User Study Design

1 Design

To guide both the design and reporting of the study, we adopt the framework proposed by [1], 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.

1.1 Purpose

This study investigates how biomedical researchers interpret explanations generated by KG-based systems in predictive settings. The explanations aim to support hypothesis generation and exploratory analysis by offering mechanistic and ontological evidence for model predictions. This situates the study within the broader category of "learn and understand" objectives, as participants were asked to evaluate the plausibility of predicted links between biomedical entities.

1.2 Users

The study involved 11 participants recruited from multiple institutions, primarily with academic or research backgrounds in biomedical and health-related sciences. The sample included PhD students, one professor, and researchers both junior and senior in academia and industry setting. Participants represented diverse fields including life sciences, neurosciences, microbiology, bioinformatics, and health sciences. While they differed in their exposure to AI or explainable systems, all had domain knowledge relevant to interpreting biomedical relationships between drugs, diseases, and genes.

Participants self-reported their academic background and domain expertise across Molecular Biology, Knowledge Graphs, and AI systems using a four-level scale ("No Knowledge" to "Expert"). They also indicated how likely they were to trust AI-generated decisions in healthcare. All participants were informed that the evaluated links were model predictions and were instructed to assess the explanations, not the correctness of the predictions themselves.

The study was conducted in both in-person and remote formats. Participants first reviewed the study goals and consented to participate using a structured form.

As no sensitive or personally identifying information was collected, and all participation was anonymous and optional, no formal ethics approval was required.

1.3 Tasks

The study was structured around two main tasks with 10 explanations each, designed to reflect real-world biomedical inference scenarios: Task 1 is Drug Repurposing and Task 2 is Drug–Target Interaction.

Each prediction was accompanied by four explanation graphs, one per system: *Minerva* (System 1), *PoLo* (System 2), *RExLight* (System 3, without ontological expansion), and *REx* (System 4, full version). Explanations connected the predicted entities (e.g., drug and disease) through intermediate nodes and labeled relations sourced from Hetionet.

Each explanation was rendered using the same visual style and layout technique to control for presentation bias and focus user attention on content differences rather than appearance. Explanations connected the predicted entities (e.g., drug and disease) through intermediate nodes and labeled relations sourced from Hetionet.

Participants rated each explanation on a 5-point Likert scale across three dimensions of explanation quality:

Scientific Validity. How scientifically correct, plausible, and coherent the explanation is based on existing biomedical knowledge.

Completeness. The extent to which the explanation provides sufficient detail to make the prediction understandable. It should be informative without being unnecessarily overwhelming.

Relevance. Relates to the usefulness and informativeness of the explanation for understanding the prediction. An explanation can be scientifically valid but still irrelevant if it does not help clarify why the prediction matters or how it connects to the task at hand.

Optional free-text comments allowed users to elaborate on strengths and weaknesses. At the end of the study, participants also rated how much they would trust predictions based on the explanations from each system.

1.4 Setup

The study was implemented using an online forms which served as both presentation interface and response form. Each participant received a personalized spreadsheet with the following structure:

- A background and consent tab;
- One tab for each task, presenting the 10 predictions with four explanation each;
- A final tab to rate overall prediction trustworthiness based on each system.

All users received the tasks in the same fixed order (Task 1 then Task 2), and explanation systems were shown in a consistent left-to-right order across predictions. No interaction logs were collected, and users completed the study at their own pace.

1.5 Procedure

After recruitment via email or personal invitation, participants were given access to their individualized study sheet. For in-person sessions, a facilitator explained the goals and responded to questions before and during the task. Remote participants received the same information in written form and could request clarification as needed via mail or online meeting.

Participants began by completing consent form and the background questionnaire, then proceeded to the task tabs. Most participants completed the study in more than 60 minutes, depending on the level of written feedback provided.

1.5.1 Analysis and Presentation of Data

We applied a mixed-methods approach. Likert-scale responses were aggregated per system and task using descriptive statistics. Results are presented through summary tables and heatmaps, capturing both overall trends and per-explanation variation.

Free-text comments were thematically analyzed using a combination of manual coding and topic modeling. Emergent themes were grouped by system and task, supporting the interpretation of quantitative trends.

References

[1] Pesquita, C., Ivanova, V., Lohmann, S., Lambrix, P.: A framework to conduct and report on empirical user studies in semantic web contexts. In: Knowledge Engineering and Knowledge Management: 21st International Conference, EKAW 2018, Nancy, France, November 12-16, 2018, Proceedings 21. pp. 567–583. Springer (2018)