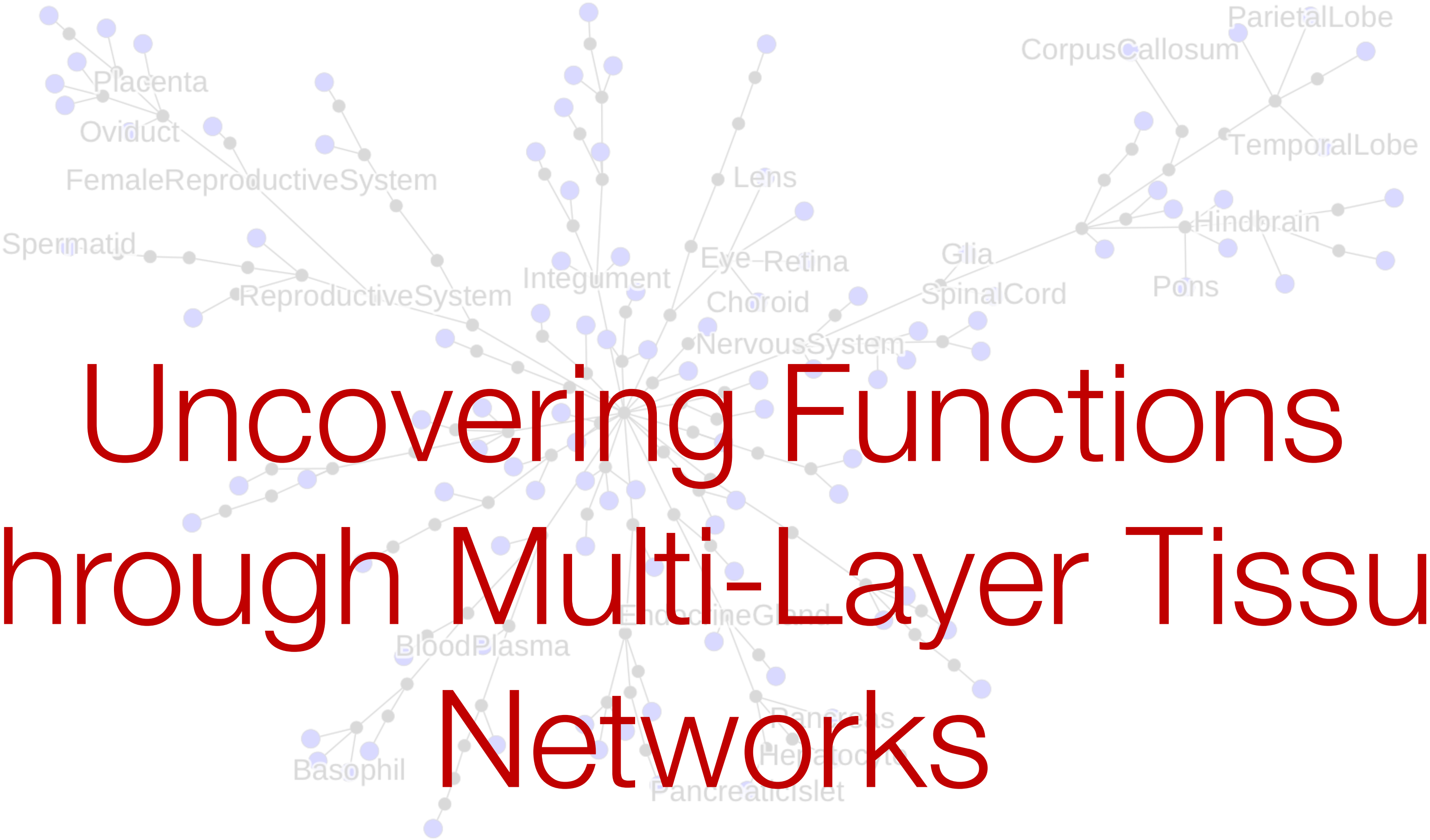


Uncovering Functions Through Multi-Layer Tissue Networks

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

Network biomedicine

Networks are a general language for describing and modeling biological systems, their structure, functions and dynamics

Why Protein Functions?

- Protein functions important for:
  - Understanding life at the molecular level
  - Biomedicine and pharmaceutical industry

- Biotechnological limits & rapid growth of sequence data: most proteins can only be annotated computationally [Clark et al. 2013, Rost et al. 2016, Greene et al. 2016]

What Does My Protein Do?

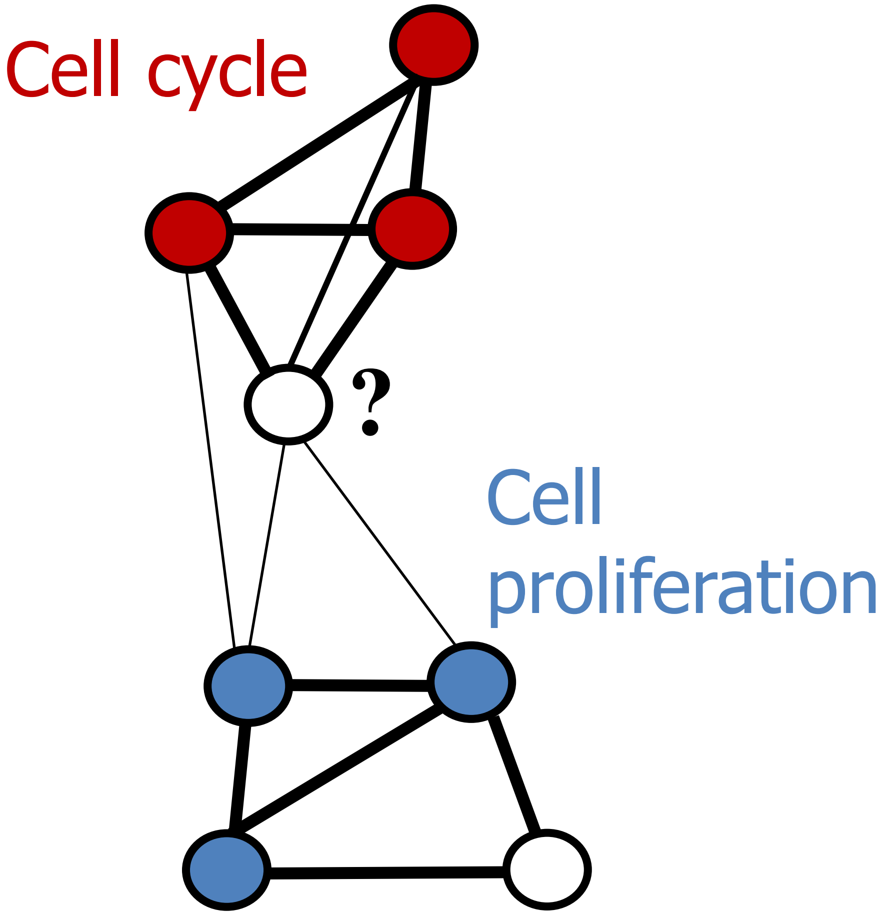
Goal: Given a set of proteins and possible functions, we want to predict each protein's association with each function:

$$\text{antn: Proteins} \times \text{Functions} \rightarrow [0,1]$$

$$\text{antn: CDC3} \times \text{Cell cycle} \rightarrow 0.9$$

$$\text{antn: RPT6} \times \text{Cell cycle} \rightarrow 0.05$$

Existing Research



"Guilty by association": protein's function is determined based on who it interacts with

- Approaches
  - Neighbor scoring
  - Indirect scoring
  - Random walks

[Zuberi et al. 2013, Radivojac et al. 2013, Kramer et al. 2014, Yu et al. 2015] and many others

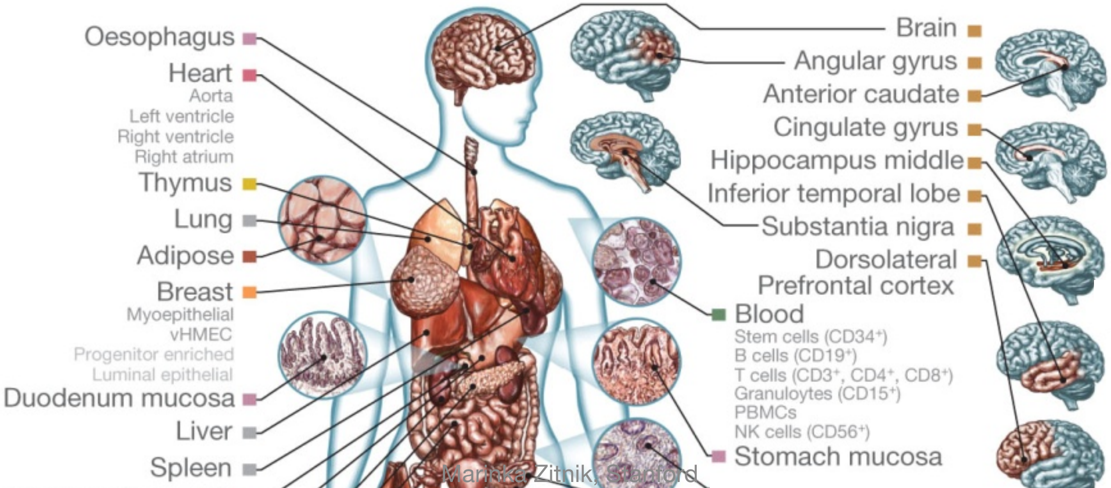
Existing Research

- Protein functions are assumed constant across organs and tissues:
  - Functions in heart are the same as in skin
  - Functions in frontal lobe are the same as in whole brain

Lack of methods to predict functions in different biological contexts

Questions for Today

1. How can we describe and model multi-layer tissue networks?
2. Can we predict protein functions in given context [e.g., tissue, organ, cell system]?
3. How functions vary across contexts?



Biotechnological Challenges

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

Computational Challenges

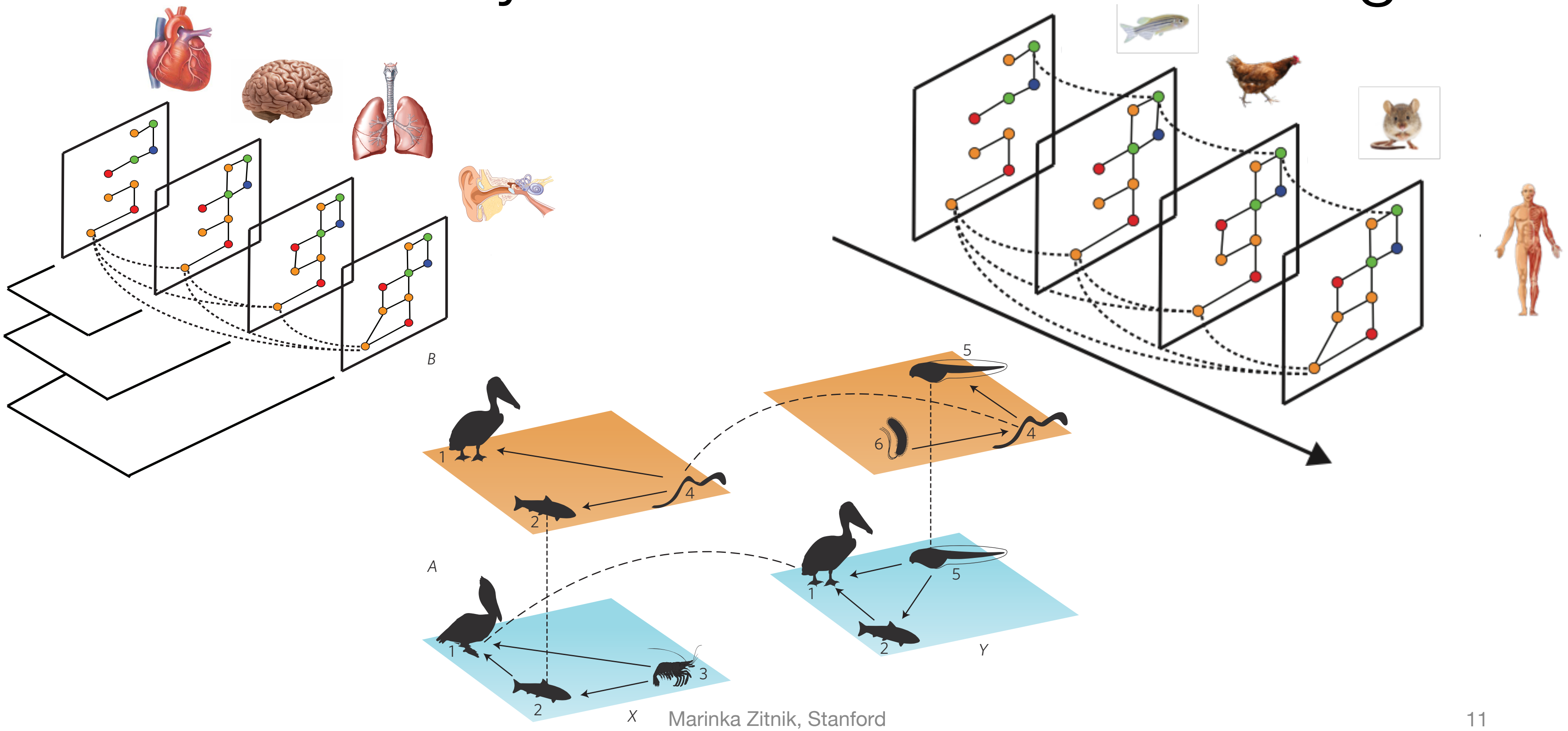
- Multi-layer network theory is only emerging at present
- Lack of formulations accounting for:
  - multiple interaction types
  - interactions vary in space, time, scale
  - interconnected networks of networks
- Nodes have different roles across layers
- Labels are extremely sparse

Part 1

The multi-layer nature of networks In biomedicine

Multi-Layer Networks

- Collections of interdependent networks
- Different layers have different meanings

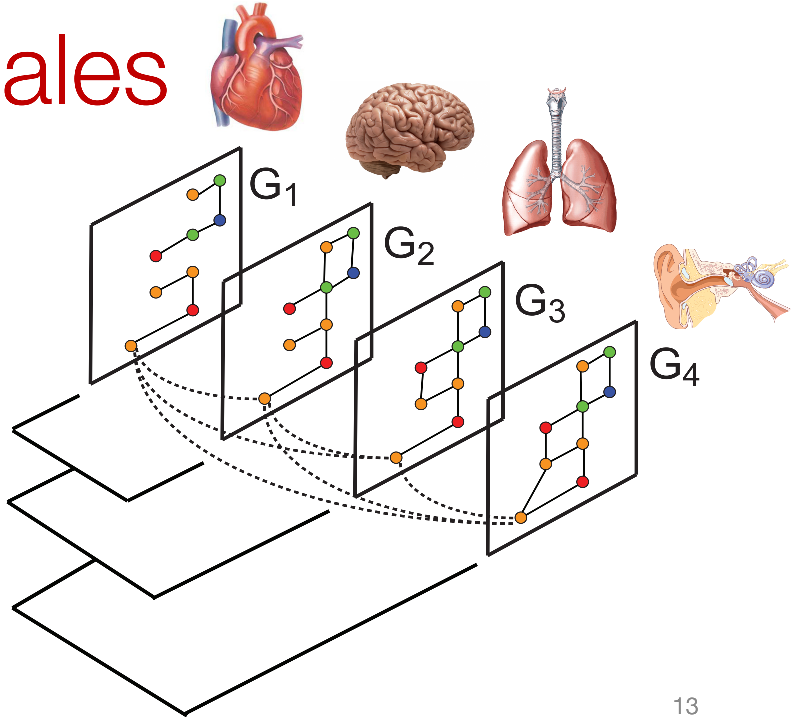


Many Network Layers

- Many networks are inherently multi-layer but the layers are:
  - Modeled independently of each other
  - Collapsed into one aggregated network
- The models must be:
  - Multi-scale: Layers at different levels of granularity
  - Scalable: Tens or hundreds of layers

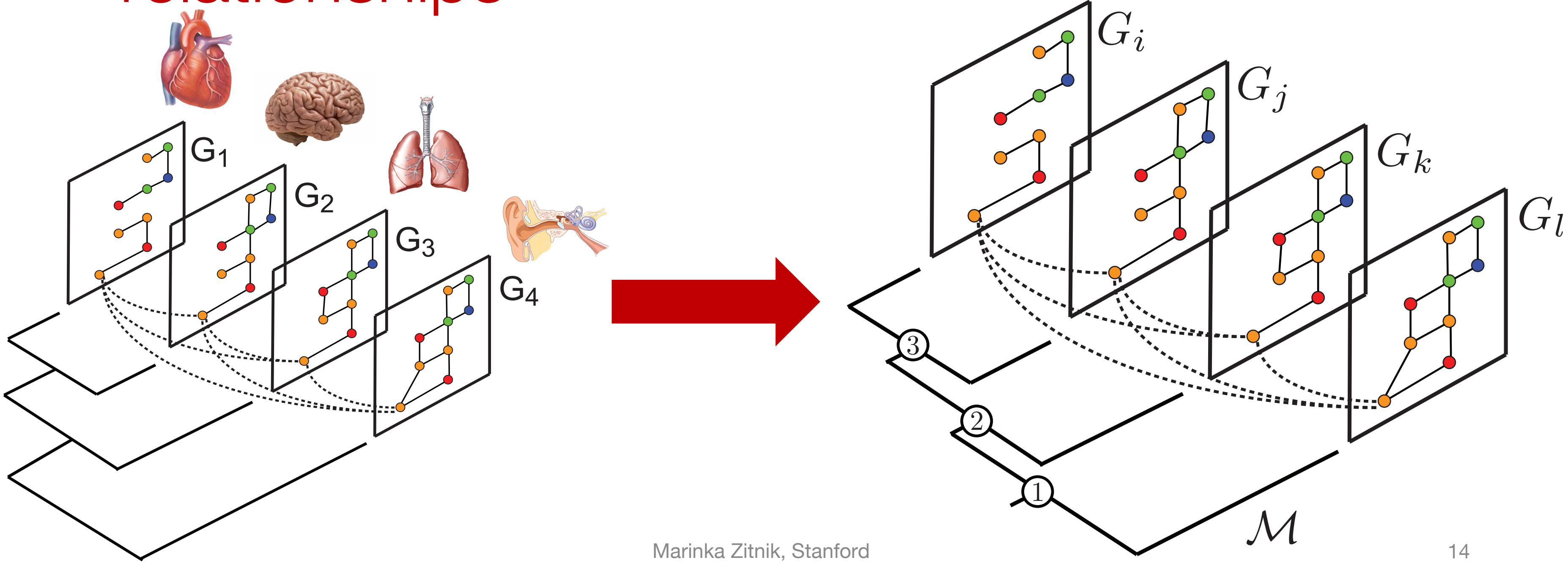
Example: Tissue Networks

- Separate protein-protein interaction network for each tissue
- Biological similarities between tissues at multiple scales



Example: Tissue Networks

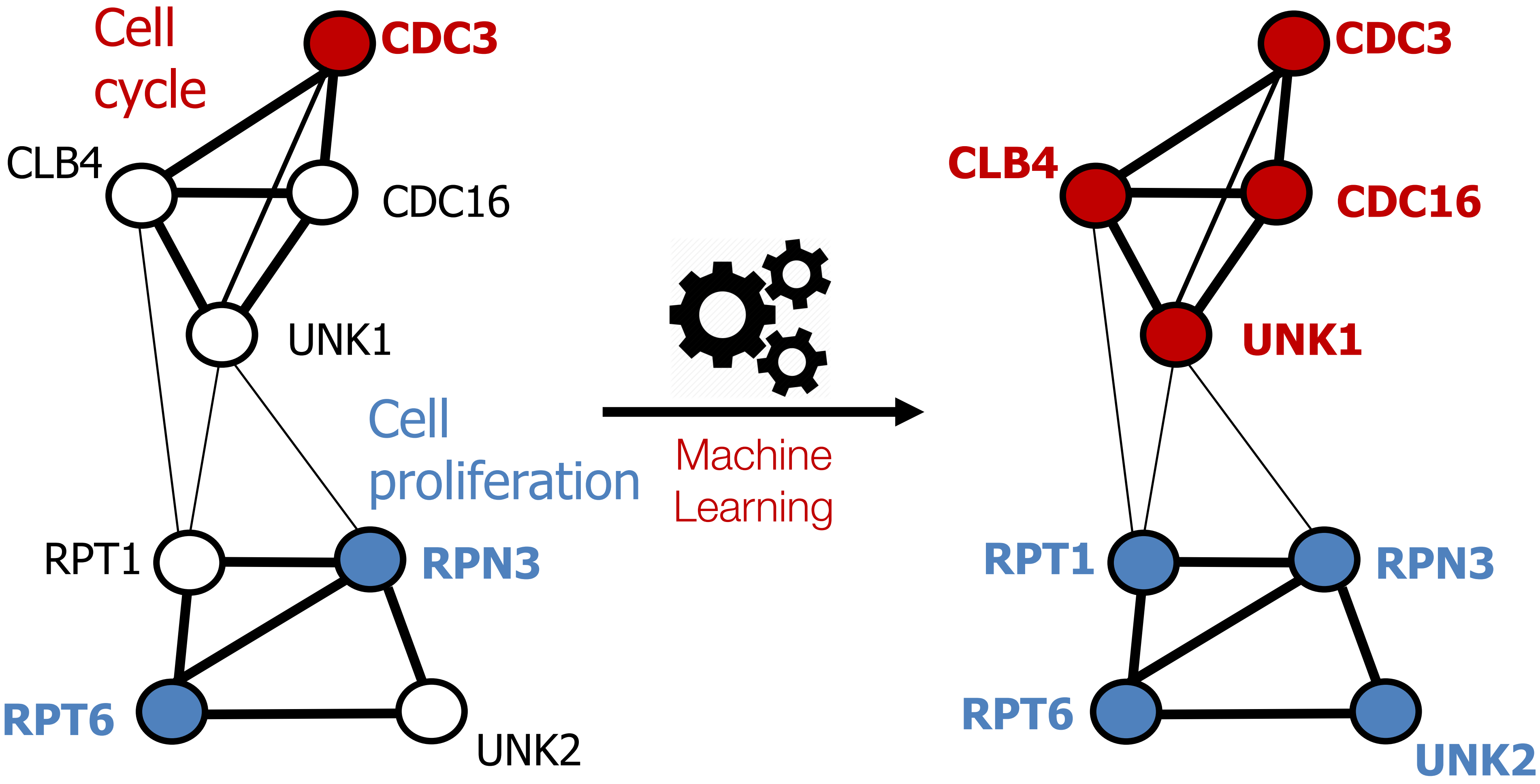
- Each PPI network is a layer $G_i = (V_i, E_i)$
- Similarities between layers are given in hierarchy $\mathcal{M}$, map π encodes parent-child relationships



Part 2

Neural embeddings for multi-layer networks

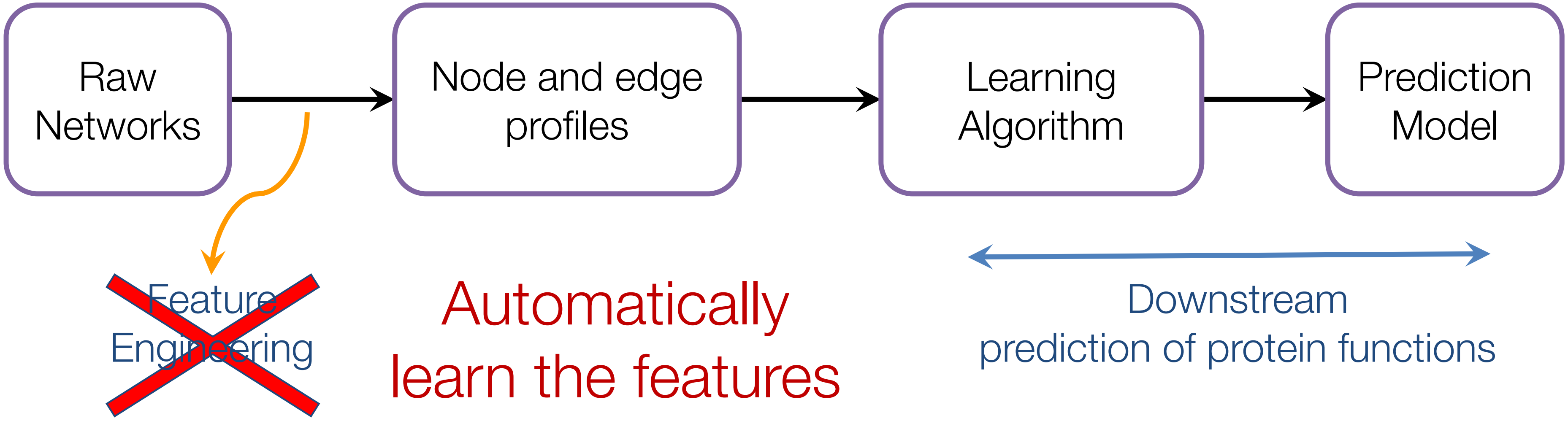
Machine Learning in Networks



Function prediction: Multi-label node classification

Machine Learning Lifecycle

- Machine Learning Lifecycle: This feature, that feature
- Every single time!



Raw Networks

Node and edge profiles

Learning Algorithm

Prediction Model

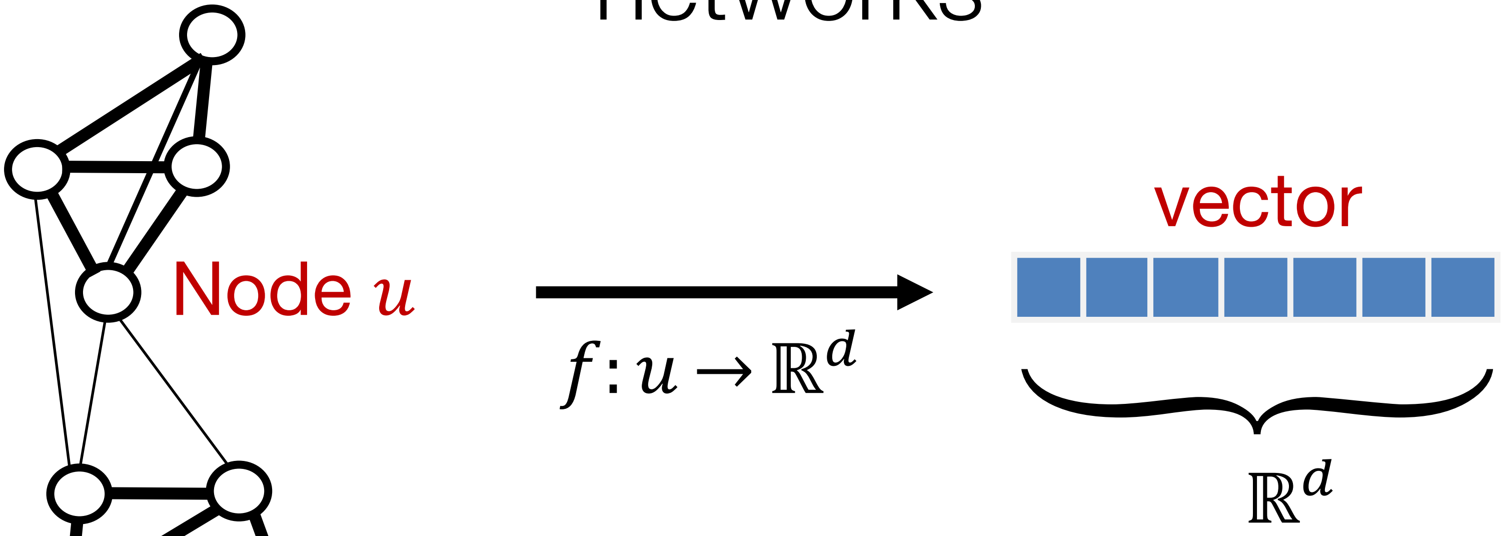
~~Feature Engineering~~

Automatically learn the features

Downstream prediction of protein functions

Feature Learning in Graphs

Efficient task-independent feature learning for machine learning in networks



Node u

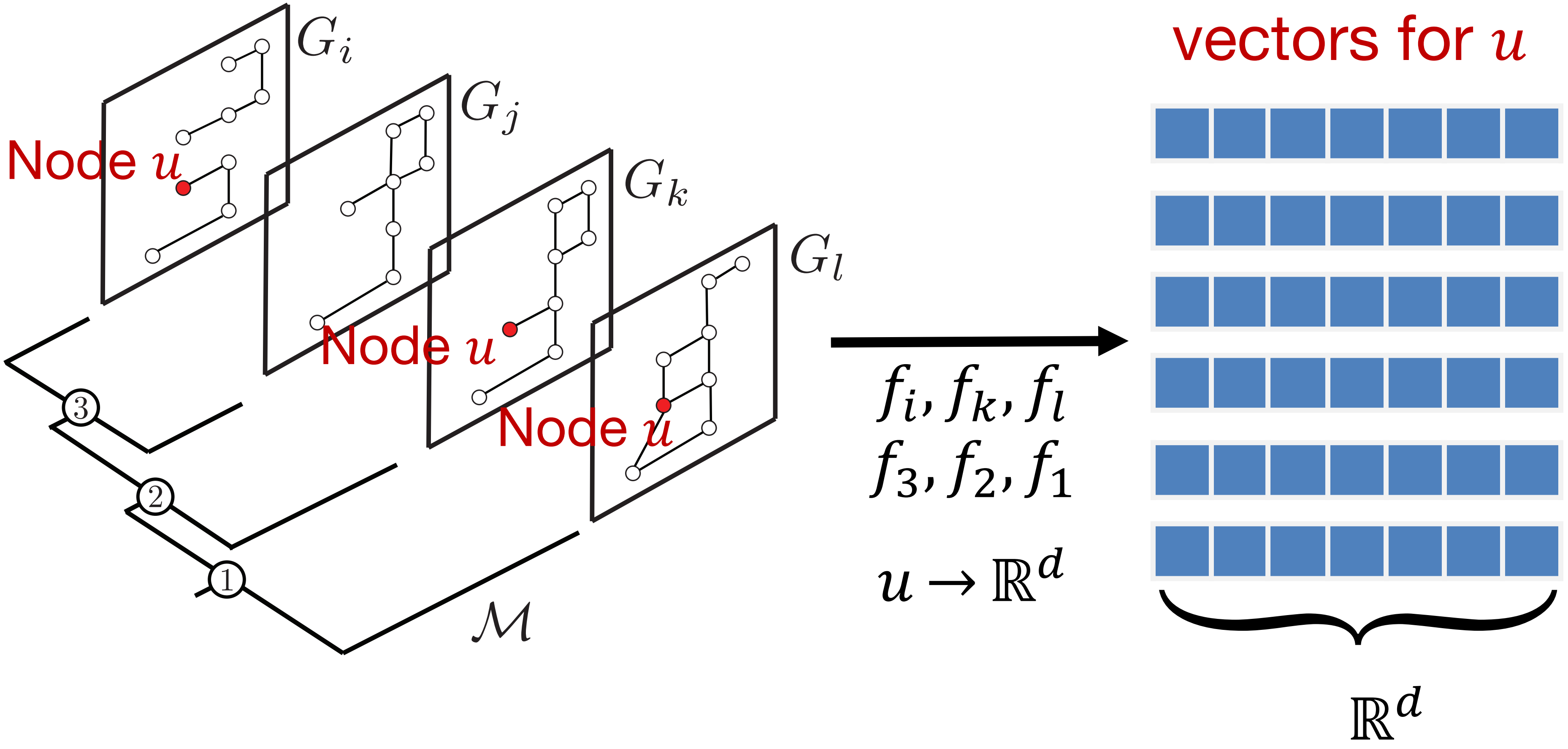
$f: u \to \mathbb{R}^d$

vector

$\mathbb{R}^d$

Feature representation, embedding

Feature Learning in Multi-Layer Nets



vectors for u

f_i, f_k, f_l

f_3, f_2, f_1

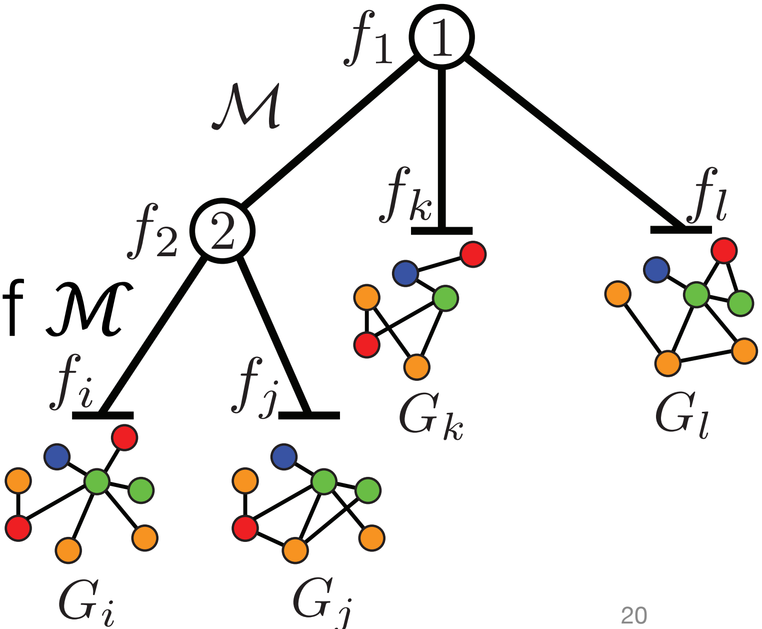
$u \to \mathbb{R}^d$

$\mathbb{R}^d$

Multi-layer, multi-scale embedding

Features in Multi-Layer Network

- Given: Layers $\{G_i\}_i$, hierarchy $\mathcal{M}$
  - Layers $\{G_i\}_{i=1..T}$ are in leaves of $\mathcal{M}$
- Goal: Learn functions: $f_i : V_i \rightarrow \mathbb{R}^d$
- Multi-scale model:
  - f_i are in leaves of $\mathcal{M}$
  - f_I are internal elements of $\mathcal{M}$



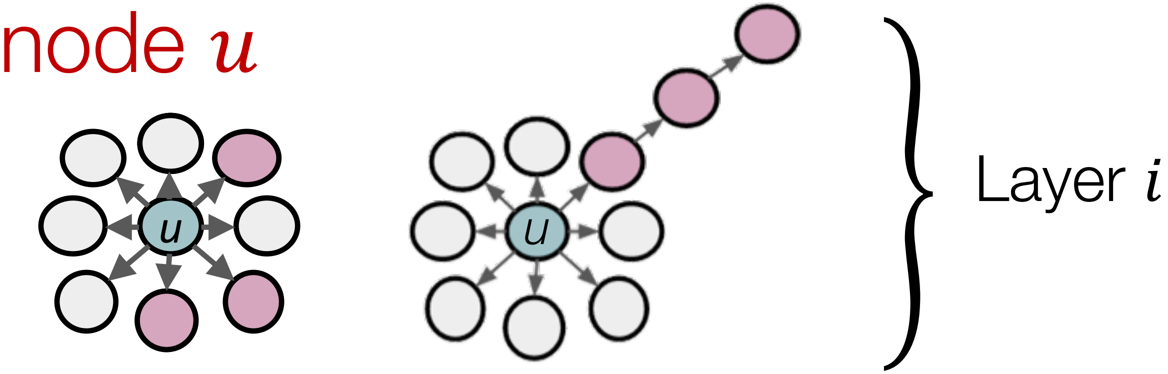
Features in Multi-Layer Network

- Approach has two components:

1. Single-layer objectives: nodes with similar neighborhoods in each layer are embedded close together
2. Hierarchical dependency objectives: nodes in nearby layers are encouraged to share similar features

Single-Layer Objectives

- Intuition: For each layer, embed nodes to d dimensions by preserving their similarity
- Approach: Nodes u and v are similar if their network neighborhoods are similar
- Given node u in layer i we define nearby nodes $N_i(u)$ based on random walks starting at node u



[Grover et al. 2016]

Single-Layer Objectives

- Given node u in layer i, learn u's representation such that it predicts nearby nodes $N_i(u)$:

$$\omega_i(u) = \log Pr(N_i(u)|f_i(u))$$

- Given T layers, maximize:

$$\Omega_i = \sum_{u \in V_i} \omega_i(u), \quad \text{for } i = 1, 2, \ldots, T$$

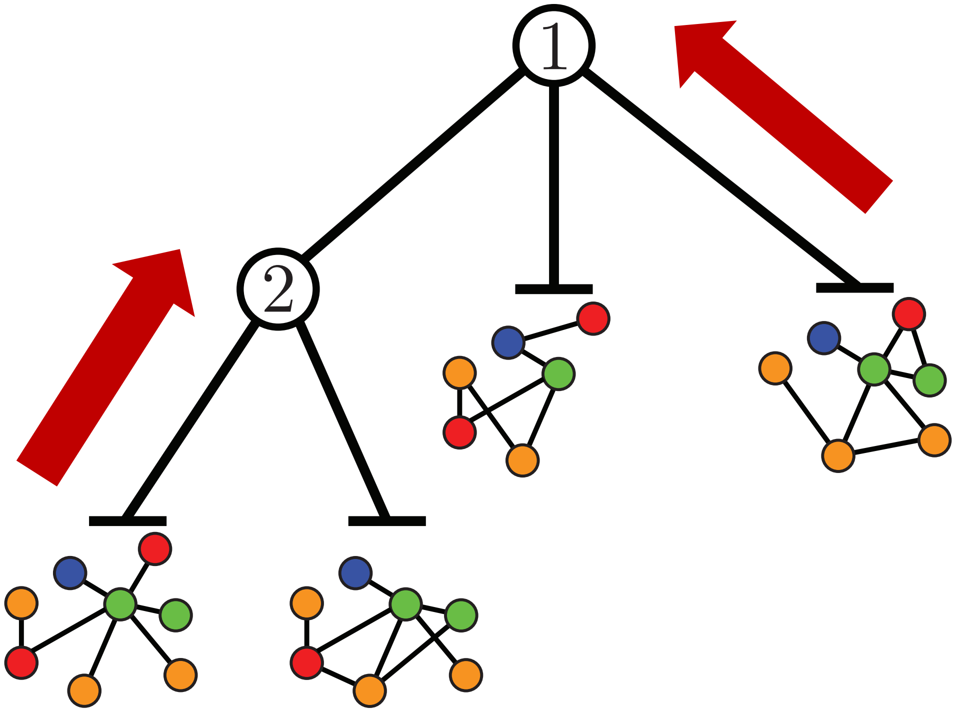
Interdependent Layers

- So far, we did not consider hierarchy $\mathcal{M}$
- Node representations in different layers are learned independently of each other

How to model dependencies between layers when learning features?

Idea: Interdependent Layers

- Encourage nodes in layers nearby in the hierarchy to be embedded close together



Relationships Between Layers

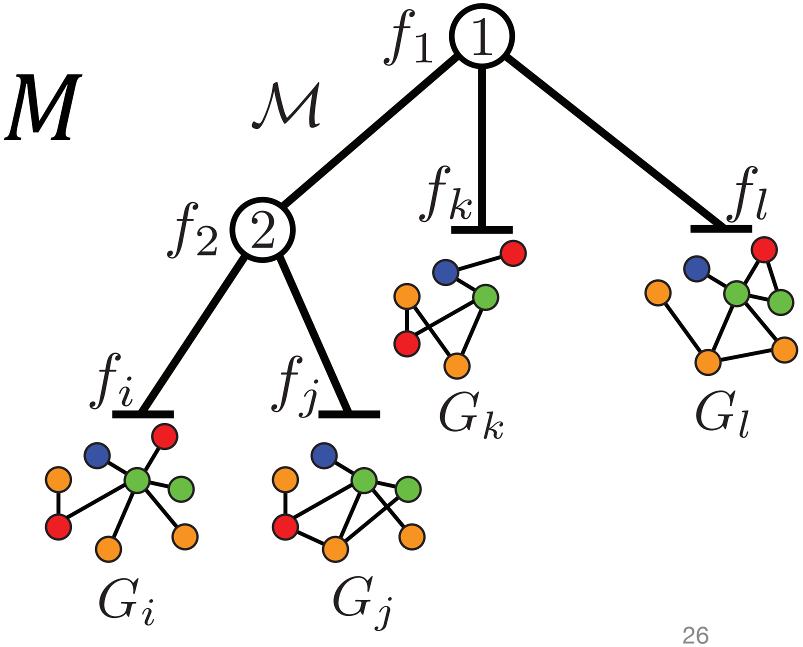
- Hierarchy M is a tree, given by the parent-child relationships:

$$\pi : M \rightarrow M$$

- $\pi(i)$ is parent of i in M

Example:

"2" is parent of G_i, G_j



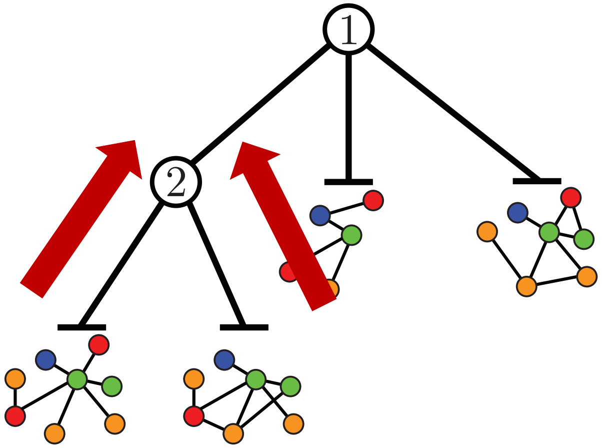
Interdependent Layers

- Given node u, learn u's representation in layer i to be close to u's representation in parent $\pi(i)$:

$$c_i(u) = \frac{1}{2}\|f_i(u) - f_{\pi(i)}(u)\|_2^2$$

- Multi-scale: Repeat at every level of $\mathcal{M}$

$$C_i = \sum_{u \in L_i} c_i(u)$$



L_i has all layers appearing in sub-hierarchy rooted at i

Final Model: *OhmNet*

Automatic feature learning in multi-layer networks

Solve maximum likelihood problem:

$$\max_{f_1,f_2,\ldots,f_{|M|}} \underbrace{\boxed{\sum_{i\in\mathcal{T}} \Omega_i}}_{\text{Single-layer objectives}} - \lambda \underbrace{\boxed{\sum_{j\in\mathcal{M}} C_j}}_{\text{Hierarchical dependency objectives}} ,$$

OhmNet Algorithm

1. For each layer, compute random walk probs.
2. For each layer, sample fixed-length random walks starting from each node u
3. Optimize the OhmNet objective using stochastic gradient descent

Scalable: No pairwise comparison of nodes from different layers

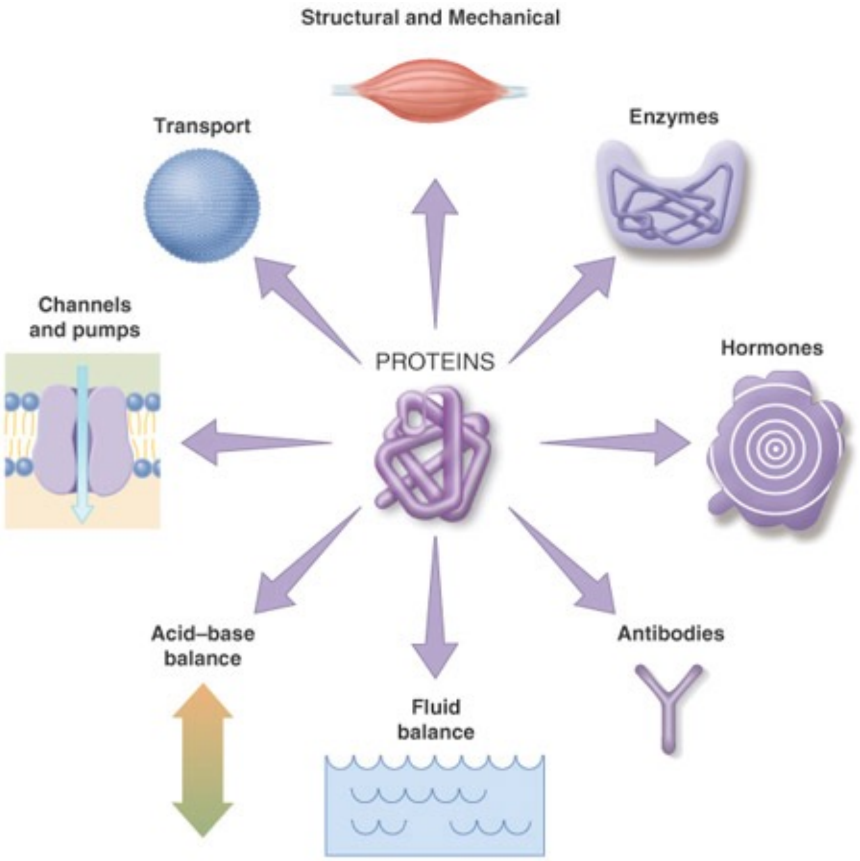
Part 3

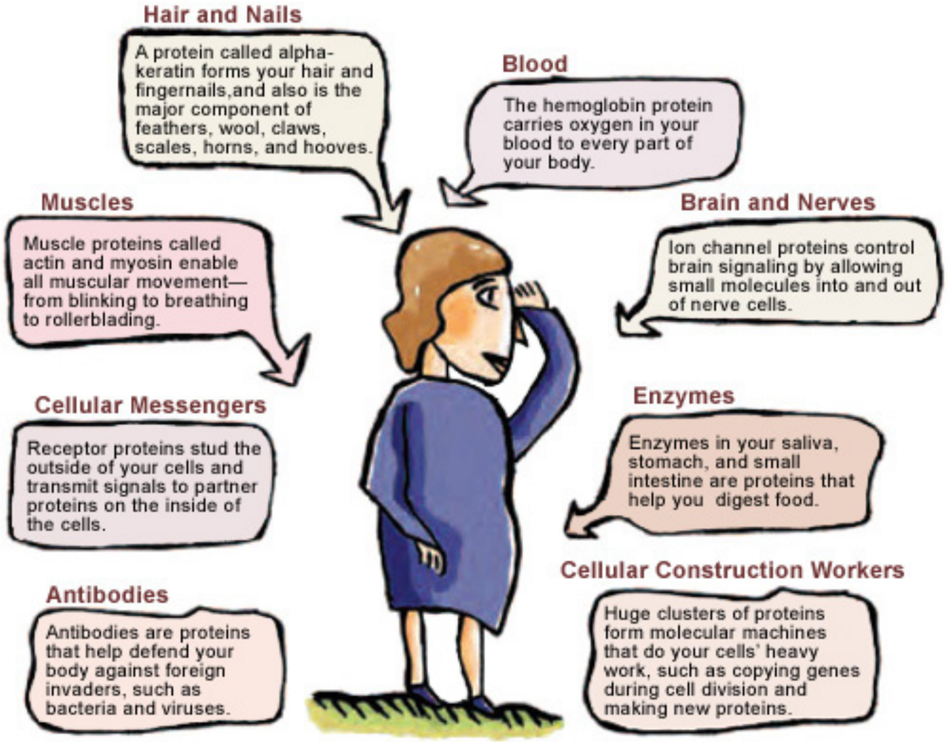
Results: Protein function prediction across tissues

Tissue-Specific Function Prediction

1. Learn features of every node and at every scale based on:
   - Edges within each layer
   - Inter-layer relationships between nodes active on different layers
2. Predict tissue-specific protein functions using the learned node features

Protein Functions and Tissues




Hair and Nails

A protein called alpha-keratin forms your hair and fingernails,and also is the major component of feathers, wool, claws, scales, horns, and hooves.

Blood

The hemoglobin protein carries oxygen in your blood to every part of your body.

Muscles

Muscle proteins called actin and myosin enable all muscular movement—from blinking to breathing to rollerblading.

Brain and Nerves

Ion channel proteins control brain signaling by allowing small molecules into and out of nerve cells.

Cellular Messengers

Receptor proteins stud the outside of your cells and transmit signals to partner proteins on the inside of the cells.

Enzymes

Enzymes in your saliva, stomach, and small intestine are proteins that help you digest food.

Antibodies

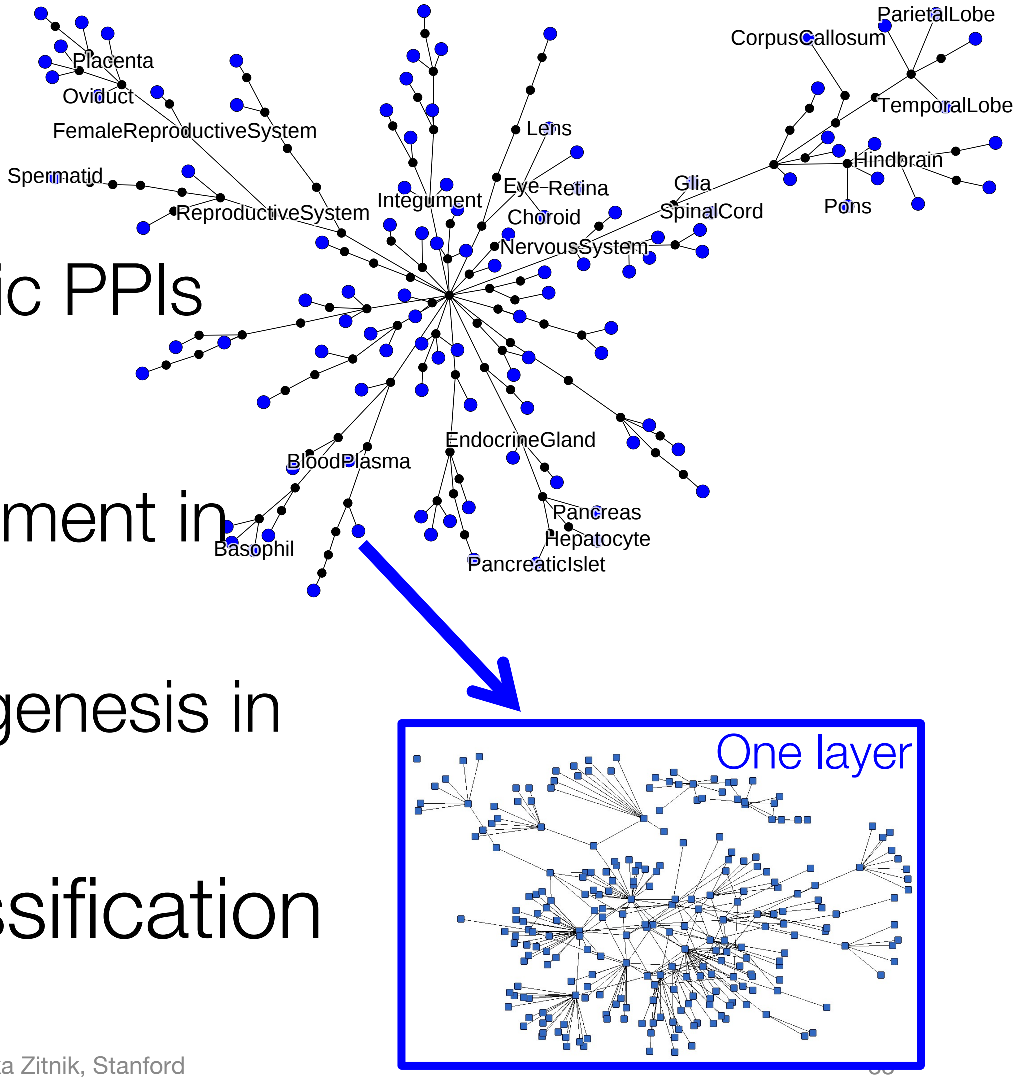
Antibodies are proteins that help defend your body against foreign invaders, such as bacteria and viruses.

Cellular Construction Workers

Huge clusters of proteins form molecular machines that do your cells' heavy work, such as copying genes during cell division and making new proteins.

Data: 107 Tissue Layers

- Layers are PPI nets:
  - Nodes: proteins
  - Edges: tissue-specific PPIs
- Node labels:
  - E.g., Cortex development in renal cortex tissue
  - E.g., Artery morphogenesis in artery tissue
- Multi-label node classification

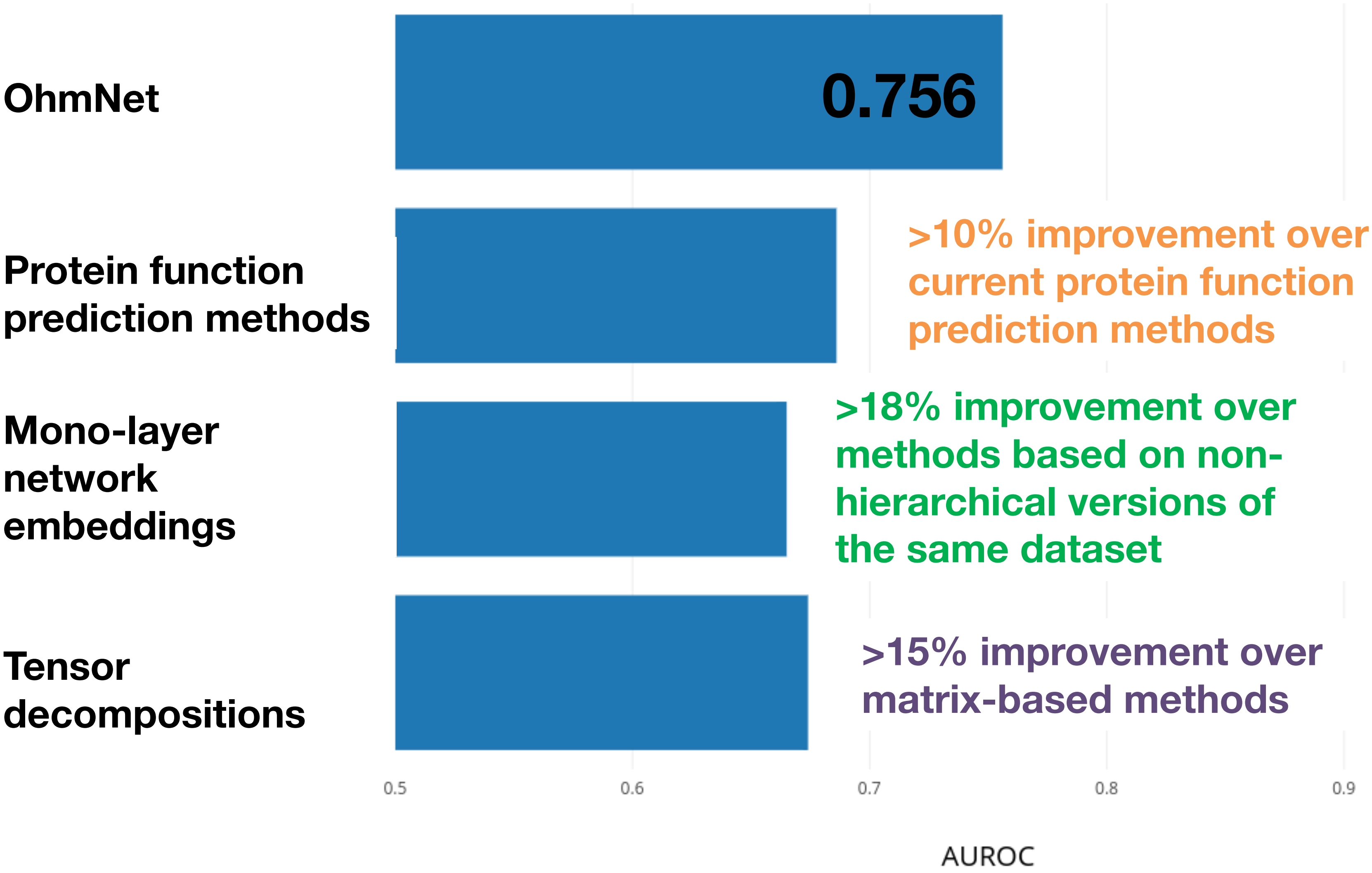


Experimental Setup

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

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

Protein Function Prediction



>10% improvement over current protein function prediction methods

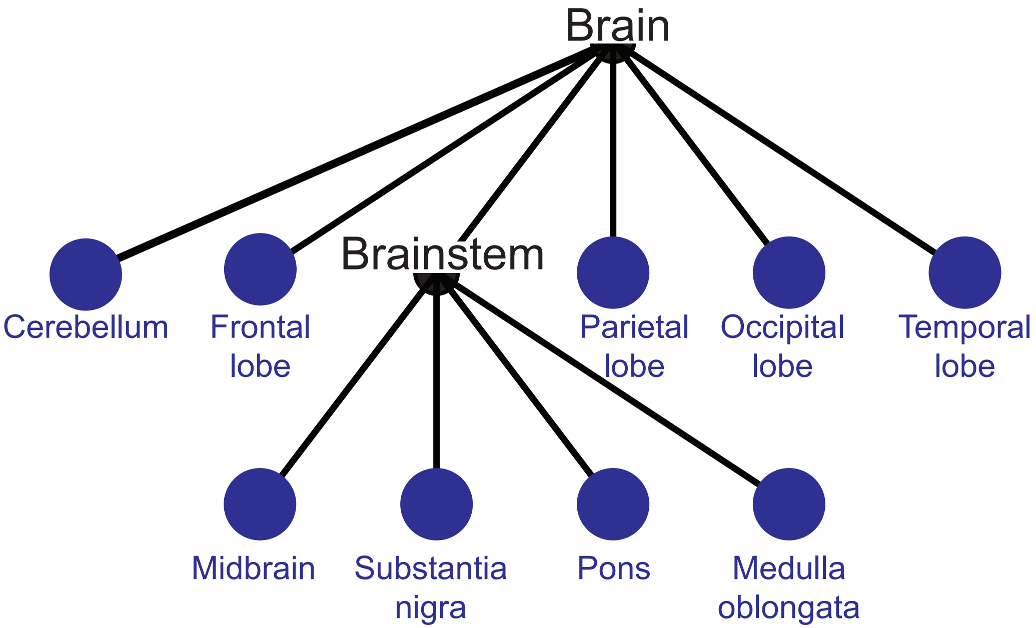
>18% improvement over methods based on non-hierarchical versions of the same dataset

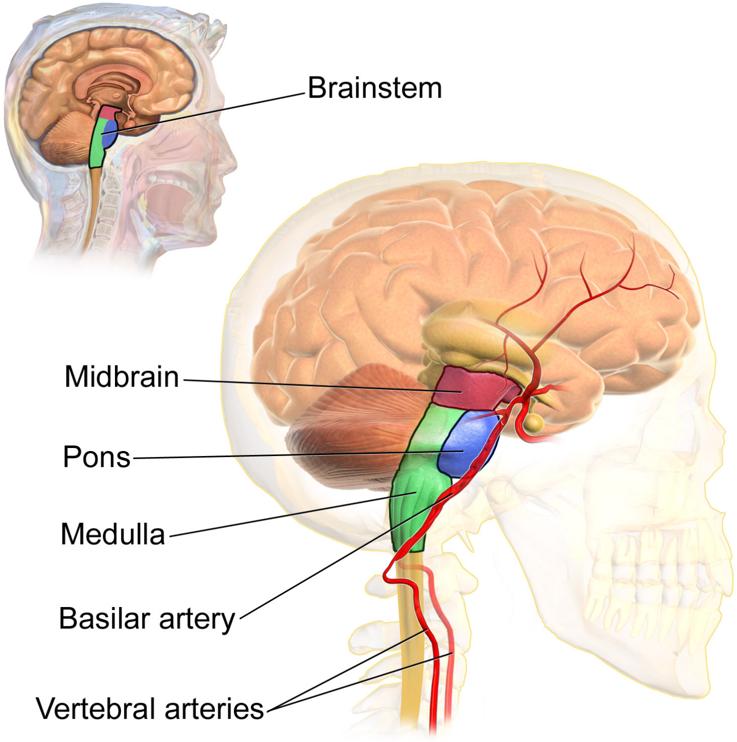
>15% improvement over matrix-based methods

Part 4

Results: Other applications

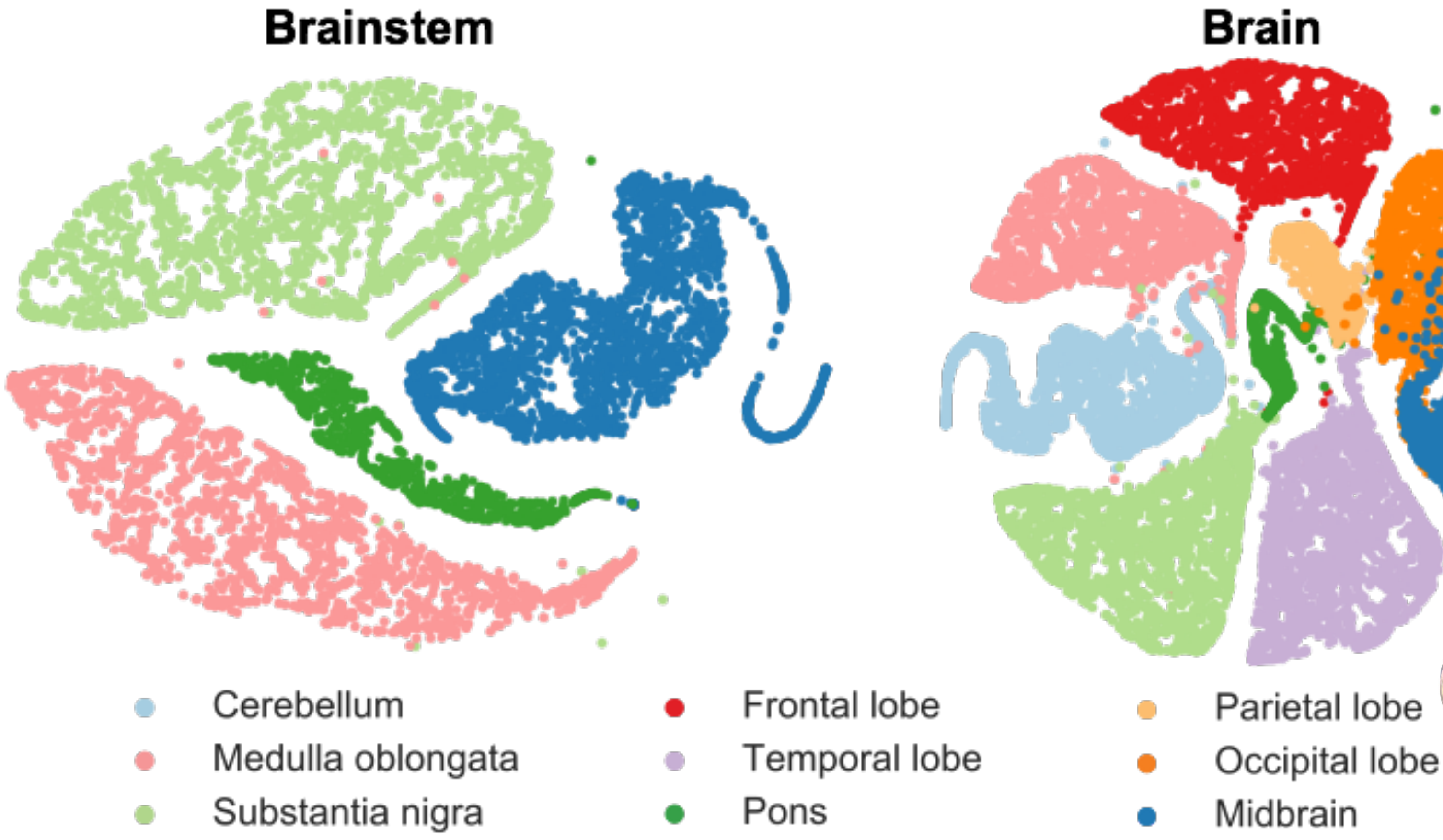
Brain Tissues

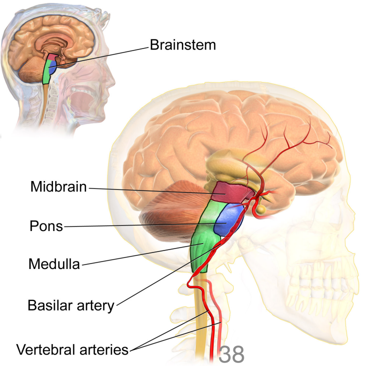




9 brain tissue PPI networks in two-level hierarchy

Meaningful Node Embeddings





Unannotated Tissues

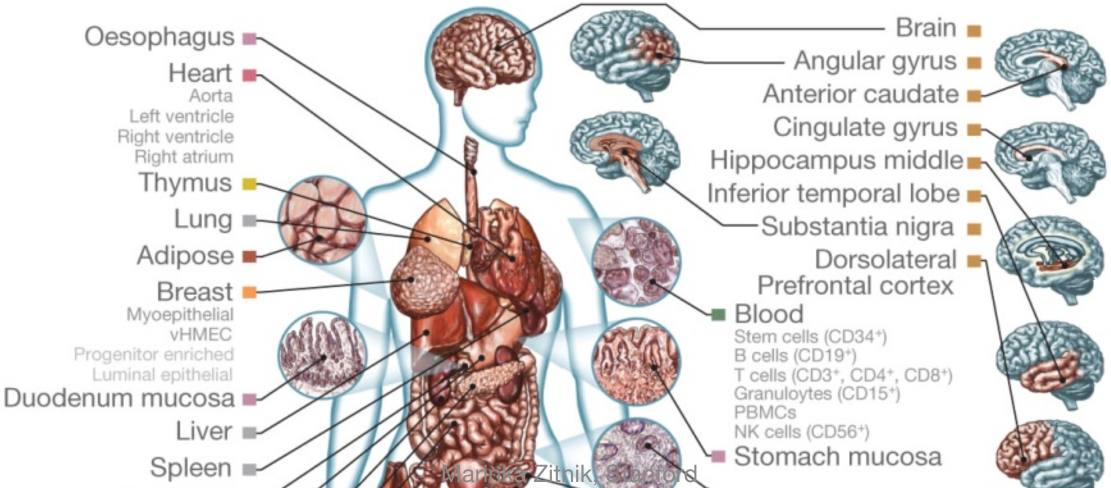
- Transfer functions to unannotated tissues
- Task: Predict functions in target tissue without access to any annotation/label in that tissue

Target tissue	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 AUC values

Revisit: Questions for Today

1. How can we describe and model multi-layer tissue networks?
2. Can we predict protein functions in given context [e.g., tissue, organ, cell system]?
3. How functions vary across contexts?



Conclusions

- Unsupervised feature learning in multi-layer networks
- Learned features can be used for any downstream prediction task: node classification, node clustering, link prediction
- Move from flat networks to large multiscale systems in biology

Thank you!

snap.stanford.edu/ohmnet

Predicting multicellular function through multi-layer tissue networks. M. Zitnik, J. Leskovec. *Bioinformatics* 2017.

To appear at ISMB/ECCB 2017

DARPA

NIH

CHAN ZUCKERBERG BIOHUB