KSBi-BIML 2026

Bioinformatics & Machine Learning(BIML) Workshop for Life Scientists

생명정보학 & 머신러닝 워크샵 (온라인) ▶

Drug discovery and development – Pharmacogenomics and beyond

남호정 _ GIST

KSBI | 한국생명정보학회
KOREAN SOCIETY FOR BIOINFORMATICS

KSBi-BIML 2026

Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists

한국생명정보학회가 주최하는 BIML-2026 동계 Bioinformatics &Machine Learning 교육 워크숍에 여러분을 초대합니다.

BIML 워크숍은 생명정보학 연구자들이 최신 AI바이오 분야의 인공지능 기반 분석 기술과 바이오 데이터 분석 기법을 이론과 실습을 통해 체계적으로 배울 수 있는 전문 교육 프로그램입니다. 2015년에 시작된 BIML 워크숍은 올해로 12년 차를 맞이하며, 국내 생명정보학 분야의 최초이자 최고 수준의 교육 프로그램으로 자리 잡았습니다. 이번 워크숍은 크게 인공지능바이오(AI바이오) 분야와 디지털바이오 분야, 두 분야로 구성됩니다.

AI바이오 분야에서는 생명정보 분석에 폭넓게 응용되고 있는 다양한 인공지능 기반 자료 모델링 기법을 다룰 예정입니다. 특히, 인공지능 심층학습을 활용한 단백질 구조 예측, 유전체 분석, 신약 개발에 대한 이론 및 실습 강의가 진행됩니다.

또한 디지털바이오 분야에서는 단일세포오믹스, 공간오믹스, 멀티오믹스, 메타오믹스에 대한 강의도 마련되어 있어, 연구자들의 분석 역량 강화에 실질적인 도움을 줄 것으로 기대됩니다.

또한 2024년부터 추가된 의료정보 자료 분석을 다루는 강의를 올해도 지속해서 운영하고자 합니다. 이는 최근 의료정보 자료 분석에 관한 연구 수요 증가를 반영한 것으로, 관련 연구를 수행하는 의과학자 및 의료정보 연구자들에게 유용한 지침을 제공할 것입니다.

또한, 올해도 생명정보학 기술의 다양화에 발맞춰 온라인 강좌를 대폭 확대했습니다. 올해는 무료 강좌 10개를 포함한 총 40개 이상의 강좌가 개설되며, 연구 주제에 맞는 강좌 추천과 강연료 할인 혜택도 제공합니다.

BIML-2026는 국내 주요 연구 중심 대학의 전임 교수 및 각 분야 최고 전문가들의 강의로 구성되어 있으며, 기초 이론부터 최신 연구 동향까지 아우르는 심도 있는 교육의 장이 될 것으로 확신합니다.

여러분의 많은 관심과 참여를 기대합니다!

2026년 2월

한국생명정보학회장 류 성 호

Drug discovery and development
- Pharmacogenomics and beyond

본 수업에서는 빅데이터와 AI 기반 신약개발 연구 동향에 초점 맞춘다. 약물 발굴 단계에서 AI 적용 분야로 유효물질 탐색, ADME/Tox 예측 등 최신 AI 기술과 빅데이터의 잠재력을 활용한 다양한 연구 기술들에 대하여 알아본다. 또한 개인별 유전자에 따른 약물 반응을 연구/예측하는데 필요한 생명정보학적 접근 방식을 알아본다.

강의는 다음의 내용을 포함한다:

- Drug discovery and development 기본 개념
- Pharmacogenomics 기본 개념
- Proteins, molecules representation features
- 최신동향 AI기반 약물 개발 연구 소개

* 교육생준비물:

강의 동영상 플레이가 가능한 컴퓨터

Google Colab 사용 가능 컴퓨터

* 강의: 남호정 교수 (광주과학기술원 전기전자컴퓨터공학부)

Curriculum Vitae

Speaker Name: Hojung Nam, Ph.D.

▶ **Personal Info**

Name	Hojung Nam
Title	Professor
Affiliation	Gwangju Institute of Science and Technology (GIST)

▶ **Contact Information**

Address	123 Cheomdangwagi-ro, Buk-gu, Gwangju, 61005, Republic of Korea
Email	hjnam@gist.ac.kr

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)
2018-2023	Associate Professor, Gwangju Institute of Science and Technology (GIST)
2023-	Professor, Gwangju Institute of Science and Technology (GIST)

Selected Publications (5 maximum)

1. Bongsung Bae, Haelee Bae, **Hojung Nam***, "LOGICS: Learning optimal generative distribution for designing de novo chemical structures", Journal of Cheminformatics 2023 Sep 7;15(1):77.

2. Haelee Bae, **Hojung Nam***, "GraphATT-DTA: attention-based novel representation of interaction to predict drug-target binding affinity", Biomedicines 2023, 11(1), 67.

3. Hansol Lee, Songyeon Lee, Ingoo Lee, **Hojung Nam***, "AMP-BERT: Prediction of Antimicrobial Peptide Function Based on a BERT Model", Protein Science, 2022 Dec 3;e4529. doi: 10.1002/pro.4529.

4. Koon Mook Kang§, Ingoo Lee§, **Hojung Nam***, Yong-Chul Kim*, "AI-Based Prediction of New Binding Site and Virtual Screening for the Discovery of Novel P2X3 Receptor Antagonists", European Journal of Medicinal Chemistry, 2022 Jul 1;240:114556.

5. Hyunho Kim, Minsu Park, Ingoo Lee, **Hojung Nam***, "BayeshERG: A Robust, Reliable, and Interpretable Deep Learning Model for Predicting hERG Channel Blockers", Briefings in Bioinformatics 2022 Jun 17;bbac211. doi: 10.1093/bib/bbac211.

KSBi-BIML 2024

Drug discovery and development - Pharmacogenomics and beyond

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

Contents

- Lecture 1
  - Introduction to pharmacogenomics
    - Drug discovery and development
  - Key data sources
  - Representations of proteins, chemicals

- Lecture 2
  - 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]
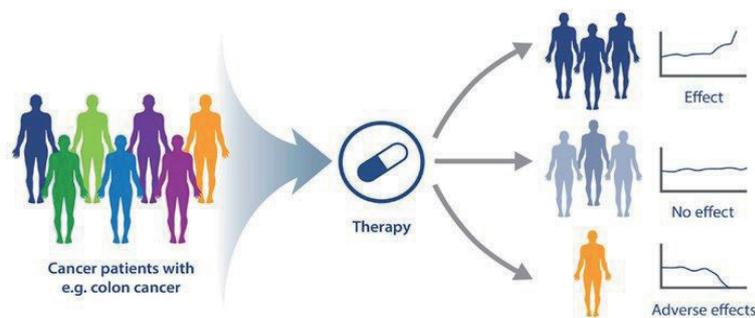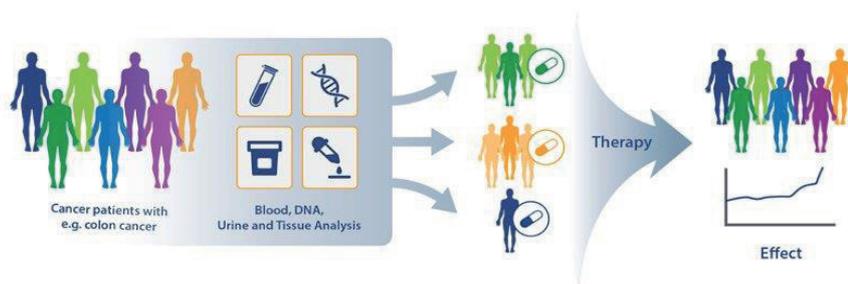
https://blog.crownbio.com/pdx-personalized-medicine#_

Drug discovery and development

Drug discovery			Pre-clinical	Clinical trials			Market
Target discovery	Hit screening	Lead optimization	Pre-clinical	Phase1	Phase2	Phase3	

High throughput screening

Literature study KO/KD test

3D Modeling SAR/QSAR

ADME/PK

Animal studies

Efficacy 100~200 patients

Safety / PK 50~150 patients

Efficacy 500~5000 patients

SBi 한국생명정보학회
Korean Society for Bioinformatics
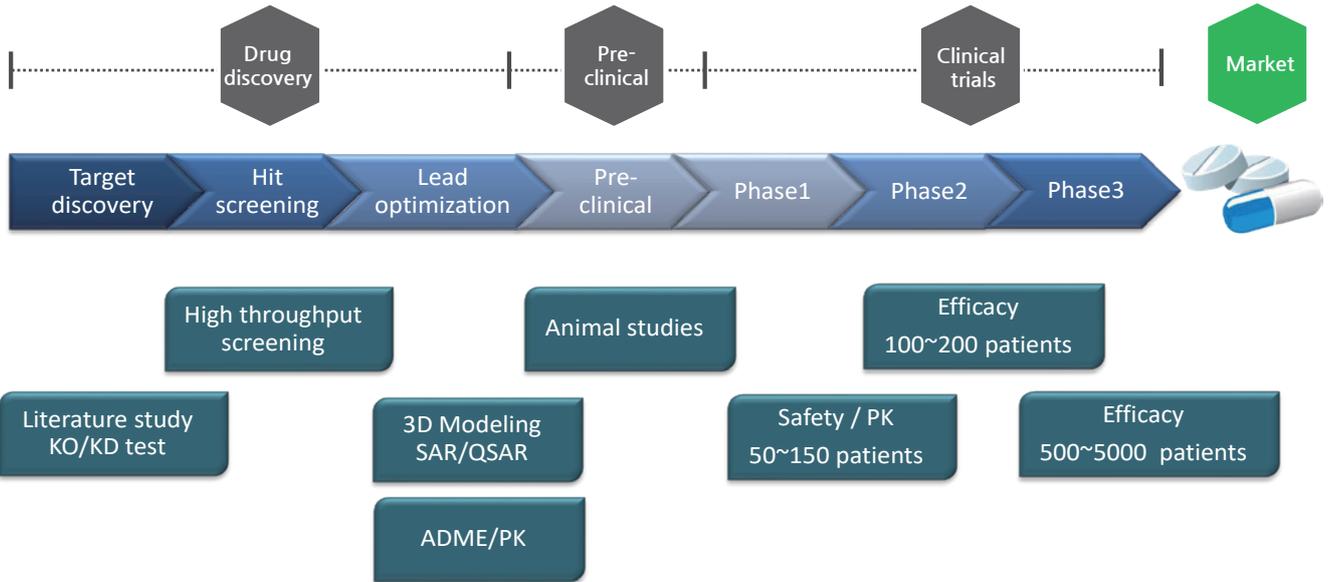
Pharmacogenomics in drug discovery and development

Drug discovery			Pre-clinical	Clinical trials			Market
Target discovery	Hit screening	Lead optimization	Pre-clinical	Phase1	Phase2	Phase3	

Suggest best candidates

Suggest best trial case

Target discovery w/ variations

ADME/T (CYP450)

Interactions w/ variations

Drug repositioning

Patients stratification

SBi 한국생명정보학회
Korean Society for Bioinformatics
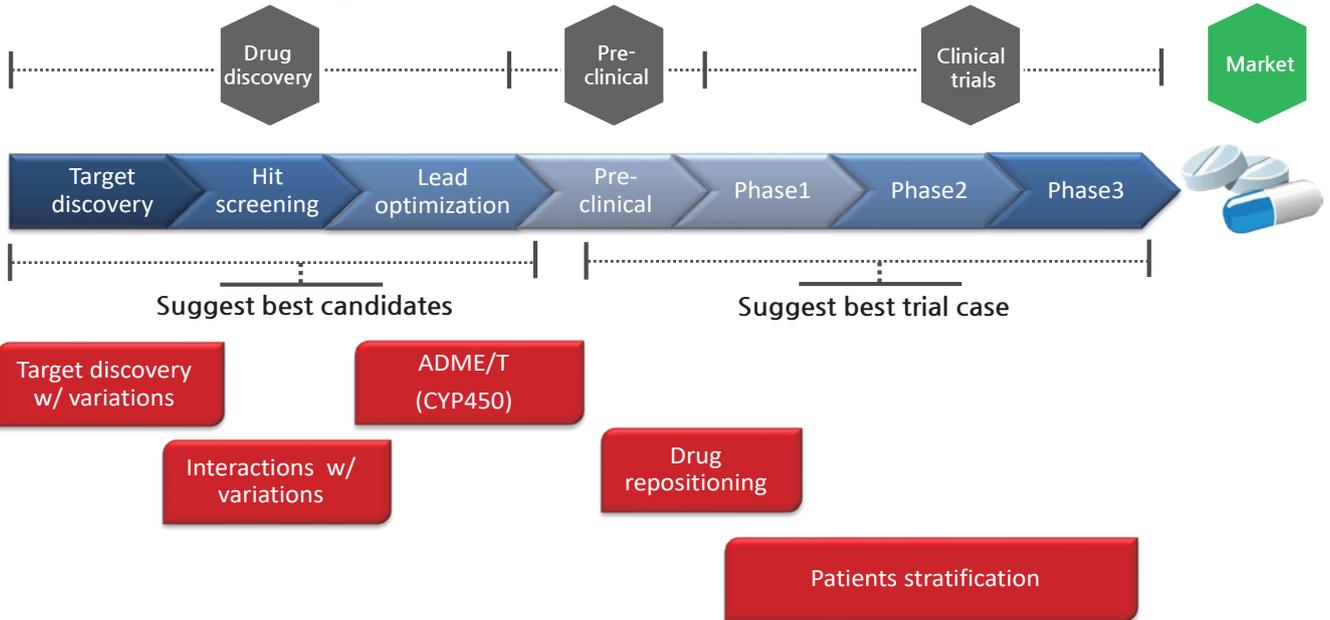
Example 1 – TPMT

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)

TPMT function

Azathioprine or 6-mercaptopurine dose	→	Normal function	→	Expected drug effect
		Decreased function	→	Risk of haematological toxicity
		No function	→	High risk of haematological toxicity

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

SBi 한국생명정보학회
Korean Society for Bioinformatics
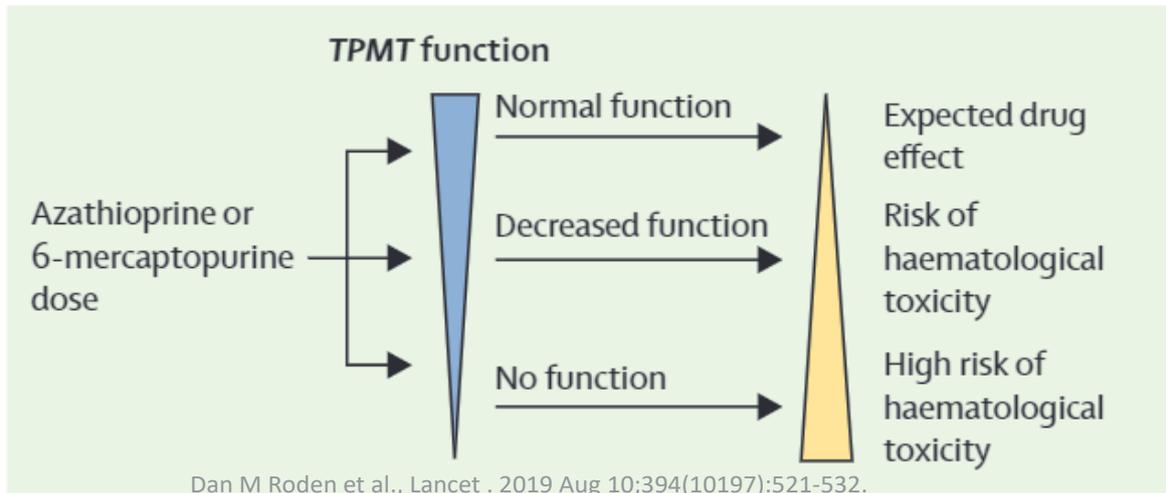
Example 2 – CYP2D6

- 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.
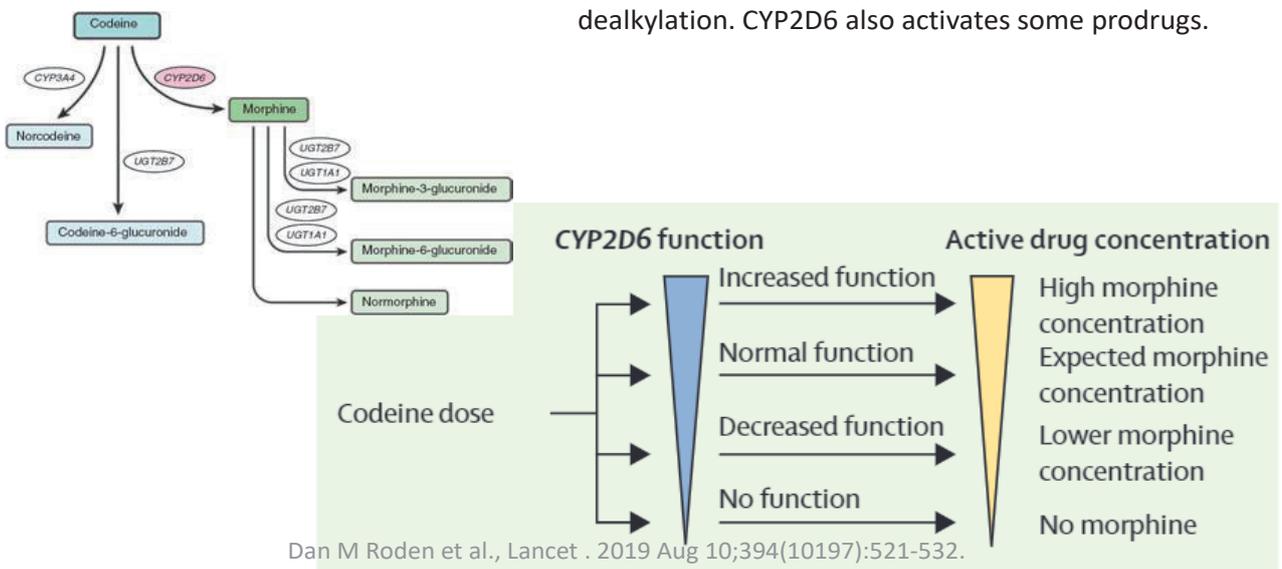
Codeine
CYP3A4 CYP2D6
Norcodeine Morphine
UGT2B7
  UGT2B7
  UGT1A1 Morphine-3-glucuronide
Codeine-6-glucuronide UGT2B7
  UGT1A1 Morphine-6-glucuronide
  Normorphine

CYP2D6 function   Active drug concentration

Codeine dose	→	Increased function	→	High morphine concentration
		Normal function	→	Expected morphine concentration
		Decreased function	→	Lower morphine concentration
		No function	→	No morphine

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

SBi 한국생명정보학회
Korean Society for Bioinformatics

- Lecture 1
  - Introduction to pharmacogenomics
    - Drug discovery and development
  - Key data sources
  - Representations of proteins, chemicals
- Lecture 2
  - 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)
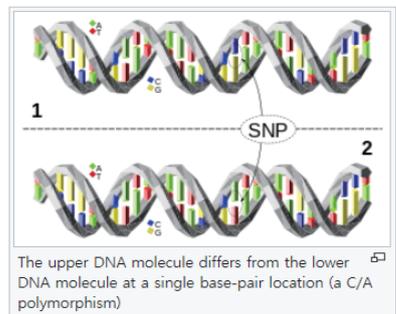
https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism

NCBI dbSNP

gnomAD

https://gnomad.broadinstitute.org/

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
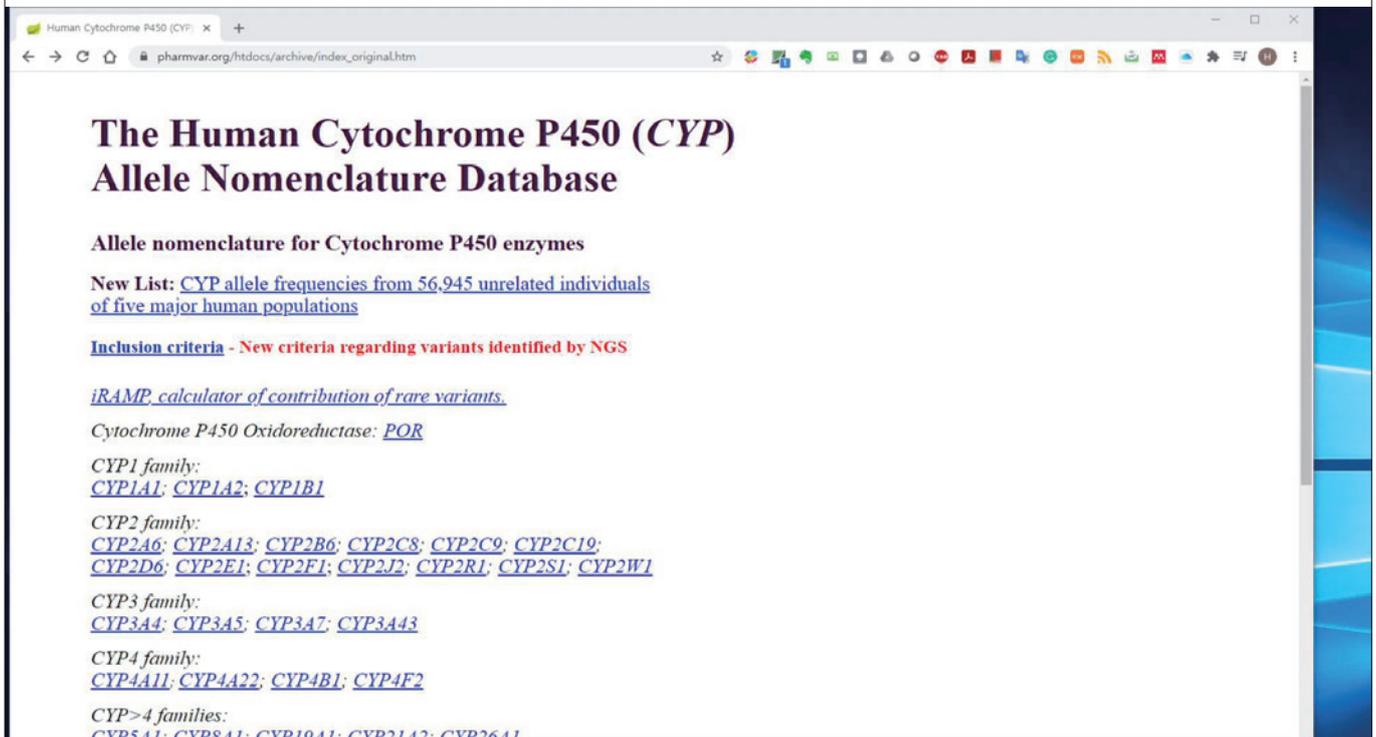
SBi 한국생명정보학회 Korean Society for Bioinformatics

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

SBi 한국생명정보학회 Korean Society for Bioinformatics   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/

The Pharmacogene Variation (PharmVar) Consortium is a central repository for pharmacogene (PGx) variation that focuses on haplotype structure and allelic variation.

The information in this resource facilitates basic and clinical research as well as the interpretation of pharmacogenetic test results to guide precision medicine.

PharmVar API Services are now available for third party use. For more information, visit the API Service Documentation Page

Follow us on Twitter

PharmVar Publications
Articles published by PharmVar are available on the resources page.

Original content from the cypalleles.ki.se site is available through the archive

https://www.pharmvar.org/

20

PHARMGKB

PHARMGKB

Publications   News   Downloads   Contact   ⑦ Help

Search PharmGKB   🔍

Search for a molecule, gene, variant, or combination

Therapeutic Resource for COVID-19

ℹ PharmGKB data are under a Creative Commons license. More details are in our Data Usage Policy. Please cite PharmGKB if you use our information or images.

Drug Label Annotations	Clinical Guideline Annotations	Curated Pathways	Annotated Drugs
📗 780	📖 165	⌥ 151	⧉ 709

https://www.pharmgkb.org/

SBi 한국생명정보학회
Korean Society for Bioinformatics

PharmGKB   × +

← → C ⌂ 🔒 pharmgkb.org

PHARMGKB

Publications   News   Downloads   Contact   ⑦ Help

Search PharmGKB   🔍

Search for a molecule, gene, variant, or combination

Therapeutic Resource for COVID-19

ℹ PharmGKB data are under a Creative Commons license. More details are in our Data Usage Policy. Please cite PharmGKB if you use our information or images.

Drug Label Annotations	Clinical Guideline Annotations	Curated Pathways	Annotated Drugs
📗 780	📖 165	⌥ 151	⧉ 709

https://www.pharmgkb.org/

SBi 한국생명정보학회
Korean Society for Bioinformatics

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.

SBi 한국생명정보학회 Korean Society for Bioinformatics

NCBI PubChem

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

SBi 한국생명정보학회 Korean Society for Bioinformatics

24

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

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/

https://go.drugbank.com/

Genomics of Drug Sensitivity in Cancer (GDSC)

https://www.cancerrxgene.org/

- Lecture 1
  - Introduction to pharmacogenomics
    - Drug discovery and development
  - Key data sources
  - Representations of proteins, chemicals

- Lecture 2
  - Studies related to pharmacogenomics based on machine learning

PROTEIN REPRESENTATIONS

Why protein representations are necessary?
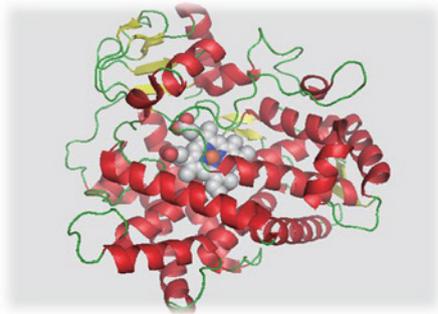
Representation of proteins for machine-learning features that fully captured wide ranges of properties of the target molecule

---

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
  - One-hot embedding
  - Knowledge graph embedding

Amino Acid Composition –AAC (20D)

Amino acid compositions for unfiltered sequences from 38 organisms
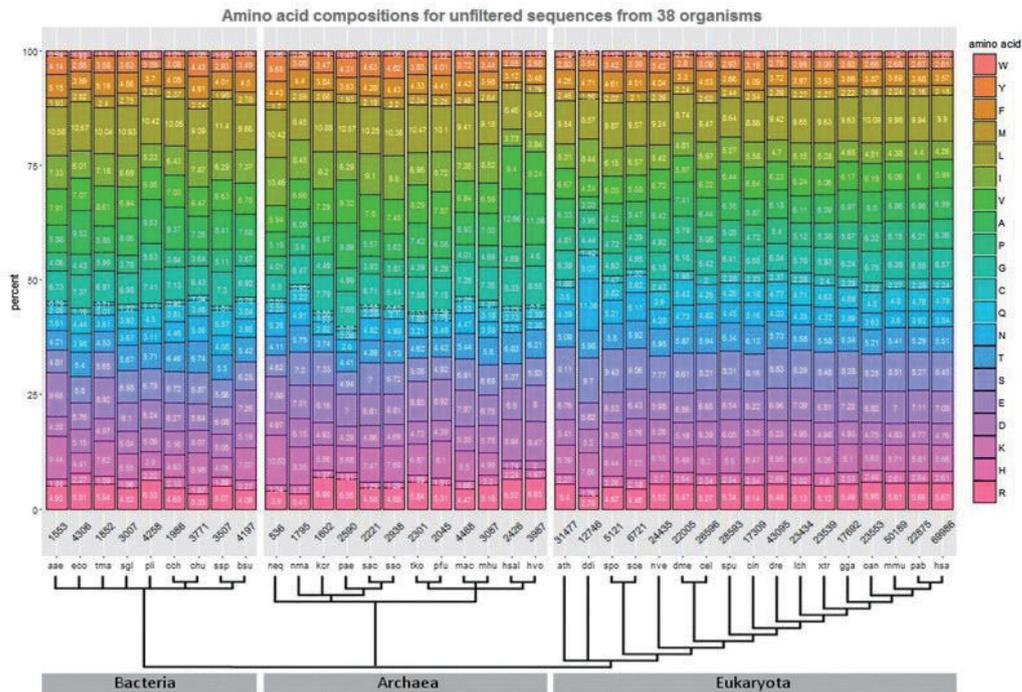
BMC Research Notes volume 11, Article number: 117 (2018)

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          RAA          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
```

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

Protein descriptors (실습코드)

https://colab.research.google.com/drive/1smQsJSVlTKsI7difhyzLco-2eG96mCCL?usp=sharing

Protein Embedding
(Convert Categorical Data to Numerical Data)

- One-Hot Encoding

- Word embedding

Knowledge Graph

- A knowledge graph is a knowledge base that uses a graph-structured data model to integrate data.
  - entities (such as objects, people, and concepts) are depicted as **nodes**
  - relationships or connections between entities are represented as **edges**
- Knowledge graphs enable enhanced **information retrieval**, **reasoning**, and **knowledge discovery**.
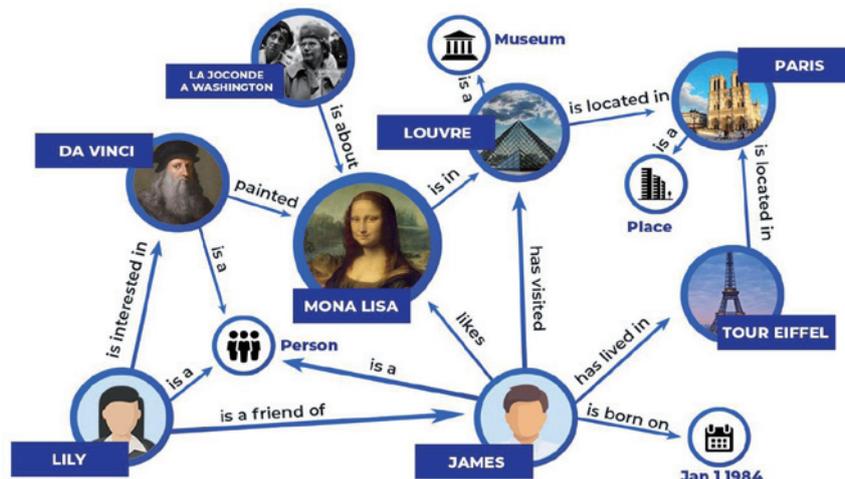
Hetionet (eLife 2017)
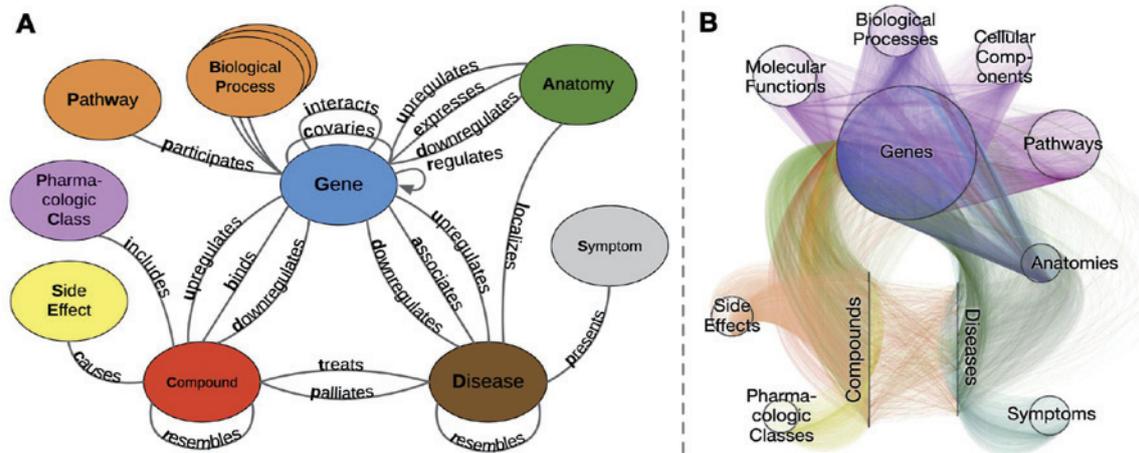
- 11 node types (metanodes), 24 edge types (metaedges)

Hetionet (eLife 2017)

Table 1. Metanodes.

Hetionet v1.0 includes 11 node types (metanodes). For each metanode, this table shows the abbreviation, number of nodes, number of nodes without any edges, and the number of metaedges connecting the metanode.

Metanode	Abbr	Nodes	Disconnected	Metaedges
Anatomy	A	402	2	4
Biological process	BP	11,381	0	1
Cellular component	CC	1391	0	1
Compound	C	1552	14	8
Disease	D	137	1	8
Gene	G	20,945	1800	16
Molecular function	MF	2884	0	1
Pathway	PW	1822	0	1
Pharmacologic class	PC	345	0	1
Side effect	SE	5734	33	1
Symptom	S	438	23	1

Table 2. Metaedges.

Hetionet v1.0 contains 24 edge types (metaedges). For each metaedge, the table reports the abbreviation, the number of edges, the number of source nodes connected by the edges, and the number of target nodes connected by the edges. Note that all metaedges besides Gene→regulates→Gene are undirected.

Metaedge	Abbr	Edges	Sources	Targets
Anatomy–downregulates–Gene	AdG	102,240	36	15,097
Anatomy–expresses–Gene	AeG	526,407	241	18,094
Anatomy–upregulates–Gene	AuG	97,848	36	15,929
Compound–binds–Gene	CbG	11,571	1389	1689
Compound–causes–Side Effect	CcSE	138,944	1071	5701
Compound–downregulates–Gene	CdG	21,102	734	2880
Compound–palliates–Disease	CpD	390	221	50
Compound–resembles–Compound	CrC	6486	1042	1054
Compound–treats–Disease	CtD	755	387	77
Compound–upregulates–Gene	CuG	18,756	703	3247
Disease–associates–Gene	DaG	12,623	134	5392
Disease–downregulates–Gene	DdG	7623	44	5745
Disease–localizes–Anatomy	DlA	3602	133	398
Disease–presents–Symptom	DpS	3357	133	415
Disease–resembles–Disease	DrD	543	112	106
Disease–upregulates–Gene	DuG	7731	44	5630
Gene–covaries–Gene	GcG	61,690	9043	9532
Gene–interacts–Gene	GiG	147,164	9526	14,084
Gene–participates–Biological Process	GpBP	559,504	14,772	11,381
Gene–participates–Cellular Component	GpCC	73,566	10,580	1391
Gene–participates–Molecular Function	GpMF	97,222	13,063	2884
Gene–participates–Pathway	GpPW	84,372	8979	1822
Gene→regulates→Gene	Gr > G	265,672	4634	7048
Pharmacologic Class–includes–Compound	PCiC	1029	345	724

Knowledge Graph Embedding (KGE)

- A Knowledge graph embedding (KGE) is a representation of a KG element into a continuous vector space.
  - The primary objective is to ensure that these embeddings capture the semantics and relations such that similar or related entities/relations are closer in the embedding space.

Image adapted from https://towardsdatascience.com/knowledge-graph-embeddings-101-2cc1ca5db44f
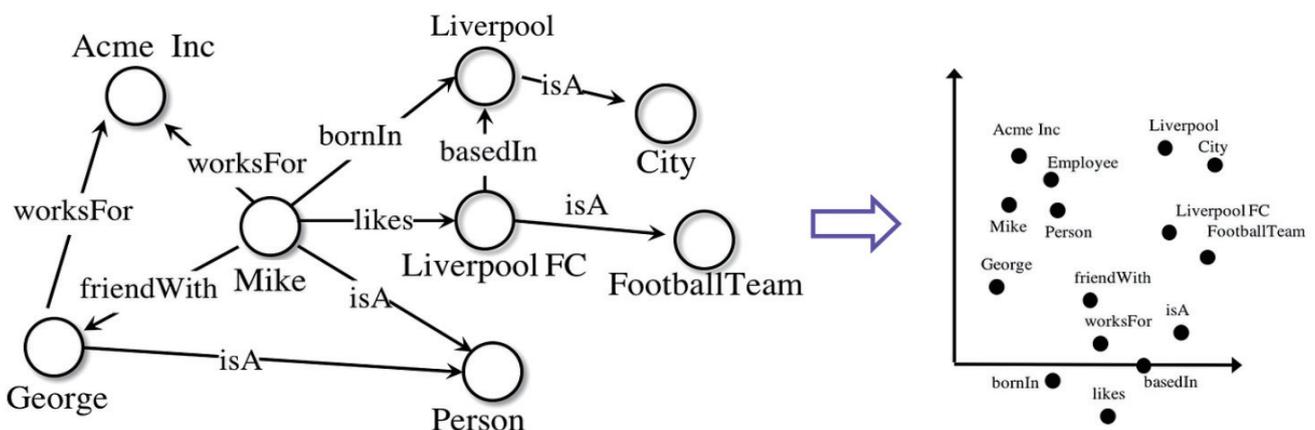
Knowledge Graph Embedding (KGE)
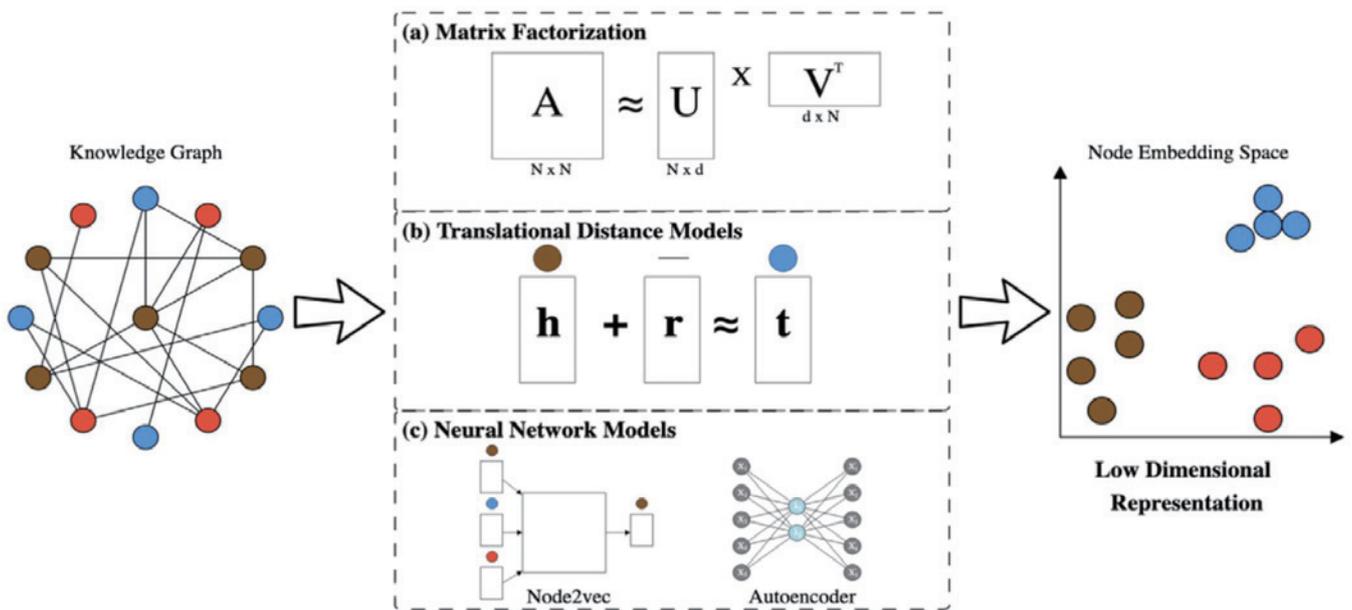
Nicholson et. al., *Comput Struct Biotechnology J* **18**, 1414–1428 (2020)

Translational Models

- TransE :
  - If two entities are related by a specific relationship, the embedding of one entity plus the embedding of the relationship should be close to the embedding of the second entity.
  - For a given triple (h,r,t) (where h is the head entity, r is the relation, and t is the tail entity), the relationship is modeled as: $h + r \approx t$
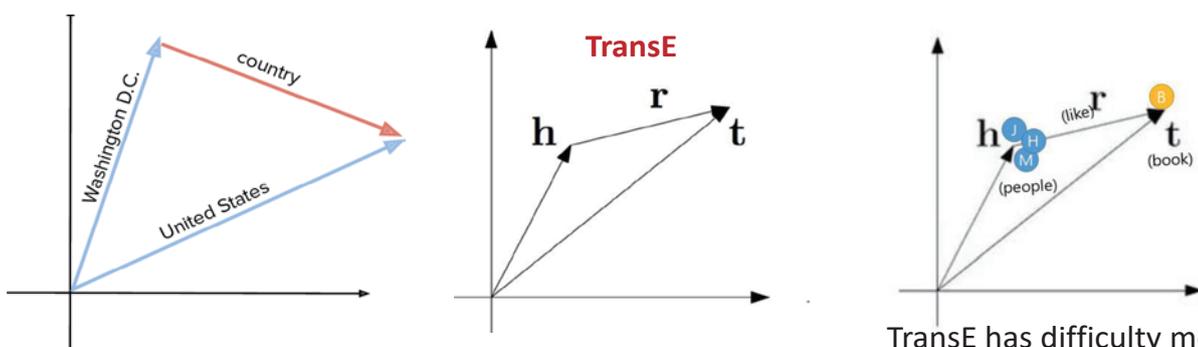
TransE has difficulty modeling 1-N, N-1, and N-N relationships.

https://learnijing.tistory.com/5

Translational Models

- TransR
  - TransR learns relation-specific embeddings. Each relationship has its own embedding space, and entities are transformed into this space before translation.
  - For each relation r, there's a transformation matrix M_r. Entities are first transformed:

    $h_r = h \cdot M_r$

    $t_r = t \cdot M_r$

    Then, the translation is applied: $h_r + r \approx t_r$
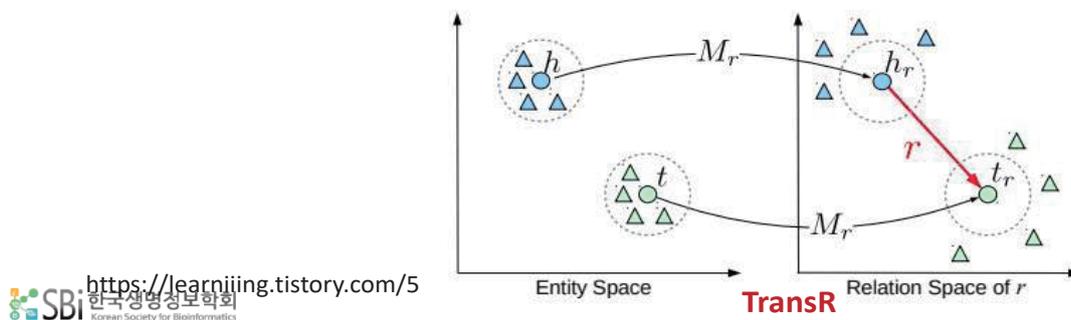
TransR

Translational Models

- RotatE
  - RotatE represents relations as rotations in the complex vector space. For a triple (h, r, t), the relation r is modeled as a rotation from h to t in the complex plane.
  - This approach is particularly powerful for capturing symmetric, antisymmetric, transitive, and inversion properties of relations.

Protein embedding (실습코드)

- Lecture 1
  - Introduction to pharmacogenomics
    - Drug discovery and development
  - Key data sources
  - Representations of proteins, chemicals

- Lecture 2
  - Studies related to pharmacogenomics based on machine learning

MOLECULAR REPRESENTATION

Why molecular representations are necessary?
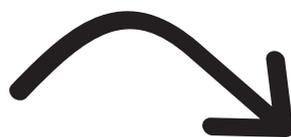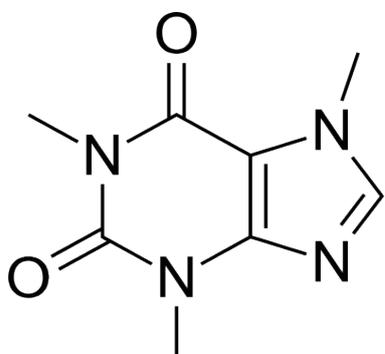
Representation of chemical compounds for machine-learning features that fully captured wide ranges of chemical and physical properties of the target molecule

SBi 한국생명정보학회
Korean Society for Bioinformatics

Types of molecular representations

- Molecular descriptors
- Molecular fingerprints
- Molecular embeddings

SBi 한국생명정보학회
Korean Society for Bioinformatics

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

MW?

194.08

Molecular descriptors

HO O

| 0D | 1D | 2D | 3D | 4D |
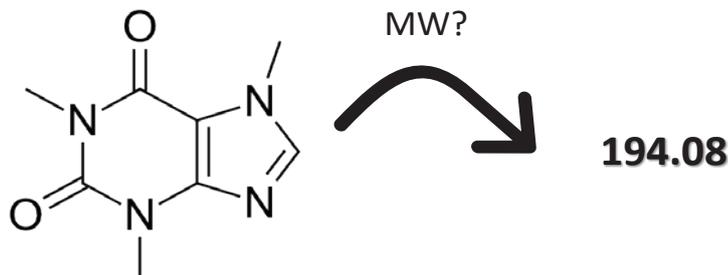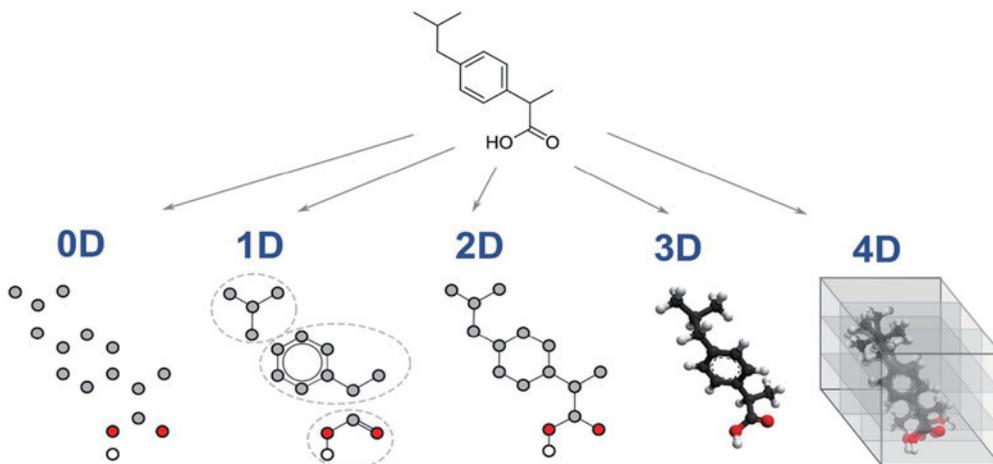
1) **0D-descriptors** (Molecular formula, i.e. Molecular weights, atom counts, bond counts),
2) **1D-descriptors** (Chemical graph, i.e. Fragment counts, functional group counts),
3) **2D-descriptors** (Structural topology, i.e. Wiener index, Balaban index, Randic index, BCUTS),
4) **3D-descriptors** (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 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
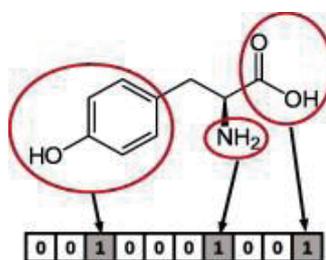
Types of fingerprints

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

SBi 한국생명정보학회
Korean Society for Bioinformatics

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

path length = 0
path length = 1
path length = 2
path length = 3
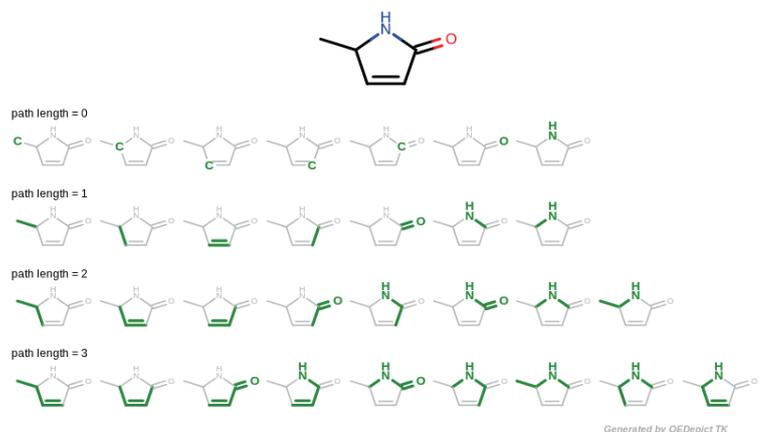
Generated by OEDepict TK

길이에 따른 fragment 추출 예시

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

SBi 한국생명정보학회
Korean Society for Bioinformatics

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

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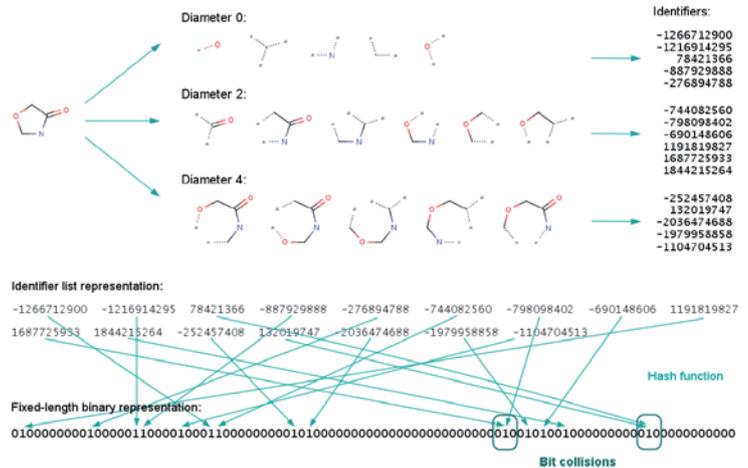
ECFP fingerprint의 산출 절차

Mol2Vec

- Mol2vec learns vector representations of molecular substructures that point in similar directions for chemically related substructures.
- Compounds can finally be encoded as vectors by summing the vectors of the individual substructures
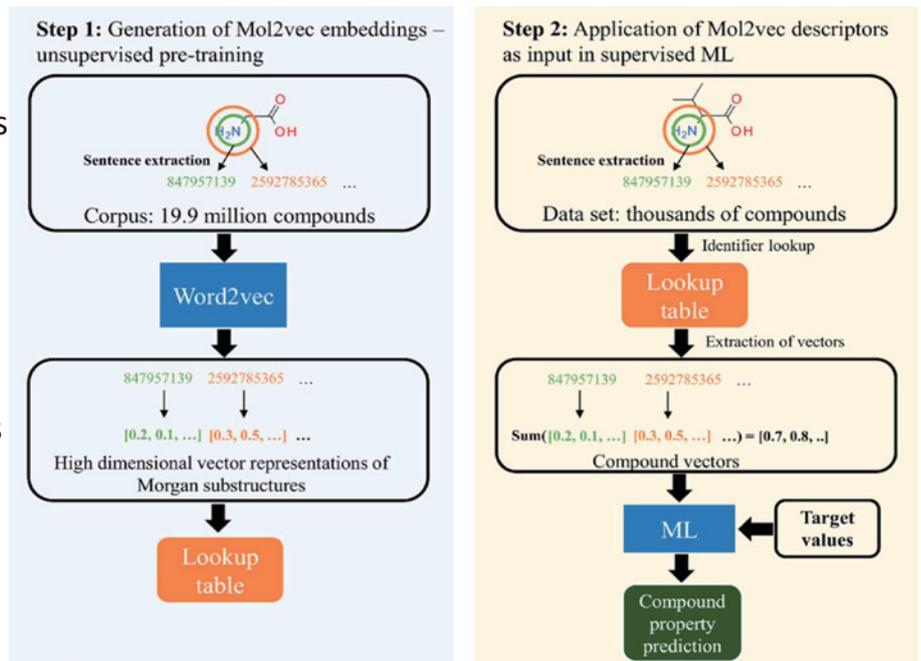
GNN

- 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.
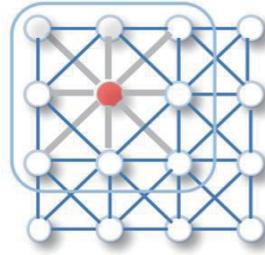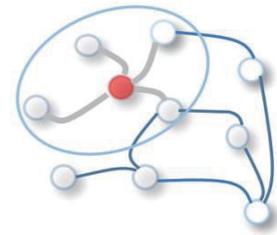
(b) Graph Convolution. To get a hidden representation of the red node, one simple solution of the graph convolutional operation is to take the average value of the node features of the red node along with its neighbors. Different from image data, the neighbors of a node are unordered and variable in 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
    - $h_G^{t+1} = readout(h_v^{t+1}, \forall v \in G)$
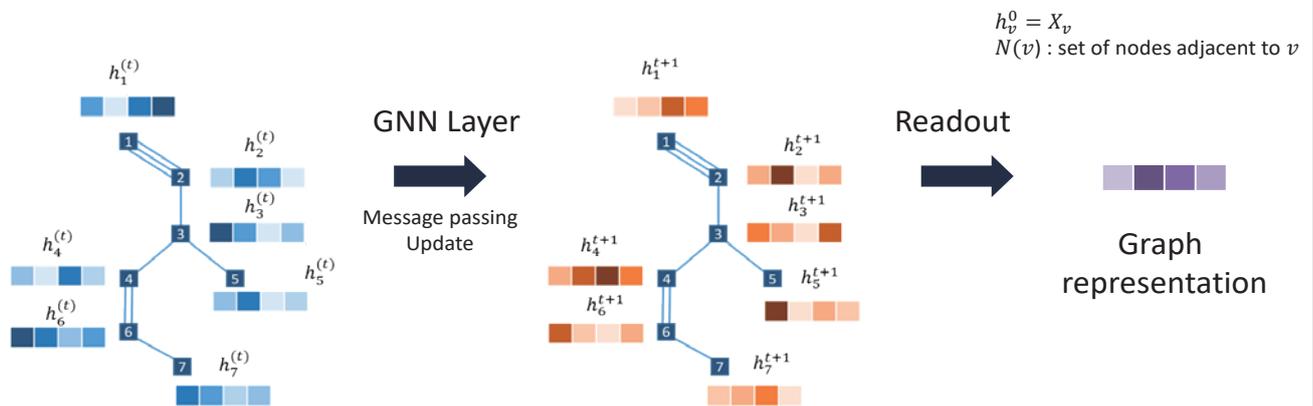
$h_v^0 = X_v$
$N(v)$: set of nodes adjacent to v

GNN Layer

Message passing
Update

Readout

Graph representation

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

60

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

  - $m_v^{(t+1)} = message\_passing(\{h_w^{(t)}, \forall w \in N(v)\})$

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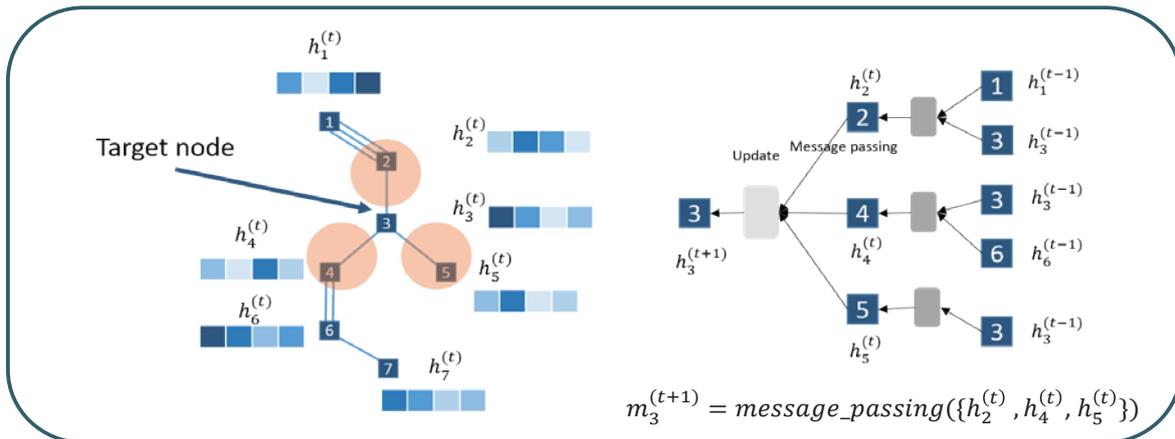
Graph Neural Network

- **Update**
  - $update$: function that update the t+1 time step hidden representation with t time step node representation and message passing

  - $h_v^{t+1} = update(m_v^{(t+1)}, h_v^{(t)})$

$$m_3^{(t+1)} = message\_passing(\{h_2^{(t)}, h_4^{(t)}, h_5^{(t)}\})$$
$$h_3^{t+1} = update(m_3^{(t+1)}, h_3^{(t)})$$
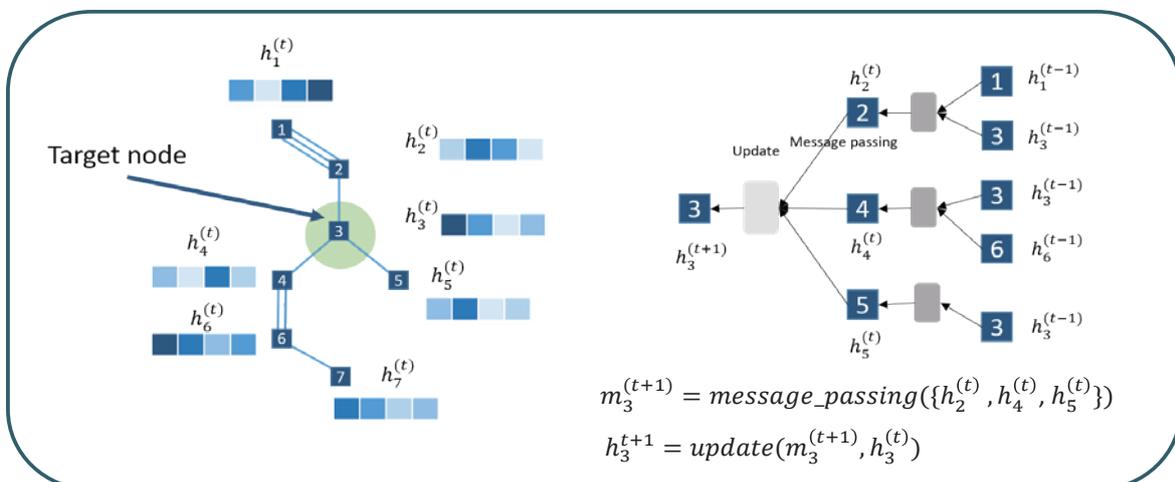
Graph Neural Network

- **Readout**
  - *readout* : function that represent the graph calculated by all hidden representations

  - $h_G^{t+1} = readout(h_v^{t+1}, \forall v \in G)$

$$h_G^{t+1} = readout(\;h_1^{t+1}\; h_2^{t+1}\; h_3^{t+1}\; h_4^{t+1}\; h_5^{t+1}\; h_6^{t+1}\; h_7^{t+1}\;) = $$
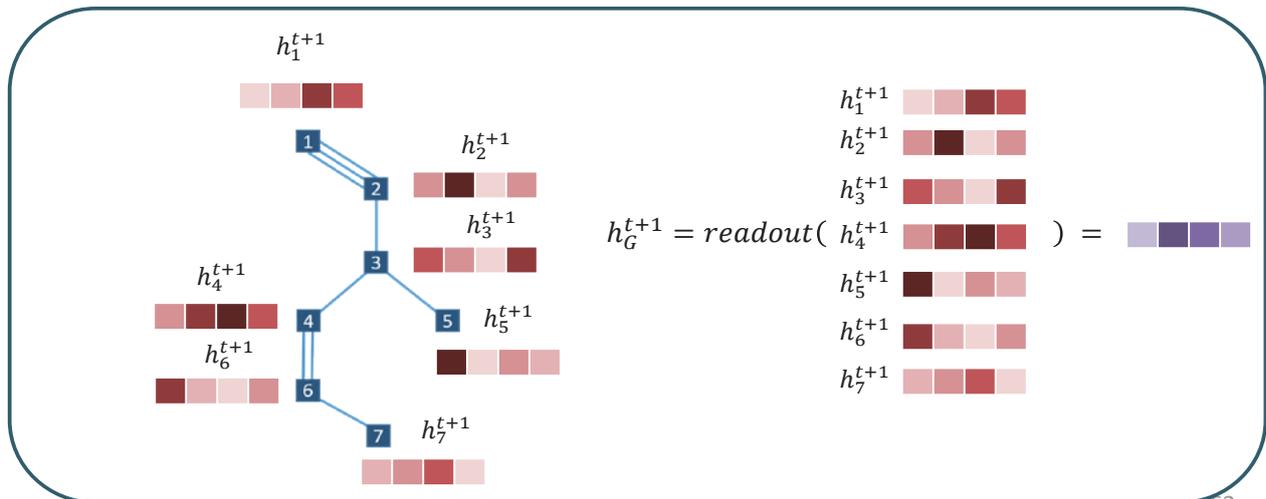
Graph Neural Network Models

- 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

Compound representation (실습코드)

https://colab.research.google.com/drive/1smQsJSVlTKsI7difhyzLco-2eG96mCCL?usp=sharing

---

Lecture 1 - END.

KSBi-BIML 2024

Drug discovery and development - Pharmacogenomics and beyond

Hojung Nam, Ph.D.

Professor
School of Electrical Engineering and Computer Science (EECS)
Gwangju Institute of Science and Technology (GIST)
Contact: hjnam@gist.ac.kr

Contents

- **Lecture 1**
  - Introduction to pharmacogenomics
    - Drug discovery and development
  - Key data sources
  - Representations of proteins, chemicals

- **Lecture 2**
  - 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

Ultra-rapid Metabolizer	Normal Metabolizer	Poor Metabolizer
↓	↓	↓
Under-dosed: Lack of efficacy	Expected response	Over-dosed: Adverse drug reactions
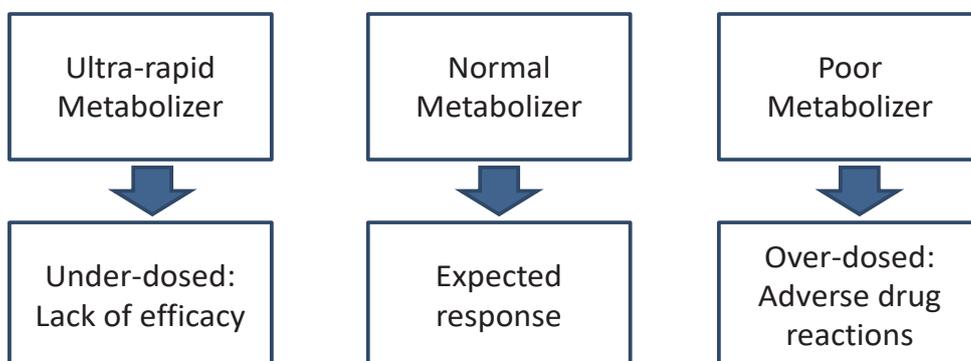
Cytochrome P450 enzymes

- 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

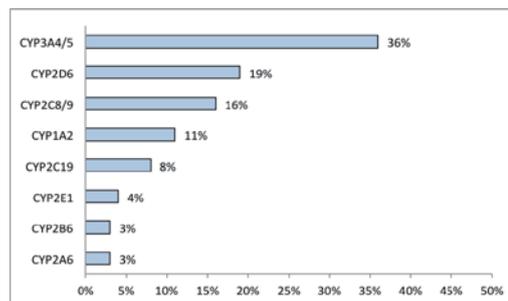| Family	Function	Members	Genes	pseudogenes
CYP1	drug and steroid (especially estrogen) metabolism, benzo[a]pyrene toxification (forming (+)-benzo[a]pyrene-7,8-dihydrodiol-9,10-epoxide)	3 subfamilies, 3 genes, 1 pseudogene	CYP1A1, CYP1A2, CYP1B1	CYP1D1P
CYP2	drug and steroid metabolism	13 subfamilies, 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
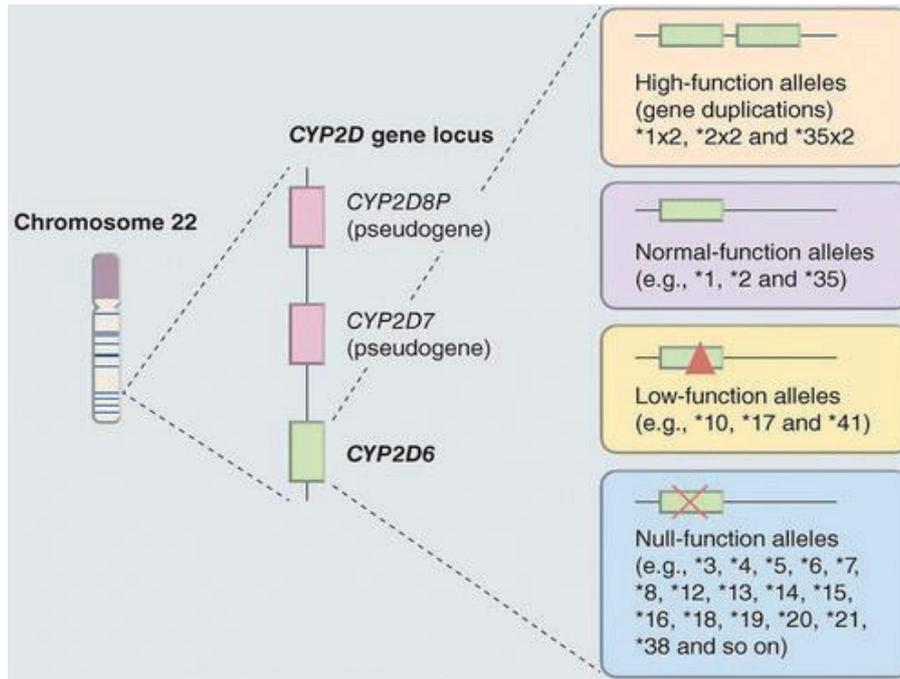
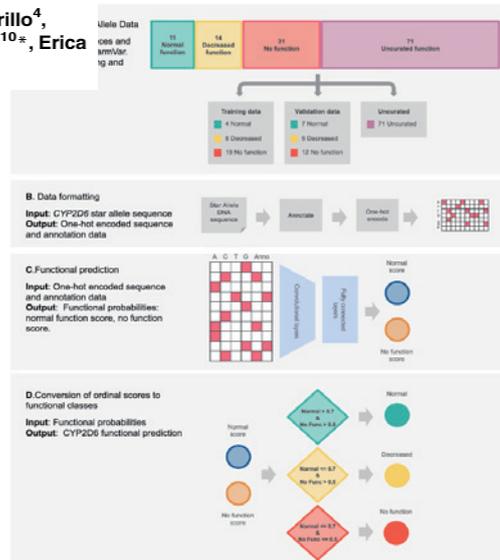
Related study:
prediction of CYP2D6 haplotype function

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]*

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
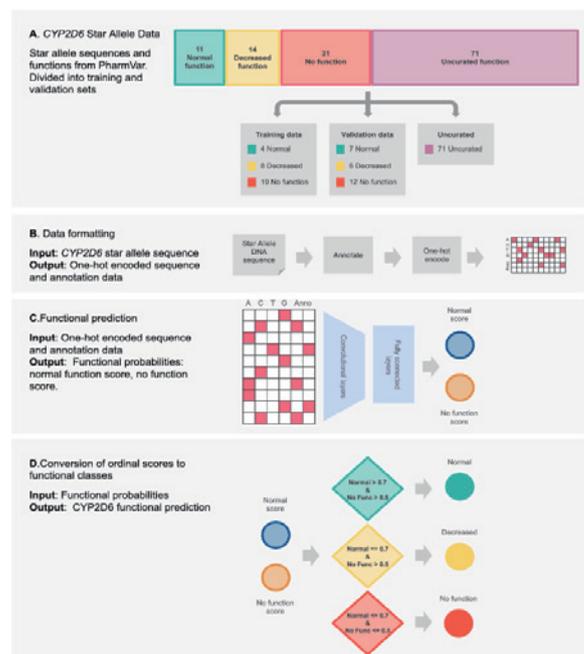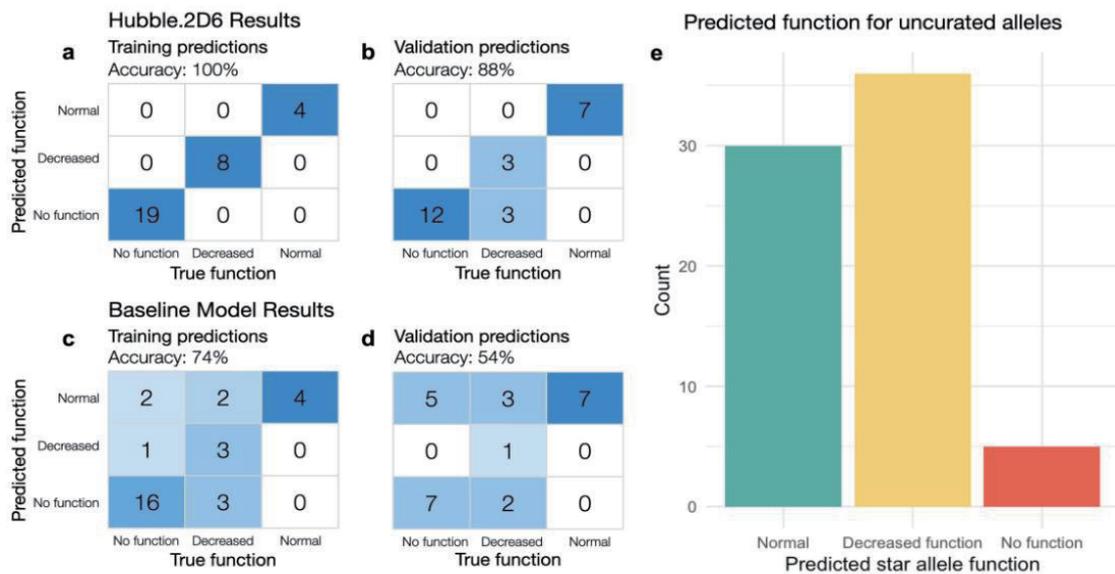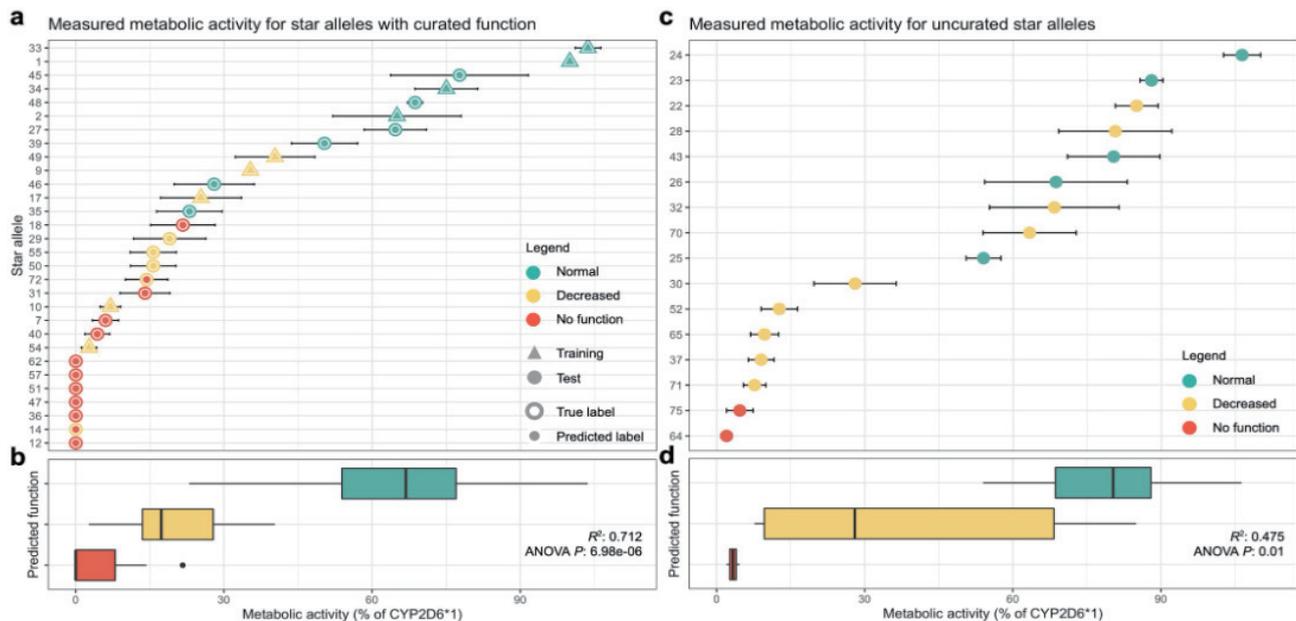

Fig 3. Prediction of star allele function with *in vitro* data. The figures summarize the distribution of metabolic activity measured *in vitro* for star alleles whose function was predicted by Hubble. The distribution of functional activity is shown in (a) and (b) for star alleles with CPIC-assigned 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

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[5] & 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.
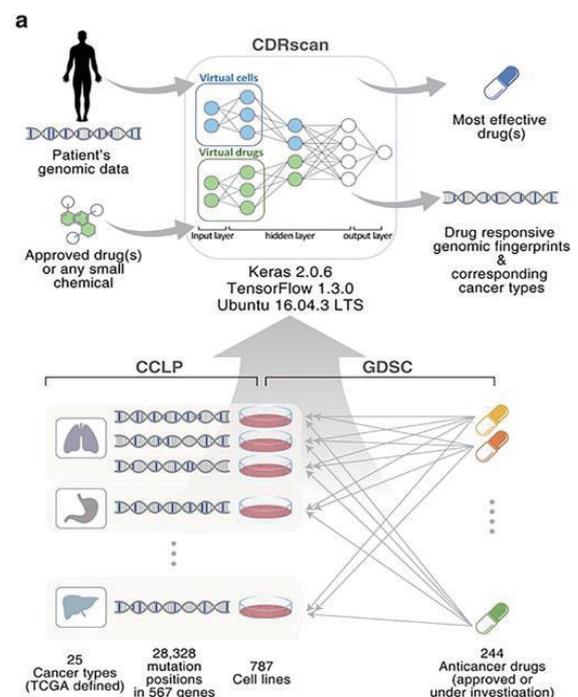
Related study:
prediction of cancer cell sensitivity to drugs

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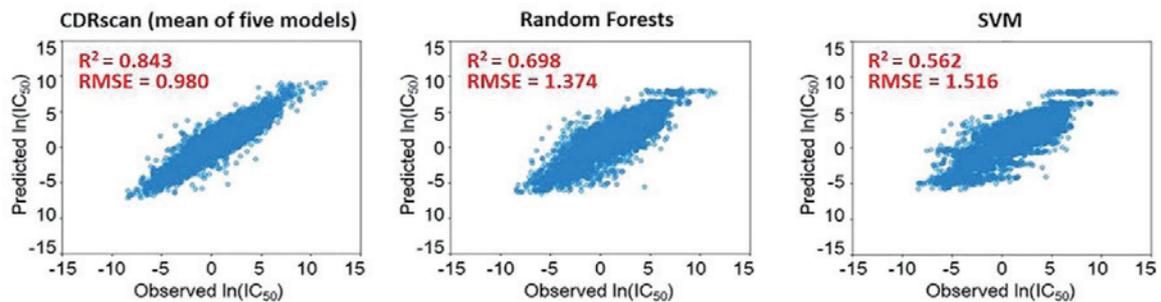
- multi-fold cross validation (five-fold with each fold)

DrugCell

Cancer Cell

Volume 38, Issue 5, 9 November 2020, Pages 672-684.e6

CellPress

Article

Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells

Brent M. Kuenzi [1, 5], Jisoo Park [1, 5], Samson H. Fong [1, 2], Kyle S. Sanchez [1], John Lee [1], Jason F. Kreisberg [1], Jianzhu Ma [4], Trey Ideker [1, 2, 3, 6]

Show more ⌄

⊰ Share   ⟢⟢ Cite

DrugCell

A

Genotype Drug

Embedding of genotype via cell model (VNN) Embedding of chemical structure (ANN)

In silico treatment of cell with drug

DrugCell

Response of cell to drug

- CTRP (Cancer Therapeutics Response Portal) v2 + GDSC
- 509,204 cell line-drug pairs, covering 684 drugs and 1,235 cell lines.

B

Binary mutations 2,086 subsystems 6 neurons / subsystem 5 layers
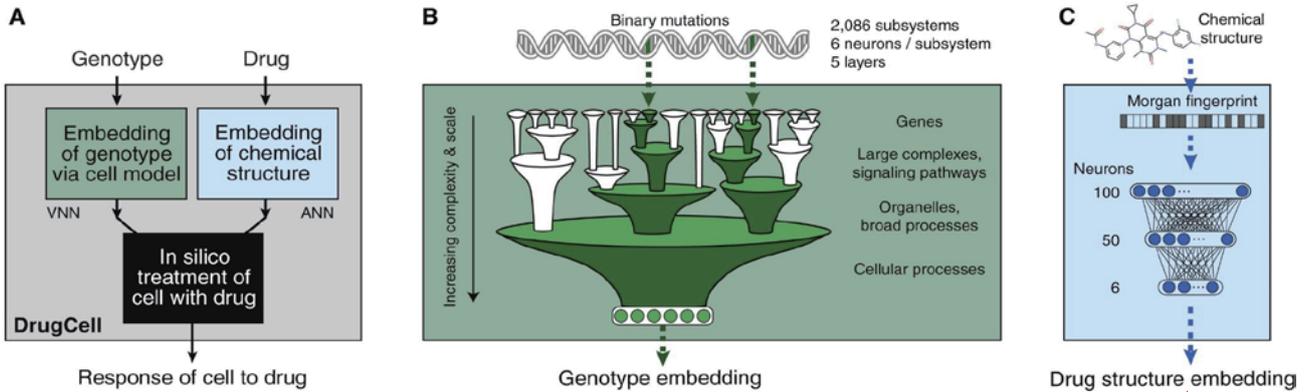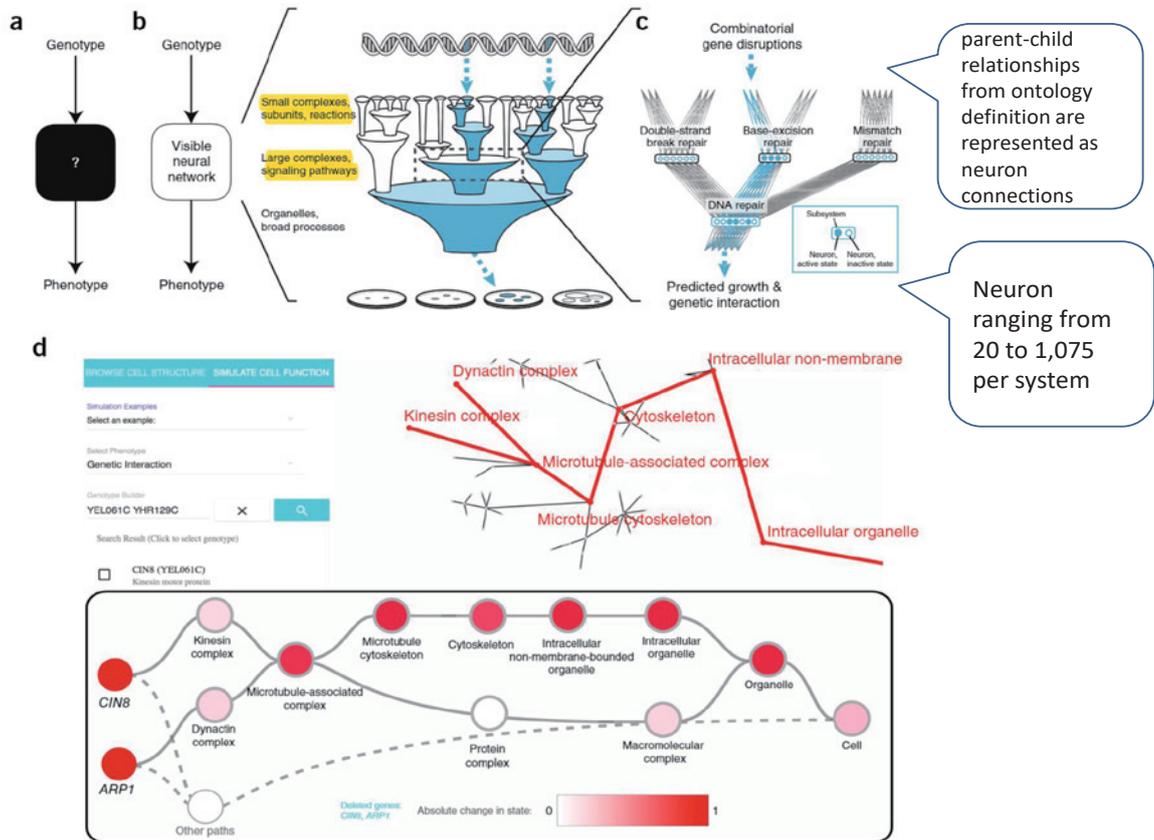
Increasing complexity & scale

Genes
Large complexes, signaling pathways
Organelles, broad processes
Cellular processes

Genotype embedding

- CCLE
- Binary vector of top 15% most frequently mutated genes -> total 3,008 genes

C

Chemical structure

Morgan fingerprint

Neurons
100
50
6

Drug structure embedding

Morgan Fingerprints (nbits = 2048, radius = 2)

Figure 1 | Modeling system structure and function with visible learning. (a) A conventional neural network translates input to output as a black box without knowledge of system structure. **(b)** In a visible neural network, input–output translation is based on prior knowledge. In DCell, gene-disruption genotypes (top) are translated to cell-growth predictions (bottom) through a hierarchy of cell subsystems (middle). **(c)** Each subsystem is embedded in the prior structure using multiple neurons per subsystem. **(d)** Screen capture of DCell online service.

parent-child relationships from ontology definition are represented as neuron connections

Neuron ranging from 20 to 1,075 per system

https://dx.doi.org/10.1038/nmeth.4627

DrugCell

A

rho = 0.8

B

p < 0.0001

C
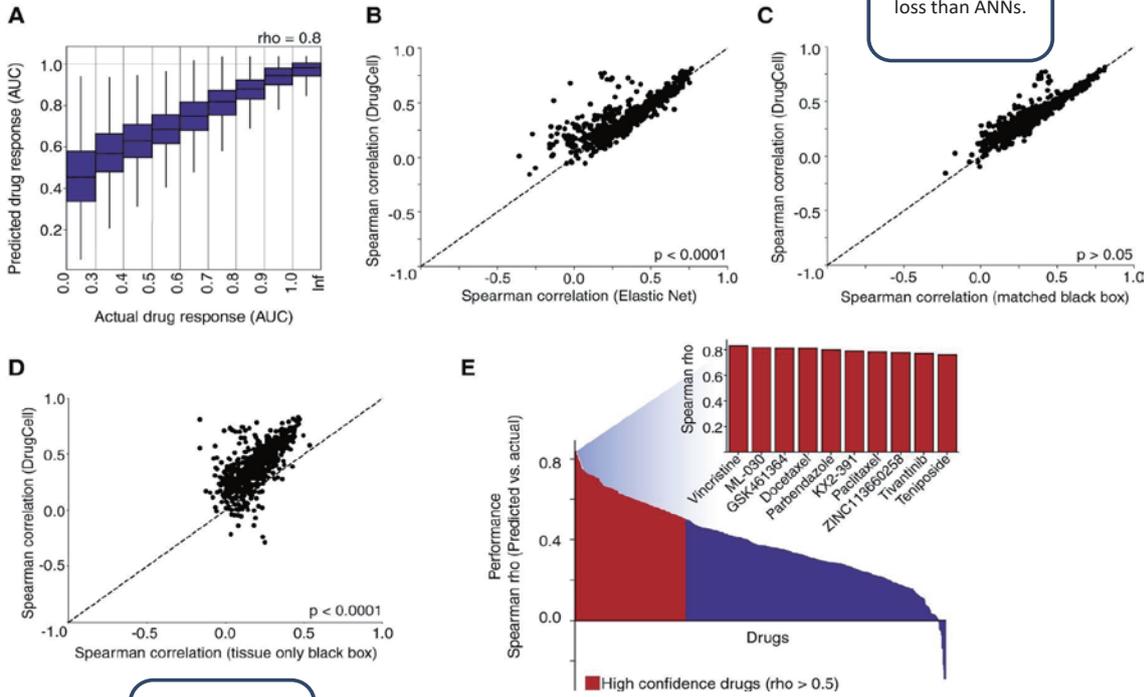
No performance loss than ANNs.

p > 0.05

D

p < 0.0001

VNN really works!

E

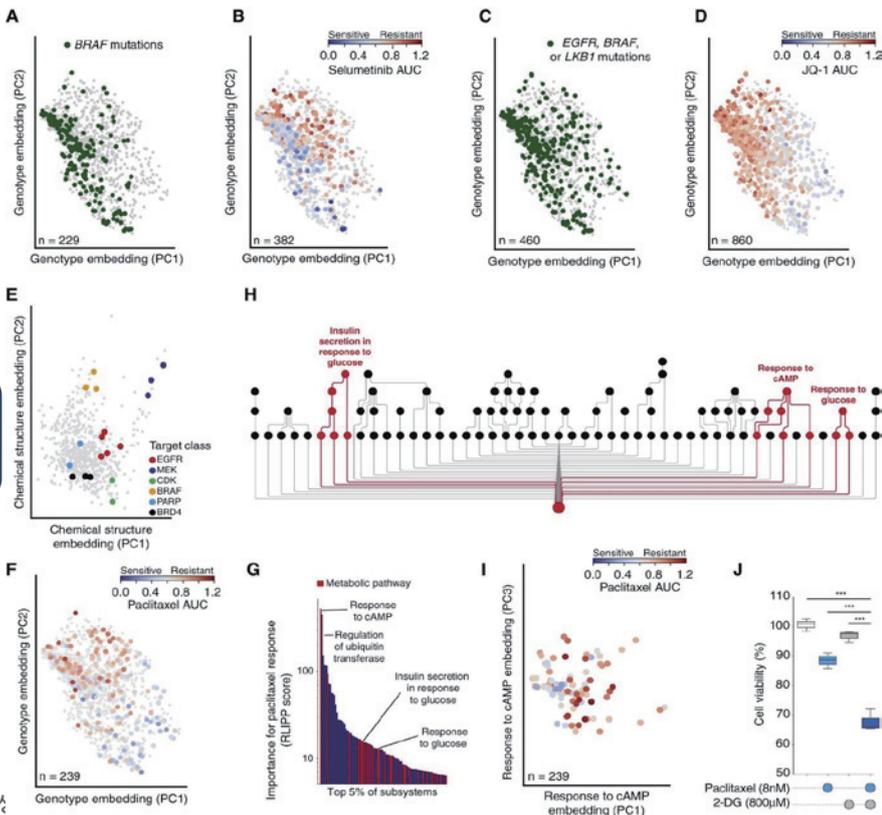Spearman rho

Vincristine, ML-030, GSK461364, Docetaxel, Parbendazole, KX2-391, Paclitaxel, ZINC113660258, Tivantinib, Teniposide

High confidence drugs (rho > 0.5)

SBi 한국생명

A — BRAF mutations — n = 229

B — Sensitive Resistant — Selumetinib AUC — n = 382

C — EGFR, BRAF, or LKB1 mutations — n = 460

D — Sensitive Resistant — JQ-1 AUC — n = 860

E — Chemical structure embedding (PC2) / Chemical structure embedding (PC1) — Target class: EGFR, MEK, CDK, BRAF, PARP, BRD4

Clustered by target class

H — Insulin secretion in response to glucose — Response to cAMP — Response to glucose

'Paclitaxel' RLIPP → Stabilize microtubules (미세소관)

F — Sensitive Resistant — Paclitaxel AUC — Genotype embedding (PC2) / Genotype embedding (PC1) — n = 239

G — Metabolic pathway — Importance for paclitaxel response (RLIPP score) — Response to cAMP, Regulation of ubiquitin transferase, Insulin secretion in response to glucose, Response to glucose — Top 5% of subsystems

I — Sensitive Resistant — Paclitaxel AUC — Response to cAMP embedding (PC3) / Response to cAMP embedding (PC1) — n = 239

J — Cell viability (%) — Paclitaxel (8nM), 2-DG (800μM)

Combination with Glycolysis inhibitor → Significantly Effective!

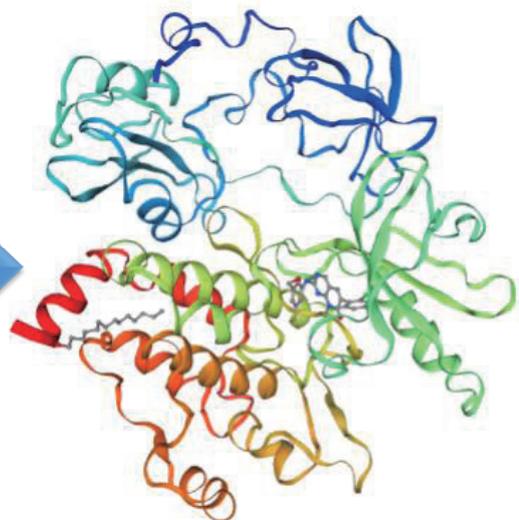SBi 한국생

- 43 -

PROTEIN SEQUENCE AND DRUG INTERACTIONS

Prediction of drug-target interaction


Imatinib

Druggable?
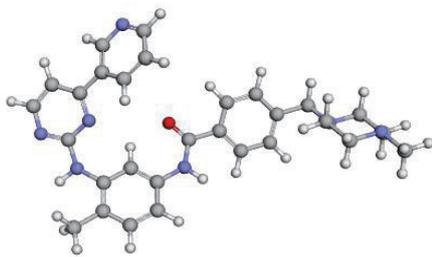
BCR/ABL fusion protein

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.
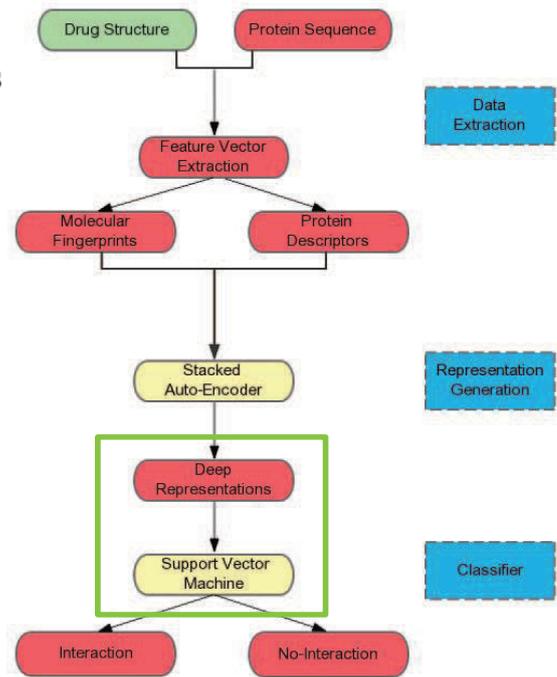
Multi-scale features deep representations inferring interactions (MFDR)
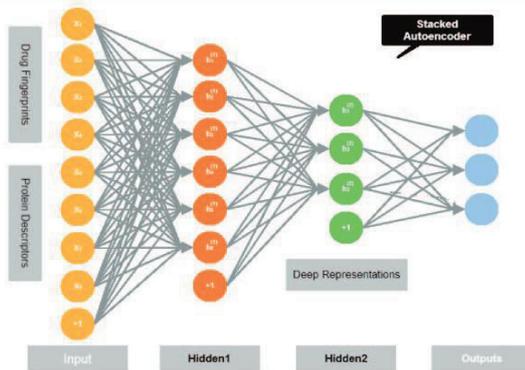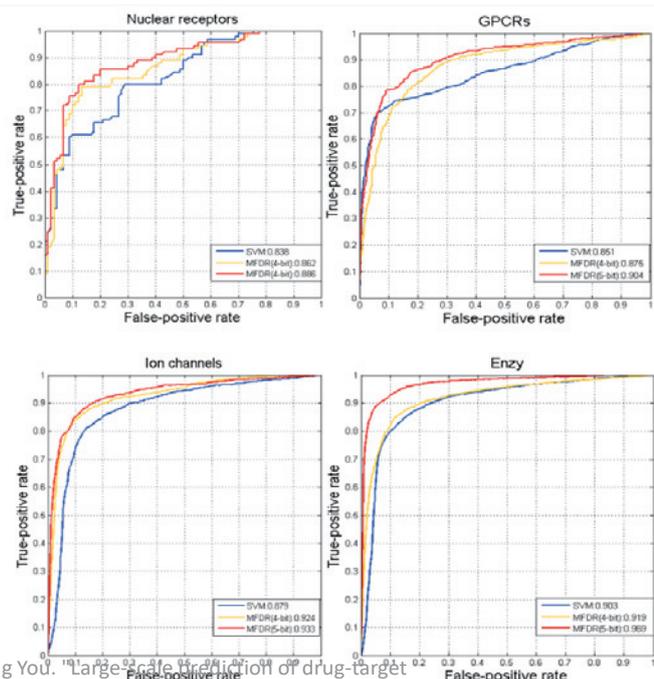

DTI prediction using protein descriptors

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

5fold cross-validation

DRUG-TARGET DATA STATISTIC

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

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

OXFORD

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**
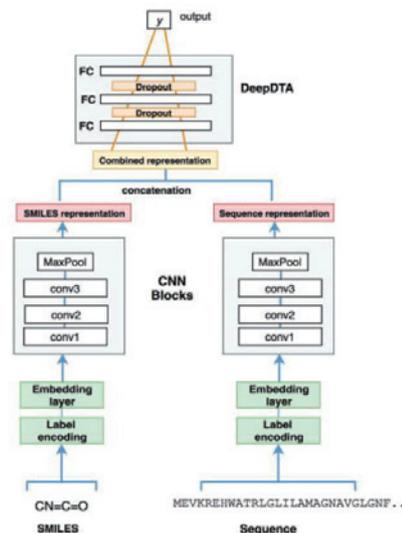 - first used CNN to learn representations of proteins

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

SBi 한국생명정보학회
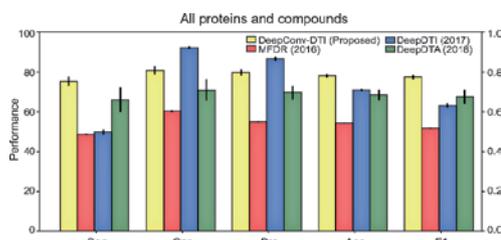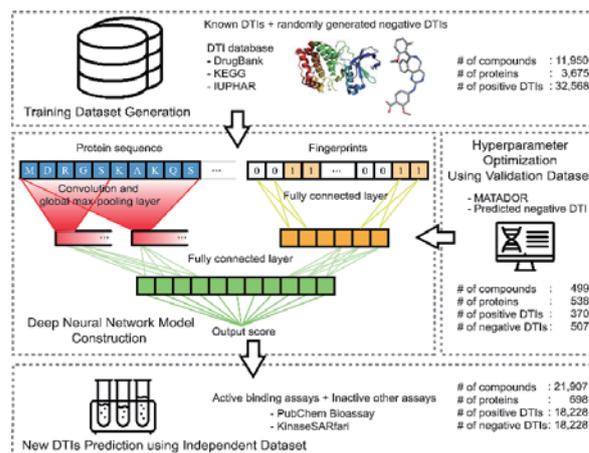Korean Society for Bioinformatics

---

DTI prediction using protein sequence

RESEARCH ARTICLE

DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences
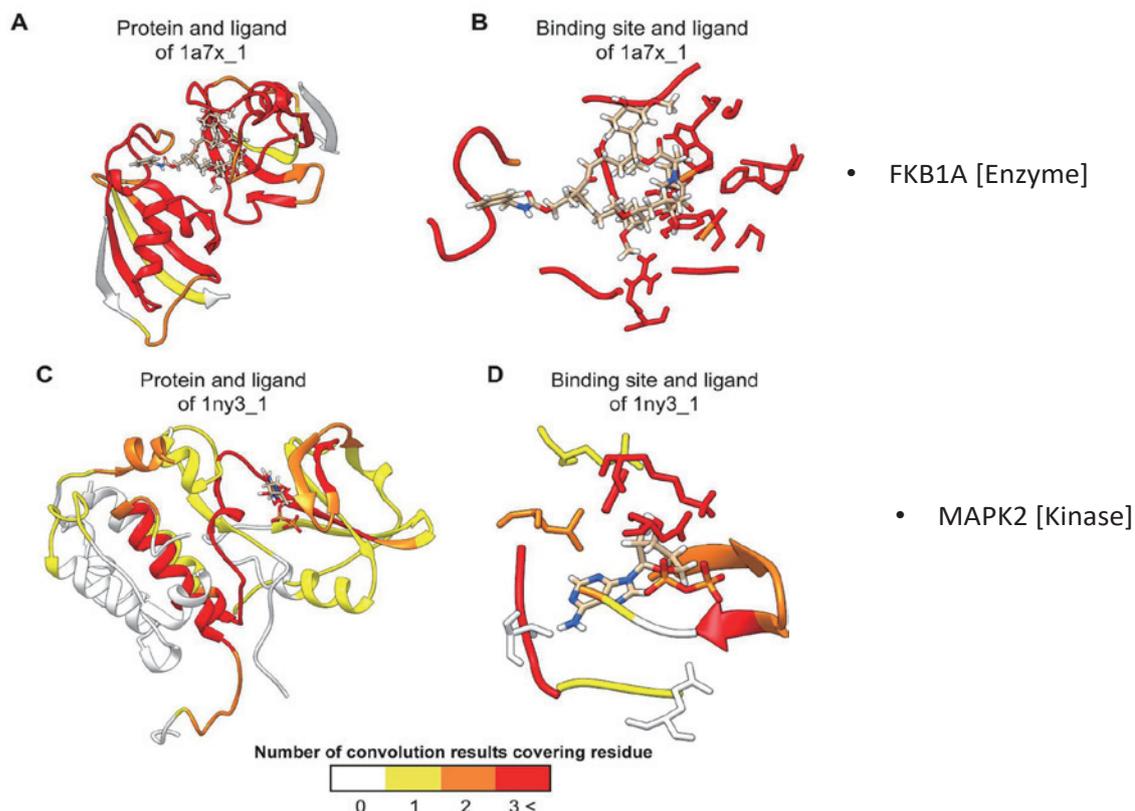
Ingoo Lee[ ], Jongsoo Keum[ ], Hojung Nam[*]

- **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

SBi 한국생명정보학회
Korean Society for Bioinformatics

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

A Protein and ligand of 1a7x_1

B Binding site and ligand of 1a7x_1

- FKB1A [Enzyme]

C Protein and ligand of 1ny3_1

D Binding site and ligand of 1ny3_1

- MAPK2 [Kinase]

Number of convolution results covering residue

0 1 2 3 <

HoTS: Highlights on Target Sequence and Prediction of Drug-Target interaction (2022)
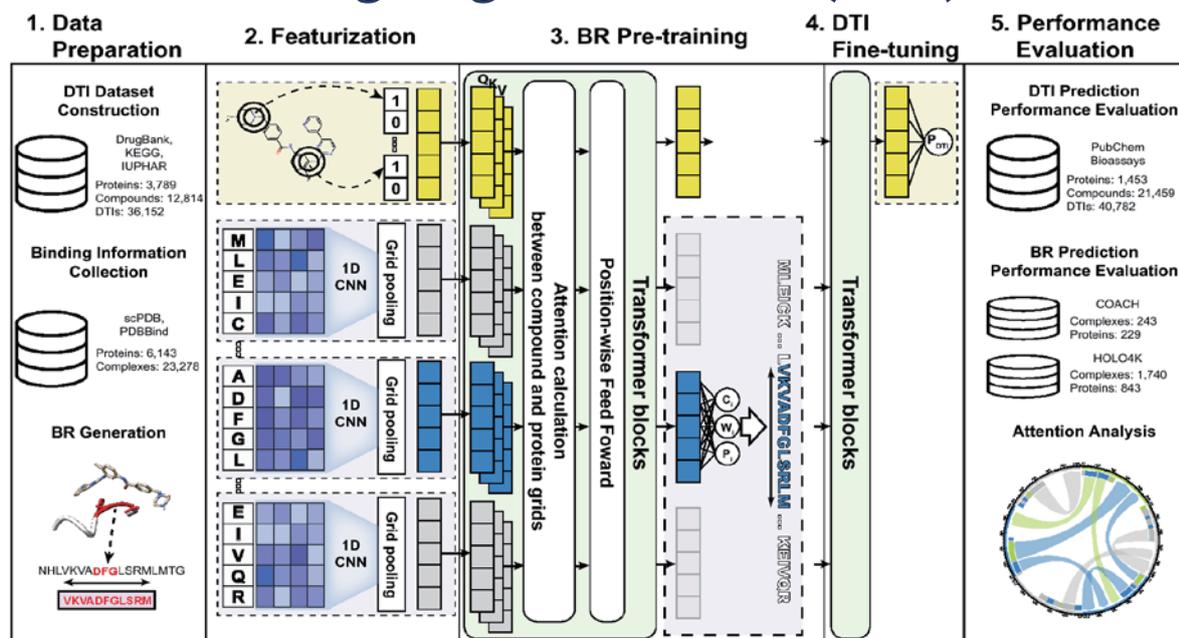
- **Prediction of binding regions** for DTIs
- Showed **better performance** in **hit identification**
- An **interpretable** deep Learning model

Ingoo Lee, Hojung Nam*,

Performance improvement in DTI prediction

- DTI test dataset 1: proteins collected from the DTI Database as general druggable targets
  - evaluate DTI prediction performance for druggable targets whose BRs have not been trained.

- DTI test dataset 2: DTIs for proteins whose SCOPe family was the same as the BR training dataset
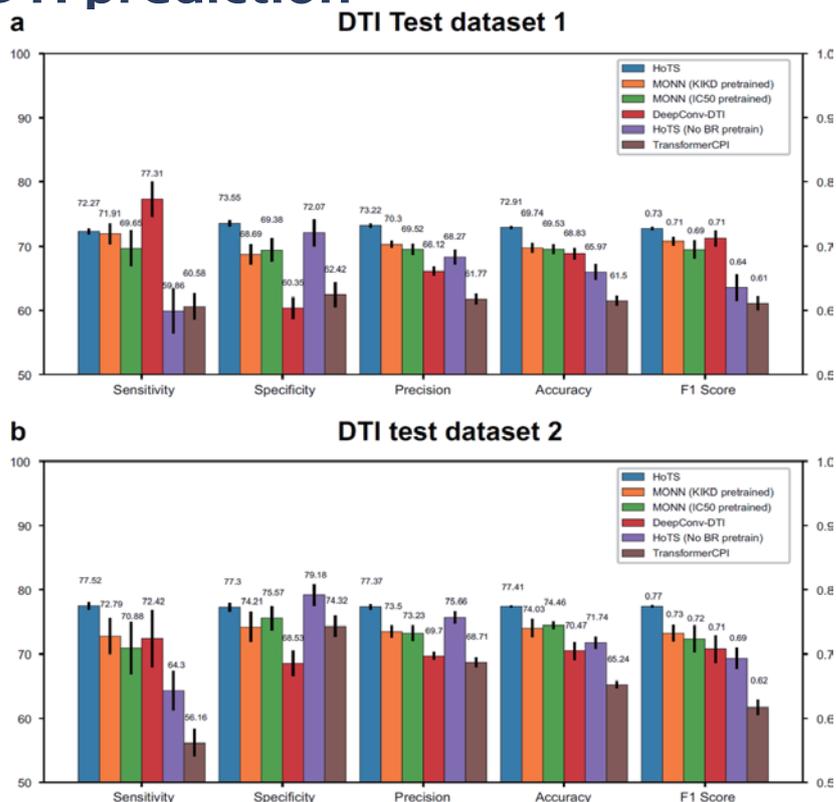 - evaluate DTI prediction performance for proteins with the same or similar interacting motifs.

Ingoo Lee, Hojung Nam*, "Sequence-based prediction of binding regions and drug-target interactions", Journal of cheminformatics 2022

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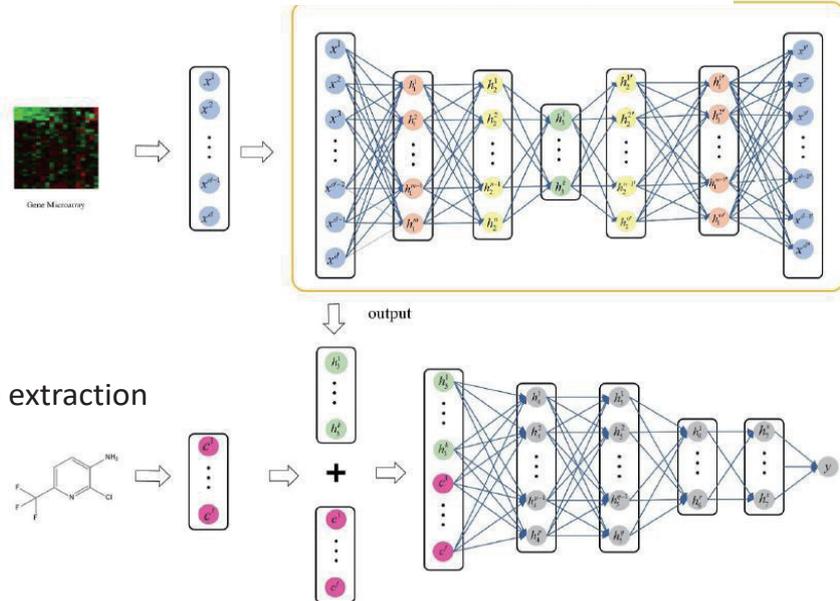
Li, Min, et al. "DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines." *IEEE/ACM transactions on computational biology and bioinformatics* (2019).

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
	R²	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
	R²	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
	R²	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:
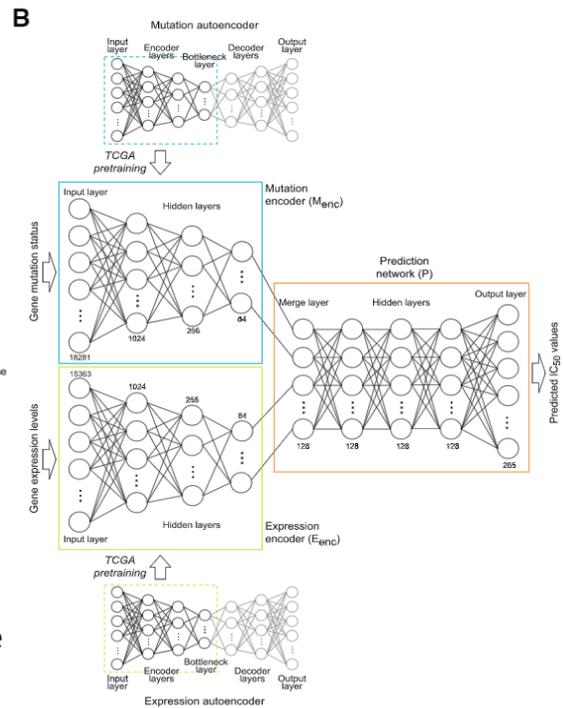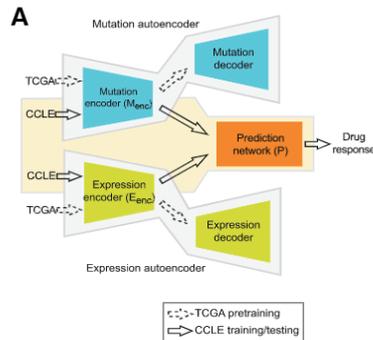# prediction of cancer cell sensitivity to drugs

- 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.

SBi 한국생명정보학회
Korean Society for Bioinformatics

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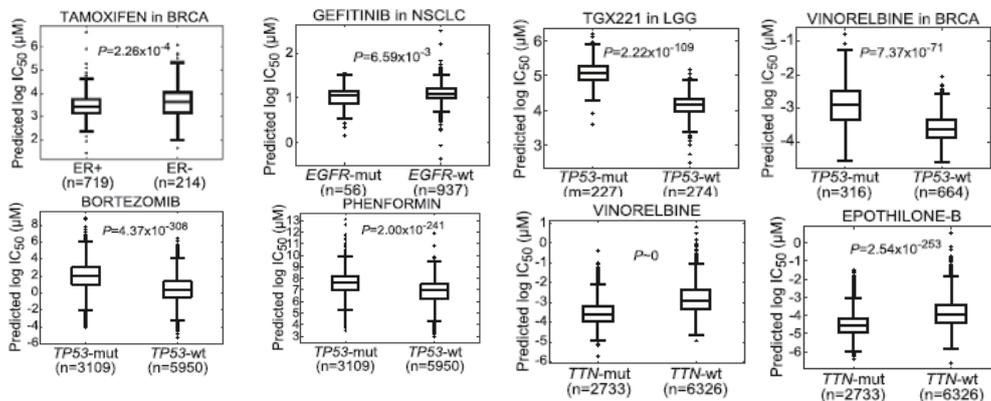
- Samples with mutation showed significantly different result compared to non-mutated samples

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.

SBi 한국생명정보학회
Korean Society for Bioinformatics

Related study:
prediction of cancer cell sensitivity to drugs

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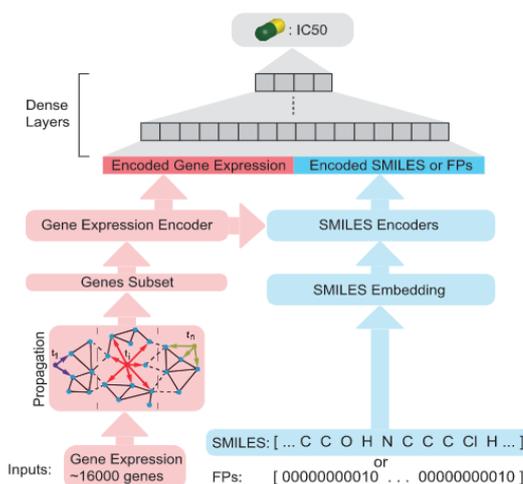
[†]IBM Research, 8803 Zürich, Switzerland
[‡]ETH Zürich, 8092 Zürich, Switzerland
[⊥]University of Zürich, 8006 Zürich, Switzerland
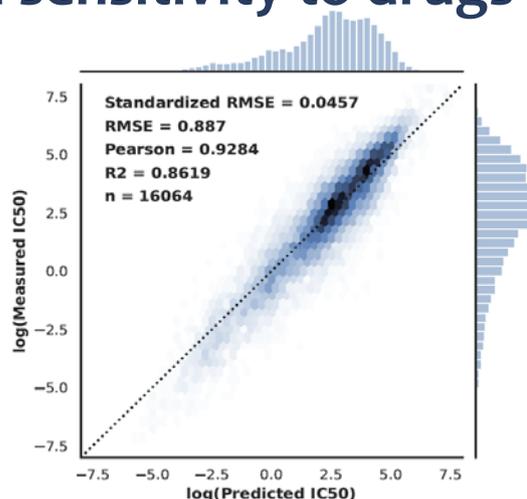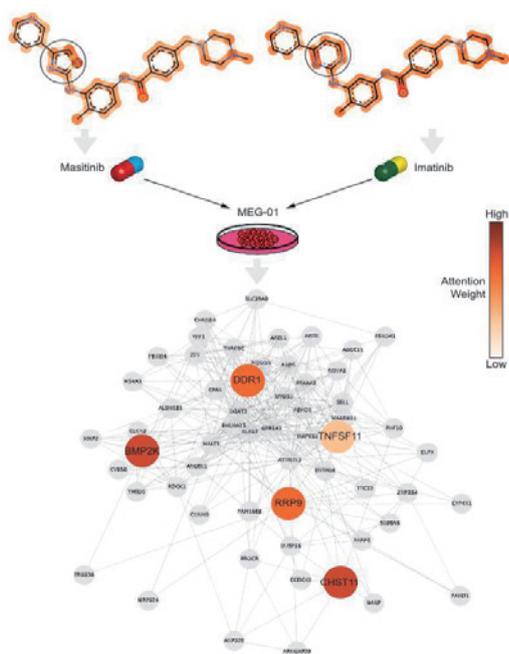[§]RWTH Aachen University, 52056 Aachen, Germany
[∥]Heidelberg University, 69047 Heidelberg, Germany

- Transcriptomic feature
- PPI for feature selection
- SMILES

- Attention based model
  - Interpretable

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.

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**
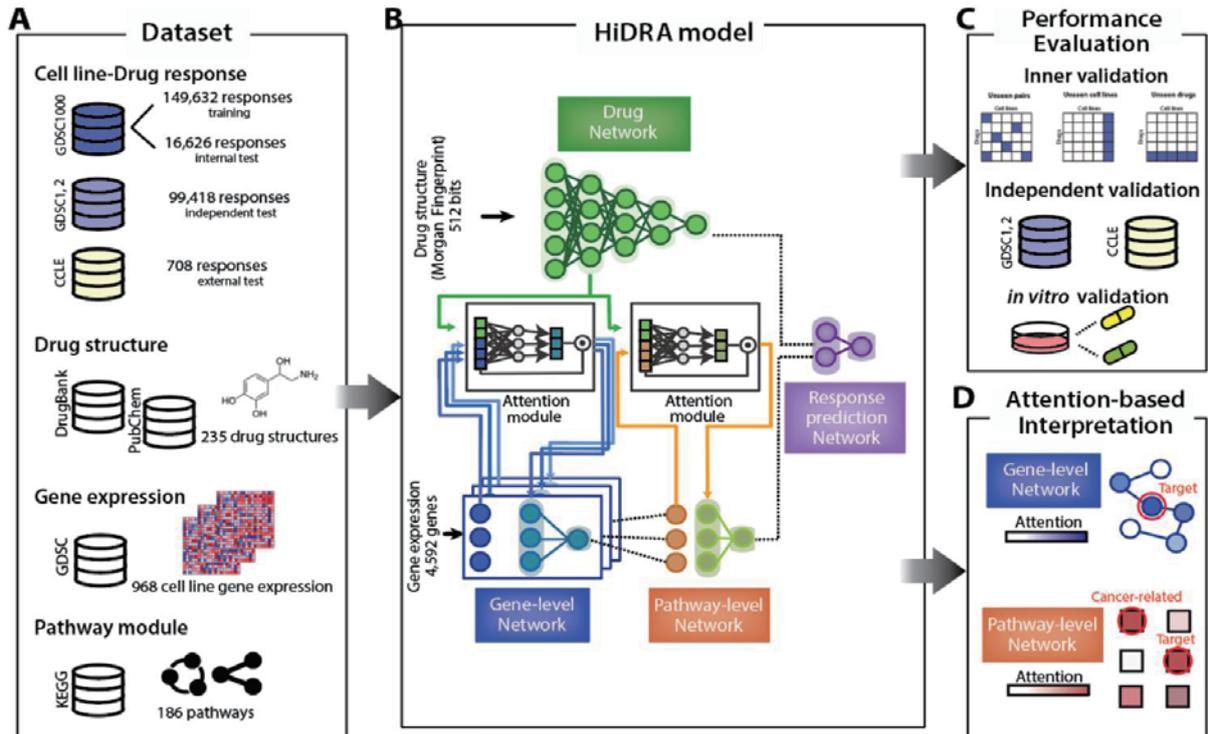
- 51 -

HiDRA

Iljung Jin, Hojung Nam, "HiDRA: Hierarchical Network for Drug Response Prediction with Attention", J. Chem. Inf. Model. 2021, 61, 3858–3867
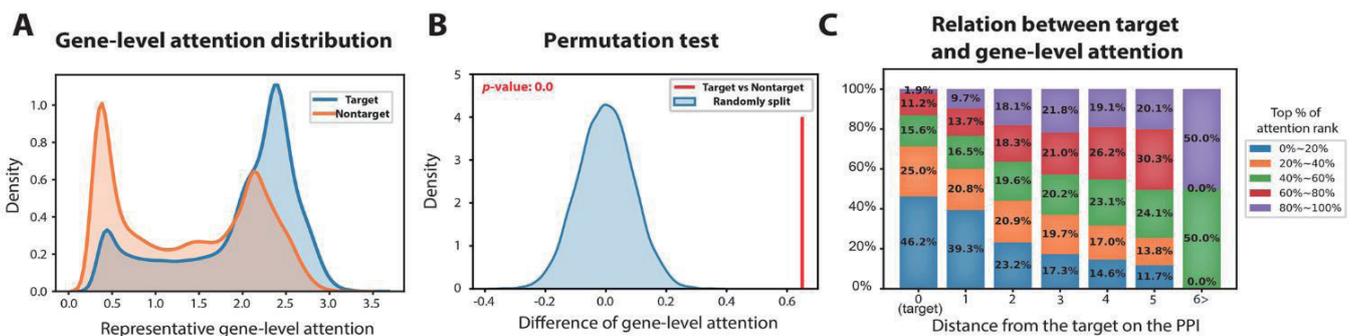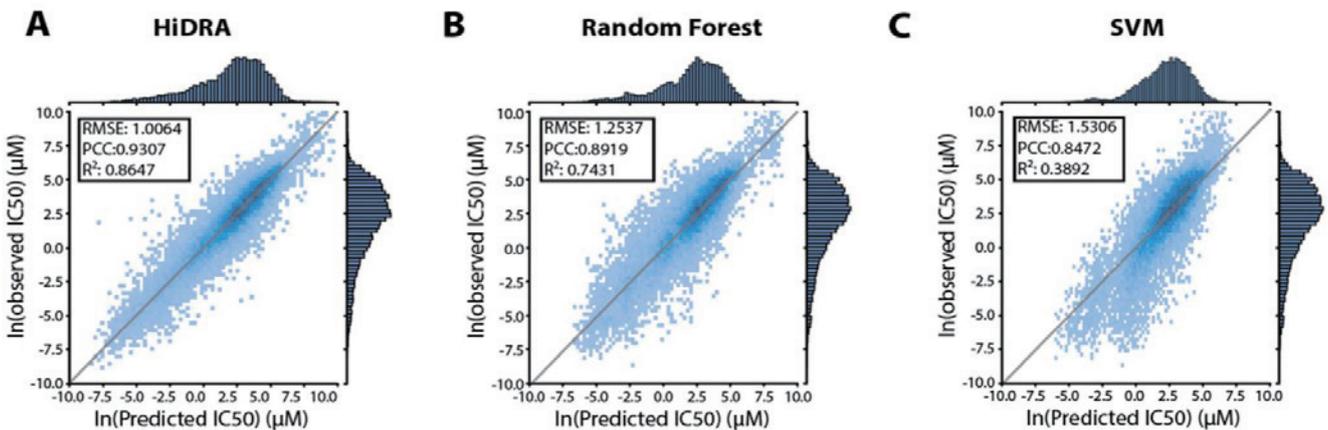
Iljung Jin, Hojung Nam, "HiDRA: Hierarchical Network for Drug Response Prediction with Attention", J. Chem. Inf. Model. 2021, 61, 3858–3867

Contents

- Lecture 1
  - Introduction to pharmacogenomics
    - Drug discovery and development
  - Key data sources
  - Representations of proteins, chemicals

- Lecture 2
  - Studies related to pharmacogenomics based on machine learning

SBi 한국생명정보학회
Korean Society for Bioinformatics

End

SBi 한국생명정보학회
Korean Society for Bioinformatics