KNOWLEDGE-AUGMENTED GRAPH MACHINE LEARNING FOR DRUG DISCOVERY: FROM PRECISION TO INTERPRETABILITY

Zhiqiang Zhong & Davide Mottin

Aarhus University

AUGUST 6 - 10
KDD2023
LONG BEACH, CA

AARHUS
UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

OUTLINE

I. Introduction and Motivation

II. Background of Drug Discovery

III. Graph Machine Learning (GML) and Knowledge Graph (KG) in Drug Discovery

IV. Knowledge-augmented Graph Machine Learning (KaGML) for Drug Discovery

V. Practical Resources

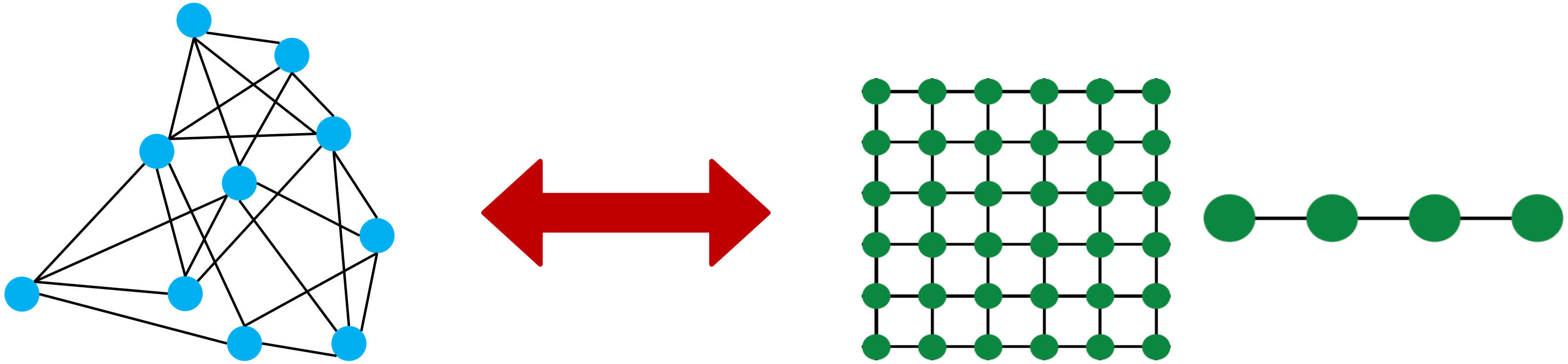
VI. Open Challenges and Future Directions

FUNDAMENTALS OF GRAPH MACHINE LEARNING (GML) AND KNOWLEDGE GRAPH (KG)

DEALING WITH GRAPHS IS DIFFICULT

Graphs are far more complex!

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



Graphs

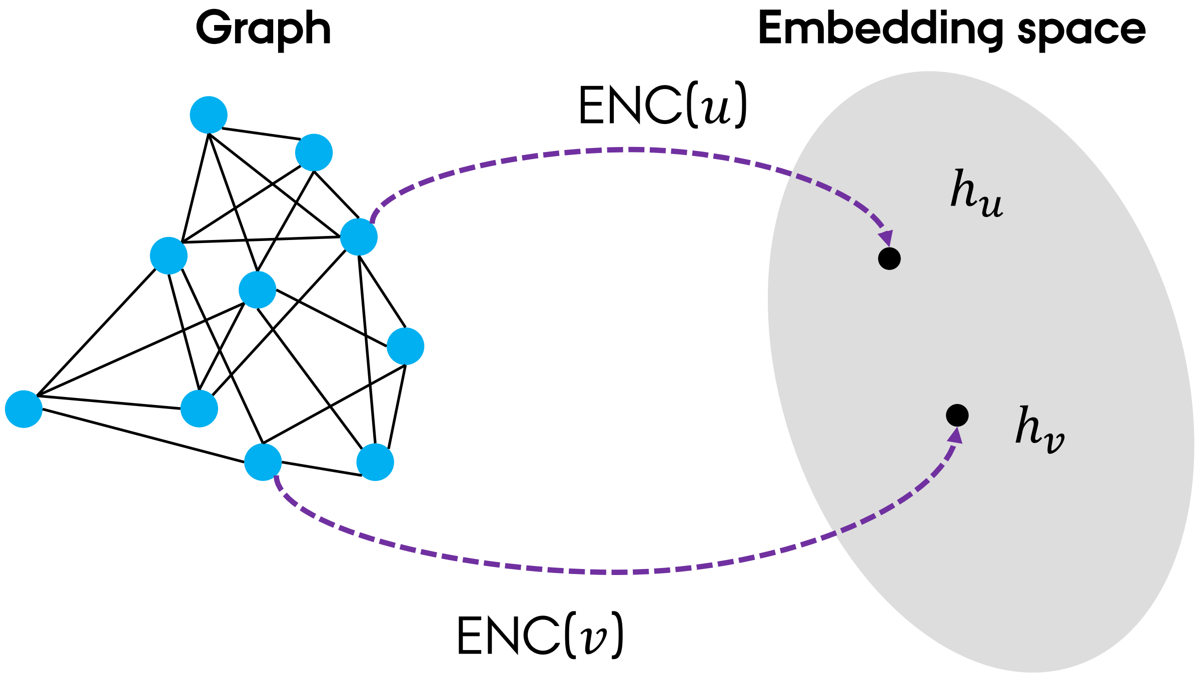
Images

Text

- No fixed node ordering
- Often dynamic and have multimodal features

(KNOWLEDGE) GRAPH MACHINE LEARNING

- (Knowledge) Graph Representation Learning



❑ Node property prediction

$\hat{Y}_u = f(h_u)$

❑ Link prediction

$\hat{Y}_{uv} = f(h_u, h_v, e_{uv})$

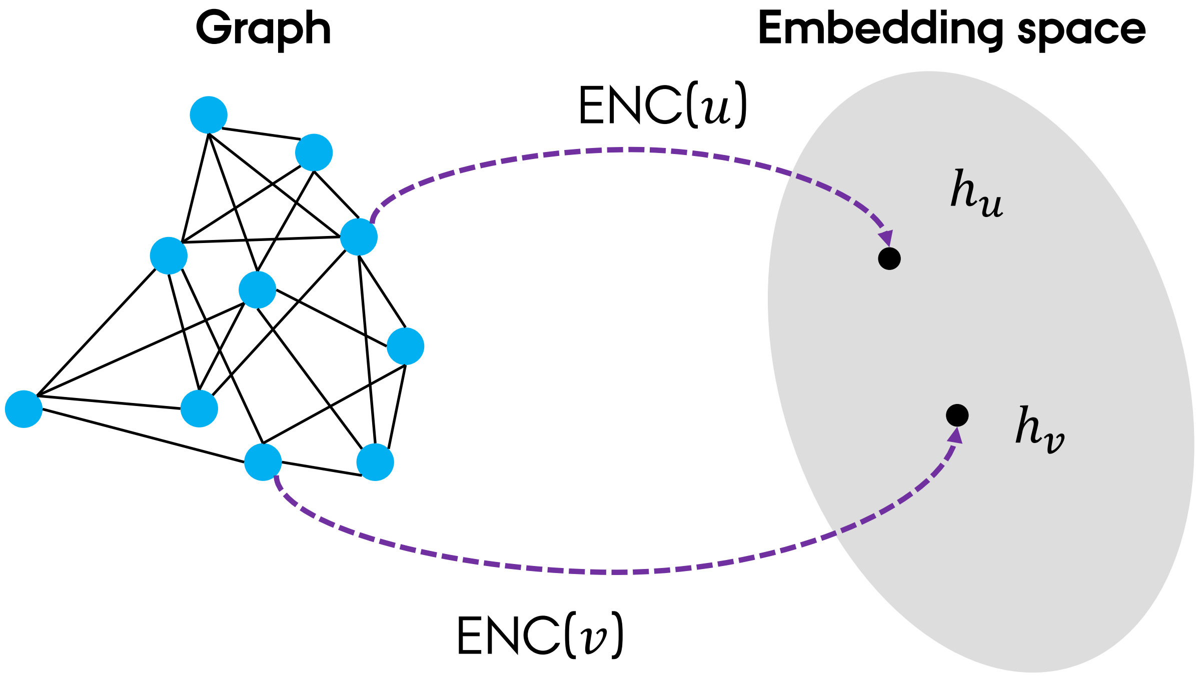
❑ Graph property prediction

$\hat{Y}_{\mathcal{G}} = f(\bigoplus_{u \in \mathcal{V}} h_u)$

❑ Etc.

(KNOWLEDGE) GRAPH MACHINE LEARNING

- (Knowledge) Graph Representation Learning



□ Node property prediction

$\hat{Y}_u = f(h_u)$ – $\textsc{PredEntity}$

□ Link prediction

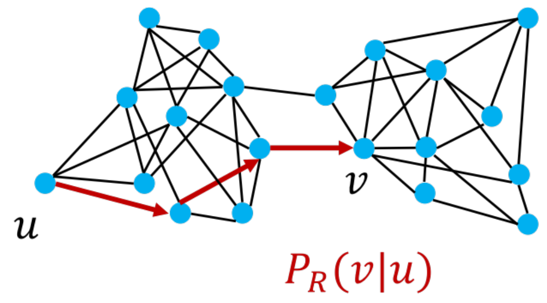
$\hat{Y}_{uv} = f(h_u, h_v, e_{uv})$ – $\textsc{PairEntity}$

□ Graph property prediction

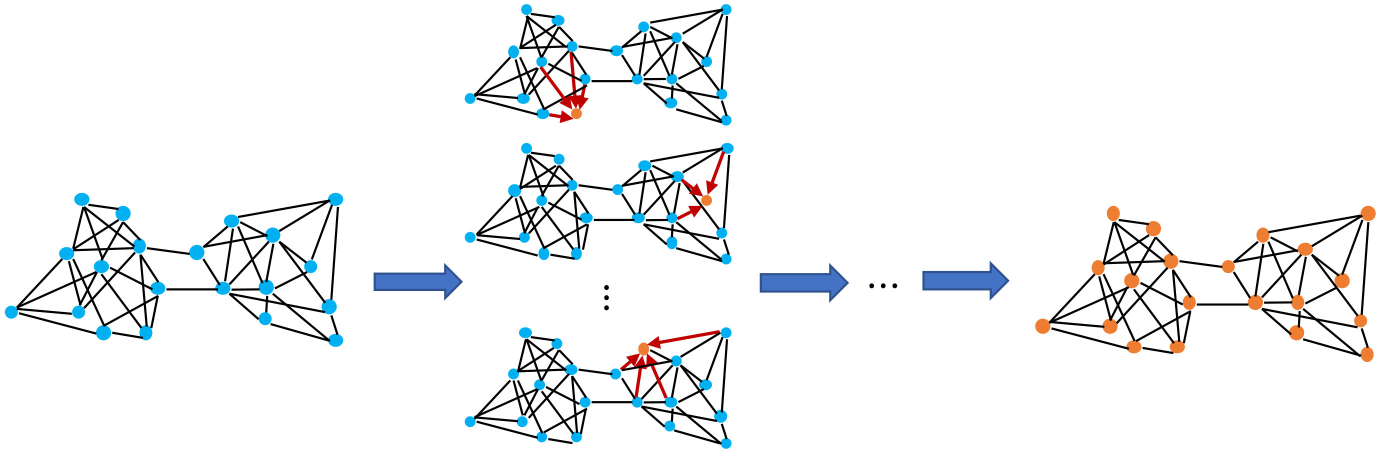
$\hat{Y}_{\mathcal{G}} = f(\bigoplus_{u \in \mathcal{V}} h_u)$ – $\textsc{Entity2Entity}$

□ Etc.

(KNOWLEDGE) GRAPH MACHINE LEARNING



Estimate the probability of visiting node v on a random walk starting from node u using some random walk strategy R.

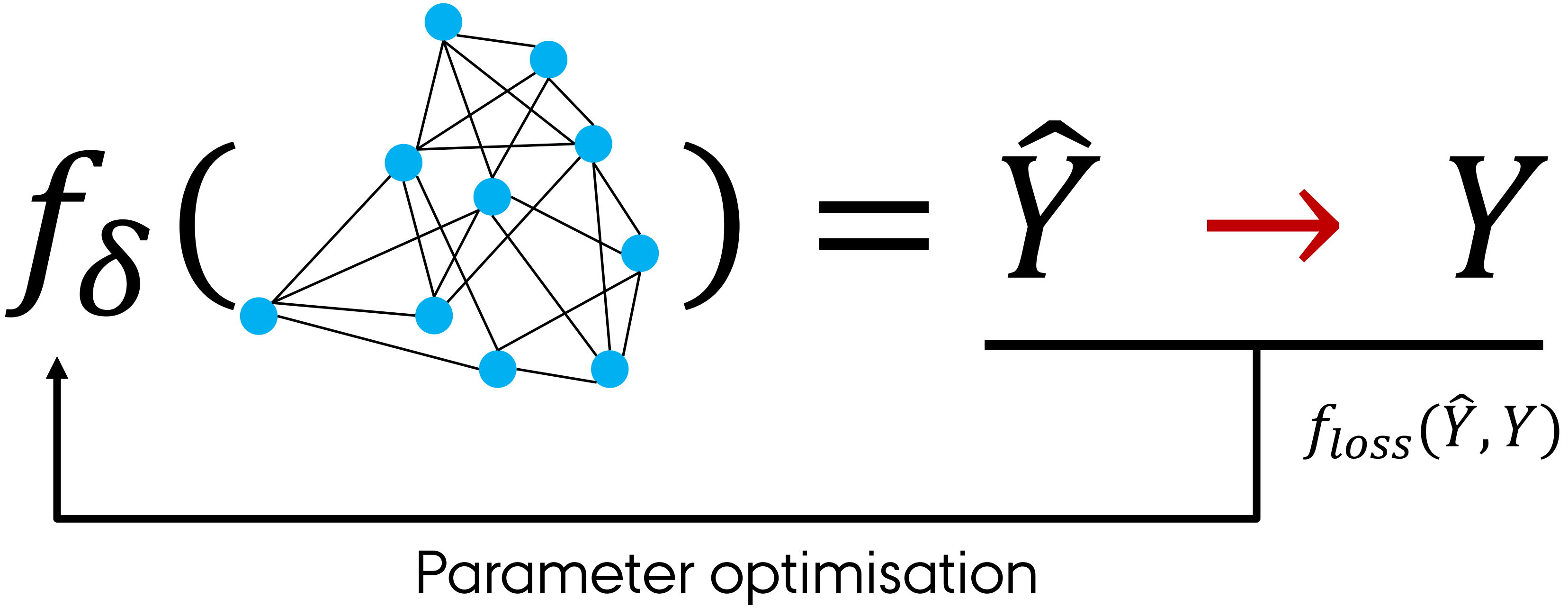


1st hop aggregation

Lth hop aggregation

"Shallow" (K)GML Approaches *vs.* "Deep" (K)GML Approaches

HOW TO TRAIN (K)GML MODELS?



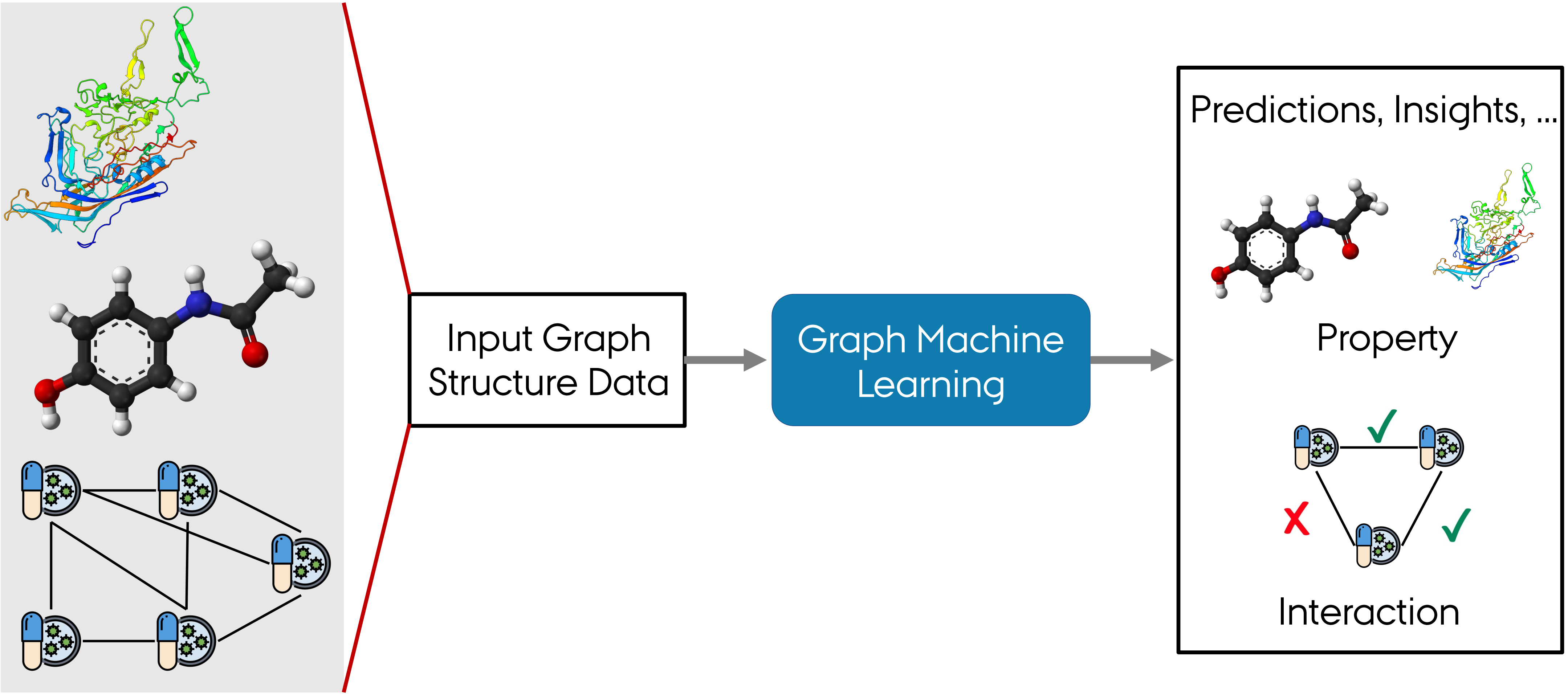
$$f_\delta(\;\cdot\;) = \hat{Y} \rightarrow Y$$

$f_{loss}(\hat{Y}, Y)$

Parameter optimisation

GML AND KG FOR DRUG DISCOVERY

GML FOR DRUG DISCOVERY



Input Graph Structure Data

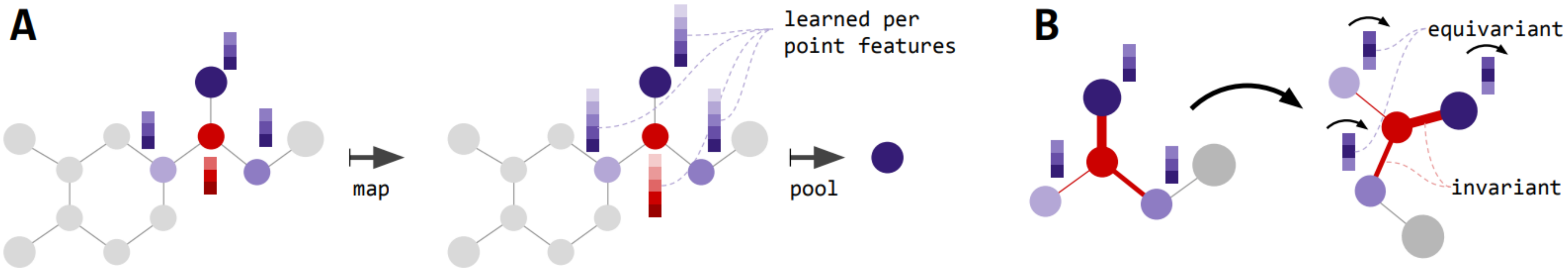
Graph Machine Learning

Predictions, Insights, …

Property

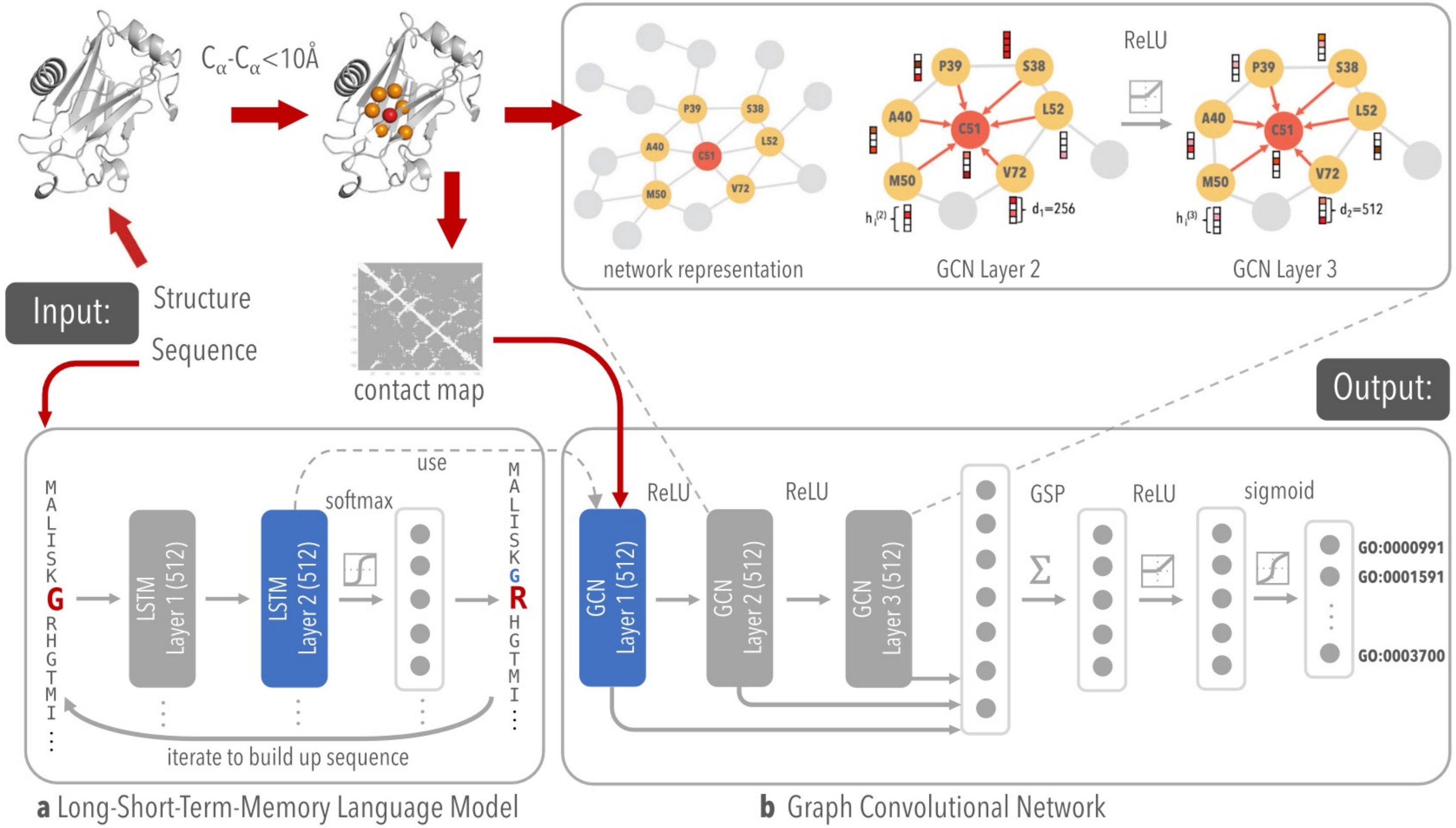
Interaction

SE(3)-TRANSFORMER (NERUIPS’20)



- Rich information about molecules can be summarised into molecular graphs
- A variant of Transformer for 3D biomedical graphs, which is equivariant under continuous 3D roto-translations

DEEPFRI (NAT. COMMUN.'21)



- One protein can be represented as a graph by connecting residues close in 3D space
- Proteins can be organised into a big graph based on their similarities
- GML encoders can capture information from different perspectives about proteins

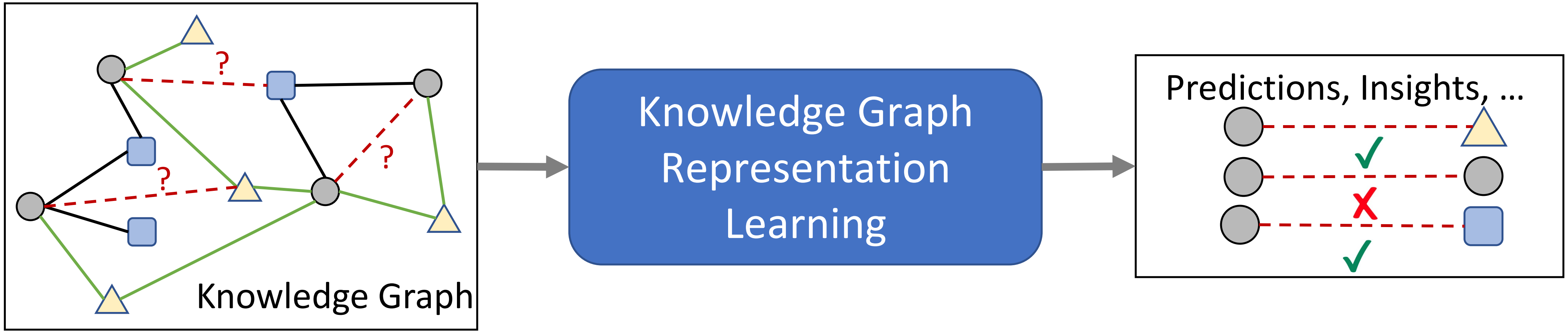
ONE MORE STEP: INVESTIGATE GML

- High data dependency
  - The effectiveness of GML depends on high-qualified training data
  - Biomedical data generation is time-consuming and expensive
- Poor generalisation
  - Uncertain performance on instances that have never been observed in training data
- Lacks interpretability
  - “Black box” damages the usability of clinical treatment

GML FOR DRUG DISCOVERY – PAPER LIST

- Survey papers
  - Utilizing graph machine learning within drug discovery and development, Brief. Bioinformatics, 2021.
  - Graph representation learning in biomedicine and healthcare, Nat. Biomed. Eng., 2022.
  - Graph-based generative models for de novo drug design, Drug Discov. Today Technol., 2019.
  - A compact review of molecular property prediction with graph neural networks, Drug Discov. Today Technol., 2020.
- Some representative papers
  - Protein sequence design with a learned potential, Nat. Commun., 2022.
  - Learning from protein structure with geometric vector perceptrons, ICLR, 2021.
  - Deep learning of high-order interactions for protein interface prediction, KDD, 2020.
  - An E(3) equivariant variational autoencoder for molecular linker design, ICML, 2022.

KG FOR DRUG DISCOVERY

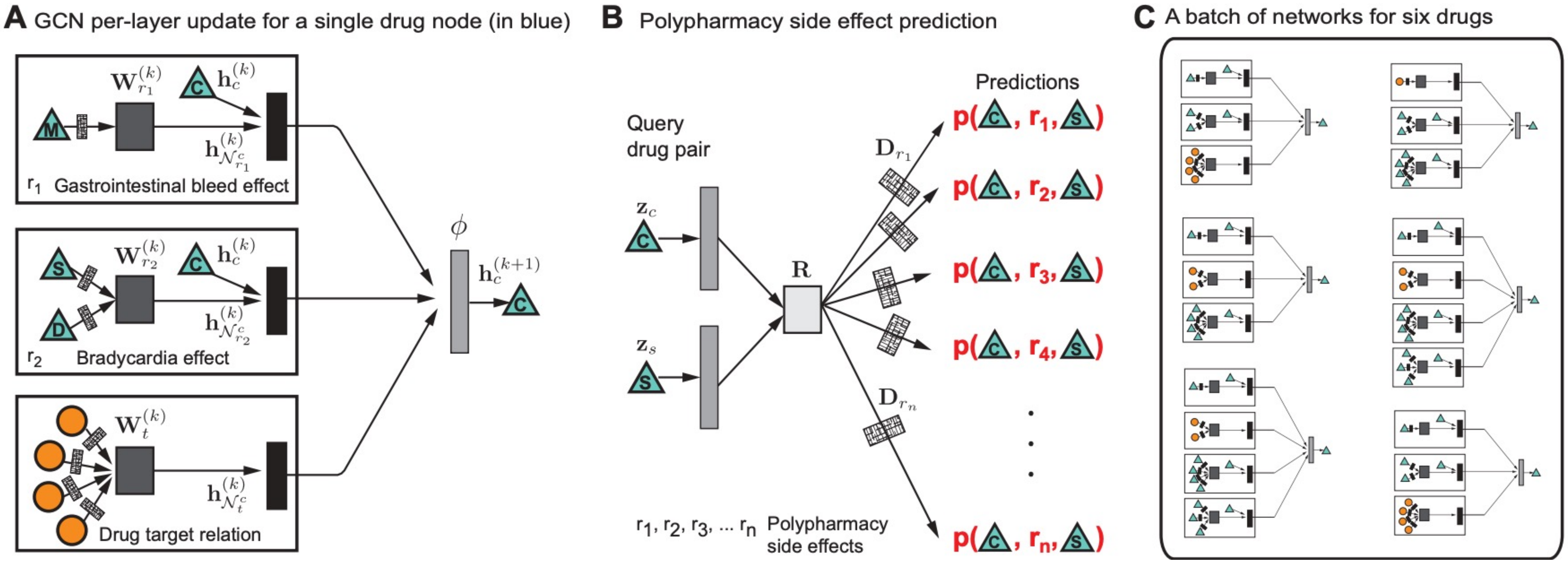


Knowledge Graph

Knowledge Graph Representation Learning

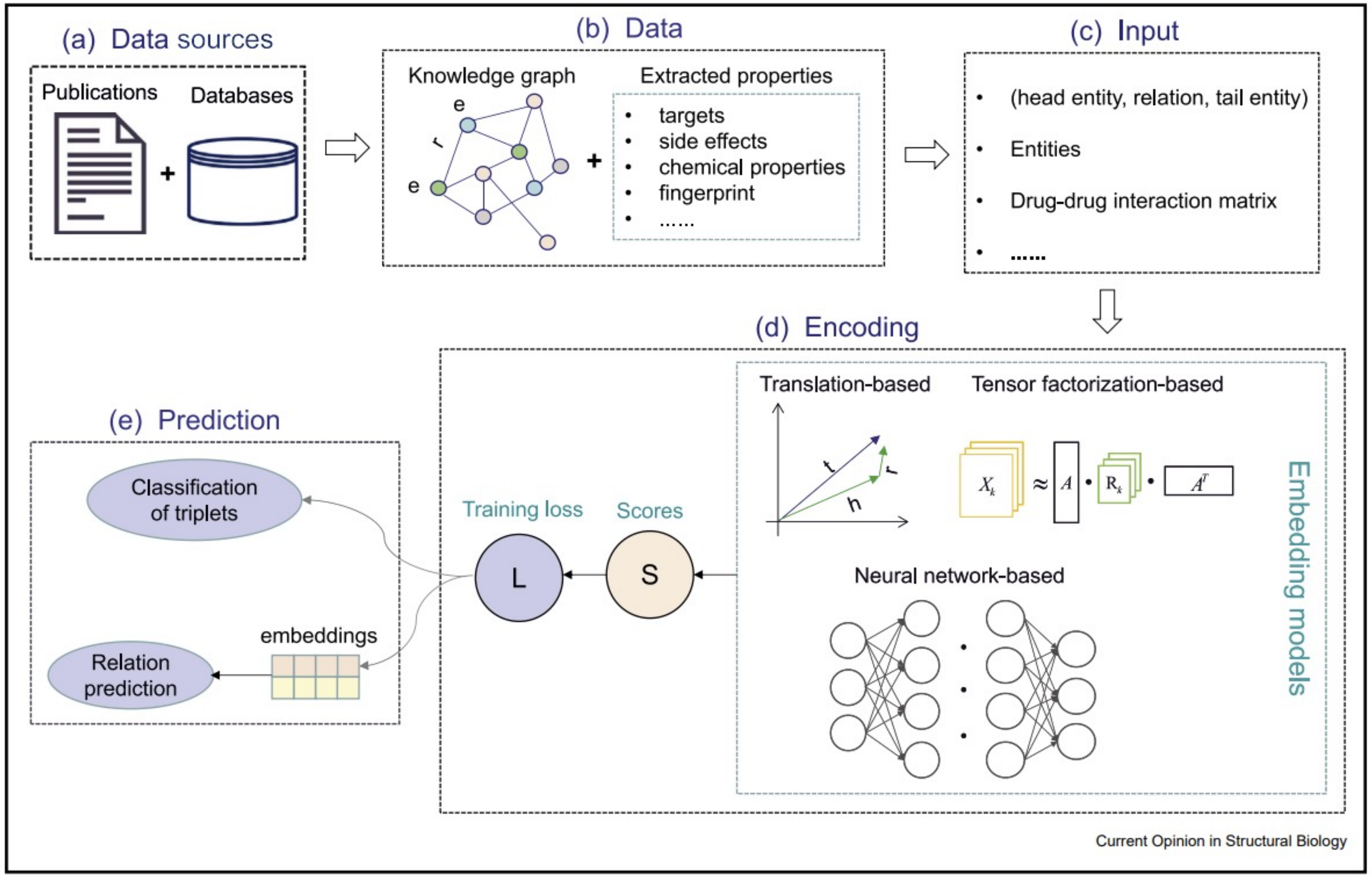
Predictions, Insights, …

DECAGON (BIOINFOM. 18)



- KGRL methods summarise information about each drug
- The side effect is predicted based on the knowledge about each drug

ZENG ET AL. (CURR. OPIN. STRUCT. BIOL.'22)



Current Opinion in Structural Biology

- A complete pipeline from raw data sources to construct KGs
- Making predictions based on knowledge from KG

ONE MORE STEP: INVESTIGATE KG

- High data dependency
  - The effectiveness of KG depends on large-scale high-qualified training data
  - Biomedical data generation is time-consuming and expensive
- Poor generalisation
  - Supported tasks are limited to the KG context

- Good interpretability
  - Results generated based on human knowledge are more reliable.

KG FOR DRUG DISCOVERY – PAPER LIST

- Survey papers
  - Building a knowledge graph to enable precision medicine, Biarxiv, 2022.
  - Toward better drug discovery with knowledge graph, Curr. Opin. Struct. Biol., 2022.
- Some representative papers
  - Drug knowledge bases and their applications in biomedical informatics research, Brief. Bioinformatics, 2019.
  - Machine learning prediction and tau-based screening identifies potential alzheimer's disease genes relevant to immunity, Commun. Biol., 2022.
  - Knowledge graph-based recommendation framework identifies drivers of resistance in EGFR mutant non-small cell lung cancer, Nat. Commun., 2022.

ACKNOWLEDGEMENT

Zhiqiang Zhong
Postdoc
Aarhus University

Davide Mottin
Asst. Prof.
Aarhus University

Our work is supported by the Horizon Europe and Danmarks Innovationsfond under the Eureka, Eurostar grant no E115712.

Thank you!

Questions?

zzhong@cs.au.dk
https://zhiqiangzhongddu.github.io/

AARHUS
UNIVERSITY