VAE Generative Network for de novo Protein Design

BY: BERNARD MOUSSAD

UNIVERSITY LIBRARIES
VIRGINIA TECH

The Problem

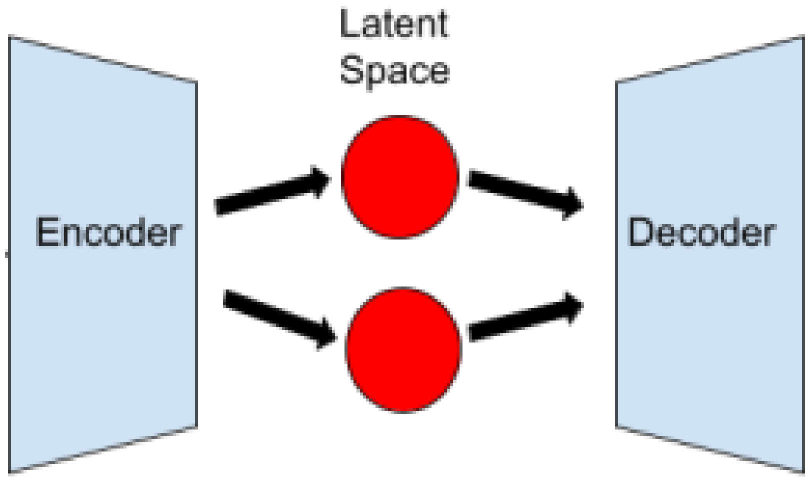
- Protein design has numerous applications that can benefit humanity
  - *Vaccinations*
  - *Enzyme Design*
- The issue is that proteins can take many conformations and shapes
- Designs by hand have been done but mainly based on known conformations

Progress in the Design Space

- Protein design has numerous applications that can benefit humanity
  - *Vaccinations*
  - *Enzyme Design*
- The issue is that proteins can take many conformations and shapes
  - Levinthal's paradox estimates 10^{300} possible conformations
- Designs by hand have been performed but mainly based on segments of known conformations

Introducing Variational Autoencoders (VAEs)

- VAEs are capable of converting high dimensional data to lower dimensions
- It is also possible to sample the latent space to generate new objects of the data it was trained on

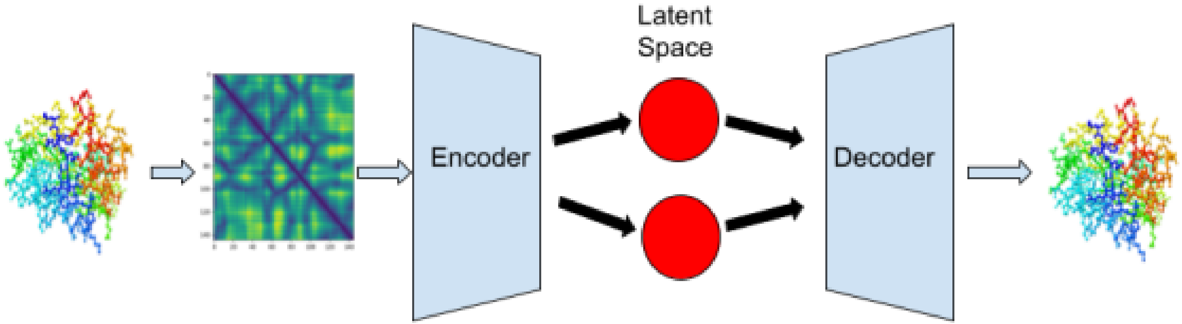


Application of VAEs to Protein Generation

- By training a VAE network on certain proteins, certain characteristics can be generated within the latent space
  - Ex: Generation of a protein capable of binding to the COVID-19 spike protein to prevent its replication

Proposal

- Train a VAE on C_{alpha} distance maps
- Reproduce distance maps fed through the network
- Produce 3D cartesian coordinates of C_{alpha} atoms instead of distance maps
- Sample latent space to view generated proteins



Metrics

- K-L Divergence (D_{KL}): Computes how similar two probability distributions are to view how much information is lost when using the encoder's distribution, Q*(x),* to represent decoder's distribution, *P(x)*

$$D_{KL}(P||Q) = \sum P(x)\log\left(\frac{P(x)}{Q(x)}\right)$$

- L2 Loss: Computes sum of the squared difference between the ground truth and the prediction

Questions