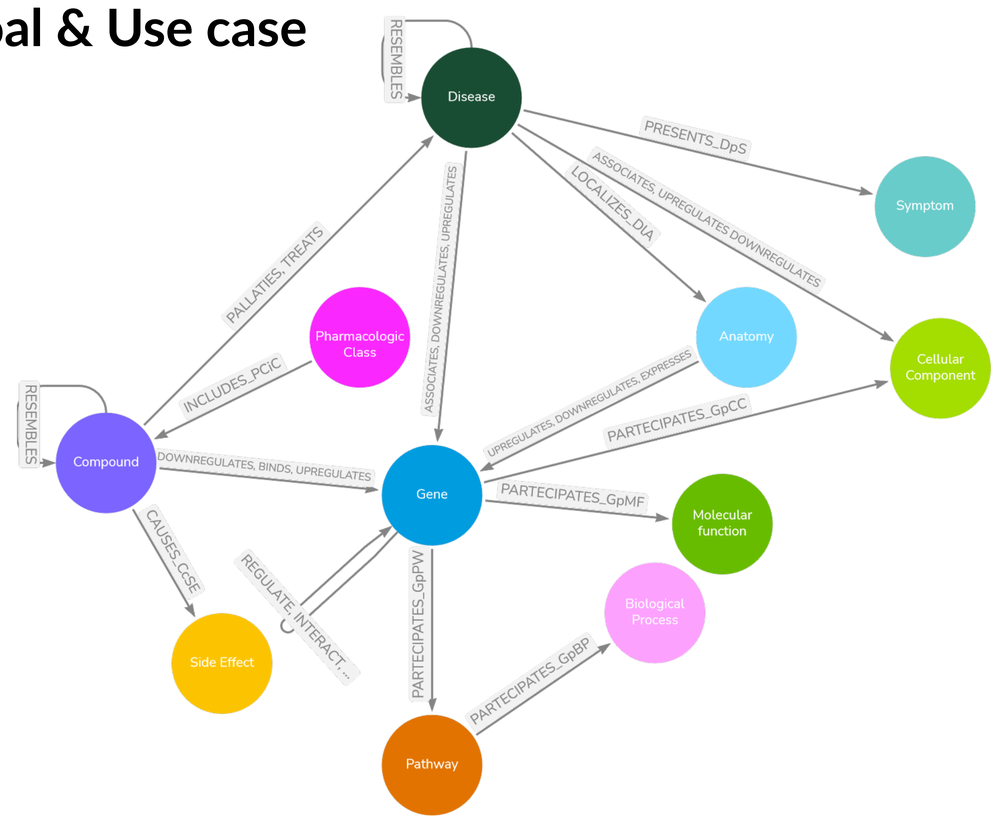
A Semi-Automatic Approach for feeding Bio-Medical KGs

Sara Bonfitto, Manuel Dileo*, Elena Casiraghi, Sabrina Gaito, Giorgio Valentini, Marco Mesiti
Department of Computer Science, University of Milan, Italy
*manuel.dileo@unimi.it

Introduction

- Bio-Medical Knowledge Graphs (BioKGs) are largely used for the representation of **heterogeneous inter-related bio-medical entities** that can be exploited for the development of **artificial intelligence in medicine**.
- The continuous feeding of BioKGs with new results obtained by laboratory analysis is of paramount importance for the **generation of massive datasets** on which the AI algorithms can be properly trained and tested.
- In this work we proposed a semi-automatic approach for the acquisition of tabular data, their **semantic annotation** according to a domain Ontology, and translation in RDF triples. After the validation of the generated graph, it can be included in BioKG.

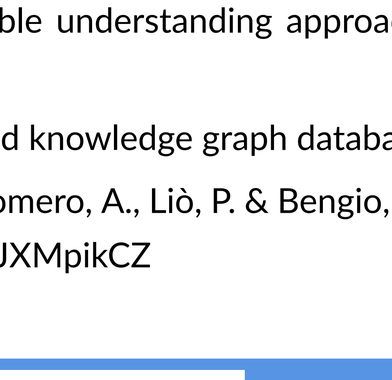
Goal & Use case



SynLethDB 2.0 [3] is a recently developed BioKG representing knowledge about **synthetic lethal interactions between gene-pairs** but also bio-medical knowledge from other bio-medical databases (linking e.g. genes to genes, genes to compounds, compounds and their side effects, etc.).

Chemical Compound	Chemical ID	Gene1	Gene2
10,11-dihydro-10,11-dihydroxy-5H-dibenzazepine-5-carboxamide	C004822	EPHX1	CTSD
10-(4'-(N-diethylamino)butyl)-2-chlorophenoxazine	C553100	TFEB	CTSD
10-(4'-(N-diethylamino)butyl)-2-chlorophenoxazine	C553100	THPO	POLQ
10-(4'-(N-diethylamino)butyl)-2-chlorophenoxazine	C553100	THPO	ADRA1D
10-nitro-oleic acid	C521487	AKT1	ADRA1D

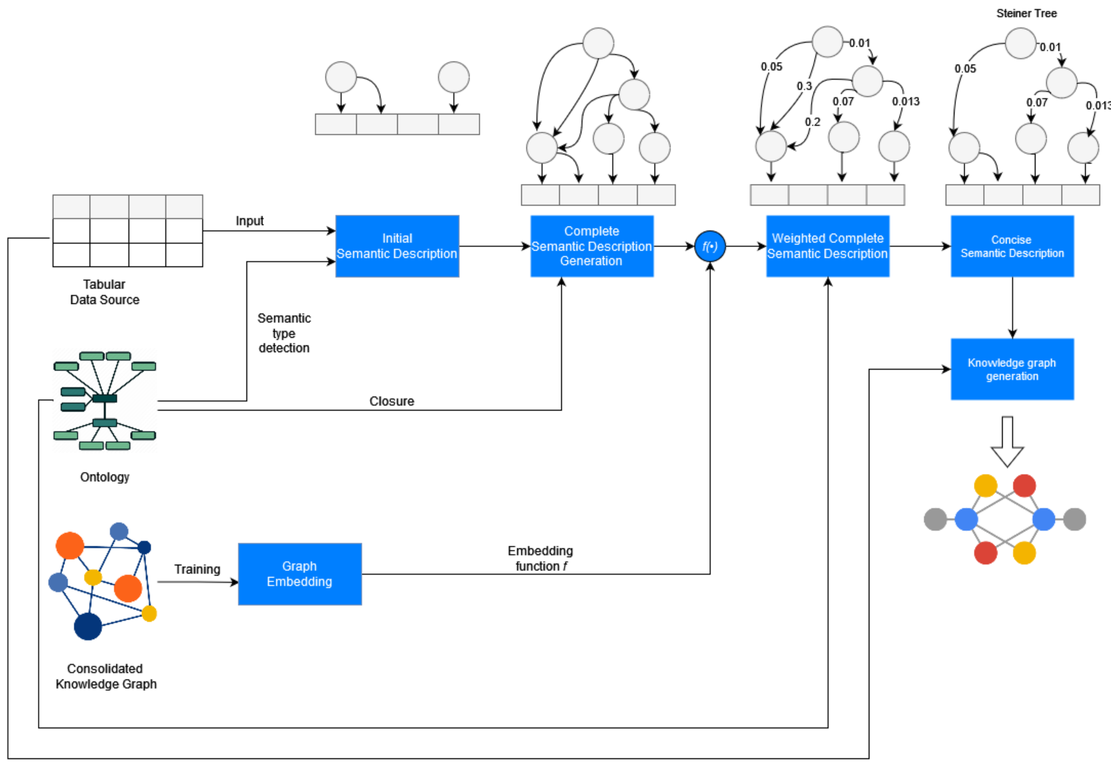
SAGA



The table above shows the **result of a laboratory experiment** that discovered that ChemicalCompound upregulates Gene1 and that Gene1 is synthetically lethal for Gene2. The goal of our work is to **automatically process the table to extract a semantic description** of its content that will be exploited for the translation of the table in the property graph on the right.

Methodology & Results

The **SAGA**^{tab} - Semantic Approach for the acquisition of tabular data is proposed that relies on Graph Attention Networks.

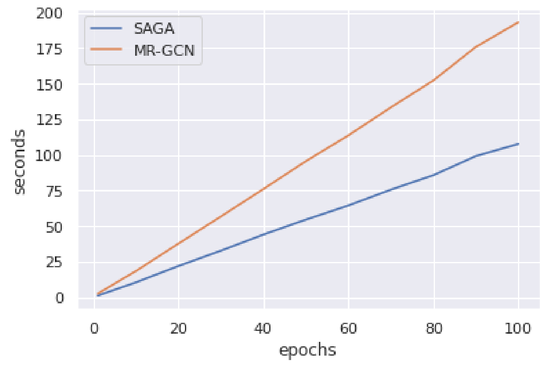


For the construction of the semantic description we exploit:

- The knowledge from a **domain Ontology** to identify *i)* concepts associated to table columns; *ii)* relationships that are valid according to the Ontology.
- A cutting edge heterogeneous attention-based **graph neural network** to embed SynLethDB into a vectorial space.

We evaluated SAGA^{tab} on a traditional **link prediction** setting using AUROC as the evaluation metric.

Datasets	SeMi	MR-GCN	SAGA
Movie-set	0.55	0.67	**0.84**
Area-set	0.5	0.77	**0.91**
PP-set	0.51	0.67	**0.81**



Conclusions & Future work

- In this work, a complex architecture has been set up for facilitating the acquisition of experimental data and their integration into a consolidated knowledge graph.
- Several **GUIs** have been developed **for supporting the user** in checking and modifying the data and the automatically predicted links.
- As future work we wish to further extend the architecture for working with different types of data and for further support the user in the incremental feeding of BioKGs.

References

[1] Taheriyan, M., Knoblock, C., Szekely, P. & Ambite, J. Learning the semantics of structured data sources. *J. Of Web Semantics*. 37-38 pp. 152-169 (2016)

[2] Bonfitto, S., Casiraghi, E. & Mesiti, M. Table understanding approaches for extracting knowledge from heterogeneous tables. *WIREs Data Mining & Knowledge Discovery*. 11 (2021)

[3] Wang, J., et al, SynLethDB 2.0: a web-based knowledge graph database on synthetic lethality for novel anticancer drug discovery. *Database*. **2022** (5)

[4] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P. & Bengio, Y. Graph Attention Networks. *International Conference On Learning Representations*. (2018), https://openreview.net/forum?id=rJXMpikCZ