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

Zhiqiang Zhong & Davide Mottin

Aarhus University

PRESENTERS

Zhiqiang Zhong
Postdoc
Aarhus University

Davide Mottin
Asst. Prof.
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

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

SIGNIFICANCE AND CHALLENGES OF DRUG DISCOVERY

DRUG DISCOVERY IS LONG AND EXPENSIVE

- It usually takes 10-15 years and costs around 2 billion US dollars

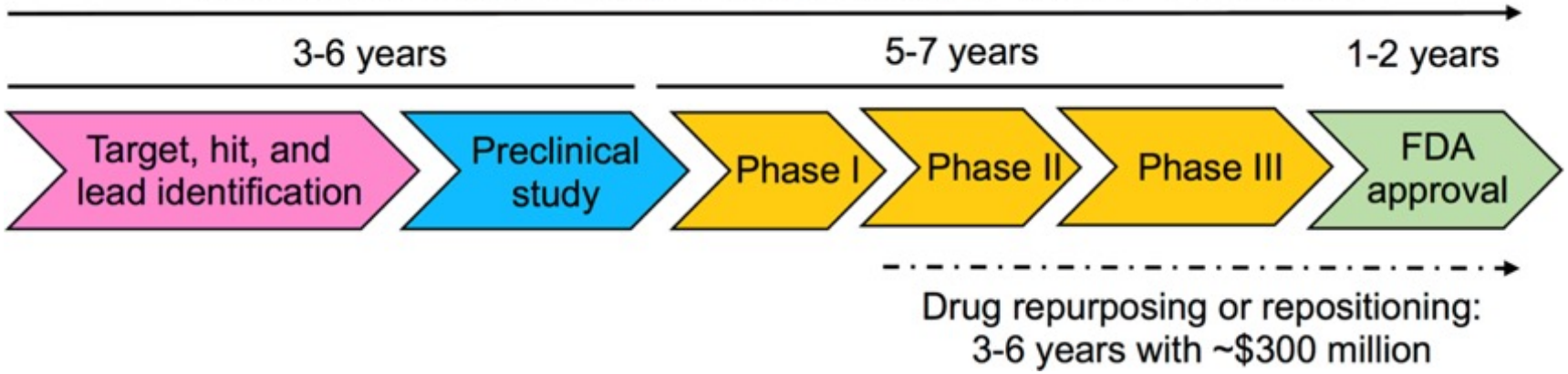
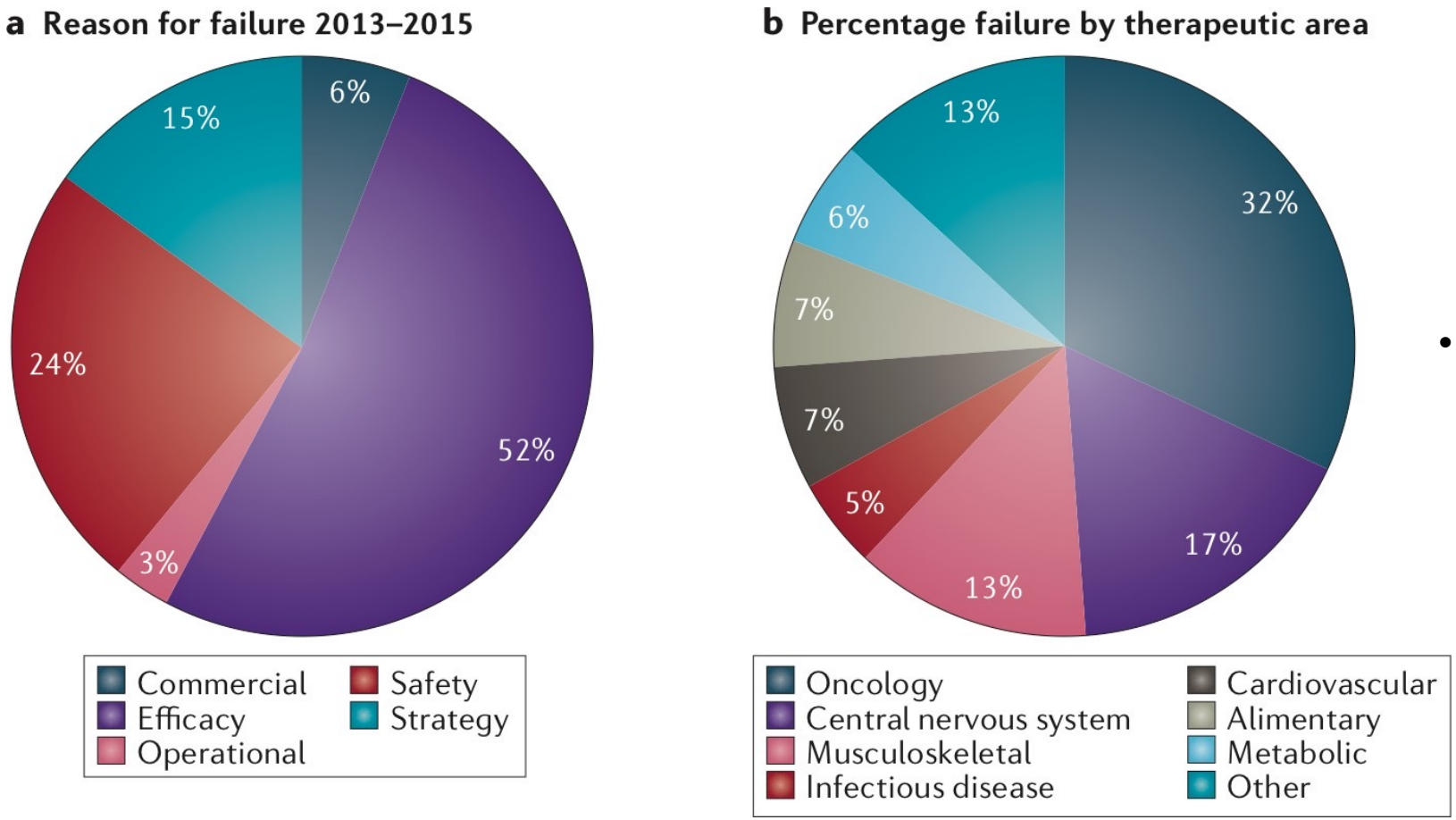


Figure Source: *Cheng F.*, Methods Mol. Biol., 2019

REASONS FOR DRUG DISCOVERY FAILURES

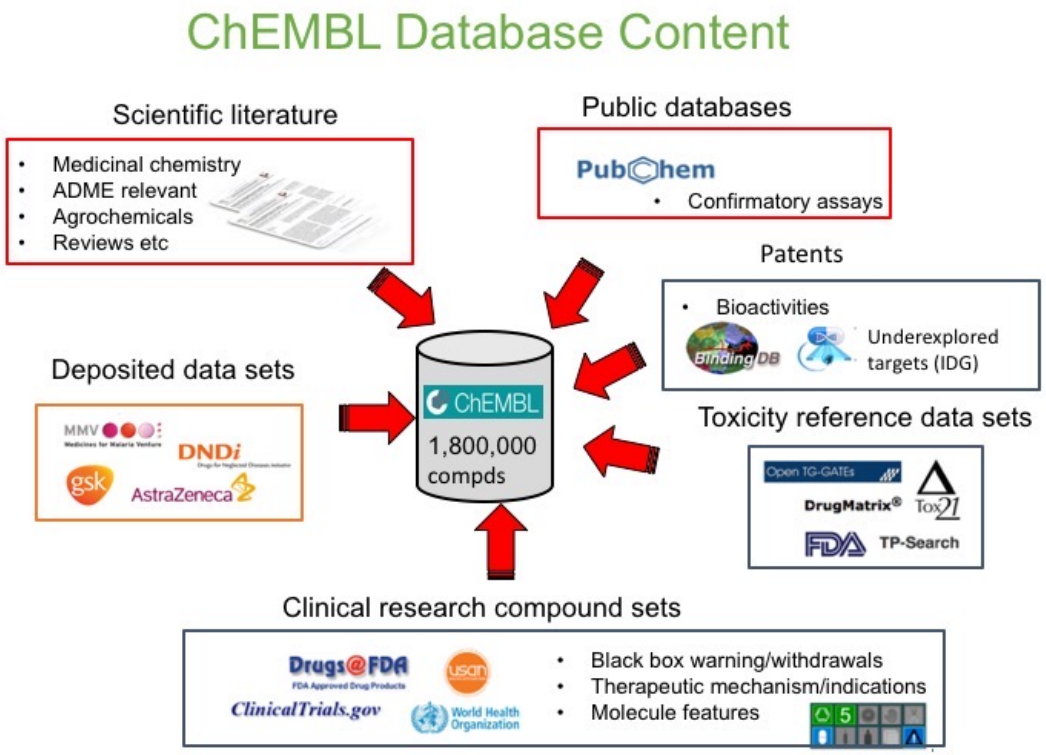


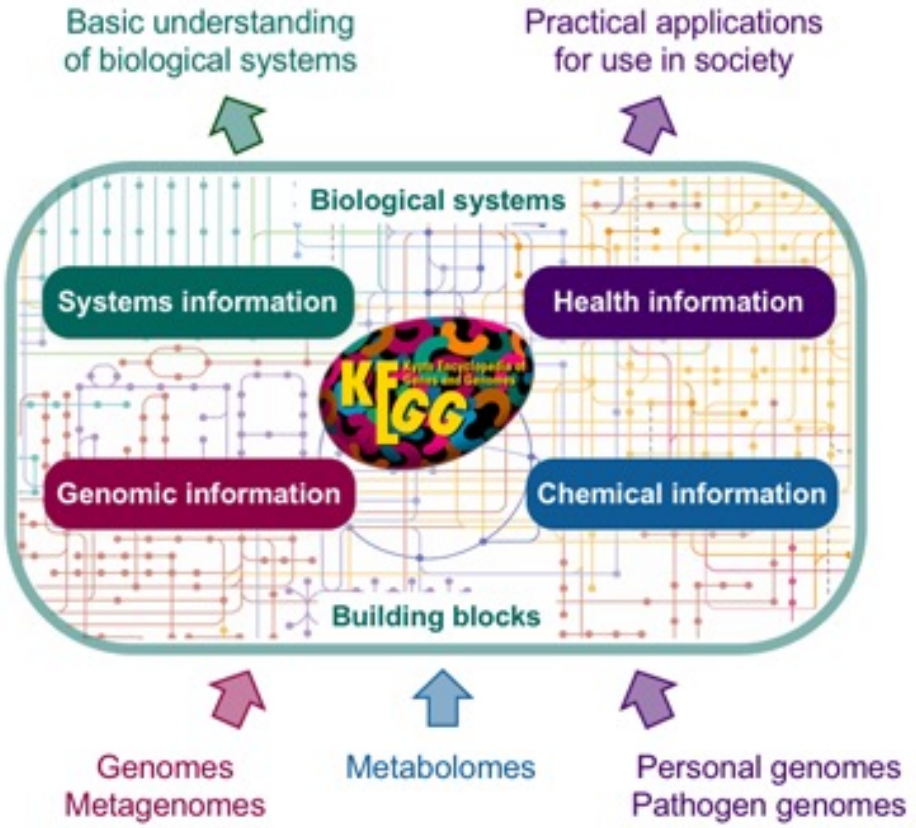
- AI techniques can potentially address major failures.

Figure Source: *Harrison RK.*, Nat Rev Drug Discovery 2020

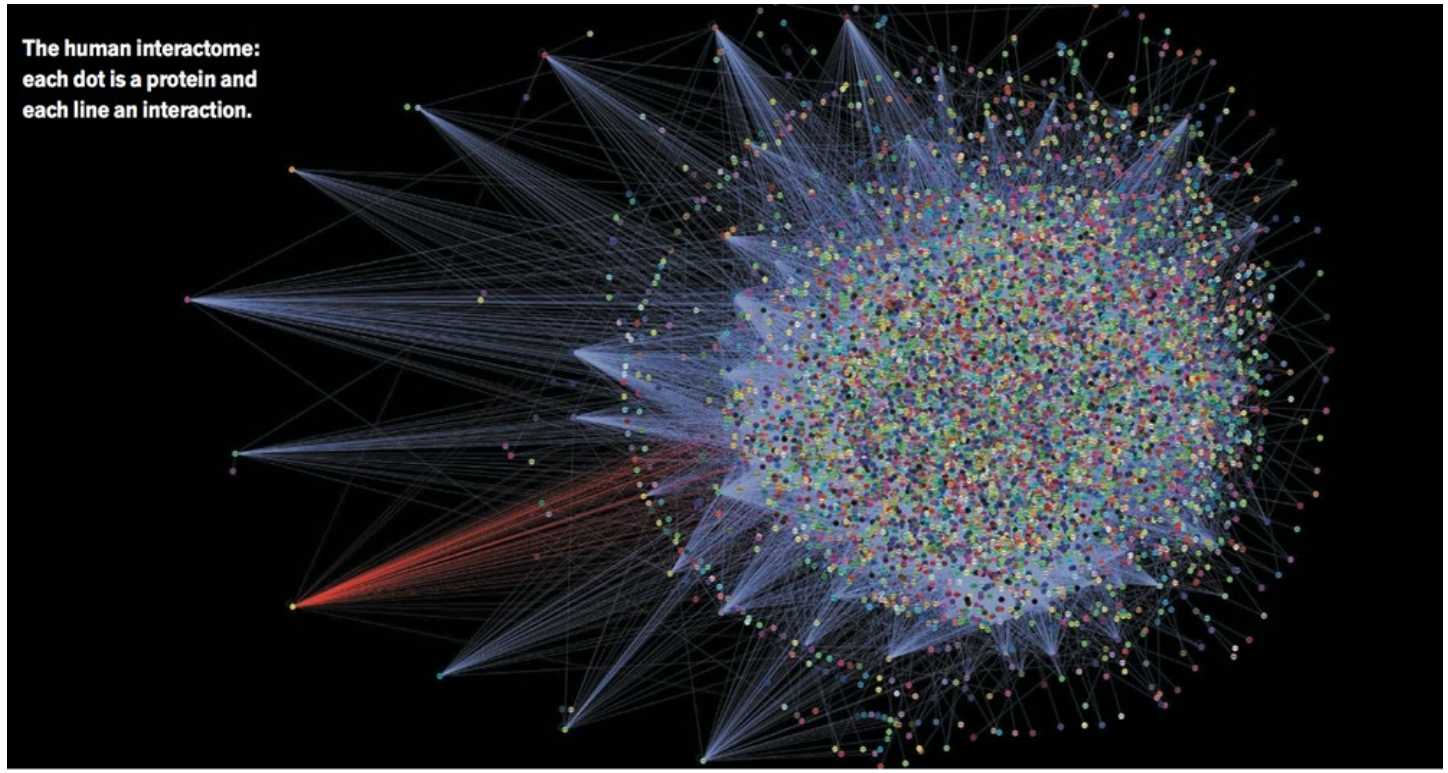
BIG OPPORTUNITY FOR AI

- A huge amount of data is generated in the biomedical domain





BIG OPPORTUNITY FOR AI



Fessenden et al., Nature 2017

Luck et al., Nature 2020

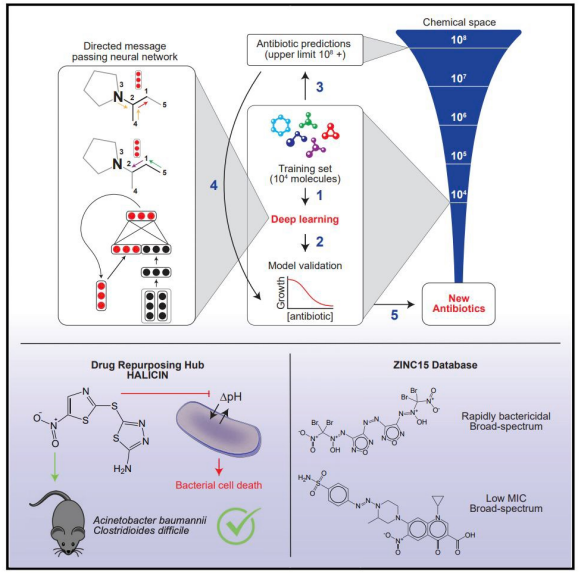
- ~13 major organ systems
- ~200-300 cell types
- ~10-50 trillion total human cells
- ~20,000 protein-coding genes
- ~3,000 metabolites
- ~300 different post-translational modifications
- ~600K binary interactions of proteins
- Over 2M nodes among # DNA, RNA, protein and variants

PROMISING AI IN DRUG DISCOVERY

Cell

A Deep Learning Approach to Antibiotic Discovery

Graphical Abstract



Authors

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, ..., Tommi S. Jaakkola, Regina Barzilay, James J. Collins

Correspondence

regina@csail.mit.edu (R.B.), jimjc@mit.edu (J.J.C.)

In Brief

A trained deep neural network predicts antibiotic activity in molecules that are structurally different from known antibiotics, among which Halicin exhibits efficacy against broad-spectrum bacterial infections in mice.

nature

Subscribe

NEWS · 20 FEBRUARY 2020

Powerful antibiotics discovered using AI

Machine learning spots molecules that work even against 'untreatable' strains of bacteria.

nature

Explore content · About the journal · Publish with us

nature › articles › article

Article | Open Access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

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 DRUG DISCOVERY

AARHUS
UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

KDD'23 TUTORIAL
9 AUGUST 2023

ZHIQIANG ZHONG AND DAVIDE MOTTIN
DATA-INTENSIVE SYSTEMS GROUP

DRUG DISCOVERY PRINCIPLES

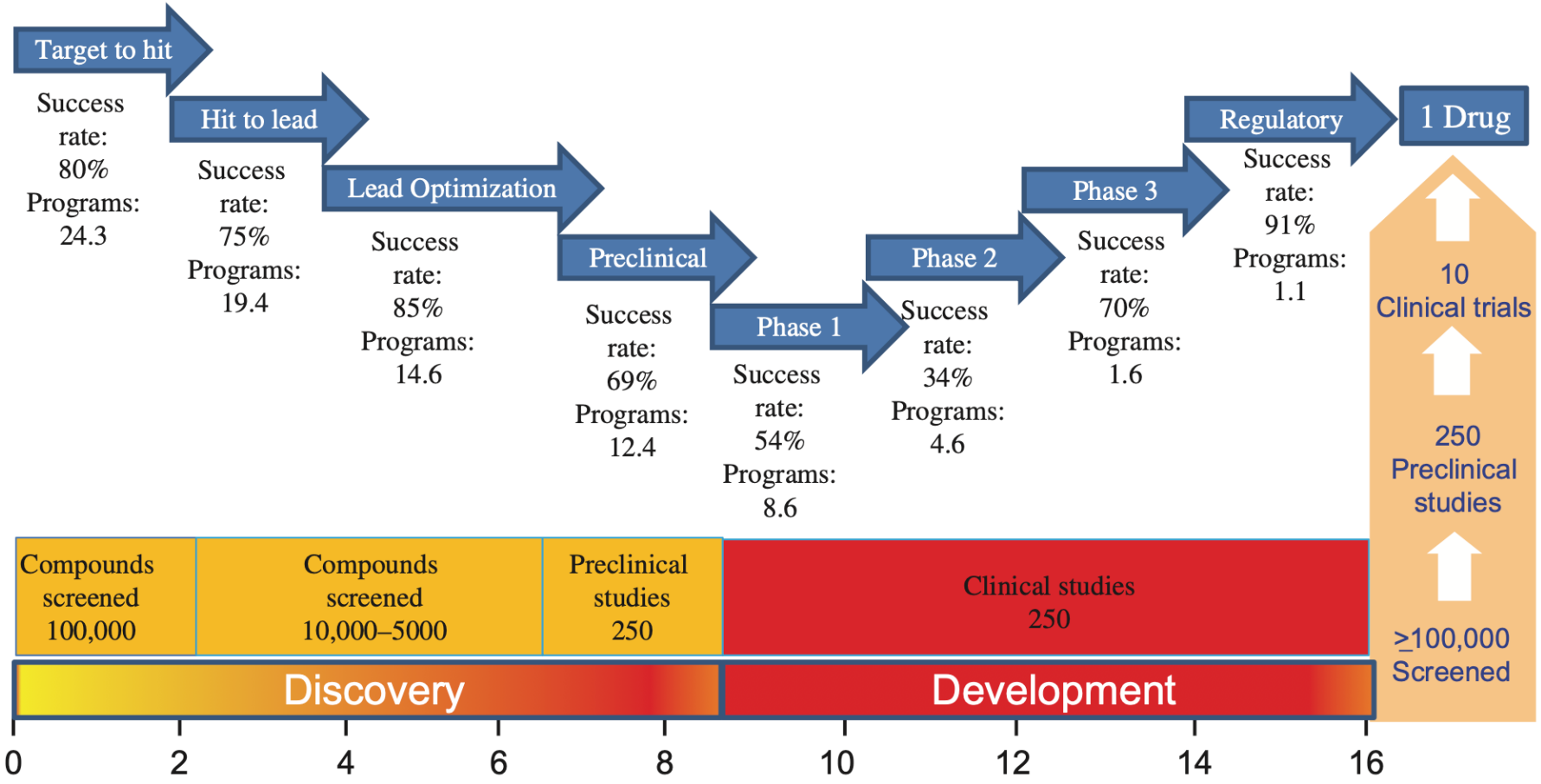


Figure Source: *Benjamin E. Blass*, Basic Principles of Drug Discovery and Development, 2022

TECHNOLOGIES FOR DRUG DISCOVERY

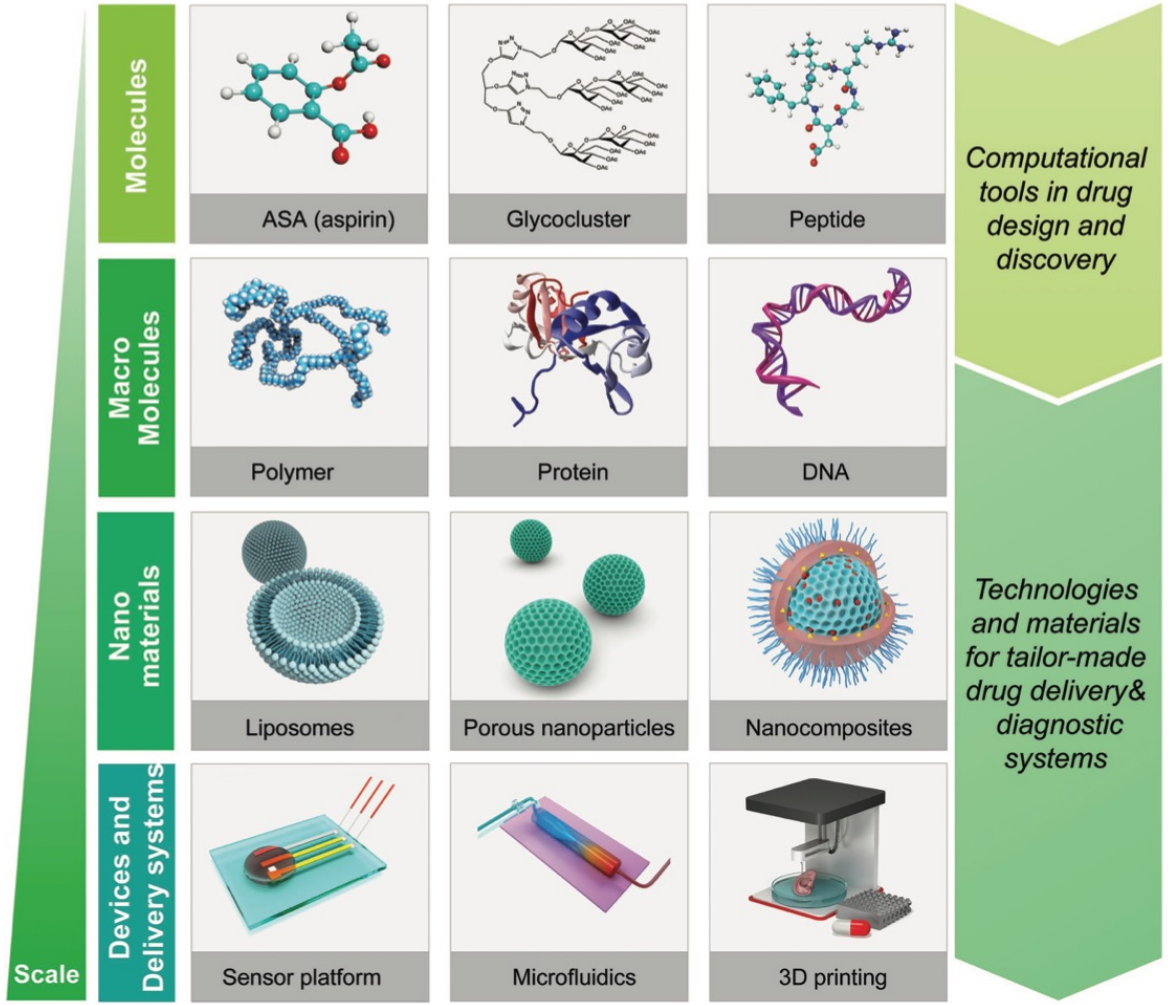


Figure Source: *Sahlgren et al.,* Advanced Healthcare Materials, 2017

EXAMPLE TOP/DOWN PROTEOMICS

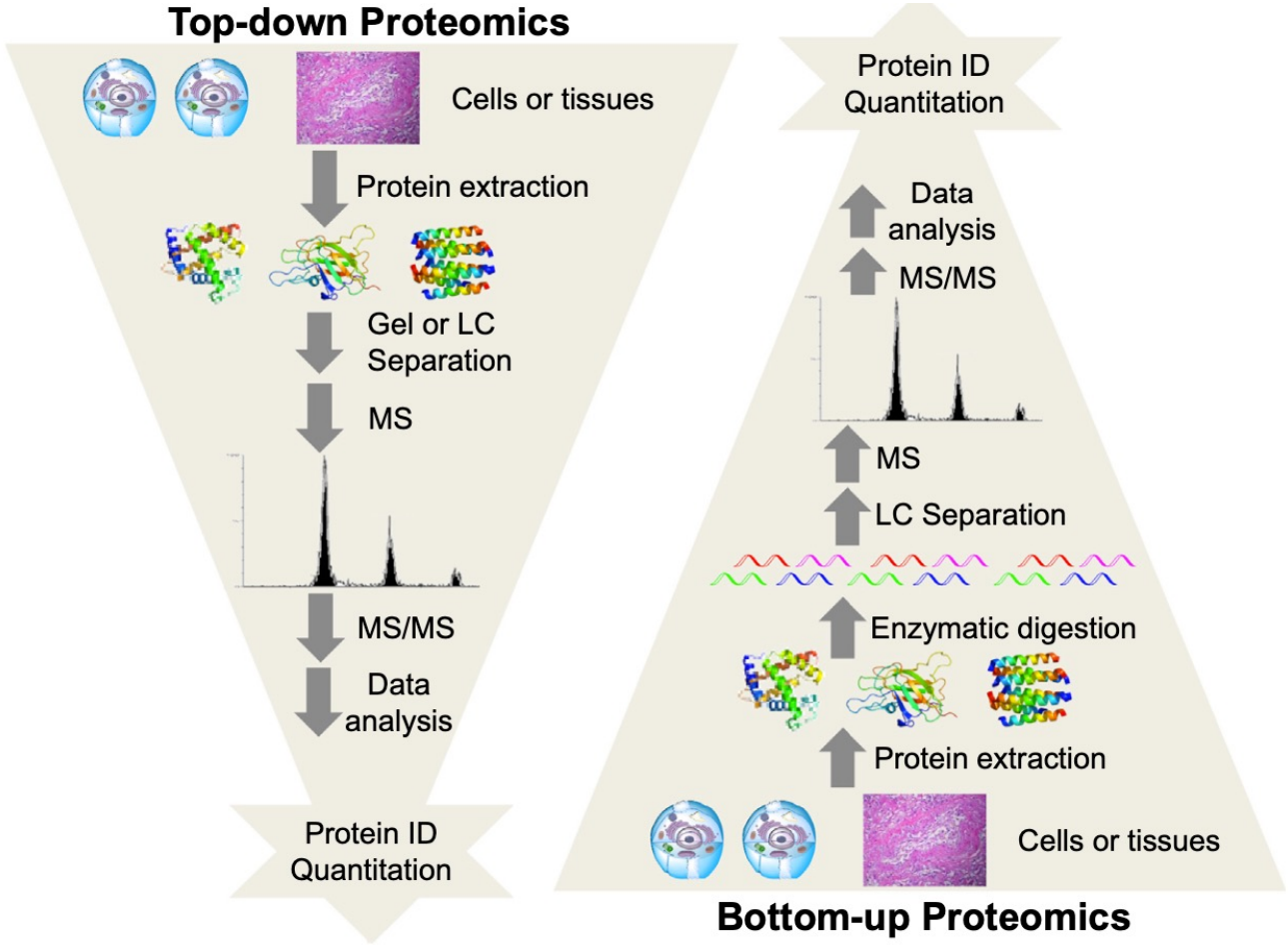


Figure Source: *Benjamin E. Blass*, Basic Principles of Drug Discovery and Development 2022

UBIQUITOUS GRAPH-STRUCTURED DATA IN DRUG DISCOVERY

AARHUS UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

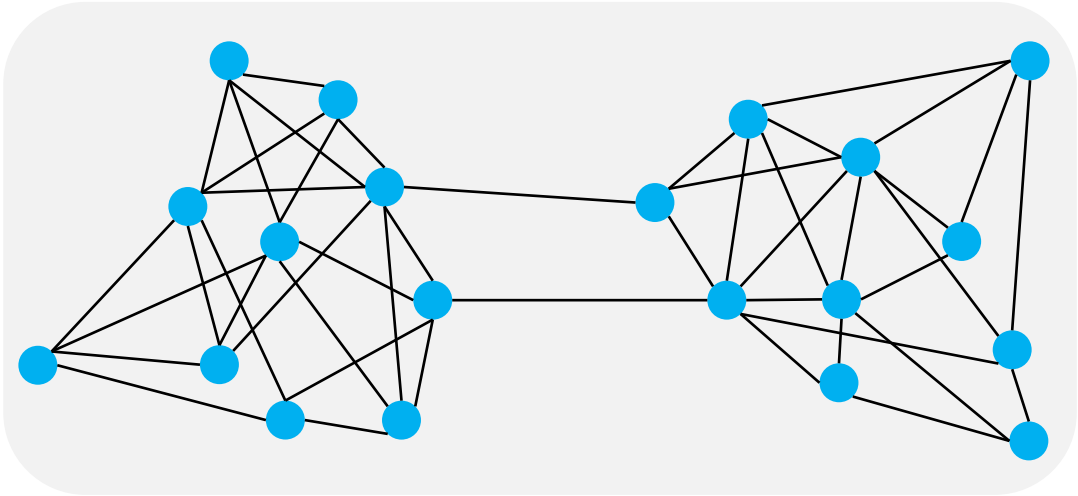
KDD'23 TUTORIAL
9 AUGUST 2023

ZHIQIANG ZHONG AND DAVIDE MOTTIN
DATA-INTENSIVE SYSTEMS GROUP

WHAT IS GRAPH?

Graph

$\mathcal{G} = (V, E, X)$



V: a set of nodes

E: a set of edges

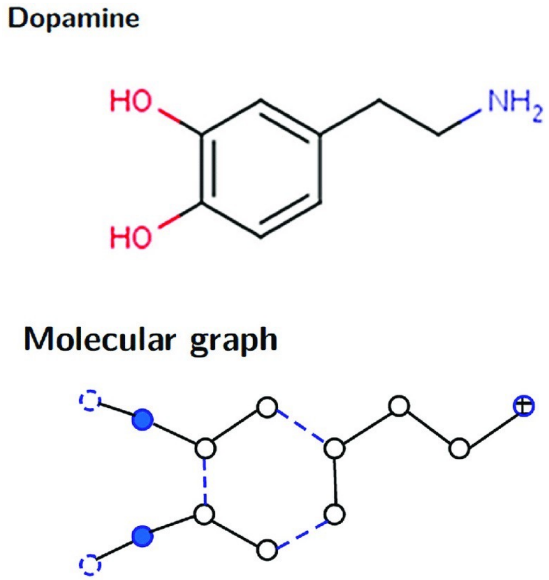
Node attribute matrix

$$\mathbf{X} = \begin{bmatrix} 0.4 & 4.4 & \cdots & 3 \\ 0.3 & 9.1 & \cdots & 6 \\ \vdots & \vdots & \ddots & \vdots \\ 0.1 & 0 & \cdots & 6 \\ 0.7 & 1.8 & \cdots & 1 \end{bmatrix}$$

Adjacency matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 & 1 \\ 0 & 1 & \cdots & 1 & 0 \end{bmatrix}$$

BIO-DATA ARE GRAPH-STRUCTURED



Small Molecules

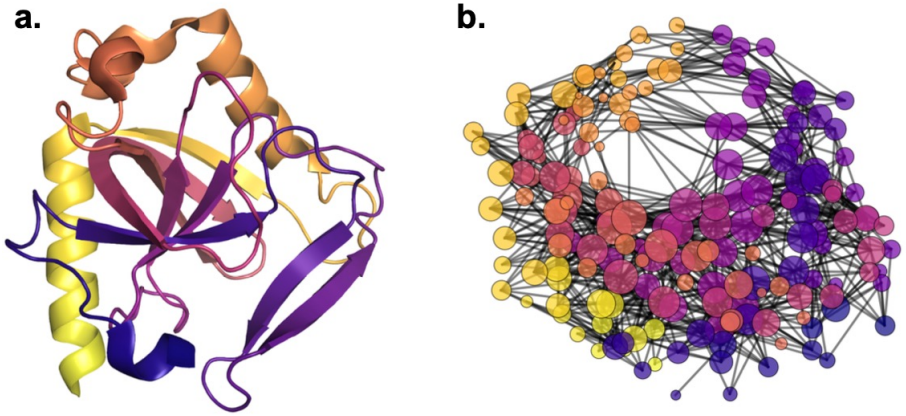
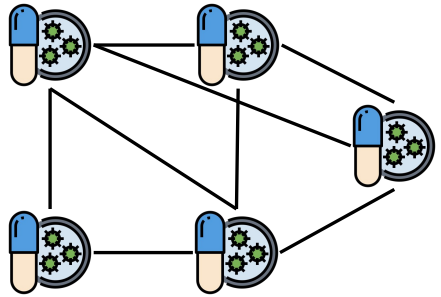


Figure 8: Illustration of **a.** a protein (PDB accession: 3EIY) and **b.** its graph representation derived based on intramolecular distance with cut-off threshold set at 10Å.

Proteins



Drug-Drug Interaction

Figure Source: *Gaudelet et al.*, Brief. Bioinformatics 2022

WHAT IS KNOWLEDGE DATABASE?

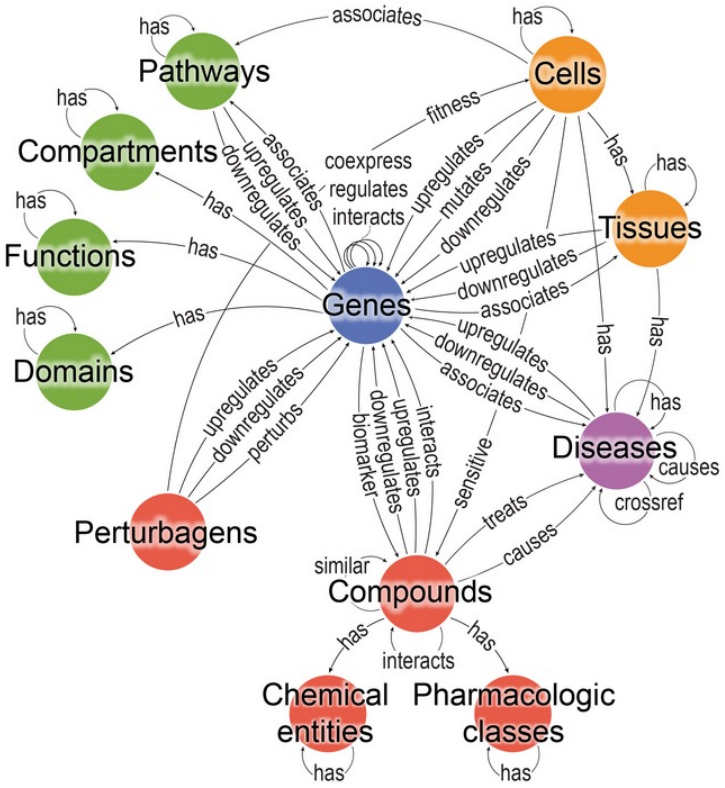
"**Knowledge** is the awareness of facts or as practical skills, the understanding of the world, and may also refer to familiarity with objects or situations" [1]

- A knowledge database $\mathcal{D}$ represents the identified knowledge in a well-organised and structured format.

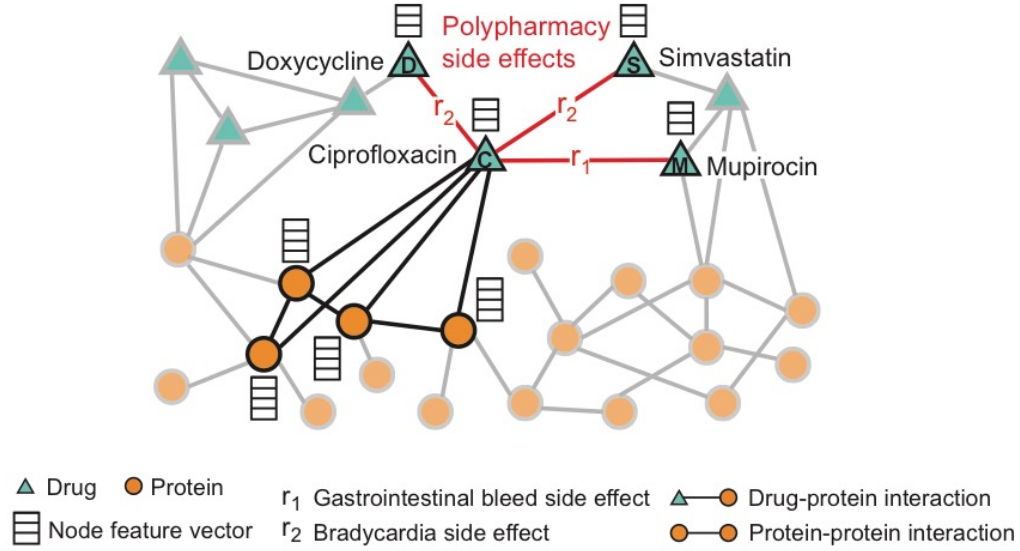
- A knowledge graph (KG) is a database represented as a directed heterogeneous graph $\mathcal{KG} = (\mathcal{V}, \mathcal{E}, \phi, \psi)$ with an entity type mapping function $\phi: \mathcal{V} \rightarrow Z^A$ and an edge type mapping function $\psi: \mathcal{E} \rightarrow Z^R$.

[1] https://dbpedia.org/page/Knowledge

GRAPH-STRUCTURED BIO-KNOWLEDGE



Bioteque

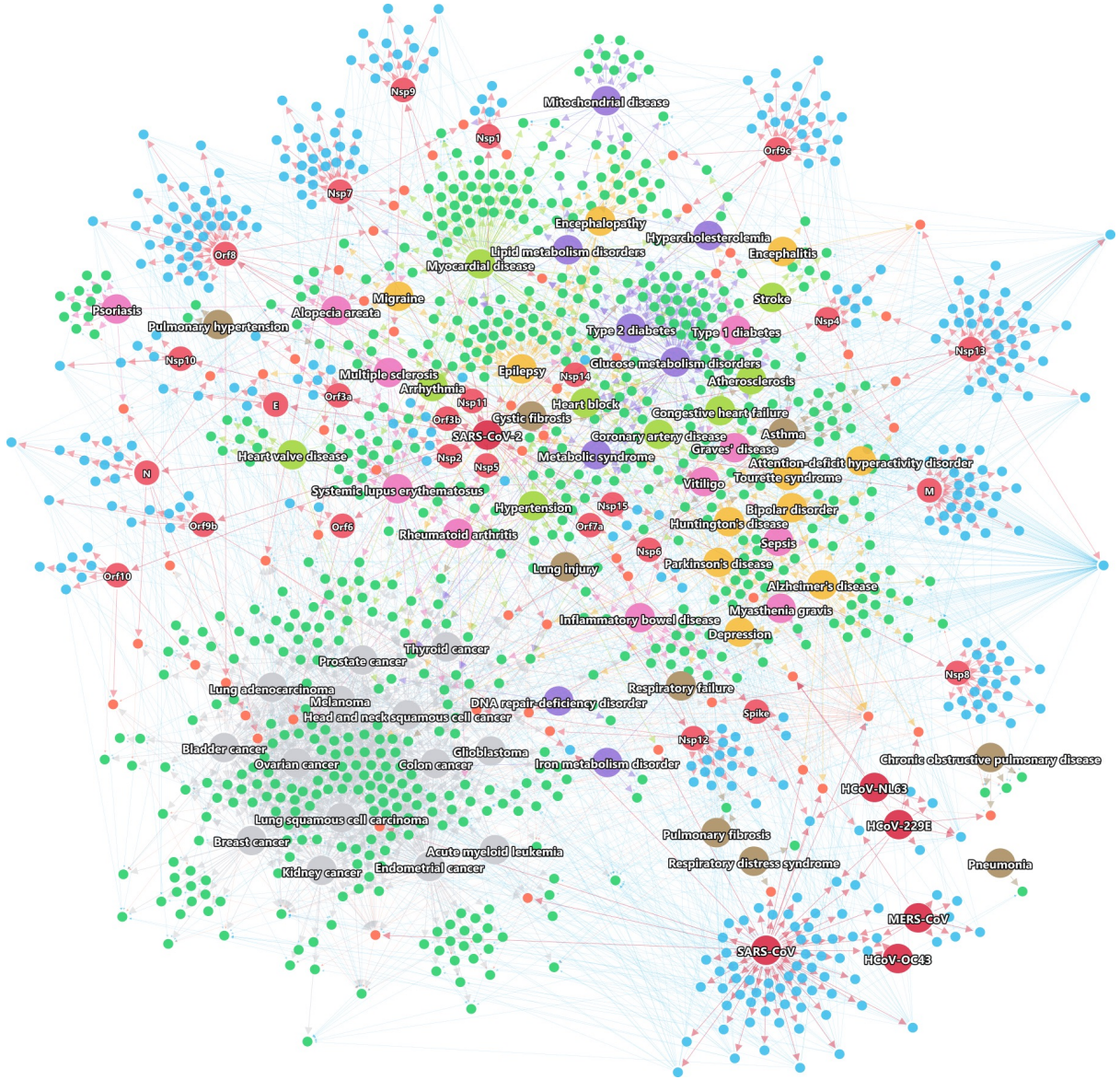


Decagon

Figure Source: *Fernández-Torra et al.*, Nature Communication 2022

Figure Source: *Zitnik, et al.*, Bioinformatics 2022

DISEASE NETWORK MAP FOR COVID-19



- Autoimmune
- Cancer
- Cardiovascular
- Metabolic
- Neurological
- Pulmonary

- Human coronaviruses (HCoVs)
- SARS-CoV-2 proteins

- Target host proteins of HCoVs
- Disease associated proteins
- Overlap of and

- Human PPIs
- Virus-human PPIs

- Disease-protein associations

Figure Source: *Zhou et al.*, PLOS Biology 2020

INTELLIGENT DRUG DISCOVERY RESEARCH QUESTIONS AND TASKS

ZHIQIANG ZHONG AND DAVIDE MOTTIN
DATA-INTENSIVE SYSTEMS GROUP

INTEGRATING AI IN DRUG DISCOVERY

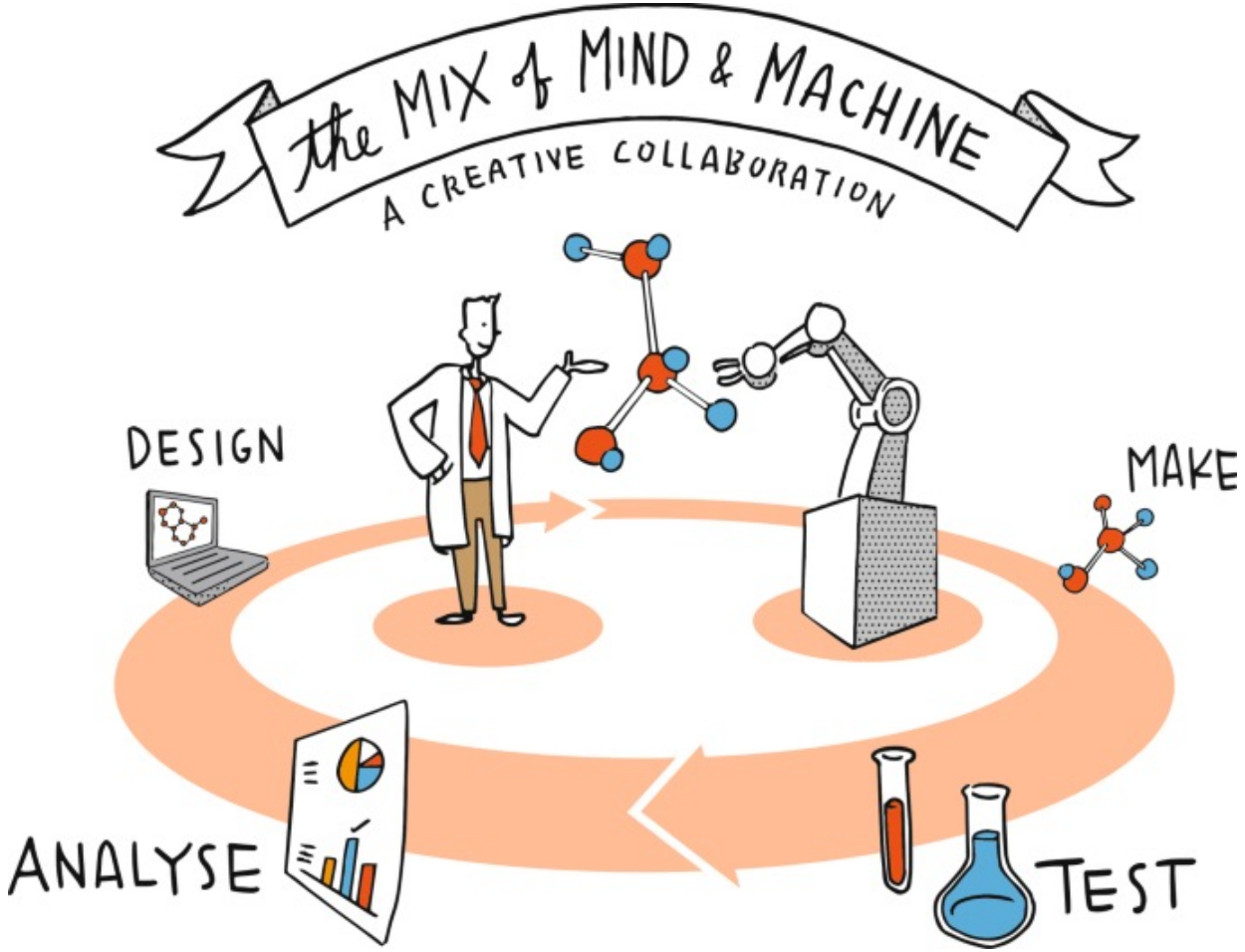
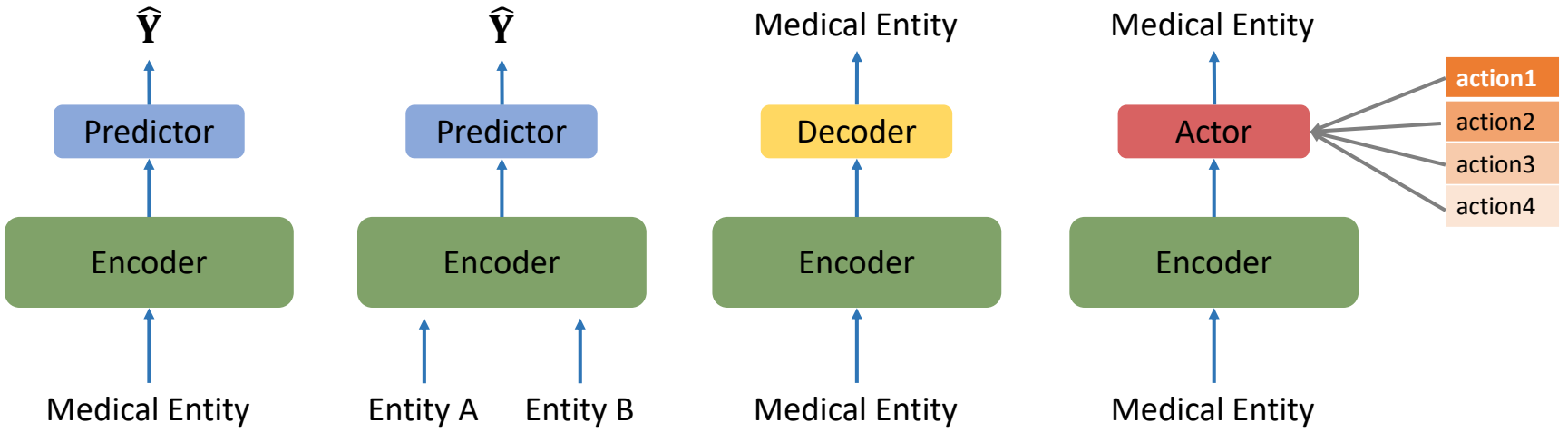


Figure Source: *Schneider, et al.*, Nature Review Drug Discovery 2019

INTELLIGENT DRUG DISCOVERY TASKS

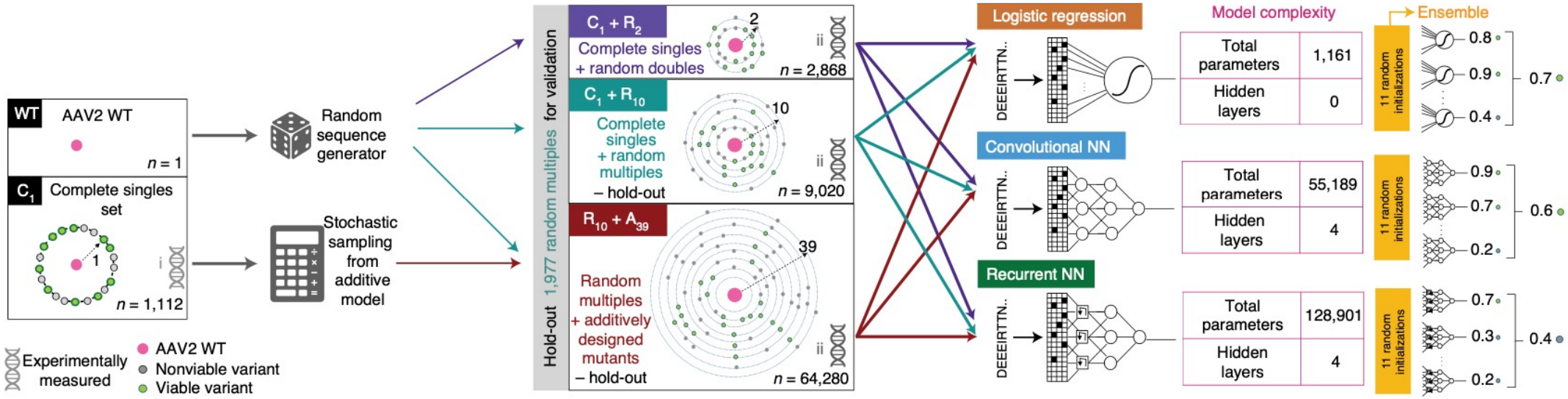
- Four schemes to realise intelligent drug discovery



INTELLIGENT ENTITY-PREDICTION TASKS

PREDENTITY

Bryant et al., (Nat. Biotechnol.'21)

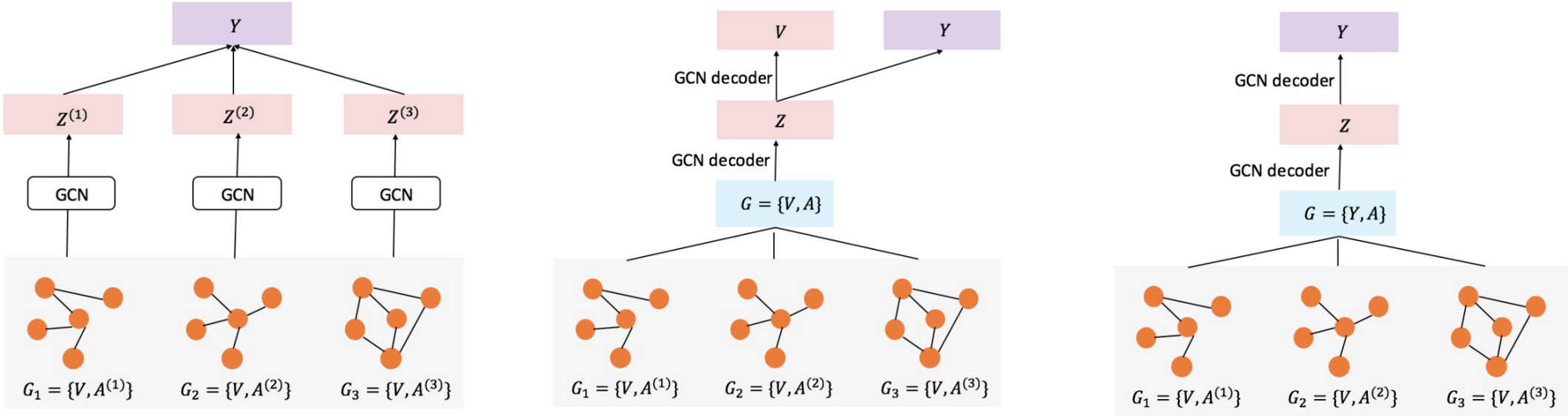


- AI algorithms take AAV proteins as input and make predictions on whether they form a valid 3D structure.

INTELLIGENT ENTITY-PAIRING TASKS

PAIRENTITY

Ma et al., (IJCAI'18)

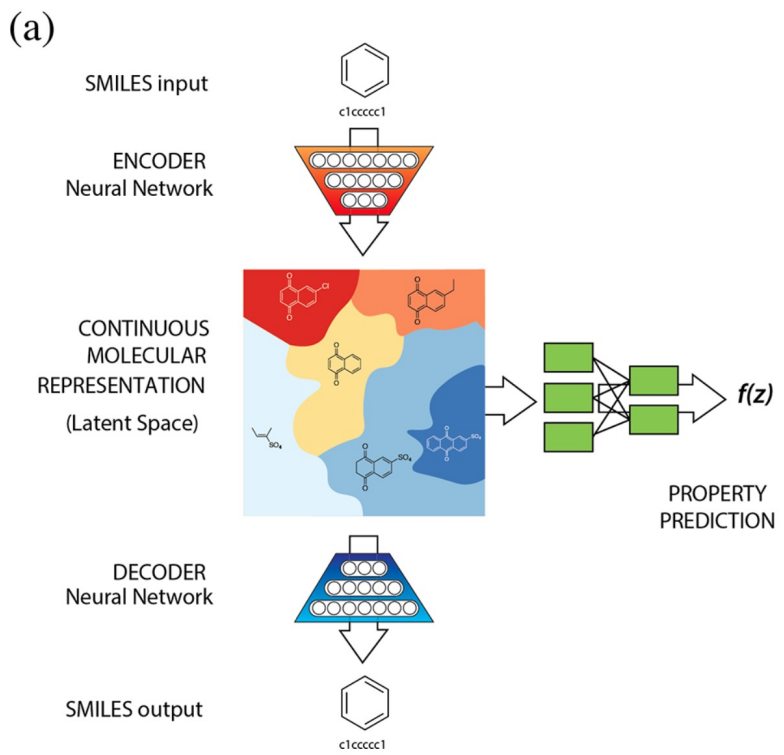


- Attentive multi-view graph auto-encoders to predict the similarity between different drug graphs.

INTELLIGENT ENTITY-TO-ENTITY TASKS

ENTITY2ENTITY

Gómez-Bombarelli et al., (ACS Cent. Sci.'18)

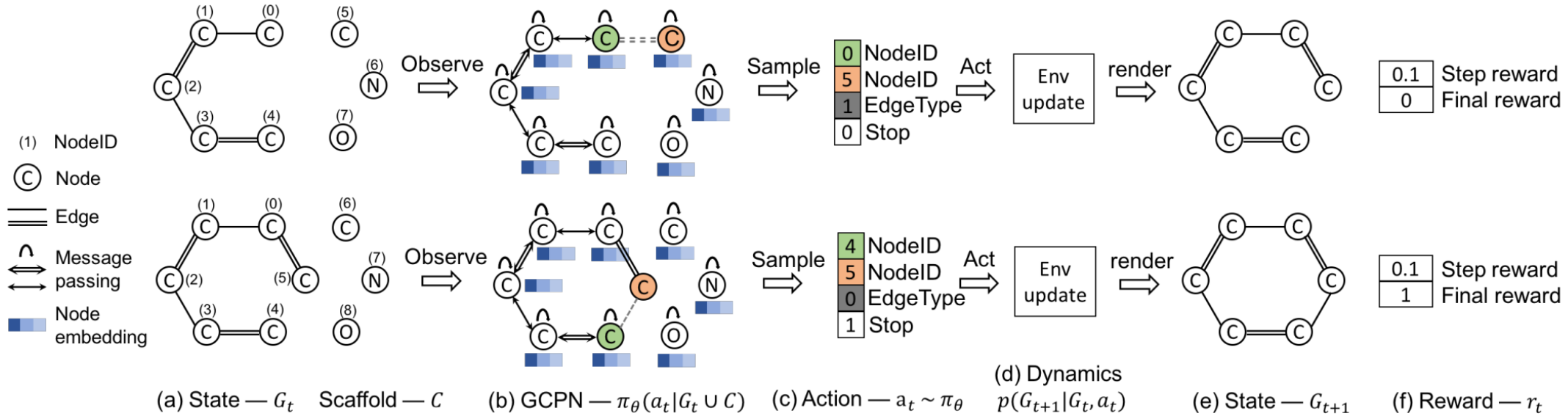


- Generate new molecules with desired new properties based on existing molecules.

INTELLIGENT ACTION-PREDICTION TASKS

ACTENTITY

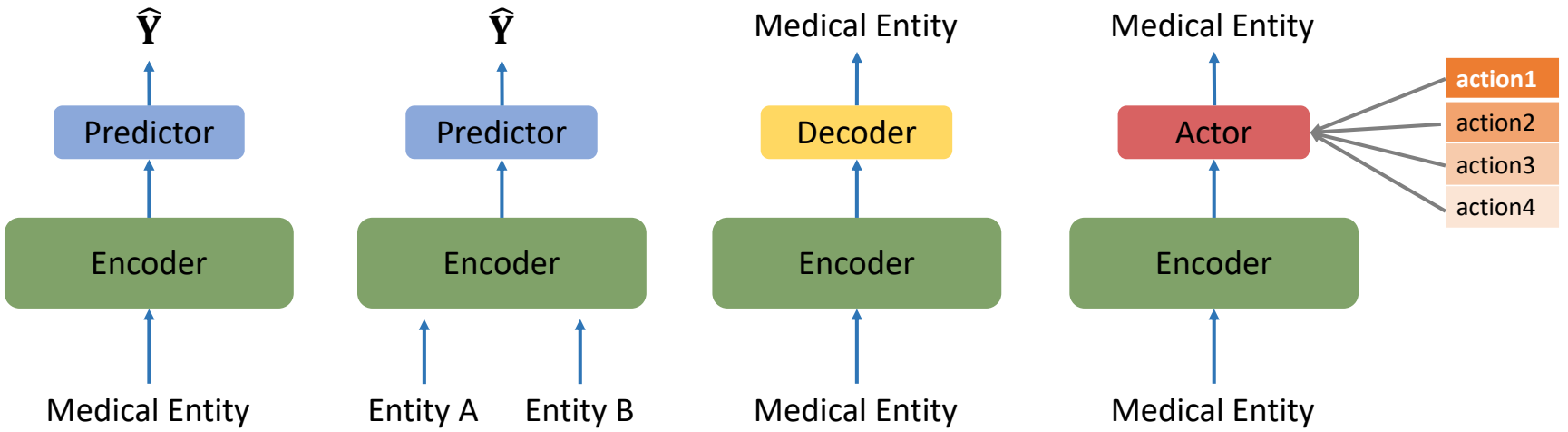
You et al., (NeurIPS'18)



- A set of actions are predicted continuously to modify existing molecules for desired states.

INTELLIGENT DRUG DISCOVERY TASKS

- Four schemes to realise intelligent drug discovery



- Different biomedical **Entity** can be classified at different levels, including molecular (e.g., **Gene**, **Molecule**), macro-molecular (e.g., **Protein**, **Antibodies**, **Enzymes**, **Receptors**, **Compound**, and organism levels (e.g., **Cell**).

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)

AARHUS UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

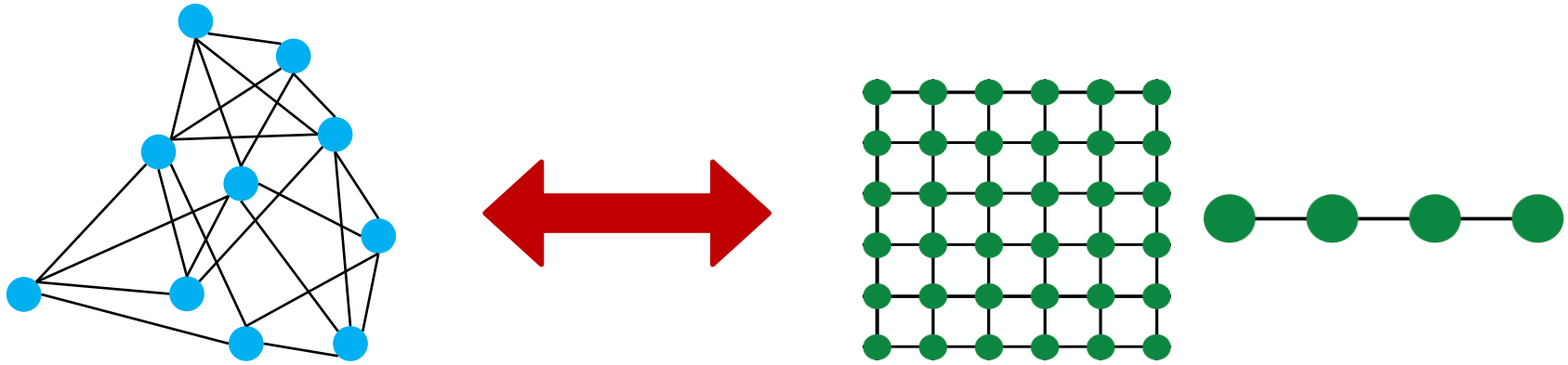
KDD'23 TUTORIAL
9 AUGUST 2023

ZHIQIANG ZHONG AND DAVIDE MOTTIN
DATA-INTENSIVE SYSTEMS GROUP

DEALING WITH GRAPHS IS DIFFICULT

Graphs are far more complex!

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

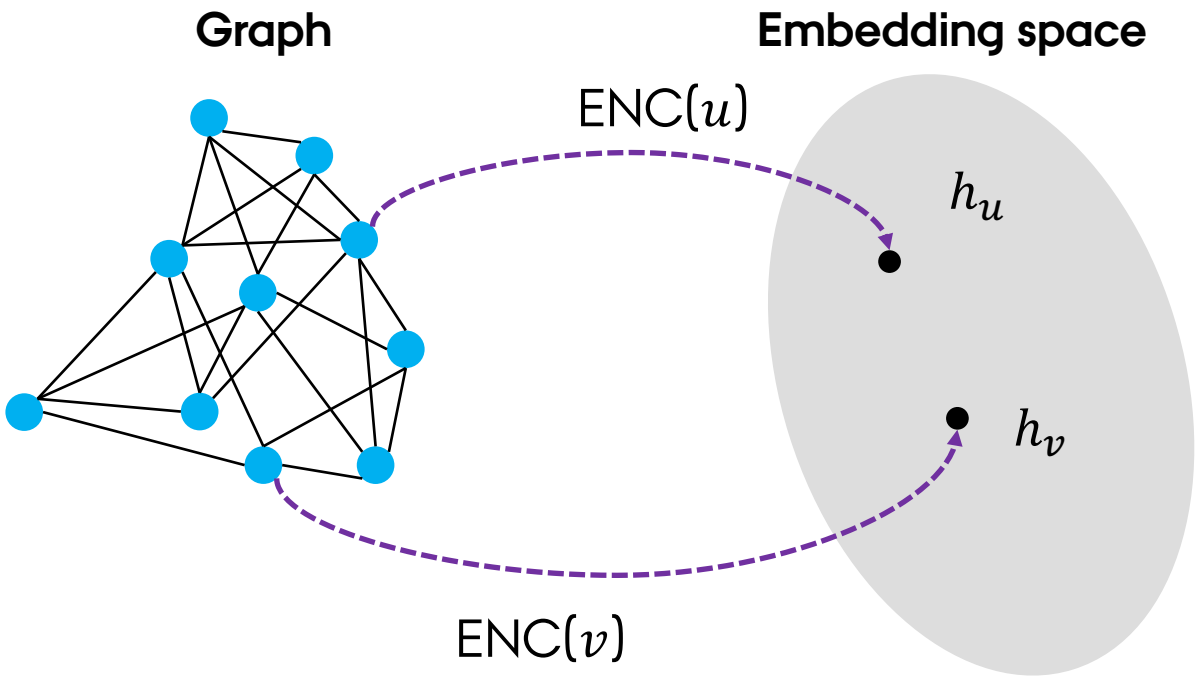


Graphs Images Text

- 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

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

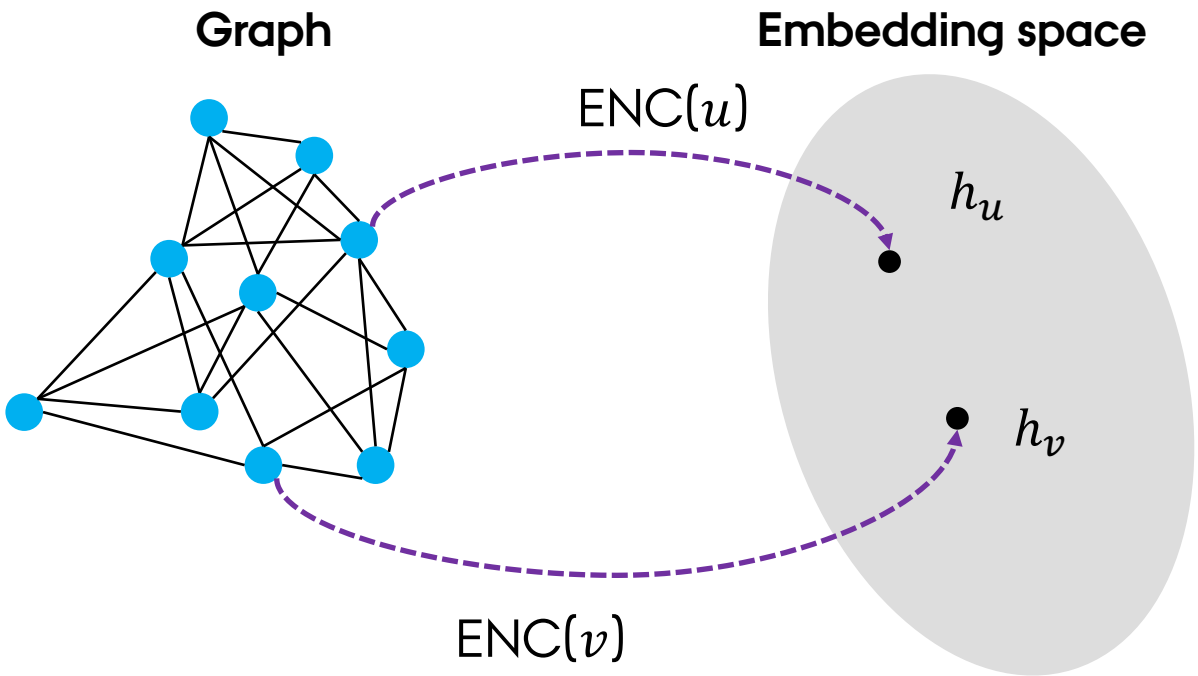
❑ Graph property prediction

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

❑ Etc.

(KNOWLEDGE) GRAPH MACHINE LEARNING

- (Knowledge) Graph Representation Learning



❑ Node property prediction

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

❑ Link prediction

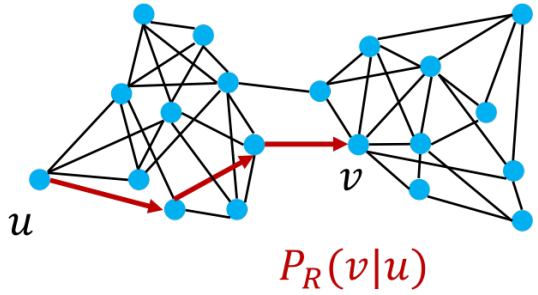
$\hat{Y}_{uv} = f(h_u, h_v, e_{uv}) - \textbf{PAIRENTITY}$

❑ Graph property prediction

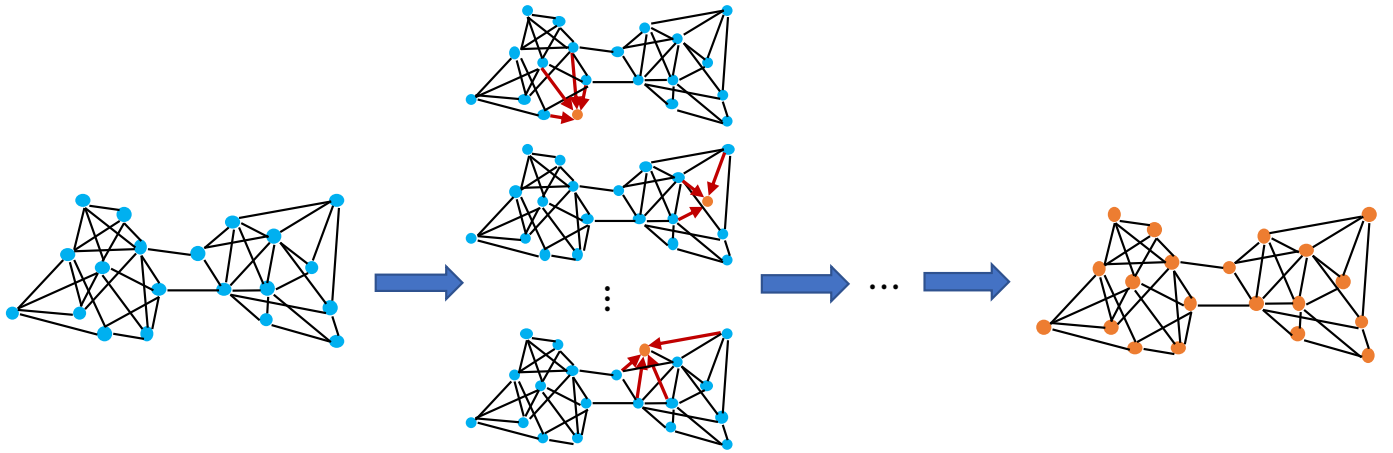
$\hat{Y}_{\mathcal{G}} = f(\bigoplus_{u \in V} h_u) - \textbf{ENTITY2ENTITY}$

❑ Etc.

(KNOWLEDGE) GRAPH MACHINE LEARNING

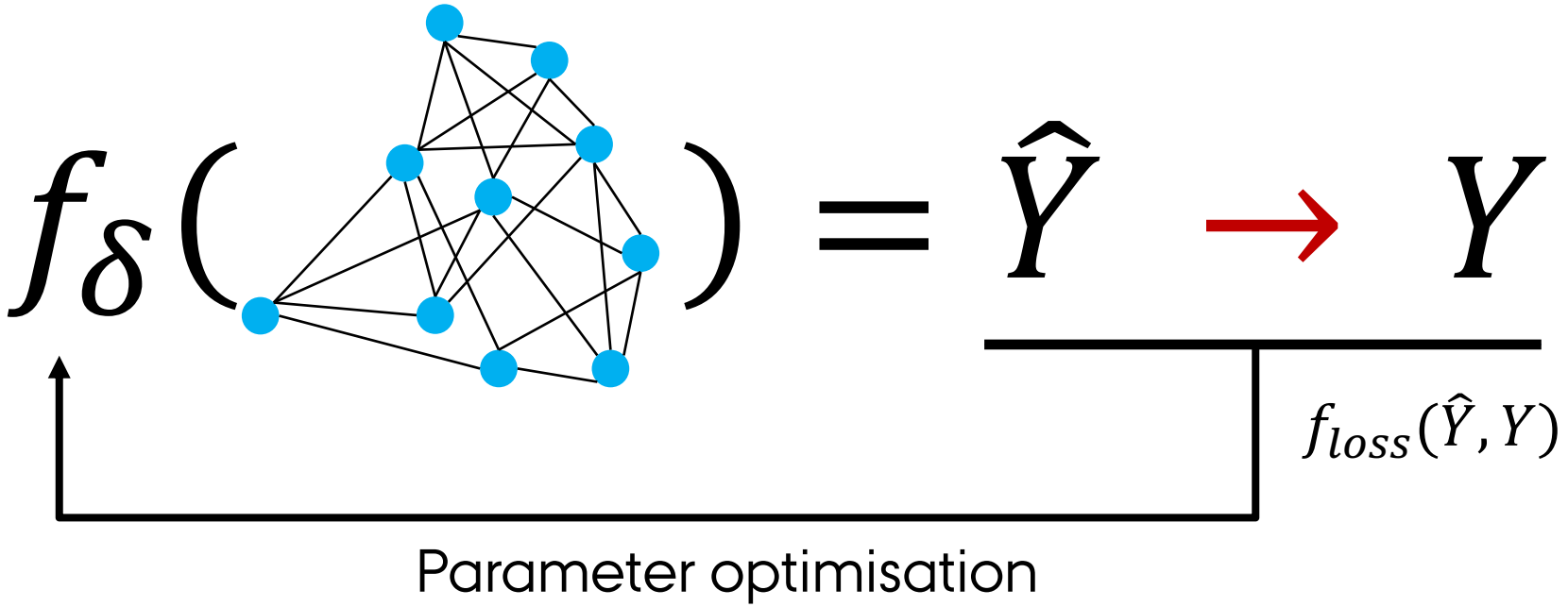


Estimate the probability of visiting node v on a random walk starting from node u using some random walk strategy R.



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

HOW TO TRAIN (K)GML MODELS?



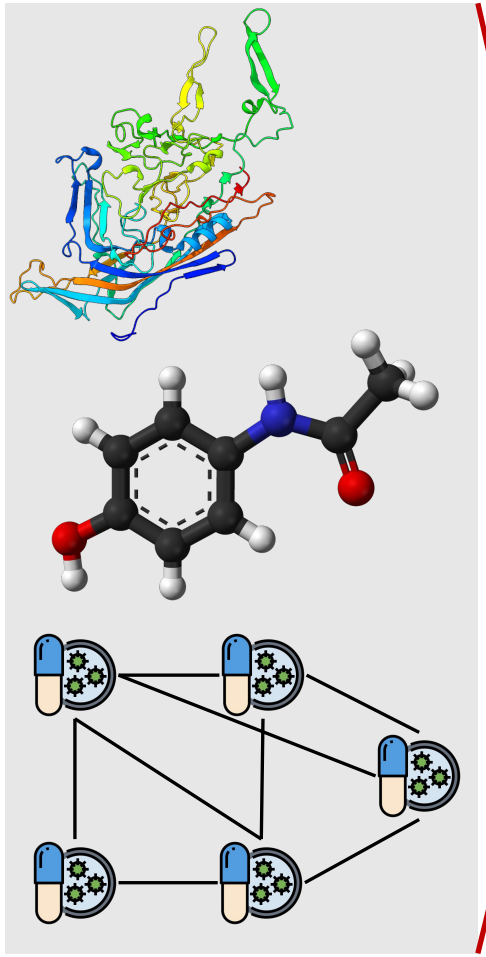
GML AND KG FOR DRUG DISCOVERY

AARHUS
UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

KDD'23 TUTORIAL
9 AUGUST 2023

ZHIQIANG ZHONG AND DAVIDE MOTTIN
DATA-INTENSIVE SYSTEMS GROUP

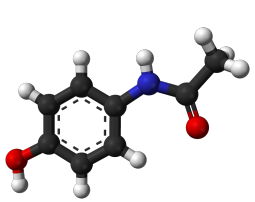
GML FOR DRUG DISCOVERY



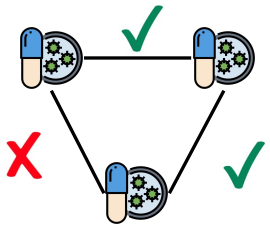
Input Graph Structure Data

Graph Machine Learning

Predictions, Insights, ...

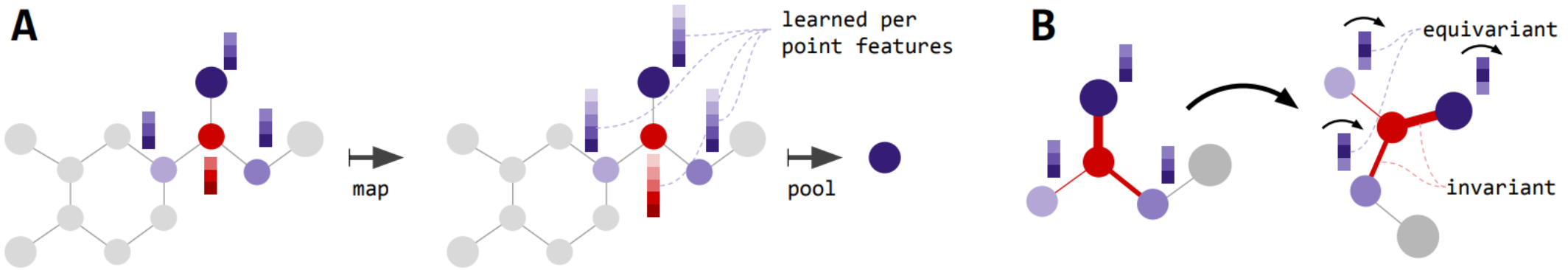


Property



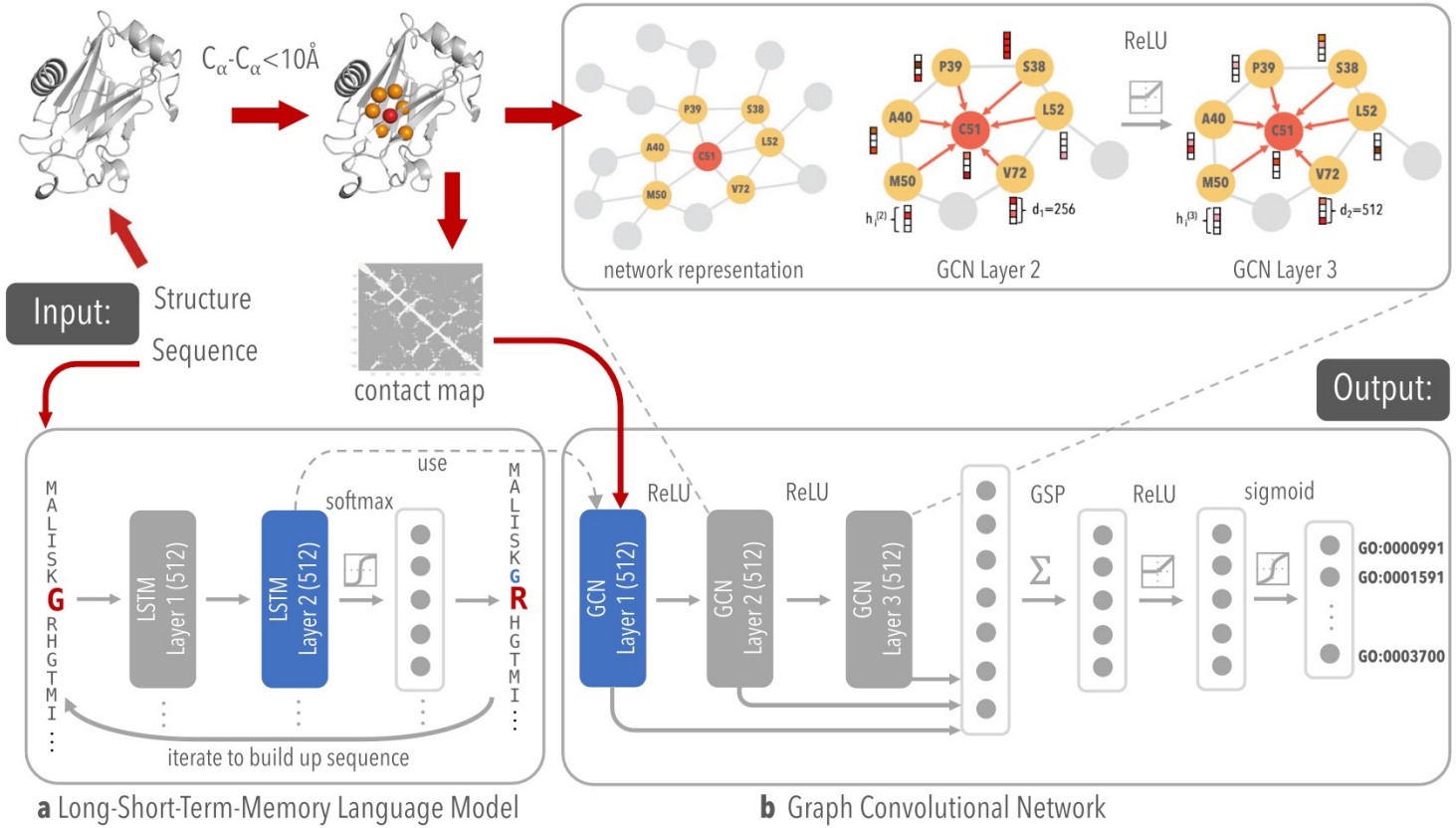
Interaction

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

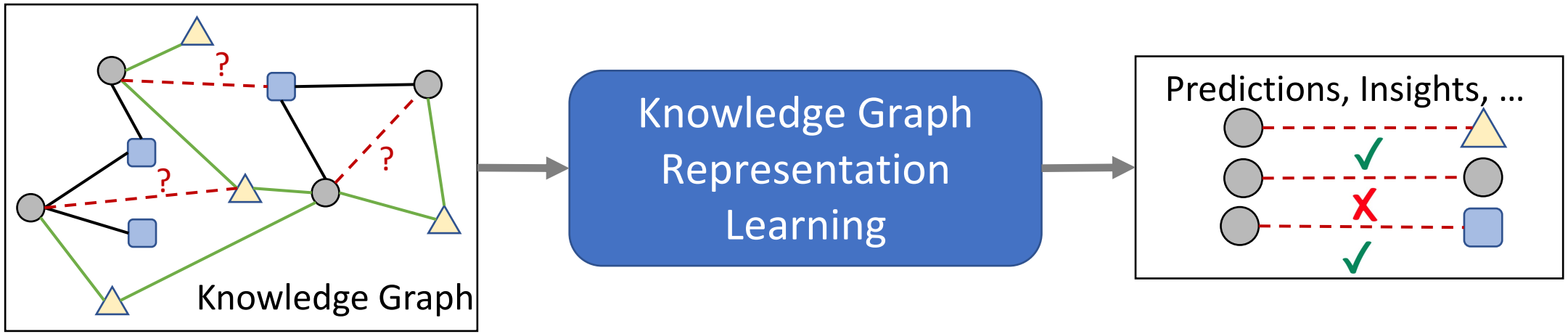
ONE MORE STEP: INVESTIGATE GML

- 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

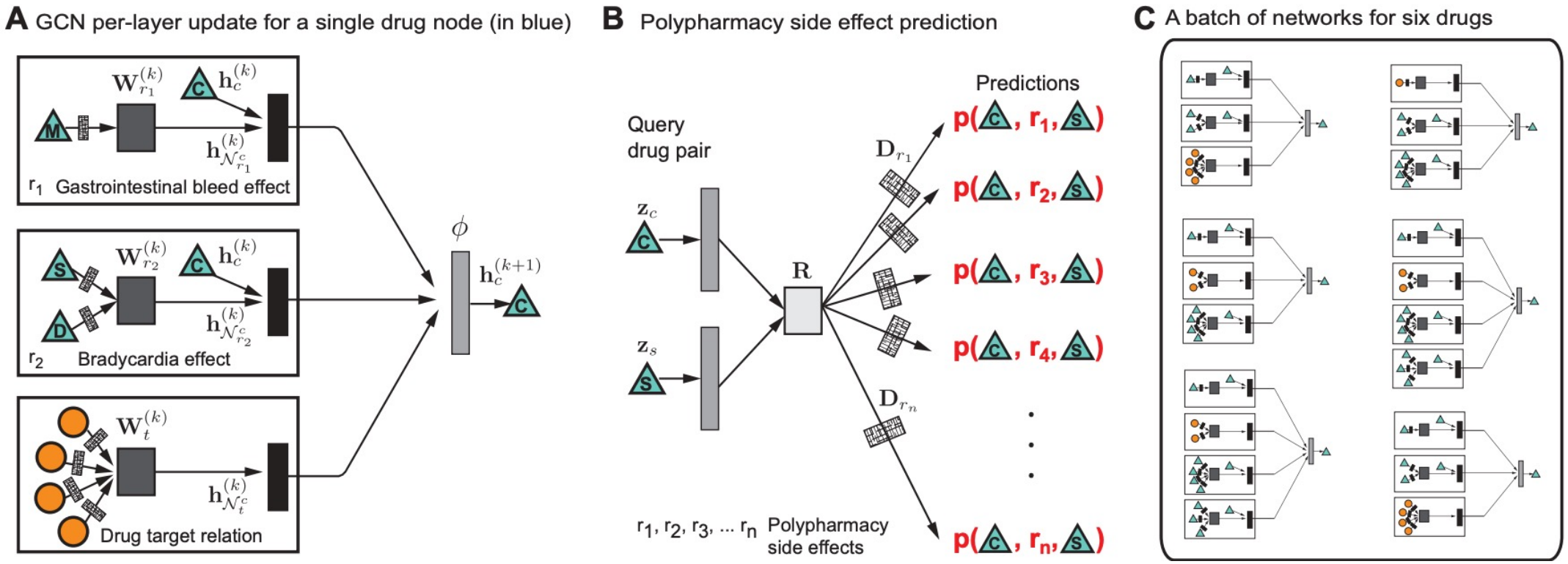
GML FOR DRUG DISCOVERY – PAPER LIST

- 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

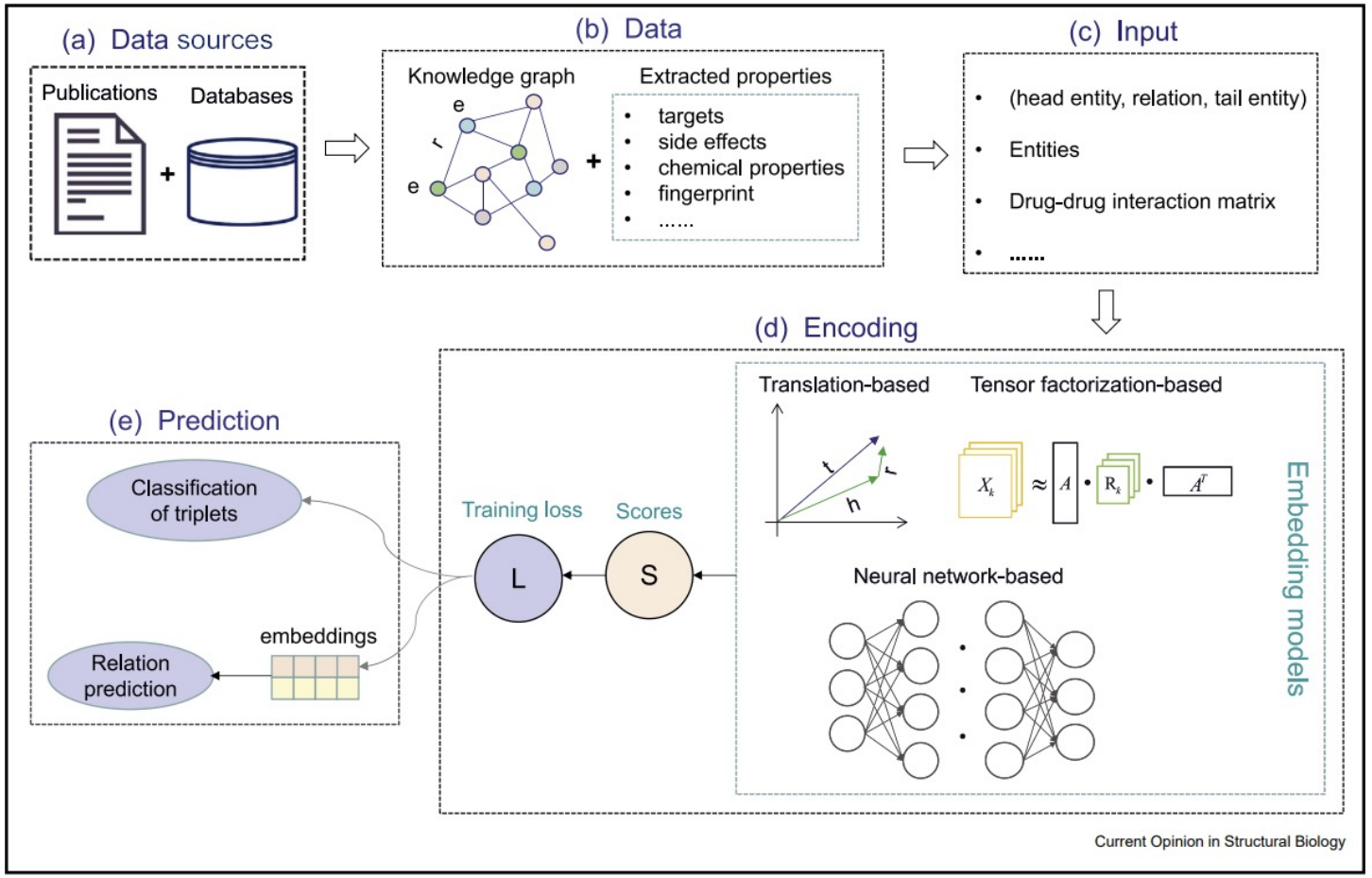


DECAGON (BIOINFOM. 18)



- 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)



- 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.

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)

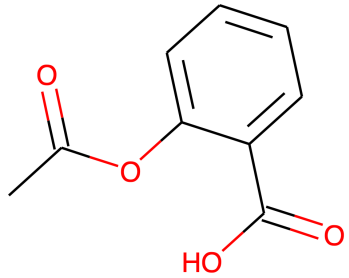
AARHUS UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

KDD'23 TUTORIAL
9 AUGUST 2023

ZHIQIANG ZHONG AND DAVIDE MOTTIN
DATA-INTENSIVE SYSTEMS GROUP

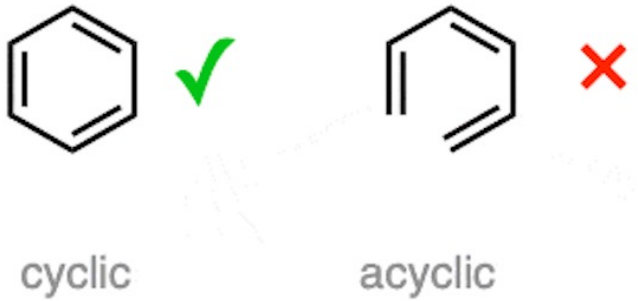
KNOWLEDGE IN DRUG DISCOVERY

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

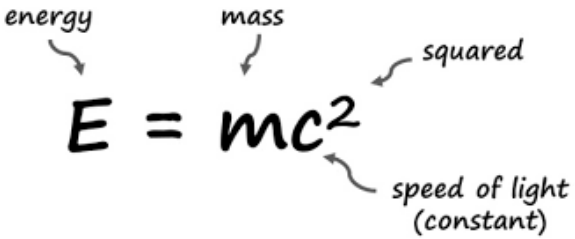


Aspirin

- 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

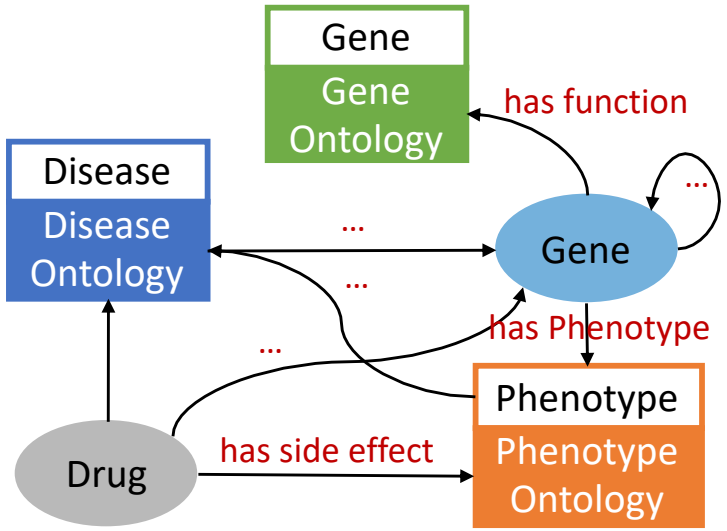


HUMAN BIOMEDICAL KNOWLEDGE



Theories and Equations

Description Context

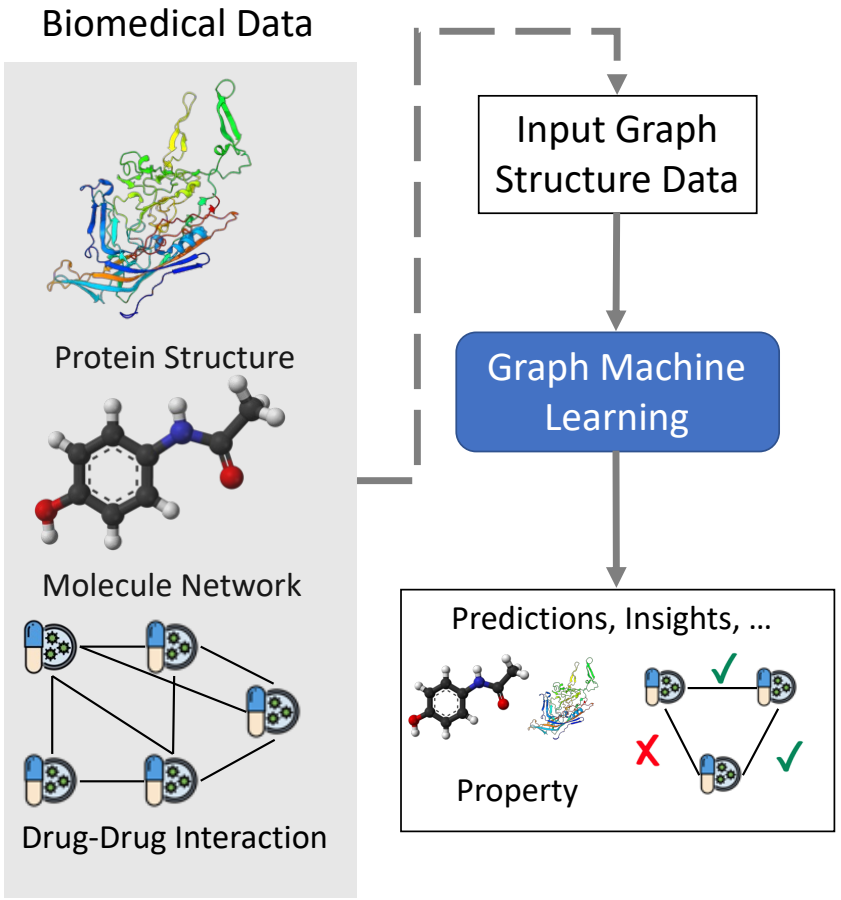


Knowledge Graph

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

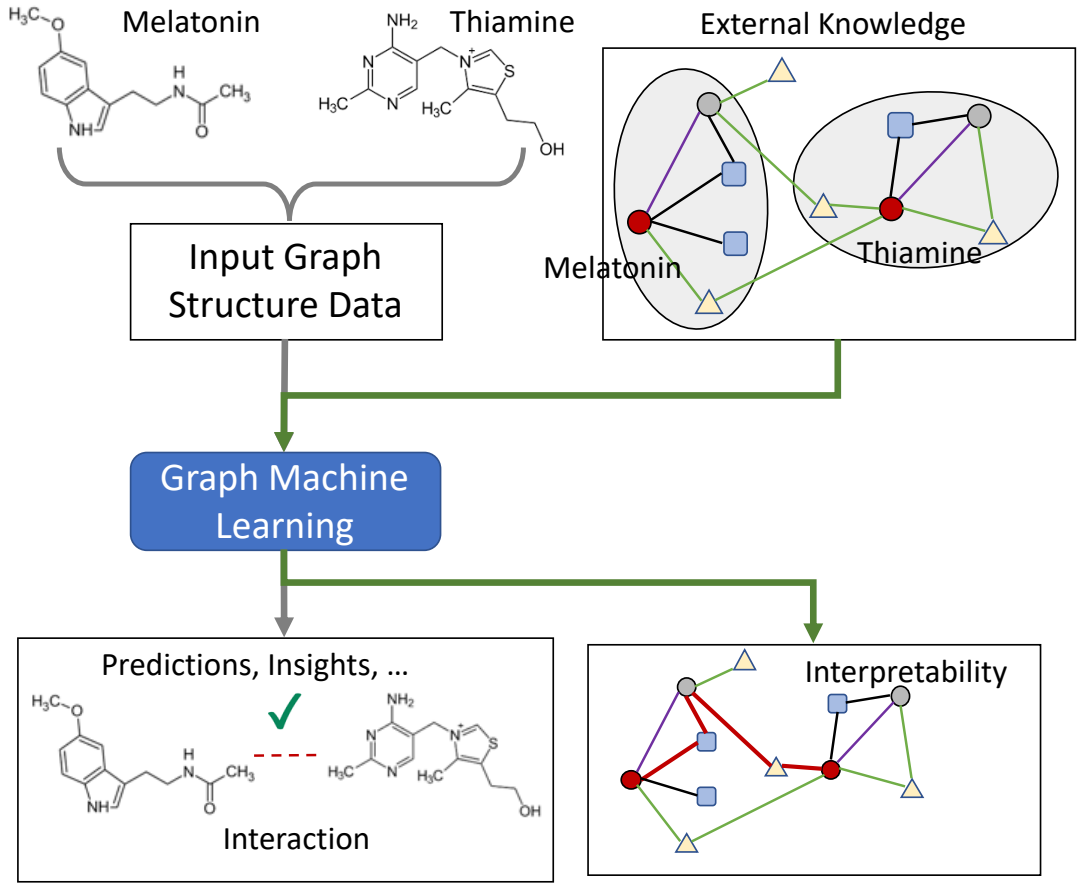
WHAT IS KNOWLEDGE-AUGMENTED GML

GML for Drug Discovery



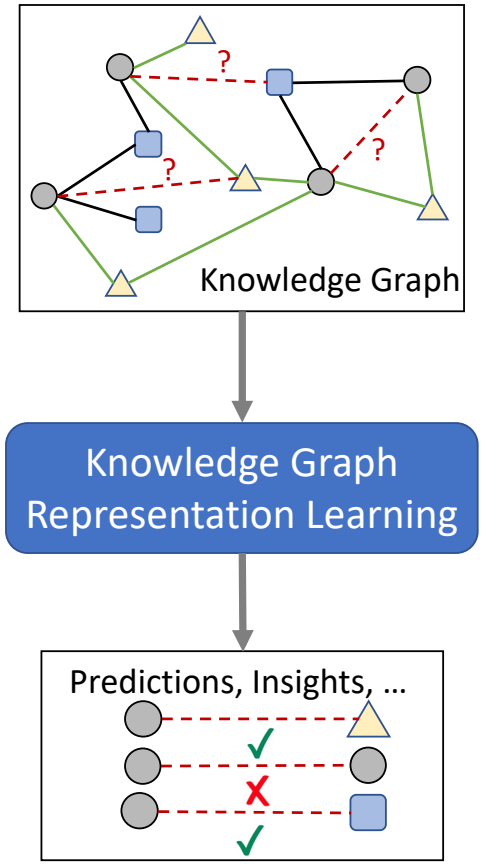
VS.

KaGML for Drug Discovery



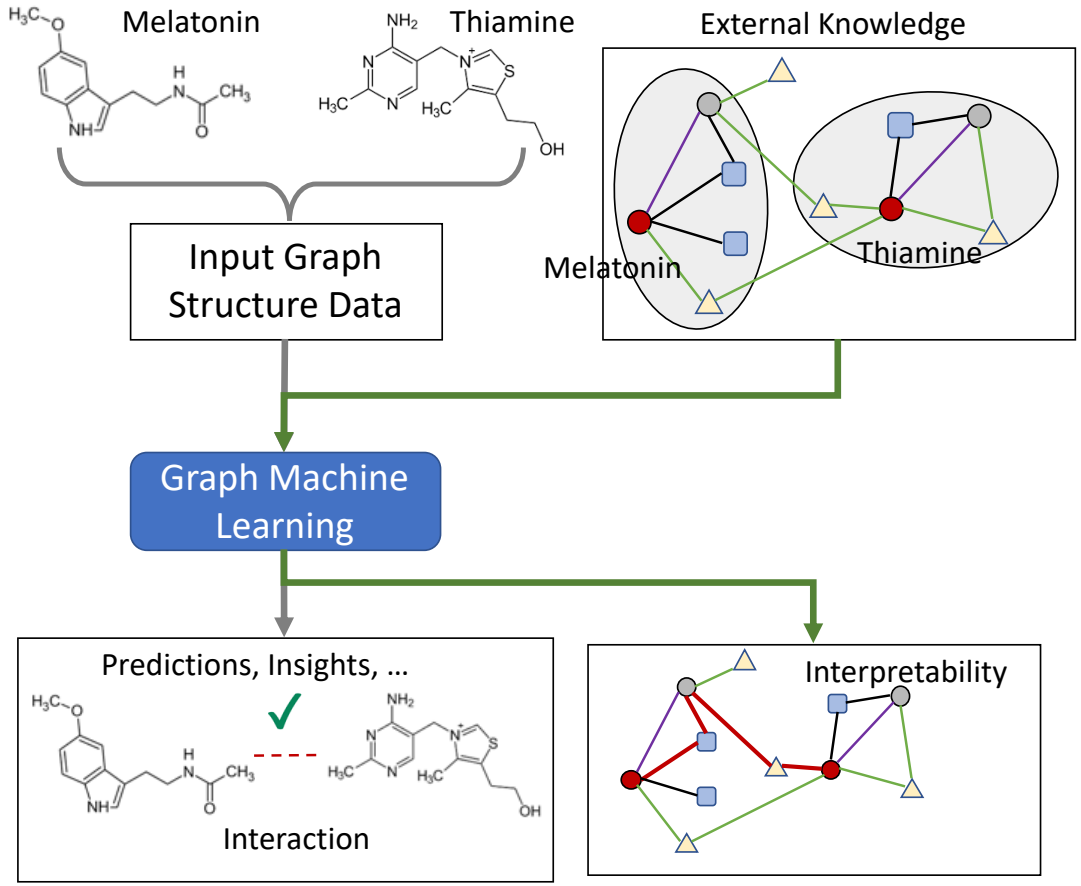
WHAT IS KNOWLEDGE-AUGMENTED GML

KG for Drug Discovery

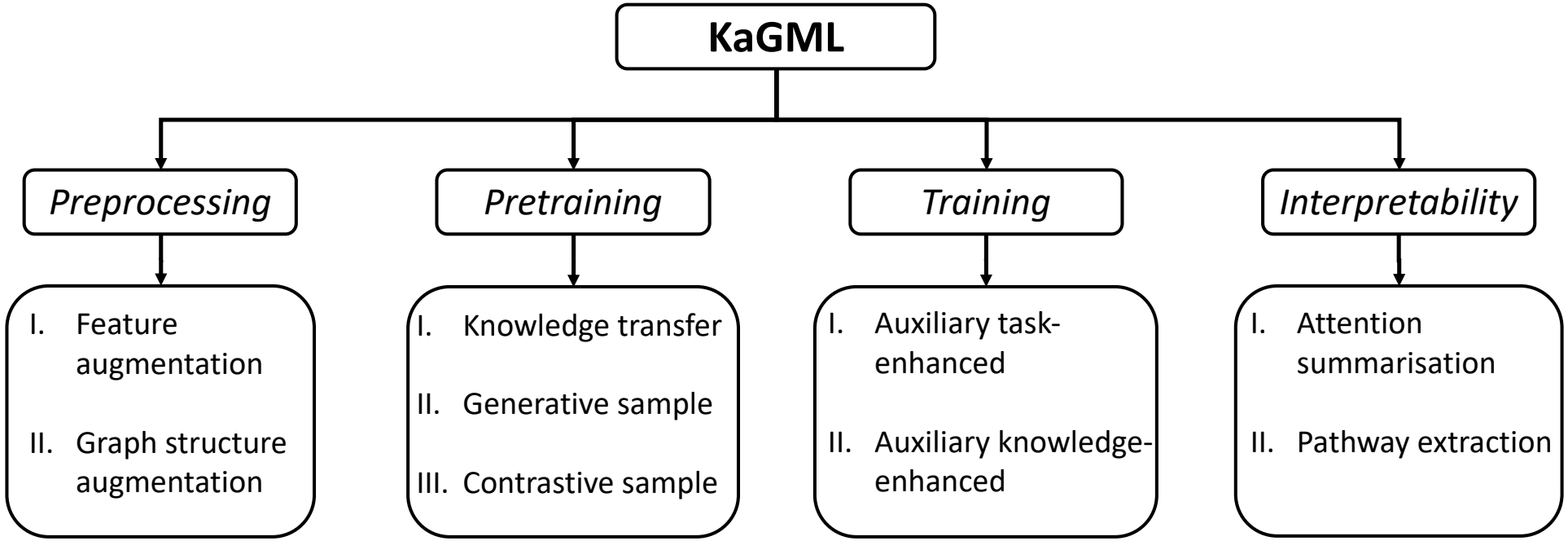


VS.

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*

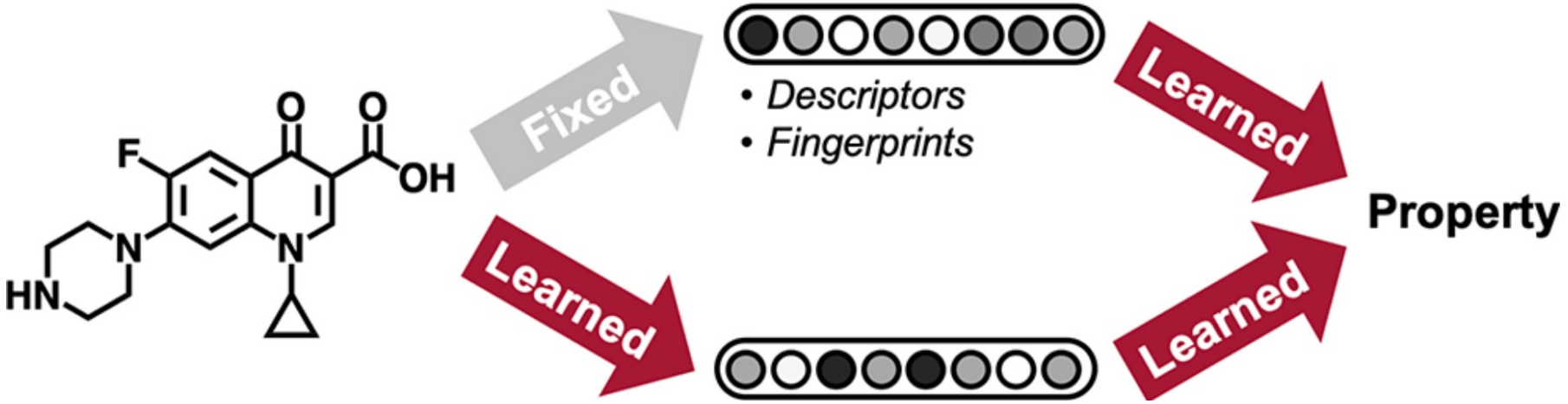
AARHUS UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

KDD'23 TUTORIAL
9 AUGUST 2023

ZHIQIANG ZHONG AND DAVIDE MOTTIN
DATA-INTENSIVE SYSTEMS GROUP

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

Pre-processing

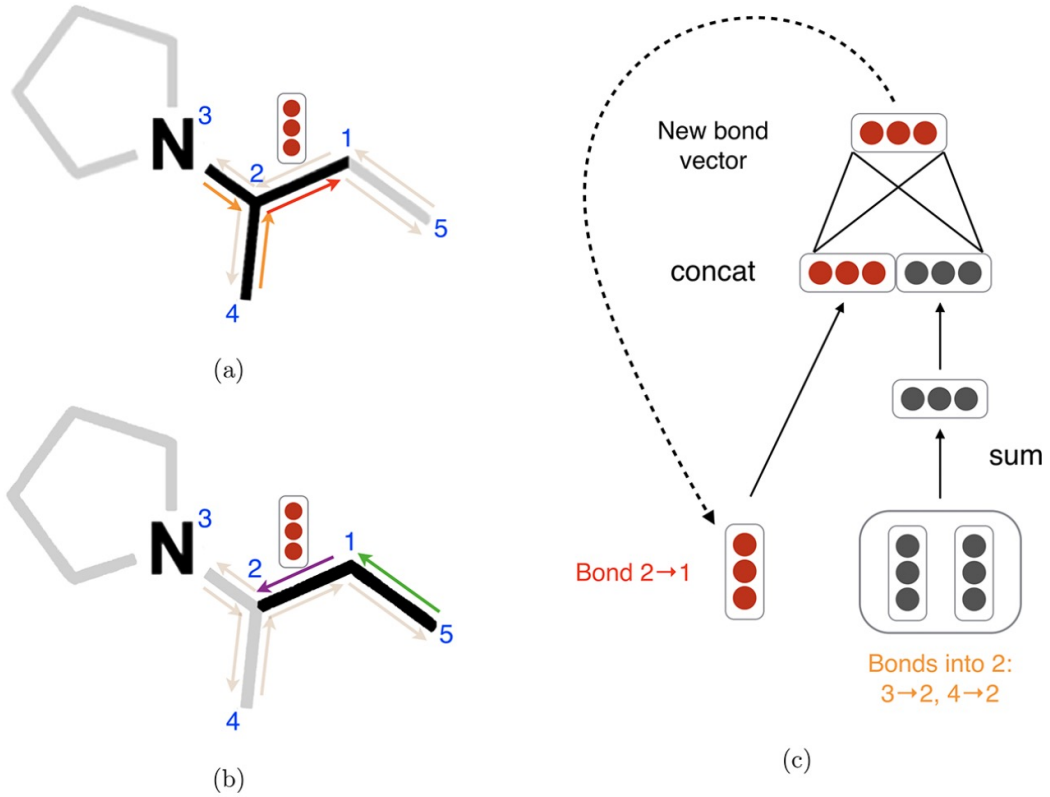


- Employ molecular node and edge features generated by chemical tools, such as RDKit[1] and UFF[2]
- Global interactions beyond L hops can be summarised into generated features

[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)

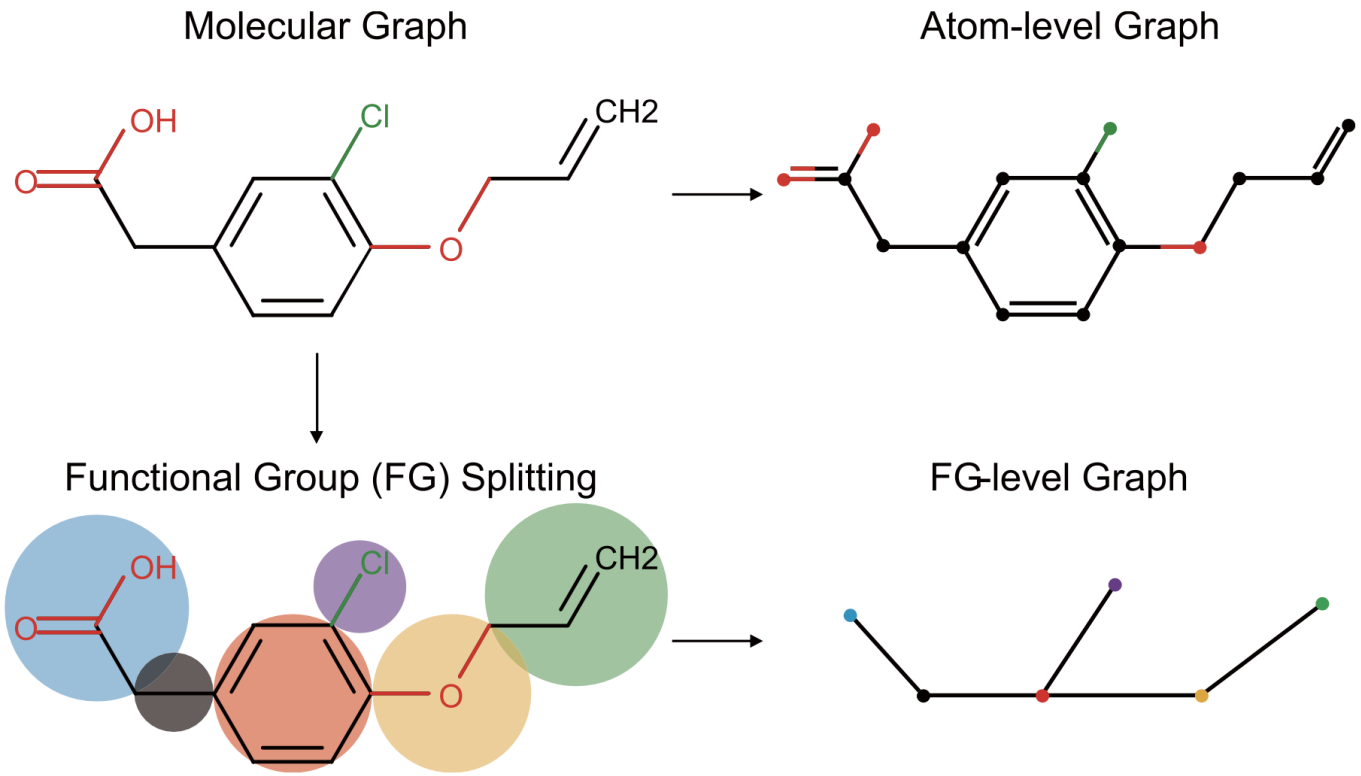
Pre-processing



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

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

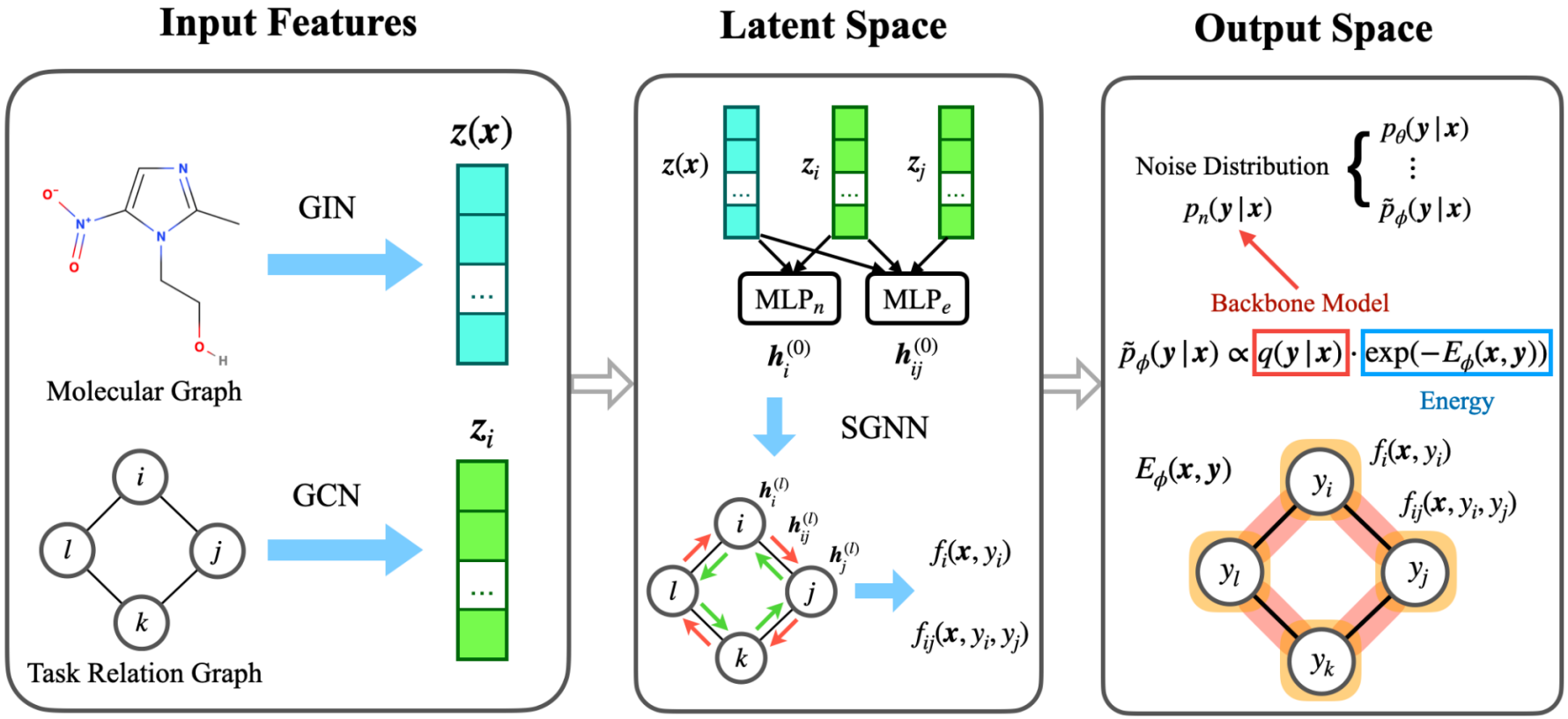
Pre-processing



- Graph augmentation based on biomedical knowledge

SGNN-EBM (AISTATS'22)

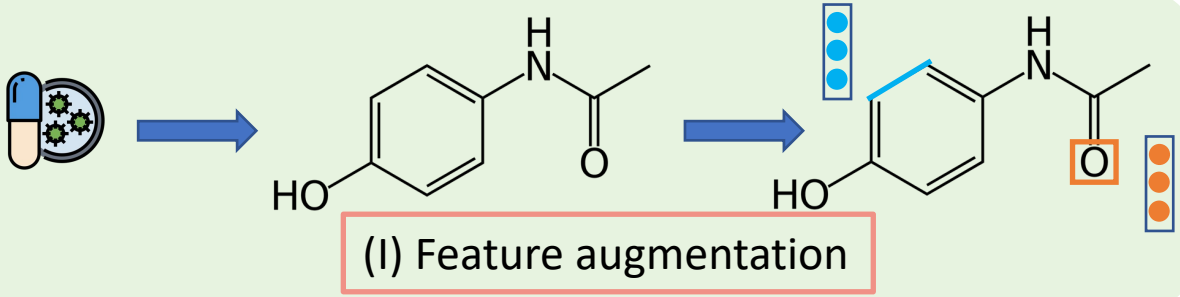
Pre-processing

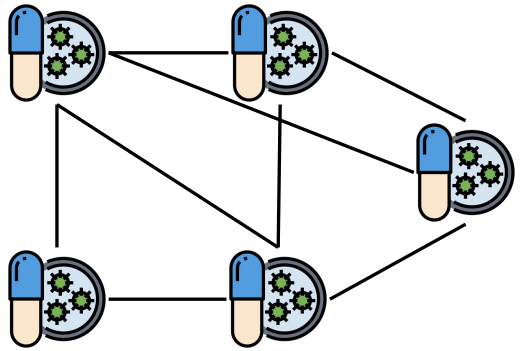


- Different tasks can be organised as task relation graphs to reveal the relationship between them.

KNOWLEDGE IN PRE-PROCESSING

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



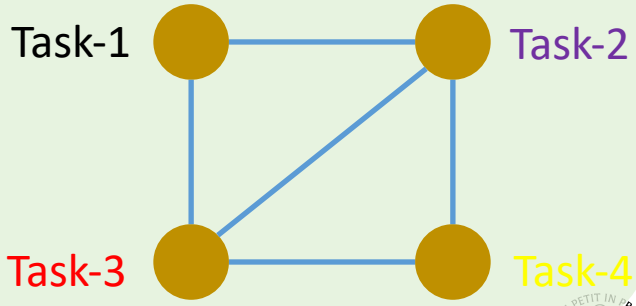


Task-1:
Drug property prediction

- **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*

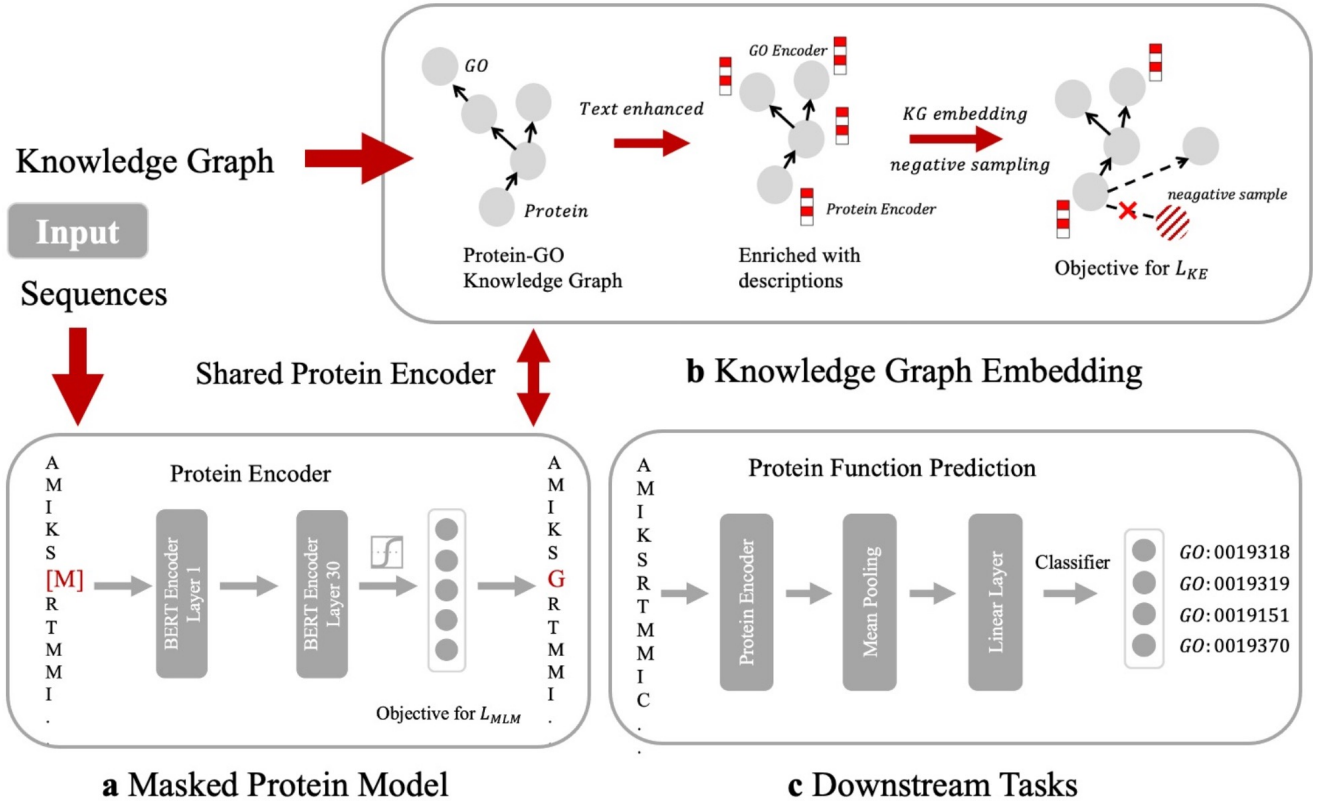
AARHUS UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

KDD'23 TUTORIAL
9 AUGUST 2023

ZHIQIANG ZHONG AND DAVIDE MOTTIN
DATA-INTENSIVE SYSTEMS GROUP

ONTOPROTEIN (ICLR'22)

Pre-Training

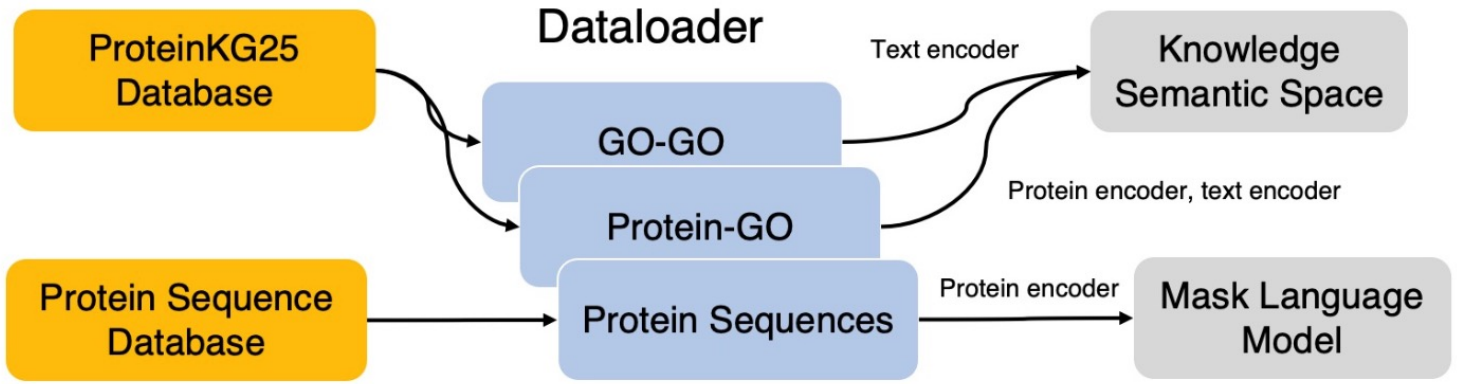


- Knowledge from Gene Ontology[1] database can optimise protein embedding during pre-training

[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

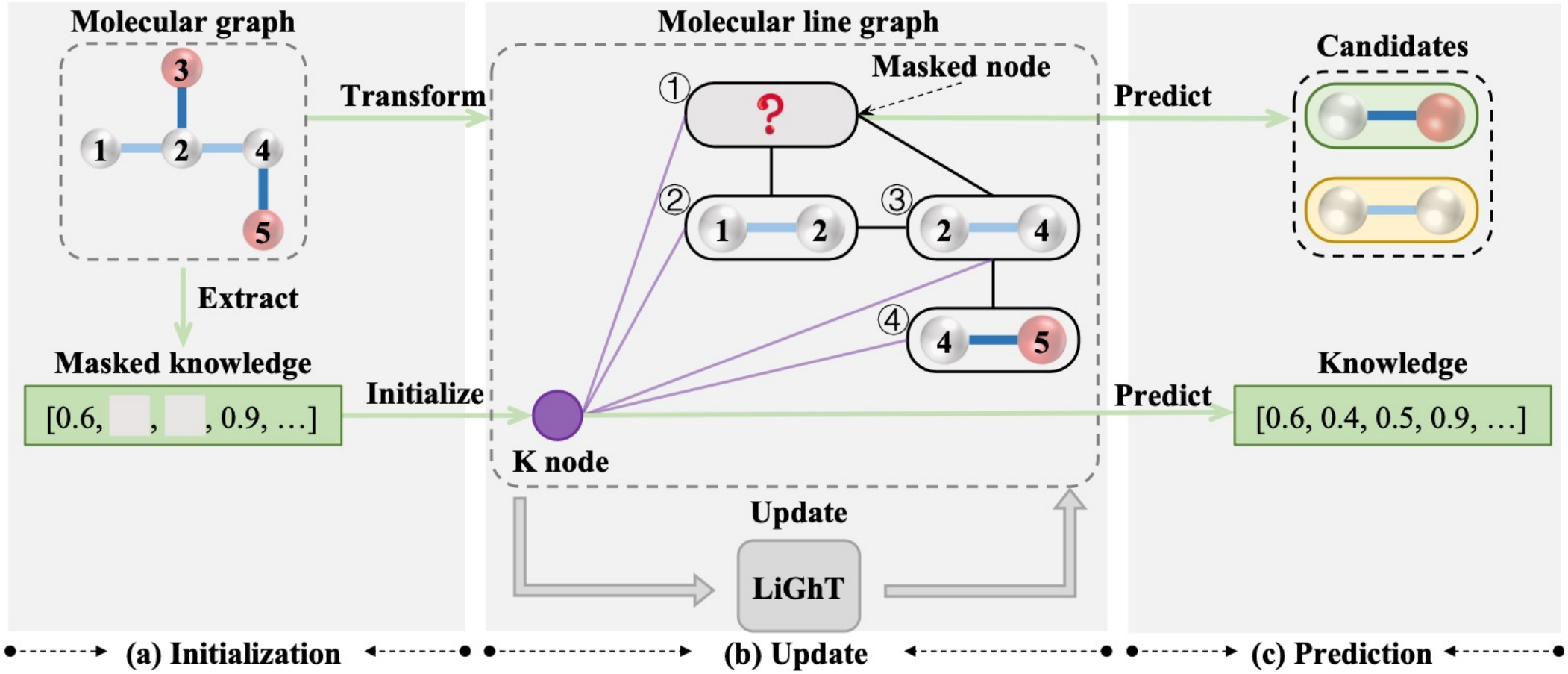
Method	Structure			Evolutionary	Engineering	
	SS-Q3	SS-Q8	Contact	Homology	Fluorescene	Stability
LSTM	0.75	0.59	0.26	0.26	0.67	0.69
TAPE Transformer	0.73	0.59	0.25	0.21	**0.68**	0.73
ResNet	0.75	0.58	0.25	0.17	0.21	0.73
MSA Transformer	-	**0.73**	**0.49**	-	-	-
ProtBert	0.81	0.67	0.35	**0.29**	0.61	**0.82**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

- Results on TAPE Benchmark

KPGT (KDD'22)

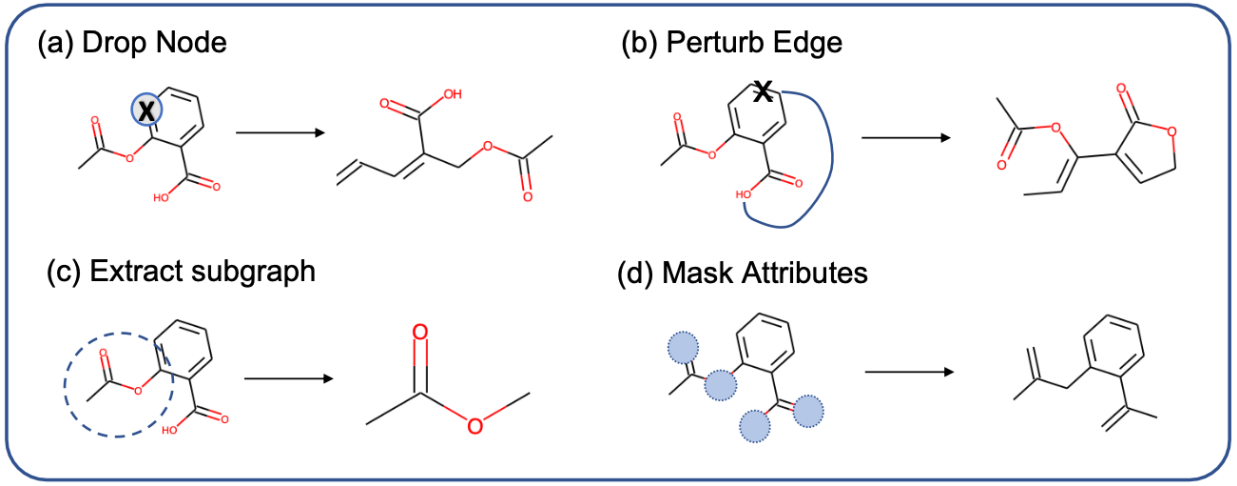
Augmented graphs

Pre-Training

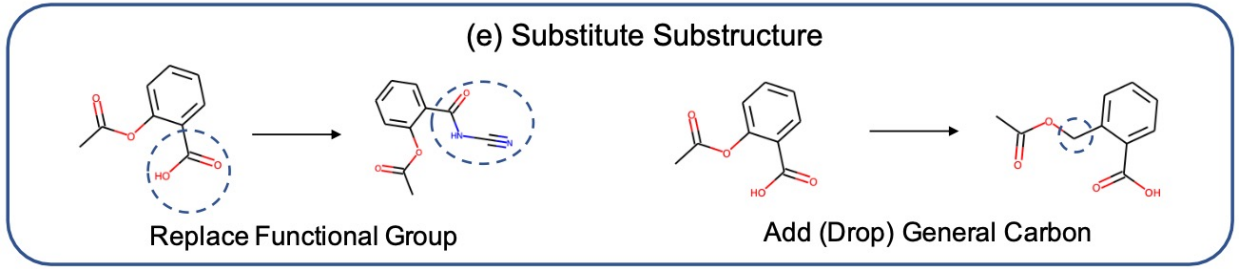


MOCL (KDD'21)

Pre-Training



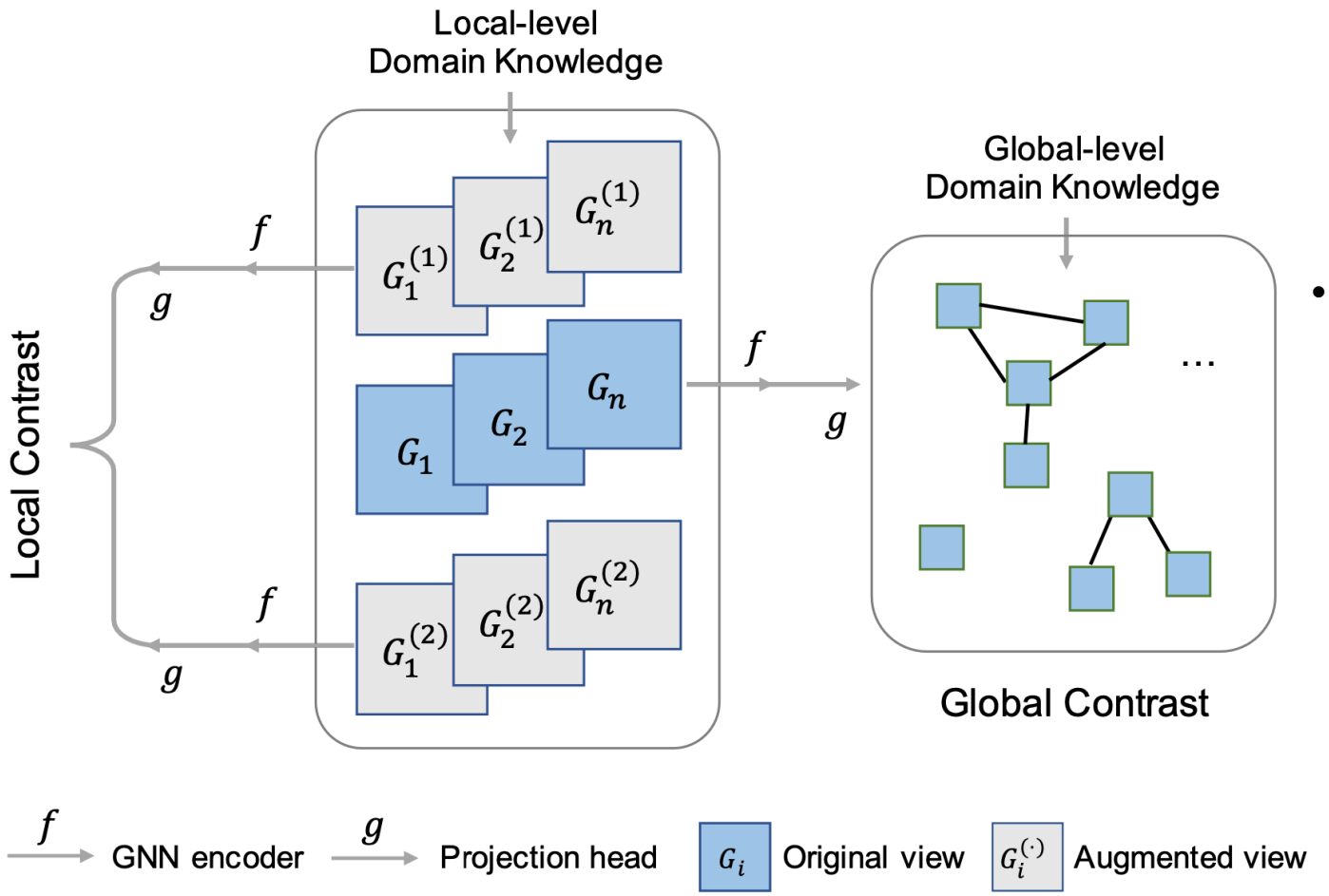
- Conventional augmentations may alter the graph semantics.



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

MOCL (KDD'21)

Pre-Training

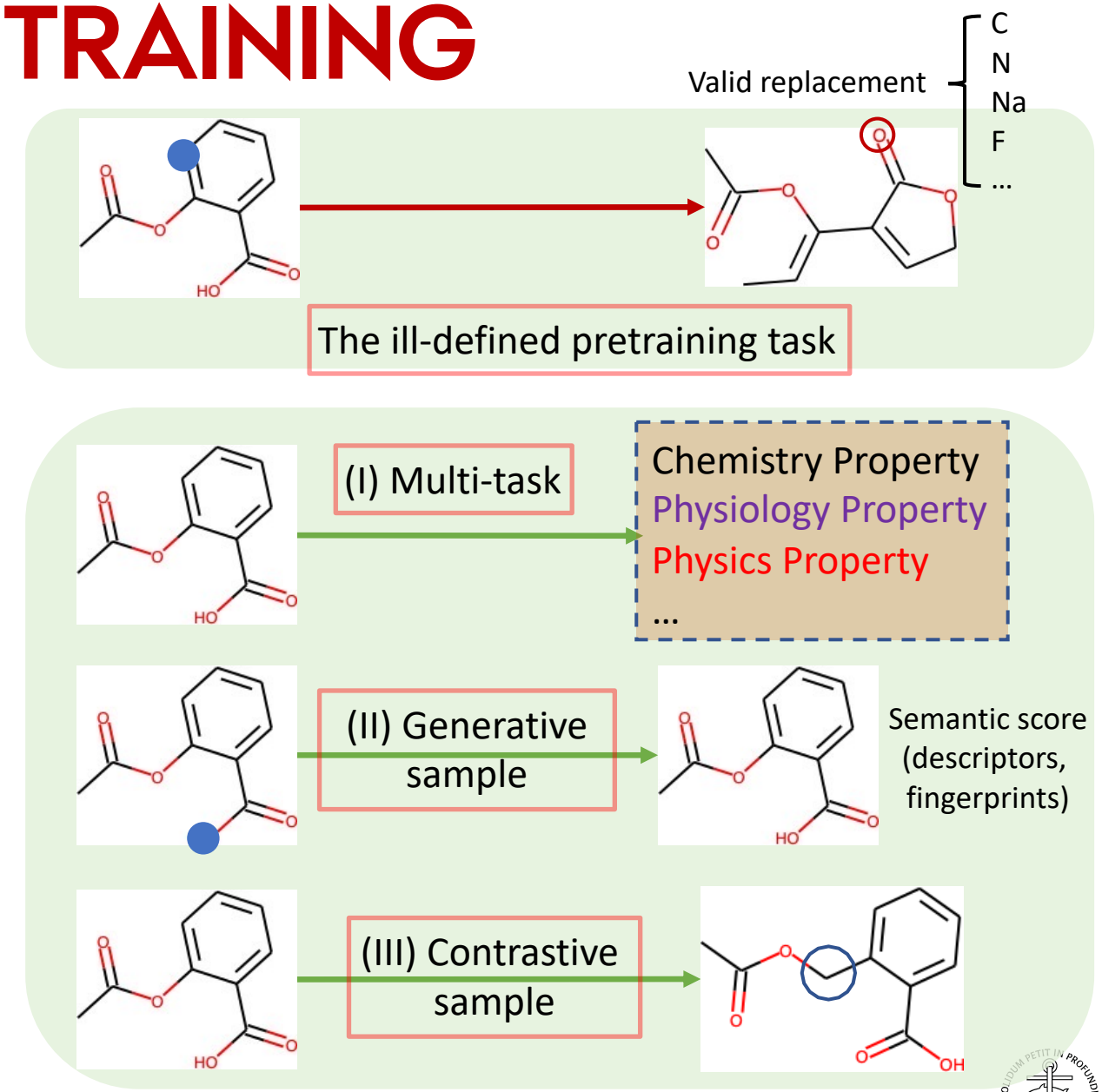


- Two augmented views are generated from local-level domain knowledge. Then, together with the original view (blue), they are fed into the GNN encoder and projection head to get global-level domain knowledge.

KNOWLEDGE IN PRE-TRAINING

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

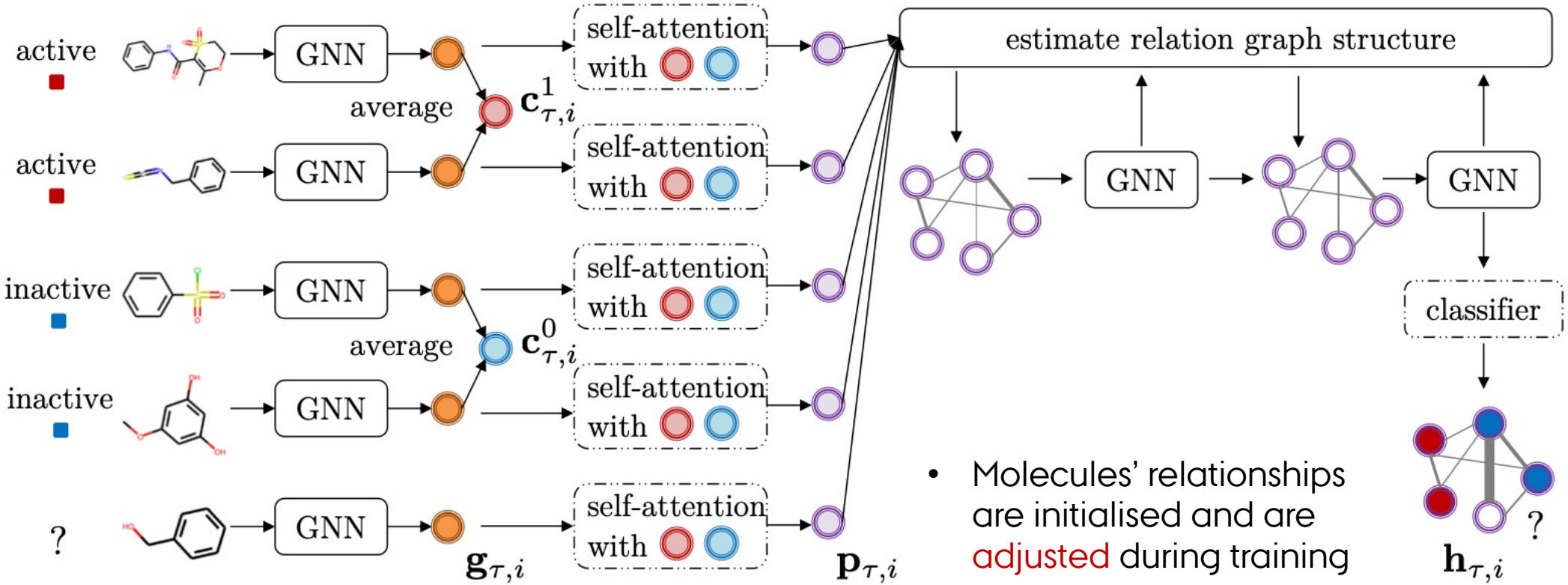
- **Knowledge-transfer**
  - Transferring external knowledge to construct pretraining objectives for GML models
- **Generative-sample**
  - Re-generating molecular graphs with the same biomedical semantics
- **Contrastive-sample**
  - Defining contrastive samples based on biomedical knowledge



INCORPORATING KNOWLEDGE IN TRAINING

PAR (NEURIPS'21)

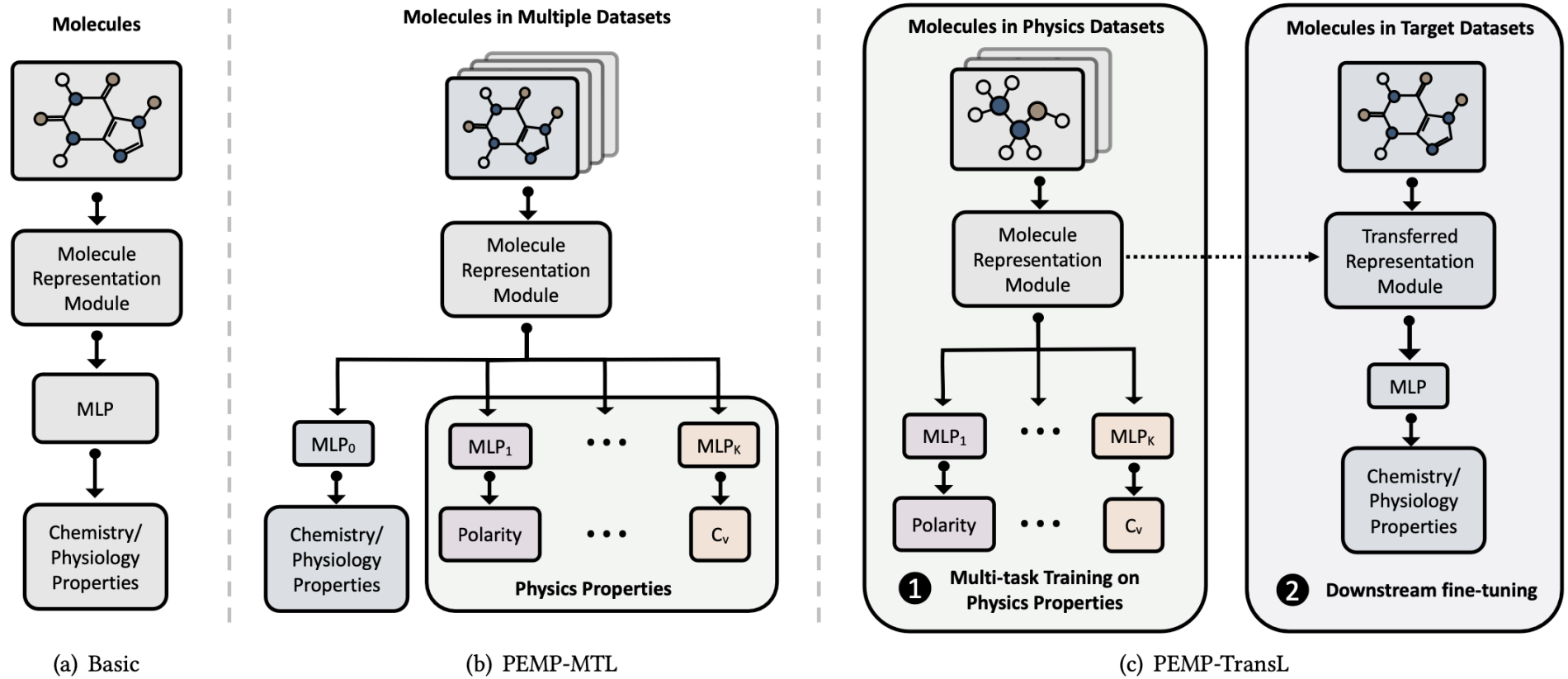
Training



- 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.

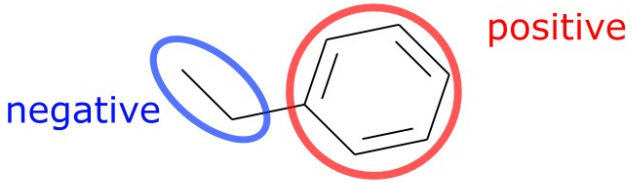
KEMPNN (ACS OMEGA'21)

Training

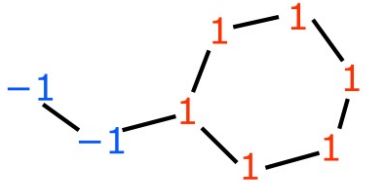
Knowledge annotation by human

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

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



Knowledge representation



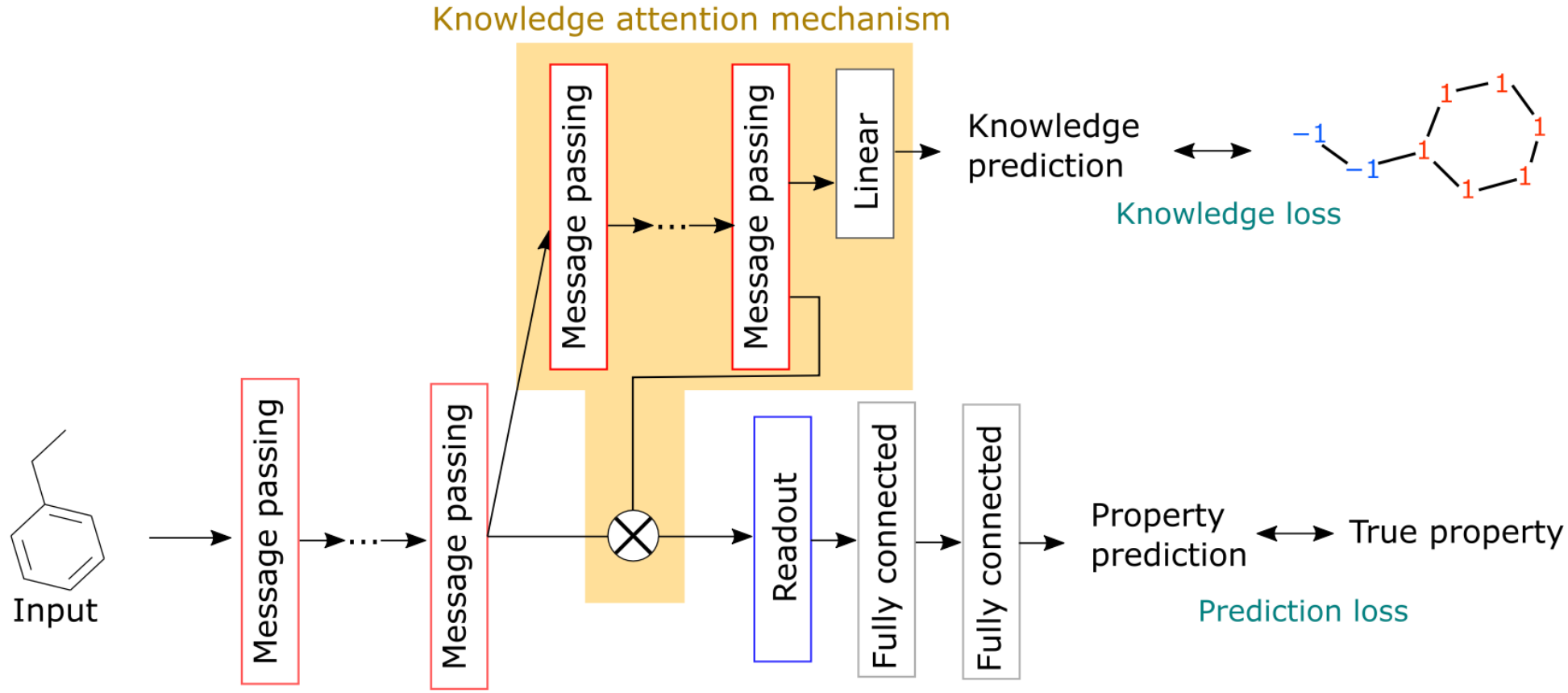
values on each atom (node)

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

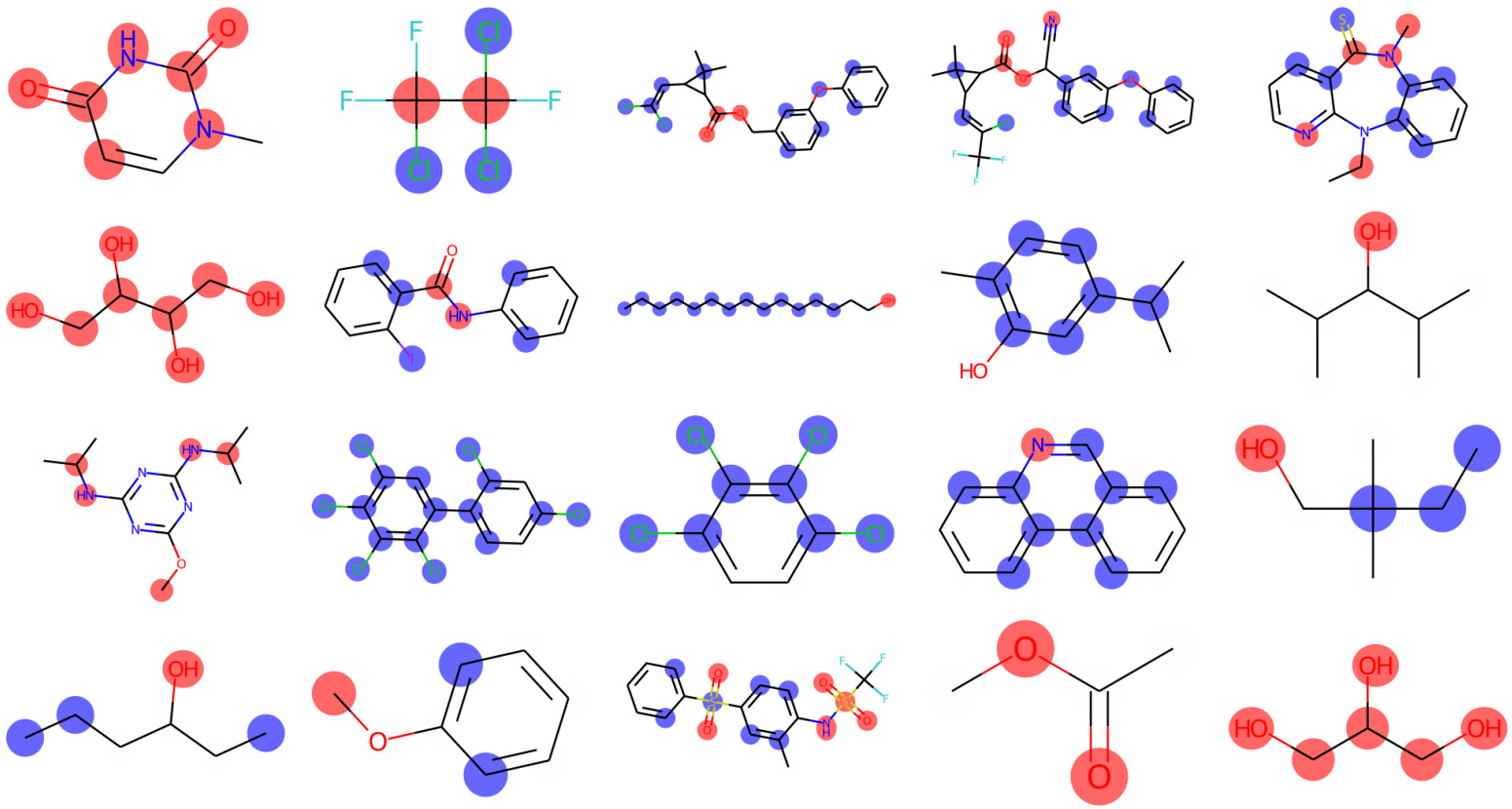
KEMPNN (ACS OMEGA'21)

Training



KEMPNN (ACS OMEGA'21)

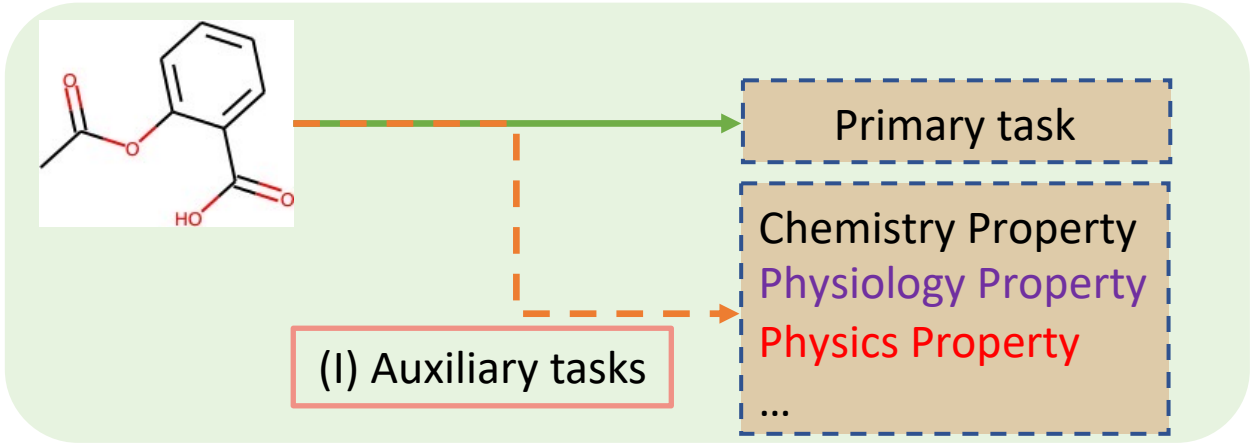
Training



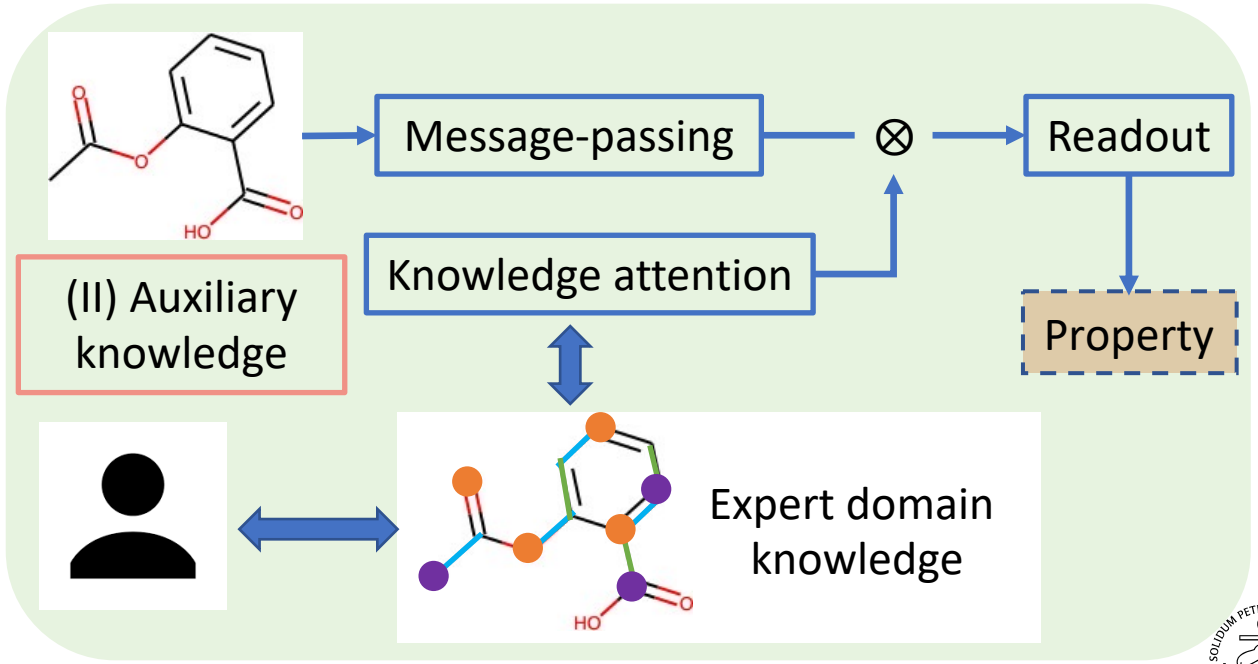
Visualisation of important compounds

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*

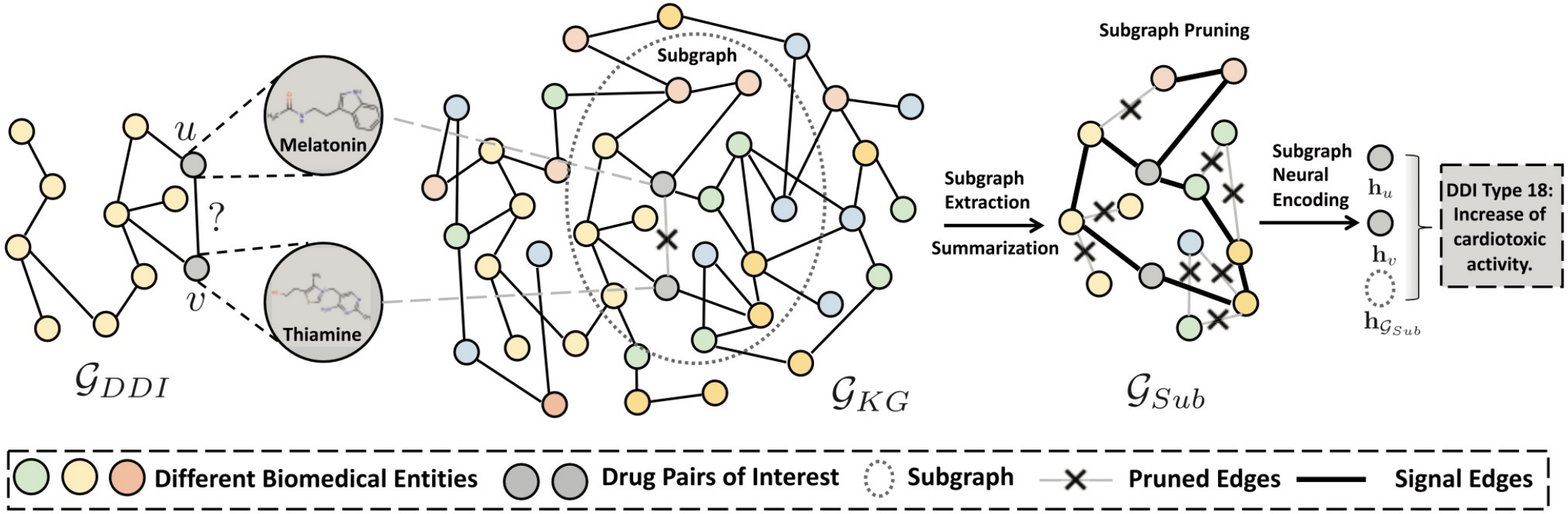
AARHUS UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

KDD'23 TUTORIAL
9 AUGUST 2023

ZHIQIANG ZHONG AND DAVIDE MOTTIN
DATA-INTENSIVE SYSTEMS GROUP

SUMGNN (BIOINFORM.'21)

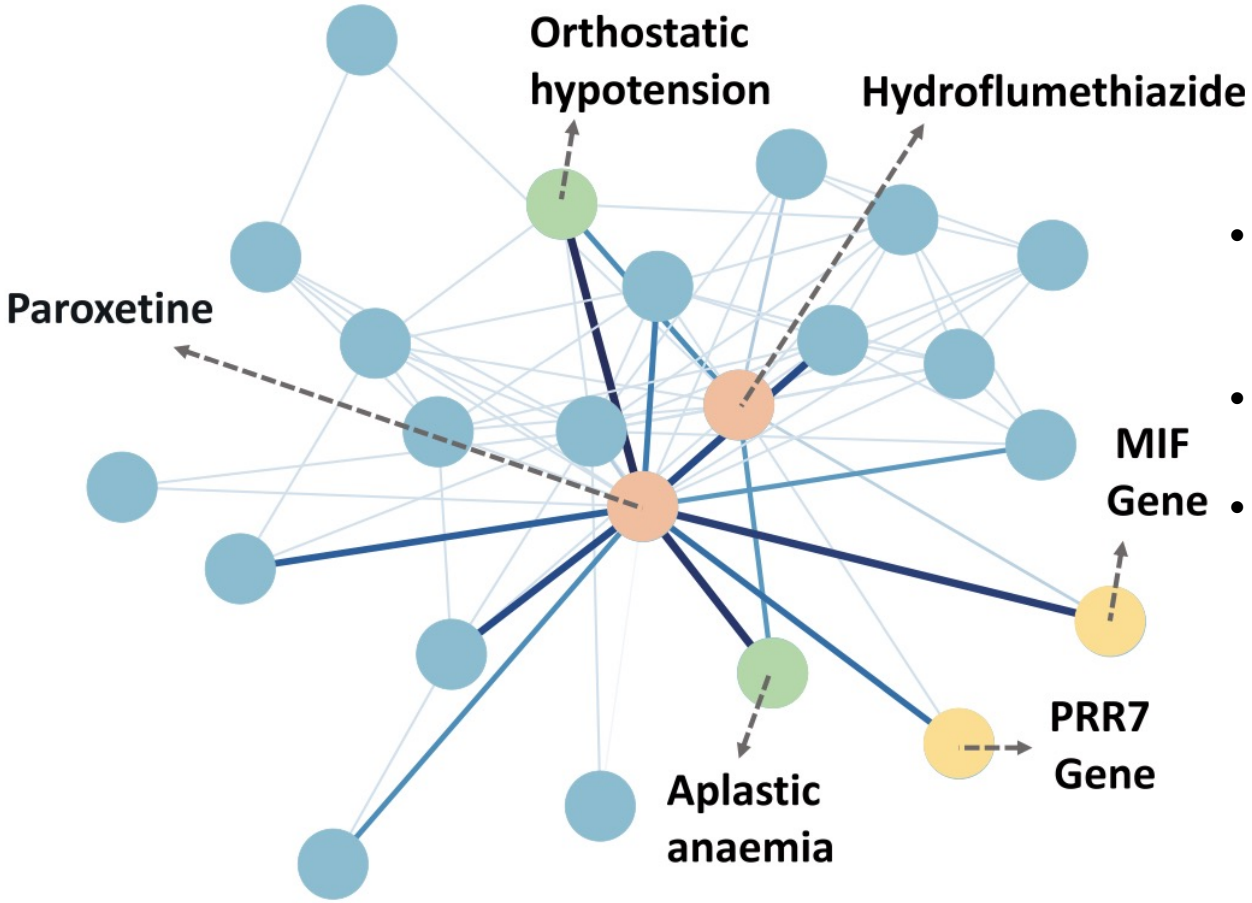
Interpretability



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

SUMGNN (BIOINFORM.'21)

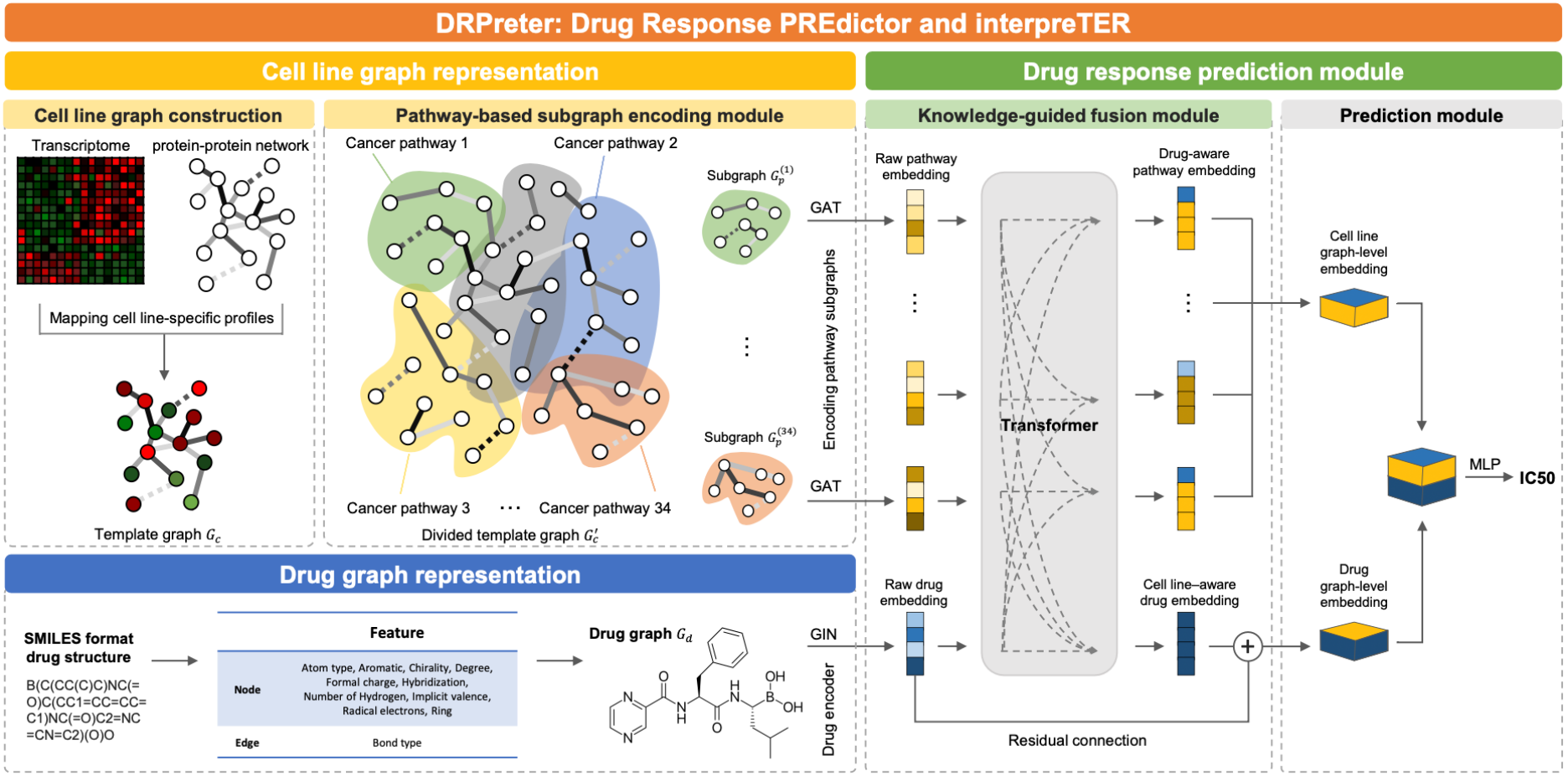
Interpretability



- 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.

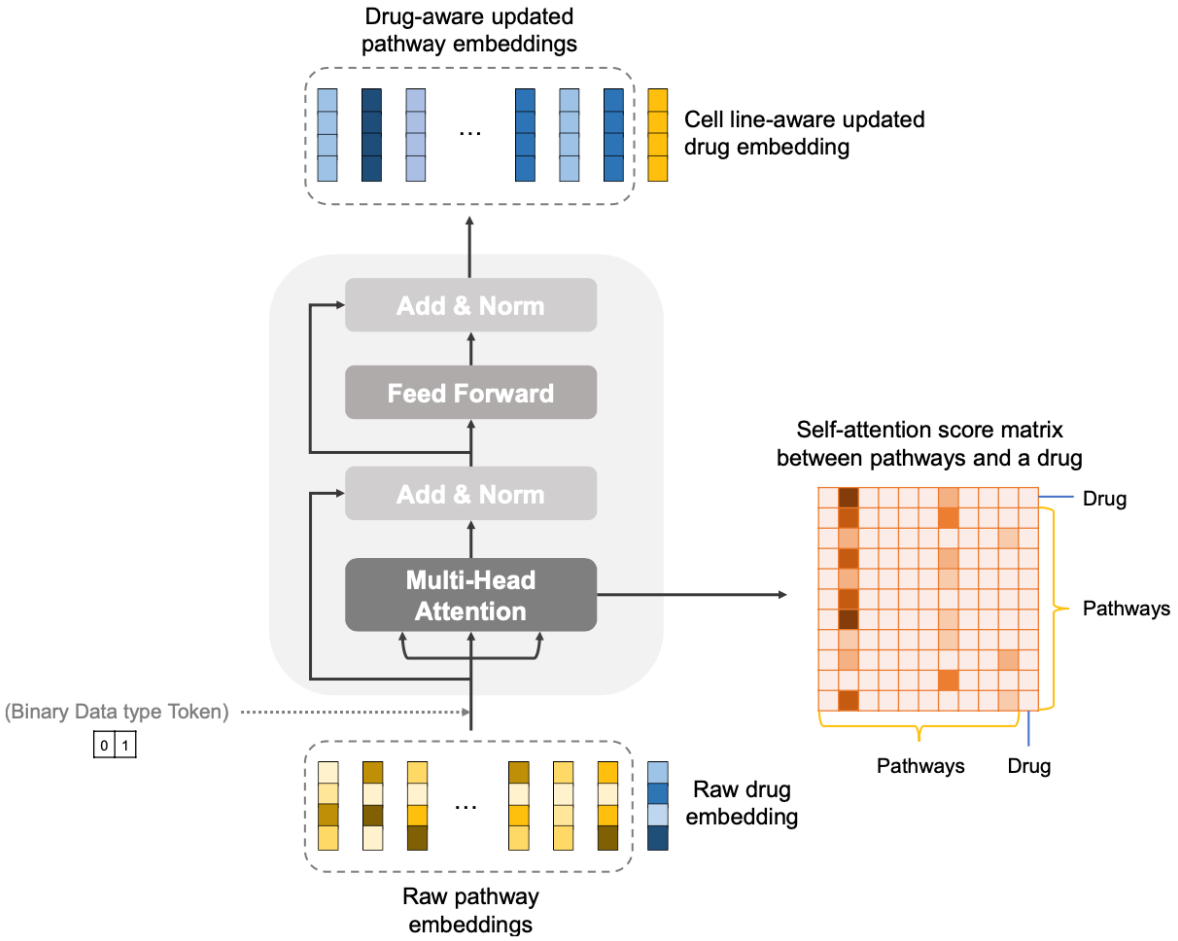
DRPRETER (NT. J. MOL. SCI.'22)

Interpretability



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

Interpretability



- Self-attention mechanism is applied to identify crucial pathways

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

Interpretability

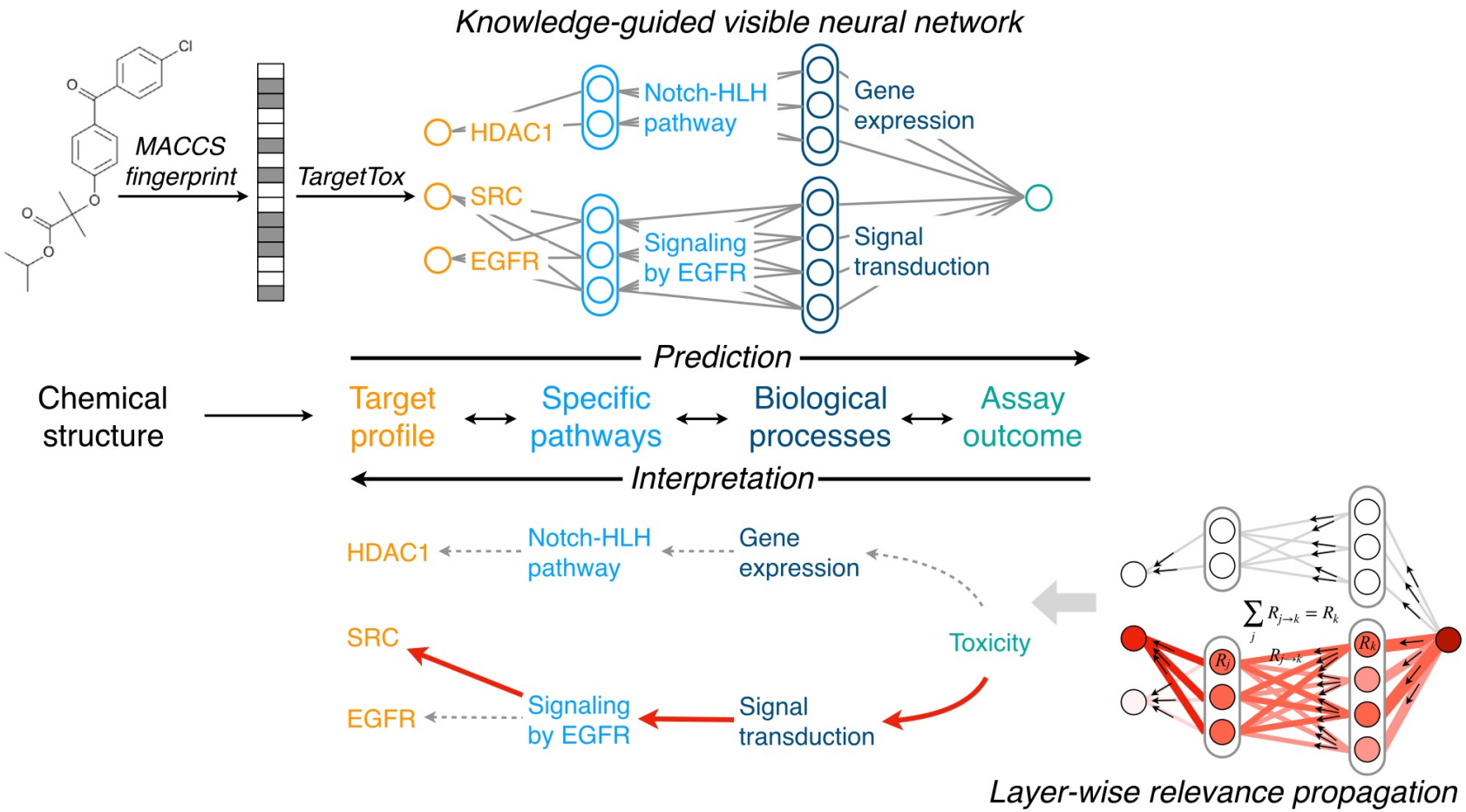
Drug	Cell Line	Disease	Top 5 Significant Genes	ln(IC50)	
				True	Predicted
Afatinib	GMS-10	Glioblastoma	*ACTR3B, PRR5, PRKCZ,* ***ERBB2****, LTBR*	0.5372	0.5324
Vinblastine	NCI-H1792	NSCLC	*CYP7A1, GTF2H2, DVL2, RAB5B,* ***TP53***	−5.9258	−5.27633
Docetaxel	PANC0327	Pancreatic cancer	***CLDN18****, 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.

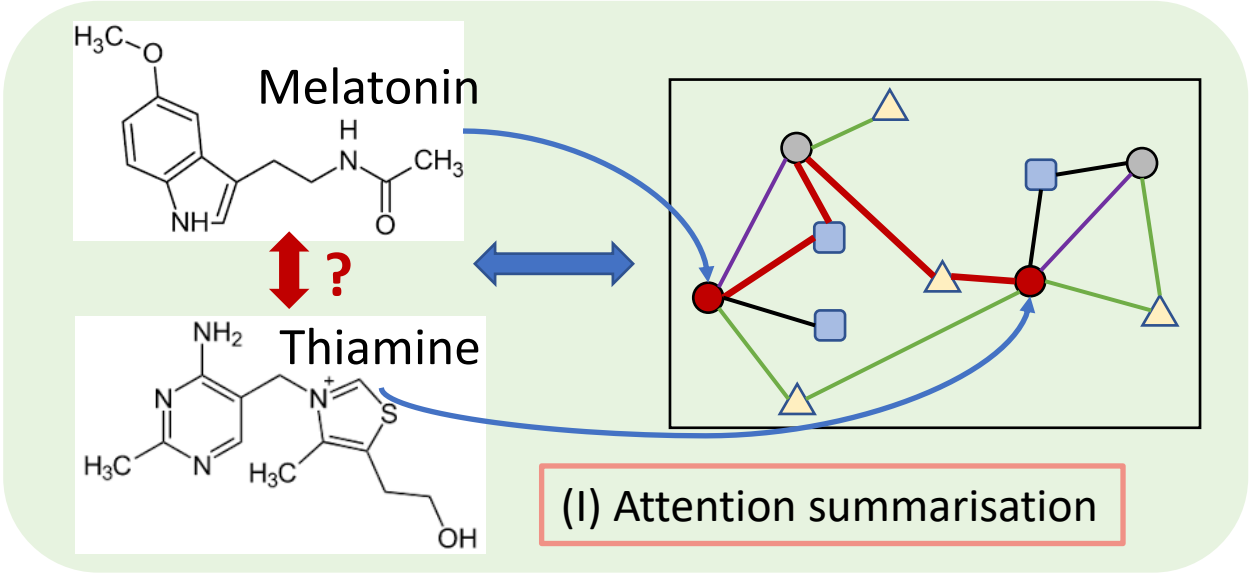
DTOX (PATTERNS'22)

Interpretability

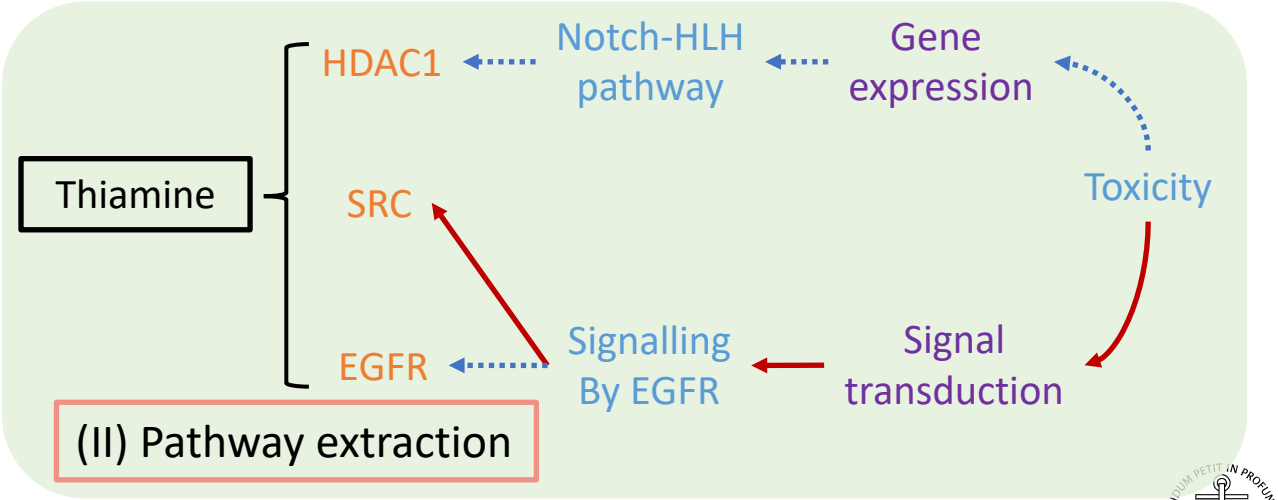


KNOWLEDGE IN INTERPRETABILITY

- **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.



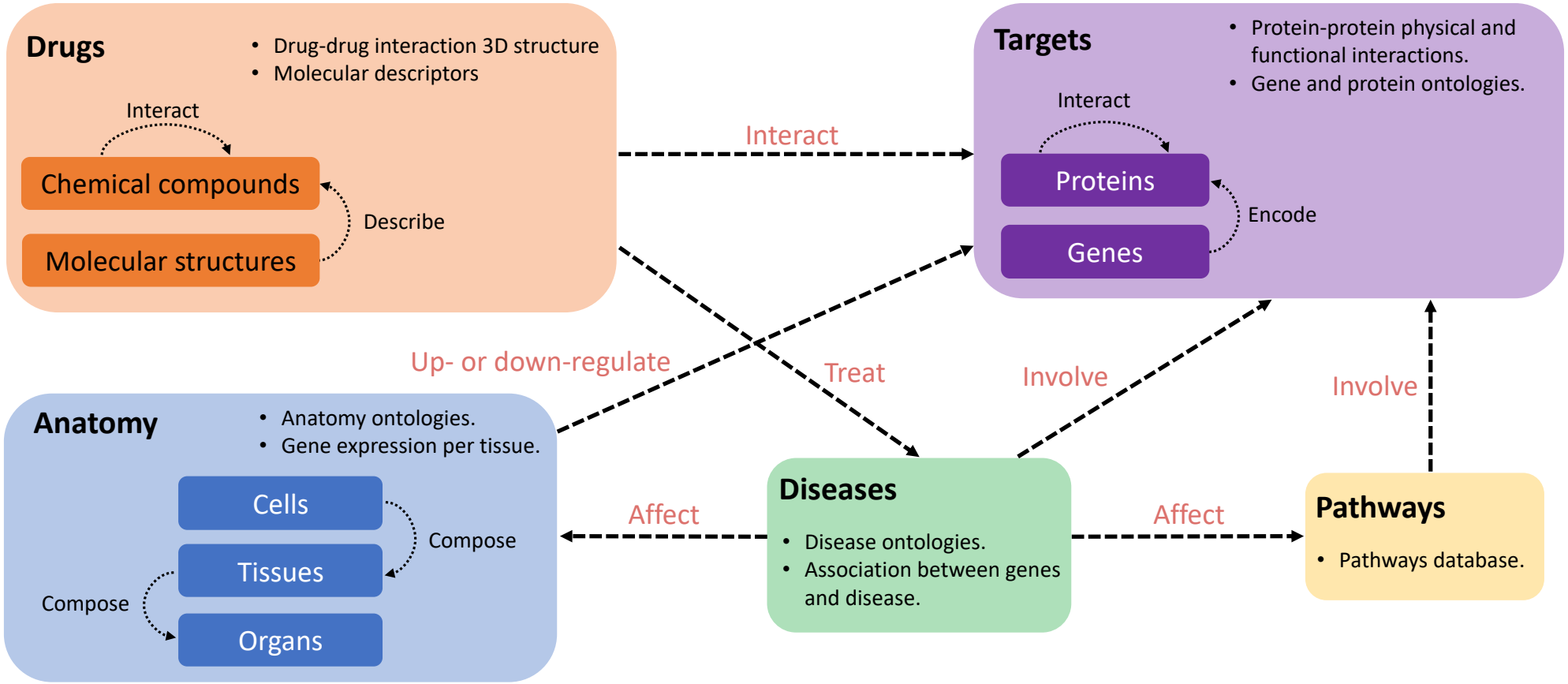
OUTLINE

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

SCHEMATIC REPRESENTATION OF MOLECULE DRUG KG CONSTRUCTION

SCHEMATA TO ORGANISE KNOWLEDGE DATABASES

- Knowledge database composition and compatibility



EXAMPLE BIOMEDICAL KNOWLEDGE DATABASES AND KNOWLEDGE GRAPHS

MOLECULAR & STRUCTURAL RESOURCES

Resource	Brief Description	Type
logP [190]	Measures of a molecule's hydrophobicity, or its partition coefficient between a nonpolar and polar solvent, and is commonly used to predict drug absorption and distribution.	Formula
rotatable bond [142]	Annotation of the (non)rotatable bond.	Formula
MolMap [151]	A method to visualise molecular structures in 3D by mapping atomic properties onto a 3D grid, allowing for the exploration and analysis of molecular interactions and properties.	Software
RDKit [177]	An open-source package to generate chemical features.	Software
UFF [178]	A molecular mechanics force field designed for the full periodic table.	Table
Mordred [191]	A tool for generating molecular descriptors, which are mathematical representations of molecular structures used for molecular property analysis.	Software
OpenBabel [192]	An open-source molecular modelling software that provides a comprehensive toolkit for molecular conversion, visualisation, and analysis.	Software
MoleculeNet [193]	A benchmark for molecular machine learning, comparing models performances on various molecular property prediction tasks such as solubility, melting point, and binding affinity.	Database
Ptable [188]	A periodic table of chemical elements classified by atomic number, electron configurations, and chemical properties into groups and periods, providing a systematic overview of elements.	Table

- There are a number of scientific tools to generate molecular features

COMPOUNDS AND DRUG & TRAGET

Resource	Brief Description	Type
	Compounds	
CheMBL [194]	A database of bioactive molecules, assays, and potency information for drug discovery and pharmaceutical research, used to facilitate target identification and selection.	Database
PubChem [195]	Open database of chemical substances that contains information on their 2D and 3D structures, identifiers, properties, biological activities and occurrence in nature.	Database
ChEBI [196]	An open-source resource for molecular biology and biochemistry, providing a systematic and standardised vocabulary of molecular entities focused on small chemical compounds.	Ontology Database
KEGG Compound [197]	A database of small molecular compounds, including their structures, reactions, pathways, and functions, used to provide information on metabolic pathways and cellular processes.	Database
DrugBank [179]	A database includes small molecular compounds, biologics, and natural products, providing information on their properties, mechanisms, and interactions used in drug discovery.	Database
	Drugs and Targets	
DDinter [198]	A database of protein-protein interactions, providing information on protein targets, their interactions, and related diseases, used to advance drug discovery and development.	Database
TCRD [199]	Database that aggregates information on proteins targeted by drugs and attributes them a development/druggability level.	Database
OpenTargets [200]	A database that integrates diverse genomic and molecular data to provide a comprehensive view of the relationships between diseases, genes, and molecular targets.	Database
TTD [201]	A publicly available database that provides information on protein and nucleic acid targets, drugs that target them and related diseases, used to advance drug discovery and development.	Database
PharmGKB [202]	A resource that provides information on the impact of human genetic variation on drug response, used to advance precision and personalised drug therapy.	Database
e-TSN [203]	A platform that integrates knowledge on disease-target associations used for target identification. These associations were extracted from literature by using NLP techniques.	Web platform
nSIDES [204]	Multiple resources made available by the Tatonetti lab on drug side effects, drug-drug interactions and pediatric drug safety.	Database
SIDER [205]	A database of marketed drugs and their side effects, providing information on the frequency, type, and severity of adverse events, used to advance drug safety and pharmacovigilance.	Database

- There are a number of knowledge databases about different biomedical entities

GENE & PROTEIN AND PATHWAYS

Resource	Brief Description	Type
	Genes and Proteins	
GeneOntology [180]	A structured and standardised ontology of gene functions, used to describe and categorise genes and gene products function in a consistent and interoperable manner.	Ontology
Entrez [206]	A database that includes nucleotide and protein sequences, genomic maps, taxonomy, and chemical compounds by referencing other databases, used to query various biomedical data.	Database
Ensembl [207]	A database that provides information on annotated genes, multiple sequence alignments and disease for a variety of species, including humans.	Database
KEGG Genes [197]	A database that provides information on genes for complete genomes, their associated pathways, and functions in various organisms.	Database
BioGRID [208]	A database of protein and genetic interactions curated from high-throughput experimental data sources in a variety of organisms. It includes a tool to create graphs of interactions.	Database
UniProt [209]	A database of protein information, including their sequences, structure, structure and post-translational modifications.	Database
STRING [210]	A database of protein-protein interactions and functional associations, integrating diverse data sources and evidence to provide a weighted network of functional relationships.	Database
HumanNet [211]	Network of protein-protein and functional gene interactions, constructed by integrating high -throughput datasets and literature, used to advance understanding of disease gene prediction.	Database
STITCH [212]	A database of known and predicted interactions between chemicals and proteins (physical and functional associations), used for the study of molecular interactions.	Database
PDB [213]	A database that provides information on the 3D structure of proteins, nucleic acids, and complex molecular assemblies, obtained experimentally or predicted.	Database
RNAcentral [214]	A repository that integrates information on non-coding RNA sequences for a variety of organisms and attributes them to a unique identifier.	Database
	Pathways	
Reactome [215]	A database that stores and curates information about the molecular pathways in humans, providing insights into cellular processes and disease mechanisms.	Database
KEGG pathways [197]	A database of curated biological pathways and interconnections between them, manually represented as pathway maps of molecular reactions and interactions.	Database
WikiPathways [216]	A database of biological pathways that integrates information from several databases, which aims to provide an overview of molecular interactions and reactions.	Database

- There are a number of knowledge databases about different biomedical entities

DISEASE

Resource	Brief Description	Type
DO [217] (Disease Ontology)	Disease Ontology (DO) is an ontology of human disease that integrates MeSH, ICD, OMIM, NCI Thesaurus and SNOMED nomenclatures.	Ontology
MonDO [218]	Semi-automatic unifying terminology between different disease ontologies.	Ontology
Orphanet [219]	A database that maintains information on rare diseases and orphan drugs using cross-references to other commonly used ontologies.	Database
OMIM [220]	A comprehensive, searchable database of gene-disease associations for Mendelian disorders.	Database
KEGG Disease [197]	A database of disease entries that are characterised by their perturbants (genetic or environmental factors, drugs, and pathogens).	Database
ICD-11 [221]	The 11th version of the international resource for recording health and clinical data in a standardised format that is constantly updated.	Ontology
Disgenet [222]	A database that integrates manually curated data from GWAS studies, animal models, and scientific literature to identify gene-disease associations. It can be used for target identification and prioritisation.	Database
DISEASES [223]	A database for disease-gene associations based on manually curated data, cancer mutation data, GWAS, and automatic text mining.	Database
GWAS Catalog [224]	Repository of published Genome-Wide Association Studies (GWAS) for investigating the impact of genomic variants on complex diseases.	Database
SemMedDB [225]	A database that provides information on the relationships between genes and diseases, extracted from the biomedical literature.	Database
OncoKB [226]	A knowledge precision database containing information on human genetic alterations detected in different cancer types.	Database
HPO [227]	The Human Phenotype Ontology (HPO) is an ontology of human phenotypes and database of disease-phenotype associations with cross-references to other relevant databases.	Ontology

PUBLICLY AVAILABLE KGS

Resource	Brief Description	Intended Usage
Hetionet [233]	An integrated KG of more than 12,000 nodes representing various biological, medical and social entities and their relationships. It is a valuable resource combining many different databases that can be used for drug discovery and repurposing.	Drug discovery Drug repurposing etc.
PharmKG [234]	A comprehensive biomedical KG integrating information from various databases, literature, and experiments. It is mainly centered around interactions between genes, diseases and drugs.	Drug discovery
DRKG [235]	A large-scale, cross-domain KG that integrates information about drugs, proteins, diseases, and chemical compounds. It is based on Hetionet, and it was used for drug repurposing for Covid-19.	Drug repurposing
CKG [236]	A KG developed for precision medicine that combines various databases and integrates clinical and omics data. It allows for automated upload and integration of new omics data with pre-existing knowledge.	Biomarker discovery Drug prioritisation.
OpenBioLink [237]	An open-source KG that integrates diverse biomedical data from various databases. It was developed to enable benchmarking of ML algorithms.	Drug discovery
BioKG [238]	A KG that integrates information about genes, proteins, diseases, drugs, and other biological entities. It aims at providing a standardised KG in a unified format with stable IDs.	Pathway discovery Drug discovery
Bioteque [64]	A KG that enables the discovery of relationships between genes, proteins, diseases, drugs, and other entities, providing an overview of biological knowledge for use in biomedical research and personalised medicine.	Broad usage
Harmonizome [239]	A KG that focuses on gene- and protein-centric information and their interactions. It provides a unified view of biological knowledge and enables the discovery of new insights fin the biomedical field.	Drug discovery Precision medicine

- Some well-organised KGs are publicly available for research

REPRESENTATIVE KAGML PAPERS

Method	Venue	Year	Task	Knowledge Usage Area			
				Preprocessing	*Pre-training*	*Training*	*Interpretability*
MPNN [137]	ICML	2017	PREDMOL	✓			
D-MPNN [138]	J. Chem. Inf. Model.	2019	PREDMOL	✓			
CMPNN [139]	IJCAI	2020	PREDMOL	✓			
KGNN [140]	IJCAI	2020	PAIRDRUGDRUG	✓			
MaSIF [141]	Nat. Methods	2020	PAIRPRTPRT	✓			
KEMPNN [142]	ACS Omega	2021	PREDMOL	✓	✓	✓	
SumGNN [143]	Bioinform.	2021	PAIRDRUGDRUG	✓			✓
FraGAT [144]	Bioinform.	2021	PREDMOL	✓			
PAINN [145]	ICML	2021	PREDMOL	✓			
MDNN [146]	IJCAI	2021	PAIRDRUGDRUG	✓			
MoCL [147]	KDD	2021	PREDMOL		✓		
AlphaFold [148]	Nature	2021	PREDPRT	✓		✓	
KGE_NFM [149]	Nat. Commun.	2021	PAIRDRUGTGT	✓			
scGCN [150]	Nat. Commun.	2021	PREDCELL	✓			
MolMapNet [151]	Nat. Mach. Intell.	2021	PREDMOL	✓			
GemNet [152]	NeurIPS	2021	PREDMOL	✓			
HOLOPROT [153]	NeurIPS	2021	PAIRPRTPRT	✓			
SynCoor [154]	NeurIPS	2021	PREDMOL	✓			
KCL [155]	AAAI	2022	PREDMOL		✓		
SGNN-EBM [156]	AISTATS	2022	PREDMOL	✓			
scGraph [157]	Bioinform.	2022	PREDGE	✓			
DTI-HETA [158]	Brief. Bioinform.	2022	PAIRDRUGTGT	✓			
PEMP [159]	CIKM	2022	PREDMOL		✓	✓	
MISU [160]	CIKM	2022	PREDMOL		✓		
GraphMVP [161]	ICLR	2022	PREDMOL		✓		
OntoProtein [162]	ICLR	2022	PREDPRT		✓		
SphereNet [163]	ICLR	2022	PREDMOL	✓			
3DInfoMax [164]	ICML	2022	PREDMOL		✓		
DRPreter [165]	Int. J. Mol. Sci.	2022	PREDDRUG	✓			✓
DENVIS [166]	J. Chem. Inf. Model.	2022	PAIRPRTPRT	✓			
ReLMole [167]	J. Chem. Inf. Model.	2022	PAIRDRUGDRUG	✓			
KPGT [168]	KDD	2022	PREDMOL		✓		
NequIP [169]	Nat. Commun.	2022	PREDMOL	✓			
GEM [170]	Nat. Mach. Intell.	2022	PREDMOL		✓		
ComENet [171]	NeurIPS	2022	PREDMOL	✓			
DTox [172]	Patterns	2022	PREDDRUG				✓
ProteinMPNN [173]	Science	2022	ACTPRT	✓			
KEMV [174]	TKDE	2022	PAIRDRUGTGT	✓			
KG-MTL [175]	TKDE	2022	PAIRMOLMOL	✓		✓	
HIGH-PPI[176]	Nat. Commun.	2023	PAIRPRTPRT	✓			

- A set of collected KaGML papers are carefully categorised into different categories based on our proposed taxonomy.

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

OPEN CHALLENGES

1. Knowledge database composition and compatibility
   i. The effectiveness of KaGML methods heavily relies on the availability of qualified knowledge databases that can provide comprehensive and sufficient information.
   ii. Harmonisation and integration of data still pose a significant challenge, as these resources are often diverse, heterogeneous, and distributed across multiple platforms. As such, addressing the lack of standardisation in data integration is a critical area for future research to enhance the power of KaGML.
   iii. Many biomedical knowledge databases have to be frequently updated and refined to stay up to date with the current research, which presents a challenge for KaGML methods. To address this, it is recommended that KaGML works store the versions of the databases used in their experiments for better reproducibility.
   iv. Important principles: FAIR[1].

[1] *Mark D Wilkinson et al.*, The fair guiding principles for scientific data management and stewardship. Scientific Data, 2016

OPEN CHALLENGES

2. Effective knowledge integration with uncertainty
   i. KaGML works have incorporated external knowledge into *preprocessing*, *pretraining*, *training*, and *interpretability* for drug discovery.
   ii. However, these approaches are typically deterministic, ignoring the underlying uncertainty of knowledge and its impact on model learning and inference.
   iii. Thus, it is an important area of future research to investigate how to effectively and systematically model knowledge uncertainties for real-world applications.

OPEN CHALLENGES

3. Advanced interpretability & careful evaluation benchmark
   i. The enhancement of the interpretability of AI models has the potential to increase the confidence and dependability of patients, as well as to enhance the applicability of the models. Nevertheless, there remains a vast scope for further research in the area of advanced interpretability, with the aim of providing more holistic and adaptable explanations, such as advanced reasoning and question-answering capabilities.
   ii. Designing a comprehensive validation pipeline, such as an explanation verification pipeline, is a promising area for future research. While KaGML approaches have been developed to address interpretability problems in drug discovery, the question of how to verify and evaluate the generated explanations remains open.

OPEN CHALLENGES

4. From drug discovery to more biomedical fields
   i. While this tutorial focuses on the recent advancements in drug discovery, other fields of biomedical research could benefit from the expanding use of KaGML techniques, including target identification and validation and gene and cell therapy.
   ii. It would be interesting to see the development of a unified KaGML framework that supports diverse healthcare services.

OPEN CHALLENGES

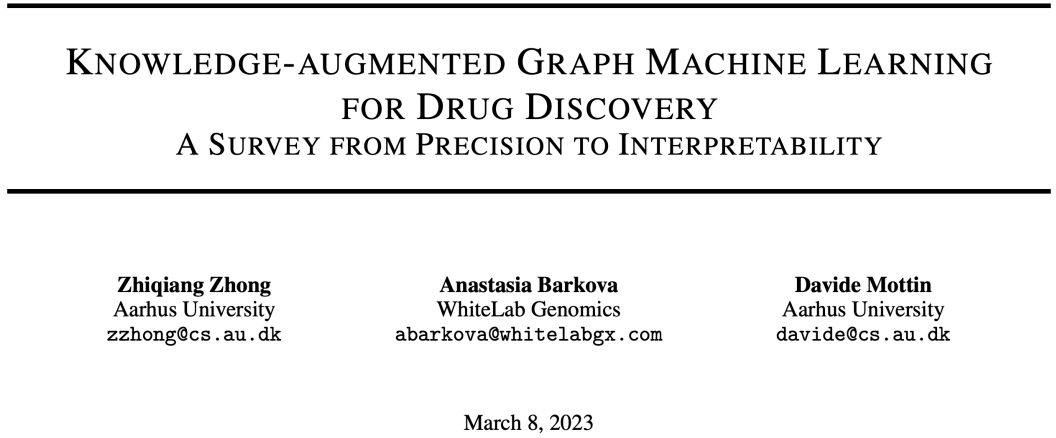
5. Security & privacy and efficiency & scalability
   i. The advancements in machine learning and growth in computational capacities have transformed the technology landscape but have also raised concerns about security and privacy.
   ii. This includes guaranteeing the ownership of knowledge databases, protecting patient-sensitive information, and ensuring the viability of models against malicious attacks.

OPEN CHALLENGES

1. Knowledge database composition and compatibility
2. Effective knowledge integration with uncertainty
3. Advanced interpretability & careful evaluation benchmark
4. From drug discovery to more biomedical fields
5. Security & privacy and efficiency & scalability

NEXT?

- A related survey manuscript is available online at:
  https://arxiv.org/abs/2302.08261



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

Zhiqiang Zhong
Aarhus University
zzhong@cs.au.dk

Anastasia Barkova
WhiteLab Genomics
abarkova@whitelabgx.com

Davide Mottin
Aarhus University
davide@cs.au.dk

March 8, 2023

- Detailed information about KaGML survey/papers/practical resources:
  https://github.com/zhiqiangzhongddu/Awesome-Knowledge-augmented-GML-for-Drug-Discovery

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