ETH zürich

BIOMEDICAL INFORMATICS

Applications of Deep Learning on Graphs

Prof. Dr. Gunnar Rätsch, Dr. Rita Kuznetsova
Institute for Machine Learning, Department of Computer Science

Teaching Staff: Biomedical Informatics Lab

Prof. Dr. Gunnar Rätsch, Lecturer

Dr. Rita Kuznetsova, Lecturer

Manuel Burger, Head TA

Fedor Sergeev, TA

(plus invited speakers for specific topics)

Course Goals

Acquire a comprehensive understanding of the fundamental concepts and building blocks of Graph Neural Networks.

Explore and dissect seminal and recent publications and breakthroughs in the domain of Graph Neural Networks.

Develop a practical understanding of Graph Neural Networks by embarking on a capstone project that demonstrates their application.

Outline for Today

1st slot (45 min):

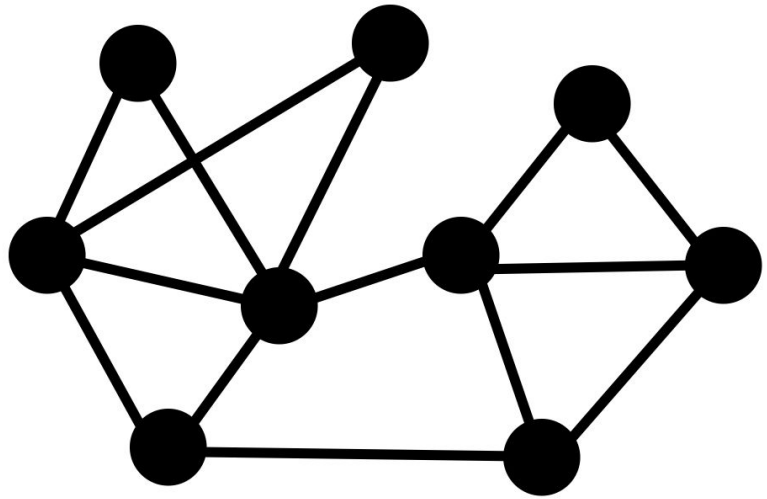
1. Goals & Motivation
2. Course Logistic
3. Dive into Graph ML Tasks & Applications: BMI Lab Expertise
    a. Learning Representations of Genome Graphs
    b. Learning Representations of Medical Concepts
    c. Knowledge Graph Representations to enhance Time-Series Predictions

2nd slot (45 min):

TA presentation about projects, tutorials, paper presentations

Why Graphs?

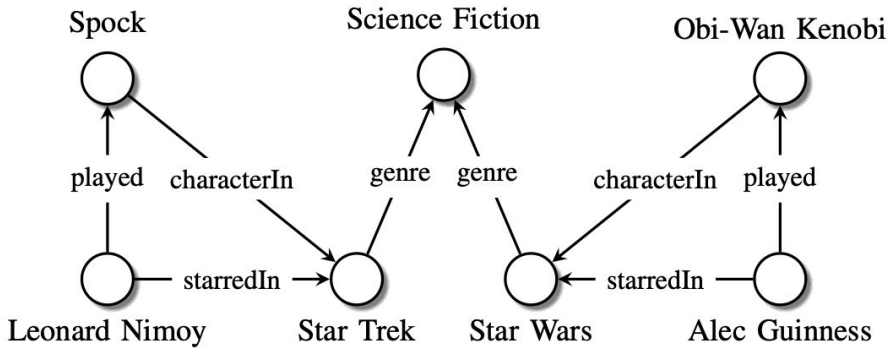
- **Graphs** are a general framework for describing and analyzing entities with relations/interactions.
- The power of the **graph formalism** lies both in its:
  - focus on relationships between entities (rather than individual entity's properties);
  - Flexibility and broad applicability of graph structures.
- Many **types of data** are represented as graphs.



Credit: [1, 4]

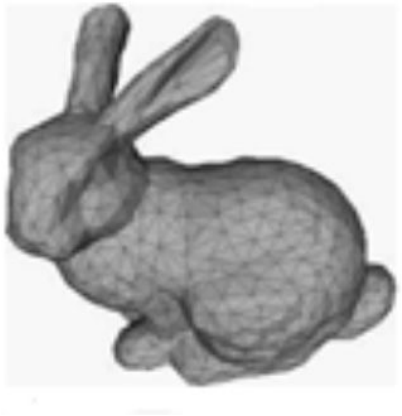
Many types of data are graphs

Social Network



Knowledge Graph

Citation Network



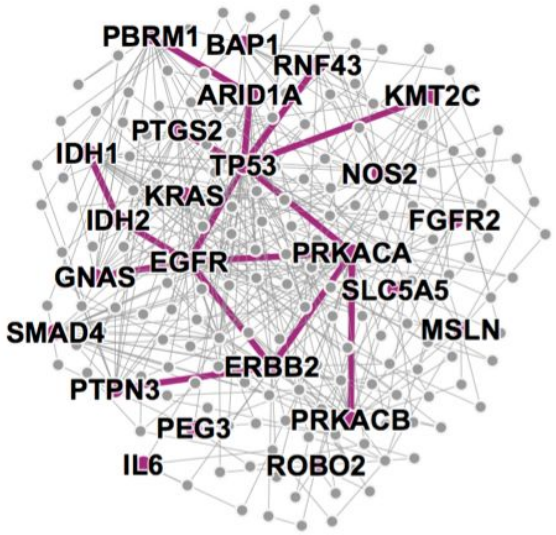
3D shape

Underground Network

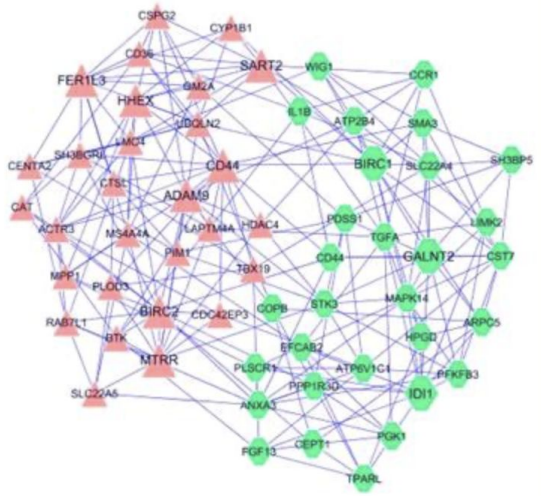
Computer Network

Credit: [1, 9, 10, 11, 12, 13]

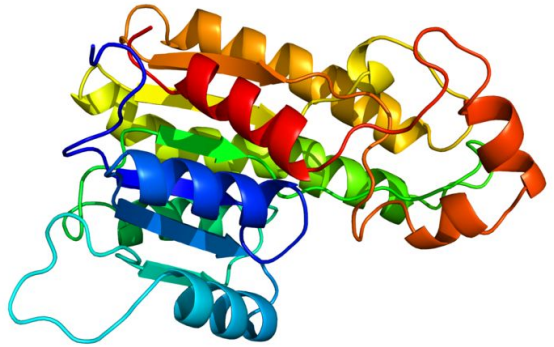
Many types of data are graphs



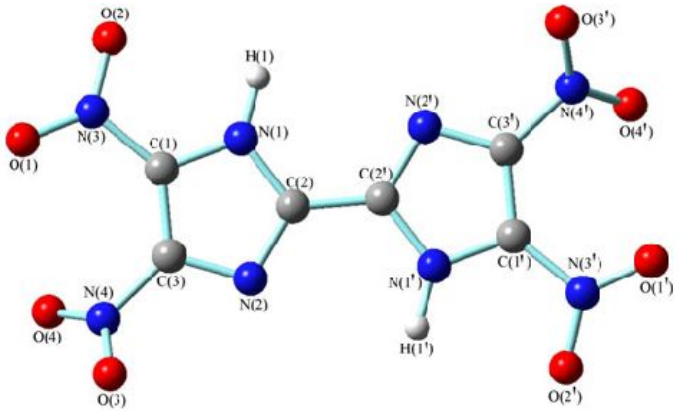
Disease Pathways



Gene Regulatory Network



Protein 3d structure



Molecule graph

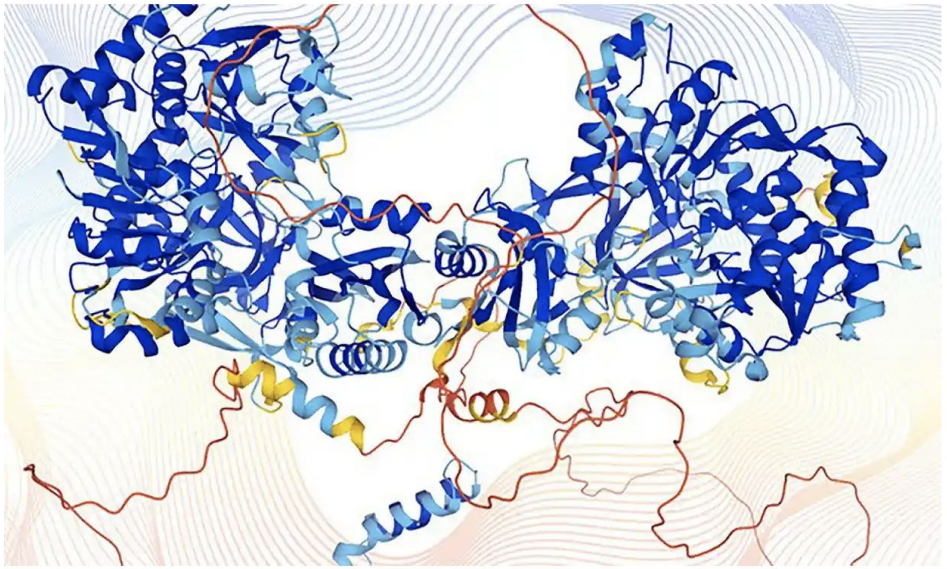
Credit: [1, 5, 6, 7, 8]

The Guardian

DeepMind AI cracks 50-year-old problem of protein folding

DeepMind uncovers structure of 200m proteins in scientific leap forward

Success of AlphaFold program could have huge impact on global problems such as famine and disease



The Economist

Menu | Weekly edition | The world in brief | Search

Podcasts | Babbage

How artificial intelligence cracked biology's biggest problem

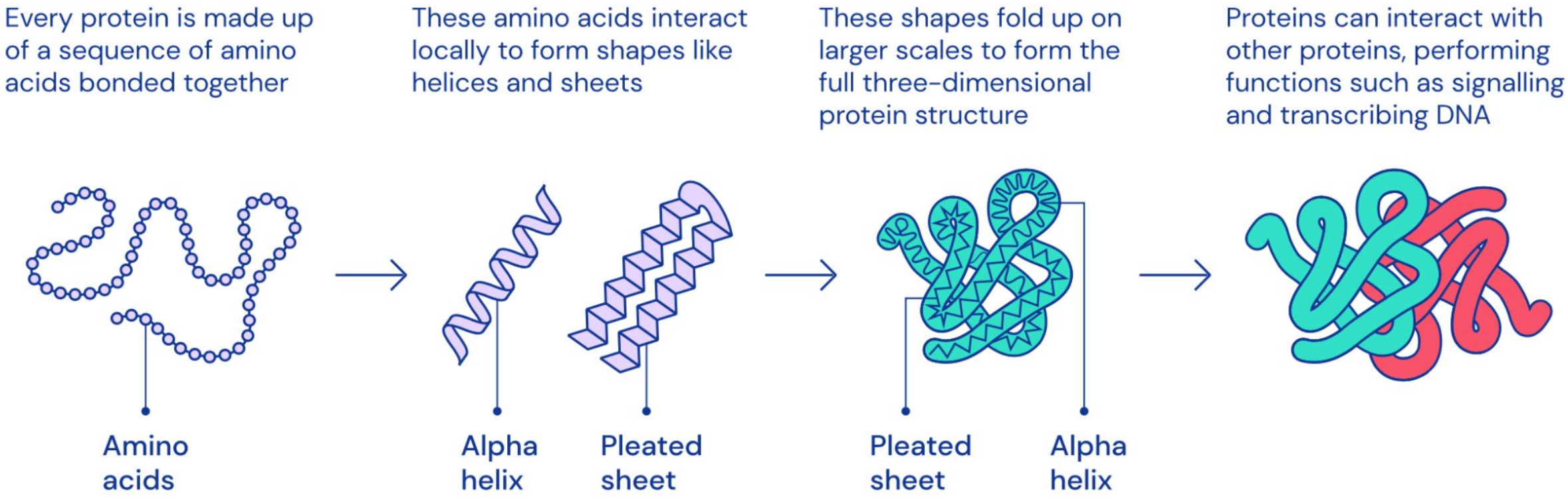
The New York Times

A.I. Predicts the Shape of Nearly Every Protein Known to Science

DeepMind has expanded its database of microscopic biological mechanisms, hoping to accelerate research into all living things.

AlphaFold

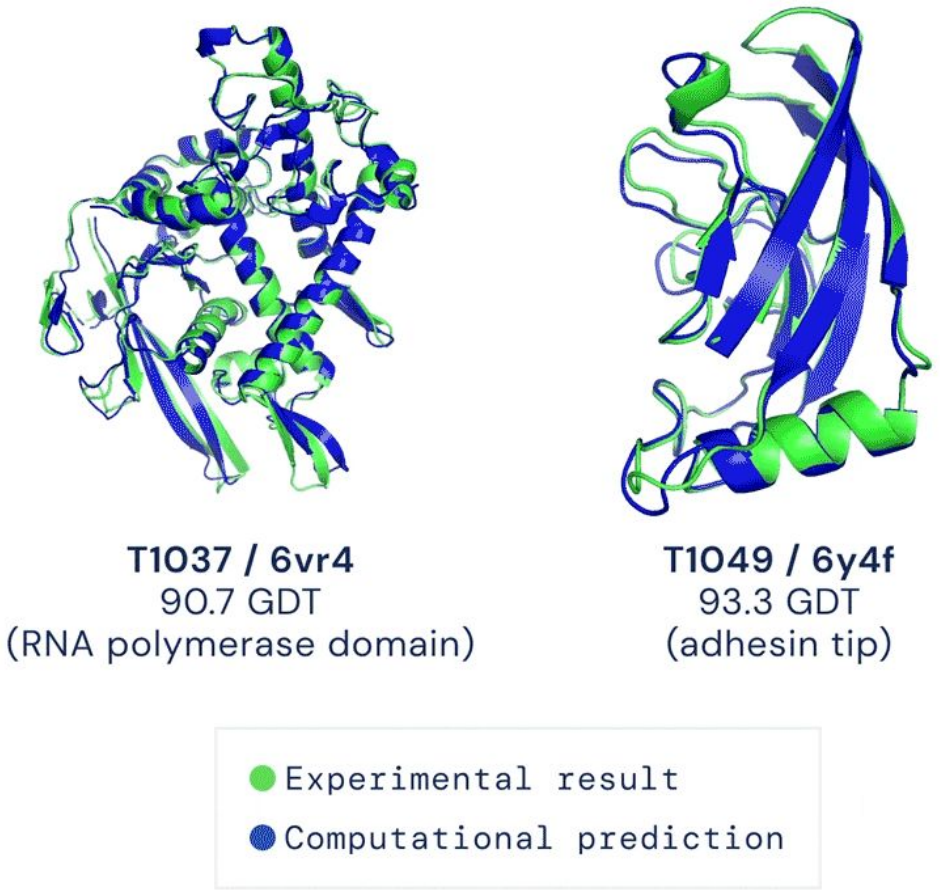
Proteins are large, complex molecules essential to all of life. Nearly every function that our body performs relies on proteins, and how they **move and change**. What any given protein can do depends on its **unique 3D structure.**



Credit: [2]

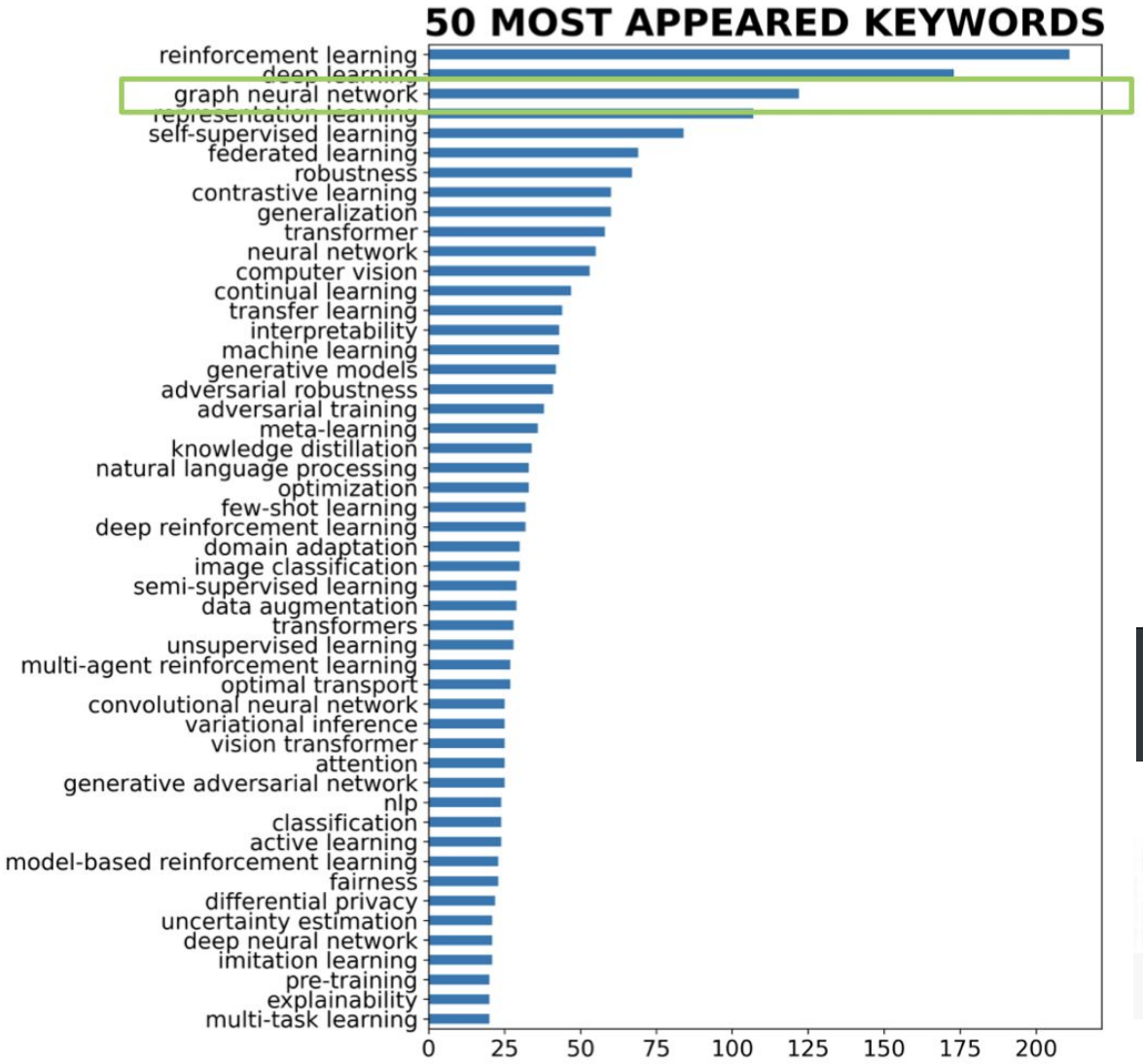
AlphaFold

The ‘protein-folding problem’ - predict a protein’s 3D structure based solely on its 1D amino acid sequence.



Why Graph Neural Networks?

ICLR 2022 keywords



LOG CONFERENCE

NEW FRONTIERS IN GRAPH LEARNING (GLFRONTIERS)

NeurIPS Workshop 2023

TGL TGLworkshop 2023

ICLR

Deep Learning on Graphs for Natural Language Processing

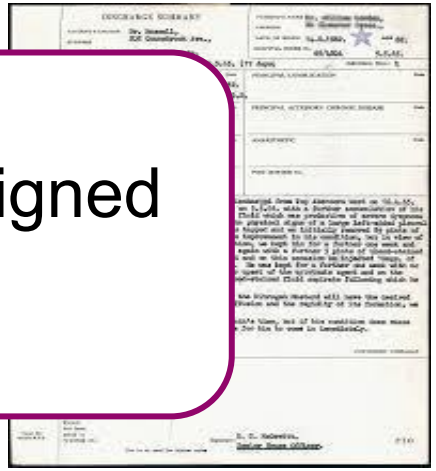
Credit: [1]

How to take an advantage of relational structure?

Complex domains have a rich relational structure, which can be represented as a **relational graph.**

By explicitly modeling relationships we achieve better performance!

Modern DL architectures are typically designed for sequences / grids!



Images

Audio signals

Text

Credit: [1]

Why Graph Neural Networks?

Not everything can be represented as a sequence or a grid.

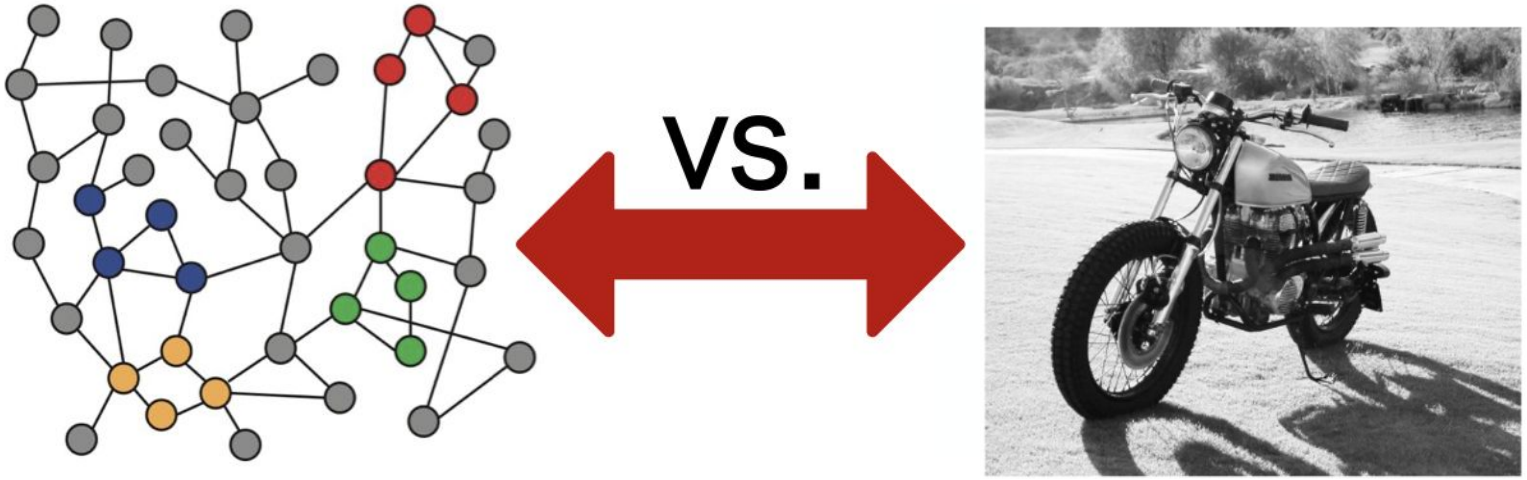
We can try, but …

SMILES (Simplified molecular-input line-entry system)

Structure is lost!

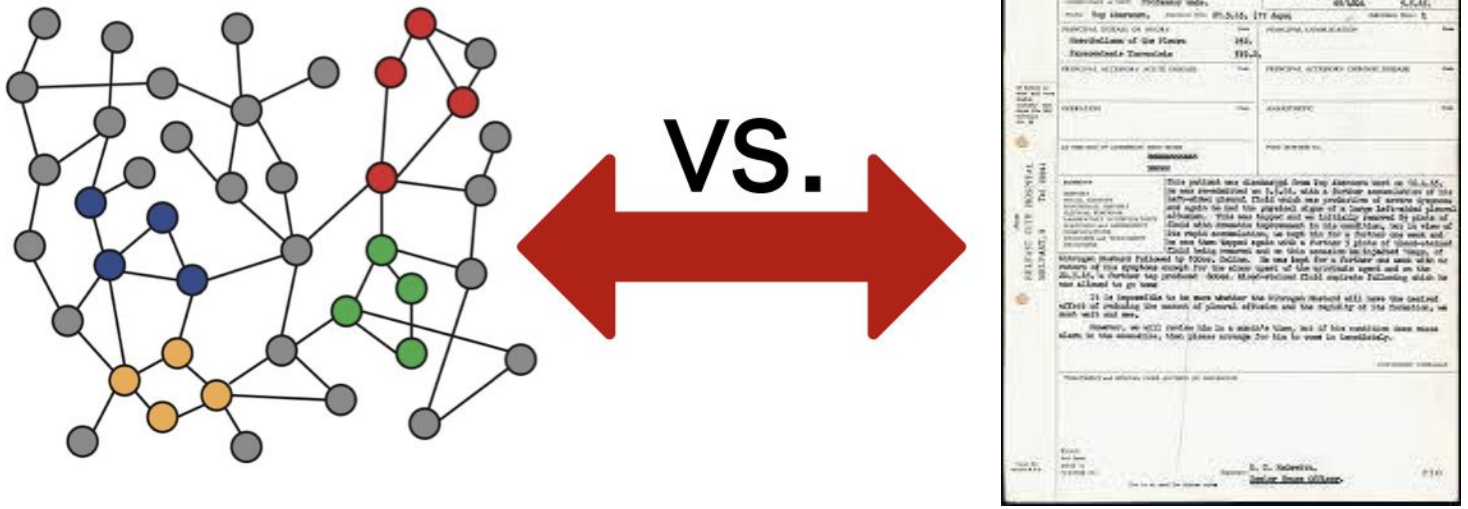
We need to develop methodology that is more broadly applicable and could keep the structural information.

Why it is hard to process Graphs/Network structure



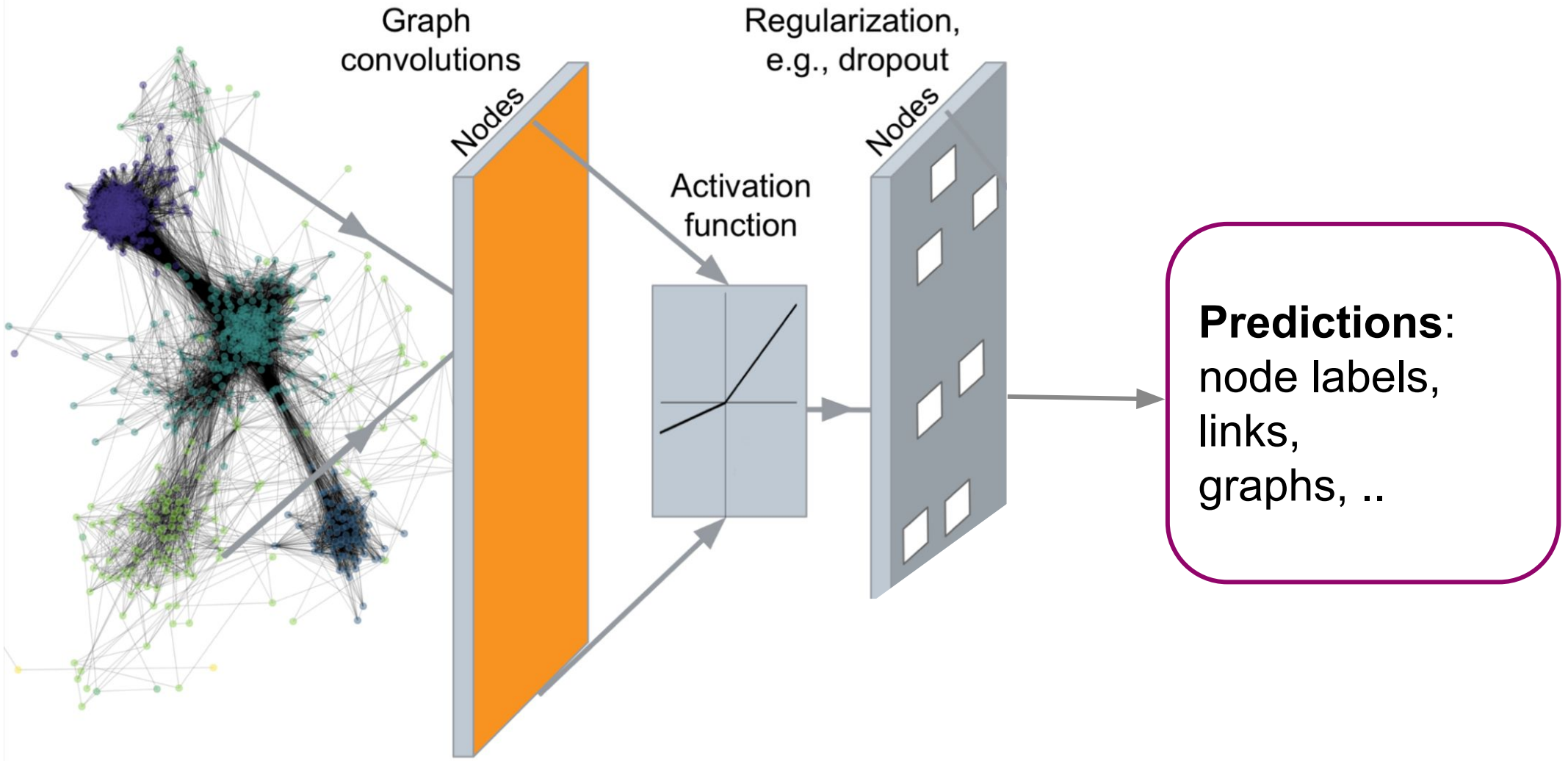
Arbitrary size and complex topological structure (*i.e.*, no spatial locality like grids)

Why it is hard to process Graphs/Network structure



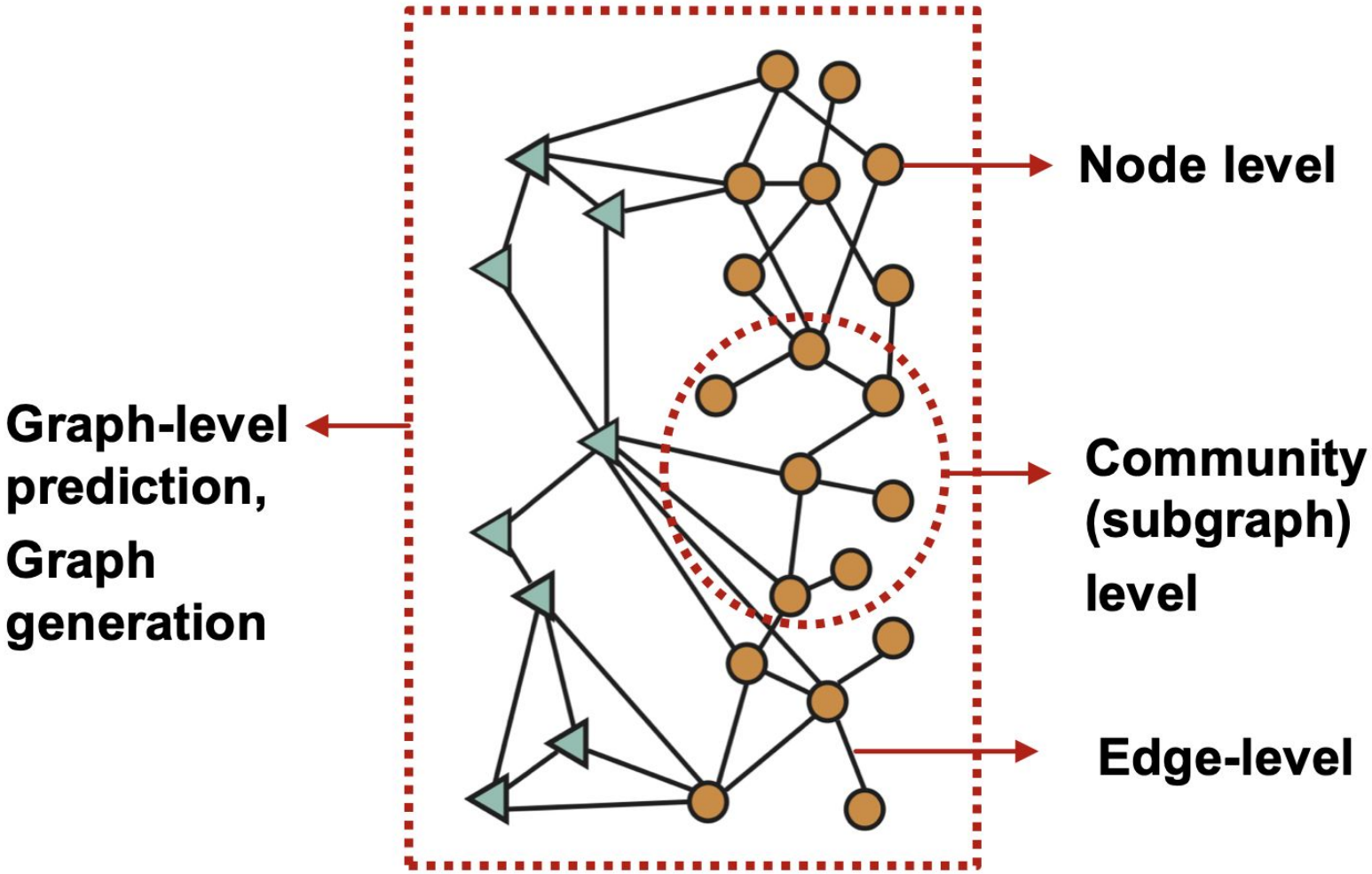
No fixed node ordering or reference point

This course



Credit: [1]

Different types of tasks



Credit: [1]

Classic Graph ML tasks

- **Node classification**: Predict a property of a node
  - **Example:** Categorize online users / items
- **Link prediction**: Predict whether there are missing links between two nodes
  - **Example:** Knowledge graph completion
- **Graph classification**: Categorize different graphs
  - **Example:** Molecular property prediction
- **Graph generation**:
  - **Example:** New molecule generation / Drug discovery

Credit: [1]

Application of Deep Learning on Graphs

Course Outline

Building concepts

1. Node Embeddings
2. Introduction to GNNs
3. Training Pipeline
4. Graph Transformers
5. Explainability
6. Scalability / Expressiveness
7. Augmentation

Applications

1. Knowledge Graphs
2. Temporal GNNs
3. Generative Modelling on Graphs
4. Drug Discovery

Questions so far?

Course Logistic

Prerequisites

Good background:

- Machine Learning
- Deep Learning
- Statistics/Probability

Programming:

- Python
- Unix Command Line

Helpful knowledge: familiarity with PyTorch

Relation to Other Courses with Similar Topics

- 263-3210-00L Deep Learning
- 263-0008-00L Computational Intelligence Lab
- 261-5120-00L Machine Learning for Health Care
- 263-5354-00L Large Language Models
- Computer Vision Courses

Helpful Resources

GNN Resources:

- Graph Representation Learning Book (William L. Hamilton)
- Geometric Deep Learning Book (Michael M. Bronstein, Joan Bruna, Taco Cohen, Petar Veličković)

ML Resources:

- Deep Learning Book (Ian Goodfellow, Yoshua Bengio, Aaron Courville)
- Pattern Recognition Book (Christopher M. Bishop)

Programming Resources:

- PyTorch Tutorial: https://pytorch.org/tutorials/beginner/basics/intro.html
- PyTorch Geometric Tutorial: https://pytorch-geometric.readthedocs.io/en/latest/

Course Format

Default: in presence teaching, Wednesday, 16.00–18.00, CAB

- About 1h lectures per week (with exceptions)
- Presentations of recent papers by students in groups of two (more details later today)
- Project work (more details later today)

Zoom: only for some (external) invited speakers

Course webpage: **https://tinyurl.com/eth-gnn-2023**

(https://bmi.inf.ethz.ch/teaching/263-5056-00l-applications-of-deep-learning-on-graphs-autumn-2023)

Course Moodle: **https://tinyurl.com/eth-gnn-2023-moodle**

(https://moodle-app2.let.ethz.ch/course/view.php?id=21170)

Communications: via Moodle

Schedule

Every Wednesday 4-6pm

Lecture

- First 45-90min, depending on week

Tutorial:

- To deepen understanding some key techniques (coding exercises, in depth explanation); not every week

Paper presentations (in groups of two):

- Opportunity to learn about a specific technique or application in greater detail
- Giving a presentation is mandatory and attending the other presentations is mandatory, too (max 2 misses in attendance of other presentations)

Grading

Exam: will cover lecture material

- Single exam
- In written form

Final grade will be composed of:

1. Projects (30%)
2. Paper presentation (10%)
3. Exam (60%)

More details about **projects** and **paper presentations** will follow by TA

Oversubscription & Waiting List

- This is the first year we offer the course. We therefore have a **limit of 50** students.

- But there are **168 (!) students** on the waiting list who really would like to take the course.

- If you decide to drop the course, please unregister as soon as possible.

- If you are still registered for the course by **October 5th** (about 2 weeks from now), we consider that you fully participate in the course and you will fail the course if you do not show up/do not deliver projects/presentations etc.

Questions so far?

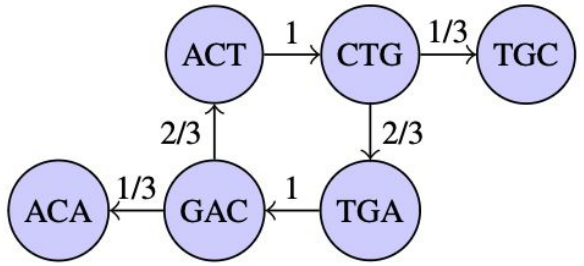
GNNs for Modeling Biomedical Data
(Biomedical Informatics Lab research)

Learning Representations of Genome Graphs

Kapusniak, Burger, Joudaki, Rätsch

- ACTGACT → ACT, CTG, TGA, GAC, ACT
- ACTGACA → ACT, CTG, TGA, GAC, ACA
- TGACTGC → TGA, GAC, ACT, CTG, TGC

Graph construction from k-mers



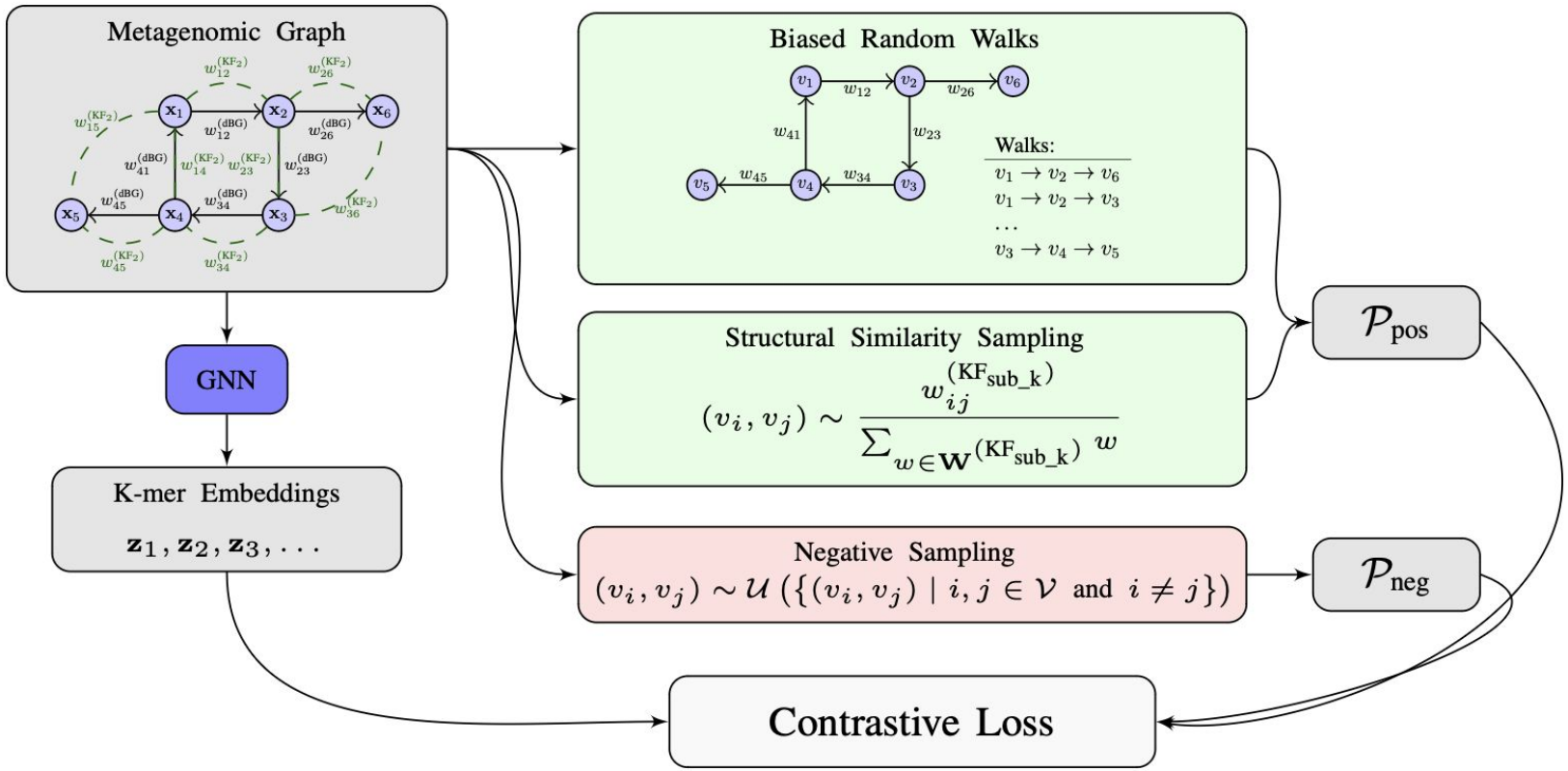
Genomic Sequence Data

Genomic De Brujin Graphs

Objective:
Learn unsupervised sequence and k-mer representations from genomic structures

Learning Representations of Genome Graphs

Kapusniak, Burger, Joudaki, Rätsch



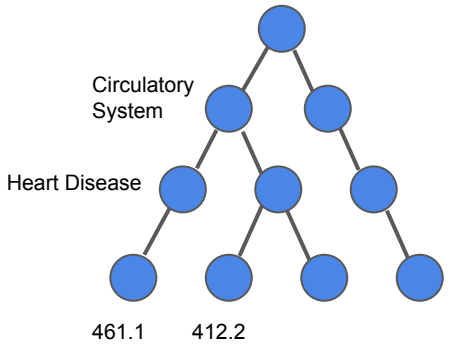
Top-1% Retrieval Performance

	Zero-Shot: Aggregated K-mer Embeddings			
k	One-Hot	Word2Vec	Node2Vec	Our CL
1	**50.3**	45 ± 0.1	45.3 ± 1.5	-
2	49.9	46.7 ± 0.1	49.9 ± 0.6	$\mathbf{52.2 \pm 0.7}$
3	46.8	48.6 ± 0.1	51.2 ± 0.2	$\mathbf{53.1 \pm 0.4}$
4	45.3	46.9 ± 0.1	49.8 ± 0.2	$\mathbf{53.3 \pm 0.3}$
5	45.4	42.3 ± 0.1	50 ± 0.4	$\mathbf{50.5 \pm 0.1}$
6	46.3	41.3 ± 0.1	49.6 ± 0.3	$\mathbf{50 \pm 0.7}$
7	44.5	-	-	$\mathbf{48.3 \pm 1.1}$
8	-	-	-	$\mathbf{50.2 \pm 0.1}$

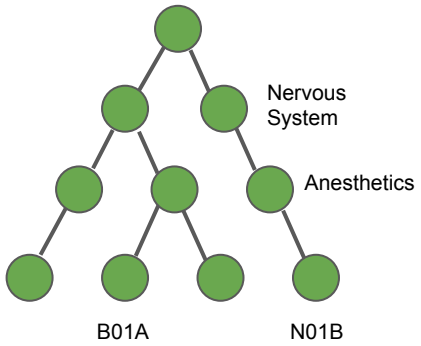
By capturing contextual as well as structural information in our k-mer representations we can see improved performance

Exploiting Prior Knowledge for Learning Representations of Concepts

Burger, Rätsch, Kuznetsova

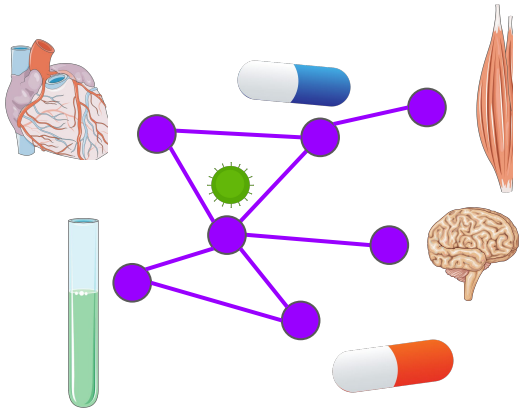


ICD9: Diseases

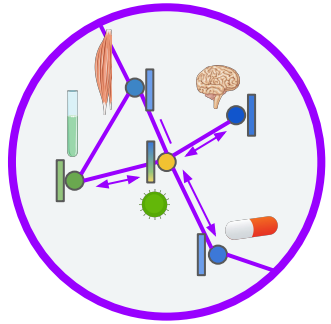


ATC: Prescriptions

aggregated



UMLS: Concepts



Graph Neural Networks

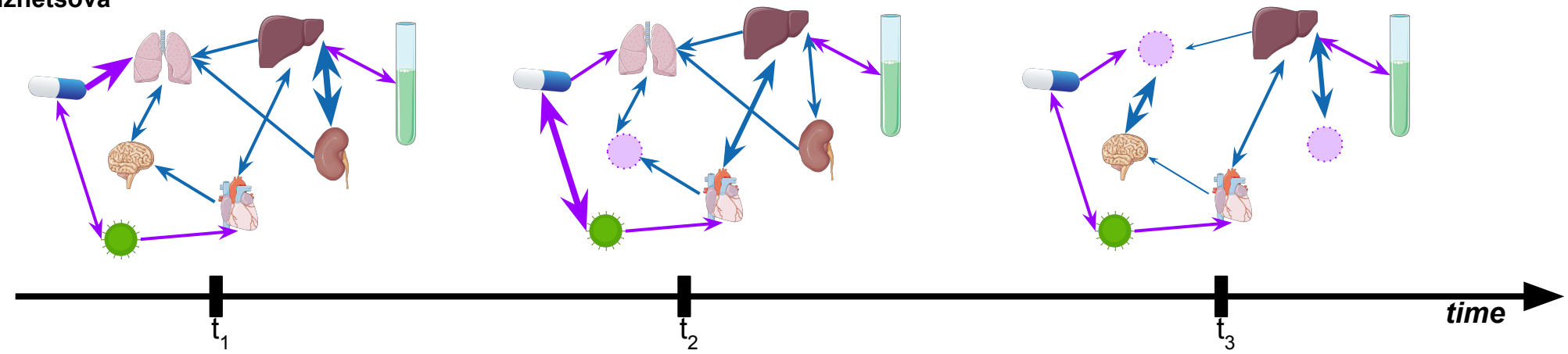
Objective:
Ground learning with **structured prior knowledge**

By grounding our representations with prior knowledge from the UMLS Metathesaurus, … improved performance [...] downstream tasks.
- Burger et al.

Method	Heart Failure		Diagnosis		
	AuROC	*F1*	*w-F1 (infl.)*	*R@20*	*R@40*
Embedding Mat.	87.02±0.49	74.26±1.00	24.87±0.24	40.17±0.12	51.41±0.29
Node2Vec *w/o* $\mathcal{C}(n)$	86.07±0.21	73.38±0.20	24.69±0.30	40.10±0.23	51.47±0.10
Cui2Vec *w/o* $\mathcal{C}(n)$	86.15±1.66	73.20±1.62	25.03±0.09	40.57±0.34	52.18±0.10
GNN ICD/ATC	87.03±0.10	74.07±0.18	24.75±0.29	40.15±0.09	51.64±0.23
GNN ICD/ATC-CO	87.05±0.15	74.16±0.04	24.59±0.41	39.95±0.22	51.50±0.34
MMUGL w/o $\mathcal{C}(n)$	86.27±0.18	73.36±0.70	25.01±0.46	40.42±0.43	51.78±0.35
Node2Vec *with* $\mathcal{C}(n)$	87.03±0.35	74.25±0.38	24.75±0.87	40.02±0.77	51.23±0.79
Cui2Vec *with* $\mathcal{C}(n)$	87.51±0.10	**74.71±0.24**	25.84±0.11	41.02±0.15	52.47±0.16
MMUGL	87.19±0.21	74.46±0.41	25.81±0.17	41.02±0.10	52.41±0.14
MMUGL, $w_{\bullet,m} = 0$	**87.60±0.40**	**74.71±0.72**	**26.40±0.02**	**41.54±0.17**	**53.02±0.37**

Using Knowledge Graph Representations to enhance ICU Time-Series Predictions

Jain, Burger, Rätsch, Kuznetsova



Objective:
Improved **time-step** representations through **structural priors**

Learn:

- Variable relationships
- Impute missing data
- Include structured prior knowledge

Method	Mortality	
	AuPRC	*AuROC*
LSTM	50.1 ± 1.3	86.1 ± 0.3
MM1DCNN	52.5 ± 1.3	86.5 ± 0.4
CMT	52.7 ± 1.0	87.1 ± 0.6
MAC	56.2 ± na	85.7 ± na
Raindrop	44.6 ± 2.1	83.7 ± 0.5
Ours	**58.7 ± 0.6**	**89.3 ± 0.1**

Questions so far?

BIOMEDICAL INFORMATICS

Slides & Image Credits

1. CS224W: Machine Learning with Graphs
2. DeepMind AlphaFold
3. Wikipedia
4. Graph Representation Learning Book
5. Disease Pathways
6. Gene Regulatory Network
7. Homology modeling
8. Molecule graph
9. Social Graph
10. Knowledge Graph
11. 3D shapes
12. Underground Network
13. Computer Network