

KSBi-BIML 2023

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

생물정보학 & 머신러닝 워크샵 (오프라인)

인공지능 신약개발 AI Drug Design

김동섭 _ KAIST



본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 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 2023

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

안녕하십니까?

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

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

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

2023년 2월

한국생명정보학회장 이 인 석

강의 시간표

DAY1 (2.6 월)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	개회사/공지사항전달			
09:30-10:50 (80)	Best practice for single-cell data analysis	박종은 교수	Introduction to ML & DNN (이론)	이상근 교수
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	Practice1: Scanpy basic workflow	김우석 김성룡 조교	CNN (이론)	이상근 교수
12:10-13:40 (90)	점심 (KOBIC 세미나)			
13:40-15:10 (90)	Public data, batch correction, cell annotation	박종은 교수	RNN, GAN, XAI (이론)	이상근 교수
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	Practice2: Advanced single-cell analysis	김우석 김성룡 조교	AI 모델 구조 정의, 학습 알고리즘 적용, 성능 평가, 시각화 방법 (Tensorflow 실습)	이정현 한성민 조교

DAY2 (2.7 화)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	공지사항전달			
09:30-10:50 (80)	Introduction to protein structure prediction - Homology modeling - Coevolution-guided modeling Early AI-based approaches	백민경 교수	Pre-trained Models for Transfer Learning (이론)	전민지 교수
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	단백질 구조 예측 실습 - MSA generation, template search - homology modeling contact prediction & modeling	백민경 교수	Pre-trained Models for Transfer Learning (실습)	정민수 조교
12:10-13:40 (90)	점심			
13:40-15:10 (90)	AI-based protein structure prediction - AlphaFold/RoseTTAFold Applications to PPI prediction & protein design	백민경 교수	Deep learning in Bioinformatics	노미나 교수
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	단백질 구조 예측 실습 II AlphaFold, RoseTTAFold 실습 및 응용	백민경 교수	Deep learning model을 이용한 실습	곽호진 박예슬 조교

DAY3 (2.8 수)

시간	강 의 서울대 자연과학대학 26동B101호	강사	강 의 서울대 자연과학대학 26동B102호	강사
09:00-09:20 (20)	등록			
09:20-09:30 (10)	공지사항전달			
09:30-10:50 (80)	화학정보학 기초(Cheminformatics) 약물특성 및 약물다움(druglikeness) Molecular Notations & Descriptors AI 신약개발을 위한 Databases AI 신약개발을 위한 Programming 기초	김동섭 교수	마이크로바이옴 기본 이론	이선재 교수
10:50-11:00 (10)	휴식			
11:00-12:10 (70)	Google Colab에 RDKit 설치 화합물 정보 읽기 실습 Bioactivity database 검색 및 정보 읽기 실습 Molecular descriptor (fingerprint) 생성 및 similarity 계산 실습	문채영 나민주 조교	16S rRNA amplicon seq. - DADA2	서영창 조준우 조교
12:10-13:40 (90)	점심 (KOBIC 세미나)			
13:40-15:10 (90)	AI 신약개발을 위한 기계학습법 기초 QSAR 모델링 기초 AI 신약개발을 위한 딥러닝 모델 Virtual screening (ligand-based, structure-based) 및 de novo design	김동섭 교수	최신 메타지놈 분석 기법의 현황	이선재 교수
15:10-15:20 (10)	휴식			
15:20-16:50 (90)	QSAR modeling 전체 과정 실습 화합물의 Bioactivity 예측 모델 개발 Virtual screening 과정을 통한 신약후보물질 발굴 실습	문채영 나민주 조교	Shotgun metagenome 분석 (Linux)	서영창 조준우 조교

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

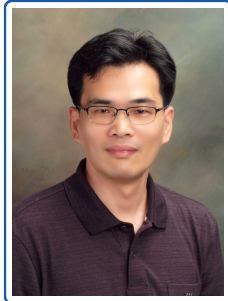
* 선수 지식: 기초 수준의 python programming

* 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 2023

인공지능 신약설계
AI Drug Design



Google Classroom

- ❑ BiML: AI 신약개발
- ❑ <https://classroom.google.com/u/o/c/NTEyMTlwODM5ODUo>
- ❑ 강의자료 및 실습용 코드 다운로드를 위해 모두 가입!



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

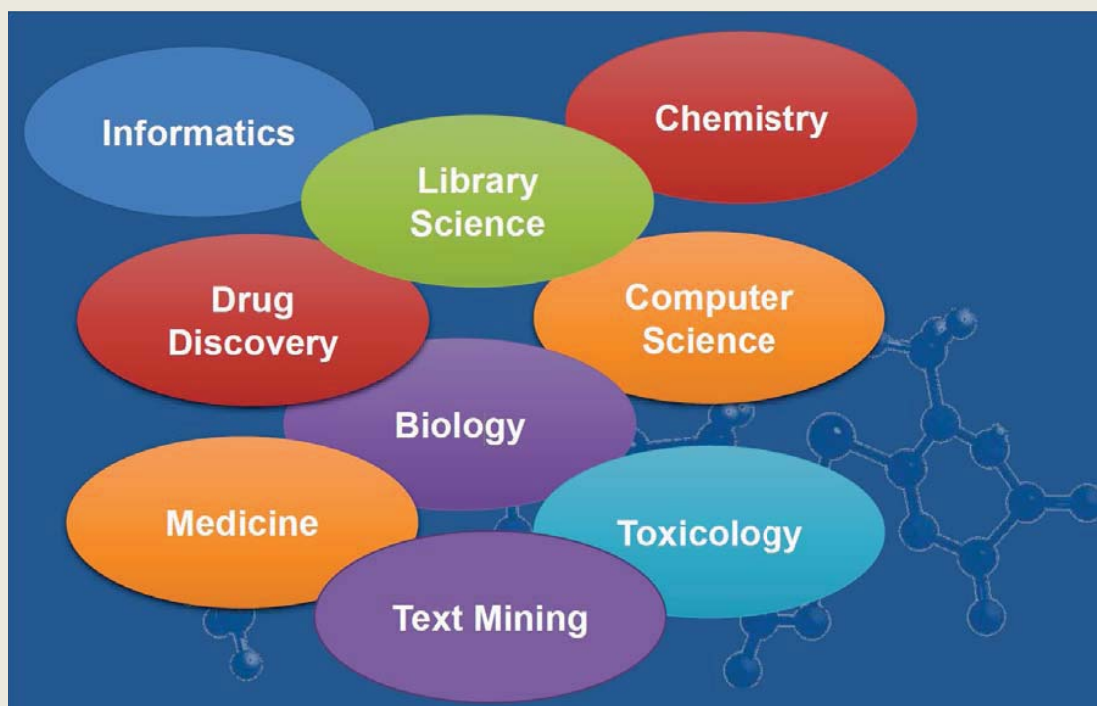


화학정보학이란

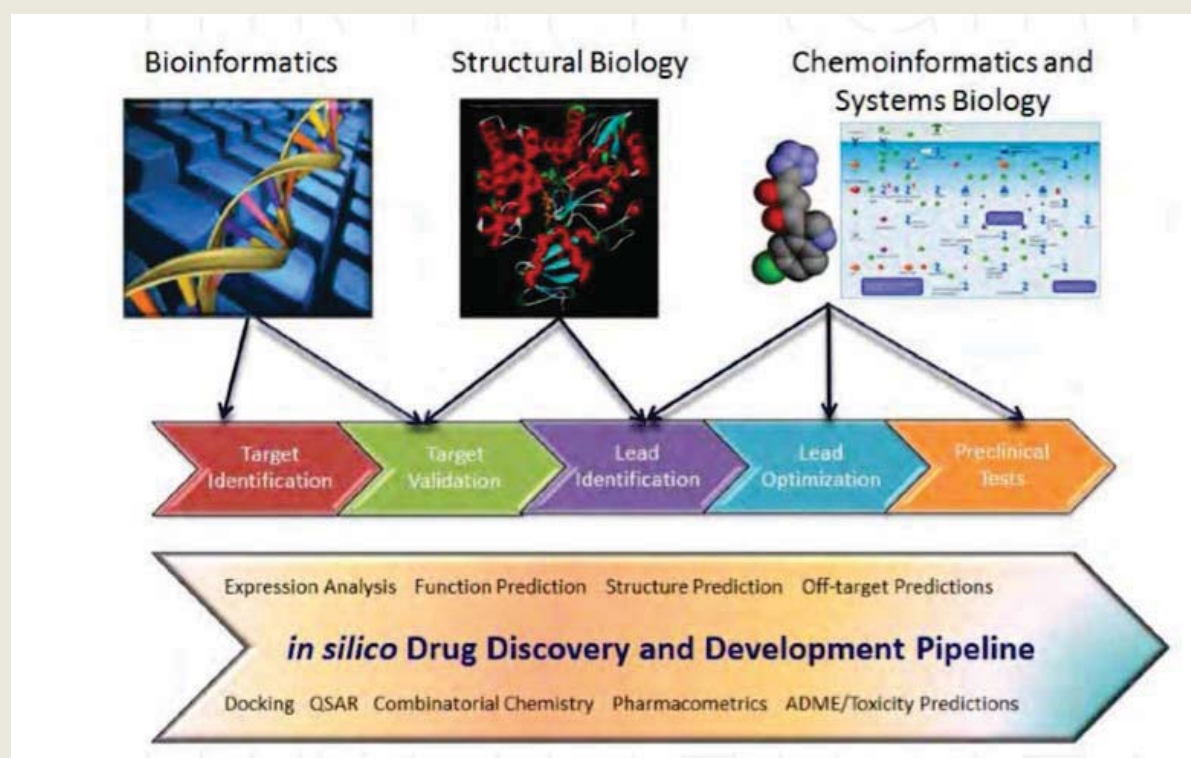
- Field of **information technology** that uses computers and computer programs to facilitate the collection, storage, analysis, and manipulation of large quantities of **chemical data**
- 여러 이름
 - Cheminformatics
 - Chemoinformatics
 - Chemical informatics
- Bioinformatics vs. Cheminformatics
 - Biological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



Interdisciplinary



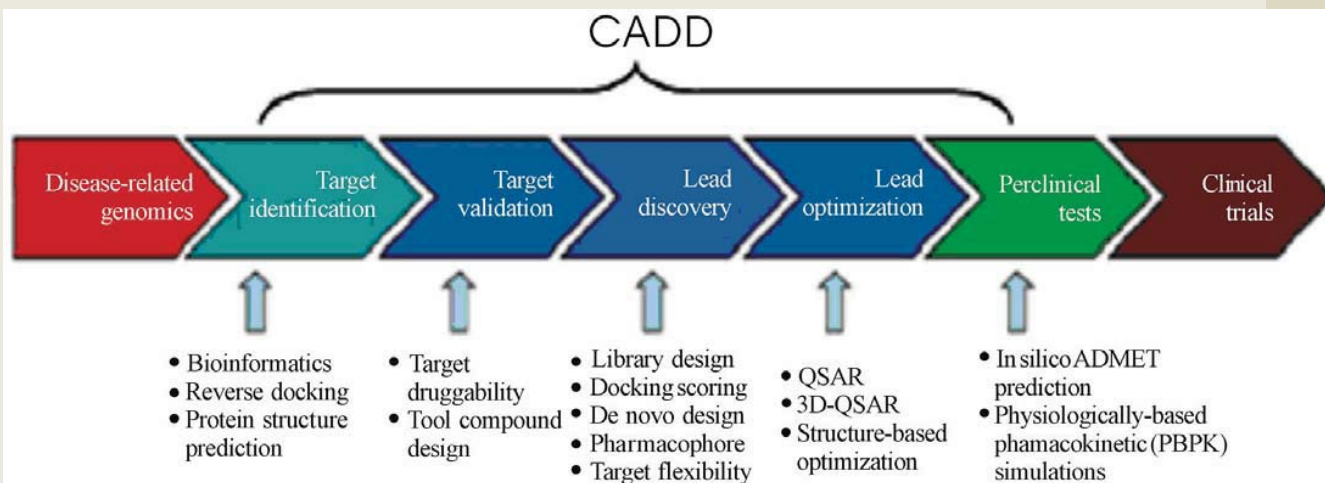
바이오정보학 vs 화학정보학





신약개발과 화학정보학

Computer-Aided Drug Design (CADD)

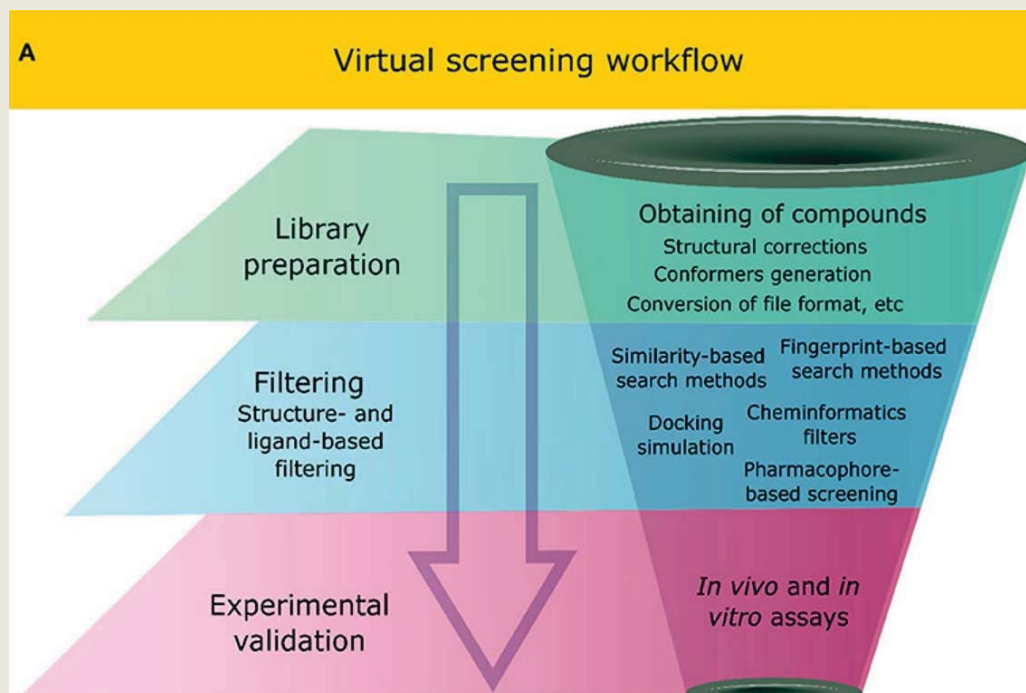


Lead Discovery & Optimization

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



가상 스크리닝



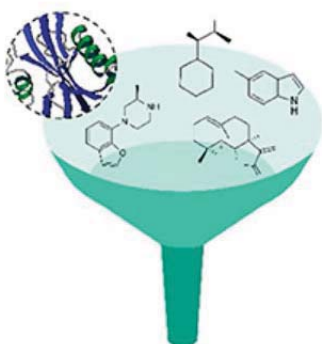
방법



B

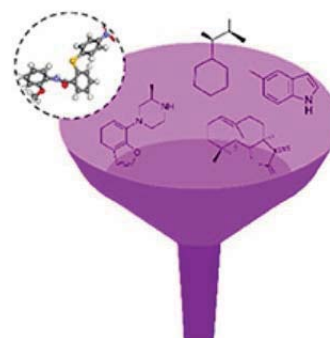
Structure-based virtual screening

- 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



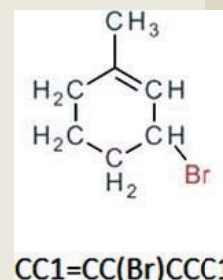
Linear Notation

- ❑ 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>



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: $\text{BH}_3 \rightarrow \text{B}$, $\text{CH}_4 \rightarrow \text{C}$, $\text{NH}_3 \rightarrow \text{N}$, $\text{H}_2\text{O} \rightarrow \text{O}$



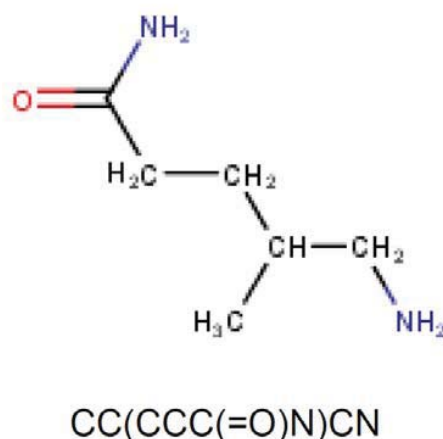
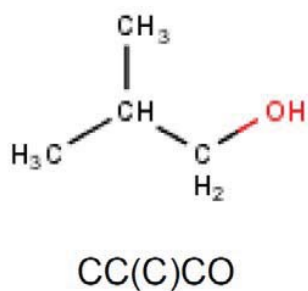
Bonds

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



Branches

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

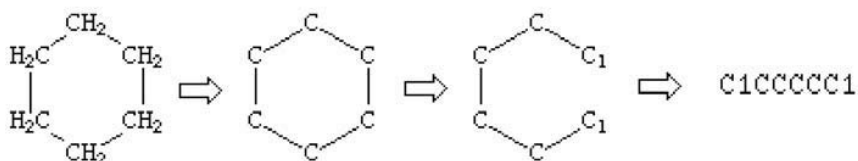




Rings

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

Cyclohexane



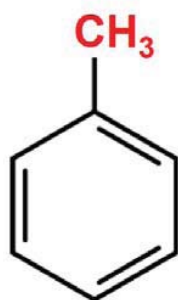
Benzene → C1=CC=CC=C1 OR c1ccccc1

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

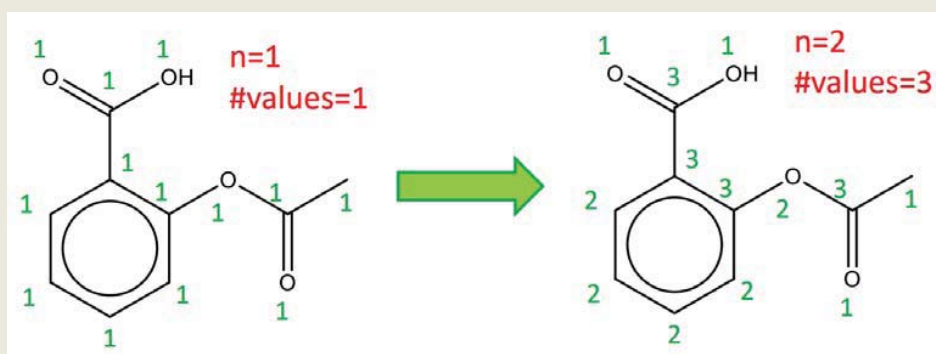
c1cccc(C)c1

c1ccccc1C



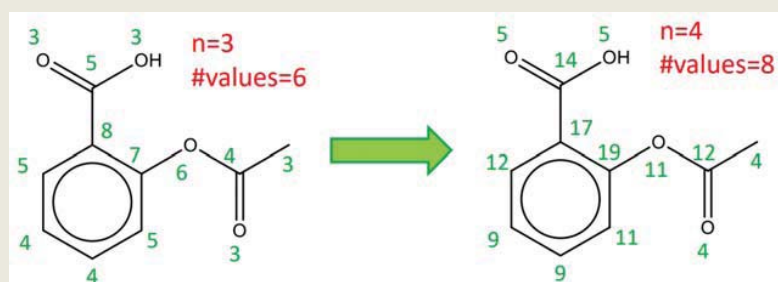
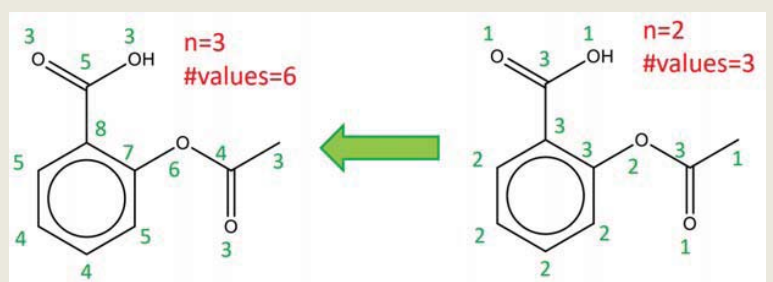
Morgan Algorithm

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



Morgan Algorithm

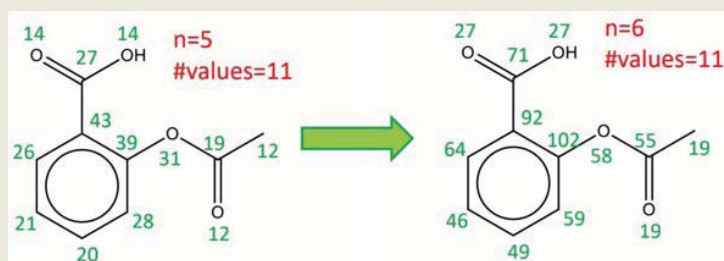
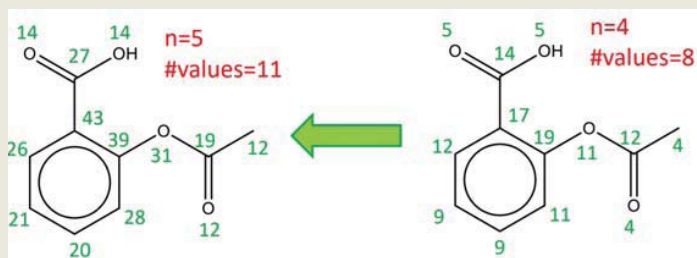
- Repeat summing of neighboring values





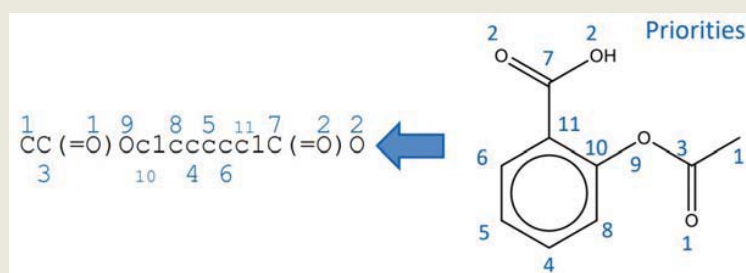
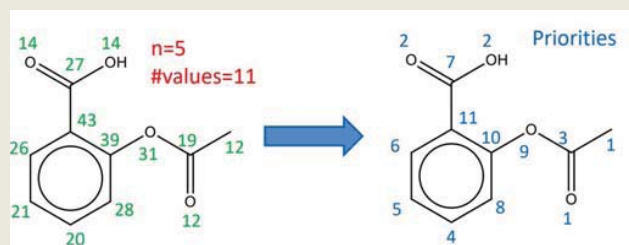
Morgan Algorithm

- 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 change 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.



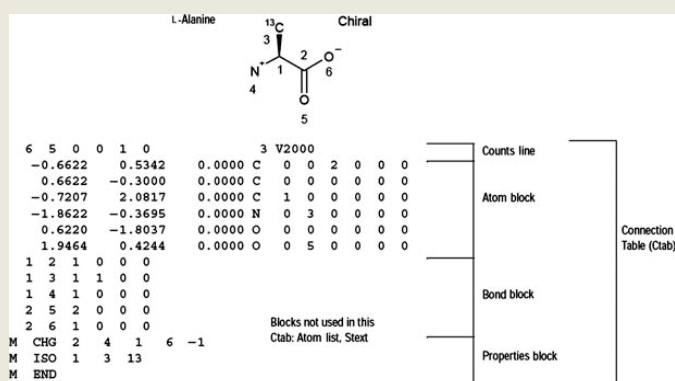
InChIKey

- ❑ 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.

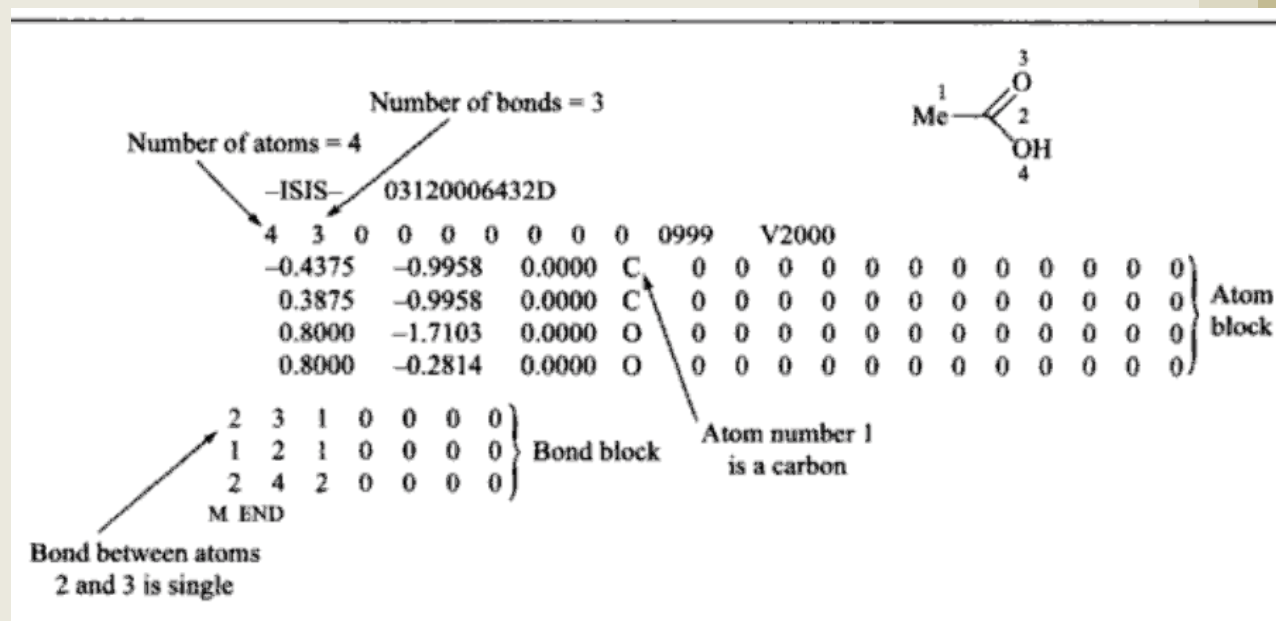


Connection Tables

- ❑ 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.



Mol file

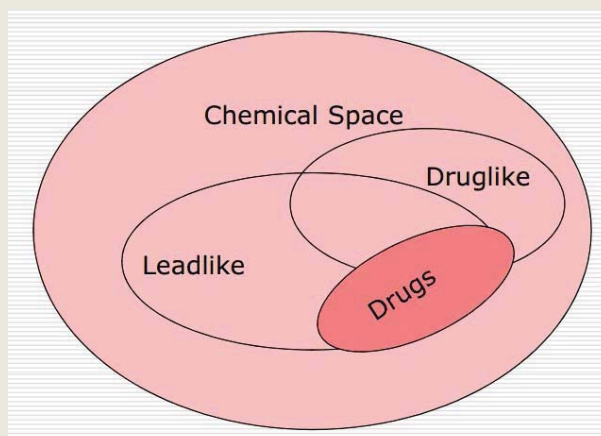


12.3: MDL mol file for acetic acid, in the hydrogen-suppressed form.



Chemical Space

- ❑ Chemical space can be viewed as being analogous to the cosmological universe.
- ❑ The total number of possible small organic molecules that populate 'chemical space' has been estimated to exceed 10^{60}
- ❑ Drug-like & Lead-like



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

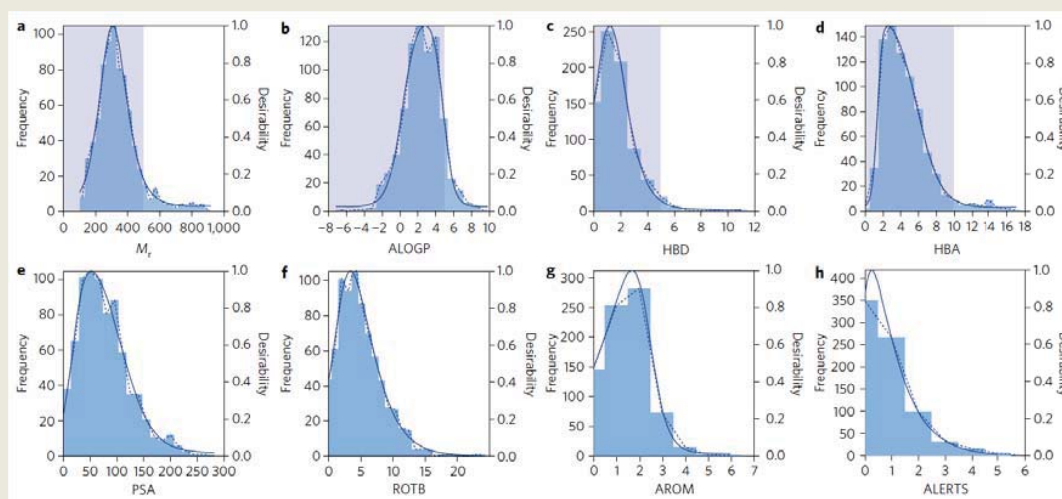
Quantifying the chemical beauty of drugs

G. Richard Bickerton¹, Gaia V. Paolini², Jérémy Besnard¹, Sorel Muresan³ and Andrew L. Hopkins^{1*}

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,

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

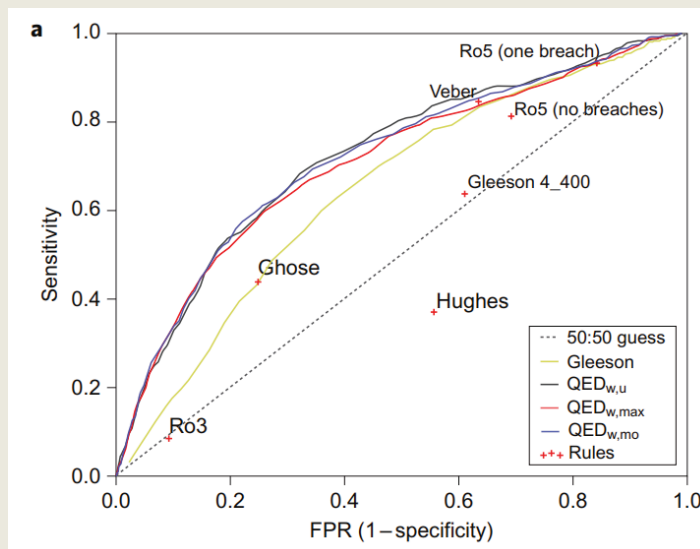
$$d(x) = a$$

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



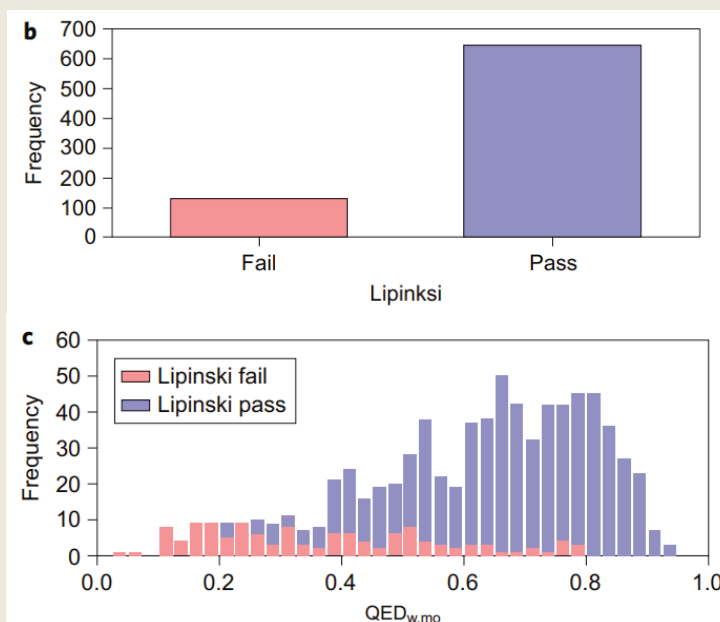
Performance

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



Rule-of-5 Comparison

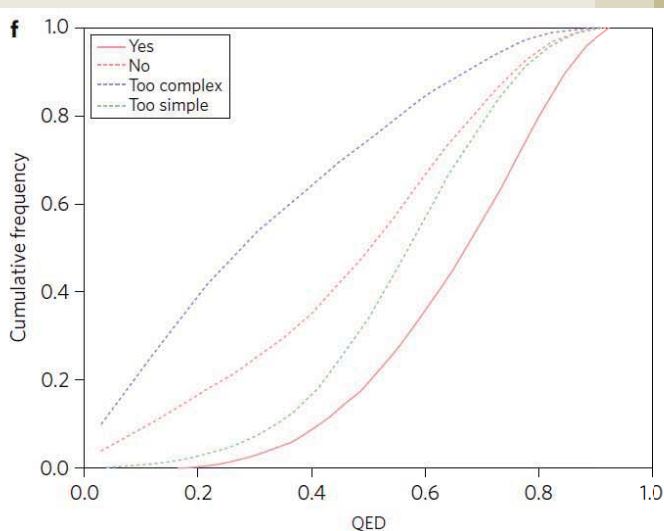
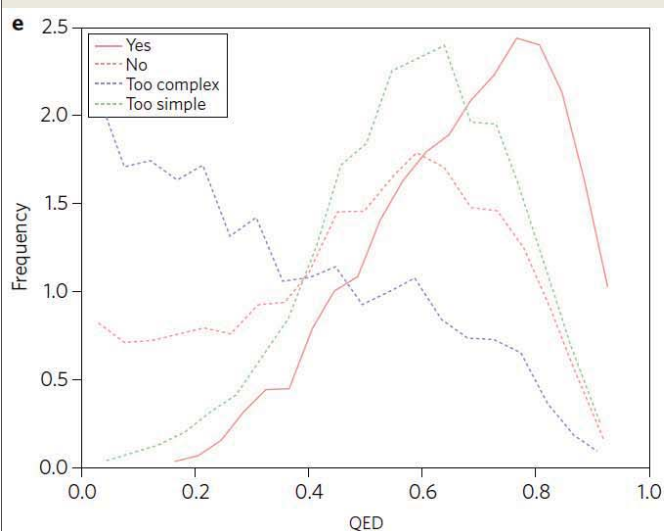
- ❑ Direct comparison of the Ro5 and QED shows the drugs failing (red) and passing (blue) Lipinski's Ro5





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
Accepted: 10 June 2009

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

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

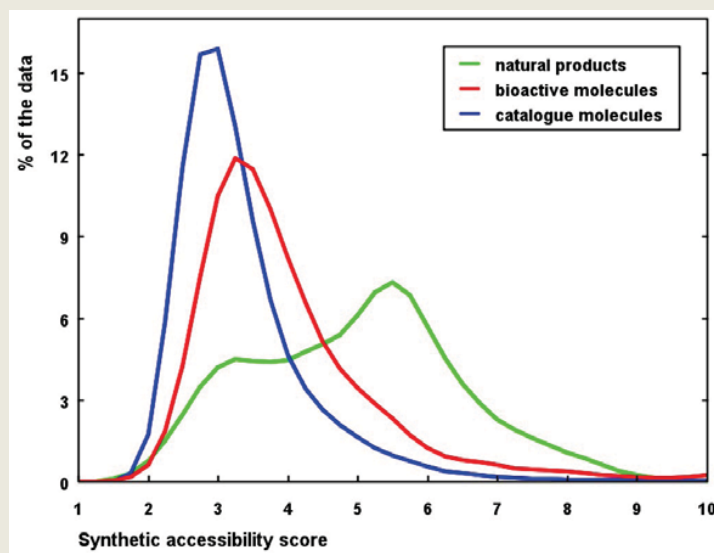
© 2009 Ertl and Schuffenhauer; licensee BioMed Central Ltd.

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.

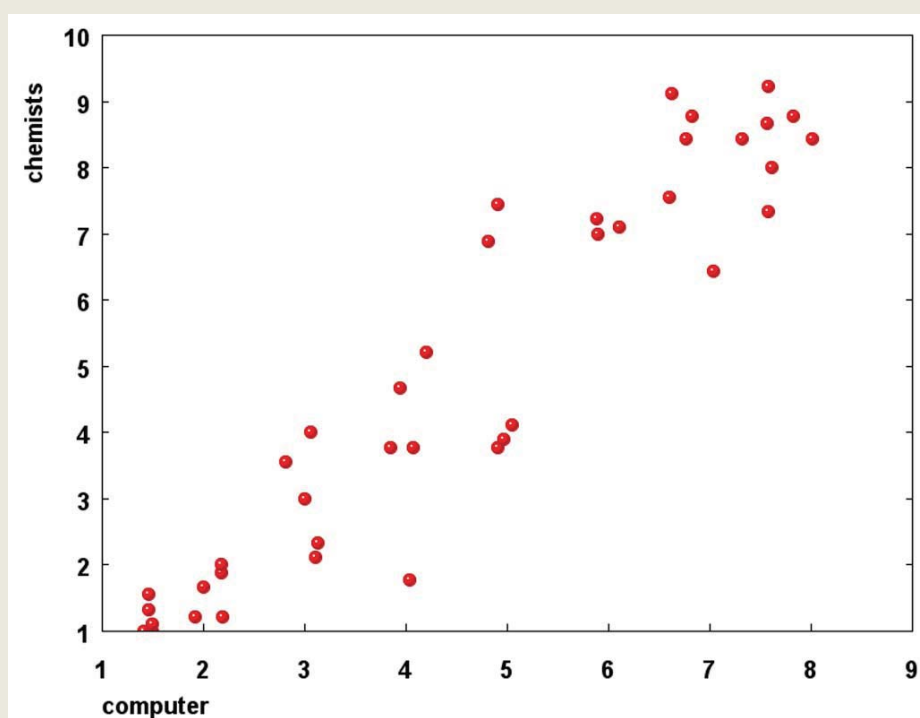


Score Distribution

- ❑ 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



Open-Source Cheminformatics
and Machine Learning



Tutorial



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

The RDKit 2021.03.1 documentation • Getting Started with the RDKit in Python previous | next | modules | index

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 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('C1=CC=CC=C1')
>>> m = Chem.MolFromMolFile('data/input.mol')
>>> stringWithMolData=open('data/input.mol','r').read()
>>> m = Chem.MolFromMolBlock(stringWithMolData)
```

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>

The screenshot shows the Google Colaboratory interface. At the top, it says 'Welcome To Colaboratory' with a menu bar containing 'File', 'Edit', 'View', 'Insert', 'Runtime', 'Tools', and 'Help'. On the right, there are 'Share', 'Settings', and 'Logout' icons. A left sidebar shows a 'Table of contents' with items like 'Getting started', 'Data science', 'Machine learning', 'More Resources', 'Machine Learning Examples', and 'Section'. The main content area is titled 'What is Colaboratory?' and includes a bulleted list of features: 'Zero configuration required', 'Free access to GPUs', and 'Easy sharing'. It also contains a paragraph about who can benefit from Colab and a 'Getting started' section.



Installing RDKit



- ❑ <http://5.9.10.113/65758226/why-do-we-have-to-append-path-for-rdkit-in-google-colab>
- ❑ !pip install rdkit-pypi



QED

```
❑ from rdkit import Chem
❑ m = Chem.MolFromSmiles('Cc1cccccl1')

❑ from rdkit.Chem import QED
❑ qed=QED.qed(m)
❑ print(qed)
```



SAS

```
❑ https://github.com/rdkit/rdkit/blob/master/Contrib/SA\_Score/sascorer.py
❑ 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
Bioactivity data			
ChEMBL	Bioactivity data from the medicinal chemistry literature	1 360 000	https://www.ebi.ac.uk/chembl/db
PubChem	Biological screening results on small molecules	49 000 000	https://pubchem.ncbi.nlm.nih.gov/
Patents			
IBM	Chemicals from full text patents	2 500 000	http://www-935.ibm.com/services/us/gbs/bao/siip/
SureChEMBL	Chemicals from full text patents	12 400 000	https://www.surechembl.org
Drugs			
DRUGBANK	Drug data and drug target information	7700	http://www.drugbank.ca
FDA/USP SRS	Substances present in FDA regulated products	34 000	http://fdasis.nlm.nih.gov/srs/srs.jsp
Availability			
ZINC	Commercially available compounds	22 700 000	http://zinc.docking.org
emolecules	Commercially available compounds	5 900 000	http://www.emolecules.com
Other			
ChEBI	Database and ontology of Chemical Entities of Biological Interest	27 000	https://www.ebi.ac.uk/chebi/
PDB	Data on biological macromolecular structures	16 000	https://www.ebi.ac.uk/pdbe/

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/>

NIH National Library of Medicine
National Center for Biotechnology Information

PubChem About Blog Submit Contact

Explore Chemistry

Quickly find chemical information from authoritative sources

Try covid-19 aspirin EGFR C9H8O4 57-17-2 C1=CC=C(C=C1)C=O InChI=1S/C3H6O/c1-3/24/11-2H3

Use Entrez Compounds Substances BioAssays

110M Compounds 272M Substances 298M Bioactivities 33M Literature 30M Patents 802 Data Sources

[See More Statistics >](#) [Explore Data Sources >](#)



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

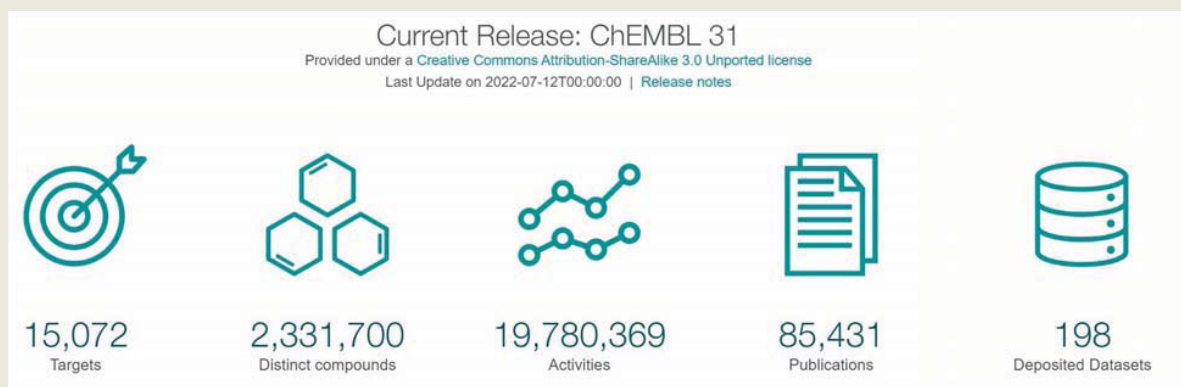
Data Collection	Live Count	Description
Compounds	110,040,027	Unique chemical structures extracted from contributed PubChem Substance records
Substances	271,907,539	Information about chemical entities provided by PubChem contributors
BioAssays	1,366,296	Biological experiments provided by PubChem contributors
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)
 - ❑ K_i (binding affinity)
 - ❑ MIC (minimum inhibitory concentration)
 - ❑ % Inhibition (of activity)



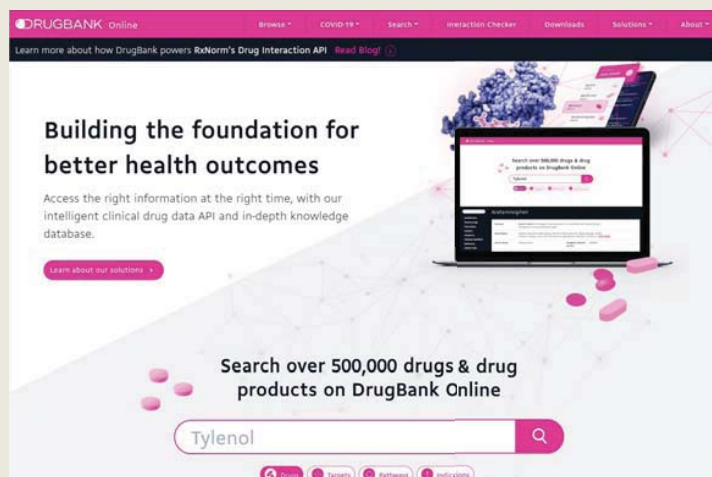
Protein Targets

- ❑ 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 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.

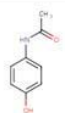
Brand Names *Acephen, Acetadryl, Allzital, Apadaz, Arthriten Inflammatory Pain, Bupap, Butapap, Cetafen, Children's Silapap, Contac Cold and Flu Non Drowsy Maximum Strength, Coricidin Hi* [READ MORE](#)

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).¹⁰ It is also used for its antipyretic effects, helping to reduce fever.²³ 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.^{15,16,23,Label}

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.¹⁹ Confusion about dosing of this drug may be caused by the availability of different formulas, strengths, and dosage instructions for children of different ages.¹⁹ 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.^{20,21,Label}

Type Small Molecule **Groups** Approved

Structure  **Weight** Average: 151.1626
Monoisotopic: 151.063328537

Chemical Formula C₉H₉NO₂

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



Targets



1. Prostaglandin E synthase 3 Details

Kind	Protein	General Function	Unfolded protein binding
Organism	Humans	Specific Function	Cytosolic prostaglandin synthase that catalyzes the oxidoreduction of prostaglandin endoperoxide H ₂ (PGH ₂) to prostaglandin E ₂ (PGE ₂) (PubMed:10922363). Molecular chaperone that localizes to genom...
Pharmacological action	Unknown	Gene Name	PTGES3
Actions	Inhibitor	Uniprot ID	Q15185
		Uniprot Name	Prostaglandin E synthase 3
		Molecular Weight	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

Kind	Protein	General Function	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
Standard Size Updated	Lead-Like 6,053,287 2014-09-29	Fragment-Like 847,909 2015-02-04	Drug-Like 17,900,742 2014-11-24	All Purchasable 22,724,825 2014-11-28	Shards 635,159 2014-05-16
Clean Size Updated	Clean Leads 4,591,276 2014-09-25	Clean Fragments 1,611,889 2014-09-24	Clean Drug-Like 13,195,699 2013-11-05	All Clean 16,403,865 2013-12-18	Clean Shards 325,950 2014-11-24
In Stock Size Updated	Leads Now 3,687,621 2014-06-25	Frag Now 704,041 2015-02-04	Drugs Now 10,639,555 2014-11-24	All Now 12,782,590 2014-05-01	Shards Now 424,775 2014-09-24
Boutique Size Updated	Boutique Leads 5,114,169 2012-12-24	Boutique Frags 2,755,555 2013-11-08	Boutique Drugs 10,292,210 2012-11-27	All Boutique 12,217,845 2012-11-27	Boutique Shards 80,698 2013-11-08
Comments/Citation	Teague, Davis, Leeson, Oprea, Angew Chem Int Ed Engl. 1999 Dec 16;38(24):3743-3748.	Carr RA, Congreve M, Murray CW, Rees DC, Drug Discov Today. 2005 Jul 15;10(14):987	Lipinski, J Pharmacol Toxicol Methods. 2000 Jul-Aug;44(1):235-49.	Purchasable chemical space	Type I binding sites
Filtering Criteria	p.mwt <= 350 and p.mwt >= 250 and p.xlogp <= 3.5 and p.rb <= 7	p.xlogp <= 3.5 and p.mwt <= 250 and p.rb <= 5	p.mwt <= 500 and p.mwt >= 150 and p.xlogp <= 5 and p.rb <= 7 and p.psa < 150 and p.n_h_donors <= 5 and p.n_h_acceptors <= 10		p.mwt < 190



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	128,648,525
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	2659674
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	186	47632	19622	2889	12273249
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 PROTEIN DATA BANK An Information Portal to 123622 Biological Macromolecular Structures

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

Advanced Search | Browse by Annotations

EMDBank Structural Biology Knowledgebase Worldwide 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.

October Molecule of the Month

Dipeptidyl Peptidase 4

Events and Activities

PUBLIC SYMPOSIUM AESTHETICS OF LIFE SCIENCES October 31, 2016 RUTGERS

PDB-101 USER SURVEY

Latest Entries As of Tuesday, Oct 19 Features & Highlights News Publications Contact Us



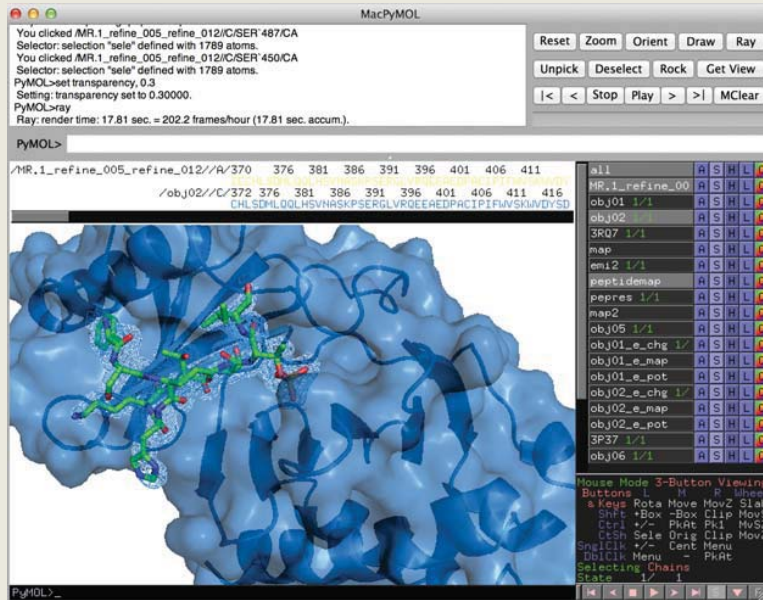
PDB ID

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

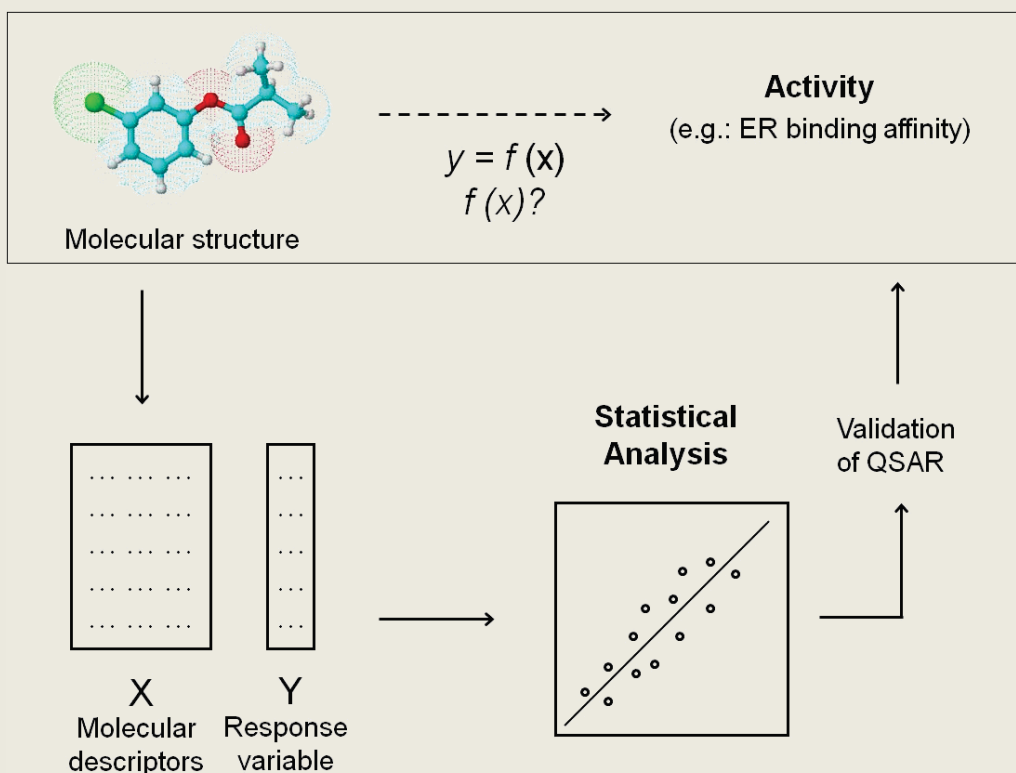


PyMOL: structure viewer

Free software (<http://pymol.org/ep>)



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

- ❑ 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	nAromAtom	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



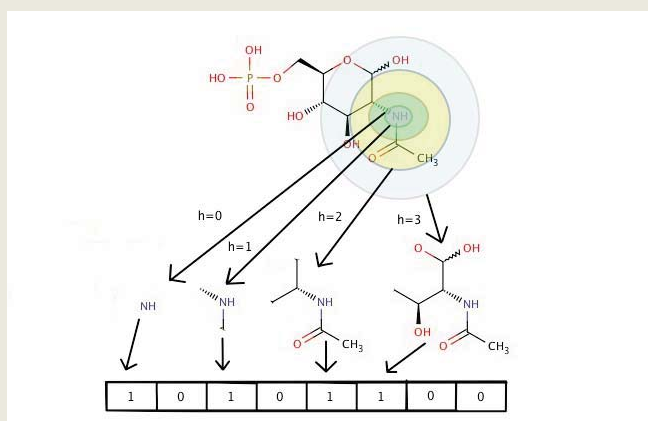
Fragment Codes

- ❑ 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 bitstring, 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



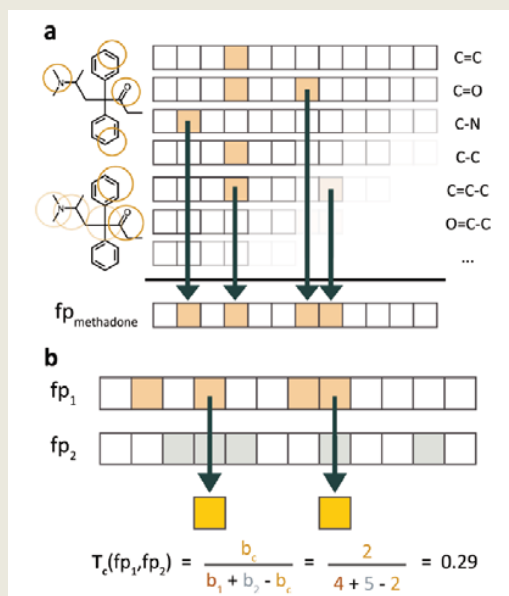
Molecular Fingerprint

- ❑ 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^{*,†} and Mathew Hahn[‡]

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

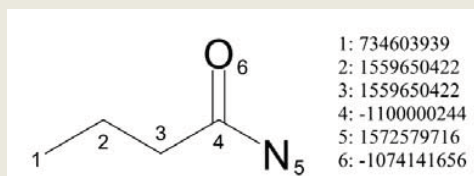
Received February 4, 2010

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.

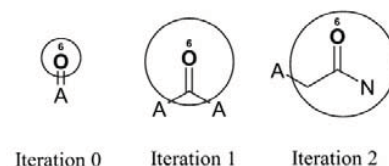
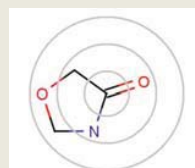


생성 과정

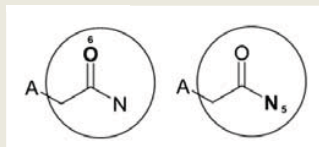
- ❑ Initial assignment of atom identifier



- ❑ Iterative updating of identifiers



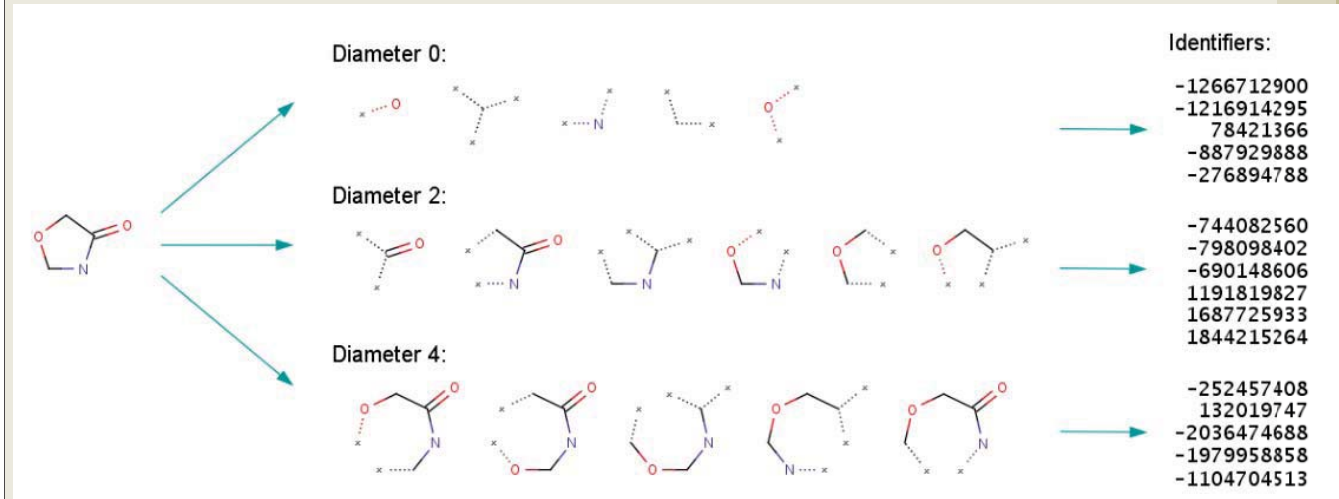
- ❑ Duplication removal





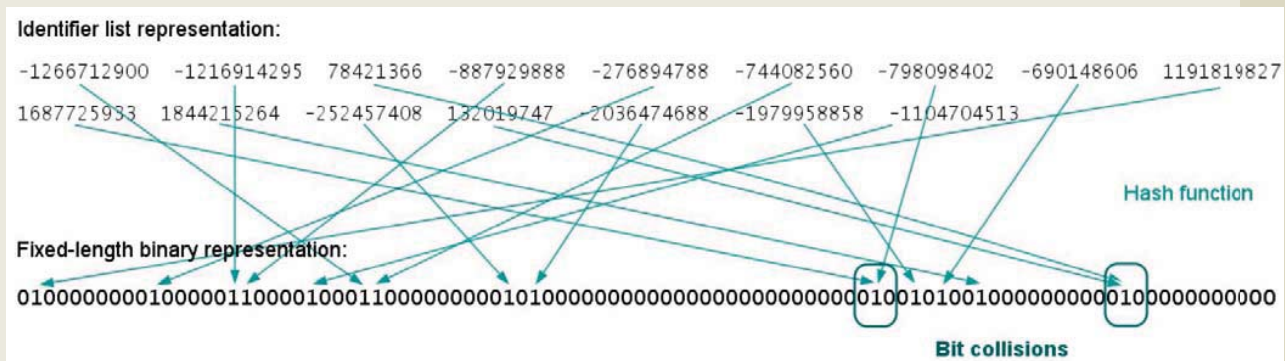
ECFP generation process

❑ Diameter (0, 2, 4, ...) or Radius (0, 1, 2, ...)



Generation of the fixed-length bit string

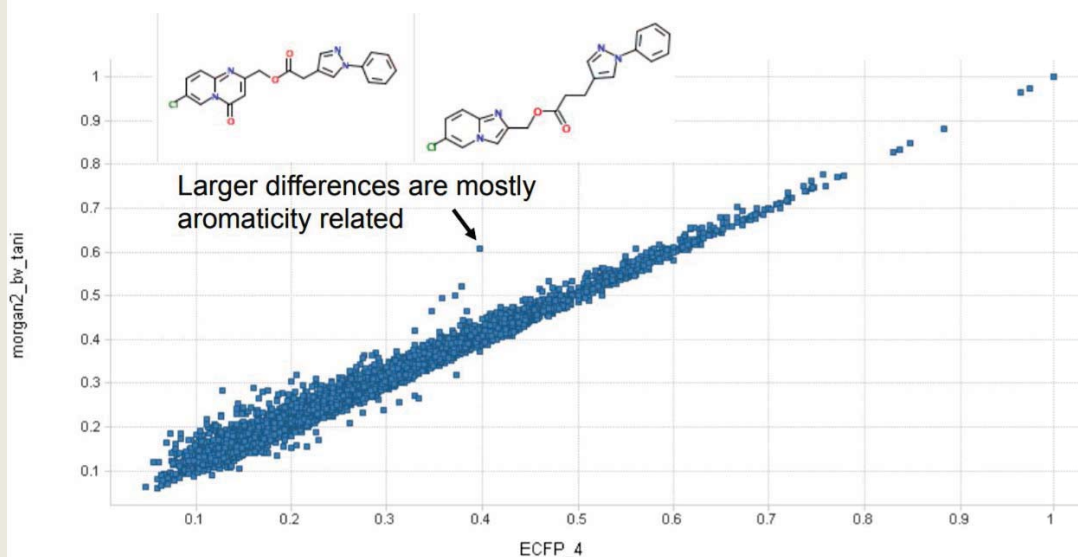
- ❑ "Folding" process
- ❑ length: 1024, 2048, ...
- ❑ Bit collisions can happen.





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 2023

인공지능 신약설계
AI Drug Design



Google Classroom

- ❑ BiML: AI 신약개발
- ❑ <https://classroom.google.com/u/o/c/NTEyMTlwODM5ODUo>
- ❑ 강의자료 및 실습용 코드 다운로드를 위해 모두 가입!



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



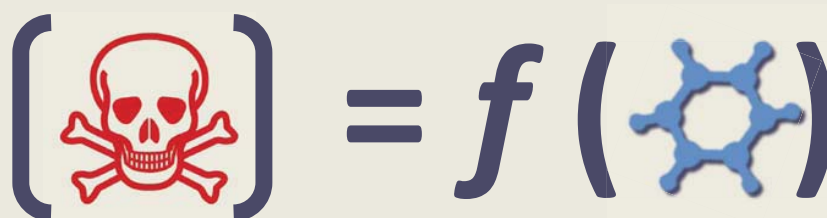
QSAR 모델링



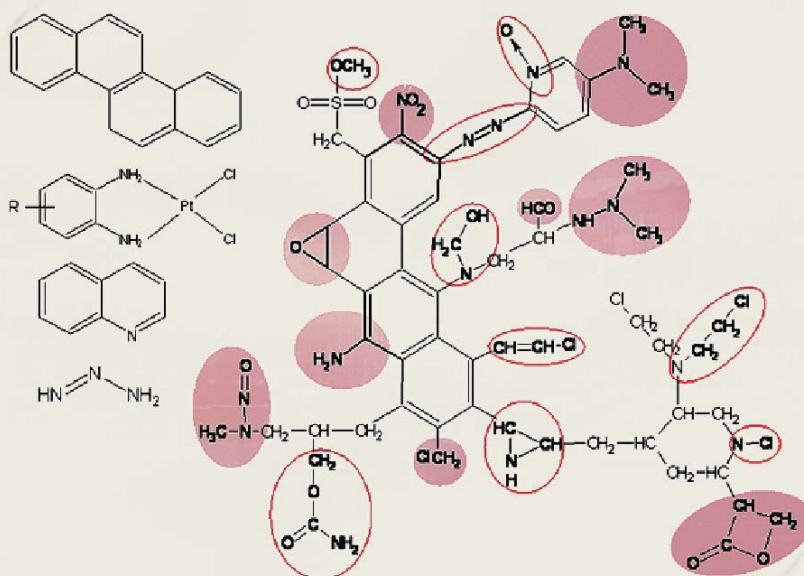


QSAR

- 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

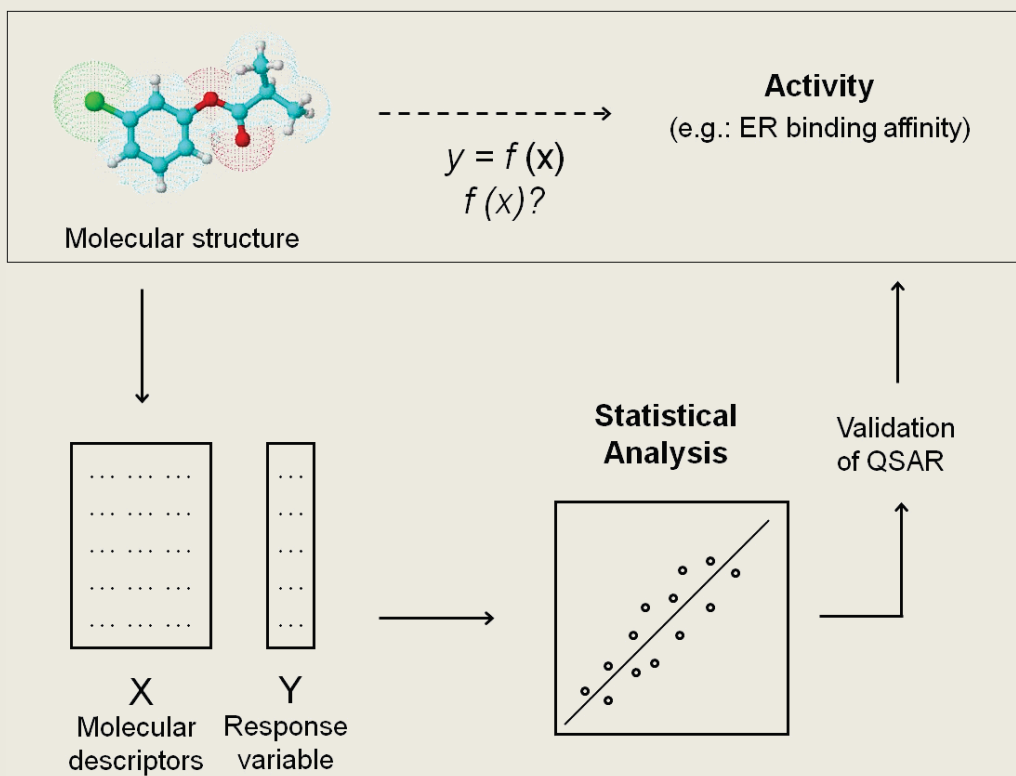


Link between toxicity and structures

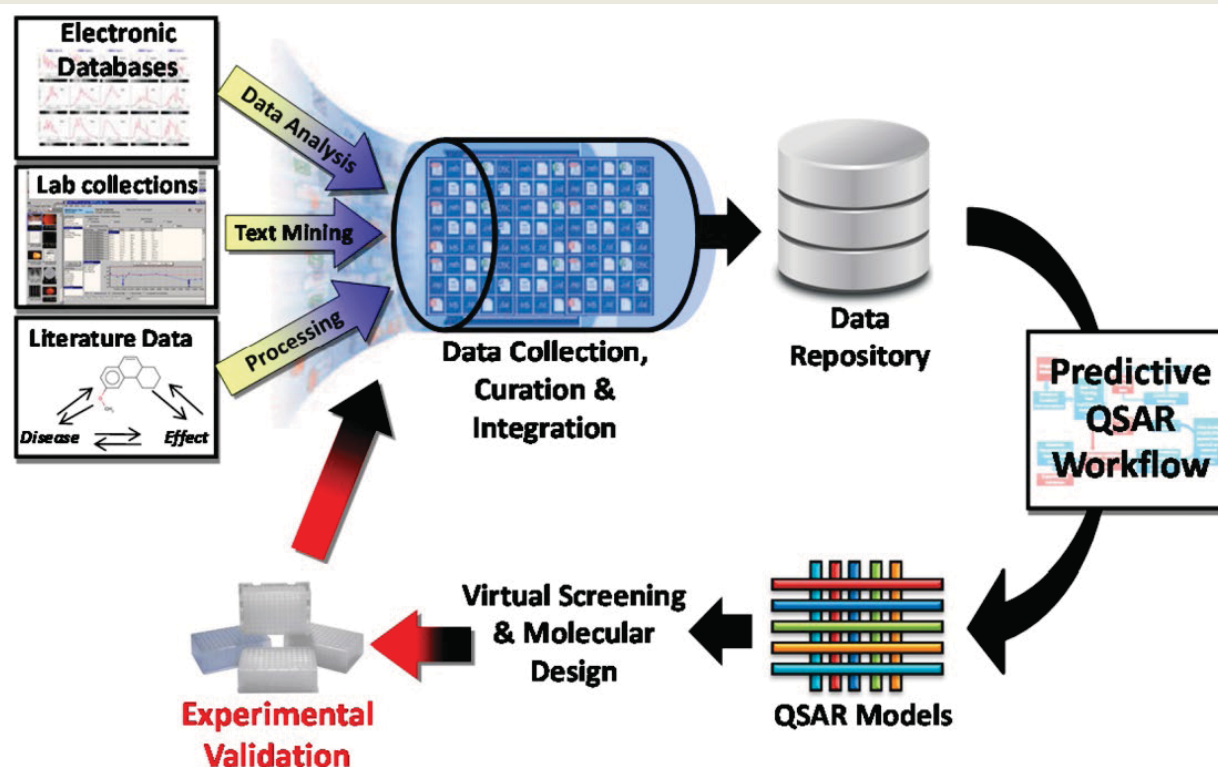




QSAR 모델링 과정

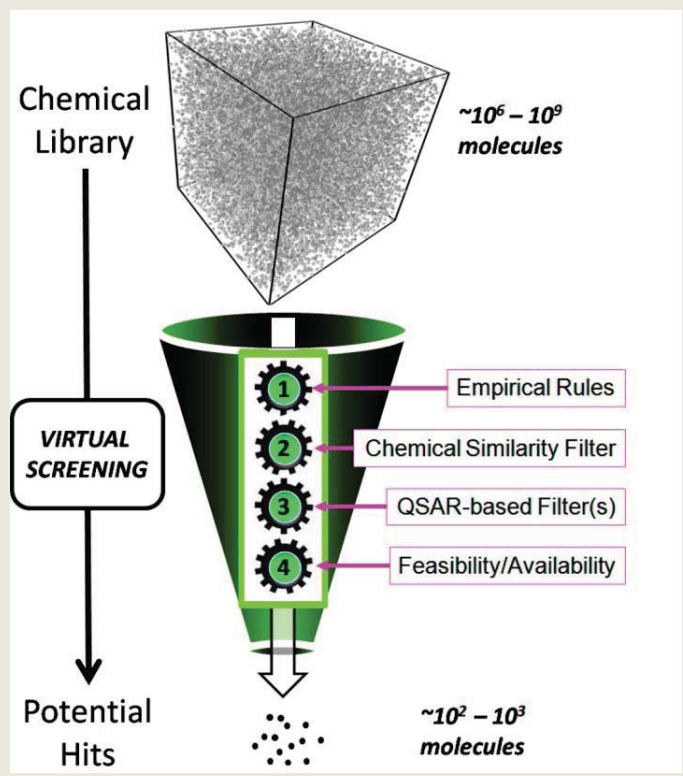


QSAR-guided drug discovery

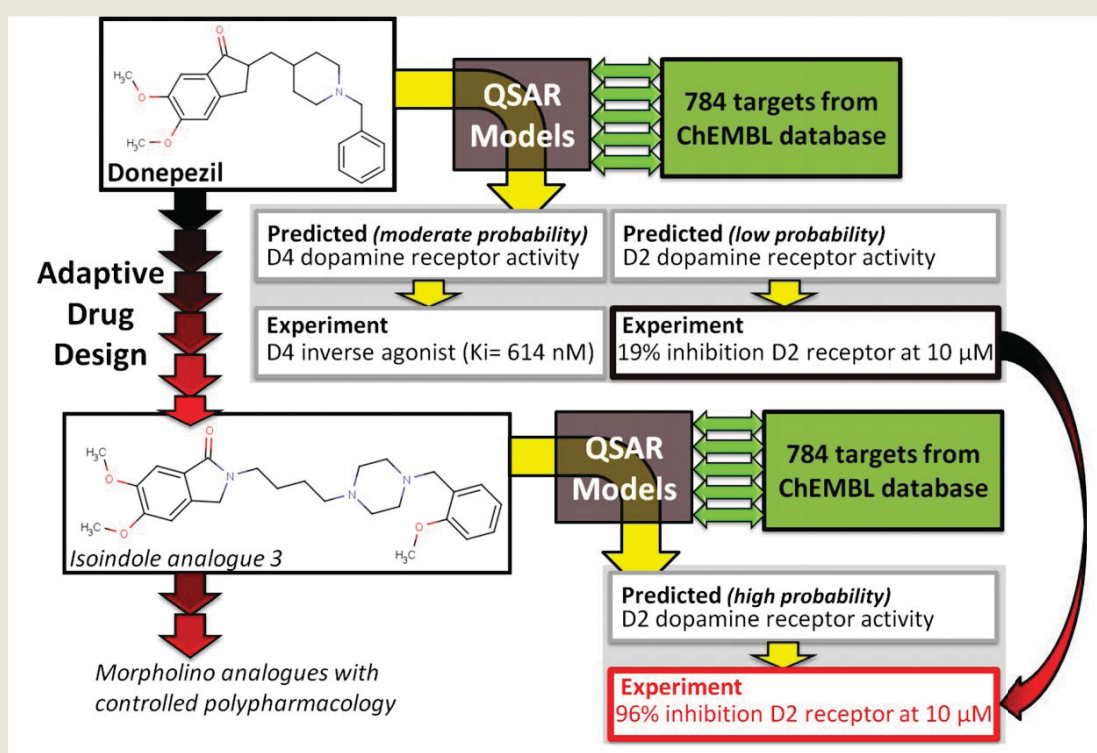




QSAR-based virtual screening



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
- ❑ **ΔG** – Gibbs free energy change associated with a chemical reaction, here a binding reaction



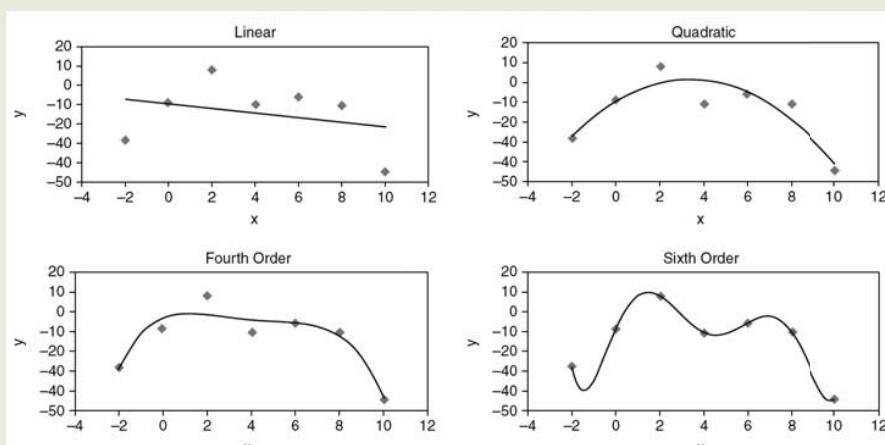
PREPARATION

- ❑ '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.



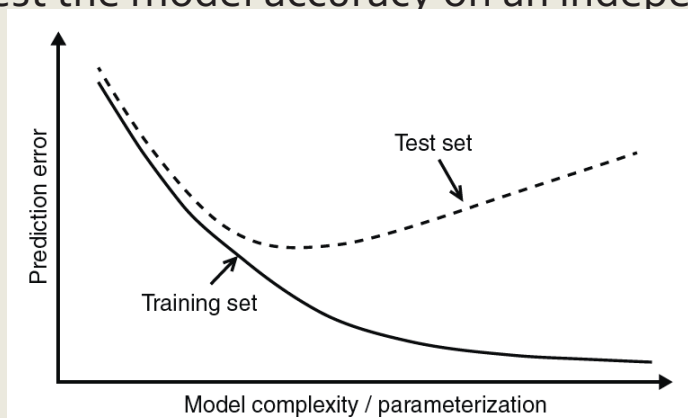
Model validation

- ❑ 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.



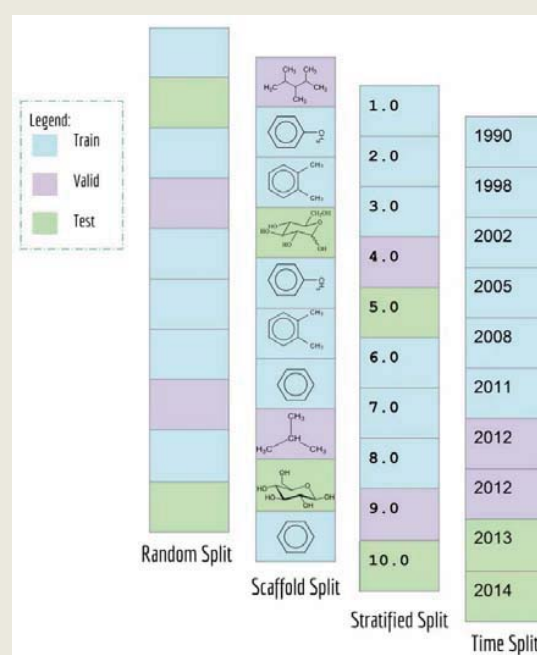
External Test Sets and Cross Validation

- ❑ 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)

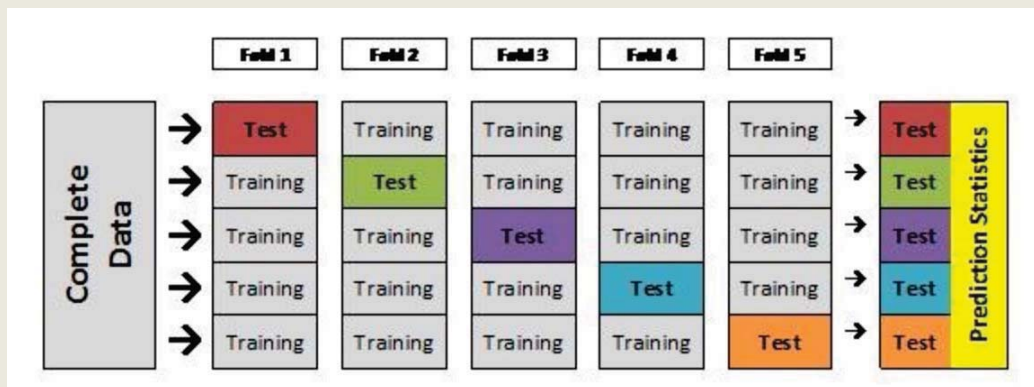




Cross Validation



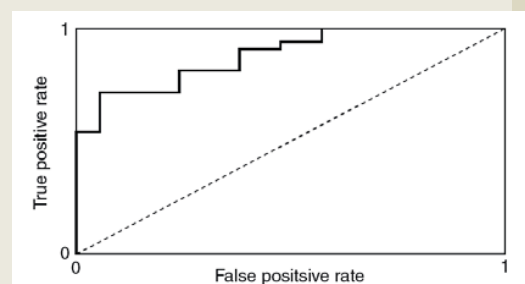
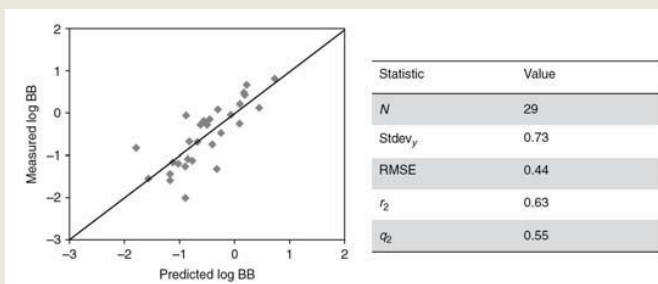
- ❑ Cross-validation
 - ❑ Leave-one-out, leave-cluster-out, n -fold cross validation
- ❑ Additional validation set

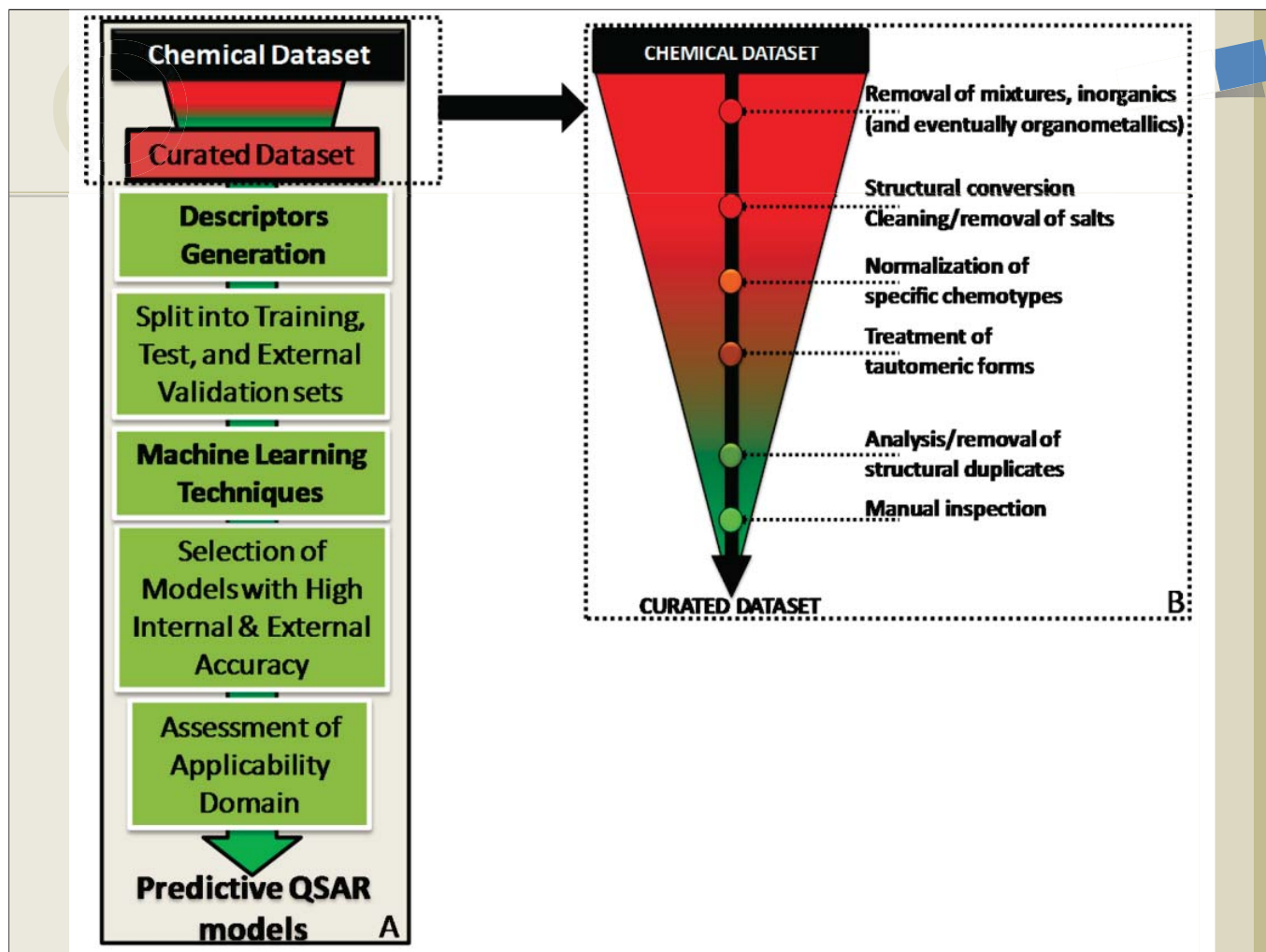


Assessing Model Performance



- ❑ <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



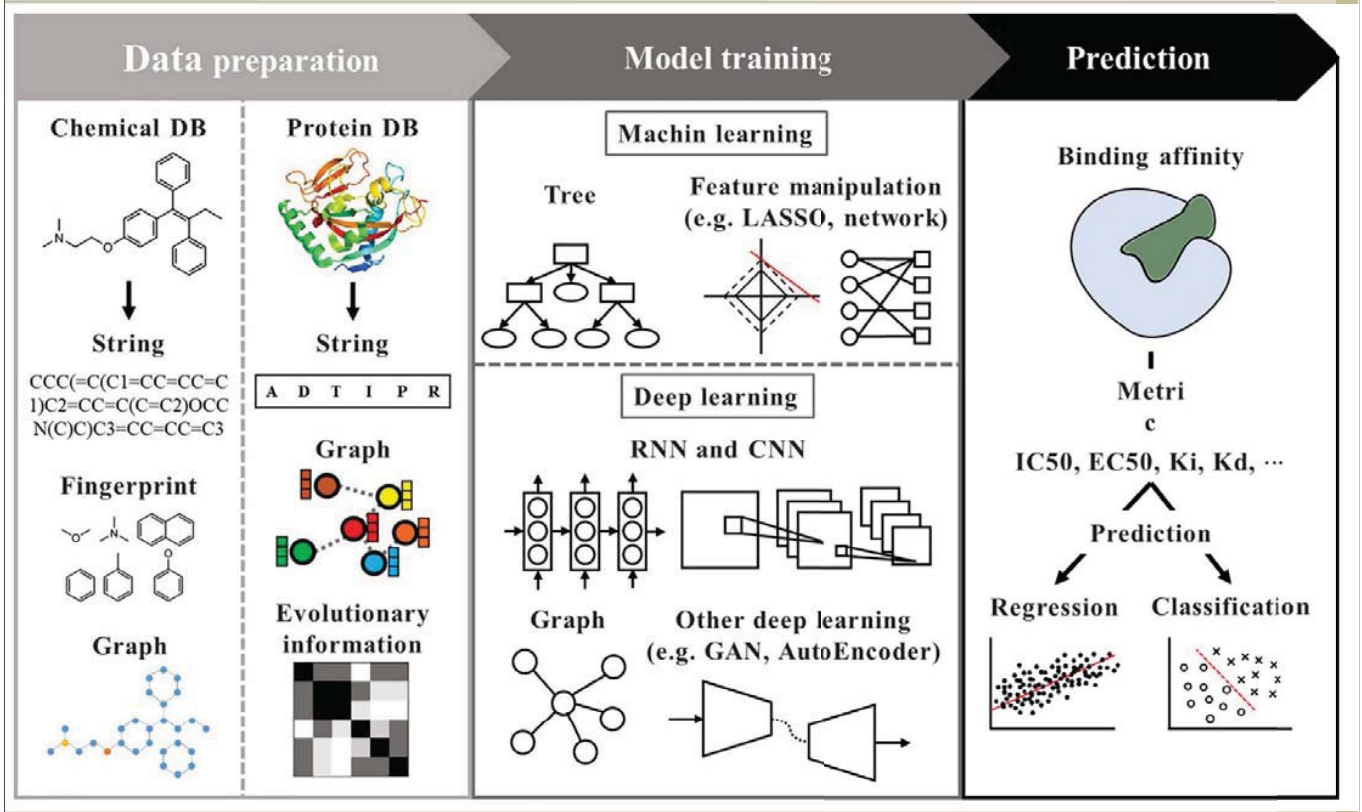


Lower limit

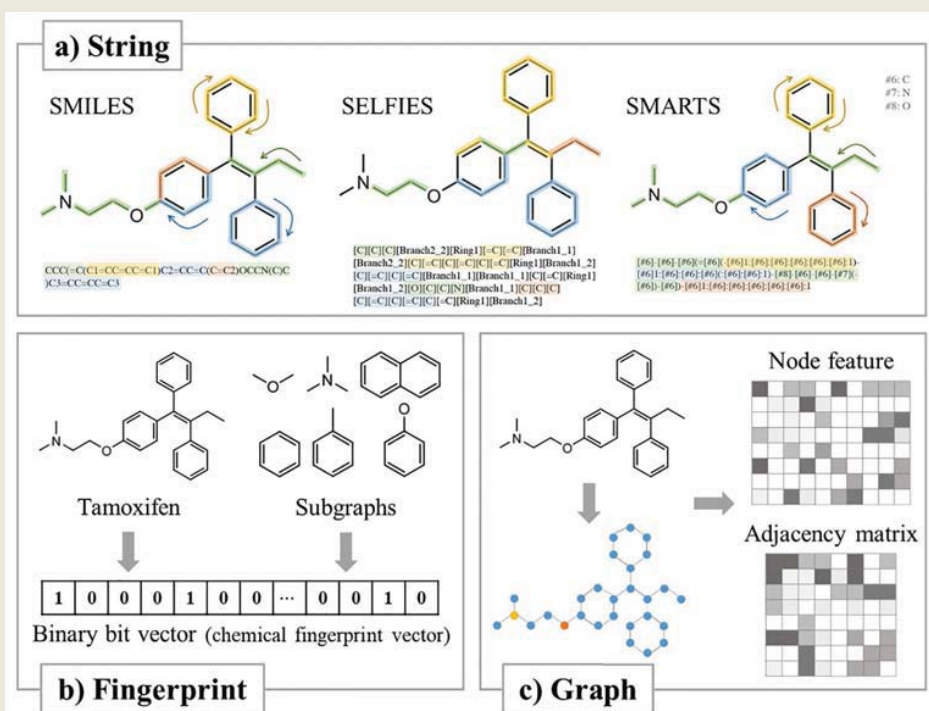
- ❑ If training sets are too small, correlation and over-fitting problem.
- ❑ Continuous response variable (activity),
 - ❑ the number of compounds in the training set should be at least 20
 - ❑ about 10 compounds should be in each of the test and external evaluation sets.
- ❑ Classification or category 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

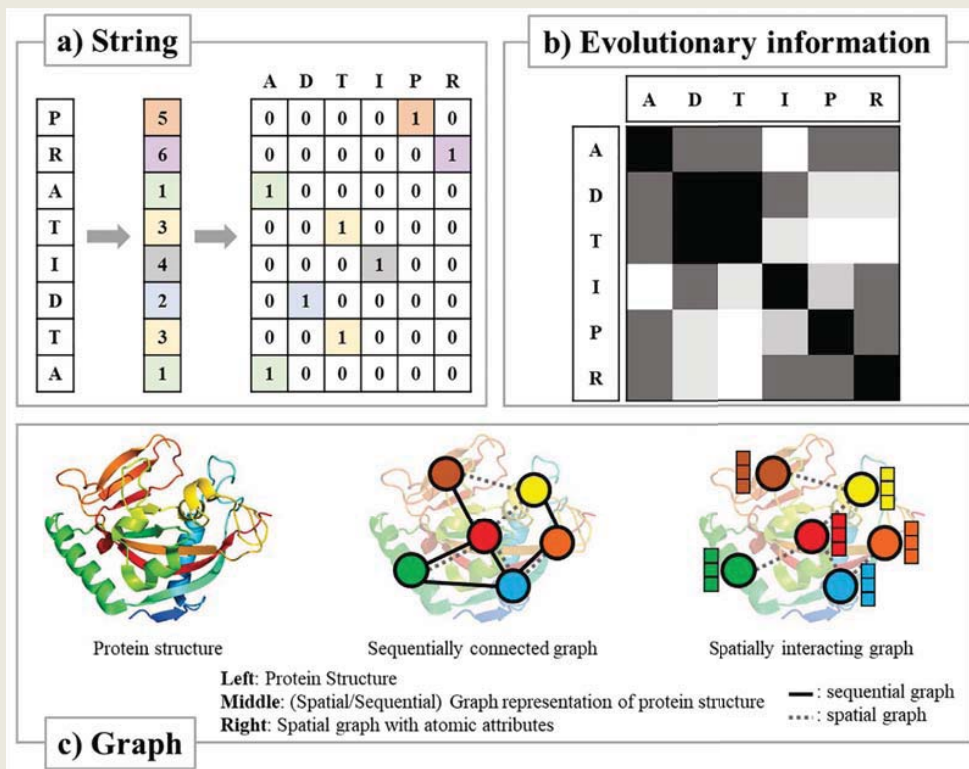


Ligand Featurization

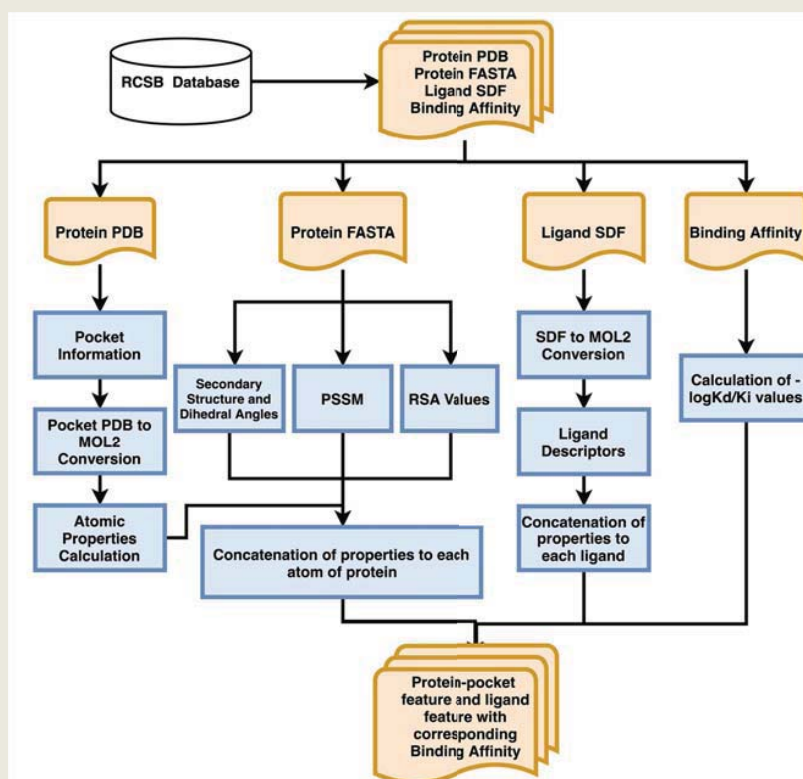




Protein Featurization



Feature extraction pipeline





Database

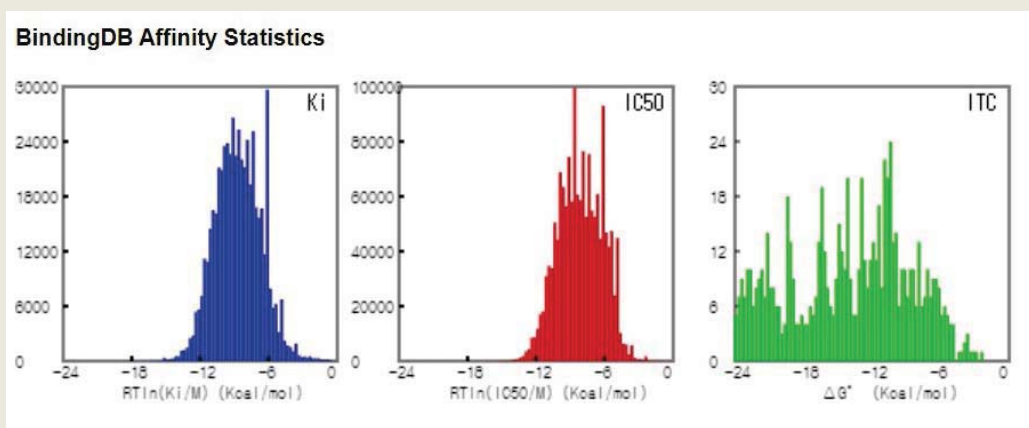
Protein-centric databases			
	Compounds	Proteins	Interactions
UniProt	-	20,385	-
Protein Data Bank	-	170,597	-
PDBbind	11,762	3,566	17,679*
Pfam	-	18,259	-
BRENDA	46	8083**	500 k

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

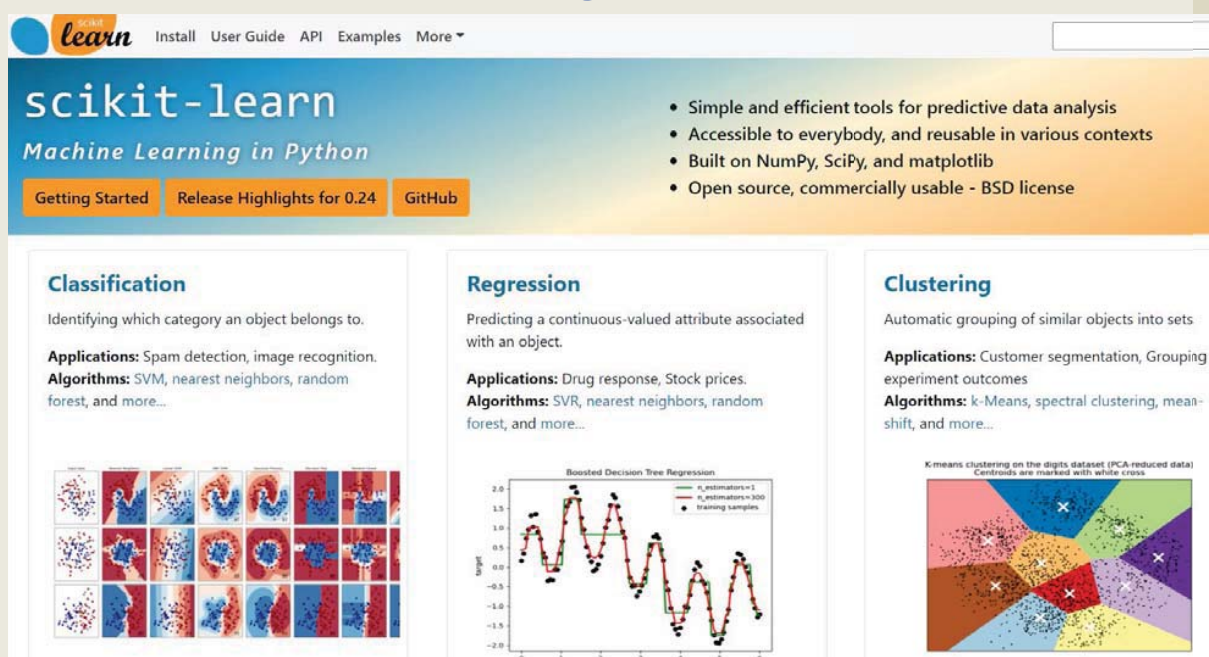
- ❑ <https://www.bindingdb.org/rwd/bind/>
- ❑ As of July 24, 2022, 2,546,129 binding data for 8,821 protein targets and 1,093,579 small molecules
- ❑ <https://www.bindingdb.org/bind/glossary.jsp>



QSAR를 위한 기계학습법

Scikit-learn

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



The screenshot shows the Scikit-learn website homepage. At the top, there is a navigation bar with links for 'Install', 'User Guide', 'API', 'Examples', and 'More'. Below the navigation bar, the 'scikit-learn' logo is prominently displayed, followed by the tagline 'Machine Learning in Python'. To the right of the logo, there are four bullet points highlighting key features: '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'. Below this, there are three main sections: 'Classification', 'Regression', and 'Clustering'. Each section provides a brief description, applications, and algorithms. The 'Classification' section includes a grid of 12 small images showing handwritten digits. The 'Regression' section features a line graph titled 'Boosted Decision Tree Regression' showing 'logit' values over time. The 'Clustering' section displays a scatter plot of 'K-means clustering on the digits dataset (PCA-reduced data)' with centroids marked by white crosses.



Textbook



- <http://www.kyobobook.co.kr/product/detailViewEng.laf?ejkGb=ENG&mallGb=ENG&barcode=9781492032649&orderClick=LEa&Kc=>

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

Aurelien Geron

가격: 99,500원
판매가: **79,600원** [20%] 19,900원 할인

이 책의 다른 상품 정보

배송비: 무료



Codes



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

ageron / handson-ml2

16.5k Stars 7.8k Forks

Code Issues 128 Pull requests 1 Actions Projects Wiki Security Insights

master 1 branch 0 tags

Go to file Code

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.

Releases: No releases published

Packages: No packages published



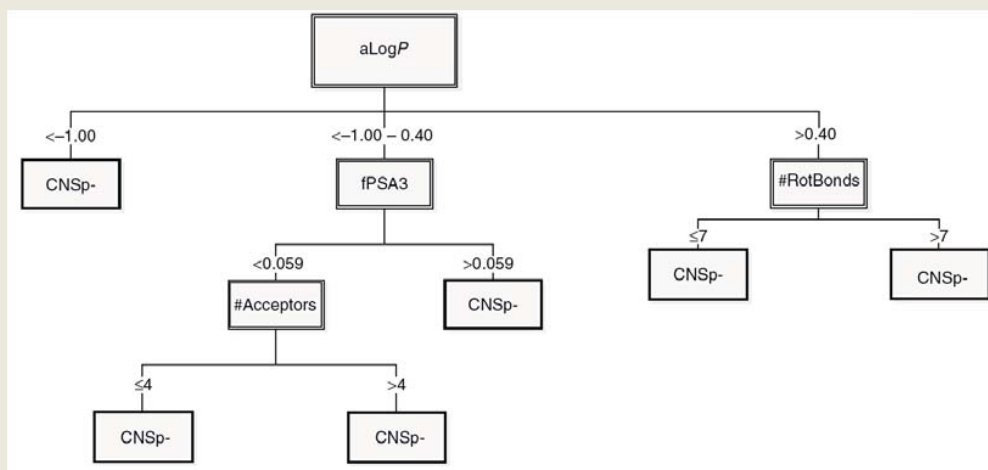
기계학습법 (Machine learning)

- 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)
```

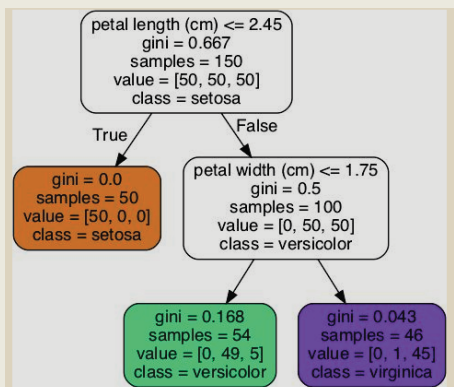
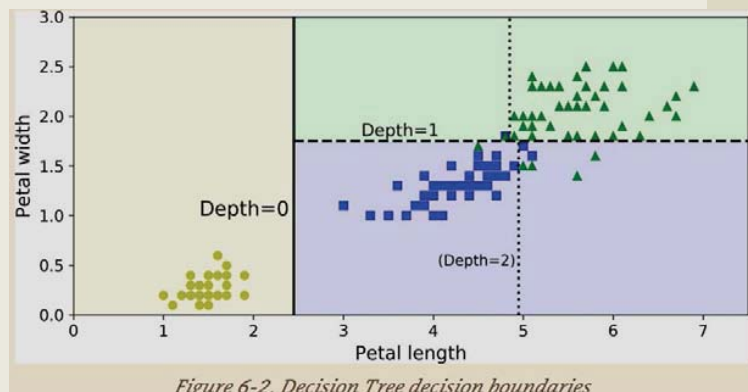


Figure 6-1. Iris Decision Tree



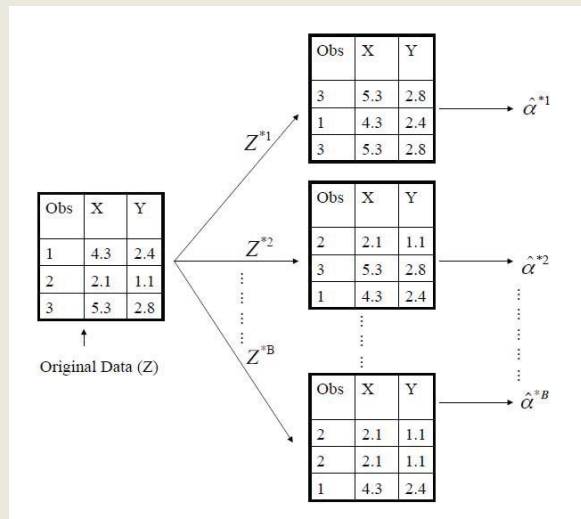
Pros and Cons

- ❑ 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**

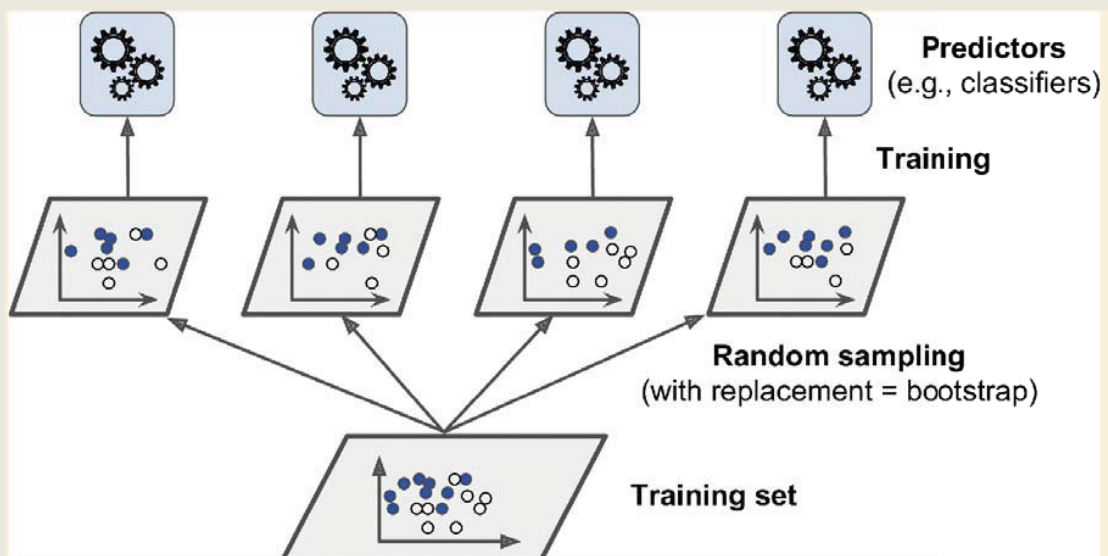
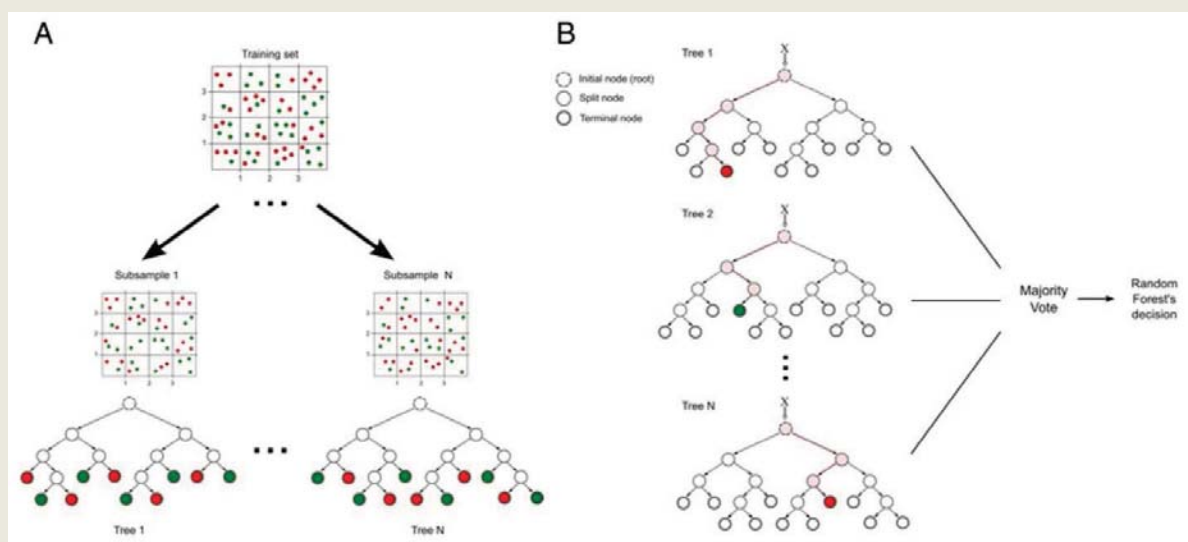


Figure 7-4. Bagging and pasting involves training several predictors on different random samples of the training set



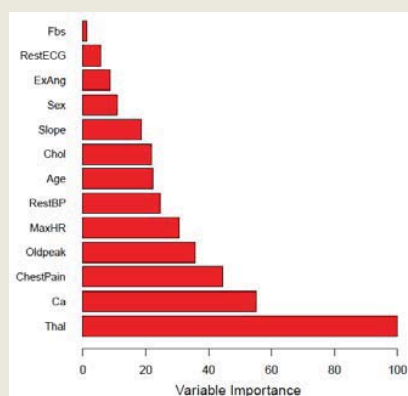
Random Forest

- ❑ 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

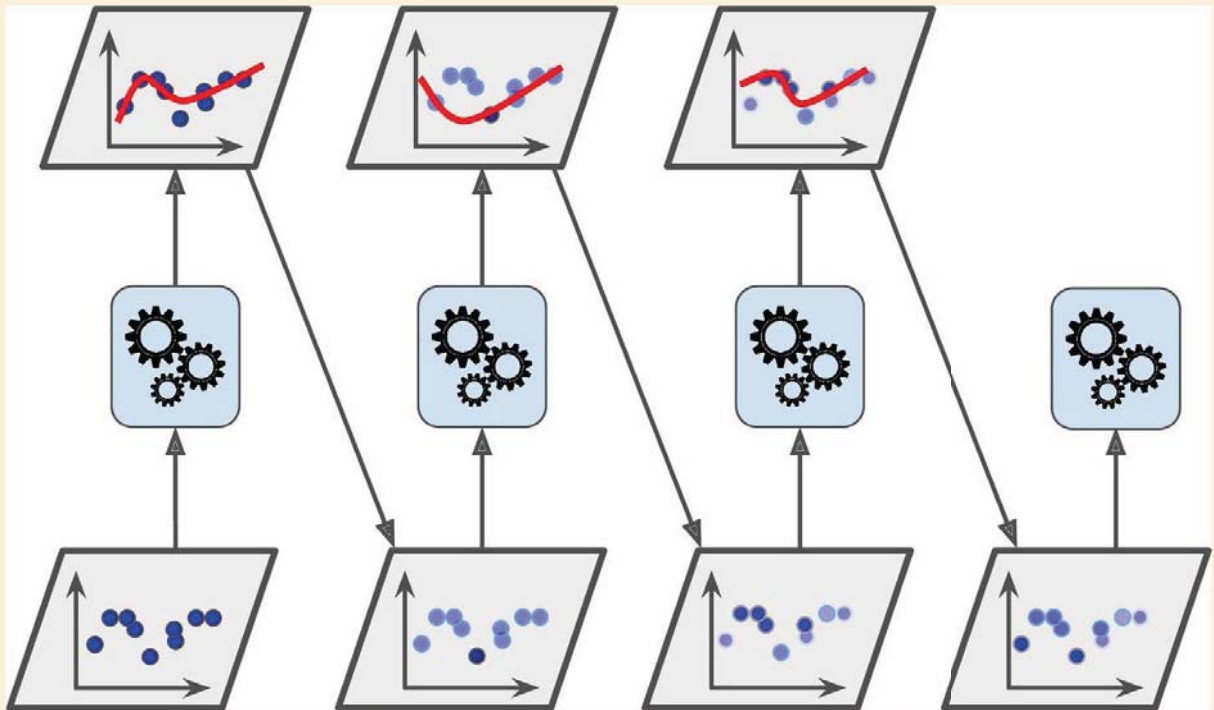


Figure 7-7. AdaBoost sequential training with instance weight updates

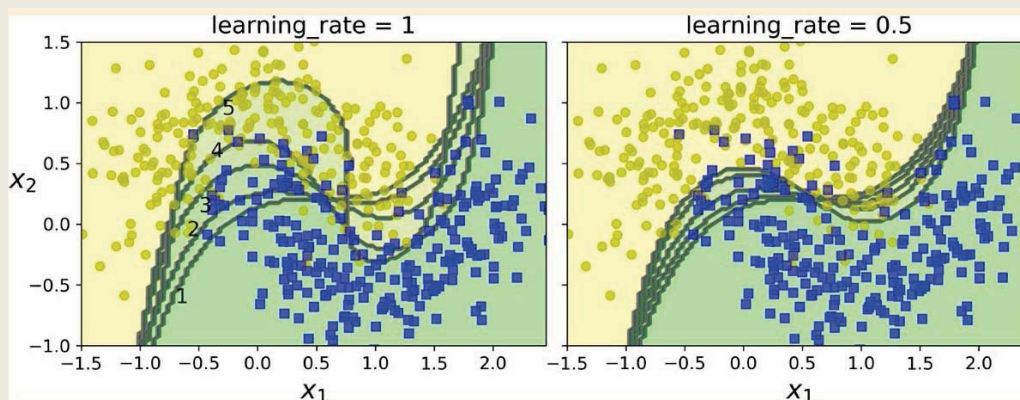


AdaBoost Code



```
from sklearn.ensemble import AdaBoostClassifier

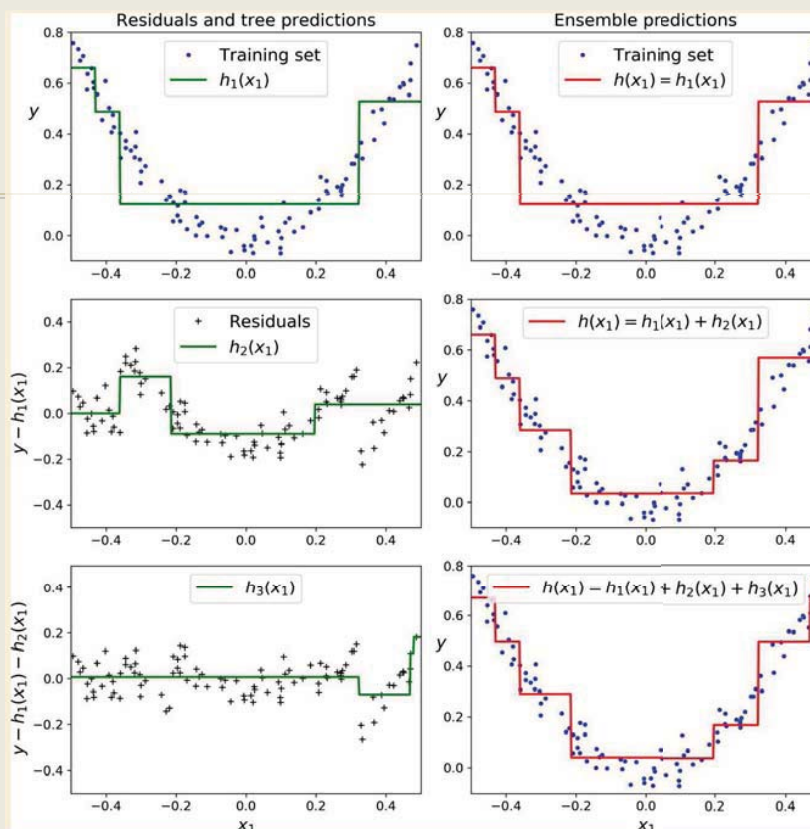
ada_clf = AdaBoostClassifier(
    DecisionTreeClassifier(max_depth=1), n_estimators=200,
    algorithm="SAMME.R", learning_rate=0.5)
ada_clf.fit(X_train, y_train)
```





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)
```

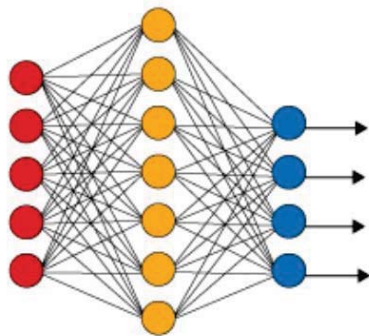
- ❑ <https://www.kaggle.com/stuarthallows/using-xgboost-with-scikit-learn>



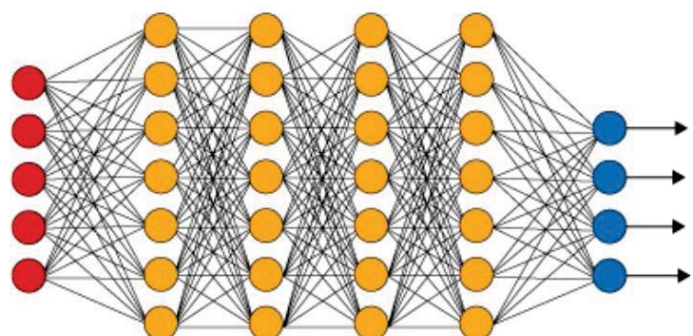
Deep learning methods



Simple Neural Network



Deep Learning Neural Network



● Input Layer ● Hidden Layer ● Output Layer



Drug Discovery



The rise of deep learning in drug discovery

Hongming Chen¹, Ola Engkvist¹, Yin Hai Wang², Marcus Olivecrona¹ and Thomas Blaschke¹



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

² 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



The screenshot shows the Kaggle page for the Merck Molecular Activity Challenge. The page layout includes a left sidebar with navigation options (Home, Competitions, Datasets, Code, Discussions, Courses, More) and a main content area. The main content area features a search bar, a 'Sign in' button, and a featured prediction competition card for the Merck Molecular Activity Challenge. The card displays the competition title, a \$40,000 prize, and a description: 'Help develop safe and effective medicines by predicting molecular activity.' Below the card, there are tabs for 'Overview', 'Data', 'Code', 'Discussion', 'Leaderboard', and 'Rules'. The 'Overview' tab is selected, showing a detailed description of the challenge, including the goal of identifying molecules that are highly active toward their intended targets but not toward other targets that might cause side effects. 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. At the bottom of the page, there is a progress bar showing the competition's status, with 'Launch' at 9 years ago and 'Close' at 9 years ago. Below the progress bar, there are statistics: 236 Teams, 269 Competitors, and 2,979 Entries. There are also links for 'Points' and 'Tiers'.

Category	Value
Teams	236
Competitors	269
Entries	2,979



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)



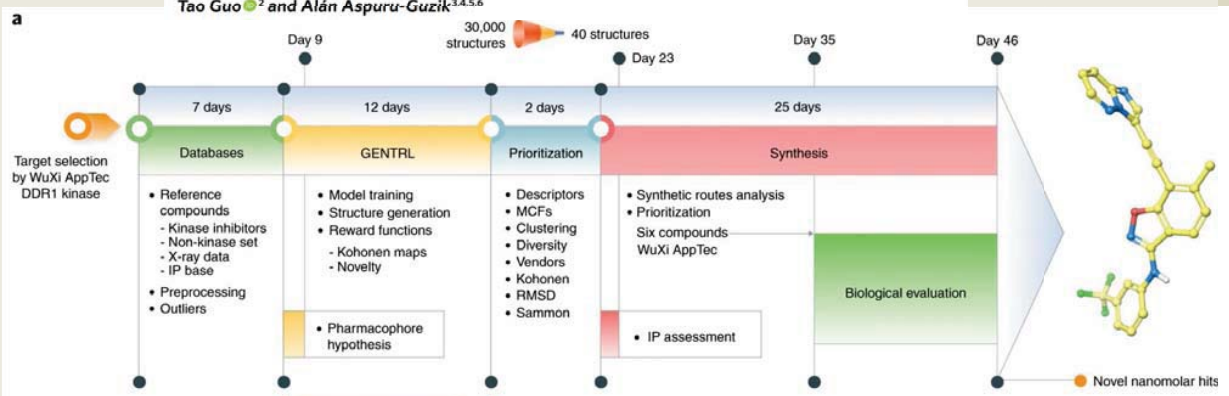
BRIEF COMMUNICATION

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

nature
biotechnology

Deep learning enables rapid identification of potent DDR1 kinase inhibitors

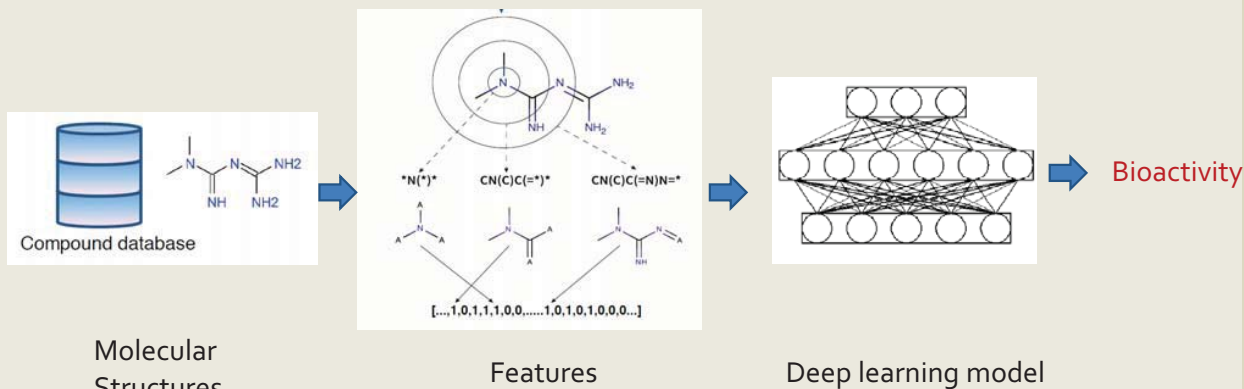
Alex Zhavoronkov^{1*}, Yan A. Ivanenkov¹, Alex Aliper¹, Mark S. Veselov¹, Vladimir A. Aladinskiy¹, Anastasiya V. Aladinskaya¹, Victor A. Terentiev¹, Daniil A. Polykovskiy¹, Maksim D. Kuznetsov¹, Arip Asadulaev¹, Yury Volkov¹, Artem Zholus¹, Rim R. Shayakhmetov¹, Alexander Zhebrak¹, Lidiya I. Minaeva¹, Bogdan A. Zagribelnyy¹, Lennart H. Lee², Richard Soll², David Madge², Li Xing², Tao Guo² and Alán Aspuru-Guzik^{3,4,5,6}





Simple Deep learning model

❑ QSAR Procedure



❑ Issues

- ❑ Featurization 방법
- ❑ DL 모델



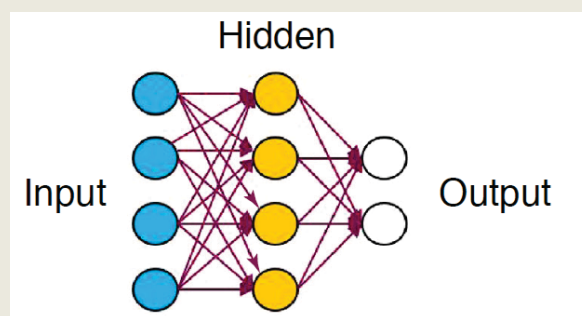
Deep Learning

- ❑ 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

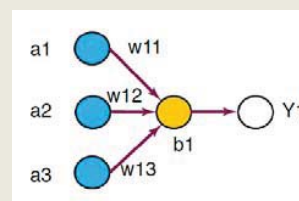
- ❑ DL uses artificial neural networks (ANNs) with many layers of nonlinear processing units for learning data representations.
- ❑ Three basic layers
 - ▣ input layer, hidden layer and output layer



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

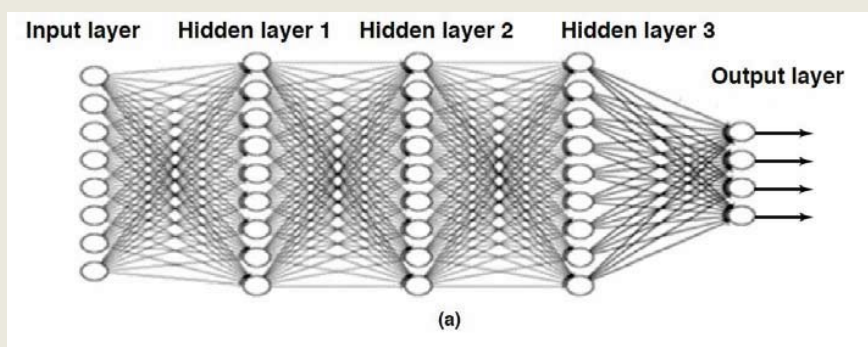


Popular Architectures

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

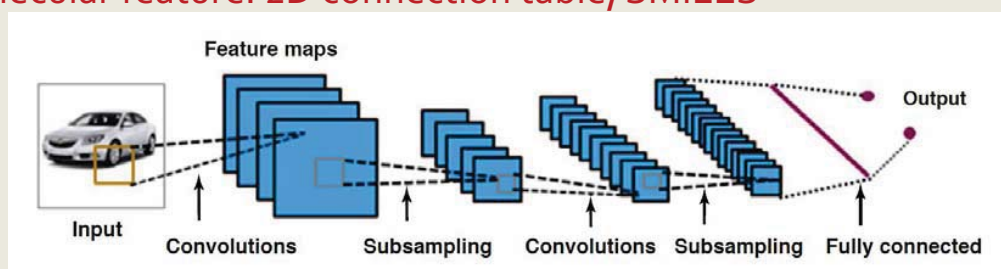
Fully connected deep neural network (FCN)

- ❑ 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)

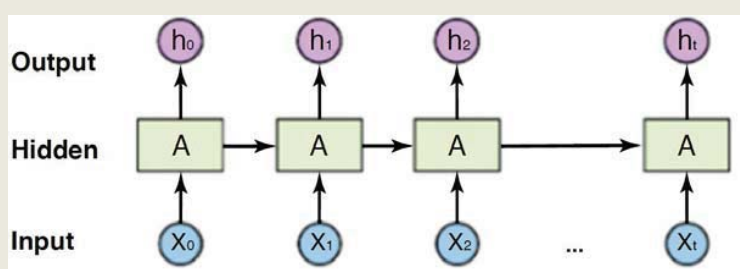
- ❑ Contains several convolution layers and subsampling layers
- ❑ The convolution layer consists of a set of filters (or kernels).
- ❑ Each filter is convoluted across the width and height of the input volume.
- ❑ The subsampling layer is used to reduce the size of feature maps.
- ❑ Owing to sharing 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**





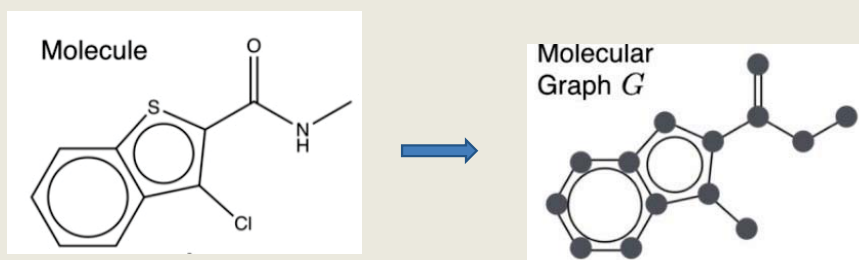
Recurrent neural network (RNN)

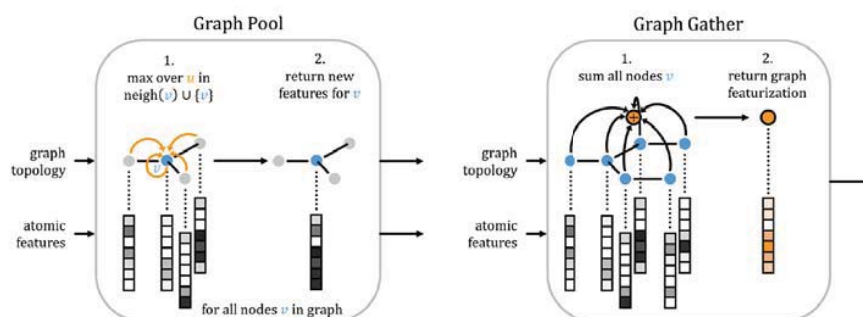
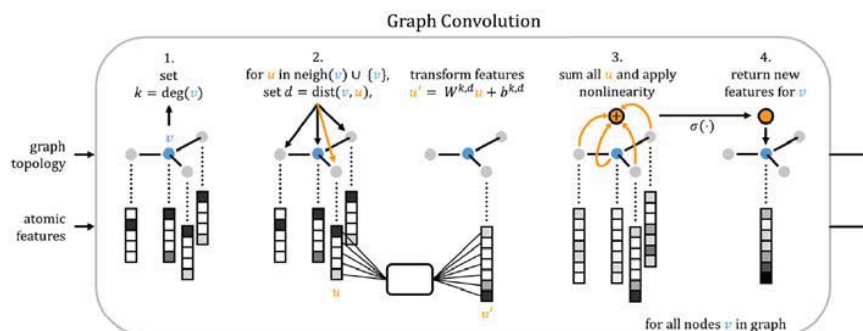
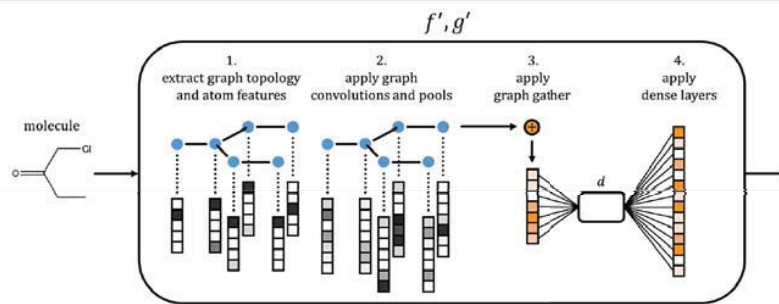
- ❑ RNNs can take sequential data as input features, which is very suitable for time-dependent tasks like language modeling.
- ❑ Using a technology called long short term memory (LSTM), RNNs can reduce the vanishing gradient problem.
- ❑ **Molecular feature: SMILES**



Graph convolution

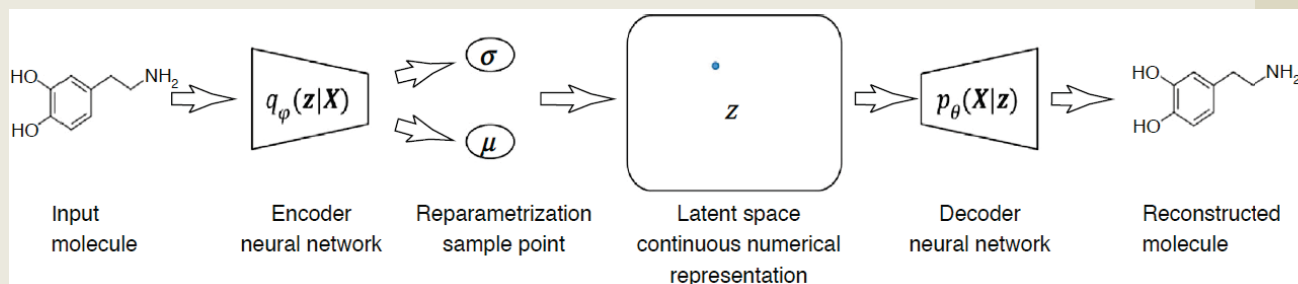
- ❑ Inspired by the Morgan circular fingerprint method
- ❑ First, the 2D molecular structure is read to form a state matrix, containing atom and bond information for each atom (Graph)
- ❑ The state matrix then goes through a convolution operation to generate a fixed length vector as the molecular representation.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



deepchem About Blog Tutorials Discuss Docs

Fork me on GitHub

DeepChem is a Python library democratizing deep learning for science.

Get Started OS Linux OSX

<https://deepchem.io/>

TensorFlow 설치 학습 API 리소스 더보기 검색 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_gumrsV3knCXnY6tjzVQtn#scrollTo=VMaLXjv77OJ3

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

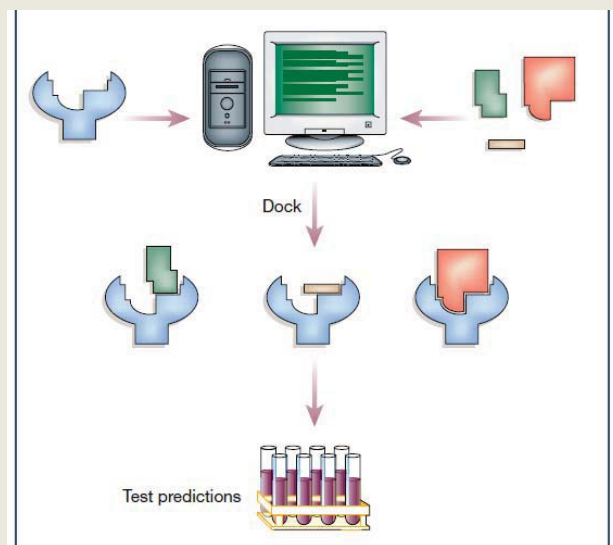


Table 1 Hit rates and drug-like properties for inhibitors discovered with high-throughput and virtual screening against the enzyme PTP-1B (ref.19)

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 μM or better inhibitors that passed all four of the drug-like criteria identified in Lipinski's 'rule of five'²⁵; †The number of compounds experimentally tested divided by the number of compounds with IC_{50} values of 100 μM 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

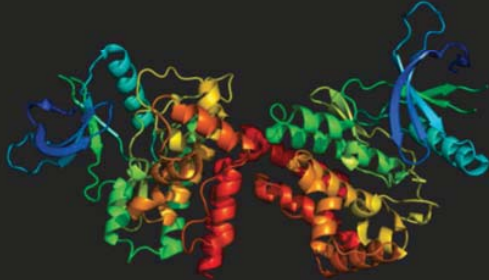
Predicting activity of JAK2 inhibitors

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/O6o674/entry>

UniProtKB · 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: 65

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⁺ + 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 entry in more detail.

Current Release: ChEMBL 31
Provided under a Creative Commons Attribution-NonCommercial 3.0 Unported license
Last Update on 2022-07-12T10:00:00 | 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

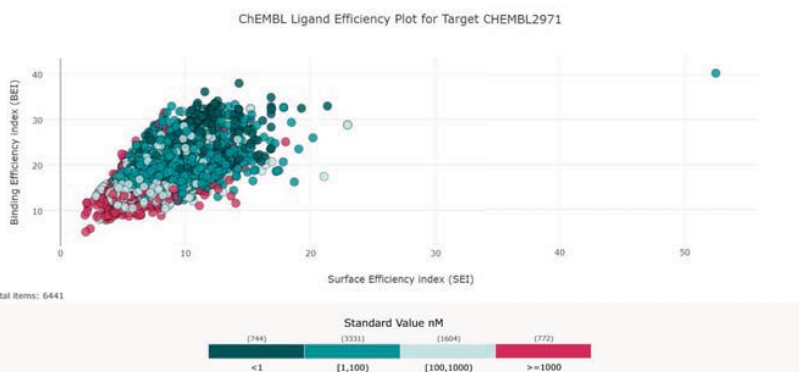
Records per page: 20
Showing 1-8 out of 8 records

CHEMBL ID	Search Hit	Name	UniProt Accessions	Type	Organism	Compounds	Activities
<input type="checkbox"/>	CHEMBL2971	Tyrosine-protein kinase JAK2	O60674	SINGLE PROTEIN	Homo sapiens	9682 By Mol. Wt.	12349 By Std. Type
<input type="checkbox"/>	CHEMBL4742263	Cereblon/Tyrosine-protein kinase JAK2	O60674, Q965W2	PROTEIN-PROTEIN INTERACTION	Homo sapiens	12 By Mol. Wt.	12 By Std. Type



Ligand Efficiencies

See all bioactivities for target CHEMBL2971 used in this visualisation



Activity Data

□ “CSV” 다운로드 및 편집 → JAK2_Chembl.csv






Browse Activities

Edit Querystring
Show Full Query

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	Sunlitrib	Kd	=	410.0	nM	6.39	No Data	CHEMBL1244467	Binding affinity to JH1 catalytic domain JAK2	single protein format	No D.
 CHEMBL1230609	EXEL-2880/GSK-1363089	Kd	=	1500.0	nM	5.82	No Data	CHEMBL1908670	Binding constant for JAK2(JH1domain-catalytic) kinase domain	single protein format	No D.
 CHEMBL1789941	INC18424	Kd	=	0.036	nM	10.44	No Data	CHEMBL1908670	Binding constant for JAK2(JH1domain-catalytic) kinase domain	single protein format	No D.
 CHEMBL607707	EKB-569	Kd	=	2000.0	nM	5.70	No Data	CHEMBL860903	Average Binding Constant for JAK2 (Kin.Dom. 2); NA=Not Active at 10 uM	single protein format	Humo
 CHEMBL1908670	PKC-412	Kd	=	94.0	nM	7.03	No Data	CHEMBL1908670	Binding constant for JAK2(JH1domain-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 heading is 'ZINC20'. Below the heading, there is a welcome message and a paragraph describing the database. To the right, there is a section for 'ZINC20 News' with a sub-heading 'ZINC20 has been released' and a 'Caveat Emptor' warning. Below the main heading, there are three columns of content: 'Getting Started' with a list of links, 'Ask Questions' with a list of questions, and 'Explore Resources' with a list of links.

ZINC20

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.

ZINC is provided by the Irwin and Shoichet Laboratories in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF). We thank NIGMS for financial support (GM71896).

To cite ZINC, please reference: Irwin, Tang, Young, Dandarchuluun, Wong, Khurebaatar, Moroz, Mayfield, Sayle, *J. Chem. Inf. Model* 2020, in press. <https://pubs.acs.org/doi/10.1021/acs.jcim.0c00675>. You may also wish to cite our previous papers: Sterling and Irwin, *J. Chem. Inf. Model*, 2015 <http://pubs.acs.org/doi/abs/10.1021/acs.jcim.5b00559>, Irwin, Sterling, Mysinger, Bolstad and Coleman, *J. Chem. Inf. Model*, 2012 DOI: 10.1021/ci3001277 or Irwin and Shoichet, *J. Chem. Inf. Model*, 2005,45(1):177-82 PDF, DOI.

Getting Started

- Getting Started
- What's New
- About ZINC 20 Resources
- Current Status / In Progress
- Why are ZINC results "estimates"?

Explore Resources

Chemistry

Tranches, Substances, 3D Representations, Rings, Patterns

And More

Catalogs, Genes, ATC Codes

Ask Questions

You can use ZINC for **general** questions such as

- How many substances in current clinical trials have PAINS patterns? (150)
- How many natural products have names in ZINC and are not for sale? (9296) get them as SMILES, names and calculated logP
- How many endogenous human metabolites are there? (47319) and how many of these can I buy? (6271) How many are FDA approved drugs? (94)
- How many compounds known to aggregate are in current clinical trials? (60)
- How many epigenetic targets have compounds known? (53) and Which of these substances can I buy? (278)
- How many ligands are there for the NMDA 1 (or channel GRIN1? (662) and How many of these are for sale? (60)
- More...

ZINC20 News

- ZINC20 has been released

Caveat Emptor: We do not guarantee the quality of any molecule for any purpose and take no responsibility for errors arising from the use of this database. ZINC is provided in the hope that it will be useful, but you must use it at your own risk.



Compound Library



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

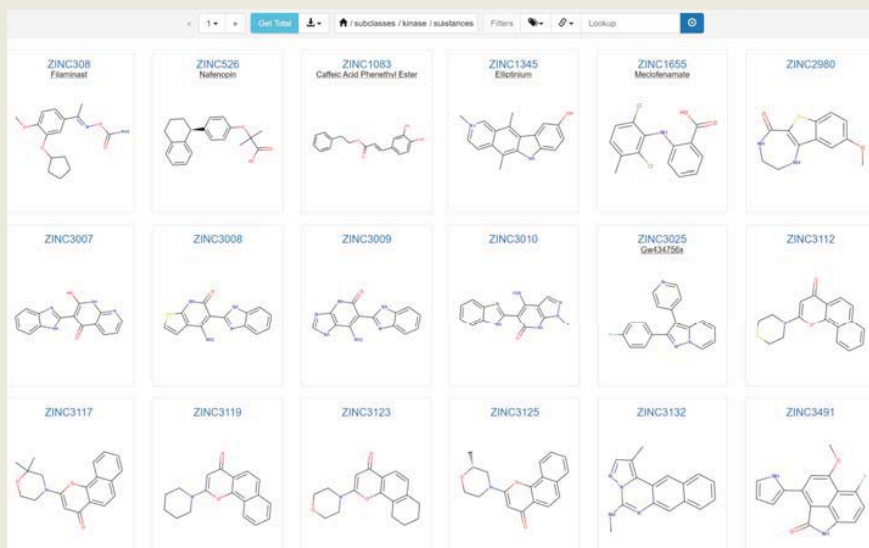
Molecular Weight (up to, Daltons)												Totals, by LogP	
	200	250	300	325	350	375	400	425	450	500	>500		
LogP (up to)	-1	27,791	172,563	710,795	1,072,978	2,241,498	786,738	276,834	116,066	92,417	77,790	7,310	5,582,780
	0	139,434	934,776	3,655,384	5,126,157	10,608,025	3,498,214	1,663,579	708,919	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/>

SwissDock

Swiss Institute of Bioinformatics

Home Target Database Submit Docking Command Line Access Help Forum Contact

What?

This website provides an access to:

- **SwissDock**, a web service to predict the molecular interactions that may occur between a target protein and a small molecule.
- **S3DB**, a database of manually curated target and ligand structures, inspired by the **Ligand-Protein Database**.

Would you like to organize a workshop? Please let us know so that we can adjust the computing resources accordingly.

Why?

- Propose a binding mode for a ligand
- Create figures for your articles
- Generate a complex to perform subsequent calculations
- Design inhibitors for the target of your choice

How?

SwissDock is based on the docking software EADock DSS, whose algorithm consists of the following 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.

Viewing SwissDock predictions in U...

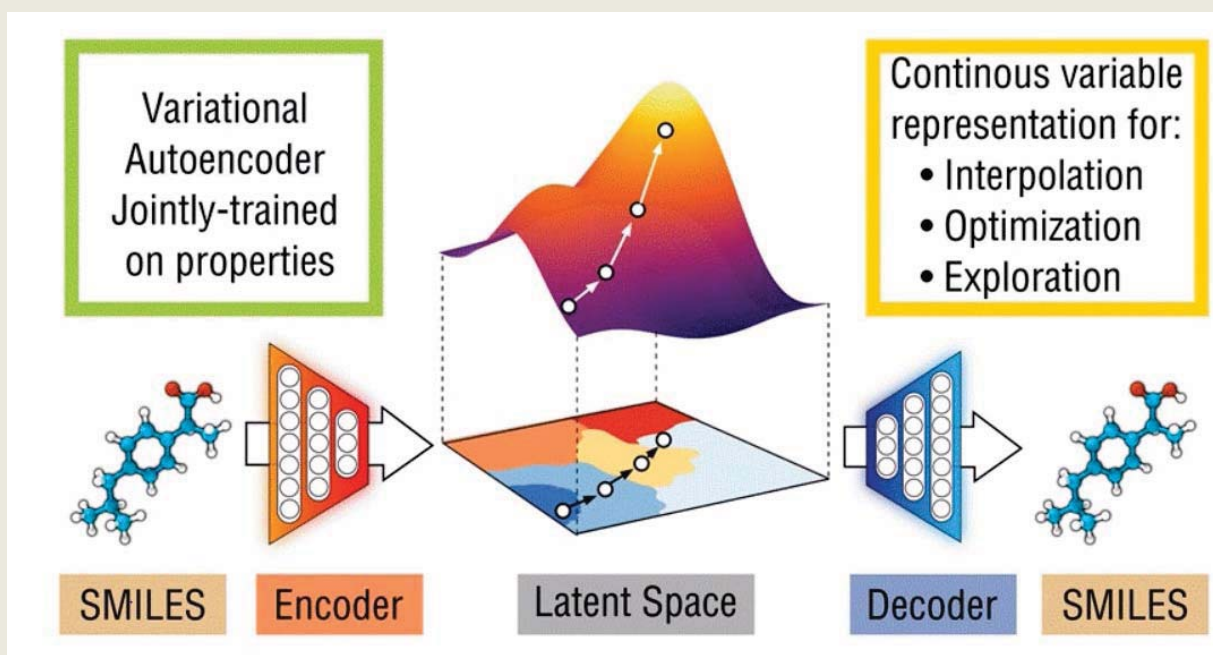


And, more

- ❑ 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?