Leveraging Foundational Models in Computational Biology: Validation, Understanding, and Innovation*

Brett Beaulieu-Jones
Department of Medicine, University of Chicago, 5841 South Maryland Avenue, MC 6092 Chicago, IL, USA
Email: beaulieujones@uchicago.edu

Steven Brenner
Department of Plant and Microbial Biology, 111 Koshland Hall, Berkeley, CA, USA
Email: brenner@compbio.berkeley.edu

Large Language Models (LLMs) have shown significant promise across a wide array of fields, including biomedical research, but face notable limitations in their current applications. While they offer a new paradigm for data analysis and hypothesis generation, their efficacy in computational biology trails other applications such as natural language processing. This workshop addresses the state of the art in LLMs, discussing their challenges and the potential for future development tailored to computational biology. Key issues include difficulties in validating LLM outputs, proprietary model limitations, and the need for expertise in critical evaluation of model failure modes.

Keywords: Generative AI, Large Language Models, Foundational Models, Computational Biology

1. Background

Large Language Models (LLMs) have demonstrated immense potential[1–9] within and outside of the biomedical domain but currently have substantial limitations when applied to biomedical research.[10,11] These models promise a new paradigm for data analysis, interpretation and hypothesis generation, but it is not clear how fully this promise will be fulfilled. LLMs are just one class of foundational models, and while they have already made a significant impact to computational biology, it is unlikely that a singular architecture geared at processing natural language will be the ideal framework for general learning in computational biology. This workshop aims to provide an understanding of the state of the art today, current challenges in the application or development of models tailored to computational biology, as well as to start a discussion of what the future holds for our community.

At present, LLMs are commonly used in attempt to directly answer complex problems in ways that are difficult to validate. Existing methods for interpretation are limited, and it is difficult without a ground truth to tell whether an answer is accurate or a "hallucination".[12] These challenges contrast

* This work is partially supported by NIH grant R00NS114850

with typical goals in biomedical research where researchers aim to understand the underlying system. Issues with LLM hallucination have been well documented and approaches for dealing with uncertainty within generative models are nascent. Proprietary models create challenges to reproducibility, privacy, and present barriers to finetuning and open sharing. The successful use of LLMs for research still requires a high degree of expertise in order to "red team", or critically interrogate and evaluate failure modes of LLMs. This process is currently poorly defined with best practices not yet widely agreed upon.

Most prior work has focused on either training LLMs or using available models (locally or via vendor provided APIs) for related tasks. A critical issue with the status quo is that the field is rapidly evolving, meaning building upon any one model is a risk and there is a constant need to retrain models and update workflows based on newly released models. Additionally, the majority of innovation has come either through using large general-purpose models (e.g., GPT4), or in training models derived from architectures designed for natural language processing. Increasingly we are seeing the development of foundational models for multimodal data in addition to more specific subfields. As a new state of the art model is released, within a relatively short period of time, researchers have developed smaller, domain or task specific models that appear to achieve comparable or slightly worse performance despite having access to vastly fewer resources. Recently, we have seen the emergence of novel architectures for foundational models trained on electronic medical record data[13,14] and multimodal models for medical-imaging and text.[15–19] While these models have demonstrated early promise, their impact does not yet compare to that of LLMs.

Topics around foundational models, specifically LLMs, have been widely covered at academic journals, conferences, and in a wide variety of other settings. However, the majority of discussions around these models have focused on the low hanging fruit, posing questions like how GPT-4 can be used as a knowledge integration tool for hypothesis generation or evaluating its capabilities against professional exams or clinical case diagnostics. There has been decidedly less attention paid to the methodological side of tailoring these models to workflows in computational biology through techniques like the programmatic generation of prompts and labels for supervised and even weakly supervised instruction fine-tuning, interpretation and/or explanation leveraging expert knowledge-based uncertainty exploration, retrieval-augmented generation strategies with "-omics" style data, multimodal approaches to include assets like clinical notes and medical imaging for phenotyping. Finally, with the rapid advancement of the larger field of foundational models, it is nearly impossible for the transdisciplinary scientists who typically attend PSB to keep up with all of the literature in a critical but separate field from their primary research.

2. Leveraging Foundational Models in Computational Biology: Workshop

LLM's and the broader field of generative AI are in period of rapid evolution. This workshop aims to help attendees of PSB differentiate between the signal and the noise. What are the breakthrough ideas, technologies, and applications that are already or are poised to have substantial impacts on the field of computational biology.

This workshop aims to provide:

1. Provide an understanding of the current state of the art for foundational models both in general and specifically within computational biology
2. Understand common failure modes and survey methods to validate results
3. Explore recent innovations in foundational models and LLMs that address prior challenges most relevant to computational biology (e.g., novel approaches for tokenization, representation of modalities outside of natural language, uncertainty estimation and explanation)
4. Showcase innovative uses of LLMs in computational biology through a "year-in-review" overview of the past years most interesting works in this area
5. Plan for the future based on invited talks by researchers on the strategies for development and utilization of the next generation of LLMs.

To do this, the workshop will be composed of three invited talks covering, “What is the current state of the art?”, “What are the Strategies for recognizing and Mitigating Failure Modes”, and a “Year-in-Review” talk based on extensive literature review. Our aim with this is to help the PSB audience determine what is worth paying attention to and which developments are simply “shining objects” that are potential distractions. Additionally, there will be a panel discussion covering the challenges and shortcomings of current approaches and what does the future look like?

3. Conclusion

LLMs hold immense potential for transforming biomedical research, but their current limitations, such as hallucinations and challenges in reproducibility, necessitate careful scrutiny. The field is evolving rapidly, with new foundational models being introduced frequently, requiring constant retraining and workflow updates. It is essential to develop methodologies specifically suited to computational biology, as general-purpose models may not be optimal for this domain. The workshop seeks to guide researchers in discerning between valuable advancements and distractions in this rapidly changing environment.

4. Acknowledgements

We would like to thank the Organizing committee of the Pacific Symposium for Bioinformatics 2024 for giving us the opportunity of organizing the proposed workshop.

References

1. Bubeck, S. *et al.* Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv [cs.CL]* (2023).
2. Singhal, K. *et al.* Large Language Models Encode Clinical Knowledge. *arXiv [cs.CL]* (2022).
3. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).

4. Eriksen Alexander V., Möller Sören & Ryg Jesper. Use of GPT-4 to Diagnose Complex Clinical Cases. *NEJM AI* **1**, AIp2300031 (2023).
5. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv [cs.CL]* (2023).
6. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
7. Varadi, M. *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).
8. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
9. Ingraham, J. B. *et al.* Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).
10. Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* **6**, 120 (2023).
11. Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. & Daneshjou, R. Large language models propagate race-based medicine. *NPJ Digit Med* **6**, 195 (2023).
12. Ji, Z. *et al.* Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **55**, 1–38 (2023).
13. Wornow, M. *et al.* The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit. Med.* **6**, 135 (2023).
14. Thapa, R., Steinberg, E. & Fries, J. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Adv. Neural Inf. Process. Syst.* (2024).
15. Azad, B. *et al.* Foundational Models in Medical Imaging: A Comprehensive Survey and Future Vision. *arXiv [cs.CV]* (2023).
16. Moor, M. *et al.* Med-Flamingo: a Multimodal Medical Few-shot Learner. in *Proceedings of the 3rd Machine Learning for Health Symposium* (eds. Hegselmann, S. et al.) vol. 225 353–367 (PMLR, 2023).
17. Jeong, J. *et al.* Multimodal Image-Text Matching Improves Retrieval-based Chest X-Ray Report Generation. *arXiv [cs.CL]* (2023).
18. Willemink, M. J., Roth, H. R. & Sandfort, V. Toward Foundational Deep Learning Models for Medical Imaging in the New Era of Transformer Networks. *Radiol Artif Intell* **4**, e210284 (2022).
19. Chambon, P., Bluethgen, C., Langlotz, C. P. & Chaudhari, A. Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains. *arXiv [cs.CV]* (2022).