

# KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML)  
Workshop for Life and Medical Scientists

생명정보학 & 머신러닝 워크숍 (오프라인)



## 인공지능 신약개발 AI Drug Design

김동섭 \_ KAIST



**KSBI**  
KOREAN SOCIETY FOR  
BIOINFORMATICS

| 한국생명정보학회



본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 워크샵 오프라인 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 2024

## Bioinformatics & Machine Learning(BIML) Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크숍인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크숍은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크숍은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의를 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의를 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구 중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의를 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 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 강사분들께 깊은 감사를 드립니다.

2024년 2월

한국생명정보학회장 이 인 석

# 강의 시간표

## DAY1 : 2월 24일 (토)

시간	강 의 (자연과학대학 28동 101호)
12:30-12:50	등록
12:50-13:00	공지사항 전달
13:00-14:30	<b>의료빅데이터/인공지능 총론</b> 김현성 교수(가톨릭대학교)
14:30-14:45	휴식
14:45-16:15	<b>의료영상 인공지능의 이해 및 의료영상 레이블링 실습</b> 백서연 교수(연석대학교)
16:15-16:30	휴식
16:30-18:00	<b>의료 정보처리 자동화 실습 / 독자적인 어플리케이션 만들기</b> 김선근 대표(원탁 주식회사), 서사도 조교

시간	강 의 (자연과학대학 28동 102호)
12:30-12:50	등록
12:50-13:00	공지사항 전달
13:00-14:20	<b>EMR 데이터를 활용한 머신러닝 기반 예후예측: Decision Tree-based Models + EMR 샘플 데이터 실습 (MIMIC sample dataset)</b> 고태훈 교수(가톨릭대학교)
14:20-14:40	휴식
14:40-16:00	<b>Chest X-ray 영상을 활용한 딥러닝 기반 폐질환 진단: Convolutional Neural Network + 의료영상 샘플 데이터 실습 (NIH Chest X-ray14)</b> 고태훈 교수(가톨릭대학교)
16:00-16:20	휴식
16:20-17:40	<b>심전도 데이터를 활용한 딥러닝 기반 부정맥 탐지: Recurrent Neural Network + Transformer + 심전도 샘플 데이터 실습 (MIT-BIH Arrhythmia Database)</b> 고태훈 교수(가톨릭대학교)

## DAY1 : 2월 26일 (월)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	<b>DNN (이론)</b> 이상근 교수(고려대학교)
10:50-11:00	휴식
11:00-12:10	<b>CNN (이론)</b> 이상근 교수(고려대학교)
12:10-13:40	점심
13:40-15:10	<b>RNN, ChatGPT, XAI (이론)</b> 이상근 교수(고려대학교)
15:10-15:20	휴식
15:20-16:50	<b>CNN/RNN 모델 구조 정의, 학습 알고리즘 적용, 성능 평가, 시각화 방법 (Tensorflow 실습)</b> 이정현 조교, 한성민 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	<b>Best practice for single-cell data analysis</b> 박종은 교수(KAIST)
11:00-11:10	휴식
11:10-12:40	<b>Practice1: Scanpy basic workflow</b> 정성민 조교, 고용준 조교
12:40-14:10	점심
14:10-15:30	<b>Public database, data integration, reference mapping, multiomics</b> 박종은 교수(KAIST)
15:30-15:40	휴식
15:40-16:50	<b>Practice2: Advanced single-cell analysis (siVI universe)</b> 정성민 조교, 고용준 조교

## DAY1 : 2월 27일 (화)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	<b>AI-based protein structure prediction</b> - Intro to protein structure prediction - Early AI-based approaches - AlphaFold and RoseTTAFold 백민경 교수(서울대학교)
10:50-11:00	휴식
11:00-12:10	<b>단백질 구조 예측 실습</b> - ColabFold를 활용한 단백질 구조 및 상호작용 예측 - Tips & Tricks for better structure modeling 백민경 교수(서울대학교)
12:10-13:40	점심
13:40-15:10	<b>AI-based protein design</b> - Intro to protein design - Protein backbone design using RFDiffusion - Protein sequence design using ProteinMPNN 백민경 교수(서울대학교)
15:10-15:20	휴식
15:20-16:50	<b>단백질 디자인 실습</b> - RFDiffusion 및 ProteinMPNN의 활용법 실습 백민경 교수(서울대학교)

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	<b>Introduction to Single-cell biology</b> 최정민 교수(고려대학교)
11:00-11:10	휴식
11:10-12:40	i. Unsupervised Spatial transcriptome analysis ii. Tumor Boundary Determination in Spatial Transcriptomics 유광민 조교, 이문영 조교
12:40-14:10	점심
14:10-15:30	i. Deconvolution Analysis Using Single-cell RNA Sequencing and Spatial Transcriptomics ii. Cell-Cell Interaction Analysis in Spatial Transcriptomics 김지현 조교, 최승지 조교
15:30-15:40	휴식
15:40-16:50	i. Open Chromatin Region Analysis and Biological Interpretation of Using scATAC-seq Dataset ii. Construction of Gene Regulatory Networks Based on Integrated Analysis of scATAC-seq and scRNA-seq Datasets 천하림 조교, 이호진 조교

## DAY1 : 2월 28일 (수)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	<b>Introduction to Transformers (이론)</b> 전민지 교수 (고려대학교)
11:00-11:10	휴식
11:10-12:40	<b>Introduction to Transformers (실습)</b> 봉현수 조교, 임우택 조교
12:40-14:10	점심
14:10-15:40	<b>Deep learning in Bioinformatics</b> 노미나 교수(한양대학교)
15:40-15:50	휴식
15:50-17:20	<b>Deep learning model을 이용한 실습</b> 박예솔 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	<b>마이크로바이옴 기본 이론</b> 이선재 교수(GIST)
10:50-11:00	휴식
11:00-12:10	<b>16S rRNA amplicon seq. - DADA2</b> 조준우 조교, 백재우 조교
12:10-13:40	점심
13:40-14:40	<b>최신 메타지놈 분석 기법의 현황</b> 이선재 교수(GIST)
14:40-14:50	휴식
14:50-16:50	<b>Shotgun metagenome 분석 (Linux)</b> 조준우 조교, 백재우 조교

## DAY1 : 2월 29일 (목)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	<b>화학정보학 기초(Cheminformatics) / 약물특성 및 약물다움(druglikeness)</b> <b>Molecular Notations &amp; Descriptors / AI 신약개발을 위한 Databases</b> <b>AI 신약개발을 위한 Programming 기초</b> 김동섭 교수(KAIST)
10:50-11:00	휴식
11:00-12:10	<b>Google Colab에 RDKit 설치 / 화합물 정보 읽기 실습</b> <b>Bioactivity database 검색 및 정보 읽기 실습</b> <b>Molecular descriptor (fingerprint) 생성 및 similarity 계산 실습</b> 정수재 조교, 나민주 조교
12:10-13:40	점심
13:40-15:10	<b>AI 신약개발을 위한 기계학습법 기초 / QSAR 모델링 기초 / AI 신약개발을 위한 딥러닝 모델</b> <b>Virtual screening (ligand-based, structure-based) 및 de novo design</b> 김동섭 교수(KAIST)
15:10-15:20	휴식
15:20-16:50	<b>QSAR modeling 전체 과정 실습 / 화합물의 Bioactivity 예측 모델 개발</b> <b>Virtual screening 과정을 통한 신약후보물질 발굴 실습</b> 정수재 조교, 나민주 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	<b>Single cell multiomics 이론 / Gene regulatory network 이론</b> 김준일 교수(숭실대학교)
11:00-11:10	휴식
11:10-12:40	<b>Seurat/Signac, ArchR, TENET+ 실습</b> 김현규 조교, 정희빈 조교
12:40-14:10	점심
14:10-15:40	<b>롱리드 시퀀싱 소개 및 유전체 조립 실습</b> 김준 교수(충남대학교)
15:40-15:50	휴식
15:50-17:20	<b>변이 분석 및 시각화 실습</b> 김준 교수(충남대학교)



## 인공지능 신약개발 AI Drug Design

신약개발에 소요되는 시간과 비용이 급속도로 증대됨에도 불구하고 신약 개발의 성공 사례는 그에 반해 날로 감소하고 있다. 이를 극복하기 위한 노력의 일환으로 다양한 종류의 인공지능 (AI) 신약개발 모델이 개발되고 있으며, 이 모델들을 활용하여 신약개발의 효율을 획기적으로 증대하고자 하는 노력들이 계속되고 있다. 이 강의에서는 이 과정에 필수적인 기초 지식인 화학정보학 (Cheminformatics) 및 기초 프로그래밍(RDKit)에 대해서 학습한 후, 인공지능 분야에서 널리 사용되는 다양한 모델들을 이용하여 신약개발에 사용되는 다양한 예측 모델 개발 방법에 대해 실습한다. 특히, 최근 그 중요성이 대두되고 있는 Deep learning 기술을 이용한 AI 신약개발 모델 개발에 대해 학습한다.

강의는 다음의 내용을 포함한다:

- 화학정보학 기초 (Introduction to cheminformatics)
- AI 신약개발을 위한 Databases
- AI 신약개발을 위한 Programming (RDKit)
- AI 신약개발을 위한 기계학습법 및 QSAR 모델링 기초
- AI 신약개발을 위한 딥러닝 모델

\* 참고 강의교재: 강의자료

\* 교육생 준비물: 노트북

\* 선수 지식: 기초 수준의 python programming

\* 강의 난이도: 초급

\* 강의: 김동섭 교수 (카이스트 바이오및뇌공학과)

# Curriculum Vitae

**Speaker Name: Dongsup Kim, Ph.D.**



## ► Personal Info

Name Dongsup Kim  
Title Professor  
Affiliation KAIST

## ► Contact Information

Address Department of Bio and Brain Engineering, KAIST, Daejeon  
Email kds@kaist.ac.kr  
Phone Number 042-350-4317

---

## Research Interest

Structural bioinformatics and computational drug development

## Educational Experience

1989 B.S., Seoul National University  
1991 M.S., Seoul National University  
1998 Ph.D., Brown University, USA

## Professional Experience

1998-2000 Post-doc research fellow, University of Pennsylvania  
2001-2002 Post-doc research fellow, Oak Ridge National Lab  
2003- Professor, Department of Bio and Brain Engineering, KAIST

## Selected Publications (5 maximum)

1. D. Yang, T. Chung, D. Kim, "DeepLUCIA: predicting tissue-specific chromatin loops using Deep Learning-based Universal Chromatin Interaction Annotator", *Bioinformatics*, 38:3501-3512 (2022)
2. H.Y. Kim, W. Jeon, D. Kim, "An enhanced variant effect predictor based on a deep generative model and the Born-Again Networks", *Scientific Reports*, 19127(2021)
3. H. Kim, D. Kim, "Prediction of mutation effects using a deep temporal convolutional network", *Bioinformatics*, 36:2047-2052 (2020)
4. A. Lee, D. Kim, "CRDS: Consensus Reverse Docking System for target fishing", *Bioinformatics*, 36:959-960 (2020)
5. W. Jeon, D. Kim, "FP2VEC: a new molecular featurizer for learning molecular properties", *Bioinformatics*, 35:4979-4985 (2019)

# KSBi-BIML 2024

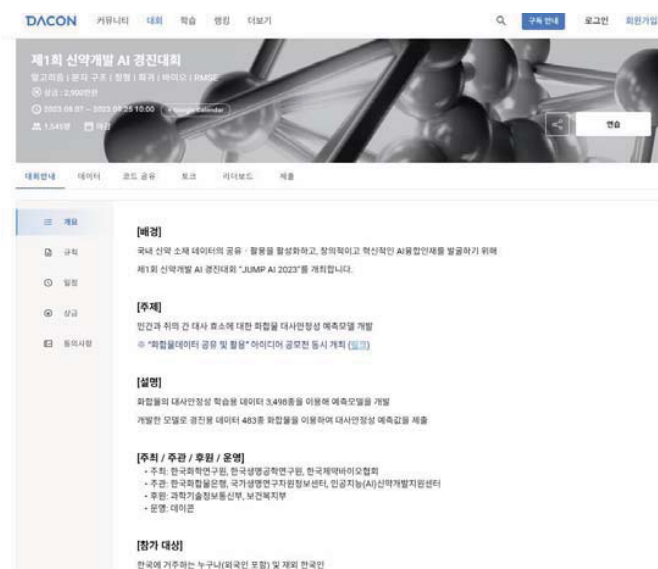
인공지능 신약설계  
AI Drug Design

## Google Classroom

- BiML: AI 신약개발
- <https://classroom.google.com/u/0/c/NjYxMDE1NTMyMTI0>
- 강의자료 및 실습용 코드 다운로드를 위해 모두 가입!

# After this lecture, you can win

- <https://dacon.io/competitions/official/236127/overview/description>



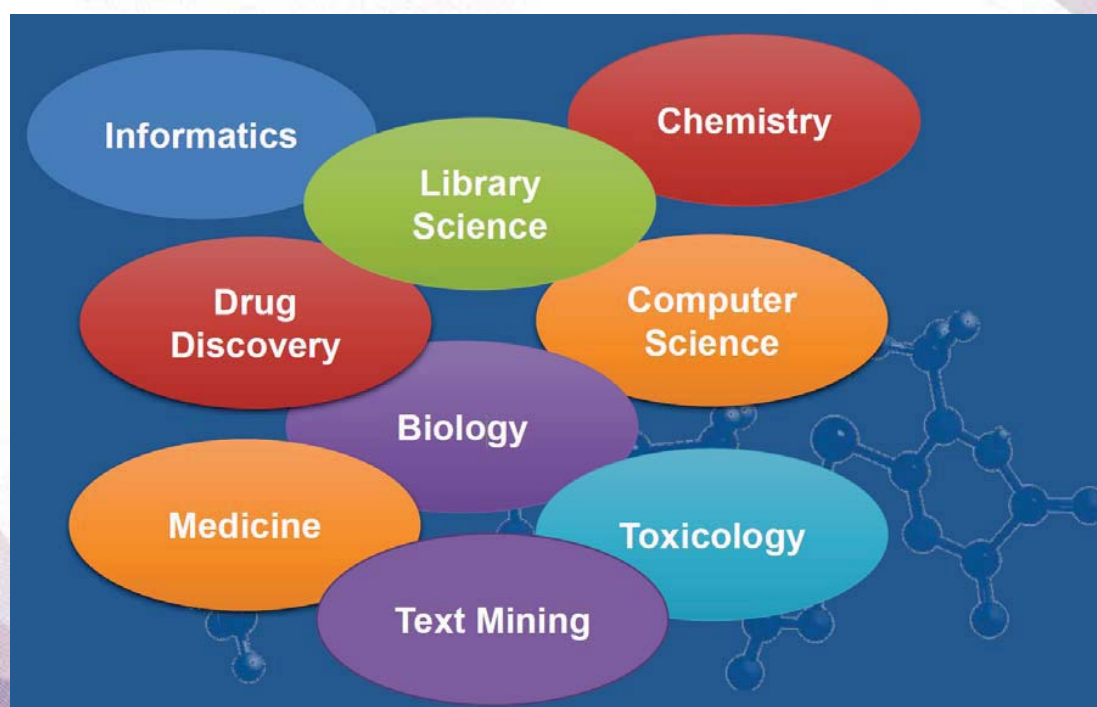
## 개요

- 강의
  - 화학정보학 기초(Cheminformatics)
  - 약물특성 및 약물다움(druglikeness)
  - Molecular Notations & Descriptors
  - AI 신약개발을 위한 Databases
  - AI 신약개발을 위한 Programming 기초
- 실습
  - Google Colab에 RDKit 설치
  - RDKit 실습: 화합물 정보 읽기 등
  - Bioactivity database 검색 및 정보 읽기 실습
  - Molecular descriptor (fingerprint) 생성 및 similarity 계산 실습

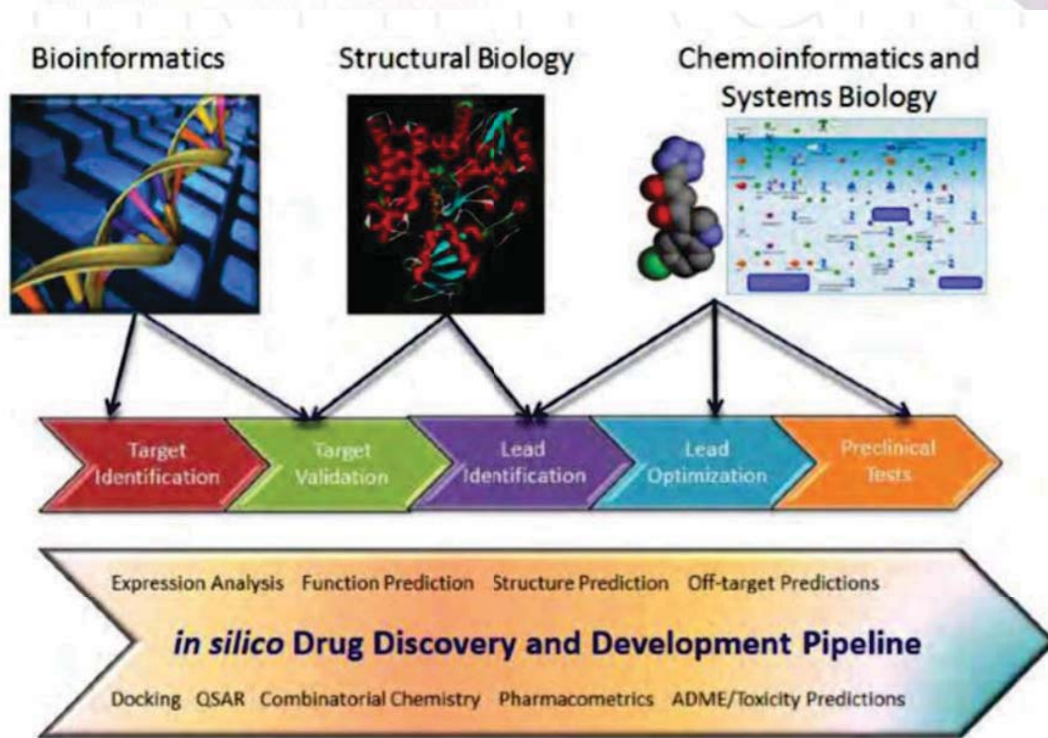
# 화학정보학이란

- Field of **information technology** that uses computers and computer programs to facilitate the collection, storage, analysis, and manipulation of large quantities of **chemical data**
- 여러 이름
  - Cheminformatics
  - Chemoinformatics
  - Chemical informatics
- Bioinformatics vs. Cheminformatics
  - Biological data: Bioinformatics
  - Chemical data: Cheminformatics
- 응용분야: 신약개발, 독성학, ...

## Interdisciplinary

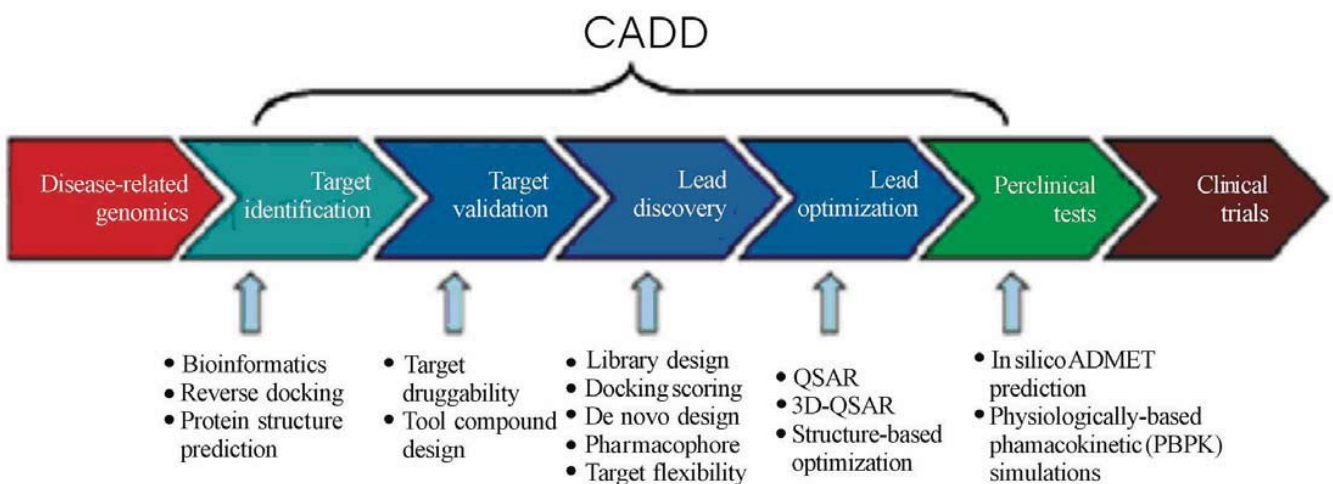


# 바이오통보학 vs 화학정보학



# 신약개발과 화학정보학

- Computer-Aided Drug Design (CADD)

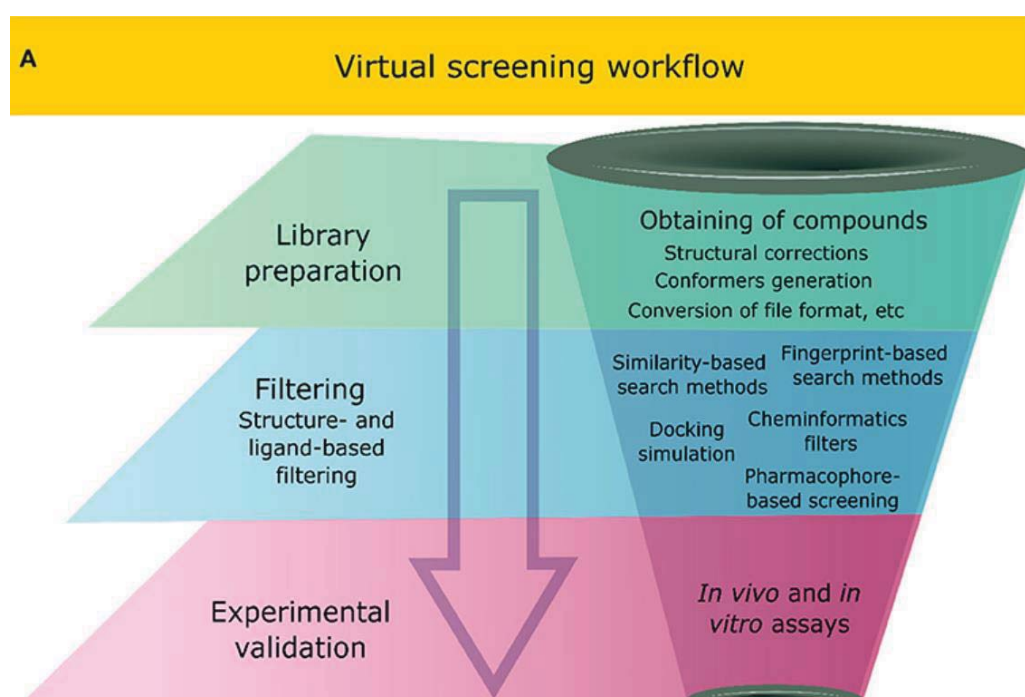




# Lead Discovery & Optimization

- Compound library design
- Virtual screening
- Docking
- Pharmacophore modeling
- QSAR (Quantitative Structure Activity Relationship)
- De novo design

## 가상 스크리닝

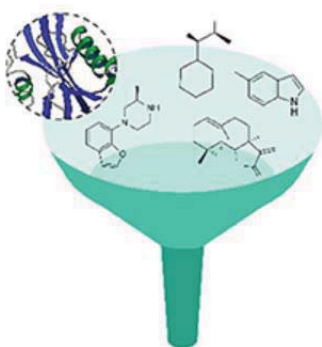


# 방법

B

## Structure-based virtual screening

- 1 Structure-based pharmacophore modeling
- 2 Molecular dynamics simulation
- 3 Molecular docking



## Ligand-based virtual screening

- 1 Ligand-based pharmacophore modeling
- 2 Machine learning algorithms
- 3 3D shape similarity search
- 4 Molecular fingerprints



## Molecular structures

- Linear notation
  - SMILES
  - InChI, InChIKey
- Connection table method
  - Molfile
  - SDF
  - MOL2

<https://www.ebi.ac.uk/chembl/dbcompound/inspect/CHEMBL413>

[http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=5284616&loc=ec\\_rcs](http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=5284616&loc=ec_rcs)

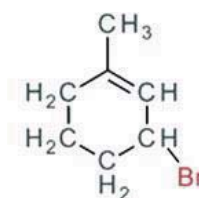


# Linear Notation

- Line notations represent structures as a linear string of alphanumeric symbols.
- Their compactness was an advantage in the early days of cheminformatics when storage space was at a premium.
- Even nowadays, it can be faster to enter a structure as a notation instead of using a chemical structure drawing program.

# SMILES

- Simplified Molecular Input Line Entry System
- A given chemical structure can have many valid and unambiguous representations (e.g., it is possible to start with any atom to derive a SMILES string).
- But for comparison purposes it is desirable to have a unique representation known as the 'canonical' one.
  - Morgan algorithm: iterative calculation of connectivity value of each atom
- <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>



CC1=CC(Br)CCC1

# Atoms

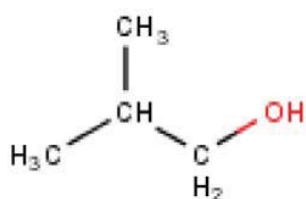
- Represented by their atomic symbols: C, N, O, and P
- The second letter of two-character atomic symbols must be in lower case: Cl (not CL), Br (not BR)
- Each non-hydrogen atom is enclosed in square brackets: [Au] or [Fe]
- Square brackets can be omitted for elements in the organic subset (B, C, N, O, P, S, F, Cl, Br, and I), if the proper number of “implicit” hydrogen atoms is assumed:  $\text{BH}_3 \rightarrow \text{B}$ ,  $\text{CH}_4 \rightarrow \text{C}$ ,  $\text{NH}_3 \rightarrow \text{N}$ ,  $\text{H}_2\text{O} \rightarrow \text{O}$

# Bonds

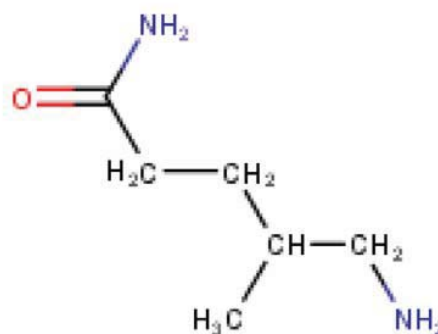
- Single bond  $\rightarrow$  “-” (can be omitted)
- Double bond  $\rightarrow$  “=”
- Triple bond  $\rightarrow$  “#”
- Aromatic bond  $\rightarrow$  “:” (can be omitted)
- Examples
  - $\text{CH}_4 \rightarrow \text{C}$
  - $\text{CH}_3\text{-CH}_3 \rightarrow \text{CC}$  (or C-C)
  - $\text{CH}_2=\text{CH}_2 \rightarrow \text{C=C}$
  - $\text{CH}\equiv\text{CH} \rightarrow \text{C\#C}$
  - $\text{CH}_3\text{OCH}_3 \rightarrow \text{COC}$
  - $\text{CH}_3\text{CH}_2\text{OH} \rightarrow \text{CCO}$
  - $\text{CH}_3\text{CH=O} \rightarrow \text{CC=O}$
  - $\text{HC}\equiv\text{N} \rightarrow \text{C\#N}$

# Branches

- Specified by enclosures in parentheses
- Can be nested or stacked



CC(C)CO

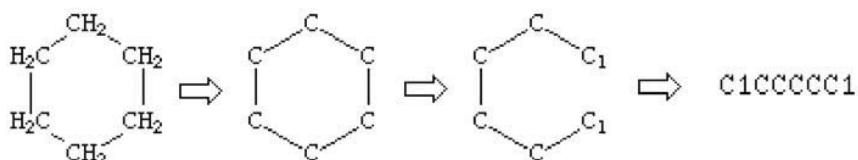


CC(CCC(=O)N)CN

# Rings

- Represented by breaking one single or aromatic bond in each ring, designating this ring-closure point with a digit

## Cyclohexane

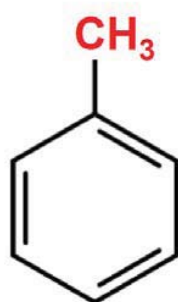


**Benzene → C1=C-C=C-C=C1 OR c1ccccc1**

**Note: Lower-case letters represent aromaticity.**

# Canonical SMILES

- Multiple SMILES representations exist for a given molecule.
- One “canonical” SMILES is selected among them: Morgan algorithm



**Cc1ccccc1**

c1(C)ccccc1

c1c(C)cccc1

c1cc(C)ccc1

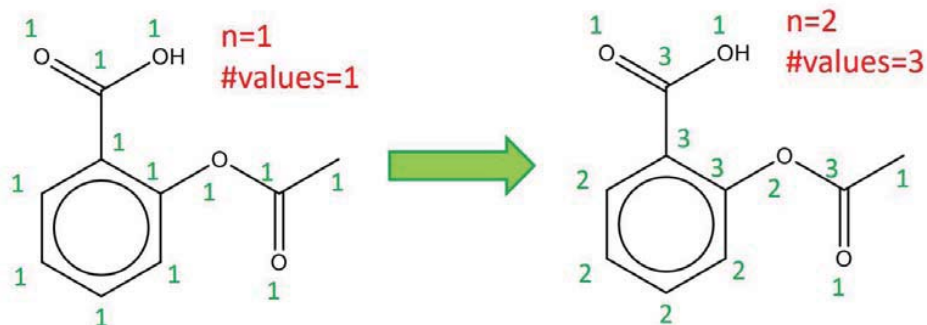
c1ccc(C)cc1

c1cccc(C)c1

c1ccccc1C

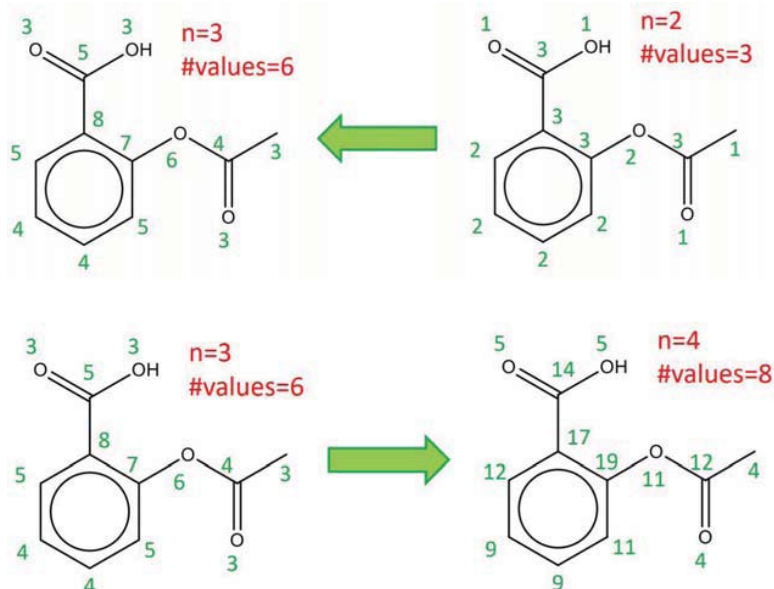
# Morgan Algorithm

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



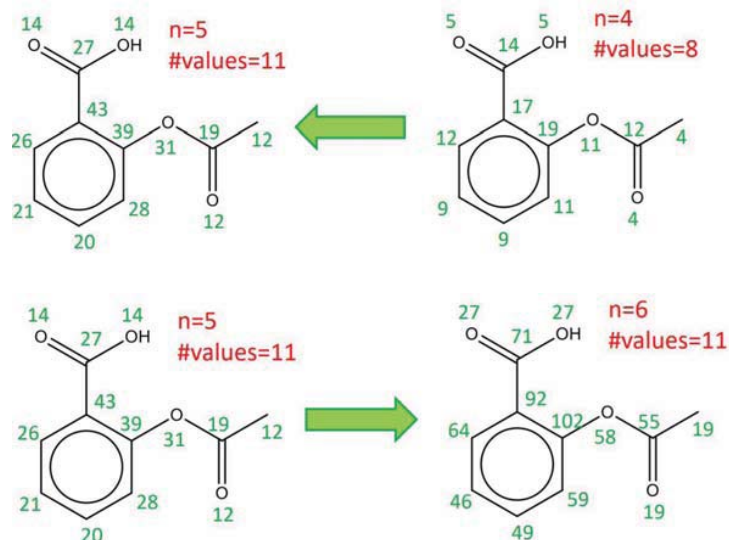
# Morgan Algorithm

- Repeat summing of neighboring values



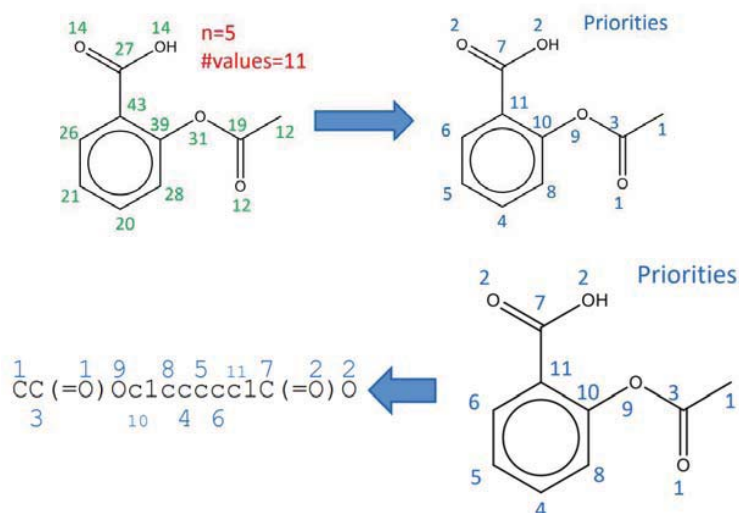
# Morgan Algorithm

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



# Morgan Algorithm

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



# Isomeric SMILES

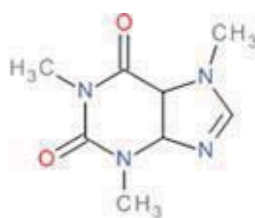
- Isotope: the integral atomic mass preceding the atomic symbol:  $^{13}\text{CH}_4 \rightarrow [13\text{CH}_4]$
- Stereochemistry
  - Atom stereo centers [(R/S)-configurations for a chiral center]
    - C[C@@H](C(=O)O)N L-Alanine
    - C[C@H](C(=O)O)N D-Alanine
  - Bond stereo centers [cis/trans-isomerism]
    - F/C=C/F or F\C=C\F (E)-1,2-difluoroethene (trans isomer)
    - F/C=C\F or F\C=C/F (Z)-1,2-difluoroethene (cis isomer)

# Limitation of SMILES

- Most SMILES encoders/decoders are proprietary.
  - Different groups implemented (slightly) different SMILES generation algorithms.
  - Not interchangeable between databases (or research groups) unless the same software is used.
- Doesn't have 2d and 3d coordinates retained, so need to changes to other formats like MOL, SDF, etc.
- Multiple smiles for one compound

# InChI

- International Chemical Identifier
- The goal of InChI is to provide a unique string representing a chemical substance of known structure.
- InChI is freely available and extensible.



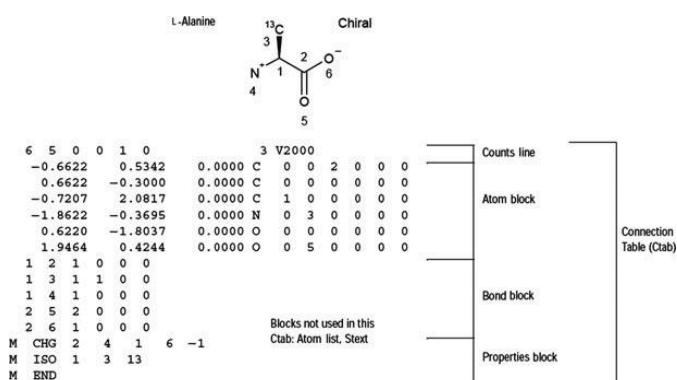
InChI = 1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3  
InChIKey = RYYVLZVUVI.JVGH-UHFFFAOYSA-N

# InChIKey

- The length of an InChI string increases with the size of the corresponding chemical structure.
- Not appropriate to use in internet search engines.
  - These search engines do not care case sensitivity nor special characters used in InChI.
- InChIKey was introduced for internet and database searching/indexing.
- A 27-character string derived from InChI, using a hashing algorithm.

# Connection Tables

- The MDL (now Symyx) connection table or CTfile, has become the *de facto* standard for exchange of datasets.
- It separates atoms and bonds into separate blocks.
- A molecule file, or 'molfile,' describes a single molecular structure that can contain disjoint fragments.
- A molfile consists of a header block and a connection table.
- Structure–data files (SDFfiles) contain structures and data for any number of molecules.







# Drug & Drug-likeness

- Drugs are an ill-defined entity from a chemical standpoint.
- Drug-like compound is defined as those compounds that have acceptable ADME/Tox properties to survive through the completion of human Phase 1 trials

## Lipinski's Rule-of-5

- The rule of five states that poor absorption or permeability are more likely when
  - cLogP (the calculated 1-octanol–water partition coefficient, a measure of lipophilicity) is  $>5$
  - molecular mass is  $>500$  Da
  - the number of hydrogen-bond donors (OH plus NH count) is  $>5$
  - the number of hydrogen-bond acceptors (O plus N atoms) is  $>10$
- Its conceptual simplicity and ease of calculation has made it the leading measure of drug-likeness.

# QED

- Quantitative Estimate of Drug-likeness

## ARTICLES

PUBLISHED ONLINE: 24 JANUARY 2012 | DOI: 10.1038/NCHEM.1243

nature  
chemistry

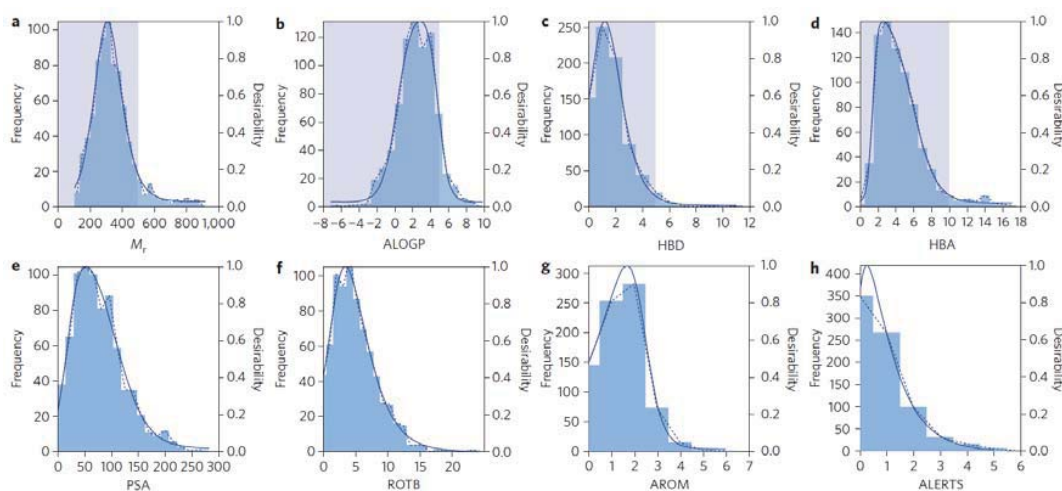
## Quantifying the chemical beauty of drugs

G. Richard Bickerton<sup>1</sup>, Gaia V. Paolini<sup>2</sup>, Jérémy Besnard<sup>1</sup>, Sorel Muresan<sup>3</sup> and Andrew L. Hopkins<sup>1\*</sup>

Drug-likeness is a key consideration when selecting compounds during the early stages of drug discovery. However, evaluation of drug-likeness in absolute terms does not reflect adequately the whole spectrum of compound quality. More worryingly, widely used rules may inadvertently foster undesirable molecular property inflation as they permit the encroachment of rule-compliant compounds towards their boundaries. We propose a measure of drug-likeness based on the concept of desirability called the quantitative estimate of drug-likeness (QED). The empirical rationale of QED reflects the underlying distribution of molecular properties. QED is intuitive, transparent, straightforward to implement in many practical settings and allows compounds to be ranked by their relative merit. We extended the utility of QED by applying it to the problem of molecular target druggability assessment by prioritizing a large set of published bioactive compounds. The measure may also capture the abstract notion of aesthetics in medicinal chemistry.

## Histograms of molecular properties

- Eight selected molecular properties for a set of 771 orally absorbed small molecule drugs



# Quantitative Estimate of Drug-likeness (QED)

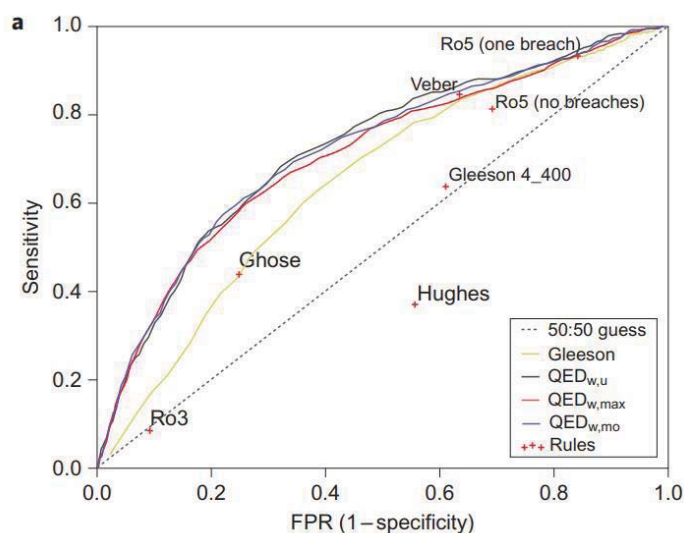
- Combining the individual desirability functions into the QED,

$$QED_w = \exp \left[ \frac{W_{MW} \ln d_{MW} + W_{ALOGP} \ln d_{ALOGP} + W_{HBA} \ln d_{HBA} + W_{HBD} \ln d_{HBD} + W_{PSA} \ln d_{PSA} + W_{ROTB} \ln d_{ROTB} + W_{AROM} \ln d_{AROM} + W_{ALERTS} \ln d_{ALERTS}}{W_{MW} + W_{ALOGP} + W_{HBA} + W_{HBD} + W_{PSA} + W_{ROTB} + W_{AROM} + W_{ALERTS}} \right]$$

$$d(x) = a + \frac{b}{\left[ 1 + \exp \left( -\frac{x - c + \frac{d}{2}}{e} \right) \right]} \left[ 1 - \frac{1}{\left[ 1 + \exp \left( -\frac{x - c - \frac{d}{2}}{f} \right) \right]} \right]$$

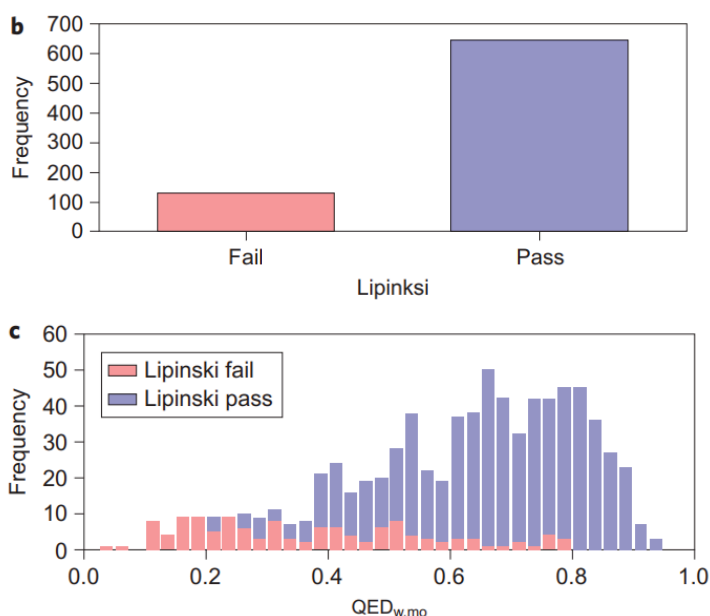
# Performance

- A receiver operating characteristic plot in classifying compounds as drug-like or otherwise



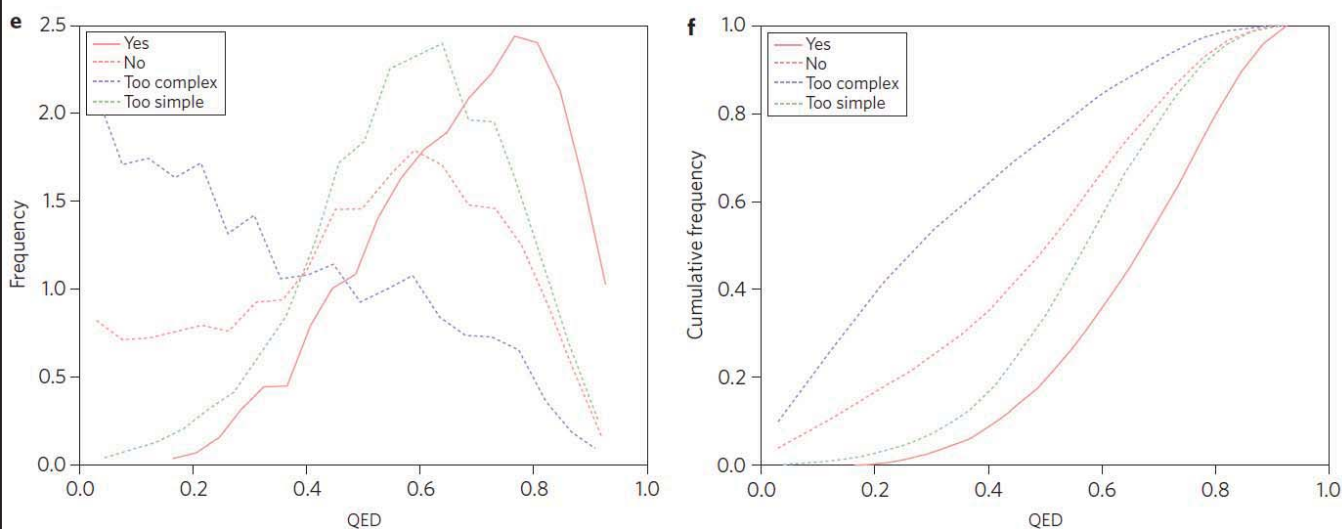
# Rule-of-5 Comparison

- Direct comparison of the Ro5 and QED shows the drugs failing (red) and passing (blue) Lipinski's Ro5



# Chemical aesthetics

- Question: “Would you undertake chemistry on this compound if it were a hit?”



# Synthetic Accessibility Score (SAS)

- Ertl et al., "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions", J. Cheminformatics, 1:8 (2009)

**Journal of Cheminformatics**



Research article

Open Access

**Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions**

Peter Ertl\* and Ansgar Schuffenhauer

Address: Novartis Institutes for BioMedical Research, Novartis Campus, CH-4002 Basel, Switzerland  
Email: Peter Ertl\* - peter.ertl@novartis.com; Ansgar Schuffenhauer - ansgar.schuffenhauer@novartis.com  
\* Corresponding author

Published: 10 June 2009

Received: 23 March 2009

Journal of Cheminformatics 2009, 1:8 doi:10.1186/1758-2946-1-8

Accepted: 10 June 2009

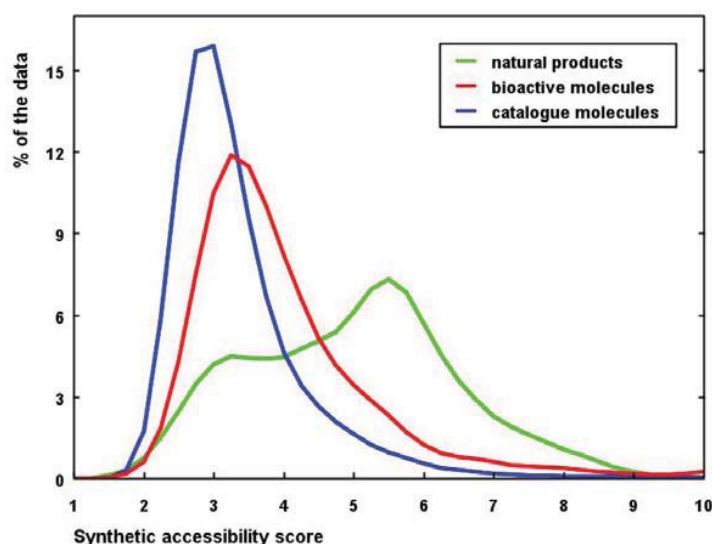
This article is available from: <http://www.jcheminf.com/content/1/1/8>

© 2009 Ertl and Schuffenhauer; licensee BioMed Central Ltd.

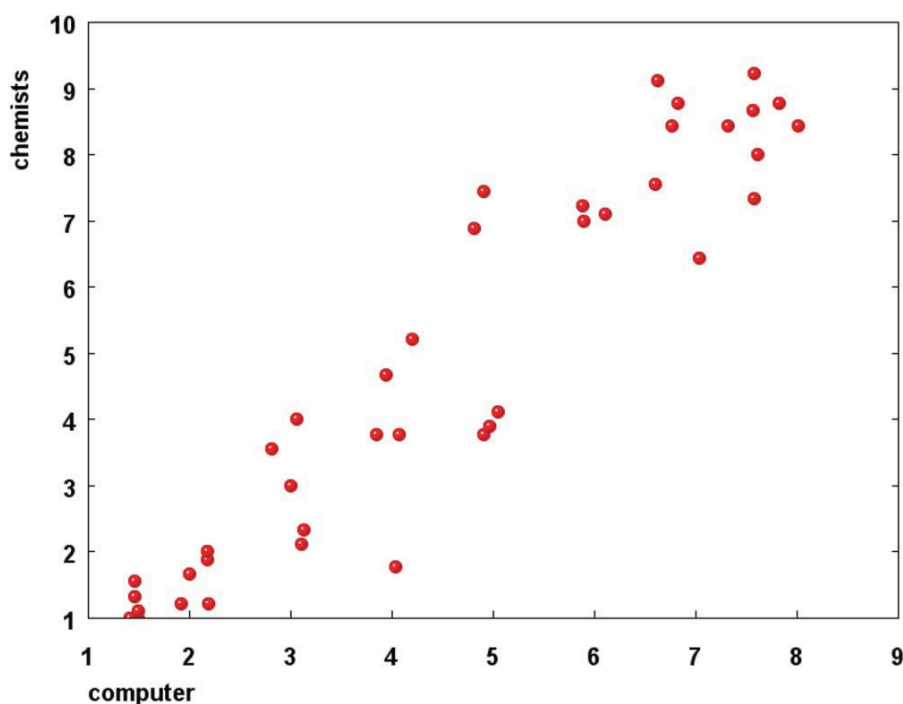
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Score Distribution

- Ease of synthesis of compounds
- SAScore = fragmentScore - complexityPenalty



# Synthetic Accessibility Score (SAS)



## RDKit

- <https://www.rdkit.org/>

## RDKit: Open-Source Cheminformatics Software

### Useful Links

- GitHub page
  - Git source code repository
  - The bug tracker
  - The releases (downloads)
- Sourceforge page
  - The mailing lists
  - Searchable archive of rdkit-discuss
  - Searchable archive of rdkit-devel
- RDKit at LinkedIn
- The RDKit Blog
- Online Documentation





# Tutorial

- <https://www.rdkit.org/docs/GettingStartedInPython.html>

The RDKit 2021.03.1 documentation » Getting Started with the RDKit in Python

Getting Started with the RDKit in Python

**Important note**  
Beginning with the 2019.03 release, the RDKit is no longer supporting Python 2. If you need to continue using Python 2, please stick with a release from the 2018.09 release cycle.

**What is this?**  
This document is intended to provide an overview of how one can use the RDKit functionality from Python. It's not comprehensive and it's not a manual.  
If you find mistakes, or have suggestions for improvements, please either fix them yourselves in the source document (the .rst file) or send them to the mailing list [rdkit-devel@lists.sourceforge.net](mailto:rdkit-devel@lists.sourceforge.net). In particular, if you find yourself spending time working out how to do something that doesn't appear to be documented please contribute by writing it up for this document. Contributing to the documentation is a great service both to the RDKit community and to your future self.

**Reading and Writing Molecules**

**Reading single molecules**  
The majority of the basic molecular functionality is found in module `rdkit.Chem`:

```
>>> from rdkit import Chem
```

Individual molecules can be constructed using a variety of approaches:

```
>>> m = Chem.MolFromSmiles('Cc1ccccc1')
>>> m = Chem.MolFromMolFile('data/input.mol')
>>> stringWithMolData=open('data/input.mol','r').read()
>>> m = Chem.MolFromMolBlock(stringWithMolData)
```

RDKit  
Open-Source Cheminformatics and Machine Learning

Table of Contents

- Getting Started with the RDKit in Python
  - Important note
  - What is this?
  - Reading and Writing Molecules
    - Reading single molecules
    - Reading sets of molecules
    - Writing molecules
    - Writing sets of molecules
  - Working with Molecules
    - Looping over Atoms and Bonds
    - Ring information
    - Modifying molecules
    - Working with 2D molecules: Generating Depictions
    - Working with 3D Molecules

# Colab

- <https://colab.research.google.com/notebooks/intro.ipynb>

Welcome To Colaboratory

File Edit View Insert Runtime Tools Help

Share Settings

Table of contents

- Getting started
- Data science
- Machine learning
- More Resources
- Machine Learning Examples
- Section

+ Code + Text Copy to Drive

Connect Editing

## What is Colaboratory?

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier. Watch [Introduction to Colab](#) to learn more, or just get started below!

### Getting started

The document you are reading is not a static web page, but an interactive environment called a **Colab notebook** that lets you write and execute code.

For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:



# Installing RDKit

- `!pip install rdkit`

## QED

- `from rdkit import Chem`
- `m = Chem.MolFromSmiles('C1CCCCC1')`
  
- `from rdkit.Chem import QED`
- `qed=QED.qed(m)`
- `print(qed)`

# SAS

- <https://mattermodeling.stackexchange.com/questions/8541/how-to-compute-the-synthetic-accessibility-score-in-python>

# Databases

# 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.

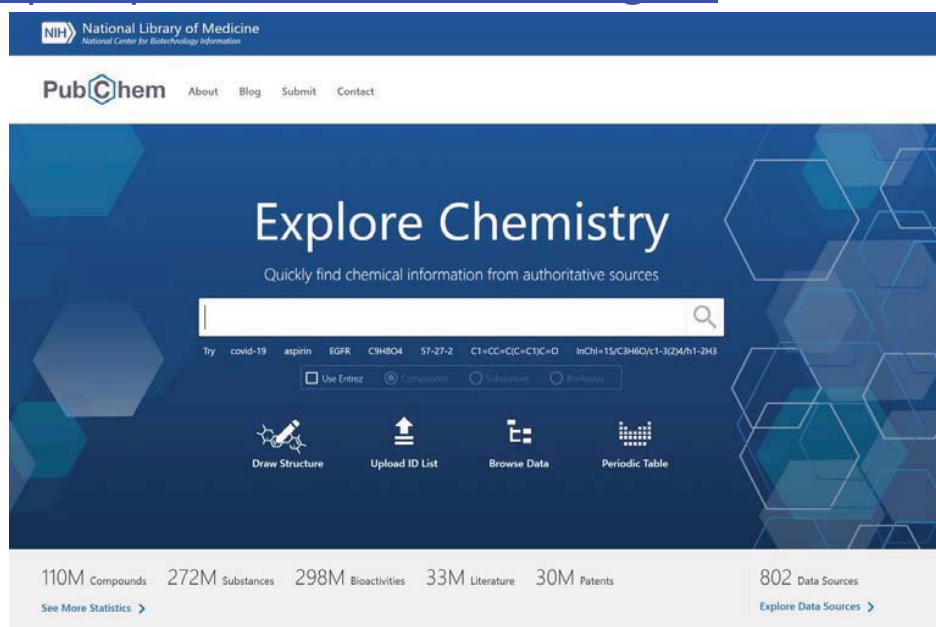
<http://dx.doi.org/10.1016/j.ddtec.2015.01.005>

# Databases

Database	Coverage (Number of entities)		
	Compounds	Proteins	Interactions
PubChem	111 m	99 k	273 m
ChEMBL	1,961,462	13,382	16,066,124
DUD-E	22,886	102	22.8 k*
DrugBank	13,791	5,696	27,954
STITCH	0.5 m	9.6 m	1.6b
TTD	2,251	3,473***	43,875
PharmGKB	708	–	–
Matador	801	2,901	15,843
DrugCentral	2,529	2,003	17,390
SuperTarget	195,770	6,219	332,828
Metz	3,858	172	258,094
MUV	93 k	17	–
ZINC	750 m**	2,864 (for eukaryotes)	638,174

# PubChem

- <https://pubchem.ncbi.nlm.nih.gov/>



## Components

- Compounds: Unique chemical structures
- Substances: Information about chemical entities
  - any combination of chemical structures, synonyms, registration IDs, descriptions, patent identifiers, protein 3D structures, and biological screening results, etc.
- Bioassay: Biological experiments
- Bioactivities

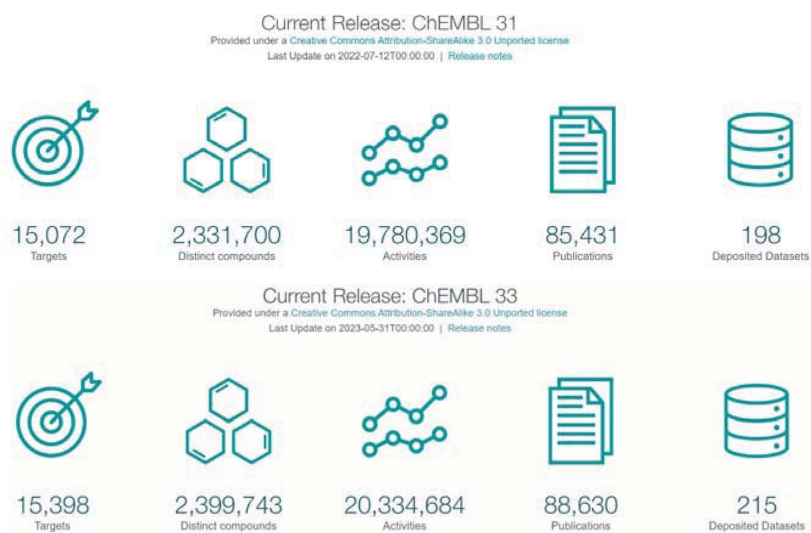
# Statistics

## PubChem Data Counts

Data Collection	Live Count	Description
Compounds	110,040,027	Unique chemical structures extracted from contributed PubChem Substance records
Substances	271,907,539	Information about chemical entities provided by PubChem contributors
BioAssays	1,366,296	Biological experiments provided by PubChem contributors
Bioactivities	298,299,306	Biological activity data points reported in PubChem BioAssays
Genes	103,715	Gene targets tested in PubChem BioAssays and those involved in PubChem Pathways
Proteins	96,561	Protein targets tested in PubChem BioAssays and those involved in PubChem Pathways
Taxonomy	112,763	Organisms of targets tested in PubChem BioAssays and those involved in PubChem Pathways
Pathways	237,925	Interactions between chemicals, genes, and proteins
Literature	32,849,900	Scientific publications with links in PubChem
Patents	29,940,379	Patents with links in PubChem
Data Sources	805	Organizations contributing data to PubChem

## ChEMBL

- <https://www.ebi.ac.uk/chembl/>
- A manually curated database of bioactive molecules with drug-like properties



# 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:
  - IC50 (half maximal inhibitory concentration)
  - Ki (binding affinity)
  - MIC (minimum inhibitory concentration)
  - % Inhibition (of activity)

## 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

- <https://go.drugbank.com/>
- Detailed drug (i.e. chemical) data with comprehensive drug target

DRUGBANK Online

Browse COVID-19 Search Interaction Checker Downloads Solutions About

Learn more about how DrugBank powers ReNorm's Drug Interaction API [Read Blog](#)

**Building the foundation for better health outcomes**

Access the right information at the right time, with our intelligent clinical drug data API and in-depth knowledge database.

[Learn about our solutions](#)

**Search over 500,000 drugs & drug products on DrugBank Online**

Tylenol

Home Targets Pathways Indications

## DrugBank example

Acetaminophen

**Identification**  
Pharmacology  
Interactions  
Products  
Categories  
Chemical Identifiers  
References  
Clinical Trials  
Pharmacoeconomics  
Properties  
Spectra  
Targets (4)  
Enzymes (15)  
Carriers (1)  
Transporters (1)

**Summary** Acetaminophen is an analgesic drug used alone or in combination with opioids for pain management, and as an antipyretic agent.

**Brand Names** Acephen, Acetadryl, Allzital, Apadaz, Arthriten Inflammatory Pain, Bupap, Butapap, Cetafen, Children's Silapap, Contac Cold and Flu Non Drowsy Maximum Strength, Coricidin Hi. [READ MORE](#)

**Generic Name** Acetaminophen **DrugBank Accession Number** DB00316

**Background** Acetaminophen (paracetamol), also commonly known as *Tylenol*, is the most commonly taken analgesic worldwide and is recommended as first-line therapy in pain conditions by the World Health Organization (WHO).<sup>10</sup> It is also used for its antipyretic effects, helping to reduce fever.<sup>23</sup> This drug was initially approved by the U.S. FDA in 1951 and is available in a variety of forms including syrup form, regular tablets, effervescent tablets, injection, suppository, and other forms.<sup>15,16,23,Label</sup>

Acetaminophen is often found combined with other drugs in more than 600 over the counter (OTC) allergy medications, cold medications, sleep medications, pain relievers, and other products.<sup>19</sup> Confusion about dosing of this drug may be caused by the availability of different formulas, strengths, and dosage instructions for children of different ages.<sup>19</sup> Due to the possibility of fatal overdose and liver failure associated with the incorrect use of acetaminophen, it is important to follow current and available national and manufacturer dosing guidelines while this drug is taken or prescribed.<sup>20,21,Label</sup>

**Type** Small Molecule **Groups** Approved

**Structure**   
[3D](#) [Download](#) [Similar Structures](#)

**Weight** Average: 151.1626  
Monoisotopic: 151.063328537

**Chemical Formula** C<sub>8</sub>H<sub>9</sub>NO<sub>2</sub>

# Targets

### 1. Prostaglandin E synthase 3 Details

<b>Kind</b>	Protein	<b>General Function</b>	Unfolded protein binding
<b>Organism</b>	Humans	<b>Specific Function</b>	Cytosolic prostaglandin synthase that catalyzes the oxidoreduction of prostaglandin endoperoxide H2 (PGH2) to prostaglandin E2 (PGE2) (PubMed:10922363). Molecular chaperone that localizes to genomi...
<b>Pharmacological action</b>	Unknown	<b>Gene Name</b>	PTGES3
<b>Actions</b>	Inhibitor	<b>Uniprot ID</b>	<a href="#">Q15185</a>
		<b>Uniprot Name</b>	Prostaglandin E synthase 3
		<b>Molecular Weight</b>	18697.195 Da

#### References

1. Botting R, Ayoub SS: COX-3 and the mechanism of action of paracetamol/acetaminophen. Prostaglandins Leukot Essent Fatty Acids. 2005 Feb;72(2):85-7. [\[Article\]](#)
2. Chandrasekharan NV, Dai H, Roos KL, Evanson NK, Tomsik J, Elton TS, Simmons DL: COX-3, a cyclooxygenase-1 variant inhibited by acetaminophen and other analgesic/antipyretic drugs: cloning, structure, and expression. Proc Natl Acad Sci U S A. 2002 Oct 15;99(21):13926-31. Epub 2002 Sep 19. [\[Article\]](#)
3. Data sheet, Acetaminophen, ebi.ac.uk [\[File\]](#)

### 2. Prostaglandin G/H synthase 2 Binding Properties Details

<b>Kind</b>	Protein	<b>General Function</b>	Prostaglandin-endoperoxide synthase activity
-------------	---------	-------------------------	--

# ZINC

- <http://zinc.docking.org/>
- ZINC was originally designed for target based virtual screening (docking)
- Now, zinc20



# (Old) ZINC subsets

	Lead-Like	Fragment-Like	Drug-Like	All	Shards
<b>Standard</b> Size Updated	<a href="#">Lead-Like</a> 6,053,287 2014-09-29	<a href="#">Fragment-Like</a> 847,909 2015-02-04	<a href="#">Drug-Like</a> 17,900,742 2014-11-24	<a href="#">All Purchasable</a> 22,724,825 2014-11-28	<a href="#">Shards</a> 635,159 2014-05-16
<b>Clean</b> Size Updated	<a href="#">Clean Leads</a> 4,591,276 2014-09-25	<a href="#">Clean Fragments</a> 1,611,889 2014-09-24	<a href="#">Clean Drug-Like</a> 13,195,609 2013-11-05	<a href="#">All Clean</a> 16,403,865 2013-12-18	<a href="#">Clean Shards</a> 325,950 2014-11-24
<b>In Stock</b> Size Updated	<a href="#">Leads Now</a> 3,687,621 2014-06-25	<a href="#">Frgs Now</a> 704,041 2015-02-04	<a href="#">Drugs Now</a> 10,639,555 2014-11-24	<a href="#">All Now</a> 12,782,590 2014-05-01	<a href="#">Shards Now</a> 424,775 2014-09-24
<b>Boutique</b> Size Updated	<a href="#">Boutique Leads</a> 5,114,169 2012-12-24	<a href="#">Boutique Frags</a> 2,755,555 2013-11-08	<a href="#">Boutique Drugs</a> 10,292,210 2012-11-27	<a href="#">All Boutique</a> 12,217,845 2012-11-27	<a href="#">Boutique Shards</a> 80,698 2013-11-08
Comments/Citation	<a href="#">Teague, Davis, Leeson, Oprea, Angew Chem Int Ed Engl, 1999 Dec 16;38(24):3743-3748.</a>	<a href="#">Carr RA, Congreve M, Murray CW, Rees DC, Drug Discov Today, 2005 Jul 15;10(14):987</a>	<a href="#">Lipinski, J Pharmacol Toxicol Methods, 2000 Jul-Aug;44(1):235-49.</a>	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

Rep. 2D 3D React. Standard Purch. Wait OK pH N/A Charge N/A

Molecular Weight (up to, Daltons)

	200	250	300	325	350	375	400	425	450	500	>500	Totals, by LogP
-1	27,791	172,563	710,795	1,072,978	2,241,498	786,738	276,834	116,066	92,417	77,790	7,310	5,582,780
0	139,434	934,776	3,655,384	5,126,157	10,608,025	3,498,214	1,663,579	708,919	570,546	507,344	4,734	27,417,112
1	362,437	2,884,636	12,030,074	16,154,544	33,650,249	11,885,957	6,807,876	3,178,487	2,648,581	2,412,998	9,940	92,025,779
2	467,220	4,584,223	22,941,208	30,908,513	65,047,385	26,752,849	17,839,254	9,349,272	8,099,970	7,686,687	24,554	193,701,135
2.5	167,513	2,136,113	12,849,121	17,977,157	38,682,058	18,584,223	13,812,274	8,111,104	7,197,414	6,979,014	24,126	126,520,117
3	90,548	1,570,772	11,037,383	16,282,627	34,831,558	19,940,391	16,037,132	10,339,743	9,362,233	9,118,717	37,422	128,648,526
3.5	36,748	929,872	7,920,574	12,490,662	27,380,104	18,703,024	16,485,194	11,784,160	10,774,472	10,693,411	58,791	117,257,012
4	9,017	369,565	4,332,131	6,472,808	10,487,856	13,034,155	14,329,253	11,683,208	10,891,465	11,003,975	86,262	82,699,695
4.5	993	86,613	1,814,492	3,457,942	6,367,225	8,853,064	10,320,054	9,945,353	9,486,869	9,825,079	117,980	60,275,664
5	150	13,393	536,018	1,405,708	3,168,584	4,995,850	6,471,525	7,025,034	6,976,742	7,325,833	144,297	38,063,134
>5	39	1,097	22,854	103,521	376,905	927,395	1,670,856	2,195,160	2,588,702	3,052,048	767,762	11,706,339
Totals, by Weight	1,301,890	13,683,623	77,850,034	111,452,617	232,841,447	127,961,860	105,713,831	74,436,506	68,689,411	68,682,896	1,283,178	884M Substances 1.9K Tranches

# Targets

- <https://zinc.docking.org/majorclasses/>

Name	# Sub Classes	# Genes	# Orthologs	# Observations	# Substances	# Purchasable	# Predictions
adhesion	1	7	11	534	292	32	79415
auxiliary transport protein	3	8	14	643	458	182	497749
cytosolic other	1	39	58	5664	4162	461	2859674
enzyme	13	1942	2819	412921	205504	23889	107386614
epigenetic regulator	3	97	103	6250	2856	510	7743379
ion channel	3	152	246	34563	22500	2638	34121578
membrane other	1	6	12	301	271	30	166343
membrane receptor	7	289	670	307365	143352	13445	79804362
Nuclear-other	1	6	8	1053	784	68	176523
Secreted	1	41	52	913	757	175	3733709
Structural	1	7	9	482	417	161	310275
surface antigen	1	14	25	444	383	74	4370725
Transcription factor	2	53	108	45550	18111	1687	5537528
Transporter	4	110	166	47632	19622	2889	12273240
Unclassified	1	540	611	11256	8680	1942	21021371

# Protein Data Bank(PDB)

- <https://www.rcsb.org/>

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB Login

**RCSB PDB** An Information Portal to 123622 Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligands

Advanced Search | Browse by Annotations

[PDB-101](#)
[PDB](#)
[EMDataBank](#)
[Bioinformatics Knowledgebase](#)
[Structural Biology Knowledgebase](#)
[Wellcome Protein Data Bank Foundation](#)

**Welcome**

**Deposit**

**Search**

**Visualize**

**Analyze**

**Download**

**Learn**

**A Structural View of Biology**

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

**Events and Activities**

**PUBLIC POSSESSION AESTHETICS OF LIFE SCIENCES** October 21, 2019 RUTGERS

**PDB-101 USER SURVEY**

**October Molecule of the Month**

**Dipeptidyl Peptidase 4**

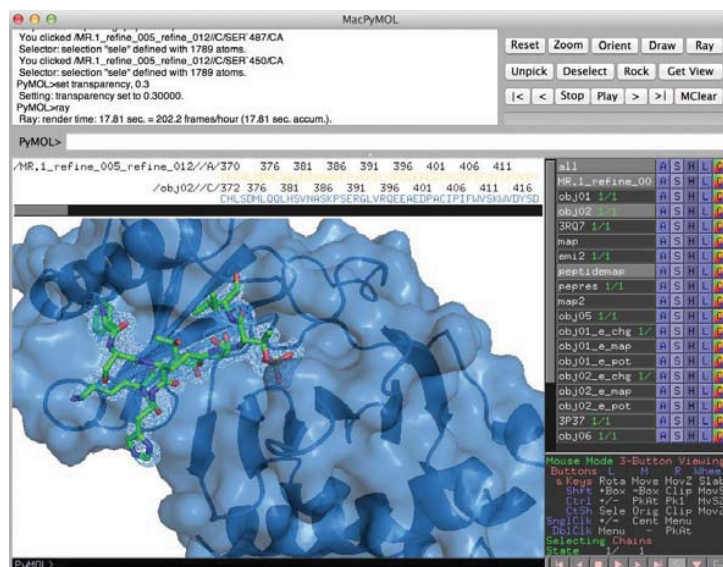
Latest Entries As of Tuesday, Oct 18 Features & Highlights News Publications Contact Us

# PDB ID

- 4-letter code
  - e.g) 12AS, 3INS
- Chain ID concatenated form
  - e.g) 12ASA

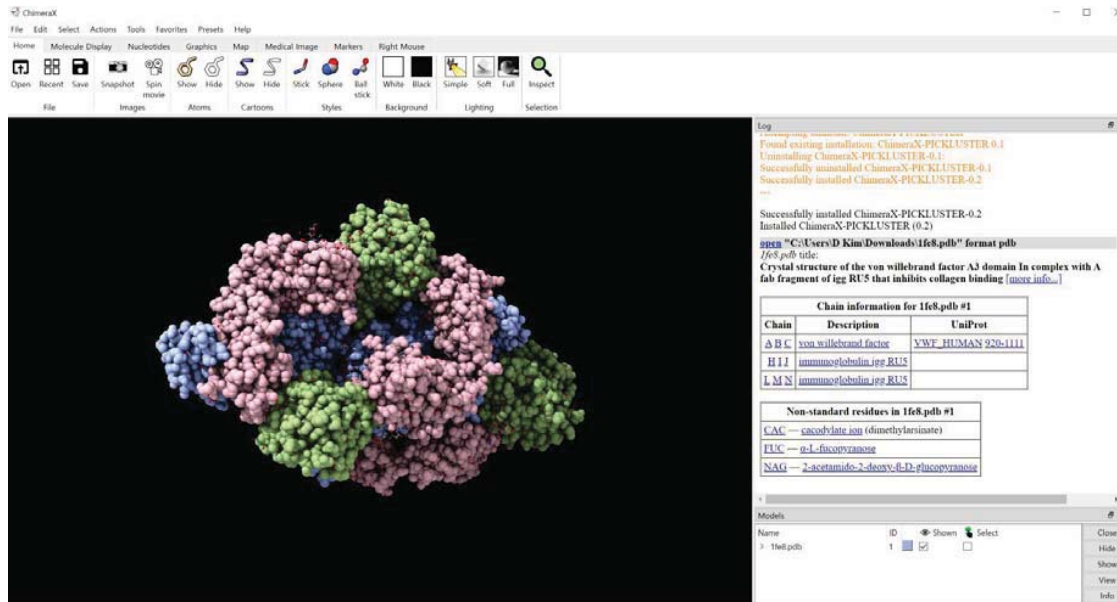
# PyMOL: structure viewer

- Free software (<http://pymol.org>)
- [https://pymolwiki.org/index.php/Windows\\_Install](https://pymolwiki.org/index.php/Windows_Install)

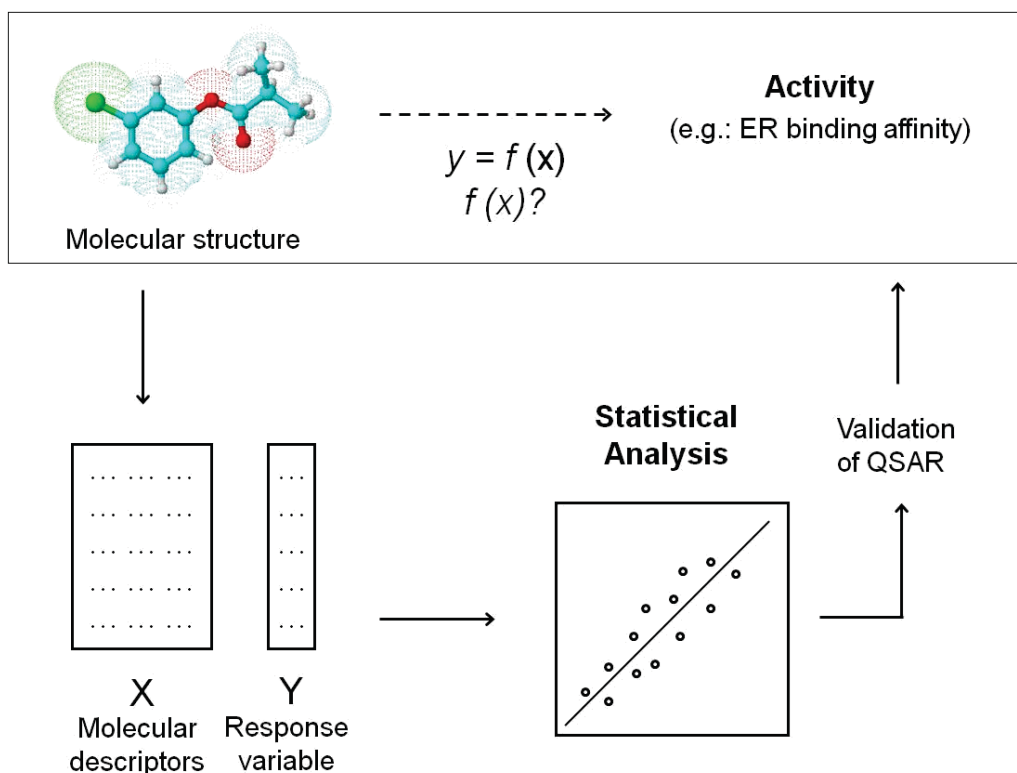


# UCSF ChimeraX

- <https://www.rbvi.ucsf.edu/chimerax/>



## QSAR 모델링 과정



# Molecular Descriptors

- Constitutional descriptors
  - molecular weight, number of chemical elements, number of H-bonds or double bonds, ...
- Physicochemical descriptors
  - lipophilicity, polarizability, ...
- Topological descriptors
  - atomic branching, ...
- Electronic, geometrical and quantum-chemical descriptors
- Fragmental/Structural keys
  - MACCS keys, ECFP

## 1D, 2D, 3D

- 1D descriptors encode numerically generic properties
  - Molecular weight, molar refractivity, and octanol/water partition coefficient, etc.
- 2D descriptors: topological representations of molecules.
  - 2D-QSAR
- 3D descriptors: obtained directly from the 3D structure of molecules
  - 3D-QSAR methods
  - Dependent on the molecular conformation

# PaDEL descriptor

- <http://www.yapcwsoft.com/dd/padeldescriptor/>
- 1875 descriptors (1444 2D\_descriptors + 431 3D\_descriptors)

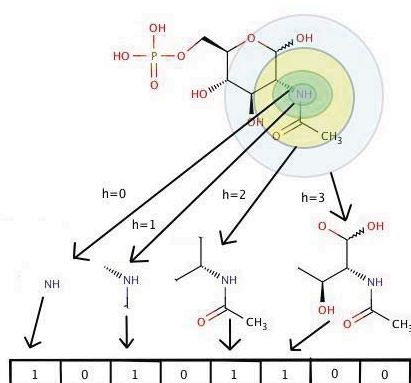
Descriptor Java Class	Descriptor	Description	Class
AcidicGroupCountDescriptor	nAcid	Number of acidic groups. The list of acidic groups is defined by these SMARTS "[O;H1]-[C,S,P]=O)", "[*]-[*]-[*];+]]]"	2D
ALOGPDescriptor	ALogP	Ghose-Crippen LogKow	2D
	ALogP2	Square of ALogP	2D
	AMR	Molar refractivity	2D
APoDescriptor	apoi	Sum of the atomic polarizabilities (including implicit hydrogens)	2D
AromaticAtomsCountDescriptor	naAromAtom	Number of aromatic atoms	2D
AromaticBondsCountDescriptor	nAromBond	Number of aromatic bonds	2D
AtomCountDescriptor	nAtom	Number of atoms	2D
	nHeavyAtom	Number of heavy atoms (i.e. not hydrogen)	2D
	nH	Number of hydrogen atoms	2D
	nB	Number of boron atoms	2D
	nC	Number of carbon atoms	2D
	nN	Number of nitrogen atoms	2D
	nO	Number of oxygen atoms	2D
	nS	Number of sulphur atoms	2D
	nP	Number of phosphorus atoms	2D
	nF	Number of fluorine atoms	2D
	nCl	Number of chlorine atoms	2D
	nBr	Number of bromine atoms	2D
	nI	Number of iodine atoms	2D
	nX	Number of halogen atoms (F, Cl, Br, I, At, Uus)	2D
AutocorrelationDescriptor	ATS0m	Broto-Moreau autocorrelation - lag 0 / weighted by mass	2D
	ATS1m	Broto-Moreau autocorrelation - lag 1 / weighted by mass	2D
	ATS2m	Broto-Moreau autocorrelation - lag 2 / weighted by mass	2D
	ATS3m	Broto-Moreau autocorrelation - lag 3 / weighted by mass	2D
	ATS4m	Broto-Moreau autocorrelation - lag 4 / weighted by mass	2D
	ATS5m	Broto-Moreau autocorrelation - lag 5 / weighted by mass	2D
	ATS6m	Broto-Moreau autocorrelation - lag 6 / weighted by mass	2D
	ATS7m	Broto-Moreau autocorrelation - lag 7 / weighted by mass	2D

# Fragment Codes

- A fragment coding system is based on a collection of small substructures or features in a closed list.
- Sub structural 'keys' from a fragment dictionary are usually recorded as a binary bit string, or fingerprint.
  - MACCS Keys
  - Comparing fingerprint bit strings is very fast.
- The alternative to structural keys is a 'hashed fingerprint.'
  - ECFPs (Extended Connectivity FingerPrints)
  - Morgan fingerprint

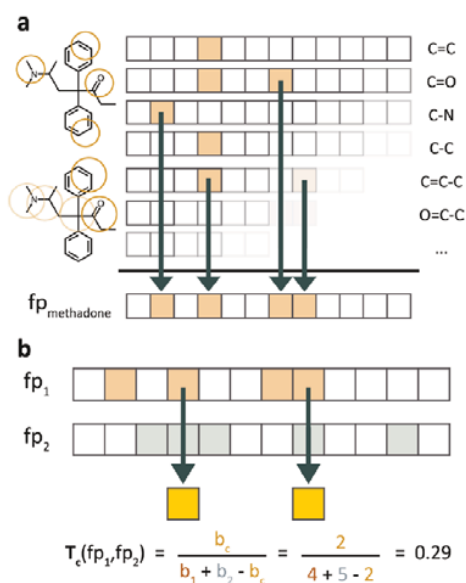
# Molecular Fingerprint

- Bit string representations of molecular structure and properties
- 2D structure features typically encoded as a vector of binary values
- ECFPs, Morgan
- Reasons for popularity in similarity searching:
  - computational efficiency
  - surprising effectiveness in detecting active compounds



# Similarity

- Tanimoto coefficient





# ECFP

- Extended Connectivity FingerPrint
- <https://docs.chemaxon.com/display/docs/extended-connectivity-fingerprint-ecfp.md>

742

*J. Chem. Inf. Model.* **2010**, *50*, 742–754

## Extended-Connectivity Fingerprints

David Rogers<sup>\*,†</sup> and Mathew Hahn<sup>‡</sup>

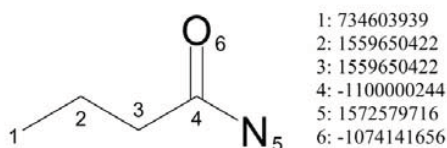
3429 North Mountain View Drive, San Diego, California 92116 and Accelrys, Incorporated, 10188 Telesis Court, Suite 100, San Diego, California 92121

Received February 4, 2010

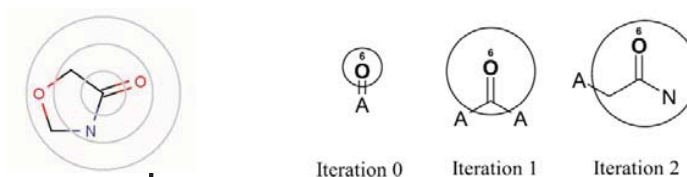
Extended-connectivity fingerprints (ECFPs) are a novel class of topological fingerprints for molecular characterization. Historically, topological fingerprints were developed for substructure and similarity searching. ECFPs were developed specifically for structure–activity modeling. ECFPs are circular fingerprints with a number of useful qualities: they can be very rapidly calculated; they are not predefined and can represent an essentially infinite number of different molecular features (including stereochemical information); their features represent the presence of particular substructures, allowing easier interpretation of analysis results; and the ECFP algorithm can be tailored to generate different types of circular fingerprints, optimized for different uses. While the use of ECFPs has been widely adopted and validated, a description of their implementation has not previously been presented in the literature.

# 생성 과정

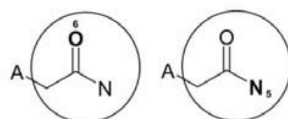
- Initial assignment of atom identifier



- Iterative updating of identifiers



- Duplication removal

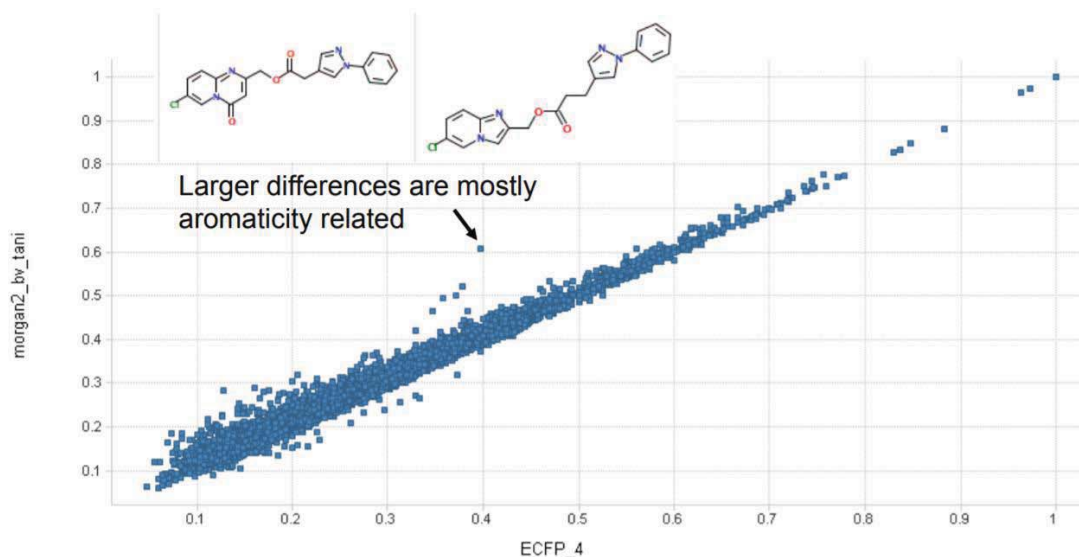






# ECFP vs. RDKit Morgan FP

## RDKit Morgan2 vs PP ECFP4



## RDKit Morgan3 vs PP ECFP6 is similar

# Morgan/Circular FP

## • Rdkit implementation of ECFP

```
>>> from rdkit.Chem import AllChem
>>> m1 = Chem.MolFromSmiles('Cc1ccccc1')
>>> fp1 = AllChem.GetMorganFingerprint(m1,2)
>>> fp1
<rdkit.DataStructs.cDataStructs.UIntSparseIntVect object at 0x...>
>>> m2 = Chem.MolFromSmiles('Cc1ncccc1')
>>> fp2 = AllChem.GetMorganFingerprint(m2,2)
>>> DataStructs.DiceSimilarity(fp1,fp2)
0.55...
```

```
>>> fp1 = AllChem.GetMorganFingerprintAsBitVect(m1,2,nBits=1024)
>>> fp1
<rdkit.DataStructs.cDataStructs.ExplicitBitVect object at 0x...>
>>> fp2 = AllChem.GetMorganFingerprintAsBitVect(m2,2,nBits=1024)
>>> DataStructs.DiceSimilarity(fp1,fp2)
0.51...
```

# KSBi-BIML 2024

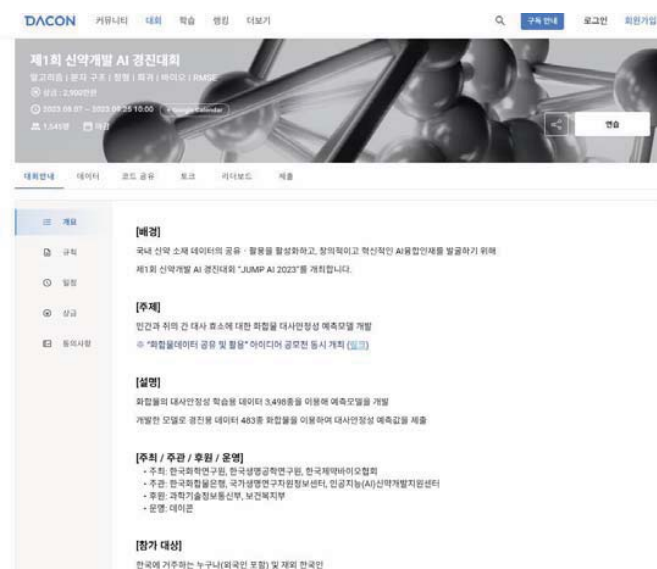
## 인공지능 신약설계 AI Drug Design

### 개요

- 강의
  - QSAR 모델링 기초
  - AI 신약개발을 위한 기계학습법 기초
  - AI 신약개발을 위한 딥러닝 모델
  - Virtual screening (ligand-based, structure-based) 및 de novo design
- 실습
  - QSAR modeling 전체 과정 실습
  - 화합물의 Bioactivity 예측 모델 개발
  - Virtual screening 과정을 통한 신약후보물질 발굴 실습

# 신약개발 AI 경진대회

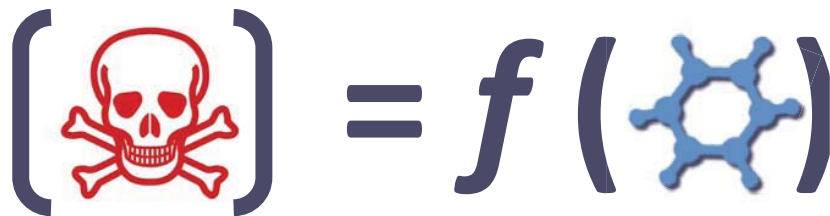
- <https://dacon.io/competitions/official/236127/overview/description>



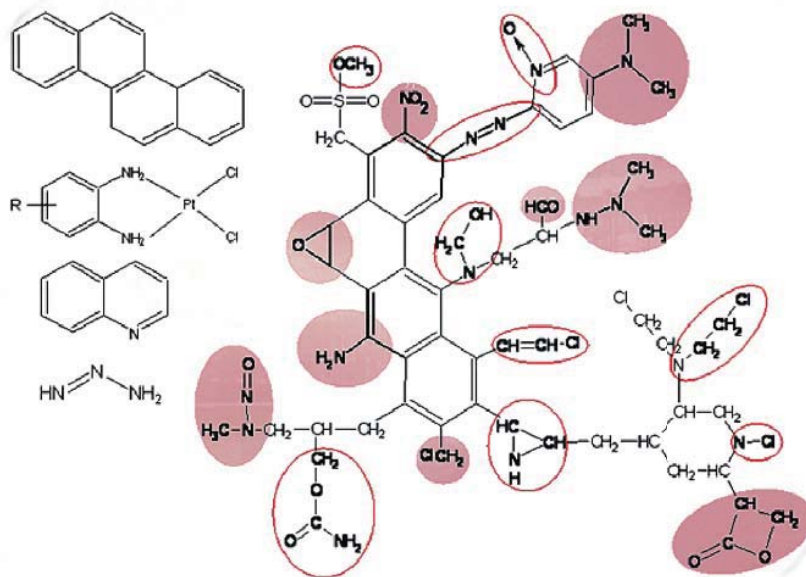
## QSAR 모델링

# QSAR

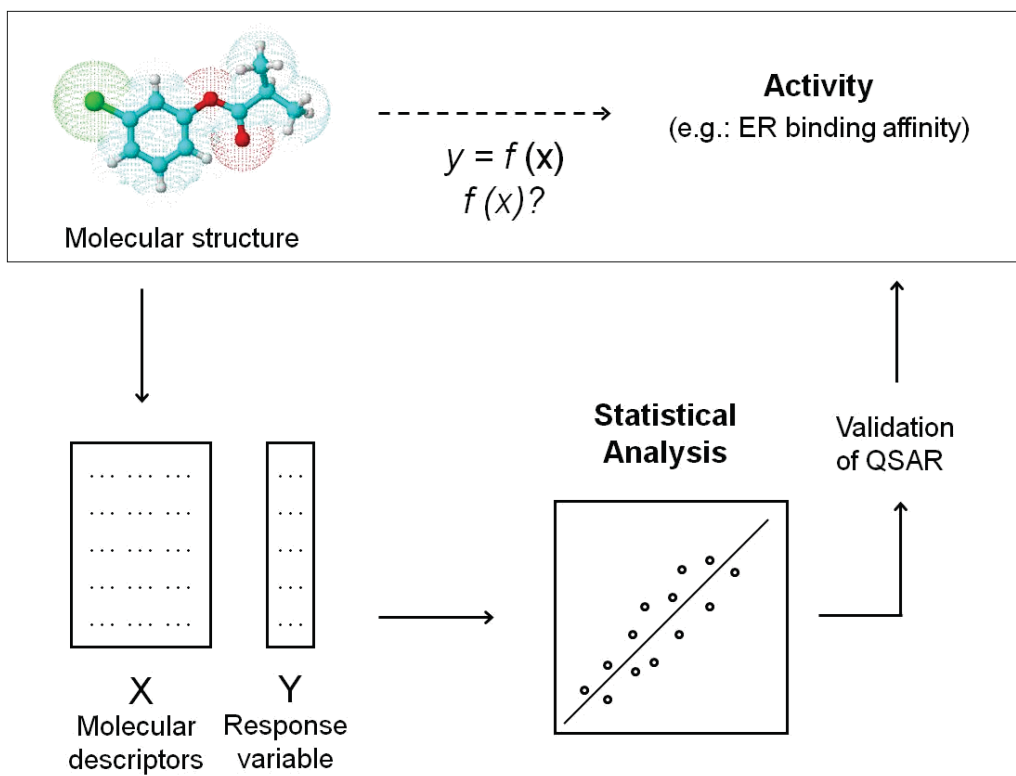
- Quantitative structure–activity relationships
- Construction of a mathematical model relating a molecular structure to a chemical property or biological effect by means of statistical techniques



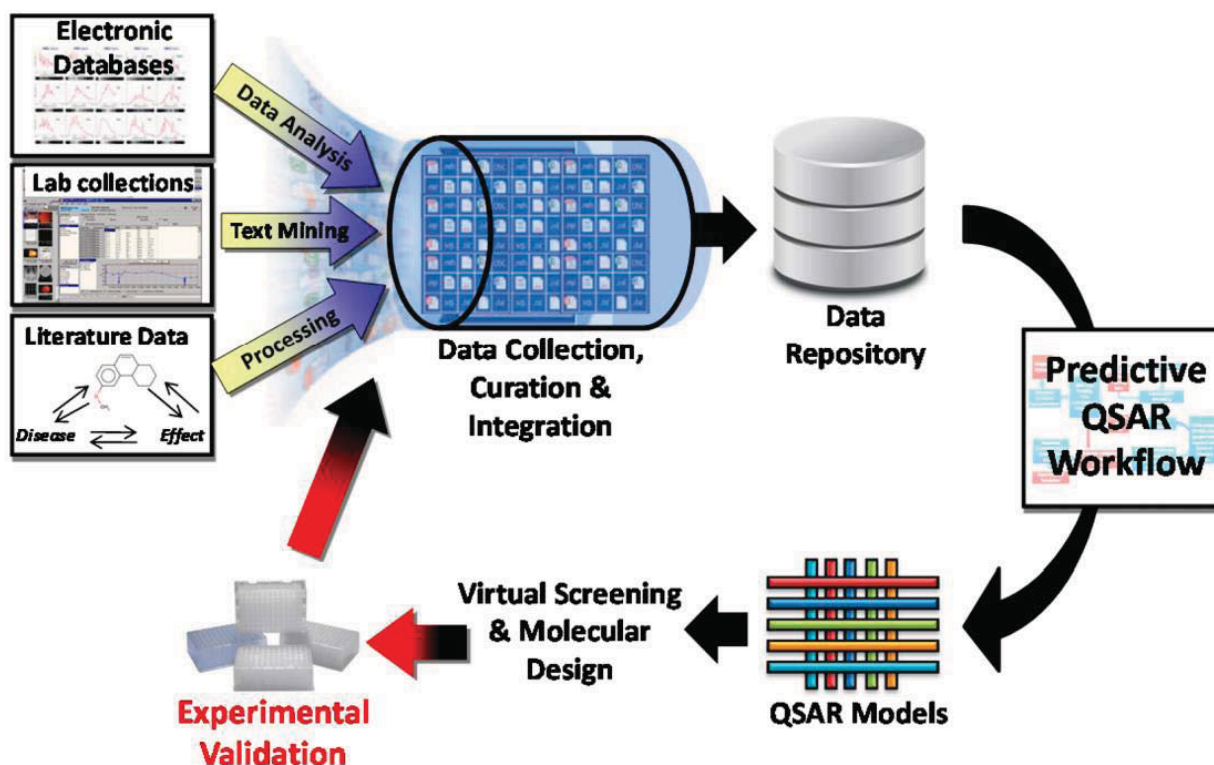
## Link between toxicity and structures



# QSAR 모델링 과정

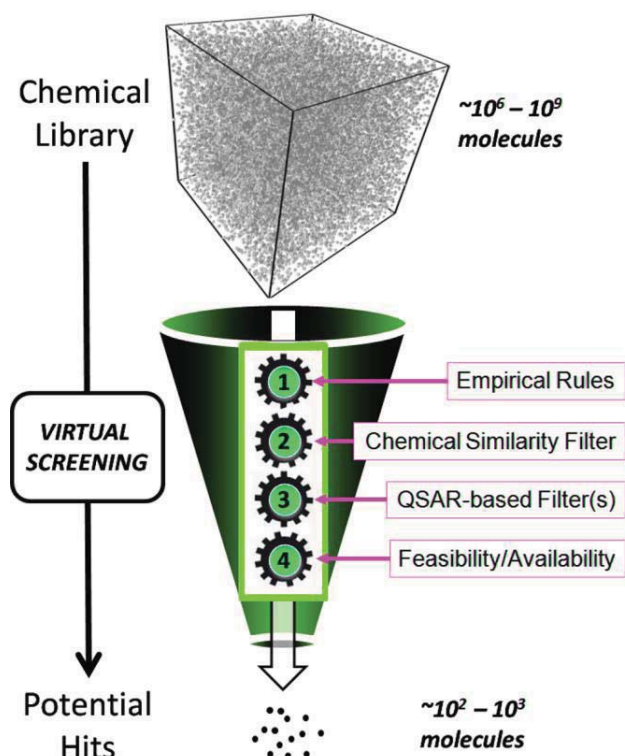


# QSAR-guided drug discovery

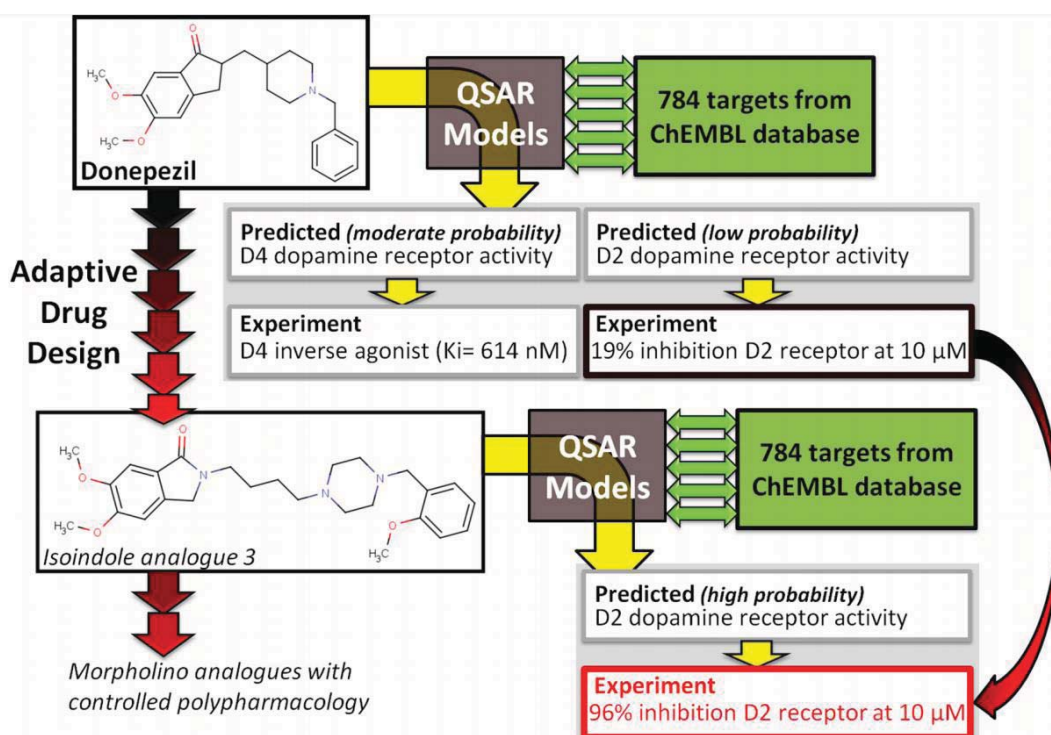




# QSAR-based virtual screening



# Target prediction and optimization



# Components

- **화합물 데이터**: a set of chemical structures that are represented by molecular descriptors
- **Activity 데이터**: a set of observed 'activities' associated with the structures.
  - Any form of experimental observation, not limited to biological activities
  - Numerical ( $IC_{50}$ ,  $K_i$ , or  $K_d$ ) or
  - Categorical labels (active/inactive; soluble/insoluble)
- A statistical modeling method to identify the key relationships between the molecular descriptors and the activities
  - Linear regression, SVM, Random forest, Deep learning

# Binding Affinity

- **IC50** - The half maximal (50%) inhibitory concentration, a measure of the potency of a substance in inhibiting a specific biological or biochemical function.
- **EC50** - Half maximal effective concentration, the concentration of compound that generates a half-maximal response in a given assay.
- **KD** – dissociation constant; the concentration of ligand that gives even odds that a given protein molecule has a ligand bound.
- **KI** - For enzyme inhibitors, this is the inhibition constant, essentially the dissociation constant KD
- **$\Delta G$**  – Gibbs free energy change associated with a chemical reaction, here a binding reaction

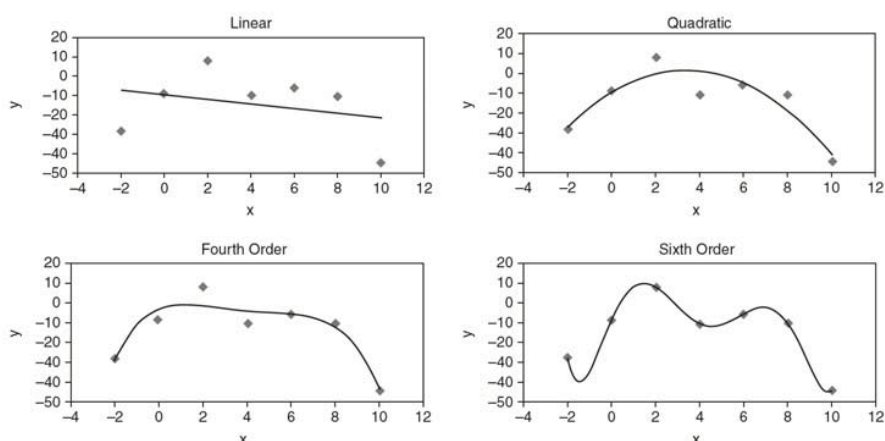


# PREPARATION

- ‘Garbage-in, garbage-out’ principle
- There are many ways in which erroneous or misleading models can be produced.
  - Data and/or Statistical method
- Check that the observations are consistent, preferably obtained from a single experimental source.
- Data taken from different assays should not be combined into a single model where possible.
- It is better to have the data points evenly spread.
- We cannot be sure that what is not reported is indeed negative.

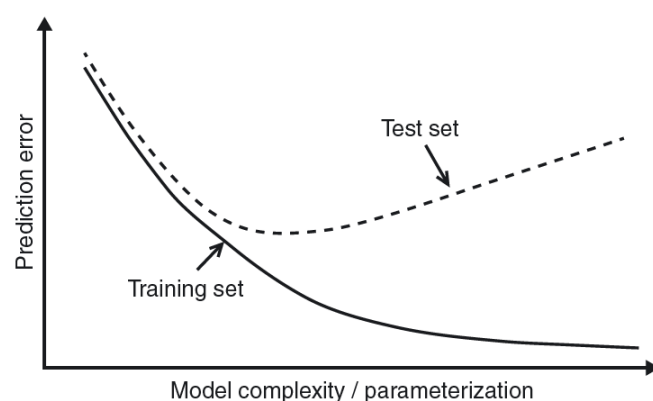
## Model validation

- Once the model is fully optimized, it is important to determine the level of prediction accuracy that can be expected when the model is applied to new compounds.
- The fit of a model to its training data is *not* a good indicator of its predictive performance for new compounds.



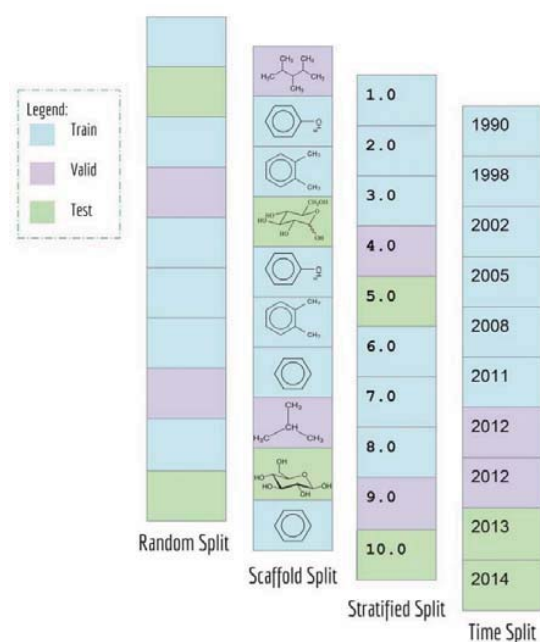
## External Test Sets and Cross Validation

- The most basic approach for assessing models involves splitting a dataset into a training set and a test set (or validation set).
- Train your model until prediction error is minimized on a test set.
- Finally test the model accuracy on an independent test set



## Data Splitting

- A number of different methods for splitting datasets
  - Random
  - Stratified
  - Cluster-based (scaffold split)
  - Temporal:
    - ChEMBL20 (training), ChEMBL21 (test)



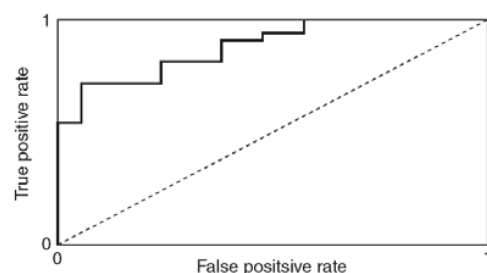
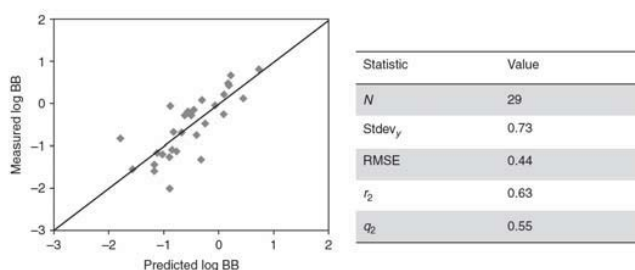
# Cross Validation

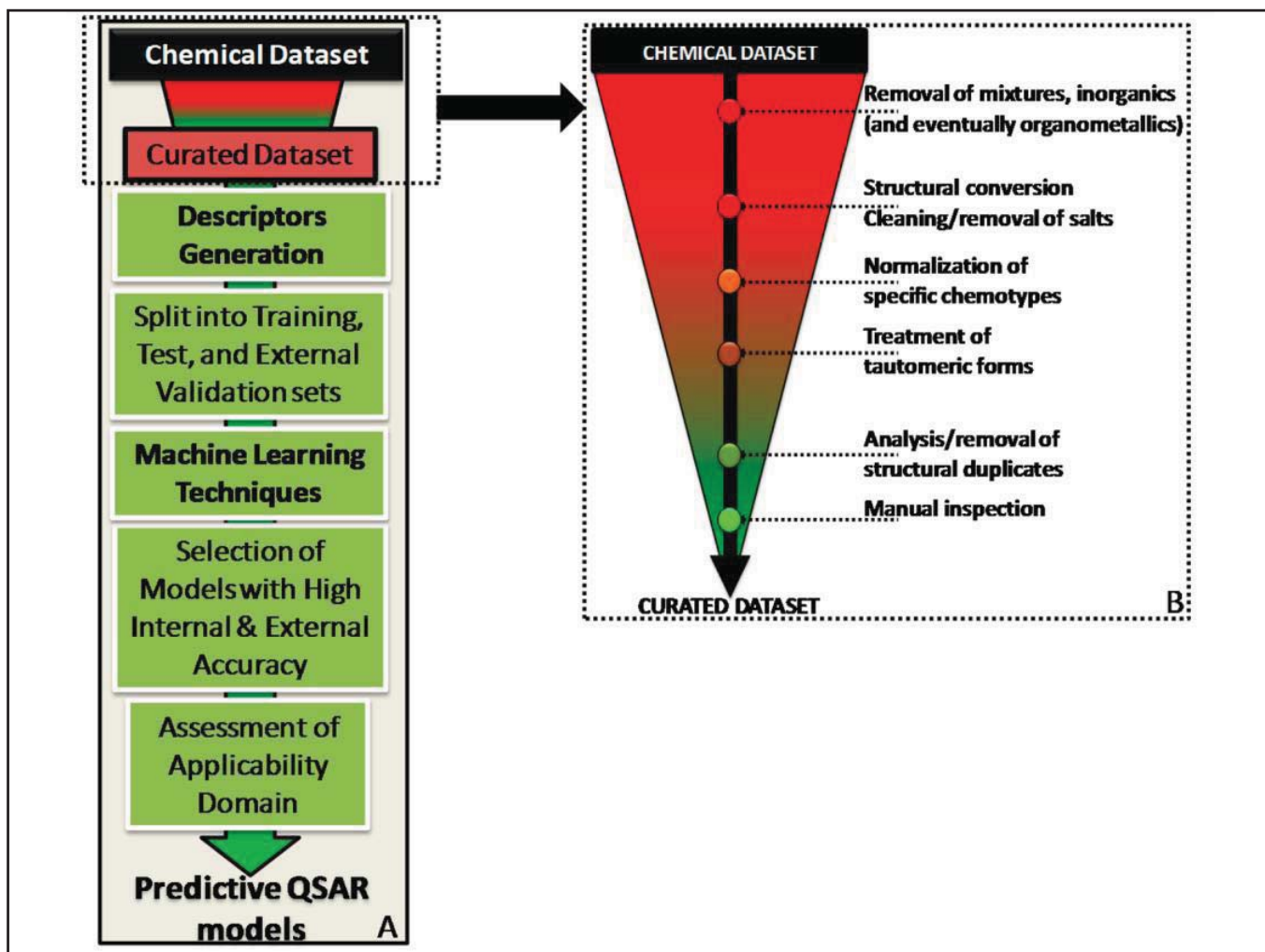
- Cross- validation
  - Leave-one-out, leave-cluster-out,  $n$ -fold cross validation
- Additional validation set



# Assessing Model Performance

- <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>
- Regression Problems
  - MAE, MSE, RMSE, Pearson correlation coefficient, Spearman Rank Correlation
- Classification Problems
  - Classification Accuracy, Precision, Recall, F1 score, AUC, PRC

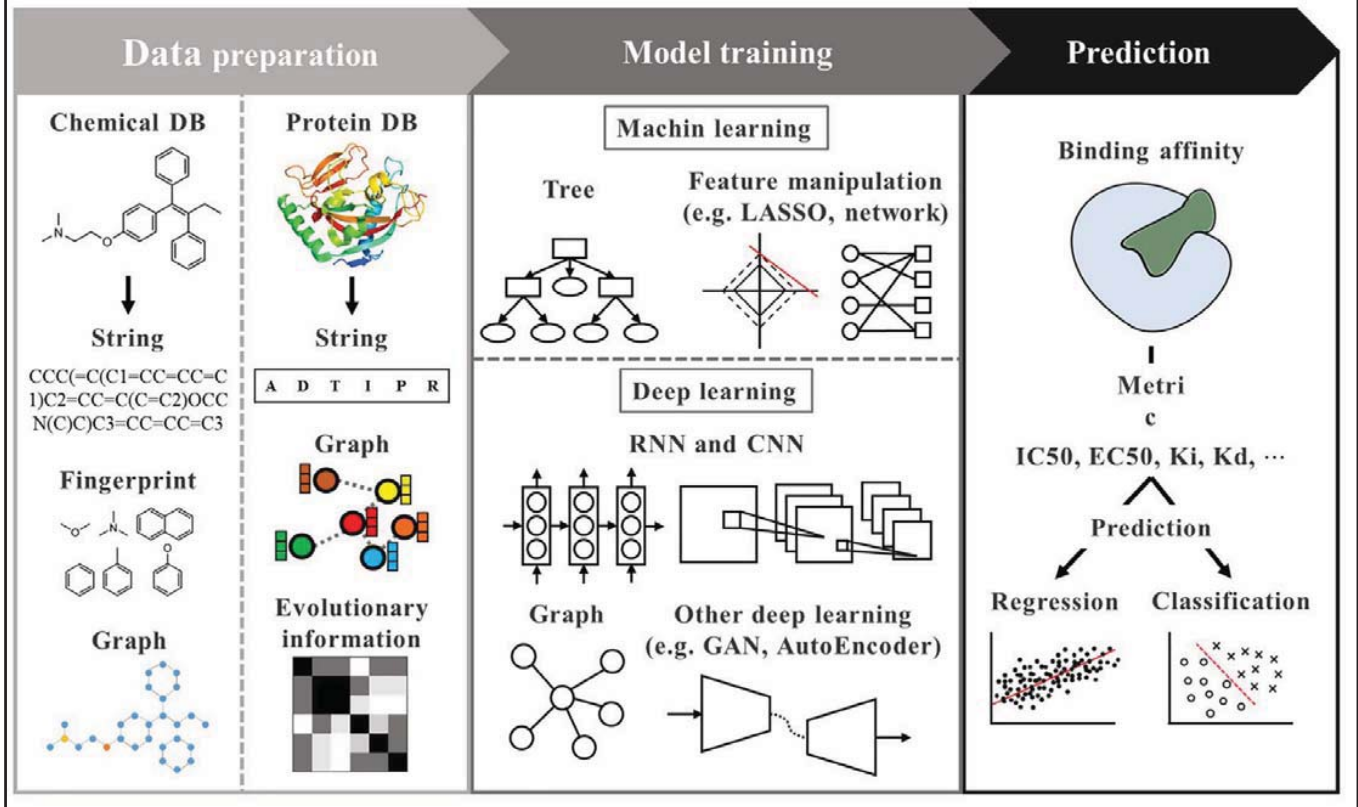




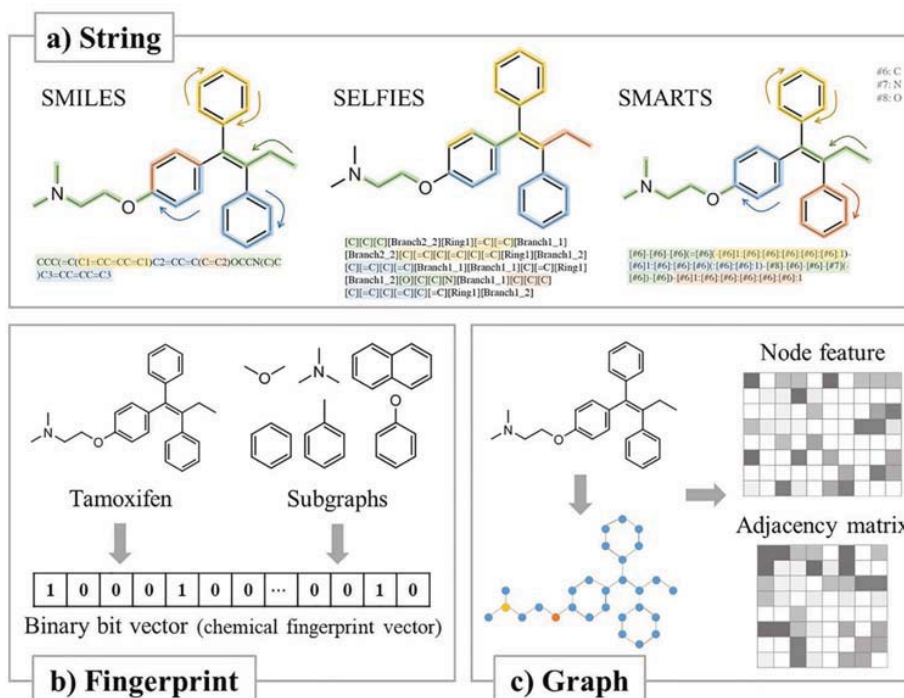
## Lower limit

- If training sets are too small, correlation and over-fitting problem.
- Continuous response variable (activity),
  - the number of compounds in the training set should be at least 20
  - about 10 compounds should be in each of the test and external evaluation sets.
- Classification or category response variable
  - training set should contain at least about 10 compounds of each class
  - test and external evaluation sets should contain no less than five compounds for each class.

# ML Approaches: Overall Process

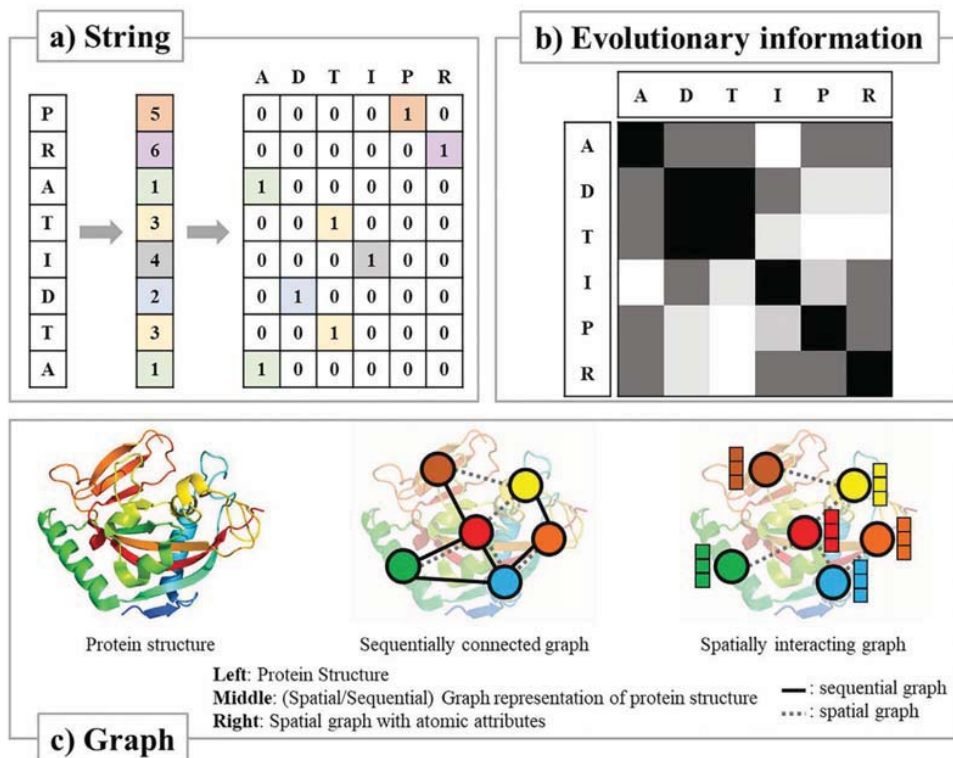


# Ligand Featurization

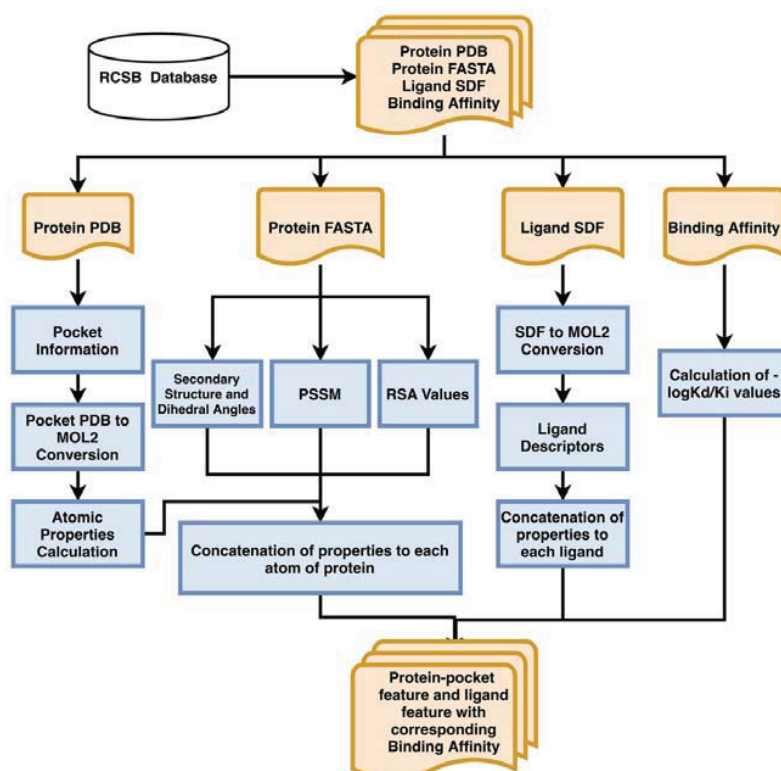




# Protein Featurization



# Feature extraction pipeline



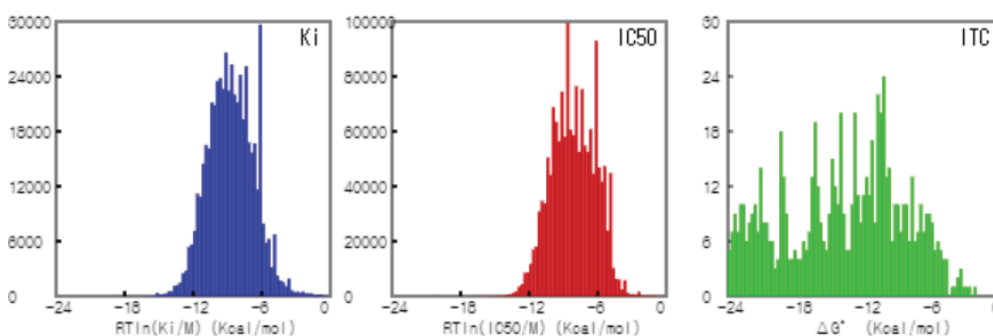
# Database

Protein-centric databases			
	Compounds	Proteins	Interactions
UniProt	-	20,385	-
Protein Data Bank	-	170,597	-
PDBbind	11,762	3,566	17,679*
Pfam	-	18,259	-
BRENDA	46	8083**	500 k
Integrated databases			
	Compounds	Proteins	Interactions
KEGG	18,749***	31,224,482****	-
BindingDB	910,479	8,161	2.1 m
Davis	72	442	30 k
K KIBA	229	211	118 k
IUPHAR/BPS	10,053	2,943	48,902

## BindingDB

- <https://www.bindingdb.org/rwd/bind/>
- As of July 24, 2022, 2,546,129 binding data for 8,821 protein targets and 1,093,579 small molecules
- <https://www.bindingdb.org/bind/glossary.jsp>

BindingDB Affinity Statistics



# QSAR를 위한 기계학습법

## Scikit-learn

- <https://scikit-learn.org/stable/>



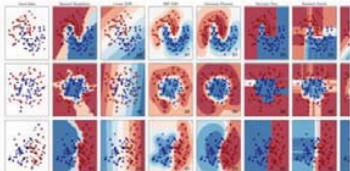
The image shows the header of the Scikit-learn website. It features the Scikit-learn logo on the left, followed by navigation links: "Install", "User Guide", "API", "Examples", and "More". Below the logo is the text "scikit-learn" and "Machine Learning in Python". On the right side, there are four bullet points: "Simple and efficient tools for predictive data analysis", "Accessible to everybody, and reusable in various contexts", "Built on NumPy, SciPy, and matplotlib", and "Open source, commercially usable - BSD license". At the bottom of the header, there are three buttons: "Getting Started", "Release Highlights for 0.24", and "GitHub".

### Classification

Identifying which category an object belongs to.

**Applications:** Spam detection, image recognition.

**Algorithms:** SVM, nearest neighbors, random forest, and more...

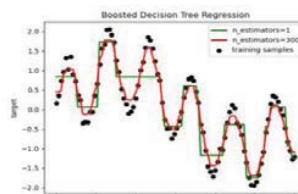


### Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.

**Algorithms:** SVR, nearest neighbors, random forest, and more...

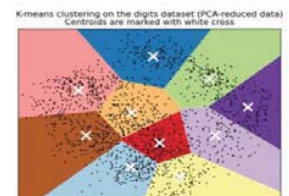


### Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** k-Means, spectral clustering, mean-shift, and more...





# Machine learning book

- <https://product.kyobobook.co.kr/detail/S000200135401>

## Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow

Concepts, Tools, and Techniques to Build Intelligent Systems 3/E | Paperback

Aurelien Geron 저자(글)  
O'Reilly Media : 2022년 01월 01일  
주관베스트 과학/기술 96위  
가장 최근에 출시된 개정판입니다. [구분보기](#)

0.0  
평가된 감성태그가 4개 없습니다  
(0개의 리뷰)

종이책 115,430원 원서/번역서 54,000원  
3% 115,430원 119,000원 [+ 할인쿠폰](#)  
적립/혜택 3,470P  
배송안내 무료배송 1  
새벽배송 내일(2/2, 금 오전 7시 전) 도착  
서울시 종로구 종로 1 변경 >  
알림 신청하시면 원하는 정보는 받아 보실 수 있습니다. [알림신청](#)  
[매장 재고-위치](#)

# Codes

- <https://github.com/ageron/handson-ml3>

Why GitHub? Team Enterprise Explore Marketplace Pricing Search Sign in Sign up

ageron / handson-ml2 Notifications Star 16.5k Fork 7.8k

Code Issues 128 Pull requests 1 Actions Projects Wiki Security Insights

master 1 branch 0 tags Go to file Code

ageron Add explanations for the first convolutional layer example 66caaa9 on 1 Jul 841 commits

File/Folder	Description	Updated
.github/ISSUE_TEMPLATE	Remove redundant issue templates	5 months ago
datasets	Fix vertical bars	2 years ago
docker	update environment gpu support	5 months ago
images	Add breakout.gif	2 years ago
.gitignore	Add .vscode to .gitignore	6 months ago
01_the_machine_learning_landscape.L...	Remove redundant code examples	2 months ago
02_end_to_end_machine_learning_pro...	Replace 'Open in Colab' button	3 months ago
03_classification.ipynb	Replace 'Open in Colab' button	3 months ago
04_training_linear_models.ipynb	Replace 'Open in Colab' button	3 months ago

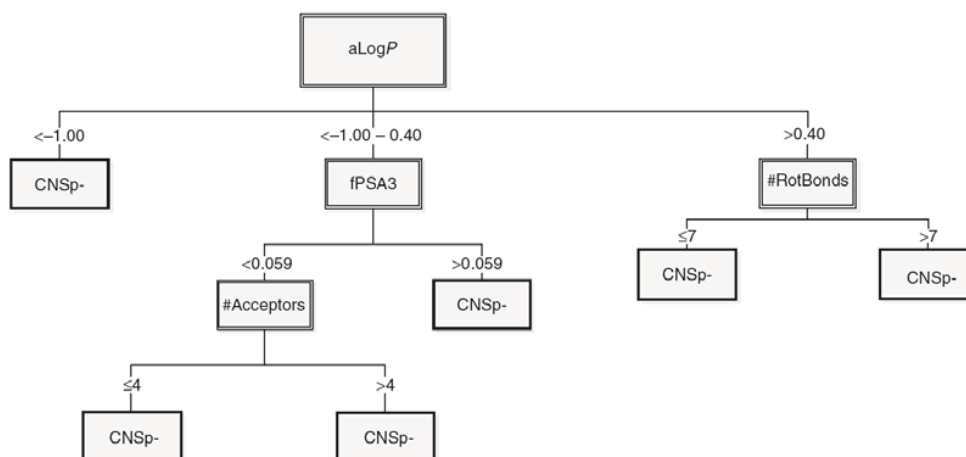
About  
A series of Jupyter notebooks that walk you through the fundamentals of Machine Learning and Deep Learning in Python using Scikit-Learn, Keras and TensorFlow 2.  
Readme  
Apache-2.0 License  
Releases  
No releases published  
Packages  
No packages published

# 기계학습법 (Machine learning)

- Simple methods
  - Linear regression-based methods
  - Decision tree
  - k-nearest neighbor (kNN)
- Nonlinear methods
  - Random Forest
  - XGboost
  - Support vector machine (SVM)
- Deep learning methods
  - Deep neural network
  - Convolutional neural network
  - Recurrent neural network
  - Graph neural network

## Decision Tree

- Decision trees are another interpretable approach to QSAR modeling that produce predictions by applying a series of descriptor-based rules to a compound.

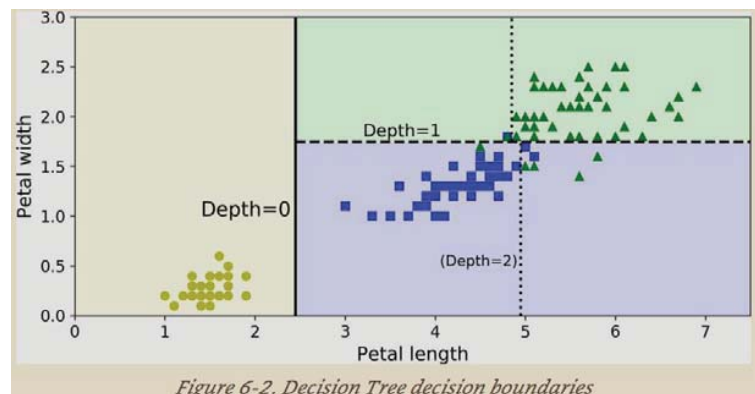
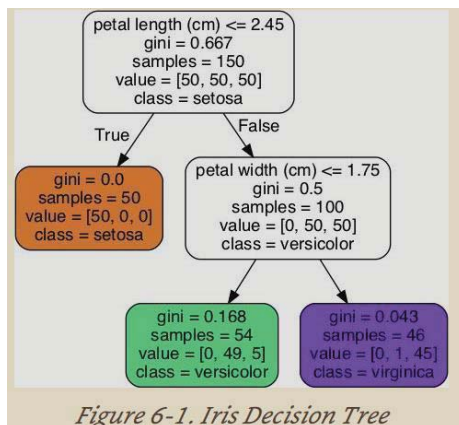


# Example

```
from sklearn.datasets import load_iris
from sklearn.tree import DecisionTreeClassifier

iris = load_iris()
X = iris.data[:, 2:] # petal length and width
y = iris.target

tree_clf = DecisionTreeClassifier(max_depth=2)
tree_clf.fit(X, y)
```

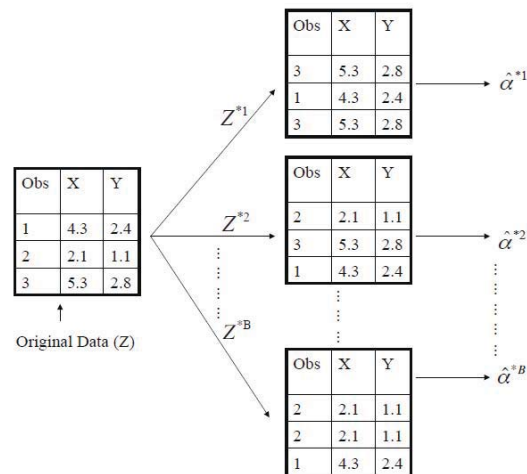


## Pros and Cons

- Tree-based methods are simple and useful for interpretation.
- However, they typically are not competitive with the best supervised learning approaches in terms of prediction accuracy.
- Bagging, random forests, and boosting methods grow multiple trees which are then combined to yield a single consensus prediction.
- Combining a large number of trees can often result in dramatic improvements in prediction accuracy, at the expense of some loss interpretation.

# Bootstrapping

- Obtain distinct data sets by repeatedly sampling observations from the original data set with *replacement*.
- Each of the “bootstrap data sets” is the same size as our original dataset.



# Bagging

- **Bootstrap aggregation, or bagging**

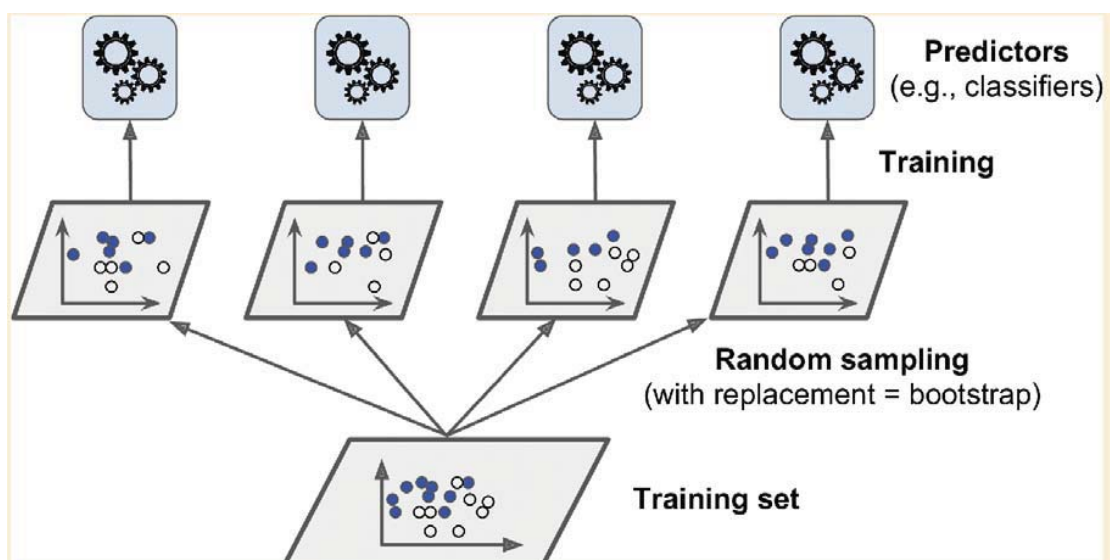
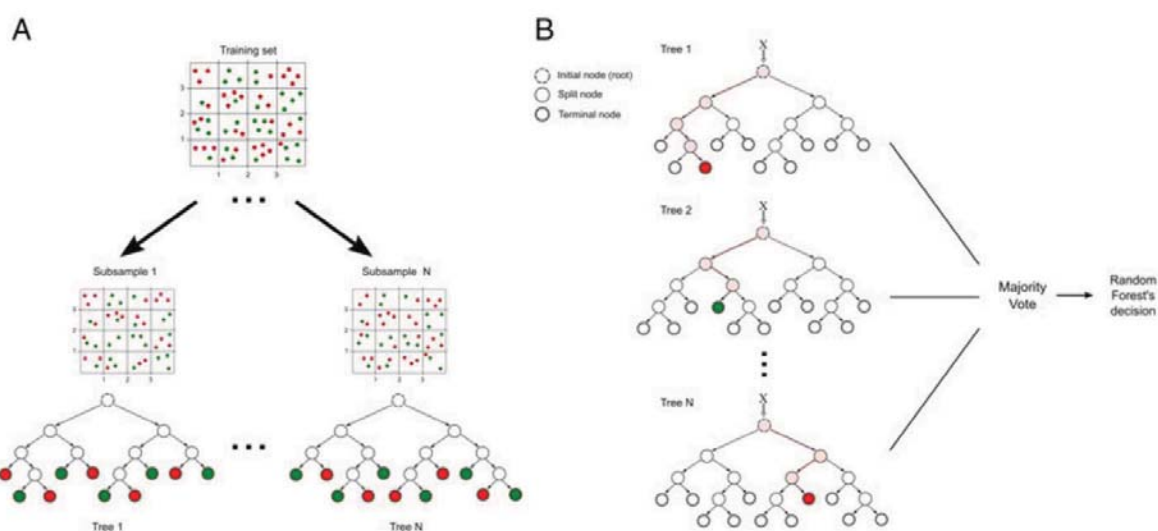


Figure 7-4. Bagging and pasting involves training several predictors on different random samples of the training set

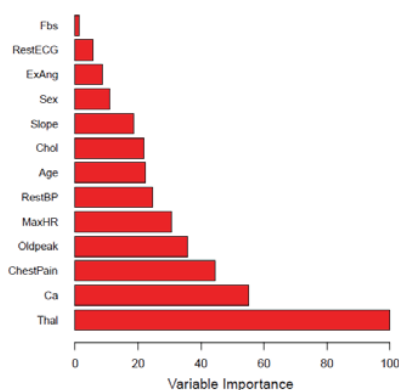
# Random Forest

- Become the industry standard method for generating global QSAR models.



## Variable importance measure

- For bagged/RF regression trees, we record the total amount that the RSS is decreased due to splits over a given predictor, averaged over all  $B$  trees.
- A large value indicates an important predictor.
- Similarly, for bagged/RF classification trees, we add up the total amount that the Gini index is decreased by splits over a given predictor, averaged over all  $B$  trees.



Variable importance plot for the **Heart** data

# RF Codes

```
from sklearn.ensemble import RandomForestClassifier

rnd_clf = RandomForestClassifier(n_estimators=500, max_leaf_nodes=16,
n_jobs=-1)
rnd_clf.fit(X_train, y_train)

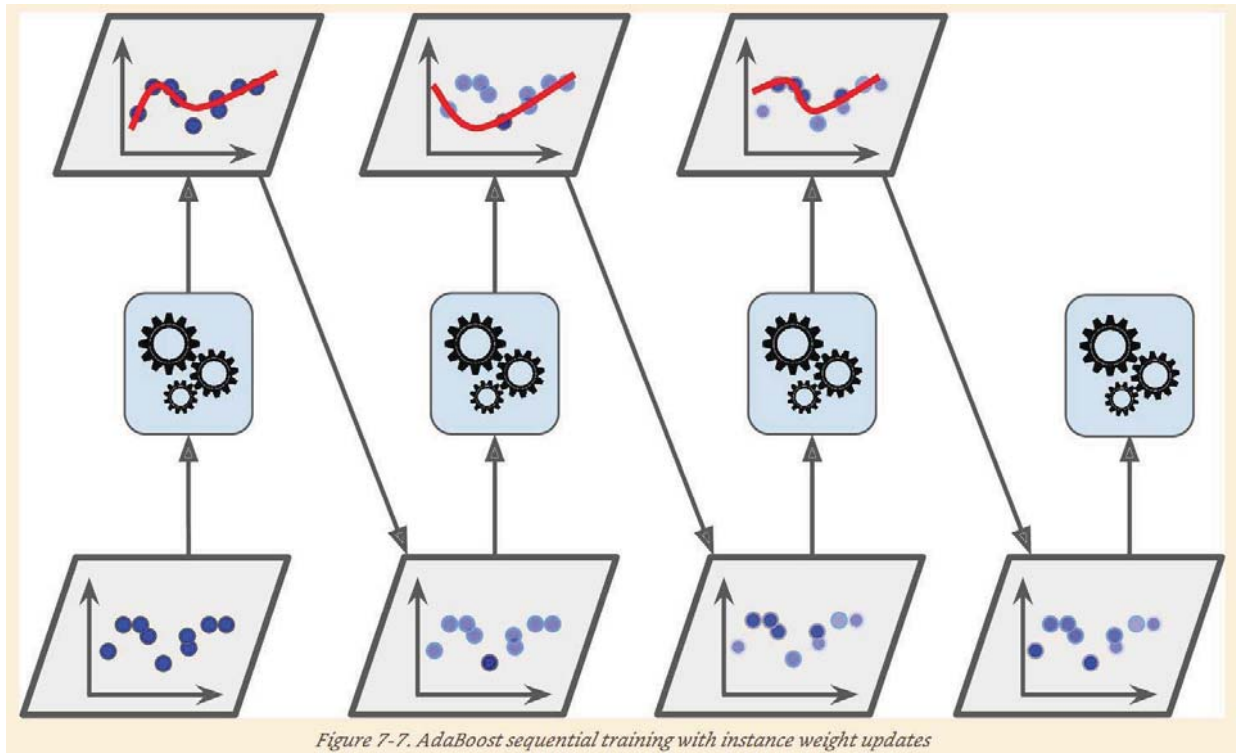
y_pred_rf = rnd_clf.predict(X_test)
```

```
>>> from sklearn.datasets import load_iris
>>> iris = load_iris()
>>> rnd_clf = RandomForestClassifier(n_estimators=500, n_jobs=-1)
>>> rnd_clf.fit(iris["data"], iris["target"])
>>> for name, score in zip(iris["feature_names"],
rnd_clf.feature_importances_):
...     print(name, score)
...
sepal length (cm) 0.112492250999
sepal width (cm) 0.0231192882825
petal length (cm) 0.441030464364
petal width (cm) 0.423357996355
```

# Boosting

- Bagging involves creating multiple copies of the original training data set using the bootstrap, fitting a separate decision tree to each copy, and then combining all of the trees in order to create a single predictive model.
- Notably, each tree is built on a bootstrap data set, independent of the other trees.
- Boosting works in a similar way, except that the trees are grown sequentially: each tree is grown using information from previously grown trees.

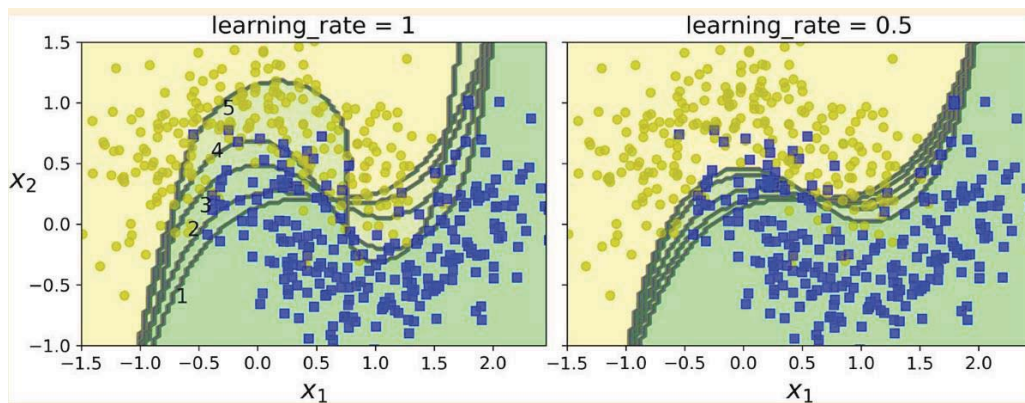
# AdaBoost



# AdaBoost Code

```
from sklearn.ensemble import AdaBoostClassifier

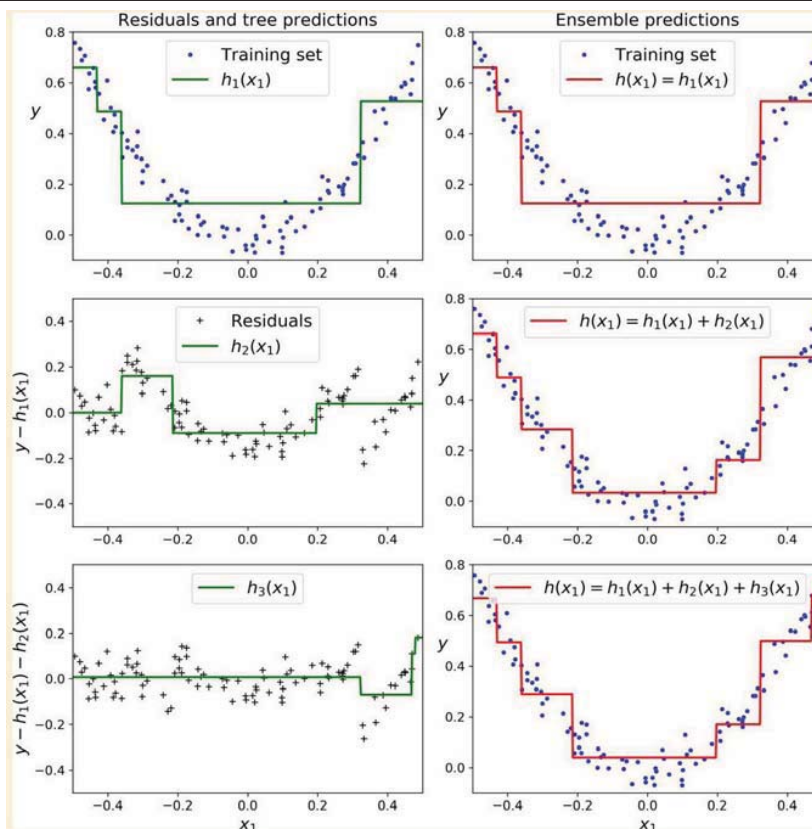
ada_clf = AdaBoostClassifier(
    DecisionTreeClassifier(max_depth=1), n_estimators=200,
    algorithm="SAMME.R", learning_rate=0.5)
ada_clf.fit(X_train, y_train)
```





# Gradient Boosting

- Just like AdaBoost, Gradient Boosting works by sequentially adding predictors to an ensemble, each one correcting its predecessor.
- However, instead of tweaking the instance weights at every iteration like AdaBoost does, this method tries to fit the new predictor to the residual errors made by the previous predictor.



```
from sklearn.ensemble import GradientBoostingRegressor
```

```
gbrt = GradientBoostingRegressor(max_depth=2, n_estimators=3, learning_rate=1.0)  
gbrt.fit(X, y)
```



# XGBoost

- Extreme Gradient Boosting
- Very popular, and known to be accurate

```
import xgboost

xgb_reg = xgboost.XGBRegressor()
xgb_reg.fit(X_train, y_train)
y_pred = xgb_reg.predict(X_val)
```

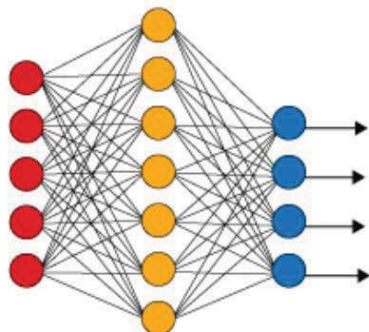
- XGBoost also offers several nice features, such as automatically taking care of early stopping:

```
xgb_reg.fit(X_train, y_train,
            eval_set=[(X_val, y_val)], early_stopping_rounds=2)
y_pred = xgb_reg.predict(X_val)
```

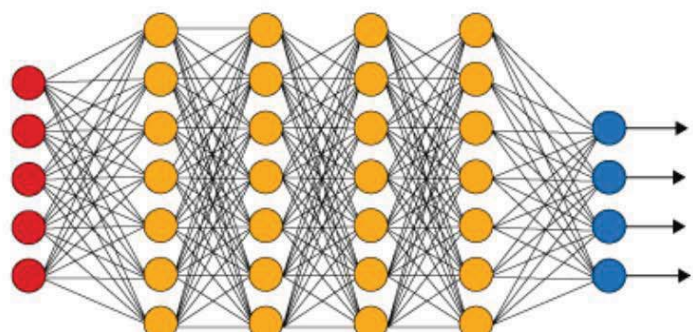
- <https://www.kaggle.com/stuarthallows/using-xgboost-with-scikit-learn>

# Deep learning methods

Simple Neural Network



Deep Learning Neural Network



● Input Layer    ● Hidden Layer    ● Output Layer

# Drug Discovery

## The rise of deep learning in drug discovery

Hongming Chen<sup>1</sup>, Ola Engkvist<sup>1</sup>, Yin Hai Wang<sup>2</sup>, Marcus Olivecrona<sup>1</sup> and Thomas Blaschke<sup>1</sup>



<sup>1</sup> Hit Discovery, Discovery Sciences, Innovative Medicines and Early Development Biotech Unit, AstraZeneca R&D Gothenburg, Mölndal 43183, Sweden

<sup>2</sup> Quantitative Biology, Discovery Sciences, Innovative Medicines and Early Development Biotech Unit, AstraZeneca, Unit 310, Cambridge Science Park, Milton Road, Cambridge CB4 0WG, UK

Over the past decade, deep learning has achieved remarkable success in various artificial intelligence research areas. Evolved from the previous research on artificial neural networks, this technology has shown superior performance to other machine learning algorithms in areas such as image and voice recognition, natural language processing, among others. The first wave of applications of deep learning in pharmaceutical research has emerged in recent years, and its utility has gone beyond bioactivity predictions and has shown promise in addressing diverse problems in drug discovery. Examples will be discussed covering bioactivity prediction, *de novo* molecular design, synthesis prediction and biological image analysis.

Drug Discovery Today, 23:1241 (2018)

# Merck Molecular Activity Challenge

Kaggle

Search

Sign In

- Home
- Competitions
- Datasets
- Code
- Discussions
- Courses
- More

Featured Prediction Competition

## Merck Molecular Activity Challenge

\$40,000

Prize Money

Help develop safe and effective medicines by predicting molecular activity.

236 teams 9 years ago

Overview Data Code Discussion Leaderboard Rules

Overview

Description

Help enable the development of safe, effective medicines.

Prizes

When **developing new medicines** it is important to identify molecules that are highly active toward their intended targets but not toward other targets that might cause side effects. The objective of this competition is to identify the best statistical techniques for predicting biological activities of different molecules, both on- and off-target, given numerical descriptors generated from their chemical structures.

Visualization-Prospect

Submission-Instructions

Winners

The challenge is based on 15 molecular activity data sets, each for a biologically relevant target. Each row corresponds to a molecule and contains descriptors derived from that molecule's chemical structure.

In addition to the prediction competition, Merck is also hosting a **visualization challenge** with a \$2,000 prize for the most insightful and elegant graphical representations of the data.

Prizes total \$40,000.

Launch

9 years ago

Close

9 years ago

236 Teams 269 Competitors 2,979 Entries

Points This competition awarded **ranking points**  
Tiers This competition counted towards **tiers**

# Winner

essentially creating 13 difficult prediction tasks in one.

## An In-the-Wild Test of Deep Learning

Competition was intense, with more than 2900 entries in just 60 days. The winners, a group of Kaggle newcomers led by graduate student George Dahl, used a deep learning model originally developed for speech recognition. The winners demonstrated that deep learning—a powerful form of artificial neural network, based on the way that the human brain learns and represents information—could provide accurate predictions with no domain specific expertise or data preprocessing. The winning result represented a 17% improvement over an industry standard benchmark and was the first time that deep learning won a Kaggle competition, opening exciting new avenues for computer-aided pharmaceutical research.

Further reading—

Industry domain	Pharmacology
Data Type	Anonymized molecular structure and activity data
Task	Predict activity levels between molecules and biologically relevant targets
Participants	269 participants on 236 teams
No. of entries	2979
Length of competition	60 days
Winning Method	Deep learning neural networks
Prizes	\$40,000

# AI 신약개발 (Deep Learning 모델)

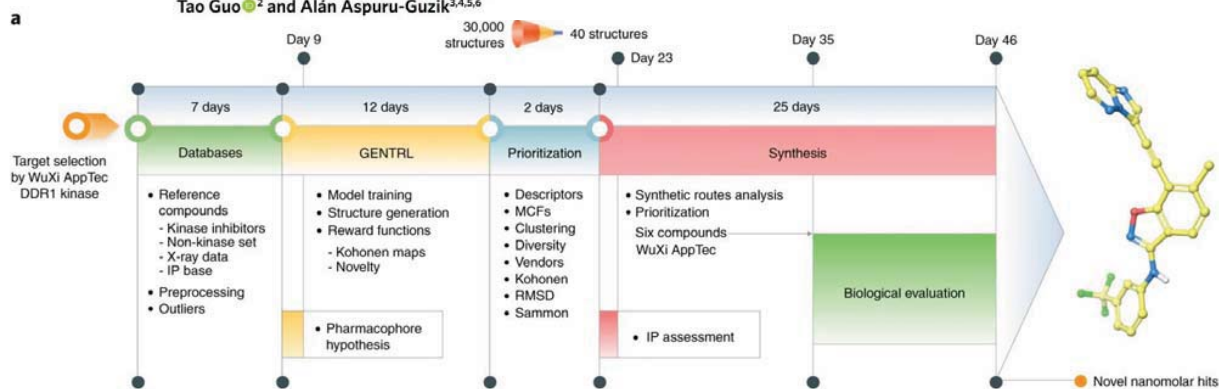
BRIEF COMMUNICATION

<https://doi.org/10.1038/s41587-019-0224-x>

nature  
biotechnology

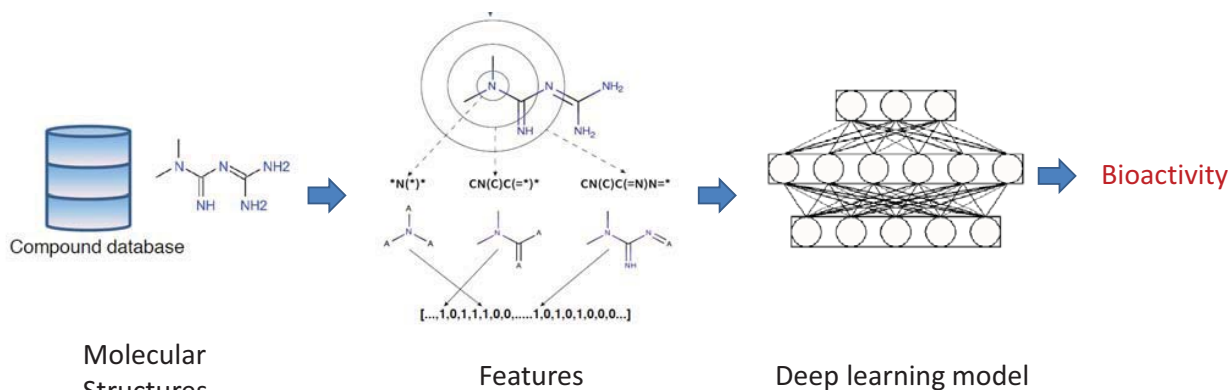
## Deep learning enables rapid identification of potent DDR1 kinase inhibitors

Alex Zhavoronkov<sup>1\*</sup>, Yan A. Ivanenkov<sup>1</sup>, Alex Aliper<sup>1</sup>, Mark S. Veselov<sup>1</sup>, Vladimir A. Aladinskiy<sup>1</sup>, Anastasiya V. Aladinskaya<sup>1</sup>, Victor A. Terentiev<sup>1</sup>, Daniil A. Polykovskiy<sup>1</sup>, Maksim D. Kuznetsov<sup>1</sup>, Arip Asadulaev<sup>1</sup>, Yury Volkov<sup>1</sup>, Artem Zholus<sup>1</sup>, Rim R. Shayakhmetov<sup>1</sup>, Alexander Zhebrak<sup>1</sup>, Lidiya I. Minaeva<sup>1</sup>, Bogdan A. Zagribelnyy<sup>1</sup>, Lennart H. Lee<sup>2</sup>, Richard Soll<sup>2</sup>, David Madge<sup>2</sup>, Li Xing<sup>2</sup>, Tao Guo<sup>2</sup> and Alán Aspuru-Guzik<sup>3,4,5,6</sup>



# Simple Deep learning model

- QSAR Procedure



- Issues

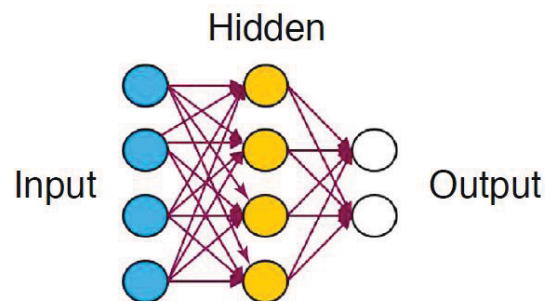
- Featurization 방법
- DL 모델

## Deep Learning

- Conventional machine learning methods for drug discovery.
  - SVM, neural networks, and random forest (RF)
- A difference between most other machine learning methods and DL is the flexibility of the NN architecture in DL.
  - fully connected feed-forward networks (FNN)
  - convolutional neural networks (CNN)
  - recurrent neural networks (RNN)
  - graph convolutional network (GCN)

# Principles of deep learning

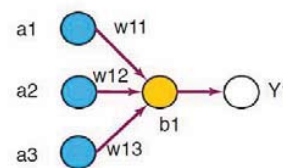
- DL uses artificial neural networks (ANNs) with many layers of nonlinear processing units for learning data representations.
- Three basic layers
  - input layer, hidden layer and output layer



# Principles of deep learning

- The interrelationship between input and output values of a hidden unit.  $Y_i$ :

$$Y_i = g \left( \sum_j W_{ij} * a_j \right)$$



- $a_j$ : the input variables
  - $W_{ij}$ : weight of input node  $j$  on node  $i$
  - $g$ : activation function, which is normally a nonlinear function (e.g., sigmoid or relu)
- The training of an ANN is done by iterative modification of the weight values through the **back-propagation** methods.

# Principles of deep learning

- Problems of traditional ANN
  - Overfitting
  - Vanishing gradients
- Algorithmic improvements in DL:
  - Dropout to address overfitting problem
  - Rectified linear unit (ReLU) to avoid vanishing gradients
  - Many novel network architectures
- Most of the DL software packages are open-sourced
  - TensorFlow, PyTorch
- Hardware: GPU, TPU
- Data, Data, Data

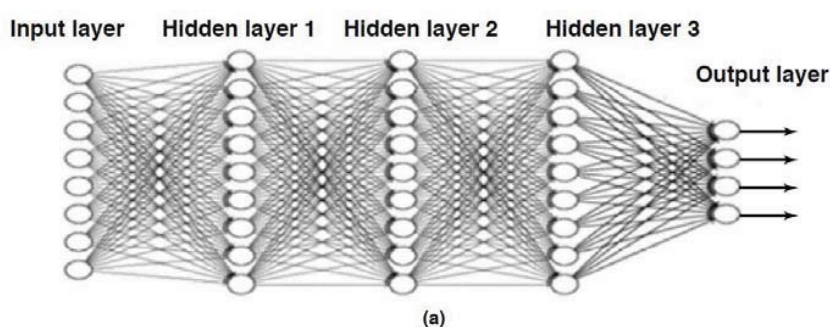
# Popular Architectures

- Fully connected deep neural network (FCN)
- Convolutional neural network (CNN)
- Recurrent neural network (RNN)
- Graph convolutional network (GCN)
- Autoencoder (AE)



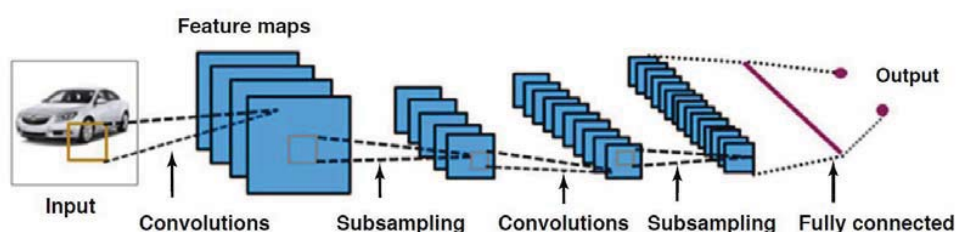
## Fully connected deep neural network (FCN)

- Contains multiple hidden layers and each layer comprises hundreds of nonlinear process units
- FCNs can take large numbers of input features.
- **Molecular Features: Fingerprint**



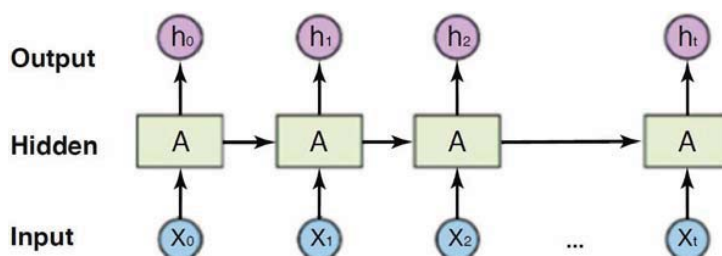
## Convolutional neural network (CNN)

- Contains several convolution layers and subsampling layers
- The convolution layer consists of a set of filters (or kernels).
- Each filter is convoluted across the width and height of the input volume.
- The subsampling layer is used to reduce the size of feature maps.
- Owing to sharing the same parameters for each filter, a CNN largely reduces the number of free parameters learned.
- It has outperformed other types of machine learning algorithms in image recognition
- **Molecular feature: 2D connection table, SMILES**



# Recurrent neural network (RNN)

- RNNs can take sequential data as input features, which is very suitable for time-dependent tasks like language modeling.
- Using a technology called long short term memory (LSTM), RNNs can reduce the vanishing gradient problem.
- **Molecular feature: SMILES**

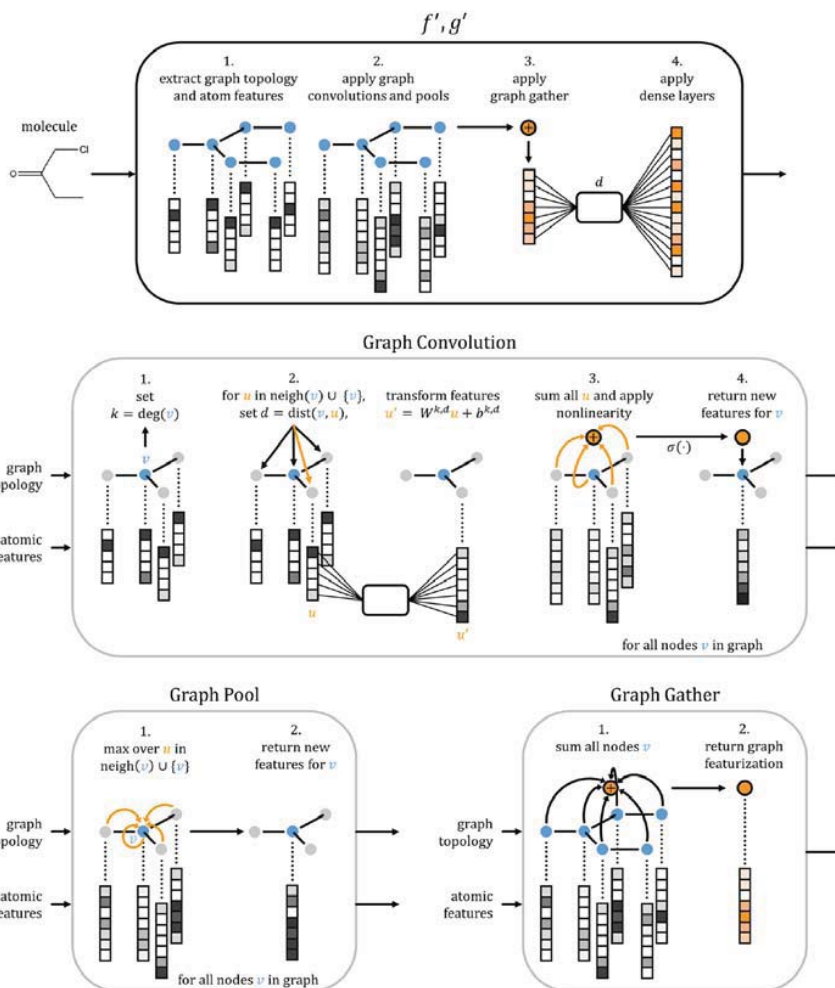


# Graph convolution

- Inspired by the Morgan circular fingerprint method
- First, the 2D molecular structure is read to form a state matrix, containing atom and bond information for each atom (Graph)
- The state matrix then goes through a convolution operation to generate a fixed length vector as the molecular representation.
- **Molecular feature: Graph**

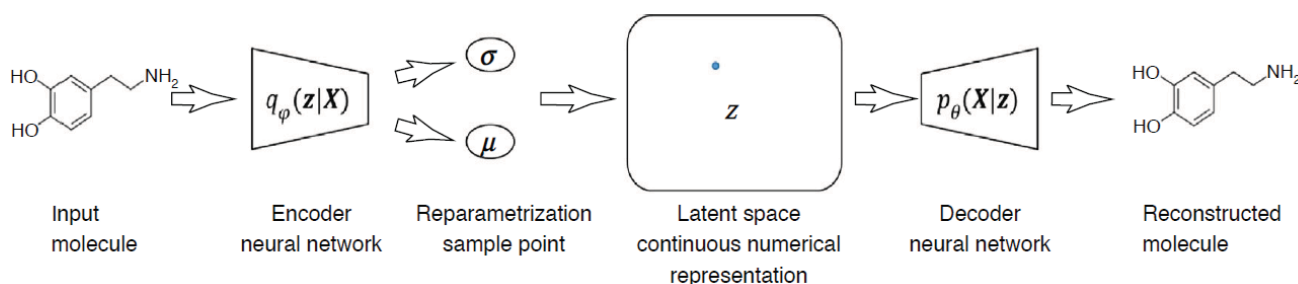






## De novo design

- Generation of new chemical structures
- Variational autoencoder (VAE) to generate chemical structures
  - Use VAE to do unsupervised learning to map chemical structures (SMILES strings) in the ZINC database into latent space
  - Latent vector in the latent space becomes a continuous representation of molecular structure
  - and can be reversibly transformed to a SMILES string through the trained VAE
  - Generation of a new structure with desirable properties



Fork me on GitHub

DeepChem is  
a Python library democratizing deep learning for  
science.

Get Started

OS

Linux

OSX

<https://deepchem.io/>

# Tensorflow

- <https://www.tensorflow.org>



TensorFlow

실지 학습 API 리소스 더보기

Language GitHub 로그인

TF 2.10이 출시되었습니다. [버전 보기](#)

TensorFlow를 사용해 프로덕션급 머신러닝 모델 만들기

선형 학습된 모델을 사용하거나 직접 모델을 학습시키기

다양한 실력 수준에 맞는 ML 솔루션 찾아 보기

연구에서 프로덕션 단계로 나아가기

[TensorFlow 알아보기](#) [생태계 살펴보기](#)

# QSAR example: HIV datasets

- The HIV dataset:
  - Ability to inhibit HIV replication for over 40,000 compounds.
- Classification task between inactive (CI) and active (CA and CM)
- The raw data csv file contains columns below:
  - “smiles”: SMILES representation of the molecular structure
  - “HIV\_active”: Binary labels for screening results: 1 (CA/CM) and 0 (CI)
- Total 41913, #pos = 1487: highly imbalanced dataset
- [https://colab.research.google.com/drive/1r4qF7DAw56\\_9umrsV3k](https://colab.research.google.com/drive/1r4qF7DAw56_9umrsV3k)

smiles	activity	HIV_active
<chem>CCC1=[O+][Cu-3]2([O+]=C(CC)C1)[O+]=C(CCI</chem>		0
<chem>C(=Cc1cccc1)C1=[O+][Cu-3]2([O+]=C(C=Cc</chem>		0
<chem>CC(=O)N1c2cccc2Sc2c1ccc1cccc21</chem>	CI	0
<chem>Nc1ccc(C=Cc2ccc(N)cc2S(=O)(=O)O)c(S(=O)</chem>	CI	0
<chem>O=S(=O)(O)CCS(=O)(=O)O</chem>	CI	0
<chem>CCOP(=O)(Nc1cccc(Cl)c1)OCC</chem>	CI	0

## Virtual Screening

# Virtual Screening

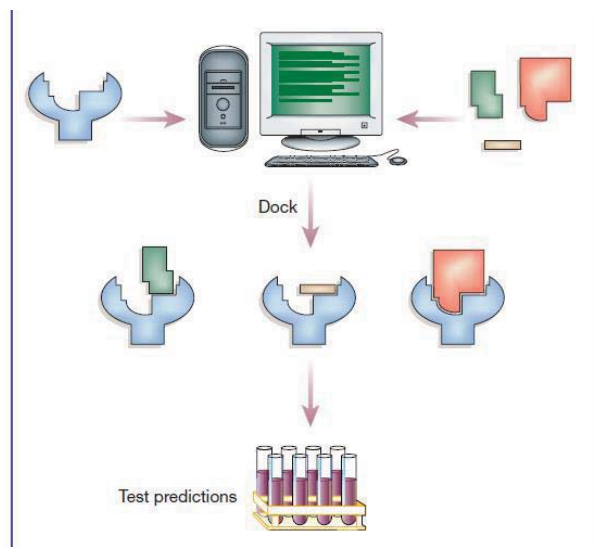


Table 1 Hit rates and drug-like properties for inhibitors discovered with high-throughput and virtual screening against the enzyme PTP-1B (ref.19)

Technique	Compounds tested	Hits with $IC_{50} < 100\mu M$	Hits with $IC_{50} < 10\mu M$	Lipinski compliant hits*	Hit rate†
HTS	400,000	85	6	23	0.021%
Docking	365‡	127	18	57	34.8%

\*Number of 100  $\mu M$  or better inhibitors that passed all four of the drug-like criteria identified in Lipinski's 'rule of five'<sup>25</sup>; †The number of compounds experimentally tested divided by the number of compounds with  $IC_{50}$  values of 100  $\mu M$  or less; ‡The number of top-scoring docking hits that were experimentally tested;  $IC_{50}$ , The concentration of inhibitor at which the enzyme is 50% inhibited.

## 리간드 기반 신약 발굴

- Ligand-based Virtual Screening
- Procedure
  - 타겟 선정
  - 타겟 단백질에 관한 정보 수집
  - ChEMBL (or BindingDB) 에서 화합물 데이터 수집
  - Binding affinity 예측 모델 개발 (QSAR)
  - ZINC에서 화합물 라이브러리 구축
  - Virtual screening으로 후보물질 선정
  - Docking 계산, Visual inspection 등을 거쳐 최종 후보물질 발굴

# 타겟

LAIDD-Practice3-Predicting\_pIC50\_of\_JAK2\_inhibitors.ipynb

File Edit View Insert Runtime Tools Help Changes will not be saved

+ Code + Text Copy to Drive

## Predicting activity of JAK2 inhibitors

### Goal of the class

- Practice the regression model using biological data

### Janus kinase

Janus kinase (JAK) is a family of intracellular, non-receptor tyrosine kinases that transduce cytokine-mediated signals via the JAK-STAT pathway. They were initially named "just another kinase" 1 and 2 (since they were just two of many discoveries in a PCR-based screen of kinases),[1] but were ultimately published as "Janus kinase". The name is taken from the two-faced Roman god of beginnings, endings and duality, Janus, because the JAKs possess two near-identical phosphate-transferring domains. One domain exhibits the kinase activity, while the other negatively regulates the kinase activity of the first.

# 타겟 단백질에 관한 정보

- UniProt: <https://www.uniprot.org/uniprotkb/O60674/entry>

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced List Search Help

## O60674 · JAK2\_HUMAN

Tyrosine-protein kinase JAK2 · Homo sapiens (Human) · EC:2.7.10.2 · Gene: JAK2 · 1132 amino acids · Evidence at protein level · Annotation score: 6.3

Entry Feature viewer Publications External links History

BLAST Align Download Add Add a publication Entry feedback

### Function

Non-receptor tyrosine kinase involved in various processes such as cell growth, development, differentiation or histone modifications. Mediates essential signaling events in both innate and adaptive immunity. In the cytoplasm, plays a pivotal role in signal transduction via its association with type I receptors such as growth hormone (GHR), prolactin (PRLR), leptin (LEPR), erythropoietin (EPOR), thrombopoietin (THPO); or type II receptors including IFN-alpha, IFN-beta, IFN-gamma and multiple interleukins (PubMed:7615558).

Following ligand-binding to cell surface receptors, phosphorylates specific tyrosine residues on the cytoplasmic tails of the receptor, creating docking sites for STATs proteins (PubMed:9618263).

Subsequently, phosphorylates the STATs proteins once they are recruited to the receptor. Phosphorylated STATs then form homodimer or heterodimers and translocate to the nucleus to activate gene transcription. For example, cell stimulation with erythropoietin (EPO) during erythropoiesis leads to JAK2 autophosphorylation, activation, and its association with erythropoietin receptor (EPOR) that becomes phosphorylated in its cytoplasmic domain. Then, STAT5 (STAT5A or STAT5B) is recruited, phosphorylated and activated by JAK2. Once activated, dimerized STAT5 translocates into the nucleus and promotes the transcription of several essential genes involved in the modulation of erythropoiesis. Part of a signaling cascade that is activated by increased cellular retinol and that leads to the activation of STAT5 (STAT5A or STAT5B) (PubMed:21368206).

In addition, JAK2 mediates angiotensin-2-induced ARHGEF1 phosphorylation (PubMed:20098430).

Plays a role in cell cycle by phosphorylating CDKN1B (PubMed:21423214).

Cooperates with TEC through reciprocal phosphorylation to mediate cytokine-driven activation of FOS transcription. In the nucleus, plays a key role in chromatin by specifically mediating phosphorylation of 'Tyr-41' of histone H3 (H3Y41ph), a specific tag that promotes exclusion of CBX5 (HP1 alpha) from chromatin (PubMed:19783980). 7 Publications

### Catalytic Activity

ATP + L-tyrosyl-[protein] = ADP + H<sup>+</sup> + O-phospho-L-tyrosyl-[protein] 1 Automatic Annotation 2 Publications

EC:2.7.10.2 (UniProtKB | ENZYME | Rhea)

Source: Rhea 10596



# ChEMBL

- <https://www.ebi.ac.uk/chembl/>

ChEMBL is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs.

Explore ChEMBL

**Description:** Shows a summary of the ChEMBL entities and quantities of data for each of them.

**Instructions:** Click on a bubble to explore a specific ChEMBL entity in more detail.

Current Release: ChEMBL 31  
Provided under a Creative Commons Attribution-ShareAlike 3.0 Unported License  
Last Update on 2022-07-12T00:00:00 | Release notes

# JAK2

Search Results

All Results 1929 Compounds 1 Targets 12 Assays 1791 Documents 125 Cells 0 Tissues 0

Targets

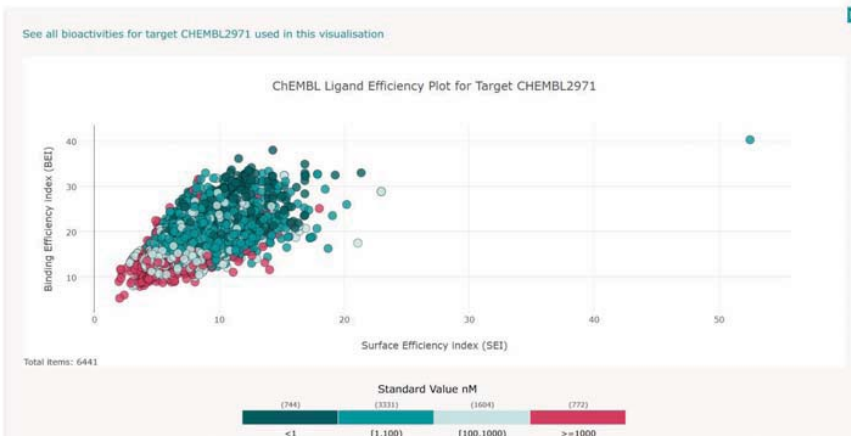
Show Full Query

8 Targets  
0 Selected - Select All  
Browse Activities

ChEMBL ID	Search Hit	Name	UniProt Accessions	Type	Organism	Compounds	Activities
CHEMBL2971		Tyrosine-protein kinase JAK2	O60674	SINGLE PROTEIN	Homo sapiens	9682	12349
CHEMBL4742263		Cereblin/Tyrosine-protein kinase JAK2	O60674, Q965W2	PROTEIN-PROTEIN INTERACTION	Homo sapiens	12	12



### Ligand Efficiencies



# Activity Data

- “csv” 다운로드 및 편집 → JAK2 ChEMBL.csv

6,451 Activities  
0 Selected - Select All  
Browse Compounds

Records per page: 20

Showing 1-20 out of 6,451 records

Molecule CHEMBL ID	Compound Key	Standard Type	Standard Relation	Standard Value	Standard Units	pCHEMBL Value	Comment	Assay CHEMBL ID	Assay Description	BAO Label	Ass Org
CHEMBL535	Suntrib	Kd	--	-410.0	nM	6.39	No Data	CHEMBL1244467	Binding affinity to JH1 catalytic domain JAK2	single protein format	No D
CHEMBL1230609	EKEL-2880/GSK-1363089	Kd	=	1500.0	nM	5.82	No Data	CHEMBL1908670	Binding constant for JAK2(JH1 domain-catalytic) kinase domain	single protein format	No D
CHEMBL1789941	JNCB18424	Kd	--	0.036	nM	10.44	No Data	CHEMBL1908670	Binding constant for JAK2(JH1 domain-catalytic) kinase domain	single protein format	No D
CHEMBL607767	EKB-569	Kd	=	2000.0	nM	5.70	No Data	CHEMBL862903	Average Binding Constant for JAK2 (Kin.Dom. 2); NA=Not Active at 10 uM	single protein format	Home
CHEMBL1908670	PKC-412	Kd	--	94.0	nM	7.03	No Data	CHEMBL1908670	Binding constant for JAK2(JH1 domain-catalytic) kinase domain	single protein format	No D

# QSAR Model 개발

- Input: Smiles
- Feature: ECFP
- Target values: pChEMBL Value
- Models: Regression model
  - Random Forest regression (Scikit-learn: RandomForestRegressor)
  - FNN (Tensorflow.keras, Deepchem)
  - Loss: Mean square error (MSE)
- Model selection:
  - Validation set

## ZINC

- <https://zinc.docking.org/>

The screenshot shows the ZINC20 website homepage. At the top, there is a navigation bar with links for ZINC, Substances, Catalogs, Tranches, Biological, and More. The main content area features a large heading 'ZINC20' and a welcome message: 'Welcome to ZINC, a free database of commercially-available compounds for virtual screening. ZINC contains over 230 million purchasable compounds in ready-to-dock, 3D formats. ZINC also contains over 750 million purchasable compounds you can search for analogs in under a minute.' To the right, there is a section for citation information, stating that ZINC is provided by the Irwin and Shoichet Laboratories in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF). Below the main content, there are three columns: 'Getting Started' with links like 'Getting Started', 'What's New', and 'About ZINC 20 Resources'; 'Ask Questions' with a list of general questions such as 'How many substances in current clinical trials have PAINS patterns?' and 'How many endogenous human metabolites are there?'; and 'ZINC20 News' with a note that 'ZINC20 has been released' and a 'Caveat Emptor' warning that the database does not guarantee the quality of any molecule for any purpose.



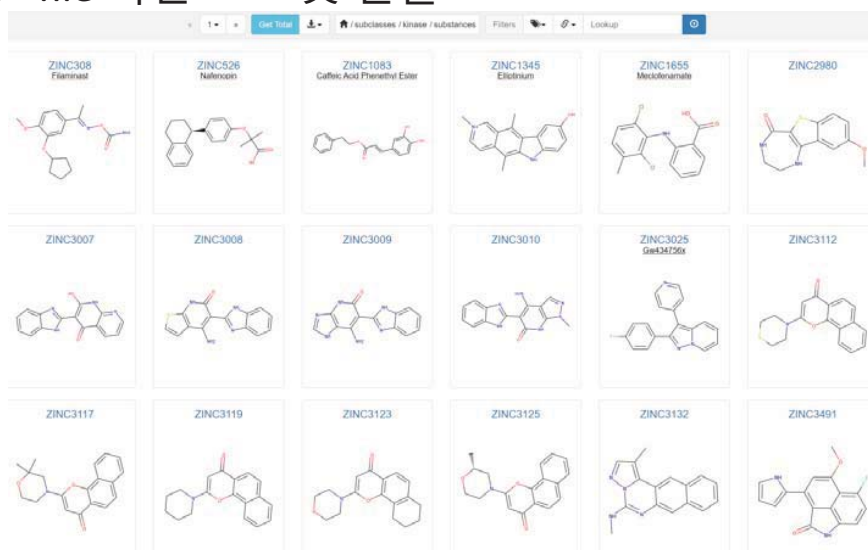
# Compound Library

- <https://zinc.docking.org/tranches/home/>

Molecular Weight (up to, Daltons)												
	200	250	300	325	350	375	400	425	450	500	>500	Totals, by LogP
-1	27,791	172,563	710,795	1,072,978	2,241,498	786,738	276,834	116,066	92,417	77,790	7,310	5,582,780
0	139,434	934,776	3,655,384	5,126,157	10,608,025	3,498,214	1,663,579	708,919	570,546	507,344	4,734	27,417,112
1	362,437	2,884,636	12,030,074	16,154,544	33,650,249	11,885,957	6,807,876	3,178,487	2,648,581	2,412,998	9,940	92,025,779
2	467,220	4,584,223	22,941,208	30,908,513	65,047,385	26,752,849	17,839,254	9,349,272	8,099,970	7,686,687	24,554	193,701,135
2.5	167,513	2,136,113	12,849,121	17,977,157	38,682,058	18,584,223	13,812,274	8,111,104	7,197,414	6,979,014	24,126	126,520,117
3	90,548	1,570,772	11,037,383	16,282,627	34,831,558	19,940,391	16,037,132	10,339,743	9,362,233	9,118,717	37,422	128,648,526
3.5	36,748	929,872	7,920,574	12,490,662	27,380,104	18,703,024	16,485,194	11,784,160	10,774,472	10,693,411	58,791	117,257,012
4	9,017	369,565	4,332,131	6,472,808	10,487,856	13,034,155	14,329,253	11,683,208	10,891,465	11,003,975	86,262	82,699,695
4.5	993	86,613	1,814,492	3,457,942	6,367,225	8,853,064	10,320,054	9,945,353	9,486,869	9,825,079	117,980	60,275,664
5	150	13,393	536,018	1,405,708	3,168,584	4,995,850	6,471,525	7,025,034	6,976,742	7,325,833	144,297	38,063,134
>5	39	1,097	22,854	103,521	376,905	927,395	1,670,856	2,195,160	2,588,702	3,052,048	767,762	11,706,339
Totals, by Weight	1,301,890	13,683,623	77,850,034	111,452,617	232,841,447	127,961,860	105,713,831	74,436,506	68,689,411	68,682,896	1,283,178	884M Substances 1.9K Tranches

# Compound Library

- Biological → Major target classes → enzyme → kinase → substances
- “csv” file 다운로드 및 변환

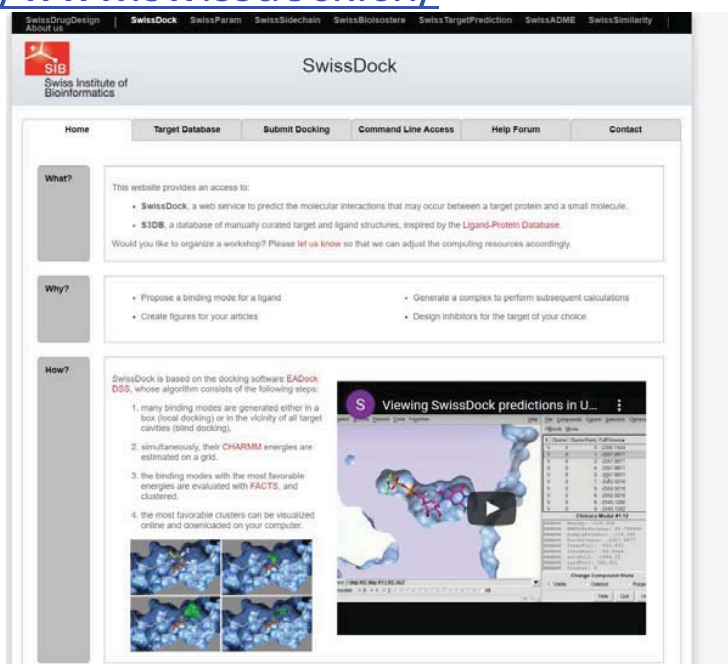


# Virtual Screening

- 개발한 QSAR regression model을 구축한 화합물 라이브러리에 적용
- Sorting
  - Prediction values
- Screening
  - 동일한 or 매우 유사 화합물 제거
  - Training data에 있는 화합물들과의 유사성 계산 (Tanimoto Coefficient or Dice Coefficient)

# Docking

- <http://www.swissdock.ch/>

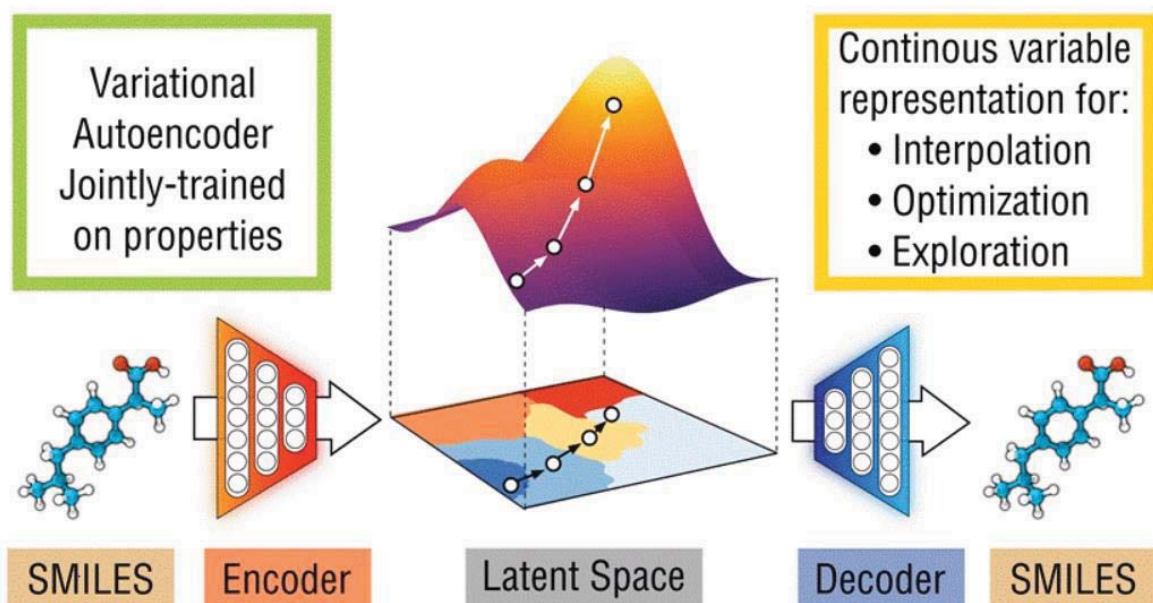


The screenshot displays the SwissDock website interface. At the top, there is a navigation bar with links for 'SwissDrugDesign', 'SwissDock', 'SwissParam', 'SwissSidechain', 'SwissBioScaffolds', 'SwissTargetPrediction', 'SwissADME', and 'SwissSimilarity'. Below this is the SIB logo and the text 'Swiss Institute of Bioinformatics'. The main navigation menu includes 'Home', 'Target Database', 'Submit Docking', 'Command Line Access', 'Help Forum', and 'Contact'. The 'What?' section explains that the website provides access to 'SwissDock', a web service for predicting molecular interactions, and 'S3DB', a database of manually curated target and ligand structures. The 'Why?' section lists reasons for using the service, such as proposing binding modes and generating complexes for calculations. The 'How?' section describes the docking process in four steps: 1. many binding modes are generated either in a box (local docking) or in the vicinity of all target cavities (blind docking); 2. simultaneously, their CHARMM energies are estimated on a grid; 3. the binding modes with the most favorable energies are evaluated with FACTS, and clustered; 4. the most favorable clusters can be visualized online and downloaded on your computer. A video player titled 'Viewing SwissDock predictions in U...' is embedded in the 'How?' section, showing a 3D molecular model of a protein-ligand complex.

## And, more

- ADME
- Toxicity 예측
- MD simulation (예, RMDS)
- Free energy ( $\Delta\Delta G$ ) 계산
- Optimization 등

## De Novo Design



# Optimization

- MORLD
- <http://morld.kaist.ac.kr/>

- Questions?