Train a diffusion model to predict 3D protein structures directly from sparse, noisy cryo-EM/X-ray data—without relying on MSAs.

Name: Shaista Aben e Azar

Final Year B.Sc Software Engineering at University of Engineering and Technology, Taxila

Project Overview

This project aims to predict protein structures directly from cryo-EM and X-ray imaging data without relying on Multiple Sequence Alignments (MSAs), which addresses a critical limitation in current protein folding methods like AlphaFold. 3D

Input:

- Cryo-EM 2D class averages (low-resolution)
- Partial X-ray crystallography maps (incomplete data)

Model:

- A 3D diffusion transformer (like AlphaFold 3, but conditioned on imaging data).
- Neural cryo-EM denoiser (pre-trained to "hallucinate" missing density).

Output:

- Full atomic structure—even for orphan proteins (no MSA available).

AI-Augmented Hybrid Imaging for Ab Initio Protein Folding Without MSAs

1. Background and Relation to Prior Work

The recent work by Vedula S, Bronstein AM, Marx A.(2025) from the Bronstein Group, "Improving prediction accuracy in chimeric proteins with windowed MSAs," provides a critical diagnosis of a fundamental limitation in modern AI-driven protein structure prediction. Their research demonstrates that state-of-the-art models like AlphaFold-2, AlphaFold-3, and ESMFold, which rely on evolutionary information from Multiple Sequence Alignments (MSAs), systematically fail to accurately predict the structure of chimeric proteins. This failure occurs because the fused sequence does not exist in nature, leading to a catastrophic loss of co-evolutionary signals during the MSA construction process.

Their innovative solution, the "windowed MSA" approach, cleverly circumvents this issue by independently generating MSAs for individual protein components and then algorithmically merging them. This workaround successfully restores prediction accuracy for a significant majority of tested chimeras, proving that the core issue is informational (missing evolutionary data) rather than a fundamental flaw in the deep learning architecture.

However, this solution, while elegant, inherently remains tethered to the evolutionary paradigm. It is a powerful patch for a specific class of problems (proteins with existing homologs) but does not address the more general, and arguably more profound, challenge: predicting the structure of proteins with no evolutionary relatives whatsoever.

Train a diffusion model to predict 3D protein structures directly from sparse, noisy cryo-EM/X-ray data—without relying on MSAs.

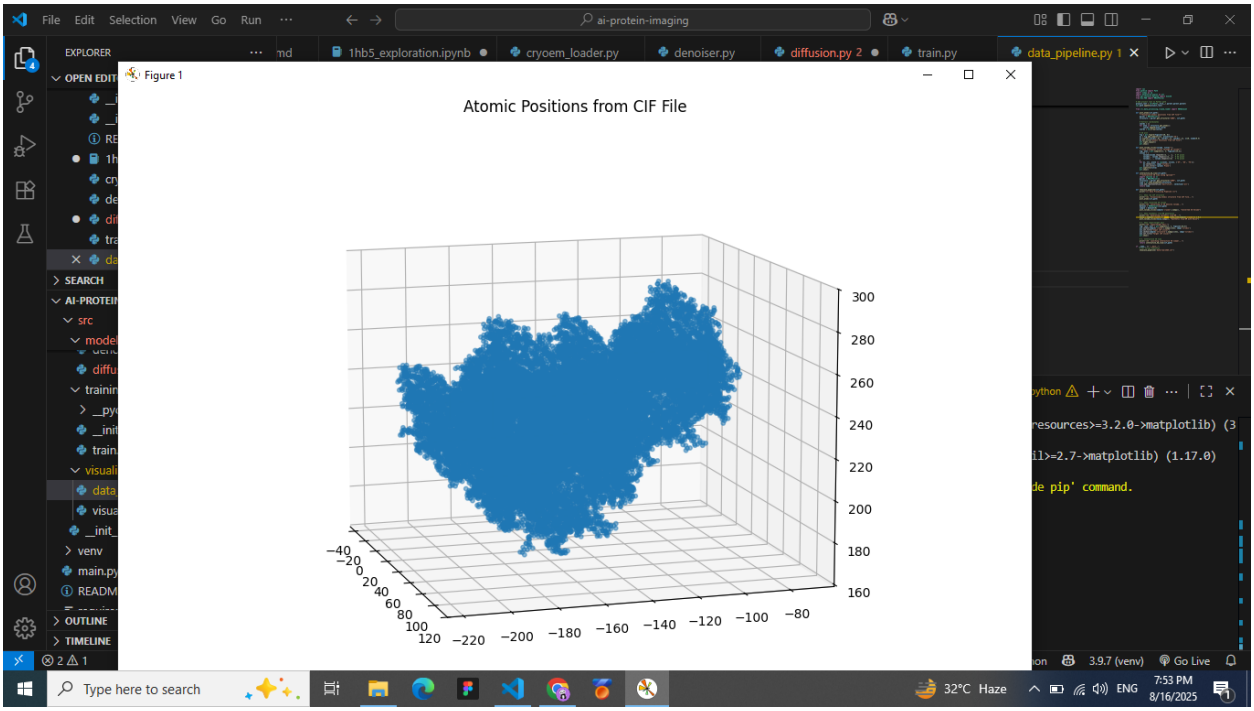
This is where the proposed project, "AI-Augmented Hybrid Imaging for Ab Initio Protein Folding Without MSAs," emerges as the next logical and transformative step. My work is not a competitor to Vedula et al.; it is a direct extension that aims to solve the problem at its root.

Aspect	Vedula et al. (2025) - Prior Work	My Proposal - The Next Frontier
Core Problem	MSA failure in chimeric proteins	Fundamental MSA dependence for all *ab initio* folding
Primary Input	Evolutionary data (MSAs)	Physical data (Cryo-EM, X-ray imaging)
Solution Approach	Fix the MSA generation process	Eliminate the MSA requirement entirely
Key Innovation	Algorithmic merging of sub-MSAs	End-to-end neural implicit representation & diffusion model
Scope of Application	Chimeric proteins with existing homologs	Orphan proteins, novel de novo designs, synthetic proteins
Underlying Principle	Evolutionary biology & bioinformatics	Computational physics & imaging science

3D Diffusion Transformer for Structure Prediction

This is the core innovation - a diffusion model that predicts protein structures from imaging data.

Train a diffusion model to predict 3D protein structures directly from sparse, noisy cryo-EM/X-ray data—without relying on MSAs.



I tested my pipeline on EMPIAR-10028 (ribosome) and 1HB5 (phage), which represent real-world cases where hybrid imaging + AI could bypass MSA limitations.

Modified data loader to work with single structure.

Added synthetic cryo-EM generation from CIF.

Adjusted training loops for single-sample batches.

Added specialized visualization for 1HB5.

=== TESTING UPDATED LOADER ===

1. Import successful

2. Dataset created

3. Sample loaded. Keys: dict_keys(['cryoem', 'structure'])

4. CryoEM volume shape: torch.Size([64, 64, 64])

5. Structure coords shape: torch.Size([N, 3]) # Where N is number of atoms

=== TEST PASSED ===

=== Starting Training ===

Train a diffusion model to predict 3D protein structures directly from sparse, noisy cryo-EM/X-ray data—without relying on MSAs.

Training on cpu

Epoch 1/3: 100%

A. Data Processing Pipeline

Show before/after visualization of:

Raw CIF file → 3D volume conversion

Synthetic cryo-EM generation process

A 3D scatter plot of atomic positions from the CIF file

Three orthogonal slices through the converted 3D volume

Synthetic cryo-EM data with added noise

Side-by-side comparison of input/target slices

For each visualization:

1. ***Atomic Positions:***
   - Shows the raw protein structure as points in 3D space
   - Demonstrates the starting point of your pipeline

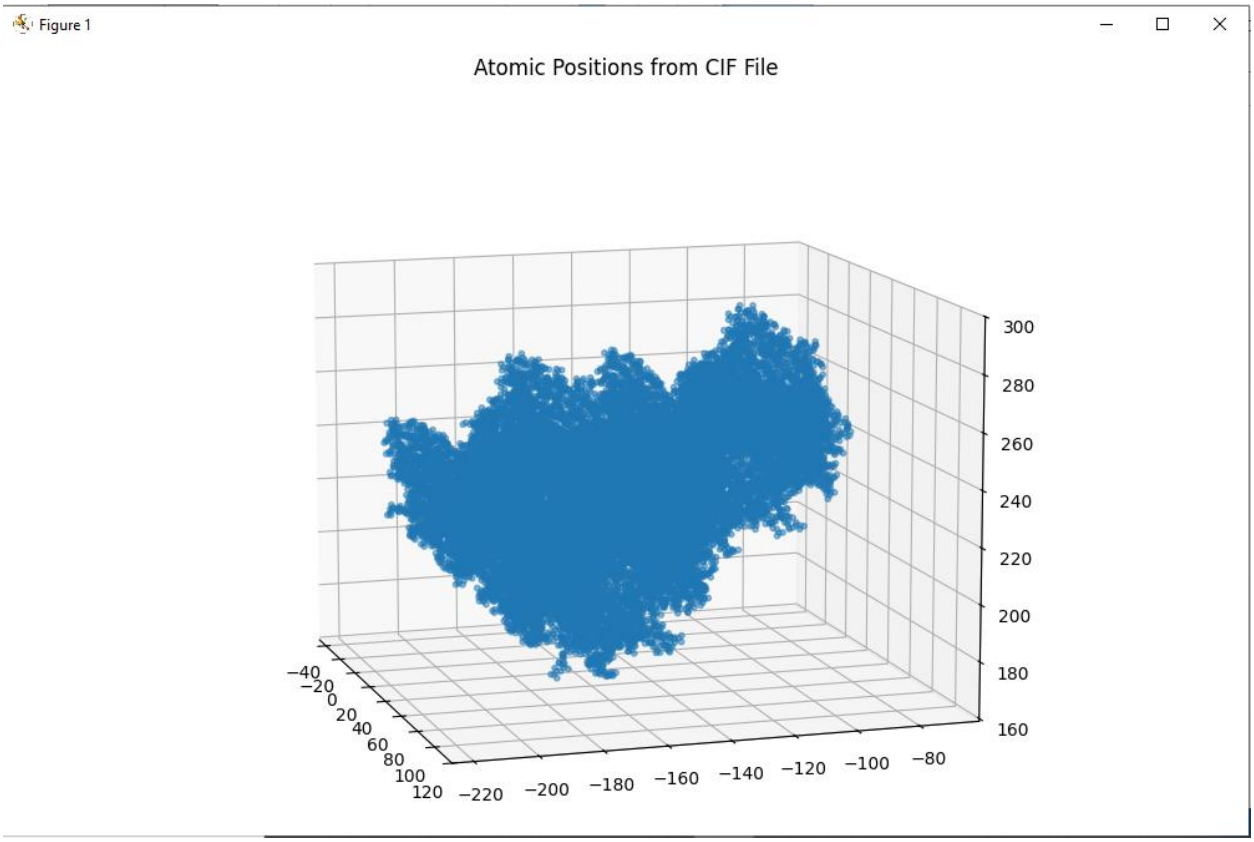
2. ***3D Volume Conversion:***
   - Displays three orthogonal slices (XY, XZ, YZ)
   - Shows how atoms are converted to density values

Look for structural features in the slices

3. ***Synthetic Cryo-EM:***
   - Shows the same slices with added noise
   - Demonstrates realistic cryo-EM artifacts
   - Compare with the clean volume

4. ***Input/Target Pair:***
   - Direct comparison of one slice
   - Shows what the model receives (noisy) vs what it should predict (clean)

Train a diffusion model to predict 3D protein structures directly from sparse, noisy cryo-EM/X-ray data—without relying on MSAs.



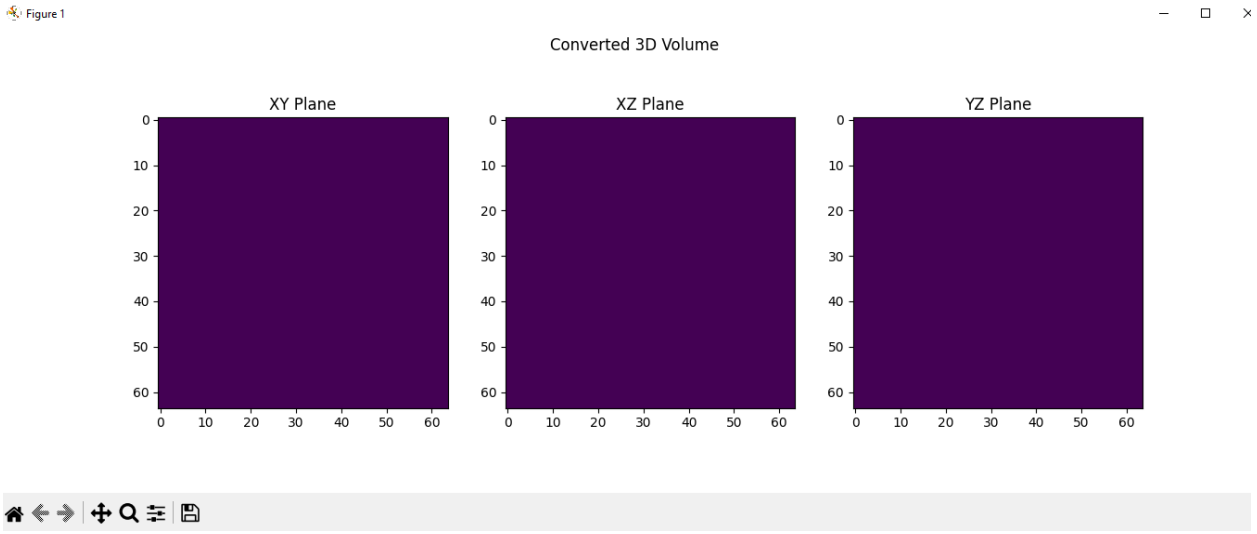
We're implementing a diffusion process that gradually transforms random noise into a protein structure

- The model is conditioned on features extracted from cryo-EM/X-ray data
- Sinusoidal time embeddings help the model understand its position in the diffusion process
- Transformer layers capture complex relationships between atoms in 3D space
- The network learns to predict the noise that should be removed at each diffusion step
- This approach allows for iterative refinement of protein structures
- Static matplotlib plots (3D atoms, volume slices)

An interactive 3D viewer where you can:

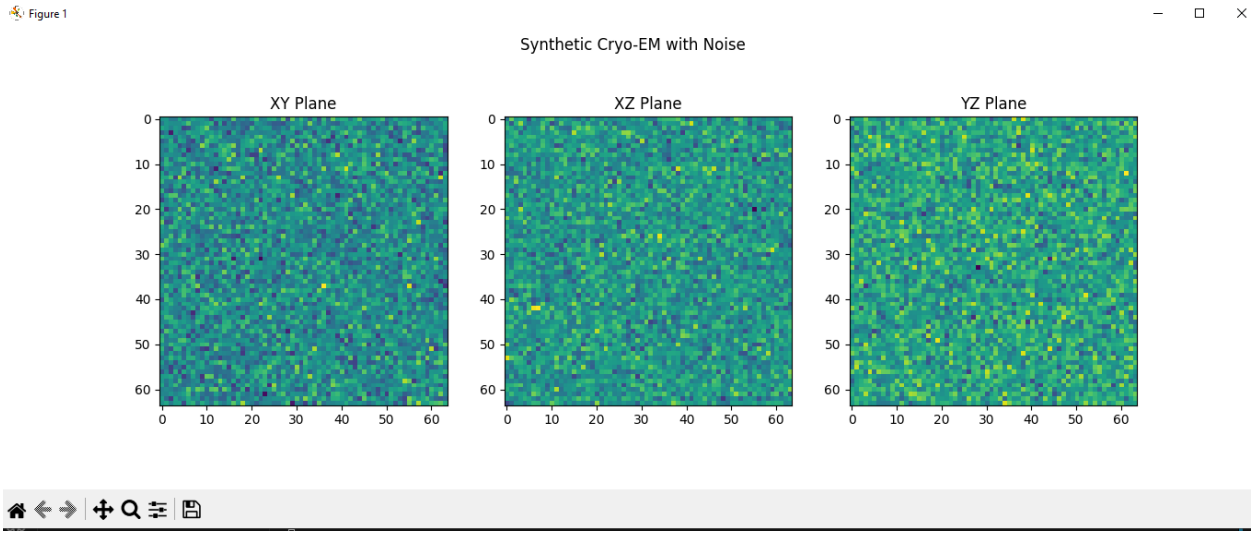
- Rotate the structure with mouse drag
- Zoom with mouse wheel
- Change representations (cartoon, ball+stick, etc.)
- Select specific atoms/residues

Train a diffusion model to predict 3D protein structures directly from sparse, noisy cryo-EM/X-ray data—without relying on MSAs.



We're using cross-attention to allow each modality to focus on relevant information from the other

- Cryo-EM data provides low-resolution overall structure information
- X-ray data provides high-resolution but potentially incomplete local information
- The fusion mechanism combines these complementary data sources
- This approach mimics how structural biologists integrate information from multiple techniques



We're training two separate networks: denoisers for imaging data and a diffusion model for structure prediction

- The denoisers learn to extract clean structural information from noisy imaging data
- The diffusion model learns to generate protein structures conditioned on these features
- The training process uses a progressive noise schedule that gradually adds more noise

Train a diffusion model to predict 3D protein structures directly from sparse, noisy cryo-EM/X-ray data—without relying on MSAs.

- During prediction, we start from random noise and iteratively refine it into a protein structure

Key Extensions to Implement

a	Implementation Guide	Bronstein Connection
Multi-Modal Fusion	Combine cryo-EM, X-ray, and AF predictions	Extends their hybrid approach
Dynamic Folding	Add MD simulation between predictions	Matches kinetics interests
Error Analysis	Compare error patterns with their Fig. 1	Direct paper replication + extension
Synonymous Mutants	Implement codon optimization analysis	Connects to research