Explaining Scientific Hypotheses in Drug Development with Knowledge Graphs

Susana Nunes(✉) and Catia Pesquita

LASIGE, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal
{scnunes,clpesquita}@ciencias.ulisboa.pt

Abstract. Artificial Intelligence (AI) methods based on Knowledge Graphs (KGs) have recently been proposed to generate explanations for scientific discovery tasks. Although these approaches offer a promising foundation for accelerating research and development, their practical adoption depends on a clearer understanding of which factors influence the usefulness and trustworthiness of the explanations they generate.

We conducted an in-depth user study with eleven biomedical researchers on KG-based explanations for two key scientific discovery tasks in drug development: drug repurposing and drug–target interaction. We evaluated the *relevance*, *completeness*, and *validity* of path-based explanations extracted from KGs by state-of-the-art methods, and compared them along two dimensions: whether they produce single vs. multiple explanatory paths, and whether they incorporate ontological information. The evaluation combined participants' ratings with a qualitative analysis of feedback to understand how experts interpret these explanations.

Our results revealed that biomedical researchers prefer explanations that integrate ontological information and present diverse, biologically plausible mechanisms over those based solely on structural connectivity. These findings shed light on key design choices for KG-based AI systems that aim to support informed candidate selection in drug development and, more broadly, improve their impact in scientific applications.

Keywords: Biomedical Knowledge Graphs · Explainable Artificial Intelligence · Drug Development

1 Introduction

While Knowledge Graphs (KGs) are increasingly used to improve reasoning in Artificial Intelligence (AI), many computational methods that leverage KGs (e.g., KG embeddings [3,32] and graph neural networks [30]) operate as black boxes. This opacity poses a significant challenge, particularly in domains such as drug development, which demand transparent and scientifically grounded predictions.

In this context, predictions serve not merely as outputs, but as scientific hypotheses that must be interpretable and justifiable before advancing to further

M. Acosta et al. (Eds.): ESWC 2026, LNCS 16549, pp. 102–119, 2026.
https://doi.org/10.1007/978-3-032-25156-5_6

validation. Biomedical KGs offer a promising solution by grounding explanations in existing scientific knowledge, enabling experts to assess whether a prediction is scientifically plausible and is worth further investigation. However, for these methods to gain broader adoption, it is essential to deepen our understanding of the key factors that shape the perceived usefulness and trustworthiness of the explanations they produce.

This need is especially pressing given the rising global demand for effective therapies and the high cost and failure rate of drug development, averaging 14 years and over US$3 billion per approved drug. While AI methods, particularly deep learning, have accelerated drug development tasks such as protein structure prediction and target discovery [23,35], many models remain black boxes, and explainability remains a key challenge. The importance of testable biological insights in drug discovery was underscored by the AstraZeneca–Sanger DREAM challenge on drug synergy prediction, which found that model interpretability had more impact than algorithmic sophistication [20], and has driven growing interest in explainable AI (XAI) for drug development research [15].

A key aspect of XAI in drug development that can mitigate the high clinical failure rates is to generate explanations that can validate a drug candidate's plausibility, considering existing biomedical knowledge. KGs can be a main driver in this area, given their ability to explicitly encode biological entities, their relationships, and underlying mechanisms [36], and enabling large-scale semantic integration, as exemplified by Hetionet [12]. In recent years, several KG-based methods have been applied to drug development tasks, including reinforcement learning-based systems such as Minerva [6], PoLo [18], and REx [24], which aim to uncover explanatory paths that support predictive links between biomedical entities.

While these algorithms have been applied in drug development settings [8], few studies have demonstrated the potential of multi-hop reasoning over biomedical KGs to generate scientifically relevant explanations, or have directly engaged domain experts or systematically compared alternative explanation strategies. To address this gap, we investigate the potential of multi-hop reasoning over KGs to support scientifically relevant explanations. To address this, we perform a systematic comparison of KG-based methods based on expert judgment. This is essential for evaluating whether KG-based explanations are meaningful and trusted by scientists, and for understanding how different approaches are valued by experts. Using established biomedical datasets, our study focuses on two key tasks in drug development: drug repurposing and drug–target interaction. We assess the perceived usefulness and trustworthiness of the explanations through expert feedback, complemented by quantitative performance metrics to contextualize system behavior. This dual perspective allows us to examine not only how well systems perform, but also how effectively their outputs support informed decision-making in real-world biomedical settings.

To ground this investigation, we propose the following research questions:

– RQ1: How do domain experts' professional backgrounds influence their preferences for KG-based explanations in drug development tasks?

- RQ2: To what extent do single-path versus multi-path explanation strategies affect experts' perceived trustworthiness and completeness of AI-generated hypotheses?
- RQ3: How does the inclusion of ontological information impact the interpretability and validity of explanations across different tasks?

The main contributions of this paper are:

1. An investigation of how domain experts' backgrounds and task contexts influence their interpretation and trust in KG-based explanations for drug development.
2. A systematic comparison of four KG-based systems applied to two central drug development tasks: drug repurposing and drug–target interaction.
3. A user study involving eleven biomedical researchers, assessing explanation quality along three dimensions: relevance, completeness, and validity.
4. A quantitative evaluation of predictive performance using standard link prediction metrics, including Mean Reciprocal Rank (MRR) and Hits@k, across multiple biomedical knowledge graphs.
5. Insights into the design of KG-based explanation strategies, informed by domain expert feedback and grounded in real-world use cases.

2 Related Work

KGs have become a core component of Explainable AI approaches in biomedical discovery due to their ability to encode structured, multi-relational data and reveal interpretable reasoning paths. In drug development, explanations must satisfy criteria beyond predictive accuracy, including biological plausibility, mechanistic coherence, and therapeutic relevance [15,27]. This makes KGs particularly suited for hypothesis-driven reasoning, as they enable the tracing of causal or mechanistic connections between biomedical entities.

Early reasoning frameworks such as DeepPath [33] and MINERVA [6] introduced reinforcement learning (RL) for multi-hop inference, optimizing agents to reach correct targets through graph traversal. However, these systems primarily focused on predictive performance rather than on the scientific interpretability of their reasoning paths. Building on this foundation, PoLo [18] integrated logical rules into the RL process, improving both predictive accuracy and interpretability by constraining agents to follow semantically coherent paths. Other works adapted KG reasoning to biomedical contexts: Ozkan et al. [25] extended the PREDICT framework [9] to rank drug–disease paths by biomedical relevance, while Stork et al. [31] enriched RL-based reasoning with phenotype annotations to improve contextual coherence.

More recently, REx [24] introduced a dual reward mechanism that balances fidelity and relevance, emphasizing scientific explainability by enriching explanatory paths with ontological context. Related approaches such as DREAMwalk [2] and TxGNN [13] have demonstrated that improving the interpretability of graph reasoning can enhance trust and clinical decision support.

User-centered evaluation of explanation strategies through expert studies is an established practice in XAI research [21,22], though it remains underexplored for KG-based explanations in scientific domains. While previous work has focused on algorithmic improvements, systematic investigation of how domain experts perceive and evaluate KG-based explanations in scientific settings remains limited. Our work addresses this gap through an empirical investigation of expert preferences, analyzing how professional backgrounds and task contexts influence the interpretation of path-based explanations in drug development.

3 Explaining Scientific Hypotheses

We investigate how KG-based methods generate meaningful explanations for two key drug development tasks: drug repurposing and drug–target interaction. The task of identifying candidate drugs, i.e., generating a hypothesis, can be modeled as a link prediction problem over a biomedical KG, where the goal is to infer missing edges that represent biologically meaningful relationships. Moreover, KGs are well-suited to support the uncovering of explanatory paths that suggest causal or mechanistic interpretations. Let us consider the following hypothesis, represented as a KG edge: *(vincristine –treats→ hematologic cancer)*. Inferring it can be based on a multi-hop path such as *vincristine –treats→ lymphatic system cancer –associated with→ TIA 1 gene –associated with→ hematologic cancer*. To generate such explanations, methods such as path and rule extraction are commonly employed within graph-based and link prediction frameworks [18,19,34]. Furthermore, such multi-hop paths can also be employed to perform link prediction, effectively working as explainable hypotheses generators. Several approaches employ reinforcement learning to handle the scalability, enhance flexibility, and promote robust generalization.

Our primary focus is to assess how explanation paths can support the underlying hypothesis by providing plausible and informative justifications that help domain experts evaluate it. To understand this, we conducted a structured user study to examine how such explanations are interpreted in biomedical decision-making scenarios, offering insights to guide the design of more effective explainable approaches. Additionally, since explainability must not come at the expense of predictive performance, particularly in high-stakes tasks like candidate drug identification, we also evaluated the predictive accuracy of the models alongside their explanatory value.

In this section, we begin by detailing the tasks in drug development and the biomedical knowledge graphs that support them. We then introduce the evaluated systems and provide an overview of the evaluation criteria applied in the evaluation.

3.1 Knowledge Graphs for Drug Development

Drug repurposing and drug-target interaction are central to accelerating therapeutic discovery while reducing cost and risk. Drug repurposing leverages exist-

ing compounds to identify new therapeutic indications, significantly shortening development timelines [1,28]. Drug–target interaction prediction helps uncover molecular mechanisms, enabling early-stage prioritization and safety assessment of candidate compounds [1,17].

Both tasks can be modeled as link prediction problems over biomedical knowledge graphs. Formally, each prediction instance takes the form of a triple (h, r, t), where h is the head entity, r the relation, and t the tail entity. In our evaluation setting, we aim to predict the missing tail t, i.e., to complete triples of the form $(h, r, ?)$.

Drug Repurposing. In this application, the task is that of predicting links between compounds and diseases. For example, a predicted link between the drug *Sunitinib* and *Papillary Renal Cell Carcinoma*, a kidney cancer type, may be supported by the path: *Sunitinib –inhibits→ Tyrosine Kinase Activity –related to→ MET gene –has mutation→ MET T540G –associated with→ Papillary Renal Cell Carcinoma*. Such paths provide interpretable justifications by tracing how drug actions relate to disease-specific molecular mechanisms.

Drug-Target Interaction. This task is instantiated as predicting links between compounds and molecular targets, typically genes or proteins. For instance, a predicted interaction between the drug *Bosutinib* and the *MAPT gene*, which is associated with neurodegenerative processes, can be justified by the path: *Bosutinib –inhibits→ FYN Kinase –associated with→ MAPT Gene*. FYN is a tyrosine kinase involved in the phosphorylation of MAPT, and this path suggests a plausible biological mechanism by which Bosutinib could influence MAPT through modulation of kinase activity.

To support our assessment, we rely on three widely used biomedical knowledge graphs that have been adopted in prior research and community benchmarks: Hetionet [12], PrimeKG [5], and OREGANO [4].

Each of these graphs contributes a unique perspective and scope to our study, as outlined below:

- **Hetionet** [12] integrates diverse biomedical data from 29 sources, covering over 45,000 entities (genes, compounds, diseases) and 4.5M triples. It supports discovering biomedical associations via graph inference, notably demonstrated in drug repurposing contexts such as Project Rephetio.
- **PrimeKG** [5] captures biological complexity across molecular, pathway, and therapeutic data with nearly 130,000 entities and 8.1M triples. Designed for precision medicine, it facilitates tasks like drug efficacy prediction and personalized diagnosis.
- **OREGANO** [4] specializes in computational drug repurposing, particularly for natural products. It integrates pharmacological and molecular data across 98,000 entities and 1.6M triples, enabling hypothesis generation for bioactive compounds and phytotherapy.

The specific relation types used for each dataset and task are summarized in Table 1.

Table 1. Triple formats for each task and dataset. The entity to be predicted is shown in **bold**, corresponding to the query of the form $(h, r, ?)$.

Task	Dataset	Triple Format
Drug Repurposing	Hetionet	(Compound, treats, **Disease**)
	PrimeKG	(Drug, drug_indication, **Disease**)
	OREGANO	(Compound, is_substance_that_treats, **Disease**)
Drug-Target Interaction	Hetionet	(Compound, binds, **Gene**)
	OREGANO	(Compound, has_target, **Target**)

3.2 Systems

We evaluate three state-of-the-art KG-based methods for explainable link prediction: Minerva [6], PoLo [18], and REx [24]. These systems were selected because they all generate explanations using reinforcement learning but differ in how they generate and prioritize explanatory paths. Figure 1 shows representative explanation paths generated by each method for the same drug–disease pair. This diversity allows us to study how different strategies affect explanation quality from the perspective of domain experts. To further explore the impact of explanation complexity, we include RExLight, an ablated version of REx in which all ontological expansions are removed from the explanation representation. A structured comparison of the evaluated methods is provided in the supplementary materials.

Minerva. [6] employs a reinforcement learning framework to perform multi-hop reasoning on knowledge graphs by exploring relevant paths connecting entity pairs. The system treats link prediction as a sequential decision-making task, where an RL agent traverses the graph, guided primarily by a reward signal based on reaching correct predictions. It employs an LSTM-based policy network to encode the traversal history and decide subsequent steps in the graph. Predictions are optimized by maximizing the likelihood of reaching correct target entities, without explicitly integrating logical rules or interpretability constraints into its reward structure.

PoLo. [18] enhances the Minerva framework by integrating logical rules directly into the reasoning process. Specifically, PoLo uses predefined logical constraints as explicit background knowledge, significantly influencing the agent's traversal paths. Its reward mechanism comprises two components: one rewarding correct predictions and another explicitly rewarding the selection of paths that correspond to established logical rules. These logical rules are assigned confidence scores based on their predictive strength, guiding the agent to prioritize rule-aligned paths. This approach not only enhances predictive accuracy but also significantly improves the interpretability and justification of predictions by explicitly grounding them in established domain knowledge.

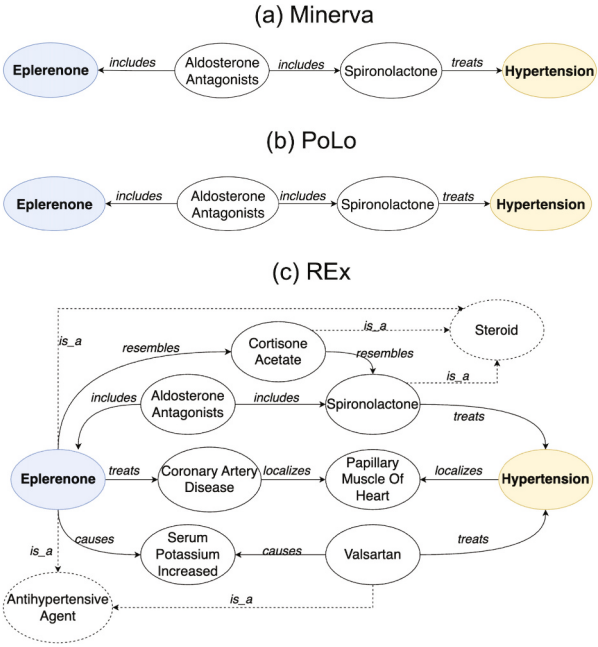

Fig. 1. Example explanation paths generated by each KG-based method for the same drug–disease prediction. Note that Minerva and PoLo yield identical explanations.

REx. [24] explicitly incorporates desirable scientific explanation criteria within its reinforcement learning framework. It employs a dual reward mechanism that balances fidelity, how the prediction matches accurately the missing target, and relevance, quantified through the information content of the nodes in the explanatory paths. In addition to this dual reward, REx integrates an early stopping mechanism that enforces simplicity by limiting path complexity. Together, these components guide the RL agent to discover paths that are not only predictive but also scientifically meaningful, concise, and semantically coherent. Unlike the other methods that produce a single logical path, REx is designed to generate multiple explanatory paths, offering complementary mechanistic insights into the prediction. Additionally, REx can enrich explanatory paths with links to relevant ontology classes, significantly enhancing the comprehensibility and coherence of the generated explanations. Following the original work, the drug repurposing task employs the National Cancer Institute Thesaurus (NCIT) [11] and the Chemical Entities of Biological Interest (ChEBI) [7] ontologies, whereas for the drug-target interaction task, we included the Gene Ontology (GO), further enriching the explanatory framework.

To assess the impact of ontological enrichment on interpretability, we also included RExLight, a variant of REx where ontological expansions are not included. By removing ontology-derived nodes from the generated paths, RExLight offers a leaner explanation structure, allowing us to isolate the effect of semantic abstraction on perceived explanation quality and complexity. All methods were trained and evaluated using the default hyperparameters provided in their original implementations. This ensured a fair and reproducible comparison across systems without additional task-specific tuning.

3.3 Explanation Evaluation Criteria

Our evaluation of explanations focuses on three core aspects of *causal adequacy*, an evidential virtue highlighted in the explanatory framework by Keas [16]. These aspects are critical for assessing whether KG-based explanations can serve as credible scientific justifications:

Relevance. Assesses whether the explanation goes beyond generic triples to provide specific, meaningful insight into the prediction. An explanation can be causally correct but still uninformative if it lacks detail about the mechanism.

Completeness. Evaluates whether the explanation presents enough causal detail to support the prediction convincingly. This often involves combining multiple contributing factors, rather than relying on a single logical path.

Validity. Captures whether the explanation provides a biologically plausible causal mechanism that aligns with current biomedical knowledge [29]. This includes checking that intermediary steps reflect known scientific processes, rather than spurious or indirect associations.

4 User Study Design

We conducted a qualitative user study with biomedical experts to evaluate the relevance, completeness, and validity of path-based explanations generated by three approaches (full details in supplementary materials). Our goal was to understand how explanation content and structure influence users' ability to interpret novel hypotheses in drug development and their impact on trust in an explainable system.

To guide both the design and reporting of the study, we adopt the framework proposed by Pesquita et al. [26], which outlines six core dimensions for conducting empirical user studies in Semantic Web contexts. These include: (i) purpose, (ii) users, (iii) tasks, (iv) setup, (v) procedure, and (vi) analysis and presentation of data. A description can be found in Table 2.

4.1 Users

The intended users of KG-based explanations in drug development are experts involved in therapeutic discovery and validation. We consider that different professional roles might value explanatory properties differently, based on their

specific domain expertise and decision-making objectives. For example, while Pharmacologists, who investigate drug effects on biological systems, are likely to prefer explanations that highlight a drug's mechanism of action and therapeutic potential, Molecular Biologists who study disease mechanisms at a molecular level might favor a broader explanatory landscape based on biochemical pathways. Furthermore, Bioinformaticians, who use computational methods to streamline drug discovery processes, may place greater value on logically coherent, quantitatively supported explanations, facilitating integration with existing computational workflows.

Table 2. Overview of the user study design following the six dimensions proposed by Pesquita et al. [26].

Dimension	Description
Purpose	Assess how biomedical researchers interpret KG-based explanations in predictive tasks, focusing on perceived usefulness and trustworthiness. Explanations aimed to support hypothesis generation and exploratory analysis.
Users	Eleven participants with backgrounds in biomedical and health sciences (including PhD students, researchers, and one professor) were recruited from multiple institutions (academia and industry) across Portugal and Spain. The sample size aligns with established qualitative usability and interpretability research practices, where approximately 10±2 expert participants are typically sufficient to identify the majority of usability patterns [14]. Given the difficulty of recruiting domain experts in biomedical research, the study prioritized depth of expert engagement over statistical generalization.
Tasks	Two tasks reflecting real-world inference settings: *Drug Repurposing* and *Drug–Target Interaction*, each with 10 predictions and 4 explanations. Participants rated each explanation on Relevance, Completeness, and Validity.
Setup	The study used an online form. Each participant received a personalized form with separate tabs for consent, tasks, and final trust ratings. All participants viewed tasks and systems in a fixed order.
Procedure	Participants completed consent and background info, then assessed the explanations. The study was conducted remotely or in person and typically lasted over 60 min.
Analysis and Presentation of Data	Quantitative data were summarized using descriptive statistics. Free-text feedback was analyzed thematically using manual coding and topic modeling.

5 Evaluation

5.1 Predictive Performance

To examine the tradeoffs between accuracy and interpretability, we investigated the predictive performance of the three evaluated systems across both tasks and datasets. Table 3 presents results using standard link prediction metrics—Hits@1, Hits@3, Hits@10, and Mean Reciprocal Rank (MRR)—to provide a fine-grained view of each method's effectiveness.

Table 3. Performance of all systems on drug repurposing and drug-target interaction across Hetionet, PrimeKG, and OREGANO, reported using Hits@1, Hits@3, Hits@10, and MRR metrics.

Task	Dataset	Method	Hits@1	Hits@3	Hits@10	MRR
Drug Repurposing	Hetionet	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**
	PrimeKG	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**
	OREGANO	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**
Drug-Target Interaction	Hetionet	MINERVA	0.079	0.193	0.375	0.170
		PoLo	0.104	0.206	0.373	0.186
		REx	**0.107**	**0.224**	**0.393**	**0.197**
	OREGANO	MINERVA	0.087	0.206	0.392	0.181
		PoLo	0.151	0.290	**0.488**	0.256
		REx	**0.174**	**0.312**	0.374	**0.264**

Across all datasets and tasks, REx consistently achieved top performance in Hits@1 and MRR. In drug repurposing, it ranked first on both Hetionet and PrimeKG, and matched PoLo on Hits@1 in OREGANO. These results highlight REx's effectiveness in prioritizing correct candidates early in the ranking.

For drug–target interaction, overall scores were lower, reflecting the increased challenge of gene-level predictions. REx again led in all metrics on Hetionet. On OREGANO, REx ranked highest on Hits@3 and MRR, whereas PoLo achieved a higher Hits@10, suggesting a tradeoff between early precision and broader recall.

While PoLo and Minerva remained competitive in some datasets, they often trailed in top-rank performance. The consistency of REx across datasets and tasks supports its robustness for link prediction in drug development tasks.

5.2 User Study Results

Users Profile and Background. We recruited eleven participants with a background in life and health sciences and different levels of familiarity with AI and KGs. 82% were aged between 25–34, while the remaining fit on the 35–49 range. Participants reported their academic background across different fields. As shown in Fig. 2, most participants had advanced training in Life Sciences, and a broad mix of experience in the other fields. Participants also self-assessed their expertise in *Knowledge Graphs*, *AI Systems*, and *Molecular Biology*. Most participants reported being competent or expert in molecular biology and AI systems, while knowledge graph familiarity was more varied, with four participants reporting no prior exposure. Overall, the participants represented a diverse and relevant sample for the study, reflecting a range of roles and expertise commonly involved in drug development.

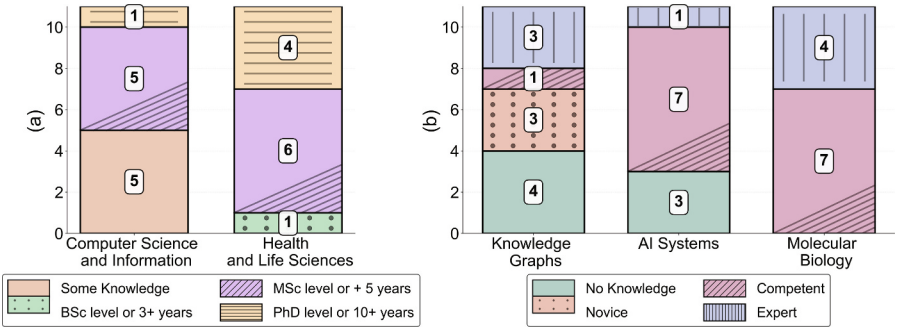

Fig. 2. Self-reported background and domain expertise. (a) Academic background across four fields. (b) Domain expertise across Knowledge Graphs, AI Systems, and Molecular Biology.

Quantitative Evaluation Across Tasks. Participants rated each explanation on relevance and validity on a scale of 1 (least) to 5 (most). Completeness was originally assessed using a midpoint-optimal scale, where three denoted *complete* and extremes (1 and 5) represented *too simple* and *too complex*, respectively. Only prediction instances correctly solved by all evaluated models were shown to experts, ensuring a fair comparison of explanation quality independent of predictive performance. Figure 3 presents heatmap ratings for all four explanation systems across both evaluated tasks. To facilitate comparison across criteria, we transformed the completeness scale in the heatmaps to reflect suitability for the task, assigning higher values to ratings closer to the midpoint.

Across both tasks, REx consistently received the highest scores in relevance and validity, suggesting it provided the most scientifically plausible and contextually informative explanations. Interestingly, both REx and RExLight have

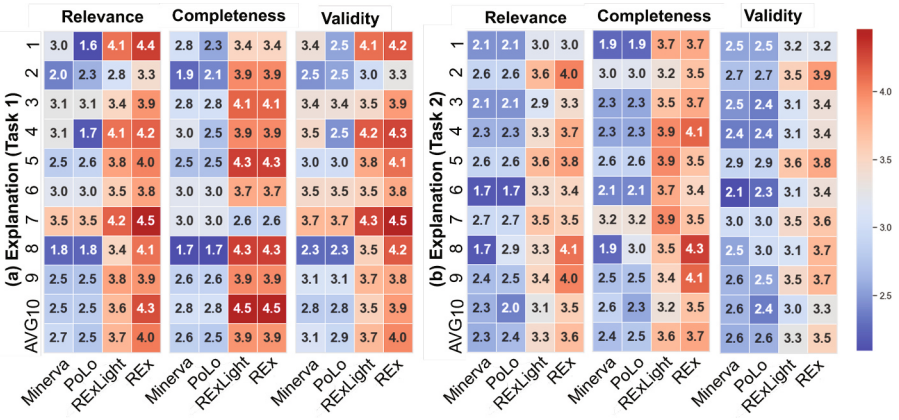

Fig. 3. Heatmaps showing participant ratings of the explanations for each system across the three evaluation criteria. Each row represents one explanation instance, and the final row is the average explanation rating.

comparable transformed completeness scores, striking a better balance between detail and conciseness. Minerva and PoLo received consistently lower ratings across all metrics in both tasks, particularly in completeness, indicating their explanations were often viewed as too simplistic.

To further elucidate the different aspects of completeness, we plotted its original distribution per system in Fig. 4. This reveals that while Minerva and PoLo produced overly simplistic explanations, REx achieved the best completeness, but occasionally verged on excessive complexity, whereas RExLight's performance was more heavily influenced by the perceived complexity of the task, with results considerably lower in drug-target interaction.

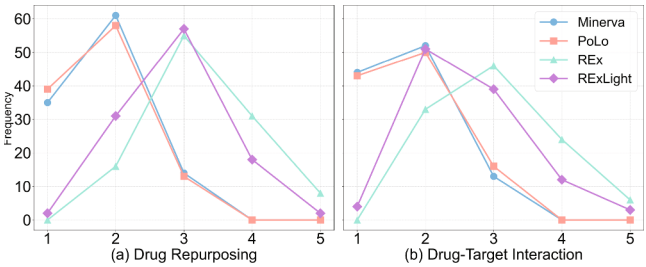

Fig. 4. Completeness ratings for each system on each task (1- too simple, 3- complete, 5 - too complex).

Qualitative Analysis. To further interpret the qualitative feedback provided by participants as free text in the user study, we applied topic modeling using the BERTopic framework [10]. BERTopic combines transformer-based embeddings (we used `all-mpnet-base-v2`) with density-based clustering (HDBSCAN) and topic representation techniques. In our case, we used the `KeyBERTInspired` module to produce more semantically meaningful keywords per topic.

We preprocessed and embedded 103 comments from both the drug repurposing and drug–target interaction tasks. After clustering, we identified three dominant topics reflecting common themes in participants' responses. Notably, 87% of participants reported a preference for mechanistic explanations, with 60 out of 103 comments explicitly reflecting this preference.

1. **Mechanistic and Molecular Similarity:** This topic includes comments that emphasize or critique mechanistic plausibility and biochemical similarities. Participants noted references to molecular classes such as prostaglandins and adenosine analogs. While some found these helpful, others questioned whether structural similarity alone justified the predicted associations.
2. **General Prediction Logic:** Comments grouped here reflected on the overall structure and clarity of the predictions. Users compared systems based on completeness, step-by-step reasoning, and whether the logic appeared deductive or merely associative. This topic also captured concerns about the visibility of key relational evidence.
3. **Ontological Reasoning and Logical Redundancy:** This theme grouped feedback critical of explanations that relied heavily on relationships like *is_ a* or *causes (side effect)* as being circular or lacking depth. Several participants also pointed out that ontological abstraction failed to capture domain-specific causality.

These themes summarize the experiences of participants. They appreciated the mechanistic depth of REx, but found its ontological expansions occasionally redundant or unclear. RExLight, by contrast, maintained interpretability despite a simpler graph representation but lacked contextual layering. Minerva and PoLo were often seen as easier to interpret, yet limited across all criteria.

Overall System Evaluation. At the end of the study, participants rated how likely they would be to trust a hypothesis based on the explanation provided by each system. Trust was assessed using a 5-point Likert scale, where 1 indicated no trust and 5 indicated full trust.

As shown in Fig. 5, participants placed greater confidence in systems that presented more detailed or structured explanations (REx and RExLight), while simpler explanations (Minerva and PoLo) were seen as insufficient to justify trust in the prediction.

The overall results show that REx outperforms other methods in both predictive performance, ranking first in Mean Reciprocal Rank (MRR) across all tasks, and explainability, obtaining the highest average scores for relevance, completeness, validity, and trustworthiness.

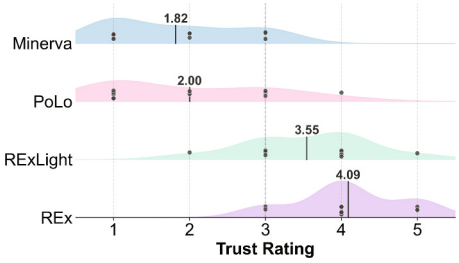

Fig. 5. Final trust ratings distribution for each system.

6 Discussion

The comparative evaluation and user study revealed clear patterns in how professional background shaped the interpretation of path-based explanations, directly informing our first research question on how domain expertise influences preferences for KG-based explanations. Participants with limited AI or KG expertise (n=4) tended to prefer explanations with ontological expansion (e.g., "I prefer explanations that include ontological expansions that are pertinent to side-effect context to strengthen prediction or association"). In contrast, participants with backgrounds in AI or KGs (n=4) were more critical, particularly regarding completeness, expecting multi-hop reasoning and structured justifications (e.g.,"I prefer explanations that establish clear mechanistic links rather than relying solely on shared symptoms"). Molecular biologists (n=4) were the most demanding in terms of biological detail, especially in the drug–target interaction task (e.g., "I prefer explanations supported by multiple mechanistic or genetic details to reinforce direct drug–disease links—because shared effects or component types alone don't guarantee therapeutic equivalence. Completeness ratings often reflected different mental models: computational experts penalized short explanations lacking intermediate reasoning, while life-science participants perceived more complex structures as more informative. Relevance ratings varied more idiosyncratically, indicating that perceived informativeness depended mainly on the specific content of each explanation.

These findings also relate to our second research question on explanation strategies, revealing how multi-path and ontologically enriched explanations affect perceived completeness and trustworthiness. Participants consistently valued explanations that aligned with known biomedical knowledge and mirrored pharmacological reasoning. The most appreciated explanations included pharmacologically meaningful information, such as alternative drugs with similar mechanisms, combined mechanistic and therapeutic links supported by signals like adverse effects or gene associations, and used ontological types to clarify less familiar nodes. For example, explanations connecting a candidate drug to another with a shared mechanism, reinforced by an adverse effect or gene-level relation, were often considered both complete and biologically grounded. Multi-

path explanations were seen as particularly effective when they offered complementary biological perspectives, rather than redundancy.

In the drug–target interaction task, participants applied stricter validity criteria, requiring explanations to reflect mechanistic evidence such as binding interactions or pathway participation—again highlighting the importance of causal adequacy. Explanations that relied on shared side effects or clinical symptoms were frequently rejected as scientifically implausible. A typical example was the rejection of an explanation connecting two drugs through the effect *diabetic coma*, which participants deemed insufficient to establish a therapeutic relationship. Users consistently preferred explanations grounded in functional or causal reasoning rather than associative or hierarchical similarities, underscoring that explanatory strength depends on biological plausibility rather than structural connectivity alone.

Ontological information also played a nuanced role in interpretability, as explored in our third research question. It was welcomed when it added implicit biological context, such as identifying a compound as a steroid, which provided interpretive cues about potential mechanisms of action. However, when ontological labels merely reiterated well-known classifications—such as marking a cancer drug as an antineoplastic agent—they were viewed as redundant. The feedback indicates that semantic abstraction should be employed selectively, offering conceptual enrichment only when it clarifies or deepens understanding.

6.1 Limitations

The study design itself highlighted some limitations. The fixed order of presentation, from simpler to more complex systems, may have introduced a perceived ranking that affected participants' evaluations. To investigate this concern, we conducted a small mixed-effects analysis focusing on Minerva and PoLo, which share identical single-path layouts, making any advantage to PoLo attributable solely to presentation order. The analysis revealed negligible differences (mean differences ≤ 0.08 Likert points) with no significant order effects (all $p \geq 0.34$), suggesting that position did not substantially influence ratings. Nevertheless, future studies should randomize or counterbalance the presentation order to mitigate potential biases. Another concern was the lack of visible data provenance: several participants felt that explanations were less credible without explicit references to curated databases such as DrugBank, even though those sources were present in the underlying KGs. Providing transparent provenance information is therefore crucial to improve the perceived trustworthiness of explainable KG systems.

7 Conclusion

This work investigates how KG–based methods generate scientifically meaningful and trustworthy explanations in drug development. Through a comparative evaluation combining predictive metrics and expert feedback, we examined how

explanation structure and user background shape perceptions of relevance, completeness, and validity in hypothesis generation.

Our results reveal distinct patterns in expert preferences across tasks and expertise profiles, showing how professional background and explanation design jointly influence perceived usefulness and trust. Experts favored explanations grounded in mechanistic reasoning and biomedical context over purely associative patterns, underscoring the need to align explainable AI with the epistemic norms of scientific reasoning.

While current systems show strong potential, future work includes refining relation selection and deepening the integration of domain knowledge to better justify biological mechanisms, incorporating expert-derived explanations, exploring personalized and adaptive strategies (e.g., task-aware ontological enrichment), and assessing the generalizability of the proposed evaluation framework in additional domains. Developing explanation strategies that adapt to expert reasoning and investigative contexts will be key to fostering trust and advancing the scientific adoption of KG-based methods in biomedical discovery.

Acknowledgments. This work was supported by FCT through the fellowship https://doi.org/10.54499/2023.00653.BD, and the LASIGE Research Unit, ref. UID/00408/2025. It was also partially supported by the KATY project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101017453, and by the CancerScan project which received funding from the European Union's Horizon Europe Research and Innovation Action (EIC Pathfinder Open) under grant agreement No. 101186829. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Innovation Council and SMEs Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Supplemental Material Statement Source. All resources are available at https://github.com/liseda-lab/KG-Explanations-Study

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdul Raheem, A.K., Dhannoon, B.N.: Comprehensive review on drug-target interaction prediction-latest developments and overview. Curr. Drug Discov. Technol. **21**(2), 56–67 (2024)
2. Bang, D., Lim, S., Lee, S., Kim, S.: Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers. Nat. Commun. **14**(1), 3570 (2023)
3. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. Adv. Neural Inf. Process. Syst. **26** (2013)
4. Boudin, M., Diallo, G., Drancé, M., Mougin, F.: The oregano knowledge graph for computational drug repurposing. Sci. Data **10**(1), 871 (2023)

5. Chandak, P., Huang, K., Zitnik, M.: Building a knowledge graph to enable precision medicine. Sci. Data **10**(1), 67 (2023)
6. Das, R., et al.: Go for a walk and arrive at the answer: reasoning over paths in knowledge bases using reinforcement learning. arXiv preprint arXiv:1711.05851 (2017)
7. Degtyarenko, K., et al.: Chebi: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. **36**(suppl_1), D344–D350 (2007)
8. Edwards, G., Nilsson, S., Rozemberczki, B., Papa, E.: Explainable biomedical recommendations via reinforcement learning reasoning on knowledge graphs. arXiv preprint arXiv:2111.10625 (2021)
9. Gottlieb, A., Stein, G.Y., Ruppin, E., Sharan, R.: Predict: a method for inferring novel drug indications with application to personalized medicine. Mol. Syst. Biol. **7**(1), 496 (2011)
10. Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022)
11. Hartel, F.W., de Coronado, S., Dionne, R., Fragoso, G., Golbeck, J.: Modeling a description logic vocabulary for cancer research. J. Biomed. Inform. **38**(2), 114–129 (2005)
12. Himmelstein, D.S., et al.: Systematic integration of biomedical knowledge prioritizes drugs for repurposing. elife **6**, e26726 (2017)
13. Huang, K., et al.: A foundation model for clinician-centered drug repurposing. Nat. Med. **30**(12), 3601–3613 (2024)
14. Hwang, W., Salvendy, G.: Number of people required for usability evaluation: the 10±2 rule. Commun. ACM **53**(5), 130–133 (2010)
15. Jiménez-Luna, J., Grisoni, F., Schneider, G.: Drug discovery with explainable artificial intelligence. Nat. Mach. Intell. **2**(10), 573–584 (2020)
16. Keas, M.N.: Systematizing the theoretical virtues. Synthese **195**(6), 2761–2793 (2018)
17. Liao, Q., et al.: Application of artificial intelligence in drug-target interactions prediction: a review. npj Biomed. Innov. **2**(1), 1 (2025)
18. Liu, Y., Hildebrandt, M., Joblin, M., Ringsquandl, M., Raissouni, R., Tresp, V.: Neural Multi-hop Reasoning with Logical Rules on Biomedical Knowledge Graphs. In: Verborgh, R., et al. (eds.) ESWC 2021. LNCS, vol. 12731, pp. 375–391. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-77385-4_22
19. Meilicke, C., Chekol, M.W., Ruffinelli, D., Stuckenschmidt, H.: Anytime bottom-up rule learning for knowledge graph completion. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, pp. 3137–3143 (2019)
20. Menden, M.P., Wang, D., Mason, M.J., Szalai, B., Bulusu, K.C., Guan, Y., Yu, T., Kang, J., Jeon, M., Wolfinger, R., et al.: Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. Nat. Commun. **10**(1), 2674 (2019)
21. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
22. Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans. Inter. Intell. Syst. (TiiS) **11**(3–4), 1–45 (2021)
23. Napolitano, F., Carrella, D., Mandriani, B., Pisonero-Vaquero, S., Sirci, F., Medina, D.L., Brunetti-Pierri, N., Di Bernardo, D.: gene2drug: a computational tool for pathway-based rational drug repositioning. Bioinformatics **34**(9), 1498–1505 (2018)

24. Nunes, S., Badreddine, S., Pesquita, C.: Rewarding explainability in drug repurposing with knowledge graphs. In: Kwok, J. (ed.) Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25, pp. 4624–4632. International Joint Conferences on Artificial Intelligence Organization (2025). https://doi.org/10.24963/ijcai.2025/515, main Track
25. Ozkan, E., Celebi, R., Yilmaz, A., Emonet, V., Dumontier, M.: Generating knowledge graph based explanations for drug repurposing predictions. In: SWAT4HCLS, pp. 22–31 (2023)
26. Pesquita, C., Ivanova, V., Lohmann, S., Lambrix, P.: A Framework to Conduct and Report on Empirical User Studies in Semantic Web Contexts. In: Faron Zucker, C., Ghidini, C., Napoli, A., Toussaint, Y. (eds.) EKAW 2018. LNCS (LNAI), vol. 11313, pp. 567–583. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-03667-6_36
27. Ponzoni, I., Páez Prosper, J.A., Campillo, N.E.: Explainable artificial intelligence: a taxonomy and guidelines for its application to drug discovery. Wiley Interdisc. Rev. Comput. Mol. Sci. **13**(6), e1681 (2023)
28. Pushpakom, S., Iorio, F., Eyers, P.A., Escott, K.J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., et al.: Drug repurposing: progress, challenges and recommendations. Nat. Rev. Drug Discovery **18**(1), 41–58 (2019)
29. Rosales, A., Morton, A.: Scientific explanation and trade-offs between explanatory virtues. Found. Sci. **26**, 1075–1087 (2021)
30. Schlichtkrull, M., Kipf, T.N., Bloem, P., Van Den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: European Semantic Web Conference, pp. 593–607. Springer (2018)
31. Stork, L., Tiddi, I., Spijker, R., ten Teije, A.: Explainable drug repurposing in context via deep reinforcement learning. In: Pesquita, C., et al. (eds.) European Semantic Web Conference, pp. 3–20. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-33455-9_1
32. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. In: International Conference on Machine Learning, pp. 2071–2080. PMLR (2016)
33. Xiong, W., Hoang, T., Wang, W.Y.: Deeppath: a reinforcement learning method for knowledge graph reasoning. arXiv preprint arXiv:1707.06690 (2017)
34. Zhang, S., et al.: Page-link: Path-based graph neural network explanation for heterogeneous link prediction. In: Proceedings of the ACM Web Conference 2023, pp. 3784–3793 (2023)
35. Zhavoronkov, A., et al.: Deep learning enables rapid identification of potent ddr1 kinase inhibitors. Nat. Biotechnol. **37**(9), 1038–1040 (2019)
36. Zhou, G., Haihong, E., Kuang, Z., Tan, L., Xie, X., Li, J., Luo, H.: Clinical decision support system for hypertension medication based on knowledge graph. Comput. Methods Programs Biomed. **227**, 107220 (2022)