VAE Generative Network for de novo Protein Design

BY: BERNARD MOUSSAD

UNIVERSITY LIBRARIES
VIRGINIA TECH

The Problem

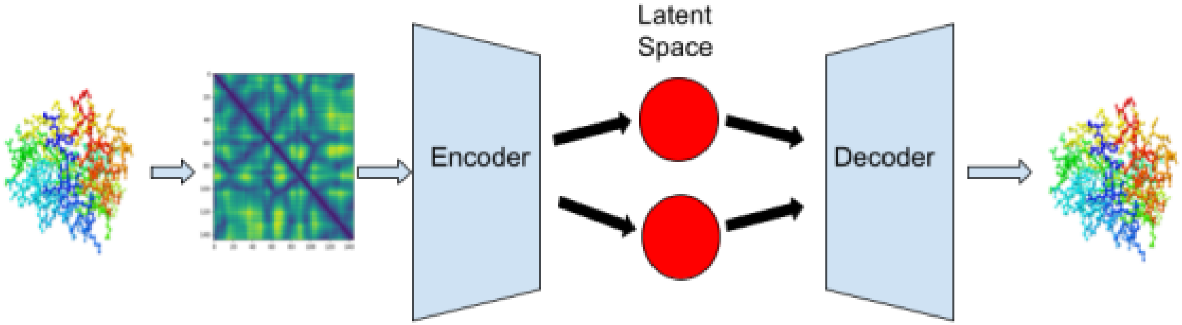
- Protein design has numerous applications that can benefit humanity
  - *Vaccinations*
  - *Enzyme Design*
- The issue is that proteins can take many conformations and shapes
- Designs by hand have been done but mainly based on known conformations

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

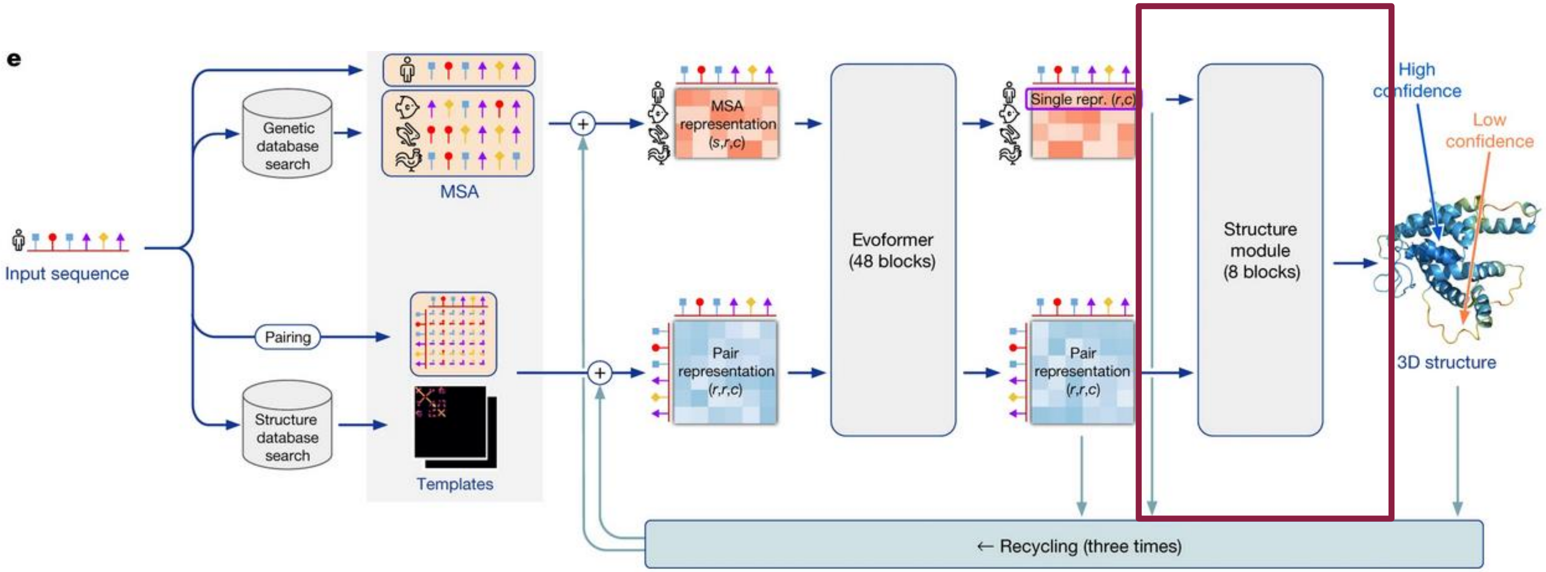
Initial Proposal

- 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



Current Progress

- Deep exploration into the architecture of the SE3 Network, a 3D Roto-Translation Equivariant Attention Networks
    - Seen in AlphaFold2's structure module.


e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

SE3 Overview

- Purpose: To apply self-attention for 3D Point cloud and graph data
- Why equivariant?
  - Because 3D data, such as protein atomic coordinates, is sensitive to translation and transformation

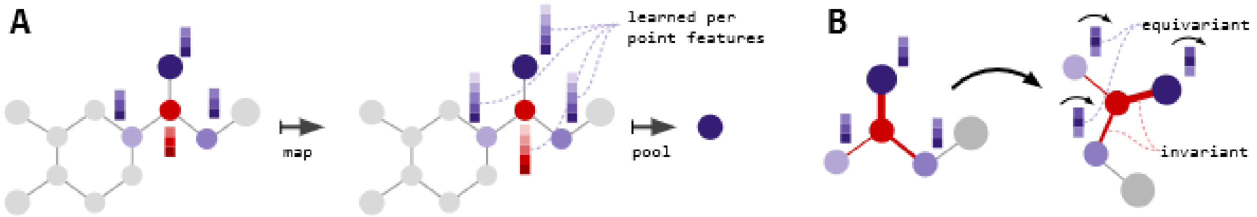


Figure 1: **A)** Each layer of the SE(3)-Transformer maps from a point cloud to a point cloud (or graph to graph) while guaranteeing equivariance. For classification, this is followed by an invariant pooling layer and an MLP. **B)** In each layer, for each node, attention is performed. Here, the red node attends to its neighbours. Attention weights (indicated by line thickness) are invariant w.r.t. input rotation.

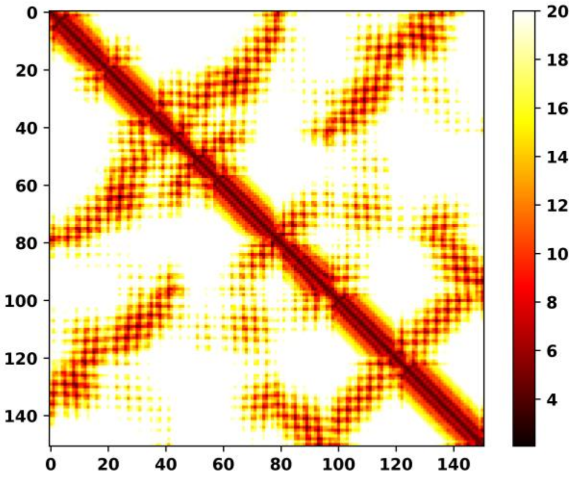
Purpose of SE3 In This Context

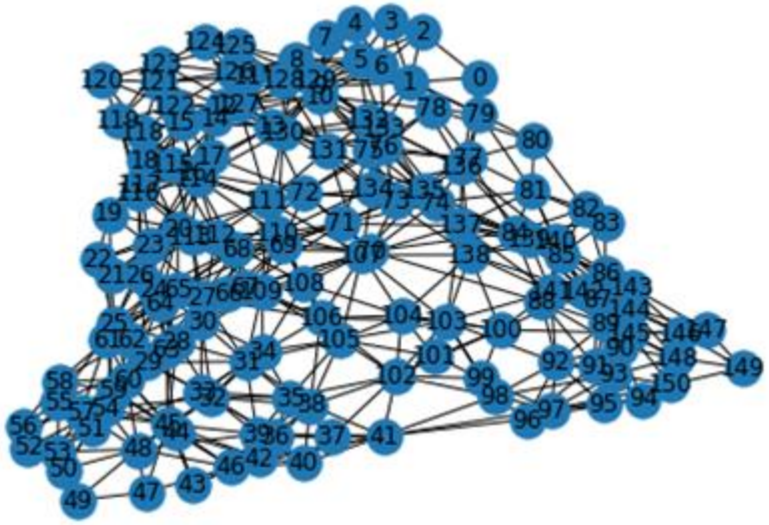
- Due to the nature of the problem, equivariance within the generated protein structure is necessary
- The main problem that can be faced:
  - Invalid Chirality (Essentially a protein is inverted)
- Intend to leverage SE3 as a structure module for the output of this network

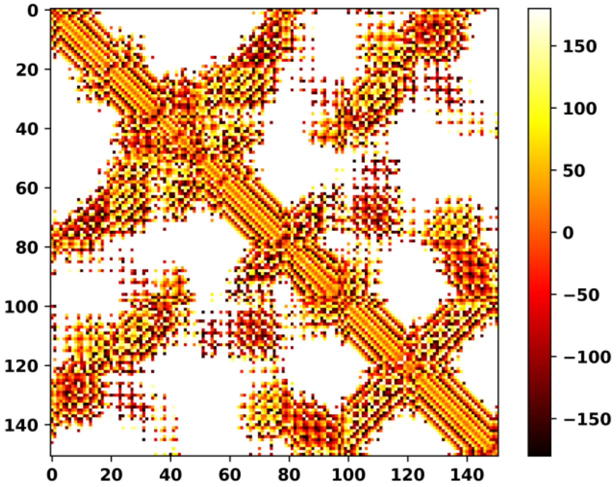
Current Architecture

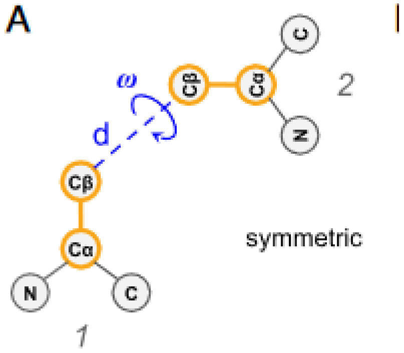
- Inputs:
  - Hypergraph depicting C-Beta atom interactions
  - Node Features
    - Binned Centrality: The number of atoms within 16 Angstroms and binned to bins of 0-6 (7 bins total)
  - Edge Features:
    - Atomic Distance: Distance between a pair of atoms
    - Inter-residue orientation angle (Omega) defined by Yang et. Al of trRosetta

Visualization of Features









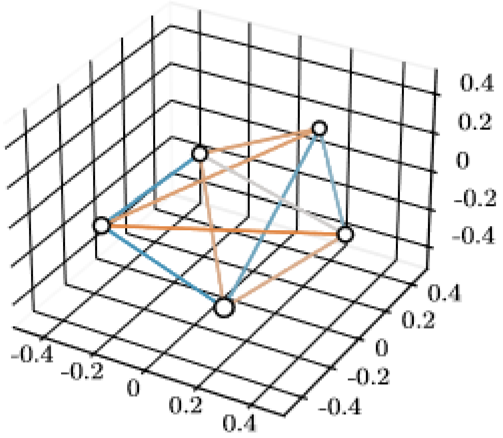
Current Architecture (cont.)

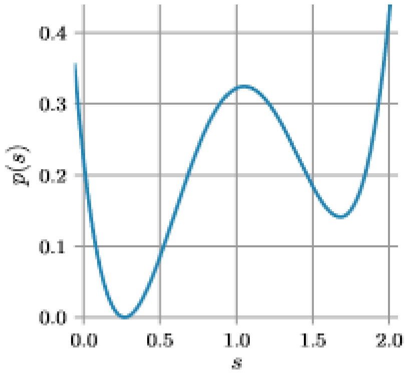
- Currently generates Nx3 matrix where N=number of samples so it’s currently in the form of a refinement problem

```
result= [[3.6821, 0.9162, 3.9393],
         [1.6956, 2.7115, 4.5427],
         [2.4745, 1.5844, 5.0162],
         [1.5174, 1.7530, 4.8103],
         [3.0756, 1.1991, 4.1712],
         [3.6285, 2.4279, 4.5034],
         [2.4163, 2.2594, 4.0556],
         [3.4414, 2.3383, 4.9026],
```

Next Steps: Iterative SE3-Transformer

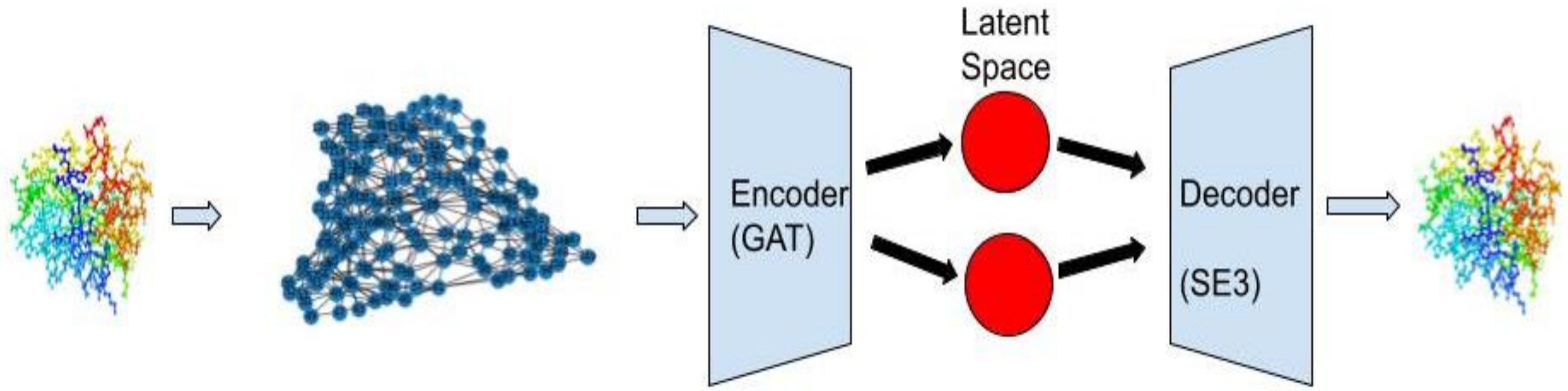
- The original authors, Fuchs et. al, demonstrated how AlphaFold2 managed to create an iterative paradigm via the SE3-Transformer
- Will be extending the current form of my architecture to leverage their findings





Overall Picture

- Maintain graph based representation of proteins
- Generate Latent Space encoding via Graph Attention (GAT)
- Reproduce atom coordinates via SE3



Questions