RETHINKING IDENTIFICATION AND SEMANTIC MEANINGS: DIFFERENCE-BASED KNOWLEDGE GRAPH EMBEDDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge graph embedding (KGE) encodes entities and relations as low-dimensional representations and relies on model architectures and scoring functions to capture link patterns, such as symmetry and inversion. However, existing methods directly use the embedding of each entity or relation for identification and as its semantic meaning. We challenge this from a difference-based perspective, which emphasize that semantic meaning is not intrinsic or self-identical, but instead arises from systematic differences among representations. Motivated by this, we propose a simple yet effective representation-level transformation that redefines the semantic representation of an entity or relation as its embedding differences with respect to all other entities or relations. This transformation can be seamlessly applied as a plug-in to a wide range of state-of-the-art KGE models, preserving overall computational efficiency while consistently improving link prediction performance across multiple benchmark datasets.

1 INTRODUCTION

Knowledge graphs (KGs) represent structured knowledge as triples (Li et al., 2022; Galkin et al., 2024), where each factual triple (h, r, t) indicates that there is a relation r between a head entity h and a tail entity t. Knowledge graph embedding (KGE) methods encode entities and relations as low-dimensional representations and rely on specific designs and scoring functions to capture structured knowledge and link patterns, such as symmetry, inversion, and composition (Bordes et al., 2013; Li et al., 2022). KGE supports a wide range of downstream tasks (Yasunaga et al., 2022), including link prediction (Gao et al., 2020; Sun et al., 2019), knowledge completion (Yang et al., 2015; Zhang et al., 2019), reasoning (Luo et al., 2025; Zhu et al., 2023), and question answering (Zhuo et al., 2025a;b).

In recent years, numerous KGE methods have been proposed, primarily differing in their model designs on the geometry of embedding spaces (Li et al., 2025; Chami et al., 2020; Do Carmo, 1992; Eisenhart, 1997; Ficken, 1939) and scoring functions (Ge et al., 2023; Cao et al., 2021; Dettmers et al., 2018; Balazevic et al., 2019; Zhang et al., 2022). Despite such model diversity, all existing methods implicitly share a common, but unmentioned, representational assumption: each embedding serves as the self-identification for an entity or a relation. In other words, the embedding of each entity or relation is directly treated as having its semantic meaning, which is tied to the absolute coordinate in the embedding space. The Figure 1 (b) illustrates this dominant paradigm, where each embedding is directly used as the representation with semantic meaning of a corresponding entity.

However, this assumption is conceptually restrictive. In a representation system, the semantic meaning of an identity is inherently based on differences: it is determined by how it differs from and interacts with all other identities, rather than by an absolute coordinate in the system. This intuition is consistent modern views of meaning, which emphasize that semantic meaning is not intrinsic or self-identical, but instead arises from systematic differences among representations. From this perspective, absolute coordinate-based embeddings fail to explicitly represent semantic meaning and contrast. In this work, we adopt this insight as a representational principle for knowledge graph embedding.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

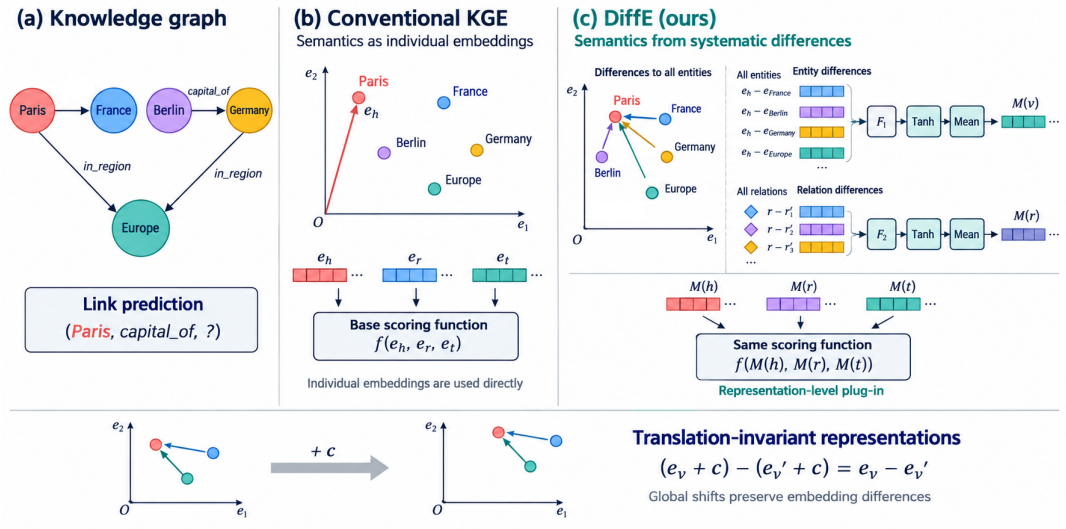

Figure 1: Overview of the proposed difference-based principle. **(a)** An example of the knowledge graph.**(b)** Existing KGE methods treat the self-identical embedding of an entity or a relation as a semantic representation for its meaning. **(c)** The proposed transformation operation, which redefines semantic representation of an entity or a relation by the embedding differences with respect to other entities or relations, and the semantic representation is then used in scoring functions.

Motivated by this difference-based perspective, we propose a plug-and-play and representation-level transformation operation for a wide range of existing KGE methods. Instead of identifying semantic meaning with a single, absolute coordinate-based embedding, we explicitly redefine the semantic representation of an entity or relation through the set of its embedding differences with respect to all other entities or relations, as illustrated in Figure 1 (c). Specifically, in the experiments, we apply a simple mean aggregation to such set of embedding differences into a compact embedding that preserves dimensionality while explicitly representing semantic meaning for each entity and relation.

Importantly, our proposed transformation operates purely at the entity and relation representation/embedding level. It preserves the embedding dimensionality of original KGE models and can be applied as a plug-and-play enhancement to existing methods without modifying their model architectures, scoring functions or training objectives, *et al.* . In addition, the overall computational complexity remains comparable to that of each original KGE model.

We empirically evaluate the proposed difference-based transformation operation across a wide range of state-of-the-art (SOTA) KGE baselines, including translation (Bordes et al., 2013), bilinear (Yang et al., 2015), rotation (Sun et al., 2019), neural network (Dettmers et al., 2018), Euclidean (Gao et al., 2020), elliptic (Li et al., 2022), and/or hyperbolic (Chami et al., 2020) geometry-based models. Across multiple benchmark datasets (Dettmers et al., 2018; Mahdisoltani et al., 2015; Toutanova & Chen, 2015) and different experimental setups (Li et al., 2025), such as fixed embedding dimensions or tuning the best embedding dimension with the validation dataset, applying the proposed transformation consistently improves link prediction performance. These results indicate that explicitly encoding systematic differences for KGE provides more informative semantic representations than absolute coordinate-based embeddings alone. In summary, our contributions are as follows.

- To the best of our knowledge, this is the first work that revisits the unmentioned representational assumption in existing knowledge graph embedding and proposes a difference-based perspective to represent semantic meaning.
- We propose a simple, plug-and-play, and representation-level transformation based on embedding differences for a wide range of KGE methods.
- Experiments on multiple benchmark datasets and experimental setups demonstrate consistent improvements in link prediction performance across SOTA KGE baselines.

2 PRELIMINARIES

2.1 KNOWLEDGE GRAPHS

A knowledge graph is a directed, edge-labeled graph that represents knowledge as a set of factual triples

$$\mathcal{G} = \{(h, r, t) \mid h, t \in \mathcal{V},\ r \in \mathcal{R}\} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V},$$

where $\mathcal{V}$ denotes the set of entities and $\mathcal{R}$ denotes the set of relations. Each triple (h, r, t) indicates that a relation r connects a head entity h and a tail entity t. Knowledge graphs are typically incomplete, motivating KGE and evaluation with the task of link prediction for completion (Bordes et al., 2013; Li et al., 2025; Sun et al., 2019).

2.2 KNOWLEDGE GRAPH EMBEDDING

Knowledge graph embedding methods encode each entity and relation as a low-dimensional representation in the embedding spaces. Formally, a KGE method learns an embedding function

$$\phi : \mathcal{V} \cup \mathcal{R} \to \mathbb{R}^{d_\mathcal{V}} \cup \mathbb{R}^{d_\mathcal{R}},$$

which maps the entity $v \in \mathcal{V}$ and the relation $r \in \mathcal{R}$ to low-dimensional embeddings $e_v = \phi(v) \in \mathbb{R}^{d_\mathcal{V}}$ and $e_r = \phi(r) \in \mathbb{R}^{d_\mathcal{R}}$, respectively. The dimension of the relations is usually dependent on the dimension of the entities (Li et al., 2025). For example, $d_\mathcal{R} = d_\mathcal{V}$ in TransE (Bordes et al., 2013). In the following paper, we denote $d = d_\mathcal{V}$ as the embedding dimension. KGE aims to learn representations that preserve expressive knowledge with scoring functions and training objectives, enabling efficient and effective downstream tasks, such as link prediction, in embedding spaces.

Scoring functions. Based on the embeddings of entities and relations, KGE models design and learn a scoring function

$$f : \mathbb{R}^{d_\mathcal{V}} \times \mathbb{R}^{d_\mathcal{R}} \times \mathbb{R}^{d_\mathcal{V}} \to \mathbb{R},$$

which calculates a real-valued score $f(e_h, e_r, e_t)$ for each triple (h, r, t). Higher scores indicate that a triple is more likely to be factual. The specific form of the scoring function $f(\cdot)$ depends on the designs. For example, translation and rotation-based methods, such as TransE (Bordes et al., 2013) and GoldE (Li et al., 2025), use geometry-based (Chami et al., 2020) similarity scoring functions, and neural network-based methods, such as ConvE (Dettmers et al., 2018), use Convolutional network-based scoring functions.

Training objectives. KGE models are typically trained based on a ranking loss or logistic loss that encourages factual triples to remain higher scores than corrupted (negative) triples (Sun et al., 2019):

$$\mathcal{L} = \sum_{(h,r,t)\in\mathcal{G},(h',r,t')\notin\mathcal{G}} \left[\gamma - f(e_h, e_r, e_t) + f(e_{h'}, e_r, e_{t'})\right]_+,$$

where $\gamma > 0$ is a margin hyperparameter and $[x]_+ = \max(0, x)$. Similarly, objectives based on regulations, softmax, and self-adversarial losses are also proposed (Li et al., 2022; 2025).

Despite many KGE methods with diversity in their model designs on the geometry, scoring functions, and training objectives, all methods implicitly share a common representational assumption: each embedding e_v or e_r is directly treated as representing the semantic meaning of the entity v or the relation r, and then embeddings e_v and e_r are directly used in the scoring function.

3 METHODOLOGY

In this section, we introduce the difference-based semantic representation framework and the transformation operations that are applicable to a wide range of existing KGE models.

3.1 OVERVIEW

Given a base KGE model and its scoring function $f(\cdot, \cdot, \cdot)$, existing methods directly use absolute coordinate-based embeddings as semantic representations to calculate $f(e_h, e_r, e_t)$.

In contrast, our approach replaces such original self-identical embeddings with difference-based semantic representations that are derived from systematic embedding differences. The resulting semantic representations are then used as inputs to the original scoring function. This transformation is performed entirely at the representation level and does not alter the functional form of $f(\cdot, \cdot, \cdot)$.

3.2 DIFFERENCE SETS

For each entity or relation $x \in \mathcal{V} \cup \mathcal{R}$, we first construct its difference set $\mathcal{D}(x)$, which encodes how x differs from all other identities in the corresponding embedding space.

We define differences within each symbol class (entities or relations) using the difference operator $\Delta(x, y) = x - y$ in the corresponding embedding space.

Definition 3.1 (Entity / Relation difference sets)**.** For any entity $v \in \mathcal{V}$ and relation $r \in \mathcal{R}$ we define

$$\mathcal{D}_{\mathcal{V}}(v) = \{\, \Delta(e_v, e_{v'}) = e_v - e_{v'} \mid v' \in \mathcal{V} \,\},\ \mathcal{D}_{\mathcal{R}}(r) = \{\, \Delta(e_r, e_{r'}) = e_r - e_{r'} \mid r' \in \mathcal{R} \,\}. \quad (1)$$

Intuitively, the difference set $\mathcal{D}_{\mathcal{V}}(v)$ of an entity v encodes how its self-identical embedding differs from all other entities in the entity embedding space, and similarly for relation difference sets.

3.3 AGGREGATION

Based on each difference set and its multiple difference embeddings, we apply a permutation-invariant aggregation operator to obtain a compact semantic representation suitable for scoring functions. Concretely, Let F be the one layer feed-forward network, Tanh be the activation function, and $mean$ be the element-wise mean aggregation that maps multiple difference-based embeddings to a representation in the semantic meaning space. Thus, for any entity $v \in \mathcal{V}$ and relation $r \in \mathcal{R}$, we get representations with semantic meaning based on the difference sets by:

$$\mathcal{M}(v) \;=\; mean\big[\text{Tanh}\big(F_1(\mathcal{D}_{\mathcal{V}}(v))\big)\big] \in \mathbb{R}^{d_{\mathcal{V}}},\ \mathcal{M}(r) \;=\; mean\big[\text{Tanh}\big(F_2(\mathcal{D}_{\mathcal{R}}(r))\big)\big] \in \mathbb{R}^{d_{\mathcal{R}}}. \quad (2)$$

We redefine $\mathcal{M}(v)$ and $\mathcal{M}(r)$ as the *difference-based semantic representations* of each entity $v \in \mathcal{V}$ and relation $r \in \mathcal{R}$ that are then used in the scoring function, instead of directly using self-identical embeddings e_v and e_r. This permutation-invariant aggregation operator preserves embedding dimensionality while explicitly encoding difference-based semantic meanings.

3.4 PLUG-AND-PLAY

Given a KGE baseline and its formulation of the scoring function $f(\cdot, \cdot, \cdot)$, we replace their self-identical embeddings by corresponding difference-based semantic representations and get the enhanced scoring function:

$$\tilde{f}(e_h, e_r, e_t) \;=\; f\big(\mathcal{M}(h),\ \mathcal{M}(r),\ \mathcal{M}(t)\big). \quad (3)$$

As all of our proposed transformation operations are differentiable, the gradients propagate back to self-identical embeddings e_v and e_r of the KGE baseline during training.

Crucially, this formulation allows the proposed difference-based principle to be applied to a wide range of state-of-the-art KGE models, including translation (Bordes et al., 2013), bilinear (Yang et al., 2015), rotation (Sun et al., 2019), neural network (Dettmers et al., 2018), Euclidean (Gao et al., 2020), elliptic (Li et al., 2022), and/or hyperbolic (Chami et al., 2020) geometry-based models.

Remarks. The model architecture, negative-sampling strategies, training objectives, regulations, normalizations, and any other optimization procedures of each KGE baseline remain unchanged. For example, TransE (Bordes et al., 2013) uses the l_2 normalization on the entity embeddings. RotatE (Sun et al., 2019) uses the self-adversarial losses and regulations. GoldE (Li et al., 2025) proposes the universal orthogonal parameterization and geometry metric for the embeddings in the Euclidean space to calculate the Elliptic and Hyperbolic based distance/similarity.

3.5 COMPLEXITY

Our transformation introduces very few additional parameters with $O({d_{\mathcal{V}}}^2 + {d_{\mathcal{R}}}^2)$, and the overall space complexity is of the same order as that of each base KGE model with $O(|\mathcal{V}|d_{\mathcal{V}} + |\mathcal{R}|d_{\mathcal{R}})$ where $|\mathcal{V}| \gg (d_{\mathcal{V}} + d_{\mathcal{R}})$. The additional time complexity for the training procedure is $O(|\mathcal{V}|^2 d_{\mathcal{V}} + |\mathcal{R}|^2 d_{\mathcal{R}})$. The training time complexity of existing KGE methods is at least $O(|\mathcal{G}|\mathcal{N}_{neg} d_{\mathcal{V}})$, where $\mathcal{N}_{neg}$ is the number of negative samples that is proportional to $|\mathcal{V}|$ and $|\mathcal{G}|\mathcal{N}_{neg} \sim (|\mathcal{V}|^2 + |\mathcal{R}|^2)$. Thus, our overall training time complexity is comparable to existing KGE models. Once the model is optimized, our testing time complexity is the same as each base KGE model as difference-based representations are obtained once after training. More experimental runtime analyses are also reported in Section 6.5.

4 DIFFERENCE-BASED PROPERTIES

In this section, we prove the formal properties of the proposed difference-based principle and compare with existing KGE methods.

4.1 INVARIANCE PROPERTY

Theorem 4.1 (The entity embedding space is invariant to any global translation $c \in \mathbb{R}^d$ given any full-column-rank feed-forward layer with an element-wise strictly increasing activation function Tanh)**.** *For any two embedding functions ϕ and ϕ', we have*

$$\mathcal{M}_{\phi}(x) = \mathcal{M}_{\phi'}(x), \ \ \forall x \in \mathcal{V} \tag{4}$$

if there exists a constant vector $c \in \mathbb{R}^d$ such that

$$\phi'(x) = \phi(x) + c, \ \ \forall x \in \mathcal{V}. \tag{5}$$

The argument for relations is embedding space identical. The proposed difference-based principle is invariant to global translations in the entity and relation embedding spaces.

Proof. If $\mathcal{M}_{\phi}(x) = \mathcal{M}_{\phi'}(x)$ for all x, then the corresponding difference sets are identical. Fix any x and $y \neq x$. We have

$$\phi(x) - \phi(y) = \phi'(x) - \phi'(y), \tag{6}$$

which implies $\phi'(x) - \phi(x) = \phi'(y) - \phi(y)$. Thus, there exists a constant vector $c = \phi'(y) - \phi(y)$ such that $\phi'(x) = \phi(x) + c$ for all x. □

4.2 BASE KGE METHODS AND COMPARISON

Existing KGE methods do not explicitly satisfy invariance properties. In particular, based on specific scoring function designs, some KGE models partially satisfy invariance properties for the entity embedding spaces, which are shown below and summarized in Table 1.

Translation-based KGE. For translation-based (Ji et al., 2015; Wang et al., 2014) methods such as TransE (Bordes et al., 2013), the scoring function $f(e_h, e_r, e_t) = \|e_h + e_r - e_t\|_2$ is invariant to a global translation applied to entity embeddings, i.e., $\phi'(v) = \phi(v) + c$ for all $v \in \mathcal{V}$. However, this invariance does not hold in the relation embedding space. As a result, TransE only exhibits a partial invariance property.

Rotation-based KGE. Rotation-based (Gao et al., 2020) methods, such as RotatE (Sun et al., 2019) and GoldE (Li et al., 2025), use relations as rotations in their scoring functions to capture more relation logical patterns and topological structure patterns. However, they are not invariant to global translations in the entity and relation embedding spaces. Consequently, rotation-based models exhibit no invariance properties.

Overall, existing KGE models do not enforce invariance properties. In contrast, the proposed difference-based representations explicitly yield the invariance property to translation for both entity and relation embedding spaces. And derived from Theorem 4.1 and keeping the same scoring function of each base KGE model, the proposed difference-based representations have the same relation logical patterns capability (Li et al., 2022) and the same topological structures capability (Chami et al., 2020), which are also summarized in Table 1.

5 IMPROVED GENERALIZATION ABILITY

In this section, we show that, under plausible translation invariance assumptions, which are commonly observed in image data (Deng et al., 2022) and graph data (Sannai et al., 2021; Bordes et al., 2013), our difference-based representations improve generalization ability for each base KGE model.

5.1 SETUPS

Let $\mathcal{S} = (\mathcal{Z}, \mathcal{Y})$ denote the space of training samples (e.g., triples with positive and negative labels), and let $\mathcal{D}$ be the data distribution (or equivalently the conditional distribution of labels given any triple (h, r, t)). Equivalently, the conditional distribution of labels or scoring values depends only on

Table 1: Comparison of invariance properties, relation logical patterns capability (Li et al., 2022) and topological structures capability (Chami et al., 2020) for representative base KGE models. "Invariance" refers to "invariance to any global translations $\phi'(x) = \phi(x) + c$". $\mathbb{E}$, $\mathbb{P}$ and $\mathbb{Q}$ refers to Euclidean, elliptic and hyperbolic geometries (Do Carmo, 1992), respectively. Details of "Symmetry, Antisymmetry, Inversion, Composition, Cycle and Hierarchy" are discussed in GoldE (Li et al., 2025). "$*$" refers to "the same as each base KGE model". Baseline KGE models typically do not hold invariance properties for entity or relation embedding spaces, whereas our difference-based representations ensure invariance and hold the other base properties.

	Invariance		Relation Logical Patterns Capability				Topological	
KGE Method	Entity	Relation	Symmetry	Antisymmetry	Inversion	Composition	Cycle	Hierarchy
TransE: $\|e_h + e_r - e_t\|_2$	✓	✗	✗	✓	✓	✓	✓	✗
DistMult: $\langle e_h, \mathrm{diag}(e_r), e_t\rangle$	✗	✗	✓	✗	✗	✗	✗	✗
RotatE: $\|e_h \circ e_r - e_t\|_2$	✗	✗	✓	✓	✓	✓	✓	✗
HousE: $\|e_h \circ e_r - e_t\|_{\mathbb{E}}$	✗	✗	✓	✓	✓	✓	✓	✗
GoldE: $\|e_h \circ e_r - e_t\|_{\mathbb{P},\mathbb{Q}}$	✗	✗	✓	✓	✓	✓	✓	✓
Ours: Difference-based	✓	✓	$*$	$*$	$*$	$*$	$*$	$*$

relative configurations of embeddings and is unchanged by adding a common offset to all embeddings. A KGE model induces a hypothesis class $\mathcal{H}$, where each $f_\phi \in \mathcal{H} : \mathcal{Z} \to \mathbb{R}$, is parameterized by a embedding function ϕ and a scoring function f directly applied on embeddings. Learning proceeds via empirical risk minimization (ERM) with training objective ℓ:

$$\hat{f}_\phi = \arg\min_{f_\phi \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f_\phi(z_i), y_i), \quad (z_i, y_i) \sim \mathcal{S}. \tag{7}$$

Assumption 5.1 (Translation invariance of the data representations)**.** Let a global translation $g \in G$, where $G \subseteq (\mathbb{R}^d, +)$, acts on embedding functions by

$$(g \cdot \phi)(x) = \phi(x) + g. \tag{8}$$

The data representations are invariant under the translation action group of G, i.e., for all $(z_i, y_i) \sim \mathcal{D}$ and all $g \in G$,

$$f_\phi(z_i) = f_{g\cdot\phi}(z_i). \tag{9}$$

This data representations invariance assumption is plausible in image data (Deng et al., 2022; Zhu et al., 2021) and graph data (Sannai et al., 2021; Ma et al., 2024; Bordes et al., 2013).

5.2 Improved Generalization Ability

Let $\mathcal{H}$ denote the hypothesis class induced by a base KGE model directly using the original embeddings. And let $\mathcal{H}_{\mathrm{inv}}$ denote the hypothesis class induced by our difference-based transformation operation.

Proposition 5.2. *$\mathcal{H}_{\mathrm{inv}}$ is invariant to any global translation action of G, whereas $\mathcal{H}$ is not, in general.*

Proof. From Theorem 4.1, $\mathcal{H}_{\mathrm{inv}}$ is invariant. From Section 4.2, $\mathcal{H}$ is not invariant, in general. □

For a base KGE model with f_ϕ, let $R(f_\phi)$ be its expected loss and $R_{|\mathcal{S}|}(f_\phi)$ be its empirical loss with $|\mathcal{S}|$ training samples. For its hypothesis class $\mathcal{H}$, the *generalization gap* or *generalization error* is given by (Sannai et al., 2021):

$$\mathscr{G}_{\mathcal{S}}(\mathcal{H}) := \sup_{f_\phi \in \mathcal{H}} |R(f_\phi) - R_{|\mathcal{S}|}(f_\phi)|.$$

Similarly, the generalization gap for the corresponding $\mathcal{H}_{\mathrm{inv}}$ induced by our difference-based transformation operation is $\mathscr{G}_{\mathcal{S}}(\mathcal{H}_{\mathrm{inv}})$.

Covering number and *Rademacher complexity* (Mohri et al., 2018) are two commonly used complexity measurements for model classes that can provide uniform generalization bounds. The covering number can also be directly used to upper bound the Rademacher complexity via Dudley's entropy integral theorem (Dudley, 1967; Mohri et al., 2018). Notation and more details are introduced in Appendix A.

Lemma 5.3 (Bounding our generalization gap with empirical Rademacher complexity $\mathscr{R}_{\mathcal{S}}(\mathcal{H})$ via the covering number N over $\mathcal{H}$ evaluated at the sample set $\mathcal{S}$ (Mohri et al., 2018)). *For any $\alpha > 0$,*

$$\mathscr{R}_{\mathcal{S}}(\mathcal{H}_{\text{inv}}) \leq 4\alpha + 12 \int_{\alpha}^{\infty} \sqrt{\frac{\log N\big(\tau, \mathcal{H}_{\text{inv}}, L_2(\mathbb{P}_{\mathcal{S}})\big)}{|\mathcal{S}|}} d\tau.$$

Following the prior work (Zhu et al., 2021), ERM over the invariant hypothesis class $\mathcal{H}_{\text{inv}}$ admits tighter generalization bounds than ERM over the baseline class $\mathcal{H}$, and for any $\epsilon > 0$,

$$N\big(\tau, \mathcal{H}_{\text{inv}}, L_2(\mathbb{P}_{\mathcal{S}})\big) \leq \inf_{\hat{\mathcal{S}}_{G,\epsilon}} N\big(\tau', \mathcal{H}, L_2(\mathbb{P}_{\hat{\mathcal{S}}_{G,\epsilon}})\big),$$

where $\tau' = \tau - \epsilon\sqrt{1 - \frac{|\hat{\mathcal{S}}_{G,\epsilon}|}{|\mathcal{S}|}}$, and by searching ϵ under the invariant assumption of G can tighten the generalization gap. More details and proofs are introduced in Appendix A.

6 Experiments

To comprehensively validate the effectiveness of the proposed difference-based representations (DiffE), we conduct experiments on link prediction task following existing KGE methods (Li et al., 2025; Bordes et al., 2013; Li et al., 2022):

(1) How does DiffE perform as a plugin compared to state-of-the-art KGE models? (Section 6.2)

(2) Can DiffE still improve base KGE models when the embedding dimensions vary? (Section 6.3)

(3) How much have difference-based entity and/or relation representations improved? (Section 6.4)

(4) Is DiffE as efficient as each base KGE model? (Section 6.5)

6.1 Experimental Setups

Datasets: Following existing works (Li et al., 2025; Sun et al., 2019; Li et al., 2022), we evaluated on three standard benchmarks: WN18RR (Dettmers et al., 2018), FB15k-237 (Toutanova & Chen, 2015), and YAGO3-10 (Mahdisoltani et al., 2015). Appendix B provides more details.

Baselines: We compare a wide range of advanced KGE baselines, including TransE (translation/Euclidean) (Bordes et al., 2013), DistMult (bilinear/Euclidean) (Yang et al., 2015), ComplEx (bilinear) (Trouillon et al., 2016), ConvE (Neural/Euclidean) (Dettmers et al., 2018), MuRP (hyperbolic) (Balazevic et al., 2019), RotatE (rotation/2D) (Sun et al., 2019), CompoundE (rotation/2D) (Ge et al., 2023), Rotate3D (rotation/3D) (Gao et al., 2020), QuatE (rotation/4D) (Zhang et al., 2019), DualE (rotation/3D) (Cao et al., 2021), ReflectE (rotation) (Zhang et al., 2022), RefH, RotH, AttH (rotation/hyperbolic/2D) (Chami et al., 2020), HousE (rotation/kD) (Li et al., 2022), GoldE (rotation/hyperbolic/Euclidean/kD) (Li et al., 2025), and KGE-CCP (Comformal) (Zhu et al., 2026).

Implementation details: For direct and fair comparisons (Sun et al., 2019; Li et al., 2025; 2022), we also fix the embedding dimension for main experiments. We use official open-source implementations for each base KGE model, and apply our difference-based representations with any other model architectures and training/tuning procedures unchanged. Our experimental protocol ensures that any observed improvements come from our difference-based representations rather than training/tuning discrepancies. We report mean rank (MR), mean reciprocal rank (MRR) and HITS at N (H@N) metrics for the link prediction performance. More details are provided in Appendix C.

6.2 Main Comparison

Table 2 reports the link prediction performance of DiffE when applied as a plug-in and shows comparison with the state-of-the-art KGE models. We focus on the relative and absolute performance gains introduced by our DiffE. We also report the statistical significance in Appendix D.

Performance gains. Across all evaluated datasets and all metrics, DiffE consistently improves a wide range of base KGE models, including translation, bilinear, rotation, Euclidean, elliptic, and hyperbolic geometry-based models, in terms of MRR and H@N. In particular, DiffE relatively improves MRR 5.5% on TransE (FB15k-237), 16% on DistMult (FB15k-237), 3% on RotatE

Table 2: Results improved by DiffE are in **bold** and the best results from base KGE models are underlined. [†]: Results are taken from (Nguyen et al., 2018); [⋄]: Results are taken from (Dettmers et al., 2018). All other results are taken from their original papers, respectively.

Model	WN18RR					FB15k-237					YAGO3-10				
	MR	MRR	H@1	H@3	H@10	MR	MRR	H@1	H@3	H@10	MR	MRR	H@1	H@3	H@10
ComplEx⋄	5261	.440	.410	.460	.510	339	.247	.158	.275	.428	6351	.360	.260	.400	.550
ConvE⋄	4187	.430	.400	.440	.520	224	.325	.237	.356	.501	1671	.440	.350	.490	.620
MuRP	-	.481	.440	.495	.566	-	.335	.243	.367	.518	-	.354	.249	.400	.567
CompoundE	-	.491	.450	.508	.576	-	.357	.264	.393	.545	-	-	-	-	-
Rotate3D	3328	.489	.442	.505	.579	165	.347	.250	.385	.543	-	-	-	-	-
QuatE	3472	.481	.436	.500	.564	176	.311	.221	.342	.495	-	-	-	-	-
DualE	-	.482	.440	.500	.561	-	.330	.237	.363	.518	-	-	-	-	-
ReflectE	1730	.488	.450	.501	.559	148	.358	.263	.396	.546	-	-	-	-	-
RefH	-	.461	.404	.485	.568	-	.346	.252	.383	.536	-	.576	.502	.619	.711
RotH	-	.496	.449	.514	.586	-	.344	.246	.380	.535	-	.570	.495	.612	.706
AttH	-	.486	.443	.499	.573	-	.348	.252	.384	.540	-	.568	.493	.612	.702
KGE-CCP	1849	.484	-	-	-	168	.338	-	-	-	-	-	-	-	-
TransE†	3384	.226	-	-	.501	357	.294	-	-	.465	-	-	-	-	-
TransE+DiffE	**2987**	**.231**	-	-	**.528**	**280**	**.310**	-	-	**.488**	-	-	-	-	-
DistMult⋄	5110	.430	.390	.440	.490	254	.241	.155	.263	.419	5926	.340	.240	.380	.540
DistMult+DiffE	**3025**	**.441**	**.398**	**.455**	**.525**	**244**	**.281**	**.201**	**.313**	**.457**	**5641**	**.347**	**.248**	**.385**	**.551**
RotatE	3340	.476	.428	.492	.571	177	.338	.241	.375	.533	1767	.495	.402	.550	.670
RotatE+DiffE	**3209**	**.483**	**.438**	**.498**	**.582**	**170**	**.347**	**.250**	**.386**	**.545**	**1540**	**.510**	**.409**	**.561**	**.687**
HousE	1303	.511	.465	.528	.602	153	.361	.266	.399	.551	1415	.571	.491	.620	.714
HousE+DiffE	**973**	**.521**	**.473**	**.536**	**.614**	**144**	**.370**	**.275**	**.405**	**.560**	**1218**	**.582**	**.500**	**.630**	**.724**
GoldE	790	.525	.476	.542	.615	142	.370	.277	.405	.558	536	.588	.510	.634	.723
GoldE+DiffE	**763**	**.537**	**.484**	**.550**	**.627**	**139**	**.380**	**.286**	**.412**	**.569**	**528**	**.601**	**.517**	**.646**	**.734**

(YAGO3-10), 2.5% on HousE (FB15k-237), and 2.7% on the best baseline GoldE (FB15k-237). DiffE also absolutely improves H@10 more than +1% across almost all datasets. These gains are non-trivial given that each improvement from previous works is also around +1%, and our compared baselines already represent strong state-of-the-art results.

Improvements for strong baselines. Notably, DiffE yields consistent improvements even when applied to highly expressive models such as HousE and GoldE. For example, on WN18RR, GoldE already achieves an MRR of 0.525, while GoldE+DiffE further improves it to 0.537. On the largest dataset YAGO3-10, GoldE+DiffE improves MRR from 0.370 to 0.380. This indicates that DiffE provides representational benefits with the invariant property and generalization ability.

Overall, the quantitative results demonstrate that DiffE serves as a general and effective enhancement to existing KGE models. The consistent gains across model families and datasets support the claim that explicitly modeling embedding differences as entity and relation senmatic representations leads to improved generalization for knowledge graph embedding.

6.3 Varying Embedding Dimensions

Table 7 in the Appendix shows the experimental results with embedding dimension restricted to 32. Such low-dimensional setting follows Chami et al. (2020); Li et al. (2025). We can observe that our DiffE plugin still improves the best base KGE model GoldE on all metrics across all datasets, demonstrating the superior effectiveness and generalization of our difference-based semantic representations.

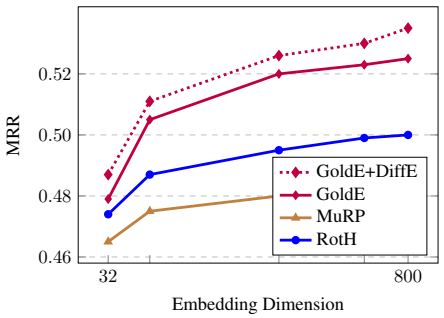

Figure 2: MRR results with the embedding dimension from $\{32, 50, 200, 500, 800\}$.

To further investigate the impact of different embedding dimensions, we conduct experiments with varying values of d. Figure 2 shows the MRR link prediction performance curves on the WN18RR dataset. We can see that DiffE consistently improves GoldE across a wide range of embedding dimensions, comprehensively showing the effectiveness of our difference-based semantic representations and the robust generalization ability of invariance properties.

Table 3: Ablation studies.

	WN18RR				FB15k-237				YAGO3-10			
Model	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
GoldE	.525	.476	.542	.615	.370	.277	.405	.558	.588	.510	.634	.723
GoldE+DiffEntity	**.528**	**.477**	**.546**	**.622**	**.373**	**.272**	**.407**	**.561**	**.590**	**.515**	**.636**	**.725**
GoldE+DiffRelation	**.532**	**.480**	**.547**	**.625**	**.377**	**.282**	**.410**	**.568**	**.597**	**.514**	**.644**	**.733**
GoldE+DiffE	**.537**	**.484**	**.550**	**.627**	**.380**	**.286**	**.412**	**.569**	**.601**	**.517**	**.646**	**.734**

6.4 Ablation Studies

We conduct the two-way ablation studies on link prediction to investigate the contributions of difference-based entity representations and difference-based relation representations. (a) +DiffEntity: only apply difference-based transformation operation for entity representations, *i.e.,* $f\big(\mathcal{M}(h),\ e_r,\ \mathcal{M}(t)\big)$; (b) +DiffRelation: only apply difference-based transformation operation for relation representations, *i.e.,* $f\big(e_h,\ \mathcal{M}(r),\ e_t\big)$. As shown in Table 3, we can observe that:

(1) both difference-based entity representations and difference-based relation representations contribute to performance improvements; (2) difference-based relation representations are more effective than difference-based entity representations; (3) combining difference-based entity representations together with difference-based relation representations shows the best results.

6.5 Efficiency

DiffE applied as a plugin maintains comparable efficiency in overall time complexity. Table 4 further reports the actual training time (10k steps) and testing time (the whole test procedure) using the same batch size settings. We can observe that DiffE applied as a plugin is as efficient as the base model.

Table 4: Efficiency comparison.

Dataset	Methods	Training	Testing
WN18RR	HousE	10.1min	20s
WN18RR	HousE+DiffE	10.5min	20s
FB15K-237	GoldE	26.1min	90s
FB15K-237	GoldE+DiffE	27min	90s

7 Related Work

Early KGE methods explore geometric operations (Ge et al., 2023; Zhang et al., 2022), including translation (Bordes et al., 2013; Wang et al., 2014; Ji et al., 2015), bilinear (Yang et al., 2015; Trouillon et al., 2016), and 2D rotations (Sun et al., 2019), to model relations. As rotations can deal with crucial logical patterns (Li et al., 2022), recent research mainly focus on two directions. Many works (Zhang et al., 2019; Gao et al., 2020; Cao et al., 2021) first seek higher-dimensional rotations for better modeling capacity. Another direction explores rotations for different (Balazevic et al., 2019; Chami et al., 2020) or combined (Xiong et al., 2022) geometries to preserve topological structures. GoldE (Li et al., 2025) generalizes existing KGE models in terms of both higher dimensions and different geometries.

However, existing methods implicitly share a common representational assumption: each embedding of entity or relation is directly treated as having its semantic meaning. Instead, we revisit this conceptually restrictive assumption and propose a difference-based perspective to represent semantic meaning.

8 Conclusion

In this paper, we proposed DiffE, a general and efficient plugin that introduces difference-based semantic representations for knowledge graph embedding. From a theoretical perspective, we show that difference-based representations induce a translation-invariant hypothesis class under mild data invariance assumptions. Empirically, extensive experiments on standard link prediction KGE benchmarks demonstrate that DiffE consistently improves a wide range of strong KGE baselines. In future work, we will explore the difference-based principle in more applications and settings, such as language models and reasoning.

AI USE STATEMENT

We have limited usage of ChatGPT (OpenAI), only to polish the writing of this paper. All suggestions from ChatGPT were carefully reviewed by the authors. No LLM tools are used to formulate research ideas, design or conduct experiments, or draw scientific conclusions. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

In this work, we have made every effort to ensure reproducibility. In Section 3, from input to output, we provide a detailed description of our method, including model structure, model training, and anomaly scores. In Appendix B, we describe all the datasets in detail. In Appendix C, we summarize all the implementation details and default hyper-parameters, including batch size, window size, learning rate, etc. Our codes will be available upon acceptance.

ETHICS AND IMPACT STATEMENT

This work does not introduce new sources of data or rely on sensitive personal information. The proposed operation is purely at the embedding/representation level. The ethical considerations and broader societal impacts are those commonly associated with existing knowledge graph embedding. We do not foresee significant negative societal impacts arising directly from this work.

This work studies difference-based semantic representations for knowledge graph embedding. The proposed principle is generic and can also be applied to a broad class of representation learning models such as language models. Potential positive impacts include drawing more attention to self-identical embeddings and semantic meaning, to improve link prediction, question answering, recommendation, or reasoning.

REFERENCES

Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. Multi-relational poincaré graph embeddings. In *Advances in Neural Information Processing Systems*, pp. 4465–4475, 2019.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 1247–1250, 2008.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, pp. 2787–2795, 2013.

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. Dual quaternion knowledge graph embeddings. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, pp. 6894–6902, 2021.

Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6901–6914, 2020.

Weijian Deng, Stephen Gould, and Liang Zheng. On the strong correlation between model invariance and generalization. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 1811–1818, 2018.

Manfredo Perdigao Do Carmo. *Riemannian geometry*. Birkhäuser, 1992.

R. M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967. ISSN 0022-1236.

Luther Pfahler Eisenhart. *Riemannian geometry*. Princeton University Press, 1997.

Frederick Arthur Ficken. The riemannian and affine differential geometry of product-spaces. *Annals of Mathematics*, pp. 892–913, 1939.

Mikhail Galkin, Xinyu Yuan, Hesham Mostafa, Jian Tang, and Zhaocheng Zhu. Towards foundation models for knowledge graph reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Chang Gao, Chengjie Sun, Lili Shan, Lei Lin, and Mingjiang Wang. Rotate3d: Representing relations as rotations in three-dimensional space for knowledge graph embedding. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 385–394, 2020.

Xiou Ge, Yun-Cheng Wang, Bin Wang, and C.-C. Jay Kuo. Compounding geometric operations for knowledge graph completion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6947–6965, 2023.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun 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*, pp. 687–696, 2015.

Rui Li, Jianan Zhao, Chaozhuo Li, Di He, Yiqi Wang, Yuming Liu, Hao Sun, Senzhang Wang, Weiwei Deng, Yanming Shen, Xing Xie, and Qi Zhang. House: Knowledge graph embedding with householder parameterization. In *International Conference on Machine Learning*, pp. 13209–13224, 2022.

Rui Li, Chaozhuo Li, Yanming Shen, Zeyu Zhang, and Xu Chen. Generalizing knowledge graph embedding with universal orthogonal parameterization. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2025.

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Yuan-Fang Li, Chen Gong, and Shirui Pan. Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025.

George Ma, Yifei Wang, Derek Lim, Stefanie Jegelka, and Yisen Wang. A canonicalization perspective on invariant and equivariant learning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. Yago3: A knowledge base from multilingual wikipedias. In *7th Biennial Conference on Innovative Data Systems Research*, 2015.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, pp. 39–41, 1995.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2018.

Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. A novel embedding model for knowledge base completion based on convolutional neural network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 327–333, 2018.

Akiyoshi Sannai, Masaaki Imaizumi, and Makoto Kawano. Improved generalization bounds of group invariant / equivariant deep networks via quotient feature spaces. In Cassio P. de Campos, Marloes H. Maathuis, and Erik Quaeghebeur (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, pp. 771–780, 2021.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations*, 2019.

Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and Their Compositionality*, pp. 57–66, 2015.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *Proceedings of the 33nd International Conference on Machine Learning*, pp. 2071–2080, 2016.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 1112–1119, 2014.

Bo Xiong, Shichao Zhu, Mojtaba Nayyeri, Chengjin Xu, Shirui Pan, Chuan Zhou, and Steffen Staab. Ultrahyperbolic knowledge graph embeddings. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2130–2139, 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 *3rd International Conference on Learning Representations*, 2015.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. In *Advances in Neural Information Processing Systems*, pp. 37309–37323, 2022.

Qianjin Zhang, Ronggui Wang, Juan Yang, and Lixia Xue. Knowledge graph embedding by reflection transformation. *Knowl. Based Syst.*, 238:107861, 2022.

Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. Quaternion knowledge graph embeddings. In *Advances in Neural Information Processing Systems*, pp. 2731–2741, 2019.

Sicheng Zhu, Bang An, and Furong Huang. Understanding the generalization benefit of model invariance from a data perspective. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 4328–4341, 2021.

Yuqicheng Zhu, Daniel Hernández, Yuan He, Zifeng Ding, Bo Xiong, Evgeny Kharlamov, and Steffen Staab. Predicate-conditional conformalized answer sets for knowledge graph embeddings. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *ACL 2025*, Findings of ACL, pp. 4145–4167. Association for Computational Linguistics, 2026.

Zhaocheng Zhu, Xinyu Yuan, Michael Galkin, Louis-Pascal A. C. Xhonneux, Ming Zhang, Maxime Gazeau, and Jian Tang. A*net: A scalable path-based reasoning approach for knowledge graphs. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Xingrui Zhuo, Shirui Pan, Jiapu Wang, Gongqing Wu, Zan Zhang, Rui Li, Zizhong Wei, and Xindong Wu. Progressive prefix-memory tuning for complex logical query answering on knowledge graphs. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pp. 3716–3724. ijcai.org, 2025a.

Xingrui Zhuo, Jiapu Wang, Gongqing Wu, Shirui Pan, and Xindong Wu. Effective instruction parsing plugin for complex logical query answering on knowledge graphs. In Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (eds.), *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pp. 4780–4792. ACM, 2025b.

A Empirical Rademacher Complexity

A.1 Generalization Error and Gap

Let $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample drawn i.i.d. from some data distribution $\mathcal{D}$, and $\mathcal{H}$ be a model class. Given a loss function $\ell : \mathbb{R} \to [0, 1]$, for a $h \in \mathcal{H}$, we define the *empirical error* as $R_{\mathcal{S}}(h) = \frac{1}{n}\sum_{i=1}^n \ell(h(\boldsymbol{x}_i), y_i)$, the *generalization error* as $R(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell(h(\boldsymbol{x}), y)]$, and the *generalization gap* as $R(h) - R_{\mathcal{S}}(h)$.

A.2 Complexity Measurements

Covering number and *Rademacher complexity* (Mohri et al., 2018) are two commonly used complexity measurements for model classes that can provide uniform generalization bounds. The covering number can also be directly used to upper bound the Rademacher complexity via Dudley's entropy integral theorem (Dudley, 1967; Mohri et al., 2018).

Covering number. Let $(\mathcal{F}, d)$ be a pseudometric space with some pseudometric[1] d. An *ϵ-cover* of a set $\mathcal{H} \subseteq \mathcal{F}$ is defined as a subset $\hat{\mathcal{H}} \subseteq \mathcal{H}$ such that for any $h \in \mathcal{H}$, there exists $\hat{h} \in \hat{\mathcal{H}}$ such that $d(h, \hat{h}) \leq \epsilon$. The covering number $N(\epsilon, \mathcal{H}, d)$ is defined as the minimum cardinality of an ϵ-cover (among all ϵ-covers) of $\mathcal{H}$. In this paper, we use the concept of covering number both for measuring model class complexities and for defining the sample covering number on datasets.

Empirical Rademacher complexity. Let $\mathcal{H}$ be a class of functions $h : \mathcal{X} \to \mathbb{R}$. Given a sample set $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$, the empirical Rademacher complexity of the model class $\mathcal{H}$ is defined as: $\mathscr{R}_{\mathcal{S}}(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i h(\boldsymbol{x}_i)\right]$ where $\boldsymbol{\sigma} = [\sigma_1, ..., \sigma_n]^\top$ is the vector of i.i.d. Rademacher random variables, each uniformly chosen from $\{-1, 1\}$.

A.3 Sample Cover Induced by Data Transformations

Existing empirical results (Deng et al., 2022; Sannai et al., 2021; Ma et al., 2024) suggest that, compared with standard models, invariant models may have certain properties to improve the performance. To identify such properties, Zhu et al. (2021) identify the related properties of the corresponding data transformations via the notion of *sample cover induced by data transformations*. We now formalize the notion following Zhu et al. (2021).

The definition of sample cover relies on the pseudometric induced by the data transformations G. Note that G generates an orbit $G(\boldsymbol{x}) \subseteq \mathcal{X}$ for each example $\boldsymbol{x} \in \mathcal{S}$. Let $\|\cdot\|$ be any norm on the input space $\mathcal{X}$. Given a set of transformations G, we define the G-induced pseudometric[2] as

$$\rho_G(\boldsymbol{x}_1, \boldsymbol{x}_2) = \inf_{\gamma\in\Gamma(\boldsymbol{x}_1,\boldsymbol{x}_2)} \int_\gamma c(\boldsymbol{r})ds, \quad \text{where } c(\boldsymbol{r}) = \begin{cases} 0, & \text{if } \boldsymbol{r} \in \cup_{\boldsymbol{x}\in\mathcal{S}} G(\boldsymbol{x}) \\ 1, & \text{otherwise} \end{cases} \tag{10}$$

where $ds = \|d\boldsymbol{r}\|$, and Γ denotes the set of all paths (curves) in $\mathcal{X}$ from $\boldsymbol{x}_1$ to $\boldsymbol{x}_2$. The ρ_G is essentially calculating the line integral along the shortest (if achievable) path γ in the scalar field c, where c can also be viewed as the "moving cost" function depending on G. The norm $\|\cdot\|$ here can be selected as any meaningful norm on the input space (e.g., Euclidean norm as in our experiments) and will later be used in defining the model's Lipschitz constant. It can be checked that ρ_G satisfies pseudometric axioms.

Definition A.1 (Sample cover induced by data transformations)**.** Let $(\mathcal{X}, \rho_G)$ be a pseudometric space and $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$ be a sample of size n. An *ϵ-sample cover* $\hat{\mathcal{S}}_{G,\epsilon}$ of the sample $\mathcal{S}$ induced by data transformations G at resolution ϵ is defined as a subset of the sample $\mathcal{S}$ such that for any $\boldsymbol{x} \in \mathcal{S}$, there exists $\hat{\boldsymbol{x}} \in \hat{\mathcal{S}}_{G,\epsilon}$ satisfying $\rho_G(\boldsymbol{x}, \hat{\boldsymbol{x}}) \leq \epsilon$.

Definition A.2 (Sample covering number induced by data transformations)**.** The *sample covering number* $N(\epsilon, \mathcal{S}, \rho_G)$ induced by data transformations G is defined as the minimum cardinality of an ϵ-sample cover:

$$N(\epsilon, \mathcal{S}, \rho_G) = \min\{|\hat{\mathcal{S}}_{G,\epsilon}| : \hat{\mathcal{S}}_{G,\epsilon} \text{ is an } \epsilon\text{-sample cover of } \mathcal{S}\}. \tag{11}$$

[1] A pseudometric is a metric if and only if it separates distinct points, namely $d(x, y) > 0$ for any $x \neq y$.

[2] Note that ρ_G is not a metric since it allows $\rho_G(x, y) = 0$ for $x \neq y$.

A.4 PROOF

Proof of Lemma 5.3. We give a self-contained proof of the Dudley entropy-integral bound used in Zhu et al. (2021), specialized to the hypothesis class $\mathcal{H}_{\mathrm{inv}}$.

Fix a nonempty sample $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$, where $n = |\mathcal{S}|$, and condition on $\mathcal{S}$ throughout the proof. Define the empirical L_2 pseudometric by

$$d_{\mathcal{S}}(h, g) := \|h - g\|_{L_2(\mathbb{P}_{\mathcal{S}})} = \left(\frac{1}{n} \sum_{i=1}^{n} \big(h(\boldsymbol{x}_i) - g(\boldsymbol{x}_i)\big)^2 \right)^{1/2}. \tag{12}$$

Recall that

$$\mathscr{R}_{\mathcal{S}}(\mathcal{H}_{\mathrm{inv}}) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}_{\mathrm{inv}}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(\boldsymbol{x}_i) \right], \tag{13}$$

where $\sigma_1, \ldots, \sigma_n$ are independent Rademacher random variables.

Fix any $\alpha > 0$. If the entropy integral in the lemma is infinite, the claimed inequality is immediate. Hence, assume that it is finite. This implies that $\mathcal{H}_{\mathrm{inv}}$ admits a finite cover at some finite radius under $d_{\mathcal{S}}$, and consequently has finite empirical diameter. Since its evaluation vectors belong to the finite-dimensional space $\mathbb{R}^n$, it then admits finite covers at every positive radius.

Step 1: A finite-set Rademacher bound. Let $\mathcal{A} \subset \mathbb{R}^n$ be a finite nonempty set satisfying $\|\mathbf{a}\|_2/\sqrt{n} \le r$ for every $\mathbf{a} \in \mathcal{A}$. For any $\lambda > 0$, independence of the Rademacher variables and the inequality $\cosh(u) \le \exp(u^2/2)$ give

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\sigma}} \left[\exp\left(\frac{\lambda}{n} \sum_{i=1}^{n} \sigma_i a_i \right) \right] &= \prod_{i=1}^{n} \cosh\left(\frac{\lambda a_i}{n} \right) \\ &\le \exp\left(\frac{\lambda^2 \|\mathbf{a}\|_2^2}{2n^2} \right) \\ &\le \exp\left(\frac{\lambda^2 r^2}{2n} \right). \end{aligned} \tag{14}$$

It follows from Jensen's inequality that

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\sigma}} \left[\max_{\mathbf{a} \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right] &\le \frac{1}{\lambda} \log \sum_{\mathbf{a} \in \mathcal{A}} \mathbb{E}_{\boldsymbol{\sigma}} \left[\exp\left(\frac{\lambda}{n} \sum_{i=1}^{n} \sigma_i a_i \right) \right] \\ &\le \frac{\log |\mathcal{A}|}{\lambda} + \frac{\lambda r^2}{2n}. \end{aligned} \tag{15}$$

Taking the infimum over $\lambda > 0$ yields

$$\mathbb{E}_{\boldsymbol{\sigma}} \left[\max_{\mathbf{a} \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right] \le r \sqrt{\frac{2 \log |\mathcal{A}|}{n}}. \tag{16}$$

For $|\mathcal{A}| = 1$, the same conclusion follows by taking $\lambda \downarrow 0$ in equation 15.

Step 2: Constructing covers at successive resolutions. Choose an arbitrary $h_0 \in \mathcal{H}_{\mathrm{inv}}$ and a finite τ_0 such that

$$\tau_0 \ge \max \left\{ 4\alpha, \sup_{h, g \in \mathcal{H}_{\mathrm{inv}}} d_{\mathcal{S}}(h, g) \right\}. \tag{17}$$

Let $\tau_j = 2^{-j} \tau_0$ for $j \ge 0$. Set $\widehat{\mathcal{H}}_0 = \{h_0\}$ and $\pi_0(h) = h_0$ for every $h \in \mathcal{H}_{\mathrm{inv}}$. For each $j \ge 1$, choose a minimum-cardinality τ_j-cover $\widehat{\mathcal{H}}_j \subseteq \mathcal{H}_{\mathrm{inv}}$, so that

$$N_j := |\widehat{\mathcal{H}}_j| = N\big(\tau_j, \mathcal{H}_{\mathrm{inv}}, L_2(\mathbb{P}_{\mathcal{S}})\big). \tag{18}$$

For every $h \in \mathcal{H}_{\mathrm{inv}}$, select $\pi_j(h) \in \widehat{\mathcal{H}}_j$ satisfying

$$d_{\mathcal{S}}\big(h, \pi_j(h)\big) \le \tau_j. \tag{19}$$

The same approximation property holds for $j = 0$ by the choice of τ_0. Moreover, $N_0 = 1$ and $N_{j-1} \le N_j$.

Choose the largest integer $J \ge 0$ such that $\tau_J \ge 2\alpha$. Then $J \ge 1$ and

$$2\alpha \le \tau_J < 4\alpha, \qquad \tau_{J+1} = \frac{\tau_J}{2} \ge \alpha. \tag{20}$$

For every $h \in \mathcal{H}_{\mathrm{inv}}$, the following telescoping identity holds:

$$h = h_0 + \sum_{j=1}^{J} \big(\pi_j(h) - \pi_{j-1}(h)\big) + \big(h - \pi_J(h)\big). \tag{21}$$

Step 3: Bounding the residual and the increments. For every realization of $\boldsymbol{\sigma}$, the Cauchy–Schwarz inequality gives

$$\begin{aligned} \sup_{h \in \mathcal{H}_{\mathrm{inv}}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \big(h(\boldsymbol{x}_i) - \pi_J(h)(\boldsymbol{x}_i)\big) &\le \sup_{h \in \mathcal{H}_{\mathrm{inv}}} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i^2\right)^{1/2} d_{\mathcal{S}}\big(h, \pi_J(h)\big) \\ &\le \tau_J. \end{aligned} \tag{22}$$

For each $j = 1, \dots, J$, define the finite set of increment vectors

$$\mathcal{A}_j = \left\{ \big(\pi_j(h)(\boldsymbol{x}_i) - \pi_{j-1}(h)(\boldsymbol{x}_i)\big)_{i=1}^{n} : h \in \mathcal{H}_{\mathrm{inv}} \right\}. \tag{23}$$

Its cardinality satisfies

$$|\mathcal{A}_j| \le N_j N_{j-1} \le N_j^2. \tag{24}$$

For every increment generated by h, the triangle inequality yields

$$\begin{aligned} d_{\mathcal{S}}\big(\pi_j(h), \pi_{j-1}(h)\big) &\le d_{\mathcal{S}}\big(\pi_j(h), h\big) + d_{\mathcal{S}}\big(h, \pi_{j-1}(h)\big) \\ &\le \tau_j + \tau_{j-1} = 3\tau_j. \end{aligned} \tag{25}$$

Applying equation 16 to $\mathcal{A}_j$ therefore gives

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}_{\mathrm{inv}}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \big(\pi_j(h)(\boldsymbol{x}_i) - \pi_{j-1}(h)(\boldsymbol{x}_i)\big) \right] \\ &\le 3\tau_j \sqrt{\frac{2 \log |\mathcal{A}_j|}{n}} \le 6\tau_j \sqrt{\frac{\log N_j}{n}}. \end{aligned} \tag{26}$$

Since h_0 is fixed independently of $\boldsymbol{\sigma}$,

$$\mathbb{E}_{\boldsymbol{\sigma}} \left[\frac{1}{n} \sum_{i=1}^{n} \sigma_i h_0(\boldsymbol{x}_i) \right] = 0. \tag{27}$$

Combining the telescoping identity, the residual bound, and the increment bounds gives

$$\mathscr{R}_{\mathcal{S}}(\mathcal{H}_{\mathrm{inv}}) \le \tau_J + 6 \sum_{j=1}^{J} \tau_j \sqrt{\frac{\log N\big(\tau_j, \mathcal{H}_{\mathrm{inv}}, L_2(\mathbb{P}_{\mathcal{S}})\big)}{n}}. \tag{28}$$

Step 4: Converting the sum into an entropy integral. The covering number is nonincreasing in its radius. Thus, for $\tau \in [\tau_{j+1}, \tau_j]$,

$$\sqrt{\frac{\log N\big(\tau, \mathcal{H}_{\mathrm{inv}}, L_2(\mathbb{P}_{\mathcal{S}})\big)}{n}} \ge \sqrt{\frac{\log N_j}{n}}. \tag{29}$$

Since $\tau_j - \tau_{j+1} = \tau_j/2$, it follows that

$$\tau_j \sqrt{\frac{\log N_j}{n}} \le 2 \int_{\tau_{j+1}}^{\tau_j} \sqrt{\frac{\log N\big(\tau, \mathcal{H}_{\mathrm{inv}}, L_2(\mathbb{P}_{\mathcal{S}})\big)}{n}}\, d\tau. \tag{30}$$

Substituting this inequality into equation 28 and using equation 20, we obtain

$$\begin{aligned}\mathscr{R}_{\mathcal{S}}(\mathcal{H}_{\mathrm{inv}}) &\leq \tau_J + 12\int_{\tau_{J+1}}^{\tau_1}\sqrt{\frac{\log N\big(\tau,\mathcal{H}_{\mathrm{inv}},L_2(\mathbb{P}_{\mathcal{S}})\big)}{n}}\,d\tau \\ &\leq 4\alpha + 12\int_{\alpha}^{\infty}\sqrt{\frac{\log N\big(\tau,\mathcal{H}_{\mathrm{inv}},L_2(\mathbb{P}_{\mathcal{S}})\big)}{|\mathcal{S}|}}\,d\tau.\end{aligned} \tag{31}$$

Because $\alpha > 0$ was arbitrary, this proves the lemma. □

Remark A.3. Lemma 5.3 bounds the empirical Rademacher complexity of $\mathcal{H}_{\mathrm{inv}}$ and does not itself require translation invariance. The role of invariance is to enable further bounds on the covering numbers appearing in the entropy integral.

B Datasets

Table 5 summarizes the detailed statistics of three datasets that are standard benchmarks for the knowledge graph embedding and link prediction.

WN18RR (Dettmers et al., 2018) and FB15k-237 (Toutanova & Chen, 2015) datasets are subsets of WN18 (Bordes et al., 2013) and FB15k (Bordes et al., 2013), respectively, with inverse relations removed to avoid test leakage. WN18 is extracted from WordNet (Miller, 1995), a database with lexical relations between words. FB15k is extracted from Freebase (Bollacker et al., 2008), a large-scale knowledge graph with general knowledge facts. YAGO3-10 is a subset of YAGO3 (Mahdisoltani et al., 2015), containing the most common 37 relations and all entities with more than 10 linked relations to ensure sufficient knowledge contents. Most of the triples in YAGO3-10 are general knowledge facts of people, such as citizenship, gender, profession and marital status.

Table 5: Statistics of five standard benchmarks.

Dataset	#Entity	#Relation	#Training	#Validation	#Test
WN18RR	40,943	11	86,835	3,034	3,134
FB15k-237	14,541	237	272,115	17,535	20,466
YAGO3-10	123,182	37	1,079,040	5,000	5,000

C Implementation Details

All experiments are conducted in Python 3.8 and PyTorch 1.8.0, executed on the Ubuntu 24.04.1 LTS server equipped with AMD EPYC 7642 48-Core CPU @ 2.30GHz, 128 GB of RAM, and NVIDIA RTX 4090 GPUs (24 GB), running CUDA 12.9 and driver version 575.64.03. To follow existing works (Sun et al., 2019; Li et al., 2025; 2022; Gao et al., 2020) and ensure fair comparisons, we fix the embedding size d of each entity as 800, 600, 1000 on WN18RR, FB15k-237 and YAGO3-10 datasets for the main experiments, respectively.

Evaluation: We follow the filtered ranking protocol (Bordes et al., 2013; Li et al., 2025; 2022) for all KGE methods during evaluation. For a test triple (h, r, t), we will rank it against all negative triplets (h, r, t') or (h', r, t) that do not appear in the knowledge graph.

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}, \quad \text{Hits@K} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\text{rank}_i \leq K),$$

where rank_i is the rank position of the correct one and $\mathbb{I}$ is the indicator function.

Hyperparameters: We fine-tune the hyperparameters on the validation datasets. The tuning of specific hyperparameters (*e.g.,* batch sizes, learning rates, margins) follows the original papers of each base KGE model. For example, the hyperparameters of GoldE (Li et al., 2025) are tuned by the random search (Bergstra & Bengio, 2012), including batch size b, self-adversarial negative sampling temperature α, loss margin γ, learning rate lr, dimension k_*, number of elliptic component spaces $m_{\mathbb{P}}$, and number of hyperbolic component spaces $m_{\mathbb{Q}}$. Their hyperparameter search spaces are shown in Table 6.

Table 6: Hyperparameter search space.

Hyperparameter	Search Space	Type
b	$\{500, 800, 1000, 1500, 2000\}$	Choice
α	$[0.5, 2.0]$	Range
γ	$\{6, 8, 10, 12, 16, 20, 24, 28\}$	Choice
lr	$\{0.0001, 0.0005, 0.001, 0.003\}$	Choice
k_*	$\{200, 400, 600, 800\}$	Choice
$m_{\mathbb{P}}, m_{\mathbb{Q}}$	$\{2, 4, 8, 10, 15, 20, 25, 30\}$	Choice

Table 7: Link prediction results with embedding dimension restricted to 32.

	WN18RR				FB15k-237				YAGO3-10			
Model	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
RotatE	.387	.330	.417	.491	.290	.208	.316	.458	-	-	-	-
QuatE	.421	.396	.430	.467	.293	.212	.320	.460	-	-	-	-
MuRE	.458	.421	.471	.525	.313	.226	.340	.489	.283	.187	.317	.478
MuRP	.465	.420	.484	.544	.323	.235	.353	.501	.230	.150	.247	.392
RefH	.447	.408	.464	.518	.312	.224	.342	.489	.381	.302	.415	.530
RotH	.472	.428	.490	.553	.314	.223	.346	.497	.393	.307	.435	.559
AttH	.466	.419	.484	.551	.324	.236	.354	.501	.397	.310	.437	.566
GoldE	.479	.441	.493	.557	.325	.238	.356	.502	.415	.324	.456	.582
GoldE+DiffE	**.487**	**.449**	**.497**	**.565**	**.338**	**.247**	**.364**	**.513**	**.425**	**.333**	**.468**	**.590**

D STATISTICAL SIGNIFICANCE

As the SOTA baseline GoldE did not report the statistical significance, we add 5-random-seed mean results with (std) using the best hyperparameters of each method. The experimental results are shown in Table 8, and we can see that DiffE significantly outperforms the SOTA baseline where the two-tailed P value is less than 0.5

Table 8: Main experimental results with statistical significance

Method	WN18RR				FB15k-237				YAGO3-10			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
GoldE	.523(.003)	.474(.003)	.540(.004)	.613(.004)	.368(.002)	.275(.003)	.402(.003)	.556(.004)	.585(.003)	.508(.004)	.631(.005)	.719(.005)
+DiffE	.534(.001)	.483(.002)	.548(.002)	.625(.003)	.378(.001)	.284(.002)	.411(.003)	.566(.003)	.598(.002)	.515(.002)	.645(.003)	.731(.004)