ESMS VAE: A Structure-Informed Variational Autoencoder for Protein Engineering

Sungmin Ahn[1*], Shihyun Moon[1], Jooyoung Jung[1], Minjae Lee[1], Jeongsu Park[1]

[1]Daegu Science High School, 154 Dongdaegu-ro, Suseong-gu, Daegu, Republic of Korea.

*Corresponding author(s). E-mail(s): ahnd6474@gmail.com;

Abstract

Sequence-based protein Variational Autoencoders (VAEs) have definitive limits on generalization, reconstruction ability, and capturing meaningful information within the latent space. To overcome these limits, we introduce ESMS VAE, a novel VAE trained with a custom loss function designed to capture structural similarity. ESMS VAE achieved a 97.17% reconstruction rate on a test set randomly sampled from UniRef50. This high reconstruction rate was maintained under small noise conditions and increased by approximately 1% with heavy noise. To test its ability to generate novel sequences, we randomly sampled and decoded latent vectors. The decoded sequences showed an identity of around 10%. The latent space's capability in downstream tasks was evaluated on fluorescent proteins (FPs). It demonstrated excellent performance as an embedding for classifying FPs and non-FPs, with a 0.987 5-fold cross-validation accuracy, and for regressing excitation and emission wavelengths, achieving RMSE values of 2.7 nm and 3.8 nm, respectively, using Gaussian Process (GP) models. FPs were projected onto the latent space and were clustered using K-means. Consensus vectors were calculated and decoded, generating sequences that largely preserved their function.

Keywords: Variational Autoencoder, Protein Engineering, Deep Learning, Generative Models, Structural Biology, ESMS

1 Introduction

A Variational Autoencoder (VAE) is a generative model capable of generating similar but different data from the input. It is usually applied to images and is used to reconstruct and generate images [1]. In protein science, VAEs have been suggested to guide the exploration of the vast and empty protein space with their latent space [1]. A VAE is made up of an encoder and a decoder. The encoder compresses data and embeds it in a latent space while the decoder reconstructs the original data from the latent space. The model is trained in a manner that minimizes information loss while compressing data and fits latent space variables to a normal distribution. However, VAEs are susceptible to issues such as posterior collapse, where the decoder ignores the latent space and generates the most common token, and KL vanishing, where an overly perfect fit of the latent space variables to a Gaussian distribution causes a loss of information in the latent space [2]. Both phenomena are associated with very low KL values, which is why a KL value of around 0.05 is often recommended.

1.1 Protein VAE

It has been demonstrated that the continuous latent space of a VAE can be applied to latent space interpolation and objective-driven amino acid optimization. Early models like DeepSequence utilized a VAE that takes a Multiple Sequence Alignment (MSA) as input [3]. It was used to predict mutations, and its latent space could be used as an embedding for other machine-learning tasks. Subsequent protein VAEs continuously evolved by taking a single sequence as input [4]; however, sequences contain smaller information compared to MSAs and these models usually showed moderate reconstruction rates. A notable advance is ProT-VAE, which uses ProT5 embeddings and a transformer architecture to capture long-distance relations between amino acids [5]. ProT-VAE showed a nearly 100% reconstruction rate on fine-tuned protein families.

1.2 ESM2 and ESMS

ESM-2 is a large language model released by Meta AI, trained with masking, and is capable of capturing long and short-distance interactions within a sequence [6]. ESM-2 was used as an embedding for ESMFold, a 3D structure prediction model known to have similar capabilities to AlphaFold 2 [7, 8]. It is commonly used as an embedding for various tasks like secondary structure prediction and thermostability prediction. Although its ability to capture structural information is undisputed, it is heavy. Needing a lighter model, Knowledge Distillation was applied to ESM embeddings. A new small transformer model, ESMS, was trained to mimic ESM2's embedding direction and size. Direction and size were each calculated with cosine similarity and RMSE. Cosine similarity and RMSE on the test set were 0.9647 and 1.2998.

2 ESMS VAE

With ProT-VAE, it has been shown that transformer architectures are capable of capturing hidden information in a sequence. Also, rather than training big transformer models, using a pre-trained model as an embedding was demonstrated to be effective.

However, only training with a sequence has clear limits on generalization, function, and versatility. It is clear that structural information must be introduced to train a fully functional VAE. ESMS VAE, in order to overcome such limits, introduced a custom loss function that explicitly forces the VAE to learn and represent structural information in the latent vector space.

2.1 Custom Loss Function and Architecture

To capture structural information, the ESMS model was used to calculate loss. ESMS, the student model of ESM2 650M, learned embeddings that capture structural information and were used to calculate the difference in structure between the original (origin) and reconstructed (recon) sequence. This way, the latent vector space was trained to capture structural information. The full ESM2 model could not be used because of limits in computational resources; thus, as a substitute, ESMS was chosen. ESMS is suitable for this task because it was trained with masking and shows how likely an amino acid will take place in a certain position. If A and B were both likely to be in a certain position, A and B would have similar embeddings. This is not only because they play a similar role and can be replaced, but also because they both cannot be replaced with certain other amino acids in that position. Thus, structural loss will not penalize amino acids that are likely to be substituted but will penalize those that are not. This also contributes to preventing posterior collapse, which usually occurs by learning the most common amino acid and relying on it. If such an event were to occur, structural loss would penalize it highly and force the VAE to rely on the latent space.

The composite loss function is defined as:

$$\begin{gathered} L = \lambda(L_{\text{MSE}} + L_{\text{COS}}) + \alpha \cdot L_{\text{CE}} + \beta \cdot L_{\text{KL}} \qquad (1) \\ \text{where } L_{\text{COS}} = 1 - \text{COS}(\text{ESMS}(\text{origin}), \text{ESMS}(\text{recon})) \\ \text{and } L_{\text{MSE}} = \text{MSE}(\text{ESMS}(\text{origin}), \text{ESMS}(\text{recon})) \end{gathered}$$

The weights are set as $\lambda = 5$, $\alpha = 30$ for epochs < 100 (then 0.1), and $\beta = 100$ for epochs < 100 (then 1).

ESMS VAE is a comparably lightweight transformer with 5.5M parameters, composed of 4-layer transformer encoders and decoders. The hyperparameters are detailed in Table 1. The Adam optimizer was used.

3 Training and Model Selection

The model was trained with two T4 GPU sessions provided by Kaggle, using a subsample of the UniRef50 dataset. Monitored learning curves did not show any signs of overfitting (Figure 1).

The model saved at epoch 500 (VAE_500) had a final reconstruction rate of 99.976%. Validation loss values were: Val CE=0.000, COS=0.003, MSE=0.007, and KL=0.002. It exhibited very small latent space values and the very low KL value suggested potential KL Vanishing (Figure 2). When noise was added to the latent space, it did not show significant changes in the reconstruction rate.

Table 1: Hyperparameters for ESMS VAE.

Parameter	Value
Vocab Size	33
d_model	256
Latent Dim	256
n_heads	4
Feed Forward	512
Dropout	0.3

Instead, the model from epoch 380 (VAE_380) was chosen because it had a KL value closer to the recommended active value of 0.05 and had the lowest validation Cross-Entropy (CE) loss. At epoch 380, the validation losses were: Val CE=0.072, COS=0.010, MSE=0.020, and KL=0.048. VAE_380 achieved a final reconstruction accuracy of 97.17% on a test set, demonstrating that ESMS VAE generalizes well to different kinds of protein (Figure 3).

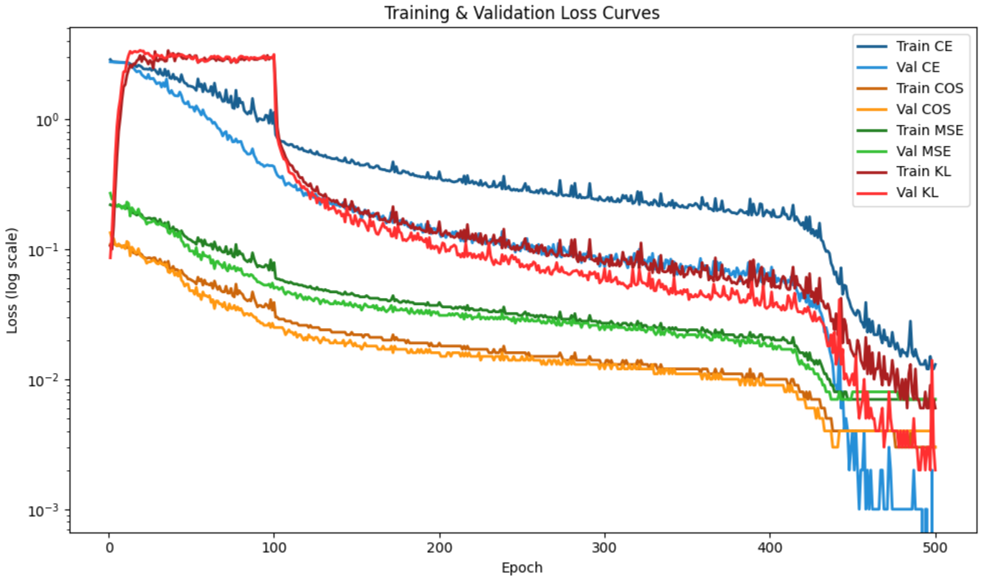
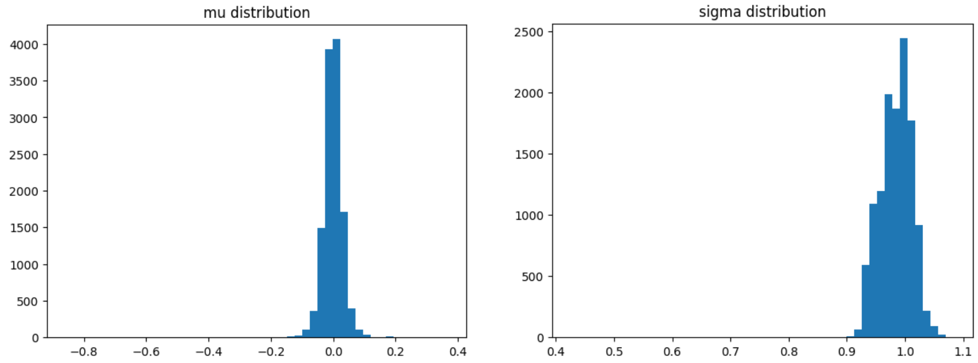


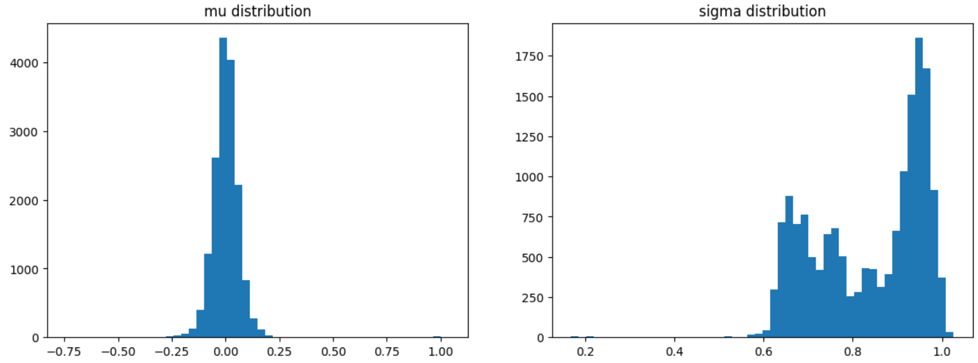
Fig. 1: Training and validation loss curves for Cross-Entropy (CE), Cosine Similarity (COS), Mean Squared Error (MSE), and KL Divergence (KL) over 500 epochs. The y-axis is on a log scale.



(a) mu distribution at epoch 500

(b) sigma distribution at epoch 500

Fig. 2: The mean (mu) and standard deviation (sigma) distribution of the latent space at epoch 500, showing signs of KL vanishing.



(a) mu distribution at epoch 380

(b) sigma distribution at epoch 380

Fig. 3: The mean (mu) and standard deviation (sigma) distribution of the latent space at epoch 380, showing healthier distributions.

4 Latent Space Validation

To check if the latent space had learned meaningful representations, the following tests were conducted in a teacher-forced manner on completely unseen sequences.

4.1 Latent Robustness Test

VAE_380's reconstruction rate was tested after adding noise to the latent space. For a test set of about 1000 sequences, the reconstruction rate increased with noise (Table 2), suggesting the VAE is capable of distinguishing and potentially correcting large noises. This denoising capability was further tested by substituting amino acids with 'X' in a random sample of 100 sequences. The ESMS cosine similarity between the original and reconstructed sequences was 0.9592, which is very close to 1, meaning the VAE is capable of identifying and eliminating noises while not altering structure.

Table 2: Reconstruction accuracy under noisy conditions.

Noise Level (σ)	Mean Reconstruction Accuracy
0.1	97.77%
1.0	98.34%
2.0	98.58%

4.2 Posterior Collapse Test

The KL divergence per dimension was checked to ensure the model had not suffered from posterior collapse. The mean KL was 0.04998 with an RMSE of 0.07027. A KL value around 0.05 is considered active, indicating that the latent space is being utilized by the decoder.

5 Latent Space Application on Fluorescent Proteins (FP)

The utility of the learned latent space was evaluated on downstream tasks involving Fluorescent Proteins (FPs). GFP sequences were collected by hand from FPbase [9].

5.1 FP Classification and Wavelength Estimation

The latent space of the VAE was applied to the classification of FPs and non-FPs, and to estimating maximum absorption and emission wavelengths. A classic Gaussian Process (GP) model was used as the model head [10]. The GP Classifier (GPC) had a 0.987 5-fold CV accuracy. The GP Regressor (GPR) had an RMSE of 2.70 nm for absorption wavelength (A_abs) and 3.80 nm for emission wavelength (xem). The classification reports on the train and test sets show the model performs well and is not overfit (Tables 3 and 4).

Table 3: Classification report on the training set.

	Precision	**Recall**	**F1-score**	**Support**
Non-FP (0)	0.9920	0.9940	0.9930	501
FP (1)	0.9880	0.9840	0.9860	250
Accuracy			0.9907	751
Macro Avg	0.9900	0.9890	0.9895	751
Weighted Avg	0.9907	0.9907	0.9907	751

Table 4: Classification report on the test set.

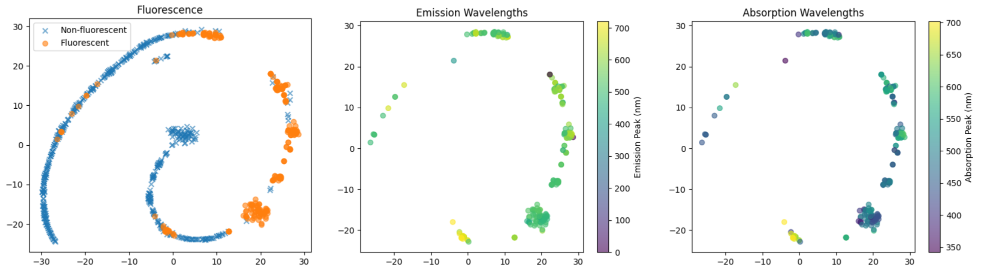
	Precision	**Recall**	**F1-score**	**Support**
Non-FP (0)	0.9840	0.9840	0.9840	125
FP (1)	0.9683	0.9683	0.9683	63
Accuracy			0.9787	188
Macro Avg	0.9761	0.9761	0.9761	188
Weighted Avg	0.9787	0.9787	0.9787	188

A t-SNE map [11] shows that the latent space successfully captures structural information, separating the two classes very well. Non-fluorescent proteins are on the inner side of two curves and fluorescent proteins are on the outside. Visualizations for emission and absorption wavelengths show continuous color gradients, meaning the latent space has separated proteins with different wavelengths and gathered proteins with similar wavelengths together (Figure 4).

5.2 FP Generation

Collected FP sequences were projected onto the latent space and clustered using K-means. Three clusters were found, with statistics detailed in Table 5. Consensus vectors were calculated based on these clusters and decoded to generate novel sequences. Generated sequences were truncated to the mean length of their cluster since VAEs usually generate meaningless sequences after the original sequence length. These sequences were classified using the trained GPC. As shown in Table 6 and Figure 5, Cluster 1 yielded a 100% success rate. This suggests that the quality and consistency of samples within a cluster are more important than the sheer number of samples for generating new functional proteins.

The 3D structures of generated vectors were predicted using the AlphaFold server. For Cluster 1 proteins, iconic beta barrels were found, maintaining a high pLDDT above 90. Proteins from Cluster 0 have two sets containing three beta strands on the



(a) Fluorescence (b) Emission Wavelengths (c) Absorption Wavelengths

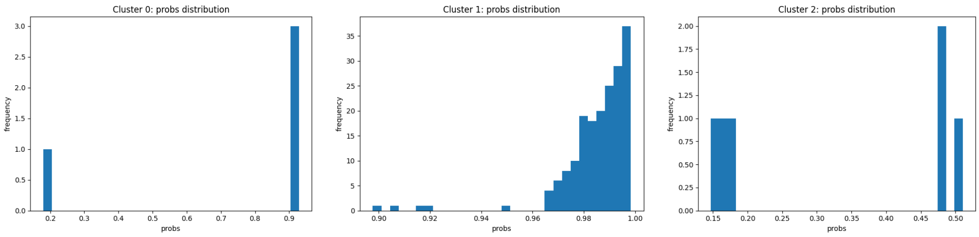
Fig. 4: t-SNE visualization of the FP latent space. (a) Clear separation of fluorescent (orange) and non-fluorescent (blue) proteins. (b, c) Continuous gradients for emission and absorption wavelengths, respectively.

Table 5: Statistics for FP clusters identified by K-means.

Cluster	n (samples)	Mean Length	Std (Length)
0	22	316	5.28
1	318	234	10.00
2	14	118	29.71

Table 6: Generation of GFP-like sequences from cluster consensus vectors.

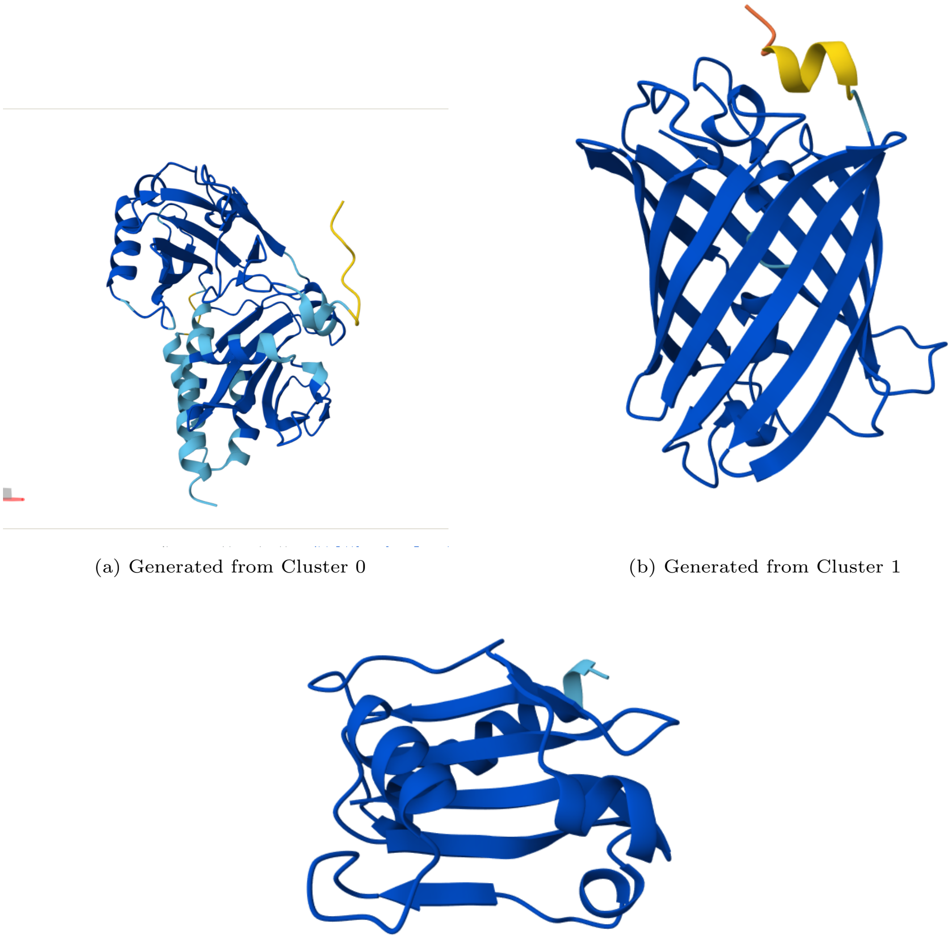
Cluster	Total Generated	Count as GFP	Ratio
0	4	3	0.7500
1	181	181	1.0000
2	6	1	0.1667



(a) Cluster 0 Probs (b) Cluster 1 Probs (c) Cluster 2 Probs

Fig. 5: Probability distribution of generated sequences being classified as GFP for each cluster. Cluster 1 shows high-confidence predictions.

inside covered by alpha helices. Proteins from Cluster 2, although not showing a beta-barrel structure, also show structural similarity among themselves, with three beta strands surrounded by alpha helices. This confirms that close vectors in the latent space encode for proteins with similar structures (Figure 6).



(a) Generated from Cluster 0

(b) Generated from Cluster 1

(c) Generated from Cluster 2

Fig. 6: Predicted 3D structures of generated proteins from different clusters show high intra-cluster structural similarity. Cluster 1 notably forms the GFP-like beta barrel.

5.3 In Vitro test of GFP

Three sequences were chosen for future in vitro validation based on: 1) GPC prediction, 2) structural likeliness to an original sequence, and 3) whether it needs other fluorescent molecules. Two sequences were chosen from Cluster 1 and one from Cluster 2. The selected sequences are:

```
>seq_cluster1_1
MASTPFKFQLKGTINGKSFTVEGEGEGNSHEGSHKGKYVCTSGKLPMSWAALGTS
FGYGMKYYTKYPSGLKNWFHEVMPEGFT
>seq_cluster1_2
MVSTGEELFTGVVPFKFQLKGTINGKSFTVEGEGEGNSHEGSHKGKYVCTSGKLP
MSWAALGTSFGYGMKYYTKYPSGLKNWFF
>seq_cluster2_1
MPRISDKLMKTRWRGFHSIPSIPPDLGGIYGIGEKTSRRKTTEHLYTGRAKDIKS
RLMKHKYGHQAIDRKIRSNIKQKKLSDLRFKFVE
```

6 Novelty of ESMS VAE

ESMS VAE is the first VAE to be trained with a loss function that considers structural information. This loss function contributes to a meaningful latent space, which leads to a generalized and stable function of the VAE. ESMS VAE, unlike DeepSequence, uses a sequence as its only input and does not require MSA alignment. While ProT-VAE relies on ProT5 embedding to learn structural information, ESMS VAE is trained explicitly to capture structural information. ProT-VAE lacks the ability to generalize without additional training; however, ESMS VAE generalizes very well on different types of proteins.

7 Conclusion and Future Directions

ESMS VAE has solved many problems, like posterior collapse, and learned a meaningful latent space. The latent space of ESMS VAE learned structural information; thus, like protein space itself, most of the latent space would hold functionless protein. The latent space of ESMS VAE is still too vast, and functional proteins are too scarce. It is highly recommended to use a regressor and a classifier together. The regressor is used to predict your objective, and the classifier will identify a functional protein. So, the regressor will guide your search while the classifier limits your search. This combination will be critical for efficiently navigating the latent space to engineer novel proteins.

Acknowledgements

This research was supported by Daegu Science High School.

References

[1] Greener, J. G., Sternberg, M. J. E. & Jones, D. T. Protein sequence-based inference of structure, function and stability using a deep vae. *Nature Communications* **9**, 3095 (2018).

[2] Lucas, J., Tucker, G., Grosse, R. & Norouzi, M. Understanding posterior collapse in generative latent variable models. *Journal of Machine Learning Research* **21**, 1–39 (2019).

[3] Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deepsequence: Deep generative models of genetic variation capture the effects of mutations. *Nature Methods* **15**, 816–822 (2018).

[4] Sinai, S., Hochreiter, S. & Obermayer, K. Lc-vae: Latent-consensus variational autoencoder for protein sequences. *arXiv preprint arXiv:1712.03346* (2017).

[5] Elnaggar, A., Heinzinger, M., Dallago, C. *et al.* Prottrans: Towards cracking the language of life's code through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[6] Rives, A., Meier, J., Sercu, T. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

[7] Lin, Z., Akin, H., Rao, R. *et al.* Esmfold: End-to-end single-sequence protein structure prediction. *bioRxiv* (2022).

[8] Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).

[9] Lambert, T. J. *et al.* Fpbase: a community-editable fluorescent protein database. *Nature Methods* **19**, 169–170 (2022).

[10] Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (MIT Press, 2006).

[11] van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).