# Adversarial Explanations for Knowledge Graph Embeddings

**Patrick Betz**, **Christian Meilicke** and **Heiner Stuckenschmidt**
University of Mannheim, Research Group Data and Web Science
{patrick, christian, heiner}@informatik.uni-mannheim.de

## Abstract

We propose a novel black-box approach for performing adversarial attacks against knowledge graph embedding models. An adversarial attack is a small perturbation of the data at training time to cause model failure at test time. We make use of an efficient rule learning approach and use abductive reasoning to identify triples which are logical explanations for a particular prediction. The proposed attack is then based on the simple idea to suppress or modify one of the triples in the most confident explanation. Although our attack scheme is model independent and only needs access to the training data, we report results on par with state-of-the-art white-box attack methods that additionally require full access to the model architecture, the learned embeddings, and the loss functions. This is a surprising result which indicates that knowledge graph embedding models can partly be explained post hoc with the help of symbolic methods.

## 1 Introduction

Knowledge graphs commonly suffer from incompleteness and many different methods have been proposed to complete the missing information which is defined as knowledge graph completion (KGC). The currently dominating techniques are based on the use of sub-symbolic representations and the respective approaches are termed knowledge graph embedding (KGE) models. A large family of KGE models has been proposed so far, e.g., [Yang *et al.*, 2015; Trouillon *et al.*, 2016; Dettmers *et al.*, 2018] and much research has been concerned with improving their performance and multiple aspects of the learning process [Ruffinelli *et al.*, 2020]. Driven by the emerging need to provide more interpretability for deep learning models and latent approaches in general, efforts to understand the causes of certain aspects of model behavior are also made in the context of KGE models [Rim *et al.*, 2021].

Adversarial attacks in machine learning, on the other hand, are concerned with decreasing the model's predictive quality by corrupting the data. In [Bhardwaj *et al.*, 2021], the authors apply this concept to KGC where an attack is defined as the modification of a single triple at training time to cause model failure at test time. While the authors motivate their work with the importance of identifying vulnerabilities and studying robustness, the task can also be understood in terms of interpretability, that is, as finding the triple that is the cause for a certain prediction. Such a triple (or a set of triples) can be referred to as an explanation.

Within this work, we propose an approach that is based on the idea to find a logical explanation for a prediction made by a KGE model. To find this explanation, we first apply a rule learning approach to learn a logical theory that describes general regularities in the knowledge graph. Then we use an efficient form of abductive reasoning to find the triple that is, together with the theory, the best explanation for the given prediction. The explanation is used as the triple that we delete or modify in the context of adversarial attacks. As we measure the quality of the explanation by its effectiveness in this context, we call it an *adversarial explanation*. We find that our method is competitive to state-of-the-art attacks while only requiring access to the training data.

The most important contributions of the work are summarized in the following. (1) We propose a novel method for constructing adversarial attacks against KGE by using abductive reasoning. Contrary to prior work, our method does not require access to the model architecture, the embeddings, or the loss function. (2) We propose a revised evaluation protocol after identifying interpretation issues in regard to attack quality in the most recent protocol used in [Bhardwaj *et al.*, 2021]. (3) We conduct extensive experiments under the new protocol and present reliable results including our method and the current state-of-the-art approaches. (4) We demonstrate that our method does not only identify an influential triple but also yields a useful symbolic explanation of the underlying statistical regularities.[1]

## 2 Preliminaries

### 2.1 Knowledge Graph Completion

A knowledge graph $G$ is a set of triples $G \subseteq E \times R \times E$ where $E$ denotes a set of entities and $R$ denotes a set of relations. The data structure can also be understood as a directed graph where each node refers to an entity and each edge is labeled by a relation. A triple $(s, r, o)$ represents the fact that the subject $s$ is in $r$-relation to $o$. From a logical point of

[1] Code and resources are available at: https://web.informatik.uni-mannheim.de/AnyBURL

view, an entity corresponds to a constant, a relation to a binary predicate, and a triple in the graph corresponds to an atomic fact $r(s, o)$ that resulted from grounding the binary predicate $r$ with the constants $s$ and $o$.

The collection of facts expresses our *knowledge* about a certain domain. This knowledge is in most cases incomplete. This means there exists a set of triples $G^* \subseteq E \times R \times E$ with $G^* \cap G = \emptyset$ describing correct but unknown triples and the field of KGC is concerned with finding these facts. While it is possible to deduce new relevant knowledge by looking at external resources, the majority of the proposed methods focus on the specific problem to use only the data already encoded in $G$.

The standard technique to evaluate these methods is based on the idea to ask for a correct candidate of an incomplete triple $(s, r, ?)$ or $(?, r, o)$. Possible answers to these queries form a candidate ranking. The common metrics hits@k or MRR (mean rank reciprocal) are based on the position of the correct candidate within the ranking. Datasets are usually split into training, validation and test sets where evaluation takes place by forming queries as described above for all the triples in the test set. Commonly, filtered variants of the evaluation metrics are presented in the experimental results where the known triples in the data are used to remove all known other true answers from a candidate ranking. We will later see that using an appropriate filter technique is, especially with respect to problem of measuring attack quality, important to get a realistic impression of the final results.

### 2.2 Sub-Symbolic and Symbolic KGC

KGE models or sub-symbolic methods represent the entities and relations of $G$ in a low dimensional vector space. They are characterised by a scoring function, which outputs real-valued confidences for individual facts, and a training method. A vast amount of research in the KGC domain is centered around KGE models and on how to train their embeddings. The DistMult [Yang *et al.*, 2015] model defines triple scores by a trilinear dot product and is extended by ComplEx [Trouillon *et al.*, 2016] towards expressing non-symmetric relationships. A number of more complicated specifications exists, for instance the scoring function of ConvE [Dettmers *et al.*, 2018] is based on convolutions. An overview over differengt KGE models and their predictive performance can be found in [Rossi *et al.*, 2021].

While latent approaches dominate KGC, the problem can also be solved with symbolic methods. An example of such an approach is called AnyBURL [Meilicke *et al.*, 2019], an anytime algorithm for learning a large set of rules that cover a substantial portion of the important regularities of the dataset. Each rule has a confidence, formally defined in [Galárraga *et al.*, 2013], which is the number of correct predictions divided by the number of all predictions of the rule.

Once a rule set $\Phi$ has been learned, AnyBURL checks for a given query $(s, r, ?)$ for each rule $\phi \in \Phi$ if there is an $o$ such that a single application of $\phi$ allows to derive $r(s, o)$. If this is the case, $o$ appears in the candidate ranking and its score is defined as the confidence of the respective rule $\phi$. If several rules generate the same candidate, its confidence is, in the standard setting of AnyBURL, the highest confidence of the respective rules.

## 3 Problem Statement

In our definition of an adversarial attack we follow the framework proposed in [Bhardwaj *et al.*, 2021]. The idea of an attack is to propose a minimal modification of a knowledge graph, i.e., the training set, with a maximal negative effect on the score of a correct prediction $t = (s_t, r_t, o_t)$ that would be ranked high without that modification. We call $t$ in the following the target of the attack. There are two different types of modifications, deleting triples or adding triples. We call the first scenario an adversarial deletion and the second scenario an adversarial addition. Throughout the work we also denote the former the *Del* setting and the latter the *Add* setting.

We measure the impact of the attack by comparing the standard metrics MRR and hits@k before and after the attack. In particular, we train a model with the original training set and apply it on the test set. Then we select a small subset from the test set for which the model achieved a good predictive performance. This subset is our set of target triples.

As we focus on small perturbations, we restrict the notion of an attack to the deletion or addition of a single triple, that is, for each of the target triples we compute a deletion and an addition triple. Then we remove the deletion set from the training set (add the addition set to the training set) and train the model again. We apply the retrained model to the completion tasks of the target set and measure the degradation of MRR and hits@k. Note that this approach is based on a batch mode where an attack of a target triple $t$ might have an impact on another target triple $t'$. If we restrict the target sets to be relatively small, such dependencies will occur rarely or not at all and can be neglected. Running a single experiment for each single target triple would be infeasible.

To make our results comparable to the methods proposed and evaluated in [Pezeshkpour *et al.*, 2019; Bhardwaj *et al.*, 2021] we consider only attacks $a = (s_a, r_a, o_a) \in G$ on a target triple $t = (s_t, r_t, o_t)$ such that $s_a = s_t$ or $s_a = o_t$ or $o_a = s_t$ or $o_a = o_t$. In other words, the attack and target need to have at least one entity in common.

In an attack scenario one can distinguish between white-box and black-box methods. To our best knowledge, so far only white-box methods have been proposed, see for example [Pezeshkpour *et al.*, 2019; Bhardwaj *et al.*, 2021; Lawrence *et al.*, 2021]. These methods have full access to the model that has been learned. This refers to two different aspects: (i) the embeddings that have been learned are accessible and (ii) the method that was used to learn these embeddings (e.g., the scoring and loss functions) is known.

Black-box methods, on the other hand, do not have access to (i) or (ii). These methods must be based on a generic concept that explains how a triple or a set of triples results into the prediction of $t$. Within this work, we propose a black-box method based on abductive reasoning [Bylander *et al.*, 1991], which can be characterized as the search for an explanation for $t$. When such an explanation, which is a set of triples, has been found, the removal (or modification) of one of its members should suppress the prediction of $t$ or lower its position in the ranking if there are several explanations.

## 4 Related Work

The approach that we propose within this paper is focused on the intersection of learning representations of KGE models and rule-based approaches. The combination of symbolic and KGE models is studied in the literature from different perspectives. [Guo *et al.*, 2016; Guo *et al.*, 2018] inject rules as background knowledge into the training of the latent representations and [Zhang *et al.*, 2019b] propose an alternating training scheme. However, these works focus on identifying or exploiting differences between the approaches to improve KGC performance whereas the presumption of this work points towards their similarity. Indeed, a result in [Meilicke *et al.*, 2021], who explore an ensemble model, suggests similarities between the symbolic and sub-symbolic models. In fact, the authors argue that the KGE models remain within the language scope of the rule-based approach. This result indicates that it might be possible to explain a prediction made by a KGE model with the help of a rule-based approach.

Methods tailored towards understanding KGE or deep learning based models in general are related to adversarial attacks when they focus on searching for influential data points. For example, [Hanawa *et al.*, 2021] investigate relevance metrics for instance-based explanations, i.e., explaining model predictions by similar training instances, and [Charpiat *et al.*, 2019] aim to express similarities from the neural network perspective. [Koh and Liang, 2017] apply influence functions in the context of image classification to trace back model predictions towards individual training images. [Lawrence *et al.*, 2021] estimate individual influences of knowledge graph facts by first tracking for every training instance the gradient-based updates induced to its parameters. The influence of a triple $t'$ on a target triple $t$ is then estimated by the difference of the original score of $t$ and the score calculated in regard to parameters where the accumulated updates induced by $t'$ are subtracted (rolled back).

In the context of adversarial attacks, influential data points are identified with the motivation to cause the most harmful effect on a particular model prediction by perturbing the data. [Pezeshkpour *et al.*, 2019] study the robustness of KGE models by defining adversarial modifications in the context of KGC queries. The change in the triple score of an attack is estimated by using a Taylor approximation and the attacking triples are selected by a parameterized decoder that maps the maximal change in scores vectors back to entities and relations in the embedding space. [Zhang *et al.*, 2019a] investigate data poisonous attacks against KGE models by defining *Direct* attacks. An embedding shifting vector of a target triple is defined as the negative gradient of the scoring function and the attacking triples are selected by calculating a perturbation benefit score based on this vector for every candidate triple. The methods proposed in [Bhardwaj *et al.*, 2021] represent the current state-of-the-art and are used as the main comparison in our experimental section. The attack setting is based on instance attribution methods and the main specifications are based on selecting the attacking triples by computing similarities to the target triples as in instance-based explanations [Hanawa *et al.*, 2021]. These similarities are either based on the embeddings or the gradients of the loss functions with respect to the candidate triples.

The aforementioned works propose white-box models which require access to the KGE architecture, the embeddings, and optionally the loss functions when gradients are computed whereas our method does not rely on any of these. Please note that this also closely resembles a more realistic scenario where an attacker might find an entry point into a system for manipulating the data but is only vaguely informed about the respective methods and protocols used.

## 5 Method

We propose a rule-based approach to solve the problem of computing an adversarial attack without having any access to the KGE model. It is based on the idea of abductive reasoning. Abductive reasoning [Mayer and Pirri, 1993] is concerned with the problem of finding an explanation $\mathcal{E}$ for an observation $t$ given a theory $\Phi$ with $\Phi \not\models t$ and $\Phi \not\models \neg t$ such that $\Phi \cup \mathcal{E} \models t$ and $\Phi \cup \mathcal{E}$ is consistent. An explanation $\mathcal{E}$ is minimal if for each $\mathcal{E}' \subset \mathcal{E}$ we have that $\Phi \cup \mathcal{E}' \not\models t$. We refer to minimal explanations when mentioning explanations in the following paragraph.

While abductive reasoning is in general intractable (NP-hard) as argued in [Bylander *et al.*, 1991], we propose a method that is incomplete but allows to efficiently compute a good explanation. In the following we present the required definitions to employ the idea in settings where rules are associated with confidences and inference is not performed in a model-theoretic notion of entailment. To that end, let $t$ be the target triple, let $\Phi$ be a set of rules that describe the regularities in the given knowledge graph $G$, and let $\mathcal{E} \subseteq G$ be a set of triples. Finally, let $min_c(\Psi)$ refer to the minimal confidence in a set of rules $\Psi \subseteq \Phi$.

**Definition 1** (Best explanation)**.** *$\mathcal{E}$ is a best explanation with respect to $\models$ iff there exists a $\Psi \subseteq \Phi$ with $\Psi \cup \mathcal{E} \models t$ such that for any other $\Psi' \subseteq \Phi$ and $\mathcal{E}' \subseteq G$ with $\Psi' \cup \mathcal{E}' \models t$ we have that $min_c(\Psi) \geq min_c(\Psi')$ .*

The definition is motivated by the idea that the strength of the explanation is determined by the weakest rule required to entail the target, however, the best explanation is not necessarily unique. Furthermore, as mentioned in Section 2.2, the application of the rule set $\Phi$ performed by AnyBURL is based on the one-time application of each rule:

**Definition 2** (One-step entailment $\models_1$)**.** *The triple $t$ is one-step entailed by $\Psi \cup \mathcal{E}$, written as $\Psi \cup \mathcal{E} \models_1 t$, iff there is a rule $h \leftarrow b$ in $\Psi$ for which a grounding exists where the grounded rule body $b$ is in $\mathcal{E}$ and the grounded head $h$ is equal to $t$.*

Contrary to general entailment, multiple or recursive applications of the same rule are not considered. Nevertheless, from Definition 2 it follows that $\Psi \cup \mathcal{E} \models t$ if $\Psi \cup \mathcal{E} \models_1 t$ for each $\mathcal{E} \subseteq G$ but the opposite direction does not hold.

In the procedure of our approach, we first have to learn a rule set $\Phi$ which has to be computed only once for any attack related to the same data set. We use AnyBURL and restrict it to only learn the following rule types:

$$h(X,Y) \leftarrow r(X,Y) \tag{1}$$

$$h(X,Y) \leftarrow r_1(X,Z) \wedge r_2(Z,Y) \tag{2}$$

$$h(X,e_1) \; \leftarrow r(X,e_2), \tag{3}$$

where $h$, $r$, $r_1$, and $r_2$ denote relations, $e_1$ and $e_2$ refer to entities (constants), and $X, Y, Z$ to variables. Constants and variables may appear at flipped positions. We only use these short rules as the considerations in [Meilicke *et al.*, 2021] indicate that longer paths might not be well represented by KGE models and are thus not helpful in explaining their behavior.

Given a target $t = (s_t, r_t, o_t)$ predicted by a KGE model as answer to a completion task $(s_t, r_t, ?)$ or $(?, r_t, o_t)$, we check for each rule $\phi \in \Phi$ if it predicts $t$ by grounding $\phi$ with respect to the target, that is, we check if the rule body is true after setting $X = s_t$ and $Y = o_t$. If the body is true, i.e., the respective triples exist in train, we say that the respective rule ***fires***. For rules of type (1) and (3) the variable binding results in a completely grounded and unique rule body with one triple and a simple look-up in the graph suffices to check if the rule fires. For rules of type (2) we need to check additionally if there exists a grounding for $Z$ such that $b_1(s_t, Z)$ and $b_2(Z, o_t)$ are true in $G$.

Let $\phi^*$ be the rule with highest confidence in the set of rules that fire. We set $\mathcal{E}$ to the triple of the unique body grounding if $\phi^*$ is of type (1) or (3) and we set $\mathcal{E}$ randomly to one body grounding if $\phi^*$ is of type (2). Therefore, $\mathcal{E}$ is a set of one or two triples. By following this procedure we construct an explanation $\mathcal{E}$ such that $\{\phi^*\} \cup \mathcal{E} \models_1 t$. Due to $\phi^* \in \Phi$ we have $\Phi \cup \mathcal{E} \models_1 t$ and due to the correctness of $\models_1$ we have also $\Phi \cup \mathcal{E} \models t$. The explanation is based on the most confident rule, therefore, for every other $\phi' \neq \phi^* \in \Phi$ with alternative explanation $\mathcal{E}'$ and $\{\phi'\} \cup \mathcal{E}' \models_1 t$ we have that $min_c(\{\phi^*\}) \geq min_c(\{\phi'\})$. It follows that $\mathcal{E}$ is the *best explanation* with respect to $\models_1$. However, it is not guaranteed to be the best explanation with respect to $\models$ . In that sense our approach can be understood as an efficient way of computing a good explanation which might not always be the best explanation.

In the **delete setting** we simply suppress the explanation $\mathcal{E}$ by deleting the triple in $\mathcal{E}$ from the training set or we randomly delete one when $\mathcal{E}$ contains two triples. Let $(s_e, r_e, o_e)$ denote that triple in the following paragraph.

In the **addition setting** we add a perturbation of $(s_e, r_e, o_e)$ to the training set by applying the following strategy. We select an entity $\alpha$ by first randomly selecting a triple from the graph and subsequently selecting the head or tail entity from this triple with equal probability. From the two possible perturbations $(\alpha, r_e, o_e)$ or $(s_e, r_e, \alpha)$, we choose the one which is in the neighbourhood of the target triple sharing an entity with the target. Assume $(\alpha, r_e, o_e)$ is the chosen triple in the neighbourhood of the target. We then check if there exists any triple in the training set satisfying $(\alpha, r_e, x)$ with $x \in E$ where $E$ is the set of all entities. If such a triple exists, we repeat the whole process otherwise we add $(\alpha, r_e, o_e)$ to the training set. To summarize, we perturb the true explanation for the target (the deletion triple) to a senseless statement about one of the entities in the target.

| | Original | No Attack | Attack |
|---|---|---|---|
| MRR | 1.0 | 0.756 | 0.619 |
| h@1 | 1.0 | 0.605 | 0.475 |

Table 1: The degradation in MRR for ComplEx on FB15k-237 in the *Del* setting when only following the existing protocol without actually performing an attack and without retraining (middle column).

Our general approach is based on the assumption that (1) the latent representation of a KGE model implicitly encodes the statistical regularities of the given knowledge graph and (2) these regularities can be explicitly represented in terms of symbolic rules. Thus, the rule with the highest confidence capable of predicting the target also should point to a triple which is influential for the KGE prediction.

# 6 Experiments

We present experimental details and results in the following sections. Moreover, we argue that the existing evaluation protocol should be revisited by showing that the largest part of the MRR degradation is caused by the protocol instead of the attack. Therefore, we propose a new protocol, re-evaluate the best performing specifications of the related work [Bhardwaj *et al.*, 2021] and compare them to our approach.

## 6.1 Experimental Settings

The general structure of our experiments follows the procedure in [Bhardwaj *et al.*, 2021]. A KGE model is trained and a subset of target triples from the test set for which the model achieved a high filtered MRR is obtained, the attack is performed by a perturbation of the training data, and in the last step the model is retrained on the perturbated training data and an evaluation is performed on the target triples.

To have a fair comparison to existing literature, we base the experiments and the comparison to related work on the public implementation of [Bhardwaj *et al.*, 2021] and we inherit the respective settings. This also holds for the methods developed in [Zhang *et al.*, 2019a]. In particular, 100 target triples which had an MRR of 1 in both directions are randomly selected from the test set. We encountered a substantial variance when running their code several times and therefore we report averages over 5 runs.

We use the KGE models ComplEx [Trouillon *et al.*, 2016], DistMult [Yang *et al.*, 2015] and ConvE [Dettmers *et al.*, 2018] and the same datasets as [Bhardwaj *et al.*, 2021], i.e., we use the common benchmarks WN18RR and FB15k-237. We compare against the best methods proposed in [Bhardwaj *et al.*, 2021]. That is, we include the feature-based and gradient-based models in [Bhardwaj *et al.*, 2021] and *Direct* attack [Zhang *et al.*, 2019a].

We integrated our own approach into the evaluation framework provided by [Bhardwaj *et al.*, 2021]. We have set the time available for AnyBURL to learn the rule set to 100 seconds. After that the computation of a single deletion and addition attack requires around 0.05 seconds. This is slightly slower than the feature-based methods and faster than the gradient-based methods in [Bhardwaj *et al.*, 2021]. Further

| | ComplEx | | | | DistMult | | | | ConvE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Del | | Add | | Del | | Add | | Del | | Add | |
| Approach | h@1 | MRR | h@1 | MRR | h@1 | MRR | h@1 | MRR | h@1 | MRR | h@1 | MRR |
| GC (cos) | **0.193** | 0.273 | 0.789 | 0.887 | **0.152** | **0.237** | 0.794 | 0.894 | **0.122** | **0.174** | 0.854 | 0.927 |
| GD (dot) | 0.222 | 0.296 | 0.794 | 0.892 | 0.182 | 0.259 | 0.797 | 0.896 | 0.261 | 0.302 | 0.885 | 0.942 |
| GL ($l_2$) | 0.208 | 0.284 | 0.803 | 0.894 | 0.175 | 0.250 | 0.796 | 0.895 | 0.129 | 0.175 | 0.831 | 0.931 |
| Cos Metric | 0.201 | 0.282 | 0.813 | 0.893 | 0.165 | 0.246 | 0.793 | 0.894 | 0.981 | 0.981 | 0.993 | 0.997 |
| Dot Metric | 0.915 | 0.928 | 0.927 | 0.958 | 0.904 | 0.917 | 0.928 | 0.963 | 0.934 | 0.937 | 0.966 | 0.983 |
| $l_2$ Metric | 0.199 | **0.272** | 0.788 | 0.890 | 0.162 | 0.243 | 0.785 | 0.890 | 0.937 | 0.960 | 0.996 | 0.998 |
| Direct | 0.949 | 0.958 | 0.964 | 0.974 | 0.965 | 0.974 | 0.984 | 0.989 | 0.786 | 0.789 | 1.0 | 1.0 |
| Rand | 0.880 | 0.890 | 0.971 | 0.982 | 0.906 | 0.916 | 0.994 | 0.995 | 0.895 | 0.901 | 0.999 | 0.999 |
| Rerun | 0.972 | 0.981 | 0.976 | 0.983 | 0.988 | 0.992 | 0.991 | 0.993 | 0.995 | 0.998 | 0.998 | 0.999 |
| Ours | 0.197 | 0.278 | **0.760** | **0.874** | 0.167 | 0.244 | **0.757** | **0.876** | 0.123 | 0.175 | **0.758** | **0.878** |

Table 2: Results for WNRR. All results are averages over five runs. Lower is better. The original MRR is 1.0 in all specifications.

details about KGE training and the code for running all experiments can be found in the supplementary material.

## 6.2 Evaluation Protocol

Unfortunately, the devil is in the detail in regard to ranking based evaluation metrics involving filtering. We will demonstrate why the existing evaluation protocol results into misleading numbers and propose a revised version. Let $T$ be the set of target triples and let $A$ be the set of attacking triples.

The **existing protocol** in [Bhardwaj *et al.*, 2021] selects $T$ by searching for a subset of triples from the test set which achieve a high filtered MRR in both directions. At this point, the full original train, valid, and test splits are used for filtering which is the standard procedure. The important aspect is the definition of the filter set after the attack; we focus on the deletion setting in the following. After selecting $A$, (1) $A$ is *removed* from the filter set and (2) likewise all the test triples from the original test split which are not contained in $T$ are removed from the filter set. The impact of the attack is then measured via the MRR calculated with the adjusted filter set. Please note that (1) allows for a trivial baseline that simply searches in the training set for a matching triple that has a higher score than a target $t$ and adds it to $A$. As it is removed from the filter set, by construction, it will be ranked higher than $t$ eventually degrading the MRR. Furthermore, (2) decreases the MRR by simply having a smaller filter set after the attack. These two effects lead to a degradation of the original MRR without a connection to any attacking scheme. We demonstrate this in Table 1 where in the middle column we apply the described protocol *without* actually performing an attack, i.e., the model is not retrained on the perturbed data after selecting $A$. Most of the degradation effect is caused by the protocol and not by the attack.

In the **new protocol**, we maintain filtering in general as it prevents model punishing when true candidates are ranked better than the current query candidates. Therefore, in the deletion setting we do not modify the filter set after training the original model which keeps the MRR on its original value when no attack is performed in contrast to Table 1. In the addition setting, we follow the same procedure but we additionally augment the filter set with the triples in $A$. This is important as otherwise models would be rewarded with overfitting the data in $A$, ranking these triples higher than the triples in $T$ which would lead to a misleading MRR degradation during test time.

## 6.3 Results

Table 2 and 3 show the results for WN18RR and FB15k-237, respectively.[2] The best (second best) results are marked in bold (underlined). The first (second) part of the tables refers to the gradient (feature)-based similarity metrics in [Bhardwaj *et al.*, 2021]. The third part contains *Direct* attack [Zhang *et al.*, 2019a], as well as a baseline which randomly removes a triple from the neighbourhood of the target *(Rand)*, and a baseline that simply retrains the model without performing an attack *(Rerun)*.

None of the previous state-of-the-art attacks is clearly dominant and when results are averaged over 5 runs and the revised protocol is used, the differences between approaches are smaller than reported in previous work. Our proposed method generates in 17 from 24 settings at least the second best result without having access to the respective model architecture, embeddings, or the loss functions. We achieve the strongest result for ConvE on WN18RR where the MRR degradation is 7.3 (4.9) percentage points (PP) higher for Hits@1 (MRR) compared to the second strongest method in the *Add* setting. On the other hand, we achieve the weakest results for ConvE on FB15k-237 where the reported MRR degradation is 2.9 PP lower than the best performing method. Overall, our results are competitive to recent state-of-the-art.

Please also note the absolute results in both tables and in particular the comparisons to the *Rerun* baseline. Although this baseline achieves a weaker degradation than the attacks in all cases, the differences are marginal in some settings on the FB15k-237 dataset (Table 3). Contrary to previous work,

[2] The codebase of [Bhardwaj *et al.*, 2021] contains an implementation for the GR method [Lawrence *et al.*, 2021] which we exclude in the final paper version in accordance with the GR authors who pointed us towards potential mistakes in that implementation.

| | ComplEx | | | | DistMult | | | | ConvE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Del | | Add | | Del | | Add | | Del | | Add | |
| Approach | h@1 | MRR | h@1 | MRR | h@1 | MRR | h@1 | MRR | h@1 | MRR | h@1 | MRR |
| GC (cos) | **0.734** | **0.814** | 0.776 | 0.853 | 0.74 | 0.823 | 0.781 | 0.856 | 0.720 | 0.784 | 0.716 | 0.805 |
| GD (dot) | 0.739 | 0.818 | 0.776 | 0.849 | 0.74 | **0.817** | 0.776 | 0.848 | 0.750 | 0.802 | 0.703 | 0.793 |
| GL ($l_2$) | 0.758 | 0.835 | 0.785 | 0.854 | 0.759 | 0.835 | 0.773 | 0.849 | 0.743 | 0.786 | 0.715 | 0.806 |
| Cos Metric | 0.747 | 0.829 | 0.789 | 0.859 | **0.730** | 0.828 | 0.764 | 0.845 | 0.768 | 0.789 | 0.714 | 0.804 |
| Dot Metric | 0.806 | 0.871 | 0.787 | 0.864 | 0.778 | 0.856 | 0.813 | 0.874 | 0.745 | 0.825 | 0.723 | 0.812 |
| $l_2$ Metric | 0.739 | 0.825 | **0.772** | **0.845** | 0.743 | 0.827 | **0.738** | **0.828** | **0.682** | 0.777 | 0.719 | 0.808 |
| Direct | 0.738 | 0.822 | 0.793 | 0.859 | 0.754 | 0.833 | 0.782 | 0.853 | 0.753 | 0.806 | **0.679** | **0.779** |
| Rand | 0.810 | 0.873 | 0.806 | 0.863 | 0.796 | 0.869 | 0.800 | 0.870 | 0.755 | 0.818 | 0.695 | 0.788 |
| Rerun | 0.773 | 0.853 | 0.800 | 0.865 | 0.795 | 0.865 | 0.795 | 0.870 | 0.762 | 0.821 | 0.739 | 0.820 |
| Ours | 0.735 | 0.817 | 0.781 | 0.856 | 0.734 | 0.822 | 0.763 | 0.843 | 0.684 | **0.775** | 0.708 | 0.799 |

Table 3: Results for FB15k-237. All results are averages over five runs. Lower is better. The original MRR is 1.0 in all specifications.

we therefore conclude in this work that the efficacy of the attack schemes is to a substantial part overshadowed by the effect of re-training the models for this dataset. Remarkably, this changes when looking at the WN18RR results (Table 2). For instance, in the *Del* setting the *Rerun* baseline achieves no degradation whereas the best attack schemes lead to MRR values of around 0.12-0.2. This means that attack efficacy and KGE robustness is highly dataset specific. We will use our approach in the next section to provide an explanation why the attacks have more influence on the WN18RR dataset.

### 6.4 Understanding the Explanation

In the following, we present two representative examples from the FB15k-237 dataset and one from WN18RR. While related work commonly interprets the attacking/influential triple as a standalone explanation for the prediction (here the target), a more complete view is provided when also regarding the underlying theory as described in Section 5. In particular, we show the target $t_i$, the explanation $\mathcal{E}_i$, and the rule $\phi_i^*$ with its confidence in brackets that led to the explanation.

(E1) $t_1$: *nationality(Sawashiro, Japan)*
$\mathcal{E}_1$: *born(Sawashiro, Tokyo)*, *located(Tokyo, Japan)*
$\phi_1^*$: *nationality(X, Y)* $\leftarrow$ *born(X,Z)* $\wedge$ *located(Z,Y)* [0.76]

(E2) $t_2$: *nutrient(Cheese, Carbohydrate)*
$\mathcal{E}_2$: *nutrient(Milk, Carbohydrate)*
$\phi_2^*$: *nutrient(Cheese, X)* $\leftarrow$ *nutrient(Milk, X)* [0.86]

(E3) $t_3$: *relatedForm(breakable, break)*
$\mathcal{E}_3$: *relatedForm(break, breakable)*
$\phi_3^*$: *relatedForm(X, Y)* $\leftarrow$ *relatedForm(Y, X)* [0.92]

All explanations listed above are plausible explanations for the targets, however, they might require us to perform an additional abstraction step, for instance applying our external knowledge about the connection of Milk and Cheese. Incorporating the rule that led to the explanation provides a more complete and transparent view by helping us to *understand the explanation*. Moreover, in complex domains, e.g. the biomedical field, we might rely on the rules to support or extend our background knowledge.

The examples also help us to understand why attacks are less influential on the FB15k-237 dataset and we will provide an explicit explanation in the following.

Consider example (E1). In the case of *Miyuki Sawashiro* there is also a triple which expresses that she is a Japanese voice actor and one that expresses that she lives in Tokyo. For both relations there are rules that allow the prediction of the nationality. This means that the perturbation or removal of the attacking triple will probably not be sufficient as other relevant evidence for her nationality remains in the data. We found these alternative explanations by performing several deletion attacks in succession, while suppressing the deleted triples of previous attacks. In FB15k-237 we could find many examples where the target triples were backed by multiple strong evidences.

Example (E3), on the other hand, is a typical example for the targets of WN18RR. For the relation *derivationalRelatedForm* there exist triples in the test set for which a correct prediction can be trivially derived from the inverse triple. Fortunately, the application of our approach helped us to discover this phenomenon. Due to the evaluation protocol, which uses only those triples as targets where the correct candidate is ranked on #1 in both directions, many triples in the target set can be explained by $\phi_3^*$.

## 7 Conclusion

We presented a black-box method for adversarial attacks against KGE models by using a special form of abductive reasoning. Experimental results showed that we achieve results on par with current state-of-the-art. Our method produces an explanation which is based on a human understandable rule reflecting a regularity in the dataset. This is a clear advantage over other methods. Our approach is applicable to any KGE model out-of-the-box while the best performing attacks in prior work have specifically been designed for the models ComplEx, ConveE, TransE and DistMult. For future research, we plan to extend the experimental setting to more sophisticated KGC architectures.

## References

[Bhardwaj *et al.*, 2021] Peru Bhardwaj, John Kelleher, Luca Costabello, and Declan O'Sullivan. Adversarial attacks on knowledge graph embeddings via instance attribution methods. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8225–8239. Association for Computational Linguistics, 2021.

[Bylander *et al.*, 1991] Tom Bylander, Dean Allemang, Michael C. Tanner, and John R. Josephson. The computational complexity of abduction. *Artificial Intelligence*, 49(1):25–60, 1991.

[Charpiat *et al.*, 2019] Guillaume Charpiat, Nicolas Girard, Loris Felardos, and Yuliya Tarabalka. Input similarity from the neural network perspective. In *NeurIPS 2019-33th Annual Conference on Neural Information Processing Systems*, 2019.

[Dettmers *et al.*, 2018] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 1811–1818, 2018.

[Galárraga *et al.*, 2013] Luis Antonio Galárraga, Christina Teflioudi, Katja Hose, and Fabian Suchanek. Amie: association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd international conference on World Wide Web*, pages 413–422, 2013.

[Guo *et al.*, 2016] Shu Guo, Quan Wang, Lihong Wang, Bin Wang, and Li Guo. Jointly embedding knowledge graphs and logical rules. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 192–202, 2016.

[Guo *et al.*, 2018] Shu Guo, Quan Wang, Lihong Wang, Bin Wang, and Li Guo. Knowledge graph embedding with iterative guidance from soft rules. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Hanawa *et al.*, 2021] Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, and Kentaro Inui. Evaluation of similarity-based explanations. In *International Conference on Learning Representations*, 2021.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

[Lawrence *et al.*, 2021] Carolin Lawrence, Timo Sztyler, and Mathias Niepert. Explaining neural matrix factorization with gradient rollback. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 4987–4995. AAAI Press, 2021.

[Mayer and Pirri, 1993] Marta Cialdea Mayer and Fiora Pirri. First order abduction via tableau and sequent calculi. *Logic Journal of the IGPL*, 1(1):99–117, 1993.

[Meilicke *et al.*, 2019] Christian Meilicke, Melisachew Wudage Chekol, Daniel Ruffinelli, and Heiner Stuckenschmidt. Anytime bottom-up rule learning for knowledge graph completion. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI/AAAI Press, 2019.

[Meilicke *et al.*, 2021] Christian Meilicke, Patrick Betz, and Heiner Stuckenschmidt. Why a naive way to combine symbolic and latent knowledge base completion works surprisingly well. In *3rd Conference on Automated Knowledge Base Construction*, 2021.

[Pezeshkpour *et al.*, 2019] Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. Investigating robustness and interpretability of link prediction via adversarial modifications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3336–3347, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Rim *et al.*, 2021] Wiem Ben Rim, Carolin Lawrence, Kiril Gashteovski, Mathias Niepert, and Naoaki Okazaki. Behavioral testing of knowledge graph embedding models for link prediction. In *3rd Conference on Automated Knowledge Base Construction*, 2021.

[Rossi *et al.*, 2021] Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2):1–49, 2021.

[Ruffinelli *et al.*, 2020] Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. You CAN teach an old dog new tricks! on training knowledge graph embeddings. In *International Conference on Learning Representations*, 2020.

[Trouillon *et al.*, 2016] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080. JMLR.org, 2016.

[Yang *et al.*, 2015] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations*, 2015.

[Zhang *et al.*, 2019a] Hengtong Zhang, Tianhang Zheng, Jing Gao, Chenglin Miao, Lu Su, Yaliang Li, and Kui Ren. Data poisoning attack against knowledge graph embedding. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4853–4859. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[Zhang *et al.*, 2019b] Wen Zhang, Bibek Paudel, Liang Wang, Jiaoyan Chen, Hai Zhu, Wei Zhang, Abraham Bernstein, and Huajun Chen. Iteratively learning embeddings and rules for knowledge graph reasoning. In *The World Wide Web Conference*, pages 2366–2377, 2019.