AUTO-REGRESSIVE WAVENET VARIATIONAL AUTOENCODERS FOR ALIGNMENT-FREE GENERATIVE PROTEIN DESIGN AND FITNESS PREDICTION.

Nikša Praljak
Graduate Program in Biophysical Sciences
University of Chicago
Chicago, IL, USA, 60637
{niksapraljak1}@uchicago.edu

Andrew L. Ferguson
The Pritzker School for Molecular Engineering
University of Chicago
Chicago, IL, USA, 60637
{andrewferguson}@uchicago.edu

ABSTRACT

Recently deep generative models (DGMs) have been highly successful in novel protein design and could enable an unprecedented level of control in therapeutic and industrial applications. One DGM approach is variational autoencoders (VAEs), which can infer higher-order amino acid dependencies for useful prediction of fitness effects of mutation. Additionally, the model infers a latent space distribution, which can learn biologically meaningful representations. Another example of a DGM approach is autoregressive models, commonly implemented in language or audio tasks that have been intensively explored in protein generation of unaligned sequences. Combining these two distinct DGM approaches for protein design and fitness prediction has not been extensively studied because VAEs are prone to posterior collapse when implemented by an expressive decoder. We explore and benchmark the use of VAEs with a WaveNet-based decoder. The advantage of WaveNet-based generators is the inexpensive training time and computation cost relative to recurrent neural networks (RNNs) and avoids vanishing gradients because WaveNets leverage dilated causal convolutions. To avoid posterior collapse, we implemented and adapted an Information Maximizing VAE (InfoVAE) loss objective instead of a standard ELBO training objective to a semi-supervised setting with an autoregressive reconstruction loss. We extend our model from unsupervised to a semi-supervised learning paradigm for fitness prediction tasks and benchmark our model's performance on FLIP and TAPE datasets for protein function prediction. To illustrate our model's performance for protein design, we have trained our models on unaligned homologous sequence libraries of the SH3 domain and AroQ Chorismate mutase enzymes. Then we deployed it to generate novel (variable-length) sequences that are computationally predicted to fold into native structures and possess natural function. Our results demonstrate that combining a semi-supervised InfoVAE model with a WaveNet-based generator provides a robust framework for functional prediction and generative protein design without requiring multiple sequence alignments.

1 INTRODUCTION

Protein sequences from non-homologous families or within homologous families with high variability and diverse lengths present challenges in the construction of multiple-sequence alignments Alley et al. (2019); Biswas et al. (2021); Shin et al. (2021). Deep generative models (DGMs) are exciting models for learning high-dimensional data distributions and generating novel data samples indistinguishable from the true data. These approaches are promising for synthetic protein design. For example, autoregressive models (i.e., language or audio generative models) have no dependency on sequence alignments, allowing these models to learn and generate novel sequences with high variability and diverse lengths. However, one major limitation of autoregressive models is the lack of ability to infer meaningful representations or conditional information (e.g., latent vectors). In contrast, another popular DGM is variational autoencoders (VAEs) which can infer a latent space

and generate novel data indistinguishable from the true data distribution. These models have been shown to effectively predict single-mutant effects Shin et al. (2021), infer a homologous family's phylogeny through the latent space, and diversify synthetic AAV capsids Sinai et al. (2021). While these models can infer a biologically meaningful latent space, they struggle to implement powerful and expressive decoders (i.e., generators) because VAEs are prone to posterior collapse Zhao et al. (2019); Chen et al. (2016); Van Den Oord et al. (2017); Yang et al. (2017). Therefore, VAEs struggle with incorporating autoregressive decoders for generating variable-length sequences and inferring alignment-free homologous protein datasets.

In this work, to successfully combine VAEs with autoregressive generators and overcome posterior collapse Bowman et al. (2015), we incorporated an Information Maximizing (InfoMax) loss objective instead of the common ELBO training objective Zhao et al. (2019). The InfoMax loss is similar to ELBO; however, prefactor weights are introduced to motivate better inference and regularization. A mutual information maximization term is introduced for explicitly encouraging high mutual information between the input vectors and latent space embeddings. We implement a WaveNet-based autoregressive generator Oord et al. (2016) for our decoder that avoids vanishing or exploding gradients by leveraging dilated causal convolutions. Previously, models have been developed that combine VAEs with dilated causal convolutions as the decoder component for text generation Yang et al. (2017); however, this approach has yet to be explored for protein design and fitness prediction. In addition, our work expands our this modeling approach by incorporating an InfoMax loss objective for improving amortized inference and avoiding posterior collapse. These convolutions are much faster than recurrent networks during training time, offer superior inference of long-range correlations, and are computationally lighter-weight than standard convolutional filters. We find that InfoVAE can infer biologically meaningful latent spaces while incorporating an expressive autoregressive generator. We extend the InfoVAE training objective to a semi-supervised learning paradigm for fitness landscape prediction. To assess the generative capacity of our model, we trained the model on two homologous family datasets, inferred meaningful latent spaces, and generated length-variable sequences. In addition, we benchmark our semi-supervised model variant four fitness landscape prediction tasks within TAPE and FLIP protein function prediction benchmarks. We find that our model predicts fitness or function better or competitively with current state-of-the-art.

2 METHODS

2.1 OVERVIEW

Figure 1 presents an overview of our approach. The proteins sequences (either aligned or unaligned) are embedded onto a lower-dimensional space (latent space) using a dilated convolutional neural network encoder $q_\phi(z|x)$. The decoder (i.e., generator) $p_\theta(x|z)$ is a WaveNet-based architecture, which samples from the latent space and predicts amino acid residues while conditioning on previous amino acids $p_\theta(x|z) = p(x_0|z)\prod_{i=1} p(x_i|\mathbf{x}_{<i}, z)$. Generally, when using a dilated causal convolution and predicting the next amino acid residue, we use teacher forcing which leverages the true labeled amino acids for previous conditional amino acids. Since the decoder is an autoregressive generator, variable length sequences can be designed. In addition, this model can be extended to a semi-supervised paradigm such that a discriminator neural network (a simple fully connected regression model; $p_\omega(y|z)$) samples the latent space and predicts fitness or function measurements. In the semi-supervised paradigm, the discriminative and generative losses are learned together.

2.2 INTEGRATING INFOVAE WITH A WAVENET-BASED GENERATOR

Traditional Variational autoencoders (VAEs) are prone to posterior collapse or poor amortized inference when implementing expressive decoders (e.g. autoregressive generators) Zhao et al. (2019); Yang et al. (2017); Chen et al. (2016); Van Den Oord et al. (2017). Here we implement an VAE model Zhao et al. (2019) to overcome posterior collapse and improve variational inference when implementing a WaveNet-based autoregressive decoder. Our unsupervised loss function is the following:

$$\mathcal{L}_{\mathcal{US}} = \xi\mathbb{E}_{z\sim q_\phi(z|x)}\big[log\ p_\theta(x|z)\big] - (1-\alpha)\mathcal{D}_{KL}\big(q_\theta(z|x)||p(z)\big) \tag{1}$$
$$-(\alpha+\lambda-1)\mathcal{D}_{MMD}\big(q_{\phi(z)}||p(z)\big)$$

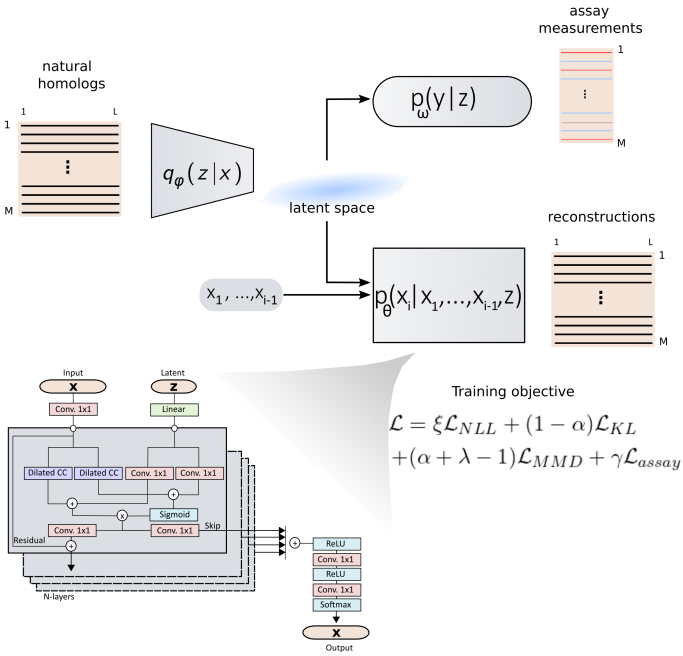

Figure 1: Schematic illustration of the model integrating an InfoMax VAE with convolutional encoder and WaveNet decoder. The architecture employs a dilated convolutional encoder $q_\varphi(z|x)$, a WaveNet decoder $p_\theta(x_i|x_1, ..., x_{i-1}, z)$, and semi-supervised regressor $p_\omega(y|z)$ to predict functional assay measurements from the learned VAE latent space.

where $p_\theta(x|z)$ is the decoder model, $\mathcal{D}_{KL}$ is the Kullback-Leibler divergence between the variational posterior approximation $q_\phi(z|x)$ and normal prior distribution $p(z)$. The third term $\mathcal{D}_{MMD}$ is the max-mean discrepancy (MMD) that helps penalize the aggregated posterior distribution and improves amortized inference. The derivation of the above loss objective can be found Zhao et al. (2019). In our work, we introduce an autoregressive decoder (WaveNet-based architecture), where $p_\theta(x|z) = p_\theta(x_0|z)\prod_{i=1} p_\theta(x_i|\mathbf{x}_{<i}, z)$. The MMD divergence term becomes $\mathcal{D}_{MMD} = \mathbb{E}_{z,z'\sim p(z),p(z')}\big[k(z,z')\big] - 2\mathbb{E}_{z,z'\sim q(z),p(z')}\big[k(z,z')\big] + \mathbb{E}_{z,z'\sim q(z),q(z')}\big[k(z,z')\big]$, where $k(\cdot,\cdot)$ is a positive definite kernel and $\mathcal{D}_{MMD} = 0$ if and only if $p(z) = q(z)$. We choose the Gaussian kernel $k(z,z') = e^{(z-z')^2/\sigma^2}$ as our characteristic kernel $k(\cdot,\cdot)$, and σ is a hyperparameter defining the bandwidth of our Gaussian kernel. The prefactor loss weights ξ, α, and λ scales the contribution of the reconstruction loss, weighs the mutual information between x and z, and scales the penalization of MMD divergence.

2.3 Extending generative model to a semi-supervised paradigm

We extend the unsupervised WaveNet VAE to a semi-supervised learning paradigm for fitness landscape prediction. The main motivation of using a semi-supervised approach is based on the idea that latent representations z can be more informative for predicting y when also used for reconstructing x. In addition, semi-supervised learning is beneficial when labels are scarce, and unlabeled data is abundant, which is generally the case for protein design over large unlabeled seqeunce databases for which a small fraction of sequences are labeled with functional assays. The semi-supervised training objective is the following:

$$\mathcal{L}_{\mathcal{SS}} = \mathcal{L}_{\mathcal{US}} + \gamma\mathbb{E}_{(x,y)\in\mathcal{D}_L}\big[log\, p_\omega(y|z)\big] \tag{2}$$

where $p_\omega(y|z)$ is a regression model (a simple fully connected neural network) parameterized with training parameters ω. In practice, we minimize the mean-squared error objective $\frac{1}{2}|y - \tilde{y}|^2$, where

y and $\tilde{y}$ is the ground truth and predicted regression value. The $(x, y) \in \mathcal{D}_L$ denotes that the samples which are fed through the supervised branch are only sequences x with assay measurements y.

2.4 TRAINING AND HYPERPARAMETER OPTIMIZATION

During training, we set ξ, α, λ, γ, and σ to 1, 0.95, 2, 1, and $\sqrt{dim(z)}$; then, we conducted hyperparameter optimization over the latent space dimensions $z \in \big[1, 20\big]$ for each fitness landscape prediction task. The optimal latent space dimension was choosen based on minimizing the negative-log likelihood and maximizing the Spearman ρ score on the validation set. In general, the prefactor loss weights can be optimized as well. The optimization algorithm used in this study was Adam Kingma & Ba (2014) with a learning rate of1E-4. For fitness prediction tasks, the number of epochs was set to 2000, and early stopping was only implemented if ρ reaches a value of 0.99 on the validation set. For simplicity, we set the minibatch size across all fitness prediction tasks to 256 samples, but this too can be optimized.

3 RESULTS: PROTEIN DESIGN

In our work, to illustrate the advantage of combining VAEs with an autoregressive WaveNet decoder, we trained our model on unaligned homologous datasets. To show that our model can handle unaligned homologous sequences, we will compare the latent embeddings and learned representations between unaligned and aligned sequence datasets.

3.1 DESCRIPTION OF THE HOMOLOG FAMILY DATASETS

We train our model and generate novel sequences from two homologous protein families: Src homologous 3 domains (SH3) and AroQ chorismate mutase (CM) enzymes. The SH3 family consists of many paralogs, which are homologous sequences that diverge due to duplication events. Since the gene is duplicated in the genome, new selective pressure can act on the duplicated gene, and subsequently, paralogs can acquire new functions. The SH3 dataset size is 5611 sequences, consisting primarily of proteins found in the fungal kingdom. Of the 5611 sequences in the basebase, 4664 are labeled with functional assay measurements for osmosensing capabilities. The CM dataset differs from the SH3 dataset because all of proteins are orthologs, which are homologous sequences that diverge due to speciation events instead of duplication events. This means most of the natural homologous CM sequences have a similar catalytic function. In addition, the CM dataset has two sets of proteins – one set corresponds to the natural homologs (1130 sequences), while the second set corresponds to synthetic designs (1618 sequences) produced using a direct coupling analysis (DCA) model that explicitly considers only pairwise epistasis Russ et al. (2020). All sequences have been functionally assayed for CM catalytic function.

3.2 LATENT SPACE INTERPRETATION

We present in Figure 2 the latent space embeddings of the SH3 dataset produced by unsupervised training (Eqn. 1) of our model operating on aligned (Figure 2A) and unaligned (Figure 2B) training data. The embedded points are colored according to an experimental select-seq assay Russ et al. (2020) that reports a proxy measure for osmosensing function termed a normalized relative enrichment (n.r.e.) score. A n.r.e. = 1 corresponds to activity comparable to the wild type SH3; a n.r.e. = 0 corresponds to activity commensurate with a null gene. In both the aligned and unaligned latent space embeddings we observe strong clustering of the highly active osmosensing sequences with n.r.e. scores of 1.0. This demonstrates that the fully unsupervised model has learned biologically meaningful representations of the sequence ensemble separating orthologs (sequences with osmosensing function; red) from paralogs (sequences lacking osmosensing activity; blue). Additional 2D projections of the SH3 dataset is shown on Figure 7.

In Figure 3 we present the aligned and unaligned latent space embeddings for the CM dataset. When training our model on the CM dataset, the training set contains only the natural homologs while the validation set contained the synthetic designs. As was observed for the SH3 data, the unsupervised model learns a meaningful latent space embedding of the training data in which we observe an emergent clustering and gradient in catalytic activity again measured by a select-seq assay Russ

et al. (2020) that provides a n.r.e. score. We observe that the model is able to generalize quite well by embedding the synthetic designs onto the latent space and clustering high fitness embeddings into a region of the space. Additional 2D elevations of the CM latent space are provided in Figure 8.

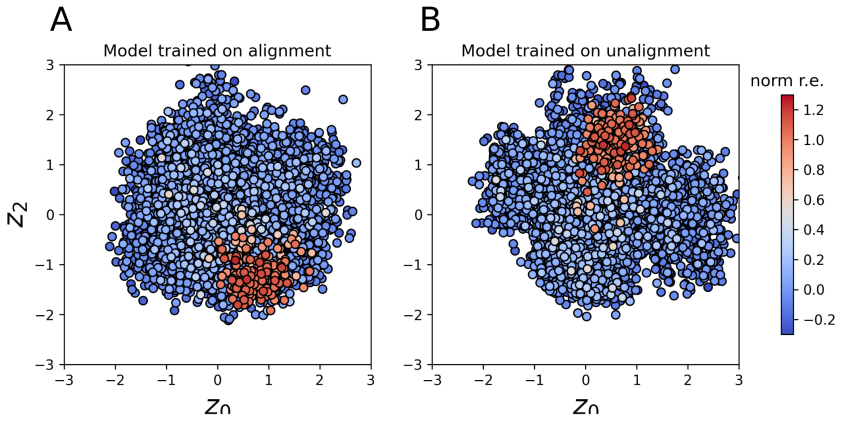

Figure 2: We plot the latent space embeddings of the natural SH3 homolog library, consisting of various different paralog groups and including the Sho1 paralog group. The colorbar represents *in vivo* fitness and indicates whether a SH3 homolog can rescue osmosensing functionality in *S. cerevisiae*. The latent spaces learned over **(A)** aligned and **(B)** unaligned sequences both provide a good separation of the high activity orthologs (red) and low activity paralogs (blue). Importantly, both models were trained using an unsupervised learning approach that was not exposed to functional assay measurements, indicating our generative model is able to learn meaning representation for designing function solely from unlabeled sequences.

3.3 Generation of novel variable length sequences with structure prediction

To illustrate the practical advantage of using an autoregressive decoder, we compared our model trained on aligned and unaligned sequence data for both the SH3 and CM datasets. We generated novel sequences by randomly sampling points within the latent spaces and decoding these through the WaveNet generator to produce novel protein sequences. For the SH3 system, we sampled and generated 5611 novel sequences, while for the CM system, we generated 1130 novel sequences. To check whether our generated sequences fold into a proper tertiary structure, we used AlphaFold2 to predict structures of four sequences for both the SH3 and CM task: the shortest generated sequence, the longest generated sequence, and two randomly selected sequences (Tables 2 and 3).

With the SH3 task, our structure predictions of the generated sequences along with the wild-type SH3 domain (PDB: 2VKN) are shown in Figure 4. All four sequences are predicted to have a very similar tertiary structure as the wild-type even though the sequence similarity to the nearest natural SH3 domain lies between only 41-43%. (We define sequence similarity to the nearest natural homolog as 1-$\frac{\eta}{L_{max}}$, where η and L_{max} are the minimum hamming distance and longest protein sequence within the natural homolog library). Interestingly,the longest SH3 domain among the generated pool is predicted to acquire an alpha helix, which was originally a hairpin loop on the WT structure (Figure 4D). It is important to note that this sequence is 11 amino acids longer than the wild-type SH3 domain (Table 2), potentially allowing the generative model to extrapolate in terms of design.

With the CM task, the two randomly sampled sequences from the generated pool (Figure 5A,B) have a very similar tertiary structure to the wild-type AroQ CM monomer in *Escherichia coli* (PDB: 1ECM). The shortest generated CM sequence is missing a whole alpha helix and significant portion of a second alpha helix compared to the wild-type but is still predicted to fold into a stable tertiary structure even though it is 45 amino acids shorter than the WT sequence (Table 3). In Figure 5D, the longest generated CM sequence seems to maintain all three alpha helices, but these helices are predicted to be longer and straighter than the wild-type helices. The four CM generated sequences have a sequence similarity to the nearest natural CM homolog of 0.18-0.21%.

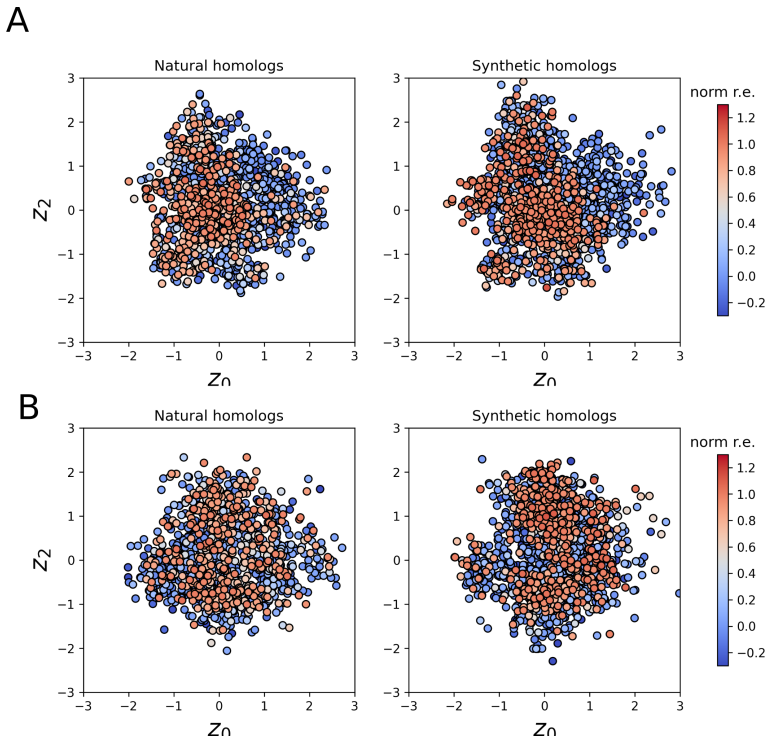

Figure 3: We plot the latent embeddings of both training (natural CM homologs) and validation (synthetic CM homologs) set. In panel **(A)**, we show the latent embeddings when the model was trained on input sequences with a multiple-sequence alignment. While in panel **(B)**, we show the latent embeddings of natural and synthetic CM sequences when using a model trained on unaligned input data. Since the CM training dataset consists of only orthologs, the latent space is more scattered in terms of high fitness embeddings (n. r.e.). However, the overall latent space is retains the Gaussian structure and learns to cluster some high fitness regions regardless whether the model was trained on aligned or unaligned input data.

In Figure 6 we analyze the entire pool of generated sequences from the SH3 and CM tasks. We computed the sequence length of each sequence for the SH3 task (Figure 6A), finding that the length variability is more diverse when trained on unaligned versus aligned input data. This was not necessarily the case for the CM homologs (Figure 6B). However, for both the SH3 and CM tasks, the model is able to generate sequences that are less sequence similar to the natural homologs when trained on unaligned input data (Figure 6A,B), illustrating a potential advantage of training generative models on unaligned sequences in generating a broader diversity of sequences that better recapitulate the natural diversity of sequence lengths while maintaining the native tertiary structure.

4 Results: Fitness prediction

One important goal of deep learning models for biology is to learn meaningful representations that can be leveraged on down-stream tasks. For instance, a major endeavor in protein design is fitness landscape prediction and representation learning for semi- and self-supervised tasks. We extended our deep generative model to a semi-supervised paradigm in the hopes of learning biologically meaningful representations for fitness landscape prediction. Our main goal is to learn a latent space z that is informative for the generative and discriminative tasks. The intuition behind this construction is based on the ideal that the representations that can be used to reconstruct the training data and generate new data indistinguishable from the training data can also be more meaningful for discriminative tasks (e.g. fitness landscape prediction). To benchmark our model's learned representations, we tested its ability to predict test datasets on four main protein systems from two popular community

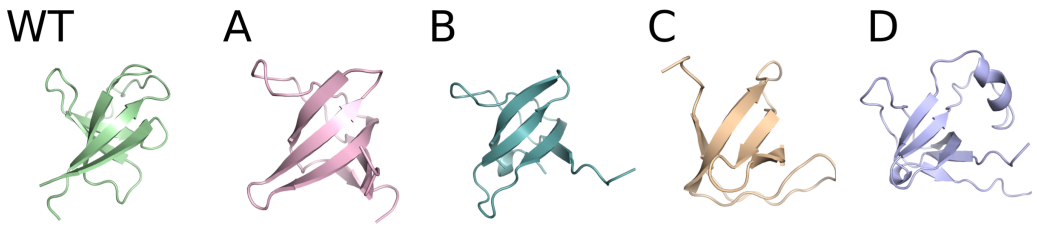

Figure 4: AlphaFold 2 predicted structures of four synthetically designed SH3 sequences. First column corresponds to the WT Sho1^{SH3} domain in *S. cerevisiae* (PDB: 2VKN) which has a length of 70 amino acids. The next four columns (**A-D**) corresponds to design sequences with variable lengths, where the **A,B** corresponds to randomly generated sequences and **C,D** corresponds to the shortest, longest generated SH3 domains. The primary structure and sequence length for these proteins are shown on Table 2

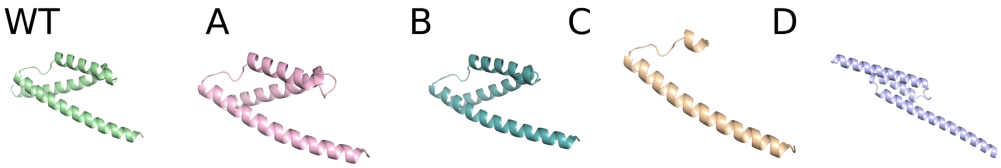

Figure 5: AlphaFold 2 predicted structures of four synthetically designed CM sequences. First column corresponds to the WT AroQ Chorismate mutase (CM) enzyme in *Escherichia coli* (PDB: 1ECM), while the next two columns (**A,B**) corresponds to design sequences randomly sampled from the generated pool of CM sequences. (**C,D**) corresponds to the remaining two generated sequences that are shortest and longest sequence within the generated pool. The primary structure and sequence length for these proteins are shown on Table 3

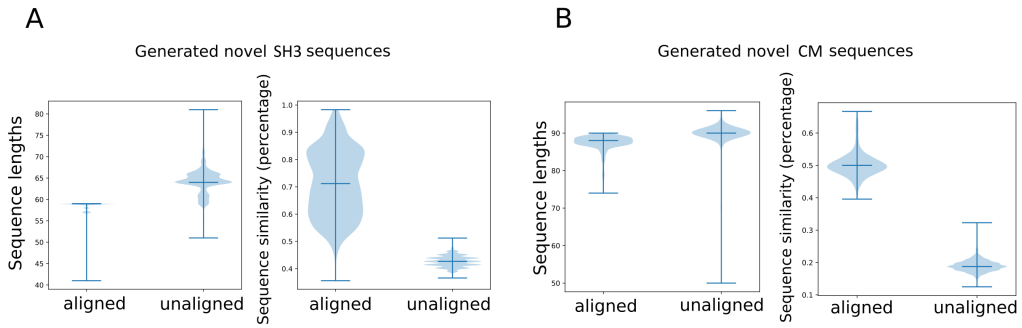

Figure 6: We generated 5611 and 1130 novel SH3 and CM sequences by randomly sampling across the latent space. In panel (**A**), we compared the generated sequences in terms of sequence length and similarity to the nearest natural SH3 homolog for models trained on aligned or unaligned input data. Similarly in (**B**), we compare the generated sequences' lengths and similarity to the nearest natural CM homolog. For both SH3 and CM, we find that sequence diversity is improved when trained on unaligned sequences, but the sequence length variability for generated CM homolog is similar regardless of aligned or unaligned input data.

benchmarks tasks: TAPE Rao et al. (2019) and FLIP Dallago et al. (2022). Thus, the four protein systems are the following:

1. Mutational screening fitness landscape of VP-1 AAV proteins (FLIP) Bryant et al. (2021); Zhang et al. (2019)

2. Highly epistatic mutational landscape GB1 (FLIP) Wu et al. (2016); Franks et al. (2006)
3. Epistatic Green Fluorescent Protein (GFP) Landscape Predictions (TAPE) Sarkisyan et al. (2016)
4. Stability Landscape Prediction (TAPE) Rocklin et al. (2017)

4.1 BENCHMARK MODEL ON LANDSCAPE PREDICTION TASKS

For the FLIP AAV capsid task, there were 7 different data split tasks that are each relevant to protein engineering scenarios. We find that semi-supervised model is able to outperform or competitively perform to current baseline scores in 5 out of 6 dataset splits (Table 4 and 5). However, we find that the semi-supervised generative model underperforms when the training set contains only low-fitness sequences and the test set contains only high-fitness sequences. For the FLIP GB1 task, we find that our model outperforms or competitively performs against the current baseline scores (Table 1). However, similar to the AAV capsid tasks, our model underperforms on the protein task where the training and testing splits contain only low- and high-fitness sequences. For the TAPE tasks, our model competitively performs against the state-of-the-art models on the GFP task (Table 6). However, the model underperforms on the stability prediction task found in Table 7. Overall, these results suggest that our generative model is not only capable of unsupervised learning and protein design but also capable of generalizing such that the model can infer fitness landscapes and predict function from sequence alone and is competitive with state-of-the-art.

Table 1: GB1 performance comparison to current baseline scores (metric: Spearman correlation).

Architecture	low-vs-high ρ	1-vs-rest ρ	2-vs-rest ρ	3-vs-rest ρ
ESM-1b Dallago et al. (2022)	**0.59**	0.28	0.55	0.79
ESM-1b Dallago et al. (2022)	0.13	**0.32**	0.36	0.54
ESM-1v Dallago et al. (2022)	0.10	**0.32**	0.32	0.77
Ridge Dallago et al. (2022)	0.34	0.28	**0.59**	0.76
CNN Dallago et al. (2022)	0.51	0.17	0.32	**0.83**
Levenshtein Dallago et al. (2022)	-0.1	-0.17	0.16	0.01
BLOSUM62 Dallago et al. (2022)	-0.13	0.15	0.14	0.01
Our model	0.42	0.28	**0.61**	**0.87**

5 CONCLUSIONS

We combined a variational autoencoder (VAE) and autoregressive generator (WaveNet) for protein design, avoiding the need of multiple-sequence aligned input data. To avoid posterior collapse when combining VAEs and WaveNet models, we implemented a Infomation Maximizing VAE (InfoVAE), adding a mutual information term to the common ELBO training objective, improving amortized inference, and forcing the decoder to use the latent conditional information. We find that generative model is able to learn meaningful latent space representations from homologous protein families, which can be leveraged to design novel functional sequences with length variability. We find that generated structures can be realized by sequence designs that are predicted by AlphaFold 2 to adopt tertiary structures in good agreement with the native fold. Additionally, we find that when the model is trained on unaligned versus aligned input sequences, the diversity of the generated sequences can improve, illustrating a potential advantage of using an autoregressive decoder. We extended our model to a semi-supervised learning paradigm and benchmarked the model on 4 different fitness landscape prediction tasks from FLIP and TAPE. We find that our model can outperform many baseline scores for the AAV and GB1 tasks. In addition, the model can compete against state-of-the-art performance for the GFP task, but underperforms in predicting the stability landscape tasks. These results and analysis suggest our deep generative model is capable of successful protein design of variable length sequences, inferring meaningful biological representations, and effectively predicting fitness from sequence alone.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the Machine Learning in the Chemical Sciences and Engineering program of The Camille and Henry Dreyfus Foundation. This work was completed in part with resources provided by the University of Chicago Research Computing Center. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No.1746045. This work was completed based upon work supported by the National Institute of Biomedical Imaging and Bioengineering under Grant No. 5T32EB009412.

REFERENCES

Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. *arXiv preprint arXiv:1902.08661*, 2019.

Surojit Biswas, Grigory Khimulya, Ethan C Alley, Kevin M Esvelt, and George M Church. Low-n protein engineering with data-efficient deep learning. *Nature methods*, 18(4):389–396, 2021.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Drew H Bryant, Ali Bashir, Sam Sinai, Nina K Jain, Pierce J Ogden, Patrick F Riley, George M Church, Lucy J Colwell, and Eric D Kelsic. Deep diversification of an aav capsid protein by machine learning. *Nature Biotechnology*, 39(6):691–696, 2021.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

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*, 2022. doi: 10.1101/2021.11.09.467890. URL https://www.biorxiv.org/content/early/2022/01/19/2021.11.09.467890.

W Trent Franks, Benjamin J Wylie, Sara A Stellfox, and Chad M Rienstra. Backbone conformational constraints in a microcrystalline u-15n-labeled protein by 3d dipolar-shift solid-state nmr spectroscopy. *Journal of the American Chemical Society*, 128(10):3154–3155, 2006.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Amy X Lu, Haoran Zhang, Marzyeh Ghassemi, and Alan M Moses. Self-supervised contrastive learning of protein representations by mutual information maximization. *BioRxiv*, 2020. doi: https://doi.org/10.1101/2020.09.04.283929.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

Gabriel J Rocklin, Tamuka M Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K Mulligan, Aaron Chevalier, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357 (6347):168–175, 2017.

William P Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, et al. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, 2020.

Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

Amir Shanehsazzadeh, David Belanger, and David Dohan. Is transfer learning necessary for protein landscape prediction? *arXiv preprint arXiv:2011.03443*, 2020.

Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Protein design and variant prediction using autoregressive generative models. *Nature communications*, 12(1):1–11, 2021.

Sam Sinai, Nina Jain, George M Church, and Eric D Kelsic. Generative aav capsid diversification by latent interpolation. *bioRxiv*, 2021. doi: https://doi.org/10.1101/2021.04.16.440236.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Nicholas C Wu, Lei Dai, C Anders Olson, James O Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife*, 5:e16965, 2016.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *International conference on machine learning*, pp. 3881–3890. PMLR, 2017.

Ran Zhang, Lin Cao, Mengtian Cui, Zixian Sun, Mingxu Hu, Rouxuan Zhang, William Stuart, Xiaochu Zhao, Zirui Yang, Xueming Li, et al. Adeno-associated virus 2 bound to its cellular receptor aavr. *Nature microbiology*, 4(4):675–682, 2019.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pp. 5885–5892, 2019.

A Appendix

Table 2: Primary structure of the generated SH3 proteins in Figure 4

Label	Primary structure	AA length
WT	DDNFIYKAKALYPYDADDDDAYEISFEQNEILQVSDIEGRWWKARRANGETGIIPSNYVQLIDGPEEMHR	70
A	ASTLFYARALYDYTAQGDDELSVAEGDLLYVLERDDDGWWKAEKDGGAGGEPAEPIELLNP	61
B	APAVETATALYDYEAQADGDLSFSFGDRITIVERTNSDDWWYGRNNRGEFGFFPANYVE	59
C	APGGVYAVVLYDFDANGDDEVDVKEGEELVILDRSNPEWFVAKNPATGEPV	51
D	APPKKVARALYDFTAEGDDELDVKEGDVVLVLEKDDGYWLVVKDDGTGGGPVVWLQSCYAVTDSSGLVPVSYVEIVPASTT	81

Table 3: Primary structure of the generated CM proteins in Figure 5

Label	Primary structure	AA length
WT	PLLALREKISALDEKLLALLAERRELAVEVGKAKLLSHRPVRDIDRERDLLERLITLGKAHHLDAHYITRLFQLIIEDSVLTQQALLQQHLNKIN	95
A	SDLEELREEIDQIDRQIIDLLAERMKRVREVGQYKISKGGPVFDPPREAEVIERLRRLAAAPLGDPERVAALLRRLIEESVLDQLDEELVK	91
B	SDLEELREEIDQIDRQIDELLAERLKLVAEVGEYKASIGLPVYDPKREAQVLDRLRELAKNAGLDPEFAELFLDFVIAEIIRHHEAIQNK	90
C	SDLEELREEIDQIDRQIIDLLNERMKIVREVGEYKISKGLPVYDPEREKQ	50
D	SDAELLELRRRQIDIDDARLELLAERRRRVAEVAALKKLANGLPRRFRREEAVLLKRLSRAAEPPGPADVAALLRRLIRAAARAQAAEAFAERRRL	96

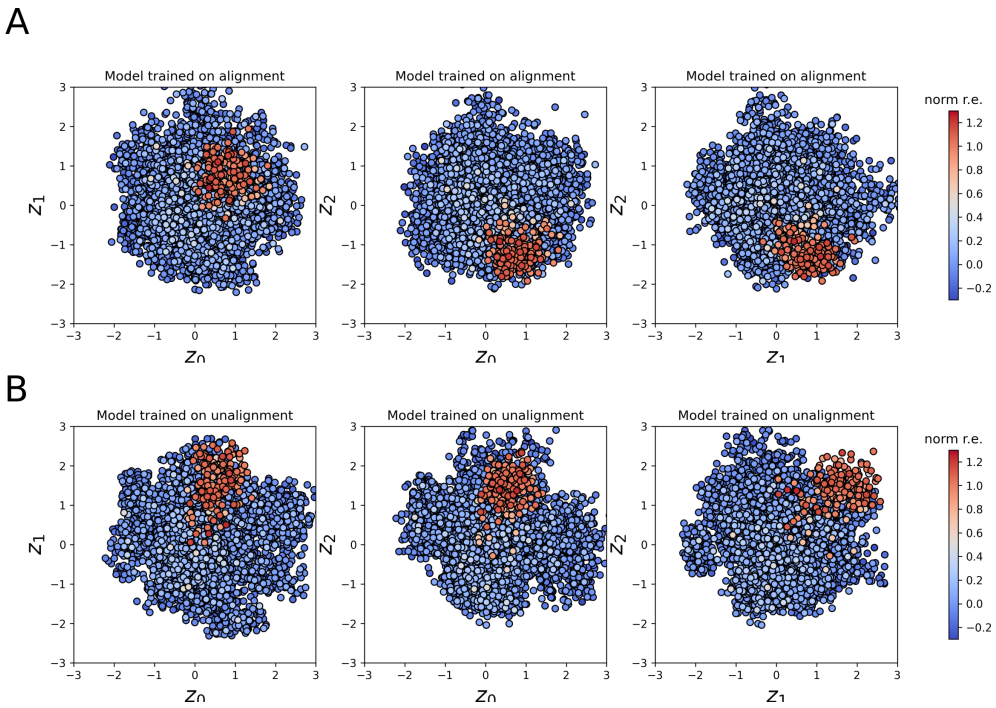

Figure 7: All of the possible 2D projections of the CM latent space.. In **(A)**, the latent embeddings of the natural homologs when the model is trained on aligned input data. While in **(B)**, the latent embeddings of natural homologs are shown when the model is trained on unaligned input data. Regardless whether the model is trained on aligned or unaligned input data, the encoder learns an embedding that discriminates between high activity orthologs (red) from low activity paralogs (blue). Importantly, no functional assay data was provided to the model during training and the unsupervised model learned this partitioning based on sequence data alone.

Table 4: AAV performance comparison to current baseline scores (metric: Spearman correlation).

Architecture	Mut-des ρ	des-mut ρ	low-vs-high ρ
ESM-1b Dallago et al. (2022)	0.76	N/A	**0.39**
ESM-1v Dallago et al. (2022)	**0.79**	N/A	0.34
Ridge Dallago et al. (2022)	0.64	0.53	0.12
CNN Dallago et al. (2022)	0.71	**0.75**	0.34
Levenshtein Dallago et al. (2022)	0.60	-0.07	0.25
BLOSUM62 Dallago et al. (2022)	N/A	N/A	N/A
Our model	**0.82**	**0.78**	0.17

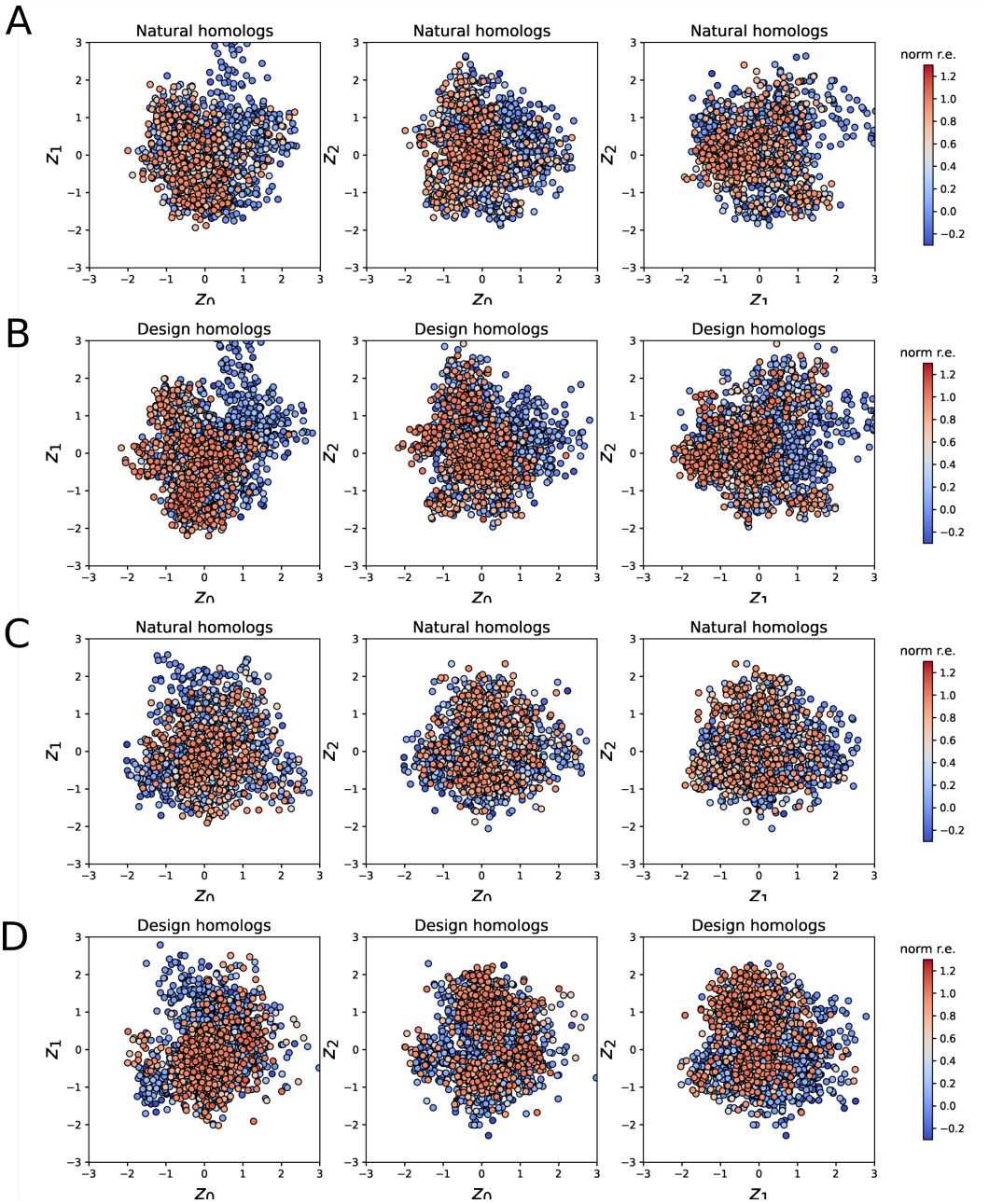

Figure 8: All of the possible 2D projections of the CM latent space.. In **(A-B)**, the latent embeddings of the natural and synthetic design homologs when the model is trained on aligned input data. While in **(C-D)**, the latent embeddings of natural and synthetic design homologs are shown when the model is trained on unaligned input data. Regardless whether the model is trained on aligned or unaligned input data, the encoder is able to learn a latent space that retains the Gaussian structure and learns representations, clusters of high fitness regions. Importantly, the synthetic design sequneces are data from the hold-out set, demonstrating the model's ability to generalize.

Table 5: AAV performance comparison to current baseline scores on mutagenesis-based dataset splits (metric: Spearman correlation).

Architecture	1-vs-rest ρ	2-vs-rest ρ	7-vs-rest ρ
ESM-1b Dallago et al. (2022)	0.03	0.65	0.65
ESM-1v Dallago et al. (2022)	0.10	0.70	0.70
Ridge Dallago et al. (2022)	0.22	0.03	0.65
CNN Dallago et al. (2022)	**0.48**	**0.74**	**0.74**
Levenshtein Dallago et al. (2022)	-0.11	0.57	0.53
BLOSUM62 Dallago et al. (2022)	N/A	N/A	N/A
Our model	**0.61**	**0.74**	0.71

Table 6: GFP state-of-the-art scores (metrics: mean squared error MSE and spearman correlation ρ). Here, the metrics are evaluated on both the bright and dark modes. In Table 8 and 9, the metrics are evaluated on the bright and dark mode of the test set.

Architecture	Pretraining	Full MSE	Full ρ
TAPE Transformer Rao et al. (2019)	no pretraining	2.59	0.22
TAPE LSTM Rao et al. (2019)	no pretraining	2.35	0.21
TAPE ResNet Rao et al. (2019)	no pretraining	2.79	-0.28
ESM Dallago et al. (2022)	masked language	N/A	**0.68**
TAPE Transformer Rao et al. (2019)	masked language	0.22	**0.68**
TAPE LSTM Rao et al. (2019)	bidirectional language	**0.19**	0.67
TAPE ResNet Rao et al. (2019)	masked language	3.04	0.21
UniRep Alley et al. (2019)	language + structure	0.20	0.67
LSTM Bepler & Berger (2019)	supervised	2.17	0.33
CPCProt Lu et al. (2020)	contrastive	N/A	**0.68**
CPCProt-LSTM Lu et al. (2020)	contrastive	N/A	**0.68**
Linear regression Shanehsazzadeh et al. (2020)	none	0.35	**0.68**
CNN Shanehsazzadeh et al. (2020)	none	0.23	**0.68**
Mutation count Dallago et al. (2022)	none	N/A	0.45
BLOSUM62 score Dallago et al. (2022)	none	N/A	0.50
Our model	no pretraining	0.21	**0.67**

Table 7: Overall stability prediction results on the test set (metrics: Spearman's correlation ρ and accuracy)

Architecture	Spearman's ρ	Accuracy
Transformer (No pretraining)	-0.06	0.5
LSTM (No pretraining)	0.28	0.6
ResNet (No pretraining)	0.61	0.68
Transformer (Pretrained)	**0.73**	**0.70**
LSTM (Pretrained)	0.69	0.69
ResNet (Pretrained)	**0.73**	0.66
Supervised	0.64	0.67
UniRep	0.73	0.69
Baseline	0.19	0.58
Our model	0.51	N/A

Table 8: GFP benchmark scores on the bright mode only (metrics: mean-squared error and Spearman's ρ).

Architecture	Pretraining	Bright MSE	Bright ρ
TAPE Transformer Rao et al. (2019)	no pretraining	0.08	0.08
TAPE LSTM Rao et al. (2019)	no pretraining	0.11	0.05
TAPE ResNet Rao et al. (2019)	no pretraining	**0.07**	-0.07
ESM Dallago et al. (2022)	masked language	N/A	N/A
TAPE Transformer Rao et al. (2019)	masked language	0.09	0.60
TAPE LSTM Rao et al. (2019)	bidirectional language	0.12	0.62
TAPE ResNet Rao et al. (2019)	masked language	0.12	0.05
UniRep Alley et al. (2019)	language + structure	0.13	0.63
LSTM Bepler & Berger (2019)	supervised	0.08	0.06
CPCProt Lu et al. (2020)	contrastive	N/A	N/A
CPCProt-LSTM Lu et al. (2020)	contrastive	N/A	N/A
Linear regression Shanehsazzadeh et al. (2020)	none	0.09	**0.68**
CNN Shanehsazzadeh et al. (2020)	none	0.12	0.66
Mutation count Dallago et al. (2022)	none	N/A	N/A
BLOSUM62 score Dallago et al. (2022)	none	N/A	N/A
Our model	no pretraining	0.118	0.54

Table 9: GFP benchmark scores on the dark mode only (metrics: mean-squared error and Spearman's ρ).

Architecture	Pretraining	Dark MSE	Dark ρ
TAPE Transformer Rao et al. (2019)	no pretraining	3.79	0
TAPE LSTM Rao et al. (2019)	no pretraining	3.43	-0.01
TAPE ResNet Rao et al. (2019)	no pretraining	4.1	-0.01
ESM Dallago et al. (2022)	masked language	N/A	N/A
TAPE Transformer Rao et al. (2019)	masked language	0.29	**0.05**
TAPE LSTM Rao et al. (2019)	bidirectional language	**0.22**	0.04
TAPE ResNet Rao et al. (2019)	masked language	4.45	0.02
UniRep Alley et al. (2019)	language + structure	0.24	0.04
LSTM Bepler & Berger (2019)	supervised	3.17	0.02
CPCProt Lu et al. (2020)	contrastive	N/A	N/A
CPCProt-LSTM Lu et al. (2020)	contrastive	N/A	N/A
Linear regression Shanehsazzadeh et al. (2020)	none	0.33	**0.05**
CNN Shanehsazzadeh et al. (2020)	none	0.28	**0.05**
Mutation count Dallago et al. (2022)	none	N/A	N/A
BLOSUM62 score Dallago et al. (2022)	none	N/A	N/A
Our model	no pretraining	0.27	**0.06**