Graph in Drug Discovery

Advanced Training
2020. 11. 11

Erkhembayar J. Ph.D

Korea AI Center for Drug Discovery and Development

https://www.nature.com/articles/nbt0717-604

AGENDA

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

	HOURS	CONTENT
GRAPH IN DRUG DISCOVERY		
INTRODUCTION	09.00 – 9.30	How Graph is used In Drug Discovery
CODING SESSSION (colab)		
CODING SESSION	09:40 - 11.00	How to Construct your own Graph? How to Construct NodeFeaturizer(Atom)? How to Construct EdgeFeaturizer(Bond)?
COVID19 SPECIAL: DRUG REPOSITIONING		
DRUG REPOSITIONING	11:00 - 11.30	Disease, Target, Drug Centric DR Knowledge Graph
CODING SESSION (colab)		
CODING SESSION	11:30 - 12.00	Knowledge –based Drug Repositioning
LUNCH		

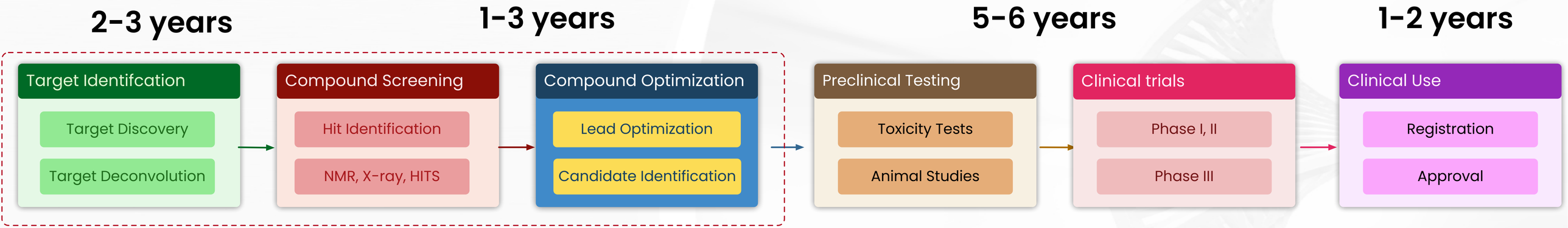
GRAPH IN DRUG DISCOVERY

General Overview of Graph Application in the Field

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

Drug Discovery

- The process of drug discovery and development are **expensive** and **time consuming** with **high failure rate**



Bringing a New Drug to Marker

Expensive

~$2.8B

DiMasi et al.: Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. J. Health Econ. 2016, 47, 20–33

Time Consuming

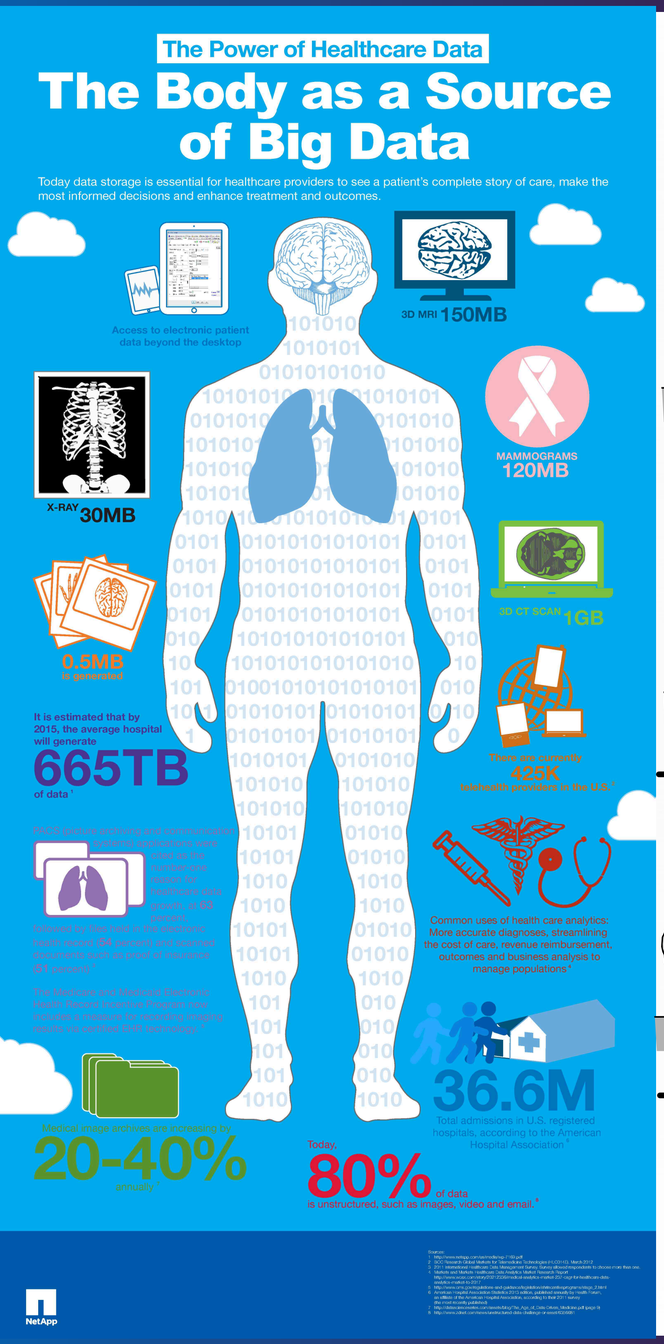
~11-16 years

Matthews, et al.:"Omics"-Informed Drug and Biomarker Discovery: Opportunities, Challenges and Future Perspectives. Proteomes 2016, 4, 28.

High Failure Rate

90%

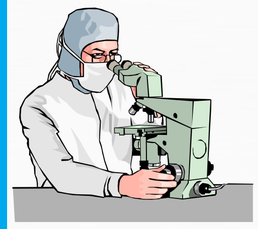
Dowden, et al.: Trends in Clinical Success Rates and Therapeutic Focus. Nat. Rev. Drug Discovery 2019, 18, 495–496.

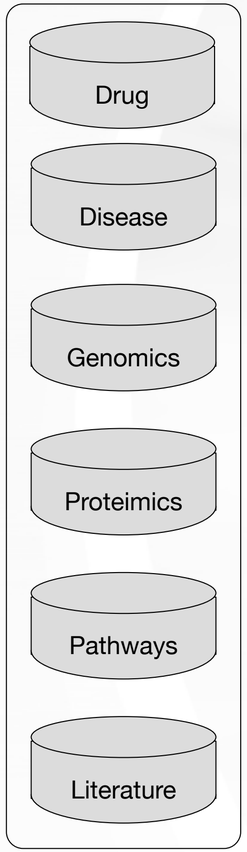


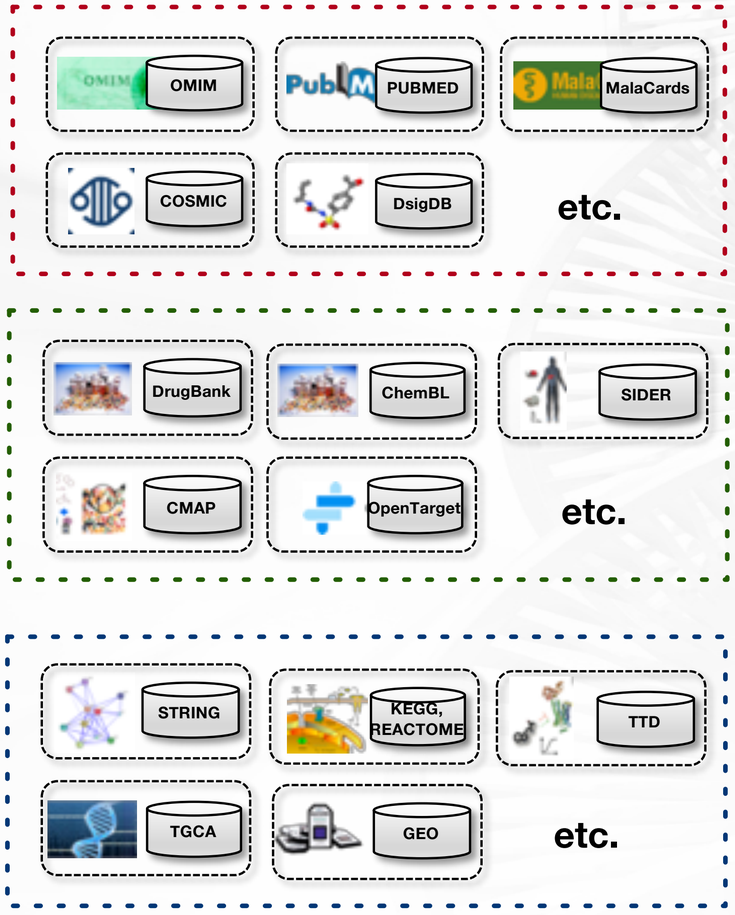
Complexity in Data Space

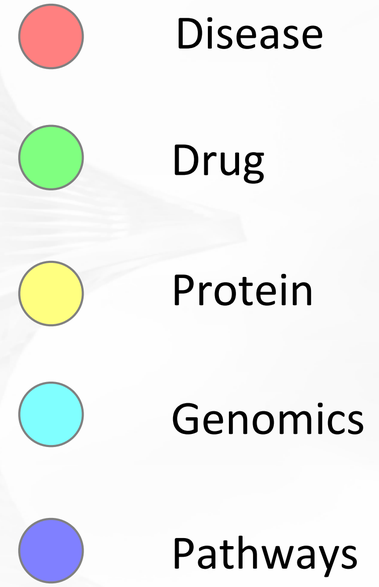
KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

high-content, high-dimensional biological data









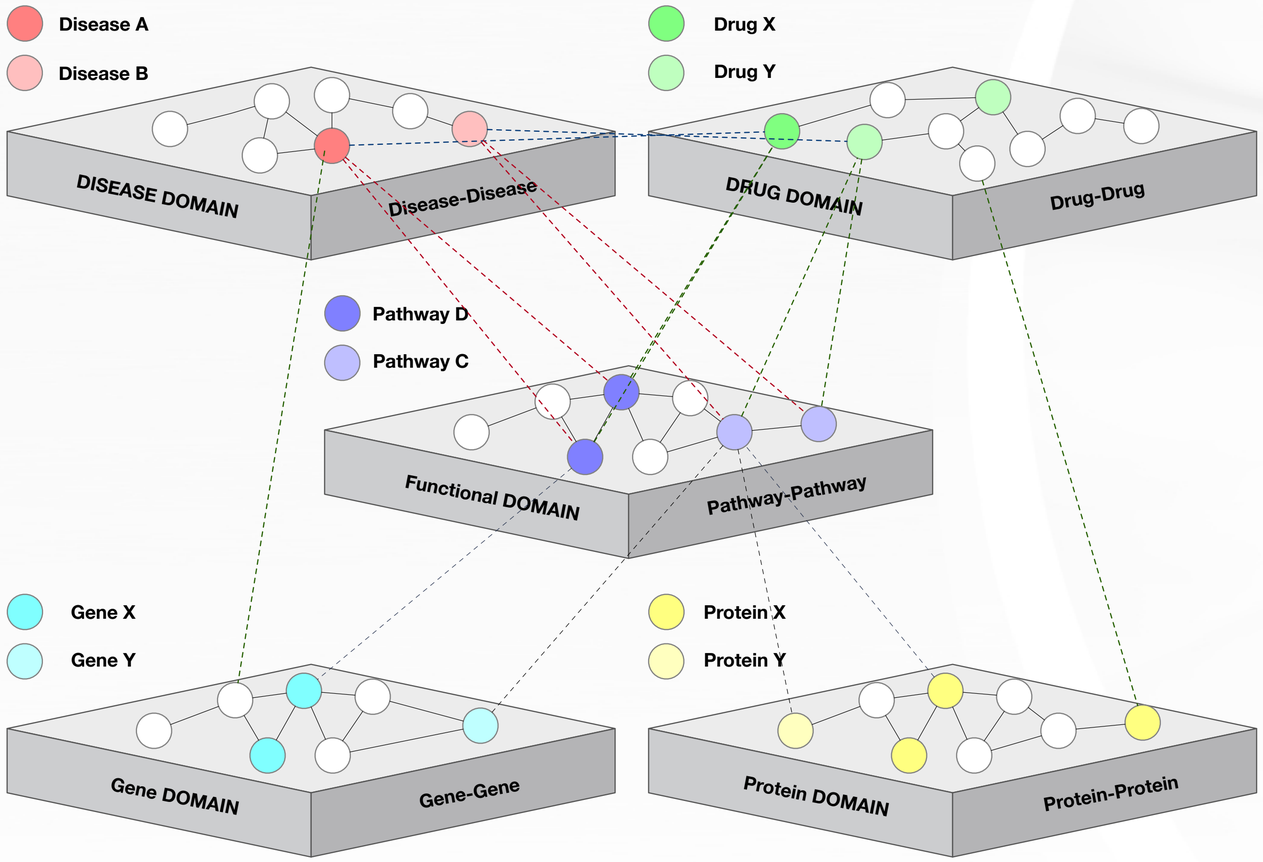
How to use these resources efficiently in an integrative way?

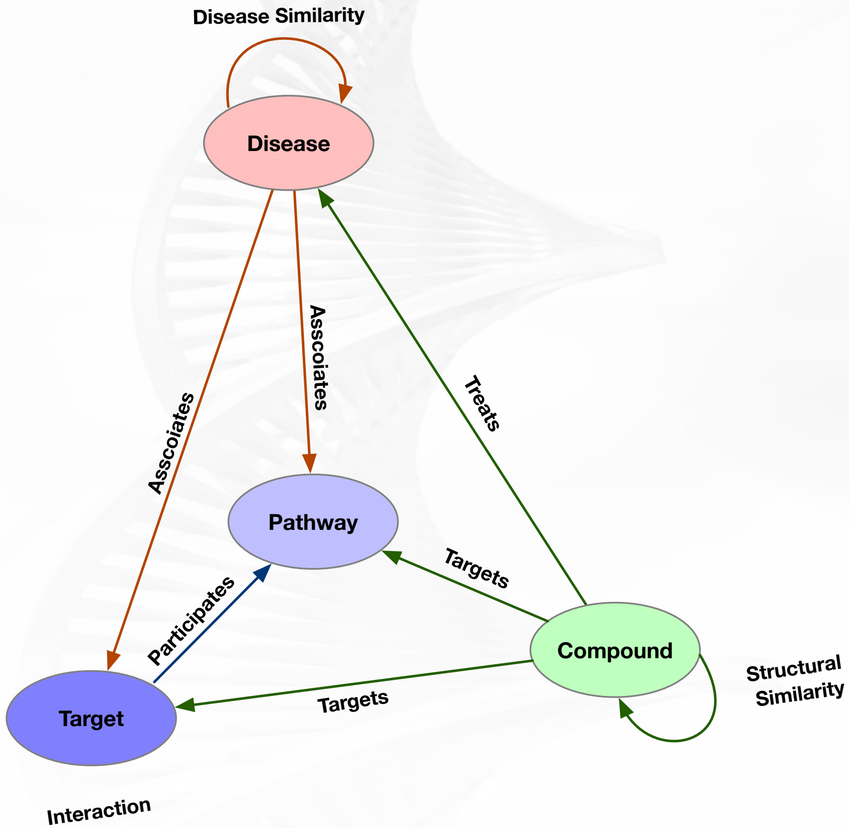
How do we take advantage of this data to discover a new drug?

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

Complexity in Representation

- Too complex to represent and interpretation



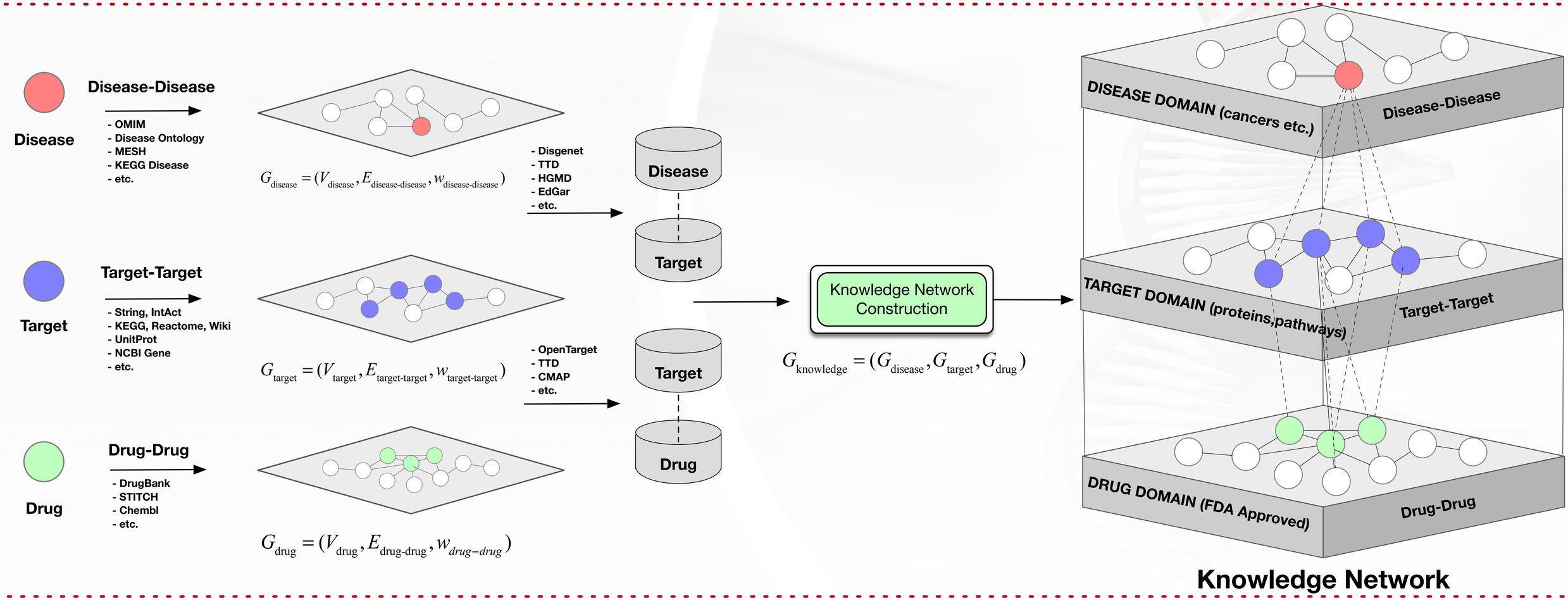


Need better representation for human interpretation ?

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

Data Representation in Graph

- Graphs integrative and representative power to solve issues



KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

Graphs in Drug Discovery

Data Scientist Perspective

Early Drug Discovery

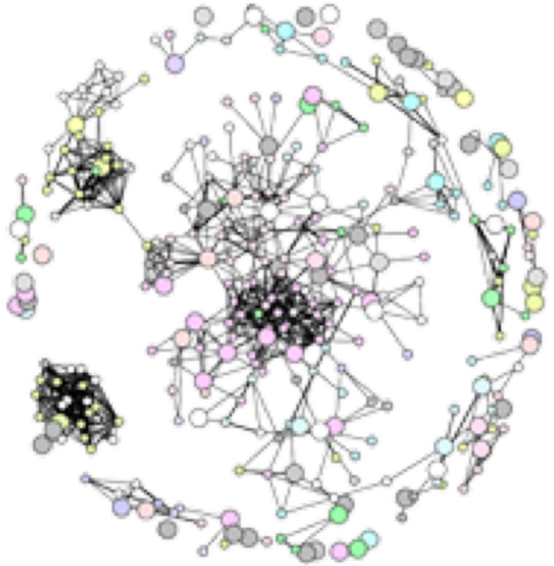
TARGET IDENTIFICATION
- Drug-Target Identification
- Target Validation
- Protein and pathway screening in-vitro

HIT IDENTIFICATION
- DRUG-TARGET INTERACTION
- QSAR
- Compound Screening

REPURPOSE EXISTING DRUGS
- Use Existing Drugs
- New Indication for Drugs

HIT TO LEAD
- Lead Identification
- Lead Optimization

Input	Output
DISEASE, TARGET(Omics, Pathways, PPI)	NEW TARGET
COMPOUND, TARGET	COMPOUND (HITs)
DRUG, TARGET RELATIONSHIP	New Indications, Drug Candidates
COMPOUND, TARGET	Drug Candidates (LEADs)



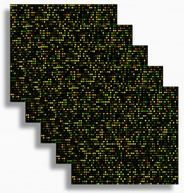
Graph

Early Stage – Biomarker Identification

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

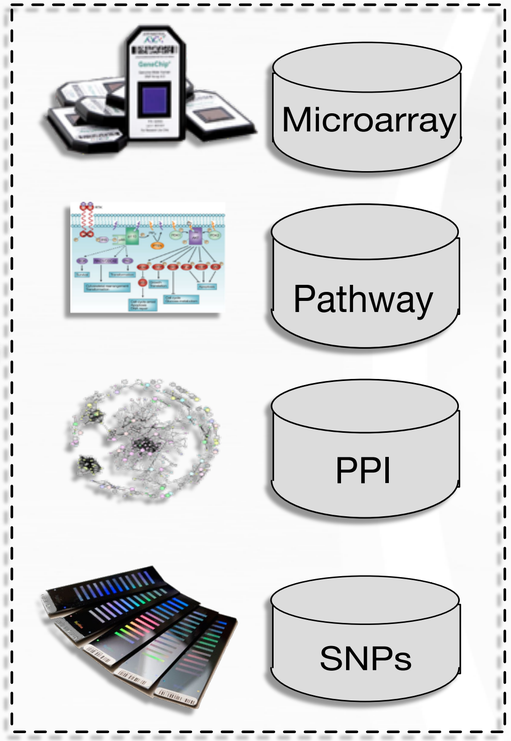
Disease group

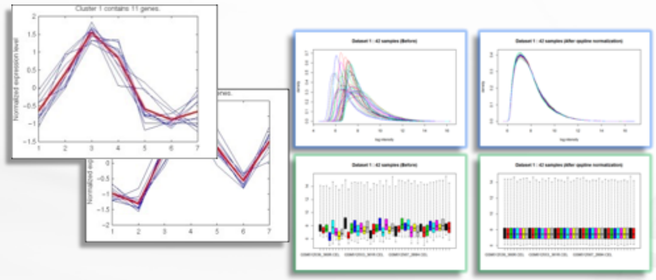
Normal group



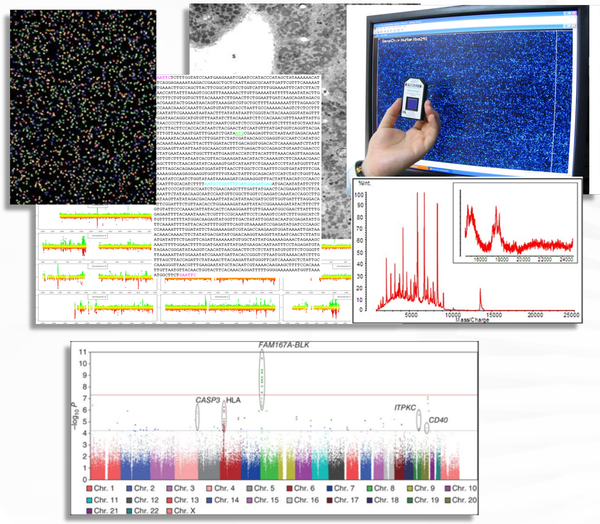
High-throughput Experiments

(e.g: microarray, pathway etc.)



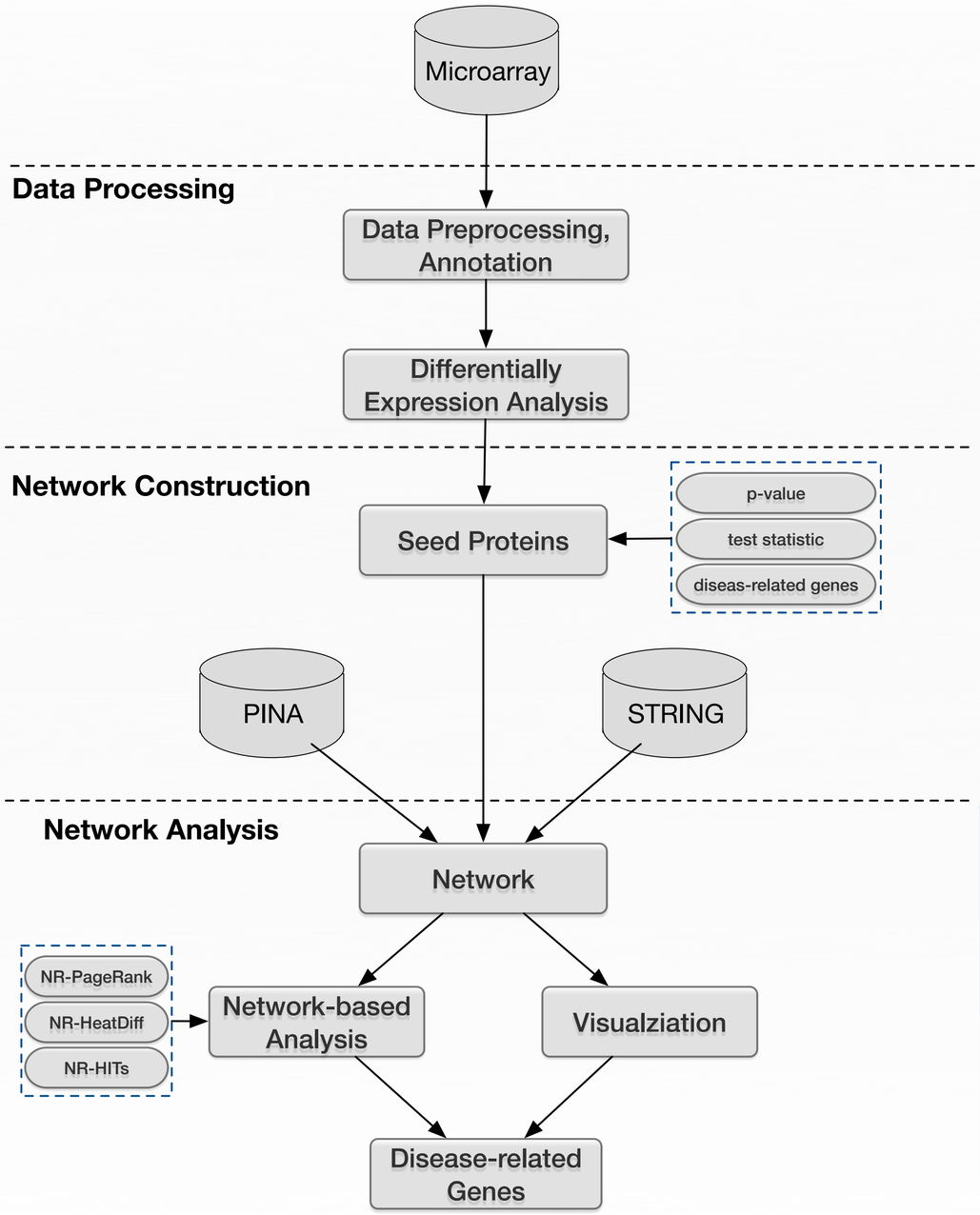


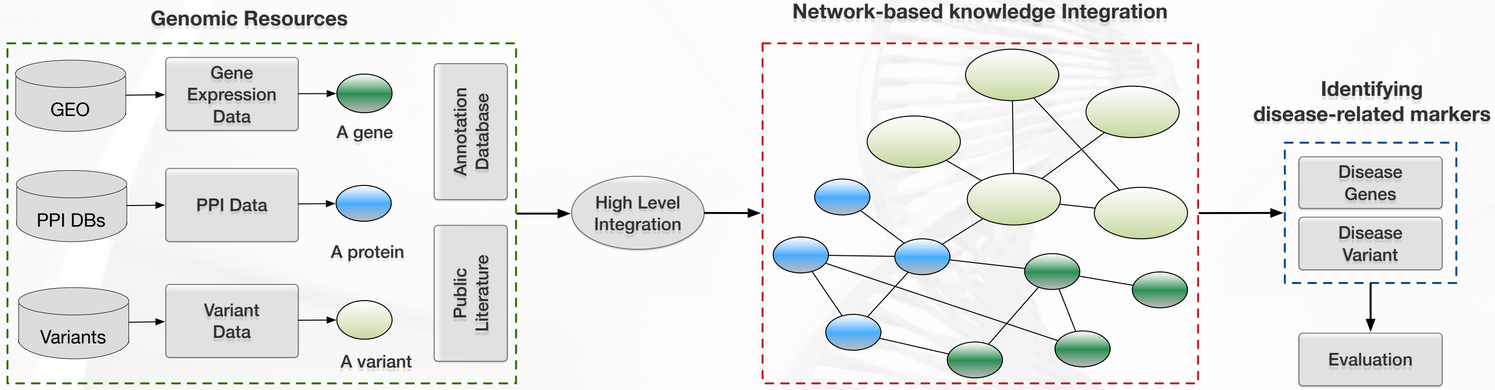
Disease Marker Identification



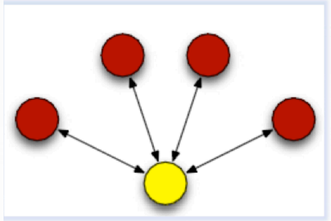
Validation

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development





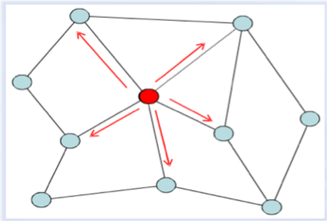
- PageRank
  - Brin S. (1998)
  - Based on idea that a web page should be highly ranked if other highly ranked pages contain hyperlinks to it.



A node is important if it is linked to other highly important nodes

$$r_j^{(n)} = (1-d) \times r_j^{(0)} + d \times \sum_{i=1}^{N} \frac{w_{ij} \times r_i^{n-1}}{\deg_i}$$

- Heat Diffusion Rank
  - Chung F. (2007)
  - The heat diffusion algorithm based on idea that different initial temperature distributions will give rise to different temperature distribution after a fixed time period

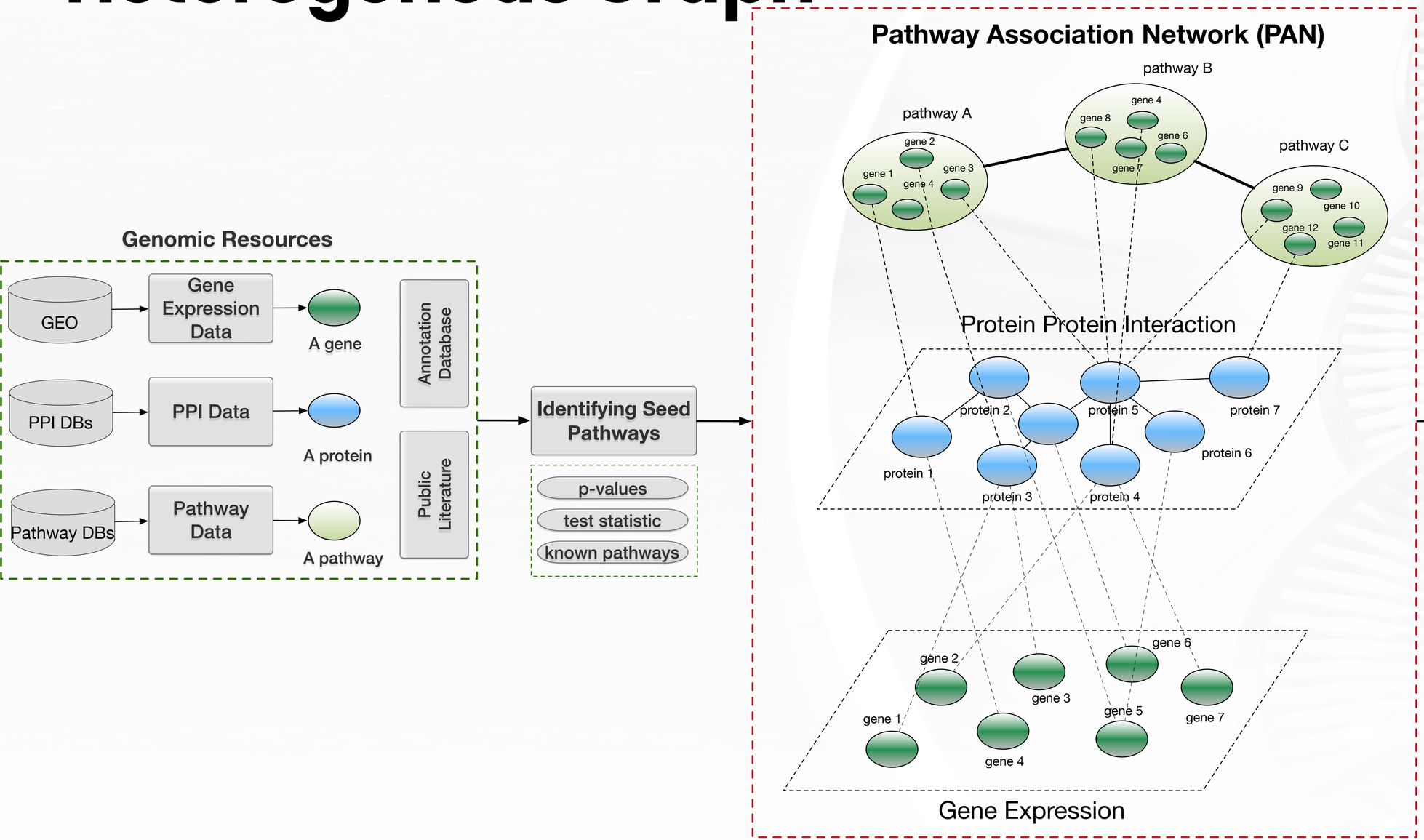


Natural Phenomon:
-Hot node diffuses heat to nearby node

$$r_v^{(i+1)} = (1-\frac{t}{N}) \cdot r_v^{(0)} + \frac{t}{N} \cdot \sum_{\{u \in V,(u,v) \in E\}} r_u^i \cdot \frac{w(u,v)}{d(u)}$$

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

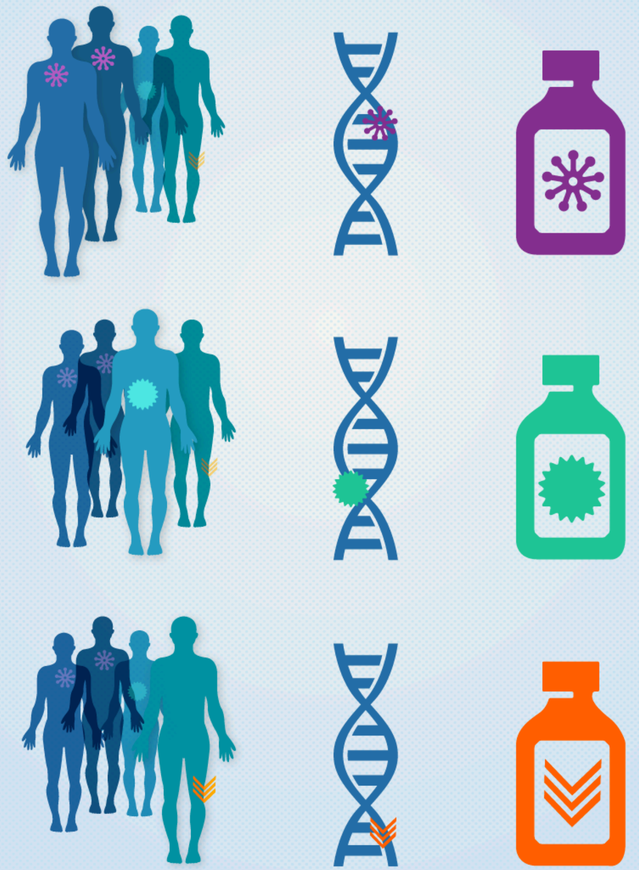
Heterogenous Graph



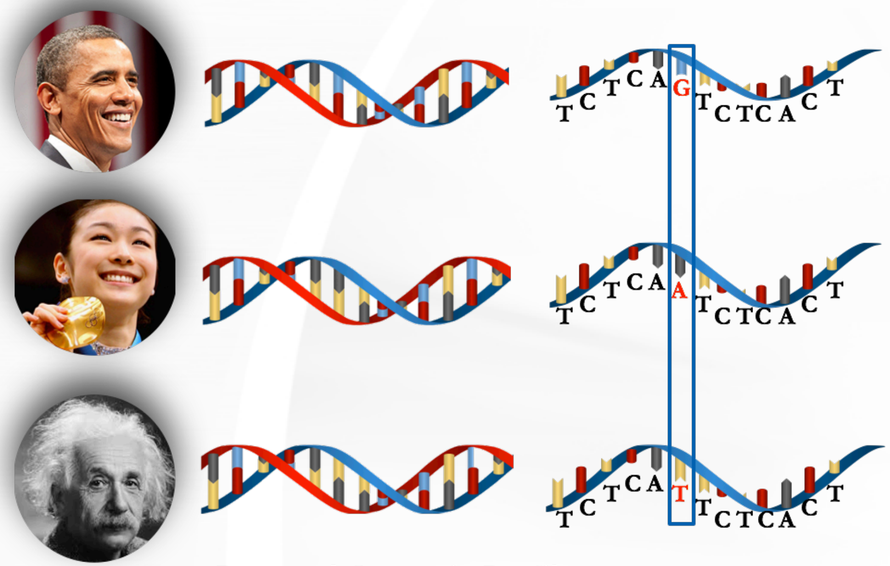
KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

NATIONAL CANCER INSTITUTE
PRECISION MEDICINE IN CANCER TREATMENT

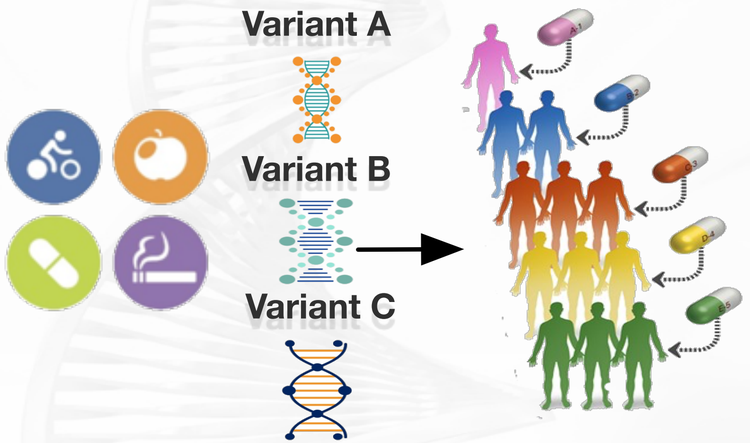
Discovering unique therapies that treat an individual's cancer based on the specific genetic abnormalities of that person's tumor.

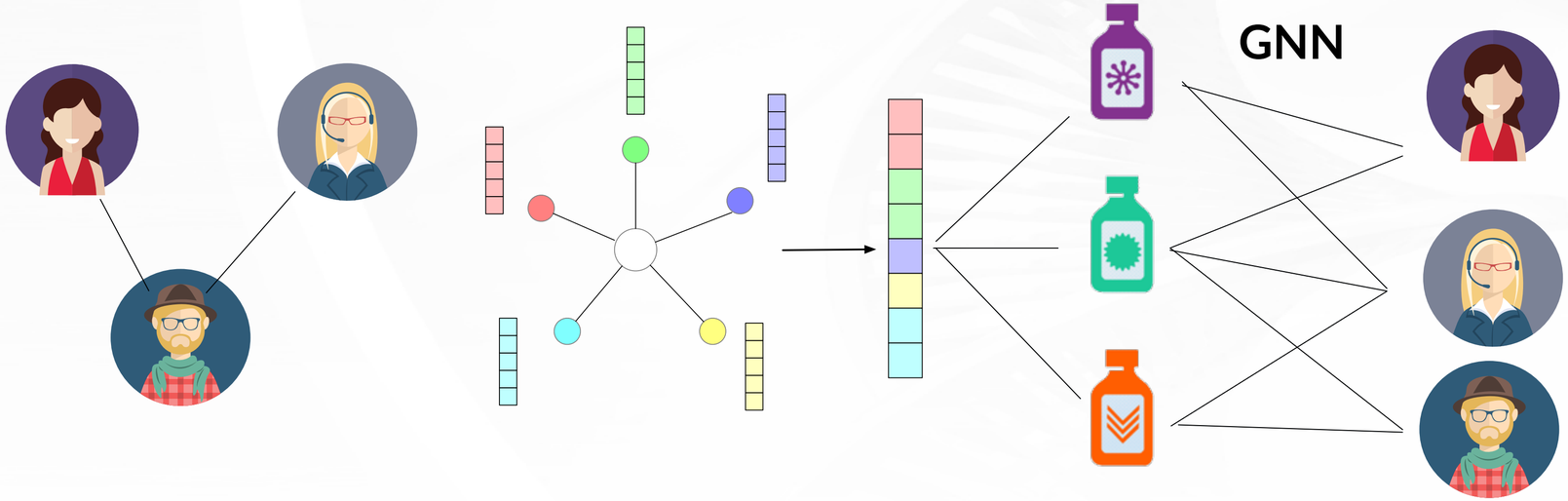


www.cancer.gov



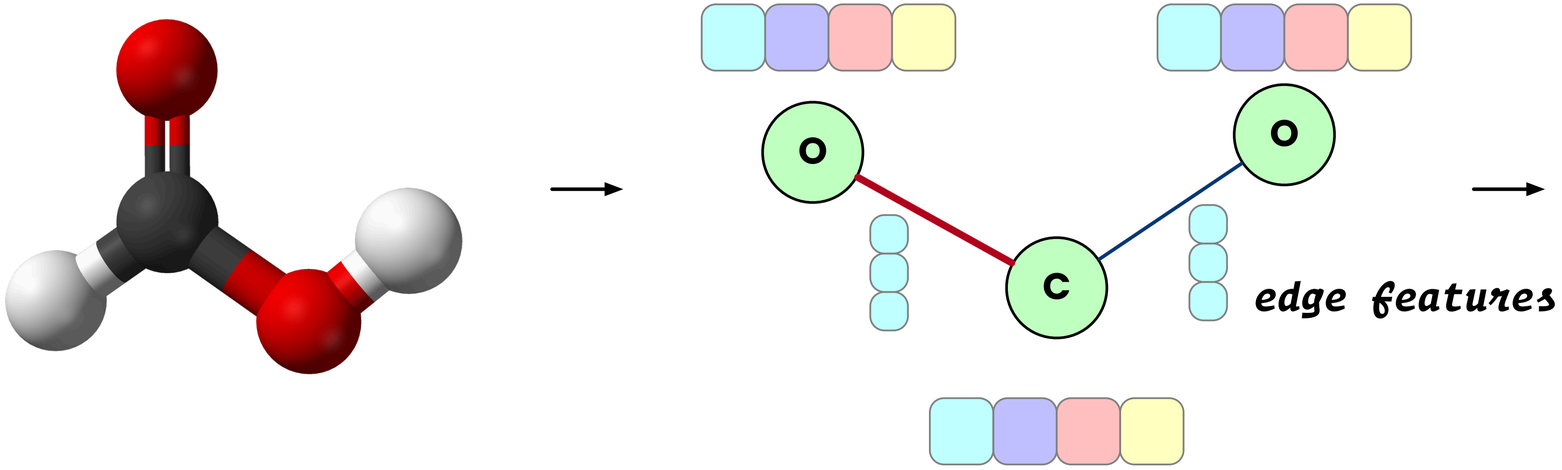
Personal Genomic Profile



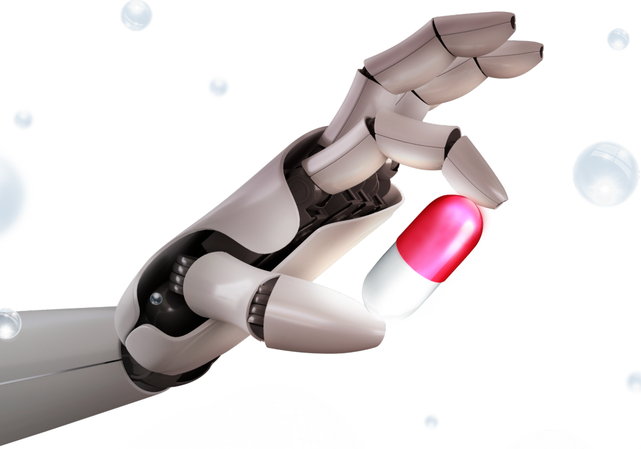


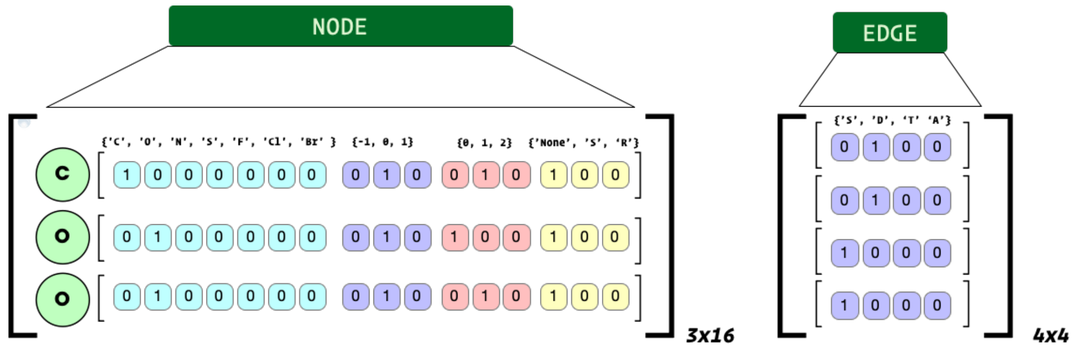
KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

How Graph is Used In Chemical Space ?



- PROPERTY PREDICTION
- DRUG-TARGET INTERACTION
- DE NOVO DRUG DESIGN



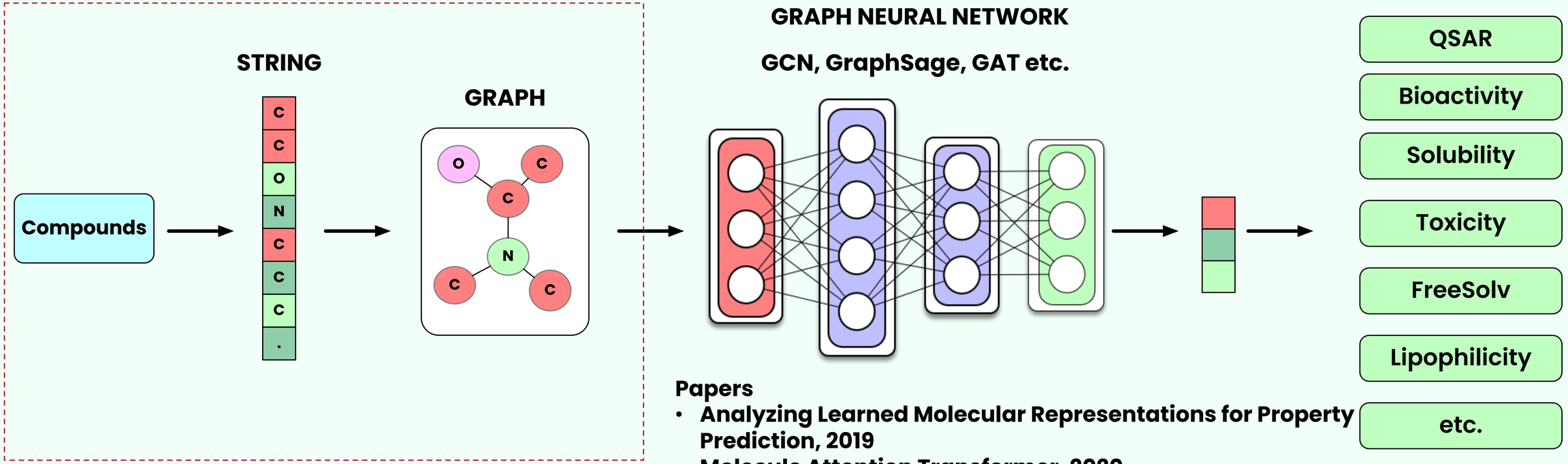


CODING SESSION: How to Make Graph Augmentation

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

DRUG PROPERTY PREDICTION

Data Augmentation

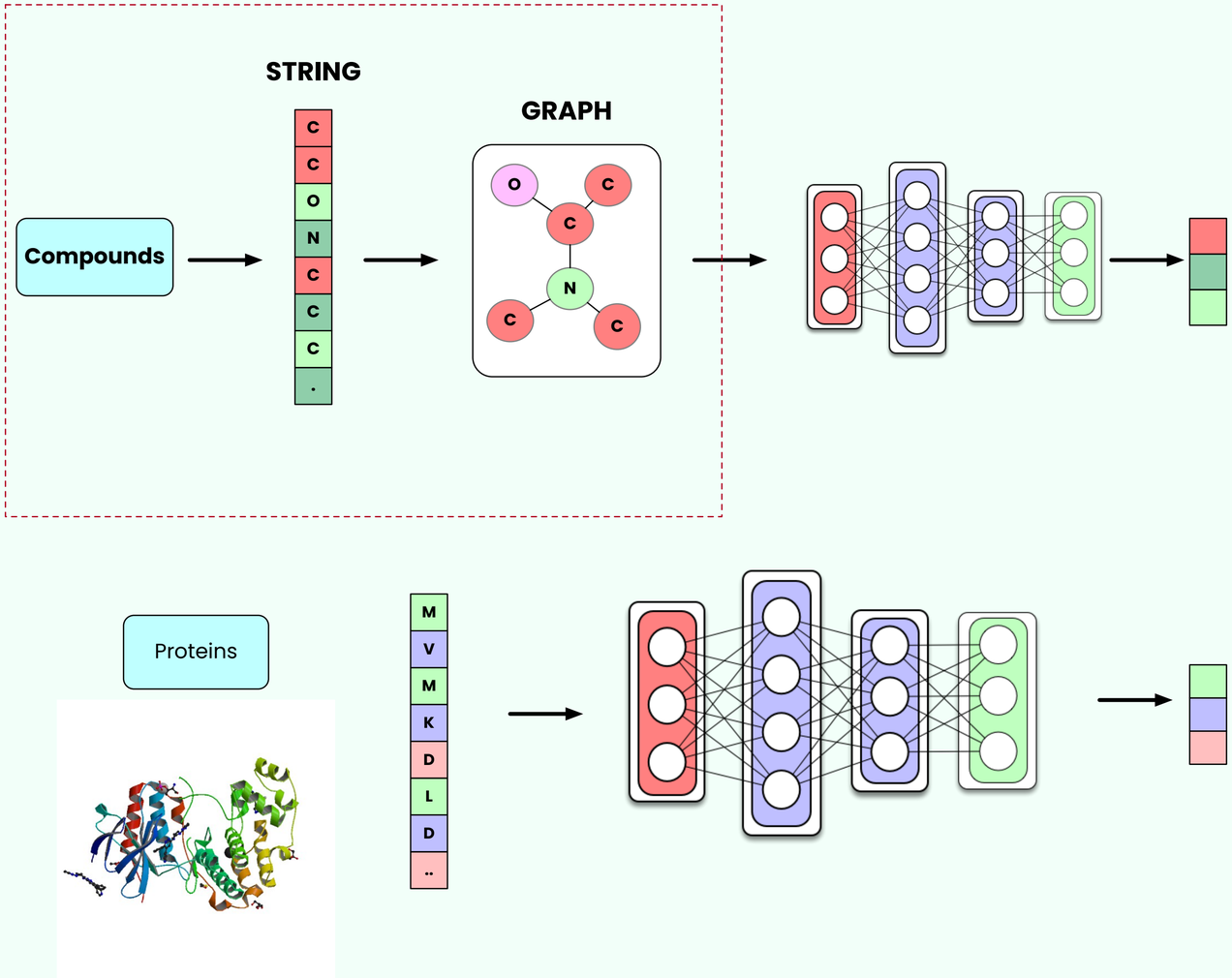


Papers

- **Analyzing Learned Molecular Representations for Property Prediction, 2019**
- **Molecule Attention Transformer, 2020**
- **Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism, 2020**

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

DRUG-TARGET INTERACTION PREDICTION



Papers

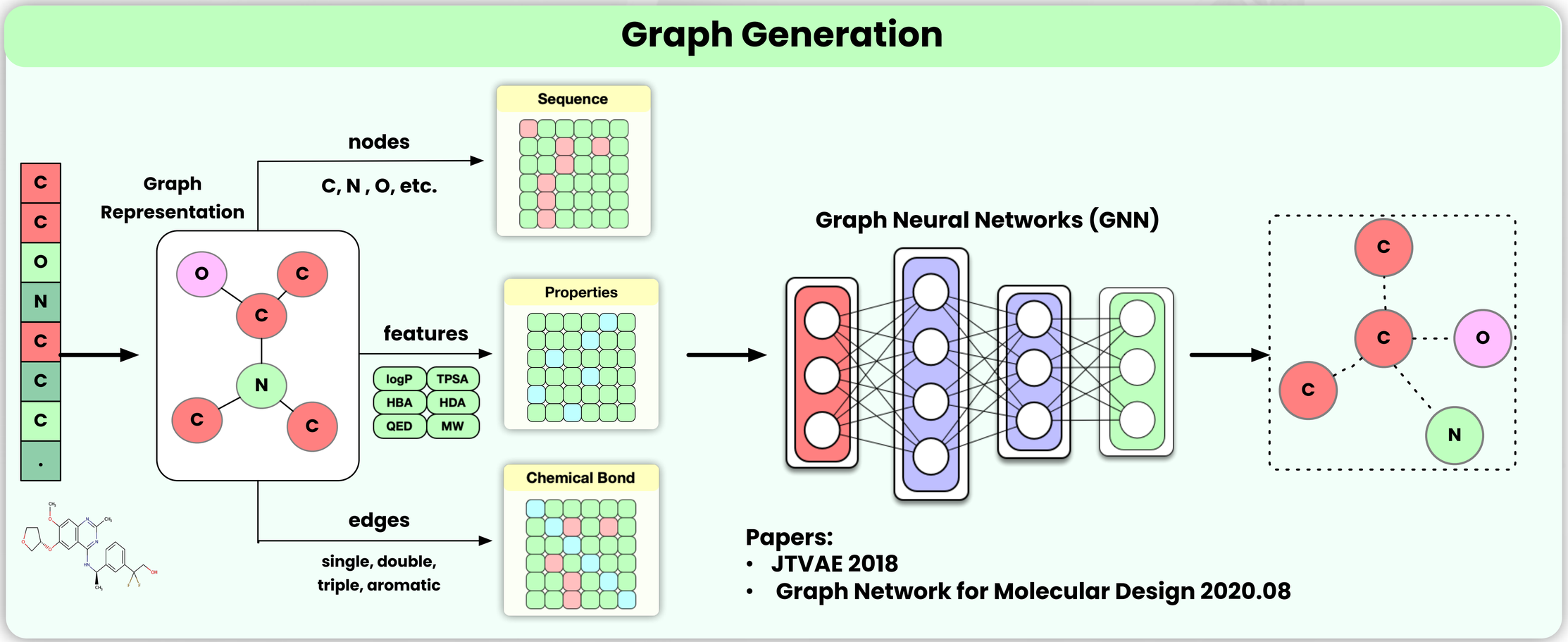
- **GraphDTA 2019**
- **MONN: A Multi-objective Neural Network for Predicting Compound-Protein Interactions and Affinities 2020**

#	Compound	Protein	IC50/EC50	Kd/Ki
1	ChemicalID	ProteinID	0.99	0.99
2	...	...	...	...
3			...	...
4			...	...
N			...	...

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

Graph Generation

- Graph Neural Networks
  - Input: Graph; Output: Graph



KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

GRAPH NETWORKS FOR MOLECULAR DESIGN

Rocío Mercado,[†1] Tobias Rastemo,[†,‡] Edvard Lindelöf,[†,‡] Günter Klambauer,[‖] Ola Engkvist,[†] Hongming Chen,[§] Esben Jannik Bjerrum[†]

[†] *Molecular AI, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, SE*
[‡] *Chalmers University of Technology, Gothenburg, SE*
[‖] *Institute of Bioinformatics, Johannes Kepler University, Linz, AU*
[§] *Centre of Chemistry and Chemical Biology, Guangzhou Regenerative Medicine and Health, Guangdong Laboratory, Guangzhou, CN*

ABSTRACT

Deep learning methods applied to chemistry can be used to accelerate the discovery of new molecules. This work introduces GraphINVENT, a platform developed for graph-based molecular design using graph neural networks (GNNs). GraphINVENT uses a tiered deep neural network architecture to probabilistically generate new molecules a single bond at a time. All models implemented in GraphINVENT can quickly learn to build molecules resembling the training set molecules without any explicit programming of chemical rules. The models have been benchmarked using the MOSES distribution-based metrics, showing how GraphINVENT models compare well with state-of-the-art generative models. This work is one of the first thorough graph-based molecular design studies, and illustrates how GNN-based models are promising tools for molecular discovery.

Keywords deep generative models · graph neural networks · drug discovery · molecular design

1 Introduction

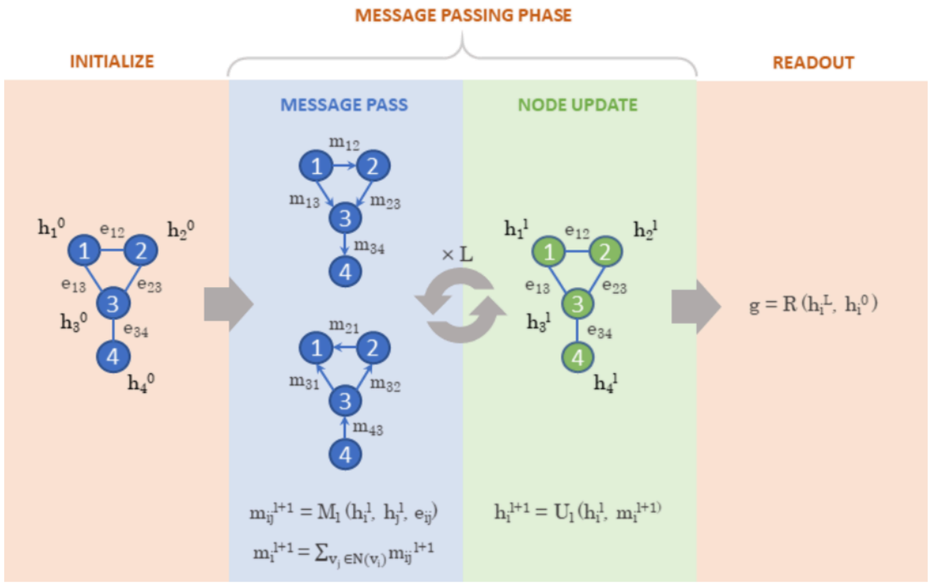
Due to the recent success of deep learning (DL) models across a wide-range of fields, it is often said that we are in the third wave of artificial intelligence (AI). [1] Some of the most utilized architectures at the forefront the recent AI boom are recurrent neural networks (RNNs), used to model sequential processes (such as speech), and convolutional neural networks (CNNs), used in computer vision tasks. [2] More recently, there has been an increase in the use of graph neural networks (GNNs), or more generally, graph networks (GN) [3], for modeling patterns in graph-structured data. Graphs are widespread mathematical structures that can be used to describe an assortment of relational information, and would seem natural choices for organic chemistry as graphs are natural data structures for describing molecular structures.

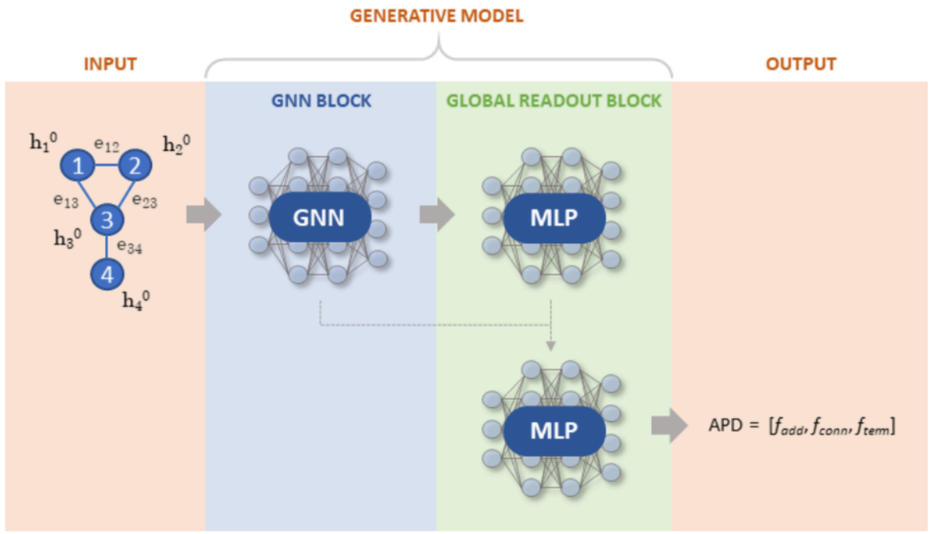
The idea of designing novel pharmaceuticals can be boiled down to generating graphs which meet all the criteria of desirable drug-like molecules. This is the guiding principle behind graph-based molecular design. *De novo* molecular design is the process of designing novel molecules with a specific set of desired pharmacological properties from scratch. This approach is the antithesis of QSAR-based high-throughput screening, where instead the structures are known and their corresponding pharmacological and physical chemical properties are unknown. Molecular generative models have emerged as promising methods for exploring the otherwise intractably large chemical space through *de novo* molecular design [4–11], especially using recurrent neural networks and variational autoencoders. Nonetheless, recent methods [4, 5, 12] have largely focused on training models to generate novel molecules encoded in the string-based SMILES format.

While string-based methods are surprisingly powerful, graphs are more natural data structures for describing molecules, and have many potential advantages over strings, especially when used with graph networks. [13–17] GNNs have the ability to 1) learn atom order permutation invariant representations, 2) encode the graph matrix representation into a latent space, and 3) efficiently train on a GPU and scale to large datasets. Some of these points are not unique to GNNs. However, the graph representation can naturally be expanded in applications where one would need more information than simply the identity and connectivity of atoms in a molecule (e.g. spatial coordinates).

Here, a new platform, GraphINVENT, is introduced for training deep generative models directly on the molecular graph representations using GNNs. First, the various elements of GraphINVENT are introduced, with similarities and differences to string-based generative models highlighted along the way. The six different GNNs used in this work are then described in detail in the methods section, together with hyperparameter tuning and training. The MOSES benchmark and other internal evaluation metrics are then used to compare model performance in training speed, reproduction of molecular properties of the train-

[1]Corresponding author: rocio.mercado@astrazeneca.com

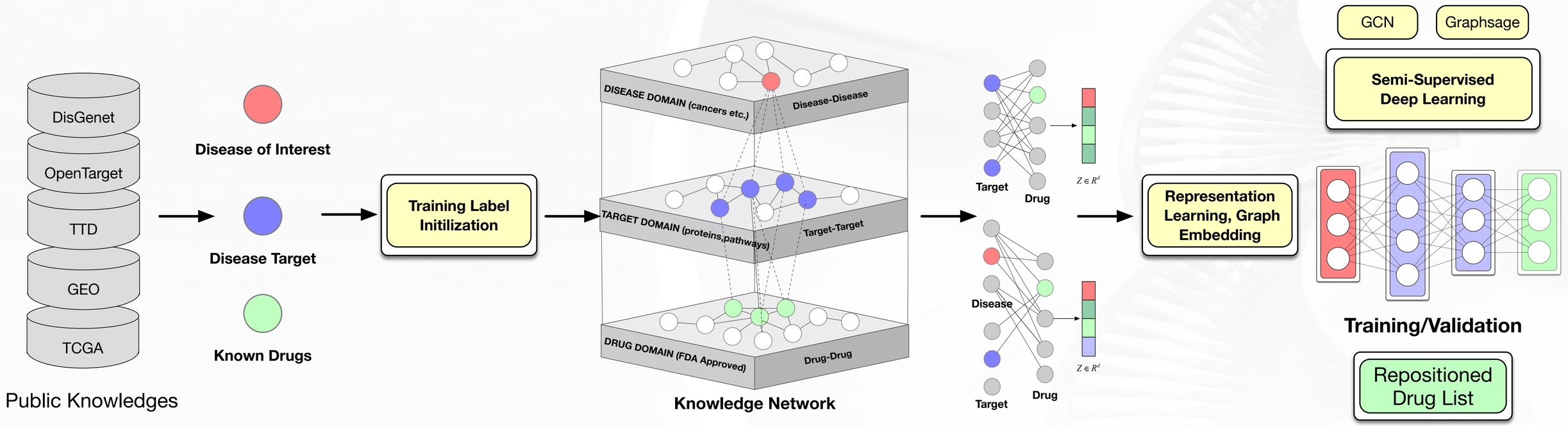




Graph is as input and Graph is generated output

KAICD
한국인공지능신약개발지원센터
Korea AI Center for Drug Discovery and Development

Drug Repositioning



Summary

- **Graph natural** representive power which helps to find patterns in data that scientist might not see
- The solution to many applications in biomedical field can be formulated as **GRAPH LEARNING PROBLEMs**
- **Graphs** are used to
  - Detect new target by taking advantages of complenterieness in variety of biomedical resources
  - Represent chemical compounds and showed promising performance
- **KNOWING GRAPH REPRESENTATION** is **IMPORTANT in DRUG DISCOVERY**

LET`S DO IT IN CODING SESSION

CODING SESSION

10 MINUTES COFFEE BREAK

Thank You