KG-EDAS: A Meta-Metric Framework for Evaluating Knowledge Graph Completion Models

Haji Gul¹, Abul Ghani Naim¹, Ajaz Ahmad Bhat^{1*}

¹School of Digital Science, Universiti Brunei Darussalam (23h1710, ghani.naim, ajaz.bhat*)@ubd.edu.bn

Abstract

Knowledge Graphs (KGs) enable applications in various domains such as semantic search, recommendation systems, and natural language processing. KGs are often incomplete, missing entities and relations, an issue addressed by Knowledge Graph Completion (KGC) methods that predict missing elements. Different evaluation metrics, such as Mean Reciprocal Rank (MRR), Mean Rank (MR), and Hit@k (e.g., Hit@1), are commonly used to assess the performance of such KGC models. A major challenge in evaluating KGC models however, lies in comparing their performance across multiple datasets and metrics. A model may outperform others on one dataset but underperform on another, making it difficult to determine overall superiority. Moreover, even within a single dataset, different metrics such as MRR and Hit@1 can yield conflicting rankings, where one model excels in MRR while another performs better in Hit@1, further complicating model selection for downstream tasks. These inconsistencies hinder holistic comparisons and highlight the need for a unified meta-metric that integrates performance across all metrics and datasets to enable a more reliable and interpretable evaluation framework. To address this need, we propose KG Evaluation based on Distance from Average Solution (EDAS), a robust and interpretable metametric that synthesizes model performance across multiple datasets and diverse evaluation criteria into a single normalized score $(M_i \in [0,1])$. Unlike traditional metrics that focus on isolated aspects of performance, EDAS offers a global perspective that supports more informed model selection and promotes fairness in cross-dataset evaluation. Experimental results on benchmark datasets such as FB15k-237 and WN18RR demonstrate that EDAS effectively integrates multi-metric, multi-dataset performance into a unified ranking, offering a consistent, robust, and generalizable framework for evaluating KGC models.

Introduction

KGs, formalized as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$ with entities \mathcal{E} , relations \mathcal{R} , and triples $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ in the form (h, r, t), encode structured real-world knowledge to enable applications such as question answering (Devlin et al. 2019), recommendation systems (Zhuang et al. 2021), and knowledge-enhanced language models. Due to their inherent incompleteness, KGC is

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

essential for predicting missing triples, such as relation prediction given (h,?,t), tail entity t prediction given (h,r,?) or head entity h given (?,r,t), using a scoring function $S_i: \mathcal{E} \times \mathcal{R} \times \mathcal{E} \to \mathbb{R}$ for each model M_i (Shu et al. 2024; Gul, Naim, and Bhat 2025).

Evaluating KGC models presents significant challenges, particularly when comparing their performance across multiple datasets and metrics. Commonly used rank-based metrics, such as MRR, MR, and Hits@k (e.g., Hits@1, Hits@3, Hits@10), assess different aspects of model performance. However, a model may excel on one dataset while underperforming on another, making it difficult to determine overall superiority(Rossi et al. 2021a). Additionally, even within a single dataset, conflicting rankings often arise when different metrics are considered. For instance, a model may achieve a high MRR but a low Hits@1 score, complicating model selection and leading to inconsistent evaluations(Sun et al. 2020). These inconsistencies across datasets and metrics highlight the need for a unified meta-metric that integrates performance across diverse evaluation criteria and benchmarks to provide a comprehensive and reliable assessment of KGC models. To address this need, we propose KG-EDAS, a multi-criteria decision-making meta-metric framework adapted from operational research (Ghorabaee et al. 2015) for KGC evaluation. KG-EDAS offers the following key capabilities:

- Unified Evaluation Framework: EDAS is the first multi-criteria evaluation metric for KGC, synthesizing performance across any KGC metrics like Hits@k and MR into a single normalized score $M_i \in [0,1]$ across any datasets, offering a single measure for model comparison.
- Enhanced Interpretability and Robustness: By balancing positive and negative deviations from average performance, EDAS provides interpretable global ranks (e.g., Rank 1, Rank 2) that resolve inconsistencies and reflect clear performance trade-offs.
- Cross-Dataset Generalizability: Unlike traditional metrics limited to single-dataset evaluations, EDAS enables comparisons both within and across datasets, facilitating a clearer assessment of model generalization and supporting robust model selection across benchmarks like FB15k-237 and WN18RR.

• Computational Efficiency: EDAS is implemented with linear time complexity $\mathcal{O}(nm)$, where n is the number of models and m is the number of evaluation criteria, ensuring scalability for large-scale KGC evaluations.

This work contributes a perspective shift in KGC evaluation by introducing a meta-metric that supports robust, interpretable and holistic comparisons of models across diverse benchmarks. Experimental results demonstrate that KG-EDAS effectively integrates multi-metric, multi-dataset performance into a unified ranking that is in consistent alignment with individual traditional metrics like MRR, MR and Hit@1 etc.

Related Work

Recent efforts in KGC have mainly focused on improving model accuracy through advanced architectures rather than refining the underlying evaluation methodologies. For instance, Wang et al. (Wang et al. 2023) introduced the Triplet Distributor Network (TDN), which demonstrated strong performance on Hits@3 but continued to rely on disjointed metrics such as MRR and Hits@k for evaluation. Similarly, Lin et al. (Lin, Socher, and Xiong 2018) proposed a multihop reasoning framework that achieved high MRR scores yet underperformed in Hit@1 evaluations, underscoring the inconsistent behaviour of traditional metrics across different criteria. Multi-task learning approaches, such as those by Kim et al. (Kim et al. 2020), aim to enhance predictive power by integrating auxiliary tasks like relation prediction; however, they still report results using isolated metrics without addressing the broader issue of metric fragmentation. Similarly, Wei et al. (Wei et al. 2023) introduced KICGPT, a large language model tailored for KGC, achieving competitive performance across multiple datasets. However, their evaluation strategy remains split across MRR, Hit@1, and Hit@10, requiring manual interpretation and potentially influencing comparative rankings. These examples illustrate a persistent reliance on conventional evaluation metrics despite growing recognition of their limitations. This fragmented approach complicates model comparison and hinders progress in the field, as researchers must manually weigh conflicting metric outcomes to make informed decisions.

In response to these challenges, recent studies have explored alternative strategies for evaluating KGC models. Ruffinelli et al. (Ruffinelli, Broscheit, and Gemulla 2020) conducted an extensive empirical review of knowledge graph embedding (KGE) models, highlighting inconsistencies in metric usage and calling for more standardized benchmarks. Sun et al. (Sun et al. 2020) emphasized the importance of incorporating uncertainty quantification into KGC evaluation, arguing that confidence estimates are essential for real-world deployment. Despite these insights, no comprehensive framework has emerged that integrates performance across multiple metrics and datasets into a single, interpretable score. Various critical gaps remain unaddressed in current KGC evaluation practices:

(1) Lack of Cross-Dataset Comparability: Most evaluation frameworks are limited to single-dataset analysis, of-

fering no mechanism to assess generalization across diverse benchmarks. As shown in recent works such as Sim-KGC (Wang et al. 2022), this limitation prevents meaningful comparisons of model robustness across varying data distributions. (2) *Underutilization of Decision Theory:* Although Multi Criteria Decision-Making (MCDM) methods like TOPSIS and VIKOR have demonstrated success in other machine learning domains (Kandakoglu, Walther, and Ben Amor 2024), their adoption in KGC remains minimal. These frameworks offer structured, principled ways to resolve conflicts among competing metrics and produce holistic model rankings, an opportunity largely overlooked in current KGC research.

To address these deficiencies, we introduce in the KG area a meta-meric KG-EDAS methodology derived from operational research for KGC evaluation. Unlike traditional scalar metrics such as MRR or Hit@k, which provide partial and often conflicting perspectives, EDAS synthesizes performance across multiple criteria and datasets into a unified, normalized score $(M_i \in [0,1])$. It computes both positive and negative deviations from average performance, enabling a balanced view of model strengths and weaknesses without relying on subjective reference points. This makes EDAS particularly well-suited for complex and uncertain environments like KGC, where ground truth rankings may be ambiguous or inconsistent. By introducing EDAS into the KGC domain, we present a principled, scalable, and interpretable meta-metric that supports fair and reproducible model comparison across diverse benchmarks.

Methodology

This section presents the KG-EDAS, a multi-criteria decision-making meta-metric framework for evaluating KGC models. By assessing performance across multiple metrics and datasets into a single interpretable score, KG-EDAS addresses the limitations of traditional scalar metrics like MRR and Hit@k, offering a unified and reproducible framework for model comparison. We begin by formulating the problem, followed by a structured explanation of how KG-EDAS is adapted to KGC evaluation.

Problem Formulation: Given a knowledge graph G = (E,R,T), where E denotes entities, R relations, and $T \subseteq E \times R \times E$ valid triples in the form (h,r,t), we focus on evaluating KGC models that predict missing entities or relations using a scoring function $S_i: E \times R \times E \to \mathbb{R}$. Let $\mathcal{M} = \{M_1, M_2, \ldots, M_n\}$ denote n KGC models, each producing a vector of scores across multiple evaluation metrics such as MRR, Hit@1, Hit@10, and MR given in Equations 1 and 2.

$$MR = \frac{1}{N} \sum_{i=1}^{N} rank_i, \quad MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank_i}, \quad (1)$$

Hits@k =
$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\operatorname{rank}_{i} \le k)$$
 (2)

This study aims to derive a unified ranking of KGC models based on their aggregated performance across all evaluation metrics and datasets. To enable a holistic compari-

son, the process begins with the construction of a performance matrix $X \in \mathbb{R}^{n \times m}$, where each entry X_{ij} represents the score of model M_i on metric j. The following sections provide a step-by-step description of the KG-EDAS metametric framework.

• Decision Matrix Construction: To perform a systematic and multi-criteria evaluation, we organize the results into a structured format called the decision matrix $X \in \mathbb{R}^{n \times m}$, where rows correspond to models and columns to metrics:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} \end{bmatrix}$$
(3)

, where $X_{ij} \in \mathbb{R}$, $i=1,\ldots,n,\ j=1,\ldots,m$ Each metric is classified as either beneficial (e.g., MRR, Hit@k) or non-beneficial (e.g., MR). This matrix serves as the foundational input for the EDAS method, transforming heterogeneous performance indicators into a uniform space suitable for computing deviations from average performance.

Average Solution Computation: Next compute the average solution Avg_i for each metric j as:

$$\text{Avg}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}, \quad j = 1, \dots, m$$
 (4)

This yields an average vector $\operatorname{Avg} = [\operatorname{Avg}_1, \operatorname{Avg}_2, \dots, \operatorname{Avg}_m] \in \mathbb{R}^m$, representing the central tendency of model performance across all criteria. The mean solution serves as an index for evaluation, which measures the relative performance of a specific model against the rest of the group. It provides consistency in ranking by removing biases caused by metrics using different scales; for instance, MRR generally spans from 0 to 1. The EDAS method provides a balanced and interpretable framework for multi-criteria KGC evaluation by normalising variations from the average evaluation.

• Positive and Negative Distance from Average (PDA and NDA): The next step involves measuring how each model deviates from the average solution, either positively or negatively, depending on whether the metric is beneficial or non-beneficial. This dual-metric approach helps EDAS to effectively evaluate both the strengths and weaknesses of KGC models, therefore enabling a balanced and interpretable multi-criteria ranking. Let X_{ij} as the performance score of the *i*-th model on the *j*-th criterion, and let Avg_j represent the average score of the *j*-th criterion across all models.

For beneficial metrics (e.g., MRR, Hit@k):

$$PDA_{ij} = \frac{\max(0, X_{ij} - Avg_j)}{Avg_i},$$
 (5)

$$NDA_{ij} = \frac{\max(0, Avg_j - X_{ij})}{Avg_j}$$
 (6)

For *non-beneficial metrics* (e.g., MR):

$$PDA_{ij} = \frac{\max(0, Avg_j - X_{ij})}{Avg_j},$$
(7)

$$NDA_{ij} = \frac{\max(0, X_{ij} - Avg_j)}{Avg_j}$$
 (8)

These normalized deviations ensure comparability across metrics with different scales, avoiding division-by-zero issues via small constant adjustments where necessary.

Weighted PDA and NDA: To incorporate the relative importance of each metric, we apply weighted aggregation. Let w_j ∈ [0, 1] denote the weight assigned to metric j, with ∑_{j=1}^m w_j = 1. In our experiments, an equal weight is assigned. The weighted positive and negative distances are computed as:

$$WPDA_i = \sum_{j=1}^{m} w_j \cdot PDA_{ij}$$
 (9)

$$WNDA_i = \sum_{j=1}^{m} w_j \cdot NDA_{ij}$$
 (10)

These values reflect how much better or worse a model performs relative to the average, weighted by the importance of each metric. As reported, this study used equal weights for each criterion to ensure balanced evaluation across complexity measures. However, depending on the downstream tasks such as recommendation systems or ranking applications, where metrics like Hit@10 or Hit@K are more valuable, users can assign higher weights to metrics that align with these objectives while assigning lower weights to less relevant metrics, to better tailor the KG-EDAS framework to task-specific requirements.

• Normalization of WPDA and WNDA: To ensure consistency and interpretability, we normalize WPDA and WNDA values to the range [0, 1]:

$$N(WPDA_i) = \frac{WPDA_i}{\max(WPDA)}$$
 (11)

$$N(WNDA_i) = \frac{WNDA_i}{\max(WNDA)}$$
 (12)

This normalization enables meaningful comparison across diverse benchmarks.

• Final Evaluation Score (M_i) : This step computes a unified performance score $M_i \in [0, 1]$ for each model:

$$M_i = \frac{1}{2} [N(WPDA_i) + (1 - N(WNDA_i))]$$
 (13)

This score balances strengths (positive deviation) and weaknesses (negative deviation), producing a single interpretable value for each model.

• Model Ranking Based on M_i : Once all models have been assigned their respective M_i scores, the final step involves generating a definitive ranking of the models based on these scores. Let $\mathbf{M} = [M_1, M_2, \dots, M_n]$ be

Table 1: Comparative time and space complexity of multi-criteria ranking methods

Method	Time Complexity	Space Complexity	Parallelizable	Notes
EDAS	$\mathcal{O}(nm)$	$\mathcal{O}(nm)$	Yes	Linear in models × metrics
TOPSIS (Kandakoglu, Walther, and Ben Amor 2024)	$\mathcal{O}(nm+n^2)$	$\mathcal{O}(nm+n^2)$	Partially	Ideal/anti-ideal vector comparisons
Pareto Frontier (Lin, Zhang, and Wang 2023)	$\mathcal{O}(n^2m)$	$\mathcal{O}(n^2)$	No	Regret Minimisation Step
Borda Count (Emerson 2023)	$\mathcal{O}(nm\log m)$	$\mathcal{O}(nm)$	Yes	Risk of inconsistency

the vector of final scores for n models. The ranking is determined by sorting this vector in descending order:

$$Rank(i) = argsort(M_i, descending=True)$$
 (14)

This yields an ordered list where the model with the highest M_i receives $Rank\ I$, indicating superior performance across all criteria. Unlike traditional metrics such as MRR and Hit@k, which often produce conflicting rankings, the M_i -based ranking resolves inconsistencies by integrating multiple criteria into a single decision-making framework. This ranking mechanism enhances interpretability, supports fair comparison, and facilitates model selection in KGC research.

Computational Complexity of KG-EDAS: As a metametric framework, KG-EDAS synthesizes diverse evaluation metrics—such as MRR, Hit@1, Hit@10, and MR—into a unified score $M_i \in [0,1]$, enabling holistic and interpretable comparisons across models and datasets. One of the key strengths of EDAS lies in its linear time complexity, which ensures scalability even when evaluating large sets of models over multiple benchmark datasets. This is particularly important given the fragmented nature of KGC evaluation, where models often exhibit inconsistent performance across different metrics and datasets. Traditional scalar metrics like MRR or Hit@k are fast to compute individually but fail to provide a comprehensive view of model effectiveness. Comparing results across these traditional metrics introduces ambiguity, requiring manual inspection that becomes increasingly impractical as the number of models and evaluation criteria grows.

KG-EDAS addresses this challenge by computing a single, interpretable ranking through a structured workflow, as summarized in Table 1. Unlike more complex multi-criteria methods such as TOPSIS or VIKOR—which rely on ideal reference points or pairwise distance matrices—EDAS uses the average performance vector as a baseline, eliminating unnecessary computational overhead while maintaining robustness and fairness.

Let n be the number of models being evaluated and m be the number of performance criteria (e.g., MRR, Hit@1, MR). Each model's performance is represented as a row in the decision matrix $X \in \mathbb{R}^{n \times m}$. The computational steps and their respective complexities are detailed below. Let T(n,m) denote the total time complexity for n models and m metrics. We now analyze the computational complexity of each step in the EDAS workflow:

Step 1: Average Calculation
$$\operatorname{Avg}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}$$
 (15)

$$T_1(n,m) = m \cdot \mathcal{O}(n) = \mathcal{O}(nm)$$

Step 2: Distance Metrics PDA_{ij}, NDA_{ij} (16)
 $T_2(n,m) = 2nm \cdot \mathcal{O}(1) = \mathcal{O}(nm)$

Step 3: Weighted Aggregation WPDA_i, WNDA_i (17) $T_3(n,m) = 2n \cdot \mathcal{O}(m) = \mathcal{O}(nm)$

Step 4: Normalization
$$N(\text{WPav}_i)$$
, $N(\text{WNav}_i)$ (18)
 $T_4(n) = \mathcal{O}(n) \text{ (max)} + \mathcal{O}(n) \text{ (division)} = \mathcal{O}(n)$

Step 5: Ranking
$$M_i = \frac{1}{2} \left[N(\text{WPav}_i) + (1 - N(\text{WNav}_i)) \right]$$
 (19)

Rank
$$(i)$$
 = argsort $(M_i, descending=True)$ (20)
 $T_5(n) = \mathcal{O}(n \log n)$

Overall:
$$T(n,m) = \underbrace{\mathcal{O}(nm)}_{\text{Step s } 1-3} + \underbrace{\mathcal{O}(n)}_{\text{Step 4}} + \underbrace{\mathcal{O}(n \log n)}_{\text{Step 5}}$$
 (21)
= $\mathcal{O}(nm + n \log n)$

For typical KGC evaluations where $m \ge \log n$ (e.g. $n = 10^4, m = 10$), this simplifies to:

$$T(n,m) \approx \mathcal{O}(nm)$$
 (22)

This linear complexity makes EDAS highly suitable for real-world applications involving large-scale model comparisons. It avoids computationally intensive operations such as iterative optimization or pairwise comparisons, further enhancing its efficiency and interpretability.

Experiments

Meta-metric KG-EDAS evaluated on widely used KG datasets:

- YAGO3-10 (Mahdisoltani, Biega, and Suchanek 2013): A subset of YAGO3 focusing on high-quality facts with entities having at least 10 relations. It contains 123,182 entities, 37 relations, 1,079,040 training, 5,000 validation, and 5,000 test triplets.
- **FB15k-237** (Bollacker et al. 2008): An updated version of FB15k with inverse triplets removed to increase difficulty. It consists of 14,541 entities, 237 relations, 272,115 training, 17,535 validation, and 20,466 test triplets.
- **FB15k** (Bollacker et al. 2008): A subset of Freebase containing general facts. It comprises 14,951 entities, 1,345 relations, 483,142 training, 50,000 validation, and 59,071 test triplets.

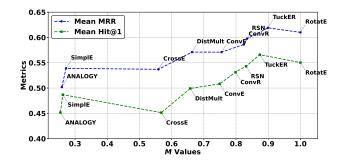
Table 2: Relation Prediction Final EDAS Scores with Model Ranking

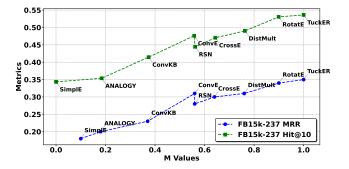
Model	WPDA_sum	WNDA_sum	NWPDA	NWNDA	M	Rank
RotatE (Sun et al. 2019)	0.2214	0.0000	0.9954	0.0000	0.9977	1
TuckER (Wang, Broscheit, and Gemulla 2019)	0.1943	0.0075	0.8735	0.0236	0.9250	<u>2</u>
RSN (Jiang, Wang, and Wang 2019)	0.1590	0.0021	0.7151	0.0065	0.8543	3
ConvR (Guo, Sun, and Hu 2019)	0.1456	0.0088	0.6547	0.0277	0.8135	4
ConvE (Dettmers et al. 2018)	0.1130	0.0051	0.5080	0.0158	0.7461	5
DistMult (Yang et al. 2015)	0.1052	0.0368	0.4730	0.1155	0.6788	6
CrossE (Zhang et al. 2019)	0.0439	0.0306	0.1974	0.0960	0.5507	7
SimplE (Kazemi and Poole 2018)	0.0333	0.1758	0.1496	0.5511	0.2992	8
ANALOGY (Liu, Wu, and Yang 2017)	0.0312	0.1995	0.1404	0.6252	0.2576	9
TorusE (Ebisu and Ichise 2018)	0.0718	0.2727	0.3227	0.8549	0.2339	10

Table 3: Link prediction results on FB15k, WN18, FB15k-237, WN18RR, and YAGO3-10. The results reported here are published in (Rossi et al. 2021b)

Models FB15k			WN18			FB15k-237			WN18RR			YAGO3-10			M	Ranks						
	MR	MRR	H@1	H@10	MR	MRR	H@1	H@10	MR	MRR	H@1	H@10	MR	MRR	H@1	H@10	MR	MRR	H@1	H@10		
RotatE	42	0.791	0.739	0.881	274	0.949	0.943	0.960	178	0.336	0.238	0.531	3318	0.475	0.426	0.573	1827	0.498	0.405	0.671	0.998	1
TuckER	39	0.788	0.729	0.889	510	0.951	0.946	0.958	162	0.352	0.259	0.536	6239	0.459	0.430	0.514	2417	0.544	0.466	0.681	0.925	2
RSN	70	0.773	0.706	0.886	471	0.950	0.946	0.959	251	0.346	0.256	0.526	5646	0.467	0.437	0.527	2582	0.527	0.446	0.673	0.854	3
ConvR	51	0.777	0.723	0.870	346	0.928	0.912	0.951	248	0.280	0.198	0.444	4210	0.395	0.346	0.483	1339	0.511	0.427	0.664	0.814	4
ConvE	51	0.688	0.595	0.849	413	0.945	0.939	0.957	281	0.305	0.219	0.476	4944	0.427	0.390	0.508	2429	0.488	0.399	0.658	0.746	5
DistMult	173	0.784	0.736	0.863	675	0.824	0.726	0.946	199	0.313	0.224	0.490	5913	0.433	0.397	0.502	1107	0.501	0.413	0.661	0.679	6
CrossE	136	0.702	0.601	0.862	441	0.834	0.733	0.950	227	0.298	0.212	0.470	5212	0.405	0.381	0.450	3839	0.446	0.331	0.654	0.551	7
SimplE	138	0.726	0.661	0.836	759	0.938	0.933	0.946	651	0.179	0.100	0.344	8764	0.398	0.383	0.427	2849	0.453	0.358	0.632	0.299	8
ANALOGY	126	0.726	0.656	0.837	808	0.934	0.926	0.944	476	0.202	0.126	0.354	9266	0.366	0.358	0.380	2423	0.283	0.192	0.457	0.258	9
TorusE	143	0.746	0.689	0.840	525	0.947	0.943	0.954	211	0.281	0.196	0.447	4873	0.463	0.427	0.534	19455	0.342	0.274	0.474	0.234	10

Figure 1: Comparison of prediction metrics across datasets. The left image shows the relation (h, ?, t) prediction comparison of mean MRR and EDAS M values across datasets: FB15k-237, FB15k, WN18, WN18RR, and YAGO3-10. The right image shows comparison of mean Hit@1 and EDAS M-values.





- WN18RR (Miller 1995): A subset of WN18, where reverse triplets are removed for increased complexity. The dataset includes 40,943 entities, 11 relations, 86,835 training, 2,924 validation, and 2,824 test triplets.
- WN18 (Miller 1995): A subset of WordNet with lexical relations. It includes 40,943 entities, 18 relations, 141,442 training, 5,000 validation, and 5,000 test triplets.

Results

By assessing performance across multiple metrics (MR, MRR, Hit@1, Hit@10) KG-EDAS produces a unified ranking that resolves inconsistencies often observed when using traditional metrics. The final EDAS score $M_i \in [0,1]$ aggre-

gates these normalized deviations (NWPDA and NWNDA), rewarding model strengths and penalizing weaknesses in a single interpretable value. This meta-metric approach offers a holistic view of model effectiveness, addressing the limitations of traditional scalar metrics like MRR, which can be inconsistent across datasets and overly sensitive to top ranks.

We first apply KG-EDAS to link (relation) prediction task results from different models across different datasets as shown in Tables 2 and 3. As Table 2 shows, RotatE achieves the highest EDAS score ($M_i=0.9977$) and is ranked first due to its consistently strong performance across all datasets, reflected in its high NWPDA (0.9954) and zero NWNDA. In contrast, models such as ANALOGY ($M_i=0.2576$)

Table 4: Correlation Coefficients and P-values between EDAS Score and Traditional Metrics

Dataset	Metric Pair	Pear	son	Kendall			
		Correlation	P-value	Correlation	P-value		
Multiple Date ante	EDAS M Values vs Mean_MRR	0.9332	0.0002	0.8733	0.0012		
Multiple Datasets	EDAS M Values vs Mean_Hit@1	0.8329	0.0053	0.8333	0.0009		
	EDAS M Values vs Hit@10	0.9834	0.0000	0.8889	0.0002		
FB15k-237	EDAS M Values vs MRR	0.9739	0.0000	0.9143	0.0007		
	EDAS M Values vs MR	-0.8372	0.0025	-0.6889	0.0047		

Table 5: Tail Prediction Final EDAS Scores with Model Ranking

Model	WPDA_sum	WNDA_sum	NWPDA	NWNDA	M	Rank
TransR (Lin et al. 2015)	0.1745	0.0331	0.9482	0.0604	0.9439	1
TransD (Ji et al. 2015)	0.1767	0.0677	0.9603	0.1236	0.9183	<u>2</u>
TransH (Wang et al. 2014)	0.1629	0.0619	0.8851	0.1130	0.8860	3
TransE (Bordes et al. 2013)	0.1362	0.0498	0.7404	0.0909	0.8248	4
ComplEx (Trouillon et al. 2016)	0.0567	0.0819	0.3080	0.1495	0.5793	5
DistMult (Yang et al. 2015)	0.0641	0.1128	0.3485	0.2060	0.5713	6
AMIE (Galárraga et al. 2015)	0.1840	0.5478	1.0000	1.0000	0.5000	7

Table 6: Tail prediction results on FB15k, WN18, FB15k-237, and WN18RR using baseline models, including aggregated metric M. The results reported here are published in (Akrami et al. 2020).

Model	FB15k				WN18			FB15k-237			WN18RR			Ranking
	MR	MRR	Hits@10	MR	MRR	Hits@10	MR	MRR	Hits@10	MR	MRR	Hits@10		
TransE	243	0.227	0.199	263	0.395	0.142	363.3	0.169	0.32	2414.7	0.176	0.47	0.944	1
TransH	211	0.177	0.234	318	0.434	0.190	398.8	0.157	0.30	2616	0.178	0.46	0.918	<u>2</u>
TransR	226	0.236	0.231	232	0.441	0.199	391.3	0.164	0.31	2847	0.184	0.48	0.886	3
TransD	211	0.179	0.234	242	0.421	0.202	391.6	0.154	0.30	2967	0.172	0.47	0.825	4
DistMult	313	0.240	0.264	915	0.558	0.80	566.3	0.151	0.30	3798.1	0.264	0.46	0.579	5
ComplEx	350.3	0.233	0.250	636.1	0.584	0.80	656.4	0.158	0.29	3755.9	0.276	0.46	0.571	6
AMIE	337	0.370	0.64	1299.8	0.931	0.094	1909	0.201	0.36	12963	0.357	0.35	0.500	7

and TorusE ($M_i=0.2339$) receive lower rankings due to higher NWNDA values, indicating more frequent underperformance relative to the group average. The experimental results summarized in Table 3 further demonstrate that KG-EDAS effectively resolves conflicts among conventional metrics, delivering a definitive and interpretable ranking of KGC models. This enables fair comparisons not only within individual benchmarks but also across them, supporting generalizable insights into model selection.

Correlation Analysis of EDAS with KGC methods given in Figure 1(a) illustrates the relationship between the proposed KG-EDAS score (M) and traditional evaluation metrics Mean MRR and Mean Hit@1—across multiple benchmark datasets including FB15k, WN18, FB15k-237, WN18RR, and YAGO3-10. When models are ranked by their EDAS scores along the x-axis, it becomes evident that both Mean MRR and Mean Hit@1 exhibit strong positive correlations with M, particularly in distinguishing topperforming models. This suggests that EDAS effectively

captures the core strengths emphasized by these widely used metrics while resolving inconsistencies that arise when models perform well in one metric but poorly in another. In contrast, isolated scalar metrics often produce conflicting rankings, making it difficult to derive a reliable overall assessment of model performance. EDAS addresses this issue by synthesizing these metrics into a single, interpretable score, offering a more balanced and consistent evaluation framework.

Figure 1(b), focusing on the FB15k-237 dataset, further reinforces the consistency of EDAS with conventional metrics such as Hit@10. The plot shows a clear pattern, indicating that models achieving higher Hit@10 values also receive higher EDAS scores. This graphical alignment supports the hypothesis that EDAS preserves and enhances the meaningful insights captured by individual metrics while eliminating ambiguity caused by conflicting rankings. Unlike traditional metrics that fluctuate independently and may misrepresent performance robustness, EDAS aggregates results across all

criteria and datasets, producing a stable and interpretable ranking that reflects true model strength.

The statistical correlation analysis presented in Table 4 quantifies this alignment. Across multiple datasets, EDAS demonstrates a strong correlation with both Mean MRR and Mean Hit@1, with Pearson coefficient values at 0.9332 and 0.8329 respectively, both statistically significant at p <0.01. Kendall's τ confirms this strong agreement, showing values of 0.8733 and 0.8333, respectively. On the FB15k-237 dataset specifically, the correlation is even stronger, with Pearson values of 0.9834 for Hit@10 and 0.9739 for MRR. These results validate that EDAS not only aligns closely with established metrics but also enhances evaluation stability by integrating them into a unified meta-metric framework. While MR exhibits a moderate negative correlation (Pearson = -0.8372), this too is expected and consistent, reflecting EDAS's ability to reward low MR values appropriately. Altogether, these findings confirm that EDAS reliably reflects model quality as assessed by traditional metrics, while offering a more holistic and reproducible evaluation approach.

Similarly, for the tail prediction task results, illustrated in Table 5, utilizing the KG-EDAS further substantiates its cross-dataset capability and unique rank allocation. Upon evaluating each method across the datasets in Table 6, it is clear that EDAS eliminates deficiencies seen in conventional metrics, providing an accurate and coherent ranking of KGC models. Moreover, the linear time complexity $\mathcal{O}(nm)$ of the EDAS method ensures scalability and efficiency, making it particularly suitable for large-scale KGC evaluations involving many models and diverse evaluation criteria.

Ablation

To evaluate the sensitivity of KG-EDAS to individual assessment metrics and confirm its robustness, we conducted an ablation study by sequentially removing one metric at a time MRR, MR, and Hit@1—and recomputing the EDAS model rankings. The results, summarized in Table 7, demonstrate that KG-EDAS produces highly consistent rankings even when a key metric is excluded.

Table 7: Model Ranking Analysis After Removing Individual Metrics

	Original	1	Remov		
Model	Rank	MRR	MR	Hit@1	Max Change
RotatE	1	1	3	1	2
TuckER	2	2	1	2	1
ConvR	3	3	2	3	1
ConvE	4	4	5	4	1
DistMult	5	5	4	5	1
CrossE	6	6	6	6	0
SimplE	7	7	7	7	0
ANALOGY	8	8	8	8	0

When MRR is removed, the rankings remain identical to the original KG-EDAS ranking for all models. This indicates that MRR, while informative, does not disproportionately influence the final ranking. Similarly, excluding Hit@1 results in no rank changes across any model, confirming that the framework effectively captures performance through complementary metrics without over-reliance on top-1 accuracy. In contrast, removing MR leads to more noticeable shifts-most notably, TuckER and ConvR swap positions, and ConvE drops from rank 4 to 5. RotatE exhibits the largest movement, shifting from rank 1 to rank 3 when MR is removed (a change of 2 positions), as reflected in its maximum rank change value. This suggests that MR plays a distinctive role in differentiating models with mid-tier performance, where subtle differences in ranking quality become more evident. Despite these changes, the majority of models show minimal variation. In fact, three models (CrossE, SimplE, ANALOGY) maintain identical rankings across all ablation settings (Max Change = 0), and no model experiences a rank shift larger than 2 positions. This further underscores the stability of the framework.

These results confirm that KG-EDAS provides a balanced and robust evaluation: it integrates multiple performance aspects into a single score without being unduly influenced by any individual metric. This makes it a reliable and consistent alternative for evaluating and ranking knowledge graph completion models, even under partial evaluation conditions.

Conclusion

In conclusion, KG-EDAS is a holistic and interpretable meta-metric framework for evaluating KGC models across multiple datasets and performance criteria. By integrating both positive and negative deviations from average performance, EDAS offers a balanced view of model strengths and weaknesses, capturing trade-offs that conventional metrics miss, such as high MRR but low Hit@1. The experimental results demonstrate that KG-EDAS aligns strongly with established metrics like mean MRR and mean Hit@1 while resolving inconsistencies among them. Correlation analysis shows that EDAS closely matches these metrics, especially MRR, while providing a stronger and more reliable way to rank results. Furthermore, ablation studies show that the framework remains largely stable even when individual metrics are removed, highlighting its resilience and comprehensive design. By looking at more than just single metrics, KG-EDAS allows for consistent comparisons of models across different datasets and helps researchers make better decisions in KGC studies. Its linear time complexity ensures scalability, making it suitable for large-scale model assessments. These advantages position KG-EDAS as a valuable tool not only for benchmarking KGC methods but also for guiding future model development and selection. This work brings a change in how to evaluate KGC, moving from scattered, specific metrics for each dataset to a clear and consistent framework for evaluation. As KGs continue to grow in size and application scope, such a standardized and interpretable evaluation methodology becomes essential for meaningful progress in the field.

Thank you for reading these instructions carefully. We look forward to receiving your electronic files!

References

- Akrami, F.; Saeef, M. S.; Zhang, Q.; Hu, W.; and Li, C. 2020. Realistic re-evaluation of knowledge graph completion methods: An experimental study. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 1995–2010.
- Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proc. of ACM SIG-MOD*, 1247–1250.
- 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.
- Dettmers, T.; Minervini, P.; Stenetorp, P.; and Riedel, S. 2018. Convolutional 2D knowledge graph embeddings. In *Proc. of AAAI*.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, volume 1, 4171–4186.
- Ebisu, T.; and Ichise, R. 2018. Toruse: Knowledge Graph Embedding on a Lie Group. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Emerson, P. 2023. From Borda to Approval Voting: A Comparative Analysis of Decision-Making Methods. *European Journal of Operational Research*, 305(1): 1–12.
- Galárraga, L. A.; Teflioudi, C.; Hose, K.; and Suchanek, F. M. 2015. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. *Proceedings of the 24th International Conference on World Wide Web*, 413–422.
- Ghorabaee, M. K.; Zavadskas, E. K.; Olfat, L.; and Turskis, Z. 2015. Multi-Criteria Inventory Classification Using a New Method of Evaluation Based on Distance from Average Solution (EDAS). *Informatica*, 26: 435–451.
- Gul, H.; Naim, A. G.; and Bhat, A. A. 2025. MuCo-KGC: Multi-context-Aware Knowledge Graph Completion. In Wu, X.; Spiliopoulou, M.; Wang, C.; Kumar, V.; Cao, L.; Zhou, X.; Pang, G.; and Gama, J., eds., *Data Science: Foundations and Applications*, 3–15. Singapore: Springer Nature Singapore. ISBN 978-981-96-8298-0.
- Guo, L.; Sun, Z.; and Hu, W. 2019. Learning to Exploit Long-term Relational Dependencies in Knowledge Graphs. In *International Conference on Machine Learning*. PMLR.
- Ji, G.; He, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. Knowledge graph embedding via dynamic mapping matrix. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 687–696.

- Jiang, X.; Wang, Q.; and Wang, B. 2019. Adaptive Convolution for Multi-relational Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).*
- Kandakoglu, M.; Walther, G.; and Ben Amor, S. 2024. The use of multi-criteria decision-making methods in project portfolio selection: a literature review and future research directions. *Annals of Operations Research*, 332(1): 807–830.
- Kazemi, S. M.; and Poole, D. 2018. Simple Embedding for Link Prediction in Knowledge Graphs. In *Advances in Neural Information Processing Systems*, volume 31.
- Kim, B.; Hong, T.; Ko, Y.; and Seo, J. 2020. Multi-Task Learning for Knowledge Graph Completion with Pretrained Language Models. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 1737–1743. Barcelona, Spain (Online): International Committee on Computational Linguistics.
- Lin, X.; Zhang, Q.; and Wang, J. 2023. Pareto Frontier Learning for Multi-Objective Optimization in Machine Learning. *IEEE Transactions on Neural Networks and Learning Systems*.
- Lin, X. V.; Socher, R.; and Xiong, C. 2018. Multi-Hop Knowledge Graph Reasoning with Reward Shaping. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3243–3253. Brussels, Belgium: Association for Computational Linguistics.
- Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.
- Liu, H.; Wu, Y.; and Yang, Y. 2017. Analogical Inference for Multi-relational Embeddings. In *International Conference on Machine Learning*. PMLR.
- Mahdisoltani, F.; Biega, J.; and Suchanek, F. M. 2013. YAGO3: A Knowledge Base from Multilingual Wikipedias. In 7th Biennial Conference on Innovative Data Systems Research (CIDR 2013). Online; accessed YAGO3-10 dataset details.
- Miller, G. A. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11).
- Rossi, A.; Barbosa, D.; Firmani, D.; Matinata, A.; and Merialdo, P. 2021a. Knowledge graph embedding for link prediction: A comparative analysis. *TKDD*.
- Rossi, A.; Barbosa, D.; Firmani, D.; Matinata, A.; and Merialdo, P. 2021b. Knowledge Graph Embedding for Link Prediction: A Comparative Analysis. *ACM Trans. Knowl. Discov. Data*, 15(2).
- Ruffinelli, D.; Broscheit, S.; and Gemulla, R. 2020. You can teach an old dog new tricks! on training knowledge graph embeddings.
- Shu, D.; Chen, T.; Jin, M.; Zhang, C.; Du, M.; and Zhang, Y. 2024. Knowledge Graph Large Language Model (KG-LLM) for Link Prediction:(ACML). *Proceedings of Machine Learning Research*, 260(1): 143.

- Sun, Z.; Deng, Z.; Nie, J.; and Tang, J. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *Proc. of ICLR*.
- Sun, Z.; Zhang, Q.; Hu, W.; Wang, C.; Chen, M.; Akrami, F.; and Li, C. 2020. A benchmarking study of embedding-based entity alignment for knowledge graphs. *arXiv* preprint *arXiv*:2003.07743.
- 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.
- Wang, J.; Wang, B.; Gao, J.; Li, X.; Hu, Y.; and Yin, B. 2023. TDN: Triplet Distributor Network for Knowledge Graph Completion. *IEEE Transactions on Knowledge and Data Engineering*, 35(12): 13002–13014.
- Wang, L.; Zhao, W.; Wei, Z.; and Liu, J. 2022. SIM-KGC: Simple Contrastive KGC with Pre-trained Language Models. In *Proc. of ACL*.
- Wang, Y.; Broscheit, S.; and Gemulla, R. 2019. A Relational Tucker Decomposition for Multi-Relational Link Prediction. *arXiv preprint*. ArXiv:1902.00898.
- Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.
- Wei, Y.; Huang, Q.; Zhang, Y.; and Kwok, J. 2023. KICGPT: Large Language Model with Knowledge in Context for Knowledge Graph Completion. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8667–8683. Singapore: Association for Computational Linguistics.
- Yang, B.; Yih, W.-t.; He, X.; Gao, J.; and Deng, L. 2015. Embedding entities and relations for learning and inference in knowledge bases. *arXiv* preprint arXiv:1412.6575.
- Zhang, W.; Paudel, B.; Zhang, W.; Bernstein, A.; and Chen, H. 2019. Interaction Embeddings for Prediction and Explanation in Knowledge Graphs. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*.
- Zhuang, L.; Wayne, L.; Ya, S.; and Jun, Z. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In Li, S.; Sun, M.; Liu, Y.; Wu, H.; Liu, K.; Che, W.; He, S.; and Rao, G., eds., *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, 1218–1227. Huhhot, China: Chinese Information Processing Society of China.