Extended Abstract

Motivation Language models trained on DNA sequences, or so-called genomic language models (gLMs), have been shown to achieve robust performance on biological prediction and generation tasks, in particular recently developed models like Evo 2 [1] and GPN [2]. When using these models for generative design tasks, it is important to ensure confidence that the pathogenic potential of novel designs is well known by researchers using the models. These models may be prompted with a sequence and asked to assign it a biological plausibility or likelihood which may be used as a proxy for pathogenicity. Aligning models using reinforcement learning by training them to assign a lower likelihood for pathogenic sequences should increase biological safety of generations.

Method Direct Preference Optimization (DPO) was utilized to encourage the model to prefer benign variants over pathogenic ones. To do so, a dataset of genes was required. ClinVar [3] was used to obtain annotated data for over 1600 genes. Data was stored by gene, including the promoter region, pathogenic variant, benign variant, and wild type variant sequences. We use the standard DPO loss with the aforementioned completions. Note that we reformulate DPO slightly for the case of MLM, in which logits are computed with bidirectional context around the mutated single nucleotide polymorphism (SNP). We define loss only over the mutated nucleotide to avoid washing out training signal. For method comparison, pretrained GPN as well as GPN that was fine-tuned with supervised finetuning (SFT) were included for comparison. Models were then assessed at their accuracy in assigning lower likelihood to pathogenic genes versus their benign counterparts for 300 held out genes.

Implementation ClinVar was accessed using the `biopython` package that has the capability to query various National Center for Biotechnology Information (NCBI) databases. A list of genes was passed into a script that read each gene and queried the ClinVar SNP database with added search terms "pathogenic" and "benign" based on the type of SNP being acquired. The wild type sequence was then obtained by querying the "nucleotide" NCBI database for the corresponding whole chromosome version that the SNP referenced, storing it locally for future use, and splicing it at the indices of the gene. For training, a custom pytorch training loop was written in conjunction with the GPN model and tokenizer hosted on Hugging Face. A custom pytorch dataloader was written for the SFT and DPO settings.

Results GPN untrained performed the best at the task with respect to having the highest accuracy compared to both SFT and DPO-trained models. DPO performed the best when comparing the sum of differences between preferred and dispreferred variant pairs, meaning that the distance between the likelihood given for one sequence versus its partner (preferred and dispreferred) was much greater during DPO evaluation than for the other two models. Evaluation on the training data indicates that the model did not overfit, where comparable accuracy was observed across train and test data. We identify 29 cases of genes which are correctly classified by the DPO model but incorrectly classified by the base model.

Discussion DPO successfully encouraged larger separation between the calibrated log probability of preferred and dispreferred sequences, suggesting that even though it is not the most accurate model, it is the most confident. A reasonable explanation for why the SFT and DPO model performed worse than the base model on the accuracy metric is the fact that ClinVar data is known to be inconsistent in quality and collected from a large number of studies. It is also biased to variants that are not too harmful, given that they must propagate in at least 1% of the population.

Conclusion It is evident that biological foundation models are able to learn much about pathogenicity from evolutionary data. It is not clear that DPO-style finetuning can improve variant effect prediction, and an actionable next step to move forward is to acquire a less-biased, higher quality data set. Experiments should also be repeated with a larger training set to observe scaling performance during training. If a promising model is produced, experimental validation would be necessary to test novel mutations predicted by the base model and finetuned model.

Improving biological safety of genomic language models via direct preference optimization

Alejandro Buendia
Department of Biomedical Data Science
Stanford University
abuen@stanford.edu

Mohini Misra
Department of Bioengineering
Stanford University
mmisra@stanford.edu

Samantha Mutiti
Department of Bioengineering
Stanford University
samutiti@stanford.edu

Abstract

Language models trained on DNA sequences have been shown to achieve robust performance on biological prediction and generation tasks, with applications to genetic risk assessment, therapeutic engineering, and synthetic biology [4, 2, 1, 5]. Genomic language models have been applied to the task of variant effect prediction, in which model likelihoods serve as a proxy for the model's inferred fitness for a genetic variant. In this paper, we seek to align a genomic language model with mutation pathogenicity data by using direct preference optimization (DPO) on clinical variants. We leverage clinically annotated data for single nucleotide polymorphisms (SNPs) from ClinVar [3] in a reinforcement learning framework. We hypothesize that aligning gLMs toward benign variants and away from pathogenic variants will enable them to make more accurate predictions for unknown variants and improve their biological safety when used downstream for de novo biological generation. After comparison on untrained, SFT-trained, and DPO-trained models we see the untrained model performed the at the highest accuracy compared to both SFT and DPO-trained models. DPO, however, introduced greater spread in the data. Evaluation on training data indicated low likelihood of overfitting which indicates that training data was not robust or representative enough for successful finetuning to this task.

1 Introduction

The genomes of organisms encode complex and diverse functions which are realized through the central dogma of transforming DNA to RNA to protein. DNA is the central "code" that governs life and gives rise to sophisticated behavior. Despite increasingly advanced biological research, to date there still remains much that is unknown about the information DNA holds and consequently how to engineer that information productively. To address this gap, there has been an effort to train language models on vast corpora of genomic DNA sequences (so called genomic language models, or gLMs). These models have been shown to achieve robust performance on biological generation and prediction tasks. For example, the recently developed model Evo 2 is capable of sequence design at the scale of whole human mitochondrial genomes, minimal bacterial genomes, and yeast chromosomes, as well as knowledge-based tasks like mutation effect prediction [1].

However, as is common for many foundation models, gLMs perform well across a range of generalized tasks but suffer a performance gap for specialized domain-specific tasks. For example, many Evo

2 generations contain unrealistic guanine-cytosine (GC) content, sequence complexity, or low-confidence translated protein structures, requiring manual filtering by researchers before use in the wet lab. Similarly, tasks like mutational prediction effect can benefit from specialized training to teach the model what makes a mutation more or less harmful beyond “blind" knowledge acquired from unlabeled patterns in the training data.

We focus on the task of pathogenicity prediction. Some mutations in genes are benign (i.e. do not harm an organism) and others are pathogenic (i.e. cause an undesirable effect on the host). Ideally, a model would be able to discriminate between pathogenic and benign variants robustly for two main reasons: (1) this would allow researchers to make more confident predictions about variants of unknown significance (VUS); and (2) when these genomic language models are used for generation, the generated sequences align more with biological safety. Current models do moderately well at variant effect prediction, but there is still room for improvement. We explore reinforcement learning methods to finetune a variant effect predictor for greater accuracy.

2 Related Work

Biological foundation models have borrowed techniques from natural language processing to pretrain large transformer-based architectures on biological sequences. Groundwork in this field was laid by protein language models (pLMs) such as ProteinBERT [6], ProGen [7], ESM [8], and ESM-2 [9]. Through self-supervised pretraining, pLMs learn representations of amino acid sequences that can be used downstream for structure and residue contact prediction as well as generative tasks such as controllable sequence design. Recent models trained on DNA sequences have shown that both masked language modeling and autoregressive pretraining strategies are effective in learning representations that can recapitulate known biology and generate de novo DNA sequences. Masked language models include DNABERT [10], Genomic Pretrained Network [4], and Nucleotide Transformer [5], while autoregressive models include HyenaDNA [11], Evo [11], and Evo 2 [1]. Evo 2 learns from sequence alone for state-of-the-art prediction performance of noncoding pathogenic mutations and clinically significant BRCA1 variants without task-specific finetuning. Supervised methods such as Enformer [12] and Borzoi [13] use a transformer architecture to train a supervised function from the genome to quantitative output tracks which measure accessibility or gene expression level (e.g. ChIP-seq, DNase-seq, ATAC-seq, CAGE).

Some recent approaches apply reinforcement learning to biological foundation models for a variety of downstream tasks. For example, Yang et al. finetunes HyenaDNA on regulatory DNA sequences for designing cis-regulatory elements [14], in which an MDP is formulated with reward given by a combination of fitness and whether a transcription factor binding site appears in the generation. ProteinRL finetunes pLMs to achieve specific desired properties [15]. ProteinDPO uses direct preference optimization to align the structure-conditioned pLM ESM-IF to experimental fitness data [16], while RLXF similarly uses supervised finetuning and proximal policy optimization to align ESM-2 to experimental fluorescence data [17].

Variant effect prediction methods span a long history and include both traditional approaches and recent deep learning-based approaches. SIFT [18], PolyPhen-2 [19], and CADD [20] use biological heuristics to estimate pathogenicity of variants. Deep-learning based approaches such as AlphaMissense [21] and EVE [22] use self-supervised pretraining to learn pathogenic amino acid substitutions. Benegas et al. present TraitGym, two non-coding variant datasets, to benchmark the performance of gLMs on a binary classification between putatively causal and non-causal genetic variants [2]. Simpler alignment-based methods such as CADD and GPN-MSA [23] are shown to excel at prediction for Mendelian traits and complex disease traits and outperform large architectures such as Evo 2. To date, no work exists for aligning genomic or protein language models with reinforcement learning to increase performance on variant effect prediction, indicating that post-training on variants may benefit large-scale architectures.

3 Method

We present a reinforcement learning framework for aligning the Genomic Pretrained Network (GPN) [4] to clinical variant data. GPN uses a standard masked language modeling task during pretraining, in which a set M of nucleotides is masked during minimization of loss

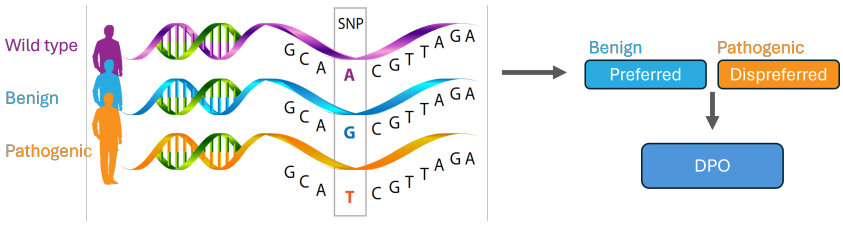


Figure 1: Benign and pathogenic SNPs are used as preferred and dispreferred completions in the finetuning workflow.

$$\mathcal{L}_{\text{MLM}}(\theta) = -\frac{1}{|M|} \sum_{i \in M} \log p_\theta(x_i \mid \tilde{\mathbf{x}}),$$

where $\tilde{\mathbf{x}}$ is the masked sequence and x_i is the i-th nucleotide position. We present two finetuning strategies to continue training the base model.

3.1 Supervised Finetuning

Supervised finetuning (SFT) serves as a baseline for the method. We finetune GPN using the cross entropy loss between the wild type sequence masked at the mutation and the wild type sequence with the benign SNP. Specifically, given wild type sequence x, we replace mutated position $i \in \{1, \ldots, L\}$ with the mask token to create $\tilde{\mathbf{x}}$. The supervised loss is formulated as

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \log p_\theta(x_i^{\text{benign}} \mid \tilde{\mathbf{x}}),$$

where x_i^{benign} is the benign SNP at nucleotide position i and $\tilde{\mathbf{x}}$ is the wild type sequence with position i masked. During training, we define loss over the mutated residues.

3.2 Direct Preference Optimization

We follow the loss formulation from [24] to finetune GPN with direct preference optimization (DPO):

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid \tilde{\mathbf{x}})}{\pi_{\text{ref}}(y_w \mid \tilde{\mathbf{x}})} - \beta \log \frac{\pi_\theta(y_l \mid \tilde{\mathbf{x}})}{\pi_{\text{ref}}(y_l \mid \tilde{\mathbf{x}})} \right) \right].$$

π_θ is the trainable policy and π_{ref} is the reference policy, i.e. the frozen GPN network. Here we define $\tilde{\mathbf{x}}$ to be the wild type sequence x masked at the mutation position. y_w is the preferred completion while y_l is the dispreferred completion. In this case, we adapt the DPO loss for use with the masked language model by feeding in the masked sequence as context to the model. y_w is defined to be the wild type sequence with the benign SNP substituted in, while y_l is the wild type sequence with the pathogenic SNP substituted in. β is the standard inverse temperature parameter controlling reliance on the reference policy. During training, we define loss over the mutated residues.

4 Experimental Setup

4.1 Dataset

Our finetuning dataset was sourced from ClinVar [3], an NIH-hosted database of gene variants and their impact on humans. We focused on coding regions (regions of DNA that translate to protein sequences, as opposed to “intergenic" regions, which are structural and regulatory regions of DNA, but not necessarily protein-encoding) from ClinVar. We pulled 1620 NCBI gene sequences and partitioned 60% for training (1056 genes), 20% for validation (264), and just under 20% for testing (300). Each gene we use has a single nucleotide polymorphism (SNP) classified as pathogenic or likely pathogenic. Considering that the base model is trained with MLM bidirectionally, we take a sequence of length 1536 (i.e. the context window for GPN) centered around the pathogenic SNP. Our

pytorch data loader returns the wild type sequence (without the pathogenic SNP) as the “preferred" completion and the wild type sequence with the pathogenic SNP substituted in as the “dispreferred" completion to be used in DPO. If a benign SNP exists at the same position, we take the preferred completion to be the wild type sequence with the benign SNP substituted in (Figure 1).

4.2 Training

We train the SFT and DPO models each for 50 epochs with AdamW optimizer and learning rate 1e-5. For DPO, we use the standard temperature setting $\beta = 0.1$. We define loss only over the mutated nucleotide to avoid washing out training signal for both settings. We show training and validation loss perplexity over 50 epochs (Figure 2) for the DPO setting. Note that a perplexity of 1 corresponds to perfect model performance, while a perplexity of 4 indicates near-random performance over the nucleotide vocab set.

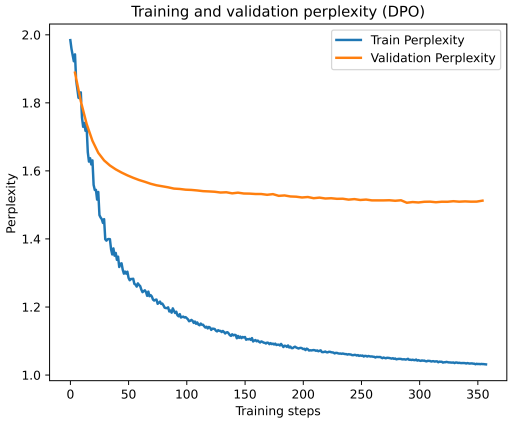


Figure 2: Training and validation perplexity curves for DPO finetuning.

4.3 Evaluation

We used the model’s predicted likelihood of a sequence as a proxy to gauge how pathogenic the model believes it to be. This makes sense biologically, because a model trained on evolutionary genomic data would have seen pathogenic variants less often (due to competitive selection) and thus consider them less likely. Here we evaluate on the 300 genes in the held-out set to check whether likelihood decreases in a dispreferred variant, with respect to the preferred variant. This can be easily checked by subtracting the likelihood of the preferred variant from the likelihood of the dispreferred variant. Positive results are desired and negatives ones indicate where the model is performing weakly.

5 Results

5.1 Quantitative Evaluation

We evaluated the base GPN, SFT, and DPO models on the likelihood assignment task described in the previous section. Initially surprisingly, GPN out of the box performed the best at the task with respect to having the highest accuracy (percent of genes where the preferred SNP had higher likelihood when compared to their dispreferred counterpart) compared to both SFT and DPO-trained models (Figure 3), indicating a strong baseline. This agrees with findings from the TraitGym benchmark [2].

This does not mean that our finetuning efforts were in vain. Another aspect of the results is *how large* the difference between the calibrated log likelihood for preferred versus dispreferred SNPs is. A model that discriminates the two variants with a high difference in likelihood can be thought of as more confident model, regardless of the accuracy, once the likelihood is calibrated to a common scale. To calculate this model confidence score, we compute:

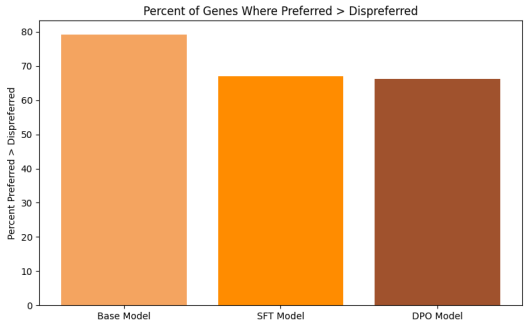


Figure 3: The accuracy of the three training modalities

$$\sum_i (l(v_{\text{preferred}}) - l(v_{\text{dispreferred}}))$$

where $l(x)$ is the log-likelihood function and v is a variant specified by its subscript. Notably, we penalize this model score for differences that are in the wrong direction (negative, meaning $l(v_{\text{dispreferred}}) > l(v_{\text{preferred}})$), so this model score is representative of confidence in the correct direction.

The results of this calculation is in Figure 4. The DPO model does the best at separating preferred and dispreferred variants overall. This indicates that the DPO model contributes some discriminatory value that may not be apparent when just looking at the accuracies. Notably, the base model does the worst by this metric, which makes sense because it is not specifically trained to discriminate between variants.

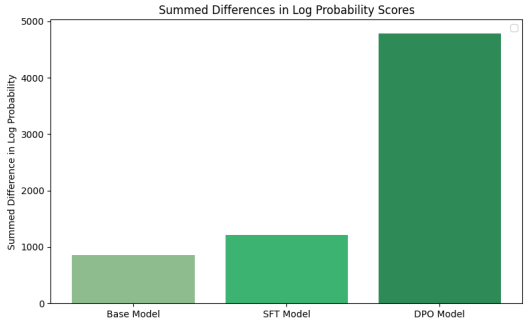


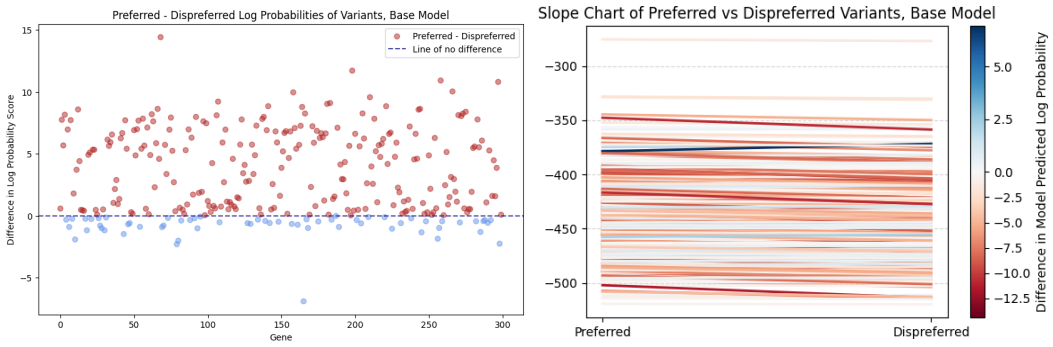
Figure 4: The accuracy of the three training modalities using the proposed model confidence score.

5.2 Qualitative Analysis

Visualizations of individual data points can be found below. Each point represents the result of

$$l(v_{\text{preferred}}) - l(v_{\text{dispreferred}}).$$

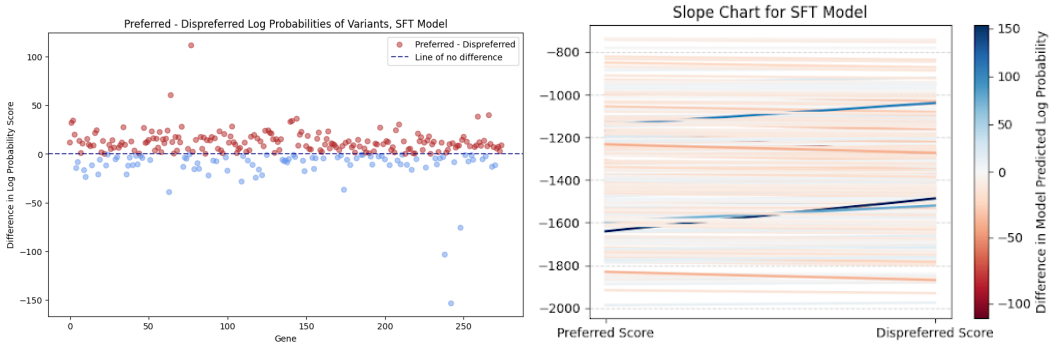
Red markers in the scatter plot and red descending trajectories in the slope chart represent success at the task (a positive difference between preferred and dispreferred likelihoods). Higher markers on the y-axis of the scatter plot and darker red and steeper slopes on the slope chart represent higher degrees of accurate separation between preferred and dispreferred variants.



(a) Difference of the log likelihoods of preferred and dispreferred variants

(b) Slope plot of base model predicted likelihood of sequences

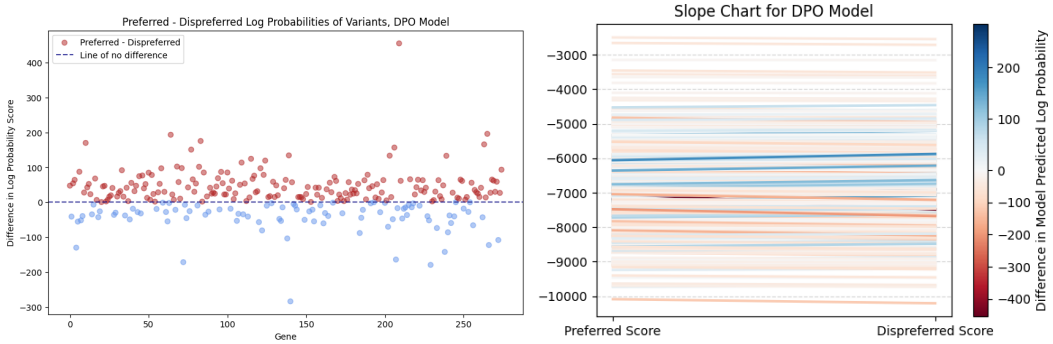
Figure 5: Base model results for the task of likelihood assigment to benign vs. pathogenic sequences.



(a) Difference of the log likelihoods for SFT

(b) Slope plot for SFT

Figure 6: Results following SFT for the task of likelihood assignment to benign vs. pathogenic sequences.



(a) Difference of the log likelihoods for DPO

(b) Slope plot for DPO

Figure 7: Results following DPO for the task of likelihood assignment to benign vs. pathogenic sequences.

Note that all of the slope plots above do not share the same color scale; if they did, we would see very faint colors for the base and SFT models, and more distinct colors for the DPO model. This is a direct result of the differences having higher magnitude in the DPO model as discussed above.

We also plotted the results similarly for the training data in Figure 8 to ensure that suboptimal performance was not due to overfitting of the model. In Table 1, the accuracies are reported for each model on each dataset split. As one can see from both, the results are not remarkably different when

comparing the training set performance and evaluation set performance for the base model, SFT model, and DPO model. This indicates that rather than overfitting, there may be some misalignment between the training objective and the downstream task used to measure model accuracy. Future work should consider how to bridge this gap.

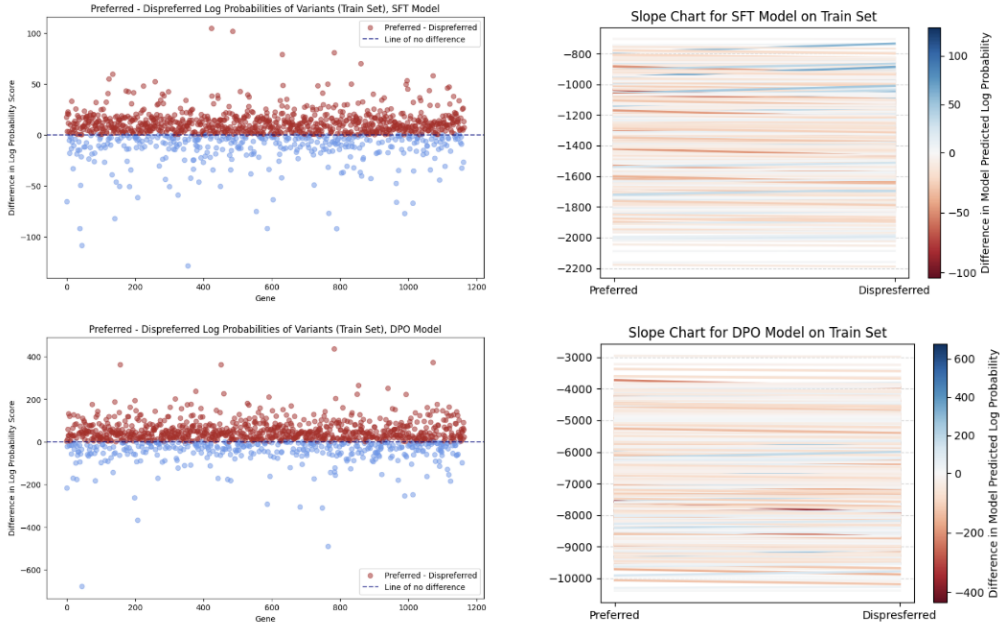


Figure 8: Results of evaluating the training data using the SFT and DPO models.

Split	Model	Task accuracy
Train	Base	0.83
Train	SFT	0.70
Train	DPO	0.69
Test	Base	0.79
Test	SFT	0.66
Test	DPO	0.66

Table 1: Train and test accuracies of the models on the variant classification task.

5.3 Introspection of genes with differential performance

We investigate the set of genes for which the DPO model produced a correct classification (i.e. difference of log likelihoods above 0) of the sequence while the base model produced an incorrect classification (i.e. difference of log likelihoods below 0). We find 29 genes which meet this criteria. We provide a representative sample of these genes and their importance in biomedical research in Table 2, showing the utility of the finetuning method for these particular cases.

6 Discussion

There are a couple of salient points to take away from the results. First, DPO encouraged more separation in the log probabilities when compared to SFT and the base model (Figure 4). As discussed above, DPO seemingly proves reliable at creating a wider disparity between assigned pathogenic and benign sequence probability, perhaps suggesting the model is more confident than the others. If it was trained to a point of higher accuracy, this would be a preferred model to use.

Gene	Research relevance
MAOA	Widely studied for links to aggression, mental health, and Brunner syndrome.
ASPM	One of the key genes controlling brain size; famous for causing primary microcephaly.
TGM1	Causes lamellar ichthyosis, a textbook skin disease.
FKTN	Central in muscular dystrophy research (Fukuyama dystrophy).
PDHA1	Causes pyruvate dehydrogenase deficiency, a widely known metabolic disorder.
GRM7	Heavily studied for its role in epilepsy, depression, and schizophrenia.
ADRA2A	Commonly referenced in studies of ADHD, blood pressure, and stress response.
PLA2G4A	Important in inflammation and cancer research; involved in lipid signaling pathways.
ANGPT2	A target in cancer and vascular disease therapies; widely studied in angiogenesis.

Table 2: Relevant genes for which the DPO model correctly classified pathogenicity but the base model did not.

Second, the pretrained GPN serves as a strong baseline for variant effect prediction and achieves state-of-the-art performance without SFT or DPO on variant classification. This represents the robustness of training models on evolutionary data in capturing some understanding of the pathogeniety of low fitness mutations. However, we would still expect finetuning to make the model better at the task at hand, not worse. A likely cause for this disparity is the quality of the data we trained on. ClinVar is known among the research community to be inconsistent in quality, and somewhat biased to common variants. Notably, since all variants must be present in at least 1% of the population, truly deleterious variants are not represented in ClinVar. This means that the data is implicitly biased toward positive samples. Thus, one may have to correct for that bias or more carefully curate the data set. We envision that this is the most important next step to creating a successful DPO-finetuned model.

7 Conclusion

Through this project, we have seen that biological information, specifically that which represents the fitness of genomic variations is accessible through genomic foundation models. While we had hoped that attempts at finetuning using DPO would have increased a genomic language model's ability to make accurate variant effect predictions, this was not the case. However, much was learned about the different approaches to finetuning and tangible next steps.

Though performance accuracy did not improve, DPO may still act as a viable method for increasing such biological preference tasks when finetuning with a more robust and potentially representative data set.

The base model's higher performance on the quantitative task demonstrates the robustness of training models on evolutionary data in capturing some understanding of the historical pathogenicity of *known* low fitness mutations. However, the question of the depth of understanding of the underlying pathogenicity of a SNP remains unanswered with the decline in performance by the two finetuned models. It remains unclear if gLMs such as GPN would be successful at performing a similar task with novel mutations. Future work to evaluate the model's pathogenicity prediction for a new SNP followed by experimental validation could serve to broaden understanding in this regard.

8 Team Contributions

- **Alejandro:** Project Ideation, Model Training, Deliverables
- **Mohini:** Project Ideation, Model Evaluation, Deliverables
- **Samantha:** Project Ideation, Dataset Generation, Deliverables

Changes from Proposal *Language Model for Experimentation, Finetuning Method, and Evaluation Method*: We initially tried to train Evo 2 with DPO. After discussing with the authors, we decided to switch to a recent gLM that is compatible with Hugging Face's API, GPN-Promoter. We note

competitive performance of GPN-based models against Evo 2 on variant effect prediction. We also initially proposed aligning the genomic language model (gLM) Evo 2 to protein fitness data (pLDDT). After noting that it was too costly to run the required AlphaFold3 forward pass for each sequence, we shifted to looking at aligning gLMs to human clinical variants to improve biological safety of generations.

References

[1] Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Etowah Adams, Stephen A. Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R.K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, Jonathan C. Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modeling and design across all domains of life with evo 2. February 2025.

[2] Gonzalo Benegas, Gökcen Eraslan, and Yun S Song. Benchmarking dna sequence models for causal regulatory variant prediction in human genetics. *bioRxiv*, pages 2025–02, 2025.

[3] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–D868, 2016.

[4] Gonzalo Benegas, Sanjit Singh Batra, and Yun S Song. Dna language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences*, 120(44):e2311219120, 2023.

[5] 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, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, and Thomas Pierrot. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, November 2024.

[6] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

[7] Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, January 2023.

[8] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[9] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023.

[10] 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, February 2021.

[11] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36:43177–43201, 2023.

[12] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, October 2021.

[13] Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R Kelley. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Nature Genetics*, pages 1–13, 2025.

[14] Zhao Yang, Bing Su, Chuan Cao, and Ji-Rong Wen. Regulatory dna sequence design with reinforcement learning, 2025.

[15] Matt Sternke and Joel Karpiak. ProteinRL: Reinforcement learning with generative protein language models for property-directed sequence design. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

[16] Talal Widatalla, Rafael Rafailov, and Brian Hie. Aligning protein generative models with experimental fitness via direct preference optimization. May 2024.

[17] Nathaniel Blalock, Srinath Seshadri, Agrim Babbar, Sarah A Fahlberg, Ameya Kulkarni, and Philip A Romero. Functional alignment of protein language models via reinforcement learning. *bioRxiv*, pages 2025–05, 2025.

[18] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, 4(7):1073–1081, 2009.

[19] Ivan Adzhubei, Daniel M Jordan, and Shamil R Sunyaev. Predicting functional effect of human missense mutations using polyphen-2. *Current protocols in human genetics*, 76(1):7–20, 2013.

[20] Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, 47(D1):D886–D894, 2019.

[21] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, et al. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664):eadg7492, 2023.

[22] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

[23] Gonzalo Benegas, Carlos Albors, Alan J Aw, Chengzhong Ye, and Yun S Song. Gpn-msa: an alignment-based dna language model for genome-wide variant effect prediction. *bioRxiv*, pages 2023–10, 2024.

[24] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.