

# KSBi-BIML 2023

Bioinformatics & Machine Learning(BIML)  
Workshop for Life Scientists, Data Scientists,  
and Bioinformaticians

생물정보학 & 머신러닝 워크샵 (온라인)



## Chemoinformatics

이민호 \_ 동국대학교



본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 워크샵 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

# KSBi-BIML 2023

## Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists, Data Scientists, and Bioinformaticians

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크숍인 BIML-2023에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크숍은 2015년에 시작하여 올해로 9차를 맞이하게 되었습니다. 지난 2년간은 심각한 코로나 대유행으로 인해 아쉽게도 모든 강의가 온라인으로 진행되어 현장 강의에서만 가능한 강의자와 수강생 사이에 다양한 소통의 기회가 없음에 대한 아쉬움이 있었습니다. 다행히도 최근 사회적 거리두기 완화로 현장 강의를 가능해져 올해는 현장 강의를 재개함으로써 온라인과 현장 강의의 장점을 모두 갖춘 프로그램을 구성할 수 있게 되었습니다.

BIML 워크숍은 전통적으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되었습니다. 올해 AI 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 심층학습(Deep learning) 기법들에 대한 현장 강의를 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체 분석, 신약개발에 대한 이론과 실습 강의를 함께 제공할 예정입니다. 또한 싱글셀오믹스 분석과 메타유전체분석 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다. 이외에 다양한 생명정보학 분야에 대하여 30개 이상의 온라인 강좌가 개설되어 제공되며 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다. 특히 BIML은 각 분야 국내 최고 전문가들의 강의로 구성되어 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의를 될 것입니다.

이번 BIML-2023을 준비하기까지 너무나 많은 수고를 해주신 BIML-2023 운영위원회의 남진우, 우현구, 백대현, 정성원, 정인경, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 연사분께 깊은 감사를 드립니다.

2023년 2월

한국생명정보학회장 이 인 석

# Chemoinformatics

본 강의에서는 생물학이나 생물정보학 전공자들이 약물이나 소분자 정보의 활용 과정에 필요한 기초 이론, 지식 및 관련 데이터베이스 정보를 전달하는 것을 목표로 한다. 화합물 데이터베이스 종류 및 DB 내에서 활용할 수 있는 정보의 가공 방법 등을 간단히 소개하며, 추후 기계학습 등에 활용할 수 있도록 분자 구조를 다차원 수치 벡터로 변환하는 기법 등을 다룬다.

강의는 다음의 내용을 포함한다:

- Representation of chemical compounds
- File formats in chemoinformatics
- Chemical databases
- Bioassay databases
- Numerical representation
- Molecular fingerprints
- Hadoop/Spark Programming

\* 강의 난이도: 초급

\* 강의: 이민호 교수 (동국대학교 생명과학과)

# Curriculum Vitae

**Speaker Name: Minhoo Lee, Ph.D.**



## ► Personal Info

Name Minhoo Lee  
Title Assistant professor  
Affiliation Dongguk University

## ► Contact Information

Address Department of Life Science, Dongguk University-Seoul,  
Ilсандong-gu, Goyang-si, Gyeonggi-do 10326,  
Republic of Korea  
Email MinhooLee@dgu.edu

---

## Research Interest

Precision medicine

## Educational Experience

2005 B.S. Dept. of BioSystems, KAIST  
2013 Ph.D. Dept. of Bio and Brain Engineering, KAIST

## Professional Experience

2013-2014 Post Doc, Information & Electronics Research Institute, KAIST  
2014-2016 Assistant professor, Dept. of Biological Sciences, Sangji University  
2016-2020 Research assistant professor, Catholic Precision Medicine Research Center,  
College of Medicine, Catholic University of Korea  
2020- Assistant Professor, Dept. of Life Science, Dongguk University

## Selected Publications (5 maximum)

1. Lee K. et al., Utilizing random Forest QSAR models with optimized parameters for target identification and its application to target-fishing server, BMC Bioinformatics, 2017
2. Lee M. et al., Genomic structures of dysplastic nodule and concurrent hepatocellular carcinoma, Human pathology, 2018
3. Lee M. et al., Whole-exome sequencing reveals differences between nail apparatus melanoma and acral melanoma, Journal of the American Academy of Dermatology, 2018
4. Lee M. et al., Circulating microRNA expression levels associated with Internet gaming disorder, Frontiers in psychiatry, 2018
5. Lee M. et al., A novel loci of the HR gene in Marie-Unna hereditary hypotrichosis using whole-exome sequencing, Indian Journal of Dermatology, Venereology, and Leprology, 2020

# KSBi-BIML

Chemoinformatics

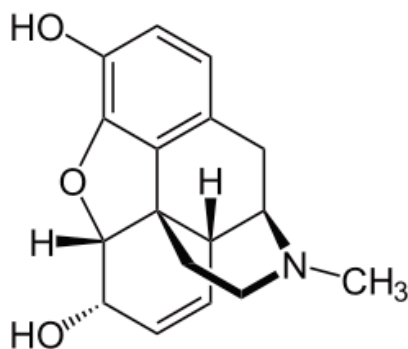
Minho Lee (MinhoLee@dgu.edu)

## Topics

- ▶ Representation of chemical compounds
  - ▶ File formats
- ▶ Chemical databases
  - ▶ Bioassay databases
- ▶ Numerical representation
  - ▶ Molecular fingerprints

## How to represent chemical information

- ▶ How can a molecular structure be stored on a computer?
  - ▶ Figure?
  - ▶ Coordinates?



## SMILES

- ▶ Simplified Molecular Input Line Entry System
  - ▶ Weininger, J Chem Inf Comput Sci, 1988, 28, 31
  - ▶ More recently, a community developed description: <http://opensmiles.org>
  - ▶ Linear format (“line notation”) that describes the connection table and stereochemistry of a molecule (i.e. 0D)
  - ▶ Convenient to enter as a query on-line, store in a database
- ▶ Basic guidelines:
  - ▶ Hydrogens are implicit
  - ▶ Parentheses indicate branches
  - ▶ Each atom is connected to the preceding atom to its left (excluding branches in-between)
  - ▶ Single bonds are implicit, = for double, # for triple

## SMILES examples

SMILES	Name	SMILES	Name
CC	ethane	[OH3+]	hydronium ion
O=C=O	carbon dioxide	[2H]O[2H]	deuterium oxide
C#N	hydrogen cyanide	[235U]	uranium-235
CCN(CC)CC	triethylamine	F/C=C/F	E-difluoroethene
CC(=O)O	acetic acid	F/C=C\F	Z-difluoroethene
C1CCCCC1	cyclohexane	N[C@@H](C)C(=O)O	L-alanine
c1ccccc1	benzene	N[C@H](C)C(=O)O	D-alanine

Reaction SMILES	Name
[I-].[Na+].C=CCBr>>[Na+].[Br-].C=CCI	displacement reaction
(C(=O)O).(OCC)>>(C(=O)OCC).(O)	intermolecular esterification

## Canonical SMILES

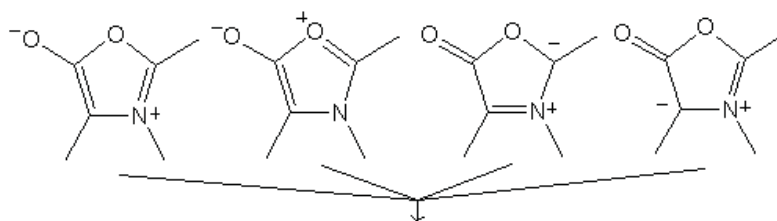
- ▶ In general, many different SMILES strings can be written for the same molecule
  - ▶ Not a unique identifier (one-to-many)
- ▶ Algorithms for producing “canonical SMILES” have been developed
  - ▶ The same unique SMILES string is always created for a particular molecule
  - ▶ One-to-one relationship between structure and representation
  - ▶ Note however, that different software implement different canonicalization algorithms



# InChI

## ▶ International Chemical Identifier

- ▶ Line notation developed by NIST and IUPAC
- ▶ Goal: An index for **uniquely** identifying a molecule
- ▶ Example



InChI=1/C6H9NO2/c1-4-6(8)9-5(2)7(4)3/h1-3H3

# InChI

## ▶ Features

- ▶ Derived from the structure
- ▶ One-to-one relationship between InChI and structure
- ▶ Layers (of specificity)
  - ▶ Can distinguish between stereoisomers, isotopes, or can leave out those layers
  - ▶ Different tautomeric forms give rise to the same InChI (unlike SMILES)

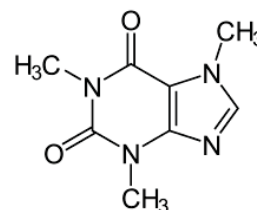
## ▶ InChIKey

- ▶ a fixed length (25 character) condensed digital representation of the InChI

## Example (Caffeine)

### ▶ SMILES

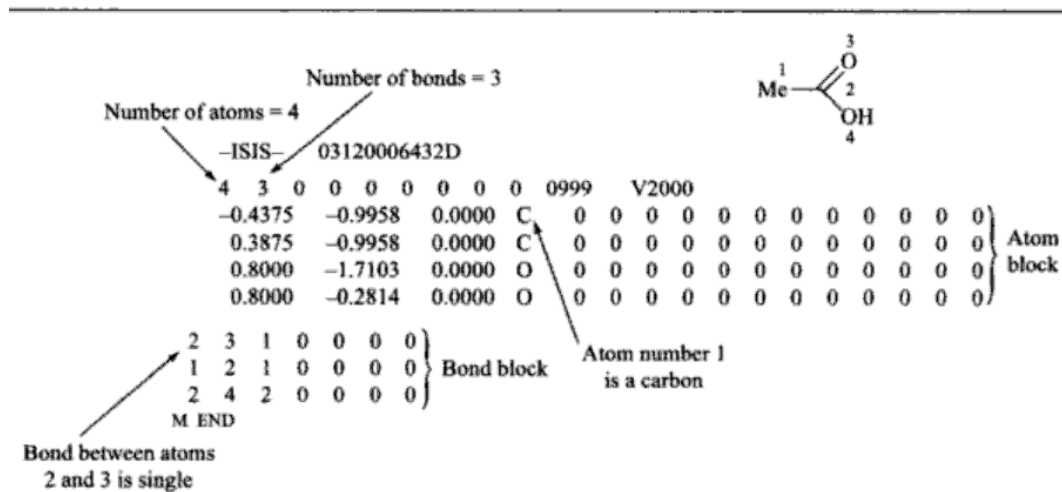
- ▶ [c]1([n+])([CH3])[c]([c]2([c]([n+]1[CH3])[n][cH][n+]2[CH3]))[O-])[O-]
- ▶ CN1C(=O)N(C)C(=O)C(N(C)C=N2)=C12
- ▶ Cn1cnc2n(C)c(=O)n(C)c(=O)c12
- ▶ Cn1cnc2c1c(=O)n(C)c(=O)n2C
- ▶ N1(C)C(=O)N(C)C2=C(C1=O)N(C)C=N2
- ▶ O=C1C2=C(N=CN2C)N(C(=O)N1C)C
- ▶ CN1C=NC2=C1C(=O)N(C)C(=O)N2C



### ▶ InChI

- ▶ InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

## Mol file



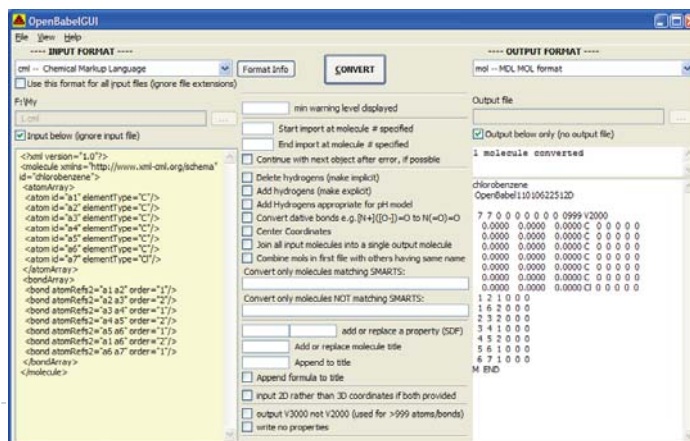
[2.3: MDL mol file for acetic acid, in the hydrogen-suppressed form.

## Other formats

- ▶ CML
- ▶ MOL2
- ▶ SDF
- ▶ PDB
- ▶ ...

## Open babel

- ▶ <http://openbabel.org>
- ▶ a chemical expert system mainly used for converting chemical file formats
- ▶ Offers CUI & GUI



# Chemical Databases

Database	Content	Size (no. of compounds)	URL
<b>Bioactivity data</b>			
ChEMBL	Bioactivity data from the medicinal chemistry literature	1 360 000	<a href="https://www.ebi.ac.uk/chembl/db">https://www.ebi.ac.uk/chembl/db</a>
PubChem	Biological screening results on small molecules	49 000 000	<a href="https://pubchem.ncbi.nlm.nih.gov/">https://pubchem.ncbi.nlm.nih.gov/</a>
<b>Patents</b>			
IBM	Chemicals from full text patents	2 500 000	<a href="http://www-935.ibm.com/services/us/gbs/bao/siip/">http://www-935.ibm.com/services/us/gbs/bao/siip/</a>
SureChEMBL	Chemicals from full text patents	12 400 000	<a href="https://www.surechembl.org">https://www.surechembl.org</a>
<b>Drugs</b>			
DRUGBANK	Drug data and drug target information	7700	<a href="http://www.drugbank.ca">http://www.drugbank.ca</a>
FDA/USP SRS	Substances present in FDA regulated products	34 000	<a href="http://fdasis.nlm.nih.gov/srs/srs.jsp">http://fdasis.nlm.nih.gov/srs/srs.jsp</a>
<b>Availability</b>			
ZINC	Commercially available compounds	22 700 000	<a href="http://zinc.docking.org">http://zinc.docking.org</a>
emolecules	Commercially available compounds	5 900 000	<a href="http://www.emolecules.com">http://www.emolecules.com</a>
<b>Other</b>			
ChEBI	Database and ontology of Chemical Entities of Biological Interest	27 000	<a href="https://www.ebi.ac.uk/chebi/">https://www.ebi.ac.uk/chebi/</a>
PDB	Data on biological macromolecular structures	16 000	<a href="https://www.ebi.ac.uk/pdbe/">https://www.ebi.ac.uk/pdbe/</a>

Note: All numbers from Apr 2014.

<http://dx.doi.org/10.1016/j.ddtec.2015.01.005>



# PubChem

▶ <https://pubchem.ncbi.nlm.nih.gov/>

Databases > Upload Services > Help more > Today's Statistics >



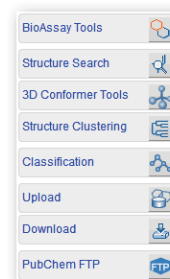
## PubChem

Try the PubChem Search Beta

**News** The PubChem Data Sources page is now updated. It helps you to find who provided what information in PubChem. [Read more...](#)

**News** PubChem Widgets 2.0f is released. It substantially updates all table-based widgets and classification widgets, while adding new capabilities and features. [Read more...](#)

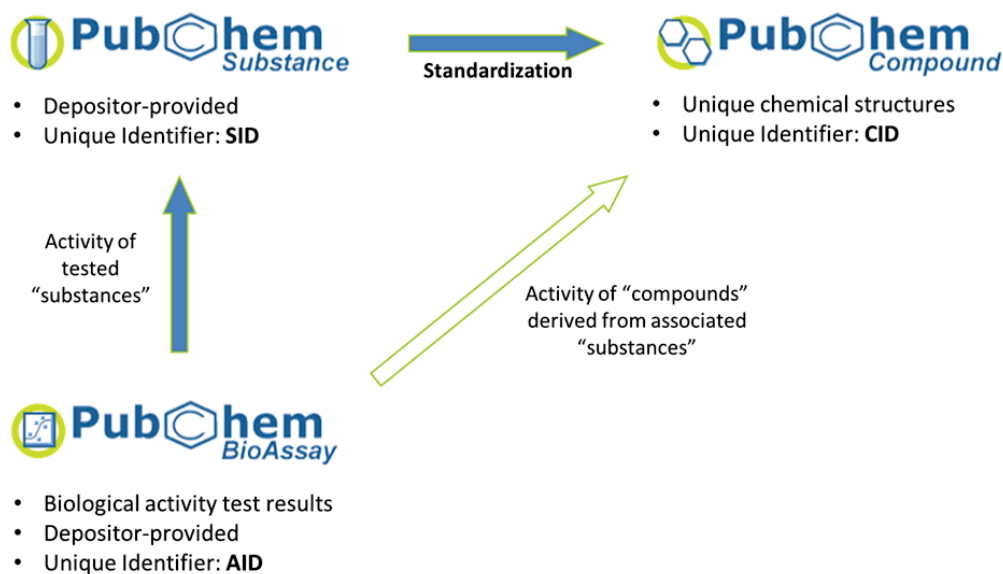
[more...](#)



Write to Helpdesk | Disclaimer | Privacy Statement | Accessibility | Data Citation Guidelines  
National Center for Biotechnology Information  
NLM | NIH | HHS



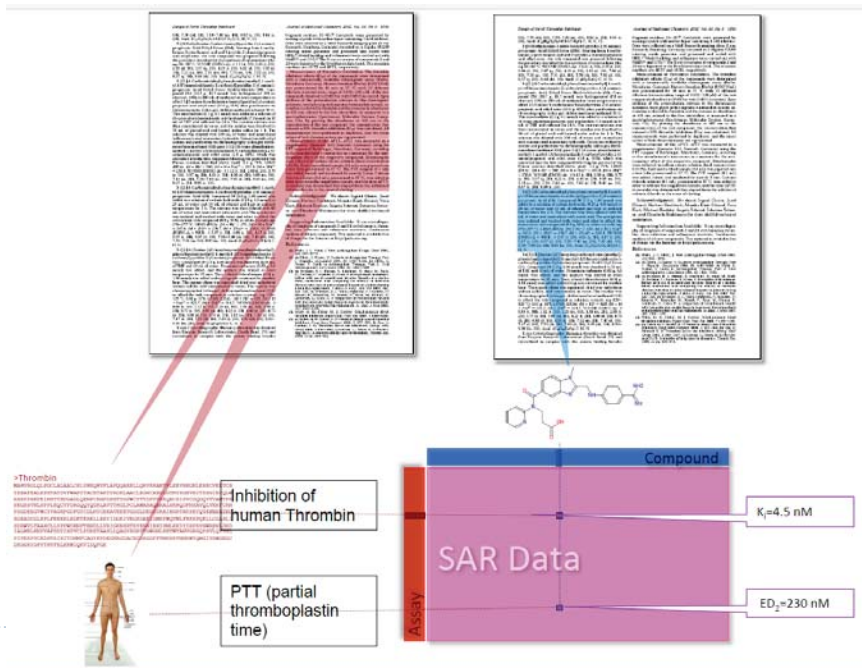
## PubChem is organized in 3 separate databases



## ChEMBL

- ▶ Open access database for drug discovery
- ▶ Freely available (searchable and downloadable)
- ▶ Content:
  - ▶ 2D structures & calculated properties (logP, MW, Lipinski, etc.)
  - ▶ Associated bioactivity data extracted from the primary medicinal chemistry journals such as J. Med. Chem.
  - ▶ Deposited data from neglected disease screening (e.g. malaria)
  - ▶ Subset of data from PubChem
- ▶ Covers ~30 years of compound synthesis and testing
- ▶ Annotated FDA-approved drugs

## ChEMBL Data generation

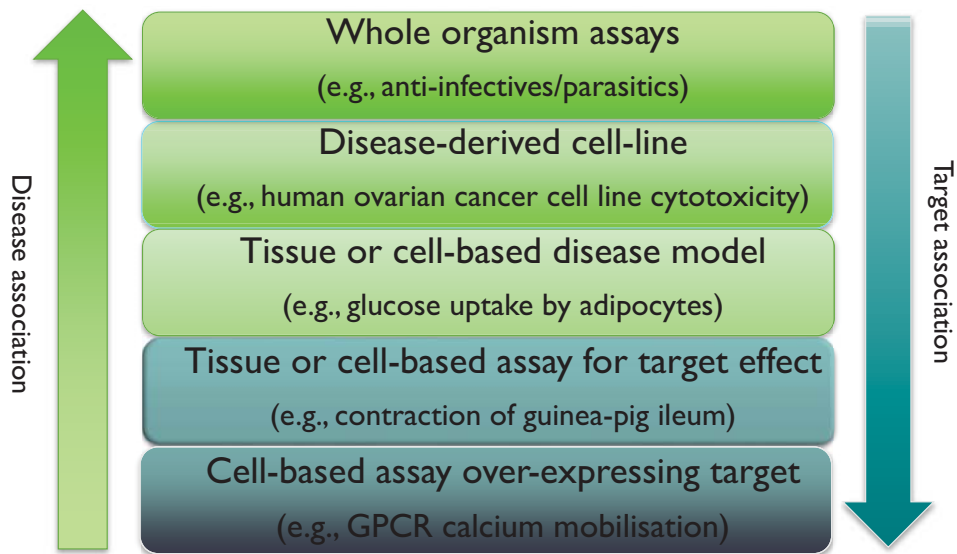


## ChEMBL Assays – Binding, Functional, ADMET

### ▶ Binding Assays

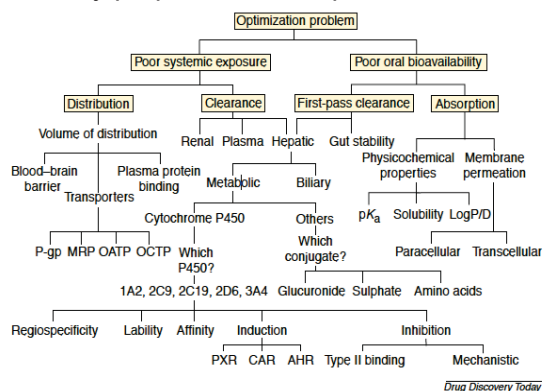
- ▶ Assays which directly measure the binding of a compound to a particular target
  - E.g., competition binding assays with a radioligand
- ▶ Various endpoints measured, but most commonly reported are:
  - ▶ IC<sub>50</sub> (half maximal inhibitory concentration)
  - ▶ K<sub>i</sub> (binding affinity)
  - ▶ MIC (minimum inhibitory concentration)
  - ▶ % Inhibition (of activity)

# Functional Assays



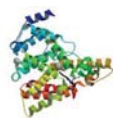
# ADMET Assays

- ▶ Assays measuring: Absorption, Distribution, Metabolism, Excretion, Toxicity properties of compounds



## ChEMBL Targets

### Protein



e.g., PDE5

### Protein complex



e.g., Nicotinic acetylcholine receptor

### Protein family



e.g., Muscarinic receptors

### Nucleic Acid



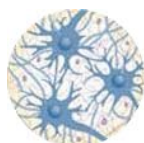
e.g., DNA

### Cell Line



e.g., HEK293 cells

### Tissue



e.g., Nervous

### Sub-cellular Fraction



e.g., Mitochondria

### Organism



e.g., Drosophila

## Protein Targets

- ▶ Each protein target linked to a sequence in UniProt
- ▶ Information from UniProt used in ChEMBL to allow searching:
  - ▶ Protein name/description
  - ▶ Synonyms and gene names
  - ▶ Organism (and NCBI Tax ID)
- ▶ Proteins in ChEMBL also classified according to family (e.g., Receptor, Kinase, Protease, Transporter etc).
  - ▶ Used for searching by target tree (Browse Targets)



# DrugBank example: Acetylsalicylic acid

Aspirin Targets (18) Enzymes (4) Transporters (3)

**IDENTIFICATION**

<b>Name</b>	Acetylsalicylic acid Commonly known or available as Aspirin	<b>Accession Number</b>	DB00945
-------------	--	-------------------------	---------

**Description**

Also known as *Aspirin*, acetylsalicylic acid (ASA) is a commonly used drug for the treatment of pain and fever due to various causes. Acetylsalicylic acid has both anti-inflammatory and antipyretic effects. This drug also inhibits platelet aggregation and is used in the prevention of blood clots stroke, and myocardial infarction (MI) <sup>Label</sup>.

Interestingly, the results of various studies have demonstrated that long-term use of acetylsalicylic acid may decrease the risk of various cancers, including colorectal, esophageal, breast, lung, prostate, liver and skin cancer <sup>15</sup>. Aspirin is classified as a *non-selective cyclooxygenase (COX) inhibitor* <sup>11,14</sup> and is available in many doses and forms, including chewable tablets, suspension, extended-release formulations, and others <sup>19</sup>.

# DrugBank: ASA targets

**TARGETS**

1. Prostaglandin G/H synthase 1 ... Binding Properties [Details](#)

<b>Kind</b>	Protein	<b>General Function</b>	Prostaglandin-endoperoxide synthase activity
<b>Organism</b>	Humans	<b>Specific Function</b>	Converts arachidonate to prostaglandin H2 (PGH2), a committed step in prostanoid synthesis. Involved in the constitutive production of prostanoids in particular in the stomach and platelets. In gas...
<b>Pharmacological action</b>	Yes	<b>Gene Name</b>	PTGS1
<b>Actions</b>	Inhibitor	<b>Uniprot ID</b>	P23219
		<b>Uniprot Name</b>	Prostaglandin G/H synthase 1
		<b>Molecular Weight</b>	68685.82 Da

**References**

1. Flipo RM: [Are the NSAIDs able to compromising the cardio-preventive efficacy of aspirin?]. Presse Med. 2006 Sep;35(9 Spec No 1):1553-60. [PubMed:17078596]
2. Schwartz KA: Aspirin resistance: a review of diagnostic methodology, mechanisms, and clinical utility. Adv Clin Chem. 2006;42:81-110. [PubMed:17131625]
3. Birnbaum Y, Ye Y, Lin Y, Freeberg SY, Huang MH, Perez-Polo JR, Uretsky BF: Aspirin augments 15-epi-lipoxin A4 production by lipopolysaccharide, but blocks the pioglitazone and atorvastatin induction of 15-epi-lipoxin A4 in the rat heart. Prostaglandins Other Lipid Mediat. 2007 Feb;83(1-2):89-98. Epub 2006 Nov 7. [PubMed:17259075]
4. Guthikonda S, Lev EI, Patel R, DeLao T, Bergeron AL, Dong JF, Kleiman NS: Reticulated platelets and uninhibited COX-1 and COX-2 decrease the antiplatelet effects of aspirin. J

# ChEBI

- ▶ <http://www.ebi.ac.uk/chebi>
- ▶ Chemical Entities of Biological Interest
- ▶ A freely available, manually curated chemistry database
- ▶ High quality, manually annotated
- ▶ Provides chemical **ontology**

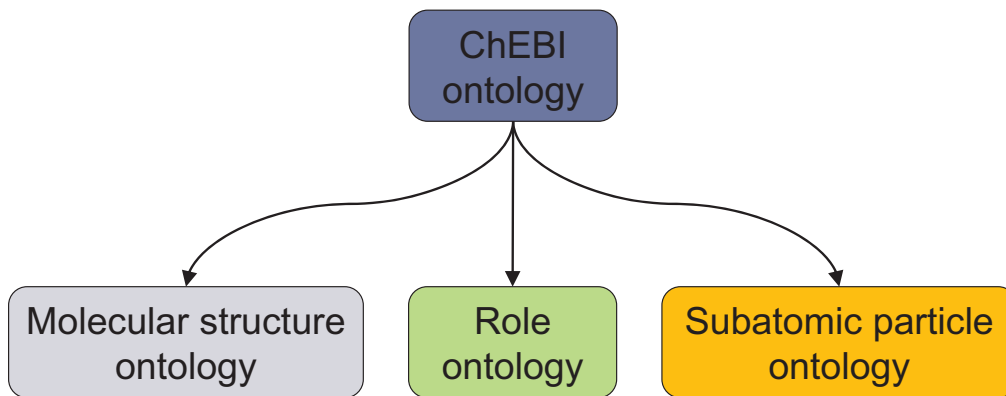
The screenshot displays the ChEBI website interface. At the top, there is a blue header with the ChEBI logo and navigation links: Home, Advanced Search, Browse, Documentation, Download, Tools, About ChEBI, Contact us, Preferences, and Submit. Below the header, a search bar is highlighted with a red box, containing the text "Search for ★★ only" and "All in ChEBI". Below the search bar, there is an example search term: "Example: iron, inChI=1SH2O/h1H2, water".

The main content area is divided into several sections:

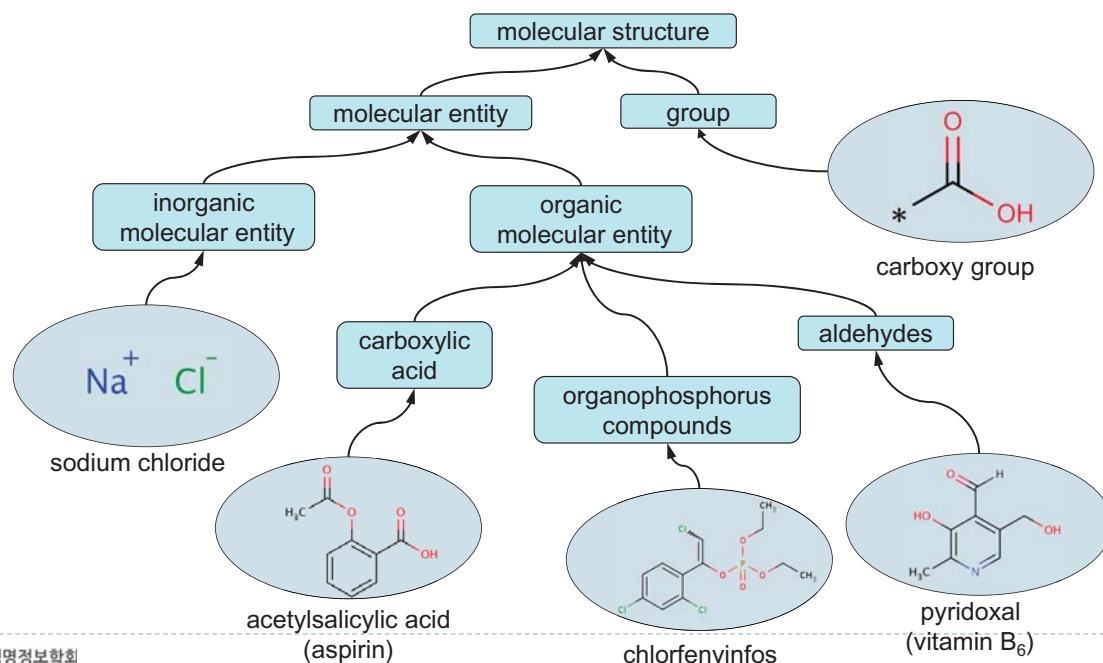
- Documentation:** Tutorial An introduction to the ChEBI database and ontology, showing users how to search and browse the web/programmatic interface. Statistics Graphs showing the growth of ChEBI, the numbers of curated, submitted and unchecked entries, the numbers of links to other resources and the sources of the information present in ChEBI. User Manual Learn more about the data fields in the ChEBI and data sources for ChEBI. [Read more...](#)
- Downloads:** SDF files ChEBI provides its chemical structures and additional data in structure-data file (SDF) format. [Read more...](#)
- Ontology files:** ChEBI ontology is provided in the W3C standard Web Ontology Language (OWL) and OBO formats. [Read more...](#)
- Database files:** ChEBI is stored in a relational database and we currently provide the ChEBI tables in flat-file tab delimited format, as an Oracle binary dumps and a generic SQL dumps for MySQL and PostgreSQL database. [Read more...](#)
- News:** Tweets by @chebi. A tweet from ChEBI (@chebi) is shown, dated 01 Mar, with the text: "Release 137 is live with 48020 fully annotated entities. See our entity of the month: WLL-vs. [bit.ly/1K1kOP](http://bit.ly/1K1kOP)". Below the tweet, there is a link to "Archived News...".
- Entity of the month:** 1st March 2016 WLL-vs. A chemical structure diagram of WLL-vs is shown, with the text "Entity of the month" and "1st March 2016 WLL-vs". Below the diagram, there is a link to "Read more ...".

At the bottom of the page, there is a footer with the SBI logo and the text "한국생명정보학회" and "Korean Society for Bioinformatics".

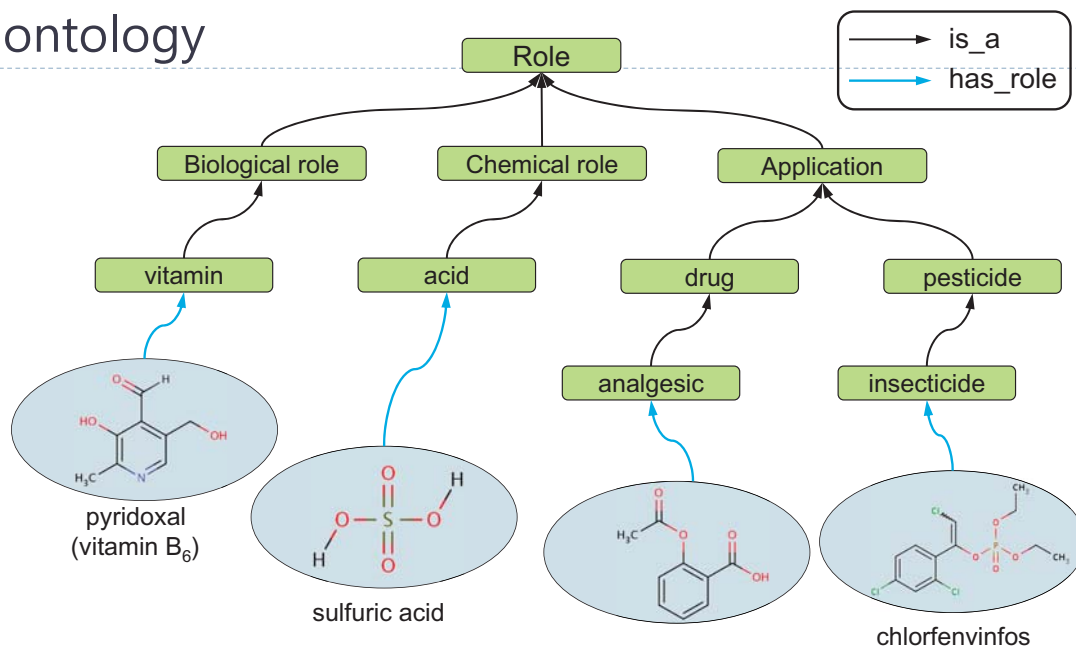
# The ChEBI ontology



# Molecular structure ontology



# Role ontology



# ChEBI ontology: aspirin

endoperoxides, precursors of prostaglandins.  
 non-narcotic analgesic  
 A drug that has principally analgesic, antipyretic and anti-inflammatory actions. Non-narcotic analgesics do not bind to opioid receptors.

View more via [ChEBI Ontology](#)

## ChEBI Ontology

- acetylsalicylic acid (CHEBI:15365) **has functional parent** salicylic acid (CHEBI:16914)
- acetylsalicylic acid (CHEBI:15365) **has role** anticoagulant (CHEBI:50249)
- acetylsalicylic acid (CHEBI:15365) **has role** antipyretic (CHEBI:35493)
- acetylsalicylic acid (CHEBI:15365) **has role** cyclooxygenase 1 inhibitor (CHEBI:50630)
- acetylsalicylic acid (CHEBI:15365) **has role** cyclooxygenase 2 inhibitor (CHEBI:50629)
- acetylsalicylic acid (CHEBI:15365) **has role** drug allergen (CHEBI:88188)
- acetylsalicylic acid (CHEBI:15365) **has role** EC 1.1.1.188 (prostaglandin-F synthase) inhibitor (CHEBI:77425)
- acetylsalicylic acid (CHEBI:15365) **has role** non-narcotic analgesic (CHEBI:35481)
- acetylsalicylic acid (CHEBI:15365) **has role** non-steroidal anti-inflammatory drug (CHEBI:35475)
- acetylsalicylic acid (CHEBI:15365) **has role** plant activator (CHEBI:73182)
- acetylsalicylic acid (CHEBI:15365) **has role** platelet aggregation inhibitor (CHEBI:50427)
- acetylsalicylic acid (CHEBI:15365) **has role** prostaglandin antagonist (CHEBI:49023)
- acetylsalicylic acid (CHEBI:15365) **has role** teratogenic agent (CHEBI:50905)
- acetylsalicylic acid (CHEBI:15365) **is a** acetate ester (CHEBI:47622)
- acetylsalicylic acid (CHEBI:15365) **is a** benzoic acids (CHEBI:22723)
- acetylsalicylic acid (CHEBI:15365) **is a** salicylates (CHEBI:26596)
- acetylsalicylic acid (CHEBI:15365) **is conjugate acid of** acetylsalicylate (CHEBI:13719)

Incoming acetylsalicylate (CHEBI:13719) **is conjugate base of** acetylsalicylic acid (CHEBI:15365)

## IUPAC Name

2-(acetyloxy)benzoic acid

## INNs

acide acétysalicylique

ácido acetilsalicílico

## Sources

ChemIDplus

NIST Chemistry WebBook

# ZINC

- ▶ <http://zinc.docking.org/>
- ▶ ZINC was originally designed for target based virtual screening (docking)
- ▶ also useful for many other things
  - ▶ finding a compound to purchase
  - ▶ downloading a library in SMILES format for ligand based virtual screening
  - ▶ find compounds by similarity to a starting
  - ▶ find compound ANNOTATED for a particular target (via ChEMBL)
  - ▶ find compounds PREDICTED for a particular target (via ChEMBL or docking)

## ZINC subsets

	Lead-Like	Fragment-Like	Drug-Like	All	Shards
<b>Standard</b> Size Updated	<b>Lead-Like</b> 6,053,287 2014-09-29	<b>Fragment-Like</b> 847,909 2015-02-04	<b>Drug-Like</b> 17,909,742 2014-11-24	<b>All Purchasable</b> 22,724,825 2014-11-28	<b>Shards</b> 625,159 2014-05-16
<b>Clean</b> Size Updated	<b>Clean Leads</b> 4,591,276 2014-09-25	<b>Clean Fragments</b> 1,611,889 2014-09-24	<b>Clean Drug-Like</b> 13,195,609 2013-11-05	<b>All Clean</b> 16,403,865 2013-12-18	<b>Clean Shards</b> 325,950 2014-11-24
<b>In Stock</b> Size Updated	<b>Leads Now</b> 3,687,621 2014-06-25	<b>Frag Now</b> 794,041 2015-02-04	<b>Drugs Now</b> 10,639,555 2014-11-24	<b>All Now</b> 12,782,590 2014-05-01	<b>Shards Now</b> 424,775 2014-09-24
<b>Boutique</b> Size Updated	<b>Boutique Leads</b> 5,114,169 2012-12-24	<b>Boutique Frags</b> 2,755,555 2013-11-08	<b>Boutique Drugs</b> 10,292,210 2012-11-27	<b>All Boutique</b> 12,217,845 2012-11-27	<b>Boutique Shards</b> 80,668 2013-11-08
Comments/Citation	Teague, Davis, Leeson, Oprea, <i>Angew Chem Int Ed Engl</i> , 1999, Dec 16;38(24):3743-3748.	Carr RA, Congreve M, Murray CW, Kees DC, <i>Drug Discov Today</i> , 2005 Jul 15;10(14):987	Lipinski, J. <i>Pharmacol Toxicol Methods</i> , 2000 Jul-Aug;44(1):235-49.	Purchasable chemical space	Type I binding sites
Filtering Criteria	p.mwt <= 350 and p.mwt >= 250 and p.xlogp <= 3.5 and p.rb <= 7	p.xlogp <= 3.5 and p.mwt <= 250 and p.rb <= 5	p.mwt <= 500 and p.mwt >= 150 and p.xlogp <= 5 and p.rb <= 7 and p.psa < 150 and p.n_h_donors <= 5 and p.n_h_acceptors <= 10		p.mwt < 190

# Protein Data Bank(PDB): structure database

## PDB file (text format)

```

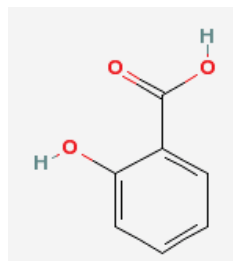
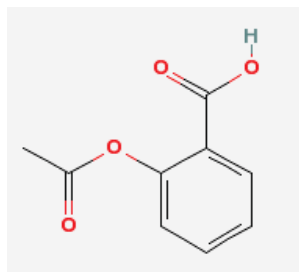
HEADER      TRANSFERASE                               17-JUN-02  1M17
TITLE      EPIDERMAL GROWTH FACTOR RECEPTOR TYROSINE KINASE DOMAIN
TITLE      2 WITH 4-ANILINOQUINAZOLINE INHIBITOR ERLOTINIB
COMPND     MOL_ID: 1;
COMPND     2 MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
COMPND     3 CHAIN: A;
COMPND     4 FRAGMENT: TYROSINE KINASE DOMAIN (RESIDUES 671-998);
COMPND     5 SYNONYM: RECEPTOR PROTEIN-TYROSINE KINASE ERBB-1;
COMPND     6 EC: 2.7.1.112;
COMPND     7 ENGINEERED: YES
SOURCE     MOL_ID: 1;
SOURCE     2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE     3 ORGANISM_COMMON: HUMAN;
SOURCE     4 GENE: EGFR;
SOURCE     5 EXPRESSION_SYSTEM: SPODOPTERA FRUGIPERDA;
SOURCE     6 EXPRESSION_SYSTEM_COMMON: FALL ARMYWORM;
SOURCE     7 EXPRESSION_SYSTEM_STRAIN: AUTOGRAPHICA
SOURCE     8 CALIFORNICA/T.NICOPLUSIA;
SOURCE     9 EXPRESSION_SYSTEM_CELL_LINE: SF9;
SOURCE     10 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE     11 EXPRESSION_SYSTEM_PLASMID: PVL1392
KEYWDS     TRANSFERASE, TYROSINE KINASE DOMAIN
EXPDTA     X-RAY DIFFRACTION
AUTHOR     J.STAMOS,M.X.SLIWKOWSKI,C.EIGENBROT
REVDAT    2 25-FEB-03 1M17 1 JRNL
REVDAT    1 04-SEP-02 1M17 0
JRNL       AUTH  J.STAMOS,M.X.SLIWKOWSKI,C.EIGENBROT
JRNL       TITL  STRUCTURE OF THE EPIDERMAL GROWTH FACTOR RECEPTOR
JRNL       TITL 2 KINASE DOMAIN ALONE AND IN COMPLEX WITH A
JRNL       TITL 3 4-ANILINOQUINAZOLINE INHIBITOR.
JRNL       REF   J.BIOL.CHEM. V. 277 46265 2002
JRNL       REFN  ASTM JBCHA3 US ISSN 0021-9258
REMARK     1
REMARK     2
REMARK     2 RESOLUTION. 2.60 ANGSTROMS.
    
```

## Molecular similarity

- ▶ Structurally similar molecules tend to have similar properties
- ▶ If we can measure similarity somehow
  - ▶ Can construct a distance matrix
    - ▶ Such matrices can be used to cluster compounds
  - ▶ Can use to find molecules in a database similar to a particular query
    - ▶ Can find unknown molecules with a similar property
  - ▶ Can use to see whether a particular property is correlated with molecular similarity

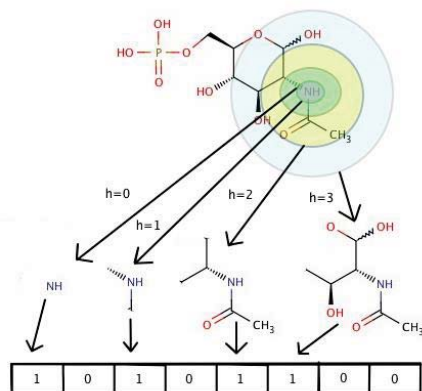
## ...But how to measure similarity?

- ▶ How similar are aspirin (A) and salicylic acid (B)?

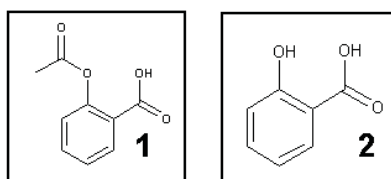


# Chemical fingerprint

- ▶ A molecular fingerprint is an encoding of the molecular structure onto a (long) binary string



# Tanimoto coefficient



<b>1</b>	<b>1</b>	<b>1</b>	<b>0</b>	<b>1</b>	<b>1</b>	<b>0</b>	<b>1</b>	<b>0</b>
<b>2</b>	<b>1</b>	<b>1</b>	<b>0</b>	<b>1</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>

A = Number of bits set in both = 3  
 B = Number of bits set in (1), but not in (2) = 2  
 C = Number of bits set in (2), but not in (1) = 0

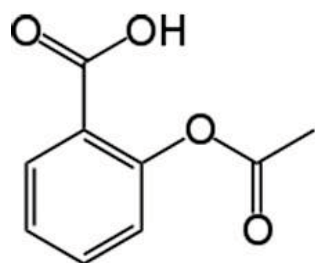
TANIMOTO COEFFICIENT =  $A / (A + B + C)$   
 =  $3 / (3 + 2 + 0) = 0.6$  or 60%



## Types of fingerprint

- ▶ PubChem
- ▶ Daylight
- ▶ Extended Connectivity Fingerprint (ECFP)
- ▶ ...

## PubChem fingerprint



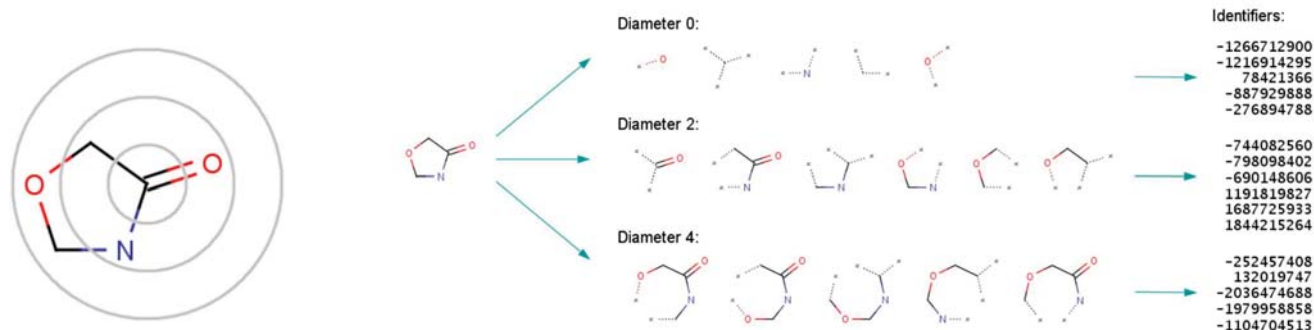
Aspirin Structure

Bit Position	Bit Substructure
0	$\geq 4$ H
1	$\geq 8$ H
2	$\geq 16$ H
3	$\geq 32$ H
4	$\geq 1$ Li
5	$\geq 2$ Li
6	$\geq 1$ B
7	$\geq 2$ B
8	$\geq 4$ B
...	...

Pubchem Molecular Fingerprint

# ECFP

## Extended Connectivity Fingerprint



### ECFP\_2, ECFP\_4, ECFP\_6, ... : depending on diameter

<https://docs.chemaxon.com/display/docs/Extended+Connectivity+Fingerprint+ECFP>

# ECFP

## Identifier list representation:

-1266712900 -1216914295 78421366 -887929888 -276894788 -744082560 -798098402 -690148606 1191819827  
1687725933 1844215264 -252457408 132019747 -2036474688 -1979958858 -1104704513

## Fixed-length binary representation:

0100000000100000110000100011000000000101000000000000000000000000000100101001000000000010000000000

Hash function

Bit collisions

### Bit string length

- ▶ Larger length decreases the likelihood of bit collision

## tools for calculating molecular fingerprints

---

- ▶ Chemistry Development Kit (CDK)

- ▶ JAVA

- ▶ <https://cdk.github.io/>

- ▶ RDKit

- ▶ C++, Python

- ▶ <https://www.rdkit.org/>

- ▶ R packages

- ▶ rcdk in CRAN

- ▶ Rcpk in bioconductor