KSBi-BIML 2021

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists

생물정보학 & 머쉰러닝 워크샵(온라인)

Pharmacogenomics in Drug Discovery and Development

남호정

SBi 한국생명정보학회 Korean Society for Bioinformatics
서울대학교 생물정보연구소 SNU BIOINFORMATICS INSTITUTE
Gene·2·us

Bioinformatics & Machine Learning for Life Scientists

BIML-2021

안녕하십니까?

한국생명정보학회의 동계 워크샵인 BIML-2021을 2월 15부터 2월 19일까지 개최합니다. 생명정보학 분야의 융합이론 보급과 실무역량 강화를 위해 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하였으며 올해로 7차를 맞이하게 되었습니다. 유례가 없는 코로나 대유행으로 인해 올해의 BIML 워크숍은 온라인으로 준비했습니다. 생생한 현장 강의에서만 느낄 수 있는 강의자와 수강생 사이의 상호교감을 가질 수 없다는 단점이 있지만, 온라인 강의의 여러 장점을 살려서 최근 생명정보학에서 주목받고 있는 거의 모든 분야를 망라한 강의를 준비했습니다. 또한 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다.

BIML 워크샵은 전통적으로 크게 생명정보학과 AI, 두 개의 분야로 구성되어오고 있으며 올해 역시 유사한 방식을 채택했습니다. AI 분야는 Probabilistic Modeling, Dimensionality Reduction, SVM 등과 같은 전통적인 Machine Learning부터 Deep Learning을 이용한 신약개발 및 유전체 연구까지 다양한 내용을 다루고 있습니다. 생명정보학 분야로는, Proteomics, Chemoinformatics, Single Cell Genomics, Cancer Genomics, Network Biology, 3D Epigenomics, RNA Biology, Microbiome 등 거의 모든 분야가 포함되어 있습니다. 연사들은 각 분야 최고의 전문가들이라 자부합니다.

이번 BIML-2021을 준비하기까지 너무나 많은 수고를 해주신 BIML-2021 운영위원회의 김태민 교수님, 류성호 교수님, 남진우 교수님, 백대현 교수님께 커다란 감사를 드립니다. 또한 재정적 도움을 주신, 김선 교수님 (AI-based Drug Discovery), 류성호 교수님, 남진우 교수님께 감사를 표시하고 싶습니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 강의자료를 만드는데 노력하셨을 뿐만 아니라 실시간 온라인 Q&A 세션까지 참여해 수고해 주시는 모든 연사분들께 깊이 감사드립니다.

2021년 2월

한국생명정보학회장 김동섭

강의개요

Pharmacogenomics in drug discovery and development

약물유전체학이란(pharmacogenomics) 유전체(genome) 수준에서 염기서열의 차이 또는 유전자 발현 차이를 분석하여 개개인이 갖는 약물 반응의 차이를 규명하는 연구분야이다. 본 수업에서는 이러한 개인별 약물 반응성을 고려한 약물 개발 과정에 대하여 알아보고 또한 개인별 유전자에 따른 약물 반응을 연구/예측하는데 필요한 생명정보학적 접근 방식을 알아본다. 구체적으로는 약물유전체학에 대한 기본 개념을 이해하고, 연구에 필요한 다양한 데이터베이스와 기본적인 생명정보학적 알고리즘들에 대해서 다룬다.

강의는 다음의 내용을 포함한다:

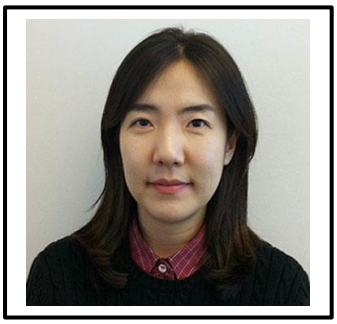
- Pharmacogenomics 기본 개념
- Drug discovery and development 기본 개념
- Protein representation features
- Molecular representation features
- 개인별 유전자 정보를 이용한 다양한 약물 개발 연구 소개

* **교육생준비물:**
  강의 동영상 플레이가 가능한 컴퓨터

* **강의: 남호정 교수 (광주과학기술원 전기전자컴퓨터공학부)**

Curriculum Vitae

Speaker Name: Hojung Nam, Ph.D.



▶**Personal Info**

Name Hojung Nam

Title Associate Professor

Affiliation Gwangju Institute of Science and Technology (GIST)

▶**Contact Information**

123 Cheomdangwagi-ro, Buk-gu, Gwangju, 61005, Republic of Korea

Email hjnam@gist.ac.kr

Phone Number 062-715-2641

Research interest : Bioinformatics, Systems Biology, Cheminformatics, Machine learning

Educational Experience

2001	B.S. in Computer Science, Sogang Univ., Seoul, Korea.
2003	M.S. in Computer Science, KAIST, Daejeon, Korea.
2009	Ph.D. in Bio and Brain Engineering, KAIST, Daejeon, Korea.

Professional Experience

2009-2013	Postdoctoral Researcher, Bioengineering, University of California, San Diego, CA USA
2013-2018	Assistant Professor, Gwangju Institute of Science and Technology (GIST)
2018-	Associate Professor, Gwangju Institute of Science and Technology (GIST)

Selected Publications (Recent two years, CA only)

1. Hyunho Kim, Eunyoung Kim, Ingoo Lee, Bongsung Bae, Minsu Park, Hojung Nam*, " Artificial Intelligence in Drug Discovery: A Comprehensive Review of Data-Driven and Machine Learning Approaches", Biotechnology and Bioprocess Engineering, volume 25, pages895–930(2020).
2. Hyunho Kim, Hojung Nam*, "hERG-Att: Self-Attention-Based Deep Neural Network for Predicting hERG Blockers", Computational Biology and Chemistry, Available online 19 May 2020, 107286.
3. Soobok Joe , Hojung Nam*, "Prediction model construction of stem cell pluripotency using CpG and non-CpG DNA methylation markers", BMC Bioinformatics, 2020 21:175.
4. Heeyeon Choi, Soobok Joe, Hojung Nam*, "Development of Tissue-Specific Age Predictors Using DNA Methylation Data", Genes 2019, 10(11), 888.
5. Ingoo Lee, Jongsoo Keum, Hojung Nam*, "DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences", PLoS Computational Biology 15(6): e1007129. https://doi.org/10.1371/journal.pcbi.1007129

KSBi-BIML 2021

Pharmacogenomics in drug discovery and development

Hojung Nam, Ph.D.
Associate Professor
School of Electrical Engineering and Computer Science (EECS)
Gwangju Institute of Science and Technology (GIST)
Contact: hjnam@gist.ac.kr

Contents

- PART1
  - Introduction to pharmacogenomics
    - Drug discovery and development
  - Key data sources
  - Representations of proteins, chemicals
- PART2
  - Studies related to pharmacogenomics based on machine learning

- PART1
  - Introduction to pharmacogenomics
    - Drug discovery and development
  - Key data sources
  - Representations of proteins, chemicals
- PART2
  - Studies related to pharmacogenomics based on machine learning

INTRODUCTION TO PHARMACOGENOMICS

Pharmacogenomic

- The term **pharmacogenetics** was coined in the 1950s and captures the idea that large effect size DNA variants contribute importantly to variable drug actions in an individual (single gene-drug).
- The term **pharmacogenomics** is now used by many to describe the idea that multiple variants across the genome that can differ across populations affect drug response. The International Conference on Harmonisation, a worldwide consortium of regulatory agencies, has defined **pharmacogenomics as the study of variations of DNA and RNA characteristics as related to drug response.**

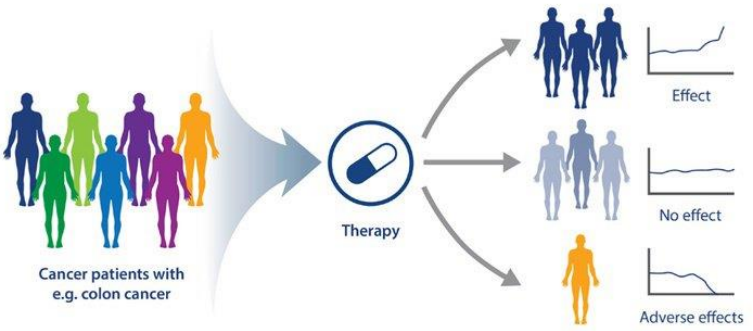
Dan M Roden et al., Lancet . 2019 Aug 10;394(10197):521-532.

Look for genetic variants that affect drug response used to treat the condition. The analysis will yield results that allow physicians to determine if their patient will have a positive response to the drug treatment.
[National Human Genome Research Institute]

Pharmacogenomics Adds Precision to the Practice of Medicine, June 15, 2015 (Vol. 35, No. 12)
https://www.genengnews.com/magazine/249/pharmacogenomics-adds-precision-to-the-practice-of-medicine/

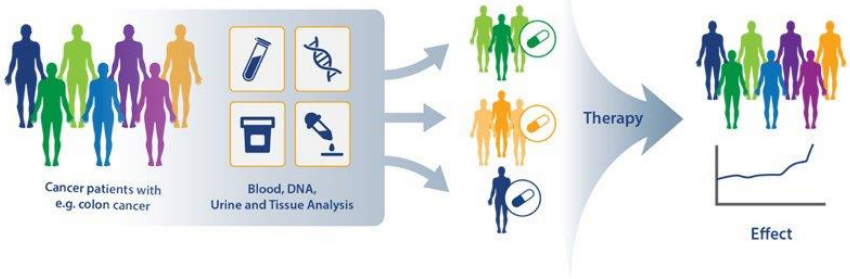
Current Medicine

One Treatment Fits All



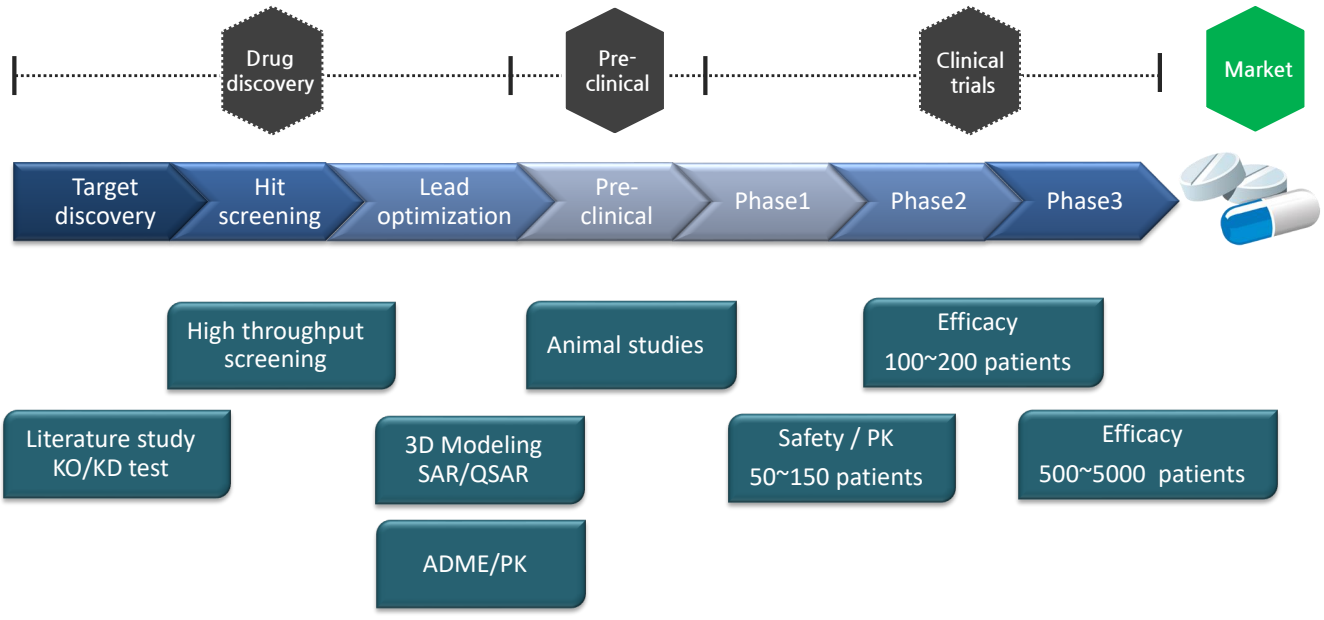
Future Medicine

More Personalized Diagnostics

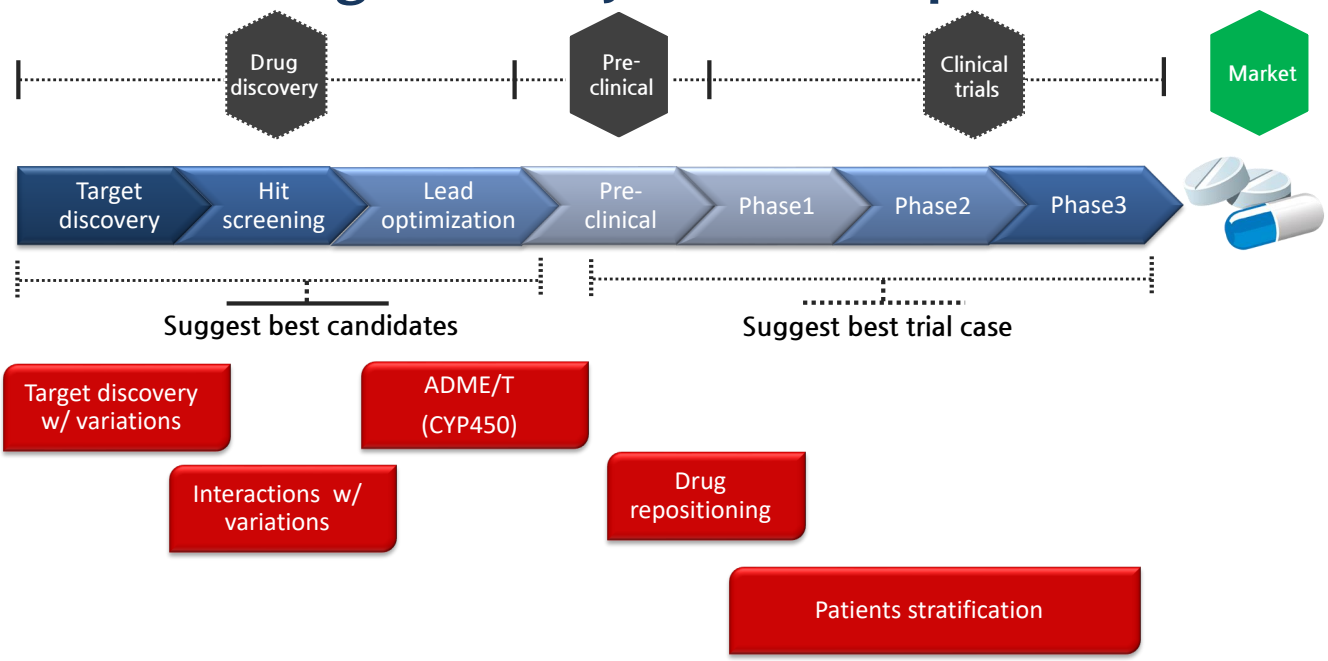


https://blog.crownbio.com/pdx-personalized-medicine#_

Drug discovery and development



Pharmacogenomics in drug discovery and development



Target discovery w/ variations

ADME/T (CYP450)

Interactions w/ variations

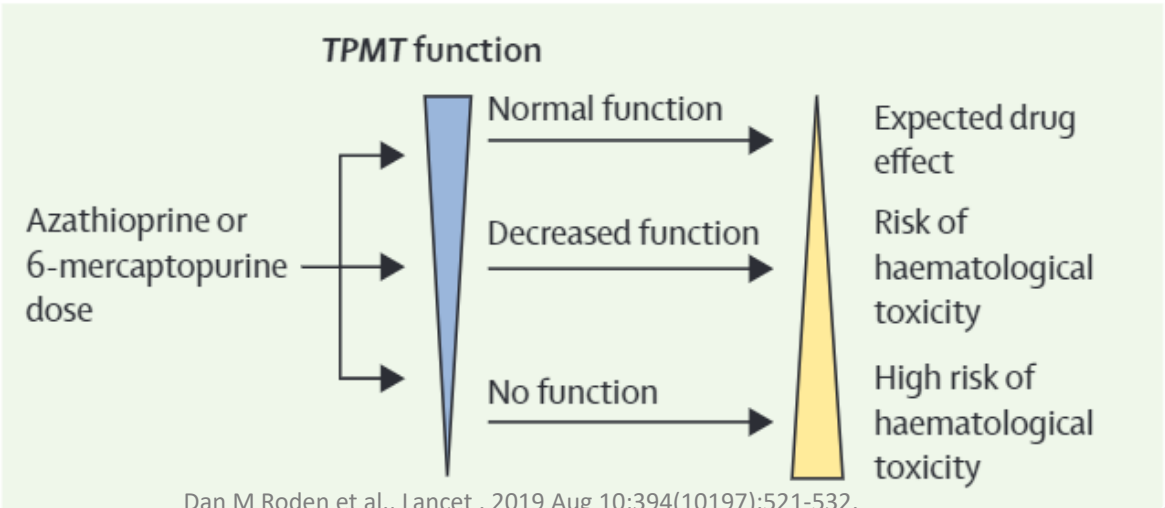
Drug repositioning

Patients stratification

Example 1 – TPMT

Pharmacogenetics in Oncology

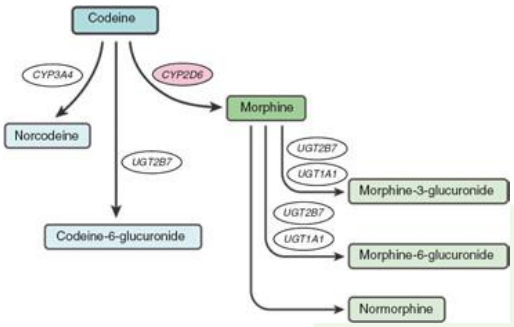
- The thiopurine S-methyltransferase (TPMT) is a metabolizer of chemotherapeutic agents 6MP and azothiopurine (used mainly in blood-based malignancies)
- TPMT deficiency leads to severe toxicity associated with treatment (potential mortality)

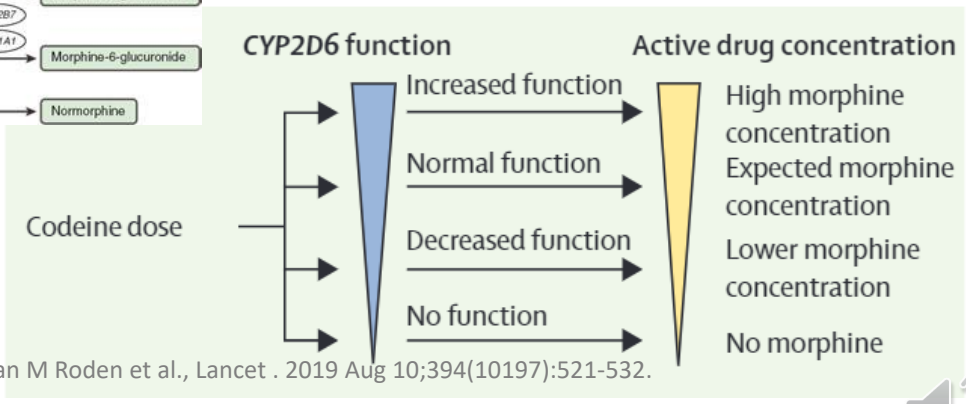


Dan M Roden et al., Lancet . 2019 Aug 10;394(10197):521-532.

Example 2 – CYP2D6

- Cytochrome P450 2D6 (CYP2D6) is an enzyme that in humans is encoded by the CYP2D6 gene. CYP2D6 is primarily expressed in the liver.
- In particular, CYP2D6 is responsible for the metabolism and elimination of approximately 25% of clinically used drugs, via the addition or removal of certain functional groups – specifically, hydroxylation, demethylation, and dealkylation. CYP2D6 also activates some prodrugs.





Dan M Roden et al., Lancet . 2019 Aug 10;394(10197):521-532.

- PART1
  - Introduction to pharmacogenomics
    - Drug discovery and development
  - Key data sources
  - Representations of proteins, chemicals
- PART2
  - Studies related to pharmacogenomics based on machine learning

KEY DATA RESOURCES

SNP (단일염기다형성)

Single-nucleotide polymorphism

From Wikipedia, the free encyclopedia

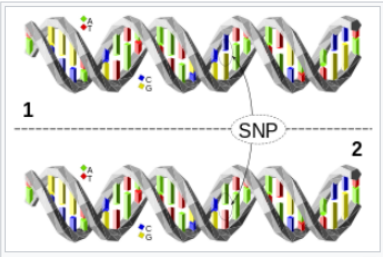
This article's **use of external links may not follow Wikipedia's policies or guidelines**. Please improve this article by removing excessive or inappropriate external links, and converting useful links where appropriate into footnote references. *(October 2012) (Learn how and when to remove this template message)*

A **single-nucleotide polymorphism**, often abbreviated to **SNP** (/snɪp/; plural /snɪps/), is a variation in a single nucleotide that occurs at a specific position in the genome, where each variation is present to some appreciable degree within a population (e.g. > 1%).[1]

For example, at a specific base position in the human genome, the C nucleotide may appear in most individuals, but in a minority of individuals, the position is occupied by an A. This means that there is a SNP at this specific position, and the two possible nucleotide variations – C or A – are said to be alleles for this position.

SNPs underlie differences in our susceptibility to disease; a wide range of human diseases, e.g. sickle-cell anemia, β-thalassemia and cystic fibrosis result from SNPs.[2][3][4] The severity of illness and the way the body responds to treatments are also manifestations of genetic variations. For example, a single-base mutation in the APOE (apolipoprotein E) gene is associated with a lower risk for Alzheimer's disease.[5]

A **single-nucleotide variant** (**SNV**) is a variation in a single nucleotide without any limitations of frequency and may arise in somatic cells. A somatic single-nucleotide variation (e.g., caused by cancer) may also be called a **single-nucleotide alteration**.

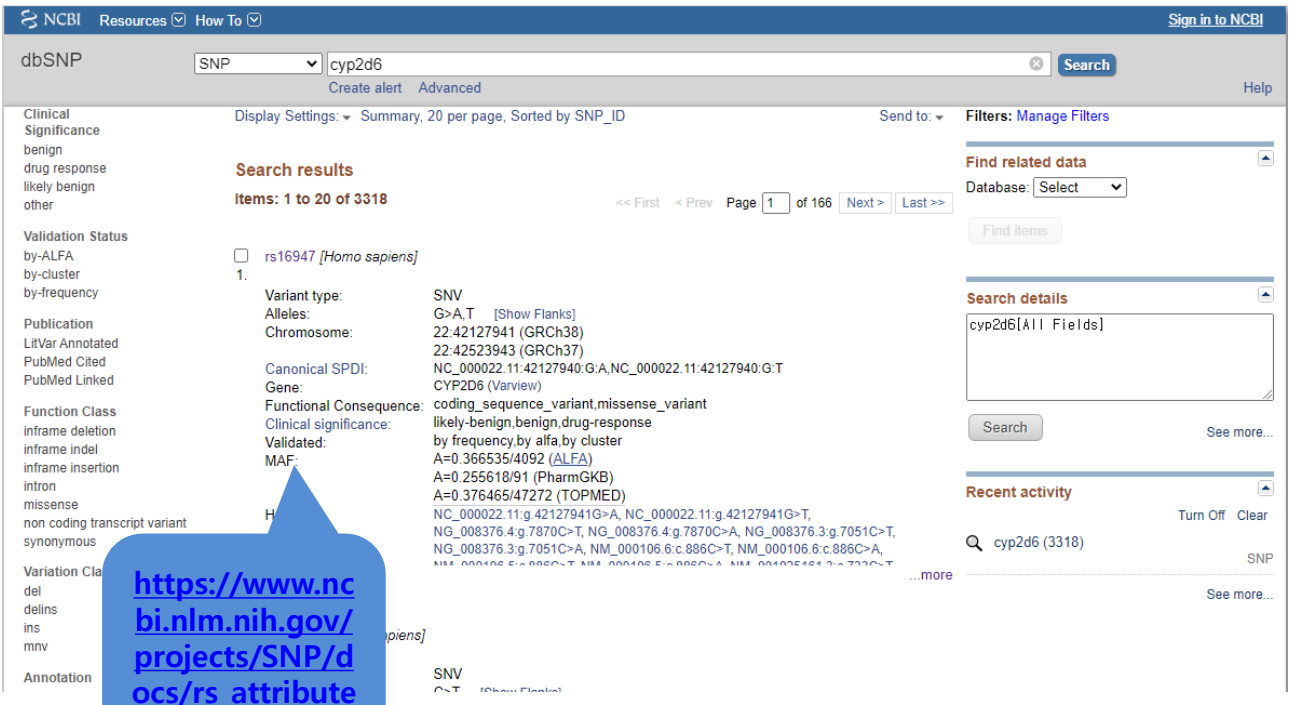


The upper DNA molecule differs from the lower DNA molecule at a single base-pair location (a C/A polymorphism)

https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism

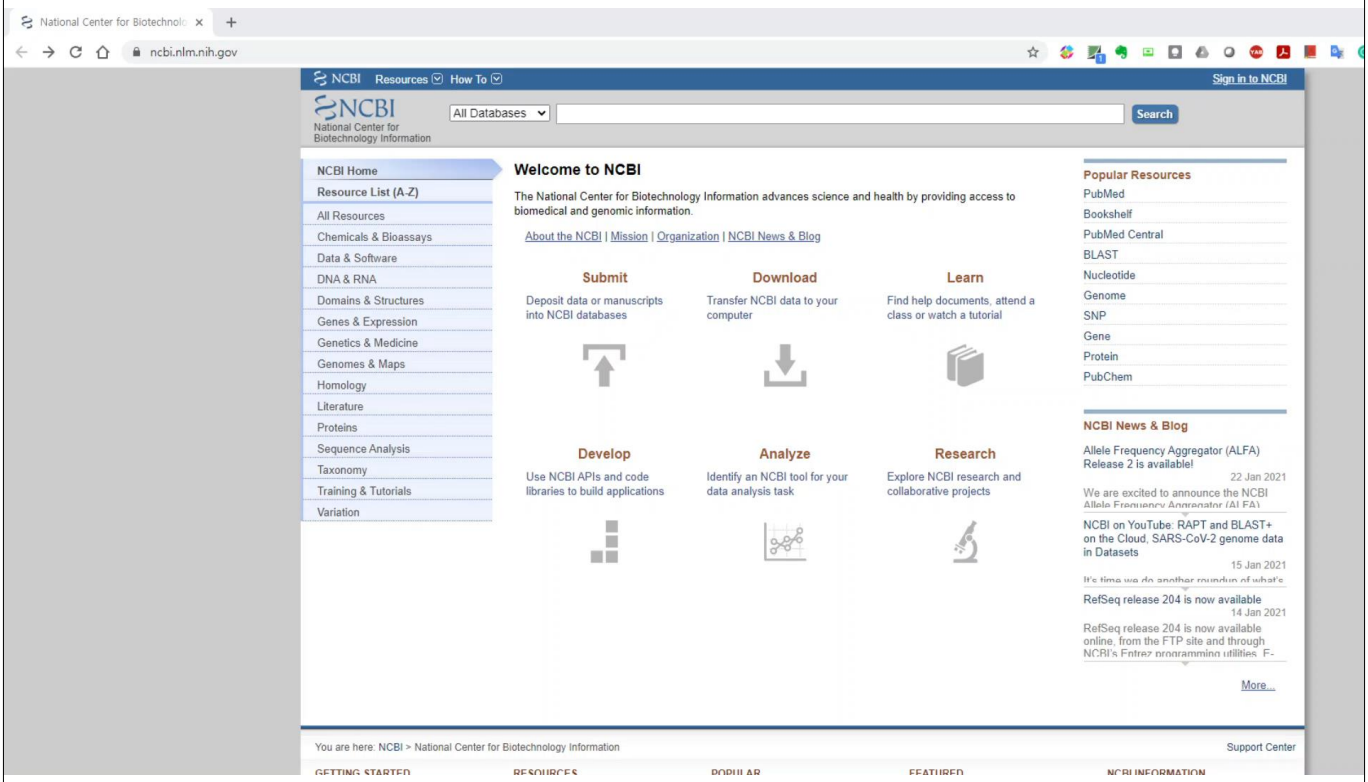
13

NCBI dbSNP



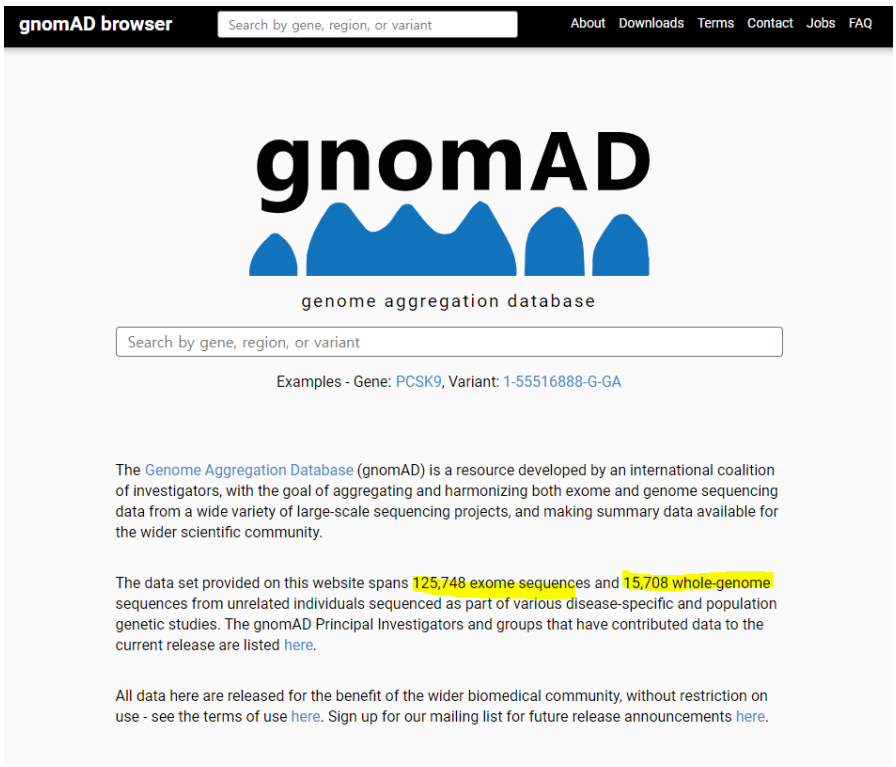
https://www.ncbi.nlm.nih.gov/snp/?term=cyp2d6

14



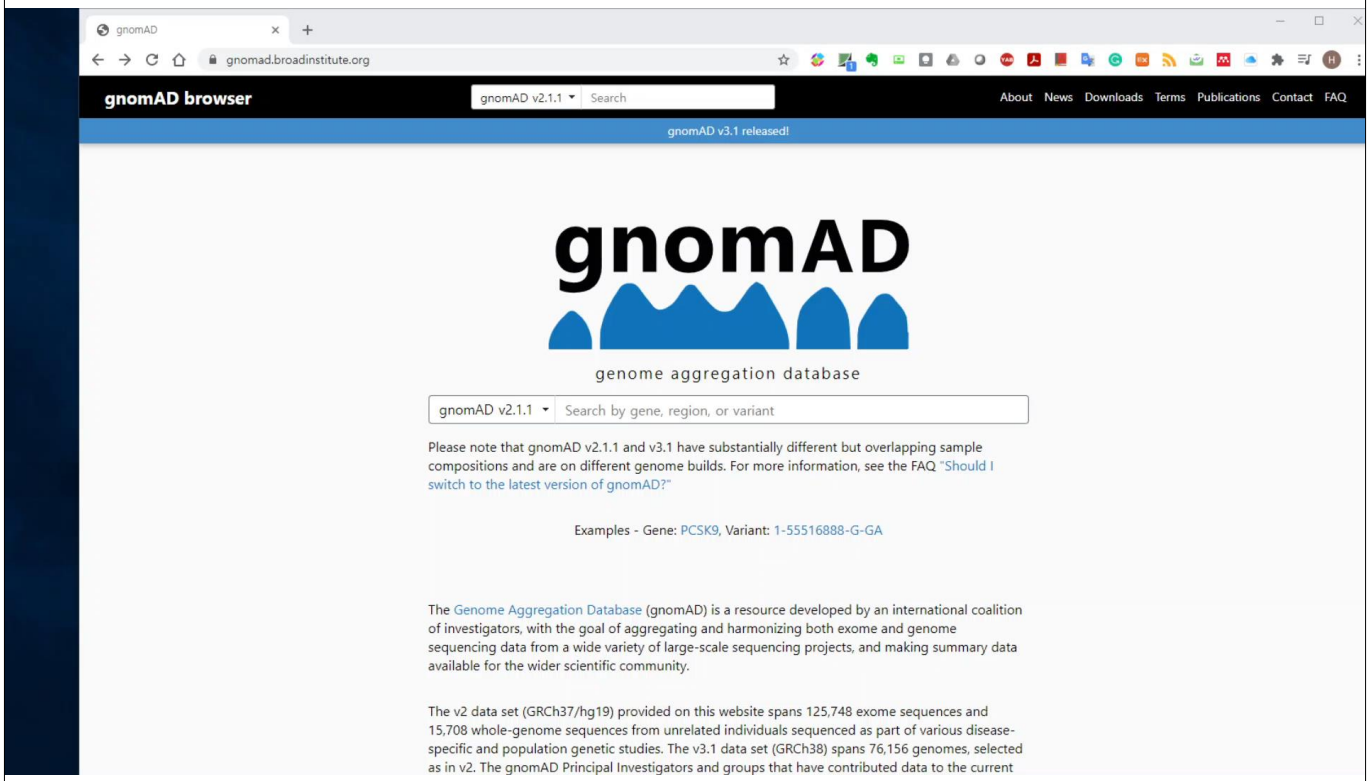
15

gnomAD



https://gnomad.broadinstitute.org/

15



https://gnomad.broadinstitute.org/

The Human Cytochrome P450 (*CYP*) Allele Nomenclature Database

Allele nomenclature for Cytochrome P450 enzymes

New List: CYP allele frequencies from 56,945 unrelated individuals of five major human populations

Inclusion criteria - **New criteria regarding variants identified by NGS**

iRAMP, calculator of contribution of rare variants.

Cytochrome P450 Oxidoreductase: *POR*

CYP1 family:
CYP1A1; *CYP1A2*; *CYP1B1*

CYP2 family:
CYP2A6; *CYP2A13*; *CYP2B6*; *CYP2C8*; *CYP2C9*; *CYP2C19*; *CYP2D6*; *CYP2E1*; *CYP2F1*; *CYP2J2*; *CYP2R1*; *CYP2S1*; *CYP2W1*

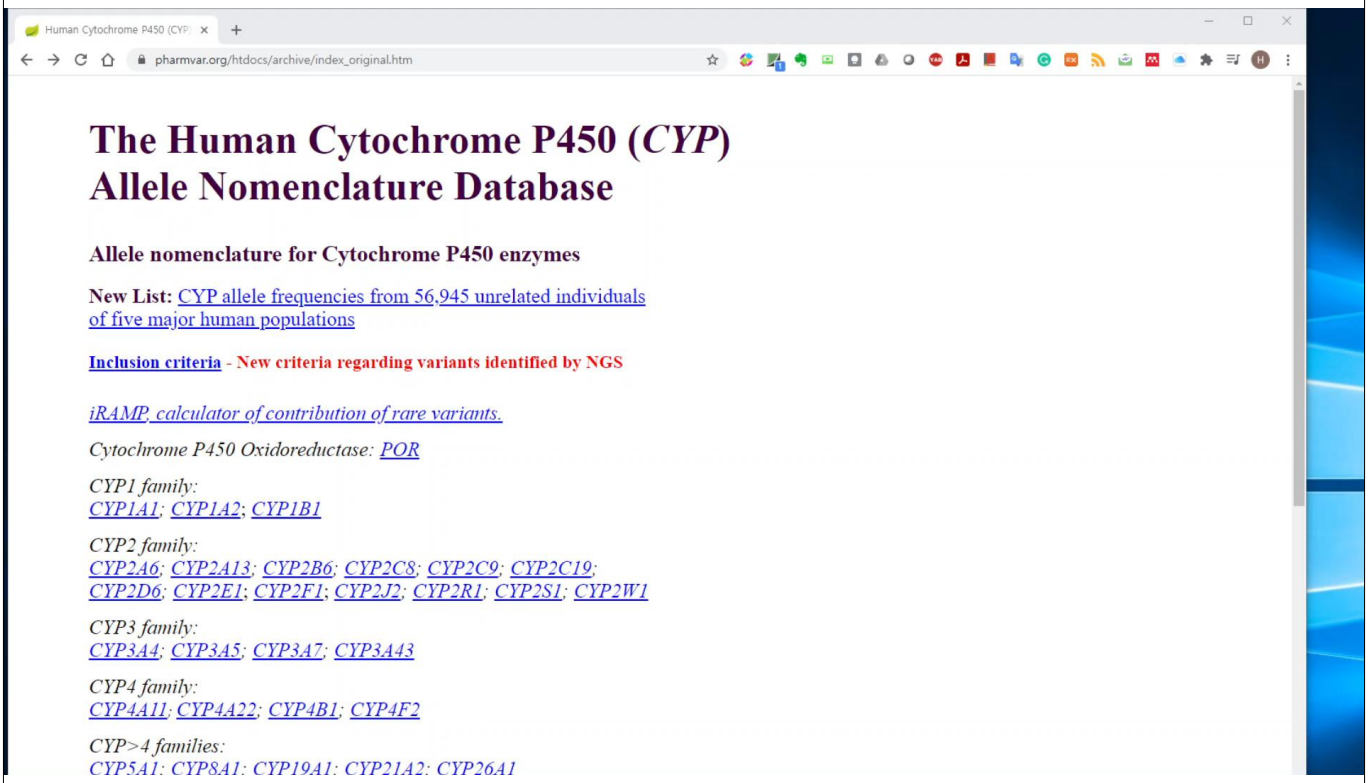
CYP3 family:
CYP3A4; *CYP3A5*; *CYP3A7*; *CYP3A43*

CYP4 family:
CYP4A11; *CYP4A22*; *CYP4B1*; *CYP4F2*

CYP>4 families:
CYP5A1; *CYP8A1*; *CYP19A1*; *CYP21A2*; *CYP26A1*

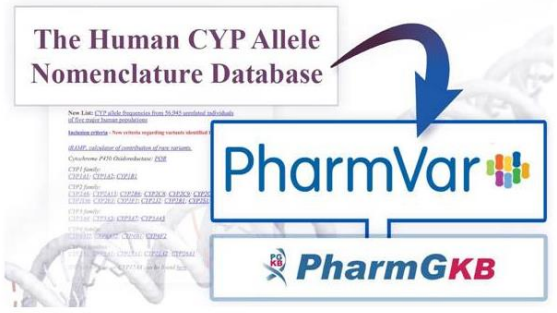
SNP information on ***CYP17A1*** *can be found here*

https://www.pharmvar.org/htdocs/archive/index_original.htm



https://www.pharmvar.org/htdocs/archive/index_original.htm

PharmVar



After more than 15 years the Human Cytochrome P450 (*CYP*) Allele Nomenclature Database has transitioned…

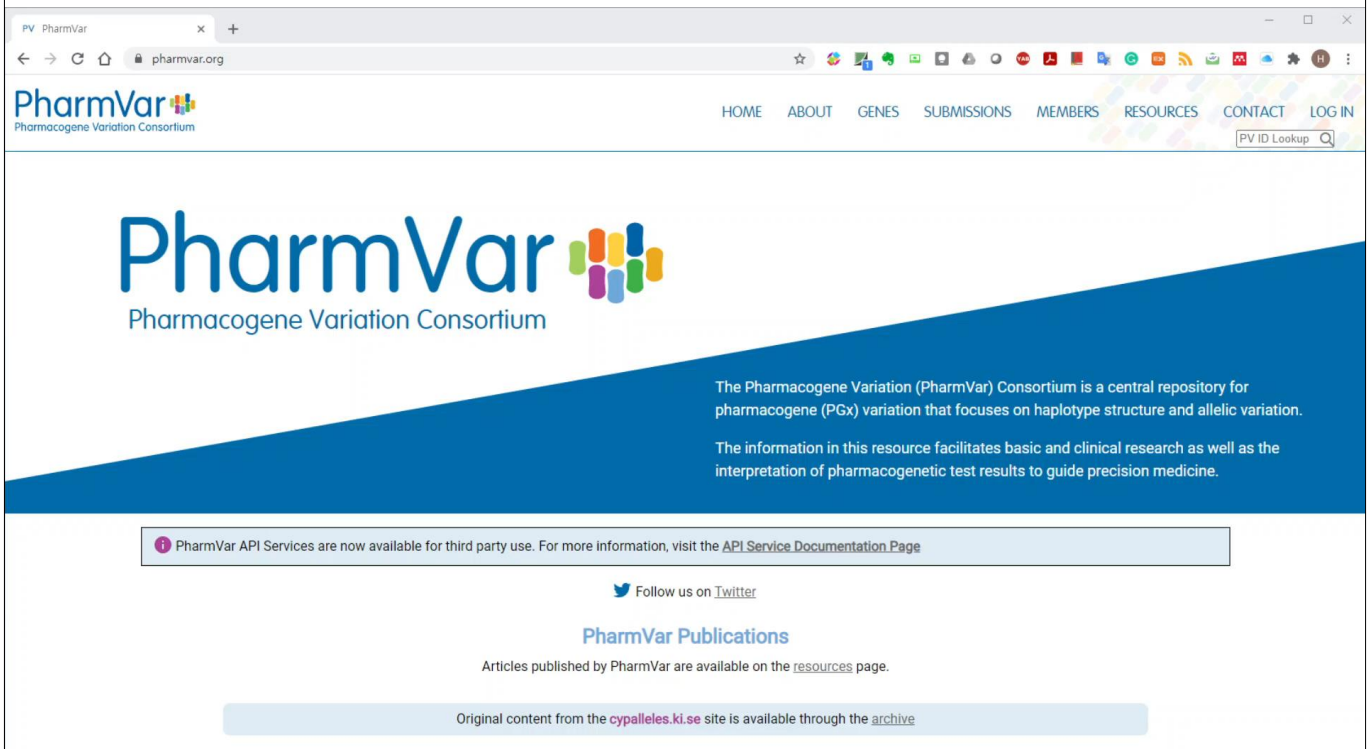
…to the **Pharm**acogene **Var**iation (**PharmVar**) Consortium at www.PharmVar.org

PharmVar will serve as a central repository for pharmacogene variation to facilitate allele (haplotype) designation and the interpretation of pharmacogenetic test results to guide precision medicine

PharmVar is a PGRN resource funded by NIGMS.

After September 26, 2017, please visit www.PharmVar.org to access content of the original P450 Nomenclature Database

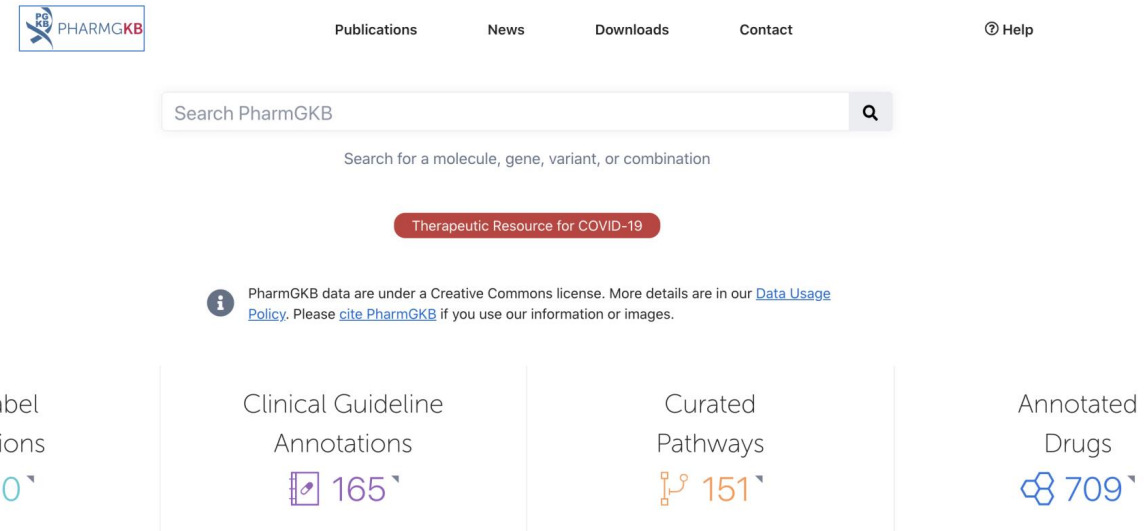
http://www.cypalleles.ki.se/



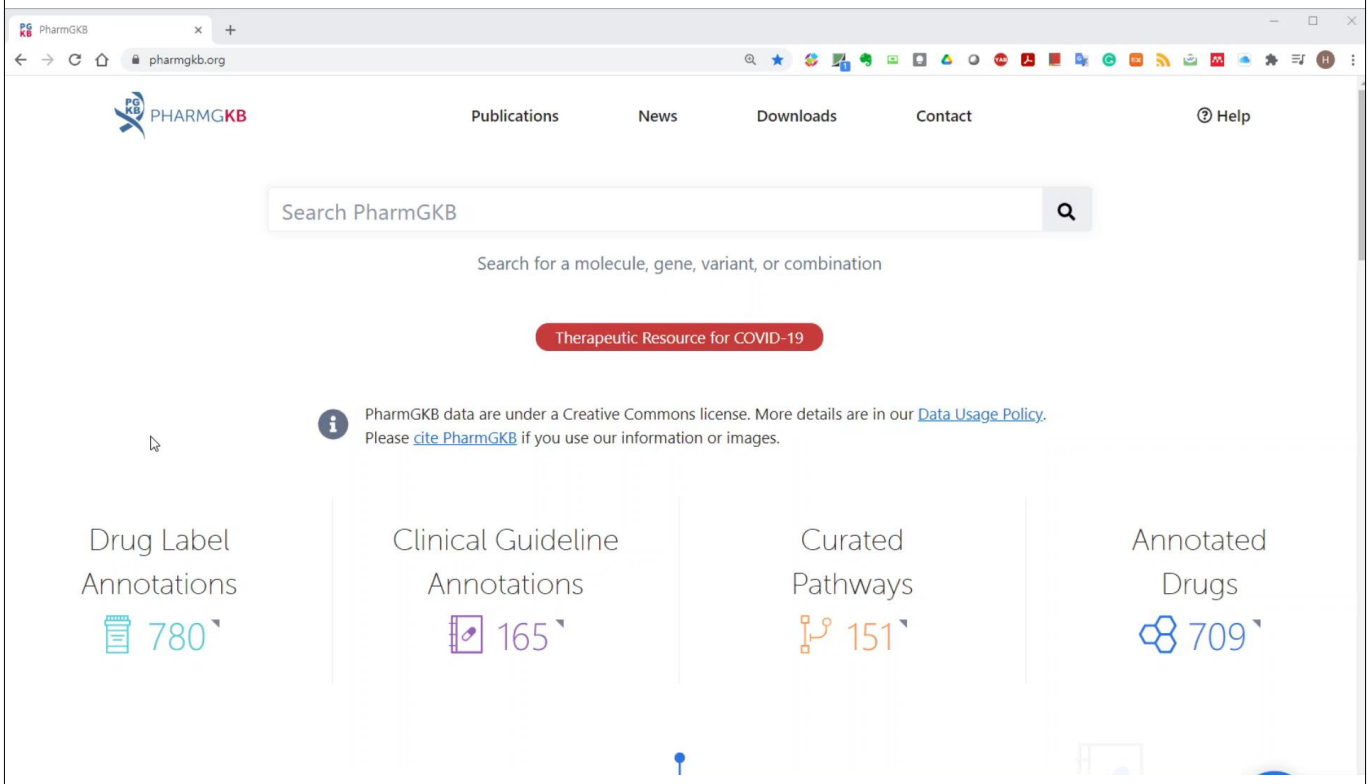
https://www.pharmvar.org/

21

PHARMGKB



https://www.pharmgkb.org/



https://www.pharmgkb.org/

Resources for pan-cancer genomics profiles and tools

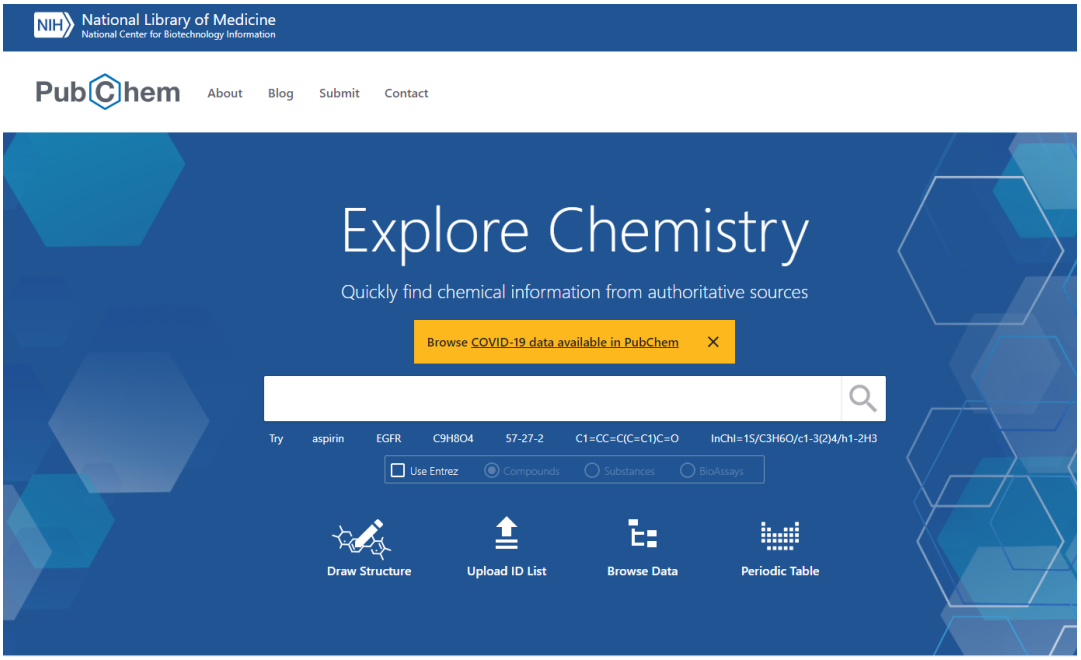
Table 2. Resources for pan-cancer genomics profiles and tools

Resource	Data type	Profiling platform	Sample size	Description	Link	References
Adult cancers						
TCGA (The Cancer Genome Atlas)	Clin, CNA, GEX, Methyl, miEX, SNV	Microarray, NGS	~11 300	Mostly primary tumors of 33 cancers	Individual cancers: https://portal.gdc.cancer.gov/ Merged pan-cancer data: https://gdc.cancer.gov/node/905/ Also downloadable by an R/Bioconductor package TCGAbiolinks [41]	[150]
MET500	CNA, SNV	NGS	500	Metastatic tumors of 30 cancers	https://met500.path.med.umich.edu/	[43]
Pediatric cancers						
TARGET (Therapeutically Applicable Research to Generate Effective Treatments)	Clin, GEX, miEX, SNV	NGS	~3200 (according to the GDC Data Portal accessed in May 2018)	6 pediatric cancers (according to the GDC Data Portal accessed in May 2018)	https://portal.gdc.cancer.gov/ Also downloaded by an R/Bioconductor package TCGAbiolinks [41]	[44]
PedPanCan (Pediatric Pan-Cancer study)	SNV	NGS	961	24 pediatric cancers	http://www.pedpancan.com	[45]
Cancer cell lines						
CCLE (Cancer Cell Line Encyclopedia)	CNA, GEX, RPPA, SNV	Microarray, NGS	~1500		https://portals.broadinstitute.org/ccle Also accessible through the Cancer Dependency Map (DepMap): https://depmap.org/portal/	[15, 151]
Curations						
ICGC (International Cancer Genome Consortium)	Clin, CNA, GEX, Methyl, miEX, SNV	Curation	~24 000	Curation of 80+ international cancer projects, including TCGA and TARGET	http://icgc.org/	[46]
COSMIC (Catalogue of Somatic Mutations in Cancer)	CNA, SNV	Curation		Summarization of cancer-related mutations across 32 000+ tumors and cancer cells curated from 25 000 papers	https://cancer.sanger.ac.uk/cosmic	[48]
Pan-cancer data visualization						
TumorMap	2D maps	Curation		Visualization of TCGA, TARGET, etc.	https://tumormap.ucsc.edu/	[47]
Gene signatures and biological pathways						
MSigDB (Molecular Signatures Database	Genes sets	Curation	~17 800 gene sets	Genes sets of cytobands, curations, motifs, computation, Gene Ontologies, oncogenic signatures and immunology	http://software.broadinstitute.org/gsea/msigdb/index.jsp	[52–54]
Pathway Commons	Biological pathways	Curation	4000+ pathways	Collection of biological pathways from 20+ databases, including KEGG and Reactome	https://www.pathwaycommons.org/	[152]
NDEx (Network Data Exchange)	Biological networks	Curation		Interactive database that allows users to query, visualize, upload, share and distribute biological networks	www.ndexbio.org/	[153]
Normal tissues						
GTEx (Genotype-Tissue Expression)	GEX	NGS	~11 700	Expression profiles of 53 non-diseased tissues across ~1000 individuals that can be used as normal controls for cancer studies	https://gtexportal.org/home/	[154, 155]

Clin, clinical data; CNA, copy number alteration; GEX, gene expression; Methyl, methylation; miEX, miRNA expression; NGS, next-generation sequencing; RPPA, reverse phase protein array; SNV, single nucleotide variant.

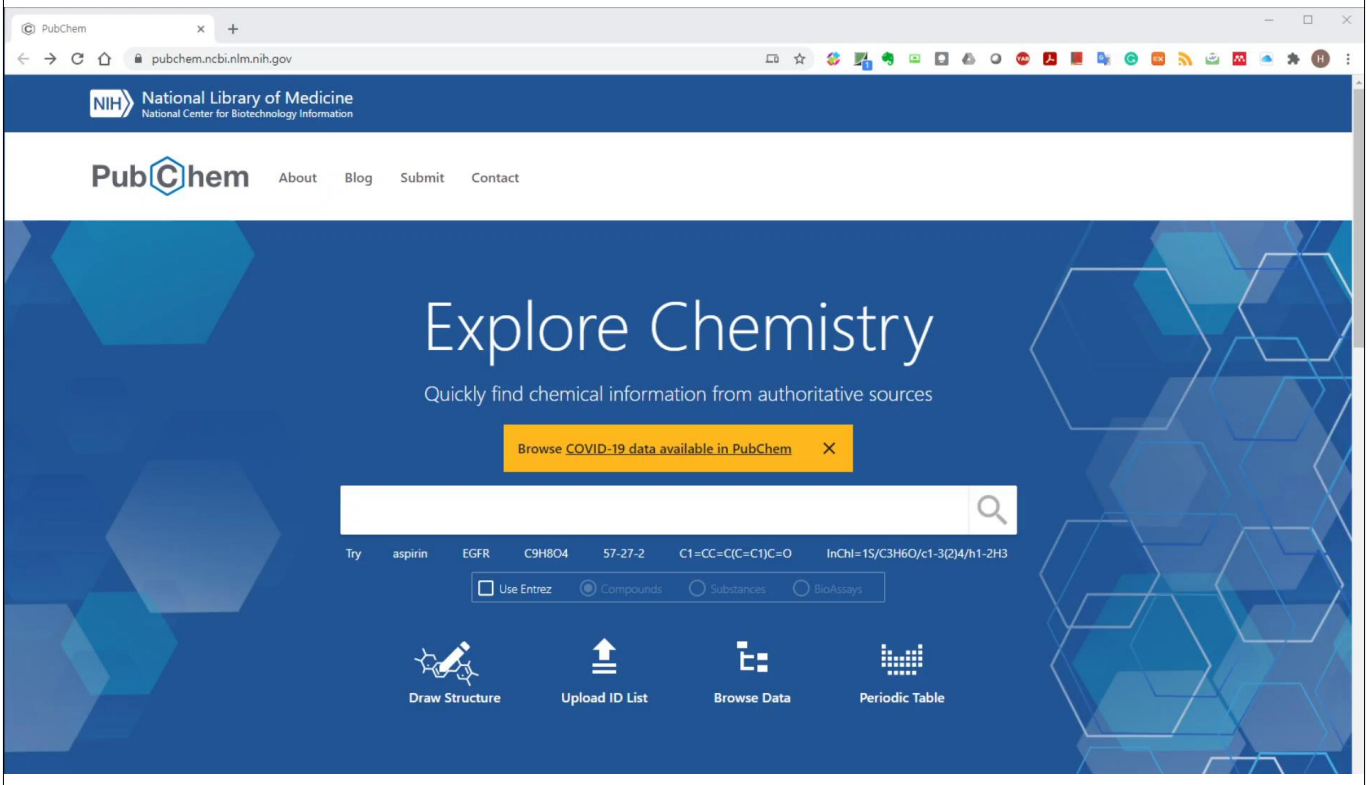
Brief Bioinform . 2020 Dec 1;21(6):2066-2083. doi: 10.1093/bib/bbz144.

NCBI PubChem



https://pubchem.ncbi.nlm.nih.gov/

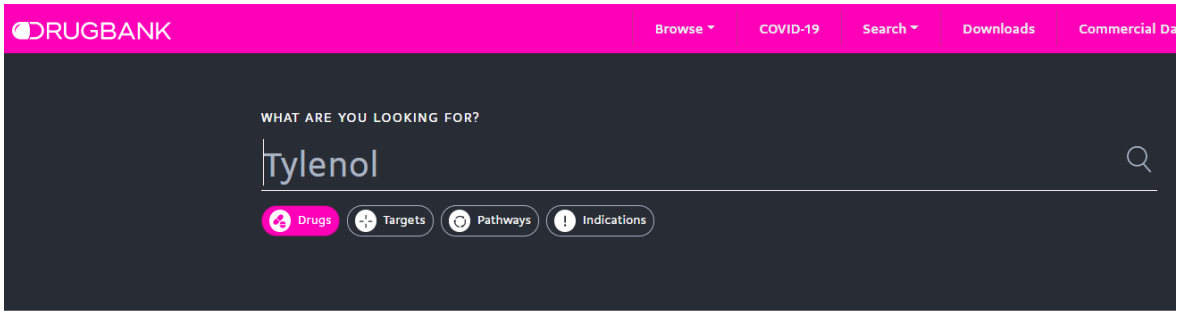
25



https://pubchem.ncbi.nlm.nih.gov/

26

DrugBank



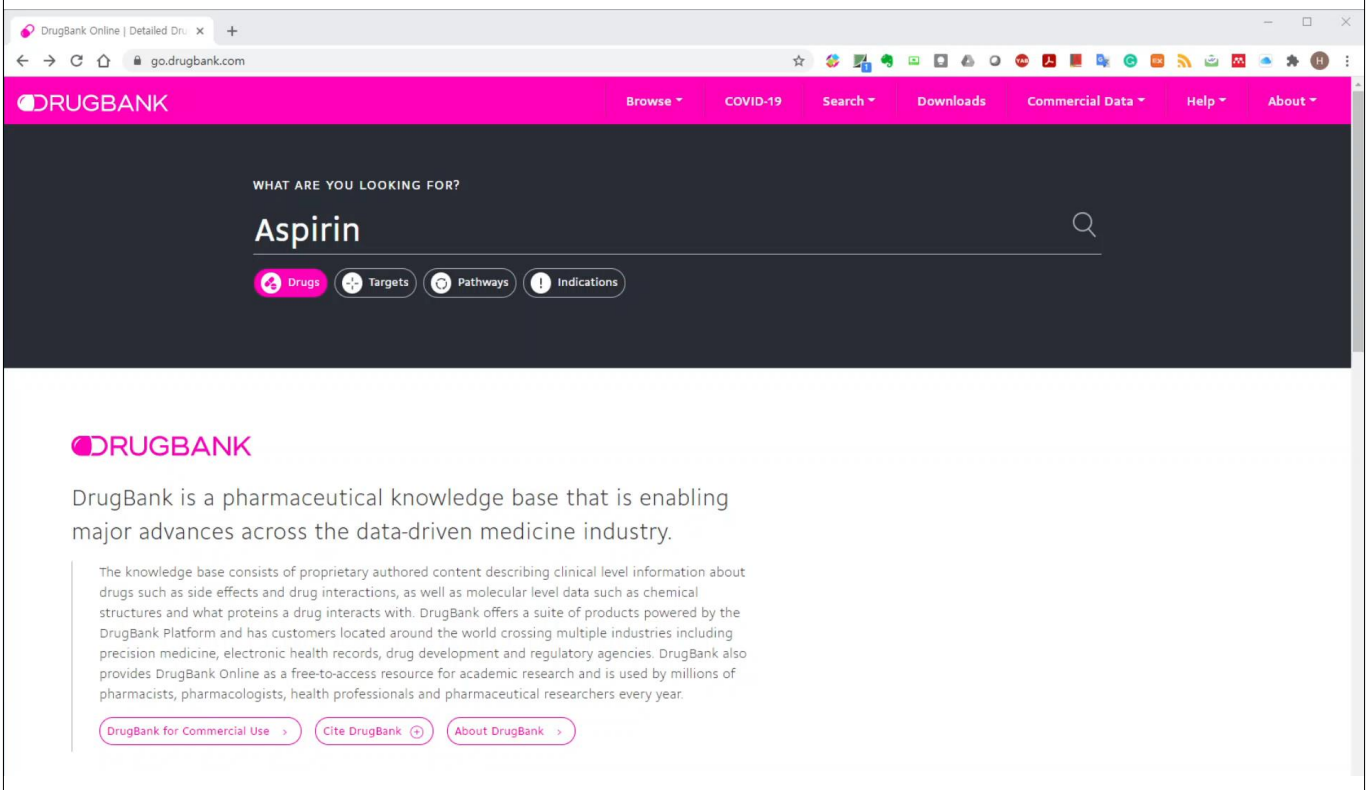
DRUGBANK

DrugBank is a pharmaceutical knowledge base that is enabling major advances across the data-driven medicine industry.

The knowledge base consists of proprietary authored content describing clinical level information about drugs such as side effects and drug interactions, as well as molecular level data such as chemical structures and what proteins a drug interacts with. DrugBank offers a suite of products powered by the DrugBank Platform and has customers located around the world crossing multiple industries including precision medicine, electronic health records, drug development and regulatory agencies. DrugBank also provides DrugBank Online as a free-to-access resource for academic research and is used by millions of pharmacists, pharmacologists, health professionals and pharmaceutical researchers every year.

DrugBank for Commercial Use › | Cite DrugBank | About DrugBank ›

https://go.drugbank.com/



DRUGBANK

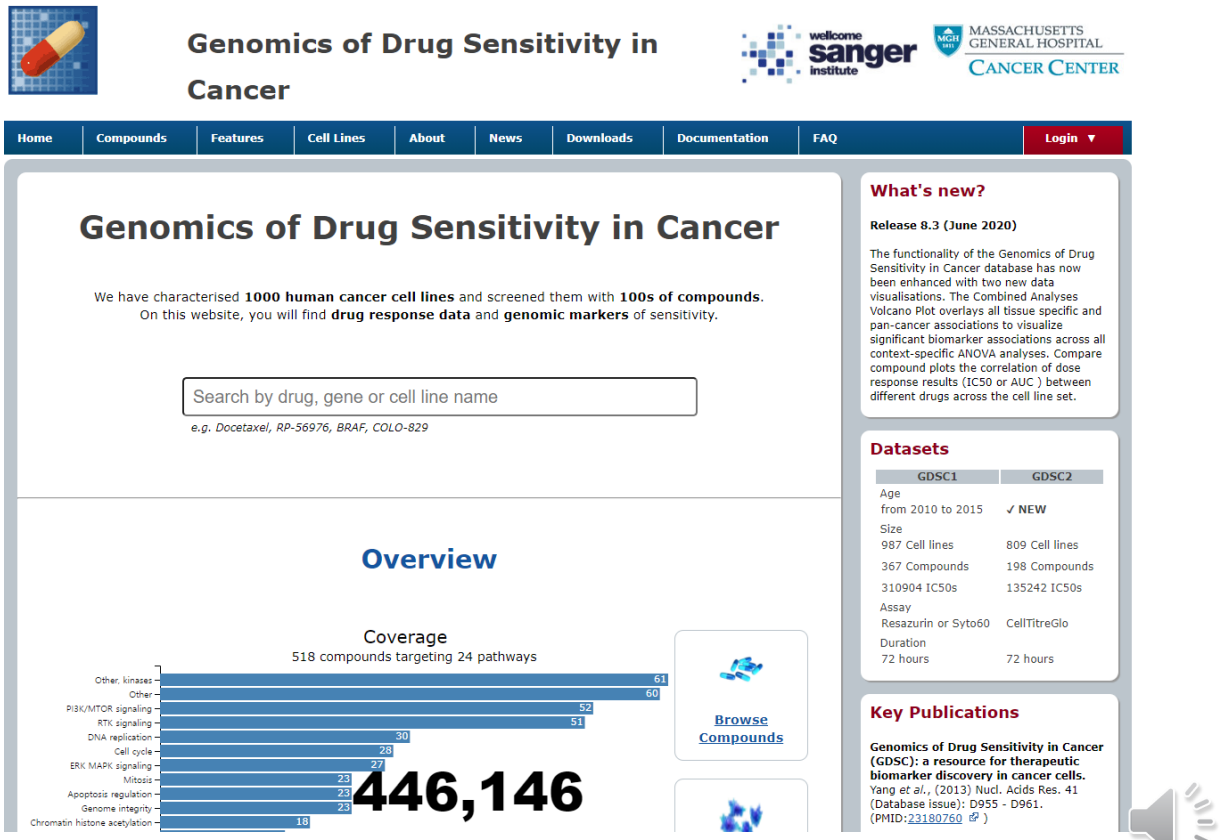
DrugBank is a pharmaceutical knowledge base that is enabling major advances across the data-driven medicine industry.

The knowledge base consists of proprietary authored content describing clinical level information about drugs such as side effects and drug interactions, as well as molecular level data such as chemical structures and what proteins a drug interacts with. DrugBank offers a suite of products powered by the DrugBank Platform and has customers located around the world crossing multiple industries including precision medicine, electronic health records, drug development and regulatory agencies. DrugBank also provides DrugBank Online as a free-to-access resource for academic research and is used by millions of pharmacists, pharmacologists, health professionals and pharmaceutical researchers every year.

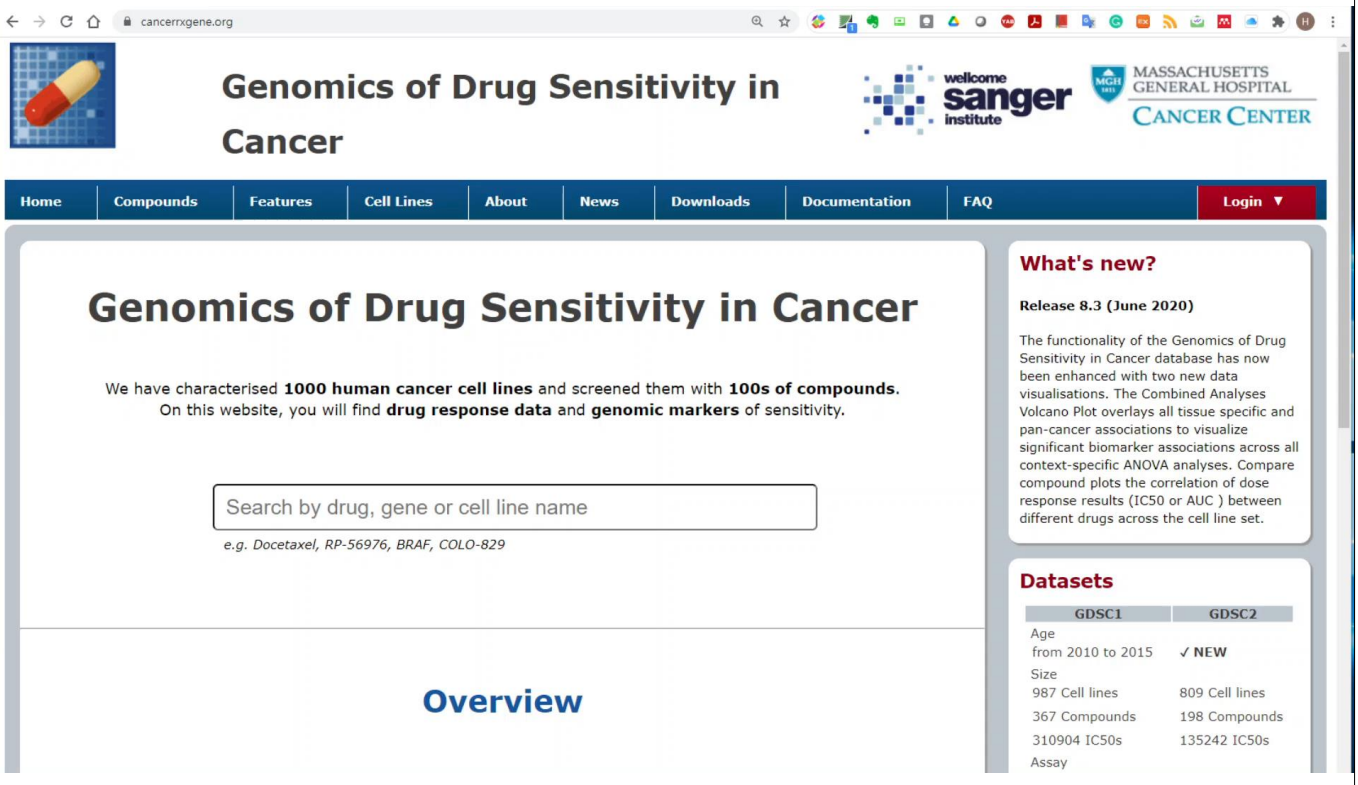
DrugBank for Commercial Use › | Cite DrugBank | About DrugBank ›

https://go.drugbank.com/

Genomics of Drug Sensitivity in Cancer (GDSC)



https://www.cancerrxgene.org/

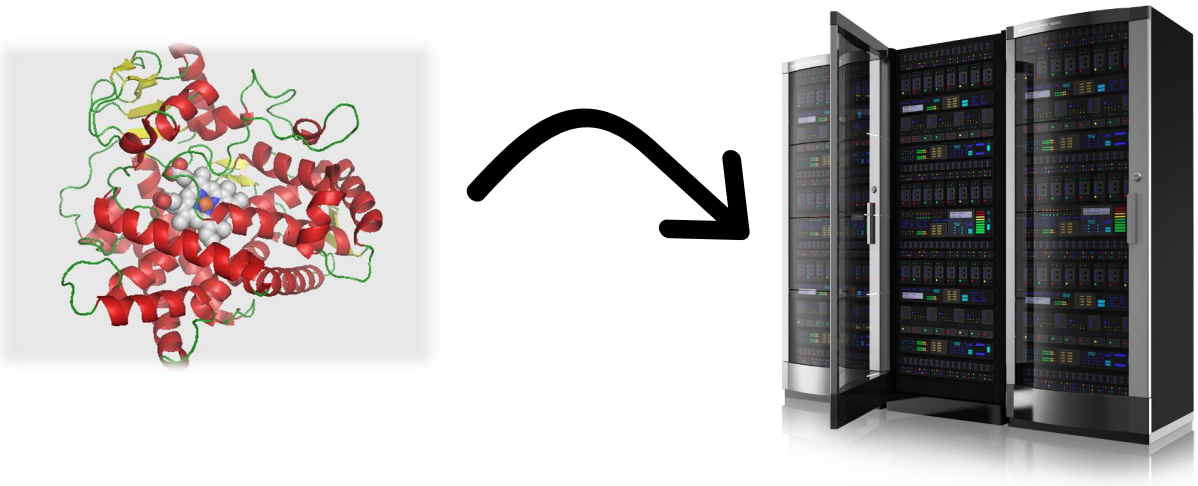


https://www.cancerrxgene.org/

- PART1
  - Introduction to pharmacogenomics
    - Drug discovery and development
  - Key data sources
  - Representations of proteins, chemicals
- PART2
  - Studies related to pharmacogenomics based on machine learning

PROTEIN REPRESENTATIONS

Why protein representations are necessary?

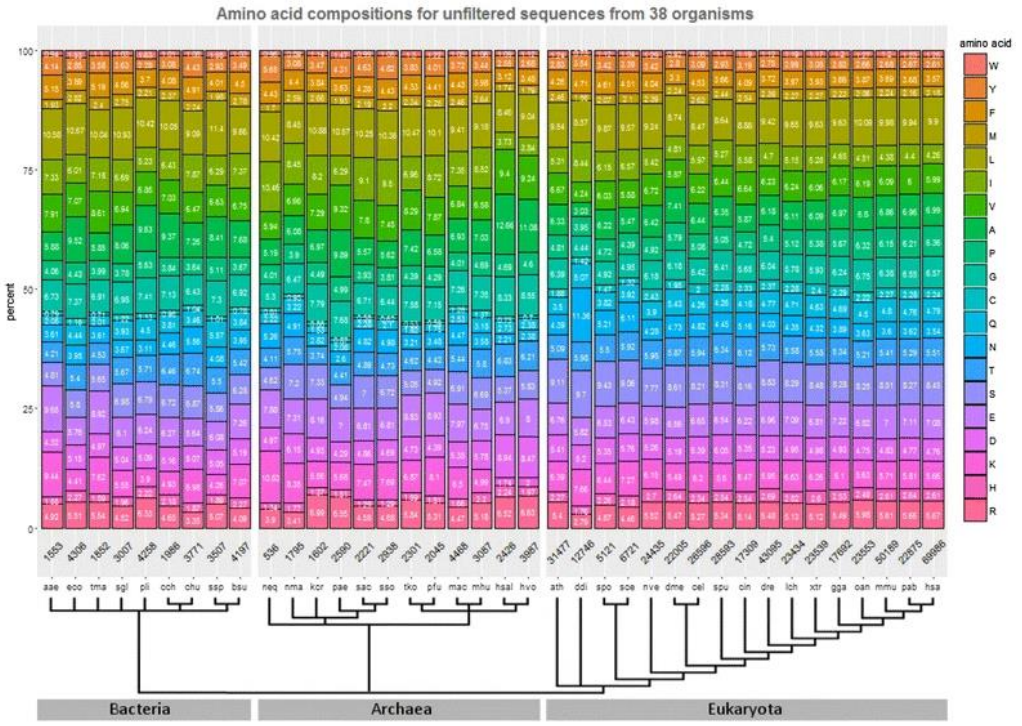


Representation of proteins for machine-learning features that fully captured wide ranges of properties of the target molecule

Types of protein representations

- Protein descriptors
  - Amino Acid Composition (AAC) - 20D
  - Dipeptide Composition Descriptor - 400D
  - Tripeptide Composition Descriptor - 8000D
  - Composition, Transition and Distribution (CTD) - 147D
- Protein embedding

Amino Acid Composition –AAC (20D)

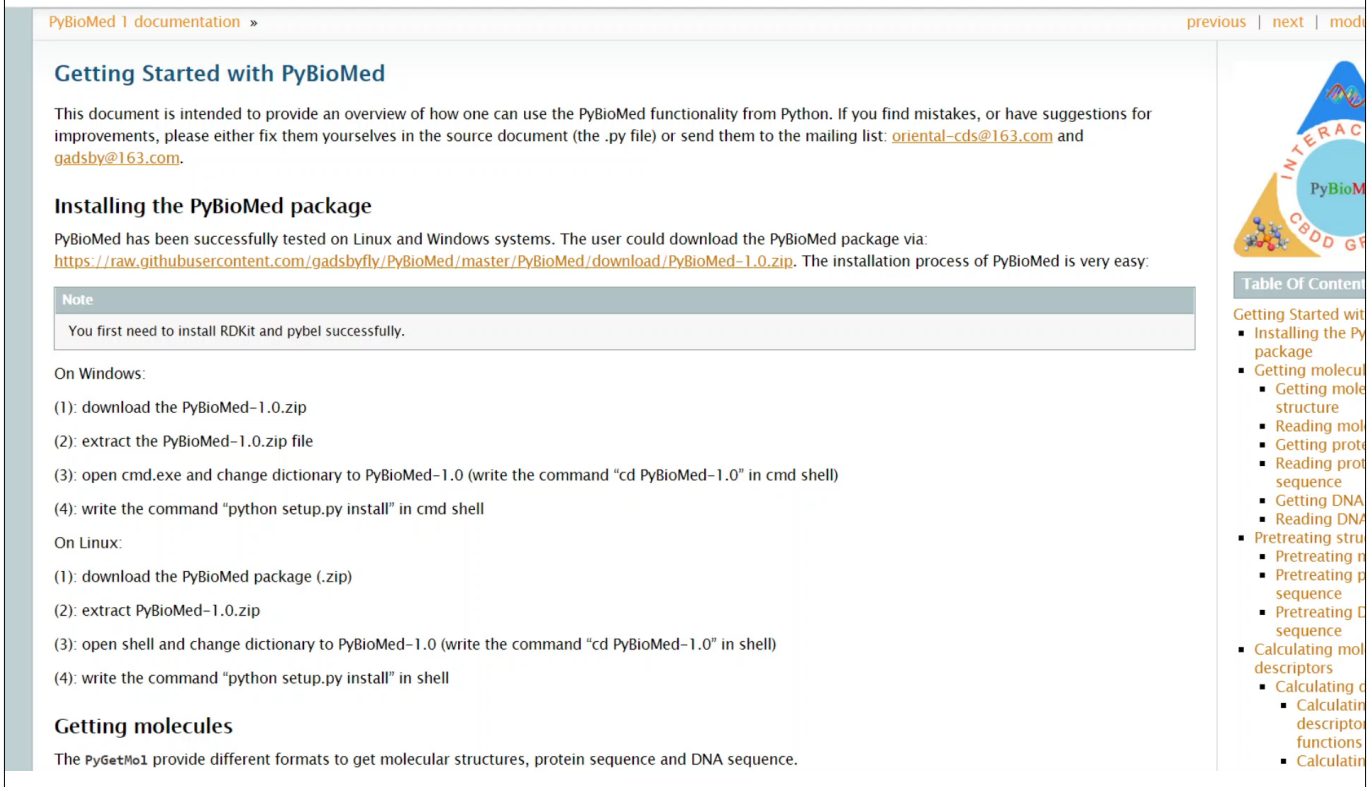


BMC Research Notes volume 11, Article number: 117 (2018)

Dipeptide (400D) / Tripeptide (8000D) Composition

```
##          AA          RA          NA          DA          CA          EA
## 0.003565062 0.003565062 0.000000000 0.007130125 0.003565062 0.003565062
##          QA          GA          HA          IA          LA          KA
## 0.007130125 0.007130125 0.001782531 0.003565062 0.001782531 0.001782531
##          MA          FA          PA          SA          TA          WA
## 0.000000000 0.005347594 0.003565062 0.007130125 0.003565062 0.000000000
##          YA          VA          AR          RR          NR          DR
## 0.000000000 0.000000000 0.003565062 0.007130125 0.005347594 0.001782531
##          CR          ER          QR          GR          HR          IR
## 0.005347594 0.005347594 0.000000000 0.007130125 0.001782531 0.003565062
```

```
##         AAA         KAA         NAA         DAA         CAA         EAA
## 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000
##         QAA         GAA         HAA         IAA         LAA         KAA
## 0.001785714 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000
##         MAA         FAA         PAA         SAA         TAA         WAA
## 0.000000000 0.000000000 0.000000000 0.001785714 0.000000000 0.000000000
##         YAA         VAA         ARA         RRA         NRA         DRA
## 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000
##         CRA         ERA         QRA         GRA         HRA         IRA
## 0.000000000 0.000000000 0.000000000 0.001785714 0.000000000 0.000000000
##         LRA         KRA         MRA         FRA         PRA         SRA
## 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000
```



PyBioMed 1 documentation »

previous | next | modu

Getting Started with PyBioMed

This document is intended to provide an overview of how one can use the PyBioMed functionality from Python. If you find mistakes, or have suggestions for improvements, please either fix them yourselves in the source document (the .py file) or send them to the mailing list: oriental-cds@163.com and gadsby@163.com.

Installing the PyBioMed package

PyBioMed has been successfully tested on Linux and Windows systems. The user could download the PyBioMed package via: https://raw.githubusercontent.com/gadsbyfly/PyBioMed/master/PyBioMed/download/PyBioMed-1.0.zip. The installation process of PyBioMed is very easy:

Note

You first need to install RDKit and pybel successfully.

On Windows:

(1): download the PyBioMed-1.0.zip

(2): extract the PyBioMed-1.0.zip file

(3): open cmd.exe and change dictionary to PyBioMed-1.0 (write the command "cd PyBioMed-1.0" in cmd shell)

(4): write the command "python setup.py install" in cmd shell

On Linux:

(1): download the PyBioMed package (.zip)

(2): extract PyBioMed-1.0.zip

(3): open shell and change dictionary to PyBioMed-1.0 (write the command "cd PyBioMed-1.0" in shell)

(4): write the command "python setup.py install" in shell

Getting molecules

The `PyGetMol` provide different formats to get molecular structures, protein sequence and DNA sequence.

Table Of Content

Getting Started wit
- Installing the Py package
- Getting molecul
  - Getting mole structure
  - Reading mol
  - Getting prote
  - Reading prot sequence
  - Getting DNA
  - Reading DNA
- Pretreating stru
  - Pretreating m
  - Pretreating p sequence
  - Pretreating D sequence
- Calculating mol descriptors
  - Calculating d
    - Calculatin descriptor functions
    - Calculatin

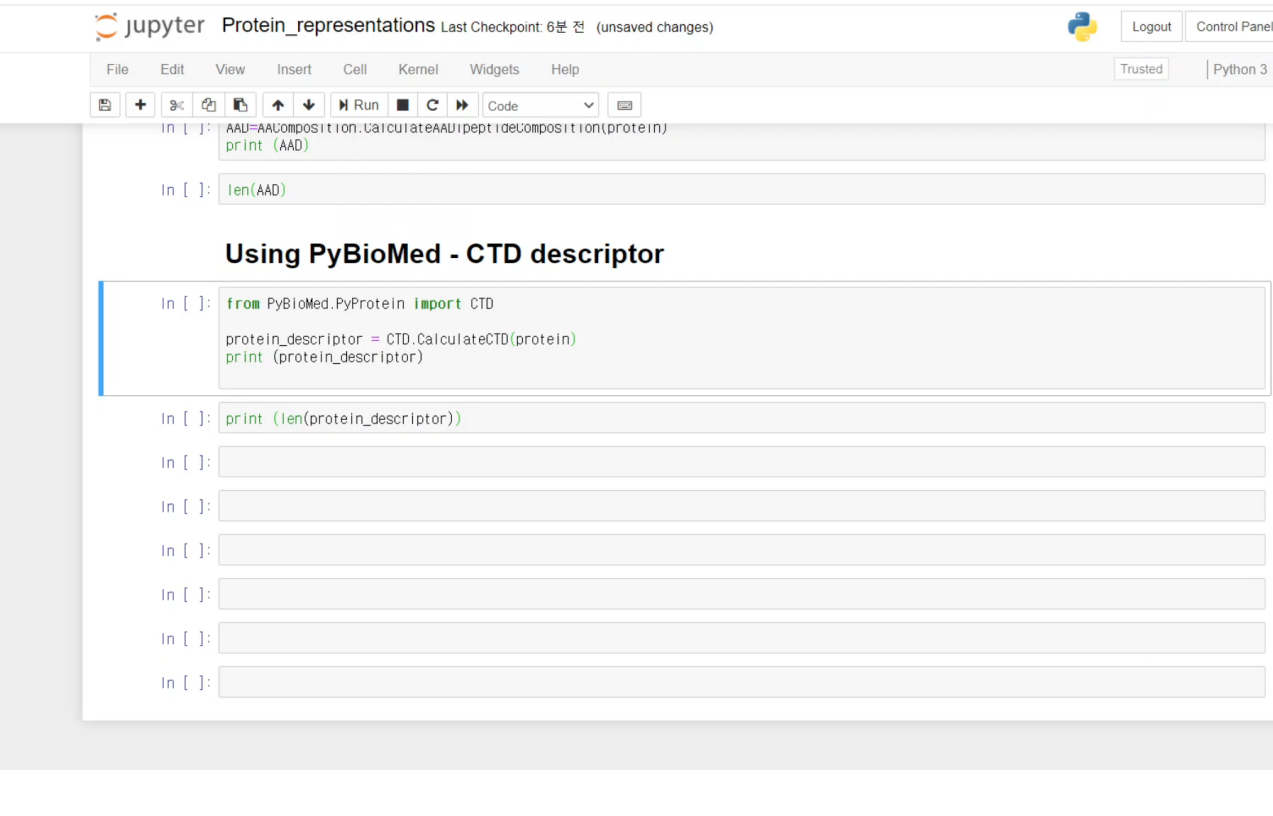
Composition, Transition and Distribution (CTD), 147D

Sequence	M	T	E	I	T	A	S	M	V	K	E	L	R	E	A	T	G	T	G	A
Sequence Index	1				5					10					15					20
Transformation	3	2	1	3	2	2	2	3	3	1	1	3	1	1	2	2	2	2	2	2
Index for 1			1							2	3		4	5						
Index for 2		1			2	3	4								5	6	7	8	9	10
Index for 3	1			2				3	4			5								
1/2 Transitions			|											|					
1/3 Transitions				|						|		|							
2/3 Transitions		|			|			|											

Table 1: Amino acid attributes, and the three-group classification of the 20 amino acids by each attribute

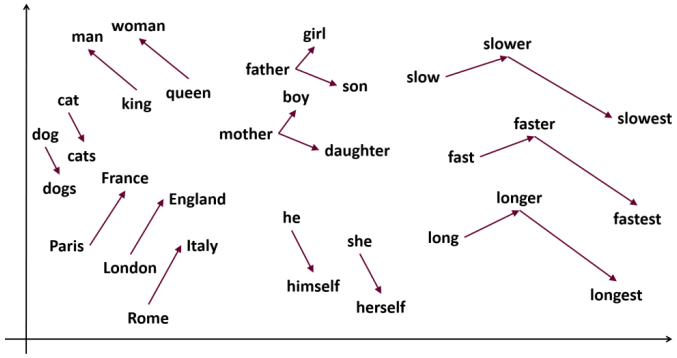
	Group 1	Group 2	Group 3
Hydrophobicity	Polar	Neutral	Hydrophobicity
	R, K, E, D, Q, N	G, A, S, T, P, H, Y	C, L, V, I, M, F, W
Normalized van der Waals Volume	0-2.78	2.95-4.0	4.03-8.08
	G, A, S, T, P, D, C	N, V, E, Q, I, L	M, H, K, F, R, Y, W
Polarity	4.9-6.2	8.0-9.2	10.4-13.0
	L, I, F, W, C, M, V, Y	P, A, T, G, S	H, Q, R, K, N, E, D
Polarizability	0-1.08	0.128-0.186	0.219-0.409
	G, A, S, D, T	C, P, N, V, E, Q, I, L	K, M, H, F, R, Y, W
Charge	Positive	Neutral	Negative
	K, R	A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V	D, E
Secondary Structure	Helix	Strand	Coil
	E, A, L, M, Q, K, R, H	V, I, Y, C, W, F, T	G, N, P, S, D
Solvent Accessibility	Buried	Exposed	Intermediate
	A, L, F, C, G, I, V, W	R, K, Q, E, N, D	M, S, P, T, H, Y

https://mran.microsoft.com/snapshot/2017-12-06/web/packages/protr/vignettes/protr.html



ProtVec (Asgari et al. ,PLoS ONE 10(11): e0141287, 2015)

- Continuous distributed representation of biological sequences for deep proteomics and genomics
  - ProtVec: “unsupervised data-driven distributed representation for biological sequences”
  - Each sequence represented as n-dimensional vector
    - Characterizes biophysical and biochemical properties
    - Determined using neural networks



Apply to proteins as well? → ProtVec

ProtVec

- Use large corpus of sequences to train representation
  - E.g.) Swiss-Prot with 546,790 manually annotated and reviewed sequences
  - Break sequences into subsequences (i.e. biological words)
  - Training of the embedding through the Skip-gram neural network
    - for protein sequences: usage of a vector size of 100 and a context size of 25
    - → every 3-gram is represented as a vector of size 100

Original Sequence

$$^{(1)}\overrightarrow{M}\,^{(2)}\overrightarrow{A}\,^{(3)}\overrightarrow{F}SAEDVLKEYDRRRRMEAL..$$

Splittings

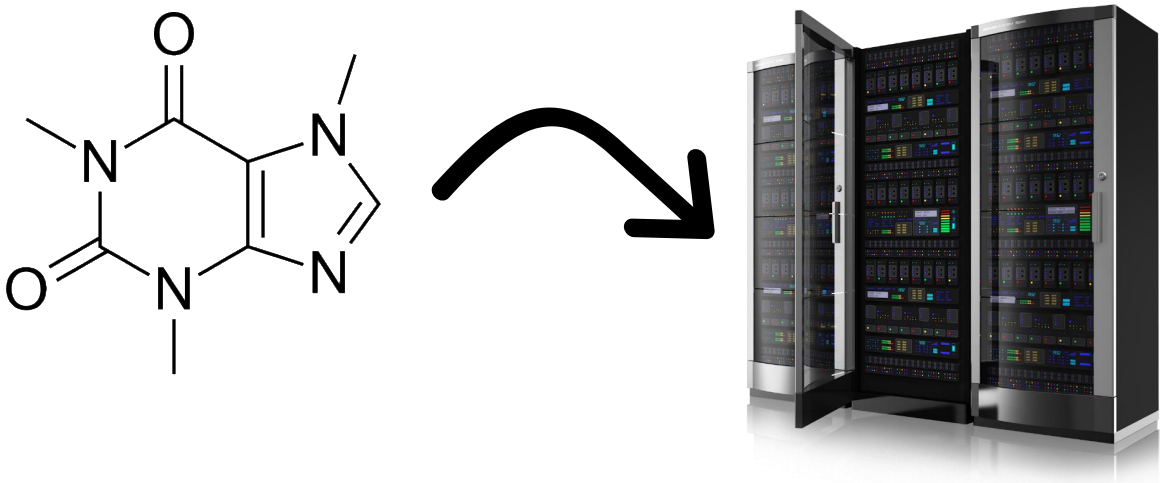
$$\begin{cases} 1) & \text{MAF, SAE, DVL, KEY, DRR, RRM, ..} \\ 2) & \text{AFS, AED, VLK, EYD, RRR, RME, ..} \\ 3) & \text{FSA ,EDV, LKE, YDR, RRR, MEA, ..} \end{cases}$$

Asgari et al. ,PLoS ONE 10(11): e0141287, 2015

- PART1
  - Introduction to pharmacogenomics
    - Drug discovery and development
  - Key data sources
  - Representations of proteins, chemicals
- PART2
  - Studies related to pharmacogenomics based on machine learning

MOLECULAR REPRESENTATION

Why molecular representations are necessary?



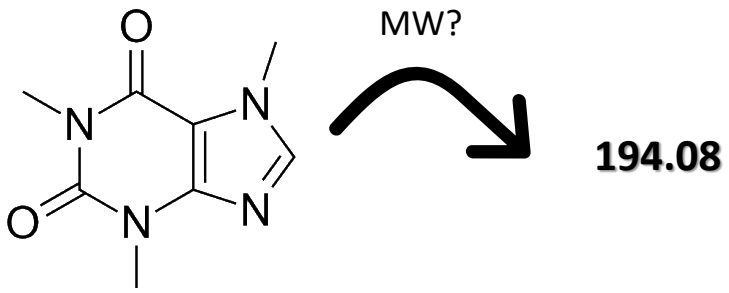
Representation of chemical compounds for machine-learning features that fully captured wide ranges of chemical and physical properties of the target molecule

Types of molecular representations

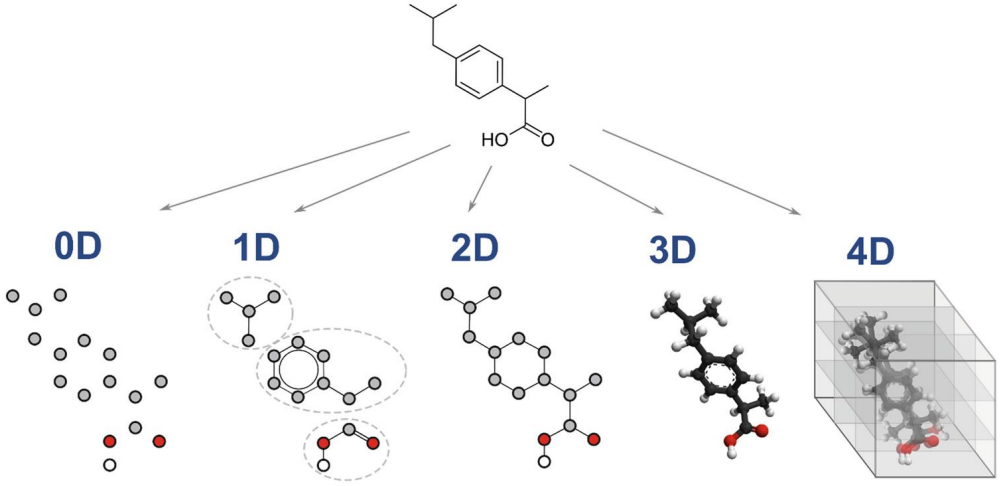
- Molecular descriptors
- Molecular fingerprints

Molecular descriptors

- Molecular descriptors are numerical values that characterize properties of molecules
- The goal of a molecular descript is to provide a numerical representation of molecular structure
- There are numbers of molecular descripts vary in complexity of encoded information



Molecular descriptors

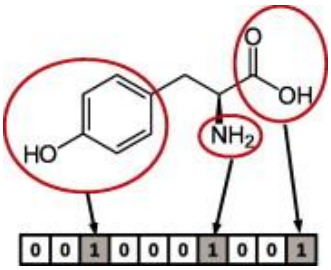


1) **0D-descriptors** (Molecular formula, i.e. Molecular weights, atom counts, bond counts),
2) **1D-descriptors** (Chemical graph, i.e. Fragment counts, functional group counts),
3) **2D-descriptors** (Structural topology, i.e. Wiener index, Balaban index, Randic index, BCUTS),
4) **3D-descriptors** (Structural geometry, i.e. WHIM, autocorrelation, 3D-MORSE, GETAWAY),
5) **4D-descriptors** (Chemical conformation, i.e. Volsurf, GRID, Raptor)

Grisoni F., Ballabio D., Todeschini R., Consonni V. (2018) Molecular Descriptors for Structure–Activity Applications: A Hands-On Approach. In: Computational Toxicology. Methods in Molecular Biology, vol 1800.

Molecular fingerprints

- Fingerprint representations of molecular structure and properties are a particularly complex form of descriptors. Fingerprints are typically encoded as binary bit strings whose settings produce, in different ways, a bit “pattern” characteristic of a given molecule.
- Fingerprints are designed to account for different sets of molecular descriptors, structural fragments, possible connectivity pathways through a molecule, or different types of pharmacophores.



https://doi.org/10.1016/j.ymeth.2014.08.005

Types of fingerprints

Class	Type	Examples
Structural based	Pattern-based FP	MACCS, PubChem, FP3, FP4
Topological	Path-based FP	Daylight, FP2
	Circular FP	ECFP2, ECFP4, ECFP6
	Pharmacophore FP	2D pharmacophore
Neural network based	Graph-based representation	GNN (graph convolutional network (GCN), graph attention network (GAT), gated graph neural network (GGNN), …)
	Molecular embedding	seq2seq, mol2vec

Pattern based fingerprints

SMARTS pattern

- 특정 SMARTS pattern 구조를 기반으로 한 지문표현자 생성 방법

Key position	Key description	Annotation
11	*1~*~*~*~1	4M Ring
12	[Cu,Zn,Ag,Cd,Au,Hg]	Group IB, IIB
13	[#8]~[#7](~[#6])~[#6]	ON(C)C
14	[#16] - [#16]	S-S
⋮	⋮	⋮

MACCS fingerprint SMARTS pattern 기준표

✓ MACCS fingerprints (166 keys)
✓ FP3, FP4 fingerprints from OpenBabel

PubChem Fingerprint

- PubChem에서 제시한 하위 구조를 기반으로 한 지문표현자 (881 bit vector)

Sections	Description
Section 1 (#0~#114)	Hierarchic element counts
Section 2 (#115~#262)	Rings in a canonic Extended Smallest Set of Smallest Rings ring set
Section 3 (#263~#326)	Simple atom pairs
Section 4 (#327~#415)	Simple atom nearest neighbors
Section 5 (#416~#459)	Detailed atom neighborhoods
Section 4 (#460~#712)	Simple SMARTS patterns
Section 4 (#713~#880)	Complex SMARTS patterns

PubChem fingerprints bit별 description

- **특징점**

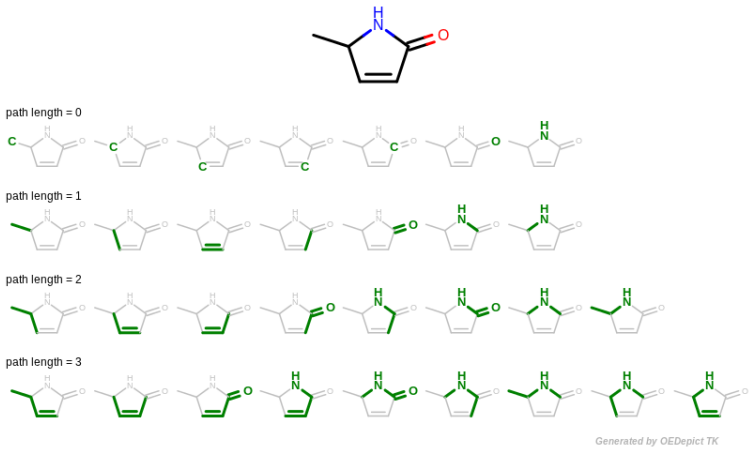
- 이미 정의된 하위 구조의 유무를 판단하여 생성되는 지문표현자로 하위 구조 검색에 유용하나 이외의 구조를 표현할 수 없음
- 상대적으로 벡터의 길이가 짧음

Path-based fingerprints

- 원자를 기준으로 모든 linear fragment 를 고려하는 방식으로 화합물 구조 그래프를 표현함
- 해싱(hashing) 알고리즘을 사용함

- 관련 Fingerprints
  - ✓ FP2 fingerprints (1,021 bit vector)
  - ✓ RDK fingerprints, Layered fingerprints (RDKit), CDK fingerprints (CDK)

- **특징점**
- 해싱 알고리즘을 사용하여 다양한 하위 구조를 표현할 수 있고 사용자가 길이 조절할 수 있음
- 하위 구조의 사전지식이 필요 없음
- 지문표현자의 resolution은 해싱 알고리즘에 따라 달라질 수 있음
- Bit collision과 bit space 낭비를 고려한 길이의 지문표현자를 찾는 것이 어려움



길이에 따른 fragment 추출 예시

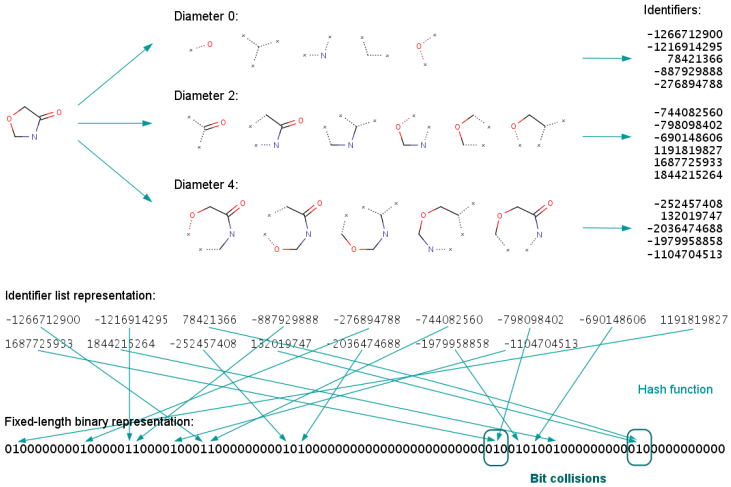
https://docs.eyesopen.com/toolkits/python/graphsimtk/fingerprint.html#section-fingerprint-path

Morgan/Circular fingerprints

- 하나의 원자를 기준으로 주어진 반경 내의 하위 구조 정보를 순차적으로 탐색하는 기법
- 해싱(hashing) 기법을 사용하여 특정 길이 내의 지문표현자로 반환하여 사용함

- 관련 Fingerprints
  - ✓ Morgan/Circular fingerprints
  - ✓ ECFPs (ECFP4, ECFP6), FCFPs

- **특징점**
- 이미 정의된 구조가 아닌 하위 구조에 대한 표현이 가능함
- 계산 속도가 빠름
- 전체적인 구조 정보를 표현하는데 유용하나 하위 구조 검색에는 적합하지 않음
- 유사성 검색에 적합함



ECFP fingerprint의 산출 절차

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

jupyter Generate_FPs Last Checkpoint: 4분 전 (autosaved)

Logout Control Panel

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

1. Using RDkit

Descriptors, MACCSkey, Morgan

In [ ]:
```
from __future__ import absolute_import
import rdkit
from rdkit import Chem
from rdkit.Chem import rdMolDescriptors # Module containing functions to compute molecular descriptors
from rdkit.Chem import Descriptors
import rdkit.rdBase
from rdkit.Chem.MACCSkeys import GenMACCSKeys
from rdkit.Chem import AllChem
from rdkit.Chem import Draw
```

In [ ]:
```
# Reading single molecules
m = Chem.MolFromSmiles("CN1C=NC2=C1C(=O)N(C(=O)N2C)C") # caffine

from rdkit.Chem.Draw import IPythonConsole #Needed to show molecules
from rdkit.Chem.Draw.MolDrawing import MolDrawing, DrawingOptions #Only needed if modifying defaults
%matplotlib inline
m
```

In [ ]:
```
rdMolDescriptors.CalcExactMolWt(m) # returns the molecule's exact molecular weight
```

In [ ]:
```
Descriptors.MolLogP(m)
```

In [ ]:
```
rdMolDescriptors.CalcMolFormula(m) # returns the molecule's formula
```

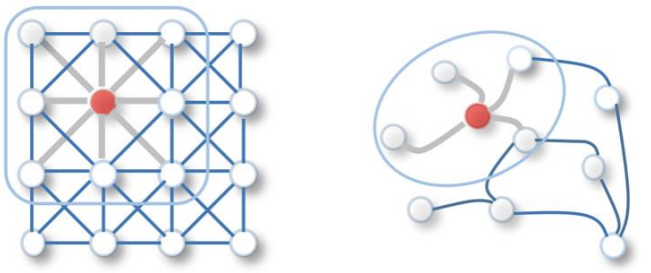
In [ ]:
```
rdMolDescriptors.CalcNumHBA(m) # returns the number of H-bond acceptors for a molecule
```

GNN

- Graph neural networks (GNNs) are connectionist models that capture the dependence of graphs via message passing between the nodes of graphs.
  - Extract features by considering the structure of the data
  - Enables automatic feature extraction from raw inputs

  → can embed the drug(molecule) into vectors which has **topological structure information** with edge and atom features

  - With end to end learning, the model can learn **data driven features**



(a) 2D Convolution. Analogous to a graph, each pixel in an image is taken as a node where neighbors are determined by the filter size. The 2D convolution takes the weighted average of pixel values of the red node along with its neighbors. The neighbors of a node are ordered and have a fixed size.

(b) Graph Convolution. To get a hidden representation of the red node, one simple solution of the graph convolutional operation is to take the average value of the node features of the red node along with its neighbors. Different from image data, the neighbors of a node are unordered and variable in size.

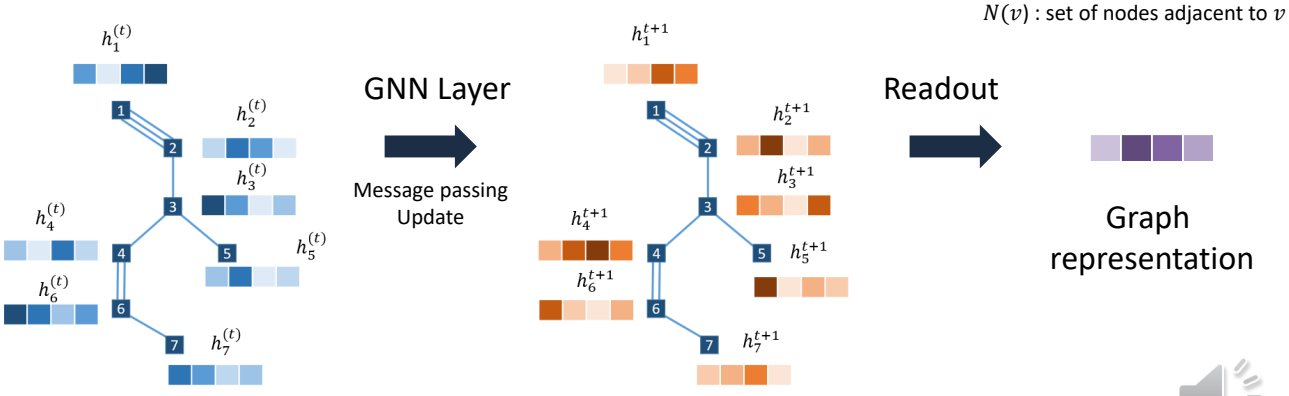
Fig. 1: 2D Convolution vs. Graph Convolution.

https://arxiv.org/abs/1901.00596

Graph Neural Network

- **Message Passing** : aggregate information from neighbors
  - $m_v^{(t+1)} = message\_passing(\{h_w^{(t)}, \forall w \in N(v)\})$
- **Update** : with message passing, update the hidden representation
  - $h_v^{t+1} = update(m_v^{(t+1)}, h_v^{(t)})$
- **Readout** : represent graph with all hidden representations
  - $h_G^{t+1} = readout(h_v^{t+1}, \forall v \in G)$

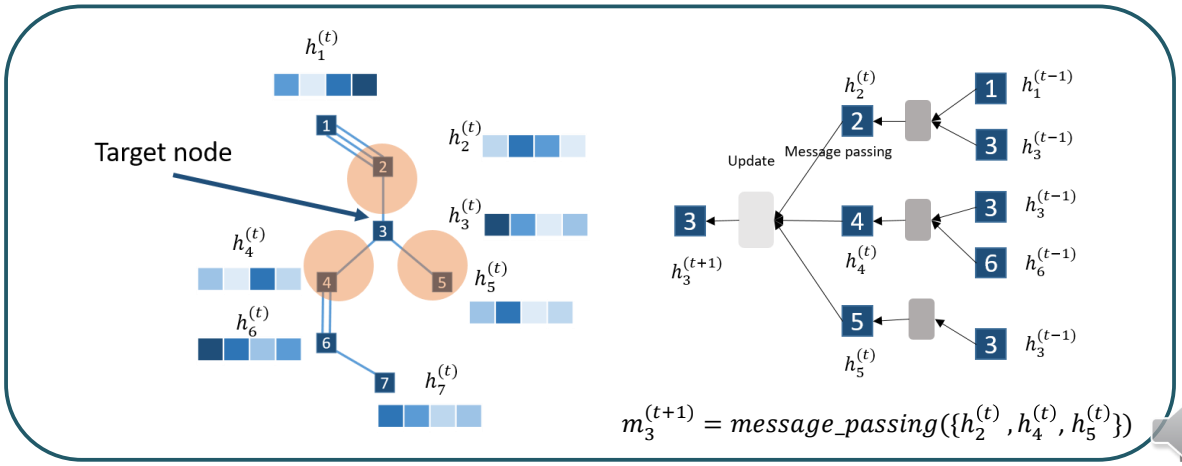
$h_v^0 = X_v$
$N(v)$: set of nodes adjacent to v



h_v^t : hidden embedding vector of node v at t-th GNN layer

Graph Neural Network

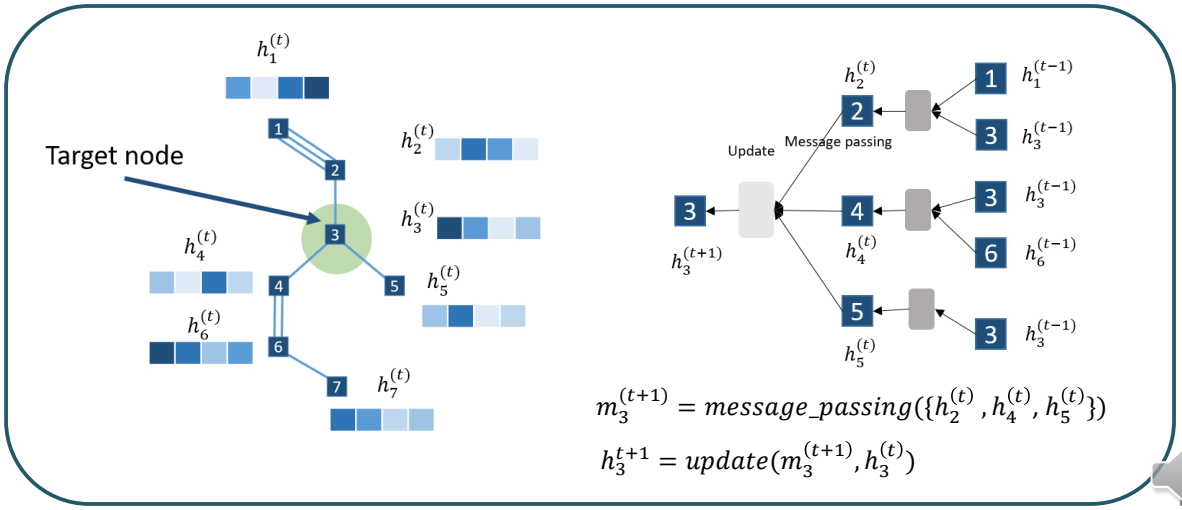
- **Message passing**
  - Message : Information that flows between neighbors and the target node
  - $message\_passing$: function that aggregate neighbor information of target node at t time step with propagation rule
  - $m_v^{(t+1)} = message\_passing(\{h_w^{(t)}, \forall w \in N(v)\})$



$m_3^{(t+1)} = message\_passing(\{h_2^{(t)}, h_4^{(t)}, h_5^{(t)}\})$

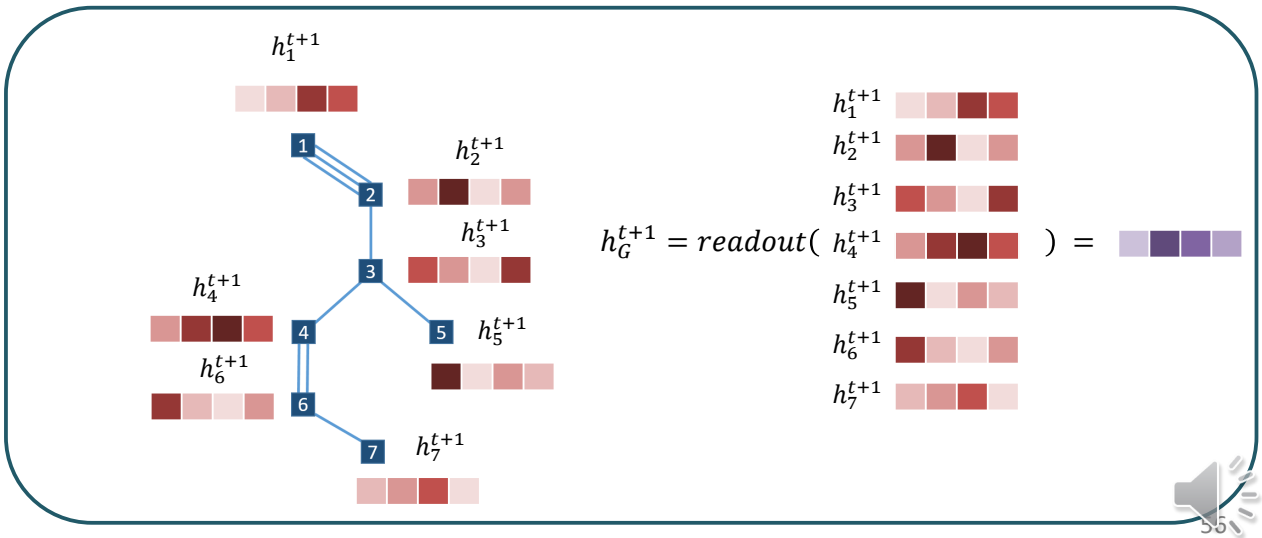
Graph Neural Network

- **Update**
  - $update$: function that update the t+1 time step hidden representation with t time step node representation and message passing
  - $h_v^{t+1} = update(m_v^{(t+1)}, h_v^{(t)})$



Graph Neural Network

- **Readout**
  - $readout$: function that represent the graph calculated by all hidden representations
  - $h_G^{t+1} = readout(h_v^{t+1}, \forall v \in G)$



Graph Neural Network Models

- Semi –Supervised Classification with Graph Convolutional Networks (**GCN**)
- Inductive Representation Learning on Large Graphs (**GraphSAGE**)
- Neural Message Passing for Quantum Chemistry (**MPNN**)
- Graph Attention Networks (**GAT**)
- How Powerful Are Graph Neural Network? (**GIN**)
- Analyzing Learned Molecular Representations for Property Prediction (**DMPNN**)

→ Various Message passing, Update, Readout function

To be continued.

1. P
2. ~~D~~R
3.

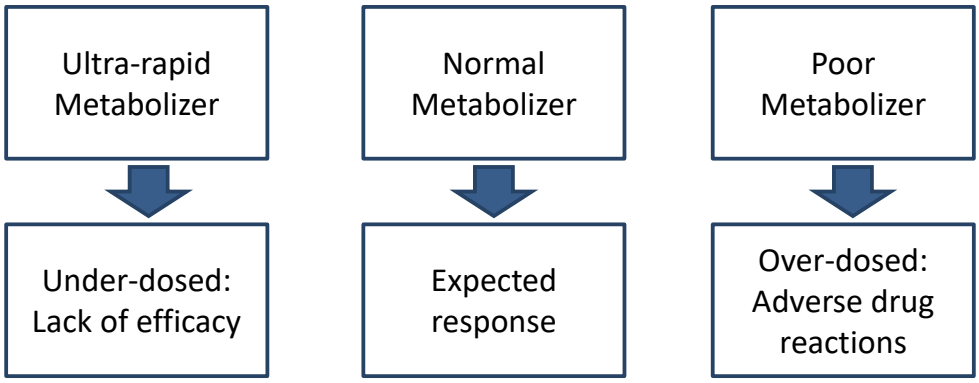
Contents

- PART1
  - Introduction to pharmacogenomics
    - Drug discovery and development
  - Key data sources
  - Representations of proteins, chemicals
- PART2
  - Studies related to pharmacogenomics based on machine learning

CYP450 VARIATIONS AND DRUG RESPONSES

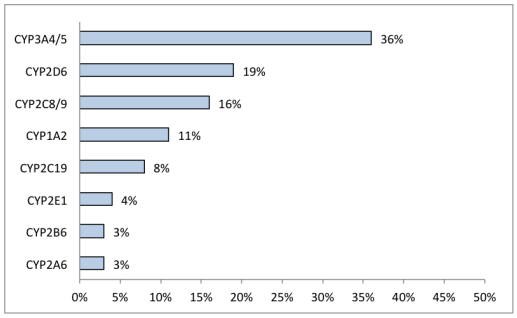
Pharmacogenomics and drug metabolism

- A patient's genetic makeup and their response to pharmaceutical drugs are seen with regards to their metabolism



Cytochrome P450 enzymes

- The super-family of cytochrome P450 enzymes has a crucial role in the metabolism of drugs
- CYPs are the major enzymes involved in drug metabolism, accounting for about 75% of the total metabolism
- Most drugs undergo deactivation by CYPs, either directly or by facilitated excretion from the body



e.g.) Proportion of antifungal drugs metabolized by different families of CYPs.

https://en.wikipedia.org/wiki/Cytochrome_P450#Drug_metabolism

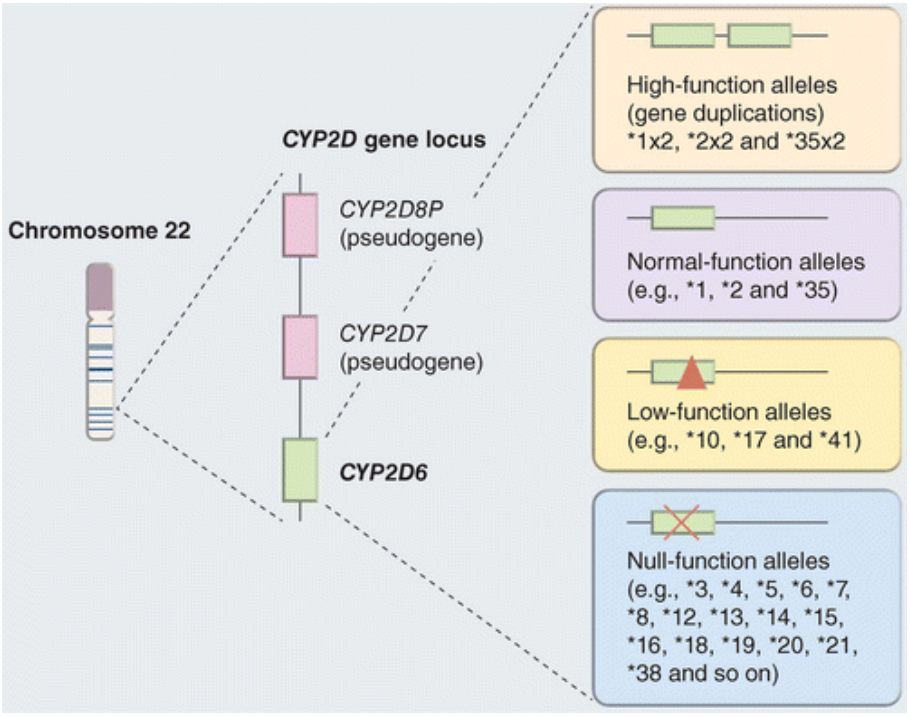
CYP450 isozymes

- Humans have 57 genes and more than 59 pseudogenes divided among 18 families of cytochrome P450 genes and 43 subfamilies

Family	Function	Members	Genes	pseudogenes
CYP1	drug and steroid (especially estrogen) metabolism, benzo[*a*]pyrene toxification (forming (+)-benzo[*a*]pyrene-7,8-dihydrodiol-9,10-epoxide)	3 subfamilies, 3 genes, 1 pseudogene	CYP1A1, CYP1A2, CYP1B1	CYP1D1P
CYP2	drug and steroid metabolism	13 subfamilies, 16 genes, 16 pseudogenes	CYP2A6, CYP2A7, CYP2A13, CYP2B6, CYP2C8, CYP2C9, CYP2C18, CYP2C19, CYP2D6, CYP2E1, CYP2F1, CYP2J2, CYP2R1, CYP2S1, CYP2U1, CYP2W1	Too many to list
CYP3	drug and steroid (including testosterone) metabolism	1 subfamily, 4 genes, 4 pseudogenes	CYP3A4, CYP3A5, CYP3A7, CYP3A43	CYP3A51P, CYP3A52P, CYP3A54P, CYP3A137P
CYP4	arachidonic acid or fatty acid metabolism	6 subfamilies, 12 genes, 10 pseudogenes	CYP4A11, CYP4A22, CYP4B1, CYP4F2, CYP4F3, CYP4F8, CYP4F11, CYP4F12, CYP4F22, CYP4V2, CYP4X1, CYP4Z1	Too many to list
CYP5	thromboxane A_2 synthase	1 subfamily, 1 gene	CYP5A1	
CYP7	bile acid biosynthesis 7-alpha hydroxylase of steroid nucleus	2 subfamilies, 2 genes	CYP7A1, CYP7B1	
CYP8	*varied*	2 subfamilies, 2 genes	CYP8A1 (prostacyclin synthase), CYP8B1 (bile acid biosynthesis)	
CYP11	steroid biosynthesis	2 subfamilies, 3 genes	CYP11A1, CYP11B1, CYP11B2	
CYP17	steroid biosynthesis, 17-alpha hydroxylase	1 subfamily, 1 gene	CYP17A1	
CYP19	steroid biosynthesis: aromatase synthesizes estrogen	1 subfamily, 1 gene	CYP19A1	
CYP20	unknown function	1 subfamily, 1 gene	CYP20A1	
CYP21	steroid biosynthesis	1 subfamilies, 1 gene, 1 pseudogene	CYP21A2	CYP21A1P
CYP24	vitamin D degradation	1 subfamily, 1 gene	CYP24A1	
CYP26	retinoic acid hydroxylase	3 subfamilies, 3 genes	CYP26A1, CYP26B1, CYP26C1	
CYP27	*varied*	3 subfamilies, 3 genes	CYP27A1 (bile acid biosynthesis), CYP27B1 (vitamin D_3 1-alpha hydroxylase, activates vitamin D_3), CYP27C1 (unknown function)	
CYP39	7-alpha hydroxylation of 24-hydroxycholesterol	1 subfamily, 1 gene	CYP39A1	
CYP46	cholesterol 24-hydroxylase	1 subfamily, 1 gene, 1 pseudogene	CYP46A1	CYP46A4P
CYP51	cholesterol biosynthesis	1 subfamily, 1 gene, 3 pseudogenes	CYP51A1 (lanosterol 14-alpha demethylase)	CYP51P1, CYP51P2, CYP51P3

https://en.wikipedia.org/wiki/Cytochrome_P450#Drug_metabolism

CYP2D6 alleles



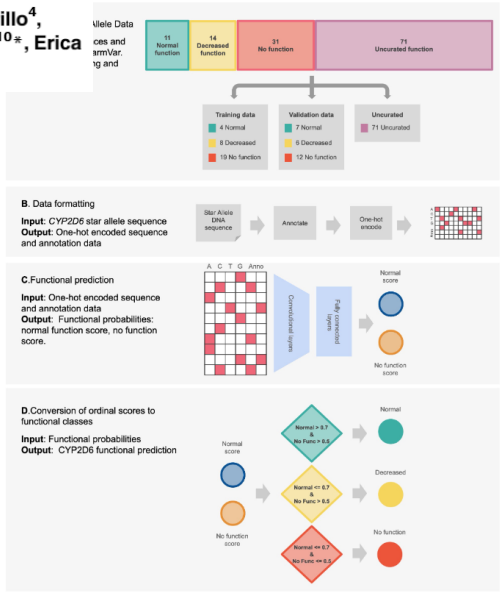
https://www.futuremedicine.com/doi/10.2217/fmeb2013.13.130

Related study: prediction of CYP2D6 haplotype function

RESEARCH ARTICLE

Transfer learning enables prediction of *CYP2D6* haplotype function

Gregory McInnes[1], Rachel Dalton[2,3], Katrin Sangkuhl[4], Michelle Whirl-Carrillo[4], Seung-been Lee[5], Philip S. Tsao[6,7], Andrea Gaedigk[8,9], Russ B. Altman[4,10]*, Erica L. Woodahl[2]*



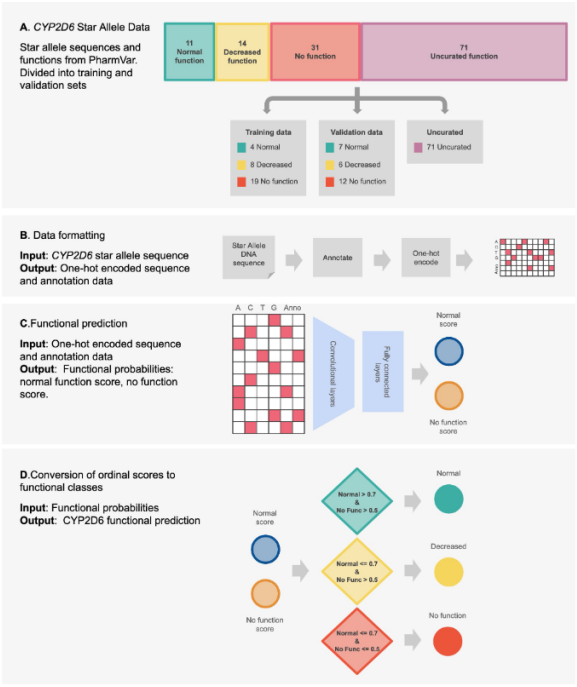
McInnes G, Dalton R, Sangkuhl K, WhirlCarrillo M, Lee S-b, Tsao PS, et al. (2020) Transfer learning enables prediction of CYP2D6 haplotype function. PLoS Comput Biol 16(11): e1008399. https://doi.org/10.1371/journal.pcbi.1008399

Related study: prediction of CYP2D6 haplotype function

- CYP2D6 is an enzyme expressed in the liver that is responsible for metabolizing more than 20% of clinically used drugs
- More than 130 haplotypes comprised of single nucleotide variants (SNVs), insertions and deletions (INDELs), and structural variants (SVs) have been discovered and catalogued in the Pharmacogene Variation Consortium

Related study: prediction of CYP2D6 haplotype function

- **Input**
  - CYP2D6 Full genomic sequence (one hot vector)
  - 9 annotations (one hot vector)
    - Coding region, rare variants, deleterious, INDEL, methylation mark, DNase hypersensitivity, TF binding site, eQTL, active site
- **Output**
  - Haplotype activity (No, Reduced, Normal activity)
- **Data**
  - Pre-training with 50,000 randomly selecting a pair of CYP2D6 star alleles with curated function, Pre-training with 314 in vivo data
  - Fine-tuning with PharmVar data
- **Model** – 3 CNN + 2 FC



McInnes G, Dalton R, Sangkuhl K, WhirlCarrillo M, Lee S-b, Tsao PS, et al. (2020) Transfer learning enables prediction of CYP2D6 haplotype function. PLoS Comput Biol 16(11): e1008399. https://doi.org/10.1371/journal.pcbi.1008399

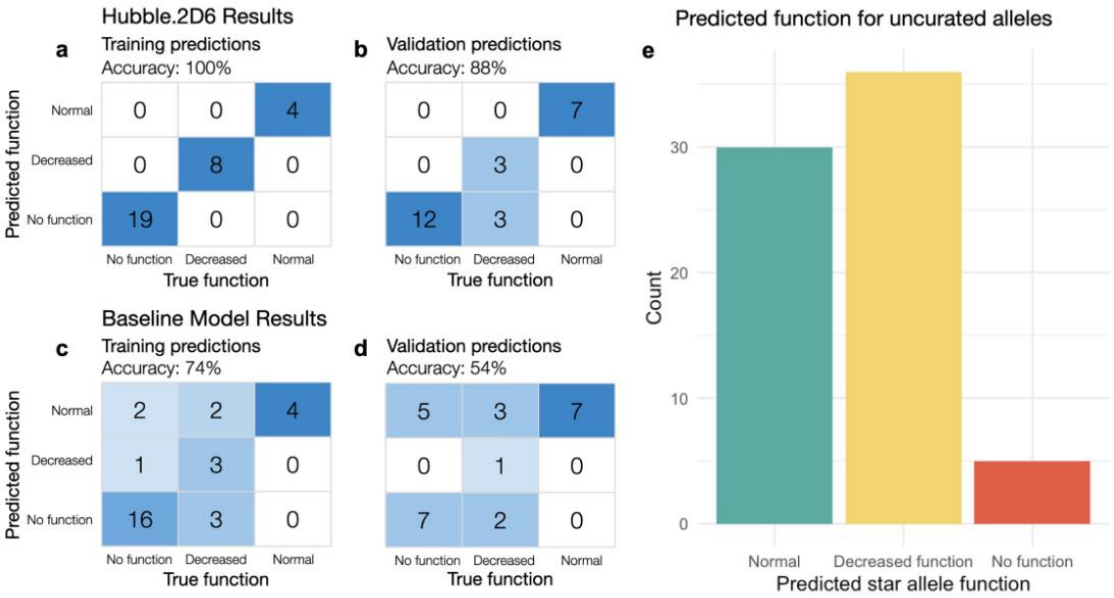


Fig 2. Star allele classification results. The figure depicts performance metrics for the prediction of star allele function in the training and validation sets; confusion matrices for class prediction in training and validation are shown in (a) and (b), for Hubble.2D6 and in (c) and (d) for the baseline model. (e) shows the frequency of predicted function for uncurated star alleles.

McInnes G, Dalton R, Sangkuhl K, WhirlCarrillo M, Lee S-b, Tsao PS, et al. (2020) Transfer learning enables prediction of CYP2D6 haplotype function. PLoS Comput Biol 16(11): e1008399. https://doi.org/10.1371/journal.pcbi.1008399

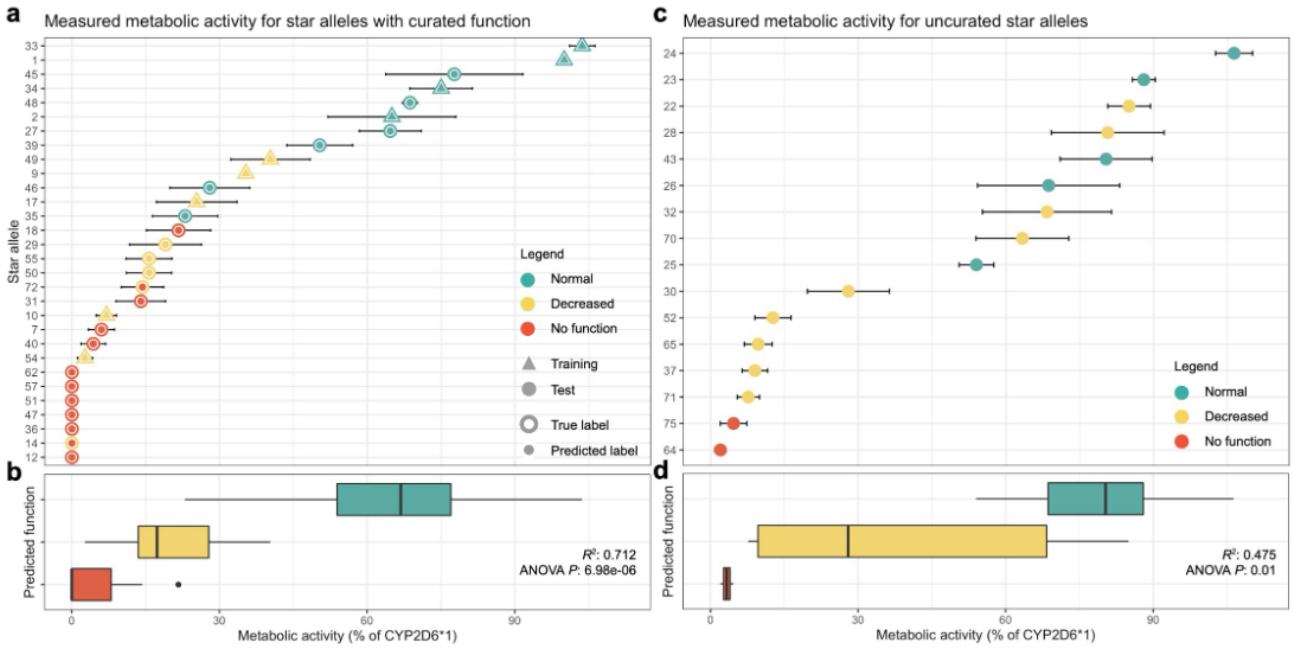


Fig 3. Prediction of star allele function with *in vitro* data. The figures summarize the distribution of metabolic activity measured *in vitro* for star alleles whose function was predicted by Hubble. The distribution of functional activity is shown in (a) and (b) for star alleles with CPIC-assigned clinical function assignments. (a) star alleles included in the training process are depicted with a triangle, and those held for testing are depicted with a circle. Error bars depict the standard error of the measured function. The outer edge of each point indicates the true, curator-assigned phenotype, while the inner color represents predicted function. (b) distribution of values for each predicted functional class for data shown in (a). (c) star alleles without assigned function status; colors represent the predicted function. (d) variance in measured activity of the star alleles for each predicted label for data shown in (c).

McInnes G, Dalton R, Sangkuhl K, WhirlCarrillo M, Lee S-b, Tsao PS, et al. (2020) Transfer learning enables prediction of CYP2D6 haplotype function. PLoS Comput Biol 16(11): e1008399. https://doi.org/10.1371/journal.pcbi.1008399

GENETIC VARIATIONS AND DRUG RESPONSES

Related study: prediction of cancer cell sensitivity to drugs

- Genomic features
  - MSI, variations, CNV
- Simple neural network

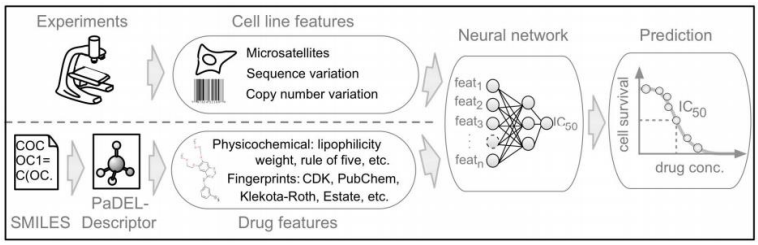
OPEN ACCESS Freely available online

PLOS ONE

Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties

Michael P. Menden[1], Francesco Iorio[1,2], Mathew Garnett[2], Ultan McDermott[2], Cyril H. Benes[3], Pedro J. Ballester[1]*, Julio Saez-Rodriguez[1]*

1 European Bioinformatics Institute, Wellcome Trust Genome Campus–Cambridge, Cambridge, United Kingdom, 2 Cancer Genome Project, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus–Cambridge, Cambridge, United Kingdom, 3 Center for Molecular Therapeutics, Massachusetts General Hospital Cancer Center and Harvard Medical School, Charlestown, Massachusetts, United States of America



Menden, Michael P., et al. "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties." PLoS one 8.4 (2013): e61318.

Related study: prediction of cancer cell sensitivity to drugs

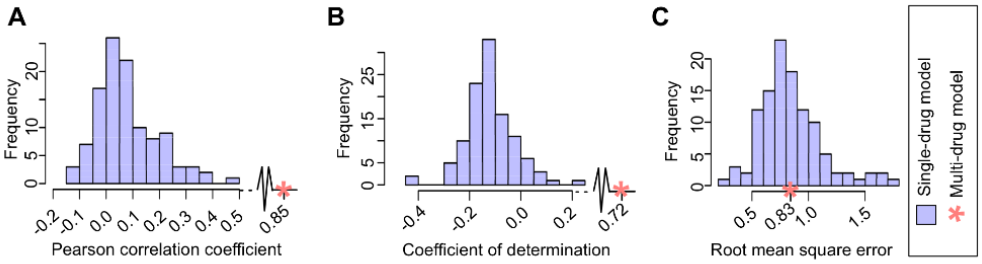
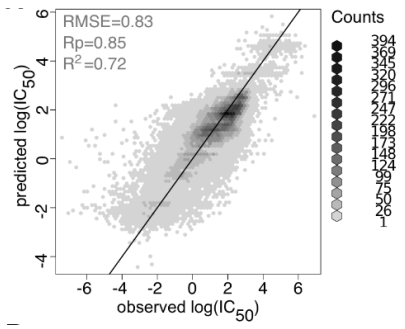


Figure 2. Comparison of single-drug models and the multi-drug model. The performance of the multi-drug model (red asterisk) and the family of 111 single-drug models (blue histogram) is represented using three different metrics: (A) Pearson correlation R_p, (B) coefficient of determination R^2, and (C) root mean square error RMSE.
doi:10.1371/journal.pone.0061318.g002



- Genomics of Drug Sensitivity in Cancer (GDSC) project
- mutational status of 77 oncogenes
- 639 cancer cell lines
- 131 drugs
- 67,488 possible drug response
- 8-fold cross-validation

Menden, Michael P., et al. "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties." PLoS one 8.4 (2013): e61318.

Related study: prediction of cancer cell sensitivity to drugs

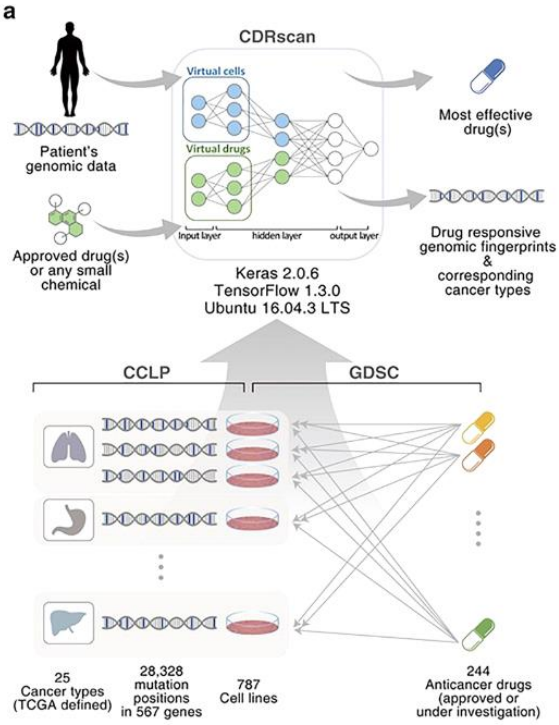
SCIENTIFIC REPORTS

OPEN

Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature

Received: 10 January 2018
Accepted: 29 May 2018
Published online: 11 June 2018

Yoosup Chang[1], Hyejin Park[1], Hyun-Jin Yang[2], Seungju Lee[1], Kwee-Yum Lee[2,3], Tae Soon Kim[2,4], Jongsun Jung[1] & Jae-Min Shin[1]

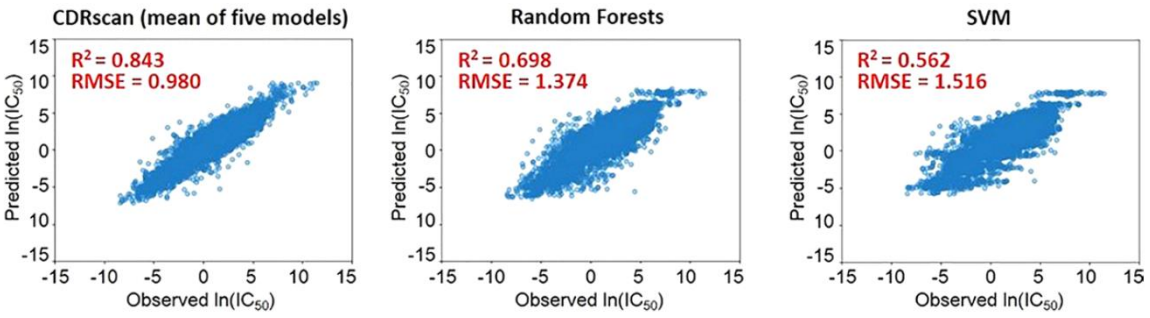


- GDSC
- 28,328 mutation positions in 567 genes
- 787 cell lines
- 244 drugs

Chang, Yoosup, et al. "Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature." Scientific reports 8.1 (2018): 8857.

Related study: prediction of cancer cell sensitivity to drugs

a

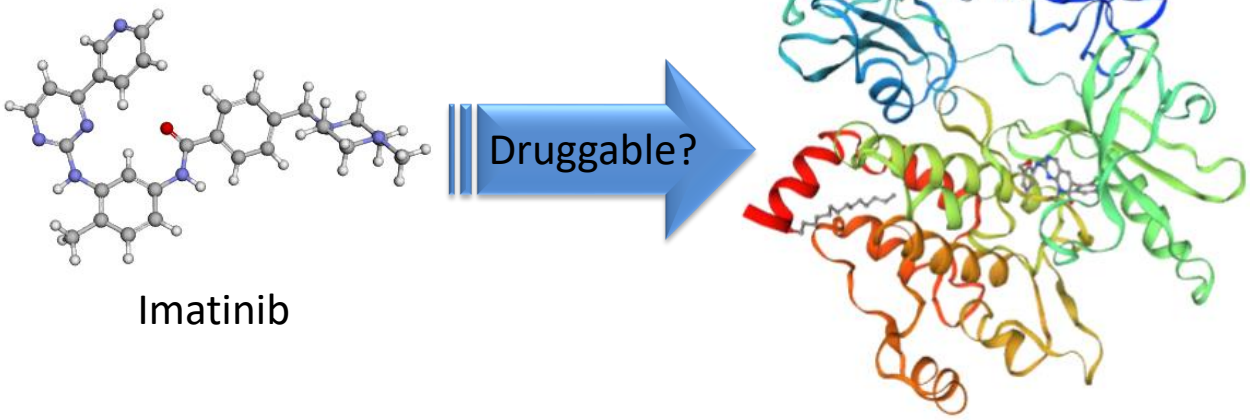


- multi-fold cross validation (five-fold with each fold)

Chang, Yoosup, et al. "Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature." Scientific reports 8.1 (2018): 8857.

PROTEIN SEQUENCE AND DRUG INTERACTIONS

Prediction of drug-target interaction



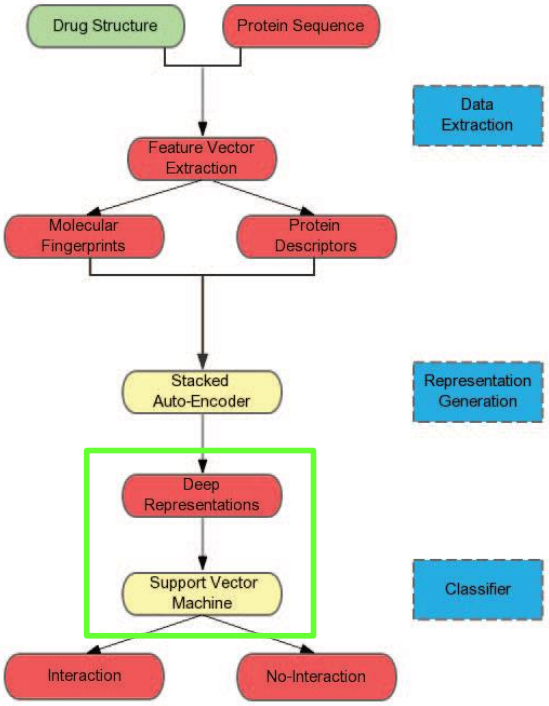
Imatinib

BCR/ABL fusion protein

DTI prediction using protein descriptors

Large-Scale Prediction of Drug-Target Interactions from Deep Representations

Peng-Wei Hu Keith C.C. Chan Zhu-Hong You
Department of Computing
Hong Kong Polytechnic University
Hung Hom, Kowloon
Hong Kong
{csphu, cskcchan, csyzhuhong }@comp.polyu.edu.hk



MFDR employed stacked Auto-Encoder(SAE) to abstract original features into a latent representation with a small dimension. With latent representation, they trained a support vector machine(SVM), which performed better than previous methods, including feature-and similarity-based methods.

Chan, Keith CC, and Zhu-Hong You. "Large-scale prediction of drug-target interactions from deep representations." *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016.

Multi-scale features deep representations inferring interactions (MFDR)

DTI prediction using protein descriptors

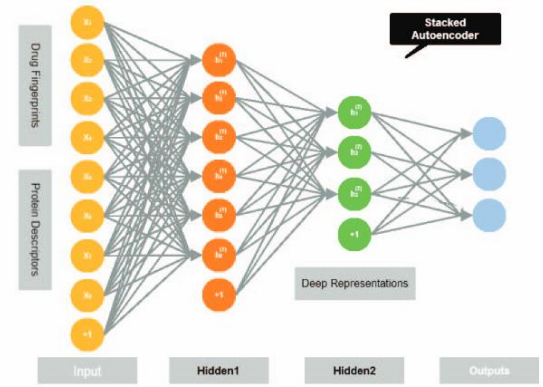
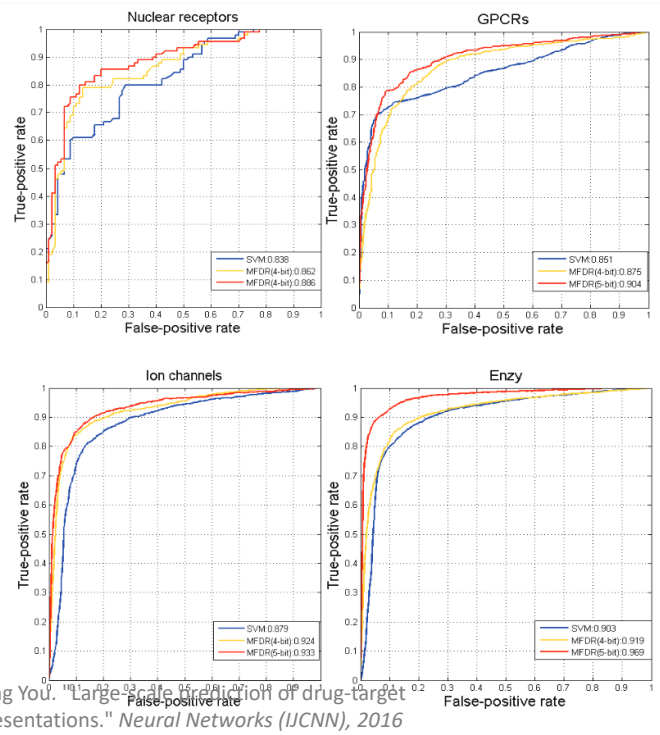


Fig. 2. A Stacked Auto-Encoder composed by two visible layers and two hidden layers

DRUG-TARGET DATA STATISTIC

Type		Ion channel	Enzyme	GPCR	Nuclear receptor
Drugs 881 bits		210	445	223	54
Target proteins		204	664	95	26
567 Descriptors	1449 Descriptors				
Positive Drug–target Interactions		1476	2926	635	90

5fold cross-validation



Chan, Keith CC, and Zhu-Hong You. "Large-scale prediction of drug-target interactions from deep representations." *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016.

DTI prediction using protein sequence

Bioinformatics, 34, 2018, i821–i829
doi: 10.1093/bioinformatics/bty593
ECCB 2018

DeepDTA: deep drug–target binding affinity prediction

Hakime Öztürk[1], Arzucan Özgür[1,*] and Elif Ozkirimli[2,*]

- **Model**
  - Input – Protein sequence, SMILES
  - Output – Binding affinity
  - Model – CNN for protein, DNN for drug
- **Contribution**
  - first used CNN to learn representations of proteins

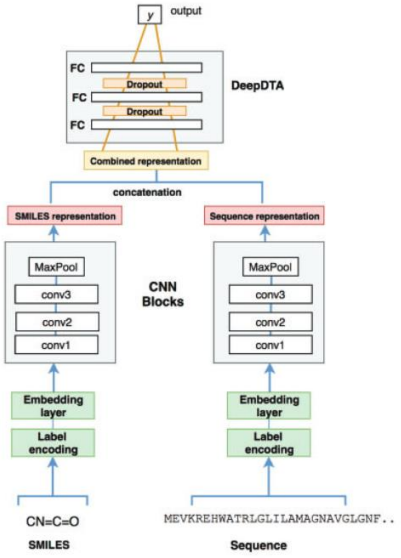


Fig. 2. DeepDTA model with two CNN blocks to learn from compound SMILES and protein sequences

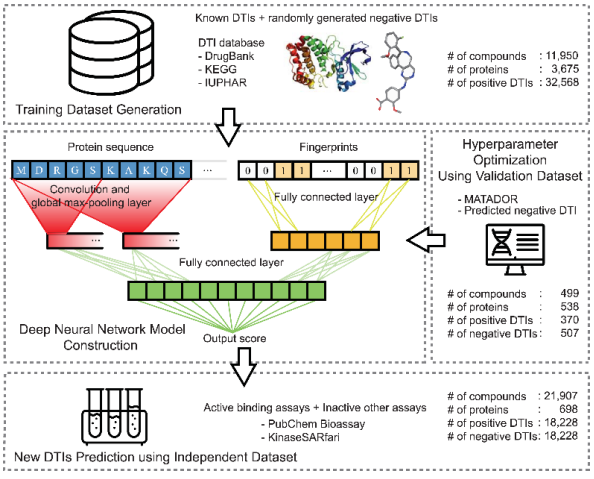
DTI prediction using protein sequence

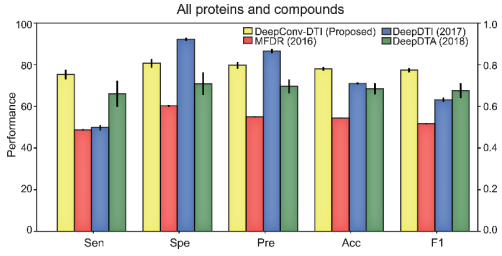
RESEARCH ARTICLE

DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences

Ingoo Lee, Jongsoo Keum, Hojung Nam*

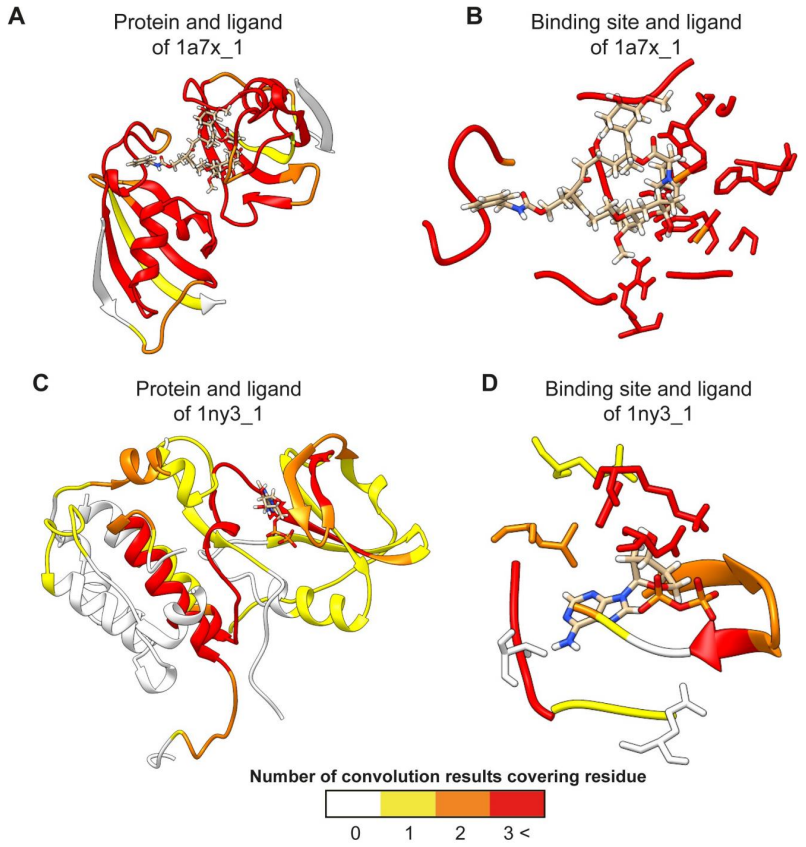
- **Model**
  - Input – Protein sequence, ECFP4
  - Output – Interaction/Non-interaction
  - Model – CNN for protein, DNN for drug
- **Contribution**
  - Embedding representation of protein works well
  - Model can capture local residue patterns





Lee I, Keum J, Nam H (2019) DeepConvDTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol 15(6): e1007129. https://doi. org/10.1371/journal.pcbi.1007129

- Compare pooled convolution result with binding sites from sc-PDB



- FKB1A [Enzyme]
- MAPK2 [Kinase]

Lee I, Keum J, Nam H (2019) DeepConvDTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol 15(6): e1007129. https://doi. org/10.1371/journal.pcbi.1007129

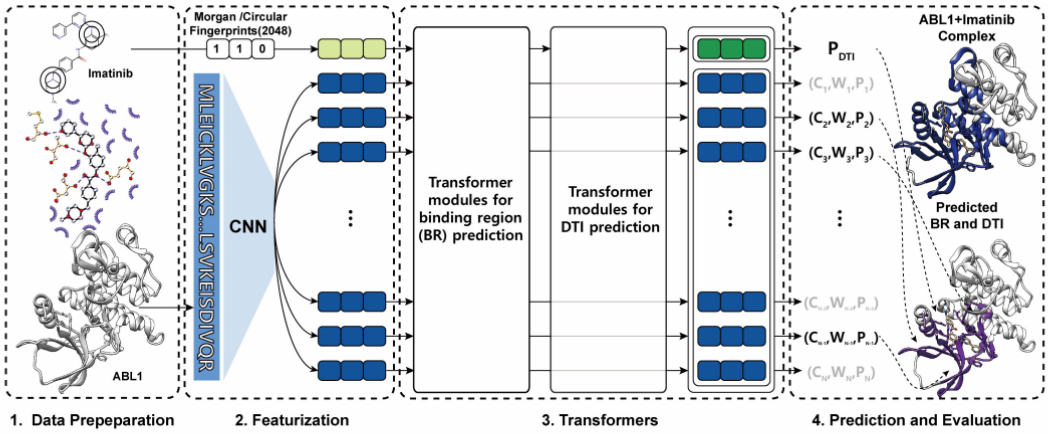


Fig. 1. HoTS model overview. HoTS considers amino acid sequences of individual proteins and Morgan/circular fingerprints of drug compounds. Therefrom, local residue patterns are extracted by a convolutional neural network, and maximum values are pooled from each protein grid. Compound and protein grids are taken into transformers to model interactions between local residue patterns and individual compounds. After passing the transformers, a compound token is used to predict DTIs, and individual protein grids are used to reflect binding regions (BR). For DTI prediction, HoTS calculates a prediction score $\mathbf{P_{DTI}}$ ranging from 0 to 1 and center (C), length (W), and confidence (P) scores for binding regions.

Ingoo Lee, Hojung Nam*, "Sequence-based prediction of binding regions and drug-target interactions", Under review.

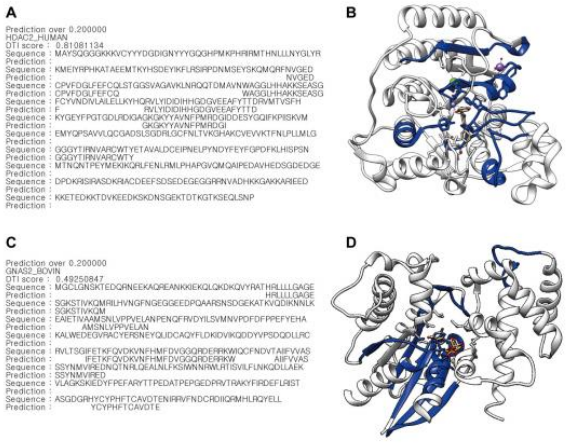


Fig. 3. Prediction and visualization of binding regions on 3D-complexes. A) Predicted binding regions for drug-target interactions between HDAC2_HUMAN and N-(4-amino-biphenyl-3-yl)benzamide (LLX). **B)** Visualization of predicted binding regions on the 3D complex of human HDAC2 complexed with LLX (Protein Data Bank: 3MAX). **C)** Predicted binding regions between GNAS2_BOVIN and 5'-guanosine-diphosphate-monothiophosphate (GSP). **D)** Visualization of predicted binding regions on the 3D complex of bovine GNAS2 complexed with GSP (Protein Data Bank: 1CUL).

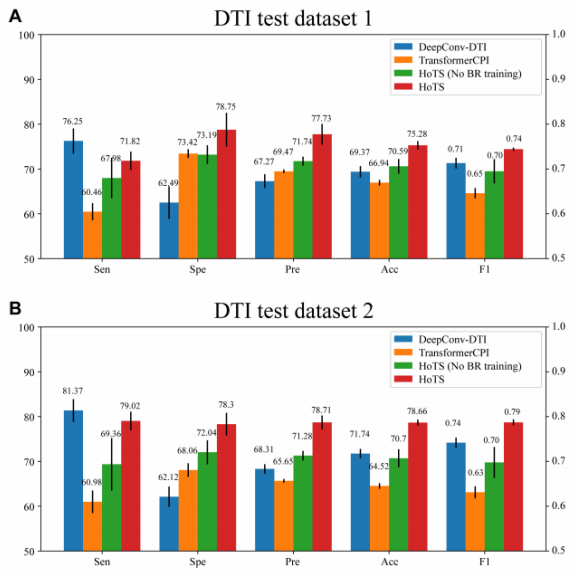


Fig. 4. Prediction performance for drug-target interactions in the independent test datasets.

Ingoo Lee, Hojung Nam*, "Sequence-based prediction of binding regions and drug-target interactions", Under review.

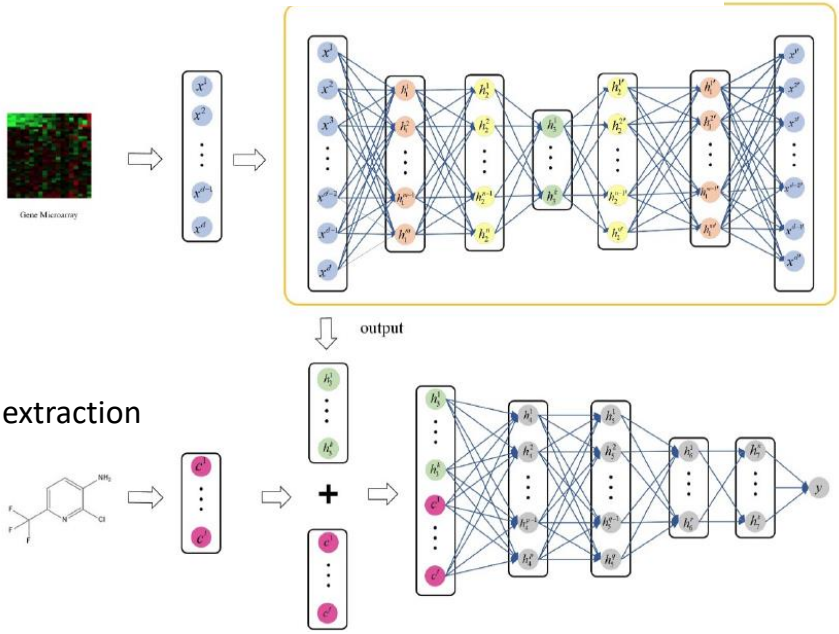
GENE EXPRESSION AND DRUG RESPONSE

Related study: prediction of cancer cell sensitivity to drugs

DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines

Min Li, Yake Wang, Ruiqing Zheng, Xinghua Shi, Yaohang Li, Fang-Xiang Wu, and Jianxin Wang

- GDSC, CCLE
- Transcriptomic feature
- Morgan fingerprint
- Autoencoder based feature extraction



Li, Min, et al. "DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines." *IEEE/ACM transactions on computational biology and bioinformatics* (2019).

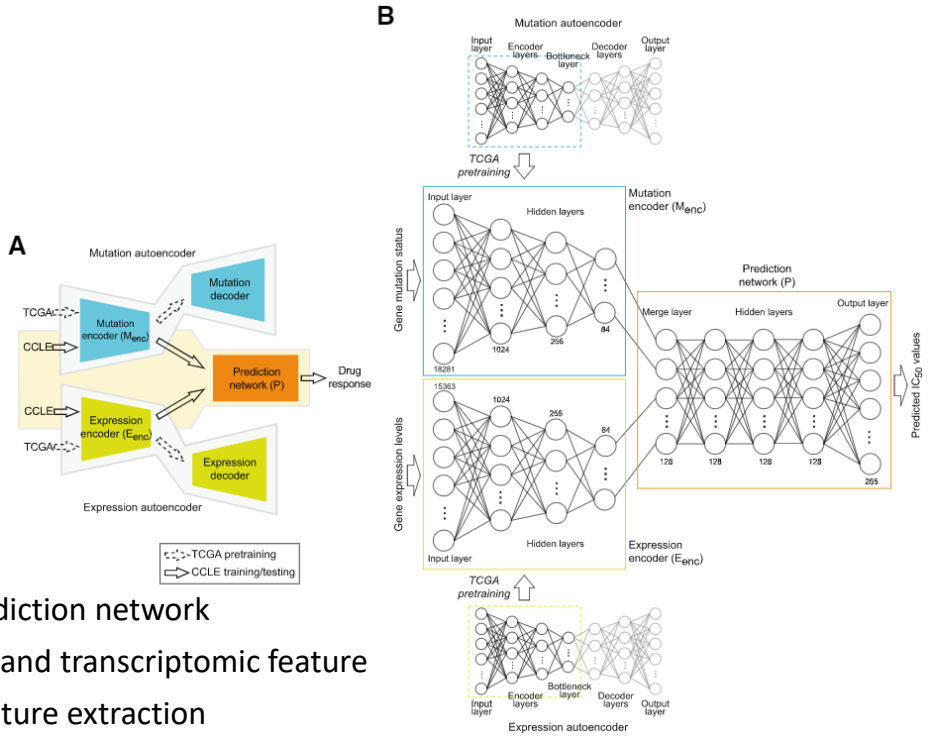
Related study: prediction of cancer cell sensitivity to drugs

	method	NN	KBMF	RF	DeepDSC
CV	RMSE	0.83	0.83+/-1.00	0.75+/-0.01	0.52+/-0.01
	R^2	0.72	0.32+/-0.37	0.74+/-0.01	0.78+/-0.01
LOTO	RMSE	0.99	NA	0.81+/-0.16	0.64+/-0.05
	R^2	0.61	NA	0.72+/-0.08	0.66+/-0.07
LOCO	RMSE	NA	0.85+/-0.41	1.40+/-0.80	1.24+/-0.74
	R^2	NA	0.52+/-0.37	0.13+/-0.11	0.04+/-0.06

- 10-fold cross-validation
- Better performance than typical machine learning methods
- Deep learning based feature extraction

Li, Min, et al. "DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines." *IEEE/ACM transactions on computational biology and bioinformatics* (2019).

Related study: prediction of cancer cell sensitivity to drugs

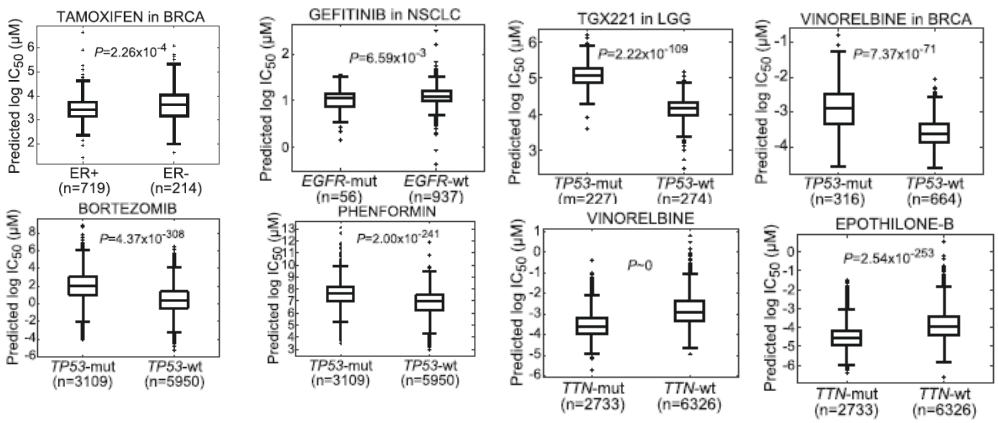


- TCGA for pre-training
- GDSC for response prediction network
- Using both of genomic and transcriptomic feature
- Autoencoder based feature extraction

Chiu, Yu-Chiao, et al. "Predicting drug response of tumors from integrated genomic profiles by deep neural networks." BMC medical genomics 12.1 (2019): 18.

Related study: prediction of cancer cell sensitivity to drugs

Measurement	DeepDR	Linear regression	SVM	Random initialization	PCA	E_{enc} only	M_{enc} only
Median MSE in testing samples[a]	1.96	10.24[b]	8.92[c]	2.30	2.44	1.96	3.09
Median number of training epochs[a]	14	–	–	9	29	17	9.5



- Samples with mutation showed significantly different result compared to non-mutated samples

Chiu, Yu-Chiao, et al. "Predicting drug response of tumors from integrated genomic profiles by deep neural networks." BMC medical genomics 12.1 (2019): 18.

Related study: prediction of cancer cell sensitivity to drugs

molecular pharmaceutics

Article

Cite This: Mol. Pharmaceutics 2019, 16, 4797−4806

pubs.acs.org/molecularpharmaceutics

Toward Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders

Matteo Manica,[†,#] Ali Oskooei,[†,#] Jannis Born,[†,‡,⊥,#] Vigneshwari Subramanian,[§] Julio Sáez-Rodríguez,[∥] and María Rodríguez Martínez*[,†]

[†]IBM Research, 8803 Zürich, Switzerland
[‡]ETH Zürich, 8092 Zürich, Switzerland
[⊥]University of Zürich, 8006 Zürich, Switzerland
[§]RWTH Aachen University, 52056 Aachen, Germany
[∥]Heidelberg University, 69047 Heidelberg, Germany

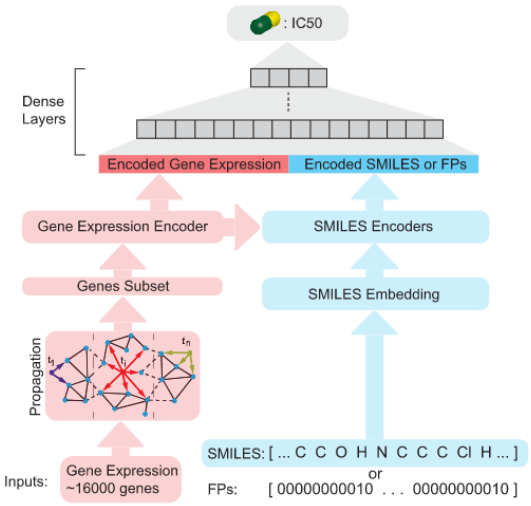


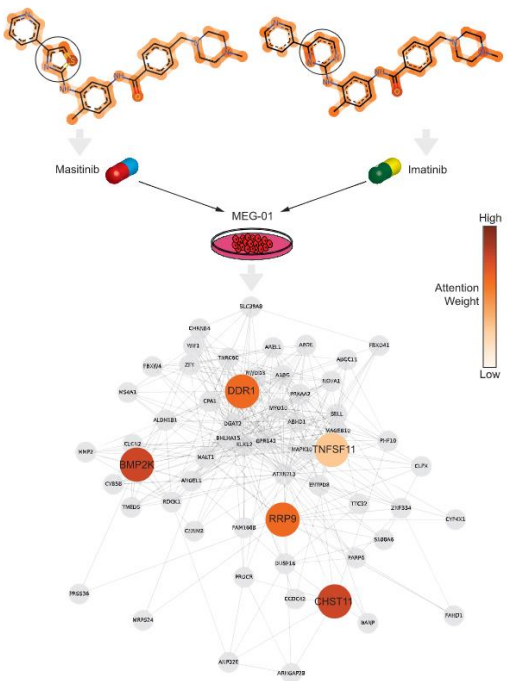
Figure 1. Multimodal end-to-end architecture of the proposed encoders. General framework for the explored architectures. Each model ingests a cell−compound pair and makes an IC50 drug sensitivity prediction. Cells are represented by the gene expression values of a subset of 2128 genes, selected according to a network propagation procedure. Compounds are represented by their SMILES string (apart from the baseline model that uses 512-bit fingerprints). The gene-vector is fed into an attention-based gene encoder that assigns higher weights to the most informative genes. To encode the SMILES strings, several neural architectures are compared (for details see section 2) and used in combination with the gene expression encoder in order to predict drug sensitivity.

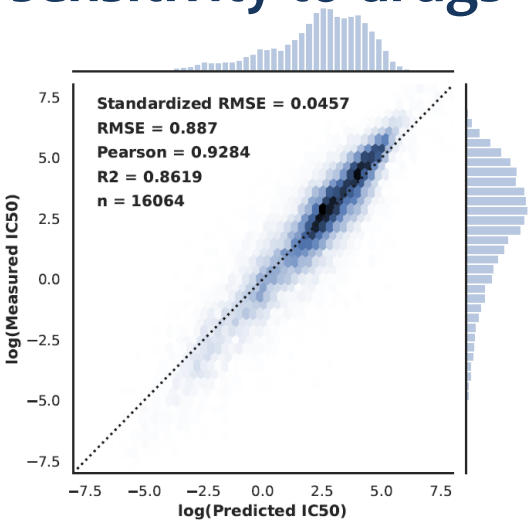
- Transcriptomic feature
- PPI for feature selection
- SMILES

- Attention based model
  - Interpretable

Manica, Matteo, et al. "Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders." Molecular Pharmaceutics (2019).

Related study: prediction of cancer cell sensitivity to drugs





Encoder type	Drug structure	Standardized RMSE Median ± IQR
Deep baseline (DNN)	Fingerprints	0.122 ± 0.010
Bidirectional recurrent (bRNN)	SMILES	0.119 ± 0.011
Stacked convolutional (SCNN)	SMILES	0.130 ± 0.006
Self-attention (SA)	SMILES	0.112* ± 0.009
Contextual attention (CA)	SMILES	0.110* ± 0.007
Multiscale convolutional attentive (MCA)	SMILES	0.109* ± 0.009
MCA (prediction averaging)	SMILES	**0.104**** ± **0.005**

Contents

- PART1
  - Introduction to pharmacogenomics
    - Drug discovery and development
  - Key data sources
  - Representations of proteins, chemicals
- PART2
  - Studies related to pharmacogenomics based on machine learning

End
-Q&A-