Rewarding Explainability in Drug Repurposing with Knowledge Graphs

Abstract

Knowledge graphs are powerful tools for modelling complex, multi-relational data and supporting hypothesis generation, particularly in applications like drug repurposing. However, for predictive methods to gain acceptance as credible scientific tools, they must ensure not only accuracy but also the capacity to offer meaningful scientific explanations.

This paper presents a novel approach REx for generating scientific explanations in knowledge graphs. It employs reward and policy mechanisms that consider desirable properties of scientific explanation to guide a reinforcement learning agent in the identification of explanatory paths within a KG. The approach further enriches explanatory paths with domain-specific ontologies, ensuring that the explanations are both insightful and grounded in established biomedical knowledge.

We evaluate our approach in drug repurposing using three popular knowledge graph benchmarks. The results clearly demonstrate its ability to generate explanations that validate predictive insights against biomedical knowledge and that outperform the state-of-the-art approaches in predictive performance, establishing REx as a relevant contribution to advance AI-driven scientific discovery.

1 Introduction

Knowledge Graphs (KGs) have emerged as versatile representations for capturing complex, multi-relational data in various scientific domains. They play a critical role in organizing, exploring, and sharing knowledge while supporting AI-based scientific discovery by providing structured, conceptually rich frameworks that can align neural model predictions with domain knowledge[D'aquin, 2024]. Link prediction has proven to be a powerful tool for hypothesis generation, enabling the discovery of novel relations between entities [Ott *et al.*, 2022; Akujuobi *et al.*, 2024]. Applications include gene-disease associations [Yuen and Jansson, 2020] and drug repurposing [Napolitano *et al.*, 2018], where new therapeutic targets are discovered for existing drugs.

To serve effectively as a scientific tool, artificial intelligence must possess the capability to generate *scientific explanations* [Durán, 2021]. This raises an important question: is there a fundamental relationship between the explanation of natural phenomena and the explanation of algorithmic outputs? This topic is subject to ongoing debate among researchers, particularly regarding how to epistemically ground the results of computational models as reliable representations of real-world phenomena. Nonetheless, there is widespread consensus that computational artefacts can facilitate the explanation of natural phenomena [Durán, 2017; Krohs, 2008; Durán, 2021].

However, recent studies highlight that widely used attribution models [Ribeiro *et al.*, 2016; Lundberg, 2017] are inadequate for achieving the level of scientific and human-level explainability required for meaningful insights [Chou *et al.*, 2022]. This is also true of state-of-the-art link prediction explainability approaches [Rossi *et al.*, 2022; Betz *et al.*, 2022], which are limited to identifying relevant features or triples, falling short of fulfilling the requirements of scientific explanations. In fact, relevant theories of scientific explanation argue that purely statistical explanations fail to identify causally relevant factors or mechanisms [Salmon, 1984].

For these predictive methods to be adopted as reliable scientific tools, they must not only deliver accurate results but also afford mechanisms for *scientific explanations*, i.e., methods that explain the scientific validity of the predictions, ensuring that they make sense with the current scientific body of knowledge and are not the result of spurious correlations [Holzinger *et al.*, 2019]. Take as an example the following explanation for a drug recommendation for the patient John Doe that is grounded on a specific inhibitory mechanism that mitigates the effects of a deleterious mutation: *John Doe –has mutation→ MET T540G –part of→ MET Gene –related to → Tyrosine Kinase Activity ←inhibits– Sunitinib.*

Moreover, the potential of KGs to support scientific insights extends beyond link prediction. As long as inputs and outputs of a hypothesis generation system can be represented within a KG, the scientific knowledge encoded therein can be explored to create scientific explanations.

This paper focuses on generating knowledge-driven scientific explanations to validate scientific hypotheses generated by AI methods. It employs reward and policy mechanisms that consider desirable properties of scientific explanation to

guide a reinforcement learning agent in the identification of explanatory paths within a KG. The approach further enriches explanatory paths with domain-specific ontologies, ensuring that the explanations are both insightful and grounded in established biomedical knowledge.

Our contributions include: (1) a method to extend Reinforcement Learning (RL) frameworks to consider scientific explainability properties when generating explanatory paths; (2) a method to calculate the relevance of explanatory paths that accounts for research bias; (3) a method to compose fully-fledged scientific explanations by integrating relevant paths with descriptive ontology classes; (4) the first evaluation of knowledge-driven drug repurposing explanation on three distinct benchmark KGs.

2 Related Work

KGs have emerged as critical tools in Explainable AI due to their ability to model multi-relational data and generate interpretable explanations. Despite significant advances, existing methods face several challenges, particularly in domains like biomedicine, where biologically relevant explanations are essential for scientific validation.

In the biomedical domain, KGs have been applied to justify AI-driven predictions for drug repurposing. For example, PoLo [Liu *et al.*, 2021] combines representation learning with logical constraints to identify interpretable reasoning paths. Similarly, Ozkan et al. [Ozkan *et al.*, 2023] extended the PREDICT framework [Gottlieb *et al.*, 2011] to rank explanatory paths by their relevance to drug indications, incorporating established biomedical relationships. Stork et al. [Stork *et al.*, 2023] proposed to improve RL-based drug repurposing using phenotype annotations. These efforts build on RL frameworks for multi-hop inference, such as DeepPath [Xiong *et al.*, 2017] and MINERVA [Das *et al.*, 2017]. However, they prioritize predictive accuracy over other desirable properties for scientific explanation, such as relevance, limiting their utility in scientific contexts.

3 Problem Definition

Consider a knowledge graph $\mathcal{G} = (\mathcal{E}, \mathcal{R}, F)$, where $\mathcal{E}$ is a set of entities, $\mathcal{R}$ is a set of relations, and $F \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is a set of triples denoted as (s, r, o) for subject, relation, and object. We will often abuse notation and directly define $\mathcal{G}$ as the set of triples such that we can write $(s, r, o) \in \mathcal{G}$.

Link prediction over knowledge graphs is often employed to discover new relationships between entities in the KG and is typically formulated as the task of finding objects o when given a tuple (s, r). Interestingly, while KGs are inherently explainable, link prediction methods typically sacrifice explainability and are limited to outputting a set of entities predicted to be linked.

Predicting new links in scientific knowledge graphs is akin to formulating a scientific hypothesis that postulates the existence of a new relationship between two already known entities, e.g., *(minoxidil, treats, hair loss)*. While classical link prediction methods that explore the KG to learn new triples can be used for generating hypotheses, it is also possible to generate hypotheses using other data sources and methods and then model the hypothesis as a new triple in a relevant KG that represents both entities. The knowledge encoded in the KG can then be used to afford explanations.

Modern theories of scientific explanation often emphasize explanatory virtues such as empirical adequacy, simplicity, scope, and coherence. We adopt the taxonomy proposed in [Keas, 2018] for scientific theories where four types of explanatory virtues are identified: (i) Evidential theoretical virtues: evidential accuracy, causal adequacy, and explanatory depth; (ii) Coherential theoretical virtues: internal consistency, internal coherence, and universal coherence; (iii) Aesthetic theoretical virtues: beauty, simplicity, and unification; (iv) Diachronic theoretical virtues: durability, fruitfulness, and applicability.

By virtue of being domain models and representing explicit relations between entities, KGs can naturally support causal adequacy and contribute to explanatory depth by identifying causal mechanisms, e.g., *vincristine –treats→ lymphatic system cancer –associated with→ TIA 1 gene ← associated with– hematologic cancer.* Path and rule extraction are common methods to offer explanations in graph theory and link prediction [Meilicke *et al.*, 2019; Liu *et al.*, 2021; Zhang *et al.*, 2023] that align well with evidential virtues, in particular causal adequacy. Moreover, KGs can also cover all coherential virtues by affording mechanisms to ensure internal consistency and coherence (i.e., an explanation's components are not contradictory) and universal coherence (i.e., the explanation fits well with extant knowledge). While aesthetic virtues are generally considered less valuable, both their pragmatic value — simpler or shorter explanations are easier to grasp — and epistemic value (v. Occam's Razor) are relevant. Finally, diachronic virtues require additional time after the initial explanation formulation and are therefore out of our scope.

Having established paths as the core of our explanation definition, we define a path of length k in $\mathcal{G}$ as a finite sequence of triples $(e_i, r_i, e_{i+1}) \in \mathcal{G}$ for $i = 1, \ldots, k-1$, which joins a sequence of distinct entities $e_1, \ldots, e_k \in \mathcal{E}$. However, not all paths connecting the subject and object of a hypothesis triple fit the criteria of scientific explainability. In what regards to evidential virtues, a first challenge is ensuring *causal detail*, which often translates to producing a chain of intermediaries linking the entities at hand [Rosales and Morton, 2021]. A second challenge is ensuring the *relevance* of explanations since an explanation that achieves causal detail can still be vague and afford little scientific insight. For example, an explanation for the hypothesis *(sunitinib, treats, renal cancer)* that takes the form *(sunitinib, is a, antineoplastic agent, treats, cancer, super class of, renal cancer)* is correct, causal, but not scientifically relevant. A third challenge lies in ensuring the *completeness* of explanations since an adequate causal account often requires the interaction of multiple factors rather than a single directed cause-and-effect path. Ensuring *universal coherence* can also be a challenge since many popular scientific KGs do not possess a schema backed by an ontology, which limits the ability to ensure and evaluate the logical coherence with the domain. Finally, a fifth challenge is related to *simplicity* or parsimony and how to ensure the pragmaticity of explanations without sacrificing

other relevant properties. Clearly, addressing the first four challenges necessarily represents a trade-off with addressing the fifth since ensuring a detailed, complete, relevant and universally coherent explanation very likely requires a larger and more complex explanation.

4 Method

4.1 Overview

Our approach generates scientific explanations for a hypothesis *h* — described as triple in a KG $\mathcal{G}$ — as a subgraph $\mathcal{G}_h$ that integrates a set of relevant explanatory paths $p \in P$ and relevant ontology classes that describe entities in the path. An explanatory path for hypothesis *h* is a path that connects the subject and object entities of *h*, respectively s_h and o_h, through a chain of relevant related entities.

Our explanation generation strategy addresses the challenges of causal detail and relevance by using reinforcement learning, conditioned on the hypothesis to validate, to find explanatory paths. It employs a reward-shaping mechanism to ensure multi-objective optimization regarding fidelity (i.e., to ensure paths successfully connect s_h and o_h) and relevance (aiming to maximize the information content (IC) of a path, a measure of the specificity of the entities composing it). Simplicity is ensured by a policy that ensures the RL agent produces paths without loops and that terminate when o_h is reached. These explanatory paths are then filtered to include those that are maximally relevant and representative of different explanation types, ensuring *completeness*. These are grouped to form the backbone of $\mathcal{G}_h$, which is then enriched by including type axioms connecting entities in $\mathcal{G}_h$ to relevant ontology classes, thereby affording a richer contextualization of the explanation subgraph, facilitating *universal coherence*

The overall approach is illustrated in Figure 1. Given a biomedical KG and a set of drug repurposing predictions to be individually explained, the method follows a three-phase process: (1) computing the information content of entities; (2) finding explanatory paths; (3) generating scientific explanations.

4.2 Information Content

An essential aspect of our method is to compute the relevance of paths. Our hypothesis is that paths involving less frequent entities are more likely to reveal meaningful relationships, resulting in explanations that are both insightful and representative of the underlying scientific knowledge. We define the relevance of a path as the average of the information content (IC) of the edges that compose it.

We formalize the concept of IC from information theory, taking into account *node degree counts*: the number of edges (relations) connected to a node (entity). First, we define the IC of an entity v that appears in a triple, either as the subject or object. The informativeness of an edge is determined by the average IC of the two entities it connects. To compute the IC of an entity, we introduce the concept of *Clustered IC*, which refines the IC calculation by analyzing a clustered graph instead of the original graph. The intuition is to diminish possible bias resulting from heterogeneous granularity levels due to over-studied and under-studied areas, as well as cases where minor variations of the same concept are included. This can be further refined by evaluating nodes within the context of specific relations, reflecting the intuition that the significance of an entity may vary depending on the nature of the relationships it engages in, providing a more detailed and accurate measure of path informativeness.

IC of a Node

Let $T = (S, R, O)$ be a random variable for the KG triples. Sampling from T means sampling a random triple from the graph.

Definition 1. The IC of a node $v \in \mathcal{E}$, denoted $\mathrm{IC}(v)$, is defined by the information content of the event $(S = v) \cup (O = v)$. It measures the surprisal of a node appearing as a subject or object in a randomly sampled triple.

We can derive the following (proof in appendix):

Theorem 1. *Considering each triple of the graph as independently and identically distributed, we have*

$$\mathrm{IC}(v) = -\log \frac{\deg(v)}{|\mathcal{G}|}. \tag{1}$$

where $\deg(v)$ *the total degree of an entity node in the knowledge graph,*

Clustered IC of a Node

The IC of a node can be modified to account for potential node degree bias due to different granularities, whereby in some subdomains, two very similar concepts are represented by different entities. This can be an effect of research bias and not necessarily translate to a scientifically meaningful measure of frequency. Let A and A' be two such similar entities and B a third entity. Two possible issues can arise: (i) non-meaningful relations between A and A' (*is a* or *synonym of*) can "boost" their node degrees, (ii) inversely, some relations highlighted only between A' and B but not between A and B (whereas, in practice, they should hold) can "hide" the true node degree of A. Such issues can artificially increase or decrease the node degree of concepts. The clustered graph aims to group semantically similar concepts together to mitigate this.

Consider the set of clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ such that each $C_i \subseteq \mathcal{E}$ and $\bigcup_{i=1}^{k} C_i = \mathcal{E}$. Consider the clustering function $\kappa : \mathcal{E} \to \mathcal{C}$. Each entity $v \in \mathcal{E}$ belongs to exactly one cluster $\kappa(v) = C_i$. The clusters are assumed to group entities per semantic similarity.

Given the KG $\mathcal{G} = (\mathcal{C}, \mathcal{R}, F)$, we derive a clustered graph $\mathcal{G}_c = (\mathcal{C}, \mathcal{R}, F_c)$ where $F_c = \{(C_i, r, C_j) | \exists (u, r, v) \in F$ s.t. $u \in C_i$ and $v \in C_j\}$. Intuitively, the clustered graph derived from $\mathcal{G}$ is a graph where the nodes are grouped into clusters and there is an edge between two clusters if any nodes in these clusters were connected by an edge in $\mathcal{G}$.

As for the original graph, we derive random variables for the clustered graph (S_c, R_c, O_c) .

Definition 2. The clustered IC (CIC) of a node $v \in \mathcal{E}$ is defined by the information content of the event $(S_c = \kappa(v)) \cup (O_c = \kappa(v))$. It measures the surprisal of a node belonging to one of the subject cluster or object cluster randomly sampled in a triple from the clustered graph $\mathcal{G}_c$.

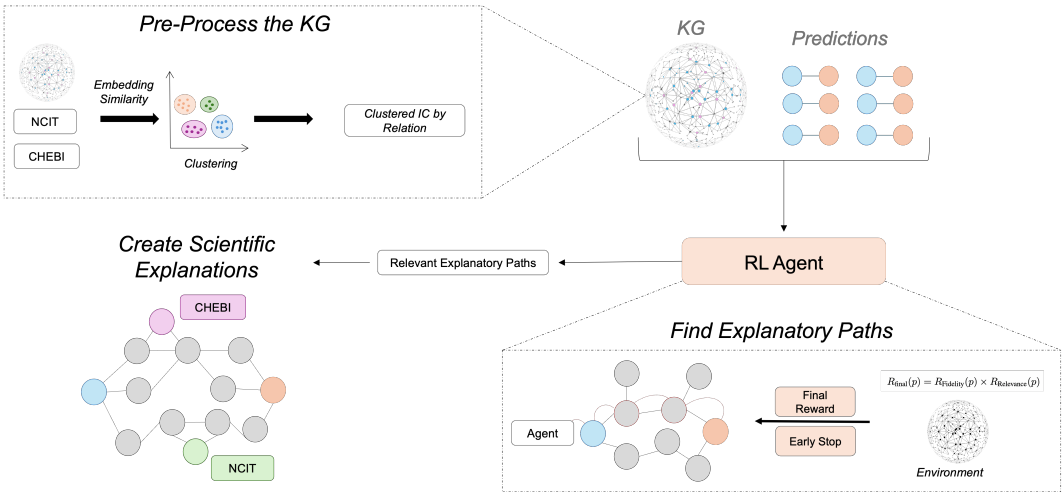

Figure 1: Overview of the approach to generate scientific explanations, with three main phases: pre-process KG, find explanatory paths, and create scientific explanations.

Similar to before, we can derive (proof in appendix):

Theorem 2. *Considering each triple of the clustered graph as independently and identically distributed, we have*

$$\mathrm{IC}_c(v) = -\log \frac{\deg(\kappa(v))}{|\mathcal{G}_c|} \tag{2}$$

where the degree of $\kappa(v)$ is calculated in the clustered graph.

The Clustered IC of a node can also be computed by relation type, whereby node degrees are calculated only with edges of a given relationship type $\deg(v, r)$.

4.3 Finding Explanatory Paths

We adapt the RL path-finding strategy proposed in [Das *et al.*, 2017] to our hypothesis validation purpose. It specifies a deterministic partially observed Markov decision process as a 5-tuple (S, O, A, T, R).

States. The state space $\mathcal{S}$ consists of all combinations in $\mathcal{E}\times\mathcal{E}\times\mathcal{E}$. Intuitively, we want a state to encode the hypothesis subject s_h and object o_h, as well as a location of exploration e (current location of the RL agent). Therefore, a state $S \in \mathcal{S}$ is represented by:

$$S = (e, s_h, o_h)$$

Observations. The complete state of the environment is not observed since the agent only knows its current location e and the hypothesis subject. Formally, the observation function $O : S \to E \times E$ is defined as:

$$O = (e, s_h)$$

Actions. The set of possible actions A_S from a state $S = (e, s_h, o_h)$ include all edges connected to the current node e in G or a decision to stop. Formally:

$$A_S = \{(e, r, e_d) \in E : S = (e, s_h, o_h), r \in R, e \in \mathcal{E}\} \cup \{\text{STOP}\}.$$

This means that at each state, the agent decides either to stop or to continue to destination node e_d.

Transition. Environment evolution is deterministic, simply updating the state to the new entity selected by the agent.

Rewards. The reward function captures two objectives:

- **Fidelity**: A scientific explanation should necessarily align with the hypothesis. Fidelity indicates whether the path-finding algorithm successfully connects s_h and o_h. Formally, if $S_T = (e, s_h, o_h)$ is the end state and $e = o_h$, $R_{\text{Fidelity}(p)} = 1$, else $R_{\text{Fidelity}(p)} = 0$.
- **Relevance**: A scientific explanation should provide detailed insights into the mechanisms underlying the hypothesis. Formally, when the end state is reached, the average IC of the path p is computed to arrive at $R_{\text{Relevance}}(p)$.

Formally, the final reward R_{final} of a path p is given by:

$$R_{\text{final}}(p) = R_{\text{Fidelity}}(p) \times R_{\text{Relevance}}(p)$$

Policy Network. We extend the policy proposed in [Das *et al.*, 2017] with an early stopping mechanism. This policy — based on LSTMs to encode the history of actions and observations — presents desirable properties for explanatory path generation on KGs, namely that it is permutation-invariant to edge ordering and history-dependent, with decisions d_t mapping the history H_t to a probability distribution over available actions A_{S_t}. The history $H_t = (H_{t-1}, A_{t-1}, O_t)$ records past actions (A_{t-1}) and observations (O_t). When the policy network chooses an action from all available actions, $\mathcal{A}_{ST}$, if $S_T = (e, s_h, o_h)$ and $e = o_h$, the agent does not take any further actions. This mechanism not only reduces unnecessary exploration but also promotes simplicity. The original policy resorted to a special action which goes from a node to itself, which resulted in possible loops and repetitiveness.

Training. We extend training to 30 rollouts, following [Liu *et al.*, 2021].

4.4 Generating Scientific Explanations

To construct scientific explanations for predictions, we begin by analyzing all explanatory paths found in the previous step and grouping them based on the pattern they follow, i.e., metapaths. For each metapath, we select the path with the highest IC. These selected paths are then merged into a graph

and enriched with the lowest common ancestors (LCA) between all consecutive entities in a path. More formally,

$$\mathcal{G}_{\text{h}} = \bigcup_{p \in \mathcal{P}} p \quad \cup \bigcup_{(e_i, e_{i+1}) \in p} \text{LCA}(e_i, e_{i+1})$$

5 Experiments

5.1 Drug repurposing

To evaluate the effectiveness of REx[1] in generating scientifically valid explanations, we applied it to the task of validating drug repurposing hypotheses. Drug repurposing identifies new therapeutic uses for existing drugs, and such a hypothesis can be formulated as a triple $(drug, treats, disease)$. As benchmarks for our experiments, we used well-known biomedical KGs that describe drugs, diseases and other relevant entities for drug repurposing[2]: Hetionet [Himmelstein *et al.*, 2017], PrimeKG [Chandak *et al.*, 2023], and OREGANO [Boudin *et al.*, 2023]. Hetionet is an integrative biomedical KG combining data from 29 sources, including genes, compounds, and diseases, and with more than 45,000 entities. PrimeKG is a precision medicine-oriented knowledge graph spanning multiple biological scales, such as pathways, phenotypes, and drug indications, with nearly 130,000 entities. OREGANO is specifically designed for drug repurposing, aligning experimental data with drug-disease associations and more than 98,000 entities. In each case, inverse edges were added when not provided in the KG. More detailed statistics can be found in the appendix.

5.2 KG enrichment

Since the benchmark KGs are semantically shallow and do not include an ontology-based schema, we enriched each KG by aligning it to relevant domain ontologies, specifically the National Cancer Institute Thesaurus (NCIT) [Hartel *et al.*, 2005] and the Chemical Entities of Biological Interest (ChEBI) [Degtyarenko *et al.*, 2007], which accurately describe drugs, diseases and other entities in the KGs. These alignments were generated using the ontology matching system AML [Faria *et al.*, 2023] (see the Appendix for further details). To support the Clustered IC computation, we generated embeddings using OWL2vec* [Chen *et al.*, 2021], which were then clustered using K-means, with the number of clusters set to 10% of the total node count.

To evaluate the relevance of the explanations, the explanatory paths identified in Hetionet were transformed into metapaths. A metapath is defined as a sequence of entity types and relations within a knowledge graph, representing a generalized pattern of connections between entities. These metapaths were then compared to a ground truth derived from the findings of Himmelstein [Himmelstein *et al.*, 2017]. A table with the ground truth paths is provided in the appendix.

6 Results and Discussion

6.1 Predictive Performance Evaluation

We evaluated the predictive performance of REx's explanatory paths against several baseline methods, including rule-based, embedding-based, neuro-symbolic, and RL-based approaches (details in Appendix). This evaluation assesses the most basic property of a scientific explanation — if it aligns with the hypothesis. The reported values for REx, PoLo, and MINERVA correspond to the calculation of a mean across five independent successful training runs, with a standard deviation between 0.004 and 0.028. All other results were reported in [Liu *et al.*, 2021].

Table 1 presents the results of various methods on the Hetionet KG. REx outperformed all state-of-the-art methods, achieving the highest MRR of 0.427. This indicates that REx effectively integrates fidelity, simplicity, and relevance to find robust explanations.

Table 1: Performance comparison of various methods for predictions on Hetionet based on Hits@1, Hits@3, Hits@10, and MRR metrics.

Method	Hits@1	Hits@3	Hits@10	MRR
AnyBURL	0.229	0.375	0.553	0.322
TransE	0.099	0.199	0.444	0.205
DistMult	0.185	0.305	0.510	0.287
ComplEx	0.152	0.285	0.470	0.250
ConvE	0.100	0.225	0.318	0.180
RESCAL	0.106	0.166	0.377	0.187
R-GCN	0.026	0.245	0.272	0.135
CompGCN	0.172	0.318	0.543	0.292
pLogicNet	0.225	0.364	0.523	0.333
MINERVA	0.264	0.409	0.593	0.370
PoLo	0.314	0.428	**0.609**	0.402
REx	**0.338**	**0.461**	**0.609**	**0.427**

Similarly, Table 2 and Table 3 highlight the results on PrimeKG and OREGANO with the best-performing methods on Hetionet. Here, REx once again surpassed the state of the art, achieving an MRR of 0.376 on PrimeKG and 0.278 on OREGANO. These results confirm that REx is able to produce explanations that have more predictive power than comparable methods across a variety of KGs.

Table 2: Performance comparison for predictions on PrimeKG based on Hits@1, Hits@3, Hits@10, and MRR metrics.

Method	Hits@1	Hits@3	Hits@10	MRR
MINERVA	0.262	0.420	**0.546**	0.359
PoLo	0.245	0.408	0.526	0.344
REx	**0.286**	**0.429**	0.544	**0.376**

To evaluate the relevance of explanatory paths, we computed the average IC for each path generated with REx, MINERVA, and PoLo. The distribution of ICs shown in Figure 2 clearly indicates that REx's paths have a higher IC in general and that it does not generate paths of low relevance, with the vast majority above 0.4. While PoLo does not fall far behind, MINERVA produces a larger portion of paths with lower IC. This underlines that the reward mechanism employed by REx effectively excludes low-relevance paths.

[1]Code available in supplementary material.

[2]Data repository links in the Appendix.

Table 3: Performance comparison for predictions on OREGANO based on Hits@1, Hits@3, Hits@10, and MRR metrics.

Method	Hits@1	Hits@3	Hits@10	MRR
MINERVA	0.133	0.200	0.489	0.220
PoLo	**0.171**	0.292	0.473	0.259
REx	**0.171**	**0.327**	**0.533**	**0.278**

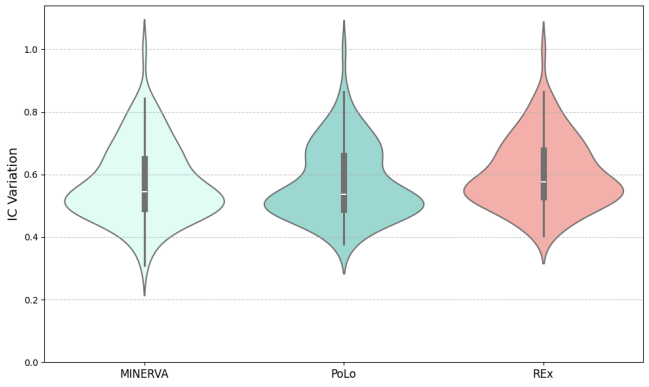

Figure 2: IC distribution for the methods MINERVA, PoLo, and REx using the Hetionet dataset.

To analyze the contribution of each component to the overall performance of REx, we conducted an ablation study by systematically removing individual components of the approach:

- **REx**: The complete version of our approach.
- **REx *-s***: A variation where the early stop mechanism is removed, sacrificing the simplicity virtue.
- **REx *-r***: A variation where the relevance is excluded from the final reward calculation.
- **REx *-rs***: A variation where both simplicity and relevance are not considered, leaving only fidelity as a final reward.

The results shown in Table 4 indicate that when the early stopping mechanism was removed (-s), the MRR dropped in all datasets. This result highlights the importance of generating concise explanations, as longer paths can increase complexity and hinder interpretability. Similarly, the removal of relevance (-r) resulted in a noticeable decrease in performance, especially for Hetionet. This demonstrates that incorporating information content to ensure biologically meaningful paths is crucial for finding scientifically valid explanations. The combination of removing both the early stopping mechanism and relevance (-rs) led to the most significant reduction in performance in the cases of PrimeKG and OREGANO, emphasizing the complementary roles of these two components.

Table 5 presents a comparative evaluation of different approaches for computing IC within the REx using Hetionet. The results show that Clustered IC by Relation yields the highest scores overall, underscoring the value of tailoring IC calculations to specific relations. Interestingly, IC outperforms the Clustered IC method in every metric, indicating that an IC that is clustered and blind to relation type loses relevant information.

Table 4: Ablation of REx based on Hits@1, Hits@3, Hits@10, and MRR metrics. Bold indicates best result, italics second best.

KG	Method	Hits@1	Hits@3	Hits@10	MRR
Hetionet	REx *-s*	*0.309*	*0.446*	**0.627**	*0.407*
	REx *-r*	0.295	0.432	0.600	0.392
	REx *-rs*	0.302	*0.446*	*0.609*	0.404
	REx	**0.338**	**0.461**	*0.609*	**0.427**
PrimeKG	REx *-s*	0.278	0.426	*0.544*	0.370
	REx *-r*	*0.284*	**0.431**	**0.554**	**0.376**
	REx *-rs*	0.277	*0.429*	0.540	0.369
	REx	**0.286**	*0.429*	*0.544*	**0.376**
OREGANO	REx *-s*	0.143	0.314	0.543	0.264
	REx *-r*	0.149	0.244	0.514	0.244
	REx *-rs*	0.105	0.222	0.498	0.209
	REx	**0.171**	**0.327**	**0.533**	**0.278**

Table 5: Performance comparison for different types of IC on REx using Hetionet based on Hits@1, Hits@3, Hits@10, and MRR metrics.

Method	Hits@1	Hits@3	Hits@10	MRR
IC	0.290	0.437	0.595	0.391
CIC	0.264	0.419	0.591	0.370
CIC by Relation	**0.338**	**0.461**	**0.609**	**0.427**

6.2 Ground-truth evaluation

To assess the relevance of the explanatory paths generated by REx, we compared them to the ground truth paths identified in [Himmelstein *et al.*, 2017], which are recognized as key mechanisms for drug repurposing. This comparative approach draws on the principle of analogy [Thagard, 1978; Thagard, 1989], where new explanatory mechanisms gain credibility when they align with well-established causal structures.

Our analysis showed that REx identified 12 distinct types of explanatory paths. Of these, 8 path types were fully consistent with the ground truth, confirming their biological plausibility and alignment with existing biomedical knowledge. The remaining 5 path types (Table 8 in Appendix), while not explicitly included in the ground truth, still provided biologically coherent insights that could help formulate novel hypotheses about drug–disease relationships. In fact, the majority of these paths are similar to ground truth ones but use *palliates* instead of *treats*, which is semantically similar. By uncovering both validated and new explanatory paths, REx demonstrates its ability to both replicate known mechanisms and offer novel and plausible explanations that may advance our understanding of drug repurposing.

Regarding path frequency in the datasets, each knowledge graph exhibits distinct mechanistic preferences (Table 9 in Appendix), with Hetionet showing a clear preference for side effect-based paths, Oregano emphasizing gene-mediated paths and PrimeKG generating a substantial number of paths solely based on (*drug – indication — disease*) chains. The distribution of path frequencies is similarly skewed across all KGs, with a few highly frequent paths and many rare ones. This suggests that while certain drug repurposing mechanisms are well-described, there may be numerous specialized pathways that might be relevant for specific cases. Further analysis revealed that multiple explanatory paths can be identified for a single prediction, with an average of 12 relevant paths per drug repurposing hypothesis in Hetionet. This diversity of explanatory paths highlights the complexity of drug repurposing and the importance of considering multiple explanatory paths when producing an explanation.

6.3 Domain Expert Evaluation

We recruited two life sciences graduates familiar with drug repurposing to evaluate the validity of 10 randomly selected full REx-generated explanations for the Hetionet dataset (Figure 5 in Appendix). Each REx explanation was presented alongside a corresponding MINERVA explanation.

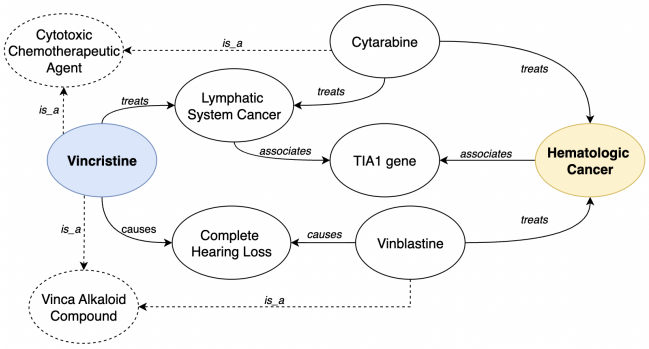

Figure 3: Explanation between Vincristine and Hematologic Cancer from Hetionet dataset.

The experts rated both explanations for each drug repurposing hypothesis on a scale from 1 to 5, reflecting their satisfaction with the explanation quality ranging from 1 (very low) to 5 (very high) and consulting any external references they deemed necessary.

Figure 4 depicts the experts' ratings for each explanation. Although both experts occasionally differ in how highly they rate the same method, the overall trend across all 10 explanations favours REx on every occasion.

To complement this analysis, we provide a literature-based validation of a REx generated explanation for the hypothesis *(vincristine, treats, hematologic cancer)* presented in Figure 3. A literature search revealed the scientific validity of this explanation, since both vincristine and vinblastine belong to the Vinca Alkaloid family, and both can cause neurotoxicity (including hearing loss) [Madsen *et al.*, 2019]. Furthermore, cytarabine and vincristine are therapeutic options for lymphatic system cancers, which are associated with the

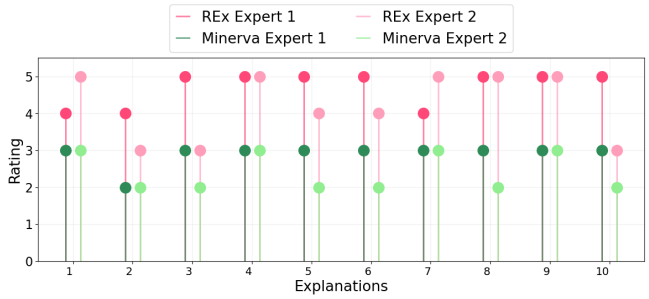

Figure 4: Comparison of the experts rating for 10 drug–disease repurposing explanations for Hetionet generated by REx (pink) and MINERVA (green).

TIA1 gene, which regulates the translation and stability of mRNAs involved in apoptosis, proliferation, and stress responses, relevant processes for cancer cell survival [Sánchez-Jiménez *et al.*, 2015]. In turn, TIA1 is associated with hematologic cancers.

7 Conclusion

We propose a novel method, REx, for generating scientific explanations of hypotheses based on KGs. Our method fulfils several desirable properties of scientific explanations, namely causal detail, relevance, completeness, coherence and simplicity.

It employs an RL-based approach guided by a dual reward that values both fidelity to the hypothesis and relevance to generate explanatory paths. The RL approach also considers an early stopping mechanism to consider simplicity. Paths are combined into a graph that is enriched with relevant ontology classes to ensure completeness and coherence. Notably, our approach can integrate a wide range of RL frameworks designed for graph-structured data, allowing our methodology to benefit from future evolutions in this field.

REx outperforms the state of the art in predictive performance, produces more relevant explanatory paths and results in explanations that are considered of better quality by experts in three benchmark tasks for drug-repurposing hypothesis validation. Nevertheless, the predictive performance of both REx and other state-of-the-art methods remains fairly modest, indicating that in many cases, no explanation can be found. This can be due to KG incompleteness or even to lack of scientific evidence, so further analysis is required to elucidate this aspect.

References

[Akujuobi *et al.*, 2024] Uchenna Akujuobi, Priyadarshini Kumari, Jihun Choi, Samy Badreddine, Kana Maruyama, Sucheendra K Palaniappan, and Tarek R Besold. Link prediction for hypothesis generation: an active curriculum learning infused temporal graph-based approach. *Artificial Intelligence Review*, 57(9):244, 2024.

[Betz *et al.*, 2022] Patrick Betz, Christian Meilicke, and Heiner Stuckenschmidt. Adversarial explanations for knowledge graph embeddings. In *IJCAI*, volume 2022, pages 2820–2826, 2022.

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

[Boudin *et al.*, 2023] Marina Boudin, Gayo Diallo, Martin Drancé, and Fleur Mougin. The oregano knowledge graph for computational drug repurposing. *Scientific data*, 10(1):871, 2023.

[Chandak *et al.*, 2023] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023.

[Chen *et al.*, 2021] Jiaoyan Chen, Pan Hu, Ernesto Jimenez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, and Ian Horrocks. Owl2vec*: Embedding of owl ontologies. *Machine Learning*, 110(7):1813–1845, 2021.

[Chou *et al.*, 2022] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81:59–83, 2022.

[Das *et al.*, 2017] Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. *arXiv preprint arXiv:1711.05851*, 2017.

[Defferrard *et al.*, 2016] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

[Degtyarenko *et al.*, 2007] Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl_1):D344–D350, 2007.

[Durán, 2017] Juan M Durán. Varying the explanatory span: scientific explanation for computer simulations. *International Studies in the Philosophy of Science*, 31(1):27–45, 2017.

[Durán, 2021] Juan M Durán. Dissecting scientific explanation in ai (sxai): A case for medicine and healthcare. *Artificial Intelligence*, 297:103498, 2021.

[D'aquin, 2024] Mathieu D'aquin. On the role of knowledge graphs in ai-based scientific discovery. *Journal of Web Semantics*, page 100854, 2024.

[Faria *et al.*, 2023] Daniel Faria, Emanuel Santos, Booma Sowkarthiga Balasubramani, Marta C Silva, Francisco M Couto, and Catia Pesquita. Agreementmakerlight. *Semantic Web*, (Preprint):1–13, 2023.

[Gottlieb *et al.*, 2011] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, 7(1):496, 2011.

[Hartel *et al.*, 2005] Frank W Hartel, Sherri de Coronado, Robert Dionne, Gilberto Fragoso, and Jennifer Golbeck. Modeling a description logic vocabulary for cancer research. *Journal of biomedical informatics*, 38(2):114–129, 2005.

[Himmelstein *et al.*, 2017] Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, 6:e26726, 2017.

[Holzinger *et al.*, 2019] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[Keas, 2018] Michael N Keas. Systematizing the theoretical virtues. *Synthese*, 195(6):2761–2793, 2018.

[Krohs, 2008] Ulrich Krohs. How digital computer simulations explain real-world processes. *International Studies in the Philosophy of Science*, 22(3):277–292, 2008.

[Liu *et al.*, 2021] Yushan Liu, Marcel Hildebrandt, Mitchell Joblin, Martin Ringsquandl, Rime Raissouni, and Volker Tresp. Neural multi-hop reasoning with logical rules on biomedical knowledge graphs. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*, pages 375–391. Springer, 2021.

[Lundberg, 2017] Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[Madsen *et al.*, 2019] Marie Lindhard Madsen, Hanne Due, Niels Ejskjær, Paw Jensen, Jakob Madsen, and Karen Dybkær. Aspects of vincristine-induced neuropathy in hematologic malignancies: a systematic review. *Cancer chemotherapy and pharmacology*, 84:471–485, 2019.

[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*, pages 3137–3143, 2019.

[Meilicke *et al.*, 2020] Christian Meilicke, Melisachew Wudage Chekol, Manuel Fink, and Heiner Stuckenschmidt. Reinforced anytime bottom up rule learning for knowledge graph completion. *arXiv preprint arXiv:2004.04412*, 2020.

[Napolitano *et al.*, 2018] Francesco Napolitano, Diego Carrella, Barbara Mandriani, Sandra Pisonero-Vaquero, Francesco Sirci, Diego L Medina, Nicola Brunetti-Pierri, and Diego Di Bernardo. gene2drug: a computational tool for pathway-based rational drug repositioning. *Bioinformatics*, 34(9):1498–1505, 2018.

[Nickel *et al.*, 2011] Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel, et al. A three-way model for collective learning on multi-relational data. In *Icml*, volume 11, pages 3104482–3104584, 2011.

[Ott *et al.*, 2022] Simon Ott, Adriano Barbosa-Silva, and Matthias Samwald. Linkexplorer: predicting, explaining and exploring links in large biomedical knowledge graphs. *Bioinformatics*, 38(8):2371–2373, 2022.

[Ozkan *et al.*, 2023] Elif Ozkan, Remzi Celebi, Arif Yilmaz, Vincent Emonet, and Michel Dumontier. Generating knowledge graph based explanations for drug repurposing predictions. In *SWAT4HCLS*, pages 22–31, 2023.

[Qu and Tang, 2019] Meng Qu and Jian Tang. Probabilistic logic neural networks for reasoning. *Advances in neural information processing systems*, 32, 2019.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[Rosales and Morton, 2021] Alirio Rosales and Adam Morton. Scientific explanation and trade-offs between explanatory virtues. *Foundations of Science*, 26:1075–1087, 2021.

[Rossi *et al.*, 2022] Andrea Rossi, Donatella Firmani, Paolo Merialdo, and Tommaso Teofili. Kelpie: an explainability framework for embedding-based link prediction models. *Proceedings of the VLDB Endowment*, 15(12):3566–3569, 2022.

[Salmon, 1984] Wesley C Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, 1984.

[Sánchez-Jiménez *et al.*, 2015] Carmen Sánchez-Jiménez, María Dolores Ludeña, and José M Izquierdo. T-cell intracellular antigens function as tumor suppressor genes. *Cell Death & Disease*, 6(3):e1669–e1669, 2015.

[Schlichtkrull *et al.*, 2018] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, pages 593–607. Springer, 2018.

[Stork *et al.*, 2023] Lise Stork, Ilaria Tiddi, René Spijker, and Annette ten Teije. Explainable drug repurposing in context via deep reinforcement learning. In *European Semantic Web Conference*, pages 3–20. Springer, 2023.

[Thagard, 1978] Paul R Thagard. The best explanation: Criteria for theory choice. *The journal of philosophy*, 75(2):76–92, 1978.

[Thagard, 1989] Paul Thagard. Explanatory coherence. *Behavioral and brain sciences*, 12(3):435–467, 1989.

[Trouillon *et al.*, 2016] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR, 2016.

[Vashishth *et al.*, 2019] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. Composition-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1911.03082*, 2019.

[Xiong *et al.*, 2017] Wenhan Xiong, Thien Hoang, and William Yang Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. *arXiv preprint arXiv:1707.06690*, 2017.

[Yang *et al.*, 2014] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[Yuen and Jansson, 2020] Ho Yin Yuen and Jesper Jansson. Better link prediction for protein-protein interaction networks. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 53–60. IEEE, 2020.

[Zhang *et al.*, 2023] Shichang Zhang, Jiani Zhang, Xiang Song, Soji Adeshina, Da Zheng, Christos Faloutsos, and Yizhou Sun. Page-link: Path-based graph neural network explanation for heterogeneous link prediction. In *Proceedings of the ACM Web Conference 2023*, pages 3784–3793, 2023.