HEIMDALL

Evaluating AI Reasoning in Protein Structural Biology: Results from a Reinforcement Learning Benchmark

Artificial intelligence systems are increasingly being applied to problems in molecular biology, drug discovery, and protein engineering. However, many modern models still struggle with agentic reasoning in protein biochemistry, particularly when tasks require analysis of protein structures rather than simple retrieval of biological facts. Real-world protein science demands the ability to interrogate structural data, interpret biochemical interactions, compute geometric relationships, and integrate experimental metadata. Current benchmarks rarely test these capabilities directly, leaving a gap between the apparent knowledge of AI systems and their ability to perform practical structural reasoning tasks.

To address this challenge, Heimdall Bio is developing a large-scale reinforcement learning (RL) training set focused on protein biochemistry and structural biology. The dataset consists of thousands of manually authored and verified questions taken from experimentally determined or predicted protein structures. These questions require models to perform tasks such as parsing crystallographic metadata, interpreting structural interactions, computing geometric features from atomic coordinates, and applying biochemical reasoning to real molecular systems. By grounding each question in experimentally validated structures, the dataset provides a deterministic and verifiable reward signal suitable for RL and evaluation of protein-focused AI systems.

As an initial benchmark, a random sample of twenty protein structural RL questions from the training set was evaluated across several independent AI models. The questions spanned a range of structural biology tasks, including secondary structure determination, torsion angle calculations, crystallographic metadata interpretation, and biochemical functional analysis. Each model was evaluated under identical conditions, and responses were scored strictly against the verified answer set. This sampling provides an early glimpse into the ability of current AI systems to perform structured reasoning in protein science.

The results highlight a substantial gap between current capabilities and the demands of practical structural biology (Figure 1). Among the tested models, Claude 4.6 achieved the highest score at a grade of 66.6%, followed by ChatGPT 5.2 at 46.6%, Gemini 3 at 41.3%, Grok 4.2 at 40.0%, Meta 4 at 8.3%, and Mistral AI 3 at 7.5%. While some models demonstrate moderate competence, none approached full accuracy, indicating that current systems frequently rely on heuristic or text-based reasoning rather than true structural analysis.

These findings illustrate a clear opportunity for improvement in AI systems regarding molecular science. By providing a large, rigorously validated set of protein-focused reinforcement learning questions, Heimdall aims to enable the next generation of AI models capable of biochemical and structural reasoning. Such capabilities are essential for applications ranging from protein engineering and enzyme design to drug discovery and biosurveillance. The Heimdall RL protein benchmark represents an important step toward training AI systems that can function not merely as biological knowledge repositories, but as practical analytical partners in modern protein science.

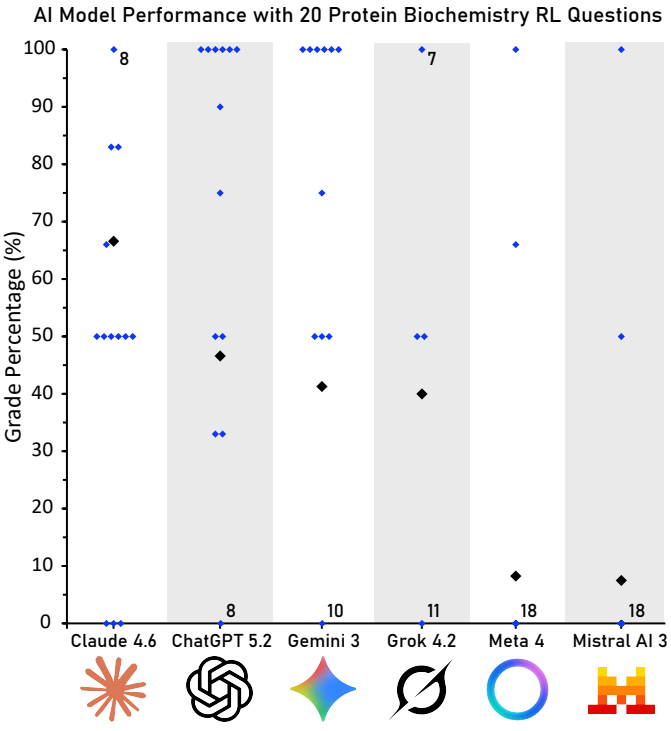

Figure 1: AI model performance with 20 protein biochemistry RL questions. Blue diamonds represent raw scores per question. If >6 raw scores cluster, a number is shown to improve readability of the graph. Overall AI performance score shown as a black diamond.