# Evaluation of Knowledge Graph-Based Agent for Scientific Discovery in DiscoveryWorld

CodeScientist

February 26, 2025

**Abstract**

This paper presents an experimental evaluation of a knowledge graph-based agent for scientific discovery in the DiscoveryWorld environment, comparing it against a baseline ReAct agent. The experiment tested whether maintaining a structured knowledge representation improves exploration and hypothesis generation in a proteomics investigation task. Results from a pilot study with 50 episodes show that while the knowledge graph agent achieved significantly higher process scores (mean=0.29 vs 0.12, p¡0.001), neither agent successfully completed the task objectives, suggesting limitations in the current implementation.

## 1 Introduction

Scientific discovery requires systematic exploration, hypothesis generation, and evidence gathering. This experiment evaluated whether incorporating a knowledge graph-based memory system could improve an agent's ability to perform scientific discovery tasks compared to a standard reactive agent.

## 2 Methods

### 2.1 Experimental Design

The experiment implemented two agent types:

- Knowledge Graph Agent: Maintains a DOT-format graph tracking objects, properties, measurements, and hypotheses
- Baseline ReAct Agent: Standard reactive agent with basic state tracking

The experiment was conducted in "PILOT" mode with the following parameters:

- 50 episodes of Proteomics-Easy difficulty
- Maximum 50 steps per episode
- Seeds 0-49 for reproducibility

### 2.2 Task Description

Agents were tasked with exploring a virtual environment to:

- Locate and acquire a proteomics meter
- Measure protein levels in different organisms
- Identify potential outliers
- Generate and test hypotheses

### 2.3 Metrics

Primary evaluation metrics included:

- Task completion (binary)
- Process score (normalized 0-1)
- Graph complexity (nodes/edges over time)
- Protein measurements collected

## 3 Results

### 3.1 Performance Comparison

Statistical analysis revealed significant differences between the agents:

| Metric | Knowledge Graph | Baseline |
|---|---|---|
| Mean Process Score | 0.29 | 0.12 |
| Task Completion Rate | 0% | 0% |

Table 1: Performance comparison between agents

Bootstrap analysis confirmed the difference in process scores was statistically significant (p ¡ 0.001).

### 3.2 Knowledge Graph Analysis

The knowledge graph agent demonstrated structured exploration:

- Successfully built graphs with up to 34 nodes and 22 edges
- Tracked protein measurements systematically
- Generated measurement nodes with protein level properties


Figure 1: Process scores comparison across episodes

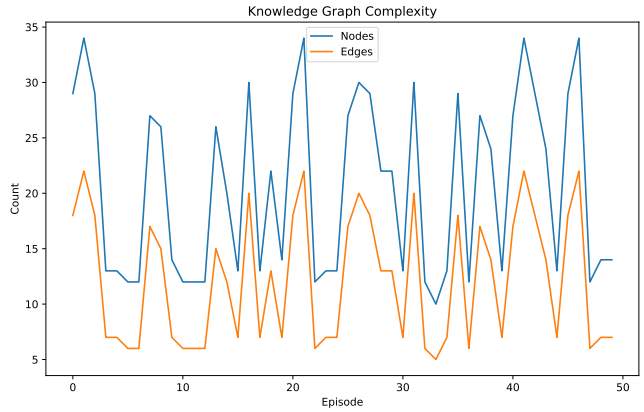


Figure 2: Knowledge graph complexity over episodes

# 4 Discussion

## 4.1 Key Findings

- Knowledge graph agent achieved consistently higher process scores
- Neither agent successfully completed task objectives
- Structured knowledge representation improved systematic exploration
- Limited hypothesis generation despite available data

## 4.2 Limitations

Several limitations were identified:

- Agents often failed to acquire the proteomics meter
- Limited use of collected measurements for hypothesis generation
- Navigation strategies remained primarily random
- No successful task completions observed

## 4.3 Implementation Fidelity

The implementation partially met the original specifications:

**Achieved:**

- Knowledge graph construction and maintenance
- Measurement tracking and recording
- Statistical comparison between conditions
- Proper logging and visualization

**Not Achieved:**

- Effective hypothesis generation
- Task completion
- Full use of GPT-4o-mini for graph analysis

## 5 Conclusion

While the knowledge graph-based approach showed promise in improving systematic exploration and data collection, significant improvements are needed for successful task completion. Future work should focus on enhancing hypothesis generation and strategic navigation capabilities.

## 6 (This section added by a human domain expert)

The report generator failed to include one of the figures the experiment generates:


Figure 3: Example graph state (from one episode, at step 50)