Approximating Probabilistic Inference in Statistical $\mathcal{EL}$ with Knowledge Graph Embeddings

Yuqicheng Zhu[1,2], Nico Potyka[3], Bo Xiong[4], Trung-Kien Tran[1], Mojtaba Nayyeri[2], Evgeny Kharlamov[1,5], Steffen Staab[2,6]

[1]Bosch Center for AI, [2]University of Stuttgart, [3]Cardiff University, [4]Stanford University,
[5]University of Oslo, [6]University of Southampton

Abstract

In domains where statistical data is collected across hierarchically organized categories, drawing valid conclusions requires reasoning jointly about proportions and the structure of the domain. Statistical $\mathcal{EL}$ ($\mathcal{SEL}$) formalizes this kind of reasoning, but exact inference is EXPTIME-hard and no implementation exists. We show how knowledge graph embeddings can approximate $\mathcal{SEL}$ inference efficiently. We prove analytical runtime and soundness guarantees, and empirically evaluate the runtime and approximation quality of our approach.

1 INTRODUCTION

Statistical information pervades decision-making in medicine, public policy, and science, yet drawing valid conclusions from it is surprisingly error-prone [Thiese et al., 2015, Brown et al., 2018]. A striking example is Simpson's paradox [Simpson, 1951]: aggregate statistics can suggest one conclusion while every relevant subgroup supports the opposite. In the 1973 UC Berkeley admissions data [Bickel et al., 1975], women appeared to be admitted at a lower overall rate than men, yet within most individual departments the pattern disappeared or even reversed. The paradox arises because the aggregate rate conflates department-level selectivity with the distribution of male and female applicants across departments. Resolving it requires reasoning jointly about statistical proportions and the structure of the domain: what subgroups exist (departments, gender), how they relate to broader categories (all applicants), and how individuals are distributed across them.

Statistical $\mathcal{EL}$ ($\mathcal{SEL}$) [Peñaloza and Potyka, 2017] provides a formal framework for exactly this combination of hierarchical structure and statistical proportions. It represents statistical statements as probabilistic conditionals $(D \mid C)[l, u]$, expressing that the proportion of C-instances that also belong to D lies in $[l, u]$. In the Berkeley setting, one can encode the department-specific admission rate of female applicants, e.g. $(\textit{Admitted} \mid \textit{DeptA\_Applicant} \sqcap \textit{Woman})[l_1, u_1]$, alongside the share of female applicants applying to each department, e.g. $(\textit{DeptA\_Applicant} \mid \textit{Woman})[l_2, u_2]$. Given such constraints, $\mathcal{SEL}$ can automatically derive tight bounds on the overall rate $(\textit{Admitted} \mid \textit{Woman})$, which can reveal when aggregate and subgroup-level comparisons are consistent or potentially divergent. We demonstrate this concretely in Appendix B. Further use cases arise in domains such as medicine and manufacturing, where statistics about different kinds of patients or products are reconciled across multiple levels of a hierarchy [Zheng et al., 2017, 2022].

Despite its practical relevance, exact reasoning in $\mathcal{SEL}$ is EXPTIME-hard [Bednarczyk, 2021]. The only theoretically exact approach constructs an exponentially large linear program via a type elimination procedure [Lutz and Schröder, 2010], but to the best of our knowledge it has never been implemented. $\mathcal{SEL}$ thus provides the right formalism for statistical reasoning over structured domains but has so far remained without tool support.

In this paper, we bridge this gap using knowledge graph embeddings. We embed $\mathcal{SEL}$ concepts as axis-aligned boxes in a vector space, with volumes encoding statistical proportions, and estimate unknown conditional probabilities via volume ratios, replacing intractable logical inference with efficient geometric computation. We prove that inference runs in linear time and is sound when the embedding loss is zero. In practice, the loss will typically be non-zero, but we hypothesize that the inference error is proportional to the embedding error. To evaluate this conjecture empirically, we construct three $\mathcal{SEL}$ ontologies from YAGO3 [Mahdisoltani et al., 2014] and evaluate the runtime, approximation error of inferred probabilities, and how it relates to the embedding error. Our main contributions are:

- We explain how knowledge graph embeddings can be used to approximate probabilistic inference efficiently.
- We apply the idea to the $\mathcal{EL}$ embedding BoxEL [Xiong

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:8312–8329.

et al., 2022], generalize embedding soundness results to $\mathcal{SEL}$ and prove novel runtime and soundness guarantees for approximate inference with $\mathcal{SEL}$ embeddings.

- We empirically evaluate the runtime, accuracy and approximation quality of our approach.

2 RELATED WORK

Probabilistic first-order logics can be classified as type 1 (statistical), type 2 (subjective) or type 3 (combined) logics [Halpern, 1990]. Type 1 probabilities represent proportions in the domain, while type 2 probabilities represent a degree of belief that a statement is true. To the best of our knowledge, most probabilistic DLs are type 2 logics, for example, Lukasiewicz and Straccia [2008], Gutiérrez-Basulto et al. [2017], Peñaloza and Potyka [2016]. The only other type 1 probabilistic DL that we are aware of has been sketched in the appendix of Lutz and Schröder [2010]. Consistency-checking is ExpTime-hard for this logic as well, and we are unaware of any implementations.

Theoretically, probabilistic reasoning in $\mathcal{SEL}$ can be performed exactly by constructing a linear optimization problem, where the numerical variables and constraints are derived from a type elimination procedure and the conditionals in the knowledge base Lutz and Schröder [2010]. However, this approach is difficult to implement and not practical because the number of types rapidly explodes. A more pragmatic approach for classical inference would be creating a proof system for $\mathcal{SEL}$ similar to those for propositional probabilistic logics [Frisch and Haddawy, 1994, Hunter and Potyka, 2023]. However, even in propositional logic, completeness of such proof systems could only be shown for very limited fragments [Frisch and Haddawy, 1994].

Knowledge graph (KG) embeddings [Bordes et al., 2013] map entities and relations into a vector space. Prominent examples include *translational* [Bordes et al., 2013] and *bilinear* embeddings [Yang et al., 2015, Liu et al., 2017]. Many KG embeddings encode only instance-level knowledge but cannot represent logical axioms. Kulmanov et al. [2019] proposed embedding $\mathcal{EL}$ concepts as n-balls. BoxEL [Xiong et al., 2022] and ELBE [Peng et al., 2022] use axis-parallel boxes instead, which is beneficial as they are closed under intersection (conjunction). Box2EL [Jackermeier et al., 2024] further embeds both concepts and roles as boxes. Other types of DLs like $\mathcal{ALC}$ were also considered [Özçep et al., 2020, Garg et al., 2019].

3 BACKGROUND ON STATISTICAL $\mathcal{EL}$

The DL $\mathcal{EL}$ [Baader, 2003] describes concepts and their relationships using a set N_C of *concept names* and a set N_R of *role/relation names*. Every concept name $A \in N_C$ and the symbol $\top$ (called *top*) are *(atomic) $\mathcal{EL}$ concepts*. If C_1, C_2 are $\mathcal{EL}$ concepts and $r \in N_R$, then so are $C_1 \sqcap C_2$ and $\exists r.C$. An $\mathcal{EL}$ interpretation $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$ consists of a non-empty set $\Delta^\mathcal{I}$ called the domain of $\mathcal{I}$ and a mapping $\cdot^\mathcal{I}$ that maps every concept name to a subset of $\Delta^\mathcal{I}$, and every relation to a relation over $\Delta^\mathcal{I} \times \Delta^\mathcal{I}$. $\cdot^\mathcal{I}$ is extended to arbitrary concepts as follows. We let $\top^\mathcal{I} = \Delta^\mathcal{I}$, $(C_1 \sqcap C_2)^\mathcal{I} = C_1^\mathcal{I} \cap C_2^\mathcal{I}$ and $(\exists r.C)^\mathcal{I} = \{a \in \Delta^\mathcal{I} \mid \exists b \in C^\mathcal{I} : (a,b) \in r^\mathcal{I}\}$. An *$\mathcal{EL}$ TBox* contains *general concept inclusions (GCIs)* of the form $C \sqsubseteq D$, where C, D are concepts. An interpretation $\mathcal{I}$ satisfies $C \sqsubseteq D$ iff $C^\mathcal{I} \subseteq D^\mathcal{I}$. We write $C \equiv D$ as a shorthand for the two GCIs $C \sqsubseteq D$ and $D \sqsubseteq C$.

Statistical $\mathcal{EL}$, $\mathcal{SEL}$ for short, is a probabilistic extension of $\mathcal{EL}$ that allows reasoning about statistical statements [Peñaloza and Potyka, 2017]. The basic syntactic elements are *(probabilistic) conditionals* $(D \mid C)[l,u]$, where C, D are $\mathcal{EL}$ concept descriptions and $l, u \in [0,1] \cap \mathbb{Q}$ are (rational) probabilities such that $l \leq u$. If $l = u$, we just write $(D \mid C)[p]$. The intuitive reading of $(D \mid C)[l,u]$ is that the conditional probability of D given C is between l and u. The probabilities are interpreted statistically. That is, $(D \mid C)[l,u]$ means that the proportion of elements in C that also belong to D is between l and u. The formal semantics of $\mathcal{SEL}$ is defined with respect to $\mathcal{EL}$ interpretations $\mathcal{I}$ with finite domain $\Delta^\mathcal{I}$. $\mathcal{I}$ satisfies $(D \mid C)[l,u]$, denoted as $\mathcal{I} \models (D \mid C)[l,u]$, iff either $C^\mathcal{I} = \emptyset$ or $\frac{|(D \sqcap C)^\mathcal{I}|}{|C^\mathcal{I}|} \in [l,u]$. This is equivalent to saying that $\mathcal{I}$ satisfies $(D \mid C)[l,u]$ iff

$$l \cdot |C^\mathcal{I}| \leq |(D \sqcap C)^\mathcal{I}| \leq u \cdot |C^\mathcal{I}|. \tag{1}$$

$\mathcal{SEL}$ generalizes $\mathcal{EL}$ in the following sense [Peñaloza and Potyka, 2017, Proposition 4].

Lemma 1. *For all $\mathcal{EL}$ interpretations $\mathcal{I}$, we have $\mathcal{I} \models C \sqsubseteq D$ iff $\mathcal{I} \models (D \mid C)[1]$.*

The consistency problem for $\mathcal{SEL}$ turned out to be EXP-complete [Baader and Ecke, Bednarczyk, 2021]. Here, we are interested in the following *inference problem* for $\mathcal{SEL}$: Given an $\mathcal{SEL}$ TBox $\mathcal{T}$ and a query $(D \mid C)$, where C, D are $\mathcal{SEL}$ concepts, find the largest lower bound l and the smallest upper bound u such that every interpretation $\mathcal{I}$ that satisfies $\mathcal{T}$ also satisfies $(D \mid C)[l,u]$.

4 ILLUSTRATIVE EXAMPLE

To foster intuition, we illustrate how embeddings can approximate statistical reasoning in $\mathcal{SEL}$ using a simplified university admissions scenario inspired by the Berkeley example from the introduction.

Example 1. *Figure 1 shows, at the top, some $\mathcal{SEL}$ axioms. Every admitted applicant and every Department A applicant (DeptA) is an applicant; 20–25% of all applicants applied to Department A; and 80% of Department A applicants were admitted. Below, we show two possible 2D embeddings, where sets of applicants are represented by boxes.*

The volumes of the boxes are shown in Table 1. For example, on the left, $\frac{Vol(DeptA \sqcap Applicant)}{Vol(Applicant)} = \frac{Vol(DeptA)}{Vol(Applicant)} = \frac{8}{40} = 0.2$, $\frac{Vol(Admitted \sqcap DeptA)}{Vol(DeptA)} = \frac{6.4}{8} = 0.8$, *in line with the proportions stated in the conditionals at the top.*

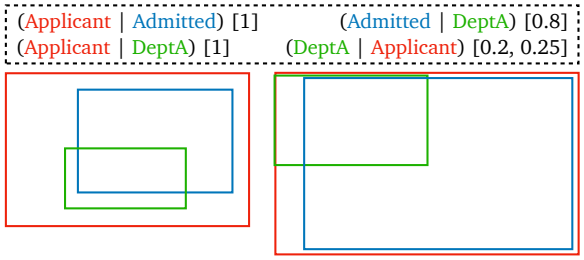

Figure 1: Two possible 2D box embeddings of the concepts Applicant (red), Admitted (blue) and DeptA (green) that maintain proportions stated in the knowledge base.

Concept	Vol. Left	Vol. Right
$Applicant$	40	60
$DeptA$	8	15
$Admitted$	17	50
$Admitted \sqcap DeptA$	6.4	12

Table 1: Volume of boxes in Figure 1.

We can now use the embeddings to estimate unknown proportions between concepts efficiently.

Example 2. *To estimate the bounds for* $(Admitted \sqcap DeptA \mid Applicant)[l_1, u_1]$, *the proportion of all applicants who were admitted via Department A, we take the minimum and maximum proportions observed across our two embeddings. Hence,* $l_1 = \min\{\frac{6.4}{40}, \frac{12}{60}\} = 0.16$ *and* $u_1 = \max\{\frac{6.4}{40}, \frac{12}{60}\} = 0.2$, *resulting in the probability interval* $[0.16, 0.2]$. *This is indeed the exact interval*[1] *as the embeddings happened to take the extreme values.*

While our estimate was exact in the previous example, the interval estimates can be too tight in other cases.

Example 3. *For the overall admission rate* $(Admitted \mid Applicant)[l_2, u_2]$, *we get* $l_2 = \min\{\frac{17}{40}, \frac{50}{60}\} = 0.43$ *and* $u_2 = \max\{\frac{17}{40}, \frac{50}{60}\} = 0.83$, *resulting in the interval estimate* $[0.43, 0.83]$. *However, the exact interval is* $[0.16, 0.96]$.

As we increase the number of embeddings, we increase the chance that one of them yields an estimate close to the extreme values of the interval. Therefore, we expect that increasing the number of embeddings decreases the gap between the true and the estimated interval.

[1]The exact intervals can be computed by solving an (exponentially large) linear optimization problem.

5 A NORMAL FORM FOR $\mathcal{SEL}$

Embeddings of the classical DL $\mathcal{EL}$ [Kulmanov et al., 2019] exploit that every $\mathcal{EL}$ knowledge base can be transformed into an equivalent normal form that contains only GCIs of the following form Baader et al. [2005]:

$$A \sqsubseteq B, A_1 \sqcap A_2 \sqsubseteq B, A \sqsubseteq \exists r.B, \exists r.A \sqsubseteq B,$$

where A, A_1, A_2, B are concept names or $\top$. Every $\mathcal{EL}$ TBox $\mathcal{T}$ can be transformed into a TBox $\mathcal{T}'$ in normal form with only a linear blowup of the size of the knowledge base [Baader et al., 2005]. We show the corresponding transformation rules in Table 6 in Appendix A. We can generalize the normal form to $\mathcal{SEL}$ as follows.

Definition 1 ($\mathcal{SEL}$ normal form)**.** *A $\mathcal{SEL}$ TBox $\mathcal{T}$ is in* normal form *if for every conditional* $(D \mid C)[l, u] \in \mathcal{T}$

- $l = u = 1$ *and* $C \sqsubseteq D$ *is in $\mathcal{EL}$ normal form or*
- $C, D \in N_C$ *are concept names.*

That is, for a normalized conditional $(D \mid C)[l, u]$, either C and D are just concept names or the conditional is deterministic ($l = u = 1$) and therefore just represents an $\mathcal{EL}$ GCI (c.f. Lemma 1). To normalize $\mathcal{SEL}$ TBoxes, we define a new rule that replaces complex concepts C, D in non-deterministic conditionals with new concept names A_1, A_2:

$$SNF0 \qquad (D \mid C)[l, u] \rightarrow (A_2 \mid A_1)[l, u], \quad C \equiv A_1, D \equiv A_2, \tag{2}$$

where $l < 1$ and C or D are complex concepts. We can then normalize an $\mathcal{SEL}$ TBox as follows:

1. Apply $SNF0$ to replace all complex concepts in conditionals with new concept names.
2. Normalize the GCIs using the $\mathcal{EL}$ transformation rules from Table 6 in Appendix A.
3. Apply Lemma 1 to transform the normalized GCIs into deterministic conditionals.

As we prove in Appendix C, the resulting $\mathcal{SEL}$ TBox is in $\mathcal{SEL}$ normal form, entailment-equivalent to the original TBox and its size is linear in the size of the original TBox.

Proposition 1. *Let $\mathcal{T}$ be an $\mathcal{SEL}$ TBox and let $\mathcal{T}'$ denote the transformed TBox. Then $\mathcal{T}'$ is in $\mathcal{SEL}$ normal form and its size is linearly bounded by the size of $\mathcal{T}$. Furthermore, for all $\mathcal{SEL}$ conditionals $(D \mid C)[l, u]$ built up over the concept names occurring in $\mathcal{T}$, we have $\mathcal{T} \models (D \mid C)[l, u]$ if and only if $\mathcal{T}' \models (D \mid C)[l, u]$.*

Note that the $\top$-concept does not have a finite box representation. We must therefore exclude it to avoid problems with the definition of proportions between (the volume of) boxes.

Definition 2. *An $\mathcal{SEL}$ TBox $\mathcal{T}'$ is called* safe *if for every probabilistic conditional $(D \mid C)[l, u] \in \mathcal{T}'$, $\mathcal{T}' \not\models (C \equiv \top)$ and $\mathcal{T}' \not\models (D \equiv \top)$.*

Let us note that $\top$ can still occur in deterministic conditionals. However, probabilistic facts $(D \mid \top)[l, u]$ that can be expressed in full $\mathcal{SEL}$ cannot be captured by our approach.

6 EMBEDDING AND APPROXIMATION

To perform approximate probabilistic reasoning, we construct geometric models for the safe normalized $\mathcal{SEL}$ ontologies $\mathcal{T}'$. In this section, we specify the mapping of concepts and roles to vector representations and formalize the loss terms that encode the probabilistic axioms.

Concept Representation. Following Xiong et al. [2022], Jackermeier et al. [2024], we represent concepts by n-dimensional *boxes*. Formally, we represent the embedding of concepts by two functions m_w, M_w that are parameterized by a learnable parameter vector w. In 2D, $m_w : N_C \to \mathbb{R}^n$ maps concept names to the lower left corner, and $M_w : N_C \to \mathbb{R}^n$ maps them to the upper right corner of their box representation. The *box associated with $C \in N_C$* is

$$\mathrm{Box}_w(C) = \{x \in \mathbb{R}^n \mid m_w(C) \leq x \leq M_w(C)\},$$

where the inequality is defined elementwise-wise. Its volume is defined as:

$$\mathrm{Vol}\left(\mathrm{Box}_w\left(C\right)\right) = \prod_{i=1}^{n} \max\left(0, M_w\left(C\right)_i - m_w\left(C\right)_i\right).$$

Role Representation. Similar to Bordes et al. [2013], Xiong et al. [2022], we associate every role name $r \in N_r$ with an affine transformation $T_w^r(x) = D_w^r x + b_w^r$, where D_w^r is an $(n \times n)$ diagonal matrix with non-negative entries and $b_w^r \in \mathbb{R}^n$ is a vector. Applying T_w^r to the box associated with a concept C results in the box $T_w^r(\mathrm{Box}_w(C)) = \{T_w^r(x) \mid x \in \mathrm{Box}_w(C)\}$ with lower corner $T_w^r(m_w(C))$ and upper corner $T_w^r(M_w(C))$.

When we compute an n-dimensional embedding, our parameter vector contains $2n$ parameters for every concept name $C \in N_C$ ($m_w(C)$ and $M_w(C)$), and $2n$ parameters for every role name $r \in N_R$ (b_w^r and diagonal entries of D_w^r). Hence, the overall size of w is $n \cdot \left(2 \cdot |N_C| + 2 \cdot |N_R|\right)$.

Loss Function. In order to compute the parameter vector w of our embedding, we minimize a loss function $\mathcal{L}(w)$. $\mathcal{L}(w)$ is composed of multiple terms that correspond to the four axiom types that can occur in the normalized TBox. We associate every axiom corresponding to a classical $\mathcal{EL}$ GCIs like in Xiong et al. [2022] by loss terms $\mathcal{L}_{NF1}(w), \mathcal{L}_{NF2}(w), \mathcal{L}_{NF3}(w), \mathcal{L}_{NF4}(w)$ corresponding to the four GCIs occurring in the classical normal form (Details can be found in Appendix D).

We will introduce a new loss term to encode the meaning of probabilistic axioms. Before we do so, let us note that every embedding can be seen as an $\mathcal{SEL}$ interpretation of the following type.

Definition 3 (Geometric Interpretation)**.** *A* geometrical $\mathcal{SEL}$ interpretation (of dimension n) $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ *is an $\mathcal{EL}$ interpretation where $\Delta^{\mathcal{I}} = \mathbb{R}^n$.*

We associate the parameter vector w of an n-dimensional $\mathcal{SEL}$ embedding with the $\mathcal{SEL}$ interpretation $\mathcal{I}_w$ defined by

- $C^{\mathcal{I}_w} = \mathrm{Box}_w(C)$ for all $C \in N_C$,
- $r^{\mathcal{I}_w} = \{(a, b) \in \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}} \mid T_w^r(a) = b\}$ for all $r \in N_R$.

We can approximate the $\mathcal{SEL}$ semantics with (infinite) geometric interpretations by replacing the cardinality of concepts with their volume. The satisfaction condition for the conditional $(D \mid C)[l, u]$ from equation (1) then becomes

$$l \cdot \mathrm{Vol}(C^{\mathcal{I}}) \leq \mathrm{Vol}(D^{\mathcal{I}} \cap C^{\mathcal{I}}) \leq u \cdot \mathrm{Vol}(C^{\mathcal{I}}). \quad (3)$$

If one volume is infinite, we regard the condition as violated. We introduce one loss term for the upper and one for the lower bound in (3) as follows:

$$\mathcal{L}^l_{(D|C)[l,u]}(w) = [l \cdot \mathrm{Vol}(C^{\mathcal{I}}) - \mathrm{Vol}(D^{\mathcal{I}} \cap C^{\mathcal{I}})]^+, \quad (4)$$

$$\mathcal{L}^u_{(D|C)[l,u]}(w) = [\mathrm{Vol}(D^{\mathcal{I}} \cap C^{\mathcal{I}}) - u \cdot \mathrm{Vol}(C^{\mathcal{I}})]^+, \quad (5)$$

where $[x]^+ = \max\{0, x\}$.

7 SOUNDNESS AND RUNTIME

We will now discuss some soundness and runtime guarantees of our approach. All proofs can be found in Appendix C. As explained in the previous section, we compute the parameter vector w of the embedding by minimizing

$$\mathcal{L}_{\mathcal{T}}(w) = \sum_{\alpha \in \mathcal{T}} \mathcal{L}_\alpha(w), \quad (6)$$

where each axiom (conditional) $\alpha \in \mathcal{T}$ is associated with a loss term $\mathcal{L}_\alpha(w)$. The loss term for probabilistic axioms is composed of the lower and upper loss terms described in (4) and (5), that is,

$$\mathcal{L}_{(D|C)[l,u]}(w) = \mathcal{L}^l_{(D|C)[l,u]}(w) + \mathcal{L}^u_{(D|C)[l,u]}(w). \quad (7)$$

An embedding is called *sound* if loss $\mathcal{L}_{\mathcal{T}}(w) = 0$ implies that all axioms are satisfied by the geometric interpretation $\mathcal{I}_w$ associated with w [Kulmanov et al., 2019]. As shown in Xiong et al. [2022], for every axiom in $\mathcal{EL}$ normal form, $\mathcal{L}_\alpha(w) = 0$ implies that $\mathcal{I}_w$ satisfies α. This implies soundness of the $\mathcal{EL}$-embedding in Xiong et al. [2022]. To prove soundness of our $\mathcal{SEL}$-embedding, it suffices to show additionally that, for every probabilistic axiom α, $\mathcal{L}_\alpha(w) = 0$ imply that $\mathcal{I}_w$ satisfies α as well.

Lemma 2. *If $\mathcal{L}_{(D|C)[l,u]}(w) = 0$, then $\mathcal{I}_w \models (D \mid C)[l, u]$.*

Together with the soundness results from Xiong et al. [2022], we obtain the following soundness guarantee.

Theorem 1 (Embedding Soundness)**.** *If $\mathcal{L}_{\mathcal{T}}(w) = 0$, then $\mathcal{I}_w \models \mathcal{T}$.*

The theorem also gives us a one-sided satisfiability test. If we can find an embedding with loss 0, then $\mathcal{T}$ can be satisfied by a geometric $\mathcal{SEL}$ Interpretation. After computing the embedding, we want to use it to perform inference. As we prove in Appendix C, this can be done efficiently.

Theorem 2 (Runtime Guarantee)**.** *For all (complex) concepts C, D built up over N_C such that $\emptyset \neq C^{\mathcal{I}_w}, D^{\mathcal{I}_w} \neq \top^{\mathcal{I}_w}$, we can compute a $p \in [0, 1]$ such that $\mathcal{I}_w \models (D \mid C)[p]$ in time $O((c + d) \cdot n)$, where c and d are the sizes[1] of C and D, respectively, and n is the embedding dimension.*

If the embedding loss is 0, our inference results are sound in the following sense.

Theorem 3 (Inference Soundness)**.** *If $\mathcal{L}_{\mathcal{T}}(w) = 0$ and $\mathcal{T} \models (D \mid C)[l, u]$, then $\mathcal{I}_w \models (D \mid C)[p]$ implies $p \in [l, u]$.*

Intuitively, the result guarantees that our computations are exact when the statistical proportions can be embedded perfectly. If there is an embedding error, this error will propagate through our probability estimates and we will evaluate the approximation quality in such cases empirically in Section 8. To efficiently generate valid consequences of a given TBox, we derive a probabilistic generalization of the classical *Modus Ponens* for $\mathcal{SEL}$ in Appendix C.

Proposition 2 (Probabilistic Modus Ponens (PMP))**.** *For all $\mathcal{EL}$ concepts C, D, E (atomic or complex) and all $\mathcal{SEL}$ interpretations $\mathcal{I}$, if $\mathcal{T} \models (D \mid C)[l_1, u_1]$ and $\mathcal{T} \models (E \mid C \sqcap D)[l_2, u_2]$, then $\mathcal{T} \models (E \mid C)[l_3, u_3]$, where $l_3 = l_1 \cdot l_2$ and $u_3 = \min\{1, u_1 \cdot u_2 + 1 - l_1\}$.*

8 EXPERIMENTS

Before presenting the experimental results, we will describe our implementation, the procedure for constructing test ontologies and our evaluation protocol.

8.1 IMPLEMENTATION

We implemented a first prototype which supports conditionals of the following types:

$$(B|A)[p], (B|A_1 \sqcap A_2)[p], (B|\exists r.A)[p], (\exists r.B|A)[p], \tag{8}$$

where $A, A_1, A_2, B \in N_C$ are concept names. As their structure corresponds to the four axioms occuring in $\mathcal{EL}$ normal form, we will refer to them as PNF1 - PNF4.

Xiong et al. [2022] added a *location regularizer* $\mathcal{L}_{loc}(w)$ to the loss function that is defined as

$$\sum_{C \in N_C} \sum_{i=1}^{n} [M_w(C)_i - \beta + \epsilon]^+ + [-m_w(C)_i - \epsilon]^+, \tag{9}$$

where β is a hyperparameter that bounds the upper and lower corner of boxes. Intuitively, $\mathcal{L}_{loc}(w)$ encourages that concepts are embedded in the hypercube $[0, \beta]^n$. As we show in the ablation study in Table 12, just regularizing the volume of boxes can yield better results. We therefore also consider the following *volume regularizer*:

$$\mathcal{L}_{vol}(w) = \sum_{i=1}^{m} [\beta^n - Vol(C_i) - \epsilon]^+ \tag{10}$$

Even though $\mathcal{L}_{loc}$ and $\mathcal{L}_{vol}$ have some redundancy, we found that their combination can work better than each term individually. The loss function that we minimize is therefore

$$\mathcal{L}(w) = \mathcal{L}_{\mathcal{T}}(w) + \mathcal{L}_{loc}(w) + \mathcal{L}_{vol}(w). \tag{11}$$

We performed hyperparameter search for embedding dimensions $n = [8, 16, 32, 64, 128]$, side lengths $\beta = [1, 10]$ and learning rates $lr = [0.001, 0.01, 0.05, 0.1]$ for Adam [Kingma and Ba, 2015]. Table 8 summarizes the best hyperparameters. Our experiments ran on a Linux machine with a 40GB NVIDIA A100 SXM4 GPU. We provide additional details in Appendix E.

8.2 DATASET CONSTRUCTION

To the best of our knowledge, there are currently no $\mathcal{SEL}$ ontologies available for evaluating our method. Therefore, we derived three datasets—*Country*, *Hybrid*, and *Person*—from the high-quality real-world knowledge base YAGO3 [Mahdisoltani et al., 2014]. These ontologies were created by computing statistics for conditionals of types PNF1 to PNF4 (c.f., Equation (8)).

Specifically, we first select a limited number of concepts and role names of interest. We then systematically generate all possible conditionals of types PNF1 to PNF4 that can be created from the selected concepts and role names, assigning probabilities based on statistical proportions. For example, to compute the probability p of the probabilistic conditional $(\exists hasChild \mid Person)[p]$, we count the number of persons who have children and divide it by the total number of persons in YAGO3. More details about the dataset construction process can be found in Appendix E.2.

It is important to note that we consider only a subset of the theoretically possible conditionals due to the following constraints:

[1] Roughly speaking, the size of a concept corresponds to the number of constructors used to create it [Baader et al.].

	#Concepts	#Role Names	#T-Box Axioms (the ratio of PNF1-4)
Person	135	6	325,600 (0.21/0.31/0.24/0.24)
Country	30	1	26,974 (0.24/0.5/0.13/0.13)
Hybrid	12	18	7,492 (0.02/0.22/0.38/0.38)

Table 2: Statistics of our benchmark ontologies.

- **Undefined Probabilities**: Conditionals with undefined probabilities are excluded. For instance, $(B|A_1 \sqcap A_2)$ is undefined if A_1 and A_2 are disjoint because the probability of the condition is 0.
- **Redundancy**: We removed redundant axioms. For example, if A_1 and A_2 are disjoint, conditionals stating that their subconcepts are disjoint are redundant.

Table 2 presents statistical information about the ontologies after the filtering process.

8.3 EXPERIMENTAL SETUP

Our reasoning task involves $\mathcal{EL}$ concepts C, D, E (atomic or complex). Given premises $(D \mid C)[l_1, u_1]$, $(E \mid C \sqcap D)[l_2, u_2]$, and a query $(E \mid C)$, our goal is to infer the associated probability interval $[l_3, u_3]$. As we illustrated in the examples in Section 4, if our embedding approximately captures the proportions stated in premise conditionals, we should be able to approximate the probability interval by calculating proportions between boxes in the embedding space. Theorem 3 implies that, whenever the embedding loss is 0, the probability interval cannot be too large. However, it can be too small as illustrated in Example 3. Furthermore, the embedding error will typically be non-zero, and so the interval can also be too large.

Data Split. We split axioms in $\mathcal{SEL}$ ontologies into two subsets: a *learning set* for training embeddings and a *query set* for evaluating approximation quality. In our experiment, we select 30% of axioms from the ontology as the query set for evaluation. Importantly, the split is non-random. For each query $(E \mid C)$, we remove the corresponding conditional $(E \mid C)[l_3, u_3]$ from the ontology and ensure that the premises $(D \mid C)[l_1, u_1]$ and $(E \mid C \sqcap D)[l_2, u_2]$ are present in the ontology to allow non-trivial inferences.

PMP Evaluation. Let q denote the number of queries. By construction, each query has corresponding premise conditionals in the ontology, allowing us to apply PMP (Proposition 2) to infer a probability interval $[l, u]$. To approximate this interval, we compute N embeddings with different random seeds, yielding N point estimates $p_1, \ldots, p_N$ (by computing proportions in the embedding) for each query. The lower and upper bounds are estimated as $\overline{l} = \min\{p_1, \ldots, p_N\}$ and $\overline{u} = \max\{p_1, \ldots, p_N\}$, respectively. In our experiments, we typically generate five query sets by varying the data split and set $N = 60$.

A detailed description of the query set construction and the pseudocode for PMP evaluation are provided in Appendix E.3.

8.4 EVALUATION PROTOCOL

Metrics. We evaluate the *embedding quality* using *Mean Absolute Error* (MAE), which measures the average absolute difference between the probabilities in the knowledge base and the corresponding proportions in the embedding.

In order to evaluate the *inference quality* of our approximation $[\overline{l}, \overline{u}]$, we consider three different metrics.

- *Soundness Accuracy* (SA): This metric measures the percentage of queries for which the approximate interval fell into the true interval.

$$SA := \frac{1}{q} \sum_{i=1}^{q} \mathbf{1}_{[\overline{l}_i, \overline{u}_i] \subseteq [l_i, u_i]}, \tag{12}$$

  where $\mathbf{1}_{S \subseteq T}$ is 1 if $S \subseteq T$ and 0 otherwise.

- *Soundness Error* (SE): This metric quantifies the extent to which the estimated interval $[\overline{l}, \overline{u}]$ extends beyond (overestimates) the true interval $[l, u]$.

$$SE := \frac{1}{q} \sum_{i=1}^{q} \left([l_i - \overline{l}_i]^+ + [\overline{u}_i - u_i]^+ \right). \tag{13}$$

- *Approximation Gap* (AG): This metric captures the absolute difference between the estimated and true lower/upper bounds.

$$AG := \frac{1}{q} \sum_{i=1}^{q} \left(|l_i - \overline{l}_i| + |u_i - \overline{u}_i| \right). \tag{14}$$

Baselines. To the best of our knowledge, there are currently no existing implementations for statistical reasoning in $\mathcal{SEL}$ that can be used for direct comparison (see the discussion in Section 2). To demonstrate that our approach gives non-trivial solutions, we consider several simple baselines to serve as points of reference for evaluating the performance of our approach:

- *Fixed Estimator*: An interval estimator that always outputs $[0, 1]$ as the predicted interval. This estimator would perform (almost) perfectly for SA, SE and AG if all true intervals were (close to) $[0, 1]$.
- *Random Estimator*: An interval estimator that randomly samples $\overline{l}, \overline{u}$ from $[0, 1]$, ensuring $\overline{l} \leq \overline{u}$, as the predicted interval. The motivation is to demonstrate that our method works significantly better than a random guess.
- *KDE Estimator*: This method estimates the probability density function (PDF) of the intervals in the learning set using a Kernel Density Estimation (KDE) method

Dataset	Method	SA ↑	SE ↓	AG ↓
Person	Fixed Estimator	4.50%	0.301	0.301
	Random Estimator	52.70%	0.262	0.572
	KDE Estimator	9.19%	0.241	0.275
	BoxSEL (ours)	**89.30%**	**0.013**	**0.233**
Country	Fixed Estimator	29.10%	0.105	**0.105**
	Random Estimator	61.30%	0.048	0.603
	KDE Estimator	3.10%	0.086	0.329
	BoxSEL (ours)	**95.60%**	**0.004**	0.241
Hybrid	Fixed Estimator	1.5%	0.734	0.734
	Random Estimator	27.8%	0.537	0.533
	KDE Estimator	1.1%	0.661	0.353
	BoxSEL (ours)	**89.70%**	**0.020**	**0.293**

Table 3: Approximation quality of baselines and BoxSEL.

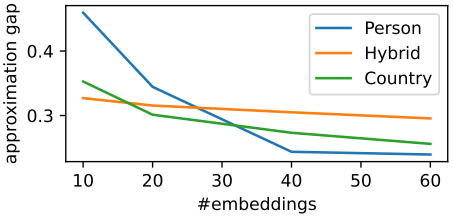

Figure 2: Approximation gap for approximation based on 10, 20, 40 and 60 embeddings.

[Davis et al., 2011]. It then samples an interval from the estimated PDF as the predicted interval. Further details are provided in Appendix E.4.

Note that existing methods (e.g. ELEm, EMEL^{++}, BoxEL and Box2EL) only handle classical $\mathcal{EL}$ knowledge bases and not $\mathcal{SEL}$ knowledge bases. Therefore, we cannot consider them as baselines for inference in $\mathcal{SEL}$. However, we perform ablation studies on alternative design choices inspired by these methods in Section 8.6.

8.5 EXPERIMENTAL RESULTS AND ANALYSIS

We now present and discuss the results of our experiments, focusing on the following questions: **RQ1**: How effectively does our approach approximate statistical reasoning in $\mathcal{SEL}$? **RQ2**: What is the relationship between embedding quality and the quality of approximate inference? **RQ3**: How efficient is the inference process of our approach?

8.5.1 Approximation Quality

Table 3 presents the approximation quality for baselines and our method, BoxSEL, across three test $\mathcal{SEL}$ ontologies: Person, Country, and Hybrid. Our approach significantly outperforms the baselines across SA and SE. Specifically, a large proportion of the approximations of our approach is sound in the sense that the estimated intervals are contained in the true intervals (SA close to 100%) and the soundness error is relatively small (2% absolute error or less). The approximation gap for BoxSEL is between 0.2 and 0.3. Let us note that the approximation gap for the fixed estimator is about 0.3 for the Person and 0.1 for the Country dataset. This shows that the ground truth intervals in these benchmarks are relatively large (the average width is 0.7 for Person and 0.9 for Country). We note that the approximation gap of our approach is not very sensitive with respect to the width of the ground truth interval. While we can see an increase in precision (SA and SE) for benchmarks with smaller intervals, it remains relatively large (above 89 %) for all benchmarks independent of the width of the ground truth intervals.

Figure 2 illustrates how the approximation gap decreases as the number of embeddings used increases. This confirms our intuition since adding more samples increases the chance of picking a sample close to the extreme values of the interval.

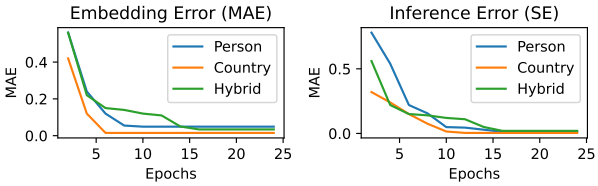

Figure 3: Relationship between embedding and inference error.

	Embedding (s/epoch)	Inference (ms/query)
Person	117.66	0.09
Country	8.96	0.23
Hybrid	12.15	0.41

Table 4: Evaluation of runtime of computing the embedding and performing inference on the computed embedding.

8.5.2 Inference vs Embedding Error

We expect that the inference error (how well do we approximate the query interval?) decreases with the embedding error (how well could the axioms in the ontology be represented by embeddings?) because errors in the embedding of proportions can cause errors in the derived estimates. In Figure 3, we illustrate this relationship by plotting embedding and inference error against the number of epochs used for computing the embeddings. We can see a clear relationship between inference and embedding error. This is a useful property in practice because the embedding error is known after computing the embedding. If the embedding

error is low, we expect good approximation results. If the embedding error is large, we should be more skeptical about the inferred probability intervals and invest more time into finding better embeddings if a higher accuracy is needed.

8.5.3 Runtime

Table 4 shows runtime results for computing embeddings (per epoch) and performing inference (per query) on the embedding. Note that each embedding has to be computed only once to transform an ontology into an embedding. Afterwards, the embedding can be used to answer an arbitrary number of queries over the ontology. Table 4 shows that even a training epoch for the large *Person* dataset can be performed in less than two minutes on average. Since embeddings are computed by gradient descent variants, the runtime per epoch depends linearly on the size of the embedding vector and the number of axioms. As we explained at the end of Section 6, the embedding vector grows linearly with respect to the individual, concept, and role names. So, overall, the runtime for computing embeddings is linear with respect to the size of the knowledge base and the number of epochs. For the experiments in Table 3, computing one embedding for the person/country/hybrid dataset took 39/3/6 minutes on average. For all datasets, inference is a matter of milliseconds as we can expect from the linear runtime guarantees stated in Theorem 2.

8.6 ABLATION STUDY

We conduct three ablation studies to investigate the impact of different concept and role representation, as well as the effect of regularization, on the approximation quality.

Impact of Concept Representation. Inspired by ELEm and EMEL^{++}, we explored an alternative representation of concepts as *n-balls*, while keeping all other components of our method unchanged. The results, presented in Table 11, show that representing concepts as boxes consistently outperforms the n-balls representation across all metrics. These findings emphasize the advantages of using boxes for concept representation, both theoretically and empirically.

Impact of Role Representation. There are several alternative methods for representing roles in the literature. ELEm and EMEL^{++} use translation, BoxEL employs affine transformation, and Box2EL utilizes dual boxes to represent roles. We compare these three approaches in Table 5. Our findings indicate that affine transformation achieves the highest performance in the Person and Country ontologies and the dual boxes representation slightly outperforms affine transformation in the Hybrid ontology. However, this improvement comes at the cost of increased model complexity. Dual boxes require $4n \cdot |N_R| + n \cdot |N_C|$ parameters to represent roles [Jackermeier et al., 2024], which is more than twice the parameters required by the affine transformation ($2n \cdot |N_R|$). Additionally, the dual boxes approach shows reduced performance in the Person and Country ontologies.

Dataset	Role Representation	SA ↑	SE ↓	AG ↓
Person	translation	87.3%	0.017	**0.232**
	affine transformation*	**89.3%**	**0.013**	0.233
	dual boxes	85.7%	0.109	0.279
Country	translation	94.7%	0.005	0.245
	affine transformation*	**95.6%**	**0.004**	**0.241**
	dual boxes	93.2%	0.013	0.261
Hybrid	translation	79.8%	0.039	0.293
	affine transformation*	89.7%	**0.020**	0.293
	dual boxes	**89.8%**	0.031	**0.287**

Table 5: Ablation study of the role representation. The design choice of our method is marked with an asterisk (*).

Impact of Regularization. We investigate the impact of the regularization terms ($\mathcal{L}_{loc}$ and $\mathcal{L}_{vol}$). Table 12 shows that volume regularization significantly improves performance. Location regularization further enhances performance and reduces the number of epochs required for convergence.

9 DISCUSSION AND FUTURE WORK

We showed how embeddings can be used to approximate probabilistic inference in $\mathcal{SEL}$ by extending the $\mathcal{EL}$-embedding BoxEL from Xiong et al. [2022]. As shown in Section 7, the resulting embeddings support point estimates for $\mathcal{SEL}$ queries in linear time, with guaranteed soundness when the embedding loss is zero. When the loss is non-zero, the inference error increases, but our experiments show a clear proportional relationship between the two, making the known embedding error a practical quality indicator. To approximate full probability intervals, we compute multiple embeddings and combine the individual point estimates. In our experiments, this procedure rarely overestimated the true intervals, and approximation quality improved steadily as the number of embeddings increased. In practice, $\mathcal{SEL}$ ontologies can be constructed fully automatically from arbitrary knowledge graphs by applying rule mining methods [Omran et al., 2018, Meilicke et al., 2019, Guimarães et al., 2021] and using the confidence values (which are usually statistical proportions) of rules for the probabilities.

Our inference soundness guarantee (Theorem 3) also sheds light on BoxEL from Xiong et al. [2022]: if all conditionals in an $\mathcal{SEL}$ ontology are deterministic, it reduces to an $\mathcal{EL}$ ontology (Lemma 1), and our $\mathcal{SEL}$ embedding becomes equivalent to the BoxEL embedding. Our soundness theorem thus guarantees that inference under BoxEL embeddings is sound as well when the loss is zero, and our empirical results suggest that the approximation error remains small when the loss is small.

10 ACKNOWLEDGEMENTS

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Yuqicheng Zhu. This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1574 - Project number 471687386, and by the EU Project SMARTY (GA 101140087).

References

Franz Baader. Least common subsumers and most specific concepts in a description logic with existential restrictions and terminological cycles. In *IJCAI*, pages 319–324, 2003.

Franz Baader and Andreas Ecke. Extending the description logic ALC with more expressive cardinality constraints on concepts. In *GCAI*, pages 6–19.

Franz Baader, Ian Horrocks, Carsten Lutz, and Ulrike Sattler. *An Introduction to Description Logic*.

Franz Baader, Sebastian Brandt, and Carsten Lutz. Pushing the EL envelope. In *IJCAI*, pages 364–369, 2005.

Bartosz Bednarczyk. Statistical EL is exptime-complete. *Information Processing Letters*, 169:106113, 2021.

Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404, 1975.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795, 2013.

Andrew W Brown, Kathryn A Kaiser, and David B Allison. Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences*, 115(11):2563–2570, 2018.

Richard A Davis, Keh-Shin Lii, and Dimitris N Politis. Remarks on some nonparametric estimates of a density function. *Selected Works of Murray Rosenblatt*, pages 95–100, 2011.

Alan M. Frisch and Peter Haddawy. Anytime deduction for probabilistic logic. *Artificial Intelligence*, 69(1-2): 93–122, 1994.

Dinesh Garg, Shajith Ikbal, Santosh K. Srivastava, Harit Vishwakarma, Hima P. Karanam, and L. Venkata Subramaniam. Quantum embedding of knowledge for reasoning. In *NeurIPS*, pages 5595–5605, 2019.

Ricardo Guimarães, Ana Ozaki, Cosimo Persia, and Baris Sertkaya. Mining EL bases with adaptable role depth. In *AAAI*, pages 6367–6374, 2021.

Víctor Gutiérrez-Basulto, Jean Christoph Jung, Carsten Lutz, and Lutz Schröder. Probabilistic description logics for subjective uncertainty. *Journal of Artificial Intelligence Research*, 58:1–66, 2017.

Joseph Y Halpern. An analysis of first-order logics of probability. *Artificial intelligence*, 46(3):311–350, 1990.

Anthony Hunter and Nico Potyka. Syntactic reasoning with conditional probabilities in deductive argumentation. *Artificial Intelligence*, 321:103934, 2023.

Mathias Jackermeier, Jiaoyan Chen, and Ian Horrocks. Dual box embeddings for the description logic EL++. In *WWW*, pages 2250—-2258, 2024.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Maxat Kulmanov, Wang Liu-Wei, Yuan Yan, and Robert Hoehndorf. EL embeddings: Geometric construction of models for the description logic EL++. In *IJCAI*, pages 6103–6109, 2019.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.

Hanxiao Liu, Yuexin Wu, and Yiming Yang. Analogical inference for multi-relational embeddings. In *ICML*, pages 2168–2178, 2017.

Thomas Lukasiewicz and Umberto Straccia. Managing uncertainty and vagueness in description logics for the semantic web. *Journal of Web Semantics*, 6(4):291–308, 2008.

Carsten Lutz and Lutz Schröder. Probabilistic description logics for subjective uncertainty. In *KR*, pages 393–403, 2010.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. YAGO3: A knowledge base from multilingual Wikipedias. In *7th biennial conference on innovative data systems research*, 2014.

Christian Meilicke, Melisachew Wudage Chekol, Daniel Ruffinelli, and Heiner Stuckenschmidt. Anytime bottom-up rule learning for knowledge graph completion. In *IJCAI*, pages 3137–3143, 2019.

Pouya Ghiasnezhad Omran, Kewen Wang, and Zhe Wang. Scalable rule learning via learning representation. In *IJCAI*, pages 2149–2155, 2018.

Özgür Lütfü Özçep, Mena Leemhuis, and Diedrich Wolter. Cone semantics for logics with negation. In *IJCAI*, pages 1820–1826, 2020.

Dhruvesh Patel, Shib Sankar Dasgupta, Michael Boratko, Xiang Li, Luke Vilnis, and Andrew McCallum. Representing joint hierarchies with box embeddings. In *AKBC*, 2020.

Rafael Peñaloza and Nico Potyka. Probabilistic reasoning in the description logic *ALCP* with the principle of maximum entropy. In *SUM*, pages 246–259, 2016.

Rafael Peñaloza and Nico Potyka. Towards statistical reasoning in description logics over finite domains. In *SUM*, pages 280–294, 2017.

Xi Peng, Zhenwei Tang, Maxat Kulmanov, Kexin Niu, and Robert Hoehndorf. Description logic EL++ embeddings with intersectional closure. *CoRR*, abs/2202.14018, 2022.

David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

Matthew S Thiese, Zachary C Arnold, and Skyler D Walker. The misuse and abuse of statistics in biomedical research. *Biochemia medica*, 25(1):5–11, 2015.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Bo Xiong, Nico Potyka, Trung-Kien Tran, Mojtaba Nayyeri, and Steffen Staab. Faithful embeddings for el++ knowledge bases. In *ISWC*, pages 22–38, 2022.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR (Poster)*, 2015.

Jie Zheng, Huan Li, Qingzhi Liu, and Yongqun He. The ontology of biological and clinical statistics (OBCS)-based statistical method standardization and meta-analysis of host responses to yellow fever vaccines. *Quantitative Biology*, 5(4):291–301, 2017.

Zhuoxun Zheng, Baifan Zhou, Dongzhuoran Zhou, Akif Quddus Khan, Ahmet Soylu, and Evgeny Kharlamov. Towards A statistic ontology for data analysis in smart manufacturing. In *ISWC (Posters/Demos/Industry)*, volume 3254 of *CEUR Workshop Proceedings*, 2022.

A EL TRANSFORMATION RULES

$$
\begin{array}{lllll}
NF0 & \hat{D} \sqsubseteq \hat{E} & \rightarrow & \hat{D} \sqsubseteq A, A \sqsubseteq \hat{E}. \\
NF1_r & C \sqcap \hat{D} \sqsubseteq B & \rightarrow & \hat{D} \sqsubseteq A, C \sqcap A \sqsubseteq B. \\
NF1_l & \hat{D} \sqcap C \sqsubseteq B & \rightarrow & \hat{D} \sqsubseteq A, A \sqcap C \sqsubseteq B. \\
NF2 & \exists r.\hat{D} \sqsubseteq B & \rightarrow & \hat{D} \sqsubseteq A, \exists r.A \sqsubseteq B. \\
NF3 & B \sqsubseteq \exists r.\hat{D} & \rightarrow & A \sqsubseteq \hat{D}, B \sqsubseteq \exists r.A. \\
NF4 & B \sqsubseteq D \sqcap E & \rightarrow & B \sqcap D, B \sqcap E.
\end{array}
$$

Table 6: Transformation rules for normalizing $\mathcal{EL}$ TBoxes from Baader et al.. C, D, E are arbitrary $\mathcal{EL}$ concepts, $\hat{D}, \hat{E}$ are non-atomic concepts, B is a concept name and A is a new concept name.

B RESOLVING SIMPSON'S PARADOX WITH $\mathcal{SEL}$

We illustrate how $\mathcal{SEL}$ can detect Simpson's paradox using a simplified version of the UC Berkeley admissions scenario from Section 4. Consider two departments, A and B, with the following conditionals about female applicants:

$$
\begin{aligned}
(\textit{Admitted} \mid \textit{DeptA} \sqcap \textit{Woman})[0.80], \\
(\textit{Admitted} \mid \textit{DeptB} \sqcap \textit{Woman})[0.90], \\
(\textit{DeptA} \mid \textit{Woman})[0.90], \\
(\textit{DeptB} \mid \textit{Woman})[0.10],
\end{aligned}
$$

and the corresponding conditionals for male applicants:

$$
\begin{aligned}
(\textit{Admitted} \mid \textit{DeptA} \sqcap \textit{Man})[0.75], \\
(\textit{Admitted} \mid \textit{DeptB} \sqcap \textit{Man})[0.85], \\
(\textit{DeptA} \mid \textit{Man})[0.10], \\
(\textit{DeptB} \mid \textit{Man})[0.90].
\end{aligned}
$$

Women are admitted at higher rates than men in *both* departments (80% vs. 75% in A, 90% vs. 85% in B). However, 90% of female applicants apply to Department A, which has the lower admission rate, while 90% of male applicants apply to Department B.

Given these conditionals, $\mathcal{SEL}$ can derive the overall admission rates by querying $(\textit{Admitted} \mid \textit{Woman})$ and $(\textit{Admitted} \mid \textit{Man})$. Since departments A and B partition the applicants of each gender, the law of total probability gives:

$$
\begin{aligned}
P(\textit{Admitted} \mid \textit{Woman}) = 0.80 \times 0.90 + 0.90 \times 0.10 = 0.81, \\
P(\textit{Admitted} \mid \textit{Man}) = 0.75 \times 0.10 + 0.85 \times 0.90 = 0.84.
\end{aligned}
$$

The aggregate male admission rate (0.84) exceeds the aggregate female rate (0.81), even though women are admitted at higher rates in every individual department. This is Simpson's paradox: the uneven distribution of applicants across departments reverses the comparison at the aggregate level.

$\mathcal{SEL}$ makes this reversal explicit. By encoding department-level admission rates and applicant distributions as conditionals, and then querying the aggregate rates, the paradox becomes a direct consequence of the derived bounds rather than a hidden artifact of aggregation.

C PROOFS OF TECHNICAL RESULTS

To improve comprehensibility of the proof, we rewrite Proposition 1 slightly by enumerating the main claims.

Proposition 1. *Let $\mathcal{T}$ be an $\mathcal{SEL}$ TBox and let $\mathcal{T}'$ denote the TBox resulting from our transformation procedure.*

1. *$\mathcal{T}'$ is in $\mathcal{SEL}$ normal form.*

2. *For all $\mathcal{SEL}$ conditionals $(D \mid C)[l, u]$ built up over the concept names occurring in $\mathcal{T}$, we have $\mathcal{T} \models (D \mid C)[l, u]$ if and only if $\mathcal{T}' \models (D \mid C)[l, u]$.*

3. *The size of $\mathcal{T}'$ is linearly bounded by the size of $\mathcal{T}$.*

Proof. 1. After step 1, all remaining conditionals are non-deterministic and contain only concept names. Therefore, they satisfy condition 2 of the normal form. In step 2 and 3, the equivalences are translated into GCIs and normalized. Step 4 translates the normalized GCIs into deterministic conditionals. Since the GCIs were in $\mathcal{EL}$ normal form, the resulting deterministic conditionals satisfy condition 1 of the normal form. Hence, $\mathcal{T}'$ is in $\mathcal{SEL}$ normal form.

2. First assume $\mathcal{T} \models (D \mid C)[l, u]$ and let $\mathcal{I}'$ be an interpretation that satisfies $\mathcal{T}'$. Then, by construction of $\mathcal{T}'$, the interpretation $\mathcal{I}$ obtained from $\mathcal{I}'$ by restricting to the concept names occurring in $\mathcal{T}$ also satisfies $\mathcal{T}$. Hence, by assumption, $\mathcal{I} \models (D \mid C)[l, u]$. Since $\mathcal{I}$ and $\mathcal{I}'$ interpret the concept names occurring in $\mathcal{T}$ equally, we also have $\mathcal{I}' \models (D \mid C)[l, u]$.

Conversely, assume $\mathcal{T}' \models (D \mid C)[l, u]$ and let $\mathcal{I}$ be an interpretation that satisfies $\mathcal{T}$. Extend $\mathcal{I}$ to an interpretation $\mathcal{I}'$ of $\mathcal{T}'$ as follows: for every new concept name A occurring in $\mathcal{T}'$ but not in $\mathcal{T}$, there must be a GCI $A \sqsubseteq C$. In particular, we can order the new concept names such that C contains only concept names that have already been interpreted (just follow the order that the normalization algorithm used when it introduced new concept names). We then let $A^{\mathcal{I}'} = C^{\mathcal{I}'}$. Then $\mathcal{I}'$ satisfies $\mathcal{T}'$ as well and, by assumption, $\mathcal{I}'$ satisfies $(D \mid C)[l, u]$. Since, $\mathcal{I}'$ extends $\mathcal{I}$, $\mathcal{I}$ satisfies $(D \mid C)[l, u]$ as well.

3. After step 1, every conditional is replaced by at most one conditional and two equivalences. The length of each of these is bounded by the length of the original conditional. Hence, the size remains linear after step 1. In step 2, each equivalence is replaced by 2 GCIs of the same length. Hence, for every conditional, we introduce at most 4 GCIs whose length is at most the length of the original conditional. We then apply the $\mathcal{EL}$ normalization rules that guarantee that the size of the resulting GCIs is linear in the size of the original GCIs. Hence, it is also linear in the size of the original conditionals (see Lemma 6.2 in Baader et al.). Hence, after translating the GCIs into deterministic conditionals in step 4, the size of $\mathcal{T}'$ is linear in the size of $\mathcal{T}$. □

Lemma 2. *If $\mathcal{L}_{(D|C)[l,u]}(w) = 0$, then $\mathcal{I}_w \models (D \mid C)[l, u]$.*

Proof. Since $\mathcal{L}^l_{(D|C)[l,u]}(w)$ and $\mathcal{L}^u_{(D|C)[l,u]}(w)$ are non-negative by definition, $\mathcal{L}_{(D|C)[l,u]}(w) = 0$ implies that both terms are equal to 0. Hence, $l \cdot \mathrm{Vol}(C^{\mathcal{I}}) \leq \mathrm{Vol}(D^{\mathcal{I}} \cap C^{\mathcal{I}})$, $\mathrm{Vol}(D^{\mathcal{I}} \cap C^{\mathcal{I}}) \leq u \cdot \mathrm{Vol}(C^{\mathcal{I}})$ and $\mathcal{I}_w \models (D \mid C)[l, u]$. □

Theorem 1 (Embedding Soundness)**.** *If $\mathcal{L}_{\mathcal{T}}(w) = 0$, then $\mathcal{I}_w \models \mathcal{T}$.*

Proof. All loss terms in $\mathcal{L}_{\mathcal{T}}(w) = \sum_{\alpha \in \mathcal{T}} \mathcal{L}_\alpha(w)$ are non-negative by definition. Hence, $\mathcal{L}_{\mathcal{T}}(w) = 0$ implies $\mathcal{L}_\alpha(w) = 0$ for all $\alpha \in \mathcal{T}$. $\mathcal{L}_\alpha(w) = 0$ together with Proposition 1-5 from Xiong et al. [2022] (for deterministic axioms) along with Lemma 2 (for probabilistic axioms) imply that $\mathcal{I}_w \models \alpha$ for all $\alpha \in \mathcal{T}$. Hence, $\mathcal{I}_w \models \mathcal{T}$. □

Theorem 2 (Runtime Guarantee)**.** *For all (complex) concepts C, D built up over N_C such that $\emptyset \neq C^{\mathcal{I}_w}, D^{\mathcal{I}_w} \neq \top^{\mathcal{I}_w}$, we can compute a $p \in [0, 1]$ such that $\mathcal{I}_w \models (D \mid C)[p]$ in time $O((c + d) \cdot n)$, where c and d is the size of C and D, respectively, and n is the embedding dimension.*

Proof. In order to compute p, we have to compute $\mathrm{Vol}(D^{\mathcal{I}} \cap C^{\mathcal{I}})$ and $\mathrm{Vol}(C^{\mathcal{I}})$ and can then let $p = \frac{\mathrm{Vol}(D^{\mathcal{I}} \cap C^{\mathcal{I}})}{\mathrm{Vol}(C^{\mathcal{I}})}$. We will now explain how we can compute the volume.

If C, D are concept names, $\mathrm{Vol}(C^{\mathcal{I}})$ can be directly computed from (6), which takes time $O(n)$. To compute $\mathrm{Vol}(D^{\mathcal{I}} \cap C^{\mathcal{I}})$, we first have to compute $\mathrm{Box}_w\left(C^{\mathcal{I}} \cap D^{\mathcal{I}}\right)$. We have $m_w(C^{\mathcal{I}} \cap D^{\mathcal{I}}) = \max\{m_w(C^{\mathcal{I}}), m_w(D^{\mathcal{I}})\}$ for the lower corner and $M_w(C^{\mathcal{I}} \cap D^{\mathcal{I}}) = \min\{M_w(C^{\mathcal{I}}), M_w(D^{\mathcal{I}})\}$ for the upper corner, where minimum and maximum are taken componentwise. If one component of $M_w(C^{\mathcal{I}} \cap D^{\mathcal{I}}) - m_w(C^{\mathcal{I}} \cap D^{\mathcal{I}})$ is negative, the intersection is empty and the volume is 0. Otherwise, we can again apply equation (6) to compute the volume in time $O(n)$.

If C or D are not concept names, we decompose them recursively. We demonstrate this for C.

If $C = C_1 \sqcap C_2$, then we can associate C with a box as in the previous case in time $O(n)$ and continue with this box.

$C = \exists r.C_1$, the box associated with C is defined by $m_w(C) = Am_w(C_1)$ and $M_w(C) = AM_w(C_1)$, where A is the inverse matrix of T^r_w. Since T^r_w is a diagonal matrix, the inverse matrix is obtained by simply inverting the diagonal entries ($x \mapsto \frac{1}{x}$). Since A is a diagonal matrix as well, we only have to multiply the diagonal entries of A by the entries in the vector, so that the overall runtime is again $O(n)$. We then continue again with the box representation of C.

Each decomposition step can be performed in time $O(n)$ and the number of decomposition steps is bounded by the size of C and D. Therefore, the overall runtime is $O((c+d)\cdot n)$. □

Theorem 3 (Inference Soundness)**.** *If $\mathcal{L}_{\mathcal{T}}(w) = 0$ and $\mathcal{T} \models (D \mid C)[l,u]$, then $\mathcal{I}_w \models (D \mid C)[p]$ implies $p \in [l,u]$.*

Proof. If $\mathcal{L}_{\mathcal{T}}(w) = 0$, our soundness guarantee implies that $\mathcal{I} \models \mathcal{T}$. Since $\mathcal{T} \models (D \mid C)[l,u]$, every model of $\mathcal{T}$ respects the bounds l and u. Hence, we must have $p \in [l,u]$. □ □

Proposition 2 (Probabilistic Modus Ponens (PMP))**.** *For all $\mathcal{EL}$ concepts C, D, E (atomic or complex) and all $\mathcal{SEL}$ interpretations $\mathcal{I}$, if $\mathcal{T} \models (D \mid C)[l_1,u_1]$ and $\mathcal{T} \models (E \mid C \sqcap D)[l_2,u_2]$, then $\mathcal{T} \models (E \mid C)[l_3,u_3]$, where $l_3 = l_1 \cdot l_2$ and $u_3 = \min\{1, u_1 \cdot u_2 + 1 - l_1\}$.*

Proof. Let us note that the proof works analogously for counting and volume semantics. We give the proof for the latter, but by just replacing the volume with cardinality, we obtain the proof for counting semantics.

Consider an arbitrary $\mathcal{SEL}$ interpretation $\mathcal{I}$ such that $\mathcal{I} \models \mathcal{T}$. To prove the claim, we have to check that the lower and upper bounds in (3) hold for the conditional $(E \mid C)[l_3,u_3]$. For the lower bound, we have

$$\begin{aligned}\mathrm{Vol}(E^{\mathcal{I}} \cap C^{\mathcal{I}}) &= \mathrm{Vol}(E^{\mathcal{I}} \cap D^{\mathcal{I}} \cap C^{\mathcal{I}}) + \\ &\quad \mathrm{Vol}(E^{\mathcal{I}} \cap (\neg D)^{\mathcal{I}} \cap C^{\mathcal{I}}) \\ &\geq l_2 \,\mathrm{Vol}(C^{\mathcal{I}} \cap D^{\mathcal{I}}) + 0 \\ &\geq (l_2 \cdot l_1) \cdot \mathrm{Vol}(C^{\mathcal{I}}),\end{aligned}$$

where we used $\mathcal{T} \models (E \mid C \sqcap D)[l_2,u_2]$ for the first and $\mathcal{T} \models (D \mid C)[l_1,u_1]$ for the second inequality.

Similarly, for the upper bound, we have

$$\begin{aligned}\mathrm{Vol}(E^{\mathcal{I}} \cap C^{\mathcal{I}}) &= \mathrm{Vol}(E^{\mathcal{I}} \cap D^{\mathcal{I}} \cap C^{\mathcal{I}}) + \\ &\quad \mathrm{Vol}(E^{\mathcal{I}} \cap (\neg D)^{\mathcal{I}} \cap C^{\mathcal{I}}) \\ &\leq u_2 \cdot \mathrm{Vol}(C^{\mathcal{I}} \cap D^{\mathcal{I}}) + \\ &\quad \big(\mathrm{Vol}(C^{\mathcal{I}}) - \mathrm{Vol}(C^{\mathcal{I}} \cap D^{\mathcal{I}})\big) \\ &\leq u_2 \cdot u_1 \cdot \mathrm{Vol}(C^{\mathcal{I}}) + \\ &\quad (\mathrm{Vol}(C^{\mathcal{I}}) - l_1 \cdot \mathrm{Vol}(C^{\mathcal{I}})) \\ &= (u_2 \cdot u_1 + 1 - l_1) \cdot \mathrm{Vol}(C^{\mathcal{I}}).\end{aligned}$$

Since $\mathrm{Vol}(E^{\mathcal{I}} \cap C^{\mathcal{I}}) \leq 1 \cdot \mathrm{Vol}(C^{\mathcal{I}})$, we can also assume that $u_3 \leq 1$. □

D ENCODING OF CLASSICAL AXIOMS (BOXEL)

The encoding of the four axiom types occurring in $\mathcal{EL}$ normal form in Xiong et al. [2022] is based on the *disjoint measurement* defined by

$$\mathrm{Disjoint}(B_1, B_2) = 1 - \frac{\mathrm{Vol}(B_1 \cap B_2)}{\mathrm{Vol}(B_1)}.$$

The measure is guaranteed to be between 0 and 1, is 0 whenever $B_1 \subseteq B_2$ and is 1 whenever $B_1 \cap B_2 = \emptyset$ [Xiong et al., 2022],

NF1: Atomic Subsumption. Axioms of the form $C \sqsubseteq D$ with $D \neq \bot$ are encoded by the loss term

$$\mathcal{L}_{C \sqsubseteq D}(w) = \mathrm{Disjoint}(\mathrm{Box}_w(C), \mathrm{Box}_w(D)). \tag{15}$$

If $D = \bot$ and C is not a nominal, that is, $C \sqsubseteq \bot$, the loss term is

$$\mathcal{L}_{C \sqsubseteq \bot}(w) = \max(0, M_w(C)_0 - m_w(C)_0 + \epsilon). \tag{16}$$

If C is a nominal, the axiom is inconsistent and this can be reported to the user.

NF2: Conjunctive Subsumption. Axioms of the form $C \sqcap D \sqsubseteq E$ with $E \neq \bot$ are encoded by the loss term

$$\mathcal{L}_{C \sqcap D \sqsubseteq E}(w) = \text{Disjoint}(\text{Box}_w(C) \cap \text{Box}_w(D), \text{Box}_w(E)). \tag{17}$$

For $E = \bot$, the loss term is

$$\mathcal{L}_{C \sqcap D \sqsubseteq \bot}(w) = \frac{Vol(\text{Box}_w(C) \cap \text{Box}_w(D))}{Vol(\text{Box}_w(C)) + Vol(\text{Box}_w(D))}. \tag{18}$$

NF3: Right Existential. Axioms of the form $C \sqsubseteq \exists r.D.$ are encoded by the loss term

$$\mathcal{L}_{C \sqsubseteq \exists r.D}(w) = \text{Disjoint}(T_w^r(\text{Box}_w(C)), \text{Box}_w(D)). \tag{19}$$

NF4: Left Existential. Axioms of the form $\exists r.C \sqsubseteq D$ are encoded by the loss term

$$\mathcal{L}_{\exists r.C \sqsubseteq D}(w) = \text{Disjoint}(T_w^{-r}(\text{Box}_w(C)), \text{Box}_w(D)), \tag{20}$$

As shown in Xiong et al. [2022], it holds for all encodings that $\mathcal{L}_\alpha(w) = 0$ implies that the corresponding axiom α is satisfied by the geometric interpretation $\mathcal{I}_w$ corresponding to the embedding.

E FURTHER EXPERIMENTAL DETAILS

E.1 SOFTPLUS VOLUME AND COOLING SCHEDULE

We minimize the loss function by gradient descent. One practical problem is that if the intersection of boxes is zero with respect to our current parameter vector (embedding) w, it will still be 0 in a small environment of w[2]. This means that the gradient will be 0, which can prevent gradient descent from bringing disjoint boxes closer together when necessary. This issue can be addressed by approximating the volume with the softplus volume [Patel et al., 2020]

$$\text{SVol}\left(\text{Box}_w\left(C\right)\right) = \prod_{i=1}^{d} \text{Softplus}_t\left(M_w\left(C\right)_i - m_w\left(C\right)_i\right), \tag{21}$$

where the *softplus function* Softplus_t is defined as $\text{softplus}_t(x) = t \log\left(1 + e^{x/t}\right)$ and t is called the *temperature parameter*. Because of the exponential growth of the exponential function, we have $1 + e^{x/t} \approx e^{x/t}$ when x/t is large, so that $\text{softplus}_t(x) \approx t \log\left(e^{x/t}\right) = x$. By letting t go to 0 as the search progresses (to let x/t become bigger), the softmax volume will converge to the real volume as the algorithm progresses. Our cooling schedule starts with a high temperature to learn fast at the beginning and decreases the temperature gradually to converge to a proper minimum. We guarantee $M_w(C) \geq m_w(C)$ for all concepts $C \in N_C$ by changing the parametrization. Instead of using $M_w(C)$, we use a vector $\delta_w(C)$ of the same dimension and let $M_w(C) = m_w(C) + \exp(\delta_w(C))$ (where the exponential function is applied element-wise).

E.2 DETAILS FOR SEL ONTOLOGY CONSTRUCTION

Why Choose YAGO as the Knowledge Base? YAGO3 is a knowledge base that combines the clean taxonomy of WordNet with the richness of the Wikipedia category system and assigns entities to more than 350,000 concepts. The accuracy of YAGO3 has been manually evaluated, with a confirmed accuracy of 95%. We chose YAGO3 rather than DBPedia [Lehmann et al., 2015] or Wikidata [Vrandečić and Krötzsch, 2014], due to their issues in differentiating concepts and entities, as well as their rather noisy concept assertions. Furthermore, YAGO3 provides complete concept assertions for each entity, such as $Person(ClaudeMonet)$, which is automatically added to the knowledge base if $Artist \sqsubseteq Person$. This allows us to avoid counting incorrect numbers of entities due to missing concept assertions and create more accurate probabilities in the dataset.

Concept & Role Name Selection: We created the conditionals by choosing subsets of concepts and roles from a particular domain. For the *Country* and *Person* dataset, we concentrate on concepts and roles belonging to ⟨*wordnet_country_108544813*⟩ and ⟨*wordnet_person_100007846*⟩, respectively. The *Hybrid* dataset is crafted by selecting a set of roles that bridge different

[2] If the distance between two boxes is d, the corresponding components in w have to be changed by at least d to let the boxes intersect. For changes "smaller than d", they will still be disjoint.

Algorithm 1 Pseudocode for the data generation

$\mathcal{L}_{concept} \leftarrow$ select concepts in certain domain.
$\mathcal{L}_{role} \leftarrow$ select roles correspondingly.

/* Generate PNF1 */
$\mathcal{D}_{PNF1} \leftarrow$ Empty List
for all $A \in \mathcal{L}_{concept}$ **do**
  for all $B \in \mathcal{L}_{concept}/A$ **do**
    $n_A \leftarrow$ count the number of entities belonging to A
    $n_{A \sqcap B} \leftarrow$ count the number of entities belonging to both A and B
    $p \leftarrow \frac{n_{A \sqcap B}}{n_A}$
    Append $(B|A)[p]$ to $\mathcal{D}_{PNF1}$

/* Generate PNF2 */
$\mathcal{D}_{PNF2} \leftarrow$ Empty List
for all $A_1 \in \mathcal{L}_{concept}$ **do**
  for all $A_2 \in \mathcal{L}_{concept}/A_1$ **do**
    for all $B \in \mathcal{L}_{concept}/\{A_1, A_2\}$ **do**
      $n_{A_1 \sqcap A_2} \leftarrow$ count the number of entities belonging to both A_1 and A_2
      $n_{A_1 \sqcap A_2 \sqcap B} \leftarrow$ count the number of entities belonging to all A_1, A_2 and B
      $p \leftarrow \frac{n_{A_1 \sqcap A_2 \sqcap B}}{n_{A_1 \sqcap A_2}}$
      Append $(B|A_1 \sqcap A_2)[p]$ to $\mathcal{D}_{PNF2}$

/* Generate PNF3 and PNF4*/
$\mathcal{D}_{PNF3} \leftarrow$ Empty List
$\mathcal{D}_{PNF4} \leftarrow$ Empty List
for all $r \in \mathcal{L}_{role}$ **do**
  for all $A \in \mathcal{L}_{concept}$ **do**
    for all $B \in \mathcal{L}_{concept}/A$ **do**
      $n_B \leftarrow$ count the number of entities belonging to B
      $n_{\exists r.A} \leftarrow$ count the number of entities related to other entities belonging to A through the role name r
      $n_{B \sqcap \exists r.A} \leftarrow$ count the number of entities belonging to B related to other entities belonging to A through the role name r
      $p_{PNF3} \leftarrow \frac{n_{B \sqcap \exists r.A}}{n_{\exists r.A}}$
      Append $(B|\exists r.A)[p_{PNF3}]$ to $\mathcal{D}_{PNF3}$
      $p_{PNF4} \leftarrow \frac{n_{B \sqcap \exists r.A}}{n_B}$
      Append $(\exists r.A|B)[p_{PNF4}]$ to $\mathcal{D}_{PNF4}$

domains and choosing the relevant top-level concepts from each domain. Specifically, roles are ranked by their fact count, with the top 18 selected. Concepts in the domain or range of these roles are then ranked by instance count, and the top 12 concepts are chosen.

Conditional Construction: We systematically traverse all conditionals of type PNF1 - PNF4 that can be generated from the selected concepts and roles and define their probabilities by statistical proportions. Algorithm 1 provides a detailed explanation of the process used to obtain probabilities for the conditionals.

E.3 PSEUDOCODE FOR THE PMP EVALUATION

We denote $(Q_2|Q_1)[p]$ as a probabilistic conditional in our entire axiom set $\mathcal{D}$, where Q_1 (body) and Q_2 (head) could be arbitrary $\mathcal{EL}$ concepts. We further define $\mathcal{D}' \subset \mathcal{D}$ as a set of conditionals whose body is the most general concept in knowledge base, i.e., $\langle wordnet\_person \rangle$ for "Person" knowledge base. We randomly select 30% of conditionals from $\mathcal{D}'$ as our query set $\mathcal{D}_{query}$. $\mathcal{D}_{model}$ refers to the set of axioms that are used to learn embeddings.

Following algorithm ensures that

- our knowledge base for learning embeddings does not contain $(Q_2|Q_1)[p]$.

Algorithm 2 Pseudocode for the PMP Evaluation

/* prepare knowledge base for learning embeddings and the query set */
$\mathcal{D}_{query} \leftarrow RandomSample(\mathcal{D}', 30\%)$
$\mathcal{D}_{model} \leftarrow Empty\ List$
for all $(Q_2|Q_1)[p] \in \mathcal{D}$ **do**
  if $D \notin \mathcal{D}_{query}$ **then**
    Append $(Q_2|Q_1)[p]$ to $\mathcal{D}_{model}$
/* Knowledge base modeling */
Learning the embeddings $\mathcal{M}$ with $\mathcal{D}_{model}$
/* PMP evaluation */
$BD \leftarrow 0, Hits \leftarrow 0, n_{eval} \leftarrow 0$
for all $(Q_2|Q_1)[p] \in \mathcal{D}$ **do**
  $n_{eval} \leftarrow n_{eval} + 1$
  if $(Q_2|Q_1)[p] \in \mathcal{D}_{query} \wedge (A|Q_1)[q1] \in \mathcal{D}_{model} \wedge (Q_2|A)[q2] \in \mathcal{D}_{model}$ **then**
    /* PMP evaluation */
    $Bound_{min} \leftarrow q1 * q2$
    $Bound_{max} \leftarrow \min(1, q1 * q2 + 1 - q2)$
    $\overline{p} \leftarrow \mathcal{M}((Q_2|Q_1))$
    if $\overline{p} < Bound_{min} \vee \overline{p} > Bound_{max}$ **then**
      $BD \leftarrow BD + \max(0, Bound_{min} - \overline{p}) + \max(0, \overline{p} - Bound_{max})$
    else
      $Hits \leftarrow Hits + 1$
$SE \leftarrow BD/n_{eval}$
$SA \leftarrow Hits/n_{eval} * 100\%$
return SE, SA

- Our knowledge base for learning embeddings contains the premisses required for PMP evaluation, i.e., $(A|Q_1)[q1]$ and $(Q_2|A)[q2]$.

		# Total	# Knowledge Base	# Query
Person	PNF1	16002	10962	3480
	PNF2	117600	75264	17856
	PNF3	96012	65772	9054
	PNF4	96012	67284	8316
Country	PNF1	870	638	176
	PNF2	24360	17864	3696
	PNF3	870	638	70
	PNF4	870	609	58
Hybrid	PNF1	156	120	30
	PNF2	1716	1320	270
	PNF3	2808	2160	194
	PNF4	2808	1968	192

Table 7: Statistics of the knowledge base/query set split for the PMP evaluation

E.4 DETAILS FOR KDE ESTIMATOR

Let $\mathbf{I_1}, \mathbf{I_2}, \ldots, \mathbf{I_m}$ denote the probability intervals of the conditionals in the learning set, where each interval $\mathbf{I_i} = [l_i, u_i]$. These intervals can be treated as samples from a 2-variate random variable with a probability density function f. To estimate

f, we use kernel density estimation, defined as

$$\bar{f}_{\mathbf{H}}(\mathbf{I}) = \frac{1}{m}\sum_{i=1}^{m} K_{\mathbf{H}}(\mathbf{I} - \mathbf{I_i}), \tag{22}$$

where $K_{\mathbf{H}}$ is the kernel function.

In our baseline, we employ Gaussian kernel to estimate the density:

$$K_{\mathbf{H}}(\mathbf{I} - \mathbf{I_i}) = \frac{1}{|\mathbf{H}|^{1/2}\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\mathbf{I} - \mathbf{I_i})^{\top}\mathbf{H}^{-1}(\mathbf{I} - \mathbf{I_i})\right), \tag{23}$$

where $\mathbf{H}$ is the bandwidth, a symmetric and positive definite $m \times m$ matrix that balances smoothness and details in the density estimate. To select the bandwidth, we apply Scott's rule [Scott, 2015]. For a 2-variate density estimate, the bandwidth matrix is given by:

$$\mathbf{H} = diag(\sigma_1, \sigma_2) \cdot m^{-\frac{1}{6}}, \tag{24}$$

where σ_1, σ_2 are the standard deviation of l_i and u_i, respectively.

We then generate predicted intervals $[\bar{l}, \bar{u}]$ for KDE estimator by sampling from the estimated density function $\bar{f}_{\mathbf{H}}(\mathbf{I})$.

F ADDITIONAL EXPERIMENTAL RESULTS

F.1 HYPERPARAMETER TUNING

We tune the main hyperparameters separately for each ontology and report the selected values in Table 8. To assess sensitivity, we vary one hyperparameter at a time around the selected configuration and report the mean and standard deviation of the resulting performance in Table 9. Overall, the results indicate that performance is relatively stable across the tested ranges.

	ED	EP	BS	LR	BR
Person	32	20	512	0.01	[0,10]
Hybrid	8	20	256	0.01	[0,1]
Country	16	30	256	0.05	[0,10]

Table 8: Best hyperparameters used for different ontologies: embedding dimension (ED), number of epochs (EP), batch size (BS), learning rate (LR), bound for regularization (BR).

Hyperparameter	**Search Range**	**Person**	**Country**	**Hybrid**
Embedding dim n	$\{8, 16, 32, 64, 128\}$	0.088 ± 0.052	0.042 ± 0.032	0.065 ± 0.025
Learning rate lr	$\{0.001, 0.01, 0.05, 0.1\}$	0.080 ± 0.025	0.048 ± 0.021	0.075 ± 0.027
Side length β	$\{1, 10\}$	0.120 ± 0.070	0.030 ± 0.010	0.065 ± 0.025

Table 9: Hyperparameter search ranges and corresponding results across datasets.

F.2 FINE-GRAINED APPROXIMATE GAP EVALUATION

Datasets	Approximation Gap				
	Total	PNF1	PNF2	PNF3	PNF4
PERSON	0.23	0.22	0.21	0.30	0.22
COUNTRY	0.25	0.24	0.24	0.28	0.24
HYBRID	0.29	0.29	0.27	0.33	0.27

Table 10: Approximation gap evaluated on 60 embeddings.

G ABLATION STUDIES

Dataset	Concept Representation	SA ↑	SE ↓	AG ↓
Person	n-balls	85.7%	0.022	0.324
	boxes*	**89.3%**	**0.013**	**0.233**
Country	n-balls	90.3%	0.018	0.359
	boxes*	**95.6%**	**0.004**	**0.241**
Hybrid	n-balls	82.9%	0.052	0.344
	boxes*	**89.7%**	**0.020**	**0.293**

Table 11: Ablation study of the concept representation. The design choice of our method is marked with an asterisk (*).

Dataset	Regularization Term	SA ↑	SE ↓	AG ↓
Person	w/o	45.1%	0.220	**0.202**
	$\mathcal{L}_{loc}$	85.7%	0.150	0.276
	$\mathcal{L}_{vol}$	87.3%	0.015	0.235
	$(\mathcal{L}_{loc} + \mathcal{L}_{vol})^*$	**89.3%**	**0.013**	0.233
Country	w/o	82.1%	0.090	0.243
	$\mathcal{L}_{loc}$	90.2%	0.110	0.255
	$\mathcal{L}_{vol}$	95.1%	0.080	0.253
	$(\mathcal{L}_{loc} + \mathcal{L}_{vol})^*$	**95.60%**	**0.004**	**0.241**
Hybrid	w/o	86.2%	0.035	0.297
	$\mathcal{L}_{loc}$	87.3%	0.018	0.319
	$\mathcal{L}_{vol}$	88.2%	**0.002**	0.301
	$(\mathcal{L}_{loc} + \mathcal{L}_{vol})^*$	**89.70%**	**0.002**	**0.293**

Table 12: Ablation study of the regularization terms. The design choice of our method is marked with an asterisk (*).