Evaluating DynaMight for Conformational Heterogeneity of the *Tetrahymena Ribozyme*

Grace Tully Rotation Project

Das Lab

June 1, 2024

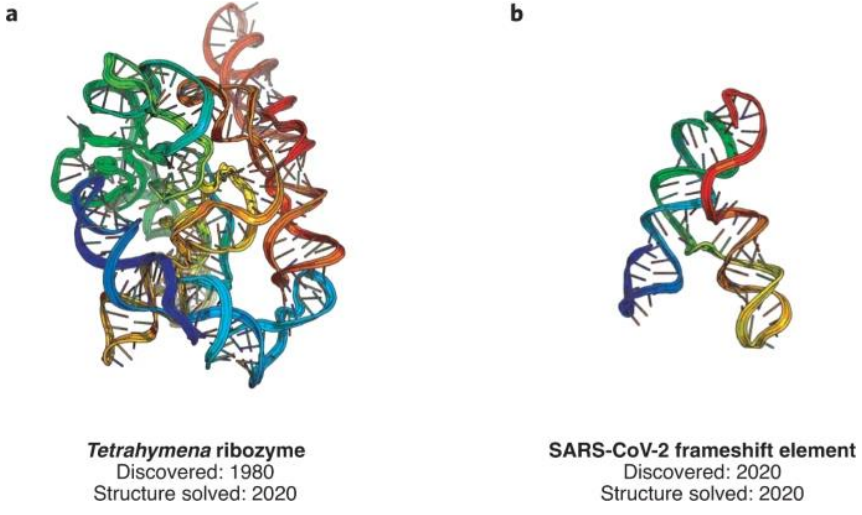
Stanford University

Main Motivational Question:

Do the latest, state-of-the-art, cryo-EM heterogeneity algorithms provide reliable information on the dynamics and structural variability of RNA-only structures?

Tertiary Structure of RNA

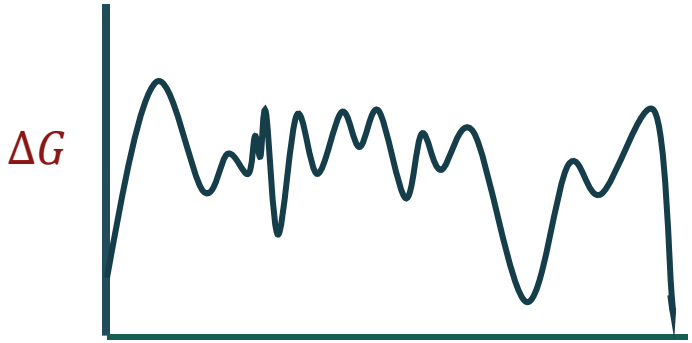
- Experimental Methods for tertiary structure:
  - X-ray Crystallography
  - Nuclear Magnetic Resonance (NMR)
  - Cryo-EM
    - RNA structure: a renaissance begins? (Das 2021)



Source: Data from PDB entries 6WLS and 6XRZ.

The Importance of Conformational Variability

- Can we use instruments to go beyond static tertiary structure to elucidate functional roles of RNAs?
  - Why AlphaFold Won't Revolutionize Drug Discovery (Lowe 2021)

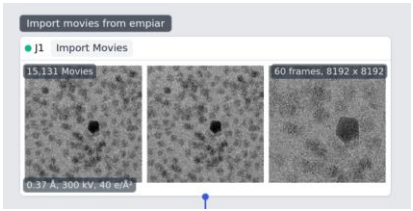


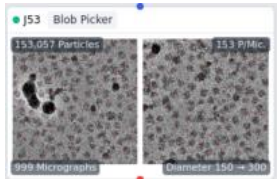
Conformational Landscape

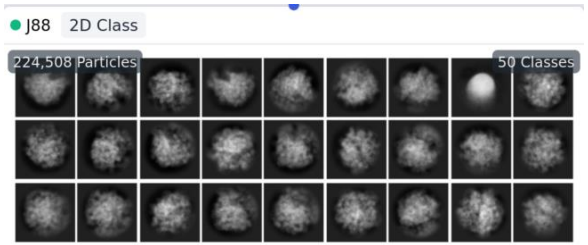
- "In this regard, the characterization of RNA structure ensembles in living cells represents a key step towards mapping the druggable transcriptome." (Spitale 2023)

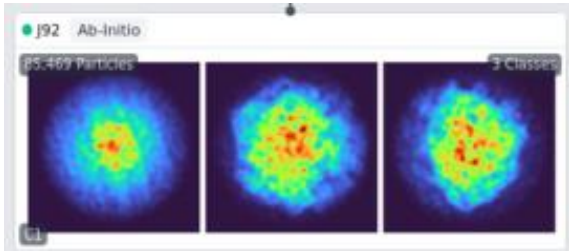
Traditional 3D reconstruction with cryoEM

- Recap of data processing workflow
- *Ab initio* 3D classification:
  - Takes a "guess" at 3D consensus volume, takes projections of the guess, then matches projections to 2D classification projections → iterative process
- Main concern: variable regions can be averaged out into final refined structure







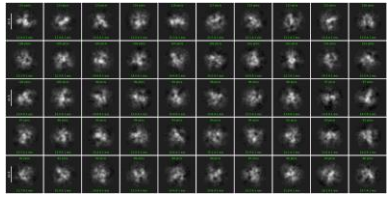


Can we take advantage of the high quantity (~1M) of particles used in the cryo-EM SPA workflow to uncover information about particle heterogeneity?

Same Ingredients – New Recipe

Traditional 3D Reconstruction:
a combination of maximum-likelihood estimation and stochastic optimization techniques

cryo-EM mircrographs



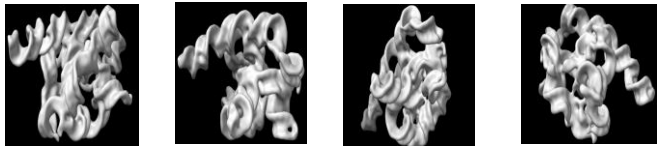
Modern ML Algorithms:
GMMs, VAEs, t-SNE, UMAP

Understanding the Computational Problem

While particle's conformational landscape can be represented by a single, non-linear 1D path, every voxel in image space takes a unique (non-linear) trajectory along that path

Can use simple algorithms (such as PCA) to *identify* variable regions, but not decode a refined trajectory of variability.

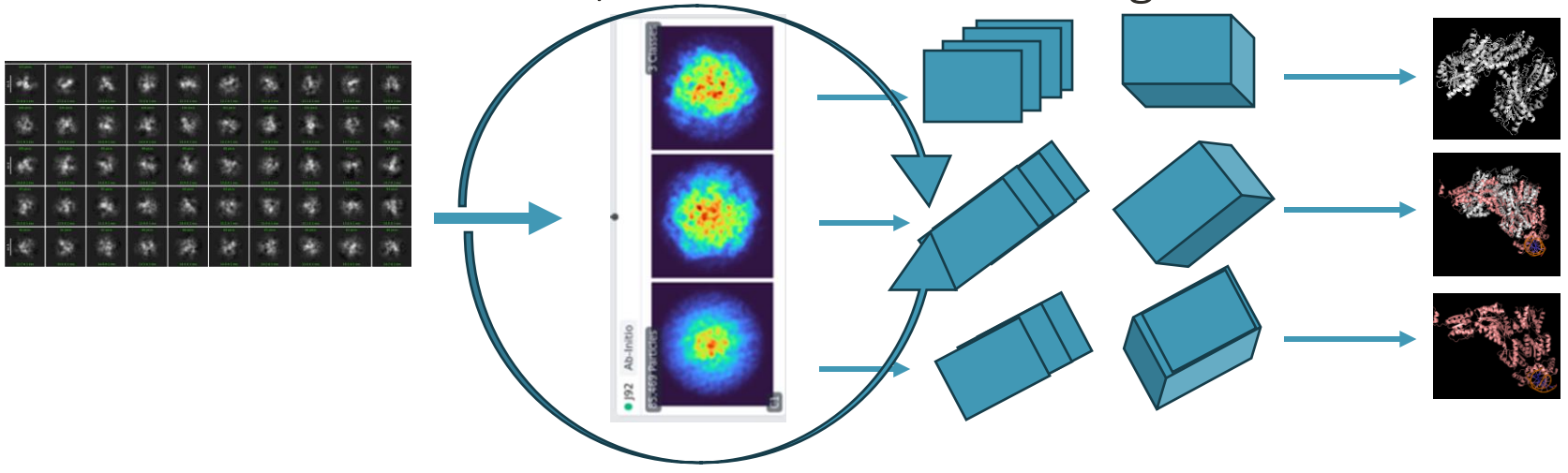
Conformational Landscape



$v \in \mathbb{R}^{1,000,000}$

Previous Methods

- Multi-model refinement/3D classification through

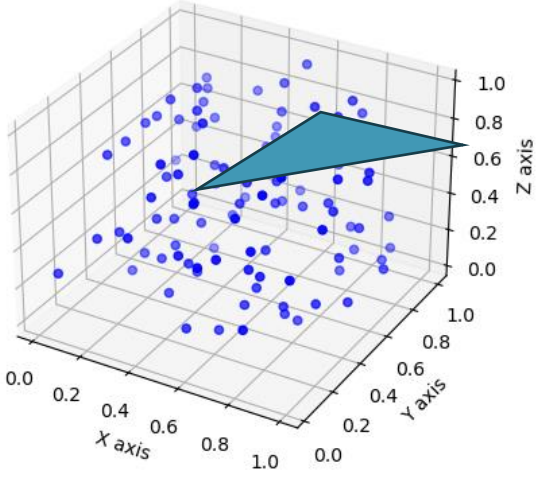


- Problems with maximum likelihood methods: 1) over or under-estimate of classes 2) assumes discrete classes exist

Manifold Embedding Techniques

- Embed 2D projections into latent space
- Separate clusters first into least variable regions (differing only in alignment)
- Then once orientations are aligned, embed regions into latent space, cluster, repeat to reveal conformational and compositional heterogeneity
- Pros and Cons

J. Frank 2016



Introduction of Variational Autoencoders using 3D Gaussian Mixture Models (GMM)

- e2gmm (EMAN2)

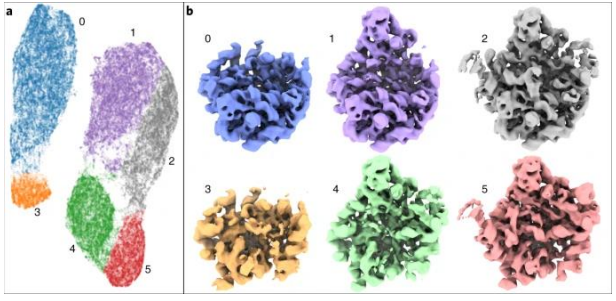


Fig. 2: Classification of assembling ribosomes. (from EMAN2)

- DynaMight (Relion-5.0)

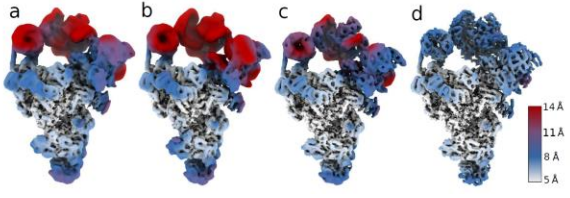
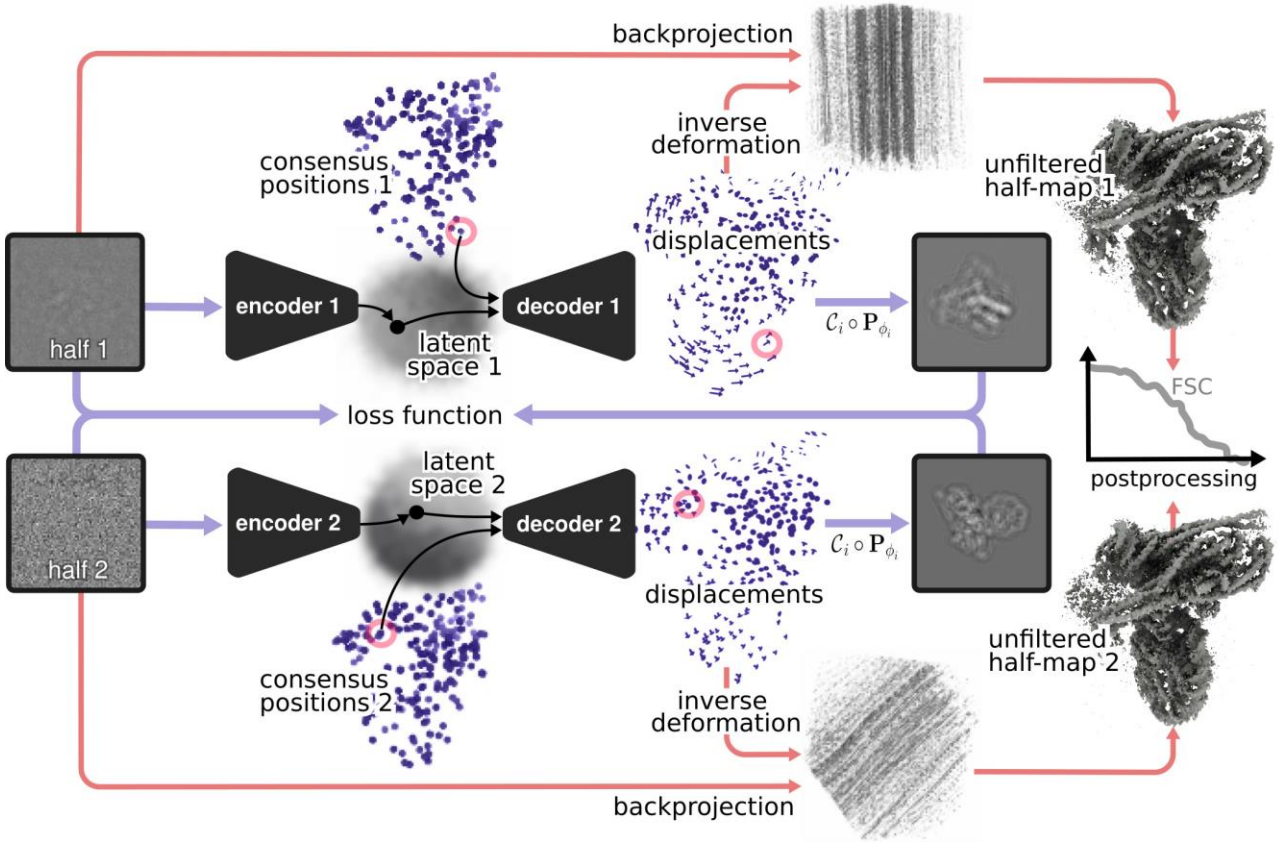


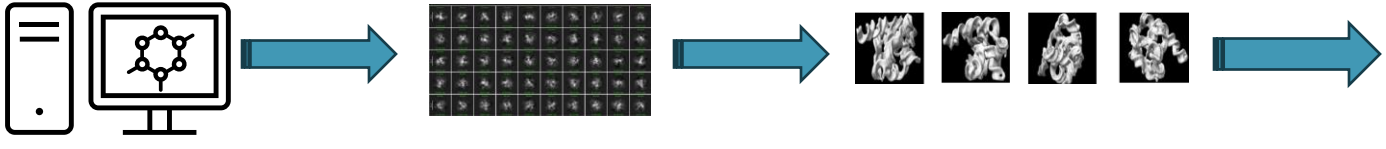
Figure 2: DynaMight reconstructions of the spliceosome subset.

DynaMight (Relion-5.0)



Do the latest, state-of-the-art, cryo-EM heterogeneity algorithms provide reliable information on the dynamics and structural variability of RNA-only structures?

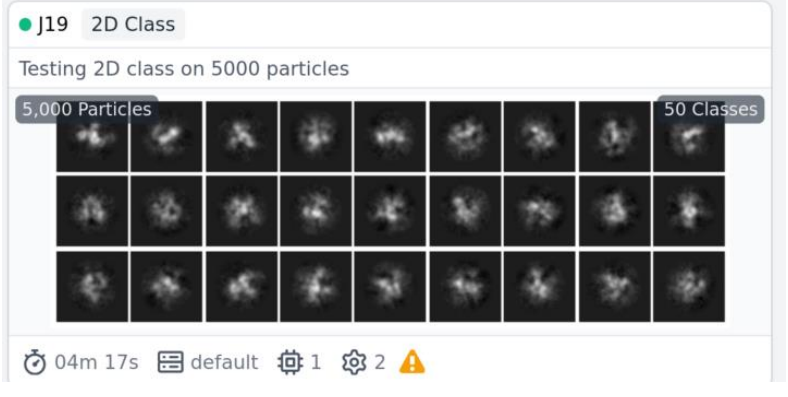
1) Create simulated particle stacks from MD simulations
2) Train DynaMight on Simulated Particle Stacks
3) Evaluate variability metrics on results for recovery of “ground truth”

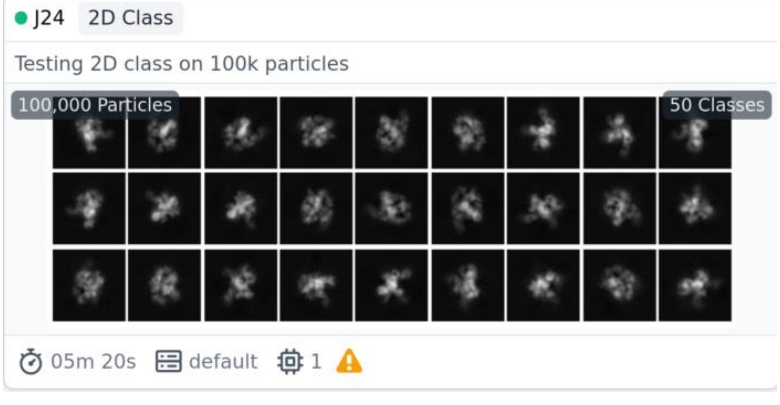


	MD	DM
RMSD	~	~
RMSF	~	~
RoG	~	~
RMSF	~	~

The simulated data

- 2000 pdbs
- Must remove ions
- Parameters --res 2.5 --box_size 224 --apix 0.86 --num_projections 10 --defocus_start 0.5 --defocus_end 1.5 --bfactor 126 --pink_noise 3.5
- Validate Results by importing into csparc

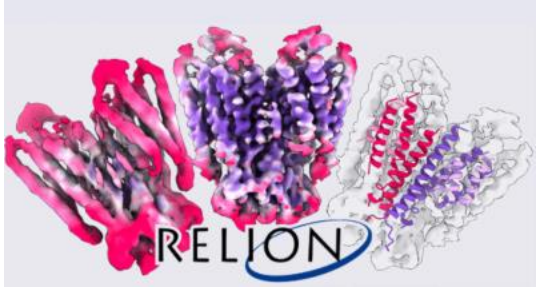




Training DynaMight with Simulated Data

- Run EMAN2 csparc2star.py to change .csg to .star
- Must change path directory in .star file to match relion project directory
- Must change .mrc to .mrcs file and change in .star file
- Note: DynaMight will not run if eman2 has been loaded (slight discrepancy in dependencies)

cryoSPARC



Technical Problems with DynaMight

1. coarse_grain util code is only written for proteins that have nucleic acid residues -- not pure nucleic acids
2. Decoder has an undefined gpu box
3. write_xyz code is not really useful (arbitrarily assigns atomic identities based on gaussian amps/widths as C/O even though initial model is CG
4. coarse_grain code mistakes uracil for a purine

Maintaining Integrity of Structures

1. Use an atomic model constraint for gaussian displacements

$$\mathcal{D}(z_i, \mathbf{c^0}) = \mathbf{c^0} + \delta_\theta(z_i, \mathbf{c^0})$$

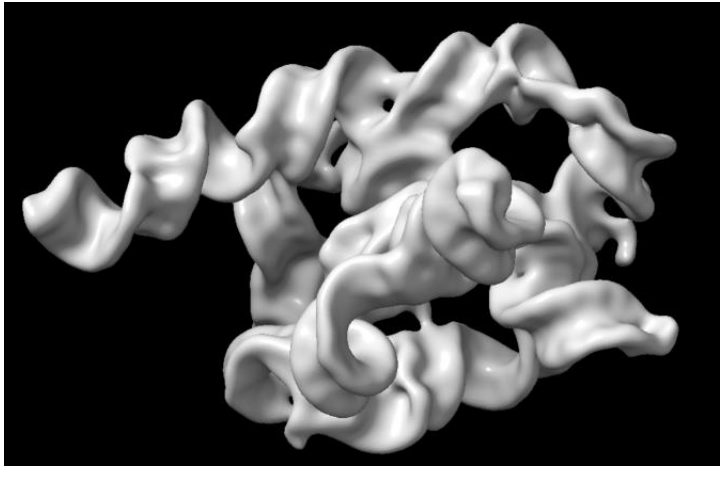
$$\mathcal{R}(E) = \sum_{\{(i,j)\,:\,E_{ij}=1\}} |d(c_i, c_j) - d(\mathcal{D}(c_i, z), \mathcal{D}(c_j, z))|^2,$$

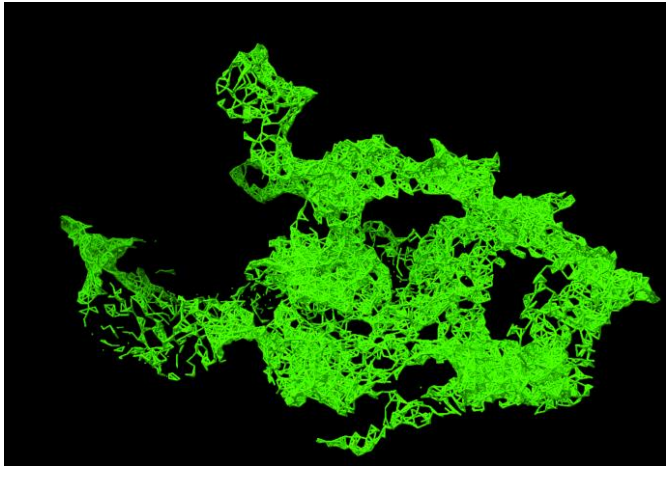
2. If gaussian pseudo-atom are within a distance of 1.5 times the average distance between all Gaussians and their two nearest neighbours, assume they are bonded and impose same constraint as above.

5000A 5000M 100kA 100kM

Computational Toolbox Required for Analysis

- Converter of a consensus .pdb structure to high resolution .mrc
- Converter for .xyz coarse-grained atomic coordinates to a pdb that can be opened in chimera



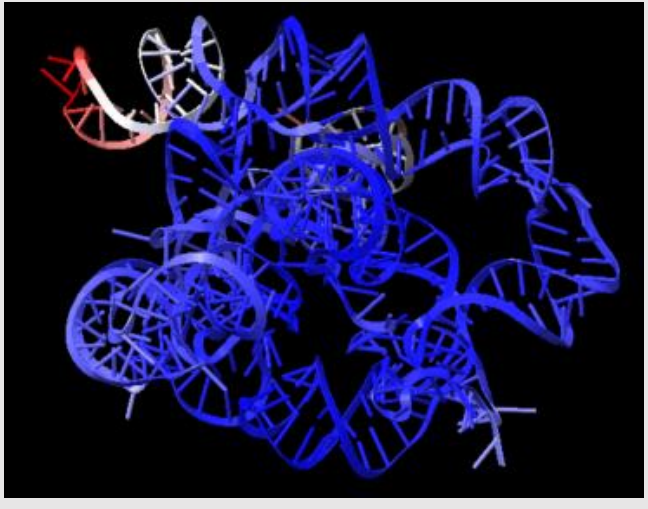


Average Global RMSD

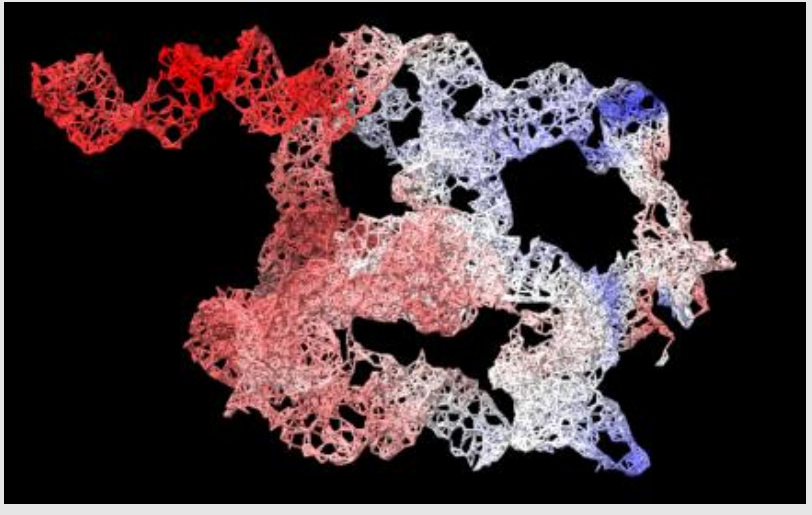
- Ground Truth : 4.3355 A

- 5000A : 0 A
- 100kA : 0 A
- 5000M :11.877206 A
- 100kM : 11.2558362 A

RMSD per atom 5000M

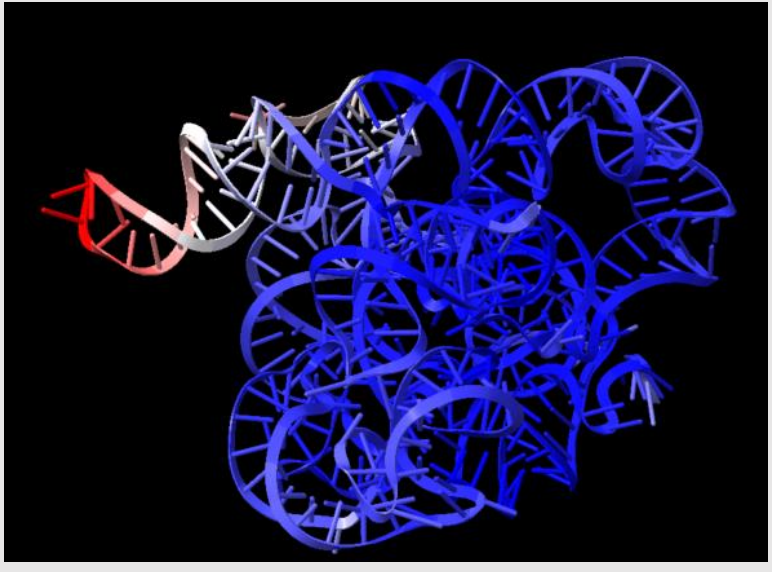


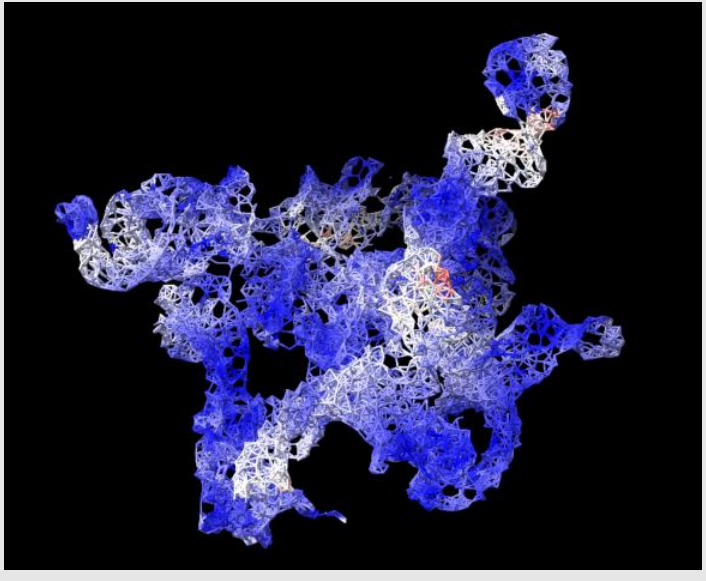
0 to 34.6 A



0 to 34.6 A

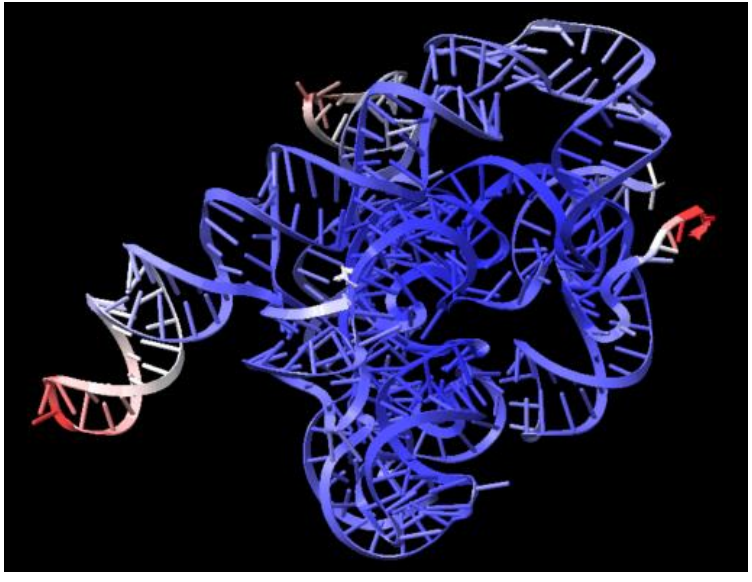
RMSD per atom 100kM

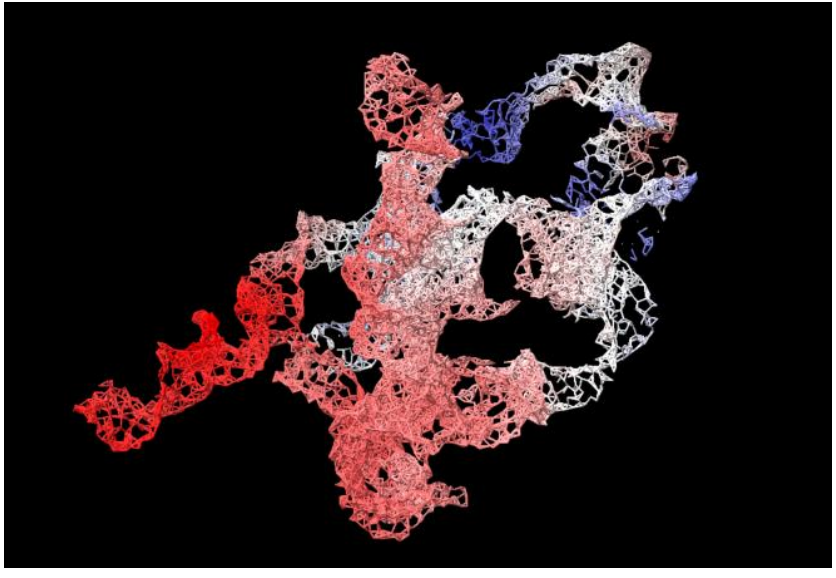




0.16 to 34.6 A

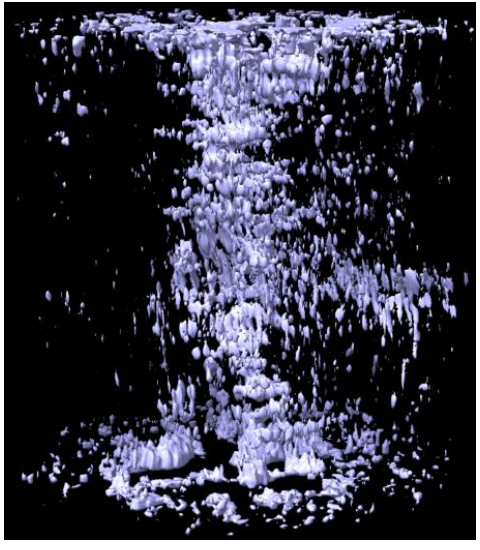
RMSF per atom 5000M

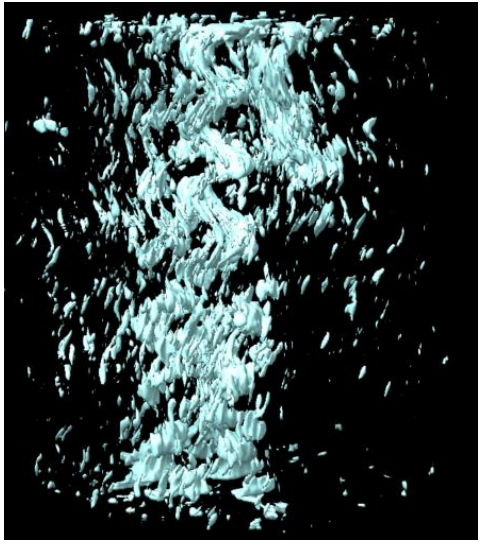


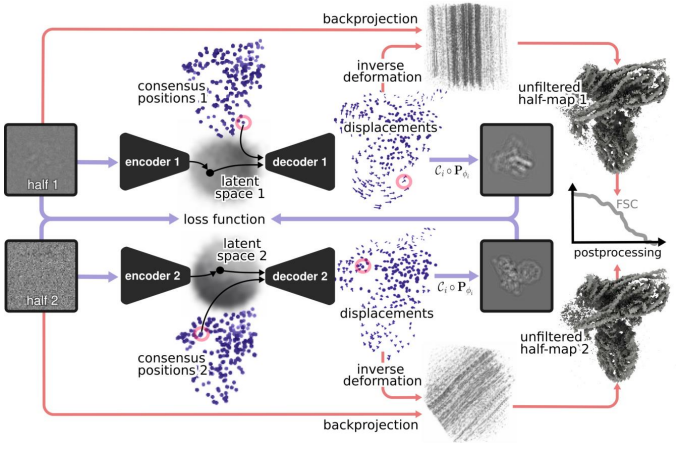


0 to 13.5 A

Results: Resolution of Final Structures 5000M







Conclusion

- DynaMight demonstrates regions of variability
- Not a reliable program to use (yet!) to understand heterogeneity of RNA-only structures

Future Work

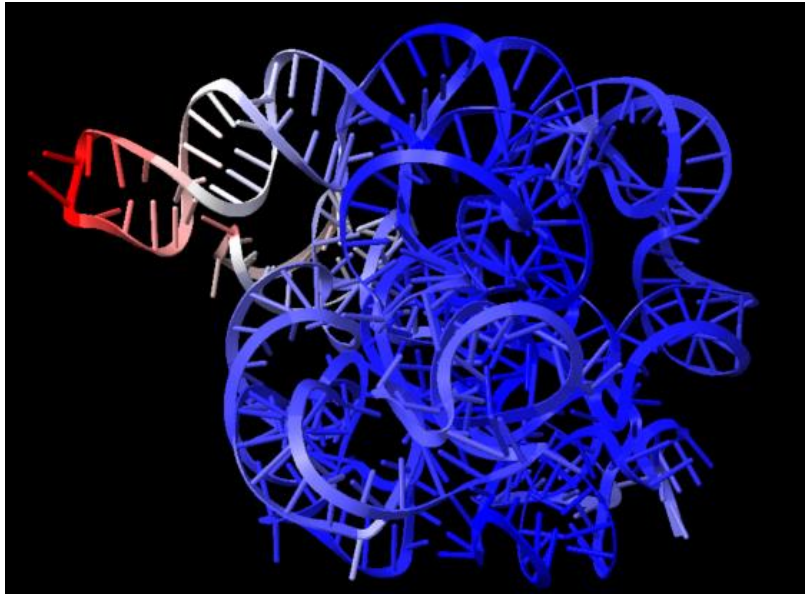
- Do a more robust analysis of DynaMight now that we have debugged many issues related to all RNA structure
  - Using Atomic model
  - Recover improved resolution of final structures
  - Compare to e2gmm
  - Improve model
  - Compare to a manifold embedding approach
- Choose an new RNA complex to study
  - *Structures of co-transcriptional RNA capping enzymes on paused transcription complex* (May 30 2024)

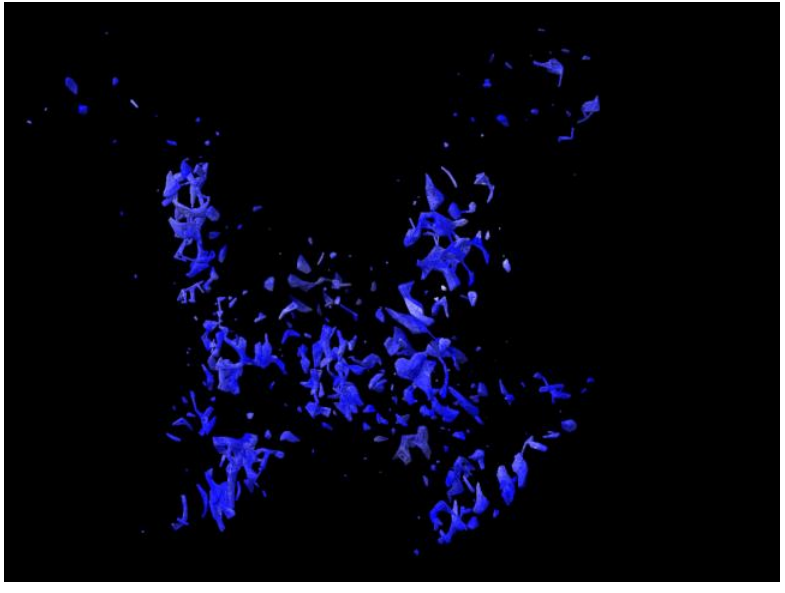
Future Work:

1. SLAC Molecular Movie ?
2. DFT calculations → MLP MD simulation → More reliable ground truths
3. GOLDD/ROOL, FMN Riboswitch

Thank You

RMSF per atom 100kM





0 16 to 34.6 A