Brief Assessment of Predictive and Generative Capabilities of Principal Component Analysis for Protein Design

Nathaniel Blalock
CBE562: Statistics
December 14th, 2023

Introduction

Proteins can address some of the greatest problems facing society including treating chronic illnesses, producing sustainable biofuels, and advancing waste-to-energy platforms to fight climate change. The scope for innovation is nearly limitless given the multifaceted roles of proteins in biological systems. Protein design spaces are remarkedly large. A relatively short 119 amino acid has a design space of 20^{119} (or 6.65×10^{154}) sequences. This means the protein design space for most proteins is larger than the estimated number of atoms in the universe. Protein engineers therefore aspire to explore protein design spaces computationally and design superior proteins.

Deep learning for protein engineering has improved our ability to model local protein design landscapes from experimental data by learning the complicated, non-linear relationship between protein sequence and function [1-2]. These supervised models while state-of-the-art are limited in their ability to extrapolate beyond the training data for regression tasks [3]. Unsupervised deep learning models have recently drawn attention for designing proteins by leveraging unlabeled natural protein sequences [4-5]. Natural selection provides selective pressure for beneficial amino acid mutations during evolutionary trajectories [6-7]. A family of natural sequences can therefore provide insight into biologically relevant constraints. These constraints can enable further exploration of the protein design space combined with insights from a supervised model and design proteins with a desirable function enhanced. A variational autoencoder (VAE) is an unsupervised deep learning model proving capable of learning from related natural sequences aligned to a protein-of-interest for designing similarly functioning, novel, and diverse protein sequences [8]. The regularized latent space of a VAE has been correlated with protein function [9]. Given that latent space dimensions can be correlated with protein function, the correlated latent space dimension can potentially help guide a stochastic search algorithm with a supervised model to design protein sequences with superior function. Simulated annealing is an effective search algorithm for exploring protein design spaces [3]. During each timestep of simulated annealing, an amino acid is randomly mutated and various models can predict the function of the generated sequence. After many timesteps, a protein with superior function can be found. Principal component analysis (PCA) provides an alternative statistical framework for correlating principal components (PCs) with limited experimental data for predicting the function of generated protein sequences. This work compares the predictive, and briefly generative, capabilities of principal component analysis and a variational autoencoder for protein design.

Methods

CreiLOV has emerged as a promising thermostable, photostable, and rapidly maturing monomeric fluorescent protein that operates independently of oxygen for novel study and engineering of enzymes and metabolic pathways in anaerobic environments including the gut microbiome, tumor environments, and high-density fermentations. A multiple sequence alignment of 243,582 natural protein sequences related to CreiLOV was curated using Jackhmmer software [10] after removing sequences less than 55% of the length of CreiLOV and the columns not corresponding to the protein-of-interest CreiLOV [8,11]. 167,613 variants of CreiLOV were obtained for assessing the correlation of principal components with fluorescent measurements [12]. The 243,582 unlabeled natural protein sequences were one-hot encoded and flattened into a 2-dimensional matrix. Principal component analysis (PCA) was performed using sklearn with a random seed of 0 for reproducibility. The top 50 principal components explained 30% of the variance in the multiple sequence alignment. The first component explained <4% of the variance in the data (Figure 1).

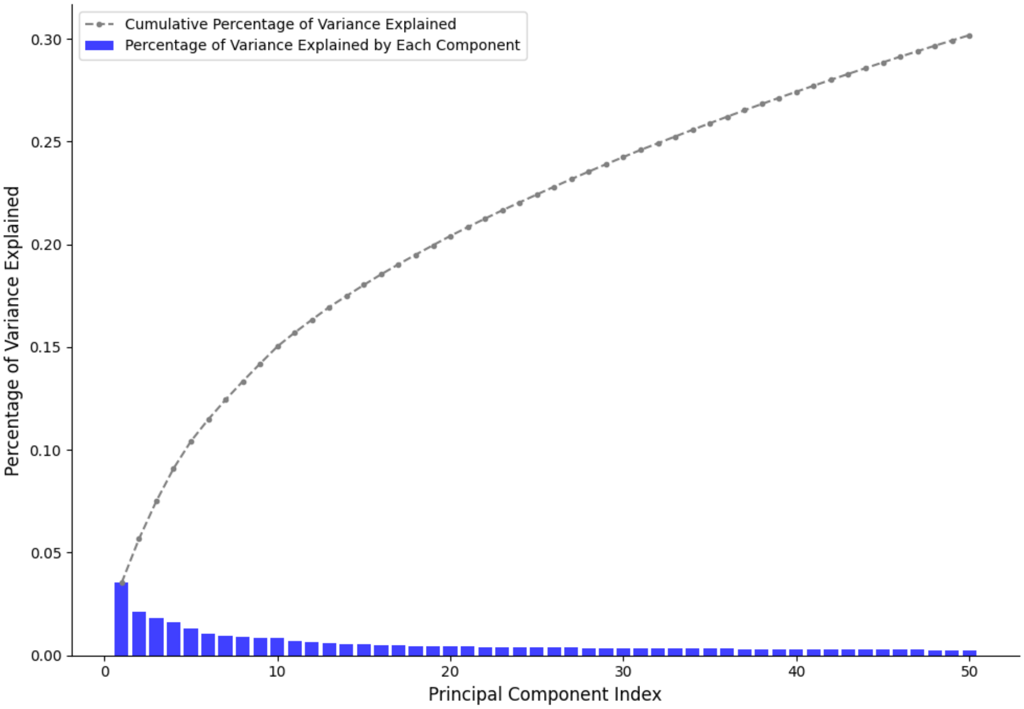
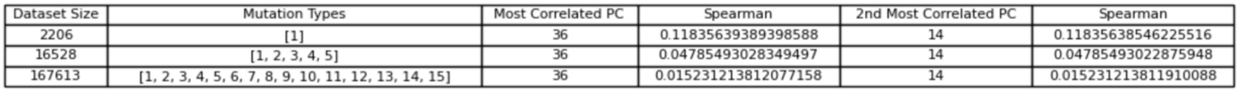


Figure 1: Variance Explained in Multiple Sequence Alignment by Principal Components

The 167,613 variants of CreiLOV were one-hot encoded, flattened into a 2-dimensional matrix, and transformed into the principal component space. The most correlated principal components were PC36 and PC14. The correlation decreased as the number of mutations increased (Table 1). Interestingly, PC36 and PC14 discretely separate the best performing variants (Figure 2).

Table 1: Correlation of Principal Components with Fluorescence



Dataset Size	Mutation Types	Most Correlated PC	Spearman	2nd Most Correlated PC	Spearman
2206	[1]	36	0.11835639389398588	14	0.11835638546225516
16528	[1, 2, 3, 4, 5]	36	0.04785493028349497	14	0.04785493022875948
167613	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]	36	0.015231213812077158	14	0.015231213811910088

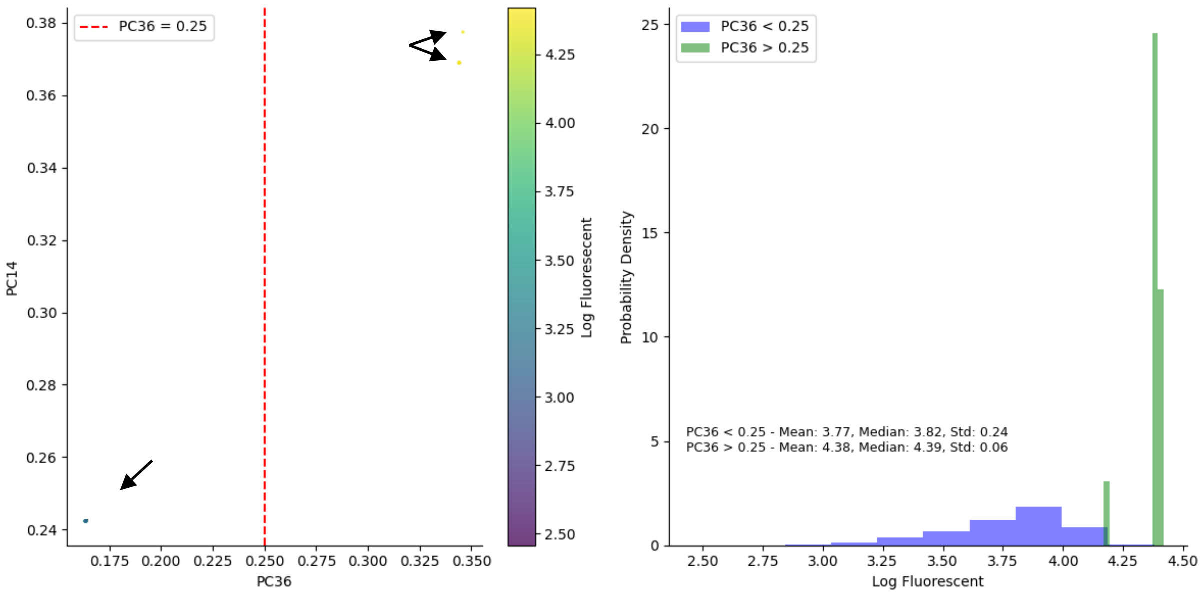


Figure 2: Principal Components Separate High-Performance CreiLOV Variants

The separated variants (with PC36 > 0.25 or PC14 > 0.3) contained the 13 most fluorescent proteins in the experimental dataset, including CreiLOV and 12 single mutation variants that interestingly almost spanned the entire length of the CreiLOV (Figure 3).

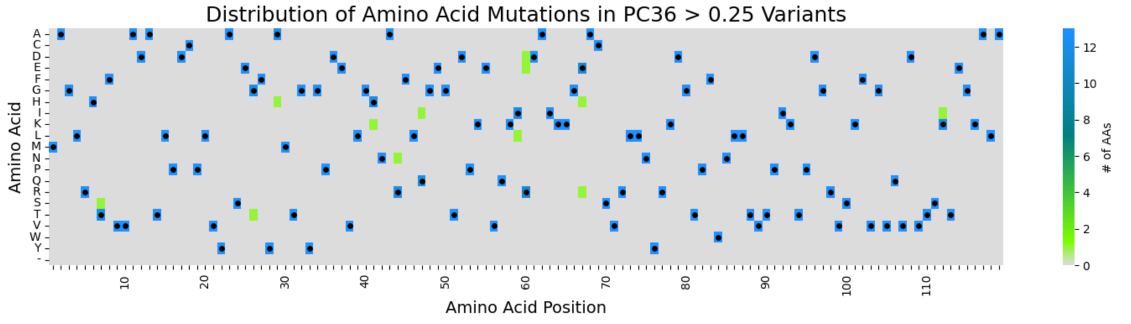
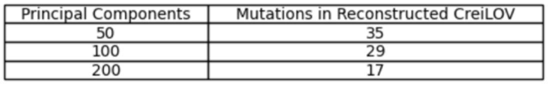


Figure 3: Distribution of Mutations in Separated Variants in Principal Component Space

The black dots represent the original CreiLOV sequence. The green squares display mutations that occurred in the separated variants in principal component space.

The original CreiLOV sequence was projected into the principal component space and transformed back into the original space using 50, 100, or 200 principal components. Sampling the amino acids with the greatest value in the reconstructed space resulted in poor reconstruction of the CreiLOV sequence (Table 2).

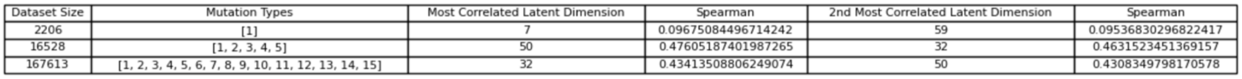
Table 2: Reconstruction Accuracy of CreiLOV Sequence using Principal Components



Principal Components	Mutations in Reconstructed CreiLOV
50	35
100	29
200	17

The VAE was trained on one-hot encoded 243,582 unlabeled natural protein sequences with corresponding phylogenetic weights in the objective function to account for uneven sampling and phylogenetic bias during stochastic gradient descent [8,11]. The data was split into training and validation sets with a 90/10 split. A VAE architecture previously optimized for the smallest final validation reconstruction loss was deployed. The most correlated latent dimensions from the VAE latent space were the 50th and 32nd latent dimensions, achieving up to a spearman correlation of 0.476 (Table 3). The 50th latent dimension had a Pearson correlation with variants containing 1-5 mutations of 0.42 as well (Figure 4).

Table 3: Correlation of VAE Latent Dimensions with Fluorescence



Dataset Size	Mutation Types	Most Correlated Latent Dimension	Spearman	2nd Most Correlated Latent Dimension	Spearman
2206	[1]	7	0.09675084496714242	59	0.09536830296822417
16528	[1, 2, 3, 4, 5]	50	0.47605187401987265	32	0.4631523451369157
167613	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]	32	0.43413508806249074	50	0.4308349798170578

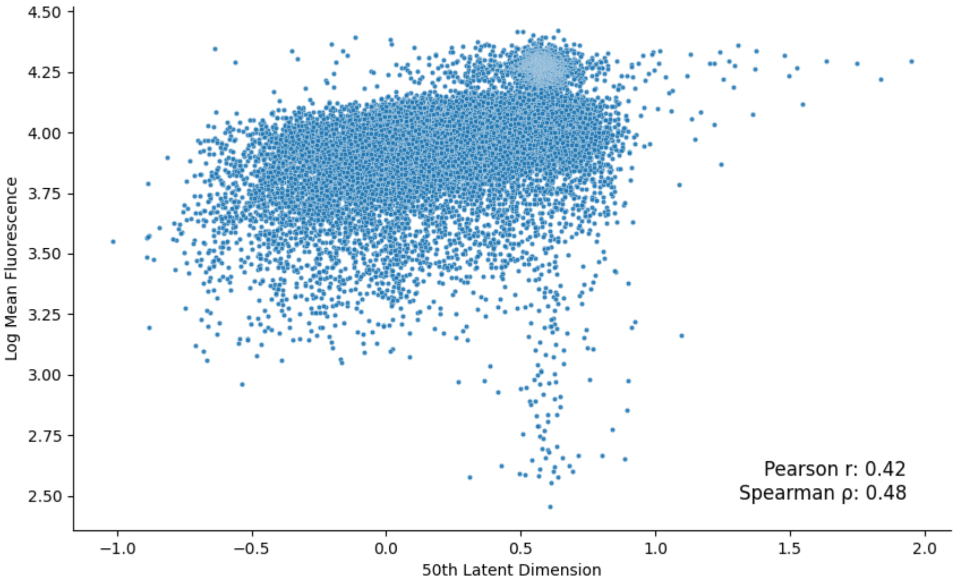


Figure 4: Correlating the 50th Latent Dimension of VAE and Fluorescence

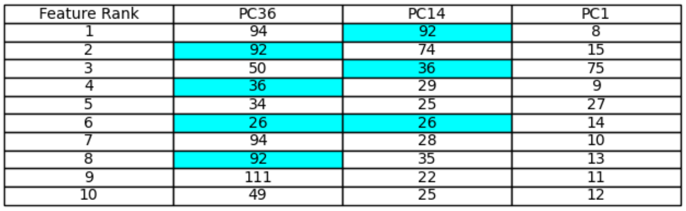
The original CreiLOV sequence was encoded into the 64-dimension latent space of the VAE and decoded back into the original dimensions. Sampling the amino acids with the greatest value in the reconstructed space resulted in the reconstruction of the CreiLOV sequence with only the first 2 amino acids and last 4 amino acids being incorrect.

Discussion

The top 50 principal components fail to capture the variance in the natural sequences perhaps because amino acid mutations affect many protein properties and therefore evolve with numerous, complicated constraints or because PCA assumes a linear independence between features and cannot effectively model complex evolutionary constraints.

PC14 and PC36 discretely separated the most fluorescence variants, but the separation was more discrete than continuous (Figure 2). The principal components would not be robust for predictive regression tasks such as scoring sequences during simulated annealing, further evident by low spearman correlations (Table 1). The number of mutations in variants and spearman correlation were negatively correlated likely because PCA is not designed to capture non-linear relationships and cannot model the non-linear interactions between amino acids. This may be why PCA is unable to effectively reconstruct sequences using the principal components as well (Table 2). Interestingly however, several one-hot encoded amino acid positions heavily contributed to the principal components most correlated with fluorescence (Table 4).

Table 4: Feature Contribution to PC14 and PC36



Feature Rank	PC36	PC14	PC1
1	94	92	8
2	92	74	15
3	50	36	75
4	36	29	9
5	34	25	27
6	26	26	14
7	94	28	10
8	92	35	13
9	111	22	11
10	49	25	12

The 10 amino acid positions that contribute the most to PC14, PC36, and PC1 are shown with the amino acid positions shared by PC14 and PC36 highlighted in blue

The 92nd amino acid is often mutated in the 167,600 variants with lower fluorescence, but the 92nd amino acid is conserved for all variants with the greatest fluorescence. The the 92nd amino acid may be in the active site or necessary for fluorescence. We can use this hypothesis to conserve the 92nd amino acid during simulated annealing to generate protein designs more likely to be fluorescent. Structural information must be obtained first to assess if the correlation is biological relevance or spurious. The amino acid positions in PC1 may be more correlated with protein stability than fluorescence as stability strongly influences evolutionary trajectories of proteins, but increasing protein stability can have a tradeoff with protein function.

The VAE proves superior for the reconstruction of sequences. The only mutations arising upon compression and reconstruction were at the beginning and end of the protein where fewer evolutionary pressures are often present, and these amino acids are often less related to protein function. Several VAE latent dimensions proved more correlated with fluorescence with a brief analysis, confirming an ability to score sequences during simulated annealing (Table 3, Figure

4). While the VAE outperforms PCA for scoring sequences using correlated latent dimensions and generating protein designs, the VAE is significantly more difficult to interpret.

This paper covers a brief assessment of the predictive and generative capabilities of PCA and VAE for protein design. While the VAE outperforms PCA for predictive and reconstruction tasks during this brief assessment, interpretation of PCA results if experimentally validated may prove meaningful for protein design and further comparisons with different MSA and DMS datasets can provide more robust conclusions.

References

[1] Gelman, S., Fahlberg, S.A., Heinzelman, P., Romero, P.A. and Gitter, A., 2021. Neural networks to learn protein sequence–function relationships from deep mutational scanning data. *Proceedings of the National Academy of Sciences*, *118*(48), p.e2104878118.

[2] Freschlin, C.R., Fahlberg, S.A. and Romero, P.A., 2022. Machine learning to navigate fitness landscapes for protein engineering. *Current Opinion in Biotechnology*, *75*, p.102713.

[3] Fahlberg, S.A., Freschlin, C.R., Heinzelman, P. and Romero, P.A., 2023. Neural network extrapolation to distant regions of the protein fitness landscape. *bioRxiv*, pp.2023-11.

[4] Hsu, C., Nisonoff, H., Fannjiang, C. and Listgarten, J., 2022. Learning protein fitness models from evolutionary and assay-labeled data. *Nature biotechnology*, *40*(7), pp.1114-1122.

[5] Wu, Z., Johnston, K.E., Arnold, F.H. and Yang, K.K., 2021. Protein sequence design with deep generative models. *Current opinion in chemical biology*, *65*, pp.18-27.

[6] Weinreich, D.M., Delaney, N.F., DePristo, M.A. and Hartl, D.L., 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. *science*, *312*(5770), pp.111-114.

[7] Salverda, M.L., Dellus, E., Gorter, F.A., Debets, A.J., Van Der Oost, J., Hoekstra, R.F., Tawfik, D.S. and de Visser, J.A.G., 2011. Initial mutations direct alternative pathways of protein evolution. *PLoS genetics*, *7*(3), p.e1001321.

[8] Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A. and Bikard, D., 2021. Generating functional protein variants with variational autoencoders. *PLoS computational biology*, *17*(2), p.e1008736.

[9] Ding, X., Zou, Z. and Brooks III, C.L., 2019. Deciphering protein evolution and fitness landscapes with latent space models. *Nature communications*, *10*(1), p.5644.

[10] Johnson, L.S., Eddy, S.R. and Portugaly, E., 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics*, *11*, pp.1-8.

[11] Riesselman, A.J., Ingraham, J.B. and Marks, D.S., 2018. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, *15*(10), pp.816-822.

[12] Chen, Y., Hu, R., Li, K., Zhang, Y., Fu, L., Zhang, J. and Si, T., 2023. Deep Mutational Scanning of an Oxygen-Independent Fluorescent Protein CreiLOV for Comprehensive Profiling of Mutational and Epistatic Effects. *ACS Synthetic Biology*, *12*(5), pp.1461-1473.

Code Availability

https://github.com/nblalock/PCA_vs_VAE.git