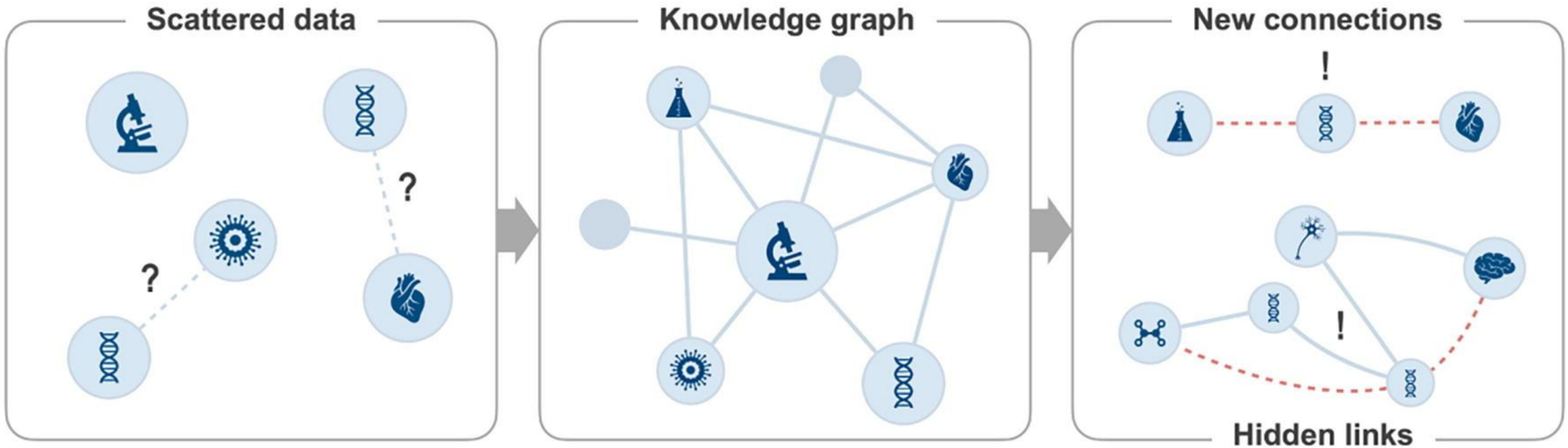
FAIR principles for knowledge graphs in SSbD

Angela Serra

Tampere University

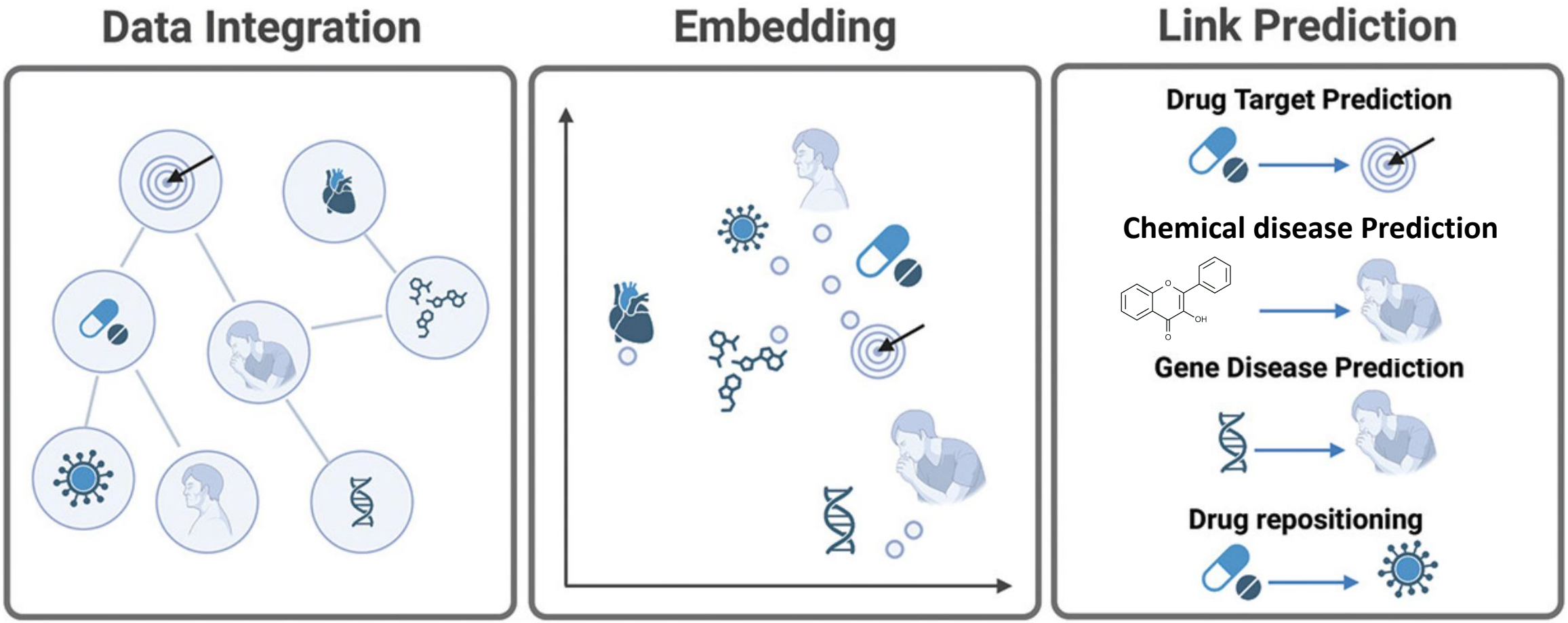
Funded by the European Union

What is a Knowledge graph



Pavel et al., CSBJ, 2022

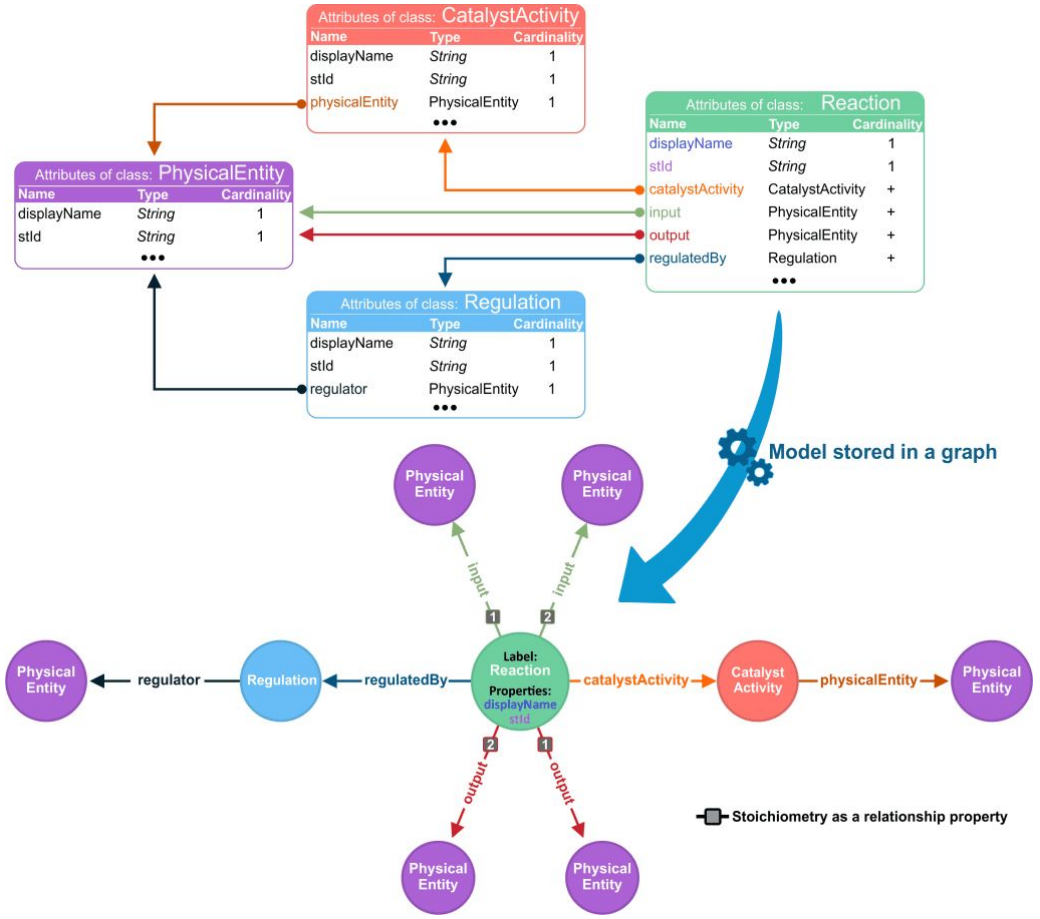
Funded by the European Union



Serra et al., Expert opinion in drug discovery, 2025

Funded by the European Union

What is a Knowledge graph



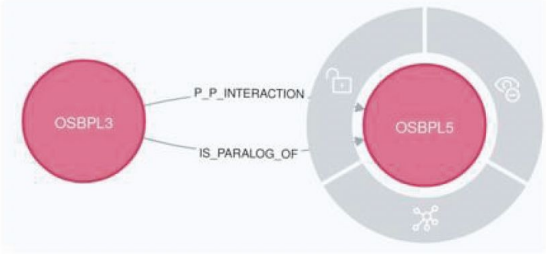
Funded by the European Union

What is a Knowledge graph

A

Node properties

GENE PROTEIN_CODING

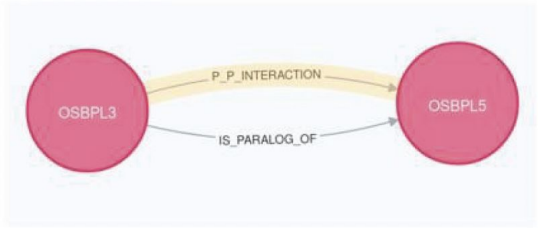


<id>: 16
Ensembl_ID: ENSG00000021762
Ensembl_Protein_ID:
ENSP00000437141,ENSP00000263650,ENSP00000374639,ENSP00000431412,ENSP00000433342,ENSP00000434265,ENSP0
0000302872,ENSP00000436950,ENSP00000433754,ENSP00000435812,ENSP00000432507,ENSP00000433222
Entrez_ID: 114879
HGNC_ID: HGNC:16392
Pfam_ID: PF00169,PF01237
Reactome_Gene_ID: R-HSA-1482801
Swiss_Prot_ID: E9PIJ6,E9PJE6,E9PLN3,E9PNH0,E9PPQ2,E9PQB4,E9PRA9,H0YCD7,Q9H0X9
assembly_name: GRCh38
biotype: protein_coding
contig: 11
created: 1570204060558
end: 3166739
gene_name: oxysterol binding protein like 5
gene_symbol: OSBPL5
gene_symbol_aliases: OBPH1,ORP5
start: 3087107
strand: -

B

Relationship properties

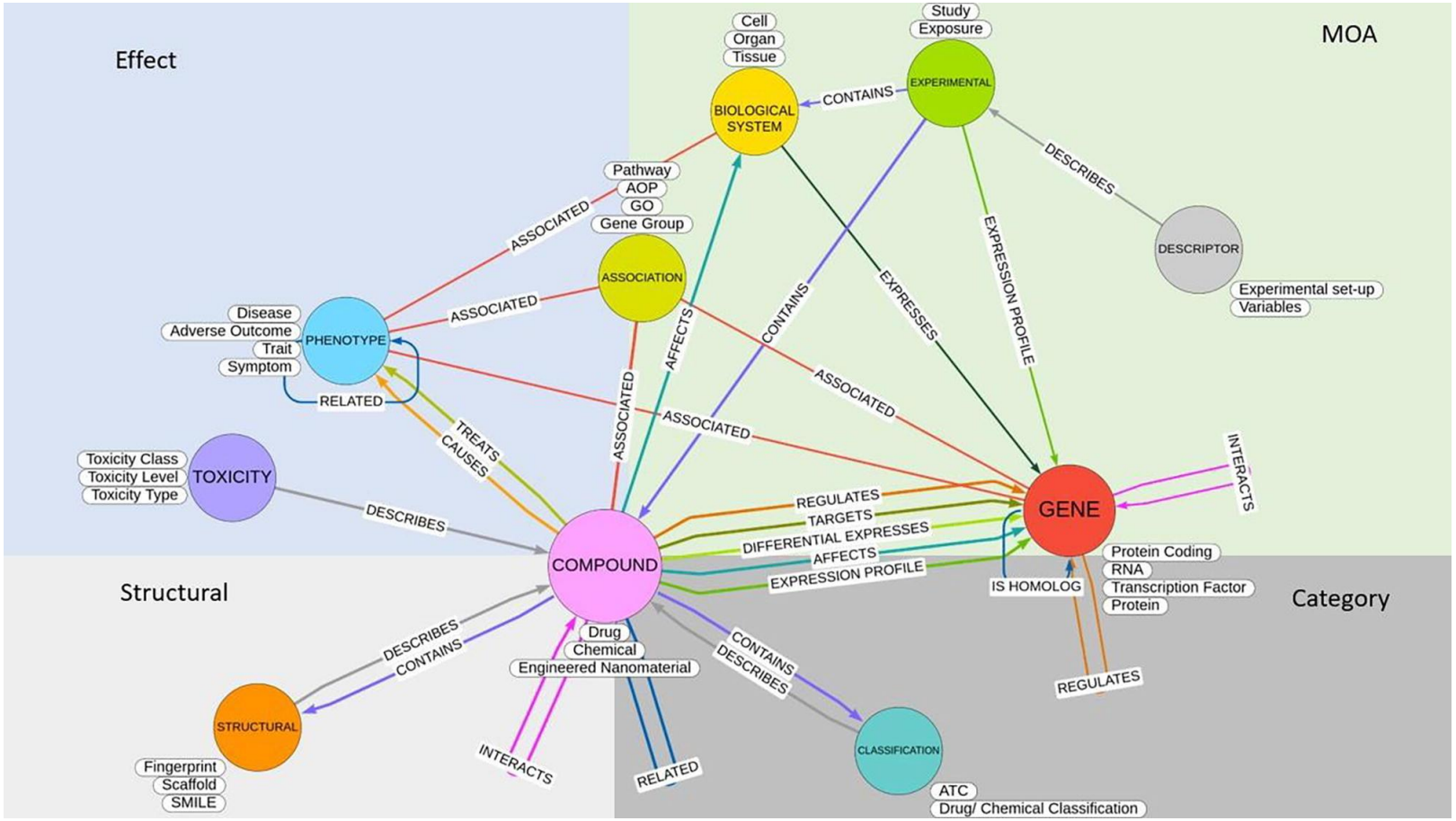
P_P_INTERACTION



<id>: 65180555
HIPPIE_Score: 0.73
HitPredict_Confidence: High
HitPredict_Interaction_score: 0.50102694538318
HitPredict_Method_Score: 0.41838000000000003
Reactome_Interaction_Type: physical association
SignaLink_Interaction_Type: PPI predicted as directed
created: 1572258044084
data_downloaded: 2019-10-24
directed: False
interaction_type: activation
source: HIPPIE,HitPredict,KEGG,STRING,Reactome,SignaLink

Funded by the European Union

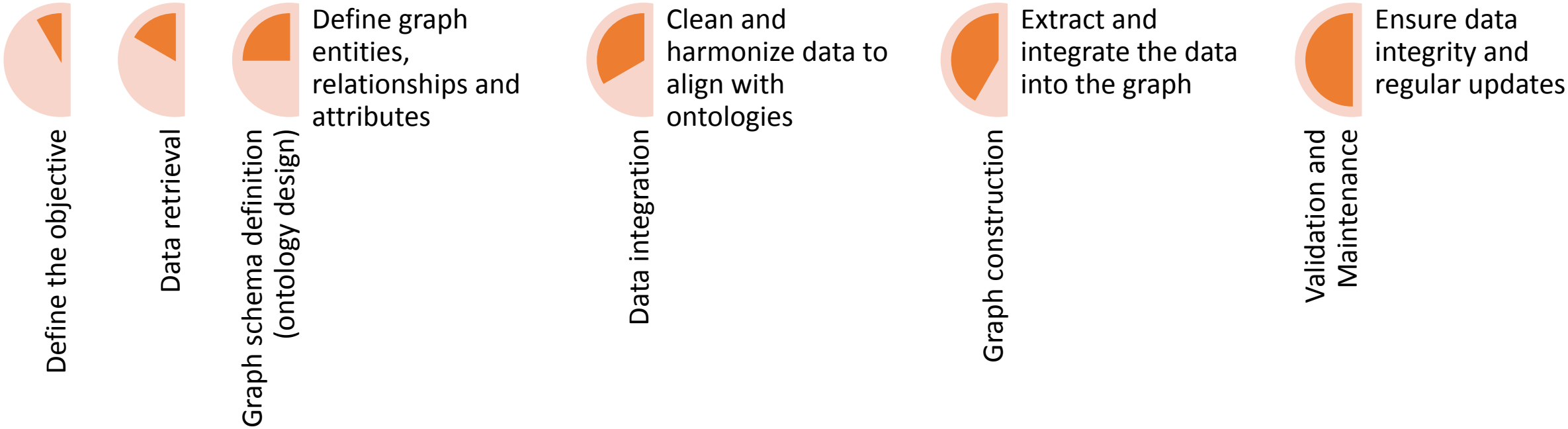
KG for chemical safety assessment



Pavel et al., CSBJ, 2022

Funded by the European Union

KG development process



Funded by the European Union

Challenges

Data Heterogeneity

- Fragmentation across multiple data sources
- Inconsistent formats and identifiers
- Lack of interoperability between sources

Modelling

- Schema design that balance flexibility and specificity to accommodate different data types and relationships

Scalability

- As the amount of data grow, so do the computational demands for graph construction and analysis

Bias

- Many graphs are biased towards well-studied entities.
- Underrepresentation of negative results (e.g., failed drug trials).

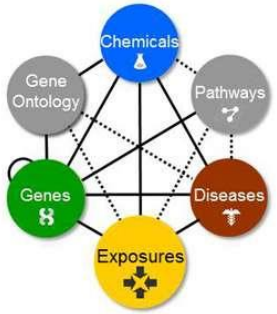
Usability

- Lack of user-friendly interfaces for non-experts.

Funded by the European Union

Possible data sources for life science

- Ontologies
- Databases
- Literature



MIMIC-III Clinical Database

Funded by the European Union

Many words for the same thing

BRCA2 DNA repair associated

Fanconi anemia, complementation group D1

P51587

HGNC:1101

BRCA2

PNCA2

FACD

FAD1

Chromosome 13: 32,315,086-32,400,268 forward strand

breast cancer 2

CCDS9344.1

GLM3

ENSG00000139618

XRCC11

Breast cancer type 2 susceptibility protein

675

BROVCA2

Funded by the European Union

Challenges

Data Heterogeneity

- Fragmentation across multiple data sources
- Inconsistent formats and identifiers
- Lack of interoperability between sources

Modelling

- Schema design that balance flexibility and specificity to accommodate different data types and relationships

Scalability

- As the amount of data grow, so do the computational demands for graph construction and analysis

Bias

- Many graphs are biased towards well-studied entities.
- Underrepresentation of negative results (e.g., failed drug trials).

Usability

- Lack of user-friendly interfaces for non-experts.

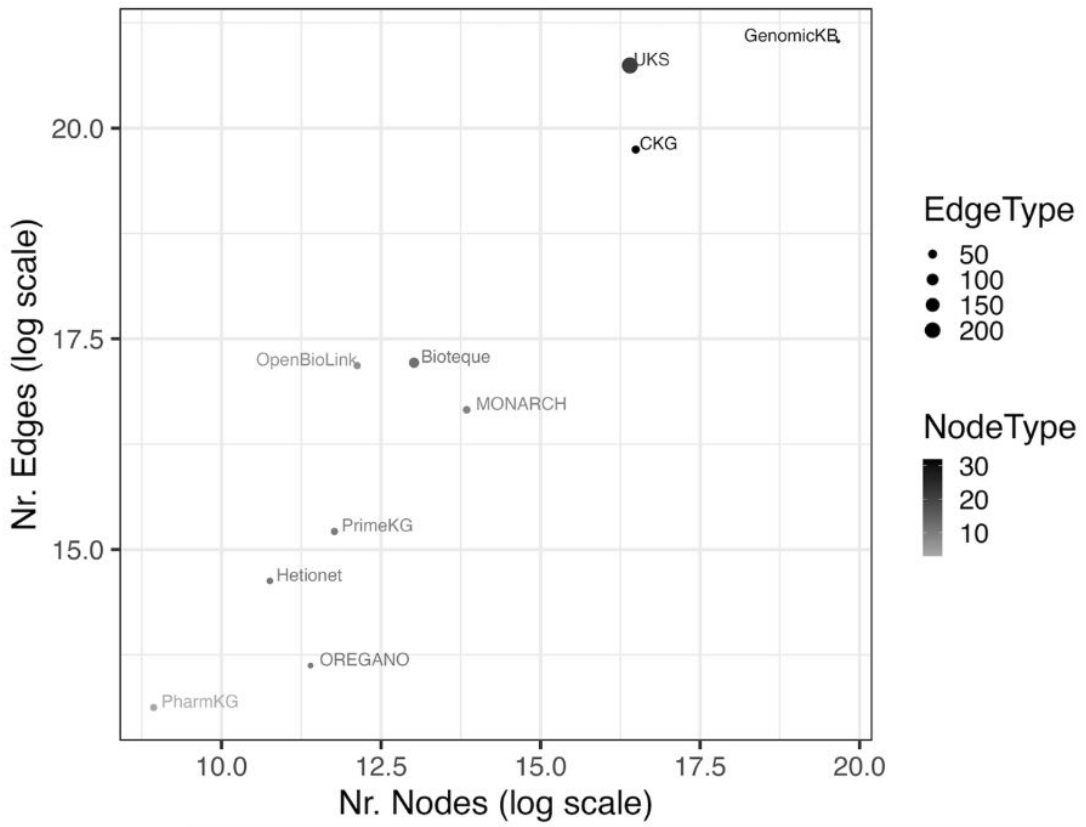
Funded by the European Union

Existing knowledge graphs

Applications:

- Search engines
- recommendation systems
- Precision medicine
- drug discovery

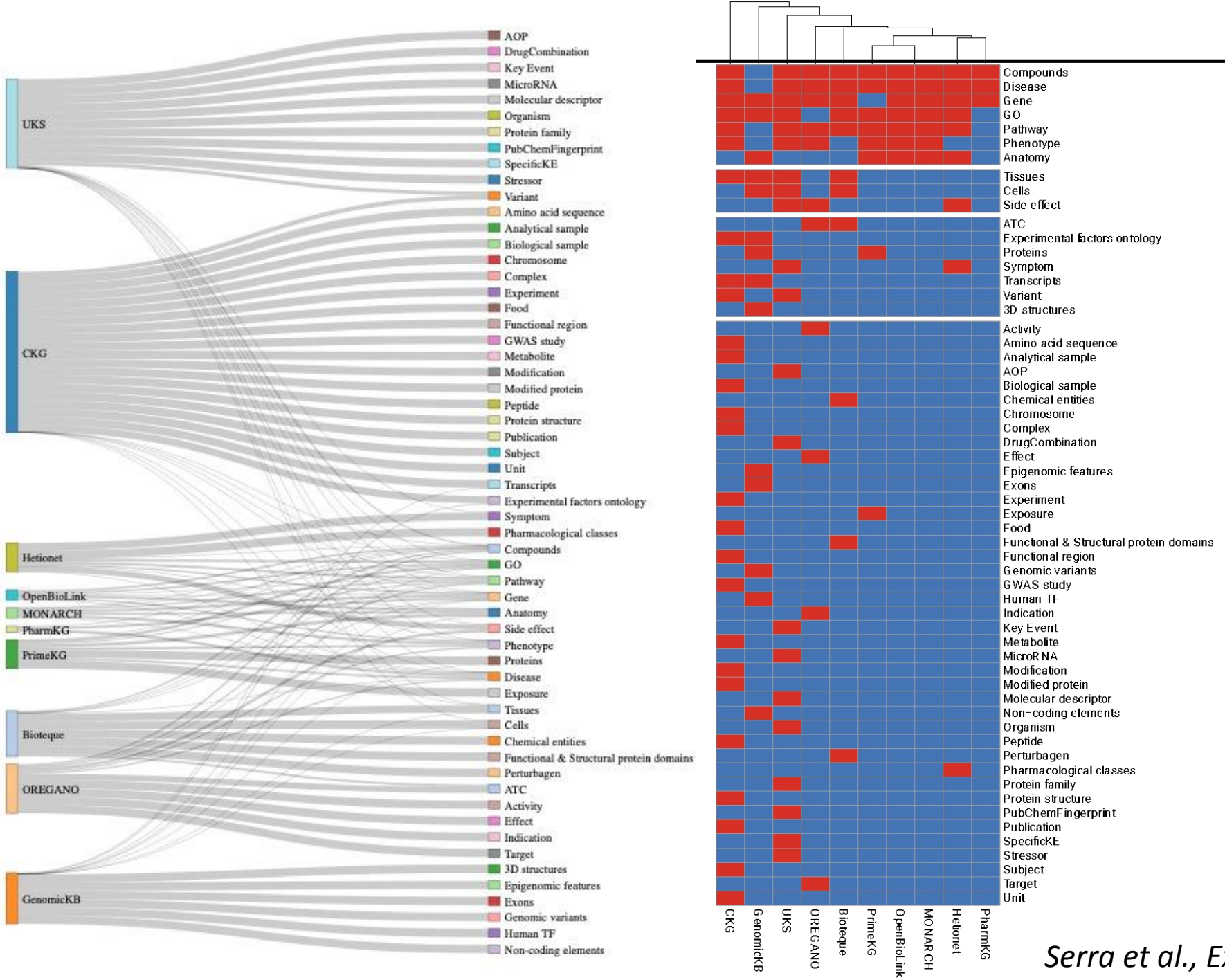
	Purpose	Publication Year
Hetionet	Drug repositiongin	2017
PharmKG	Biomedical data mining	2020
OpenBioLink	Benchmark framework	2020
UKS	Toxicology	2021
CKG	Clinical decision making	2022
Bioteque	Biomedical knowledge graph for embedding	2022
GenomicKB	Human genome database	2023
OREGANO	Drug repositioning	2023
PrimeKG	Precision medicine	2023
MONARCH	General	2023



Serra et al., Expert opinion in drug discovery, 2025

Funded by the European Union

Compound
Chemical
Drug



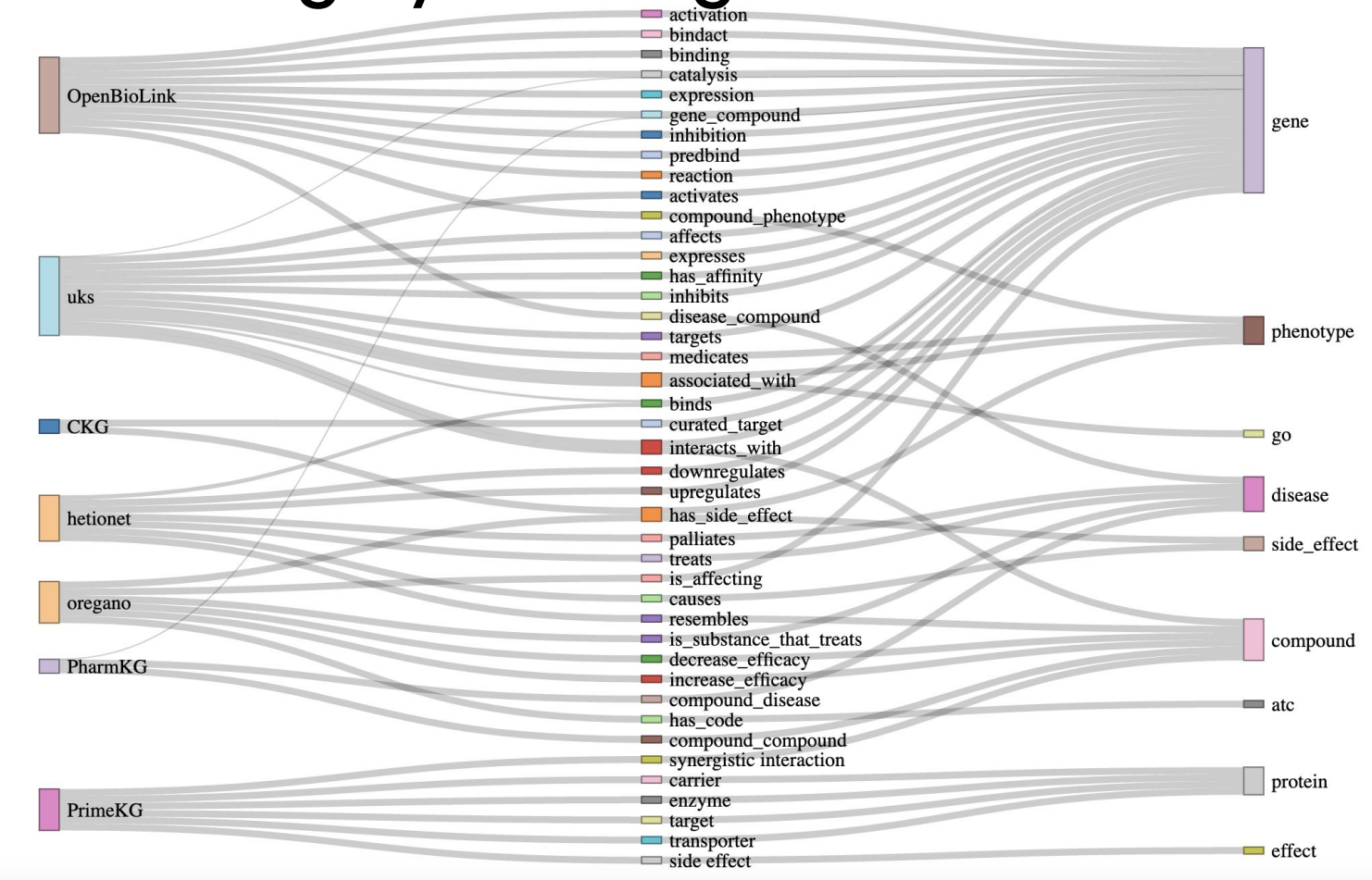
Serra et al., Expert opinion in drug discovery, 2025

	Hetionet	PharmKG	OpenBioLink	UKS	CKG	Bioteque	GenomicKB	OREGANO	PrimeKG	MONARCH
Gene/Products	Entrez	Entrez	STRING	Ensemble		UniProt	GENCODE	DrugBank	Entrez	HGNC
Proteins					UniProt	UniProt	GENECODE			Protein Ontology
Compound	DrugBank	PubChem	PubChem	Pubchem	DrugBank			DrugBank	DrunBank	CHEBI
Disease	DO	MeSH	DO	Concept ID	DO	DO		HPO	MONDO	MONDO
Sympthoms	MeSH ID									UBERON
Sideeffect	SIDER		HPO					SIDER		
Phenotype			HPO	Concept ID	HPO			HPO	HPO	
Effect								DrugBank	HPO	
Anatomy	Uberon		Uberon				Uberon		Uberon	
Pathways	WikiPathways Reactome		KEGG / Reactome	KEGG, Reactome, WikiPathways		Reactome		Reactome	Reactome	GO
GO	GO		GO	GO	GO	GO	GO		GO	
Cell						Cellosaurus	Cell Ontology			
Tissue						Tissue Ontology				

Serra et al., Expert opinion in drug discovery, 2025

Funded by the European Union

Edge labels are highly heterogeneous



Serra et al., Expert opinion in drug discovery, 2025

Funded by the European Union

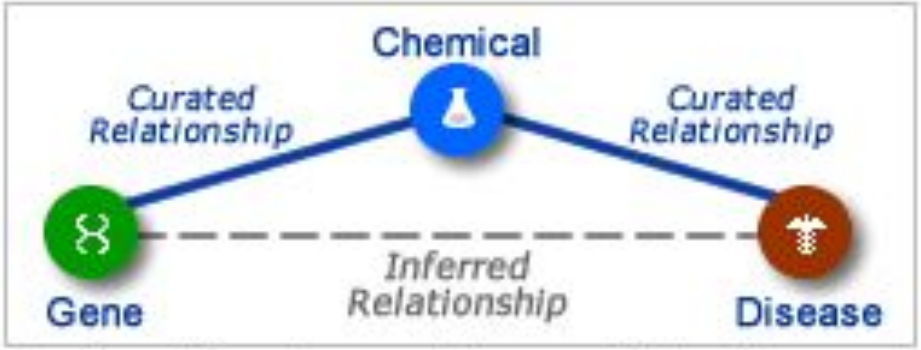
Quality of edges

Even less structured than nodes

Quantitative evaluation of edges is often missing

Curated vs inferred information

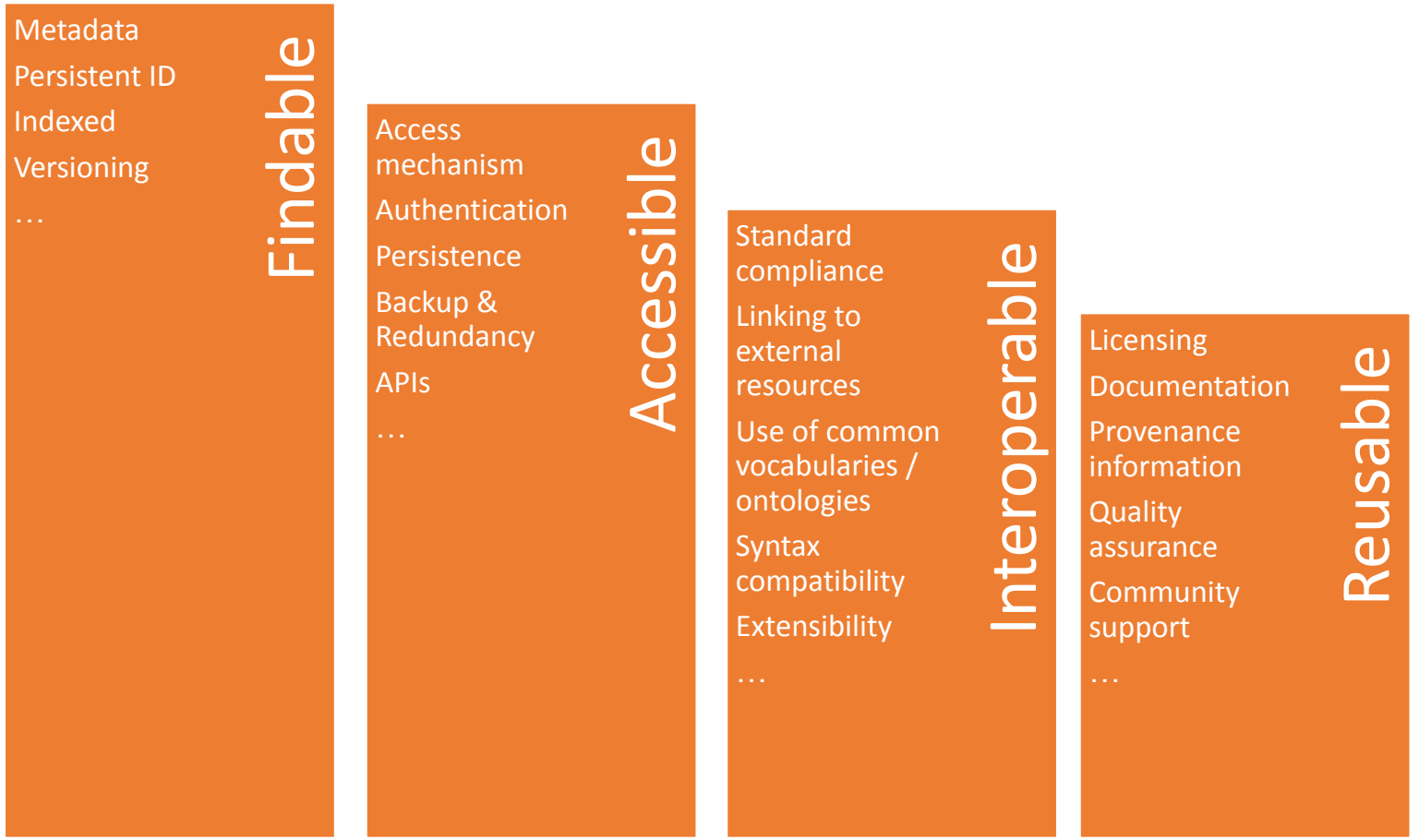
Missing contextualization (e.g. biological system, dose, time)



Creating inferred gene-disease relationships.

Funded by the European Union

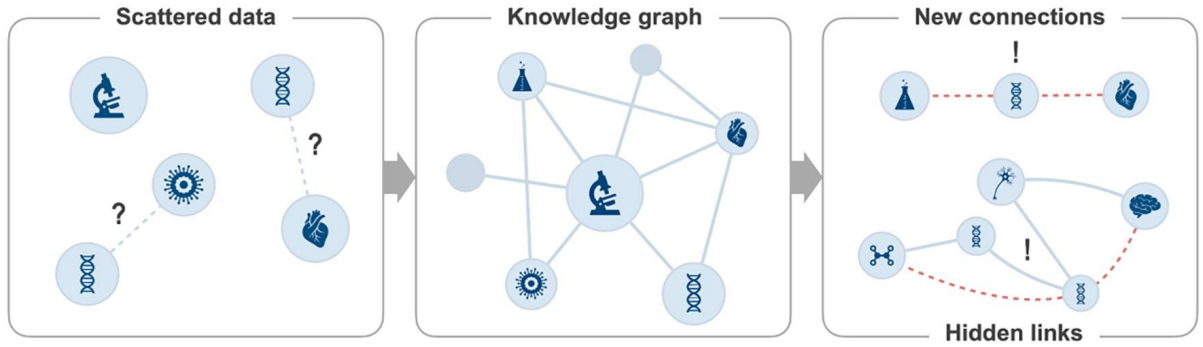
FAIR criteria



Funded by the European Union

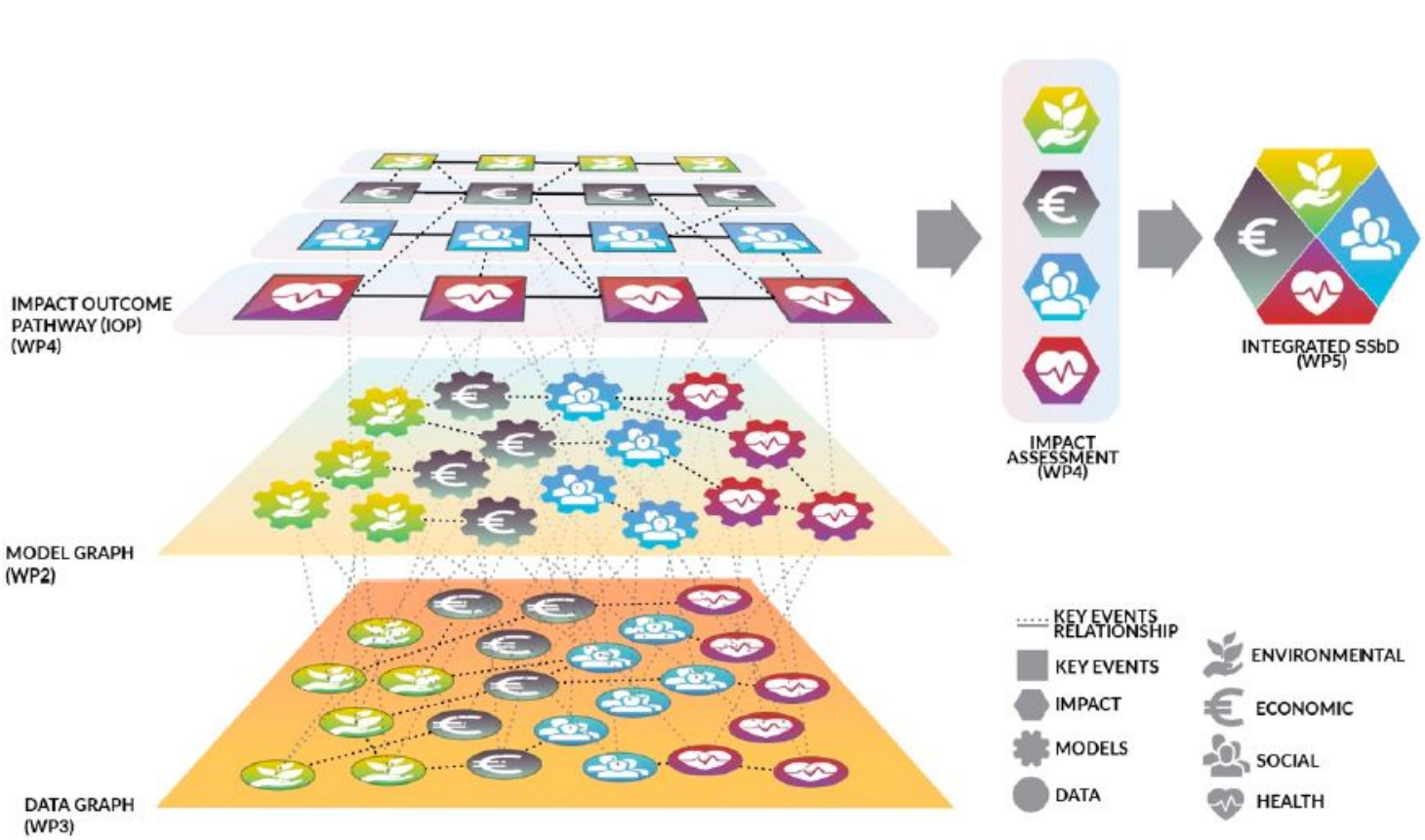
FAIRness of KG components

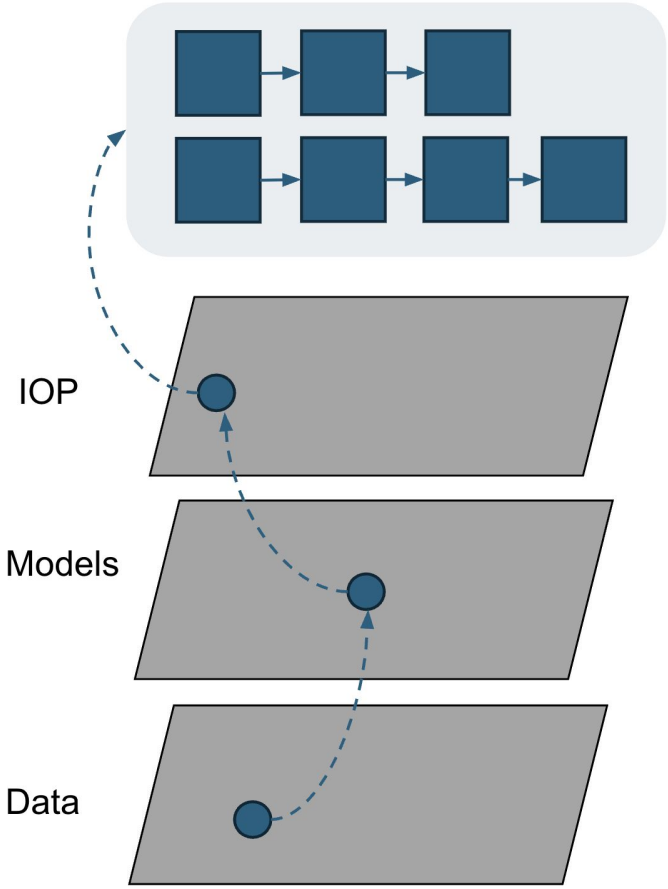
- FAIRness of the nodes
- FAIRness of the edges
- FAIRness of the extracted new knowledge



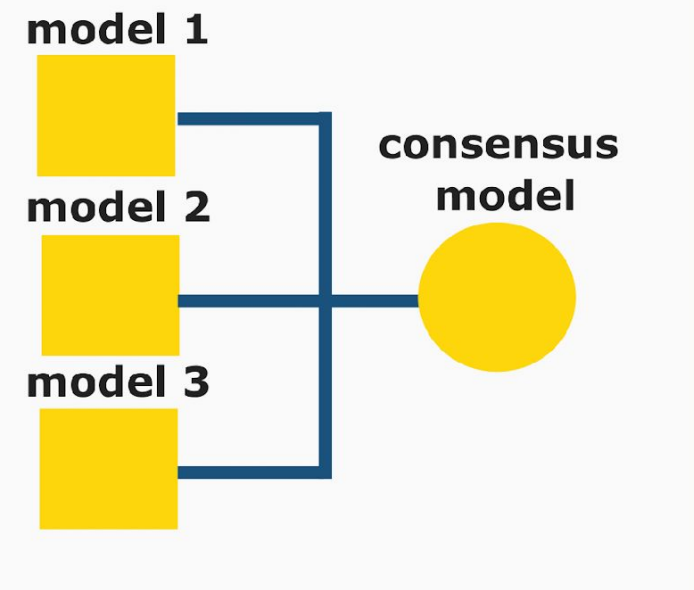
Funded by the European Union

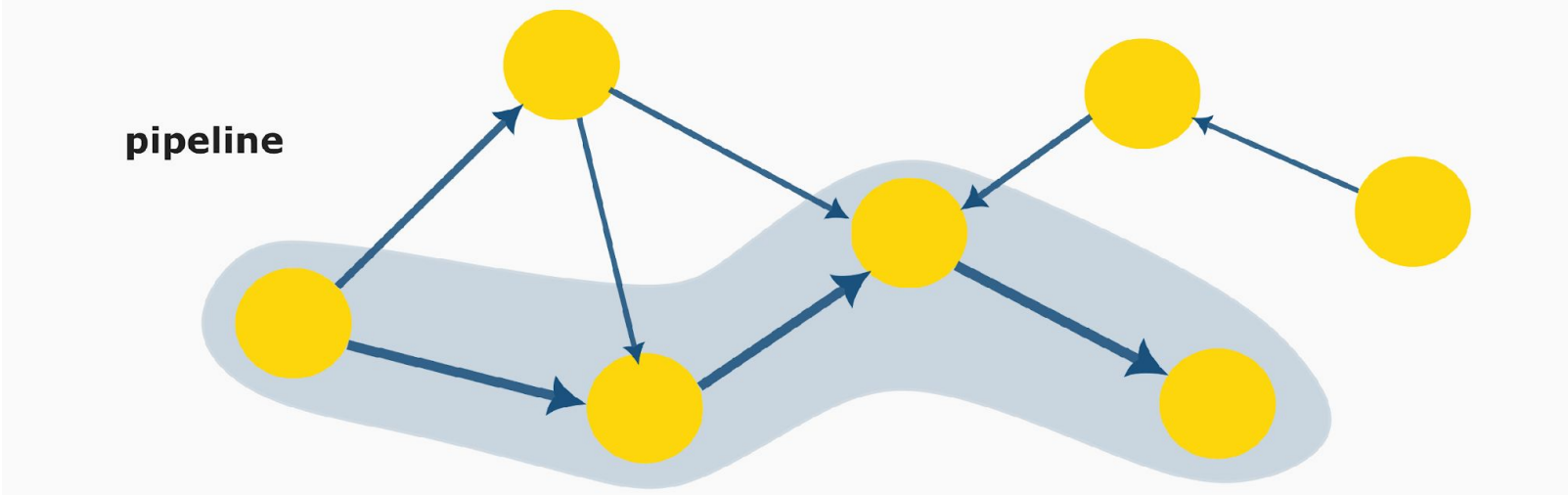
INSIGHT Knowledge Graph





Funded by the European Union





Serra et al., ES&T, submitted

Funded by the European Union

Overview

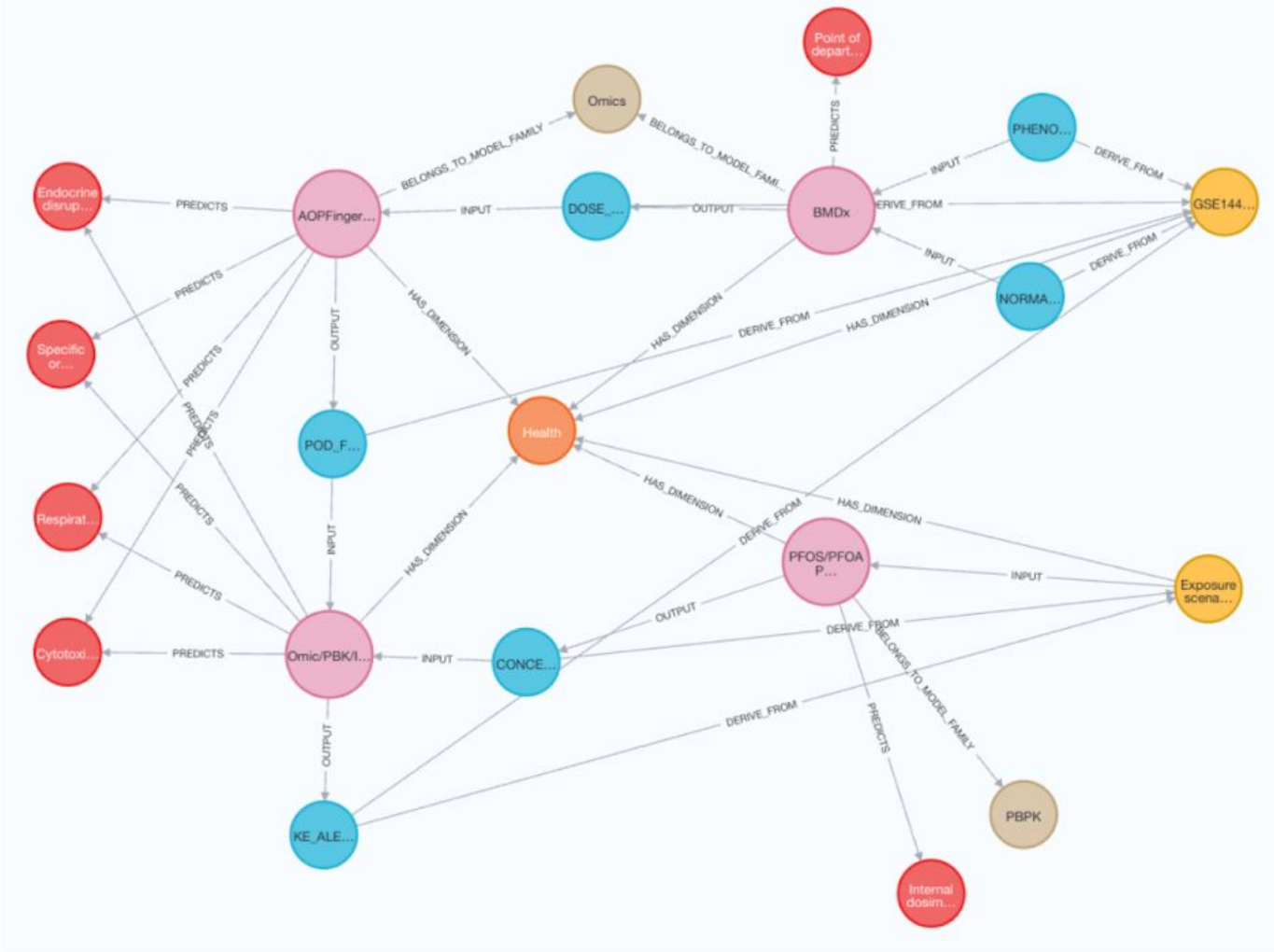
Node labels

* (19) | DerivedDataset (6) | Model (4) | ModelFamily (2) | Indicator (6) | Dimension (1) | ExperimentalDataset (2)

Relationship types

* (36) | HAS_DIMENSION (6) | DERIVE_FROM (7) | INPUT (6) | BELONGS_TO_MODEL_FAMILY (3) | PREDICTS (10) | OUTPUT (4)

Displaying 19 nodes, 33 relationships.



Node prop

Model

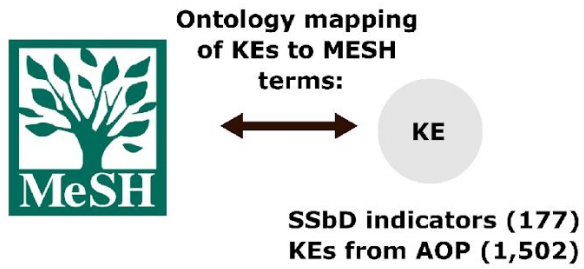
<elementId>
<id>
harURLdocumentation
hasAPI
hasAim
hasContactInfo
hasDocker
hasEstimatedStatistic
hasExpectedInputType
hasExpectedOutputType
hasGUI
hasLicence
hasName
hasProgrammingLanguage
hasReference
hasURL
hasUncertaintyEstimator
id

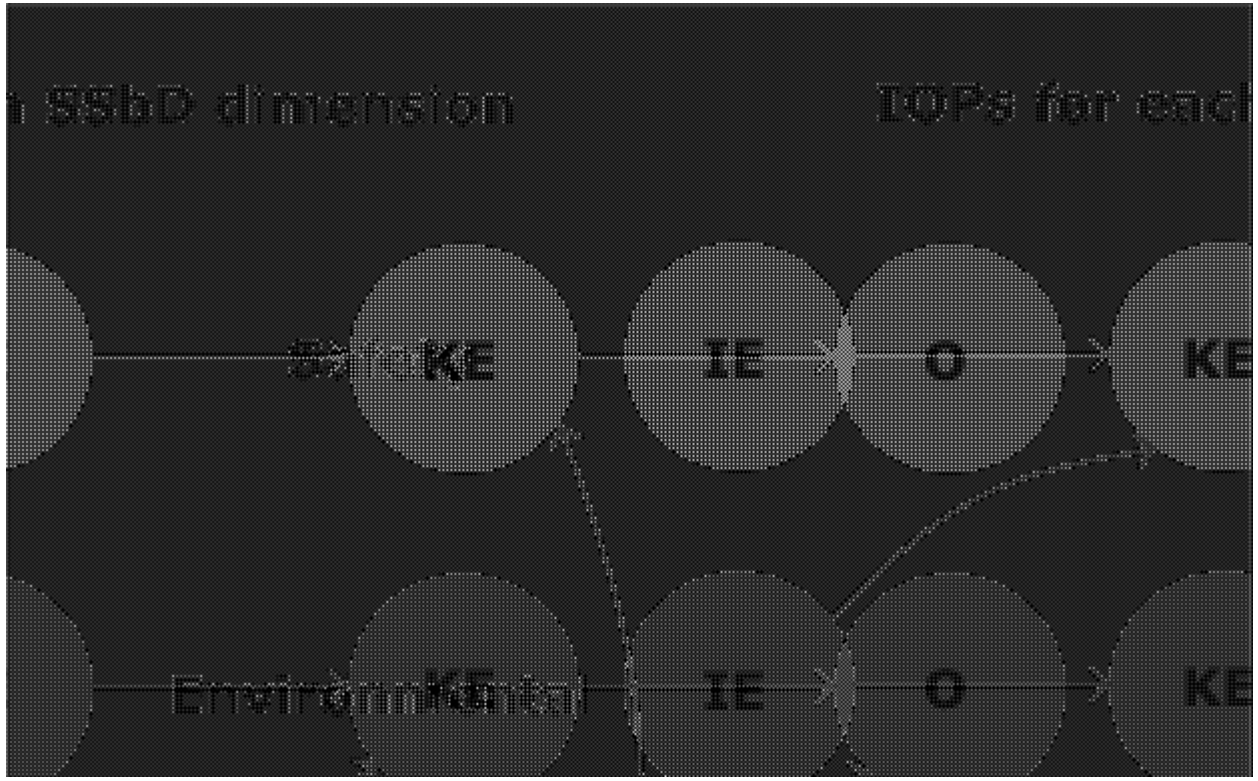
Serra et al,. In prep

Funded by the European Union

Impact Outcome Pathway (IOP)

- Indicators
- Impact categories
- KEs from AOP-wiki
- SSbD

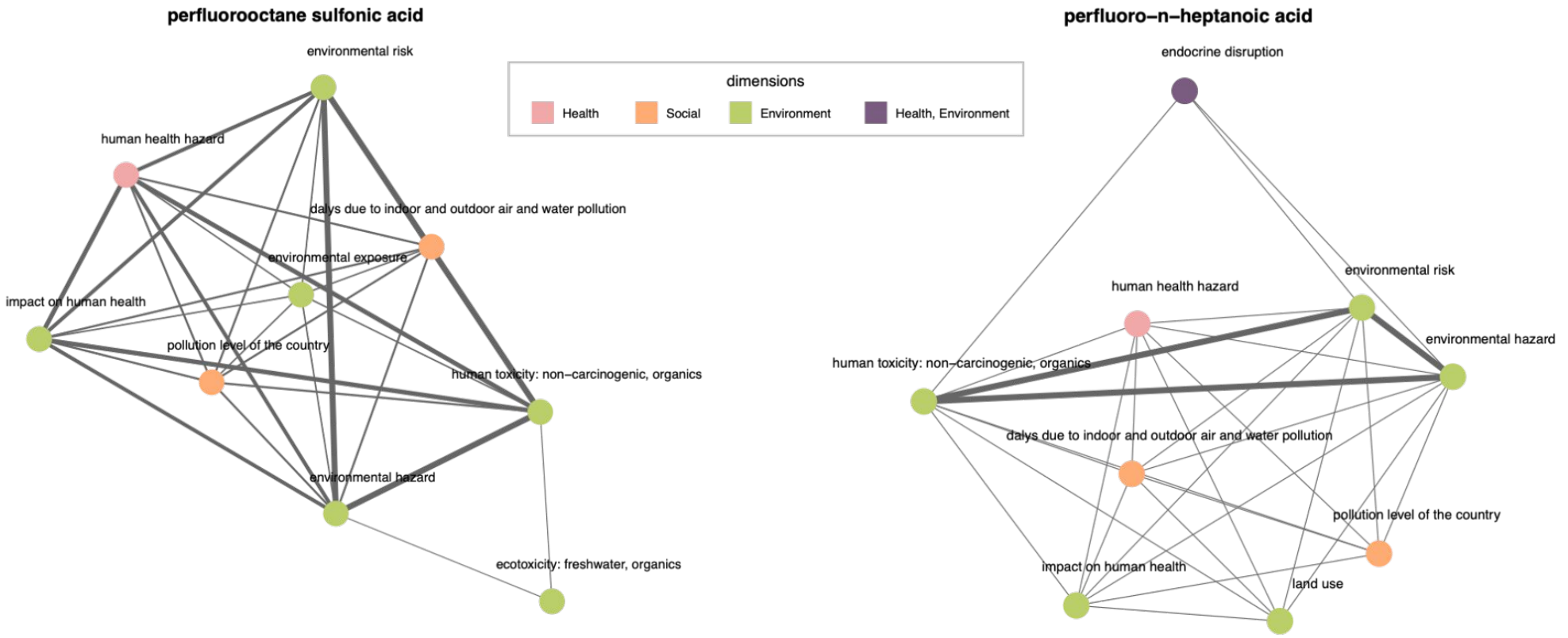




Torres et al,. In prep

Funded by
the European Union

Impact Outcome Pathway (IOP)



Torres et al,. In prep

Funded by the European Union

Future Directions

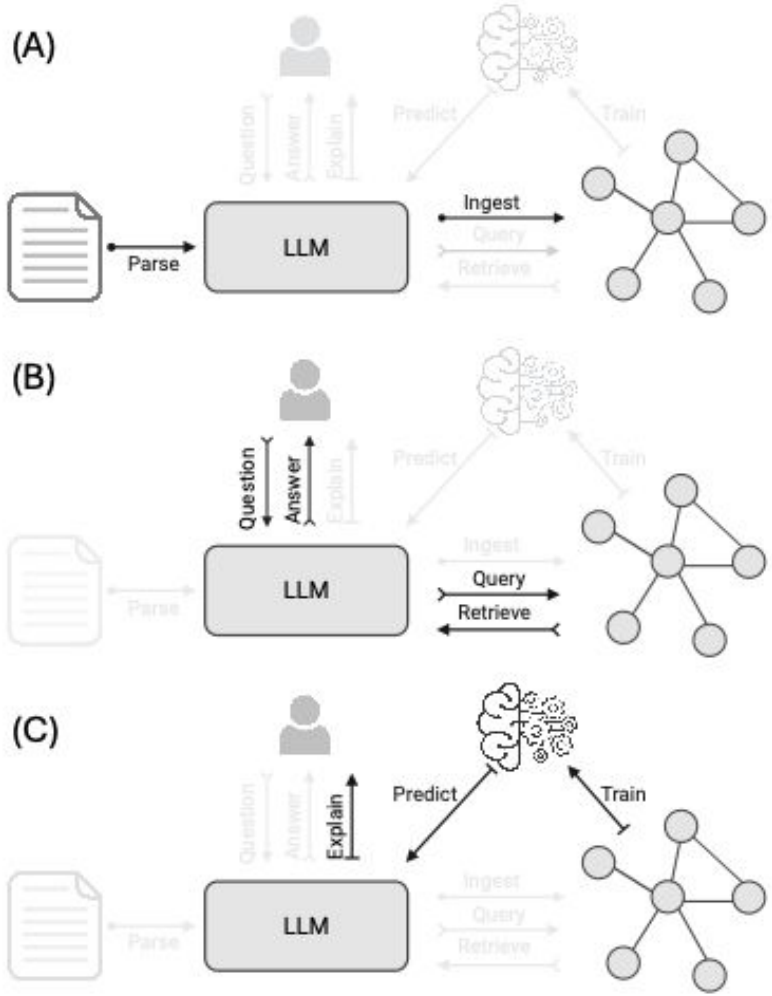
Leveraging AI and Large Language Models (LLMs):

Enhanced data harmonization and interpretability.

Natural language interfaces for KG querying.

Open Science Frameworks:

Encouraging collaborative, community-driven KG development to improve interoperability



Serra et al., Expert opinion in drug discovery, 2025

Funded by the European Union

Summary

KGs are powerful tools for FAIR data representation, and not only

Data must be harmonized before injected into knowledge graphs

Encouraging collaborative, community-driven KG development to improve interoperability

Growing role of AI also as interphase with KG

Funded by the European Union

Thank you!

Funded by the European Union