Generative models for protein design and evaluation

How to choose one generative model (stick) over the other options?

Scout Van den Bergh
Supervisor: Gaetan De Waele
Promotor: Prof. Dr. Willem Waegeman

Presentation Thesis first semester
Ghent University

Generative models for protein design and evaluation

Overview

01 Generative AI meets Proteins

02 Generative model

03 New evaluation metric

Generative AI meets Proteins

Autoregressive Models

Probability of sampling the current residue depends on previously sampled residues

Generative models for protein design and evaluation

Generative AI meets Proteins

Autoregressive Models

Probability of sampling the current residue depends on previously sampled residues

Diffusion Models

Iterative denoising mechanism, learn to predict the noise of your data.

Generative models for protein design and evaluation

Generative AI meets Proteins

Autoregressive Models

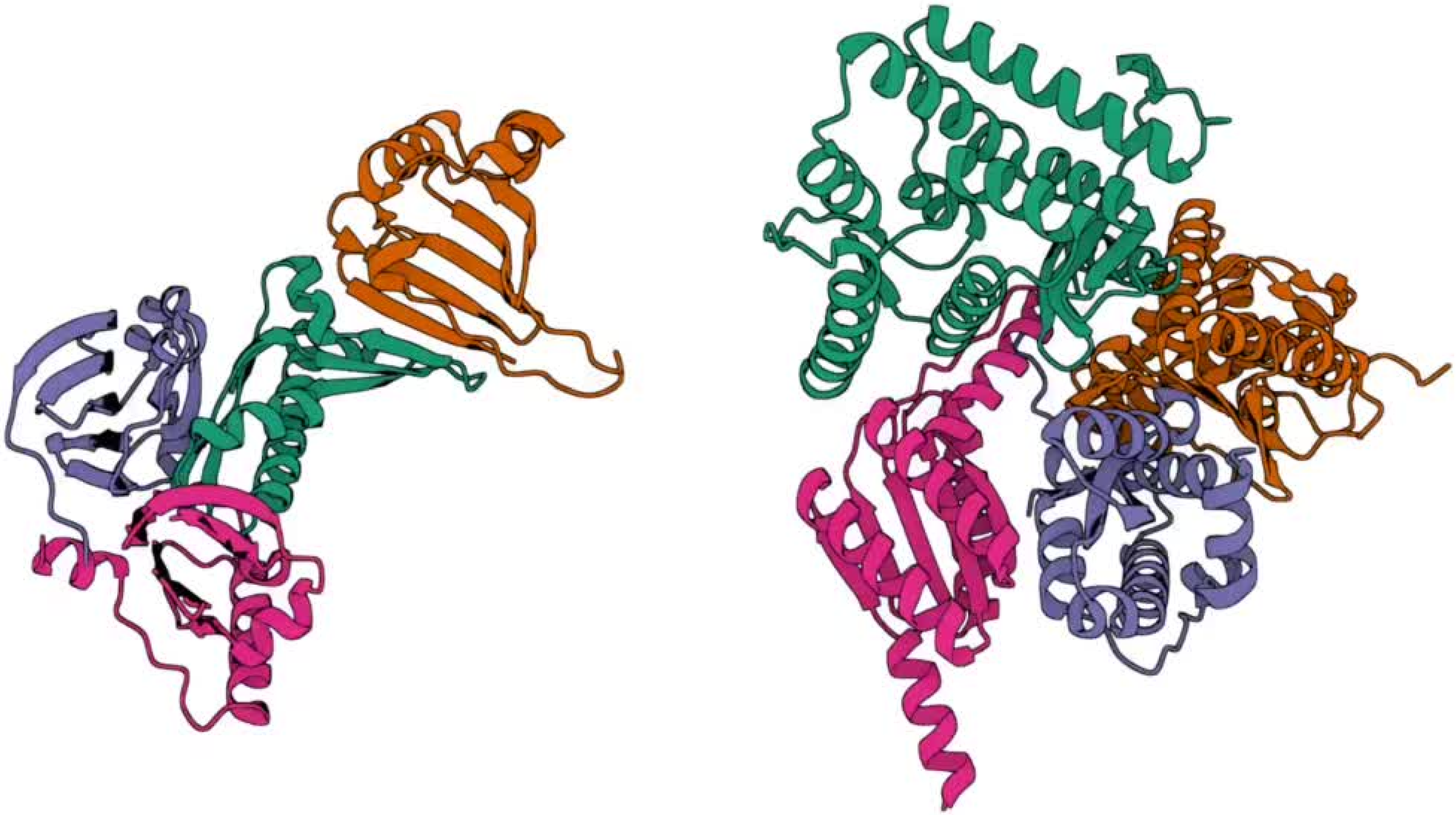
Probability of sampling the current residue depends on previously sampled residues

Diffusion Models

Iterative denoising mechanism, learn to predict the noise of your data.

Bi-directional masking protein language models

BERT-like approach.



Generative model

Dataset

Subset of Protein Data Bank (PDB)
Contains > 400k protein sequences

Generative model

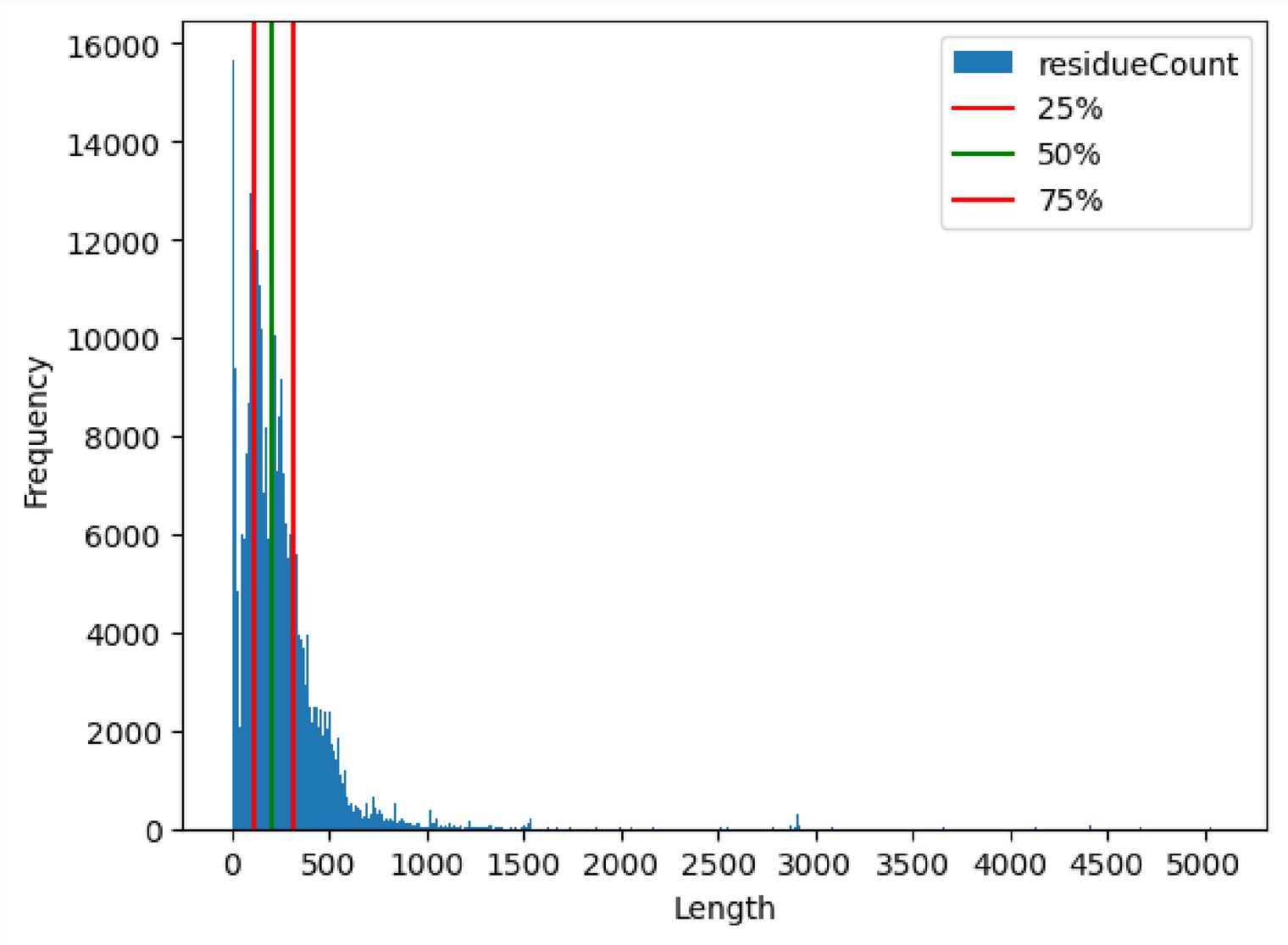
Dataset

Subset of Protein Data Bank (PDB)
Contains > 400k protein sequences

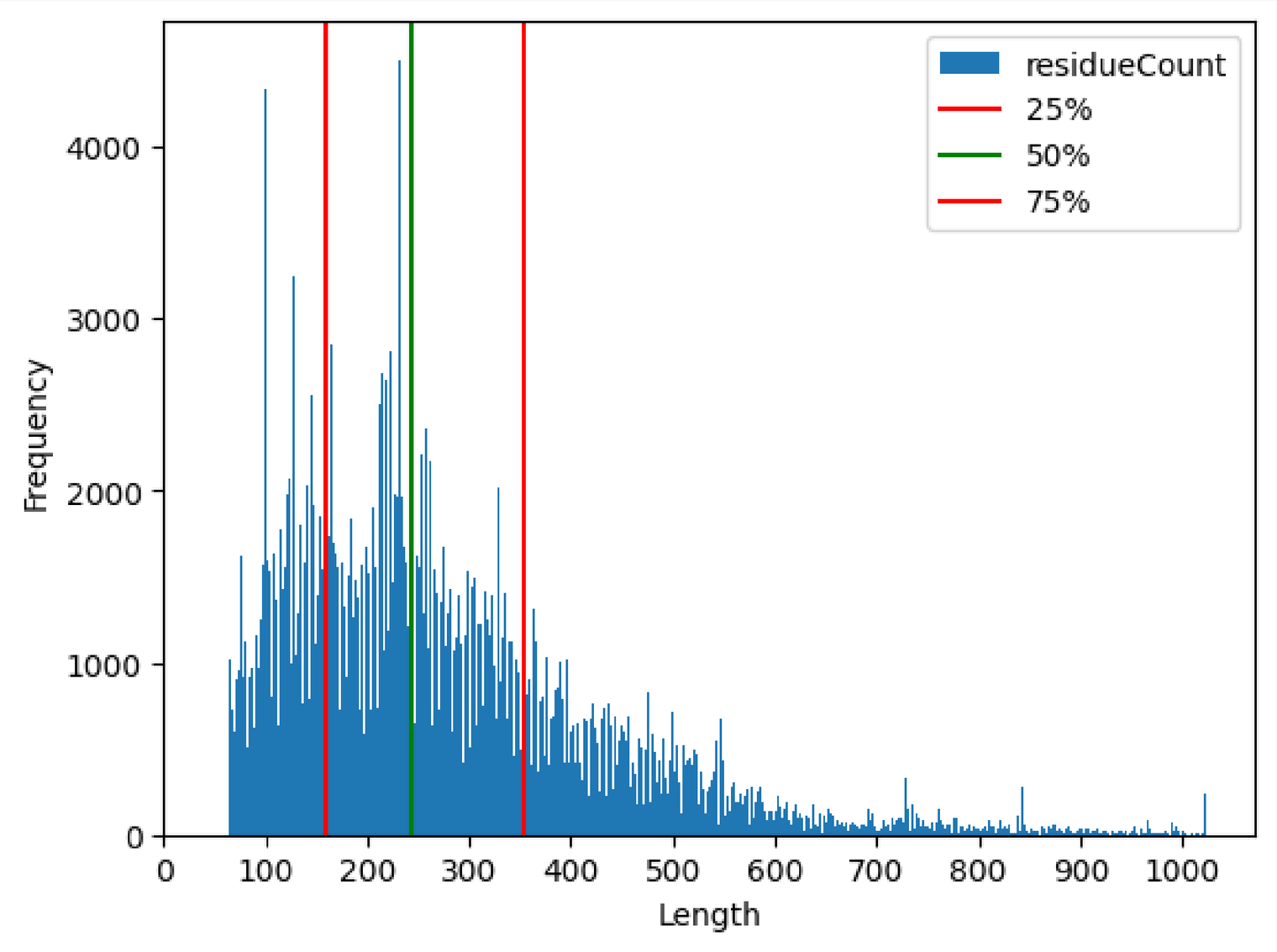
Varational Autoencoder

Learns μ of distribution
Learns σ of distribution
Reparameterization trick
Reconstruction loss + KL divergence

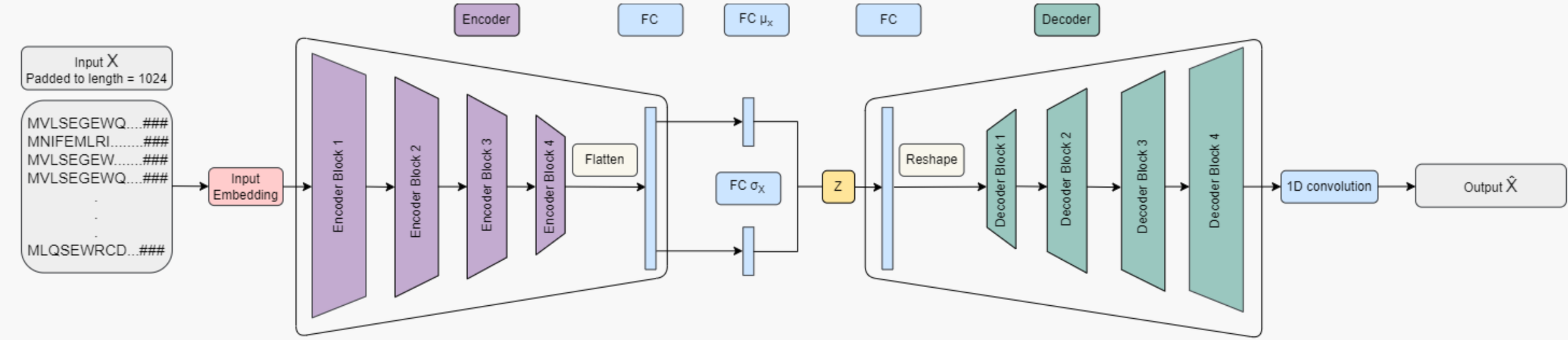
Distribution of dataset

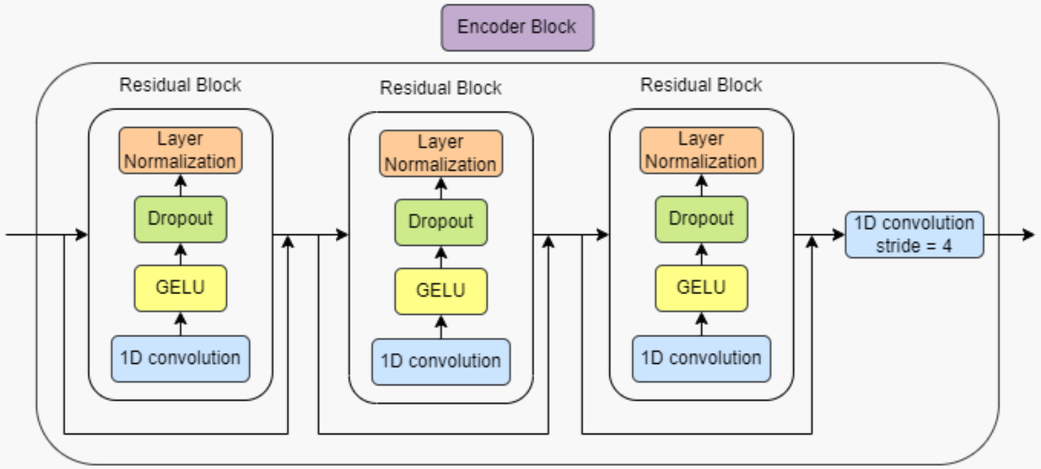


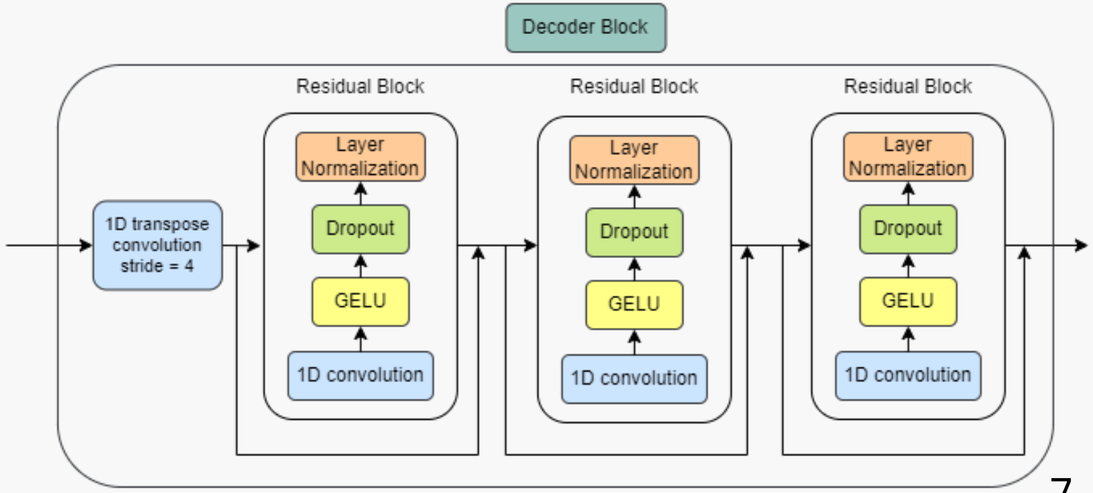
Distribution of dataset

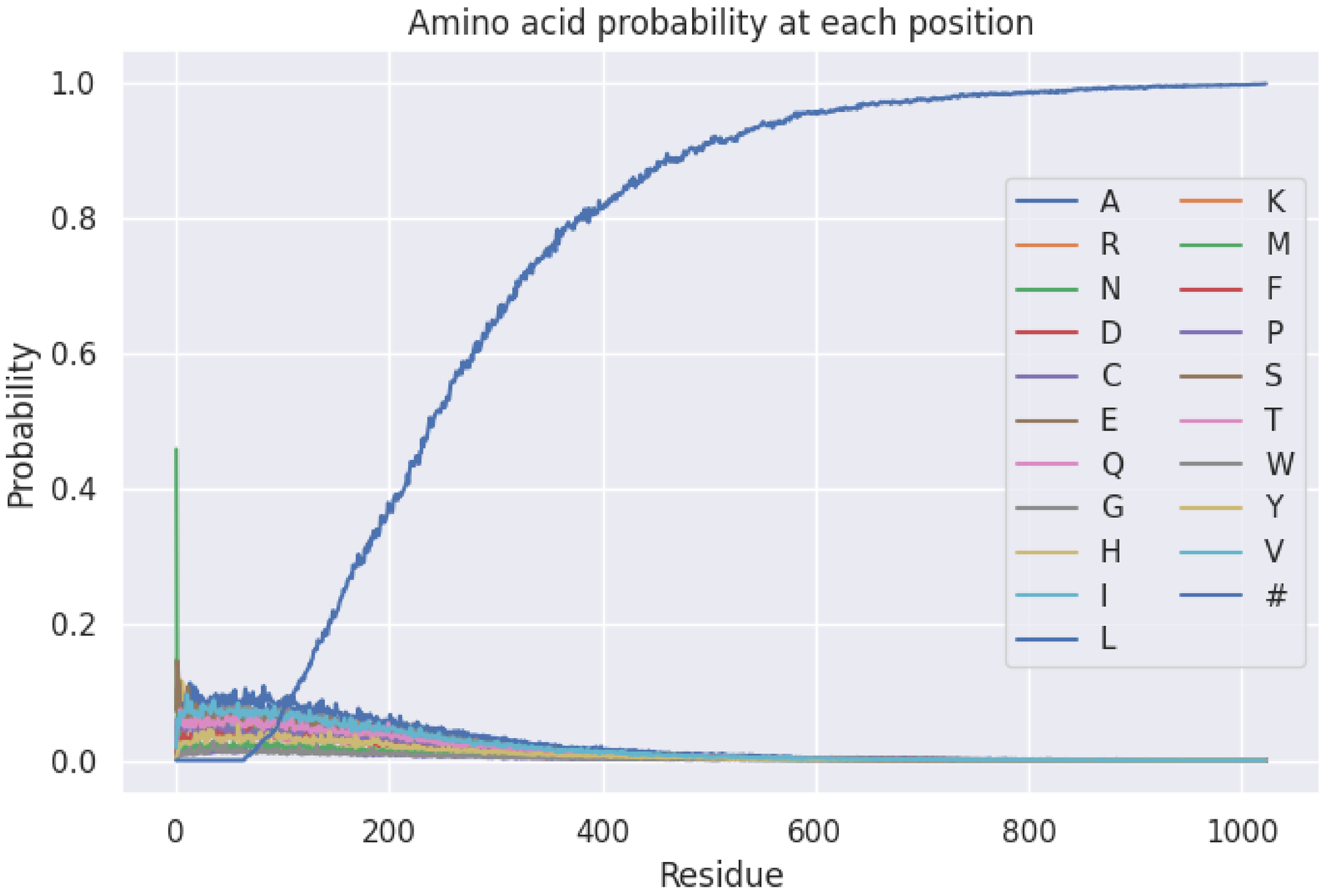


VAE model


Input X
Padded to length = 1024
MVLSEGEWQ....###
MNIFEMLRI.....###
MVLSEGEW.....###
MVLSEGEWQ....###
MLQSEWRCD...###
Input Embedding
Encoder
Encoder Block 1
Encoder Block 2
Encoder Block 3
Encoder Block 4
Flatten
FC
FC μ_x
FC σ_x
Z
FC
Reshape
Decoder
Decoder Block 1
Decoder Block 2
Decoder Block 3
Decoder Block 4
1D convolution
Output $\hat{X}$


Encoder Block
Residual Block
Residual Block
Residual Block
Layer Normalization
Dropout
GELU
1D convolution
1D convolution stride = 4

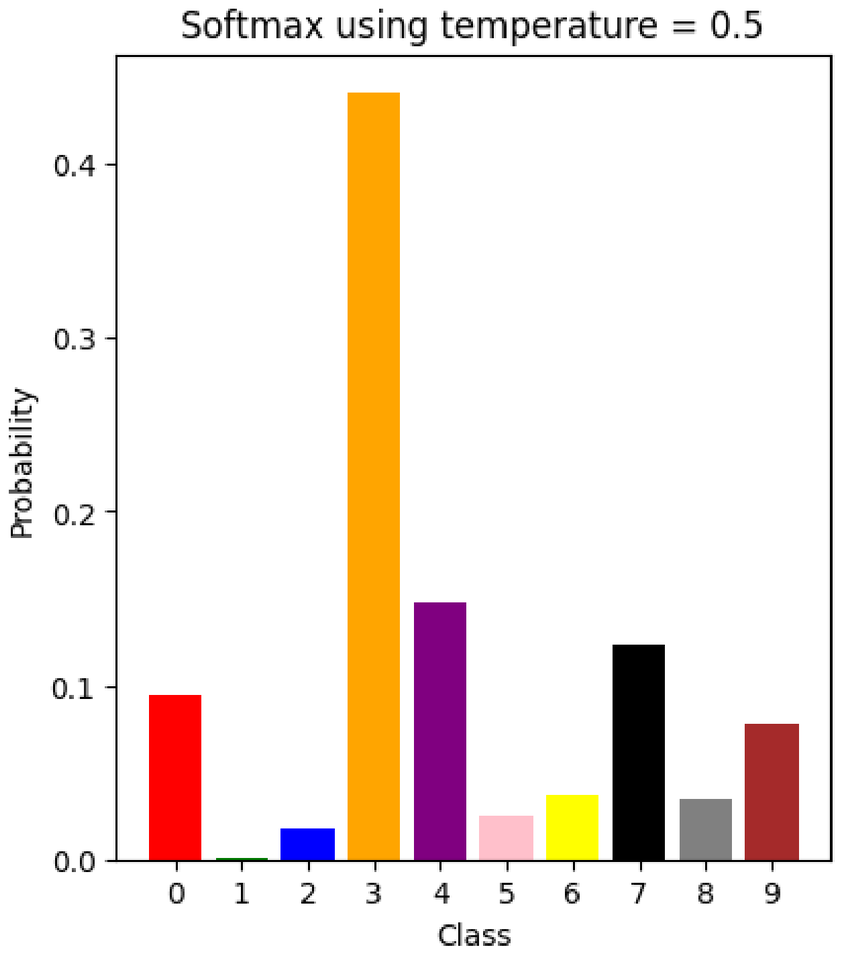

Decoder Block
Residual Block
Residual Block
Residual Block
1D transpose convolution stride = 4
Layer Normalization
Dropout
GELU
1D convolution

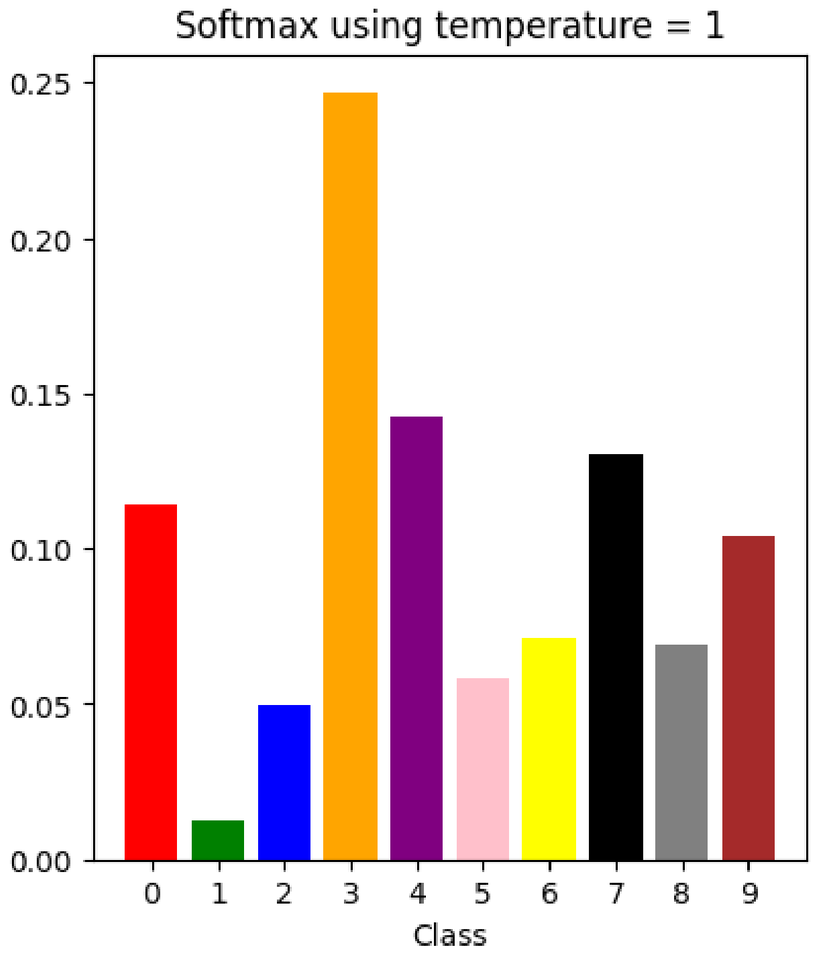

Amino acid probability at each position
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
Residue
A
R
N
D
C
E
Q
G
H
I
L
K
M
F
P
S
T
W
Y
V
#

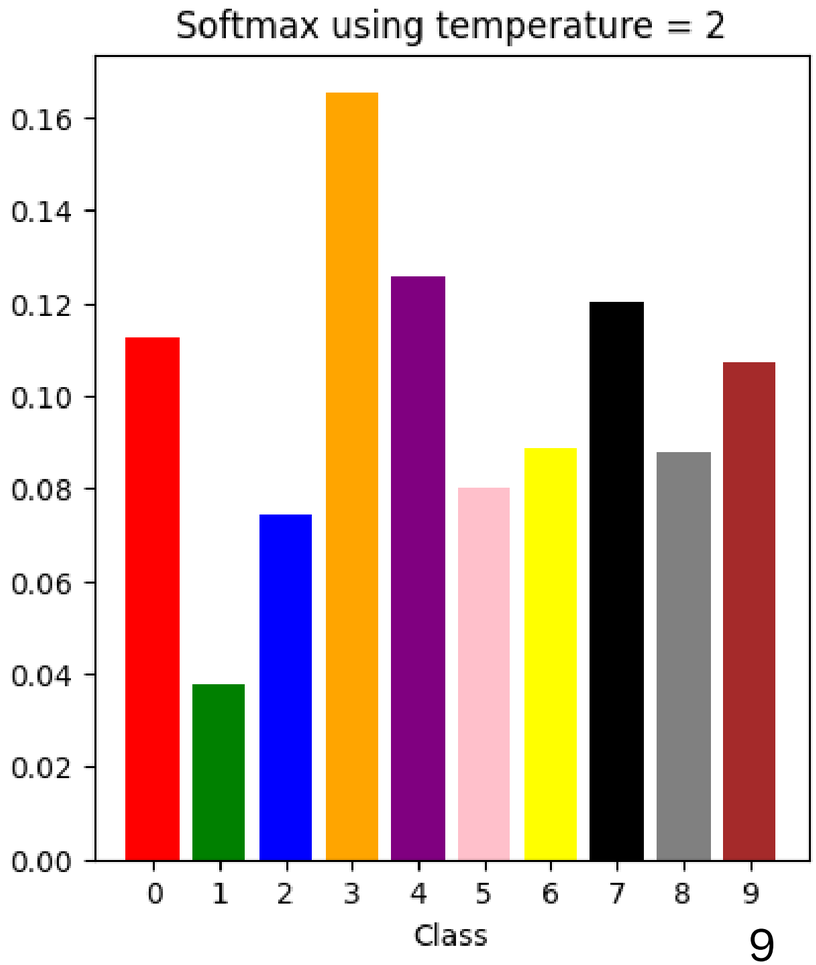
Softmax sampling using temperature

$$\texttt{Softmax}(y)_i = \frac{e^{y_i}}{\sum_j e^{y_j}}$$

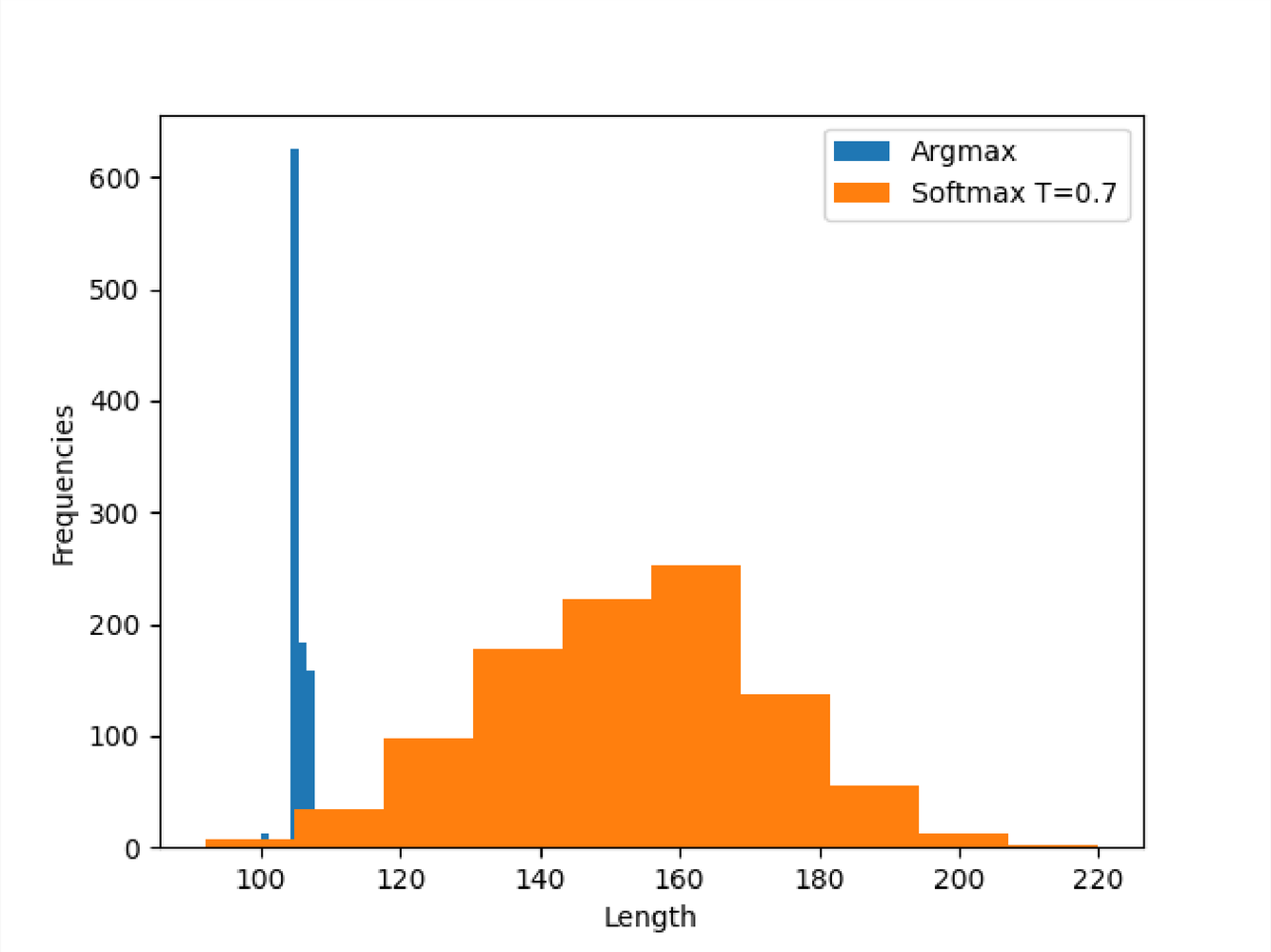
$$\texttt{Temp-Softmax}(y)_i = \frac{e^{y_i/T}}{\sum_j e^{y_j/T}}$$



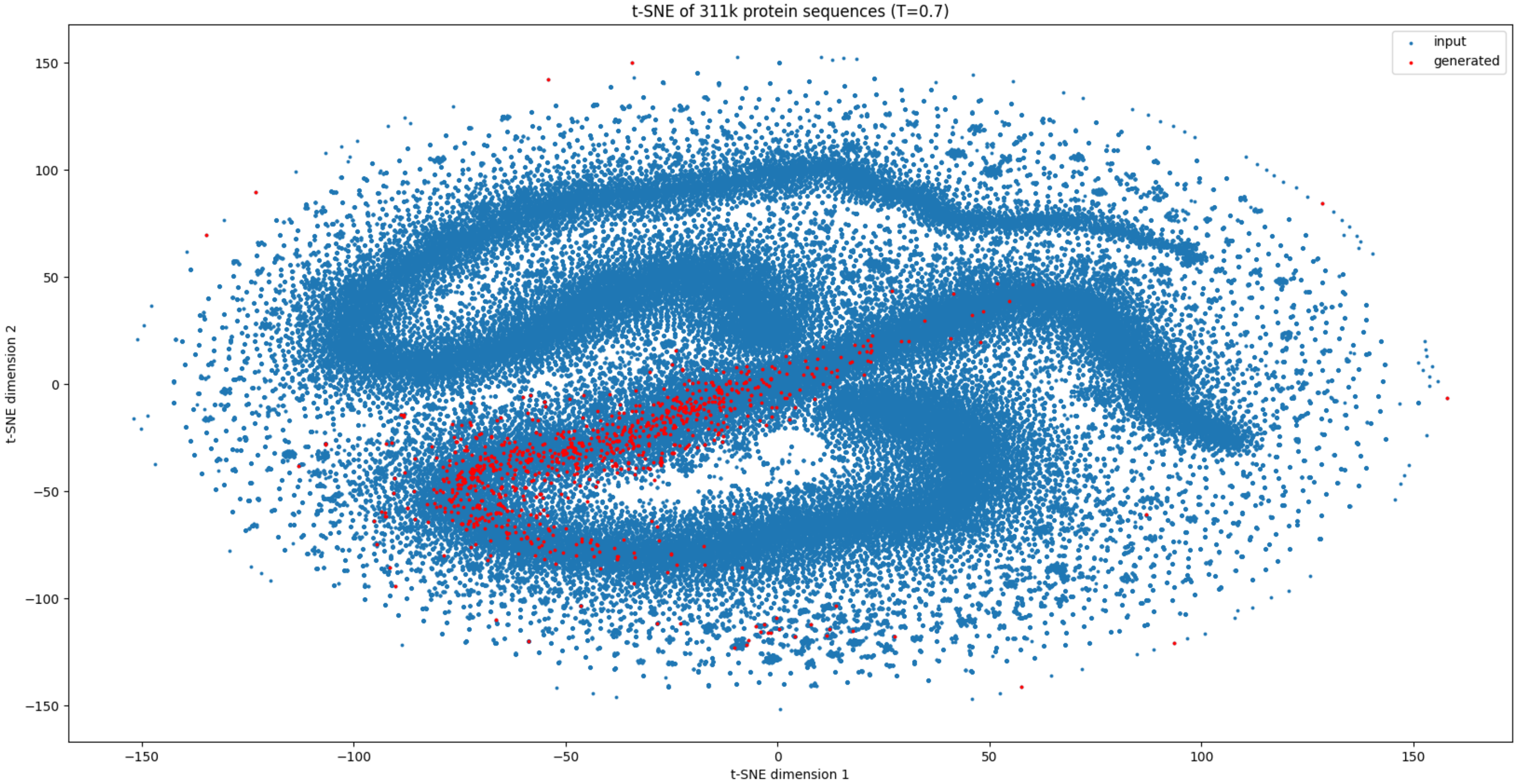




Histogram of length of 1000 generated proteins



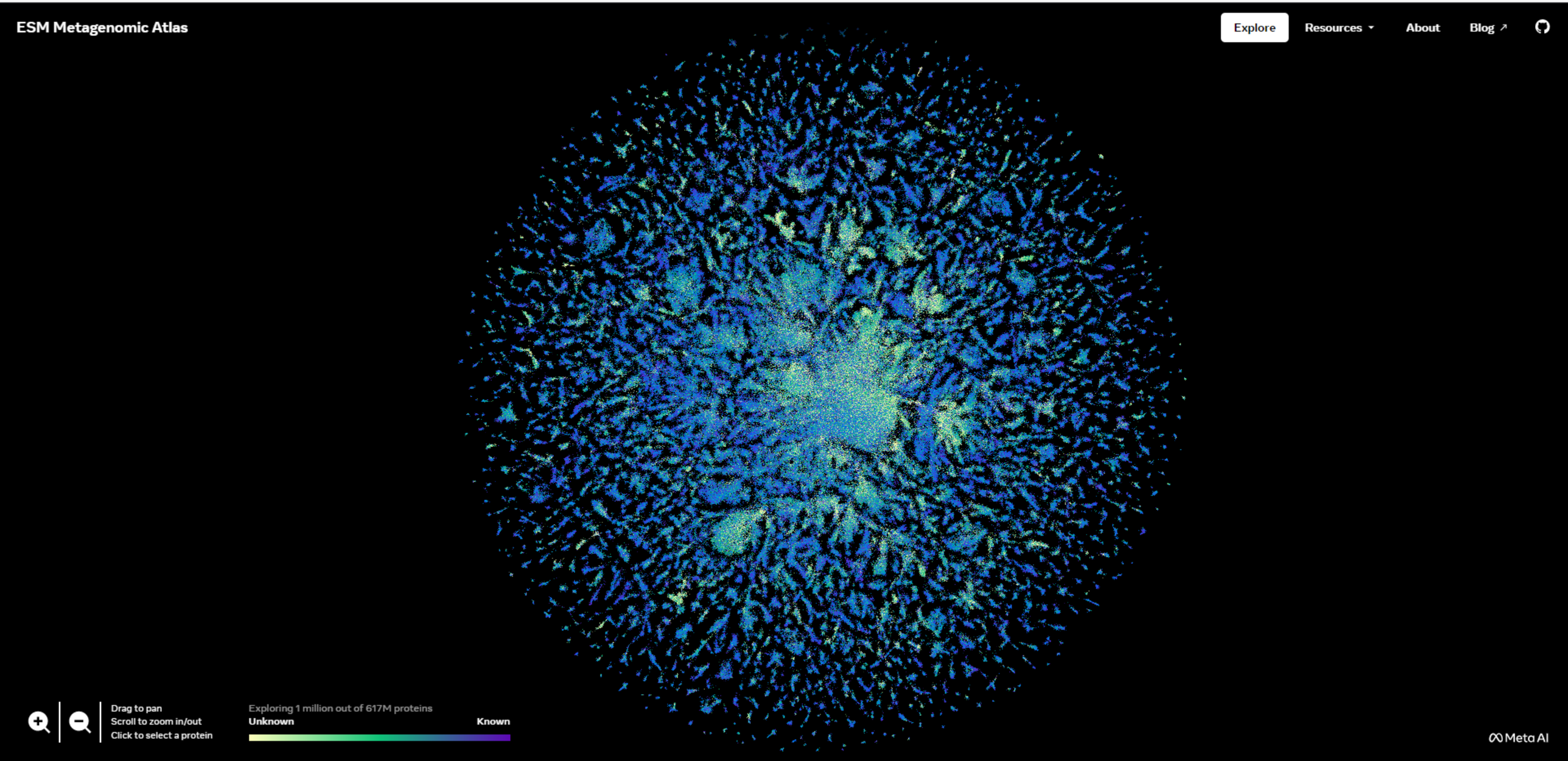
t-SNE visualization



Next steps

1. More data --> UniRef (311k --> 43M)
2. Implement Transformers
    a. Use rotary positional embeddings
    b. Use SwiGLU activation function
3. Implement Gradient Descent: The Ultimate Optimizer
4. TBD

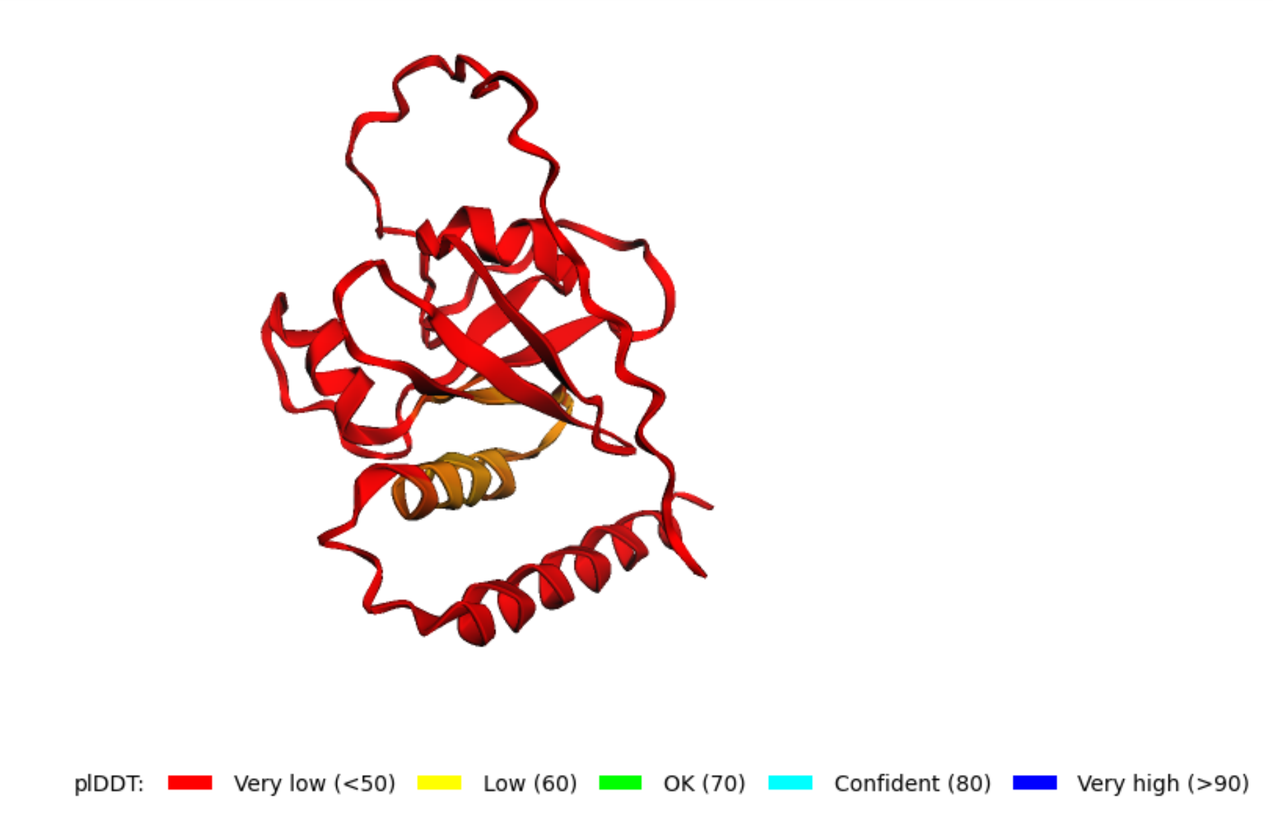
ESMFold


ESM Metagenomic Atlas
Explore
Resources
About
Blog
Drag to pan
Scroll to zoom in/out
Click to select a protein
Exploring 1 million out of 617M proteins
Unknown
Known
Meta AI

AlphaFold 2 vs ESMFold

214 683 829 structures
Requires MSA

~617 000 000 structures
Single sequence
60x faster



The CASP15 rankings

#1 PEZYFoldings AF2-based. Diverse MSAs. Custom, fine-tuned AF2 refinement

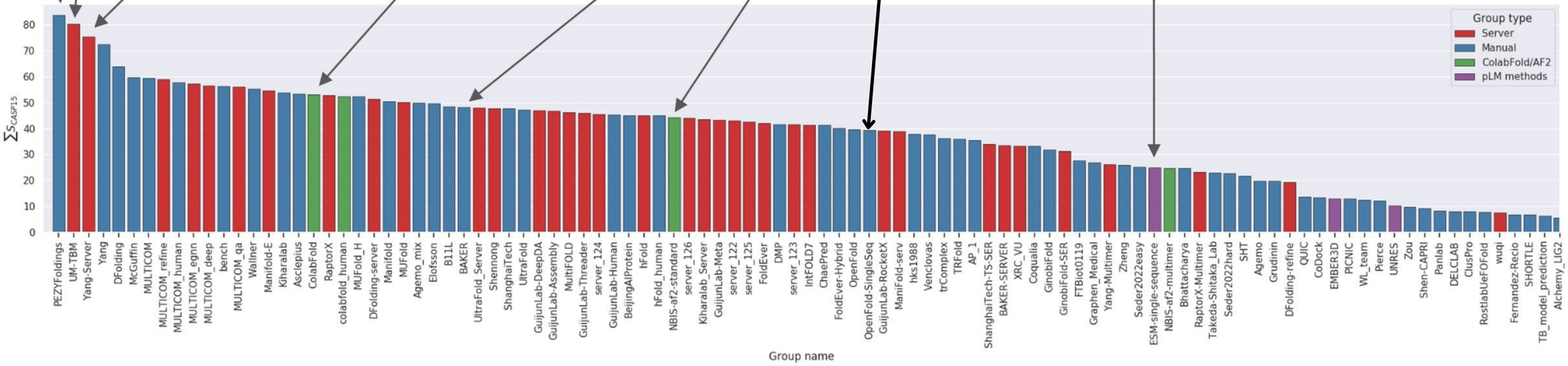
#2 UM-TBM Diverse MSAs. Threading then AF2 predictions guide I-TASSER REMC

#3 Yang-Server Diverse MSAs. AF2 predictions fed to trRosettaX2

ColabFold and NBIS-af2-standard

BAKER top non-AF2 method

ESM-singlesequence is the top pure pLM method by this metric



Evaluation

Current evaluation

- HMMR score
- DCA statistical energy score
- z-score
- pLDDT
- GDT
- PM
- TM

Evaluation

Current evaluation

- HMMR score
- DCA statistical energy score
- z-score
- pLDDT
- GDT
- PM
- TM

Fréchet Inception Distance

- Metric to evaluate performance of generative model.
- Measure of similarity between input and generated data.
- Fréchet Distance = distance between two probability distributions

Evaluation

Current evaluation

- HMMR score
- DCA statistical energy score
- z-score
- pLDDT
- GDT
- PM
- TM

Fréchet Inception Distance

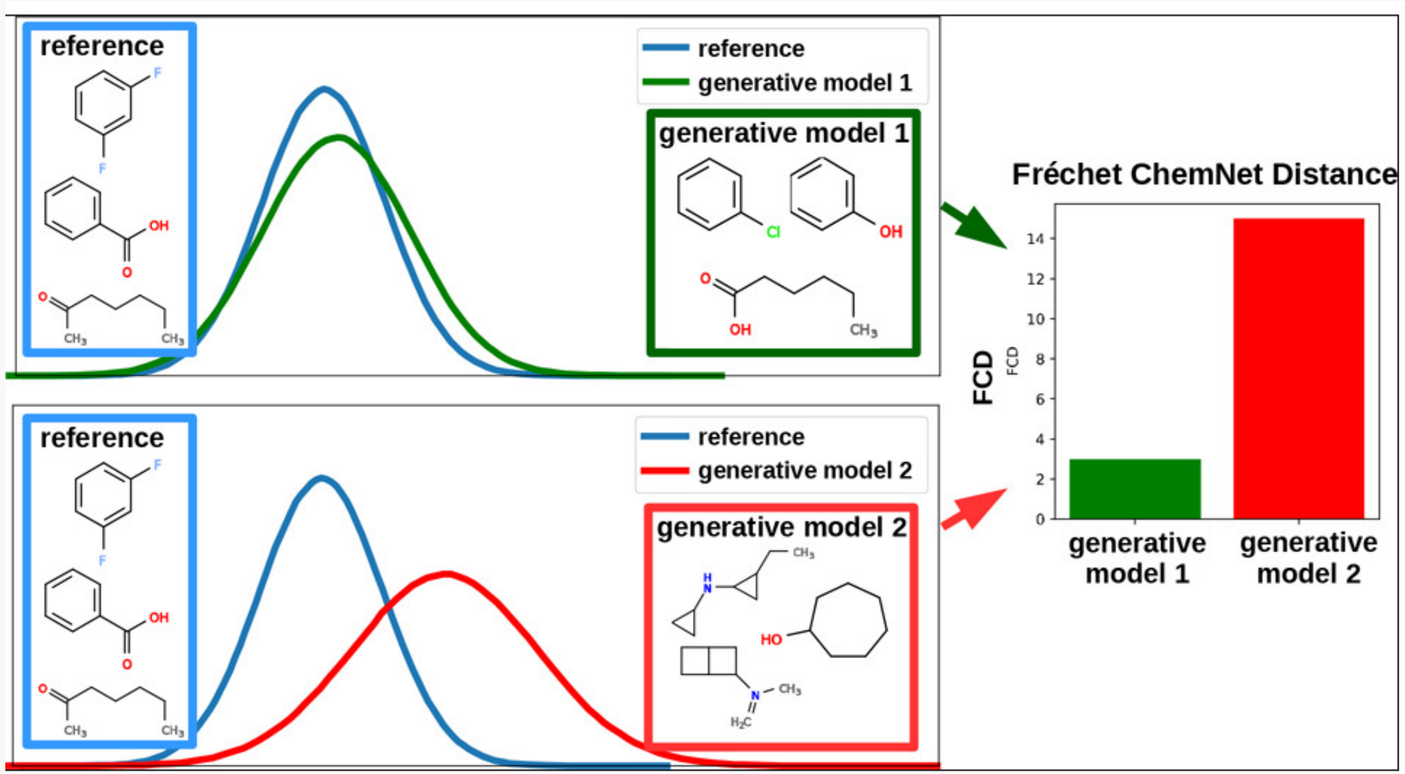
- Metric to evaluate performance of generative model.
- Measure of similarity between input and generated data.
- Fréchet Distance = distance between two probability distributions

Fréchet ESMFold Distance

Based on SOTA structure models

Fréchet Inception Distance

- Fréchet Inception Distance
- Fréchet Audio Distance
- Fréchet Video Distance
- Fréchet ChemNet Distance



Questions

Post-processing of generated sequence

Generated protein

MASEHAHRHDTVESEKFMVGVAEIEEYLFILLPECLAHGLKMKTKCVSVADTLFIQPMTDHVESN
VGPKGGPQAEGTP#AGDTSLKNLT##YDDASNSETIHS#YI#FLLRILFPARNSSSAMIIALSG#GR
QLPD##SPL#VKGLPGAD####V#VFIPF##M#A##T##E#####GARSR###YT#DFN##A#R
A#T##C#VIGLA#T#H#RNQ####RF####Q#######S###HD#NL#####G###########
K##G###F##P####T##Q######L
####S###NP##
##
##############H###
##
##
##
##
##
##
##
##

Post-processed protein

MASEHAHRHDTVESEKFMVGVAEIEEYLFILLPECLAHGLKMKTKCVSVADTLFIQPMTDHVESN
VGPKGGPQAEGTPKAGDTSLKNLTKKYDDASNSETIHSKYIKFLLRILFPARNSSSAMIIALSGKGR
QLPDKKSPLKVKGLPGAD