AUTOMATING LARGE-SCALE *in-silico* BENCHMARKING FOR GENOMIC FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The advancements in artificial intelligence in recent years, such as Large Language Models (LLMs), have fueled expectations for breakthroughs in genomic foundation models (GFMs). The code of nature, hidden in diverse genomes since the very beginning of life's evolution, holds immense potential for impacting humans and ecosystems through genome modeling. Recent breakthroughs in GFMs, such as Evo, have attracted significant investment and attention to genomic modeling, as they address long-standing challenges and transform *in-silico* genomic studies into automated, reliable, and efficient paradigms. In the context of this flourishing era of consecutive technological revolutions in genomics, GFM studies face two major challenges: the lack of GFM benchmarking tools and the absence of open-source software for diverse genomics. These challenges hinder the rapid evolution of GFMs and their wide application in tasks such as understanding and synthesizing genomes, problems that have persisted for decades. To address these challenges, we introduce `GFMBench`, a framework dedicated to GFM-oriented benchmarking. `GFMBench` standardizes benchmark suites and automates benchmarking for a wide range of open-source GFMs. It integrates millions of genomic sequences across hundreds of genomic tasks from four large-scale benchmarks, democratizing GFMs for a wide range of *in-silico* genomic applications. Additionally, `GFMBench` is released as open-source software, offering user-friendly interfaces and diverse tutorials, applicable for `AutoBench` and complex tasks like RNA design and structure prediction. To facilitate further advancements in genome modeling, we have launched a public leaderboard showcasing the benchmark performance derived from `AutoBench`. `GFMBench` represents a step toward standardizing GFM benchmarking and democratizing GFM applications.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Manato Akiyama and Yasubumi Sakakibara. Informative rna base embedding for rna structural alignment and clustering by deep representation learning. *NAR genomics and bioinformatics*, 4 (1):lqac012, 2022.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

Žiga Avsec, Roman Kreuzhuber, Johnny Israeli, Nancy Xu, Jun Cheng, Avanti Shrikumar, Abhimanyu Banerjee, Daniel S Kim, Thorsten Beier, Lara Urban, et al. The kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nature biotechnology*, 37(6):592–600, 2019.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *bioRxiv*, pp. 2022–08, 2022.

Ken Chen, Yue Zhou, Maolin Ding, Yu Wang, Zhixiang Ren, and Yuedong Yang. Self-supervised learning on millions of pre-mrna sequences improves sequence-based rna splicing prediction. *bioRxiv*, pp. 2023–01, 2023.

Yanyi Chu, Dan Yu, Yupeng Li, Kaixuan Huang, Yue Shen, Le Cong, Jason Zhang, and Mengdi Wang. A 5' utr language model for decoding untranslated regions of mrna and function predictions. *Nature Machine Intelligence*, pp. 1–12, 2024.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pp. 2023–01, 2023.

Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, pp. 2021–11, 2021.

Padideh Danaee, Mason Rouches, Michelle Wiley, Dezhong Deng, Liang Huang, and David Hendrix. bprna: large-scale automated annotation and analysis of rna secondary structure. *Nucleic acids research*, 46(11):5381–5394, 2018.

Bernardo P de Almeida, Hugo Dalla-Torre, Guillaume Richard, Christopher Blum, Lorenz Hexemer, Maxence Gélard, Javier Mendoza-Revilla, Priyanka Pandey, Stefan Laurent, Marie Lopez, et al. Segmentnt: annotating the genome at single-nucleotide resolution with dna foundation models. *bioRxiv*, pp. 2024–03, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2021. doi: 10.1101/2021.10.04.463034. URL https://www.biorxiv.org/content/early/2021/10/04/2021.10.04.463034.

Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25, 2023.

Logan Hallee, Nikolaos Rafailidis, and Jason P Gleghorn. cdsbert-extending protein language models with codon awareness. *bioRxiv*, 2023.

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

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Ioanna Kalvari, Eric P Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasinska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, Claire Toffano-Nioche, Daniel Gautheret, Zasha Weinberg, et al. Rfam 14: expanded coverage of metagenomic, viral and microrna families. *Nucleic Acids Research*, 49(D1):D192–D200, 2021.

Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

Zicheng Liu, Jiahui Li, Siyuan Li, Zelin Zang, Cheng Tan, Yufei Huang, Yajing Bai, and Stan Z Li. Genbench: A benchmarking suite for systematic evaluation of genomic foundation models. *arXiv preprint arXiv:2406.01627*, 2024.

Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6:1–14, 2011.

David H Mathews. How to benchmark rna secondary structure prediction accuracy. *Methods*, 162: 60–67, 2019.

Javier Mendoza-Revilla, Evan Trop, Liam Gonzalez, Masa Roller, Hugo Dalla-Torre, Bernardo P de Almeida, Guillaume Richard, Jonathan Caton, Nicolas Lopez Carranza, Marcin Skwark, et al. A foundational large language model for edible plant genomes. *bioRxiv*, pp. 2023–10, 2023.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin W. Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton M. Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Christopher Ré. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *CoRR*, abs/2306.15794, 2023. doi: 10.48550/ARXIV.2306.15794. URL https://doi.org/10.48550/arXiv.2306.15794.

Eric Nguyen, Michael Poli, Matthew G Durrant, Armin W Thomas, Brian Kang, Jeremy Sullivan, Madelena Y Ng, Ashley Lewis, Aman Patel, Aaron Lou, et al. Sequence modeling and design from molecular to genome scale with evo. *bioRxiv*, pp. 2024–02, 2024.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36, 2024.

Yuchen Ren, Zhiyuan Chen, Lifeng Qiao, Hongtai Jing, Yuchen Cai, Sheng Xu, Peng Ye, Xinzhu Ma, Siqi Sun, Hongliang Yan, Dong Yuan, Wanli Ouyang, and Xihui Liu. BEACON: benchmark for comprehensive RNA tasks and language models. *CoRR*, abs/2406.10391, 2024. doi: 10.48550/ARXIV.2406.10391. URL https://doi.org/10.48550/arXiv.2406.10391.

Guillaume Richard, Bernardo P de Almeida, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Marie P Lopez, Alexander Laterre, Maren Lang, et al. Chatnt: A multimodal conversational agent for dna, rna and protein tasks. *bioRxiv*, pp. 2024–04, 2024.

Frederic Runge, Karim Farid, Jorg KH Franke, and Frank Hutter. Rnabench: A comprehensive library for in silico rna modelling. *bioRxiv*, pp. 2024–01, 2024.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*, 2024.

Zhen Tan, Yinghan Fu, Gaurav Sharma, and David H Mathews. Turbofold ii: Rna structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic acids research*, 45(20):11570–11581, 2017.

Ning Wang, Jiang Bian, Yuchen Li, Xuhong Li, Shahid Mumtaz, Linghe Kong, and Haoyi Xiong. Multi-purpose rna language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, pp. 1–10, 2024.

Xi Wang, Ruichu Gu, Zhiyuan Chen, Yongge Li, Xiaohong Ji, Guolin Ke, and Han Wen. Uni-rna: universal pre-trained models revolutionize rna research. *bioRxiv*, pp. 2023–07, 2023.

Jacob West-Roberts, Joshua Kravitz, Nishant Jha, Andre Cornman, and Yunha Hwang. Diverse genomic embedding benchmark for functional evaluation across the tree of life. *bioRxiv*, pp. 2024–07, 2024.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. Peer: a comprehensive and multi-task benchmark for protein sequence understanding. *Advances in Neural Information Processing Systems*, 35:35156–35173, 2022.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nat. Mac. Intell.*, 4(10):852–866, 2022. doi: 10.1038/S42256-022-00534-Z. URL https://doi.org/10.1038/s42256-022-00534-z.

Yuning Yang, Gen Li, Kuan Pang, Wuxinhao Cao, Xiangtao Li, and Zhaolei Zhang. Deciphering 3'utr mediated gene regulation using interpretable deep representation learning. *bioRxiv*, pp. 2023–09, 2023.

Yikun Zhang, Mei Lang, Jiuhong Jiang, Zhiqiang Gao, Fan Xu, Thomas Litfin, Ke Chen, Jaswinder Singh, Xiansong Huang, Guoli Song, et al. Multiple sequence alignment-based rna language model and its application to structural inference. *Nucleic Acids Research*, 52(1):e3–e3, 2024.

Ying Zhang, Fang Ge, Fuyi Li, Xibei Yang, Jiangning Song, and Dong-Jun Yu. Prediction of multiple types of rna modifications via biological language model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2023.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V. Davuluri, and Han Liu. DNABERT-2: efficient foundation model and benchmark for multi-species genome. *CoRR*, abs/2306.15006, 2023. doi: 10.48550/ARXIV.2306.15006. URL https://doi.org/10.48550/arXiv.2306.15006.

A Related Works

A.1 Benchmark

Recognizing the critical role of benchmarking in genomic modeling, several tools have been developed to evaluate genomic models. Among these are RNABench (Runge et al., 2024), GenBench (Liu et al., 2024), BEACON (Ren et al., 2024), and DEGB (West-Roberts et al., 2024).

RNABench focuses on a set of benchmarks, such as RNA secondary structure prediction, and lacks support for evaluating the latest pre-trained models. GenBench is a modular DNA benchmarking framework that provides a DNA evaluation solution but does not extend to RNA benchmarking, and it may not prioritize user-friendliness. BEACON is a recent benchmarking tool aimed at RNA foundational models, offering some RNA evaluation datasets. However, it may lack benchmarking scalability and the complexity of its environment setup poses challenges for novices. DEGB serves as an evaluation benchmark for genomic embeddings, supporting both amino acids and nucleic acids. Its main limitation lies in the small scale of its evaluation benchmarks, and it does not support downstream applications of GFMs. Classic genomic modeling tools like Kipoi[1] (Avsec et al., 2019) have been developed to standardize access to trained models for genomic sequence analysis, offering a repository of models. However, Kipoi focuses on providing access to classic models, not GFMs, rather than benchmarking comprehensively.

There are some protein benchmarking tools, such as ProteinGym (Notin et al., 2024), Flip (Dallago et al., 2021) and Peer (Xu et al., 2022), to name a few. ProteinGym is a large-scale benchmarking tool focused on protein fitness prediction and design. It provides over 250 deep mutational scanning assays, offering a standardized dataset to evaluate machine learning models across millions of mutated protein sequences. ProteinGym is designed to assess both zero-shot and supervised models, particularly in predicting the effects of mutations and aiding protein engineering for applications like genetic disease, agriculture, and healthcare. Flip provides a benchmark for predicting the protein sequence-function relationship, a critical aspect of protein engineering. It includes data for tasks

[1] https://kipoi.org

such as adeno-associated virus stability, protein domain stability, and thermostability from multiple protein families. Flip is designed to evaluate model generalization under various conditions, such as low-resource or extrapolative scenarios. Its datasets are curated to assess the capacity of models to predict functional properties of proteins in real-world protein engineering tasks. Peer is a comprehensive multi-task benchmark that offers 17 tasks across five categories, including protein function prediction, localization, structure, and interaction predictions. It evaluates a wide range of machine learning methods, from traditional approaches to large pre-trained protein language models. Peers' broad scope helps assess model performance in different protein-related tasks, contributing to advancements in protein sequence understanding and engineering.

Existing tools do not adequately address the challenges of comprehensive, large-scale evaluation of RNA and DNA GFMs. They often lack support for downstream applications and do not facilitate the ease of use or scalability necessary to catalyses the democratization and revolution of GFM research. This gap has motivated the development of a new benchmarking tool designed to cover a broad spectrum of foundational DNA and RNA models and provide an extensive benchmarking suite.

A.2 Genomic Foundation Models

In recent years, the modeling of biological sequences, including DNA, RNA, and proteins, has garnered significant attention. Protein modeling, exemplified by works such as AlphaFold (Jumper et al., 2021; Evans et al., 2021; Abramson et al., 2024) and ESM (Lin et al., 2022), has advanced considerably over the past years, outpacing developments in DNA and RNA modeling.

In the domain of genomic sequence modeling, early efforts focused on adapting natural language processing architectures to handle genomic data. For instance, DNABERT (Ji et al., 2021) repurposed the BERT (Devlin et al., 2019) architecture for genomic sequences, demonstrating preliminary success on *in-silico* genomic tasks. Building upon this, DNABERT2 (Zhou et al., 2023) introduced improvements by replacing k-mer tokenization with byte-pair encoding (BPE) tokenization, enhancing model performance across multiple species.

To explore the capabilities of large-scale foundation models (FMs), the Nucleotide Transformers V2 (Dalla-Torre et al., 2023), AgroNT (Mendoza-Revilla et al., 2023), and SegmentNT (de Almeida et al., 2024) scaled models to billions of parameters. These models achieved promising results in understanding DNA genomes, with parameter counts reaching up to 2.5 billion and 1 billion, respectively. AgroNT, pre-trained on multi-species edible plant DNA sequences, however, did not transfer effectively to RNA sequence modeling in subsequent experiments. Addressing the challenge posed by the considerable length of genomic sequences, recent works have emphasized long-range sequence modeling and introduced auto-regressive FMs, such as HyenaDNA (Nguyen et al., 2023) and Evo (Nguyen et al., 2024).

In the context of RNA genomic modeling, several preliminary studies have emerged, including scBERT (Yang et al., 2022), RNABERT (Akiyama & Sakakibara, 2022), RNA-FM (Chen et al., 2022), RNA-MSM (Zhang et al., 2023), and RNAErnie (Wang et al., 2024). These models, however, are typically trained on limited-scale databases due to the scarcity and expense of obtaining RNA sequences. Some FMs focus on specific RNA types, such as coding sequences (CDS)(Hallee et al., 2023), 5' untranslated regions (5'UTR)(Chu et al., 2024), 3' untranslated regions (3'UTR)(Yang et al., 2023), or precursor mRNA sequences(Chen et al., 2023), which constrains their ability to capture the full diversity of RNA sequences. Uni-RNA (Wang et al., 2023) has been reported to achieve strong performance owing to its large-scale model and extensive database. However, it is not open-sourced, precluding direct comparison in experiments. ChatNT (Richard et al., 2024) is a multimodal conversational agent designed to assist with tasks involving DNA, RNA, and protein sequences. It can handle diverse genomic and proteomic tasks, such as predicting sequence structures, simulating biological processes, or interacting with foundational models. ChatNT integrates advanced AI models to facilitate research in genomic data processing, enhancing accessibility and scalability in tasks across multiple biological modalities.

Table 1: The brief statistics of subtasks in the RGB. These benchmark datasets are held out or not included in the pretraining database. The numbers of examples in training, validation and testing sets are separated by "/". * indicates the datasets are used for zero-shot performance evaluation only.

Task	Task Type	# of examples	# of classes	Metric	Sequence length	Source
SNMD	Token classification	$8,000/1,000/1,000$	2	AUC	200	This work
SNMR	Token classification	$8,000/1,000/1,000$	4	macro F1	200	This work
mRNA	Token regression	$1,735/193/192$	—	RMSE	107	Kaggle
bpRNA	Token classification	$10,814/1,300/1,305$	3	macro F1	≤ 512	(Danaee et al., 2018)
AchiveII	Token classification	$2278/285/285$	3	macro F1	≤ 500	(Mathews, 2019)
RNAStrAlign	Token classification	$17483/2186/2185$	3	macro F1	≤ 500	(Tan et al., 2017)

B Benchmark Details

B.1 RNA Genomic Benchmark

The detailed task descriptions for each nucleic acid and species, including the number of examples, classes, evaluation metric, and sequence length, are outlined in Table 1. Each task is carefully curated to reflect the complexity and variety inherent in genomic data, providing a robust framework for assessing the nuanced capabilities of state-of-the-art RNA FMs. RGB contains 6 SN-level tasks that are curated or collected from published articles. The purpose of RGB is to benchmark genomic FMs in challenging SN-level modeling tasks such as the detection and repair of SN mutations, mRNA sequence degradation rates, and RNA secondary structure prediction. Due to the lack of a plant RNA benchmark dataset, RGB includes the modeling of RNA sequences from a variety of species, e.g., plant and human. The sequence length in RGB ranges from 107 to 512, which is sufficient for most RNA understanding tasks. In summary, these multi-species and SN-level tasks in RGB serve as the first comprehensive benchmark utilized to assess the RNA sequence modeling capabilities of `GFMBench` and its baseline models. The brief introduction of the datasets in RGB is as follows:

- **Single-Nucleotide Mutation Detection (SNMD)**: We developed a plant RNA dataset synthesizing the single-nucleotide mutations. Focused on identifying potential single nucleotide changes, this task is essential for detecting mutations linked to genetic disorders. The SNMD dataset introduces up to 10 random mutations in the original sequences, regardless of variation ratios. Cross-entropy is utilized as the loss function for this binary token classification task.
- **Single-Nucleotide Mutation Repair (SNMR)**: This task challenges the model to suggest corrective actions at the single nucleotide level, aiding in gene therapy approaches. The SNMR dataset mirrors the SNMD dataset, with cross-entropy as the loss function, indicating a token 4-way (i.e., `A`, `U`, `C`, `G`) classification task.
- **mRNA Degrade Rate Prediction (mRNA)**: Estimating the decay rate of nucleotides in mRNA sequences, this task is vital for deciphering gene expression and regulation. The dataset originates from the Kaggle COVID-19 vaccine design competition[2], focusing solely on sequence-based degradation rate prediction and excluding RNA structures. It's a token regression task using MSE as the loss function, with the dataset re-split into training, validation, and testing sets for evaluation.
- **RNA Secondary Structure Prediction (bpRNA & Archive2 & RNAStralign)**: Aiming to predict RNA folding into secondary structures, this task is fundamental to RNA functionality and interactions. We evaluated `GFMBench` on four datasets, bpRNA (Danaee et al., 2018) (TR0, VL0, TS0 sets), ArchiveII (Mathews, 2019), RNAStralign (Tan et al., 2017) and Rfam (Kalvari et al., 2021). Following existing works, we have excluded sequences over 512 bases and complex structures, simplifying to three symbols: `'('`, `'.'`, `')'`.Ṙesults may not directly compare with other studies due to these modifications. Cross-entropy serves as the loss function.

Please find the appendix for the input and output examples of each subtask in RGB. The detailed task descriptions for each nucleic acid and species, including the number of examples, classes, evaluation metric, and sequence length, are outlined in Table 1. Each task is carefully curated to reflect the complexity and variety inherent in genomic data, providing a robust framework for assessing the nuanced capabilities of state-of-the-art RNA FMs.

[2] `https://www.kaggle.com/competitions/stanford-covid-vaccine`

Table 2 show the virtual examples of different datasets in RGB. Please refer to our supplementary materials to find the datasets for more details.

Table 2: The virtual input and output examples in the four benchmarks. The "..." represents the sequences that are omitted for better presentation and the red color indicates the wrong prediction in classification tasks. In the mRNA dataset, all single nucleotides have three values to predict. Note that "T" and "U" can be regarded as the same symbol in RNA sequences and depend on different datasets.

Genome Type	Dataset	Column	Examples
RNA	SNMD	Input Sequence	G A G T A ... T T G A G
		True Label	0 0 1 0 0 ... 0 0 1 0 0
		Prediction	0 0 0 0 0 ... 0 0 1 0 0
	SNMR	Input Sequence	T A C G A ... C T G A T
		True Label	T A C A A ... G T A A T
		Prediction	T A C A A ... C T G A T
	mRNA	Input Sequence	G G ... A C
		True Label	[0.1,0.3,0.2] [0.8,0.4,0.1]... [0.9,0.4,0.3] [0.5,0.2,0.6]
		Prediction	[0.1,0.3,0.2] [0.8,0.4,0.1]... [0.9,0.4,0.3] [0.5,0.2,0.6]
	bpRNA	Input Sequence	G G C G A ... C U U U U
		True Label	(((· ·...· ·)))
		Prediction	((((·...·))))
DNA	Classification	Input Sequence	A T C G A ... T A G
		True Label	1
		Prediction	0
	Regression	Input Sequence	G C C A T ... G C T
		True Label	2.56
		Prediction	2.45
	Chrom Acc (Multi-label)	Input Sequence	A T C G ... C T G
		True Label	[1, 0, 1, 1, 0, 1, 1, 0, 1]
		Prediction	[1, 1, 1, 1, 0, 1, 1, 0, 1]

B.2 Plant Genomic Benchmark

PGB (Mendoza-Revilla et al., 2023) provides a comprehensive suite of datasets designed to evaluate and improve the predictive capabilities of GFMs in plant biology. This benchmark, as shown in Table 3, encompasses a range of critical genomic tasks, including binary classification, single and multi-variable regression, and multi-label classification, addressing various aspects of plant genomics such as RNA processing, gene expression, and chromatin accessibility. By integrating diverse genomic tasks, the PGB aims to facilitate advanced research and development in plant genomics, offering a robust platform for the assessment and enhancement of model performance across different plant species. To obtain a detailed description of PGB, please refer to Agro-NT (Mendoza-Revilla et al., 2023).

B.3 Genomic Understanding Evaluation

GUE (Zhou et al., 2023) serves as a DNA genomic benchmark, encompassing 36 datasets across nine crucial genome analysis tasks applicable to a variety of species. Similar to PGB and GB, it is used for evaluating the generalizability of `GFMBench` on DNA genome benchmarking. To thoroughly assess the capabilities of genome foundation models across sequences of varying lengths, tasks have been chosen with input lengths spanning from 70 to $10,000$. The brief statistics for each dataset included in the GUE benchmark are displayed in Table 4, and the task descriptions are available in Zhang et al. (2023). Due to resource limitations, we do not include large-scale FMs in this benchmark, e.g., Agro-NT. Besides, we run the evaluation on a subset of GUE, where for each task we randomly select at most 10k samples from the original splits, e.g., training, testing and validation (if any) sets.

B.4 Genomic Benchmarks

GB is also a DNA-oriented FM benchmark suite, which can be used for generalizability evaluation of OmniGenome. It contains a well-curated collection of datasets designed for the classification of genomic sequences, focusing on regulatory elements across multiple model organisms. This collection facilitates robust comparative analysis and development of genomic FMs. The task names in the original repository are complex, we abbreviate the names as follows:

Table 3: The genomic tasks in the Plant Genomic Benchmark. This table briefly enumerates each task by name, the number of datasets available, the type of classification or regression analysis required, the range of sequence lengths, and the total number of samples in each dataset. Please find the dataset details of PGB in Agro-NT.

Task	# of datasets	Task Type	Total # of examples	# of classes	Metric	Sequence length
Polyadenylation	6	Sequence classification	738,918	2	macro F1	400
Splice site	2	Sequence classification	4,920,835	2	macro F1	398
LncRNA	2	Sequence classification	58,062	6	macro F1	$101-6000$
Promoter strength	2	Sequence regression	147,966	—	RMSE	170
Terminator strength	2	Sequence regression	106,818	—	RMSE	170
Chromatin accessibility	7	Multi-label classification	5,149,696	$9-19$	macro F1	1,000
Gene expression	6	Multi-variable regression	206,358	—	RMSE	6,000
Enhancer region	1	Sequence classification	18,893	2	macro F1	1,000

Table 4: Statistics of tasks in the GUE, these details can be found in Section B.2. from Zhang et al. (2023).

Task	Metric	Datasets	Training	Validation	Testing
Core Promoter Detection	macro F1	tata	4,904	613	613
		notata	42,452	5,307	5,307
		all	47,356	5,920	5,920
Promoter Detection	macro F1	tata	4,904	613	613
		notata	42,452	5,307	5,307
		all	47,356	5,920	5,920
Transcription Factor Prediction (Human)	macro F1	wgEncodeEH000552	32,378	1,000	1,000
		wgEncodeEH000606	30,672	1,000	1,000
		wgEncodeEH001546	19,000	1,000	1,000
		wgEncodeEH001776	27,497	1,000	1,000
		wgEncodeEH002829	19,000	1,000	1,000
Splice Site Prediction	macro F1	reconstructed	36,496	4,562	4,562
Transcription Factor Prediction (Mouse)	macro F1	Ch12Nrf2\iggrab	6,478	810	810
		Ch12Zrf384hpa004051\iggrab	5,395	674	674
		MelJun\iggrab	2,620	328	328
		MelMafkDm2p5dStd	1,904	239	239
		MelNelf\iggrab	15,064	1,883	1,883
Epigenetic Marks Prediction	macro F1	H3	11,971	1,497	1,497
		H3K14ac	26,438	3,305	3,305
		H3K36me3	29,704	3,488	3,488
		H3K4me1	25,341	3,168	3,168
		H3K4me2	24,545	3,069	3,069
		H3K4me3	29,439	3,680	3,680
		H3K79me3	23,069	2,884	2,884
		H3K9ac	22,224	2,779	2,779
		H4	11,679	1,461	1,461
		H4ac	27,275	3,410	3,410
Covid Variant Classification	macro F1	Covid	77,669	7,000	7,000
Enhancer Promoter Interaction	macro F1	GM12878	10,000	2,000	2,000
		HeLa-S3	10,000	2,000	2,000
		HUVEC	10,000	2,000	2,000
		IMR90	10,000	2,000	2,000
		K562	10,000	2,000	2,000
		NHEK	10,000	2,000	2,000
Species Classification	macro F1	fungi	8,000	1,000	1,000
		virus	4,000	500	500

- DEM corresponds to "Demo Coding vs Intergenomic Seqs"
- DOW is for "Demo Human or Worm"
- DRE represents "Drosophila Enhancers Stark"
- HCE is short for "Human Enhancers Cohn"
- HEE denotes "Human Enhancers Ensembl"
- HRE abbreviates "Human Ensembl Regulatory"
- HNP shortens "Human Nontata Promoters"

- HOR is an abbreviation for "Human Ocr Ensembl"
- DME simplifies "Dummy Mouse Enhancers Ensembl"

The brief statistics for each dataset included in the GUE benchmark are displayed in Table 4. Similar to GUE, we run the evaluation on a subset of GB, where for each task we randomly select at most 10k samples from the original splits, e.g., training, testing and validation (if any) sets.

Table 5: The brief statistics of datasets reported in the genomic benchmark (Grešová et al., 2023).

Task	# of Sequences	# of Classes	Class Ratio	Median Length	Standard Deviation
DME	$1,210$	2	1.0	$2,381$	984.4
DEM	$100,000$	2	1.0	200	0.0
DOW	$100,000$	2	1.0	200	0.0
DRE	$6,914$	2	1.0	$2,142$	285.5
HCE	$27,791$	2	1.0	500	0.0
HEE	$154,842$	2	1.0	269	122.6
HRE	$289,061$	3	1.2	401	184.3
HNP	$36,131$	2	1.2	251	0.0
HOR	$174,456$	2	1.0	315	108.1

C Data Filtering in Benchmarking

The pertaining involves RNA sequences and structures prediction, we take the data and annotation leakage problem seriously.

- To avoid structure annotation leakage of downstream benchmarks, the secondary structure predictors for all FMs were randomly initialized for fair comparisons, which means the pre-trained structure predictor of `GFMBench` was not used in benchmarks, except for zero-shot SSP experiments. Please find the source codes for details.
- To reduce sequence leakage caused by evolutionary conservative sequences across multiple species, we use the ch-hit-est tool to calculate the sequence similarity between sequences from the OneKP database and downstream tasks. We adopt the similarity threshold of 80% for ch-hit-est (Li & Godzik, 2006) to eliminate sequences whose homogeneous sequences appeared in the OneKP database. Subsequently, we exploit the blastn (Altschul et al., 1990) tool to query potentially leaked sequences in downstream benchmark datasets and further alleviate the data leakage problem. The e-value has been set to 1 for rigorous sequence filtering.

C.1 Experiment Settings

In this experiment, we carefully selected a set of key hyperparameters to optimize model performance. Below are the main hyperparameter settings along with detailed explanations:

- **Dropout**: To prevent the model from overfitting during training, we set the Dropout value to 0, meaning that no random neuron dropout is applied during training. This choice was made based on our consideration of model stability and generalization ability.
- **Learning Rate**: We set the learning rate to 2e-5, which is a relatively small value to ensure stable convergence, especially in complex training tasks. A smaller learning rate helps to avoid drastic fluctuations during the training process, leading to more precise optimization.
- **Weight Decay**: We applied a weight decay of 0.01 to control model complexity and prevent overfitting. Weight decay is a regularization technique that effectively constrains the growth of model parameters, maintaining the model's generalization capability.
- **Adam Optimizer**: We used the Adam optimizer with its parameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The Adam optimizer combines the benefits of momentum and adaptive learning rates, accelerating convergence and adapting to different gradient changes, thereby improving the efficiency and effectiveness of model training.

Table 6: The brief statistics of RNA and DNA FM baselines. Please note that the pertaining data scales cannot be directly compared because the measurements are different in various publications. The detailed introduction of these FMs can be found in original publications.

Model	Tokenization	# of Params.	Pre-training Data Scale	Pre-training Data Source	Species	Sequence Type
DNABERT-2	BPE	117M	32.49B Tokens	The 1000 Genomes Project	Human + 135 Species	DNA
NT-V2-100M	k-mers	96M	300B Tokens	The 1000 Genomes Project, etc.	Human + 850 Species	DNA
HyenaDNA-Large	SNT	47M	3.2B Tokens	Genome Reference Consortium	Human	DNA
Caduceus	SNT	1.9M	35B Tokens	Genome Reference Consortium	Human	DNA
Agro-NT-1B	k-mers	985M	472.5B Tokens	Ensembl Plants Database	48 Edible Plants	DNA
SpliceBERT	SNT	19M	2M Sequences	UCSC Genome Browser	Multi-Vertebrates	precursor-mRNA
RNA-BERT	SNT	0.5M	4,069 RNA Families	The RNA Families Database	Multi-Species	ncRNA
RNA-MSM	SNT	96M	4,069 RNA Families	The RNA Families Database	Multi-Species	ncRNA
RNA-FM	SNT	96M	23M Sequences	RNAcentral Database	Multi-Species	ncRNA
3UTRBERT	k-mers	86M	20,362 Sequences	The GENCODE Project	Human	mRNA 3'UTR
OmniGenome	SNT	186M	54.2B Tokens	The OneKP Initiative	1124 Plant Species	mRNA, CDS, UTR

- **Learning Rate Scheduler**: We opted for a linear decay learning rate scheduler, allowing the learning rate to gradually decrease during training. This strategy helps the model make smaller adjustments as it approaches the optimal solution, ensuring a better convergence outcome.
- **Batch Size**: The batch size was set to 8. This relatively small batch size helps to efficiently train the model within limited memory resources, particularly when handling large-scale data, enabling a balance between model performance and computational resource usage.
- **# of Epochs**: We set the number of training epochs to 20. This setting ensures that the model can fully learn the features within the data while avoiding the negative effects of overtraining.
- **Early Stopping**: We implemented an early stopping mechanism, terminating the training early if the validation performance does not improve for 5 consecutive epochs. This mechanism effectively prevents model overfitting and saves training time.

It is important to note that for different tasks, some hyperparameter settings may be adjusted. To obtain accurate experimental results, please refer to the detailed parameter configurations in the compiled dataset specific to each task.

C.2 Development Environment

The benchmark experiments based on `GFMBench` were conducted on a dedicated Linux computation node, equipped with 2 NVIDIA RTX 4090 GPUs. For distributed model training, we employed version 4.44.0 of the Transformers library alongside version 0.28.3 of the Accelerate library. Our implementation framework of choice for `GFMBench` was PyTorch, specifically version 2.1.0. The ViennaRNA version is 2.6.4 in our experiments. While some existing code was adapted for the modules within `GFMBench`, the majority of the codebase, such as genomic sequences preprocessing, model pre-training, objective functions, and experiments, was meticulously crafted from scratch.

C.3 Evaluation Baselines

To comprehensively evaluate the performance of the existing GFMs across the integrated benchmarks, i.e., RGB, PGB, GUE and GB, we have obtained the results of existing GFMs based on `GFMBench`.

Please note that it is assumed that the structure annotation from ViennaRNA is always available for structure-contextualized modeling to enhance OmniGenome. In SSP tasks, we can also use the ViennaRNA's structure annotations as contexts to improve downstream SSP performance. Please refer to Appendix C.3 for brief introductions of these FMs.

We can compare `GFMBench` with the following RNA and DNA FMs shown in Table 6 as baselines to help evaluate the performance of `GFMBench`. We are aware that some FMs are also developed for RNA, such as Uni-RNA (Wang et al., 2023), 5UTR-LM (Chu et al., 2024), etc. However, we cannot compare `GFMBench` with them because their source codes are very hard to work with in our efforts or are not publicly available. To help understand the baseline FMs, we briefly summaries the FM in the following sections. Please find the method and experiment details of these FMs in the original publications.

- ViennaRNA (Lorenz et al., 2011). ViennaRNA is a comprehensive genomic analysis tool that includes a diverse set of interfaces, such as RNAFold[3] and RNAInverse[4] design. ViennaRNA serves as the baseline for RNA structure prediction and RNA design in our experiments.
- DNABERT2 (Zhou et al., 2023). DNABERT2 is one of the latest DNA FMs which improves the performance of DNABERT. The main modification of DNABERT2 is the tokenization method, which was changed to BPE from k-mers.
- HyenaDNA (Nguyen et al., 2023). HyenaDNA is an autoregressive FM optimized for long-range genome data processing. HyenaDNA is based on the Hyena convolution architecture and capable of handling sequences up to 1M bases in length.
- Caduceus (Schiff et al., 2024). Caduceus[5] is an advanced DNA language model built on the MambaDNA architecture, designed to address challenges in genomic sequence modeling, such as long-range token interactions and reverse complementarity (RC).
- Nucleotide Transformer (NT) V2 (Dalla-Torre et al., 2023). The NT FMs were trained on DNA data, including the human reference genome and multi-species DNA sequences. They aim to capture the complex patterns within nucleotide sequences for various genome modeling applications.
- Agricultural Nucleotide Transformer (Agro-NT) (Mendoza-Revilla et al., 2023). Agro-NT is a large-scale DNA FM (1B parameters) akin to the Nucleotide Transformers but with a focus on plant DNA.
- SpliceBERT (Chen et al., 2023). It was trained on 2M precursor messenger RNA (pre-mRNA) and specialised in RNA splicing of pre-mRNA sequences.
- 3UTRBERT (Yang et al., 2023). This model was trained on 20k 3'UTRs for 3'UTR-mediated gene regulation tasks. It uses k-mers tokenization instead of SNT. RNA-BERT (Akiyama & Sakakibara, 2022). RNA-BERT is a BERT-style model pre-trained on a large corpus of non-coding RNA sequences. It uses masked language modeling (MLM) as its primary training objective. The model is designed to predict RNA structural alignments and can be fine-tuned for various RNA sequence classification and regression tasks
- RNA-MSM (Zhang et al., 2024) RNA-MSM is an unsupervised RNA language model based on multiple sequence alignment (MSA). It is the first model of its kind to produce embeddings and attention maps that directly correlate with RNA secondary structure and solvent accessibility. RNA-MSM is particularly effective for tasks involving evolutionary relationships in RNA sequences.
- RNA-FM (Chen et al., 2022) RNA-FM is a BERT-based RNA foundation model trained on a vast dataset of non-coding RNA sequences. The model excels in predicting RNA structure and function by leveraging masked language modeling (MLM) during pre-training. RNA-FM's training data is sourced from the RNAcentral database, providing it with extensive knowledge across diverse RNA species.
- `GFMBench`. `GFMBench` is the RNA genome FM that advocates the importance of sequence-structure alignment. Moreover, it is the first FM which addressed the *in-silico* RNA design task.
- **OmniGenome**: A FM dedicated to RNA genome modeling. This model leverages the computation-based structure to enhance the genome modeling ability and archives impressive performance on both RNA and DNA genomes.

D PUBLIC LEADERBOARD

The public leaderboard has been launched with the manuscript, and the current layout of the leaderboard is illustrated in Figure 1. We have included the results of open-source GFMs among four benchmark suites, and new results can be expected from the community. We are still working to include the performance of recent GFMs, and refine the leaderboard interface with better integrity.

[3] `https://www.tbi.univie.ac.at/RNA/RNAfold.1.html`

[4] `https://www.tbi.univie.ac.at/RNA/RNAinverse.1.html`

[5] `https://huggingface.co/kuleshov-group/caduceus-ps_seqlen-131k_d_model-256_n_layer-16`

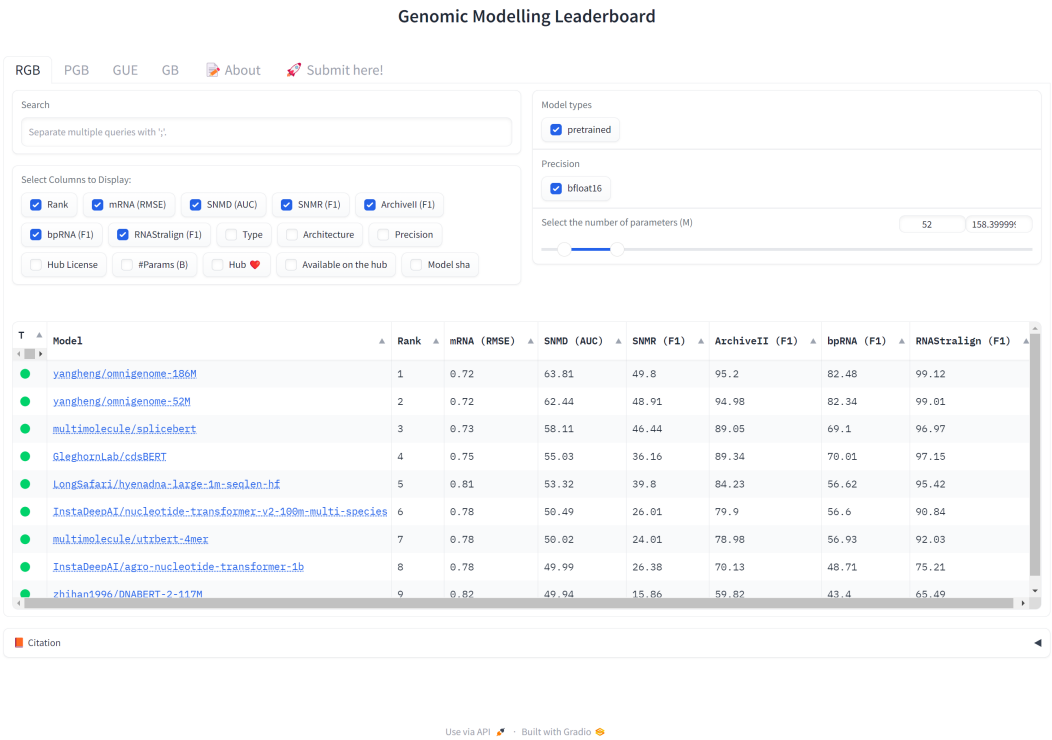

Figure 1: The current webpage interface of the public leaderboard.

E LIMITATIONS

The GFM benchmarking may not reflect the accurate performance in biology reality, we attribute the limitations of benchmarking to two major aspects:

- **Lack of *in-vivo* Data**: One of the critical limitations of GFMs lies in the absence of *in-vivo* verified genome data. While GFMs perform well in *in-silico* environments, where computational models and simulations are used to predict biological processes, these models are rarely validated against *in-vivo* data, which refers to experimental data obtained from living organisms. This presents a significant challenge for accurately translating model predictions to real-world biological applications. To be more specific, the complexity of biological systems, including interactions within cells, tissues, and organisms, often introduces variables that are not fully captured in computational simulations. For example, gene regulation, environmental factors, and cellular responses to genetic modifications may behave differently in living organisms than predicted by models trained on *in-silico* data. As a result, GFMs might not fully capture the biological complexity, leading to discrepancies between predicted and actual outcomes.
- **Model Scale Constraints**: The second major limitation is the model scales in benchmarking. As GFMs become larger and more sophisticated, their performance improves, but this scaling comes at a significant cost. Training as well as benchmarking large-scale GFMs requires immense computational resources, including high-performance GPUs or TPUs, massive memory allocation, and extensive storage for datasets. The cost of acquiring and maintaining this infrastructure can be prohibitive for many research institutions or companies, limiting access to cutting-edge GFMs.

F ETHIC STATEMENT

The development of GFMs presents various ethical challenges that must be carefully considered. As we push the boundaries of what is possible with large-scale GFMs, such as Evo, it is crucial to establish a responsible framework for their development and application. GFMs enable advanced capabilities like generating and predicting DNA sequences at a whole-genome scale, which opens the

door to significant breakthroughs in fields such as genetic engineering and therapeutic development. However, these same capabilities pose risks related to bio-security, inequality, and environmental disruption.

Safety and Ethical Implications: GFMs like OmniGenome could be misused by malicious actors for harmful purposes, such as creating synthetic organisms that could threaten bio-safety. It is essential to establish strict guidelines on access and use, including the development of safety guardrails, access controls, and audits to monitor queries and research outcomes.

Health and Social Inequity: While the open-source nature of GFMs promotes transparency and accessibility, there are concerns that the benefits of these tools may disproportionately favor well-resourced organizations, such as pharmaceutical companies, which could lead to further inequalities in global health. Intellectual property considerations also arise, as companies using open-source tools might monopolize treatments or set prohibitive costs, exacerbating health disparities.

Environmental Impact: The enhanced capabilities for genetic manipulation that GFMs enable could disrupt natural ecosystems, leading to potential loss of biodiversity or the emergence of harmful species. Additionally, the computational demands of training large models have environmental costs, such as increased carbon footprints, that must be weighed against the benefits of the scientific advancements.

In response to these concerns, we are committed to promoting ethical guidelines, transparency, and the responsible use of GFMs. We will collaborate with the community to continually refine these guidelines as the field evolves.

G SOCIAL IMPACT

The societal impact of GFMs is substantial, with applications ranging from personalized medicine to environmental management. These models have the potential to revolutionize fields such as healthcare and agriculture by providing deeper insights into genetic data, enabling the discovery of new biomarkers, and assisting in the development of more effective therapies. In healthcare, GFMs can drive advancements in precision medicine, allowing for personalized treatments based on individual genetic profiles, which could drastically improve patient outcomes for conditions such as cancer or rare genetic disorders. In agriculture, GFMs can contribute to sustainable practices by improving crop yields and resistance to disease. However, careful consideration must be given to the ecological balance, as genetic modifications could have unforeseen consequences on ecosystems. As GFMs continue to evolve, their responsible development and deployment will be crucial to ensuring that their societal impact is positive and equitable.

However, there are also risks associated with the unequal access to these powerful tools. Entities with more resources and technical expertise may benefit disproportionately from GFMs, accelerating their research and economic returns while leaving lower-resourced institutions and countries at a disadvantage. To mitigate this, it is critical to ensure that access to GFMs is democratized through open-source initiatives, global collaboration, and capacity-building efforts in low-resource settings.