VISE: Validated and Invalidated Symbolic Explanations for Knowledge Graph Integrity

Disha Purohit[1,2,*,†], Yashrajsinh Chudasama[1,2,†], Maria Torrente[4] and Maria-Esther Vidal[1,2,3]

[1]TIB-Leibniz Information Centre for Science and Technology, Hannover, Germany
[2]Leibniz University Hannover, Germany
[3]L3S Research Center, Hannover, Germany
[4]Hospital Universitario Puertade Hierro-Majadahonda, Spain

Abstract

Knowledge graphs (KGs) are naturally capable of capturing the convergence of data and knowledge, thereby making them highly expressive frameworks for describing and integrating heterogeneous data in a coherent and interconnected manner. However, based on the Open World Assumption (OWA), the absence of information within KGs does not indicate falsity or non-existence; it merely reflects incompleteness. The process of inductive learning over KGs involves predicting new relationships based on existing factual statements in the KG, utilizing either numerical or symbolic learning models. Recently, Knowledge Graph Embedding (KGE) and symbolic learning have received considerable attention in various downstream tasks, including Link Prediction (LP). LP techniques employ latent vector representations of entities and their relationships in KGs to infer missing links. Furthermore, as the quantity of data generated by KGs continues to increase, the necessity for additional quality assessment and validation efforts becomes more apparent. Nevertheless, state-of-the-art KG completion approaches fail to consider the quality constraints while generating predictions, resulting in the completion of KGs with erroneous relationships. The generation of accurate data and insights is of vital importance in the context of healthcare decision-making, including the processes of diagnosis, the formulation of treatment strategies, and the implementation of preventive actions. We propose a hybrid approach, *VISE*, which adopts the integration of symbolic learning, constraint validation, and numerical learning techniques. *VISE* leverages KGE to capture implicit knowledge and represent negation in KGs, thereby enhancing the predictive performance of numerical models. Our experimental results demonstrate the effectiveness of this hybrid strategy, which combines the strengths of symbolic, numerical, and constraint validation paradigms. *VISE* implementation is publicly accessible on GitHub (https://github.com/SDM-TIB/VISE).

Keywords

Knowledge Graphs, Symbolic Learning, SHACL Constraints, Numerical Learning, Explainability

EXPLIMED - First Workshop on Explainable Artificial Intelligence for the medical domain - 19-20 October 2024, Santiago de Compostela, Spain

*Corresponding author.
†These authors contributed equally.
disha.purohit@tib.eu (D. Purohit); yashrajsinh.chudasama@tib.eu (Y. Chudasama); mtorrente80@gmail.com (M. Torrente); maria.vidal@tib.eu (M. Vidal)
0000-0002-1442-335X (D. Purohit); 0000-0003-3422-366X (Y. Chudasama); 0000-0003-1160-8727 (M. Vidal)

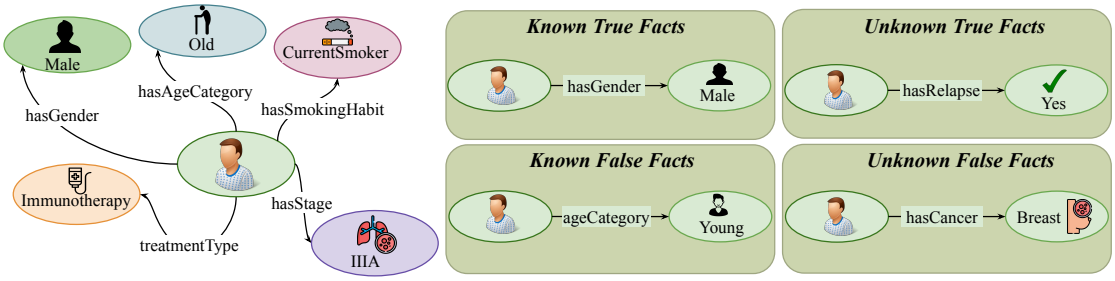


Figure 1: Prediction under Incompleteness The Lung Cancer KG use case employed in the current study demonstrates how a lung cancer patient is defined in the context of prediction under incompleteness. Furthermore, the quadrants on the right indicate the predictions or missing facts that can be classified as either known true facts, known false facts, unknown true facts, or unknown false facts.

1. Introduction

Knowledge Graphs (*KGs*) are rich structured data model that represents real-world information in the form of entities and relations that effectively merge data and knowledge through factual statements [1, 2, 3]. However, KGs are not complete based on the Open World Assumption (OWA) [4] principle. The process of inductive learning over KGs encompasses a variety of techniques for the acquisition of knowledge within KGs that will facilitate the completion of KGs. Inductive learning is crucial for detecting missing links in KGs; it includes deducing patterns and relationships from the existing KG. Established approaches can learn symbolic or numerical representations of KGs' patterns, which correspond to the fundamental building blocks for inferring missing links [1], thus, completing KGs effectively.

KGE methods project entities and relations from a KG into a lower-dimensional vector space while preserving their semantic significance. Existing KGE approaches [5, 6, 7, 8] have demonstrated promising results in various knowledge acquisition tasks, including link prediction, entity recognition, relation extraction, etc. Training KGE models typically involves ranking observed *(positive)* instances higher than unobserved *(negative)* instances. However, since KGs only provide positive instances, it becomes essential to generate negative instances [9] that can enable the model to learn intricate and valuable semantics. As illustrated in Figure 1, the potential for prediction under the incomplete nature of KGs is demonstrated by the example of a lung cancer patient and the relationships between various characteristics of this patient in the KG. The four quadrants are depicted on the right of Figure 1, which includes predictions or missing facts that can be classified as either known true facts, known false facts, unknown true facts, or unknown false facts. The category of *Known True Facts* represents the facts of the patient that are already present in the KG. For example, we know that the patient is male. The category of *Known False Facts* refers to facts that are known but are not true. The categories of *Unkown True Facts* and *Unkown False Facts* are the missing facts that are often predicted by symbolic learning or numerical learning. Conversely, traditional KGs do not explicitly represent negated facts or relationships. Instead, they concentrate on representing positive facts or relationships between entities in the KG. While this strategy simplifies the representation and querying processes, it also excludes the representation of negated facts in KGs, which impairs

the performance of downstream tasks, for example, link prediction (LP) for KG completion. Inductive learning techniques struggle to learn from only positive data in the KGs, resulting in poor predicting performance. For instance, knowing the positive facts ⟨*Patient X, hasAgeCategory, Young*⟩, *and* ⟨*Immunotherapy, hasDrug, Vinorelbine*⟩, a KGE model could predict ⟨*Patient X, hasRelapse, No Relapse*⟩. Nonetheless, the latent vector representation of the entities and their relationships are not self-explanatory. Extracting explanations efficiently for the inductive abilities remains an outstanding research challenge.

The problem of explaining the LP has received significant attention in critical domains like healthcare. Various approaches [7, 10, 11, 12] attempt to understand the inner mechanism of such inductive learning techniques, but they are unable to capture the insights of the model behavior with negated facts. We follow Rossi et al. [7] vocabulary and extract explanations for LP problems. The necessity and sufficiency of explanations can be characterized in several ways. For instance, the addition of a set of facts to a knowledge graph (KG) for an entity can lead to the model making a prediction, whereas the absence of a set of facts cannot.

In Figure 2, an exemplar sub-graph depicts the task of predicting a missing tail entity ⟨*Patient 1, patientDrug, Nivolumab*⟩. If the known facts about the head entity *Patient 1*, i.e., ⟨*Patient 1, hasStage, IIIA*⟩, *and* ⟨*Patient 1, hasSmokingHabit, CurrentSmoker*⟩ are removed from the training graph, the model's predicted tail changes. Hence, the model relies on these *necessary* facts to forecast *Nivolumab*, a plausible tail entity. In *sufficient* scenario, for instance, adding the fact ⟨*Patient 1, treatmentType, Immunotherapy*⟩ *and* ⟨*Patient 1, hasStage, IIIA*⟩ to the training graph, can lead the model to predict their drug as *Nivolumab*.

Several studies demonstrate that generating high-quality negatives is a difficult but critical step in improving KGE. As a result, negative sampling (NS) [9] has become an essential component of knowledge representation learning, considerably improving the performance of KGE models through effective negative selection. The current inductive learning approaches, such as symbolic learning [13, 14] and numerical learning [15, 16], fail to consider the validity and invalidity of constraints when anticipating missing links. This results in the addition of connections to KG graphs that do not meet domain requirements. Constraints can be validated using the Shapes Constraint Language (SHACL)- W3C standardized shape constraint language. SHACL constraints are symbolic constraints that provide explanations for the validity and integrity of data in a KG. SHACL constraints serve as a set of rules or guidelines, defining the permissible shapes that data instances can take within the graph. These constraints validate the content, and relationships of entities, ensuring compliance with predefined standards or expectations. In essence, SHACL constraints offer a symbolic framework for evaluating the correctness and coherence of data, contributing to the overall quality and reliability of the KG.

Our approach *VISE* tackles the challenge of KG completion by introducing a hybrid approach that utilizes symbolic learning, symbolic constraints validation, and numerical learning to avail the best of all the paradigms. *VISE* enhances the capabilities of KGE models by incorporating symbolic learning inferences and constraints validation, thereby further transforming the input KG by rewriting the relationships to specify negations in the KG. Thus, *VISE* helps numerical learning, i.e., KGE models excel in predictive performance empowering KG completion. Additionally, extracting two types of rationales *necessary* and *sufficient* facts for LP tasks.

The rest of the paper is organized as follows: Section 2 motivates the KG completion problem and defines the basic concepts of inductive learning. Section 3 presents the problem statement

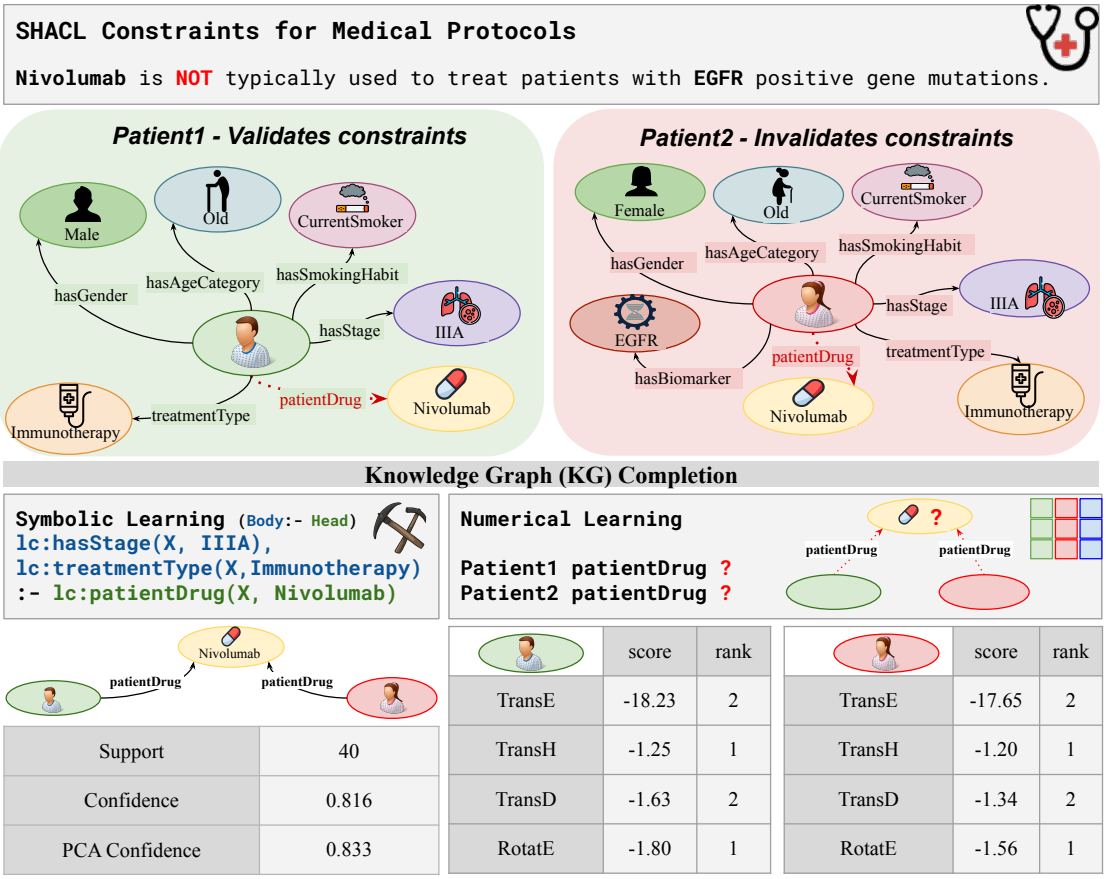


Figure 2: Motivating Example. depicts an exemplar sub-graph of the Lung Cancer (LC) KG. Domain experts have established clinical guidelines to determine the type of treatment or drug *(Nivolumab)* a patient can receive based on their genetic mutation *EGFR*. *Symbolic Learning* and *Numerical Learning* approaches are utilized to perform a prediction task to enrich incompleteness in KGs.

and defines our approach, *VISE*, using a hybrid design pattern. Section 4 evaluates the approach and benchmarks. Section 5 reports the results of the experimental study. Section 6 discusses the state of the art. Finally, section 8 presents the conclusions and future work.

2. Motivation and Background

This section uses an example to illustrate the problem of KG completion and the basic concepts necessary to understand the approach presented in this paper.

2.1. Motivating Example

The motivation for our work arises from the fact that the KG completion methods do not consider symbolic constraints to ensure KG integrity. The addition of missing relationships to KGs for the completion of incomplete KGs without ensuring that the links added to satisfy the

domain constraints may correspond to unknown false facts. The state-of-the-art KG completion approaches are deficient in considering the SHACL validation results in order to avoid completing KGs with spurious relations. Figure 2 illustrates the lung cancer use case presented in the current work. Domain experts (e.g., oncologists, medical doctors, or medical researchers) specify clinical guidelines or protocols, for example, it is recommended that the drug *Nivolumab* should be avoided for lung cancer patients mutated with *EGFR Positive* biomarker. These recommendations are defined in terms of SHACL constraints to determine whether or not patients are adhering to the clinical guidelines. The outcomes of performing SHACL constraints show if a lung cancer patient validates or invalidates the constraints, i.e., if a patient mutated with *EGFR Positive* is treated with *Nivolumab* drug. Therefore, SHACL validation reports verify the data utilized by the KG completion procedures to complete the incomplete KGs, ensuring integrity. Figure 2 shows the lung cancer KG that utilizes a set of variables to describe the main characteristics of a lung cancer patient. These include the patient identifier (also known as the electronic health record of a patient), gender, age, cancer stage (also known as the cancer stage), smoking habits (also known as the smoking habit), lung cancer biomarkers, drugs and treatments given to the patients. The OWA principle is used for representing KG in real-world scenarios. KG completion approaches such as symbolic learning and numerical learning are used to complete the missing relationships between entities of the KG. *Symbolic learning*, allows the capture of explicit patterns from the KGs and the generation of Horn rules to derive insights from the KGs. For example, as shown in Figure 2, a Horn Rule: `lc:hasStage(X, IIIA), lc:treatmentType(X, Immunotherapy):- lc:patient(X, Nivolumab)` states that if a patient has stage IIIA and receives immunotherapy, then it is most likely that the patient is being treated with the drug Nivolumab. *Numerical Learning*, i.e., KGE models predicted missing links by describing entities and their relations in a low-dimensional vector space. For example, by taking into account the patient's neighborhood, predicting the drug that a patient can receive. Figure 2 showing the *Patient 1* in green, validating the constraints since the patient is not EGFR-mutated and hence can take Nivolumab following the clinical guidelines. *Patient 2*, shown in red invalidates the constraints, i.e., does not adhere to the clinical guidelines. KG completion approaches, such as symbolic and numerical learning, still predict that the patient should be given *Nivolumab*, as they fall short in assessing whether the predicted missing links validate or invalidate the clinical guidelines given by the domain experts. As shown, numerical learning approaches such as *TransH* and *RotatE* predict patients (e.g., *Patient 2*) not adhering to clinical guidelines with higher rank and score.

2.2. Preliminaries

This section introduces basic preliminaries to understand our approach, i.e., shape, constraints, shape evaluation, SHACL, knowledge graph embedding, Horn rule, heuristic-based negative edges, support, confidence, PCA confidence, Hits@K, MRR, necessary and sufficient explanation. More details about these preliminary concepts in [1, 13, 17, 18].

Knowledge Graphs. A *knowledge graph (KG)* is a directed *edge-labeled graph* $KG = (V, E, L)$, where *Con* is a set of countable infinite constants. $V \subseteq Con$ is a set of nodes, $L \subseteq Con$ is a set of edge labels, and $E \subseteq V \times L \times V$ is a set of edges.

Constraints. A *constraint* corresponds to a rule that imposes restrictions on the values taken

for target nodes in V with a given edge E.

Shapes. A *shape* corresponds to a conjunction of constraints that a set of nodes in a knowledge graph must satisfy. A *shape* ϕ is inductively defined as follows:

ϕ::=T represents the value True;

| Δ_N nodes belongs to the set of nodes N;

| ψ_{cond} a node satisfies the *Boolean* condition *cond*;

| $\phi_1 \wedge \phi_2$ is conjunction of shape ϕ_1 and shape ϕ_2;

| $\neg\phi$ represents the negation of shape ϕ;

| $\rightarrow^p \phi\{min, max\}$ is cardinality on outward edges with label p to nodes satisfying ϕ; min and max are natural numbers.

Shape Schema. A shapes schema is defined as a tuple $\sum = (\varphi, S, \lambda)$, where:

- φ is a set of shapes;
- S is a set of shape labels;
- λ: S $\rightarrow \varphi$ is a total function from labels to shapes.

Shape Target. Given a shapes schema $\Sigma = (\varphi, S, \lambda)$ and a directed edge-labelled graph KG $= (V, E, L)$, $\theta(\phi, V)$ corresponds to the subset of nodes in V which are targets of $\phi \in \varphi$.

Shape Schema Evaluation. Given a shapes schema $\Sigma = (\varphi, S, \lambda)$ and a directed edge-labelled graph KG $= (V, E, L)$, a node $v \in V$. Given a shape $\phi \in \varphi$, the shape evaluation function $[\phi]^{KG,v} \in \{0, 1\}$ states the results of evaluating ϕ in a node v from V in KG.

- $[T]^{\text{KG},v} = 1$
- $[\Delta_N]^{\text{KG},v} = 1$ iff $v \in N$
- $[\psi_{cond}]^{\text{KG},v} = 1$
- $[\phi_1 \wedge \phi_2]^{KG,v} = \min\{[\phi_1]^{KG,v}, [\phi_2]^{KG,v}\}$
- $[\neg\phi]^{KG,v} = 1 - [\phi]^{KG,v}$
- $[\rightarrow^p \phi\{min, max\}]^{KG,v} = 1$ iff $\min \leq |\{(v, p, u) \in E \mid [\phi]^{KG,v} = 1\}| \leq \max$

Shape Schema Validation. Given a shapes schema $\Sigma = (\varphi, S, \lambda)$ and a directed edge-labelled graph KG $= (V, E, L)$, A node $v \in V$ validates φ, i.e., $v \models \varphi$, iff $[\phi]^{KG,v} = 1$ for all $\phi \in \varphi$ and v in $\theta(\phi, V)$. KG satisfies the shape schema $\Sigma = (\varphi, S, \lambda)$, iff for all v in V, $v \models \varphi$.

Example 2.1. *A shape schema $\sum$ of lung cancer patients is given as follows:*

- $\sum = (\varphi, S, \lambda)$,
- φ = $\{\rightarrow^{lc:hasGender}\ string\{1,1\},\ \wedge,\ \{\rightarrow^{lc:hasAge}\ string\{1,1\}\ \rightarrow^{lc:treatmentType}\ string\{1,^*\}\}$,
- $S = \{exS : Patient, exS : Treatment\}$,
- $\lambda(exS : Patient)$ = $\{\rightarrow^{hasGender}\ string\{1,1\}\} \wedge \{\rightarrow^{lc:hasAge}\ string\{1,1\}$,
- $\lambda(exS : Treatment)$ = $\rightarrow^{lc:treatmentType}\ string\{1,^*\}$.

The evaluation of the shape schema $\sum = (\varphi, S, \lambda)$ of lung cancer patients represented in KG, validates the nodes of patients with only one gender and age. Additionally, each lung cancer patient should receive at least one treatment.

Shapes Constraint Language (SHACL). SHACL [19] is the World Wide Web Consortium (W3C) recommendation language for the declarative specification of integrity constraints over RDF KGs. A SHACL shape represents a set of constraints that apply over the same entities; it can refer to another shape, to represent constraints between entities of two types.

Knowledge Graph Embedding (KGE). Given a directed *edge-labeled graph*, *KG = (V, E, L)* and set of vectors Γ. A *KGE* of *KG* is a pair of mappings (ϵ, σ) such that

- ϵ: $V \rightarrow \Gamma$, i.e., $\epsilon(e)$ maps a entity e in V to a vector in Γ, and
- σ: L $\rightarrow \Gamma$, i.e., $\sigma(l)$ maps a directed edge l to a vector in Γ.

A score function ϕ: $V \times L \times V \rightarrow \mathbb{R}$ is used to measure the plausibility of candidate triples represented in low-dimensional vector space, triples t = $\langle s, p, o \rangle$ with the higher score $\theta(\epsilon(s), \sigma(p), \epsilon(o))$ values conveys better plausibility. The objective of KGE is to learn the embeddings in (ϵ,σ) that maximize the plausibility of positive edges in E^+ and minimize the plausibility of negative edges in E^-. The set of positive edges, E^+, corresponds to the edges in T. The set of negative edges, E^-, corresponds to the edges in $V \times \mathrm{L} \times V \notin E^+$.

Hits@K. Given a tail prediction $p(s, ?)$ over the directed *edge-labeled graph* $KG = (V, E, L)$, the model predicts a list of K entities that might be related to the tail entity. $Hits@K$ determines the fraction of plausible entities that appear in the top K predictions.

$$\boxed{Hits@K = \frac{|\text{NumberOfPlausibleEntities} \leq K|}{K}} \tag{1}$$

Mean Reciprocal Rank (MRR). Given the prediction problem either from a head or tail perspective over the directed *edge-labeled graph* $KG = (V, E, L)$, the model ranks the plausible entities and calculates the reciprocal rank of the plausible entity for each predictive task.

$$\boxed{MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank_i}} \tag{2}$$

Horn Rule. A Horn rule is a logical implication defined as follows: *Body* $\Rightarrow$ *Head*. The body of the rule is comprised of predicate facts. The head is a predicate fact of a single atom. All the variables in the *Head* are terms of at least one predicate fact in the *Body*. Every two predicate facts in *Body* share at least one variable. We say a rule $R : B_1 \wedge B_2 \wedge \cdots \wedge B_n \implies R(x, y)$ where *Head* represents $R(x, y)$ and *Body* is $B_1 \wedge B_2 \wedge B_3 \wedge ... \wedge B_n$.

Entailment of a Mined Rule [20]. Given a directed *edge-labeled graph KG= (V,E,L)* and a mined rule $R : Body \Rightarrow Head$, the entailed facts of R corresponds to the instantiations of the predicate fact in *Head* on substituting the variables in *Body*, i.e., positive instantiations of the conjunction of predicates in *Body*. That implies $\forall\, V2$ such that, *Body[Z:=V2]* is a positive predicate fact, and *Head[Z:=V2]* corresponds to an entailed fact of R. We can defined a predicate fact as positive entailed fact $E^+(R)$, if $Head[Z := V2] = p(s, o') \in unknownE^+$.

Support of a Horn Rule. Given a directed *edge-labeled graph KG= (V,E,L)* and a mined rule $R : Body \Rightarrow Head$, the support of R indicates the number of positive entailed facts of *Head*.

Confidence of a Horn Rule. Given a directed *edge-labeled graph KG= (V,E,L)* and a mined rule $R : Body \Rightarrow Head$, the confidence of R is defined as a proportion of the positive predicate facts of *Head* that are positive entailed facts based on R.

Heuristic-based Negative Edges (hE$^-$). Given a directed *edge-labeled graph KG= (V,E,L)* and a mined rule $R : Body \Rightarrow Head$, where *Head* is $p(s, o)$. A heuristic-based negative edges hE^- corresponds to the set of instantiations $p(s, o')$ that do not belong to E, but

- exists $p(s, o) \in E^+(R)$,
- $p(s, o')$ is entailed by *Body*.

$$hE^-(R) = \{p(s,o') | p(s,o') \notin E^+(R) \ \wedge \ p(s,o) \in E^+(R) \ \wedge \ \text{p(s,o') is entailed by Body}\}$$

The set of hE^- comprises triples to be predicted following this heuristic.

PCA Confidence score of a Horn Rule. Given a directed *edge-labeled graph KG= (V,E,L)* and a mined rule $R : Body \Rightarrow Head$, where *Head* is $p(s,o)$. The Partial Completeness Assumption (PCA) score of R corresponds to the ratio of $support(R)$ to the cardinality of the union of E^+ and hE^-. PCA confidence score quantifies the number of triples of the form $p(s,o')$ from E^- that can be deduced following the heuristics edges. PCA (R) score $\in [0,1]$, where the score indicates the amount of triples can be inferred.

$$\boxed{PCA(R) = \frac{support(r)}{|E^+ \cup hE^-|}} \tag{3}$$

A Necessary and Sufficient Explanation. Given a directed *edge-labeled graph KG= (V,E,L)* and a mined rule $R : Body \Rightarrow Head$, where *Head* is $p(s,o)$. Given a score function θ: $V \times L \times V \rightarrow \mathbb{R}$ is used to measure the plausibility of predicted triples in *KG= (V,E,L)*.

- A necessary explanation corresponds to a set of predicate facts $p(s,o) \in E^+$, if removed from KG^{train} leads to a decrease in score function θ.
- A sufficient explanation corresponds to a set of predicate facts $p(s,o) \notin E^+$, if added to KG^{train}, leads to an increase in score function θ.

3. Our Approach

This section states the problem addressed in this paper and introduces the VISE framework, which integrates symbolic learning, constraint validation, and numerical learning to create more explainable, and reliable systems. The objective is to create a framework that is designed to consider the semantics of symbolic systems.

3.1. Problem Statement

Consider *edge-labeled graph KG = (V, E, L)*, such that each node $e \in V$ represents an entity, and each $p \in L$ represents a unique relation between the entities. Let $\sum = (\varphi, S, \lambda)$ be a shape schema over KG, and $\theta(s,p,o')$ be a scoring function quantifying the plausibility of a triple (s,p,o'). The problem of link prediction over KG, i.e., a tail prediction $\langle s,p,? \rangle$, such that s is the subject entity in V and predicate p in L corresponds to the optimization problem of identifying an entity o' that produces the most plausible candidates for the incomplete triple (s,p,o') and s and o' validate $\sum = (\varphi, S, \lambda)$.

$$\boxed{o' = \underset{e \in V}{\arg\min}\, \theta(s,\ p,\ e) \wedge s \models \varphi \wedge e \models \varphi} \tag{4}$$

The aim is to find the most plausible entities o by inferring heuristic-based negative edges hE^- based on the positive edges E^+ in KG and validates the shape schema $\sum$, i.e., $(s,p,o') \models \varphi$.

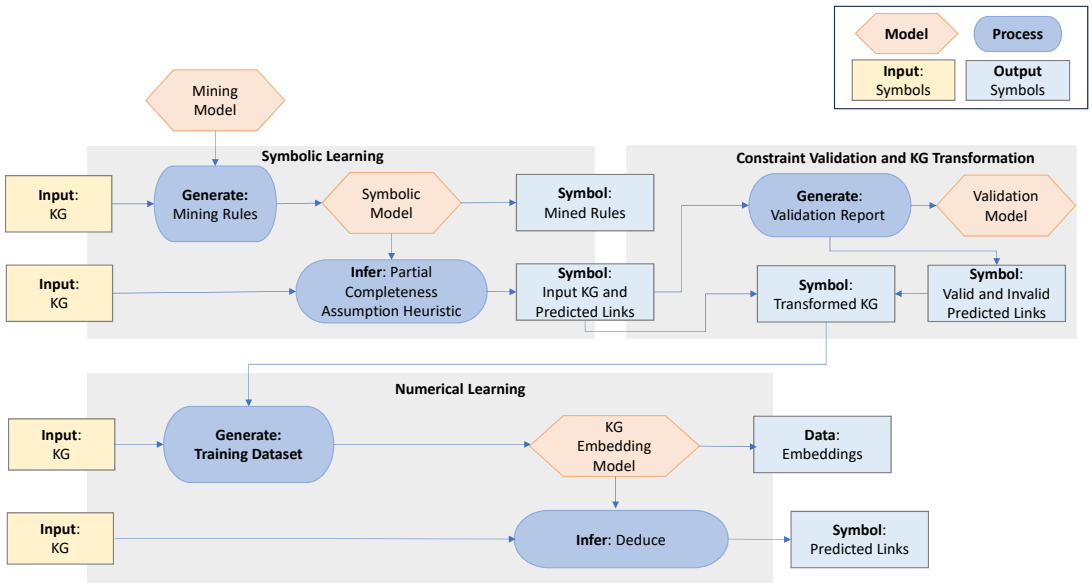


Figure 3: VISE Design Pattern: Hybrid design pattern to demonstrate the use of symbolic rules from the *Symbolic Learning* component and *Constraint Validation and KG Transformation* component in combination with *Numerical Learning* to enhance the predictive performance of KGE models.

3.2. The VISE Framework

VISE encompasses a hybrid approach that showcases the impact of considering the semantics of the symbolic system over the numerical learning approaches. *VISE* follows the hybrid design pattern as illustrated in Figure 3, strategically combining numerical learning with symbolic learning and constraints validation methods.

Symbolic learning is applied to the input KG, resulting in the generation of logical rules and PCA heuristic-based edges. The learned heuristic-based edges serve as prior knowledge, improving numerical learning approaches such as KGE models combined with constraints validation and KG transformation. During the process of symbolic learning, VISE utilizes extracted horn rules in conjunction with PCA Confidence in order to infer heuristic-based negative edges. The mined rules are subsequently employed to generate predictions regarding the missing relationships in the input KG. These predictions are based on logical inference, which is used to calculate the entailment of the mined rules. SPARQL queries are employed to infer the entailment of mined rules and construct heuristic-based negative edges (hE^-).
The predictions generated by the symbolic learning system in conjunction with the input KG are then fed to the *Constraints Validation and KG Transformation* component, where the predicted links are evaluated to determine whether they validate or invalidate the SHACL constraints. Furthermore, the generated validation report is utilized to transform or rewrite the

SHACL Constraints for Medical Protocols

Nivolumab is **NOT** typically used to treat patients with **EGFR** positive gene mutations.

Symbolic Learning (Body:- Head)

```
lc:hasStage(X, IIIA), lc:treatmentType(X,Immunotherapy) :-
lc:patientDrug(X, Nivolumab)
```

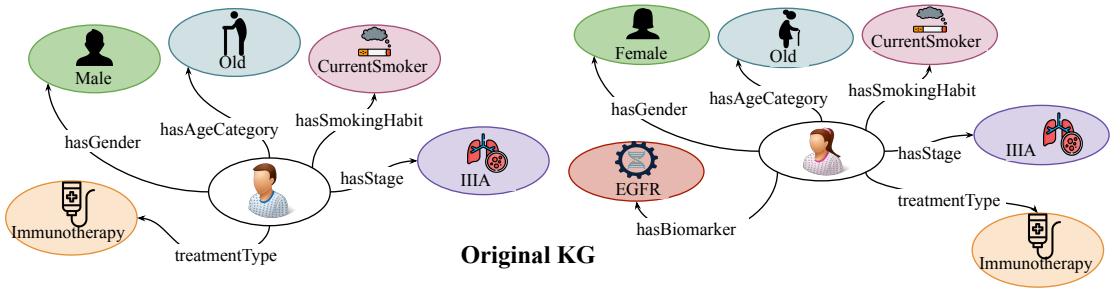
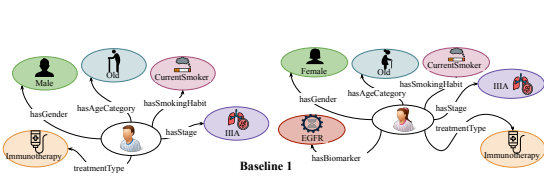
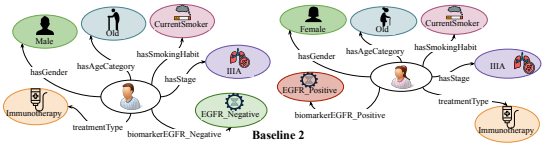


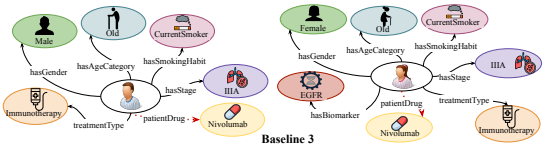
Figure 4: The figure illustrates the SHACL constraint, horn rule, and a subgraph of the original KG.



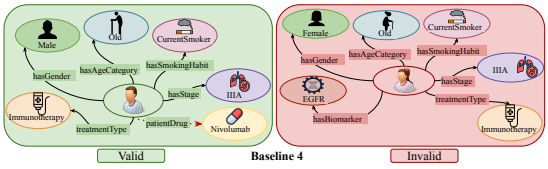
(a) Subgraph of original lung cancer KG



(b) Transformation of KG predicates without symbolic learning enrichment



(c) KG enrichment with predictions generated using symbolic learning rules



(d) KG enrichment with predictions of symbolic rules and SHACL validation

Figure 5: Different baseline approaches to show the transformation process of KGs

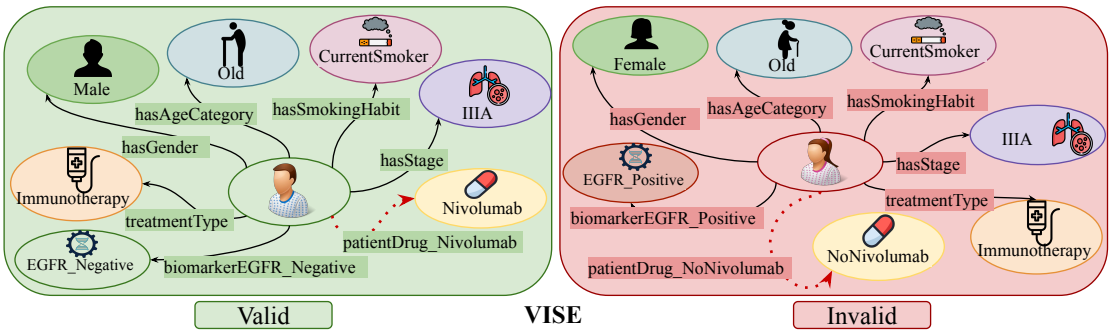


Figure 6: The *VISE* translation the KG predicates for explicitly stating negation in the KG.

KG to contain the information resulting from the constraints validation. The transformed KG is then provided as input to the numerical learning models, i.e., KGE models, during the training phase. This is achieved by processing the data into a low-dimensional space. The process of numerical learning is capable of predicting missing links, thereby completing the KGs with missing links that validate the constraints at a higher rank and with a greater probability of accuracy. *Tranformation or rewriting of KGs* before giving as input to the numerical learning component transforms the KGs to contain negated facts, allowing the KGE model to learn in all the four quadrants as shown in Figure 1, enhancing the performance of the models and empowering KG completion. Several studies demonstrated the need for negated facts in KGs to boost the performance of KGE models. *VISE* employs a two-fold rewriting process. First, it evaluates the links predicted by symbolic learning using constraints. Second, depending upon the validation report of the predicted link. If the patient in the lung cancer KG invalidates the constraint, the links that resemble the patient characteristics in the KG are added with negation.

3.3. Running Example

As discussed in subsection 2.1, a SHACL constraint stipulates that a patient who has undergone a mutation involving the EGFR gene should not be treated with the drug Nivolumab. Symbolic learning, on the other hand, has identified a rule that states that if a patient is in the advanced stage of lung cancer (stage IV) and has undergone immunotherapy treatment, there is a higher probability that the patient will receive the drug *Nivolumab*. PCA heuristics enabled symbolic learning to predict that the patient should receive *Nivolumab*.
The SHACL validation revealed that this patient violated the constraint if the predicted link was added to the KG. In the transformation process, *VISE* is capable of adding a fact to the KG indicating that the patient should not receive *Nivolumab*. As a result, the transformed KG explicitly represents positive and negative facts. This new way of modeling statements further enhances the performance of the models by assigning high values of plausibility score o' to triples that are likely to be true and low scores to triples that are likely to be false.
To illustrate, a running example demonstrates the process of transforming the KG to explicitly add negated facts to the KG, thereby further enhancing numerical learning and performance. Figure 4 shows the SHACL constraints based on the clinical guidelines, horn rules mined over the original KG, and a sub-figure that resembles the original lung cancer KG. Figure 5a shows the *Baseline 1* approach involves inputting the original KG into numerical learning, which is then processed by KGE models without the addition of inferred facts from symbolic learning for KG enrichment. Figure 5b illustrates the transformation of KG predicates implemented over the original KG (*Baseline 2*), thereby demonstrating the necessity to rewrite the negated facts to the KG for enhancement of numerical learning. For instance, the predicate may be altered to either `EGFR_Positive` if a patient is positively mutated for EGFR or `EGFR_Negative` if a patient is negatively mutated for EGFR mutation. The results of the transformation of KG are presented in Section 5 and demonstrate the impact on the performance of KGE models.
Figure 5c shows the *Baseline 3* approach that involves enriching the KG with symbolic learning. The KG is enriched with symbolic learning rules that incorporate PCA heuristics to generate heuristics-based negative (hE^-) edges. As discussed before, the rule can predict that if a stage IV lung cancer patient receives immunotherapy treatment then that patient is more likely to

Table 1
KGs Statistics. Table depicts the statistics of benchmarks, #triples – Number of RDF triples in the KG, #entities – Number of distinct entities in the KG, #predicates – Number of distinct predicates in the KG.

KGs	# triples	# entities	# relations
KG_1	1871	93	43
KG_2	13097	267	43
KG_3	20581	383	43

receive drug *Nivolumab* which is added as a fact as shown in the Figure 5c with `patientDrug Nivolumab` in the KG given as input to KGE models. The *Baseline 3* approach is in alignment with the state-of-the-art methodology of *SPaRKLE* [20]. Furthermore, Figure 5d displays *Baseline 4*, which highlights the influence of constraint validation. To show the need for constraint validation, inferred facts resulting from symbolic learning were added to the KG, which is only fed into KGE models once a patient validates the constraint. As shown in Figure 5d, the patient in red invalidates the constraint. As a result, the inferred fact `patientDrug Nivolumab` is not included for that patient. Figure 6 illustrates the transformation of KG implemented in *VISE*. In *VISE*, we aim to emphasize the significance of the hybrid approach in considering the impact of symbolic systems. The transformation of KG is achieved by explicitly incorporating the inferred facts predicted by symbolic rules, thereby validating the constraints for the predicted facts. To illustrate, in the figure referenced in the text, the patient in green validates the constraint. Consequently, the rewriting in the transformed KG includes `biomarkerEGFR_Negative`, `EGFR_Negative`, `patientDrug_Nivolumab`, and `Nivolumab`. For the patient in red, invalidates the constraint. Consequently, the following transformation is performed as follows `EGFR_Positive`, `EGFR_Positive`, `patientDrug_NoNivolumab`, and `NoNivolumab`. The results of the transformation of *VISE* KG are presented in 5, which demonstrates the impact on the performance of KGE models. Figure 5a, Figure 5b, Figure 5c, and Figure 5d present the benchmarks (resp., Baseline 1, Baseline 2, Baseline 3, and Baseline 4) utilized in the experiments.

4. Evaluation

The aforementioned section outlines our proposed framework *VISE* and its components. In this section, we report the experiment settings, benchmark description, observed results, involved baselines, and models. We empirically assess the effectiveness of *VISE* in the LP problem over the Lung Cancer KG. *VISE* provides comprehensive explanations: *necessary* and *sufficient* for LP task. For instance, an LP task can be *"Whether a lung cancer patient is in Relapse?"*. Thus, given a head entity and relation predicting the tail entity, i.e., ⟨Patient X, hasRelapse, ?⟩. The empirical evaluation aims to answer the following research questions: **RQ1)** What is the impact of negated facts on the KGE model's performance and its explainability? **RQ2)** How do symbolic rules and constraints enhance the explanations of the KGE model's behavior?
Benchmark. We evaluate *VISE* approach on three anonymized Lung Cancer KGs: KG_1, KG_2, and KG_3. Table 1 shows the statistics of all the benchmarks. The Lung Cancer KG comprises medical records about a lung cancer patient from heterogeneous data sources. Each medical record describes the characteristics of a patient suffering from lung cancer. The medical charac-

teristics include a cancer stage (*e.g., Stage IVB*), age, gender, smoking habit (*e.g., Current Smoker*), type of mutation (*e.g., EGFR Negative*), recommended drug for treatment (*e.g.,* Vinorelbine), the occurrence of relapse (*e.g., Relapse or Progression or No Relapse*), and types of treatment (*e.g., Immnunotherapy*) for curing the cancer. The prediction problem is a link prediction to predict the *Relapse* of a lung cancer patient, which can be *Relaspe* or *No Relaspe*. We utilize SHACL constraints as medical protocols that recommend when a drug should be prescribed according to a patient's mutations; we defined one shape schema with four different SHACL constraints, for instance, a constraint stating that *"If a patient mutated with EGFR negative should not take Afatinib, and if a patient mutated with EGFR positive should not take Nivolumab"*.

Baselines. We evaluate and compare four baselines for our *VISE* approach. Baseline 1 includes the evaluation of the state-of-the-art KGE models for the KG completion. Baseline 2 reveals the evaluation of transformed KG with KGE models. Baseline 3 utilizes the hybrid approach, SPaRKLE [20], which employs symbolic learning techniques to enhance the performance of KGE models. Baseline 4 combines SPaRKLE with the results of patients who satisfy the medical protocols. *VISE* approach integrates the fusion of SPaRKLE with the transformed KG including validation and violation results to enhance the performance of KGE models in LP tasks. The current implementation utilizes various state-of-the-art KGE models from the PyKEEN [21] pipeline, which includes TransE [16], TransD [22], TransH [23], and RotatE [24]. We conducted an ablation study to tune hyperparameters for KGE models based on benchmark KGs. Translation-distance space models, including TransE, TransD, and TransD, translate the head entity's geometric embedding space with a given relation closer to the tail entity. RotatE, a popular model for learning embeddings in Euclidean space, has attracted attention for learning symmetric, asymmetric, 1-1, 1-N, N-1, and M-to-N relationships. Table 2, Table 3 and Table 4 demonstrates the comparison between baselines and *VISE* approach for KG completion.

Implementation. *VISE* is implemented in a virtual machine on Google Colab with 40 GiB VRAM and 1 GPU NVIDIA A100-SMX4, with CUDA version 12.2 (Driver 535.104.05) using Python 3.9. The source code of *VISE* approach, the benchmark KGs, and the trained KGE models are publicly available in our GitHub repository [1]. Figure 3 depicts the hybrid design pattern, integrating inductive learning with symbolic learning techniques. Symbolic learning includes logical horn rules (R) and SHACL constraints (ϕ). Symbolic learning is performed over the input KG, resulting in rules, heuristic-based edges, and SHACL validation. Thus, the inferred heuristic edges with validation results are utilized as implicit knowledge to enhance inductive learning, i.e., KGE models. The predictions generated from the symbolic rules and constraints materialized in the input KG and fed as input to inductive learning. The benchmark KGs are divided into 80-20 train-test splits. The model's efficacy in the LP problem is evaluated using *Hits@K* and *MRR*. Both metrics have values between 0 and 1, and higher conveys better. VISE relies on [20] and [25] for symbolic learning methods. Furthermore, our approach is model-agnostic and compatible with other symbolic and inductive learning approaches.

[1] https://github.com/SDM-TIB/VISE

Table 2
KG_1 **Evaluation.** Empirical evaluation of various KGE models on KG_1. Hits@1, Hits@3, Hits@5, Hits@10 and MRR are reported. Four baselines and VISE (in light green color) indicates the impact of considering the captured knowledge in the prediction tasks. The values in bold convey better results.

Approaches	Results for KG_1					
	Model	Hits@1	Hits@3	Hits@5	Hits@10	MRR
Baseline 1	**TransE**	0.000	0.444	0.622	0.822	0.269
	TransD	0.000	0.755	**0.866**	0.911	0.361
	TransH	**0.422**	**0.777**	0.844	**0.955**	**0.625**
	RotatE	**0.422**	0.511	0.555	0.555	0.485
Baseline 2	**TransE**	0.000	0.644	0.777	0.844	0.330
	TransD	0.000	0.844	0.866	0.889	0.406
	TransH	**0.711**	**0.866**	**0.867**	**0.911**	**0.785**
	RotatE	0.489	0.533	0.622	0.667	0.542
Baseline 3	**TransE**	0.000	0.458	0.604	0.812	0.266
	TransD	0.000	0.583	0.729	**0.895**	0.323
	TransH	**0.479**	**0.687**	**0.770**	0.875	**0.618**
	RotatE	0.354	0.416	0.541	0.625	0.430
Baseline 4	**TransE**	0.000	0.478	0.586	0.739	0.276
	TransD	0.000	0.630	0.826	0.956	0.340
	TransH	**0.413**	**0.739**	**0.913**	**0.978**	**0.606**
	RotatE	**0.413**	0.521	0.543	0.586	0.479
VISE	**TransE**	0.000	0.667	0.770	0.916	0.352
	TransD	0.000	0.854	0.937	**1.000**	0.421
	TransH	**0.791**	**0.958**	**1.000**	**1.000**	**0.877**
	RotatE	0.479	0.500	0.604	0.729	0.538

5. Results

In this empirical study, we evaluate the efficacy of numerical inductive and symbolic learning approaches in terms of the evaluation metrics proposed by Akrami et al. [26]. These empirical studies aim to address the research questions **RQ1** in Section 5.1 and **RQ2** in Section 5.2.

5.1. Impact of Negated Facts on KGE Model Behavior

We report the effectiveness of *VISE* approach, focusing on KGE models- TransE, TransD, TransH, and RotatE in the context of lung cancer relapse prediction problems. The comprehensive analysis revealed a robust performance compared to baselines. KGE models are trained over the different benchmark KGs, i.e., positive edges E^+, to predict missing links. The evaluation report presented in Table 2, 3, and 4 are obtained using the optimized hyperparameters provided by the PyKEEN pipeline. The impact of negated facts is assessed with Hits@1, Hits@3, Hits@5, Hits@10, and MRR in KG completion. TransE, a basic translation model, emerged as performing worst in all baselines with benchmarks respectively. Nevertheless, highlighting the limitations of TransE in modeling 1-N relationships leads to poor performance, particularly in predicting the correct tail at the topmost position. TransH model results support the claim in [23], that it outperforms TransE and TransD models. In KG_1 and KG_2, TransH performance contributes to

promising results in capturing complex geometric relationships with score values ranging from 0.413 to 0.865. TransD, which uses relation-specific projections to translate the embedding space, yields slightly lower values than TransH and TransE. However, RotatE indicates the best performance in all the testbeds except in KG_1. In KG_2 and KG_3, the values of Hits@1 range from 0.489 to 0.887. We can observe that the evaluation of benchmark KGs in different experimental testbeds, *VISE* outperforms compared to the other baseline approaches. The experimental evaluation comprises 100 testbeds per KGs, amounting to a total of 300 testbeds. In summary, the evaluation results underline the robust performance of TransH and RotatE for KG completion in lung cancer relapse prediction tasks.

However, the rationale behind the inner workings of KGE models may be difficult to understand. The experimental results demonstrate the need for explanations and assistance to understand KGE model behavior. *VISE* shows improved KGE model performance and provides two types of post hoc explanation for the prediction problem. In *VISE* approach, KGE models showed marginally better performance compared to Baseline 1. We categorize our explanations as *necessary* and *sufficient*. The heuristic-based negative edges (hE^-) generated by symbolic learning demonstrate the importance of enhancing the performance of *VISE*. The addition of hE^- edges to KG has been deemed a sufficient explanation, as evidenced by the improved performance of the KGE model in terms of Hits@K and MRR. For example, Table 5 displays examples of mined Horn rules that were chosen based on the SHACL constraints, i.e., clinical guidelines used to infer the hE^- edges. Moreover, the removal of these edges from the KG resulted in a notable decline in performance, which can be attributed to the necessity of these facts, i.e., necessary facts to explain the prediction performance thereby answering **RQ1**.

5.2. Effectiveness of Symbolic Rules and Constraints on LP task

The Horn rules mined by AMIE [13] over LC KG are used to help doctors screen for and identify persons who are at high risk of acquiring lung cancer. Mined rules are examined in terms of biomarkers, medications, and therapies, and ranked according to the *PCA confidence* score. The effectiveness of *VISE* is evaluated in terms of the impact of validating constraints for the missing link being predicted by the symbolic learning technique. As described in Section 3 the heuristic-based negative edges (hE$^-$) are predicted using the Partial Completness Assumption (PCA) heuristics from the input KG. The PCA Confidence of a Horn rule, which indicates the amount of incompleteness in a knowledge graph (KG), is employed to infer new links and predictions. These predictions are validated by applying the SHACL constraints to determine the validity of the inferred links. The results demonstrated in Table 5 indicate the amount of valid and invalid predictions produced by the symbolic learning techniques. Table 5 shows examples of the symbolic rules, for example, `stage(?a, IV)`, `treatment(?a, Immunotherapy)` $\Rightarrow$ `drug(?a, Nivolumab)` stating that if a stage IV lung cancer patient received *Immunotherapy* treatment then it is more likely that the patient receives Nivolumab is with the *PCA Confidence* score of 0.833. As mentioned before, the heuristics-based negative edges (hE$^-$) or predictions are validated using SHACL constraints, and Table 5 shows the number of valid ($\#v$) and invalid ($\#in$) links for each of the LC KGs used as a benchmark in *VISE*.

Furthermore, the symbolic rules are used to represent the studies reported in the literature. Table 6 provides examples of mined Horn rules and supporting literature. Consequently, the

Table 3
KG_2 **Evaluation.** Empirical evaluation of various KGE models on KG_2. Hits@1, Hits@3, Hits@5, Hits@10 and MRR are reported. Four baselines and VISE (in light green color) indicates the impact of considering the captured knowledge in the prediction tasks. The values in bold convey better results.

Approaches	Results for KG_2					
	Model	Hits@1	Hits@3	Hits@5	Hits@10	MRR
Baseline 1	**TransE**	0.000	0.514	0.711	0.870	0.296
	TransD	0.010	0.615	0.796	0.952	0.349
	TransH	**0.540**	**0.834**	**0.917**	**0.974**	**0.702**
	RotatE	0.536	0.761	0.819	0.869	0.663
Baseline 2	**TransE**	0.000	0.714	0.860	0.939	0.378
	TransD	0.015	0.758	0.898	0.965	0.403
	TransH	**0.850**	**0.965**	**0.987**	**1.000**	**0.911**
	RotatE	0.831	0.971	0.990	0.990	0.903
Baseline 3	**TransE**	0.000	0.550	0.722	0.898	0.313
	TransD	0.012	0.582	0.767	0.934	0.342
	TransH	0.546	0.823	0.901	0.946	0.695
	RotatE	**0.617**	**0.913**	**0.958**	**0.982**	**0.765**
Baseline 4	**TransE**	0.000	0.587	0.778	0.904	0.323
	TransD	0.006	0.606	0.784	0.907	0.338
	TransH	0.584	0.864	0.913	0.947	0.728
	RotatE	**0.701**	**0.920**	**0.950**	**0.978**	**0.814**
VISE	**TransE**	0.000	0.734	0.880	0.955	0.384
	TransD	0.012	0.767	0.880	0.958	0.405
	TransH	**0.865**	0.970	0.982	**1.000**	**0.918**
	RotatE	0.856	**0.973**	**0.988**	0.997	0.916

impact of symbolic rules and constraints utilized to explain the KGE models is demonstrated, thereby enabling an answer to be provided to the research question **RQ2**.

6. Related Work

The integration of symbolic and numerical learning into KGs enhances their utility and interpretability. Symbolic techniques employ rules and logic to identify missing relationships, whereas numerical approaches utilize low-dimensional vector spaces to discern connections between entities in large KGs. Symbolic constraint validation, in isolation, identifies inaccuracies in KGs that can be employed to assess data quality. It is of vital importance to explain the predictions in the healthcare domain, as this helps domain experts in decision-making, identifying the interactions between drugs and their side effects, and patient diagnosis.
The majority of KGE methods employ triples from KGs as input, with the embeddings being trained using vector space assumptions (e.g., translational, neural network, complex space) [15]. Furthermore, the embeddings are obtained to perform the link prediction task [21], as outlined in Rivas et al. [6]. Rivas et al. propose a neuro-symbolic perception for drug treatment response to enhance the link prediction capabilities of KGE models by deducing implicit knowledge using datalog rules. Akrami et al. [26] present a study that employs a realistic and updated assessment

Table 4
KG_3 **Evaluation.** Empirical evaluation of various KGE models on KG_3. Hits@1, Hits@3, Hits@5, Hits@10 and MRR are reported. Four baselines and VISE (in light green color) indicates the impact of considering the captured knowledge in the prediction tasks. The values in bold convey better results.

Approaches	Results for KG_3					
	Model	Hits@1	Hits@3	Hits@5	Hits@10	MRR
Baseline 1	**TransE**	0.000	0.560	0.795	0.943	0.324
	TransD	0.002	0.551	0.690	0.872	0.310
	TransH	0.622	0.864	0.943	0.983	0.756
	RotatE	**0.696**	**0.933**	**0.969**	**0.987**	**0.820**
Baseline 2	**TransE**	0.000	0.713	0.840	0.931	0.376
	TransD	0.008	0.694	0.824	0.935	0.379
	TransH	**0.882**	**0.969**	**0.997**	**1.000**	**0.929**
	RotatE	0.864	0.987	0.995	1.000	0.924
Baseline 3	**TransE**	0.000	0.519	0.747	0.923	0.310
	TransD	0.011	0.551	0.716	0.884	0.322
	TransH	0.596	0.876	0.925	0.977	0.740
	RotatE	**0.714**	**0.941**	**0.969**	**0.990**	**0.829**
Baseline 4	**TransE**	0.000	0.536	0.735	0.931	0.311
	TransD	0.002	0.551	0.733	0.870	0.318
	TransH	0.542	0.849	0.908	0.974	0.702
	RotatE	**0.700**	**0.945**	**0.972**	**0.992**	**0.818**
VISE	**TransE**	0.000	0.760	0.878	0.948	0.388
	TransD	0.013	0.684	0.762	0.884	0.368
	TransH	0.868	0.980	0.994	**1.000**	0.924
	RotatE	**0.887**	**0.986**	**0.996**	0.998	**0.936**

of various KG completion techniques. The objective is to establish their usefulness in improving KG completeness and quality. The findings of the study, as presented in work [26], indicate that the embedding models may have been biased toward learning reverse relations for LP due to the presence of data redundancy and Cartesian product relations.

Furthermore, it was demonstrated that simple models, such as symbolic learning approaches, outperform numerical models when data contains reverse relations or data redundancy. Consequently, we aim to showcase the combination of symbolic and numerical methodologies that frequently result in enhanced performance and outcomes in a variety of activities, including KG completion in our proposed approach *VISE*. Moreover, the validation of constraints over the predicted links from symbolic learning approaches can assist in identifying whether the predicted links validate or invalidate the constraints. This process can enhance the system's performance by providing information about the constraint validation for LP tasks. While both symbolic and numerical techniques have advantages and disadvantages, a hybrid approach that combines them can mitigate shortcomings while leveraging the complementary benefits of these KG completion methods to further empower KGs.

In a related study, Lajus et al. [13] present a symbolic learning technique that captures the co-occurrence of relationships, rules, and logical dependencies within KGs. Among numerous KG completion approaches, this method employs the OWA to extract association rules from

Table 5
Exemplary Mined Horn Rules.

Exemplary Mined Horn Rules	PCA Conf.	KG_1 hE^-		KG_2 hE^-		KG_3 hE^-	
		#v	#in	#v	#in	#v	#in
drug(?a, Nivolumab) ⇐ stage(?a, IV), treatment(?a, Immunotherapy)	0.833	5	4	30	24	50	40
drug(?a, Nivolumab) ⇐ treatment(?a, Immunotherapy), treatment(?a, Intravenous_Chemotherapy)	0.722	13	10	78	60	130	100
biomarker(?a, EGFR_Negative) ⇐ relapseProgression(?a, Progression), drug(?a, Pembrolizumab)	0.971	0	1	0	6	0	10
biomarker(?a, EGFR_Negative) ⇐ biomarker(?a, ALK_Negative), treatment(?a, Radiotherapy_To_Bone)	0.921	2	2	12	12	20	20

Table 6
Mined Horn Rules. Exemplary mined and statements Reported in the Literature and their Relationship with the Analysis Outcomes.

Exemplary Mined Horn Rules	Statements
drug(?a, Nivolumab) ⇐ stage(?a, IV), treatment(?a, Immunotherapy)	Lung cancer patients in stage IV[27] and receive Immunotherapy treatment are more likely to revive *Nivolumab*[28].
drug(?a, Nivolumab) ⇐ treatment(?a, Immunotherapy), treatment(?a, Intravenous_Chemotherapy)	Non-small cell lung cancer patients receive *Nivolumab*[28, 29] as first-line Chemotherapy and Immunotherapy treatments for progression-free survival.
biomarker(?a, EGFR_Negative) ⇐ relapseProgression(?a, Progression), drug(?a, Pembrolizumab)	EGFR[30] negative lung cancer patients are more likely to experience progression and are treated with pembrolizumab[31] drug.
biomarker(?a, EGFR_Negative) ⇐ biomarker(?a, ALK_Negative), treatment(?a, Radiotherapy_To_Bone)	Lung cancer patients mutated with ALK negative and receives treatment radiotherapy[32, 33] to bone are more likely to be also mutated with EGFR[30] negative.

KGs. AMIE [13] enhances the quality and completeness of KGs by deducing missing linkages and connections, while also considering semantics. AnyBURL [14] (Anytime Bottom-Up Rule Learning) is a state-of-the-art system that entails initiating specific instances within the KGs and subsequently generalizing them to generate more encompassing logical rules that can be applied across the KGs. Khajeh Nassiri et al. [34] propose a symbolic learning technique that emphasizes the use of logical rules, including numerical predicates. This method enables KGs to recognize and mine correlations between numerical values, measures, and other quantitative properties, resulting in a more expansive and precise representation of real-world knowledge. Chudasama et al. [35] demonstrated in one of the related studies that SHACL technologies may

be utilized to evaluate data over KGs for quality assessment, as well as in predictive modeling analysis to improve model interpretability. SHACL technologies can be used to validate data and then used with ensemble approaches, such as Random Forest and Decision Trees, to interpret the behavior of Machine Learning (ML) models, which can help understand the outcomes generated by prediction models. Rabbani et al. [36] employs a technique to extract validating shapes from large KGs. Furthermore, an efficient SHACL validation engine [25] shows the best performance in planning and executing SHACL shape schema to determine whether entities from KGs comply with specific medical protocols. *VISE* is system-agnostic, allowing straightforward integration with any existing KGE model or symbolic system. To achieve integration, the mining horn rules for symbolic systems, as well as the computation of PCA and prediction scores, and a set of SHACL constraints can be utilized for validation.

7. Discussion

VISE framework demonstrates the effectiveness of considering PCA heuristics for LP tasks and generates explanations. However, the proposed approach has limitations in terms of incorporating the semantics of KGs. By employing symbolic reasoning, implicit facts can be deduced, which can be utilized to enhance the neighborhood of an entity. Consequently, considering the semantics of KGs would provide a comprehensive picture of the scalability of each KGE model in real-world use cases. Furthermore, investigating the computational overhead observed for each model to capture complex relationships can also be conducted in future studies. The experimental results demonstrate that KGE models do not fully account for the contextual knowledge of entities, such as entity validation in the context of medical protocols. Nevertheless, our approach, *VISE*, is domain-agnostic and can be used to enhance and explain the behavior of KGE models in LP tasks. The mining of rules, validation of entities, and training of the KGE models scale with the size of KGs. Thus, exploiting the minimal neighborhood with specific rules for negative sampling will aid in solving the scalability issues. Lastly, state-of-the-art KGE models are commonly employed for a range of downstream tasks. However, their latent vector representations lack self-interpretability. Consequently, future studies may benefit from leveraging the enriched contextual information considering the semantics of KGs to enhance the explanations generated by Large Language Models (LLMs). This could prove valuable in critical domains such as healthcare, facilitating more efficient decision-making processes.

8. Conclusions and Future Works

VISE avails the advantages of the PCA heuristic, which improves predictions regarding missing links. Constraint validation helps include additional symbolic system semantics in numerical learning approaches. Empirical evidence indicates that the integration of symbolic learning approaches with constraint validation can enhance the performance of KGE models, particularly when prior knowledge is taken into account. Consequently, *VISE* exemplifies the advantages of integrating symbolic and numerical methodologies into a hybrid or neuro-symbolic AI system. This integration allows academics and practitioners to combine these two conceptual AI approaches, thereby achieving accurate solutions for KG completion.

Moreover, *VISE* demonstrated the necessity of rewriting the KGs to include negative edges using SHACL constraints' results rather than randomly generating negative samples for the numerical learning approaches. It is important to note that hybrid approaches do come with limitations. This work provides evidence that hybrid methods necessitate the integration of various components, which can result in increased computational complexity. The processing of symbolic systems may result in the mining of rules and the inference of triples that are unnecessary for numerical models, which may not utilize them. This opens the door for future research to efficiently execute hybrid systems and to fully leverage their benefits.

Acknowledgments

This work has been partially supported by TrustKG- Transforming Data in Trustable Insights with grant P99/2020 and the EraMed project P4-LUCAT (GA No. 53000015).

References

[1] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge Graphs, Synthesis Lectures on Data, Semantics, and Knowledge, Morgan & Claypool Publishers, 2021.

[2] C. Gutierrez, J. F. Sequeda, Knowledge graphs, Commun. ACM 64 (2021) 96–104. doi:10.1145/3418294.

[3] F. Aisopos, S. Jozashoori, E. Niazmand, D. Purohit, A. Rivas, A. Sakor, E. Iglesias, D. Vogiatzis, E. Menasalvas, A. R. González, G. Vigueras, D. Gómez-Bravo, M. Torrente, R. H. López, M. P. Pulla, A. Dalianis, A. Triantafillou, G. Paliouras, M. Vidal, Knowledge graphs for enhancing transparency in health data ecosystems, Semantic Web 14 (2023) 943–976. URL: https://doi.org/10.3233/SW-223294. doi:10.3233/SW-223294.

[4] Y. Loyer, U. Straccia, Any-world assumptions in logic programming, Theoretical Computer Science 342 (2005) 351–381. URL: https://www.sciencedirect.com/science/article/pii/S030439750500304X. doi:https://doi.org/10.1016/j.tcs.2005.04.005.

[5] F. Akrami, L. Guo, W. Hu, C. Li, Re-evaluating embedding-based knowledge graph completion methods, in: CIKM, 2018. doi:10.1145/3269206.3269266.

[6] A. Rivas, D. Collarana, M. Torrente, M.-E. Vidal, A neuro-symbolic system over knowledge graphs for link prediction, Semantic Web Journal. Special Issue on Neuro-Symbolic Artificial Intelligence and the Semantic Web (2023) 1–25. doi:10.3233/SW-233324.

[7] A. Rossi, D. Firmani, P. Merialdo, T. Teofili, Explaining link prediction systems based on knowledge graph embeddings, in: SIGMOD, 2022.

[8] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge graph embedding: A survey of approaches and applications, IEEE Transactions on Knowledge and Data Engineering 29 (2017) 2724–2743. URL: http://dx.doi.org/10.1109/TKDE.2017.2754499. doi:10.1109/tkde.2017.2754499.

[9] T. Madushanka, R. Ichise, Negative sampling in knowledge graph representation learning:

A review, CoRR abs/2402.19195 (2024). URL: https://doi.org/10.48550/arXiv.2402.19195. doi:10.48550/ARXIV.2402.19195. arXiv:2402.19195.

[10] Y. Chudasama, Exploiting semantics for explaining link prediction over knowledge graphs, in: C. Pesquita, H. Skaf-Molli, V. Efthymiou, S. Kirrane, A. Ngonga, D. Collarana, R. Cerqueira, M. Alam, C. Trojahn, S. Hertling (Eds.), The Semantic Web: ESWC 2023 Satellite Events - Hersonissos, Crete, Greece, May 28 - June 1, 2023, Proceedings, volume 13998 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 321–330. URL: https://doi.org/10.1007/978-3-031-43458-7_50. doi:10.1007/978-3-031-43458-7_50.

[11] H. Zhang, T. Zheng, J. Gao, C. Miao, L. Su, Y. Li, K. Ren, Data poisoning attack against knowledge graph embedding, in: IJCAI, 2019.

[12] D. Purohit, M. Vidal, Mining symbolic rules to explain lung cancer treatments, in: C. Pesquita, H. Skaf-Molli, V. Efthymiou, S. Kirrane, A. Ngonga, D. Collarana, R. Cerqueira, M. Alam, C. Trojahn, S. Hertling (Eds.), The Semantic Web: ESWC 2023 Satellite Events - Hersonissos, Crete, Greece, May 28 - June 1, 2023, Proceedings, volume 13998 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 69–74. URL: https://doi.org/10.1007/978-3-031-43458-7_13. doi:10.1007/978-3-031-43458-7_13.

[13] J. Lajus, L. Galárraga, F. Suchanek, Fast and Exact Rule Mining with AMIE 3, in: The Semantic Web, 2020.

[14] C. Meilicke, M. W. Chekol, D. Ruffinelli, H. Stuckenschmidt, Anytime bottom-up rule learning for knowledge graph completion, in: IJCAI-19, 2019. doi:10.24963/ijcai.2019/435.

[15] M. Ali, M. Berrendorf, C. T. Hoyt, L. Vermue, M. Galkin, S. Sharifzadeh, A. Fischer, V. Tresp, J. Lehmann, Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework, CoRR (2020).

[16] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, NIPS'13, Curran Associates Inc., Red Hook, NY, USA, 2013, p. 2787–2795.

[17] J. E. Labra-Gayo, H. García-González, D. Fernández-Alvarez, E. Prud'hommeaux, Challenges in RDF Validation, Springer International Publishing, 2019, p. 121–151. URL: http://dx.doi.org/10.1007/978-3-030-06149-4_6. doi:10.1007/978-3-030-06149-4_6.

[18] A. Rossi, D. Firmani, P. Merialdo, T. Teofili, Explaining link prediction systems based on knowledge graph embeddings, in: Proceedings of the 2022 International Conference on Management of Data, SIGMOD/PODS '22, ACM, 2022. URL: http://dx.doi.org/10.1145/3514221.3517887. doi:10.1145/3514221.3517887.

[19] H. Knublauch, D. Kontokostas, Shapes Constraint Language (SHACL), W3C Recommendation, 2017. URL: https://www.w3.org/TR/2017/REC-shacl-20170720/.

[20] D. Purohit, Y. Chudasama, A. Rivas, M.-E. Vidal, Sparkle: Symbolic capturing of knowledge for knowledge graph enrichment with learning, in: Proceedings of the 12th Knowledge Capture Conference 2023, K-CAP '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 44–52. URL: https://doi.org/10.1145/3587259.3627547. doi:10.1145/3587259.3627547.

[21] M. Ali, M. Berrendorf, C. T. Hoyt, L. Vermue, S. Sharifzadeh, V. Tresp, J. Lehmann, PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings, Journal of Machine Learning Research (2021). URL: http://jmlr.org/papers/v22/20-825.html.

[22] G. Ji, S. He, L. Xu, K. Liu, 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), 2015, pp. 687–696. URL: https://aclanthology.org/P15-1067.pdf. doi:10.3115/v1/P15-1067.

[23] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes, Proceedings of the AAAI Conference on Artificial Intelligence 28 (2014). URL: https://ojs.aaai.org/index.php/AAAI/article/view/8870. doi:10.1609/aaai.v28i1.8870.

[24] Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang, Rotate: Knowledge graph embedding by relational rotation in complex space, 2019. URL: https://arxiv.org/abs/1902.10197. doi:10.48550/ARXIV.1902.10197.

[25] M. Figuera, P. D. Rohde, M.-E. Vidal, Trav-SHACL: Efficiently Validating Networks of SHACL Constraints, in: The Web Conference, ACM, New York, NY, USA, 2021.

[26] F. Akrami, M. S. Saeef, Q. Zhang, W. Hu, C. Li, Realistic re-evaluation of knowledge graph completion methods: An experimental study, in: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, SIGMOD '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1995–2010. doi:10.1145/3318464.3380599.

[27] A. Karaboué, T. Collon, V. Bodiguel, J. Cucherousset, P. Innominato, M. Bouchahda, R. Adam, F. Levi, Nivolumab timing as a major survival predictor in patients with stage iv non–small cell lung cancer., Journal of Clinical Oncology 40 (2022) 9058–9058. doi:10.1200/JCO.2022.40.16_suppl.9058.

[28] D. P. Carbone, M. Reck, L. Paz-Ares, B. Creelan, L. Horn, M. Steins, E. Felip, M. M. van den Heuvel, T.-E. Ciuleanu, F. Badin, N. Ready, T. J. N. Hiltermann, S. Nair, R. Juergens, S. Peters, E. Minenza, J. M. Wrangle, D. Rodriguez-Abreu, H. Borghaei, G. R. Blumenschein, L. C. Villaruz, L. Havel, J. Krejci, J. C. Jaime, H. Chang, W. J. Geese, P. Bhagavatheeswaran, A. C. Chen, M. A. Socinski, First-line nivolumab in stage iv or recurrent non–small-cell lung cancer, New England Journal of Medicine 376 (2017) 2415–2426. doi:10.1056/NEJMoa1613493.

[29] M. D. Hellmann, L. Paz-Ares, R. B. Caro, B. Zurawski, S.-W. Kim, E. C. Costa, K. Park, A. Alexandru, L. Lupinacci, E. de la Mora Jimenez, H. Sakai, I. Albert, A. Vergnenegre, S. Peters, K. Syrigos, F. Barlesi, M. Reck, H. Borghaei, J. R. Brahmer, K. J. O'Byrne, W. J. Geese, P. Bhagavatheeswaran, S. K. Rabindran, R. S. Kasinathan, F. E. Nathan, S. S. Ramalingam, Nivolumab plus ipilimumab in advanced non–small-cell lung cancer, New England Journal of Medicine 381 (2019) 2020–2031. doi:10.1056/NEJMoa1910231.

[30] M. A. Velez, T. F. Burns, Is the game over for pd-1 inhibitors in egfr mutant non-small cell lung cancer?, Translational Lung Cancer Research 8 (2019). URL: https://tlcr.amegroups.org/article/view/28523.

[31] M. J. Hadfield, A. Turshudzhyan, K. Shalaby, A. Reddy, Response with pembrolizumab in a patient with egfr mutated non-small cell lung cancer harbouring insertion mutations in v834l and l858r, Journal of Oncology Pharmacy Practice 28 (2022) 717–721. doi:10.1177/10781552211057867, pMID: 34783273.

[32] V. Nardone, C. Romeo, E. D'Ippolito, P. Pastina, M. D'Apolito, L. Pirtoli, M. Caraglia, L. Mutti, G. Bianco, A. Falzea, R. Giannicola, A. Giordano, P. Tagliaferri, C. Vinciguerra,

I. Desideri, M. Loi, A. Reginelli, P. Tassone, P. Correale, The role of brain radiotherapy for egfr- and alk-positive non-small-cell lung cancer with brain metastases: a review, La Radiologia medica 128 (2023). doi:10.1007/s11547-023-01602-z.

[33] A. Wrona, R. Dziadziuszko, J. Jassem, Combining radiotherapy with targeted therapies in non-small cell lung cancer: focus on anti-egfr, anti-alk and anti-angiogenic agents, Translational Lung Cancer Research 10 (2021). URL: https://tlcr.amegroups.org/article/view/49653.

[34] A. K. Nassiri, N. Pernelle, F. Saïs, REGNUM: generating logical rules with numerical predicates in knowledge graphs, in: The Semantic Web - 20th International Conference, ESWC 2023, Hersonissos, Crete, Greece, May 28 - June 1, 2023, Proceedings, volume 13870, Springer, 2023, pp. 139–155. doi:10.1007/978-3-031-33455-9_9.

[35] Y. Chudasama, D. Purohit, P. D. Rohde, M.-E. Vidal, Enhancing interpretability of machine learning models over knowledge graphs, in: N. Keshan, S. Neumaier, A. L. Gentile, S. Vahdati (Eds.), Proceedings of the Posters and Demo Track of the 19th International Conference on Semantic Systems co-located with 19th International Conference on Semantic Systems (SEMANTiCS 2023), Leipzig, Germany, September 20 to 22, 2023, volume 3526 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3526/paper-05.pdf.

[36] K. Rabbani, M. Lissandrini, K. Hose, Extraction of validating shapes from very large knowledge graphs, Proc. VLDB Endow. 16 (2023) 1023–1032. URL: https://www.vldb.org/pvldb/vol16/p1023-rabbani.pdf. doi:10.14778/3579075.3579078.