Graph Neural Networks in Life Sciences: Opportunities and Solutions

Zichen Wang, Vassilis N. Ioannidis, Huzefa Rangwala, Tatsuya Arai, Ryan Brand, Mufei Li, Yohei Nakayama

Amazon

Outline of this tutorial

1. Overview of Graph ML in biomedical science [lecture]
2. Making sense of small molecules with GNNs [hands-on]
    1. how to construct features from atom graphs for small organic compounds
    2. molecular property prediction
3. Prediction of COVID-19 mRNA vaccine degradation with GCN
    1. protein function prediction
4. Going beyond single graph, bi-graph based binding affinity prediction for protein-ligand pairs [hands-on]
5. Organizing and generating new knowledge for drug discovery and repurposing with knowledge graphs (KGs) [hands-on]

Overview

- Graphs are ubiquitous in life sciences
- Various ML applications have been developed for biomedical graphs
- Refresher on graph theory and graph neural networks

Graphs in biomedical sciences (by entities)

Molecules

- RNAs, proteins, small molecules

Omics: genomics, transcriptomics, proteomics, metabolomics

- PPI network, gene regulatory network (GRN), biological pathways

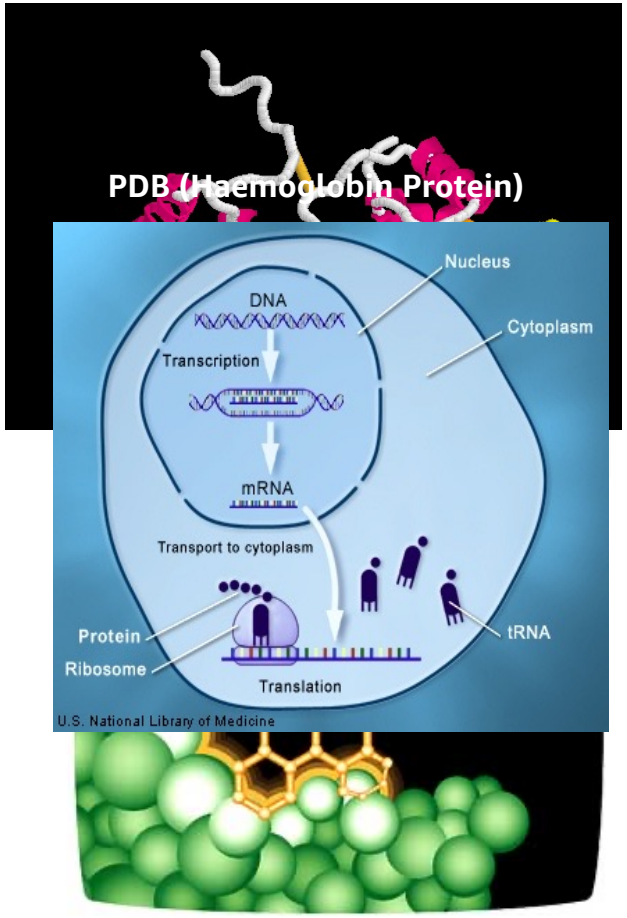
Therapies: (perturbation, phenotype)

- Drug-disease networks, disease-disease similarity networks

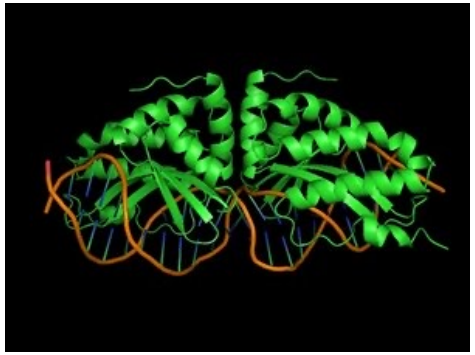
Healthcare: histopathological images, EHRs, Clinical Trials, Knowledge Graphs (KGs)

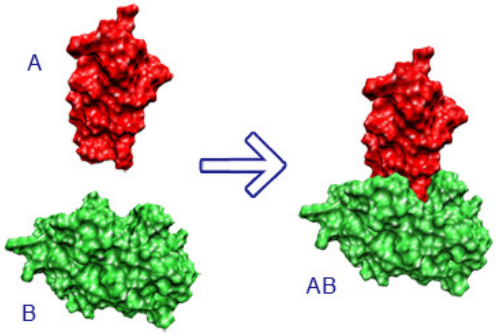
Molecules: DNA, mRNA, Proteins

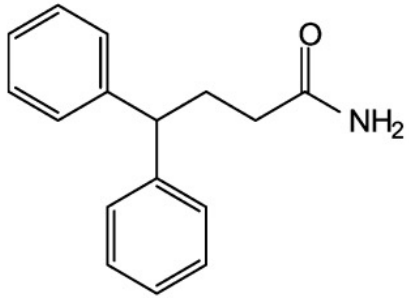
- Proteins
  - poly-peptides ~ 70-3000 amino-acids.
  - Central Dogma: DNA -> mRNA -> Protein.
- Proteins govern several processes
  - Oxygen binding/release is controlled by conformational changes.
  - Bind to other DNA/RNA/proteins and molecule.
- Several other functions – dependent on the global or local structure.

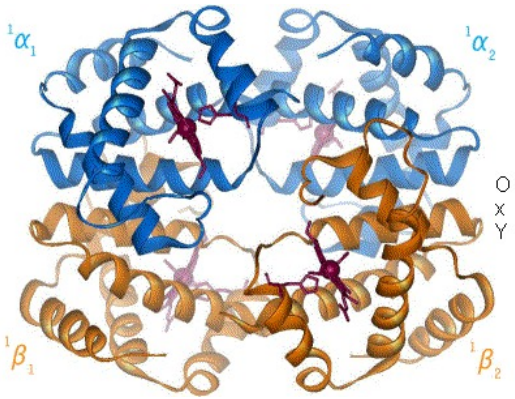


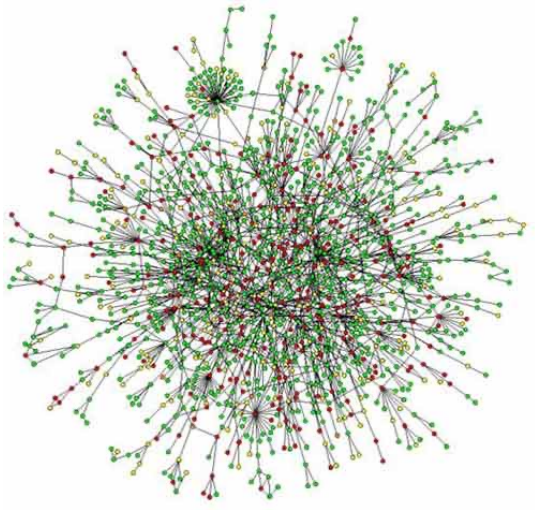
Molecular Interactions









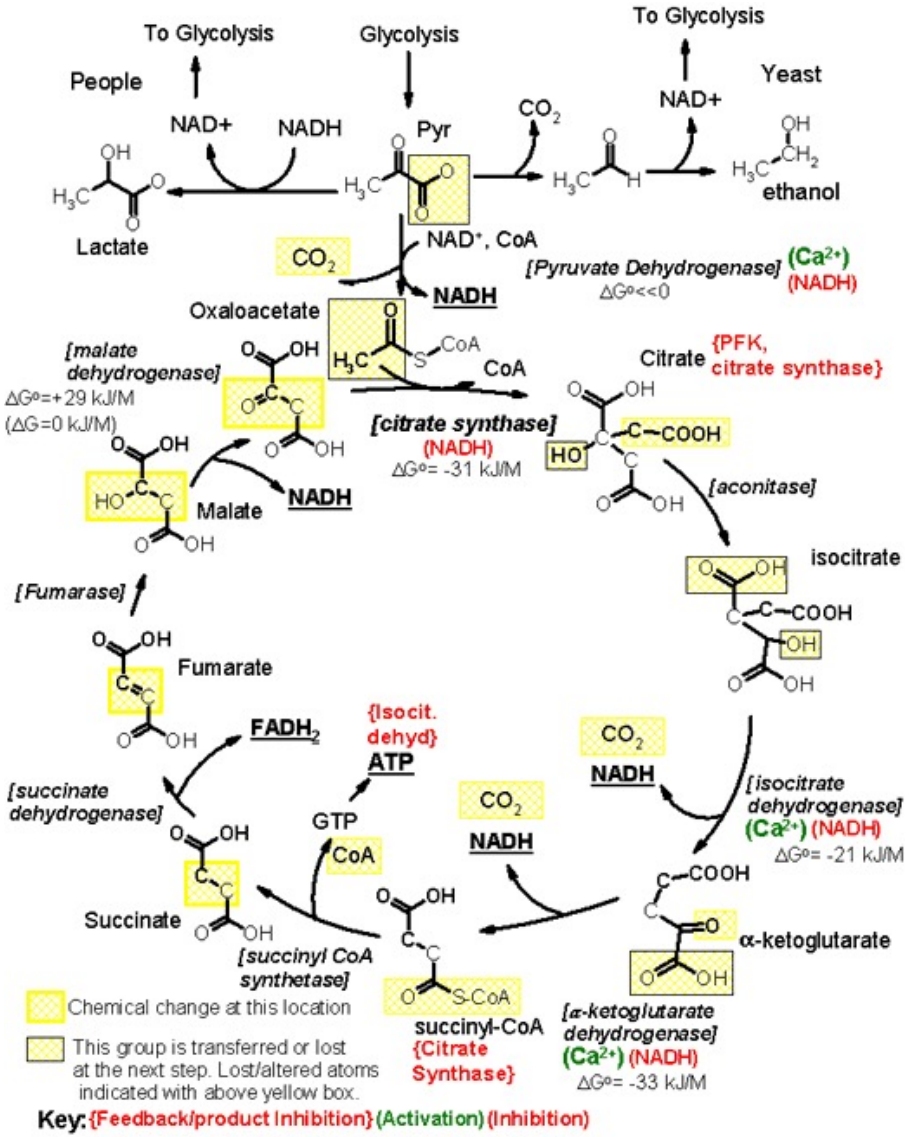


Interactions at different levels, multiple partners, and different partners.

Molecular Interactions: metabolic networks

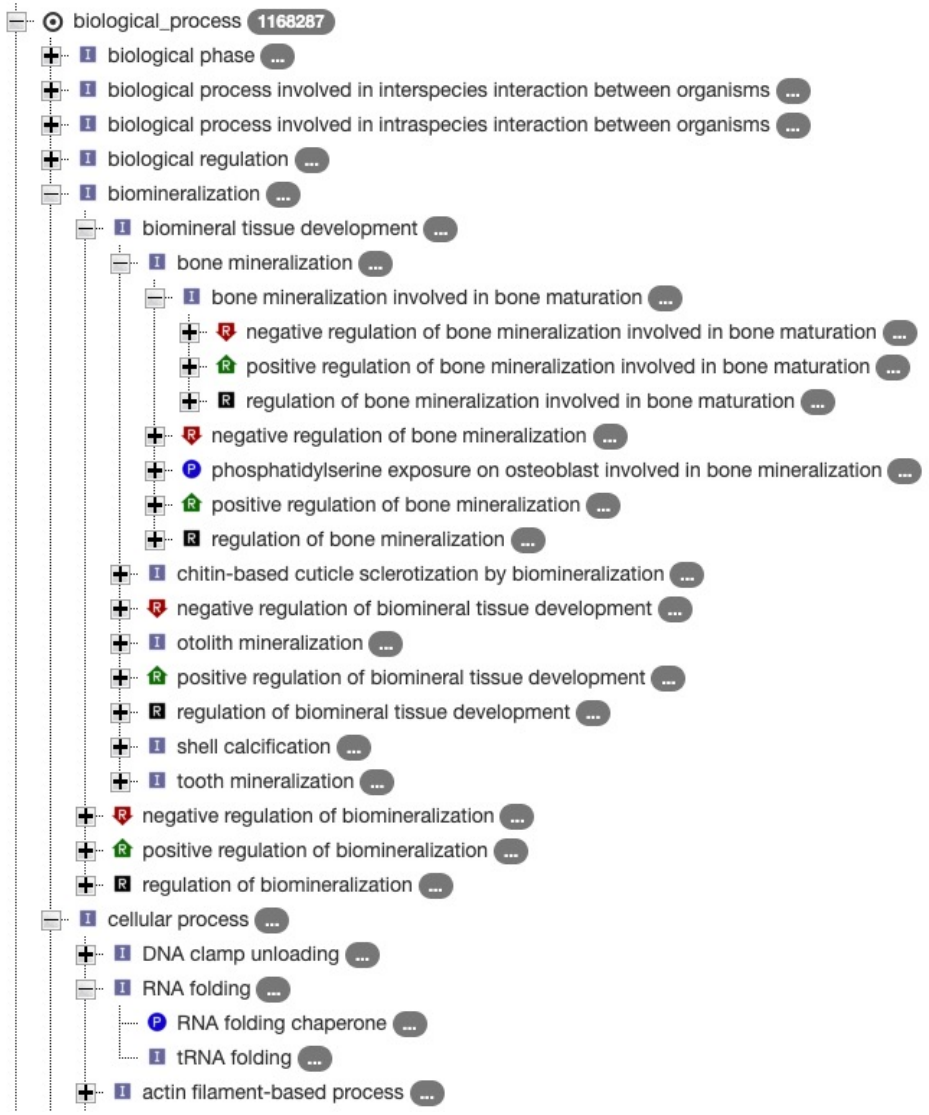
Metabolites connected by chemical reactions

(Citric acid cycle)



Graphs in biomedical sciences (less obvious ones)

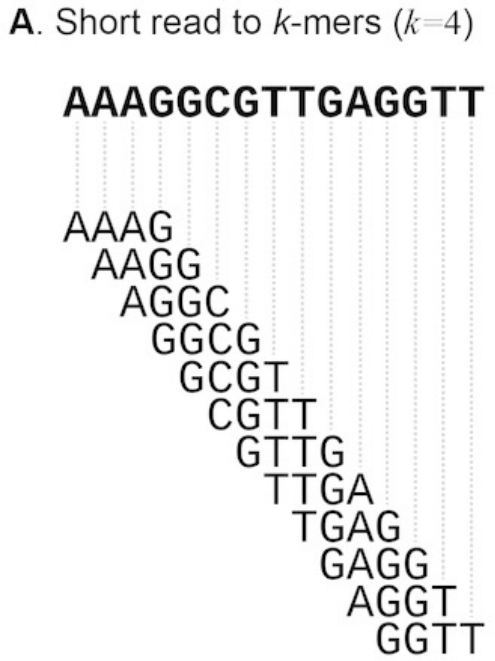
Biomedical ontologies are directed acyclic graphs (DAGs)

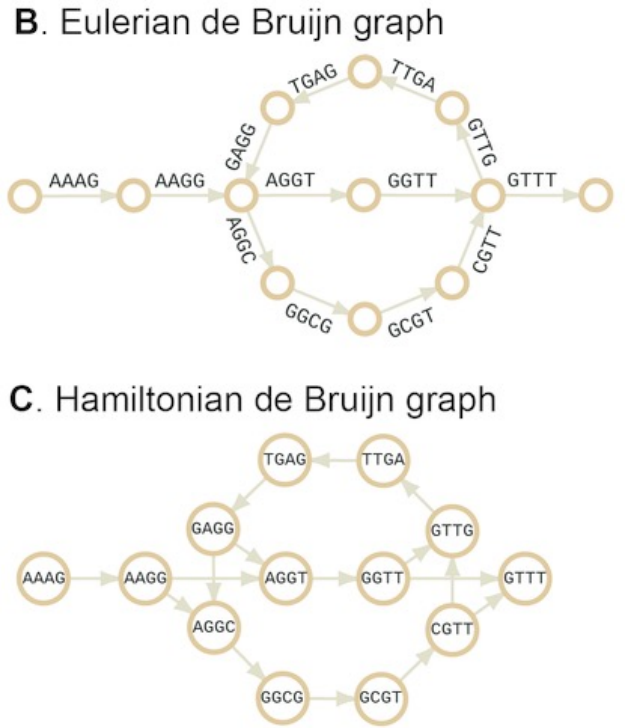


Graphs in biomedical sciences (less obvious ones)

Biomedical ontologies are DAGs

De Brujin graph used in sequence alignment





Sohn JI and Nam JW (2016): The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*

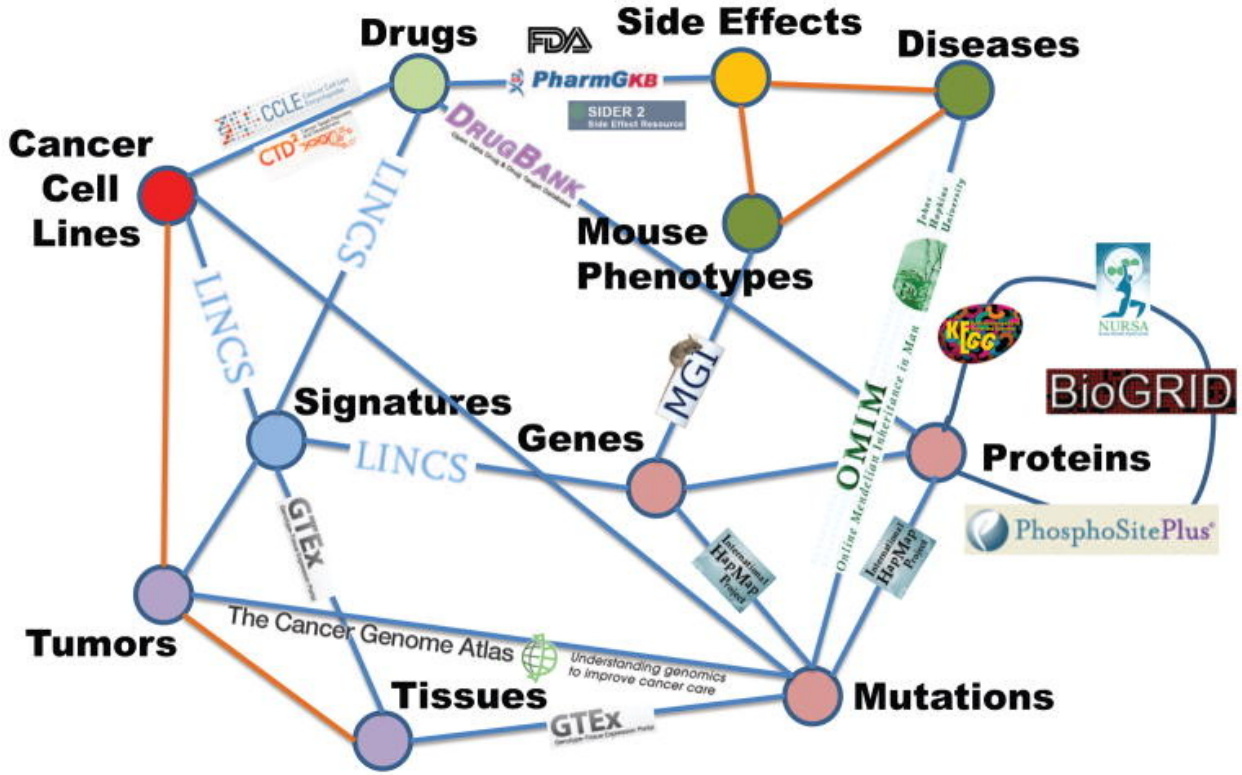
Graphs in biomedical sciences (less obvious ones)

Biomedical ontologies are DAGs

De Brujin graph for sequence alignment

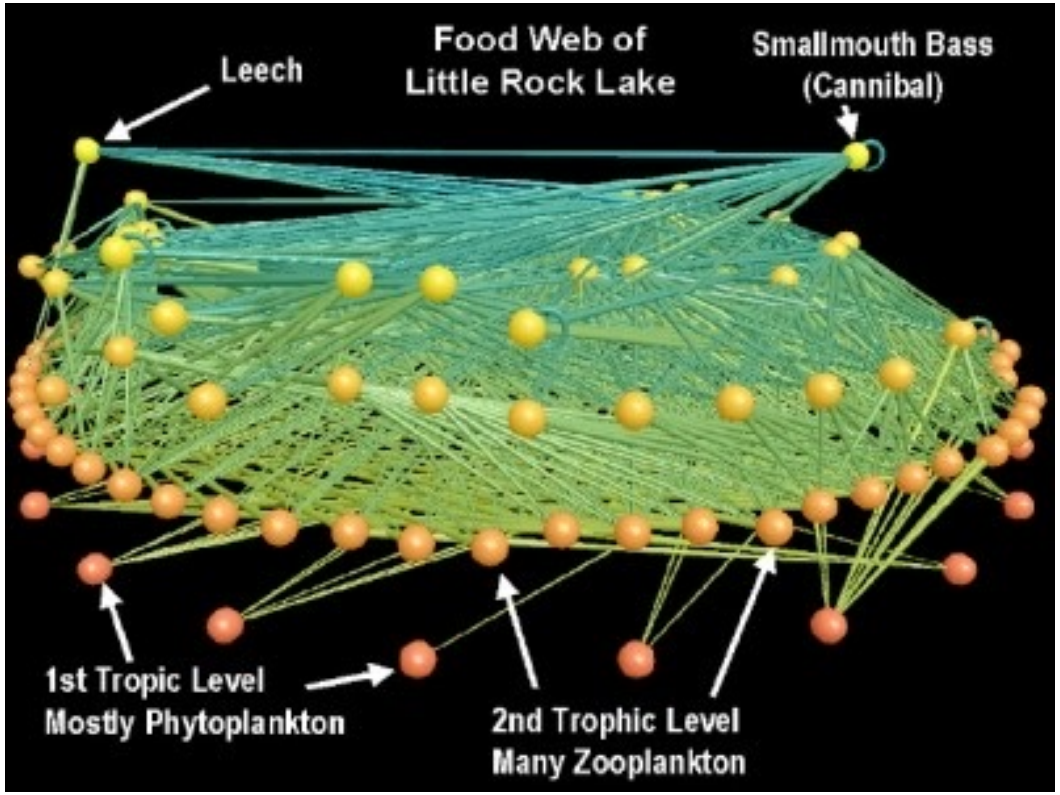
Integrative graphs across entities:

- Different entities can be connected by data and/or knowledge

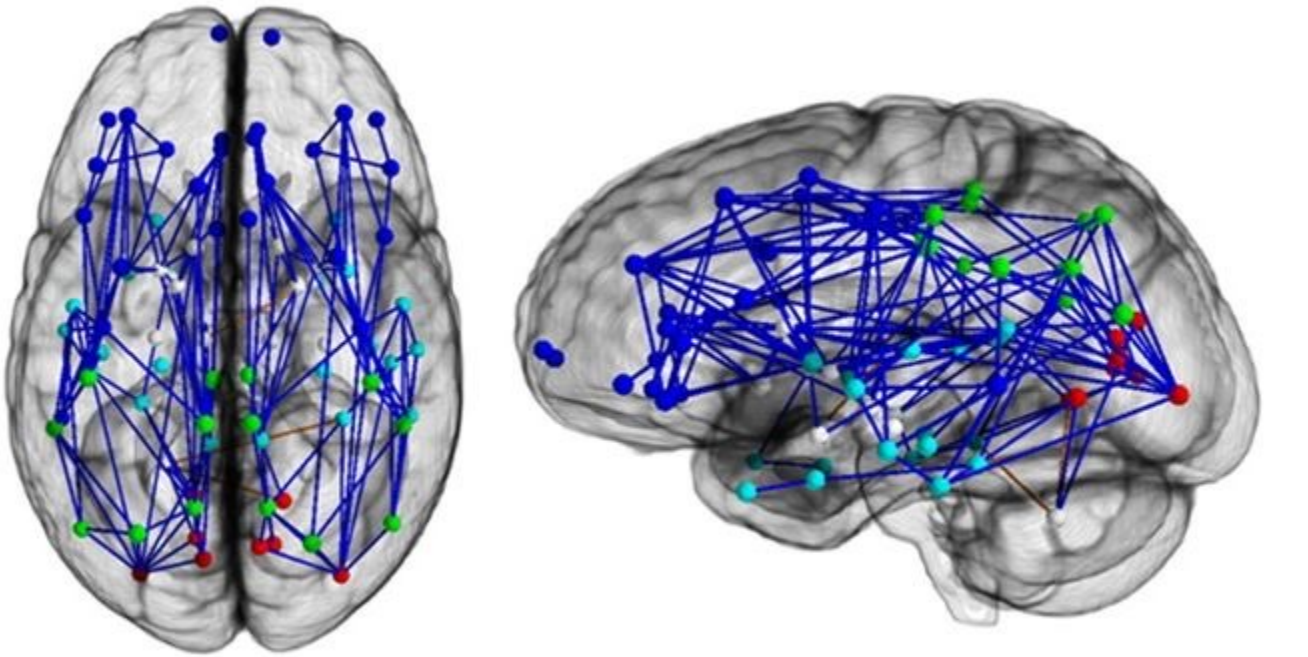


Ma'ayan A, Rouillard AD, Clark NR, Wang Z, et al. (2014): Lean Big Data Integration in Systems Biology and Systems Pharmacology. *Trends Pharmacol Sci.*

Graphs in biomedical sciences (less obvious ones): food web



Graphs in biomedical sciences (less obvious ones): Brain Networks



Courtesy Image

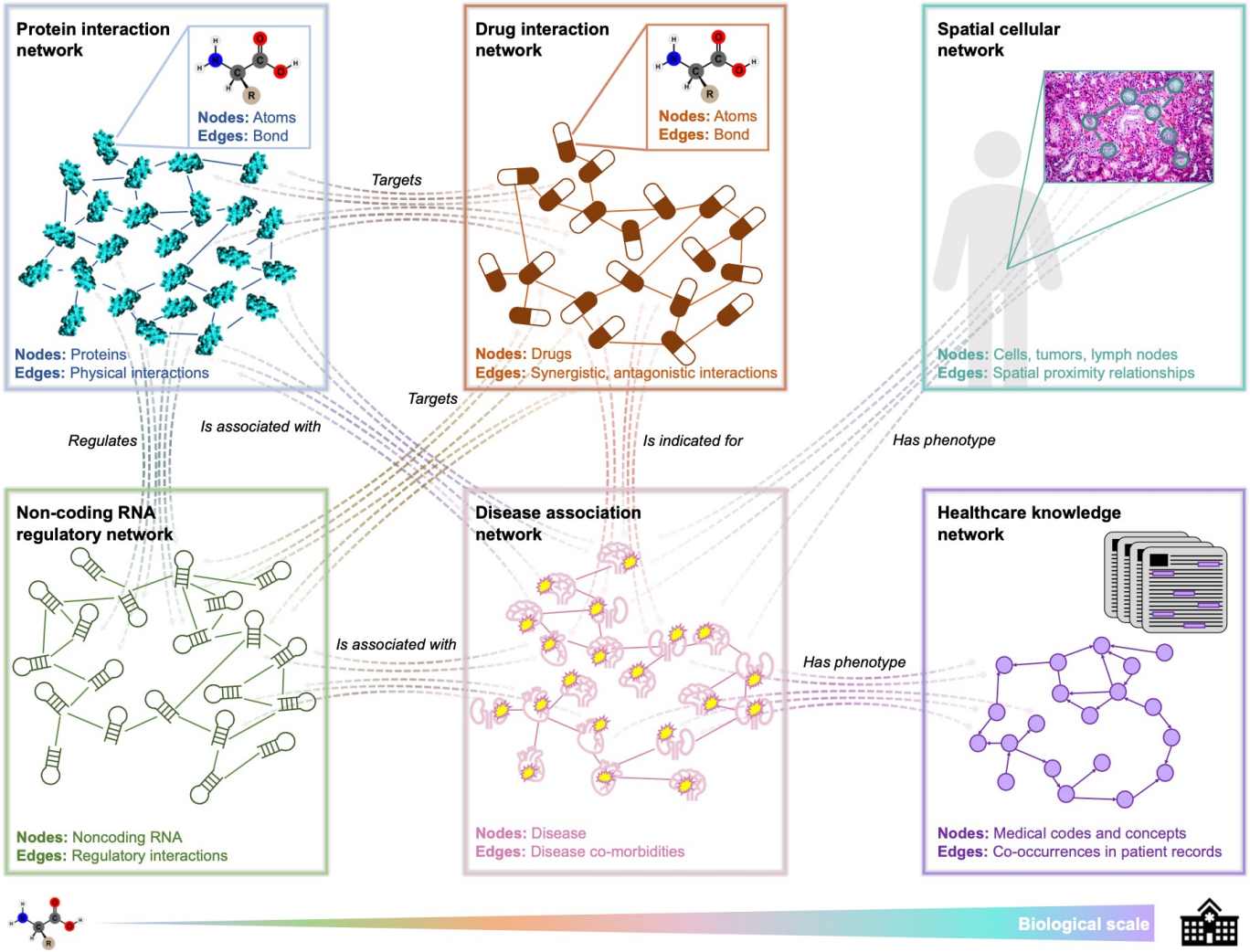
Graph ML applications in biomedicine

Graph representing:

- Molecules
- Omics
- Therapies
- Healthcare

Multi-scale and hierarchical

- Heterogenous graph connecting biomedical entities across multiple scales
- Graph of graphs (nodes are graphs)



Li MM, Huang K, Zitnik M (2021): Graph Representation Learning in Biomedicine. *arXiv:2104.04883*

Graph ML applications in biomedicine

Molecules

- **Property/function prediction**, generation of novel molecules, **molecular interaction prediction**

Omics

- Gene embeddings, perturbation effect prediction with RNA-seq data

Therapies

- **Drug repurposing**, adverse drug event predictions

Healthcare

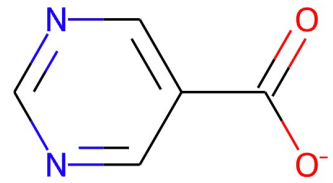
- Disease diagnosis with biomedical KG
- Clinical trial site selection/outcome predictions

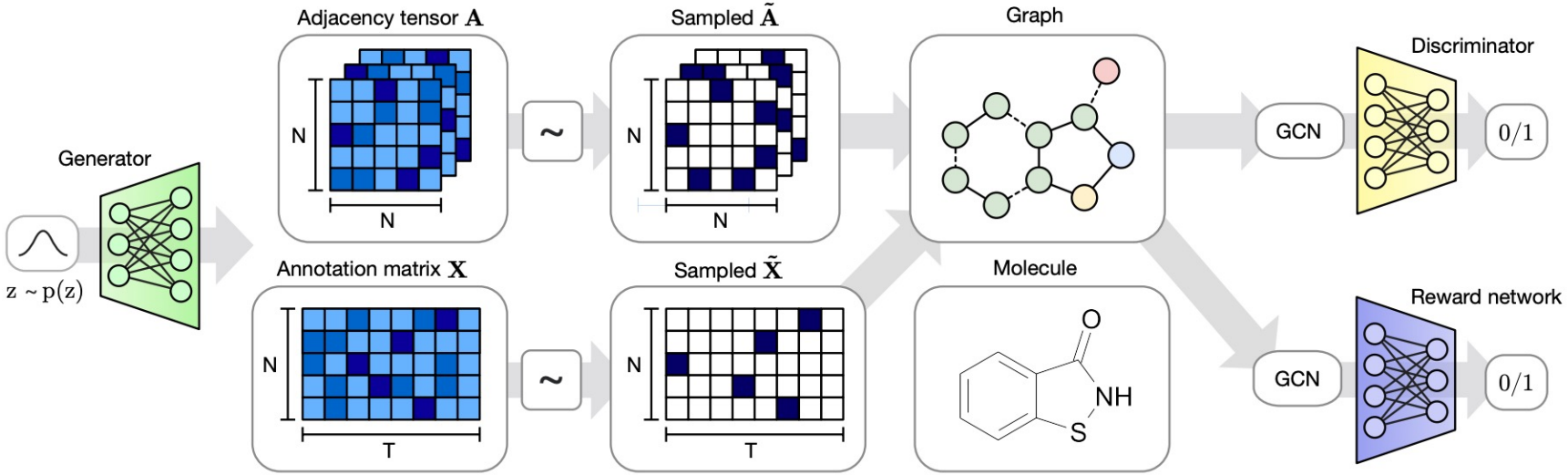
Molecule: generative models for *de novo* drug design

- SMILES string based
- ChemVAE (Gómez-Bombarelli et al., 2016)
- GrammarVAE (Kusner et al, 2017)

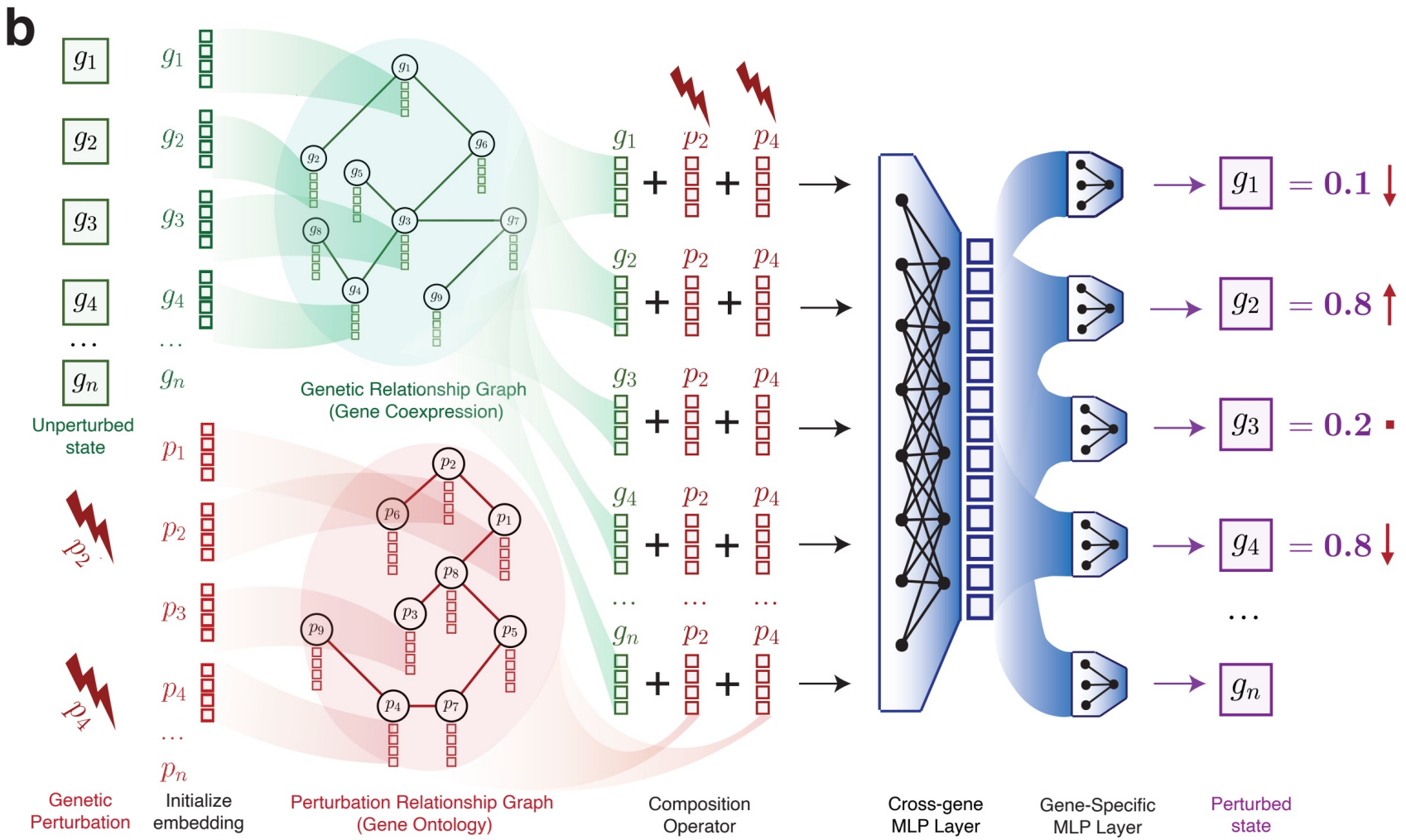
c1ncncc1C(=O)[O-]

- Graph-based
  - GraphVAE (Simonovsky and Komodakis, 2018)
  - DGMG: step-wise graph generation (Li et al. 2018)
  - MolGAN (De Cao and Kipf, 2018)
  - Masked graph model (MGM) (Mahmood et al., 2021)
  - Equivariant Diffusion Model (Hoogeboom et al., 2022)



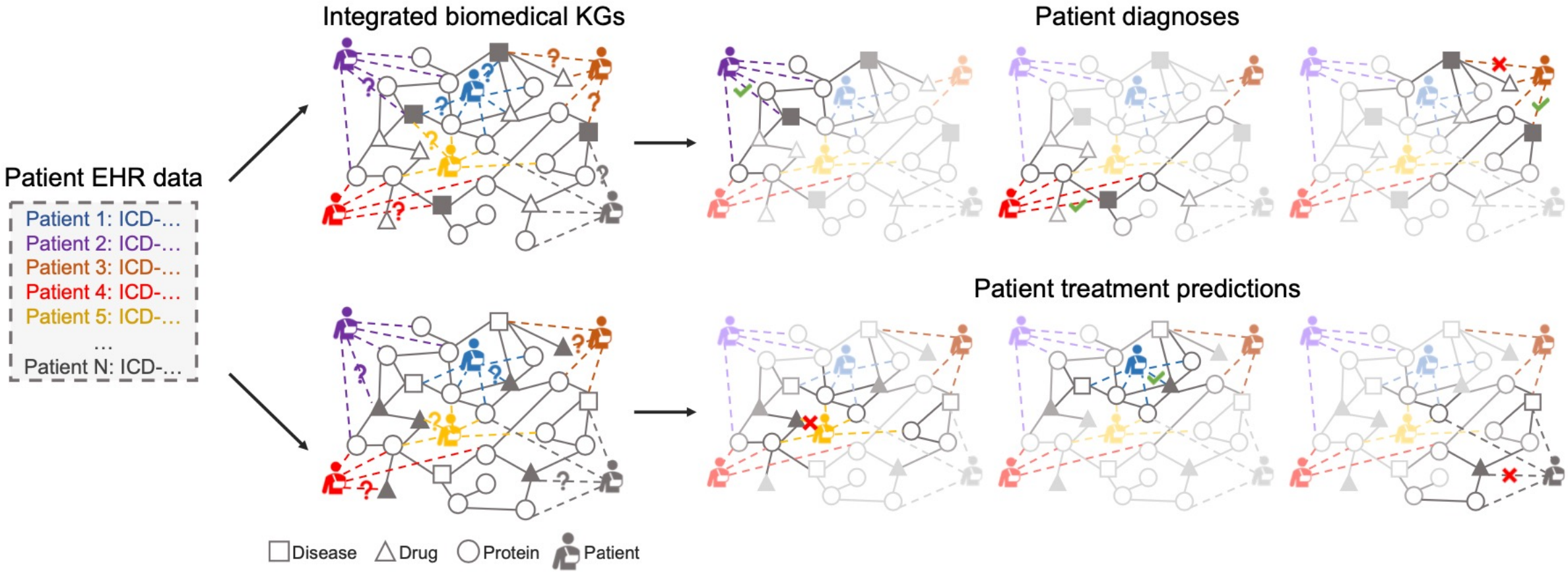


OMICS: Predicting perturbation effects using graphs of genes



Roohani Y, Huang K, Leskovec J (2022): GEARS: Predicting transcriptional outcomes of novel multi-gene perturbations. *bioRxiv 2022.07.12.499735v2*

Healthcare: integration of health data into knowledge graphs to predict patient diagnoses or treatments via link prediction



Li MM, Huang K, Zitnik M (2021): Graph Representation Learning in Biomedicine. *arXiv:2104.04883*

Healthcare: Clinical trial outcome prediction with heterogenous biomedical knowledge graph

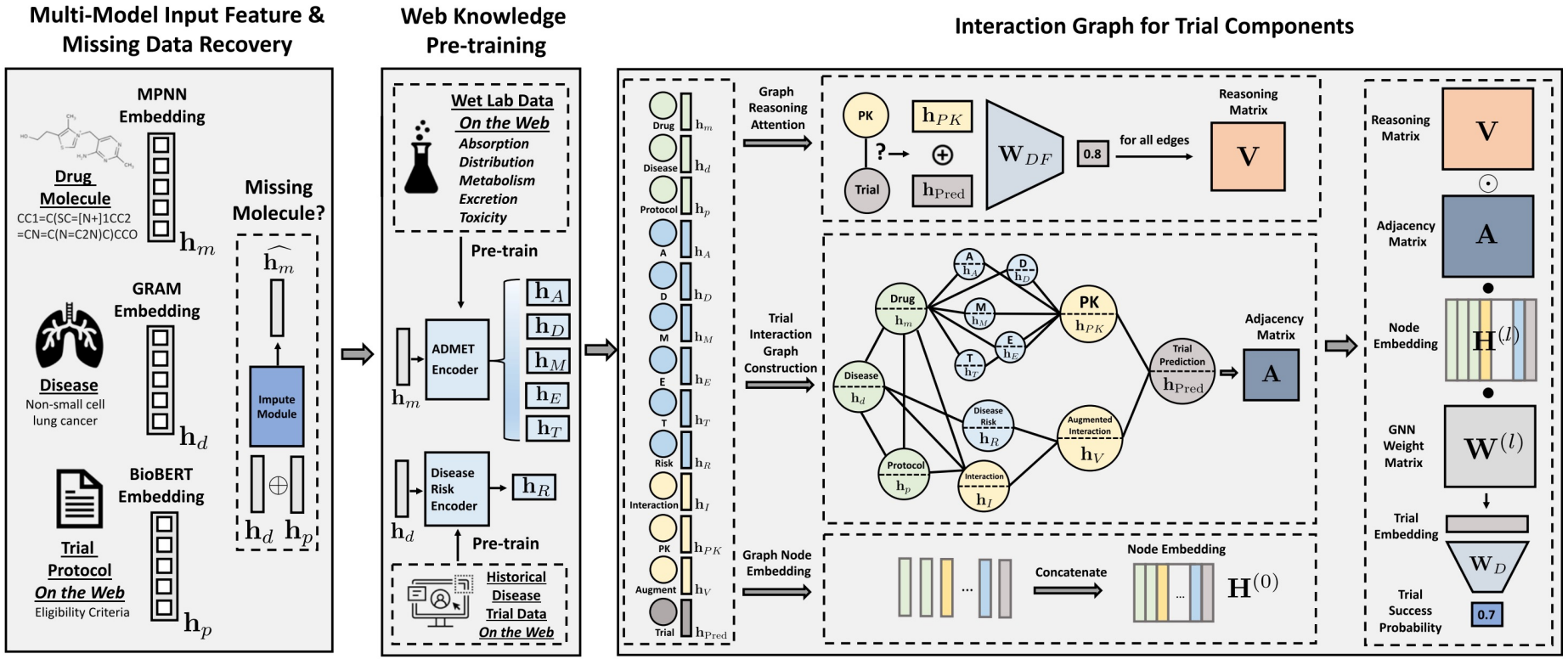


Figure 5. `HINT` **framework**

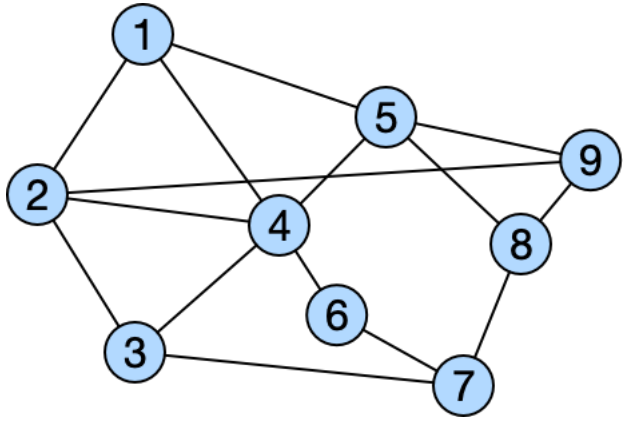
`HINT` is an end-to-end neural network pipeline with the following components: drug molecule embedding $\mathbf{h}_m$, disease embedding $\mathbf{h}_d$, and trial eligibility criteria embedding $\mathbf{h}_p$. Before constructing an interaction graph using these components, `HINT` pretrain some embeddings (blue nodes) using external knowledge about drug properties and disease risks. Then, we construct an interaction graph to characterize interactions between various trial components. Trial embeddings are learned based on the interaction graph to capture both trial components and their interactions. Based on the learned representation and the dynamic attentive graph neural network (Equation 13), we make trial outcome predictions.

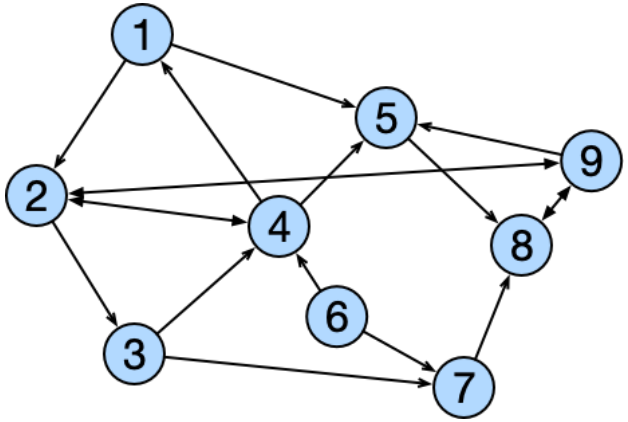
Fu T et al. (2022): HINT: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns*

Refresher on graph theories: notations

A graph is a set of nodes and edges $\mathrm{G} = \{\mathrm{V}, \mathrm{E}\}$

- Nodes == Vertices == Entities:
  - $V = \{1, \ldots, \mathrm{n}\}$
- Edges == Links == Relations:
  - Undirected graphs: $E = \{\{i, j\}: \ i, j \in V\} \subseteq V \times V$
    - Node degree is the number of incident edges
  - Directed graphs: $E = \{(\mathrm{i}, \mathrm{j}) : i, \mathrm{j} \in V\} \subseteq V \times V$
    - In-degree: number of incoming edges
    - Out-degree: number of outgoing edges





Refresher on graph theories: neighborhood and features

Neighborhood:

- $\mathcal{N}(i) = \{j: (i,j) \in E\}$

Degree:

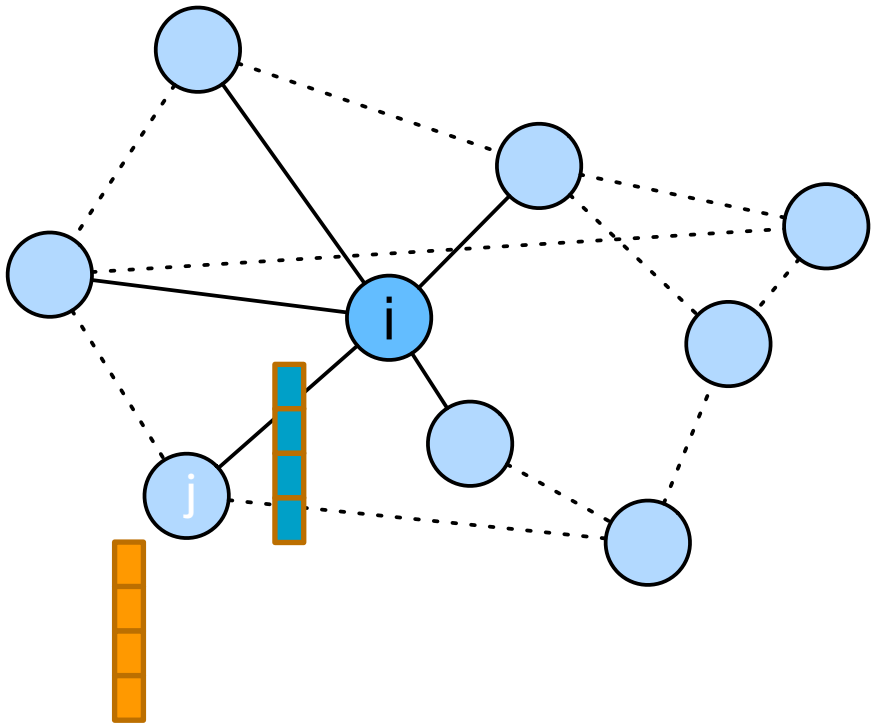
- $d_i = |\mathcal{N}(i)|$

Node features:

- $x : \mathcal{V} \rightarrow \mathbb{R}^k$
- $\mathrm{X} = (x_1, \ldots, x_n)^T$

Edge features:

- $e : E \rightarrow \mathbb{R}^{k'}$



Refresher on graph theories

Adjacency matrix $\mathrm{A} \in \mathbb{R}^{n \times n}$ encodes graph structure

- $\mathrm{A_{ij}} = 1 \Leftrightarrow$ there is an edge (i, j); 0 otherwise
- Symmetric for undirected graphs
- Adjacency Matrix for weighted graph stores the weights of the edges.

Laplace matrix $\mathrm{L} = \mathrm{diag}(d_1, \ldots, d_n) - \mathrm{A}$

- Relates to many useful graph properties: spectral decomposition and graph-based signal processing.

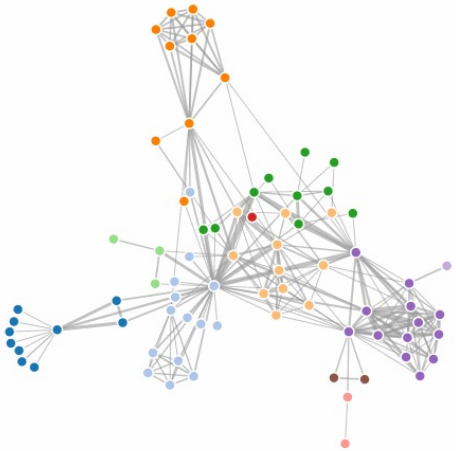
Refresher on graph theories

Dimensionality reduction:

- Spectral embeddings (aka Laplacian eigenmaps), performs PCA on L

Graph visualization/layout/drawing:

- Computes the low-dim coordinates of nodes from a graph to preserve their local and global topological structures.
- Force-based layout (e.g. Fruchterman-Re algorithm)
  - Attractive forces between connected nod
  - Repulsive forces among all nodes



Refresher on graph theories: random graph generation and biology-inspired graph models

Erdős–Rényi model

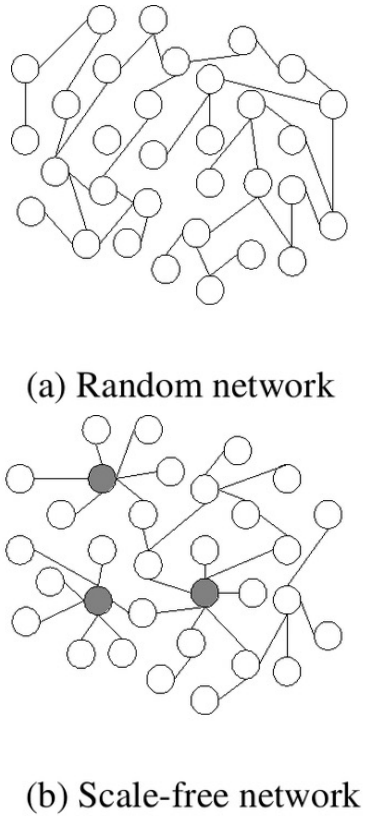
- Generating graph by sampling edges from all possible edges uniform randomly

Barabási–Albert model

- Generating scale-free network with preferential attachment mechanism
  - Scale-free network: degree distribution follows a power law $p(d_i = k) \sim k^{-\gamma}$
  - Preferential attachment: a new node connecting to existing nodes in the network is proportional to their degrees

Duplication-divergence model

- Genes duplicate and diverge to gain or lose functions in evolution
- Nodes duplicate and inherit edges from parent nodes
- New nodes diverge by gaining and losing inherited edges



(a) Random network

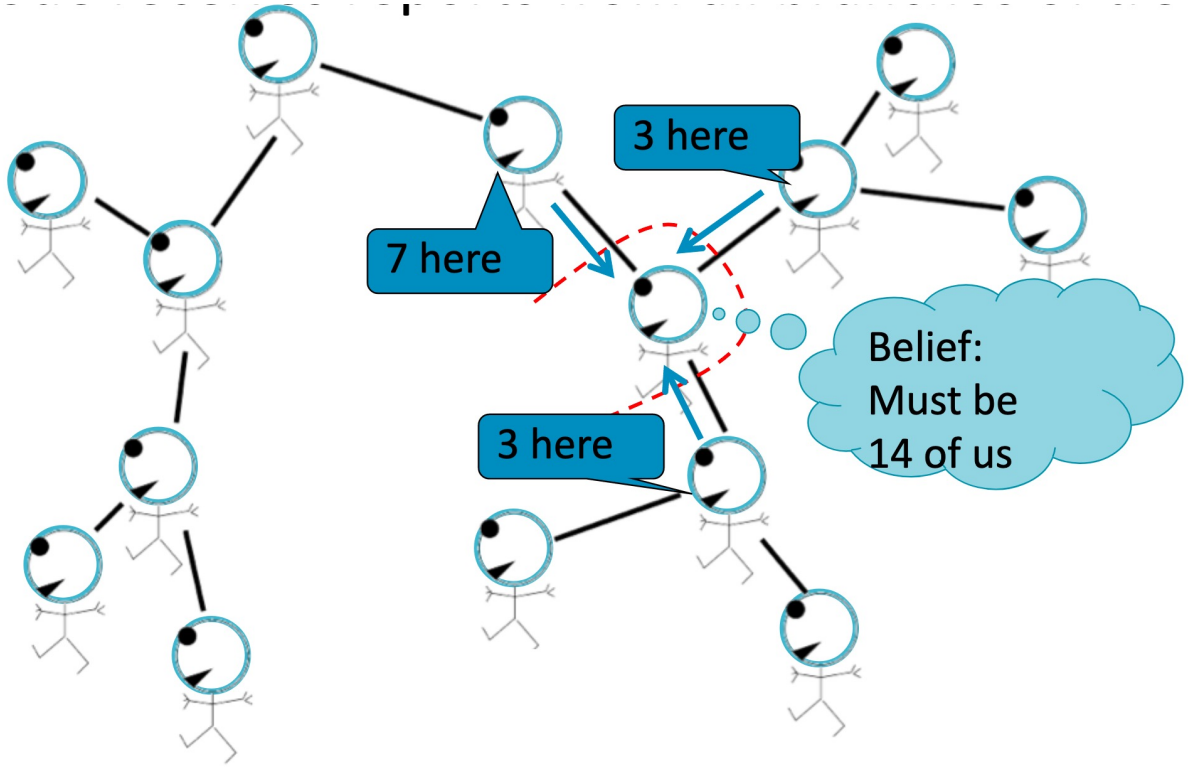
(b) Scale-free network

Barabási, AL, Albert R (1999): Emergence of scaling in random networks. *Science*
Ispolatov I, Krapivsky PL, Yuryev A (2005): Duplication-divergence model of protein interaction network. *Phys Rev E*

Machine Learning on Graphs: GNN basics

Message passing:

- Task: Count the number of nodes in the graph
- Condition: Each node only interacts with its neighbors

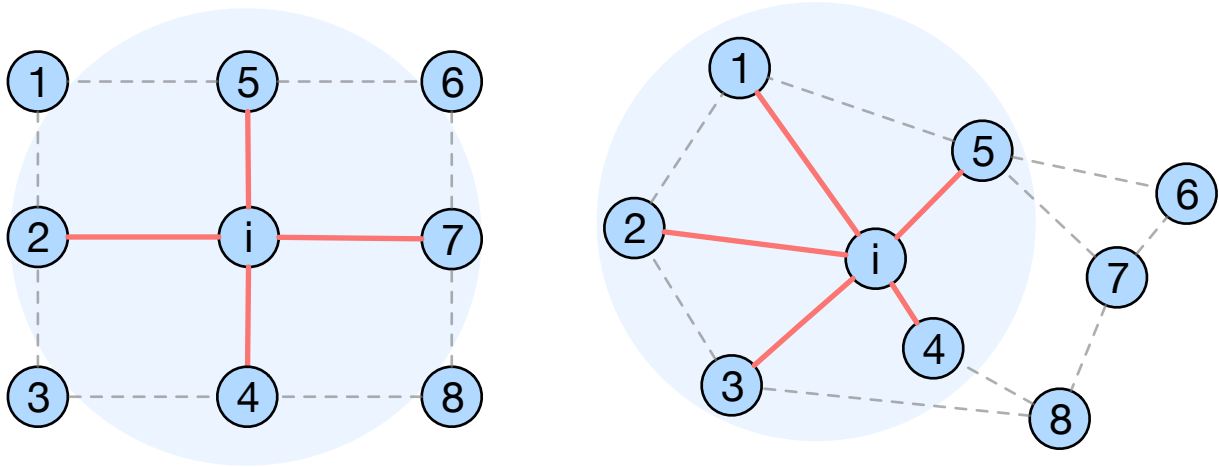


Machine Learning on Graphs: GNN basics

Message passing:

- Message: $\boldsymbol{m}_{ij} = M(\boldsymbol{h}_i, \boldsymbol{h}_j, \boldsymbol{e}_{ij})$
- Aggregation: $\boldsymbol{m}_i = AGG_{j \in \mathcal{N}(i)} \boldsymbol{m}_{ij}$, should be permutation invariant
- Update: $\boldsymbol{h}_i \leftarrow U(\boldsymbol{h}_i, \boldsymbol{m}_i)$

Learn node-level representations $\boldsymbol{h}_i$



Machine Learning on Graphs: levels of tasks

Node: $f(\boldsymbol{h}_i)$

- Node classification/regression
- Node clustering, community detection

Edge: $f(\boldsymbol{h}_i, \boldsymbol{h}_j)$

- Link prediction

(sub)Graph: $f(AGG_{i \in V} \boldsymbol{h}_i)$

- Graph classification/regression

Multi-graph: $f\left(AGG_{i \in V_1} \boldsymbol{h}_i, AGG_{j \in V_2} \boldsymbol{h}_j\right)$

Hands-on set up

Head to https://dashboard.eventengine.run/login and enter the event engine hash **f843-10d7089734-9a**. You will be asked to login with an email to receive the OTP to get an AWS Account. Follow the instructions on the website:

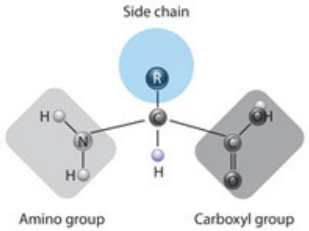
1. Click on AWS Console
2. On the window popup, select Open Console. This will open an AWS Console. The AWS Management Console is a browser-based GUI for Amazon Web Services (AWS). Through the console, a customer can manage their cloud computing, cloud storage and other resources running on the Amazon Web Services infrastructure.
3. On the top right, ensure N.Virginia (us-east-1) is selected. If for any reason, you are logged into the different region, please switch to N.Virginia.
4. On the top of the Console, type in SageMaker in the search bar
5. On the left sidebar, head to Notebook > Notebook instances
6. You will see an instance set up. Select Open Jupyter on the right side of the page under Actions. This will open a Jupyter notebook interface hosted on Amazon SageMaker

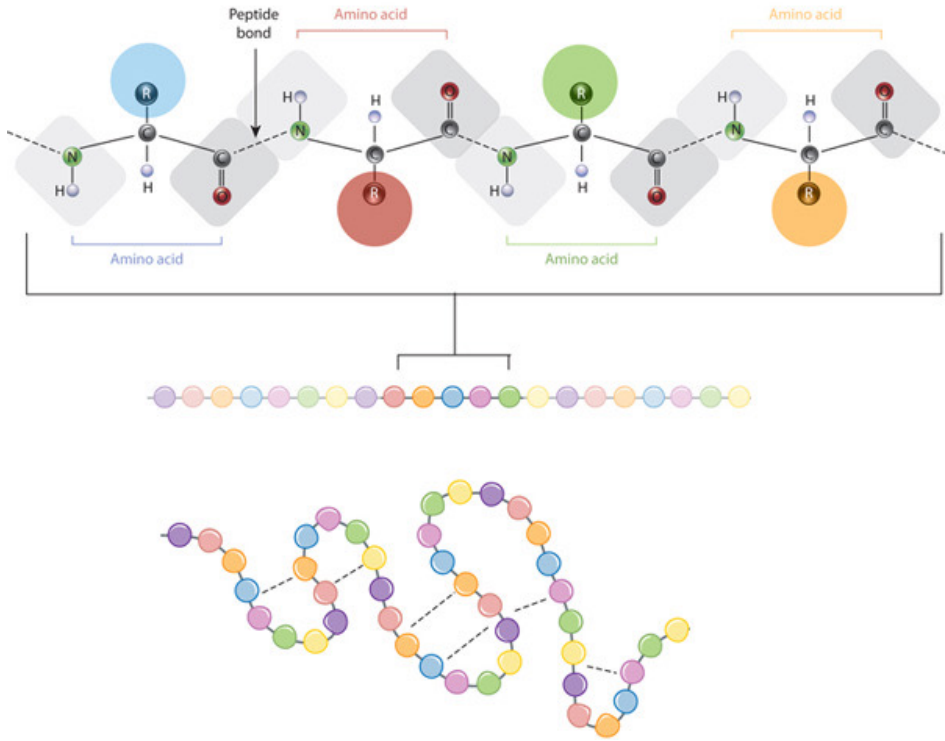
Section 2: Making sense of small molecules with GNNs

Section 3: Making sense of macro-molecules with GNNs

Introduction - Proteins

- Major macromolecules carrying out the essential functions in biology
- Levels of protein structures
    - Primary: sequence of amino acids (AAs) connected by peptide bonds. 20 types of common AAs.





Essentials of Cell Biology (2010)
https://www.nature.com/scitable/ebooks/essentials-of-cell-biology-14749010/122996920/

Introduction - Proteins

- Major macromolecules carrying out the essential functions in biology
- Levels of protein structures
    - Primary: sequence of amino acids (AAs) connected by peptide bonds
    - Secondary: alpha-helix, beta-sheet
    - Tertiary: 3D structure of individual peptide chain
    - Quaternary: 3D structure of protein complex (>= 2 peptide chains)
- Protein folding: the process of primary sequence folding into tertiary structure. Sequence *usually* determines structure.
- Tertiary structures dictate functions

Essentials of Cell Biology (2010)
https://www.nature.com/scitable/ebooks/essentials-of-cell-biology-14749010/122996920/

Related work – Protein LMs

- Protein sequences are sentences of amino acids

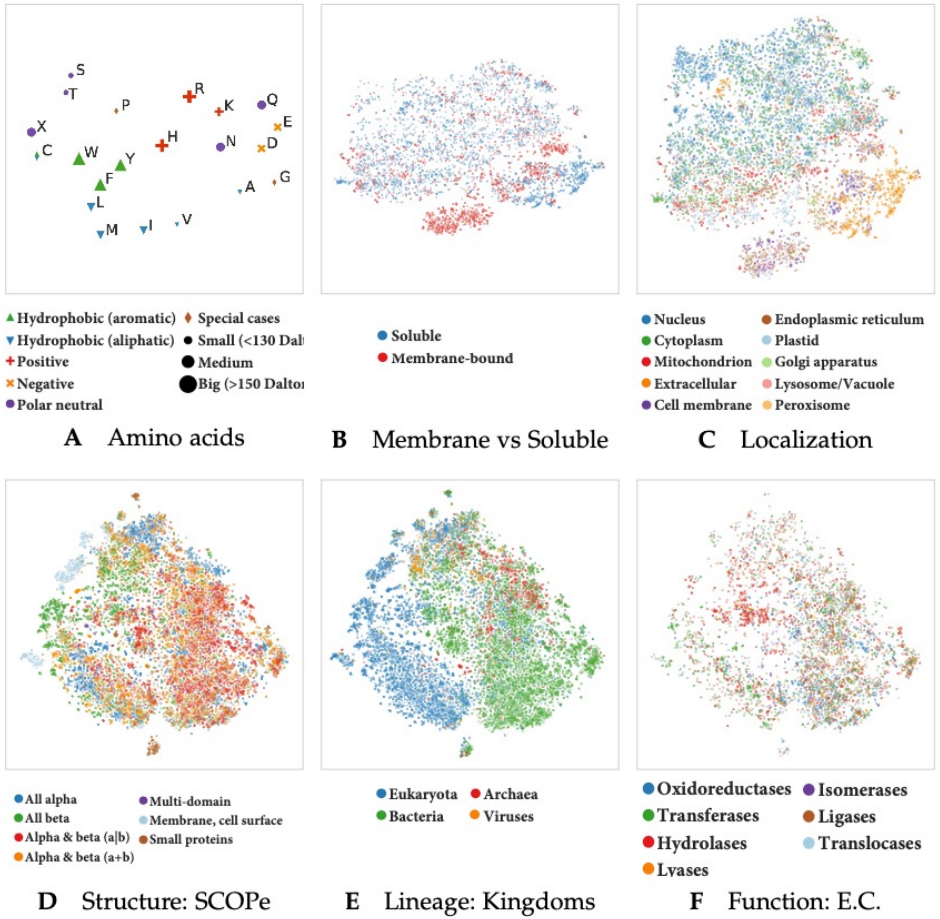
```
>1E2Q_1|Chain A|THYMIDYLATE KINASE|HOMO SAPIENS (9606)
GSHMAARRGALIVLEGVDRAGKSTQSRKLVEALCAAGHRAELLRFPERSTEIGKLLSSYLQKKSDVEDHSVHLLFSANRWEQV
PLIKEKLSQGVTLVVDRYAFSGVAFTGAKENFSLDWCKQPDVGLPKPDLVLFLQLQLADAAKRGAFGHERYENGAFQERALRC
FHQLMKDTTLNWKMVDASKSIEAVHEDIRVLSEDAIATATEKPLGELWK
```

- Protein LMs can be trained using Masked LM (MLM) or auto-regressive objectives.
    - MLM: ProtBert[1], ProtAlbert[1], ESM[2]
    - Auto-regressive: ProtTXL[1], ProtXLNet[1]
- Some protein LMs are bigger than natural LMs because of bigger datasets:
    - UniRef: 216M sequences (22x English Wikipedia)
    - BFD: 2.1B sequences (112x English Wikipedia)

[1] Elnaggar *et al.* (2020) *arXiv:2007.06225*
[2] Rives *et al.* (2021) *PNAS*

Related work – Protein LMs

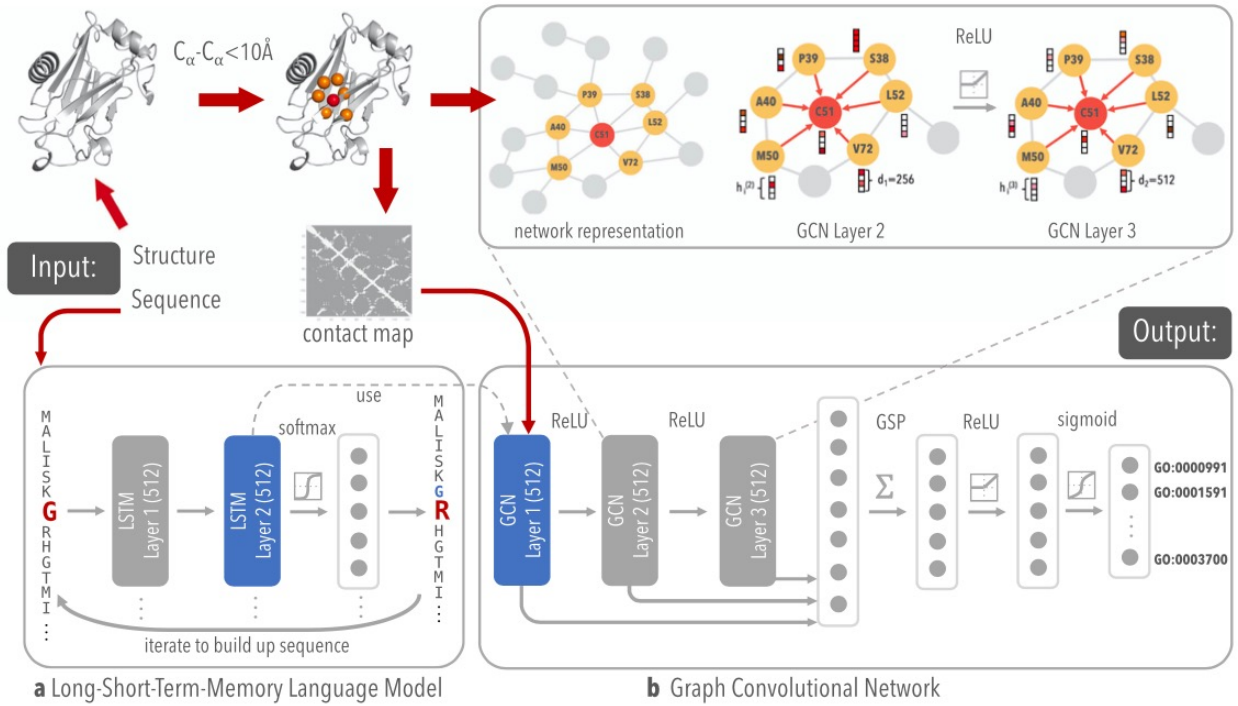
- Protein LMs capture various features of proteins



Elnaggar *et al.* (2020) "ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing" *arXiv:2007.06225*

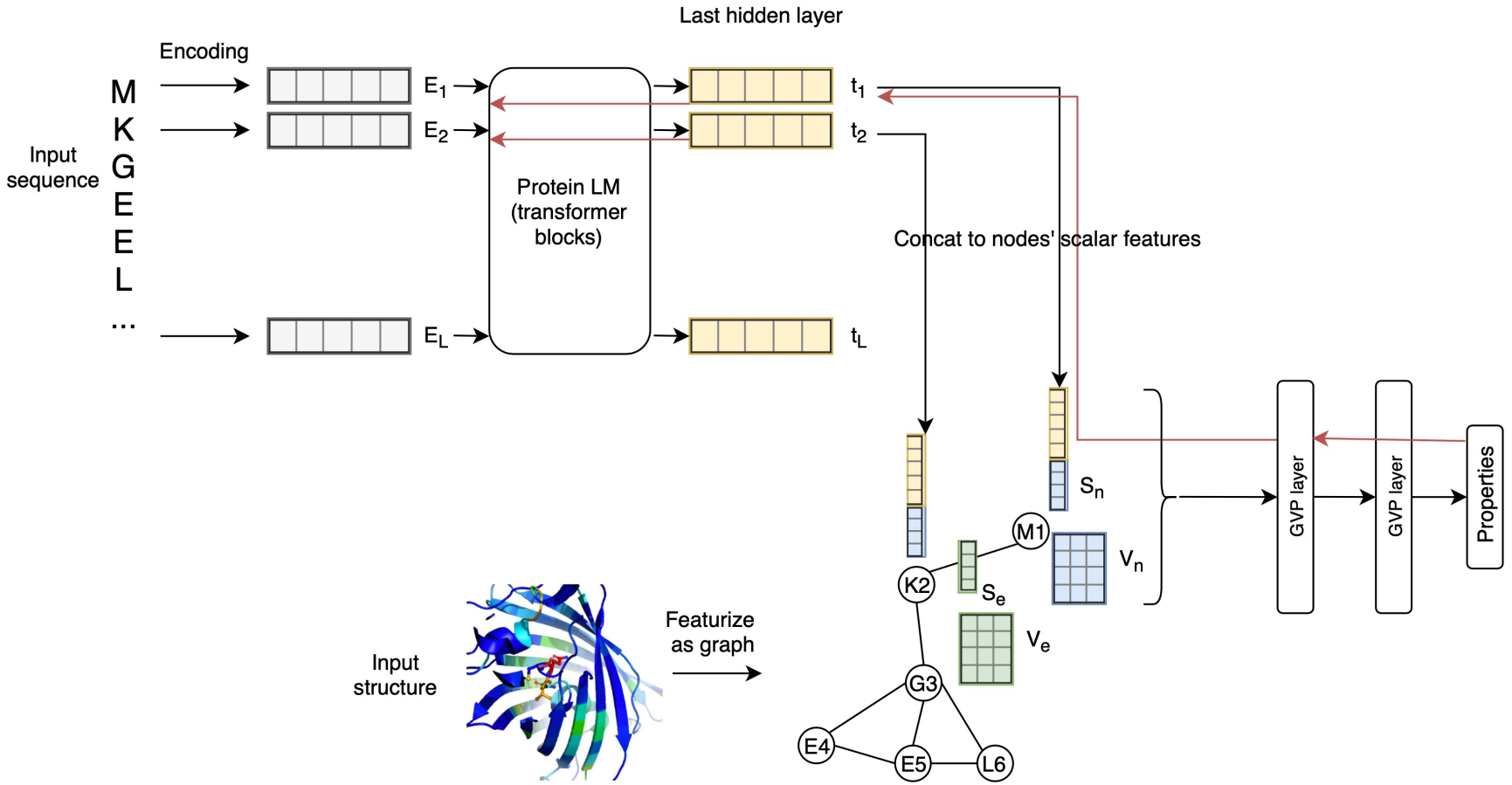
Related work – Modeling of protein structures using GNN

- DeepFRI:
  - Uses embeddings of AAs from protein LM as node features
  - Converts protein structure to graphs of AAs
  - Trains GNN to predict protein properties



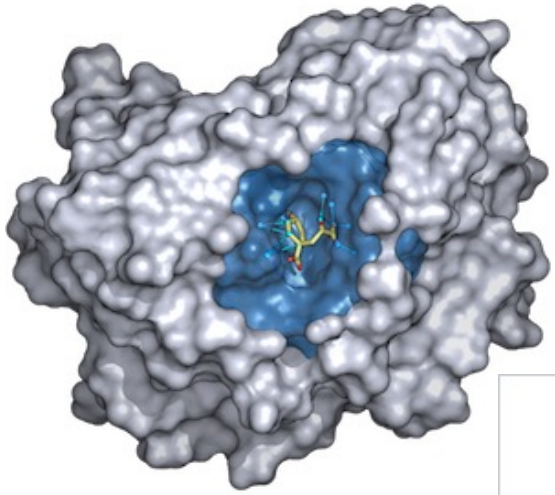
Gligorijević *et al.* (2021) "Structure-based protein function prediction using graph convolutional networks" *Nature Comm.*

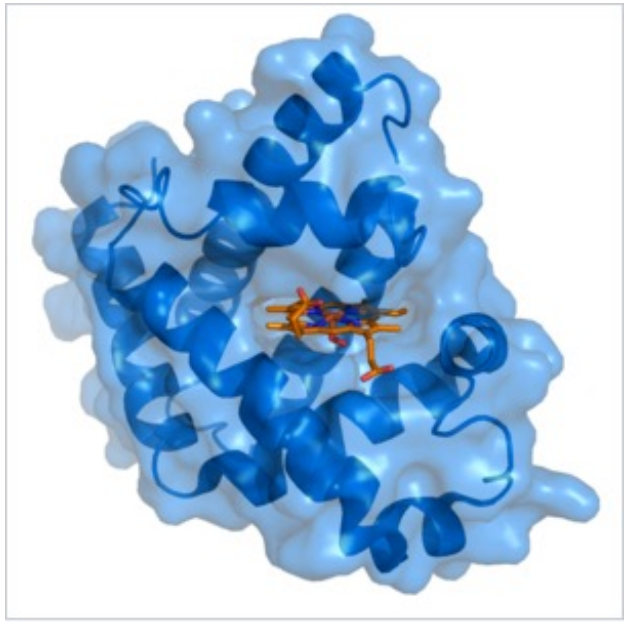
Our method: LM-GVP



Wang, Combs, Brand *et al*. (2022) "LM-GVP: an extensible sequence and structure informed deep learning framework for protein property prediction" *Sci. Rep.*

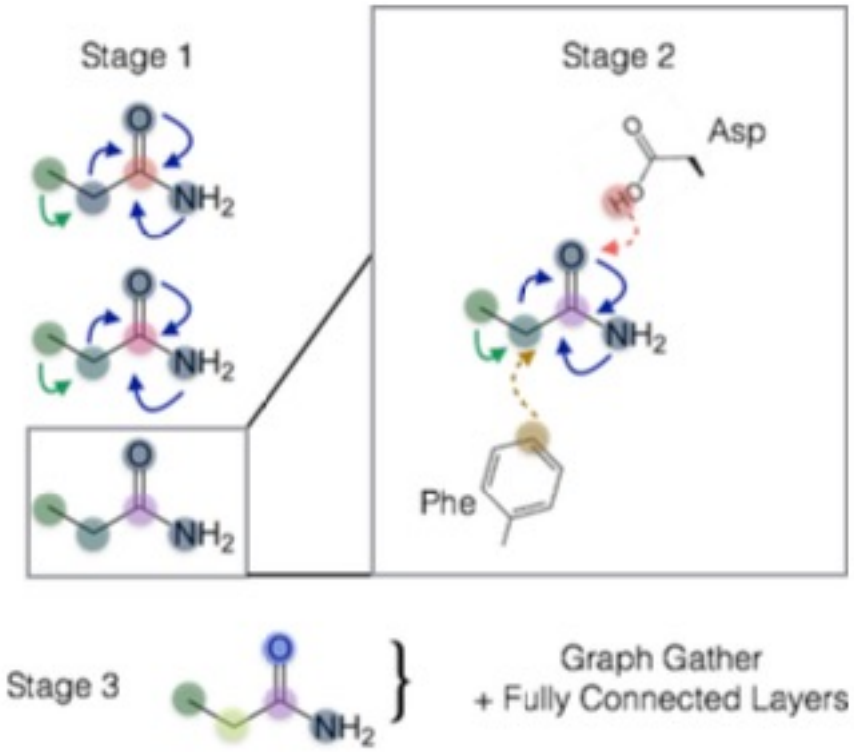
Section 4: Going beyond single graph, bi-graph based binding affinity prediction for protein-ligand pairs





```
#  ==========================================================================================
# List of the protein-ligand complexes in the PDBbind refined set v.2020
# 5316 protein-ligand complexes in total, which are ranked by binding data
# Latest update: July 2021
# PDB code, resolution, release year, -logKd/Ki, Kd/Ki, reference, ligand name
#  ==========================================================================================
2r58  2.00  2007   2.00  Kd=10mM       // 2r58.pdf (MLY)
3c2f  2.35  2008   2.00  Kd=10.1mM     // 3c2f.pdf (PRP)
3g2y  1.31  2009   2.00  Ki=10mM       // 3g2y.pdf (GF4)
3pce  2.06  1998   2.00  Ki=10mM       // 3pce.pdf (3HP)
4qsu  1.90  2014   2.00  Kd=10mM       // 4qsu.pdf (TDR)
4qsv  1.90  2014   2.00  Kd=10mM       // 4qsv.pdf (THM)
4u54  2.41  2015   2.06  Kd=8.7mM      // 4u54.pdf (3C5)
3ao4  1.95  2011   2.07  Kd=8.5mM      // 3ao1.pdf (833)
4cs9  2.01  2014   2.10  Kd=8mM        // 4cs8.pdf (AMP)
2w8w  2.14  2009   2.12  Kd=7.5mM      // 2w8j.pdf (PLS)
3gv9  1.80  2009   2.12  Ki=7.5mM      // 3gqz.pdf (GV9)
6r9u  1.26  2019   2.12  Kd=7.5mM      // 6r8l.pdf (JVQ)
6abx  1.70  2019   2.14  Ki=7.19mM     // 6abx.pdf (FLC)
4q90  1.54  2015   2.15  Ki=7.0mM      // 4q7p.pdf (4H2)
5cs3  2.50  2015   2.15  Kd=7.0mM      // 5cs3.pdf (EP1)
4tim  2.40  1992   2.16  Ki=6.9mM      // 4tim.pdf (2PG)
5fe6  1.77  2016   2.16  Kd=6910uM     // 5fdz.pdf (5WZ)
6ghj  2.26  2018   2.16  Kd=6.89mM     // 6ghj.pdf (3-mer)
3gqz  1.80  2009   2.17  Ki=6.7mM      // 3gqz.pdf (GF7)
```

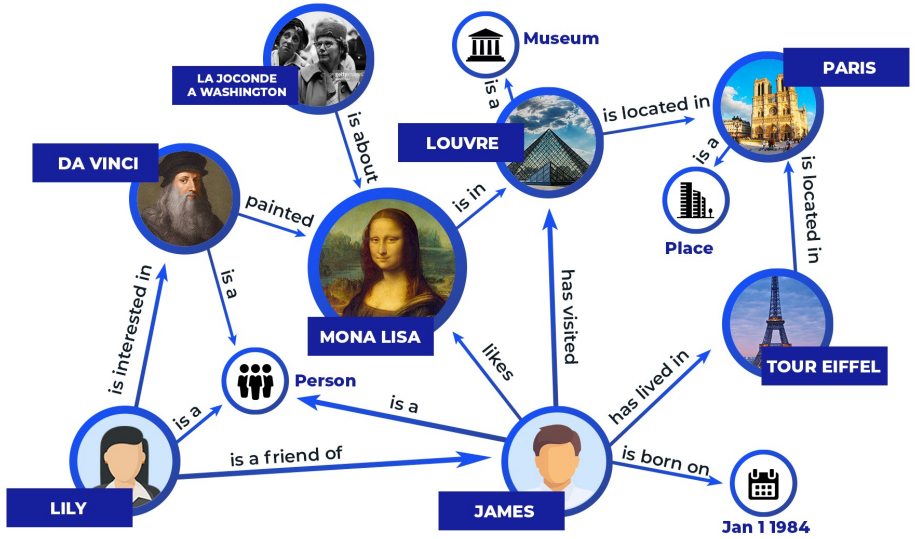
PotentialNet : Binding affinity prediction

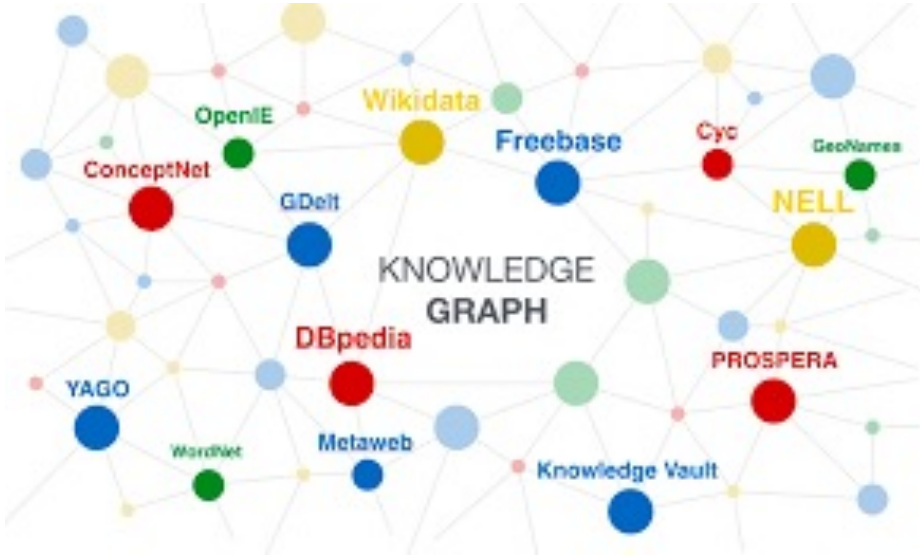


Section 5: Organizing and generating new knowledge for drug discovery and repurposing with knowledge graphs

Knowledge graphs

- Collection of entities and relations connecting entities





- Major knowledge graph (KG) efforts by Google, Amazon, Apple, and Facebook

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo et al. "Knowledge graphs." *arXiv preprint arXiv:2003.02320* (2020).

Knowledge graph use cases

- Google KG powers and enriches search results with related entities
- IBM’s question answering (QA) system Watson, which was able to beat human experts
- Siri, Cortana, Alexa, or Google Now operate using KGs
- Bio2RDF and LinkedLifeData are bio KGs used for QA and decision support in the life sciences.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo et al. "Knowledge graphs." *arXiv preprint arXiv:2003.02320* (2020).

Knowledge graphs as heterogeneous graphs

- Heterogeneous graph (unweighted) $\mathcal{G} := \{\{\mathcal{V}_t\}_{t=1}^{T}, \{\mathcal{E}_r\}_{r=1}^{R}\}$
  - Node-types $\mathcal{V}_t := \{v_n^t\}_{n=1}^{N_t}$
  - Relation-types $\mathcal{E}_r := \{(v_n^t, v_n^{t'}) \in \mathcal{N}^t \times \mathcal{N}^{t'}\}$

- Every edge in *G* represented as a triple $(v_n^t, r, v_n^{t'})$

  (Covid, causes, ARDS)

  Covid $\in$ Disease

  ARDS $\in$ Symptom

  (Vassilis, is, Greek)

  Vassilis $\in$ People

  Greek $\in$ Nationality

ARDS: acute respiratory distress syndrome

- Billions of nodes and edges

Knowledge graph construction

- KG construction approaches
  - Curated by team of experts
  - Automated by natural language processing and machine learning models typically applied to large documents

Leonard Nimoy was an actor who played the character Spock in the science-fiction movie Star Trek

subject	*predicate*	*object*
(LeonardNimoy,	*profession,*	*Actor)*
(LeonardNimoy,	*starredIn,*	*StarTrek)*
(LeonardNimoy,	*played,*	*Spock)*

- Open vs Closed work assumption
  - Closed world assumption: non-existing triples indicate false relationships
  - Open world assumption: non-existing triples might be true relationships

Nickel, Maximilian, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. "A review of relational machine learning for knowledge graphs." *Proceedings of the IEEE* 104, no. 1 (2015): 11-33.

Important KG tasks

- Link prediction (Knowledge graph completion)
  - Given *G* predict if $(v_n^t, r, v_n^{t'})$ exists
- Entity resolution
  - Given *G* decide whether two entities represent the same entity, e.g., Obama, Barack Obama
- KG cleaning / detecting anomalies in KGs
  - Automated KG construction may introduce noisy/anomalous links in the KG
- Task addressed by machine learning approaches or rule based approaches
  - Rules have to be designed and not always applicable

(Vassilis, is_born_in, Athens)

(Vassilis, is, Greek)

Nickel, Maximilian, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. "A review of relational machine learning for knowledge graphs." *Proceedings of the IEEE* 104, no. 1 (2015): 11-33.

Representation learning on knowledge graphs

- Learn *D x 1* embeddings for nodes and relation-types in the KG

$$(v_n^t, r, v_n^{t'}) \longrightarrow \mathbf{h}_{n_t},\ \mathbf{h}_r,\ \mathbf{h}_{n_{t'}}$$

- KG embedding methods (KGE) define different scoring functions for triples $f(\mathbf{h}_{n_t}, \mathbf{h}_r, \mathbf{h}_{n_{t'}})$
  - TransE $-\|\mathbf{h}_{n_t} + \mathbf{h}_r - \mathbf{h}_{n_{t'}}\|$
  - DistMult $\mathbf{h}_{n_t}^\top \mathrm{diag}(\mathbf{h}_r)\mathbf{h}_{n_{t'}}$

$$\min_{\{\mathbf{h}_{n_t}, \mathbf{h}_r, \mathbf{h}_{n_{t'}}\}} \sum_{n_t, r, n_{t'} \in \mathbb{D}^+ \cup \mathbb{D}^-} \log(1 + \exp(-y \times f(\mathbf{h}_{n_t}, \mathbf{h}_r, \mathbf{h}_{n_{t'}})))$$

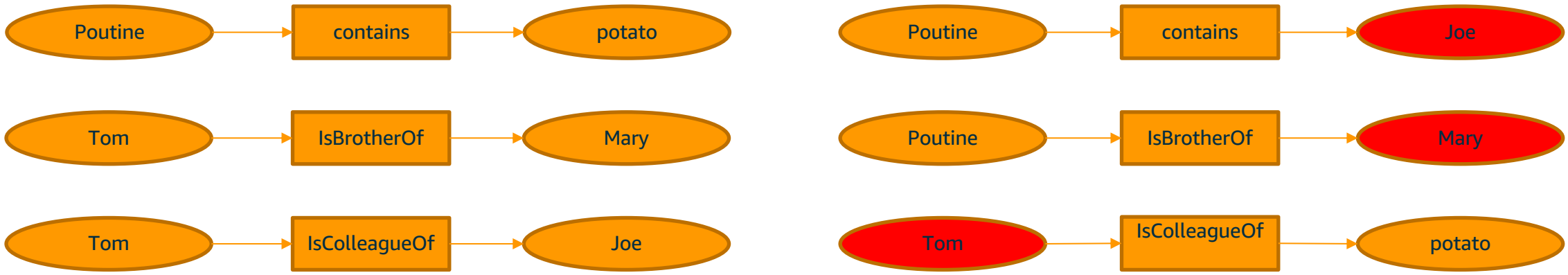
- *y* is 1 if the triple exists in the graph (positive triple) and -1 otherwise (negative triple)

- DGL-KE an optimized library for learning KGE models https://github.com/awslabs/dgl-ke

Zheng, D., Song, X., Ma, C., Tan, Z., Ye, Z., Dong, J.,Xiong, H., Zhang, Z., and Karypis, G. "Dgl-ke: Trainingknowledge graph embeddings at scale" arXiv:2004.08532, 2020

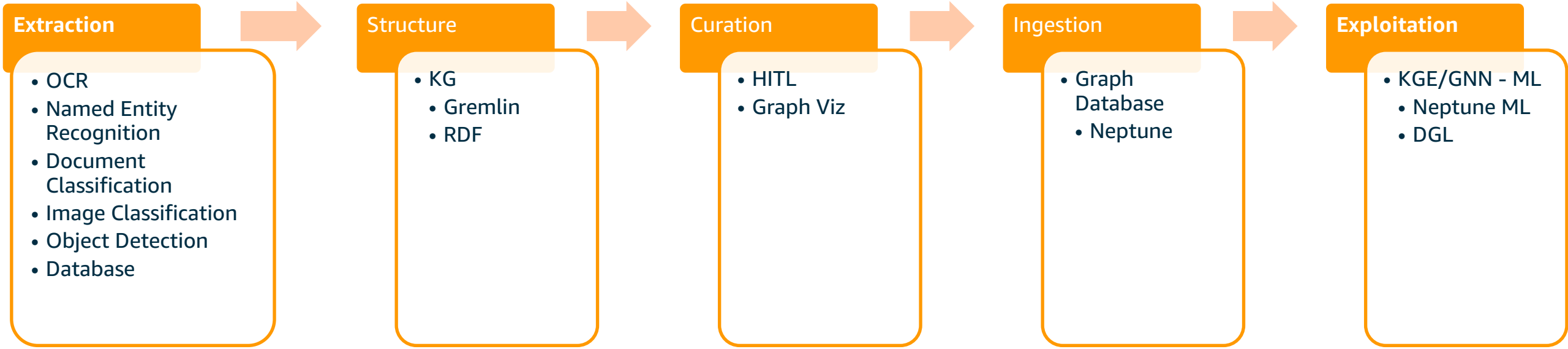
Training KGEs with negative sampling

- Creating *K* negative triples per positive triple by randomly swapping head and tail entities



- Why not considering as negative triples all the non existent edges in *G* ?
  - Scalability : One million nodes corresponds to one billion negative examples
  - Open world assumption (OWA) : Some of the negative triples might be in fact positive
  - Generalization : Sampling works better in practice because of OWA

General Process of Knowledge Graph Construction

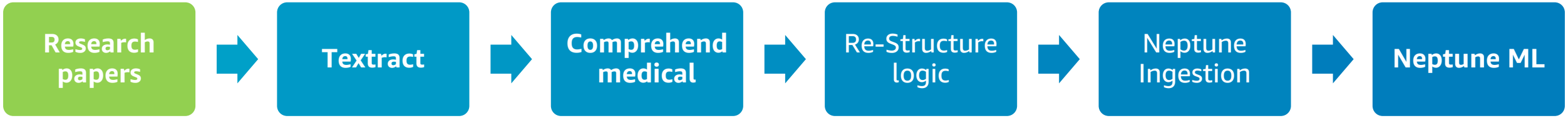


Case studies of KGs in life sciences

1. Automated KG construction from PubMed literatures

2. Drug repurposing with KG

Proposed Knowledge Graph Construction Pipeline for HCLS



Drug repurposing knowledge graph

- Motivation
  - Improve the efficiency and speed of discovering new treatments
  - De novo drug discovery is a lengthy and costly process
- Drug repurposing
  - Redevelop existing drugs for use in novel diseases
  - Relies on identifying novel interactions among biological entities like genes and compounds

DRKG data sources

- Abundance of data available in different data sources

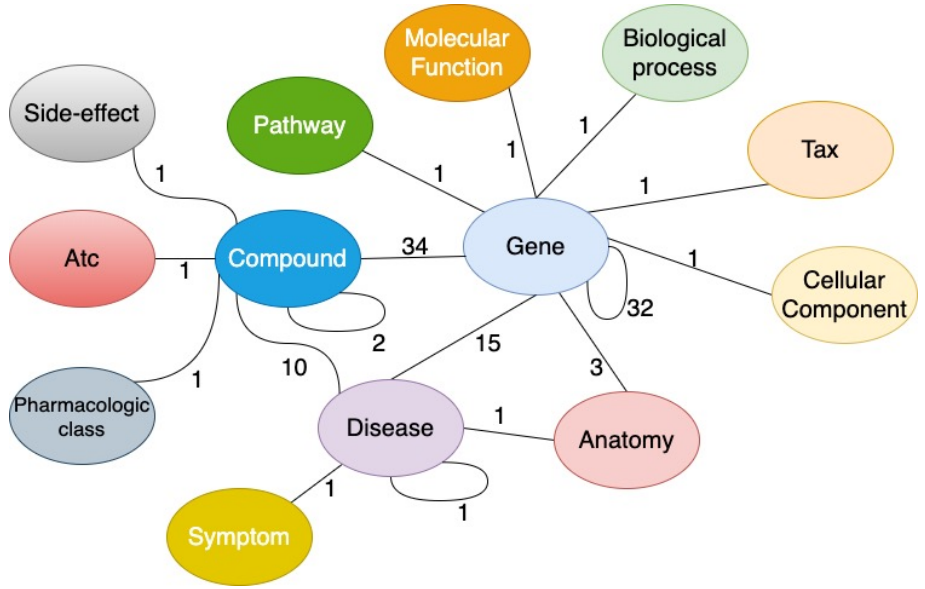
- Data sources use one of several ID spaces to represent genes, compounds, diseases and others.
  - DrugBank : drugs and properties
  - STRING : protein to protein interactions
  - GNBR : relations among biological entities extracted from text
  - Recent bibliographic resources concerning Covid-19 proteins and treatments

Bethany Percha and Russ B Altman. “A global net-work of biomedical relationships derived from text”. *Bioinformatics* (2018), pp. 2614–2624.

DRKG

- Github: https://github.com/gnn4dr/DRKG

Entity type	Drugbank	GNBR	Hetionet	STRING	IntAct	DGIdb	Bibliography	Total Entities
Anatomy	-	-	400	-	-	-	-	400
Atc	4,048	-	-	-	-	-	-	4,048
Biological Process	-	-	11,381	-	-	-	-	11,381
Cellular Component	-	-	1,391	-	-	-	-	1,391
Compound	9,708	11,961	1,538	-	153	6,348	6,250	24,313
Disease	1,182	4,746	257	-	-	-	33	5,103
Gene	4,973	27,111	19,145	18,316	16,321	2,551	3,181	39,220
Molecular Function	-	-	2,884	-	-	-	-	2,884
Pathway	-	-	1,822	-	-	-	-	1,822
Pharmacologic Class	-	-	345	-	-	-	-	345
Side Effect	-	-	5,701	-	-	-	-	5,701
Symptom	-	-	415	-	-	-	-	415
Tax	-	215	-	-	-	-	-	215
Total	19,911	44,033	45,279	18,316	16,474	8,899	9,464	97,238



Head entity	relations	tail entity
Gene::6118	STRING::BINDING::Gene:Gene	Gene::1111
Gene::5433	STRING::CATALYSIS::Gene:Gene	Gene::8148
Gene::84617	DGIDB::INHIBITOR::Gene:Compound	Compound::DB05773

Vassilis N. Ioannidis, Xiang Song, Saurav Manchanda, Mufei Li, Xiaoqin Pan, Da Zheng, Xia Ning, Xiangxiang Zeng, and George Karypis, 'DRKG - Drug Repurposing Knowledge Graph for Covid-19', (to appear), 2020

DR task vis a vis link prediction

- DR predicts whether existing drugs can be used to treat a novel disease
    - Predict whether a drug interacts with a disease via a treatment relation
    - Predict whether a drug interacts with a gene related to the target disease via an inhibit relation

- DR as a link prediction task in the DRKG using the learned KGE to
    - Score and rank the possible drugs for treating the target disease
    - Score and rank the possible drugs for inhibiting the relevant genes

DR for Covid-19 using treatment relation

- Validation drugs are 32 clinical trial drugs for Covid-19

- Target relation type is drug-treats-disease

Drug name	Score	Ranking in top-100
Ribavirin	-0.21	0
Dexamethasone	-1.00	4
Colchicine	-1.08	8
Methylprednisolone	-1.16	16
Oseltamivir	-1.39	49
Deferoxamine	-1.51	87

- 6 drugs ranked in the top 100 predicted treatments

DR for Covid-19 using inhibit relation

- Validation drugs are 32 clinical trial drugs for Covid-19
- Extracted 442 Covid-19 related genes
- Target relation type is drug-inhibits-gene
- Aggregated top 100 results across genes

Drug name	Number of hits	Frequency
Dexamethasone	401	17.42
Thalidomide	336	9.52
Chloroquine	258	5.28
Deferoxamine	111	2.38
Colchicine	108	1.94
Methylprednisolone	105	1.68
Losartan	92	1.99
Ribavirin	92	2.03
Ruxolitinib	47	0.77
Tofacitinib	33	0.46
Hydroxychloroquine	14	0.20
Piclidenoson	6	0.15
Azithromycin	5	0.058
Oseltamivir	1	0.2
Sargramostim	1	0.011