[Dataset Proposal]

Closing the Omics Gap: A Benchmark for Unified Evaluation of Biomolecular Foundation Models

Joseph G. Wakim[*], **Vinayak Gupta**[*], **Jose Manuel Marti**, **Jonathan E. Allen**,
Brian Bartoldson, **Bhavya Kailkhura**[†]

Lawrence Livermore National Laboratory, Livermore, CA, USA
[*]Equal contribution; [†]Corresponding author: `kailkhura1@llnl.gov`

Abstract

In recent years, biomolecular foundation models (bioFMs) have been trained on massive amounts of omics data to encode complex patterns in biological sequences. These models have shown remarkable predictive performance across a broad range of applications in biotechnology, as well as the ability to generate novel, viable sequences. However, the vast majority of existing bioFMs are unimodal, trained exclusively on nucleotide or amino acid sequences, and are evaluated on tasks specific to their sequence type. With few benchmarks incorporating multi-omics data, opportunities for cross-modal evaluations are limited. To address this gap, we propose a novel cross-modal benchmark that links nucleotide and amino acid sequences to common biological outcomes. Given a pair of genes, our benchmark will pose questions such as: *Do the encoded proteins co-localize or share similar functions? Are the genes associated with a common disease or linked to common drug targets?* By providing a common platform for evaluation, our benchmark will support comparisons of unimodal and multimodal bioFMs, offering a foundation for tracking their capabilities and informing appropriate safety oversight.

Although DNA is composed of just four types of nucleotides, this macromolecule forms a "biological language" that describes all life. The expressiveness of this language arises from the conversion of DNA into proteins; triplets of nucleotides on DNA encode amino acids, which fold into proteins that confer phenotypes and enable diverse biological functions. Biomolecular foundation models (bioFMs), adopting architectures from natural language processing, have been trained on vast amounts of omics data to learn the complex patterns in nucleotide and amino acid sequences [1, 2, 3, 4, 5]. By deciphering the biological language, these models offer powerful tools for applications in personalized medicine [6, 7, 8], drug discovery [9, 10], and protein engineering [11, 12].

The majority of existing bioFMs are unimodal, with genome foundation models (gFMs) trained exclusively on nucleotide sequences and protein foundation models (pFMs) trained on amino acid sequences. Despite the direct link between DNA and proteins, gFMs and pFMs excel in different areas: pFMs better capture protein-level functional attributes, while gFMs capture additional regulatory information from non-coding nucleotides [13]. However, as gFMs continue to improve, the performance gap in function prediction may narrow. Recent evidence also suggests that integrating nucleotide and amino acid sequences during training can improve biological understanding captured by bioFMs [2, 14]. To promote the cross-modal evaluation of bioFMs, we propose a novel dataset aligning nucleotide sequences, amino acid sequences, and phenotypic labels across several species. Such a dataset would not only support model benchmarking but also help monitor emerging capabilities and related safety considerations in bioFMs.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science: The Reach and Limits of AI for Scientific Discovery.

Proposed Dataset for Cross-Modal Evaluation of BioFMs

We propose a benchmark of matched genes and proteins that are labeled with phenotypic information to support cross-modal evaluation of gFMs and pFMs. Our dataset will include nucleotide sequences spanning coding segments and intervening introns, as well as the encoded amino acid sequences. We will collect sequence information from public sources, such as RefSeq [15] and UniProt [16]. Corresponding sequences will be assigned an identifier, forming gene-protein records. For each record, we will collect information about the structure, localization, functions, interactions, and biomedical implications of the encoded protein. Structural attributes will be derived from the Protein Data Bank [17], while localization details will be obtained from UniProt [16]. Functional and pathway-related information will be drawn from the Human Protein Atlas [18], Reactome [19], and the Gene Ontology [20]. Protein-protein interactions will be collected from STRING [21], while protein-ligand interactions will be obtained from ChEMBL [22], PubChem [23], and DrugBank [24]. Biomedical implications will be pulled from The Cancer Genome Atlas [25] and DepMap [26].

We will form prediction tasks that each ask a binary question about individual records or pairs of records. For example, given an individual record, we will ask: *Are melanoma tumor cells dependent on the associated gene?* Meanwhile, given two records, we will ask: *Do the encoded proteins share a subcellular location? Do they have similar functions? Do they interact? Are they known to bind a common ligand?* Our prediction tasks will tend to be protein centric and will not probe a model's understanding of non-coding or regulatory features of the genome. Since pFMs are specialized for protein-centric prediction tasks and our nucleotide sequences will include introns, pFMs may hold an advantage on our benchmark and can serve as a reference point for future gFM development. As gFMs improve in their ability to distinguish coding regions from introns and infer protein properties, the performance gap may narrow. To evaluate bioFMs, we will train separate prediction heads for each model and prediction task, then quantify performance on the held-out examples.

Our benchmark will test whether leading gFMs and pFMs capture biological information beyond what can be obtained from sequence alignment alone. For prediction tasks involving pairs of gene-protein records, we will compare the sequence identity of the positive and negative examples and will sample the two classes to ensure overlapping sequence-identity distributions. We will also identify a subset of "challenge cases" where records share attributes but have dissimilar sequences, or conversely, have similar sequences but differ in their attributes.

Preliminary Results

We present GENO-PROT as a demonstration of our proposed dataset [13]. GENO-PROT includes nine prediction tasks like those described above, which reference human nucleotide and amino acid sequences. We evaluate leading bioFMs ranging from 6.6 million to seven billion parameters in size [1, 2, 3, 5, 27, 28], as well as simple ensembles that combine predictions from gFMs and pFMs. As a baseline comparison, we also score the zero-shot performance of general-purpose large language models (LLMs), including GPT-4.1 and GPT-4o. We find that pFMs tend to outperform gFMs on protein-centric prediction tasks; further gFM development is needed for reliable protein inference. Meanwhile, general-purpose LLMs tend to perform at or near chance levels. Ensembles of gFMs and pFMs match or outperform the individual models they contain, suggesting improved robustness. Our proposed dataset will extend GENO-PROT to additional species and prediction tasks, while also controlling for sequence similarity in the positive and negative classes [13].

Summary

BioFMs encode complex relationships in nucleotide and amino acid sequences that underlie biological phenotypes. However, existing benchmarks for evaluating these models are largely unimodal, limiting cross-modal comparisons. We propose a novel benchmark that links nucleotide and amino acid sequences to phenotypic labels, formulated as binary prediction tasks designed to probe a model's biological understanding. The benchmark will enable direct comparisons of gFMs and pFMs. By controlling for sequence similarity, our benchmark will highlight cases where traditional alignment-based approaches are poorly suited. Preliminary results using human sequences suggest that current pFMs tend to outperform gFMs on protein-centric prediction tasks, while both tend to outperform general-purpose LLMs. When used in multimodal ensembles, bioFMs tend to achieve more robust performance [13]. However, as bioFMs continue to develop, the performance landscape may vary, and our benchmark will serve as a resource for tracking modeling capabilities.

Acknowledgments and Disclosure of Funding

This work has been authored by Lawrence Livermore National Security, LLC under Contract No. DE-AC52-07NA27344 with the U.S. Department of Energy. This material is based upon work supported by the Department of Energy, Office of Science, Office of Advanced Scientific Computing Research. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for United States Government purposes. Release Number: LLNL-PROP-2010644.

References

[1] ESM Team. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning, 2024.

[2] Yong He, Pan Fang, Yongtao Shan, Yuanfei Pan, Yanhong Wei, Yichang Chen, et al. Generalized biological foundation model with unified nucleic acid and protein language. *Nature Machine Intelligence*, 7(6):942–953, 2025.

[3] Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, et al. Genome modeling and design across all domains of life with Evo 2. *bioRxiv*, page 2025.02.18.638918, 2025.

[4] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 2024.

[5] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in Neural Information Processing Systems*, 36:43177–43201, 2023.

[6] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

[7] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[8] Mohan Timilsina, Samuele Buosi, Muhammad Asif Razzaq, Rafiqul Haque, Conor Judge, and Edward Curry. Harmonizing foundation models in healthcare: A comprehensive survey of their roles, relationships, and impact in artificial intelligence's advancing terrain. *Computers in Biology and Medicine*, 189:109925, 2025.

[9] Kexin Huang, Payal Chandak, Qianwen Wang, Shreyas Havaldar, Akhil Vaid, Jure Leskovec, Girish N Nadkarni, Benjamin S Glicksberg, Nils Gehlenborg, and Marinka Zitnik. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30(12):3601–3613, 2024.

[10] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical AI. *Nature Medicine*, 28(9):1773–1784, 2022.

[11] Chai Discovery. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024.

[12] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

[13] Joseph G. Wakim, Vinayak Gupta, Jose Manuel Marti, Jonathan E. Allen, Brian Bartoldson, and Bhavya Kailkhura. Benchmarking biomolecular foundation models for cross-modal genomics-proteomics. In *NeurIPS 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences*, 2025.

[14] Amina Mollaysa, Artem Moskale, Pushpak Pati, Tommaso Mansi, Mangal Prakash, and Rui Liao. BioLangFusion: Multimodal fusion of DNA, mRNA, and protein language models. arXiv:2506.08936, 2025.

[15] Tamara Goldfarb, Vamsi K. Kodali, Shashikant Pujar, Vyacheslav Brover, Barbara Robbertse, et al. NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Research*, 53(D1):D243–D257, 2025.

[16] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025.

[17] Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V Crichlow, Cole H Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M Duarte, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research*, 49(D1):D437–D451, 2021.

[18] Fredrik Pontén, Karin Jirström, and Matthias Uhlen. The Human Protein Atlas—a tool for pathology. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 216(4):387–393, 2008.

[19] Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, Robert Petryszak, Eliot Ragueneau, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Thawfeek Varusai, Joel Weiser, Adam Wright, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52(D1):D672–D678, 2024.

[20] The Gene Ontology Consortium, Suzi A. Aleksander, James Balhoff, Seth Carbon, J. Michael Cherry, et al. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1), 2023.

[21] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023.

[22] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, 2019.

[23] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023.

[24] Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, 2024.

[25] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

[26] DepMap, Broad. DepMap Public 25Q2. Dataset, 2025.

[27] Theo Sanderson, Maxwell L Bileschi, David Belanger, and Lucy J Colwell. ProteInfer, deep neural networks for protein functional inference. *eLife*, 12:e80942, 2023.

[28] Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, et al. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723):eado9336, 2024.