Computational and Human-based methods for Knowledge Discovery over Knowledge Graphs

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades

Doktor-Ingenieur
(abgekürzt Dr.-Ing.)
genehmigte Dissertation

von Herrn

M.Sc. Ariam Rivas Méndez
geboren am 26.06.1990
in Holguín, Kuba

2023

Referentin:	Prof. Dr. Maria-Esther Vidal, Leibniz Universität Hannover
Korreferentin:	Prof. Dr. Sandra Geisler, RWTH Aachen University
Tag der Promotion:	09.05.2023

Abstract

The modern world has evolved, accompanied by the huge exploitation of data and information. Daily, increasing volumes of data from various sources and formats are stored, resulting in a challenging strategy to manage and integrate them to discover new knowledge. The appropriate use of data in various sectors of society, such as education, healthcare, e-commerce, and industry, provides advantages for decision support in these areas. However, knowledge discovery becomes challenging since data may come from heterogeneous sources with important information hidden. Thus, new approaches that adapt to the new challenges of knowledge discovery in such heterogeneous data environments are required. The semantic web and knowledge graphs (KGs) are becoming increasingly relevant on the road to knowledge discovery. This thesis tackles the problem of knowledge discovery over KGs built from heterogeneous data sources. We provide a neuro-symbolic artificial intelligence system that integrates symbolic and sub-symbolic frameworks to exploit the semantics encoded in a KG and its structure. The symbolic system relies on existing approaches of deductive databases to make explicit, implicit knowledge encoded in a KG. The proposed deductive database DS can derive new statements to ego networks given an abstract target prediction. Thus, DS minimizes data sparsity in KGs. In addition, a sub-symbolic system relies on knowledge graph embedding (KGE) models. KGE models are commonly applied in the KG completion task to represent entities in a KG in a low-dimensional vector space. However, KGE models are known to suffer from data sparsity, and a symbolic system assists in overcoming this fact. The proposed approach discovers knowledge given a target prediction in a KG and extracts unknown implicit information related to the target prediction. As a proof of concept, we have implemented the neuro-symbolic system on top of a KG for lung cancer to predict polypharmacy treatment effectiveness. The symbolic system implements a deductive system to deduce pharmacokinetic drug-drug interactions encoded in a set of rules through the Datalog program. Additionally, the sub-symbolic system predicts treatment effectiveness using a KGE model, which preserves the KG structure. An ablation study on the components of our approach is conducted, considering state-of-the-art KGE methods. The observed results provide evidence for the benefits of the neuro-symbolic integration of our approach, where the neuro-symbolic system for an abstract target prediction exhibits improved results. The enhancement of the results occurs because the symbolic system increases the prediction capacity of the sub-symbolic system. Moreover, the proposed neuro-symbolic artificial intelligence system in Industry 4.0 (I4.0) is evaluated, demonstrating its effectiveness in determining relatedness among standards and analyzing their properties to detect unknown relations in the I4.0KG. The results achieved allow us to conclude that the proposed neuro-symbolic approach for an abstract target prediction improves the prediction capability of KGE models by minimizing data sparsity in KGs.

Keywords *Neuro-Symbolic System, Sub-Symbolic System, Symbolic System, Knowledge Graph Embedding, Datalog*

Zusammenfassung

Die moderne Welt hat sich weiterentwickelt, begleitet von einer enormen Verwertung von Daten und Informationen. Täglich werden immer größere Datenmengen aus verschiedenen Quellen und Formaten gespeichert, was zu einer anspruchsvollen Strategie für die Verwaltung und Integration dieser Daten führt, um neues Wissen zu entdecken.Die angemessene Nutzung von Daten in verschiedenen Bereichen der Gesellschaft, wie z. B. im Bildungs- und Gesundheitswesen, im elektronischen Handel und in der Industrie, bietet Vorteile für die Entscheidungsfindung in diesen Bereichen. Die Wissensentdeckung stellt jedoch eine Herausforderung dar, da die Daten aus heterogenen Quellen stammen können, in denen wichtige Informationen verborgen sind. Daher sind neue Ansätze erforderlich, die sich an die neuen Herausforderungen der Wissensentdeckung in solchen heterogenen Datenumgebungen anpassen. Das semantische Web und Wissensgraphen (KGs) gewinnen auf dem Weg zur Wissensentdeckung zunehmend an Bedeutung. Diese Arbeit befasst sich mit dem Problem der Wissensentdeckung über Wissensgraphen, die aus heterogenen Datenquellen aufgebaut sind. Wir stellen ein neurosymbolisches System der künstlichen Intelligenz zur Verfügung, das symbolische und subsymbolische Frameworks integriert, um die in einem Wissensgraphen kodierte Semantik und seine Struktur zu nutzen. Das symbolische System stützt sich auf bestehende Ansätze deduktiver Datenbanksysteme, um implizite Wissen, die in einem KG kodiert sind, explizit zu machen. Das vorgeschlagene deduktive System DS kann aus einer abstrakten Zielvorhersage neue Aussagen für das Ego-Netzwerk ableiten. Dadurch minimiert DS die Datenarmut im Wissensgraphen. Darüber hinaus stützt sich das subsymbolische System auf Wissensgrapheneinbettungsmodelle (KGE). KGE-Modelle werden üblicherweise in der KG-Vervollständigung eingesetzt, um Entitäten in einem KG in einem niedrigdimensionalen Vektorraum darzustellen. Es ist jedoch bekannt, dass KGE-Modelle unter Datenarmut leiden, und das symbolische System hilft dabei, diese Tatsache zu überwinden. Der vorgeschlagene Ansatz entdeckt Wissen anhand einer Zielvorhersage in einem Wissensgraphen und extrahiert unbekannte implizite Informationen in Bezug auf die Zielvorhersage. Als Proof of Concept haben wir das neuro-symbolische System auf einem KG für Lungenkrebs implementiert, um die Wirksamkeit einer Polypharmazie-Behandlung vorherzusagen. Das symbolische System implementiert ein deduktives System zur Ableitung von pharmakokinetischen Wechselwirkungen zwischen Medikamenten, die in einem Regelsatz durch ein Datalog-Programm kodiert sind. Im Gegensatz dazu sagt das subsymbolische System die Wirksamkeit der Behandlung anhand eines KGE-Modells voraus. Zusätzlich sagt das subsymbolische System die Wirksamkeit der Behandlung mit Hilfe eines KGE-Modells voraus, das die KG-Struktur beibehält. Für die Komponenten unseres Ansatzes wurde eine Ablationsstudie durchgeführt, bei der modernste KG-Einbettungsmethoden berücksichtigt wurden. Die beobachteten Ergebnisse belegen die Vorteile der neuro-symbolischen Integration unseres Ansatzes, wobei das neuro-symbolische System für eine abstrakte Zielvorhersage verbesserte Ergebnisse aufweist. Die Verbesserung der Ergebnisse erfolgt, weil das symbolische System die Vorhersagekapazität des subsymbolischen Systems erhöht. Darüber hinaus wurde das vorgeschlagene neuro-symbolische System der künstlichen Intelligenz in der Industrie 4.0 (I4.0) evaluiert und seine Effektivität bei der Bestimmung der Verwandtschaft zwischen Normen und der Analyse ihrer Eigenschaften zur Erkennung unbekannter Beziehungen im I4.0KG demonstriert. Die erzielten Ergebnisse lassen den Schluss zu, dass der vorgeschlagene neurosymbolische Ansatz für eine abstrakte Zielvorhersage die Vorhersagefähigkeit von KGE-Modellen verbessert, indem er die Datenarmut in KGs minimiert.

Schlüsselwörter *Neurosymbolisches System, subsymbolisches System, symbolisches System, Wissensgrapheneinbettung, Datalog.*

Acknowledgements

First of all, God, I would like to thank you for allowing me to undergo this exciting experience of doing a Ph.D. You have always been with me, giving me strength, motivation, and faith to achieve your purposes. The bigger the challenge, the closer I felt to you. My wife, my love, Maria Isabel, this is an achievement for both of us. You are my engine who motivated me to continue from my academic beginnings. My parents, Maira and Jorge, who are always concerned and attentive in everything they can help, even from a distance. Also, my in-laws Javie and Zoraida, who had words of encouragement every weekend to continue on this journey. Thanks to all my family for their prayers and especially to my uncle Wildo and Angelica for their support and for being there for me.

I would like to thank Prof. Dr. Maria-Esther Vidal for motivating me to apply for a DAAD scholarship which we successfully received. Her supervision along the way has been invaluable and always inspiring to do academic research. Thank you very much, Prof. Dr. Sören Auer, for all your support during my fellowship. German Academic Exchange Service, thank you for your award and full support during the scholarship. To all the Professors of the Biomedas program, thank you very much. Many thanks, Katja, for helping me since my arrival, and you still continue to help me. Thank you, Irlan Grangel, for guiding me from my beginnings and achieving our first paper together, as well as Diego Collarana; thank you very much. Throughout the exciting academic journey, I have had the opportunity to meet people who have inspired me over the years. Special thanks to Ahmad Sakor, Akhilesh Vyas, Alireza Mohammadi, David Chaves, Disha Purohit, Emetis, Enrique Iglesias, Farah Karim, Hao Haung, Jose Alberto, Kemele, Lucie-Aimee, Maira Russo, Mauricio Brunet, Philipp Rohde, Samaneh, Shahi Dost, Simone Matern, Trupti, Yaser, and Yashraj, with whom I had the pleasure of sharing during my Ph.D. studies.

Ariam Rivas Méndez

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Data and knowledge have become critical assets throughout the ongoing digitization process. Digital activities generate vast amounts of data in all knowledge domains, e.g., education, health, e-commerce, and industry. Data are a top priority for companies whose business processes require data processing. Companies will only be able to acquire and maintain competitive advantages through appropriate knowledge. The data generated in these digitization processes are paramount for improving numerous areas of human development. However, potentially important information is hidden in all this data, which is seldom made explicit or exploited. Knowledge discovery becomes cumbersome because data may come from heterogeneous data sources represented at different levels of structure; they could also either be incomplete or have missing associations. Furthermore, as systems for data generation and ingestion progress, knowledge discovery problems are increasing in complexity, i.e., the solution to the problem depends on an extensive amount of data. Therefore, new techniques are required to adapt to the emerging challenges of discovering knowledge in heterogeneous data settings.

Much research has recently investigated critical aspects of discovering knowledge on real-world data. The Semantic Web created by Tim Berners-Lee [12] and converted into a standard by the World Wide Web Consortium (W3C)[1] is becoming increasingly relevant on the road to knowledge discovery. The Semantic Web aims to enrich the web of data with machine-understandable semantics. Furthermore, semantic technologies aim to represent knowledge from raw data sources and to form semantic networks [13, 53]. A knowledge representation paradigm, knowledge graph (KG), has emerged, contributing to solving knowledge discovery problems. Knowledge graphs are a data structure for organizing real-world knowledge and integrating information from multiple heterogeneous data sources. KG is a directed labeled graph representing connections among data entities with nodes

[1] `https://www.w3.org/`

and edges. Nodes denote real-world entities, and edges represent the relationships between two nodes. Thus, KGs have been adopted as data structures representing data and metadata (contextual details such as data source, authorship, quality measures, and properties associated with the entities); they are becoming central in improving the predictions of artificial intelligence (AI) models by providing them with the knowledge represented in KGs as input.

In Industry 4.0, AI models are applied on top of KGs to extend the knowledge of the standardization frameworks and discover new relationships between standards issues [9, 41, 56, 84]. On the other hand, in the medical domain, AI techniques over KGs, such as Karim et.al. [64], Marinka et.al. [152], and Raziyeh et.al. [83], are successfully utilized in relevant problems, e.g., prediction of Drug-Drug Interactions, forecast of treatment responses by analyzing images [24, 61, 71], prediction of the existence of disease [126, 150], for drug repurposing [111, 129, 147], or analytic support of diagnosis from exposure symptoms [120]. The AI models on top of KGs are widely used for their high performance in supporting expert decision-making. However, they are limited by the data's lack of symbolic representations of reasoning and semantic description. The symbolic representations are valuable because they allow for the representation of complex concepts and relationships that may not be easily captured purely statistically. Symbolic representations can capture abstract concepts, logical rules, and symbolic relationships, enabling more advanced reasoning and inference capabilities.

This thesis aims at providing a knowledge discovery approach over knowledge graphs. We provide a neuro-symbolic approach that integrates symbolic and sub-symbolic representation, reasoning, and learning. The symbolic system resorts to a deductive database to make explicit, implicit information encoded in a knowledge graph. Furthermore, the sub-symbolic system is enhanced by the symbolic system with potentially useful implicit information. The sub-symbolic system implements the knowledge graph embedding (KGE) models, focusing on learning distributed vector representations for entities and relations in knowledge graphs. The implementation of our approach facilitates the uncovering of actionable insights. The approach developed in this thesis is applied in the biomedical and industry 4.0 domains to assess predictive power. The observed results put into perspective the benefits of empowering KGE with facts from the deductive database. The following section motivates the main problem and challenges of this thesis.

1.1 Motivation

We illustrate the work reported in this dissertation in the context of the healthcare domain. The Electronic Health Records (EHRs) of lung cancer patients con-

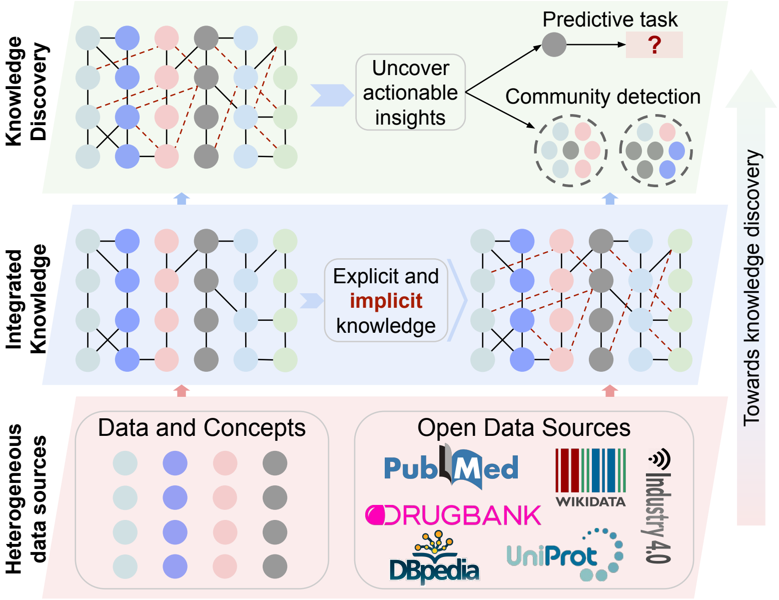


Figure 1.1: **Motivating example**. Steps to knowledge discovery over knowledge graphs. Real-world data and concepts have incomplete data and are in heterogeneous data sources (Layer 1). To uncover the missing associations, knowledge is integrated semantically where the implicit and explicit knowledge are together (Layer 2). Finally, at the third layer, actionable insights can be uncovered on top of a model with knowledge integrated (Layer 3).

tain patients' conditions for effective diagnoses and treatment prescriptions. EHRs may be in various formats, e.g., relational tables storing demographics data, flat files storing liquid biopsies, or clinical notes. Clinicians must validate EHRs and external data sources to effectively diagnose disease, prescribe effective treatments, and foresee adverse effects, e.g., drug interactions or side effects. Multiple open data sources provide crucial knowledge for a complete description of patients' diseases and an understanding of genetic factors related to diseases. These dispersed data from different open sources must be integrated to describe the disease comprehensively. However, open data sources also use a variety of formats, from structured to unstructured. The clinicians will have to search through multiple data sources and identify potential adverse effects and patient characteristics to detect events that may impact the effectiveness of treatment. Data complexity issues such as volume and diversity present a significant challenge to the successful integration of knowledge needed to discover new insights, such as predicting the effectiveness of treatment. Achieving actionable knowledge from real-world data and concepts requires a novel approach that allows us to explore and mine data to uncover actionable insights.

Figure 1.1 depicts the need for knowledge integrated to uncover actionable insights from real-world data and concepts. The first layer comprises heterogeneous data sources that represent data in original formats. For example, the information about patients and drugs might be distributed across different sources on the Web, such as DrugBank[2], Wikidata[3], and UniProt[4], and EHRs as flat files. The data sources may contain missing associations and implicit facts. The second layer, integrated knowledge, shows the heterogeneous data sources represented in a homogeneous model through explicit semantic labeling of concepts and their relationships. However, the implicit knowledge represented by red dashed lines is unknown and should be integrated into the data model in a machine-readable format. The third layer shows a formal representation of explicit and implicit knowledge. The knowledge discovery techniques need to exploit the semantics encoded in the data representation to produce actionable insights. Those actionable insights can be used for decision-making. Challenges must be addressed at each layer to provide a unified view of the heterogeneous data sources and uncover actionable insights. The following section discusses the main problem and challenges that guided this dissertation.

1.2 Problem Statement and Challenges

Technology has facilitated the storage and publication of increasingly large volumes of data at different levels of structure. However, they require to be processed to extract relevant knowledge. More data encodes facts and inherent insights that must be made explicit to understand and analyze them properly. At the conceptual level, we face a knowledge integration and discovery problem, i.e., merge and make explicit the knowledge about the entities spread in different open data sources and uncover relevant information. The research problem that guided our work can be stated as follows: We investigate how to improve discovering actionable knowledge through better knowledge integration. This thesis aims to integrate data and concepts semantically and provides a neuro-symbolic AI approach, enabling the uncovering of helpful knowledge. Several challenges must be overcome to achieve such integration and knowledge discovery. Figure 1.2 illustrates the three main challenges tackled by this thesis.

[2] https://go.DrugBank.com/

[3] https://www.wikidata.org/wiki/Wikidata:Mainpage

[4] https://www.uniprot.org/

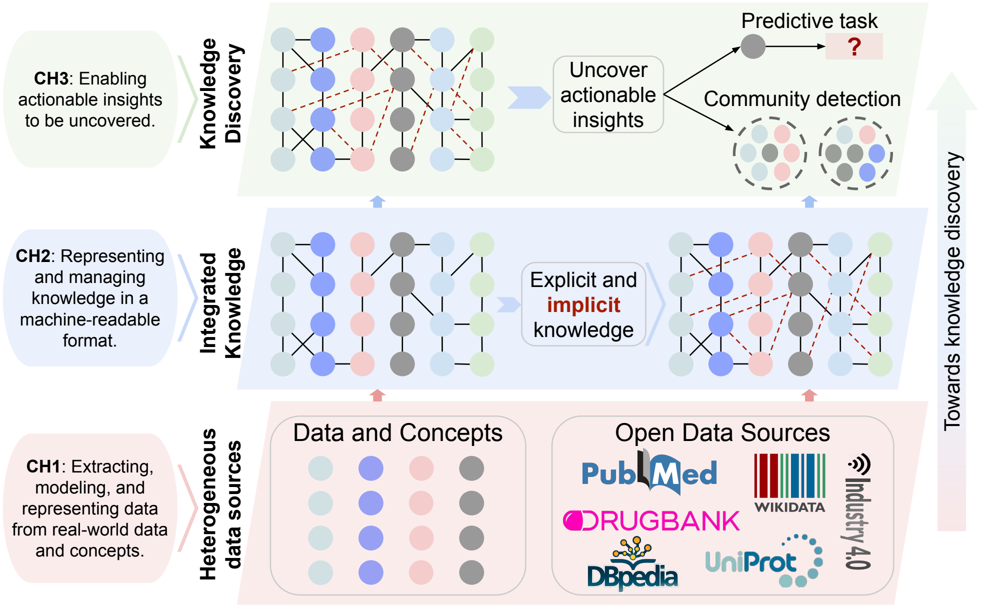

CH3: Enabling actionable insights to be uncovered.
Knowledge Discovery
Uncover actionable insights
Predictive task
?
Community detection
CH2: Representing and managing knowledge in a machine-readable format.
Integrated Knowledge
Explicit and implicit knowledge
CH1: Extracting, modeling, and representing data from real-world data and concepts.
Heterogeneous data sources
Data and Concepts
Open Data Sources
PubMed
WIKIDATA
DRUGBANK
DBpedia
UniProt
Industry 4.0
Towards knowledge discovery

Figure 1.2: **Challenges**. Three main challenges are identified in this thesis toward knowledge discovery from real-world data and concepts. The first challenge **CH1**: modeling data from heterogeneous data sources. The second challenge **CH2**: managing knowledge in a machine-readable format. The third challenge **CH3**: enabling actionable insights to be uncovered and explored.

Challenge 1: Extracting, modeling, and representing data from heterogeneous data sources

Real-world data and concepts are the main sources of knowledge in healthcare, business, scientific and technical domains. However, the data are from heterogeneous data sources and may be incomplete and have missing associations to a given context; moreover, the data is generally not machine-readable. Data can be represented at different levels of structure, e.g., structured, semi-structured, and unstructured, and open data sources provide information at these three levels of structure. Data alone does not generate knowledge, while their connections in a specific context provide useful insights for knowledge discovery. Representing this knowledge in a machine-readable format that allows for the identification and semantic representation of entities is crucial for the work in this thesis. The modeling and knowledge representation mechanism selection defines the downstream constraints concerning machine reasoning, interoperability, and communication interfaces. The data and concepts represented in RDF are machine-readable and enable semantics to be added to the data and converted into meaningful actions to provide applications with new capabilities and facilitate knowledge exchange and richer experiences. Therefore, we need a unified knowledge representation that addresses the *structuredness* conflicts of heterogeneous data, such as different levels of structure, granularity or aggregation, and file formats.

Challenge 2: Representing and managing knowledge in a machine-readable format

Once the data and concepts have been transformed into a homogeneous model, the main challenge is to enrich the data with meaning and context. Adopting the RDF and OWL data models [102, 139], knowledge is represented by A-Box and T-Box composing the knowledge graph representing explicit knowledge. The A-Box (assertion box) represents the factual knowledge or data in a domain, and the T-Box (terminological box) contains terminological axioms that define the concepts and relationships within the domain. It specifies the general knowledge or the structural constraints that define the classes, properties, and relationships between them. The knowledge graphs constructed from real-world data suffer from data sparsity issues, i.e., explicit relations account for only a tiny part of all possible relations. However, by assuming the Open World Assumption (OWA), i.e., what is not known to be true is just unknown, KGs allow ameliorating the data sparsity issue. Making implicit knowledge explicit and machine-readable enables the knowledge graph to become meaningful in discovery tasks. Given human cognitive capabilities and experience-based knowledge, they can recognize the implicit knowledge in data. Assuming that the variable P_1 is a person and P_1 has a sibling P_2, explicitly, the fact that P_2 is a person is unknown, but implicitly it is. One of the tasks at hand in knowledge management is to transfer this implicit knowledge to knowledge graphs explicitly. However, the large number of data, properties, and concepts that the knowledge graph represents making implicit knowledge explicit have a high computational cost.

Challenge 3: Enabling actionable insights to be uncovered and explored

Knowledge discovery models require the capability to understand the structure of the data model, preserve its semantic meaning and infer new facts. The knowledge discovery task aims to find patterns that can be considered knowledge about data. This task becomes even more complex in the presence of heterogeneous data sources lacking an integrated model. Furthermore, once the knowledge is uncovered, it needs to be transformed into aggregate data that can be effectively utilized for further analysis and decision-making. This conversion process involves consolidating the discovered knowledge into a coherent representation that allows for easy interpretation and extraction of insights. To enhance the knowledge discovery process is crucial to leverage the implicit knowledge explicitly embedded within the knowledge graph. This implicit knowledge may include latent relationships, hidden patterns, or contextual information, which plays an essential role in uncovering actionable insights.

1.3 Research Questions

We derive the following research questions based on the main problem and associated challenges.

RQ1: How can metadata encoding data meaning be exploited to discover relationships in knowledge graphs?

To address this question, we provide a Neuro-Symbolic Artificial Intelligence approach that contributes features of symbolic and sub-symbolic systems. Our proposal aims to benefit from the advantages of symbolic and sub-symbolic paradigms. Our hybrid approach discovers knowledge given a target prediction in a knowledge graph and extracts unknown implicit information related to the target prediction. The symbolic system is implemented by a deductive database defined for a target prediction over a KG. Furthermore, the symbolic system enhances the predictive capacity of the sub-symbolic system implemented by a KGE model.

RQ2: How can heterogeneous data sources be integrated to obtain a unified knowledge representation?

We use a knowledge graph approach to answer this research question, considering the metadata describing the semantics encoded in the data. Semantic technologies can provide a comprehensive basis for building a knowledge model for linked data. We employ the concept of the knowledge graph to represent the data and discuss its benefits. Furthermore, knowledge graphs created from raw data are incomplete and need to be enriched with external knowledge to uncover missing associations. We define mapping assertions specified in the RDF Mapping Language (RML) [100] to generate the RDF graph.

RQ3: How can implicit knowledge be used to enhance knowledge discovery tasks?

To address this question, we have defined a symbolic system that relies on existing approaches of deductive database systems. The symbolic system corresponds to deductive databases that can derive new statements, e.g., conclude new facts, from inference rules and facts stored in the extensional database. Furthermore, the deductive database proposed is addressed to a target prediction which renders the computational complexity polynomial time. The evaluation of inference rules, implemented as an intensional database, concludes implicit knowledge on top of an extensional database. Then, the symbolic system implemented over the KG applies deductive reasoning to enhance KG completeness. Thus, data sparsity is reduced. Next, knowledge discovery models benefit from the facts deduced by the symbolic system. As a result, more accurate associations are uncovered.

RQ4: What is the impact of deductive reasoning on accurately uncovering knowledge?

To address this question, we empirically evaluate the effectiveness of our neuro-symbolic AI approach. We are interested in how deductive reasoning contributes to accurate knowledge discovery. We conduct an ablation study on the components of our system, considering state-of-the-art KG embedding methods. We also evaluate the use of deductive reasoning and non-reasoning. The observed results prove the benefits of deductive reasoning over deductive databases that accurately represent implicit and explicit knowledge.

RQ5: How can the proposed approach be applied to real-world cases?

To address this research question, we applied our symbolic system over three KGs, DE4LungCancer KG [3], Knowledge4COVID-19 KG [115], and iASiS KG [131]. We are interested in presenting the significant benefit of the discovery task on a knowledge graph. We illustrate the applicability of our approach to real scenarios in the biomedical domain, specifically in four projects, iASiS[5], BigMedilytics[6], P4-LUCAT[7], and H2020 CLARIFY[8]. The observed results indicate that our method can be relevant in all these applications.

1.4 Thesis Overview

In this section, we present an overview of the main contributions of this thesis and references to scientific publications supporting this work. Figure 1.3 summarizes the main contributions of this thesis.

[5] https://project-iasis.eu/

[6] https://www.bigmedilytics.eu/

[7] https://p4-lucat.eu/

[8] https://www.clarify2020.eu/

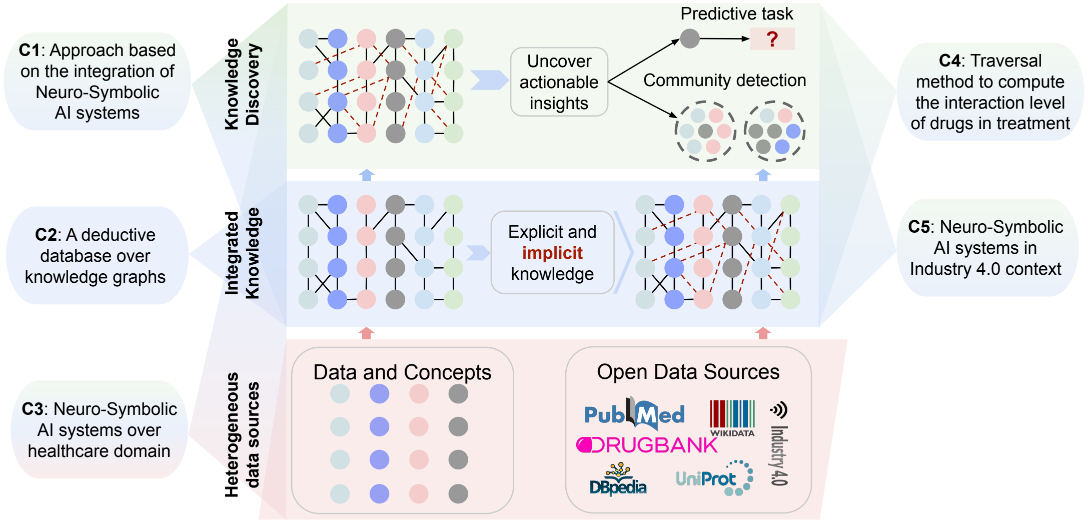

C1: Approach based on the integration of Neuro-Symbolic AI systems
C2: A deductive database over knowledge graphs
C3: Neuro-Symbolic AI systems over healthcare domain
Knowledge Discovery
Integrated Knowledge
Heterogeneous data sources
Uncover actionable insights
Predictive task
?
Community detection
Explicit and implicit knowledge
Data and Concepts
Open Data Sources
PubMed
WIKIDATA
DRUGBANK
DBpedia
UniProt
Industry 4.0
C4: Traversal method to compute the interaction level of drugs in treatment
C5: Neuro-Symbolic AI systems in Industry 4.0 context

Figure 1.3: **Thesis Contributions**. Four main contributions of this thesis towards uncovering actionable insights including **C1**: An approach based on the integration of neuro-symbolic artificial intelligence systems; **C2**: the implementation of our approach to predict polypharmacy treatment effectiveness; **C3**: a technique to compute the drug interaction score in polypharmacy treatment; and **C4**: determining relatedness across I4.0 standards.

1.4.1 Contributions

Contribution 1: *An approach based on integrating Neuro-Symbolic Artificial Intelligence systems.* The symbolic systems are the most prominent tools for modeling behavior, while sub-symbolic systems are for modeling cognition based on a vector representation. We propose a neuro-symbolic AI system that integrates symbolic-subsymbolic systems on top of knowledge graphs. This hybrid approach enhances the predictive capacity of the AI models on the knowledge graph. The symbolic system transfers the implicit knowledge to the knowledge graph explicitly. The symbolic system is specified in a deductive database implemented in Datalog that derives deductions. Thus, the data sparsity issue is minimized by considering the symbolic system. Embedding models implement the sub-symbolic system. Moreover, KGE models are known to suffer from data sparsity, and the symbolic system assists in overcoming this fact. Our approach enhances the properties of the adjacent vertices to the entities in a target prediction through the symbolic system. Therefore, KGE models better represent the entities in the KG. The neuro-symbolic AI system proposed is domain-agnostic and could be applied to any predictive task on KG. This proposal addresses the research question **RQ1**.

Contribution 2: *A deductive database over knowledge graphs.* We propose a deductive database over a knowledge graph based on the existing approaches of deductive database systems. The deductive database is implemented in Datalog. A deductive database is a system that can derive deductions, e.g., conclude new facts, from inference rules and facts stored in the extensional database. Datalog considers two sets of clauses: a set of ground facts called the Extensional Database (EDB) and a Datalog program P called the Intensional Database (IDB). In the proposed deductive system, the Datalog rules stated in the IDB represent the experts' knowledge explicitly transferred to the knowledge graph. Since Datalog is a declarative language, the rules are defined declaratively, facilitating their definition. Thus, the deductive system allows writing Datalog rules over an RDF graph. Furthermore, the proposed deductive system represents the experts' knowledge and explainability with logical rules. This contribution aims to answer the research question **RQ3**.

Contribution 3: *Neuro-Symbolic systems over healthcare domain.* As a proof of concept, we have implemented our neuro-symbolic system on top of a KG for lung cancer to predict polypharmacy treatment effectiveness. Polypharmacy is the concurrent use of multiple drugs in treatments, and it is a standard procedure to treat severe diseases, e.g., lung cancer. We integrate treatments, their prescribed drugs, drug-drug interactions, and drug-protein interactions into a knowledge graph. The knowledge graph of polypharmacy treatment responses is populated with descriptions of more than 420 oncological treatments. The missing associations and incompleteness of the data are overcome by the integration with open data sources. Finally, the knowledge graph has been linked to existing open web sources such as DrugBank[2], Wikidata[3], Uniprot[4], DBpedia[9], and Pubmed[10]. The symbolic system implements a deductive system to infer pharmacokinetic drug-drug interactions. The intensional database comprises Horn rules that model the different types of pharmacokinetic drug-drug interactions and the effects of combining them. As a result, pharmacokinetics drug-drug interactions (DDIs) can be deduced in medical treatments. A pharmacokinetic DDI is deduced when a set of drugs are part of a treatment, and the rule applied is valid. Our deductive system captures the knowledge represented in the RDF graph and deduces the unknown DDIs encoded in a set of rules through the Datalog program. Empirical evaluations demonstrate the effectiveness of the symbolic system deducing pharmacokinetics DDIs in treatments. The implemented symbolic system is integrated into the sub-symbolic system performed by KGE models. The empirical results put the deduction power of deductive databases into perspective; they improve the predictive capacity of existing KGE models, answering the research questions **RQ1**, **RQ2**, and **RQ4**.

Contribution 4. *Traversal method to compute the interaction score of a drug in*

[9] https://www.dbpedia.org/

[10] https://pubmed.ncbi.nlm.nih.gov/

treatment. We propose a method based on the computation of wedges [142] in a knowledge graph to measure the interaction score of drugs in treatment. A wedge w is a path with two edges in a directed labeled graph; w is composed of three vertices $\{a, b, c\}$ and two ordered pairs of edges $\{(a, b), (b, c)\}$ of the directed labeled graph. The vertex b is the middle vertex of w. We apply the wedge concept to the DDIs knowledge graph, where the edges of a wedge represent DDIs. The proposed method computes the distribution of the middle-vertex of wedges. The middle vertex is particularly important in the wedge because it is both the object drug of one interaction and the precipitating drug of another. Our method consists of a deductive database implemented in Datalog. Furthermore, the method provides a ranking of drugs measuring the interaction score in treatment. We empirically evaluate the effectiveness of our metric on treatments of three diseases, COVID-19, Alzheimer's, and Parkinson's disease. The experiments use a knowledge graph of 216 polypharmacy COVID-19 treatments that comprise COVID-19 drugs and drugs for the most common comorbidities that impact the survival of COVID-19 patients [31]. The observed results show the benefits of our traversal method concerning four interaction checker web tools, suggesting that drugs with the higher frequency of middle vertex have a higher interaction score in treatment. Furthermore, experts in the domain evaluated with successful outcomes the results of our method with oncological treatments. With this proposal, research question **RQ3** is addressed.

Contribution 5. *Neuro-Symbolic Artificial Intelligence systems in Industry 4.0 context.* Industry 4.0 (I4.0) standards and standardization frameworks provide a unified way to describe smart factories. Standards specify the components, systems, and processes inside a smart factory and their interaction. Different industrial communities have defined standardization frameworks aligning standards according to their features and expressiveness. As a result, interoperability conflicts are generated whenever smart factories are described with miss-classified standards. We address the problem of determining relatedness across I4.0 standards described in terms of their main features and standardization frameworks. Our goal is to uncover alignments among related standards, i.e., standards that define the same type of smart factory components. The proposed neuro-symbolic AI system evaluated in Industry 4.0 demonstrates its effectiveness in determining relatedness among standards and analyzing their properties to detect unknown relations. Furthermore, the symbolic system deduces new relations between standards based on their properties' characteristics and experts' knowledge. Thus, the symbolic system explicitly transfers the implicit knowledge concerning standards to the knowledge graph. The sub-symbolic system resorts to KGE to determine relatedness among standards based on similarity metrics. Next, community detection algorithms can automatically create communities of highly similar standards based on the similarity values. Thus, alignments across standards are predicted, and the standardization frameworks containing the new alignments are able to minimize the interoperability issues across the smart factories. The empirical evaluation is performed on a knowledge graph of 249 I4.0 standards. Our results suggest that relations among standards can be detected accurately. This contribution allows us to answer the research question **RQ1**.

1.4.2 List of Publications

This thesis is based on the following publications.

Papers in Proceedings of Peer-Reviewed Conferences

- **Ariam Rivas**, Maria-Esther Vidal: *Capturing Knowledge about Drug-Drug Interactions to Enhance Treatment Effectiveness.* K-CAP '21: Proceedings of the 11th on Knowledge Capture Conference (2021). Ariam Rivas is the first author of this article. Ariam Rivas defined the problem and motivating example, the development of the approach, the revision of the state-of-the-art approaches, the development of the software, and the execution and analysis of the experiments and results. *Nominated to the best student paper.*

- **Ariam Rivas**, Irlan Grangel-Gonzalez, Diego Collarana, Jens Lehmann, and Maria-Esther Vidal: *Unveiling Relations in the Industry 4.0 Standards Landscape Based on Knowledge Graph Embeddings.* In Proceeding of the 31st International Conference of Database and Expert Systems Applications (DEXA 2020). Ariam Rivas is the first author of this article. Ariam Rivas defined the problem and motivating example, the development and implementation of the approach, the revision of the state-of-the-art approaches, and the execution and analysis of the experiments and results.

Peer-Reviewed International Journals

- **Ariam Rivas**, Diego Collarana, Maria Torrente, Maria-Esther Vidal. *A Neuro-Symbolic System over Knowledge Graphs for Link Prediction.* In: Semantic Web Journal (2022). Ariam Rivas is the first author of this article. He defined the problem definition and motivating example, the development, and implementation of the approach, the development of the polypharmacy treatment knowledge graph, the revision of the state-of-the-art approaches, and the execution and analysis of the experiments and results.

- Fotis Aisopos, Samaneh Jozashoori, Emetis Niazmand, Disha Purohit, **Ariam Rivas**, Ahmad Sakor, Enrique Iglesias, Dimitrios Vogiatzis, Ernestina Menasalvas, Alejandro Rodriguez Gonzalez, Guillermo Vigueras, Daniel Gomez Bravo, Maria Torrente, Roberto Hernández López, Mariano Provencio Pulla, Athanasios Dalianis, Ana Triantafillou, Georgios Paliouras and Maria-Esther Vidal. *Knowledge Graphs for Enhancing Transparency in Health Data Ecosystems.* In: Semantic Web Journal (2022). Ariam Rivas is one of the first authors of this paper. He contributed to the knowledge-driven ecosystem in the context of lung cancer and the DE4LungCancer knowledge graph creation pipeline. Moreover, I contributed to the exploration of DE4LungCancer KG, assessing the impact of DDIs on the effectiveness of lung cancer treatment's response, computational analysis, statistical tests, writing–reviews, and editing.

- Ahmad Sakor, Samaneh Jozashoori, Emetis Niazmand, **Ariam Rivas**, Konstantinos Bougiatiotis, Fotis Aisopos, Enrique Iglesias, Philipp D. Rohde, Trupti Padiya, Anastasia Krithara, Georgios Paliouras, Maria-Esther Vidal: *Knowledge4COVID-19: A Semantic-based Approach for Constructing a COVID-19 related Knowledge Graph from Various Sources and Analysing Treatments' Toxicities.* Journal of Web Semantics (2022). Ariam Rivas contributed to the definition of the Scientific Open Data Ecosystem, the Knowledge4COVID-19 knowledge graph creation pipeline, and exploration for detecting relevant adverse effects on Knowledge4COVID-19, to apply the deductive system to deduce DDIs from 216 polypharmacy COVID-19 treatments, and the computational analysis, evaluation, validation, writing–review, and editing.

- **Ariam Rivas**, Irlan Grangel-Gonzalez, Diego Collarana, Jens Lehmann, and Maria-Esther Vidal: *Discover Relations in the Industry 4.0 Standards Via Unsupervised Learning on Knowledge Graph Embeddings.* Journal of Data Intelligence (2020). Ariam Rivas is the first author of this article. He defined the problem and motivating example, the development and implementation of the approach, the revision of the state-of-the-art approaches, and the execution and analysis of the experiments and results.

- Maria-Esther Vidal, Kemele M. Endris, Samaneh Jazashoori, Ahmad Sakor, **Ariam Rivas**: *Transforming Heterogeneous Data into Knowledge for Personalized Treatments - A Use Case.* Datenbank-Spektrum volume 19, pages 95–106 (2019). Ariam Rivas contributed to the knowledge-driven framework for supporting personalized medicine, evaluation, knowledge discovery, and statistical test. Ariam defined and implemented the similarity measure to quantify the similarity between patients and to perform the semEP [94] community detection algorithm to discover patterns between patients that share similar properties in the iASiS KG [131].

1.5 Thesis Structure

This thesis is structured in seven chapters which are outlined as follows:

- Chapter 1 covers the main research problem and challenges, research questions, the contributions that address research questions, and a list of published scientific articles.

- Chapter 2 introduces the basic concepts in the field of the Semantic Web, knowledge graph, and symbolic and sub-symbolic systems that are required to understand the work of this thesis.

- Chapter 3 examines the current state-of-the-art research work to provide a clearer picture of the research conducted in this thesis. First, we discuss a complete view of generic neuro-symbolic AI approaches. Secondly, we present techniques and models to represent data and discover knowledge. Finally, we show existing methods for uncovering actionable insights in Industry 4.0 and the healthcare domain.

- Chapter 4 presents a neuro-symbolic artificial intelligence approach over KGs. We show a domain-agnostic approach able to capture the implicit knowledge in a KG by a symbolic system and enhance the predictive capacity of sub-symbolic systems. We implement the neuro-symbolic system proposed on top of a KG for lung cancer to predict polypharmacy treatment effectiveness as a proof of concept. We create the KG of polypharmacy lung cancer treatments and perform an extensive evaluation of the symbolic-subsymbolic system in state-of-the-art KGE models.

- Chapter 5 presents the problem of finding relations among I4.0 standards described in terms of their main features and standardization frameworks. The neuro-symbolic AI system proposed is evaluated in I4.0, demonstrating its effectiveness in determining relatedness among standards and analyzing their properties to detect unknown relations.

- Chapter 6 presents the deductive database *DS* to compute the interaction score of drugs in treatment based on the wedge concept. We apply the *DS* in the biomedical domain, specifically in four projects, iASiS[5], BigMedilytics[6], P4-LUCAT[7], and H2020 CLARIFY[8] for assessing the impact of DDIs on the effectiveness of lung cancer and COVID-19 treatments. Furthermore, we present a similarity measure that evaluates patient similarity, and a knowledge discovery technique is used to uncover patterns in iASiS KG.

- Chapter7 finalizes the thesis with a summary of the main results and contributions to the problem of uncovering actionable insights from heterogeneous data sources. In addition, we discuss the limitations of the work and propose possible directions for future work.

Chapter 2

Background

In this chapter, we present the basic concepts and theoretical foundations of the research conducted in this thesis. Section 2.1 describes the basic concepts and formalisms of the Semantic Web; they include standards and models such as Resource Description Framework (RDF), SPARQL query language, and knowledge graphs. Next, Section 2.2 examines deductive database systems.Finally, section 2.3 introduces neuro-symbolic artificial intelligence and describes the symbolic and sub-symbolic systems.

2.1 The Semantic Web

The *Semantic Web* is an extension of the existing *Web of Documents*, where documents are transformed into objects by annotating them and rendering their meanings explicit. Tim Berners-Lee first proposed the idea of the Semantic Web in 2001 to add context information within the data itself that describes concepts in the real world [12]. The Semantic Web exists as a vision to extend principles of the existing Web to the Web of Data, such that computers can be more useful work [118]. The Semantic Web provides a framework for representing and accessing data. Semantic Technologies represent a set of standards, protocols, and technologies to create data stores, vocabularies, and rules written for handling data. The core technology stack of the Semantic Web consists of a set of standards: Resources Description Framework (RDF)[102] is a data model for data exchange, Resource Description Framework Schema (RDFS) which provides a syntax for defining schemes, and the Web Ontology Language (OWL) for expressing logical axioms. These standards adopt the principles of knowledge representation languages in the context of the Web. Linked Data is created following a set of principles proposed by Tim Berners-Lee for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web. Linked Data makes

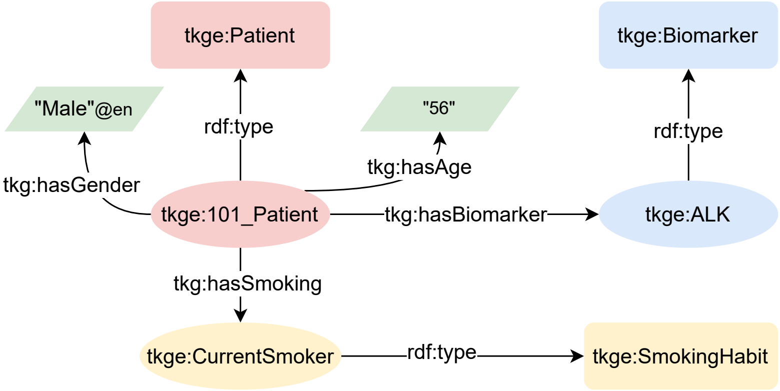


Figure 2.1: **Example of RDF graph**.The resource `101_Patient` is a patient which is connected through the predicate `hasBiomarker` to the resource `ALK` which is defined as a biomarker and through the predicate, `hasSmoking` to the entity `CurrentSmoker` defined as a smoking habit. In addition, age and gender are known for the resource `101_Patient`.

available semi-structured data sources on the Web for both machines and humans. Linked Open Data (LOD) is linked data released under an open licence [11]. The inception of LOD encouraged data providers to publish large linked datasets from diverse domains, which has led to the creation of a semantically linked global data space, called the Linked Open Data Cloud (LOD Cloud) [15]. LOD Cloud has grown substantially: in 2007, there were 12 datasets with 19 links, and in August 2021, there were 1,512 datasets with 16,174 with around 413,734,019,304 triples[1].

2.1.1 The Resource Description Framework

RDF is a recommended standard by the World Wide Web Consortium (W3C) [47] that specifies the architecture, syntax, and semantics describing resources on the Web. RDF is conceived as a machine-readable format whose syntax, grammar, and semantics are interoperable across different architectures. The main building block of RDF is a triple. RDF triple is a positive statement and is composed of subject, predicate, and object, where:

- A subject denotes a resource described by predicate and object; only URIs or blank nodes can be subjects in RDF.
- A predicate represents a property that relates the subject to the object; only URIs can be predicated in RDF.

[1] `https://lod-cloud.net/`

- An object specifies a value of the predicate; URIs, blank nodes, or literals can represent a value of the predicate.

Formally, RDF triple is defined as follows [6, 97]:

Definition 2.1.1 (RDF triple [6])**.** *Let* $\mathbf{I}$, $\mathbf{B}$, $\mathbf{L}$ *be disjoint infinite sets of URIs, blank nodes, and literals, respectively. A tuple* $(s\ p\ o) \in (\mathbf{I} \cup \mathbf{B}) \times \mathbf{I} \times (\mathbf{I} \cup \mathbf{B} \cup \mathbf{L})$ *is an RDF triple, where* s *is the subject,* p *is the property, and* o *is the object.*

Example 2.1.1. *A set of RDF triples is called an RDF dataset. Figure 2.1 illustrates an example of an RDF graph, where the edge (`tkge:101_Patient tkg:hasBiomarker tkge:ALK`) represents an RDF triple. The entity `tkge:101_Patient` corresponds to the subject, `tkg:hasBiomarker` and `tkge:ALK` represent a property and an object, respectively.*

Definition 2.1.2 (RDF Graph [97])**.** *An RDF graph* $G = (V, E, L)$ *is a labeled directed graph where nodes represent resources, and labels stand for properties:*

- *An RDF triple* $(s\ p\ o) \in E$*, corresponds to an edge in* E *from node* s *to node* o*;* p *is the label of the edge denoting the property that relates both nodes;*
- s*,* $o \in V$*,* s *corresponds to a subject and* o *corresponds to an object; and*
- $p \in L$ *is an edge label corresponding to a property.*

Example 2.1.2. *Figure 2.1 illustrates a portion of an RDF graph that describes a lung cancer patient. Nodes correspond to resources, and edges represent properties of RDF vocabularies. The graph contains URIs, e.g., `tkge:Patient` as a class, `tkge:101_Patient` as an instance of this class, and literals e.g., the gender of a patient expressed in English.*

2.1.2 The SPARQL Language

The SPARQL Protocol and RDF Query Language (SPARQL)is recommended by W3C [34] for querying RDF data through graph patterns. SPARQL queries consist of three parts [6]; pattern matching, solution modifiers, and output type. Pattern matching includes several graph matching features, such as optional, join, nesting, filtering values, and choice of the data source to be matched to the pattern. Solution modifiers allow modifying the values computed by the pattern matching part. The solution modifiers define the projection, distinct, group, order, and limit operators. Finally, the output type can be "yes/no", pattern-matching variables values, new RDF data construction, and resource descriptions. A SPARQL query contains a head and a body, where the body is an RDF graph pattern expression with the possibility to include triple patterns, conjunctions, optional parts, and constraints over the values of the variables. The head of the query indicates how to construct the answer to the query. The evaluation of a SPARQL query against an RDF graph is performed in two steps. First, the query body is matched to the RDF graph to obtain a set of bindings for the variables in the query body. Next, using the header information of the query, these bindings are processed by applying classical relational operators to produce the query answer. SPARQL defines operators, *OPTIONAL*, *UNION*, *FILTER*, and *AND*, to construct graph pattern expressions. The syntax of the SPARQL graph pattern is defined as:

Definition 2.1.3 (SPARQL Graph Pattern Expression [6])**.** *Let F be an infinite set of variables disjoint from $I \cup B \cup L$. A SPARQL graph pattern expression is defined recursively as follows:*

1. *A triple pattern $t \in (I \cup B \cup F) \times (I \cup F) \times (I \cup B \cup L \cup F)$ is a graph expression,*

2. *If P_1 and P_2 are graph patterns, then expressions $(P_1\ AND\ P_2)$, $(P_1\ OPT\ P_2)$, and $(P_1\ UNION\ P_2)$ are graph patterns,*

3. *If P is a graph pattern and R is a SPARQL built-in filter condition, then the expression $(P\ FILTER\ R)$ is a graph pattern.*

Example 2.1.3. *illustrates a SPARQL graph pattern composed of a set of RDF triples patterns, OPTIONAL, FILTER, and AND (.) operators.*

```
{
    ?patient a tkge:Patient .
    ?patient tkg:hasGender ?gender .
    ?patient tkg:hasAge ?age .
    OPTIONAL{?patient tkg:hasBiomarker ?biomarker .}
    FILTER(?age > "50")
}
```

Furthermore, SPARQL defines four query forms: *SELECT*, *CONSTRUCT*, *ASK*, and *DESCRIBE*. The *SELECT* query is the most commonly used form, returning a set of bound variables, i.e., RDF terms of the graph that satisfy the given graph pattern. *CONSTRUCT* query returns an RDF graph specified according to the given graph pattern. *ASK* query returns a Boolean value TRUE if the given graph pattern has solutions in the target RDF dataset or FALSE otherwise. The result of a *DESCRIBE* query is an RDF graph with relevant information about the URIs in a graph pattern. Search engines are in charge of determining this relevant information. In this thesis, we focus on SPARQL *SELECT* queries formally defined as follows:

Definition 2.1.4 (SPARQL SELECT query [117])**.** *Let Q be a SPARQL expression and $Z \subset F$ be a finite set of variables. A SPARQL select query is an expression of the form $SELECT_Z(Q)$.*

Example 2.1.4. *represents a SPARQL SELECT query composed of a graph pattern expression, AND, OPTIONAL, and FILTER operators, and solution modifier DISTINCT. The SPARQL query retrieves clinical data about the patients. The query projects the unique values that are mapped to the variables ?gender ?age and ?biomarker in the graph pattern expression.*

```
PREFIX tkg: <http://research.tib.eu/lung-cancer/vocab/>
PREFIX tkge: <http://research.tib.eu/lung-cancer/entity/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?gender ?age ?biomarker
WHERE {
        ?patient a tkge:Patient .
        ?patient tkg:hasGender ?gender .
        ?patient tkg:hasAge ?age .
        OPTIONAL {?patient tkg:hasBiomarker ?biomarker .}
        FILTER (?age > "50")
     }
```

2.1.3 Knowledge Graphs

An RDF knowledge graph (KG) is a data structure representing factual knowledge with entities and their relationships using a data graph [49]. Knowledge graphs enable the description of the meaning of data, the integration of data from heterogeneous sources, and the discovery of unknown patterns. KGs are used in countless domains because of their ability to model data in a machine-readable form. Thus, the knowledge semantically represented in knowledge graphs can be exploited to solve a broad range of problems. The meaning of the data is stored together with the data in the graph in the form of ontologies. This makes knowledge graphs self-descriptive, a unique location for finding and understanding data. The semantics of data are explicit and include formalisms for supporting inferencing. Google used the knowledge graph concept in 2012 and launched the Google knowledge graph [121]. There are several publicly available knowledge graphs such as Wikidata[3] [133], DBpedia[9], UniProt[4], etc. DBpedia [73] is a knowledge graph that uses Wikipedia to extract information to be represented in RDF format.

Although there is no agreement upon a formal definition of knowledge graphs, in this thesis, we use the definition presented by Paulheim [27]. A knowledge graph is described as follows:

- mainly describes real-world entities and their interrelations, organized in a graph;
- defines possible classes and relations of entities in a schema;
- allows for potentially interrelating arbitrary entities with each other;
- covers various topical domains.

Following the description presented by Paulheim, a data model is required to build a knowledge graph. Given that RDF is a data model to describe resources on the Web of Data, we define a KG as an RDF Graph.

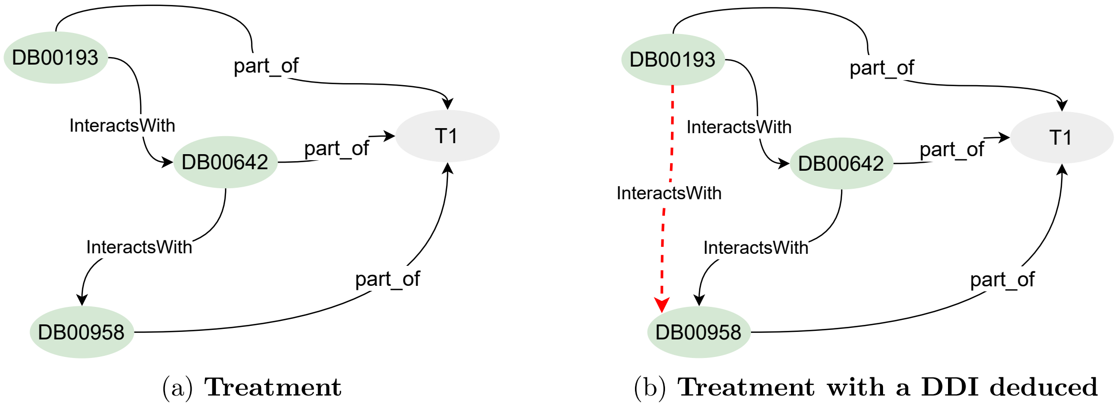
2.2 Deductive Systems

Deductive database systems rely on a query language designed around a logical data model. Relationships are considered as the value of a logical predicate. Deductive database systems are an advanced form of relational systems [101]. Deductive database systems are best suited for applications where a huge amount of data must be accessed and complex queries must be performed. Deductive systems share with relational systems the important property of being declarative, i.e., to allow the user to query or update by his/her requests instead of how to operate. A deductive database is a system that can derive deductions, e.g., conclude new facts, from inference rules and facts stored in the database [101]. Deductive systems maintain deductive qualities in the rules that are stated in the system.

2.2.1 Notation of Deductive Database System

Deductive database systems structure information into two categories, data or ground facts and rules. Facts are represented by a predicate with constant arguments and are assertions about a relevant part of the domain, such as `Joe has biomarker ALK positive`. Rules are sentences that deduce facts from other facts, such as `If X has biomarker positive, then X has Cancer`.

The facts and rules are represented as Horn clauses [22]. Horn clauses are represented in the following notation: $L_0 \Leftarrow L_1, ..., L_n$, where each L_i is a literal of the form $p_i(t_{i1}, ..., t_{ik})$. P_i is a predicate symbol, and t_i are terms. A term is either a constant or a variable. The left-hand side of a rule is the head, and the right-hand side is its body. Clauses with an empty body represent facts, while clauses with at least one literal in the body represent rules. The fact `Joe has biomarker ALK positive` can be represented as $biomarker(alk, joe)$. The rule `If X has biomarker positive, then X has Cancer` can be represented as



(a) **Treatment**

(b) **Treatment with a DDI deduced**

Figure 2.2: **Example of deduction.** Figure 2.2a shows a treatment, *T1*, represented in an labelled directed graph. The drugs *DB00193*, *DB00642*, and *DB00958* are part of *T1*. The drug-drug interactions are represented by the property *interactsWith*. Figure 2.2b depicts the ideal labeled directed graph, where a symbolic system generates a new DDI between *DB00193* and *DB00958*.

$cancer(X) \Leftarrow biomarker(Y, X)$. The symbols *biomarker* and *cancer* are predicate symbols. The symbols *alk* and *joe* are constants, and the symbols X and Y are variables. The language used in this thesis to specify facts, rules, and queries in deductive databases is Datalog.

2.2.2 Datalog a rule-based language

Datalog is a declarative programming language used to work with deductive databases. Datalog is specifically designed to interact with large databases, and a significant contribution comes from the integration of logic programming and databases. A Datalog program is a set of rules represented as Horn clauses [22]. A Datalog program P must satisfy the following safety conditions; each fact of P is ground, and each variable that occurs in the head of a rule of P must also occur in the body of the same rule. A rule is safe if all its variables are bound, where any variable appearing as an argument in a body predicate is bound. Those conditions ensure that all facts deduced from P are finite. Datalog considers two sets of clauses: a set of ground facts called the Extensional Database (EDB) and a Datalog program P called the Intensional Database (IDB). The predicates in the EDB and IDB are divided into two disjoint sets, EDB-predicates, which occur in the EDB, and the IDB-predicates, which occur in P but not in the EDB. Furthermore, the head predicated of each clause in P is an IDB-predicate, and the

EDB-predicate can occur in the body of the rule. The IDB-predicates of P can be identified with relations, called IDB-relations, which are not stored explicitly.

An example of EDB is the set of facts $E = \{interactsWith(DB00193, DB00642),\ interactsWith(DB00642, DB00958)\}$ expressed as an labelled directed graph in Figure 2.2a. The predicate $interactsWith$ represents interactions between two drugs. Let P_t be a Datalog program containing the following clauses:

$$\begin{aligned} interactsWith(A, X) \Rightarrow& \\ \mathit{inferredInteraction}(A, X).& \quad (r1) \\ interactsWith(A, B), \mathit{inferredInteraction}(B, X) \Rightarrow& \\ \mathit{inferredInteraction}(A, X).& \quad (r2) \end{aligned}$$

Rule $r2$ states that exist an *inferredInteraction* between drug A and X, if there is another drug B which interacts with A with the predicate $interactsWith$, and there is an *inferredInteraction* from B to X. The evaluation results of $r2$ is $\{\mathit{inferredInteraction}(DB00193, DB00958)\}$, which is observed in Figure 2.2b. Program P_t (1) can be considered as a query against the EDB producing as an answer the predicate *inferredInteraction*. Thus, the differentiation between the two sets of clauses, EDB, and P_t, has a clear meaning. The EDB is considered a time-varying collection of data. On the other hand, the program P_t (1) is a time-invariant mapping that relates a result to each possible database state.

2.2.3 Semantics of Datalog

A Herbrand Base (HB) [22] is the set of all ground facts that can be expressed in all the predicates and constants in P_t. The extensional part of the Herbrand base is denoted by EHB, i.e., all literals of HB whose predicate is an EDB-predicate. Similarly, the set of all literals of HB whose predicate is an IDB-predicate denotes IHB. Let S be a finite set of Datalog clauses, we denote by $cons(S)$ the set of all the facts that are logical consequences of S. The semantics of a Datalog program can be described as mapping $\mathcal{M}p$ from EHB to IHB. $\mathcal{M}p$ mapping each possible extension database $E \in EHB$ associates the set of $\mathcal{M}p(E)$ of intensional results facts defined by $\mathcal{M}p(E) = cons(P_t \cup E) \cap IHB$. Let I and J be two literals. I subsumes J, denoted by $I \triangleright J$, if exist a substitution θ of variables such that $I\theta = J$, i.e., applying θ to I gives J. If $I \triangleright J$, we can say that J is an instance of I, e.g., $p(a, b, b)$ and $p(c, c, c)$ are both instances of $p(X, Y, Y)$ while $p(b, b, a)$ is not.

Model theory is a branch of mathematical logic that defines the semantics of formal systems. In the context of Datalog, an interpretation consists of assigning a specific meaning to constant and predicate symbols. The concept of logical consequence can be defined as follows: a fact F follows logically from a set S of clauses, if and only if (iff) each interpretation satisfying every clause of S also satisfies F. If F follows from S, we write $S \models F$. Datalog considers a particular interpretation, Herbrand Interpretation, which assigns to each constant symbol a lexicographic entity. Therefore, two non-identical Herbrand interpretations only differ in the respective interpretations of the predicate symbols. Herbrand interpretation can be identified with a subset $\mathcal{I}$ of the Herbrand base HB. This subset contains all the ground facts which are true under the interpretation. Thus, a ground fact $p(c_1, ..., c_n)$ is true under the interpretation $\mathcal{I}$, iff $p(c_1, ..., p_n) \in \mathcal{I}$. A Datalog rule of the form $L_0 \Leftarrow L_1, ..., L_n$ is true under $\mathcal{I}$ iff for each substitution θ which replaces variables by constants, whenever $L_1\theta \in \mathcal{I} \wedge ... \wedge l_n\theta \in \mathcal{I}$ then it also holds that $L_0\theta \in \mathcal{I}$. A Herbrand interpretation that satisfies a clause C is called a Herbrand model for C. Considering the following Herbrand interpretations:

$$\begin{aligned}\mathcal{I}_1 = \{&interactsWith(DB00338, DB00361), interactsWith(DB00361, DB00642),\\ &interactsWith(DB00642, DB00641), \mathit{inferredInteraction}(DB00338, DB00361),\\ &\mathit{inferredInteraction}(DB00361, DB00642), \mathit{inferredInteraction}(DB00642, DB00641)\}.\end{aligned}$$

we can observe that $\mathcal{I}_1$ is not a Herbrand model of the program P_t (1), while Herbrand model $\mathcal{I}_2$ is a Herbrand model of P_t (1), where:

$$\begin{aligned}\mathcal{I}_2 = \mathcal{I}_1 \cup \{&\mathit{inferredInteraction}(DB00338, DB00642),\\ &\mathit{inferredInteraction}(DB00361, DB00641),\\ &\mathit{inferredInteraction}(DB00338, DB00641)\}.\end{aligned}$$

2.2.4 Inference of Datalog Rules

Datalog rules enable new facts to be produced from given facts. Datalog allows inferring all ground facts that are a consequence of a finite set of Datalog clauses. Let a Datalog rule R of the form $L_0 \Leftarrow L_1, ..., L_n$ and a list of ground facts $F_1, ..., F_n$. We can infer in one step the fact $L_0\theta$ from the rule R and from the facts $F_1, ..., F_n$, if a substitution θ exists such that for each $1 \leq i \leq n$ $L_i\theta = F_i$. The inferred fact may be a new fact or a known fact. This general inference rule is the Elementary Production Principle (EPP). EPP produces new facts from given rules and facts. Considering the Datalog rule $r1$ of program P_t (1) and the EDB E_1, where:

$$\begin{aligned}E_1 = \{&interactsWith(DB00338, DB00361),\\ &interactsWith(DB00361, DB00642), interactsWith(DB00642, DB00641)\}.\end{aligned}$$

We can infer in one step *inferredInteraction(DB00642,DB00641)* from the fact $interactsWith(DB00642, DB00641)$. The substitution used was $\theta = \{X \leftarrow$

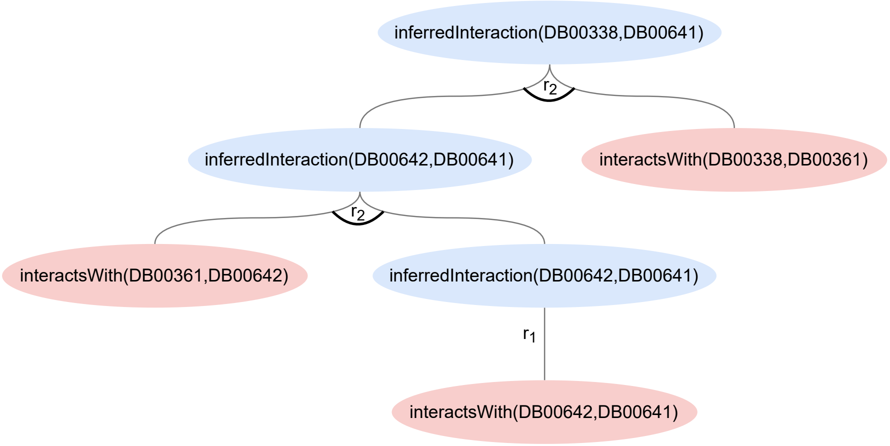

inferredInteraction(DB00338,DB00641)
r2
inferredInteraction(DB00642,DB00641)
interactsWith(DB00338,DB00361)
r2
interactsWith(DB00361,DB00642)
inferredInteraction(DB00642,DB00641)
r1
interactsWith(DB00642,DB00641)

Figure 2.3: **Proof tree**. The sequence of applications of EPP to infer the ground fact *inferredInteraction(DB00338,DB00641)* in the Datalog program P_t (1).

$DB00642, A \leftarrow DB00641\}$. Considering the rule $r2$ and the facts *interactsWith(DB00361, DB00642)* and *interactsWith(DB00642, DB00641)*, we can infer in one step *inferredInteraction(DB00361,DB00641)* by applying EPP and using the substitution $\theta = \{A \leftarrow DB00361, X \leftarrow DB00641, B \leftarrow DB00642\}$.

A ground fact F can be inferred from a set of Datalog clauses S, denoted by $S \vdash F \equiv F \in S$, or F can be obtained by applying the inference rule EPP a finite number of times. The relationship $\vdash$ is defined as follows:

- $S \vdash F$ if $F \in S$.
- $S \vdash F$ if a rule $R \in S$ and ground facts $F_1, ..., F_n$ exist such that $\forall 1 \leq i \leq n$ $S \vdash F_i$ and F can be inferred in one step by the application of EPP to R and $F_1, ..., F_n$.

The sequence of EPP applications to inferring a ground fact F from S is called a proof of F from S. A proof can be represented as a proof tree with different levels and with the derived fact F at the root node. Let S_t denote the set of all the clauses in the program P_t and the EDB E_1, i.e., $S_t = P_t \cup E_1$. Figure 2.3 shows the proof tree of $S_t \vdash$ *inferredInteraction(DB00338,DB00641)*.

Stefano Ceri et.al. [22] present a method of computing $cons(S)$, the set of all the facts that are logical consequences of a finite set of Datalog clauses:

Algorithm 1 Infer Algorithm

```
Input: a finite set $S$ of Datalog clauses.
Output: $cons(S)$.
```

```
$W \Leftarrow S$;
while EPP applies to some rule and facts of $W$ producing a new ground fact $F \notin W$ do
    $W \Leftarrow W \cup \{F\}$;
end while
return $(W \cap HB)$
```

Algorithm 1 terminates and outputs a finite set of facts $cons(S)$. The order of algorithm 1 in generating new facts corresponds to the bottom-up order of the proof tree. The principle on which algorithm 1 relies is called bottom-up evaluation. This principle in artificial intelligence is known as forward chaining because Datalog rules are processed forward, from premises to conclusions.

2.3 Neuro-Symbolic Artificial Intelligence

Neuro-Symbolic Artificial Intelligence is a field of Artificial Intelligence (AI) that combines symbolic and sub-symbolic AI models [14, 39, 116]. The symbolic models refer to AI approaches based on handling explicit symbols that refer to representations of reasoning and explainability, while sub-symbolic creates distributed vector-based representations of data rather than logical or symbolic representations [14, 54]. Neuro-symbolic AI focuses on integrating symbolic and sub-symbolic systems. The goals are to provide a unifying view of logic and connectionism, contribute to the modeling and understanding of cognition, and produce better models for integrating machine learning and reasoning. However, symbolic and sub-symbolic systems differ fundamentally in how they represent data and information.

According to research [14, 39, 51, 59], information at different levels of abstraction differs in structure and composition, where higher levels of abstraction are symbolic and lower levels of abstraction are sub-symbolic. Symbolic systems typically use structured representation languages from formal logic, and sub-symbolic systems usually use representations based on vector space. Neuro-symbolic integration comprises translating symbolic knowledge into the sub-symbolic system, learning additional knowledge by the sub-symbolic system, and extracting symbolic knowledge from the sub-symbolic system. Knowledge extraction provides an incremental explanation and learning of the neuro-symbolic system. Several approaches employ translation algorithms from a symbolic representation to a subsymbolic representation and vice versa [14, 82].Our work integrates a domain-agnostic symbolic system with a Knowledge Graph Embeddings (KGE) model to reduce the KG sparsity towards improving the model's predictive capability. Thus, we broaden the scope and applicability in several domains of neural-symbolic integration.

2.3.1 Symbolic Systems

The symbolic systems rely on explicit knowledge representation through formal or logical languages [54]. The symbolic systems address high-level deductive reasoning, logic inference, and rule-based search algorithms to solve a specified model. Furthermore, they depend on embedding human knowledge and rules of procedure in the system. Symbolic systems feature in the formal modeling of complex tasks and human behavior. Such systems are based on complex symbolic data structures, e.g., graphs, trees, shapes and grammar, and symbolic logic. Expert systems represent an attempt to work on symbolic reasoning [112]. Expert systems typically consist of two components, a knowledge base that stores facts and rules and an inference engine that performs the actual reasoning. Therefore, expert systems have significant expressive power and are straightforward to interpret and validate [122]. However, inference engine models have an algorithmic complexity non-deterministic polynomial-time hardness, which restricts them in dealing with complex problems [82]. Symbolic systems meet difficulty in modeling uncertainty and ambiguity. Moreover, problem resolution in big data spaces is a challenging task [51].

2.3.2 Subsymbolic Systems

Subsymbolic models are Artificial Intelligence systems and usually use representations based on vector space for representing data and information. Subsymbolic methods are generally robust to noise in the data and have been demonstrated to outperform human performance on tasks involving video, audio, and text. Subsymbolic approaches, particularly deep learning, emulate the process of neural connections in the human brain to build models [106, 145]. It suggests that deep learning models can model the implicit correlations within the data. However, these models cannot provide explicit inference evidence to explain the results. In addition, neural link prediction models are based on subsymbolic representations called embeddings. These models have been widely applied to the knowledge graphs completion task. Entity and relationship embeddings are learned by maximizing a scoring function over valid factual triples.

Knowledge Graph Embeddings

A method for building KG embeddings is a machine learning model that learns latent vector representations of entities $v \in V$ and relations $e \in E$ in a KG,

Table 2.1: **Scoring function and complexity of embedding models**. Adapted from [110]

Model	Scoring function	Complexity
HolE	$r(h \star t)$	$\mathcal{O}(\lvert E\rvert d + \lvert\mathcal{R}\rvert d)$
RESCAL	$h^T W_r t = \sum_{i=1}^{d} \sum_{j=1}^{d} w_{ij}^{(r)} h_i t_j$	$\mathcal{O}(\lvert E\rvert d + \lvert\mathcal{R}\rvert d^2)$
DistMult	$h^T W_r t = \sum_{i=1}^{d} h_i diag(W_r)_i t_i$	$\mathcal{O}(\lvert E\rvert d + \lvert\mathcal{R}\rvert d)$
RotatE	$-\lvert\lvert h \circ r - t\rvert\rvert$	$\mathcal{O}(\lvert E\rvert d + \lvert\mathcal{R}\rvert d)$
QuatE	$(h \otimes r)t$	$\mathcal{O}(\lvert E\rvert d + \lvert\mathcal{R}\rvert d)$
TransE	$\lVert h + r - t\rVert$	$\mathcal{O}(\lvert E\rvert d + \lvert\mathcal{R}\rvert d)$
TransH	$\lVert h_\perp + r - t_\perp\rVert$	$\mathcal{O}(\lvert E\rvert d + 2\lvert\mathcal{R}\rvert d)$
TransR	$\lVert h_r + r - t_r\rVert$	$\mathcal{O}(\lvert E\rvert d + \lvert\mathcal{R}\rvert d^2)$
TransD	$\lVert h_\perp + r - t_\perp\rVert$	$\mathcal{O}(2\lvert\mathcal{E}\rvert d + 2\lvert\mathcal{R}\rvert d)$
UM	$\lVert h - t\rVert$	$\mathcal{O}(\lvert E\rvert d)$
SE	$\lVert M_{r,1}h - M_{r,2}t\rVert$	$\mathcal{O}(\lvert E\rvert d + 2\lvert\mathcal{R}\rvert d^2)$
ERMLP	$w^T g(W[h; r; t])$	$\mathcal{O}(\lvert E\rvert d + \lvert\mathcal{R}\rvert d + k(3d + 2) + 1)$
ConvKB	$g([h; r; t] \circledast \omega)W$	$\mathcal{O}(\lvert E\rvert d + \lvert\mathcal{R}\rvert d + 4\tau$

preserving their semantic meaning [110]. In cases where KGs are incomplete, new facts have to be identified to add to the KGs. This task is known as Knowledge Graph Completion and can be done by inferring new facts from those already in the KG. This approach, called Link Prediction, exploits the KG to learn low-dimensional representations named Knowledge Graph Embeddings (KGE) and is used to infer new facts. The embeddings are numerical vectors that represent any element. *KGE* models define a scoring function ϕ to estimate the plausibility of any triple, $\langle h, r, t\rangle$, where $h, t \in V$ and $r \in E$ [109]. The model learns embeddings that optimize the score of known facts and considers unseen facts highly plausible, where higher ϕ values yield higher plausibility. Link predictions are performed by identifying which entities provide the best scores if added to the incomplete triples as heads or tails. If $prediction = tail$, the link prediction task finds t as the best scoring tail for the incomplete triple $\langle h, r, ?\,\rangle$:

$$\operatorname*{argmax}_{t \in V} \phi(h, r, t)$$

If $prediction = head$, it can be defined analogously. The data sparsity issue may negatively impact the state-of-the-art of KGE methods, i.e., true triples that can be used as positive samples to guide KGE training represent only a minor portion. We analyze thirteen embedding models from different families [110] to compute latent representations, e.g., vectors, of entities and relations in the KG and then employ them to infer new facts. In particular, we examine three main families of models:

- Tensor Decomposition models such as *HolE*, *RESCAL*, and *DistMult*.
- Geometric models such as *RotatE*, *QuatE*, and the Trans* family models *TransE*, *TransH*, *TransD*, and *TransR*.
- Deep Learning models such as *UM*, *SE*, *ERMLP*, and *ConvKB*.

Tensor Decomposition models

Tensor Decomposition models are mathematical models used to analyze and represent higher-order data structures known as tensors [69]. Tensors are generalizations of matrices capable of representing multi-dimensional data with more than two dimensions. This tensor can be decomposed into a combination of low-dimensional vectors, i.e., the embeddings of entities and relations. Embeddings are learned a posteriori from known facts and are able to generalize and associate high scores with unseen facts. Moreover, these models typically employ few or no shared parameters, which makes them especially lightweight and simple to train.

Hole: Holographic embeddings (*HolE*) [90] combines the expressive power of the tensor product with the efficiency and simplicity of *TransE*. *HolE* computes circular correlation, denotes by $\star$ in Table 2.1, between the embeddings of head (h) and tail (t) entities and multiply it by the relation (r) embedding.

RESCAL: *RESCAL* [91] is an algorithm of relational learning based on a tensor factorization where models entities as vectors and relations as matrices. In *RESCAL*, the relation matrices W_r contain weights $w_{i,j}$ between the i-th factor of h and j-th factor of t.

DistMult: *DistMult* [141] represents the relation embedding as a bidimensional matrix given a head and tail embeddings. *DistMult* is a simplification of *RESCAL*. The scoring function is the bilinear product where the relation embeddings are restricted to diagonal matrices. Therefore, the scoring function is commutative and considers all relations symmetric.

Geometric models

Geometric models interpret relations as geometric operations in latent space [134]. The head embedding undergoes a spatial transformation ξ as a function of the values of the relation embedding. The fact score is the distance between the resultant vector and the tail vector and is computed using a distance function δ, e.g., Manhattan Distance or Euclidean norm. $\phi(h, r, t) = \delta(\xi(h, r), t)$.

RotatE: *RotatE* [127] represents each relation as a rotation from the head entity to the tail entity in the complex latent space. *RotatE* maps the head and tail entities to the complex embeddings, i.e., $h, t \in \mathbb{C}^k$. The rotation r is applied to h by operating a Hadamard product (denoted by $\circ$ in Table 2.1).

QuatE: *QuatE* [148] operates on the quaternion space and learns hypercomplex valued embeddings (quaternion embeddings) to represent entities and relations. The relation is used to rotate the head entity h, where $\otimes$ represents the Hamilton product, see Table 2.1. Hypercomplex representations extend complex representations by representing each number with one real and three imaginary components.

TransE: *TransE* [19] proposes a geometric interpretation of the latent space and interprets relation vectors as translations in vector space, $h + r \approx t$. *TransE* can not naturally model 1-n, n-1 and n-m relationships. Suppose a relation r with cardinality 1-n, $(h, r, t_1), (h, r, t_2)$ then the model fits the embeddings in order to ensure $h + r \approx t_1$ and $h + r \approx t_2$, i.e. $t_1 \approx t_2$.

TransH: *TransH* [138] is an extension of *TransE* that aims to overcome the limitations of *TransE*. Furthermore, in *TransH*, each relation is represented by a normal vector of a hyperplane, where the variables $h_\perp$ and $t_\perp$ denote a projection to the hyperplane w_r of the labeled relation r, where r is the vector of a relation-specific translation in the hyperplane w_r.

TransR: *TransR* [78] represents entities and relations in distinct vector spaces and learns embeddings by translation between projected entities. $h_r = h * M_r$, where M_r corresponds to a projection matrix $M_r \in \mathbb{R}^{dxk}$ that projects entities from the entity space to the relation space; further $r \in \mathbb{R}^k$.

TransD: *TransD* [60] employs separate projection vectors for each entity and relation. In score function of *TransD* the variables $h_\perp$ and $t_\perp$ are defined as, $h_\perp = M_{rh}h$ and $t_\perp = M_{rt}t$, where M_{rh}, $M_{rt} \in \mathbb{R}^{m \times n}$ are two mapping matrices defined as follows: $M_{rh} = r_p h_p + I^{m \times n}$ and $M_{rt} = r_p t_p + I^{m \times n}$. The subscript p means the projection vectors, and $I^{m \times n}$ denotes the identity matrix of size $m \times n$.

UM: The Unstructured Model (*UM*) [17] is a simplified version of *TransE* where it does not consider differences in relations and only models entities as embeddings. This model can be beneficial in KGs containing only a single relationship type.

SE: Structured Embedding (*SE*) [18] model defines two matrices $M_{r,1}$ and $M_{r,2}$ to project head and tail entities for each relation. *SE* can discern between the subject and object roles of an entity since it employs different projections for the embeddings of the head and tail entities.

Deep Learning models

Deep learning models use deep neural networks to build the KGE model. Neural networks learn parameters such as weights and biases and combine them with the input data to recognize meaningful patterns. Deep networks organize the parameters in distinct layers, usually interspersed with nonlinear activation functions. Several types of layers have been developed, such as dense layers that only combine the input data X with the weights W and add a bias b: $W \times X + b$.

ERMLP: *ERMLP* [32] is a model based on a multi-layer perceptron and uses a single hidden layer. In the score function, the variable $W \in \mathbb{R}^{k \times 3d}$ represents the weight matrix of the hidden layer, the variable $w \in \mathbb{R}^k$ represents the weights of the output layer, and g is the activation function. In Table 2.1, the variable k corresponds to the number of neurons in the hidden layer.

ConvKB: *ConvKB* [89] employing a convolutional neural network. Each embedding triple (h, r, t) is represented as a matrix $A = [h; r; t] \in \mathbb{R}^{d \times 3}$, where each column represents the embedding for h, r, and t. In the convolution layer, a set of convolutional filters $\omega_i \in \mathbb{R}^{1 \times 3}, i = 1, ..., \tau$, where τ corresponds to the number of convolutional filters in a layer, are applied on the input in order to examine the global relationships between same dimensional entries of the embedding triple and to generalize the transitional characteristics. Each ω_i is repeatedly operated over every row of A to finally generate a feature map $v = [v_{i,1}, ..., v_{i,d}] \in \mathbb{R}^d$ as $v_i = g(\omega A + b)$, where $b \in \mathbb{R}$ is a bias term and g is an activation function. In the score function, the variable $w \in \mathbb{R}^{\tau d \times 1}$ is a shared weight vector.

2.4 Summary

The research problem of discovering actionable insights from heterogeneous data sources and the particular challenges stated in Chapter 1 requires comprehensive solutions from different angles. This chapter's existing concepts and technologies provide a basis for addressing the raised challenges. The neuro-symbolic AI systems introduced in Section 2.3 present the basic principles for solving the knowledge discovery task. Specifically, we rely on deductive databases for the symbolic system and KGE models for the sub-symbolic system to answer the research questions **RQ1** and **RQ4**. The Semantic Web described in Section 2.1 provides a framework for representing and accessing data. Section 2.1 defines the basis for addressing the task of integrating data from heterogeneous data sources into a homogeneous model, contributing to answering the research questions **RQ2**. In Section 2.2, the deductive systems are introduced. We present the notation and properties of deductive systems that allow deriving deductions. We present a symbolic system that relies on Datalog to enhance the semantics encoded in the data for addressing research question **RQ3**.

Chapter 3

Related Work

This chapter presents a detailed analysis of the state-of-the-art approaches related to the main research problems and research questions defined in Chapter 1. First, we show the topics identified for the review of existing approaches in Figure 3.1. We present for each topic an overview of approaches and limitations within the scope of the challenges defined by this thesis. Section 3.1 introduces state-of-the-art approaches proposed by research communities for neuro-symbolic AI systems. We point out the shortcomings of the state-of-the-art techniques on the problem of integrating neuro-symbolic systems. Next, in Section 3.2 we review the existing approaches in the area of sub-symbolic systems focusing on different problems in the healthcare context. Then, Section 3.3 shows approaches of symbolic systems in which human and machine experts are able to interpret information without ambiguity. Finally, Section 3.4 presents the most recent approaches to resolving semantic interoperability issues in the knowledge discovery field. We carried out a review and critical discussion of the current approaches for discovering communities of I4.0 standards.

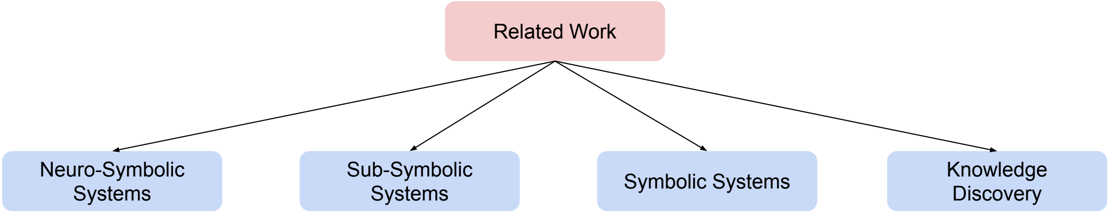


Figure 3.1: **Related Work Topics**. Figure 3.1 illustrates the work related to this thesis in four areas: neuro-symbolic systems, sub-symbolic systems, symbolic systems, and knowledge discovery.

3.1 Generic Neuro-Symbolic Artificial Intelligence Approaches

Generic Neuro-Symbolic AI systems contribute features of symbolic and sub-symbolic systems regardless of domains. A neuro-symbolic AI system provides a neural-symbolic implementation of logic, a logical characterization of a neural system, or a hybrid learning system [39, 40, 128]. Real applications are possible in areas with social relevance and high economic impacts, such as bioinformatics, robotics, fraud prevention, and the Semantic Web [14]. Methods utilized in neural-symbolic integration in some of the aforementioned applications include translation algorithms between logic and networks. Also, the community has focused on studying the systems empirically through case studies and real-world applications. One of the potential applications of neuro-symbolic AI is the diagnosis and prediction of ophthalmic diseases [51].

Neuro-Symbolic AI is used to assist the representation of deep learning in processing visual question answering (NS-VQA) [75, 80, 81, 143], contributing more transparency to the reasoning process. NS-VQA is composed of three parts to separate reasoning from visual perception and textual comprehension. The core part relies on the sub-symbolic system to reason both visual and textual representations obtained by deep learning and question answering in clinical eye scenarios. Thereby, neuro-symbolic has promising application prospects in medical images, specifically in classifying ocular diseases. Furthermore, Zhao et al. [150] analyze the state-of-the-art automatic disease diagnosis from clinical data, concluding that the problem of feature sparsity and missing values are affecting disease diagnosis. They propose a knowledge-guided graph attention network for disease prediction and effectively generate embeddings to accurately predict both general and rare diseases. Following the same line, as Zhao et al. [150], Sun et al. [126] propose an innovative graph neural network (GNN)-based model for disease prediction, using external knowledge bases to augment the insufficient data from the clinical data. They evaluate the approach to predicting chronic obstructive pulmonary disease, illustrating the effectiveness of the proposed model.

In the field of vision-based tasks, such as semantic image labeling, high-performance systems have been produced. Karpathy et al. [65] propose an approach for the recognition and labeling tasks for the content of different regions of the images; it combines Convolutional Neural Networks over the image regions together with bidirectional Recurrent Neural Networks over sentences. Once this mapping of images and sentences in the embedding space has been established, a structured objective is introduced that aligns the two modalities through multimodal embedding. The emerging system performs better than classical approaches, where tasks involving semantic descriptions are associated with databases that contain background knowledge, and computer image processing approaches were based on rule-based techniques.

Despite the progress of Neuro-Symbolic Artificial Intelligence, the scope and applicability of symbol processing are limited. Furthermore, these systems do not examine polynomial overload when integrating both paradigms. Our work leverages the symbolic system, independent of the application domain, and improves the predictive capability of KGE models. Moreover, our approach addresses the deductive database to an abstract target prediction, rendering the computational complexity polynomial time. Thus, we show the positive impact of completing KG via a deductive system on the overall performance of a predictive model implemented using KGEs.

3.2 Sub-symbolic Systems in Healthcare Contexts

Knowledge graphs are becoming increasingly important in the biomedical field. Discovering new and reliable facts from existing knowledge using KGE is a cutting-edge method. KG allows a variety of additional information to be added to aid reasoning and obtain better predictions. Zhu et al. [151] develop a process for constructing and reasoning multimodal Specific Disease Knowledge Graphs (SDKG). SDKG is based on five cancers and six non-cancer diseases. The principal purpose is to discover reliable knowledge and provide a pre-trained universal model in that specific disease field. The model is built in three parts: structure embedding (S) with TransE, TransD, and ConvKB, category embedding (C), and description embedding (D) with BioBERT to convert description annotations into vectors. The best results are obtained when description embedding is combined with structure embedding, specifically with the ConvKB embedding model. Karim et al. [64] propose a new machine-learning approach for predicting DDIs based on multiple data sources. They integrated drug-related information such as diseases, pathways, proteins, enzymes, and chemical structures from different sources into a KG. Then different embedding techniques are used to create a dense vector representation for each entity in the KG. These representations are introduced in traditional machine learning classifiers and a neural network architecture based on a convolutional LSTM (Conv-LSTM) modified to predict DDIs. The results show that the combination of KGE and Conv-LSTM performs state-of-the-art results.

Meng et.al. [135] propose a framework, Predicting Rich DDI (PRD), to predict multilabel of DDIs. The framework PRD relies on knowledge graph embedding techniques and predicts DDI as a linked prediction task. PRD uses a drug knowledge graph generated from different sources and biomedical texts with descriptions of the DDIs in the predictive task. They provide a joint translation-based embedding model to learn DDIs by integrating drug knowledge graphs and biomedical texts into the same semantic space. PRD framework aims to represent the triples from drug KG and the rich DDI triples from the biomedical text in a unified joint embedding model. Thus, the DDI prediction problem is addressed as a link prediction task. In the experimental study, the authors compare three DDIs methods, including multitasking dyadic drug-drug interaction prediction (MDDP) [62], considered one of the best baselines for multiple DDI type predictions [135], and two KG embeddings techniques. PRD achieves improvement over all baselines.

The above-mentioned research aims to discover reliable knowledge based on knowledge graphs using KGE models. However, they are limited by the data sparsity issue of the KGE models and the lack of symbolic reasoning. We overcome this limitation by integrating a Neuro-Symbolic AI system enabling reasoning and robust learning to improve the predictive capability of KGE models.

3.2.1 Polypharmacy Side Effect Prediction and Drug-Drug Interactions Prediction

A framework to predict DDIs is presented in [36]; they exploit information from multiple linked data sources to create various drug similarity measures. Then, they build a large-scale and distributed linear regression learning model to predict DDIs. They evaluate their model to predict the existence of drug interactions, considering the DDIs as symmetric. A neural network-based method for drug-drug interaction prediction is proposed in [108]. They use various drug data sources in order to compute multiple drug similarities. They computed drug similarity based on drug substructure, target, side effect, off-label side effect, pathway, transporter, and indication data. The proposed method first performs similarity selection and then integrates the selected similarities with a nonlinear similarity fusion method to obtain high-level features. Thus, they represent each drug by a feature vector and are used as input to the neural network to predict DDIs.

Other approaches focus on predicting DDIs and their effects [72, 83, 113, 152]. Beyond knowing that a pair of drugs interact, it is essential to know the effect of DDI in polypharmacy treatments. In [72], propose a novel deep learning model to predict DDIs and their effects. They use additional features based on structural similarity profiles (SSP), Gene Ontology term similarity profiles (GSP), and target gene similarity profiles (TSP) to increase the classification accuracy. The proposed model uses an autoencoder to reduce the dimension of the resulting vector from the combination of SSP, TSP, and GSP. The benchmark used has 1597 drugs and 188'258 DDIs with 106 different types. The model works as a multi-label classification model where the deep feed-forward network has an output layer of size 106, representing the number of DDI types. The results show that the model obtains equal or better results in 101 out of 106 DDI types than baseline methods. Also, they demonstrate how adding the features GSP and TSP increases the accuracy of DDIs prediction. Marinka Zitnik et al. [152] present Decagon, an approach for predicting the side effects of drug pairs. The approach develops a new convolutional graph neural network for link prediction. They construct a multi-modal graph of protein-protein interactions, drug-protein target interactions, and the DDI side effects. The graph encoder model produces embeddings for each node in the graph. They proposed a new model that assigns separate processing channels for each relation type and returns an embedding for each node in the graph. Then, the Decagon decoder for polypharmacy side effects relation types takes pairs of embeddings and produces a score. Thus, Decagon can predict the side effect of a pair of drugs.

All the approaches mentioned above are limited to predicting DDIs and their effects between pairs of drugs. However, in our view, the interactions and their effects need to be considered as a whole and not in pairs in polypharmacy treatments. Our symbolic system resorts to a set of rules that state the implicit definition of new DDIs generated as a result of the combination of multiple drugs in treatment. Since cancer treatment schemes are usually composed of more than two drugs, and patients may have several co-existing diseases requiring additional medications, it is of significant relevance in holistically deducing DDIs.

3.2.2 Treatment Response Prediction

Deep learning is extensively used in medical applications, focusing on disease detection and diagnosis [5, 25]. However, there are limited studies on predicting treatment responses. Recent studies have employed deep learning to predict treatment responses by analyzing images [61, 71]. Cheng Jin [24] presents a multitask deep learning approach that allows for simultaneous tumor segmentation and response prediction. The model is trained with magnetic resonance images of rectal cancer patients to predict pathologic response after neoadjuvant chemoradiotherapy. In addition, Watts [30] presents a review paper about machine learning techniques for predicting treatment response using Electroencephalography (EEG) in major depressive disorder (MDD). Watts points to the promising use of EEG within machine learning models to predict treatment responses in MDD. All the above approaches aim to predict treatment response through image analysis. However, we propose to predict treatment response based on the drugs and their description, i.e., DDIs with their effects, DPIs, and the genes encoded by the proteins. Then, our neuro-symbolic system that enables expressive reasoning and robust learning enhances the predictive capacity of KGE models. Thus, our approach supports clinicians in having a treatment response at the time of treatment.

3.3 Symbolic Systems

The ability to formally capture semantics, such as that of symbolic systems, enables human and machine experts to interpret information unambiguously. Furthermore, symbolic systems allow further data enrichment using symbolic inference mechanisms, such as Descriptive Logic inference [8] and rule-based reasoning [16, 23]. Stavropoulos et al. [125] present a rule-based approach for detecting health-related problems from wearable sensor lifestyle data that aggregate clinical value to make informed decisions on monitoring and intervention. To achieve interoperability at different levels, they use OWL 2 ontologies [46] as the underlying knowledge representation formalism, generating interoperable Knowledge Graphs (KGs). The KG is further enriched with a set of preconfigured rules to derive logical consequences and semantically enrich KGs. They use rules to provide expert knowledge in the form of constraints and SHACL rules [48] to recognize patterns, anomalies, and situations of interest based on predefined rules and conditions. Thus, the system is able to alert of patterns and anomalies in patients, and clinicians can make quick decisions regarding interventions and follow-up.

In recent years, there has been a growing interest in Pharmacovigilance. Extensive research has been conducted to predict potential DDI. One approach to predicting potential DDI is based on similarity [36, 124, 132, 146], with the core idea of predicting the existence of a DDI by comparing candidate drug pairs with known interacting drug pairs. These approaches define a wide variety of drug similarity measures for comparison. The known DDIs that are very similar to a candidate pair provide evidence for the presence of a DDI between the candidate pair drugs. Sridhar et al. [124] propose a probabilistic approach for inferring unknown DDIs from a network of multiple drug-based similarities and known DDIs. They used the probabilistic programming framework Probabilistic Soft Logic. This symbolic approach predicts three types of interacctions [124], CYP-related interactions (CRDs), where both drugs are metabolized by the same CYP enzyme, NCRDs, where no CYP is shared between the drugs and general DDI from Drugbank. Furthermore, they consider seven drug-drug similarities. Thus, they found five novels DDIs validated by external sources.

Albeit representing domain-specific knowledge, the approaches mentioned above cannot solve the problem of predicting unknown relationships in KG and are limited to the defined rules. We provide a neuro-symbolic system that integrates a symbolic and sub-symbolic system considering semantics in the KG and is able to predict unknown relationships.

3.4 Knowledge Discovery in Industry 4.0 Domains

In recent years, a great deal of research has been investigating key aspects of discovering standards communities. Furthermore, many approaches are proposed to corroborate and extend the knowledge of the standardization frameworks and resolve semantic interoperability issues.

3.4.1 Unsupervised Learning

Unsupervised learning techniques do not rely on the class attribute for model building; instead, they extract knowledge from discovering interrelationships between data elements. The main types of unsupervised learning tasks are clustering and association. These algorithms discover hidden patterns or clusters of data without the need for human intervention. Its ability to discover similarities and differences in information makes it the ideal solution for exploratory data analysis. SemEP [94] is an unsupervised semantics-based edge partitioning method. SemEP combines a data mining framework for link prediction, semantic similarities, and an algorithmic approach to partition the edges of a graph. Thus, the semEP problem is to create a minimal partitioning of the edges such that the cluster density of each subset of edges is maximal. An advantage of semEP edge clustering is that it allows a node to participate in more than one cluster.
METIS [66] is a model for partitioning large irregular graphs and computing fill-reducing orderings of sparse matrices. The algorithm implemented in METIS is based on the multilevel graph partitioning paradigm [67], which quickly produces high-quality partitioning. METIS can partition an unstructured graph into a user-specified number k of parts.
The KMeans algorithm [7] clusters the data by separating the samples into equal variance groups. This algorithm requires the number of clusters to be specified. KMeans has three steps; the first one chooses the initial centroids. The second step is a loop between the following two steps. First, it assigns each sample to its nearest centroid. The second step creates new centroids by taking the average value of all samples assigned to each previous centroid. Then, the difference between the old and new centroids is calculated, and the algorithm repeats this loop until the difference is less than a threshold.

3.4.2 Solving Interoperability in I4.0

Zeid *et al.* [144] study different approaches to achieve interoperability of different standardization frameworks. In this work, the current landscape for smart manufacturing is described by highlighting the existing standardization frameworks in different regions of the globe. Lin *et al.* [76] present similarities and differences between the RAMI4.0 model and the IIRA architecture. Based on the study of these similarities and differences, the authors proposed a functional alignment among layers in RAMI4.0 with the functional domains and crosscutting functions in IIRA. Monteiro *et al.* [85] and Velazquez *et al.* [130] further report on the comparison of the RAMI4.0 and IIRA frameworks. This work presents a cooperation model to align both standardization frameworks. Furthermore, mappings between RAMI4.0 IT Layers and the IIRA functional domain are established. Moreover, the IIRA and RAMI4.0 frameworks are compared based on different features, e.g., country of origin, source organization, basic characteristics, application scope, and structure. It further details where correspondences exist between the IIRA viewpoints and RAMI4.0 layers. In [29], Darmois *et al.* present the main contributions to the analysis of IoT standardization. This work has defined knowledge areas used for the classification of standards and identifies the standardization gaps. The purpose is to support interoperability in complex IoT systems and provide guidelines contributing to semantic interoperability approaches. Aligning standardization frameworks is useful for solving interoperability problems, but not all standards are classified by layers in the standardization frameworks. However, these approaches aim to solve interoperability problems by mapping the different frameworks without creating a common vocabulary that semantically represents the standards. This thesis proposes an approach to solve interoperability problems among I4.0 standards by discovering unknown relationships.

3.4.3 Ontology-based Approaches in I4.0

Ontology-based approaches have contributed to creating a shared understanding of the I4.0 domain. Lelli *et al.* [74] propose the reuse of existing ontologies as one of the main principles in ontology design. For this purpose, they make use of Linked Open Vocabulary (LOV) and collect 22 ontologies related to IoT. They state that project developers in the IoT community do not reuse existing works, damaging the attempt to define a shared understanding of smart interoperability. Kovalenko and Euzenat [70] have equipped data integration with diverse methods for ontology alignment. They examine the problems of ontological correspondence in the context of engineering knowledge integration. Kovalenko and Euzenat present technologies for defining mappings between ontologies to support data integration. Finally, they illustrate how mappings can be generated from definitions in the Expressive and Declarative Ontology Alignment Language (EDOAL). These approaches are limited to representing the existing characteristics of the knowledge domain in ontologies, which is useful because it enables data integration in Industry 4.0. However, there are standards that are not classified in any standardization framework, and this limits the solution of the interoperability problem. In this work, we employ the Standard Ontology (STO) for representing the main properties of standards and standardization frameworks, as well as relationships among them [45].

3.4.4 Knowledge Graphs and Semantic Data Integration

Sebastian *et al.* [9] propose a semantically annotated knowledge graph for Industry 4.0-related standards, norms, and frameworks. The I4.0 knowledge graph helps overcome Industry 4.0 challenges requiring comprehensive knowledge of the different standards. Furthermore, the I4.0 knowledge graph considers the semantics and relations between standards and the standardization framework. Garofalo *et al.* [41] outline Knowledge Graph Embeddings for I4.0 use cases. Existing techniques for generating embeddings on top of KG are examined. Further, the analysis of how these techniques can be applied to the I4.0 domain is described; specifically, it identifies predictive maintenance, quality control, and context-aware robots as the most promising areas to apply the combination of KGs with embeddings. These approaches mentioned above support data-driven pipelines to transform industrial data into actionable knowledge in smart factories. Galinski [38] examines the problem of semantic data integration and interoperability between standards. This work emphasizes the need for metadata, data models, and metamodels for standards. It also presents an interesting description of which data to consider when describing a standard. Hodges *et al.* [57] propose an approach for semantic integration of standards to achieve interoperability between them by means of ontologies; relevant standards and well-known ontologies to represent standards are also identified. Albeit representing domain-specific knowledge, the approaches mentioned above cannot discover alignments across I4.0 standards. We overcome this limitation by exploiting embeddings over a knowledge graph of I4.0 standards to predict relatedness among standards.

The approaches presented in this section describe and characterize existing knowledge in the I4.0 domain. However, in our view, two directions need to be considered to enhance the knowledge in the domain; 1) the use of a KG-based approach to encoding the semantics, and 2) the use of machine learning techniques to discover and predict new communities of standards based on their relations. Our goal is to uncover alignments among related standards. Nevertheless, finding alignments across I4.0 requires the encoding of domain-specific knowledge represented in standards of diverse nature.

3.5 Summary

Based on the above-mentioned analysis of the existing approaches, this thesis focuses on integrating neuro-symbolic AI systems to uncover reliable knowledge from heterogeneous data. These systems should be able to discover new relationships and patterns hidden in the data. The representation of the heterogeneous data and their meaning needs to be preserved to improve the performance of data analytics and predictive models. More importantly, a suitable representation of the data and appropriate semantic enrichment of data are required to extract more powerful and relevant insights.

Chapter 4

A Neuro-Symbolic Artificial Intelligence System over Knowledge Graphs

Neuro-Symbolic AI is a highly active area that has been studied for decades [39] and endeavors to combine symbolic and sub-symbolic AI models. AI aims to simulate human behavior, which is often driven by cognition and mental processing. The symbolic systems are the most prominent tools for modeling behavior, and the sub-symbolic systems are for modeling cognition and the brain. Complex problem-solving using AI requires a significantly enriched language. Symbolic and sub-symbolic systems differ fundamentally in how they represent data and information. Symbolic systems typically use structured representation languages from formal logic, and sub-symbolic systems usually use representations based on vector space. Thus, neuro-symbolic integration aims to bridge the gap between symbolic and sub-symbolic systems.

Integrating neuro-symbolic into real-world applications is a challenging task. Even in controlled environments, e.g., training simulators, neuro-symbolic integration may not be completed successfully [52]. For instance, Fernlund et al. [35] describe systems that use machine learning to learn relations from expert observations. While these systems are successful in learning, they lack the expressive power of symbolic systems. Another example of neuro-symbolic systems in bioinformatics is Connectionist Inductive Learning and Logic Programming (CILP) [40], which combines connectionist learning (neural networks) with logic programming. CILP models may struggle with limited data availability. Connectionist learning relies heavily on having sufficient labeled examples for training, and logic programming may require substantial domain-specific knowledge. CILP models can face scalability challenges, especially when dealing with large-scale datasets. The training and inference processes of neural networks can be computationally expensive, particularly when combined with logical reasoning. Furthermore, Karpathy et al. [65] combine convolutional neural networks (CNNs) with bidirectional recurrent neural networks (RNNs) over sentences to recognize and label image regions. Combining CNNs with bidirectional RNNs has shown promising results in various tasks, including image captioning and visual question answering. Despite advances in neuro-symbolic AI integration, symbol processing currently has limited scope and applicability. In this chapter, we propose an approach that integrates a domain-agnostic symbolic system with a Knowledge Graph Embedding (KGE) model to reduce the KG sparsity towards improving the model's predictive capability. Thus, we broaden the scope and applicability in several domains of neuro-symbolic integration. Figure 4.1 shows the main challenges tackled in this chapter and the contribution to addressing the challenges. The content of this chapter is based on the publication [104]. The results of this chapter provide an answer to the following research questions:

RQ1: How can metadata encoding data meaning be exploited to discover relationships in knowledge graphs?

RQ2: How can heterogeneous data sources be integrated to obtain a unified knowledge representation?

RQ3: How can implicit knowledge be used to enhance knowledge discovery tasks?

RQ4: What is the impact of deductive reasoning on accurately uncovering knowledge?

To address research questions **RQ1** and **RQ3**, we present an approach based on the integration of Neuro-Symbolic AI systems. The symbolic system is implemented by deductive databases, enhancing the predictive capacity of sub-symbolic systems implemented by KGE models. The deductive databases are defined for an abstract target prediction over a knowledge graph. Our proposed solution builds the ego networks of the entities that correspond to the head and tail of the abstract target prediction to deduce new relationships and enhance the ego networks. The ego network consists of three main components. The ego represents the node of interest from which connections are examined and analyzed. The neighboring nodes are the entities directly connected to the ego and the relationships between their neighbors. The sub-symbolic systems benefit from the enhanced ego networks and perform better predictions. To address research questions **RQ2** and **RQ4**, we

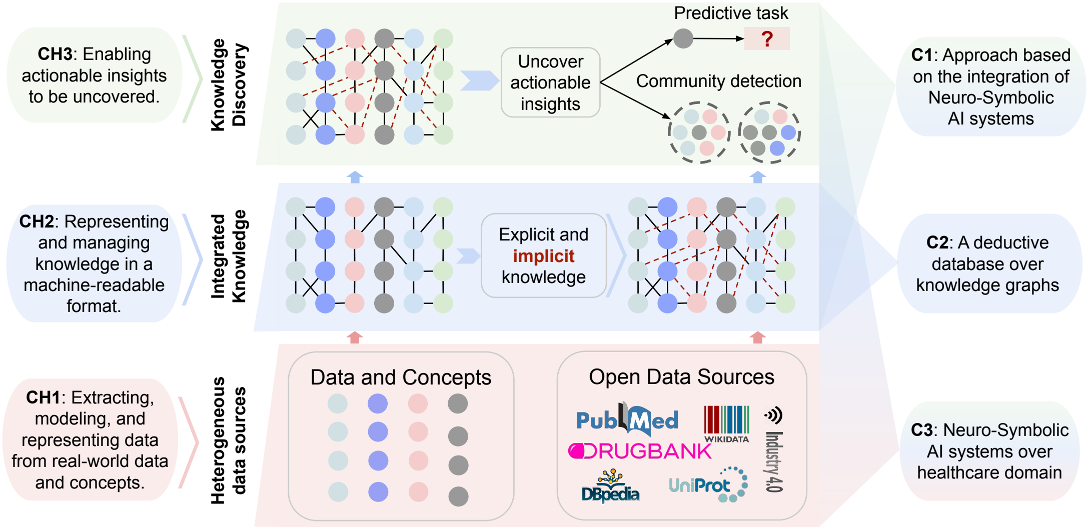

CH3: Enabling actionable insights to be uncovered.
Knowledge Discovery
Predictive task
?
Uncover actionable insights
Community detection
C1: Approach based on the integration of Neuro-Symbolic AI systems
CH2: Representing and managing knowledge in a machine-readable format.
Integrated Knowledge
Explicit and implicit knowledge
C2: A deductive database over knowledge graphs
CH1: Extracting, modeling, and representing data from real-world data and concepts.
Heterogeneous data sources
Data and Concepts
Open Data Sources
PubMed
WIKIDATA
DRUGBANK
Industry 4.0
DBpedia
UniProt
C3: Neuro-Symbolic AI systems over healthcare domain

Figure 4.1: **Challenges and contributions**. This chapter focuses on discovering relationships in a knowledge graph and proposes an approach based on the integration of Neuro-Symbolic AI systems to solve the problem.

build a knowledge graph to integrate heterogeneous data sources, enabling the description of the meaning of data. We empirically evaluate the effectiveness of our Neuro-Symbolic AI approach. We conduct an ablation study on the components of our approach, considering state-of-the-art KG embedding methods. The deductive system reduces the data sparsity issues in the KG, enabling the knowledge graph to become meaningful in the discovery task and enhancing the KGE algorithms in the link prediction task. We summarize the contributions of this chapter as follows:

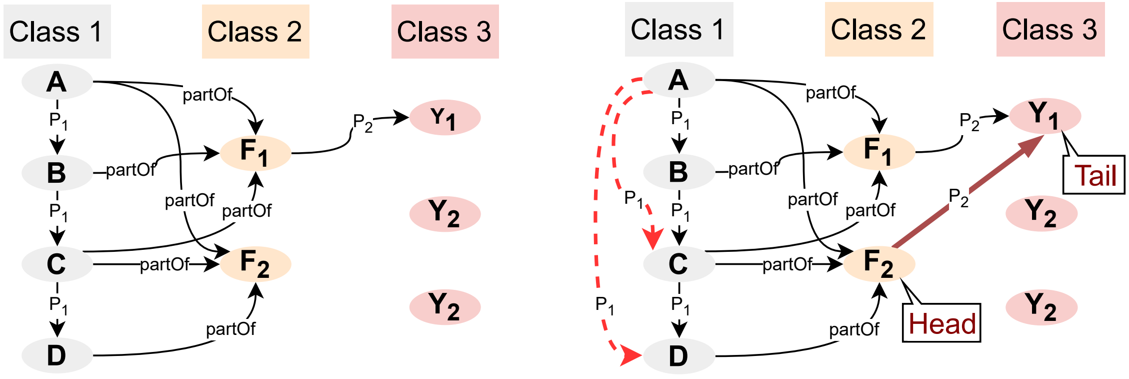
- A domain-agnostic approach able to empower the predictive capacity of sub-symbolic systems with a deductive database system. The deductive system minimizes the data sparsity issues by deducing the implicit relationships in the KG.
- A formalization of the symbolic systems for an abstract target prediction over a knowledge graph. The symbolic system enhances the ego networks of the target prediction with implicit relations.
- An extensive evaluation of symbolic-sub-symbolic systems in the state-of-the-art KGE algorithms. We demonstrate the benefit of integrating symbolic and sub-symbolic systems.

The remainder of this chapter is structured as follows. Section 4.1 presents a motivating example showing the data sparsity issue in a knowledge graph for the link prediction task. We illustrate the need to minimize the data sparsity issue to obtain accurate predictions. Next, Section 4.2 presents the Neuro-Symbolic AI approach, addressing the research question **RQ1**. The approach assumes that a link prediction problem is defined in terms of an abstract target prediction over a KG. We integrate a deductive database system with the sub-symbolic system, where the deductive system alleviates the data sparsity issues in KG and allows the sub-symbolic system to predict links accurately. To address the research question **RQ3**, we present a deductive database system for an abstract target prediction over a KG based on the existing approaches of deductive database systems. The deductive system is implemented in Datalog and can derive deductions, e.g., conclude new facts. Furthermore, the proposed deductive system explicitly states the experts' knowledge by rules and transfers the implicit knowledge to the knowledge graph. In Section 4.3, we address the problem of predicting polypharmacy treatment effectiveness by applying our Neuro-Symbolic AI approach. We integrate heterogeneous data sources and build a polypharmacy treatment knowledge graph solving the **RQ2**. Then, the problem of predicting treatment effectiveness is modeled as a problem of link prediction. Next, Section 4.4 presents an empirical evaluation assessing the impact of the Neuro-Symbolic AI system proposed. The empirical results put the deduction power of deductive databases into perspective; they improve the predictive capacity of existing KGE models, answering the research question **RQ4**. Finally, Section4.5 presents the closing remarks of this chapter.

4.1 Motivating Example

We motivate the problem addressed in this chapter through a knowledge graph with sparsity issues. Currently, we face severe data sparsity issues in KGs because of privacy concerns or limited information available in the Open World Assumption. Data sparsity negatively impacts the tasks of knowledge discovery on top of KGs. We aim to show that minimizing the data sparsity in the KG can lead to better prediction in the knowledge discovery task.

Figure 4.2a illustrates a knowledge graph with entities belonging to three classes. The entities of *Class 1* are related by property P_1, e.g., $\langle A, P_1, B\rangle$. In addition, the entities of *Class 2* can contain entities of *Class 1* by the property *partOf*, e.g., $\langle A, partOf, F_1\rangle$. Moreover, entities of *Class 2* can be related to entities of *Class 3* by the property P_2, e.g., $\langle F_1, P_2, Y_1\rangle$. Figure 4.2b presents an ideal knowledge graph where all the implicit relations are explicitly represented, whereas the dotted arrows represent the implicit relations. A symbolic system, e.g., a Deductive Database, transfers the implicit deduced knowledge to the KG explicitly.



(a) **Knowledge Graph**

(b) **Ideal Knowledge Graph**

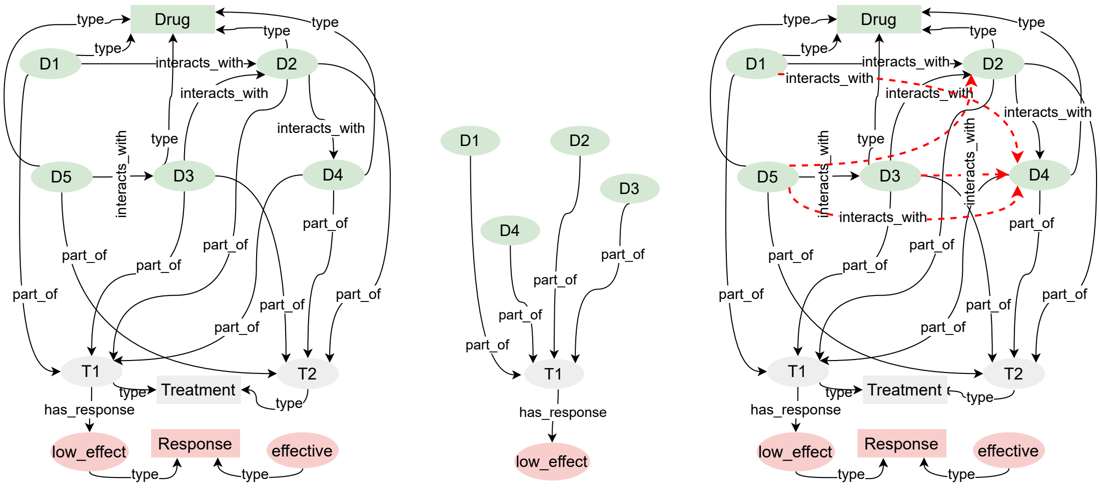
Figure 4.2: **Motivating Example.** Figure 4.2a illustrates a portion of a KG with data sparsity issues. Figure 4.2b shows the ideal KG with implicit knowledge, and an AI model predicts relationships between entities of *Class 2* and *Class 3*. The model learns latent vector representations of entities and relations in the ideal KG, and the link $\langle F_2, P_2, Y_1 \rangle$ is predicted.

The symbolic system is expressed in a deductive database system implemented in Datalog. The rules that define how new relations are generated can be stated, e.g., in a Datalog program, and represent the experts' knowledge explicitly transferred to the knowledge graph. Thus, the data sparsity issue is minimized by considering the symbolic system. The deduced relations $\langle A, P1, C \rangle$, and $\langle A, P1, D \rangle$ increase the descriptions of the entities of *Class 2* and make F_1 and F_2 share more relationships. Then, a sub-symbolic system, e.g., a KGE model, can better learn latent vector representations of entities, minimizing the data sparsity issue. The KGE model can predict the missing link $\langle F_2, P_2, Y_1 \rangle$, considering that the model represents the entity F_1 and F_2 nearby in the embedding space, and predict the triple. We aim to integrate a neuro-symbolic system where the predictive capability of the sub-symbolic systems is empowered by the symbolic system, reducing the data sparsity issue.

4.2 Neuro-Symbolic Artificial Intelligence Approach

4.2.1 Preliminaries

A knowledge graph is a data structure representing factual statements with entities and their relationships using a data graph [49]. KGs are used in countless



(a) Knowledge Graph $\mathcal{KG}$

(b) $ego(T1)$: ego network of $T1$

(c) Knowledge Graph $\mathcal{KG}$ applying DS

Figure 4.3: **Example Knowledge Graph.** Figure 4.3a shows a KG with three classes, five green entities belonging to class *Drug*, two gray entities belonging to class $Treatment$, and two red entities belonging to class $Response$. Figure 4.3b illustrates the ego network for the entity $T1$, where the entities $D1, D2, D3, D4$, and *low_ effect* are the neighbors of $T1$. Figure 4.3c shows the KG resulting from DS. The red arrows represent the new deduced links in the ego network $ego(.)$.

domains because of their ability to model data in a machine-readable form [58]. Let $\mathcal{KG} = (V, E, L, C, I, D, R, \mathcal{N}, ego, \alpha)$ be a KG, where:

- V is a set of nodes that correspond to concepts (e.g., classes and entities).
- $E \subseteq V \times L \times V$ is a set of edges representing relationships, i.e., triples (s, p, o), between concepts.
- L is a set of properties.
- C is a set of classes $C \subseteq V$.
- $I : V \rightarrow C$ is a function that maps each entity in V to a class C.
- $D : L \rightarrow C$ maps a property to the class that corresponds to the domain of the property.
- $R : L \rightarrow C$ maps each property to a class that corresponds to the range of the property.
- $\mathcal{N} : V \rightarrow 2^V$, where 2^V represents the power set of nodes V. $\mathcal{N}(v)$ defines the neighbors of the entity v, i.e., $\mathcal{N}(v) = \{v_i | (v, r, v_i) \in E \vee (v_i, r, v) \in E\}$.

- $ego : V \rightarrow 2^{V \times L \times V}$, the function $ego(.)$ represents ego networks in the knowledge graph. $ego(v)$ assigns to each concept in V the set of labeled edges, where v is in the subject or object position. $ego(v) = \{(u_1, r, u_2)|(u_1, r, u_2) \in E \wedge (u_1 = v \vee u_2 = v)\}$. The $ego(v)$ defines the ego network of the entity v.
- $\alpha : 2^V \rightarrow 2^{V \times L \times V}$. The function $\alpha(.)$ returns a set of triples between the pairs of elements in the input.If F is a set of entities in V, $\alpha(F) = \{(v_1, r, v_2)|(v_1, r, v_2) \in E \wedge v_1 \in F \wedge v_2 \in F\}$. The function $\alpha(.)$ returns the edges between pairs of entities in the input set F.

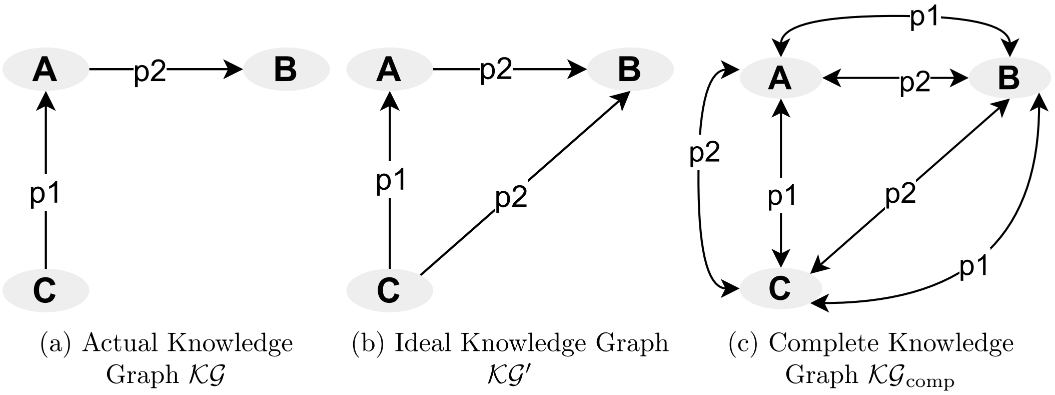
Figure 4.3a depicts a $\mathcal{KG}$, where the set of classes are represented by $C = \{Drug, Treatment, Response\}$. The class for each entity is represented by the function $I(v)$, e.g., $I(T1) = Treatment$. For the property $has\_response \in L$, the domain is defined by the function $D(has\_response) = Treatment$, while the range is $R(has\_response) = Response$. Figure 4.3b illustrates the ego network of the entity $T1$, where the neighbors of the entity $T1$ are defined by $\mathcal{N}(T1) = \{D1, D2, D3, D4, low\_effect\}$. Furthermore, the set of edges between pairs of entities in the set of neighbors of entity $T1$ is defined by $\alpha(\mathcal{N}(T1)) = \{(D1, interacts\_with, D2), (D2, interacts\_with, D4), (D3, interacts\_with, D2)\}$, where we can observe the three triples in Figure 4.3a. Note that although low_effect is in the ego network of the entity $T1$, this entity is not related to any other entity in this ego network.

An ideal knowledge graph: An ideal knowledge graph is a knowledge graph $\mathcal{KG}' = (V, E', L, C, I, D, R, \mathcal{N}, ego, \alpha)$ that contains all the true existing relations between entities in V. The Closed World Assumption (CWA) is assumed on $\mathcal{KG}'$, i.e., what is unknown to be true in $\mathcal{KG}'$ is false.

An actual knowledge graph: An actual knowledge graph $\mathcal{KG} = (V, E, L, C, I, D, R, \mathcal{N}, ego, \alpha)$ is a knowledge graph that follows the assumption Open World Assumption (OWA), i.e., what is not known to be true is just unknown and may be true.

A complete knowledge graph: A complete knowledge graph $\mathcal{KG}_{\text{comp}} = (V, E_{comp}, L, C, I, D, R\mathcal{N}, ego, \alpha)$ is a knowledge graph, which includes a relation for each possible combination of entities in V. Note that not all relationships in $\mathcal{KG}_{\text{comp}}$ are necessarily true.

A knowledge graph $\mathcal{KG}$ may only contain a portion of the edges represented in $\mathcal{KG}'$, i.e., $E \subseteq E'$; it represents those relations that are known and are not necessarily complete. On the other hand, since $\mathcal{KG}_{\text{comp}}$ is a complete knowledge graph, $E \subseteq E' \subseteq E_{\text{comp}}$. The set of missing edges in $\mathcal{KG}$ is defined as $\Delta(E', E) = E' - E$, i.e., it is the set of relations existing in the ideal knowledge graph $\mathcal{KG}'$ that are not represented in $\mathcal{KG}$. Figure 4.4 illustrates three knowledge graphs. Figure 4.4b is an ideal knowledge graph that states that only three relationships are true. The actual knowledge graph, presented in Figure 4.4a, is incomplete and only includes two relationships; $(C, p2, B)$ is unknown and is not part of the current knowledge graph. Figure 4.4c illustrates a complete knowledge graph, with a relation for each combination of entities in V and properties in L. All the possible relationships are included in this graph.



(a) Actual Knowledge Graph $\mathcal{KG}$

(b) Ideal Knowledge Graph $\mathcal{KG}'$

(c) Complete Knowledge Graph $\mathcal{KG}_{\text{comp}}$

Figure 4.4: **Example of actual, ideal, and complete knowledge graph**.

An **abstract target prediction** over a $\mathcal{KG}$ is defined in terms of a tuple $\tau = \langle \mathcal{KG}, r, prediction, DS, KGE \rangle$:

- $\mathcal{KG}$ is a knowledge graph $\mathcal{KG} = (V, E, L, C, I, D, R, \mathcal{N}, ego, \alpha)$.
- r represents a prediction property, $r \in L$.
- *prediction* indicates the head or the tail of triples to predict. A tail prediction of triples $\langle h, r, t \rangle$ is the process of finding t for the incomplete triple $\langle h, r, ? \rangle$, head predictions can be defined analogously.
- DS is the deductive database system over $\mathcal{KG}$.
- KGE is the knowledge graph embedding over $\mathcal{KG}$.

The deductive system DS derives new facts from inference rules and facts stored in a database [101]; it is expressed as a set of extensional and intensional rules in Datalog. Datalog considers two sets of clauses: a set of ground facts called the Extensional Database (EDB) and a Datalog program P called the Intensional Database (IDB). The predicates in the EDB and IDB are divided into two disjoint sets, EDB predicates, which occur in the EDB, and the IDB predicates, which occur in IDB. The head predicate of each clause in P is an IDB predicate, and the EDB predicate can occur in the body of the rule. If C_1 and C_2 are the domain and range of r

respectively, then EDB comprises ground facts of the form: $p(s, o)$ where the triple $(s, p, o) \in \mathit{ego}(v) \cup \alpha(\mathcal{N}(v))$, and $I(v) \in \{C_1, C_2\}$. The EDB in our DS contains ground facts from the ego networks and from their neighbors. Given a prediction property $r = \mathit{has\_response}$ we know the domain $D(\mathit{has\_response}) = Treatment$ and range $R(\mathit{has\_response}) = \mathit{Response}$. Figure 4.3a shows entities of type $Treatment$ and entities of type $\mathit{Response}$ for the domain and range of the property $\mathit{has\_response}$, respectively. The EDB comprises all the ground facts defined by the ego networks: $\mathit{ego}(T1), \mathit{ego}(T2), \mathit{ego}(\mathit{low\_ effect})$, and $\mathit{ego}(\mathit{effective})$, and their neighbors $\alpha(\mathcal{N}(T1)), \alpha(\mathcal{N}(T2)), \alpha(\mathcal{N}(Tlow\_ \mathit{effect}))$, and $\alpha(\mathcal{N}(\mathit{effective}))$, where entities $T1$ and $T2$ belong to class $Treatment$, and $\mathit{low\_ effect}$ and $\mathit{effective}$ belong to the class $\mathit{Response}$.

An example of EDB is the set of facts $\{interacts\_with(D1, D2), interacts\_with(D2, D4)\}$, where the property $interacts\_with \in L$ and the entities $\{D1, D2, D4\} \subseteq V$. The predicate $interacts\_with$ represents interactions between two drugs. Let P be a Datalog program containing the following clauses:

$$\begin{array}{lrl} r1 & interactsWith(A, X) & \Rightarrow \mathit{inferredInteraction}(A, X). \\ r2 & \mathit{inferredInteraction}(B, X), interactsWith(A, B) & \Rightarrow \mathit{inferredInteraction}(A, X). \end{array}$$

Rule $r2$ states that exist an $\mathit{inferred\_ interaction}$ between drug A and X, if there is another drug B which interacts with A with the predicate $interacts\_with$, and there is an $\mathit{inferred\_ interaction}$ from B to X. The evaluation results of $r2$ is $\{\mathit{inferred\_ interaction}(D1, D4)\}$, shown in Figure 4.3c with a red arrow.

KGE is a machine learning model that learns vector representation (i.e., KG embeddings) in a low dimensional continuous vector space for entities $v \in V$ and relations $e \in E$ in a $\mathcal{KG}$. KGE model exploits the $\mathcal{KG}$ structure to predict new relations in E. The KGE model resorts to a scoring function ϕ to estimate the plausibility of the vector representation of a triple, where higher ϕ values yield higher plausibility [110]. Link prediction is performed by identifying which vector representation of an entity provides the best values of the scoring function ϕ; these entities are added to the incomplete triples as heads or tails. If $\mathit{prediction} = \mathit{tail}$, then the link prediction task is the process of finding t as the best scoring tail for the incomplete triple $\langle h, r, ?\rangle$:

$$\underset{t \in V}{\operatorname{argmax}} \, \phi(h, r, t).$$

If $\mathit{prediction} = \mathit{head}$, it can be defined analogously. The state of the art of KGE methods may be negatively impacted by the data sparsity issue, i.e., ground facts that can be used as positive samples to guide KGE training represent only a minor portion. The proposed deductive database system for abstract target prediction alleviates the data sparsity issue by enhancing links in the ego network $ego(v)$, which are managed as new ground facts.

Suppose the abstract target prediction is defined for the current knowledge graph $\mathcal{KG}$ presented in Figure4.3a where the prediction property is $r = has\_response$, and the prediction corresponds to the tail, i.e., $prediction = tail$. The link prediction task predicts incomplete triples $\langle h, r, ? \rangle$, where the head h represents entities of class $Treatment$, i.e., entities h in V such that $I(h) = Treatment$, and the relation is $r = has\_response$.

4.2.2 Problem Statement

Given an actual knowledge graph $\mathcal{KG} = (V, E, L, C, I, D, R, \mathcal{N}, ego, \alpha)$ and its corresponding ideal knowledge graph $\mathcal{KG}' = (V, E', L, C, I, D, R, \mathcal{N}, ego, \alpha)$. Given an abstract target prediction over an actual knowledge graph $\mathcal{KG}$, $\tau = \langle \mathcal{KG}, r, prediction, DS, KGE \rangle$, we tackle the *problem of predicting relationships over* $\mathcal{KG}$.

Given a relation, $e \in \Delta(E_{\text{comp}}, E)$ (i.e., the set of missing edges in $\mathcal{KG}$), the problem of predicting relationships consists of determining whether $e \in E'$, i.e., if a relation e corresponds to an existing relation in the ideal knowledge graph $\mathcal{KG}'$. We are interested in finding the maximal set of relationships or edges E_a that belongs to the ideal $\mathcal{KG}'$, i.e., find a set E_a that corresponds to a solution of the following optimization problem:

$$\underset{E_a \subseteq E_{comp}}{\text{argmax}} \, |E_a \cap E'|.$$

4.2.3 Proposed Solution

Our proposed solution resorts to a symbolic system implemented by a deductive database to enhance the predictive capacity of the link prediction task solved by knowledge graph embedding models. The approach assumes that a link prediction problem is defined in terms of an abstract target prediction $\tau = \langle \mathcal{KG}, r, prediction, DS, KGE \rangle$ over a knowledge graph $\mathcal{KG} = (V, E, L, C, I, D, R, \mathcal{N}, ego, \alpha)$.

A Symbolic System: Deductive system DS corresponds to the deductive databases where the EDB comprises ground facts of the form: $p(s, o)$, where the triple $\langle s, p, o \rangle \in ego(v) \cup \alpha(\mathcal{N}(v))$, $I(v) \in \{C_1, C_2\}$, $C_1 = D(r)$, and $C_2 = R(r)$. The variables C_1 and C_2 represent the domain and range of the property r, respectively. The IDB contains rules that allow deducing new relationships in the ego network $ego(v)$. Then, the stratified negation is the fragment of Datalog used, which is a restricted form of negation. In stratified Datalog, rules are organized into layers or strata, where each stratum contains rules that depend only on rules

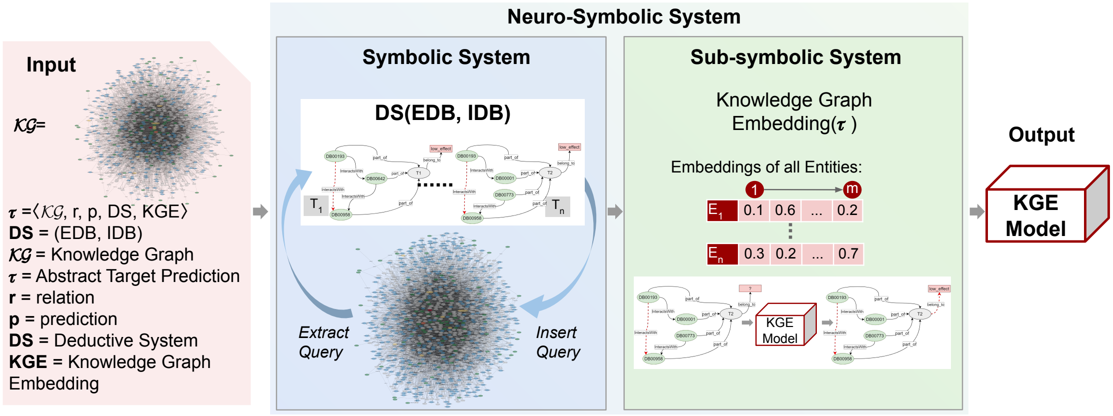


Figure 4.5: **Approach**. The input is a knowledge graph ($\mathcal{KG}$), an abstract target prediction τ, and a deductive system, and returns a KGE model. The symbolic system is implemented by a deductive system $DS(EDB, IDB)$ that deduces new relationships in the ego network $ego(v)$ and between their neighbors $\alpha(\mathcal{N}(v))$. Then, the sub-symbolic system implemented by a KGE model employs the $\mathcal{KG}$ with the deduced new relationships to predict incomplete triples. KGE solves the abstract target prediction τ for the relation r and the *prediction* head or tail.

in lower strata. It allows for the negation of facts or rules that are in the lower strata but not in the same or higher strata as a rule using negation. This restriction ensures that there are no circular dependencies between rules and that the program can be evaluated using a bottom-up approach. The computational method executed to empower the ego networks $ego(v)$ is built on the results of deductive databases to compute the minimal model of the deductive database[22]. This minimal model is defined in terms of the fixed-point assignment $\sigma^{ego(.)}_{\text{MINFIX}}$, that deduces relationships between entities v_i and v_j in the neighbors $\mathcal{N}(.)$. The minimal model for $DS(EDB, IDB)$ can be computed in polynomial time in the overall size of the ego network $ego(v)$ and the neighbors $\alpha(\mathcal{N}(v))$ for all the entities v where $I(v) \in \{C_1, C_2\}$, $C_1 = D(r)$, and $C_2 = R(r)$.

A Sub-symbolic System: A model to learn Knowledge Graph Embeddings solves the abstract target prediction τ over $\mathcal{KG}$ for the relation r and the *prediction* head or tail. The sub-symbolic system predicts incomplete triples of the way $\langle h, r, ?\, \rangle$ if $prediction = tail$ and $\langle ?\, , r, t\rangle$ if $prediction = head$.

The Integration of Symbolic and Sub-symbolic Systems: The ego network $ego(v)$ and the edges between their neighbors $\alpha(\mathcal{N}(v))$ are extended with explicit relationships among entities in the neighbors $\mathcal{N}(v)$ by the deductive system $DS(EDB, IDB)$. As a result, the symbolic system implemented by $DS(EDB, IDB)$ alleviates the data sparsity issues in $\mathcal{KG}$ that may negatively affect the learning of the KGE in the abstract target prediction τ.

4.2.4 The Symbolic and Sub-symbolic System Architecture

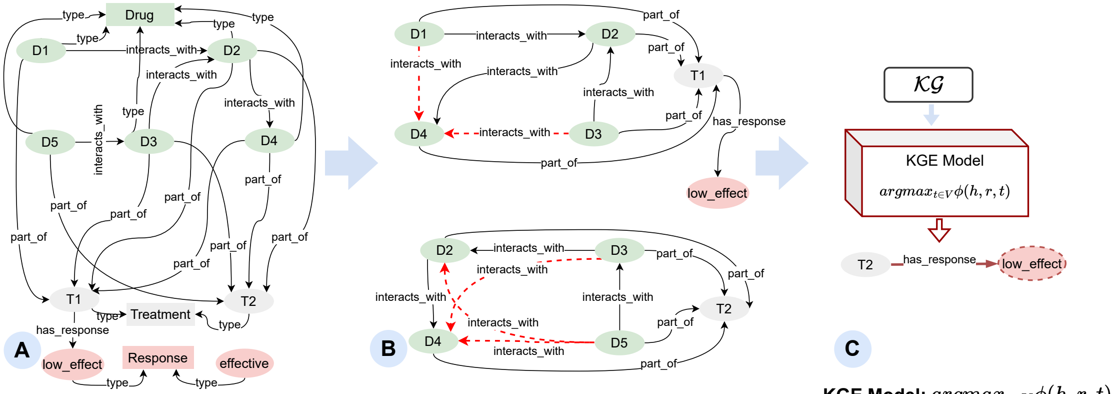
Figure 4.5 depicts the architecture that implements the proposed approach. The architecture receives a knowledge graph $\mathcal{KG} = (V, E, L, C, I, D, R, \mathcal{N}, ego, \alpha)$ and an abstract target prediction $\tau = \langle \mathcal{KG}, r, prediction, DS, KGE \rangle$, where $\mathcal{KG}$ is the knowledge graph, r is a property, *prediction* represents the head or tail of triples to predict, DS is the deductive system, and KGE is the knowledge graph embedding. The architecture returns a learned model of embeddings. These embeddings are used to solve the target prediction task defined by τ.

The architecture is composed of two main steps. First, the relationships implicitly defined by the deductive system are deduced by means of a Datalog program. Second, once $\mathcal{KG}$ is augmented with new deduced relationships, KGE learns a latent representation of entities and properties of $\mathcal{KG}$ in a low-dimensional space. The architecture is agnostic of the method to learn the embeddings. Moreover, our approach is domain-agnostic. For example, it can be applied in the context of Industry 4.0 to discover relations between standards and thus solve interoperability issues between standardization frameworks [105, 107].

4.2.5 Abstract Target Prediction Task. Running example

Albeit illustrated in the context of treatment response, the proposed method is domain-agnostic. It only requires the definition of the deductive system to enhance the relationships in the ego network of the entities v where $I(v) \in \{C_1, C_2\}$, $C_1 = D(r)$, and $C_2 = R(r)$. Figure 4.6 illustrates the proposed steps to enhance the predictive capacity by knowledge graph embedding models. The $\mathcal{KG}$ shown in Figure 4.6**(A)** is the same as in Figure 4.3a. Assuming we receive as input the abstract target prediction $\tau = \langle \mathcal{KG}, r, prediction, DS, KGE \rangle$, where the $\mathcal{KG}$ is represented in Figure 4.6**(A)**, the property is $r = has\_response$, the $prediction = tail$, DS is the deductive system, and KGE is the KGE algorithm. The EDB of the DS comprises all the ground facts of the form: $p(s, o)$, where the triple $(s, p, o) \in ego(v) \cup \alpha(\mathcal{N}(v))$, $I(v) \in \{C_1, C_2\}$, $C_1 = D(has\_response)$, and $C_2 = R(has\_response)$. Then, the domain and range of the property $r = has\_response$ are $Treatment$ and $Response$, respectively. In addition, the entity type for v in ego network $ego(v)$ is $Treatment$ or $Response$. The entities of type $Response$ are *low_ effect* and *effective*, and $T1$ and $T2$ are entities of type $Treatment$.

The EDB comprises all the ground facts defined by the ego networks: $ego(T1), ego(T2), ego(low\_\ effect)$, and $ego(effective)$, and their neighbors $\alpha(\mathcal{N}(T1)), \alpha(\mathcal{N}(T2)), \alpha(\mathcal{N}(low\_\ effect))$, and $\alpha(\mathcal{N}(effective))$. Figure 4.6**(B)** shows the ego networks $ego(T1)$ and $ego(T2)$ with the set of edges between pairs of entities in the set of neighbors of entity $T1$ and $T2$ defined by



Original Knowledge Graph $\mathcal{KG}$: contains five green entities belonging to class Drug, two gray entities belonging to class Treatment, and two red entities belonging to class Response.
The **target prediction** corresponds to links between entities of type Treatment and Response. Thus, the head of a target prediction is a treatment, while the tail is a response.

Enhancing links between entities in the ego network ego(T1) and ego(T2). Both ego networks contain the relations between the set of neighbors of T1 and T2 defined by $\alpha(\mathcal{N}(T1))$ and $\alpha(\mathcal{N}(T2))$.
The deductive system DS deduces new relationships in ego(T1) and ego(T2). The red arrows represent the deduced relationships. The deductive system enriches the $\mathcal{KG}$.

KGE Model: $argmax_{t\in V}\phi(h,r,t)$
KGE model infers missing links for each triple using the highest score of the function $\phi(h,r,t)$.
KGE model places T1 and T2 nearby in the embedding space after deducing new facts in their ego networks. The KGE model predicts the tail = low_effect for the missing link $\langle T2, has\_response, ?\rangle$

Figure 4.6: **Running example**. Figure 4.6 illustrates the proposed steps to enhance the predictive capacity by KGE models. **Step A**: given a KG and an abstract target prediction τ the ego networks $ego(v)$ are defined. **Step B**: illustrates the ego network $ego(T1)$ and $ego(T2)$ and a deductive system deduces new relationships to enhances the $ego(v)$ in $\mathcal{KG}$. **Step C**: depicts a KGE model in which predictive capability is enhanced by symbolic reasoning. The relationships in E and the new facts deduced by DS improve the link prediction task.

$\alpha(\mathcal{N}(T1))$ and $\alpha(\mathcal{N}(T2))$, respectively. Then, DS deduces new relationships enhancing the links in the $ego(T1)$ and $ego(T2)$; red arrows represent the deduced relationships. Considering the Datalog program $P(1)$ as the IDB for DS, the facts $inferred\_interaction(D1, D4)$, $inferred\_interaction(D3, D4)$, $inferred\_interaction(D5, D4)$, and $inferred\_interaction(D5, D2)$ are deduced enhancing the ego network.

The SPARQL query in Listing 1 extracts the ego network $ego(T1)$ and the set of edges between pairs of entities in the set of neighbors of entity $T1$ defined as $\alpha(\mathcal{N}(T1))$. Listing 1 illustrates a CONSTRUCT query that returns RDF triples in the form of subject, predicate, and object and represents the ground facts of the EDB. The predicate represents the ground predicated in the EDB, the subject represents the first term of the ground predicated, and the object represents the second term.

The IDB described by the Datalog program $P(1)$ allows deducing new relationships and increasing the ego networks $ego(T1)$ and $ego(T2)$. The deduced relations are inserted into the $\mathcal{KG}$ through the SPARQL query in Listing 2. The deduced relationship e belongs to E', i.e., $e \in E'$.

```
PREFIX ex: <http://example/vocab/>
CONSTRUCT {?A <interacts_with> ?B} WHERE {
        ?A ex:part_of <T1> .
        ?B ex:part_of <T1> .
        ?A ex:interacts_with ?B .
```

Listing 1: **SPARQL query to ground the extensional predicate** $\mathit{interacts\_with}(A, B)$

```
PREFIX ex: <http://example/vocab/>
INSERT DATA {
        <A> ex:interacts_with <X>
        }
```

Listing 2: **SPARQL query to insert the deduced relationships from the intensional predicate** *inferred_ interaction(A,X)*

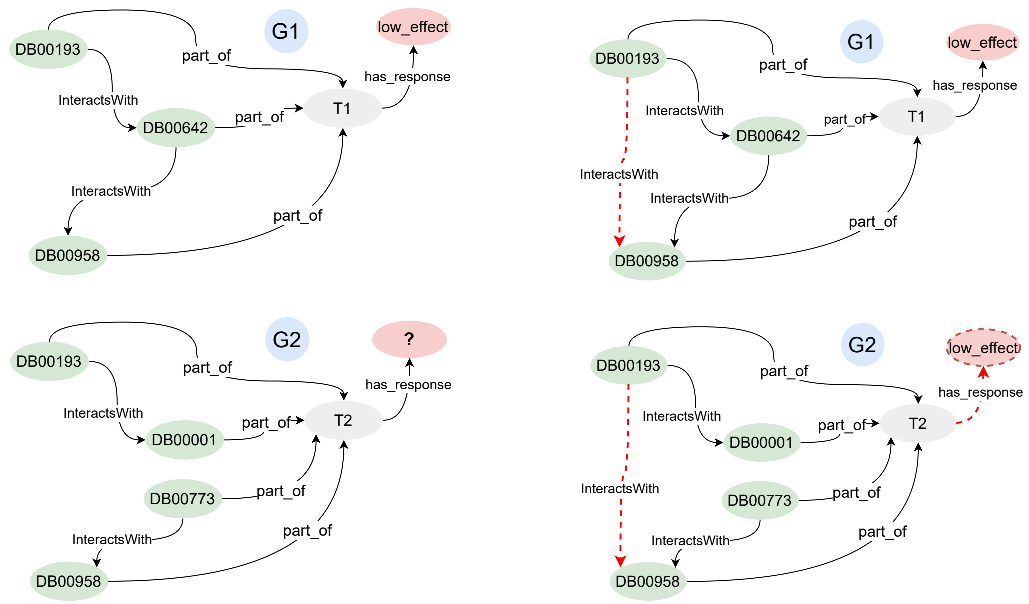
Figure 4.6**(C)** illustrates a KGE model in which the symbolic system enhances the predictive capacity. The DS increases the relationships E in $\mathcal{KG}$, alleviating the data sparsity issues in $\mathcal{KG}$. Thus, the $\mathcal{KG}$ that contains new facts deduced by DS guides the KGE model, improving the link prediction task for $r = has\_response$ and $prediction = tail$. Figure 4.6**(C)** shows the link prediction task of finding t as the best scoring tail for the incomplete triple $(T2, has\_response, ?\,)$: $\text{argmax}_{t \in V}\, \phi(T2, has\_response, t)$. Treatment $T2$ is predicted to have a response *low_ effect* (*T2, has_ response, low_ effect*), i.e., $T1$ and $T2$ are nearby in the embedding space after enhancing the $ego(v)$ in $\mathcal{KG}$.

4.3 Polypharmacy Treatment Effectiveness Prediction with a Neuro-Symbolic AI System

As a proof concept, we apply our Neuro-Symbolic AI approach to address the problem of predicting polypharmacy treatment effectiveness. We have implemented a deductive system on top of a Treatment Knowledge Graph ($\mathcal{KG}$). The technique aims to identify the combination of drugs whose interactions may affect the treatment's effectiveness. Then, the problem of predicting treatment effectiveness is modeled as a problem of link prediction between treatments and the responses: *low-effect* or *effective*.

4.3.1 Motivating Example

We motivate our work in healthcare, specifically for predicting polypharmacy treatment response. Polypharmacy is the concurrent use of multiple drugs in treatments, and it is a standard procedure to treat severe diseases, e.g., lung cancer. Polypharmacy is a topic of concern due to the increasing number of unknown drug-drug interactions (DDIs) that may affect the response to medical treatments. There are two types of DDIs, pharmacodynamics, i.e., *the effect of a drug in the body*, and pharmacokinetics, i.e., *the course of a drug in the body*. Pharmacokinetics DDIs alter a drug's absorption, distribution, metabolism, or excretion. For example, an increase in absorption will increase the object drug's bioavailability and vice versa. If a DDI affects the object's drug distribution, the drug transport by plasma proteins is altered. Moreover, a drug's therapeutic efficacy and toxicity are affected when a pharmacokinetics DDI alters the object's drug metabolism.



(a) **Oncological Treatment**

(b) **Predicting Response of Oncological Treatment**

Figure 4.7: **Motivating Example.** Figure 4.7a shows two polypharmacy oncological treatments, *T1* and *T2*, represented in RDF. The drugs *DB00193*, *DB00642*, and *DB00958* are part of *T1*, and the drug-drug interactions are represented by the property *InteractsWith*. The therapeutic response of *T1* is annotated as *low_ effect* by the property *has_ response*, while the therapeutic response of *T2* is unknown. Figure 4.7b depicts the ideal RDF graph, where a symbolic system generates a new DDI between *DB00193* and *DB00958*. Ideally, a sub-symbolic system detects that both treatments are similar and predicts the effectiveness of *T2* as low effective.

Lastly, if the excretion of an object drug is reduced, the drug's elimination half-life will be increased. Notice that the pharmacokinetic interactions can be encoded in a symbolic system.

Figure 4.7a shows two polypharmacy oncological treatments encoded in RDF. We extract the known DDIs between the drugs of these treatments from DrugBank[2]. However, polypharmacy therapies produce unforeseen DDIs due to drug interactions in the treatment. Since DDIs affect the effectiveness of a treatment, there is a great interest in uncovering these DDIs. Figure 4.7b depicts an ideal RDF graph where all the true relations are explicitly represented. Dotted red arrows represent DDI between the drugs `DB00193` and `DB00958` that are generated as the result of DDIs among drugs in the treatment. Rules that specify how these DDIs are generated can be represented in a Datalog program where the extensional database corresponds to facts representing explicit relationships. On the other hand, the implicit DDIs can be deduced via the intensional rules of the deductive system. The DDI between `DB00193` and `DB00958` increases the description of treatments `T1` and `T2`, enabling both treatments to share more relationships. Then, a sub-symbolic system, e.g., implemented using a KGE model, can explore this enhanced and make a more accurate prediction of the treatment response by employing the deduced DDIs. For example, the geometric model *TransH* places `T1` and `T2` nearby in the embedding space after deducing DDIs and predicts the therapeutic response of `T2`. As a result, this neuro-symbolic system enhances treatment information by identifying drug combinations whose interactions may affect treatment effectiveness. We propose an approach that resorts to symbolic reasoning implemented by a Datalog database and stage-of-the-art KGE models; it deduces DDIs within a treatment. Then, the KGE model embeds all the knowledge in the graph and predicts treatment responses. Although we depict the method in the context of treatment effectiveness, this approach is domain-agnostic and could be applied to any other link prediction task.

4.3.2 Treatment Knowledge Graph Creation

The P4-LUCAT consortium[1] collected heterogeneous data sources that comprise clinical records, drugs, and scientific publications and built a knowledge graph that provides an integrated view of these data. The KG is built with the aim of personalized medicine for Lung Cancer treatments. The treatments are extracted from Electronic Health Records (EHRs) from the Hospital Universitario Puerta del Hierro of Majadahonda of Madrid (HUPHM). Furthermore, the DDIs are extracted from DrugBank, in the approved category. The interactions' type and effect are extracted using named entity and linking methods implemented by Sakor et

[1] `https://p4-lucat.eu/`

Table 4.1: **Summary of the Lung Cancer Knowledge Graph**.

Knowledge Graph for Lung Cancer	**Records**
Lung Cancer Patients	1'242
Lung Cancer Drug	45
Chemotherapy Drug	7
Immunotherapy Drug	3
Antiangiogenic Drug	2
Tki Drug	5
Non Oncological Drug	41
Oncological Surgery	9
Tumor Stage	6
Publications	178'265
Drugs	8'453
Drug-Drug Interactions	1'550'586

al. [114]. These methods have also been used to extract DDIs in covid-19 and lung cancer treatments [115, 131]. Table 4.1 contains a summary of the number of annotations by classes in the Lung Cancer Knowledge Graph.

Figure 4.8 describes a Lung Cancer patient in the Lung Cancer Knowledge Graph. The patient *P1* is in stage II and has surgery. Also, *P1* received treatment on *10.07.2020* with an effective therapeutic response. In that treatment, *P1* is treated with a combination of chemotherapy drugs and one non-oncological drug. Drug-Drug Interactions with the effect and their impact is reported.

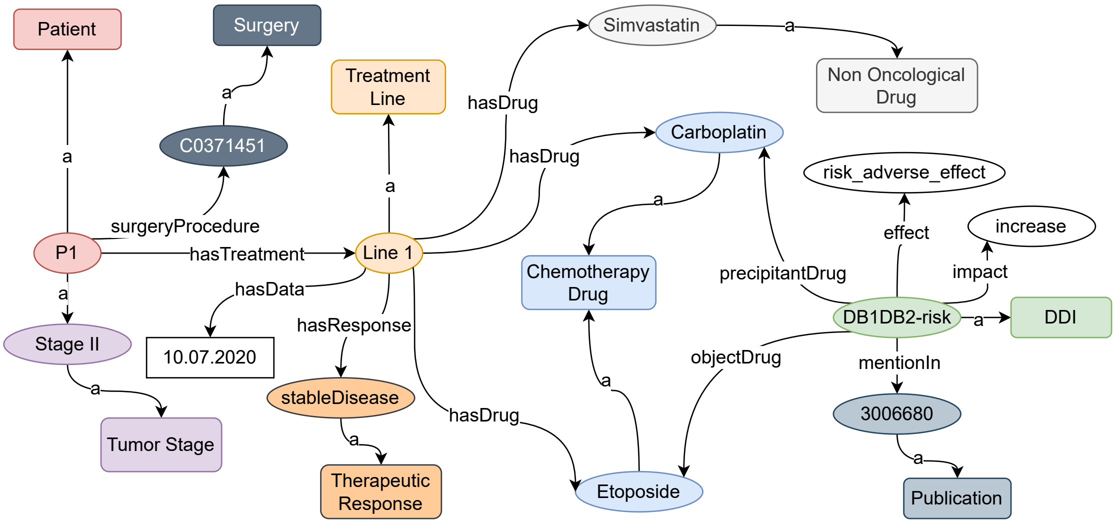
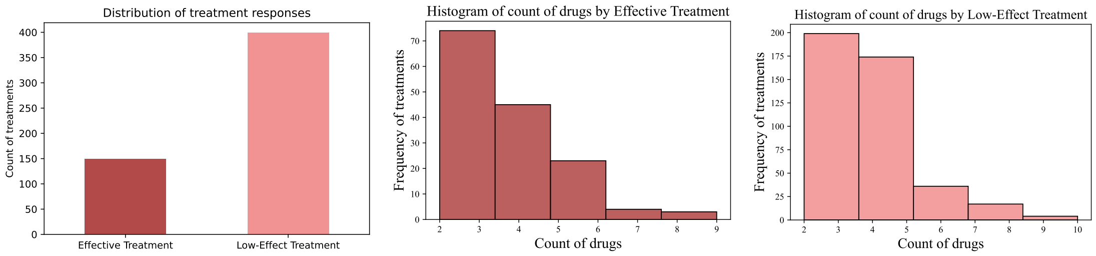


Figure 4.8: **Representation of a patient in the Lung Cancer Knowledge Graph**.

The input $\mathcal{KG}$ in our use case contains 548 polypharmacy cancer treatments $\mathcal{T}$ extracted from lung cancer clinical records, with the therapeutic response from each of them and the known Drug-Drug Interactions. The therapeutic response is the target class and can be set to the value *low-effect* or *effective* treatment. The meaning of an *effective* treatment is because of a complete therapeutic response or stable disease. A *low-effect* treatment means a partial therapeutic response or disease progression. Figure 4.9 depicts a descriptive analysis of the treatment response according to the data extracted from the clinical records. Figure 4.9a shows the treatment response distribution, where there are 149 *effective* treatments and 399 *low-effect* treatments. Figure 4.9b and 4.9c present the histogram for the class *effective* and *low-effect*, respectively. We can observe that there are treatments with nine and ten drugs in both treatments' response classes. Also, the most *low-effect* treatments are composed of more drugs than *effective* treatments. The rate of drugs between five and ten can be explained by the fact that in patients with multiple comorbidities, multiple drugs are prescribed to treat the disease.



(a) **Distribution of classes** (b) **Histogram Effective Treatment** (c) **Histogram Low-Effect Treatment**

Figure 4.9: **Descriptive analysis of the treatment responses.**

For each treatment, $t_i \in \mathcal{T}$, the DDIs and their effect are known from DrugBank [140]. Then, the treatments, the treatment response, the drugs, DDIs, and DDI effects for each treatment are managed in $\mathcal{KG}$. The polypharmacy treatment knowledge graph $\mathcal{KG} = (V, E, L, C, I, D, R, \mathcal{N}, ego, \alpha)$ is defined as follows:

- The types Drug, Treatment, DDI, Effect of DDI, and Treatment Response belong to C.
- Drugs, Treatments, DDIs, Effect of DDI, and Treatment Response are represented as instances of V.
- Edges in E that belong to $V \times V$ represent relations about drugs into a treatment.
- Properties *tgk:has_response*, *tgk:part_of*, *tgk:precipitant_drug*, *tgk:object_drug*, *tgk:ddiEffect*, *tgk:hasInteraction*, and *rdf:type* correspond to labels in L.

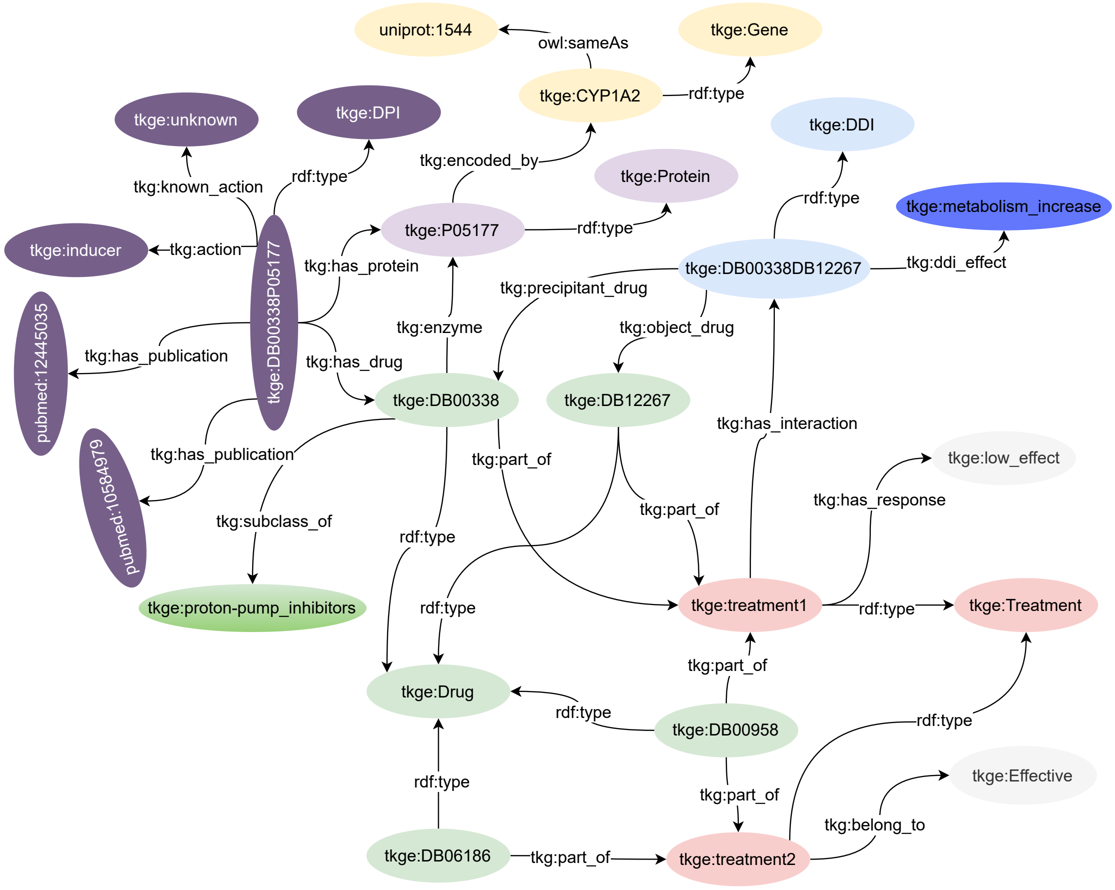


Figure 4.10: **Portion of Polypharmacy Treatment Knowledge Graph $\mathcal{KG}$.** The entity treatment *tkge:treatment1*, is composed by three drugs represented by the nodes, *tkge:DB00338*, *tkge:DB12267*, and *tkge:DB00958*. The entity treatment *tkge:treatment2*, contains two drugs and shares *tkge:DB00958* with *tkge:treatment1*. The node *tkge:DB00338DB12267* represents a DDI in the *treatment1* where the *tkge:DB00338* is the precipitant, and *tkge:DB12267* is the object drug. The effect of the DDI is represented by the node *tkge:metabolism_increase*. The treatment *tkge:treatment1* has a low effective response represented by the property *tkg:has_response*.

Figure 4.10 shows a portion of $\mathcal{KG}$. We model treatments, their prescribed drugs, drug-drug interactions, drug-protein interactions, publications related to the drug-protein interactions, and the gene that encodes the proteins in a knowledge graph.

4.3.3 Symbolic System. Deductive Database

Let $\tau = \langle \mathcal{KG}, r, prediction, DS, KGE \rangle$ be the input abstract target prediction, where $\mathcal{KG}$ is the polypharmacy treatment knowledge graph, $r = has\_response$, $prediction = tail$, DS the deductive database system, and KGE the knowledge graph embedding algorithm. The IDB of the DS comprises a set of rules to deduce new DDIs in treatments. A DDI is deduced when a set of drugs are taken together and is represented as a relation in the minimal model of the deductive database DS. The extensional database corresponds to statements about interactions between drugs stated in $\mathcal{KG}$. The ground predicates included in the EDB are the following; they are extracted from the KG by executing SPARQL queries:

$rule_1(serum, increase).$

$rule_1(metabolism, decrease).$

$rule_1(absorption, increase).$

$rule_1(excretion, decrease).$

precipitant(DB00958DB06186, DB00958).

effect(DB00958DB06186, excretion).

$rule_2(serum, decrease).$

$rule_2(metabolism, increase).$

$rule_2(absorption, decrease).$

$rule_2(excretion, increase).$

object(DB00958DB06186, DB06186).

impact(DB00958DB06186, decrease).

SPARQL queries in Listing 3 and Listing 4 declaratively define the ground $rule_1$ and $rule_2$ in the EDB. Both queries are executed on top of the $\mathcal{KG}$; the CONSTRUCT query returns RDF triples in the form of subject, predicate, and object. The predicate in the RDF triples represents the ground predicate in the EDB.

```
PREFIX tkg: <http://research.tib.eu/lung-cancer/vocab/>
PREFIX tkge: <http://research.tib.eu/lung-cancer/entity/>
CONSTRUCT {?E <rule1> ?I} WHERE {
        ?ddi a tkge:DDI .
        ?ddi tkg:effect ?E .
        ?ddi tkg:impact ?I .
        FILTER((?E in (tkge:serum, tkge:absorption) && ?I="increase") ||
               (?E in (tkge:metabolism, tkge:excreation) && ?I="decrease")) }
```

Listing 3: **SPARQL query to ground the extensional predicate** $rule_1(E, I)$

The facts included in the ground predicates *precipitant, object, effect*, and *impact* from the EDB are extracted using the CONSTRUCT query of Listing 5. The EDB contains thousands of facts for those predicates; therefore, only a few ground facts are presented.

The above-mentioned $rule_1$ identifies the combinations of effect and impact that alter the toxicity of an object drug, while $rule_2$ extracts the combinations of effect and impact that alter the effectiveness of an object drug. The intensional database (a.k.a. IDB) comprises Horn rules that state when a new DDI can be

```
PREFIX tkg: <http://research.tib.eu/lung-cancer/vocab/>
PREFIX tkge: <http://research.tib.eu/lung-cancer/entity/>
CONSTRUCT {?E <rule2> ?I} WHERE {
        ?ddi a tkge:DDI .
        ?ddi tkg:effect ?E .
        ?ddi tkg:impact ?I .
        FILTER((?E in (tkge:serum, tkge:absorption) && ?I="decrease") ||
               (?E in (tkge:metabolism, tkge:excreation) && ?I="increase")) }
```

Listing 4: **SPARQL query to ground the extensional predicate** $rule_2(E, I)$

```
PREFIX tkg: <http://research.tib.eu/lung-cancer/vocab/>
PREFIX tkge: <http://research.tib.eu/lung-cancer/entity/>
CONSTRUCT { ?ddi <precipitant> ?A .
            ?ddi <object> ?B .
            ?ddi <effect> ?E .
            ?ddi <impact> ?I } WHERE {
    ?ddi a tkge:DDI .
    ?ddi tkg:precipitant ?A .
    ?ddi tkg:object ?B .
    ?ddi tkg:effect ?E .
    ?ddi tkg:impact ?I }
```

Listing 5: **SPARQL query to extract the ground the extensional predicates** *precipitant(ddi,A)*, *object(ddi,B)*, *effect(ddi,E)*, and *impact(ddi,I)*

deduced as a result of the combination of the treatment's drug. These rules are negation free; thus, the interpretation of the deductive database corresponds to the minimal model of the EDB and IDB. The intensional database relies on the fact that pharmacokinetic DDIs cause the concentration of one of the interacting drugs (a.k.a. object) to be altered when combined with the other drug (a.k.a. precipitant). Thus, the absorption, distribution, metabolism, or excretion rate of the object drug is affected. Whenever the object drug absorption is decreased (resp. increased), the bioavailability of the drug is also affected. Furthermore, any alteration in the metabolism or excretion of the object drug has consequences on the therapeutic efficacy and toxicity of the drug. The following Datalog rules state the effect of pharmacokinetic DDIs:

$$\textit{precipitant(ID, A), object(ID, B), effect(ID, E), impact(ID, I)} \Rightarrow ddi(A,E,I,B). \quad (1)$$

$$ddi(A,E,I,B) \Rightarrow inferred\_ddi(A,E,I,B). \quad (2)$$

$$inferred\_ddi(A,E_2,I_2,B), ddi(B,E,I,C), rule_1(E,I), rule_1(E_2,I_2), (A! = C) \Rightarrow inferred\_ddi(A,E,I,C). \quad (3)$$

$$inferred\_ddi(A,E2,I2,B), ddi(B,E,I,C), rule_2(E,I), rule_2(E_2,I_2), (A! = C) \Rightarrow inferred\_ddi(A,E,I,C). \quad (4)$$

Rule (2) states the base case of the IDB. The predicate symbol *ddi* represents the DDIs with their effect and impact in $\mathcal{KG}$. Precipitant drug A generates effect E (e.g., absorption, excretion, metabolism, serum concentration) with impact I (e.g., increase or decrease) in object drug B. The predicate symbol *inferred_ ddi* expresses a deduced DDI, where the first term is the precipitant drug, the second and third terms represent the value of the property effect and impact of the DDIs deduced, and the last term is the object drug. Rule (3) and (4) define the effects of combining drugs that interact in a polypharmacy treatment and comprises the clauses to deduce relationships encoded in $\mathcal{KG}$. The head predicate *inferred_ ddi* becomes valid when the predicate symbols in the body of the rule are also valid. The DDIs deduced from the Rule (3) increase the toxicity of the object drug, and the DDIs deduced from Rule (4) alter the effectiveness of the object drug. Those deduced DDIs are aggregated to the $\mathcal{KG}$; they represent valuable insights into each treatment. Each DDI deduced, which is part of the minimal model of the IDB predicate *inferred_ ddi(A,E,I,C)*, is inserted into the $\mathcal{KG}$ using the query shown in Listing 6. From the motivating example, we can observe that by applying the DDI deductive system to the treatment `T1` in Figure 4.7a, a new DDI is deduced in Figure 4.7b; it represents a new true triple enhancing the treatment information, reducing thus, data sparsity.

4.3.4 Sub-Symbolic System. Knowledge Graph Embedding Model

Once the deductive system DS deduces new DDIs, the Knowledge Graph Embedding algorithm KGE is applied to learn a latent representation of the entities in a low-dimensional space. The DS increases the relationships in the ego networks $ego(v)$ such as $I(v) \in \{C_1, C_2\}$, $C_1 = D(has\_response)$, and $C_2 = R(has\_response)$. The DS minimizes the data sparsity issues by augmenting the description of the treatments with newly deduced DDIs. Then, KGE

```
PREFIX tkg: <http://research.tib.eu/lung-cancer/vocab/>
INSERT DATA {
        <ddi> tkg:precipitant <A> .
        <ddi> tkg:object <C> .
        <ddi> tkg:effect <E> .
        <ddi> tkg:impact <I>
        }
```

Listing 6: **SPARQL query to insert the deduced DDI from the intensional predicate** *inferred_ddi(A,E,I,C)*

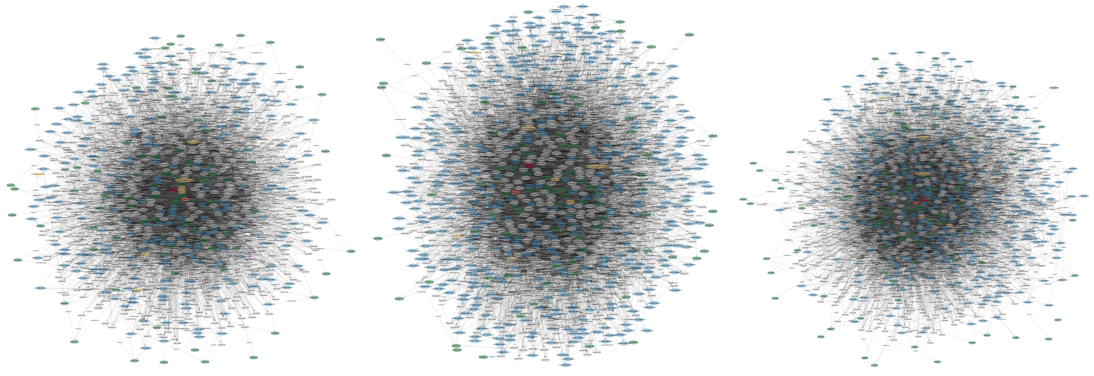
is able to improve the entities' representation in the embedding space. Thus, the scoring function $\phi(h, r, t)$ of the KGE is improved, and the link prediction task infers missing links that correspond to triples $\langle h, r, ? \rangle$, where $I(h) = D(r)$ and $r = has\_response$. Symbolic and sub-symbolic systems are highly complementary to each other. Sub-symbolic AI systems are able to solve complex problems that humans cannot analyze to draw conclusions or make predictions. Sub-symbolic methods are generally robust to data noise, while symbolic systems are vulnerable to data noise, which contrasts with the strength of sub-symbolic approaches.

4.4 Experimental Study

In this chapter, we empirically assess the impact of the DDIs encoded in $\mathcal{KG}$ on our approach's behavior. In particular, this chapter explores the following research questions: **Q1** Can the problem of predicting treatment effectiveness be effectively modeled as a problem of link prediction? **Q2** Can the Symbolic System for an abstract target prediction improve the link prediction capacity of the KGEs? **Q3** Can knowledge encoded in drug-drug interactions enhance the accuracy of the predictive task?

4.4.1 Experiment Setup

We empirically evaluate the effectiveness of our approach to capture knowledge encoded in $\mathcal{KG}$ and predict polypharmacy treatment response.



(a) $\mathcal{KG}_{basic}$ (b) $\mathcal{KG}$ (c) $\mathcal{KG}_{random}$

Figure 4.11: **Benchmarks to evaluate.** Figure 4.11a represents the $\mathcal{KG}_{basic}$, and it includes treatments from clinical records and pharmacokinetic DDI extracted from Drugbank. Figure 4.11b represents the $\mathcal{KG}$ and includes treatments from clinical records, pharmacokinetic DDI extracted from Drugbank, and a new set of pharmacokinetic DDI deduced by the DDI Deductive System. Figure 4.11c represents the $\mathcal{KG}_{random}$ and includes treatments from clinical records, pharmacokinetic DDI extracted from Drugbank, and the same number of new links deduced in $\mathcal{KG}$ is generated randomly.

Benchmarks

We conduct our evaluation over three Knowledge Graphs represented in Figure 4.11. $\mathcal{KG}_{basic}$ is the Knowledge Graph which only contains for each polypharmacy treatment the DDIs and their effect extracted from Drugbank. The second Knowledge Graph, $\mathcal{KG}$, includes not only the DDIs extracted from DrugBank but also the ones deduced by Deductive Database, i.e., it contains new deduced DDIs and their effects. Lastly, the third Knowledge Graph, $\mathcal{KG}_{random}$ is created from $\mathcal{KG}_{basic}$; it also includes the same number of links included in $\mathcal{KG}$, but these links are randomly generated, i.e., they correspond to false or true relations. We **aim** to validate whether the links discovered by our DDI Deductive System improve the prediction of treatment responses.

Knowledge Graph Embedding Models

We utilize eleven models to compute latent representations, e.g., vectors, of entities and relations in the three KGs and then employ them to infer new facts. In particular, we utilize three main families of models:

- Tensor Decomposition models such as *HolE* and *RESCAL*.

Table 4.2: **Statistics of Knowledge Graph**. Metrics to measure size, diversity, and sparsity in Knowledge Graph

KG	T	E	R	RE	EE	RD	ED
$\mathcal{KG}_{basic}$	5630	1069	7	1.615	10.846	804.286	10.533
$\mathcal{KG}$	6675	1069	7	1.726	10.989	953.571	12.488
$\mathcal{KG}_{random}$	6675	1069	7	1.710	11.291	953.571	12.488

- Geometric models such as *RotatE*, *QuatE*, and the Trans* family models *TransE*, *TransH*, *TransD*, and *TransR*.
- Deep Learning models such as *UM*, *SE* and *ERMLP*.

The PyKEEN (Python KnowlEdge EmbeddiNgs) framework [4] is used to learn the embeddings. The hyper-parameters utilized to train the model are epoch number 200 and training loops: stochastic local closed world assumption (sLCWA). The negative sampling techniques used are Uniform negative sampling and Bernoulli negative sampling. The embedding dimensions and the rest of the parameters are set by default. To assure statistical robustness, we apply 5-fold cross-validation. For evaluating the performance of embeddings methods, we measure the metrics: $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$, $F1\text{-}Score = \frac{2*(precision*recall)}{(precision+recall)}$.

Implementations

The pipeline for predicting polypharmacy treatment response has been implemented in Python 3.9. Experiments are executed using 12 CPUs Intel® Xeon(R) W-2133 at 3.60GHz, 64 GB RAM, and 1 GPU GeForce GTX 1080 Ti/PCIe/SSE2 with 12 GB VRAM. We used the library pyDatalog[2] to develop the Deductive System and the library PyKEEN[3], to learn the embeddings.

4.4.2 Metrics to Characterize the Benchmarks

Table 4.2 shows the statistics of the three KGs. We considered the metrics, Number of Triples (T), Entities (E), and Relations or properties (R), to measure the size in KG. The metrics Relation entropy (RE) and Entity entropy (EE) are considered to measure diversity and Relational density (RD) and Entity density (ED) to measure sparsity in Knowledge Graph.

The metrics RE and EE measure the distribution of relationships and entities in the KG, respectively. Higher values of RE mean that all possible relations

[2] https://sites.google.com/site/pydatalog/home

[3] https://pykeen.readthedocs.io/en/stable/index.html

are equally probable, and lower values mean one or more relations have a high probability. The values of the metric RE mean that all possible relations in $\mathcal{KG}$ are more equally probable than all possible relations in $\mathcal{KG}_{basic}$ and $\mathcal{KG}_{random}$. The three KGs have a higher EE value than RE as they use a small set of manually defined relations but contain many entities. The metrics RD and ED measure the sparsity of entities and relationships in the KG, respectively. We measure sparsity as information density, where RD means average triples per relation and ED is the average triples per entity. $\mathcal{KG}_{basic}$ has the lower average triples per relation and entity while $\mathcal{KG}$ and $\mathcal{KG}_{random}$ have the higher average triples per entity. The metrics evaluated in Table 4.2 are defined in the paper [99], implemented in our GitHub[4].

4.4.3 Impact of Capturing Symbolic Knowledge

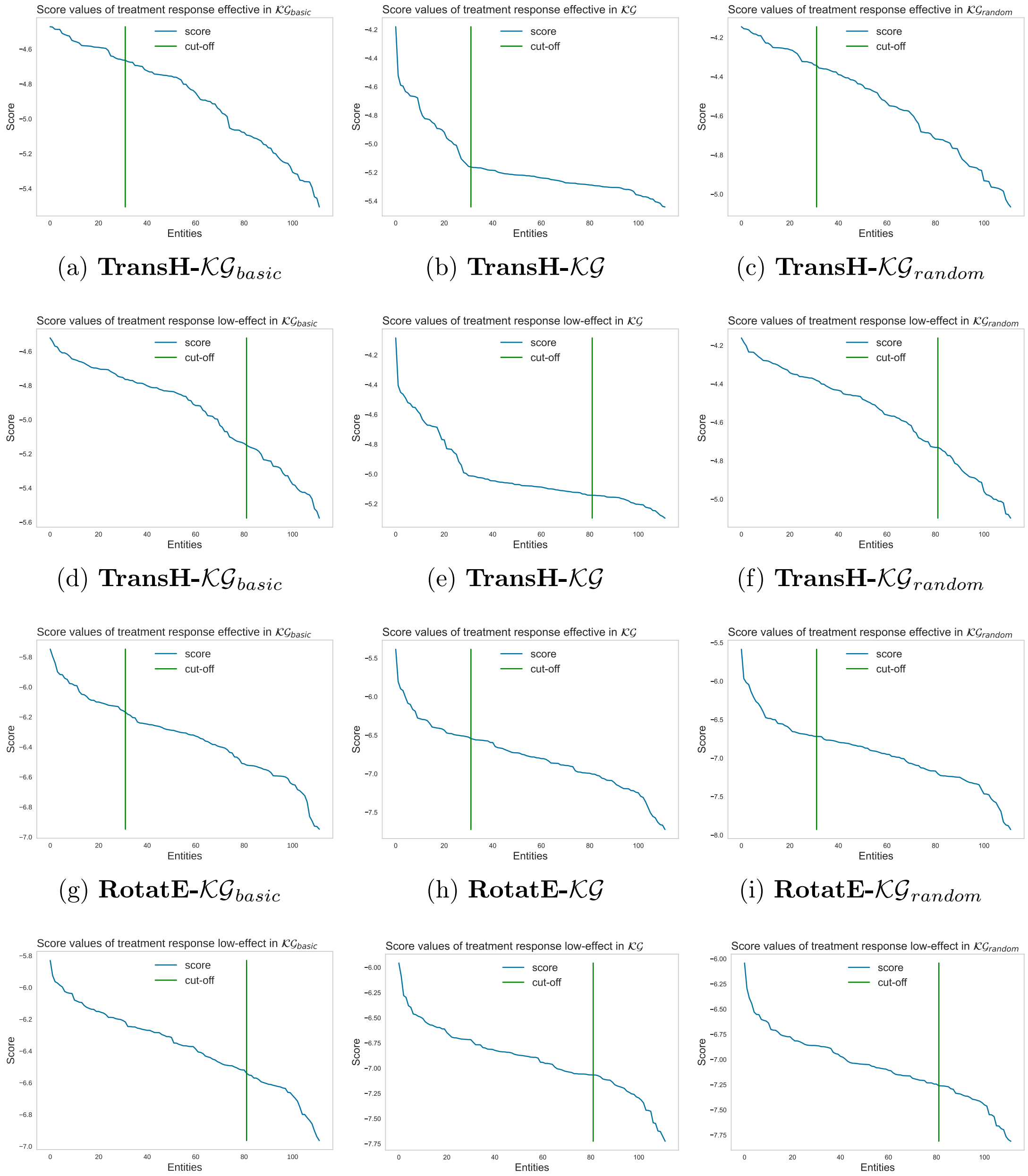
Figure 4.12 shows the behavior of the scoring function for the entities predicted by *TransH* and *RotatE* embedding models. For the purpose of brevity, we only show the score value results for two embedding models. The evaluation material is available in the GitHub repository[5]. We can notice how DS for the prediction property $r = has\_response$ is impacting the KGE models. Figure 4.12a to 4.12c and Figure 4.12g to 4.12i show the score values of the entities predicted on the link prediction task given the predicate *ex:has_response* and object *effective* by the *TransH* and *RotatE* models, respectively. Figure 4.12d to 4.12f and Figure 4.12j to 4.12l report on the score values of the entities predicted given the predicate *ex:has_response* and object *low-effect* by the *TransH* and *RotatE* models, respectively. We can observe that both models have different behaviors for each benchmark ($\mathcal{KG}_{basic}, \mathcal{KG}, \mathcal{KG}_{random}$). The vertical line in each plot represents the cut-off in a specific percentile. The percentile used for each KG was based on the percentage of links to the entity *effective* and *low-effect* in the KG. The portion of entities predicted, delimited by the vertical line, is evaluated in terms of precision, recall, and f1-score.

4.4.4 Evaluating the performance of our integrated Symbolic-Sub-symbolic System

The selected portions of entities predicted are measured with precision, recall, and f1-score on average because of cross-validation. Figure 4.13 and Figure 4.14 show the evaluation of the Link Prediction task through Uniform negative sampling and Bernoulli negative sampling, respectively. Uniform sampling randomly chooses

[4] `https://github.com/SDM-TIB/Statistics`$_{\text{K}}$`nowledgeGraph`

[5] `https://github.com/arivasm/Neuro-Symbolic`$_{\text{T}}$`reatment-Response.git`



(a) **TransH-$\mathcal{KG}_{basic}$** (b) **TransH-$\mathcal{KG}$** (c) **TransH-$\mathcal{KG}_{random}$**

(d) **TransH-$\mathcal{KG}_{basic}$** (e) **TransH-$\mathcal{KG}$** (f) **TransH-$\mathcal{KG}_{random}$**

(g) **RotatE-$\mathcal{KG}_{basic}$** (h) **RotatE-$\mathcal{KG}$** (i) **RotatE-$\mathcal{KG}_{random}$**

(j) **RotatE-$\mathcal{KG}_{basic}$** (k) **RotatE-$\mathcal{KG}$** (l) **RotatE-$\mathcal{KG}_{random}$**

Figure 4.12: **Score value of the predicted entities.** The green line represents the cut-off at the 27 and 73 percentiles for the three KGs.

the candidate entity based on a uniform probability between all possible entities. Bernoulli sampling corrupts the head with probability p and the tail with $1 - p$, where p is an average number of unique tail entities per unique head entities given a relation r. The relation with cardinally *1-n* has a higher probability of corrupting the head, and relations *n-1* have a higher probability of corrupting the tail. Figure 4.13 and 4.14 show the results of the three KG benchmarks. Each plot depicts the results of a metric for each embedding model and KG. The best performing embedding model in the three metrics is *TransH*. The KGE models have all better performance in $\mathcal{KG}$ obtained in the three metrics in both negative sampling techniques. In addition, the worst performance is observed in $\mathcal{KG}_{random}$. These results suggest that the deduced DDIs by the Deductive System are meaningful to the treatment responses. More importantly, they put the crucial role of the deduced relations into perspective.

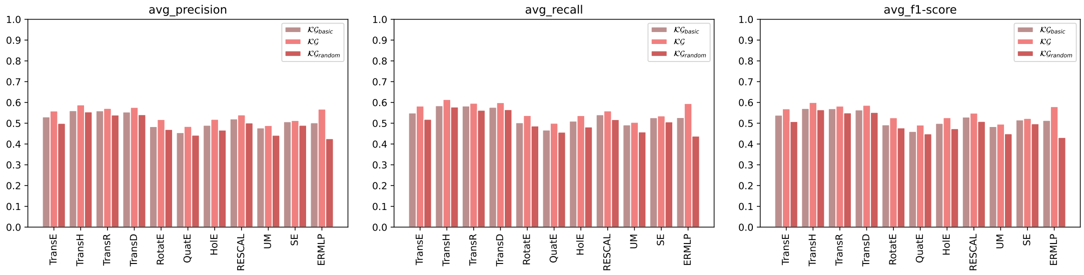


Figure 4.13: Evaluation of the Link Prediction task in terms of precision, recall, and f-measure. Utilizing Uniform negative sampling.

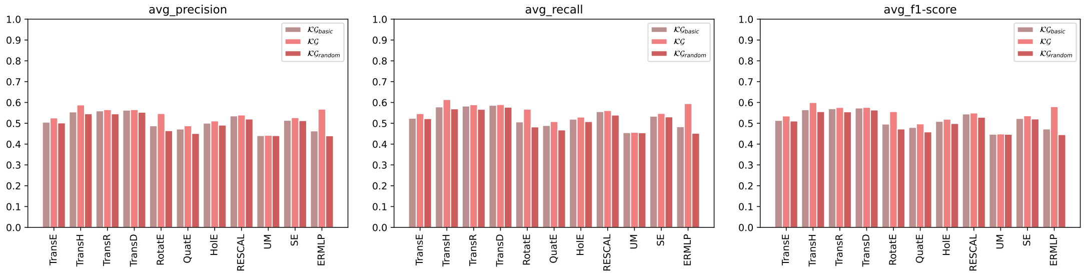


Figure 4.14: Evaluation of the Link Prediction task in terms of precision, recall, and f-measure. Utilizing Bernoulli negative sampling.

4.4.5 Discussion

The techniques proposed in this chapter rely on known relations between entities to predict novel links in the KG. During the experimental study, we observe that these techniques could improve the prediction of treatment effectiveness. Figure 4.15 shows a box plot of cosine similarity. Five treatments with a low-effect

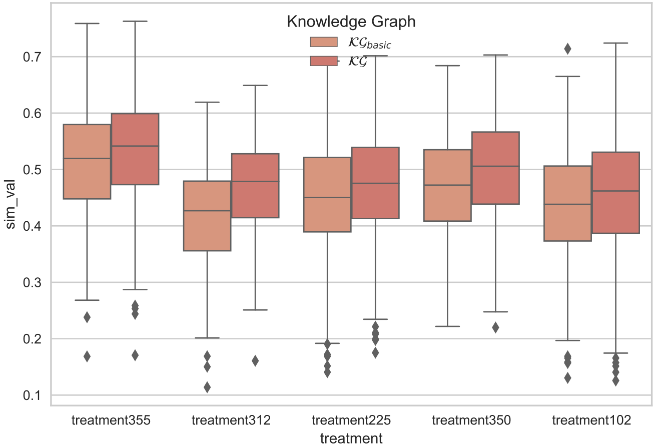


Figure 4.15: **Box-plot of Cosine Similarity**. The boxplot illustrates the distribution of cosine similarity values between treatments in x-axe with a list of treatments. We observe the five treatments in the x-axe are more similar to the treatments in $\mathcal{KG}$ than in $\mathcal{KG}_{basic}$.

response are selected, and $\mathcal{KG}_{basic}$ misclassify them, but $\mathcal{KG}$ predicts them correctly. Next, all the treatments with a low-effect response are selected. Thus, the cosine similarity is computed between the selected treatment and the list of treatments with the same response. We can observe that the five treatments are more similar to the list of treatments in $\mathcal{KG}$ than in $\mathcal{KG}_{basic}$. The first quartile, median, and third quartile values in the boxplot are higher in $\mathcal{KG}$ than in $\mathcal{KG}_{basic}$. Therefore, these outcomes put in evidence the quality of the deduced links in $\mathcal{KG}$ and their impact on the accuracy of the KGE models in the resolution of the task of predicting treatment effectiveness.

Figure 4.16 shows the distribution of DDIs by treatment in $\mathcal{KG}_{basic}$, $\mathcal{KG}$, and $\mathcal{KG}_{random}$. The x-axis represents the count of DDIs in treatment, and the y-axis represents the density of treatments in the KG with a specific x value. We utilized the Kernel Density Estimation (KDE) function to compute the probability density of the count of DDIs in each KG. We can observe for both treatment response *effective* and *low-effect* that $\mathcal{KG}$ have less density for treatments with five or fewer DDIs than the other two KGs and more density for treatments with more than five DDIs than the rest of the KGs. Furthermore, most treatments with *effective* response contain less than five DDIs while treatments with *low-effect* response contain more than five DDIs. These outcomes evidence the crucial role implicit DDIs have on a treatment's response and the need to deduce them using symbolic systems.

Analysis of deduced DDI by Treatment classes: Figure 4.17 exhibits the

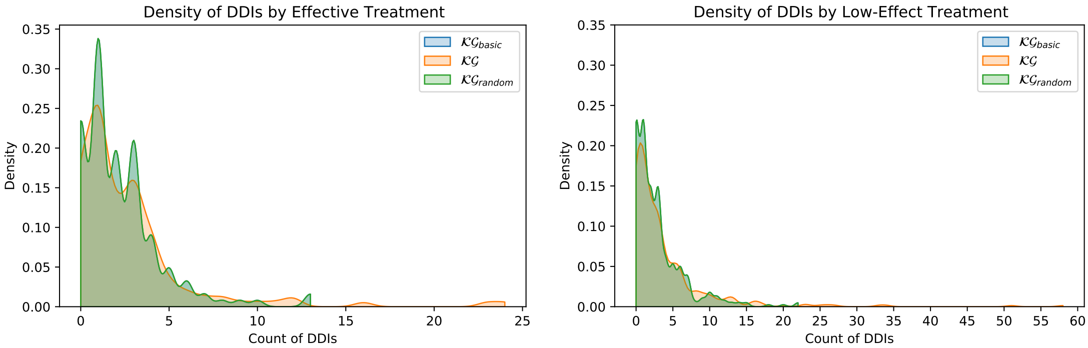


Figure 4.16: **The distribution of DDIs by treatment for each KG**. Figure 4.16a shows the density of treatments by DDIs for the treatment response *effective* in $\mathcal{KG}_{basic}$, $\mathcal{KG}$, and $\mathcal{KG}_{random}$. Figure 4.16b shows the density of treatments by DDIs for the treatment response *low-effect*.

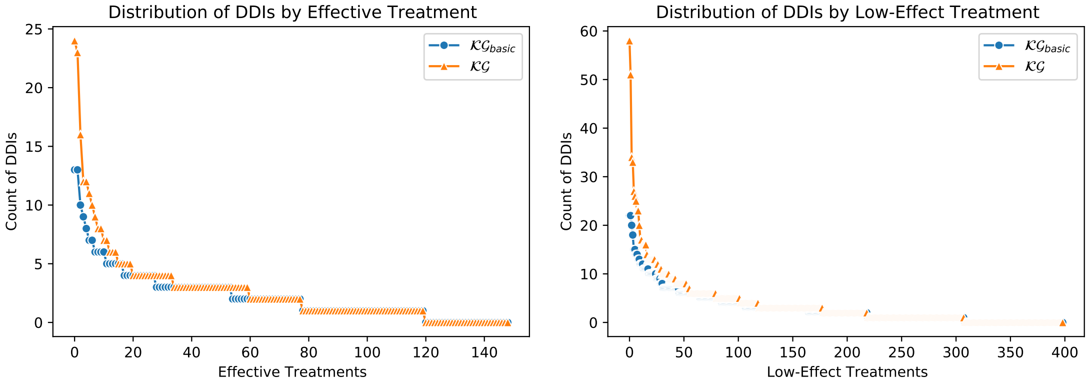


Figure 4.17: Distribution of DDIs by treatment response.

distribution of DDIs by treatment response in both $\mathcal{KG}_{basic}$ and $\mathcal{KG}$. The DDI Deductive System deduces new DDIs in 23.1% of treatments with *low-effect* responses while only 10.7% of treatments with *effective* responses deduce new DDIs. This analysis indicates that the DDI Deductive System deduces more than twice the number of DDIs in *low-effect* response treatments than in *effective* response treatments.

4.5 Summary

In this chapter, we present an approach based on integrating Neuro-Symbolic Artificial Intelligence systems and propose a deductive database over a knowledge graph based on the existing approaches of deductive database systems. Knowledge graphs enable the description of the meaning of data, the integration of data from heterogeneous sources, and the discovery of unknown patterns. However, they are limited by the data sparsity issues for knowledge discovery. The proposed deductive database makes implicit knowledge explicit and machine-readable. Our proposed solution builds ego networks of an abstract target prediction to deduce new relationships and enhances the ego networks. Thus, the deductive database reduces the data sparsity issue, enabling the knowledge graph to become meaningful in the discovery task. The proposed neuro-symbolic AI system integrates the deductive database with KGE models, which benefit from the symbolic system as it assists in overcoming data sparsity issues. We assess the performance of our approach in a KG for lung cancer to discover treatment effectiveness. The deductive system improves existing embedding models by performing the treatment prediction task. Results of a 5-fold cross-validation process demonstrate that our Neuro-Symbolic AI approach improves the state-of-the-art KGE models evaluated. Overall with the proposed approach, knowledge represented semantically in knowledge graphs can be exploited to solve a broad range of problems.

Chapter 5

Discover Relations across Industry 4.0 Standards with a Neuro-Symbolic AI System

The main objective of the fourth industrial revolution, Industry 4.0 (I4.0), is the creation of *smart factories* by combining the Internet of Things (IoT), Internet of Services (IoS), and Cyber-Physical Systems (CPS). In smart factories, humans, machines, materials, and CPS cooperate intelligently to produce individualized products. This cooperation requires effective communication and the resolution of interoperability issues generated whenever the same products are described with different standards. Different industrial communities have defined standardization frameworks aligning standards according to their features and expressiveness. Relevant examples are the Reference Architecture for Industry 4.0 (RAMI4.0) [2] or the Industrial Internet Connectivity Framework (IICF) in the US [77]. Despite the capacity to categorize existing standards, standardization frameworks may present divergent interpretations of the same standard. Mismatches among standard classifications generate semantic interoperability conflicts that negatively impact communication effectiveness in smart factories.

Database and Semantic Web communities have extensively studied the problem of data integration [42, 70, 86], and various approaches have been proposed to support data-driven pipelines to transform industrial data into actionable knowledge in smart factories [9, 57, 92]. Ontology-based approaches have also contributed to creating a shared understanding of the domain [74], specifically Kovalenko and Euzenat [70] have equipped data integration with diverse methods for ontology alignment. Furthermore, Lin *et al.* [76] identify interoperability conflicts across domain-specific standards (e.g., RAMI4.0 model and the IICF architecture), while works by Grangel-Gonzalez *et al.* [43, 44, 45] show the relevant role that Descriptive Logic and Datalog play in liaising I4.0 standards. Certainly, the extensive literature in data integration provides the foundations for enabling the semantic description and alignment of "similar" things in a smart factory. Nevertheless, finding alignments across I4.0 requires encoding domain-specific knowledge represented in standards of diverse nature and standardization frameworks defined with different industrial goals. We rely on state-of-the-art knowledge representation and discovery approaches to embedding meaningful associations and features of the I4.0 landscape to enable interoperability.

In this chapter, we address the problem of determining relatedness across I4.0 standards described in terms of their main features and standardization frameworks. Our goal is to uncover alignments among related standards, i.e., standards that define the same type of smart factory components. Moreover, we aim to provide a precise classification of the standards and contribute to a more precise categorization in the standardization frameworks. Figure 5.1 shows the challenges we tackle in this chapter and the contributions to address the challenges. The research work presented in this chapter is based on the publications [105, 107]. This chapter addresses the following research questions:

RQ1: How can metadata encoding data meaning be exploited to discover relationships in knowledge graphs?

RQ3: How can implicit knowledge be used to enhance knowledge discovery tasks?

To answer the research questions **RQ1** and **RQ3**, we propose a knowledge-driven approach able to integrate standards and standardization frameworks into a knowledge graph for discovering relations. The Neuro-Symbolic AI approach presented in Chapter 4 is employed to discover relationships between standards in the KG. We present a symbolic system implemented by a deductive database to enhance the performance of knowledge graph embedding models. The features of the standards represented in a knowledge graph are exploited to build latent representations in a low-dimensional space, i.e., embeddings. Values of similarity metrics between embeddings are used in conjunction with state-of-the-art community detection algorithms to identify patterns among standards. Then, the *homophily* prediction principle is performed in each community to discover new links between standards and frameworks. The observed results demonstrate the benefits of exploiting knowledge graphs for the computation of alignments across standards. These outcomes provide evidence of the accuracy of the uncovered patterns and the discovered relations The main contributions of this chapter are:

- A formalization of the problem of finding relations among I4.0 standards. It presents $I4.0\mathcal{RD}$, a knowledge-driven approach to unveil these relations.

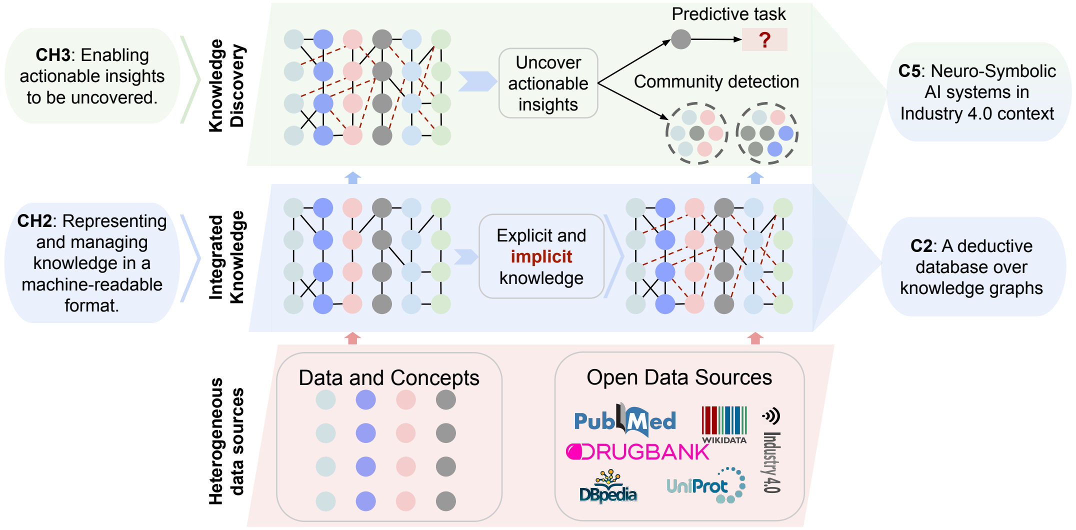

CH3: Enabling actionable insights to be uncovered.
Knowledge Discovery
Uncover actionable insights
Predictive task
?
Community detection
C5: Neuro-Symbolic AI systems in Industry 4.0 context
CH2: Representing and managing knowledge in a machine-readable format.
Integrated Knowledge
Explicit and implicit knowledge
C2: A deductive database over knowledge graphs
Heterogeneous data sources
Data and Concepts
Open Data Sources
PubMed
WIKIDATA
DRUGBANK
DBpedia
UniProt
Industry 4.0

Figure 5.1: **Challenges and contributions**. This chapter focuses on discovering relationships in a knowledge graph and proposes an approach based on the integration of Neuro-Symbolic AI systems to solve the problem.

I4.0RD exploits the semantic description encoded in a knowledge graph via a symbolic system and the creation of embeddings to identify the communities of standards that should be related.

- An extensive evaluation of *I4.0RD* in different embeddings learning models, similarity measures, and community detection algorithms. The evaluation material is available at `https://github.com/i40-Tools/I40KG-Embeddings`.

This chapter is structured as follows: Section 5.1 motivates the work presented in this chapter by illustrating the interoperability problem presented in Industry 4.0. Section 5.2 formalizes the problem statement and proposed solution addressed in this chapter. Section 5.3 describes our *I4.0RD* architecture that is proposed to address the problem presented in Section 5.2. Section 5.4 presents an empirical evaluation of *I4.0RD* and an analysis of the obtained results. The observed results reveal the benefits of *I4.0RD* for the computation of alignments across standards. Finally, concluding remarks for this chapter are presented in Section 5.5.

5.1 Motivating Example

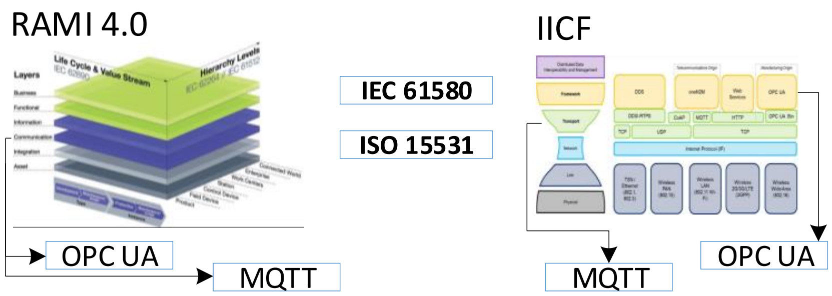


Figure 5.2: **Motivating Example**. The RAMI4.0 and IICF standardization frameworks are developed for diverse industrial goals; they classify standards in layers according to their functions, e.g., OPC UA and MQTT under the communication layer in RAMI4.0, and OPC UA and MQTT in the framework and transport layers in IICF, respectively. Further, some standards, e.g., IEC 61580 and ISO 15531, are not classified yet.

Existing efforts to achieve interoperability in I4.0 mainly focus on the definition of standardization frameworks. A standardization framework defines different layers of group-related I4.0 standards based on their functions and main characteristics. Typically, classifying existing standards in a certain layer is not a trivial task, and it is biased by the point of view of the community that developed the framework. RAMI4.0 and IICF are exemplar frameworks. The former is developed in Germany while the latter is in the US; they meet specific I4.0 requirements of certain locations around the globe. RAMI4.0 classifies OPC UA and MQTT standards into the Communication layer, stating that both standards are similar. Contrary, IICF presents OPC UA and MQTT at distinct layers, i.e., the framework and the transport layers, respectively. Furthermore, independently of the classification of the standards made by standardization frameworks, standards have relations based on their functions. Therefore, IEC 61580 is an international standard defining communication protocol for intelligent electronic devices, and ISO 15531 is a standard for industrial automation systems. Both standards are not classified at all. Figure 5.2 depicts these relations across the frameworks RAMI4.0 and IICF and the standards; it illustrates interoperability issues in the I4.0 landscape.

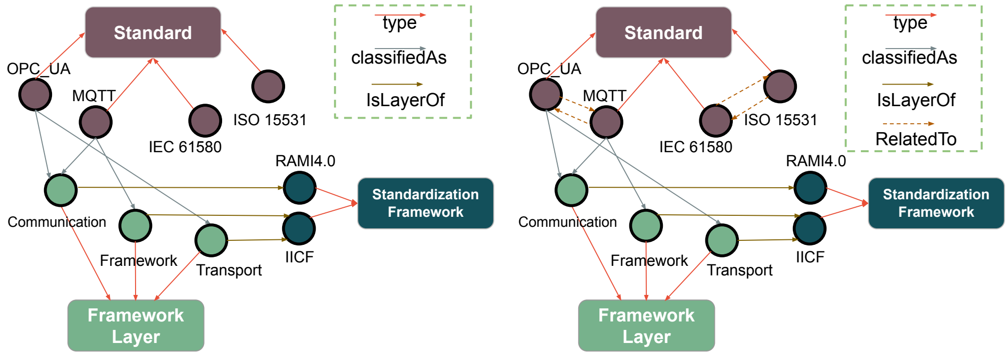
Existing data integration approaches rely on the description of the characteristics of entities to solve interoperability by discovering alignments among them. Specifically, in the context of I4.0, semantic-based approaches have been proposed to represent standards, known relations among them, as well as their classification according to existing frameworks [10, 26, 76, 79]. Despite the information, the structured modeling of the I4.0 landscape only provides the foundations for detecting interoperability issues. We propose *I4.0RD*, an approach capable of discovering relations over I4.0 knowledge graphs to identify unknown relations among standards. Our proposed methods exploit relations represented in an I4.0 knowledge graph to compute the similarity of the modeled standards. Then, an unsupervised graph partitioning method determines similar communities of similar standards. *I4.0RD* explores communities to identify possible relations of standards, thus enhancing interoperability.

5.2 Problem Statement and Proposed Solution

In this chapter, we tackle the problem of unveiling relations between I4.0 standards. Relations among standards and standardization frameworks (e.g., in Figure 5.3a) are represented in a knowledge graph named I4.0KG. To populate the I40KG, Grangel-Gonzalez, Bader *et al.* [9] has surveyed and analyzed the standards landscape from a semantic perspective, and the resulting I40KG represents knowledge expressed in over 200 industry-related documents, including technical reports and research papers. Nodes in the I4.0KG correspond to standards and frameworks; edges represent relations among standards, as well as the standards group in a framework layer. An I4.0KG is defined as follows:

Given sets V_e and V_t of entities and types, respectively, a set E of labeled edges representing relations and a set L of labels. An I.40KG is defined as $\mathcal{G} = (V_e \cup V_t, E, L, I, D, R, ego, \mathcal{N}, \alpha)$:

- The types Standard, Frameworks, and Framework Layer belong to V_t.
- I4.0 standards, frameworks, and layers are represented as instances of V_e.
- The types of the entities in V_e are represented as edges in E that belong to $V_e \times V_t$.
- Edges in E that belong to $V_e \times V_e$ represent relations between standards and their classifications into layers according to a framework.
- Properties *relatedTo*, *Type*, *classifiedAs*, *isLayerOf* correspond to labels in L that represent the relations between standards, their type, their classification into layers, and the layers of a framework, respectively.
- $I : V_e \rightarrow V_t$ maps each entity to a class.
- $D : L \rightarrow C$ maps a property to a domain class.



(a) Actual I4.0 KG

(b) Ideal I4.0 KG

Figure 5.3: **Example of I4.0KGs**. (a) shows known relationships among standards to Framework Layer and Standardization Framework. (b) depicts all the ideal relationships between the standards expressed with the property *relatedTo*. Standards OPC UA and MQTT are related, as well as the standards IEC 61968 and IEC 61400. Our aim is discovering relations *relatedTo* in (b).

- $R : L \to C$ maps each property to a class range.
- $ego : V_e \to 2^{V_e \times L \times V_e}$, the function $ego(.)$ represents ego networks in the knowledge graph. $ego(v)$ assigns to each concept in V_e the set of labeled edges where v is in the subject or object position.
- $ego(v) = \{(u_1, r, u_2) | (u_1, r, u_2) \in E \wedge (u_1 = v \vee u_2 = v)\}$. The $ego(v)$ defines the ego network of the entity v.
- $\mathcal{N}(v) = \{v_i | (v, r, v_i) \in E \vee (v_i, r, v) \in E\}$. The function $\mathcal{N}(v)$ defines the neighbors of the entity v. $\mathcal{N}(v)$ assigns to each concept in V_e the set of concepts $\mathcal{N}(v)$, where v and each element of $\mathcal{N}(v)$ are in the subject or object position of E.
- $\alpha : 2^{V_e} \to 2^{V_e \times L \times V_e}$. The function $\alpha(.)$ returns a set of edges between the pairs of elements in the input, where 2^{V_e} represents the power set of entities in V_e.
- $\alpha(T) = \{(v_1, r, v_2) | (v_1, r, v_2) \in E \wedge v_1 \in T \wedge v_2 \in T\}$. The function $\alpha(.)$ returns the set of edges between pairs of entities in the input set T.

An ideal knowledge graph: Let $\mathcal{G}' = (V_e \cup V_t, E', L, I, D, R, ego, \mathcal{N}, \alpha)$ be an *ideal* I4.0 knowledge graph that contains all the *existing relations* between standard entities and frameworks in V_e, i.e., an oracle that knows whether two standard entities are related or not, and to which layer they should belong. Figure 5.3b illustrates a portion of an ideal I4.0KG, where the relations between standards are explicitly represented. $\mathcal{G}'$ assumes the CWA, i.e., what is unknown to be true must be false.

An Actual knowledge graph: Let $\mathcal{G} = (V_e \cup V_t, E, L, I, D, R, ego, \mathcal{N}, \alpha)$ be the *actual* I4.0KG, e.g., in Figure 5.3a, that follows the assumption OWA, i.e., what is not known to be true is just unknown.

A complete knowledge graph: Let $\mathcal{G}_{\text{comp}} = (V_e \cup V_t, E_{\text{comp}}, L, I, D, R, ego, \mathcal{N}, \alpha)$ be a *complete* knowledge graph which includes a relation for each possible combination of elements in V_e and labels in L, i.e., $E \subseteq E' \subseteq E_{\text{comp}}$, where not all relationships are necessarily true.

$\mathcal{G}$ only contains a portion of the relations represented in $\mathcal{G}'$, i.e., $E \subseteq E'$; it represents those relations that are known and are not necessarily complete. Let $\Delta(E', E) = E' - E$ be the set of relations existing in the ideal knowledge graph $\mathcal{G}'$ that is not represented in $\mathcal{G}$.

5.2.1 Problem Statement

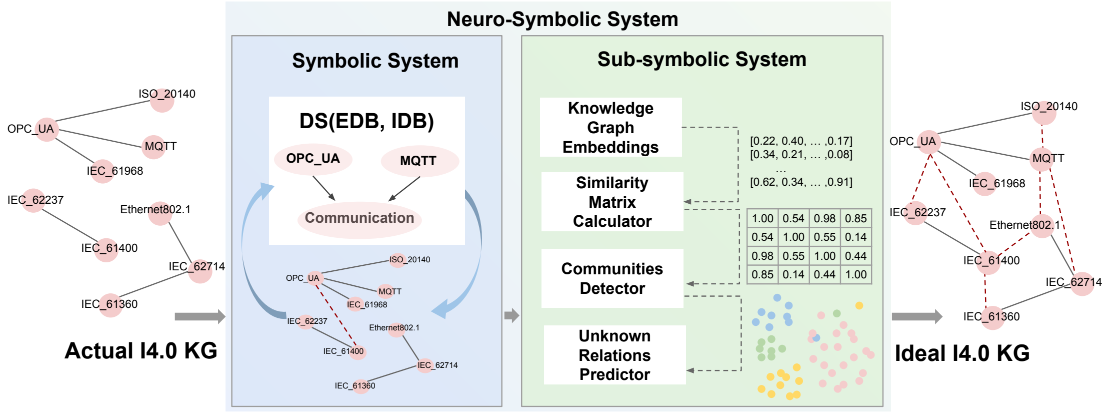
Given a relation $e \in \Delta(E_{\text{comp}}, E)$, the problem of discovering relations consists of determining whether $e \in E'$, i.e., if a relation represented by an edge $r{=}(e_i\ l\ e_j)$ corresponds to an existing relation in the ideal knowledge graph $\mathcal{G}'$. Specifically, we focus on the problem of discovering *relations* between standards in $\mathcal{G} = (V_e \cup V_t, E, L, I, D, R, ego, \mathcal{N}, \alpha)$. We are interested in finding the maximal set of relationships or edges E_a that belongs to the ideal I4.0KG, i.e., find a set E_a that corresponds to a solution of the following optimization problem:

$$\underset{E_a \subseteq E_{comp}}{\operatorname{argmax}} |E_a \cap E'|$$

Considering the knowledge graphs depicted in Figures 5.3a and Figures 5.3b, the problem addressed in this work corresponds to the identification of edges in the ideal knowledge graph that corresponds to unknown relations between standards.

5.2.2 Proposed Solution

We present a neuro-symbolic AI approach to discover relationships between standards in Industry 4.0 KG. We propose a symbolic system implemented by a deductive database to enhance the performance of knowledge graph embedding models. The EDB of the deductive database DS contains ground facts of the ego networks $ego(v)$, where $I(v) \in \{C_1, C_2\}, C_1 \in D(r), C_2 \in R(r)$, and $r = relatedTo$. The variables C_1 and C_2 represent the domain and range of the property r, respectively. The IDB comprises rules for deducing relations in $ego(v)$. The



1.00	0.54	0.98	0.85
0.54	1.00	0.55	0.14
0.98	0.55	1.00	0.44
0.85	0.14	0.44	1.00

Figure 5.4: **The *I4.0RD* Architecture.** *I4.0RD* receives the actual I4.0 KG and outputs an extended version of the I4.0KG including novel relations. The symbolic system is implemented by a deductive system $DS(EDB, IDB)$ that deduces new relationships in the ego networks $ego(v)$. Then, the sub-symbolic system implemented by a KGE model receives the I4.0KG enhanced by the symbolic system. Trans* family of models creates embeddings for each standard, and similarity values between embeddings are computed; these values are used to partition standards into communities. Finally, the homophily prediction principle is applied to each community to discover unknown relations. A knowledge graph closer to the ideal I4.0 KG is generated.

sub-symbolic systems implemented by KGE models learn a latent representation of entities and relations and exploit relations represented in an I4.0KG. Further, an unsupervised graph partitioning method determines the parts of the I4.0KG or communities of standards that are similar. Then, the *homophily* prediction principle is applied in a way that similar standards in a community are considered to be related.

5.3 The *I4.0RD* Architecture

We call *I4.0RD* the proposed architecture capable of discovering relations over I4.0 knowledge graphs to identify unknown relations among standards. Figure 5.4 presents *I4.0RD*, a pipeline that implements the proposed approach. *I4.0RD* receives an I4.0KG $\mathcal{G}$ and returns an I4.0KG $\mathcal{G}'$ that corresponds to a solution to the problem of discovering relations between standards. First, the symbolic system DS deduces new relationships in the $ego(v)$ minimizing the data sparsity issues in the I4.0KG. Second, in order to compute the similarity values between the entities an I4.0KG, the sub-symbolic system learns a latent representation of the standards in a low-dimensional space. Our approach resorts to the Trans* family of KGE models to compute the embeddings of the standards. Then, a distance metric for vector spaces is applied to compute the values of similarity between standards. Next, community detection algorithms are applied to identify communities of related standards. METIS [66], KMeans [7], and SemEP [95] are methods included in the pipeline to produce different communities of standards. Finally, *I4.0RD* applies the *homophily* principle to each community to predict relations or alignments among standards.

5.3.1 Symbolic System in I4.0KG. Deductive Database

The EDB of our deductive database system DS corresponds to statements in the ego networks $ego(v)$, where $I(v) \in \{C_1, C_2\}, C_1 \in D(r), C_2 \in R(r)$, $r = relatedTo$, and C_1 and C_2 represent the domain and range of the property r, respectively. The IDB contains a set of rules to deduce the implicit relationships between standards in the ego networks. The relation *relatedTo* is extracted from the literature and represents a relation that connects two standards. Beside *relatedTo* is an equivalence relation that satisfies three properties, i.e., the relation is reflexive, symmetric, and transitive. They are defined as follows:

- Reflexive: $\forall e_i \in V_e(e_i, relatedTo, e_i)$
- Symmetric: $\forall e_i, e_j \in V_e((e_i, relatedTo, e_j) \Leftrightarrow (e_j, relatedTo, e_i))$
- Transitive: $\forall e_i, e_j, e_k \in V_e : ((e_i, relatedTo, e_j) \wedge (e_j, relatedTo, e_k)) \Rightarrow (e_i, relatedTo, e_k)$

An example of the transitivity property of *relatedTo* is presented with the following three standards: *IEC 61310 P3 E2; IEC 61310 P1 E2; IEC 61310 P2 E2.* Those standards describe electrical features and entitle machinery safety - Indication, marking, and actuation. From the literature the next relations are known: *(IEC 61310 P3 E2, relatedTo, IEC 61310 P1 E2)* ∧ *(IE 61310 P1 E2, relatedTo, IEC 61310 P2 E2)* and that implies: *(IEC 61310 P3 E2, relatedTo, IEC 61310 P2 E2).* Since the property *relatedTo* between standards is an equivalent relation, the transitive closure of the relations is materialized in I4.0KG. Thus, we can capture implicit relations between I4.0 standards. The following Datalog rules state when the property *relatedTo* relates pairs of standards.

$$relatedTO(A,\ B) \Rightarrow inferred\_relatedTO(A,\ B). \tag{5}$$

$$type(A,\ Standard) \Rightarrow inferred\_relatedTO(A,A). \tag{6}$$

$$relatedTO(B,\ A) \Rightarrow inferred\_relatedTO(A,\ B). \tag{7}$$

$$inferred\_relatedTO(A,\ B), relatedTO(B,\ C) \Rightarrow inferred\_relatedTO(A,\ C). \tag{8}$$

$$classifiedAs(A,\ C), classifiedAs(B,\ C) \Rightarrow inferred\_relatedTO(A,\ B). \tag{9}$$

Rule (5) states the base case of the IDB, where the predicate *inferred_ relatedTO* contains the facts of the predicate *relatedTO*, variables A and B are related standards. Rule (6) states the reflexive property of *relatedTo*, where the variable A is of type *Standard.* Rule (7) defines the symmetric property of *relatedTo*, and Rule (8) declares the transitive property of *relatedTo*, where the variables A, B, and C are of type *Standard.* Rule (9) states that standards belonging to the same class are related by the *relatedTo* property. Figure 5.5 shows the relation *relatedTo* before applying the deductive system in I4.0KG (cf. Figure 5.5a) and after being applied Figure 5.5b. Figure 5.5b illustrates how the I4.0 standards knowledge graph is more connected after the symbolic system DS deduces new relationships. The graphs are plotted using Cytoscape[1] .

5.3.2 Sub-Symbolic System in I4.0KG

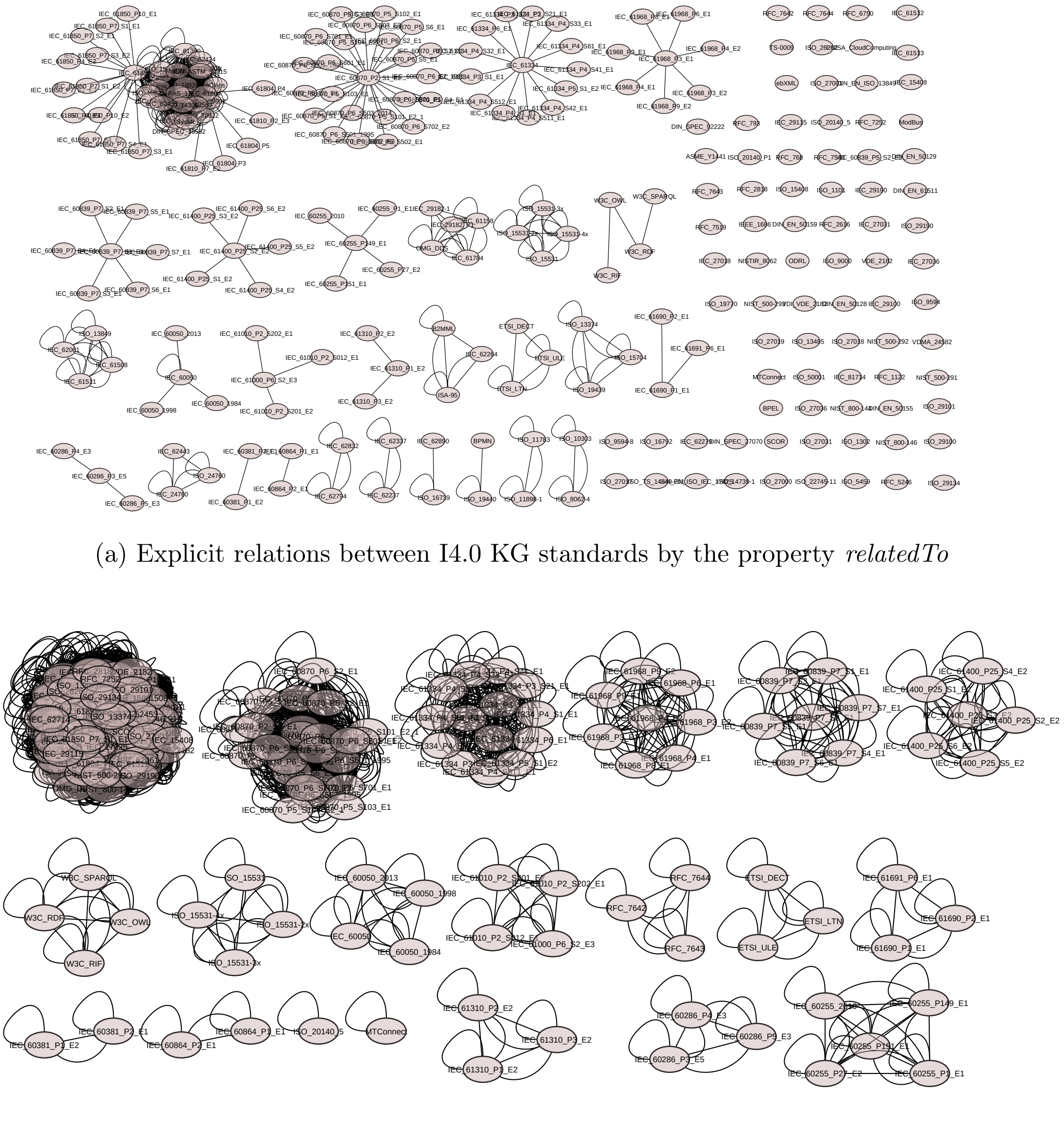
Learning Latent Representations of Standards. *I4.0RD* utilizes the Trans* family of models to compute latent representations, e.g., vectors, of entities and relations in an I4.0 knowledge graph. In particular, *I4.0RD* utilizes TransE, TransD, TransH, and TransR. These models differ in the representation of the embeddings for the entities and relations (Wang et al. [136]). Suppose e_i, e_j, and p denote the vectorial representation of two entities related by the labeled edge p in an I4.0 knowledge graph. Furthermore, $\|x\|_2$ represents the Euclidean norm.

TransE, TransH, and TranR represent the entity embeddings as $(e_i, e_j \in \mathbb{R}^d)$, while TransD characterizes the entity embeddings as $(e_i, w_{e_i} \in \mathbb{R}^d - e_i, w_{e_j} \in \mathbb{R}^d)$. As a consequence of different embedding representations, the scoring function also varies. For example, TransE is defined in terms of the score function $\|e_i + p - e_j\|_2^2$, while $\|M_p e_i + p - M_p e_j\|_2^2$ defines TransR[2]. Furthermore, TransH score function corresponds to $\|e_{i\perp} + d_p - e_{j\perp}\|_2^2$, where the variables $e_{i\perp}$ and $e_{j\perp}$ denote a projection to the hyperplane w_p of the labeled relation p, and d_p is the vector of a relation-specific translation in the hyperplane w_p. To learn the embeddings, *I4.0RD* resorts to the PyKeen (Python KnowlEdge EmbeddiNgs) framework [4]. As hyperparameters for the models of the Trans* family, we use the ones specified in the original papers of the models. The hyperparameters include embedding dimension (set to 50), number of epochs (set to 500), batch size (set to 64), seed (set to 0), learning rate (set to 0.01), scoring function (set to 1 for TransE, and 2 for the rest), margin loss (set to 1 for TransE and 0.05 for the rest). All the configuration classes and hyperparameters are open in GitHub [3].

[1] `https://cytoscape.org/`

[2] M_p corresponds to a projection matrix $M_p \in \mathbb{R}^{dxk}$ that projects entities from the entity space to the relation space; further $p \in \mathbb{R}^k$.

[3] `https://github.com/i40-Tools/I4.0KG-Embeddings`



(a) Explicit relations between I4.0 KG standards by the property *relatedTo*

(b) Explicit and implicit relations between I4.0 KG standards by the property *relatedTo*

Figure 5.5: **Relations between I4.0 KG standards**. (a) Using explicit relations between standards in I4.0 KG, 109 connected components are found. (b) Applying the symbolic system *DS* for the property *relatedTo*, 20 connected components are found, 89 less than in (a). Standards in I4.0 KG are more connected, and new relations in the connected components correspond to meaningful relations.

Computing Similarity Values Between Standards. Once the algorithm–Trans* family–that computes the embeddings reaches a termination condition, e.g., the maximum number of epochs, the I4.0KG embeddings are learned. As the next step, *I4.0RD* calculates a *similarity symmetric matrix* between the embeddings that represent the I4.0 standards. Any distance metric for vector spaces can be utilized to calculate this value. However, as a proof of concept, *I4.0RD* applies the Cosine Similarity and the Inverse Euclidean Distance. Let u be an embedding of the Standard-A and v an embedding of the Standard-B; the similarity score between both standards is defined by Cosine Similarity[4] as follows:

$$cosine(u, v) = \frac{u.v}{||u||_2||v||_2}$$

The Inverse Euclidean Distance[5] between the vectors u and v, is defined as follows:

$$d(u, v) = 1 - ||u - v||_2$$

After building the *similarity symmetric matrix*, *I4.0RD* applies a threshold to restrict the similarity values. *I4.0RD* relies on percentiles to calculate the value of such a threshold. Further, *I4.0RD* utilizes the function Kernel Density Estimation (KDE) to compute the density of both similarity measures, Cosine Similarity and Inverse Euclidean Distance; it sets to zero the similarity values lower than the given threshold.

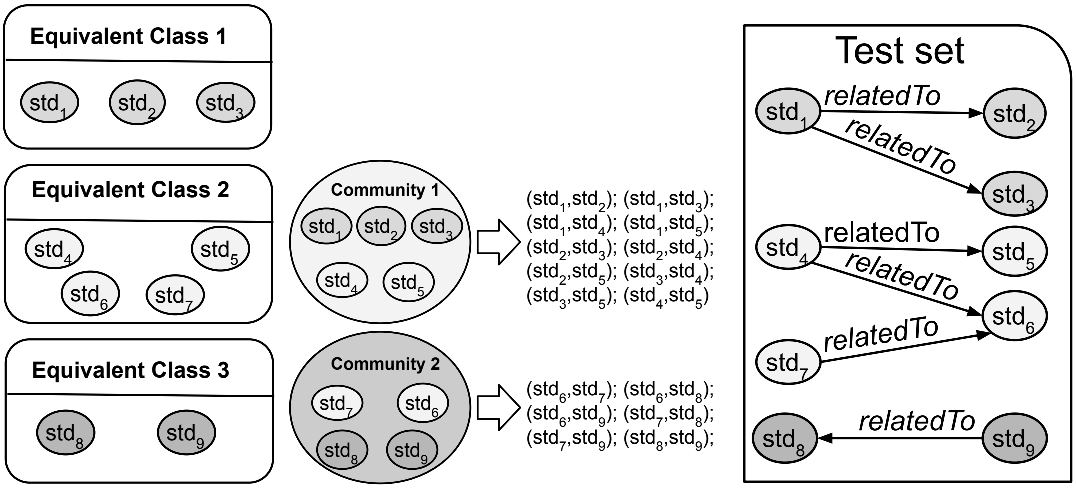
Detecting Communities of Standards. *I4.0RD* maps the problem of computing groups of potentially related standards to the problem of community detection. Once the embeddings are learned, the standards are represented in a vectorial way according to their functions, preserving their semantic characteristics. Using the embeddings, *I4.0RD* computes the similarity between the standards in the I4.0 KG, as mentioned in the previous section. The values of similarity between standards are utilized to partition the set of standards in a way that standards in a community are highly similar but dissimilar to the standards in other communities. As proof of concept, three state-of-the-art community detection algorithms have been used in *I4.0RD*: SemEP, METIS, and KMeans. They implement diverse strategies for partitioning a set based on the values of similarity, and our goal is to evaluate which of the three is more suitable to identify meaningful connections between standards.

Discovering Relations Between Standards. New relations between standards are discovered in this step; the *homophily* prediction principle is applied

[4] https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.cosine.html

[5] https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.euclidean.html

over each of the communities, and all the standards in a community are assumed to be related. Figure 5.6 depicts an example where new relations are computed from two communities; unknown relations correspond to connections between standards in a community that does not exist in the input I4.0KG ($\mathcal{G}$). Figure 5.6a shows the equivalent classes of the I4.0KG example. Community 1 has five standards where three of them belong to Equivalent Class 1, and the other two belong to Equivalent Class 2. Applying the homophily prediction principle to Community 1, six new relations are found between standards from Equivalent Class 1 and Equivalent Class 2; these are $(std_1, std_4), (std_2, std_4), (std_3, std_4), (std_1, std_5), (std_2, std_5), (std_3, std_5)$. These new relations are evaluated by experts to proof that they correspond to meaningful relations.



(a) Equivalent classes induced by the property *relatedTo*

(b) Application of the Homophily Prediction Principle

(c) Known Relations used to determine discovered relations between standards

Figure 5.6: **Discovering relations between standards**. (a) The homophily prediction principle is applied to two communities. As a result, 16 relations between standards are found. (b) Five out of the 16 found relations correspond to meaningful relations.

5.4 Experimental Evaluation

We use the equivalent classes induced by the property *relatedTo* as a baseline. An equivalent class is induced by equivalent relations like *relatedTo* that satisfy three properties, i.e., the relation is reflexive, symmetric, and transitive. The equivalent classes are partitions of the set of standards induced by the relation *relatedTo*. Figure 5.7 shows the number of partitions and how many standards each partition of our baseline has. Equivalent Class 1 has the highest number of standards, with 148. All the standards in each equivalent class are related to each other but isolated from the other equivalent classes. Assuming that the different combinations of similarity measures and the community detection algorithms are effective predictors of the standards communities, the distances between the equivalent classes and the communities discovered should be close. The Average Category-based Score measure assesses the distance between Communities and the baseline.

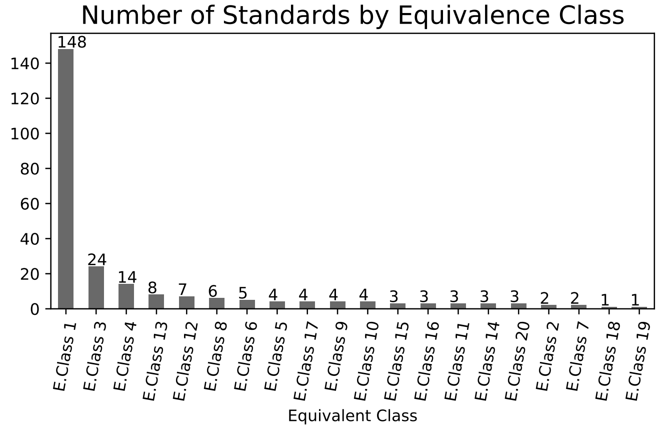


Figure 5.7: **Baseline of Equivalent Classes.** I4.0KG has 20 Equivalent Classes, and most of them have less than 10 standards except the Equivalent Classes 1, 2, and 5.

We report on the impact that the knowledge encoded in the I4.0 knowledge graph has in the behavior of *I4.0RD*. In particular, we assess the following research questions:

Q1) How the function used to determine the relatedness between standards impact the outcome of the problem of uncovering relations among standards?

Q2) Does a semantic community-based analysis on I4.0KG allow for improving the quality of predicting new relations on the I4.0 standards landscape?

Q3) What is the effect of combining distinct similarity measures, embedding techniques, and community detection algorithms in the task of detecting the relatedness among standards?

Experiment Setup: Four embedding algorithms are considered to build the standards embedding. Each of these algorithms is evaluated independently. Next, a similarity matrix for the standards embedding is computed. Cosine Similarity and Inverse Euclidean Distance are considered similarity measures. The similarity matrix is required for applying the community detection algorithms. In our experiments, three algorithms are used to compute the Communities. In overall, we evaluate twenty-four combinations between embedding algorithms, similarity measures, and community detection algorithms. To assure statistical robustness, we execute 5-fold cross-validation with one run. For the purposes of understanding how the Trans* methods, similarity measures, and community detection algorithms are performing, we evaluate the similarity density of the standards by Trans* methods, also the quality of the generated Communities, the accuracy of the Communities in discovering new relationships and the distance between the Communities and the baseline using Cosine and Inverse Euclidean Distance.

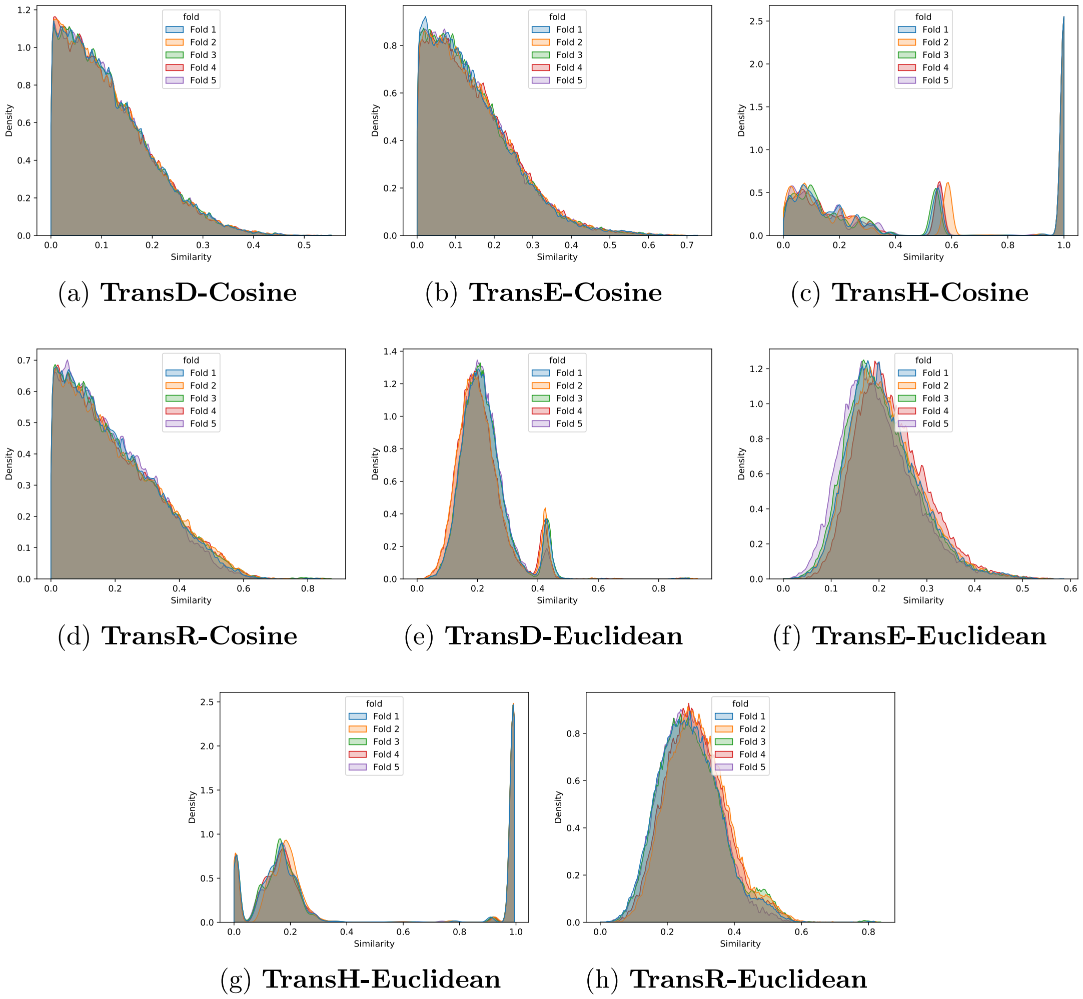
Implementation: Our proposed approach is implemented in Python 2.7 and integrated with the PyKeen (Python KnowlEdge EmbeddiNgs) framework [4], METIS 5.1 [6], SemEP [7], and Kmeans [8]. The experiments were executed on a GPU server with ten chips Intel(R) Xeon(R) CPU E5-2660, two chips GeForce GTX 108, and 100 GB RAM.

Thresholds for Computing Values of Similarity. Figure 5.8 depicts the density function of each fold for each embedding algorithm using the similarity metrics Cosine Similarity and Inverse Euclidean Distance. We notice that Inverse Euclidean Distance finds a higher density of similar standards than the Cosine Similarity metric in all Trans* methods. Figures 5.8a and 5.8b show the values of the folds of TransD and TransE in Cosine Similarity, where all the similarity values are close to 0.0, i.e., all the standards are different. Figure 5.8d suggests that all the folds have similar behavior with values between 0.0 and 0.5 and a short group of standards with similarity values in 0.8. Figure 5.8c and Figure 5.8g show a group of standards similar with values close to 1.0 and the rest of the standards between 0.0 and 0.4. The percentile of the similarity matrix is computed with a threshold of 0.85. That means all values of the similarity matrix, which are less than the percentile computed, are filled with 0.0, and then, these two standards are dissimilar. After analyzing the density of each fold (cf. Figure 5.8), the thresholds of TransH and TransR using Cosine Similarity are set to 0.50 and 0.75, respectively. The reason is that the two cases with a high threshold find all similar standards, and creating more than one Community of standards will not be possible. The thresholds of the similarity matrix using Inverse Euclidean Distance are also modified for the same reason. TransD, TransH, and TransR are set to 0.95, 0.60, and 0.75, respectively. In the case of TransH, there is a high density of

[6] http://glaros.dtc.umn.edu/gkhome/metis/metis/download

[7] https://github.com/SDM-TIB/semEP

[8] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html



(a) **TransD-Cosine** (b) **TransE-Cosine** (c) **TransH-Cosine**

(d) **TransR-Cosine** (e) **TransD-Euclidean** (f) **TransE-Euclidean**

(g) **TransH-Euclidean** (h) **TransR-Euclidean**

Figure 5.8: **Similarity density by Cosine and Inverse Euclidean Distance of each fold per Trans* methods**. Results from the Inverse Euclidean Distance in all the Trans* methods have higher similarity values than Cosine similarity. Figures 5.8a, 5.8b, and 5.8d show that all folds have values close to zero, i.e., with embeddings created by TransD, TransE, and TransR the standards are very different from each other. However, TransH in both similarity measures (cf. Figure 5.8c and Figure 5.8g) exploits properties of the standards and generates embeddings with a different distribution of similarity, i.e., values between 0.0 and 0.4, as well as values close to 1.0. According to known characteristics of the I4.0 standards, the TransH distribution of similarity using both Cosine Similarity and Inverse Euclidean Distance better represents their relatedness.

values close to 1.0; it indicates that for a threshold of 0.85, the percentile computed is almost 1.0. the values of the similarity matrix less than the threshold are filled with 0.0; values of 0.0 represent that the compared standards are not similar.

5.4.1 Impact of Metrics for Determining Relatedness among Standards

There are a variety of metrics to evaluate the quality of clusters. We used five recognized cluster metrics to estimate the quality of the communities from the I4.0KG embeddings. All the metrics are normalized in the range [0,1] where higher is a better score.

a) **Conductance (InvC)**: measures relatedness of entities in a community and how different they are to entities outside the community [37]. The inverse of Conductance is reported: $1 - Conductance(K)$, where $K = \{k_1, k_2, ..., k_n\}$ the set of standards communities obtained by the clustering algorithm, and k_i are the computed clusters.

b) **Performance (P)**: sums up the number of intra-community relationships, plus the number of non-existent relationships between communities [37]. Higher values indicate that a cluster is both internally dense and externally sparse.

c) **Total Cut (InvTC)**: sums up all similarities among entities in different communities [21]. The Total Cut values are normalized by dividing the sum of the similarities between the entities. The inverse of Total Cut is reported as follows: $1 - NormTotalCut(K)$

d) **Modularity (M)**: is the value of the intra-community similarities between the entities divided by the sum of all the similarities between the entities minus the sum of the similarities among the entities in different communities, in case they are randomly distributed in the communities [88]. The value of the Modularity is in the range of $[-0.5, 1]$, which can be scaled to $[0, 1]$ by computing: $\frac{Modularity(K)+0.5}{1.5}$.

e) **Coverage (Co)**: compares the fraction of intra-community similarities between entities to the sum of all similarities between entities [37]. Higher coverage values mean that there are more edges within clusters than edges linking different clusters.

5.4.2 Quality of the Predicted Relations among Standards

The quality of the predicted relations among standards is evaluated by accuracy. In order to measure the accuracy of the predicted relations in the communities, we are comparing them with the relations in the test set. The test set (TS) is used to validate the results, and it is represented as $TS = \{\langle s, p, o\rangle | s, o \in V_e, p \in \mathit{relatedTo}\}$ and V_e are standards (cf. Figure 5.6c). Considering we are applying the homophily prediction principle in the communities, all the standards in a community (c) are related to each other (cf. Figure 5.6b). Homophily prediction in a community is defined as follows: $H(c) = \{\langle s, p, o\rangle | s, o \in c \wedge p \in \mathit{relatedTo} \wedge s \neq o\}$. Then, we are selecting from TS the set of triples $\langle s, p, o\rangle$ where s or o are standards from cluster c; it is defined as follows: $S(c, TS) = \{\langle s, p, o\rangle | \langle s, p, o\rangle \in TS \wedge (s \in c \vee o \in c)\}$. Finally, is evaluated the percentage of predicted relations $acc(c)$ among standards in the community c; $acc(c) = \frac{|S(c,TS) \cup H(c)|}{|S(c,TS)|}$, where the numerator corresponds to the number of discovered relations from c. Since we are executing 5-fold cross-validation with one run, is reported the average accuracy.

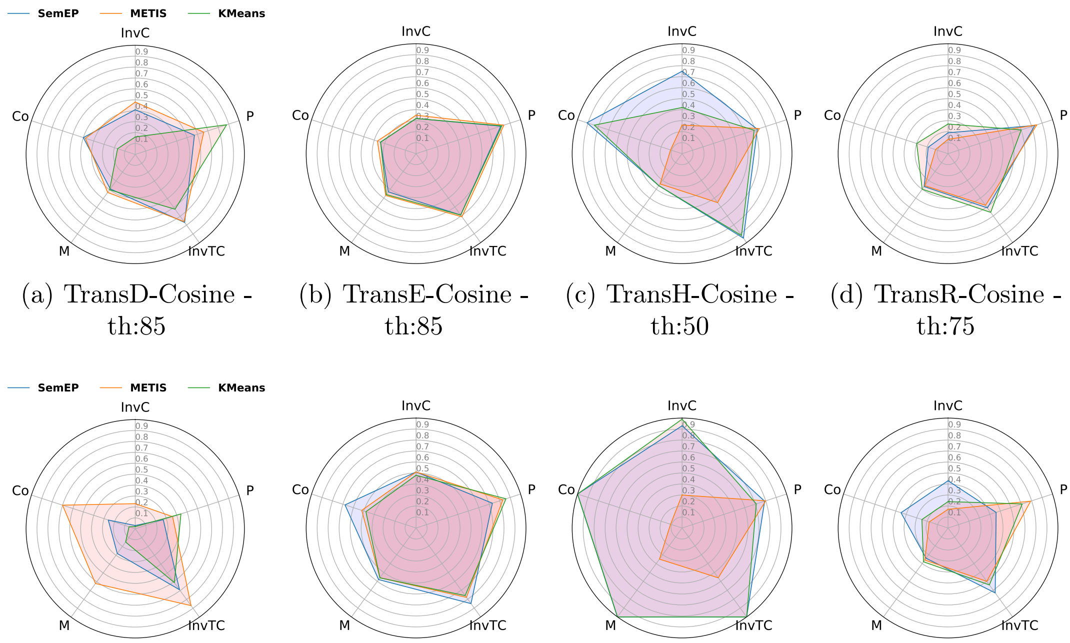
5.4.3 Impact of Community Detection Methods

Average Category-based Score: We compared our baseline, Equivalent Classes, with the communities generated by the community detection algorithms. Given a Community **C** of standards, the average Category-based Score, $\mathcal{C}(C)$, corresponds to the average of the 'Category-based' measure for each pair of standards in the clusters of **C**. Values of $\mathcal{C}(C)$ are in the ranges between 0.0 and 1.0. A value equal to 0.0 indicates that there is no intersection between the classes of equivalence of the pairs of standards in the clusters of **C**, whereas a value close to 1.0 represents that almost all the pairs of standards in each cluster of **C** share exactly the same classes of equivalence. Let EC be the Equivalent classes, EC_i be the set of standards in the Equivalent Class i, C_k be the set of standards in the Community k, and $Comb(n)$ represents the number of pair of standards given a set of standards with cardinality n; it is computed by the number of two combinations of a set of n elements, $Comb(n, r = 2) = \frac{n!}{(n-2)! * 2!} = \frac{n*(n-1)}{2}$. The Average Category-based Score is defined as follows:

$$\mathcal{C}(C_k) = \frac{\sum_{i=1}^{|EC|} Comb(|C_k \cup EC_i|)}{Comb(|C_k|)}$$

$$avg(\mathcal{C}) = \frac{\sum_{k=1}^{|C|} \mathcal{C}(C_k)}{|C|}$$

Quality of the communities: We evaluated three community detection algorithms with two different similarity metrics and four Trans* methods. Considering



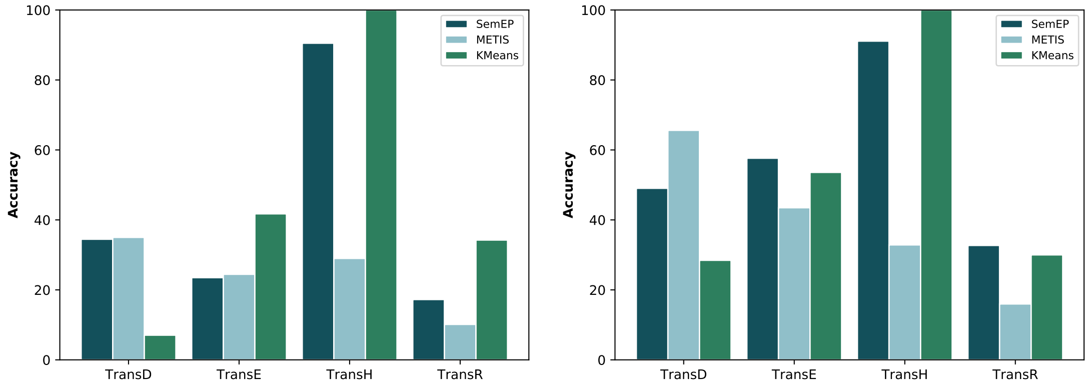
(a) TransD-Cosine - th:85 (b) TransE-Cosine - th:85 (c) TransH-Cosine - th:50 (d) TransR-Cosine - th:75

(e) TransD-Euclidean - th:95 (f) TransE-Euclidean - th:85 (g) TransH-Euclidean - th:60 (h) TransR-Euclidean - th:75

Figure 5.9: **Quality of the generated communities**. The communities are evaluated in terms of prediction metrics using the SemEP, METIS, and KMeans algorithms. Communities are derived for each combination of the Trans* method and similarity measure. In this case, higher values are better. Our approach exhibits the best performance with TransH embeddings in Cosine Similarity and Inverse Euclidean Distance, i.e., Figure 5.9c and Figure 5.9g. SemEP achieves the highest values in the five evaluated parameters using Inverse Euclidean Distance and in four of the five evaluated parameters with Cosine Similarity.

the five metrics for assessing the communities, the best communities are obtained by Inverse Euclidean Distance, TransH, and with the SemEP and KMeans algorithms. Figure 5.9g shows how the InvTC, M, and Co have values close to one for SemEP and KMeans. The Performance (P) for SemEP and KMeans is 0.8 and 0.7, respectively, meaning that communities built by KMeans have more external links to other communities than those built by SemEP. The inverse of Conductance (InvC) is high in SemEP and KMeans, with 0.93 and 0.99, respectively. This metric measures the relatedness of standards in a community and how different they are from standards outside the community.

The *I4.0RD* accuracy: Figure 5.10b shows the best performance for TransH-KMeans achieving 100% of accuracy. However, KMeans is only able to discover three communities of standards while our baseline is already known to have twenty equivalence classes. This means that KMeans is clustering our 249 standards into just three clusters. K-Means finds the optimal number of clusters by computing the K-Elbow curve, but the results are not close to our baseline. Nevertheless, SemEP achieves an accuracy of over 90% in both similarity measures, and furthermore, the number of communities discovered is very close to our baseline, reaching a mean of 16 communities. All the communities are assessed against the baseline to validate their closeness to the equivalence classes.

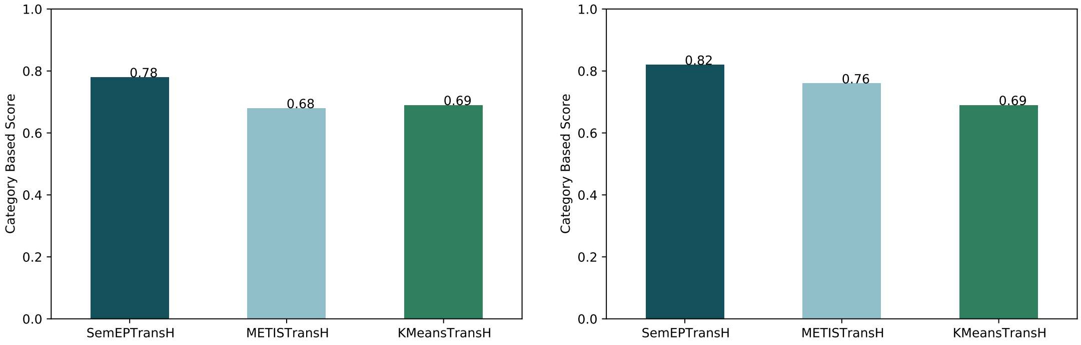


(a) Cosine Similarity Measure

(b) Inverse Euclidean Distance Measure

Figure 5.10: The ***I4.0RD* accuracy**. Percentage of the test set for the property *relatedTo* is achieved in each cluster. Figure 5.10a and Figure 5.10b show the precision of the community detection algorithms by the measure Cosine Similarity and Inverse Euclidean Distance, respectively. Our approach exhibits the best performance using TransH embedding and with the SemEP and KMeans algorithms in both similarity measures reaching an accuracy of up to 90%.

Baseline: TransH is selected as the best embedding according to the results achieved in the metrics for determining relatedness among Standards (cf. Figure 5.9) and quality of the predicted relations among standards (cf. Figure 5.10). Taking TransH as the best embedding of the communities generated by the three community detection algorithms, the two similarity measures are evaluated. Figure 5.11 depicts the results of the measure Average Category-based Score for both similarity measures. The combination SemEP and TransH achieved the best performance in both similarity measures; see Figure 5.11a and Figure 5.11b. Although KMeans has the highest accuracy, the performance in the measure Average Category-based Score, where it is compared with the baseline, is one of the lowest. In contrast, SemEP has the highest values for this measure and is also over 90% accuracy, which means that the communities discovered by SemEP are the closest to our baseline and with high accuracy.



(a) Cosine Similarity Measure (b) Inverse Euclidean Distance Measure

Figure 5.11: **Average Category Based Score respect to Equivalence Class**. Figure 5.11a and Figure 5.11b show how similar our communities are to the baseline. Our approach exhibits the best performance with Inverse Euclidean Distance and SemEP, achieving 82%.

Network analysis: The I4.0KG is updated with the communities found by the combination of TransH, Inverse Euclidean Distance, and SemEP, which is the best performer for the metrics evaluated. With the updated I4.0KG, we are adding new links predicted by the communities. Table 5.1 shows the analysis of I4.0KG with new predicted links against our baseline. We improve the standards connectivity by predicting new links.

Q1 - Corroborating the quality of communities in I4.0KG. We executed a five-fold cross-validation procedure to compute the accuracy of $\mathcal{I4.0RD}$. To that end, the data set is divided into five consecutive folds shuffling the data before splitting into folds. Each fold is used once as validation, i.e., the test set, while the remaining fourth folds form the training set. Figure 5.9 depicts the impact of metrics for evaluating communities. The best results are obtained with the combination of the Inverse Euclidean Distance and TransH with SemEP and KMeans algorithms; see Figure 5.9g. The values obtained for this combination for both SemEP and KMeans are high except for the metric **Performance** (P). SemEP and KMeans have values of 0.8 and 0.7, respectively, which means that communities built by KMeans have more external links to other communities than communities by SemEP.

Q2 - Predicting new relations between standards. In order to assess the second research question, the data set is divided into five consecutive folds. Each fold comprises 20% of the relationships between standards. Next, precision mea-

Table 5.1: **Network connectivity analysis**. Table 5.1 shows the statistics for I4.0KG after transitive closure of the property *relatedTo* between standards and the statistics I4.0KG with the new links predicted by combining TransH, Inverse Euclidean Distance, and SemEP. Results reveal a general improvement in connectivity when predicting new links. The Number of edges, Avg. number of neighbors, and Network density increase predicting new links, allowing for fewer connected components and improving data integration. Measures that improve are highlighted in **bold**. The network analysis was performed by Cytoscape [119].

Statistic	Baseline	TransH-Inv.Euclidean-SemEP
Number of nodes	249	249
Number of edges	22,969	**23,207**
Avg. number of neighbors	91.245	**92.201**
Network diameter	1	**3**
Network radius	1	1
Characteristic path length	1.000	1.001
Clustering coefficient	0.976	0.974
Network density	0.368	**0.372**
Connected components	20	**13**
Multi-edge node pairs	11,360	**11,479**
Number of self-loops	249	249

surement is applied to evaluate the main objective: to unveil uncovered associations and, at the same time, corroborate knowledge patterns that are already known. As shown in Figure 5.10, the best performances for the property *relatedTo* are achieved by TransH embeddings in combination with the SemEP and KMeans algorithm in both similarity measures. KMeans reaches higher accuracy than SemEP; however, KMeans discover only three communities of standards while our baseline is already known to have twenty Equivalence Classes. On the other hand, the number of communities discovered by SemEP is very close to our baseline, reaching a mean of 16 communities. The communities of standards discovered using TransH embeddings, Inverse Euclidean Distance, and the SemEP algorithm contribute to the resolution of interoperability in I4.0 standards. To provide an example of this, we observe a resulting cluster with the standards *IEC 60255 P27 E2, IEC 60255 P151 E1, IEC 60255 2010, IEC 60255 P1 E1, IEC 60255 P149 E1* and *MTConnect*. The former provides an information model for describing manufacturing data. The latter offers a vocabulary for manufacturing equipment. It is important to note that the standard *MTConnect* is not related to the training set nor in I4.0KG. The membership of those standards in the cluster means that they should be classified together in the standardization frameworks. Besides, it also suggests to the creators of the standards that they might look after possible existing synergies between them. This example suggests that the techniques employed in this work are capable of discovering new communities of standards. These communities can be used to improve the classification that the standardization frameworks provide for the standards.

Q3 - Comparison with the baseline of equivalent classes. From the combination of four Trans embeddings, two similarity measures, and three community detection algorithms, we asses 24 results. In both the evaluation of the quality of the communities and the accuracy of new relations, the best results are reached with the TransH embedding, SemEP, and KMeans as cluster algorithms and both similarity metrics. Finally, in the evaluation with the baseline, the best similarity metric is Inverse Euclidean Distance, and the best clustering algorithm is SemEP. Figure 5.11b shows the Average Category Based Score achieved by SemEP with respect to Equivalence Class. We reach high values, meaning that almost all the pairs of standards in each community share the same equivalence classes.

5.4.4 Discussion

The techniques proposed in this chapter rely on known relations between I4.0 standards to discover novel patterns and new relations. During the experimental study, we observed that these techniques could group together standards that are known to be related and standards whose relatedness is implicitly represented in the I4.0KG. This feature facilitates the detection of high-quality communities as reported in Figure 5.9, as well as for an accurate discovery of relations between standards (cf. Figure 5.10) and for the evaluation with the baseline of equivalent classes, as shown in Figure 5.11. As observed, the accuracy of the approach can be benefited from the application of the Trans* family algorithms, e.g., TransH, and from similarity measures, e.g., Inverse Euclidean Distance. Additionally, SemEP groups in the same communities have highly similar standards and lead our approach to high-quality discoveries. Our results suggest that the techniques TransH, Inverse Euclidean Distance, and SemEP uncover meaningful communities with high quality because the performance of the five metrics for evaluating communities are close to one, which means that standards in a community are different from standards outside the community, and there are more edges within communities than edges linking different communities. Also, the accuracy is up 90%, which means that are discovered over 90% of the relationships and evaluating with the baseline achieving 82%, i.e., almost all the pairs of standards in each community share exactly the same equivalence classes. Moreover, the number of communities is close to the number of equivalence classes in the baseline.

We analyze both techniques in detail to understand why the aforementioned combination of TransH, Inverse Euclidean Distance and SemEP produces the best results. TransH introduces the mechanism of projecting the relation to a specific hyperplane [137], enabling, thus, the representation of relations with cardinality many to many. Since the materialization of transitivity and symmetry of the property *relatedTo* corresponds to many to many relations, the instances of this materialization are taken into account during the generation of the embeddings, specifically during the translating operation on a hyperplane. Thus, even though semantics is not explicitly utilized during the computation of the embeddings, considering different types of relations empowers the embeddings generated by TransH. Moreover, it allows for a more precise encoding of the standards represented in I4.0KG. Figures 5.8c and 5.8g illustrate groups of standards in the similarity intervals $[0.9, 1.0]$, and $[0.0, 0.4]$. Inverse Euclidean Distance is able to find in all the Trans* methods a higher density of similar standards than Cosine Similarity. The SemEP algorithm can detect these similarities and represent them in high-precision communities. The other three embedding models, i.e., TransD, TransE, and TransR, do not represent the standards in the best way with either of the two similarity measures. TransD, TransE, and TransR report that most of the standards are in the similarity interval $[0.0, 0.4]$ (cf. Figure 5.8). This means that no community detection algorithm could be able to discover communities with high quality. Reported results indicate that the presented approach enables – in average– for discovering communities of standards by up to 90%. As an example of a relevant community, we observed a resulting cluster with the standards *IEC 60255 P27 E2, IEC 60255 P151 E1, IEC 60255 2010, IEC 60255 P1 E1, IEC 60255 P149 E1*, and *MTConnect*. All of them are related to product safety requirements and vocabulary for manufacturing equipment. It is important to note that the *MTConnect* standard is in a different equivalent class than the other community standards. However, our approach *I4.0RD* is able to create a community grouping all of them together. Although these results required the validation of experts in the domain, an initial evaluation suggests the results are accurate.

5.5 Summary

This chapter presents the *I4.0RD* approach that combines a deductive database system with knowledge graph embedding to discover relationships between I4.0 standards. We addressed the problem of exploiting encoded data in knowledge graphs to discover relationships by applying our Neuro-Symbolic AI approach. Our approach resorts to I4.0KG to discover relations between standards; I4.0KG represents relations between standards extracted from the literature or defined according to the classifications stated by the standardization frameworks.

The symbolic system implemented by a deductive database makes implicit relationships explicit and minimizes the data sparsity issues in the KG. The IDB establishes rules for the equivalence relation of the *relatedTo* property and rules for declaring that standards belonging to the same classes are related by the *relatedTo* property. We integrate the deductive database system with the sub-symbolic system implemented by KGE models. The KGE models benefit from the symbolic system as it enhances the ego networks of standards, assisting in overcoming the data sparsity issues in I4.0KG. Different algorithms for generating embeddings are applied on top of I4.0KG. Two similarity measures are applied to assess the similarity of the standards. We employed three community detection algorithms, i.e., SemEP, METIS, and KMeans, to identify similar standards, i.e., communities of standards, as well as to analyze their properties. Additionally, by applying the homophily prediction principle, novel relations between standards are discovered. We empirically evaluated the quality of the proposed techniques over 249 standards, initially related through 736 instances of the property *relatedTo*. The deductive database system makes 22,233 implicit relationships explicit in the I4.0KG. Furthermore, the equivalent classes induced by the property *relatedTo* are used as a baseline in the evaluation process. The Trans* family of embedding models is used to identify a low-dimensional representation of the standards according to the materialized instances of *relatedTo*. Results of a 5-fold cross-validation process suggest that our approach is able to identify novel relations between standards effectively. In addition, the Inverse Euclidean Distance enables identifying relations with higher precision. Thus, our work broadens the repertoire of knowledge-driven frameworks for understanding I4.0 standards facilitating the resolution of the existing interoperability issues in the I4.0 landscape.

Chapter 6

Applications

In this chapter, we present the use of the Deductive System DS on three different KGs and diseases. We investigate the applicability of the deductive database system in the biomedical domain, specifically in four projects, iASiS[5], BigMedilytics[6], P4-LUCAT[7], and H2020 CLARIFY[8]. iASiS is a European Union Horizon 2020-funded project that seeks to pave the way for precision medicine by utilizing patient data insights. iASiS focuses on two disease use cases: lung cancer and dementia. BigMedilytics is an H2020 project aiming to develop innovative data-driven solutions to improve the healthcare system in Europe. BigMedilytics covers a wide range of chronic diseases and frequent cancers (e.g., prostate, lung, and breast). We show the potential for discovering patterns that can enable the explanation of treatment interactions and patient characterization. Thus, we broaden the scope and applicability of DS in several domains. Figure 6.1 depicts the main challenges, and the contribution tackled in this chapter. The content of this chapter is based on the publications [3, 103, 104, 115, 131]. The results of this chapter provide an answer to the following research question:

RQ5: How can the proposed approach be applied to real-world cases?

We present the main results that show a significant benefit in the discovery task on a knowledge graph in each application. The remainder of this chapter is structured as follows: Section 6.1 presents the deductive database DS to deduce DDIs and compute the interaction score of drugs in treatment based on the wedge concept. The results are evaluated in Knowledge4COVID-19 KG and real lung cancer treatments. Section 6.2 explains the adverse effect of Covid-19 treatments retrieved on top of the Knowledge4COVID-19 KG. Then, Section 6.3 assesses the impact of DDIs on the effectiveness of lung cancer treatments DE4LungCancer KG. Next, Section 6.4 presents a similarity measure that evaluates the similarity

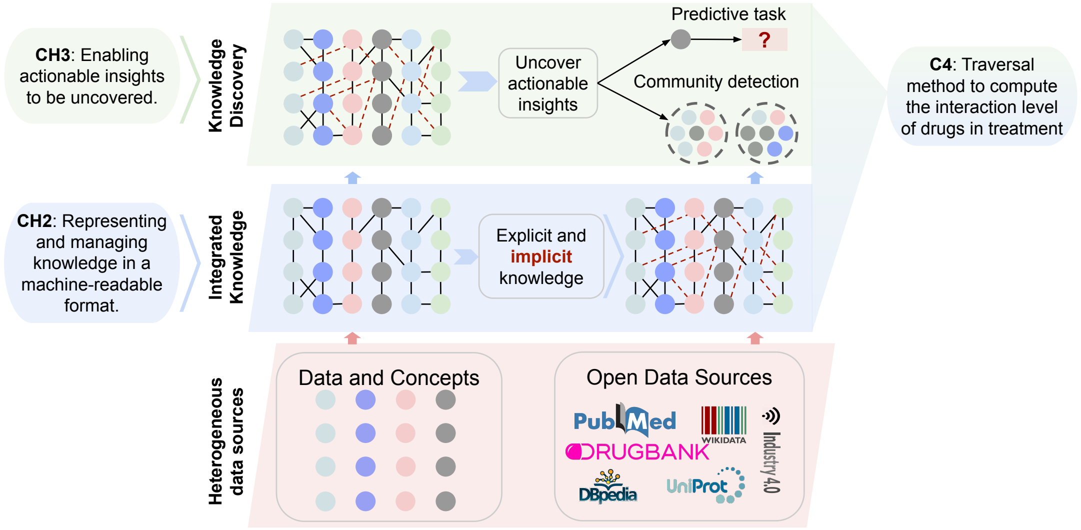

CH3: Enabling actionable insights to be uncovered.
Knowledge Discovery
Uncover actionable insights
Predictive task
?
Community detection
C4: Traversal method to compute the interaction level of drugs in treatment
CH2: Representing and managing knowledge in a machine-readable format.
Integrated Knowledge
Explicit and implicit knowledge
Heterogeneous data sources
Data and Concepts
Open Data Sources
PubMed
WIKIDATA
DRUGBANK
Industry 4.0
DBpedia
UniProt

Figure 6.1: **Challenges and contributions**. This chapter focuses on discovering relationships in a knowledge graph that can enable the explanation of treatment interactions and patient characterization.

between patients, and a knowledge discovery technique is used to uncover patterns in iASiS KG [131]. Finally, Section 6.5 presents the closing remarks of the chapter.

6.1 Traversal method to compute the interaction score of a drug in treatment

Drug treatments have been of great interest over the years to ensure that they are administered safely and with maximum benefit to any given patient. Patient safety can be affected by exposure to combinations of drugs that could interact with and cause toxicity or treatment failure. Nowadays, multi-drug treatments are common, and identifying potential Drug-Drug Interactions is crucial. Public drug databases and semi-structured resources provide a wealth of information on drugs that can be exploited to enhance tasks, e.g., data mining, ranking, and query answering. These databases mainly focus on drug-drug pair interactions, while DDIs remain unknown when in multi-drug treatments. Unknown DDIs in multi-drug treatments need to be revealed to enable clinicians to assess the effectiveness of treatments and anticipate the toxicities. Database and Semantic Web communities have extensively studied the problem of DDIs [68, 140, 149], and various approaches have been proposed to support the detection of potential DDIs in treatments [55, 98]. Ontology- and knowledge graph-based approaches have also contributed to integrating DDIs and creating a minimal information model for describing potential DDIs [55]. However, finding new DDIs in a multi-drug treatment requires capturing knowledge about individual DDIs among drug pairs and analyzing the effects of these DDIs in the whole treatment.

We model treatments, their prescribed drugs, and the DDIs in a knowledge graph and address the *problem of discovering relationships over knowledge graphs*. We propose a symbolic system to deduce relationships that considers the semantics encoded in the knowledge graph and the connectivity of the relations. The proposed symbolic system implemented by a deductive database deduces the unknown relationships encoded in a set of rules through a Datalog program. Further, the deductive database can compute the interaction score of drugs in treatment through a graph traversal method. The traversal method relies on the computation of wedges in a knowledge graph and then computes the distribution of the middle-vertex of wedges. A middle vertex is particularly important in a wedge because it is the object drug of one DDI and the precipitant drug of another DDI. Thus, drugs that correspond to the middle vertex of several wedges, represent drugs whose presence in the treatment may negatively impact effectiveness. We summarize the contributions of this section as follows:

1. A deductive database system based on Datalog that is capable of deducing implicit DDI represented in a knowledge graph and providing a ranking of interaction score of drugs in treatment.
2. An extensive evaluation of our approach to treatments of different diseases. The evaluation material is available in CK-DDI[1].

6.1.1 Symbolic System

The symbolic system corresponds to a deductive system DS, where the EDB comprises ground facts. The deductive system proposed is based on the Datalog program (1) presented in chapter 4. The extensional database corresponds to statements about interactions between drugs stated in a KG, and the IDB of the DS comprises a set of rules to deduce new DDIs and wedges in treatments. We consider the same ground predicates included in the EDB of the deductive database (1) together with the following predicate:

$$\begin{array}{ll}
rule_3(serum, increase). & rule_3(serum, decrease). \\
rule_3(metabolism, decrease). & rule_3(metabolism, increase). \\
rule_3(absorption, increase). & rule_3(absorption, decrease). \\
rule_3(excretion, decrease). & rule_3(excretion, increase).
\end{array}$$

[1] https://mybinder.org/v2/gh/arivasm/KCAP$_{\mathrm{D}}$EMO/HEAD?urlpath=voila%2Frender%2Fcomputation$_{\mathrm{d}}$rug$_{\mathrm{w}}$edge$_{\mathrm{C}}$OVID.ipynb

The above-mentioned $rule_3$ identifies the combinations of effect and impact that alter the toxicity or effectiveness of an object drug representing the pharmacokinetic DDIs. The following IDB comprises Horn rules that state when a new DDI can be deduced and when a wedge exists in a treatment:

$$precipitant(ID, A), object(ID, B), \mathit{effect(ID,\ E)}, impact(ID, I) \Rightarrow ddi(A, E, I, B).(10)$$

$$ddi(A, E, I, B) \Rightarrow inferred\_ddi(A, E, I, B).(11)$$

$$inferred\_ddi(A, E_2, I_2, B), ddi(B, E, I, C), rule_1(E, I), rule_1(E_2, I_2), (A! = C) \Rightarrow inferred\_ddi(A, E, I, C).(12)$$

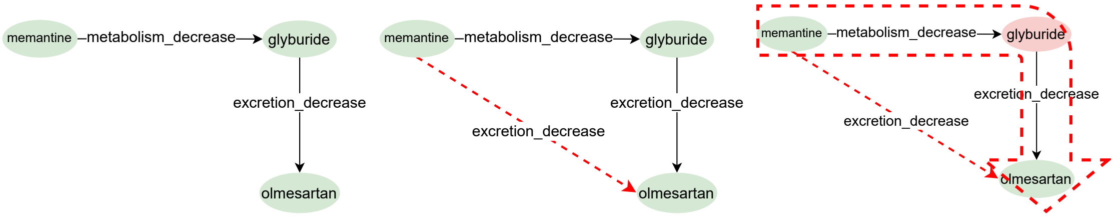
$$inferred\_ddi(A, E2, I2, B), ddi(B, E, I, C), rule_2(E, I), rule_2(E_2, I_2), (A! = C) \Rightarrow inferred\_ddi(A, E, I, C).(13)$$

$$inferred\_ddi(A, E, I, B), inferred\_ddi(B, E_2, I_2, C), (A! = C) \Rightarrow wedge(A, B, C, E, I, E_2, I_2).(14)$$

$$inf erred\_ddi(A, E, I, B), inf erred\_ddi(B, E_2, I_2, C), rule_3(E, I), rule_3(E_2, I_2), (A! = C) \Rightarrow wedge\_pk(A, B, C, E, I, E_2, I_2).(15)$$

A wedge w is a path with two edges in a directed labeled graphs [142]; w is composed of three vertices $\{a, b, c\}$ and two ordered pairs of edges $\{(a, b), (b, c)\}$ of the directed labeled graph. The vertex b is the middle vertex of w. We apply the wedge concept to the DDIs knowledge graph, where the edges of a wedge represent DDIs. The middle vertex is both the object drug of one interaction and the precipitant drug of the other interaction. A wedge w is defined as the following: $w =$ vertex triplet(a,b,c), where $\{a, b, c\} \subseteq V$ and $\{(a, b), (b, c)\} \subseteq E$. The node b is the middle-vertex of w.

The rules for deducing DDIs in treatments are the same as presented in the deductive database (1), and we include the Rule (15) for deducing the wedges in treatments considering pharmacokinetic DDIs. Rule (14) determines the wedges in treatments where the DDI can be pharmacokinetic or pharmacodynamic DDIs. The head predicate *wedge* represents wedges, where the first three terms correspond to the wedge vertexes. The second term is the middle vertex of the wedge. The last four terms represent the effects and impacts of the DDIs in the wedge, where the first two are the effect and impact of the first DDI, and the last two represent the effect and impact of the second DDI. Figure 6.2 illustrates a multidrug treatment with two DDIs. Figure 6.2b shows in red color a DDIs deduced by the Rule (12) in from the DS (10) and Figure 6.2c depicts a wedge deduced by the Rule (14) and Rule (15) highlighted in red color where the drug glyburide represents the middle vertex of the wedge.



(a) Graph representing DDIs in treatment

(b) Graph representing a deduced DDIs in treatment

(c) Graph representing a wedge in treatment

Figure 6.2: **Wedge in a treatment**. Figure 6.2a shows a treatment with two interactions, drug memantine decreases the metabolism of the drug glyburide, and the drug glyburide decreases the excretion of the drug olmesartan. Figure 6.2b illustrates a deduced DDI in red color by DS; the drug memantine decreases the excretion of the drug olmesartan with the Rule (12). Figure 6.2c represents the deduced wedge(memantine, glyburide, olmesartan) with Rule (14) and Rule (15). The drug glyburide is the middle vertex of the wedge.

6.1.2 Experimental Study

We empirically evaluate the effectiveness of our approach to deduce DDIs in a knowledge graph and compute the interaction score of drugs in treatment. The Knowledge4COVID-19 KG [115] is a unique source of knowledge to identify patterns in the integrated networks of interactions, biomedical entities, and publications, e.g. adverse events generated by combining COVID-19 drugs and drugs prescribed for pre-existing conditions. In particular, we aim to answer the following research questions: **Q1)** Can our approach be able to uncover DDIs in a multi-drug treatment? **Q2)** What is the impact of the symbolic system implemented in Datalog on the computation of the interaction score of drugs in treatment? We configure the following empirical study to assess these questions.

Benchmark: We conduct our evaluation over ten real treatments for three different diseases. The first seven treatments (T1-T7) are for patients with COVID-19 who are treated with concomitant medications for an underlying medical condition. The treatment for COVID-19 is extracted from [28]. The concomitant medications for the first treatment, $T1$ are for the comorbidities antihypertensive: Beta-blockers, statins, and Type 2 Diabetes, and for the second treatment, $T2$, are asthma, statins, and Type 2 Diabetes. The comorbidities in the third treatment, $T3$, are for the comorbidities asthma, high cholesterol, and pneumonia, and for the fourth treatment, $T4$, are diabetes, hypertension, and pneumonia. The comorbidities in the fifth treatment, $T5$, are diabetes, high cholesterol, and hypertension. The comorbidities in the sixth treatment, $T6$, are asthma and hypertension, and for the seventh treatment, $T7$, are renal diseases, obesity, and hypertension. The

treatment $T8$ is for Alzheimer's disease and is extracted from the paper [63]. The concomitant medications for $T8$ are Hypertension and Type 2 Diabetes comorbidities. The treatment for Hypertension is extracted from [1]; it is a therapy with three antihypertensive agents: Angiotensin Receptor Blockers (ARB), a thiazide diuretic, and amlodipine, known as triple therapies. The treatment for Type 2 Diabetes is a typical drug to help your body secrete more insulin. The last two treatments are for Parkinson's disease; they are extracted from [87] and [93]. Parkinson's disease is often accompanied by problems, which may be treatable (e.g, depression, excessive sweating, and urinary incontinence [96]).
Metrics: We measure the deduced percentage of edges (D); it is defined as follows: $D = (|E_2|-|E_1|)/|E_2| * 100$, where E_2 corresponds to the edges result of applying the deductive system over Knowledge4COVID-19 KG, and E_1 is the actual edges in Knowledge4COVID-19 KG. D is a "higher is better" metric representing the percentage of new edges added to the deduced graph.

Impact of Symbolic System on Identifying DDIs in Treatments

In this section, we evaluate the drug-drug interactions that can be deduced over the Knowledge4COVID-19 KG and the effects of these interactions. Table 6.1 shows the percentage of DDIs deduced (D) and wedge absolute frequency (F) for each middle-vertex by the method [103] in existing treatments. The middle vertex of a wedge is highly important because the middle vertex is both the object drug for one interaction and the precipitant drug for another interaction. Thus, drugs that correspond to the middle-vertex of wedges, represent drugs whose presence in the treatment may negatively impact effectiveness and toxicity.

Table 6.1: **Ten multi-drug Treatments**. Frequency distribution of wedges with the symbolic system. Treatments are evaluated in four interaction checker tools: COVID-19 (C-19), WebMD (WD), Medscape (MS), and DrugBank (DB) (May 2nd, 2022). Each tool shows the DDI-Reduction percentage that indicates how many DDIs are avoided in a treatment when the middle-vertex drug is removed. The DDI-reduction percentage is a higher-is-better metric. The symbol "-" indicates that the treatment is not part of the interaction checker tools. Middle-vertex drugs reduce the DDIs, suggesting, thus, wedges and their middle vertices are part of DDIs that affect treatment effectiveness and toxicities. Best values in **bold**.

T	Symbolic System		D	DDI-Reduction Percentage			
	Middle-Vertex	F		C-19	WD	MS	DB
T1	**hydroxychloroquine**	**22**	52.17	**66.7**	50.0	50.0	50.0
	azithromycin	18					

	dapagliflozin	15					
	lovastatin	12					
	metoprolol	12					
T2	**hydroxychloroquine**	**22**	56.52	**100.0**	33.3	33.3	50.0
	azithromycin	18					
	glyburide	15					
	simvastatin	12					
	montelukast	12					
T3	**Azithromycin**	**9**	45.45	**100.0**	**100.0**	**100.0**	42.9
	Montelukast	4					
	Lovastatin	4					
	Hydroxychloroquine	0					
	Doxycycline	0					
T4	**Ciprofloxacin**	**12**	52.17	33.3	**75.0**	**75.0**	44.4
	Metoprolol	**12**		33.3	25.0	25.0	33.3
	Hydroxychloroquine	9					
	Azithromycin	9					
	Linagliptin	7					
T5	**Hydroxychloroquine**	**5**	33.33	**100.0**	25.0	25.0	60.0
	Glyburide	**5**		0.0	50.0	50.0	60.0
	Simvastatin	3					
	Azithromycin	3					
	Ramipril	0					
T6	**Propranolol**	**8**	15.38	**100.0**	50.0	50.0	60.0
	Hydroxychloroquine	5					
	Azithromycin	5					
	Theophylline	4					
	Ramipril	1					
T7	**Timolol**	**11**	38.89	**50.0**	**50.0**	**50.0**	44.4
	Cyclophosphamide	**11**		0.0	0.0	0.0	44.4
	Azithromycin	7					
	Hydroxychloroquine	7					
	Bupropion	6					
T8	**glyburide**	**15**	26.92	–	33.3	33.3	**83.3**
	amlodipine	10					
	memantine	8					
	olmesartan	0					

	donepezil	0					
	hydrochlorothiazide	0					
T9	**venlafaxine**	**21**	36.84	–	20.0	20.0	**45.5**
	darifenacin	6					
	oxybutynin	6					
	safinamide	4					
	levodopa	2					
	carbidopa	0					
T10	**imipramine**	**45**	30.77	–	40.0	40.0	**45.5**
	carbidopa	12					
	tolterodine	12					
	oxybutynin	8					
	entacapone	0					
	levodopa	0					

We can observe in Table 6.1 that over 15% of new DDIs are deduced in all the treatments. The middle vertexes with higher wedge absolute frequency are highlighted in bold, which are the ones that could decrease the effectiveness of the treatment. Table 6.1 shows the DDI-Reduction percentage for the drugs with higher wedge absolute frequency (F) for each treatment. The DDI-Reduction percentage is evaluated in four interaction checker tools on May 2nd, 2022, Liverpool COVID-19 Interactions[2], WebMD[3], Medscape[4], and Drugbank[5]. The validation is done on the versions of Liverpool COVID-19 Interactions and Drugbank, which correspond to 2022-04-13 and 2022-01-04, respectively. Existing tools (e.g., COVID-19 Drug Interactions for the University of Liverpool) only identify pairwise interactions. DDI-Reduction percentage is measured, and it indicates how many DDIs are avoided in a treatment when the middle-vertex drug is withdrawn. The evaluation suggests that withdrawing the middle vertex with higher absolute frequency reduces most interactions. Thus, wedges and their middle-vertex represent DDIs affecting treatment effectiveness and toxicity. When more than one drug contains the higher wedge absolute frequency (F) in treatment, clinicians must decide which drug is withdrawn. The third COVID-19 treatment contains concomitant drugs for asthma, high cholesterol, and pneumonia comorbidities. The method proposed by [103] indicates Azithromycin as the drug with the highest absolute frequency of being the wedges middle-vertex. Therefore, it represents

[2] https://www.covid19-druginteractions.org/checker

[3] https://www.webmd.com/interaction-checker/default.htm

[4] https://reference.medscape.com/drug-interactionchecker

[5] https://go.drugbank.com/drug-interaction-checker

the DDIs that affect treatment effectiveness and toxicities, and withdrawing it reduces most interactions.

Ranking of Interaction score of Drugs in Treatment

A graph traversal method computes the wedges and the distribution of the middle-vertex of wedges. The maximal possible number of wedges centered at vertex v is defined as $\mu = |x| * \binom{n}{r} = |x| * \frac{(n)!}{r!(n-r)!}$, where n: represents the number of drugs in the treatment, r represents a pair of drugs involved in each different DDIs, and x represents the set of effects and impacts of DDIs. μ computes the combinations of pairs of drugs in the treatment multiplied by the cardinality of the set of effects and impact of DDIs. The interaction score centered at each drug v in treatment is computed by: $\Upsilon_v = \frac{W_v}{\mu}$, where W_v is the amount of wedge in the vertices v. The interaction score represents drugs whose presence in the treatment may negatively impact effectiveness and toxicity. The range of the interaction score of Υ_v is $[0, 1]$, where higher values mean drugs that correspond to the middle vertex of several wedges and may negatively impact the treatment because they participate in multiple DDIs as both precipitant and object drug. A zero value in Υ_v means that the drug v is not the object drug for one interaction and the precipitant drug for another interaction. Thus, drug v is not a middle-vertex of the wedges, i.e., $W_v = 0$.

Since our method distinguishes between pharmacokinetic and pharmacodynamic DDIs, the method produces two interaction-score rankings, one for pharmacokinetic DDIs and one for both pharmacokinetic and pharmacodynamic DDIs. We evaluate the following three real lung cancer treatments composed of oncological drugs non Oncological drugs:

- First lung cancer treatment (LCT1):
  - Oncological drugs: Gemcitabine, Nivolumab.
  - Non-Oncological drugs: Ranitidine, Ciprofloxacin, Furosemide, Gabapentin.
- Second lung cancer treatment (LCT2):
  - Oncological drugs: Pemetrexed.
  - Non-Oncological drugs: Omeprazole, Lormetazepam, Ondansetron, Metoclopramide, Tamsulosin.
- Third lung cancer treatment (LCT3):
  - Oncological drugs: Alectinib.

- Non-Oncological drugs: Atorvastatin, Diazepam, Folic acid, Atenolol.

We computed the interaction score for each drug in the three treatments, and clinicians evaluated the results. Table 6.2 shows the results of our *DS* computing the interaction score of drugs in treatments. We can observe that non-oncological drugs have higher interaction scores than oncological drugs when the treatment contains comorbidities drugs. The behavior of the interaction score is similar, either considering the pharmacokinetic DDIs or both pharmacokinetic and pharmacodynamic DDIs. The oncological drug Nivolumab and the non-oncological drug Gabapentin in the LCT1 treatment are not in the interaction score because both drugs are not a middle vertex of a wedge in the treatment. We have developed an API[6] to execute our method. The interaction score method is used for the clinicians from the Hospital Universitario Puerta del Hierro of Majadahonda of Madrid (HUPHM) in the project CLARIFY[8].

Table 6.2: **Interaction Score of Drugs in Treatments**. The first column represents the treatments, the second column represents the interaction score considering pharmacokinetic DDIs (PK), and the third column depicts the score considering pharmacokinetic and pharmacodynamic DDIs (PK-PD). Drugs with the higher score are in bold.

Treatment	**Interaction Score (PK)**		**Interaction Score (PK-PD)**	
	Drug	Score	Drug	Score
LCT1	**Ciprofloxacin**	**0.333**	**Furosemide**	**0.762**
	Furosemide	0.267	Ciprofloxacin	0.333
	Gemcitabine	0.000	Gemcitabine	0.000
	Ranitidine	0.000	Ranitidine	0.000
LCT2	**Tamsulosin**	**0.133**	**Ondansetron**	**0.095**
	Metoclopramide	0.083	Metoclopramide	0.086
	Ondansetron	0.050	Tamsulosin	0.076
	Pemetrexed	0.000	Pemetrexed	0.000
	Omeprazole	0.000	Omeprazole	0.000
	Lormetazepam	0.000	Lormetazepam	0.000
LCT3	**Folic acid**	**0.333**	**Folic acid**	**0.333**
	Diazepam	0.333	Diazepam	0.333
	Alectinib	0.200	Alectinib	0.200
	Atenolol	0.000	Atenolol	0.000
	Atorvastatin	0.000	Atorvastatin	0.000

[6] `https://github.com/SDM-TIB/CLARIFY`$_K$G$_E$`xploration`$_A$PI

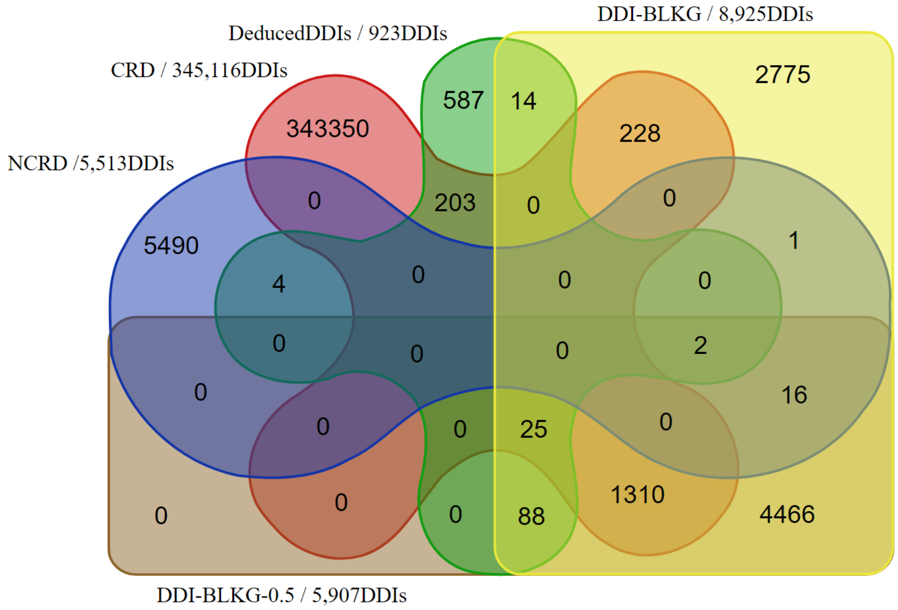


Figure 6.3: **Venn Diagram depicts the overlap among five sets of DDIs**. 345,116 CRD pairs of drugs targeting at least one protein of the family CYP. 5,513 NCRD are pairs of drugs targeting a No CYP protein. 8,925 DDI-BLKG are DDIs predicted by the DDI-BLKG method, while 5,907 DDI-BLKG-05 represents the subset of DDIs in DDI-BLKG with a score equal to or greater than 0.5. 923 DeducedDDIs generated by the deductive system.

6.1.3 Effectiveness of the predictive tasks for DDI identification

The Knowledge4COVID-19 KG integrates 216 COVID-19 treatments that comprise COVID-19 drugs and drugs for the most common comorbidities that impact the survival of COVID-19 patients [31]. In addition, Knowledge4COVID-19 KG incorporates 345,116 CRD, where CRD are drugs from DrugBank that target at least one protein of the family CYP [123] and 5,513 NCRD pairs of drugs, where NCRD drugs target at least one protein but are not of the family CYP [123]. Furthermore, Knowledge4COVID-19 KG integrates 923 deduced DDIs (a.k.a. DeducedDDIs), where DeducedDDIs are deduced DDIs by our *DS* (10), and 8,925 predicted DDIs generated by the DDI-BLKG method, 5,907 have a score equal to or greater than 0.5 (a.k.a. DDI-BLKG-0.5). DDI-BLKG is proposed by Bougiatiotis et.al. [20] and predicts DDIs based on scientific publications. This method analyses the paths connecting interacting and non-interacting drug pairs in this Knowledge4COVID-19 KG and trains a machine learning algorithm (random forest) to discriminate between those two classes. Based on the trained model is then apply predictions to all non-interacting pairs to identify potential DDIs that were not previously known based on the resulting prediction confidence scores.

We compute the DDI over the 216 COVID-19 treatments in knowledge4COVID-19 KG with the methods DDI-BLKG and DeducedDDIs. Then, we analyzed the DDIs predicted with the CRD and NCRD drug pairs. Figure 6.3 reports on the overlap between the DDIs deduced on the drugs of the COVID-19 treatments (a.k.a. DeducedDDIs), DDI-BLKG, DDI-BLKG-0.5 (DDI-BLKG with a prediction score equal or greater than 0.5), CRD, and NCRD. It is essential to highlight that CRD and NCRD are computed from the whole DrugBank dataset of drugs, while DDI-BLKG and DeducedDDIs are limited to COVID-19 drugs. The percentages of overlap of DeducedDDIs, DDI-BLKG, and DDI-BLKG-0.5 with CRD are 24.70%, 17.51%, and 22.60 %. Thus, both methods (i.e., the deductive system and DDI-BLKG) can identify DDIs between drugs mediated by the CYP enzyme family, i.e., CRD pairs of drugs. CYP enzymes play an important role in catalyzing the metabolism of pharmaceuticals, and their inhibition or induction causes clinically significant pharmacokinetic drug-drug interactions [50]. Thus, these results suggest that even though these methods do not exploit any information about the drug's target enzymes, they can identify a good proportion of DDIs that are part of the CRD group.

6.2 Relevant Adverse Effects Detected on Knowledge4COVID-19

This section describes the adverse effect of Covid-19 treatments retrieved on top of the Knowledge4COVID-19 KG. We aim to provide support for analyzing relevant adverse effects that may be produced as a result of interactions among drugs to treat COVID-19 and conditions. As a proof of concept, we illustrate the results of the analysis of the most common comorbidities, i.e., hypertension, asthma, and diabetes. These comorbidities are linked to the ACE-2 receptor expression and may facilitate the entry of the virus into the host cells as a consequence of releasing the proprotein convertase. More importantly, this effect may fire a "vicious infectious circle," which may result in an increase in morbidity and mortality [33]. Nevertheless, a more detailed analysis of the impact of the combination of drugs can be executed on the publicly available Jupyter Notebook[7]. Exemplar drug-drug interactions represented in the Knowledge4COVID-19 KG can also be visualized[8].

[7] https://colab.research.google.com/drive/146-oQTxDpZQoOifKY6iafaEwuupH7q3t?usp=sharing

[8] https://youtu.be/7YsTYJzRfR0

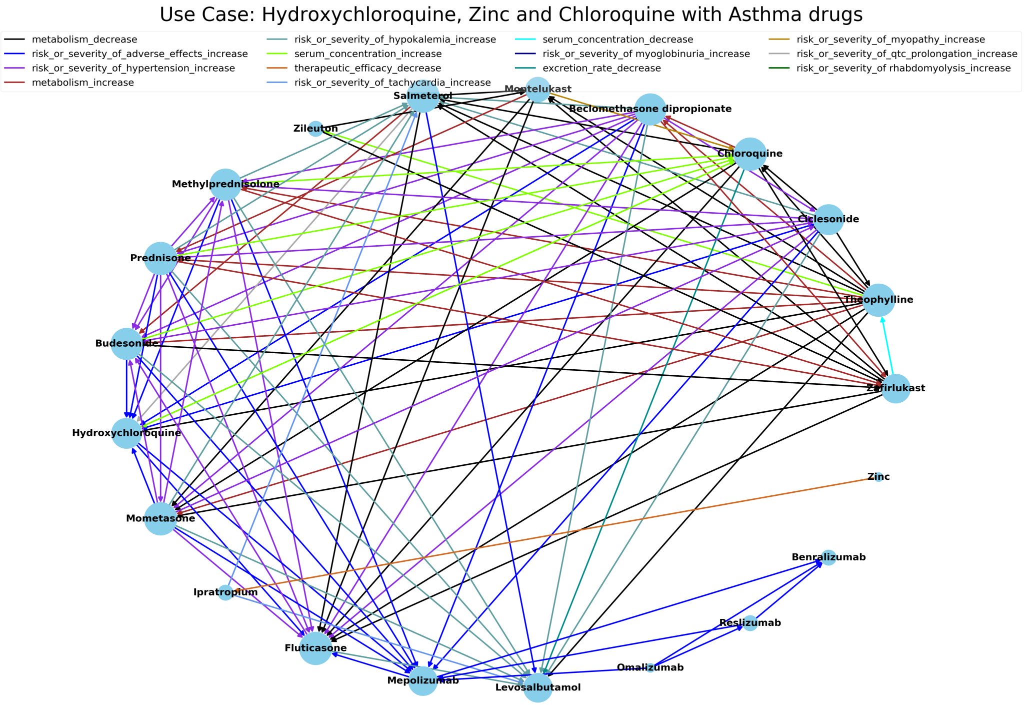

Use Case: Hydroxychloroquine, Zinc and Chloroquine with Asthma drugs
metabolism_decrease
risk_or_severity_of_adverse_effects_increase
risk_or_severity_of_hypertension_increase
metabolism_increase
risk_or_severity_of_hypokalemia_increase
serum_concentration_increase
therapeutic_efficacy_decrease
risk_or_severity_of_tachycardia_increase
serum_concentration_decrease
risk_or_severity_of myoglobinuria_increase
excretion_rate_decrease
risk_or_severity_of_myopathy_increase
risk_or_severity_of_qtc_prolongation_increase
risk_or_severity_of rhabdomyolysis_increase
Salmeterol
Montelukast
Beclomethasone dipropionate
Zileuton
Chloroquine
Methylprednisolone
Ciclesonide
Prednisone
Theophylline
Budesonide
Zafirlukast
Hydroxychloroquine
Zinc
Mometasone
Benralizumab
Ipratropium
Reslizumab
Fluticasone
Omalizumab
Mepolizumab
Levosalbutamol

Figure 6.4: The adverse effects generated as the result of the interactions among COVID-19 drugs (Hydroxychloroquine, Zinc, and Chloroquine) with treatments for Asthma. Relations retrieved from the Knowledge4COVID-19 KG

Figure 6.4, Figure 6.5, and Figure 6.6 depict adverse effects that can be triggered in COVID-19 patients who receive treatments for hypertension, asthma, or diabetes. Each plot reports a labeled directed graph; nodes represent drugs, and an edge between two drugs, represents an interaction. The label of an edge, denoted by the line color and the figure legend, indicates the type of side effect.

Figure 6.4 presents 14 types of drug-drug interactions that may occur among the COVID-19 drugs Hydroxychloriquine, Zinc, and Chloroquine, and asthma drugs. The pharmacokinetic drug-drug interactions between a pair of drugs, A and B, indicate that A impacts B's absorption, metabolism, and excretion when both drugs are administrated together. As a result, A may reduce the effectiveness or increase toxicities. The rest of the interactions are pharmacodynamic, i.e., their pharmacological outcome may be affected. Six out of the 14 reported drug-drug interactions are pharmacokinetic. Chloroquine may reduce the metabolism of Zafirlukast, Mometasone, and Fluticasone; it can also decrease the excretion rate of Levosalbutamol. Hydroxychloriquine also impacts the metabolism of Theophylline. Furthermore, the serum concentration of Chloroquine may be increased with asthma drugs by Methylprednisolone, Prednisone, and Budesonide. Thus, the

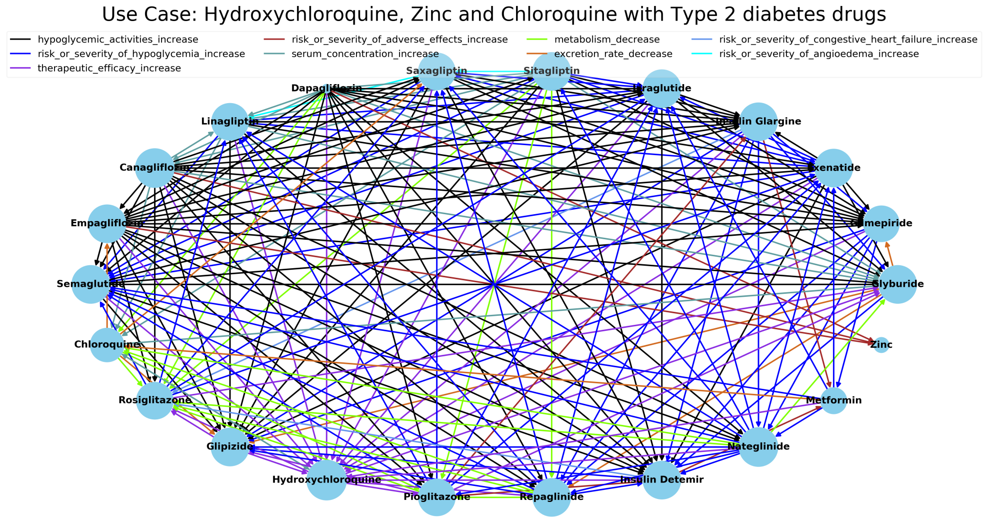

Use Case: Hydroxychloroquine, Zinc and Chloroquine with Type 2 diabetes drugs
hypoglycemic_activities_increase
risk_or_severity_of_hypoglycemia_increase
therapeutic_efficacy_increase
risk_or_severity_of_adverse_effects_increase
serum_concentration_increase
metabolism_decrease
excretion_rate_decrease
risk_or_severity_of_congestive_heart_failure_increase
risk_or_severity_of_angioedema_increase
Saxagliptin
Sitagliptin
Dapagliflozin
Liraglutide
Linagliptin
Insulin Glargine
Canagliflozin
Exenatide
Empagliflozin
Glimepiride
Semaglutide
Glyburide
Chloroquine
Zinc
Rosiglitazone
Metformin
Glipizide
Nateglinide
Hydroxychloroquine
Insulin Detemir
Pioglitazone
Repaglinide

Figure 6.5: The adverse effects generated as the result of the interactions among COVID-19 drugs (Hydroxychloroquine, Zinc, and Chloroquine) with treatments for Type 2 Diabetes. Relations retrieved from the Knowledge4COVID-19 KG

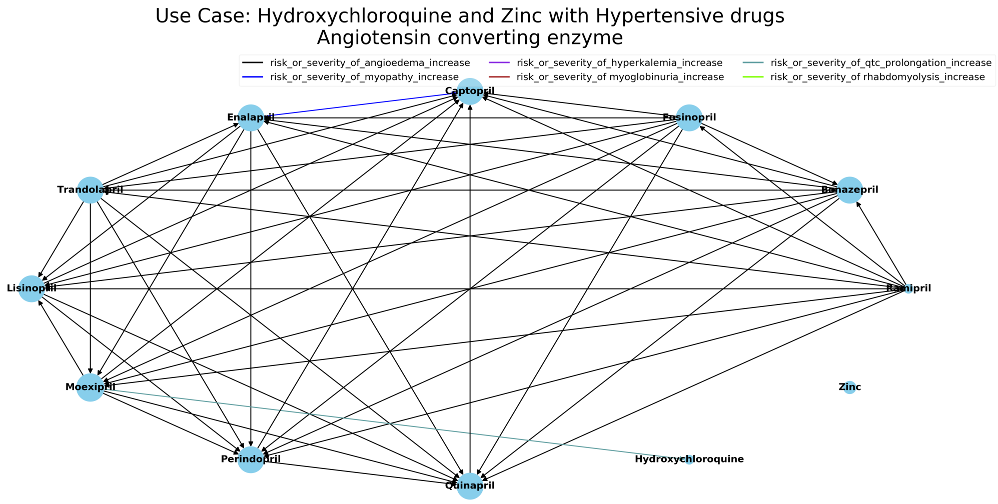
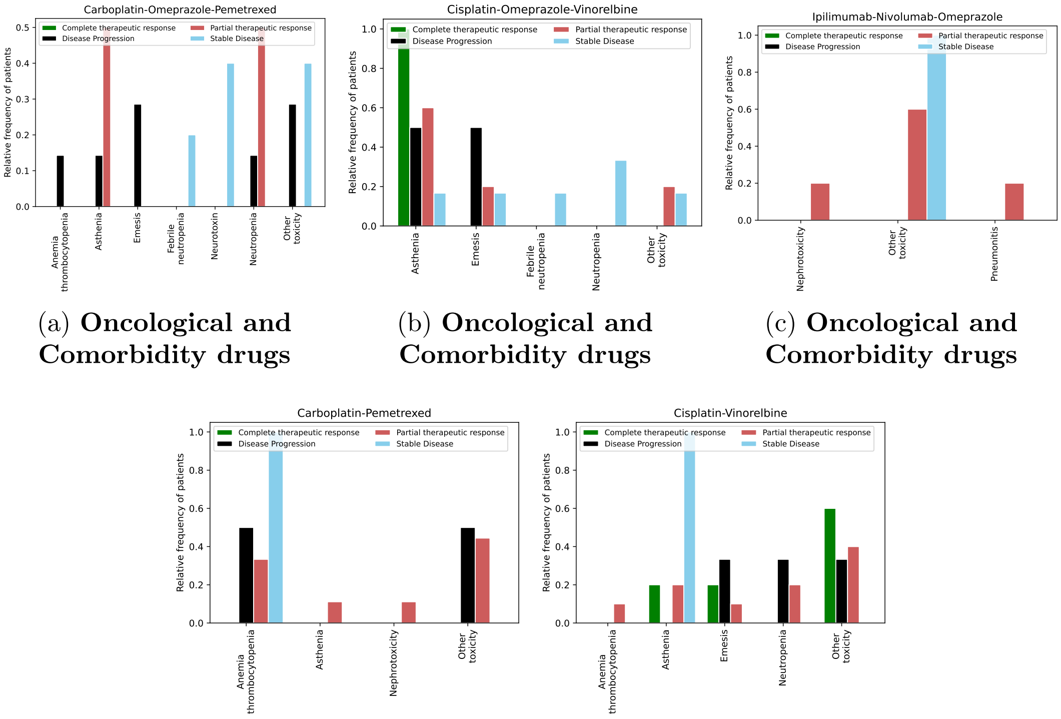

Use Case: Hydroxychloroquine and Zinc with Hypertensive drugs
Angiotensin converting enzyme
risk_or_severity_of_angioedema_increase
risk_or_severity_of_myopathy_increase
risk_or_severity_of_hyperkalemia_increase
risk_or_severity_of myoglobinuria_increase
risk_or_severity_of_qtc_prolongation_increase
risk_or_severity_of rhabdomyolysis_increase
Captopril
Enalapril
Fosinopril
Trandolapril
Benazepril
Lisinopril
Ramipril
Moexipril
Zinc
Perindopril
Hydroxychloroquine
Quinapril

Figure 6.6: The adverse effects generated as the result of the interactions among COVID-19 drugs (Hydroxychloroquine, Zinc, and Chloroquine) with treatments for Hypertension. Relations retrieved from the Knowledge4COVID-19 KG

effectiveness of the treatment was negatively affected. Four drugs may increase the severity of the side effects of Hydroxychloriquine. At the pharmacodynamic level, it can be observed that Montelukast and Chloroquine may increase the risk of myopathy, and Salmeterol and Hydroxychloriquine may increase the risk of QT prolongation. Since the risk of cardiac events during QT syndrome is high, these results suggest that the combinations of the treatments need to be administrated with great precaution. Similarly, Figure 6.5 reveals a more significant number of interactions among the drugs Hydroxychloriquine, Zinc, and Chloroquine and the drugs typically prescribed to Type 2 diabetes patients. All the drugs affect the efficacy of Hydroxychloriquine and the combination of Rosiglitazone in treatments with Insulin Determir or Insulin Glargine. Additionally, the therapeutic efficacy of Rosiglitazone can be increased when used in combination with Hydroxychloroquine, and Chloroquine may reduce the effectiveness of Metformin. They should be administrated with precaution because their therapeutic efficacy may be reduced. Drug interactions of hypertension treatments based on drugs Angiotensin-converting enzyme, with the drugs Hydroxychloriquine and Zinc, are reported in Figure 6.6. As reported, the combination of these drugs may cause pharmacodynamic interactions that can critically affect the function of nerve and muscle cells, including those in the heart. The above results suggest that COVID-19 patients receiving treatments for pre-existing conditions need to be carefully treated.

6.3 Assessment of the Impact of DDI in the Effectiveness of the Lung Cancer Treatments

In this section, we evaluate the impact of DDIs on the effectiveness of lung cancer treatments registered on a knowledge-driven data ecosystem (DE) named DE4LungCancer. DE4LungCancer has been applied in the context of iASiS[5], BigMedilytics[6], P4-LUCAT[7], and EU H2020 CLARIFY[8]. The knowledge represented in the DE4LungCancer KG is exploited to understand the impact of the interactions between a treatment's drugs on the effectiveness of the treatment. The evaluation of treatments' effectiveness is performed based on the number of toxicities observed in the lung cancer patients and the assessment of a treatment's response provided by the patients' oncologists; these results are part of the clinical records processed by the Clinical DE and integrated into the DE4LungCancer KG. The DDIs in treatment are computed based on three computational methods. The first method (*DrugBank*) computes the number of DDIs in treatment based on the DDIs reported on DrugBank. We extracted the DDIs from DrugBank and included them in our DE4LungCancer KG. The second method (*DS*) proposed by Rivas and Vidal [103] deduces new DDIs based on a deductive system implemented in Datalog on top of KG. The third method (DDI-BLKG) proposed by Bougiatiotis et.al. [20] predicts DDIs based on scientific publications; in this section, we named *Literature*.



(a) **Oncological and Comorbidity drugs**

(b) **Oncological and Comorbidity drugs**

(c) **Oncological and Comorbidity drugs**

(d) **Oncological drugs**

(e) **Oncological drugs**

Figure 6.7: **Toxicity analysis of oncological treatments.** Figure 6.7 shows five bar plots of the toxicities produced by treatments in lung cancer patients. The treatment responses are differentiated by color. The oncological treatments with comorbidity drugs generate more toxicities than those without comorbidity drugs.

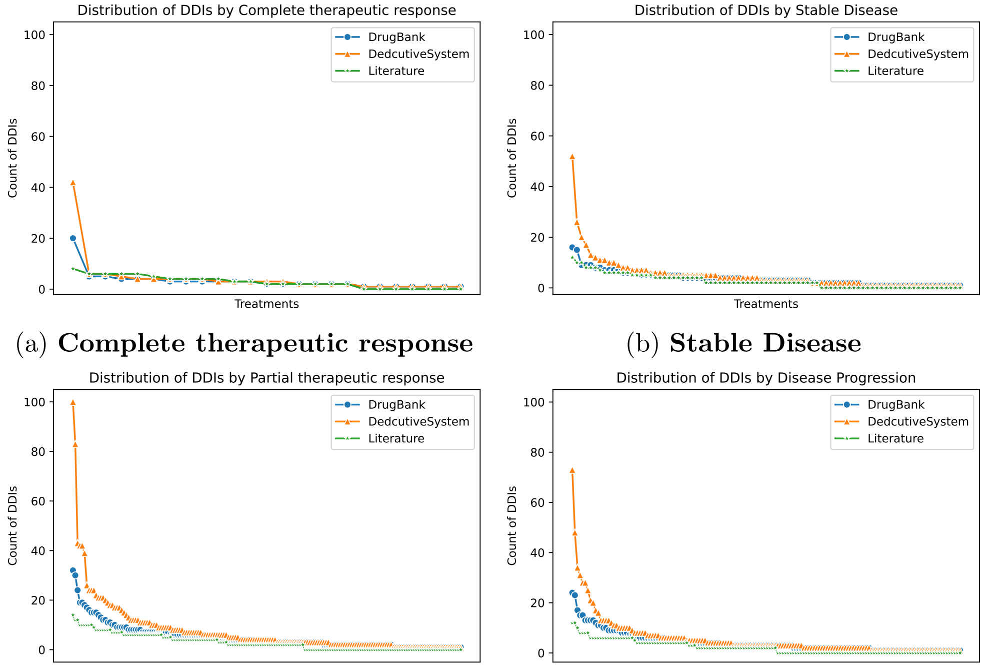
6.3.1 Treatment Toxicity Analysis

We have selected the most frequent oncological treatments for analyzing their toxicities from DE4LungCancer KG. The treatments in Figure 6.7a, Figure 6.7b, and Figure 6.7c contain oncological and comorbidity drugs. Figure 6.7d and Figure 6.7e show the same treatments as Figure 6.7a and Figure 6.7b without comorbidity drugs. The x-axes represent the toxicities of patients receiving the treatment, and the y-axes are the relative frequency of patients having toxicity. The treatment responses are evaluated in four categories: *complete therapeutic response, stable disease, partial therapeutic response*, and *disease progression*, where a *complete therapeutic response* is the desired response, and *disease progression* is the worst expected response. We observed that oncology treatments without comorbidity drugs cause fewer toxicities in patients than oncology treatments together with comorbidity drugs. In addition, for patients taking the treatment represented in

Figure 6.7c without comorbidity drugs, no toxicity was caused. Furthermore, the patients with a *complete therapeutic response* have fewer toxicities than the other treatment response.

6.3.2 DDIs Analysis

This section describes the adverse effect of Covid-19 treatments retrieved on top of the Knowledge4COVID-19 KG. This section describes the adverse effect of Covid-19 treatments retrieved on top of the Knowledge4COVID-19 KG. We have extracted from DE4LungCancer KG the lung cancer treatments with their respective responses. Our purpose is to compute the distribution of DDIs for each treatment response. The hypothesis is that treatments with a *complete therapeutic response* or *stable disease* have fewer DDIs than treatments with *partial therapeutic response* and *disease progression*. The data are processed to have the treatments in four disjoint sets of treatment responses. For treatments with different responses, the most frequent response is selected. Thus, each treatment is classified into a single response class. The DDIs for each treatment are computed based on the DDIs reported on DrugBank, and two computational methods.



(a) **Complete therapeutic response**

(b) **Stable Disease**

(c) **Partial therapeutic response**

(d) **Disease Progression**

Figure 6.8: **Distribution of DDIs by treatment response.**

Figure 6.8 shows the distribution of DDIs by each treatment response. The x-axis represents each treatment, and the y-axis represents the count of DDIs in treatment. The three color lines on the plots represent the three methods employed to compute the DDIs. We observe that for the three methods used, the distribution of DDIs for treatments with a *complete therapeutic response* (Figure 6.8a) or *stable disease* (Figure 6.8b) have fewer DDIs than treatments with *partial therapeutic response* (Figure 6.8c) and *disease progression* (Figure 6.8d), corroborating our hypothesis.

6.3.3 Correlation Analysis between DDIs and Treatment Responses

This section describes the adverse effect of Covid-19 treatments retrieved on top of the Knowledge4COVID-19 KG. We are interested in computing the correlation between DDI in treatment and the number of patients with a specific response to the treatment. The treatment responses are evaluated in four categories: *complete therapeutic response* and *stable disease* are positive responses to treatment, while *partial therapeutic response* and *disease progression* are negative responses. Our hypothesis is to detect a negative correlation between DDI in treatment and the number of patients with *complete therapeutic response* or *stable disease*. A negative correlation, in this case, means more patients with positive responses and less DDI in the treatment. Moreover, we expect to identify a positive correlation between DDI in treatment and the number of patients with a *partial therapeutic response* or *disease progression*. We have extracted the lung cancer treatments with their respective response from DE4LungCancer KG. Then, the number of DDIs for each treatment is computed based on the DDIs reported on DrugBank, and two computational methods, *DS* and *Literature*. Also, we compute the number of patients by treatment response for each treatment. Finally, we perform a spearman correlation analysis between the four therapeutic responses and the three computational methods for computing the DDIs.

Table 6.3: Spearman correlation coefficient analysis between DDIs and responses over DE4LungCancer KG. Complete therapeutic response (CTR), Stable Disease (SD), Partial therapeutic response (PTR), Disease Progression (DP).

Response	**DrugBank**		**DS**		**Literature**	
	correlation	p-value	correlation	p-value	correlation	p-value
CTR	-0.31658	0.11509	-0.30451	0.13041	0.18642	0.36187
SD	-0.20782	0.09150	-0.21407	0.08194	-0.09353	0.45156
PTR	-0.33183	0.00027	-0.32374	0.00039	-0.29062	0.00155
DP	-0.38461	0.00018	-0.39093	0.00014	-0.25746	0.01429

Table 6.3 shows spearman's results of the correlation analysis. We can observe a negative correlation for all the combinations between treatment responses and DDI methods except for *complete therapeutic response* and DDI based on Literature but with a high p-value. Considering the data on DE4LungCancer KG, we do not identify a positive correlation between the number of DDIs in treatment and the number of patients with a *partial therapeutic response* or *disease progression*.

6.3.4 Correlation Analysis between Drugs and DDIs in Treatment

We analyzed the correlation in lung cancer treatments between the number of drugs and the number of DDIs. The hypothesis is that increasing the number of drugs in treatment increases the number of DDIs. Therefore, a positive correlation should be identified. We retrieved the lung cancer treatments from DE4LungCancer KG. Then, we counted the number of drugs by treatment. The number of DDIs for each treatment is computed based on the drug-drug interactions reported by the three following computational methods *DrugBank*, *DS*, and *Literature*. Table 6.4 illustrates the strong positive correlation between the number of drugs and the number of DDIs in treatments, i.e., the higher the number of drugs in treatment, the higher the number of treatment interactions. Although the Spearman correlation coefficient for the *Literature* method is low, it exhibits a positive correlation.

Table 6.4: Spearman correlation coefficient analysis between the number of drugs in the treatment and the number of DDIs among these drugs.

DrugBank		**DS**		**Literature**	
correlation	p-value	correlation	p-value	correlation	p-value
0.75418	1.89e-21	0.76469	2.44e-22	0.13050	0.24860

6.4 Evaluation and Knowledge Discovery over the IASIS KG

We propose a framework that resorts to computational extraction methods for mining knowledge from data sources, e.g., clinical notes, images, or scientific publications. The proposed framework is used in the context of the EU H2020-funded project iASiS [5] with the aim of paving the way for accurate diagnostics and personalized treatments. Knowledge discovery techniques are used to uncover patterns in the iASiS knowledge graph. Patterns include common characteristics of patients depending on their toxic habits, familial antecedents, or comorbidities.

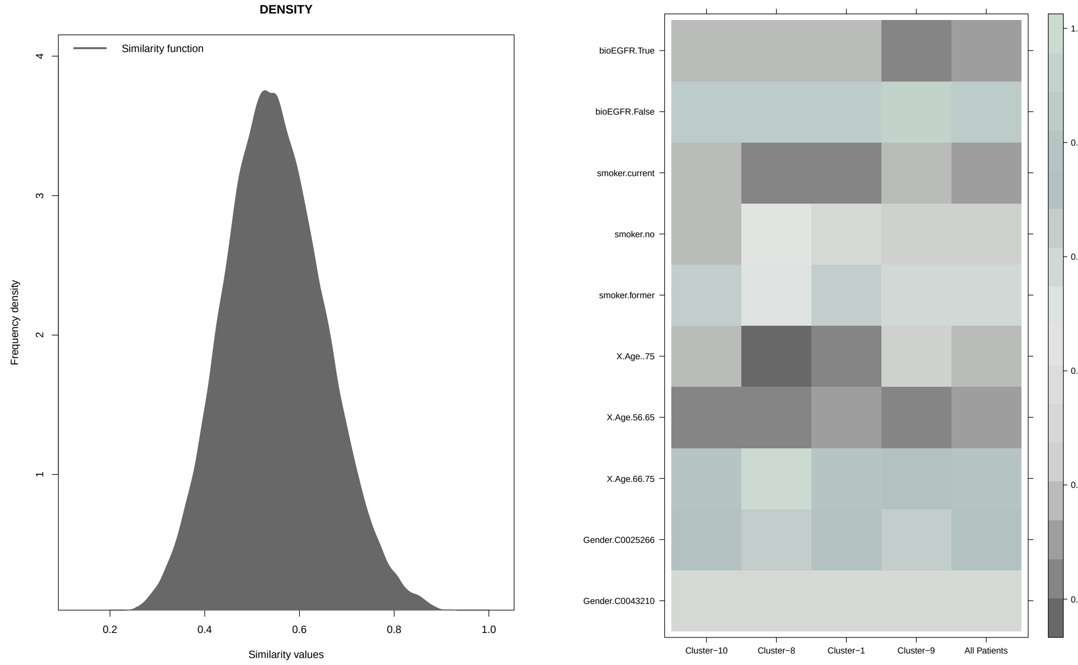
We define a similarity measure as a function that quantifies the similarity between two patients. The patient similarity combines similarity values of the main characteristics of the two patients: age, gender, mutated genes, toxic habits, the evolution of a tumor, the mutations, and the patient performance status (ecog). Similarity values between these characteristics are computed based on different similarity measures: i) Lists are compared using Spearman's rho while the Jaccard similarity coefficient is utilized for sets; ii) similarity between drugs is computed based on the chemical structure of the drugs (SIMCOMP) [9]; iii) side effects are compared using the Human Phenotype Ontology similarity (HPOSim)[10], and iv) The UMLS similarity measure[11] is used for UMLS terms.

The similarity values are combined in terms of a triangular norm. Figure 6.9a depicts the density distribution of the similarity values for pairs of lung cancer patients in the iASiS KG. We can observe that a considerably large portion of the patient population has relatively high values of similarity, suggesting that a large number of patients have similar reactions to the prescribed treatments. Further analysis with clinical partners is required to validate the meaning of observed values of similarity. Furthermore, we apply community detection algorithms to discover patterns between patients that share similar properties in the iASiS knowledge graph. We resort to semEP (Semantics Based Edge Partitioning Problem) [94] for computing patients' communities based on the similarity values. It creates a minimal partitioning of the input graph, such that the density of each community is maximal. The community density represents the degree of similarity of the entities in a community. Figure 6.9b reports on the results of computing semEP against the iASiS knowledge graph. The main properties of the patients involve mutations of lung cancer-related genes, e.g., EGFR; demographic attributes, smoking habits, treatments, and tumor stages. The studied population is composed of 739 patients. The goal of the study is to identify the four communities of patients– out of 13 communities – with characteristics that differed from the whole population; the Kolmogorov-Smirnov test was used to rank the communities. Figure 6.9b reports on four communities of patients; using a heatmap plot, the percentage of patients in each community or cluster is described in terms of age, gender, EGFR mutation, and smoking habits. For example, patients in Cluster-1 are not current smokers, and a considerable number of them are non-smokers; in addition, the biomarker EGFR is negative for many of them. The results are initial and require further

[9] http://www.genome.jp/tools/simcomp/

[10] https://sourceforge.net/projects/hposim/

[11] http://www.d.umn.edu/~tpederse/umls-similarity.html



(a) Density Distribution Patient Similarity (b) Communities of Lung Cancer Patients

Figure 6.9: **Knowledge Analytics**. (a) A function able to quantify the similarity between two lung cancer patients is described in terms of frequency density; the function takes into account treatments and the evolution of the tumors, mutations, and patient performance. The reported results suggest that a large number of patients react similarly to the treatments. However, more studies are required to validate this observation. (b) Communities of lung cancer patients and the summary of the observed features age, toxic habits, and EGFR mutations. Distributions of the observed features differ from the whole population, enabling the study of patients with unique characteristics.

study from the clinical partners of the project. However, they suggest that these techniques have the power to uncover patterns in the observed features of patients.

6.5 Summary

In this chapter, we present the applicability of our deductive database system to real scenarios. The proposed deductive database makes implicit knowledge explicit deducing new relationships *DS* relies on known relations between drugs to deduce the relationships encoded in a set of rules through a Datalog program. The deduction of wedges on top of the knowledge graph enables uncovering of combinations of drugs whose interactions may reduce the effectiveness of treatment and to compute the interaction score of drugs in treatment. The interaction score is proving particularly useful for clinicians. Furthermore, *DS* reduces the data sparsity issue, enabling the knowledge graph to become meaningful in the discovery task, e.g., we observe that treatments with *partial therapeutic response* or *disease progression* have more DDI than treatments with *complete therapeutic response* or *stable disease*. We show the benefit of our approach on Knowledge4COVID-19 KG and lung cancer treatments by computing the DDI-Reduction Percentage and interaction score of drugs in treatment. Moreover, we illustrate the benefit of *DS* evaluating the effectiveness of the lung cancer treatments on top of DE4LungCancer KG. In addition, we provide a framework that resorts to computational extraction methods for mining knowledge from data sources. The proposed framework is used in iASiS project. We define a similarity measure that quantifies the similarity between two lung cancer patients. Then, we apply community detection algorithms to discover patterns between patients that share similar properties in the iASiS knowledge graph. We identified communities of patients with characteristics that differed from the whole population. Overall, the results indicate that our method can be of significant relevance in all these applications.

Chapter 7

Conclusions and Future Directions

In this thesis, we studied the problem of knowledge discovery over knowledge graphs built from heterogeneous data sources. We proposed a Neuro-Symbolic Artificial Intelligence approach that discovers knowledge given a target prediction in a knowledge graph and predicts unknown links in the KG. In particular, the discussion of the research problem, research questions, challenges, and contributions to address the challenges are presented in Chapter 1. Fundamental background concepts are examined in Chapter 2. An overview of state-of-the-art approaches related to the problem tackled in this thesis is analyzed in Chapter 3. Then, the following three chapters, Chapter 4, Chapter 5, and Chapter 6, describe and evaluate the proposed solution to the challenges in discovering knowledge over knowledge graphs. Finally, in this Chapter 7, we review the research questions and examine the achieved results. Furthermore, we examine the limitations of the work and outline possible future directions for future work.

7.1 Revising the Research Questions

RQ1: How can metadata encoding data meaning be exploited to discover relationships in knowledge graphs?

To answer this research question, we proposed a neuro-symbolic artificial intelligence approach over KGs that attempts to combine symbolic and sub-symbolic AI models. In Chapter 4 and Chapter 5 this research question is addressed. We show a domain-agnostic approach able to capture the implicit knowledge in a KG by a symbolic system and enhance the predictive capacity of sub-symbolic systems. The symbolic system is implemented by deductive databases defined for an abstract target prediction over a knowledge graph. Our proposed solution builds the ego networks of the entities that correspond to the domain and range of the abstract target prediction to deduce new relationships enhancing the ego networks. Sub-symbolic systems benefit from improved ego networks and perform the link prediction problem. The approach assumes that a link prediction problem is defined in terms of an abstract target prediction over a KG. We evaluate the approach in the biomedical domain, predicting polypharmacy treatment effectiveness and the Industry 4.0 context, demonstrating its effectiveness in determining relatedness among standards and analyzing their properties to detect unknown relations. We observed that by exploiting the data and metadata that encode the meaning of the data with a symbolic system, we deduce relationships that improve the behavior of the KGE models.

RQ2: How can heterogeneous data sources be integrated to obtain a unified knowledge representation?

Chapter 4 answers this research question using a knowledge graph approach that considers the metadata describing the semantics encoded in the data in the biomedical context. We integrated treatments, their prescribed drugs, drug-drug interactions, drug-protein interactions, publications related to the drug-protein interactions, and the gene that encodes the proteins in a knowledge graph. The knowledge graph of polypharmacy treatment responses is populated with descriptions of 548 oncological treatments. Overall, the final knowledge graph is a unified knowledge representation where we overcome the integration challenges. Integration tackles missing associations and incompleteness of heterogeneous data with open data sources. The knowledge graph was linked to existing open web sources such as DrugBank[2], Wikidata[3], Uniprot[4], DBpedia[9], and Pubmed[10]. The built KG provides the advantage of retrieving new knowledge from open web sources using the owl:sameAs OWL property via a federated query engine.

RQ3: How can implicit knowledge be used to enhance knowledge discovery tasks?

This research question is addressed in Chapter 4 and Chapter 5. We presented a formalization of the symbolic system DS that relies on existing approaches of deductive database systems. DS is proposed for an abstract target prediction over a knowledge graph that can derive new statements; it concludes new facts, from inference rules and facts stored in the extensional database. The EDB of DS comprises ground facts of the form $s(p, o)$. The triples $\langle s, p, o\rangle$ belong to the ego network v or the relations between their neighbors, where v belongs to the domain or range of the abstract target prediction. The IDB of *DS* contains rules that allow deducing new relationships in the ego networks of the target prediction. In summary, the deductive system *DS* minimizes the data sparsity issues by adding the deduced implicit relationships in the KG. The new implicit relationships incorporated into the KG are related to the abstract target prediction and represent relevant relations to the predictive task.

RQ4: What is the impact of deductive reasoning on accurately uncovering knowledge?

This research question is addressed in Chapter 4. We empirically evaluate the effectiveness of our neuro-symbolic artificial intelligence approach. We conduct an ablation study on the components of our system, considering state-of-the-art KG embedding methods. The deductive system *DS* reduces the data sparsity issues in the knowledge graph, enhancing the KGE methods in the link prediction task. We evaluate using deductive reasoning, non-reasoning, and randomly adding the same number of links deduced. The observed results provide evidence of the advantages of our approach in improving the state-of-the-art KG embedding methods analyzed. *DS* deduces new relationships accurately that represent implicit and explicit knowledge.

RQ5: How can the proposed approach be applied to real-world cases?

Chapter 6 answers this research question by applying our symbolic system over three KGs, DE4LungCancer KG [3], Knowledge4COVID-19 KG [115], and iASiS KG [131]. We illustrate the applicability of our approach in four projects in the biomedical context, iASiS[5], BigMedilytics[6], P4-LUCAT[7], and H2020 CLARIFY[8]. We presented the *DS* to deduce DDIs and compute the interaction score of drugs in treatment based on the wedge concept in different diseases. We show the potential of *DS* for discovering patterns that can enable the explanation of treatment interactions and patient characterization. Overall, we presented the benefit of the discovery task on a knowledge graph.

7.2 Limitations

Despite the overall achieved research objectives, we acknowledge that there are limitations that have yet to be covered in the scope of this thesis. First, we are not applying a semantic reasoner to take advantage of the Web Ontology Language (OWL) reasoning capabilities, including RDF/RDFS reasoning capabilities. OWL allows expressing other schema definitions in RDF, e.g., expressing the equality of individuals, owl:sameAs, or equivalence or disjointness of properties and classes, owl:equivalentClass, owl:equivalentProperty, owl:disjointWith, owl:propertyDisjointWith. Thus, reasoning can be done on the set of OWL properties, and the KG can be enriched with new implicit triples. The second limitation is that the approach assumes the user can write the rules representing the experts' knowledge. Third, the sub-symbolic systems, i.e., KGE models, considered in our approach do not distinguish between the data and the metadata represented in the KG, where metadata describes the data by defining classes and properties. KGE models assume all the triples in the KG as data while considering the metadata in the KGE model scoring function can help improve predictions.

7.3 Future Directions

Based on our findings, and the contributions made in this thesis, we now present some of the future directions of this work for the research community:

- KGE models define a scoring function $\phi(h, r, t)$ for estimating the plausibility of a triple $\langle h, r, t\rangle$ based on the embeddings of their elements. The scoring function of the KGE models relies on statistical reasoning, e.g., neural networks, tensor decomposition, or geometric models. Although the results are convincing, their inference mechanisms of KGE suffer from low interpretability caused by high dimensionality. We envision extending our approach where the scoring function of the KGE models considers the DS to assess the plausibility of a triple. This neuro-symbolic combination can significantly improve the interpretability of how a given model reached a particular response.

- Exploit semantic reasoning to leverage the capabilities of OWL. The properties of OWL to express equality between entities as well as to express equivalence or disjointness of classes and properties allow reasoning new relationships and enrich the knowledge graph.

- Extend the symbolic system with algorithms of learning rules from KGs. The rule learning algorithm requires implementing a strategy to evaluate the quality of the rules mined from the knowledge graph, such that meaningful rules are generated and combinatorial explosion is avoided.

7.4 Closing Remarks

With the growing amount of heterogeneous data in the ongoing digitization process, the problem of knowledge discovery is constantly facing new prospects and challenges. In this thesis, we have shown the benefits of semantic knowledge integration to successfully tackle the problem of uncovering helpful knowledge. We have integrated data and concepts semantically and provide a neuro-symbolic AI approach, enabling the uncovering of relevant knowledge. The symbolic system deduces implicit knowledge and makes it explicit in the KG, alleviating data sparsity issues and enhancing KGE models. Additionally, the proposed approach in this thesis is applied in four projects, demonstrating the significant relevance in all these applications. Future research work can build upon the presented contributions to devise more interpretable and comprehensive discovery approaches.

Appendices

Appendix A

List of Publications

Papers in Proceedings of Peer-Reviewed Conferences

- **Ariam Rivas**, Maria-Esther Vidal: *Capturing Knowledge about Drug-Drug Interactions to Enhance Treatment Effectiveness.* K-CAP '21: Proceedings of the 11th on Knowledge Capture Conference (2021). *Nominated to the best student paper.*

- **Ariam Rivas**, Irlan Grangel-Gonzalez, Diego Collarana, Jens Lehmann, and Maria-Esther Vidal: *Unveiling Relations in the Industry 4.0 Standards Landscape Based on Knowledge Graph Embeddings.* In Proceeding of the 31st International Conference of Database and Expert Systems Applications (DEXA 2020).

Peer-Reviewed International Journals

- **Ariam Rivas**, Diego Collarana, Maria Torrente, Maria-Esther Vidal. *A Neuro-Symbolic System over Knowledge Graphs for Link Prediction.* In: Semantic Web Journal (2022). (Under-Review).

- Fotis Aisopos, Samaneh Jozashoori, Emetis Niazmand, Disha Purohit, **Ariam Rivas**, Ahmad Sakor, Enrique Iglesias, Dimitrios Vogiatzis, Ernestina Menasalvas, Alejandro Rodriguez Gonzalez, Guillermo Vigueras, Daniel Gomez Bravo, Maria Torrente, Roberto Hernández López, Mariano Provencio Pulla, Athanasios Dalianis, Ana Triantafillou, Georgios Paliouras and Maria-Esther Vidal. *Knowledge Graphs for Enhancing Transparency in Health Data Ecosystems.* In: Semantic Web Journal (2022). (Under-Review)

- Alejandro Rodriguez Gonzalez, Ernestina Menasalvas, Fotis Aisopos, Dimitrios Vogiatzis, Anastasia Krithara, Georgios Paliouras, Samaneh Jozashoori, **Ariam Rivas**, Ahmad Sakor, Maria-Esther Vidal, Maria Torrente, Mariano Provencio Pulla, Anna Trinatafyllou, Athanasios Dalianis. *Lung Cancer Pilot.* Book Chapter. 2022 (Under-Review).

- Ahmad Sakor, Samaneh Jozashoori, Emetis Niazmand, **Ariam Rivas**, Konstantinos Bougiatiotis, Fotis Aisopos, Enrique Iglesias, Philipp D. Rohde, Trupti Padiya, Anastasia Krithara, Georgios Paliouras, Maria-Esther Vidal: *Knowledge4COVID-19: A Semantic-based Approach for Constructing a COVID-19 related Knowledge Graph from Various Sources and Analysing Treatments' Toxicities.* Journal of Web Semantics (2022).

- **Ariam Rivas**, Irlan Grangel-Gonzalez, Diego Collarana, Jens Lehmann, and Maria-Esther Vidal: *Discover Relations in the Industry 4.0 Standards Via Unsupervised Learning on Knowledge Graph Embeddings.* Journal of Data Intelligence (2020).

- Maria-Esther Vidal, Kemele M. Endris, Samaneh Jazashoori, Ahmad Sakor, **Ariam Rivas**: *Transforming Heterogeneous Data into Knowledge for Personalized Treatments - A Use Case.* Datenbank-Spektrum volume 19, pages 95–106 (2019).

Bibliography

[1] H.M.A. Abraham, C.M. White, and W.B. White. “The Comparative Efficacy and Safety of the Angiotensin Receptor Blockers in the Management of Hypertension and Other Cardiovascular Diseases.” In: *Drug Saf* 38 (2015), pp. 33–54. DOI: https://doi.org/10.1007/s40264-014-0239-7.

[2] Peter Adolphs et al. *Structure of the Administration Shell*. Status Report. ZVEI and VDI, 2016.

[3] Fotis Aisopos et al. “Knowledge Graphs for Enhancing Transparency in Health Data Ecosystems”. In: 2022. URL: https://www.semantic-web-journal.net/system/files/swj3294.pdf.

[4] Mehdi Ali et al. “The KEEN Universe: An ecosystem for knowledge graph embeddings with a focus on reproducibility and transferability”. (in press). 2020.

[5] Diego Ardila et al. “End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography”. In: *Nature medicine* 25.6 (2019), pp. 954–961.

[6] Marcelo Arenas, Claudio Gutierrez, and Jorge Pérez. “Foundations of RDF Databases”. In: *Reasoning Web. Semantic Technologies for Information Systems: 5th International Summer School 2009, Brixen-Bressanone, Italy, August 30 - September 4, 2009, Tutorial Lectures*. Ed. by Sergio Tessaris et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 158–204. ISBN: 978-3-642-03754-2. DOI: 10.1007/978-3-642-03754-2_4. URL: https://doi.org/10.1007/978-3-642-03754-2_4.

[7] David Arthur and Sergei Vassilvitskii. “K-means++: The Advantages of Careful Seeding”. In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA ’07. New Orleans, Louisiana: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035. ISBN: 978-0-898716-24-5. URL: http://dl.acm.org/citation.cfm?id=1283383.1283494.

[8] Franz Baader, Ian Horrocks, and Ulrike Sattler. “Description Logics”. In: *Handbook on Ontologies*. Ed. by Steffen Staab and Rudi Studer. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 3–28. ISBN: 978-3-540-24750-0. DOI: 10.1007/978-3-540-24750-0_1. URL: https://doi.org/10.1007/978-3-540-24750-0_1.

[9] Sebastian R. Bader et al. “A Knowledge Graph for Industry 4.0”. In: *The Semantic Web* 12123 (2020), pp. 465–480.

[10] Sebastian R. Bader et al. “Structuring the Industry 4.0 Landscape”. In: *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. 2019, pp. 224–231. DOI: 10.1109/ETFA.2019.8869268.

[11] Tim Berners-Lee. *Linked Data*. URL: https://www.w3.org/DesignIssues/LinkedData.html.

[12] Tim Berners-Lee, James Hendler, and Ora Lassila. "The Semantic Web". In: *Scientific American* 284.5 (May 2001), pp. 34–43. URL: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

[13] Abraham Bernstein, James Hendler, and Natalya Noy. "A New Look at the Semantic Web". In: *Commun. ACM* 59.9 (Aug. 2016), pp. 35–37. ISSN: 0001-0782. DOI: 10.1145/2890489. URL: https://doi.org/10.1145/2890489.

[14] Tarek R. Besold et al. "Chapter 1. Neural-Symbolic Learning and Reasoning: A Survey and Interpretation1". In: *Neuro-Symbolic Artificial Intelligence: The State of the Art* (Dec. 2021). ISSN: 1879-8314. DOI: 10.3233/faia210348. URL: http://dx.doi.org/10.3233/faia210348.

[15] Christian Bizer, Tom Heath, and Tim Berners-Lee. "Linked Data - The Story So Far". In: *Int. J. Semantic Web Inf. Syst.* 5 (2009), pp. 1–22.

[16] Piero A. Bonatti and Daniel Olmedilla. "Rule-Based Policy Representation and Reasoning for the Semantic Web". In: *Reasoning Web: Third International Summer School 2007, Dresden, Germany, September 3-7, 2007, Tutorial Lectures*. Ed. by Grigoris Antoniou et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 240–268. ISBN: 978-3-540-74615-7. DOI: 10.1007/978-3-540-74615-7_4. URL: https://doi.org/10.1007/978-3-540-74615-7_4.

[17] Antoine Bordes et al. "A Semantic Matching Energy Function for Learning with Multi-relational Data". In: *Machine Learning* (2014), pp. 1–30. DOI: 10.1007/s10994-013-5363-6. URL: https://hal.archives-ouvertes.fr/hal-00835282.

[18] Antoine Bordes et al. "Learning Structured Embeddings of Knowledge Bases". In: *25th Conference on Artificial Intelligence (AAAI)*. San Francisco, United States, Aug. 2011, pp. 301–306. URL: https://hal.archives-ouvertes.fr/hal-00752498.

[19] Antoine Bordes et al. "Translating Embeddings for Modeling Multi-relational Data". In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges et al. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.

[20] Konstantinos Bougiatiotis et al. "Drug-Drug Interaction Prediction on a Biomedical Literature Knowledge Graph". In: *International Conference on Artificial Intelligence in Medicine*. 2020.

[21] Aydin Buluç et al. "Recent Advances in Graph Partitioning". In: *Algorithm Engineering - Selected Results and Surveys*. 2016, pp. 117–158. DOI: 10.1007/978-3-319-49487-6_4. URL: https://doi.org/10.1007/978-3-319-49487-6%5C_4.

[22] S. Ceri, G. Gottlob, and L. Tanca. "What you always wanted to know about Datalog (and never dared to ask)". In: *IEEE Transactions on Knowledge and Data Engineering* 1.1 (1989), pp. 146–166. DOI: 10.1109/69.43410.

[23] Xiaojun Chen, Shengbin Jia, and Yang Xiang. "A review: Knowledge reasoning over knowledge graph". In: *Expert Systems with Applications* 141 (2020), p. 112948. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2019.112948. URL: https://www.sciencedirect.com/science/article/pii/S0957417419306669.

[24] Jin Cheng et al. "Predicting treatment response from longitudinal images using multi-task deep learning". In: *Nature Communications* 12.1 (2021). DOI: 10.1038/s41467-021-22188-y.

[25] Sasank Chilamkurthy et al. "Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study". In: *The Lancet* 392.10162 (2018), pp. 2388–2396.

[26] Nitishal Chungoora et al. "Towards the ontology-based consolidation of production-centric standards". In: (Jan. 2013). URL: https://repository.lboro.ac.uk/articles/journal_contribution/Towards_the_ontology-based_consolidation_of_production-centric_standards/9573980.

[27] Philipp Cimiano and Heiko Paulheim. "Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods". In: *Semant. Web* 8.3 (Jan. 2017), pp. 489–508. ISSN: 1570-0844. DOI: 10.3233/SW-160218. URL: https://doi.org/10.3233/SW-160218.

[28] Steven Harris, Andy Seaborne, and Eric Prud'hommeaux. *Chloroquine or Hydroxychloroquine and/or Azithromycin.* 2021. URL: https://www.covid19treatmentguidelines.nih.gov/therapies/antiviral-therapy/chloroquine-or-hydroxychloroquine-and-or-azithromycin/.

[29] Emmanuel Darmois et al. *Digitising the Industry - Internet of Things Connecting the Physical, Digital and Virtual Worlds.* River Publishers, 2016. ISBN: 9788793379817. DOI: https://doi.org/10.13052/rp-9788793379824. URL: https://www.riverpublishers.com/pdf/ebook/RE_E9788793379824.pdf.

[30] Watts Devon et al. "Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis". English. In: *Translational Psychiatry* 12.1 (2022). URL: https://www.proquest.com/scholarly-journals/predicting-treatment-response-using-eeg-major/docview/2701336455/se-2.

[31] Irawaty Djaharuddin et al. "Comorbidities and mortality in COVID-19 patients". In: vol. 35. 2021. DOI: 10.1016/j.gaceta.2021.10.085.

[32] Xin Dong et al. "Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '14. New York, New York, USA: Association for Computing Machinery, 2014, pp. 601–610. ISBN: 9781450329569. DOI: 10.1145/2623330.2623623. URL: https://doi.org/10.1145/2623330.2623623.

[33] Hasan Ejaz et al. "COVID-19 and comorbidities: Deleterious impact on infected patients". In: *Journal of infection and public health* (2020).

[34] Prud'hommeaux Eric and Seaborne Andy. *SPARQL Query Language for RDF.* URL: https://www.w3.org/TR/rdf-sparql-query/.

[35] H.K.G. Fernlund et al. "Learning tactical human behavior through observation of human performance". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 36.1 (2006), pp. 128–140. DOI: 10.1109/TSMCB.2005.855568.

[36] Achille Fokoue et al. "Predicting Drug-Drug Interactions Through Large-Scale Similarity-Based Link Prediction". In: *The Semantic Web. Latest Advances and New Domains.* Springer International Publishing, 2016. ISBN: 978-3-319-34129-3.

[37] Marco Gaertler and Thomas Erlebach. "Clustering". In: *Network Analysis: Methodological Foundations*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 178–215. ISBN: 978-3-540-31955-9. DOI: 10.1007/978-3-540-31955-9_8. URL: https://doi.org/10.1007/978-3-540-31955-9_8.

[38] Christian Galinski. *Interoperability of metadata. semantic interoperability.* Tech. rep. Austria, 2005.

[39] Artur d'Avila Garcez and Luis C. Lamb. *Neurosymbolic AI: The 3rd Wave.* 2020. arXiv: 2012.05876 [cs.AI].

[40] Artur S. d'Avila Garcez, Krysia Broda, and Dov M. Gabbay. "Neural-symbolic learning systems - foundations and applications". In: *Perspectives in neural computing.* 2002.

[41] Martina Garofalo et al. *Leveraging Knowledge Graph Embedding Techniques for Industry 4.0 Use Cases.* 2018. DOI: 10.48550/ARXIV.1808.00434. URL: https://arxiv.org/abs/1808.00434.

[42] Behzad Golshan et al. "Data Integration: After the Teenage Years". In: *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017, Chicago, IL, USA, May 14-19, 2017.* 2017, pp. 101–106.

[43] Irlán Grangel-González et al. "Alligator: A Deductive Approach for the Integration of Industry 4.0 Standards". In: *20th Int. Conf. on Knowledge Engineering and Knowledge Management, EKAW.* 2016, pp. 272–287.

[44] Irlán Grangel-González et al. "Knowledge Graphs for Semantically Integrating Cyber-Physical Systems". In: *Database and Expert Systems Applications - 29th International Conference, DEXA, Regensburg, Germany, September 3-6, Proceedings, Part I.* 2018, pp. 184–199.

[45] Irlán Grangel-González et al. "The industry 4.0 standards landscape from a semantic integration perspective". In: *2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)* (2017), pp. 1–8.

[46] Bernardo Cuenca Grau et al. "OWL 2: The next step for OWL". In: *Journal of Web Semantics* 6.4 (2008). Semantic Web Challenge 2006/2007, pp. 309–322. ISSN: 1570-8268. DOI: https://doi.org/10.1016/j.websem.2008.05.001. URL: https://www.sciencedirect.com/science/article/pii/S1570826808000413.

[47] RDF Working Group. *Resource Description Framework (RDF).* 2014. URL: https://www.w3.org/RDF/.

[48] W3C Working Group. *SHACL Advanced Features.* 2017. URL: https://www.w3.org/TR/shacl-af/.

[49] Claudio Gutiérrez and Juan F Sequeda. "Knowledge graphs". In: *Communications of the ACM* 64.3 (2021), pp. 96–104.

[50] J. Hakkola, J. Hukkanen, and M. Turpeinen et al. "Inhibition and induction of CYP enzymes in humans: an update." In: *Arch Toxicol* 94 (2020), pp. 3671–3722. DOI: https://doi.org/10.1007/s00204-020-02936-7.

[51] Muhammad Hassan et al. *Neuro-Symbolic Learning: Principles and Applications in Ophthalmology.* 2022. DOI: 10.48550/ARXIV.2208.00374. URL: https://arxiv.org/abs/2208.00374.

[52] A. Heuvelink. "Cognitive Models for Training Simulations". English. PhD thesis. Vrije Universiteit Amsterdam, 2009.

[53] Pascal Hitzler. "A Review of the Semantic Web Field". In: *Commun. ACM* 64.2 (Jan. 2021), pp. 76–83. ISSN: 0001-0782. DOI: 10.1145/3397512. URL: https://doi.org/10.1145/3397512.

[54] Pascal Hitzler et al. "Neuro-symbolic approaches in artificial intelligence". In: *National Science Review* 9.6 (Mar. 2022). nwac035. ISSN: 2095-5138. DOI: 10.1093/nsr/nwac035. eprint: https://academic.oup.com/nsr/article-pdf/9/6/nwac035/43952953/nwac035.pdf. URL: https://doi.org/10.1093/nsr/nwac035.

[55] Harry Hochheiser et al. "A Minimal Information Model for Potential Drug-Drug Interactions". In: *Frontiers in Pharmacology* 11 (2021), p. 2477. ISSN: 1663-9812. DOI: 10.3389/fphar.2020.608068. URL: https://www.frontiersin.org/article/10.3389/fphar.2020.608068.

[56] J. Hodges, K. Garcia, and S. Ray. "Semantic Development and Integration of Standards for Adoption and Interoperability". In: *Computer* 50.11 (Nov. 2017), pp. 26–36. ISSN: 1558-0814. DOI: 10.1109/MC.2017.4041353.

[57] Jack Hodges, Kimberly García, and Steven R. Ray. "Semantic Development and Integration of Standards for Adoption and Interoperability". In: *Computer* 50 (2017), pp. 26–36.

[58] Aidan Hogan et al. "Knowledge Graphs". In: *ACM Comput. Surv.* 54.4 (July 2021). ISSN: 0360-0300. DOI: 10.1145/3447772. URL: https://doi.org/10.1145/3447772.

[59] Krzysztof Janowicz et al. "Neural-Symbolic Integration and the Semantic Web". In: *Semant. Web* 11.1 (Jan. 2020), pp. 3–11. ISSN: 1570-0844. DOI: 10.3233/SW-190368. URL: https://doi.org/10.3233/SW-190368.

[60] Guoliang Ji et al. "Knowledge Graph Embedding via Dynamic Mapping Matrix". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Beijing, China: Association for Computational Linguistics, July 2015, pp. 687–696. DOI: 10.3115/v1/P15-1067. URL: https://aclanthology.org/P15-1067.

[61] Yuming Jiang et al. "Noninvasive prediction of occult peritoneal metastasis in gastric cancer using deep learning". In: *JAMA network open* 4.1 (2021), e2032269–e2032269.

[62] Bo Jin et al. "Multitask Dyadic Prediction and Its Application in Prediction of Adverse Drug-Drug Interaction". In: *AAAI.* 2017.

[63] Md. Tanvir Kabir et al. "Combination Drug Therapy for the Management of Alzheimer's Disease". In: *International Journal of Molecular Sciences* 21.9 (2020). ISSN: 1422-0067. DOI: 10.3390/ijms21093272. URL: https://www.mdpi.com/1422-0067/21/9/3272.

[64] Md. Rezaul Karim et al. "Drug-Drug Interaction Prediction Based on Knowledge Graph Embeddings and Convolutional-LSTM Network". In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB 2019, Niagara Falls, NY, USA, September 7-10, 2019.* Ed. by Xinghua Mindy Shi et al. ACM, 2019, pp. 113–123. DOI: 10.1145/3307339.3342161. URL: https://doi.org/10.1145/3307339.3342161.

[65] Andrej Karpathy and Li Fei-Fei. “Deep Visual-Semantic Alignments for Generating Image Descriptions”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.4 (Apr. 2017), pp. 664–676. ISSN: 2160-9292. DOI: 10.1109/tpami.2016.2598339. URL: http://dx.doi.org/10.1109/TPAMI.2016.2598339.

[66] George Karypis and Vipin Kumar. “A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs”. In: *SIAM J. Sci. Comput.* 20.1 (Dec. 1998), pp. 359–392. ISSN: 1064-8275. DOI: 10.1137/S1064827595287997. URL: http://dx.doi.org/10.1137/S1064827595287997.

[67] George Karypis and Vipin Kumar. “Multilevelk-way Partitioning Scheme for Irregular Graphs”. In: *Journal of Parallel and Distributed Computing* 48.1 (1998), pp. 96–129. ISSN: 0743-7315. DOI: https://doi.org/10.1006/jpdc.1997.1404. URL: https://www.sciencedirect.com/science/article/pii/S0743731597914040.

[68] Andrej Kastrin, Polonca Ferk, and Brane Leskošek. “Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning”. In: *PLOS ONE* 13.5 (May 2018), pp. 1–23. DOI: 10.1371/journal.pone.0196865. URL: https://doi.org/10.1371/journal.pone.0196865.

[69] Tamara G. Kolda and Brett W. Bader. “Tensor Decompositions and Applications”. In: *SIAM Review* 51.3 (2009), pp. 455–500. DOI: 10.1137/07070111X. eprint: https://doi.org/10.1137/07070111X. URL: https://doi.org/10.1137/07070111X.

[70] Olga Kovalenko and Jérôme Euzenat. “Semantic Matching of Engineering Data Structures”. In: *Semantic Web Technologies for Intelligent Engineering Applications*. Ed. by Stefan Biffl and Marta Sabou. Cham: Springer International Publishing, 2016, pp. 137–157. ISBN: 978-3-319-41490-4. DOI: 10.1007/978-3-319-41490-4_6. URL: https://doi.org/10.1007/978-3-319-41490-4_6.

[71] Changhee Lee, Jinsung Yoon, and Mihaela Van Der Schaar. “Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data”. In: *IEEE Transactions on Biomedical Engineering* 67.1 (2019), pp. 122–133.

[72] Geonhee Lee, Chihyun Park, and Jaegyoon Ahn. “Novel deep learning model for more accurate prediction of drug-drug interaction effects”. In: *BMC Bioinformatics* 20 (Aug. 2019). ISSN: 1471-2105. DOI: 10.1186/s12859-019-3013-0. URL: https://doi.org/10.1186/s12859-019-3013-0.

[73] Jens Lehmann et al. “DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia”. In: *Semantic Web* 6 (2015), pp. 167–195. DOI: 10.3233/SW-140134.

[74] Francesco Lelli. “Interoperability of the Time of Industry 4.0 and the Internet of Things”. In: *Future Internet* 11.2 (2019). ISSN: 1999-5903. DOI: 10.3390/fi11020036. URL: https://www.mdpi.com/1999-5903/11/2/36.

[75] Weixin Liang et al. *LRTA: A Transparent Neural-Symbolic Reasoning Framework with Modular Supervision for Visual Question Answering*. 2020. DOI: 10.48550/ARXIV.2011.10731. URL: https://arxiv.org/abs/2011.10731.

[76] Shi-Wan Lin et al. *Reference Architectural Model Industrie 4.0 (RAMI 4.0)*. Tech. rep. Industrial Internet Consortium and Plattform Industrie 4.0, 2017. URL: http://www.iiconsortium.org/pdf/JTG2_Whitepaper_final_20171205.pdf.

[77] Shi-Wan Lin et al. *The Industrial Internet of Things Volume G1: Reference Architecture*. White Paper IIC:PUB:G1:V1.80:20170131. Industrial Internet Consortium, 2017.

[78] Yankai Lin et al. "Learning Entity and Relation Embeddings for Knowledge Graph Completion". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 29.1 (Feb. 2015). URL: https://ojs.aaai.org/index.php/AAAI/article/view/9491.

[79] Yan Lu, K C Morris, and Simon Frechette. "Standards landscape and directions for smart manufacturing systems". In: *2015 IEEE International Conference on Automation Science and Engineering (CASE)*. 2015, pp. 998–1005. DOI: 10.1109/CoASE.2015.7294229.

[80] Jiayuan Mao et al. *The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision*. 2019. DOI: 10.48550/ARXIV.1904.12584. URL: https://arxiv.org/abs/1904.12584.

[81] Kenneth Marino et al. "KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 14111–14121.

[82] Giuseppe Marra. "Bridging symbolic and subsymbolic reasoning with minimax entropy models". In: *Intelligenza Artificiale* 15 (Feb. 2022), pp. 71–90. DOI: 10.3233/IA-210088.

[83] Raziyeh Masumshah, Rosa Aghdam, and Changiz Eslahchi. "A neural network-based method for polypharmacy side effects prediction". In: *BMC Bioinformatics* 22.385 (July 2021). ISSN: 1471-2105. DOI: 10.1186/s12859-021-04298-y. eprint: https://academic.oup.com/bioinformatics/article-pdf/34/13/i457/25098429/bty294.pdf. URL: https://doi.org/10.1186/s12859-021-04298-y.

[84] Nicola Melluso, Irlan Grangel-González, and Gualtiero Fantoni. "Enhancing Industry 4.0 standards interoperability via knowledge graphs with natural language processing". In: *Computers in Industry* 140 (2022), p. 103676. ISSN: 0166-3615. DOI: https://doi.org/10.1016/j.compind.2022.103676. URL: https://www.sciencedirect.com/science/article/pii/S0166361522000732.

[85] Paula Monteiro et al. "Adoption of Architecture Reference Models for Industrial Information Management Systems". In: *Int. Conf. on Intelligent Systems (IS)*. IEEE. 2018, pp. 763–770.

[86] Michalis Mountantonakis and Yannis Tzitzikas. "Large-scale Semantic Integration of Linked Data: A Survey". In: *ACM Comput. Surv.* 52.5 (Sept. 2019), 103:1–103:40. ISSN: 0360-0300. DOI: 10.1145/3345551. URL: http://doi.acm.org/10.1145/3345551.

[87] Shermyn Neo and Sheng et al. Wong. "Evolution of Initial Pharmacologic Treatment of Newly Diagnosed Parkinson's Disease Patients over a Decade in Singapore". In: *Parkinson's Disease* 2020.6293124 (2020). DOI: 10.1155/2020/6293124.

[88] M. E. J. Newman. "Modularity and community structure in networks". In: *Proceedings of the National Academy of Sciences* 103.23 (2006), pp. 8577–8582. ISSN: 0027-8424. DOI: 10.1073/pnas.0601602103. eprint: https://www.pnas.org/content/103/23/8577.full.pdf. URL: https://www.pnas.org/content/103/23/8577.

[89] Dai Quoc Nguyen et al. "A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 327–333. DOI: 10.18653/v1/N18-2053. URL: https://aclanthology.org/N18-2053.

[90] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. *Holographic Embeddings of Knowledge Graphs.* 2015. DOI: 10.48550/ARXIV.1510.04935. URL: https://arxiv.org/abs/1510.04935.

[91] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. "A Three-Way Model for Collective Learning on Multi-Relational Data". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning.* ICML'11. Bellevue, Washington, USA: Omnipress, 2011, pp. 809–816. ISBN: 9781450306195.

[92] P. O'Donovan, K. Leahy, and K. Bruton. "An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities". In: *Journal of Big Data 2* 25 (2015).

[93] Solla P et al. "Therapeutic interventions and adjustments in the management of Parkinson disease: role of combined carbidopa/levodopa/entacapone (Stalevo®)." In: *Neuropsychiatr Dis Treat.* 6.1 (2010). URL: https://doi.org/10.2147/NDT.S5190.

[94] Guillermo Palma, Maria-Esther Vidal, and Louiqa Raschid. "Drug-Target Interaction Prediction Using Semantic Similarity and Edge Partitioning". In: *The Semantic Web – ISWC 2014.* Ed. by Peter Mika et al. Cham: Springer International Publishing, 2014, pp. 131–146. ISBN: 978-3-319-11964-9.

[95] Guillermo Palma, Maria-Esther Vidal, and Louiqa Raschid. "Drug-Target Interaction Prediction Using Semantic Similarity and Edge Partitioning". In: *Proc. of the 13th Int. Semantic Web Conf. - Part I.* ISWC '14. NY, USA: Springer-Verlag New York, Inc., 2014, pp. 131–146. ISBN: 978-3-319-11963-2. DOI: 10.1007/978-3-319-11964-9_9. URL: http://dx.doi.org/10.1007/978-3-319-11964-9_9.

[96] *Symptoms - Parkinson's disease.* 2019. URL: https://www.nhs.uk/conditions/parkinsons-disease/symptoms/.

[97] Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. "Semantics and Complexity of SPARQL". In: *ACM Trans. Database Syst.* 34.3 (Sept. 2009). ISSN: 0362-5915. DOI: 10.1145/1567274.1567278. URL: https://doi.org/10.1145/1567274.1567278.

[98] Habibollah Pirnejad et al. "Preventing potential drug-drug interactions through alerting decision support systems: A clinical context based methodology". In: *International Journal of Medical Informatics* 127 (2019), pp. 18–26. ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2019.04.006. URL: https://www.sciencedirect.com/science/article/pii/S1386505618303095.

[99] Jay Pujara, Eriq Augustine, and Lise Getoor. "Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short". In: *EMNLP.* 2017.

[100] *R2RML: RDB to RDF Mapping Language.* https://www.w3.org/TR/r2rml/. Accessed: 28-10-2022.

[101] Raghu Ramakrishnan and Jeffrey D. Ullman. "A survey of deductive database systems". In: *The Journal of Logic Programming* 23.2 (1995), pp. 125–149. ISSN: 0743-1066. DOI: https://doi.org/10.1016/0743-1066(94)00039-9. URL: https://www.sciencedirect.com/science/article/pii/0743106694000399.

[102] *RDF 1.1 Concepts and Abstract Syntax.* URL: https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

[103] Ariam Rivas and Maria-Esther Vidal. "Capturing Knowledge about Drug-Drug Interactions to Enhance Treatment Effectiveness". In: *Proceedings of the 11th on Knowledge Capture Conference*. K-CAP '21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 33–40. ISBN: 9781450384575. DOI: 10.1145/3460210.3493560. URL: https://doi.org/10.1145/3460210.3493560.

[104] Ariam Rivas et al. "A Neuro-Symbolic System over Knowledge Graphs for Link Prediction". In: 2022. URL: https://www.semantic-web-journal.net/content/neuro-symbolic-system-over-knowledge-graphs-link-prediction.

[105] Ariam Rivas et al. "Discover Relations in the Industry 4.0 Standards Via Unsupervised Learning on Knowledge Graph Embeddings". In: *Journal of Data Intelligence* 2.3 (Sept. 2021), pp. 336–347. DOI: 10.26421/jdi2.3-2. URL: https://doi.org/10.26421/JDI2.3-2.

[106] Ariam Rivas et al. "Multi-granulation Strategy via Feature Subset Extraction by Using a Genetic Algorithm and a Rough Sets-Based Measure of Dependence". In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Ed. by Ruben Vera-Rodriguez, Julian Fierrez, and Aythami Morales. Cham: Springer International Publishing, 2019, pp. 203–211. ISBN: 978-3-030-13469-3.

[107] Ariam Rivas et al. "Unveiling Relations in the Industry 4.0 Standards Landscape Based on Knowledge Graph Embeddings". In: *Database and Expert Systems Applications*. 2020. DOI: 10.1007/978-3-030-59051-2_12.

[108] Narjes Rohani and Changiz Eslahchi. "Drug-Drug Interaction Predicting by Neural Network Using Integrated Similarity". In: *Scientific Reports* 9 (Sept. 2019). ISSN: 2045-2322. DOI: 10.1038/s41598-019-50121-3. URL: https://doi.org/10.1038/s41598-019-50121-3.

[109] Andrea Rossi et al. "Explaining Link Prediction Systems Based on Knowledge Graph Embeddings". In: *Proceedings of the 2022 International Conference on Management of Data*. SIGMOD '22. Philadelphia, PA, USA: Association for Computing Machinery, 2022, pp. 2062–2075. ISBN: 9781450392495. DOI: 10.1145/3514221.3517887. URL: https://doi.org/10.1145/3514221.3517887.

[110] Andrea Rossi et al. "Knowledge Graph Embedding for Link Prediction: A Comparative Analysis". In: *ACM Trans. Knowl. Discov. Data* 15.2 (Jan. 2021). ISSN: 1556-4681. DOI: 10.1145/3424672. URL: https://doi.org/10.1145/3424672.

[111] Mithun Rudrapal, Shubham J. Khairnar, and Anil G. Jadhav. "Drug Repurposing (DR): An Emerging Approach in Drug Discovery". In: *Drug Repurposing*. Ed. by Farid A. Badria. Rijeka: IntechOpen, 2020. Chap. 1. DOI: 10.5772/intechopen.93193. URL: https://doi.org/10.5772/intechopen.93193.

[112] Stuart J. Russell and Peter Norvig. *Artificial Intelligence - A Modern Approach, Third International Edition*. Pearson Education, 2010. ISBN: 978-0-13-207148-2. URL: http://vig.pearsoned.com/store/product/1,1207,store-12521%5C_isbn-0136042597,00.html.

[113] Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. "Deep learning improves prediction of drug&#x2013;drug and drug&#x2013;food interactions". In: *Proceedings of the National Academy of Sciences* 115.18 (2018), E4304–E4311. DOI: 10.1073/pnas.1803294115. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1803294115. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1803294115.

[114] Ahmad Sakor et al. “Falcon 2.0: An Entity and Relation Linking Tool over Wikidata”. In: *CIKM ’20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020.* Ed. by Mathieu d’Aquin et al. ACM, 2020, pp. 3141–3148. DOI: 10.1145/3340531.3412777. URL: https://doi.org/10.1145/3340531.3412777.

[115] Ahmad Sakor et al. “Knowledge4COVID-19: A Semantic-based Approach for Constructing a COVID-19 related Knowledge Graph from Various Sources and Analysing Treatments’ Toxicities”. In: *arXiv preprint arXiv:2206.07375* (2022). DOI: https://doi.org/10.48550/arXiv.2206.07375.

[116] Md Kamruzzaman Sarker et al. *Neuro-Symbolic Artificial Intelligence: Current Trends.* 2021. DOI: 10.48550/ARXIV.2105.05330. URL: https://arxiv.org/abs/2105.05330.

[117] Michael Schmidt, Michael Meier, and Georg Lausen. “Foundations of SPARQL Query Optimization”. In: *Proceedings of the 13th International Conference on Database Theory.* ICDT ’10. Lausanne, Switzerland: Association for Computing Machinery, 2010, pp. 4–33. ISBN: 9781605589473. DOI: 10.1145/1804669.1804675. URL: https://doi.org/10.1145/1804669.1804675.

[118] *Semantic Web.* URL: https://www.w3.org/standards/semanticweb/.

[119] Paul Shannon et al. “Cytoscape: a software environment for integrated models of biomolecular interaction networks”. In: *Genome research* 13.11 (2003), pp. 2498–2504.

[120] Eunji Shin et al. “Knowledge graph embedding and reasoning for real-time analytics support of chemical diagnosis from exposure symptoms”. In: *Process Safety and Environmental Protection* 157 (2022), pp. 92–105. ISSN: 0957-5820. DOI: https://doi.org/10.1016/j.psep.2021.11.002. URL: https://www.sciencedirect.com/science/article/pii/S095758202100598X.

[121] Amit Singhal. *Introducing the Knowledge Graph: things, not strings.* 2012. URL: https://www.blog.google/products/search/introducing-knowledge-graph-things-not/.

[122] Edward E. Smith and Stephen M. Kosslyn. “Cognitive Psychology: Mind and Brain”. In: 2006.

[123] Dhanya Sridhar, Shobeir Fakhraei, and Lise Getoor. “A probabilistic approach for collective similarity-based drug-drug interaction prediction”. In: *Bioinform.* 32.20 (2016), pp. 3175–3182.

[124] Dhanya Sridhar, Shobeir Fakhraei, and Lise Getoor. “A probabilistic approach for collective similarity-based drug–drug interaction prediction”. In: *Bioinformatics* 32.20 (June 2016), pp. 3175–3182. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw342. eprint: https://academic.oup.com/bioinformatics/article-pdf/32/20/3175/25040491/btw342.pdf. URL: https://doi.org/10.1093/bioinformatics/btw342.

[125] Thanos G. Stavropoulos et al. “Detection of Health-Related Events and Behaviours from Wearable Sensor Lifestyle Data Using Symbolic Intelligence: A Proof-of-Concept Application in the Care of Multiple Sclerosis”. In: *Sensors* 21.18 (2021). ISSN: 1424-8220. DOI: 10.3390/s21186230. URL: https://www.mdpi.com/1424-8220/21/18/6230.

[126] Zhenchao Sun et al. “Disease Prediction via Graph Neural Networks”. In: *IEEE Journal of Biomedical and Health Informatics* 25.3 (2021), pp. 818–826. DOI: 10.1109/JBHI.2020.3004143.

[127] Zhiqing Sun et al. *RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space.* 2019. DOI: 10.48550/ARXIV.1902.10197. URL: https://arxiv.org/abs/1902.10197.

[128] Zachary Susskind et al. "Neuro-Symbolic AI: An Emerging Class of AI Workloads and their Characterization". In: *CoRR* abs/2109.06133 (2021). arXiv: 2109.06133. URL: https://arxiv.org/abs/2109.06133.

[129] Lucreţia Udrescu et al. "Clustering drug-drug interaction networks with energy model layouts: community analysis and drug repurposing". In: *Scientific reports* 6.1 (2016), pp. 1–10.

[130] Nancy Velasquez, Elsa Estevez, and Patricia Pesado. "Cloud Computing, Big Data and the Industry 4.0 Reference Architectures". In: *Journal of Computer Science and Technology* 18.03 (2018), e29.

[131] Maria-Esther Vidal et al. "Transforming Heterogeneous Data into Knowledge for Personalized Treatments - A Use Case". In: *Datenbank-Spektrum* 19.2 (2019), pp. 95–106. DOI: 10.1007/s13222-019-00312-z. URL: https://doi.org/10.1007/s13222-019-00312-z.

[132] Santiago Vilar et al. "Similarity-based modeling in large-scale prediction of drug-drug interactions". In: *Nature Protocols* 9 (Sept. 2014). ISSN: 1750-2799. DOI: 10.1038/nprot.2014.151. URL: https://doi.org/10.1038/nprot.2014.151.

[133] Denny Vrandečić. "Wikidata: A New Platform for Collaborative Data Collection". In: *Proceedings of the 21st International Conference on World Wide Web*. WWW '12 Companion. Lyon, France: Association for Computing Machinery, 2012, pp. 1063–1064. ISBN: 9781450312301. DOI: 10.1145/2187980.2188242. URL: https://doi.org/10.1145/2187980.2188242.

[134] Meihong Wang, Linling Qiu, and Xiaoli Wang. "A Survey on Knowledge Graph Embeddings for Link Prediction". In: *Symmetry* 13.3 (2021). ISSN: 2073-8994. DOI: 10.3390/sym13030485. URL: https://www.mdpi.com/2073-8994/13/3/485.

[135] Meng Wang et al. "Drug-Drug Interaction Predictions via Knowledge Graph and Text Embedding: Instrument Validation Study". In: *JMIR Medical Informatics* 9 (2021). DOI: 10.2196/28277.

[136] Quan Wang et al. "Knowledge Graph Embedding: A Survey of Approaches and Applications". In: *IEEE Trans. Knowl. Data Eng.* 29.12 (2017), pp. 2724–2743. DOI: 10.1109/TKDE.2017.2754499.

[137] Zhen Wang et al. "Knowledge Graph Embedding by Translating on Hyperplanes". In: *AAAI.* 2014.

[138] Zhen Wang et al. "Knowledge Graph Embedding by Translating on Hyperplanes". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 28.1 (Jan. 2014). URL: https://ojs.aaai.org/index.php/AAAI/article/view/8870.

[139] *Web Ontology Language (OWL).* 2012. URL: https://www.w3.org/OWL/.

[140] David S. Wishart et al. "DrugBank: a comprehensive resource for in silico drug discovery and exploration". In: *Nucleic Acids Research* 34 (Jan. 2006), pp. 668–672. ISSN: 0305-1048. DOI: 10.1093/nar/gkj067. URL: https://doi.org/10.1093/nar/gkj067.

[141] Bishan Yang et al. *Embedding Entities and Relations for Learning and Inference in Knowledge Bases.* 2014. DOI: 10.48550/ARXIV.1412.6575. URL: https://arxiv.org/abs/1412.6575.

[142] Yixing Yang et al. "Efficient Bi-Triangle Counting for Large Bipartite Networks". In: *Proc. VLDB Endow.* 14.6 (Feb. 2021), pp. 984–996. ISSN: 2150-8097. DOI: 10.14778/3447689.3447702. URL: https://doi.org/10.14778/3447689.3447702.

[143] Kexin Yi et al. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding". In: *Advances in Neural Information Processing Systems.* Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/5e388103a391daabe3de1d76a6739ccd-Paper.pdf.

[144] Abe Zeid et al. "Interoperability in Smart Manufacturing: Research Challenges". In: *Machines* 7.2 (2019), p. 21.

[145] Jing Zhang et al. "Neural, symbolic and neural-symbolic reasoning on knowledge graphs". In: *AI Open* 2 (2021), pp. 14–35.

[146] Ping Zhang et al. "Label Propagation Prediction of Drug-Drug Interactions Based on Clinical Side Effects". In: *Scientific Reports* 5 (July 2015). ISSN: 2045-2322. DOI: 10.1038/srep12339. URL: https://doi.org/10.1038/srep12339.

[147] Rui Zhang et al. "Drug repurposing for COVID-19 via knowledge graph completion". In: *Journal of Biomedical Informatics* 115 (2021), p. 103696. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2021.103696. URL: https://www.sciencedirect.com/science/article/pii/S1532046421000253.

[148] Shuai Zhang et al. "Quaternion Knowledge Graph Embeddings." In: *NeurIPS.* 2019, pp. 2731–2741. URL: http://papers.nips.cc/paper/8541-quaternion-knowledge-graph-embeddings.

[149] Wen Zhang et al. "Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data". In: *BMC Bioinformatics* 18 (Jan. 2017). ISSN: 1471-2105. DOI: 10.1186/s12859-016-1415-9. URL: https://doi.org/10.1186/s12859-016-1415-9.

[150] Qing Zhao et al. "Knowledge Guided Feature Aggregation for the Prediction of Chronic Obstructive Pulmonary Disease With Chinese EMRs". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2022), pp. 1–10. DOI: 10.1109/TCBB.2022.3198798.

[151] Chaoyu Zhu et al. "Multimodal reasoning based on knowledge graph embedding for specific diseases". In: *Bioinformatics* 38.8 (Feb. 2022), pp. 2235–2245. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac085. eprint: https://academic.oup.com/bioinformatics/article-pdf/38/8/2235/43370161/btac085.pdf. URL: https://doi.org/10.1093/bioinformatics/btac085.

[152] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. "Modeling polypharmacy side effects with graph convolutional networks". In: *Bioinformatics* 34.13 (June 2018), pp. i457–i466. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty294. eprint: https://academic.oup.com/bioinformatics/article-pdf/34/13/i457/25098429/bty294.pdf. URL: https://doi.org/10.1093/bioinformatics/bty294.

Curriculum Vitae

Personal information

First name(s) / Surname(s)	**Rivas Mendez, Ariam**
Address(es)	Hannover, Lower Saxony, Germany
Telephone(s)	+49 172 2678522
Web(es)	linkedin.com/in/ariam-rivas, https://www.researchgate.net/profile/Ariam_Rivas_Mendez
E-mail	ariam.rivas@tib.eu
Nationality	Cuban
Date of birth	26.06.1990
Gender	Male

Desired employment / Occupational field

Research Assistant

Work experience

Dates	December 2018 onwards
Occupation or position held	Research assistant
Main activities and responsibilities	Researcher, working in European projects of Bioinformatics. Artificial Intelligence, Knowledge Discovery, and Knowledge Graph Embeddings.
Name and address of the employer	Leibniz University of Hannover, L3S Research Center, Lange Laube 28, 30159 Hannover, Germany
Type of business or sector	University (Education)
Dates	October 2018 onwards
Occupation or position held	Ph.D. Student at Leibniz University of Hannover
Main activities and responsibilities	Research, working on projects
Name and address of the employer	Leibniz Information Center for Science & Technology (TIB), Leibniz University of Hannover
Type of business or sector	University (Education)
Dates	September 2014 until May 2018
Occupation or position held	Research assistant professor
Main activities and responsibilities	Teaching, Research, Modelling, Simulation, Optimization of processes, Software Development
Name and address of the employer	University of Holguín, Avenida XX Aniversario, Piedra Blanca, Holguín, Cuba. http://www.uho.edu.cu
Type of business or sector	University (Education)

Education

Dates	2014-2017

Title of qualification awarded	Master's degree (MSc) in Applied Mathematics and Informatics for Management
Principal subjects/occupational skills covered	Advanced Machine Learning - Data Mining in Enterprise Systems - Intelligent systems applied to the Administration - Statistics - Elements of Mathematical Modeling - Organizational Management - Enterprise Software Engineering - Basic Elements of Administration Thesis Title: “Procedure for the forecasting of pharmaceutical product using regression models. Case study: EMCOMED Holguín”
Name and type of organization providing education and training	University of Holguín
Dates	2009-2014
Title of qualification awarded	Bachelor's degree (B.Sc.) in Computer Engineering (Graduated with Honours)
Principal subjects/occupational skills covered	Software engineer | Data base | Programmer |Artificial Intelligence Thesis Title: “Experimental study on rule-based supervised classification algorithms in high-dimensional datasets”
Name and type of organization providing education and training	University of Holguín.
Personal skills and competences	• **Commitment and responsibility:** I show my commitment and willingness to give my best, to get involved, and work hard. • **Logical reasoning:** thinking logically is paramount. In order to extract accurate insights from data, you need to stick to the insights that the data has produced. • **Critical thinking**: At every stage of the data science process, I have had to reach critical decisions in order to achieve the desired results. • **Organisation:** helps me to structure and organize the code well and in the management of tasks.
Social skills and competences	• **Teamwork**: work experience in intercultural, international, and interdisciplinary teams on seven research projects in Germany and Cuba (including four European projects, Horizon 2020, and a VLIR project with Belgium Universities). • **Communication and Teaching**: developed during three years as a Teaching Assistant in Higher Education, working on international projects, writing project deliverables, software documentation, and scientific papers.
Methodical skills and competences	• **Self-discipline**: I set priorities, identify what is important and urgent to do, and set deadlines for tasks. • **Problem-solving skills**: I have the skill to recognize difficulties and react quickly to them and find an appropriate solution before further damage occurs. • **Self-management:** I clearly overview pending tasks and prioritize the important ones. Then I can perform tasks better and solve problems and challenges. • **Reliability**: I have worked for three years as a data scientistic (clinical data of lung cancer and dementia patients) in Germany. Furthermore, I worked for more than two years on the control of student withdrawal in the historical archive of the System for the Management of the New University (SIGENU) in Cuba.
Mother tongue(s)	**Spanish**
Other language(s)	

Self-assessment	**Understanding**				**Speaking**				**Writing**	
European level (*)	Listening		Reading		Spoken interaction		Spoken production			
English	B2	Proficient User	B2	Proficient User	B2	Proficient User	B2	Proficient User	B2	Proficient User
German	B1	Medium User	B1	Medium User	B1	Medium User	B1	Medium User	B1	Medium User

(*) Common European Framework of Reference for Languages

Certifications and Training

1. Scientific Database Programming. Certificate earned on 07/2022.
2. Introduction to Biologic and Medical Concepts. Certificate earned on 03/2022.
3. Semantic Data Modeling. Certificate earned on 12/2021.
4. Introduction to mechanistic mathematical modeling approach in biology and medicine. Certificate earned on 11/2021
5. Introduction to Machine Learning and Data Mining. Certificate earned on 09/2021.
6. Knowledge Engineering and Semantic Web. Certificate earned on 07/2021.
7. Responsable Data Management. Certificate earned on 06/2021.
8. Introduction to Molecular Data Science. Certificate earned on 04/2021.
9. Introduction to bioinformatics data types and analysis techniques. Certificate earned on 04/2021.
10. Introduction to Scientific Databases. Certificate earned on 03/2021.
11. Ethics and Scientific Integrity. Certificate earned on 02/2021.
12. Mathematical Tools - R. Certificate earned on 05/2017.
13. Advanced topics in Artificial Intelligence. Certificate earned on 05/2017.
14. Heuristic Approaches to Problem Solving. Certificate earned on 04/2017
15. Data visualization and user interfaces. Certificate earned on 03/2017
16. Data Modeling. Certificate earned on 03/2017
17. Computational Mathematics. Certificate earned on 02/2017
18. Business process modeling. Certificate earned on 02/2017

Technical skills and competences

- Research Engineer - Artificial Intelligence - Machine Learning - Knowledge Graph - Knowledge Discovery - Knowledge Engineering - Data Scientist - Big Data Analisis - Data Visualization - Data Management - Relational Database - Software Engineer - Project Management - Statistics - Metaheuristics - Rough Set Theory

Computer skills and competences

- Deep Learning (TensorFlow, Pytorch, PyKeen)
- Knowledge Graph Embeddings
- Machine Learning (Sklearn, WEKA, RapidMiner)
- Metaheuristics (Genetic Algorithm, Particle Swarm Optimization, Ant colony optimization)
- Rough Set Theory (Rough Sets Data Explorer (ROSE2))
- Statistical Analysis (Python, R)
- Analysis and Design of Algorithms
- Data Visualization (Python, R, Cytoscape)
- Knowledge Graph (RML), Knowledge Engineer (RDF, RDFS, OWL, SPARQL), Semantic Technologies, Semantic Data Integration, Linked Data
- Programming (Python, R, Java, JavaScript, JQuery)
- Data Base Modelling, Construction (PostgreSQL, MySQL)
- Software Engineer (UML, RUP, ICONIX, BPM, BPMN)
- Simulation (Arena Simulation Software)

Additional information

Main publications

1. Knowledge4COVID-19: A Semantic-based Approach for Constructing a COVID-19 related Knowledge Graph from Various Sources and Analysing Treatments' Toxicities. Ahmad Sakor, Samaneh Jozashoori, Emetis Niazmand, **Ariam Rivas**, Konstantinos Bougiatiotis, Fotis Aisopos, Enrique Iglesias, Philipp D. Rohde, Trupti Padiya, Anastasia Krithara, Georgios Paliouras, Maria-Esther Vidal. https://doi.org/10.48550/arXiv.2206.07375 (Journal of Web Semantics 2022)
2. Capturing Knowledge about Drug-Drug Interactions to Enhance Treatment Effectiveness. **Ariam Rivas**, Maria-Esther Vidal. In Proceedings of the 11th on Knowledge Capture Conference (K-CAP 2021). https://doi.org/10.1145/3460210.3493560
3. Discover Relations in the Industry 4.0 Standards Via Unsupervised Learning on Knowledge Graph Embeddings. **Ariam Rivas**, Irlan Grangel-González, Diego Collarana, Jens Lehmann and Maria-Esther Vidal. Journal of Data Intelligence ISSN: 2577-610X. Vol.2 No.3 September 2021. https://doi.org/10.26421/JDI2.3-2
4. Unveiling Relations in the Industry 4.0 Standards Landscape Based on Knowledge Graph Embeddings. **Ariam Rivas**, Irlan Grangel-González, Diego Collarana, Jens Lehmann and Maria-Esther Vidal. Database and Expert Systems Applications. DEXA 2020. https://doi.org/10.1007/978-3-030-59051-2_12

5. Knowledge Graphs Evolution and Preservation - A Technical Report from ISWS 2019. https://doi.org/10.48550/arXiv.2012.11936
6. Transforming Heterogeneous Data into Knowledge for Personalized Treatments—A Use Case. Maria-Esther Vidal, Kemele M. Endris, Samaneh Jazashoori, Ahmad Sakor, **Ariam Rivas.** *Datenbank Spektrum* **19**, 95–106 (2019). https://doi.org/10.1007/s13222-019-00312-z
7. Multi-granulation Strategy via Feature Subset Extraction by Using a Genetic Algorithm and a Rough Sets-Based Measure of Dependence. **Ariam Rivas**, Ricardo Navarro, Chyon Hae Kim, Rafael Bello. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2018. https://doi.org/10.1007/978-3-030-13469-3_24
8. Application of Machine Learning Methods in a System Based on Ontology. María Isabel Castellanos, **Ariam Rivas**, Emilio Lucas. 3rd International Workshop of Semantic Web, 2018, Cuba. ISSN: 1613-0073, Vol-2096, http://ceur-ws.org/Vol-2096/
9. Prediction of limit values of environmental indicators using artificial neural networks in an ontology-based system. María Isabel Castellanos Domínguez, **Ariam Rivas**, Emilio Enrique Lucas López. 17th International Convention and Fair Informática 2018.
10. Procedure for forecasting demand through artificial neural networks. Yosvani Orlando Lao-León, **Ariam Rivas**, Milagros Caridad Pérez-Pravia, Fernando Marrero-Delgado. Electronic journal "Ciencias Holguín", ISSN: 1027-2127. 2017. http://www.redalyc.org/articulo.oa?id=181549596004
11. Digital Library with automatic document classification techniques.**Ariam Rivas**, Eduardo Pérez-Perdomo, María Isabel Castellanos Domínguez, Victor Gongora Zaldivar. Proceedings of the VIII International Scientific Conference, 2017. ISBN: 978-959-16-3272-2.
12. Procedure for the forecasting of demand for pharmaceuticals using the multiple linear regression models. **Ariam Rivas**, Yosvani Orlando Lao-León, Rafael Bello. Proceedings of the VIII International Scientific Conference, University of Holguin, ISBN: 978-959-16-3272-2. April 2017
13. Study of the process of knowledge discovery in data. **Ariam Rivas**, Eduardo Pérez-Perdomo. Proceedings of the VII International Scientific Conference, University of Holguin, ISBN: 978-959-16-2472-7 April 27, 2015

Projects

- Research Assistant: P4-LUCAT Personalized medicine for lung cancer treatment:(2020-present) https://p4-lucat.eu/
- Research Assistant: H2020 CLARIFY Big data and Artificial Intelligence for monitoring health status and quality of life after cancer treatment (2020-present). https://www.clarify2020.eu/
- Granted a DAAD scholarship in the program Research Grants - Doctoral Programmes in Germany. 2018-2022
- Research Assistant: H2020 BigMedilytics - Big Data for Medical Analytic (December 2018 to February 2021, Germany). https://www.bigmedilytics.eu/
- Research Assistant: H2020 iASiS-Integration and analysis of heterogeneous big data for precision medicine and suggested treatments for different types of patients (December 2018 to October 2020, Germany). https://www.project-iasis.eu/
- Research Assistant: Knowledge management through techniques of Artificial Intelligence in organizational processes, with emphasis on environmental management. University of Holguín and Central University of Las Villas “Marta Abreu”. 2016-2018
- Research Assistant of international project: Strengthening of the IT roll in Cuban universities for society development. Cooperation program of CUBA – VLIR network (in association with Belgium universities). Subproject 1: IT Researches and associated sciences. 2016-2017

Honours and awards

- DAAD scholarship in the program Research Grants - Doctoral Programmes in Germany. 2018
- Graduated with Honours. University of Holguin, Cuba. July 2014
- Award 2ndplace in the Final University of Holguín. International Collegiate Programming Contest (ICPC), 2013
- Award 1st place in the Final University of Holguín. International Collegiate Programming Contest (ICPC), 2012
- Award 2nd place in the Cuban final. International Collegiate Programming Contest (ICPC), 2011