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

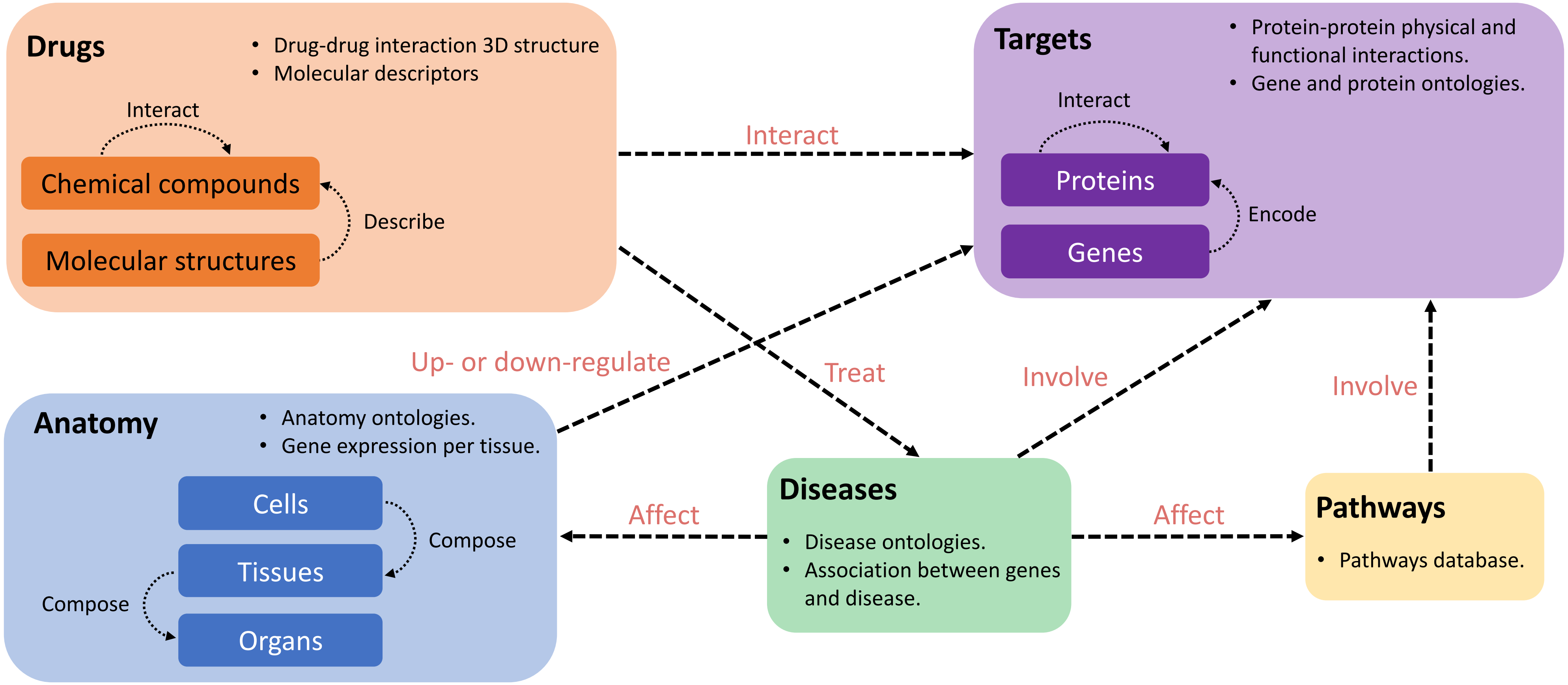
V. Practical Resources

VI. Open Challenges and Future Directions

SCHEMATIC REPRESENTATION OF MOLECULE DRUG KG CONSTRUCTION

SCHEMATA TO ORGANISE KNOWLEDGE DATABASES

- Knowledge database composition and compatibility



Drugs

- Drug-drug interaction 3D structure
- Molecular descriptors

Targets

- Protein-protein physical and functional interactions.
- Gene and protein ontologies.

Anatomy

- Anatomy ontologies.
- Gene expression per tissue.

Diseases

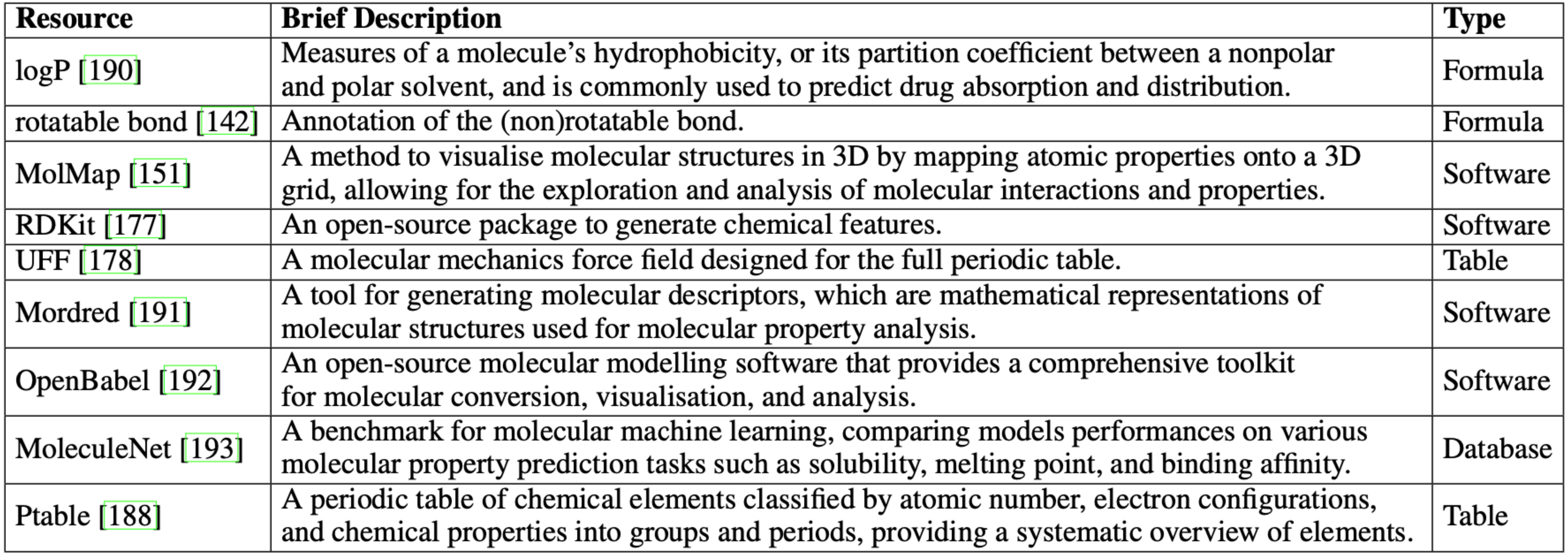
- Disease ontologies.
- Association between genes and disease.

Pathways

- Pathways database.

EXAMPLE BIOMEDICAL KNOWLEDGE DATABASES AND KNOWLEDGE GRAPHS

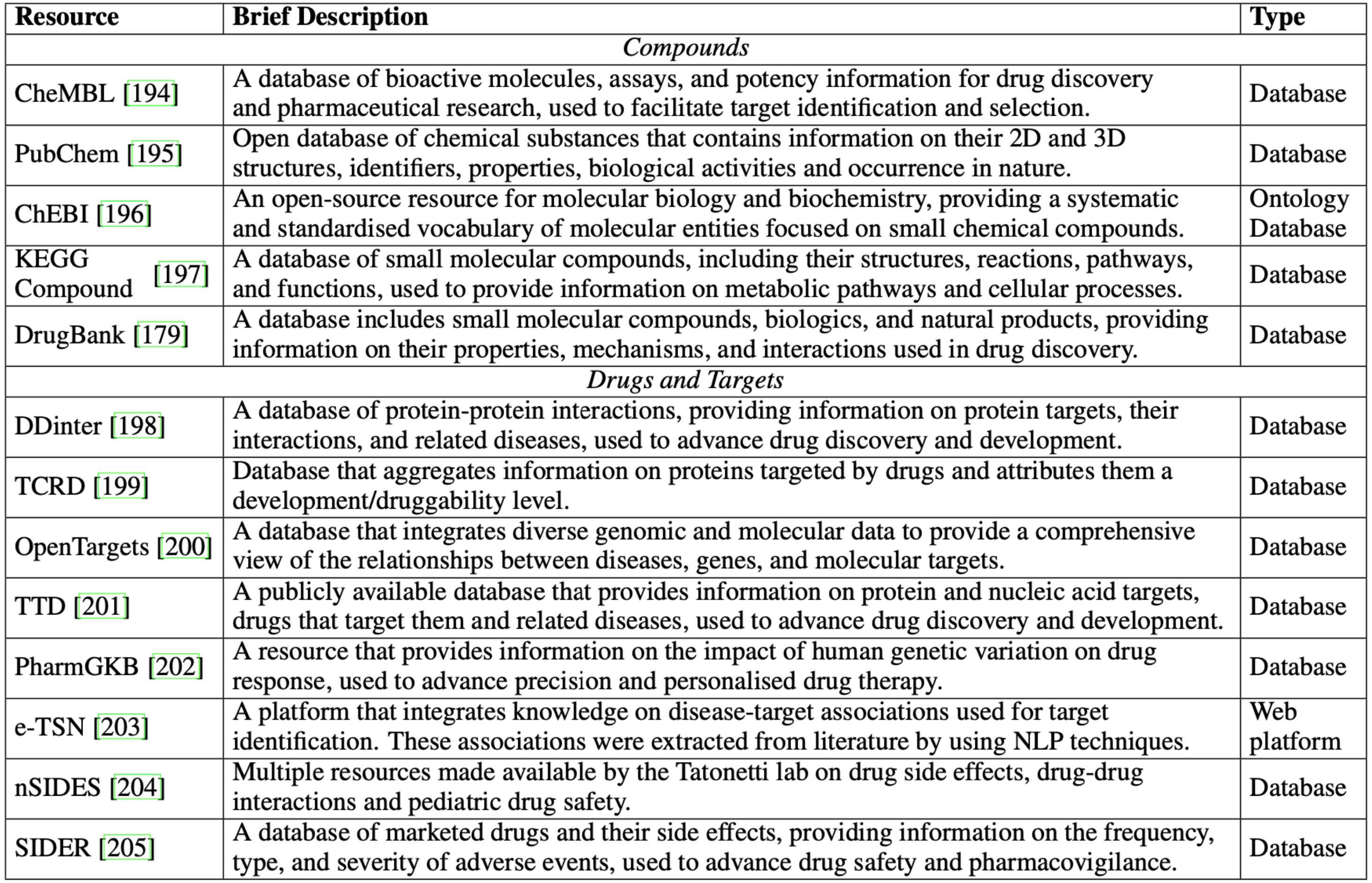
MOLECULAR & STRUCTURAL RESOURCES



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

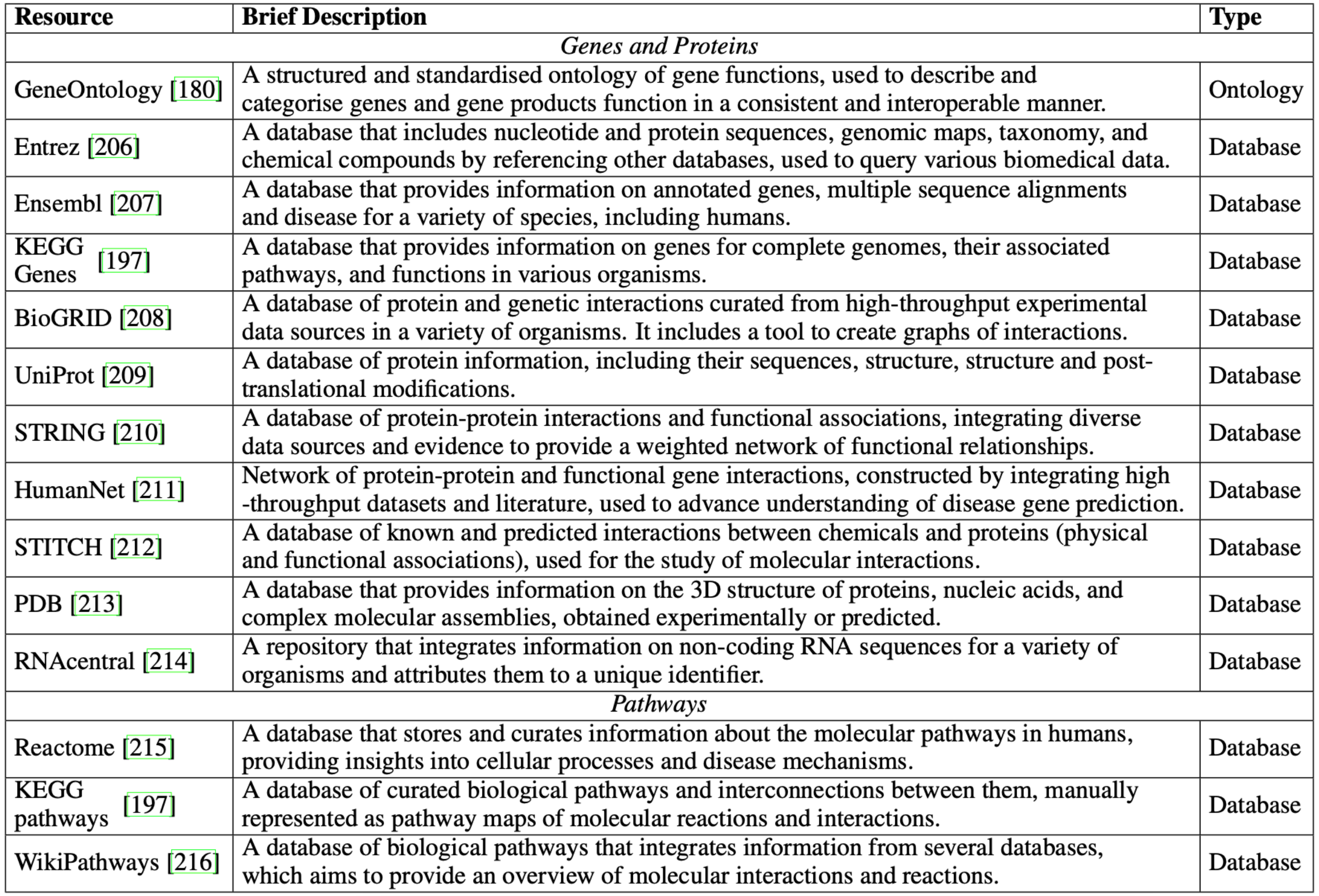
COMPOUNDS AND DRUG & TRAGET



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

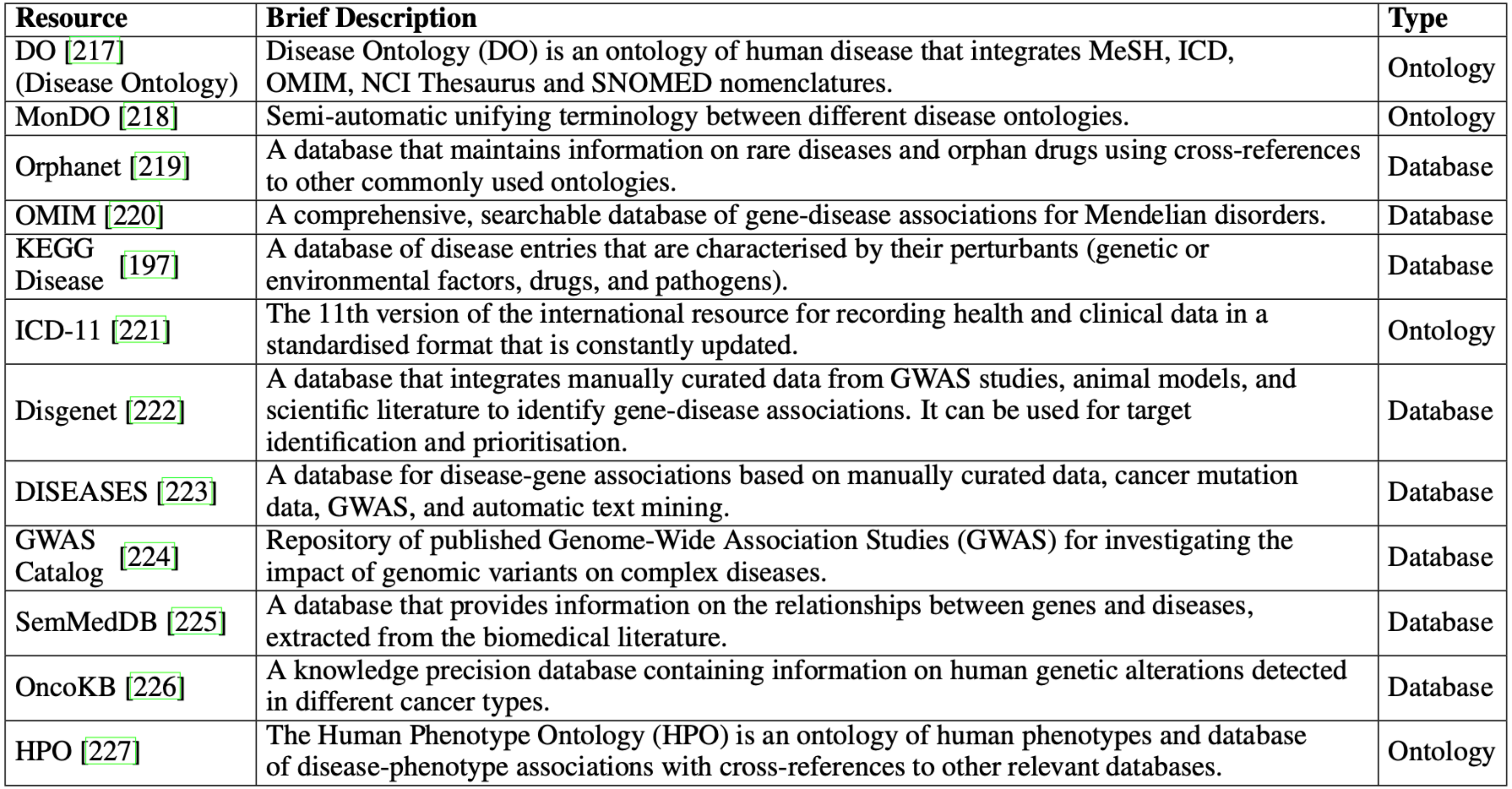
GENE & PROTEIN AND PATHWAYS



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

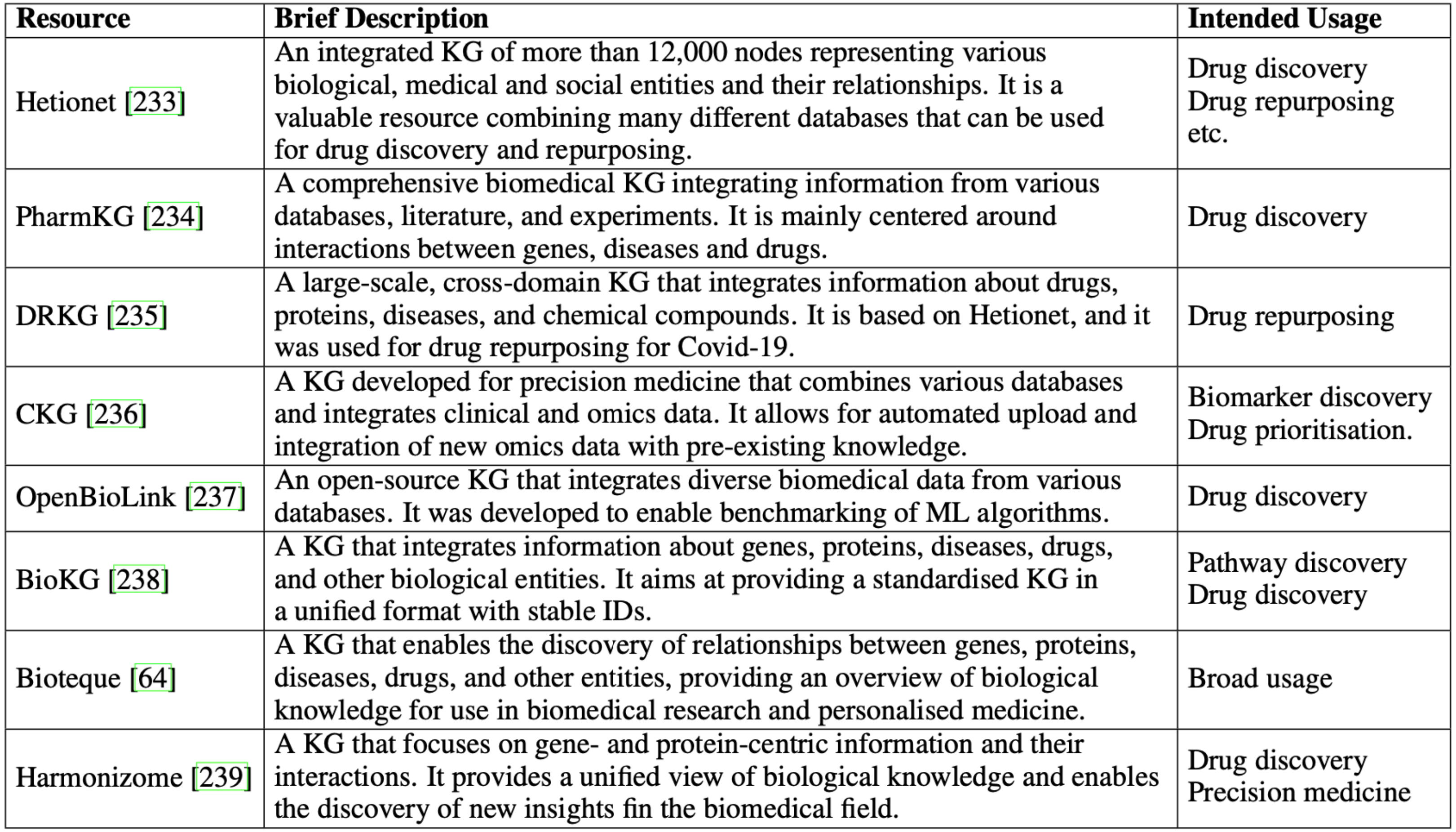
- There are a number of knowledge databases about different biomedical entities

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

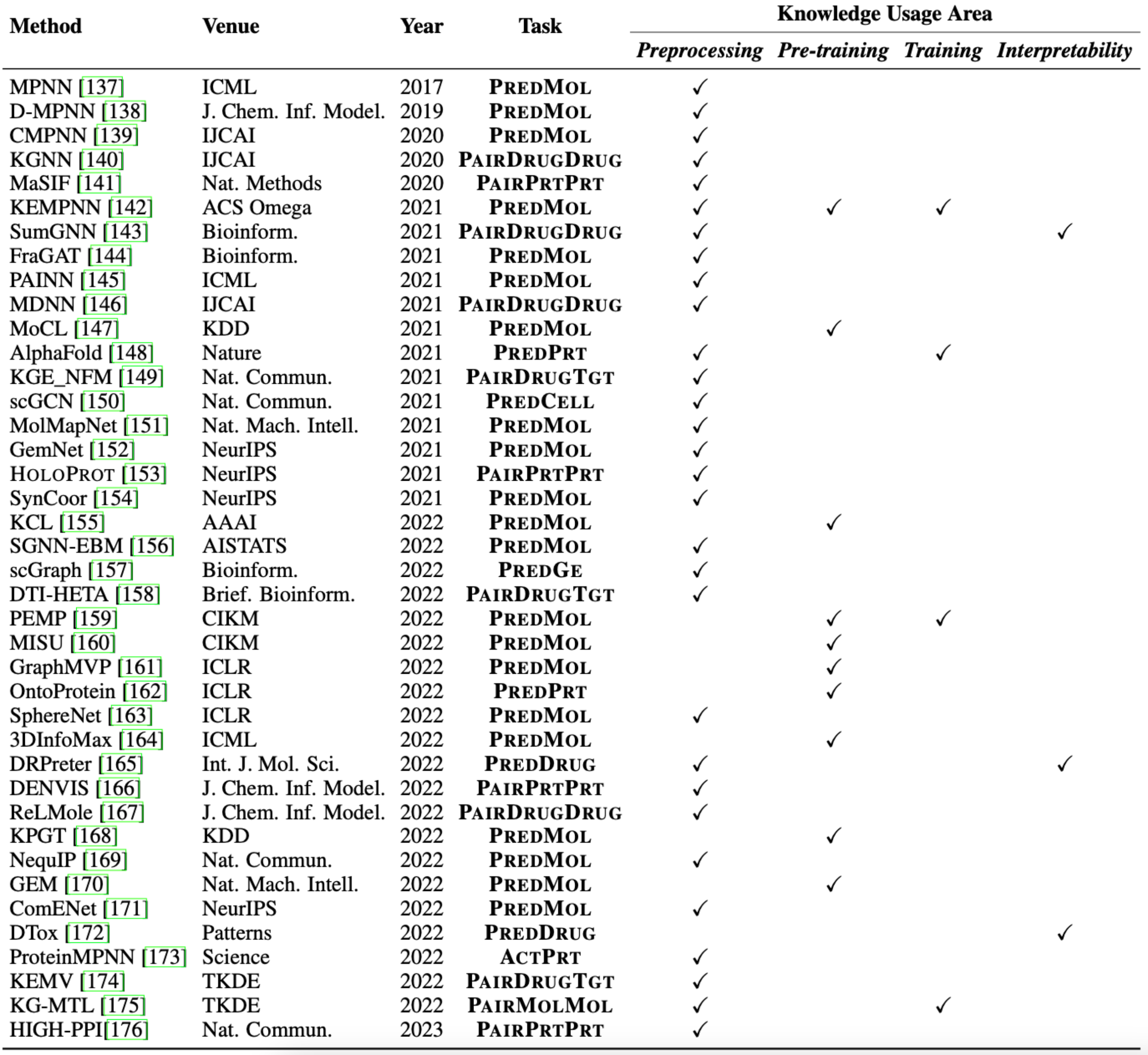
PUBLICLY AVAILABLE KGS



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.

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