타내는 비트 문자열로, 구조 간의 빠른 비교

데이터 처리 및 관리

- 화학 데이터의 저장 및 검색을 관리하여 대규모 데이터 세트를 효율적으로 처리
  - tautomeric forms, stereochemistry, isotopes 등 관리
- 트랜잭션 관리를 지원

Cartridge 용도

화학 구조 저장

- SMILES(간단한 분자 입력 라인 입력 시스템), InChI(국제 화학 식별자) 및 MOL/SDF(분자/구조 데이터 파일)와 같은 다양한 화학 파일 형식으로 분자 구조를 저장
- 중복 구조 체크

Structure 검색

- Exact Search
- Substructure Search/Superstructure Search
- Similarity Search

구조 정규화 및 표준화

- 분자 구조를 자동으로 표준화하여 (예: 구조 이성질체를 일반적인 형태로 변환, 전하 정규화, Salt 제거) 데이터베이스 전체에서 일관성을 유지

구조-특성 계산

- 분자량, logP(분배계수), 극성 표면적 등과 같은 분자 특성(Descriptor)

화학 반응 처리

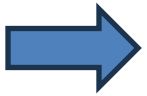
- 일부 카트리지는 반응 검색을 지원하며 화학반응, 반응물, 생성물, 반응조건을 관리

Cartridge 사용 예제

Chemaxon JCHEM

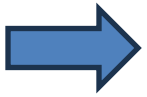
```
SELECT id FROM mystructures where jc_compare(smiles,
'Brc1ccccc1C=CCc1cccc(Br)c1CCc1ccccc1Br', 't:d') = 1
```

```
select jc_compare('c1ccccc1', 'Brc1ccccc1', 't:u') from dual;
```



1

```
select jc_compare('Clc1ccccc1', 'c1ccccc1', 't:u') from dual;
```



0

분자 구조 검색

분자 구조 검색 유형

Exact Search

- 정의: 원자 유형, 결합 순서 및 입체화학을 포함하여 쿼리 구조와 정확하게 일치하는 분자 검색
- 사용 사례: 특정 분자가 데이터베이스에 이미 존재하는지 확인

Substructure Search

- 정의
  - 특정 하위 구조 또는 기능 그룹을 포함하는 분자를 식별
  - 이는 쿼리가 대상 구조의 하위 그래프인 그래프 기반 검색
- 사용 사례: 특정 약리단, functional group 또는 scaffold 가 있는 분자를 검색

Superstructure Search

- 정의: Substructure Search와는 반대로 쿼리가 대상 구조의 상위 그래프인 그래프 기반 검색

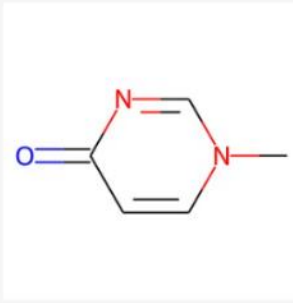
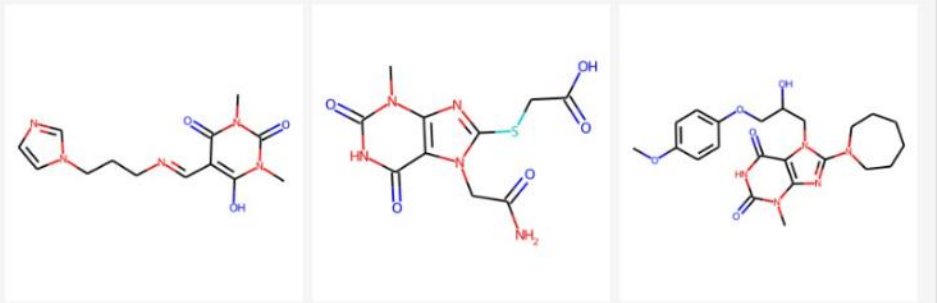
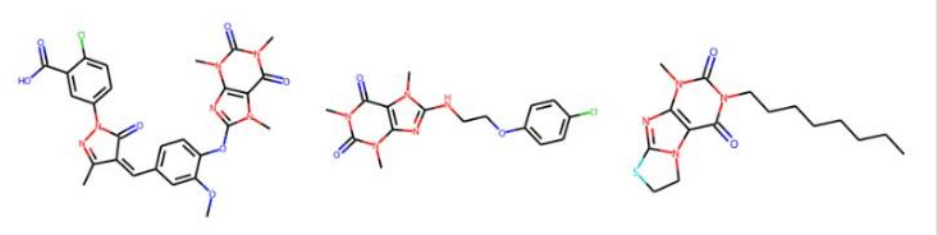
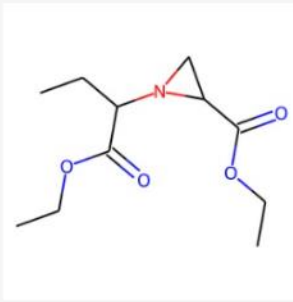
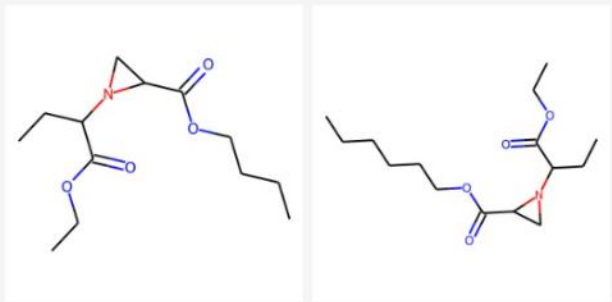
Similarity Search

- 정의: 사전 정의된 유사성 측정항목(예: 타니모토 계수)을 기반으로 쿼리 구조와 유사한 분자를 검색
- 사용 사례: 알려진 활성 화합물과 유사한 생물학적 활성을 갖는 화합물을 식별

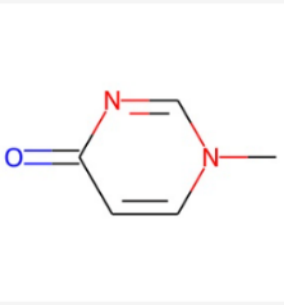
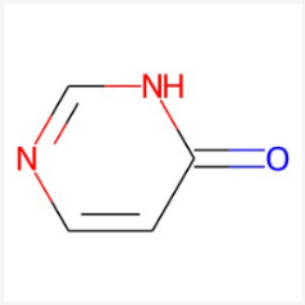
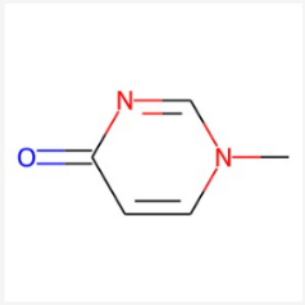
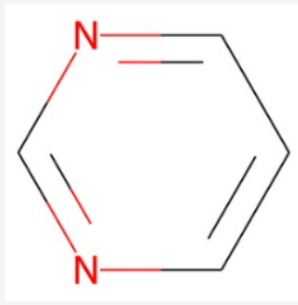
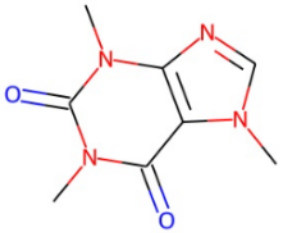
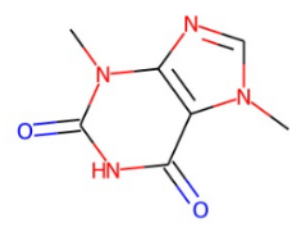
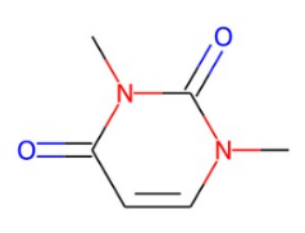
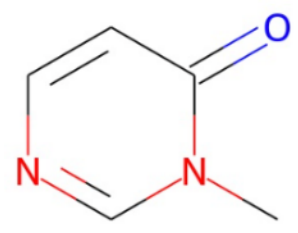
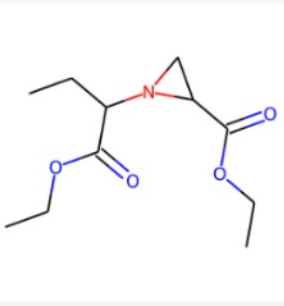
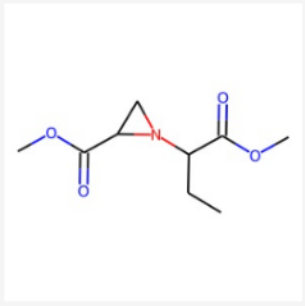
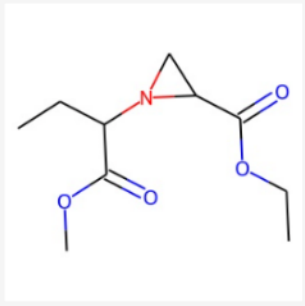
SMARTS Pattern Search ([X3&H0], [c,n;H1], [F,Cl,Br,I])

- 정의
  - SMARTS(SMiles ARbitrary Target Spec) 을 이용하여 복잡한 화학 패턴을 지정하여 검색
    - SMARTS(SMiles ARbitrary Target Spec) : SMILES(Simplified Molecular Input Line Entry System)를 확장하여 하위 구조 검색을 위한 패턴을 정의하는 언어
- 사용 사례: 입체화학 및 원자 속성 필터를 포함하는 고급 하위 구조 검색
- 참조 사이트 : https://daylight.com/dayhtml/doc/theory/theory.smarts.html

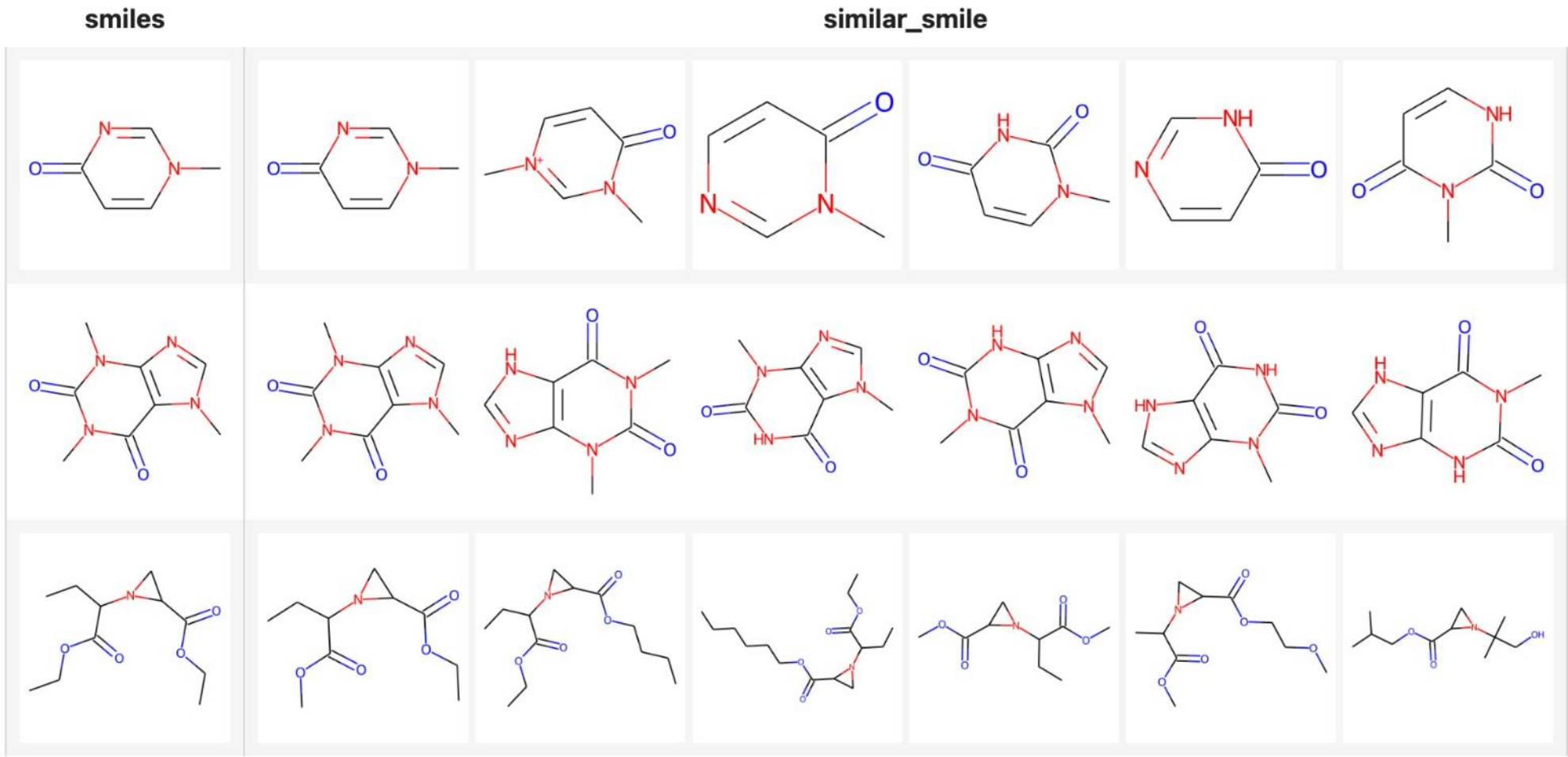
Substructure search

smiles	is_substructure
	
	
	

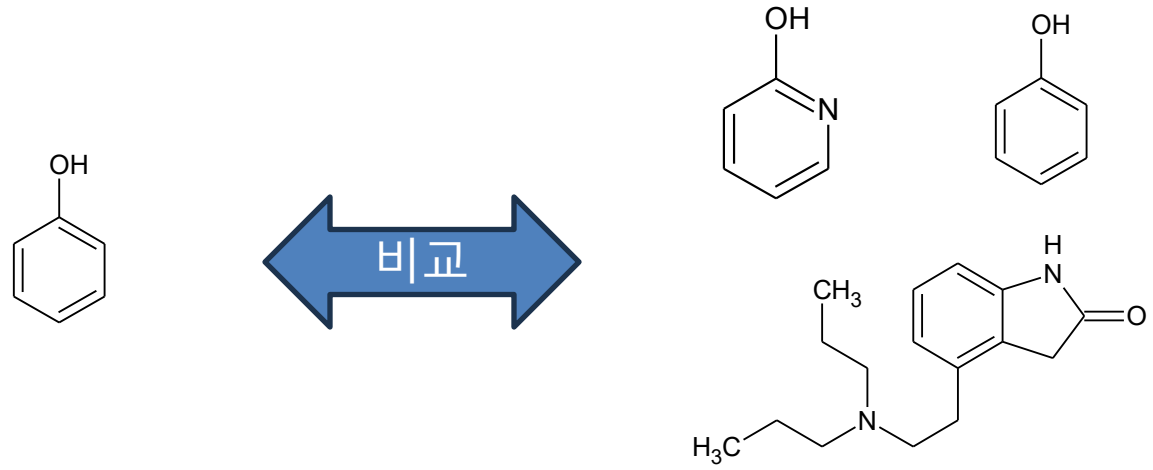
Superstructure search

smiles	is_superstructure		
			
			
			

Similarity Search

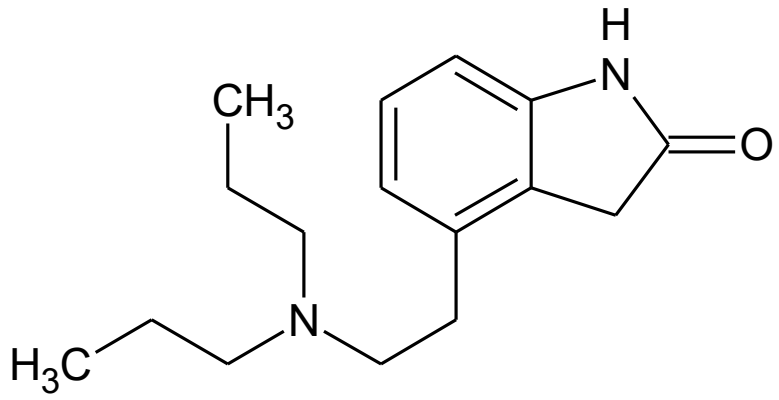


molecule 간의 비교



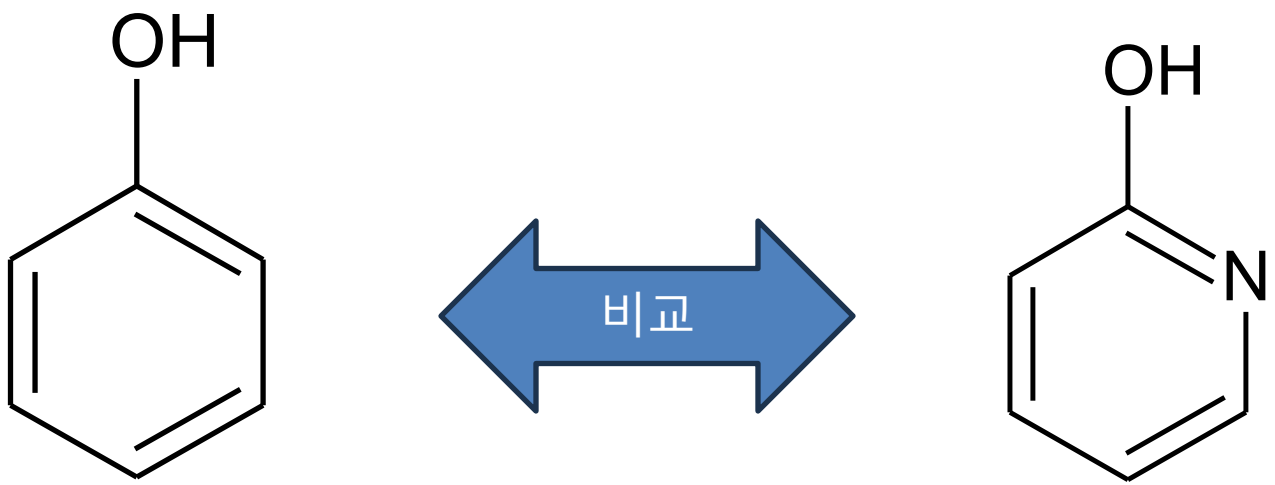
- Direct comparison of molecular structures is reasonably straightforward for a trained chemist, but more difficult for a computer
- The simplest approach, to compare every atom in one molecule to every one in a second molecule is slow, particularly if we wish to search a database containing thousands or even millions of entries
- We therefore need a quick way (even approximate) way to compare two molecules
- We can illustrate the idea of molecular fragments using SMILES strings.

Comparing molecules with fingerprints



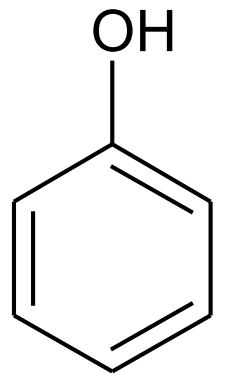
N	C	O	S	P	NH	Phenyl
1	1	1	0	0	1	1

Phenol vs 2-hydroxypyridine



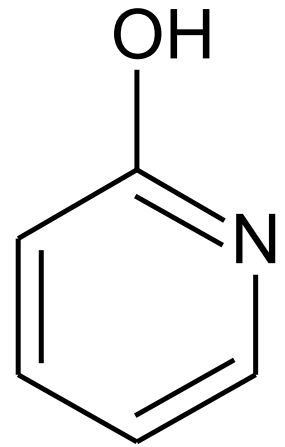
Fragments of phenol

Fragments of phenol	*SMILES*
One atom fragments:	c, O
Two atom fragments:	cc, cO
Three atom fragments	ccc, ccO
Four atom fragments:	cccc, cccO
Five atom fragments:	ccccc, ccccO
Six atom fragments:	c1ccccc1, cccccO
Seven atom fragments:	Oc1ccccc1



Fragments of 2-hydroxypyridine

- One atom fragments: c, n, O
- Two atom fragments: cn, cc, cO
- Three atom fragments: ccn, cnc, ccc, ncO, ccO
- Four atom fragments: cccn, ccnc, nccc, cccO, cccc, cncO
- Five atom fragments: ncccc, cnccc, ccncc, ccccO, cncO
- Six atom fragments: n1ccccc1, Ocnccc, Occccn
- Seven atom fragments: n1ccc(O)cc1



Molecular fingerprints

- A set of molecular fragments for a particular molecule can be assembled to form a molecular fingerprint.
- A fingerprint is a binary number made up of the digits 1 and 0.
- Each position (bit) in the string denotes a possible molecular fragment.
- The digit is set to 1 to denote that a particular fragment is present in the molecule

Comparing fingerprints

- For example, bit strings for phenol (**Ph**) and 2-hydroxypyridine (**2HPy**) might look like this:

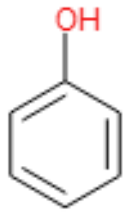
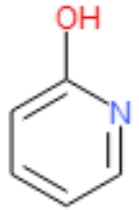
	one atom fragments			two atom fragments				three atom fragments									
	c	n	O	cc	cn	cO	nO	ccc	ccn	cnc	ccO	cnO	ncO	nnn	nnc	ncn	ccO
Ph	1	0	1	1	0	1	0	1	0	0	1	0	0	0	0	0	1
2HPy	1	1	1	1	1	1	0	1	1	1	1	0	1	0	0	1	1

Fingerprint

- MDL mol key (166 key

Key	Description	Key	Description	Key	Description	Key	Description
1	ISOTOPE	41	CTN	81	SA(A)A	121	N HETEROCYCLE
2	103 < ATOMIC NO. < 256	42	F	82	ACH2QH	122	AN(A)A
3	GROUP IVA, VA, VIA PERIODS 4-6 (GE..)	43	QHAQH	83	QAAAA@1	123	OCO
4	ACTINIDE	44	OTHER	84	NH2	124	QQ
5	GROUP IIIB,IVB (SC..)	45	C=CN	85	CN(C)C	125	AROMATIC RING>1
6	LANTHANIDE	46	BR	86	CH2QCH2	126	A!O!A
7	GROUP VB, VIB, VIIB (V..)	47	SAN	87	X!A$A	127	A$A!O>1 (&..)
8	QAAA@1	48	OQ(O)O	88	S	128	ACH2AAACH2A
9	GROUP VIII (FE..)	49	CHARGE	89	OAAAO	129	ACH2AACH2A
10	GROUP IIA (ALKALINE EARTH)	50	C=C(C)C	90	QHAACH2A	130	QQ>1 (&..)
11	4M RING	51	CSO	91	QHAAACH2A	131	QH>1
12	GROUP IB, IIB (CU..)	52	NN	92	OC(N)C	132	OACH2A
13	ON(C)C	53	QHAAAQH	93	QCH3	133	A$A!N
14	S-S	54	QHAAQH	94	QN	134	X (HALOGEN)
15	OC(O)O	55	OSO	95	NAAO	135	Nnot%A%A
16	QAA@1	56	ON(O)C	96	5M RING	136	O=A.1
17	CTC	57	O HETEROCYCLE	97	NAAAO	137	HETEROCYCLE
18	GROUP IIIA (B..)	58	QSQ	98	QAAAAA@1	138	QCH2A>1 (&..)
19	7M RING	59	Snot%A%A	99	C=C	139	OH
20	Si	60	S=O	100	ACH2N	140	O>3 (&..)
21	C=C(Q)Q	61	AS(A)A	101	8M RING OR LARGER	141	CH3>2 (&..)
22	3M RING	62	A$A!A$A	102	QO	142	N>1
23	NC(O)O	63	N=O	103	CL	143	A$A!O
24	N-O	64	A$A!S	104	QHACH2A	144	Anot%A%Anot%A
25	NC(N)N	65	C%N	105	A$A($A)$A	145	6M RING>1
26	C$=C($A)$A	66	CQ(C)(C)A	106	QA(Q)Q	146	O>2
27	I	67	QS	107	XA(A)A	147	ACH2CH2A
28	QCH2Q	68	QHQH (&..)	108	CH3AAACH2A	148	AQ(A)A
29	P	69	QQH	109	ACH2O	149	CH3>1
30	CQ(C)(C)A	70	QNQ	110	NCO	150	A!A$A!A
31	QX	71	NO	111	NACH2A	151	NH
32	CSN	72	OAAO	112	AA(A)(A)A	152	OC(C)C
33	NS	73	S=A	113	Onot%A%A	153	QCH2A
34	CH2=A	74	CH3ACH3	114	CH3CH2A	154	C=O
35	GROUP IA (ALKALI METAL)	75	A!N$A	115	CH3ACH2A	155	A!CH2!A
36	S HETEROCYCLE	76	C=C(A)A	116	CH3AACH2A	156	NA(A)A
37	NC(O)N	77	NAN	117	NAO	157	C-O
38	NC(C)N	78	C=N	118	ACH2CH2A>1	158	C-N
39	OS(O)O	79	NAAN	119	N=A	159	O>1
40	S-O	80	NAAAN	120	HETEROCYCLIC ATOM>1 (&..)	160	CH3
						161	N
						162	AROMATIC
						163	6M RING
						164	O
						165	RING
						166	FRAGMENTS

MDL key example

Data Image	name	MDL2DKeys166
 phenol	phenol	113 139 143 152 157 162 163 164 165
 2-hydroxypyridine	2-hydroxypyridine	65 92 98 110 113 117 121 137 139 143 156 157 161 162 163 164 165

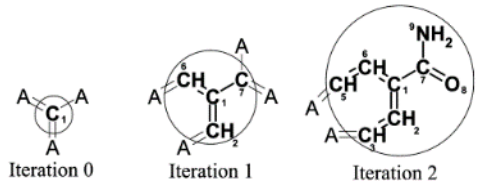
Fingerprint 의 종류

Structural(Path-Based) Fingerprint

- 정의
  - 분자 내의 선형 경로(원자와 결합의 순서)를 지정, 각 경로는 fingerprint의 비트로 인코딩
- 예 : Daylignt Fingerprint, RDKIT Fingerprint

Circular(Extended Connectivity) Fingerprint

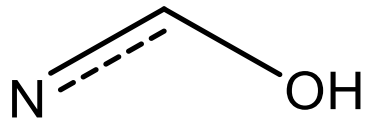
- 정의 : 원자의 원형 이웃을 인코딩, 특정 반경까지 각 원자 주변의 환경을 기반으로 특징을 지정
- 예
  - ECFP4 or ECFP6 : 반경이 2(ECFP4) 및 3(ECFP6)인 원형 fingerprint의 특징 구현
  - FCFP(Functional Connectivity Fingerprints) : Functional group에 중점



Substructure(Key-Based) Fingerprint

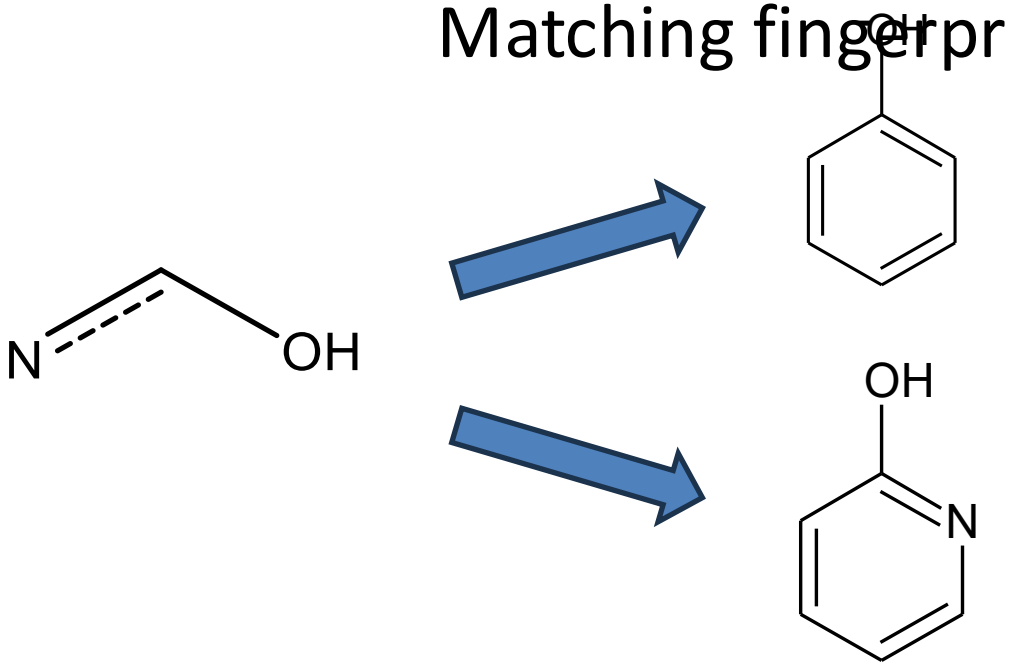
- 정의 : 사전 정의된 substructure 구조 또는 분자 단편 세트를 기반
- 예
  - MACCS : 방향족 고리, 특정 원자 유형 등과 같은 일반적인 하위 구조를 나타내는 사전 정의된 166개의 구조 키로 구성된 공통 세트
  - PubChem Fingerprint : Pubchem Database에 사용되는 사전 정의된 세트

Substructure fingerprints



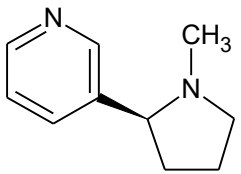
	one atom fragments			two atom fragments				three atom fragments									
	c	**n**	**O**	**cc**	**cn**	**cO**	**nO**	**ccc**	**ccn**	**cnc**	**ccO**	**cnO**	**ncO**	**nnn**	**nnc**	**ncn**	**ccO**
S	1	1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0

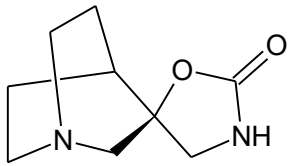
Matching fingerprints

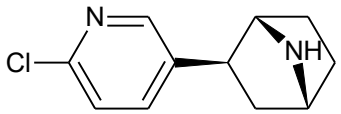


Ph	1	0	1	1	0	1	0	1	0	0	1	0	0	0	0	0	1
2HPh	1	1	1	1	1	1	0	1	1	1	1	0	1	0	0	1	1
S	1	1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0

How similar is similar?





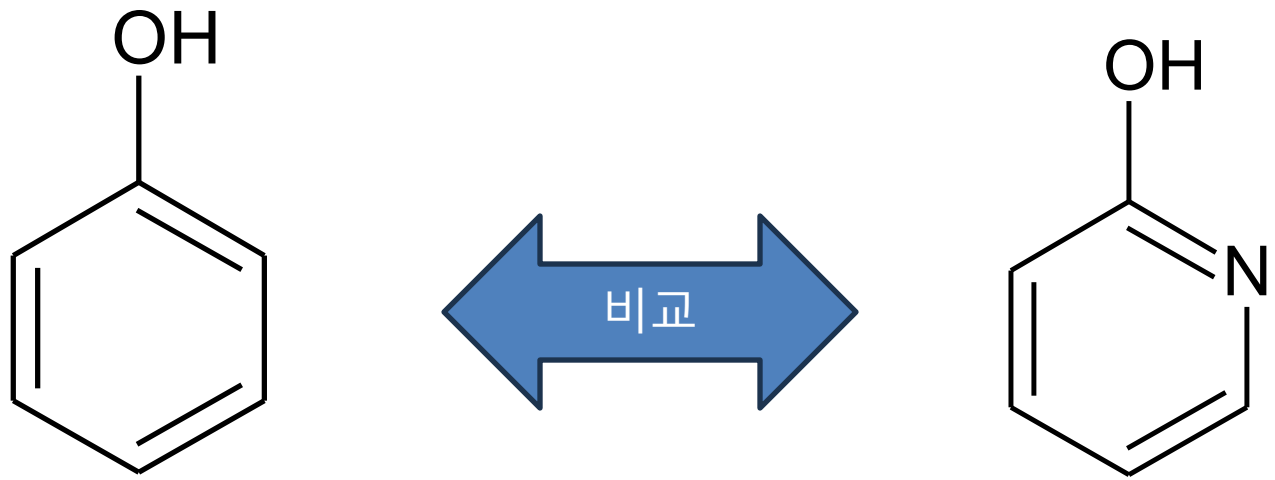


Discuss

Molecular similarity

- Similarity 원리:
  - 구조적으로 유사한 구조는 유사한 속성을 가진다.
    - Properties: biological activity, solubility, color and so on
- Similarity 측정으로 할 수 있는 것
  - distance matrix 구축
    - Distance = inverse of similarity
    - Compound Clustering
    - Such matrices can be used to cluster compounds, to create a 2D depiction showing the spread of molecular structures in a dataset, to select a diverse subset
  - Database에서 특정 쿼리와 유사한 구조를 검색
    - 유사한 속성을 가진 알려지지 않은 구조 검색
  - 특정 속성과 분자 구조 유사성의 상관 관계를 확인
- Similarity 측정 방법 ➔ Molecular fingerprint

Molecular Similarity



	one atom fragments			two atom fragments				three atom fragments									
	c	n	O	cc	cn	cO	nO	ccc	ccn	cnc	ccO	cnO	ncO	nnn	nnc	ncn	ccO
Ph	1	0	1	1	0	1	0	1	0	0	1	0	0	0	0	0	1
2HPy	1	1	1	1	1	1	0	1	1	1	1	0	1	0	0	1	1

Chemical Diversity and Similarity

- Molecular (dis)similarity can be measured in a large number of different ways.
- One important method, *the* ***Tanimoto Coefficient***, makes use of molecular fragments.

$$T = \frac{N_c}{N_1 + N_2 - N_c}$$

- This is a value between 0 and 1 that describes the number of common structural fragments in a molecule.
- N_1 Number of fragments in molecule 1
- N_2 Number of fragments in molecule 2
- N_c Number of fragments common to both molecules 1 and 2

Fi

- SA: Number of fingerprint features common to both records. (Number of "and" bits.)
- SB: Number of features present in the first record but absent from the second.
- SC: Number of features present in the second record but absent from the first.

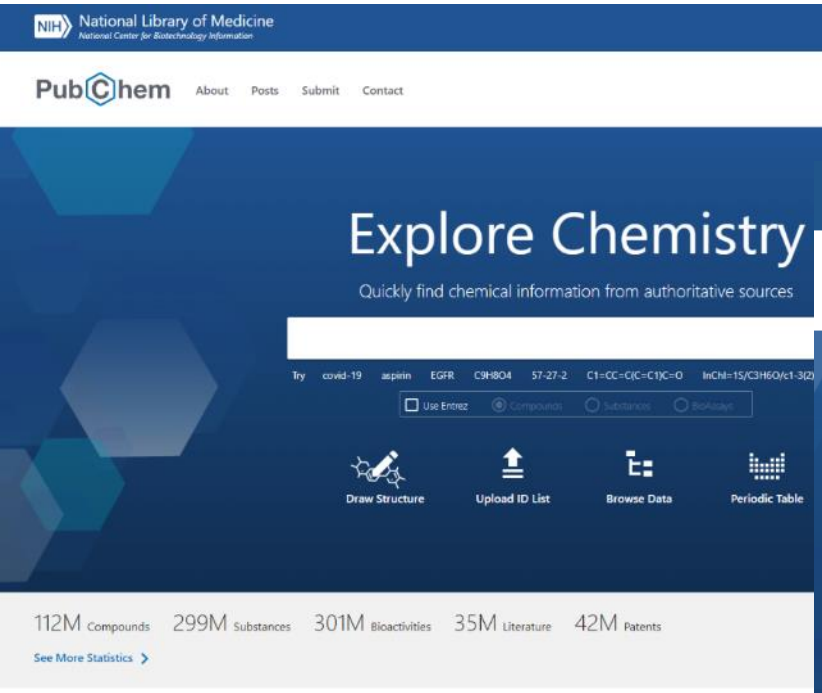
Option	Mathematical Definition
Cosine	Distance (D) is 1 minus the cosine of the angle between the two fingerprints when treated as binary vectors (vectors whose elements must be 0 or 1). $D = 1 - \frac{SA}{[(SA+SB)(SA+SC)]^{1/2}}$ Possible values range from 0 to 1.
Dice	Distance is 1 minus the number of "and" bits scaled by the mean number of "on" bits for both records. $D = 1 - \frac{2SA}{(2SA+SB+SC)}$ Possible values range from 0 to 1.
Tanimoto	Distance is 1 minus the number of "and" bits scaled by the number of "or" bits. $D = 1 - \frac{SA}{(SA+SB+SC)}$ Possible values range from 0 to 1.
Euclidean	Distance is the square-root of the number of "on" bits in the squared difference between the two fingerprints when treated as vectors. For a binary fingerprint, this can be expressed as: $D = (SB + SC)^{1/2}$ Possible values range from 0 to the total number of features possible in a single fingerprint.
Manhattan	Distance is the number of "on" bits in the absolute difference between the two fingerprints when treated as vectors. For a binary fingerprint, this can be expressed as: $D = SB + SC$ Possible values range from 0 to the total number of features possible in a single fingerprint.

공개 데이터베이스

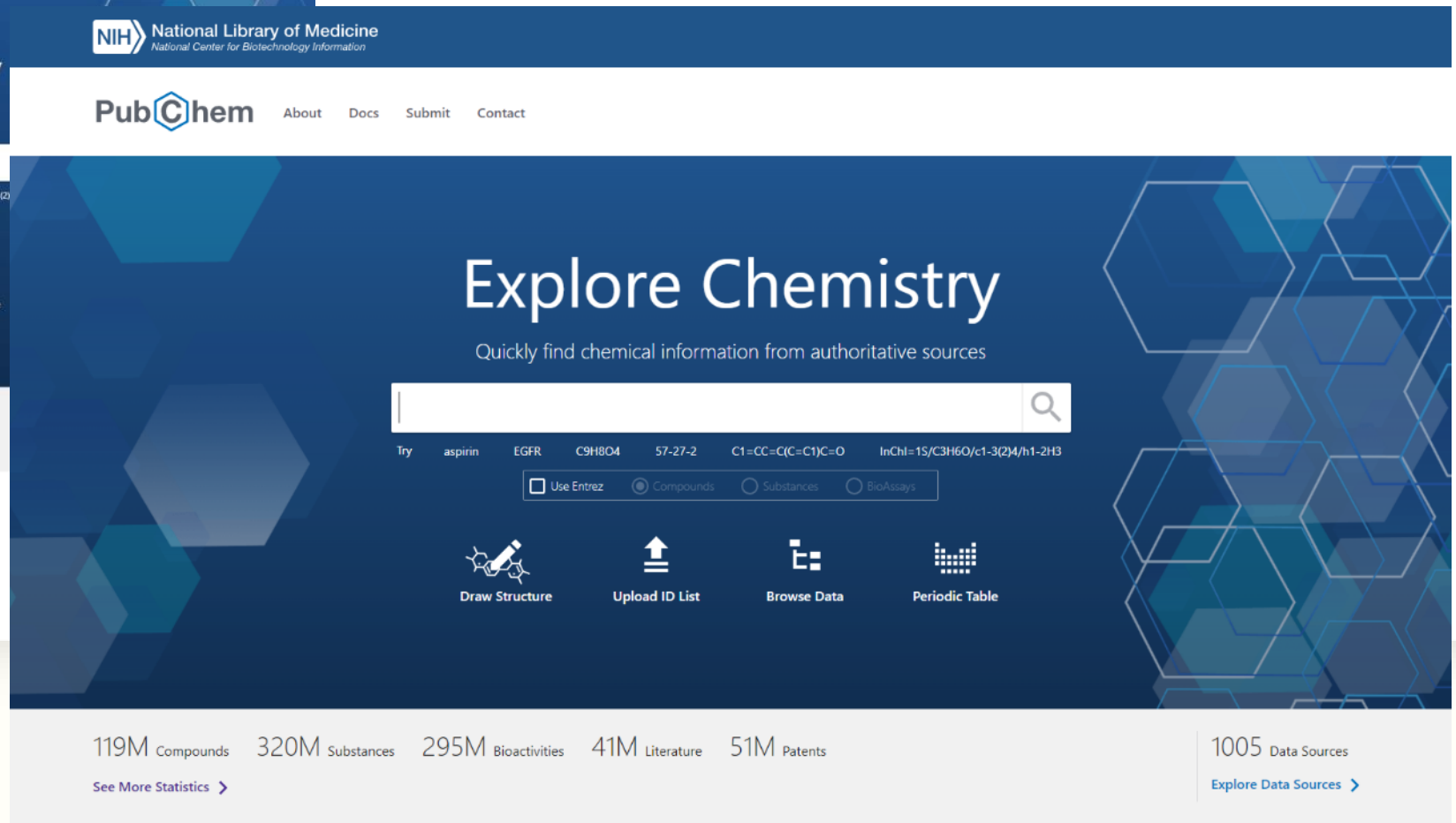
2022

Name	Description	Compound Counts
PubChem Compounds	PubChem 데이터베이스에서 정규화된 PubChem 화합물(CID)의 데이터베이스	110283434
Mcule	가상 스크리닝 및 분자 모델링 서비스를 제공하는 온라인 Drug Discovery 플랫폼	32919693
SureChEMBL	모든 주요 특허 문서에서 화학구조를 자동 추출하여 만든 DB. 화합물은 텍스트 또는 화학 이미지에서 발견되는 화학 이름에서 파생 UniChem 로딩 규칙을 위반한 제품을 제외한 모든 SureChEMBL 화합물이 포함	22094771
ZINC	가상 스크리닝을 위해 상업적으로 이용 가능한 화합물의 무료 데이터베이스로, 캘리포니아 대학교 샌프란시스코 (UCSF) 제약 화학과의 Shoichet 연구소에서 제공 [Irwin and Shoichet, J. Chem. Inf. Model. 2005;45(1):177-82]	16886865
MolPort	사용자가 화합물의 상업적 공급원을 찾을 수 있도록 설계된 데이터베이스.	7803496
eMolecules	수백만 개의 공개 도메인 구조를 포함하는 무료 화학 구조 검색 엔진. 가격, 가용성 및 공급업체 정보를 이용하려면 eMolecules Plus 구독이 필요	5168336
ChEMBL	과학 문헌에서 추출한 생리 활성 약물 소분자 및 생체 활성 데이터베이스.	2303816
BindingDB	약물 표적으로 간주되는 단백질과 작은 약물 유사 분자의 상호 작용에 중점을 둔DB로 측정된 결합 친화도에 대한 웹에서 액세스할 수 있는 공개 데이터베이스	1000223

PubChem (https://pubchem.ncbi.nlm.nih.gov/)



2022

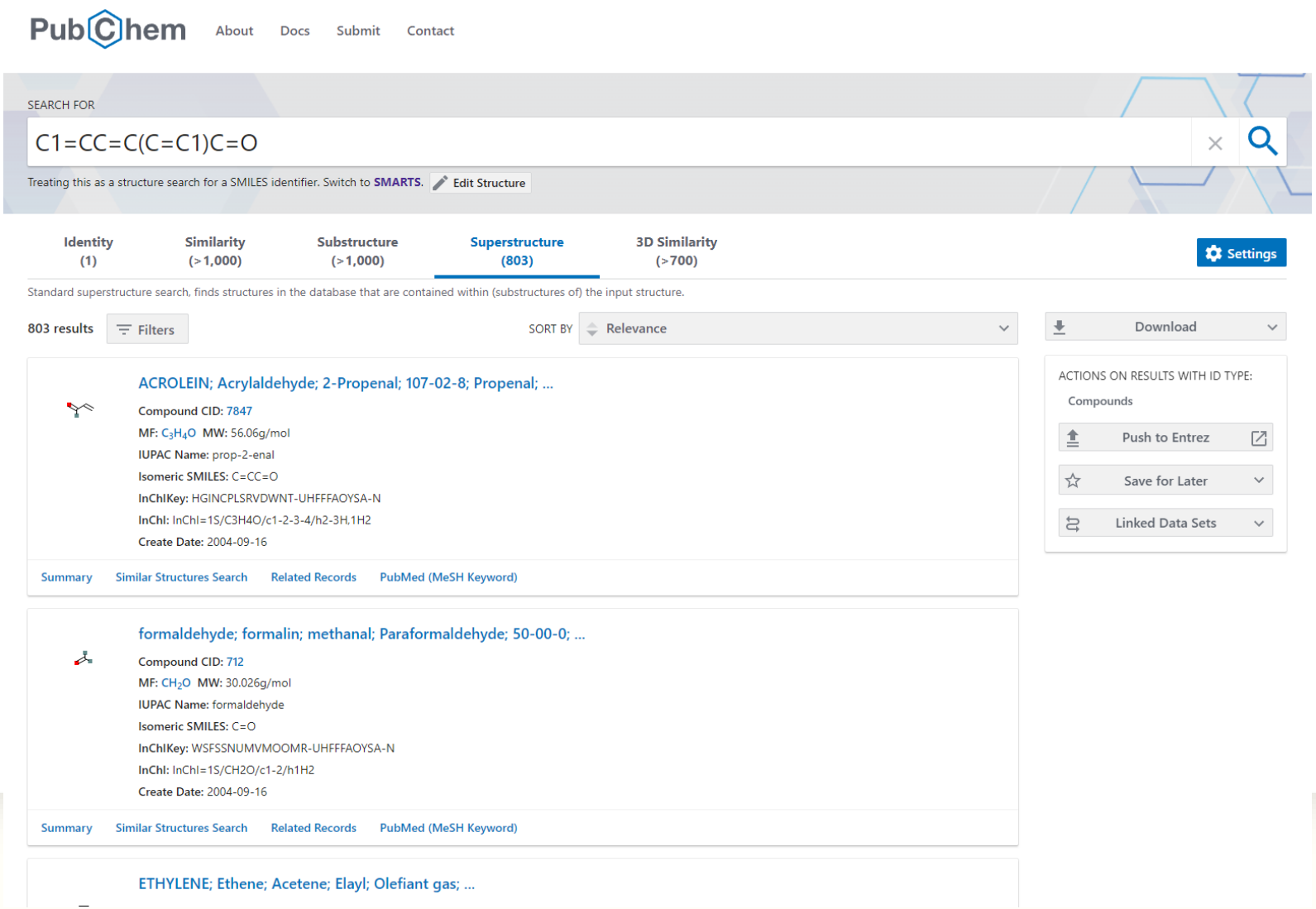


PubChem

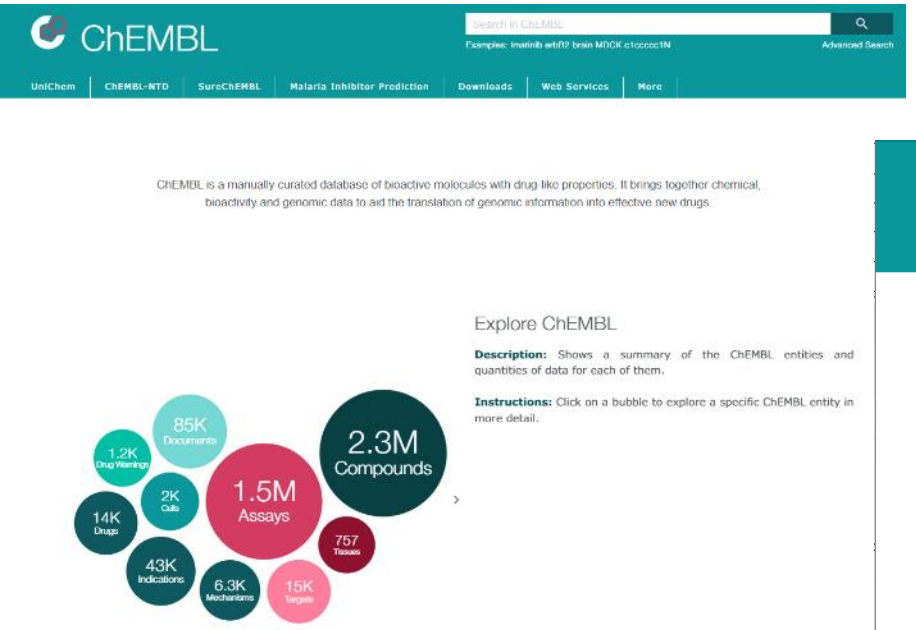
PubChem Data Counts

Collection	Live Count	Description
Periodic Table of Elements	118	Interactive periodic table with up-to-date element property data collected from authoritative sources
Compounds	118,576,011	Unique chemical structures extracted from contributed PubChem Substance records
Substances	319,906,730	Information about chemical entities provided by PubChem contributors
BioAssays	1,671,314	Biological experiments provided by PubChem contributors
Bioactivities	295,339,102	Biological activity data points reported in PubChem BioAssays
Genes	113,242	Gene targets tested in PubChem BioAssays and those involved in PubChem Pathways
Proteins	247,990	Protein targets tested in PubChem BioAssays and those involved in PubChem Pathways
Taxonomy	108,188	Organisms of targets tested in PubChem BioAssays and those involved in PubChem Pathways
Pathways	241,163	Interactions between chemicals, genes, and proteins
Cell Lines	2,005	Information about cell lines
Literature	41,383,055	Scientific publications with links in PubChem
Patents	50,836,952	Patents with links in PubChem
Data Classifications	73	Browse the distribution of PubChem data among nodes in the hierarchy of interest
Data Sources	1,005	Organizations contributing data to PubChem

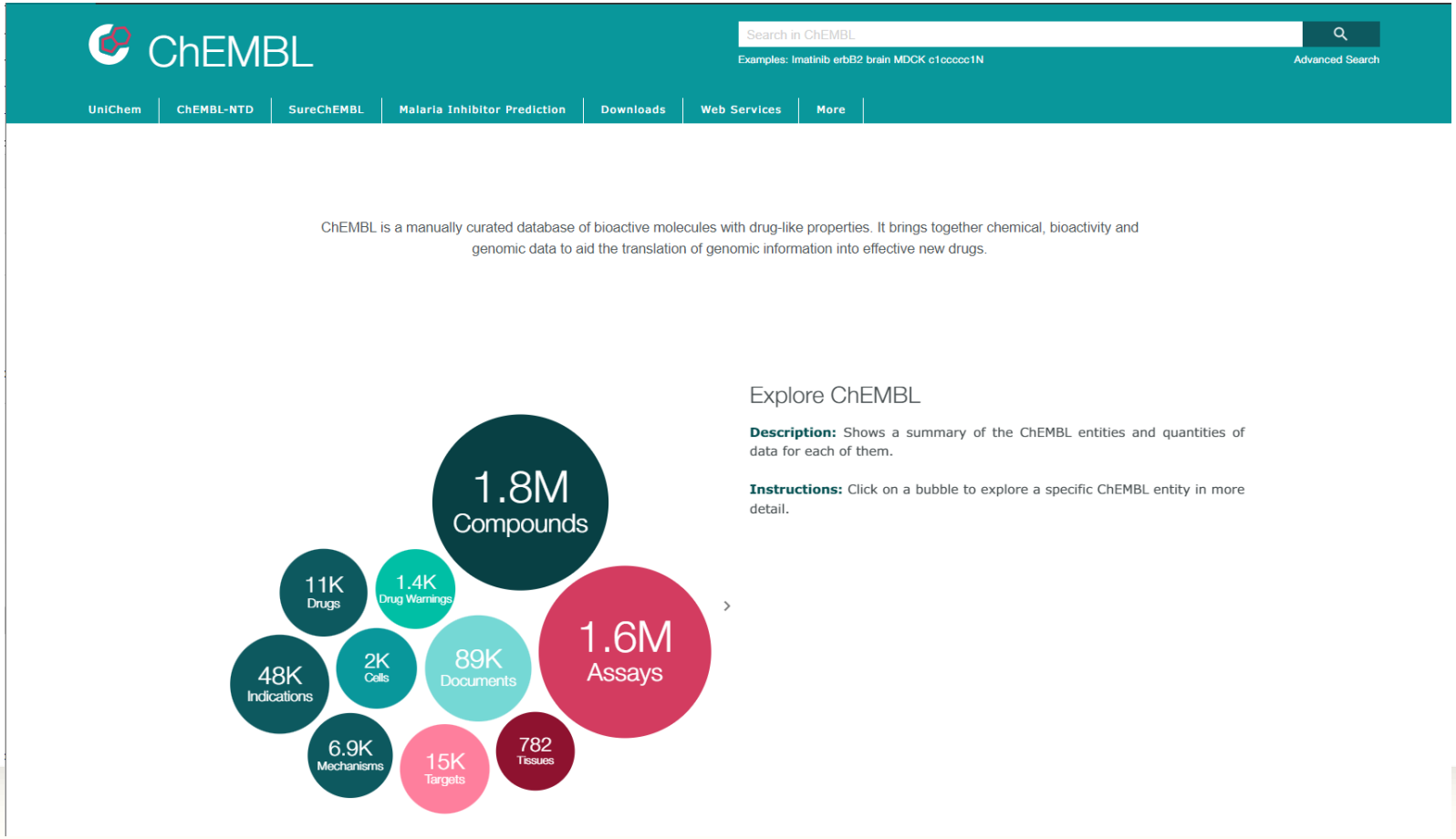
Pubchem


PubChem
About
Docs
Submit
Contact
SEARCH FOR
C1=CC=C(C=C1)C=O
Treating this as a structure search for a SMILES identifier. Switch to SMARTS.
Edit Structure
Identity (1)
Similarity (>1,000)
Substructure (>1,000)
Superstructure (803)
3D Similarity (>700)
Settings
Standard superstructure search, finds structures in the database that are contained within (substructures of) the input structure.
803 results
Filters
SORT BY
Relevance
Download
ACTIONS ON RESULTS WITH ID TYPE:
Compounds
Push to Entrez
Save for Later
Linked Data Sets
ACROLEIN; Acrylaldehyde; 2-Propenal; 107-02-8; Propenal; ...
Compound CID: 7847
MF: C3H4O MW: 56.06g/mol
IUPAC Name: prop-2-enal
Isomeric SMILES: C=CC=O
InChIKey: HGINCPLSRVDWNT-UHFFFAOYSA-N
InChI: InChI=1S/C3H4O/c1-2-3-4/h2-3H,1H2
Create Date: 2004-09-16
Summary
Similar Structures Search
Related Records
PubMed (MeSH Keyword)
formaldehyde; formalin; methanal; Paraformaldehyde; 50-00-0; ...
Compound CID: 712
MF: CH2O MW: 30.026g/mol
IUPAC Name: formaldehyde
Isomeric SMILES: C=O
InChIKey: WSFSSNUMVMOOMR-UHFFFAOYSA-N
InChI: InChI=1S/CH2O/c1-2/h1H2
Create Date: 2004-09-16
Summary
Similar Structures Search
Related Records
PubMed (MeSH Keyword)
ETHYLENE; Ethene; Acetene; Elayl; Olefiant gas; ...

ChEMBL(https://www.ebi.ac.uk/chembl/)



2022



ChEMBL

Current Release: ChEMBL 34

Last Update on 2024-03-28 | Release notes

Long-term data preservation

15,598
Targets

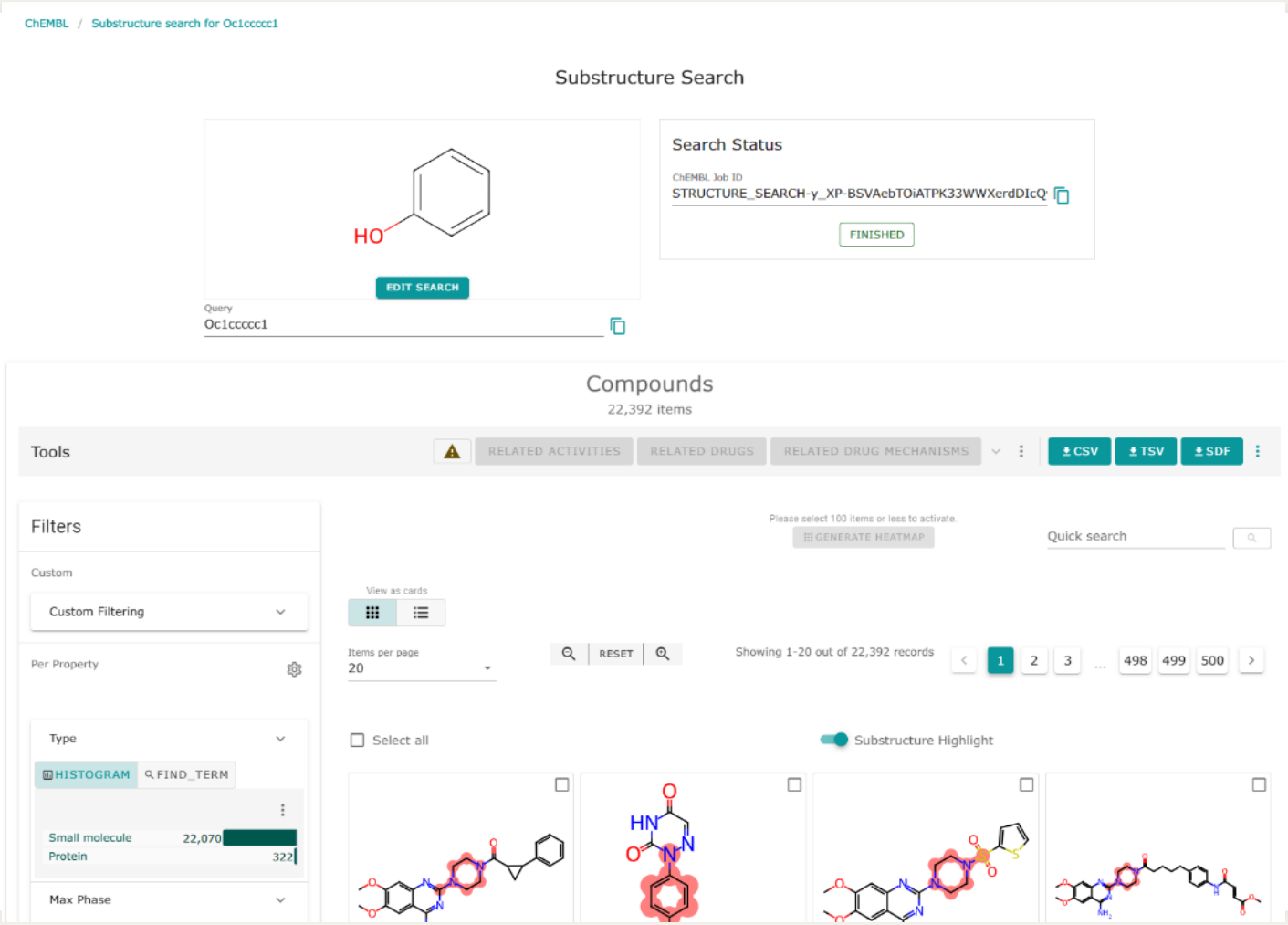
2,431,025
Distinct compounds

20,772,701
Activities

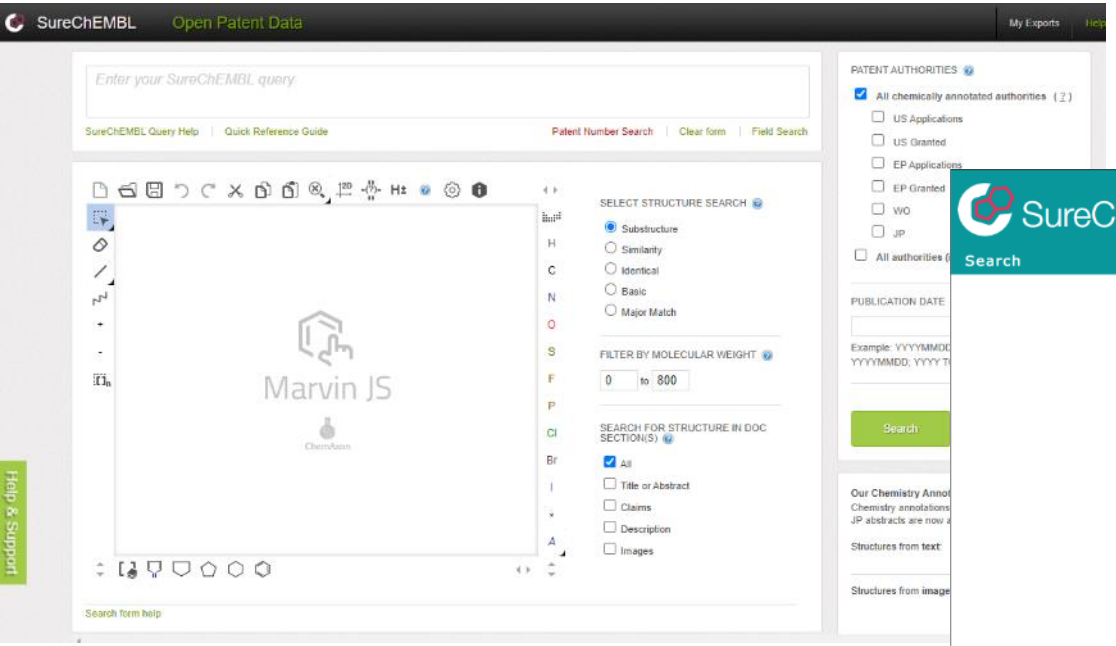
89,892
Publications

262
Deposited Datasets

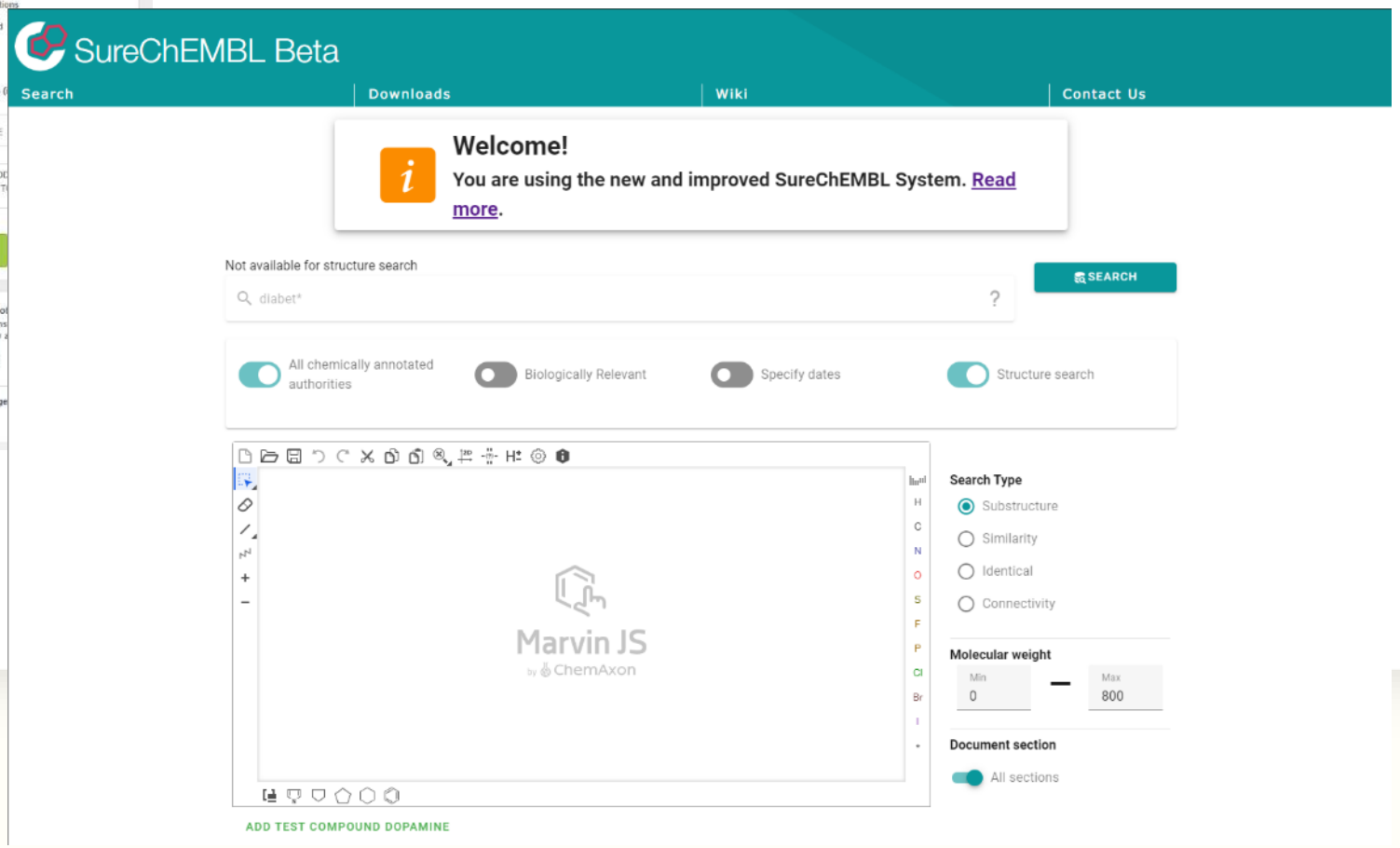
ChEMBL


ChEMBL / Substructure search for Oc1ccccc1
Substructure Search
HO
EDIT SEARCH
Query
Oc1ccccc1
Search Status
ChEMBL Job ID
STRUCTURE_SEARCH-y_XP-BSVAebTOiATPK33WWXerdDIcQ
FINISHED
Compounds
22,392 items
Tools
RELATED ACTIVITIES
RELATED DRUGS
RELATED DRUG MECHANISMS
CSV
TSV
SDF
Filters
Custom
Custom Filtering
Per Property
Type
HISTOGRAM
FIND_TERM
Small molecule 22,070
Protein 322
Max Phase
Please select 100 items or less to activate.
GENERATE HEATMAP
Quick search
View as cards
Items per page
20
RESET
Showing 1-20 out of 22,392 records
1
2
3
498
499
500
Select all
Substructure Highlight

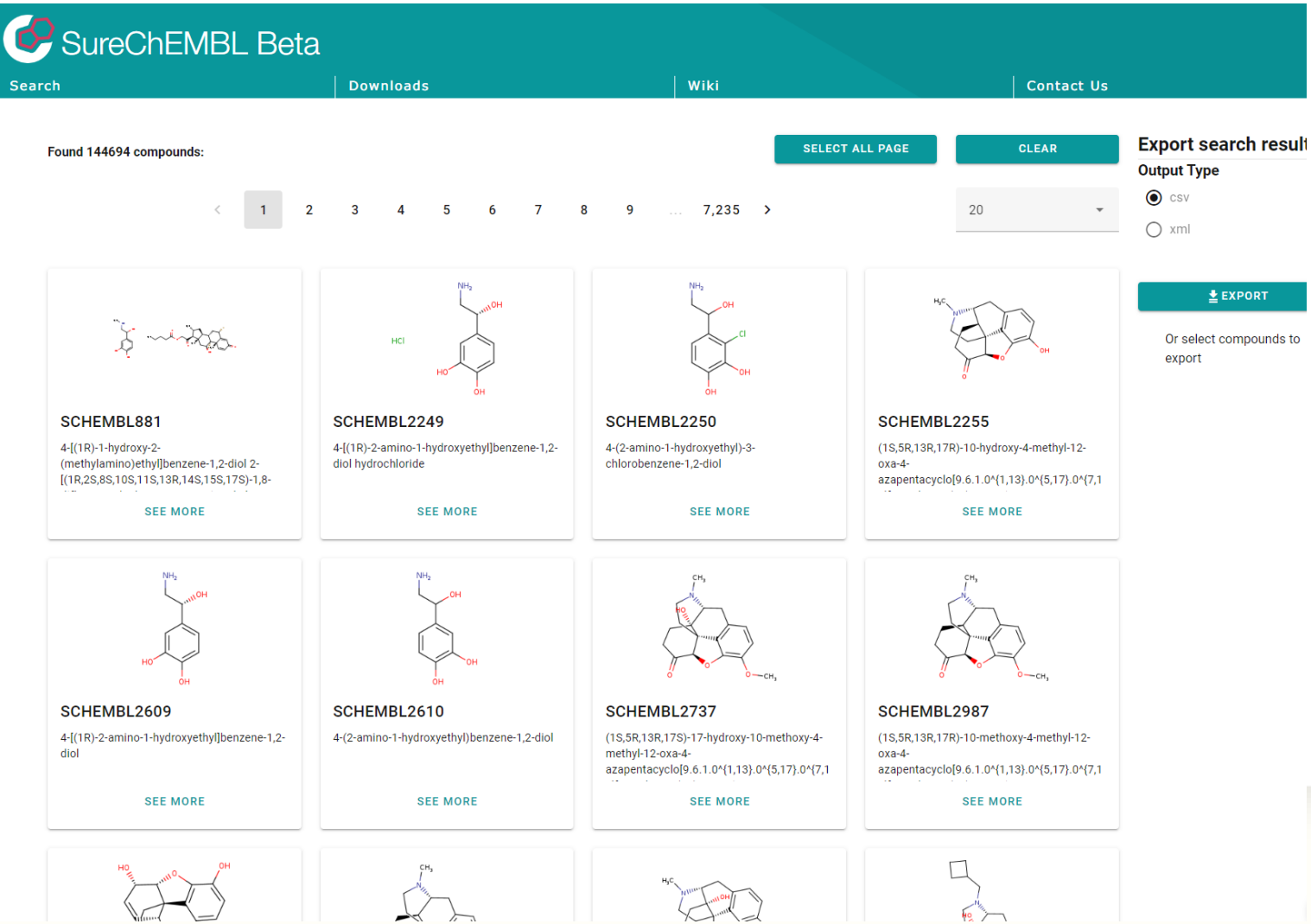
SureChEMBL (https://www.surechembl.org/)



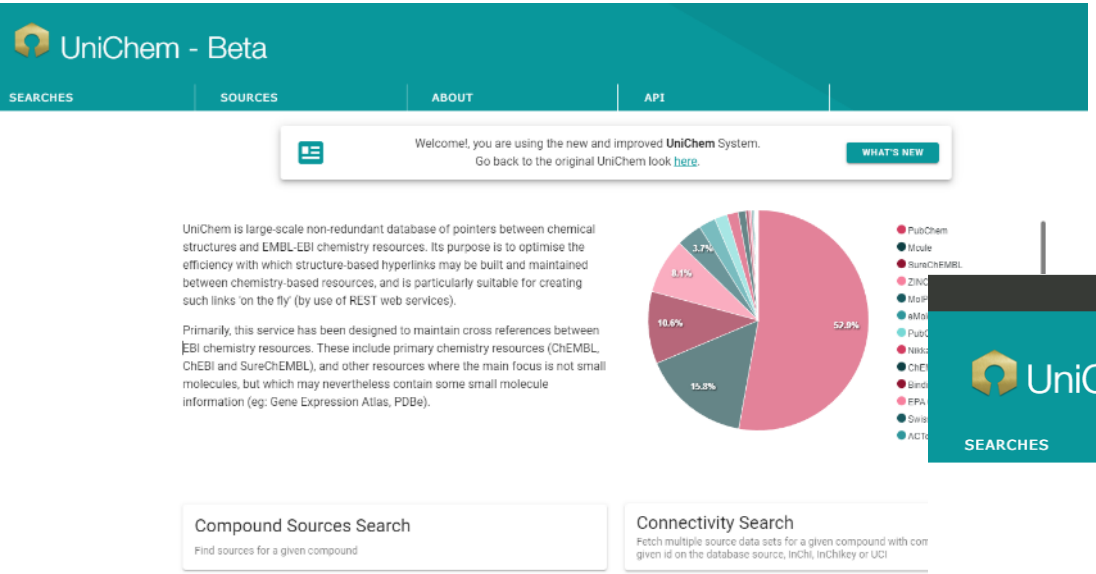
2022



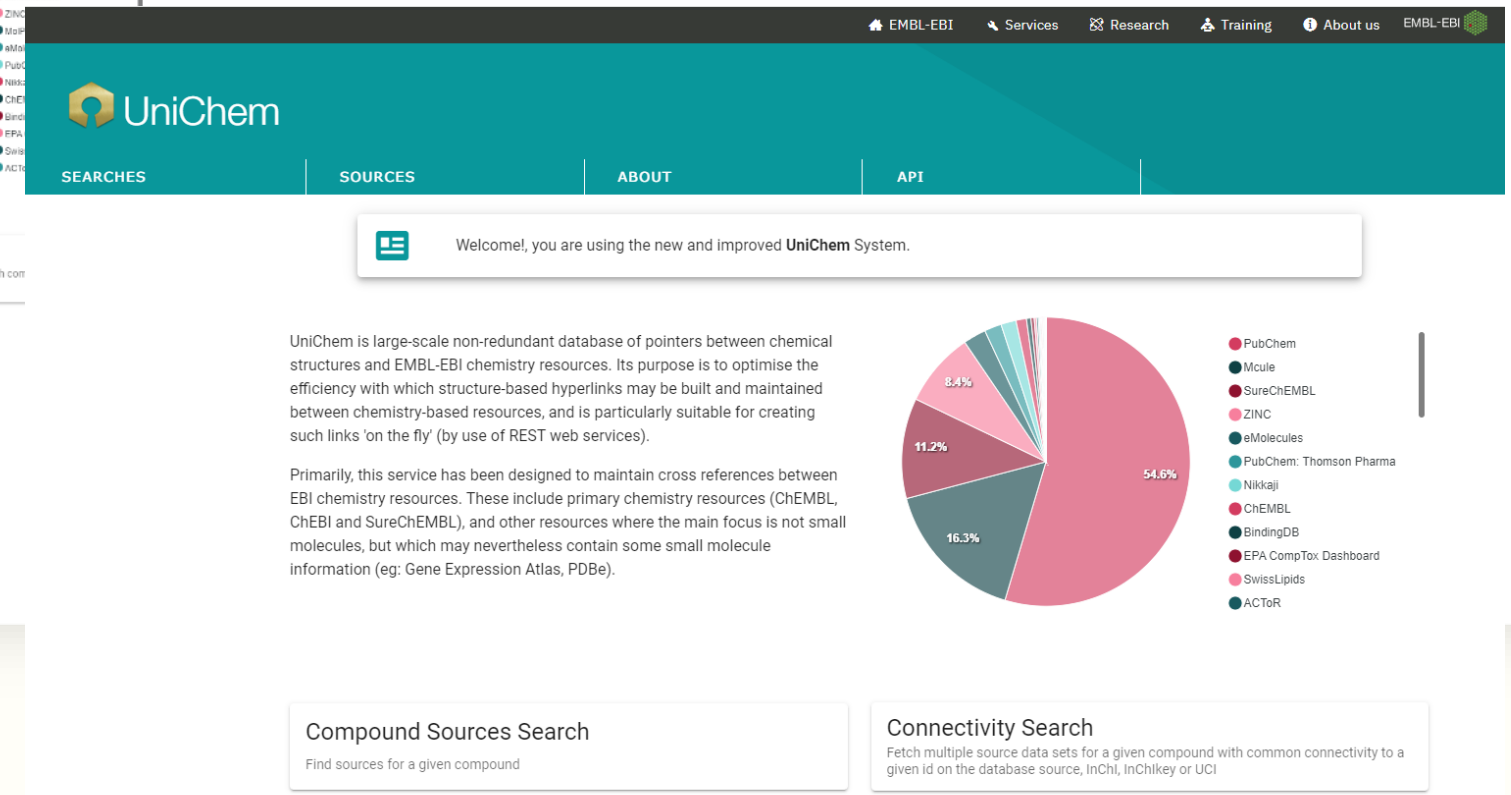
SureChEMBL


SureChEMBL Beta
Search
Downloads
Wiki
Contact Us
Found 144694 compounds:
SELECT ALL PAGE
CLEAR
1 2 3 4 5 6 7 8 9 ... 7,235
20
Export search result
Output Type
csv
xml
EXPORT
Or select compounds to export
SCHEMBL881
4-[(1R)-1-hydroxy-2-(methylamino)ethyl]benzene-1,2-diol 2-[(1R,2S,8S,10S,11S,13R,14S,15S,17S)-1,8-
SEE MORE
SCHEMBL2249
4-[(1R)-2-amino-1-hydroxyethyl]benzene-1,2-diol hydrochloride
SEE MORE
SCHEMBL2250
4-(2-amino-1-hydroxyethyl)-3-chlorobenzene-1,2-diol
SEE MORE
SCHEMBL2255
(1S,5R,13R,17R)-10-hydroxy-4-methyl-12-oxa-4-azapentacyclo[9.6.1.0^{1,13}.0^{5,17}.0^{7,1
SEE MORE
SCHEMBL2609
4-[(1R)-2-amino-1-hydroxyethyl]benzene-1,2-diol
SEE MORE
SCHEMBL2610
4-(2-amino-1-hydroxyethyl)benzene-1,2-diol
SEE MORE
SCHEMBL2737
(1S,5R,13R,17S)-17-hydroxy-10-methoxy-4-methyl-12-oxa-4-azapentacyclo[9.6.1.0^{1,13}.0^{5,17}.0^{7,1
SEE MORE
SCHEMBL2987
(1S,5R,13R,17R)-10-methoxy-4-methyl-12-oxa-4-azapentacyclo[9.6.1.0^{1,13}.0^{5,17}.0^{7,1
SEE MORE

UNICHEM



2022



UniChem

Data sources

UniChem currently contains data from the sources listed below.

Search

Source ID	Name	Description	Compounds Count ↓		
22	PubChem Compounds	A database of normalized PubChem compounds (CIDs) from the PubChem Database.	110283434	^	•••
Process of Data Acquisition: Standard InChIs and Keys provided on ftp site. **Last updated:** 2021-09-07					
23	Mcule	An online drug discovery platform with virtual screening and molecular modelling services.	32919693	v	•••
15	SureChEMBL	SureChEMBL automatically extracts chemistry from the full text of all major patent authorities. Compounds are derived from either chemical names found in text or in chemical depictions. All SureChEMBL compounds are included, except those failing UniChem loading rules.	22690638	v	•••
9	ZINC	A free database of commercially-available compounds for virtual screening, provided by the Shoichet Laboratory in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF). [Irwin and Shoichet, J. Chem. Inf. Model. 2005;45(1):177-82]	16886865	v	•••
10	eMolecules	A free chemical structure search engine containing millions of public domain structures. Pricing, availabilities, and vendor information requires an eMolecules Plus subscription.	5168336	v	•••
21	PubChem ('Thomson Pharma' subset)	A subset of the PubChem DB: from the original depositor 'Thomson Pharma'.	3858588	v	•••
29	Nikkaji	Nakkaji (The Japan Chemical Substance Dictionary) is an organic compound dictionary database prepared by the Japan Science and Technology Agency (JST).	3439411	v	•••
1	ChEMBL	A database of bioactive drug-like small molecules and bioactivities abstracted from the scientific literature.	2371556	v	•••
31	BindingDB	A public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of proteins considered to be drug-targets with small, drug-like molecules	1000223	v	•••
32	EPA (Environmental Protection Agency) CompTox Dashboard	The foundation of chemical safety testing relies on chemistry information such as high-quality chemical structures and physicochemical properties. This information is used by scientists to predict the potential health risks of chemicals.The CompTox Dashboard is part of a suite of dashboards developed by EPA to help evaluate the safety of chemicals. It provides access to a variety of data and information on over 700,000 chemicals currently in use and of interest to environmental researchers. Within the CompTox Dashboard, users can access chemical structures, experimental and predicted physicochemical and toxicity data, and additional links to relevant websites and applications. It maps curated physicochemical property data associated with chemical substances to their corresponding chemical structures	742310	v	•••

Rows per page: 10 1-10 of 41

Unichem

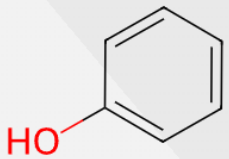
Compound Sources Search

Search by InChI

InChI=1S/C6H6O/c7-6-4-2-1-3-5-6/h1-5,7H

- (•) InChI
- ( ) InChIKey
- ( ) Source Compound ID
- ( ) UniChem Compound ID (UCI)

DRAW MOL SEARCH



Searched Compound

UCI: 31085

InChI: InChI=1S/C6H6O/c7-6-4-2-1-3-5-6/h1-5,7H

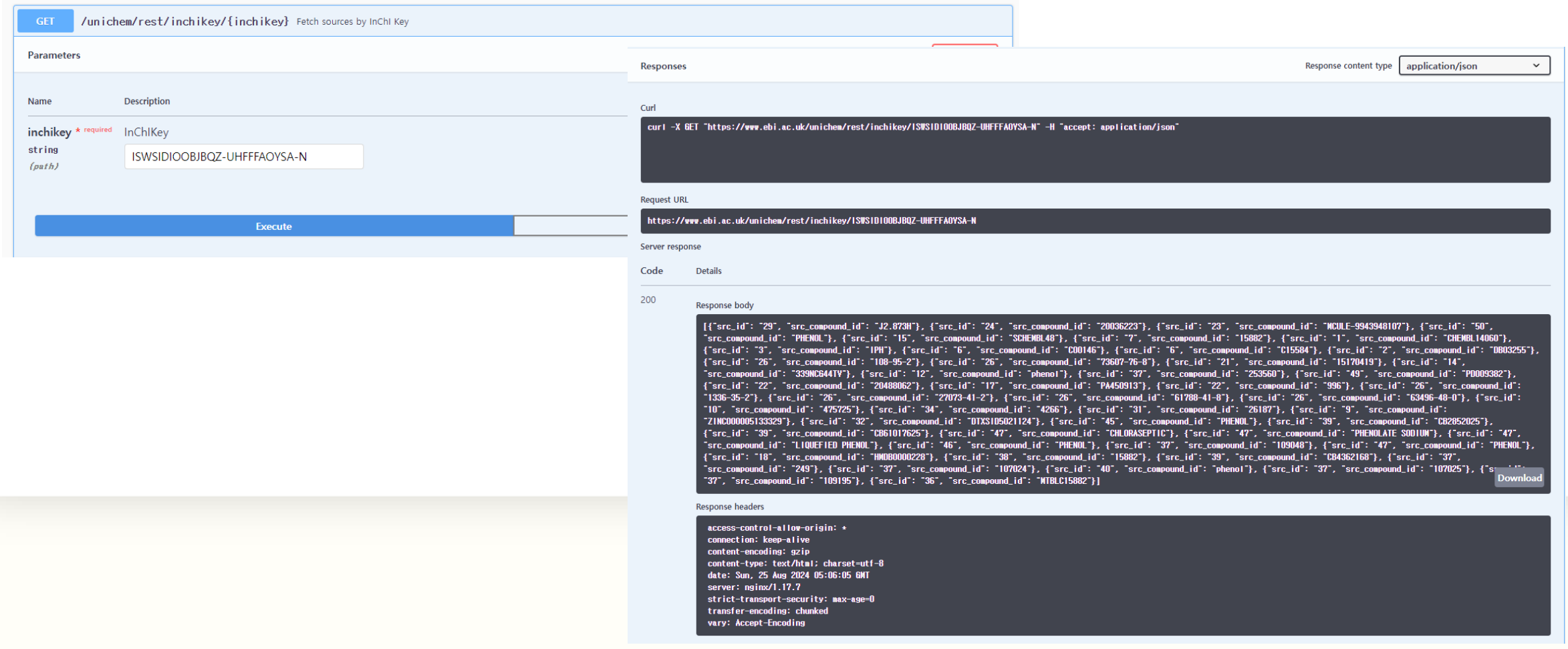
InChI Key: ISWSIDIOOBJBQZ-UHFFFAOYSA-N

Search

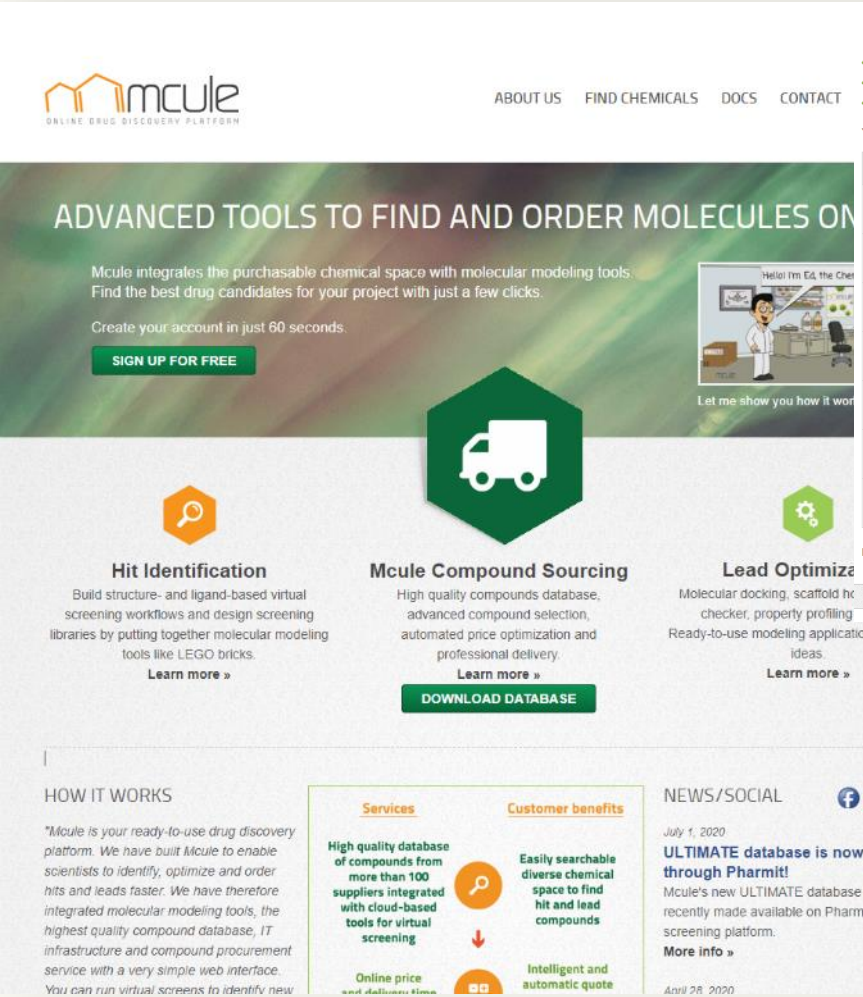
Source Compound ID	Source Name	Source ID ↑	
CHEMBL14060	ChEMBL	1	
DB03255	DrugBank	2	
IPH	PDBe (Protein Data Bank Europe)	3	
C00146	KEGG (Kyoto Encyclopedia of Genes and Genomes) Ligand	6	
C15584	KEGG (Kyoto Encyclopedia of Genes and Genomes) Ligand	6	
15882	ChEBI (Chemical Entities of Biological Interest).	7	
ZINC000005133329	ZINC	9	

UNICHEM API

- 검색 API URL : https://www.ebi.ac.uk/unichem/rest/inchikey/<<inchikey>>



Mcule (https://mcule.com)



2022

MCULE DATABASE

- Screen in-house and **purchase the best virtual hits** by just a few clicks
- Expand your in-house library with unique and diverse compounds
- High quality database and professional procurement service

You can freely download the following collections of the high quality Mcule database including 2D structures and Mcule IDs:

Subset name	Description	Last updated	Nb. of compounds	Filetype	Filesize	Download
Mcule Full	Purchasable Mcule supplier & ULTIMATE catalogs	Nov. 11, 2022	40,156,284	2D SDF (sdf.gz) SMILES (smi.gz)	9.1 GB 511.0 MB	Download Download
Mcule In Stock	Purchasable Mcule supplier in stock catalogs	Oct. 18, 2022	5,590,972	2D SDF (sdf.gz) SMILES (smi.gz)	1.3 GB 82.0 MB	Download Download
Mcule Building Blocks	Purchasable Mcule supplier building block catalogs	Oct. 19, 2022	3,452,192	2D SDF (sdf.gz) SMILES (smi.gz)	458.0 MB 42.0 MB	Download Download
Mcule Known Stock Amounts	Purchasable Mcule supplier in stock catalogs with known stock amount	Oct. 19, 2022	4,603,113	2D SDF (sdf.gz) SMILES (smi.gz)	1.1 GB 68.0 MB	Download Download
Mcule Virtual	Purchasable Mcule supplier virtual catalogs & ULTIMATE catalogs	Oct. 18, 2022	34,181,272	2D SDF (sdf.gz) SMILES (smi.gz)	7.8 GB 424.0 MB	Download Download
Mcule ULTIMATE Express	Purchasable Mcule ULTIMATE Express catalogs with fast delivery	Aug. 28, 2022	15,618,017	2D SDF (sdf.gz) SMILES (smi.gz)	3.0 GB 186.0 MB	Download Download
Mcule ULTIMATE Express 1	Purchasable Mcule ULTIMATE Express 1 catalog with fast delivery	Aug. 28, 2022	486,387	2D SDF (sdf.gz) SMILES (smi.gz)	65.0 MB 6.0 MB	Download Download

If you need the database in smaller chunks, please visit the documentation site where you can find detailed

Mcule database

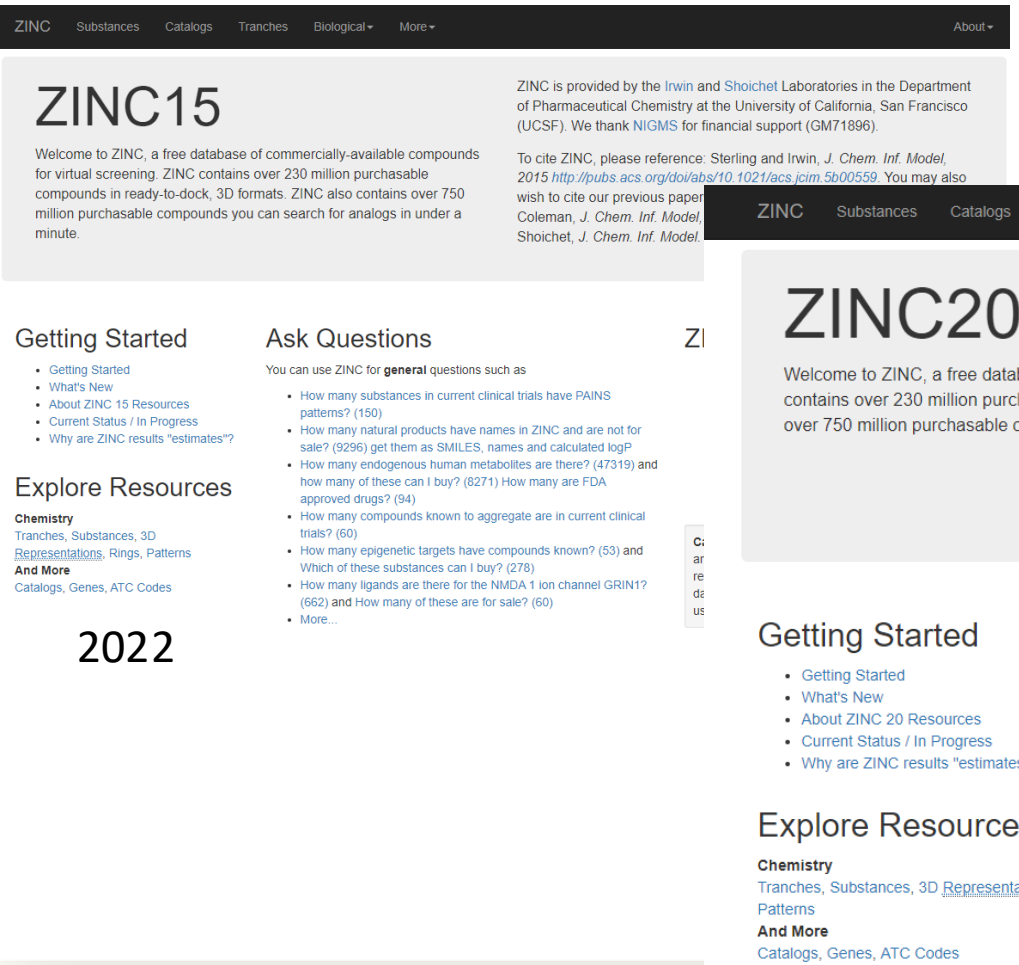
- Screen in-house and **purchase the best virtual hits** by just a few clicks
- Expand your in-house library with unique and diverse compounds
- High quality database and professional procurement service

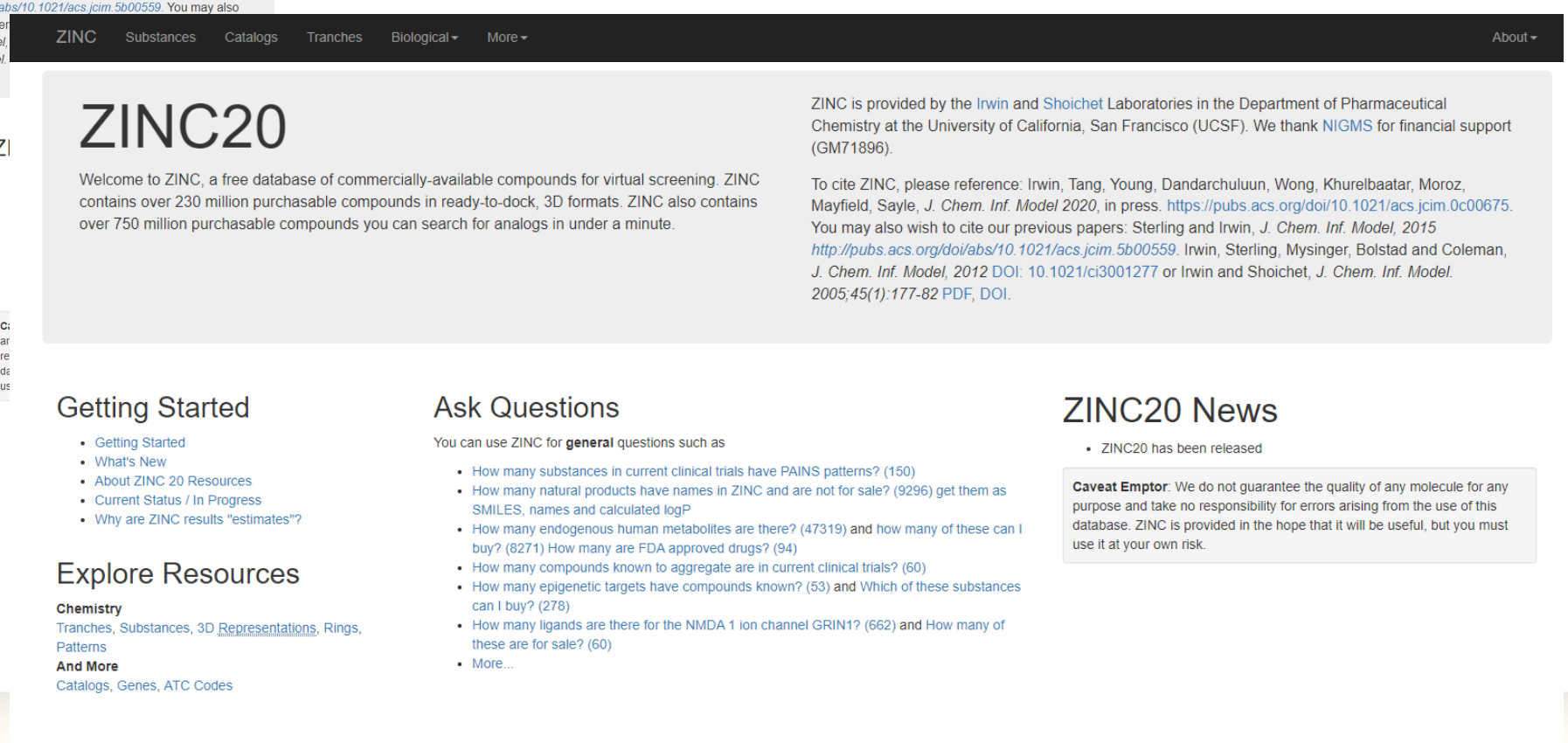
You can freely download the following collections of the high quality Mcule database including 2D structures and Mcule IDs:

Subset name	Description	Last updated	Nb. of compounds	Filetype	Filesize	Download
Mcule Full	Purchasable Mcule supplier & ULTIMATE catalogs	Aug. 23, 2024	42,169,343	2D SDF (sdf.gz) SMILES (smi.gz)	9.5 GB 537.0 MB	Download Download
Mcule In Stock	Purchasable Mcule supplier in stock catalogs	Aug. 24, 2024	5,921,591	2D SDF (sdf.gz) SMILES (smi.gz)	1.3 GB 88.0 MB	Download Download
Mcule Full Building Blocks	Purchasable Mcule supplier building block catalogs	Aug. 24, 2024	4,996,314	2D SDF (sdf.gz) SMILES (smi.gz)	694.0 MB 62.0 MB	Download Download
Mcule Known Stock Amounts	Purchasable Mcule supplier in stock catalogs with known stock amount	Aug. 24, 2024	4,981,804	2D SDF (sdf.gz) SMILES (smi.gz)	1.2 GB 75.0 MB	Download Download
Mcule Virtual	Purchasable Mcule supplier virtual catalogs & ULTIMATE catalogs	Aug. 24, 2024	36,073,649	2D SDF (sdf.gz) SMILES (smi.gz)	8.1 GB 447.0 MB	Download Download
Mcule ULTIMATE Express	Purchasable Mcule ULTIMATE Express catalogs with fast delivery	Aug. 28, 2022	15,618,017	2D SDF (sdf.gz) SMILES (smi.gz)	3.0 GB 186.0 MB	Download Download
Mcule ULTIMATE Express 1	Purchasable Mcule ULTIMATE Express 1 catalog with fast delivery	Aug. 28, 2022	486,387	2D SDF (sdf.gz) SMILES (smi.gz)	65.0 MB 6.0 MB	Download Download

If you need the database in smaller chunks, please visit the documentation site where you can find detailed instructions on how to create smaller files.

ZINC(https://zinc.docking.org/)





BindingDB(https://www.bindingdb.org/rwd/bind/index.jsp)