

# Replacing Paths With Connection-Biased Attention for Knowledge Graph Completion

Sharmishtha Dutta<sup>1</sup>, Alex Gittens<sup>1</sup>, Mohammed J. Zaki<sup>1</sup>, Charu C. Aggarwal<sup>2</sup>

<sup>1</sup>Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, USA

<sup>2</sup>IBM T. J. Watson Research Center, Yorktown Heights, NY, USA

duttas@rpi.edu, gittea@rpi.edu, zaki@cs.rpi.edu, charu.us@ibm.com

## Abstract

Knowledge graph (KG) completion aims to identify additional facts that can be inferred from the existing facts in the KG. Recent developments in this field have explored this task in the inductive setting, where at test time one sees entities that were not present during training; the most performant models in the inductive setting have employed path encoding modules in addition to standard subgraph encoding modules. This work similarly focuses on KG completion in the inductive setting, without the explicit use of path encodings, which can be time-consuming and introduces several hyperparameters that require costly hyperparameter optimization. Our approach uses a Transformer-based subgraph encoding module only; we introduce connection-biased attention and entity role embeddings into the subgraph encoding module to eliminate the need for an expensive and time-consuming path encoding module. Evaluations on standard inductive KG completion benchmark datasets demonstrate that our **Connection-Biased Link Prediction (CBLiP)** model has superior performance to models that do not use path information. Compared to models that utilize path information, CBLiP shows competitive or superior performance while being faster. Additionally, to show that the effectiveness of connection-biased attention and entity role embeddings also holds in the transductive setting, we compare CBLiP’s performance on the relation prediction task in the transductive setting.

## Introduction

Knowledge graphs (KGs) store facts expressed in the forms of relationships between entities. Each fact is represented as a triple  $(h,r,t)$  or  $(head, relation, tail)$ . Here, the *head* and *tail* represent entities such as people, places, and institutions, and the *relation* represents the relation between the two entities. These facts are modeled as a directed graph with labeled edges, where each entity is a vertex in the graph, the relations are the edge labels, and edges are directed from the head to the tail entities.

KGs are often constructed by crowdsourced data, or by using off-the-shelf fact extraction tools. Therefore, in addition to containing spurious information, they can omit facts that are implicit in the observed data. These omissions hinder the usefulness of KGs in various downstream tasks such as question answering in search engines. This has motivated

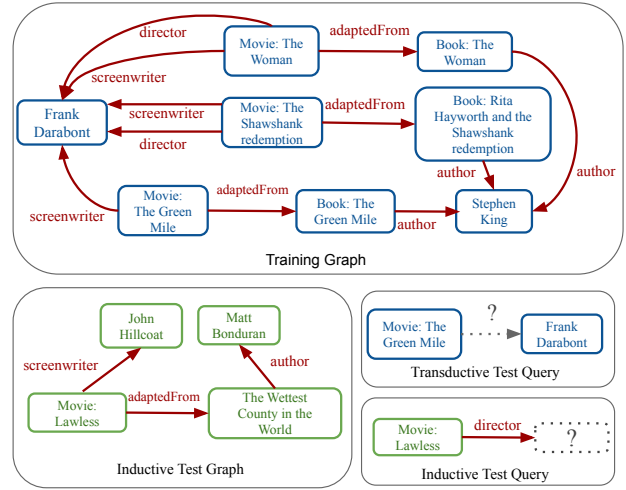


Figure 1: Example of training graph and test queries for KG completion in transductive and inductive settings.

researchers to use statistical relational learning to infer missing entities or relations from incomplete triples. Earlier approaches to KG completion were developed for the transductive setting, where the triples to be completed consist of entities and relations seen during training. While this stream of research has yielded numerous high-performing models, these models cannot be used when the triples to be completed contain entities that were not seen at training time.

Real-world KGs grow continuously as new facts are added, and the set of entities may grow over time, meaning that at any given point, completion models may need to be used on entities that were not seen during training. Inductive KG Completion assumes the relations between entities remain the same in the newly added facts. To illustrate this setting, Figure 1 depicts a training graph that reflects the correlation between the role of a director and screenwriter. The same artist often plays these two roles when the screenplay is adapted from an existing novel. In the transductive KG completion setting, a relation prediction task involves queries where both entities are present in the training graph and the most plausible relation is to be determined. In the inductive setting, the test graph contains entities added to

the graph after the model was trained, and the inductive KG completion task to find the best option for a missing entity involves entities unseen during training (marked by purple ink in the figure).

Naturally, in the inductive setting, graph-learning-based approaches rely greatly on the information provided by neighbors of the incomplete triple of interest. This has led to several Graph Convolution Network (GCN) based models that encode the surrounding subgraph to learn about relation interactions. Paths between entities have also been utilized in several models (Lin et al. 2022; Li, Wang, and Mao 2023; Pan et al. 2022; Zhu et al. 2021), as explicitly encoding path information has resulted in a great performance boost. Here, path information is usually defined as an ordered sequence of relations between two entities. For example, two paths between Stanley Kubrick and Stephen King, according to Figure 1, are  $(\text{screenwriter}^{-1}, \text{adaptedFrom}, \text{author})$  and  $(\text{director}^{-1}, \text{adaptedFrom}, \text{author})$ . One could also find a path of length 6 between Stanley Kubrick and Frank Darabont.

The use of paths necessitates multiple decisions, such as determining the ideal path length to obtain meaningful information and exclude noise, the number of paths to be extracted, and various design choices for combining this information with the subgraph information. Additionally, path extraction between entities on the fly and representing them in the model adds overhead to the training time and parameter count.

Transformers have replaced Recurrent Neural Networks for sequence modeling precisely because appropriate positional encoding allows attention alone to suffice; we hypothesize that when represented as sequences of triples, appropriate positional encodings similarly unleash the full power of attention, obviating the need for explicit and costly modeling of path information. Indeed, by adapting the connection-biased attention from GRAN (Wang et al. 2021)—where it was used to better represent single  $n$ -ary facts in a knowledge base—to the representation of subgraphs of a KG, we demonstrate that Transformers alone, without specialized submodules for path representation, suffice to perform accurate KG completion.

**Main Contributions.** Our key contributions are:

1. We introduce CBLiP: a context-aware Transformer-based model, with a novel connection-biased attention module at its core for reasoning in KGs.
2. We introduce entity roles, a simple and effective construct to represent unseen entities in a subgraph, as an alternative to conventional relative distance-based entity labeling in the inductive link prediction setting.
3. We demonstrate the effectiveness of CBLiP by comparing its performance on the entity prediction task in the inductive setting with that of state-of-the-art models on benchmark datasets and showing that it achieves best-performing or competitive results.
4. We highlight CBLiP’s effectiveness across settings by similarly evaluating its performance on a transductive relation prediction task.

## Related Work

Knowledge graph completion is often based on learning continuous vector embedding of entities and relations. KG completion garnered attention in the early 2010s with TransE (Bordes et al. 2013) where the distance between learned embedding vectors determined the plausibility of a triple. GraIL (Teru, Denis, and Hamilton 2020) introduced the inductive learning task and the standard benchmark datasets for this task and has garnered much attention as a more practical approach. In this section, we discuss the development of both settings.

### Transductive Learning Models

Initial work in the transductive setting focused on triple-based models, exploiting entities’ inherent properties. Recent work has focused on gathering information about the neighborhood surrounding an entity.

**Translational Distance-based Models** Research in translation-based models focused on learning embeddings that satisfy specific properties. For example, TransE (Bordes et al. 2013) models relations as translation vectors between head and tail entities, aiming to maintain  $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ . Despite its limitations in representing many-to-one and one-to-many relations, TransE remains a lightweight and straightforward model. Subsequent models like RotatE (Sun et al. 2018) and QuatE (Zhang et al. 2019) embed entities in complex and quaternion spaces, respectively.

**Factorization Based Models** Another family of models aims to capture semantic similarity by observing pairwise interactions between entities. RESCAL (Nickel, Tresp, and Kriegel 2011) was an early model with this idea. It was followed by a simplified variation DistMult (Yang et al. 2015) and a complex number variation ComplEx (Trouillon et al. 2016). Simple (Kazemi and Poole 2018) extends Canonical Polyadic (CP) decomposition by removing the independence between learned entity representations.

**Neural Models** PathCon (Wang, Ren, and Leskovec 2021) utilizes a message-passing mechanism to aggregate edge-based local context and paths between node pairs for relation prediction. It does not learn any node embeddings, limiting its applicability for entity prediction.

### Inductive Learning Models

Inductive KG completion aims to extend the task to unseen entities. Models first relied on Graph Neural Networks (GNNs) and more recently on Transformers to encode context information by aggregating neighborhood interactions. Many models have leveraged path information between the *head* and *tail* entities by incorporating a path encoding module in addition to the subgraph module that aggregates neighborhood interactions.

**Models without Path Information** GraIL (Teru, Denis, and Hamilton 2020) selects the common neighbors of the target head and tail entities as the context of a target triple to be scored. The model employs double-radius labeling (distance from head, distance from tail) to denote each entity’s

relative position and learns embeddings for relations through attention computation inside an Relational GCN module. The final score is a function of the triple in consideration and the encoded subgraph surrounding it. CoMPiLE (Mai et al. 2021) extends GraIL by considering directedness in its subgraph encoding module and computing edge (triple) attentions to learn edge embedding. TACT (Chen et al. 2021) expands GraIL by proposing an additional relation correlation module by learning 6 predefined categories of interactions via unique linear transformations for each kind. This categorization of TACT is closely related to our approach to constructing a connection-biased adjacency matrix of 7 categories. However, TACT incorporates this information into its subgraph encoding module whereas we use it to compute connection-biased attention in the transformer layers.

**Models with Path Information** ConGLR (Lin et al. 2022) expands on GraIL by modifying the subgraph encoding module. Additionally, it constructs a context graph that uses the relational paths involving the neighborhood entities. A combination function and a weighted aggregation function are employed to encode the paths represented as a sequence of relations and to combine the paths, respectively. The final scoring function integrates context information and path-based logical reasoning. Report (Li, Wang, and Mao 2023) uses successive stacks of transformer layers for context encoding and path encoding. A hierarchical structure of transformer layers is utilized to fuse the query and the representations of context and path to compute a final score. (Pan et al. 2022) proposed LogCo where a GCN-based subgraph module is complemented with a path encoding module. Each path representation is compared with the target relation to compute an attention score based on similarity. The model uses positive and negative path samples for a contrastive training regime. These models add the overhead of path representation and path aggregation during training and path extraction during training and inference time.

NBFNet (Zhu et al. 2021) offers a more scalable solution to this by generalizing Bellman-Ford algorithm for finding the shortest paths. The model learns entity pair representations as well as path representations utilizing the distributive properties of the generalized operators. This allows parallel scoring of query triples that share the same entity-relation pairs and the models suffer from time complexity as well as memory overhead due to additional support needed to represent paths. While these models perform better than those without path information, they come with a drawback of real-time path extraction for inference triples. Since the core of inductive link prediction is to conduct reasoning in unseen entities during training, preprocessing of paths is not a realistic choice for inference triples.

### Models with Rule Extraction

RPJE (Niu et al. 2020) combines rules and paths by injecting length-2 rules into KG embeddings. DRUM (Sadeghian et al. 2019) and NeuralLP (Yang, Yang, and Cohen 2017) extract probabilistic first-order logic rules to assign weights to paths between entities. RuleN (Meilicke et al. 2018) assigns confidence to rules by randomized process. While these

methods are efficient in learning short and simple rules, they suffer from scalability issues while finding frequent patterns in large graphs.

## Problem Formulation

Relational data can be modeled as a directed heterogeneous graph  $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})$  where  $\mathcal{E}$  and  $\mathcal{R}$  represent the set of entities and relations modeled as nodes and edge types in the graph, respectively.  $\mathcal{F} \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E}$  represents the labeled edges or fact triples represented as ordered tuples of head-entity, relation, tail-entity.

The inductive KG completion task comprises the following components.

1. **Training graph:**  $\mathcal{G}_{\text{train}} = (\mathcal{E}_{\text{train}}, \mathcal{R}, \mathcal{F}_{\text{train}})$  where  $\mathcal{F}_{\text{train}} \subset \mathcal{E}_{\text{train}} \times \mathcal{R} \times \mathcal{E}_{\text{train}}$  is the set of training facts.
2. **Validation triples:**  $\mathcal{F}_{\text{valid}} \subset \mathcal{E}_{\text{train}} \times \mathcal{R} \times \mathcal{E}_{\text{train}}$
3. **Test graph:**  $\mathcal{G}_{\text{test}} = (\mathcal{E}_{\text{test}}, \mathcal{R}, \mathcal{F}_{\text{test}})$  where  $\mathcal{F}_{\text{test}} \subset \mathcal{E}_{\text{test}} \times \mathcal{R} \times \mathcal{E}_{\text{test}}$ . This serves as the fact graph for the test-time inference triples.
4. **Test-time inference triples:**  $\mathcal{F}_{\text{infer}} \subset \mathcal{E}_{\text{test}} \times \mathcal{R} \times \mathcal{E}_{\text{test}}$ . Given an incomplete fact from  $\mathcal{F}_{\text{infer}}$ , the aim is to complete it using information from  $\mathcal{G}_{\text{test}}$  and the model trained on  $\mathcal{G}_{\text{train}}$ .

The inductive setting is characterized by the fact that  $\mathcal{E}_{\text{train}} \cap \mathcal{E}_{\text{test}} = \emptyset$ , i.e., the entities seen at test time were unseen at training time.

Our goal is to find a model that computes a score  $s$  for a triple  $\langle h, r, t \rangle$ . We hypothesize that a plausibility score of a triple can be determined using information present in the ego graphs of the *head* and *tail* entities. The  $k$ -hop ego graph  $\mathcal{N}_e$  of entity  $e$  consists of the triples in its  $k$ -hop enclosing subgraph. Thus, we model the score with

$$s = g(h, r, t, \mathcal{N}_h, \mathcal{N}_t) \quad (1)$$

Here,  $g$  is a function of the triple and its local contexts; in this work,  $g$  is given by the CBLIP architecture introduced in the next section. During training, we corrupt either the head or tail of a true triple ( $p_i$ ) and obtain a corrupted triple ( $n_i$ ). We then train the model to assign higher scores to true triples using a margin-based ranking loss:

$$\mathcal{L} = \sum_{i=1}^{|\mathcal{F}_{\text{train}}|} \max(0, g(n_i) - g(p_i) + \gamma)$$

Here,  $\gamma$  is the margin that separates the true and corrupted facts and allows for flexibility in training. At test time, triples with higher scores are considered to be more plausible completions.

### Transductive Setting

In the transductive setting, there is no separate test graph and the test-time inference triples satisfy  $\mathcal{F}_{\text{infer}} \subset (\mathcal{E}_{\text{train}} \times \mathcal{R} \times \mathcal{E}_{\text{train}})$ . That is, the entities seen at test time are all present in the training graph. Our goal in relation prediction is to, given putative head and tail entities, predict the relation between them. There are typically many fewer relations than entities (a few hundred vs. tens of thousands), so instead of a scoring

function, we explicitly model the likelihood of each relation given the putative head and tail entities:

$$\mathbb{P}(r|h, t) \propto g(h, t, \mathcal{N}_h, \mathcal{N}_t) \quad (2)$$

One advantage of explicitly modeling  $\mathbb{P}(r|h, t)$  is that there is no need for negative samples. The model is trained by minimizing the cross-entropy loss between the log-likelihood of our estimation and the ground truth observation  $r$  over the training data. CBLiP is again used as the architecture for  $g$  in this setting.

## Model Overview

This section describes the architecture used for  $g$  in the CBLiP model. Given  $\langle h, r, t \rangle$ , we find the neighboring triples of both  $h$  and  $t$  as the context of that triple. In the inductive setting, each triple in our model is represented using learned entity role vectors for its *head* and *tail* entity along with a *relation* vector embedding. The resulting vector encodings of the triples in the neighborhood are used as inputs to a connection-biased Transformer, and the output sequence is passed through a linear transformation to obtain the final score. Below, we explain these components in detail and the modifications made for the transductive setting.

### Entity Role Vectors

Inductive learning is aided by learning entity behaviors and interactions to facilitate inference on the unseen (during training) entities during test time. To this end, we represent each unseen entity by a vector representing its role. Traditional GNN-based models (Teru, Denis, and Hamilton 2020; Chen et al. 2021; Lin et al. 2022) distinguish neighbor entities in a subgraph by assigning relative distance from the target *head* or *tail* entity and by initializing the values with one-hot vector encoding.

Instead, we represent entities in the local neighborhood of the putative triple of interest by assigning roles to them. Such an entity has one of three roles:  $\{\textit{head}, \textit{tail}, \textit{and other}\}$ . The role simply represents whether the neighbor entity is in fact the putative *head* entity, the putative *tail* entity, or some other entity. Despite the simplicity of the role embeddings, this distinction allows the model to distinguish between triples that are immediate neighbors or distant neighbors of the putative triple, thus improving the model’s performance. This approach also ensures a shared representation of these roles across the model, unlike the local updates of relative-distance-based labeling. We denote the role vector of an entity  $e$  succinctly with  $\text{ROLE}(e)$ .

### Context Embedding and Representation of a Triple

We employ a connection-biased Transformer Encoder to obtain contextual embeddings for a given target triple  $\langle h, r, t \rangle$ . We define the neighborhood of  $h$  and  $t$  as:

$$\mathcal{N} = \{f | f \in \mathcal{N}_h \oplus \mathcal{N}_t\}$$

and use a breadth-first search algorithm to collect the triples in  $\mathcal{N}$ . Here  $\oplus$  is the union or intersection operation. We select up to  $m$  neighboring triples from  $\mathcal{N}$ ; here  $m$  is a hyperparameter. We obtain the embedding of each triple

	(h,r,t)	(h,r1,e1)	(e1,r2,e2)	(t,r3,e3)	(e3,r4,e4)	(e4,r5,t)
(h,r,t)	-	+	-	×	-	◆
(h,r1,e1)	+	-	×	-	-	-
(e1,r2,e2)	-	◆	-	-	-	-
(t,r3,e3)	◆	-	-	-	×	◆
(e3,r4,e4)	-	-	-	◆	-	×
(e4,r5,t)	◆	-	-	×	◆	-

Legend + head-head ◆ tail-tail ◆ head-tail × tail-head

Figure 2: An example of constructing a connection-biased adjacency matrix. The icons denote the presence of a particular kind of overlap of entities between triples.

$f = \langle e_1, r, e_2 \rangle$  in  $\mathcal{N}$  by aggregating its entity and relation embeddings:

$$\mathbf{f} = \mathcal{A}(\text{ROLE}(e_1), \mathbf{r}, \text{ROLE}(e_2))$$

We explored two options for the aggregation function  $\mathcal{A}$ : concatenation and mean. The connection-biased aspect of the Transformer Encoder is described in the following section.

### Input Sequence

With different pieces of information at hand, we can construct the final contextual representation of the target triple  $\langle h, r, t \rangle$  for scoring. For each triple in  $\mathcal{T}_{\text{train}}$ , we construct a sequence of tokens encoding its neighborhood,  $\mathcal{S}_{in}$ :

$$\mathcal{S}_{in} = [\mathbf{f}^*, \mathbf{f}_{\mathcal{N}}^1, \dots, \mathbf{f}_{\mathcal{N}}^m] \quad (3)$$

Here,  $\mathbf{f}^*$  is the embedding of the putative completed triple to be scored. This embedding is additionally aggregated (using  $\oplus$ ) with a special `target` vector embedding to distinguish it from the embeddings for triples from  $\mathcal{N}$ .

### Connection-Biased Adjacency Matrix

Paths are often used to learn relation interaction patterns in a (sub)graph. We want to avoid the design decisions, time, and memory complexity that come with this addition to a model. By constructing a connection-biased adjacency matrix, we aim to learn implicit knowledge of paths, distance, and shared neighborhoods, which is instrumental in correctly predicting the final relation.

While GRAN (Wang et al. 2021) constructs a similar matrix for each of its  $n$ -ary fact’s components, we build this for members in a subgraph. The connection types in our model depict the overlap of entities between neighboring triples whereas GRAN distinguishes connections between  $n$ -ary fact components such as entity-value, attribute-value, and so on.

We construct a connection-biased adjacency matrix  $\mathbf{C}$  for the triples in a subgraph. We do so by comparing whether these triples share the same head and tail entity and in which manner. Each entry  $c_{ij} \in \mathbf{C}$  represents the kind of connections two triples  $f_i$  and  $f_j$  can have:

1.  $f_i$ 's head is  $f_j$ 's head
2.  $f_i$ 's tail is  $f_j$ 's tail
3.  $f_i$ 's tail is  $f_j$ 's head
4.  $f_i$ 's head is  $f_j$ 's tail
5.  $f_i$ 's head is  $f_j$ 's head AND  $f_i$ 's tail is  $f_j$ 's tail (implying there are two parallel edges between the same pair of entities)
6.  $f_i$ 's head is  $f_j$ 's tail AND  $f_i$ 's tail is  $f_j$ 's head (inverse relations, e.g., `sonOf` and `motherOf`)
7.  $f_i$  and  $f_j$  share neither head nor tail

This approach serves three main purposes:

1. It informs the model whether the head and tail entities have shared neighbor entities.
2. It serves as an implicit method of encoding paths as we just need to know about shared entities, and their order in a path. For example, we can have a pair of length-2 paths where the involved entities and relations are the same but their directions are different as follows:

- $\langle e_1, r_1, e_2 \rangle$  and  $\langle e_2, r_2, e_3 \rangle$
- $\langle e_1, r_1, e_2 \rangle$  and  $\langle e_3, r_2, e_2 \rangle$

Our approach can distinguish between these two and any other combinations of directions.

3. It implicitly captures the relative distance between all entity pairs. Most importantly, it informs the model of 1-hop neighbors and neighbors farther hops away from the target head entity (or from the tail entity).

This implicit knowledge of paths, distance, and shared neighborhood is instrumental to correctly predicting the final relation. Figure 2 shows an example of finding four kinds of connections.

### Connection-Biased Attention in Transformer Encoder

Transformers create key, query, and value vectors  $\mathbf{K}$ ,  $\mathbf{Q}$ ,  $\mathbf{V}$  for tokens by a linear transformation with corresponding learnable weight matrices  $\mathbf{W}^K$ ,  $\mathbf{W}^Q$ ,  $\mathbf{W}^V$  (Vaswani et al. 2017). With slight abuse of notation, we represent any two tokens in an input sequence at  $x_i$  and  $x_j$ . We define the connection-biased similarity between  $x_i$  and  $x_j$  as:

$$\alpha_{ij} = \frac{(\mathbf{W}^Q \mathbf{x}_i)^\top (\mathbf{W}^K \mathbf{x}_j + \mathbf{c}_{ij}^K)}{d_y} \quad (4)$$

Here,  $\mathbf{c}_{ij}^K$  is the Key-specific bias vector for connection type  $c_{ij}$ . The corresponding output vector  $\mathbf{y}_i$  is computed by modifying the attention computation:

$$\mathbf{y}_i = \sum_{j=1}^{m+1} \frac{\exp(\alpha_{ij})}{\sum_{k=1}^{m+1} \exp(\alpha_{ik})} (\mathbf{W}^V \mathbf{x}_j + \mathbf{c}_{ij}^V) \quad (5)$$

Here,  $\mathbf{c}_{ij}^V$  is the Value-specific bias vector for connection type  $c_{ij}$ .

We denote the output sequence as:

$$\mathcal{S}_{out} = [\mathbf{y}^*, \mathbf{y}_{\mathcal{N}}^1, \dots, \mathbf{y}_{\mathcal{N}}^m] \quad (6)$$

The architecture for connection-biased adjacency and the overall architecture of the proposed model is depicted in Figure 3.

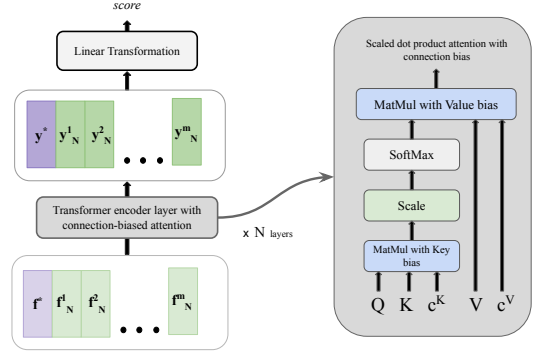


Figure 3: Connection-biased attention computation. The left diagram shows an overview of the input and output sequence, and the right one elaborates the enhanced encoder layer with connection bias.

### Transductive Training

For using our model in the transductive training, we replace  $\text{ROLE}(e)$  with entity-specific learnable vector embeddings. Note that, in this setting, all test entities are seen during training, and a relative representation is not needed. Another change in the transductive version of the model is the distinction between neighbors of *head* and *tail*. We choose  $m$  neighbors from each entity and construct a specific interpretation of the input token sequence in Eq. (3).

Note that in this case,  $\mathbf{f}^*$  is the target entity pair representation without the relation. Two separate vector embeddings specify the roles of each triple (neighbor of *head* or neighbor of *tail*) in the neighborhood. All respective equations are modified to reflect the presence of  $2m + 1$  triples in the input sequence. We make a final change by applying the softmax function to the output of the linear transformation to get a probability distribution over all possible relations.

### Experimental Evaluation

In this section, we evaluate our model on the entity prediction task in the inductive setting and present the performances in three KG datasets (12 versions). Additionally, we present relation prediction results in the transductive setting. All experiments were conducted on Quadro RTX 6000 (with NVLink) GPU with 32 GB memory.

#### Inductive Entity Prediction

**Baselines** NeuralLP (Yang, Yang, and Cohen 2017), DRUM (Sadeghian et al. 2019), and RuleN (Meilicke et al. 2018) are rule-extraction-based methods. GraIL (Teru, Dennis, and Hamilton 2020), CoMPILE (Mai et al. 2021) and TACT (Chen et al. 2021) are graph-based models that use only the subgraph information surrounding a target entity pair and exclude explicit path information. ConGLR (Lin et al. 2022), Report (Liu et al. 2023), LogCo (Pan et al. 2022), NBFNet (Zhu et al. 2021) are graph-based models that utilize path information between target *head* and *tail* entities. Report (Li, Wang, and Mao 2023) uses vanilla Transformers to encode context and path and is the most similar

Table 1: Hits@10 for entity prediction in inductive KG dataset splits; Bold and underlined text represents the best and 2nd best results, respectively. All results are sourced from the original papers except for TACT (taken from ConGLR).

Methods		WN18RR				FB15K-237				NELL995			
		v1	v2	v3	v4	v1	v2	v3	v4	v1	v2	v3	v4
Rule-based	NeuralLP	74.37	68.93	46.18	67.13	52.92	58.94	52.90	55.88	40.78	78.73	82.71	80.58
	DRUM	74.37	68.93	46.18	67.13	52.92	58.73	52.90	55.88	19.42	78.55	82.71	80.58
	RuleN	80.85	78.23	53.39	71.59	49.76	77.82	87.69	85.60	53.50	81.75	77.26	61.35
Graph-based (w/o path)	GraIL	82.45	78.68	58.43	73.41	64.15	81.80	82.83	89.29	59.50	93.25	91.41	73.19
	CoMPILE	83.60	79.82	60.69	75.49	67.64	82.98	84.67	87.44	58.38	<u>93.87</u>	92.77	75.19
	TACT*	84.04	81.63	67.97	76.56	65.76	83.56	85.20	88.31	79.80	88.97	94.02	73.78
Graph-based (with path)	ConGLR	85.64	<u>92.93</u>	70.74	<u>92.90</u>	68.29	85.98	88.61	89.31	<u>81.07</u>	<b>94.92</b>	94.36	<u>81.61</u>
	Report	88.03	85.83	72.31	81.46	71.69	88.91	<u>91.62</u>	<u>92.28</u>	-	-	-	-
	LogCo	90.16	86.73	68.68	79.08	73.90	84.21	86.47	89.22	61.75	93.48	<u>94.44</u>	80.82
	NBFNet	94.80	90.50	<b>89.30</b>	89.00	83.40	<b>94.90</b>	<b>95.10</b>	<b>96.00</b>	-	-	-	-
	<b>CBLiP</b>	<b>97.30</b>	<b>94.10</b>	<u>81.30</u>	<b>96.40</b>	<b>89.30</b>	<u>94.10</u>	84.20	80.10	<b>88.00</b>	93.70	<b>97.70</b>	<b>87.60</b>

Table 2: MRR for entity prediction in inductive KG dataset splits; Bold and underlined text represent best results and 2nd best results, respectively.

Methods		WN18RR				FB15K-237			
		v1	v2	v3	v4	v1	v2	v3	v4
Rule-based	NeuralLP	71.74	68.54	44.23	67.14	46.13	51.85	48.7	49.54
	DRUM	72.46	68.82	44.96	67.27	47.55	52.78	49.64	50.43
	RuleN	79.15	77.82	51.53	71.65	45.97	59.08	<b>73.68</b>	<b>74.19</b>
Graph-based (w/o path)	GraIL	80.45	78.13	54.11	73.84	48.56	62.54	70.35	70.6
	CoMPILE	78.28	79.61	53.97	75.34	50.52	<u>64.54</u>	66.95	63.69
Graph-based (with path)	Report	<u>80.95</u>	<u>82.01</u>	<u>58.38</u>	<u>77.34</u>	<u>53.22</u>	<b>70.62</b>	<u>71.51</u>	<u>71.28</u>
	<b>CBLiP</b>	<b>87.70</b>	<b>87.00</b>	<b>60.50</b>	<b>88.00</b>	<b>59.30</b>	63.70	56.10	53.40

to our model in terms of choice of encoding architecture.

**Datasets** (Teru, Denis, and Hamilton 2020) extracted 12 inductive datasets from three popular KG benchmark datasets – Wordnet, Freebase, and Nell. We present the dataset statistics in supplementary material (?).

**Experimental Setup** For inductive entity prediction, we corrupt a triple by replacing its head/tail with a randomly chosen entity during training. For inference, we want to see how the model ranks variations of corrupted triples by scoring  $\langle h, r, ? \rangle$  or  $\langle ?, r, t \rangle$  and finding the rank of the true triple. We follow the existing literature and use 50 randomly chosen entities from  $\mathcal{E}_{\text{test}}$  to corrupt test triples. The model scores the true triples and the corrupted triples, and the rank of the true triple is recorded. The ranks of all true test triples contribute to finding Hits@ $n$  (ratio of correct hits in the top  $n$  sorted predictions), which is a commonly used metric in rank-based experimental studies. We report Hits@10 and mean reciprocal rank (MRR) and evaluate the performance of our model. The hyperparameter selection is described in the appendix.

**Entity Prediction Results** The results of the entity prediction task according to Hits@10 are presented in Table 1. Our model CBLiP achieves state-of-the-art performance in 7

dataset splits and the performance is consistently better than the rule-extraction and no-path baselines. The model struggles in the Freebase dataset, where a path-based model best utilizes the extremely rich data density. The MRR results of this experiment are presented in Table 2, where we notice a similar trend of dominating performance in Wordnet and not in Freebase.

### Transductive Relation Prediction

The transductive entity prediction task is expensive due to the full vocabulary testing. It is also interchangeable with relation prediction task (Wang, Ren, and Leskovec 2021). We choose relation prediction as it is a faster measure of how well a model can learn to reason over a KG in the transductive setting.

We remove the target relation  $r$  from all triples and generate an input sequence  $S_{in}$ . The output probabilities of each relation type are sorted and the position of true relation  $r$  is retrieved. This serves as the rank of the relation. We report Hits@1, Hits@3, and MRR for this task in Table 3. We demonstrate the results on three primary datasets: WN18RR, FB15K-237, and NELL995. Our supplementary material (?) contains the dataset statistics, hyperparameter configuration, and additional experiments on other datasets.

We compare CBLiP with a neural baseline PathCon

Table 3: Relation prediction in transductive KG datasets; Bold and underlined text represents best and 2nd best results, respectively. All results are sourced from PathCon.

Methods	MRR	WN18RR			FB15K-237			NELL995		
		Hits@1	Hits@3	MRR	Hits@1	Hits@3	MRR	Hits@1	Hits@3	MRR
Models w/o path	TransE	0.784	0.669	0.870	0.966	0.946	0.984	0.841	0.781	0.889
	ComplEx	0.840	0.777	0.880	0.924	0.879	0.970	0.703	0.625	0.765
	DistMult	0.847	0.787	0.891	0.875	0.806	0.936	0.634	0.524	0.720
	RotatE	0.799	0.735	0.823	0.970	0.951	0.980	0.729	0.691	0.756
	SimplE	0.730	0.659	0.755	0.971	0.955	0.987	0.716	0.671	0.748
	QuatE	0.823	0.767	0.852	<u>0.974</u>	<u>0.958</u>	0.988	0.752	0.706	0.783
	DRUM	0.854	0.778	0.912	0.959	0.905	0.958	0.715	0.640	0.740
Model with path	PathCon	<u>0.974</u>	<u>0.954</u>	<b>0.994</b>	<b>0.979</b>	<b>0.964</b>	<b>0.994</b>	0.896	<u>0.844</u>	<u>0.941</u>
	<b>CBLiP</b>	<b>0.976</b>	<b>0.960</b>	<u>0.993</u>	0.971	0.949	<u>0.992</u>	<b>0.919</b>	<b>0.868</b>	<b>0.964</b>

Table 4: Hits@10 and MRR for entity prediction in Wordnet dataset splits; Bold text represents better results. CBLiP-vanilla denotes our model without the connection-biased attention component.

	v1	v2	v3	v4
	Hits@10			
CBLiP	<b>97.30</b>	<b>94.10</b>	<b>81.30</b>	<b>96.40</b>
CBLiP-vanilla	92.00	71.70	69.30	90.90
	MRR			
CBLiP	<b>87.70</b>	<b>87.00</b>	<b>60.50</b>	<b>88.00</b>
CBLiP-vanilla	76.70	66.20	53.50	77.90

(Wang, Ren, and Leskovec 2021) that uses path information explicitly in the model. Other baseline models include a variety of translation-based, factorization-based, and rule-extraction-based models.

Table 3 shows the relation prediction results. We see again that CBLiP performs better than all models that explicitly do not utilize paths. Similarly, it achieves better or more competitive performance against PathCon.

### Ablation Studies

We study the effectiveness of the proposed attention by comparing it with vanilla attention in Transformers. We call this variation CBLiP-vanilla. We run these experiments on the Wordnet dataset and report the results in Table 4. We observe that the performance of the model in both metrics drops significantly when we eliminate the connection bias from the transformer encoder.

### Discussion

We notice that CBLiP performs well in Wordnet and Nell across settings. This could be due to the dense degree of Freebase (on average, each entity has a large number of immediate neighbors), which a path-based model could best utilize.

However, we argue that omitting the use of paths has its strengths. Firstly, computing paths between target pairs on the fly is highly expensive and researchers often fix a path

length and precompute a set of paths between target pairs of test set. However, this can be a limitation in real applications, as one will come across a new target entity pair during test time and has to compute paths between them. This also comes with an extra set of hyperparameter tuning regarding path length, how many paths to use, how to represent inverse relations, and how to aggregate each path.

In contrast to these, our proposed method adds only a handful of new learnable bias vectors for each connection type between tokens. Adding the connection bias to the Transformer encoder block does not add significant computational overhead since these bias computations do not require additional matrix multiplications. The scaled dot product attention module (that we modify by adding connection-bias vectors) retains its complexity of  $\mathcal{O}(N^2d)$  for a sequence of  $N$  tokens with  $d$  feature dimension.

The entity role introduced in our paper is also an intuitive construct and adds negligible overhead to represent all entities in the model. Since these two constructs capture the similarity of involved entities, the model implicitly learns relative distance information along with possible path information between tokens.

## Conclusion and Future Work

We propose CBLiP, a KG link prediction model using inexpensive and intuitive use of entity role and connection-bias in a subgraph. We show the effectiveness of our model in two KG settings in different tasks where CBLiP showcases excellent performance while being intuitive and relatively simpler to its contemporaries. Future work can address the fully inductive setting, where entities *and* relations may be seen at test time that were not present at training time.

### A. Codebase

We provide the pytorch implementation of our code and hyperparameter configurations for all our experiments at: <https://github.com/shoron-dutta/CBLiP>.

Table 5: Statistics of datasets used in inductive link prediction experiments

Table 6: WN18RR

Version	Split	$\mathcal{R}$	$\mathcal{E}$	#TR1	#TR2
v1	train	9	2746	5410	630
	test	9	922	1618	188
v2	train	10	6954	15262	1838
	test	10	2923	4011	441
v3	train	11	12078	25901	3097
	test	11	5084	6327	605
v4	train	9	3861	7940	934
	test	9	7208	12334	1429

Table 7: FB15k-237

Version	Split	$\mathcal{R}$	$\mathcal{E}$	#TR1	#TR2
v1	train	183	2000	4245	489
	test	146	1500	1993	205
v2	train	203	3000	9739	1166
	test	176	2000	4145	478
v3	train	218	4000	17986	2194
	test	187	3000	7406	865
v4	train	222	5000	27203	3352
	test	204	3500	11714	1424

Table 8: NELL-995

Version	Split	$\mathcal{R}$	$\mathcal{E}$	#TR1	#TR2
v1	train	14	10915	4687	414
	test	14	225	833	100
v2	train	88	2564	8219	922
	test	79	4937	4586	476
v3	train	142	4647	16393	1851
	test	122	4921	8048	809
v4	train	77	4922	7546	876
	test	61	3294	7073	731

## B. Inductive Entity Prediction

### B1. Hyperparameter Selection

We select at most  $m$  neighbors from  $k$ -hop subgraphs of entities. We include the closest neighbors first and then allow for neighbors farther away. We select feature dimension  $d$  from the set  $\{20, 32, 40, 64, 80, 128\}$ . We use multi-headed attention with the number of attention heads  $\{2, 4\}$  and select the number of encoder layers from  $\{2, 3, 4\}$ . We use Adam optimizer with learning rate from the set  $\{0.01, 0.001, 0.008, 0.0005\}$ . The specific hyperparameter setting for each dataset can be found in our GitHub repository.

### B2. Dataset

The inductive dataset splits provided by (Teru, Denis, and Hamilton 2020) are widely used for the inductive KG completion tasks. We present the dataset statistics in Table 5.

Table 9: Statistics of transductive KG datasets

	WN18	WN18RR
#nodes	40,943	40,943
#relations	18	11
#training	141,442	86,835
#validation	5,000	2,824
#test	5,000	2,924

	FB15K	FB15K-237
#nodes	14,951	14,541
#relations	1,345	237
#training	483,142	272,115
#validation	50,000	17,526
#test	59,071	20,438

	NELL995	DDB14
#nodes	63,917	9,203
#relations	198	14
#training	137,465	36,561
#validation	3,907	3,897
#test	3,964	3,882

The terms #TR1 and #TR2 refer to the following:

- In train mode:
  - #TR1 refers to  $\mathcal{F}_{\text{train}}$ , the set of triples used for training
  - #TR2 refers to  $\mathcal{F}_{\text{valid}}$ , the set of triples we use for validation
- In test mode:
  - #TR1 refers to  $\mathcal{F}_{\text{test}}$ , the set of triples used as a fact graph (to collect topological and neighborhood data) for inference triples
  - #TR2 refers to  $\mathcal{F}_{\text{infer}}$ , the set of triples to test the model’s inference capabilities

## C. Transductive Relation Prediction

### C1. Datasets

Initial KG experiments were done on Freebase and Wordnet datasets, FB15k and WN18, whose test sets contained inverse triples of training triples. It caused simpler models to perform well during test by memorizing training data. (Toutanova and Chen 2015) proposed a corrected versions of these datasets- FB15K-237 and WN18RR, respectively. We have presented results on FB15k-237, WN18RR, and NELL995 in the main text of our paper.

PathCon (Wang, Ren, and Leskovec 2021) proposed DDB14 dataset which is generated from the disease database. The statistics of all 6 datasets are presented in Table 9.

### C2. Experimental Evaluation

Here, we present results in the other 3 datasets. Table ?? shows CBLiP’s performance in comparison with the relation prediction baseline models. CBLiP performs achieves



Table 10: Transductive relation prediction in KG datasets

Methods	MRR	Hits@1	Hits@3
FB15K			
TransE	0.962	0.940	0.982
ComplEx	0.901	0.844	0.952
DistMult	0.661	0.439	0.868
RotatE	0.979	0.967	0.986
Simple	0.983	0.972	0.991
QuatE	0.983	0.972	0.991
DRUM	0.945	0.945	0.978
PathCon	<b>0.984</b>	<b>0.974</b>	<b>0.995</b>
CBLiP	0.863	0.763	0.962
WN18			
TransE	0.971	0.955	0.984
ComplEx	0.985	0.979	0.991
DistMult	0.786	0.584	0.987
RotatE	0.984	0.979	0.986
Simple	0.972	0.964	0.976
QuatE	0.981	0.975	0.983
DRUM	0.969	0.956	0.980
PathCon	<b>0.993</b>	<b>0.988</b>	<b>0.998</b>
CBLiP	0.991	0.985	0.996
DDB14			
TransE	0.966	0.948	0.980
ComplEx	0.953	0.931	0.968
DistMult	0.927	0.886	0.961
RotatE	0.953	0.934	0.964
Simple	0.924	0.892	0.948
QuatE	0.946	0.922	0.962
DRUM	0.958	0.930	0.987
PathCon	0.980	0.966	0.995
CBLiP	<b>0.981</b>	<b>0.967</b>	<b>0.995</b>

competitive performance in WN18 and DDB14. It struggles in FB15K. We think the high triples to entities ratio in this dataset could be the reason for this. The dense neighborhood information is best captured with a path-based model like PathCon (Wang, Ren, and Leskovec 2021).

## References

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Chen, J.; He, H.; Wu, F.; and Wang, J. 2021. Topology-aware correlations between relations for inductive link prediction in knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6271–6278.

Kazemi, S. M.; and Poole, D. 2018. Simple embedding for link prediction in knowledge graphs. *Advances in neural information processing systems*, 31.

Li, J.; Wang, Q.; and Mao, Z. 2023. Inductive relation prediction from relational paths and context with hierarchical transformers. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Lin, Q.; Liu, J.; Xu, F.; Pan, Y.; Zhu, Y.; Zhang, L.; and Zhao, T. 2022. Incorporating context graph with logical reasoning for inductive relation prediction. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 893–903.

Liu, H.; Chen, Y.; He, P.; Zhang, C.; Wu, H.; and Zhang, J. 2023. An inductive knowledge graph embedding via combination of subgraph and type information. *Scientific Reports*, 13(1): 21228.

Mai, S.; Zheng, S.; Yang, Y.; and Hu, H. 2021. Communicative message passing for inductive relation reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4294–4302.

Meilicke, C.; Fink, M.; Wang, Y.; Ruffinelli, D.; Gemulla, R.; and Stuckenschmidt, H. 2018. Fine-grained evaluation of rule-and embedding-based systems for knowledge graph completion. In *The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I 17*, 3–20. Springer.

Nickel, M.; Tresp, V.; and Kriegel, H.-P. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 809–816.

Niu, G.; Zhang, Y.; Li, B.; Cui, P.; Liu, S.; Li, J.; and Zhang, X. 2020. Rule-guided compositional representation learning on knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 2950–2958.

Pan, Y.; Liu, J.; Zhang, L.; Zhao, T.; Lin, Q.; Hu, X.; and Wang, Q. 2022. Inductive relation prediction with logical reasoning using contrastive representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 4261–4274.

Sadeghian, A.; Armandpour, M.; Ding, P.; and Wang, D. Z. 2019. Drum: End-to-end differentiable rule mining on knowledge graphs. *Advances in Neural Information Processing Systems*, 32.

Sun, Z.; Deng, Z.-H.; Nie, J.-Y.; and Tang, J. 2018. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *International Conference on Learning Representations*.

Teru, K.; Denis, E.; and Hamilton, W. 2020. Inductive relation prediction by subgraph reasoning. In *International Conference on Machine Learning*, 9448–9457. PMLR.

Toutanova, K.; and Chen, D. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, 57–66.

Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; and Bouchard, G. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*, 2071–2080. PMLR.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

- Wang, H.; Ren, H.; and Leskovec, J. 2021. Relational message passing for knowledge graph completion. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1697–1707.
- Wang, Q.; Wang, H.; Lyu, Y.; and Zhu, Y. 2021. Link Prediction on N-ary Relational Facts: A Graph-based Approach. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 396–407.
- Yang, B.; Yih, S. W.-t.; He, X.; Gao, J.; and Deng, L. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*.
- Yang, F.; Yang, Z.; and Cohen, W. W. 2017. Differentiable learning of logical rules for knowledge base reasoning. *Advances in neural information processing systems*, 30.
- Zhang, S.; Tay, Y.; Yao, L.; and Liu, Q. 2019. Quaternion knowledge graph embeddings. *Advances in neural information processing systems*, 32.
- Zhu, Z.; Zhang, Z.; Xhonneux, L.-P.; and Tang, J. 2021. Neural bellman-ford networks: A general graph neural network framework for link prediction. *Advances in Neural Information Processing Systems*, 34: 29476–29490.