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

Zhiqiang Zhong & Davide Mottin

Aarhus University

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

TAXONOMY OF KNOWLEDGE-AUGMENTED GRAPH MACHINE LEARNING (KAGML)

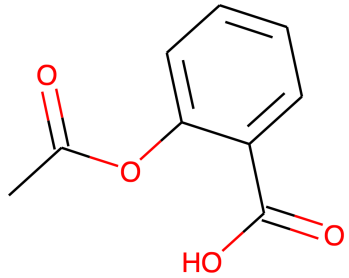
AARHUS UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

KDD'23 TUTORIAL
9 AUGUST 2023

ZHIQIANG ZHONG AND DAVIDE MOTTIN
DATA-INTENSIVE SYSTEMS GROUP

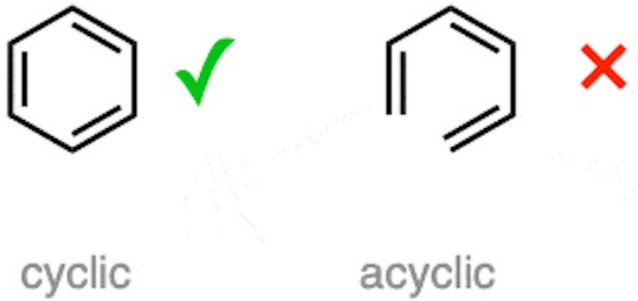
KNOWLEDGE IN DRUG DISCOVERY

- A graph-structured data presents primary information
  - Aspirin contains a phenyl ring ✓
  - Aspirin does not contain a phenyl ring ✗

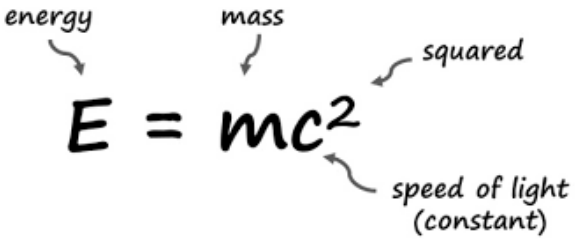


Aspirin

- External knowledge indicates deeper insights
  - Dropping a carbon atom in the phenyl ring of aspirin can lead to a dramatic change in the aromatic system and result in an alkene chain

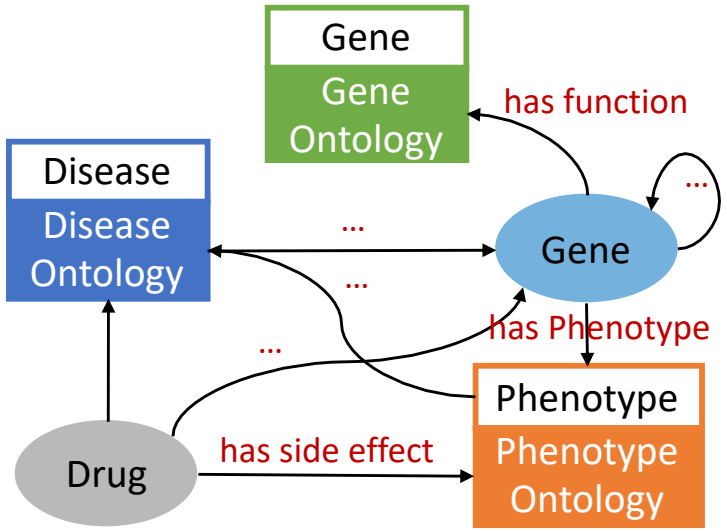


HUMAN BIOMEDICAL KNOWLEDGE



Theories and Equations

Description Context

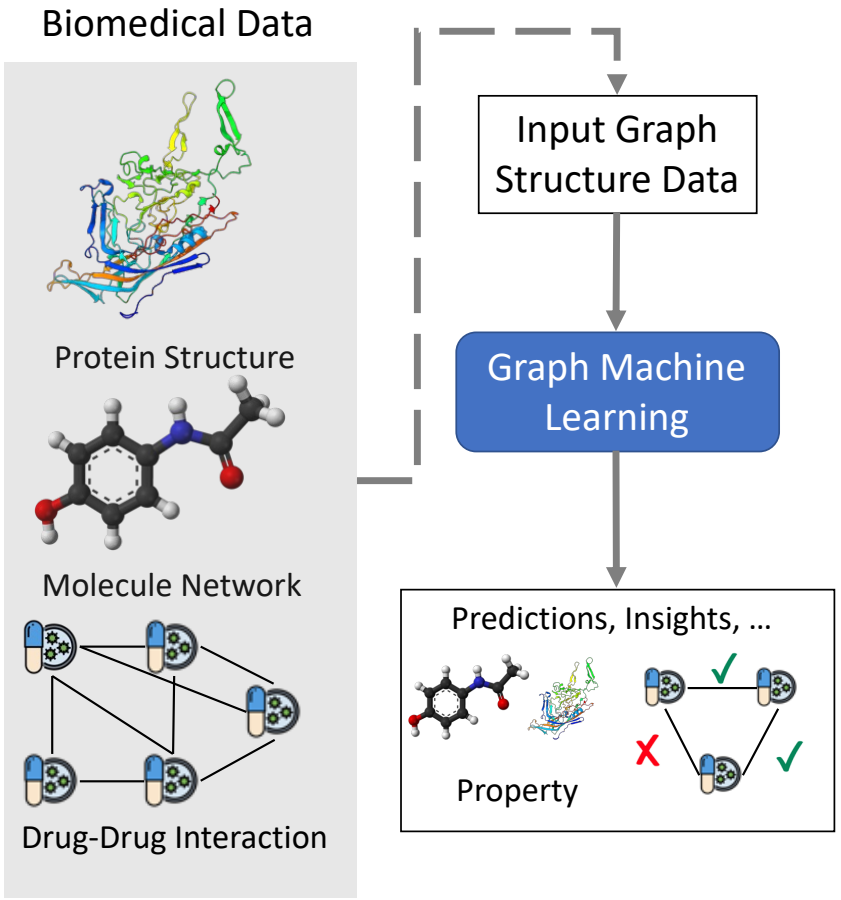


Knowledge Graph

- Knowledge is any external information absent from the input graph but helpful for generating the output

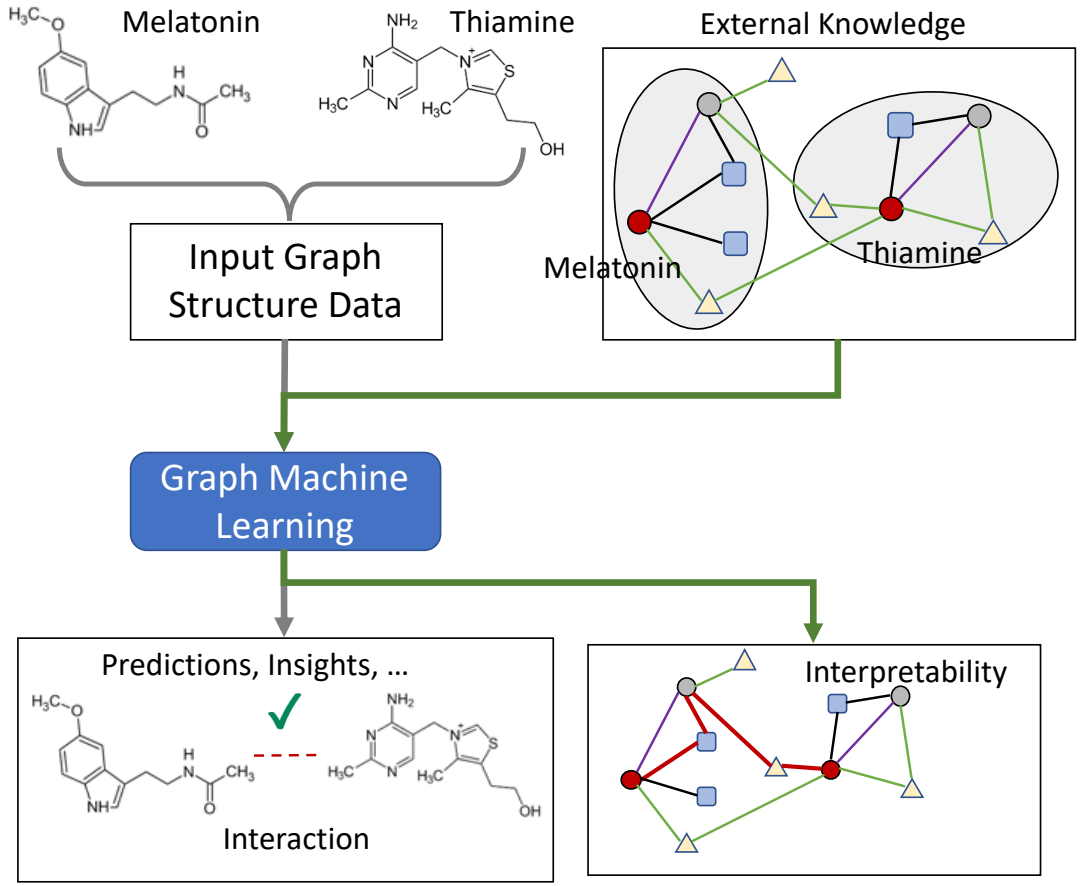
WHAT IS KNOWLEDGE-AUGMENTED GML

GML for Drug Discovery



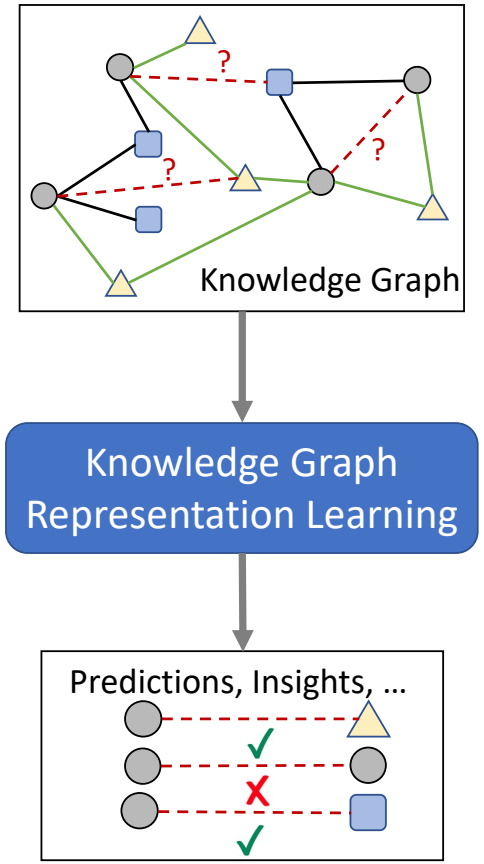
VS.

KaGML for Drug Discovery



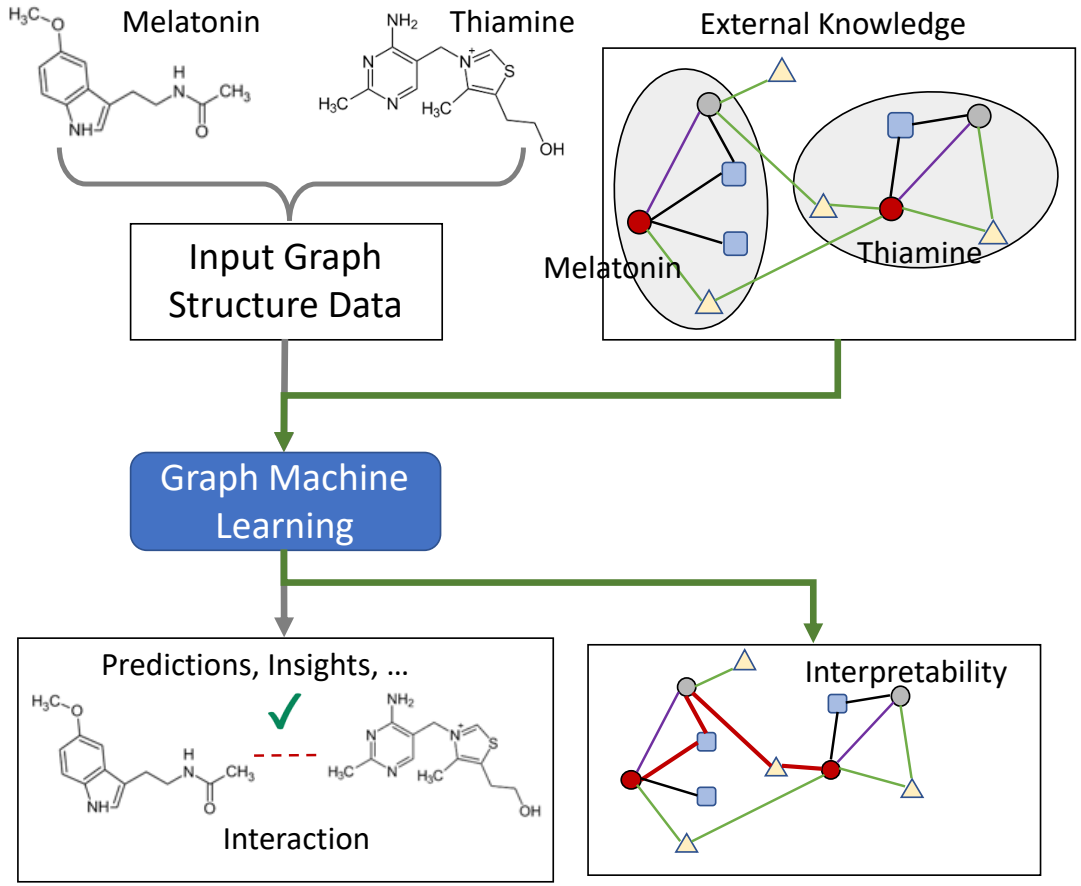
WHAT IS KNOWLEDGE-AUGMENTED GML

KG for Drug Discovery

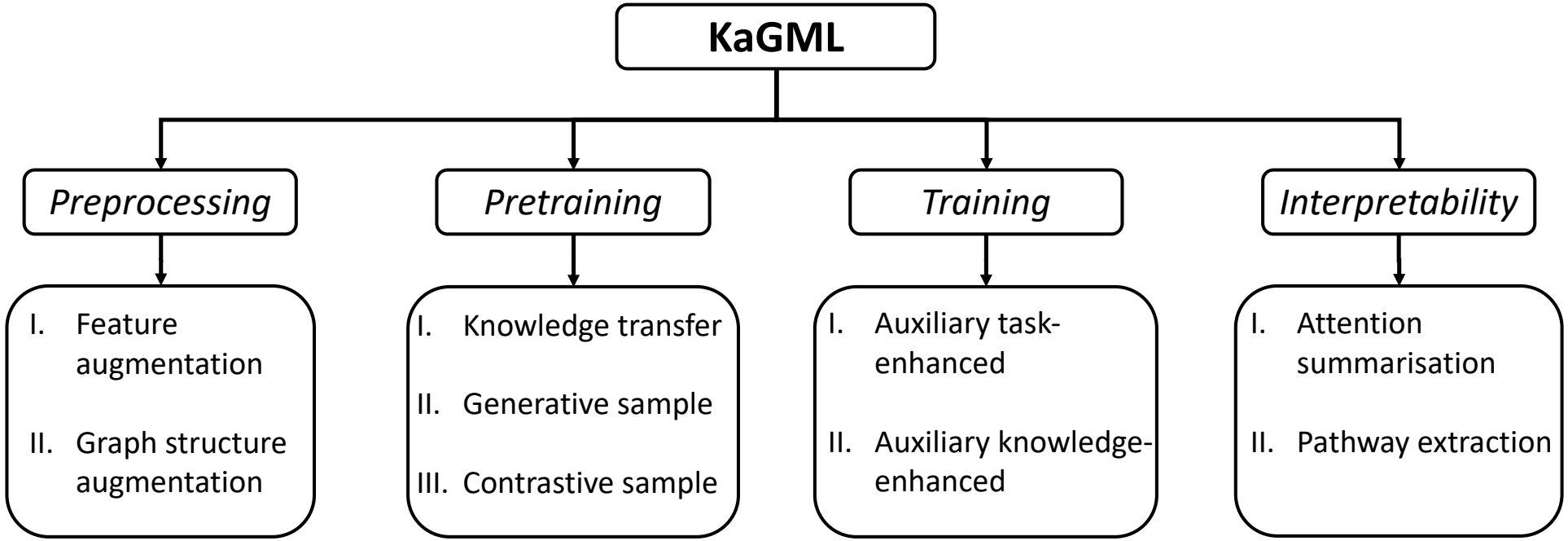


VS.

KaGML for Drug Discovery



TAXONOMY OF KAGML



- ML venues: ICLR, ICML, NeurIPS, AAAI, KDD, AISTATS, IJCAI, TKDE, CIKM, etc.
- Biomedical venues: Bioinform., J. Chem. Inf. Model., ACS Omega, Int. J. Mol. Sci., etc.
- Interdisciplinary venues: Nature, Science, Nat. Mach. Intell., Nat. Methods, Patterns, etc.

INCORPORATING KNOWLEDGE IN *PRE-PROCESSING*

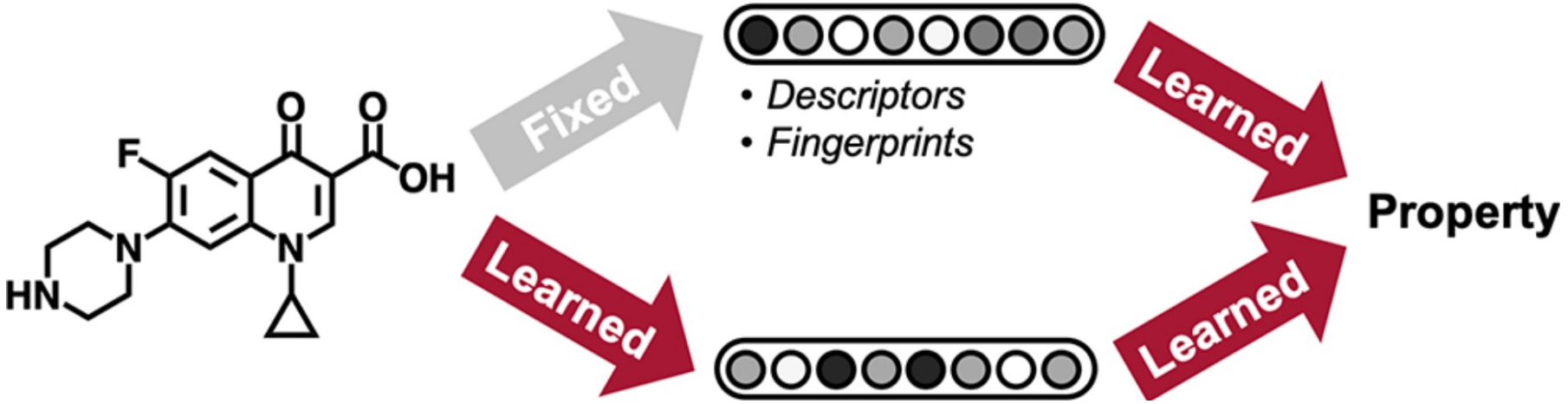
AARHUS UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

KDD'23 TUTORIAL
9 AUGUST 2023

ZHIQIANG ZHONG AND DAVIDE MOTTIN
DATA-INTENSIVE SYSTEMS GROUP

D-MPNN (J. CHEM. INF. MODEL.'19)

Pre-processing

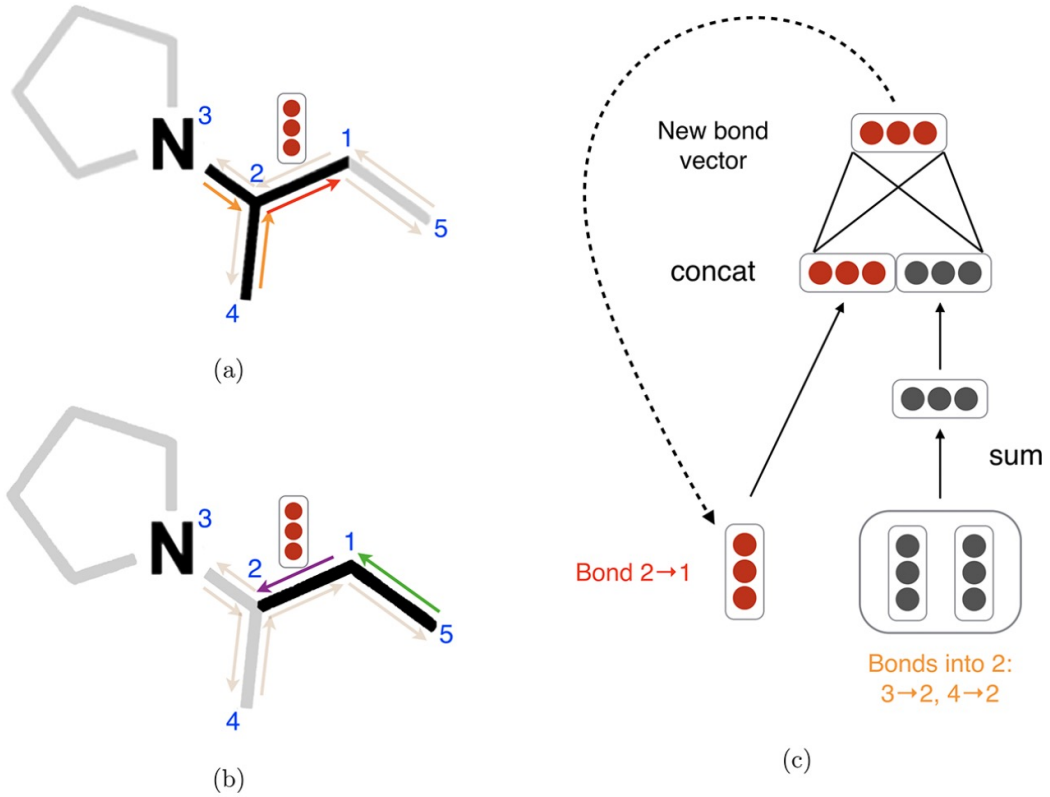


- Employ molecular node and edge features generated by chemical tools, such as RDKit[1] and UFF[2]
- Global interactions beyond L hops can be summarised into generated features

[1] *A Patrícia Bento, et al.*. An open source chemical structure curation pipeline using rdkit. J. Cheminformatics, 2020
[2] *Anthony K Rappé, at al.*. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations, J. Am. Chem. Soc., 1992

D-MPNN (J. CHEM. INF. MODEL.'19)

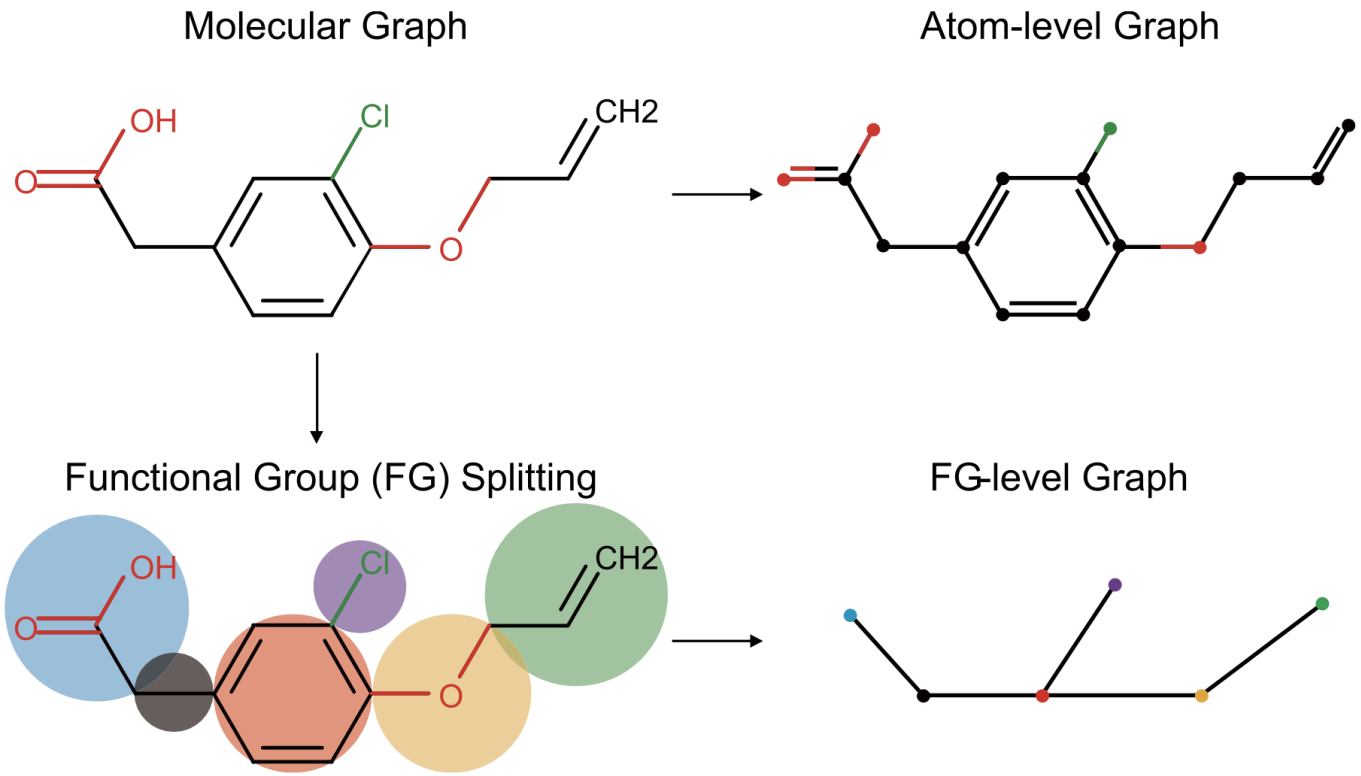
Pre-processing



- Generated node and edge features can be naturally integrated into the pipeline of GML algorithms.

RELMOLE (J. CHEM. INF. MODEL.'22)

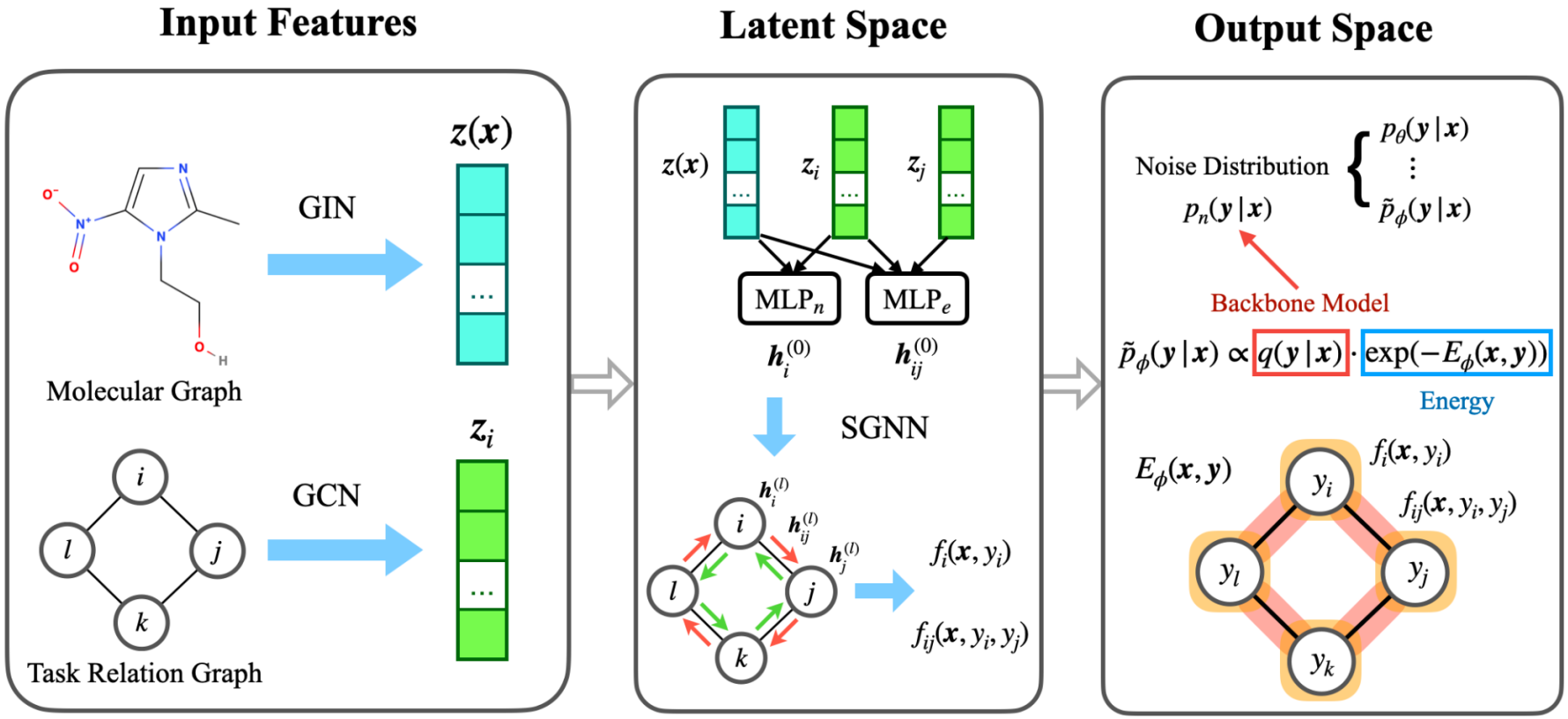
Pre-processing



- Graph augmentation based on biomedical knowledge

SGNN-EBM (AISTATS'22)

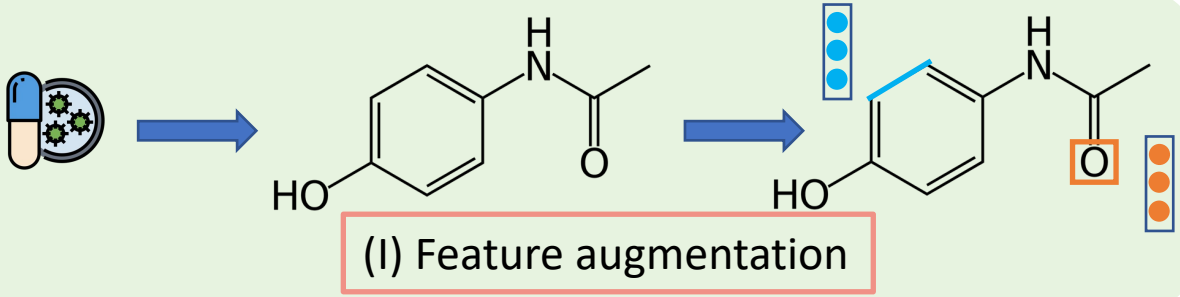
Pre-processing

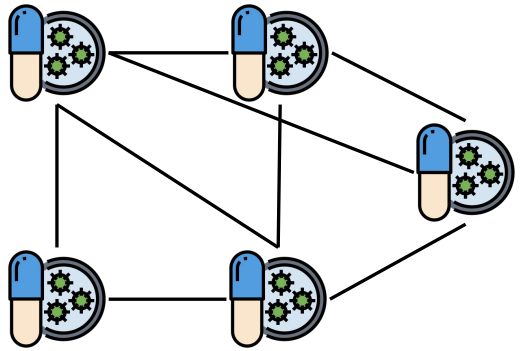


- Different tasks can be organised as task relation graphs to reveal the relationship between them.

KNOWLEDGE IN PRE-PROCESSING

- **Feature augmentation**
  - Protein backbone dihedral angles
  - Expert-engineered descriptors or molecular fingerprints, e.g., Dragon descriptors or Morgan fingerprints



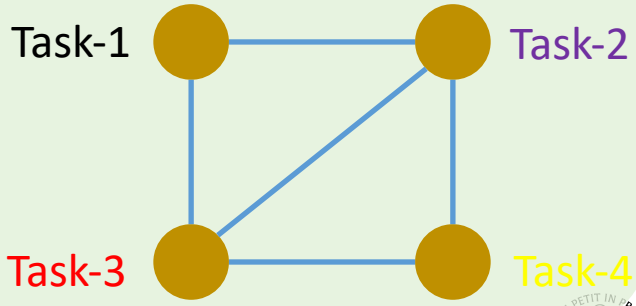


Task-1:
Drug property prediction

- **Graph structure augmentation**
  - 3D protein distance
  - A task graph can be constructed by counting the number of corresponding proteins

(II) Graph structure augmentation

Task 1: UniProt-...
Task 2: UniProt-...
Task 3: UniProt-...
Task 4: UniTrot-...
...



INCORPORATING KNOWLEDGE IN *PRE-TRAINING*

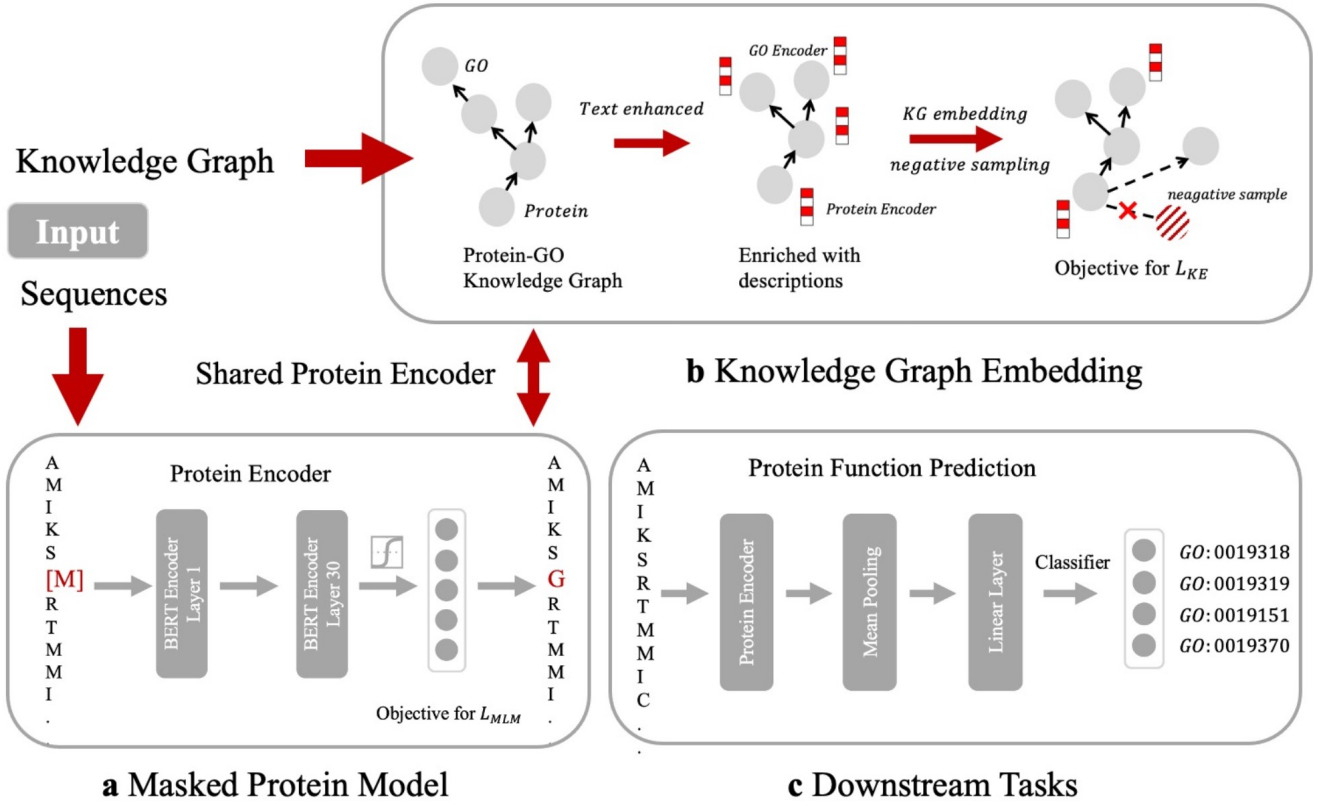
AARHUS UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

KDD'23 TUTORIAL
9 AUGUST 2023

ZHIQIANG ZHONG AND DAVIDE MOTTIN
DATA-INTENSIVE SYSTEMS GROUP

ONTOPROTEIN (ICLR'22)

Pre-Training

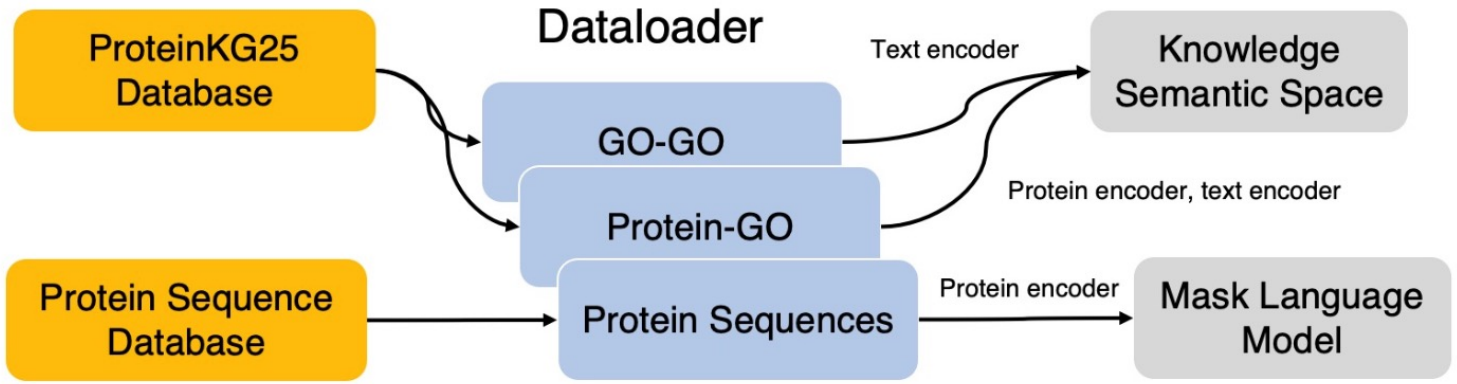


- Knowledge from Gene Ontology[1] database can optimise protein embedding during pre-training

[1] The Gene Ontology Consortium. The gene ontology resource: enriching a gold mine. Nucleic Acids Research, 2021

ONTOPROTEIN (ICLR'22)

Pre-Training



The dataflow of OntoProtein

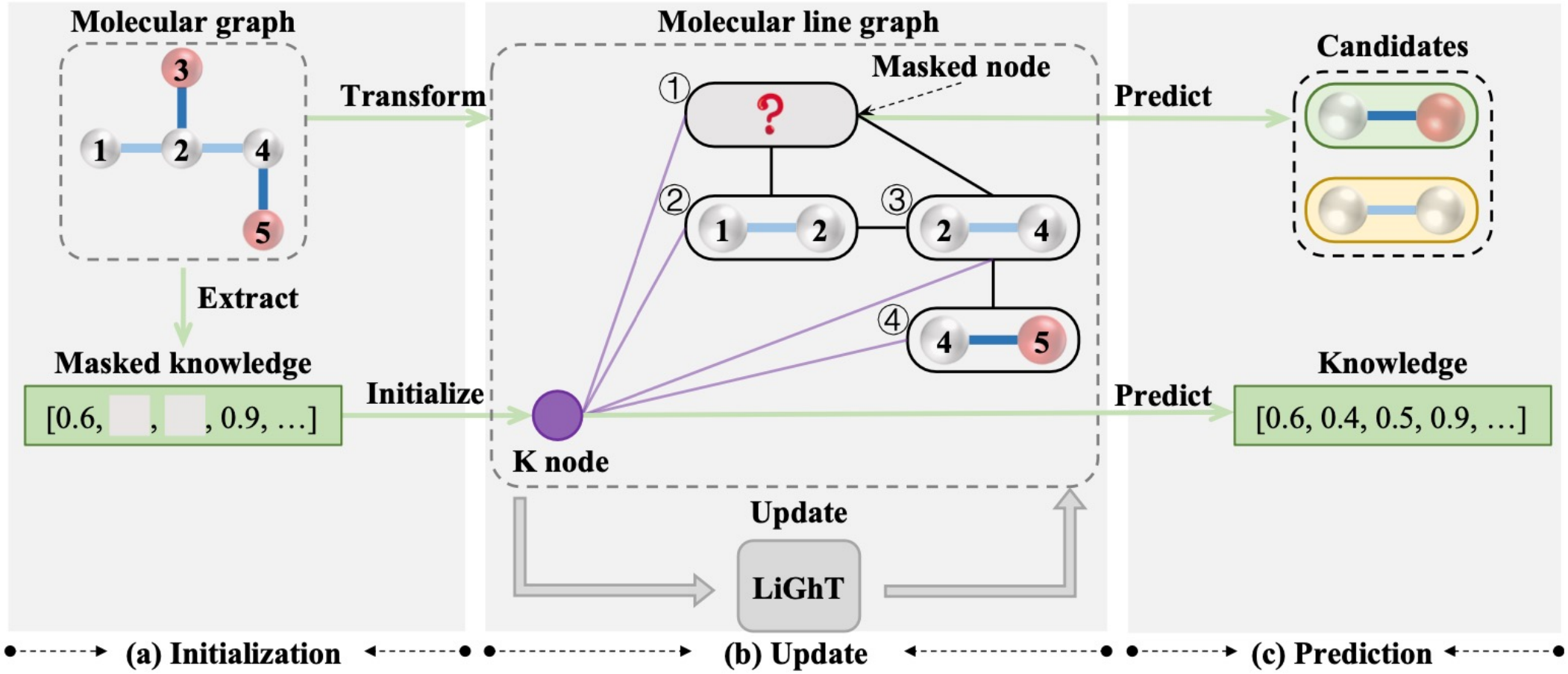
Method	Structure			Evolutionary	Engineering	
	SS-Q3	SS-Q8	Contact	Homology	Fluorescene	Stability
LSTM	0.75	0.59	0.26	0.26	0.67	0.69
TAPE Transformer	0.73	0.59	0.25	0.21	**0.68**	0.73
ResNet	0.75	0.58	0.25	0.17	0.21	0.73
MSA Transformer	-	**0.73**	**0.49**	-	-	-
ProtBert	0.81	0.67	0.35	**0.29**	0.61	**0.82**
OntoProtein	**0.82**	0.68	0.40	0.24	0.66	0.75

- Results on TAPE Benchmark

KPGT (KDD'22)

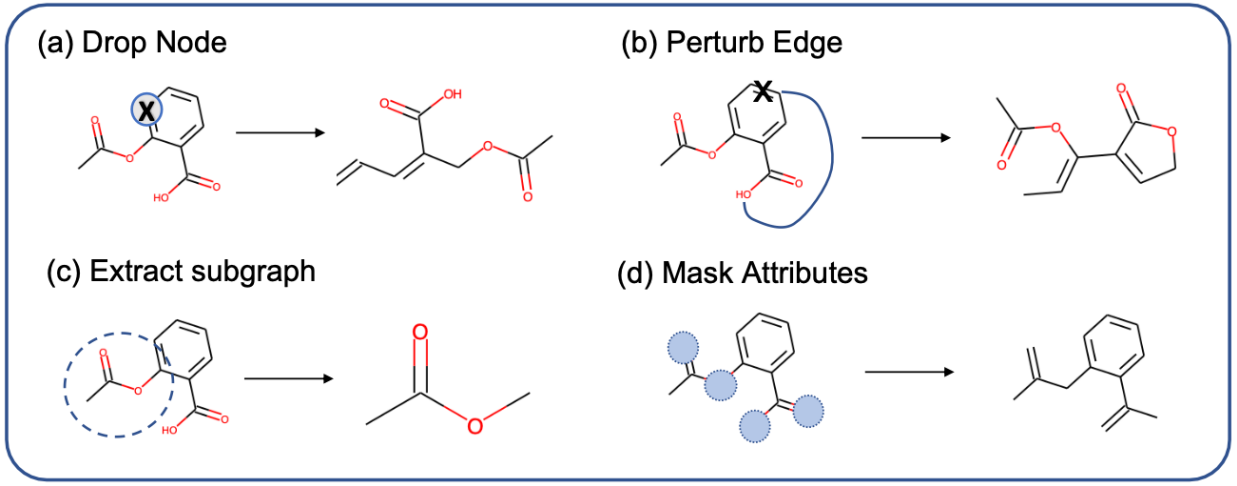
Augmented graphs

Pre-Training

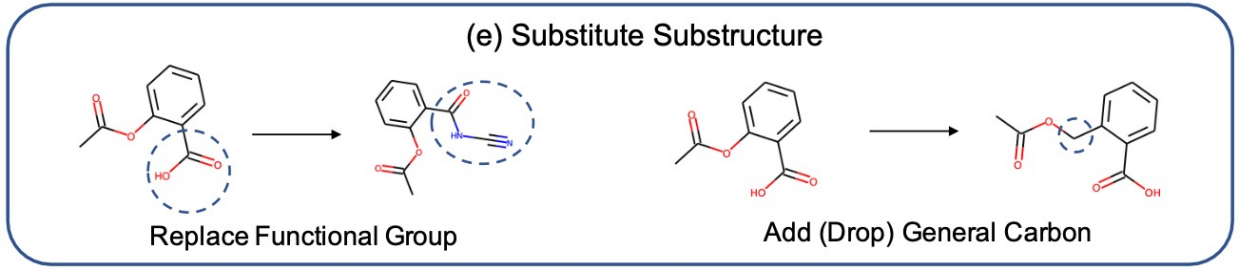


MOCL (KDD'21)

Pre-Training



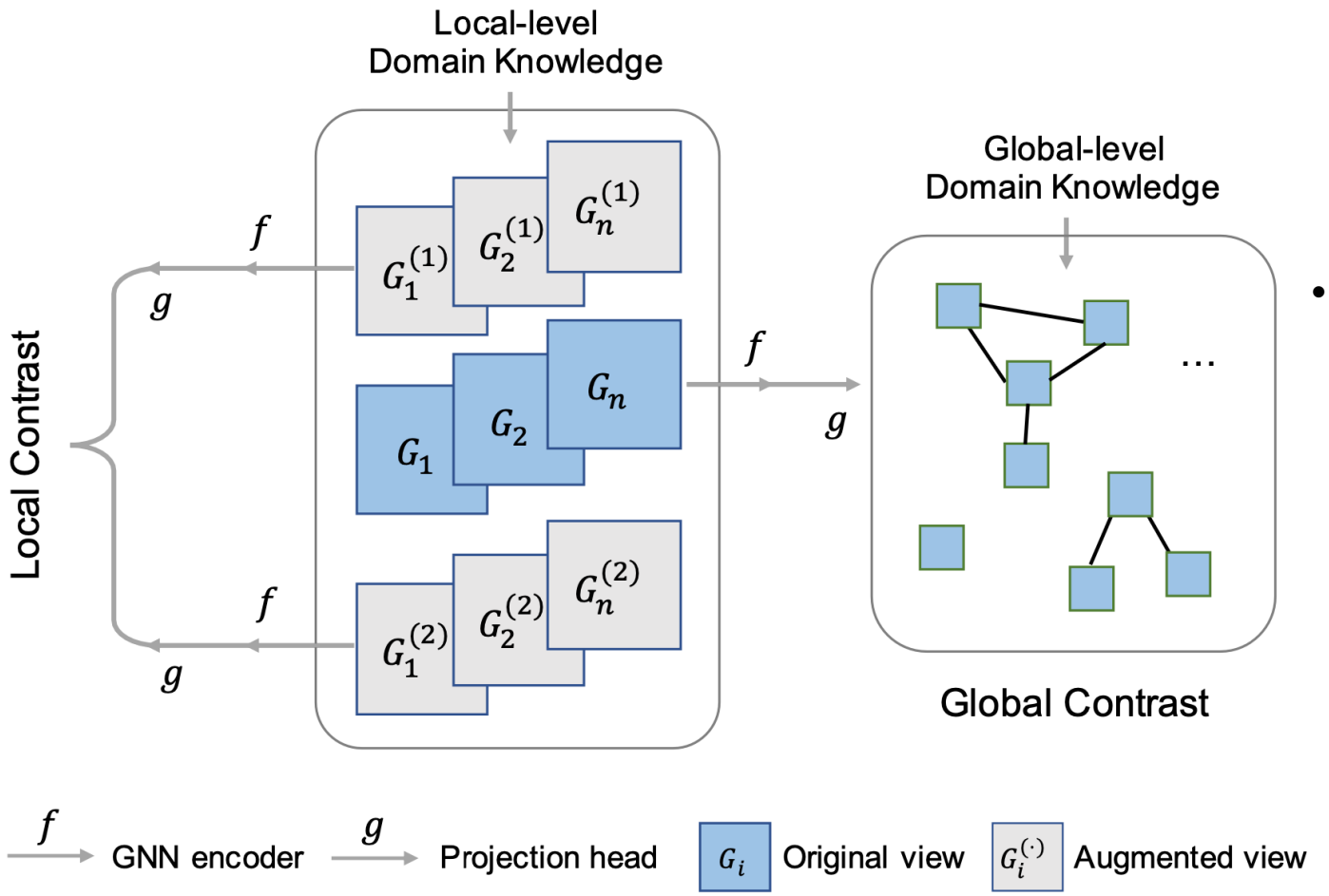
- Conventional augmentations may alter the graph semantics.



- Proposed augmentation in which valid substructures are replaced by bioisosteres that share similar properties.

MOCL (KDD'21)

Pre-Training

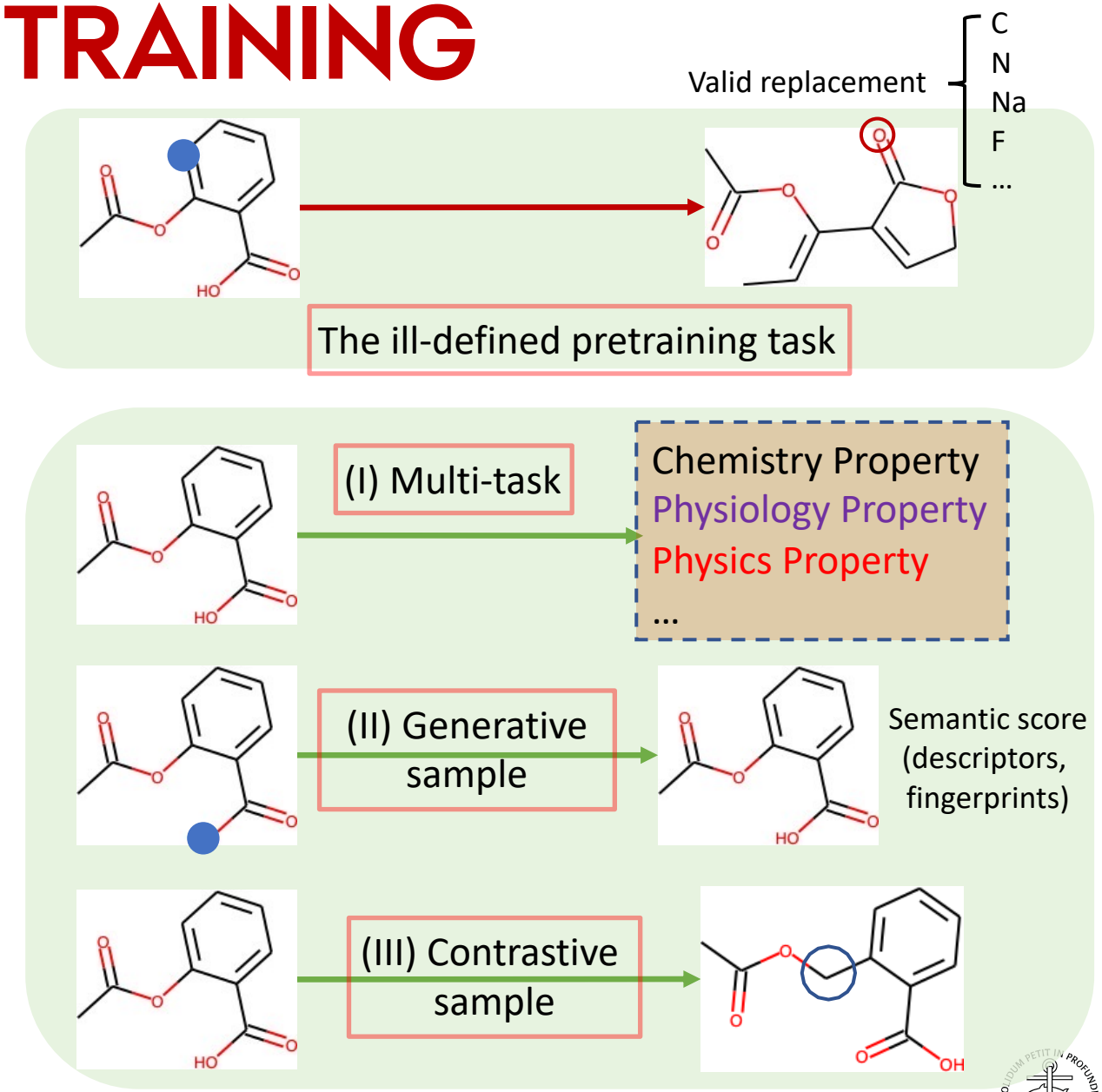


- Two augmented views are generated from local-level domain knowledge. Then, together with the original view (blue), they are fed into the GNN encoder and projection head to get global-level domain knowledge.

KNOWLEDGE IN PRE-TRAINING

- Classic pre-training strategies may damage drug discovery tasks
  - Invalid masking
  - Invalid prediction target

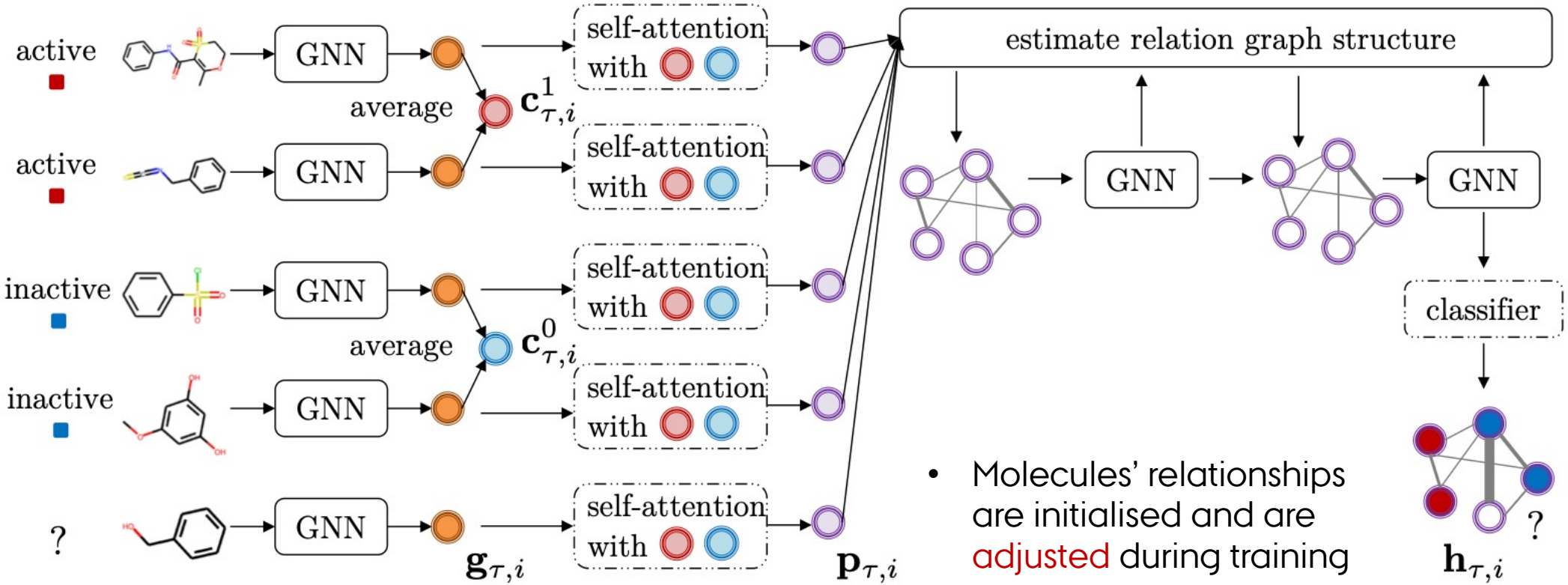
- **Knowledge-transfer**
  - Transferring external knowledge to construct pretraining objectives for GML models
- **Generative-sample**
  - Re-generating molecular graphs with the same biomedical semantics
- **Contrastive-sample**
  - Defining contrastive samples based on biomedical knowledge



INCORPORATING KNOWLEDGE IN TRAINING

PAR (NEURIPS'21)

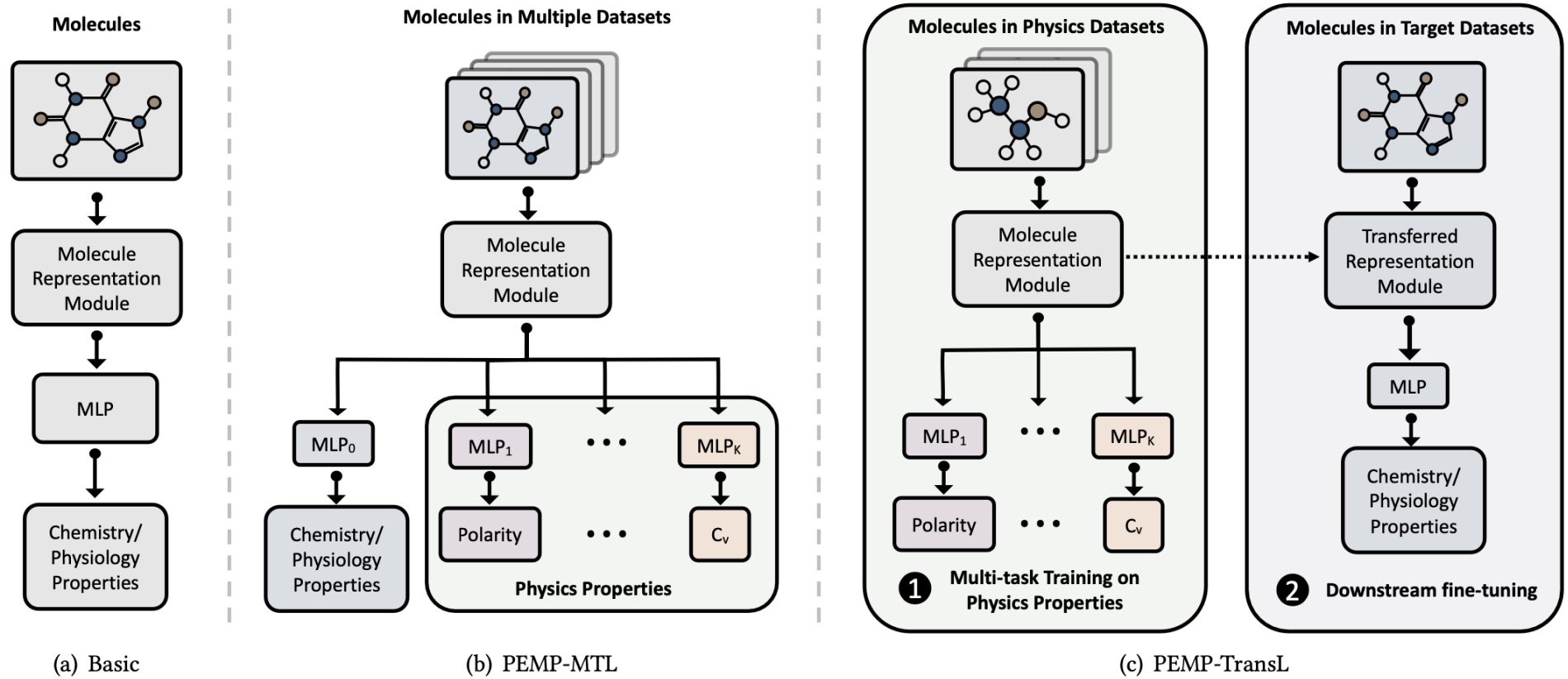
Training



- Molecules' relationships are initialised and are adjusted during training

PEMP (CIKM'22)

Training



(a) Basic (b) PEMP-MTL (c) PEMP-TransL

- Physics properties can be used as training objectives to optimise the embedding during training.

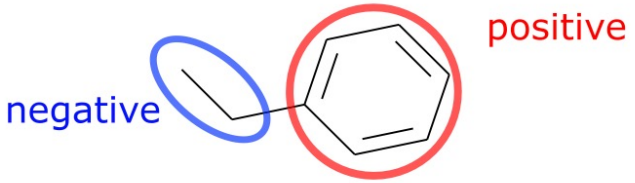
KEMPNN (ACS OMEGA'21)

Training

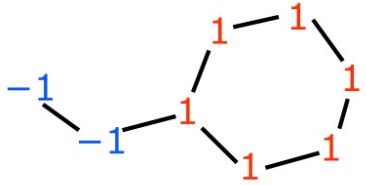
Knowledge annotation by human

- annotate by making substructure based rule
- annotate manually one-by-one

Substructure	Effect on property
	positive
C-C	negative
C=C	positive



Knowledge representation



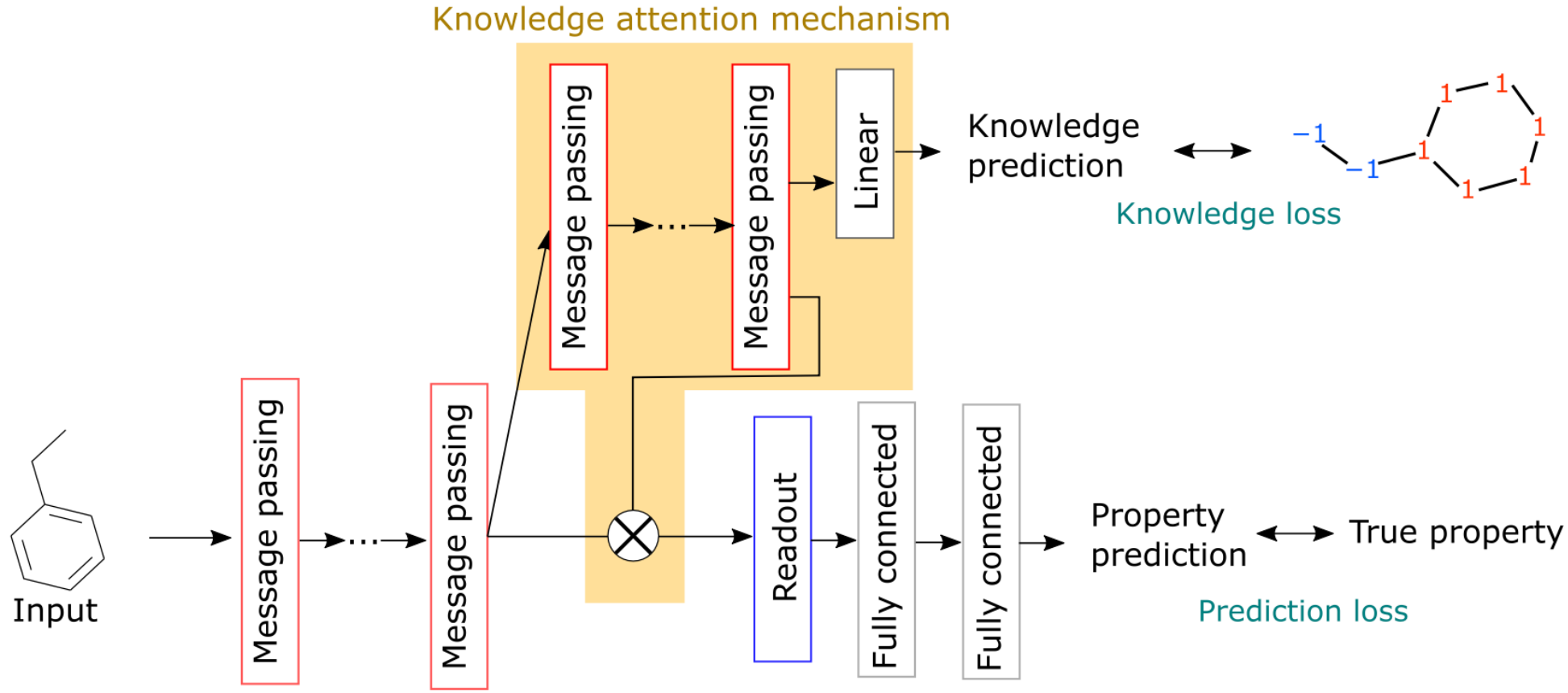
values on each atom (node)

1 → positive effect on property
−1 → negative effect on property
0 → no effect on property

- Knowledge can directly ameliorate GML models within the information propagation process

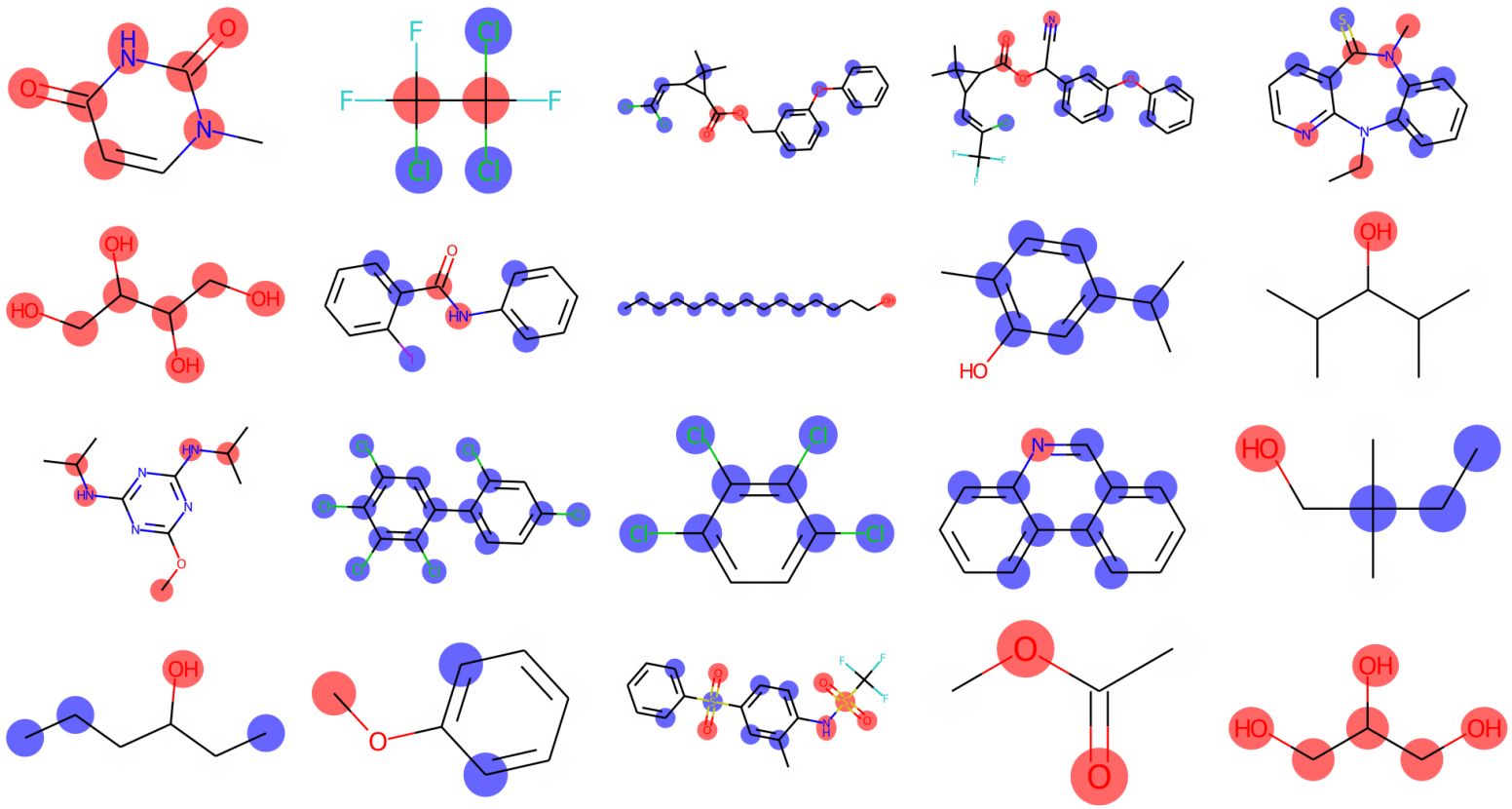
KEMPNN (ACS OMEGA'21)

Training



KEMPNN (ACS OMEGA'21)

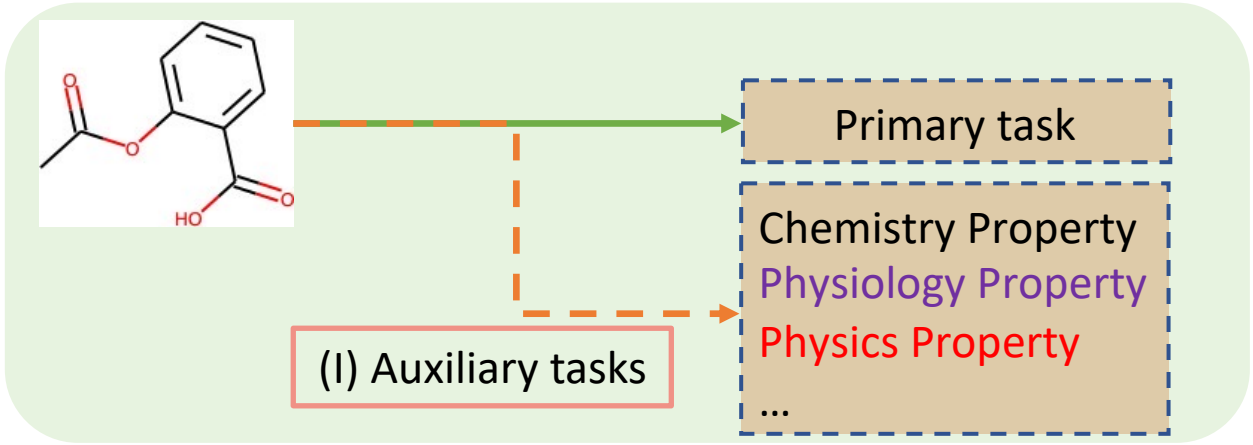
Training



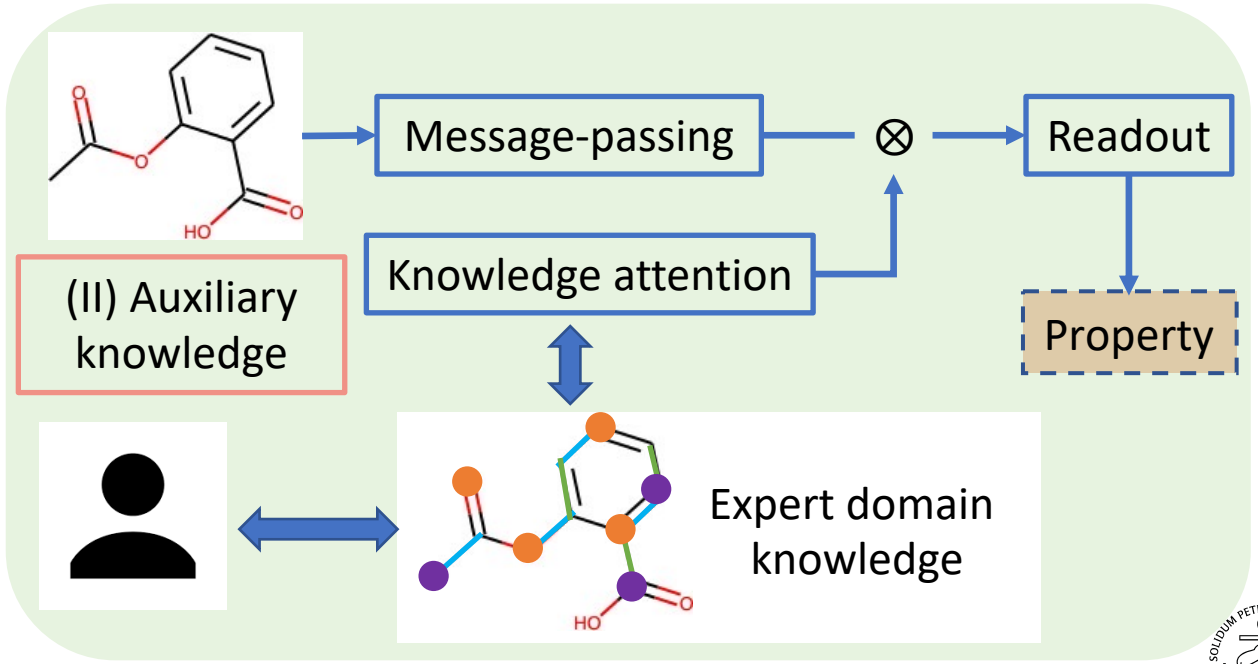
Visualisation of important compounds

KNOWLEDGE IN TRAINING

- **Auxiliary task-enhanced training**
  - Using relevant labelling information of target entities as additional training signals in addition to the primary downstream task labelling data.



- **Auxiliary knowledge-enhanced training**
  - Leveraging domain knowledge to adjust the internal processes of GML models to reduce the dependency on training data



INCORPORATING KNOWLEDGE IN *INTERPRETABILITY*

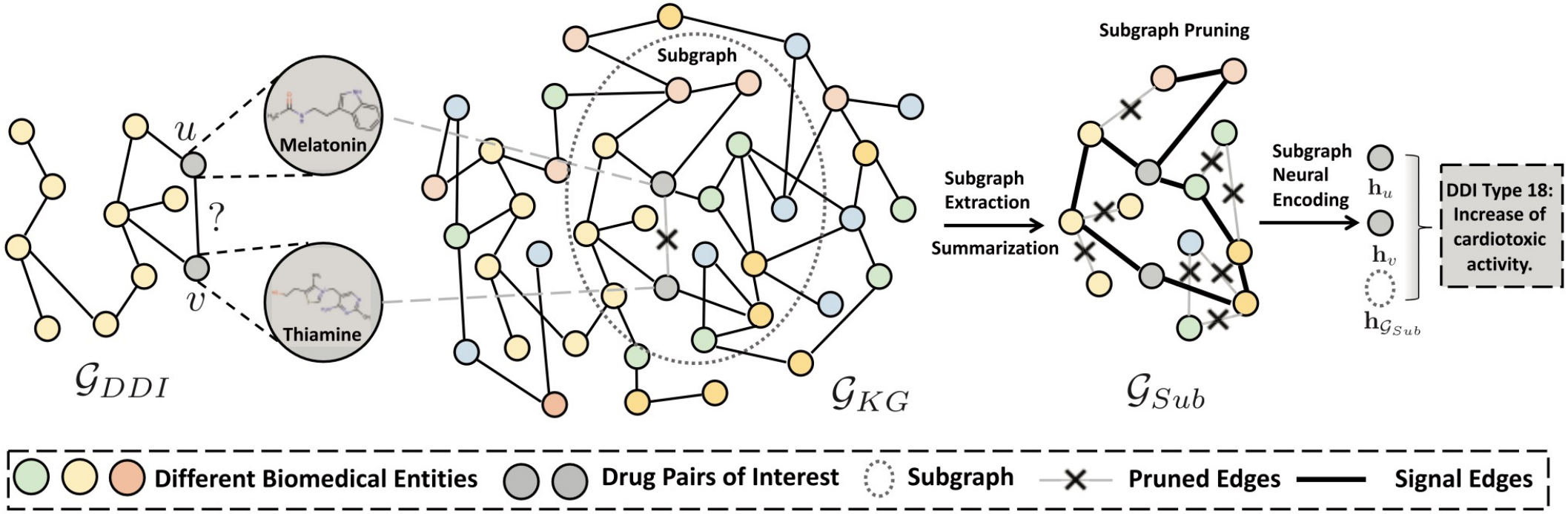
AARHUS UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

KDD'23 TUTORIAL
9 AUGUST 2023

ZHIQIANG ZHONG AND DAVIDE MOTTIN
DATA-INTENSIVE SYSTEMS GROUP

SUMGNN (BIOINFORM.'21)

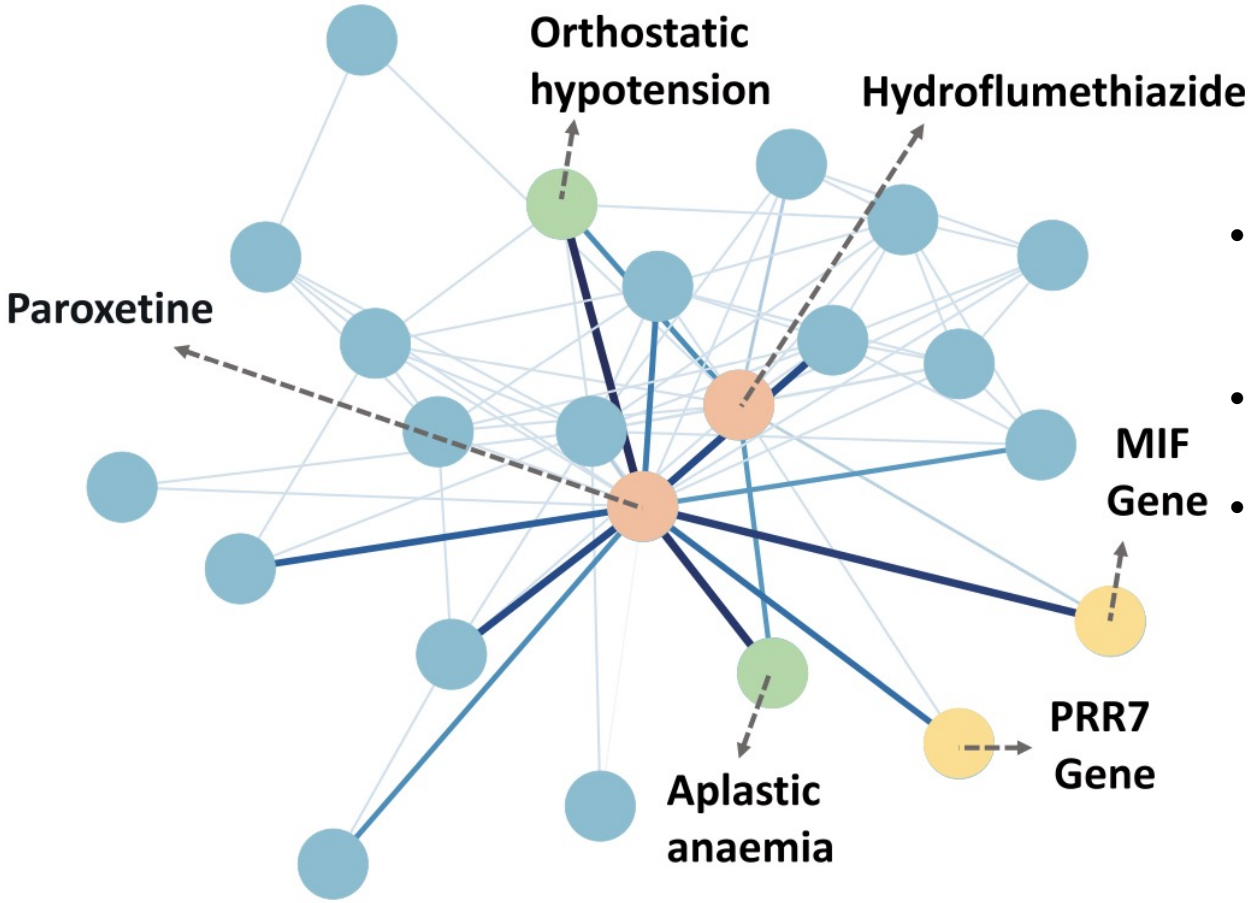
Interpretability



- Important subgraphs are extracted to explain the relationship between the two drugs

SUMGNN (BIOINFORM.'21)

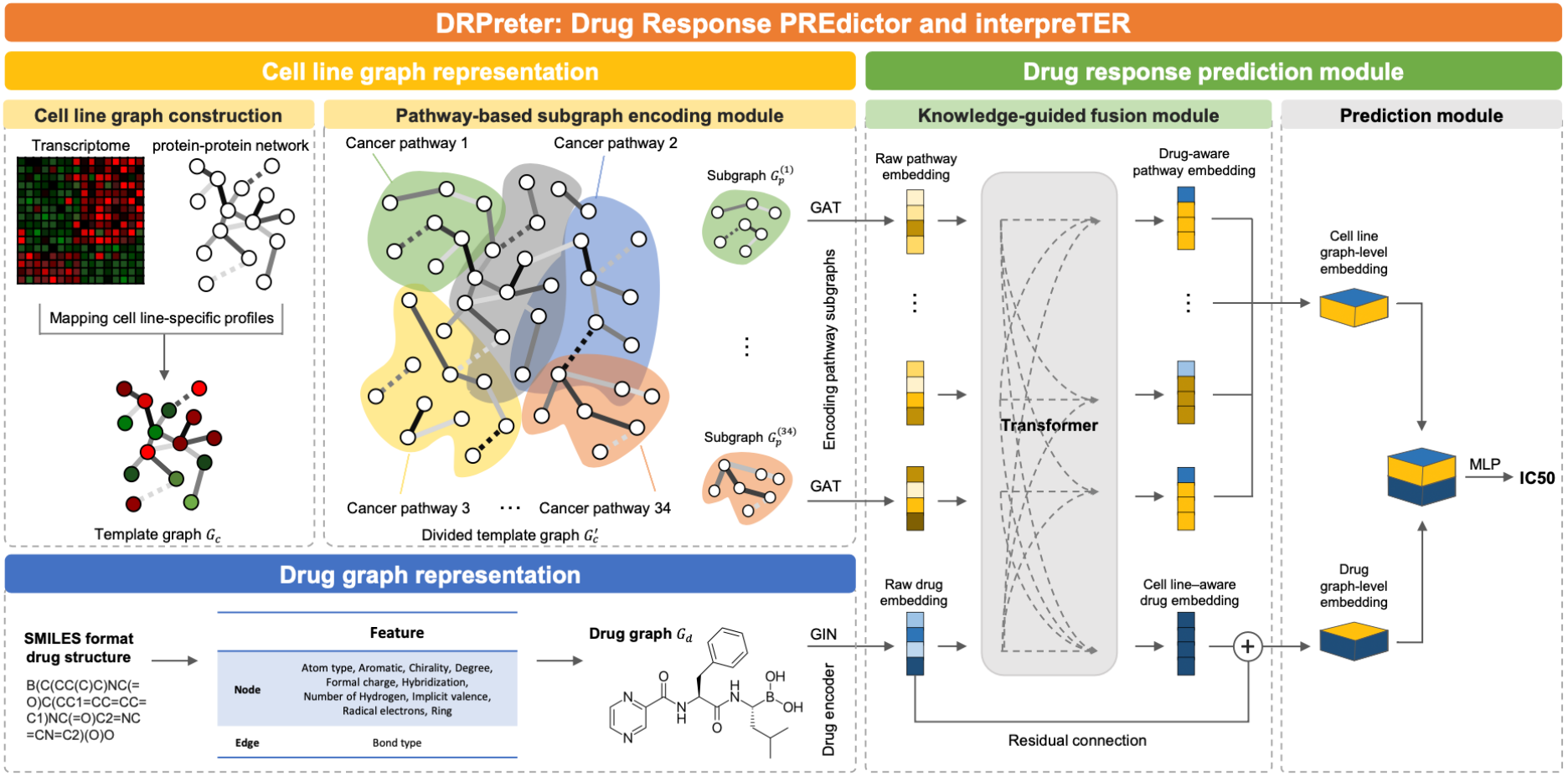
Interpretability



- SumGNN generates a short reasoning path to provide clues for understanding drug interactions.
- The shade of colour indicates the strength of attention weight.
- Low-weight edges in the extracted subgraph are pruned by SumGNN, and SumGNN focuses on a sparse set of signal edges and nodes.

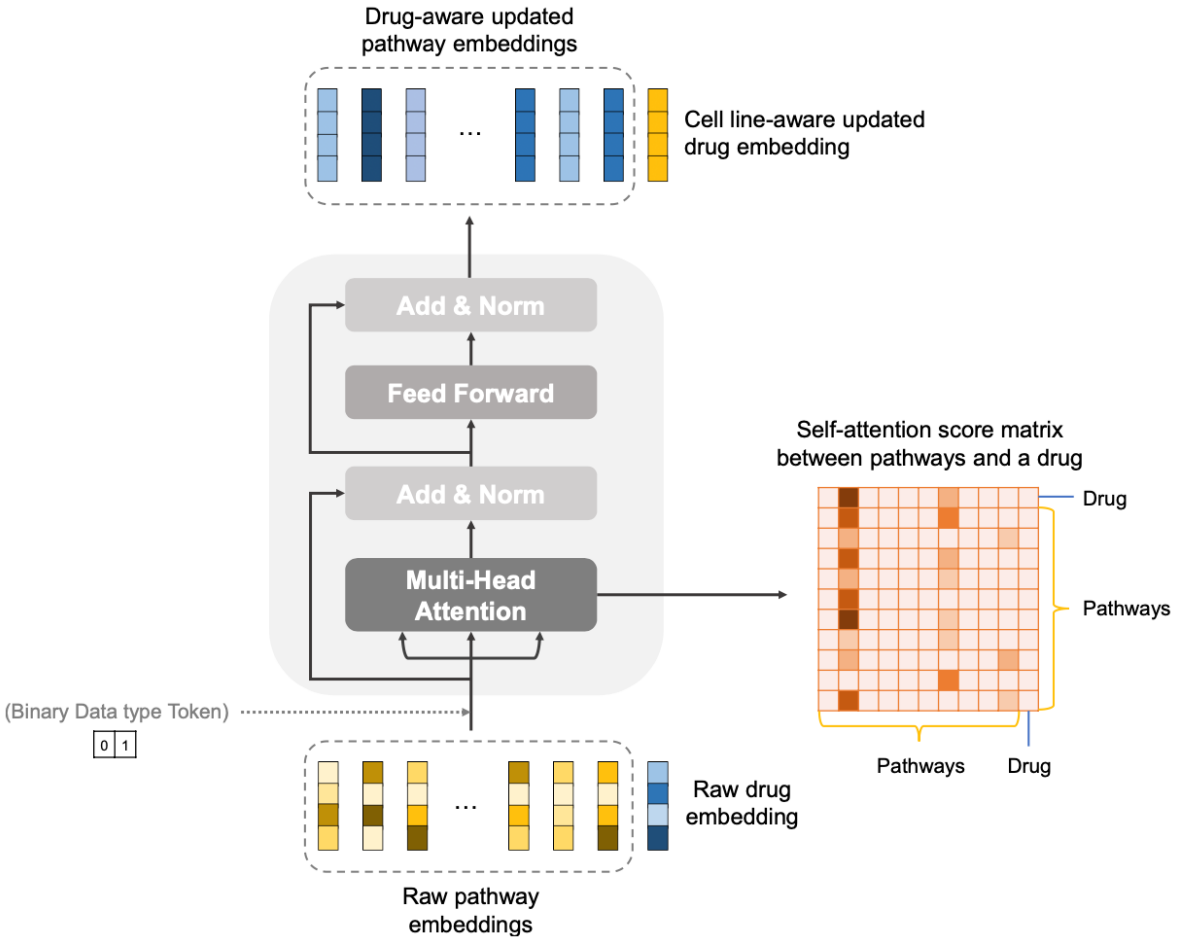
DRPRETER (NT. J. MOL. SCI.'22)

Interpretability



DRPRETER (NT. J. MOL. SCI.'22)

Interpretability



- Self-attention mechanism is applied to identify crucial pathways

DRPRETER (NT. J. MOL. SCI.'22)

Interpretability

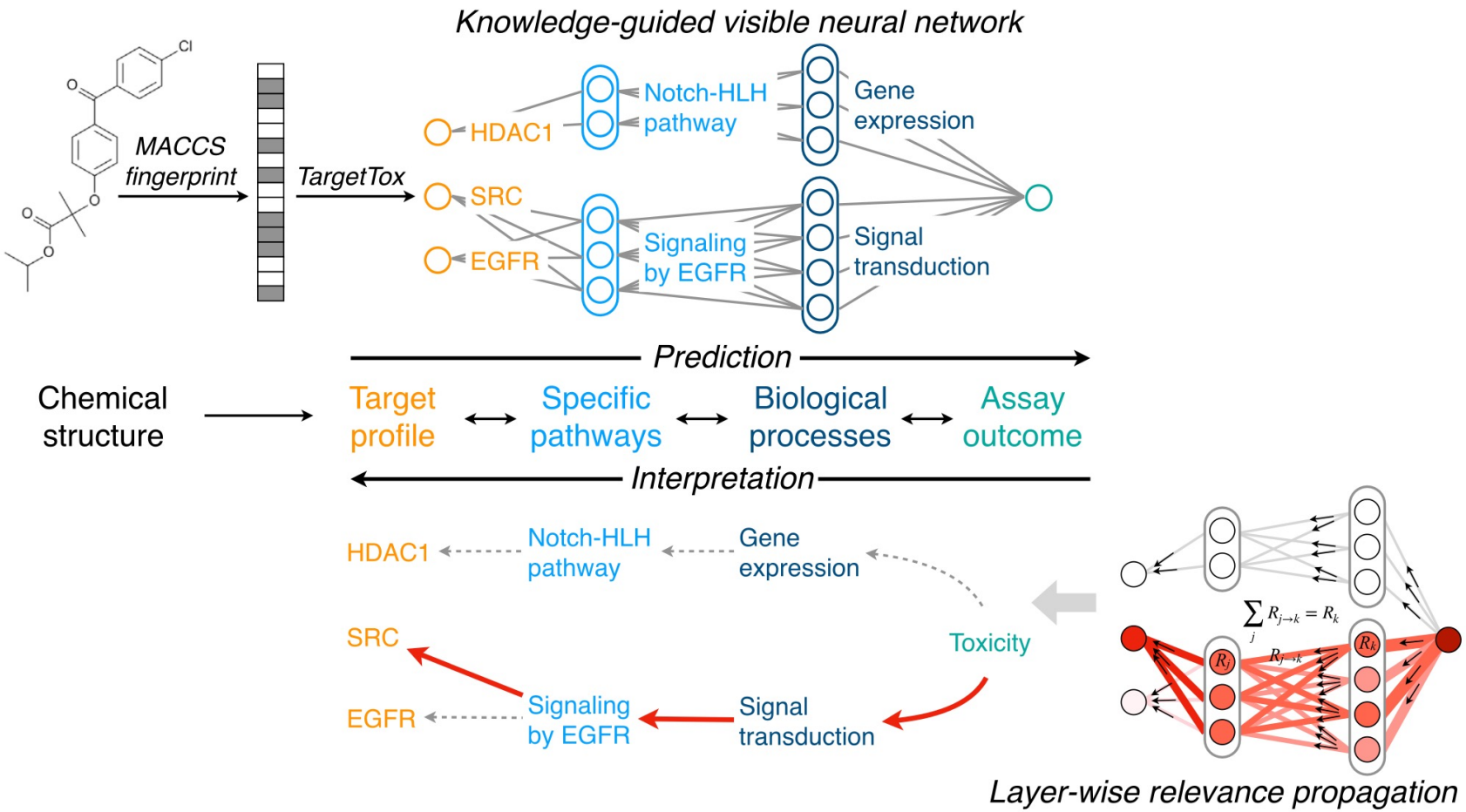
Drug	Cell Line	Disease	Top 5 Significant Genes	ln(IC50)	
				True	Predicted
Afatinib	GMS-10	Glioblastoma	*ACTR3B, PRR5, PRKCZ,* ***ERBB2****, LTBR*	0.5372	0.5324
Vinblastine	NCI-H1792	NSCLC	*CYP7A1, GTF2H2, DVL2, RAB5B,* ***TP53***	−5.9258	−5.27633
Docetaxel	PANC0327	Pancreatic cancer	***CLDN18****, SOX17, FGF19, WNT7A, CDH5*	−3.7668	−3.8204
Rapamycin	IGR1	Melanoma	*TYRP1, DCT, TYR, FRZB,* ***CDK2***	−1.6747	−1.7651
Bortezomib	EBC-1	Lung squamous cell carcinoma Derived from metastatic site: Skin	*SHC4, TNR, IL17RA,* ***MAPK12****, SMURF1*	−5.7714	−6.0714

Genes in bold are direct targets of drugs, are involved in target pathways, or are biomarkers of disease.

Gradient-based gene importance analysis.

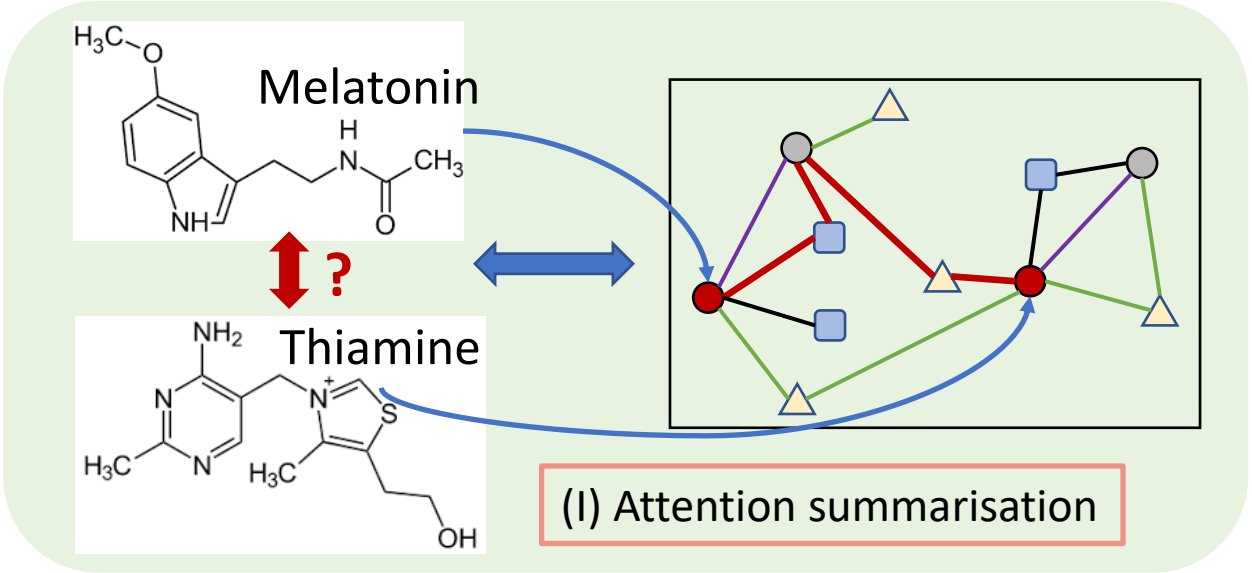
DTOX (PATTERNS'22)

Interpretability

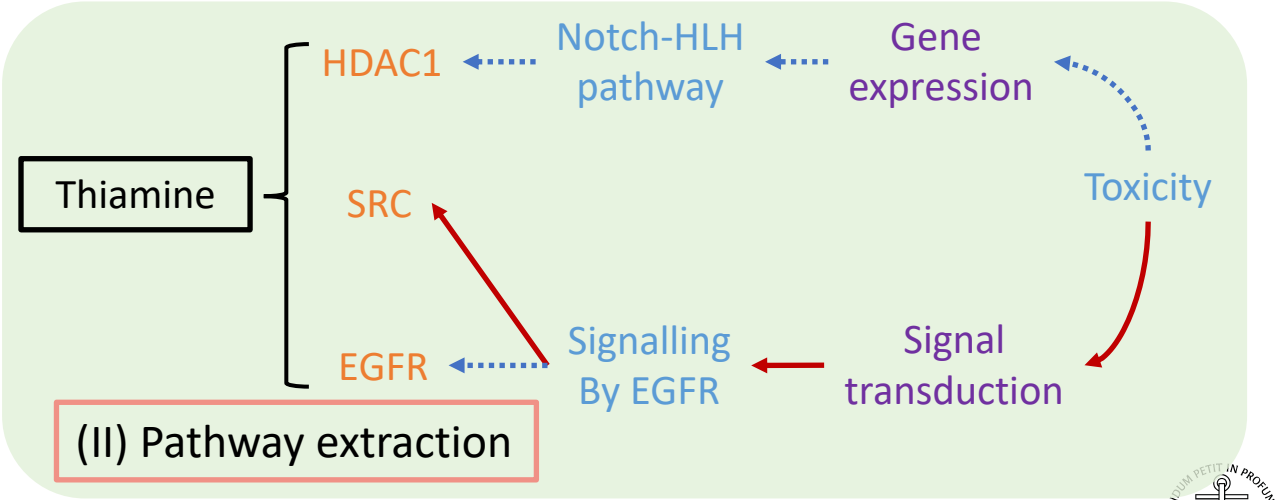


KNOWLEDGE IN INTERPRETABILITY

- **Attention summarisation**
  - Machine-readable domain knowledge represented in KGs can be highlighted for downstream applications.



- **Pathway extraction**
  - Adaptively infer pathways pertaining to target biomedical entities from pathway datasets and highlight the key pathways for the explanation.



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