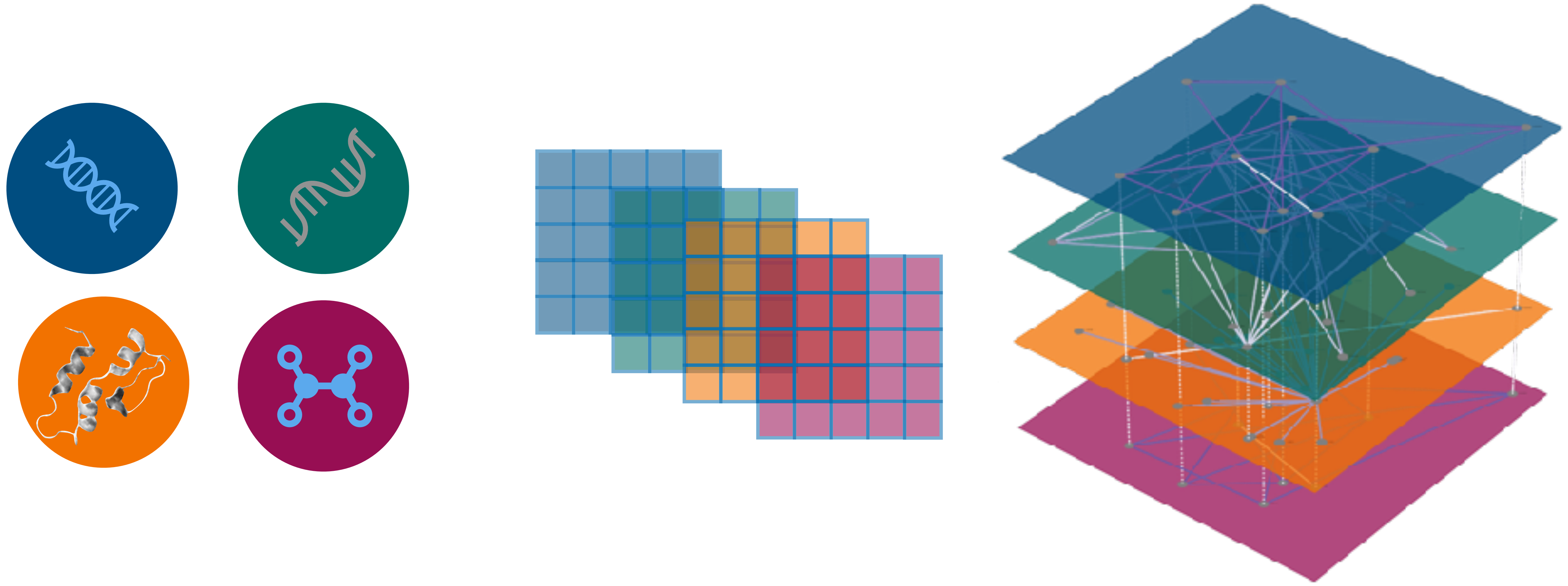


Graph Machine Learning

AI-guided Protein Science

Alberto Santos — Multi-omics Network Analytics (MoNA)

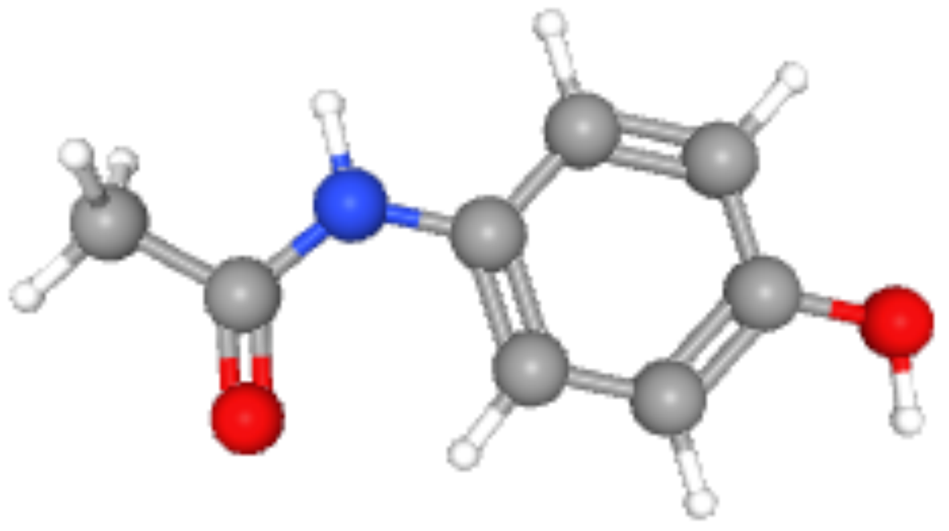
The Novo Nordisk Foundation
Center for Biosustainability

Limitations of Deep Learning

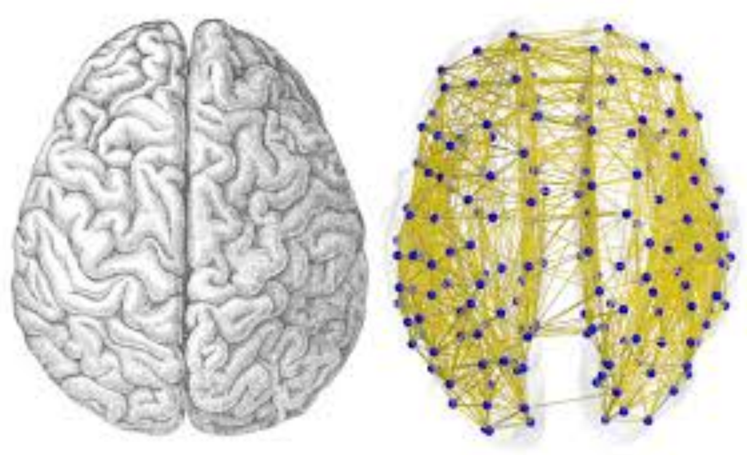
- Standard DL (CNNs, RNNs, Transformers) is limited to work with data that is **structured** in regular formats, such as **images**, **sequences**, and **grids**.

- What about **different structures**?

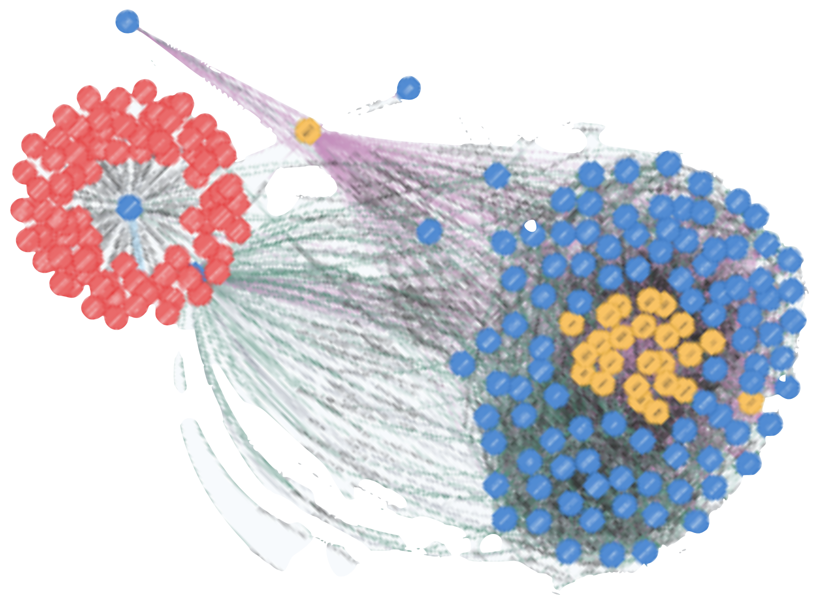
Graphs are everywhere and help describe complex systems



Molecules



Connectivity



Knowledge graphs

PPIs

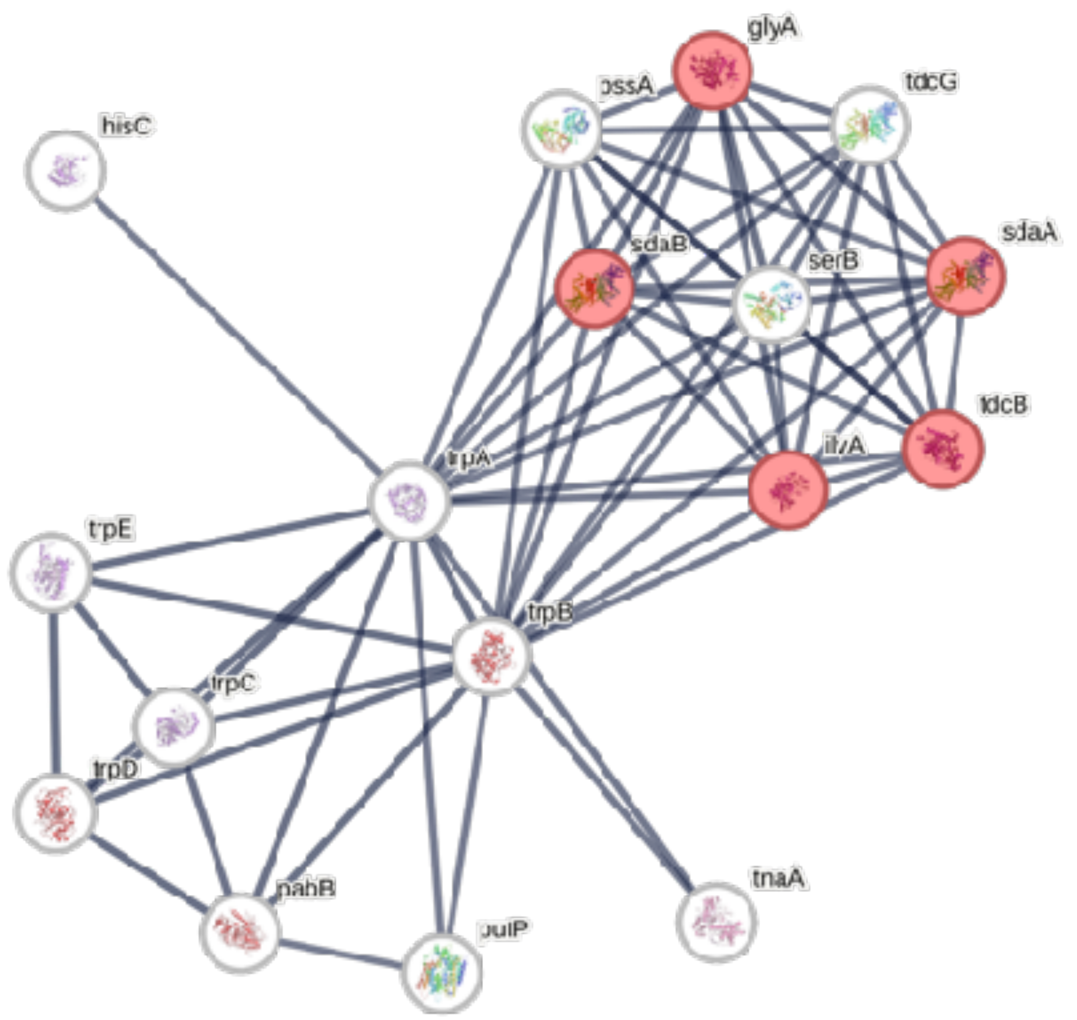
Machine Learning on Graphs

- Three types of prediction tasks:
  - **Node**-level: predict the identity, role or features of nodes within a graph.
  - **Edge**-level: predict relationships between nodes or features of these relationships.
  - **Graph**-level: predict the property of an entire graph.
- Graphs contain different information that we can use to make predictions: nodes, edges, global-context and connectivity.

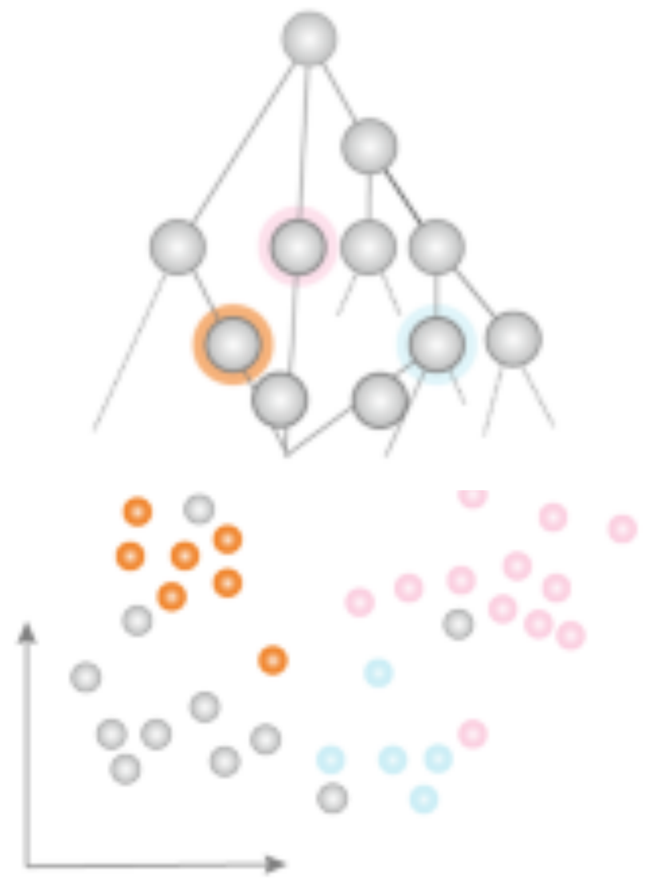
Node Classification

- Predict the label (type, category or attribute) associated with all the nodes in the graph

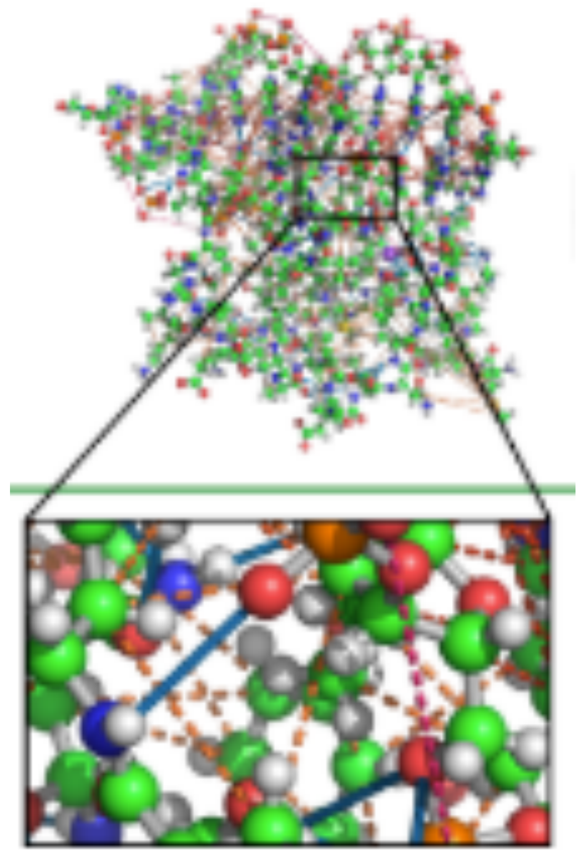
Examples



Classifying the function of proteins



Classifying cell types



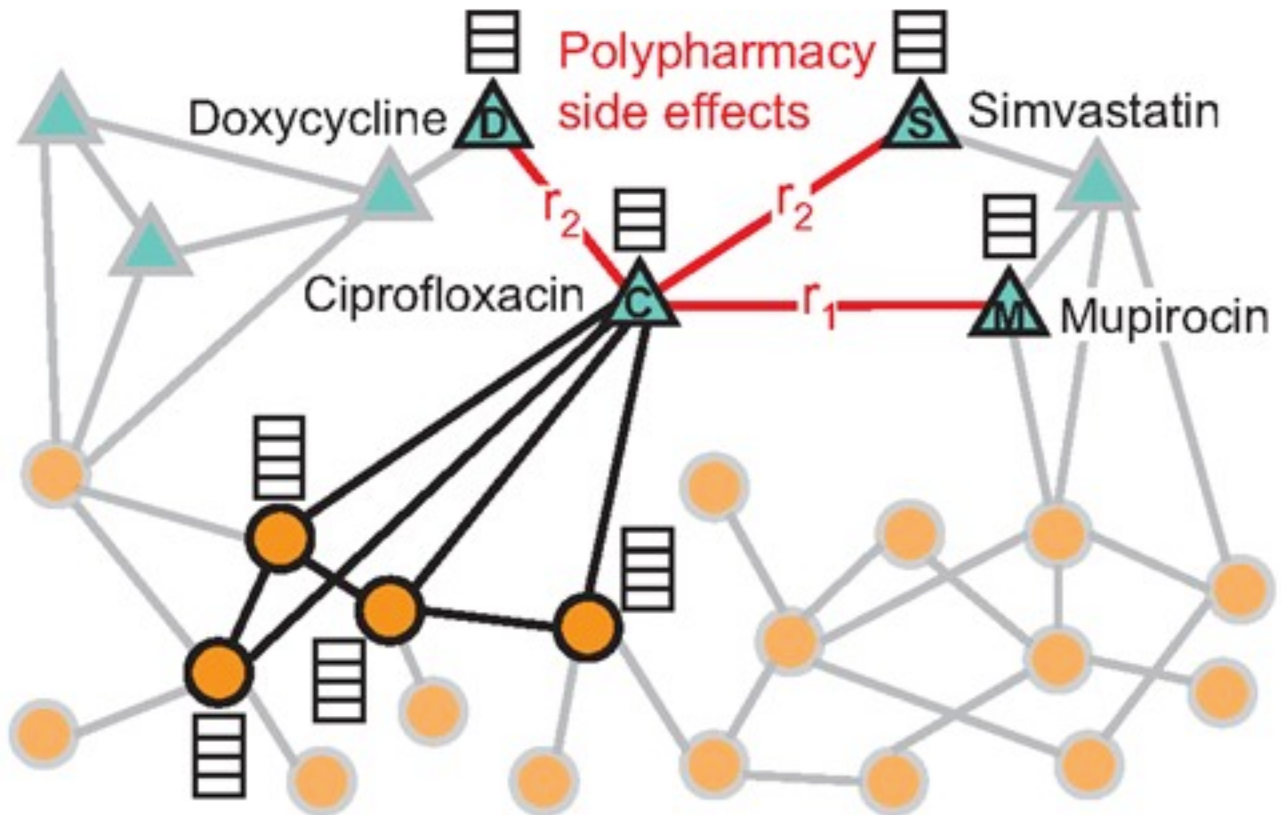
Classifying amino acids

Leveraging the Cell Ontology to classify unseen cell types. 2021. Sheng Wang, Angela Oliveira Pisco, Aaron McGeever, Maria Brbic, Marinka Zitnik, Spyros Darmanis, Jure Leskovec, Jim Karkanias & Russ B. Altman

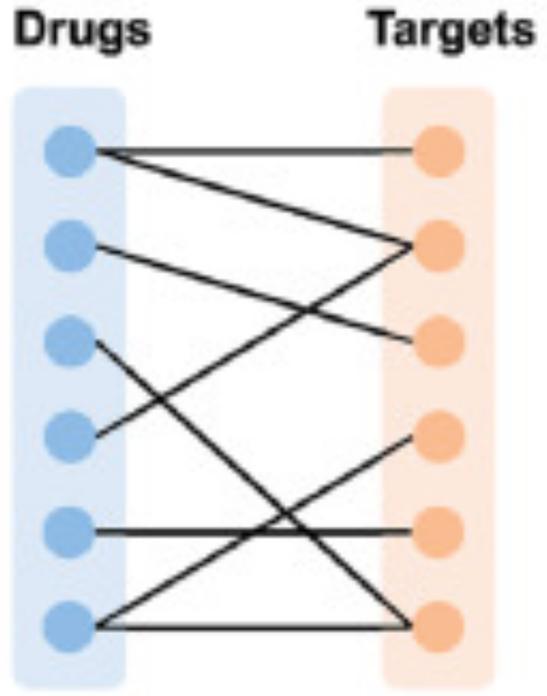
Relation Prediction

- Given a set of nodes and an incomplete set of edges between these nodes, we want to infer the missing edges using the structure of the graph

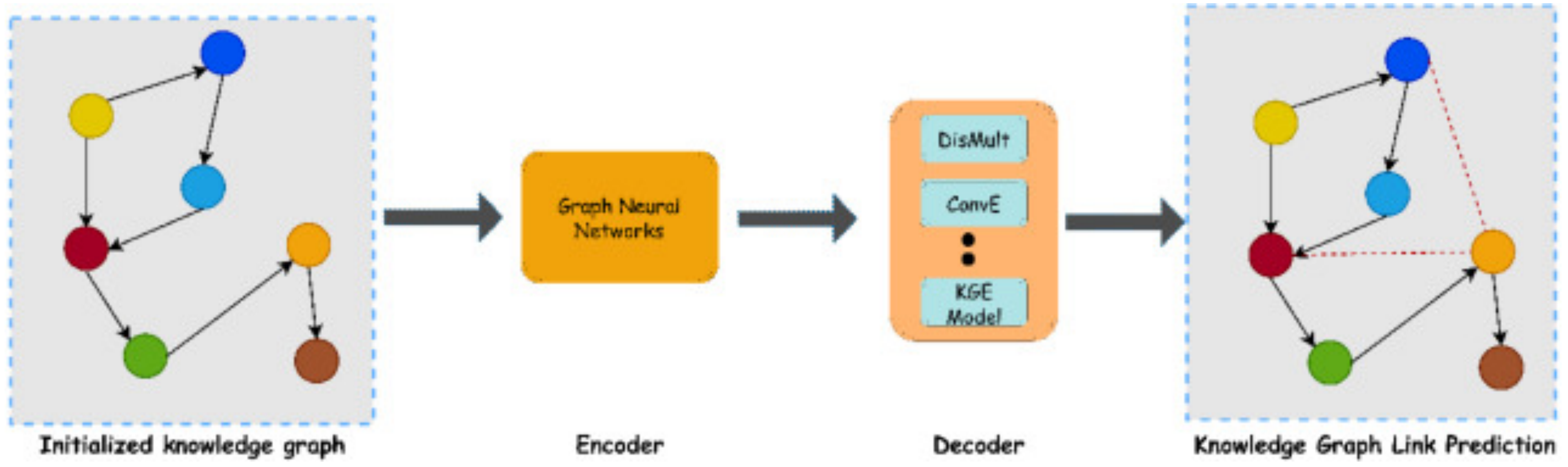
Examples



Drug-Side effect prediction



Drug-Target prediction



Knowledge Graph completion

Modeling polypharmacy side effects with graph convolutional networks. 2018. Marinka Zitnik, Monica Agrawal, Jure Leskovec

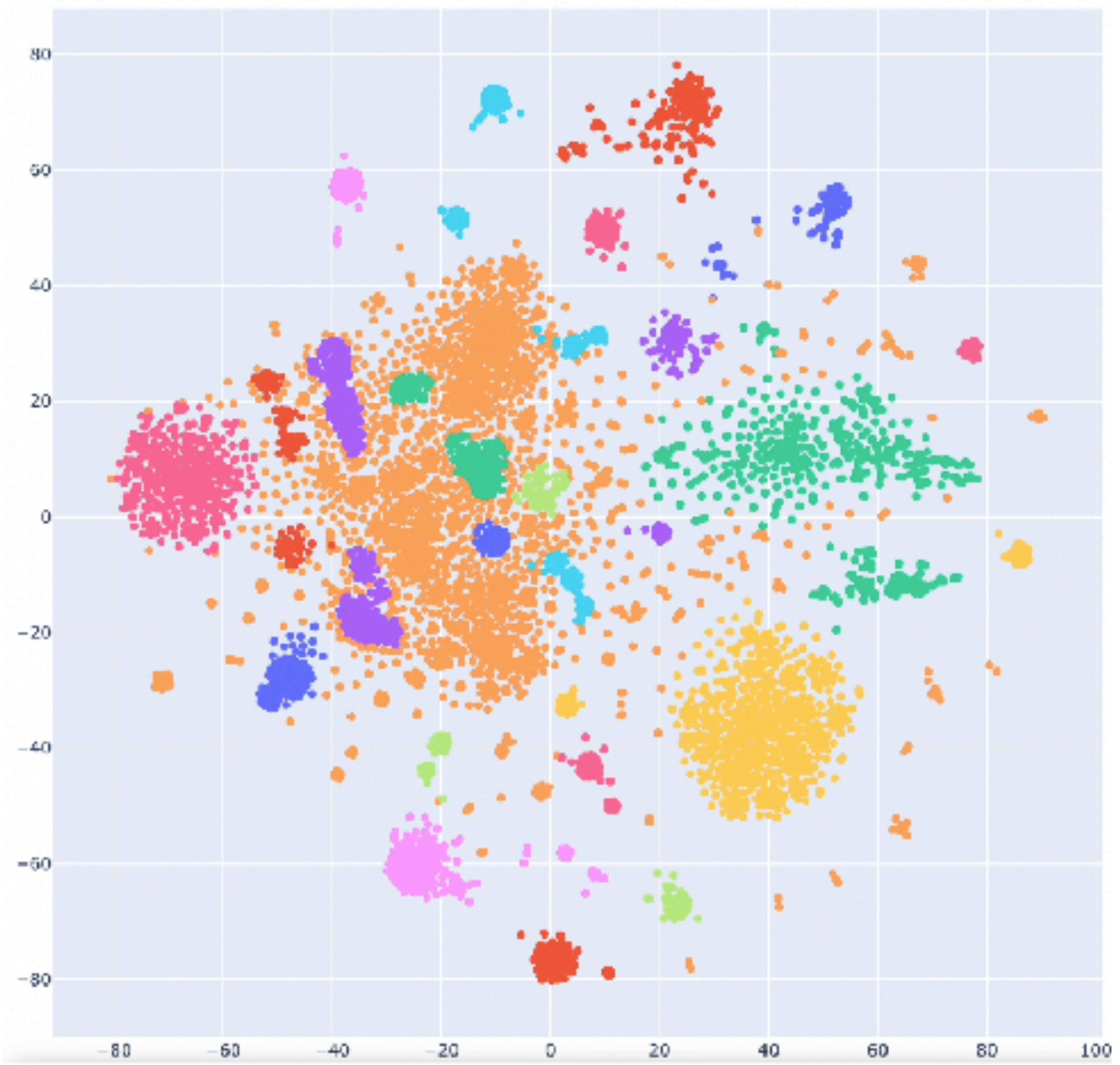
Graph neural network approaches for drug-target interactions. 2022. Zehong Zhang, Lifan Chen , Feisheng Zhong, Dingyan Wang, Jiaxin Jiang, Sulin Zhang, Hualiang Jiang, Mingyue Zheng, Xutong Li

A knowledge graph completion model based on contrastive learning and relation enhancement method.2022. LinYu Li, Xuan Zhang, YuBin Ma, Chen Gao, Jishu Wang, Yong Yu, Zihao Yuan, Qiuying Ma

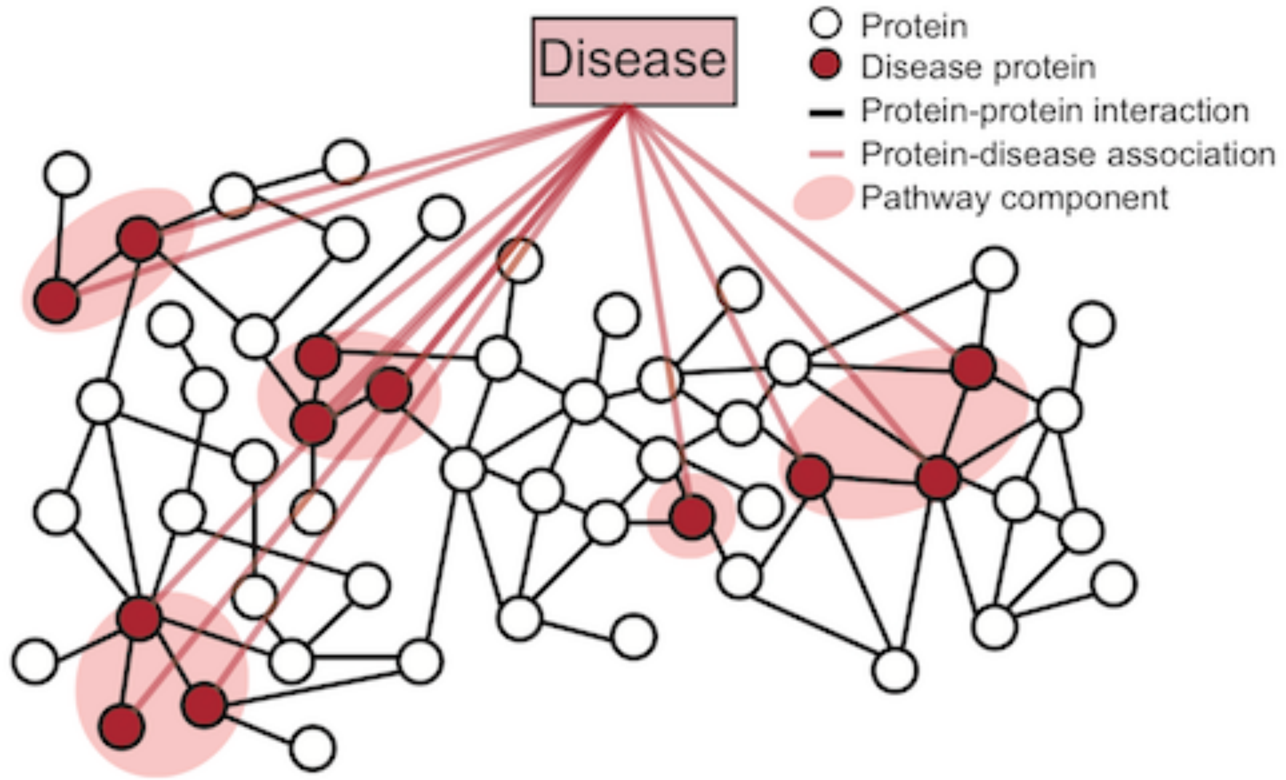
Community Detection

- Identify clusters where nodes are more likely to form edges

Examples



Detecting microbial communities



Disease pathways

Large-scale analysis of disease pathways in the human interactome. 2018. Monica Agrawal, Marinka Zitnik, Jure Leskovec.

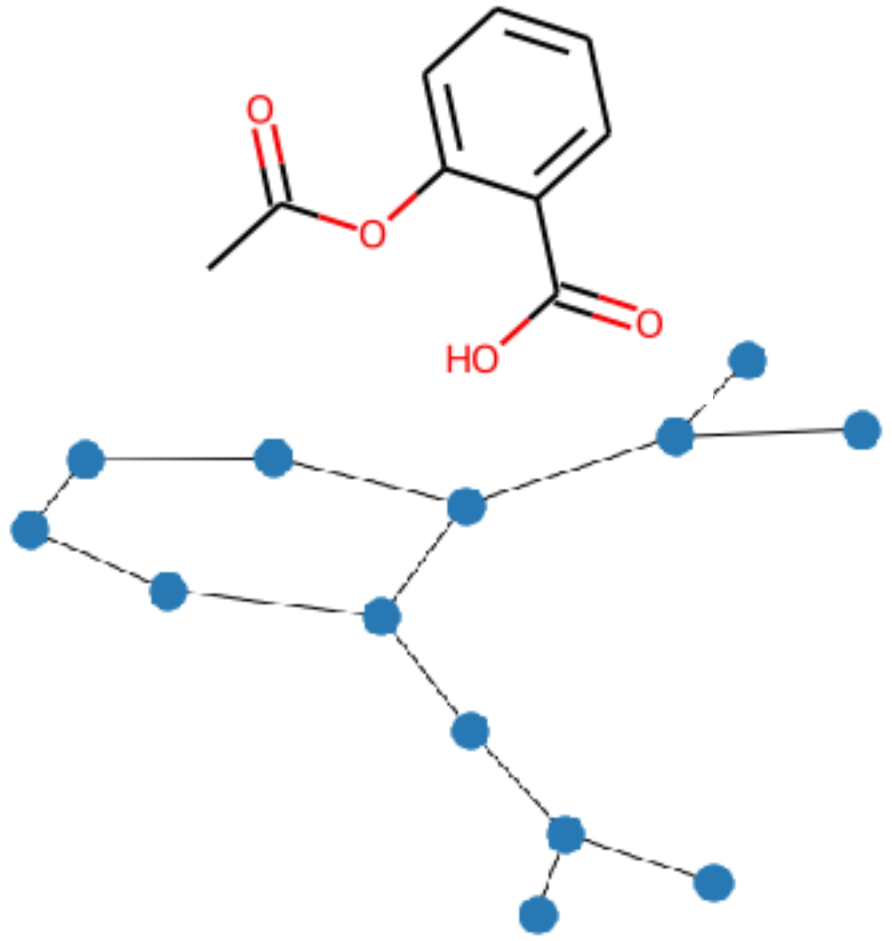
https://micw2graph.streamlit.app/Microbial_association_networks

https://github.com/benedekrozemberczki/awesome-community-detection/blob/master/chapters/deep_learning.md

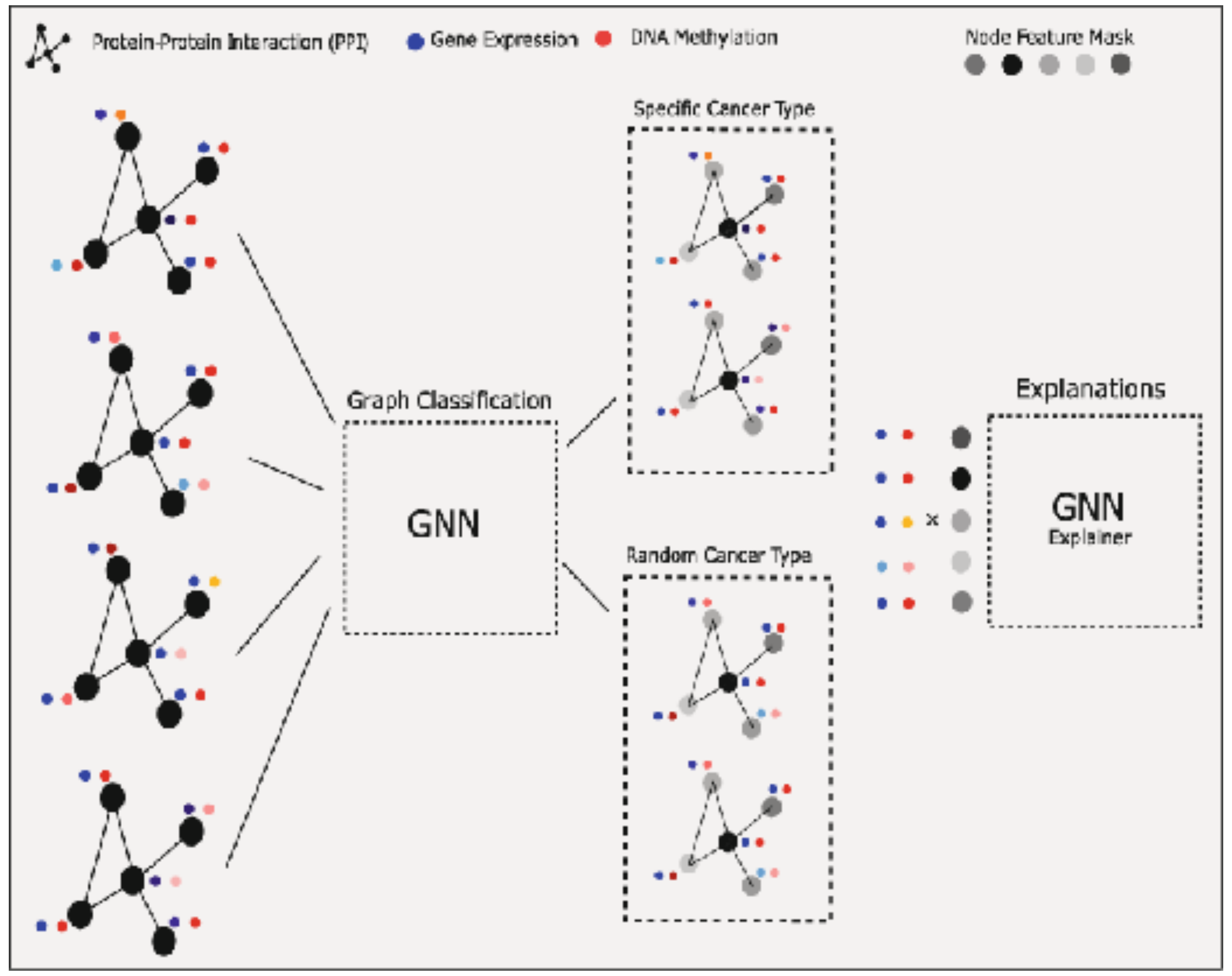
Graph Classification/Regression

- Learn over graph data and make independent predictions for each graph

Examples



Predicting molecular properties



Patient stratification

Human-in-the-Loop Integration with Domain-Knowledge Graphs for Explainable Federated Deep Learning. 2023. Andreas Holzinger, Anna Saranti, Anne-Christin Hauschild, Jacqueline Beinecke, Dominik Heider, Richard Roettger, Heimo Mueller, Jan Baumbach, Bastian Pfeifer

Important Concepts

- **Independence** — points are not independent and identically distributed (interconnected nodes)

- **Homophily** — tendency of nodes to share attributes with their neighbours

- **Structural equivalence** — Nodes with similar local structures will share similar labels

Traditional Approaches

Node-level statistics

- Extracting some **statistics or features** from the graph and sing these features as input to an standard machine learning classifier
- **Node-level**:

  - Node degree - Node centrality - Clustering coefficient - Motifs or graphlets

- **Edge-level**:
  - <u>Local</u> overlap: Number of common neighbours two nodes share (e.g., Jaccard overlap)
  - <u>Global</u> overlap:
    - *Katz Index*: number of paths of all lengths between two nodes
    - *Leicht, Holme, and Newman* Similarity: ratio between the number of observed paths and the number of expected paths between two nodes.
    - *Random walk methods*: considering random walks instead of exact counts of paths (e.g., Personalised Page Rank, probability that a random walk starting at one node visits another).

Exploiting graphlet decomposition to explain the structure of complex networks: the GHuST framework. 2020. Rafael Espejo, Guillermo Mestre, Fernando Postigo, Sara Lumbreras, Andres Ramos, Tao Huang & Ettore Bompard

Traditional Approaches

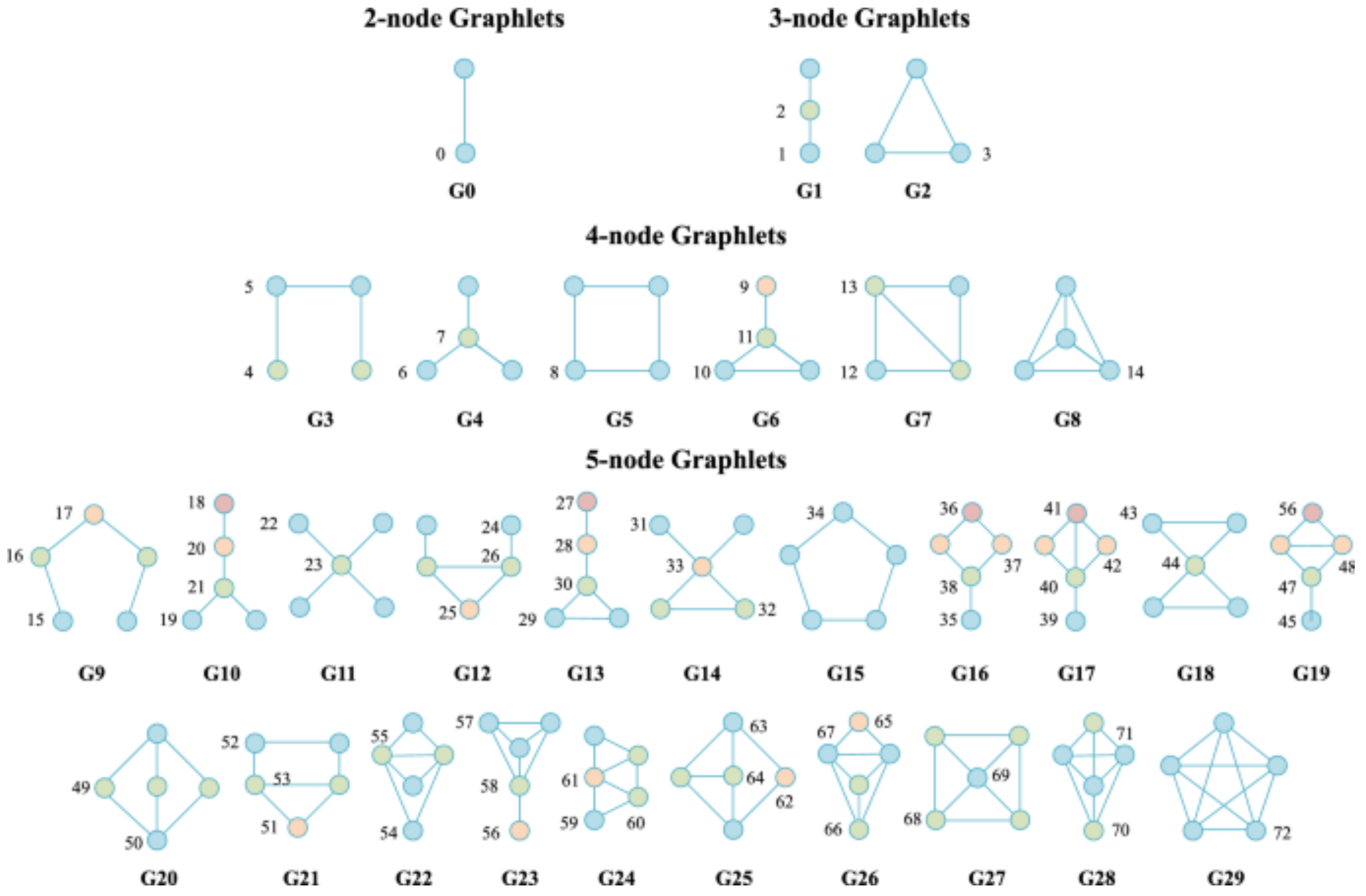
Graph-level statistics

- **Bag of nodes**: aggregating node-level statistics (summary statistics based on node stats)

- **Iterative Neighbour Aggregation**: extract node-level features that contain rich information about their neighbours and aggregate it at the graph-level — e.g., The Weisfieler-Lehman Kernel

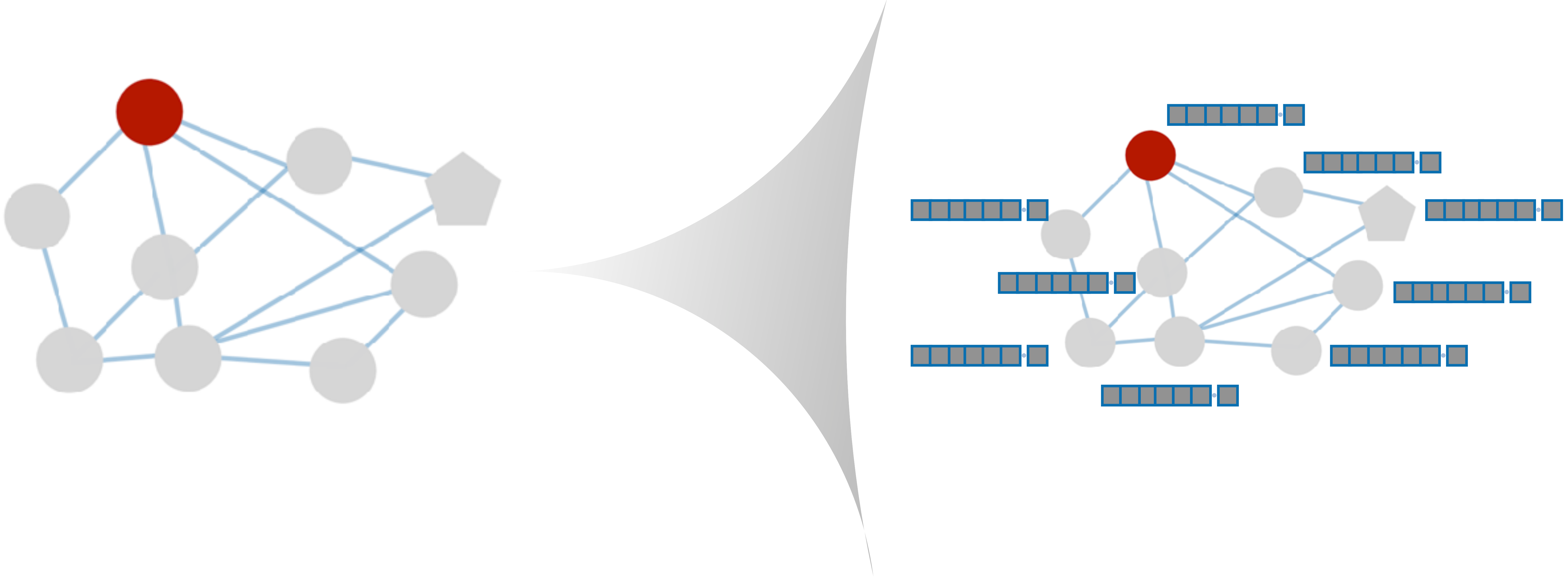
- **Graphlets**: count the occurrence of different small subgraph structures (graphlets)

- **Path-based methods:** evaluates the different kinds of paths that occur in the graph (e.g., random walks/ shortest paths counting occurrence of sequences of node labels (degree))



Representation Learning

- The node- and graph-level statistics approached are limited
- Graph Representation Learning — learn representations that encode structural information about the graph to use them downstream

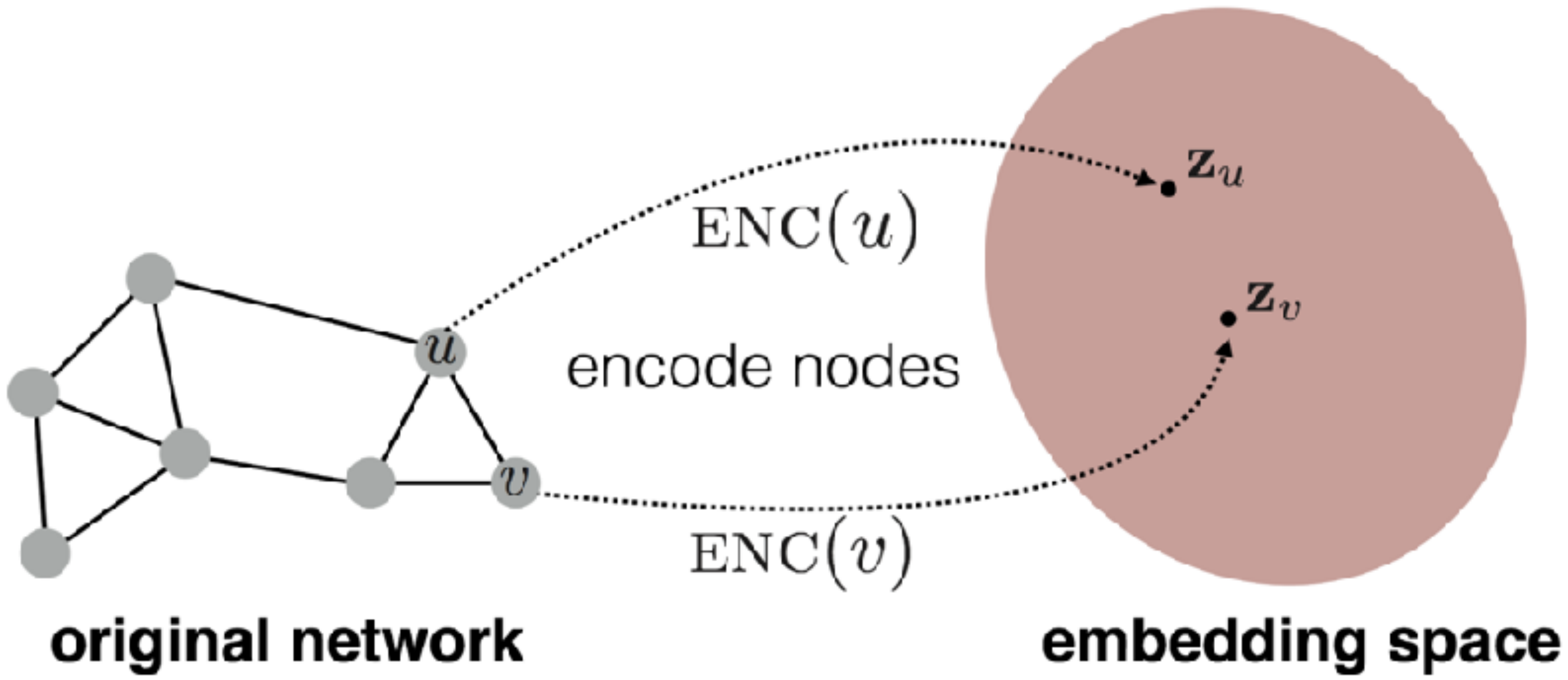


Node Embeddings

- Encode nodes as low-dimensional vectors summarising their position and structure of the local graph neighbourhood

- Similarity in the embedding space (e.g., dot product) approximates similarity in the original graph

- Encoder-decoder methods:



Encoder (ENC): goal learn a mapping from nodes to embeddings. We Define a node similarity function and optimise the parameters of the encoder so that similarity

$$S[u, v] \approx \mathbf{z}_u{}^T\mathbf{z}_v$$

Decoder (DEC): reconstruct certain graph statistics from the node embedding (e.g., pairwise decoders predict whether two nodes are neighbours in the graph).

$$\mathrm{DEC}(\mathbf{z}_u, \mathbf{z}_v) \approx S[u, v]$$

Shallow Embedding Methods

Type	Method	Decoder	Similarity measure	Loss function
Matrix Factorisation	Laplacian Eigenmaps	$\|z_u - z_v\|_2^2$	General	$\mathrm{DEC}(z_u, z_v) \cdot S[u, v]$
	Graph Factorisation	$z_u^\top z_v$	$A[u, v]$	$\|\mathrm{DEC}(z_u, z_v) \cdot S[u, v]\|_2^2$
	GraRep	$z_u^\top z_v$	$A[u, v], A^2[u, v], \ldots, A^k[u, v]$	$\|\mathrm{DEC}(z_u, z_v) \cdot S[u, v]\|_2^2$
	Hope	$z_u^\top z_v$	General	$\|\mathrm{DEC}(z_u, z_v) \cdot S[u, v]\|_2^2$
Random walk	DeepWalk	$\frac{e^{z_u^\top z_v}}{\sum_{k \in \mathcal{V}} e^{z_u^\top z_k}}$	$p_{\mathcal{G}}(v \mid u)$	$-S[u, v] \cdot \log(\mathrm{DEC}(z_u, z_v))$
	node2vec	$\frac{e^{z_u^\top z_v}}{\sum_{k \in \mathcal{V}} e^{z_u^\top z_k}}$	$p_{\mathcal{G}}(v \mid u)$ (biased)	$-S[u, v] \cdot \log(\mathrm{DEC}(z_u, z_v))$

$p_{\mathcal{G}}(v \mid u)$ Probability of visiting *v* on a fixed length random walk starting from *u*

Representation Learning on Graphs: Methods and Applications.2017. William L. Hamilton, Rex Ying, Jure Leskovec

Random Walk Methods

- Two nodes have similar embeddings if they cooccur on short random walks

Type	Method	Decoder	Similarity measure	Loss function
Random walk	DeepWalk	$\frac{e^{z_u^\top z_v}}{\sum_{k\in\mathcal{V}} e^{z_u^\top z_k}}$	$p_{\mathcal{G}}(v \mid u)$	$-S[u,v]\cdot\log(\mathrm{DEC}(z_u, z_v))$
	node2vec	$\frac{e^{z_u^\top z_v}}{\sum_{k\in\mathcal{V}} e^{z_u^\top z_k}}$	$p_{\mathcal{G}}(v \mid u)$ (biased)	$-S[u,v]\cdot\log(\mathrm{DEC}(z_u, z_v))$

- These methods differ on how they overcome the complexity of evaluating the loss function ($O(|V|)$).

Limitations of Shallow Embeddings

- **Complexity grows** with the number of nodes ($O(|V|)$) — problem with large graphs

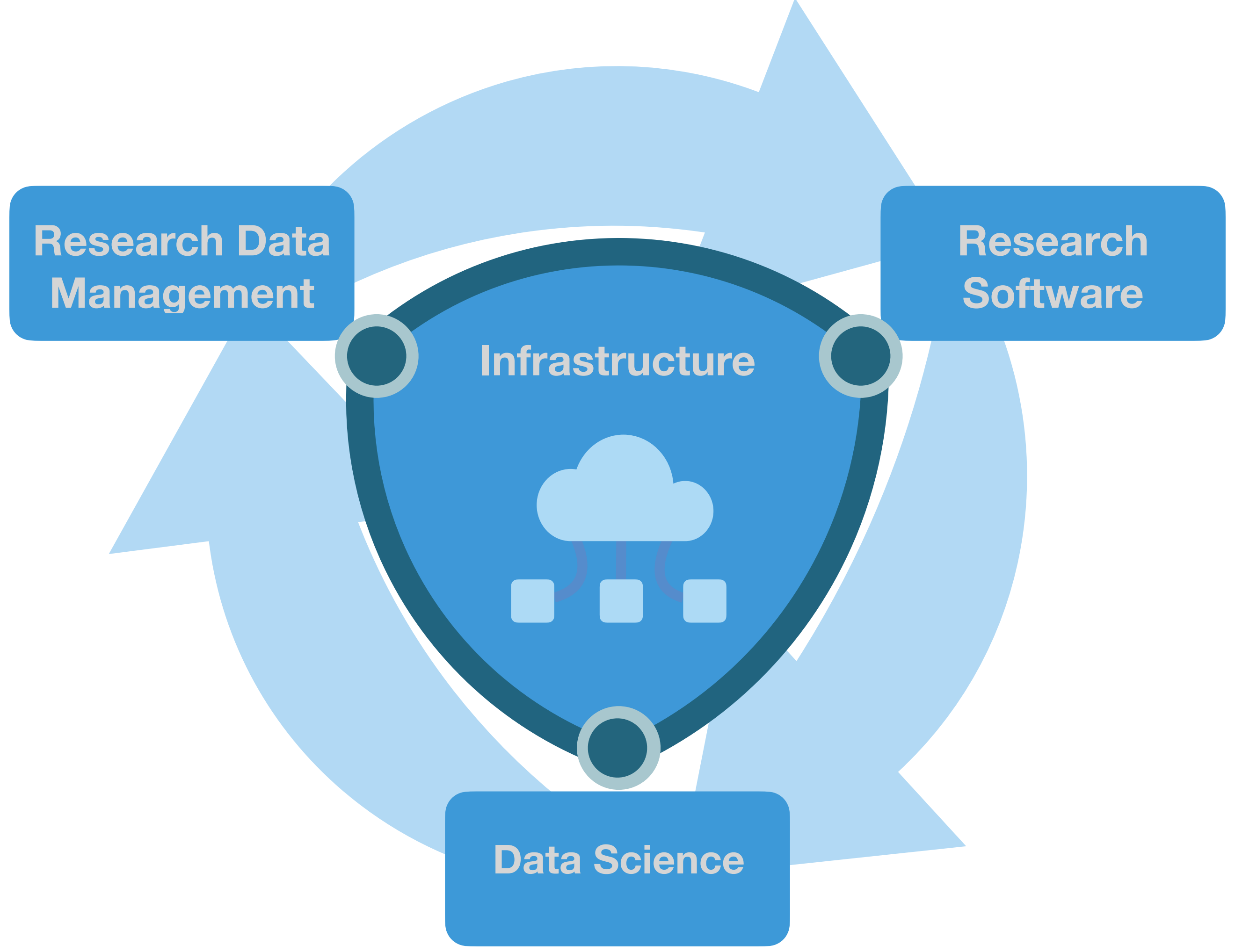
- They **do not use node features**

- They are **transductive** — can only generate embeddings for nodes present during the training phase (no generalizable to unseen nodes)

Thank you

Multi-omics Network Analytics Research Group

Informatics Platform



The Novo Nordisk Foundation Center for Biosustainability

novo nordisk fonden

PIONEER CENTRE FOR ARTIFICIAL INTELLIGENCE

✉ albsad@dtu.dk

@albsantosdel

https://github.com/Multiomics-Analytics-Group

https://multiomics-analytics-group.github.io/

The Novo Nordisk Foundation Center for Biosustainability