Cost-Effective Biological Data Analysis via a Benchmark and Ensemble of Large Language Models

Vinayak Gupta[1] (gupta20@llnl.gov), Brian Bartoldson[1] (bartoldson1@llnl.gov), Joseph Wakim[1] (wakim1@llnl.gov), Jonathan Allen[1] (allen99@llnl.gov), Jose Manuel Marti[1] (martimartine1@llnl.gov), Tianlong Chen[2] (tianlong@unc.edu), Bhavya Kailkhura[1] (kailkhura1@llnl.gov)
[1]Lawrence Livermore National Laboratory, [2]University of North Carolina at Chapel Hill

TOPIC: Foundation Models, Multiscale/Multimodal Modeling, and Automated Labs.

CHALLENGES: AI is profoundly transforming biology research, enabling breakthroughs that were once out of reach, such as solving longstanding challenges like protein folding, which has paved the way for designing synthetic proteins, de novo vaccines, and precision cancer therapies [1, 2]. Initiatives like the Cell Maps for AI program are using language models to create detailed spatiotemporal maps that link genotype to phenotype, advancing our understanding of cellular functions [3]. In drug discovery, AI-driven models streamline workflows by identifying causal targets and predicting drug efficacy with human-level accuracy, enhancing precision medicine [4]. Large language models (LLMs) have become increasingly popular for harnessing deep knowledge of biomolecules and powering various bio-data analysis tasks, including understanding protein folding, analyzing molecular structures, and modeling protein-gene interactions. However, the performance of LLMs is generally proportional to their size, necessitating more computational resources for improved training and inference. These models can encompass billions of variables; for instance, ESM-3 [5] is equipped with a remarkable 98 billion parameters, and OpenAI's GPT models with more than a trillion parameters. This situation is further exacerbated by increasing competition in the industry, driving companies to spend billions on acquiring more computational power, which inflates the costs associated with computing chips. Given these trends, there will be exponential growth in the demand for resources for keeping the biological scientific community at the forefront of the AI revolution. It will also have a significant environmental impact, as the resources required for training and developing hardware for these models are substantial, resulting in a considerable carbon footprint, which will impact the Department of Energy's commitment to a zero-carbon future.

OPPORTUNITY: Integrating AI models into the bioscience community is crucial for advancing research in areas such as virus analysis, disease understanding, and protein-gene interactions. This importance is highlighted by the 2024 Nobel Prize in Chemistry, awarded to the creators of an AI model that predicts protein structures with performance surpassing human capabilities. Notably, the model, AlphaFold [1], determines the three-dimensional configurations of proteins from their amino acid sequences, models protein-protein interactions, analyzes how proteins engage with one another, and evaluates protein stability under various conditions. Despite these advancements, the rapid evolution of AI technology has introduced new contenders in the AI-bio race, including LLMs that can outperform AlphaFold and tackle innovative tasks such as generating new fluorescent proteins and even identifying previously unrecognized viruses [5,6]. The extensive scope of biological research, characterized by its diverse objectives, complicates the task of effectively applying models and prioritizing them. It is particularly challenging to identify models that not only meet research needs but also have the least carbon footprint. This consideration is crucial, as the environmental impact of these models can be significant, contributing to an increased carbon footprint that undermines sustainability efforts. Here, we highlight several open opportunities within the biological community. O1: *Given a specific bio-analysis task, which LLM is best suited for that task?* O2: *How can we determine which model, large or small, is most effective for a given task that doesn't always require a larger model?* And O3: *In scenarios with limited resources, can we leverage a combination or ensemble of smaller LLMs to approximate the performance of a larger model?* Addressing these challenges will not only benefit researchers but also sponsors of scientific endeavors, enabling more efficient allocation of resources. To address the above-defined questions within the biological research community, we discuss a structured, step-by-step methodology to identify pain points, survey existing literature, design effective solutions, and gain deeper insights. Specifically, in response to O1 on identifying the most suitable LLMs for a given biological analysis task, we recommend a comprehensive benchmarking study of current bio-function models. This study should involve evaluating both publicly available models, such as the protein-folding LLM Chai-1[6], and privately held models like ESM-3[5] on a wide-range of tasks. The benchmarking process should evaluate each model's performance using energy-sensitive metrics often overlooked in the literature, such as computational requirements and operational costs [7]. Furthermore, it should include ranking these models on standardized datasets, such as protein banks datasets, while considering factors like input size and processing capabilities. Such a benchmarking initiative would provide researchers with a valuable tool for selecting the most appropriate LLM based on specific needs and privacy requirements. This would not only save significant time and resources for researchers but also help the bioinformatics community with clear guidelines and strategic insights.

In response to O2 on prioritizing smaller language models for simpler tasks, the challenge arises from the observation that while larger models generally outperform smaller models in highly specialized tasks that require deep understanding of data, the performance difference for many simpler tasks can be minimal. For a wide range of straightforward tasks, the performance benefits of using a larger models may not be significant [6, 8]. This leads to the next step in the benchmarking process for O1. Specifically, for tasks related to biological data analysis, it is intuitive that not every piece of data requires querying a larger model. Many analyses could be conducted with good accuracy using a smaller, more efficient model. This abstraction layer -- determining when to shift from larger models to smaller models -- would save valuable resources, computing time, and, importantly, reduce the carbon footprint of research. This task requires a working framework that: (i) can judge and identify the most suitable LLM for each task; (ii) load the appropriate smaller model (public) or private API onto the machine; and (iii) convert input data into the correct formats before querying the LLM for results. Although this would require an initial engineering effort, the long-term savings in compute and power consumption would be considerable.

In response to O3 on designing an ensemble of smaller language models, we recommend developing an algorithm to pretrain a collection of smaller models and implement a polling methodology that could outperform a larger language model. This would build on O2, where resource optimization through multiple smaller models is the focus, requiring design solutions across all stages of the AI pipeline: training, testing, and deployment. During the training phase, a set of smaller models can be trained synchronously to mimic the capabilities of a larger model. Recent research has shown that integrating smaller models with the precision derived from a larger model during training can substantially improve accuracy and noise robustness, outperforming a single model [5]. Similarly, during testing, the ensemble can be compared with the predictions from the larger model to identify the most effective polling techniques. We believe that, especially in resource-constrained environments, using an ensemble of smaller models offers distinct advantages. Modifying a smaller model requires fewer resources and can be more easily updated, making it a more flexible and sustainable option. Such an ensemble is aimed to effectively provide AI resources to research groups with limited resources, ensuring more equitable access to AI technologies. This research could also extend into applications in hospitals with limited resources.

TIMELINESS: The push for using bigger AI models in biological data analysis is growing, as its potential to transform fields like protein folding, drug discovery, and disease understanding becomes increasingly apparent [9]. While there are challenges in selecting the most appropriate models for specific tasks, the opportunities for growth significantly outweigh these obstacles. AI has already been instrumental in achieving major breakthroughs, and several areas show great promise for further advancement. These include identifying the best LLMs for bio-analysis tasks, utilizing smaller models for simpler tasks, and developing ensembles of smaller models to optimize resources and improve performance. Addressing these opportunities will enhance research efficiency, reduce computational costs, and minimize the environmental impact of biological studies. Additionally, these innovations will make AI more accessible to resource-constrained research groups, ensuring more equitable opportunities for growth in the field.

REFERENCES: **[1]** Jumper. et al. "Highly accurate protein structure prediction with AlphaFold." Nature, 2021. **[2]** Prabhune. "Infusion of Artificial Intelligence in Biology". The Scientist 2024. **[3]** https://cm4ai.org/ **[4]** Zhao et al. "Cancer Mutations Converge on a Collection of Protein Assemblies to Predict Resistance to Replication Stress". Cancer Discovery. **[5]** Hayes et al. "Simulating 500 million years of evolution with a language model". 2024 **[6]** "Chai-1: Decoding the molecular interactions of life". 2024. **[7]** www.biomap.com/sota/ **[8]** Bartoldson et al. Compute-Efficient Deep Learning: Algorithmic Trends and Opportunities. JMLR 2024. **[9]** DOE Advancements in Artificial Intelligence for Science Program. colorado.edu/researchinnovation/doe-advancements-artificial-intelligence-science.