Protein Design with Agent Rosetta: A Case Study for Specialized Scientific Agents

J. Teneggi*, S.M. B. A. Turzo*, T. Marwah, A. Bietti, P. D. Renfrew, V. K. Mulligan, S. Golkar

Polymathic AI

This research was supported in part by the AI2050 program at Schmidt Sciences (Grant G-25-70028) and the Simons Foundation.

Research Question

Can LLMs use niche software for real-world scientific tasks?

Background

Protein design is a central task with broad implications for real-world scientific pipelines (e.g., pharmaceuticals, material science).

Rosetta

Rosetta [1] is the leading collection of physics-based software for heteropolymer design, docking, and structure prediction.

It is particularly useful in data-scarce applications where training deep learning models may be prohibitive. For example, when designing with non-canonical amino acids or geometries.

While powerful, Rosetta can be difficult to use.

RosettaScripts

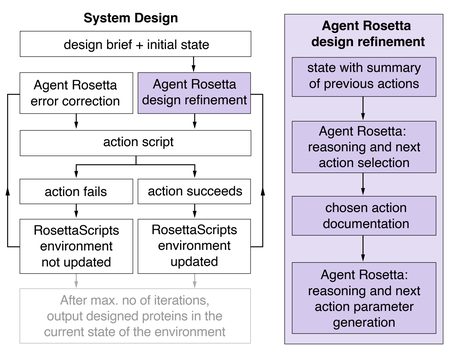
RosettaScripts [2] is Rosetta's XML scripting interface. It is a natural testbed for AI agents because:

- It leverages the knowledge and reasoning capabilities of LLMs.
- It combines reasoning with a structured output format.
- It supports scaling over thousands of compute nodes.

An LLM agent fluent in RosettaScripts would enable broad design pipelines while increasing accessibility.

Agent Rosetta

Agent Rosetta achieves user-defined goals by iterative refinement of protein designs with a tailored RosettaScripts environment.



The Importance of Environment Design

Prompting alone was not sufficient to integrate frontier LLMs with niche scientific software

Generating Valid RosettaScripts Syntax

We evaluated different LLMs at generating compositional penalty blocks, fundamental to guide protein design in Rosetta.

All LLMs struggled at translating their (correct) reasoning into valid compositional penalty blocks. To solve this issue, we abstracted RosettaScripts' syntax to align with the LLMs reasoning:

Prompt

Write a compositional penalty block that penalizes more than 5 prolines. Use a linear boundary with slope of 10.

Gemini 2.5 Flash

```
PENALTY_DEFINITION
TYPE PRO
ABSOLUTE 5
DELTA_START 1
DELTA_END 2
PENALTIES 10 20
BEFORE_FUNCTION CONSTANT
AFTER_FUNCTION LINEAR
END_PENALTY_DEFINITION
```

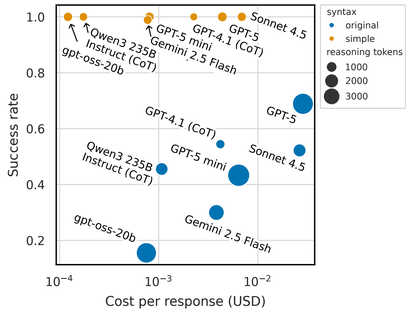
Claude Sonnet 4.5

```
PENALTY_DEFINITION
TYPE PRO
ABSOLUTE 5
DELTA_START -5
DELTA_END 0
PENALTIES 0 0 0 0 -10 0
BEFORE_FUNCTION CONSTANT
AFTER_FUNCTION LINEAR
END_PENALTY_DEFINITION
```

Simplified syntax

```
PENALTY_DEFINITION
TYPE P
SHAPE ABOVE
TARGET 5
RADIUS 0
BOUNDARY LINEAR
STRENGTH 10
END_PENALTY_DEFINITION
```

Our simplified syntax reduces costs and improves performance across all model sizes, highlighting the importance of environment design to interface LLMs with scientific software.



Environment Ablation

RosettaScripts is a complex collection of software that requires generalization beyond the context of the base LLM.

We tested frontier LLMs at generating RosettaScripts protocols from scratch, without the scaffolding automated by our environment.

All LLMs failed sharply when the task required generalization beyond the prompting instructions:

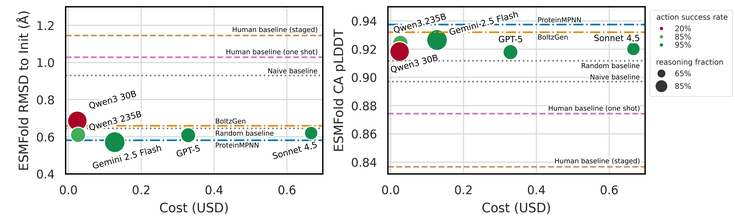
Model	Success rate (%)			
	Task 1 (easy)	Task 2 (medium)	Task 3 (medium)	Task 4 (hard)
GPT-5	100	100	40	0
Gemini 2.5 Flash	90	10	0	0
Qwen3 235B Instruct (CoT)	90	20	0	0
Sonnet 4.5	90	100	90	0

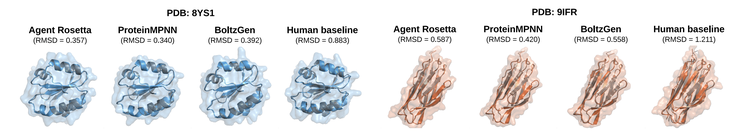
Matching Deep Learning Models

We benchmarked Agent Rosetta on fixed-backbone sequence design with canonical amino acids against existing deep learning models.

We found Agent Rosetta generated designs with similar predicted RMSD and pLDDT.

Agent Rosetta also outperformed fixed design protocols written by scientists with extensive RosettaScripts knowledge.



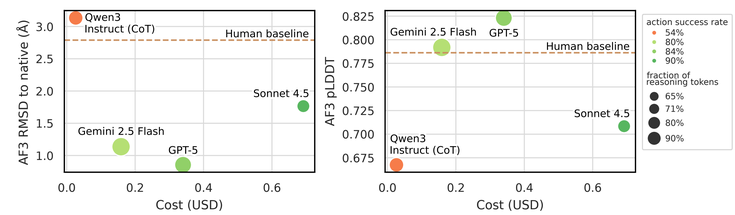


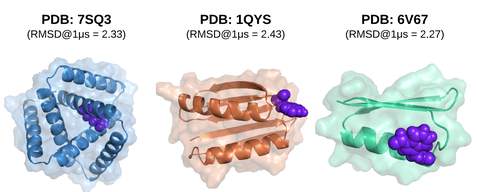
Going Beyond Machine Learning Models

We verified Agent Rosetta can perform tasks with non-canonical amino acids, difficult to model with deep learning approaches.

We considered inserting a modified tryptophan in the core of de novo structures. We validated designs with MD simulations.

Agent Rosetta outperformed a static human-written protocol, the only baseline for this task.





References

[1] Leaver-Fay et al. "Rosetta 3", 2011.
[2] Fleishman et al. "Rosettascripts: a scripting language interface to the Rosetta macromolecular modeling suite.", 2011.