ARE GENOMIC LANGUAGE MODELS ALL YOU NEED? EXPLORING GENOMIC LANGUAGE MODELS ON PROTEIN DOWNSTREAM TASKS

Sam Boshar
Massachusetts Institute of Technology, InstaDeep
sboshar@mit.edu, s.boshar@instadeep.com

Evan Trop, Bernado Almeida, Thomas Pierrot*
InstaDeep
{e.trop, b.dealmeida, 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 (gLMs) hold the potential to make downstream predictions not only about DNA sequences but also about proteins. However, the performance of gLMs on protein tasks remains unknown, mostly due to the lack of evaluation tasks with paired proteins and their true coding DNA sequences (CDS) that can be processed by gLMs. In this work, we curated five such datasets and use them to evaluate the performance of multiple state-of-the-art genomic and proteomic language models (pLMs). We found that, despite their pre-training on largely non-coding sequences, gLMs are competitive and even outperform their pLMs counterparts on some tasks. The best performance was achieved using the retrieved "true" CDS compared to alternative sampling strategies. The application of gLMs to proteomics offers the potential to leverage rich CDS data, and in the spirit of the central dogma, the possibility of a unified and synergistic approach to genomics and proteomics.

1 INTRODUCTION

Large language models (LLMs), have revolutionized the field of Natural Language Processing (NLP) thanks to their capability to learn through self-supervision from unlabeled data (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 nature of biological sequences, such as proteins, DNA and RNA, paired with the abundance of unlabeled data obtained through high-throughput sequencing, make it a perfect application for these methods to thrive. This effort started first in proteomics, where several works showed that training large Transformer models to recover masked amino-acids in protein sequences leads to powerful representations that can then be used to solve diverse downstream tasks with state-of-the-art performance (Lin et al., 2022; Elnaggar et al., 2021; Jumper et al., 2021; Lin et al., 2023). More recently, similar models were developed for genomics and trained over the human reference genome as well as hundreds of reference genomes from different species to recover masked consecutive nucleotides in chunks (Dalla-Torre et al., 2023; Zhou et al., 2023; Ji et al., 2021; Nguyen et al., 2023; Benegas et al., 2022; Nguyen et al., 2024). 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 down-

*Corresponding Author

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.

stream tasks with improved performance,including predicting diverse DNA molecular phenotypes related to splicing, 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, based on codons, was recently introduced (Outeiral & Deane, 2022; Li et al., 2023; Hallee et al., 2023). These models were trained on large datasets made of coding sequences (CDS) by reconstructing masked codons - instead of masked amino-acids. Notably, the Codon Adaptation Language Model (CaLM) showed that cLMs can outperform their amino-acid based counterparts on several downstream tasks of interest such as species recognition, prediction of protein and transcript abundance or melting point estimation when controlling for model size (Outeiral & Deane, 2022). This improved performance seems to be related to the ability of codon-based language models (cLMs) to capture patterns of codon usage across DNA sequences.

Inspired by these recent results, we aim to study to what extent genomic language models (gLMs) can be used as a general unified approach to solve tasks in both domains - genomics and proteomics (Supplementary Fig. 1). In opposition to cLMs, gLMs have been trained over full raw genomes and as such can be used to analyze non-coding regions as well as full genes including exons and intronic regions. While this makes gLMs widely capable for genomics tasks, their capacity to solve protein tasks from their corresponding CDS has not been explored. Since they have never seen "true" CDS per se during training, as exons are always separated by introns in eukaryotic species genomes, and coding sequences represent on average only $\sim 1.5\%$ of the human genome data used for training (Lander, 2011), it is unclear to what extent these models can be competitive with protein language models (pLMs).

In this paper, we present a comprehensive analysis of gLMs applied to protein-related tasks. We established a benchmark of five common protein analysis tasks and curated CDS sequences for a fair comparison between pLMs and gLMs. Our evaluation of two state-of-the-art pLMs and gLMs revealed that gLMs outperformed or matched pLMs on three out of five tasks, while underperforming on the remaining two. Notably, careful curation of matched CDS sequences was crucial for optimal gLM performance. The two tasks where pLMs significantly outperformed gLMs required sensitivity to codon-level changes. Intriguingly, gLMs significantly outperformed pLMs in predicting protein melting points – a trend also observed with cLMs. Further investigation revealed that gLMs achieve this by leveraging GC-content and species information from nucleotide sequences, features not captured by protein-based models.

2 Protein Downstream Tasks

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 et

al. (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 Rao et al. (2019). 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.

3 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 while 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 degenerate sequence 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 sequences translate into unique amino-acids, we are able to match the dataset splits presented in TAPE (Rao et al., 2019).

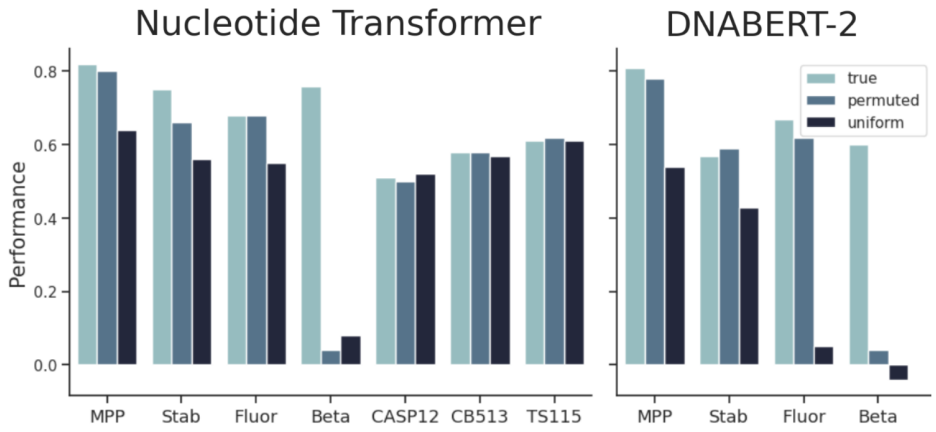


Figure 1: The impact of three codon sampling strategies on the performance of NT-v2 and DNABERT2 over 5 tasks (CASP12, CB513 and TS115 are the different test sets for the SSP task). The strategies include uniformly sampling codons, permuting synonymous codons, and no sampling (true CDS). Performance is measured as Spearman correlation for Fluorescence, Beta-Lactamase, and Stability, R^2 for Melting Point, and accuracy for SSP classification task.

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.

4 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.

5 Impact of Codon Usage on Genomics Models

We initiated our study by evaluating the impact of having access to the true CDS sequence on genomic language models performance. To answer that question, we follow a procedure similar to the one presented in CaLM (Outeiral & 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.

6 Genomic vs Protein Language Models

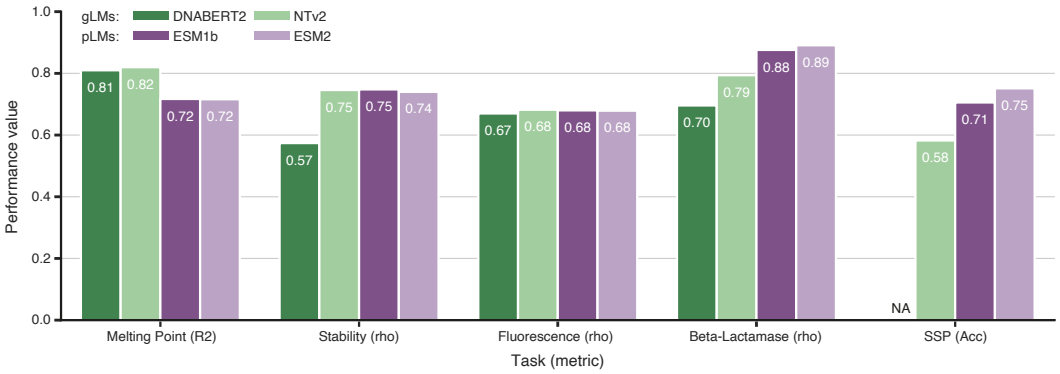


Figure 2: **Genomic Language Models are Competitive on Protein Tasks**. Evaluation results of Nucleotide Transformer v2 500M, DNABERT2, ESM2 650M, and ESM1-b 650M on the test datasets of the proposed tasks. The metrics used to measure performance were chosen to match previous benchmarks and include Spearman correlation ρ, R^2, and accuracy, with a higher value indicating better performance for all metrics. Notably NTv2, matches or supersedes pLMs on 3 of the 5 tasks.

We compared the four aforementioned models over the five tasks and reported the performance in Figure 2 (see also Supplementary Table 1). 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 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 section.

7 Melting Point Prediction Task Analysis

We showed that gLMs outperforms significantly their pLM counterpart on the melting point prediction task. A similar behavior has been reported for cLMs (Outeiral & Deane, 2022). This motivated us to analyse the disparity between gLMs and pLMs performance on this task. In particular, we explored whether the superior performance of gLMs can be attributed to a biological phenomenon such as codon usage, or whether it is exploiting a “superficial” feature unique to CDS data. Here we define superficial as information readily available that does not contribute to a better understanding of proteins.

In investigating impact of codon usage reported in Figure 1, we found that in the absence of codon usage information the NT-v2 performance drops below that of ESM, the gLM achieving an accuracy of only 0.64 compared to the pLM's 0.72. This result suggests that evaluated on the true CDS NT-v2 is utilizing codon frequencies. We further explored if the improved performance on the melting point prediction task could be related to additional sequence features. One indication that the NT-v2 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. To test this we developed two hypotheses around the use of global sequence information.

7.1 The GC-content Hypothesis.

We hypothesized that the NT-v2 may use GC-content to influence protein melting point 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 hypothesis, we augmented both ESM-2 and NT-v2 with the sequence's 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 NT-v2 (Supplementary Fig. 2a). Augmenting NT-v2 with the same information does not lead to any increase in performance. This suggests that NT-v2 already has access to GC-content information.

7.2 The Species-level Conditioning Hypothesis.

We next explored if the NT-v2 may exploit codon usage bias to condition on the species the sequence was derived from. The melting point prediction dataset consists of proteins from thirteen different species ranging from unicellular E. coli, to mice and humans. Proteins of different species have distinct melting point profiles and identifiable codon preferences (Supplementary Fig. 2b). To test this hypothesis, first we verify that gLM can better identify species from sequence. We finetuned NT-v2 and ESM-2 on the task of species identification and found that NT achieves an accuracy of 0.95 while ESM-2 achieves an accuracy of only 0.81 (Supplementary Fig. 2d,e). Additionally, we showed that the t-SNE for pretrained embeddings of models reveal that gLM embeddings are strongly structured by species while pLM are not (Supplementary Fig. 2c).

To test whether species information may account for the difference in performance we augmented both ESM-2 and NT-v2 with the species information of each sequence and evaluate test set performance. This augmentation was done by appending a one-hot species-identifying vector to the embeddings of each model. We find that augmenting ESM-2 with species information increases performance from an R^2 value of 0.72 to 0.79 (Supplementary Fig. 2a). This closes most of the gap with the NT-v2 trained from curated CDS and brings the model to the performance of NT-v2 trained with permutated codons (no local information) which has an R^2 of 0.80. In contrast, augmenting NT-v2 with species does not result in an improvement in performance, suggesting that NT-v2 achieves the majority of its advantage via conditioning on species information, which it learned during pre-training.

Using these findings we finally tested if augmenting ESM-2 with both global attributes (GC-content and species) could recover the performance of NT-v2. We found ESM-2 achieved a similar performance of 0.79, while, once more, NT-v2 showed no change in performance. Our results demonstrate that although there are additive benefits for ESM-2 from having both features, there still exists a gap in performance with NT-v2 (Supplementary Fig. 2a). We presume this advantage is coming from local codon interaction information present in the coding sequences.

8 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 evidence 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 obtaining lower performance

on the remaining 2 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.

ACKNOWLEDGMENTS

We would like to thank Liviu Copoiu and Patrick Bordes for help on assembling the CDS dataset for the MPP task.

REFERENCES

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

Gonzalo Benegas, Sanjit Singh Batra, and Yun S Song. Dna language models are powerful zero-shot predictors of non-coding variant effects. *BioRxiv*, pp. 2022–08, 2022.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Hassan Sirelkhatim, Guillaume Richard, 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*, 2021. doi: 10.1101/2021.11.09.467890. URL https://www.biorxiv.org/content/early/2021/11/11/2021.11.09.467890.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

Elad Firnberg, Jason W. Labonte, Jeffrey J. Gray, and Marc Ostermeier. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution*, 31(6):1581–1592, 02 2014. ISSN 0737-4038. doi: 10.1093/molbev/msu081. URL https://doi.org/10.1093/molbev/msu081.

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

Magnus Haraldson Høie, Erik Nicolas Kiehl, Bent Petersen, Morten Nielsen, Ole Winther, Henrik Nielsen, Jeppe Hallgren, and Paolo Marcatili. 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, 2022.

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.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Carola Kanz, Philippe Aldebert, Nicola Althorpe, Wendy Baker, Alastair Baldwin, Kirsty Bates, Paul Browne, Alexandra van den Broek, Matias Castro, Guy Cochrane, et al. The embl nucleotide sequence database. *Nucleic acids research*, 33(suppl_1):D29–D33, 2005.

E. Lander. Initial impact of the sequencing of the human genome. *Nature*, 470:187–197, 2011. doi: 10.1038/nature09792. URL https://doi.org/10.1038/nature09792.

Sizhen Li, Saeed Moayedpour, Ruijiang Li, Michael Bailey, Saleh Riahi, Milad Miladi, Jacob Miner, Dinghai Zheng, Jun Wang, Akshay Balsubramani, Khang Tran, Minnie Zacharia, Monica Wu, Xiaobo Gu, Ryan Clinton, Carla Asquith, Joseph Skalesk, Lianne Boeglin, Sudha Chivukula, Anusha Dias, Fernando Ulloa Montoya, Vikram Agarwal, Ziv Bar-Joseph, and Sven Jager. Codonbert: Large language models for mrna design and optimization. *bioRxiv*, 2023. doi: 10.1101/2023.09.09.556981. URL https://www.biorxiv.org/content/early/2023/09/12/2023.09.09.556981.

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.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

Y. Liu. A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Commun Signal*, 18(1):145, Sep 2020. doi: 10.1186/s12964-020-00642-6.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint arXiv:2306.15794*, 2023.

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.

Carlos Outeiral and Charlotte Deane. Codon language embeddings provide strong signals for protein engineering. *bioRxiv*, pp. 2022–12, 2022.

Margaux M Pinney, Daniel A Mokhtari, Eyal Akiva, Filip Yabukarski, David M Sanchez, Ruibin Liang, Tzanko Doukov, Todd J Martinez, Patricia C Babbitt, and Daniel Herschlag. Parallel molecular mechanisms for enzyme temperature adaptation. *Science*, 371(6533), 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with tape, 2019.

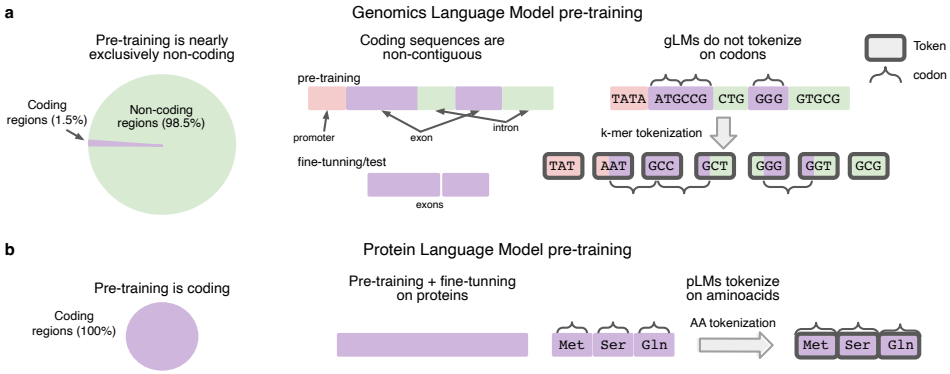
Gabriel J. Rocklin, Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K. Mulligan, Aaron Chevalier, Cheryl H. Arrowsmith, and David Baker. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, 2017. doi: 10.1126/science.aan0693. URL https://www.science.org/doi/abs/10.1126/science.aan0693.

KS Sarkisyan, DA Bolotin, MV Meer, DR Usmanova, AS Mishin, GV Sharonov, DN Ivankov, NG Bozhanova, MS Baranov, O Soylemez, NS Bogatyreva, PK Vlasov, ES Egorov, MD Logacheva, AS Kondrashov, DM Chudakov, EV Putintseva, IZ Mamedov, DS Tawfik, KA Lukyanov, and FA Kondrashov. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, May 2016. doi: 10.1038/nature17995.

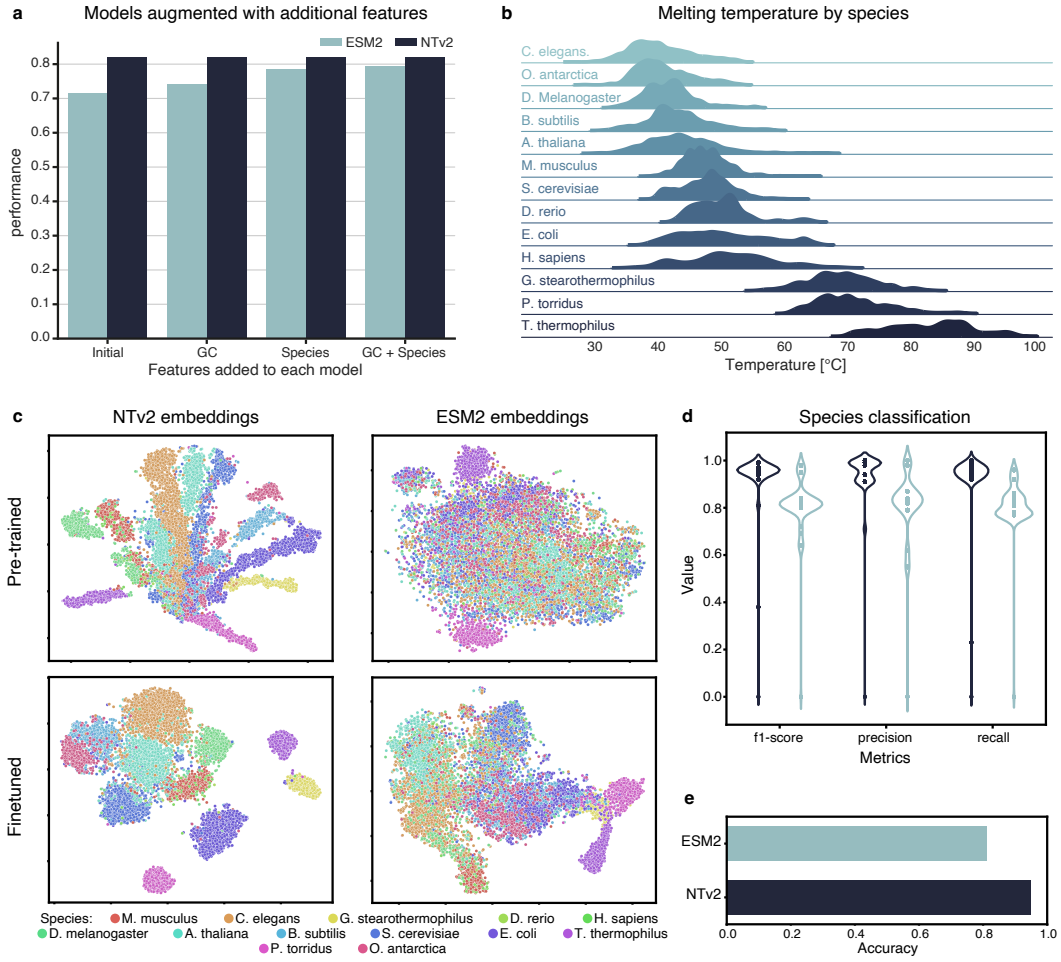
Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. Peer: A comprehensive and multi-task benchmark for protein sequence understanding, 2022.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome, 2023.

A APPENDIX



Supplementary Figure 1: **Differences between genomic and protein language model pre-training**. We outline key differences between **(a)** gLM and **(b)** pLM pre-training that make the task of building robust protein representations more difficult for gLMs. Unlike pLMs and cLMs, gLM pre-training is predominantly ($\sim 99\%$) on non-coding regions of the genome, the vast majority of which (barring prokaryotic genomes) are non-contiguous, while fine-tuning and inference are carried out with contiguous CDS. Additionally, coding regions during pre-training are not tokenized on codons, making amino acid representations non-trivial.



Supplementary Figure 2: **Genomic language models use species codon-usage bias to outperform protein language models on melting point prediction**. **a)** The results of appending combinations of GC-content and Species to NT-v2 and ESM-2 embeddings during fine-tuning on the melting point prediction task. We find that species information accounts for the majority of the disparity of performance between ESM-2 and NT-v2. We also augment NT with the same information but see no change in performance indicating NT-v2 already has access to this information. **b)** The distribution of melting points for each species in the dataset show distinct profiles. **c)** Dimensionality reduction via t-SNE of the pre-trained and fine-tuned NT-v2 and ESM-2 models demonstrates that the gLM captures the structure of species information to a greater degree than pLM and initially acquired this knowledge from its pre-training. **d)** We train ESM-2 and NT-v2 models to predict the species from sequence via fine-tuning with a single layer classification head. We plot the f1-score, precision and recall across species. **e)** Bar plot for the species classification accuracy weighted by the number of sequences for each species. Results from both d and e confirm that NT is superior at identifying species.

	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
Joint (ESM + NT)	0.68	0.89	0.89	0.83	0.82	N/A	N/A	N/A

Supplementary Table 1: Evaluation results of Nucleotide Transformer v2 (500M), DNABERT2, ESM2 (650M), ESM1-b (650M), and joint ESM-NTv2 models on the test sets of the different tasks. The metrics used to measure performance in each task 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 different labels, but we still performed the evaluation for completeness. DNABERT2 was not evaluated on SSP since its BPE tokenization prevents residue-level predictions.