Dense representation of the functional protein space with WAEs

Julien Horwood, Basile Dura and Nicolas Bélanger

Université de Montréal

Introduction

Proteins play a major role in the way living organisms develop, and their sequencing is a very active field of research. However, much like most sequences of letters do not form a word, an overwhelmingly large proportion of proteins do not have a particular function. In this work, we aim to find a dense low-dimensional representation of the otherwise sparse space of functional proteins by leveraging the Wasserstein Auto-Encoder framework. Wasserstein Auto-Encoders represent a very general new class of Auto-Encoders which derive their optimization objective from optimal transport theory. The result of this objective is a regularization pushing the encoder's *marginal* distribution towards the latent space prior, which differs from the element-wise regularization of Variational Auto-Encoders and provides smoother embeddings.

Definitions

- $\mathcal{X}$: Data space (protein sequences)
- P_X: True distribution of data
- P_G: Model distribution of data

We would like to build a generative model G with $P_G \approx P_X$. We use a distance measure between probability distributions taken from *optimal transport theory*. Given a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathcal{R}_+$, the optimal transport problem $W_c(P_X, P_G)$ is defined as finding

$$\inf_{P \in \mathcal{P}(X,\tilde{X}) | X \sim P_X, \tilde{X} \sim P_G} \mathbb{E}_{(X,\tilde{X}) \sim P}(c(X, \tilde{X})).$$

Since the model relies on an auto-encoder architecture, let's also define

- $\mathcal{Z}$: Latent space
- P_Z : Prior distribution over latent space
- $Q(Z|X)$: Encoding distribution
- $G : \mathcal{Z} \to \mathcal{X}$: Decoder mapping

Objective Reformulation

Assuming the decoder G is deterministic, and $P_G(X)$ denotes the marginal distribution of X under the given model, the optimal transport problem can be reformulated as

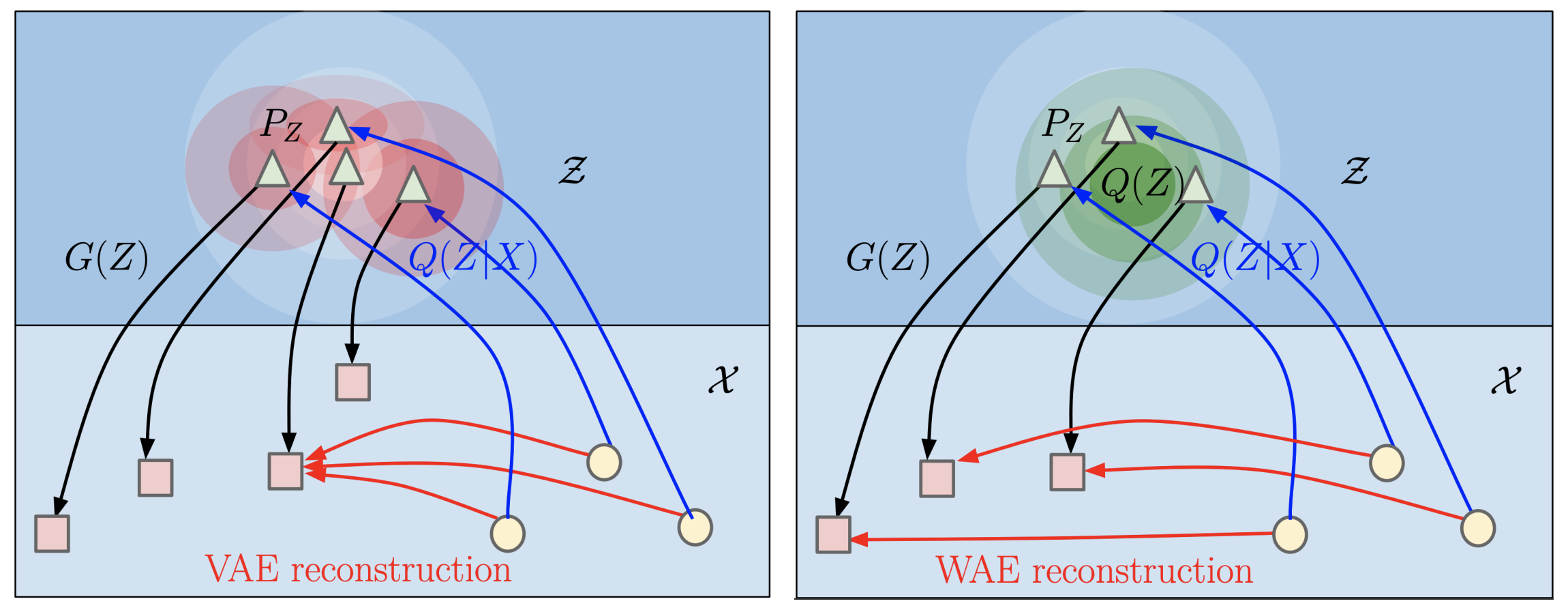
$$W_c(P_X, P_G) = \inf_{Q : Q_Z = P_Z} \mathbb{E}_{P_X}[\mathbb{E}_{Q(Z|X)}[c(X, G(Z))]]$$

The notable advantage of this reformulation is that we can optimize over the encoders instead of arbitrary joint distributions.

Constraint Relaxation

Relaxing constraint $Q_Z = P_Z$ by applying a penalty λ over a given divergence function $\mathcal{D}_Z(Q_Z, P_Z)$, the WAE objective to minimize becomes

$$\inf_{Q(Z|X)} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}[c(X, G(Z))] + \lambda \mathcal{D}_Z(Q_Z, P_Z)]$$



(a) VAE reconstruction (b) WAE reconstruction

Figure: Illustration from Tolstikhin et al. [1]

The notable difference with the VAE framework is the penalization of the marginal rather than the conditional distribution of the encoder.

WAE Algorithm

Given a regularization coefficient $\lambda > 0$, a divergence measure $\mathcal{D}_\mathcal{Z}$, a cost function c defining the distance W_c and initial parameter values of the encoder Q_ϕ and the decoder G_θ,

```
while $(\phi, \theta)$ not converged
    Sample $\{z_1, \ldots, z_n\}$ from the prior $P_Z$
    Sample $\tilde{z}_i$ from $Q_\phi(Z|x_i)$ for $i = 1, \ldots, n$
    Update $Q_\phi$ and $G_\theta$ by descending the WAE objective function
endwhile
```

MNIST Application

First, we applied our WAE implementation on the MNIST Dataset in order to check its validity by producing results obtained in the original paper, albeit in smaller dimension.

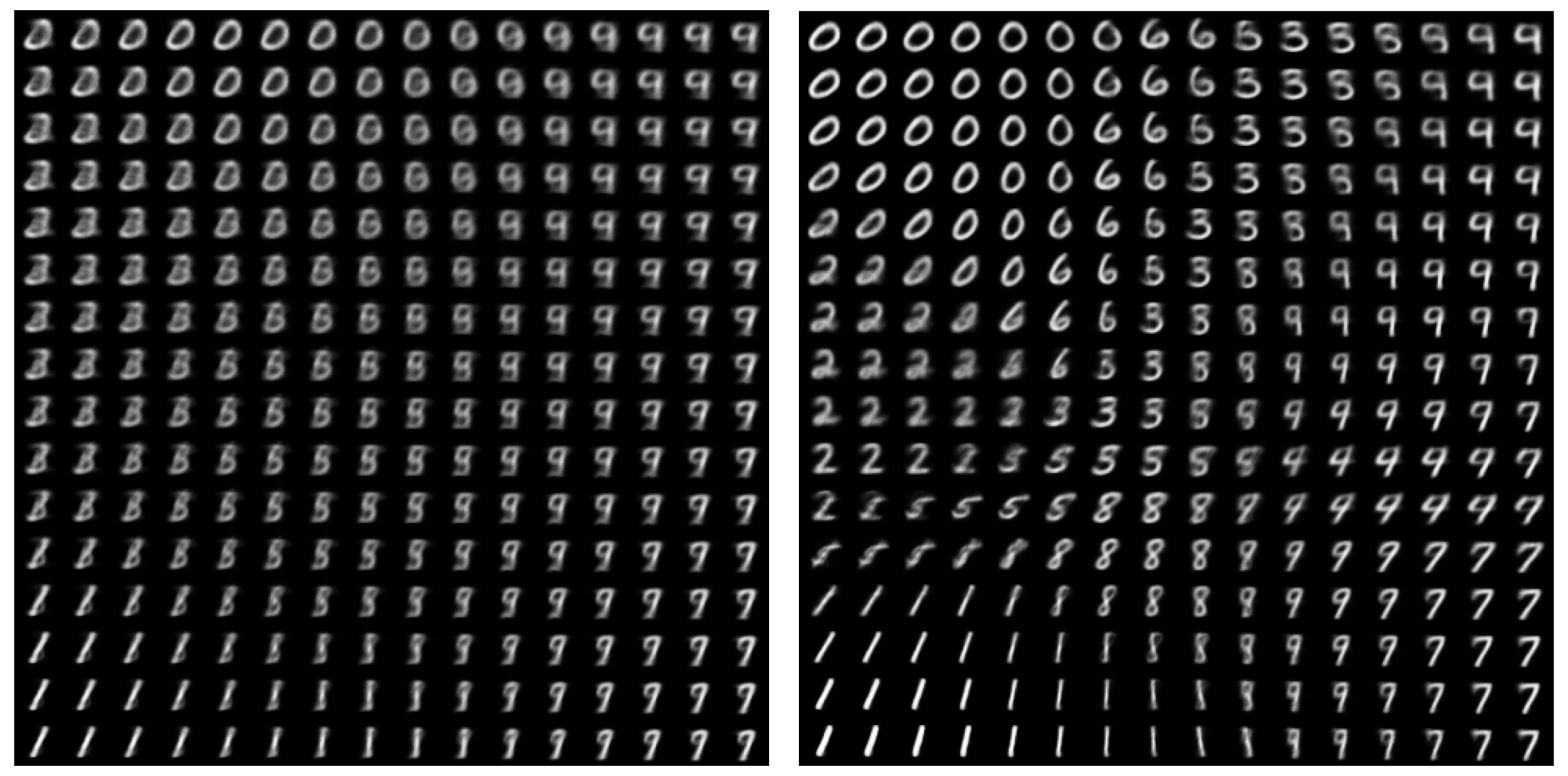
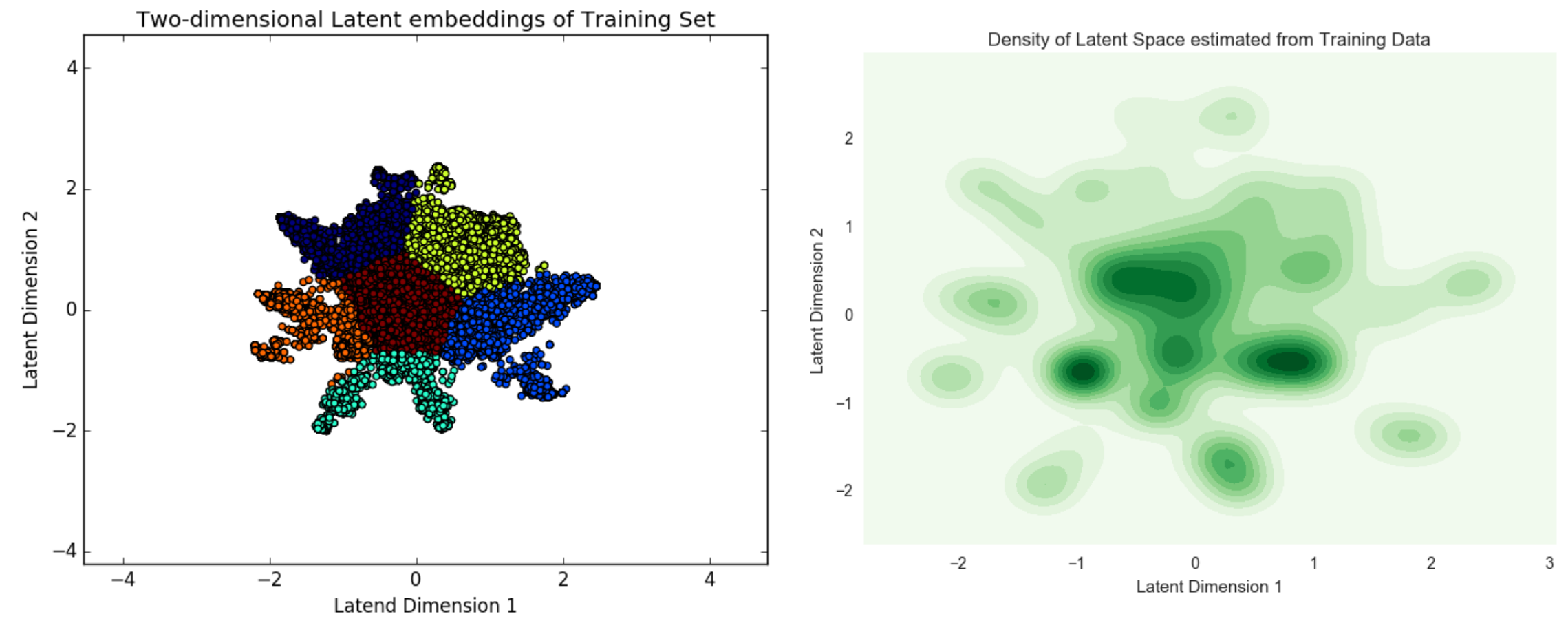


Figure: Visualization of the 2D latent space for MNIST, both before and after training

Protein Space Embedding

The biological research on protein functionality is a very costly process, notably due to the exponentially large space of protein sequences and involved testing process. The general research hypothesis is that sequences with minor mutations to functional proteins are more likely to encode biological functions. Thus it may be interesting to identify dense areas of functional proteins within the sequence space. We attempt to identify such high density functional protein regions by embedding known functional proteins into a latent space using WAEs. The ability to sample new sequences from these high-density areas of the latent space may then be of high value for guiding new research.



(a) Scatter plot (b) Density Estimation

Reconstruction Performance

Once the model is trained, we can look at its reconstruction performance on the protein set.

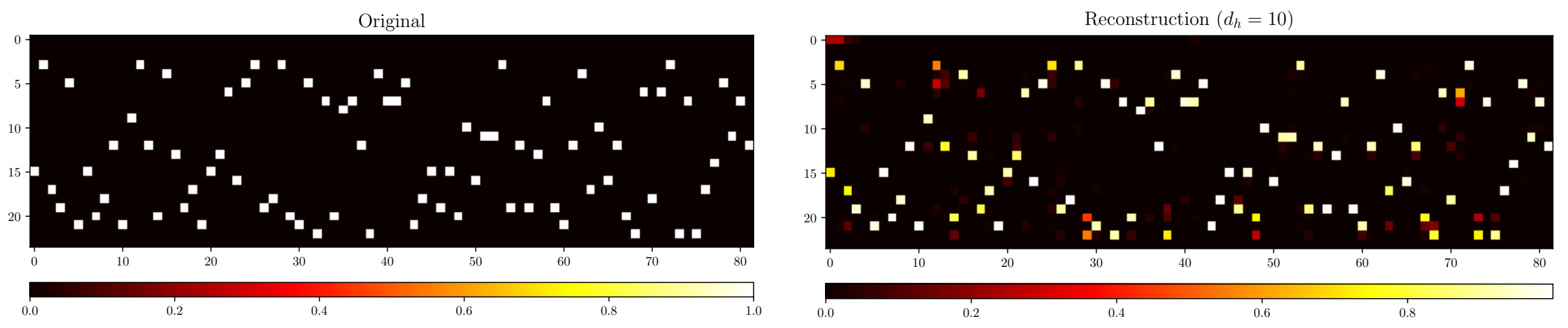


Figure: Reconstruction of the sequence by the WAE

In order to test the performance of our model for protein generation, we follow Sinai et al.'s [2] work and compare the log-probability of known sequences (output by the network) with experimentally measured *fitness*. We made this comparison using Spearman correlation.

Model	Correlation
VAE	0.271
WAE	0.311

Table: Spearman Correlation between the *fitness* and the log-probability

Conclusion

In this work, we have explored the usefulness of Wasserstein Auto-Encoders as generative models. In particular, we have demonstrated their functionality by interpolating between the digit embeddings of MNIST. Furthermore, we inspired ourselves of the work of Sinai et al. [2] on learning dense representations of the protein space by applying WAEs to this problem.

Selected References

[1] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf.
Wasserstein auto-encoders.
arXiv preprint arXiv:1711.01558, 2017.

[2] Sam Sinai, Eric Kelsic, George M Church, and Martin A Nowak.
Variational auto-encoding of protein sequences.
arXiv preprint arXiv:1712.03346, 2017.