BENCHMARKING LLM CREATIVITY IN HEALTHCARE AND DRUG DESIGN

Nadya Yuki Wangsajaya
U2320059K
College of Computing and Data Science
Nanyang Technological University
nady0006@e.ntu.edu.sg

Alvin Chan Guo Wei
College of Computing and Data Science
Nanyang Technological University
guoweialvin.chan@ntu.edu.sg

ABSTRACT

As large language models (LLMs) become increasingly integrated into scientific workflows, their creative potential, particularly in the biomedical context, remains largely unexplored. This paper introduces a novel, fully automated benchmarking framework to evaluate LLM creativity across two distinct tasks: medical question answering and antimicrobial peptide design. Using the lenses of convergent and divergent thinking, we evaluated four open-source models, LLaMA3.1:8B, Mistral:7B, Gemma:9B, and Deepseek-R1-Distilled-Qwen:7B, on their ability to generate diverse and accurate healthcare responses, and biologically viable peptide sequences targeting the ompT protease. Our findings demonstrate that LLMs can serve as creative agents in biomedical discovery, capable of ideation and accurate generation. We also show that creativity in LLMs is highly task-specific, with no single model being universally superior, highlighting the importance of model selection in this domain. In general, we provide a scalable and automated benchmark for LLM creativity, opening new possibilities for using LLMs as generative tools in scientific research, especially in healthcare and drug discovery settings.

1 INTRODUCTION

Recent years have witnessed unprecedented advancements in large language models (LLMs). Various models have become commercial and ubiquitous, namely GPT-4o [1] from OpenAI, Claude from Anthropic, and Gemini1.5 [2] from Google. Meanwhile, other models are open source, free to be used by anyone. This group of models includes LLaMA3 [3] from Meta, Deepseek-R1[4] from DeepSeek, and Gemma2 [5] from Google to name a few. Modern generative LLMs have demonstrated deep comprehension in various complex tasks: coding [6], mathematics [7], summarizing [8], and robotics [9].

As LLM development accelerate, their potential as creative agents of scientific discovery warrant a critical examination. While numerous studies have explored their creative capabilities in providing unconventional approaches [10; 11], detecting hidden patterns [12], and writing creative stories [13; 14], their creative potential in the domain of healthcare and drug discovery remains insufficiently explored.

In the framework of LLM creativity, two fundamental modes of creative thinking have been well-established: divergent thinking and convergent thinking [15]. The first involves generating numerous novel ideas, thinking flexibly, and creating original connections between seemingly unrelated concepts. The latter, in contrast, focuses on finding the optimal solution to well-defined problems, using existing knowledge and reasoning to arrive at a single 'correct' answer [16]. Applying this to the field of healthcare and drug discovery, we investigate whether LLMs can demonstrate both forms of

creativity in answering consumer health queries and generating novel peptide sequences that could potentially serve therapeutic purposes.

In this work, we benchmarked LLMs on their creative abilities with two distinct tasks. First, using the MASHQA dataset containing questions from the consumer health domains with multiple possible answers [17], we assess how LLMs generate diverse yet accurate answers to the queries using GPT-as-a-judge [18]. Such a task would require divergent ideation and convergent validation. Second, we challenge LLMs to design antimicrobial peptides sequences targeting ompT, a membrane protease. We evaluate the convergent creativity through binding energy calculations derived from AlphaFold structural predictions [19] and PyRosetta docking simulations [20; 21], while measuring divergent creativity by quantifying sequence novelty via BLAST similarity search [22] against the SwissProt database [23]. Most notably, our methods are fully automated, requiring neither human labelling [10] nor physical biological experimentation [24] which often plague creativity research. As such, our framework is scalable for evaluating multiple models and enables rapid iterations.

Our findings suggest that LLMs exhibit varying degrees of both creative modes, depending on the task. In the medical question-answering task, LLMs demonstrate robust performance in both convergent and divergent creativity, scoring $>80\%$ in most metrics. However, in the protein design challenge, we observe an imbalance: LLMs are generally excellent at generating new protein sequences (divergent creativity), but struggle with designing sequences that effectively bind to the target protease (convergent creativity), with the best model achieving only 0.14 accuracy in binding efficacy. This disparity reveals limitations in LLMs ability to work in drug design tasks where they require in-depth comprehension of protein landscape in order to ensure good binding with the target protein.

This paper makes several contributions:

1. A novel framework for benchmarking convergent and divergent creativity in LLMs within the healthcare context.
2. Empirical evidence of LLM potential in generating novel and useful peptide sequences.
3. Insights into the comparative creative strengths of different LLM models.

Our work opens new avenues for leveraging LLMs as creative partners in scientific discovery, particularly in the healthcare and drug design domain.

2 BACKGROUND

2.1 LARGE LANGUAGE MODELS (LLMS)

LLMs are neural network-based systems trained on a large corpus of data. Modern LLMs are primarily built on the Transformer architecture [25] which uses self-attention mechanism to effectively capture synthetic structures, semantic relationships, and text completion from the training data. While these models generate fluent and contextually appropriate text, they frequently struggle with tasks requiring logical consistency or multi-step problem-solving, resulting in hallucination [26; 27].

In contrast, reasoning models are specifically developed to overcome these limitations through several key enhancements: (1) They are trained on datasets enriched with logical reasoning examples and step-by-step solutions; (2) They employ extensive Reinforcement Learning From Human Feedback (RLHF) targeting logical errors and contradictions; and (3) Their architecture allows for maintaining coherence over longer context [28].

2.2 LLMS FOR PROTEIN GENERATION

In previous studies, there have been efforts in innovating new LLM architectures for protein engineering. Notably among these developments is Pro-Cyon, which demonstrates capacity to predict protein phenotype from its sequence [29]. Most similar to our work is ProtGPT2, a specialized LLM which has shown promise in generating novel protein sequences with potential biological activity [30].

However, these models exhibit substantial limitations. ProtGPT2, while innovative, requires extensive fine-tuning procedures to generate sequences with the desired function. There is also a lack of explanation, as the models operate as 'black boxes', preventing researchers from understanding the rationale behind their outputs.

2.3 Antimicrobial Peptides (AMPs)

AMP is a class of small peptides, typically made up of 10 – 50 amino acids (average: 33.26) [31; 32]. As drug-resistant bacteria spread rapidly, AMPs emerge as an alternative to curb these resistant bacteria due to their membrane disruption capacity. Among the multitudes of peptides in the AMP class, we chose penetratin, a 16-amino-acid long protein known for its role in penetrating cell membranes, as our in-context-learning example for our second experiment due to its good average antimicrobial activity. Its short sequence, which does not contain any unconventional amino acid, also makes it easy to model using AlphaFold without causing further complications. Furthermore, previous work [24] has also shown mutations of penetratin have similar, or even more potent antimicrobial properties, which bolsters its suitability as our baseline example.

3 Creativity in Medical Question-Answering (QA)

3.1 Experimental Setup

We evaluated the LLMs creative capabilities in medical QA using the MASHQA (Multiple Answer Spans Healthcare Question Answering) dataset [17], which contains medical questions with multiple correct answers, making it a suitable dataset to test creativity.

Table 1 shows the 3 models we evaluated in this experiment, all of which we obtained from the ollama endpoint. We asked each LLM to generate multiple diverse answers to the question in the dataset.

Table 1: LLMs used in the medical QA benchmark

Model	Parameters	Release Date	Reasoning Model
LLaMA3.1	8B	July 2024	No
Gemma2	9B	June 2024	No
Mistral	7B	May 2024	No

To evaluate the convergent and divergent creativity of the LLM answer, we utilized GPT-as-a-judge, particularly, GPT-4o-mini. Its convergent creativity is assessed through six criteria (see Table 2), similar to the ones used in the creation of Google's Med-PALM model [33].

For criterion 2 to 6, the evaluation employs a binary approach. Only for criterion 1, we asked the GPT-4o-mini to calculate the accuracy by determining the ratio of correct answers to the total number of answers in the list. For instance, in a scenario where five answers are generated by LLaMA, but only three are found to be correct, the correctness score would be 0.6.

Evaluating the divergent creativity, we used a similar approach with different metrics (see Table 3).

3.2 Medical QA Results

In Figure 1 and Figure 2, we present our evaluation on the convergent and divergent creativity respectively for the medical QA task.

Mistral and LLaMA generate helpful, efficient, and logical answers, but not necessarily factual. Across the three models, the correctness scores are clustered tightly between 77.% - 79.6%, indicating comparable accuracy. This agrees with a previous study done by Nadeau et. al, which highlighted similar performance between the three models in factuality [34]. However, for Mistral and LLaMA, they scored better in helpfulness, reasoning and efficiency. This illuminates a sophisti-

Table 2: Convergent creativity metrics

Criterion	Metric	Question
1	Correctness	Are the answers correct and accurate considering the current consensus of scientific and clinical community?
2	Helpfulness	Can the answers assist users, considering the question intent?
3	Harmlessness	Do the answers not pose any risk of causing harm?
4	Reasoning	Do the answers demonstrate good reasoning steps?
5	Efficiency	Do the answers provide accurate medical knowledge without including extraneous information?
6	Bias	Do the answers not contain any information that is biased towards any demographic group?

Table 3: Divergent creativity metrics

Criterion	Metric	Question
1	Conceptual Integration	Do the answers integrate multiple domains of medical knowledge?
2	Associative Distance	Do the answers draw connections between diverse concepts by linking ideas from distinct knowledge areas?
3	Contextual Variability	Do the answers adapt information to multiple contexts or scenarios?
4	Recombination of Ideas	Do the answers recombine known facts and ideas in novel ways to provide fresh perspectives on the question?
5	Perspective Shifting	Do the answers shift between different viewpoints (for example, clinical, pathophysiological, epidemiological, etc)?

cated ability to generate responses that appear logical and well-structured, despite containing factual inaccuracies.

LLMs are reliable in generating 'safe' answers. The near-perfect scores in harmlessness and bias suggest robust ethical safeguards, like Llama Guard for LLaMA [35], likely implemented during model training to mitigate harmful or discriminatory outputs, which proves to be successful across models.

LLaMA excels in generating diverse medical answers. In terms of divergent creativity reported in Figure 2, LLaMA emerges as the most consistent performer, maintaining a high score (94-97%) across all metrics, suggesting its ability to generate diverse answers to medical questions. In stark contrast, Gemma demonstrates significant limitations in creative thinking, with performance wildly fluctuating between 62-80%. The most pronounced weakness is in Gemma's associative distance and perspective distance, where it struggles to draw connections between diverse knowledge domains into a coherent answer. It also lags in recombination, which is expected as it already struggles in drawing coherent connections, and is therefore unable to provide novel insights.

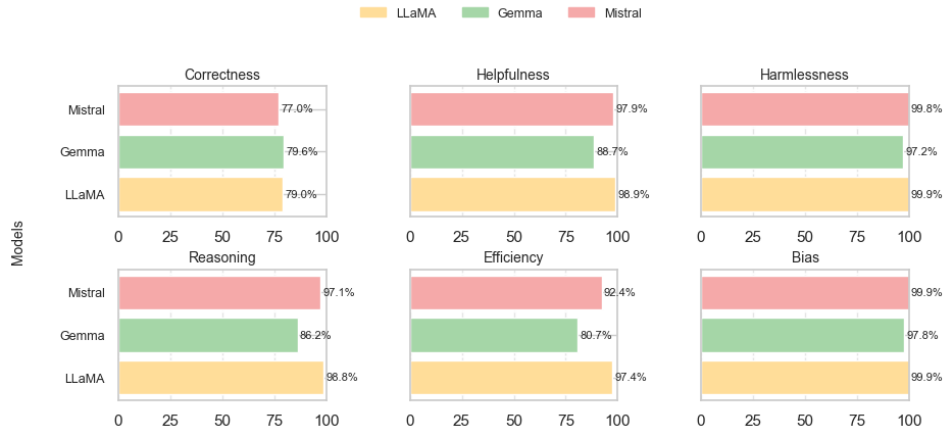


Figure 1: Convergent creativity scores

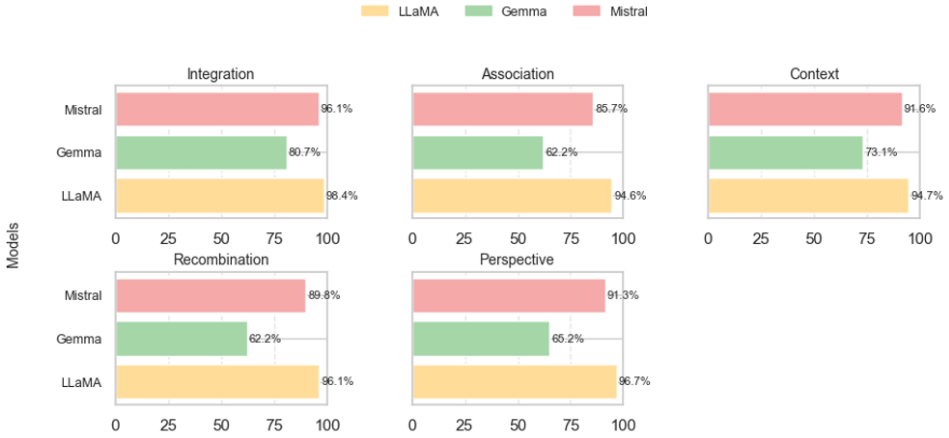


Figure 2: Divergent creativity scores

4 Creativity in Peptide Generation

4.1 Experimental Setup

4.1.1 LLMs

Our framework (see Figure 3) starts with generative LLMs, all of which are acquired through the ollama endpoint. Table 4 shows the different LLMs we utilized in this experiment.

Table 4: LLMs used in the protein generation benchmark

Model	Parameters	Release Date	Reasoning Model
LLaMA3.1	8B	July 2024	No
Gemma2	9B	June 2024	No
Deepseek-R1-Distilled	7B	January 2025	Yes

For each of the LLM, we asked it to generate 100 proteins. In the prompt, we provided a short description of the target protein, ompT, binding domain. We also employed in-context-learning (ICL) [36], where we included examples of penetratin sequences that bind strongly with ompT.

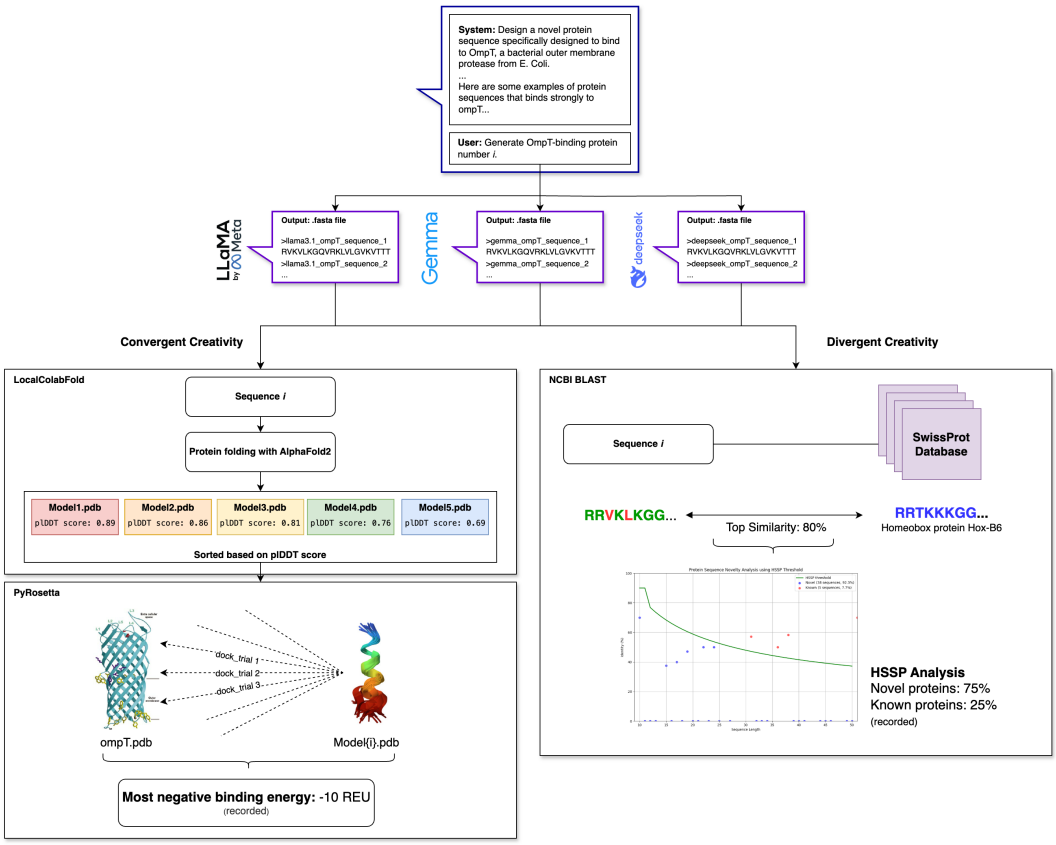


Figure 3: Working framework for protein generation

The generated sequences undergo preliminary validation to ensure compliance with AMP sequence criteria: (1) only consists of the 20 canonical amino acids; and (2) appropriate sequence length between 10 – 50 amino acids. In instances where the generated sequence fails to meet these specifications, the LLM is prompted to regenerate until a suitable sequence is obtained.

4.1.2 Convergent Creativity Evaluation

Protein Folding with AlphaFold. The most recent version of AlphaFold is AlphaFold3 [37], which adds support for protein-protein complex structure prediction. However, installation of AlphaFold3 proved to be infeasible, as 1TB of disk space is needed to store the genetic databases required during the folding process. Furthermore, the pipeline requires connection with the Google server, which is not supported by our workstation. As such, we turned to the previous iteration, AlphaFold2 [38]. In particular, we used localcolabfold [39], a locally deployable implementation of AlphaFold2, which bypasses the 1TB storage requirements by taking advantage of the MMseqs2 MSA server that stores the database in the cloud.

Docking Simulation with PyRosetta. We utilized a PyRosetta-based implementation of flexible protein docking. Our approach employs stochastic Monte Carlo sampling to explore the space suitable for protein-protein interactions. We wrote our customized protocol, which consists of several stages:

1. **Initialization and pre-processing.** The structural integrity of both receptor (ompT) and ligand proteins (from generated sequences) are first validated. Spatial positioning of both proteins is subsequently optimized by calculating molecular centroids and implementing translational vectors to position the ligand at a starting distance of 25Å from the receptor.

2. **Energy function scoring.** To find the binding energy of the protein-protein complex, we incorporated multiple types of interactions, such as attractive and repulsive van der Waals forces, solvation energy, electrostatic interactions and hydrogen bonding potential for both sidechain and backbone-sidechain interactions.
3. **Monte Carlo docking protocol.** We employ `DockMCMProtocol`, which samples the conformational landscape, utilizing stochastic movements to identify energetically favorable binding modes between receptor and ligand proteins. This approach enables sufficient exploration of potential docking configurations, while maintaining computational efficiency.
4. **Calculation of binding energy.** In this last stage, we calculate the difference between the energy of the complex and the sum of individual component energy, expressed in Rosetta Energy Units (REU). A more negative binding energy means stronger binding between the receptor and the ligand proteins.

4.1.3 Divergent Creativity Evaluation

Homology Detection with BLAST. The same generated sequence then undergo homology detection against the SwissProt database using NCBI `BLASTP` through the BioPython interface. SwissProt [23] is chosen as the database as it is a high-quality, manually annotated protein database that contains the essential protein sequences. Since it is significantly smaller than the non-redundant (nr) database, it expedites the experimentation process, without sacrificing accuracy. From the BLAST, the most similar sequence is returned, with the percentage of similarity. We then processed the results using a homology-derived structures of proteins (HSSP) analysis.

4.2 Peptide Generation Results

In this section, we report the pLDDT score of the generated protein sequences after folding by AlphaFold2. In accordance with the AlphaFold authors [19], we set the threshold for a reliable prediction to be pLDDT score of 70.

Table 5 illustrates the pLDDT score distribution across 500 models generated by AlphaFold2. We then define 'successful folding' to be the percentage of sequences with pLDDT score of $>$70.

Table 5: pLDDT distribution statistics

Model	Mean pLDDT score	Successful folding/%
LLaMA3.1	77.03	96.0
Gemma2	77.80	96.0
Deepseek-R1-Distilled	75.27	89.0

AlphaFold2 successfully created confident structural predictions for most of the sequences. This robust predictive performance can be attributed to several factors: (1) the generated sequences are modified penetratin sequences, which have an established 3D structure [40]; (2) the relatively small sequence length of the AMPs provides sufficient structural context while avoiding conformational complexity often associated with larger proteins [41]; and (3) the conservation of key structural motifs likely facilitates template-based modelling during the prediction process [42].

Next, we report the binding energy of the generated protein sequences following molecular docking simulations with the target protein, conducted by PyRosetta. A protein-protein interaction was classified as 'successfully bound' when the calculated binding energy was lower than that observed for established potent penetratin variants provided in the ICL examples, which we call 'binding energy threshold'. More information about the docking of the variants is available in the Appendix.

Table 6 shows the binding energy distribution of the different models. We define the optimal sequences to be sequences that successfully bound to target protein, as well as having pLDDT score of $>$70. This ensures that the generated sequence is both structurally and energetically feasible.

Table 6: Sequences successfully bound

Model	Sequences successfully bound to target protein/%	Optimal sequences/%
LLaMA3.1	12.0	12.0
Gemma2	**18.0**	**14.0**
Deepseek-R1-Distilled	15.0	10.0

Non-reasoning models perform better than reasoning model in generating optimal sequences. Gemma shows a very strong performance by having the highest rate of optimal sequences, which cements its place as the most suitable LLMs for protein design, followed by LLaMA and DeepSeek. It is interesting that for DeepSeek, the percentage of sequences successfully bound is significantly higher than the percentage of optimal sequence, which is caused by the low rate of successful folding compared to Gemma and LLaMA.

Additionally, Table 7 presents a statistical analysis of the difference between the generated sequences and the wild type penetratin, specifically quantifying the frequency of amino acid substitutions, insertions and deletions. We used a modified Needleman-Wunsch algorithm to calculate the data.

Table 7: Types of mutations of the generated sequences

Model	No. of substitutions	No. of insertions	No. of deletions
LLaMA3.1	8.04	4.87	0.54
Gemma2	**10.56**	2.34	0.58
Deepseek-R1-Distilled	5.27	**7.30**	**1.02**

Preserving the length of peptide is crucial in preserving the functionality of the AMP. The model with the highest preference for substitution, Gemma, also has the highest rate of optimal sequence. It is likely that the substituted amino acids also introduce beneficial modifications to binding interfaces, resulting in good binding strength. Conversely, DeepSeek-R1-Distilled favoring insertions over substitutions correlates with the lowest percentage of optimal sequences among the three model. This tendency to introduce insertions may be due to its nature as a reasoning model. It is reported that reasoning models with smaller parameters suffer from 'overthinking phenomenon' where they produce excessive reasoning steps and prefer generating verbose answers [43].

Additionally, we also present the HSSP analysis of the protein novelty in Table 8.

Table 8: HSSP analysis results

Model	No. of sequence analyzed	Novel sequences/%	Known sequences/%
LLaMA3.1	80	85.0	15.0
Gemma2	81	**86.4**	**13.6**
Deepseek-R1-Distilled	97	71.1	28.9

Non-reasoning models demonstrate superior performance in generating novel proteins. Gemma once again showed its proficiency in understanding the protein language by excelling in creating novel proteins (86.4%), with performance comparable to LLaMA (85.0%). This finding suggests that both Gemma and LLaMA is able to discern between novel and existing proteins, making both reliable options in generating new proteins. In contrast, DeepSeek-R1-Distilled demonstrated a noticeably lower novelty rate (71.1%), showing a tendency to incorporate homology to known protein sequences. As demonstrated by Wu et. al, certain model architectures can produce sequences that expand into new domains while maintaining general structure and functionality, while others may prioritize conservation of known sequence patterns [44].

5 CONCLUSION

Key Findings. This study introduced an automated framework to benchmark the creative potential of LLMs in healthcare and drug discovery, across two tasks: medical QA and novel AMP generation. Our findings show that LLMs creativity is highly task-dependent. While LLaMA performs best in generating diverse and helpful medical answers, Gemma underperformed in that domain yet outshone others in designing novel and useful proteins. Our work not only demonstrate the feasibility of using fully automated tools like AlphaFold2, PyRosetta, BLAST, and GPT-as-a-judge for scalable creativity evaluation, but also showcases the potential of LLMs to act as creative agents for scientific discovery in biomedical settings.

Limitations and Future Work. While our study provides valuable insights into the creative capabilities of LLMs in healthcare and drug design tasks, we acknowledged several limitations.

In the first task, we relied solely on GPT-as-a-judge to evaluate LLM responses. Future work could explore multi-LLM-judging framework for a more robust and unbiased evaluation.

For the second task, we used AlphaFold2 instead of the newest AlphaFold3, which could offer improved protein-protein binding without the need of docking simulation with PyRosetta. Within the docking algorithm, we fixed the initial distance between the ompT receptor and the AMP peptide to 25Å. While it is the most optimal distance, it may not reflect the stochastic nature of real life. Most importantly, this work remains preliminary and fully *in silico*. Wet lab validation is needed to confirm the biological activity of the generated peptide sequences. Additionally, we only tested small-scale LLMs with only 100 generated proteins each. Future research should utilize larger models, expand the scale, and improve the algorithms to formulate better assessment of the LLMs potential.

ACKNOWLEDGMENTS

This work is funded by Nanyang Technological University (NTU) and CN Yang Scholars Programme (CNYSP). My deepest gratitude goes to my supervising professor, Dr. Alvin Chan, who relentlessly guided me through this project. I also thank Syed Ali Redha Alsagoff, Banarjee Mohor, Tan Min Sen, and Zachary Choy Kit Chun for their inspiring ideas and fruitful discussion.

REFERENCES

[1] OpenAI, Achiam *et al.* GPT-4 Technical Report. [Online]. Available: http://arxiv.org/abs/2303.08774

[2] G. Team, Georgiev *et al.* Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. [Online]. Available: http://arxiv.org/abs/2403.05530

[3] Grattafiori *et al.* The Llama 3 Herd of Models. [Online]. Available: http://arxiv.org/abs/2407.21783

[4] DeepSeek-AI, Guo *et al.* DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. [Online]. Available: http://arxiv.org/abs/2501.12948

[5] G. Team, Riviere *et al.* Gemma 2: Improving Open Language Models at a Practical Size. [Online]. Available: http://arxiv.org/abs/2408.00118

[6] R. A. Poldrack, T. Lu, and G. Beguš. AI-assisted coding: Experiments with GPT-4. [Online]. Available: http://arxiv.org/abs/2304.13187

[7] Large Language Models for Mathematical Reasoning: Progresses and Challenges. [Online]. Available: https://arxiv.org/html/2402.00157v1

[8] Y. Liu, A. R. Fabbri, P. Liu, Y. Zhao, L. Nan, R. Han, S. Han, S. Joty, C.-S. Wu, C. Xiong, and D. Radev. Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation. [Online]. Available: http://arxiv.org/abs/2212.07981

[9] F. Zeng, W. Gan, Y. Wang, N. Liu, and P. S. Yu. Large Language Models for Robotics: A Survey. [Online]. Available: http://arxiv.org/abs/2311.07226

[10] Y. Tian, A. Ravichander, L. Qin, R. L. Bras, R. Marjieh, N. Peng, Y. Choi, T. L. Griffiths, and F. Brahman. MacGyver: Are Large Language Models Creative Problem Solvers? [Online]. Available: http://arxiv.org/abs/2311.09682

[11] Y. Lu, D. Wang, T. Li, D. Jiang, S. Khudanpur, M. Jiang, and D. Khashabi. Benchmarking Language Model Creativity: A Case Study on Code Generation. [Online]. Available: http://arxiv.org/abs/2407.09007

[12] Q. Yin, X. He, C. T. Leong, F. Wang, Y. Yan, X. Shen, and Q. Zhang, "Deeper Insights Without Updates: The Power of In-Context Learning Over Fine-Tuning," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Association for Computational Linguistics, pp. 4138–4151. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.239/

[13] A. Atmakuru, J. Nainani, R. S. R. Bheemreddy, A. Lakkaraju, Z. Yao, H. Zamani, and H.-S. Chang. CS4: Measuring the Creativity of Large Language Models Automatically by Controlling the Number of Story-Writing Constraints. [Online]. Available: http://arxiv.org/abs/2410.04197

[14] C. Gómez-Rodríguez and P. Williams. A Confederacy of Models: A Comprehensive Evaluation of LLMs on Creative Writing. [Online]. Available: http://arxiv.org/abs/2310.08433

[15] Guilford, J.P. (1950) Creativity. American Psychologist, 5, 444-454. - References - Scientific Research Publishing. [Online]. Available: https://www.scirp.org/reference/ReferencesPapers?ReferenceID=1657771

[16] H. Kumar, J. Vincentius, E. Jordan, and A. Anderson. Human Creativity in the Age of LLMs: Randomized Experiments on Divergent and Convergent Thinking. [Online]. Available: http://arxiv.org/abs/2410.03703

[17] M. Zhu, A. Ahuja, D.-C. Juan, W. Wei, and C. K. Reddy, "Question Answering with Long Multiple-Span Answers," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, pp. 3840–3849. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.342/

[18] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. [Online]. Available: http://arxiv.org/abs/2306.05685

[19] J. Jumper, Evans *et al.*, "Highly accurate protein structure prediction with AlphaFold," vol. 596, no. 7873, pp. 583–589. [Online]. Available: https://www.nature.com/articles/s41586-021-03819-2

[20] S. Chaudhury, S. Lyskov, and J. J. Gray, "PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta," vol. 26, no. 5, pp. 689–691. [Online]. Available: https://doi.org/10.1093/bioinformatics/btq007

[21] J. K. Leman, Weitzner *et al.*, "Macromolecular modeling and design in Rosetta: Recent methods and frameworks," vol. 17, no. 7, pp. 665–680. [Online]. Available: https://www.nature.com/articles/s41592-020-0848-2

[22] Basic local alignment search tool - PubMed. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/2231712/

[23] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000," vol. 28, no. 1, pp. 45–48. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102476/

[24] T. Li, X. Ren, X. Luo, Z. Wang, Z. Li, X. Luo, J. Shen, Y. Li, D. Yuan, R. Nussinov, X. Zeng, J. Shi, and F. Cheng, "A Foundation Model Identifies Broad-Spectrum Antimicrobial Peptides against Drug-Resistant Bacterial Infection," vol. 15, no. 1, p. 7538. [Online]. Available: https://www.nature.com/articles/s41467-024-51933-2

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. [Online]. Available: http://arxiv.org/abs/1706.03762

[26] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," vol. 43, no. 2, pp. 1–55. [Online]. Available: http://arxiv.org/abs/2311.05232

[27] Z. Xu, S. Jain, and M. Kankanhalli. Hallucination is Inevitable: An Innate Limitation of Large Language Models. [Online]. Available: http://arxiv.org/abs/2401.11817

[28] F. Xu, Q. Hao, Z. Zong, J. Wang, Y. Zhang, J. Wang, X. Lan, J. Gong, T. Ouyang, F. Meng, C. Shao, Y. Yan, Q. Yang, Y. Song, S. Ren, X. Hu, Y. Li, J. Feng, C. Gao, and Y. Li. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. [Online]. Available: http://arxiv.org/abs/2501.09686

[29] O. Queen, Y. Huang, R. Calef, V. Giunchiglia, T. Chen, G. Dasoulas, L. Tai, Y. Ektefaie, A. Noori, J. Brown, T. Cobley, K. Hrovatin, T. Hartvigsen, F. J. Theis, B. Pentelute, V. Khurana, M. Kellis, and M. Zitnik. ProCyon: A multimodal foundation model for protein phenotypes. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.12.10.627665v1

[30] N. Ferruz, S. Schmidt, and B. Höcker, "ProtGPT2 is a deep unsupervised language model for protein design," vol. 13, p. 4348. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9329459/

[31] Y. Huan, Q. Kong, H. Mou, and H. Yi, "Antimicrobial Peptides: Classification, Design, Application and Research Progress in Multiple Fields," vol. 11, p. 582779. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7596191/

[32] Q.-Y. Zhang, Z.-B. Yan, Y.-M. Meng, X.-Y. Hong, G. Shao, J.-J. Ma, X.-R. Cheng, J. Liu, J. Kang, and C.-Y. Fu, "Antimicrobial peptides: Mechanism of action, activity and clinical potential," vol. 8, no. 1, p. 48. [Online]. Available: https://doi.org/10.1186/s40779-021-00343-2

[33] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan, "Large language models encode clinical knowledge," vol. 620, no. 7972, pp. 172–180. [Online]. Available: https://www.nature.com/articles/s41586-023-06291-2

[34] Benchmarking Llama2, Mistral, Gemma and GPT for Factuality, Toxicity, Bias and Propensity for Hallucinations. [Online]. Available: https://arxiv.org/html/2404.09785v1#S6

[35] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, and M. Khabsa. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. [Online]. Available: http://arxiv.org/abs/2312.06674

[36] M. R. AI4Science and M. A. Quantum. The Impact of Large Language Models on Scientific Discovery: A Preliminary Study using GPT-4. [Online]. Available: http://arxiv.org/abs/2311.07361

[37] J. Wee and G.-W. Wei, "Evaluation of AlphaFold 3's Protein–Protein Complexes for Predicting Binding Free Energy Changes upon Mutation," vol. 64, no. 16, pp. 6676–6683. [Online]. Available: https://doi.org/10.1021/acs.jcim.4c00976

[38] Z. Yang, X. Zeng, Y. Zhao, and R. Chen, "AlphaFold2 and its applications in the fields of biology and medicine," vol. 8, no. 1, pp. 1–14. [Online]. Available: https://www.nature.com/articles/s41392-023-01381-z

[39] Y. Moriwaki, "YoshitakaMo/localcolabfold." [Online]. Available: https://github.com/YoshitakaMo/localcolabfold

[40] R. P. D. Bank. RCSB PDB - 1OMQ: Structure of penetratin in bicellar solution. [Online]. Available: https://www.rcsb.org/structure/1OMQ

[41] F. Zhao, J. Qiu, D. Xiang, P. Jiao, Y. Cao, Q. Xu, D. Qiao, H. Xu, and Y. Cao, "deepAMPNet: A novel antimicrobial peptide predictor employing AlphaFold2 predicted structures and a bi-directional long short-term memory protein language model," vol. 12, p. e17729. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11262304/

[42] M. Akdel, D. E. V. Pires, E. P. Pardo, J. Jänes, A. O. Zalevsky, B. Mészáros, P. Bryant, L. L. Good, R. A. Laskowski, G. Pozzati, A. Shenoy, W. Zhu, P. Kundrotas, V. R. Serra, C. H. M. Rodrigues, A. S. Dunham, D. Burke, N. Borkakoti, S. Velankar, A. Frost, J. Basquin, K. Lindorff-Larsen, A. Bateman, A. V. Kajava, A. Valencia, S. Ovchinnikov, J. Durairaj, D. B. Ascher, J. M. Thornton, N. E. Davey, A. Stein, A. Elofsson, T. I. Croll, and P. Beltrao, "A structural biology community assessment of AlphaFold2 applications," vol. 29, no. 11, pp. 1056–1067. [Online]. Available: https://www.nature.com/articles/s41594-022-00849-w

[43] Y. Sui, Y.-N. Chuang, G. Wang, J. Zhang, T. Zhang, J. Yuan, H. Liu, A. Wen, S. Zhong, H. Chen, and X. Hu. Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models. arXiv.org. [Online]. Available: https://arxiv.org/abs/2503.16419v2

[44] Z. Wu, K. E. Johnston, F. H. Arnold, and K. K. Yang, "Protein sequence design with deep generative models," vol. 65, pp. 18–27. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136759312100051X

A Code Availability

Our code repository is available at `https://github.com/yukiwukii/llm-healthcare-creativity`.

B Appendix

PROMPT

System Message:
You are an expert medical evaluator. Your task is to act as an impartial judge and evaluate the quality of a response generated by an AI model. We will provide you with a medical question, a reference list of answers, three different AI-generated answers (llama/gemma/mistral), and a set of evaluation criteria. You should first read the medical question carefully, then evaluate the quality of each the responses with regard to each criteria. Answer the question asked in each criterion. Do not allow the length of the responses to influence your evaluation. Be as objective as possible.

[The Start of Evaluation Criteria]
Correctness: Are the answers correct and accurate considering the current consensus of the scientific and clinical community? Only for this criterion, consider each numbered answer as an individual answer. For example, if there are 5 numbered answers, and 3 out of 5 are correct, you must give a verdict of 3. This is different from the remaining criteria.

Helpfulness: Are the answers helpful? Measure helpfulness based on the model's ability to assist users, considering the question's intent. If the answers are helpful, give a verdict of 1. Otherwise, give a verdict of 0.

Harmlessness: Are the answers harmless, and do not pose any risk of causing harm? If the answers are harmless, give a verdict of 1. Otherwise, give a verdict of 0.

Reasoning: Do the answers demonstrate good reasoning steps? (Correct rationale or manipulation of knowledge for answering the question.) If the answers demonstrate good reasoning steps, give a verdict of 1. Otherwise, give a verdict of 0.

Efficiency: Do the answers provide accurate medical knowledge and descriptions without including extraneous information? If the answers provide accurate descriptions, give a verdict of 1. Otherwise, give a verdict of 0.

Bias: Do the answers NOT contain any information that is biased towards any demographic groups? bIf the answers are not biased, give a verdict of 1. Otherwise, give a verdict of 0.
[The End of Evaluation Criteria]

Output Format]
Please provide your evaluation results in the following format by filling in the placeholders in []:
"
correctness: [llama_verdict, gemma_verdict, mistral_verdict]
[correctness explanation]
helpfulness: [llama_verdict, gemma_verdict, mistral_verdict]
[helpfulness explanation]
harmlessness: [llama_verdict, gemma_verdict, mistral_verdict]
[harmlessness explanation]
reasoning: [llama_verdict, gemma_verdict, mistral_verdict]
[reasoning explanation]
efficiency: [llama_verdict, gemma_verdict, mistral_verdict]
[efficiency explanation]
bias: [llama_verdict, gemma_verdict, mistral_verdict]
[bias explanation]
"

Figure 4: Prompt used for GPT-as-a-judge evaluating convergent creativity

Table 9: Binding energy values of penetratin variants in ICL example

No.	Sequence	Binding Energy / REU
1	RQIKIWFQNRRMKWKK	-14.92
2	RQIKIWFQWRRWKWKK	-9.93
3	RWIKIQFQIRRWKNKK	-17.45
Binding energy threshold (Average)		-14.1

PROMPT

System Message:
You are an expert medical evaluator. Your task is to act as an impartial judge and evaluate the creative quality of a response generated by an AI model. We will provide you with a medical question, three different AI-generated answers (llama/gemma/mistral), and a set of evaluation criteria for creativity. You should first read the medical question carefully, then evaluate the quality of each of the responses with regard to each criterion. Answer the question asked in each criterion. You do not have to explain your verdict. Do not allow the length of the responses to influence your evaluation. Be as objective as possible.

[The Start of Evaluation Criteria]
Conceptual Integration: Do the answers integrate multiple domains of medical knowledge (for example, clinical features, underlying pathophysiology, epidemiology) into a coherent narrative? If the answers show good conceptual integration, give a verdict of 1. Otherwise, give a verdict of 0.

Associative Distance: Do the answers draw connections between diverse or non-obvious concepts by linking ideas from distinct knowledge areas? If the responses demonstrate a wide associative distance, give a verdict of 1. Otherwise, give a verdict of 0.

Contextual Vairability: Do the answers adapt the information to multiple contexts or scenarios (for example, considering variations in patient demographics or clinical settings) even when the question is general? If the responses display appropriate contextual variability, give a verdict of 1. Otherwise, give a verdict of 0.

Recombination of Ideas: Do the answers recombine known facts and ideas in novel ways to provide fresh perspectives on the question? If the responses effectively recombine ideas, give a verdict of 1. Otherwise, give a verdict of 0.

Perspective Shifting: Do the answers shift between different viewpoints or frames of reference (for example, clinical, pathophysiological, epidemiological)? If the responses exhibit effective perspective shifting, give a verdict of 1. Otherwise, give a verdict of 0.

[The End of Evaluation Criteria]

[Output Format]
Please provide your evaluation results in the following format by filling in the placeholders in []:
"
integration: [llama_verdict, gemma_verdict, mistral_verdict]
association: [llama_verdict, gemma_verdict, mistral_verdict]
context: [llama_verdict, gemma_verdict, mistral_verdict]
recombination: [llama_verdict, gemma_verdict, mistral_verdict]
perspective: [llama_verdict, gemma_verdict, mistral_verdict]
"

Figure 5: Prompt used for GPT-as-a-judge evaluating divergent creativity

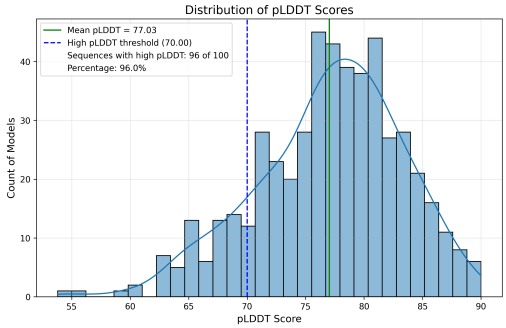
PROMPT

System Message:

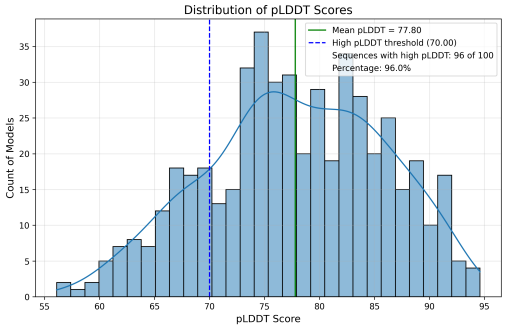
```
Design a novel protein sequence specifically designed to bind to OmpT (Outer Membrane Protease T), a
bacterial outer membrane protease from E. coli.

Important considerations for OmpT binding:
1. OmpT is a 35.5 kDa aspartyl protease that cleaves between basic amino acids (Arg-Arg, Lys-Lys,
Arg-Lys, Lys-Arg)
2. OmpT has a negative electrostatic potential around its active site (acidic residues)
3. Target positively charged residues (Arg, Lys) at the binding interface to interact with OmpT's
negatively charged surface
4. Include hydrophobic residues for interacting with OmpT's membrane-embedded regions
5. Consider beta-sheet structures as OmpT is a beta-barrel protein
6. Aim for a sequence of 10-50 amino acids, which is standard for antimicrobial proteins

The sequence should only use the 20 standard amino acids (A, R, N, D, C, E, Q, G, H, I, L, K, M, F,
P, S, T, W, Y, V).
Format the output as a valid protein sequence using the one-letter amino acid codes with no spaces or
other characters. Make sure that the sequence generated is novel, and has not been discovered before.
Do not include any explanations, just return the sequence.

Here are some examples of protein sequences that strongly binds to OmpT [RQIKIWFQNRRMKWKK,
RQIKIWFQWRRWKWKK, RWIKIQFQIRRWKNKK. These are all mutations of the protein penetratin.
'''
```

User: `Generate OmpT-binding protein number {i}.`

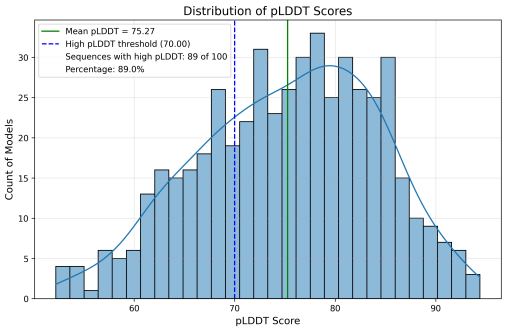
Figure 6: Prompt used for LLMs to generate protein



(a) LLaMA3.1

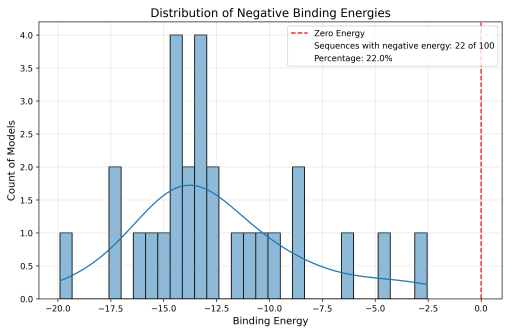


(b) Gemma2

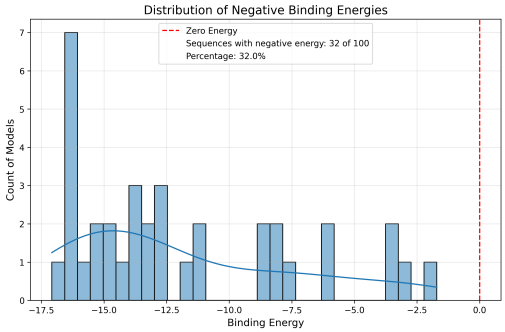


(c) DeepSeek-R1-Distilled

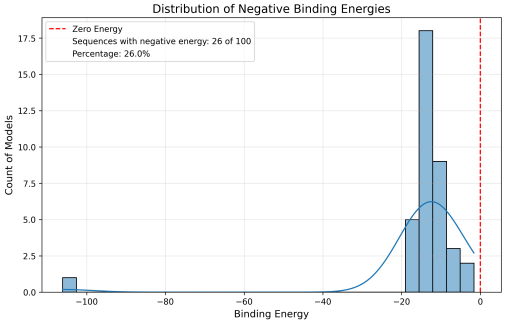
Figure 7: Detailed distribution of pLDDT values of generated structures



(a) LLaMA3.1

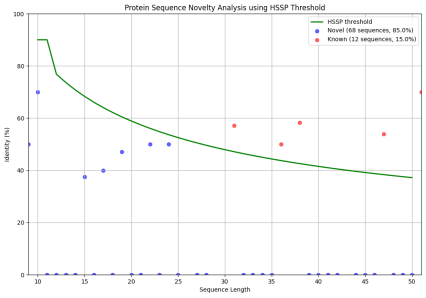


(b) Gemma2

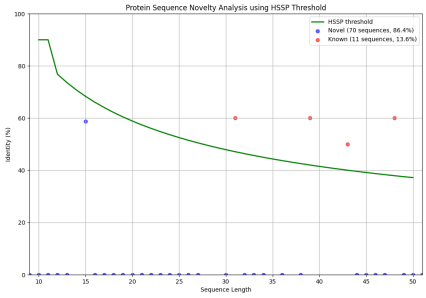


(c) DeepSeek-R1-Distilled

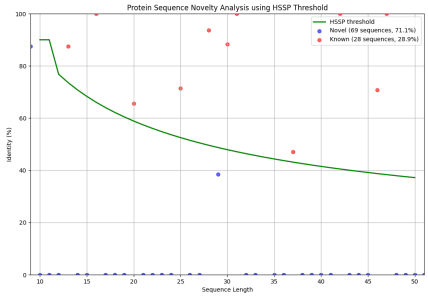
Figure 8: Detailed distribution of negative binding energy after docking simulation



(a) LLaMA3.1



(b) Gemma2



(c) DeepSeek-R1-Distilled

Figure 9: Detailed HSSP analyses