A Neuro-Symbolic System over Knowledge Graphs for Link Prediction

Ariam Rivas [a,*], Diego Collarana [b,c], Maria Torrente [d] and Maria-Esther Vidal [a,e]
[a] *Leibniz University of Hannover, Germany*
E-mail: ariam.rivas@tib.eu
[b] *Fraunhofer Institute for Intelligent Analysis and Information Systems, Dresden, Germany*
E-mail: diego.collarana.vargas@iais.fraunhofer.de
[c] *Universidad Privada Boliviana, Cochabamba, Bolivia*
[d] *Oncología Médica, Hospital Universitario Puerta de Hierro-Majadahonda, Spain*
E-mail: mtorrente80@gmail.com
[e] *TIB Leibniz Information Centre for Science and Technology, Germany*
E-mail: maria.vidal@tib.eu

Abstract. Neuro-Symbolic focuses on integrating symbolic and sub-symbolic systems. The aim is to provide a neural-symbolic implementation of logic, a logical characterization of a neural system, or a hybrid learning system that contributes features of symbolic and sub-symbolic systems. They differ fundamentally in how they represent data and information. Neuro-symbolic systems have recently received significant attention in the scientific communities. However, despite efforts in neural-symbolic integration, symbol processing currently has limited scope and applicability. This work leverages the symbolic system, independent of the application domain, and improves the predictive capability of Knowledge Graph Embeddings (KGE). We tackle the problem of Neuro-Symbolic AI integration, enabling expressive reasoning and robust learning to discover relationships over a knowledge graph. We present a novel approach to integrating Neuro-Symbolic AI systems. Deductive databases implement the symbolic system for an abstract target prediction over a knowledge graph. The symbolic system enhances the predictive capacity of the subsymbolic systems implemented by KGE models. Our approach builds the ego networks of the head and tail of the abstract target prediction, and the symbolic system deduces new relationships enhancing the ego networks. Thus, the subsymbolic systems increase the predictive capacity of the abstract target prediction. As a proof of concept, we have implemented our neuro-symbolic system on top of a KG for lung cancer to predict treatment effectiveness. Our empirical results put the deduction power of deductive databases into perspective; they suggest that enhancing the neighborhoods of the entities on the head or tail of a target prediction can improve the predictive capacity of existing KGE models.

Keywords: Neuro-Symbolic Artificial Intelligence, Deductive Systems, Knowledge Graph Embeddings, Drug-Drug Interactions

1. Introduction

Neural-symbolic computing is an active research area that attempts to combine the division of symbolic and sub-symbolic models. The symbolic models refer to representations of reasoning and explainability, while sub-symbolic models are Artificial Intelligence systems. Complex problem-solving using Artificial Intelligence (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

*Corresponding author. E-mail: ariam.rivas@tib.eu.

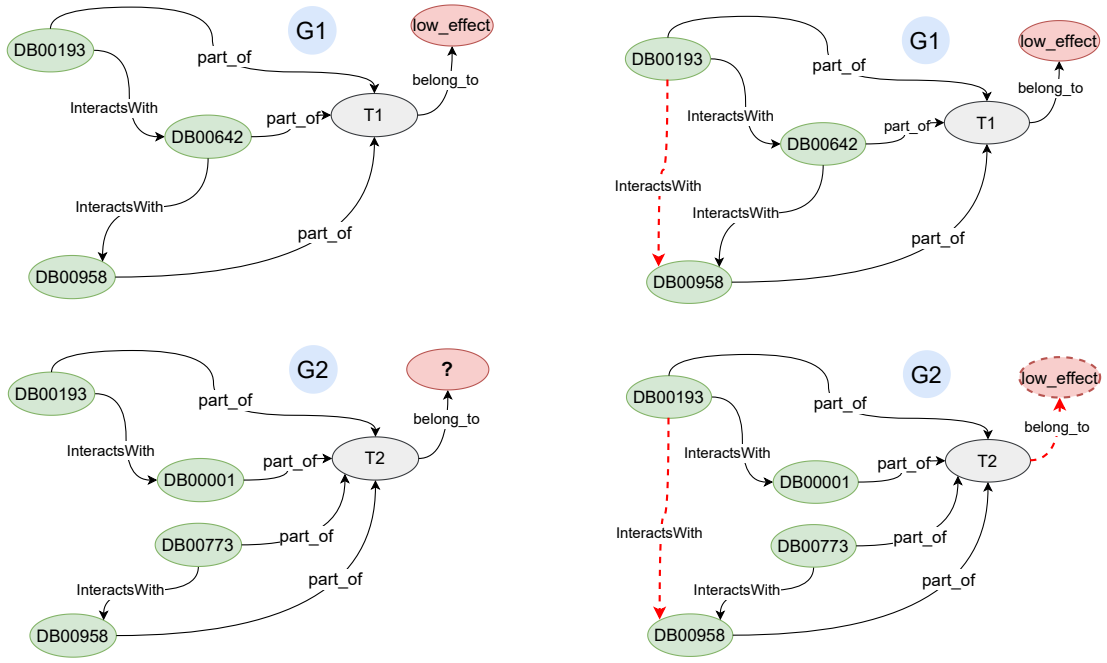
sub-symbolic systems usually use representations based on vector space. Thus, neuro-symbolic integration aims to bridge the gap between symbolic and sub-symbolic systems.
Integrating neural-symbolic into real-world applications is a challenging task. Even in controlled environments, e.g., training simulators, neural-symbolic integration may not be completed successfully [1]. For instance, Fernlund et al. [2] 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 neural-symbolic systems in bioinformatics is the Connectionist Inductive Learning and Logic Programming (CILP) [3]. Furthermore, Karpathy et al. [4] combine convolutional neural networks with bidirectional recurrent neural networks over sentences to recognize and label image regions. Despite advances in neural-symbolic IA integration, symbol processing currently has limited scope and applicability. 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.
Problem: We tackle the problem of Neuro-symbolic AI integration, enabling expressive reasoning and robust learning to discover relationships over knowledge graphs.
Proposed Solution: We present a novel approach based on the integration of Neuro-Symbolic AI systems. The symbolic system is implemented by deductive databases, enhancing the predictive capacity of subsymbolic systems implemented as KGE models. The deductive databases are defined for an abstract target prediction over a Knowledge Graph (KG). Our proposed solution builds the ego networks of the entities that correspond the head and tail of the abstract target prediction to deduce new relationships and enhance the ego networks. The KG is completed with relations implicitly defined in the deductive systems. As a result, the subsymbolic system works over a larger set of positives relations and is able to more precisely predict new links.
Results: We assess the performance of the proposed neuro-symbolic system on top of a KG of lung cancer treatments; the predictive task is to predict treatment effectiveness. The experiments are executed following different configurations and baselines. Results of a 5-fold cross-validation process demonstrate that our integrated system, improves the prediction accuracy of eleven state-of-the-art KGE models. Thus, the outcomes of this experimental study put the power of deductive databases into perspective, showing thus, how they can empower the predictive capacity of KGE models.
Contributions: This paper relies on our previous work [5] where we propose a deductive system over knowledge graphs to formalize the process of pharmacokinetic DDIs. Built on these results, we present a hybrid approach able to combine symbolic reasoning expressed by deductive systems with the subsymbolic expressiveness of KG embeddings, to enhance prediction accuracy. As a proof of concept, we assess the power of our proposal on the problem of predicting treatment effectiveness. In a nutshell, our contributions are:

1. An approach able to empower KGEs with symbolic deductive systems.
2. A domain-agnostic approach able to capture the knowledge represented in a knowledge graph and deduce relationships and their properties.
3. The assessment of the proposed system to the problem of predicting the effectiveness of lung-cancer treatments composed of multiple drugs (i.e., polypharmacy treatments).
4. An extensive evaluation of symbolic-subsymbolic system in state-of-art KGE models.

The rest of the paper is structured as follows: Section 2 illustrates a motivating example. Section 3 presents preliminaries and details of our proposed approach. Section 4 illustrates how the proposed hybrid method can be applied in the context of predicting the effectiveness of polypharmacy lung cancer treatments. Results of the empirical evaluation of our method are reported in Section 5. Section 6 analyses the state of the art. Finally, we close with the conclusion and future work in Section 7.

2. Motivating Example

We motivate our work in the healthcare context, 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



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

Fig. 1. **Motivating Example.** Figure 1a 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 *belong_to*, while the therapeutic response of *T2* is unknown. Figure 1b 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.

interactions (DDIs) that may affect the response of a medical treatment. 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 bioavailability, and vice versa. If a DDI affects the object 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 drug metabolism. 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 1a shows two polypharmacy oncological treatments encoded in RDF. We extract the known DDIs between the drugs of these treatments from DrugBank[1]. 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 1b 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 represents in a Datalog program were 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 parts of the KG to share more relations the information required to consider both treatments similar. Then, a subsymbolic system, e.g., implemented using a KGE model, can explore this enhanced 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

[1]https://go.drugbank.com

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.

3. Our Proposed Symbolic and Subsymbolic System

3.1. Preliminaries

Knowledge Graphs: A knowledge graph (KG) is a data structure that represents factual knowledge with entities and their relationships using a data graph [6]. KGs are used in countless domains because of their ability to model data in a machine-readable form. Let $\mathcal{T}\_\mathcal{KG} = (O, \mathcal{G})$ be a KG, where O is the set of classes and properties of a unified ontology $O = (Classes, Properties)$. $\mathcal{G} = \langle V, E, L\rangle$ is the data graph where, $L \subseteq Properties$; $Classes \subseteq V$; and for each $e \in V \wedge e \in Classes$; if $(e, type, C) \in E$ then e is of type C.
Ego Networks in Knowledge Graphs: An **ego network** of an ego entity v is defined as $Ego_{\mathcal{G}}(v) = \{v_i | (v, r, v_i) \in E \vee (v_i, r, v) \in E\}$. Figure 2(A) shows the ego network for the ego entity $T2$ as $Ego_{\mathcal{G}}(T2) = \{D2, D3, D4, D5\}$.
Neighborhoods induced by Ego Networks: The **neighborhood** of an ego network $Ego_{\mathcal{G}}(v)$ is represented by $\mathcal{N}_{\mathcal{G}}(Ego_{\mathcal{G}}(v)) = \{(x, r, y) | x \in Ego_{\mathcal{G}}(v) \wedge y \in Ego_{\mathcal{G}}(v) \wedge (x, r, y) \in E\}$. Figure 2(B) shows the neighborhood of the ego networks $Ego_{\mathcal{G}}(T1)$ and $Ego_{\mathcal{G}}(T2)$.
Abstract Target Prediction: An abstract target prediction over a $\mathcal{T}\_\mathcal{KG}$ is defined in terms of a triple $\tau = \langle \tau_h, r, \tau_t \rangle$:

- $\tau_h, \tau_t \in Classes$;
- $r \in Properties$; and
- $\exists v_1 \in \tau_h, v_2 \in \tau_t | (v_1, v_2) \in V$

Figure 2(A) shows a running example for the abstract target prediction $\tau = \langle$*Treatment, belong_to, Response*$\rangle$ where $\tau_h =$ class Treatment and $\tau_t =$ class Response.
Projections in a Knowledge Graph based on an Abstract Target Prediction. Let $\mathcal{G}|_\tau$ be a projection of $\mathcal{T}\_\mathcal{KG}$ by an abstract target prediction τ defined by $\mathcal{G}|_\tau = \{(v_h, r, v_t) | v_h \in \tau_h \wedge v_t \in \tau_t \wedge (v_h, r, v_t) \in E\}$. Figure 2(B) depicts an example of $\mathcal{G}|_\tau$ by the abstract target prediction $\tau = \langle$*Treatment, belong_to, Response*$\rangle$; it is the triple $\langle$*T1, belong_to, low_effect*$\rangle$.
Deductive Databases: A deductive database is a system that can derive deductions, e.g., conclude new facts, from inference rules and facts stored in the database [7]. Deductive systems maintain deductive qualities in the rules that are stated in the system. The language commonly used to specify facts, rules and queries in deductive databases is Datalog. A Datalog program is a set of rules represented as Horn clauses [8]. Horn clauses are represented in the following shape: $L_0 \Leftarrow L_1, ..., L_n$, where each L_i is a literal of the form $p_i(t_1, ..., tk_i)$. P_i is a predicate symbol and t_j are terms. A term is either a constant or a variable. The left-hand side of a Datalog clause is a head, and the right-hand side is its body. Clauses with an empty body represent facts. A Datalog program P must satisfy the following safety conditions; each fact of P is ground, and each variable which 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 limited, where any variable that appears as an argument in a predicate of the body is limited. 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). An example of EDB is the set of facts $InteractsWith(DB00193, DB00642), InteractsWith(DB00642, DB00958)$ expressed in Figure 1a, $G1$. The predicate *InteractsWith* represents interactions between two drugs. Let P be a Datalog program containing the following clauses:

$$r1 : inferred\_interaction(A, X) \Leftarrow InteractsWith(A, X).$$

$$r2 : inferred\_interaction(A, X) \Leftarrow InteractsWith(A, B), inferred\_interaction(B, X).$$

Table 1
Scoring function and complexity of embedding models. Adapted from [9]

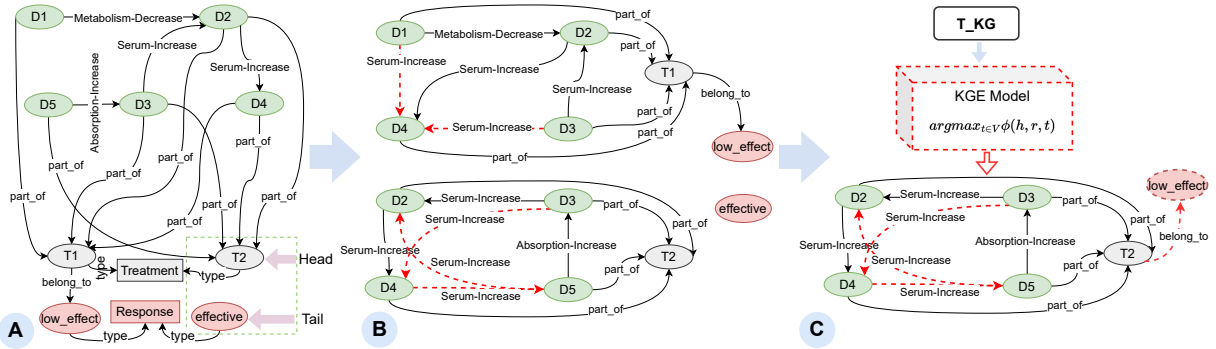
Embedding model	Scoring function	Complexity
HolE	$(h \star t) \times r$	$\mathcal{O}(\|E\|d + \|\mathcal{R}\|d)$
RESCAL	$h^T W_r t = \sum_{i=1}^{d} \sum_{i=1}^{d} w_{ij}^{(r)} h_i t_j$	$\mathcal{O}(\|E\|d + \|\mathcal{R}\|d^2)$
RotatE	$-\|\|h \circ r - r\|\|$	$\mathcal{O}(\|E\|d + \|\mathcal{R}\|d)$
QuatE	$h \times r * t$	$\mathcal{O}(\|E\|d + \|\mathcal{R}\|d)$
TransE	$\|\|h + r - t\|\|$	$\mathcal{O}(\|E\|d + \|\mathcal{R}\|d)$
TransH	$\|\|h_\perp + r - t_\perp\|\|$	$\mathcal{O}(\|E\|d + 2\|\mathcal{R}\|d)$
TransD	$\|\|h_\perp + r - t_\perp\|\|$	$\mathcal{O}(2\|\mathcal{E}\|d + 2\|\mathcal{R}\|d)$
TransR	$\|\|h_r + r - t_r\|\|$	$\mathcal{O}(\|E\|d + \|\mathcal{R}\|d^2)$
UM	$\|\|h - t\|\|$	$\mathcal{O}(\|E\|d)$
SE	$\|\|M_{r,1} h - M_{r,2} t\|\|$	$\mathcal{O}(\|E\|d + 2\|\mathcal{R}\|d^2)$
ERMLP	$w^T g(W[h; r; t])$	$\mathcal{O}(\|E\|d + \|\mathcal{R}\|d + k(3d + 2) + 1$

Rule $r2$ states that exist an *inferred_interaction* between drug A and X, if there is another drug B which interacts with A with the predicate *InteractsWith*, and there is an *inferred_interaction* from B to X. The evaluation results of $r2$ is $\{inferred\_interaction(DB00193, DB00958)\}$, which is observed in Figure 1b, G2.

Deductive Databases for Abstract Target Predictions: A deductive database for an abstract target prediction is defined by $DS|_{\tau_h}(EDB, IDB)$ and $DS|_{\tau_t}(EDB, IDB)$, where $DS|_{\tau_h}(.,.)$ represents the Deductive System for the class τ_h in the abstract target prediction $\langle \tau_h, r, \tau_t \rangle$ and $DS|_{\tau_t}(.,.)$ for the class τ_t. EDB is a subset of $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v))$, i.e., $EDB \subseteq \mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v))$ and the IDB contains the IDB-predicates that allow to deduce new relationships and enhance the ego network, i.e., increasing $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v))$.

Knowledge Graph Embeddings for Abstract Target Predictions: KG embeddings is a machine learning task that learn latent vector representations of entities $v \in V$ and relations $e \in E$ in a KG, preserving their semantic meaning. 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 high-dimensional representations named Knowledge Graph Embeddings (KGE) and is used to infer new facts. The state of the art of KGE methods may be negatively impacted by the data sparsity issue, i.e., true triples that can be used as positive samples to guide KGE training represent only a minor portion. The proposed symbolic system implemented by a Deductive database for abstract target prediction τ alleviates the data sparsity issue by enhancing links in the $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v))$, which are managed as positive triples. The positive triples are used as positive samples to guide the KGE model, improving the performance of the scoring function, see Figure 2(C).

Knowledge Graph Embedding Models: We implemented our symbolic-subsymbolic system in eleven embedding models from different families [9]. Holographic embeddings (*HolE*) [10] computes circular correlation, denotes by $\star$ in Table 1, between the embeddings of head and tail entities. *RESCAL* [11] 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. *RotatE* [12] represents each relation as a rotation from the head entity to the tail entity in the complex latent space. The rotation r is applied to h by operating a Hadamard product (denoted by $\circ$ in Table 1). *QuatE* [13] operates on the quaternion space and learns hypercomplex valued embeddings (quaternion embeddings) to represent entities and relations. *TransE* [14] 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* [15] 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 this hyperplane, where the variables $h_\perp$ and $t_\perp$ denote a projection to the hyperplane w_r of the labeled relation r, and r is the vector of a relation-specific translation in the hyperplane w_r.



Original Knowledge Graph (T_KG): contains polypharmacy cancer treatments from clinical records with the therapeutic response from each of them and DDIs. 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 treatment response value (e.g., low_effect or effective). An ego network of an entity e (head *h* or tail *t*) is the **set of entities connected** to the entity through incoming or out coming links. The eco network of T2 (a.k.a. Ego(T2)) is the set {D2,D3,D4,D5}.

Enhancing links between entities in the ego network of head and tail of target predictions. The ego networks are built for the entity types of interest, i.e. the treatments T1 and T2. Then, define rules that enrich ego networks with neighborhood relationships of head and tail of target predictions.

Build a deductive system based on the defined rules to deduce new relationships in the neighborhood. The red arrows represent the deduced relationships. The deductive system enriches the neighborhood of the entities of interest with positive triples. The positive triples are represented by the set of relations E and the new deduced relationships.

Knowledge Graph Embeddings Model (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 $\phi(h,r,t)$

Positive triples are used as positive samples to guide KGE Model.

KGE model places T1 and T2 nearby in the embedding space after deducing positive triples in their ego networks and predicts the therapeutic response of T2.

Fig. 2. **Running example**. Figure 2 illustrates the proposed steps to enhance the predictive capacity by KGE models. **Step A**: given a KG and an abstract target prediction $\tau = \langle Treatment, belong\_to, Response \rangle$ the ego network $Ego_{\mathcal{G}}(v)$ is defined. **Step B**: illustrates the neighborhood of the ego network $Ego_{\mathcal{G}}(T1)$ and $Ego_{\mathcal{G}}(T2)$ and a deductive system based on the head and tail of τ deduces new relationships to enhances the neighborhoods $\mathcal{N}_{\mathcal{G}}(Ego_{\mathcal{G}}(v))$. **Step C**: depicts a KGE model in which predictive capability is enhanced by symbolic reasoning. The relationships in $\mathcal{N}_{\mathcal{G}}(Ego_{\mathcal{G}}(v))$ improving the link prediction task in $\mathcal{G}|_{\tau}$.

TransR [16] 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* [17] 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 $m \times n$. The Unstructured Model (*UM*) [18] 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 that contain only a single type of relationship. Structured Embedding (*SE*) [19] model defines two matrices $M_{r,1}$ and $M_{r,2}$ to project head and tail entities for each relation. SE can discern between subject and object roles of an entity since it employs different projections for the embeddings of the head and tail entities. *ERMLP* [20] is a model based on 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 1, the variable k corresponds to the number of neurons in the hidden layer.

3.2. Problem Statement

We tackle the *problem of discovering relationships over a KG*. Given a KG and an abstract target prediction, we aim to enhance the predictive capacity of the link prediction task. Let $\mathcal{T}\_\mathcal{KG}' = (O, \mathcal{G}')$ be an ideal knowledge graph that contains all the existing relations between entities in V, where $\mathcal{G}' = \langle V, E', L \rangle$ is the data graph. $\mathcal{T}\_\mathcal{KG}$ is the actual knowledge graph which only contains a portion of the edges represented in $\mathcal{T}\_\mathcal{KG}'$, i.e., $E \subseteq E'$; it represents those relations that are known and is not necessarily complete. Let $\Delta(E', E) = E' - E$ be the set of relations existing in the ideal knowledge graph $\mathcal{T}\_\mathcal{KG}'$ that are not represented in $\mathcal{T}\_\mathcal{KG}$. Let $\mathcal{T}\_\mathcal{KG}_{\text{comp}} = (O, \mathcal{G}_{\text{comp}})$ be a *complete* knowledge graph where $\mathcal{G}_{\text{comp}} = \langle V, E_{comp}, L \rangle$ is a data graph, which includes a relation for each possible combination of entities in V, i.e., $E \subseteq E' \subseteq E_{\text{comp}}$. Let $\mathcal{G}'|_{\tau}$ be a projection of $\mathcal{G}'$ by an abstract target prediction τ.

Given a relation $e \in \Delta(E_{\text{comp}}, E)$ and an abstract target prediction τ, the problem of discovering relations consists of determining whether $e \in E'$, i.e., if a relation e corresponds to an existing relation in the ideal graph $\mathcal{G}'|_{\tau}$. We are

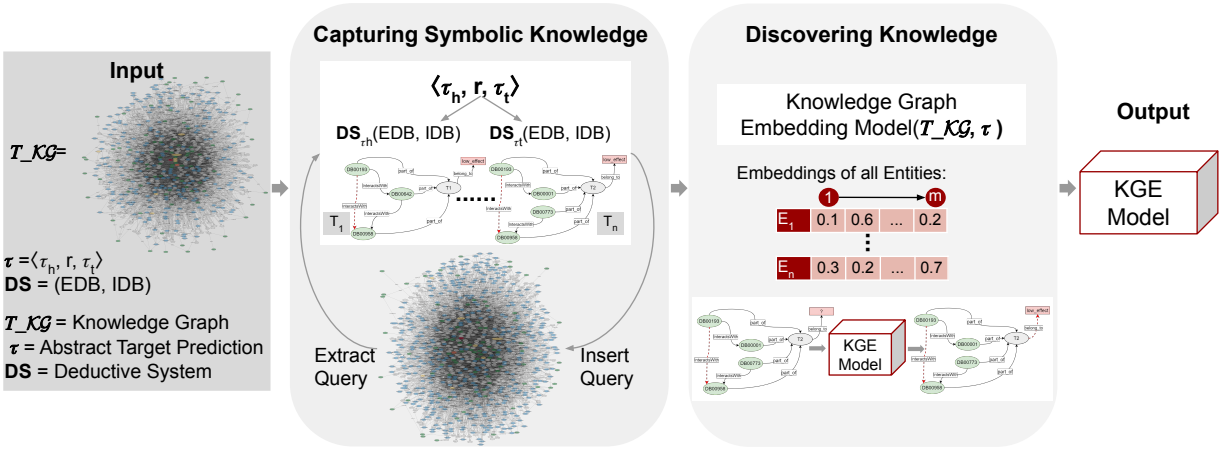


Fig. 3. **Approach**. The input is a Knowledge Graph ($\mathcal{T}\_\mathcal{KG}$), an abstract target prediction τ, and a deductive system for τ, and returns a KGE model. Capturing symbolic knowledge stage, the Deductive System $DS|_{\tau_h}(EDB, IDB)$ and $DS|_{\tau_t}(EDB, IDB)$ deduces relationships in the neighborhoods $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v))$ of the ego network $Ego_\mathcal{G}(v)$. Then, in discovering knowledge stage, $\mathcal{T}\_\mathcal{KG}$ and the resulting neighborhoods $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v))$ are embedded by a KGE algorithm to solve the link prediction task in $G|_\tau$.

interested in finding the maximal set of relationships or edges E_a that belong to the ideal $\mathcal{G}'|_\tau$, i.e., find a set E_a that corresponds to a solution of the following optimization problem:

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

3.3. Proposed Solution

Our proposed solution resorts to symbolic reasoning implemented by a deductive database to enhance the predictive capacity of the link prediction task solved by knowledge graph embedding (KGE) models. The approach assumes that a link prediction problem is defined in terms of an abstract target prediction $\tau = \langle \tau_h, r, \tau_t \rangle$ over a knowledge graph $\mathcal{T}\_\mathcal{KG} = (O, \mathcal{G})$.

A Symbolic System: Deductive systems $DS|_{\tau_h}(EDB, IDB)$ and $DS|_{\tau_t}(EDB, IDB)$ correspond to the deductive databases for the abstract target prediction τ. Thus, for each entity v_h in τ_h (resp. v_t in τ_t), $DS|_{\tau_h}(EDB, IDB)$ (resp. $DS|_{\tau_t}(EDB, IDB)$) defines relations between entities in the neighborhoods $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v_t))$ (resp. $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v_t))$), induced by the ego network $Ego_\mathcal{G}(v_h)$. The computational method executed to empower a neighborhood $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v))$ is built on the results of deductive databases to compute the minimal model of the deductive database[8]. This minimal model is defined in terms of the fixed-point assignment $\sigma^{\mathcal{N}_\mathcal{G}(.)}_{\text{MINFIX}}$, that deduces relationships between entities v_i and v_j in the neighborhoods $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v_h))$ for each entity v_h in τ_h (resp. v_t in τ_t). The minimal model for $DS|_{\tau_h}(EDB, IDB)$ (resp. $DS|_{\tau_t}(EDB, IDB)$) can be computed in polynomial time in the overall size of the neighborhoods $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v))$ for all the entities in τ_h (resp. τ_t).

A Subsymbolic System: A model to learn Knowledge Graph Embeddings solves the abstract target prediction τ and completes the $\mathcal{T}\_\mathcal{KG} = (O, \mathcal{G})$ with links of the type $\langle \tau_h, r, \tau_t \rangle$.

The Integration of Symbolic and Subsymbolic Systems: The neighborhoods $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v_t))$ and $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v_t))$ are extended with explicit relationships among entities in the ego networks of entities in $Ego_\mathcal{G}(v_h)$ (resp. $Ego_\mathcal{G}(v_h)$). As a result, the symbolic system implemented by $DS|_{\tau_h}(EDB, IDB)$ and $DS|_{\tau_t}(EDB, IDB)$ alleviate the data sparsity issues in $G|_\tau$ that may negatively affect the process of learning the KGE for the abstract target prediction τ.

3.4. The Symbolic and Subsymbolic System Architecture

Figure 3 depicts the architecture that implements the proposed approach. The architecture receives a Knowledge Graph $\mathcal{T}\_\mathcal{KG} = (O, \mathcal{G})$, an abstract target prediction $\tau = \langle \tau_h, r, \tau_t \rangle$, and **Deductive Databases for Abstract Target**

Table 2
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
Oncologycal Surgery	9
Tumor Stage	6
Publications	178'265
Drugs	8'453
Drug-Drug Interactions	1'550'586

Predictions, and 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 Systems for an abstract target prediction τ are deduced by means of Datalog programs. The minimal model of the Datalog programs correspond to in the neighborhoods of the entities in τ_h and τ_t. Once $\mathcal{T}\_\mathcal{KG}$ is augmented with new relationships, KGE learns a latent representation of the entities τ_h and τ_t in a high-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 [21, 22]. In this paper, we illustrate the use of our proposed symbolic and subsymbolic system if in the task of predicting treatment effectiveness, as shown next.

4. A Use Case: Prediction of Polypharmacy Treatment Effectiveness

As a proof concept, we have implemented a Deductive System on top of a Treatment Knowledge Graph ($\mathcal{T}\_\mathcal{KG}$). 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*.

4.1. A Knowledge Graph for Lung Cancer

The P4-LUCAT consortium[2] 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; only approved interactions are added. The type and effect of the interactions are extracted by using named entity and linking methods implemented by Sakor et al. [23]. These methods have also been used to extract DDIs in covid-19 and lung cancer [24, 25]. Table 2 contains a summary of the number of annotations by classes in the Lung Cancer Knowledge Graph.

Figure 4 describes a Lung Cancer patient in the Lung Cancer Knowledge Graph. The patient *P1* is in stage II and has a surgery. Also, *P1* received a treatment on *10.07.2020* with an effective therapeutic response. In that treatment, *P1* was treated with a combination of chemotherapy drugs and one non-oncological drug. A Drug-Drug Interactions with the effect and the impact was reported.

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

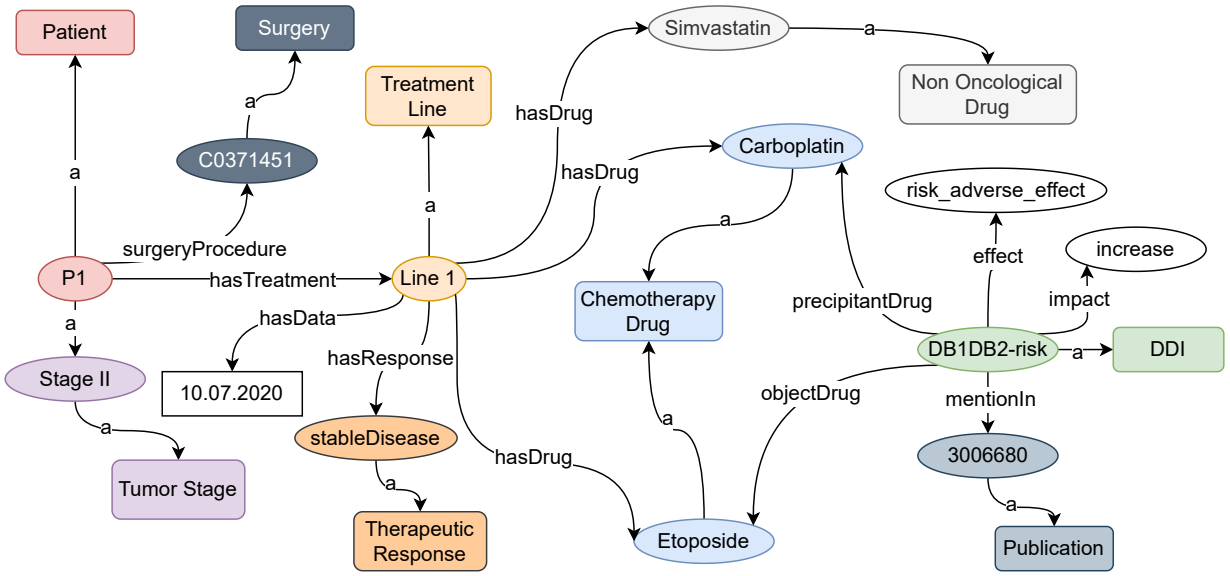
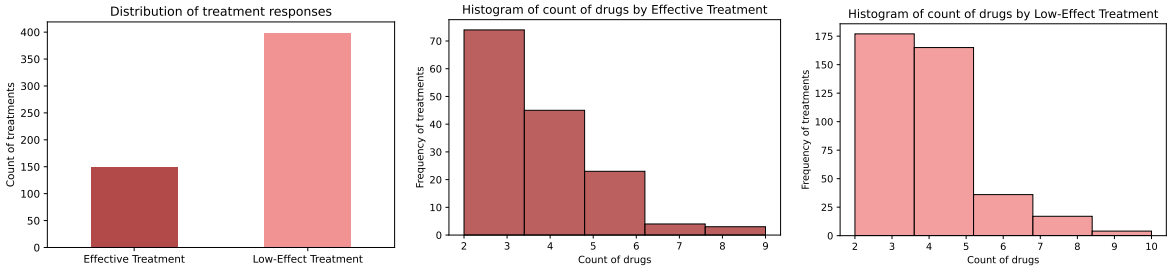


Fig. 4. **Representation of a patient in the Lung Cancer Knowledge Graph**.



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

Fig. 5. **Descriptive analysis of the treatment responses.**

The input $\mathcal{T}\_\mathcal{KG}$ 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 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 5 depicts a descriptive analysis of the treatment response according to the data extracted from the clinical records. Figure 5a shows the treatment response distribution, where there are 149 *effective* treatments and 399 *low-effect* treatments. Figure 5b and 5c 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.

For each treatment $t_i \in \mathcal{T}$, the DDIs and their effect are known from DrugBank [26]. Then, the treatments, the treatment response, the drugs, DDIs and effects by each treatment are managed in $\mathcal{T}\_\mathcal{KG}$. Figure 6 shows a portion of $\mathcal{T}\_\mathcal{KG}$. The node colors correspond to the type of entity, and the edges represent relationships amount the drugs grouped in a treatment. The $\mathcal{T}\_\mathcal{KG} = (O, \mathcal{G})$ is defined as follows:

- The types Drug, Treatment, DDI, Effect of DDI, and Treatment Response belong to *Classes*.

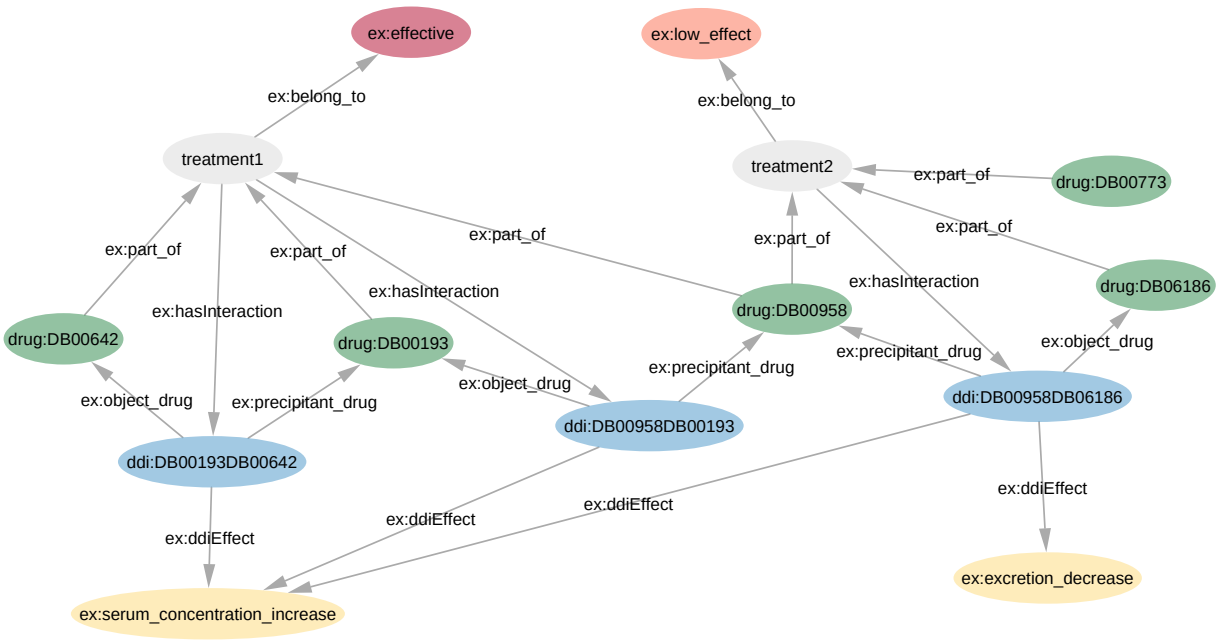


Fig. 6. **Portion of Treatment Knowledge Graph** $\mathcal{T}\_\mathcal{KG}$. The *treatment1* is composed by three drugs represented by the nodes, *drug:DB00642*, *drug:DB00193*, and *drug:DB00958*. The *treatment2* also contains three drugs and shares *drug:DB00958* with *treatment1*. The blue node *ddi:DB00958DB06186* represents a DDI in the *treatment2* where the *drug:DB00958* is the precipitant and *drug:DB06186* is the object drug. The effect of this DDI is represented by the yellow node *ex:excretion_decrease*. Then, the treatment *treatment2* has a low effective response represented by the property *ex:belong_to*.

- 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 *ex:belong_to*, *part_of*, *ex:precipitant_drug*, *ex:object_drug*, *ex:ddiEffect*, and *ex:hasInteraction* correspond to labels in L.

4.2. The Abstract Target Prediction Task

Albeit illustrated in the context of treatment-response, the proposed method is domain-agnostic. It only requires the definition of deductive systems to enhance the relationships among the entities in the ego networks of the head and tail classes of an abstract predictive task that define the link prediction problem. Figure 2 illustrates the proposed steps to enhance the predictive capacity by knowledge graph embedding models. The input in Figure 2 **step (A)** is a KG and an abstract target prediction $\tau = \langle$*Treatment, belong_to, Response*$\rangle$; the head of a target prediction τ_h is a treatment, while the tail τ_t is a response. Thus, the link prediction task predicts links in the projection of a graph $G|_\tau$. The triple $\langle$*T1, belong_to, low_effect*$\rangle$ is an example of a triple to predict. Then, the ego network of an entity v (head h or tail t) is computed. The ego networks for the entities $T1$ and $T2$ are composed by $Ego_\mathcal{G}(T1) = \{D1, D2, D3, D4\}$ and $Ego_\mathcal{G}(T2) = \{D2, D3, D4, D5\}$, respectively. **Step (B)** in Figure 2 depicts the neighborhood of the ego networks $Ego_\mathcal{G}(T1)$ and $Ego_\mathcal{G}(T2)$. Then, a Deductive System based on the τ_h and τ_t deduces new relationships to enhance the links in the $Ego_\mathcal{G}(v)$; arrows colored in red represent the deduced relationships. EDB is a subset of $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v))$, i.e., $EDB \subseteq \mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v))$ and the IDB allows deducing new relationships and enhance the ego network, i.e., increasing $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v))$. **Step (C)** in Figure 2 illustrates a KGE model in which the predictive capacity is enhanced by the symbolic reasoning. The relationships in $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v))$, used as a positive example, guide the KGE model, improving the link prediction task in the $G|_\tau$. Treatment $T2$ is predicted to have a response *low_effect* $\langle$*T2, belong_to, low_effect*$\rangle$, i.e., $T1$ and $T2$ are nearby in the embedding space after enhancing the $\mathcal{N}_\mathcal{G}(Ego_\mathcal{G}(v))$.

4.3. Deductive Databases for Abstract Target Predictions about Treatment Effectiveness

A DDI is deduced when a set of drugs are taking together and is represented as a relation in the minimal model of the deductive database. The extensional database corresponds to statements about interactions between drugs stated in $\mathcal{T}\_\mathcal{KG}$. The ground predicates included in the Extensional Database are the following; they are extracted from the KG by executing SPARQL queries:

$rule_1(serum, increase).$	$rule_2(serum, decrease).$	*precipitant(DB00958DB06186, DB00958).*
$rule_1(metabolism, decrease).$	$rule_2(metabolism, increase).$	*object(DB00958DB06186, DB06186).*
$rule_1(absorption, increase).$	$rule_2(absorption, decrease).$	*effect(DB00958DB06186, excretion).*
$rule_1(excretion, decrease).$	$rule_2(excretion, increase).$	*impact(DB00958DB06186, decrease).*

SPARQL queries in Listing 1 and Listing 2 declaratively define the ground $rule_1$ and $rule_2$ in the EDB. Both queries are executed on top of the $\mathcal{T}\_\mathcal{KG}$; 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 1: **SPARQL query to ground the extensional predicate** $rule_1(E, I)$

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

The facts included in the ground predicates *precipitant, object, effect*, and *impact* from the EDB are extracted using the construct query of Listing 3. The EDB contains thousands of facts for the those predicate; therefore, only 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 deduced as a result of the combination of treatment's drug. These rules are negation free; thus, the interpretation of the deductive database correspond 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 rate of absorption, distribution, metabolism, or excretion 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

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

consequences on the therapeutic efficacy and toxicity of the drug. The following Datalog rules state the effect of pharmacokinetic DDIs:

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

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

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

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

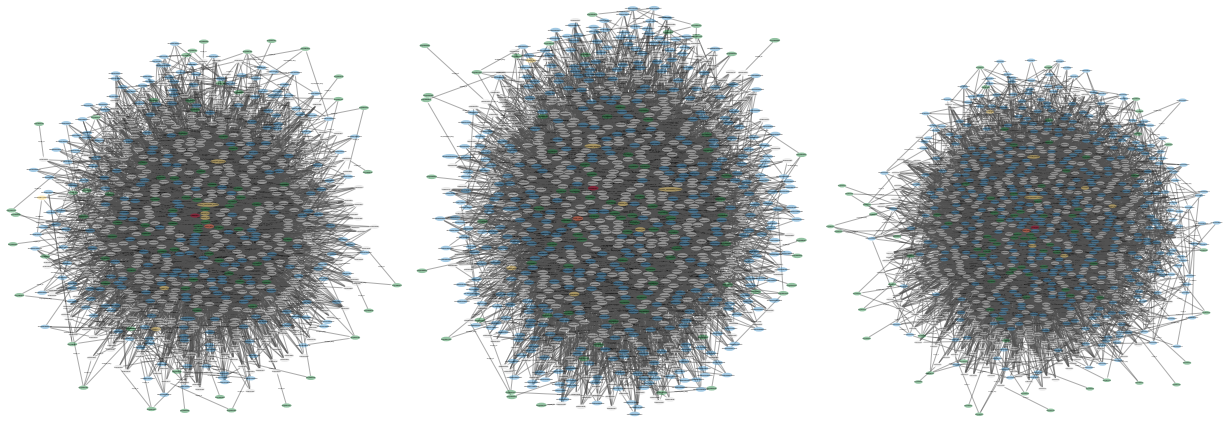
Rule (2) states the base case of the IDB. The predicate symbol *ddi* represents the DDIs with their effect and impact in $\mathcal{T}\_\mathcal{KG}$. Precipitant drug A generates effect E (e.g., absorption, excretion, metabolism, serum concentration) with impact I (e.g., increase or decrease) in object drug B. The predicate symbol $inferred\_ddi$ expresses a deduced DDI, where the first term is the precipitant drug, the second and third term 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 $\mathcal{T}\_\mathcal{KG}$. 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 $\mathcal{T}\_\mathcal{KG}$; 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 $\mathcal{T}\_\mathcal{KG}$ using the query shown in Listing 4. From the motivating example, we can observe that by applying the DDI deductive system to the treatment `T1` in Figure 1a, a new DDI is deduced in Figure 1b; it represents a new true triple enhancing the treatment information, reducing thus, data sparsity.

4.4. Link Prediction based on Knowledge Graph Embedding Models

Once the deductive system deduces new DDIs, a Knowledge Graph Embedding model is applied to learn a latent representation of the entities in a high-dimensional space. Meaning that in each treatment where the Deductive System deduce new DDIs the embeddings of the treatments may change because of the scoring function of the

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



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

Fig. 7. **Benchmarks to evaluate.** Figure 7a represents the $\mathcal{T}\_\mathcal{KG}_{basic}$ and it includes treatments from clinical records and pharmacokinetic DDI extracted from Drugbank. Figure 7b represents the $\mathcal{T}\_\mathcal{KG}$ and it includes treatments from clinical records, pharmacokinetic DDI extracted from Drugbank, and a new set of pharmacokinetic DDI deduced by the DDI Deductive System. Figure 7c represents the $\mathcal{T}\_\mathcal{KG}_{random}$ and includes treatments from clinical records, pharmacokinetic DDI extracted from Drugbank, and the same number of new links deduced in $\mathcal{T}\_\mathcal{KG}$ is generated randomly.

embedding model. Symbolic and subsymbolic system are highly complementary to each other. Subsymbolic AI systems are able to solve complex problems which are impossible to analyze by humans to draw conclusions or make predictions. Subsymbolic methods are generally robust to data noise, while symbolic systems are vulnerable to data noise, which contrasts with the strength of subsymbolic approaches.

5. Experimental Study

We empirically assess the impact that the DDIs encoded in $\mathcal{T}\_\mathcal{KG}$ has on the behavior of our approach. In particular, this work explores the following research questions: **RQ1** Can the problem of predicting treatment effectiveness be effectively modeled as a problem of link prediction? **RQ2** Can the Symbolic System for an abstract target prediction improve the link prediction capacity of the KGEs? **RQ3** Can knowledge encoded in drug-drug interactions enhance the accuracy of the predictive task?

5.1. Experiment Setup

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

Table 3
Statistics of Knowledge Graph. Metrics to measure size, diversity, and sparsity in Knowledge Graph

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

5.1.1. Benchmarks

We conduct our evaluation over three Knowledge Graphs represented in Figure 7. $\mathcal{T}\_\mathcal{KG}_{basic}$ is the Knowledge Graph which only contains for each polypharmacy treatment the DDIs and their effect extracted from Drugbank. The second Knowledge Graph, $\mathcal{T}\_\mathcal{KG}$ includes not only the DDIs extracted from DrugBank, but also the ones deduced by Deductive Database for the abstract target prediction $\tau = \langle$*Treatment, belong_to, Response*$\rangle$, i.e., it contains new deduced DDIs and their effects. Lastly, the third Knowledge Graph, $\mathcal{T}\_\mathcal{KG}_{random}$ is created from $\mathcal{T}\_\mathcal{KG}_{basic}$; it also includes the same number of links included in $\mathcal{T}\_\mathcal{KG}$ 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.

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

The PyKEEN (Python KnowlEdge EmbeddiNgs) framework [27] 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 a statistical robustness, we apply 5-folds cross-validation. For evaluate the performance of embeddings methods, we measure the metrics: $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$, $F1\text{-}Score = \frac{2*(precision*recall)}{(precision+recall)}$.

5.1.3. Implementations

The pipeline for predicting polypharmacy treatment response has been implemented in Python 3.9. Experiments were 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[3] to develop the Deductive System and the library PyKEEN[4], to learn the embeddings.

5.2. Metrics to Characterize the Benchmarks

Table 3 shows the statistics of the three KGs. We considered the metrics, Number of Triples (T), Entities (E), and Relations (R) to measure 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 means that all possible relations are equally probable and lower values means one or more relations have a high probability. The values of the metric RE means that all possible relations in $\mathcal{T}\_\mathcal{KG}$ are more equally probable than all possible relations in $\mathcal{T}\_\mathcal{KG}_{basic}$ and $\mathcal{T}\_\mathcal{KG}_{random}$. 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

[3]https://sites.google.com/site/pydatalog/home
[4]https://pykeen.readthedocs.io/en/stable/index.html

RD means average triples per relation and *ED* is the average triples per entity. $\mathcal{T}\_\mathcal{KG}_{basic}$ has the lower average triples per relation and entity while $\mathcal{T}\_\mathcal{KG}$ and $\mathcal{T}\_\mathcal{KG}_{random}$ have the higher average triples per entity. The metrics evaluated in Table 3 are defined in the paper [28], implemented by us and available at [5].

5.3. Impact of Capturing Symbolic Knowledge

Figure 8 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 [6]. We can notice how $DS|_{\tau_h}(EDB, IDB)$ for the abstract target prediction $\tau = \langle$*Treatment, belong_to, Response*$\rangle$ is impacting on the KGE models. Figure 8a to 8c and Figure 8g to 8i show the score values of the entities predicted on the link prediction task given the predicate *ex:belong_to* and object *effective* by the *TransH* and *RotatE* models, respectively. Figure 8d to 8f and Figure 8j to 8l report on the score values of the entities predicted given the predicate *ex:belong_to* and object *low-effect* by the *TransH* and *RotatE* models, respectively. We can observe that the models have different behaviors for each KG. 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.

5.4. Evaluating the performance of our integrated Symbolic-Subsymbolic System

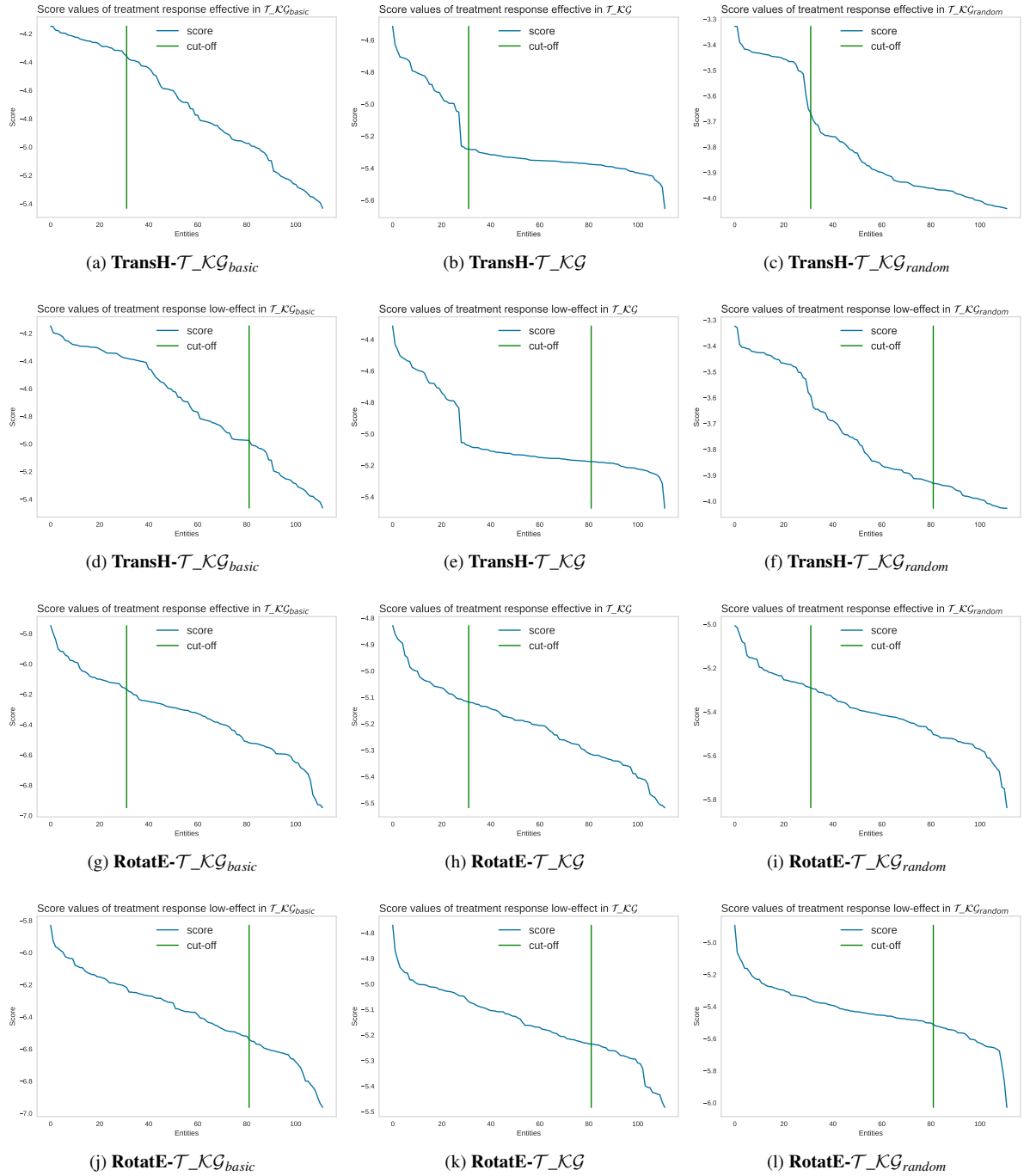
The selected portions of entities predicted were measured precision, recall and f1-score on average because of cross-validation. Figure 9 and Figure 10 show the evaluation of the Link Prediction task through Uniform negative sampling and Bernoulli negative sampling, respectively. Uniform sampling randomly chooses 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 a 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 higher probability of corrupting the tail. Figure 9 and 10 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 $\mathcal{T}\_\mathcal{KG}$ obtained in the three metrics in both negative sampling techniques. In addition, the worst performance is observed in $\mathcal{T}\_\mathcal{KG}_{random}$. 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.

5.5. Discussion

The techniques proposed in this paper 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 11 shows a box plot of cosine similarity. Five treatments with a low-effect response were selected and $\mathcal{T}\_\mathcal{KG}_{basic}$ misclassifies them, but $\mathcal{T}\_\mathcal{KG}$ 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 $\mathcal{T}\_\mathcal{KG}$ than in $\mathcal{T}\_\mathcal{KG}_{basic}$. The first quartile, median and third quartile values in the boxplot are higher in $\mathcal{T}\_\mathcal{KG}$ than in $\mathcal{T}\_\mathcal{KG}_{basic}$. Therefore, these outcomes put in evidence the quality of the deduced links in $\mathcal{T}\_\mathcal{KG}$ and their impact on the accuracy of the KGE models in the resolution of the task of predicting treatment effectiveness.

Figure 12 shows the distribution of DDIs by treatment in $\mathcal{T}\_\mathcal{KG}_{basic}$, $\mathcal{T}\_\mathcal{KG}$, and $\mathcal{T}\_\mathcal{KG}_{random}$. 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 function Kernel Density Estimation (KDE) 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 $\mathcal{T}\_\mathcal{KG}$ have less density

[5] https://github.com/SDM-TIB/Statistics_KnowledgeGraph
[6] https://github.com/arivasm/Neuro-Symbolic_Treatment-Response.git



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

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

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

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

Fig. 8. **Score value of the predicted entities.** The green line represents the cut-off at the 27 and 73 percentiles for the three KGs.

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 put in evidence the crucial role that implicit DDIs have on a treatment's response and the need of deducing them using symbolic systems.

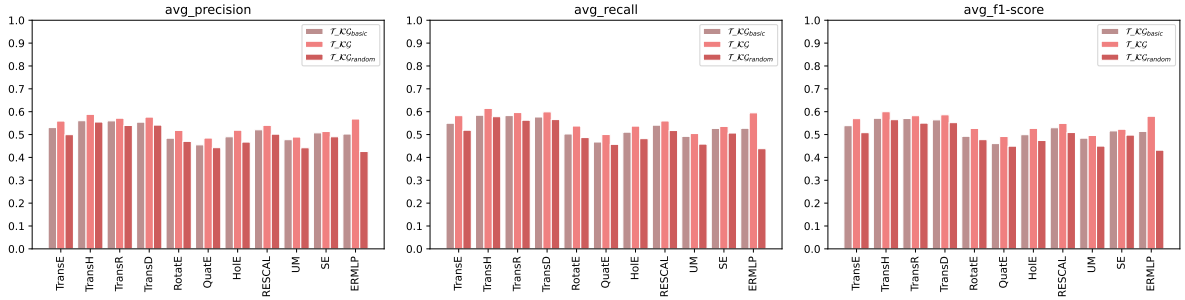


Fig. 9. Evaluation of the Link Prediction task in terms of precision, recall and f-measure. Utilizing Uniform negative sampling.

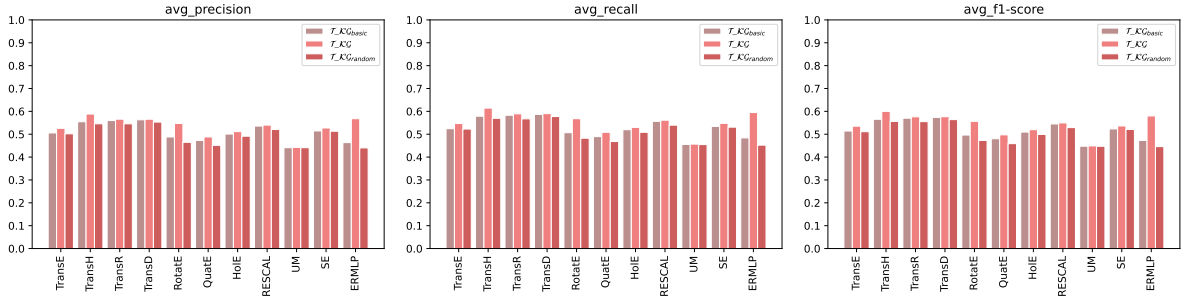


Fig. 10. Evaluation of the Link Prediction task in terms of precision, recall and f-measure. Utilizing Bernoulli negative sampling.

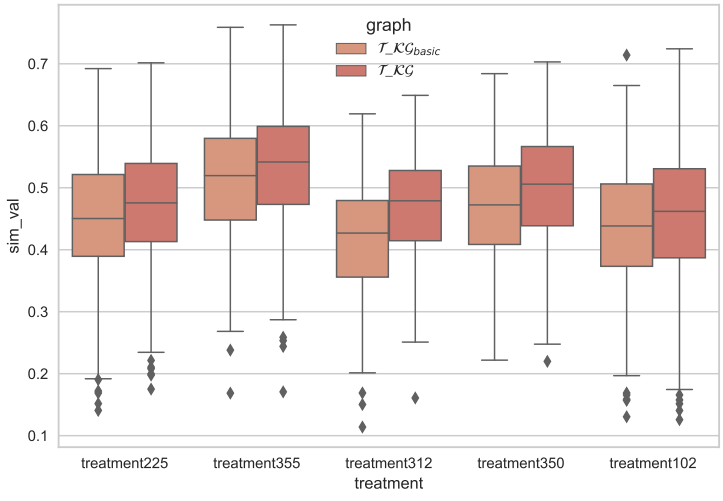


Fig. 11. **Boxplot of Cosine Similarity**.

Analysis of deduced DDI by Treatment classes: Figure 13 exhibits the distribution of DDIs by treatment response in both $\mathcal{T}\_\mathcal{KG}_{basic}$ and $\mathcal{T}\_\mathcal{KG}$. 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.

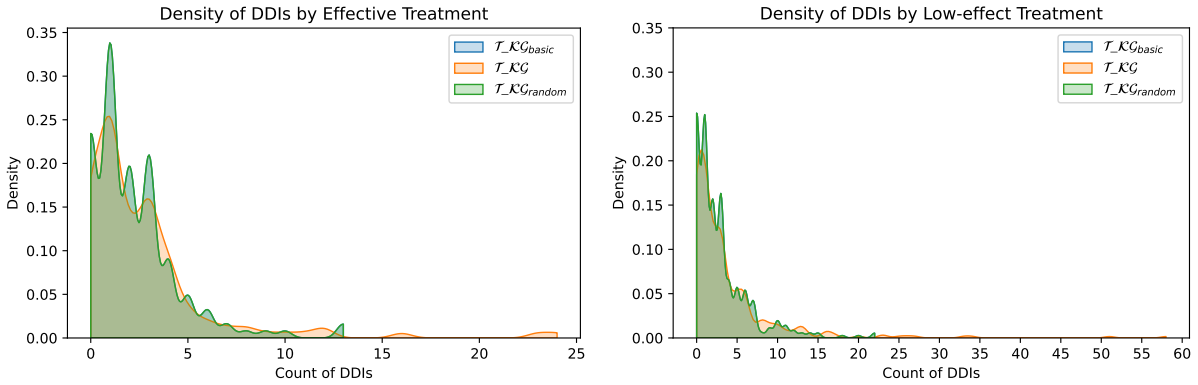


Fig. 12. **The distribution of DDIs by treatment for each KG**. Figure 12a shows the density of treatments by DDIs for the treatment response *effective* in $\mathcal{T}\_\mathcal{KG}_{basic}$, $\mathcal{T}\_\mathcal{KG}$, and $\mathcal{T}\_\mathcal{KG}_{random}$. Figure 12b shows the density of treatments by DDIs for the treatment response *low-effect*.

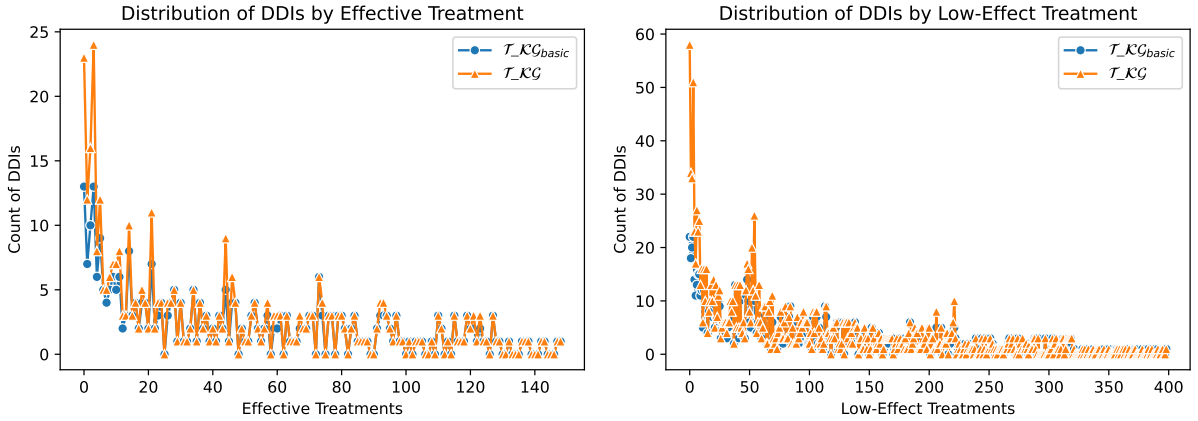


Fig. 13. Distribution of DDIs by treatment response.

6. Related Work

6.1. Neuro-Symbolic Artificial Intelligence

Neuro-Symbolic Artificial Intelligence is a highly active area that has been studied for decades [29]. Neuro-symbolic AI focuses on integrating symbolic and sub-symbolic systems. Several approaches employ translation algorithms from a symbolic representation to a subsymbolic representation and vice versa [30]. The aim is to provide a neural-symbolic implementation of logic, a logical characterisation of a neural system, or a hybrid learning system that contributes features of symbolic and sub-symbolic systems [29, 31]. Real applications are possible in areas with social relevance and high economic impacts, such as bioinformatics, robotics, fraud prevention, and the semantic web [30]. 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. An example of a neural-symbolic system in the field of bioinformatics is the Connectionist Inductive Learning and Logic Programming (CILP) [3]. In the field of vision-based tasks, such as semantic image labelling, high-performance systems have been produced. Karpathy et al. [4] propose an approach is introduced for the recognition and labelling tasks for 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 a multimodal embedding. The emerging system performs better than classical approaches where tasks involving semantic descriptions are associated with databases that contained 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, in our approach, the deductive database is addressed to an abstract target prediction which renders the computational complexity polynomial-time. Thus, we show the positive impact that completing KG via deductive system have in the overall performance of a predictive model implemented using KGEs.

6.2. Knowledge Graph Embedding in Biomedical field

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. [32] develop a process for constructing and reasoning a 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 by 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 ConvKB embedding model. Karim et al. [33] 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), which was modified to predict DDIs. The results show that the combination of KGE and Conv-LSTM performs the state-of-the-art results.

The above-mentioned research aims to discover reliable knowledge based on knowledge graph 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 Neuro-Symbolic AI system enabling expressive reasoning and robust learning to improve the predictive capability of KGE models.

6.3. Polypharmacy side effect prediction and Drug-Drug Interactions prediction

In recent years, there has been a growing interest in Pharmacovigilance. Extensive research has been conducted to predict potential DDI. One of the approaches to predict potential DDI is based on similarity [34–37], 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. [34] propose a probabilistic approach for inferring unknown DDIs from a network of multiple drug-based similarities and known DDIs. They used probabilistic programming framework Probabilistic Soft Logic. This symbolic approach predicts three types of interacctions [34], 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. Futhermore, they considering seven drug–drug similarities. Thus, they found five novel DDIs validated by external sources. A framework to predict DDIs is presented in [37], 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 [38]. 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 [39–42]. Beyond knowing that a pair of drugs interact, it is essential to know the effect of DDI in polypharmacy treatments. In [40] 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 in 101 out of 106 DDI types the model obtains equal or better results than baseline methods. Also, they demonstrate how adding the features GSP and TSP increases the accuracy of DDIs prediction. Marinka Zitnik et al. [39] presents Decagon, an approach for predicting 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 predict DDIs and their effects between pair 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 implic defintion of new DDIs generated as a result of the combination of multiple drugs in a treatment. Since cancer treatment schemas are usually composed for more than one drug, and also patients may have several co-existing disease that require additional medications, it is of significant relevance holistically deducing DDIs.

7. Conclusions and Future Work

This paper addresses the problem of Neuro-Symbolic AI integration, enabling expressive reasoning and robust learning to discover relationships over knowledge graphs. We have presented a novel approach that integrates symbolic-subsymbolic systems to enhance the predictive capacity of the abstract target prediction in KGE models. The symbolic system is implemented by a deductive database defined for an abstract target prediction over a KG. Our proposed solution builds the ego networks of the head and tail of the abstract target prediction to deduce new relationships and enhance the ego network; it is able enhances the ego networks of the abstract target prediction and effectively predict treatment effectiveness. Further, the subsymbolic system implemented by a KGE model enhances the predictive capacity of the abstract target prediction and completes the KG. We assess the performance of our approach in a knowledge graph for lung cancer to discover treatment effectiveness. Predicting treatment effectiveness is effectively modelled as a problem of link prediction and exploiting DDI Deductive System improves existing embedding models by performing the treatment prediction task. Results of a 5-fold cross-validation process demonstrate that our approach, integrating neuro-symbolic systems, improves the eleven KGE models evaluated. Our approach using the reasoning of the symbolic system is able enhances the ego networks of the abstract target prediction and effectively predict treatment effectiveness. Thus, our work broadens the repertoire of Neuro-Symbolic AI systems for discovering relationships over a KG. As for future work, we envision having a more fine-grained description of the DDIs and a descriptive profile of the patients and improving the model.

Acknowledgements

Ariam Rivas is supported by the German Academic Exchange Service (DAAD). The authors thank the BIOMEDAS program for training. This work has been partially supported by the EU H2020 RIA funded projects

CLARIFY with grant agreement No 875160, EraMed P4-LUCAT No 53000015, and Opertus Mundi GA 870228, as well as, the Federal Ministry for Economic Affairs and Energy (BMWi) project SPEAKER (FKZ 01MK20011A).

References

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

[2] H.K.G. Fernlund, A.J. Gonzalez, M. Georgiopoulos and R.F. DeMara, Learning tactical human behavior through observation of human performance, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **36**(1) (2006), 128–140. doi:10.1109/TSMCB.2005.855568.

[3] A.S. d'Avila Garcez, K. Broda and D.M. Gabbay, Neural-symbolic learning systems - foundations and applications, in: *Perspectives in neural computing*, 2002.

[4] A. Karpathy and L. Fei-Fei, Deep Visual-Semantic Alignments for Generating Image Descriptions, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(4) (2017), 664–676–. doi:10.1109/tpami.2016.2598339. http://dx.doi.org/10.1109/TPAMI.2016.2598339.

[5] A. Rivas and M.-E. Vidal, Capturing Knowledge about Drug-Drug Interactions to Enhance Treatment Effectiveness, in: *Proceedings of the 11th on Knowledge Capture Conference*, K-CAP '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 33–40–. ISBN 9781450384575. doi:10.1145/3460210.3493560.

[6] C. Gutiérrez and J.F. Sequeda, Knowledge graphs, *Communications of the ACM* **64**(3) (2021), 96–104.

[7] R. Ramakrishnan and J.D. Ullman, A survey of deductive database systems, *The Journal of Logic Programming* **23**(2) (1995), 125–149. doi:https://doi.org/10.1016/0743-1066(94)00039-9. https://www.sciencedirect.com/science/article/pii/0743106694000399.

[8] S. Ceri, G. Gottlob and L. Tanca, What you always wanted to know about Datalog (and never dared to ask), *IEEE Transactions on Knowledge and Data Engineering* **1**(1) (1989), 146–166. doi:10.1109/69.43410.

[9] A. Rossi, D. Barbosa, D. Firmani, A. Matinata and P. Merialdo, Knowledge Graph Embedding for Link Prediction: A Comparative Analysis, *ACM Trans. Knowl. Discov. Data* **15**(2) (2021). doi:10.1145/3424672.

[10] M. Nickel, L. Rosasco and T. Poggio, Holographic Embeddings of Knowledge Graphs, arXiv, 2015. doi:10.48550/ARXIV.1510.04935. https://arxiv.org/abs/1510.04935.

[11] M. Nickel, V. Tresp and H.-P. 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, Omnipress, Madison, WI, USA, 2011, pp. 809–816–. ISBN 9781450306195.

[12] Z. Sun, Z.-H. Deng, J.-Y. Nie and J. Tang, RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space, arXiv, 2019. doi:10.48550/ARXIV.1902.10197. https://arxiv.org/abs/1902.10197.

[13] S. Zhang, Y. Tay, L. Yao and Q. Liu, Quaternion Knowledge Graph Embeddings., in: *NeurIPS*, 2019, pp. 2731–2741. http://papers.nips.cc/paper/8541-quaternion-knowledge-graph-embeddings.

[14] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston and O. Yakhnenko, Translating Embeddings for Modeling Multi-relational Data, in: *Advances in Neural Information Processing Systems*, Vol. 26, C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger, eds, Curran Associates, Inc., 2013. https://proceedings.neurips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.

[15] Z. Wang, J. Zhang, J. Feng and Z. Chen, Knowledge Graph Embedding by Translating on Hyperplanes, *Proceedings of the AAAI Conference on Artificial Intelligence* **28**(1) (2014). https://ojs.aaai.org/index.php/AAAI/article/view/8870.

[16] Y. Lin, Z. Liu, M. Sun, Y. Liu and X. Zhu, Learning Entity and Relation Embeddings for Knowledge Graph Completion, *Proceedings of the AAAI Conference on Artificial Intelligence* **29**(1) (2015). https://ojs.aaai.org/index.php/AAAI/article/view/9491.

[17] G. Ji, S. He, L. Xu, K. Liu and J. Zhao, 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)*, Association for Computational Linguistics, Beijing, China, 2015, pp. 687–696. doi:10.3115/v1/P15-1067. https://aclanthology.org/P15-1067.

[18] A. Bordes, X. Glorot, J. Weston and Y. Bengio, A Semantic Matching Energy Function for Learning with Multi-relational Data, *Machine Learning* (2014), 1–30. doi:10.1007/s10994-013-5363-6. https://hal.archives-ouvertes.fr/hal-00835282.

[19] A. Bordes, J. Weston, R. Collobert and Y. Bengio, Learning Structured Embeddings of Knowledge Bases, in: *25th Conference on Artificial Intelligence (AAAI)*, San Francisco, United States, 2011, pp. 301–306. https://hal.archives-ouvertes.fr/hal-00752498.

[20] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun and W. Zhang, 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, Association for Computing Machinery, New York, NY, USA, 2014, pp. 601–610–. ISBN 9781450329569. doi:10.1145/2623330.2623623.

[21] A. Rivas, I. Grangel-González, D. Collarana, J. Lehmann and M.-E. Vidal, 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.

[22] A. Rivas, I. Grangel-Gonzalez, D. Collarana, J. Lehmann and M.-e. Vidal, Discover Relations in the Industry 4.0 Standards Via Unsupervised Learning on Knowledge Graph Embeddings, *Journal of Data Intelligence* **2**(3) (2021), 336–347. doi:10.26421/jdi2.3-2. https://doi.org/10.26421/JDI2.3-2.

[23] A. Sakor, K. Singh, A. Patel and M. Vidal, 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*, M. d'Aquin, S. Dietze, C. Hauff, E. Curry and P. Cudré-Mauroux, eds, ACM, 2020, pp. 3141–3148. doi:10.1145/3340531.3412777.

[24] A. Sakor, S. Jozashoori, E. Niazmand, A. Rivas, K. Bougiatiotis, F. Aisopos, E. Iglesias, P.D. Rohde, T. Padiya, A. Krithara et al., Knowledge4COVID-19: A Semantic-based Approach for Constructing a COVID-19 related Knowledge Graph from Various Sources and Analysing Treatments' Toxicities, *arXiv preprint arXiv:2206.07375* (2022). doi:https://doi.org/10.48550/arXiv.2206.07375.
[25] M. Vidal, K.M. Endris, S. Jazashoori, A. Sakor and A. Rivas, Transforming Heterogeneous Data into Knowledge for Personalized Treatments - A Use Case, *Datenbank-Spektrum* **19**(2) (2019), 95–106. doi:10.1007/s13222-019-00312-z.
[26] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration, *Nucleic Acids Research* **34** (2006), 668–672–. doi:10.1093/nar/gkj067.
[27] M. Ali, H. Jabeen, C.T. Hoyt and J. Lehmann, The KEEN Universe: An ecosystem for knowledge graph embeddings with a focus on reproducibility and transferability, 2020, (in press).
[28] J. Pujara, E. Augustine and L. Getoor, Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short, in: *EMNLP*, 2017.
[29] A. d'Avila Garcez and L.C. Lamb, Neurosymbolic AI: The 3rd Wave, 2020.
[30] T.R. Besold, A. d'Avila Garcez, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K.-U. Kühnberger, L.C. Lamb, P.M.V. Lima, L. de Penning, G. Pinkas, H. Poon and G. Zaverucha, Chapter 1. Neural-Symbolic Learning and Reasoning: A Survey and Interpretation1, *Neuro-Symbolic Artificial Intelligence: The State of the Art* (2021). doi:10.3233/faia210348.
[31] Z. Susskind, B. Arden, L.K. John, P. Stockton and E.B. John, Neuro-Symbolic AI: An Emerging Class of AI Workloads and their Characterization, *CoRR* **abs/2109.06133** (2021). https://arxiv.org/abs/2109.06133.
[32] C. Zhu, Z. Yang, X. Xia, N. Li, F. Zhong and L. Liu, Multimodal reasoning based on knowledge graph embedding for specific diseases, *Bioinformatics* **38**(8) (2022), 2235–2245. doi:10.1093/bioinformatics/btac085.
[33] M.R. Karim, M. Cochez, J.B. Jares, M. Uddin, O.D. Beyan and S. Decker, 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*, X.M. Shi, M. Buck, J. Ma and P. Veltri, eds, ACM, 2019, pp. 113–123. doi:10.1145/3307339.3342161.
[34] D. Sridhar, S. Fakhraei and L. Getoor, A probabilistic approach for collective similarity-based drug–drug interaction prediction, *Bioinformatics* **32**(20) (2016), 3175–3182. doi:10.1093/bioinformatics/btw342.
[35] P. Zhang, F. Wang, J. Hu and R. Sorrentino, Label Propagation Prediction of Drug-Drug Interactions Based on Clinical Side Effects, *Scientific Reports* **5** (2015). doi:10.1038/srep12339.
[36] S. Vilar, E. Uriarte, L. Santana, T. Lorberbaum, G. Hripcsak, C. Friedman and N.P. Tatonetti, Similarity-based modeling in large-scale prediction of drug-drug interactions, *Nature Protocols* **9** (2014). doi:10.1038/nprot.2014.151.
[37] A. Fokoue, M. Sadoghi, O. Hassanzadeh and P. Zhang, 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.
[38] N. Rohani and C. Eslahchi, Drug-Drug Interaction Predicting by Neural Network Using Integrated Similarity, *Scientific Reports* **9** (2019). doi:10.1038/s41598-019-50121-3.
[39] M. Zitnik, M. Agrawal and J. Leskovec, Modeling polypharmacy side effects with graph convolutional networks, *Bioinformatics* **34**(13) (2018), i457–i466. doi:10.1093/bioinformatics/bty294.
[40] G. Lee, C. Park and J. Ahn, Novel deep learning model for more accurate prediction of drug-drug interaction effects, *BMC Bioinformatics* **20** (2019). doi:10.1186/s12859-019-3013-0.
[41] R. Masumshah, R. Aghdam and C. Eslahchi, A neural network-based method for polypharmacy side effects prediction, *BMC Bioinformatics* **22**(385) (2021). doi:10.1186/s12859-021-04298-y.
[42] J.Y. Ryu, H.U. Kim and S.Y. Lee, Deep learning improves prediction of drug–drug and drug–food interactions, *Proceedings of the National Academy of Sciences* **115**(18) (2018), E4304–E4311. doi:10.1073/pnas.1803294115.
[43] Y. Yang, Y. Fang, M.E. Orlowska, W. Zhang and X. Lin, Efficient Bi-Triangle Counting for Large Bipartite Networks, *Proc. VLDB Endow.* **14**(6) (2021), 984–996–. doi:10.14778/3447689.3447702.
[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* **12** (2011), 2825–2830.
[45] Y. Wu, X. Yang, A. Plaza, F. Qiao, L. Gao, B. Zhang and Y. Cui, Approximate computing of remotely sensed data: SVM hyperspectral image classification as a case study, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **9**(12) (2016).
[46] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi and S. Homayouni, Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **13** (2020), 6308–6325. doi:10.1109/JSTARS.2020.3026724.
[47] L. Breiman, Random Forests, *Machine Learning* **45** (2001), 5–32. doi:10.1023/A:1010933404324.
[48] C. Cortes and V. Vapnik, Support-vector networks, *Machine learning* **20**(3) (1995), 273–297.
[49] W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian and X. Li, Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data, *BMC Bioinformatics* **18** (2017). doi:10.1186/s12859-016-1415-9.
[50] A. Kastrin, P. Ferk and B. Leskošek, Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning, *PLOS ONE* **13**(5) (2018), 1–23. doi:10.1371/journal.pone.0196865.
[51] H. Pirnejad, P. Amiri, Z. Niazkhani, A. Shiva and et al., Preventing potential drug-drug interactions through alerting decision support systems: A clinical context based methodology, *International Journal of Medical Informatics* **127** (2019), 18–26. doi:https://doi.org/10.1016/j.ijmedinf.2019.04.006. https://www.sciencedirect.com/science/article/pii/S1386505618303095.
[52] H. Hochheiser, X. Jing, E.A. Garcia, S. Ayvaz, R. Sahay, M. Dumontier and et al., A Minimal Information Model for Potential Drug-Drug Interactions, *Frontiers in Pharmacology* **11** (2021), 2477. doi:10.3389/fphar.2020.608068.