Are Genomic Language Models All You Need? Exploring Genomic Language Models on Protein Downstream Tasks

Sam Boshar[1, 2], **Evan Trop**[1], **Bernardo P. de Almeida**[1], **Thomas Pierrot**[1]

[1]InstaDeep
[2]Massachusetts Institute of Technology
sboshar@mit.edu, e.trop@instadeep.com, b.dealmeida@instadeep.com, t.pierrot@instadeep.com

Abstract

In recent years, large language models trained on enormous corpora of unlabeled biological sequence data have demonstrated state-of-the-art performance on a variety of downstream tasks. These LLMs have been successful in modeling both genomic and proteomic sequences and their representations have been used to outperform specialized models in a myriad of tasks. Since the genome contains the information to encode all proteins, genomic language models hold the potential to make downstream predictions about proteins as well as DNA. This observation motivates a model that can perform well on both genomic and proteomic downstream tasks. However, since there are few tasks which pair proteins to their true coding DNA sequences, it is difficult to compare the two model types. In this work we curate five such datasets and use them to evaluate the performance of multiple state-of-the-art genomic and proteomic models. We find that, despite their pre-training on largely non-coding sequences, genomics language models are competitive and even outperform their protein counterparts on some tasks.

Introduction

Large language models (LLMs), thanks to their capability to learn through self-supervision from unlabeled data, have recently revolutionized the field of NLP (Devlin et al. 2018; Brown et al. 2020; Raffel et al. 2020). Recently, the same techniques have been applied to learn from biological data. The discrete and sequential structure of biological sequences such as proteins or DNA and RNA paired with the abundance of unlabeled data, obtained through high-throughput sequencing, makes it a perfect application field for these methods to strive.

This effort started first in proteomics where several works (Lin et al. 2022; Elnaggar et al. 2021) showed that training large Transformer models to recover masked amino-acids in protein sequences yield these models to produce powerful representations that can then be used to solve diverse downstream tasks with state-of-the-art performance (Lin et al. 2023). More recently, similar models were developed for genomics (Dalla-Torre et al. 2023; Zhou et al. 2023; Ji et al. 2021; Nguyen et al. 2023), and trained over the human reference genome as well as hundreds of reference genomes from different species to recover masked consecutive nucleotides in chunks. These DNA models, while more recent and still less mature than their protein counterparts, have also showed the ability to build strong representations of nucleic acid sequences to solve downstream tasks with improved performance. Notably, these models were mainly evaluated to be able to predict diverse DNA molecular phenotypes such as splice sites, regulatory elements and chromatin profiles.

Motivated by the central dogma of biology which states that the genome encodes all protein information and by the fact that codon usage can influence protein structure and function (Liu 2020), a third class of models, codon-based, was recently introduced with models such as CaLM and CodonBERT (Outeiral and Deane 2022; Li et al. 2023). These models were trained on large datasets made of coding sequences (CDS) by reconstructing masked codons - instead of masked amino-acids. These models were then showed to outperform their amino-acid based counterparts on several protein downstream tasks of interest.

Inspired by these recent results, we aim to study to what extent genomics language models (gLMs) can be used as a general unified approach to solve tasks in both worlds - genomics and proteomics. In opposition to codon-based language models (cLMs), genomics language models have been trained over full raw genomes from which coding sequences represent on average only 2% of the data. In addition, these models have never seen "true" CDS per se during training as exons are always separated by introns in eukaryotic species genomes. Therefore, it is an interesting question to what extent these models can be competitive with protein language models (pLMs).

However, this comparison is not straightforward due to the scarcity of labeled datasets containing both proteins and their associated CDS. Identifying and retrieving the CDS for a protein dataset can be a difficult, expensive and unfruitful process. While prior works such as CaLM and CodonBERT built such tasks to evaluate their codon-based models they did not release the datasets nor the weights of the models, hindering the reproducibility of such results. In this paper, we study five standard protein datasets from the protein language models literature and provide the curated CDS to enable a systematic comparison of protein and genomics language models. We decided to focus our attention on the ESM2 (650M parameters) and ESM1b (650M) models for

Dataset	Task Type	# Train Samples	# Validation Samples	# Test Samples	Mean Sequence Length (bp)
Fluorescence	Regression	21464	5366	27217	714
Beta-Lactamase (Unique)	Regression	3457	865	1080	858
Beta-Lactamase (Complete)	Regression	11252	2814	1080	858
Stability	Regression	53700	2512	12851	135
Melting Point	Regression	9432	1064	1648	1176
Secondary Structure Prediction	Per AA Classification	6224	1556	334	724

Table 1: Overview of the proposed tasks. Samples in each dataset contain protein sequences paired with nucleotide sequences. Total sampled over all 3 test sets is provided for SSP.

proteins and the Nucleotide Transformer v2 (500M) and DNABERT2 (117M) for genomics, as these are considered as the state-of-the-art models in their respective fields. After showing the importance of getting access to true CDS versus reconstructed ones through uniform or biased sampling, we compare these models over the five tasks. We show that genomics language models outperform their protein counterparts on three out of five tasks. Finally, we carry out further analysis on one of the tasks, seeking to understand the disparity of Nucleotide Transformer's increased performance against ESM.

Protein Downstream Tasks Datasets

We study five protein tasks of interest that are frequent in the literature. This collection includes sequence- and residue-level tasks, spanning regression and multi-label classification. We detail and motivate below these five tasks. See Table 1 for an overview of these tasks.

Secondary Structure Prediction (SSP): Understanding the structure of proteins is integral to understanding their function. This task tests a model's ability to learn local secondary structure. The task is a multi-label classification task where each input amino-acid is associated with one of 8 labels, denoting which secondary structure that residue is a part of. Following the work of Klausen (Høie et al. 2022) we used splits filtered at 25% sequence identity to ensure generalization, and evaluated on 3 test sets: CASP12, CB513, TS115.

Melting Point Prediction (MPP): Predicting protein melting point can be a challenging task as even single residue mutations can have large impacts (Pinney et al. 2021). Melting point prediction is a sequence-level regression task that evaluates a model's ability to predict a measure of melting temperature. We follow the same "mixed" splits described in FLIP (Dallago et al. 2021) which seek to avoid over-emphasis of large clusters. Sequences are clustered at 20% identity with 80% of clusters assigned to the train dataset and 20% of clusters assigned to the test dataset.

Fluorescence Prediction: Estimating the fitness landscape of proteins which are many mutations away from the wildtype sequence is one of the core challenges of protein design. This task evaluates a model's ability to predict log-fluorescence of higher-order mutant green fluorescent protein (GFP) sequences. Original data is from an experimental study of the GFP fitness landscape (Sarkisyan et al. 2016). Inspired from the TAPE and PEER benchmarks (Rao et al. 2019; Xu et al. 2022), we restrict the training set to amino-acid sequences with three or fewer mutations from parent GFP sequences, while the test set is all sequences with four or more mutations.

Beta-lactamase Activity Prediction: It is also important for models to have the precision to accurately predict the effects of single amino-acid mutations (Xu et al. 2022). Beta-Lactamase is a regression task consisting of sequences from a study exploring the fitness landscape of all single codon substitutions in the TEM-1 gene (Firnberg et al. 2014). Labels indicate the ability of mutant genes to confer ampicillin resistance.

Protein Stability Prediction: It is important for models trained on diverse sequences to be able to accurately predict a small region of the fitness landscape. This task evaluates how well models predict stability around a small region of high-fitness sequences. Labels indicate a peptide's ability to maintain structure at increasing levels of protease, which serves as a proxy for stability.

Retrieving and Curating Coding Sequences

One main contribution of this work is to retrieve, curate, and share consolidated CDS datasets for the five protein tasks of interest to allow the comparison of nucleic acid- and amino-acid-based models. We detail in this paragraph how these CDS were collected for each task.

For MPP, we used the Uniprot(uni 2023) ID mapping tool to map the Uniprot ID's associated with each protein, available from the TAPE benchmark (Rao et al. 2019), to the DNA sequence database of EMBL CDS (Kanz et al. 2005). Any retrieved CDS from EMBL whose translation did not match the original amino-acid sequences were filtered out.

In SSP, we used protein sequences with associated PDB ID's (Berman et al. 2000) from the dataset hosted by NetsurfP-3.0 (Høie et al. 2022). To collect the CDS we first used the RCSB 1D Coordinate Server (Berman et al. 2000) which assembles alignments between structure and sequence databases, to find alignments to protein sequences from the Uniprot database. Returned alignments to Uniprot were filtered out if there was not complete coverage. The

remaining Uniprot id's were then mapped to the sequence database EMBL CDS using the same process as for MPP described above.

For the beta-lactamase task, all sequences corresponded to the same gene. We obtained the TEM-1 reference gene as well as the mutations from supplementary material of ref. (Rocklin et al. 2017). This original fluorescence dataset contains many degenerate coding sequences. In PEER (Xu et al. 2022) labels were averaged over degenerate coding sequences in the original dataset. This process removes much data and does not allow us to study gLMs on degenerate sequences. Consequently, we propose two training datasets, sharing a single test set. The *Complete* set contains all CDS samples except those that are degenerate with respect to any CDS in the test set. The *Unique* set contains a random, maximal, subset of the non-degenerate coding sequences. This Unique set allows comparison between the gLMs and pLMs since all translated sequences are unique, while the Complete set demonstrates the impact of data availability on gLM performance.

For the stability prediction task, coding sequences were taken from supplementary material of the original experimental study (Rocklin et al. 2017). Since all CDS translate into unique amino-acids, we are able to match the dataset splits presented in TAPE (Rao et al. 2019).

Finally, for the fluorescence task we obtained the reference GFP gene, as well as its mutations from the reference of the original data (Sarkisyan et al. 2016). We chose to take the *Unique* subset as described above since the dataset was mostly non-degenerate.

Evaluation Methodology

The two pre-trained gLMs, DNABERT2 and NT-v2, and the two pre-trained pLMs, ESM1b and ESM2, were respectively evaluated with corresponding CDS and protein sequences as input and fine-tuned in similar conditions for a fair comparison. In opposition to all the other tasks that are regression tasks at the sequence level, the SSP task is a classification task at the amino-acid level. This is simply performed by pLMs by predicting for each amino-acid embedding a secondary structure from the 8 possible classes. For the Nucleotide Transformer, as tokens represent 6-mers, each token embedding is mapped to two classification predictions corresponding to the two amino-acids that the 6-mer represents. As DNABert2 uses Byte Pair Encoding to tokenize nucleotides sequences, we couldn't retrieve any systematic mapping from tokens to amino-acids and thus couldn't evaluate this model over the SSP task.

Fine-tuning of the models was done using IA^3 (Liu et al. 2022) parameter-efficient fine-tuning, along with a single-layer classification or regression head. IA^3 scales activations by a learnable vector, introducing a number of parameters approximately 0.1% of the total number of parameters. Models were fine-tuned with a batch size of 8. Adam optimizer was used with a learning rate of 0.003. Models were evaluated at fixed intervals over the validation set during training. Checkpoints with the highest R^2 for regression and lowest cross-entropy loss for classification over the validation set were saved and evaluated on the test set.

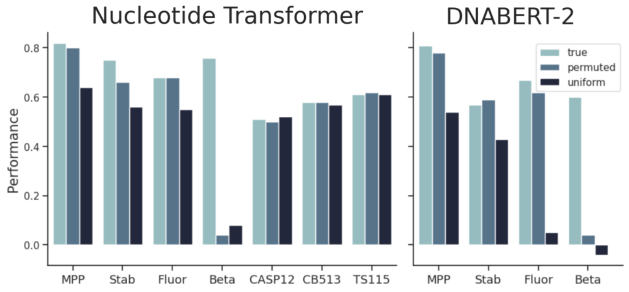

Figure 1: The impact of three codon sampling strategies on gLM performance over 5 tasks (CASP12, CB513 and TS115 are the different test sets for the SSP task). The strategies include uniformly sampling codons, permuting synonymous codons, and no sampling (true CDS). Performance is measured as Spearman correlation for Fluorescence, Beta-Lactamase, and Stability, R^2 for Melting Point, and accuracy for SSP classification task.

Impact of Codon Usage on Genomics Models

We initiated our study by evaluating the impact of having access to the true CDS sequence on genomics language models performance. To answer that question, we follow a procedure similar to the one presented in CaLM (Outeiral and Deane 2022). We fine-tune the genomics language models on all tasks, excluding SSP for DNABert2 as it couldn't be evaluated on that task, in three different settings: (1) on "true" curated CDS, (2) on sequences obtained by respecting codon frequencies from the true CDS but by permuting codons and (3) on sequences obtained by uniformly sampling codons. We report the obtained performance on the test sets of each task in Figure 1.

We observe that on most tasks, having access to the "true" CDS improves the performance over sequences obtained by sampling codons from their natural frequencies, thus justifying the need for our curated dataset. We also observe that randomly sampling codons yields degraded and close to zero performance on the Beta-Lactamase prediction task. Interestingly, we observe that Nucleotide Transformer v2 seems to be more robust than DNABERT to the codon distributions shift which might be explained by respectively the usage of 6-mers tokenization compared to BPE.

Genomic vs Protein Language Models

We compared the four aforementioned models over the five tasks and reported the performance in Table 2. First, we observe that the Nucleotide Transformer v2 matches or outperforms its DNABERT2 gLM counterpart on all the protein downstream tasks, confirming the recently published results on genomics downstream tasks (Dalla-Torre et al. 2023). Interestingly, we also observe that ESM2 and ESM1b seem to have comparable performance over these five tasks.

We observe that the Nucleotide Transformer matches the performance of its pLMs counterparts on the fluorescence prediction and stability prediction tasks. This suggests that despite the distribution shift between the raw genes seen during training by gLMs and the true CDS sequences, these

	Fluorescence (ρ)	Beta- Lactamase (ρ)		Stability (ρ)	Melting Point (R^2)	SSP (Acc)		
Train Dataset	All	Complete	Unique	All	All	All		
Test Dataset	All	All		All	All	CASP12	CB513	TS115
Nucleotide Transformer v2	0.68	0.79	0.76	0.75	0.82	0.51	0.58	0.61
DNABERT2	0.67	0.70	0.60	0.57	0.81	N/A	N/A	N/A
ESM2	0.68	0.89	0.89	0.74	0.72	0.63	0.76	0.76
ESM1-b	0.68	0.88	0.88	0.75	0.72	0.61	0.73	0.72

Table 2: Evaluation results of Nucleotide Transformer v2 500M, DNABERT2, ESM2 650M, and ESM1-b 650M on the test datasets of the proposed tasks. The metrics used to measure performance were chosen to match previous benchmarks and include Spearman correlation ρ, R^2, and accuracy, with a higher value indicating better performance for all metrics. We note, that protein models evaluated on *Complete* splits see identical proteins with differing labels, however we perform the evaluation for completeness. DNABERT2 was not evaluated on SSP since its BPE tokenization prevents residue-level predictions.

models are able to capture protein features to the same extent than protein models. However, the Nucleotide Transformer and DNABERT2 models underperform on the beta-lactamase activity prediction and SSP tasks. This might suggest that gLMs can capture global patterns in protein sequences but fail to capture finer-grain effects such as structure or the impact of single point mutations.

Finally, we observe that both the Nucleotide Transformer and DNABERT2 models outperform significantly ESM models on the melting point prediction tasks. We propose detailed analysis about this result in the next paragraph.

Melting Point Prediction Task Analysis

In this section we seek to understand whether gLMs performance on the MPP task can be attributed to a biological phenomenon regarding codon usage, or whether it is exploiting a "superficial" feature unique to CDS. Here we define superficial as information readily available that does not contribute to a better understanding of proteins.

Local vs Global Information. As per the impact of codon usage study reported in Figure 1, we observed that in the absence of codon usage information, the Nucleotide Transformer performance drop's below that of ESM, suggesting that the Nucleotide Transformer is utilizing codon frequencies. One indication that the Nucleotide Transformer might be exploiting superficial features of CDS would be if it can achieve the similar performance using only global sequence information. The motivation is that a biological phenomenon regarding codon usage would likely depend on their absolute and relative locations.

The GC-content Hypothesis. We then hypothesized that the Nucleotide Transformer may use GC-content to influence protein melting temperature prediction. The GC-content of a genomic sequence indicates the proportion of guanine (G) or cytosine (C) bases. G-C base pairs, featuring three hydrogen bonds, are more stable than A-T base pairs with two hydrogen bonds. Higher GC-content leads to higher melting temperatures in equal-length sequences. To test this, we augment ESM2 with the sequence GC-content information by appending the normalized GC-content to the embeddings before making the melting point prediction. Although this addition moderately improves performance with an increase in R^2 from 0.72 to 0.74, the model still lags behind Nucleotide Transformer. This suggests that Nucleotide Transformer's predictive ability goes beyond mere adjustment via GC-content.

The Species-level Conditioning Hypothesis. We then explored if the Nucleotide Transformer may exploit codon usage bias to condition on the species from which the sequence was derived. The MPP dataset consists of proteins from thirteen species ranging from unicellular E. coli, to mice and humans. These species have distinct melting point profiles and identifiable codon preferences. Codon bias across species is a well-documented phenomenon that reflects mutational and selective pressures, and is evident in the MPP dataset. To test this assumption, we augment both ESM and the Nucleotide Transformer with the species information of each sequence and evaluate test set performance. We append a one-hot species-identifying vector to the embeddings of each model. We find that augmenting ESM with species information increases performance from an R^2 value of 0.72 to 0.79, equal to that of NT on permuted codons. In contrast, augmenting Nucleotide Transformer with species only improves performance from R^2 of 0.81 to 0.83. This might suggest that Nucleotide Transformer achieves the majority of its advantage via condition on species information, which it learned during pre-training, and the remaining performance could come from local codon interactions.

Conclusion

After retrieving and curating CDS datasets for five protein downstream tasks of interest, we evaluated two pLMs and two gLMs over all tasks using a standardized fine-tuning strategy. After reporting evidences that true CDS are required for gLMs to obtain good performance, we observe that these models match and even outperform pLMs on 3 out of the 5 tasks while being strongly dominated on the 2 other tasks. This suggests that gLMs might be a good starting point to build unified foundational models for biology, but it leaves the door open to better understand how to improve these models on tasks such as SSP. We hope that the collection and release of the five CDS datasets will help the community to keep making progress in this field.

Acknowledgements

We would like to thank Liviu Copoiu and Patrick Bordes for helping us to obtain the CDS samples corresponding to the protein sequences for the MPP task.

References

2023. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1): D523–D531.

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; and Bourne, P. E. 2000. The protein data bank. *Nucleic acids research*, 28(1): 235–242.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Carranza, N. L.; Grzywaczewski, A. H.; Oteri, F.; Dallago, C.; Trop, E.; Sirelkhatim, H.; Richard, G.; et al. 2023. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv*, 2023–01.

Dallago, C.; Mou, J.; Johnston, K. E.; Wittmann, B. J.; Bhattacharya, N.; Goldman, S.; Madani, A.; and Yang, K. K. 2021. FLIP: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10): 7112–7127.

Firnberg, E.; Labonte, J. W.; Gray, J. J.; and Ostermeier, M. 2014. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution*, 31(6): 1581–1592.

Høie, M. H.; Kiehl, E. N.; Petersen, B.; Nielsen, M.; Winther, O.; Nielsen, H.; Hallgren, J.; and Marcatili, P. 2022. NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic acids research*, 50(W1): W510–W515.

Ji, Y.; Zhou, Z.; Liu, H.; and Davuluri, R. V. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15): 2112–2120.

Kanz, C.; Aldebert, P.; Althorpe, N.; Baker, W.; Baldwin, A.; Bates, K.; Browne, P.; van den Broek, A.; Castro, M.; Cochrane, G.; et al. 2005. The EMBL nucleotide sequence database. *Nucleic acids research*, 33(suppl_1): D29–D33.

Li, S.; Moayedpour, S.; Li, R.; Bailey, M.; Riahi, S.; Miladi, M.; Miner, J.; Zheng, D.; Wang, J.; Balsubramani, A.; Tran, K.; Zacharia, M.; Wu, M.; Gu, X.; Clinton, R.; Asquith, C.; Skalesk, J.; Boeglin, L.; Chivukula, S.; Dias, A.; Montoya, F. U.; Agarwal, V.; Bar-Joseph, Z.; and Jager, S. 2023. CodonBERT: Large Language Models for mRNA design and optimization. *bioRxiv*.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022: 500902.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130.

Liu, H.; Tam, D.; Muqeeth, M.; Mohta, J.; Huang, T.; Bansal, M.; and Raffel, C. A. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35: 1950–1965.

Liu, Y. 2020. A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Commun Signal*, 18(1): 145.

Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A.; Birch-Sykes, C.; Wornow, M.; Patel, A.; Rabideau, C.; Massaroli, S.; Bengio, Y.; et al. 2023. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint arXiv:2306.15794*.

Outeiral, C.; and Deane, C. 2022. Codon language embeddings provide strong signals for protein engineering. *bioRxiv*, 2022–12.

Pinney, M. M.; Mokhtari, D. A.; Akiva, E.; Yabukarski, F.; Sanchez, D. M.; Liang, R.; Doukov, T.; Martinez, T. J.; Babbitt, P. C.; and Herschlag, D. 2021. Parallel molecular mechanisms for enzyme temperature adaptation. *Science*, 371(6533).

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, X.; Canny, J.; Abbeel, P.; and Song, Y. S. 2019. Evaluating Protein Transfer Learning with TAPE. arXiv:1906.08230.

Rocklin, G. J.; Chidyausiku, T. M.; Goreshnik, I.; Ford, A.; Houliston, S.; Lemak, A.; Carter, L.; Ravichandran, R.; Mulligan, V. K.; Chevalier, A.; Arrowsmith, C. H.; and Baker, D. 2017. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347): 168–175.

Sarkisyan, K.; Bolotin, D.; Meer, M.; Usmanova, D.; Mishin, A.; Sharonov, G.; Ivankov, D.; Bozhanova, N.; Baranov, M.; Soylemez, O.; Bogatyreva, N.; Vlasov, P.; Egorov, E.; Logacheva, M.; Kondrashov, A.; Chudakov, D.; Putintseva, E.; Mamedov, I.; Tawfik, D.; Lukyanov, K.; and

Kondrashov, F. 2016. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603): 397–401.

Xu, M.; Zhang, Z.; Lu, J.; Zhu, Z.; Zhang, Y.; Ma, C.; Liu, R.; and Tang, J. 2022. PEER: A Comprehensive and Multi-Task Benchmark for Protein Sequence Understanding. arXiv:2206.02096.

Zhou, Z.; Ji, Y.; Li, W.; Dutta, P.; Davuluri, R.; and Liu, H. 2023. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. arXiv:2306.15006.