KSBi-BIML 2026

Bioinformatics & Machine Learning(BIML) Workshop for Life Scientists

생명정보학 & 머신러닝 워크샵 (온라인)

Drug target prediction and drug repositioning with graph learning

김선 _ 서울대학교 / 이상선 _ 인하대학교

KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

⚠

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2026 워크샵을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 행위자 본인에게 있음**을 알립니다.

KSBi-BIML 2026

Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists

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

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

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

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

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

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

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

여러분의 많은 관심과 참여를 기대합니다!

2026년 2월

한국생명정보학회장 류 성 호

강의개요

Drug Target Prediction and Drug Repositioning with Graph Learning

약물-표적 관계 예측은 신약 개발 초기 단계에 필수적인 기술이며, 기존의 약물을 재활용하는 약물 재창출 분야에도 밀접한 관련이 있는 기술이다. 그렇다면, 약물의 표적은 어떻게 예측할 수 있을까? 이를 바탕으로 약물 재창출은 어떻게 할 수 있을까? In silico 기반의 약물-표적 관계 예측은 약물과 약물, 약물과 질병, 질병과 유전자 등 여러 가지 상호작용을 고려해야 하기에 많은 어려움이 따른다.

본 강의에서는 약물, 질병, 유전자 간 상호작용을 그래프로 학습하여 약물-표적 예측 및 약물 재창출을 설명한다. 먼저 Random walk, Network propagation, Graph neural network 등 기본적인 그래프 분석 기법들을 배우고, 이를 약물-표적 상관관계 분석/예측 및 약물 재창출 분야에서 효율적이고 효과적으로 활용한 최신 사례를 소개한다.

강의는 다음의 내용을 포함한다.

- 그래프 마이닝 알고리즘
- Graph neural network 기반의 딥러닝 기술
- 약물-표적 관계 예측(Drug-Target Interaction) 사례 및 기술
- 약물 재창출(Drug repositioning) 사례 및 기술

* 교육생준비물: X (이론강의)

* 강의 난이도: 중급

* 강의: 김선 교수 (서울대학교 컴퓨터공학부) / 이상선 컴퓨터공학 박사

Curriculum Vitae

Speaker Name: Sun Kim, Ph.D.

▶ **Personal Info**

Name	Sun Kim
Title	Professor
Affiliation	Seoul National University (SNU)

▶ **Contact Information**

Address	Department of Computer Science and Engineering, 301-421, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, 08826
Email	sunkim.bioinfo@snu.ac.kr

Research Interest

Machine Learning, Deep Learning, Multi-omics, Bioinformatics, AI-drug discovery

Educational Experience

1985	B.S., Computer Science, Seoul National University
1987	M.S., Computer Science, KAIST
1997	Ph.D., Computer Science, University of Iowa

Professional Experience

1998-2001	Senior Computer Scientist, DuPont Central Research
2001-2011	Assistant/Associate Professor, School of Informatics and Computing, Indiana University
2009-2011	Chair, School of Informatics and Computing, Indiana University
2011-2021	Director, Bioinformatics Institute, Seoul National University
2011-	Professor, Department of Computer Science and Engineering & Interdisciplinary Program in Bioinformatics, Seoul National University
2022-	Research Director, MOGAM Institute for Biomedical Research

Selected Publications (5 maximum)

1. Lee, D., Yang, J., & Kim, S. (2022). Learning the histone codes with large genomic windows and three-dimensional chromatin interactions using transformer. Nature Communications, 13(1), 1-19.
2. Lim, S., Lu, Y., Cho, C. Y., Sung, I., Kim, J., Kim, Y., ... & Kim, S. (2021). A review on compound-protein interaction prediction methods: data, format, representation and model. Computational and Structural Biotechnology Journal, 19, 1541-1556.
3. Rhee, S., Seo, S., & Kim, S. (2018, July). Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (pp. 3527-3534).
4. Seo, S., Oh, M., Park, Y., & Kim, S. (2018). DeepFam: deep learning based alignment-free method for protein family modeling and prediction. Bioinformatics, 34(13), i254-i262.
5. Jo, K., Jung, I., Moon, J. H., & Kim, S. (2016). Influence maximization in time bounded network identifies transcription factors regulating perturbed pathways. Bioinformatics, 32(12), i128-i136.

Curriculum Vitae

Speaker Name: Sangseon Lee, Ph.D.

▶ **Personal Info**

Name	Sangseon Lee
Title	Assistant Proferssor
Affiliation	Inha University

▶ **Contact Information**

Address	100, Inha-Ro, Michuhol-Gu, Incheon, 22212
Email	ss.lee@inha.ac.rk

Research Interest

Graph Learining, Bioinformatics, Explainable Deep Learning

Educational Experience

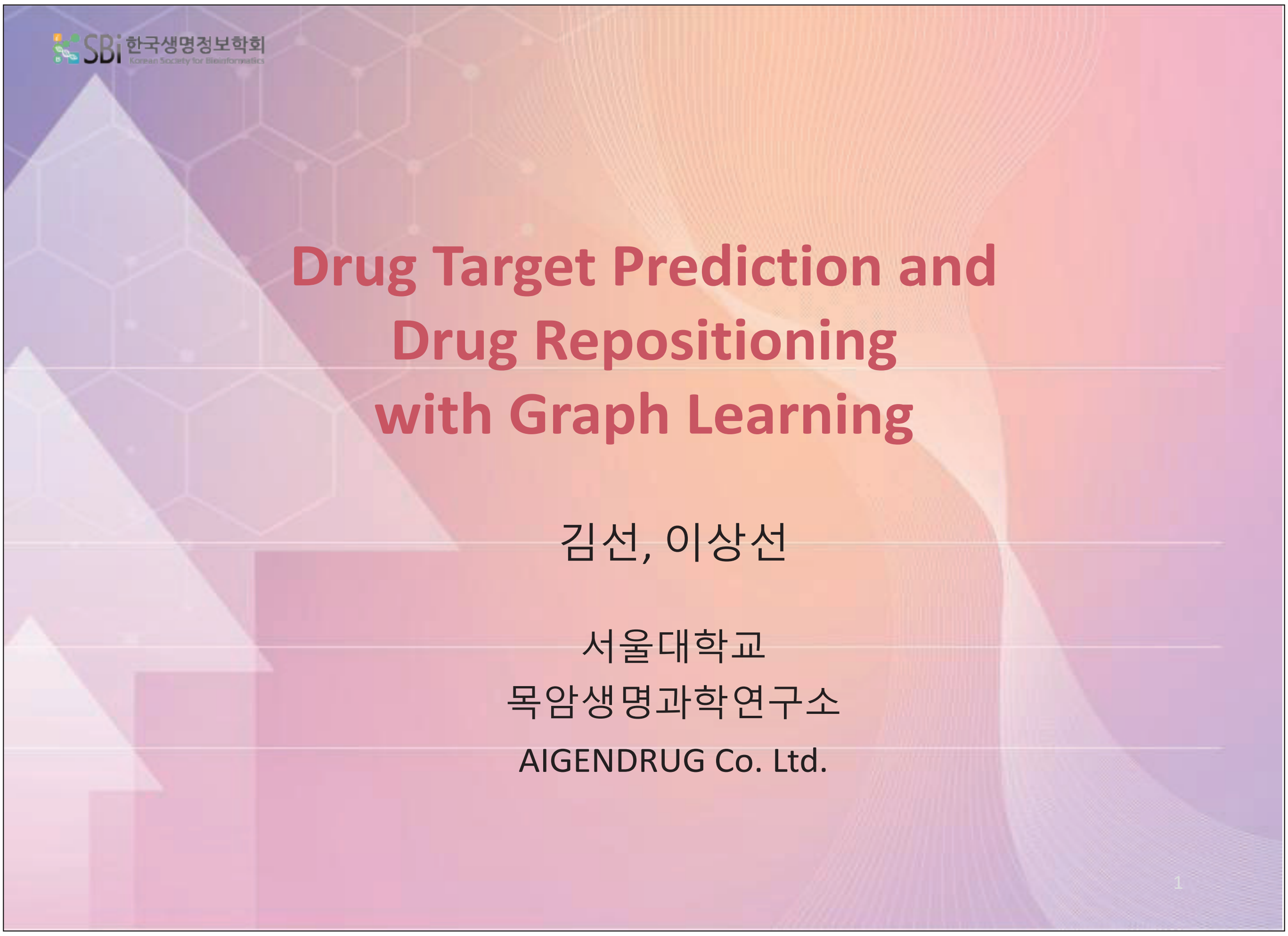
2013	B.S. in Computer Engineering, Seoul National University, Korea
2020	Ph.D. in Computer Engineering, Seoul National University, Korea

Professional Experience

2020-2020	Post-doc research fellow, Bioinfmatics Institute, Seoul National University, Korea
2020-2021	Post-doc research fellow, BK21 FOUR Intelligence Computing, Seoul National University, Korea
2021-2024	Post-doc research fellow, Institute of Computer Technology, Seoul National University, Korea
2014-	Assistant Professor, Inha University

Selected Publications (3 maximum)

1. Sangseon Lee, Sangsoo Lim, Taeheon Lee, Inyoung Sung, Sun Kim, Cancer subtype classification and modeling by pathway attention and propagation, Bioinformatics, 36(12), 2020.
2. Sangseon Lee, Joonhyeong Park, Yinhua Piao, Dohoon Lee, Danyeong Lee, Sun Kim, Multi-layered knowledge graph neural network reveals pathway-level agreement of three breast cancer multi-gene assays, Computational and Structural Biotechnology Journal, 23, 2024.
3. Hyunjun Lim, Sun Kim, Sangseon Lee, CheapNet: Cross-Attention on Hierarchical Representations For Efficient Protein-Ligand Binding Affinity Prediction, ICLR 2025.



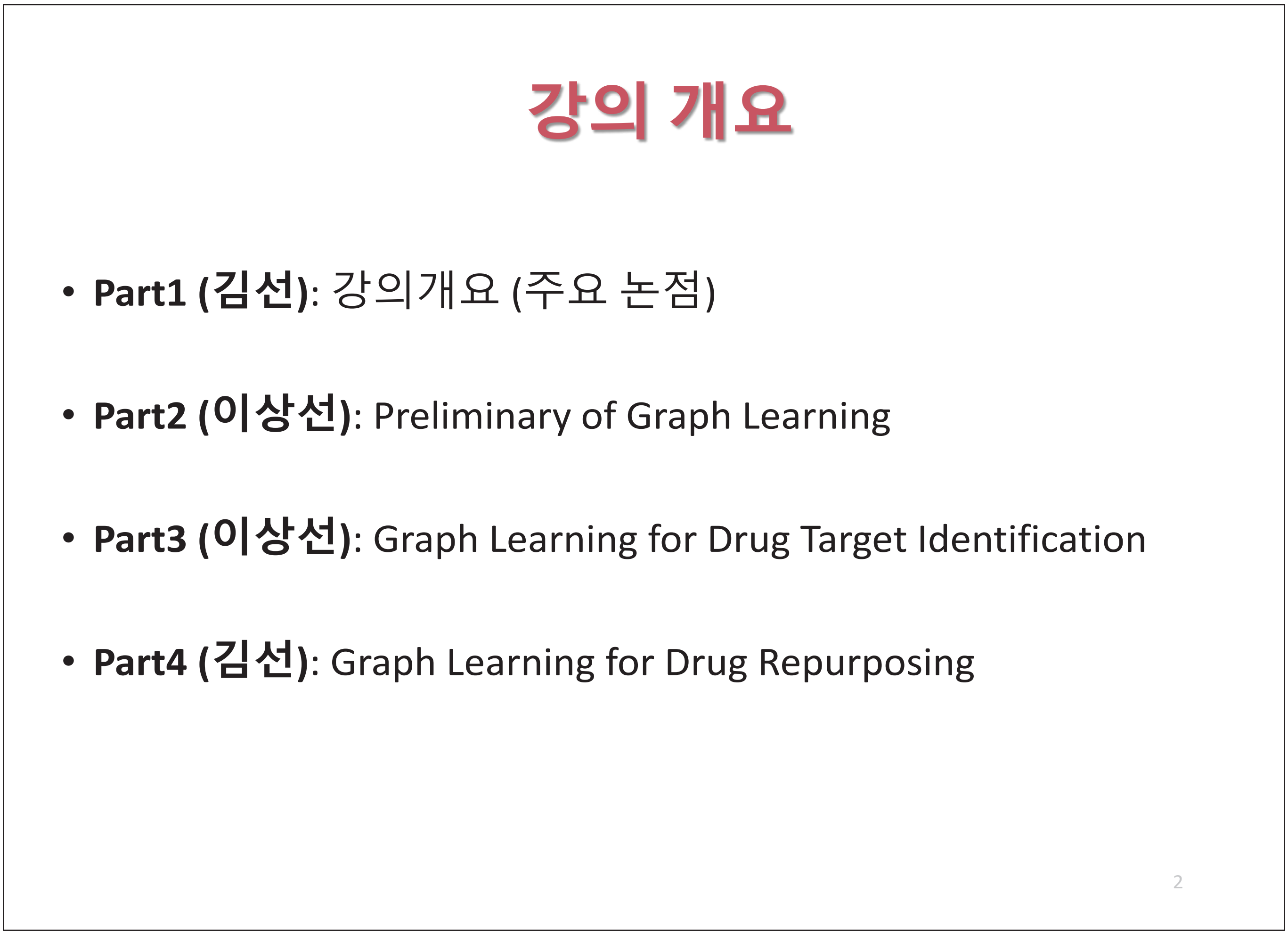
KSBi 한국생명정보학회 Korean Society for Bioinformatics

Drug Target Prediction and Drug Repositioning with Graph Learning

김선, 이상선

서울대학교
목암생명과학연구소
AIGENDRUG Co. Ltd.

1



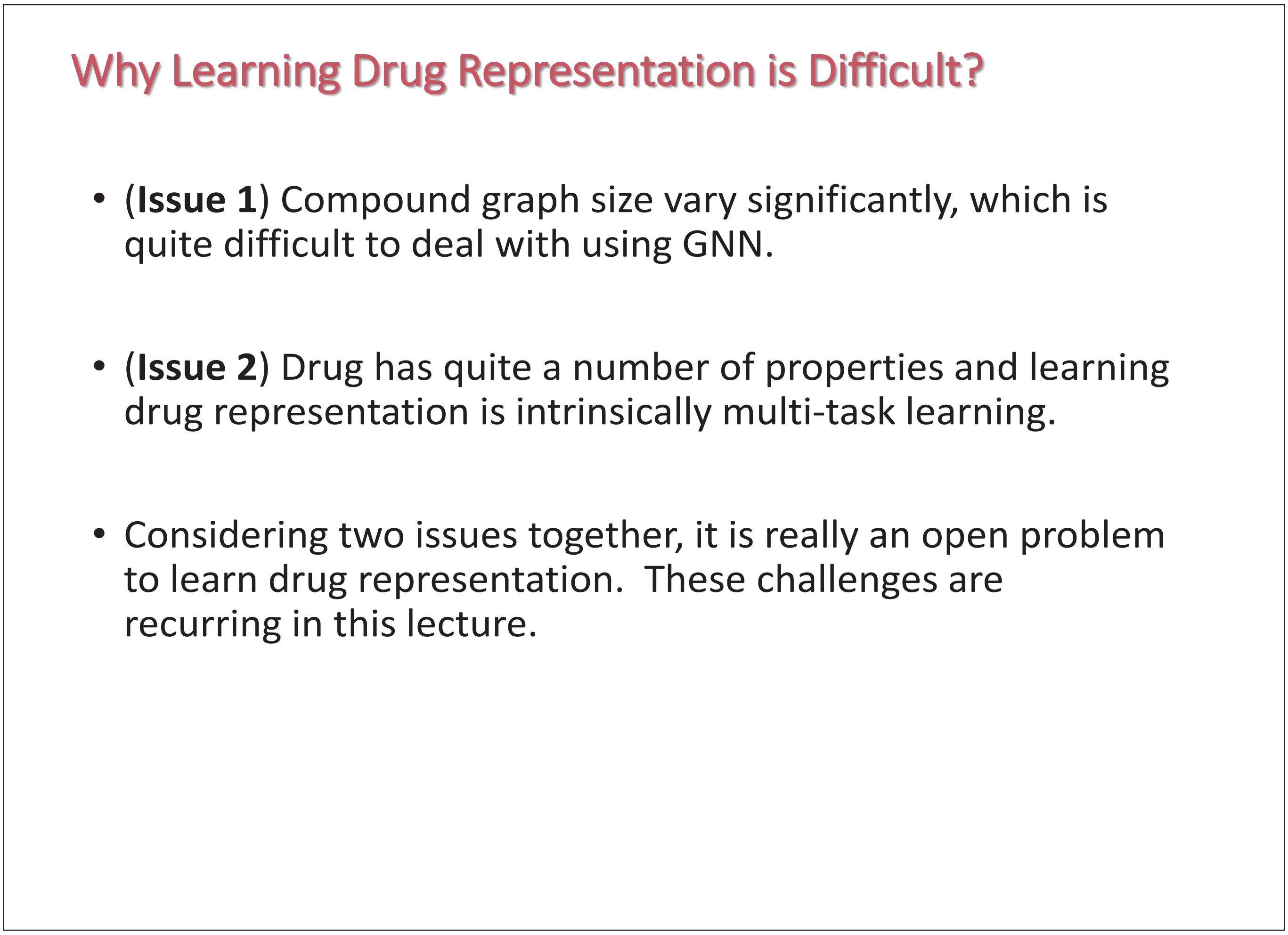
강의 개요

- **Part1 (김선)**: 강의개요 (주요 논점)
- **Part2 (이상선)**: Preliminary of Graph Learning
- **Part3 (이상선)**: Graph Learning for Drug Target Identification
- **Part4 (김선)**: Graph Learning for Drug Repurposing

2

PART 1
강연 개요

3



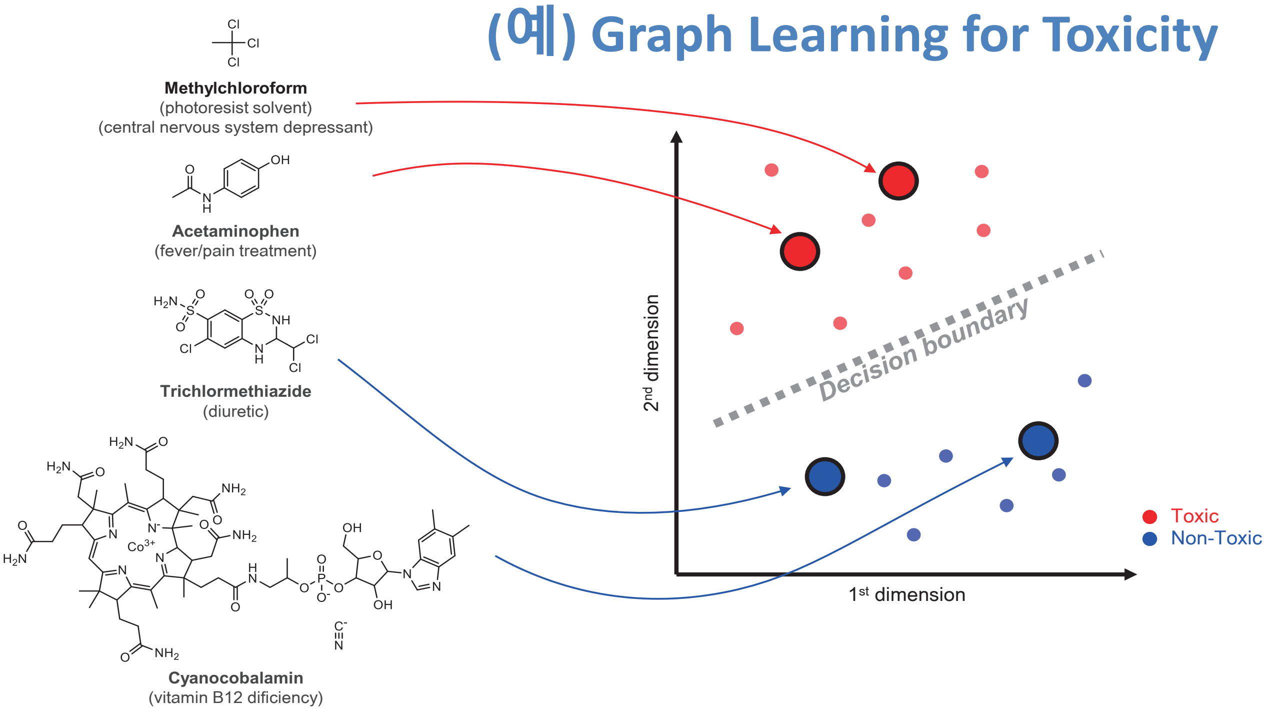
Why Learning Drug Representation is Difficult?

- (**Issue 1**) Compound graph size vary significantly, which is quite difficult to deal with using GNN.

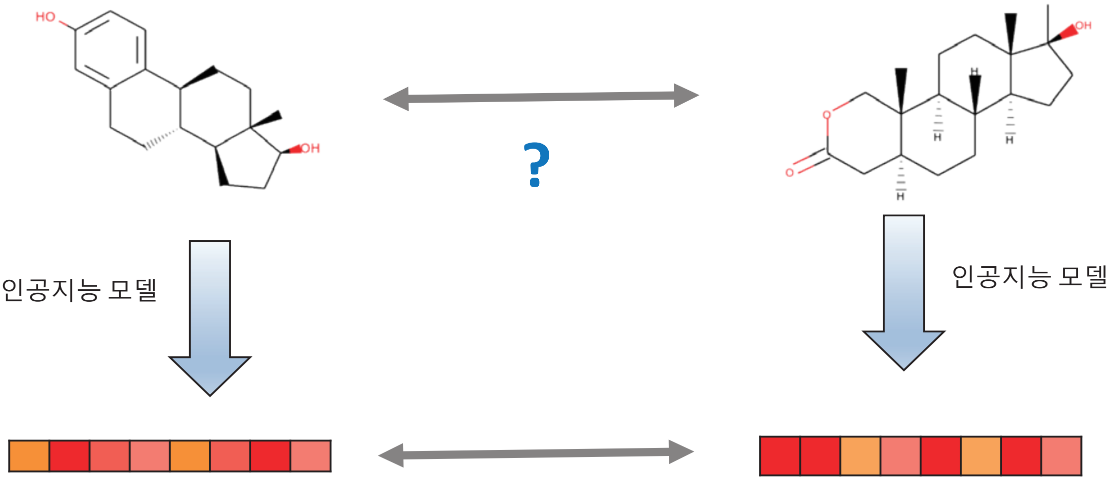
- (**Issue 2**) Drug has quite a number of properties and learning drug representation is intrinsically multi-task learning.

- Considering two issues together, it is really an open problem to learn drug representation. These challenges are recurring in this lecture.

(예) Graph Learning for Toxicity



Why Learning Drug Representation is Useful?



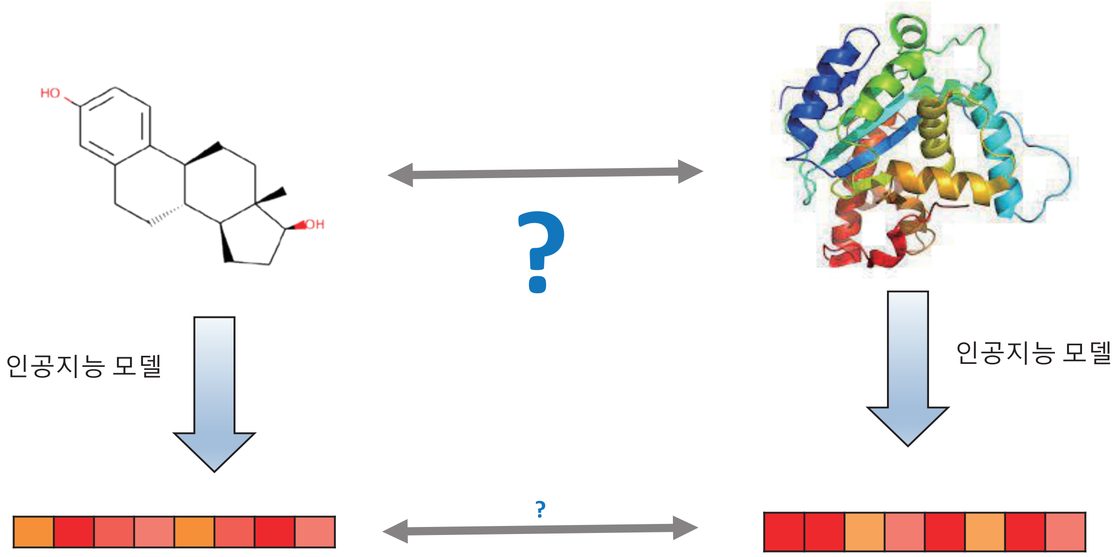
similarity: 0.583

Learning Drug-Target Interaction

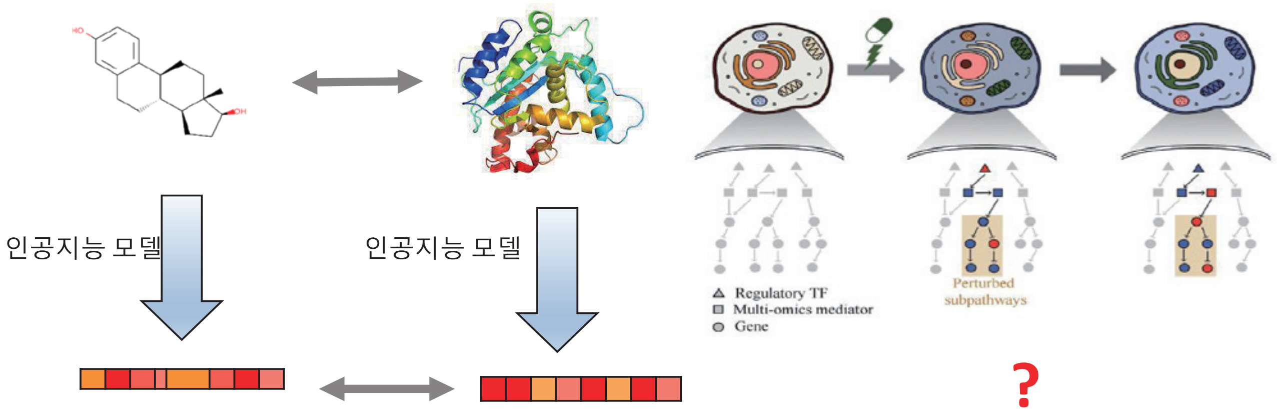
- Given that learning drug representation is difficult, it becomes even more difficult to learn drug-target interaction (DTI) because
  - Drug representation needed to be learned.
  - Representation of target proteins needs to be learned.

- Well, another very complicating factor.
  - DTI should consider what happens after a drug targets a protein (gene) because genes function as a group in a very complex interaction.

Summary: Drug-Target Interaction



True DTI: Compound-Protein-Cell



인공지능 모델

인공지능 모델

?

Drug Re-positioning is Learning Representation of Heterogenous Networks.

- Drug repositionign is to discover unknown association between drug and disease.
- Association between drug and disease is to discover distant relationship.
- Thus, we need help!
- Forutnately, we can use gene networks for this.
- Weel, this becomes to learn representation of three heterogenous networks: drug – gene – disease.

PART 2
Preliminary of Graph Learning

Contents

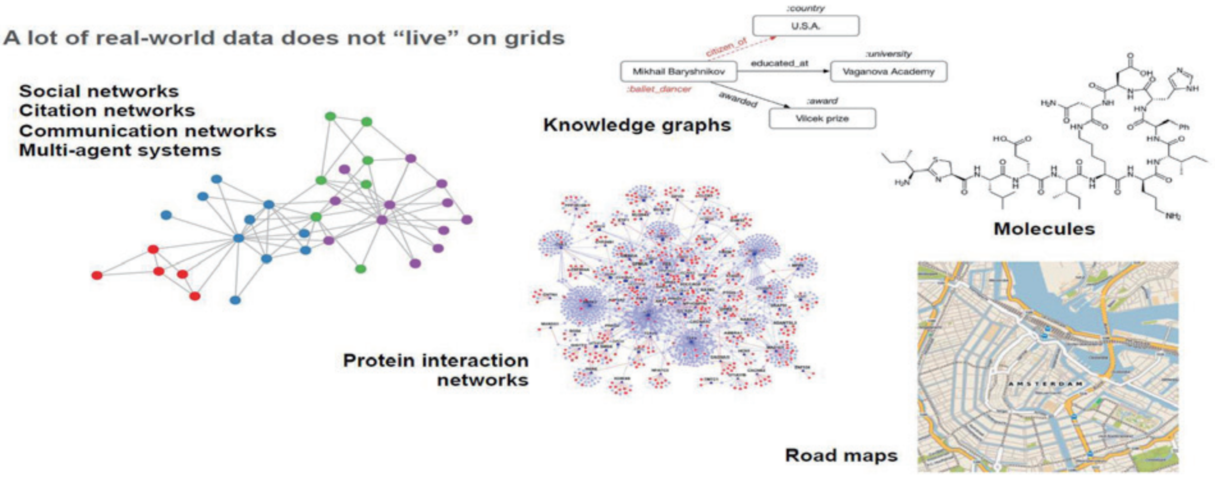
- **What is Graphs?**
  - Example of Graphs in Bioinformatics
- **Preliminary**
  - Random Walk-Based Node Embedding
  - Network Propagation
  - Network Centralities / Clustering
  - VAE / Collective VAE
  - Matrix Factorization
  - Graph Neural Network

What is Graph?

- General concept of graph
- Example of graphs in Bioinformatics

Graph

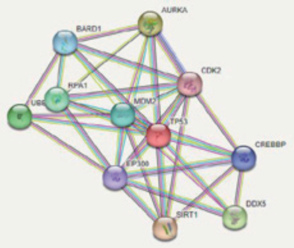
- [Mathematics] A structure made of vertices and edges, G=(V, E)
- [Abstract Data Type] An abstract data type representing relations or connections



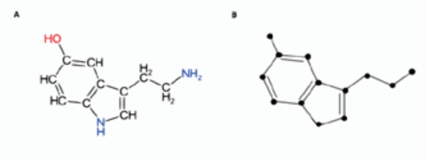
*slide from Thomas Kipf, **University of Amsterdam**

Example of Graphs in Bioinformatics - related to DTI & DR

- Relationships between genes, drugs, or diseases

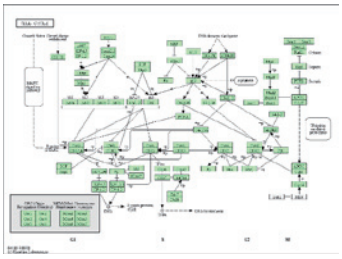


Protein-Protein Interaction (PPI) Network



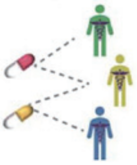
Molecular Graph

Protein-Disease Network



Biological Pathway

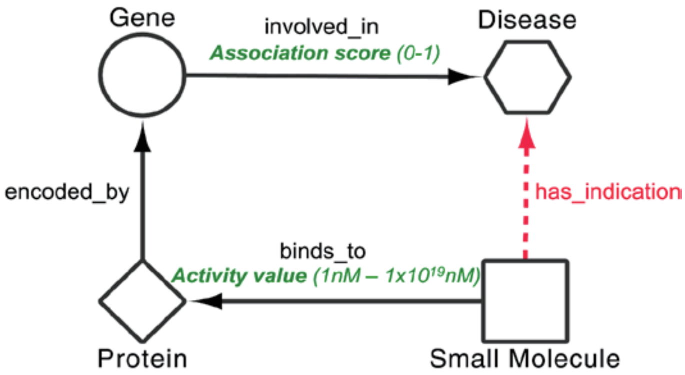
Drug-Drug Network



Drug-Disease Network

Example of Graphs in Bioinformatics - related to DTI & DR

- PPI network & Biological pathway
  - Represents biological mechanisms via gene interactions
  - Can be utilized for learning states of data (ex. patient, cell-line, …)
- Roles in the DTI & DR tasks
  - Identification of patients or cell-lines through multi-omics data
  - Bridge between drugs and disease



Gene

involved_in

Association score (0-1)

Disease

encoded_by

has_indication

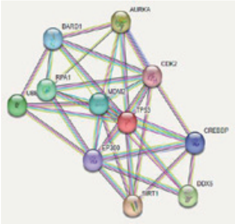
binds_to

Activity value (1nM – 1x10^{19}nM)

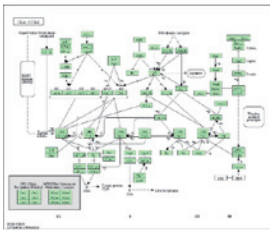
Protein

Small Molecule

(Mullen, Joseph, et al., *PloS One*, 2016)

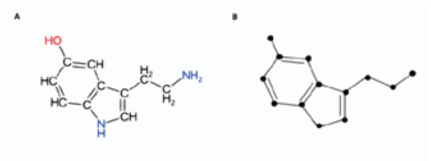


Protein-Protein Interaction (PPI) Network



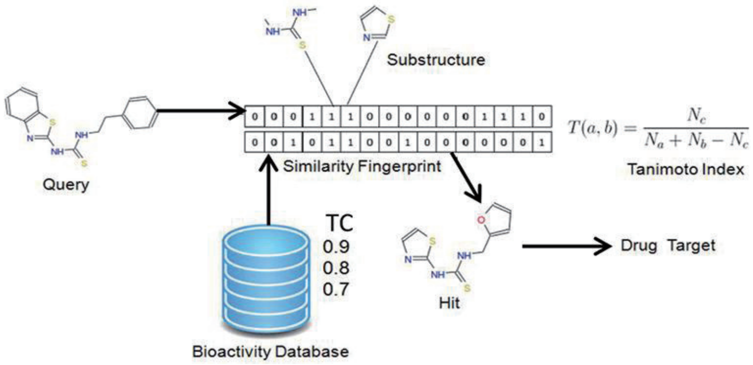
Biological Pathway

Example of Graphs in Bioinformatics - related to DTI & DR



Molecular Graph

- <u>Molecular Graph</u>
  - Represents information of drug or small molecule itself
  - Atom types, Bond types, Atom-Atom distance, Bond-Bond angles, …
- <u>Roles in the DTI & DR tasks</u>
  - Used as inputs for learning drug’s structure, function, properties, ..
  - Used as ingredients for calculating drug-drug similarities



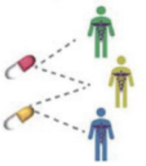
(https://www.intechopen.com/chapters/52373)

Example of Graphs in Bioinformatics - related to DTI & DR

- <u>Drug, Gene, Disease Network</u>
  - Association between drugs, genes, and diseases
- <u>Roles in the DTI & DR tasks</u>
  - Main inputs for learning drug targets and repurposing diseases
  - DTI: which drugs and genes interact?
  - DR: which drugs are used for other diseases?
    - Drug-disease association
    - Discover novel or new targets of approved drugs

Drug-Drug Network

Protein-Disease Network



Drug-Disease Network

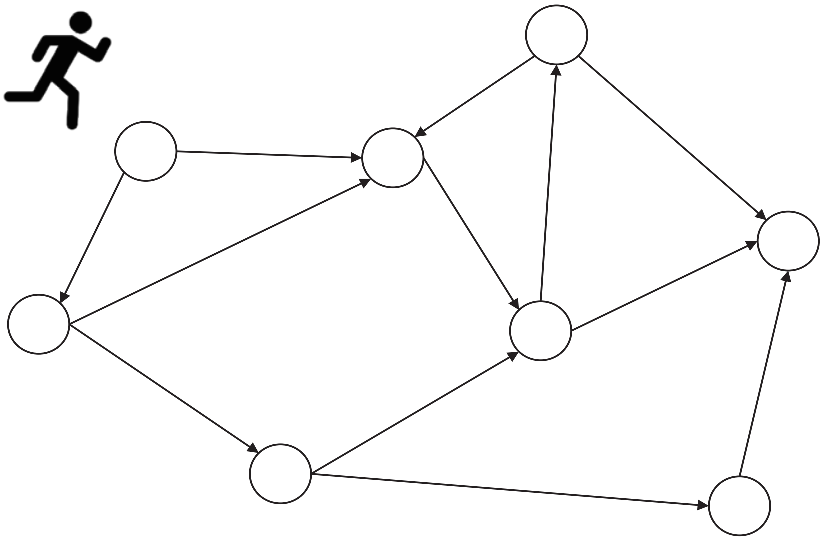
Preliminary for Graph Learning

- Random Walk-based Node Embedding
- Network Propagation
- Network Centralities / Clustering
- VAE / Collective VAE
- Matrix Factorization
- Graph Neural Network

Random Walk-based Node Embedding

Random walk

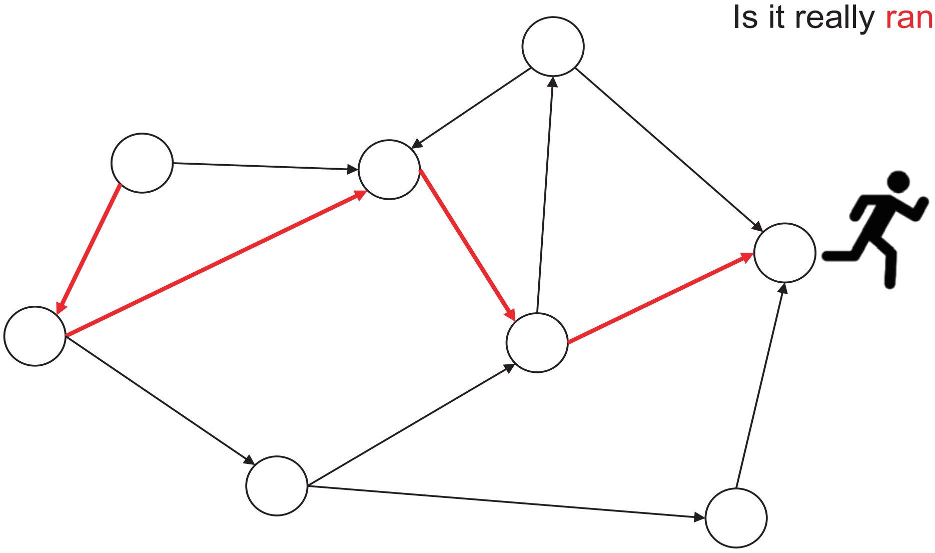
- An agent in the graph moves “randomly” along the graph topology to explore different nodes.



Random walk

- An agent in the graph moves “randomly” along the graph topology to explore different nodes.

Is it really random?

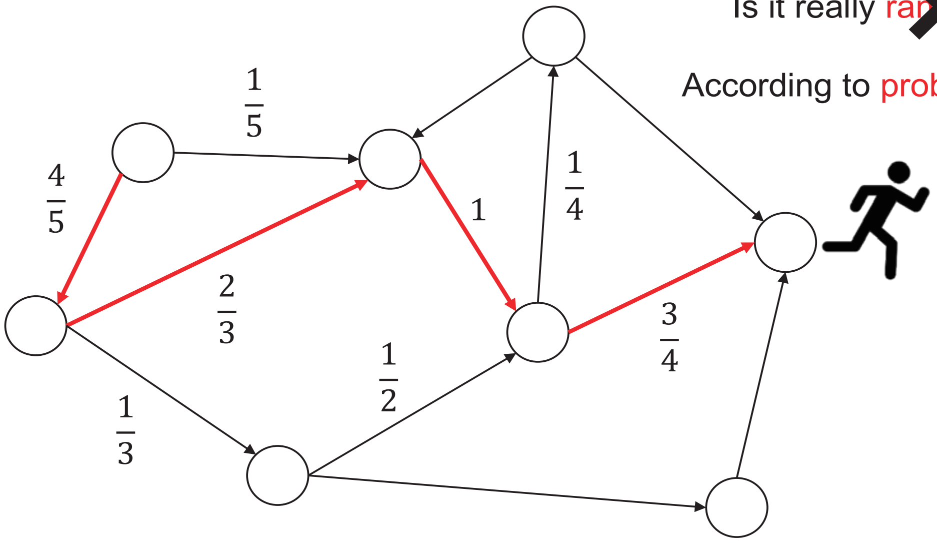


Random walk

- An agent in the graph moves “randomly” along the graph topology to explore different nodes.

Is it really ~~random~~?

According to probabilities!

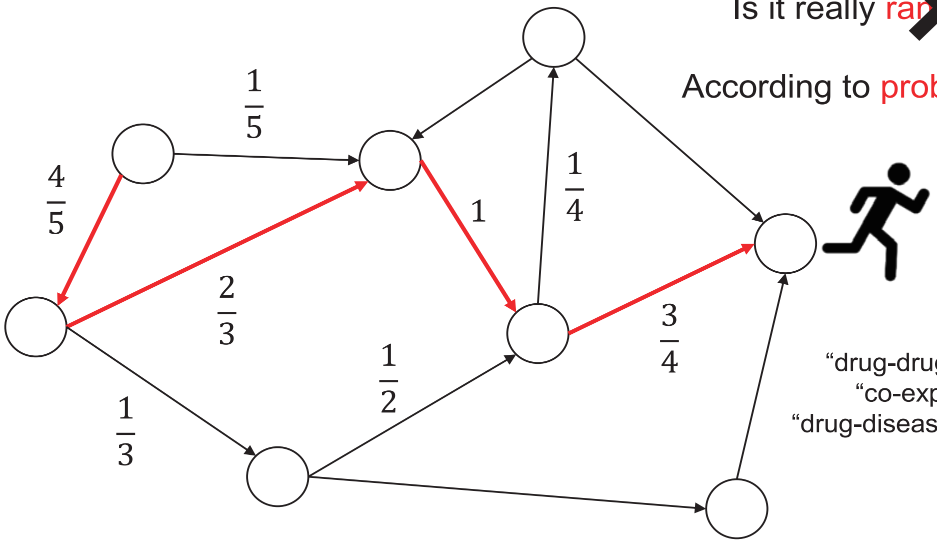


Random walk

- An agent in the graph moves “randomly” along the graph topology to explore different nodes.

Is it really ~~random~~?

According to probabilities!

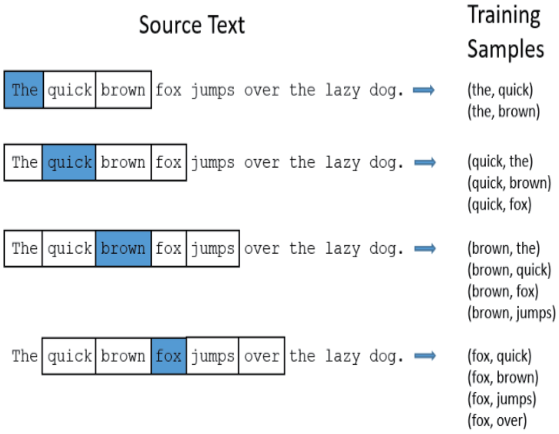


“drug-drug similarity”
“co-expression”
“drug-disease association”
⋮

Random walk-based Node Embedding

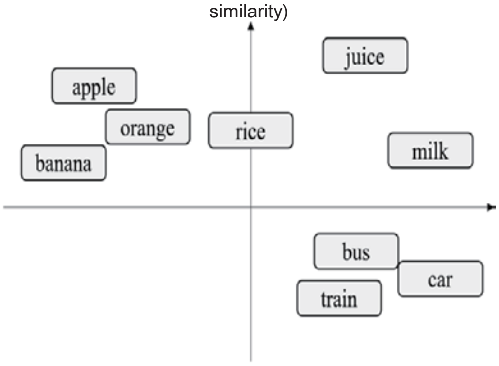
- Inspired by word embedding in natural language processing
  - word2vec: learn word representations by co-occurrence in the sentences
  - Predict context words using a center word

Example of word2vec input



(http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/)

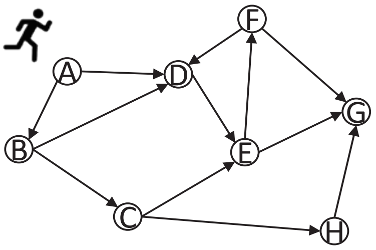
Example of word embedding (by similarity)



(Li, Bofang, et al., *Data Science and Engineering*, 2019)

Random walk-based Node Embedding

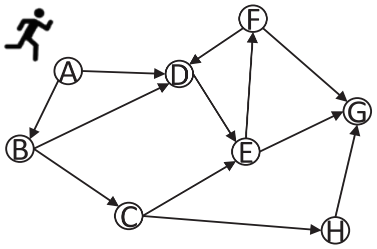
- How to get the sentences from a graph?
  - Random walk!



A → B → D → E → G
A → D → E → F → D → E → G
A → B → C → H → G
A → D → E → G
⋮

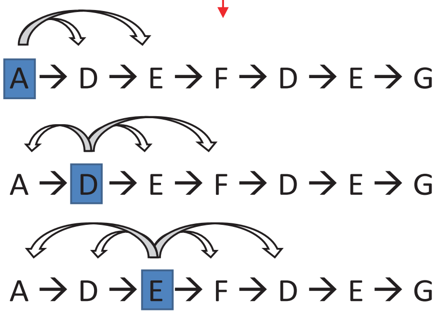
Random walk-based Node Embedding

- How to get the sentences from a graph?
  - Random walk!



A → B → D → E → G
A → D → E → F → D → E → G
A → B → C → H → G
A → D → E → G

⋮


A → D → E → F → D → E → G

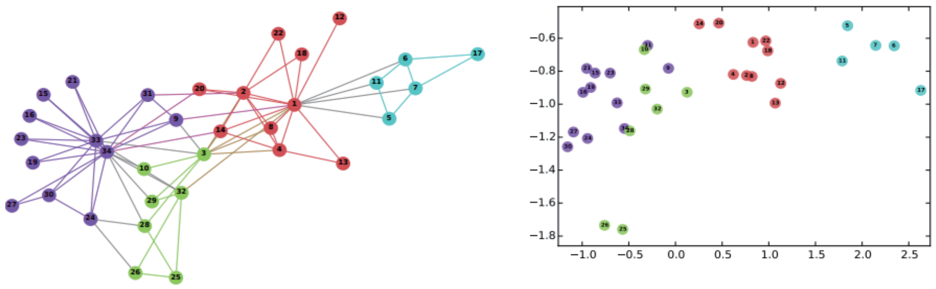
A → D → E → F → D → E → G

A → D → E → F → D → E → G

Make sentences by considering node co-occurrences

Random walk-based Node Embedding

- DeepWalk
  - Generate node embeddings using random walks



(a) Input: Karate Graph (b) Output: Representation

(Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena., *ACM SIGKDD*, 2014.)

Random walk-based Node Embedding

- Exploration of graph
  - DFS: Depth-First Search
  - BFS: Breadth-First Search

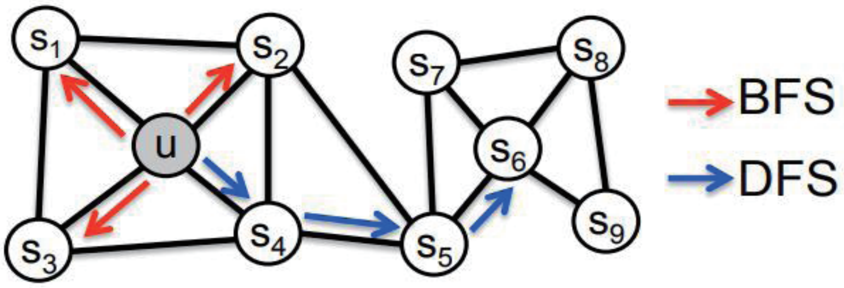
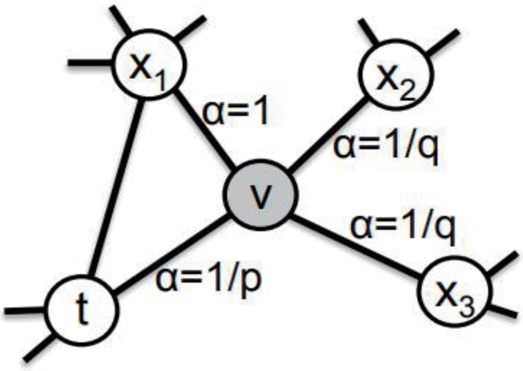


Figure 1: BFS and DFS search strategies from node u ($k = 3$).

(Grover, Aditya, and Jure Leskovec., ACM SIGKDD, 2016.)

Random walk-based Node Embedding

- Exploration of graph with different probabilities
  - The walk just transitioned from t to v and is now evaluating its next step out of node v.
  - Edge labels indicate search biases α.

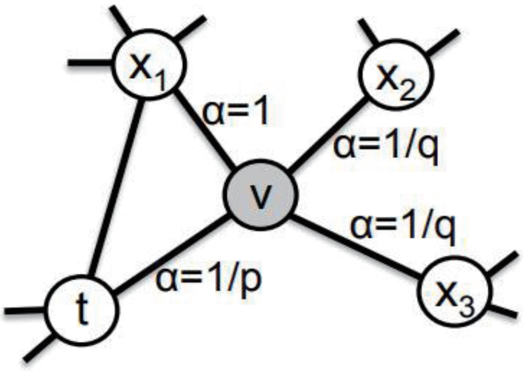


(Grover, Aditya, and Jure Leskovec., ACM SIGKDD, 2016.)

Random walk-based Node Embedding

- Exploration of graph with different probabilities
  - The walk just transitioned from t to v and is now evaluating its next step out of node v.
  - Edge labels indicate search biases α.

$p = q = 1$
(special case; DeepWalk)



31

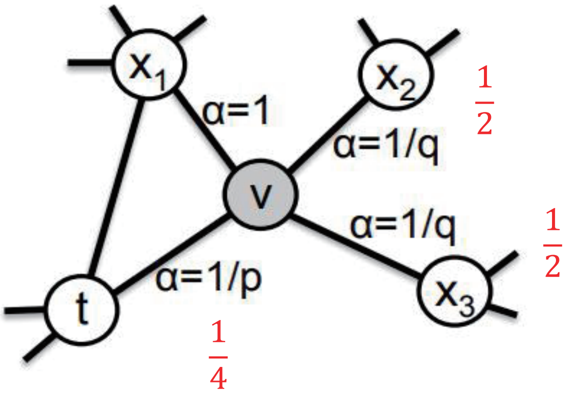
(Grover, Aditya, and Jure Leskovec., ACM SIGKDD, 2016.)

Random walk-based Node Embedding

- Exploration of graph with different probabilities
  - The walk just transitioned from t to v and is now evaluating its next step out of node v.
  - Edge labels indicate search biases α.

$p = q = 1$
(special case; DeepWalk)

$p > q$
(More explore)



32

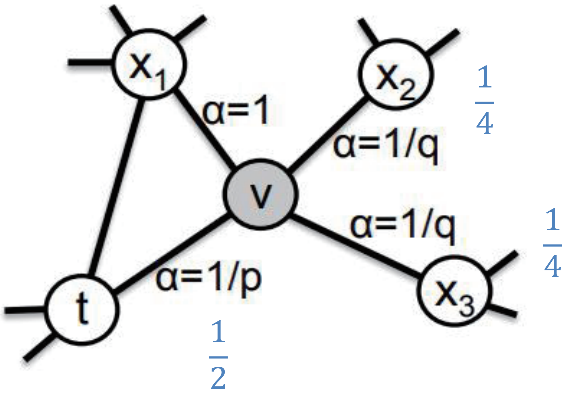
(Grover, Aditya, and Jure Leskovec., ACM SIGKDD, 2016.)

Random walk-based Node Embedding

- Exploration of graph with different probabilities
  - The walk just transitioned from t to v and is now evaluating its next step out of node v.
  - Edge labels indicate search biases α.

$p = q = 1$
(special case; DeepWalk)

$p > q$
(More explore)



$p < q$
(walk local)

(Grover, Aditya, and Jure Leskovec., ACM SIGKDD, 2016.)

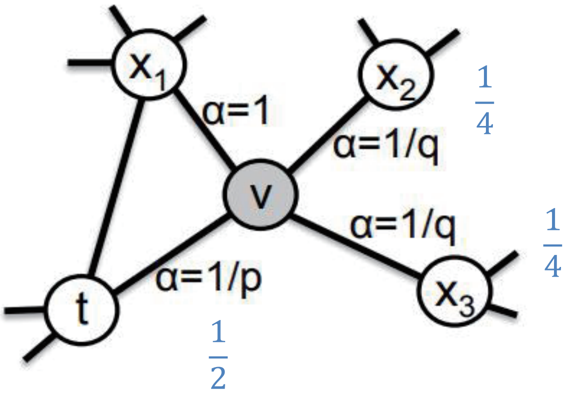
Random walk-based Node Embedding

- Exploration of graph with different probabilities
  - The walk just transitioned from t to v and is now evaluating its next step out of node v.
  - Edge labels indicate search biases α.

“node2vec”

$p = q = 1$
(special case; DeepWalk)

$p > q$
(More explore)

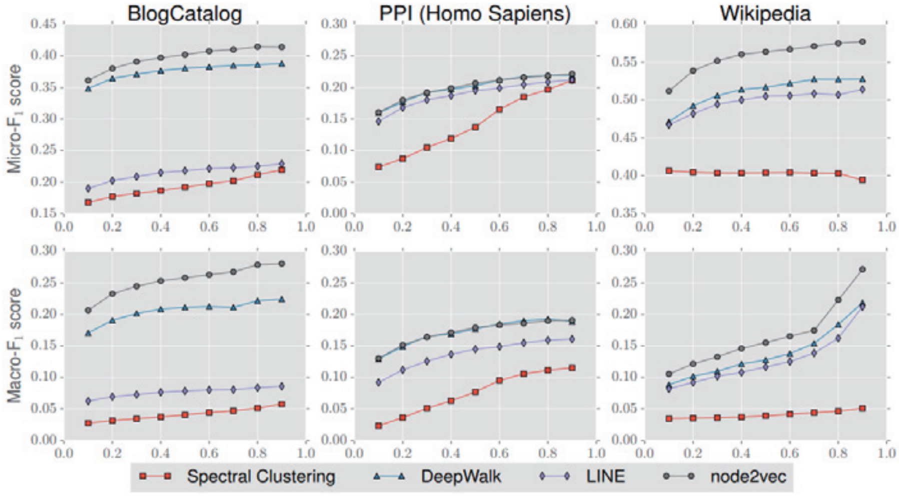


$p < q$
(walk local)

(Grover, Aditya, and Jure Leskovec., ACM SIGKDD, 2016.)

Random walk-based Node Embedding

- Exploration of graph with different probabilities
  - The walk just transitioned from t to v and is now evaluating its next step out of node v.
  - Edge labels indicate search biases α.

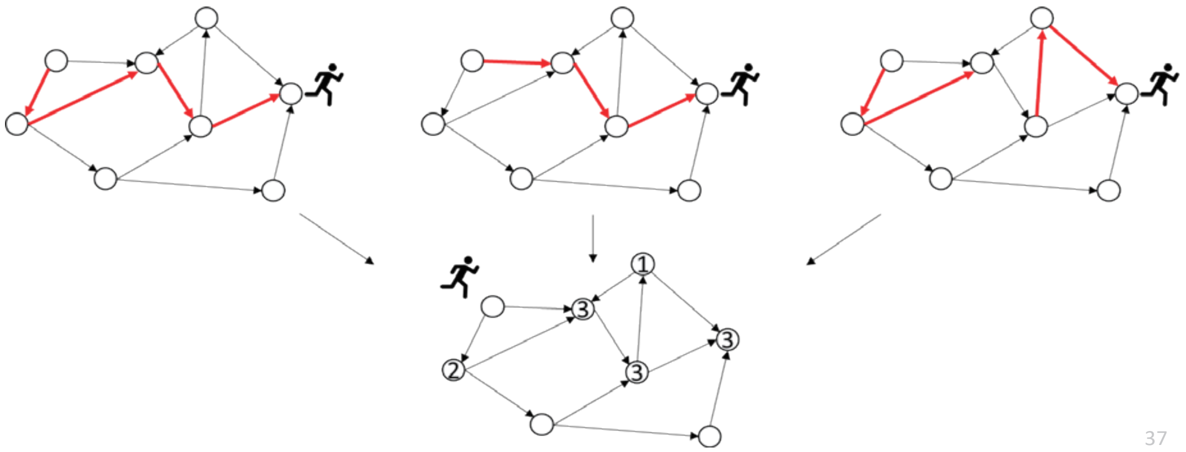


(Grover, Aditya, and Jure Leskovec., ACM SIGKDD, 2016.)

Network Propagation

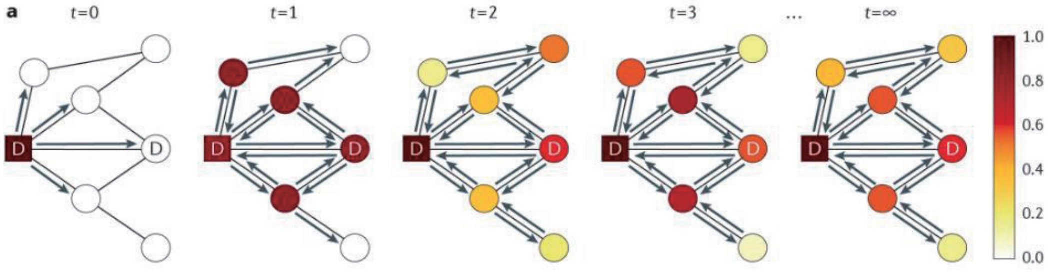
Network Propagation

- Random walks are generated by transition probabilities.
  - The number of random walks is the number of samples used by the model (DeepWalk, node2vec).
- So what if we create an infinite number of random walks of a certain length from one starting point and then measure the frequency of nodes observed in the walks?



Network Propagation

- Propagate information of known nodes (= seeds) via network topology
- Until certain steps, the amount of information (or flow) will be converged



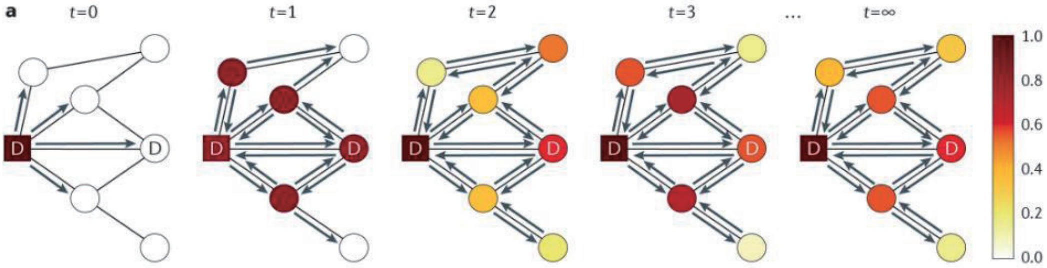
(Cowen, Lenore, et al., *Nature Reviews Genetics*, 2017)

Network Propagation

- Propagate information of known nodes (= seeds) via network topology
- Until certain steps, the amount of information (or flow) will be converged

- Random walk with re-start (RWR)

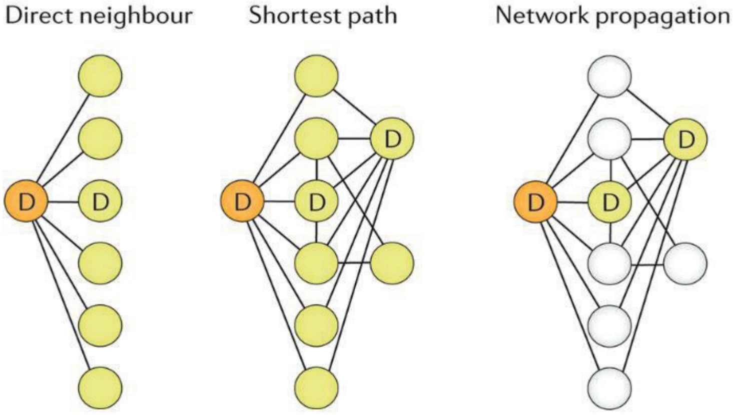
$$p(t+1) = \alpha \times p(0) + (1-\alpha) \times W \times p(t)$$



(Cowen, Lenore, et al., *Nature Reviews Genetics*, 2017)

Advantages of Network Propagation

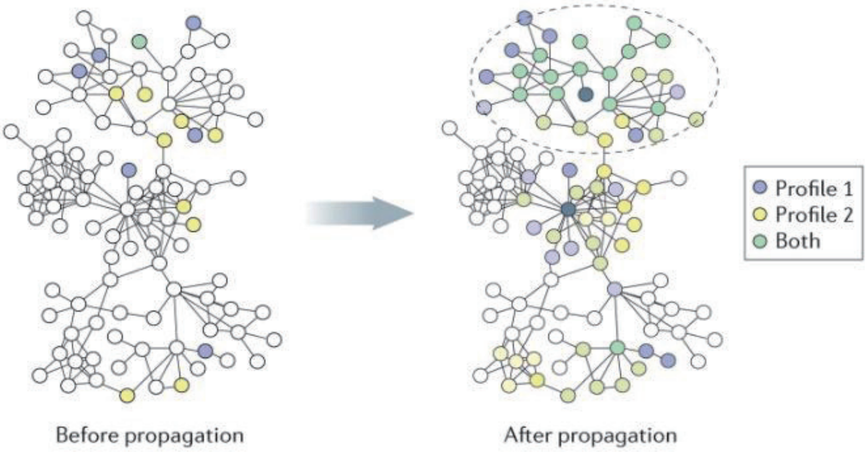
- Looking at more distant neighbours that are up to two steps away (yellow; middle panel) again introduces many false positives.
- Network propagation overcomes these problems by simultaneously considering all paths between genes (yellow; right panel).



(Cowen, Lenore, et al., *Nature Reviews Genetics*, 2017)

Advantages of Network Propagation

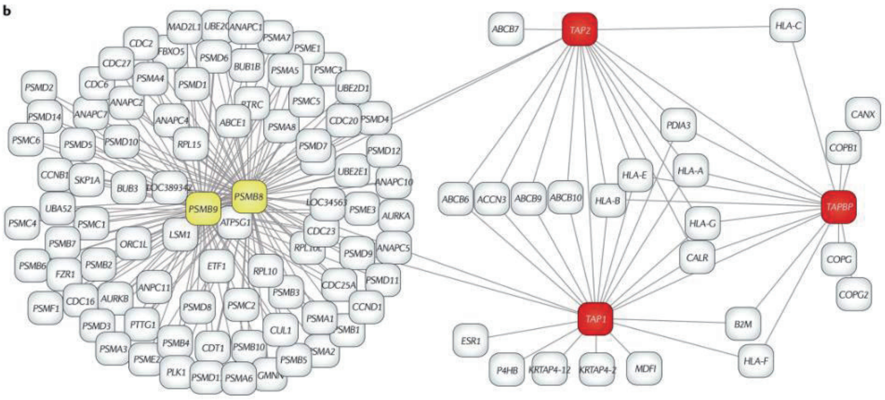
- Network propagation considers and aggregates influence of all seeds via network topology
- It can capture informative clusters of interest



(Cowen, Lenore, et al., *Nature Reviews Genetics*, 2017)

Advantages of Network Propagation

- Propagation of the signal from any of the three known disease genes (red) ranks the other known disease genes very highly, owing to the many paths between them.
- Genes in yellow are ranked highly by alternative network analysis methods (which consider direct neighbours or shortest paths); however, these are false positives.

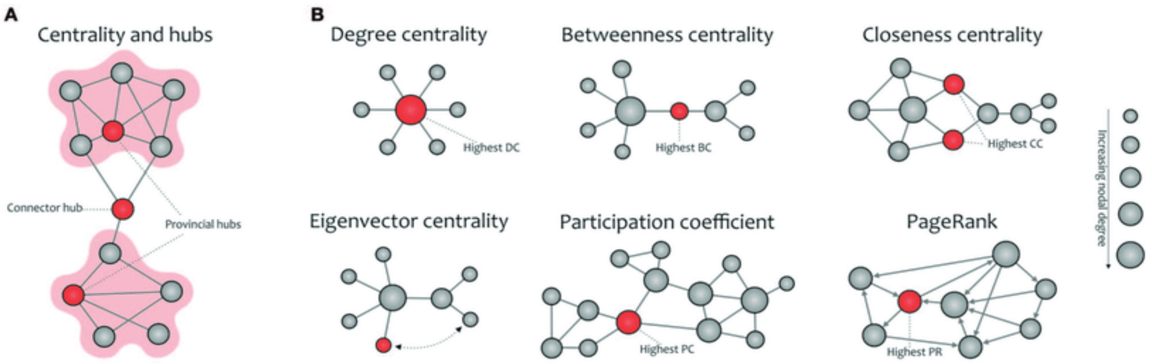


(Cowen, Lenore, et al., *Nature Reviews Genetics*, 2017)

Network Centralities / Clustering

Network Centralities

- **Centrality** assign numbers or rankings to nodes within a graph corresponding to their network position.
- "What characterizes an important vertex?” → How to define “important”?



Farahani, Farzad V., Waldemar Karwowski, and Nichole R. Lighthall. "Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review." frontiers in Neuroscience (2019)

Network Centralities

1. Degree Centrality
- defined as the number of links incident upon a node

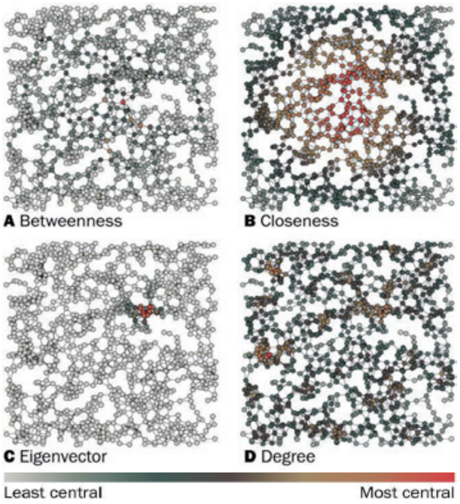
2. Closeness Centrality
- is the average length of the shortest path between the node and all other nodes in the graph.

3. Betweenness Centrality
- the number of times a node acts as a bridge along the shortest path between two other nodes.

4. Eigenvector Centrality
- Measure of the influence of a node in a network.
- Measured by calculating the eigenvector of adjacency matrix
- Google's PageRank is based on the normalized eigenvector centrality

Farahani, Farzad V., Waldemar Karwowski, and Nichole R. Lighthall. "Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review." frontiers in Neuroscience (2019)

Network Centralities



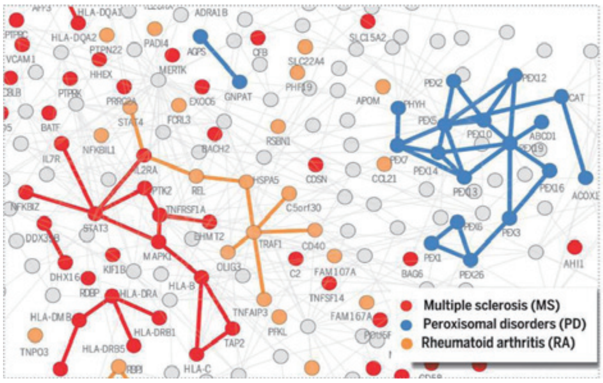
Different scores are assigned for different centralities
- A centrality which is optimal for one application is often sub-optimal for a different application.
- The optimal measure depends on the network structure of the most important vertices
- Complex networks (e.g. disease networks) have heterogeneous topology; ranking its nodes with centrality possesses limitations [2].

[1] Wikipedia: Network Centrality (https://en.wikipedia.org/wiki/Centrality#/media/File:Wp-01.png, retrieved 2022-11-15)
[2] Lawyer, Glenn. "Understanding the influence of all nodes in a network." Scientific reports 5.1 (2015): 1-9.

Network Clustering

Disease are interplay of multiple molecular processes

- Disease-associated proteins interact with each other and cluster to form **disease modules**
- Network clustering methods are utilized for detecting communities and modules



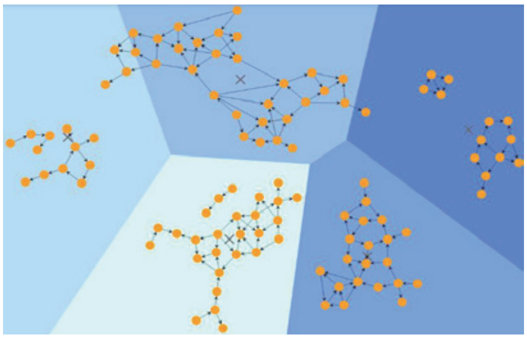
Menche, Jörg, et al. "Uncovering disease-disease relationships through the incomplete interactome." Science 347.6224 (2015): 1257601.

Network Clustering

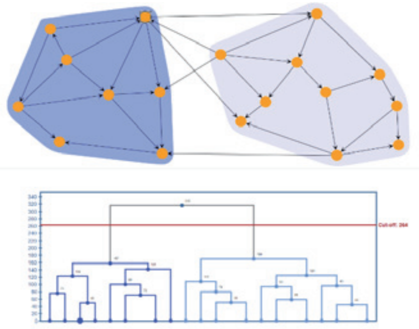
Widely-used Network clustering algorithms

1. **k-means clustering**
   - partitions the graph into k clusters based on the location of the nodes such that their distance from the cluster's mean (centroid) is minimum
   - The distance is defined using various metrics as Euclidean distance, Euclidean-squared distance, Manhattan distance, or Chebyshev distance.

k-means clustering



Hierarchical clustering



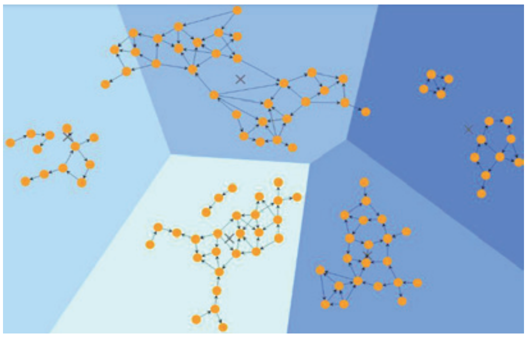
yworks: Clustering Graphs and Networks, https://www.yworks.com/pages/clustering-graphs-and-networks)

Network Clustering

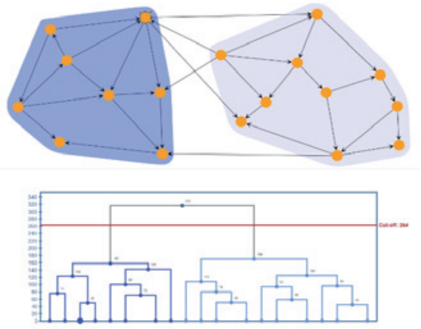
Widely-used Network clustering algorithms
2. Hierarchical clustering

- Partitions the graph into a hierarchy of clusters.
- The result is a dendrogram which can be cut based on a given cut-off value.

k-means clustering



Hierarchical clustering

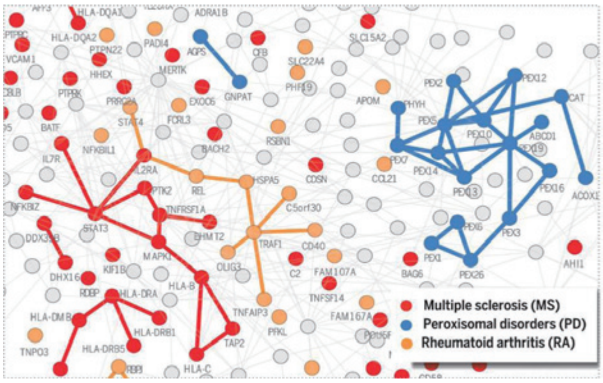


yworks: Clustering Graphs and Networks, https://www.yworks.com/pages/clustering-graphs-and-networks)

Network Clustering

*** Limitations of disease module-based approaches**

- available interactome and disease-related gene information are incomplete, and do have sufficient coverage to map out disease modules

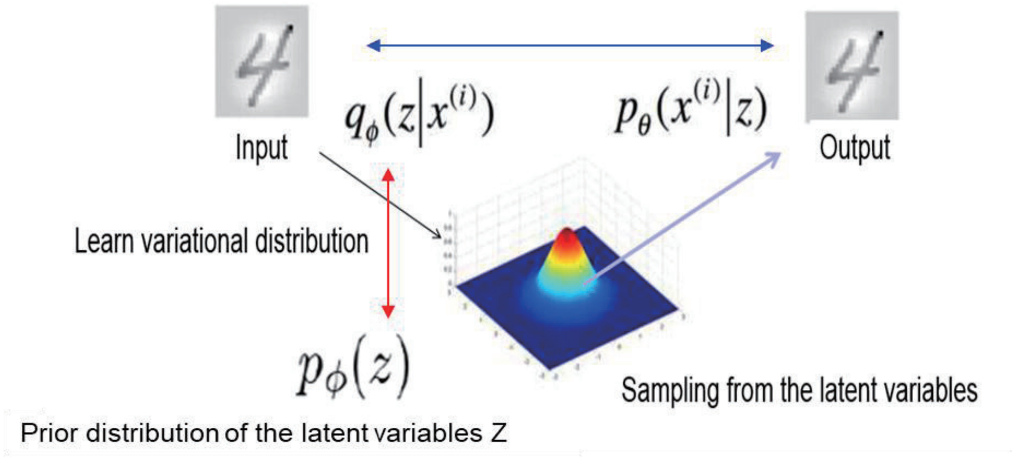


Menche, Jörg, et al. "Uncovering disease-disease relationships through the incomplete interactome." Science 347.6224 (2015): 1257601.

VAE / Collective VAE

Variational Auto-Encoder (VAE)

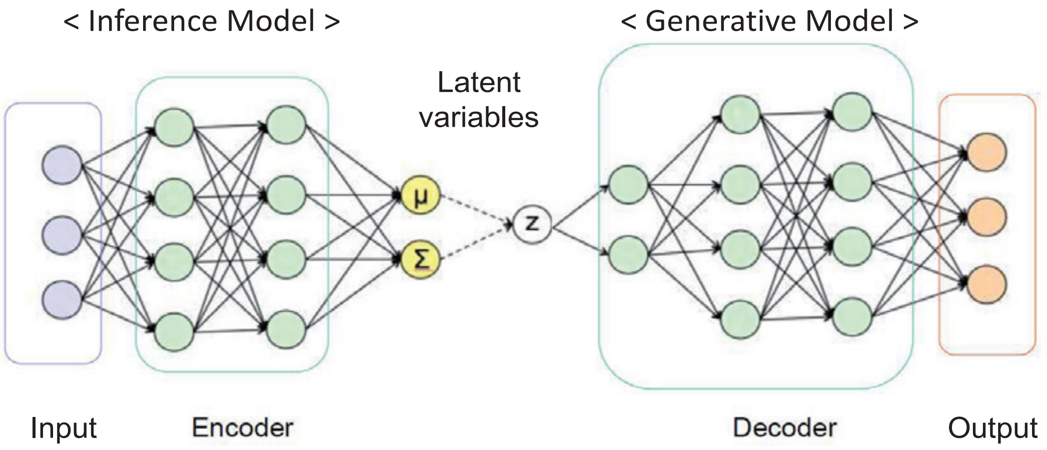
- A generative model that reconstructs input data from latent variables


Input
$q_\phi(z|x^{(i)})$
$p_\theta(x^{(i)}|z)$
Output
Learn variational distribution
$p_\phi(z)$
Sampling from the latent variables
Prior distribution of the latent variables Z

Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

Variational Auto-Encoder (VAE)

- A generative model that reconstructs input data from latent variables



< Inference Model >

< Generative Model >

Latent variables

Input

Encoder

Decoder

Output

Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

Collective VAE

- Proposed model for item recommendation
- Simultaneously recover user ratings (main task) and side information
- Can be utilized for DTI & DR
  - Main task: drug-disease association
  - Side Information: drug information

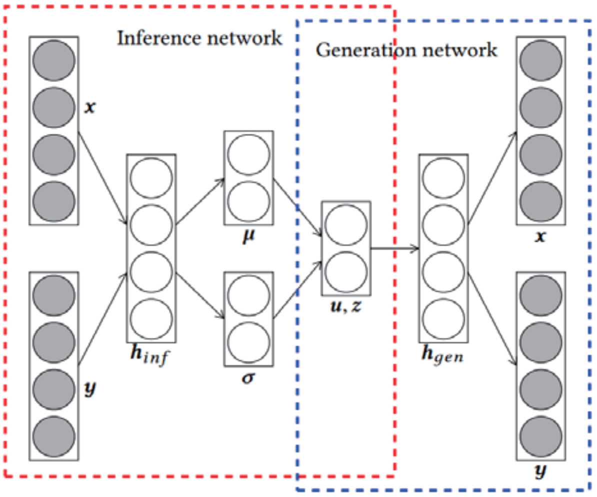


Figure 1: Collective Variational Autoencoder

(Chen, Yifan, and Maarten de Rijke, *Proceedings of the 3rd workshop on deep learning for recommender systems*, 2018.)

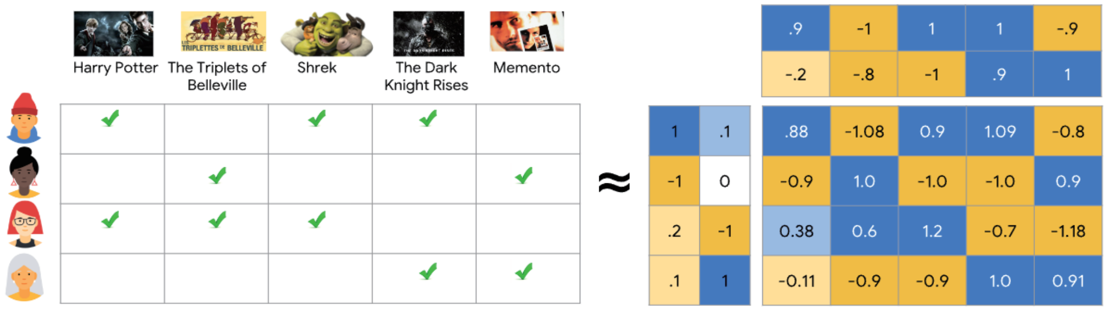
Matrix Factorization

Matrix Factorization

- A class of collaborative filtering algorithms used in recommender systems.
- Decompose a matrix into tow lower dimensional matrices
  - Learn low dimensional latent embeddings of row/column

$$A \approx UV^T$$

$$A \in R^{m \times n} \qquad U \in R^{m \times d} \qquad V \in R^{n \times d} \qquad m, n \gg d$$



	Harry Potter	The Triplets of Belleville	Shrek	The Dark Knight Rises	Memento
	✓		✓	✓	
		✓			✓
	✓	✓	✓		
				✓	✓

≈

1	.1
-1	0
.2	-1
.1	1

.9	-1	1	1	-.9
-.2	-.8	-1	.9	1

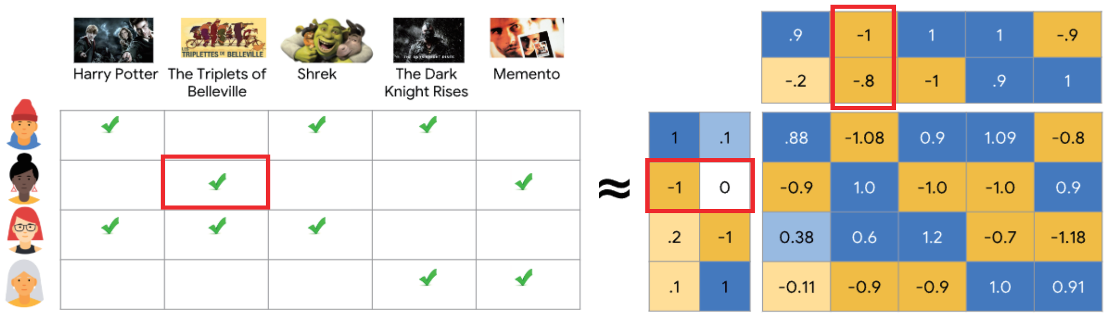
.88	-1.08	0.9	1.09	-0.8
-0.9	1.0	-1.0	-1.0	0.9
0.38	0.6	1.2	-0.7	-1.18
-0.11	-0.9	-0.9	1.0	0.91

https://developers.google.com/machine-learning/recommendation/collaborative/matrix

Matrix Factorization

- Minimize difference of A and UV^T

$$\min_{U \in \mathbb{R}^{m \times d},\ V \in \mathbb{R}^{n \times d}} \sum_{(i,j) \in \text{obs}} (A_{ij} - \langle U_i, V_j \rangle)^2$$



Harry Potter	The Triplets of Belleville	Shrek	The Dark Knight Rises	Memento
✓		✓	✓	
	✓			✓
✓	✓	✓		
			✓	✓

$$\approx \begin{bmatrix} 1 & .1 \\ -1 & 0 \\ .2 & -1 \\ .1 & 1 \end{bmatrix} \begin{bmatrix} .9 & -1 & 1 & 1 & -.9 \\ -.2 & -.8 & -1 & .9 & 1 \end{bmatrix} = \begin{bmatrix} .88 & -1.08 & 0.9 & 1.09 & -0.8 \\ -0.9 & 1.0 & -1.0 & -1.0 & 0.9 \\ 0.38 & 0.6 & 1.2 & -0.7 & -1.18 \\ -0.11 & -0.9 & -0.9 & 1.0 & 0.91 \end{bmatrix}$$

https://developers.google.com/machine-learning/recommendation/collaborative/matrix

Matrix Factorization

- Minimize difference of A and UV^T
- How to handle unobserved cases?
  - Assume the value as 0.
  - Minimize the loss function with different weights

$$\min_{U \in \mathbb{R}^{m \times d},\ V \in \mathbb{R}^{n \times d}} \sum_{(i,j) \in \text{obs}} (A_{ij} - \langle U_i, V_j \rangle)^2 + w_0 \sum_{(i,j) \notin \text{obs}} (\langle U_i, V_j \rangle)^2$$

Observed Only MF

1		1	1	
	1			1
1	1	1		
			1	1

$\Sigma_{(i,\ j)\ \in\ \text{obs}}\ (A_{ij} - U_i\ .\ V_j)^2$

Weighted MF

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

$\Sigma_{(i,\ j)\ \in\ \text{obs}}\ (A_{ij} - U_i\ .\ V_j)^2 +$

$w_0\ \Sigma_{(i,\ j)\ \notin\ \text{obs}}\ (0 - U_i\ .\ V_j)^2$

https://developers.google.com/machine-learning/recommendation/collaborative/matrix

Matrix Factorization (≈ Matrix Completion)

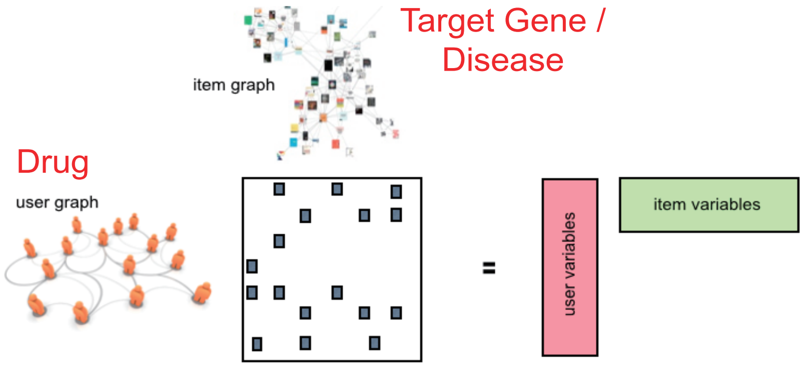
- Standard matrix factorization is transductive.

$$\min_{W,H} \sum_{(i,j)\in\Omega} \left(P_{ij} - \left(WH^T\right)_{ij}\right)^2 + \underline{\frac{\lambda}{2}\left(\|W\|_F{}^2 + \|H\|_F{}^2\right)}$$

To prevent overfitting

Target Gene / Disease

Drug



Example of item recommendation

Beyond Low Rank Matrix Factorization | Center for Big Data Analytics (utexas.edu)

Matrix Factorization (≈ Matrix Completion)

- Standard matrix factorization is transductive.

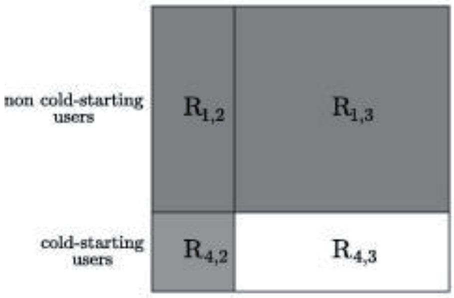
$$\min_{W,H} \sum_{(i,j)\in\Omega} \left(P_{ij} - \left(WH^T\right)_{ij}\right)^2 + \underline{\frac{\lambda}{2}\left(\|W\|_F{}^2 + \|H\|_F{}^2\right)}$$

To prevent overfitting

- All matrix completion approaches suffer from extreme sparsity of the observed matrix and the cold-start problem.

Easy to learn & predict

Hard to learn & predict



(Ocepek, Uroš, Jože Rugelj, and Zoran Bosnić., Expert Systems with Applications, 2015.)

Matrix Factorization (≈ Matrix Completion)

- Standard matrix factorization is transductive.

$$\min_{W,H} \sum_{(i,j)\in\Omega} \left(P_{ij} - \left(WH^T\right)_{ij}\right)^2 + \underline{\frac{\lambda}{2}\left(\|W\|_F{}^2 + \|H\|_F{}^2\right)}$$

To prevent overfitting

- Inductive Matrix Factorization (or Completion)
    - Can be interpreted as a generalization of the transductive multi-label formulation

$$\min_{W,H} \sum_{(i,j)\in\Omega} \ell\left(P_{ij},\ \underline{x_i{}^T W}\underline{H^T y_j}\right) + \frac{\lambda}{2}\left(\|W\|_F{}^2 + \|H\|_F{}^2\right)$$

Matrix Factorization (≈ Matrix Completion)

- Inductive Matrix Factorization (or Completion)
    - Can be interpreted as a generalization of the transductive multi-label formulation

$$\min_{W,H} \sum_{(i,j)\in\Omega} \ell\left(P_{ij},\ x_i{}^T WH^T y_j\right) + \frac{\lambda}{2}\left(\|W\|_F{}^2 + \|H\|_F{}^2\right)$$

- Positive-Unlabeled (PU) Matrix Completion
    - In case of DTI task, we collect positive pairs of drug and target protein.
    - It is difficult to “well-defined negative” data.

$$\min_{W,H} \sum_{(i,j)\in\Omega^+} \left(P_{ij} - x_i WH^T y_j{}^T\right)^2 + \underline{\alpha \sum_{(i,j)\in\Omega^-} \left(P_{ij} - x_i WH^T y_j{}^T\right)^2}$$
$$+ \lambda\left(\|W\|_F{}^2 + \|H\|_F{}^2\right)$$

α: the penalty of the unobserved entries toward zero

(Zeng, Xiangxiang, et al., Chemical Science, 2020)

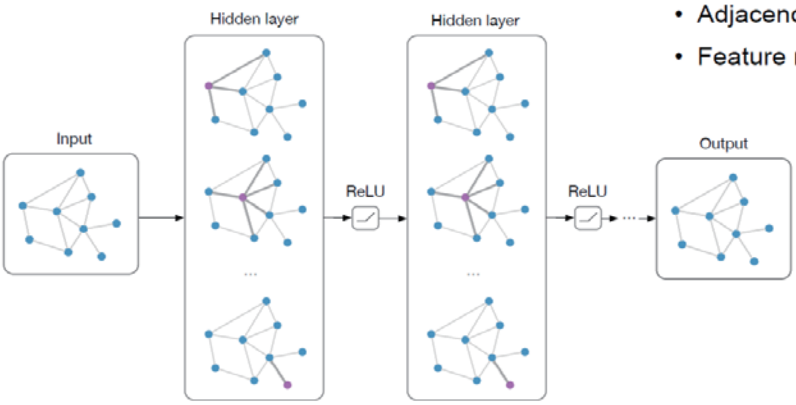
Graph Neural Network

Graph Neural Network

The bigger picture:

Notation: $\mathcal{G} = (\mathbf{A}, \mathbf{X})$

- Adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$
- Feature matrix $\mathbf{X} \in \mathbb{R}^{N \times F}$

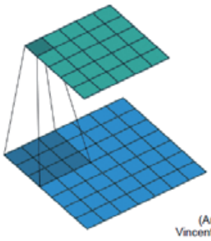


Main idea: Pass messages between pairs of nodes & agglomerate

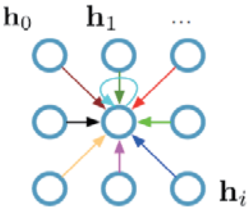
*slide from Thomas Kipf, **University of Amsterdam**

Recap: Convolutional Neural Networks (on grids)

Single CNN layer with 3x3 filter:



(Animation by Vincent Dumoulin)



Update for a single pixel:

- Transform messages individually $\mathbf{W}_i \mathbf{h}_i$
- Add everything up $\sum_i \mathbf{W}_i \mathbf{h}_i$

$\mathbf{h}_i \in \mathbb{R}^{F'}$ are (hidden layer) activations of a pixel/node

Full update:

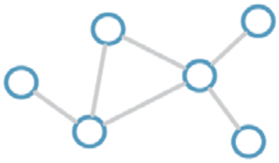
$$\mathbf{h}_4^{(l+1)} = \sigma\left(\mathbf{W}_0^{(l)}\mathbf{h}_0^{(l)} + \mathbf{W}_1^{(l)}\mathbf{h}_1^{(l)} + \cdots + \mathbf{W}_8^{(l)}\mathbf{h}_8^{(l)}\right)$$

*slide from Thomas Kipf, **University of Amsterdam**

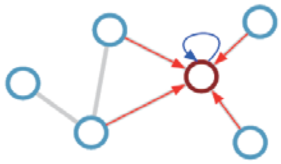
Graph Convolutional Networks (GCNs)

Kipf & Welling (ICLR 2017), related previous works by Duvenaud et al. (NIPS 2015) and Li et al. (ICLR 2016)

Consider this undirected graph:



Calculate update for node in red:



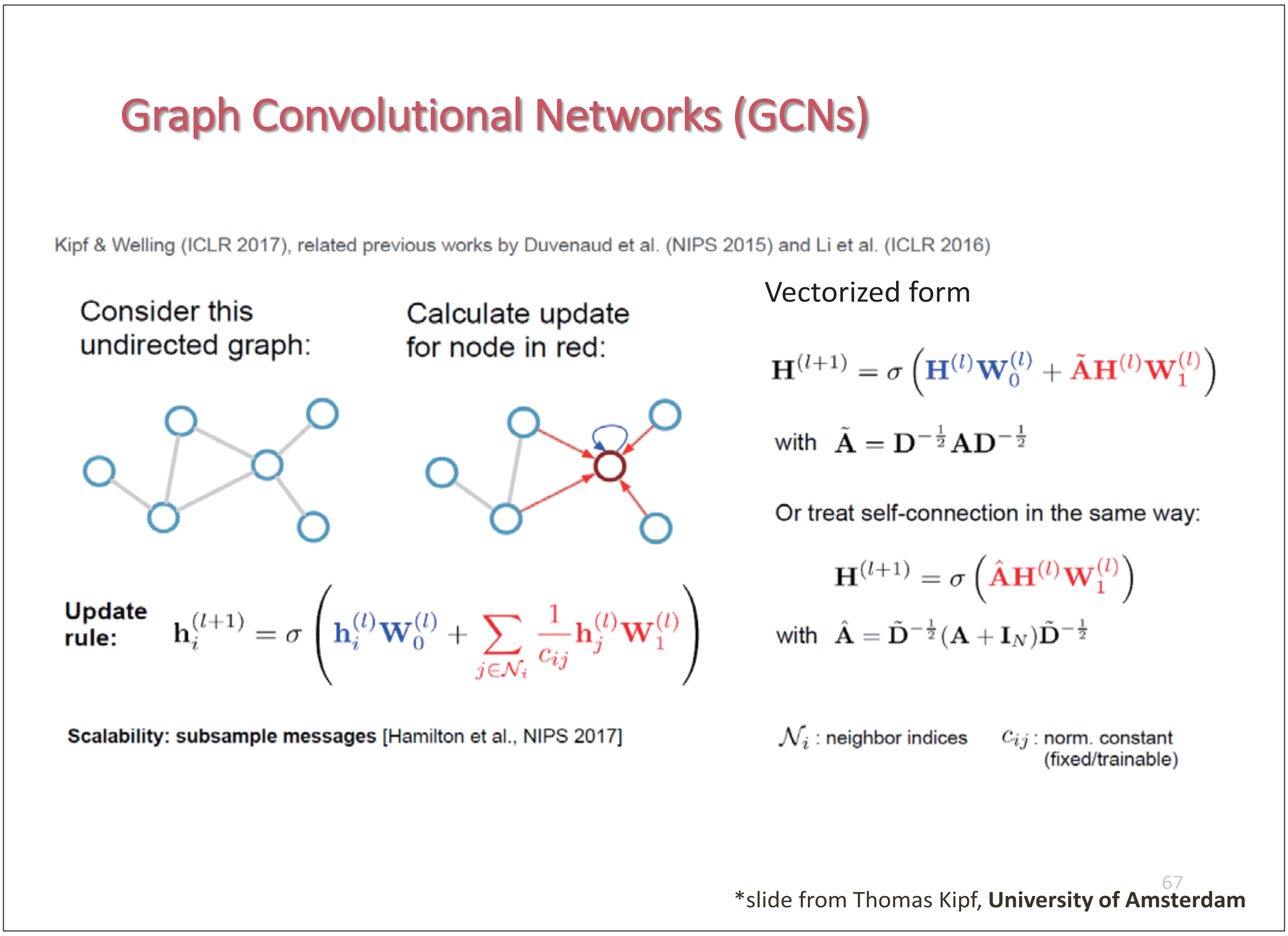
Update rule: $$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{h}_i^{(l)}\mathbf{W}_0^{(l)} + \sum_{j \in \mathcal{N}_i} \frac{1}{c_{ij}} \mathbf{h}_j^{(l)}\mathbf{W}_1^{(l)}\right)$$

Scalability: subsample messages [Hamilton et al., NIPS 2017]

$\mathcal{N}_i$: neighbor indices

c_{ij} : norm. constant (fixed/trainable)

*slide from Thomas Kipf, **University of Amsterdam**



Graph Convolutional Networks (GCNs)

Kipf & Welling (ICLR 2017), related previous works by Duvenaud et al. (NIPS 2015) and Li et al. (ICLR 2016)

Consider this undirected graph:

Calculate update for node in red:

Update rule: $$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{h}_i^{(l)}\mathbf{W}_0^{(l)} + \sum_{j\in\mathcal{N}_i}\frac{1}{c_{ij}}\mathbf{h}_j^{(l)}\mathbf{W}_1^{(l)}\right)$$

Scalability: subsample messages [Hamilton et al., NIPS 2017]

Vectorized form

$$\mathbf{H}^{(l+1)} = \sigma\left(\mathbf{H}^{(l)}\mathbf{W}_0^{(l)} + \tilde{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}_1^{(l)}\right)$$

with $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$

Or treat self-connection in the same way:

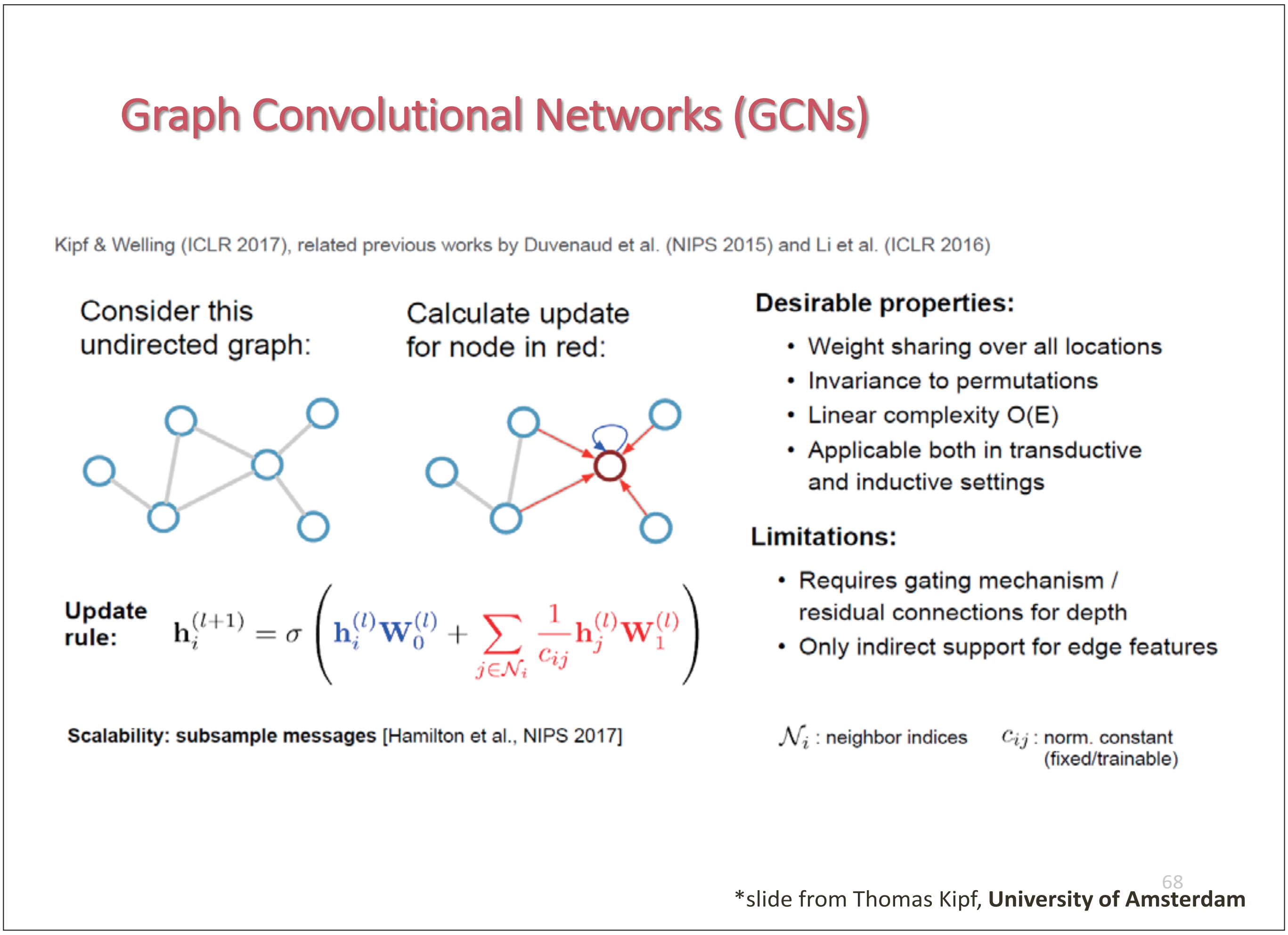
$$\mathbf{H}^{(l+1)} = \sigma\left(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}_1^{(l)}\right)$$

with $\hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I}_N)\tilde{\mathbf{D}}^{-\frac{1}{2}}$

$\mathcal{N}_i$: neighbor indices c_{ij} : norm. constant (fixed/trainable)

67

*slide from Thomas Kipf, **University of Amsterdam**



Graph Convolutional Networks (GCNs)

Kipf & Welling (ICLR 2017), related previous works by Duvenaud et al. (NIPS 2015) and Li et al. (ICLR 2016)

Consider this undirected graph:

Calculate update for node in red:

Desirable properties:

- Weight sharing over all locations
- Invariance to permutations
- Linear complexity O(E)
- Applicable both in transductive and inductive settings

Limitations:

- Requires gating mechanism / residual connections for depth
- Only indirect support for edge features

Update rule: $$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{h}_i^{(l)}\mathbf{W}_0^{(l)} + \sum_{j\in\mathcal{N}_i}\frac{1}{c_{ij}}\mathbf{h}_j^{(l)}\mathbf{W}_1^{(l)}\right)$$

Scalability: subsample messages [Hamilton et al., NIPS 2017]

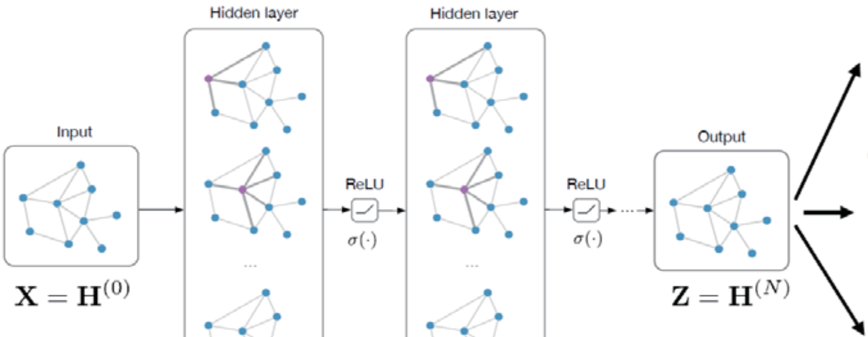
$\mathcal{N}_i$: neighbor indices c_{ij} : norm. constant (fixed/trainable)

68

*slide from Thomas Kipf, **University of Amsterdam**

Classification and link prediction with GNNs/GCNs

Input: Feature matrix $\mathbf{X} \in \mathbb{R}^{N \times E}$, preprocessed adjacency matrix $\hat{\mathbf{A}}$



$$\mathbf{H}^{(l+1)} = \sigma\left(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right)$$

Node classification:

$\mathrm{softmax}(\mathbf{z_n})$

e.g. Kipf & Welling (ICLR 2017)

Graph classification:

$\mathrm{softmax}(\sum_n \mathbf{z_n})$

e.g. Duvenaud et al. (NIPS 2015)

Link prediction:

$p(A_{ij}) = \sigma(\mathbf{z_i^T z_j})$

Kipf & Welling (NIPS BDL 2016)
"Graph Auto-Encoders"

*slide from Thomas Kipf, **University of Amsterdam**

Various GNNs - Isotropic

- Different *Aggregation* and *Update* functions are utilized for GNNs

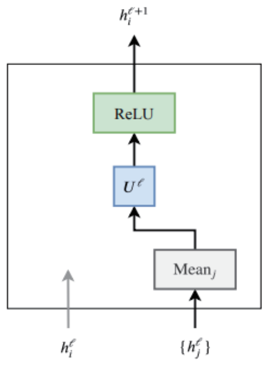


Figure 6. GCN Layer

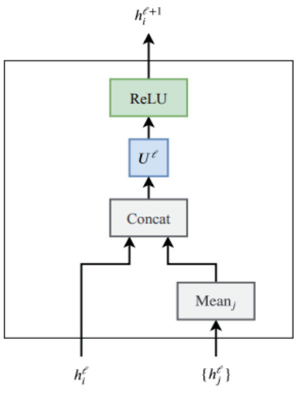


Figure 7. GraphSage Layer

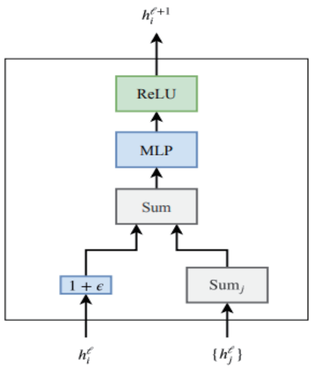


Figure 8. GIN Layer

Dwivedi, Vijay Prakash, et al. "Benchmarking graph neural networks." arXiv preprint arXiv:2003.00982 (2020).

Various GNNs - Anisotropic

- Different *Aggregation* and *Update* functions are utilized for GNNs
- Learn weights of neighborhoods

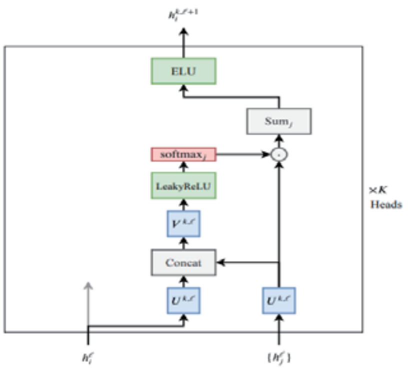


Figure 9. GAT Layer

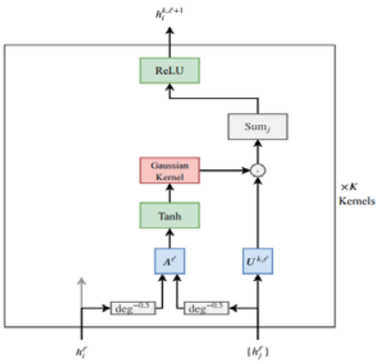


Figure 10. MoNet Layer

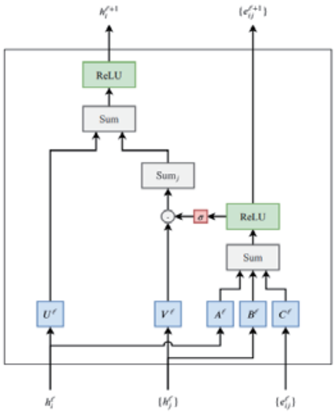


Figure 11. GatedGCN Layer

Dwivedi, Vijay Prakash, et al. "Benchmarking graph neural networks." arXiv preprint arXiv:2003.00982 (2020).

Summary of Part2

Summary

- **Graph**
    - A collection of interactions
    - Contains relationships between drugs, genes, and diseases
    - Heterogenous data types provide rich information but also cause technical challenges
- **Technologies**
    - Random Walk-Based Node Embedding
    - Network Propagation
    - Network Centralities / Clustering
    - VAE / Collective VAE
    - Matrix Factorization
    - Graph Neural Network

PART 3
Graph Learning for Drug Target Identification

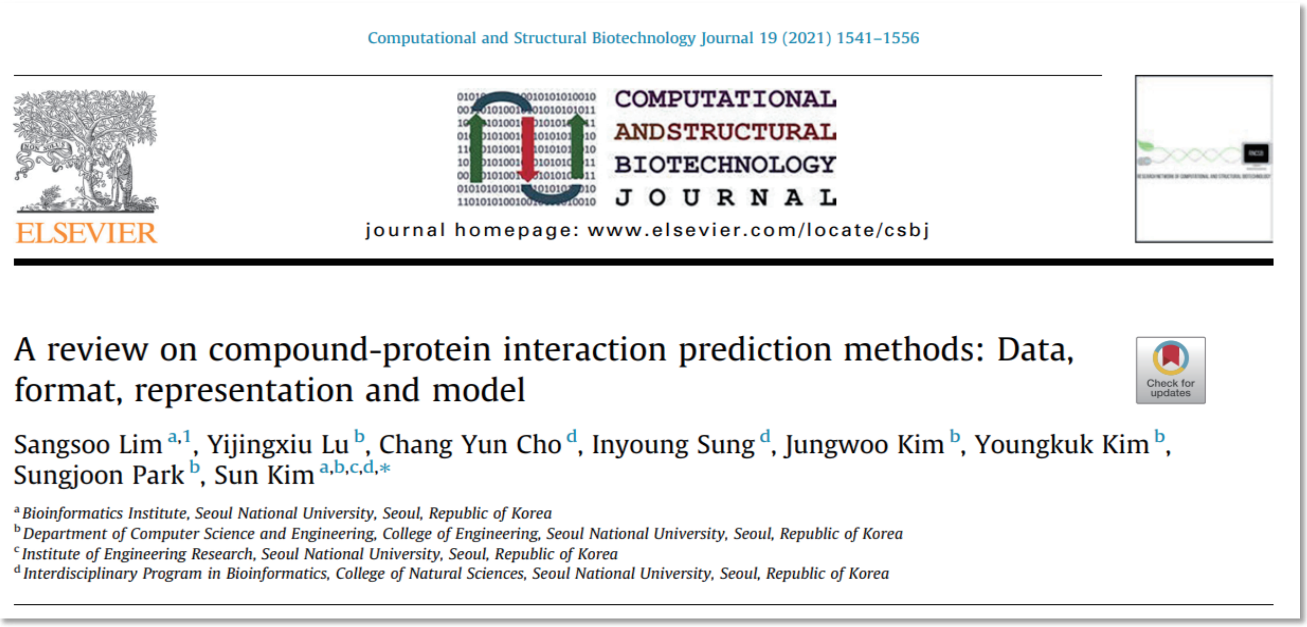
Contents

- **Current researches in DTI prediction**

- **Future directions in DTI prediction**
  - Heterogenous drug, gene, disease information
  - Downstream effect of drugs

- **Technologies for DTI**
  - deepDTnet (Chemical Science, 2020)
  - Drug embedding with target information
    (Briefings in Bioinformatics, accepted)

Current researches in DTI prediction



Computational and Structural Biotechnology Journal 19 (2021) 1541–1556

ELSEVIER

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A review on compound-protein interaction prediction methods: Data, format, representation and model

Sangsoo Lim [a,1], Yijingxiu Lu [b], Chang Yun Cho [d], Inyoung Sung [d], Jungwoo Kim [b], Youngkuk Kim [b], Sungjoon Park [b], Sun Kim [a,b,c,d,*]

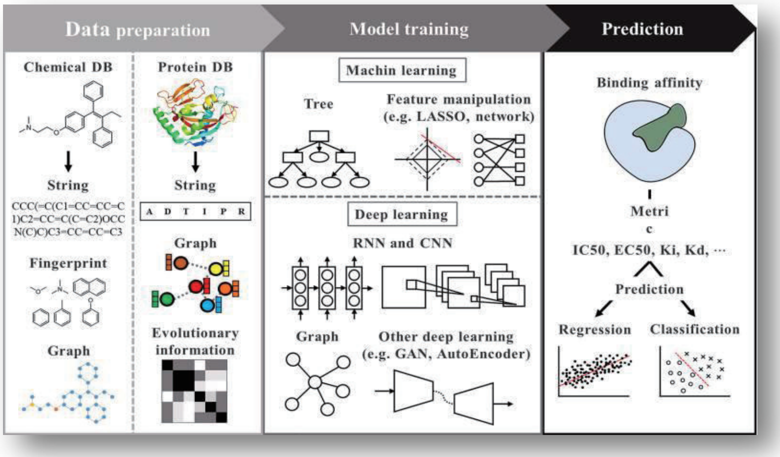
[a] Bioinformatics Institute, Seoul National University, Seoul, Republic of Korea
[b] Department of Computer Science and Engineering, College of Engineering, Seoul National University, Seoul, Republic of Korea
[c] Institute of Engineering Research, Seoul National University, Seoul, Republic of Korea
[d] Interdisciplinary Program in Bioinformatics, College of Natural Sciences, Seoul National University, Seoul, Republic of Korea

Computational and Structural Biotechnology Journal, 2021 (cited 25 times)

Review on DTI research

- **Background**:
  - AI approaches such as kernel-based, tree-based classifications, and neural network variations are recently applied to predicting affinity or interactions between small molecular drugs and protein targets.
  - DTI researches could be separated into three major parts: data preparation, model training, and prediction.

Overview of DTI prediction processes



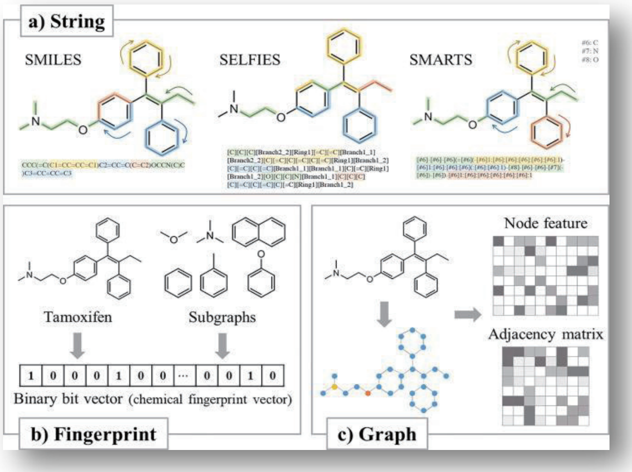
Review on DTI research

- **Data preparation:**
- Compounds:
  - Chemical compounds can be described naturally in a human-readable format such as strings, graphs, or images.
  - Chemical fingerprints that represents the existence of constitutive substructures/scaffolds or common functional groups are also widely used.
- Proteins:
  - Protein are represented as sequence of amino acids in most recent AI-based DTI researches.
  - To utilize protein 3D structures, it is common to convert it as chemically attributed spatial graphs.
  - Compared to the number of known amino acid sequences, number of known protein structures are much smaller.

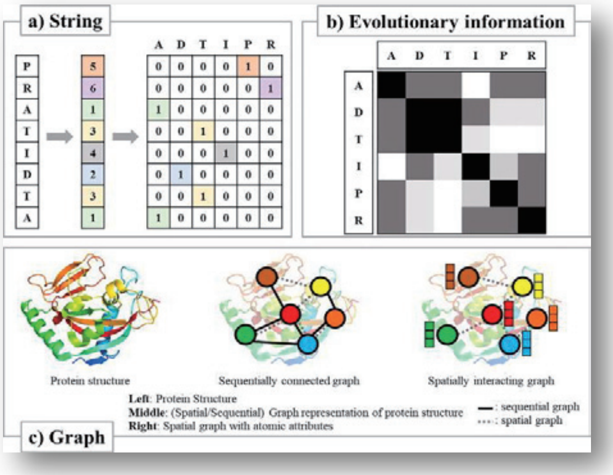
Review on DTI research

- **Data preparation:**

Formats and encoding schemes of compounds



Formats and encoding schemes of proteins


a) String
P
R
A
T
I
D
T
A
5
6
1
3
4
2
3
1
A D T I P R
b) Evolutionary information
A D T I P R
A
D
T
I
P
R
Protein structure
Sequentially connected graph
Spatially interacting graph
Left: Protein Structure
Middle: (Spatial/Sequential) Graph representation of protein structure
Right: Spatial graph with atomic attributes
: sequential graph
: spatial graph
c) Graph

Review on DTI research

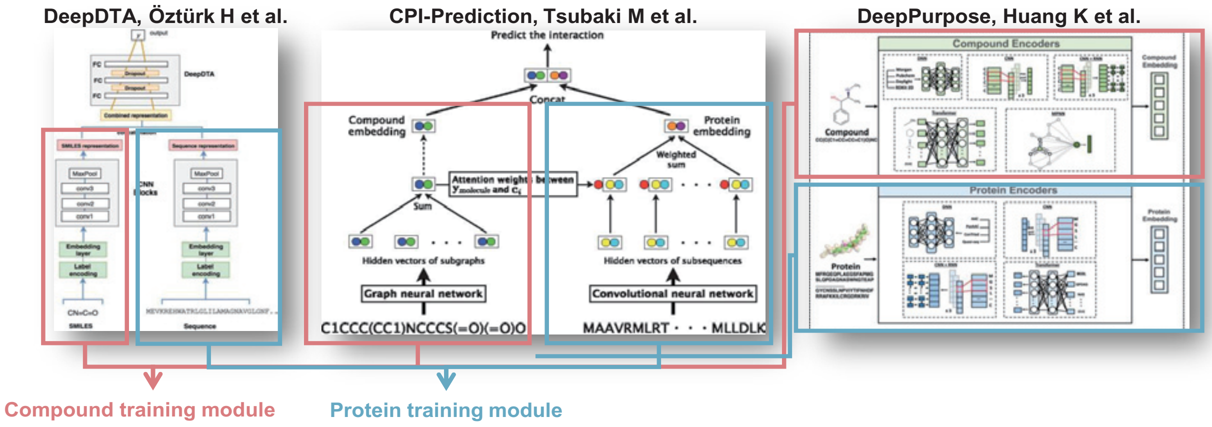
- **Model training:**

- Machine learning-based methods:
  - Decision tree, random forest
  - Support vector machine
  - Heterogeneous network

- Deep learning-based methods:
  - Recurrent neural network (RNN), Natural language processing (NLP)
  - Convolutional neural network (CNN)
  - Graph neural network (GNN)
  - Variational autoencoder (VAE) or generative adversarial network (GAN)

Typical model architectures for DTI

- **Train compounds and proteins** separately with two independent deep learning modules.
- **Combine latent vectors** of compounds and proteins for interaction prediction.



DeepDTA, Öztürk H et al.

CPI-Prediction, Tsubaki M et al.

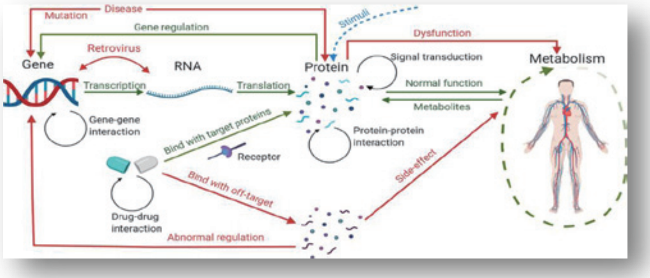
DeepPurpose, Huang K et al.

..ztụrk H, ..zgụr A, Ozkirimli E. *Bioinformatics,* 2018
Tsubaki M, Tomii K, Sese J. *Bioinformatics,* 2019,.
Huang K, Fu T, Glass L M, et al. *Bioinformatics,* 2020,.

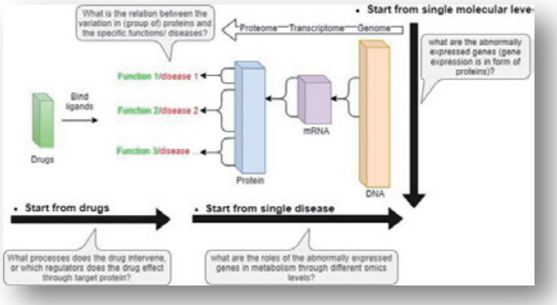
Drug – Target interaction
needs to consider
downstream effects, gene expressions!

Drug-target interactions and gene expression

- Drug molecules intervene in the regulatory process by binding with specific target ligands.
- Traditional treatment design based on physical parameters and external modalities or simple drug-target interactions are not sufficient for meeting clinical drug safety criteria or specifying variability among individuals.
- Modeling of the integrated clinical data and multi-layer molecular interactions makes the drug responses predictable.



A systemic view of disease


What is the relation between the variation in (group of) proteins and the specific functions/ diseases?
• Start from single molecular leve
Proteome—Transcriptome—Genome
what are the abnormally expressed genes (gene expression is in form of proteins)?
Function 1/disease 1
Bind ligands
Function 2/disease 2
mRNA
Function 3/disease ...
Drugs
Protein
DNA
• Start from drugs
• Start from single disease
What processes does the drug intervene, or which regulators does the drug effect through target protein?
what are the roles of the abnormally expressed genes in metabolism through different omics levels?

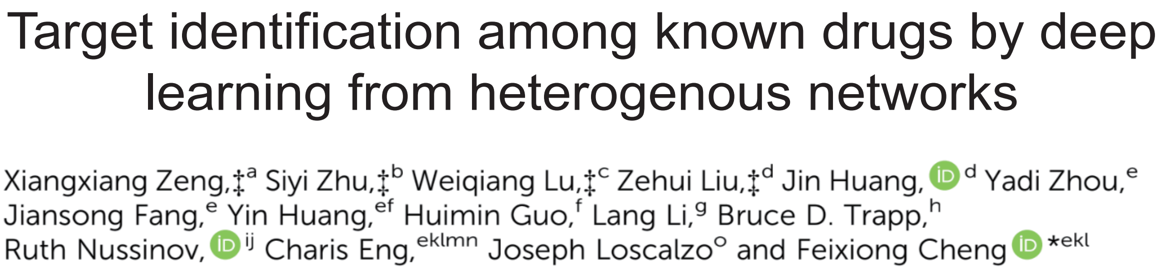
Analysis of disease and drug effect

Yue R, Dutta A. Computational systems biology in disease modeling and control, review and perspectives[J]. npj Systems Biology and Applications, 2022, 8(1): 1-16.

Technologies for DTI

- deepDTnet (Chemical Science, 2020)
- Drug embedding with target information (Briefings in Bioinformatics, accepted)

Chemical Science, 2020



Target identification among known drugs by deep learning from heterogenous networks

Xiangxiang Zeng,‡[a] Siyi Zhu,‡[b] Weiqiang Lu,‡[c] Zehui Liu,‡[d] Jin Huang, [d] Yadi Zhou,[e] Jiansong Fang,[e] Yin Huang,[ef] Huimin Guo,[f] Lang Li,[g] Bruce D. Trapp,[h] Ruth Nussinov, [ij] Charis Eng,[eklmn] Joseph Loscalzo[o] and Feixiong Cheng *[ekl]

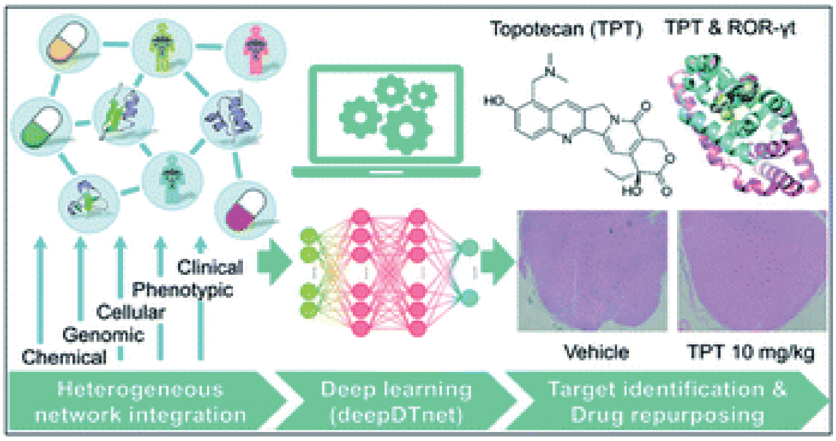
Motivation

- Drug target identification is a crucial process for drug discovery and effective treatment of human diseases
- Unintended therapeutic effects or multiple drug-target interactions leading to off-target toxicities and suboptimal effectiveness
- Experimental determination of drug-target interactions is costly and time-consuming
- **Challenge**
  - the features learned from the unsupervised learning procedure did not capture non-linearity
  - randomly selected drug–target pairs as negative samples often cause potential false positive rate
- **Approach**: a network-based deep learning for *in silico* identification of molecular targets for known drugs
  - Embeds 15 types of chemical, genomic, phenotypic, and cellular networks
  - Generate biologically and pharmacologically relevant features through learning low-dimensional but informative vectors for both drugs and targets
  - To address the lack of negative samples, they utilized Positive-Unlabeled (PU) setting

DeepDTnet

- **DeepDTnet** is a deep learning methodology for new target identification and drug repurposing in a heterogeneous drug–gene–disease network embedding 15 types of chemical, genomic, phenotypic, and cellular network profiles.

Overview of deepDTnet


Topotecan (TPT)
TPT & ROR-γt
HO
N
O
Clinical
Phenotypic
Cellular
Genomic
Chemical
Vehicle
TPT 10 mg/kg
Heterogeneous network integration
Deep learning (deepDTnet)
Target identification & Drug repurposing

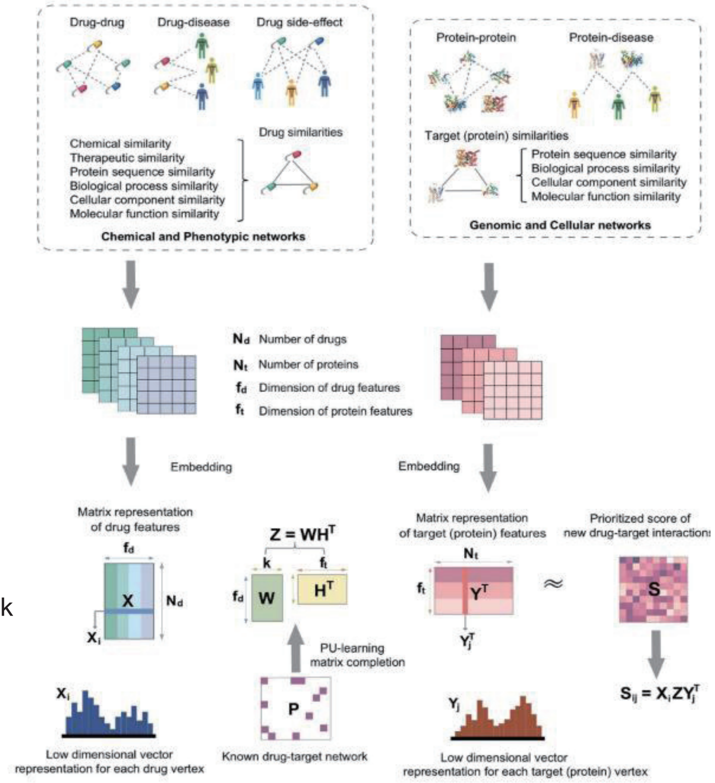
Model overview

- **Input**:
  - 15 types of chemical, genomic, phenotypic, and cellular networks for 732 drugs and 1,178 targets.
- **Output**:
  - The likelihood of the pairwise interaction score between drugs and targets.
- **Methodology**:
  - DeepDTnet learns low-dimensional vector representation of the features for each node in the heterogeneous network.
  - After learning the feature matrix for drugs and targets, deepDTnet applies PU-matrix completion to find the best projection from the drug space onto target (protein) space.
  - Finally, deepDTnet infers new targets for a drug ranked by geometric proximity to the projected feature vector of the drug in the projected space.

Model overview

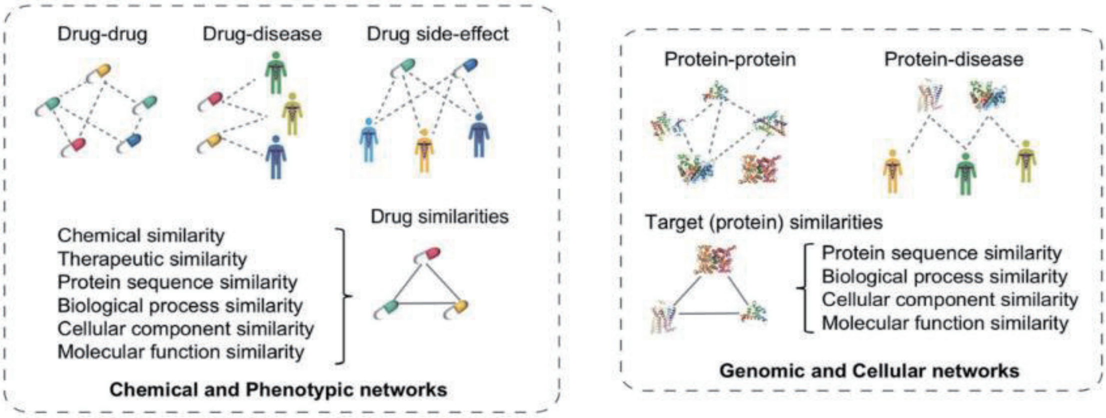
Learn the low-dimensional vectors for drugs, diseases

PU-matrix completion algorithm for the lack of publicly available negative samples


Drug-drug
Drug-disease
Drug side-effect
Chemical similarity
Therapeutic similarity
Protein sequence similarity
Biological process similarity
Cellular component similarity
Molecular function similarity
Drug similarities
Chemical and Phenotypic networks
Protein-protein
Protein-disease
Target (protein) similarities
Protein sequence similarity
Biological process similarity
Cellular component similarity
Molecular function similarity
Genomic and Cellular networks
N_d Number of drugs
N_t Number of proteins
f_d Dimension of drug features
f_t Dimension of protein features
Embedding
Embedding
Matrix representation of drug features
f_d
X
N_d
X_i
$Z = WH^T$
k
f_t
f_d
W
H^T
PU-learning matrix completion
P
Known drug-target network
Matrix representation of target (protein) features
N_t
f_t
Y^T
Y_j^T
≈
Prioritized score of new drug-target interactions
S
$S_{ij} = X_i Z Y_j^T$
X_i
Low dimensional vector representation for each drug vertex
Y_j
Low dimensional vector representation for each target (protein) vertex

Heterogenous networks

- Various databases are collected and utilized
  - Ex) drug-target network: DrugBank, Therapeutic Target Database, PharmGKB
  - Ex) disease-gene network: OMIM, CTD, HuGE navigator



Drug-drug

Drug-disease

Drug side-effect

Drug similarities

Chemical similarity
Therapeutic similarity
Protein sequence similarity
Biological process similarity
Cellular component similarity
Molecular function similarity

Chemical and Phenotypic networks

Protein-protein

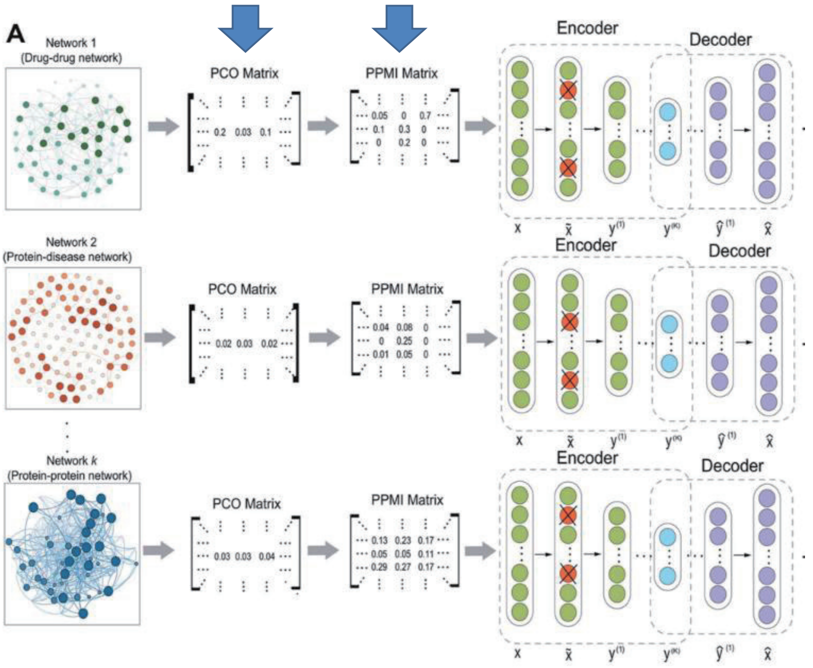
Protein-disease

Target (protein) similarities

Protein sequence similarity
Biological process similarity
Cellular component similarity
Molecular function similarity

Genomic and Cellular networks

Step1: low-dimensional representaions



Probabilistic Co-Occurrence matrix & Positive Pointwise Mutual Information

- Network propagation learns both local and global topological information
- After k step, a **probabilistic co-occurrence matrix** is obtained for each network

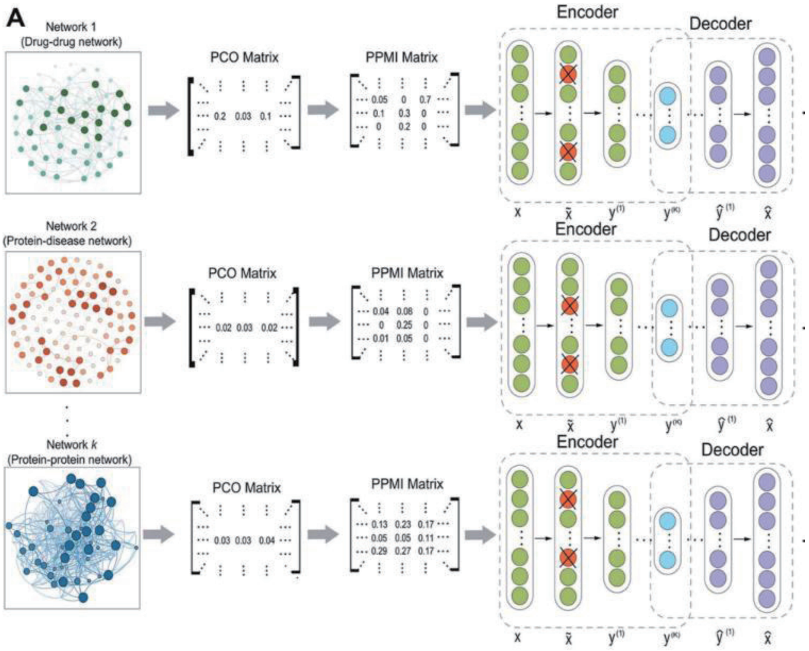
$$p_k = \omega \cdot p_{k-1}A + (1-\omega)p_0$$

- **A positive pointwise mutual information (PPMI)** matrix is calculated to obtain drug representaions

$$\mathrm{PPMI} = \max\left(\log \frac{M(i,j) * \sum_{i}^{N_r} \sum_{j}^{N_c} M(i,j)}{\sum_{i}^{N_r} M(i,j) * \sum_{j}^{N_c} M(i,j)},\ 0 \right)$$

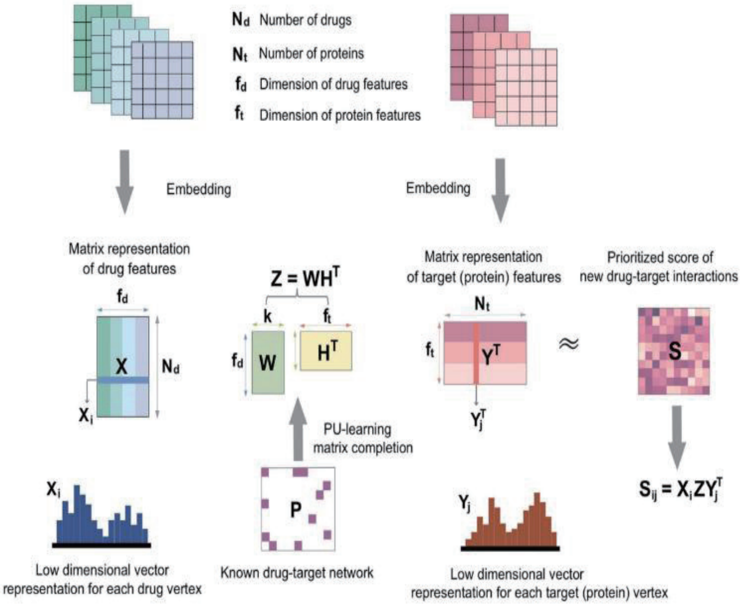
M : the original co-occurrence matrix,
N_r : the number of rows
N_c : the number of columns.

Step1: low-dimensional representaions



Stacked denoising autoencoder is utilized for learning low-dimensional vectors

Step2: PU-based matrix completion



N_d Number of drugs
N_t Number of proteins
f_d Dimension of drug features
f_t Dimension of protein features
Embedding
Embedding
Matrix representation of drug features
f_d
X
N_d
X_i
$Z = WH^T$
k
f_t
f_d
W
H^T
Matrix representation of target (protein) features
N_t
f_t
Y^T
Y_j^T
≈
Prioritized score of new drug-target interactions
S
PU-learning matrix completion
X_i
P
Y_j
$S_{ij} = X_i Z Y_j^T$
Low dimensional vector representation for each drug vertex
Known drug-target network
Low dimensional vector representation for each target (protein) vertex

Inductive matrix completion

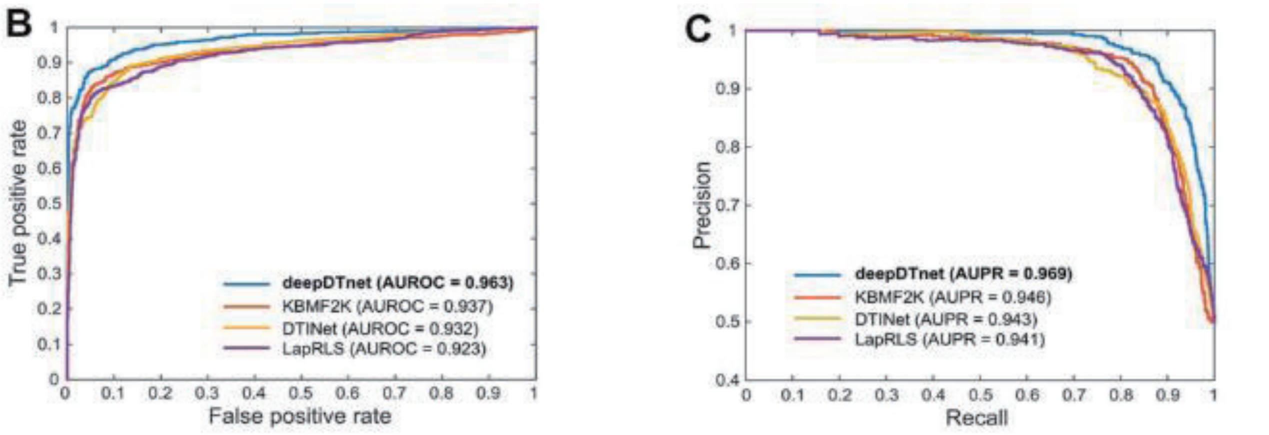
$$\min_{W,H} \sum_{(i,j)\in\Omega} \ell(P_{ij},\ x_i^T W H^T y_j) + \frac{\lambda}{2}\left(\|W\|_F^2 + \|H\|_F^2\right)$$

PU-matrix completion

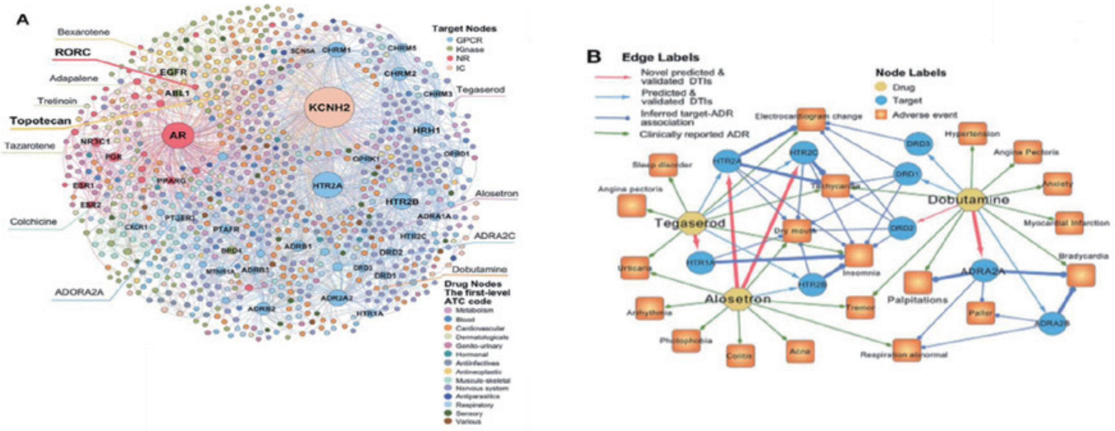
$$\min_{W,H} \sum_{(i,j)\in\Omega^+} \left(P_{ij} - x_i W H^T y_j^T\right)^2 + \alpha \sum_{(i,j)\in\Omega^-} \left(P_{ij} - x_i W H^T y_j^T\right)^2$$
$$+ \lambda\left(\|W\|_F^2 + \|H\|_F^2\right)$$

α: the penalty of the unobserved entries toward zero

Results: Perfomance of DTI prediction



Results: The uncovered drug-target network



Summary

- Deep learning model for learning heterogeneous drug-gene-disease newtork
- Key points
  - Learn multiple chemical & genomic information as low-dimensional embeddings
  - Apply PU-matrix completion to address sparsity of postivie samples and lack of negative samples in DTI

Briefings in Bioinformatics, 2023, accepted

Improved Drug Response Prediction by Drug Target Data Integration via Network-based Profiling

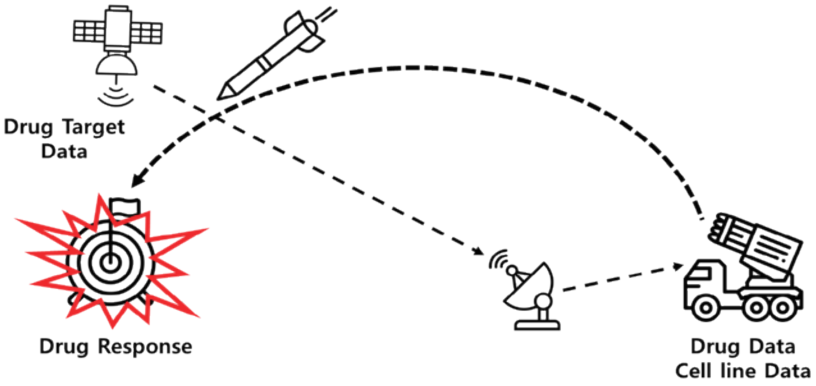
Minwoo Pak,[1,†] Sangseon Lee,[2,†] Inyoung Sung[3] and Sun Kim[1,3,4,*]

Motivation

- Drug response prediction is important for precision medicine in that it can help predict how a patient would react to a drug before the actual administration
- Intuitively, use of drug target interaction (DTI) information can be useful for drug response prediction
- **Challenge**: use of DTI is difficult because existing drug response database such as CCLE and GDSC do not have information about transcriptome after drug treatment
- **Approach**: framework, NetGP that can improve existing deep learning-based drug response prediction models by effectively utilizing drug target information.
  - a module to compute gene perturbation scores by the network propagation technique on a Protein-Protein Interaction (PPI) network
  - NetGP with the network propagation technique produces perturbation effects by the pharmacologic modulation of target gene
  - a model-agnostic way so that any existing DTI tool can be incorporated.

Motivation

- **Drug response** prediction is highly significant in precision medicine in that it can help predict how a patient would react to a drug before the actual administration.
- **Drug target information** represents the mechanism of the drug affecting a cell thereby bridging the relationship between the two.



*GDSC: Genomics of Drug Sensitivity in Cancer
*CADD: Chemoinformatics Tools and User Services

Model overview

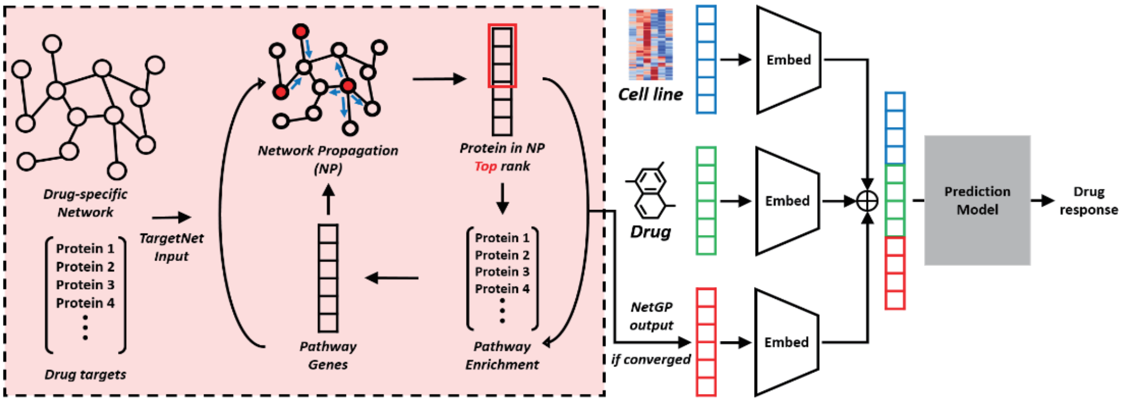
- **Input**
  - Drug response information from GDSC
  - Drug SMILES data from CADD
  - Protein-protein interaction network from STRING
  - Drug target information from GDSC and DrugBank
- **Model**
  - **TargetNet**: drug target profile extraction algorithm
  - **Placeholder** drug response prediction method
- **Output**
  - Drug response: IC50 or area under dose-response curve value

Overview of NetGP

- Integration with existing tools in terms of embedding vector (in model-agnostic way)

NetGP: Drug Target Profile Extraction Algorithm

Drug Target integration for Drug Response Prediction

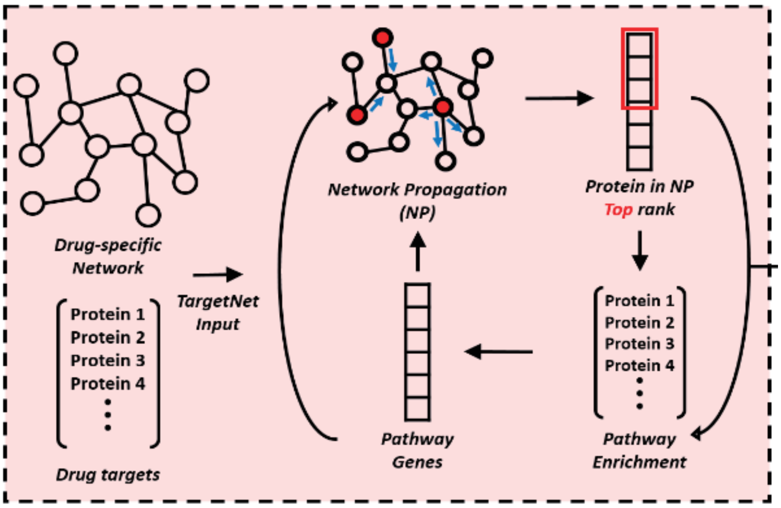


- Simulate a perturbation effect of a given drug using drug target information and PPI network → network propagation

NetGP: Model detail

- **Phase 1:** network-based drug target profile extraction phase

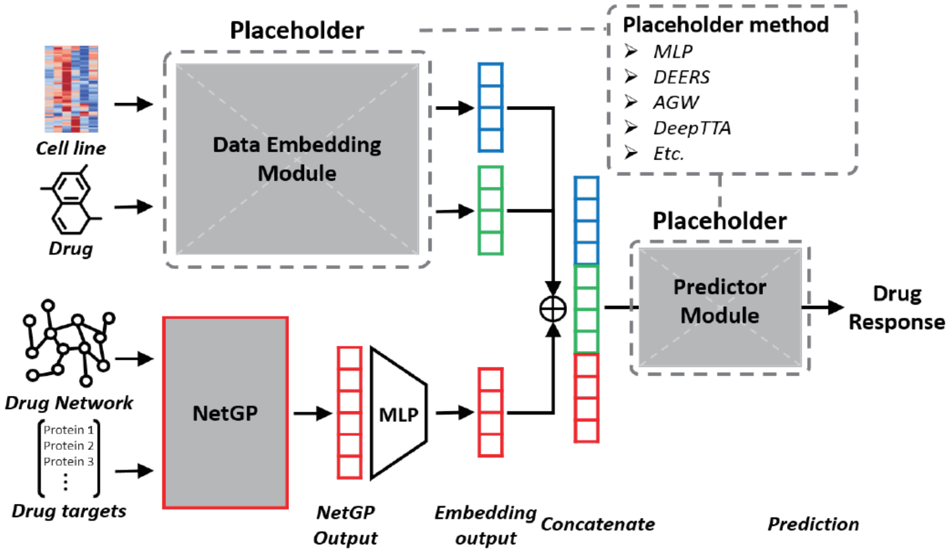
NetGP: Drug Target Profile Extraction Algorithm



- Network propagation identifies affected candidate genes from drug target genes
- Iteratively perform network propagation with enriched biological mechanisms
  - Network propagation prunes to biased seeds and network topology
  - Iteration will remove noises

NetGP: Model detail

- **Phase 2**: drug target profile integration
  - Embed cell line, drug and drug target profile from NetGP
  - Any deep learning model can be replaced with Placeholder



Placeholder

Data Embedding Module

Cell line

Drug

Drug Network

Protein 1
Protein 2
Protein 3
⋮

Drug targets

NetGP

MLP

NetGP Output

Embedding output

Concatenate

Placeholder method
- MLP
- DEERS
- AGW
- DeepTTA
- Etc.

Placeholder

Predictor Module

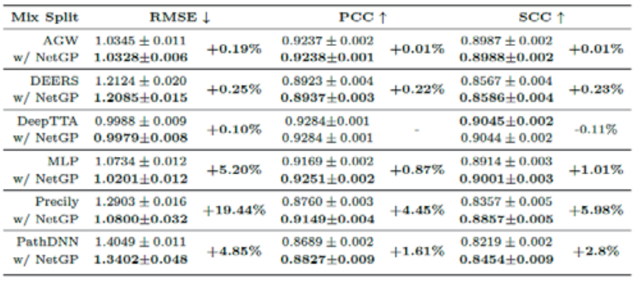
Drug Response

Prediction

Results: Performance of Drug Response Prediction

- Drug response prediction performance gain by integrating TargetNet
  - 1st row: Placeholder method
  - 2nd row: Placeholder method + NetGP

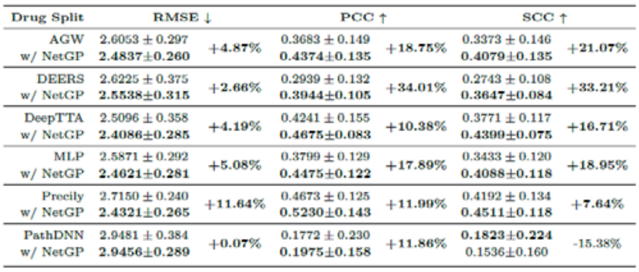
- Traditional evaluation scheme



Mix Split	RMSE ↓		PCC ↑		SCC ↑	
AGW	1.0345 ± 0.011	+0.19%	0.9237 ± 0.002	+0.01%	0.8987 ± 0.002	+0.01%
w/ NetGP	**1.0328±0.006**		**0.9238±0.001**		**0.8988±0.002**	
DEERS	1.2124 ± 0.020	+0.25%	0.8923 ± 0.004	+0.22%	0.8567 ± 0.004	+0.23%
w/ NetGP	**1.2085±0.015**		**0.8937±0.003**		**0.8586±0.004**	
DeepTTA	0.9988 ± 0.009	+0.10%	0.9284±0.001	-	**0.9045±0.002**	-0.11%
w/ NetGP	**0.9979±0.008**		0.9284 ± 0.001		0.9044 ± 0.002	
MLP	1.0734 ± 0.012	+5.20%	0.9169 ± 0.002	+0.87%	0.8914 ± 0.003	+1.01%
w/ NetGP	**1.0201±0.012**		**0.9251±0.002**		**0.9001±0.003**	
Precily	1.2903 ± 0.016	+19.44%	0.8760 ± 0.003	+4.45%	0.8357 ± 0.005	+5.98%
w/ NetGP	**1.0800±0.032**		**0.9149±0.004**		**0.8857±0.005**	
PathDNN	1.4049 ± 0.011	+4.85%	0.8689 ± 0.002	+1.61%	0.8219 ± 0.002	+2.8%
w/ NetGP	**1.3402±0.048**		**0.8827±0.009**		**0.8454±0.009**	

(b) Mix Split

- Unseen drugs during training

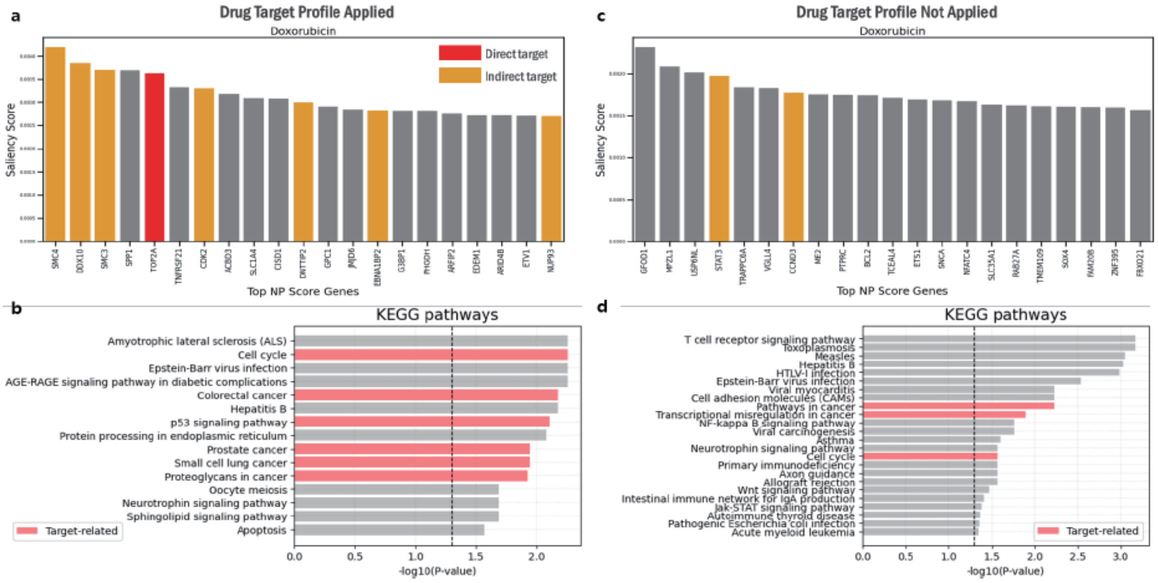


Drug Split	RMSE ↓		PCC ↑		SCC ↑	
AGW	2.6053 ± 0.297	+4.87%	0.3683 ± 0.149	+18.75%	0.3373 ± 0.146	+21.07%
w/ NetGP	**2.4837±0.260**		**0.4374±0.135**		**0.4079±0.135**	
DEERS	2.6225 ± 0.375	+2.66%	0.2939 ± 0.132	+34.01%	0.2743 ± 0.108	+33.21%
w/ NetGP	**2.5538±0.315**		**0.3944±0.105**		**0.3647±0.084**	
DeepTTA	2.5096 ± 0.358	+4.19%	0.4241 ± 0.155	+10.38%	0.3771 ± 0.117	+16.71%
w/ NetGP	**2.4086±0.285**		**0.4675±0.083**		**0.4399±0.075**	
MLP	2.5871 ± 0.292	+5.08%	0.3799 ± 0.129	+17.89%	0.3433 ± 0.120	+18.95%
w/ NetGP	**2.4621±0.281**		**0.4475±0.122**		**0.4088±0.118**	
Precily	2.7150 ± 0.240	+11.64%	0.4673 ± 0.125	+11.99%	0.4192 ± 0.134	+7.64%
w/ NetGP	**2.4321±0.265**		**0.5230±0.143**		**0.4511±0.118**	
PathDNN	2.9481 ± 0.384	+0.07%	0.1772 ± 0.230	+11.86%	**0.1823±0.224**	-15.38%
w/ NetGP	**2.9456±0.289**		**0.1975±0.158**		0.1536±0.160	

(a) Drug Split

Results: Gene importance analysis

- Drug example: Doxorubicin

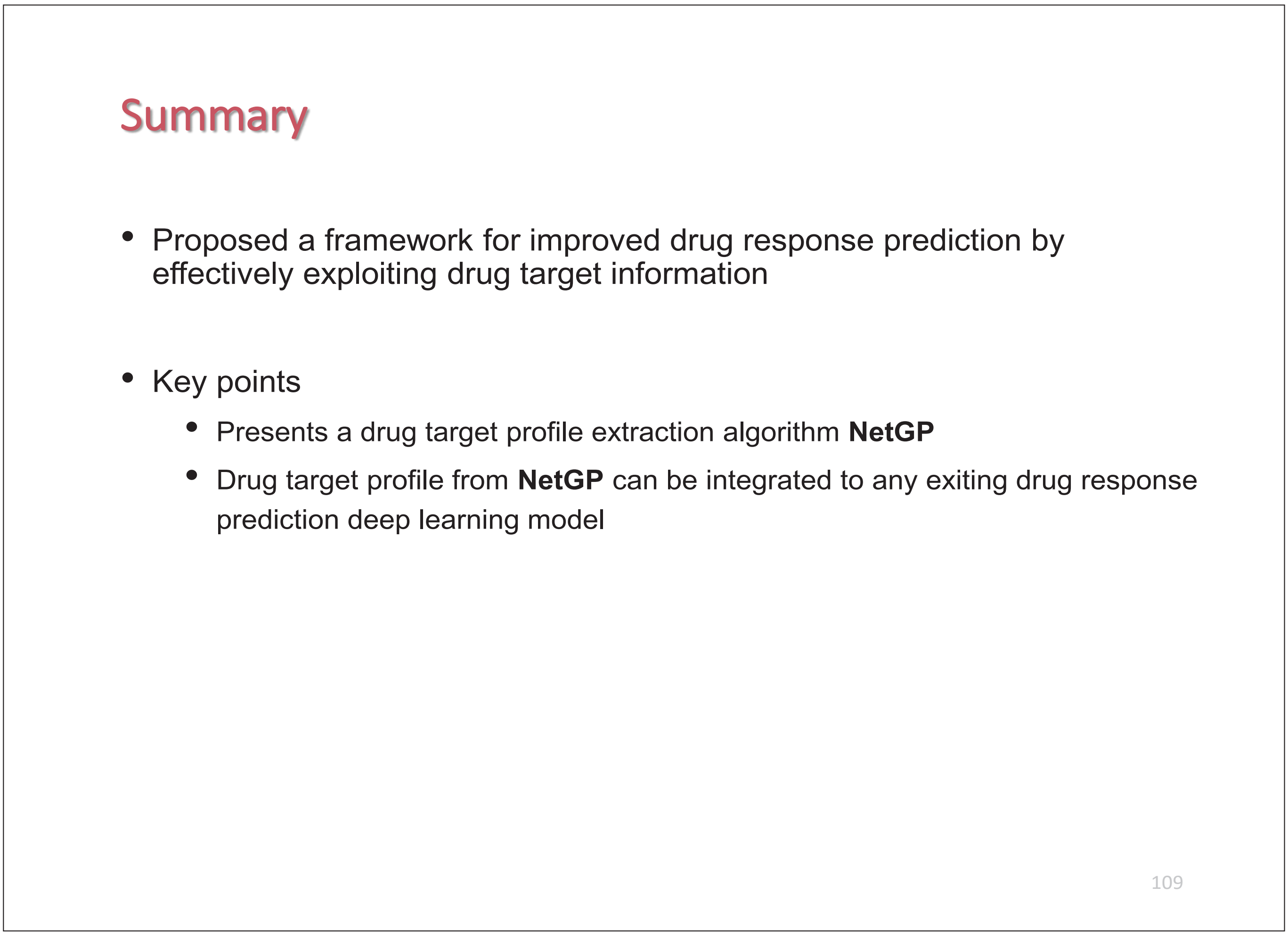


Results: Effect of Drug Target Information

- Use of drug target profile boosts prediction performance, especially for drugs with explicit target proteins known

Table 3. Explicit Target Drugs vs. Non-explicit Target Drugs. * indicates explicit target pathway.

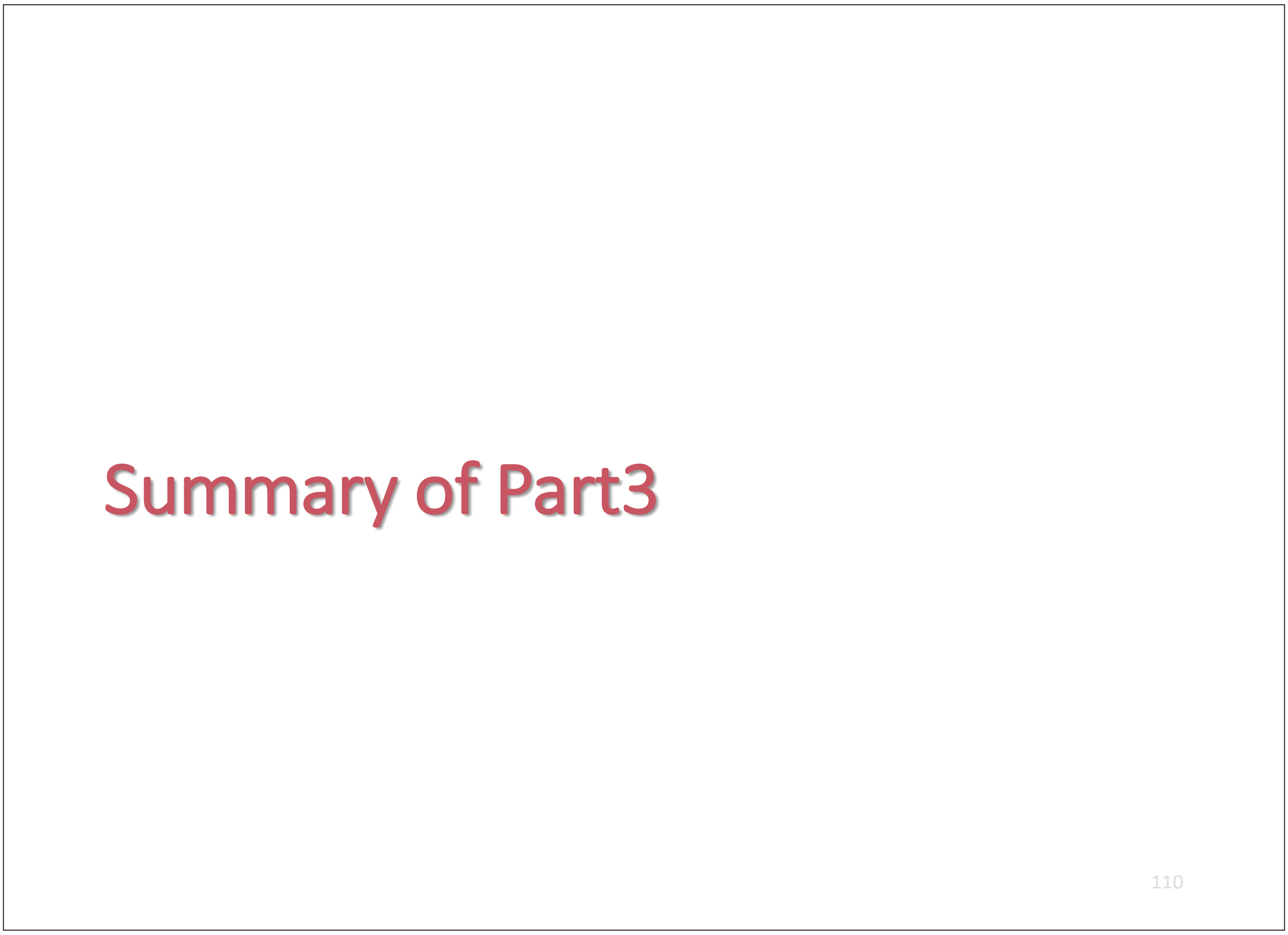
Category	Default	Framework Applied	Difference
↑ Default	0.3154	0.4532	**+43.69%**
↑ DNA Replication	0.3794	0.3835	**+1.08%**
↑ Mitosis	0.7467	0.7542	**+1.00%**
↑ *Other, Kinase	0.3930	0.6959	**+77.07%**



Summary

- Proposed a framework for improved drug response prediction by effectively exploiting drug target information

- Key points
  - Presents a drug target profile extraction algorithm **NetGP**
  - Drug target profile from **NetGP** can be integrated to any exiting drug response prediction deep learning model



Summary of Part3

Summary

- **Graph Learning for DTI**
  - Current DTI studies focus only drugs and targets of interest.
  - Learning heterogenous relationships between drugs, genes, and diseases is important.
  - Downstream effects of drugs will improve drug-target idenfication and drug response prediction.

PART 4
Drug Repurposing

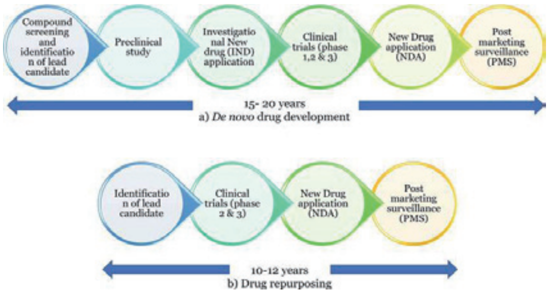
Contents

- **Introduction**
- **Examples**
  - Baricitinib
  - DrugCell
  - Deep learning approach to Antibiotic discovery
  - Literature-based approaches
- **Networks and Databases**
  - Networks
    - PPI – STRING, BioGRID
    - Biological pathways – KEGG, Reactome
    - Disease networks – Diseasome, HDN, DGN
    - Comprehensive heterogeneous networks – Hetio, MSI
  - Databases
    - Drug Repurposing Hub
    - RepoDB
    - CTD
    - PharmacoDB
- **Technologies**
  - Network analysis
    - Network centralities
    - Network clustering – K-means, Hierarchical
    - Network propagation – PropaNet, MLDEG
  - Network representation learning
    - word2vec – DeepWalk, node2vec, DREAMwalk
    - Graph Neural Network
- **Network-based drug repurposing: cases**
  - SNF-cVAE (Knowledge-Based Systems, 2021)
  - CBPred (Cells, 2019)
  - DeepDR (Bioinformatics, 2019)
  - BiFusion (ISMB 2020)
  - DreamWalk (in review)

Drug repositioning (or repurposing)

- Repurposing of old drugs to treat diseases is increasingly becoming an attractive proposition.
- Advantages of repurposing drugs
  - Risk of failure is lower
  - Time frame can be reduced
  - Less investment is needed

  → Less risky and more rapid return in investment!



Compound screening and identificatio n of lead candidate
Preclinical study
Investigatio nal New drug (IND) application
Clinical trials (phase 1,2 & 3)
New Drug application (NDA)
Post marketing surveillance (PMS)
15- 20 years
a) De novo drug development
Identificatio n of lead candidate
Clinical trials (phase 2 & 3)
New Drug application (NDA)
Post marketing surveillance (PMS)
10-12 years
b) Drug repurposing

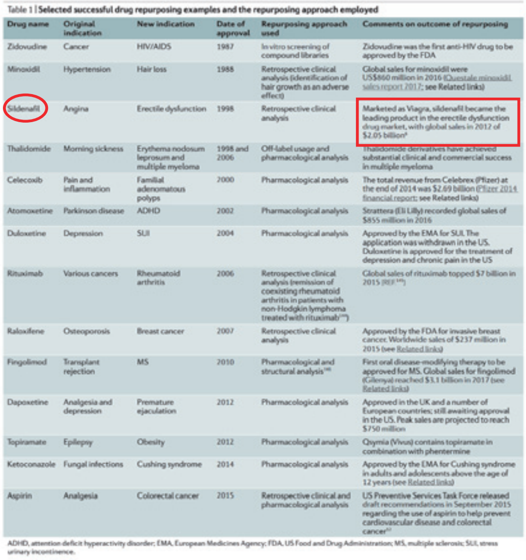
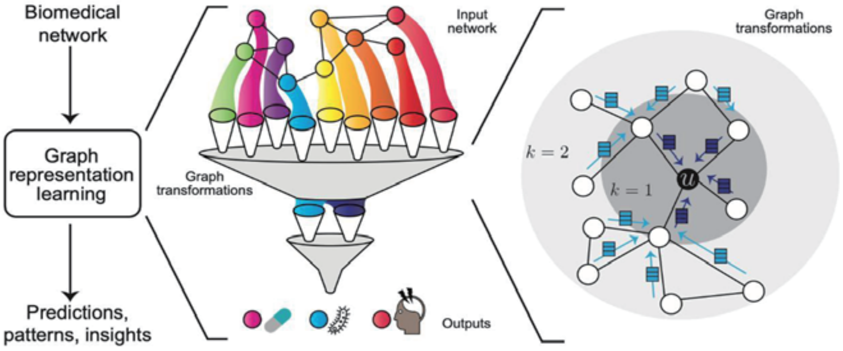


Table 1 | Selected successful drug repurposing examples and the repurposing approach employed

Drug name	Original indication	New indication	Date of approval	Repurposing approach used	Comments on outcome of repurposing
Zidovudine	Cancer	HIV/AIDS	1987	In vitro screening of compound libraries	Zidovudine was the first anti-HIV drug to be approved by the FDA
Minoxidil	Hypertension	Hair loss	1988	Retrospective clinical analysis (identification of hair growth as an adverse effect)	Global sales for minoxidil were US$860 million in 2016 (Questale minoxidil sales report 2017; see Related links)
Sildenafil	Angina	Erectile dysfunction	1998	Retrospective clinical analysis	Marketed as Viagra, sildenafil became the leading product in the erectile dysfunction drug market, with global sales in 2012 of $2.05 billion[illegible]
Thalidomide	Morning sickness	Erythema nodosum leprosum and multiple myeloma	1998 and 2006	Off-label usage and pharmacological analysis	Thalidomide derivatives have achieved substantial clinical and commercial success in multiple myeloma
Celecoxib	Pain and inflammation	Familial adenomatous polyps	2000	Pharmacological analysis	The total revenue from Celebrex (Pfizer) at the end of 2014 was $2.69 billion (Pfizer 2014 financial report; see Related links)
Atomoxetine	Parkinson disease	ADHD	2002	Pharmacological analysis	Strattera (Eli Lilly) recorded global sales of $855 million in 2016
Duloxetine	Depression	SUI	2004	Pharmacological analysis	Approved by the EMA for SUI. The application was withdrawn in the US. Duloxetine is approved for the treatment of depression and chronic pain in the US
Rituximab	Various cancers	Rheumatoid arthritis	2006	Retrospective clinical analysis (remission of coexisting rheumatoid arthritis in patients with non-Hodgkin lymphoma treated with rituximab[illegible])	Global sales of rituximab topped $7 billion in 2015 (REF [illegible])
Raloxifene	Osteoporosis	Breast cancer	2007	Retrospective clinical analysis	Approved by the FDA for invasive breast cancer. Worldwide sales of $237 million in 2015 (see Related links)
Fingolimod	Transplant rejection	MS	2010	Pharmacological and structural analysis[illegible]	First oral disease-modifying therapy to be approved for MS. Global sales for fingolimod (Gilenya) reached $3.1 billion in 2017 (see Related links)
Dapoxetine	Analgesia and depression	Premature ejaculation	2012	Pharmacological analysis	Approved in the UK and a number of European countries; still awaiting approval in the US. Peak sales are projected to reach $750 million
Topiramate	Epilepsy	Obesity	2012	Pharmacological analysis	Qsymia (Vivus) contains topiramate in combination with phentermine
Ketoconazole	Fungal infections	Cushing syndrome	2014	Pharmacological analysis	Approved by the EMA for Cushing syndrome in adults and adolescents above the age of 12 years (see Related links)
Aspirin	Analgesia	Colorectal cancer	2015	Retrospective clinical and pharmacological analysis	US Preventive Services Task Force released draft recommendations in September 2015 regarding the use of aspirin to help prevent cardiovascular disease and colorectal cancer[illegible]

ADHD, attention deficit hyperactivity disorder; EMA, European Medicines Agency; FDA, US Food and Drug Administration; MS, multiple sclerosis; SUI, stress urinary incontinence.

Pushpakom, S., Iorio, F., Eyers, P. et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 18, 41–58 (2019)



Representation learning for networks in biology and medicine.

Li, Michelle M., Kexin Huang, and Marinka Zitnik. "Graph Representation Learning in Biomedicine." arXiv preprint arXiv:2104.04883 (2021).

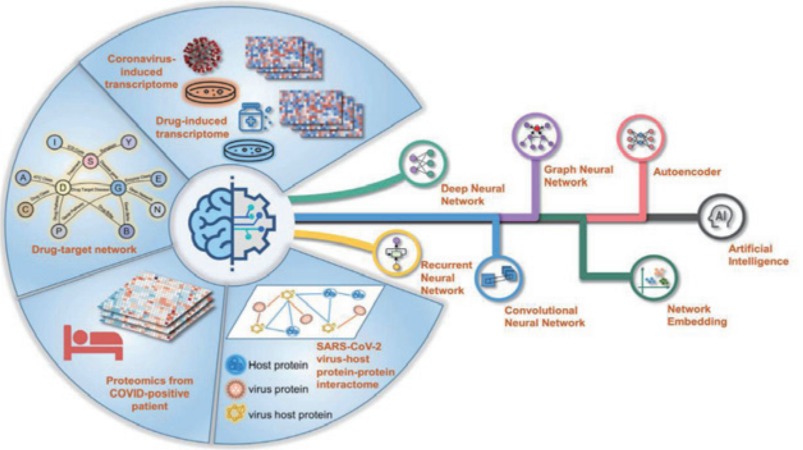
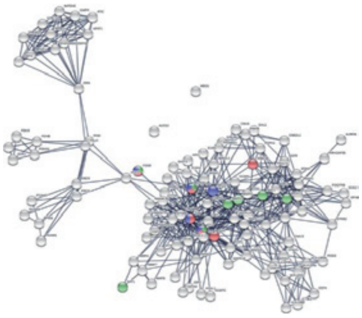


FIGURE 5 A diagram illustrating deep learning-based drug repurposing infrastructure for emerging development of host-targeting therapies to fight COVID-19 and future pandemic. We posited that approved drugs that specific human proteins/targets may offer potential host-targeting therapies for COVID-19 as COVID-19 may share biology with human cells and tissues from the SARS-CoV-2 virus–host protein–protein interactome perspective[3,4,6]

Pan, Xiaoqin, et al. "Deep learning for drug repurposing: Methods, databases, and applications." Wiley Interdisciplinary Reviews: Computational Molecular Science (2022)

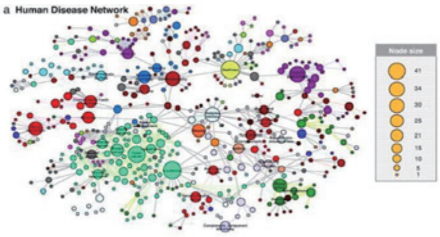
Disease and Biological Networks

- Networks is a method of representing systemic biological interactions between various biological objects.
- These networks or graphs are used to capture relationships between biological entities.



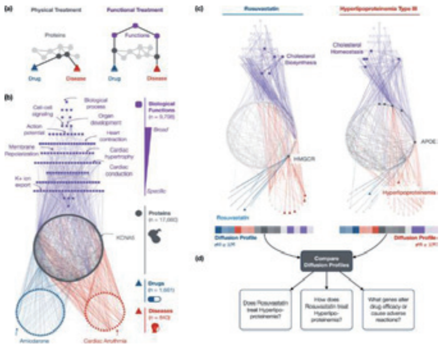
Protein-Protein Interaction Network of Heroin Use Disorder

Chen, SJ., Liao, DL., Chen, CH. et al. Construction and Analysis of Protein-Protein Interaction Network of Heroin Use Disorder. Sci Rep 9, 4980 (2019)



Human Disease Network

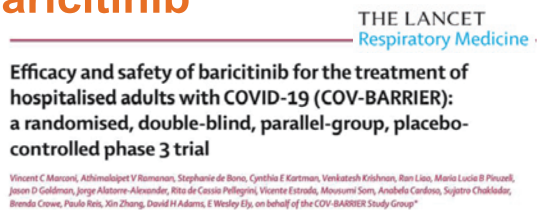
Goh, Kwang-Il, et al. "The human disease network." Proceedings of the National Academy of Sciences 104.21 (2007): 8685-8690.



Multi-scale Interactome Network

Ruiz, C., Zitnik, M. & Leskovec, J. Identification of disease treatment mechanisms through the multiscale interactome. Nat Commun 12, 1796 (2021)

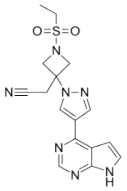
Baricitinib



THE LANCET
Respiratory Medicine

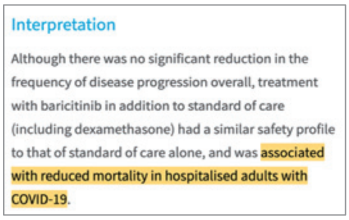
Efficacy and safety of baricitinib for the treatment of hospitalised adults with COVID-19 (COV-BARRIER): a randomised, double-blind, parallel-group, placebo-controlled phase 3 trial

*Vincent C Marconi, Athimalaipet V Ramanan, Stephanie de Bono, Cynthia E Kartman, Venkatesh Krishnan, Ran Liao, Maria Lucia B Piruzeli, Jason D Goldman, Jorge Alatorre-Alexander, Rita de Cassia Pellegrini, Vicente Estrada, Mousumi Som, Anabela Cardoso, Sujatro Chakladar, Brenda Crowe, Paulo Reis, Xin Zhang, David H Adams, E Wesley Ely, on behalf of the COV-BARRIER Study Group**



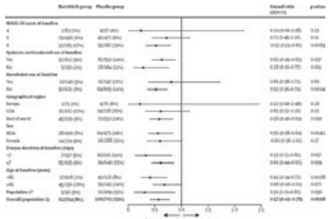
Baricitinib

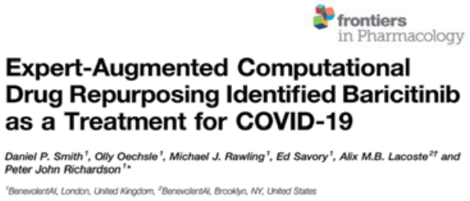
- Originally used for rheumatoid arthritis (RA).
- Inhibitor of Janus Kinase (JAK).



Interpretation

Although there was no significant reduction in the frequency of disease progression overall, treatment with baricitinib in addition to standard of care (including dexamethasone) had a similar safety profile to that of standard of care alone, and was associated with reduced mortality in hospitalised adults with COVID-19.



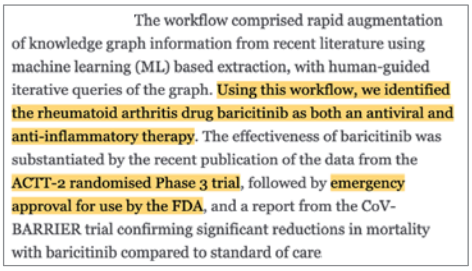


frontiers in Pharmacology

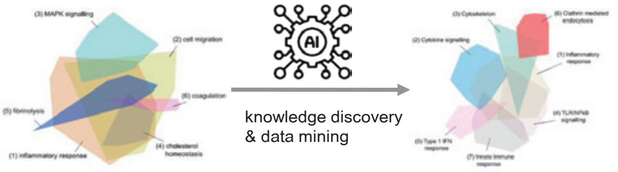
Expert-Augmented Computational Drug Repurposing Identified Baricitinib as a Treatment for COVID-19

Daniel P. Smith[1], Olly Oechsle[1], Michael J. Rawling[1], Ed Savory[1], Alix M.B. Lacoste[2†] and Peter John Richardson[1]*

[1]BenevolentAI, London, United Kingdom, [2]BenevolentAI, Brooklyn, NY, United States

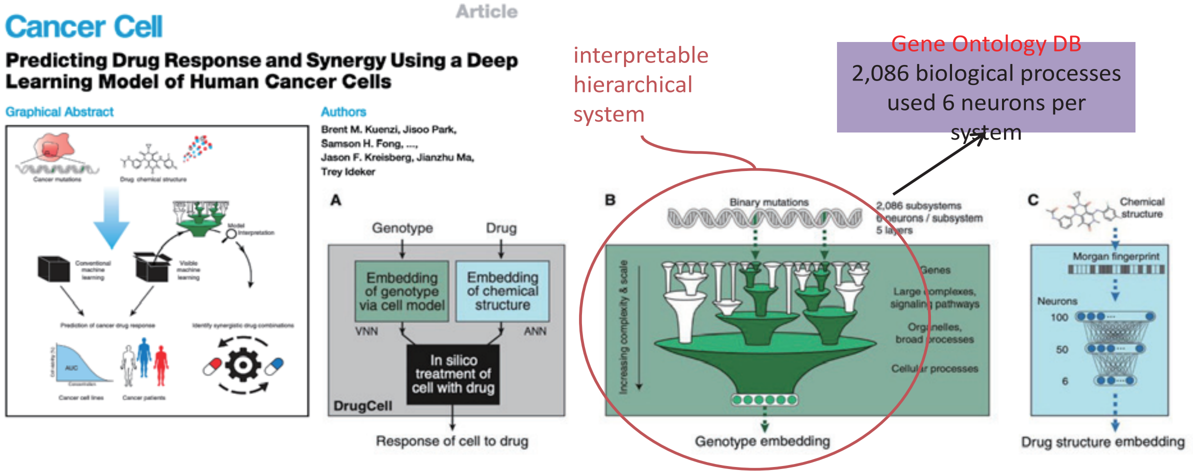


The workflow comprised rapid augmentation of knowledge graph information from recent literature using machine learning (ML) based extraction, with human-guided iterative queries of the graph. Using this workflow, we identified the rheumatoid arthritis drug baricitinib as both an antiviral and anti-inflammatory therapy. The effectiveness of baricitinib was substantiated by the recent publication of the data from the ACTT-2 randomised Phase 3 trial, followed by emergency approval for use by the FDA, and a report from the CoV-BARRIER trial confirming significant reductions in mortality with baricitinib compared to standard of care



DrugCell

- DrugCell is an interpretable deep learning model that simulates the response of human cancer cells to therapy.
- DrugCell predictions might generalize to patient tumors and can be used to design synergistic drug combinations that significantly improve treatment outcomes.



Cancer Cell

Article

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

Graphical Abstract

Authors

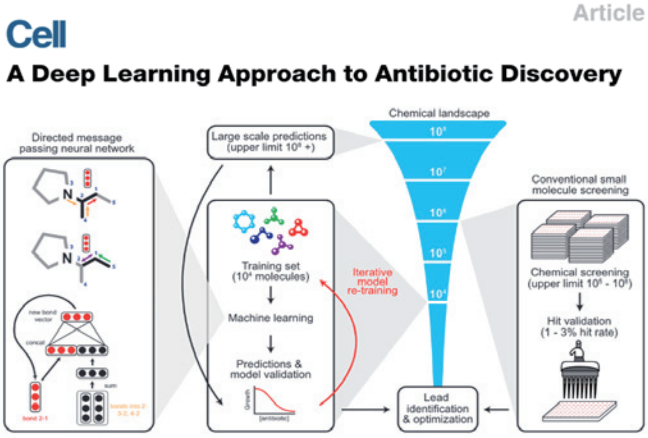
Brent M. Kuenzi, Jisoo Park,
Samson H. Fong, ...,
Jason F. Kreisberg, Jianzhu Ma,
Trey Ideker

interpretable hierarchical system

Gene Ontology DB
2,086 biological processes used 6 neurons per system

Kuenzi, Brent M., et al. "Predicting drug response and synergy using a deep learning model of human cancer cells." Cancer cell 38.5 (2020): 672-684.

Antibiotic discovery



Cell

Article

A Deep Learning Approach to Antibiotic Discovery

Directed message passing neural network
new bond vector
concat
sum
bond 2-1
bonds into 2: 3-2, 4-2
Large scale predictions (upper limit 10⁸ +)
Training set (10⁴ molecules)
Machine learning
Predictions & model validation
Growth
[antibiotic]
Iterative model re-training
Chemical landscape
10⁸
10⁷
10⁶
10⁵
10⁴
Lead identification & optimization
Conventional small molecule screening
Chemical screening (upper limit 10⁵ - 10⁶)
Hit validation (1 - 3% hit rate)

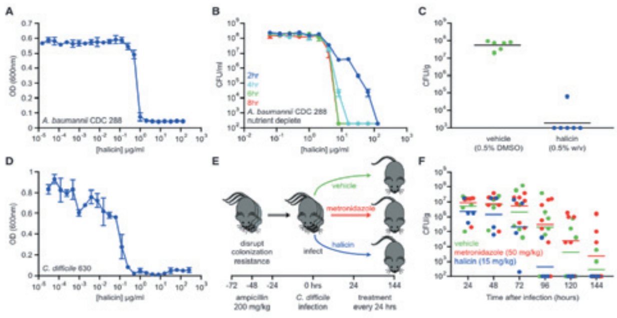
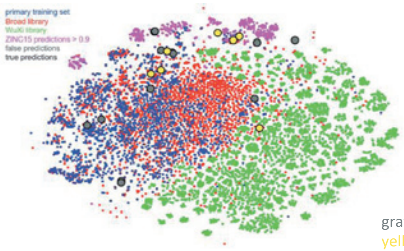


Figure 5. Halicin Displays Efficacy in Murine Models of Infection



gray: false positive predictions
yellow: true positive predictions

Stokes, Jonathan M., et al. "A deep learning approach to antibiotic discovery." Cell 180.4 (2020): 688-702.

Dsicovery of structurally divergent antibiotics

- Here, we demonstrate how the combination of *in silico* predictions and empirical investigations can lead to the discovery of new antibiotics.

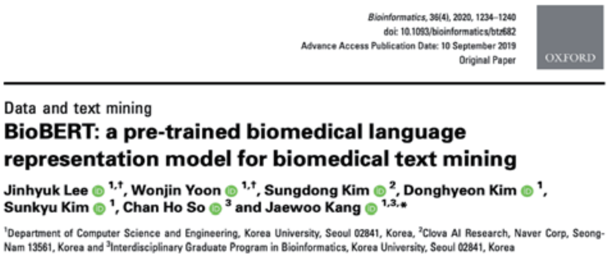
- First, we trained a deep neural network model to predict *growth inhibition of Escherichia coli using a collection of 2,335 molecules.*

- Second, we applied the resulting model to several *discrete chemical libraries, comprising >107 million molecules*, to identify potential lead compounds with activity against *E. coli*.

- After *ranking the compounds* according to the model's predicted score, we lastly selected a list of candidates based on a pre-specified prediction score threshold, chemical structure, and availability.

- Through this approach, from the Drug Repurposing Hub, we identified the c-Jun N-terminal kinase inhibitor SU3327 (De et al.,

Literature-based approaches



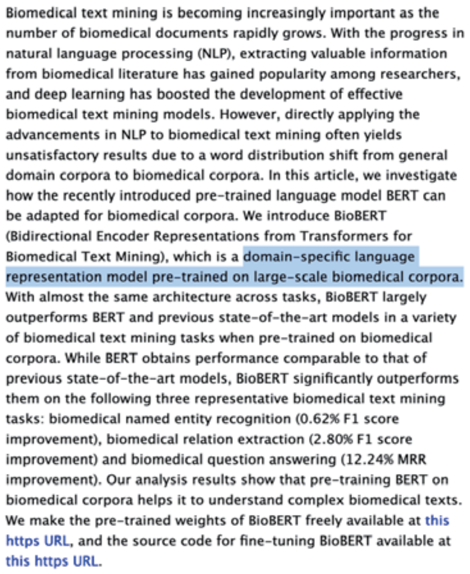
Bioinformatics, 36(4), 2020, 1234–1240
doi: 10.1093/bioinformatics/btz682
Advance Access Publication Date: 10 September 2019
Original Paper

Data and text mining

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

Jinhyuk Lee [1,†], Wonjin Yoon [1,†], Sungdong Kim [2], Donghyeon Kim [1], Sunkyu Kim [1], Chan Ho So [3] and Jaewoo Kang [1,3,*]

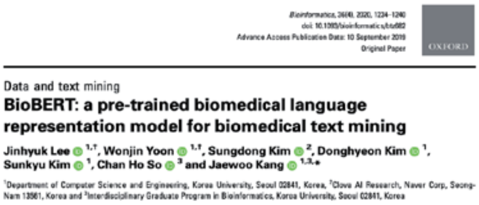
[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, Korea, [2]Clova AI Research, Naver Corp, Seong-Nam 13561, Korea and [3]Interdisciplinary Graduate Program in Bioinformatics, Korea University, Seoul 02841, Korea



Biomedical text mining is becoming increasingly important as the number of biomedical documents rapidly grows. With the progress in natural language processing (NLP), extracting valuable information from biomedical literature has gained popularity among researchers, and deep learning has boosted the development of effective biomedical text mining models. However, directly applying the advancements in NLP to biomedical text mining often yields unsatisfactory results due to a word distribution shift from general domain corpora to biomedical corpora. In this article, we investigate how the recently introduced pre-trained language model BERT can be adapted for biomedical corpora. We introduce BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining), which is a domain-specific language representation model pre-trained on large-scale biomedical corpora. With almost the same architecture across tasks, BioBERT largely outperforms BERT and previous state-of-the-art models in a variety of biomedical text mining tasks when pre-trained on biomedical corpora. While BERT obtains performance comparable to that of previous state-of-the-art models, BioBERT significantly outperforms them on the following three representative biomedical text mining tasks: biomedical named entity recognition (0.62% F1 score improvement), biomedical relation extraction (2.80% F1 score improvement) and biomedical question answering (12.24% MRR improvement). Our analysis results show that pre-training BERT on biomedical corpora helps it to understand complex biomedical texts. We make the pre-trained weights of BioBERT freely available at this https URL, and the source code for fine-tuning BioBERT available at this https URL.

Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.

Literature-based approaches



Bioinformatics, 36(4), 2020, 1234–1240
doi: 10.1093/bioinformatics/btz682
Advance Access Publication Date: 10 September 2019
Original Paper

OXFORD

Data and text mining

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

Jinhyuk Lee [1,†], Wonjin Yoon [1,†], Sungdong Kim [2], Donghyeon Kim [1], Sunkyu Kim [1], Chan Ho So [3] and Jaewoo Kang [1,3,*]

[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, Korea, [2]Clova AI Research, Naver Corp, Seong-Nam 13561, Korea and [3]Interdisciplinary Graduate Program in Bioinformatics, Korea University, Seoul 02841, Korea

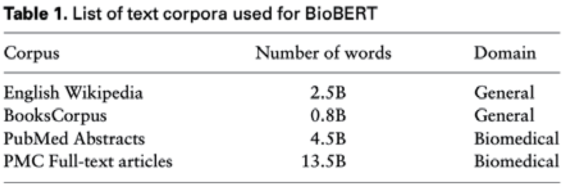
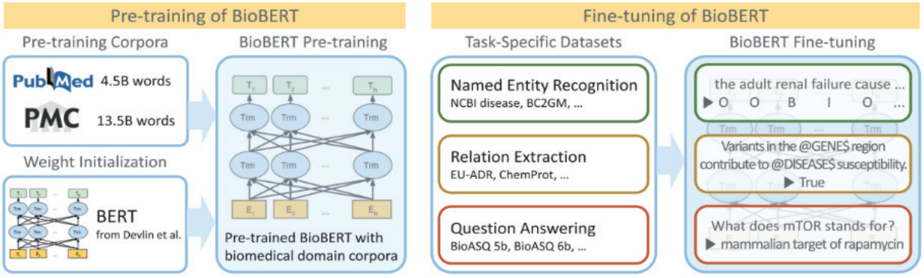


Table 1. List of text corpora used for BioBERT

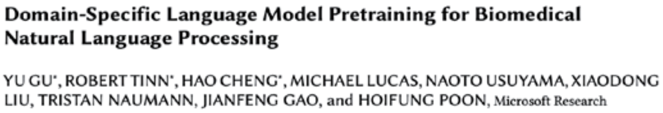
Corpus	Number of words	Domain
English Wikipedia	2.5B	General
BooksCorpus	0.8B	General
PubMed Abstracts	4.5B	Biomedical
PMC Full-text articles	13.5B	Biomedical



Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.

Literature-based approaches

PubMedBERT



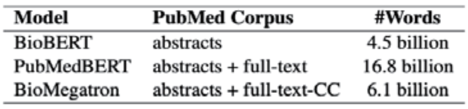
Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing

YU GU*, ROBERT TINN*, HAO CHENG*, MICHAEL LUCAS, NAOTO USUYAMA, XIAODONG LIU, TRISTAN NAUMANN, JIANFENG GAO, and HOIFUNG POON, Microsoft Research

BioMegatron

BioMegatron: Larger Biomedical Domain Language Model

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina,
Raul Puri, Mostofa Patwary, Mohammad Shoeybi, Raghav Mani
NVIDIA / Santa Clara, California, USA
hshin@nvidia.com



Model	PubMed Corpus	#Words
BioBERT	abstracts	4.5 billion
PubMedBERT	abstracts + full-text	16.8 billion
BioMegatron	abstracts + full-text-CC	6.1 billion

Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.

Gu, Yu, et al. "Domain-specific language model pretraining for biomedical natural language processing." ACM Transactions on Computing for Healthcare (HEALTH) 3.1 (2021): 1-23.

Shin, Hoo-Chang, et al. "BioMegatron: Larger biomedical domain language model." arXiv preprint arXiv:2010.06060 (2020).

Networks

Commonly used biological networks and disease networks

Protein-protein interaction network (PPI) – STRING, BioGRID
Biological pathways network – KEGG, Reactome
Disease networks – Diseasome, HDN, DGN
Comprehensive heterogeneous network – HetioNet, MSI

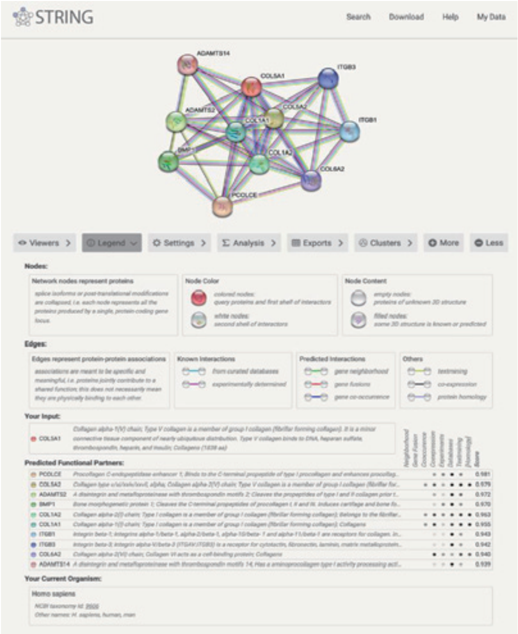
PPI Network - STRING

STRING

- Search Tool for the Retrieval of Interacting Genes/Proteins
- Integrates all publicly available sources of protein-protein interaction information.
    - Automated text mining
    - Interaction experiments
    - Computational interaction predictions from co-expression
- Statistics of latest version of STRING

Category	Count
Organisms	14,094
Proteins	67,592,464
Interactions	20,052,394,041

Browse COL5A1 protein in STRING



Szklarczyk, Damian et al. "The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets." Nucleic acids research vol. 49,D1 (2021)

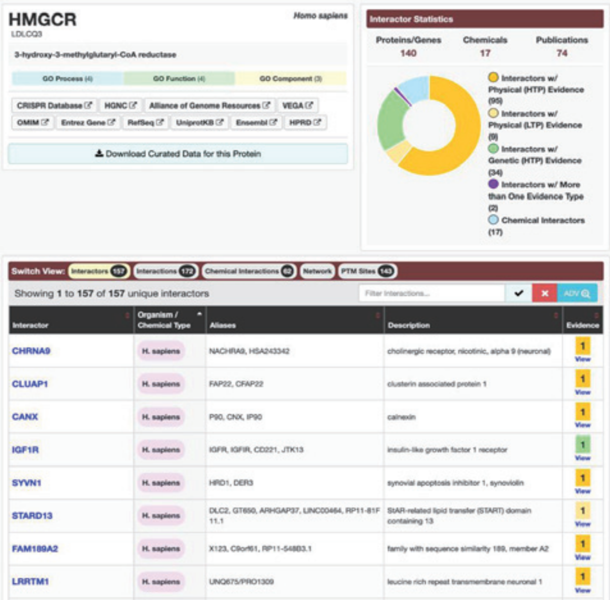
PPI Network - BioGRID

BioGRID

- Biological General Repository for Interaction Datasets
- Archives genetic and protein interaction data from various organisms.

Category	Count
Protein/Genetic interactions	2,551,504
Chemical interactions	29,417
Post translational modifications	1,128,339

Browse HMGCR protein in BioGRID



HMGCR
LDLCQ3
Homo sapiens
3-hydroxy-3-methylglutaryl-CoA reductase

GO Process (4)
GO Function (4)
GO Component (3)
CRISPR Database
HGNC
Alliance of Genome Resources
VEGA
OMIM
Entrez Gene
RefSeq
UniprotKB
Ensembl
HPRD
Download Curated Data for this Protein

Interactor Statistics

Proteins/Genes	Chemicals	Publications
140	17	74

- Interactors w/ Physical (HTP) Evidence (95)
- Interactors w/ Physical (LTP) Evidence (9)
- Interactors w/ Genetic (HTP) Evidence (34)
- Interactors w/ More than One Evidence Type (2)
- Chemical Interactors (17)

Switch View: Interactors 157 Interactions 172 Chemical Interactions 62 Network PTM Sites 143
Filter Interactions...
ADV

Showing 1 to 157 of 157 unique interactors

Interactor	Organism / Chemical Type	Aliases	Description	Evidence
CHRNA9	H. sapiens	NACHRA9, HSA243342	cholinergic receptor, nicotinic, alpha 9 (neuronal)	1 View
CLUAP1	H. sapiens	FAP22, CFAP22	clusterin associated protein 1	1 View
CANX	H. sapiens	P90, CNX, IP90	calnexin	1 View
IGF1R	H. sapiens	IGFR, IGFIR, CD221, JTK13	insulin-like growth factor 1 receptor	1 View
SYVN1	H. sapiens	HRD1, DER3	synovial apoptosis inhibitor 1, synoviolin	1 View
STARD13	H. sapiens	DLC2, GT650, ARHGAP37, LINC00464, RP11-81F 11.1	StAR-related lipid transfer (START) domain containing 13	1 View
FAM189A2	H. sapiens	X123, C9orf61, RP11-548B3.1	family with sequence similarity 189, member A2	1 View
LRRTM1	H. sapiens	UNQ675/PRO1309	leucine rich repeat transmembrane neuronal 1	1 View

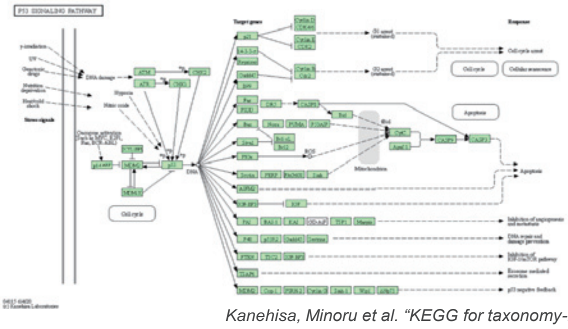
Oughtred, Rose et al. "The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions." Protein science : a publication of the Protein Society vol. 30,1 (2021)

Biological Pathways Network - KEGG

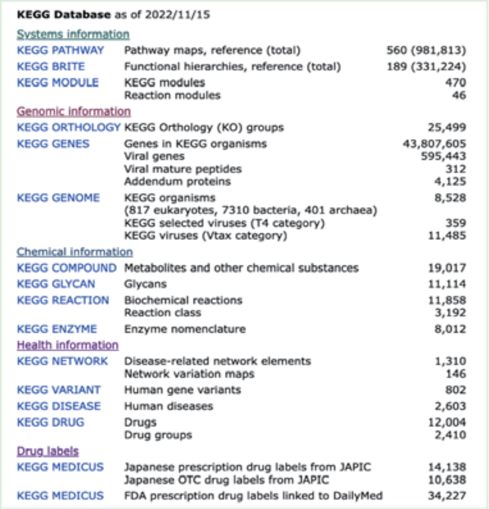
KEGG

- Kyoto Encyclopedia of Genes and Genomes
- A curated collection of biological information compiled from published material.
- Includes information on genes, proteins, metabolic pathways, molecular interactions, and biochemical reactions associated with specific organisms.
- Provides a relationship for how these components are organized in a cellular structure or reaction pathway.

p53 signaling pathway from KEGG



Statistics of KEGG



KEGG Database as of 2022/11/15

Systems information		
KEGG PATHWAY	Pathway maps, reference (total)	560 (981,813)
KEGG BRITE	Functional hierarchies, reference (total)	189 (331,224)
KEGG MODULE	KEGG modules	470
	Reaction modules	46
Genomic information		
KEGG ORTHOLOGY	KEGG Orthology (KO) groups	25,499
KEGG GENES	Genes in KEGG organisms	43,807,605
	Viral genes	595,443
	Viral mature peptides	312
	Addendum proteins	4,125
KEGG GENOME	KEGG organisms (817 eukaryotes, 7310 bacteria, 401 archaea)	8,528
	KEGG selected viruses (T4 category)	359
	KEGG viruses (Vtax category)	11,485
Chemical information		
KEGG COMPOUND	Metabolites and other chemical substances	19,017
KEGG GLYCAN	Glycans	11,114
KEGG REACTION	Biochemical reactions	11,858
	Reaction class	3,192
KEGG ENZYME	Enzyme nomenclature	8,012
Health information		
KEGG NETWORK	Disease-related network elements	1,310
	Network variation maps	146
KEGG VARIANT	Human gene variants	802
KEGG DISEASE	Human diseases	2,603
KEGG DRUG	Drugs	12,004
	Drug groups	2,410
Drug labels		
KEGG MEDICUS	Japanese prescription drug labels from JAPIC	14,138
	Japanese OTC drug labels from JAPIC	10,638
KEGG MEDICUS	FDA prescription drug labels linked to DailyMed	34,227

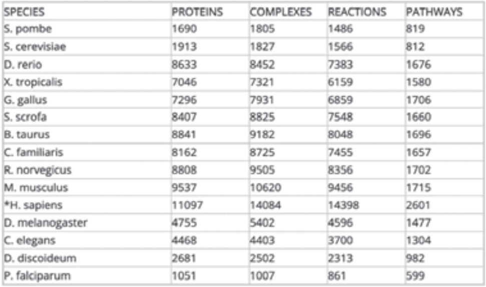
Kanehisa, Minoru et al. "KEGG for taxonomy-based analysis of pathways and genomes." Nucleic acids research, gkac963. 27 Oct. 2022

Biological Pathways Network - Reactome

Reactome

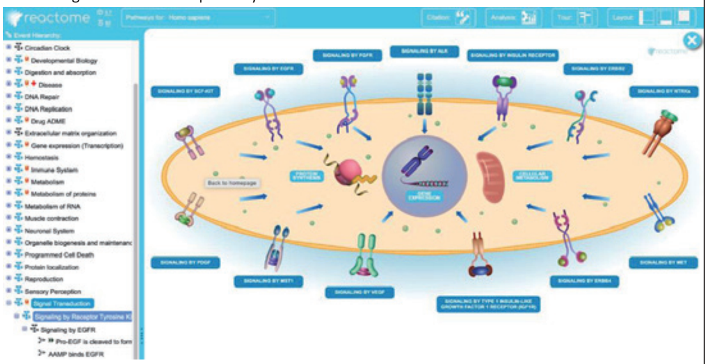
- Open source pathway database
- Curated human pathways encompassing metabolism, signaling, and other biological processes.
- Every pathway is traceable to primary literature.
- Cross-reference to many other bioinformatics databases.
- Provides data analysis and visualization tools.

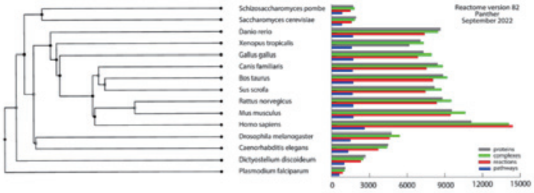
Statistics of Reactome



SPECIES	PROTEINS	COMPLEXES	REACTIONS	PATHWAYS
S. pombe	1690	1805	1486	819
S. cerevisiae	1913	1827	1566	812
D. rerio	8633	8452	7383	1676
X. tropicalis	7046	7321	6159	1580
G. gallus	7296	7931	6859	1706
S. scrofa	8407	8825	7548	1660
B. taurus	8841	9182	8048	1696
C. familiaris	8162	8725	7455	1657
R. norvegicus	8808	9505	8356	1702
M. musculus	9537	10620	9456	1715
*H. sapiens	11097	14084	14398	2601
D. melanogaster	4755	5402	4596	1477
C. elegans	4468	4403	3700	1304
D. discoideum	2681	2502	2313	982
P. falciparum	1051	1007	861	599

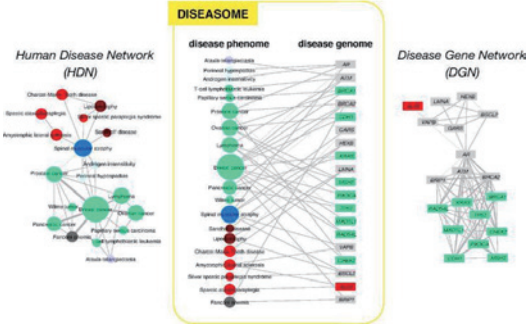
Browse Signal Transduction pathway in Reactome

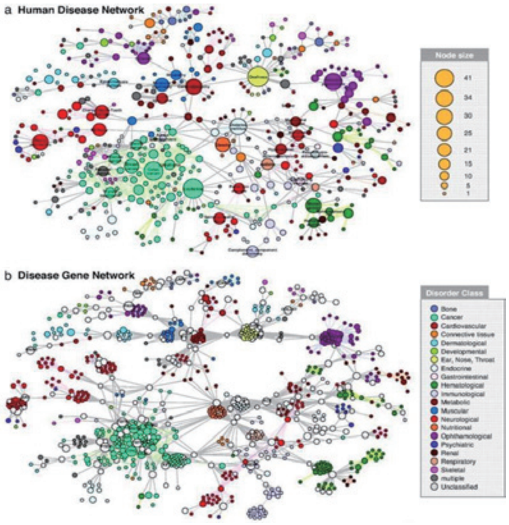




Gillespie, Marc et al. "The reactome pathway knowledgebase 2022." Nucleic acids research vol. 50,D1 (2022)

Disease Networks – Diseasome, HDN and DGN





Diseasome

- A small subset of OMIM-based disease gene association.

HDN: Human Disease Network

- Projection of the diseasome bipartite graph.
- Two diseases are connected if there is a gene that is implicated in both.

DGN: Disease Gene Network

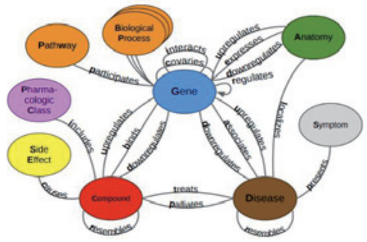
- Two genes are connected if they are involved in the same disease.

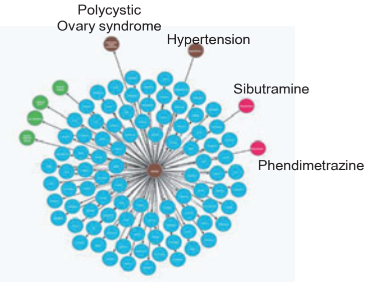
Goh, Kwang-Il, et al. "The human disease network." Proceedings of the National Academy of Sciences 104.21 (2007): 8685-8690.

Comprehensive Heterogeneous Networks - HetioNet

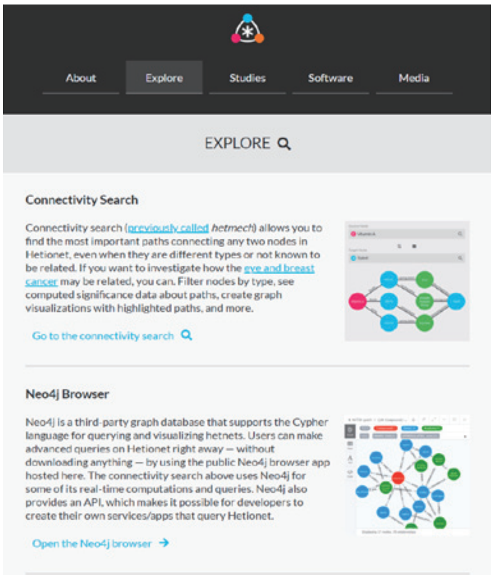
HetioNet

- An integrative network encoding knowledge from millions of biomedical studies.
- Data were integrated from 29 public resources to connect meta-nodes.
- Meta nodes (11 types): anatomy, biological process, cellular component, compound, disease, gene, molecular function, pathway, pharmacologic class, side effect, symptom
- Meta edges (24 types)





HetionNet Web Interface



Connectivity Search

Connectivity search (previously called *hetmech*) allows you to find the most important paths connecting any two nodes in Hetionet, even when they are different types or not known to be related. If you want to investigate how the eye and breast cancer may be related, you can. Filter nodes by type, see computed significance data about paths, create graph visualizations with highlighted paths, and more.

Go to the connectivity search

Neo4j Browser

Neo4j is a third-party graph database that supports the Cypher language for querying and visualizing hetnets. Users can make advanced queries on Hetionet right away — without downloading anything — by using the public Neo4j browser app hosted here. The connectivity search above uses Neo4j for some of its real-time computations and queries. Neo4j also provides an API, which makes it possible for developers to create their own services/apps that query Hetionet.

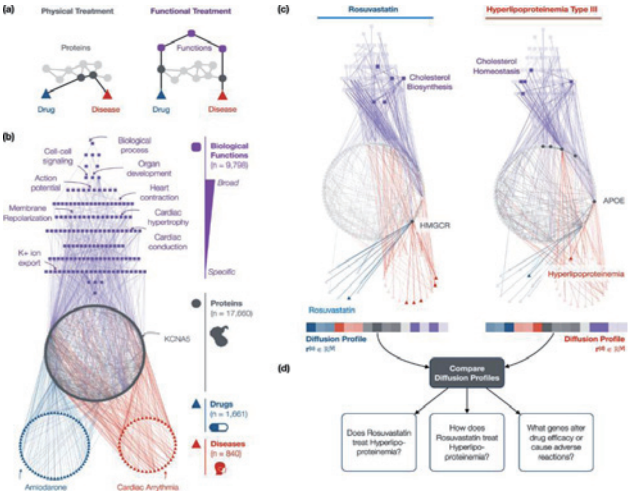
Open the Neo4j browser →

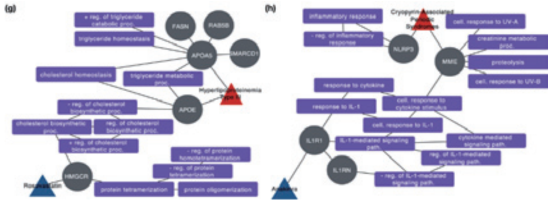
Himmelstein, Daniel Scott et al. "Systematic integration of biomedical knowledge prioritizes drugs for repurposing." eLife vol. 6 e26726. 22 Sep. 2017

Comprehensive Heterogeneous Networks - MSI

MSI

- Multiscale Interactome network
- An integrative network of disease, proteins, biological functions and drugs.
- Data were retrieved from 19 public databases.
- Random walk-based method can be applied to capture the effects of drugs through a hierarchy of biological functions and protein-protein interactions.




(g)
+ reg. of triglyceride catabolic proc.
triglyceride homeostasis
FASN
RAB5B
APOA5
SMARCD1
cholesterol homeostasis
triglyceride metabolic proc.
Hyperlipoproteinemia Type III
- reg. of cholesterol biosynthetic proc.
APOE
cholesterol biosynthetic proc.
reg. of cholesterol biosynthetic proc.
+ reg. of cholesterol biosynthetic proc.
- reg. of protein homotetramerization
- reg. of protein tetramerization
HMGCR
Rosuvastatin
protein tetramerization
protein oligomerization
(h)
inflammatory response
Cryopyrin-Associated Periodic Syndromes
cell. response to UV-A
- reg. of inflammatory response
NLRP3
creatinine metabolic proc.
MME
proteolysis
cell. response to UV-B
response to cytokine
response to IL-1
cell. response to cytokine stimulus
cell. response to IL-1
IL1R1
IL-1-mediated signaling path.
cytokine mediated signaling path.
reg. of IL-1-mediated signaling path.
IL1RN
- reg. of IL-1-mediated signaling path.
Anakinra

Ruiz, C., Zitnik, M. & Leskovec, J. Identification of disease treatment mechanisms through the multiscale interactome. Nat Commun 12, 1796 (2021)

Databases

Commonly used databases for Drug repositioning

Drug Repurposing Hub
repoDB
CTD
PharmacoDB

Database Overview (graph view)

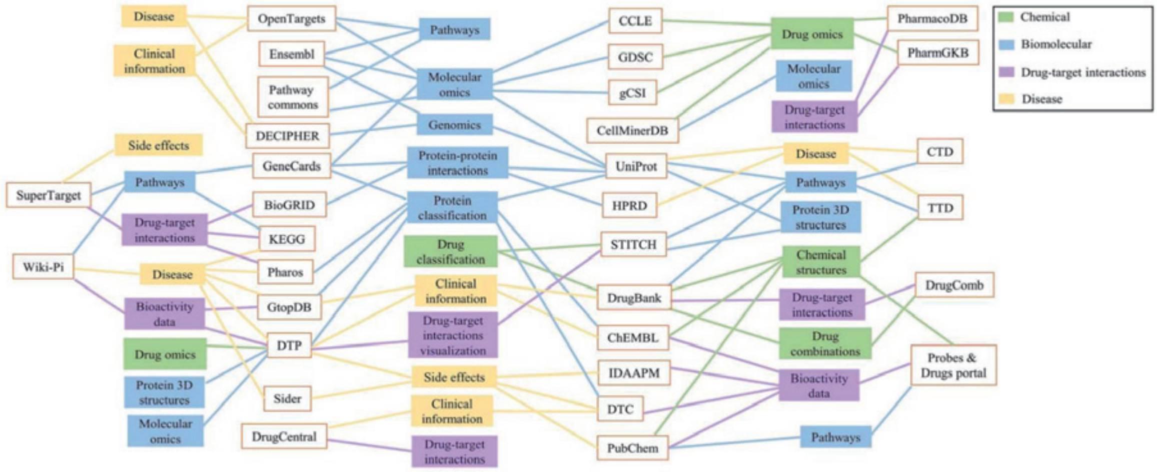


Figure 1. Drug repositioning databases categorized into more than one subcategory. Some subcategories are shown more than once in order to facilitate the interpretation of database relationships.

Tanoli, Ziaurrehman, et al. "Exploration of databases and methods supporting drug repurposing: a comprehensive survey." Briefings in bioinformatics 22.2 (2021): 1656-1678.

Database Overview (table view)

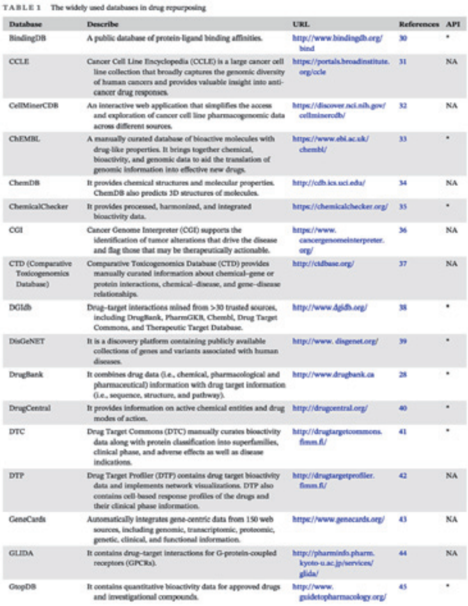


TABLE 1 The widely used databases in drug repurposing

Database	Describe	URL	References	API
BindingDB	A public database of protein-ligand binding affinities.	http://www.bindingdb.org/bind	30	*
CCLE	Cancer Cell Line Encyclopedia (CCLE) is a large cancer cell line collection that broadly captures the genomic diversity of human cancers and provides valuable insight into anti-cancer drug responses.	https://portals.broadinstitute.org/ccle	31	NA
CellMinerCDB	An interactive web application that simplifies the access and exploration of cancer cell line pharmacogenomic data across different sources.	https://discover.nci.nih.gov/cellminercdb/	32	NA
ChEMBL	A manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity, and genomic data to aid the translation of genomic information into effective new drugs.	https://www.ebi.ac.uk/chembl/	33	*
ChemDB	It provides chemical structures and molecular properties. ChemDB also predicts 3D structures of molecules.	http://cdb.ics.uci.edu/	34	NA
ChemicalChecker	It provides processed, harmonized, and integrated bioactivity data.	https://chemicalchecker.org/	35	*
CGI	Cancer Genome Interpreter (CGI) supports the identification of tumor alterations that drive the disease and flag those that may be therapeutically actionable.	https://www.cancergenomeinterpreter.org/	36	NA
CTD (Comparative Toxicogenomics Database)	Comparative Toxicogenomics Database (CTD) provides manually curated information about chemical–gene or protein interactions, chemical–disease, and gene–disease relationships.	http://ctdbase.org/	37	NA
DGIdb	Drug–target interactions mined from >30 trusted sources, including DrugBank, PharmGKB, Chembl, Drug Target Commons, and Therapeutic Target Database.	http://www.dgidb.org/	38	*
DisGeNET	It is a discovery platform containing publicly available collections of genes and variants associated with human diseases.	http://www. disgenet.org/	39	*
DrugBank	It combines drug data (i.e., chemical, pharmacological and pharmaceutical) information with drug target information (i.e., sequence, structure, and pathway).	http://www.drugbank.ca	28	*
DrugCentral	It provides information on active chemical entities and drug modes of action.	http://drugcentral.org/	40	*
DTC	Drug Target Commons (DTC) manually curates bioactivity data along with protein classification into superfamilies, clinical phase, and adverse effects as well as disease indications.	http://drugtargetcommons.fimm.fi/	41	*
DTP	Drug Target Profiler (DTP) contains drug target bioactivity data and implements network visualizations. DTP also contains cell-based response profiles of the drugs and their clinical phase information.	http://drugtargetprofiler.fimm.fi/	42	NA
GeneCards	Automatically integrates gene-centric data from 150 web sources, including genomic, transcriptomic, proteomic, genetic, clinical, and functional information.	https://www.genecards.org/	43	NA
GLIDA	It contains drug–target interactions for G-protein-coupled receptors (GPCRs).	http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/	44	NA
GtopDB	It contains quantitative bioactivity data for approved drugs and investigational compounds.	http://www.guidetopharmacology.org/	45	*

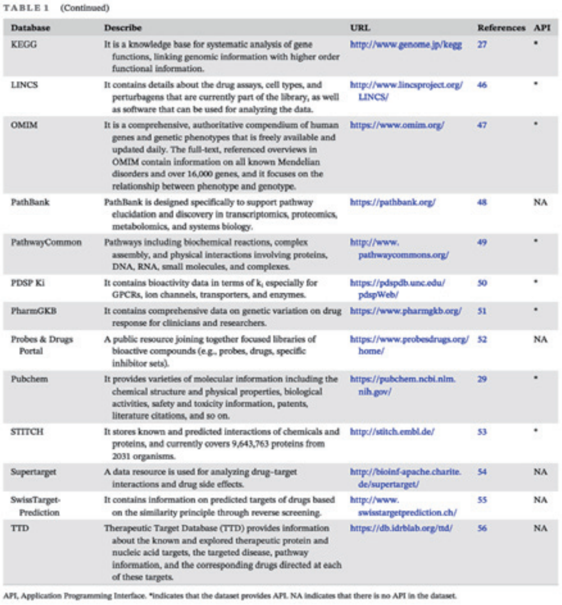


TABLE 1 (Continued)

Database	Describe	URL	References	API
KEGG	It is a knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information.	http://www.genome.jp/kegg	27	*
LINCS	It contains details about the drug assays, cell types, and perturbagens that are currently part of the library, as well as software that can be used for analyzing the data.	http://www.lincsproject.org/LINCS/	46	*
OMIM	It is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. The full-text, referenced overviews in OMIM contain information on all known Mendelian disorders and over 16,000 genes, and it focuses on the relationship between phenotype and genotype.	https://www.omim.org/	47	*
PathBank	PathBank is designed specifically to support pathway elucidation and discovery in transcriptomics, proteomics, metabolomics, and systems biology.	https://pathbank.org/	48	NA
PathwayCommon	Pathways including biochemical reactions, complex assembly, and physical interactions involving proteins, DNA, RNA, small molecules, and complexes.	http://www.pathwaycommons.org/	49	*
PDSP Ki	It contains bioactivity data in terms of k_i especially for GPCRs, ion channels, transporters, and enzymes.	https://pdspdb.unc.edu/pdspWeb/	50	*
PharmGKB	It contains comprehensive data on genetic variation on drug response for clinicians and researchers.	https://www.pharmgkb.org/	51	*
Probes & Drugs Portal	A public resource joining together focused libraries of bioactive compounds (e.g., probes, drugs, specific inhibitor sets).	https://www.probesdrugs.org/home/	52	NA
Pubchem	It provides varieties of molecular information including the chemical structure and physical properties, biological activities, safety and toxicity information, patents, literature citations, and so on.	https://pubchem.ncbi.nlm.nih.gov/	29	*
STITCH	It stores known and predicted interactions of chemicals and proteins, and currently covers 9,643,763 proteins from 2031 organisms.	http://stitch.embl.de/	53	*
Supertarget	A data resource is used for analyzing drug–target interactions and drug side effects.	http://bioinf-apache.charite.de/supertarget/	54	NA
SwissTarget-Prediction	It contains information on predicted targets of drugs based on the similarity principle through reverse screening.	http://www.swisstargetprediction.ch/	55	NA
TTD	Therapeutic Target Database (TTD) provides information about the known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information, and the corresponding drugs directed at each of these targets.	https://db.idrblab.org/ttd/	56	NA

API, Application Programming Interface. *indicates that the dataset provides API. NA indicates that there is no API in the dataset.

Pan, Xiaoqin, et al. "Deep learning for drug repurposing: Methods, databases, and applications." Wiley Interdisciplinary Reviews: Computational Molecular Science (2022)

Databases: Drug Repurposing Hub

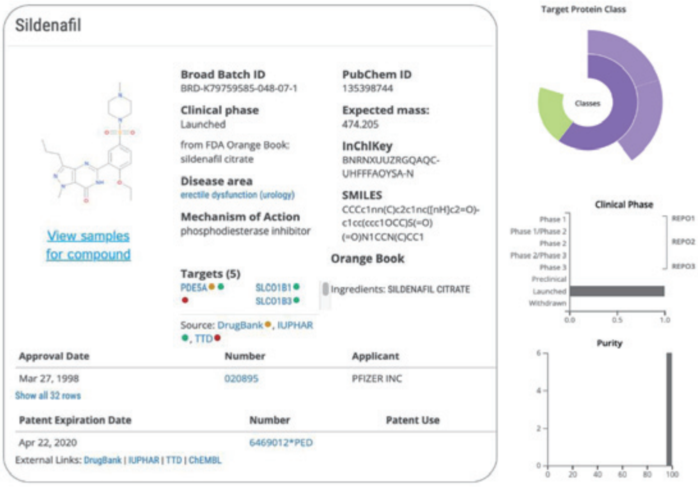
Drug Repurposing Hub

- A curated and annotated collection of FDA-approved drugs, clinical trial drugs, and pre-clinical tool compounds with a companion information resource.
- Hand-curated collection of compounds were experimentally confirmed and annotated with literature-reported targets.
- Each drug information includes compound name, clinical phase, mechanism of action, and protein target.

Statistics of Drug Repurposing Hub

Category	Count
Total samples	16,826
Protein targets	2,183
Unique compounds	7,934
Drug indications	670

Browse Sildenafil in Drug Repurposing Hub Web app



Sildenafil

View samples for compound

Broad Batch ID
BRD-K79759585-048-07-1

Clinical phase
Launched

from FDA Orange Book:
sildenafil citrate

Disease area
erectile dysfunction (urology)

Mechanism of Action
phosphodiesterase inhibitor

Targets (5)
PDE5A ● ● ● SLCO1B1 ● SLCO1B3 ●

Source: DrugBank ●, IUPHAR ●, TTD ●

PubChem ID
135398744

Expected mass:
474.205

InChIKey
BNRNXUUZRGQAQC-UHFFFAOYSA-N

SMILES
CCCc1nn(C)c2c1nc([nH]c2=O)-c1cc(ccc1OCC)S(=O)(=O)N1CCN(C)CC1

Orange Book

Ingredients: SILDENAFIL CITRATE

Approval Date	Number	Applicant
Mar 27, 1998	020895	PFIZER INC

Show all 32 rows

Patent Expiration Date	Number	Patent Use
Apr 22, 2020	6469012*PED	

External Links: DrugBank \| IUPHAR \| TTD \| ChEMBL

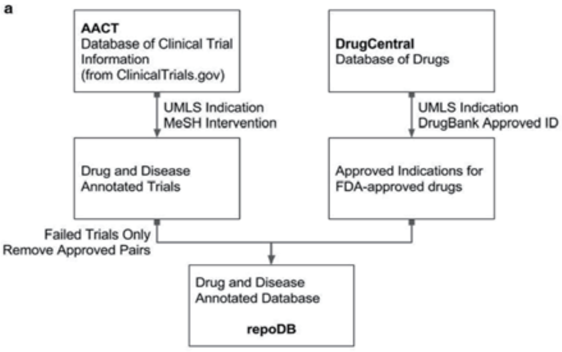
Corsello, Steven M et al. "The Drug Repurposing Hub: a next-generation drug library and information resource." Nature medicine vol. 23,4 (2017)

Databases: repoDB

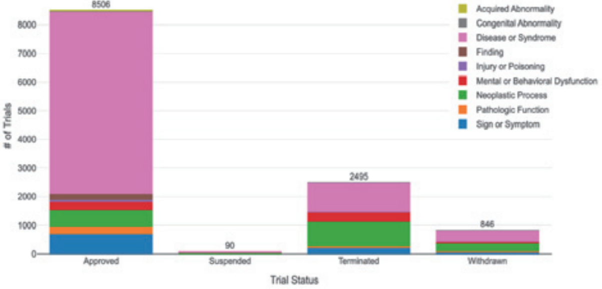
repoDB Drug Repositioning Database

repoDB

- A standard set of drug repositioning <u>successes and failures</u> that can be used to fairly and reproducibly benchmark computational repositioning methods.
- Data were extracted from DrugCentral and ClinicalTrials.gov.
- Each drug information includes compound name, <u>clinical phase</u> and disease name.


a
AACT
Database of Clinical Trial Information
(from ClinicalTrials.gov)
DrugCentral
Database of Drugs
UMLS Indication
MeSH Intervention
UMLS Indication
DrugBank Approved ID
Drug and Disease Annotated Trials
Approved Indications for FDA-approved drugs
Failed Trials Only
Remove Approved Pairs
Drug and Disease Annotated Database
repoDB

Statistics of repoDB



Category (status)	Drug count
Approved	2,162
Suspended	78
Terminated	518
Withdrawn	336

Brown, A., Patel, C. A standard database for drug repositioning. Sci Data 4, 170029 (2017)

Databases: CTD

ctd

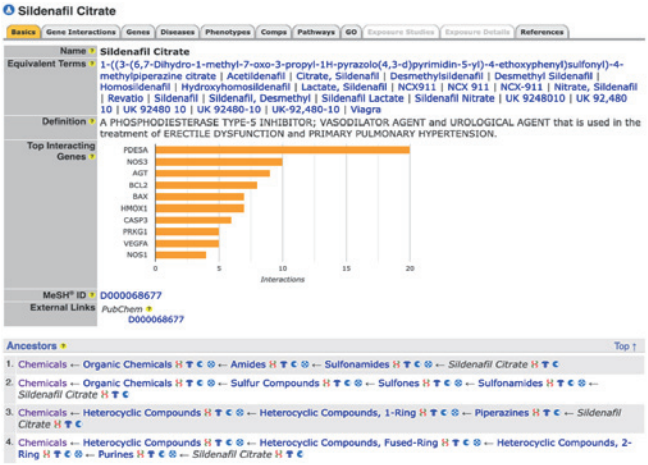
Chemicals
Pathways
Diseases
Exposures
Organisms
Genes
GO & Phenotypes
Anatomy
References

Curated Exposure Statements	204,467
Unique Chemicals	1,500
Unique Genes	1,084
Unique Diseases	488
Unique GO Terms	484
Curated Exposure References	3,300

CTD

- Comparative Toxicogenomics Database
- Provides manually curated information about <u>chemical-gene or protein interactions, chemical-disease, and gene-disease relationships</u>.
- Recent version of CTD offers a <u>CTD Tetramer tool</u> that generates potential molecular mechanistic pathways.

Browse Sildenafil in CTD Web app



Sildenafil Citrate

Name: Sildenafil Citrate

Equivalent Terms: 1-((3-(6,7-Dihydro-1-methyl-7-oxo-3-propyl-1H-pyrazolo(4,3-d)pyrimidin-5-yl)-4-ethoxyphenyl)sulfonyl)-4-methylpiperazine citrate | Acetildenafil | Citrate, Sildenafil | Desmethylsildenafil | Desmethyl Sildenafil | Homosildenafil | Hydroxyhomosildenafil | Lactate, Sildenafil | NCX911 | NCX 911 | NCX-911 | Nitrate, Sildenafil | Revatio | Sildenafil | Sildenafil, Desmethyl | Sildenafil Lactate | Sildenafil Nitrate | UK 9248010 | UK 92,480 10 | UK 92480 10 | UK 92480-10 | UK-92,480-10 | Viagra

Definition: A PHOSPHODIESTERASE TYPE-5 INHIBITOR; VASODILATOR AGENT and UROLOGICAL AGENT that is used in the treatment of ERECTILE DYSFUNCTION and PRIMARY PULMONARY HYPERTENSION.

Top Interacting Genes

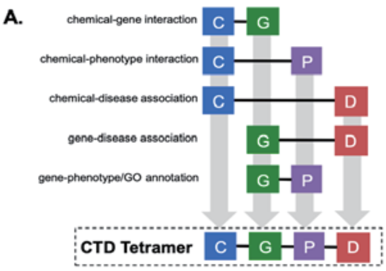
MeSH® ID: D000068677

External Links: PubChem D000068677

Ancestors

1. Chemicals ← Organic Chemicals ← Amides ← Sulfonamides ← *Sildenafil Citrate*
2. Chemicals ← Organic Chemicals ← Sulfur Compounds ← Sulfones ← Sulfonamides ← *Sildenafil Citrate*
3. Chemicals ← Heterocyclic Compounds ← Heterocyclic Compounds, 1-Ring ← Piperazines ← *Sildenafil Citrate*
4. Chemicals ← Heterocyclic Compounds ← Heterocyclic Compounds, Fused-Ring ← Heterocyclic Compounds, 2-Ring ← Purines ← *Sildenafil Citrate*

CTD Tetramer tool


A.
chemical-gene interaction
chemical-phenotype interaction
chemical-disease association
gene-disease association
gene-phenotype/GO annotation
CTD Tetramer
C
G
P
D

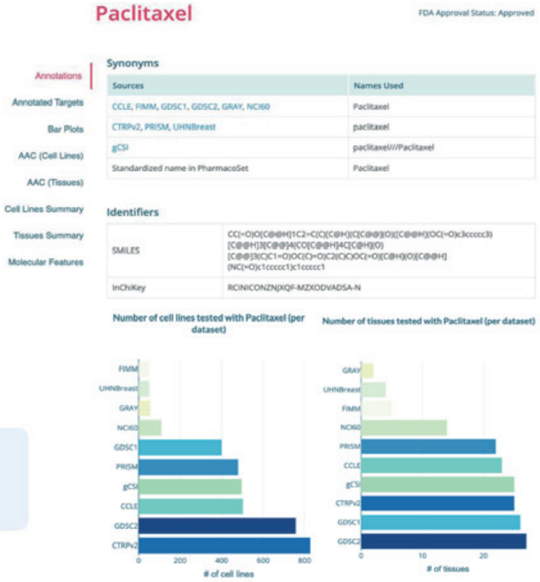
Davis, Allan Peter et al. "Comparative Toxicogenomics Database (CTD): update 2023." Nucleic acids research, gkac833. 28 Sep. 2022

Databases: PharmacoDB

PharmacoDB

- A web-application database that integrates multiple cancer pharmacogenomics datasets profiling approved and investigational drugs across cell lines from diverse tissue types.
- Offers a standardized cell line, drug identifiers and data format for drug sensitivity measurements.
- Included cell line data from..
  - CCLE, CTRPv2, FIMM, GDSC1, GDSC2, GRAY, NCI60, PRISM, UHNBreast, gCSI

Browse Paclitaxel in PharmacoDB Web app



Paclitaxel

FDA Approval Status: Approved

Synonyms

Sources	Names Used
CCLE, FIMM, GDSC1, GDSC2, GRAY, NCI60	Paclitaxel
CTRPv2, PRISM, UHNBreast	paclitaxel
gCSI	paclitaxel///Paclitaxel
Standardized name in PharmacoSet	Paclitaxel

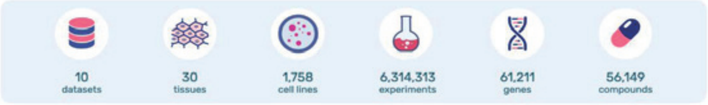
Identifiers

SMILES	CC(=O)O[C@@H]1C2=C(C)[C@H](C[C@@](O)([C@@H](OC(=O)c3ccccc3)[C@@H]3[C@@]4(CO[C@@H]4C[C@H](O)[C@@]3(C)C1=O)OC(C)=O)C2(C)C)OC(=O)[C@H](O)[C@@H](NC(=O)c1ccccc1)c1ccccc1
InChiKey	RCINICONZNJXQF-MZXODVADSA-N

Number of cell lines tested with Paclitaxel (per dataset)

Number of tissues tested with Paclitaxel (per dataset)

Statistics of PharmacoDB



Feizi, Nikta et al. "PharmacoDB 2.0: improving scalability and transparency of in vitro pharmacogenomics analysis." Nucleic acids research vol. 50,D1 (2022)

Technology

Network analysis technologies

Network analysis technologies

Analytical algorithms describing human gene networks have been developed for three major tasks in disease research:

1. Disease gene prioritization,
2. Disease module discovery, and
3. Stratification of complex diseases.

Network-based Drug Repurposing Technologies

SNF-cVAE (Knowledge-Based Systems, 2021)
CBPred (Cells, 2019)
DeepDR (Bioinformatics, 2019)
BiFusion (ISMB 2020)
Semantic Teleport (in revision)

The Main Issue for Network-based Drug Repurposing

Discover drug-disease relationship using
- Drug network
- Gene network
- Disease network

Hetionet database:

- drug-drug network: 1552 nodes, 6,486 edges
- disease-disease network: 137 nodes, 543 edges
- **gene-gene network: 20,945 nodes, ~200,000 edges**

- drug-gene edges: ~50,000
- diseass-gene edges: ~30,000

Major Issues for Drug Repurposing

- There are multiple ways to learn embedding vectors for drug
  - Drug-centered embeddings from Drug-drug, Drug-target, Drug-disease.
  - Then, **how to combine different views on drugs**?

- Three-way relationship among drug-gene-disease cannot be learned at once.
- In the end, we need to deduce **drug-disease binary relationship**.
  - Basically, binary relationships are somehow combined on different layers, hierarchically.

Network-based Drug Repurposing Technologies

SNF-cVAE (Knowledge-Based Systems, 2021)
CBPred (Cells, 2019)
DeepDR (Bioinformatics, 2019)
BiFusion (ISMB 2020)
Semantic Teleport (BioRxiv. In review)

Network-based Drug Repurposing: Cases

Knowledge-Based Systems 212 (2021) 106585

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

SNF–CVAE: Computational method to predict drug–disease interactions using similarity network fusion and collective variational autoencoder

Tamer N. Jarada [a], Jon G. Rokne [a], Reda Alhajj [a,b,c,*]

[a] Department of Computer Science, University of Calgary, Calgary, Alberta, Canada
[b] Department of Computer Engineering, Istanbul Medipol University, Istanbul, Turkey
[c] Department of Health Informatics, University of Southern Denmark, Odense, Denmark

Jarada, Tamer N., Jon G. Rokne, and Reda Alhajj. "SNF–CVAE: computational method to predict drug–disease interactions using similarity network fusion and collective variational autoencoder." Knowledge-Based Systems 212 (2021): 106585.

Network-based Drug Repurposing: Cases

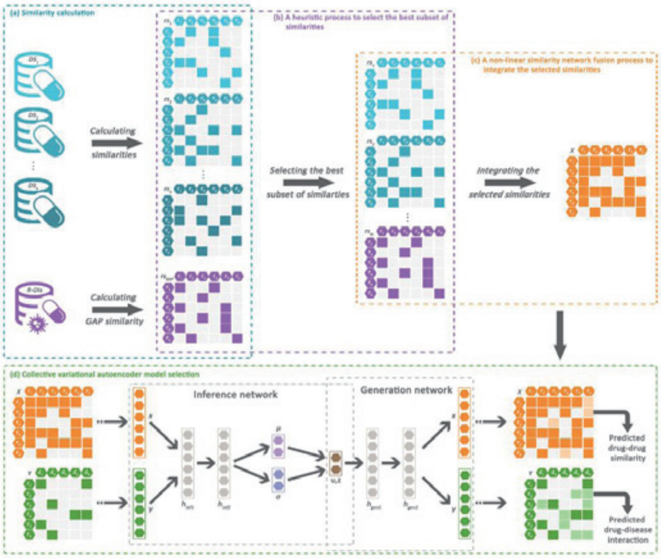
SNF-CVAE

- Input:
  - Drug-related similarity information
  - Drug-disease interactions
- Method:
  - **Similarity network fusion (SNF)**
    - Drug similarity network using drug-related data sets and drug-disease interaction dataset.
  - **Collective variational autoencoder (CVAE)**
    - Training cVAE with drug similarity (from above) and drug-disease interaction.
- Predicted drug candidates for potentially treating Alzheimer’s disease and Juvenile rheumatoid arthritis.

Jarada, Tamer N., Jon G. Rokne, and Reda Alhajj. "SNF–CVAE: computational method to predict drug–disease interactions using similarity network fusion and collective variational autoencoder." Knowledge-Based Systems 212 (2021): 106585.

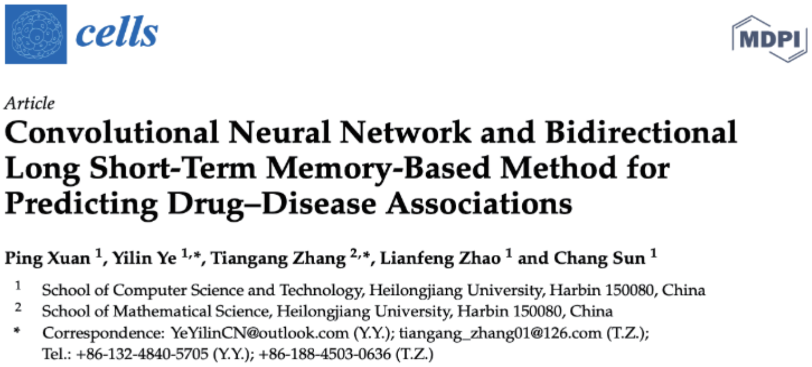
Network-based Drug Repurposing: Cases

SNF-CVAE



Jarada, Tamer N., Jon G. Rokne, and Reda Alhajj. "SNF–CVAE: computational method to predict drug–disease interactions using similarity network fusion and collective variational autoencoder." Knowledge-Based Systems 212 (2021): 106585.

Network-based Drug Repurposing: Cases



cells

MDPI

Article

Convolutional Neural Network and Bidirectional Long Short-Term Memory-Based Method for Predicting Drug–Disease Associations

Ping Xuan [1], Yilin Ye [1,*], Tiangang Zhang [2,*], Lianfeng Zhao [1] and Chang Sun [1]

[1] School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China

[2] School of Mathematical Science, Heilongjiang University, Harbin 150080, China

[*] Correspondence: YeYilinCN@outlook.com (Y.Y.); tiangang_zhang01@126.com (T.Z.); Tel.: +86-132-4840-5705 (Y.Y.); +86-188-4503-0636 (T.Z.)

Xuan, Ping, et al. "Convolutional neural network and bidirectional long short-term memory-based method for predicting drug–disease associations." Cells 8.7 (2019): 705.

Network-based Drug Repurposing: Cases

CBPred

- Input:
  - Drug similarity matrix (fingerprint-based)
  - Disease similarity matrix (MeSH-based)
- Goal:
  - Enrich paths between drugs and diseases
- Method:
  - **Convolutional Neural Network (CNN)**
    - Learn the association representation of drug-disease pairs from their similarities and associations.
  - **Bidirectional LSTM (BiLSTM)**
    - Learns path representation of drug-disease pair.
- Provided a list of novel drug-disease associations for drug repositioning

Xuan, Ping, et al. "Convolutional neural network and bidirectional long short-term memory-based method for predicting drug–disease associations." Cells 8.7 (2019): 705.

Network-based Drug Repurposing: Cases

CBPred

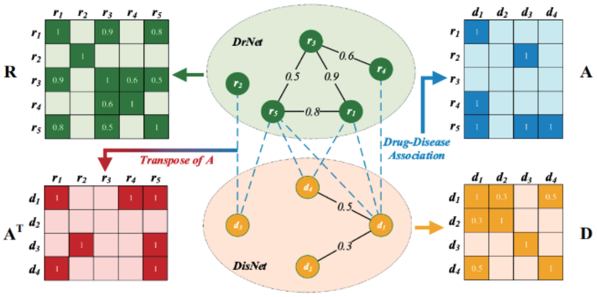


Figure 1. Construction of drug-disease heterogeneous network DrDisNet. *R* and *D* are the similarity matrix of drugs and diseases, respectively. *A* is the association matrix between drugs and diseases, while A^T is the transpose of *A*.

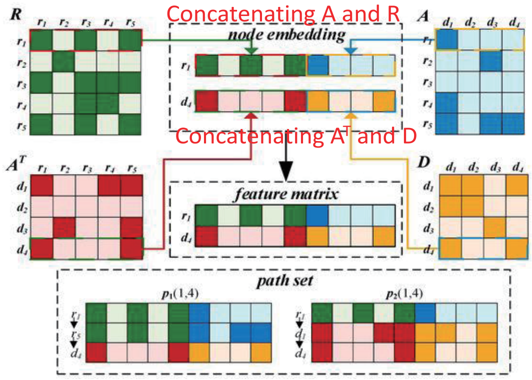
R and D are easily constructed
by comparing rows and colums as vectors.

A is from prior knowledge.

Xuan, Ping, et al. "Convolutional neural network and bidirectional long short-term memory-based method for predicting drug–disease associations." Cells 8.7 (2019): 705.

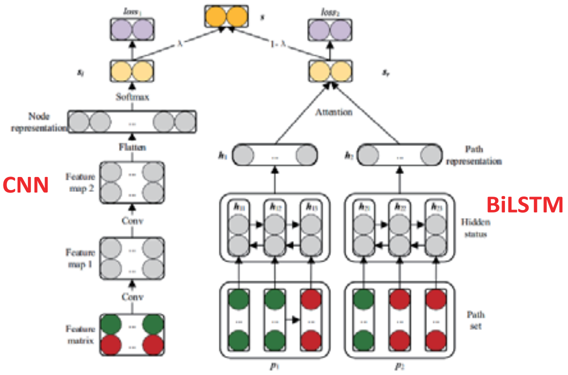
Network-based Drug Repurposing: Cases

CBPred



Concatenating A and R

Concatenating A^T and D



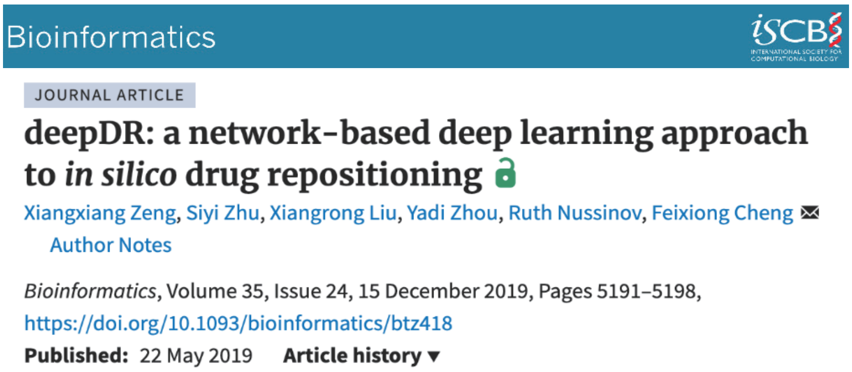
CNN

BiLSTM

Figure 2. Construction of the framework based on the convolutional neural network and bidirectional long short-term memory for learning the original and path representations.

Xuan, Ping, et al. "Convolutional neural network and bidirectional long short-term memory-based method for predicting drug–disease associations." Cells 8.7 (2019): 705.

Network-based Drug Repurposing: DeepDR



Bioinformatics

JOURNAL ARTICLE

deepDR: a network-based deep learning approach to *in silico* drug repositioning

Xiangxiang Zeng, Siyi Zhu, Xiangrong Liu, Yadi Zhou, Ruth Nussinov, Feixiong Cheng

Author Notes

Bioinformatics, Volume 35, Issue 24, 15 December 2019, Pages 5191–5198,

https://doi.org/10.1093/bioinformatics/btz418

Published: 22 May 2019 **Article history ▾**

Zeng, Xiangxiang, et al. "deepDR: a network-based deep learning approach to in silico drug repositioning." Bioinformatics 35.24 (2019): 5191-5198.

Network-based Drug Repurposing: DeepDR

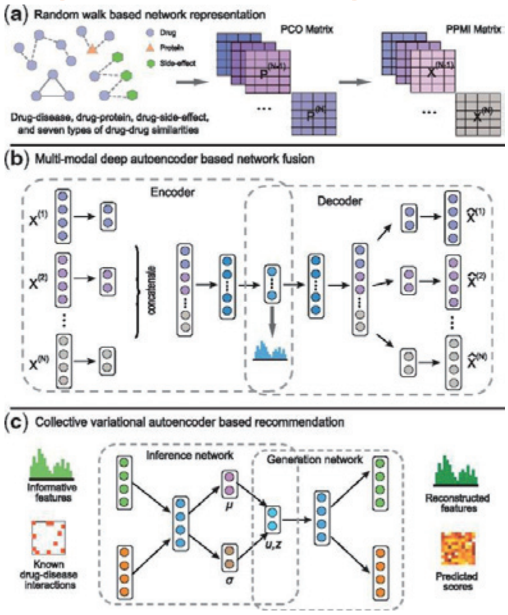
- **Input**: Integrated network of 10 different networks:
  - one drug-disease,
  - one drug-side-effect,
  - one drug-target and
  - seven drug-drug networks

- **Method**: A three-step approach for drug repurposing
  1. **Random walk-based representation of 10 networks**
     1. Probabilistic co-occurrence matrix construction by random walks
     2. Shifted pointwise mutual information (PPMI) → **factorization of co-occurrence matrix** for network representation.
  2. **Multi-modal deep autoencoder (MDA) based network fusion** of 10 network representations
  3. **Collective VAE** for *new* drug-disease association prediction: uses
     1. Extracted features from MDA (side (*auxiliary?*) information)
     2. Known drug-disease associations

- The predicted drug-disease associations were validated by the *ClinicalTrials.gov* database

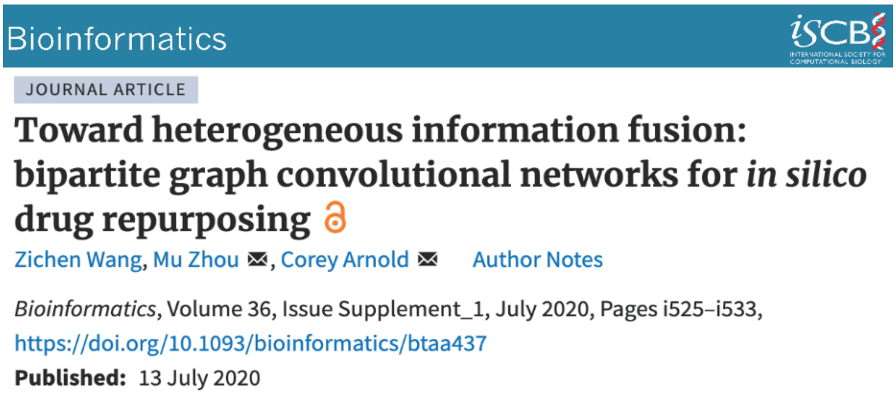
Zeng, Xiangxiang, et al. "deepDR: a network-based deep learning approach to in silico drug repositioning." Bioinformatics 35.24 (2019): 5191-5198.

Network-based Drug Repurposing: DeepDR


(a) Random walk based network representation
Drug
Protein
Side-effect
Drug-disease, drug-protein, drug-side-effect, and seven types of drug-drug similarities
PCO Matrix
PPMI Matrix
(b) Multi-modal deep autoencoder based network fusion
Encoder
Decoder
concatenate
(c) Collective variational autoencoder based recommendation
Informative features
Known drug-disease interactions
Inference network
Generation network
μ
σ
u,z
Reconstructed features
Predicted scores

Zeng, Xiangxiang, et al. "deepDR: a network-based deep learning approach to in silico drug repositioning." Bioinformatics 35.24 (2019): 5191-5198.

Network-based Drug Repurposing: BiFusion



Bioinformatics

JOURNAL ARTICLE

Toward heterogeneous information fusion: bipartite graph convolutional networks for *in silico* drug repurposing

Zichen Wang, Mu Zhou ✉, Corey Arnold ✉ Author Notes

Bioinformatics, Volume 36, Issue Supplement_1, July 2020, Pages i525–i533, https://doi.org/10.1093/bioinformatics/btaa437

Published: 13 July 2020

Wang, Zichen, Mu Zhou, and Corey Arnold. "Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing." Bioinformatics 36.Supplement_1 (2020): i525-i533.

Network-based Drug Repurposing: BiFusion

BiFusion (Wang et al., ISMB 2020)

- **Input:**
  - Drug-protein-disease heterogeneous network

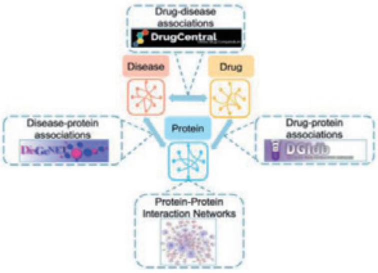
- **Method: 3-step deep learning framework**
  - **A bipartite GCN encoder for drug-disease pair embedding**
  - **Bipartite graph attention to protein** (*gene or protein centric*)
    - disease→protein
    - drug → protein
  - **Bipartite graph attention from protein** (*gene or protein centric*)
    - protein → disease
    - protein → drug

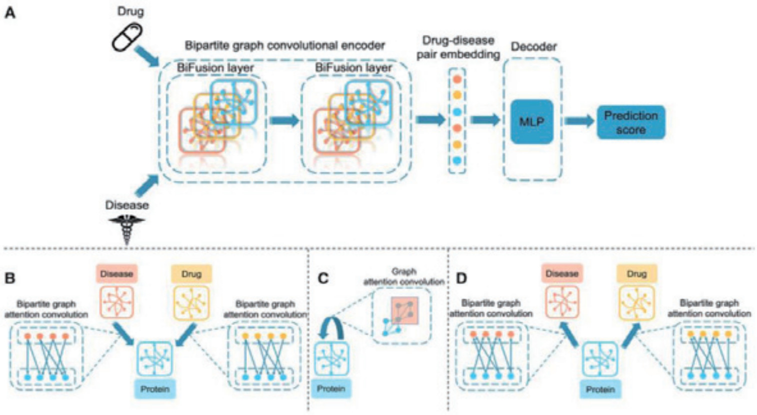
Wang, Zichen, Mu Zhou, and Corey Arnold. "Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing." Bioinformatics 36.Supplement_1 (2020): i525-i533.

Network-based Drug Repurposing: BiFusion

BiFusion (Wang et al., ISMB 2020)

Heterogeneous network of BiFusion

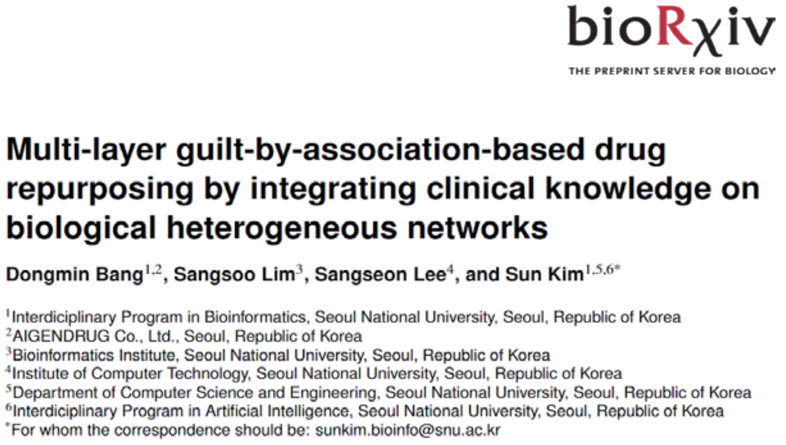

Drug-disease associations
DrugCentral
Disease
Drug
Protein
Disease-protein associations
DisGeNET
Drug-protein associations
DGIdb
Protein-Protein Interaction Networks


A
Drug
Bipartite graph convolutional encoder
BiFusion layer
BiFusion layer
Drug-disease pair embedding
Decoder
MLP
Prediction score
Disease
B
Disease
Drug
Bipartite graph attention convolution
Bipartite graph attention convolution
Protein
C
Graph attention convolution
Protein
D
Disease
Drug
Bipartite graph attention convolution
Bipartite graph attention convolution
Protein

Wang, Zichen, Mu Zhou, and Corey Arnold. "Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing." Bioinformatics 36.Supplement_1 (2020): i525-i533.

Network-based Drug Repurposing: DREAMwalk

DREAMwalk (Bang et al., *in revision*)


bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Multi-layer guilt-by-association-based drug repurposing by integrating clinical knowledge on biological heterogeneous networks

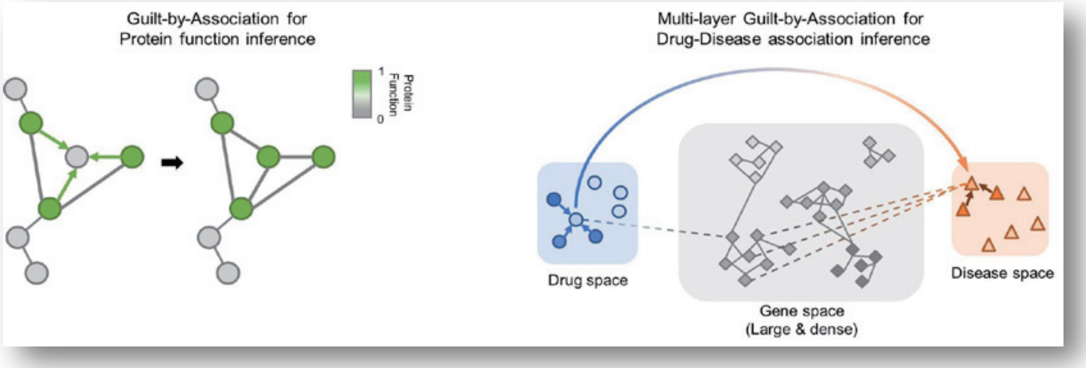
Dongmin Bang[1,2], **Sangsoo Lim**[3], **Sangseon Lee**[4], **and Sun Kim**[1,5,6*]

[1]Interdiciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea
[2]AIGENDRUG Co., Ltd., Seoul, Republic of Korea
[3]Bioinformatics Institute, Seoul National University, Seoul, Republic of Korea
[4]Institute of Computer Technology, Seoul National University, Seoul, Republic of Korea
[5]Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea
[6]Interdiciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea
[*]For whom the correspondence should be: sunkim.bioinfo@snu.ac.kr

Network-based Drug Repurposing: DREAMwalk

DREAMwalk (Bang et al., *in preparation*)

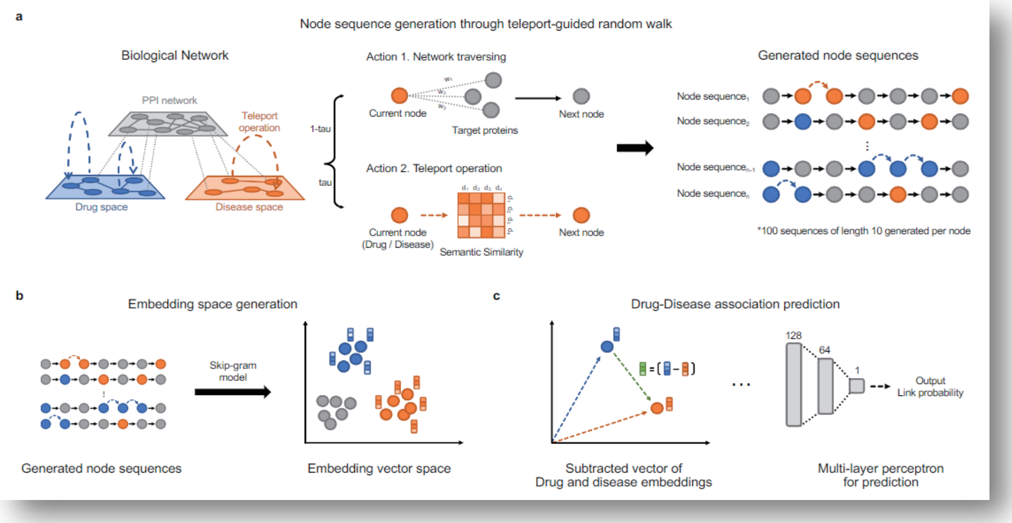
- Input:
  - Drug-gene-disease heterogeneous network
- Method:
  - *Semantic multi-layer Guilt-by-association*
  - Implemented by random walk with **clinical knowledge-guided teleport**
  - Teleport is performed to semantically similar neighbor drug/diseases


Guilt-by-Association for
Protein function inference
1
Protein Function
0
Multi-layer Guilt-by-Association for
Drug-Disease association inference
Drug space
Gene space
(Large & dense)
Disease space

Network-based Drug Repurposing: DREAMwalk

DREAMwalk (Bang et al., *in preparation*)

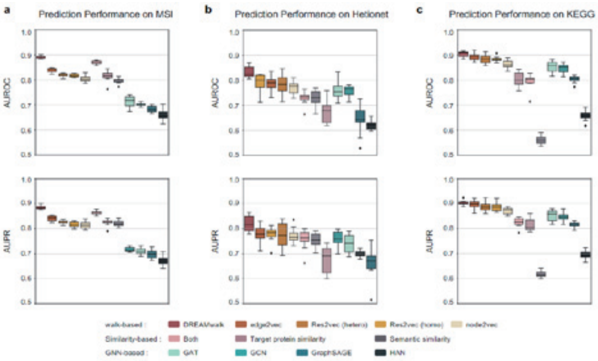
- Method overview

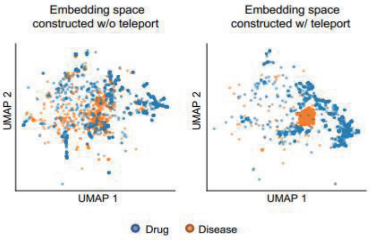

a
Node sequence generation through teleport-guided random walk
Biological Network
PPI network
Teleport operation
Drug space
Disease space
1-tau
tau
Action 1. Network traversing
Current node
Target proteins
Next node
Action 2. Teleport operation
Current node (Drug / Disease)
Semantic Similarity
Next node
Generated node sequences
Node sequence$_1$
Node sequence$_2$
Node sequence$_{n-1}$
Node sequence$_n$
*100 sequences of length 10 generated per node
b
Embedding space generation
Skip-gram model
Generated node sequences
Embedding vector space
c
Drug-Disease association prediction
128
64
1
Output Link probability
Subtracted vector of Drug and disease embeddings
Multi-layer perceptron for prediction

Network-based Drug Repurposing: DREAMwalk

DREAMwalk (Bang et al., *in preparation*)

- Results:
  - State-of-the-art drug-disease association prediction
  - Harmonious embedding space of both clinical and biological contexts

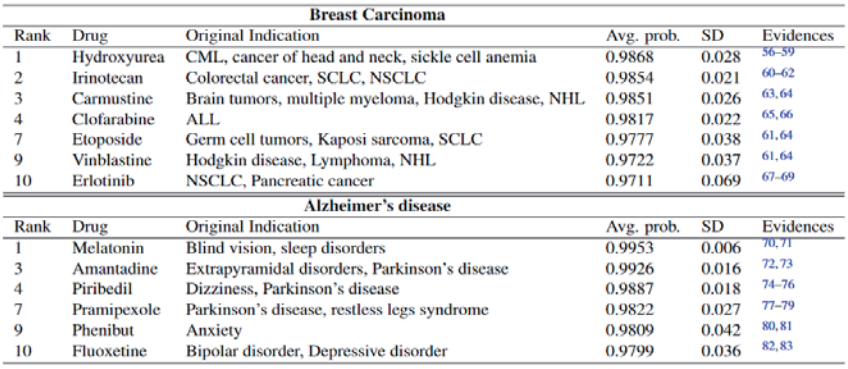




Network-based Drug Repurposing: DREAMwalk

DREAMwalk (Bang et al., *in preparation*)

- Results:
    - Drug repurposing for breast carcinoma and Alzheimer's disease: well supported by literatures



Breast Carcinoma

Rank	Drug	Original Indication	Avg. prob.	SD	Evidences
1	Hydroxyurea	CML, cancer of head and neck, sickle cell anemia	0.9868	0.028	56–59
2	Irinotecan	Colorectal cancer, SCLC, NSCLC	0.9854	0.021	60–62
3	Carmustine	Brain tumors, multiple myeloma, Hodgkin disease, NHL	0.9851	0.026	63, 64
4	Clofarabine	ALL	0.9817	0.022	65, 66
7	Etoposide	Germ cell tumors, Kaposi sarcoma, SCLC	0.9777	0.038	61, 64
9	Vinblastine	Hodgkin disease, Lymphoma, NHL	0.9722	0.037	61, 64
10	Erlotinib	NSCLC, Pancreatic cancer	0.9711	0.069	67–69

Alzheimer's disease

Rank	Drug	Original Indication	Avg. prob.	SD	Evidences
1	Melatonin	Blind vision, sleep disorders	0.9953	0.006	70, 71
3	Amantadine	Extrapyramidal disorders, Parkinson's disease	0.9926	0.016	72, 73
4	Piribedil	Dizziness, Parkinson's disease	0.9887	0.018	74–76
7	Pramipexole	Parkinson's disease, restless legs syndrome	0.9822	0.027	77–79
9	Phenibut	Anxiety	0.9809	0.042	80, 81
10	Fluoxetine	Bipolar disorder, Depressive disorder	0.9799	0.036	82, 83

Summary of Drug Repurposing

- There are multiple ways to learn embedding vectors for drug
    - Drug-centered embeddings from Drug-drug, Drug-target, Drug-disease.
    - Then, how to combine different views on drugs?
    - **deepDR**: Multi-modal deep autoencoder
    - **SNF-cVAE**: similarity network fusion
    - **DreamWalk**: semantic random walks

- Three-way relationship among drug-gene-disease cannot be learned at once.
- In the end, we need to deduce **drug-disease binary relationship**.
    - Basically, binary relationships are somehow combined on different layers, hierarchically.
    - **deepDR**: Multi-modal deep autoencoder; then cVAE for drug-disease
    - **SNF-cVAE**: similarity network fusion; then cVAE for drug-disease
    - **BiFusion**: protein-centric bipartite graph attention twice; then MLP for drug-disease
    - **Zhang, Zhao et. al**: row pairing from drug-drug, drug-disease, disease-drug matrices; path generation by aligning paired vectors; then CNN + LSTM for drug-disease
    - **DreamWalk**: semantic random walks; then drug-disease embedding in the same space; then similarity between drug vector and disease vector for drug-disease

감사합니다!