Deep Learning for Molecular Dynamics Prediction
CS 229B Course Project

Robin Cai (rcai2@stanford.edu)
Leah Reeder (lreeder@stanford.edu)

Problem Formulation

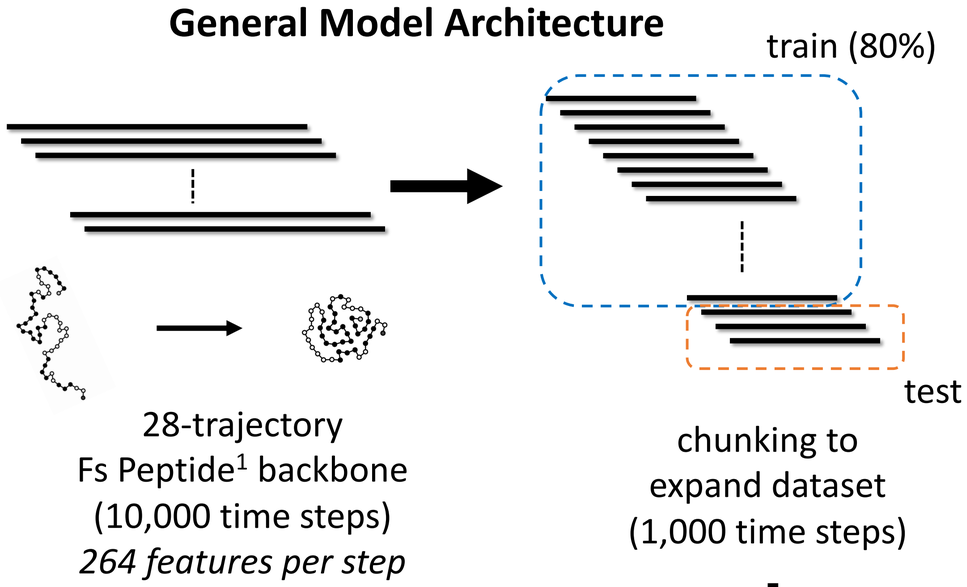
Simulating complex physical equations at short time steps, **molecular dynamics** (MD) typically has **high computational demands**, making it challenging to simulate biological processes over **extended timescales** of interest.

We explore the efficacy of deep learning approaches to learn:

1. molecular structures
2. underlying physical patterns

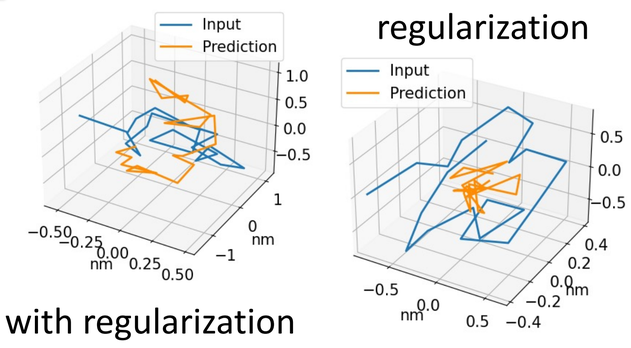
and thus, **predict future sequences** based on a **single starting structure**.

General Model Architecture



Key design component:

Apply **regularization** to prevent **dense clustering** at the center of mass.



Main evaluation criteria: root mean square deviation (RMSD)

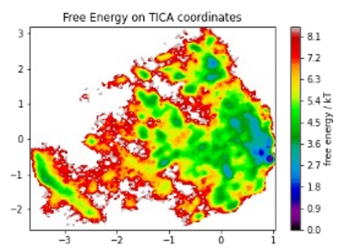
Traditional Techniques for MD Analysis

Markov State Models (MSMs) effectively analyze molecular dynamics but require:

1. the modeler's **deep biological insight**
2. an **accurate estimation** of the Markovian metastable states.

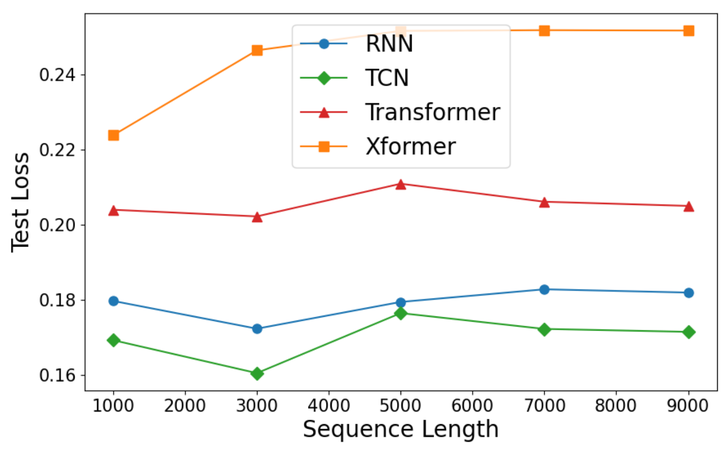
Workflow[3]:

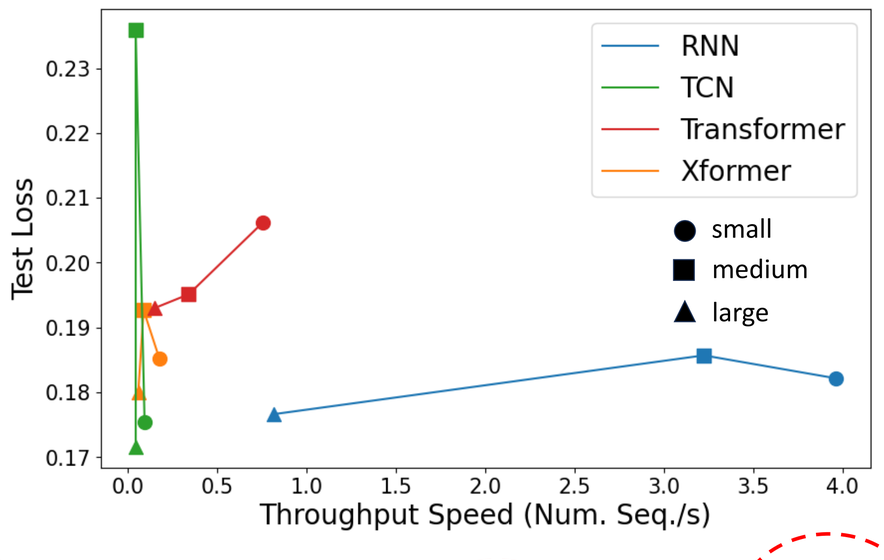
1. **Feature extraction** from trajectories
2. **Reduce dimensions** with time-independent component analysis (**TICA**)
3. Apply **K-means clustering**
4. Construct MSM
5. Aggregate into **metastable states**
6. Predict with **transition probability matrix** of metastable states

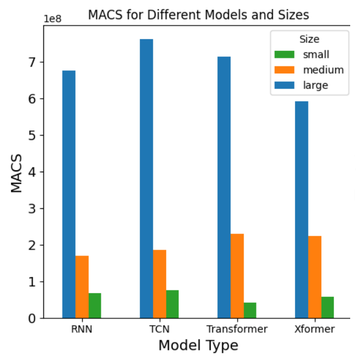


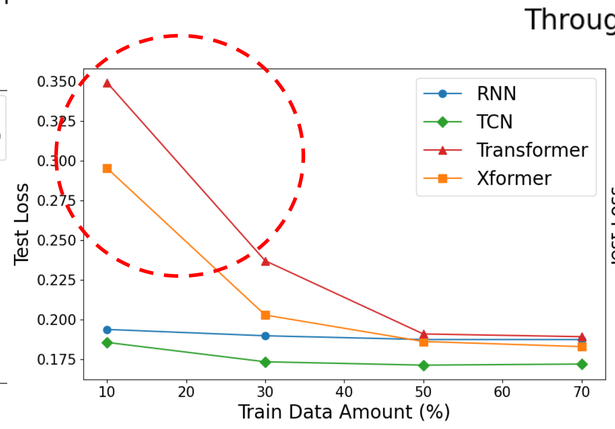
Avg RMSD ~ 8Å

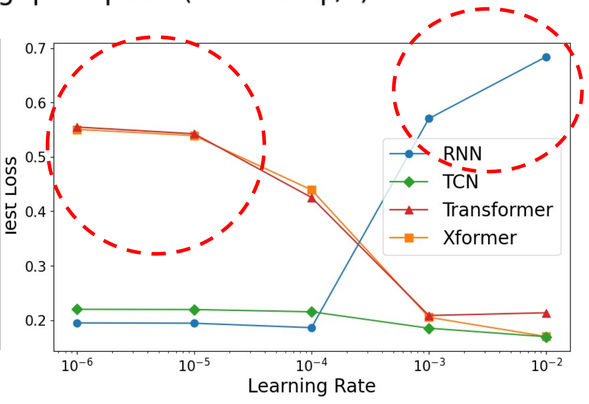
Model Comparisons











Strength:

- RNN — Computationally efficient
- TCN — Robust to the choice of hyperparameter

Weakness:

- Transformer
- Xformer

Sensitivity to amount of training data

Avg RMSD ~ 4Å

Representation Learning

Our models attain RMSDs near native folds (~4Å), but still do not match the test input very well. To further refine our results, we tackled the complexity introduced by **rotational movements** in the molecular trajectories, which can complicate coordinate-based learning. By integrating representation learning with **Graph Neural Networks (GNNs)**, we adopted a **rotationally equivariant approach to representing coordinates**, enhancing the model's ability to learn from the data without being confounded by orientation changes.

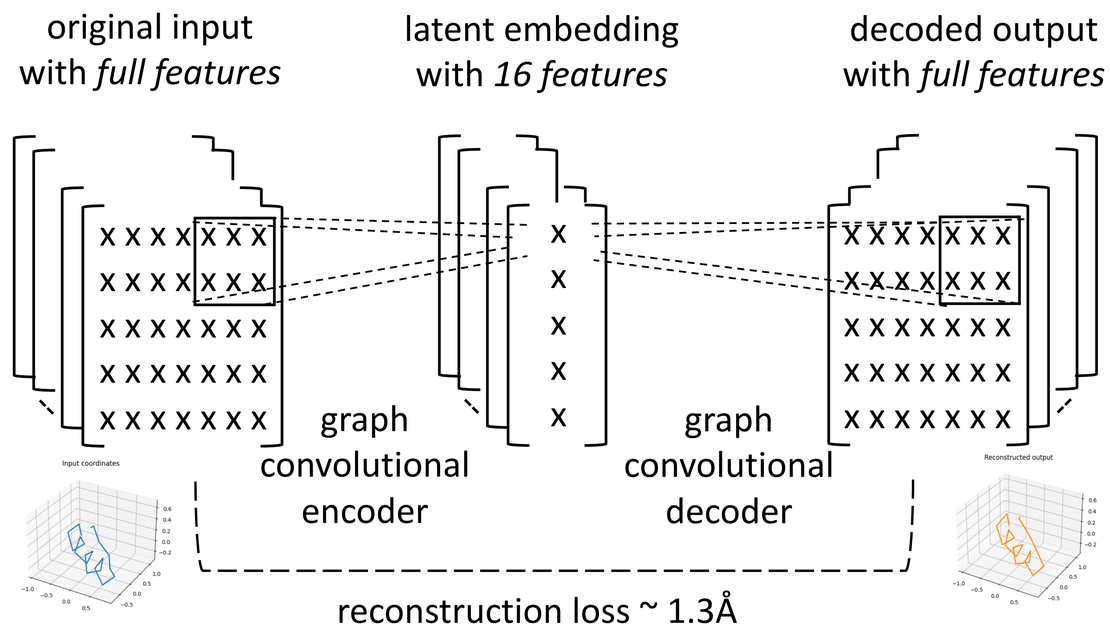
Node and edge embeddings[5] are created using:

1. C-α coordinates and orientations
2. Dihedral angles of backbone
3. Distances between sets of atoms

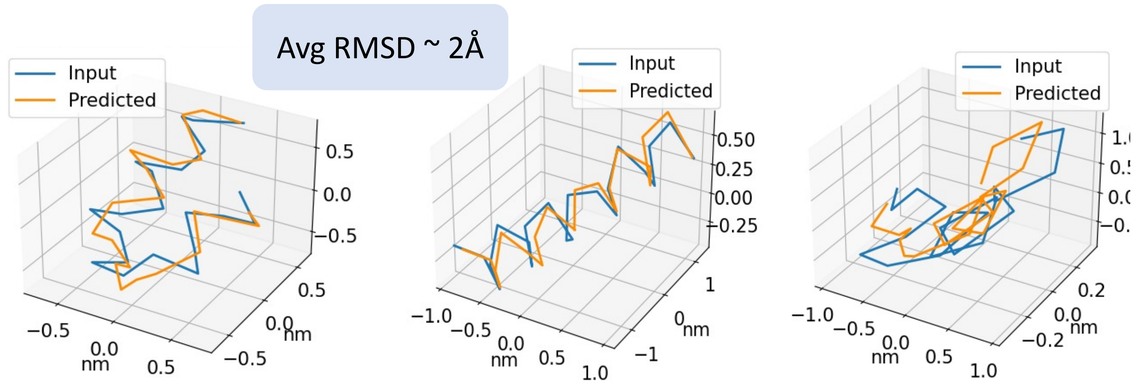
Our reconstruction loss was considered over the C-α coordinates.

The latent embedding becomes our new input to the sequence models. We can see that our predictions have a lower average RMSD (~2Å) and have the correct orientation.

Graph Convolutional Autoencoder (GCAE)



Avg RMSD ~ 2Å



Citations:
[1] R.T. McGibbon, *Fs MD Trajectories*, 2014. Data retrieved from figshare.
[2] B. Leufaudeux et. al. *xFormers: A modular and hackable Transformer modelling library.* 2022.
[3] M.K. Scherer et. al. *PyEMMA 2: A software package for estimation, validation, and analysis of markov models.* 2015.
[4] S. Bai et. al. *An empirical evaluation of generic convolutional and recurrent networks for sequence modeling.* 2018.
[5] B. Jing et. al. *Equivariant Graph Neural Networks for 3D Macromolecular Structure.* 2021.

Stanford | ENGINEERING
Chemical Engineering

ICME