KSBi-BIML 2021

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists

생물정보학 & 머쉰러닝 워크샵(온라인)

AI-based Drug Discovery

김선

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

AI-based drug discovery

Sun Kim group
Department of Computer Science and Engineering, Bioinformatics Institute, Seoul National University, Seoul, Korea

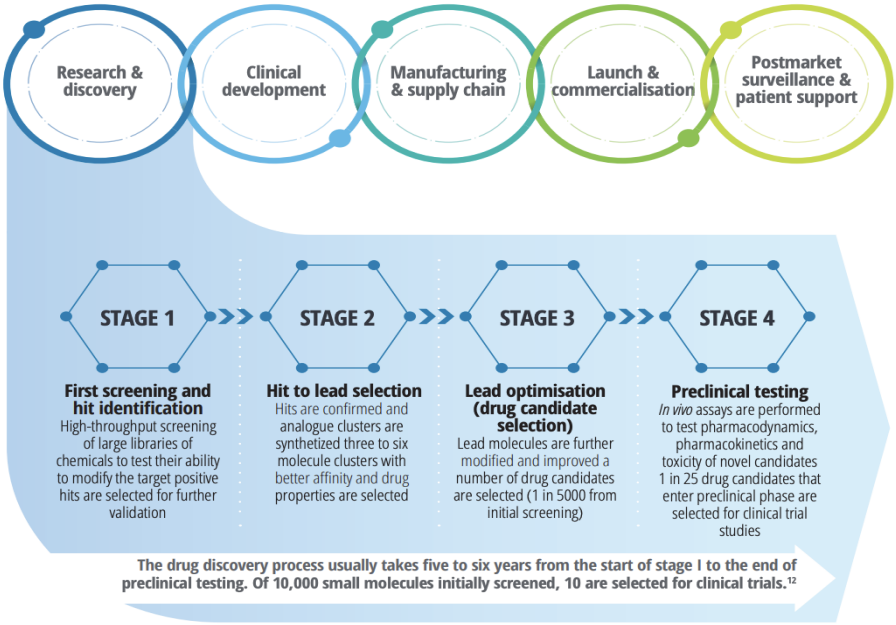
In this tutorial, we will deliver recent developments in AI-based drug discovery. Since drug discovery is a very wide and complicated area of research, we will begin by explaining basic concepts and database resources on small-molecule drug or compound; target of drug; molecular signature before and after drug treatment; and phenotype such as drug sensitivity, toxicity, side effect, LADME (liberation, absorption, distribution, metabolism, and excretion). Then, in Part 2 of this tutorial, we will spent time to explain why AI-based drug discovery has emerged. Traditional drug discovery focused on predicting targets and phenotypes directly. Related research topics have been extensively investigated in the context of valid compound design, pharmacodynamics and pharmacokinetics. However, gap between compounds and phenotypes are big and wide. As molecular profiling techniques from genome and epigenome sequencing have been developed rapidly over the years, a relatively new concept called pharmacogenomics has emerged and has been extensively studied. In fact, information at the molecular level can be a bridge between compounds and phenotypes, which can be an innovative technology for drug discovery. However, computational analysis of data for drug discovery has become much more challenging since traditional concepts, such as valid compound design, pharmacodynamics and pharmacokinetics, already difficult computational problems and adding genomics dimension increases search space dramatically on top of already extremely large search space of the drug discovery problem. Fortunately, recent development of AI, deep learning, and graph mining technologies has begun to shed light on this daunting computational problem. In Part 3, we will introduce some of the representative examples of AI-based drug discovery technologies. A list of examples are: reinforcement learning for de novo molecule design, GAN and autoencoder for compound design, deep learning models for drug activity prediction, junction tree variational auto encoder for generating valid molecules, deep learning and symbolic AI for planning chemical syntheses, mixture representation learning for toxicity prediction, deep learning models for drug target interaction, GAN model for generating compounds from molecular biology data, and deep learning model for pharmacogenomcs study.

Recent Amazing Progress on AI-based Drug Discovery

- We were amazed how rapidly AI-based drug discovery techniques have been developing!

- Today, we present
  - Our view on AI-based drug discovery
  - Introducing recent exemplary successes in this field.

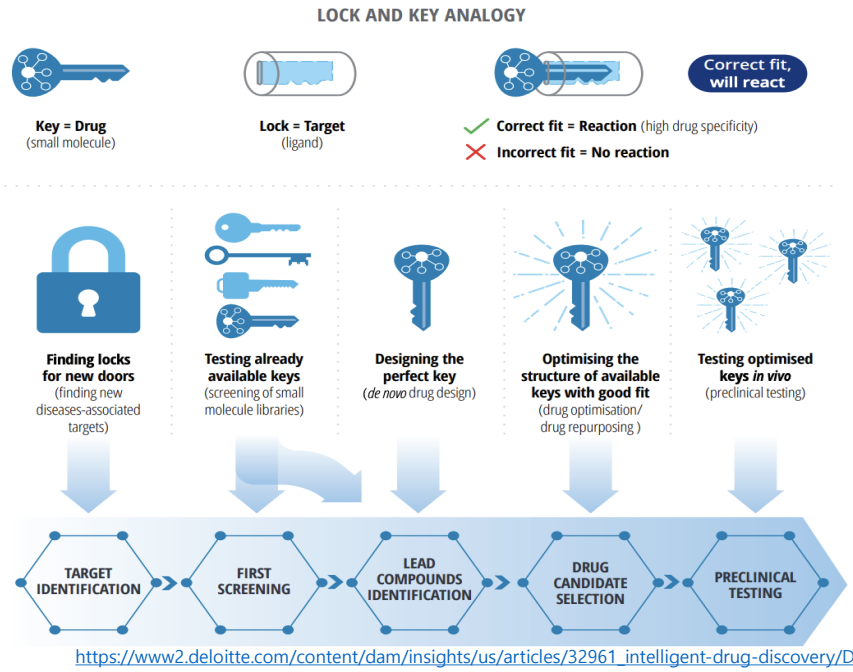
The biopharma value chain for drug discovery



Source: Deloitte analysis.
https://www2.deloitte.com/content/dam/insights/us/articles/32961_intelligent-drug-discovery/DI_Intelligent-Drug-Discovery.pdf

FIGURE 5

Lock and key analogy showing the five main challenges for AI in drug discovery



https://www2.deloitte.com/content/dam/insights/us/articles/32961_intelligent-drug-discovery/DI_Intelligent-Drug-Discovery.pdf

Source: Deloitte analysis.

Drug Discovery

- Drug discovery is resource intensive, and involves typical time-lines of 10–20 years and costs that range from US$0.5 billion to US$2.6 billion. Artificial intelligence promises to accelerate this process and reduce costs by facilitating the rapid identification of compounds.
- https://www.nature.com/articles/s41587-019-0224-x
- 오늘 강연은 small molecule drug discovery 관련 내용만 입니다.
- 약물을 성공적으로 만드는 것은 많은 요소를 고려해야 합니다.

A Computer Scientist's View

- **Drug discovery = exploring huge search space**
- Single tools cannot solve this complex problem of dealing with daunting search space
  - Compound space
  - Target gene space
  - Genetic space
  - Phenotype space
  - Combination of all above

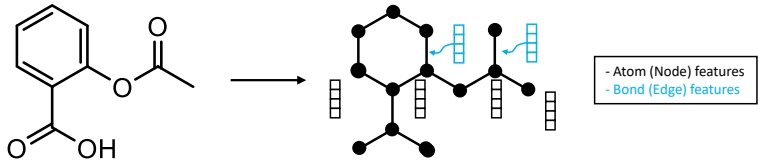
Exploring Compound Space (and also reaction space)

Chemical Space

- Sequence — SMILES, SMARTS, SELFIES

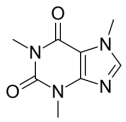
  CC(=O)OC1=CC=CC=C1C(=O)O

- Graph — 2D/3D Graph structure



- Other — Substructures or fingerprints, Image, etc.

Linear string representation: Chemical Sequence-based Descriptors



Caffeine

1. **SMILES** (**S**implified **M**olecular-**I**nput **L**ine **E**ntry **S**ystem) — *JCIM*, 1988

   a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings

   ex) CN1C=NC2=C1C(=O)N(C(=O)N2C)C

2. **SMARTS** (**SM**ILES **AR**bitrary **T**arget **S**pecification) — *Daylight Chem Info Systems*

   a language for specifying substructural patterns in molecules

   ex) [#6]-[#7]1:[#6]:[#7]:[#6]2:[#6]:1:[#6](=[#8]):[#7](:[#6](=[#8]):[#7]:2-[#6])-[#6]

3. **SELFIES** (**SELF**-referenc**I**ng **E**mbedded **S**trings) — *NIPS*, 2019

   ex) [C][N][C][=N][C][=C][Ring1][Ring2][C][Branch1_3][epsilon][=O][N][Branch1_3][Branch2_2][C][Branch1_3][epsilon][=O][N][Ring1][Branch1_3][C][C]

Two major issues with compound

- Search space
  - A chemical space often referred to in cheminformatics is that of potential pharmacologically active molecules. Its size is estimated to be in the order of 10^{60} molecules.
  - https://en.wikipedia.org/wiki/Chemical_space
- Synthesizability?
  - Is it possible to synthesize a given compound?
  - What is the best planning for synthesizing the compound?
- These two issues will be explored with two recent deep learning papers shortly.

Exploring Target (gene) Space

Proteins as Drug Targets

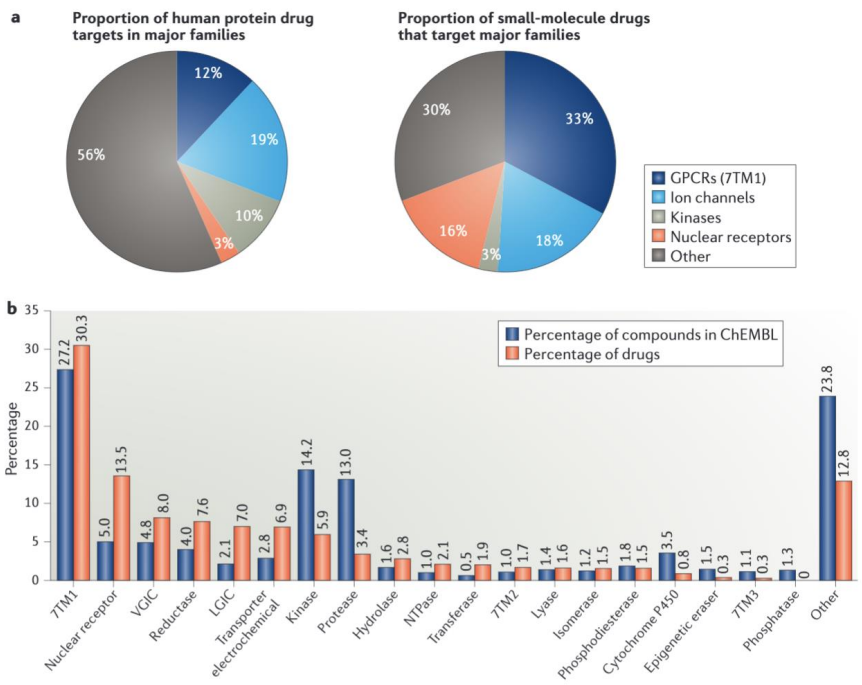


Figure 1 | **Major protein families as drug targets. a** | Distribution of human drug targets by gene family (left) and

A comprehensive map of molecular drug targets, *Nature Reviews Drug Discovery*. 2017

Target Protein Space

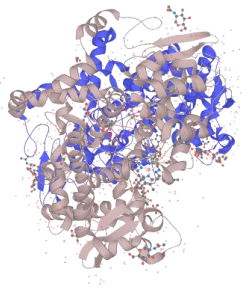
- Sequence — <u>Amino Acid Sequence</u>

  MLARALLLCA VLALSHTANP…

  Sequential graph (K-mer)

- Graph — <u>3D structure, spatial graph</u>

- Other — <u>Domain, Image, etc.</u>



Major issues

- Understanding protein sequence and family hierarchy
  - GPCR proteins:
    - Deep learning based alignment-free method for protein family modeling and prediction. *Bioinformatics/ISMB*. 2018
    - Deep Hierarchical Embedding for Simultaneous Modeling of GPCR Proteins in a Unified Metric Space. In review
- Structure of proteins
  - AlphaFold/AlphaFold2
    - Improved protein structure prediction using potentials from deep learning. *Nature* 2020

Exploring Genetic Space

Major tasks and challenges

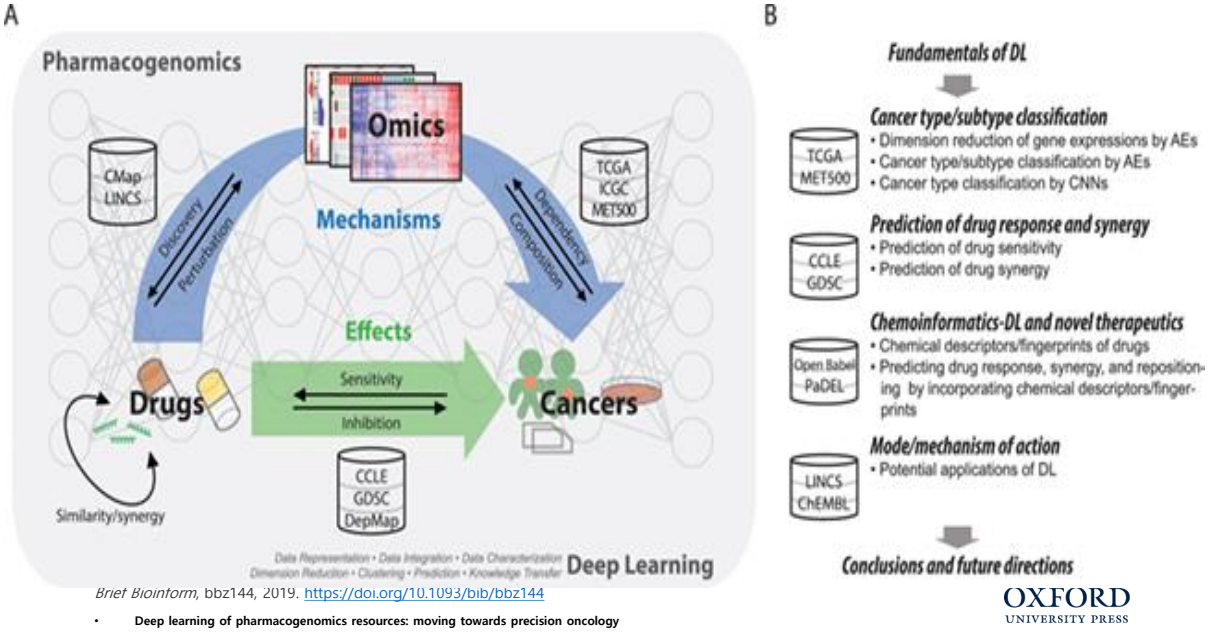
- Genome (DNA)-based drug efficacy and sensitivity prediction
- Synthetic lethality
- Targeted cancer therapy

- Transcriptome (gene expression) based drug effect prediction

- Multi-omics based drug study

- The number of dimensions is huge.
  - 20,000 + tens of millions

Figure 1 Pharmacogenomics in cancer and overview of this review.
(A) Cancer pharmacogenomics and DL. Cancer ...

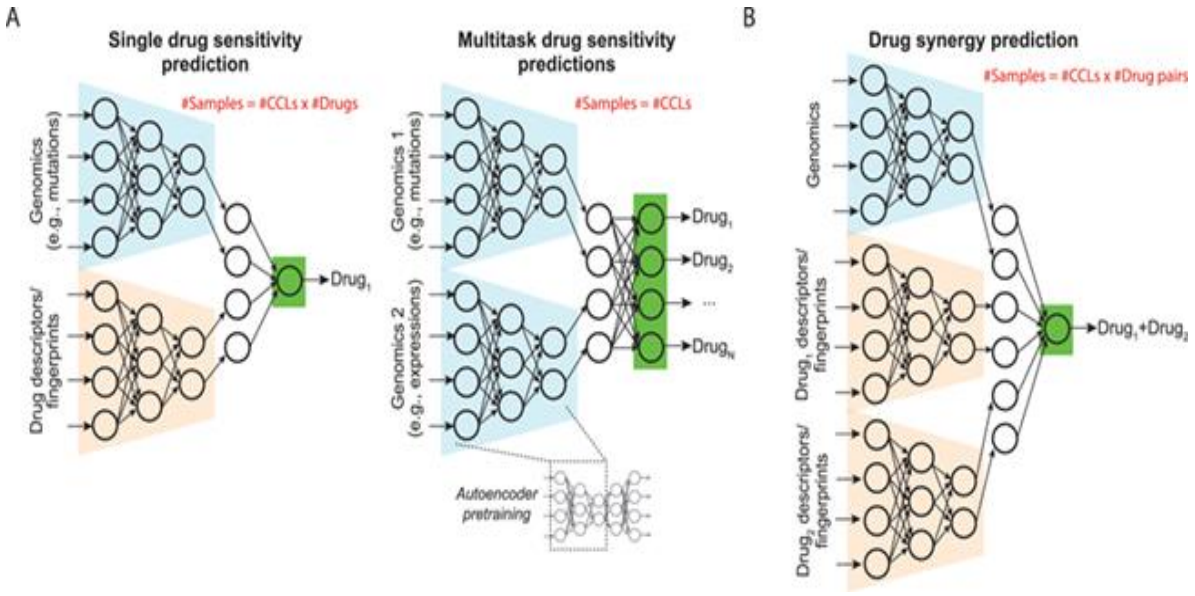


Brief Bioinform, bbz144, 2019. https://doi.org/10.1093/bib/bbz144

- **Deep learning of pharmacogenomics resources: moving towards precision oncology**

OXFORD UNIVERSITY PRESS

Figure 4 DL models for predicting drug response and synergy. (A) Models for predicting drug sensitivity of a single ...



Brief Bioinform, bbz144, https://doi.org/10.1093/bib/bbz144

Databases (genome level)

- **PharmGKB** : pharmacogenomics resource sponsored by NIH
  - Collects information on human genetic variation and drug responses.
  - **'Clinical evidence-centered'** Gene variant data
  - **Annotation data** downloadable
  - Contains non-CYP450 enzyme data

- **PharmVar** : catalogue on allelic variation of ADME genes
  - Co-working with PharmGKB
  - Contains more **'sequence-centered'** variation data
  - Sequence data downloadable

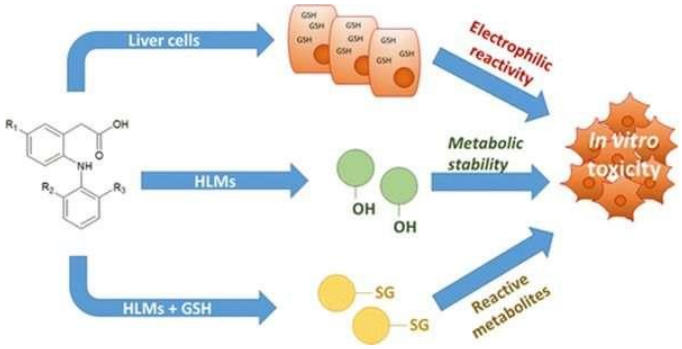
Databases (multi-omics)

- Multi-omics data before drug treatment with drug response information (IC50 or AUC) :
  - GDSC, CCLE, NCI-60
- Time-series gene expression data after drug treatment :
  - NCI TPW, NCI-DREAM

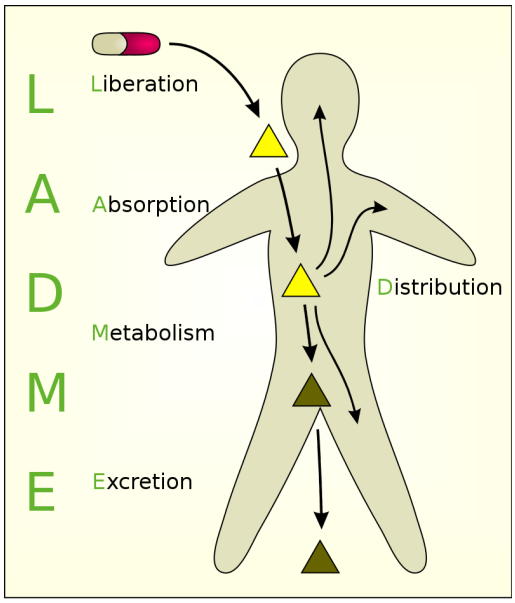
Exploring Phenotype Space

Phenotypes

- ADME
- Toxicity



images.google

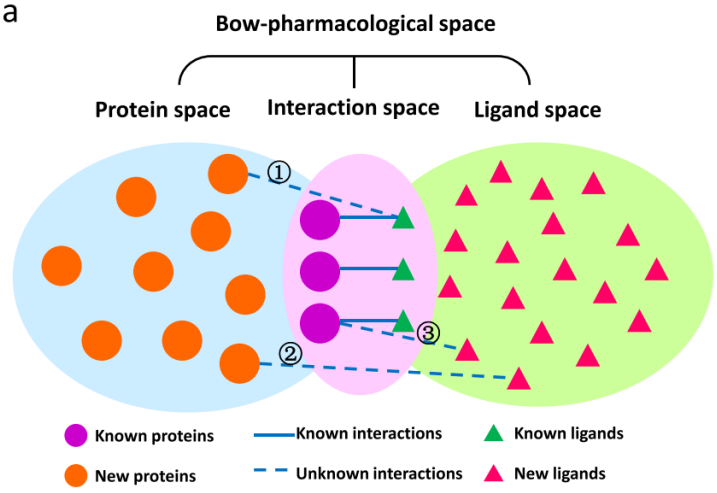


images.google

Exploring Compound Space and Target Gene Space

Concept of ‘space’ – search space

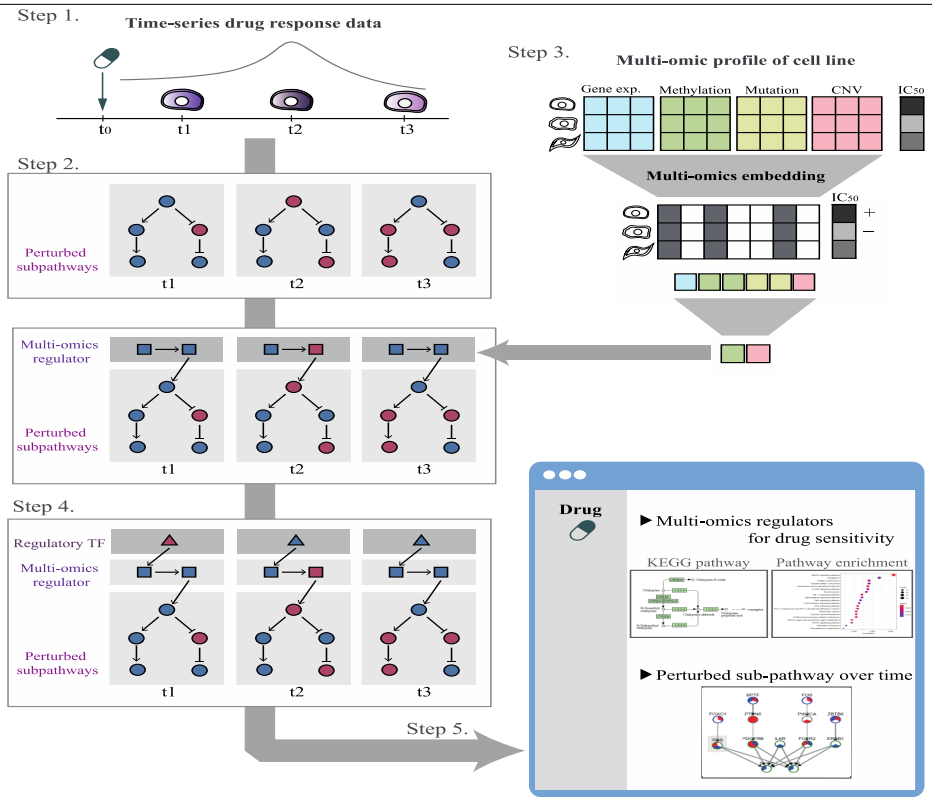
a



Li, Li, et al. "Predicting protein-ligand interactions based on bow-pharmacological space and Bayesian additive regression trees." *Scientific reports* 9.1 (2019): 1-12.

Exploring Compound Space and Genetic Space

DRIM: A web-based system for investigating drug response at the molecular level by condition-specific multi-omics data integration, *Frontiers in Genetics*. In press



Examples for Exploring Compound Space

Two major issues with compound space

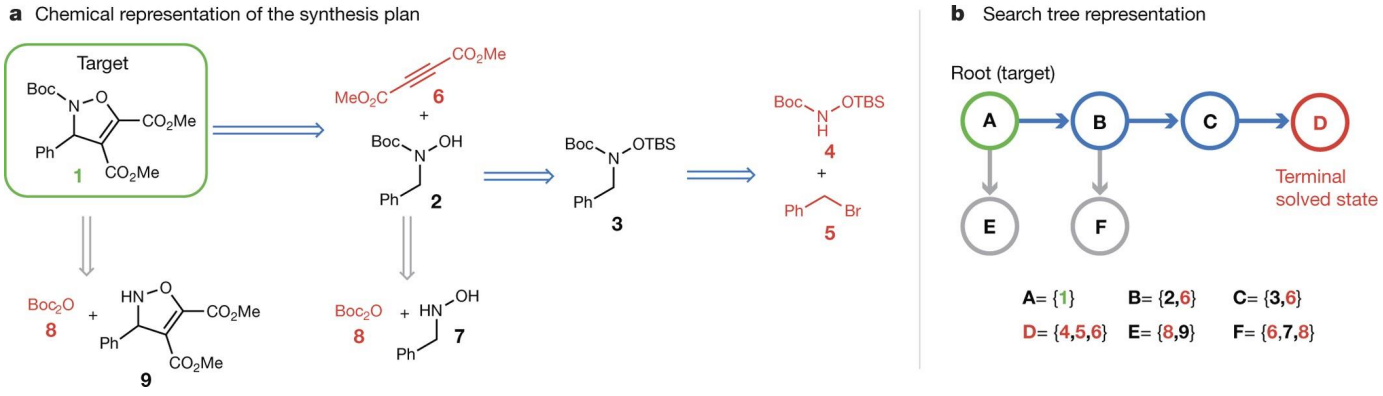
- **Search space**
  - A chemical space often referred to in cheminformatics is that of potential pharmacologically active molecules. Its size is estimated to be in the order of 10^{60} molecules.
  - https://en.wikipedia.org/wiki/Chemical_space
- **Synthesizable?**
  - Is it possible to synthesize a given compound?
  - What is the best planning for synthesizing the compound?

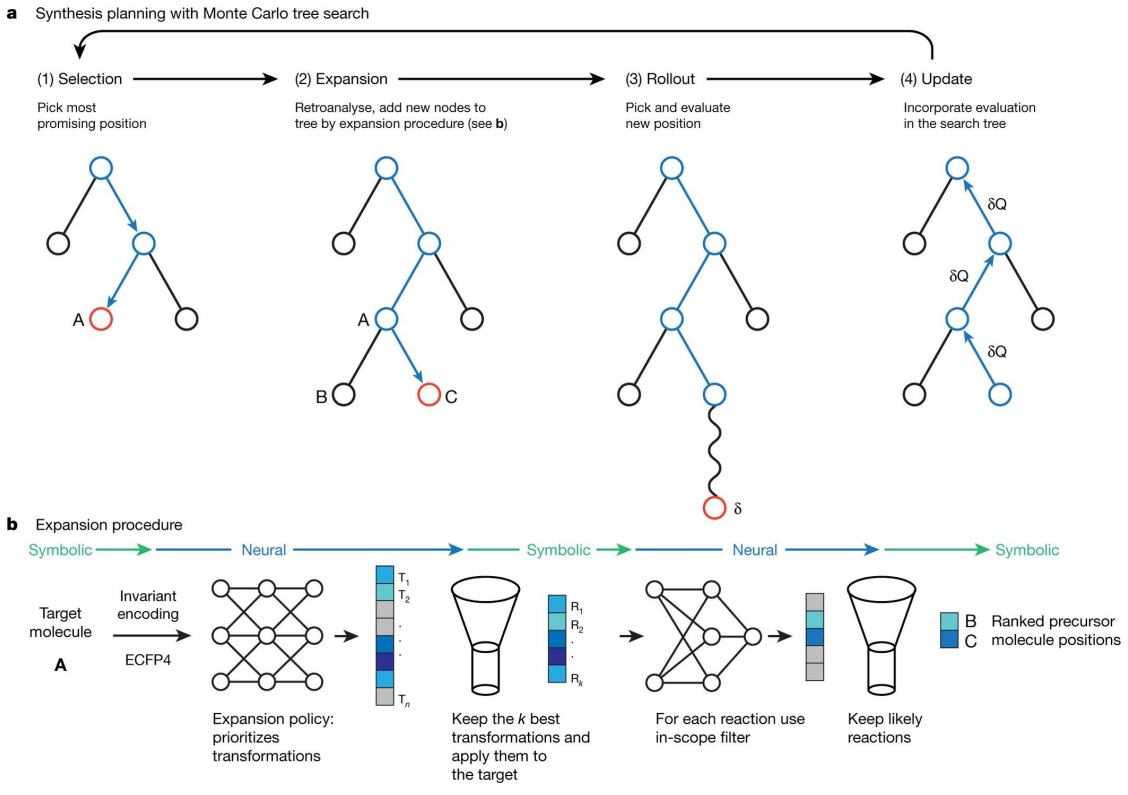
Dealing with How to Synthesize Compound Space

Planning chemical syntheses with deep neural networks and symbolic AI

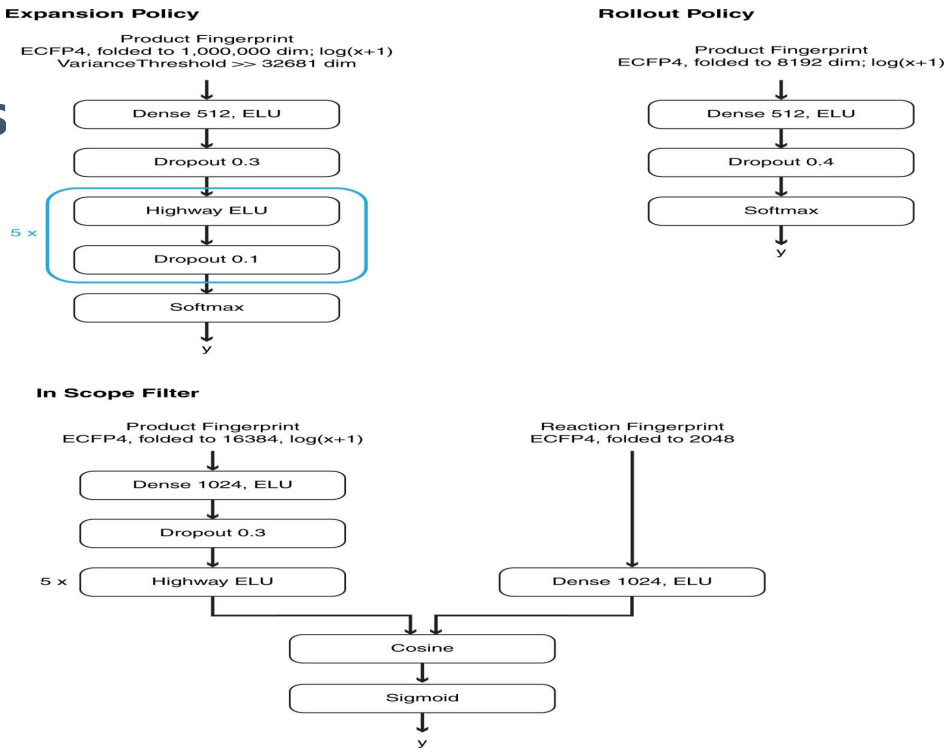
Nature 2018

Translation of the traditional chemists' retrosynthetic route representation to the search tree representation.

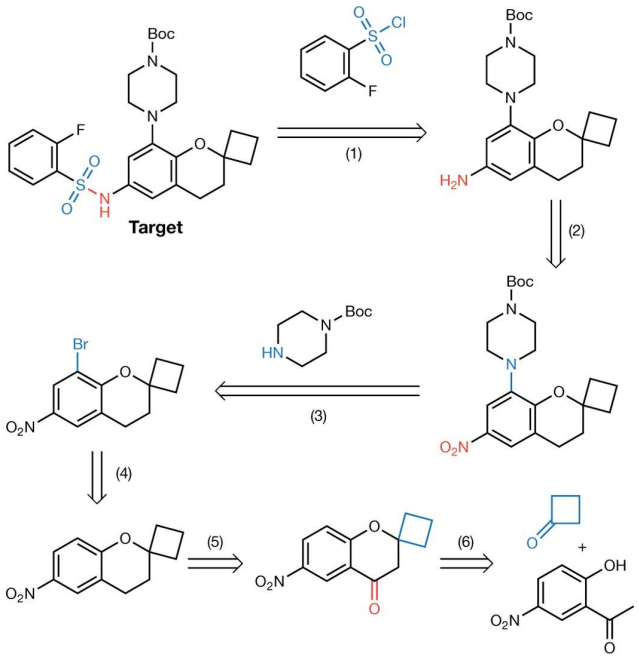




Architectures of the employed neural networks.



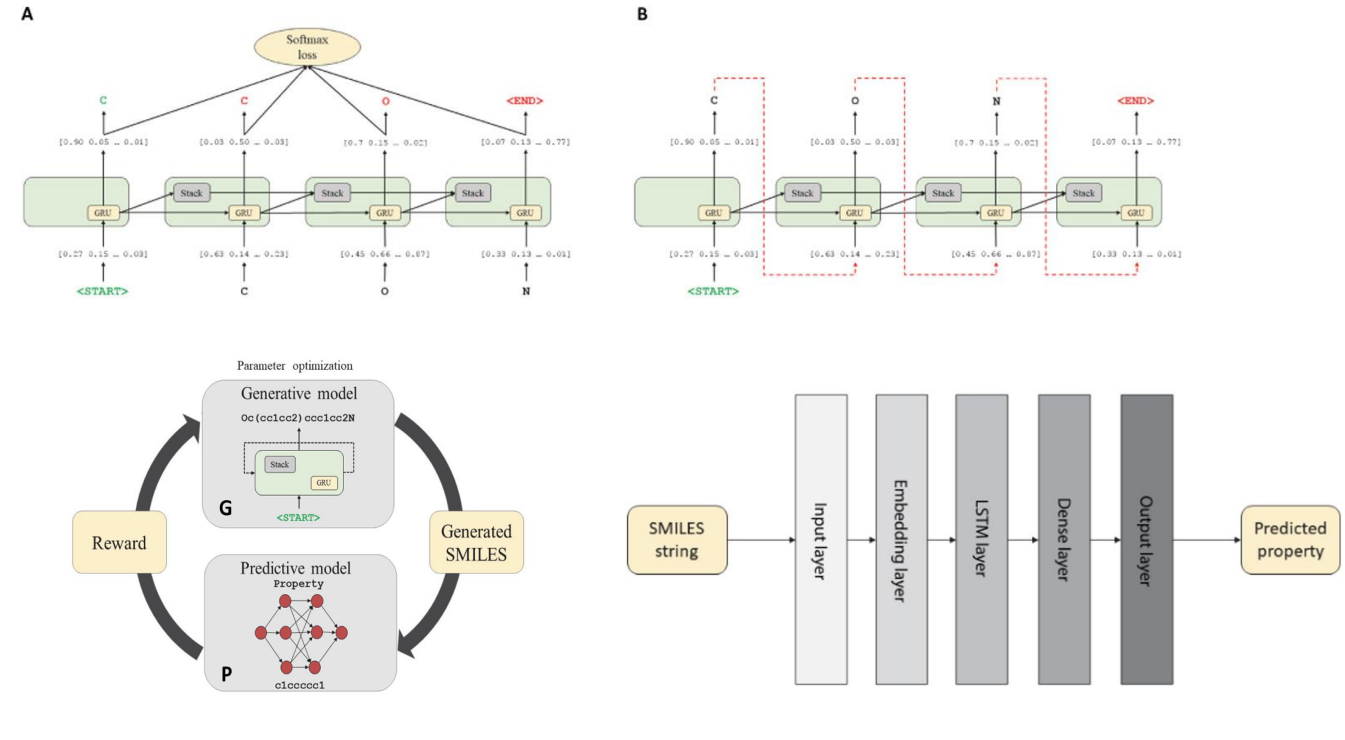
An exemplary six-step synthesis route for an intermediate in a drug candidate synthesis.



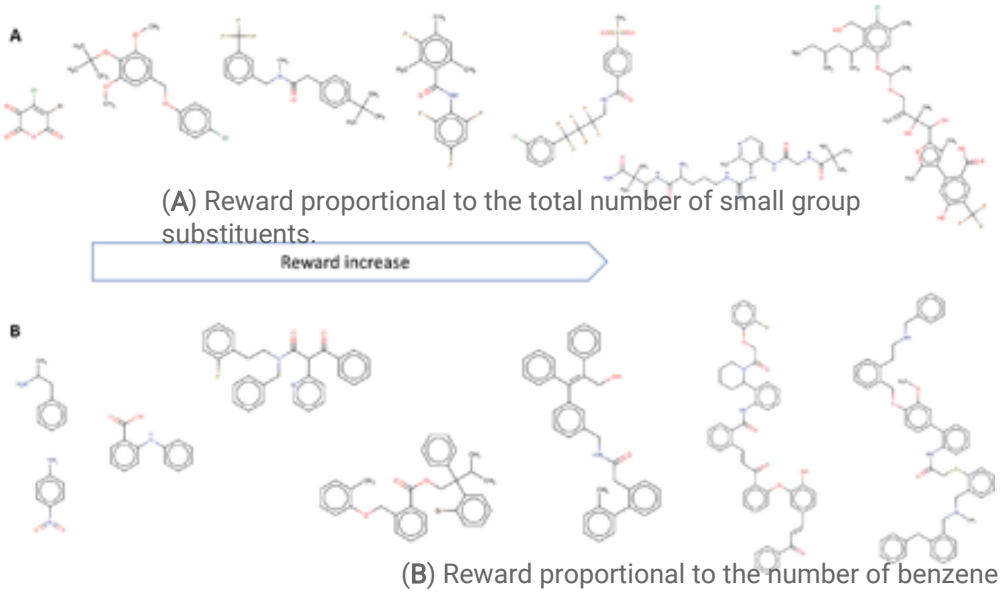
Dealing with Big Compound Space (1)

Deep reinforcement learning for de novo drug design

Science Advances 25 Jul 2018



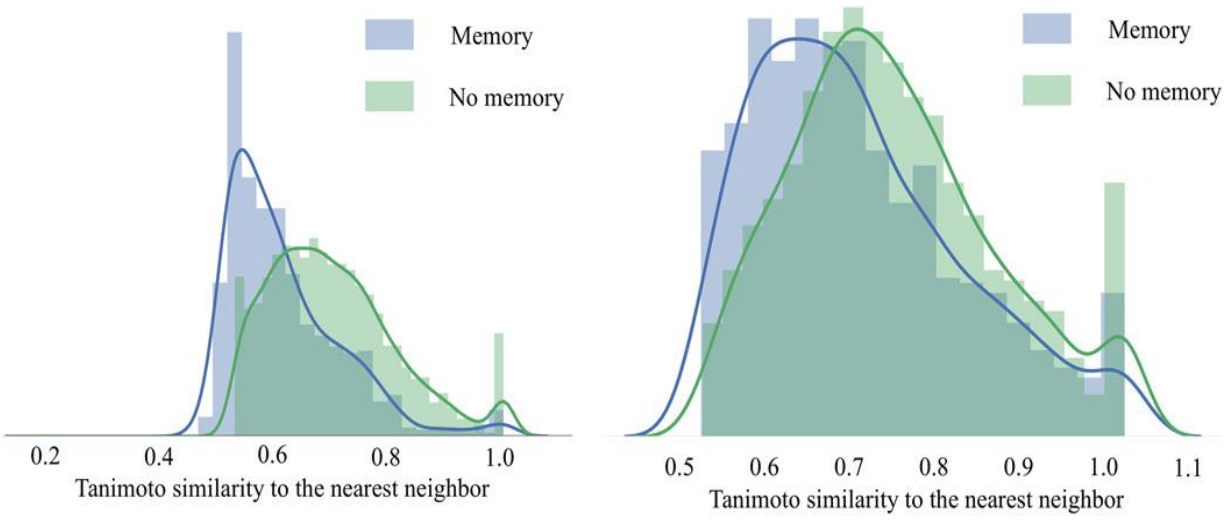
Evolution of generated structures as chemical substructure reward increases.



(A) Reward proportional to the total number of small group substituents.

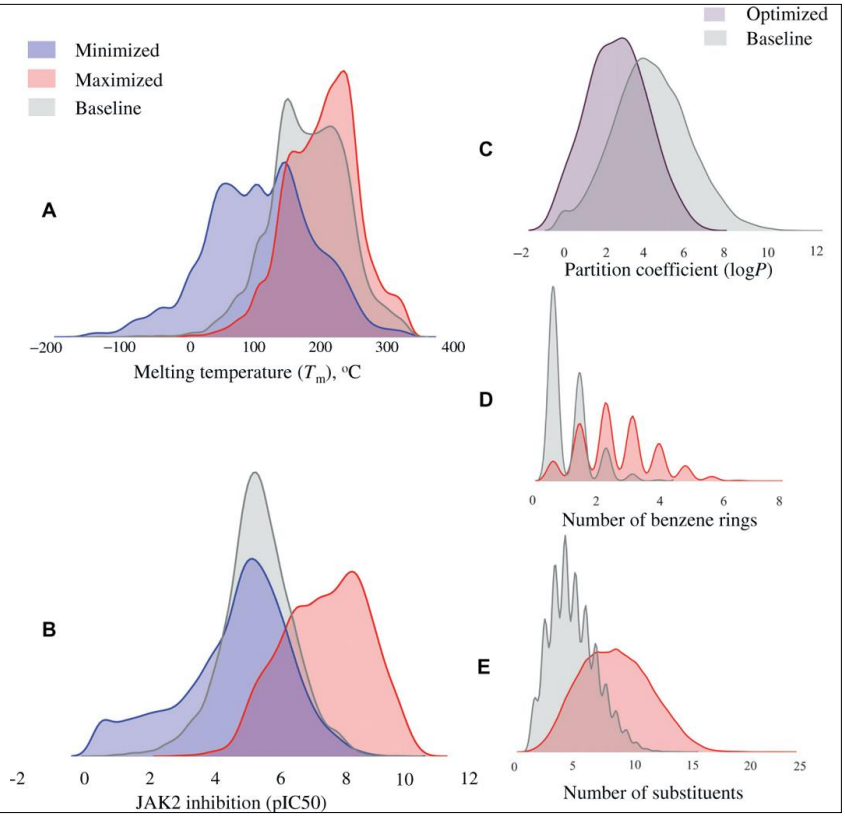
(B) Reward proportional to the number of benzene rings.

Stack RNN generates diverse molecules



(A) Internal diversity of generated libraries. (B) Similarity of the generated libraries to the training data set from the ChEMBL database.

RL generates valid molecules



Dealing with Big Compound Space (2)

Junction Tree Variational Autoencoder for Molecular Graph Generation

ICML 2018

Main Idea to Reduce Search Space

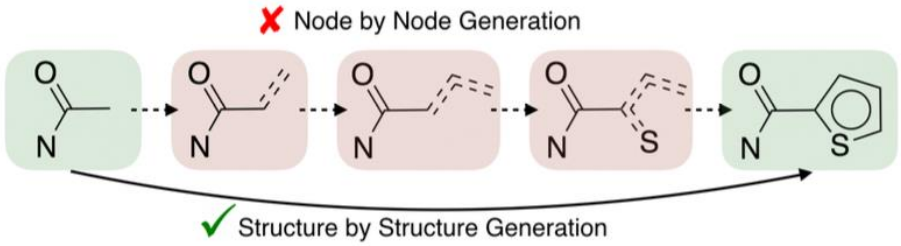
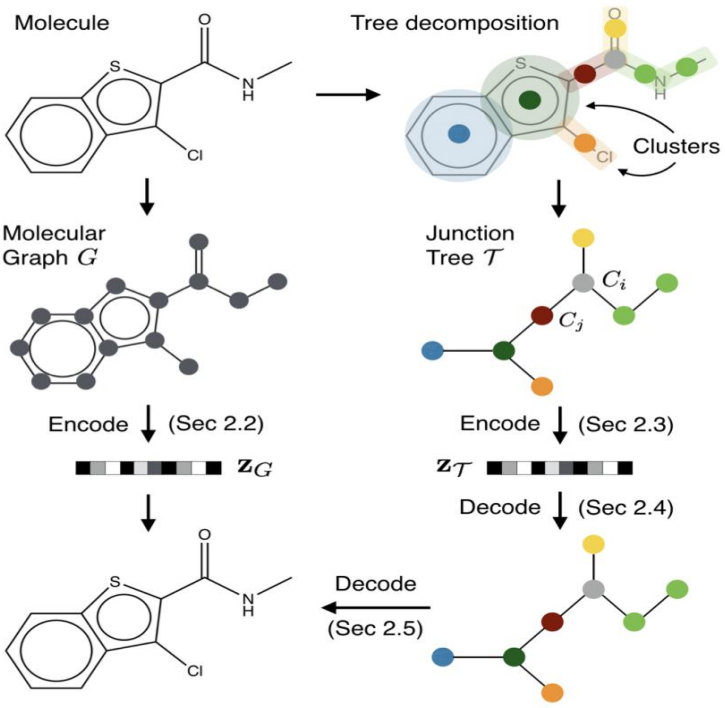


Figure 2. Comparison of two graph generation schemes: Structure by structure approach is preferred as it avoids invalid intermediate states (marked in red) encountered in node by node approach.

Junction Tree Molecular Generation



Tree Decoding

Junction Tree Variational Autoencoder for Molecular Graph Generation

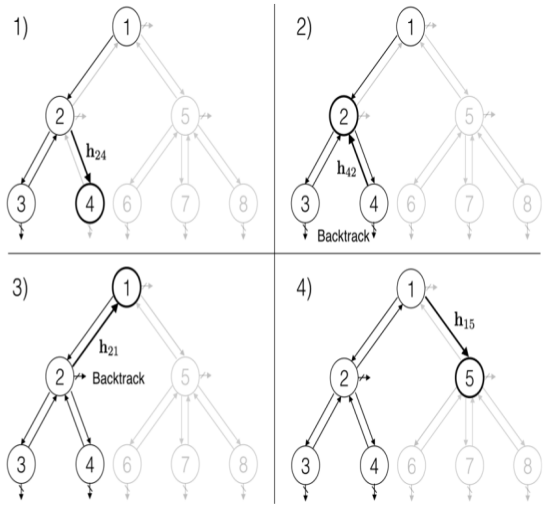


Figure 4. Illustration of the tree decoding process. Nodes are la-

Algorithm 1 Tree decoding at sampling time

Require: Latent representation $\mathbf{z}_{\mathcal{T}}$

Initialize: Tree $\widehat{\mathcal{T}} \leftarrow \emptyset$
function SampleTree(i, t)
  Set $\mathcal{X}_i \leftarrow$ all cluster labels that are chemically compatible with node i and its current neighbors.
  Set $d_t \leftarrow expand$ with probability p_t. ▷ Eq.(11)
  if $d_t = expand$ **and** $\mathcal{X}_i \neq \emptyset$ **then**
    Create a node j and add it to tree $\widehat{\mathcal{T}}$.
    Sample the label of node j from $\mathcal{X}_i$ ▷ Eq.(12)
    SampleTree$(j, t+1)$
  end if
end function

Decoding a Molecule from a Junction Tree

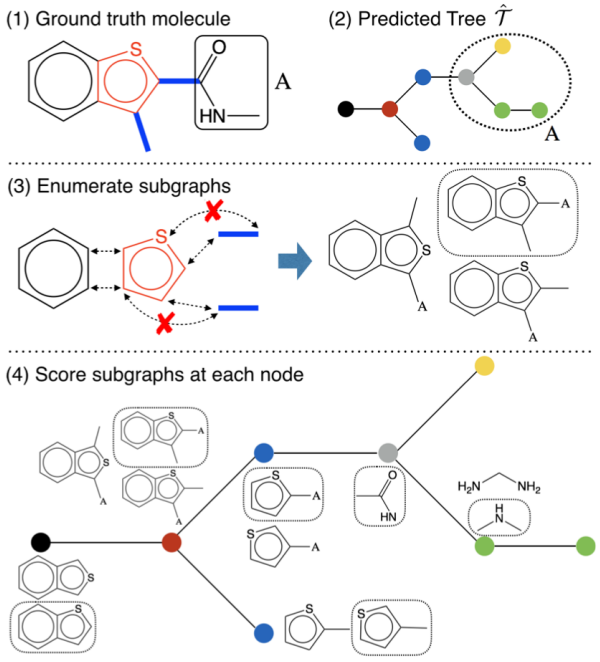


Figure 5. Decode a molecule from a junction tree. 1) Ground truth

Graph Decoder

Let $\mathcal{G}(\mathcal{T})$ be the set of graphs whose junction tree is $\mathcal{T}$. Decoding graph $\hat{G}$ from $\widehat{\mathcal{T}} = (\widehat{\mathcal{V}}, \widehat{\mathcal{E}})$ is a structured prediction:

$$\hat{G} = \arg \max_{G' \in \mathcal{G}(\widehat{\mathcal{T}})} f^a(G') \qquad (14)$$

where f^a is a scoring function over candidate graphs. We only consider scoring functions that decompose across the clusters and their neighbors. In other words, each term in the scoring function depends only on how a cluster C_i is attached to its neighboring clusters C_j, $j \in N_{\widehat{\mathcal{T}}}(i)$ in the tree $\widehat{\mathcal{T}}$. The problem of finding the highest scoring graph $\hat{G}$ –

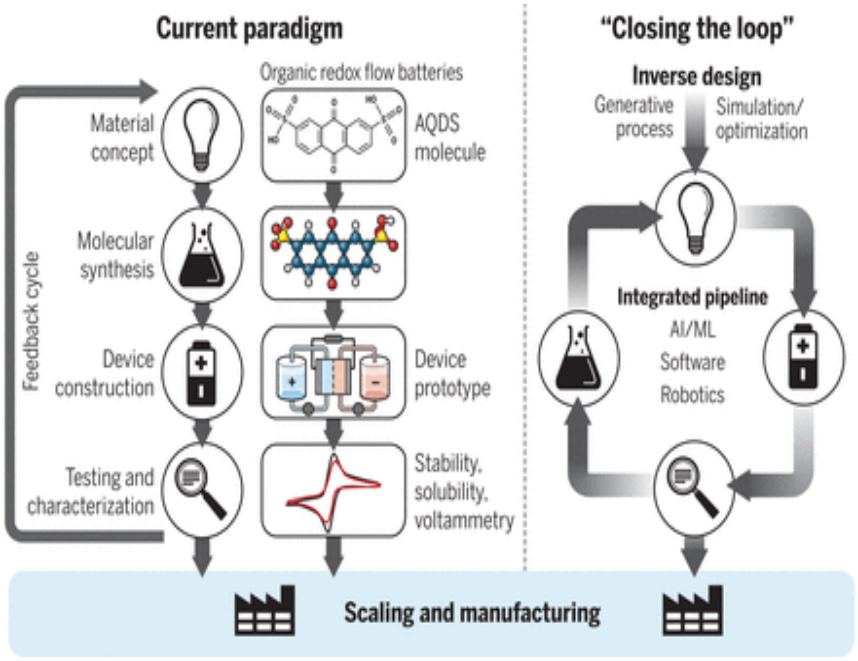
bors C_j, $j \in N_{\widehat{\mathcal{T}}}(i)$. We score G_i as a candidate subgraph by first deriving a vector representation $\mathbf{h}_{G_i}$ and then using $f_i^a(G_i) = \mathbf{h}_{G_i} \cdot \mathbf{z}_G$ as the subgraph score. To this end,

Dealing with Big Compound Space (3)

Inverse molecular design using machine learning: Generative models for matter engineering

Science 27 Jul 2018

Closing the loop requires incorporating inverse design, smart software (*93*), AI/ML, embedded systems, and robotics (*87*) into an integrated ecosystem.



Example for Exploring Target (gene) Space

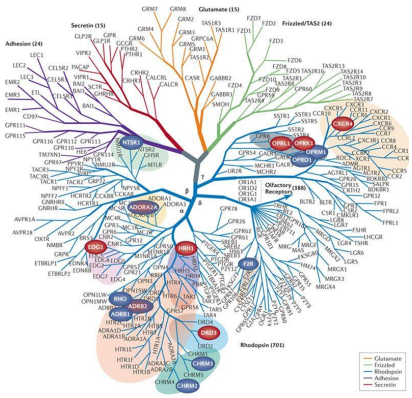
Deep Hierarchical Embedding for Simultaneous Modeling of GPCR Proteins in a Unified Metric Space

Taeheon Lee, et al
in review

Backgrounds

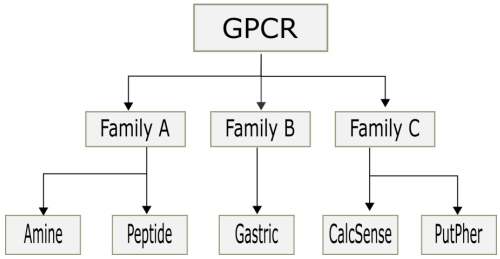
G-protein Coupled Receptors

- largest transmembrane protein family
- important drug targets
- widely diverged protein family
- hierarchical class structure



Previous Studies

Classifying / Modeling each **hierarchical classes**



Limits

Set of disconnected models for each subparts
No unified representation for GPCR sequences

Stevens, Raymond C., et al. "The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. " *Nature reviews Drug discovery* 12.1 (2013): 25.

Introduction

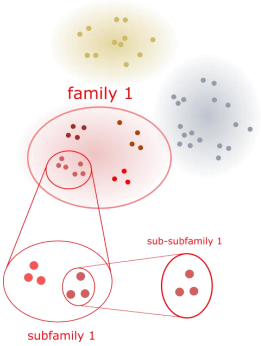
Our work

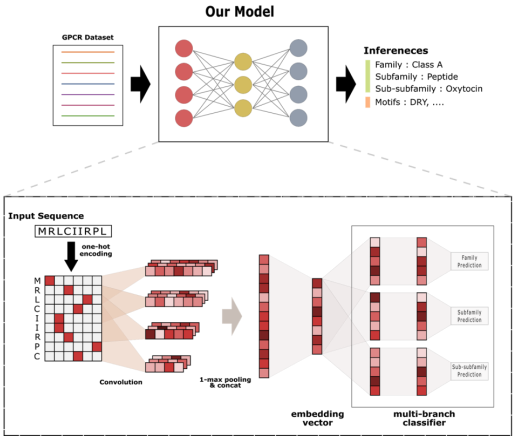
Modeling **hierarchical class structure** in GPCR

- with a **unified model**
- into a **single metric space**
- using **deep learning**

Key contributions

- vector embedding
- metric distances between vectors





Methods

Neural Network Architecture

Data representation

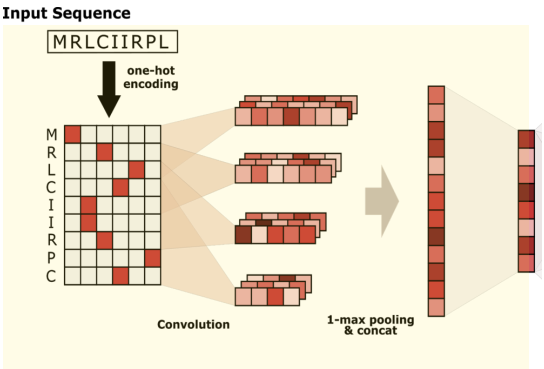
One-hot encoding

Feature extractor with CNN

1-D motif discovery convolutional filter [DeepBind]

Various window lengths convolutional filter [DeepFam]

Local & significant sequence feature is learned



1-max pooling & concatenation

Existence of learned motifs in the sequence [DeepBind]

Alipanahi, Babak, et al. "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning." *Nature bio technology* 33.8 (2015): 831.

Seo, Seokjun, et al. "DeepFam: deep learning based alignment-free method for protein family modeling and prediction." *Bioinfo rmatics* 34.13 (2018): i254-i262.

Methods

Neural Network Architecture

Embedding layer

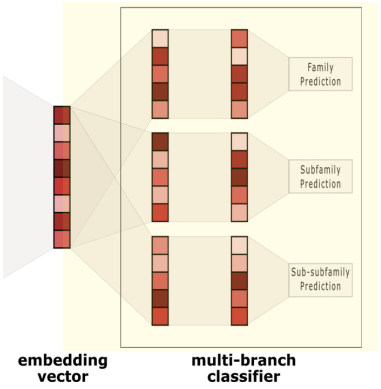
Dimension reduction on features from CNN layer

Representation of the input sequence

Multiple branch classifier [MDNET]

s Unification of features from three hierarchical levels

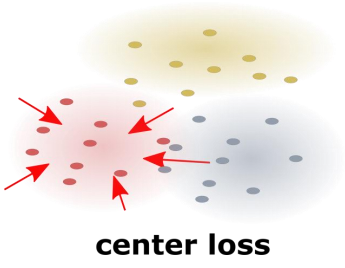
Generation of shared features from three different levels

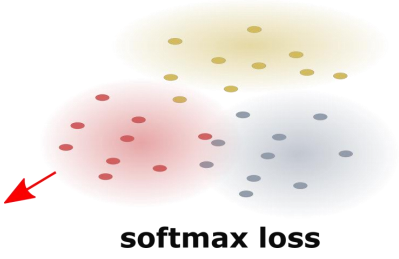


Nam, Hyeonseob, and Bohyung Han. "Learning multi-domain convolutional neural networks for visual tracking." *Proceedings of the I EEE Conference on Computer Vision and Pattern Recognition*. 2016.

Methods

Loss Function





Center loss [Center loss]

$$L_c = \sum_{i=1}^{n} \|d(x_i) - \mu_C(x_i)\|_2 \longrightarrow L_c = \max\left(\sum_{i=1}^{n} \|d(x_i) - \mu_C(x_i)\|_2, m\right)$$

$d(x_i)$: representation of the sequence x_i in the neural network

$\mu_C(x_i)$: center representation of the class that the sequence x_i belongs to

m : class boundary margin (configured for each level)

n : number of data points

Compact representation in terms of distances

Softmax loss (with cross entropy)

softmax $\sigma(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}$

cross entropy $L_S = -\sum_{i=1}^{n} y_i \log \sigma(x_i)$

n : number of data points

K : number of classes

y_i : class label of data i

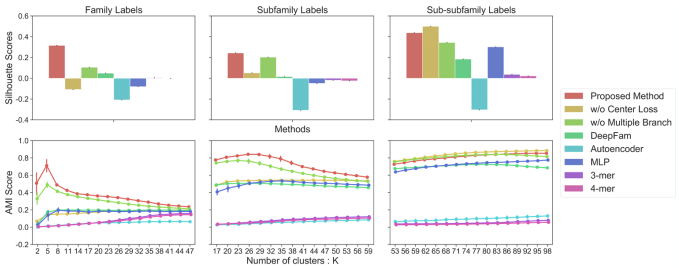
Separable representation of data in different classes

Wen, Yandong, et al. "A discriminative feature learning approach for deep face recognition." *European conference on computer vision*. Springer, Cham, 2016.

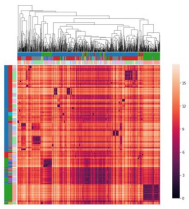
Results

Analysis on Embedding Vectors

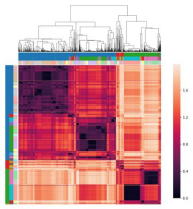
Cluster Analysis



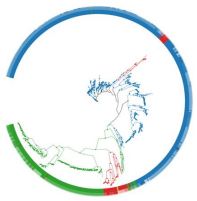
Phylogenetic Structure



Distance matrix from DeepFam



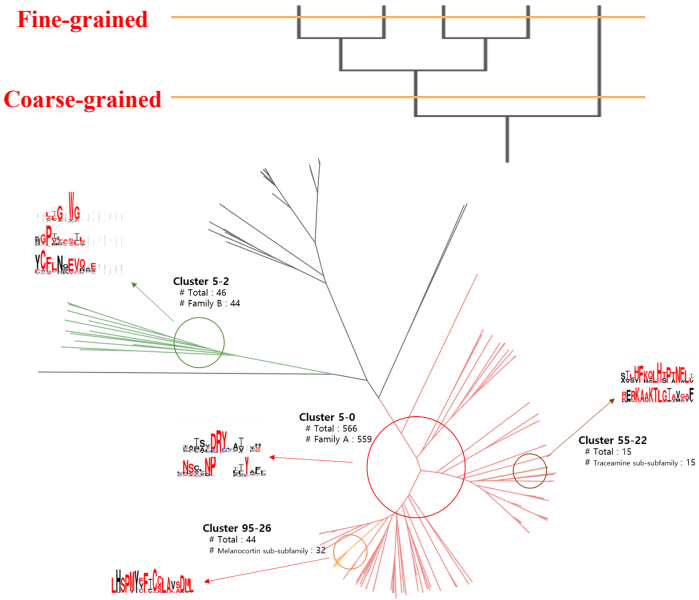
Distance matrix from the Proposed Method



Phlyogenetic Tree from the Proposed Method

Results

Motif Analysis from Embeddin Vectors



Motif analysis

motif discovery

dataset selection

Observations

Coarse-grained

- DRY, NSxxNPxxY : Family A
- LIGWG, GPVLASLL, CFLxEVQ : Family B

Fine-grained

- RKAAKTLG, FKQLHXPTM : Traceamine(A)
- SPMxCCLAxDML : Melanocortin(A)

Example for
Exploring
Target (gene) Space

Improved protein structure prediction using potentials from deep learning

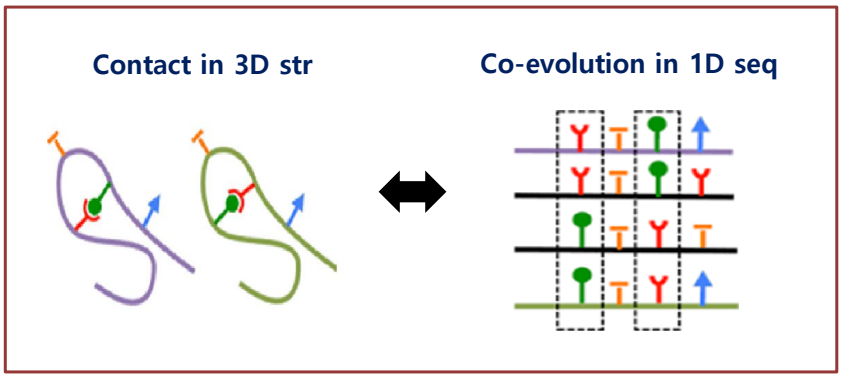
AlphaFold

Nature **volume 577**, pages706–710(2020)

Success of Bioinformatics: contact prediction (2014)

David Baker, U of Washington, Seattle

Related works by many other authors since 1999



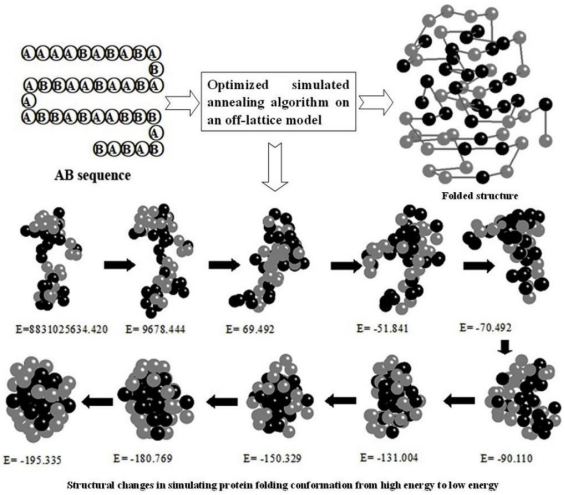
Slide from Chaok Seok @ SNU

Introduction: Determining Protein Structure

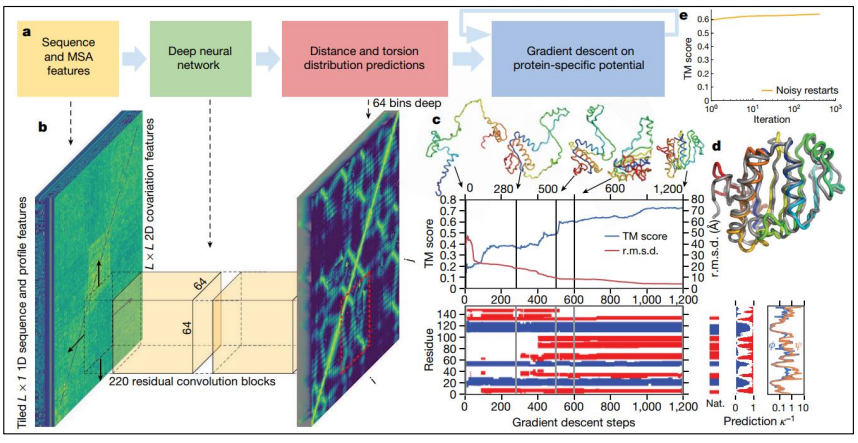
- Structure with low potential is stable!
- Protein structure prediction is to find a lowest potential structure with a given sequence.
- Quite a number of techniques are combined.
- **Structure with low potential are searched with deep learning models for predicting torsion angles and distances.**

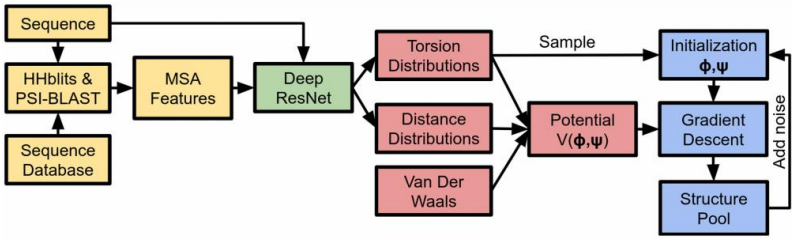
Simulated Annealing

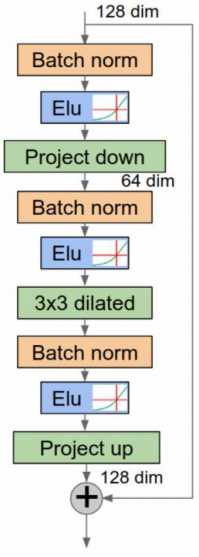
- Randomly choose positions / lower the probability of moving at each step



Structural changes in simulating protein folding conformation from high energy to low energy





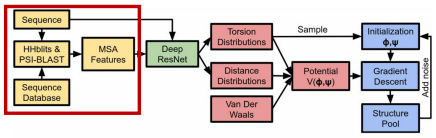


The layers used in one block of the deep residual convolutional network

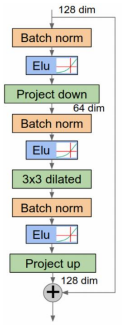
Training Set Preprocessing

- Extract MSA
  - Uniclust30 dataset -> search with HHblits
  - Position-specific substitution probabilities & covariation features

- Input features
  - 1-hot amino acid type (21 features)
  - Profile : PSI-BLAST, HHblits profiles, HMM profile
  - Biases, deletion probability, residue index
  - Covariation : Potts model parameters (484 + 1 (Frobenius norm))

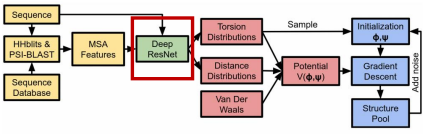


Distance Prediction Model

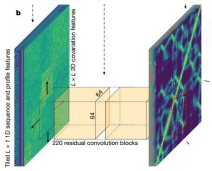


- 220 layer ResNet
  - 7 * 4(dilation 1, 2, 4, 8) : 256 channels
  - 48 * 4(dilation 1, 2, 4, 8) : 128 channels

  - Target : distance between the Cβ atoms of the residues; divided the range 2–22 Å into 64 equal bins
  - Auxiliary loss : secondary structure 0.005; accessible surface area : 0.001

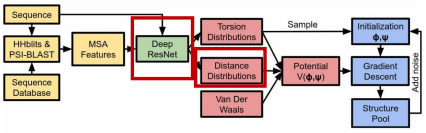


Distogram Prediction(1)



- Divided L x L distance matrix into non-overlapping 64 x 64 crops
  - Outcome for each 1x1 bin is probability distribution

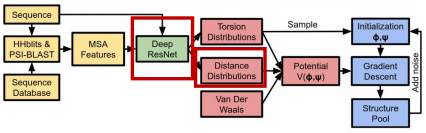
- Data Augmentation
  - Randomize the offset of the crops (many thousands of different training samples from single protein)
  - Add noise proportional to the ground truth resolution to the atom coordinates



Distogram Prediction(2)

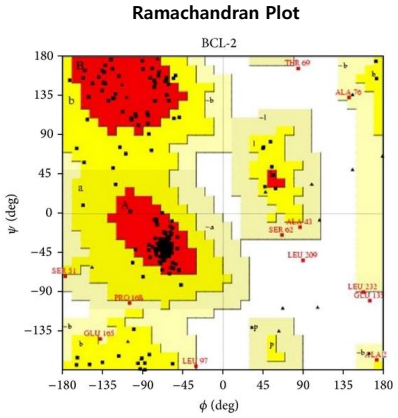
- To predict each L x L residue pairs, many 64 x 64 crops are combines
  - Several tilings are produced and averaged together
    - 64 x 64 different possible tilings
    - heavier weighting for the predictions near the center of the crop
  - Four separate models with slightly different hyperparameters are averaged together

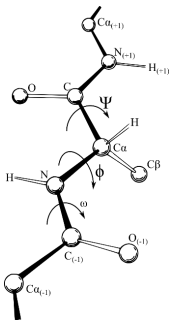
- Mode of combined distogram is used as prediction (figure on next page)

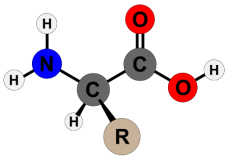


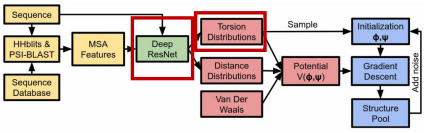
Tortion Prediction

- Predicts marginal Ramachandran distribution for each residue
  - Probability distribution is divided into 10° x 10° bins (36 x 36 bins)

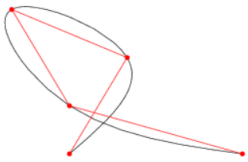




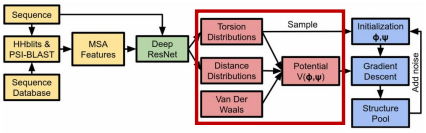




Building differentiable potential



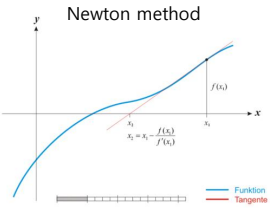
- Discrete distance distribution
  - Interpolated with a cubic spline
  - $V_{\text{distance}}(\mathbf{x}) = -\sum_{i,j,\ i \neq j} \log P(d_{ij} \mid \mathcal{S}, \text{MSA}(\mathcal{S})) - \log P(d_{ij} \mid \text{length}, \delta_{\alpha\beta})$
- Tortion distributions
  - Each marginal predictions are fitted with a unimodal von Mises Distribution
- Rosetta's $V_{\text{score2\_smooth}}$
  - Van der Waals term to prevent steric clashes

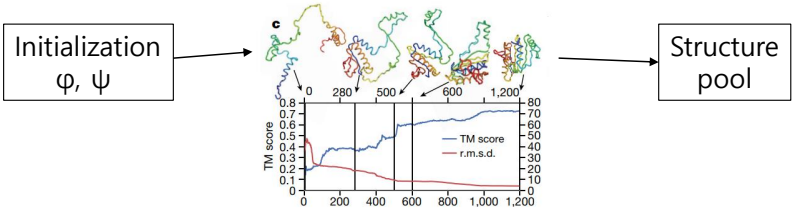


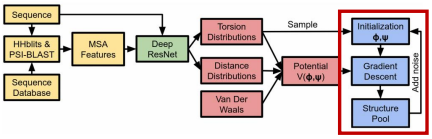
Structure realization

- We can now compute Differentiable Potential given torsion angle initialization!
  - $V_{\text{total}}(\phi, \psi) = V_{\text{distance}}(G(\phi, \psi)) + V_{\text{torsion}}(\phi, \psi) + V_{\texttt{score2\_smooth}}(G(\phi, \psi))$

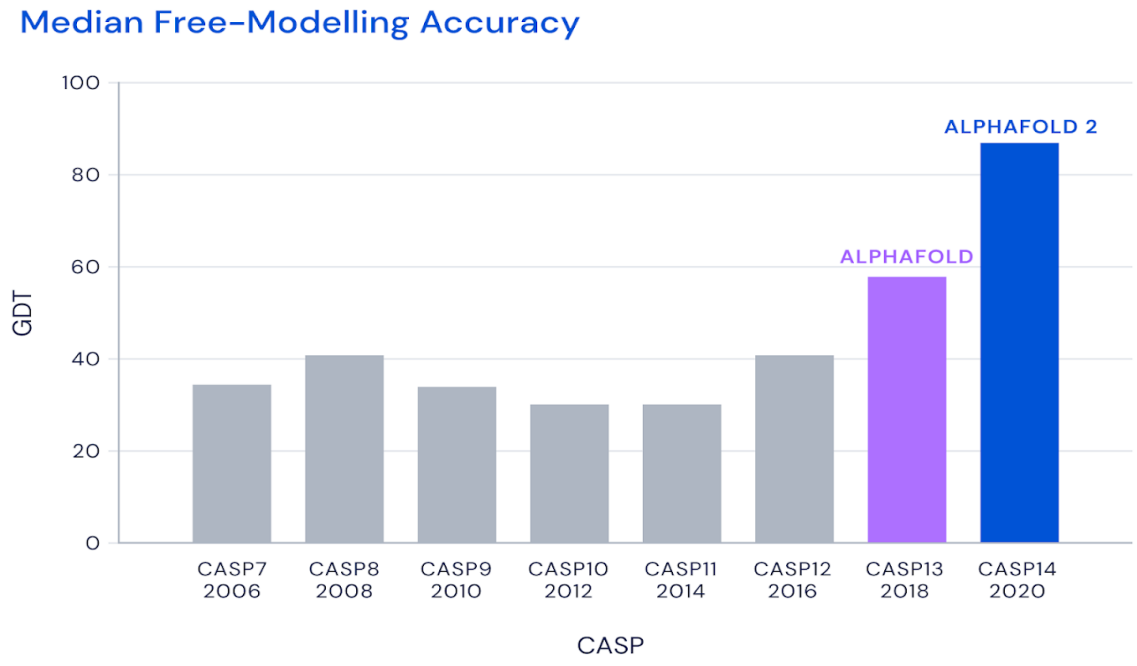
- Minimize V_{total} using gradient descent
  - L-BFGS, a variation of quasi-newton method, is used







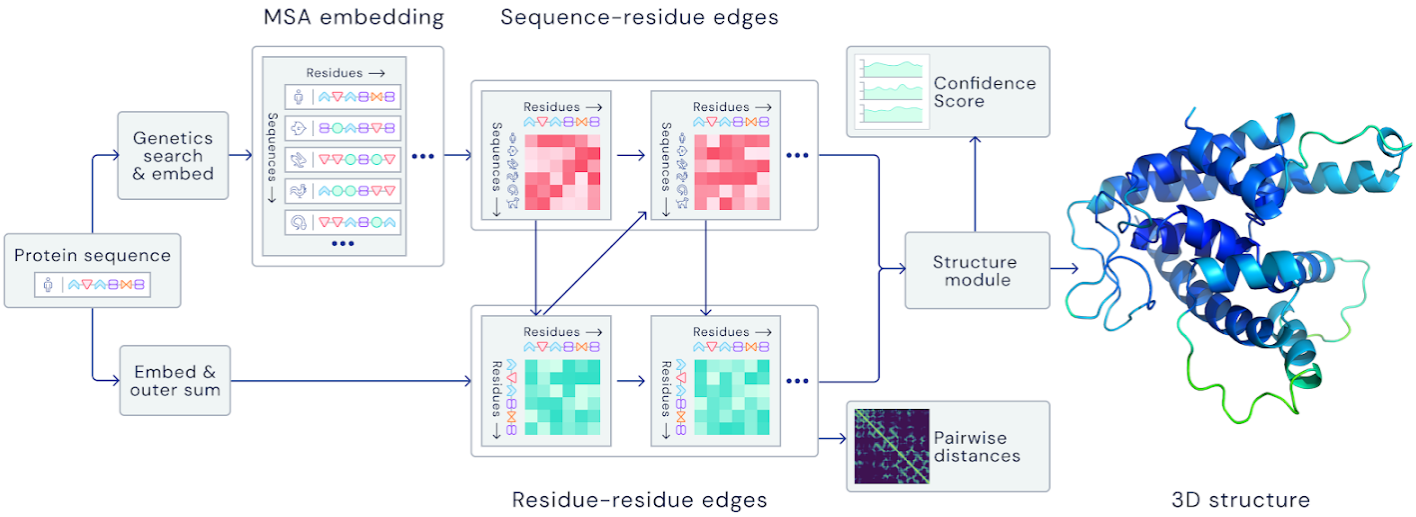
AlphaFold2 @ CASP14, 2020



Slide from Chaok Seok @ SNU

AlphaFold2: Deep learning의 끝판왕 (2020)
John Jumper

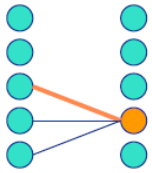
(Mentioned physics & geometry in his talk)



Slide from Chaok Seok @ SNU

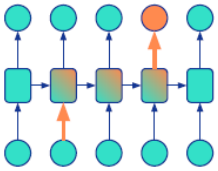
John Jumper: Data가 충분하지 않으므로 학습을 잘 할 수 있는 network 이 필요하다

Inductive Bias for Deep Learning Models



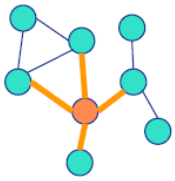
Convolutional Networks (e.g. computer vision)

- data in regular grid
- information flow to local neighbours



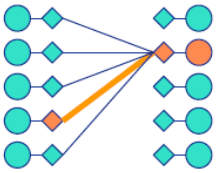
Recurrent Networks (e.g. language)

- data in ordered sequence
- information flow sequentially



Graph Networks (e.g. recommender systems or molecules)

- data in fixed graph structure
- information flow along fixed edges



Attention Module (e.g. language)

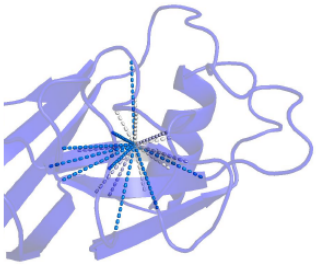
- data in unordered set
- information flow dynamically controlled by the network (via keys and queries)

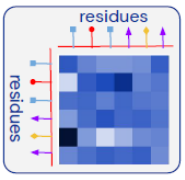
Slide from Chaok Seok @ SNU

John Jumper: Physical insights are built into the network

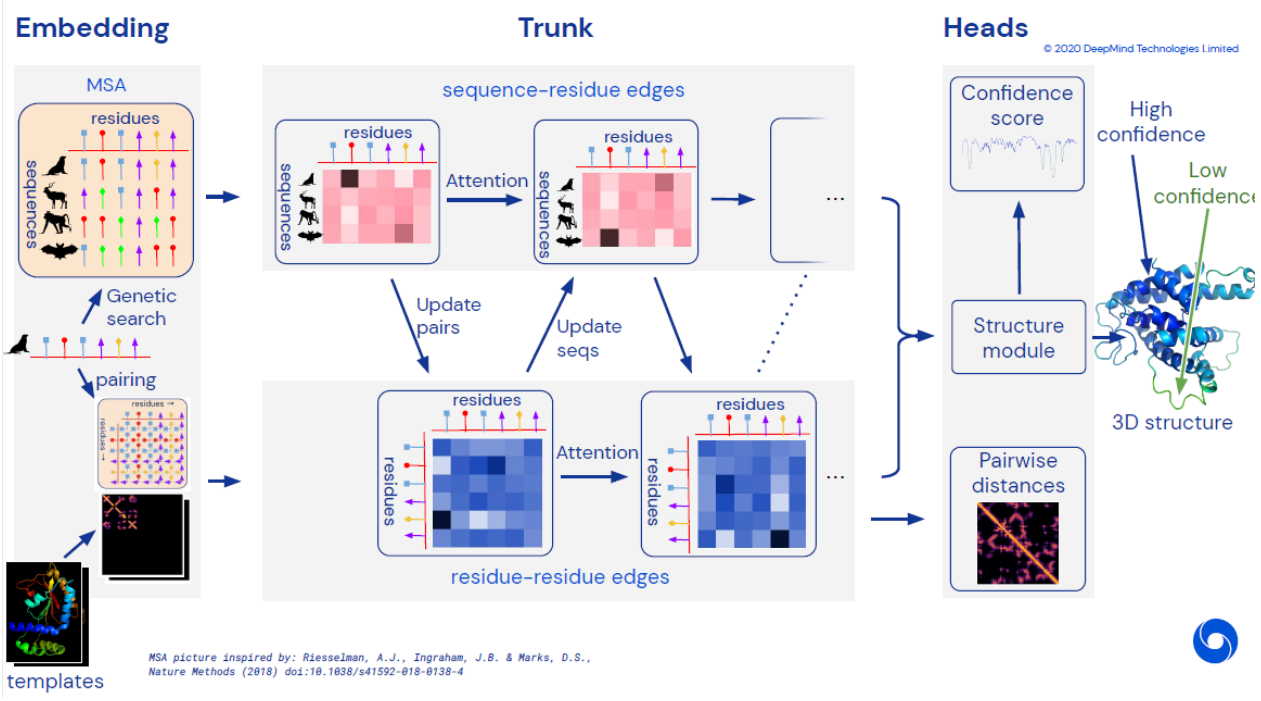
Putting our protein knowledge into the model

- Physical insights are built into the network structure, not just a process around it
- End-to-end system directly producing a structure instead of inter-residue distances
- Inductive biases reflect our knowledge of protein physics and geometry
  - The positions of residues in the sequence are de-emphasized
  - Instead residues that are close in the folded protein need to communicate
  - The network iteratively learns a graph of which residues are close, while reasoning over this implicit graph as it is being built





Slide from Chaok Seok @ SNU

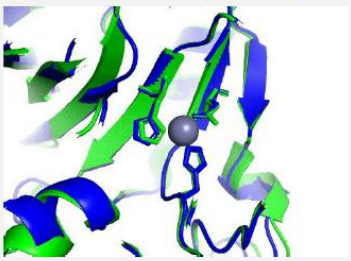


Slide from Chaok Seok @ SNU

Biological context

- Computational structure prediction is typically underspecified
  - Oligomeric state, ligands, DNA-binding, experimental conditions, multiple conformations etc.
- Our networks implicitly models the missing context
- Uses a variety of physical and evolutionary information (e.g. profile-only is still pretty accurate)

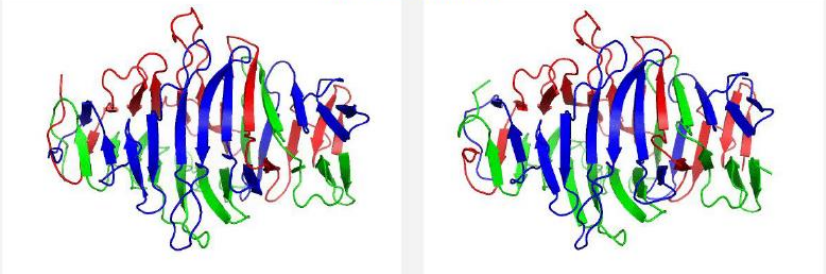
T1056 (zinc binding)



AlphaFold / Experiment

TBM-hard, 98.2 GDT

T1080 (trimer)



AlphaFold (monomer prediction x3)

FM/TBM, 85.9 GDT

Experimental structure

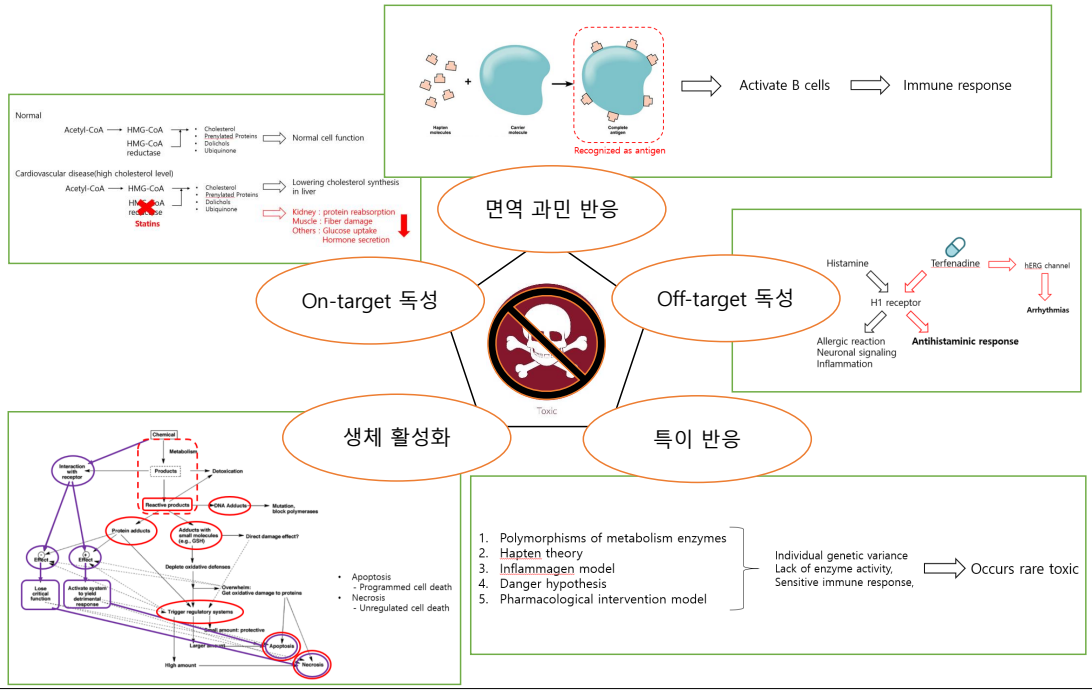
Slide from Chaok Seok @ SNU

Example for Exploring Phenotype Space (Toxicity)

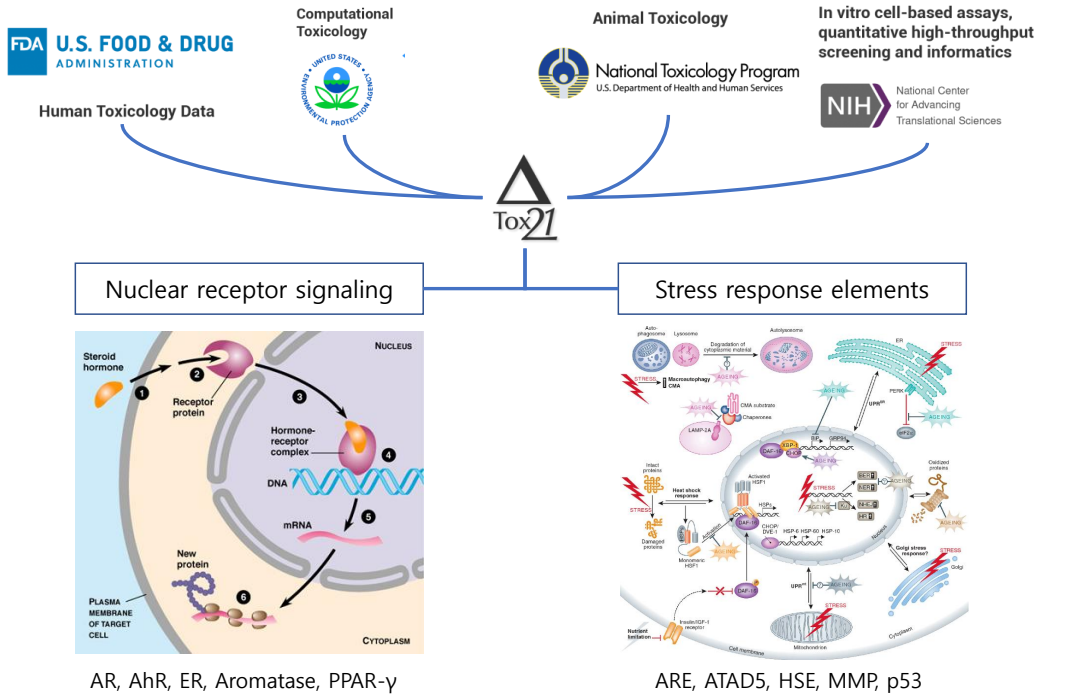
TOP: A deep mixture representation learning method for boosting molecular toxicity prediction

Methods. 2020

약물의 대표적인 독성 발현 기전



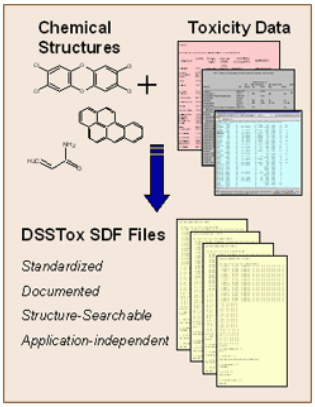
세포 수준의 독성 관련 수용체 및 패스웨이 – Tox21



독성 관련 공공 데이터베이스

독성에 관련된 공공 데이터베이스 리스트.
데이터베이스마다 독성에 대한 기준과 사용한 기법이 다름
연구 목적에 맞게 적절한 데이터 베이스 활용 예정.

Database	특징	# of compounds
ToxCast	다양한 세포주, 생물학적 활성 등을 고려해 약 700 종류의 *in vitro* assay를 high-throughput screening으로 측정한 데이터 베이스	약 8500개
Tox21	12개의 주요 세포 독성 타겟에 대한 화학물질의 반응을 Luciferase assay 등을 통해 정성적으로 측정한 데이터 베이스	약 8000개
DSSTOX	화학물질의 물리화학적 특징과 Tox21, ToxCast의 생물학적 실험 데이터를 연동해 제공하는 데이터 베이스	약 740,000개
ClinTox	FDA에 승인된 약물과 독성 문제로 임상 시험에 실패한 약물 비교	약 1500개
SIDER	시판 중인 약물의 부작용에 대한 통합 정보 데이터 보유. 약물 부작용이 보고된 논문 및 실험 데이터를 빈도와 심각도에 따라 분류해 제공	약 1500개
ECOTOX	13,000여 종의 생물에 대한 화학물질의 독성 실험 데이터를 통합해 제공. 독성은 EC50, IC50, NOEL 등을 기준으로 평가하고 관련 논문 링크 수록.	약 12,000개

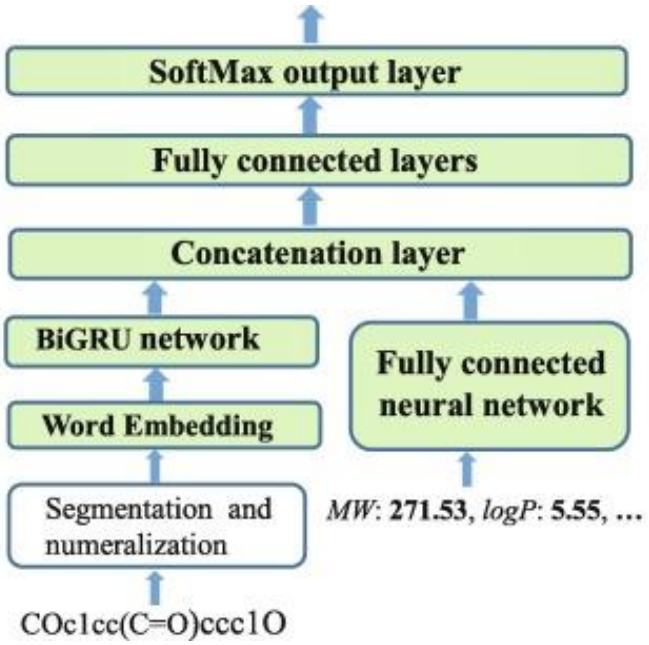


ToxCast: Chemical Research in Toxicology **2011** 24 (8), 1251-1262
Tox21: National Research Council. 2007. *Toxicity Testing in the 21st Century: A Vision and a Strategy.*
DSSTOX: Computational Toxicology, 12 (2019), p. 100096
ClinTox: PLOS ONE, 2013, 8
SIDER:Nucleic Acids Research, 2016, Vol. 44, Database issue D1075–D1079
ECOTOX: Environmental toxicology and chemistry 30.8 (2011)

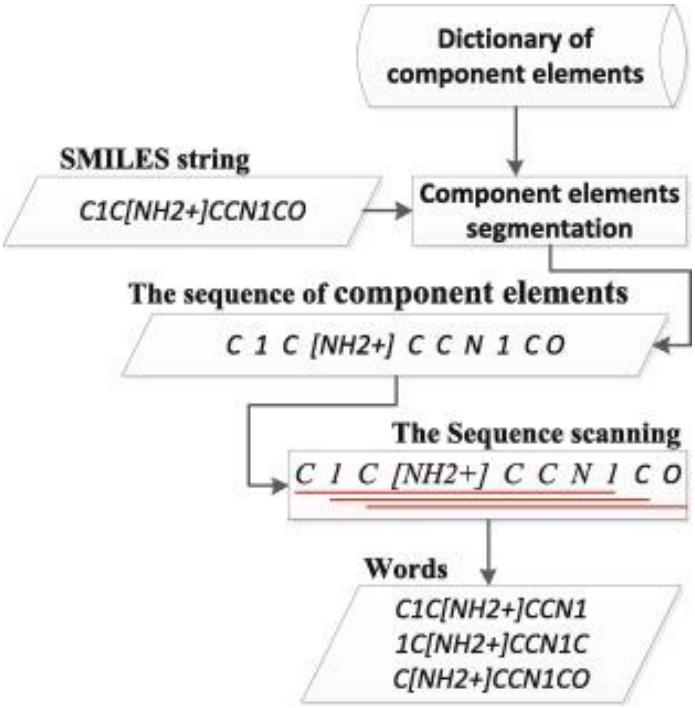
Problem to be solved here for Toxicity

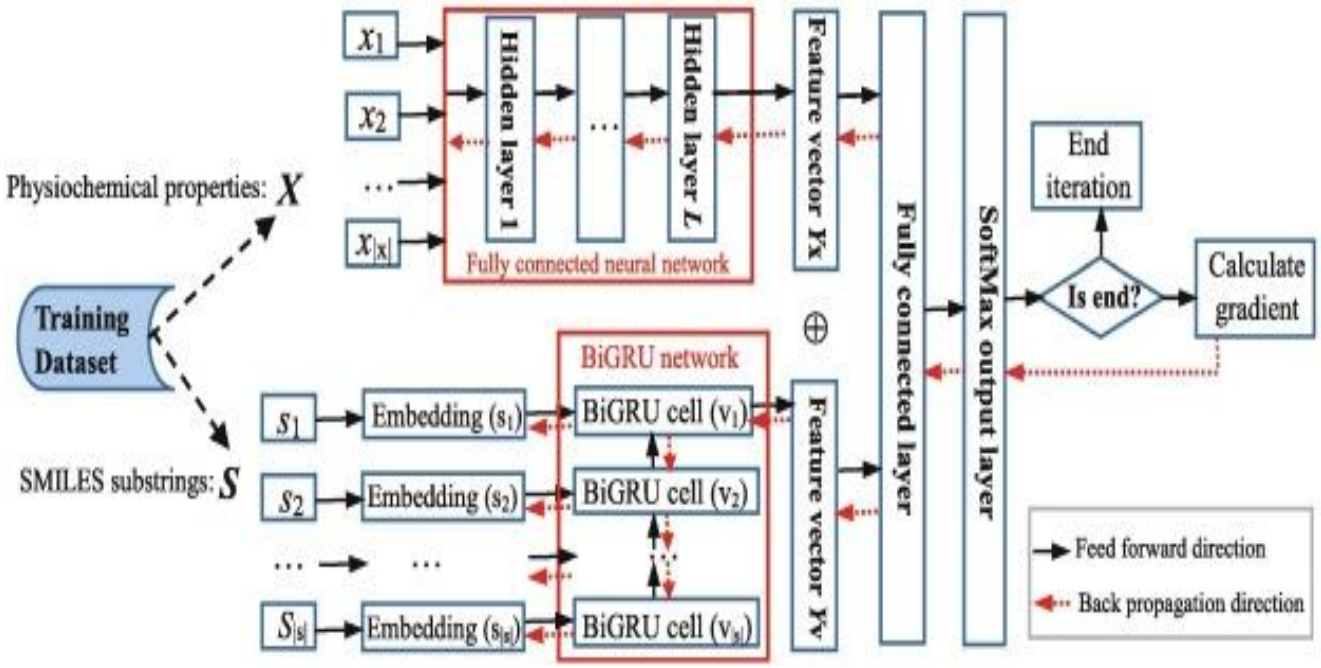
- Any compounds that target very important genes are toxic.
- Example) 12 proteins in Tox21 database

Top Architecture

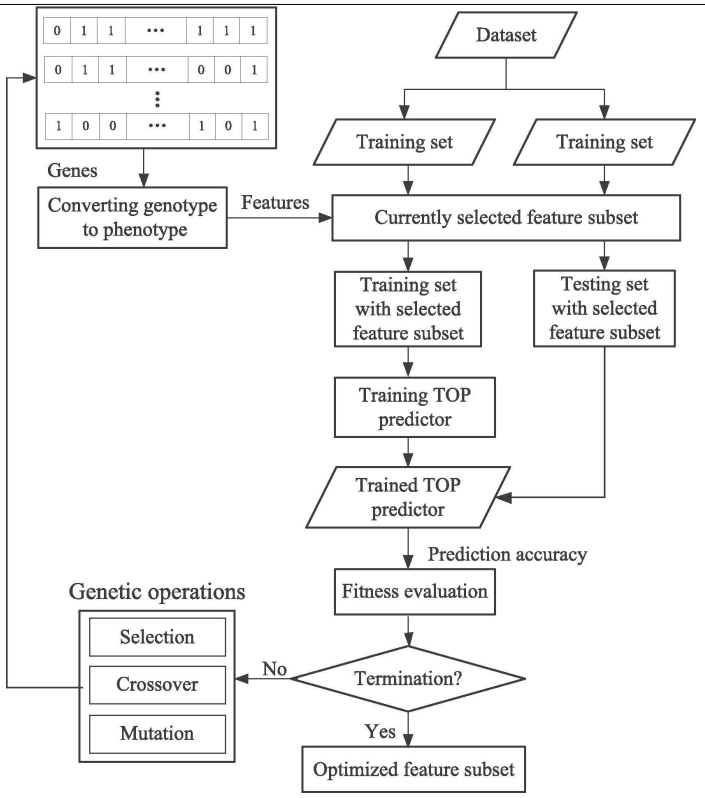


Segmentation





GA-based physiochemical feature selection



An ablation study on ToxB

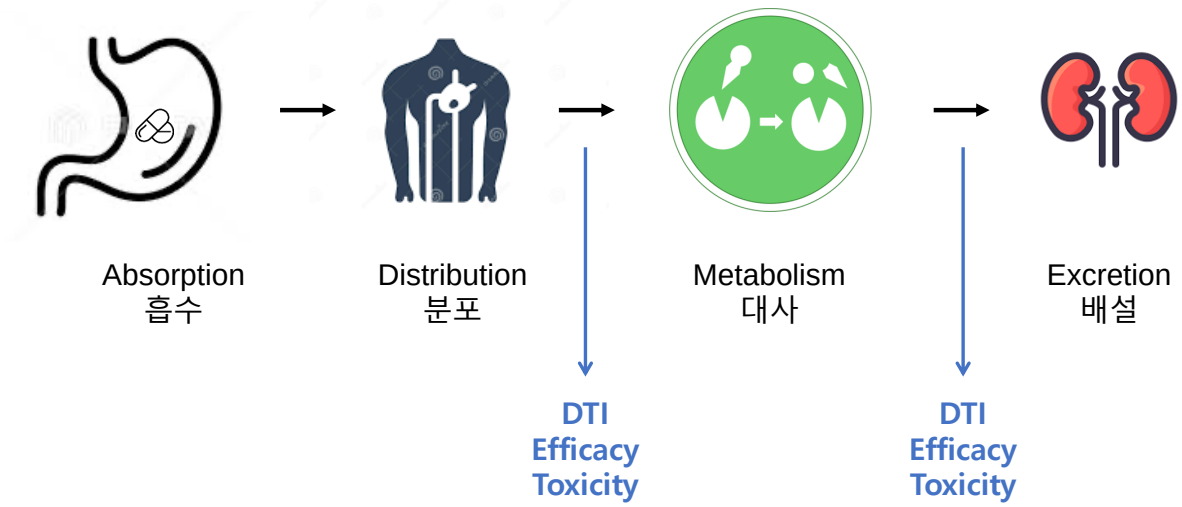
Model	AUC
TOP with full components	0.912 ± 0.005
Without samples augmentation	0.908 ± 0.003
Without physiochemical properties learning	0.717 ± 0.012
Without BiGRU-based SMILES string learning	0.632 ± 0.001
Replacing BiGRU with GRU	0.899 ± 0.007
Replacing BiGRU with BiRNN	0.903 ± 0.006

Example for Exploring Phenotype Space (Drug Metabolosm)

Slide made by Dongmin Bang

Backgroud: ADME

- **ADME** : Disposition of a drug within an **organism**



Background: metabolism & CYP450

- **Metabolism** : transformation of xenobiotics into excretable form

- **Cytochrome P450** : “The most important enzyme” in metabolism
  - Mostly expressed in the liver and small intestine
  - Superfamily of at least 57 isoforms
  - Activity of CYP450 directly effects the drug efficacy/toxicity
    - Expression pattern
    - Inhibition level – Drug-drug interaction
    - **Genetic polymorphism**

Background: metabolism & CYP450

- CYP450 and genetic polymorphism
  - Contains Highly polymorphic genes
  - **Activity Score(AS) system : translates genotype to phenotype**
    - Classifies phenotypes into **"Decreased/Normal/Increased function"**
  - Adopted by CPIC(Clinical Pharmacogenetics Implementation Consortium)
    - "Guidelines for CYP3A5 Genotype and Tacrolimus Dosing"

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

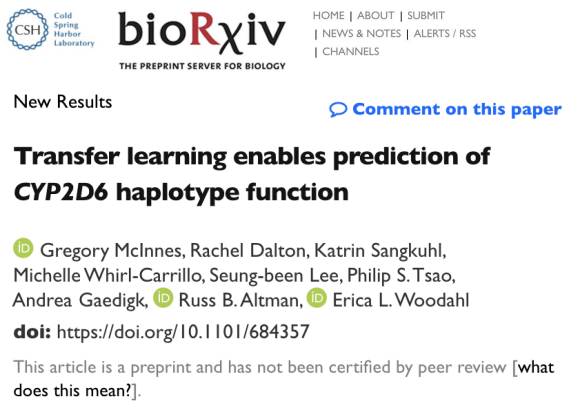
introduction : metabolism & CYP450

- CYP450 and genetic polymorphism
  - Increased function : gene duplication
  - Decreased function : gene deletion, frameshift mutation, CNV

- Example : CYP2D6 – metabolizes 25-30% of drugs
  - Wild-type : CYP2D6***1** → Normal function
  - 46% of Asian population : CYP2D6***10** → Decreased function
  - 5% of Western population : CYP2D6* **x N** → Increased function
    - "Ultrarapid metabolizer, UM"

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

Metabolism Prediction by Transfer Learning



Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | NEWS & NOTES | ALERTS / RSS | CHANNELS

New Results

Comment on this paper

Transfer learning enables prediction of *CYP2D6* haplotype function

Gregory McInnes, Rachel Dalton, Katrin Sangkuhl, Michelle Whirl-Carrillo, Seung-been Lee, Philip S. Tsao, Andrea Gaedigk, Russ B. Altman, Erica L. Woodahl

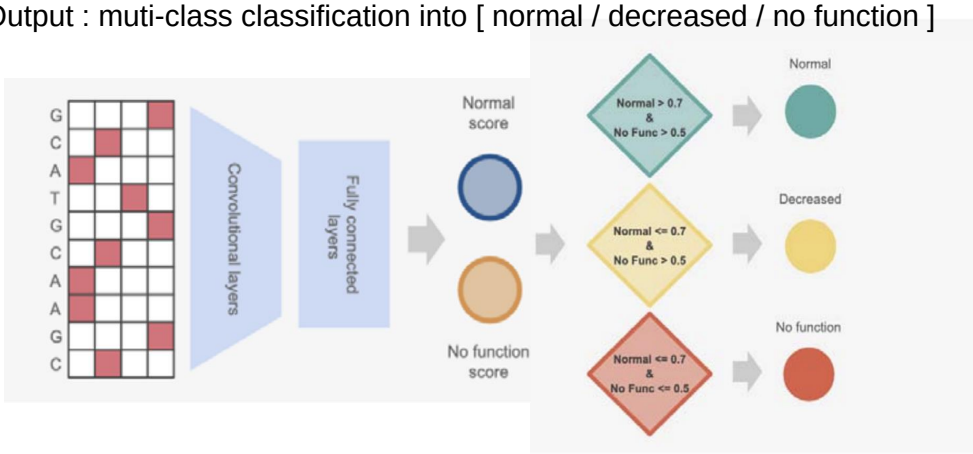
doi: https://doi.org/10.1101/684357

This article is a preprint and has not been certified by peer review [what does this mean?].

- bioRxiv, Feb 2020
  - Dept. of Biomedical Data Science, S tanford Univ
  - Dept of Biomedical Science, Univ of Montana
- Prediction of CYP2D6 function from DNA sequence with transfer learning

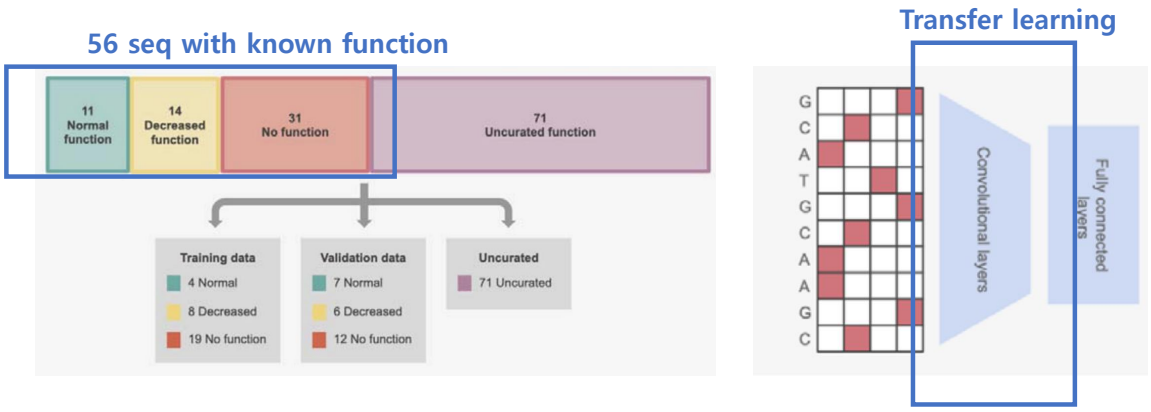
Metabolism prediction via deep learning

- Prediction of CYP2D6 function from DNA sequence
  - Input : One-hot encoded DNA sequence(4) + annotation data(8)
    - Total 7417 x 12 matrix
  - Output : muti-class classification into [normal / decreased / no function]



Metabolism prediction via deep learning

- Transfer learning improves prediction accuracy (40% → 88%)
  - Too small dataset : Only 56 sequence w/ curated functions
  - Applied transfer learning on 3-layer CNN



1. Simulated diploid's activity score classification 5만개의 simulated CYP2D6 diploid 조합 from 'gnomAD' database

input: simulated CYP2D6 diploid sequence → cnn → mlp

output: Activity score **classification**
- 0
- 0.5
- 1
- 1.5
- 2

2. Liver microsome sample activity regression 314개의 실제 liver microsome sample data

input: microsome CYP2D6 diploid sequence → cnn → mlp

output: Dexamethasone metabolism rate **regression**
- 0 ~ 1

3. CYP2D6 genome sequence data activity classification

input: CYP2D6 haploid sequence → cnn → mlp

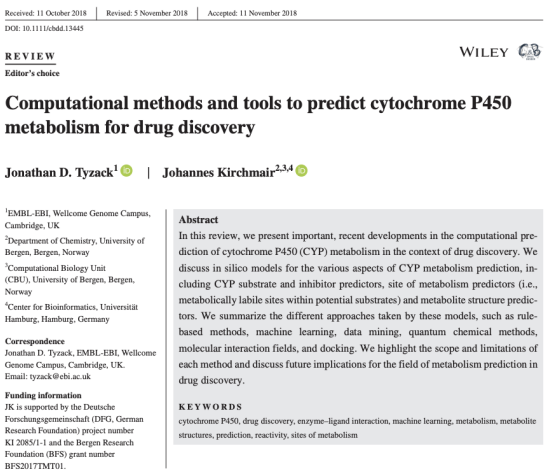
output: Metabolism activity **classification**
- Increased
- Normal
- Decreased

Slide made by Dongmin Bang

Sumamry:
Metabolism prediction via deep learning

- Transfer learning improves prediction accuracy (40% → 88%)
  - Step 1 : predicting activity score of 50,000 simulated sequences
    - Created by introducing SNVs and INDELs on sites with low importance
  - Step 2 : regression model of predicting functional activity data from actual liver micros ome samples
    - Sequenced liver microsome data with measured metabolic activity
  - Weights from pretrained network are copied and used as starting weight

Survey: methods for prediction of metabolism



Received: 11 October 2018 | Revised: 5 November 2018 | Accepted: 11 November 2018
DOI: 10.1111/cbdd.13445

REVIEW
Editor's choice

WILEY

Computational methods and tools to predict cytochrome P450 metabolism for drug discovery

Jonathan D. Tyzack[1] | Johannes Kirchmair[2,3,4]

[1]EMBL-EBI, Wellcome Genome Campus, Cambridge, UK
[2]Department of Chemistry, University of Bergen, Bergen, Norway
[3]Computational Biology Unit (CBU), University of Bergen, Bergen, Norway
[4]Center for Bioinformatics, Universität Hamburg, Hamburg, Germany

Correspondence
Jonathan D. Tyzack, EMBL-EBI, Wellcome Genome Campus, Cambridge, UK.
Email: tyzack@ebi.ac.uk

Funding information
JK is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number KI 2085/1-1 and the Bergen Research Foundation (BFS) grant number BFS2017TMT01.

Abstract
In this review, we present important, recent developments in the computational prediction of cytochrome P450 (CYP) metabolism in the context of drug discovery. We discuss in silico models for the various aspects of CYP metabolism prediction, including CYP substrate and inhibitor predictors, site of metabolism predictors (i.e., metabolically labile sites within potential substrates) and metabolite structure predictors. We summarize the different approaches taken by these models, such as rule-based methods, machine learning, data mining, quantum chemical methods, molecular interaction fields, and docking. We highlight the scope and limitations of each method and discuss future implications for the field of metabolism prediction in drug discovery.

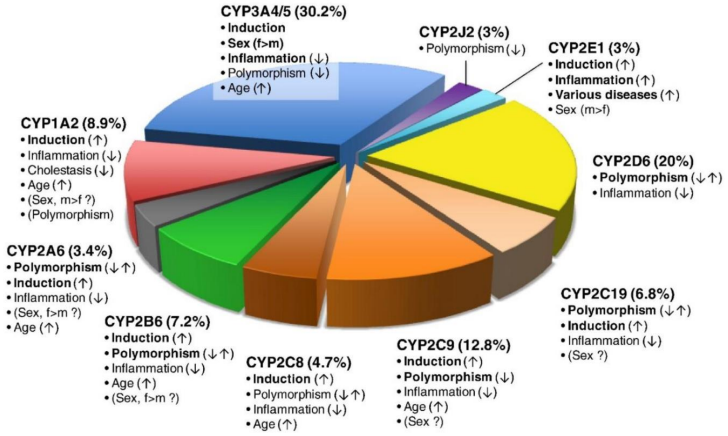
KEYWORDS
cytochrome P450, drug discovery, enzyme–ligand interaction, machine learning, metabolism, metabolite structures, prediction, reactivity, sites of metabolism

- *Chem Biol Drug Des*. 2019 Apr
  - Wellcome Genome Campus, Cambridge, UK

- **3 categories of prediction tools** :
  - Prediction of substrates/inhibitors of CYP
  - Site of Metabolism(SoM) prediction
  - Metabolite structure prediction

Survey: methods for prediction of metabolism

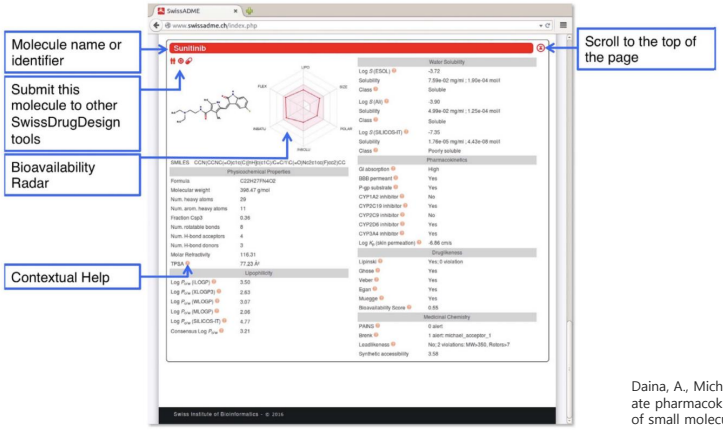
- 1) Prediction of substrates/inhibitors of CYP
  - Understanding the specificity of individual CYP isoforms
  - Assists prediction of Drug-Drug Interactions(DDI)



Olubadewa A. Fatunde et al, The Role of CYP450 Drug Metabolism in Precision Cardio-Oncology, *Int. J. Mol. Sci.* 2020 Feb

Survey: methods for prediction of metabolism

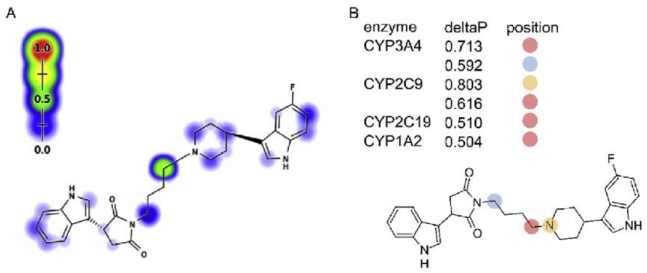
- 1) Prediction of substrates/inhibitors of CYP
  - Docking methods – 3D modeling of substrate binding
  - Machine learning methods based on enzyme-substate/inhibitor interaction data
    - SwissADME(Daina et al. 2017) : web-based SVM model for prediction of CYP inhibitors



Daina, A., Michielin, O. & Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci Rep.* 2017 Mar

Survey: methods for prediction of metabolism

- 2) Site of Metabolism(SoM) prediction
  - 2-step model : Reactivity prediction & Accessibility prediction

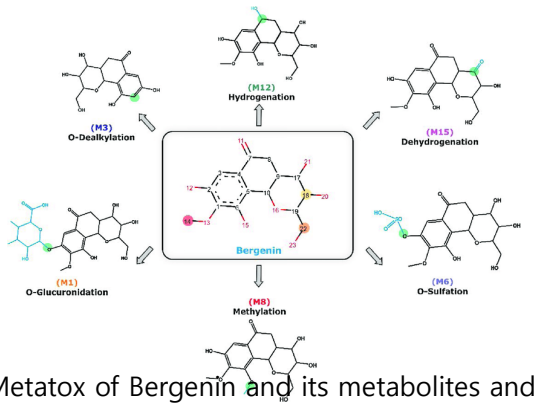


"Results for in silico SoM prediction for MW005"

Martyna Z Wróbel et al. Synthesis and biological evaluation of new multi-target 3-(1H-indol-3-yl)pyrrolidine-2,5-dione derivatives with potential antidepressant effect, *Eur J Med Chem.* Dec 2019

Survey: methods for prediction of metabolism

- 3) Metabolite structure prediction
  - MetaTox(Rudik et al., 2017) : Prediction of metabolite and estimates its **toxicity**
  - Meteor Nexus(Marchant et al., 2008) : SoM & metabolite prediction via k-nearest neighbor approach



Prediction in the Metatox of Bergenin and its metabolites and their respective chemical reactions.

Rudik et al., MetaTox: Web Application for Predicting Structure and Toxicity of Xenobiotics' Metabolites, *J Chem Inf Model.* 2017 Apr

Examples for Exploring Genetic (genome) Space

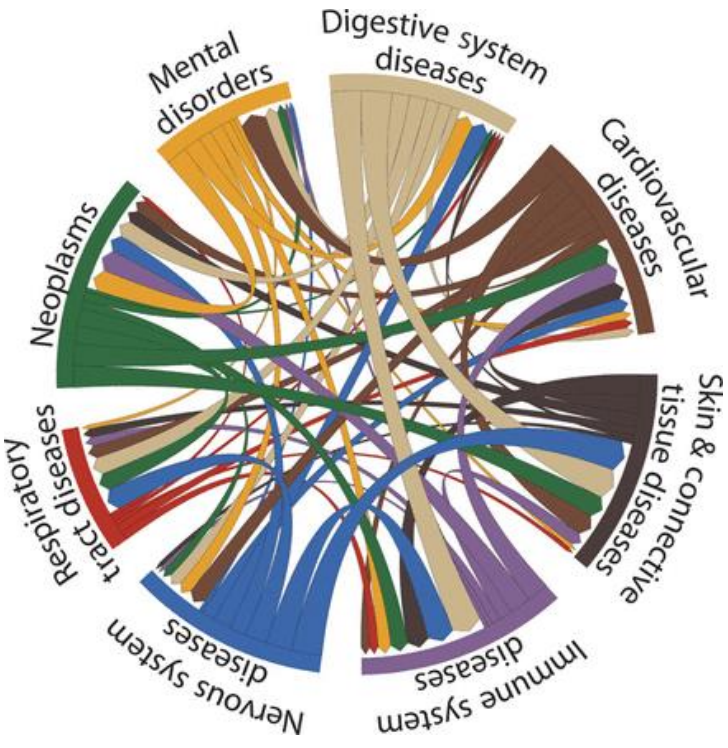
The druggable genome and support for target identification and validation in drug development

- *Science Translational Medicine* 29 Mar 2017

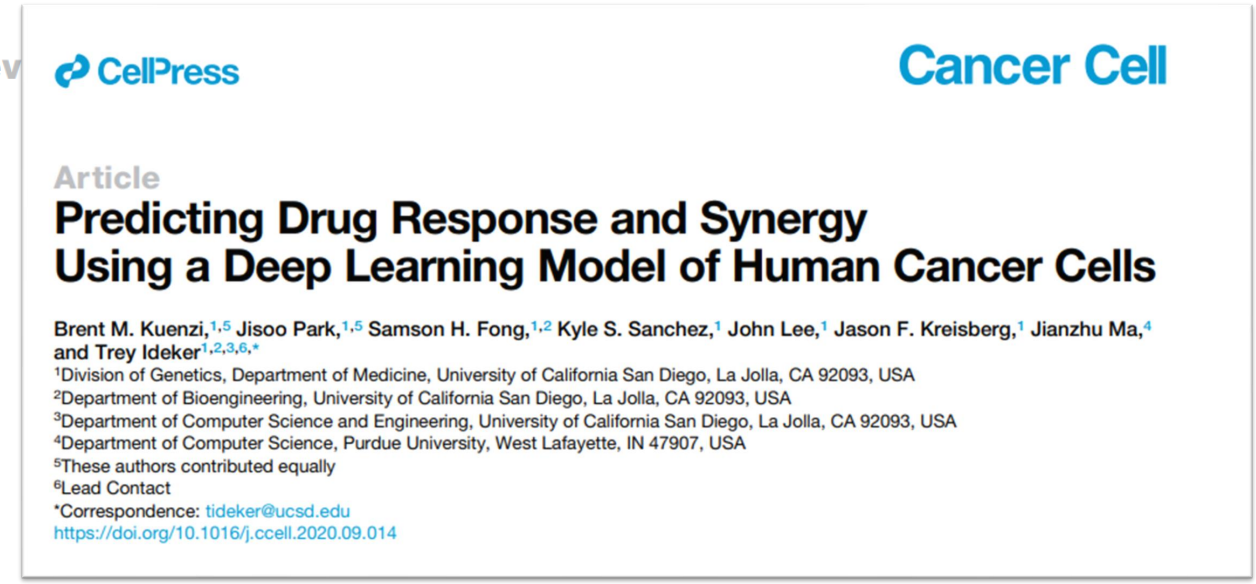
- **Abstract**
- Target identification (determining the correct drug targets for a disease) and target validation (demonstrating an effect of target perturbation on disease biomarkers and disease end points) are important steps in drug development. Clinically relevant associations of variants in genes encoding drug targets model the effect of modifying the same targets pharmacologically. To delineate drug development (including repurposing) opportunities arising from this paradigm, we connected complex disease- and biomarker-associated loci from genome-wide association studies to an updated set of genes encoding druggable human proteins, to agents with bioactivity against these targets, and, where there were licensed drugs, to clinical indications. We used this set of genes to inform the design of a new genotyping array, which will enable association studies of druggable genes for drug target selection and validation in human disease.

Potential repurposing opportunities from the discordant GWAS phenotype/drug indication matches

This connection is determined by a drug target gene occurring within 50 kbp of a GWAS association



Cancer Cell



CellPress Cancer Cell

Article

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

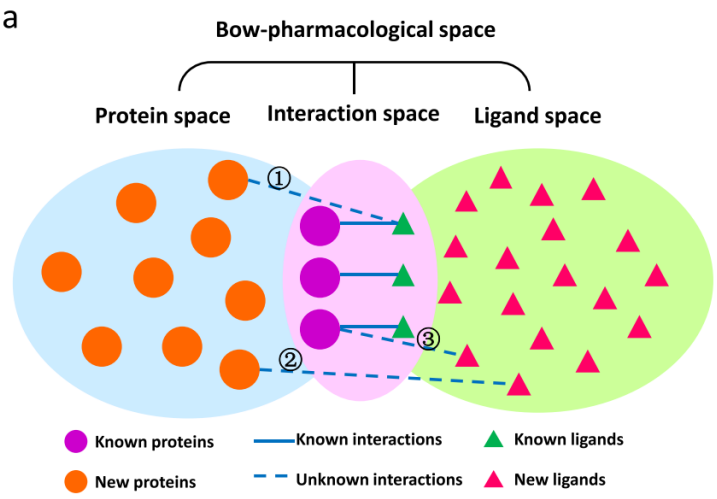
Brent M. Kuenzi,[1,5] Jisoo Park,[1,5] Samson H. Fong,[1,2] Kyle S. Sanchez,[1] John Lee,[1] Jason F. Kreisberg,[1] Jianzhu Ma,[4] and Trey Ideker[1,2,3,6,*]

[1]Division of Genetics, Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA
[2]Department of Bioengineering, University of California San Diego, La Jolla, CA 92093, USA
[3]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093, USA
[4]Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA
[5]These authors contributed equally
[6]Lead Contact
*Correspondence: tideker@ucsd.edu
https://doi.org/10.1016/j.ccell.2020.09.014

Example for Exploring Compound Space and Target Gene Space

Concept of ‘space’ – search space

a



Li, Li, et al. "Predicting protein-ligand interactions based on bow-pharmacological space and Bayesian additive regression trees." *Scientific reports* 9.1 (2019): 1-12.

Deep learning enables rapid identification of potent DDR1 kinase inhibitors

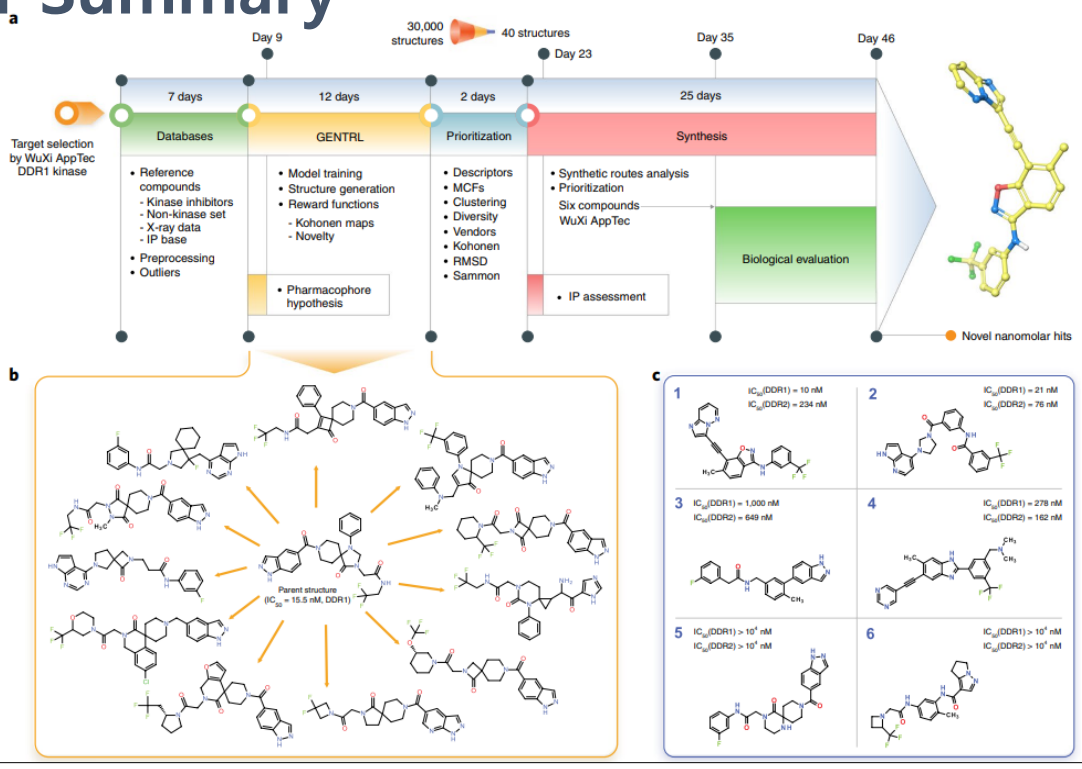
Nature Biotechnology 2019

111

Introduction

- To design DDR1 kinase inhibitor, used are
  - VAE with strong prior from tensor decomposition
  - RL with three SOMs (self organizing maps)

Brief Summary

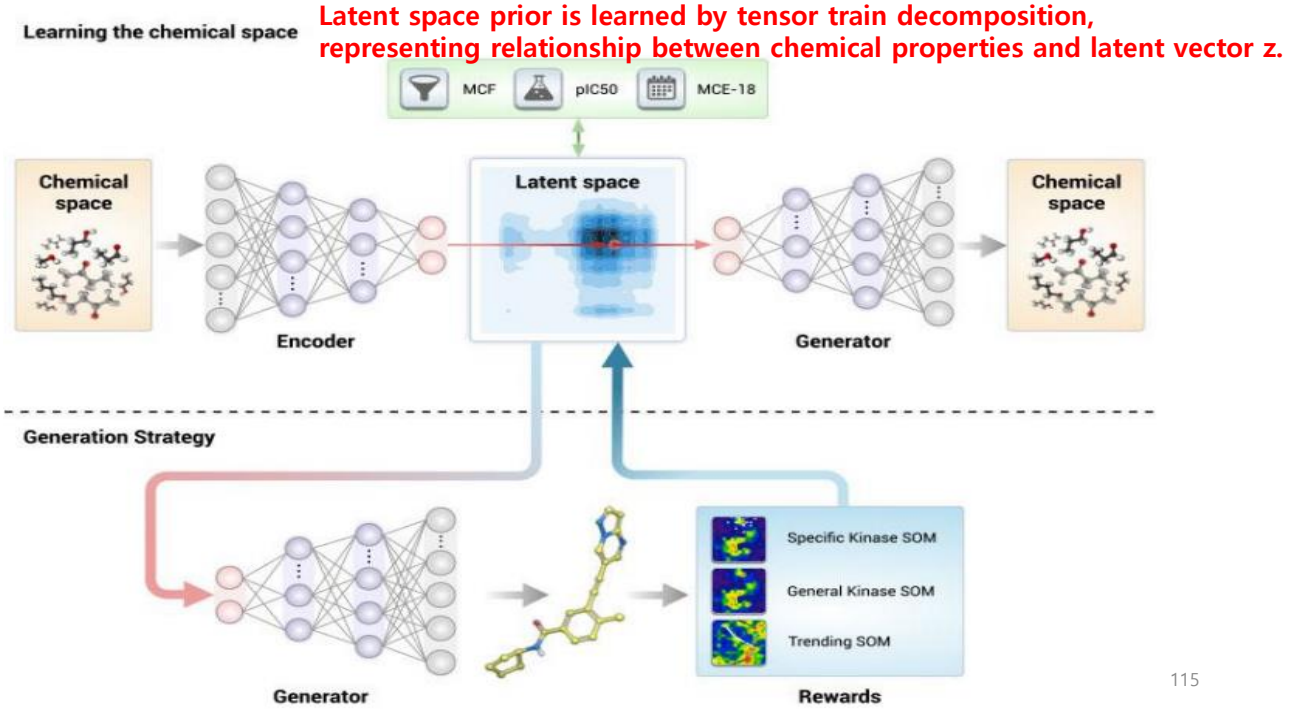


Database

- 1. Zinc Clean Leads collection(1,936,962 molecules)
  - molecular weight in the range from 250 to 350 Daltons, a number of rotatable bonds not greater than 7, and XlogP less than or equal to 3.5. We removed molecules containing charged atoms or atoms besides C, N, S, O, F, Cl, Br, H or cycles longer than 8 atoms. The molecules were filtered via medicinal chemistry filters (MCFs) and PAINS filters.(https://github.com/molecularsets/moses)
- 2. Known DDR1 Kinase Inhibitors
- 3. Common Kinase Inhibitors (positive)
- 4. Molecules that act on Non-Kinase Targets (negative)
  - 2~4: from ChemBL dataset
- 5. Patent Data for claimed molecules
  - www.globaldata.com 2017년까지 특허 등록된 약물17,000종
- 6. 3D Structure of DDR1 Inhibitors

114

Model Architecture



115

Reinforcement Learning

Reward function

1.**general kinase SOM**, $R_{general}$)
predict the activity of compounds against kinases

2.**specific kinase SOM,** $R_{specific}$)
select compounds located in neurons associated with DDR1 inhibitors within the whole kinase map

3.**trending SOM,** $R_{trending}$)
assess the novelty of chemical structures.

$$\max_{\psi} \mathbb{E}_{\mathbf{z}\sim p_{\psi}(\mathbf{z})} R(\mathbf{z}), \quad R(\mathbf{z}) = \mathbb{E}_{\mathbf{x}\sim p_{\theta}(\mathbf{x}|\mathbf{z})} [R_{\text{general}}(\mathbf{x}) + R_{\text{specific}}(\mathbf{x}) + R_{\text{trending}}(\mathbf{x})]$$

$$\nabla_{\psi} \mathbb{E}_{\mathbf{z}\sim p_{\psi}(\mathbf{z})} R(\mathbf{z}) = \mathbb{E}_{\mathbf{z}\sim p_{\psi}(\mathbf{z})} \nabla_{\psi} \log p_{\psi}(\mathbf{z}) \cdot \mathrm{R}(\mathbf{z})$$

For reducing the variance of the gradient
→using 'baseline' technique.
The rewards for each molecule in a batch are calculated and averaged, and the average reward is then subtracted from each individual reward:

$$\nabla_{\psi} \mathbb{E}_{\mathbf{z}\sim p_{\psi}(\mathbf{z})} R(\mathbf{z}) \approx \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \nabla_{\psi} \log p_{\psi}(\mathbf{z}_i) \left[R(\mathbf{z}_i) - \frac{1}{N} \sum_{j=1}^{N} R(\mathbf{z}_j) \right]$$

116

Experiments

- **Docking simulation**
  in the Maestro suite (https://www.schrodinger.com). PDB structure 3ZOS was preprocessed and energy minimized using the Prep module
- **In vitro activity assays**
  The activity of the molecules against human DDR1 and human DDR2 kinases was assessed using KinaseProfiler (Eurofins Scientific).
- **Cell-culture activity assay**
  To measure autophosphorylation, the gene encoding human DDR1b with a hemagglutinin tag was cloned into pCMV Tet-On vector (Clontech), and stable inducible cell lines established in U2OS were used for the IC50 test. DDR1 expression was induced for 48h before DDR1 activation by rat tail collagen I (Sigma 11179179001). The cells were detached with trypsinization and transferred to a 15 ml tube. Then after pretreatment with the compound for 0.5h, the cells were treated with compounds in the presence of 10μg ml−1 rat tail collagen I for 1.5h at 37 °C.

117

Experiments

- **Cell-culture fibrosis assay**

  MRC-5 or human hepatic LX-2 cells were grown in reduced serum medium and treated with compounds for 30minutes. Subsequently, the cells were stimulated with 10ng ml–1 or 4ng ml–1 TGF-β (R&D Systems, 240- B-002) for 48 or 72h. The cells were lysed in radioimmunoprecipitation assay buffer and cell lysate of each sample was loaded onto a Wes automated western blot system (ProteinSimple, a Bio-Techne brand).

- Cytochrome inhibition
- Microsomal stability
- Pharmacokinetic studies
- Statistics and reproducibility

118

Example for Exploring Compound Space and Target Gene Space

(3 different approaches are reviewed)

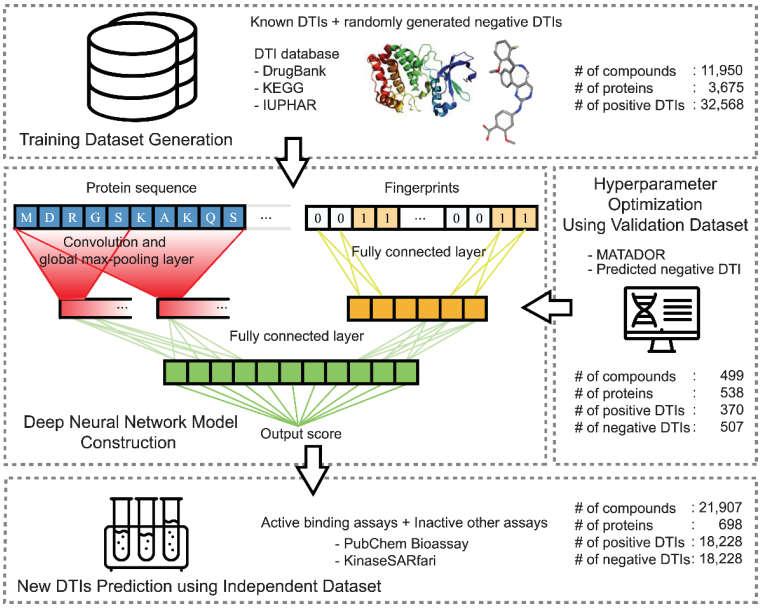
DeepConv-DTI
Prediction of drug-target interactions via deep learning with convolution on protein sequences

Ingoo Lee ,Jongsoo Keum ,Hojung Nam
PLoS Computational Biology, 2019

DeepConvDTI

Prediction of drug-target interaction via deep learning with convolution on protein sequences

- **Framework**:
  - CNN:
    - protein encoding
    - capture local residue patterns
    - globally max pooling
  - FC layer
    - drug encoding
- **Input**:
  - Protein: sequence
  - Drug: Morgan(Circular) fingerprints
- **Output**:
  - Interaction probability
- **Hyper parameters**:
  - Protein input: 2500
  - Drug input: 2048
  - protein layer size: 128
  - drug layer size: 128
  - CNN window size: 5,10,15,20,25



Data for DeepConvDTI

Source: *DrugBank, KEGG, IUPHAR, MATADOR, PubChem, CHEMBL*			
	# of DTIs	# of compounds	# of proteins
Training	Union **DrugBank, KEGG, IUPHAR** 32,568 positive 65,136 negative	11,950	3,657
Validation	**MATADOR**: 370 positive **Liu et al.** : 507 negative	499	538
Test	**PubChem Bioassays**: 18,228 positive + 18,228 negative **CHEMBL KinaseSARfari**: 3,835 positive + 5,520 negative	PubChem: 21,907 KinaseSARfari: 3,379	PubChem: 698 KinaseSARfari: 389

DeepConvDTI

Data:

Source: *DrugBank, KEGG, IUPHAR, MATADOR, PubChem, CHEMBL*			
	# of DTIs	# of compounds	# of proteins
Training	Union **DrugBank, KEGG, IUPHAR** 32,568 positive 65,136 negative	11,950	3,657
Validation	**MATADOR**: 370 positive **Liu et al.** : 507 negative	499	538
Test	**PubChem Bioassays**: 18,228 positive + 18,228 negative **CHEMBL KinaseSARfari**: 3,835 positive + 5,520 negative	PubChem: 21,907 KinaseSARfari: 3,379	PubChem: 698 KinaseSARfari: 389

Direct and indirect interactions between protein and chemical

DeepConvDTI

Data:

Source: *DrugBank, KEGG, IUPHAR, MATADOR, PubChem, CHEMBL*			
	# of DTIs	# of compounds	# of proteins
Training	Union **DrugBank, KEGG, IUPHAR** 32,568 positive 65,136 negative	11,950	3,657
Validation	**MATADOR**: 370 positive **Liu et al.** : 507 negative	499	538
Test	**PubChem Bioassays**: 18,228 positive + 18,228 negative **CHEMBL KinaseSARfari**:	PubChem: 21,907 KinaseSARfari: 3,379	PubChem: 698 KinaseSARfari: 389

Highly negative credible samples

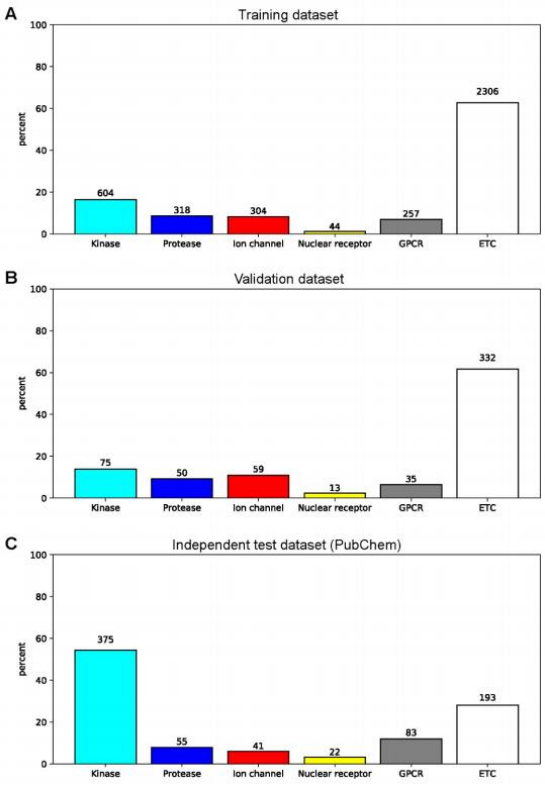
Improving compound–protein interaction prediction by building up highly credible negative samples

Hui Liu[1,2,†], Jianjiang Sun[3,†], Jihong Guan[4], Jie Zheng[2] and Shuigeng Zhou[3,*]

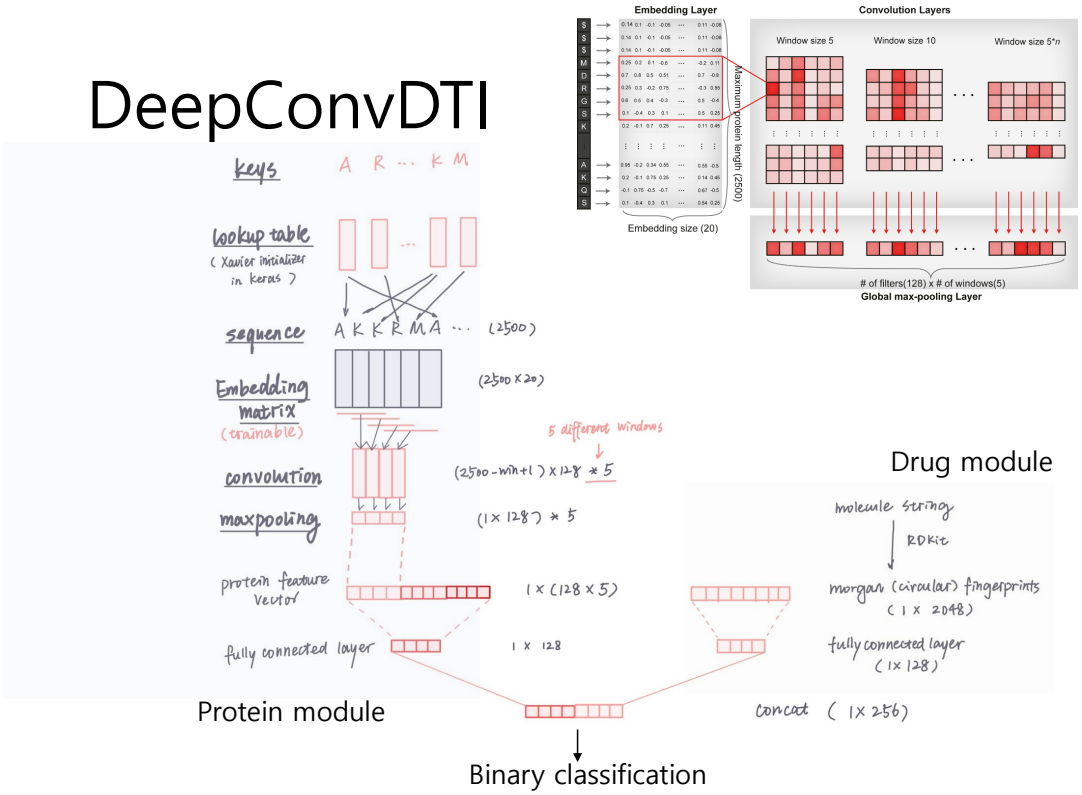
DeepConvDTI

Data:

- **Protein classes distribution**



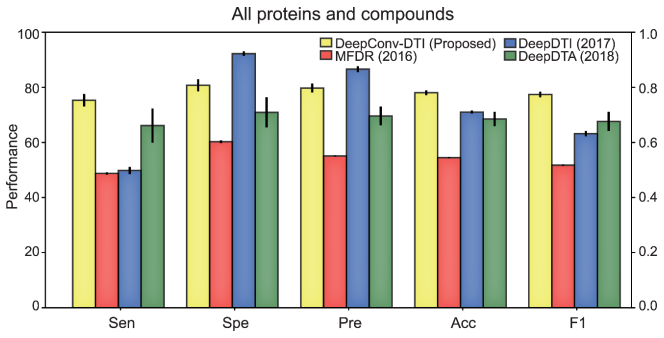
DeepConvDTI



DeepConvDTI

Performance:

- Comparison with other models
  - Multi-scale Features Deep Representation (MFDR) (based on SAE)
  - DeepDTI (based on DBN)
  - DeepDTA (based on CNN)
- 18,228 positive + 18,228 negative; 21,907 compounds, 698 proteins



Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences

Masashi Tsubaki, Kentaro Tomii, Jun Sese
Bioinformatics, 2018

Tsubaki et al.

Data

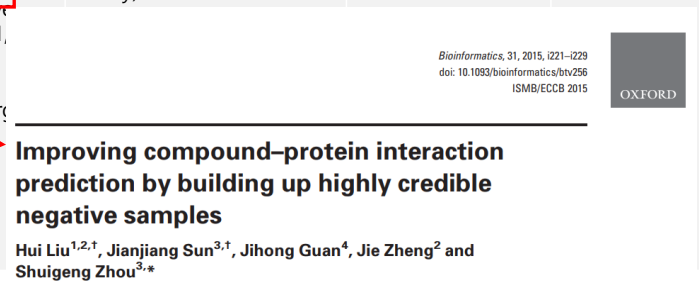
Source: *Liu et al. (from DrugBank, MATADOR), DUD-E*				
	# of DTIs		# of compounds	# of proteins
Training	**Human:** 3,369 positive **C.elegans:** 4,000 positive positive : negative = 1:1, 1:3, 1:5 **DUD-E:** DTIs for 102 targets	For **Human, C.elegans**: 80% of the dataset For **DUD-E**: DTIs for 72 targets (22,886 active + 22886 decoy)	**Human:** 1,052 **C.elegans:** 1,434	**Human:** 852 **C.elegans**: 2,504
Validation		10% of the dataset		
Test		For **Human, C.elegans**: 10% of the dataset For **DUD-E**: DTIs for 30 targets		

Tsubaki et al.

Data

Source: *Liu et al. (from DrugBank, MATADOR), DUD-E*

	# of DTIs		# of compounds	# of proteins
Training	**Human:** 3,369 positive **C.elegans:** 4,000 positive positive : negative = 1:1, 1:3, 1:5 **DUD-E:** DTIs for 102 targets	For **Human, C.elegans**: 80% of the dataset For **DUD-E**: DTIs for 72 targets (22,886 active + 22886 decoy)		
Validation				
Test				

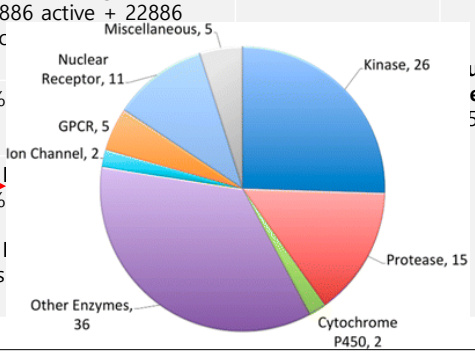


Tsubaki et al.

Data

Source: *Liu et al. (from DrugBank, MATADOR), DUD-E*

	# of DTIs		# of compounds	# of proteins
Training	**Human:** 3,369 positive **C.elegans:** 4,000 positive positive : negative = 1:1, 1:3, 1:5 **DUD-E:** DTIs for 102 targets	For **Human, C.elegans**: 80% of the dataset For **DUD-E**: DTIs for 72 targets (22,886 active + 22886 decc		
Validation		10%		
Test		For I 10% For I DTIs		

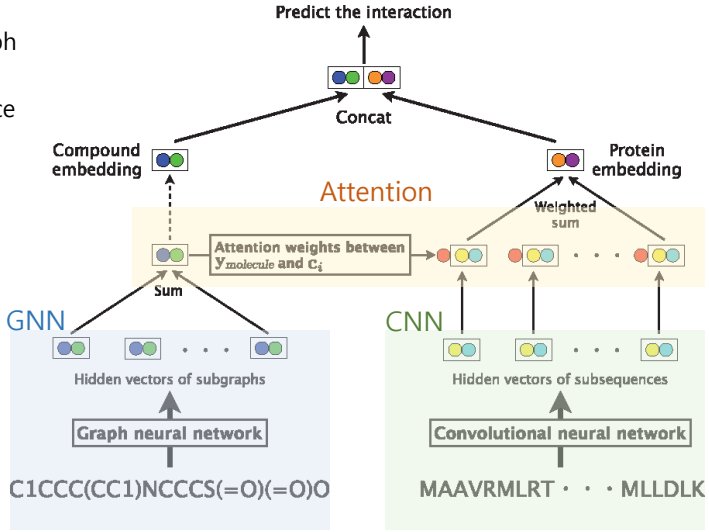


uman: 852
elegans: 504

Tsubaki et al.

Compound–protein interaction prediction with end-to-end learning of neural ne tworks for graphs and sequences

- **Framework**:
  - GNN
    - Encode compound graph
  - CNN
    - Encode protein sequence
  - Attention mechanism
    - Capture interaction site
- **Input**:
  - Protein: sequence
  - Compound: G = (V, E)
- **Output**:
  - GNN
    subgraph vector representation
  - CNN
    residue vector representation
  - Interaction probability



Tsubaki et al.

For molecule embedding

1) Consider the use of r-radius subgraph

$$\mathcal{V}_i^{(r)} = \{v_j \mid j \in \mathcal{N}(i,r)\},$$
$$\mathcal{E}_i^{(r)} = \{e_{mn} \in \mathcal{E} \mid (m,n) \in \mathcal{N}(i,r) \times \mathcal{N}(i,r-1)\}$$

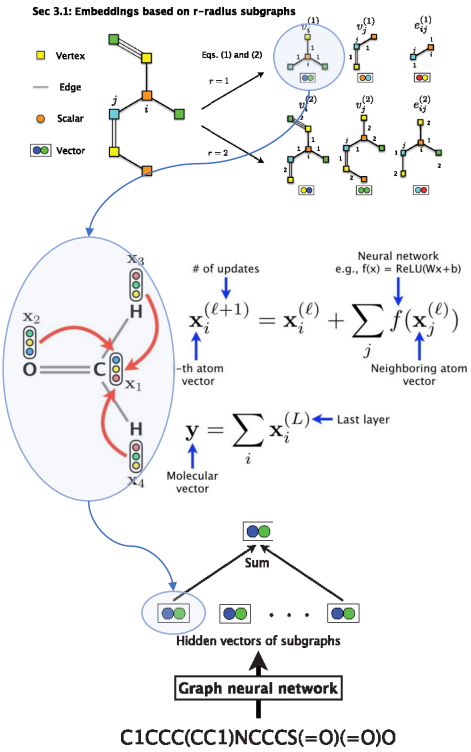
2) For each subgraph: update the vertex and edge vectors

$$\mathbf{v}_i^{(t+1)} = \sigma\left(\mathbf{v}_i^{(t)} + \sum_{j \in \mathcal{N}(i)} \mathbf{h}_{ij}^{(t)}\right)$$
$$\mathbf{e}_{ij}^{(t+1)} = \sigma\left(\mathbf{e}_{ij}^{(t)} + \mathbf{g}_{ij}^{(t)}\right)$$

3) Represent subgraph with summation of the hidden vectors of each vertices

4) Represent molecule with summation of the hidden vectors of each subgraphs

$$\mathbf{y}_{molecule} = \frac{1}{|\mathcal{V}|} \sum_{i=1}^{|\mathcal{V}|} \mathbf{v}_i^{(t)}$$

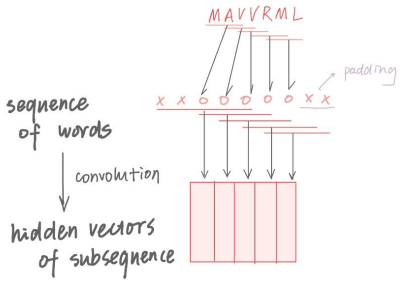


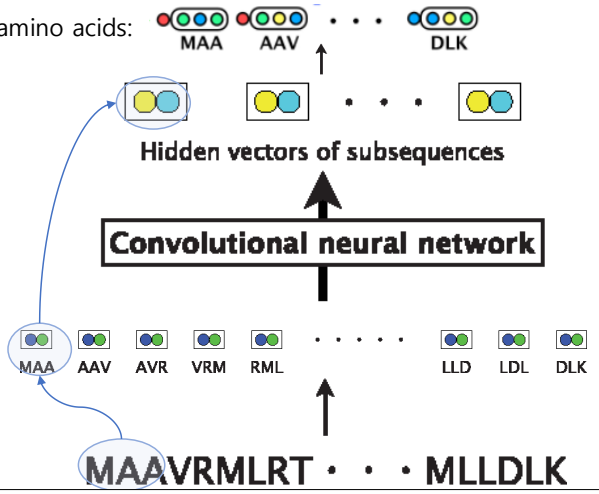
Tsubaki et al.

For protein embedding

1) Consider each 3-gram amino acids as a w
ord
$[\mathbf{x}_1;\mathbf{x}_2;\mathbf{x}_3],[\mathbf{x}_2;\mathbf{x}_3;\mathbf{x}_4],\ldots,[\mathbf{x}_{|\mathcal{S}|-2};\mathbf{x}_{|\mathcal{S}|-1};\mathbf{x}_{|\mathcal{S}|}]$

2) A … CNN … 3-gram amino acids:
$\mathbf{x}_{i:i+w-1} = \mathbf{c}_i^{(0)} \in \mathbb{R}^{dw}$
$\mathbf{c}_i^{(1)} = f(\mathbf{W}_{conv}\mathbf{c}_i^{(0)} + \mathbf{b}_{conv})$





Tsubaki et al.

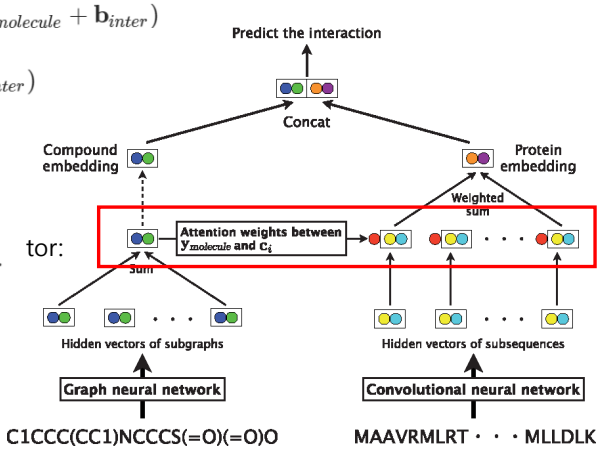
For capturing interaction sites (attention mechanism)

1) Compute the dot product values as weights:
For m $\mathbf{h}_{molecule} = f(\mathbf{W}_{inter}\mathbf{y}_{molecule} + \mathbf{b}_{inter})$

For res $\mathbf{h}_i = f(\mathbf{W}_{inter}\mathbf{c}_i + \mathbf{b}_{inter})$

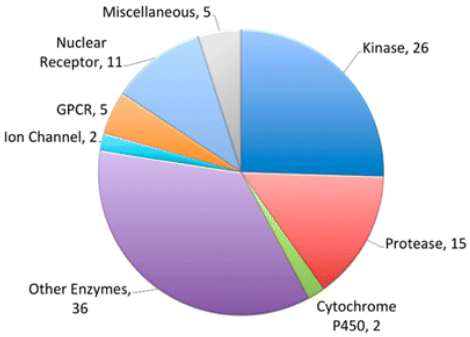
Calcula $\alpha_i = \sigma\left(\mathbf{h}_{molecule}^{\top}\mathbf{h}_i\right)$

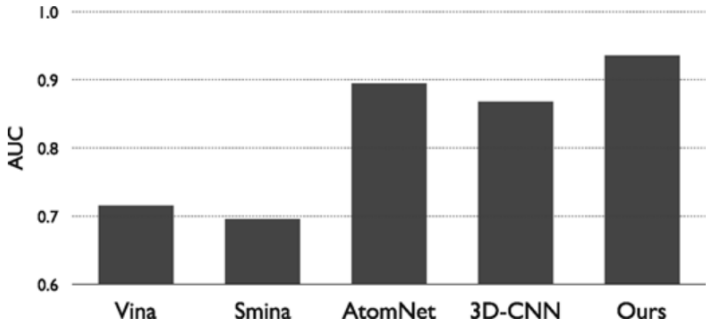
2) Calcula $\mathbf{y}_{protein} = \sum_{i=1}^{|\mathcal{C}|}\alpha_i\mathbf{h}_i$. tor:



Tsubaki et al.

Performance:



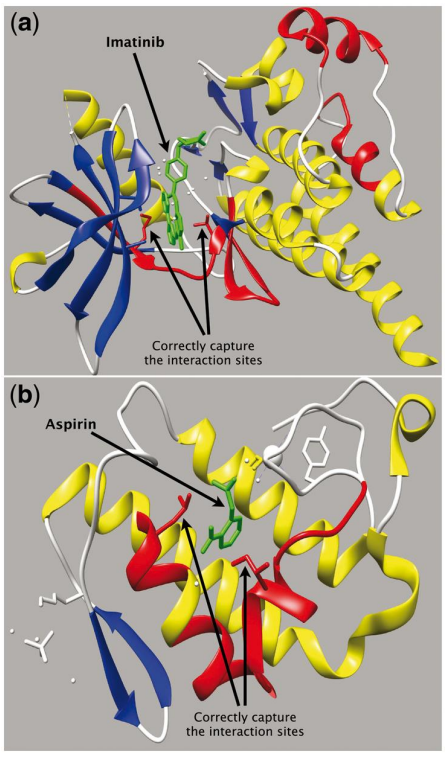


DUD-E

Tsubaki et al.

Explanation of attention module:

- visualization of CPIs with attention weights.
  - **Green**: drug compounds
  - **Red**: high weighted residue
- The neural attention mechanism can indicate important regions in a protein for interactions between a drug compound and a protein by highlighting high-value attention weights.



Graph Convolutional Neural Networks for Predicting Drug-Target Interactions

Wen Torng and Russ B. Altman

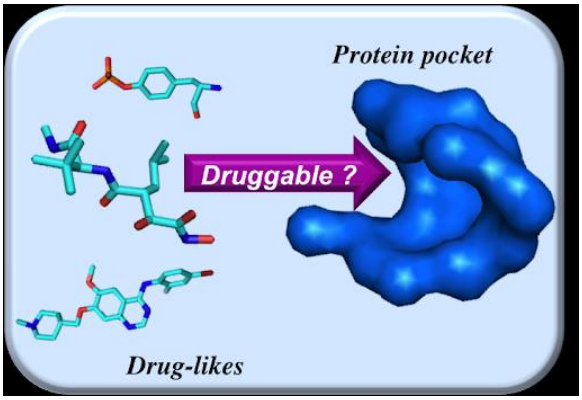
Department of Bioengineering, Stanford University, Stanford, California 94305, United States

Department of Genetics, Stanford University, Stanford, California 94305, United States

Yijingxiu Lu

Beforehand

- Protein pocket:
  - An important, ambiguous feature of protein surface that determine what interactions are possible with ligands and other macromolecules.

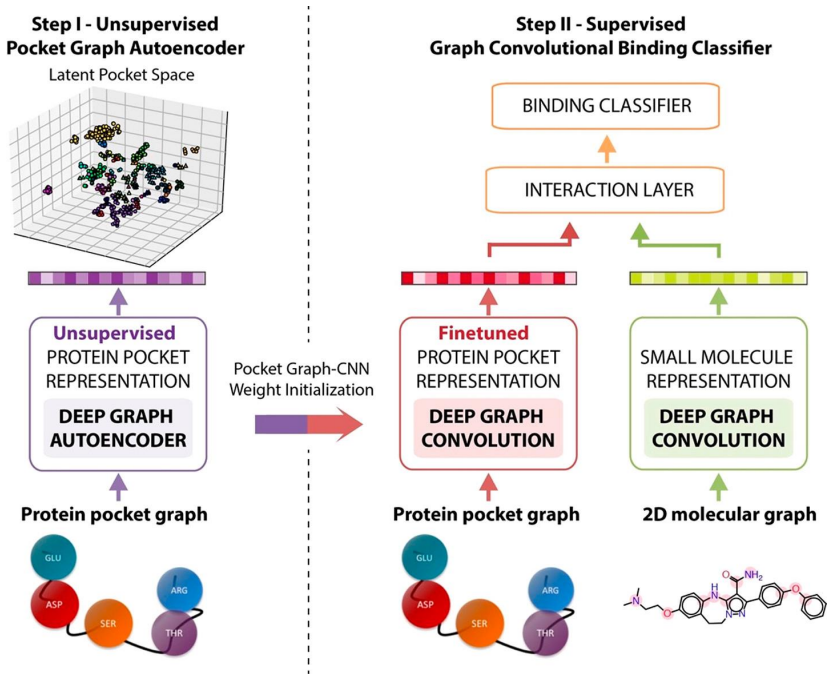


Summary

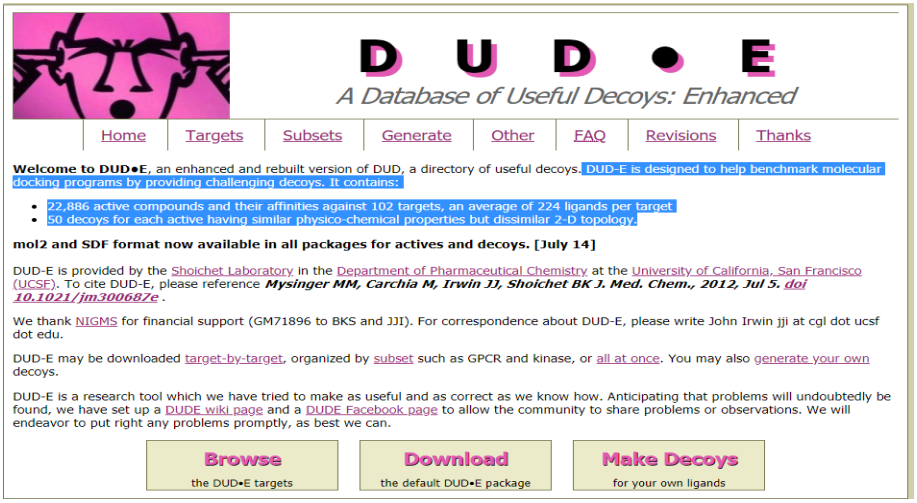
- Task:
  - Learn fixed-size representations of protein pockets
  - predict protein-ligand interactions
- Framework:
  - Graph Auto-encoder(AE): Trained on a representative druggable pocket set to lear n general pocket features
  - Graph-CNN: Trained to extract features from the pocket graphs and 2D ligand gra phs
- Datasets:
  - For training AE: 965 pockets (Druggable pocket sets + DUD-E (Database of Useful Decoys-Enhanced))
  - For training Graph-CNN: DUD-E data set
  - For validating Graph-CNN:
    - 4-fold validation model: DUD-E data set
    - Full data set model: maximum unbiased validation (MUV) data set

Framework

- Left: **Step I** - Graph Auto-enco der to generate general pocke t features
- Right: **Step II** - Graph CNN to predict binding label of a pair of drug-target
  - In the Graph-CNN framework, **Pocket GCN** and **Molecule G CN** are constructed to extract features from pocket graph a nd molecule graph separately.
  - Pocket GCN is initialized with learnt weights from Step I



DUD-E



- Contains 22,886 active compounds and their affinities against 102 targets
- An enhanced and rebuilt version of DUD
  - **DUD: Directory of Useful Decoys**
  - Decoy: a compound that has similar physical properties but dissimilar topology with an active compound against a target
- On average, each target has 224 actives (positive examples) and over 10,000 decoys (negative exa mples)

http://dude.docking.org/

Maximum Unbiased Validation (MUV) data set

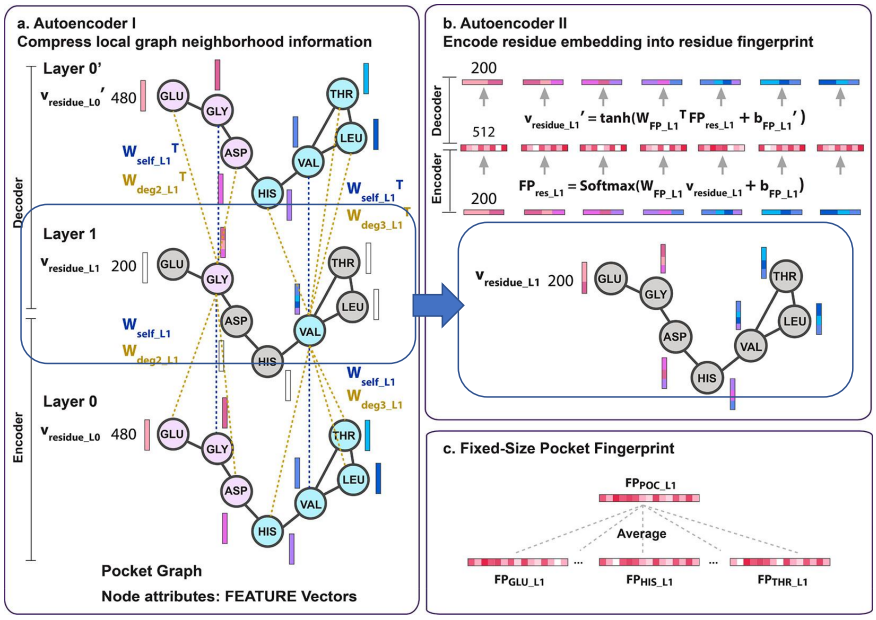
- Benchmark data sets that designed using "refined nearest neigh bor" analysis for virtual screening based on PubChem bioactivity data
- Contains a collection of datasets of actives and corresponding decoy datasets
- Unbiased with regard to both analogue bias and artificial enrich ment
  - 17 active classes, 93,087 molecules

Input Featurization & Preprocessing

- **Protein pockets**:
  - Represent each protein pocket as a graph of key residues
  - For each PDB co-crystal structure, identify the pocket residues by retrie ving residues that have any atom within 6 Å of the bound ligand
  - Node -> pocket residue
  - Node attribute -> local amino acid microenvironment that described as fix ed-size vectors generated using FEATURE program
  - Edge -> Distance between nodes

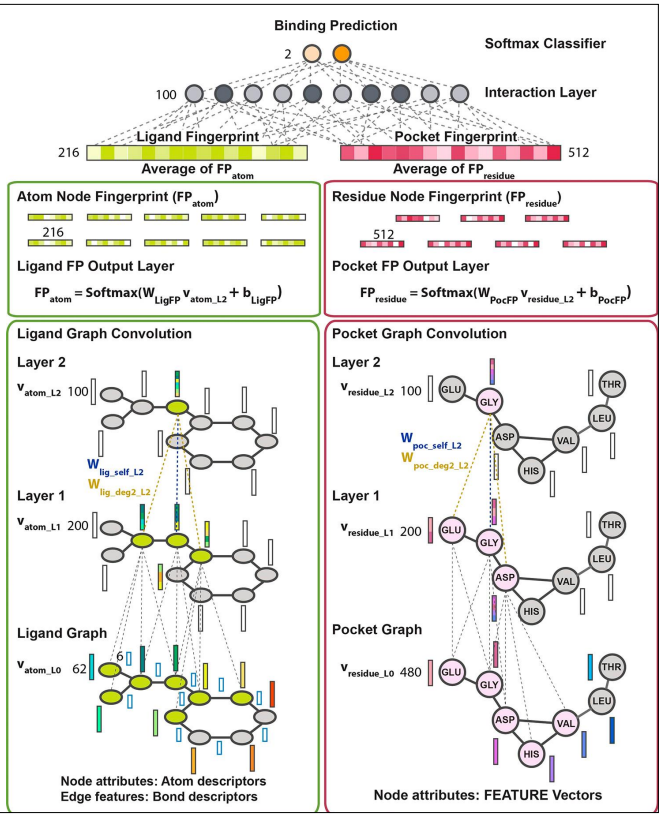
- **Small molecules**:
  - RDKit package
  - Node -> atom
  - Edge -> bond

Step I: Auto-encoder



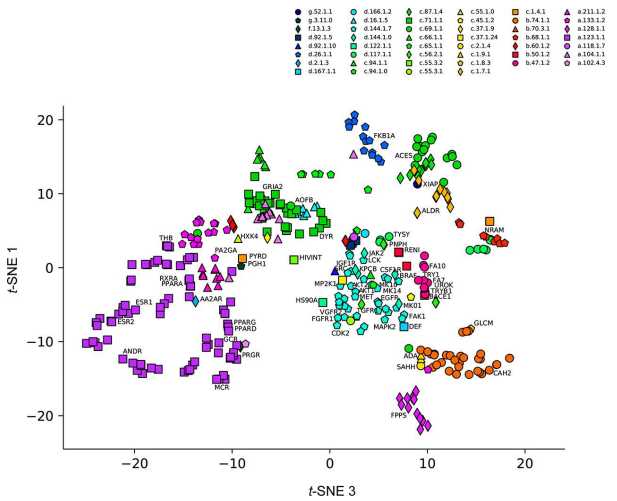
Step II: Graph-CNN

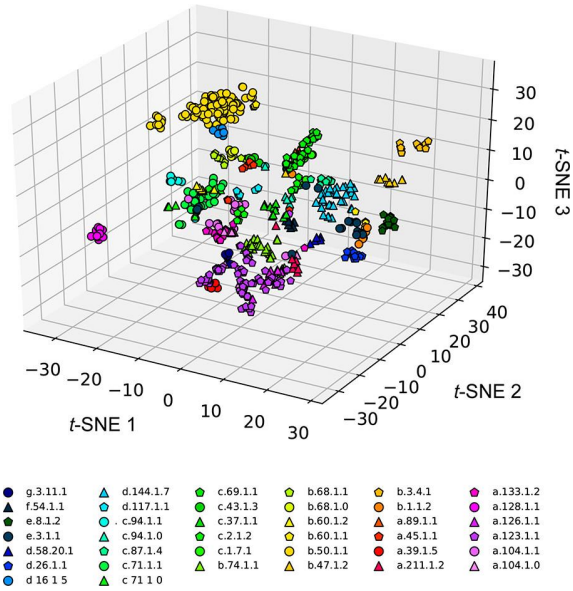
- Comprises
  - (1) Molecule graph convolution module
    - 2 GCN layers with 200, 100 filters
    - Learn molecular fingerprints of length 216
  - (2) Pocket graph convolution module
    - 2 GCN layers with 200, 100 filters
    - Learn pocket fingerprints of length 512
    - Has the same architecture of the "Encoder I" and "Encoder II" in Step I
  - (3) Interaction module
  - (4) Softmax classifier
- Adapting the "graph convolution operations" pr oposed by *Duvenaud et al.* to generalize the Gr aph-CNN framework onto graphs, and learn fix ed-size molecular fingerprints



Evaluation & Results

- 2D, 3D t-SNE visualization of fixed-size pocket fingerprints learned by Graph-AE





Example for Exploring Compound Space and Genetic Space

nature COMMUNICATIONS

De novo generation of hit-like molecules from gene expression signatures using artificial intelligence

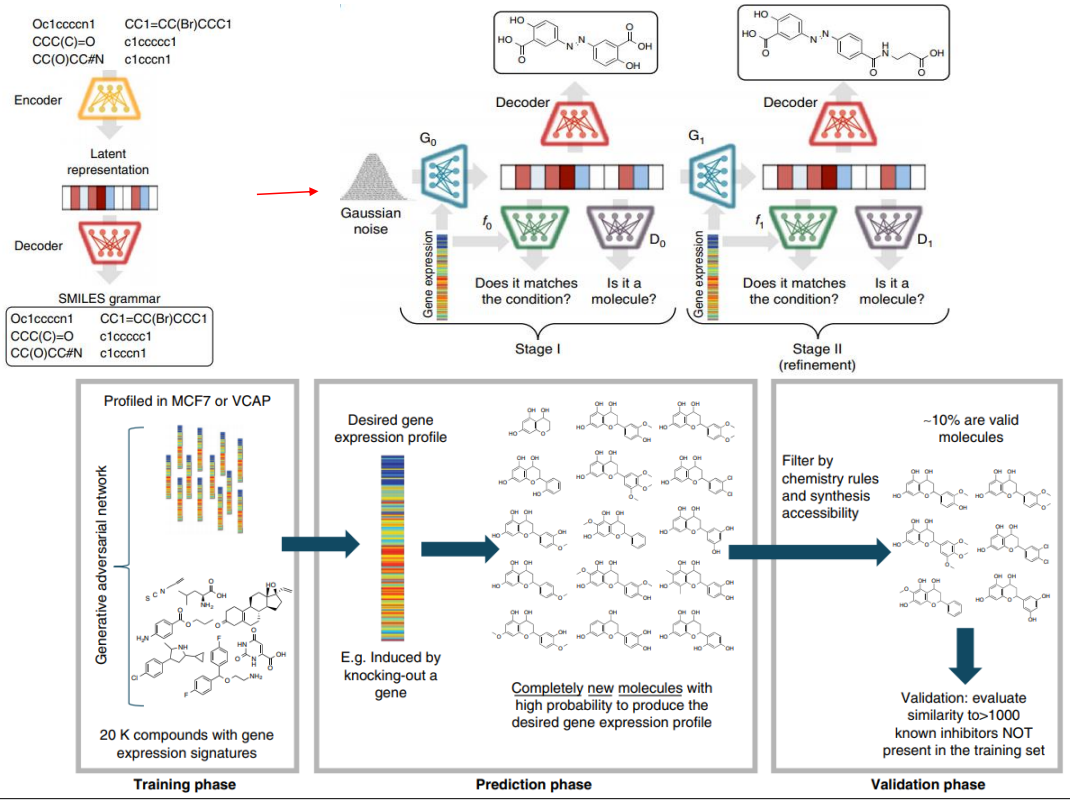
Oscar Méndez-Lucio1,2*, Benoit Baillif 1 , Djork-Arné Clevert3, David Rouquié1,5* & Joerg Wichard4,5*

1 Bayer SAS, Bayer Crop Science, 355 rue Dostoïevski, CS 90153, 06906 Valbonne, Sophia Antipolis Cedex, France.
2 Bloomoon, 13 Avenue Albert Einstein, 69100 Villeurbanne, France.
3 Department of Machine Learning Research, Bayer AG, 13353 Berlin, Germany.
4 Department of Genetic Toxicology, Bayer AG, 13353 Berlin, Germany.
5 These authors jointly supervised this work: David Rouquié, Joerg Wichard.

Introduction

- Connecting chemistry and biology through gene expression without the need of previous activity labels.
- Conditioning a GAN with transcriptomic data

Model



Datasets

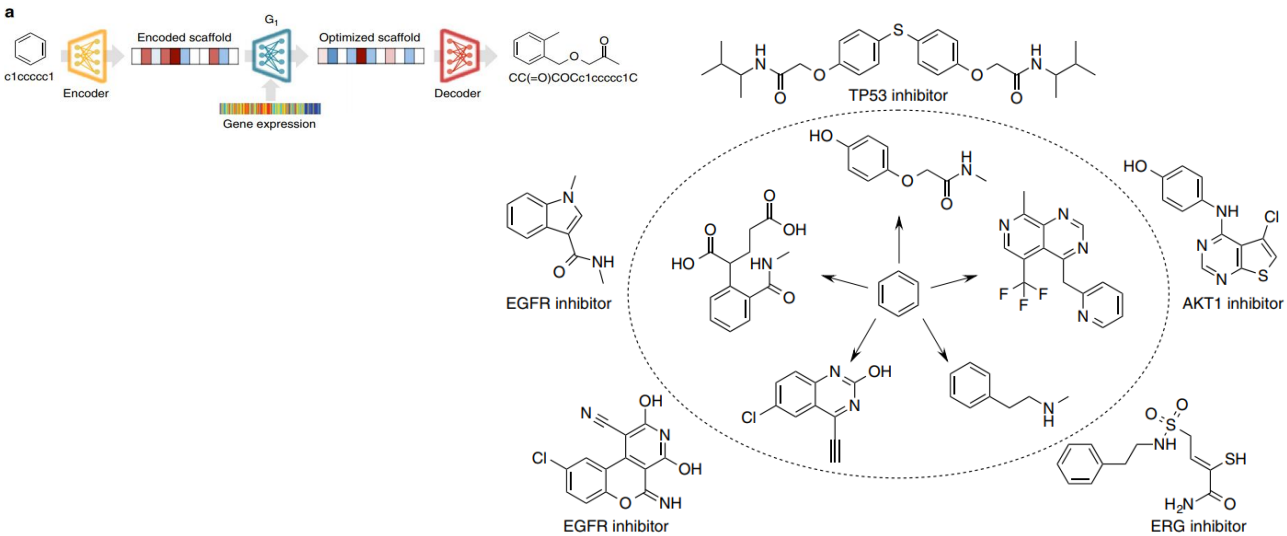
- L1000 from GEO
- LINCS: phase 1 (GSE92742), phase 2 (GSE70138)

- Landmark genes coming from perturbagens tested at 5 or 10 μM either on MCF7 or VCAP cell lines after 24 h of exposure.

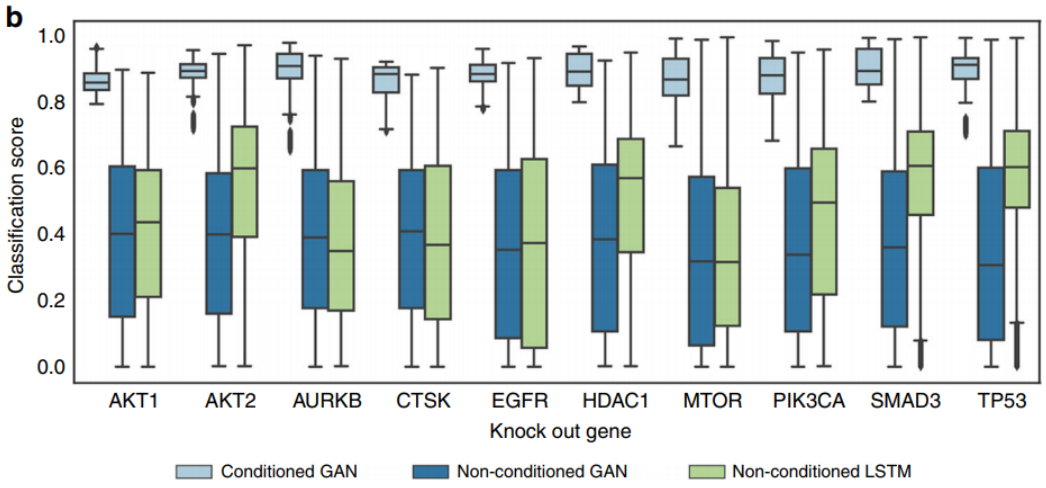
- 19,768 compounds and 31,821 gene expression signatures.

Experiments 3

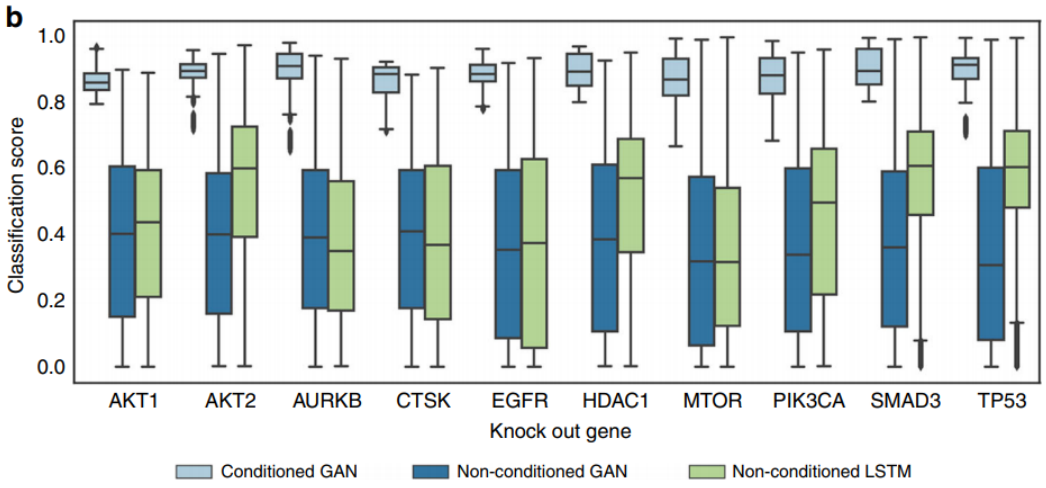
Conditioned GAN focus on specific areas of the chemical space.



Result



Result



Example for Exploring Compound Space and Genetic Space

DRIM: A web-based system for investigating drug response at the molecular level by condition-specific multi-omics data integration

Minsik Oh, Sungjoon Park, Sangseon Lee, Dohoon Lee, Sangsoo Lim, Dabin Jeong, Kyuri Jo, Inuk Jung, and Sun Kim

Frontiers in Genetics, 2020

Bio & Health Informatics Lab

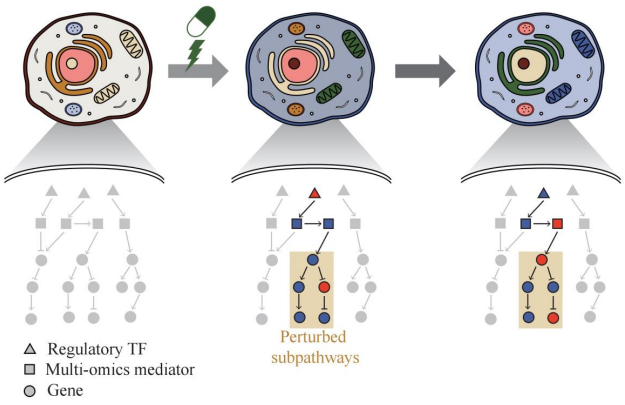
Introduction

- Pharmacogenomics is the study of how genes affect a person's response to drugs.
- The variability in drug responses among cells is a major challenge in cancer drug therapy, thus personalized drug response research is much needed.
- With the recent advances in instrument technologies, drug response analysis at the molecular level has become possible.
  - Multi-omics data before drug treatment with drug response (IC50 or AUC) : GDSC [1], CCLE [2], NCI-60 [3]
  - Time-series gene expression data after drug treatment : NCI TPW [4], NCI-DREAM [5]

- We have an opportunity to investigate relationship between drug response phenotypes and corresponding molecular data.

Introduction

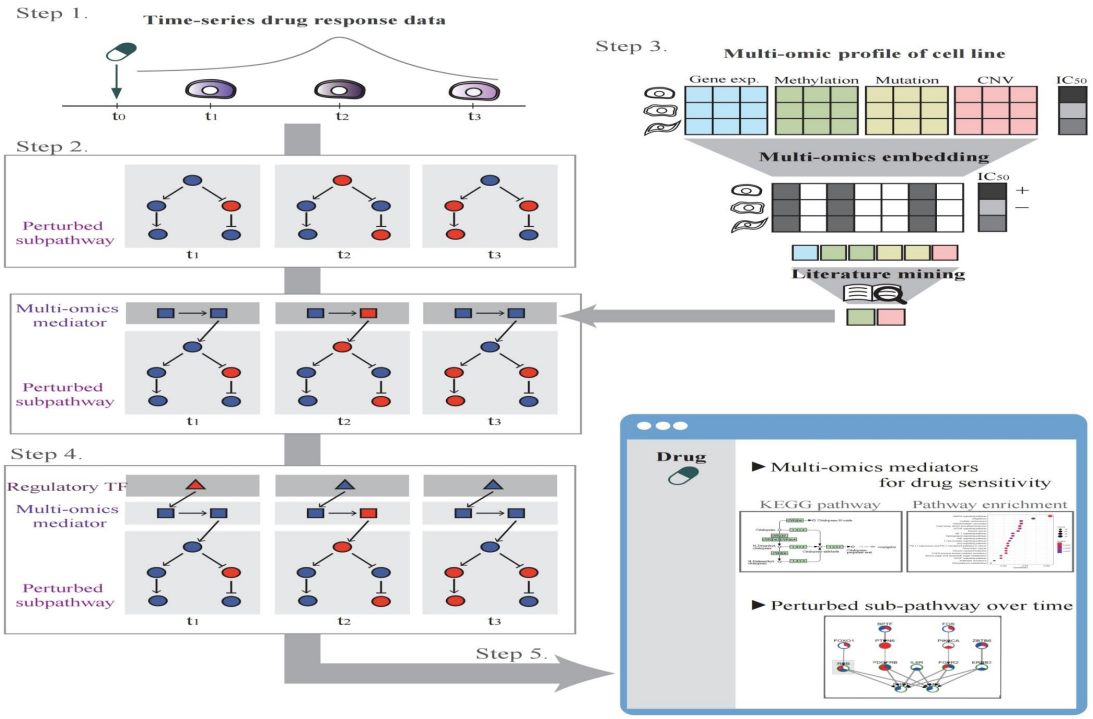
Drug response at the molecular level need to be done by

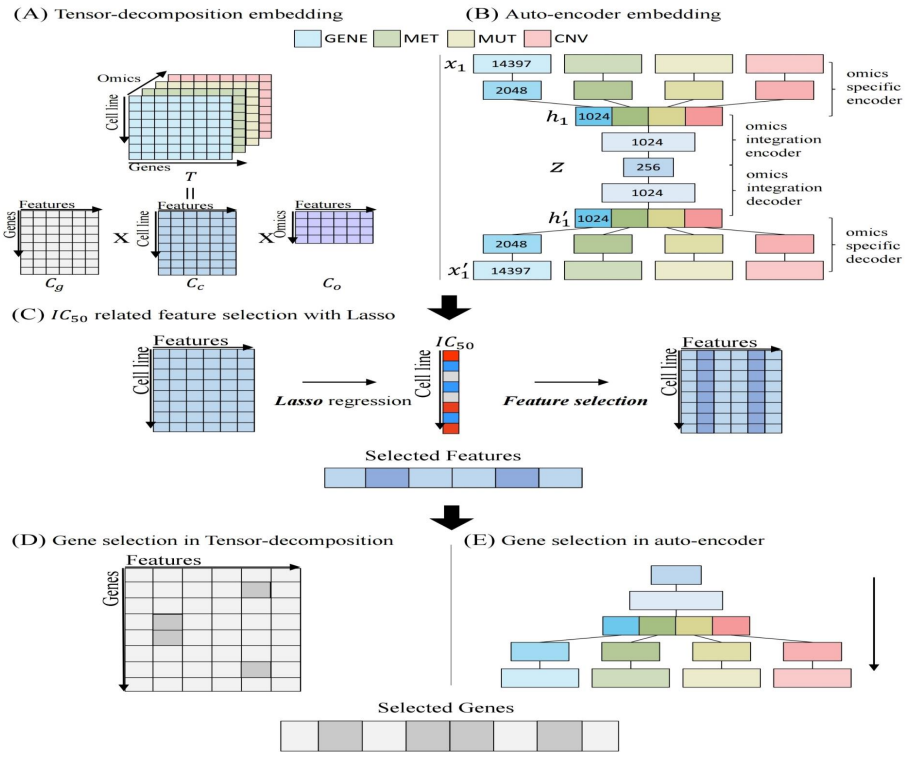


- Pathway-level analysis
  - Drug responses can be better explained at the biological pathway level, rather than at the gene level.

- Multi-omics level analysis
  - Integrative analysis of multi-omics data can help understand cell line-specific gene regulation mechanisms for pathway activation and it can be used as a signature for drug response sub-pathway identification

DRIM workflow





Multi-omics embedding

- DRIM integrates four multi-omics data before drug treatment as **gene-centric vector from CCLE database**.
- IC50 related features selection with Lasso regression
- Gene selection related to associated features.
  - Using (Gene x Features) matrix
  - Activation and propagation from selected features to the omics data later in the decoder.

Multi-omics data analysis result on the web

(A) Multi-omics analysis table

Cell line IC50 value

Cell line	IC50
BT549_BREAST	2.01
T47D_BREAST	2.9
MCF7_BREAST	3.03
MDAMB468_BREAST	3.76
MDAMB231_BREAST	6.5

Cell-line

Multi-omics potential mediator genes

Gene	Score
ERBB3	8.008
VEGFA	6.113
PGR	5.960
CDAN1	5.958
ABCG2	5.827

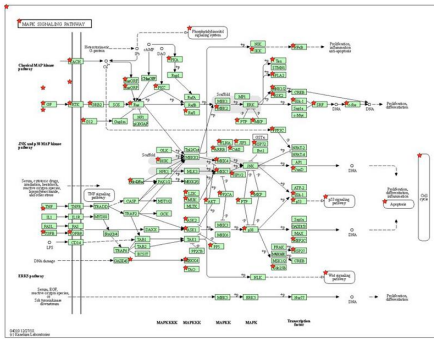
Mediator gene

Perturbed pathway

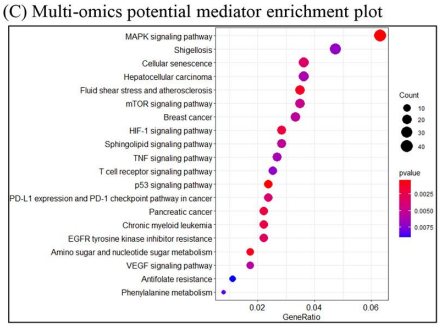
Perturbed pathway	P-value
p53 signaling pathway	0.005
Amino sugar and nucleotide sugar metabolism	0.011
T cell receptor signaling pathway	0.022
Chronic myeloid leukemia	0.025

Perturbed pathway

(B) KEGG-pathway plot



(C) Multi-omics potential mediator enrichment plot



- DRIM provides multi-omics data analysis tables (Figure A)
  - Cell line IC50 value
  - Multi-omics potential mediator genes
  - Perturbed pathway list with P-value
- KEGG-pathway plot for perturbed pathway (Figure B)
- Multi-omics potential mediator pathway enrichment plot (Figure C)

Time-series gene expression data analysis result

(A) Cell line & sub-pathway selector

Configure

Cellline: MDA-MB-231

Subpathway: EGFR tyrosine ki

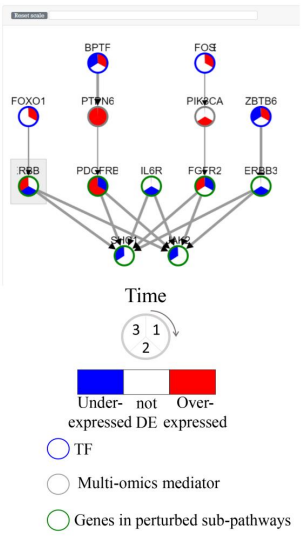
(B) Perturbed sub-pathways of cell line

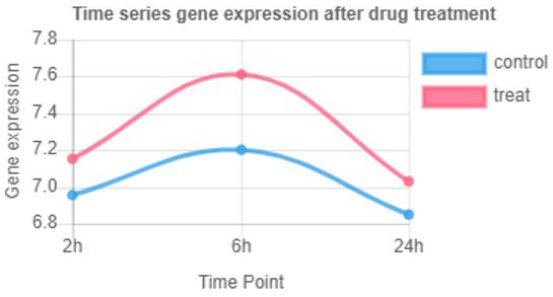
Perturbed sub-pathway

Perturbed sub-pathway	P-value
Hepatitis B	0
EGFR tyrosine kinase inhibitor...	0
ErbB signaling pathway	0
PPAR signaling pathway	0
Inflammatory	0

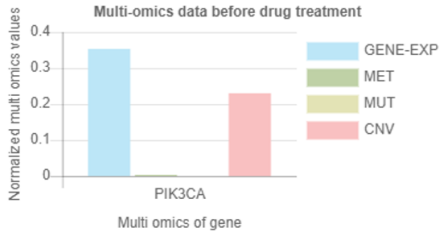
Perturbed sub-pathway

(C) Perturbed sub-pathway network



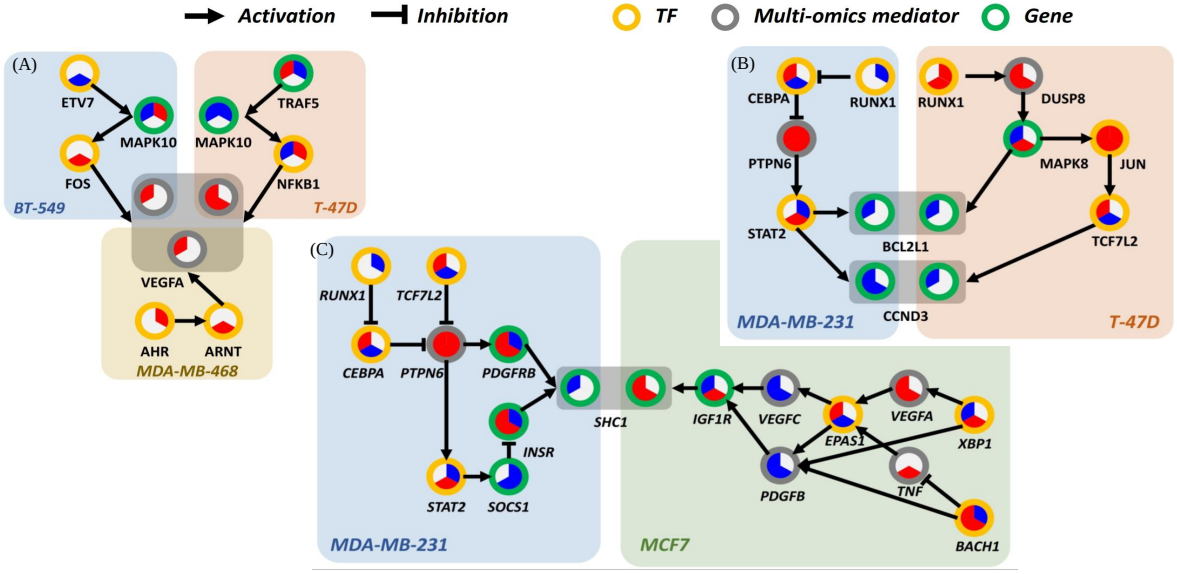


(D) Multi-omics data and time series gene expression data plot



Perturbed sub-pathway of each cell line.

Personalized perturbed sub-pathway analysis



Drug Discovery and Cells

NATURE REVIEWS | DRUG DISCOVERY
VOLUME 18 | JANUARY 2019

Drug repurposing: progress, challenges and recommendations

*Sudeep Pushpakom[1], Francesco Iorio[2], Patrick A. Eyers[3], K. Jane Escott[4], Shirley Hopper[5], Andrew Wells[6], Andrew Doig[7], Tim Guilliams[8], Joanna Latimer[9], Christine McNamee[1], Alan Norris[1], Philippe Sanseau[10], David Cavalla[11] and Munir Pirmohamed[1]**

Abstract | Given the high attrition rates, substantial costs and slow pace of new drug discovery and development, repurposing of 'old' drugs to treat both common and rare diseases is increasingly becoming an attractive proposition because it involves the use of de-risked compounds, with potentially lower overall development costs and shorter development timelines. Various data-driven and experimental approaches have been suggested for the identification of repurposable drug candidates; however, there are also major technological and regulatory challenges that need to be addressed. In this Review, we present approaches used for drug repurposing (also known as drug repositioning), discuss the challenges faced by the repurposing community and recommend innovative ways by which these challenges could be addressed to help realize the full potential of drug repurposing.

CONNECTIVITY MAP (CMAP)

HTTPS://WWW.BROADINSTITUTE.ORG/CONNECTIVITY-MAP-CMAP

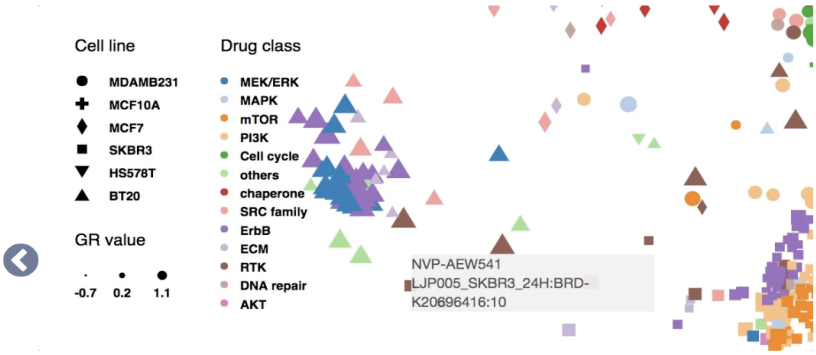
- To date, CMap has generated a library containing over 1.5M gene expression profiles from ~5,000 small-molecule compounds, and ~3,000 genetic reagents, tested in multiple cell types. To produce data of that scale, we've developed L1000, a relatively inexpensive and rapid high-throughput gene expression profiling technology. Expression data are processed through a computational pipeline that converts raw fluorescence intensity into signatures, which can be used to query the CMap database for perturbations that give a related gene expression response.
- Funding for our work comes from the NIH LINCS (Library of Integrated Cellular Signatures) project

The LINCS Consortium

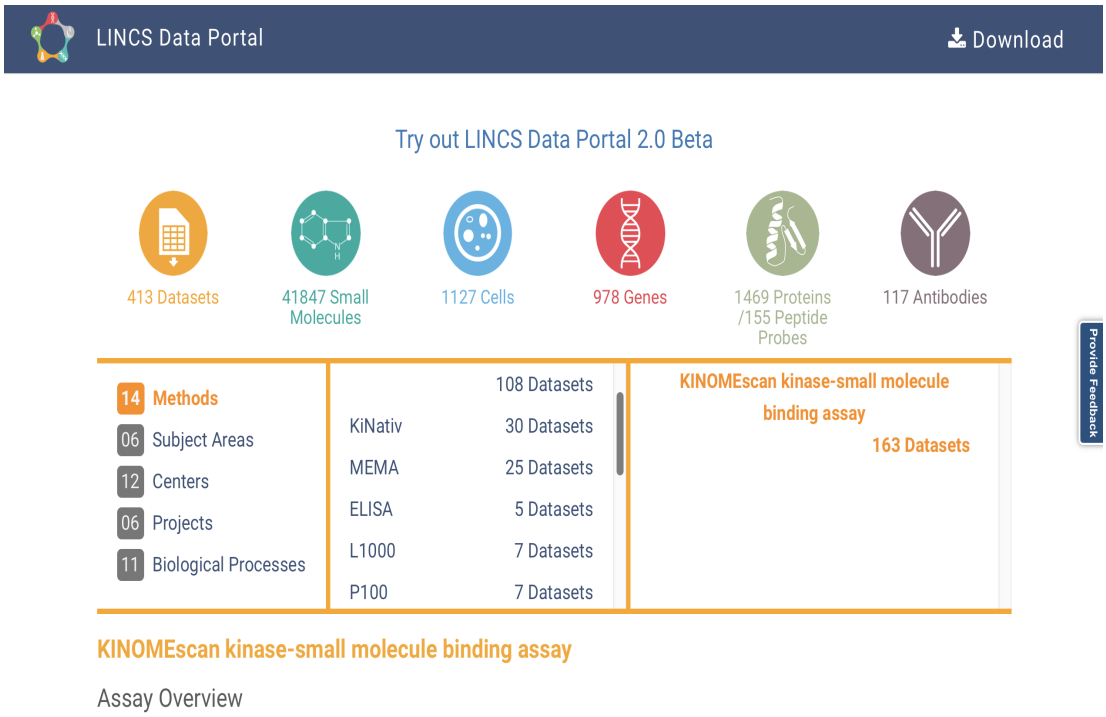
- By generating and making public data that indicates how cells respond to various genetic and environmental stressors, the LINCS project will help us gain a more detailed understanding of cell pathways and aid efforts to develop therapies that might restore perturbed pathways and networks to their normal states. The LINCS website is a source of information for the research community and general public about the LINCS project. This website along with the LINCS Data Portal contains details about the assays, cell types, and perturbagens that are currently part of the library, as well as links to participating sites, data releases from the sites, and software that can be used for analyzing the data.

Featured Interactive Data Visualization

LINCS Joint Project (LJP)
Breast Cancer Network Browser (BCNB)



In a recent study published in Nature Communications, the HMS LINCS Center, in collaboration with the LINCS Transcriptomics Center and the BD2K-LINCS DCIC, analyzed the gene expression and phenotypic response of six breast cancer cell lines to over a hundred drugs and pre-clinical small molecules. The perturbations were applied in different concentrations while gene expression was measured at different time points using the L1000 technology. Under the same conditions, the cells were imaged for cell viability. The LINCS Joint Project-Breast Cancer Network Browser (LJP-BCNB) is an interactive visualization of 2344 signatures created from this dataset



Drug–disease similarity approach to identify topiramate in IBD

- Dudley and colleagues compared the gene expression signature of inflammatory bowel disease (IBD) derived from publicly available data obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus with the gene expression profile of 164 drugs obtained from the Connectivity Map (cMap).
- Therapeutic predictions for drug–disease pairs were derived based on the extent of negative correlation between the gene expression signature of the drug and that of the disease.

Drug–drug similarity approach to identify the potential use of fasudil in amyotrophic lateral sclerosis

- Iorio and colleagues used the 'guilt by association' principle to construct a drug network using publicly available transcriptomic profiles of drugs, which allowed them to identify drugs with a similar transcriptional signature and therefore a perceived similar mechanism of action.
- Using gene expression profiles of each drug across multiple treatments on different cell lines and/or at different dosages obtained from the Connectivity Map (cMap), they computed a representative transcriptional response for each drug.
- A drug network was then constructed in which two drugs were connected to each other if their optimal transcriptional responses were similar according to a similarity measure developed by the authors (called drug distance).

Use of GWAS-identified targets for potential repurposing of denosumab in Crohn's disease

- This prompted Sanseau and colleagues to speculate about a potential role for denosumab in Crohn's disease.
- Using human B-lymphoblastoid cells and osteoblasts, they found that the Crohn's disease-associated *TNFSF11* variant was associated with the differential expression of *TNFSF11* and was able to explain population variation in *TNFSF11* expression in both cell types representing distinct cellular lineages relevant for both inflammatory and bone disease.

The challenges of big data

- Advances in technology such as next-generation sequencing and continuously reducing costs mean that researchers can generate large quantities of experimental data; these include data generated by high-throughput DNA and RNA sequencing, mass spectrometry, metabolomics and transcriptomic data, phenotyping and many more. Added to this are large amounts of clinical data that are increasingly becoming available from electronic health records (EHRs), clinical trials and biobanks. Such data are often referred to as big data — data sets that are so large or complex that traditional data processing methods are inadequate

Cancer Cell

Prev

CellPress

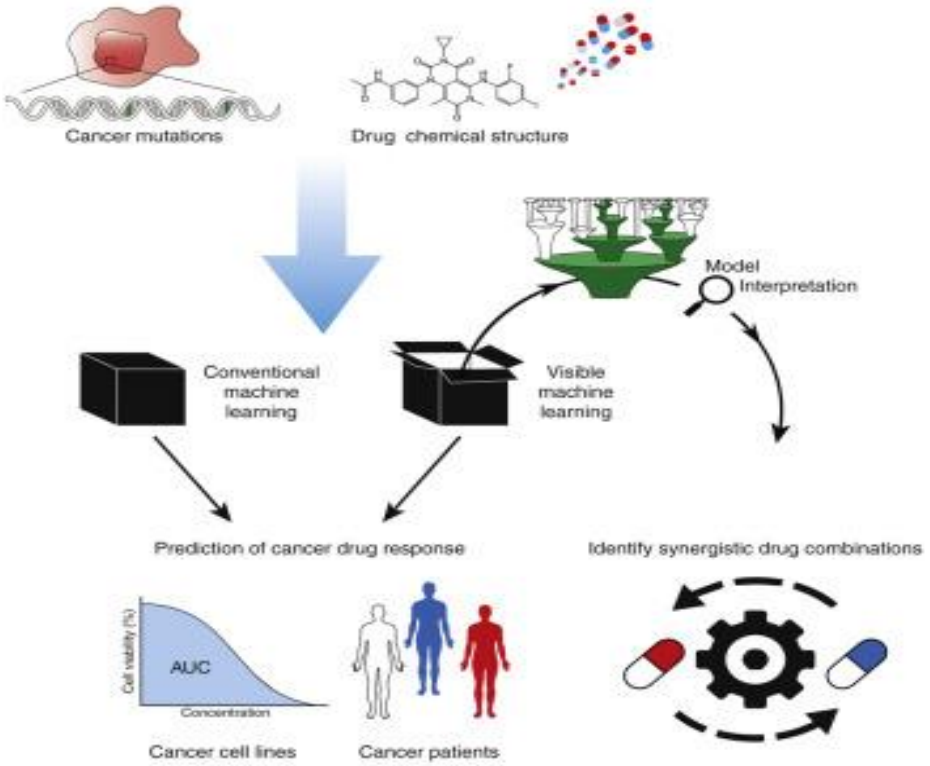
Cancer Cell

Article

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

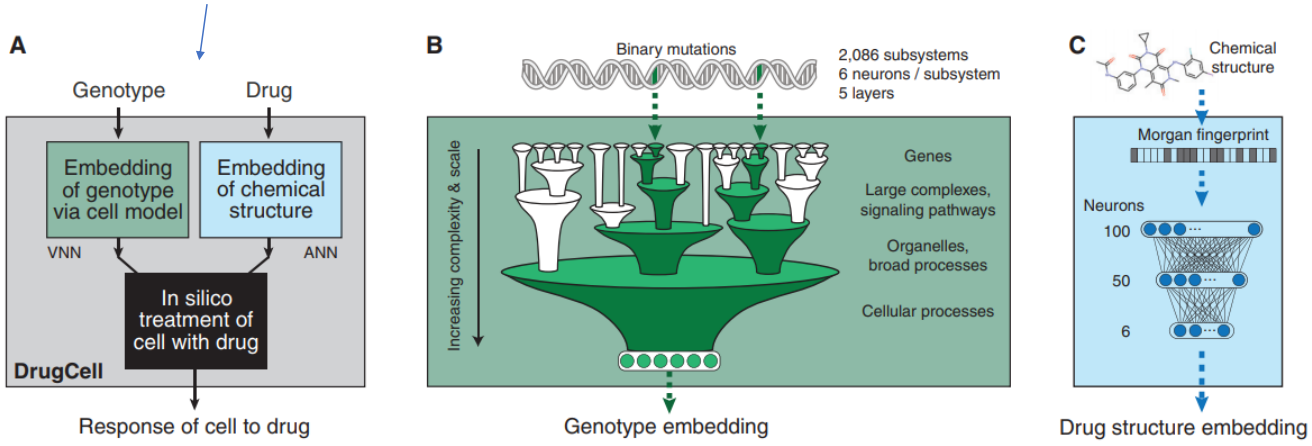
Brent M. Kuenzi,[1,5] Jisoo Park,[1,5] Samson H. Fong,[1,2] Kyle S. Sanchez,[1] John Lee,[1] Jason F. Kreisberg,[1] Jianzhu Ma,[4] and Trey Ideker[1,2,3,6,*]

[1]Division of Genetics, Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA
[2]Department of Bioengineering, University of California San Diego, La Jolla, CA 92093, USA
[3]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093, USA
[4]Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA
[5]These authors contributed equally
[6]Lead Contact
*Correspondence: tideker@ucsd.edu
https://doi.org/10.1016/j.ccell.2020.09.014



Visible Layers (with Gene Ontology)

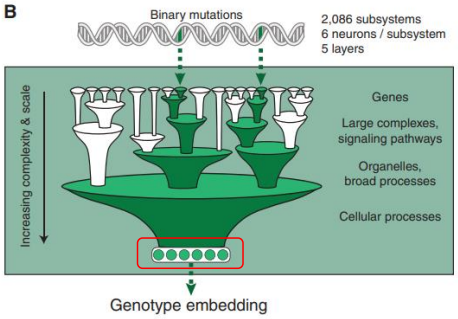
Genomics of Drug Sensitivity in Cancer database (GDSC)
Cancer Therapeutics Response Portal v2 (CTRP)

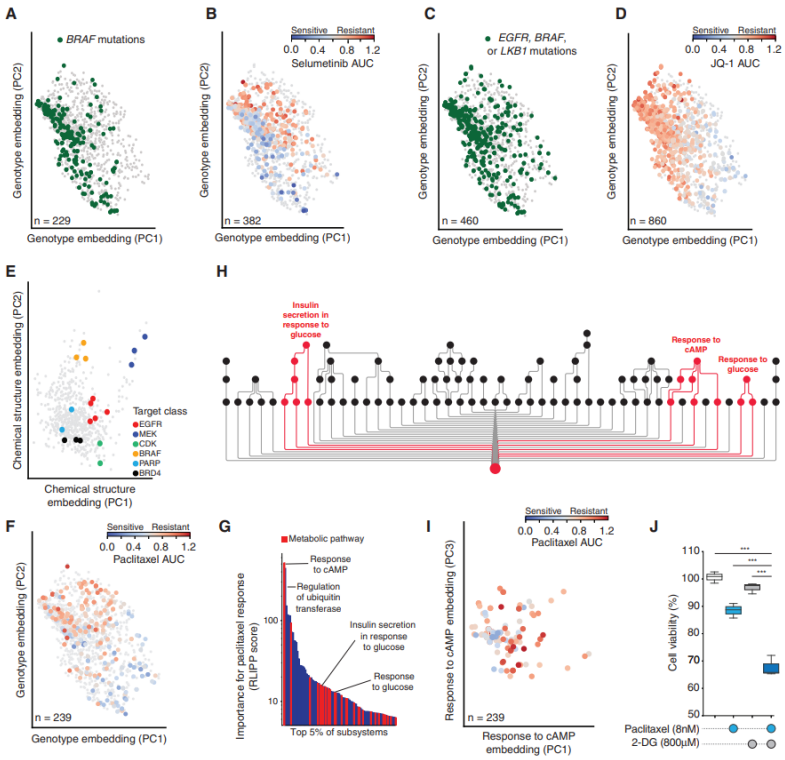


Defining a Hierarchy of Genes and Cellular Subsystems

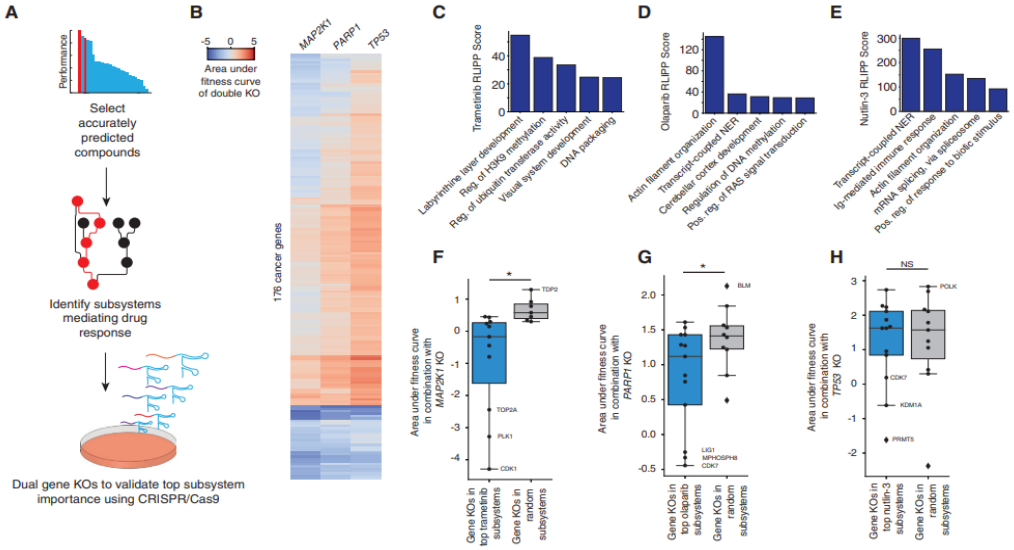
- we selected the top 15% most frequently mutated genes in human cancers accord- ing to the Cancer Cell Line Encyclopedia (CCLE) among genes annotated to Gene Ontology (GO) terms.
- This procedure yielded 3,008 genes, henceforth called '**DrugCell genes**', which were used in model construction. These genes were organized into a hierarchy of nested gene sets, representing cellular subsystems at different scales, based on terms extracted from the GO Biological Process hierarchy.
- To further reduce model complexity, we restricted the hierarchy to a maximal depth of five subsystems by removing all subsystems more than five parent-child relations above the bottom layer subsystems of the hierarchy (subsystems without any children).
- The resulting hierarchy, composed of 2,086 subsystems, defined the branch of DrugCell for embedding of genotype (left branch in Figure 1A, also called the VNN; Figure 1B).

Characterization of Cancer Cell States Learned by DrugCell





Systematic Validation of Identified Mechanisms of Sensitivity Using CRISPR/Cas9



Literature Mining

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

DigChem: Identification of disease-gene-chemical relationships from Medline abstracts

Jeongkyun Kim[1], Jung-jae Kim[2], Hyunju Lee[1]*

1 Gwangju Institute of Science and Technology, School of Electrical Engineering and Computer Science, Gwangju, Korea, **2** Institute for Infocomm Research, A-STAR, 138632, Singapore

* hyunjulee@gist.ac.kr

Abstract

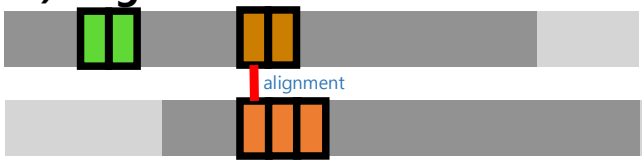
disease-gene-chemical triplet relationship from Medline abstracts

- There are few studies that extract **disease-gene-chemical** relation ships from biomedical literature at a PubMed scale
- Authors proposed a DL model based on **Bi-LSTM** to identify the evidence sentences of relationships from **Medline abstracts**
- Developed a search engine called DigChem
  - http://gcancer.org/digchem -> but not available right now..

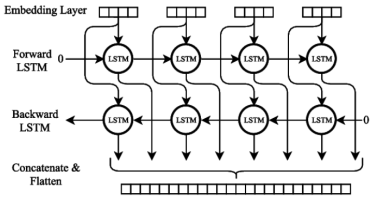
Keywords

- **Disease-Gene-Chemical Triplet Relationship from literature**

- **Horizontal(parallel) alignment of sentences**



- **Bidirectional LSTM**



Materials & Methods

Gold standard evidence sentences

- Authors assumed **two sentences** together represent a triplet relationship

Gene-Chemical sentence — Disease sentence

- if (three elements appear in the same sentence):
  - **duplicate** the sentence into two identical sentences
- if (a sentence has multiple mentions of gene and of chemical):
  - **each pair is extracted** to form either positive or negative triplet with di sease mention
- Authors randomly selected sentence pairs from Medline abstracts, and **ma nually evaluated** them as positive or negative(1,000 pairs each)
  - Among 2,000 pairs, half of them were from the same sentence

Materials & Methods

Positive / Negative relationship

	Positive relationship	Negative relationship
Name of chemical, gene, and disease	exist	exist
Relationship	direct or indirect	none

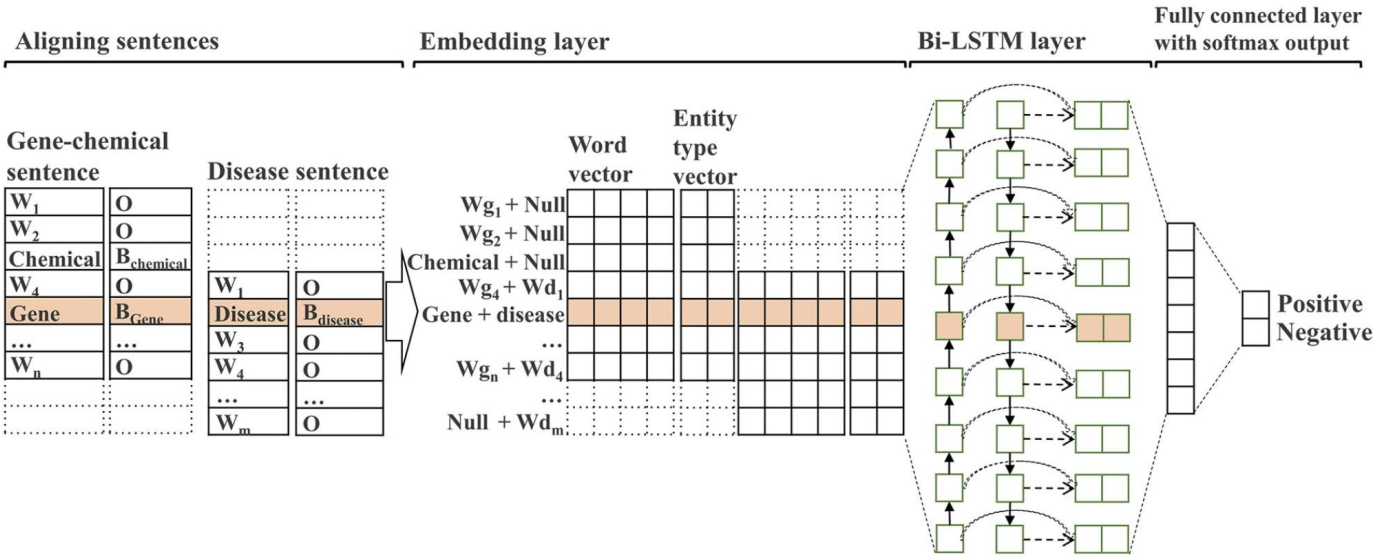
Materials & Methods

Positive / Negative sentence pair example(from article)

- Positive sentence pairs for gene *BNP*, chemical *SUN*, and disease *renal cell carcinoma* (PMID: 24984876).
  - Sentence 1: "At the protein level, Western blot analysis showed that *SUN* increased *BNP* and b-MHC, while it inhibited a-MHC protein levels in a concentration-dependent manner."
  - Sentence 2: "Sunitinib (*SUN*) is a multi-targeted tyrosine kinase inhibitor used for the treatment of gastrointestinal stromal tumors and *renal cell carcinoma*."
- Negative sentence pairs for gene *ACE*, chemical *hydralazine*, and disease *glomerulosclerosis* (PMID: 25143333).
  - Sentence 1: "CONCLUSION: The results show following an abrupt decline in podocyte number, the initiation of *ACE*-inhibition but not *hydralazine*, was accompanied by higher podocyte number in the absence of proliferation."
  - Sentence 2: "OBJECTIVE: The objective of this article is to test the effects of angiotensin-converting enzyme (*ACE*)-inhibition on glomerular epithelial cell number in an inducible experimental model of focal segmental *glomerulosclerosis* (FSGS)."

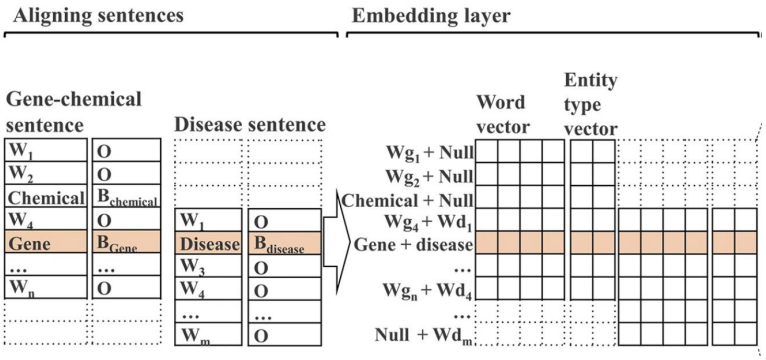
Materials & Methods

Relation classification model



Materials & Methods

Word embedding



- Used two embedding features
  - Word representation vectors
    - applied Word2Vec to Medline data set
    - vector size: 200
  - Entity type representation vectors
    - used NER tools and tagged with BIO format
    - 7 tags in total(B and I for gene, chemical and disease each, and one O)
    - vector size: 20

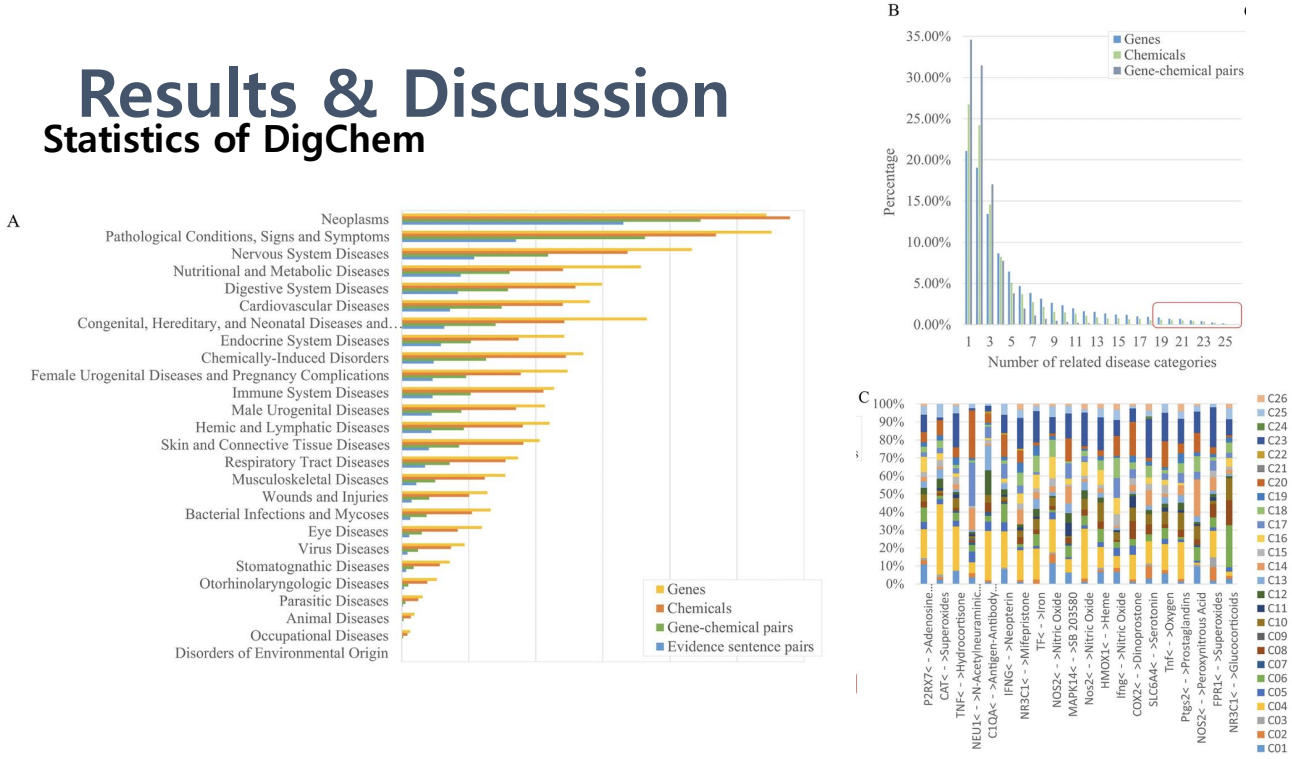
Materials & Methods

Post processing

- Before post-processing, the false positive rates were 35.8%
- Authors constructed five rules to filter out false positive sentences
  1. filter out when recognized mentions are not contained in synonyms of entities in dictionaries after the recognized mentions are normalized into entity names
  2. filter out if any mention is recognized as more than one entity type
  3. filter out if it contains hyponyms of 'study' (it may express a purpose of research)
  4. filter out if gene name and chemical nmae are connected by a conjunction in the dependency parse tree

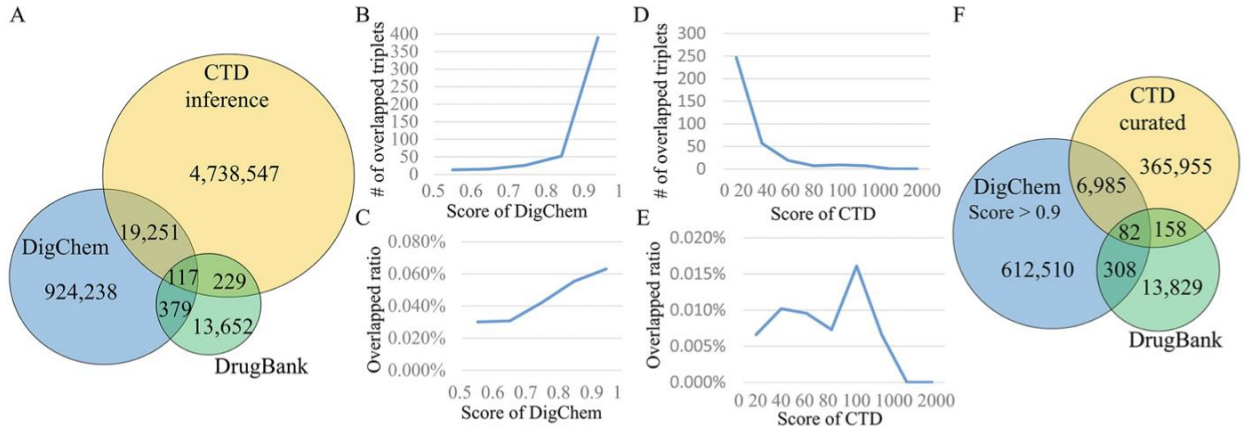
Results & Discussion

Statistics of DigChem



Results & Discussion

Comparison results with CTD & DrugBank



BioBERT and Its Applications

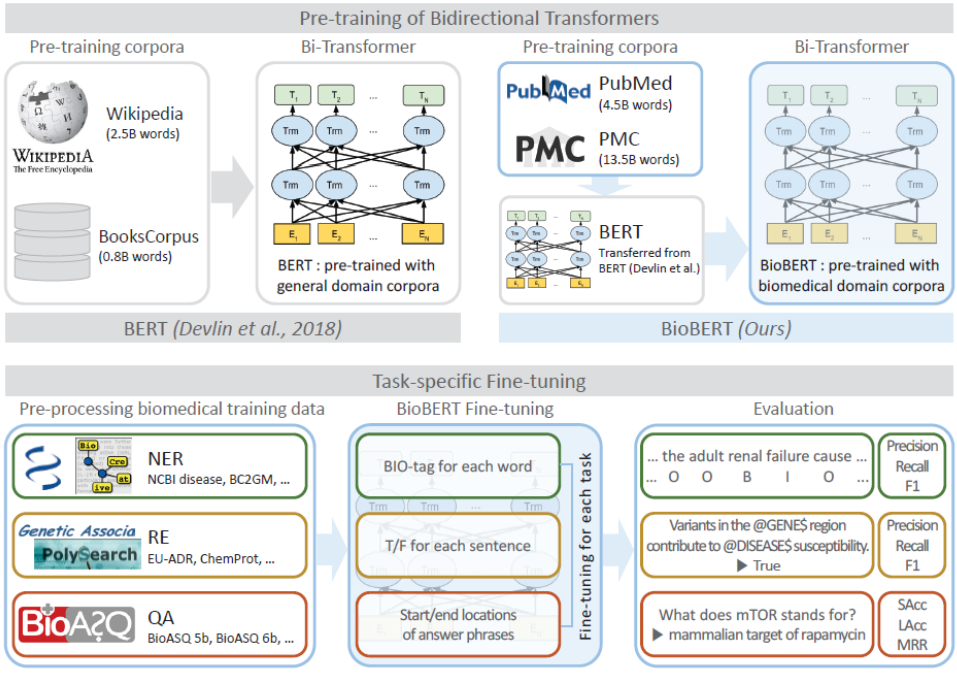
Slides By Prof. Jaewoo Kang @ Korea University

BioBERT: a pre-trained biomedical language representation model for biomedical text mining

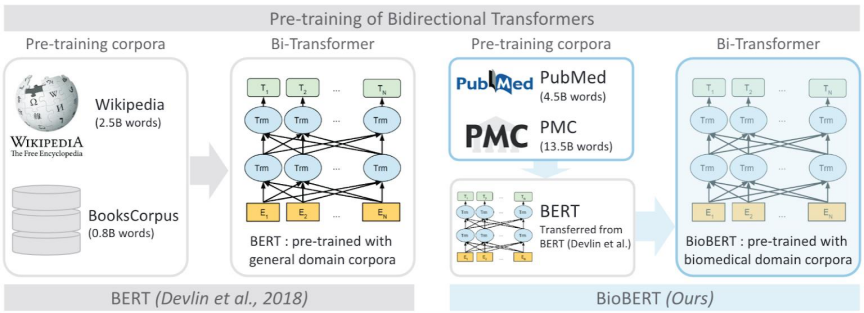
w/ Jinhyuk Lee†, Wonjin Yoon†, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So

[*Bioinformatics 2020*]

Overview of BioBERT



Pre-training of BioBERT



– Pre-trained Bidirectional Transformer on PubMed + PMC (18B words) on top of BERT (3.3B words) for more than **20 days with 8 V100 (32GB) GPUs**.

– Keeps the same architecture throughout the tasks except the last softmax layer

Named Entity Recognition

DMIS

- Named entity recognition (NER) is a task of recognizing proper nouns in a corpus.
  – In the biomedical domain, detecting domain-specific entity types such as disease, chemical, gene and protein, is a main objective of the task.

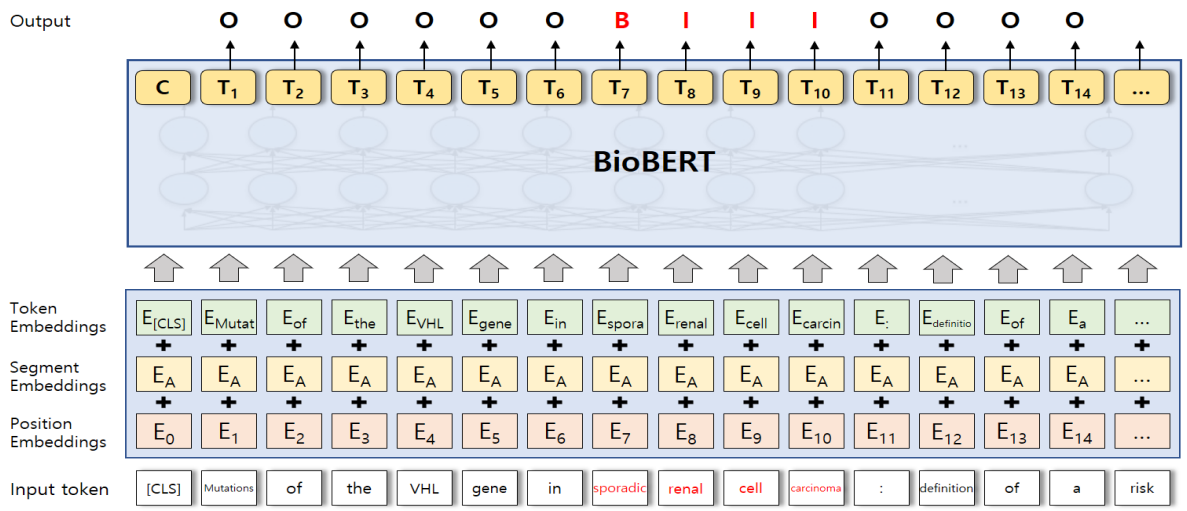
- Example of BioNER: NCBI-disease
  – Sequence tagging task – annotate with B, I, O
  – Polysemious words

Mutations	of	the	VHL	gene	in	sporadic	renal	cell	carcinoma	:	definition
O	O	O	O	O	O	B	I	I	I	O	O
of	a	risk	factor	for	VHL	patients	to	develop	an	*RCC*	.
O	O	O	O	O	B	O	O	O	O	B	O

67

Named Entity Recognition

- (Bio)BERT for sequence tagging task:
  - Utilized the last layer of LM(BioBERT) as a contextualized representation
  - Representations are fed into the output layer (1 layer feed-forward neural network)

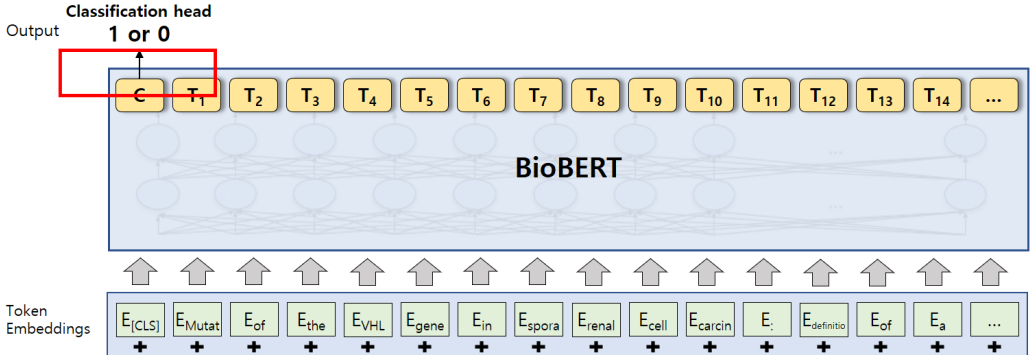


99

Relation Extraction

- Relation extraction in the biomedical domain is a task of classifying relations of named entities in a biomedical corpus.
  - Binary or multi-class classification task on a sentence using [CLS] token.
- Example:
  - "C1167 polymorphism in the @GENE$ gene and D6S366 near the SOD2 gene are not associated with the development of @DISEASE$ and diabetic retinopathy in IDDM. "

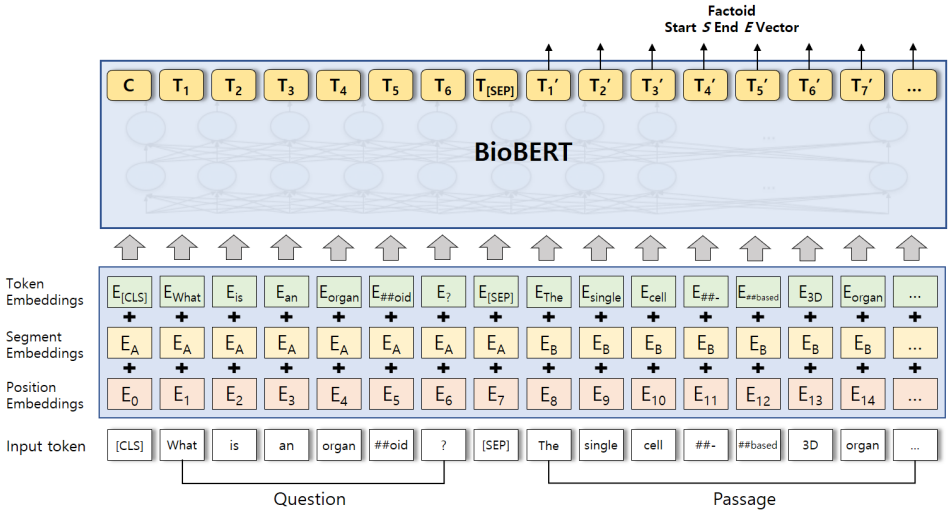
    -> Output: 0 : @GENE$ and @DISEASE$ are not related



200

Question Answering

- Representations are fed into the output layer (1 layer FFNN; e.g. $W \in d^{768\times2}$)
- We utilized multi-level transfer learning strategy : BioBERT -> SQuAD -> BioASQ
  - Fine-tuning on SQuAD : Understand the structure of Question Answering
  - Fine-tuning on BioASQ : Enhance the model by target-domain data supervision



Performance of BioBERT on QA

Table 8. Biomedical question answering test results

Datasets	Metrics	SOTA	BERT (Wiki + Books)	BioBERT v1.0 (+ PubMed)	BioBERT v1.0 (+ PMC)	BioBERT v1.0 (+ PubMed + PMC)	BioBERT v1.1 (+ PubMed)
BioASQ 4b	S	20.01	27.33	25.47	26.09	**28.57**	27.95
	L	28.81	44.72	44.72	42.24	**47.82**	44.10
	M	23.52	33.77	33.28	32.42	**35.17**	34.72
BioASQ 5b	S	41.33	39.33	41.33	42.00	44.00	**46.00**
	L	56.67	52.67	55.33	54.67	56.67	**60.00**
	M	47.24	44.27	46.73	46.93	49.38	**51.64**
BioASQ 6b	S	24.22	33.54	**43.48**	41.61	40.37	42.86
	L	37.89	51.55	55.90	55.28	57.77	57.77
	M	27.84	40.88	48.11	47.02	47.48	**48.43**

Notes: Strict Accuracy (S), Lenient Accuracy (L) and Mean Reciprocal Rank (M) scores on each dataset are reported. The best scores are in bold, and the second best scores are underlined. The best BioASQ 4b/5b/6b scores were obtained from the BioASQ leaderboard (http://participants-area.bioasq.org).

- **12.24 MRR** improvement on average

- Achieves state-of-the-art of performances on **3 out of 3** datasets (BioASQ 4b, 5b, 6b)

Selected as one of top-3 best papers

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2020 in the section 'Natural Language Processing'. The articles are listed in alphabetical order of the first author's surname.

Section Natural Language Processing
▪ Guan J, Li R, Yu S, Zhang X. A Method for Generating Synthetic Electronic Medical Record Text. IEEE/ACM Trans Comput Biol Bioinform 2019.
▪ Lee J, Yoon W, Kim S, Kim D, Kim S, Ho So C, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2019;36(4):1234-40.
▪ Rosemblat G, Fiszman M, Shin D, Kılıçoğlu H. Towards a characterization of apparent contradictions in the biomedical literature using context analysis. J Biomed Inform 2019;98:103275.

203

Wrap-up

Main Issues

- **Representation**
- **Search space**
- Truly inter-disciplinary field including computer science, chemistry, biology, pharmacology, medicine, animal sciences, etc.
- We, computer scientists, have a lot to do!

2020 DAY1

+ **interns in my lab** contributed a lot.

감사합니다!

질문 부탁드립니다.