# GraphDTA

Prediction of drug-target binding affinity using graph convolutional networks

2020.10.28

Hyeonsu Lee KAICD





Time	Title	Contents		
09:00 ~ 09:40	Introduction - GraphDTA	Introduction, Data Representation		
10:00 ~ 10:40	Graph Structure	Graph, Graph Representation		
11:00 ~ 12:00	Graph Convolutional Networks	Graph Convolution, Graph Convolutional Networks		
12:00 ~ 13:00	Lunch Time			
13:00 ~ 13:40		Data Structure		
14:00 ~ 14:40	GraphDTA Practice(Google Colab)	Model Structure		
15:00 ~ 15:40		Predict with Pretrained Model		
16:00 ~ 16:40	Introduction – Auto Encoder, Generative Models	AE, Stacked AE, VAE, feat.GAN		
17:00 ~ 18:00	VAE Demo(Google Colab)	CHEM VAE Demo		

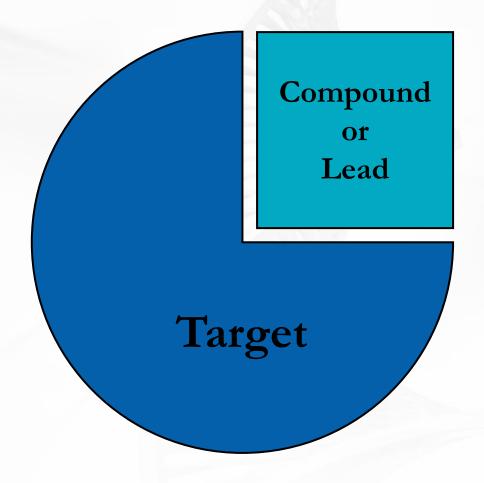


# Drug-Target Interaction



## DTI

- 용어
  - ▶ Target : 질병과 관련된 특정 유기체 내의 분자 (단백질)
  - ➤ Compound or Lead : Target과 결합하여 Target을
    제어할 수 있는 화학 물질

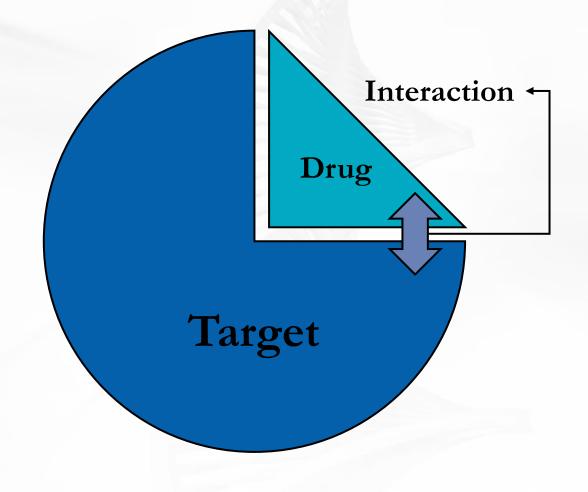


# 보고인공지능신약개발지원센터

## DTI

#### • 용어

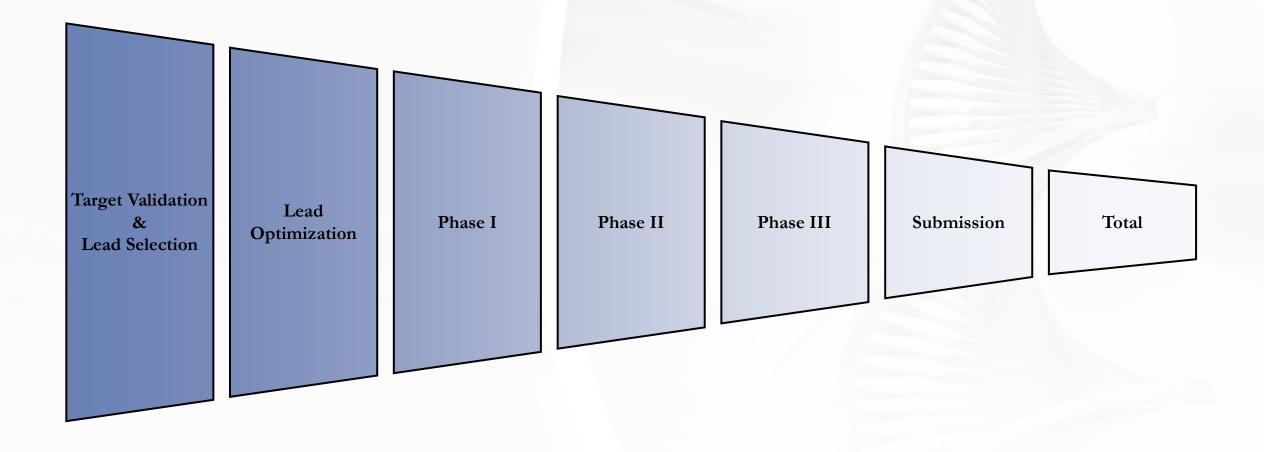
- ▶ Properties : 질병 치료를 위해 Compound or Lead
   에서 원하는 모든 특성(뛰어난 효과, 낮은 부작
   용, 낮은 독성)
- ➤ Drug: Properties가 완전히 최적화된 Compound
- ➤ Interaction : Target과 Drug 사이의 상호 작용
- ➤ DTI : Interaction 예측을 통한 신약후보물질 도출
- > DTA: Binding Affinity를 통해 interaction을 예측



DTI SUMMARY

## **DTI**

• 개요



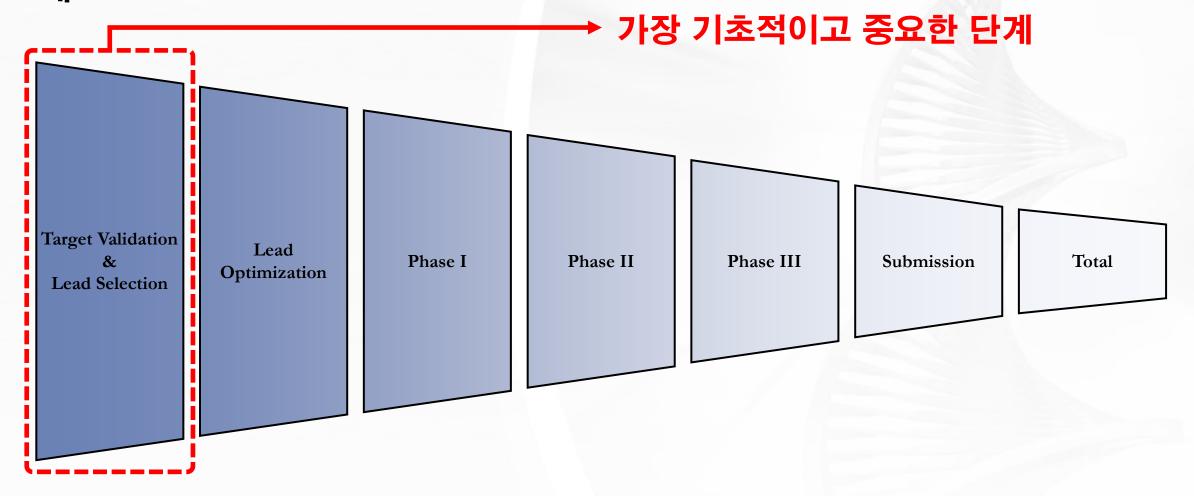
DTI SUMMARY





October 28, 2020

• 개요

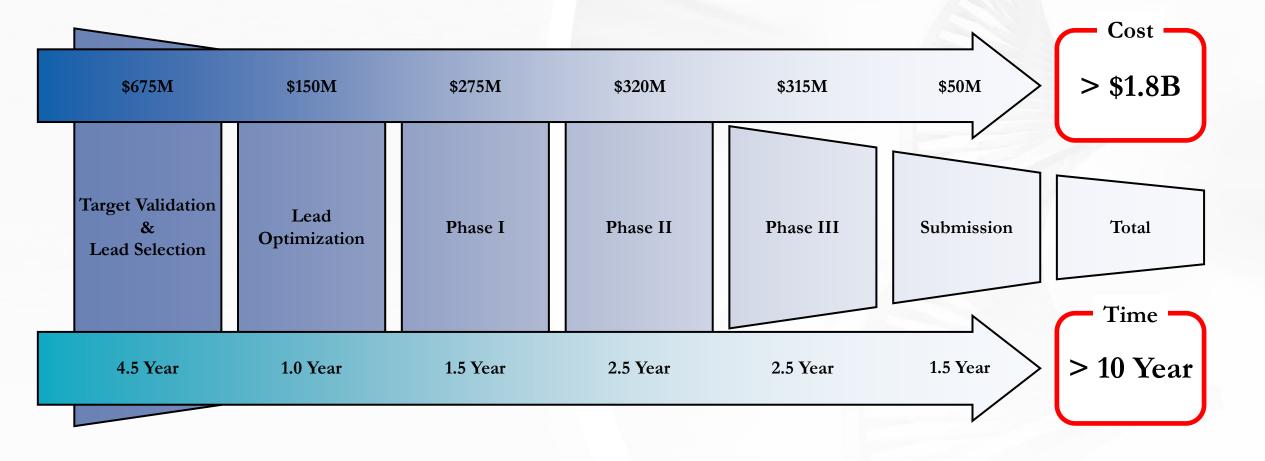


DTI LIMITATION

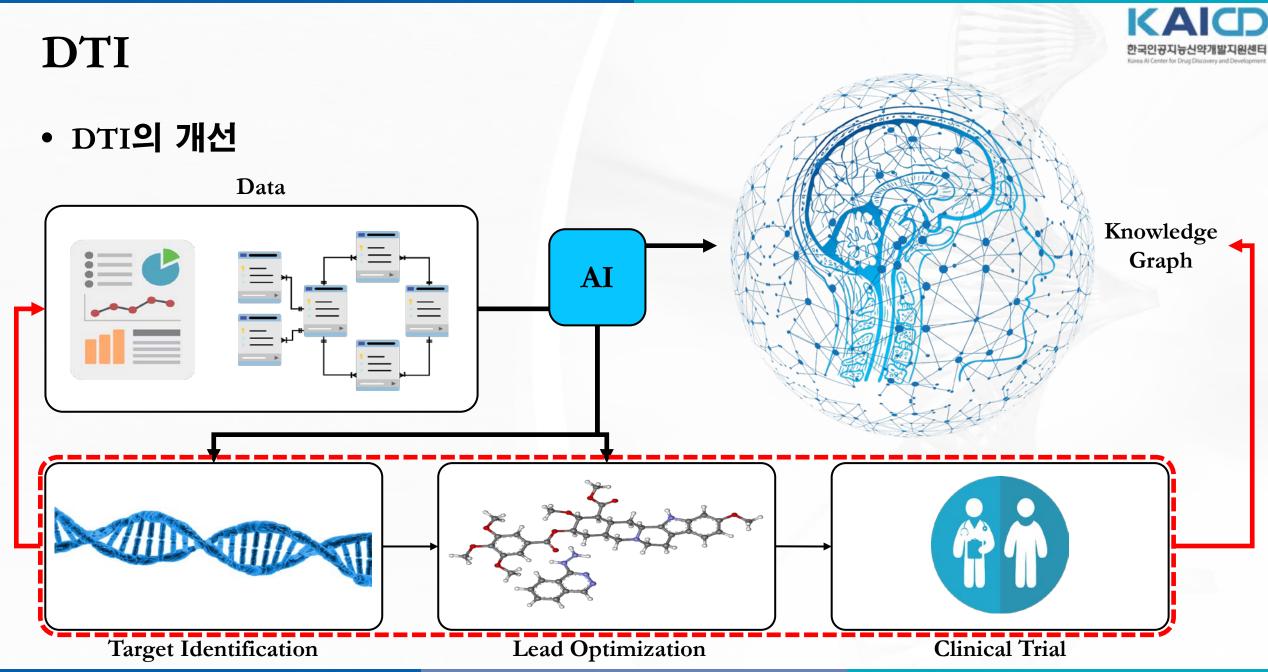


# KAICD 한국인공지능신약개발지원센트 Karea Al Center for Drug Discovery and Directogener

## • 한계



DTI IMPROVEMENT



#### DTI

#### • 기존 연구



DESEABOH ARTICLE

Computational-experimental approach to drug-target interaction mapping: A case study on kinase inhibitors

Anna Cichonska<sup>1,2</sup>\*, Balaguru Ravikumar<sup>2</sup>, Elina Parri<sup>2</sup>, Sanna Timonen<sup>2</sup>, Tapio Pahikkala<sup>3</sup>, Antti Airola<sup>3</sup>, Krister Wennerberg<sup>2</sup>, Juho Rousu<sup>1</sup>, Tero Aittokallio<sup>2,4</sup>\*

 Helsinki Institute for Information Technology HillT, Department of Computer Science, Aallo University, Espop, Finland, 2 Institute for Molecular Modicine Finland FillMI, University of Helsinki, Helsinki, Finland,
3 Department of Information Technology, University of Turku, Turku, Finland, 4 Department of Mathematics and Statistics, University of Turku, Turku, Finland

\* anna.cichonska@helsinki.fi (AC); tero.aittokalio@helsinki.fi (TA)

#### Abstract

Citatioe: Cichonska A, Ravikumar B, Parri E, Timonen S, Pahikasla T, Arota A, et al. (2017) Computational-perprimental approach to drugtarget interaction mapping: A case study on kinase inhibitors. PLoS Comput Biol 13(5): e1005678. https://doi.org/10.1371/s/cuma.pobi.1005678.

Editor: Avner Schlessinger, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: February 3, 2017
Accepted: July 11, 2017

Accepted: July 11, 2017

Published: August 7, 2017

Copyright: 0 2017 Cichonska et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and according to any machine, provided the principal

author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information

Funding: This work was financially supported by the Helsinki Doctoral Education Network in Information and Communications Technology HIGT to AC, Academy of Fishand (285496 to JR, 2698692, 27432, 265594 and 315070 to TA, 272577, 277293 to KW, 289903 to AM), Cancer Society of Fishand to KW and TA, Signid Juselius Souday of Fishand to KW and TA, Signid Juselius Soudayton KW, and the Biocentim Helsinki Due to relatively high costs and labor required for experimental profiling of the full target space of chemical compounds, various machine learning models have been proposed as cost-effective means to advance this process in terms of predicting the most potent compound-target interactions for subsequent verification. However, most of the model predictions lack direct experimental validation in the laboratory, making their practical benefits for drug discovery or repurposing applications largely unknown. Here, we therefore introduce and carefully test a systematic computational-experimental framework for the prediction and pre-clinical verification of drug-target interactions using a well-established kernel-based regression algorithm as the prediction model. To evaluate its performance, we first predicted unmeasured binding affinities in a large-scale kinase inhibitor profiling study, and then experimentally tested 100 compound-kinase pairs. The relatively high correlation of 0.77 (p < 0.0001) between the predicted and measured bioactivities supports the potential of the model for filling the experimental gaps in existing compound-target interaction maps. Further, we subjected the model to a more challenging task of predicting target interactions for such a new candidate drug compound that lacks prior binding profile information. As a specific case study, we used tivozanib, an investigational VEGF receptor inhibitor with currently unknown off-target profile. Among 7 kinases with high predicted affinity, we experimentally validated 4 new off-targets of tivozanib, namely the Src-family kinases FRK and FYN A, the non-receptor tyrosine kinase ABL1, and the serine/threonine kinase SLK. Our sub-sequent experimental validation protocol effectively avoids any possible information leakage between the training and validation data, and therefore enables rigorous model validation for practical applications. These results demonstrate that the kernel-based modeling approach offers practical benefits for probing novel insights into the mode of action of investigational compounds, and for the identification of new target selectivities for drug repurposing applications.

PLOS Computational Biology | https://doi.org/10.1371/journal.pcbi.1005678 August 7, 2017

KronRLS

#### RESEARCH ARTICLE SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines Tong He<sup>1†</sup>, Marten Heidemeyer<sup>1†</sup>, Fuqiang Ban<sup>2</sup>, Artem Cherkasov<sup>2‡</sup> and Martin Ester<sup>1††</sup> omputational prediction of the interaction between drugs and targets is a standing challenge in the field of drug scovery. A number of rather accurate predictions were reported for various binary drug-target benchmark datats. However, a notable drawback of a binary representation of interaction data is that missing endpoints for non nteracting drug-target pairs are not differentiated from inactive cases, and that predicted levels of activity depend n pre-defined binarization thresholds. In this paper, we present a method called SimBoost that predicts continuous (non-binary) values of binding affinities of compounds and proteins and thus incorporates the whole interaction spectrum from true negative to true positive interactions. Additionally, we propose a version of the method called mBoostQuant which computes a prediction interval in order to assess the confidence of the predicted affinity, thus lefining the Applicability Domain metrics explicitly. We evaluate SimBoost and SimBoostQuant on two established drug-target interaction benchmark datasets and one new dataset that we propose to use as a benchmark for readoss cheminformatics applications. We demonstrate that our methods outperform the previously reported models cross the studied datasets Keywords: Read-across, Gradient boosting, Drug-target interaction, Prediction interval, Applicability Domain, QSAR Recent studies [1] have demonstrated that machine Finding a compound that selectively binds to a particu-learning-based approaches have the potential to predict lar protein is a highly challenging and typically expensive procedure in the drug development process, where more than 90% of candidate compounds fail due to cross-information on the similarity among compounds and the process of th reactivity and/or toxicity issues. It is therefore an important topic in drug research to gain knowledge about the drugs and between targets to infer the interaction of interaction of compounds and target proteins through untested drug-target pairs is the essence of the readcomputational methods. Such in silico approaches are capable of speeding up the experimental wet lab work by tematically prioritizing the most potent compounds I help predicting their potential side effects. evaluation of such machine learning-based prediction methods are the Enzymes, Ion Channels, Nuclear Recepand help predicting their potential side effects. tor, and G Protein-Coupled Receptor datasets [3]. These datasets contain binary labels $Y_{(i,j)} = 1$ if drug-target pair $(d_i, t_i)$ is known to interact (as shown by wet lab experiments) and $Y_{(i,j)} = 0$ if either $(d_i, t_j)$ is known to \*Correspondence: ester@cs.sfu.ca; ester@sfu.ca \*Tong He and Marten Heldemeyer contributed equally to this work not interact or if the interaction of $(d_i, t_i)$ is unknown. The datasets tend to be biased towards drugs and targets that are considered to be more important or easier ✓ Springer Open

Journal of Cheminformatics

Korea Al Center for Drug Discovery and Developmen

#### **SimBoost**

#### DTI

#### • 최근 연구

Bioinformatics, 34, 2018, i821-i829 ECCB 2018



#### DeepDTA: deep drug-target binding affinity prediction

Hakime Öztürk<sup>1</sup>, Arzucan Özgür<sup>1,\*</sup> and Elif Ozkirimli<sup>2,\*</sup>

<sup>1</sup>Department of Computer Engineering and <sup>2</sup>Department of Chemical Engineering, Bogazici University, Istanbul 34342. Turkey

\*To whom correspondence should be addressed

Motivation: The identification of novel drug-target (DT) interactions is a substantial part of the drug discovery process. Most of the computational methods that have been proposed to predict DT interactions have focused on binary classification, where the goal is to determine whether a DT pair interacts or not. However, protein-ligand interactions assume a continuum of binding strength values, also called binding affinity and predicting this value still remains a challenge. The increase in the affinity data available in DT knowledge-bases allows the use of advanced learning techniques such as deep learning architectures in the prediction of binding affinities. In this study, we propose a deep-learning based model that uses only sequence information of both targets and drugs to predict DT interaction binding affinities. The few studies that focus on DT binding affinity prediction use either 3D structures of protein-ligand complexes or 2D features of compounds. One novel approach used in this work is the modeling of protein sequences and compound 1D representations with convolutional neural networks (CNNs).

Results: The results show that the proposed deep learning based model that uses the 1D representations. tations of targets and drugs is an effective approach for drug target binding affinity prediction. The model in which high-level representations of a drug and a target are constructed via CNNs achieved the best Concordance Index (CI) performance in one of our larger benchmark datasets, outperforming the KronRLS algorithm and SimBoost, a state-of-the-art method for DT binding af finity prediction.

Contact: arzucan.ozgur@boun.edu.tr or elif.ozkirimli@boun.edu.tr

Supplementary information: Supplementary data are available at Bioinformatics online.

#### 1 Introduction

The successful identification of drug-target interactions (DTI) is a critical step in drug discovery. As the field of drug discovery expands with the discovery of new drugs, repurposing of existing drugs and identification of novel interacting partners for approved drugs is also gaining interest (Oprea and Mestres, 2012). Until recently, DTI prediction was approached as a binary classification problem Bleakley and Yamanishi, 2009; Cao et al., 2014, 2012; Cobanoglu et al., 2013; Gönen, 2012; Öztürk et al., 2016; Yamanishi et al., information about protein-ligand interactions, namely the binding affinity values. Binding affinity provides information on the strength of the interaction between a drug-target (DT) pair and it is usually expressed in measures such as dissociation constant (K<sub>d</sub>), inhibition and Zeng, 2016). Formulating the DT prediction task as a binding

Similarly, low Ki values indicate high binding affinity, Kd and Ki values are usually represented in terms of pK<sub>d</sub> or pK<sub>s</sub>, the negative loga-rithm of the dissociation or inhibition constants.

In binary classification based DTI prediction studies, construcnegative (not-binding) samples directly affects the performance of the model. As of last decade, most of the DTI studies utilized four major datasets by Yamanishi et al. (2008) in which DT pairs with 2008; van Laarhoven et al., 2011), neglecting an important piece of no known binding information are treated as negative (not-binding) samples. Recently, DTI studies that rely on databases with binding affinity information have been providing more realistic binary constant (Ka) or the half maximal inhibitory concentration (IC50). affinity prediction problem enables the creation of more realistic IC<sub>50</sub> depends on the concentration of the target and ligand datasets, where the binding affinity scores are directly used.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.oru/licenses/bv-nc/4.0/l. which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact

#### **DeepDTA**

#### WIDEDTA: PREDICTION OF DRUG-TARGET BINDING AFFINITY

#### A PREPRINT

Hakime Öztürk

Department of Computer Engineering Bogazici University 34342 Istanbul Turket

Elif Ozkirimli Department of Chemical Engineering Bogazici University 34342 Istanbul Turkey

Arzucan Özgür Department of Computer Engineering Bogazici University 34342 Istanbul, Turkey arzucan.ozgur@boun.edu.tr

February 13, 2019

#### ABSTRACT

Motivation: Prediction of the interaction affinity between proteins and compounds is a major chall lenge in the drug discovery process. WideDTA is a deep-learning based prediction model that em-ploys chemical and biological textual sequence information to predict binding affinity. Results: WideDTA uses four text-based information sources namely the protein sequence, ligand SMILES, protein domains and motifs, and maximum common substructure words to predict binding affinity. WideDTA outperformed one of the state of the art deep learning methods for drug-target binding affinity prediction, DeepDTA on the KIBA dataset with a statistical significance. This indionlung animy prediction, prediction, prepared to the KIDA dataset with a statistical significance. This indicates that the word-based sequence representation adapted by WideDTA is a promising alternative to the character-based sequence representation approach in deep learning models for binding affinity prediction, such as the one used in DeepDTA. In addition, the results showed that, given the protein sequence and ligand SMILES, the inclusion of protein domain and motif information as well as ligand maximum common substructure words do not provide additional useful information for the deep learning model. Interestingly, however, using only domain and motif information to represent proteins achieved similar performance to using the full protein sequence, suggesting that important binding relevant information is contained within the protein motifs and domains

#### 1 Introduction

4 Feb

[q-bio.QM]

arXiv:1902.04166v1

Discovery of potential drugs for new targets is an expensive and time consuming process. Even though over 97M compounds are deposited in PubChem database [1] (accession date: Jan 2019), the latest version (version 5.1.2) of DrugBank [2] reports only around 12K drug entries. Considering the expansive search space, the development of methodologies to predict the interactions between drugs and targets with high precision can be accelerated by the recent advances in the artificial intelligence applications in chemical research.

Over the last decade, most studies employing traditional machine learning algorithms modelled the prediction of the interaction between compounds and proteins as a binary classification problem (interacts or not) [3, 4, 5, 6, 7]. Recently, deep learning architectures have also become a popular choice in drug discovery studies. The success of the first studies [8, 9] that employed deep neural networks (DNN) to model the interaction between proteins and compounds over traditional machine-learning methods motivated later studies that adopted new architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) [10, 11] and stacked-autoencoders [12].

**WideDTA** 





Korea Al Center for Drug Discovery and Developme

#### **GANsDTA: Predicting Drug-Target Binding Affinity Using GANs**

Lingling Zhao 1\*, Junjie Wang 1, Long Pang 2, Yang Liu 1 and Jun Zhang 3

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, <sup>2</sup> Institute of Space Environment and Material Science, Harbin Institute of Technology, Harbin, China, <sup>3</sup> Department of Rehabilitation, Hailongjang

The computational prediction of interactions between drugs and targets is a standing challenge in drug discovery. State-of-the-art methods for drug-target interaction prediction are primarily based on supervised machine learning with known label information. However, in biomedicine, obtaining labeled training data is an expensive and a laborious process. This paper proposes a semi-supervised generative adversarial networks (GANs)-based method to predict binding affinity. Our method comprises two parts, two GANs for feature extraction and a regression network for prediction. The semisupervised mechanism allows our model to learn proteins drugs features of both labeled and unlabeled data. We evaluate the performance of our method using multiple public datasets. Experimental results demonstrate that our method achieves competitive performance while utilizing freely available unlabeled data. Our results suggest that utilizing such unlabeled data can considerably help improve performance in various biomedical relation extraction processes, for example, Drug-Target interaction and protein-protein interaction, particularly when only limited labeled data are available in such tasks. To our best knowledge, this is the first semi-supervised GANs-based method

#### Specialty section:

Technology of China, China

Frontiers in Genetics

Received: 23 August 2019

Predicting Drug-Target Binding

INTRODUCTION

to predict binding affinity.

A basic task in the field of new drug design and development is to model the interaction between known drugs and target proteins and to identify drugs with a high affinity for specific disease proteins (Cheng et al., 2018a; Cheng et al., 2019b). However, this is a rather challenging and xpensive process even when only approximately 97M compounds reported by the PubChem database (Bolton et al., 2008) and 12K drug entries reported by the DrugBank (Wishart et al., 2006 are considered. Computational methods, especially machine learning models, can considerably accelerate the drug development process and save costs by guiding biological experiments.

Drug-target interaction (DTI) prediction (Yamanishi et al., 2010; Liu et al., 2016; Nascim et al., 2016: Keum and Nam, 2017) was modeled as a binary classification problem and solved by a few traditional machine learning methods in recent decades. These methods have achieved remarkable performancehowever, they still exhibit limitations because of their strong dependence doi: 10.3389/tjone.2019.01243 on handcrafted features.

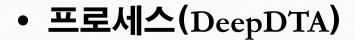
Frontiers in Genetics | www.frontiersin.org

**GANsDTA** 

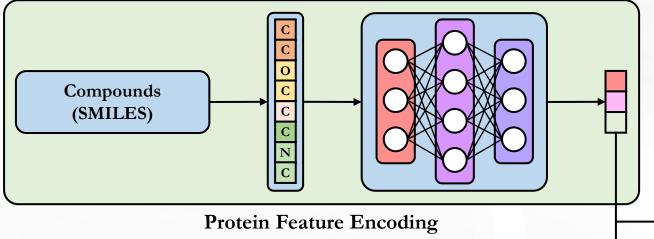
DTI PROCESS



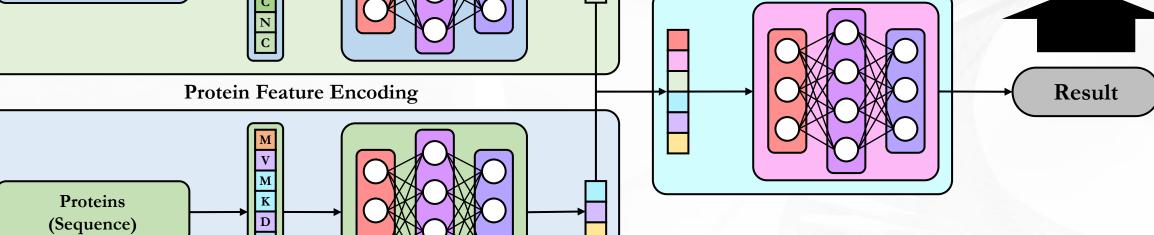








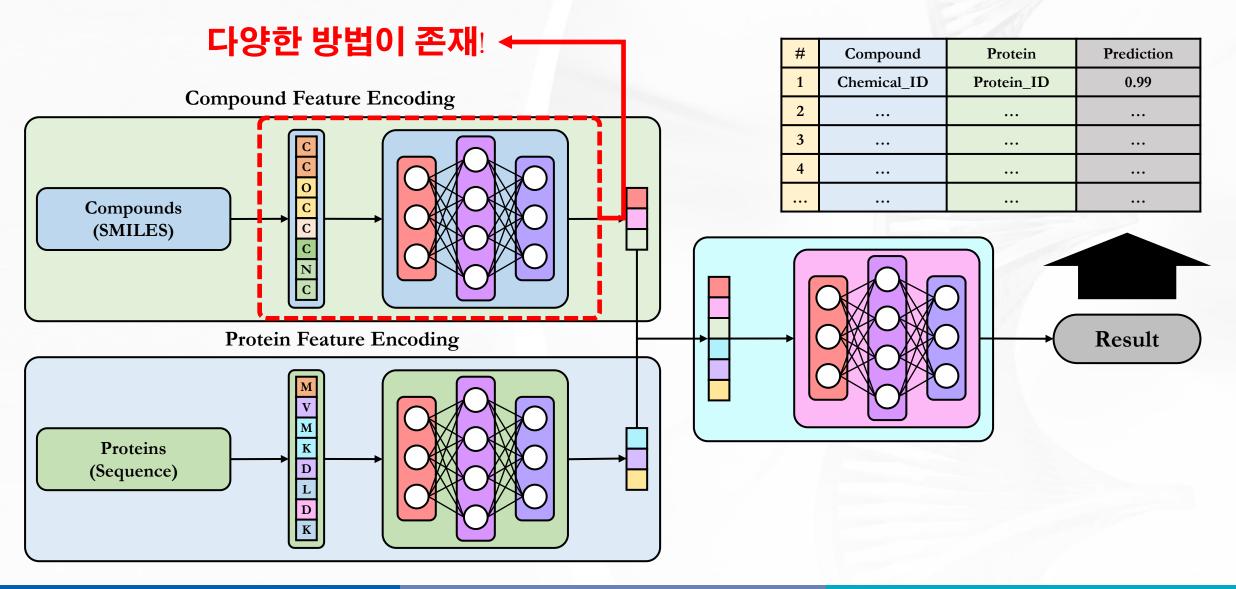
#	Compound	Protein	Prediction						
1	Chemical_ID	Protein_ID	0.99						
2	•••								
3									
4									



DTI PROCESS

## **DTI**





DTI PROPOSED MODEL





#### • Proposed Model - GraphDTA

bioRxiv preprint doi: https://doi.org/10.1101/684662, this version posted July 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

#### GraphDTA: prediction of drug-target binding affinity using graph convolutional networks

Thin Nguyen<sup>1\*</sup>, Hang Le<sup>2</sup>, Svetha Venkatesh<sup>1</sup>

<sup>1</sup>Applied Artificial Intelligence Institute, Deakin University, Australia;

<sup>2</sup>Faculty of Information Technology, Nha Trang University, Viet Nam

\*Correspondence: thin nguyen@deakin.edu.au

#### Abstract

While the development of new drugs is costly, time consuming, and often accompanied with safety issues, drug repurposing, where old drugs with established safety are used for medical conditions other than originally developed, is an attrac-tive alternative. Then, how the old drugs work on new targets becomes a crucial part of drug repurposing and gains much of interest. Several statistical and machine learning models have been proposed to estimate drug-target binding affinity and deep learning approaches have been shown to be among state-of-the-art meth ods. However, drugs and targets in these models were commonly represented in 1D strings, regardless the fact that molecules are by nature formed by the chemical bonding of atoms. In this work, we propose GraphDTA to capture the struc tural information of drugs, possibly enhancing the predictive power of the affin ity. In particular, unlike competing methods, drugs are represented as graphs and graph convolutional networks are used to learn drug-target binding affinity. We trial our method on two benchmark drug-target binding affinity datasets and compare the performance with state-of-the-art models in the field. The results show that our proposed method can not only predict the affinity better than non-deep learning models, but also outperform competing deep learning approaches. This demonstrates the practical advantages of graph-based representation for molecules in providing accurate prediction of drug-target binding affinity. The application may also include any recommendation systems where either or both of the userand product-like sides can be represented in graphs.

Availability and implementation The proposed models are implemented in Python. Related data, pre-trained models, and source code are publicly available at https://github.com/thinng/GraphDTA.

#### 1 Introduction

It costs 2.6 billion US dollars to develop a *de novo* drug [22] and takes about 10–17 years for the drug to be accepted/rejected by US FDA [1, 29]. Repurposing/repositioning a drug – identifying new use for an existing approved drug [33]—would reduce the time

▶ 입력 : 그래프 형식으로 나타낸 화학 화합물(Drug) 정보

▶ 입력 : 시퀀스 형태의 단백질(Target) 정보

凌력 : 화학 화합물과 단백질 사이의 결합 친화도 예측 점수

▶ 기존 모델보다 더 뛰어난 성능을 보임

Method	Protein rep.	Compound rep.	CI	MSE		
Baseline models						
DeepDTA	Smith-Waterman	Pubchem-Sim	0.790	0.608		
DeepDTA	Smith-Waterman	1D	0.886	0.420		
DeepDTA	1D	Pubchem-Sim	0.835	0.419		
KronRLS	Smith-Waterman	Pubchem-Sim	0.871	0.379		
SimBoost	SimBoost Smith-Waterman		0.872	0.282		
DeepDTA	1D	1D	0.878	0.261		
WideDTA	1D + PDM	1D + LMCS	0.886	0.262		
Proposed model - GraphDTA						
GCN [17]	1D	Graph	0.880	0.254		
GAT_GCN	1D	Graph	0.881	0.245		
GAT [37]	1D	Graph	0.892	0.232		
GIN [40]	1D	Graph	0.893	0.229		

HOW?

PROPOSED MODEL

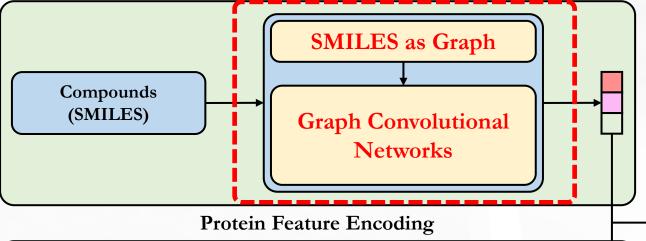




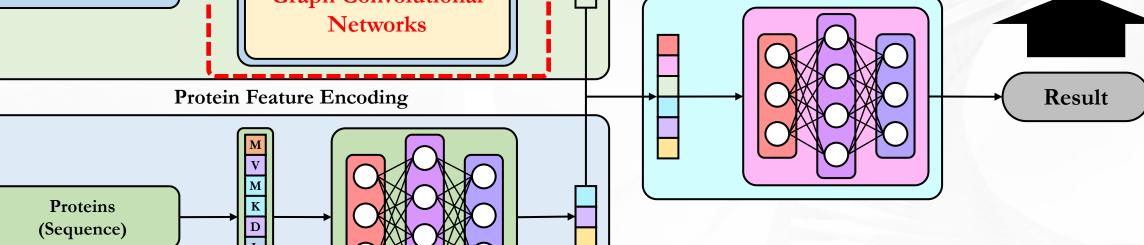




DTI



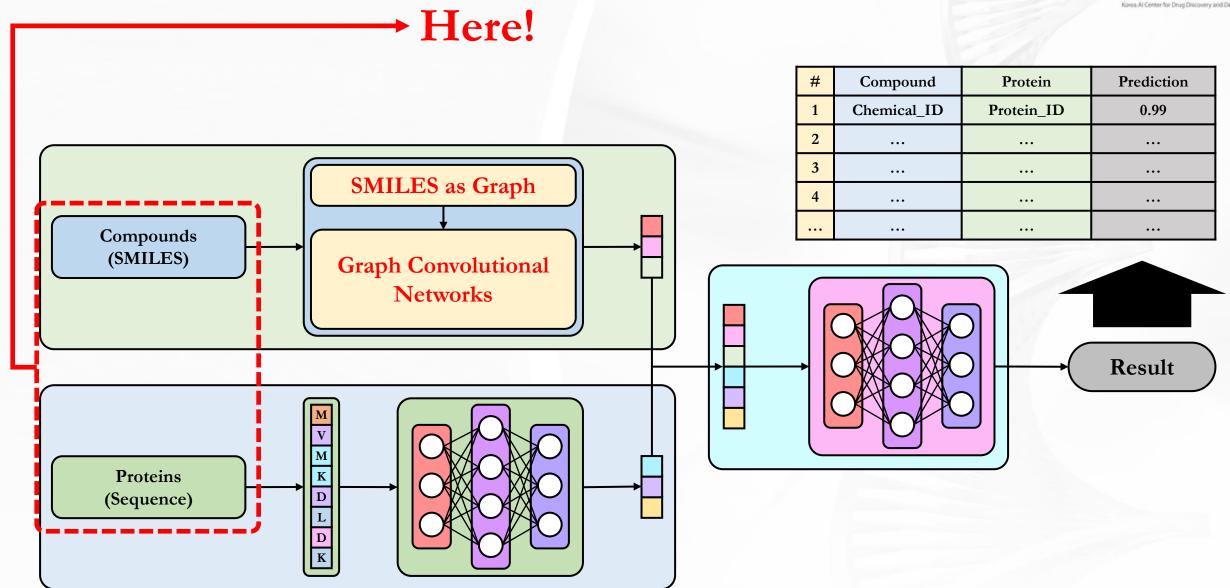
#	Compound	Protein	Prediction 0.99					
1	Chemical_ID	Protein_ID						
2								
3			•••					
4								
•••	•••	•••						





# Datasets





DATASETS DAVIS & KIBA



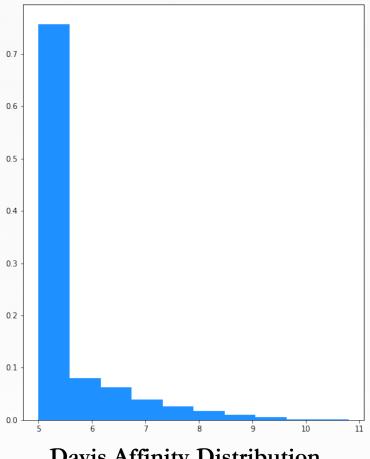
## **Datasets**

- Davis
  - ➢ 결합 친화도 점수 : 평형 이온화 상수 K<sub>d</sub>
- Kiba
  - $\triangleright$  결합 친화도 점수 :  $K_i$ ,  $K_d$ ,  $IC_{50} \rightarrow$  statistically optimized and combined  $\rightarrow$  KIBA score

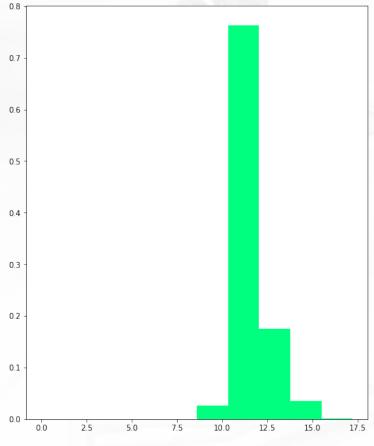
Dataset Name	Proteins	Compounds	Affinity value	Score	
Davis	442	72	5.0 ~ 10.8	$K_{\!d}$	
Kiba	229	2,116	0.0 ~ 17.2	KIBA	

## **Datasets**

• Affinity Distribution



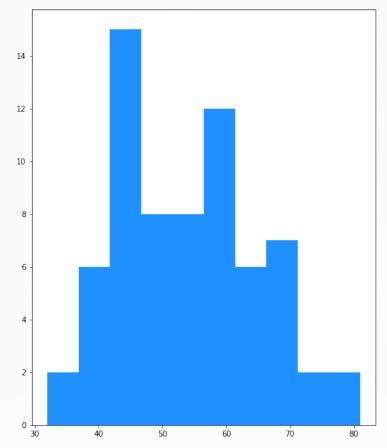
**Davis Affinity Distribution** 



Kiba Affinity Distribution

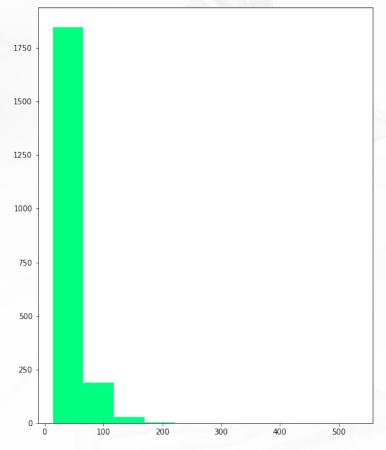
## **Datasets**

• Length of SMILES Distribution



Davis Length of SMILES Distribution



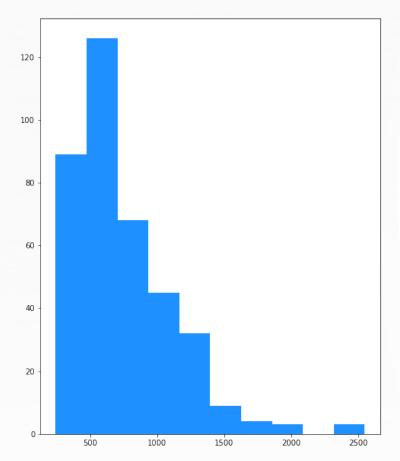


Kiba Length of SMILES Distribution

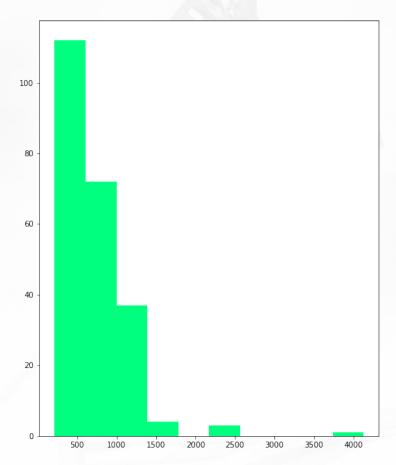
# ICAICD 한국인공지능신약개발지원센터 Nonea Al Center for Drug Discovery and Development

## **Datasets**

• Length of Protein Sequence Distribution



Davis Length of Protein Sequence Distribution



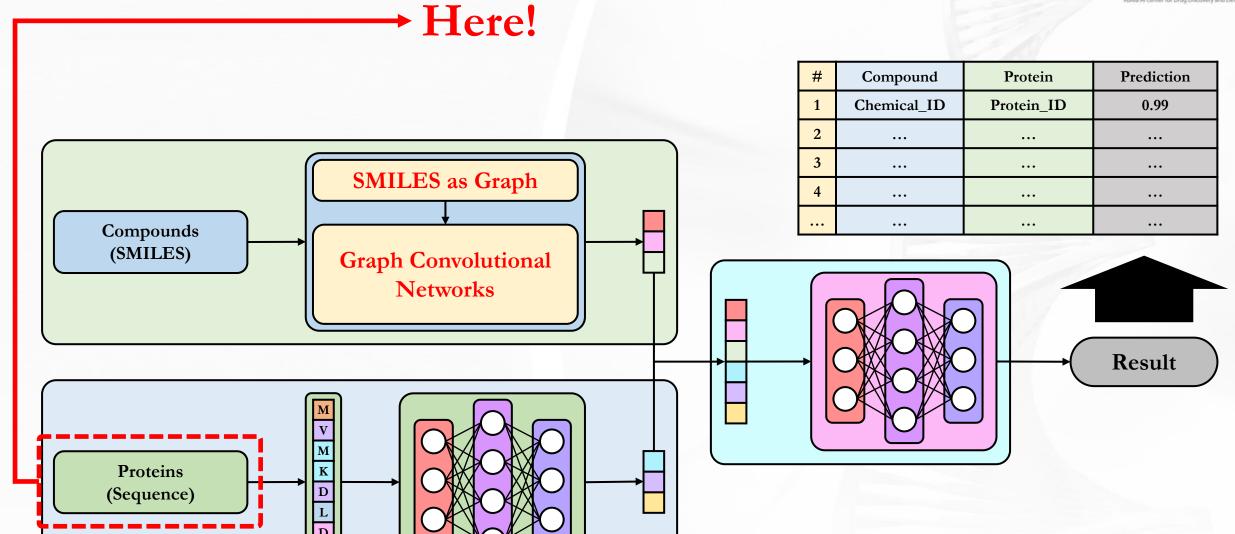
Kiba Length of Protein Sequence Distribution



# Protein Representation

- UniProt -

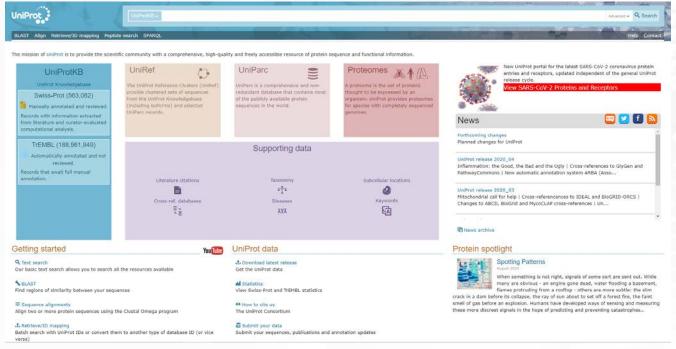




#### 한국인공지능신약개발지원센터 Koss Al Center for Drug Discovery and Development

## Protein Representation

- UniProt
  - ▶ 단백질 서열 및 기능 정보에 대한 자유롭게 액세스 할 수 있는 데이터베이스



https://www.uniprot.org/



## Protein Representation

- UniProt
  - ▶ 알파벳 'J'를 제외한 25개의 알파벳으로 표현된 아미노산 코드를 통해 단백질의 염기 서열 표현
  - ➤ 표준 아미노산 20개, 추가 아미노산 2개, 중복 표현 2개, Unspecified amino acid 1개

A	Alanine	F	Phenylalanine	L	Leucine	Q	Glutamine	V	Valine
В	D or N	G	Glycine	M	Methionine	R	Arginine	W	Tryptophan
С	Cysteine	Н	Histidine	N	Asparagine	s	Serine	X	Unspecified amino acid
D	Aspartic acid	I	Isoleucine	O	Pyrrolysine	Т	Threonine	Y	Tyrosine
E	Glutamic acid	K	Lysine	P	Proline	U	Selenocysteine	Z	E or Q

## KAICD 한국인공지능신약개발지원센터

## Protein Representation

• 예시 - Real Data



→ Protein\_Sequence

"PFWKILNPLLERGTYYYFMGQQPGKVLGDQRRPSLPALHFIKGAGKKESSRHGGPHCNVFVEHEALQRPVASDFEPQGLSEARWNSKENLLAGPSENDPNLFVALYDFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQTKNGQGWVPSNYITPVNSLEKHSWYHGPVSRNAAEYLLSSGINGSFLVRESES QRSISLRYEGRVYHYRINTASDGKLVVSSESBFNTLAELVHHHSTUHHFIVTVAKKPTVYGVSPNYDKWEMERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAKVTIKEDTMEVEEFLKEAAVMKEIKHPNLVQLLGVGTREPPFYIITETHTYGNLLDVLRECNRGEVNAVVLLYMATGISSAMEVIKFIHRDLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKSDVWAFGVLLWEIATYGMSPYPGIDLSQVYVELLKEDYTWEFVEEFLKEACAWWGWNPSDRPSFAEIHQAFETHMEQESSISDEVEKELGKQGVGCAVSTLLQAPELPTKTRTSRRAAEHRTTDYPEMFT KAGGESDPLDHEPAVSPLLPRKERGPPEGGLNEGLLPKDKKTNLFSALIKKKKKTAPTPPKRSSSFREMDGQPERRGAGEEEGRDISNGALAFTPLDTADPAKSPKPSNGAGVPNGALRESGGSGFRSPHLWKKSSTLTSSRLATGEEEGGGSSSKRFLRSCSASCVPHGAKDTEWRSVTLPRDLQSTGRQFFTFGGHKSEKPALPRKRAGENRSDQVTRGTVTPPPRLVKKNEEADEVFKDIMESSPGSSPPNLTPKPLRRQVTVAPASGLPHKEEAGKGSALGTPAAAEPVTPTSKAGSGAPGGTSKGPAEESRVRRHKHSSESPGRDKGKLSRLKPAPPPPPAASAGKAGGKPSQSPSQEAAGEAVLGAKTKATSLVDAVNSDAPSGEGLKKPVLPATPKPQSAKPSGTPISPAPVPSTLPSASSALAGDQPSSTAFIPLISTRVSLRKTRQPPERIASGAITKGVVLDSTEALCLAISRNSEQMASHSAVLEAGKNLYTFCVSYVDSIQQMRNKFAFREAINKLENNLRELQICPATAGSGPAATQDFSKLLSSVKEISDIVQR", "ABL1 (F3171)p":

"PFWKILNPLLERGTYYYFMGQQPGKVLGDQRRPSLPALHFIKGAGKKESSRHGGPHCNVFVEHEALQRPVASDFEPQGLSEAARWNSKENLLAGPSENDPNLFVALYDFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQTKNGQGWVPSNYITPVNSLEKHSWYHGPVSRNAAEYLLSSGINGSFLVRESESS QRSISLRYEGRVYHYRINTASDGKLYVSSESBFNTLAELVHHHSTVADGLITTLHYPAPKRNKPTVYGVSPNYDKWEMERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIKHPNLVQLLGOVTREPPFYIITEFMTYGNLLDYRECNRGVWAVVLYMATQISSAMEYL KKFIHRDLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKSDVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEGCPEKVYELMRACWGWNPSDRPSFAEIHQAFETMFGDTYTAHAGAKFPIKWTAPEKTATAHAGAKFPIKTTANGAKFPLDTADPAKSPKPSNGAGVPNSALRESGGSGFRSPHLWKKSSTLTSSRLSGEGGGSSSKRFLRSCSASCVPHGANTEWRSVTLPRDLQSTGRQFD TFGGHKSEKPALPRKRAGENRSDQVTRGTVTPPPRLVKKNEEAADEVFKDIMESSPGSSPPNLTPKPLRRQVTVAPASGLPHKEEAGKGSALGTPAAAEPVTPTSKAGSGAPGGTSKGPAESRVRRHKHSSESPGRDKGKLSRLKPAPPPPPPAASAGKAGGKPSQSPSQEAAGEAVLGAKTKATSLVDAVNSDA PSQPGGELKKPVLPATPKPQSAKPSGTPISPAPVPSTLPSASSALAGDQPSSTAFIPLISTRVSLRKTRQPPERIASGAITKGVVLDSTEALCLAISRNSEQMASHSAVLEAGKNLYTFCVSYVDSIQQMRNKFAFREAINKLENNLRELQICPATAGSGPAATQDFSKLLSSVKEISDIVQR", "ABL1 (23.11)".

"PFWKILNPLLERGTYYYFMGQQPGKVLGDQRRPSLPALHFIKGAGKKESSRHGGPHCNVFVEHEALQRPVASDFEPQGLSEAARWNSKENLLAGPSENDPNLFVALYDFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQTKNGQGWVPSNYITPVNSLEKHSWYHGPVSRNAAEYLLSSGINGSFLVRESES
QRSISLRYEGRYYHYRINTASDGKLYVSSESRENTLAELVHHHSTVADGLITTLHYPAPKRNKPTVYGVSPNYDKWEMERTDITMKHKLGGGQYGEVYGGWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIKHPNLVQLLGVCTREPPFYIITEFMTYGNLDVLRECNRQEVNAVVLLYMATQISSAMEY
KNFIHRDLAARNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKTAPESLAYNKFSIKSDVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMEPGCPEKVYLEHRACKWQNPSDRPSFAEIHQAFETHFQLGAFTHDFWATGHTMYPEM
KGQGESDPLDHEPAVSPLLPRKERGPPEGGLNEDERLLPKDKKKNLFSALIKKKKKAPTPPKRSSSFREMDGQPERRGAGEEGGFDISNGALAFTPLDTADPAKSPKPSNGAGVPNGALRESGGSGFRSPHLWKKSSTLTSSRLATGEEEGGGSSSKRFLRSCSASCVPHGARDTEWRSVTLPRDLQSTGRQP
TFGGHKSEKPALPRKRAGENRSDQVTRGTVTPPPRLVKKNEEAADEVFKDIMESSPGSSPPNLTPKPLRQVTVAPASGLPHKEEAGKGSALGTPAAAEPVTPTSKAGSGAPGGTSKGPAEESRVRRHKHSSESPGRDKGKLSRLKPAPPPPPAASAGKAGGKPSQSPSQEAAGEAVLGAKTKATSLVDAVNSD
PSQPGEGLKKPVLPATPKPQSAKPSGTPISPAPVPSTLPSASSALAGDQPSSTAFIPLISTRVSLRKTRQPPERIASGAITKGVVLDSTEALCLAISRNSEQMASHSAVLEAGKNLYTFCVSYVDSIQQMRNKFAFREAINKLENNLRELQICPATAGSGPAATQDFSKLLSSVKEISDIVQR", "ABL1

"PFWKILNPLLERGTYYYFMGQQPGKVLGDQRRPSLPALHFIKGAGKKESSRHGGPHCNVFVEHEALQRPVASDFEPQGLSEAARWNSKENLLAGPSENDPNLFVALYDFVASGDNTLSITKGEKLRVLGYNHNGEWCEAQTKNGQGWVPSNYITPVNSLEKHSWYHGPVSRNAAEYLLSSGINGSFLVRESES ORSISLRYEGRVYHYRINTASDGKLYVSSESRFNTLAELVHHHSTVADGLITTLHYPAPKRNKPTVYGVSPNYDKWEMERTDITMKHKLGGGOYGEVYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIKHPNLVQLLGVCTREPPFYIITEFMTYGNLLDYLRECNROEVNAVVLLYMATOISSAMEY

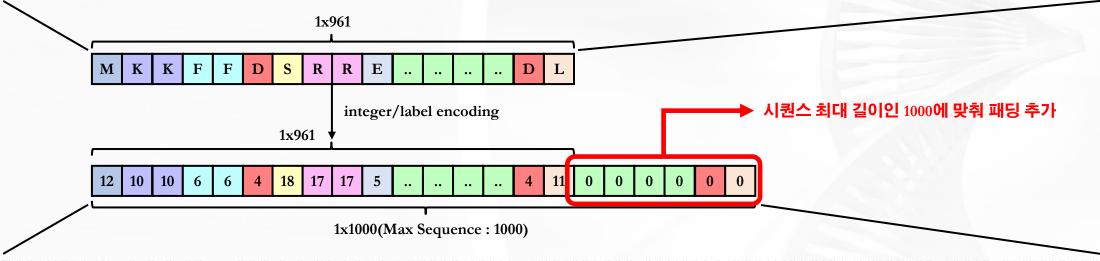
## Protein Representation



• Proteins to Input Presentation(Integer/Label Encoding)

#### AAK1:

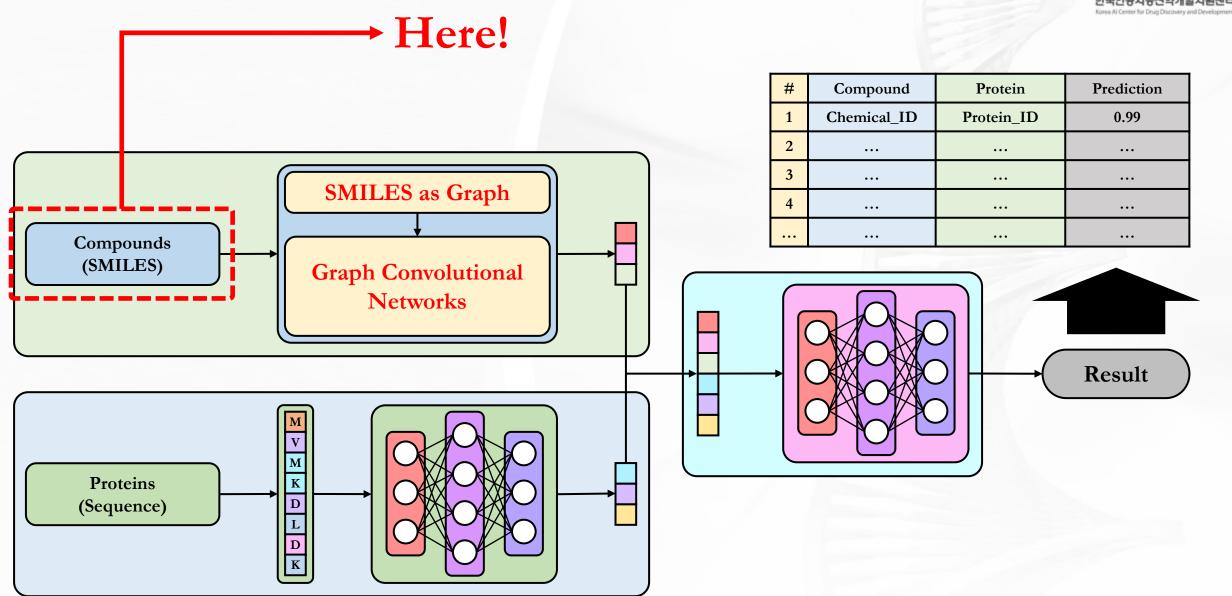
"MKKFFDSRREQGSGLGSGSSGGGSTSGLGSGYIGRVFGIGRQQVTVDEVLAEGGFAIVFLVRTSNGMKCALKRMFVNNEHDLQVCKREIQIMRDLSGHKNIVGYIDSSINNVSSGDVWEVLILMDFCRGQQVVNLMNQRLQTGFTENEVLQIFCDTCEAVARLHQCKTPIIHRDLKVENILLHDRGHYVLCDFGSATNK FQNPQTEGVNAVEDEIKKYTTI.SYRAPEMYNLYSGKIITTKADIWALGCLLYKLCYFTLPFGESQVAICDGNFTIPDNSRYSQDMHCLIRYMLEPDPDKRPDIYQVSYFSFKLLKKECPIPNVQNSPIPAKLPEPVKASEAAAKKTQPKARLTDPIPTTETSIAPRQRPKAGQTQPNPGILPIQPALTPRKRATVQPPPQAGGSNQ PGLLASVPQPKPQAPPSQPLPQTPSTQAGQLPAQAQATPQHQQQQQQQQQQQQQQLALATALHQQQLMTQQABQATPQHQQAQATPQHQAQAPPAQQPAQTPQFQAAAPQAPAPQAPAQPAQPAQPAGFVQQQQQQQQQQQQQQQQQQLALATALHQQQLMTQQAAQQQQQQQQQQQQQLATALHQQQLMTQQAALQKPTMAAGQQPQPQPAAAPQPAPAQEPAIQAPVRQQFKVQTPPPAVQGKVGSLTPPSSPKTQRAGHRRILSDVTHSAVFGVPASKSTQLLQAAAAEASLNKSKSATTTPSGSPRTSQQNVYNPSEGSTWNPFDDDNFSKLTAEELLNDFAKLGGSAESLIPGGSTQGDAFATTSFSAGTAEKRKGGQTVDSGLPLLSVSDPFIPLQVPDAPEKLIEGLKSPDTSLLLPDLLPMTDPFGSTSDAVIEKADVAVESLIPGLEPPVVQRLPSQTESVTSNRTDSLTGEDSLLCSLLSNPTTDLLEEFAPTAISAPVHKAAEDSNLISGFDVPEGSDKVAEDEFDPIPVLITKNPQGGHSRNSSGSSESSLPNLARSLLLVDQLIDL"





# Molecule Representation - SMILES -





# KAIOD 한국인공지능신약개발지원센터

## Molecule Representation

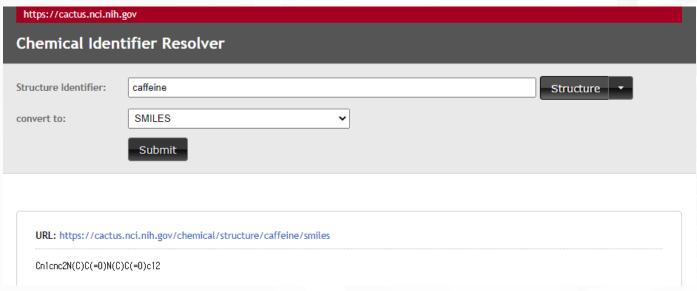
#### • 개요

- ▶ 화학 물질의 구조를 표현하는 방법으로 문자열, 그래프 등 다양한 형태가 존재
- ▶ 문자열로 표현하는 방법에서 대표적으로 WLN, ROSDAL, SMILES가 존재
- Davis & Kiba Datasets에서는 SMILES를 통해 화학 물질을 표현

# KAICO 한국인공지능신약개발지원센터

## Molecule Representation

- SMILES
  - > Simplified Molecular Input Line Entry System의 약자
  - 화학물질의 구조를 문자열로 나타내는 방법
  - ▶ 구조를 문자열로 변환할 때, 6가지 규칙 적용



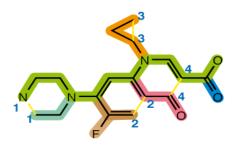
https://cactus.nci.nih.gov/chemical/structure

# ICAICD 한국인공지능신약개발지원센틱 Kores Al Center for Drug Discovery and Development

## Molecule Representation

#### • 규칙

- ▶ 원자는 표준 원소 기호로 표기
- ▶ 수소 원자는 가능한 모든 곳에 연결되어 있다고 가정, 표기 생략
- ▶ 이웃한 원자는 인접해서 표기
- ▶ 2중 결합은 '=', 3중 결합은 '#'으로 표기
- ▶ 가지는 '()'로 표기
- 고리는 고리를 생성하는 두 원자에 숫자를 표기(방향족 고리는 원자를 소문자로 표기)



N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

## Molecule Representation

#### • 특성

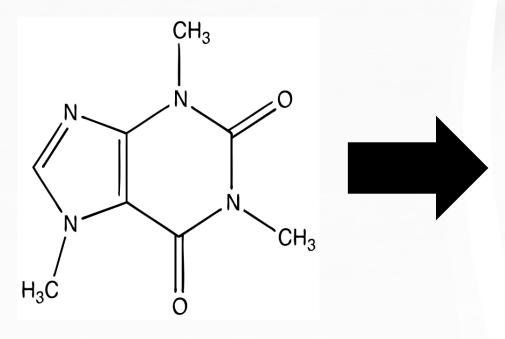
- > 원자(Atoms)
- ➤ 결합(Bonds)
- ▶ 고리(Rings)
- > 방향족성(Aromaticity)
- ➤ 분기(Branching)
- > 입체배열(Stereochemistry)
- > 동위원소(Isotopes)



# ICAICD 한국인공지능신약개발지원센터 Kana Al Center for Drug Discovery and Development

## Molecule Representation

예시



Cn1cnc2N(C)C(=O)N(C)C(=O)c12
[SMILES]

Caffeine - C<sub>8</sub>H<sub>10</sub>N<sub>4</sub>O<sub>2</sub>

[Molecule Structure]

# KAICD 한국인공지능신약개발지원센터 None Al Center for Drug Discovery and Development

## Molecule Representation

• 예시 - Real Data



0C4cccc4 , 5287969 : CN1CCC(C(C1)0)C2=C(C=C(C3=C20C(=CC3=C)C4CC)0)C2=C(=0)N(C2=CN=C)N(C3=C)C1=CC=C=C)N(C3=CN=C)N(C2=CN=C)N(C3= NC4=C(C=C(C=C4)C(=0)NC5CCN(CC5)C)OC)C", "9926054": "CC1=CC2=C(C=C1)N=C(C3=NC=C(N23)C)NCCN.Cl", "16007391": "CCN(CCC0C1=CC2=C(C=C1)C(=NC=N2)NC3=NNC(=C3)CC(=0)NC4=CC(=CC=C4)F)CC0", "5328940": "CN1CCN(CC1)CCCOC2=C(C=C3C(=C2)N=CC(=C3NC4=CC(=C(C=C4Cl)Cl)OC)C#N)OC", "11234052": "CC1=CC2=C(N1)C=CC(=C2F)OC3=NC=NN4C3=C(C(=C4)OCC(C)O)C", "11656518": "CN1C2=C(C=C(C=C2)OC3=CC(=NC=C3)C4=NC=C(N4)C=C(N4)C=C(N4)C=C(C)O(N4) (F)(F)F)N=C1NC5=CC=C(C=C5)C(F)(F)F", "6918454": "C1CC1CONC(=0)C2=C(C(=C2)F)F)NC3=C(C=C2)F)F)NC3=C(C=C3)I)Cl", "156414": "C=CC(=0)NC1=C(C=C2)C(=C1)C(=N2)NC3=CC(=C3)F)Cl)OCCCN4CCOCC4", "9933475": "CC1=CC2=C(N1)C=CC(=C2F)OC3=NC=NC4=CC(=C(C=C43)OC)OCCCN5CCCC5", "11626560": "CC(C1=C(C=CC(=C1Cl)F)Cl)OC2=C(N=CC(=C2)C3+CN(N=C3)C4CCNCC4)N", "3062316": "CC1=C(C(=CC=C1)Cl)NC(=0)C2=CN=C(S2)NC3=NC (=C3)N4CCN(CC4)CCO)C", "156422": "CC1=CC=C(C=C1)N2C(=CC(=N2)C(C)C)NC(=O)NC3=CC=C(C4=CC=CC=C43)OCCN5CCOCC5", "44150621": "CC(C(=0)0)0.CN1CCN(CC1)C2=CC3=C(C=C2)NC(=C4C(=C5C(=NC4=0)C=CC=C5F)N)N3 .0", "176167": "CN1C=C(C2=CC=CC1)C3=C(C(=0)NC3=0)C4=CN(C5=CC=CC=C54)C6CCN(CC6)CC7=CC=CC=N7", "176870": "COCCOC1=C(C=C2)(=C1)C(=NC=N2)NC3=CC=CC(=C3)C#C)OCCOC", "42642645": "COC1=CC2=C  $\frac{C2 - CC3 - C(C - C2)N - CN - C3NC4 - CC(-C(C - C4)OCC5 - CC(-CC - C5)F)Cl", "126565": "CC12C(CC(O1)N3C4 - CC - CC2)C3 - CC2)C3 - CC2 - CC2)N - CC3 - CC3 - CC3 - CC4 - CC3 - CC4 - CC4$ (=CC=C3)NN=C4N", "9929127": "CC1=C(C=CC=N1)C(=0)NC2=C3C(=CC(=C2OC)Cl)C4=C(N3)C=NC=C4", "11712649": "C1C2=CN=C(N=C2C3=C(C=C(C=C3)Cl)C(=N1)C4=C(C=CC=C4F)F)NC5=CC=C(C=C5)C(=0)0", "10074640": "CC1=C(C=C(C=C1)NC(=0)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC(=CS4)C5=CN=CC=C5", "51004351": "CC12C(C(CC(01)N3C4=CC=CC=C4C5=C6C(=C7C8=CC=C8N2C7=C53)CNC6=0)N(C)C(=0)C9=CC=CC=C9)OC", "11667893": "CC1(CNC2=C1C=CC(=C2)NC(=0)C3=C(N=CC=C3)NCC4=CC=NC=C4)C", "9915743": "CC0C1=C(C=C2C(=C1)N=CC(=C2NC3=CC(=C(C=C3)0CC4=CC=CC=N4)Cl)C#N)NC(=0)C=CCN(C)C", "644241": "CC1=C(C=C(C=C1)C(=C)NC2=CC(=CC(=C1)C( N3C=C(N=C3)C)C(F)(F)F)NC4=NC=CC(=N4)C5=CN=CC=C5", "447077": "CN1C2=NC(=NC=C2C=C(C1=0)C3=C(C=CC=C3Cl)Cl)NC4=CC(=CC=C4)SC", "10461815": "CC1=C(NC(=C1C(=0)N2CCCC2CN3CCCC3)C)C=C4C5=C(C=C4C(=C)S(=0)N2CCCC2CN3CCCC3)C)C=C4C5=C(C=C5)S(=0)N2CCCC2CN3CCCC3)CNCCCC3 CC6=C(C=CC=C6Cl)Cl)NC4=0", "9884685": "C1C0CCN1C2=NC(=NC3=C20C4=C3C=CC=N4)C5=CC(=CC=C5)0", "24180719": "CCCS(=0)(=0)NC1=C(C(=C(C=C1)F)C(=0)C2=CNC3=NC=C(C=C23)Cl)F", "25243800": "CC(C)N1C2=C(C=C1)F)C(=0)C2=CNC3=NC=C1)F)C(=0)C2=CNC3=NC=C1)F)C(=0)C2=CNC3=NC=C1)F)C(=0)C2=CNC3=NC=C1)F)C(=0)C2=CNC3=NC=C1)F)C(=0)C2=CNC3=NC=C1)F)C(=0)C2=CNC3=NC=C1)F)C(=0)C2=CNC3=NC=C1)F)C(=0)C N6CCOCC6", "11984591": "CC1(C(=0)NC2=C(01)C=CC(=N2)NC3=NC(=NC=C3F)NC4=CC(=C(C(=C4)OC)OC)OC)C.C1=CC=C(=C1)S(=0)(=0)0", "153999": "CN(C)CC1CCN2C=C(C3=CC=CC=C32)C4=C(C5=CN(CC01)C6=CC=CC=C65)C(=0) NC4=0", "25127112": "C1CCC(C1)C(CC#N)N2C=C(C=N2)C3=C4C=CNC4=NC=N3.0P(=0)(0)0", "176155": "CS(=0)C1=CC=C(C=C1)C2=NC(=2)C3=CC=NC=C3)C4=CC=C(C=C4)F", "24779724": "CN1C=C(C=N1)C2=NN3C (=NN=C3SC4=CC5=C(C=C4)N=CC=C5)C=C2", "3025986": "CC(C)(C)C1=CN=C(01)CSC2=CN=C(S2)NC(=0)C3CCNCC3", "10138260": "CC1=C(NC(=C1C(=0)NCC(CN2CCCC2)0)C)C=C3C4=C(C=CC(=C4)F)NC3=0", "10127622": "CN1C=NC2=C1C=C(C(=C2F)NC3=C(C=C(C=C3)Br)Cl)C(=0)NOCCO", "216239": "CNC(=0)C1=NC=CC(=C1)OC2=CC=C(C=C2)NC(=0)NC3=CC(=C(0=C3)Cl)C(F)(F)F", "44259": "CC12C(C(CC(01)N3C4=CC=CC=C4C5=C6C (=C7C8=CC=CC8N2C7=C53)CNC6=0)NC)OC", "5329102": "CCN(CC)CCNC(=0)C1=C(NC(=C1C)C=C2C3=C(C=CC(=C3)F)NC2=0)C", "16038120": "CC(C)S(=0)(=0)C1=CC=CC=C1NC2=NC(=NC=C2C1)NC3=C(C=C(C=C3)N4CCC(CC4)N5CCN (CC5)C)OC", "10427712": "C1=CC(=C1)O)C2=NC3=C(N=C2C4=CC(=CC+O))N=C(N=C3N)N", "16722836": "CC1=CN=C(N=C1NC2=CC(=CC+C2)S(=0)(=0)NC(C)(C)C)NC3=CC=C(C=C3)OCCN4CCCC4", "3038522": "CC(C)OC1=CC=C4)OC1=C4)OC1=C4)OC1=C4)OC1=C5OC1=C5)OC1=C5,OC1=C5 (C=C1)NC(=0)N2CCN(CC2)C3=NC=NC4=CC(=C(C=C43)OC)OCCCN5CCCCC5", "9926791": "CC1CCN(CC1N(C)C2=NC=NC3=C2C=CN3)C(=0)CC#N", "5494449": "CC1=CC(=NN1)NC2=NC(=C(=C2)N3CCN(CC3)C)SC4=CC=C(C=C4)NC(=0)C5CC

LIMITATION

#### 한국인공지능신약개발지원센틱 Korea Al Center for Drug Discovery and Development

## Molecule Representation

#### • 한계점

- ▶ 문자열로 이루어진 SMILES 특성 상, 복잡한 화학 구조를 제대로 반영하지 못함
- ▶ 화학 구조를 제대로 반영하지 못했다는 것은 정보 유실을 의미
- ▶ 방대한 화합물 정보를 제대로 이용하지 못함
- ▶ 신약후보물질 예측의 한계가 존재

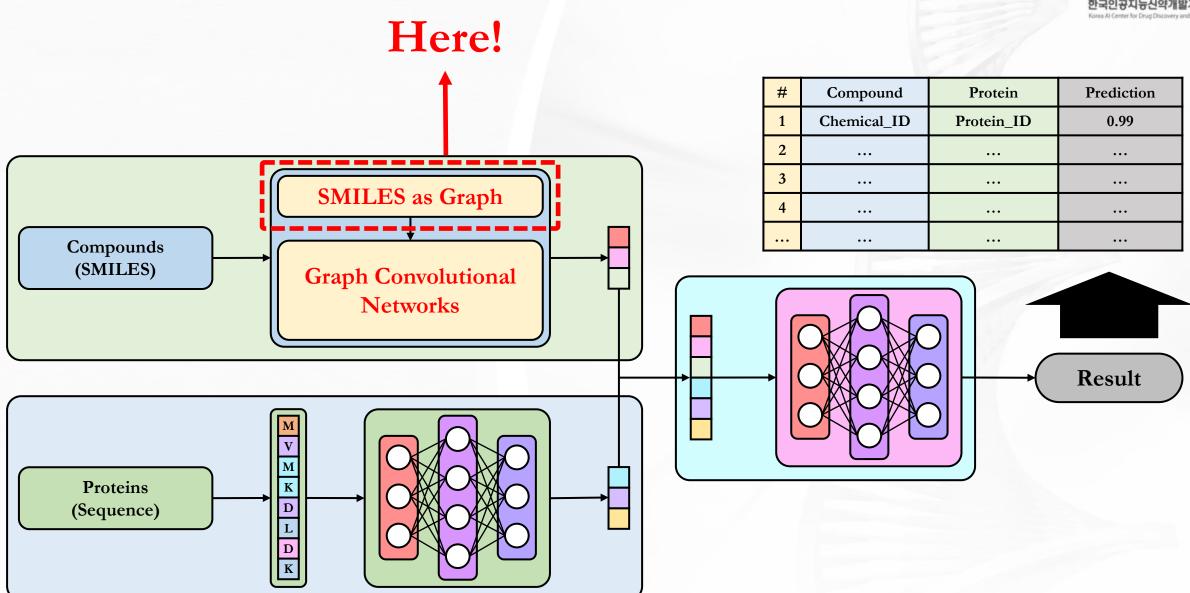


# Molecule Representation

- Graph -

Hyeonsu Lee (KAICD) GraphDTA October 28, 2020

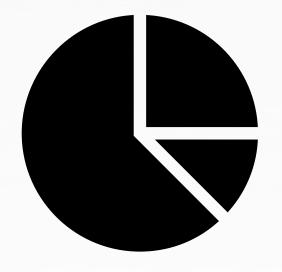


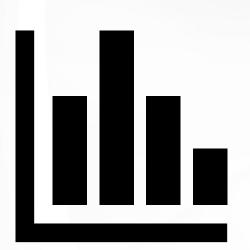


#### KAICD 한국인공지능신약개발지원센터 Kone Al Center for Drug Discovery and Development

#### Graph

- Graph Statistical Graph
  - > 어떠한 데이터들을 그림상으로 시각화하여 나타낸 것을 의미



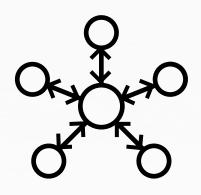




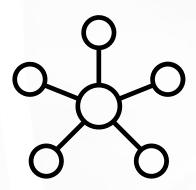
October 28, 2020

#### KAICD 한국인공지능신약개발지원센터 Korea Al Center for Drug Discovery and Development

- Graph Mathematical Graph
  - > 일부 객체들의 쌍들이 서로 연관된 객체의 집합을 이루는 구조
  - ▶ 방향 / 無방향, 가중치 그래프로 구분



Directed graph

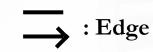


Undirected graph



Weighted graph



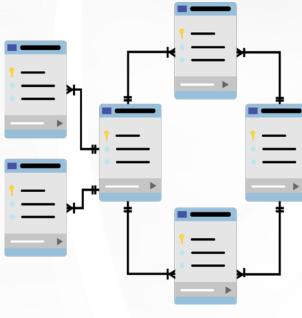


**REAL CASE GRAPH** 

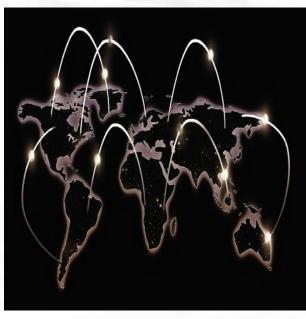
- Graph Real Case
  - ▶ 객체를 정점, 객체 사이의 관계를 간선으로 표현



사회 관계망

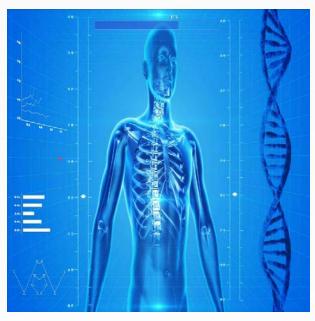


관계형 데이터베이스



지도

- Graph Real Case
  - ▶ 객체를 정점, 객체 사이의 관계를 간선으로 표현



인체 구조



지식

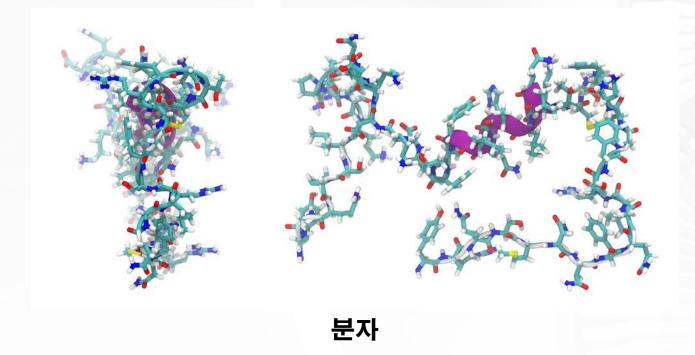


우주

REAL CASE GRAPH

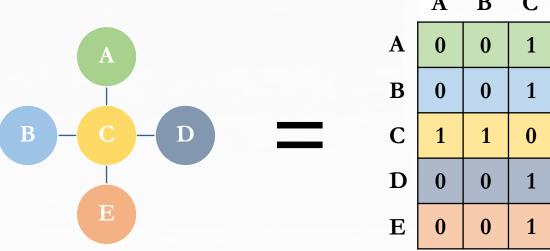
#### 는 AICD 한국인공지능신약개발지원센터 Kates AI Center for Drug Discovery and Development

- Graph Real Case
  - ▶ 객체를 정점, 객체 사이의 관계를 간선으로 표현





- Graph Graph Data Structure
  - ▶ 정점(Vertex or Node)과 그 정점을 연결하는 간선(Edge)을 하나로 모아 놓은 비선형 자료 구조
  - 정점 정보는 특성 행렬, 간선 정보는 인접 행렬로 표현



	A	В	С	D	$\mathbf{E}$	
A	0	0	1	0	0	
В	0	0	1	0	0	
C	1	1	0	1	1	
D	0	0	1	0	0	
$\mathbf{E}$	0	0	1	0	0	

인접 행렬

	1	2	3	4	5	6
A	0	1	0	1	0	1
В	1	1	1	0	0	1
C	1	0	0	0	1	0
D	1	1	1	0	0	1
E	0	0	1	1	0	0

특성 행렬

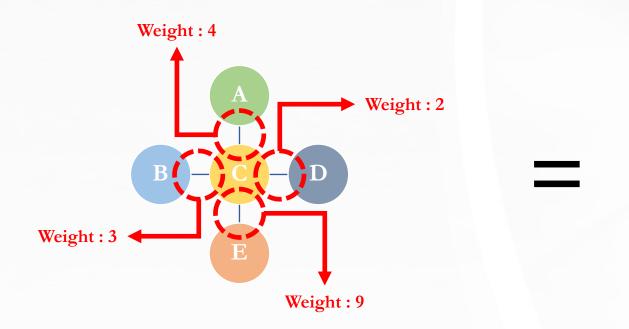


- Graph Adjacency & Feature
  - ▶ 인접 행렬 : 그래프에서 어느 정점들이 간선으로 연결되었는지 나타내는 정사각 행렬
  - ▶ 특성 행렬 : 그래프에서 정점들이 가지고 있는 특성들을 나타내는 정사각 행렬



#### KAICD 한국인광지능신약개발지원센터 Kote Al Center for Drug Discovery and Development

- Graph Weight Matrix
  - ▶ 가중치 행렬 : 인접 행렬을 이용하여 표현
  - ▶ 간선이 존재하지 않는 경우, 가중치 범위 밖의 값 INF 등으로 표현



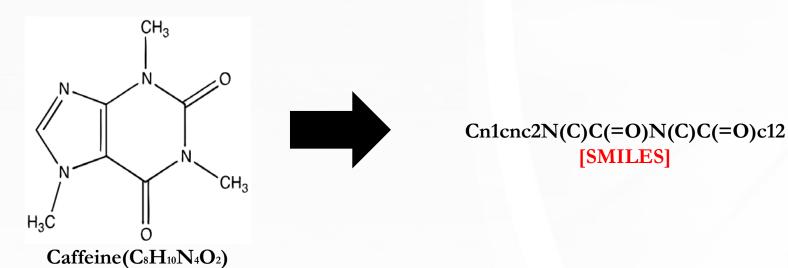
	A	В	C	D	E
A	INF	INF	4	INF	INF
В	INF	INF	3	INF	INF
C	4	3	INF	2	9
D	INF	INF	2	INF	INF
$\mathbf{E}$	INF	INF	9	INF	INF

인접 행렬(가중치)

## ICAICD 한국인공지능신약개발지원센터 Name Al Center for Drug Discovery and Development

### Graph - Molecule Representation

- SMILES as Graph
  - ▶ RDKit 라이브러리를 활용
  - ▶ 화학 물질의 정보를 담고 있는 데이터를 활용하여 구조 이미지(구조 식) 생성
  - ▶ 문자열로 구성된 SMILES를 신경망 학습을 위한 그래프 형식으로 변환





0	0	1	0	0
0	0	1	0	0
1	1	0	1	1
0	0	1	0	0
0	0	1	0	0

0	1	0	1	0
1	1	1	0	0
1	0	0	0	1
1	1	1	0	0
0	0	1	1	0

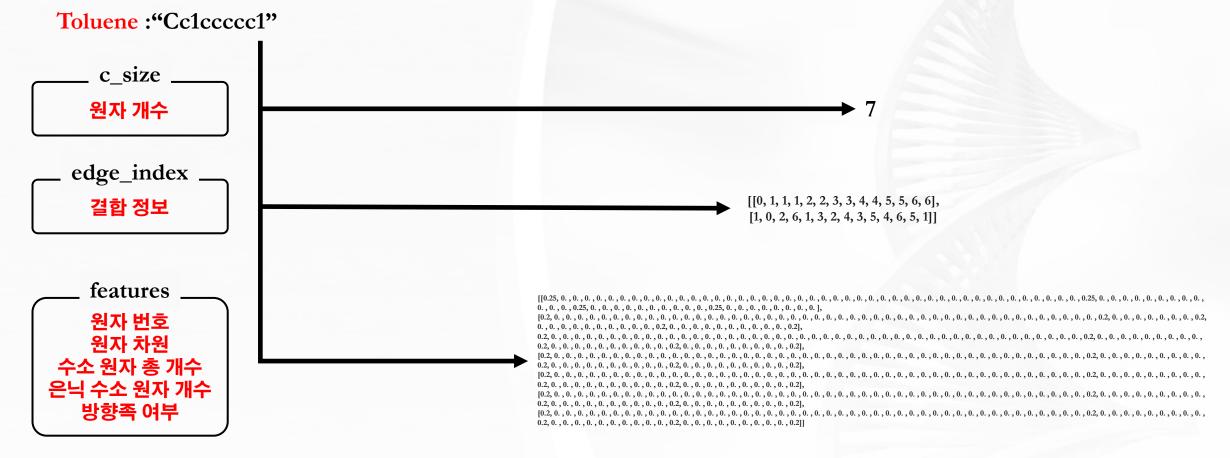
[Matrix]

[Molecule Structure]

## 

### Graph - Molecule Representation

• SMILES to Input Presentation(Graph Representation)



{"Cc1cccc1": c\_size, edge\_index, features}

## KAICD 한국인공지능신약개발지원센트 Notes Al Center for Drug Discovery and Divelogment

- Graph Problem & Solution
  - ▶ 분자처럼 원자의 속성, 연결의 종류 등을 고려해야하는 경우, 속성 그래프로 데이터를 표현해야 함
  - > 속성 그래프 데이터는 기존 방식으로 벡터의 형태로 변환하는 것이 불가능
  - ▶ 벡터의 형태로 데이터를 입력 받는 기존의 인공신경망으로는 분자 그래프 데이터를 처리할 수 없음
  - ➤ Graph convolution을 이용하여 정점 또는 그래프 자체를 벡터 형태의 데이터로 변환하여 해결



# Graph Neural Networks

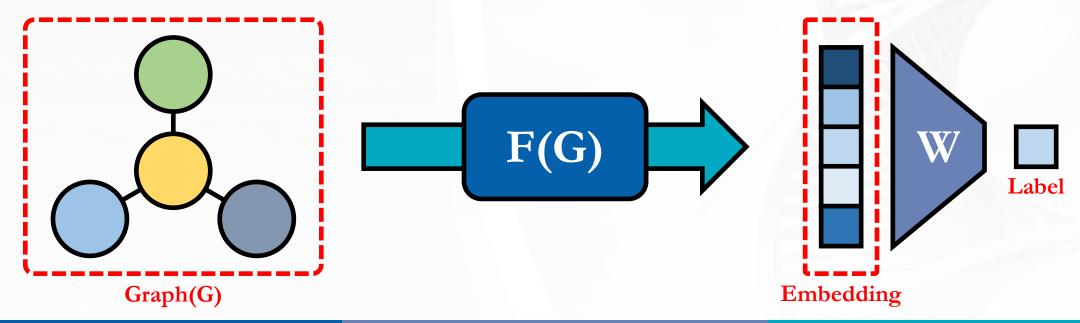
Hyeonsu Lee (KAICD) GraphDTA October 28, 2020



#### Graph Neural Network

#### • 개요

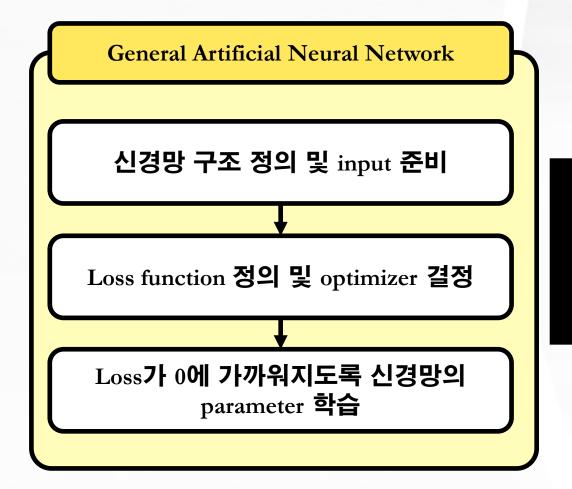
- 그래프 구조에서 사용하는 인공 신경망
- ▷ 입력이 그래프 구조라는 특징을 지님
- ▶ 그래프에 존재하는 정점 사이의 관계를 모델링하고, 이에 대한 표현을 생성하는 것이 핵심



## 한국인공지능신약개발지원센터 Kasa Al Center for Drug Discovery and Development

#### Graph Neural Network

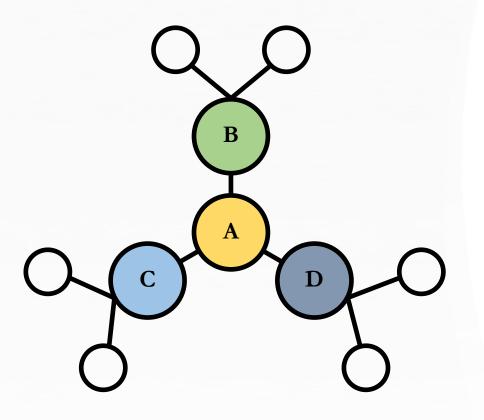
• 학습 과정

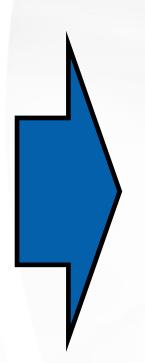


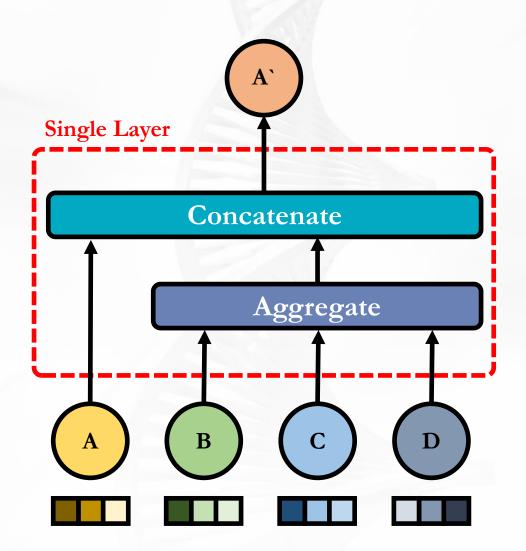
# **Graph Neural Network** Aggregate, Concatenate 함수 정의 및 graph input 준비 Loss function 정의 및 optimizer 결정 Loss가 0에 가까워지도록 신경망의 Aggregate, Concatenate parameter 학습

### Graph Neural Network

• Neighborhoods Aggregation



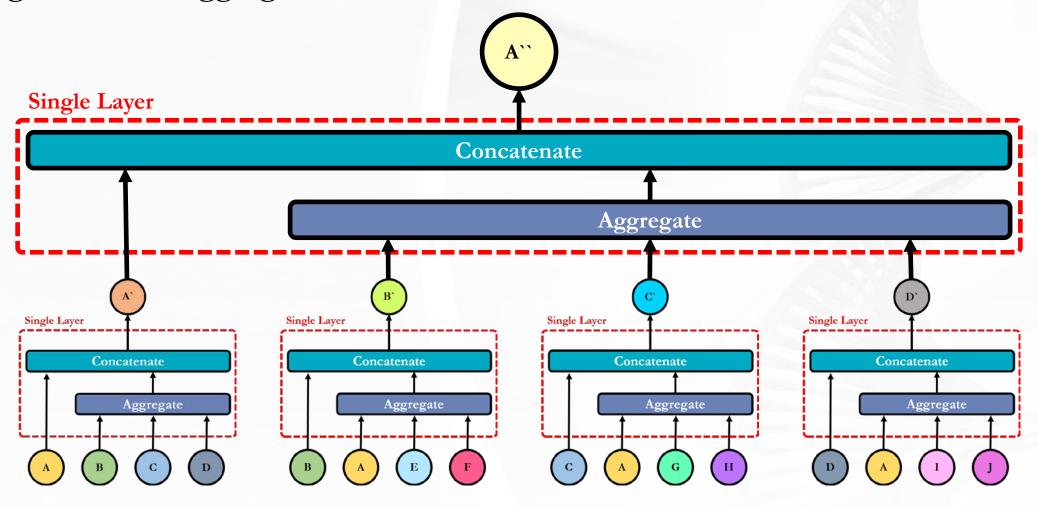




#### 한국인공지능신약개발지원센터 Kises Al Center for Drug Discovery and Development

### Graph Neural Network

• Neighborhoods Aggregation





#### Graph Neural Network

• Algorithm

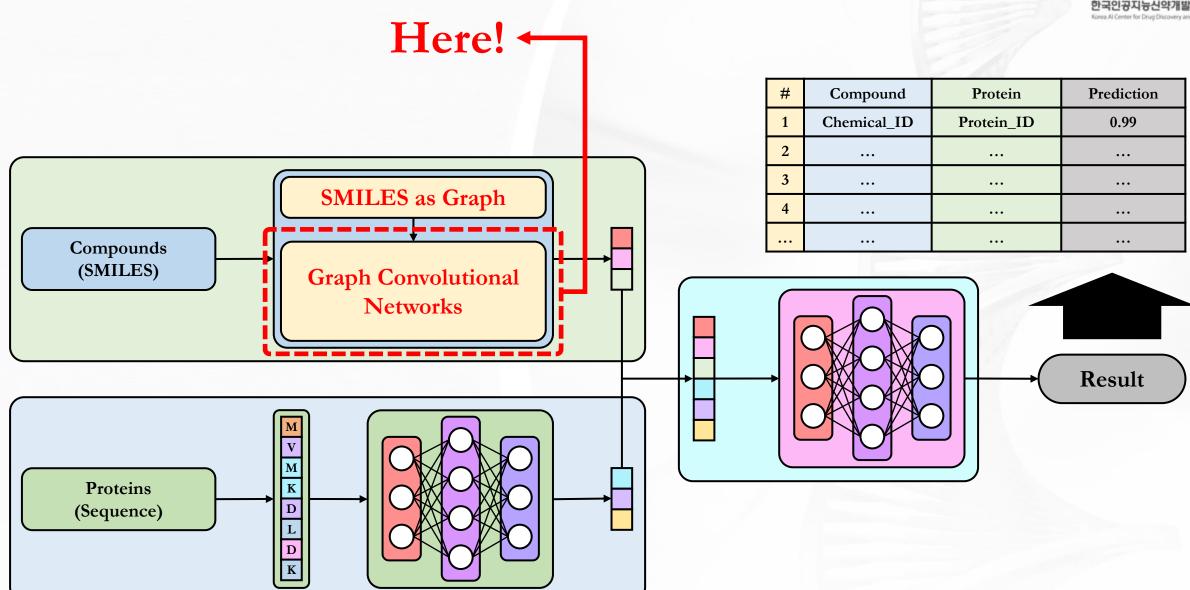
```
for v in V:
       h_{v}^{0} = v's feature vector
for i in range(1, k+1):
        for v in V:
                a = Aggregate(\{h_{u}^{i-1}\} \mid \{u, v\} \text{ in } E\})
                h_{v}^{i} = Concatenate(h_{v}^{i-1}, a)
```



# Graph Convolutional Networks

Hyeonsu Lee (KAICD) GraphDTA October 28, 2020







#### Convolution

#### • 개요

- ▶ 행렬 내에서 특성을 뽑기 위한 연산
- ▶ 입력 행렬과 특성 탐지를 통해 입력에 대한 특성 지도를 생성
- ▶ 특성 지도를 통해 입력 행렬의 패턴 파악





0	5	0	5	0
0	7	1	7	0
1	4	9	4	1
0	3	2	3	0
0	1	1	1	0



1	3	0
0	2	1
0	2	3

45	73	
17	93	
68	63	

35

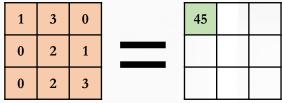


#### Convolution

• 합성곱 연산 과정(Matrix)

0	)	5	0	5	0
0	)	7	1	7	0
1		4	9	4	1
0	)	3	2	3	0
0	)	1	1	1	0

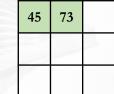




0	5	0	5	0
0	7	1	7	0
1	4	9	4	1
0	3	2	3	0
0	1	1	1	0



			_
1	3	0	
0	2	1	
0	2	3	



0	5	0	5	0
0	7	1	7	0
1	4	9	4	1
0	3	2	3	0
0	1	1	1	0



1	3	0	45	73	12
0	2	1			
0	2	3			

0	5	0	5	0
0	7	1	7	0
1	4	9	4	1
0	3	2	3	0
0	1	1	1	0

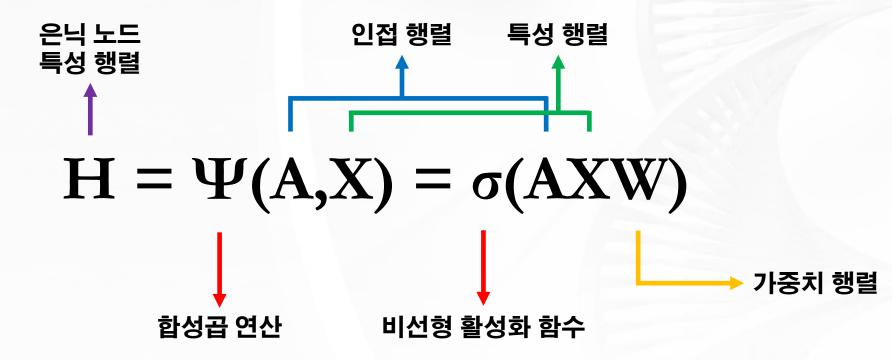


1	3	0	
0	2	1	
0	2	3	

## ICAICD 한국인공지능신약개발지원센터 Kites Al Center for Drug Discovery and Development

#### **Graph Convolution**

- 특정 노드의 은닉 벡터를 해당 노드의 이웃 정보를 통해 표현
  - > 그래프에 포함된 노드나 그래프 자체를 벡터 형태의 데이터로 변환
  - ▶ 노드의 특성과 연결 정보만을 고려하기 때문에 간선 특성은 사용되지 않음

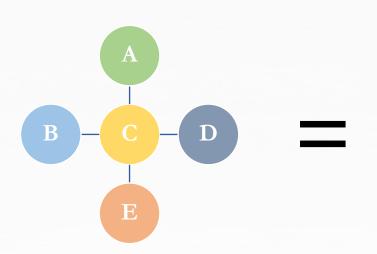


#### ICAICD 한국인공지능신약개발지원센터 Kanes Al Center for Drug Discovery and Development.

#### **Graph Convolution**

• 행렬 계산 예시

$$G = (A, X)$$



	A	В	С	D	E
A	0	0	1	0	0
В	0	0	1	0	0
C	1	1	0	1	1
D	0	0	1	0	0
$\mathbf{E}$	0	0	1	0	0
				_	

인접 행렬(A)

		1	2	3	4	5	6
	A	0	1	0	1	0	1
	В	1	1	1	0	0	1
,	C	1	0	0	0	1	0
	D	1	1	1	0	0	1
	E	0	0	1	1	0	0

특성 행렬(X)

October 28, 2020

## ICAICD 한국인공지능신약개발지원센터 Kona Al Center for Drug Discovery and Development

#### **Graph Convolution**

• 행렬 계산 예시

$$A \times X = H$$

	A	В	С	D	E
A	0	0	1	0	0
В	0	0	1	0	0
C	1	1	0	1	1
D	0	0	1	0	0
$\mathbf{E}$	0	0	1	0	0

X

	1	2	3	4	5	6
A	0	1	0	1	0	1
В	1	1	1	0	0	1
C	1	0	0	0	1	0
D	1	1	1	0	0	1
$\mathbf{E}$	0	0	1	1	0	0

특성 행렬(X)

	1	2	3	4	5	6
A	1	0	0	0	1	0
В	1	0	0	0	1	0
C	2	3	3	2	0	3
D	1	0	0	0	1	0
E	1	0	0	0	1	0

은닉 정점 특성 행렬(H)

인접 행렬(A)



#### **Graph Convolution**

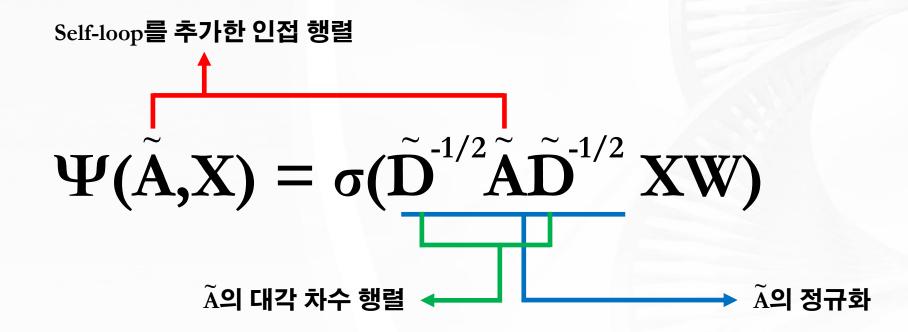
- 문제점
  - ▶ 인접 행렬에는 노드의 연결만 표현되어 있음
    - 합성곱 연산에서 각 노드 자체에 대한 정보 유실
  - ▶ 인접 행렬은 정규화 되어 있지 않음
    - 특성 벡터와 곱 연산을 수행할 경우 크기가 <mark>불안정</mark>하게 변할 수 있음

## 인접 행렬의 한계

## ICAICD 한국인공지능신약개발지원센터 Kites Al Center for Drug Discovery and Development

### **Graph Convolution**

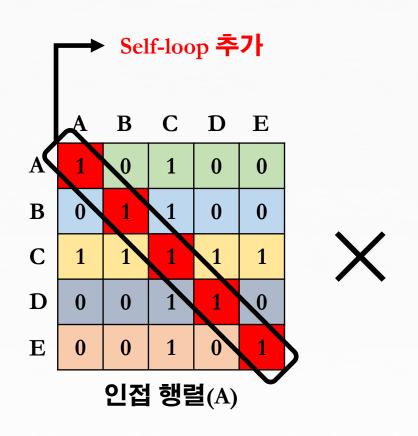
- 해결 방안 적용
  - ➤ 인접 행렬에 self-loop 추가
  - ▶ 인접 행렬을 D<sup>-1/2</sup>AD 로 정규화



## ICAICD 한국인공지능신약개발지원센터 Kones Al Center for Drug Discovery and Development

### **Graph Convolution**

• Self-loop 추가





특성 행렬(X)

0

0

D

 $\mathbf{E}$ 

	1	2	3	4	5	6
A	1	1	0	1	1	1
В	1	1	1	0	1	1
C	3	3	3	2	1	3
D	1	1	1	0	1	1
E	1	0	1	1	1	0

은닉 정점 특성 행렬(H)

#### **Graph Convolution**

• 인접 행렬 정규화



В

C D E

0

0

0

0

0

	A	В	C	D	E		
A	$\frac{1}{\sqrt{2}}$	0	0	0	0		
В	0	$\frac{1}{\sqrt{2}}$	0	0	0		
C	0	0	$\frac{1}{\sqrt{5}}$	0	0		
D	0	0	0	$\frac{1}{\sqrt{2}}$	0		
$\mathbf{E}$	0	0	0	0	$\frac{1}{\sqrt{2}}$		
차수 행렬( $\tilde{ extbf{D}}^{ ext{-}^{1/2}}$ )							

0 B D 0  $\mathbf{E}$ 인접 행렬 $(\tilde{A})$ 

	A	В	C	D	$\mathbf{E}$
A	$\frac{1}{\sqrt{2}}$	0	0	0	0
В	0	$\frac{1}{\sqrt{2}}$	0	0	0
C	0	0	$\frac{1}{\sqrt{5}}$	0	0
D	0	0	0	$\frac{1}{\sqrt{2}}$	0
E	0	0	0	0	$\frac{1}{\sqrt{2}}$

차수 행렬 $(\tilde{\mathbf{D}}^{-1/2})$ 

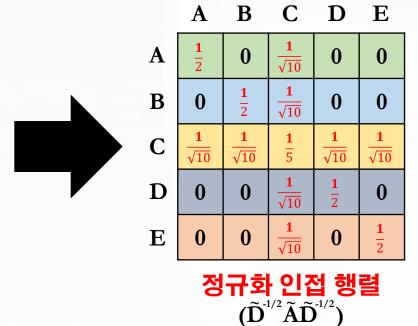
## ICAICD 한국인공지능신약개발지원센터 Kones Al Center for Drug Discovery and Development

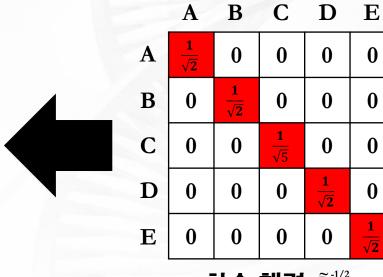
#### **Graph Convolution**

• 인접 행렬 정규화

$$\mathbf{p}^{\sim -1/2} \overset{\sim}{\mathbf{A}} \overset{\sim}{\mathbf{p}}^{-1/2}$$

	A	В	C	D	$\mathbf{E}$	
A	$\frac{1}{\sqrt{2}}$	0	0	0	0	
В	0	$\frac{1}{\sqrt{2}}$	0	0	0	
C	0	0	$\frac{1}{\sqrt{5}}$	0	0	
D	0	0	0	$\frac{1}{\sqrt{2}}$	0	
$\mathbf{E}$	0	0	0	0	$\frac{1}{\sqrt{2}}$	
차수 행렬( $\tilde{ ext{D}}^{ ext{-}1/2}$ )						





## ICAICD 한국인공지능신약개발지원센터 Name Al Center for Drug Discovery and Development

#### **Graph Convolution**

• 최종 계산

$$\mathbf{D}^{\tilde{\mathbf{X}}} \overset{1/2}{\mathbf{A}} \times \tilde{\mathbf{D}}^{1/2} \times \mathbf{X} = \mathbf{H}$$

	A	В	C	D	E
A	$\frac{1}{2}$	0	$\frac{1}{\sqrt{10}}$	0	0
В	0	$\frac{1}{2}$	$\frac{1}{\sqrt{10}}$	0	0
C	$\frac{1}{\sqrt{10}}$	$\frac{1}{\sqrt{10}}$	1/5	$\frac{1}{\sqrt{10}}$	$\frac{1}{\sqrt{10}}$
D	0	0	$\frac{1}{\sqrt{10}}$	$\frac{1}{2}$	0
$\mathbf{E}$	0	0	$\frac{1}{\sqrt{10}}$	0	$\frac{1}{2}$

정규화 인접 행렬

	1	Z	3	4	5	0
A	0	1	0	1	0	1
В	1	1	1	0	0	1
C	1	0	0	0	1	0
D	1	1	1	0	0	1
$\mathbf{E}$	0	0	1	1	0	0

특성 행렬(X)

	-	_			•	Ĭ
A	$\frac{1}{\sqrt{10}}$	1/2	0	1/2	$\frac{1}{\sqrt{10}}$	1/2
В	$\frac{2+\sqrt{10}}{2\sqrt{10}}$	1 2	$\frac{1}{2}$	0	$\frac{1}{\sqrt{10}}$	$\frac{1}{2}$
C	$\frac{10 + \sqrt{10}}{5\sqrt{10}}$	$\frac{3}{\sqrt{10}}$	$\frac{3}{\sqrt{10}}$	$\frac{2}{\sqrt{10}}$	1 5	$\frac{3}{\sqrt{10}}$
D	$\frac{2+\sqrt{10}}{2\sqrt{10}}$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{\sqrt{10}}$	$\frac{1}{2}$
E	$\frac{1}{\sqrt{10}}$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{\sqrt{10}}$	0

3

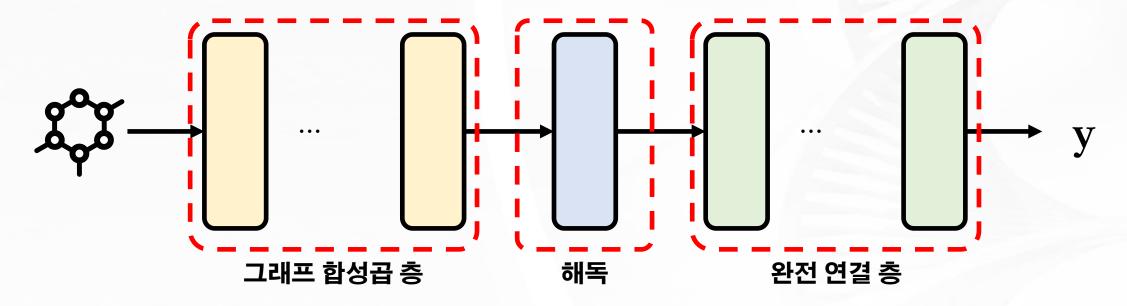
은닉 정점 특성 행렬(H)

#### | CAICD 한국인광지능신약개발지원센터 Notes Al Center for Drug Discovery and Development

### **Graph Convolutional Networks**

#### • 개요

- ▶ 일반적으로 그래프 합성곱 층과 완전 연결 층으로 구성
- ▶ 그래프 분류 및 회귀 문제에서는 해독 과정이 필수
- > 노드 분류 또는 링크 예측의 경우 해독 과정이 필요 없음



#### | CAICD 한국인공지능신약개발지원센터 Notes Al Center for Drug Discovery and Development

#### Graph Convolutional Networks

#### • 연산

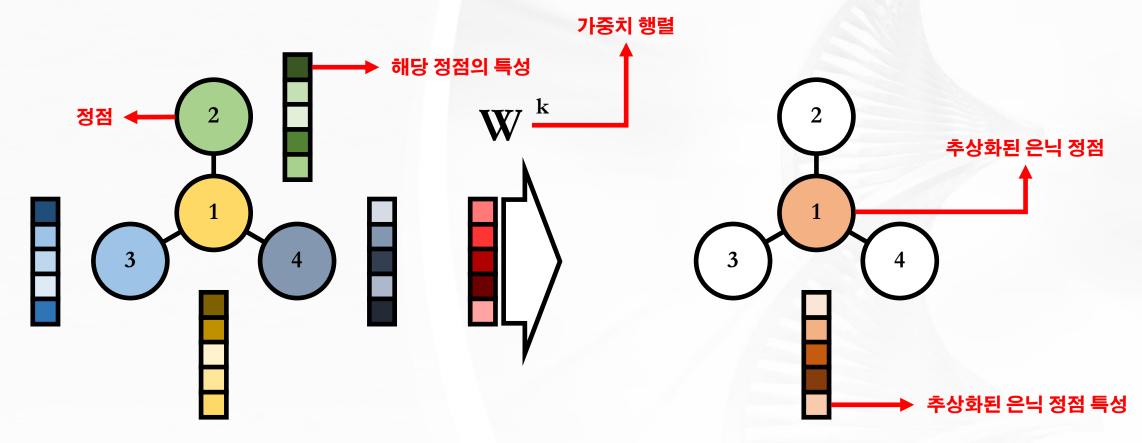
- > 각 그래프 합성곱 층은 그래프 합성곱 연산으로 정의
- ▶ 그래프 합성곱 층을 반복적으로 적용
- > 그래프에 포함된 정점에서 추상화된 은닉 특성을 추출

 $\mathbf{H}^{(k)} = \sigma(\mathbf{\hat{D}}^{-1/2} \mathbf{\hat{A}} \mathbf{\hat{D}}^{-1/2} \mathbf{H}^{(k-1)} \mathbf{W}^{(k)})$   $\mathbf{\hat{L}}$  번째 그래프 합성곱 층 출력 이전 그래프 합성곱 층 출력

## 한국인공지능신약개발지원센터 Kites Al Center for Drug Discovery and Development

#### Graph Convolutional Networks

• 정점 특성 학습 과정

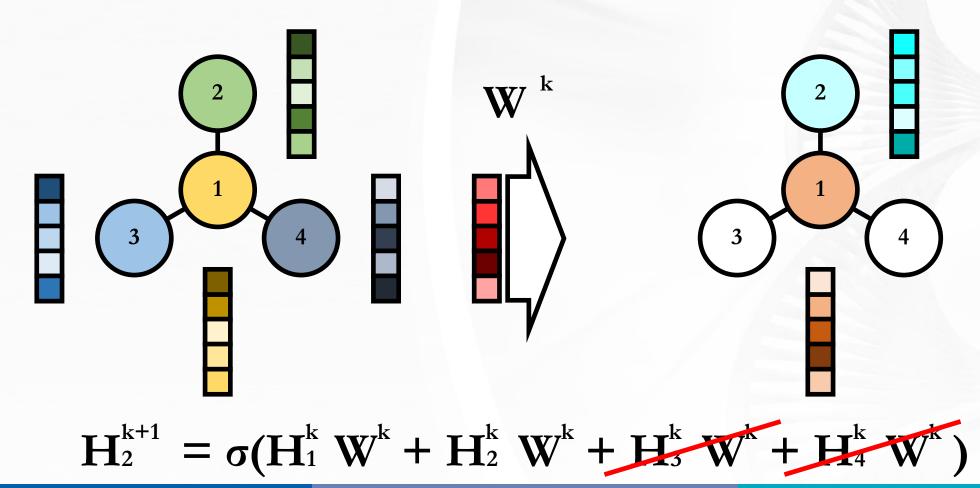


 $\mathbf{H}_{1}^{k+1} = \sigma(\mathbf{H}_{1}^{k} \mathbf{W}^{k} + \mathbf{H}_{2}^{k} \mathbf{W}^{k} + \mathbf{H}_{3}^{k} \mathbf{W}^{k} + \mathbf{H}_{4}^{k} \mathbf{W}^{k})$ 

## KAICD 한국인공지능신약개발지원센터 Kates Al Center for Drug Discovery and Development

### Graph Convolutional Networks

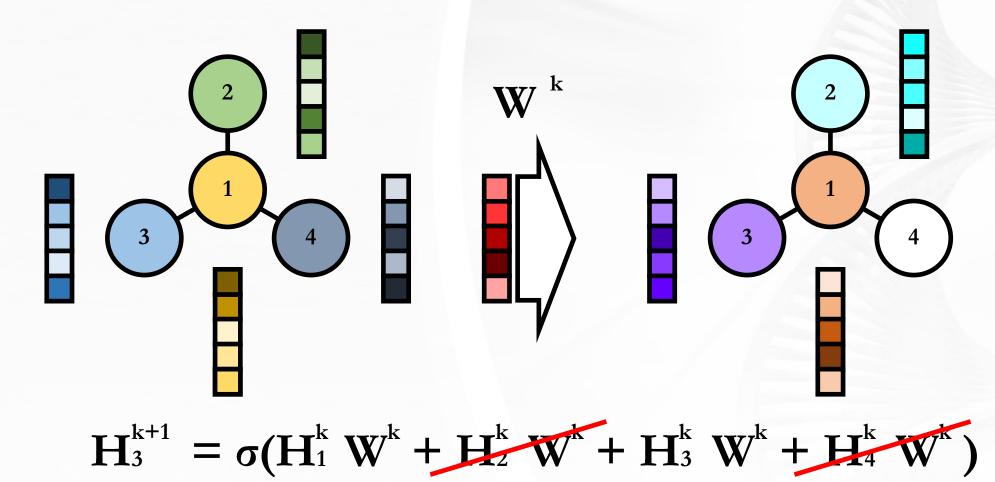
• 정점 특성 학습 과정



## ICAICD 한국인공지능신약개발지원센터 Kana Al Center for Drug Discovery and Development

#### Graph Convolutional Networks

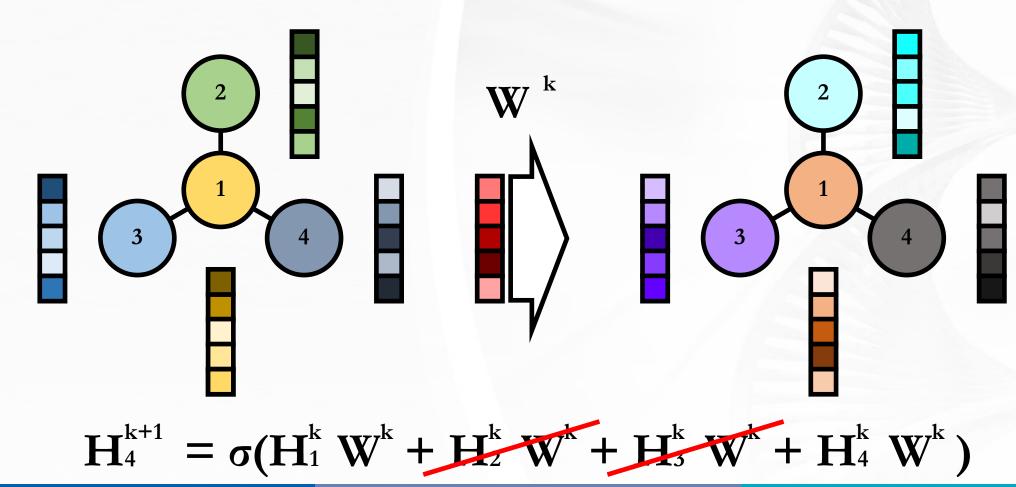
• 정점 특성 학습 과정



## ICAICD 한국인공지능신약개발지원센터 Kons Al Center for Drug Discovery and Development

### Graph Convolutional Networks

• 정점 특성 학습 과정



#### 한국인공지능신약개발지원센터 Kises Al Center for Drug Discovery and Development

#### Graph Convolutional Networks

- GCN 구현
  - > 직접 구현 시, 인접 행렬 및 특성 행렬 추출부터 batch 생성까지 많은 어려움이 따름
  - ▶ PyTorch 기준, Deep Graph Library(DGL)와 PyTorch Geometric를 통해 간편하게 구현 가능
  - ➤ 본 강의(논문)에서는 PyTorch Geometric으로 GCN-based Model을 구현

```
import torch
import torch.nn as nn
import torch.nn.functional as F
from torch_geometric.nn import GCNConv, global_max_pool as gmp
class GCNNet(torch.nn.Module):
    def __init__(self, n_output=1, n_filters=32, embed_dim=128,num_features_xd=78, num_features_xt=25, output_dim=128, dropout=0.2):
        super(GCNNet, self).__init__()
       self.n_output = n_output
        self.conv1 = GCNConv(num_features_xd, num_features_xd)
        self.conv2 = GCNConv(num_features_xd, num_features_xd*2)
        self.conv3 = GCNConv(num_features_xd*2, num_features_xd * 4)
        self.fc_g1 = torch.nn.Linear(num_features_xd+4, 1024)
        self.fc_g2 = torch.nn.Linear(1024, output_dim)
        self.relu = nn.ReLU()
        self.dropout = nn.Dropout(dropout)
        self.embedding_xt = nn.Embedding(num_features_xt + 1, embed_dim)
        self.conv_xt_1 = nn.Conv1d(in_channels=1000, out_channels=n_filters, kernel_size=8)
        self.fc1_xt = nn.Linear(32*121, output_dim)
        self.fc1 = nn.Linear(2*output_dim, 1024)
        self.fc2 = nn.Linear(1024, 512)
        self.out = nn.Linear(512, self.n_output)
```

코드 구현 예시



# GraphDTA Model

Hyeonsu Lee (KAICD) GraphDTA October 28, 2020

**MODEL PROCESS** 



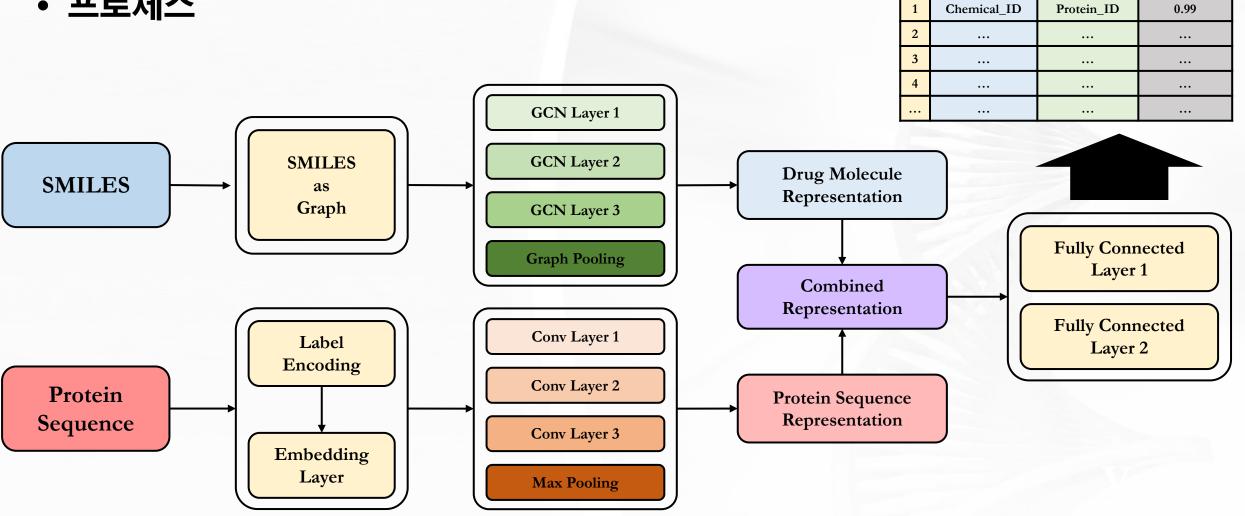
Prediction

Compound

Protein

#### Model

• 프로세스



MODEL EXPERIMENTS



#### Model

#### • 실험 설정

- Davis, Kiba 두 데이터셋에서 DeepDTA, WideDTA와 동일한 train/test split을 사용함
- ▶ 데이터 인스턴스의 80%를 훈련에 사용했으며, 나머지 20%를 테스트에 사용함
- > 기존 논문과 동일 선상에서 성능을 비교하기 위해 MSE와 CI로 성능을 평가함
- > MSE : Mean Square Error, 평균 제곱 오차(낮을수록 좋은 성능)
- ➢ CI : Concordance Index, 일치 지수(높을수록 좋은 성능)
- ▶ 실습 코드에선 다양한 성능평가지표의 시각화를 위해 RMSE, PCC, SRCC 추가
- ▶ 하이퍼 파라미터 설정은 코드 리뷰에서 다룰 예정

MODEL PAPER RESULT





#### • 결과

Davis, KIBA 두 데이터 셋에서 Baseline(DeepDTA, KronRLS, SimBoost)보다 좋은 성능을 도출함

Method	Protein rep. Compound rep.		CI	MSE					
Baseline models									
DeepDTA	Smith-Waterman	Pubchem-Sim	0.790	0.608					
DeepDTA	Smith-Waterman	1D	0.886	0.420					
DeepDTA	1D	Pubchem-Sim	0.835	0.419					
KronRLS	Smith-Waterman	Pubchem-Sim	0.871	0.379					
SimBoost	Smith-Waterman	Pubchem-Sim	0.872	0.282					
DeepDTA	1D	1D	0.878	0.261					
WideDTA	1D + PDM	1D + LMCS	0.886	0.262					
	Proposed model - GraphDTA								
GCN [17]	1D	Graph	0.880	0.254					
GAT_GCN	1D	Graph	0.881	0.245					
GAT [37]	1D	Graph	0.892	0.232					
GIN [40]	1D	Graph	0.893	0.229					

**Davis Dataset Prediction Performance** 

Method	Protein rep.	Compound rep.	CI	MSE					
Baseline models									
DeepDTA	1D	Pubchem-Sim	0.718	0.571					
DeepDTA	Smith-Waterman	Pubchem-Sim	0.710	0.502					
KronRLS	Smith-Waterman	Pubchem-Sim	0.782	0.411					
SimBoost	Smith-Waterman	Pubchem-Sim	0.836	0.222					
DeepDTA	Smith-Waterman	1D	0.854	0.204					
DeepDTA	1D	1D	0.863	0.194					
WideDTA	1D + PDM	1D + LMCS	0.875	0.179					
	Proposed model - GraphDTA								
GAT [37]	1D	Graph	0.866	0.179					
GIN [40]	1D	Graph	0.882	0.147					
GCN [17]	1D	Graph	0.889	0.139					
GAT_GCN	1D	Graph	0.891	0.139					

Kiba Dataset Prediction Performance



## Code Review

Hyeonsu Lee (KAICD) GraphDTA October 28, 2020

#### **Code Review**

- Google Colab
  - https://bit.ly/33aJfZp
  - ▶ 반드시 드라이브에 사본 저장 후 진행해주세요!

