Opportunities and Challenges in GraphML for Biomedicine

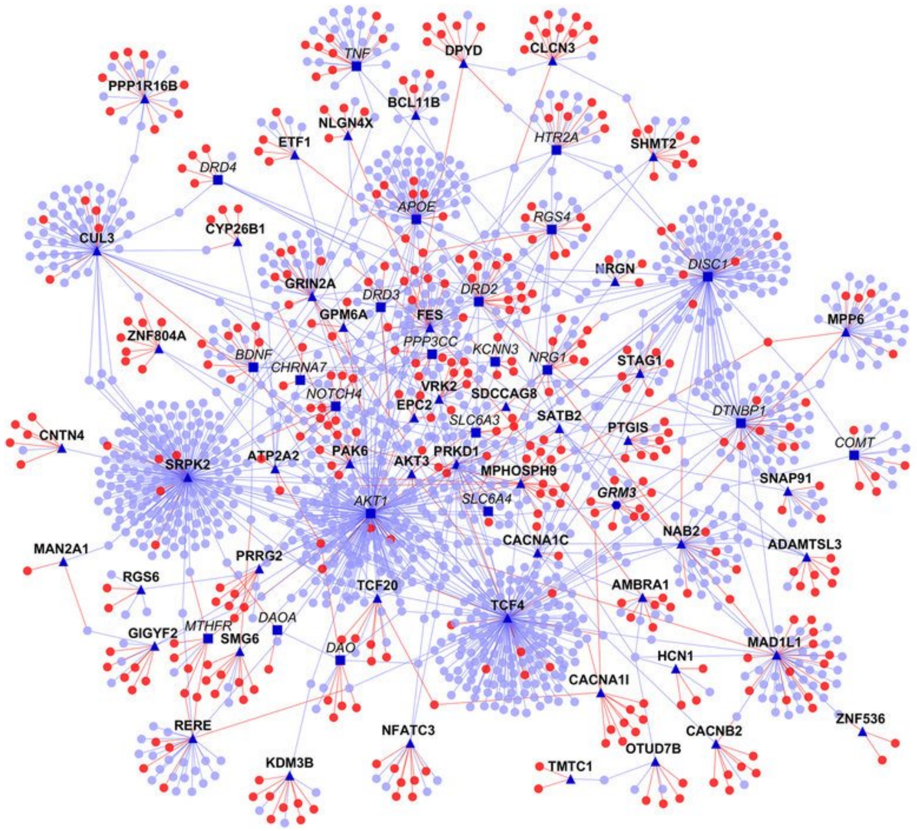
Megha Khosla

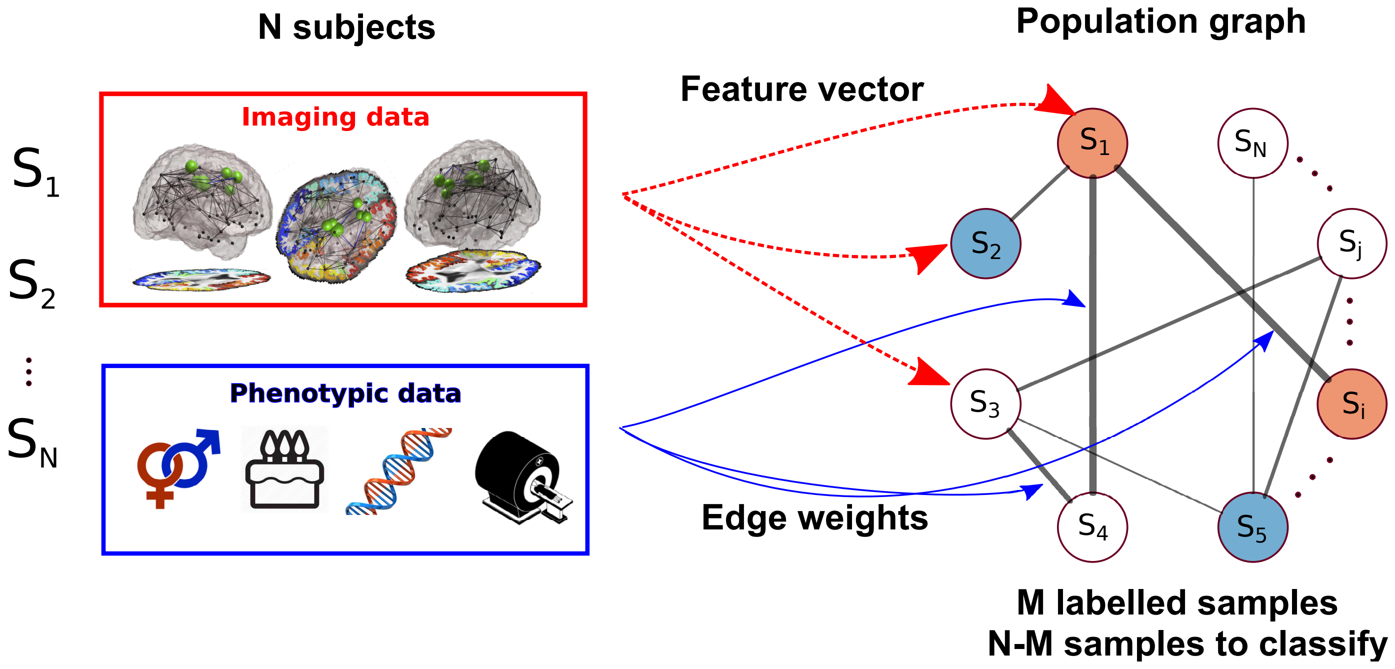
https://khosla.github.io

m.khosla@tudelft.nl

TUDelft

Graphs in biomedicine




N subjects

Population graph

Feature vector

Imaging data

Phenotypic data

Edge weights

M labelled samples
N-M samples to classify

Protein interaction network

Image Source : wikipedia

Patient Network

Image Source : Parisot et al.

Graph Machine Learning (GraphML)

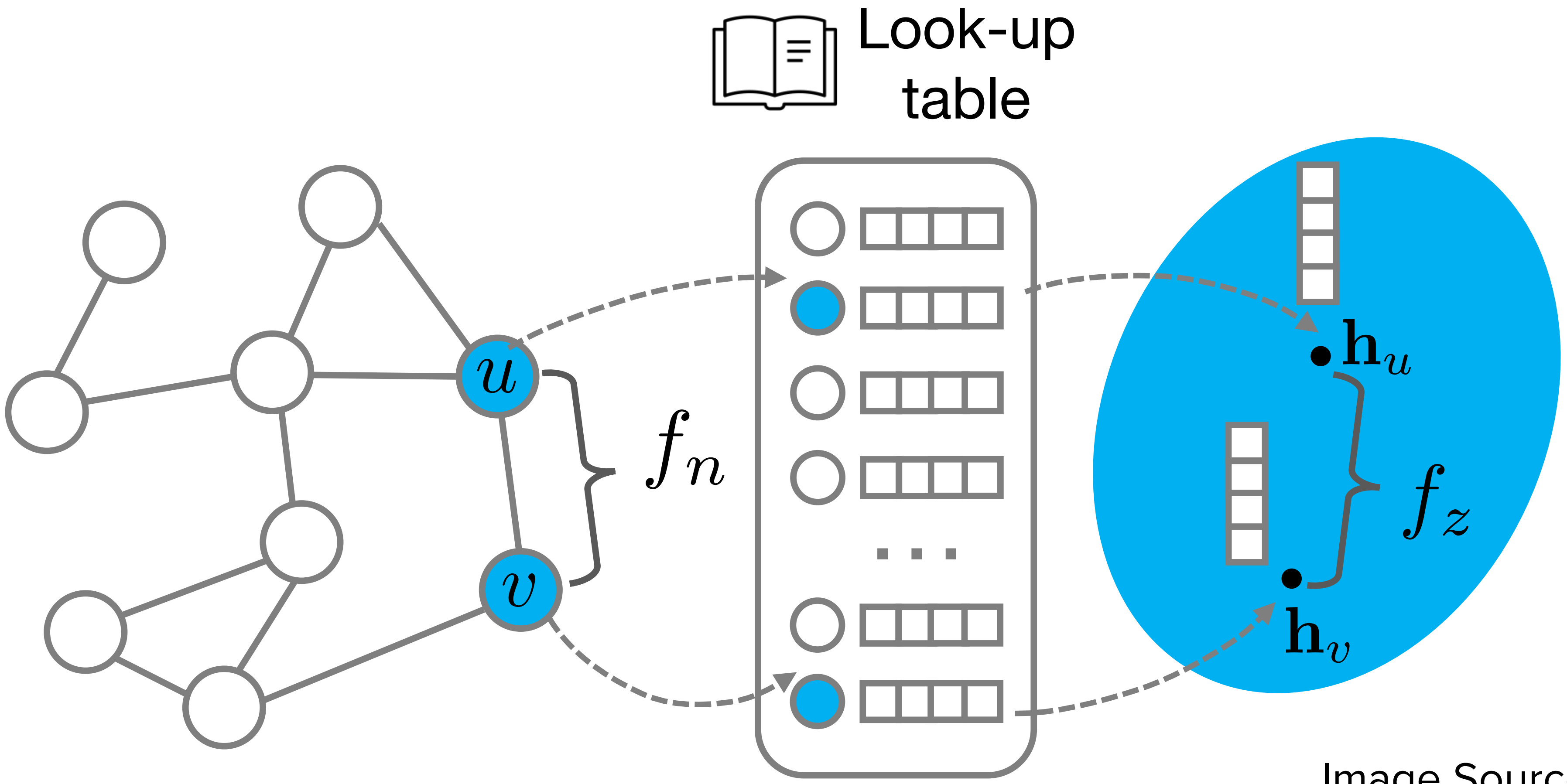
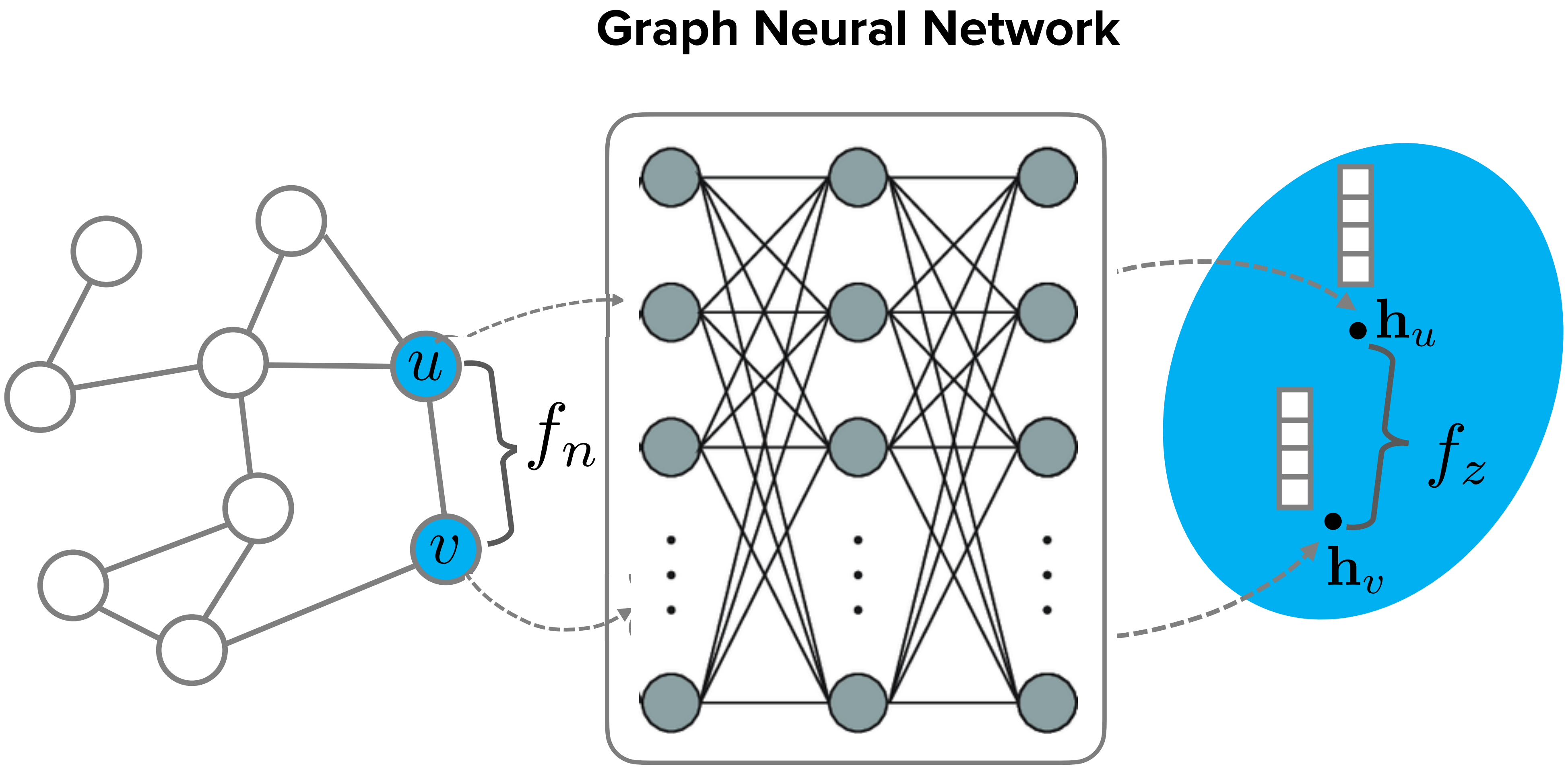


Image Source: [Li et al., 2022]

Shallow Network Embedding Methods

Examples : **DeepWalk, Node2Vec, NERD, HOPE**

Graph Machine Learning



Examples :

GCN, GAT, GIN

Applications of GraphML in Biomedicine

Biological problems

- Predict new human-pathogen protein interactions
- Predict new miRNA-disease associations

Main Challenges

- Data scarcity
- Data Bias

Other common issues

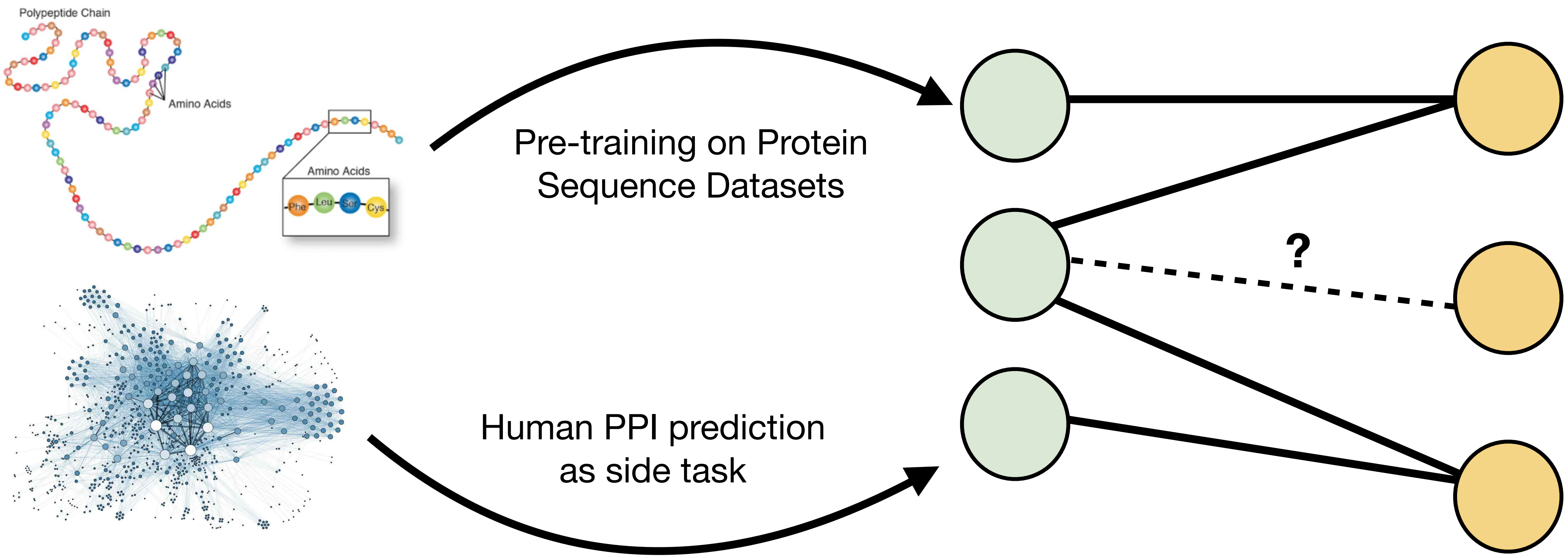
- Wrong evaluation setups leading to data leakage
- Limited and biased train-test data

- N. Dong, J. Schrader, S. Mücke, M. Khosla, "*A Message Passing framework with Multiple data integration for miRNA-Disease association prediction*", In **Scientific Reports**, 2022.
- N. Dong, S. Mücke, M. Khosla, "*MuCoMiD: A Multitask graph Convolutional Learning Framework for miRNA-Disease Association Prediction*", in IEEE/ACM **Transactions on Computational Biology and Bioinformatics** 2022
- N. Dong, G. Brogden, G. Gerold, M. Khosla, "A multi-task transfer learning framework for the prediction of virus-human protein-protein interactions", **BMC Bioinformatics**, 2021.
- N.Dong, M.Khosla, Towards a consistent evaluation of miRNA-disease association prediction models. In IEEE International Conference on Bioinformatics and Biomedicine (**BIBM**), 2020

Predicting protein-pathogen protein interactions

How to use inductive biases from multiple sources of information to overcome challenges of learning under low data regimes?

Fine Tuning



Pre-training on Protein Sequence Datasets

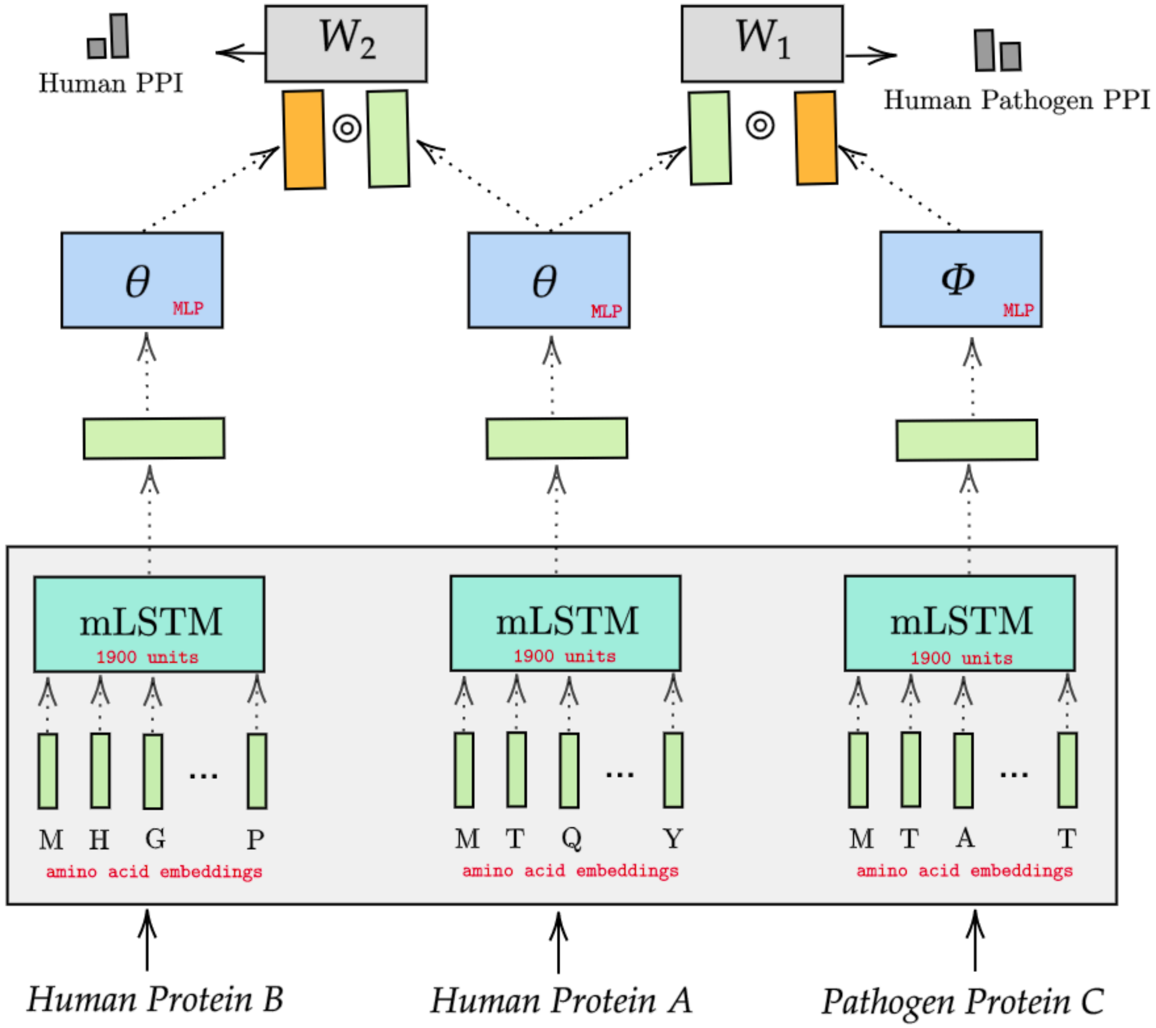
Human PPI prediction as side task

Join learning framework

- Powerful input protein representations learnt over 24 million protein sequences

- Multitask learning framework using graph reconstruction losses

- Besides strong results on public datasets we could identify COVID 19 top receptor



https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04484-y

Predicting miRNA-disease associations

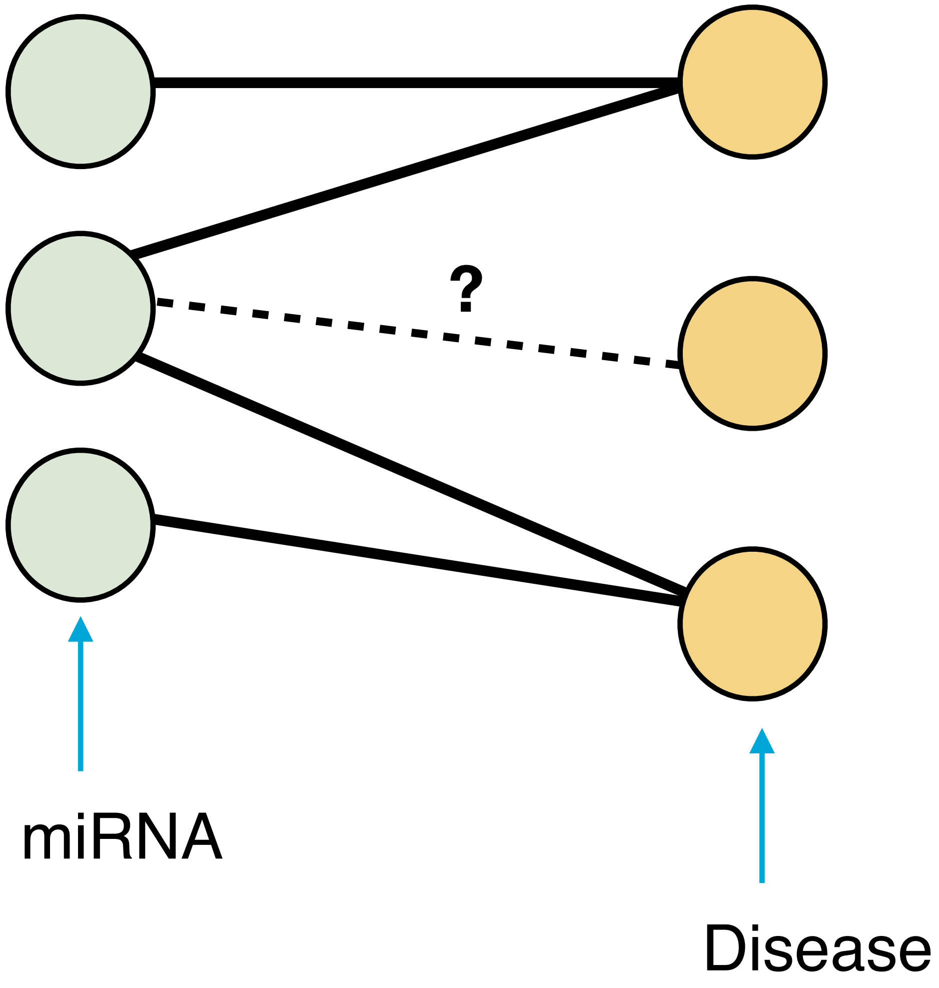
Data bias

20% of the diseases account for 80% of associations

Data scarcity

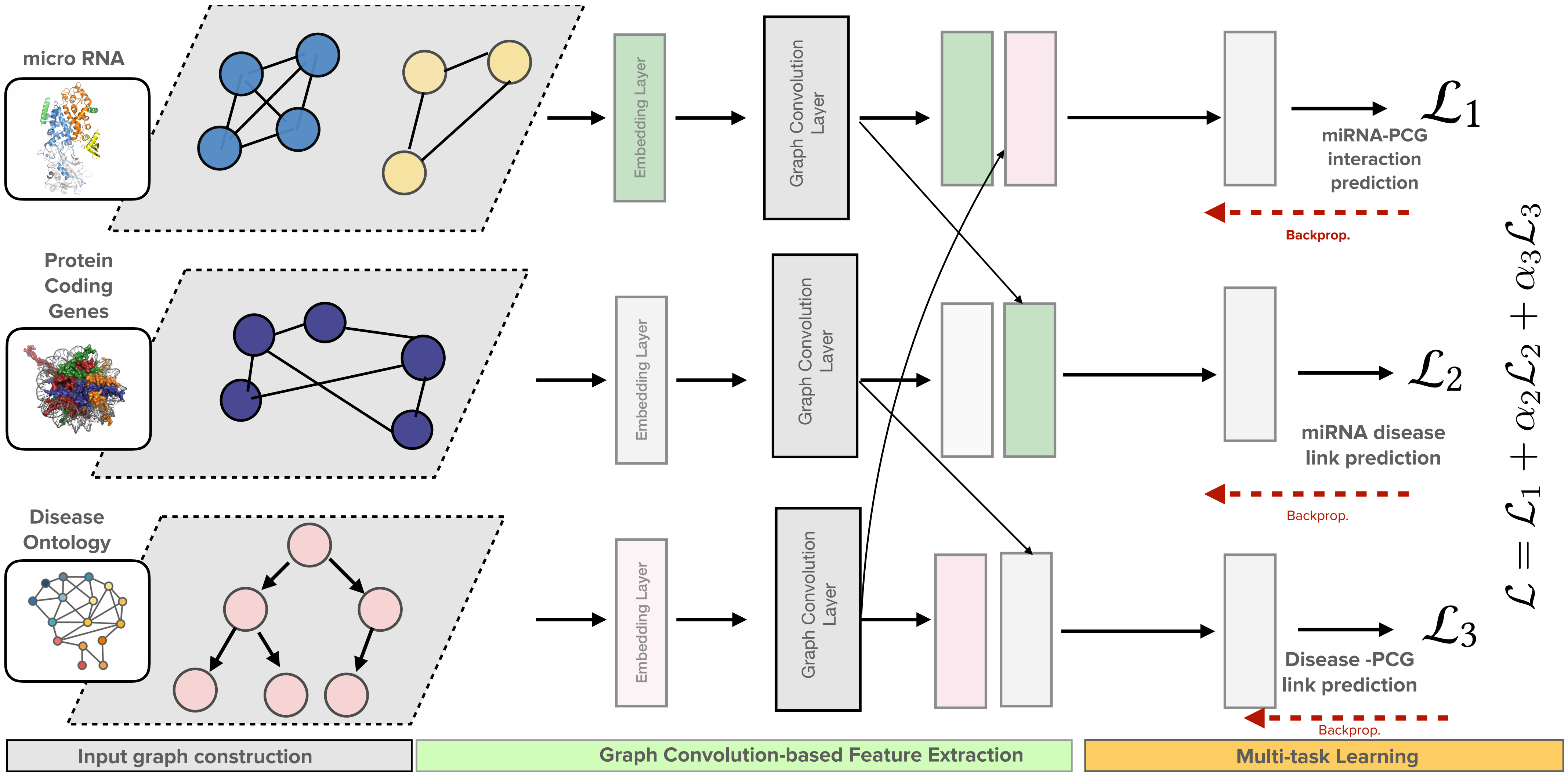
Sparse bipartite graph with small number of nodes

High number of false positives in training data



Overall strategy: Learn jointly from miRNA family, miRNA-gene, disease-gene interactions and disease ontology information

Join learning framework


micro RNA
Protein Coding Genes
Disease Ontology
Embedding Layer
Graph Convolution Layer
$\mathcal{L}_1$
miRNA-PCG interaction prediction
Backprop.
$\mathcal{L}_2$
miRNA disease link prediction
Backprop.
$\mathcal{L}_3$
Disease -PCG link prediction
Backprop.
$\mathcal{L} = \mathcal{L}_1 + \alpha_2\mathcal{L}_2 + \alpha_3\mathcal{L}_3$
Input graph construction
Graph Convolution-based Feature Extraction
Multi-task Learning

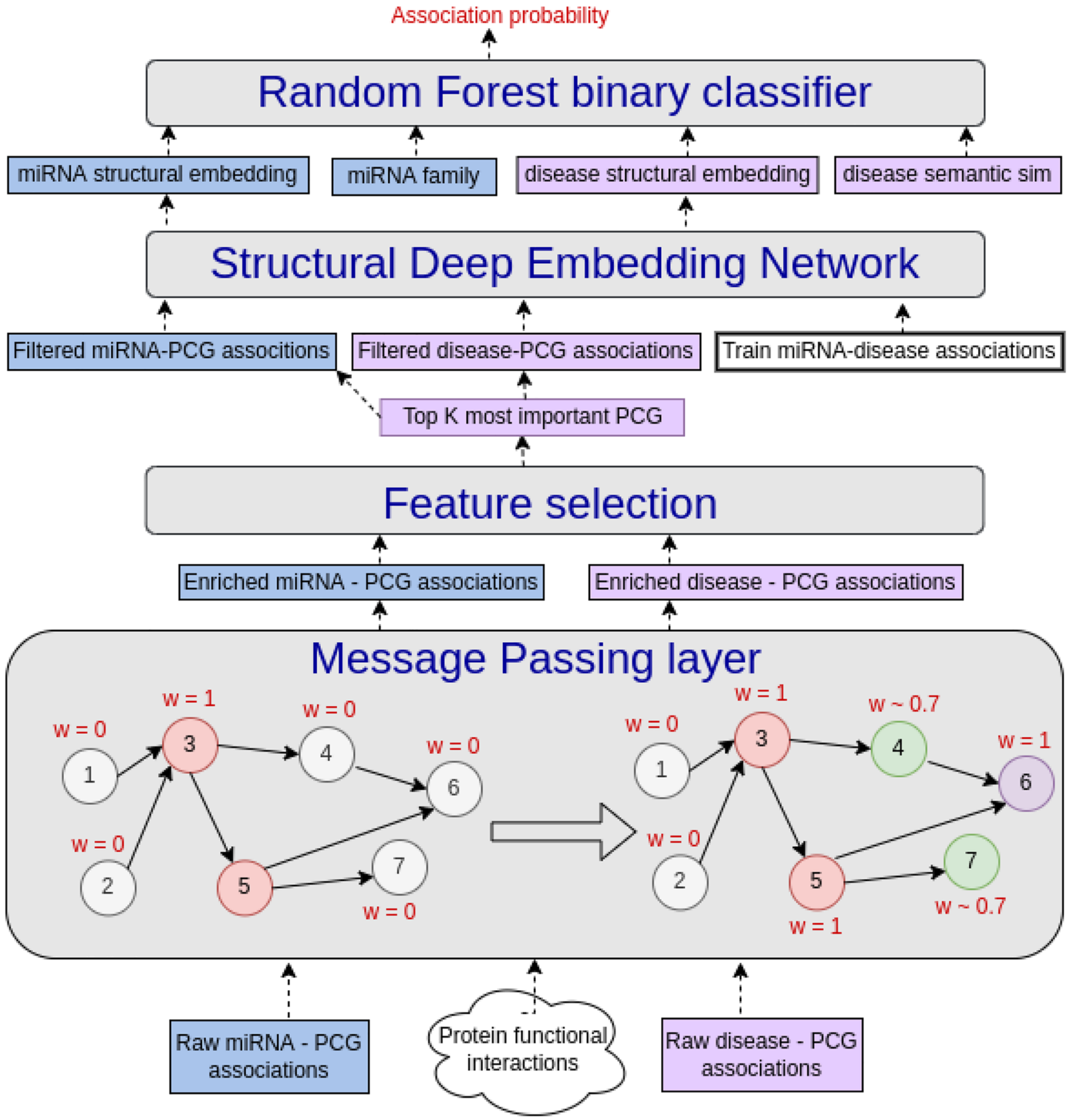
https://ieeexplore.ieee.org/abstract/document/9779549

How to filter training data

- Message passing to enrich gene associations

- Feature selection to select important genes

- Shallow network embeddings over Heterogeneous association network

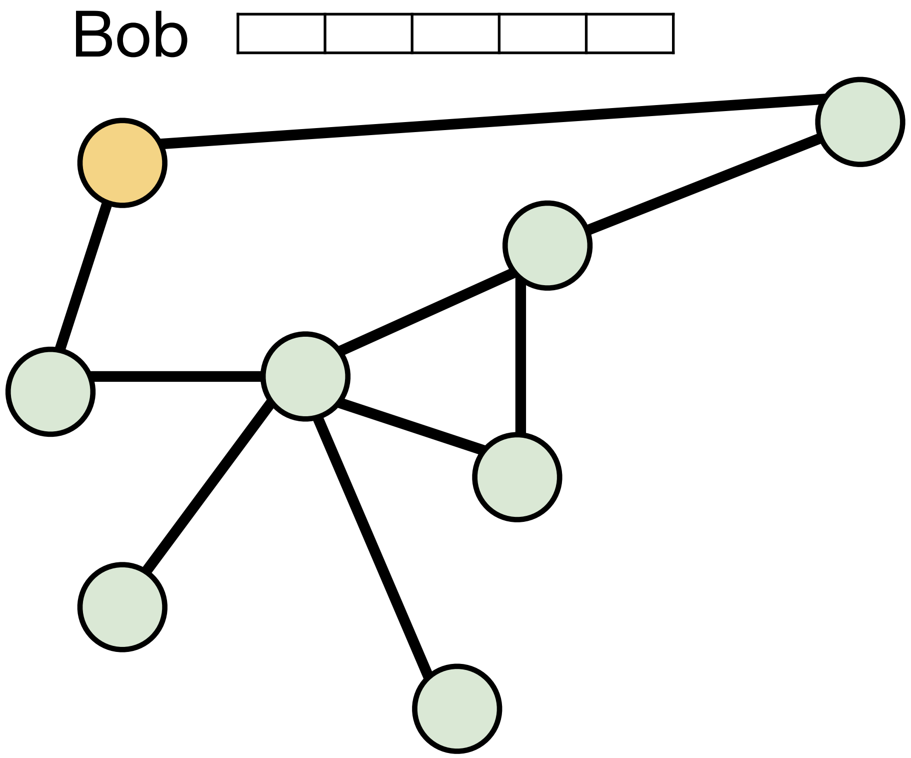

Association probability
Random Forest binary classifier
miRNA structural embedding
miRNA family
disease structural embedding
disease semantic sim
Structural Deep Embedding Network
Filtered miRNA-PCG associtions
Filtered disease-PCG associations
Train miRNA-disease associations
Top K most important PCG
Feature selection
Enriched miRNA - PCG associations
Enriched disease - PCG associations
Message Passing layer
w = 0
w = 1
w = 0
w = 0
w = 0
w = 0
1
2
3
4
5
6
7
w = 0
w = 1
w ~ 0.7
w = 1
w = 0
w = 1
w ~ 0.7
Raw miRNA - PCG associations
Protein functional interactions
Raw disease - PCG associations

https://www.nature.com/articles/s41598-022-20529-5

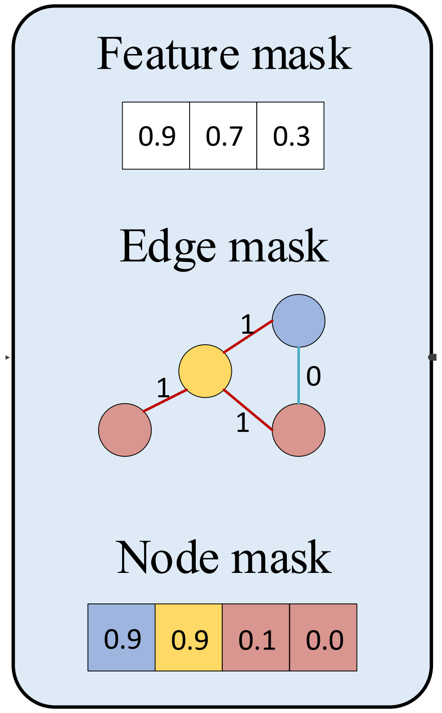
Challenges of transparency and privacy

Transparency

Why was Bob’s loan denied?

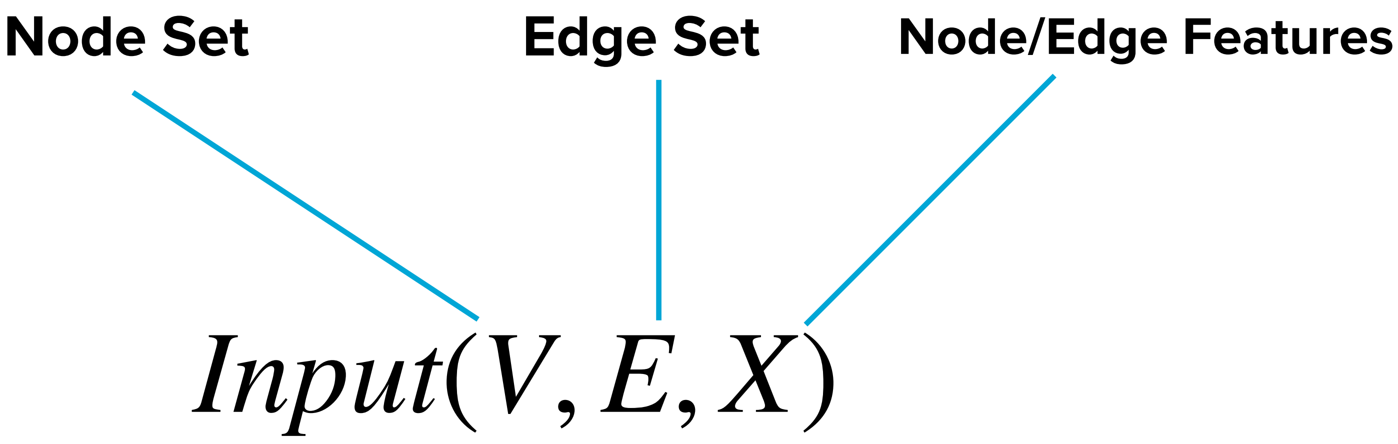


Explanation

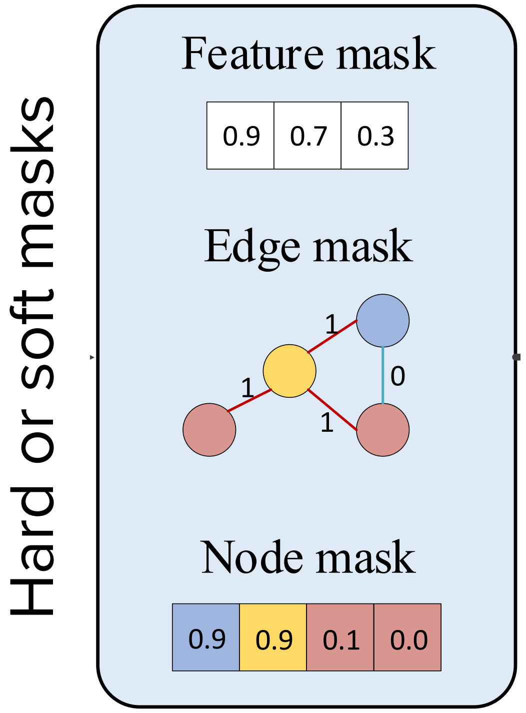


Decision has to be explained not only in terms of features but also graph structure. General explainability methods cannot be trivially applied for graphs.

Post-hoc explanations



Explanation types



Examples: GNNExplainer, Zorro, PGExplainer

Explanation types:

Feature explanations in terms of most relevant features $X' \subset X$

Structure explanations in terms of most relevant nodes $(V' \subset V)$ or edges $(E' \subset E)$

We are interested in finding both feature and structure explanations which effectively capture interplay of structure and features in model's decision making.

Privacy

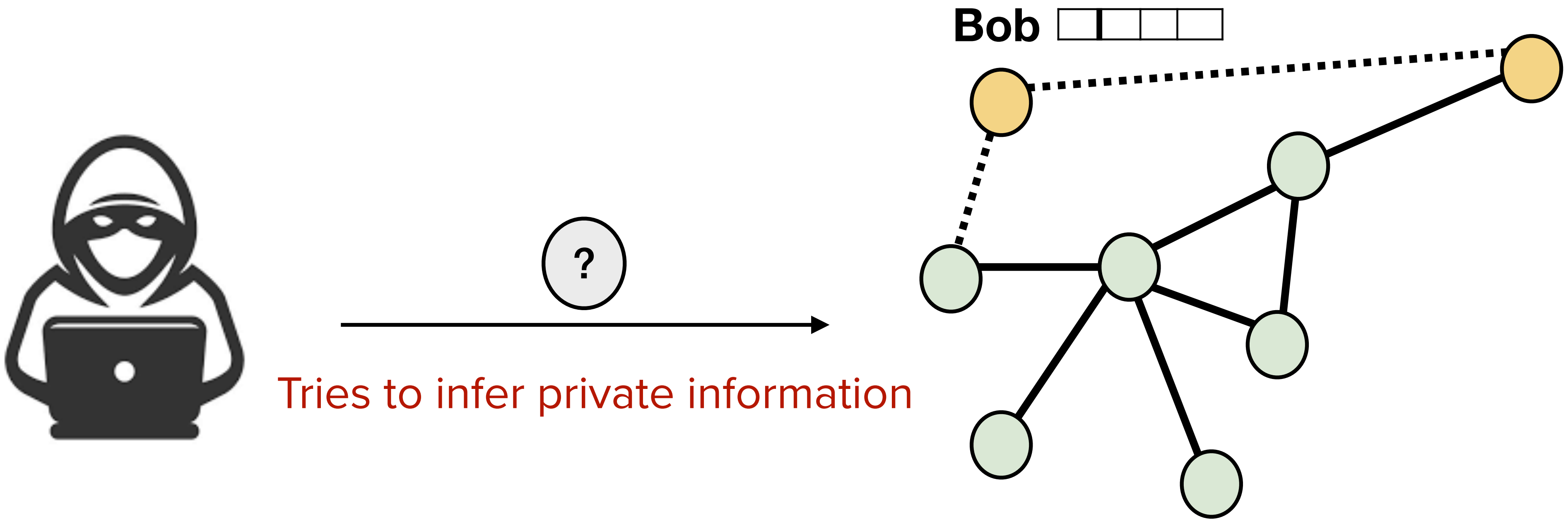
Graphs can contain sensitive information

- User's sensitive attributes
- Sensitive relations

GNNs encode relation information within the model, could memorise such information

- Your identity could be revealed because of your neighbour

Privacy



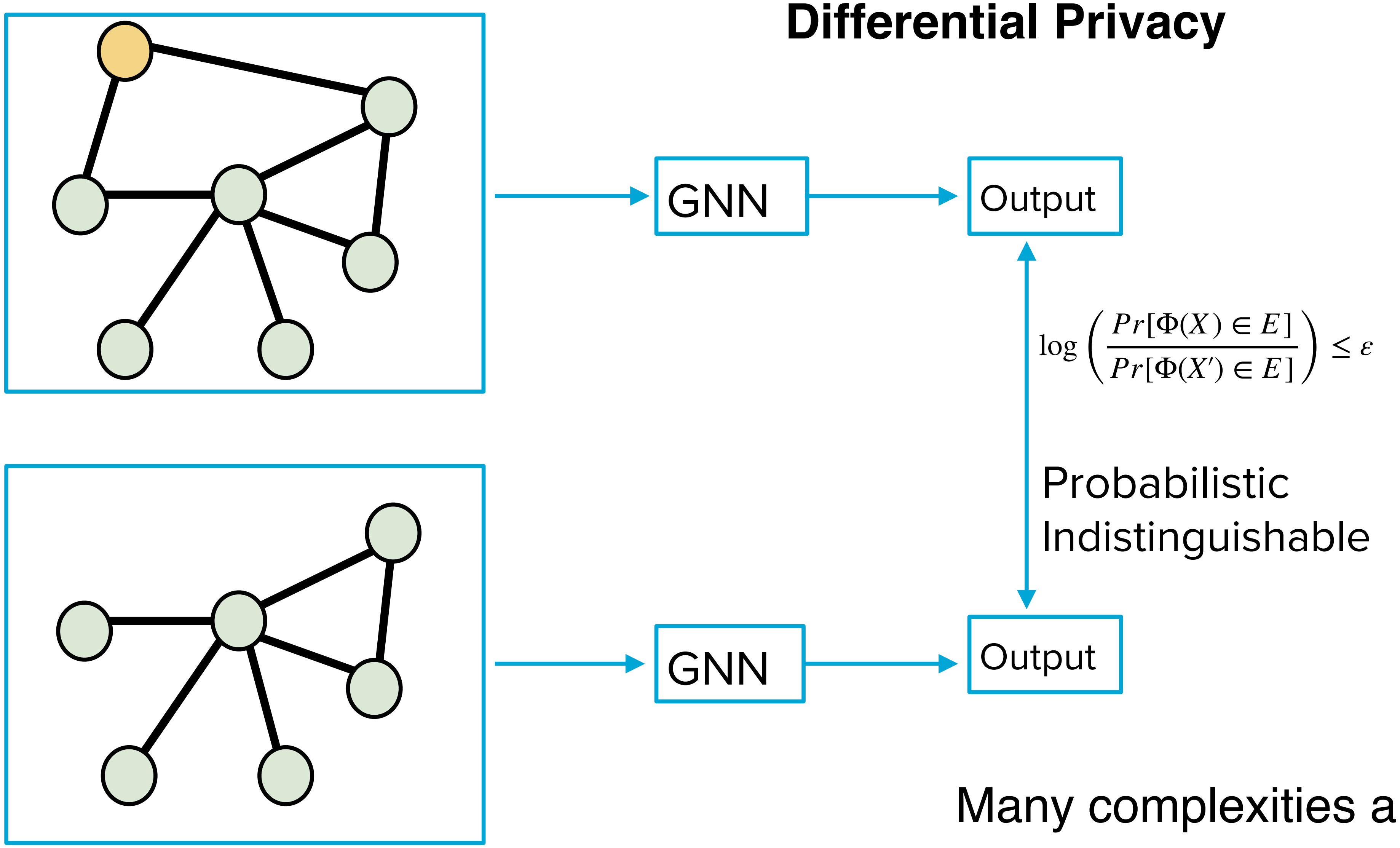
Node Membership Inference : Is Bob a part of training data? [Olatunji et al., ’21] [Duddu et al., ’20]

Relation reconstruction : Who are friends of Bob? [He et al., ’21] [Zhang et al., ’20]

Attribute Inference : Does Bob smoke?

Building Private GNN Models

Differential Privacy



$$\log\left(\frac{Pr[\Phi(X) \in E]}{Pr[\Phi(X') \in E]}\right) \leq \varepsilon$$

Probabilistic Indistinguishable

Many complexities arise…..

Building Privacy Preserving models for graphs

A direct application of techniques like DP-SGD is not possible due to

- Unbounded sensitivity (think of the effect of leaving out or adding one node in a graph)

- Violation of i.i.d. assumption

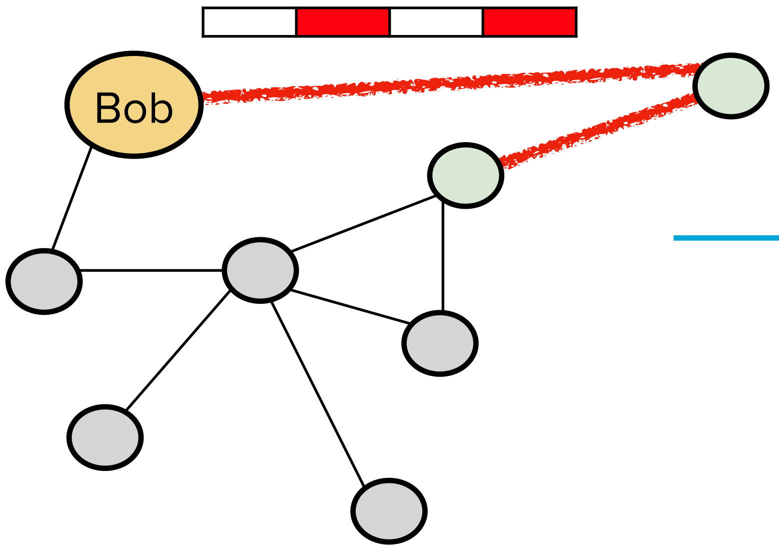
- Need for inference privacy (as training data might be used during inference)

PrivGNN (Olatunji, Funke, Khosla, 2021), GAP (Sajadmanesh, Shamsabadi, A, Bellet, et al. 2022)

Transparency - Privacy Tradeoffs

But we want our models to be **transparent** and **private** simultaneously

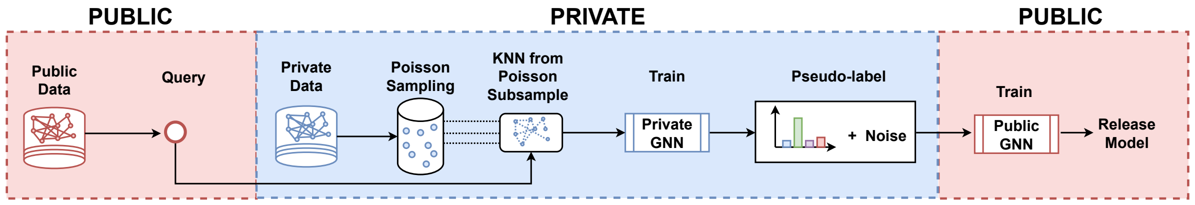
Transparency



Can we explain private models? How?

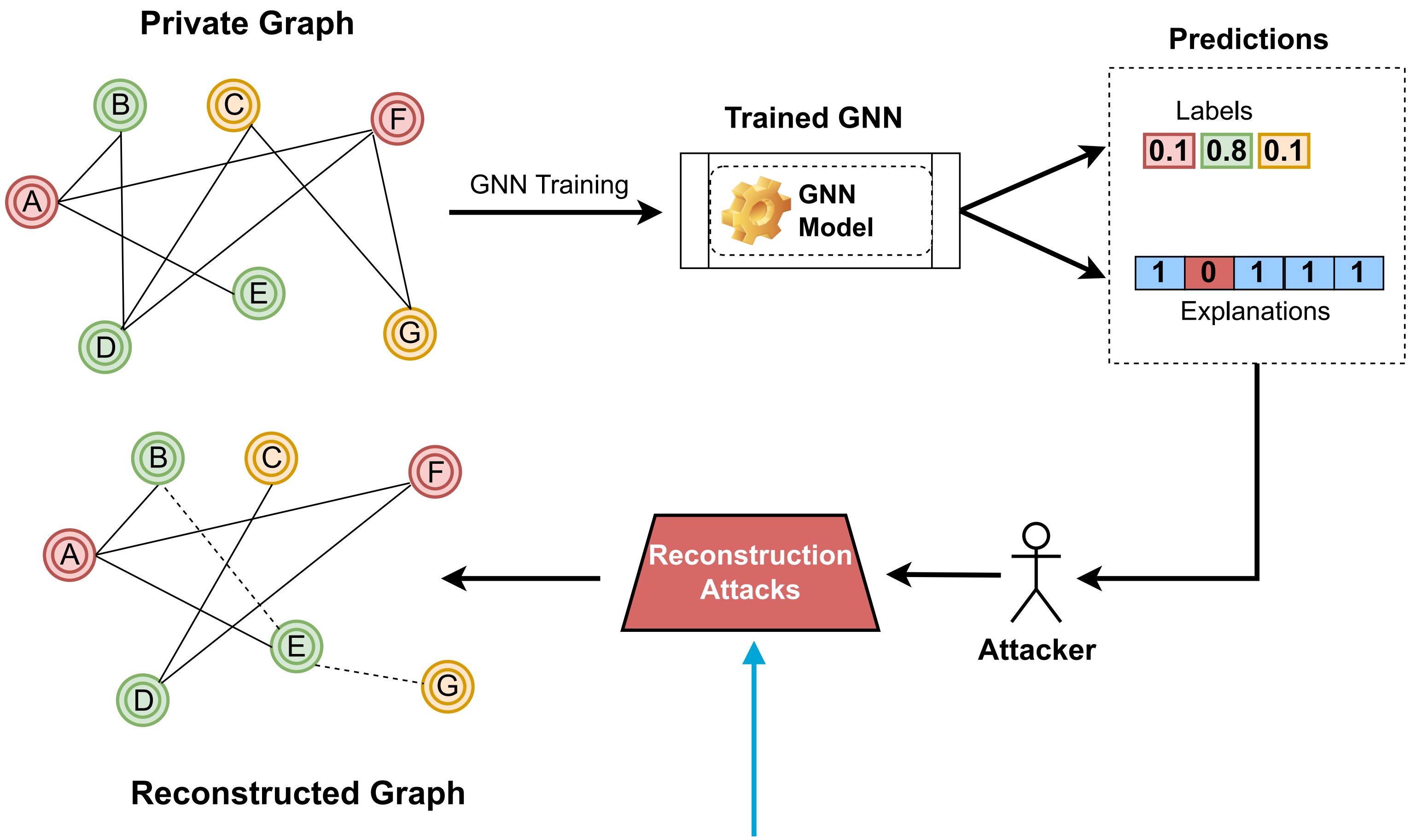
Privacy

How much can the privacy be hurt?
How to preserve privacy?



Complex privacy preserving mechanism

Reconstructing graphs from feature explanations



Graph Structure Learning

Private Graph Extraction via Feature Explanations [Olatunji et al. 2022]
https://arxiv.org/abs/2206.14724

Some interesting findings

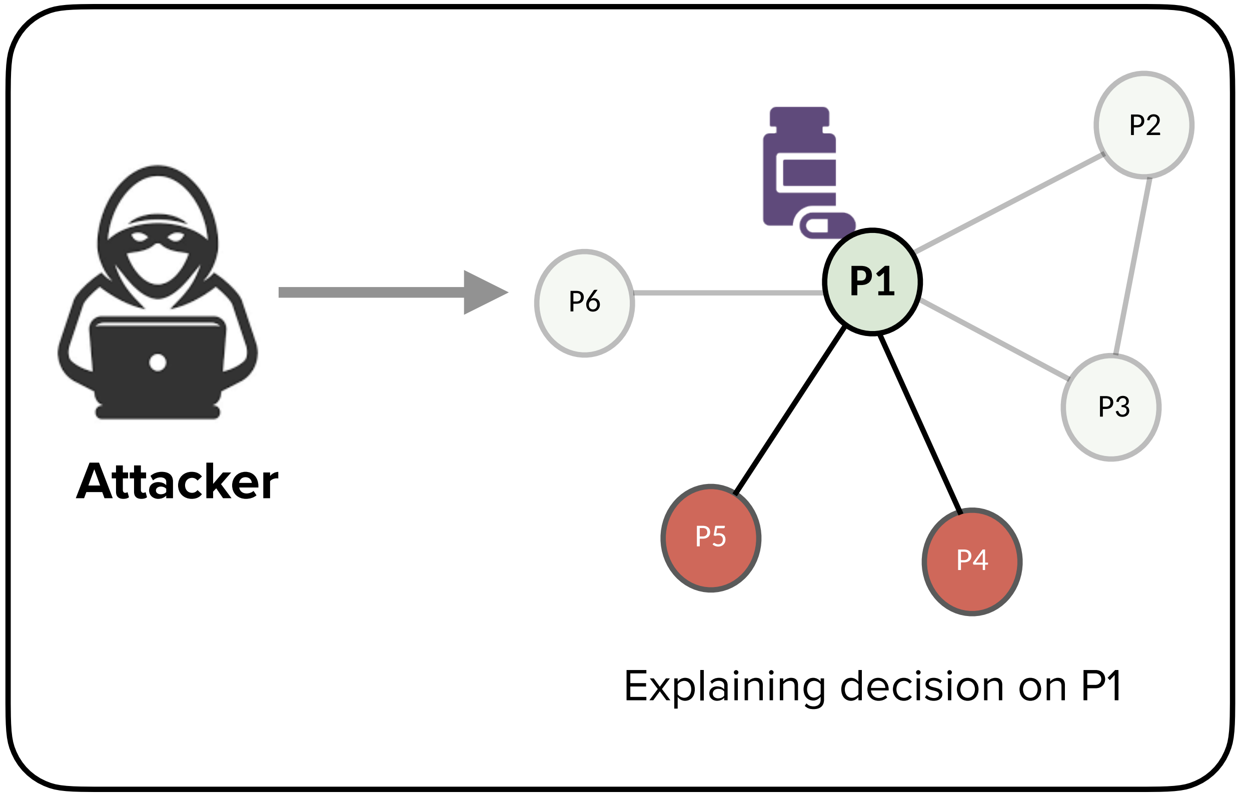
- Training graph could be reconstructed using alone the feature explanations and the labels

- Certain explanations leak more information than others

- Gradient based explanations incur high privacy loss while showing low **utility** (quantified by high faithfulness and sparsity)

Challenges

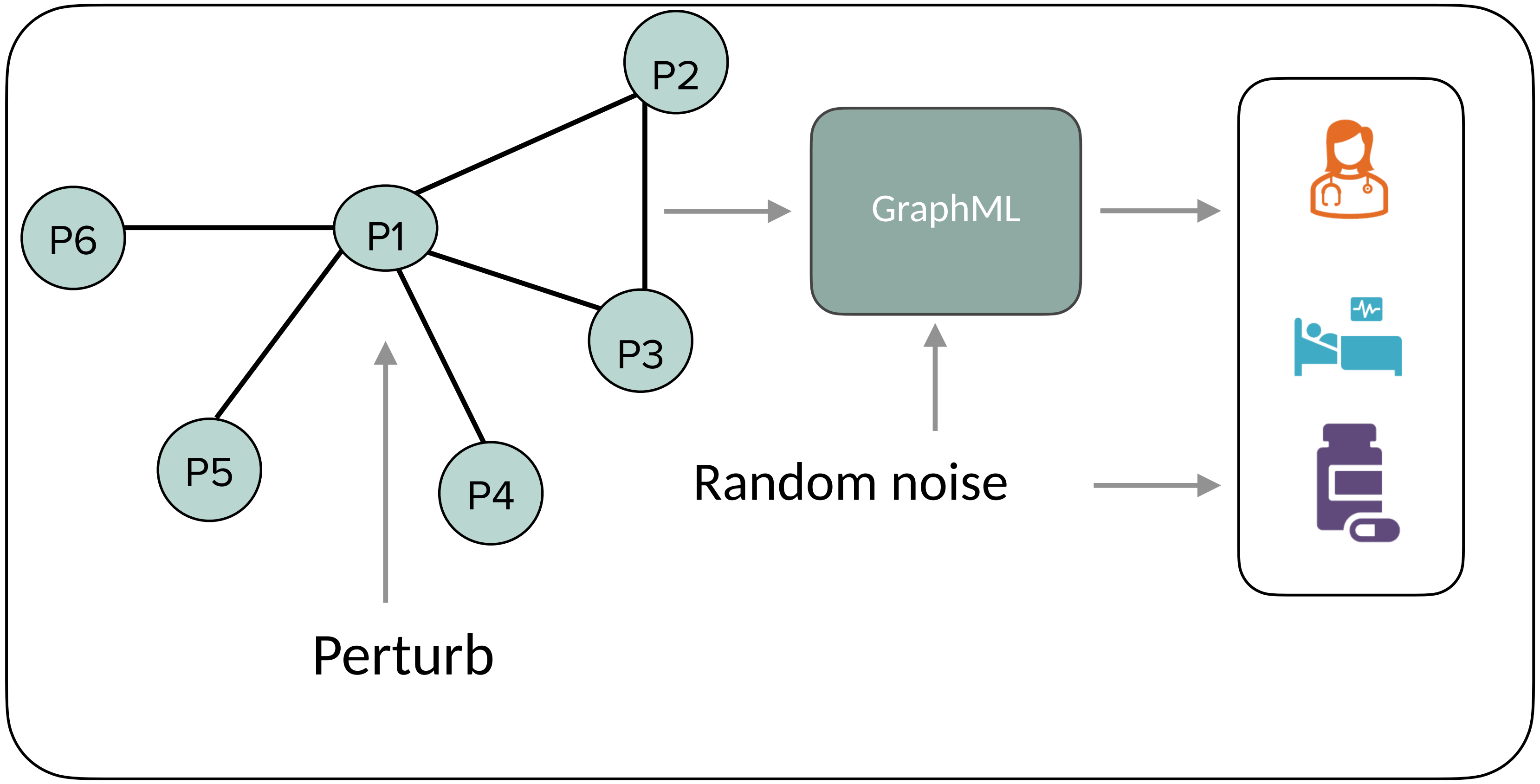
Structure explanations can directly reveal information about neighbours



Explanations of neighbouring datapoints would be correlated

Challenges

Private learning over graphs is more complex than that in standard ML



How to define explanation for a private model?

Research Directions and Open Questions

Quantification of privacy leakage in presence of different explanation types

- How can we measure information leakage due to different explanation types?
- Risk-utility assessment of different explainers/explanations
- Can we release explanations privately while still maintaining their utility?

Explaining the decisions of privacy-preserving models

- What should be the properties of an explanation for a privacy-preserving model?
    - Such properties might need to be defined based on the private learning strategy
- How to release such explanations in a private manner?

Joint optimization of privacy and transparency

- How can we optimise for the combined requirements of privacy and transparency in GraphML?

Privacy and Transparency in Graph Machine Learning: A Unified Perspective, M.Khosla. In AIMLAI@CIKM’22
https://arxiv.org/abs/2207.10896