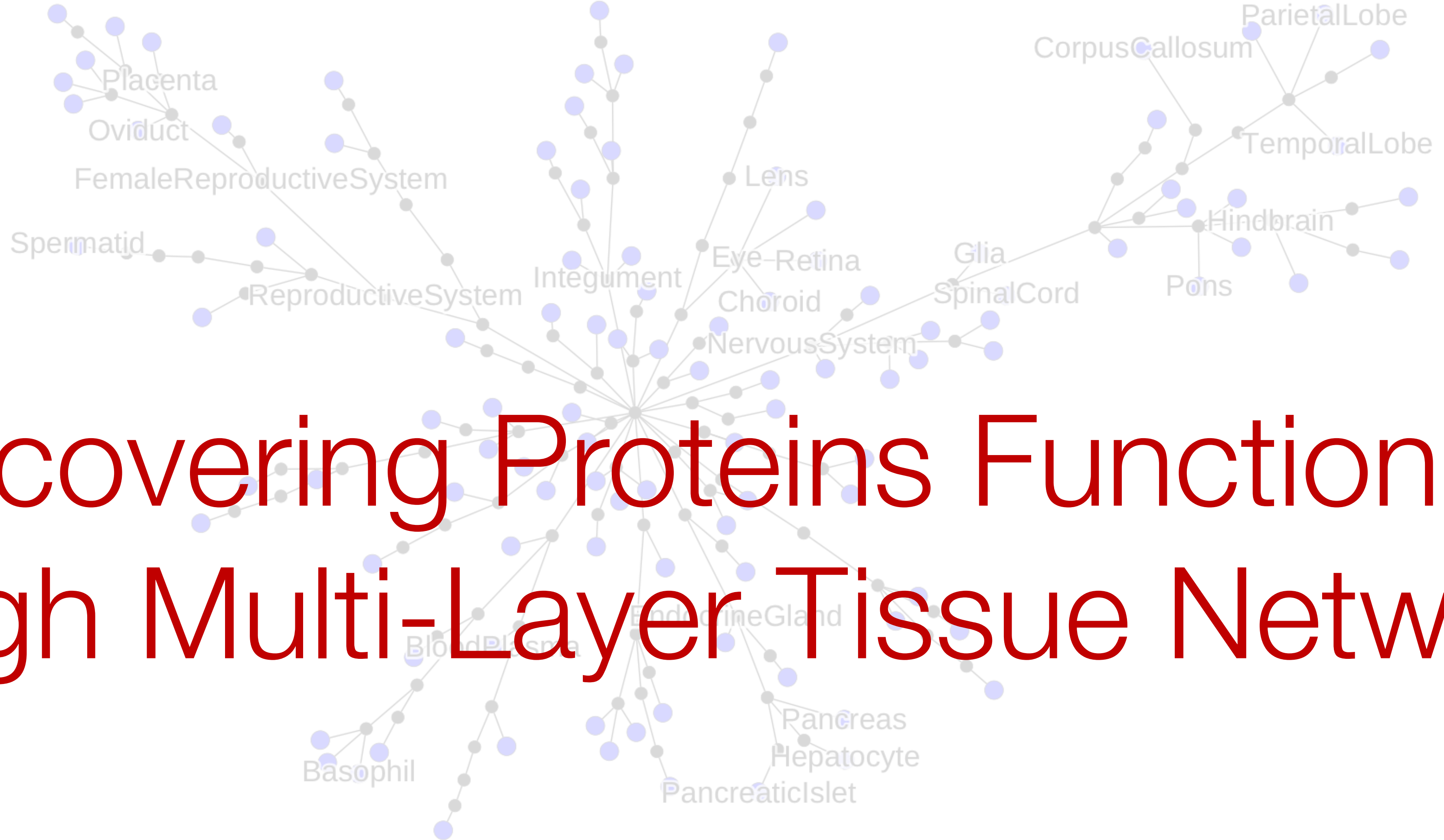


Uncovering Proteins Functions Through Multi-Layer Tissue Networks

Marinka Zitnik
marinka@cs.stanford.edu
Joint work with Jure Leskovec

Why tissues?

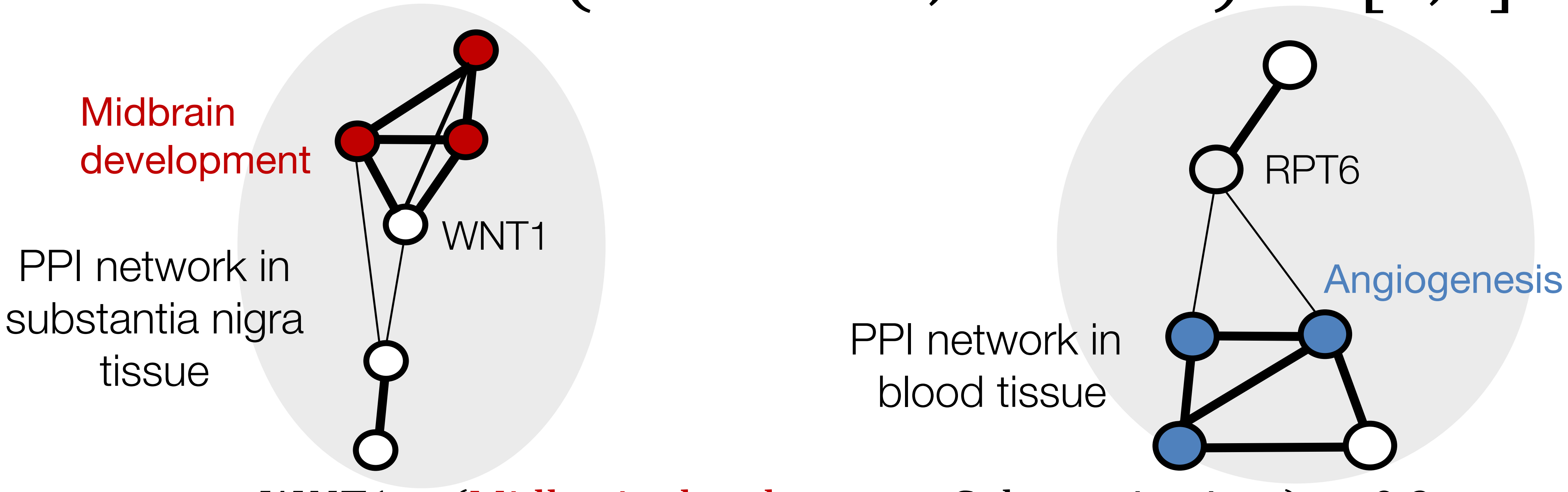
A unified view of cellular functions across human tissues is essential for understanding biology, interpreting genetic variation, and developing therapeutic strategies

[Greene et al. 2015, Yeger & Sharan 2015, GTEx and others]

What Does My Protein Do?

Goal: Given a set of proteins and possible functions, predict each protein's association with each function

$$\text{Proteins} \times (\text{Functions}, \text{Tissues}) \rightarrow [0,1]$$



$$WNT1 \times (\text{Midbrain development}, \text{Substantia nigra}) \rightarrow 0.9$$

$$\text{RPT6} \times (\text{Angiogenesis}, \text{Blood}) \rightarrow 0.05$$

Existing Research

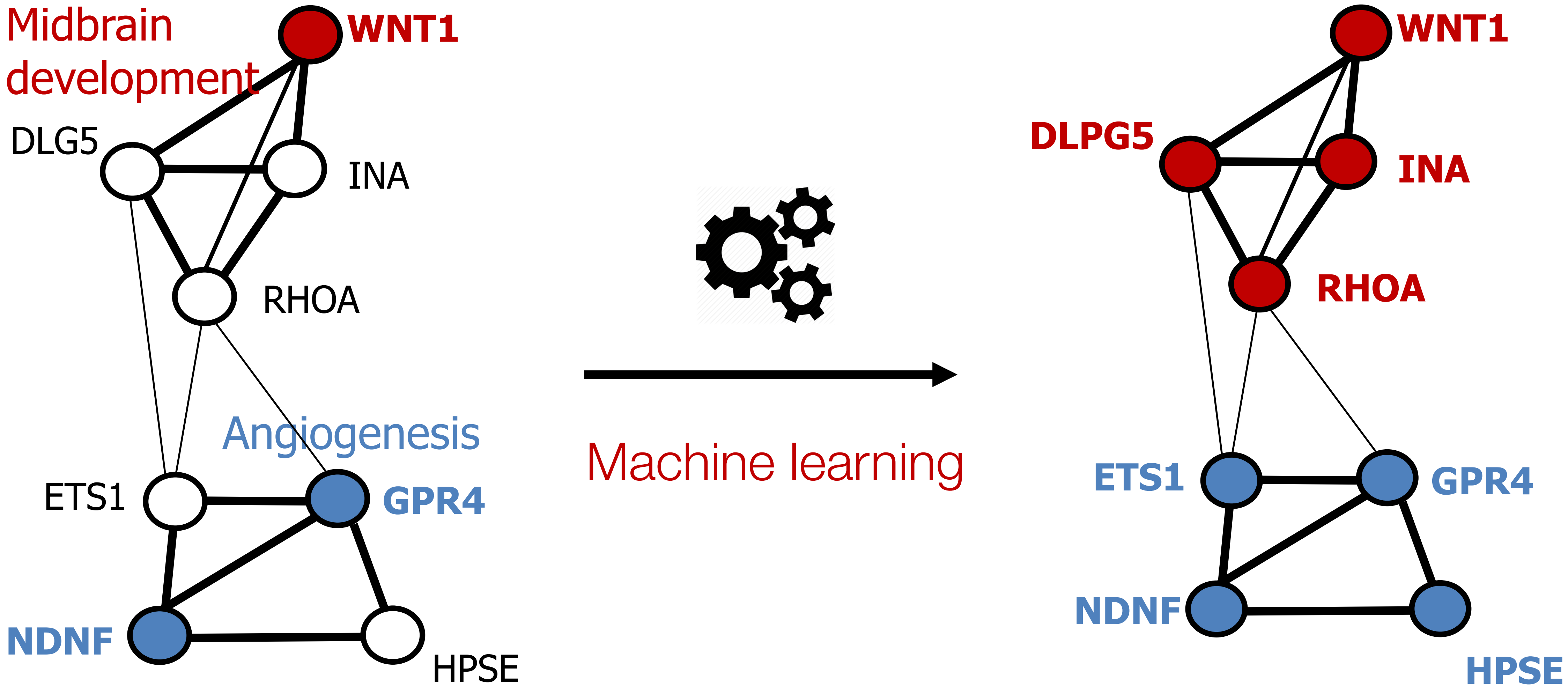
- Guilty by association: protein's function is determined based on who it interacts with [Zuberi et al. 2013, Radivojac et al. 2013, Kramer et al. 2014, Yu et al. 2015] and many others]
  - No tissue-specificity
- Protein functions are assumed constant across organs and tissues:
  - Functions in heart are the same as in skin

Lack of methods for predicting protein functions in different biological contexts

Challenges

- Tissues have inherently multiscale, hierarchical organization
- Tissues are related to each other:
  - Proteins in biologically similar tissues have similar functions [Greene et al. 2015, ENCODE 2016]
  - Proteins are missing in some tissues
- Interaction networks are tissue-specific
- Many tissues have no annotations

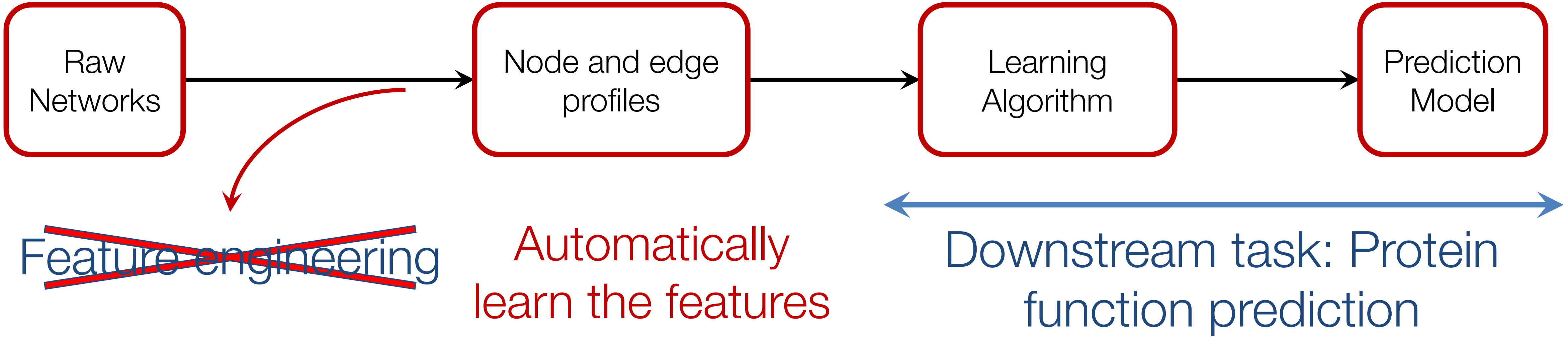
Machine Learning in Networks



Multi-label node classification: midbrain development, angiogenesis, etc.

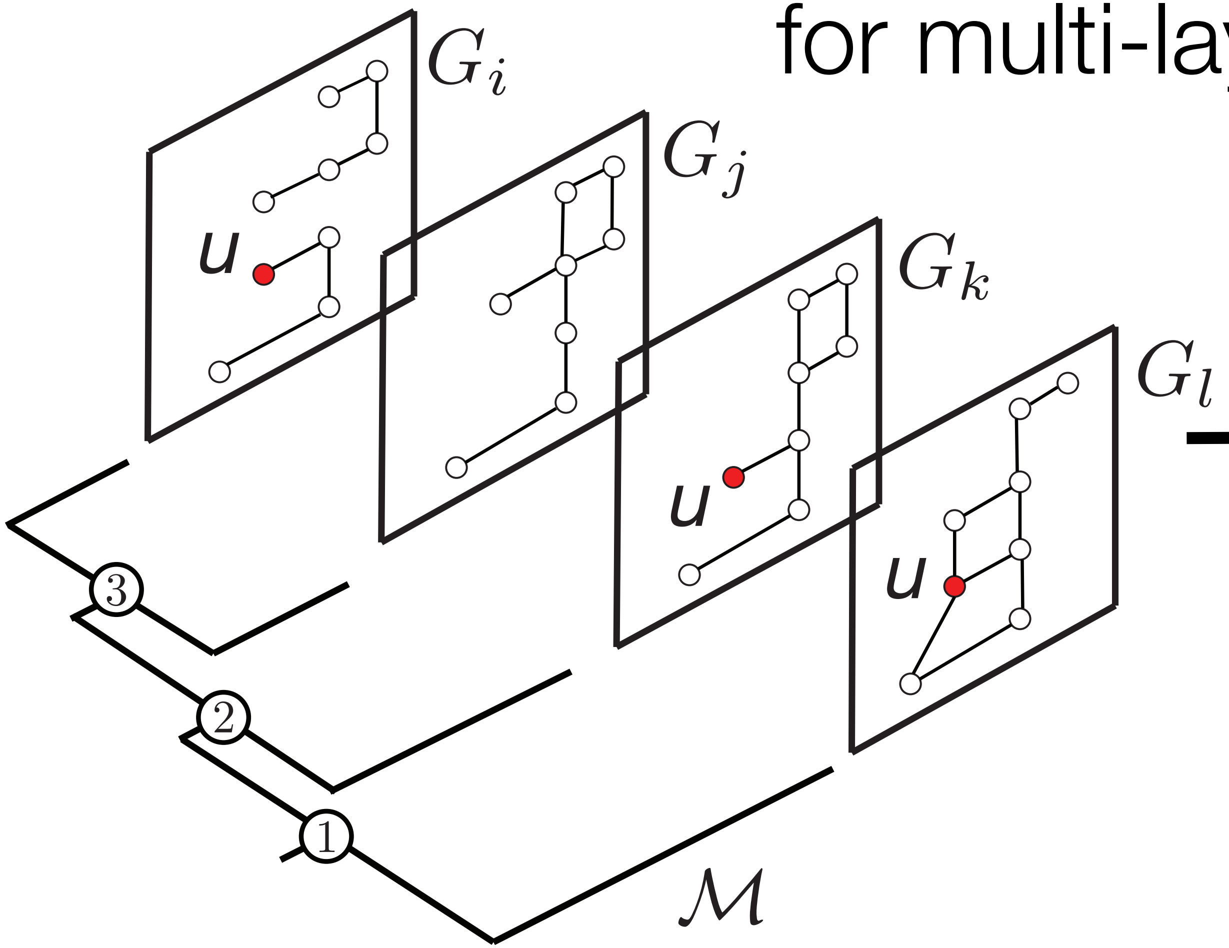
Machine Learning Lifecycle

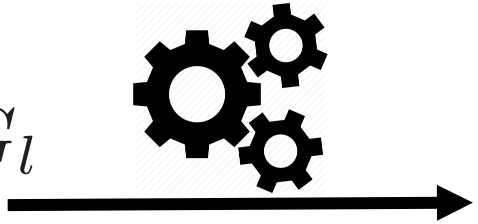
- Machine learning lifecycle: This feature, that feature
- Every single time!



Feature Learning in Multi-Layer Graphs

OhmNet: Unsupervised feature learning for multi-layer networks

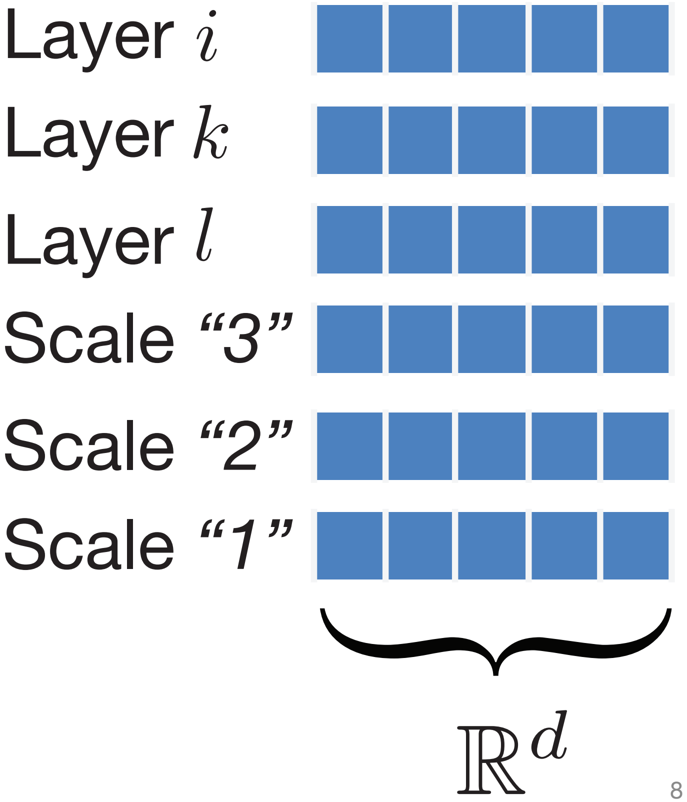




f_i, f_k, f_l
f_3, f_2, f_1

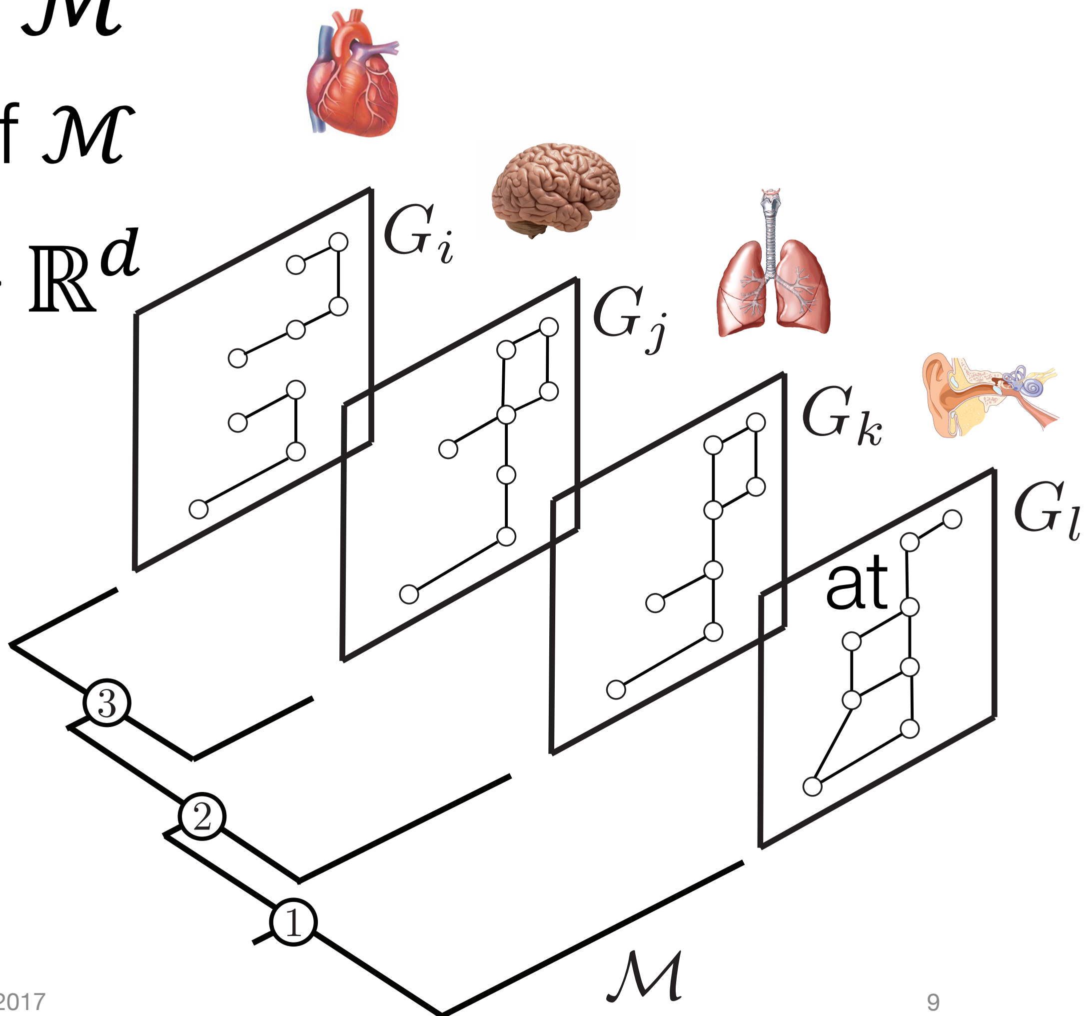
$u \rightarrow \mathbb{R}^d$

Vectors, node embeddings



Features in Multi-Layer Tissue Network

- Given: Layers $\{G_i\}_i$, hierarchy $\mathcal{M}$
  - Layers $\{G_i\}_{i=1..T}$ are in leaves of $\mathcal{M}$
- Goal: Learn functions: $f_i\colon V_i \to \mathbb{R}^d$
- Multi-scale model:
  - Learn node embeddings at each possible scale
  - Layers i, j, k, l
  - Scales "3", "2", "1"



OhmNet Learning Approach

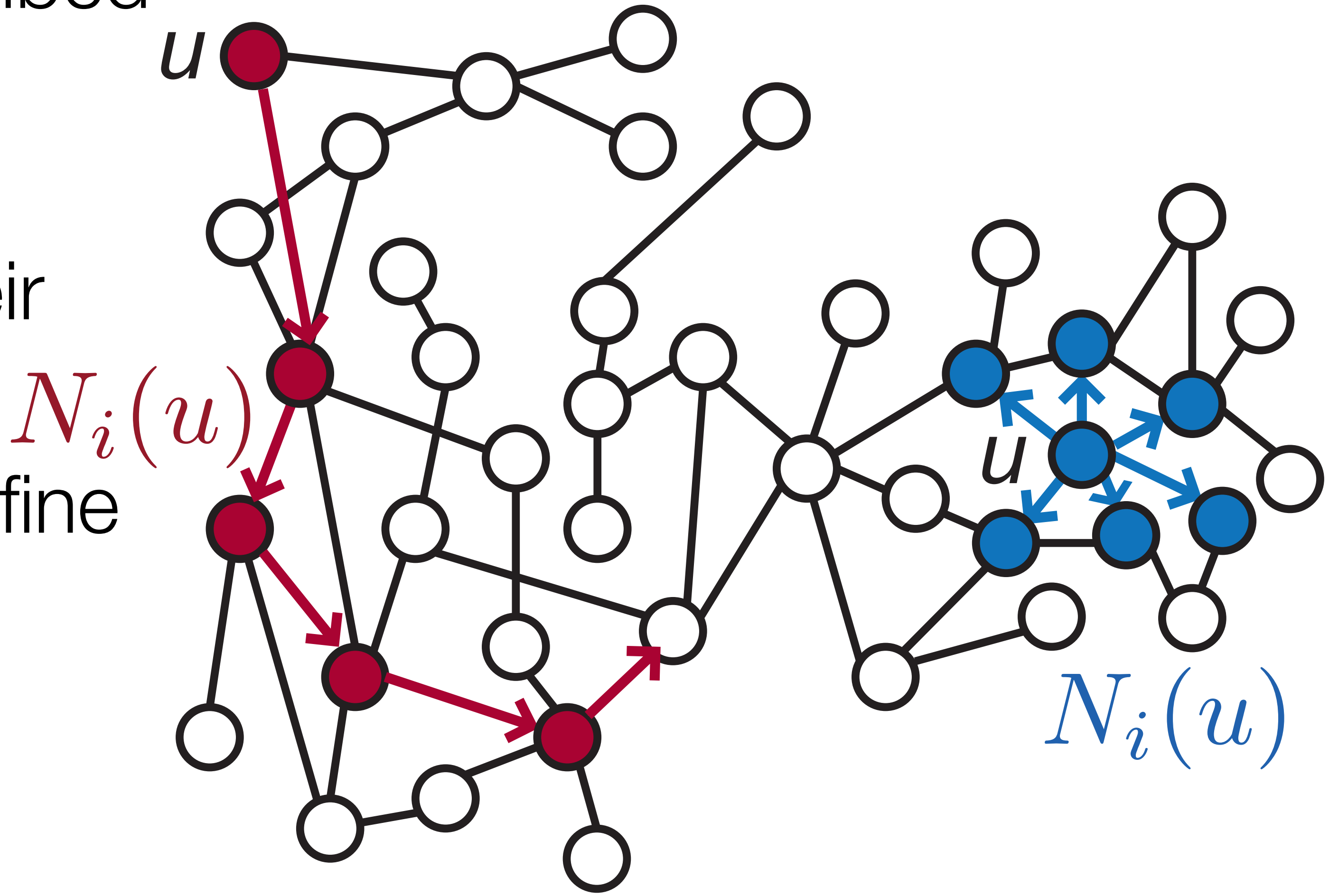
OhmNet has two components:

1. Single-layer objectives
   Nodes with similar network neighborhoods in each layer are embedded close together

2. Hierarchical dependency objectives
   Nodes in nearby network layers in the hierarchy share similar features

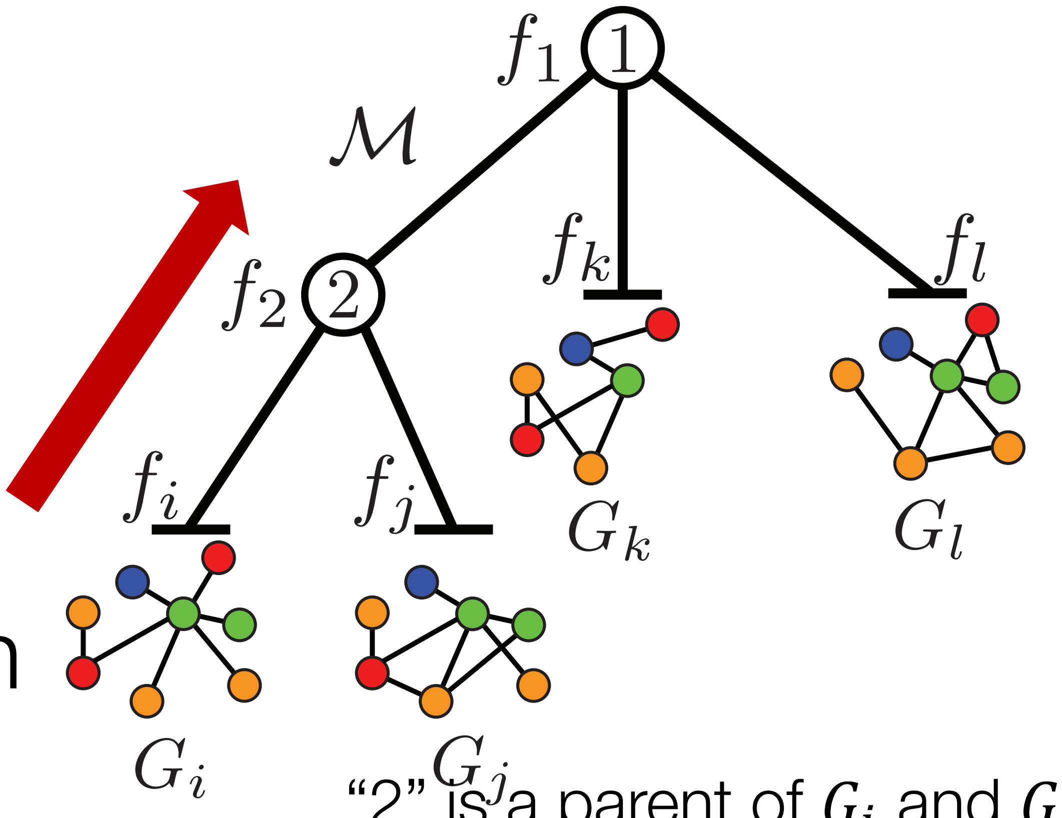
Single-Layer Objectives

- Intuition: For each layer, embed nodes to d dimensions by preserving their similarity
- Two nodes are similar if their neighborhoods are similar
- For node u in layer i we define nearby nodes as nodes in G_i visited by random walks starting at u



Dependencies Between Network Layers

- Intuition: Proteins in biologically similar tissues share similar features
- Use tissue hierarchy to recursively regularize features at i to be similar to features in i's parent

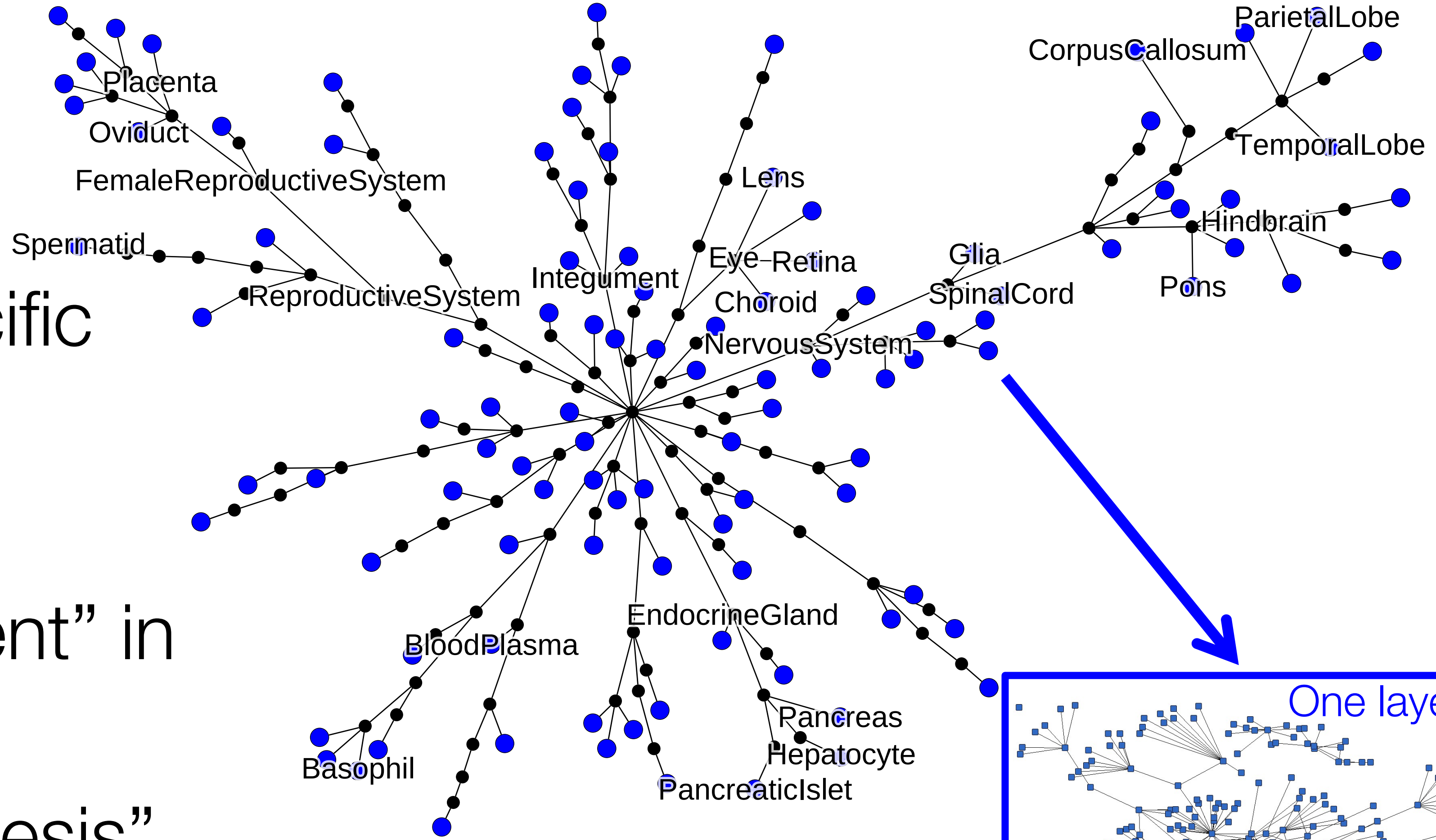


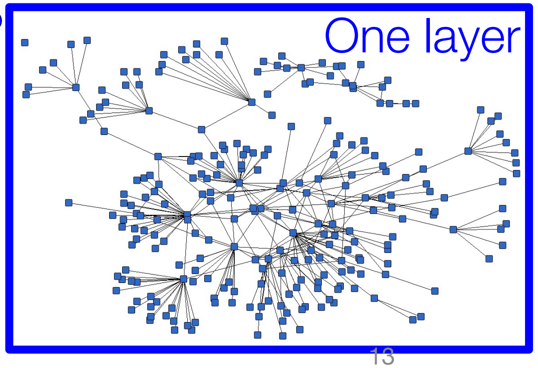
"2" is a parent of G_i and G_j

OhmNet generates multi-scale node embeddings

Data: 107 Tissue Layers

- Layers are PPI nets:
  - Nodes: proteins
  - Edges: tissue-specific PPIs
- Node labels:
  - “Cortex development” in renal cortex tissue
  - “Artery morphogenesis” in artery tissue





Experimental Setup

- Protein function prediction is a multi-label node classification task
- Every node (protein) is assigned one or more labels (functions)

- Setup:
  - Learn OhmNet embeddings for multi-layer tissue network
  - Train a classifier for each function based on a fraction of proteins and all their functions
  - Predict functions for new proteins

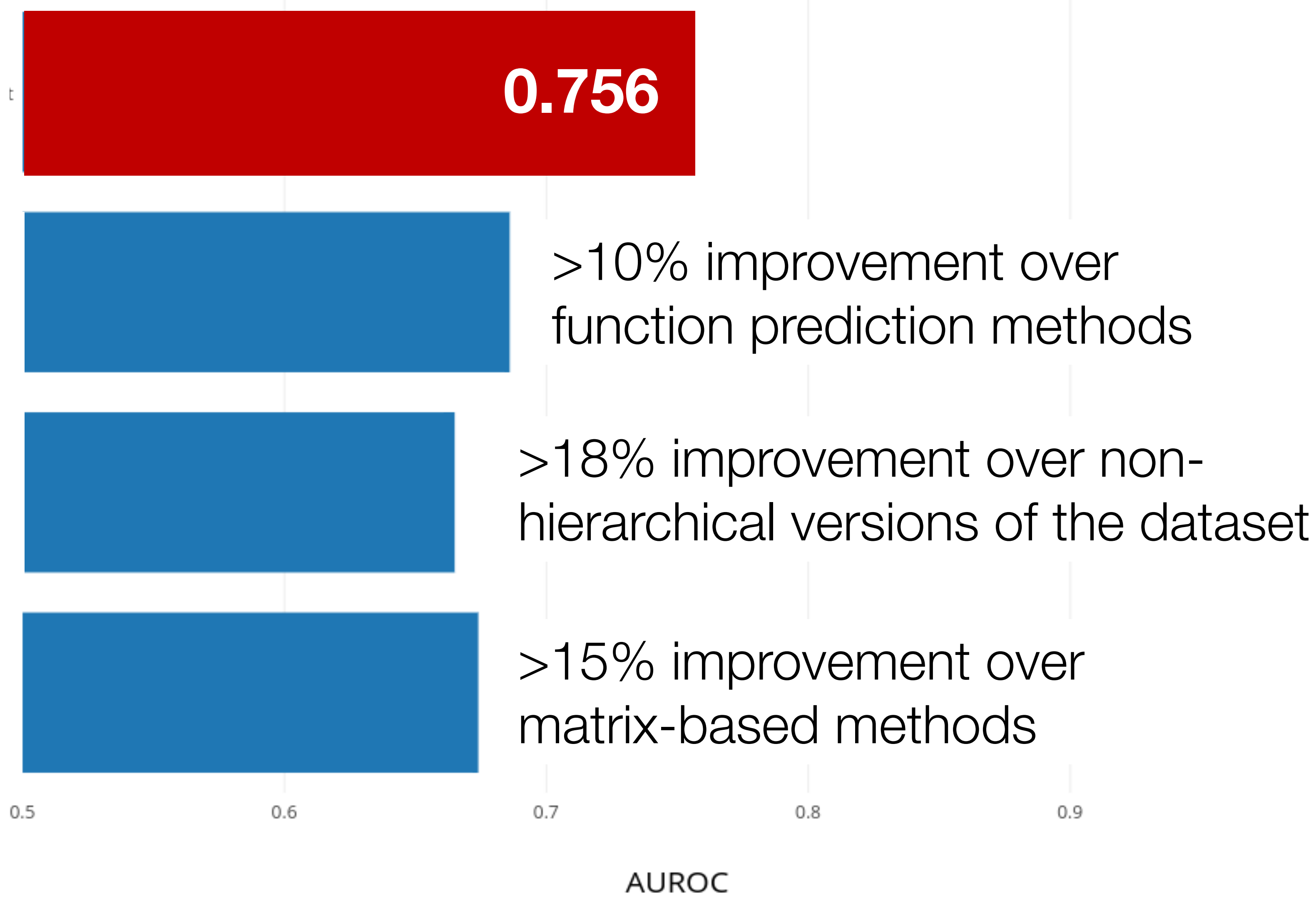
Tissue-Specific Protein Functions

OhmNet

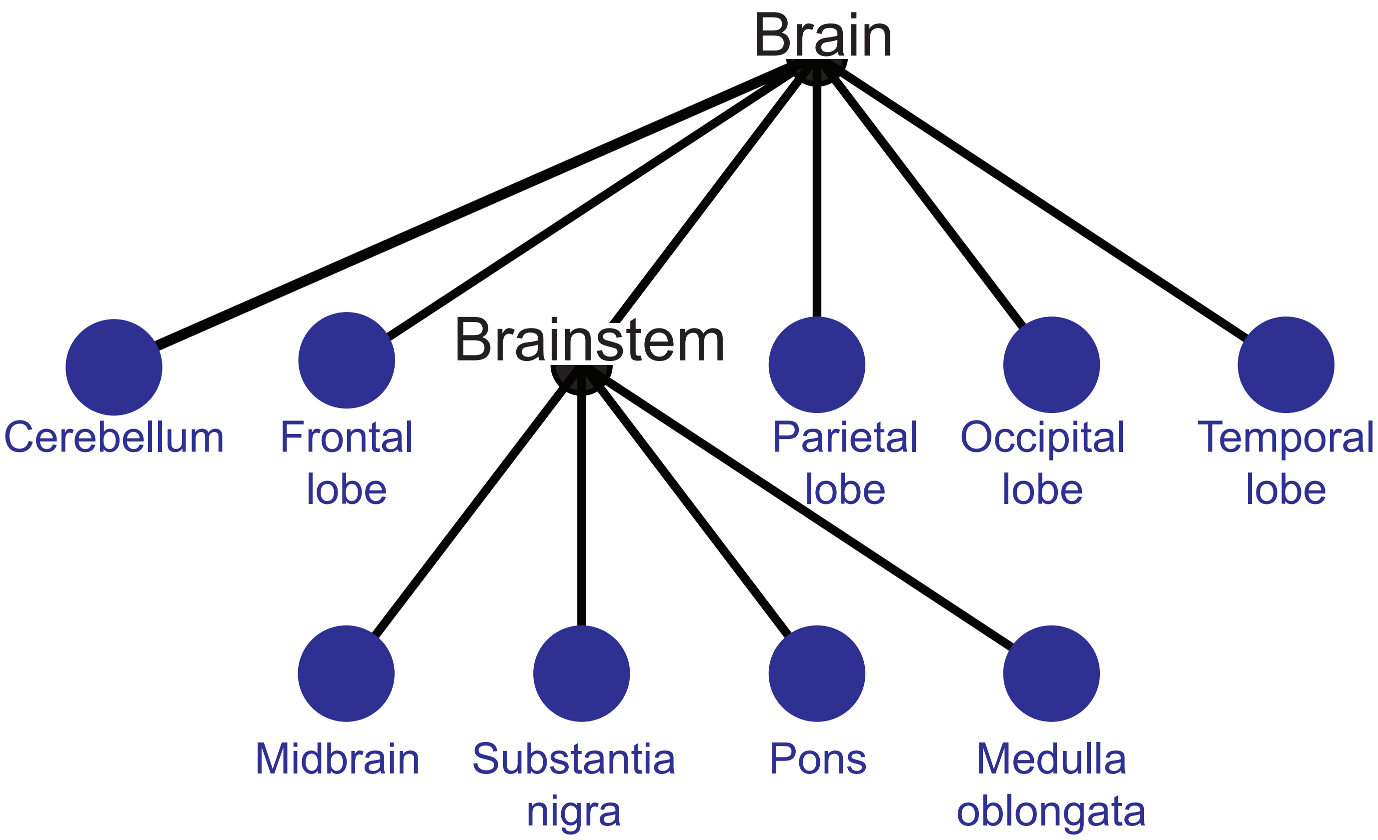
Protein function prediction methods

Mono-layer network embeddings

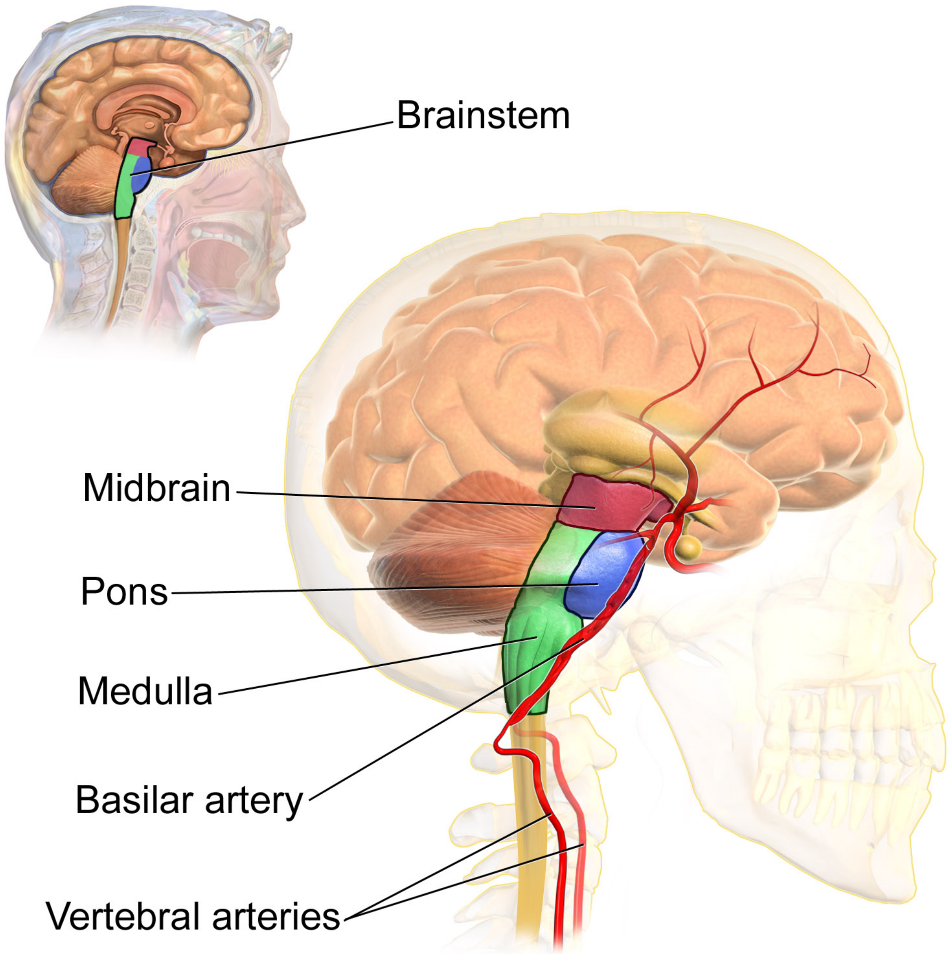
Tensor decompositions



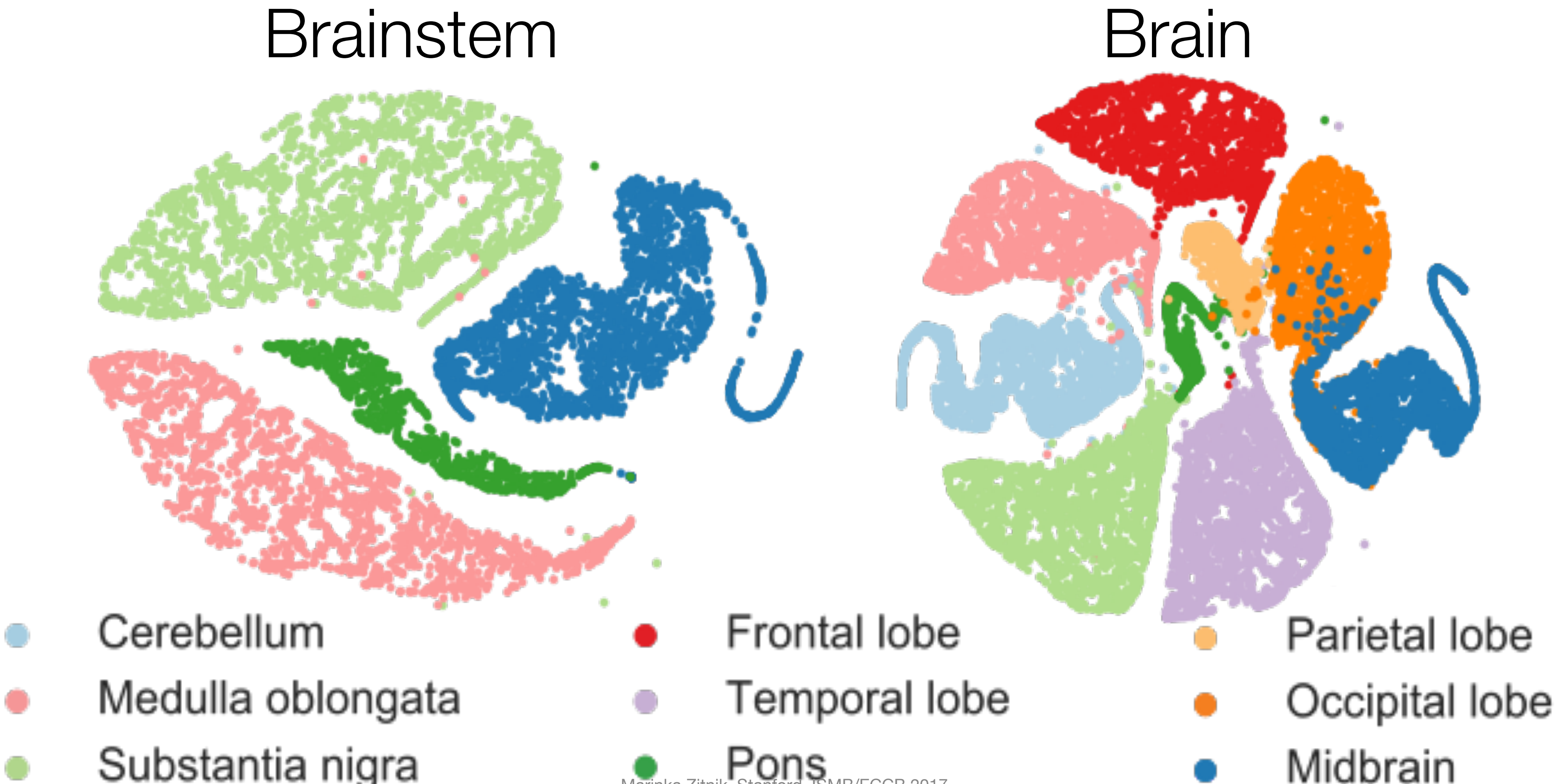
Case Study: 9 Brain Tissues



9 brain tissue PPI networks
in two-level hierarchy



Multi-Scale Node Embeddings



Annotating Proteins in a New Tissue

- Transfer protein functions to an unannotated tissue
- Task: Predict functions in target tissue without access to any annotation/label in that tissue

Target tissue	Tissue-specific (OhmNet)	Tissue non-specific	Improvement
Placenta	0.758	0.684	**11%**
Spleen	0.779	0.712	**10%**
Liver	0.741	0.553	**34%**
Forebrain	0.755	0.632	**20%**
Blood plasma	0.703	0.540	**40%**
Smooth muscle	0.729	0.583	**25%**
Average	0.746	0.617	**21%**

Reported are AUROC values (see paper for other metrics)

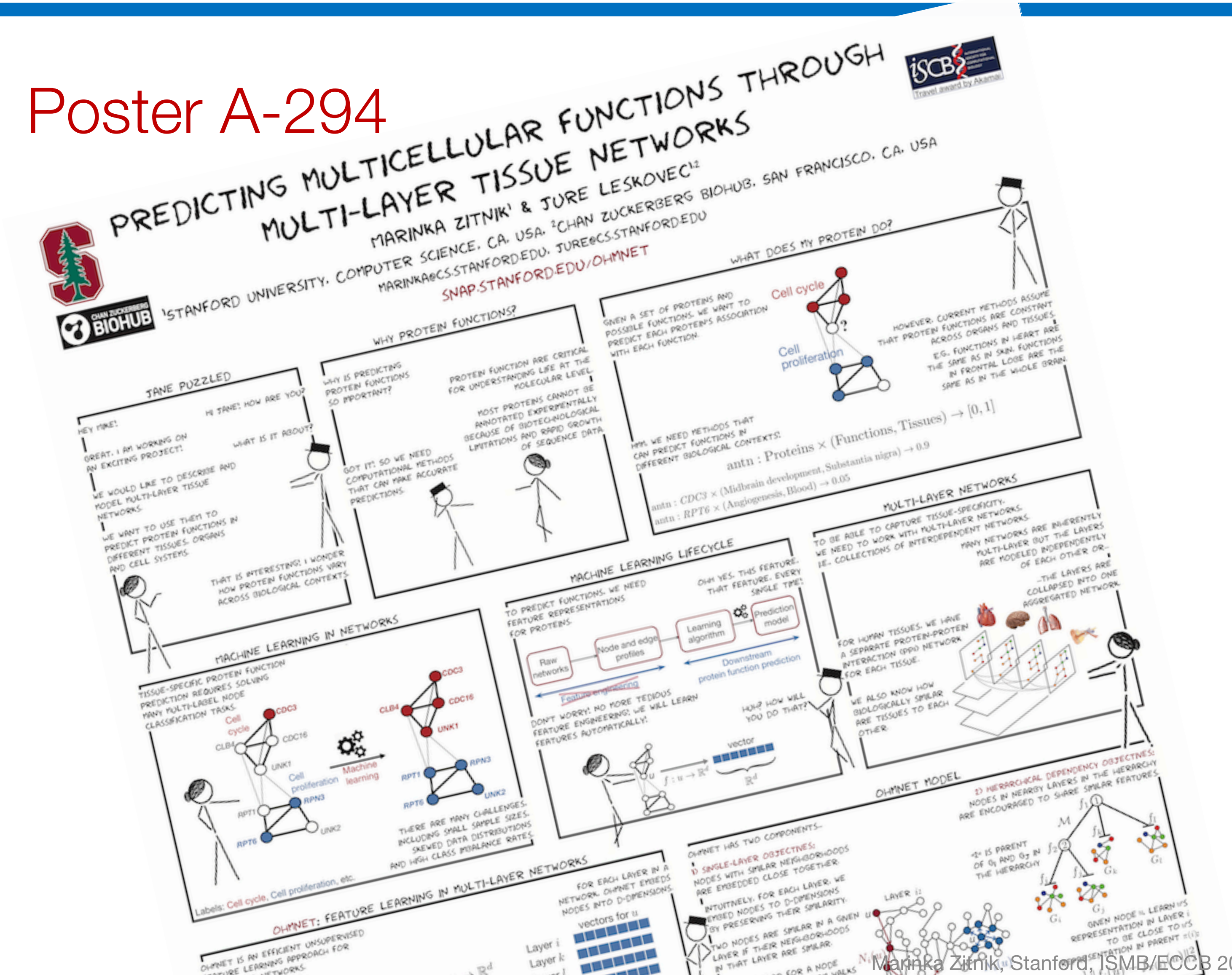
Conclusions

- Unsupervised feature learning for multi-layer networks
- Learned embeddings can be used for any downstream prediction task: node classification, node clustering, link prediction
- OhmNet predicts protein functions across biological contexts

A shift from flat networks to large multiscale systems in biology

snap.stanford.edu/ohmnet

Poster A-294



CHAN ZUCKERBERG BIOHUB

NIH

NSF

DARPA

Travel Award

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY