Deep Learning Models for Protein Assembly

Denoising AutoEncoder for Protein Reconstruction

MATH-592: Semester Project CSE II

Francesca Venturi

Professor: Matteo Dal Peraro

Supervisor: Lucien Fabrice Krapp

Outline of the project

- Aim of the project
- Possible applications
- Approach:
  - Classification task
  - Autoencoder for protein reconstruction
- Protocol
- Results
- Improvements and Future work

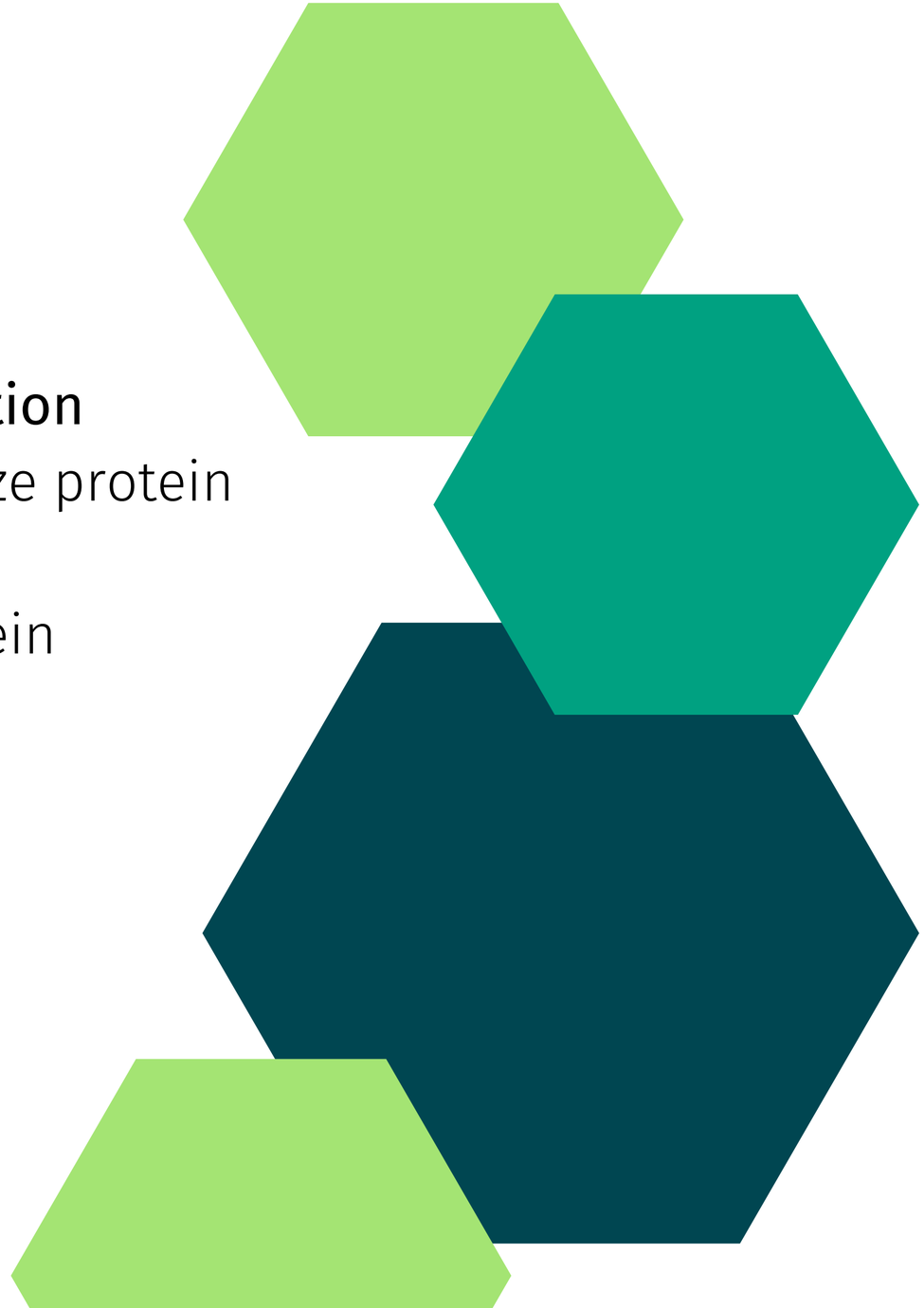
Aim of the Project

What: Structural Embedding

- Lower-dimensional space: retain **relevant information**
- Compare and classify protein arrangements, analyze protein structure and function
- From atomic to residue representation of the protein
- Data interpretability

How: Autoencoder

- Geometric Tranformer (PeSTo)
- Pooling Layer
- Unpooling Layer



Possible applications

- Faster than MD simulation
- Designing new proteins (hallucination)
- Latent Database

Faster than MD simulations

Conformation sampling:

- Generating molecular structures: small changes to the starting conformation (e.g. rotating bonds)
- Explore the **different possible arrangements** of a biomolecule and identify the most energetically favourable

Goal:

Collect information to represent the behaviour of the system.
The same goal can be achieved by means of **MD simulations**, which are, however, **computationally inefficient** and time-consuming.

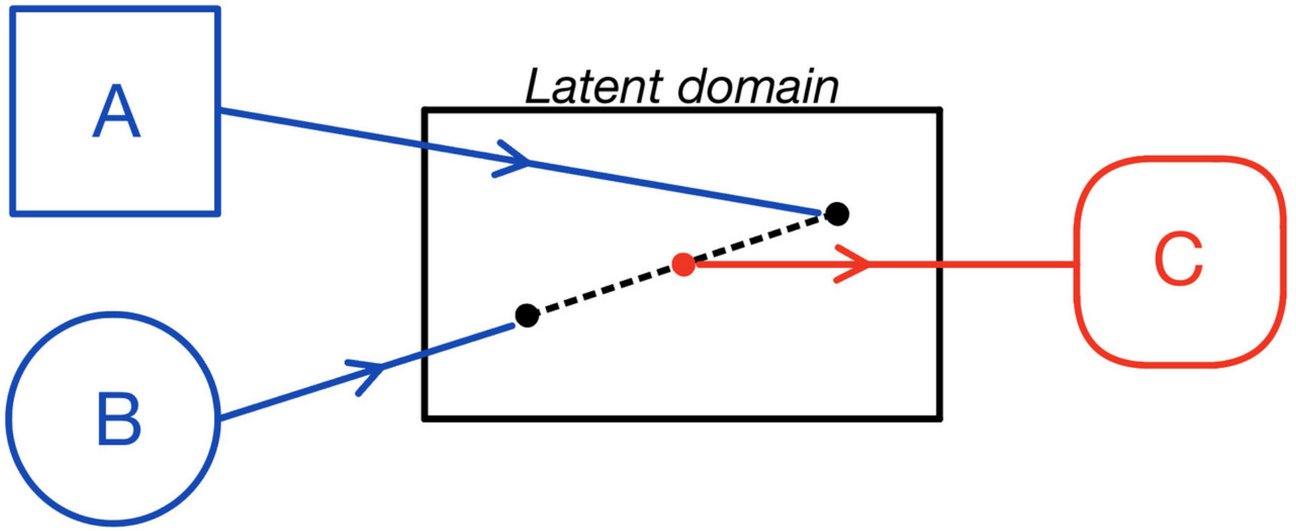
Example

Initial situation: 2 configurations A and B

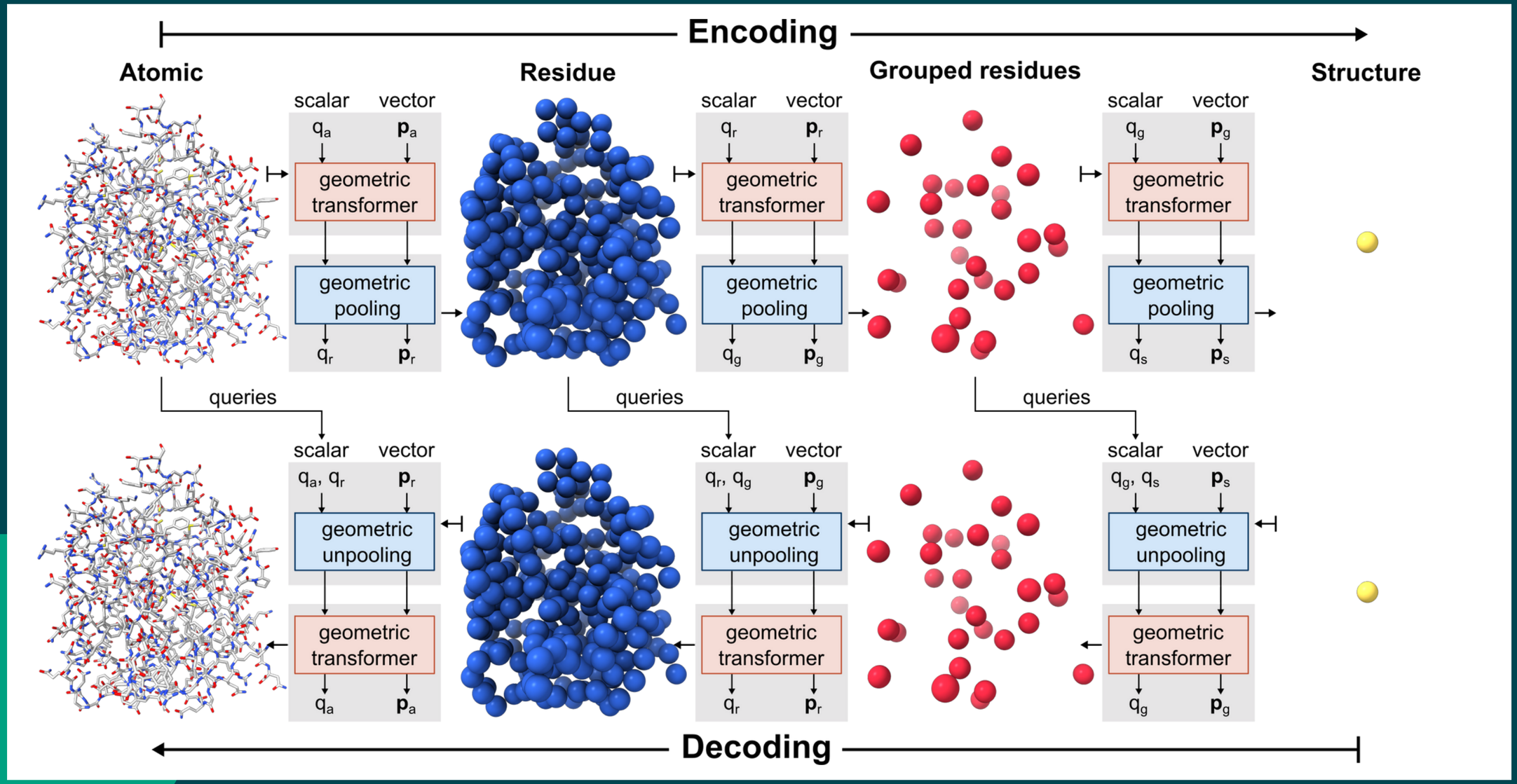
Goal: Sample an intermediate conformation C

How:

- **Latent representation** of A and B
- Intermediate step C in the dimensionally reduced domain
- Finally, use the decoder to produce artificially the expanded (atomic) representation of configuration C

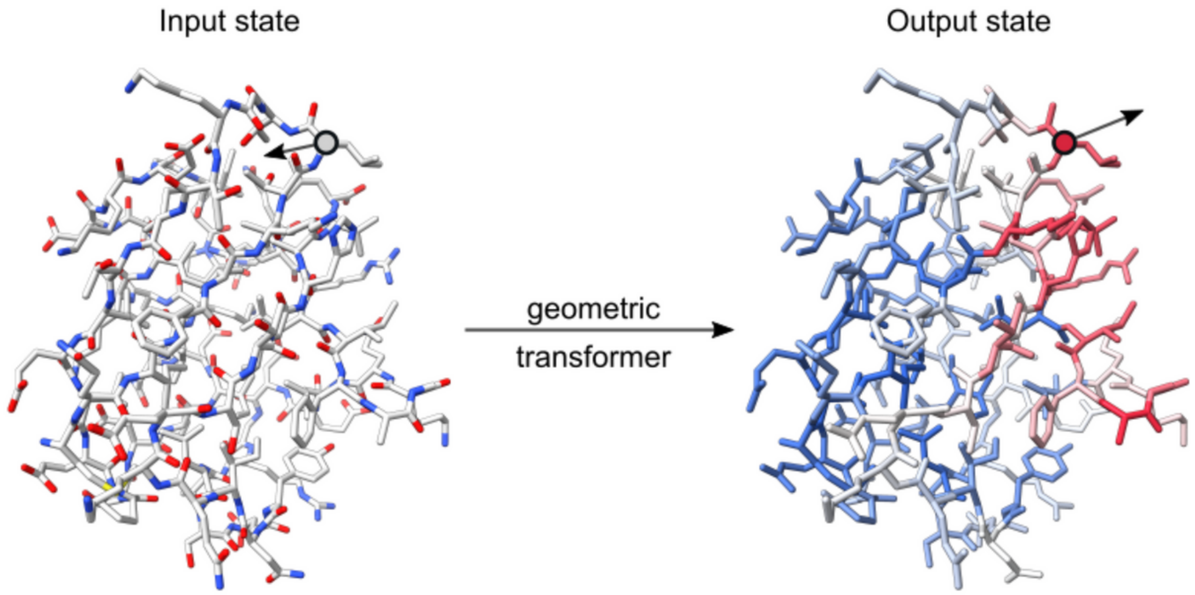


Approach: the AutoEncoder



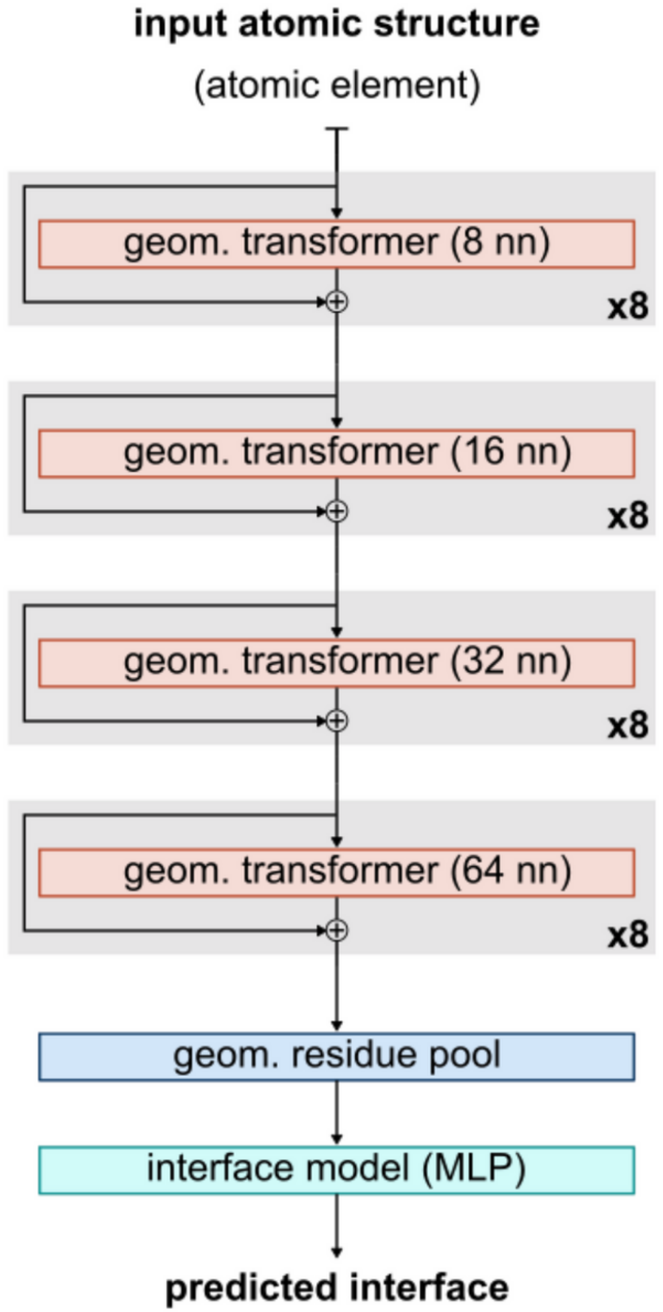
Geometric Transformers

- Scalar (q) and vector states (**p**)
- (Self-) Attention mechanism
- Progressive update through geometric chain transformers



- ever-widening neighbourhood

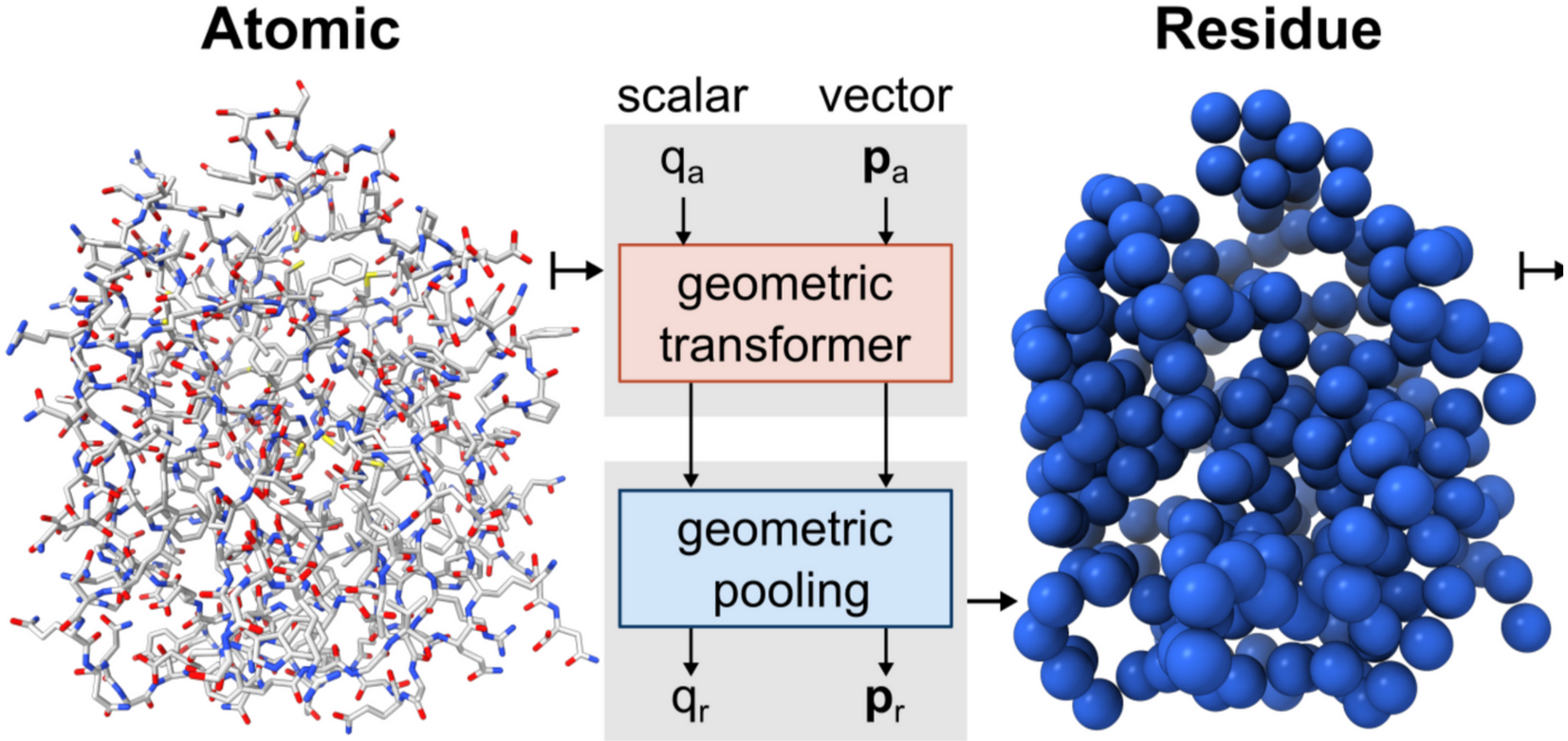
→ propagate information to a greater distance



The Pooling Layer

Compression: from atomic-level to residue-level representation of the protein

How: exploiting **attention mechanisms** and the membership of an atom in a residue



This transition is independent of the number of atoms within a residue.

Chemical Characteristics Prediction

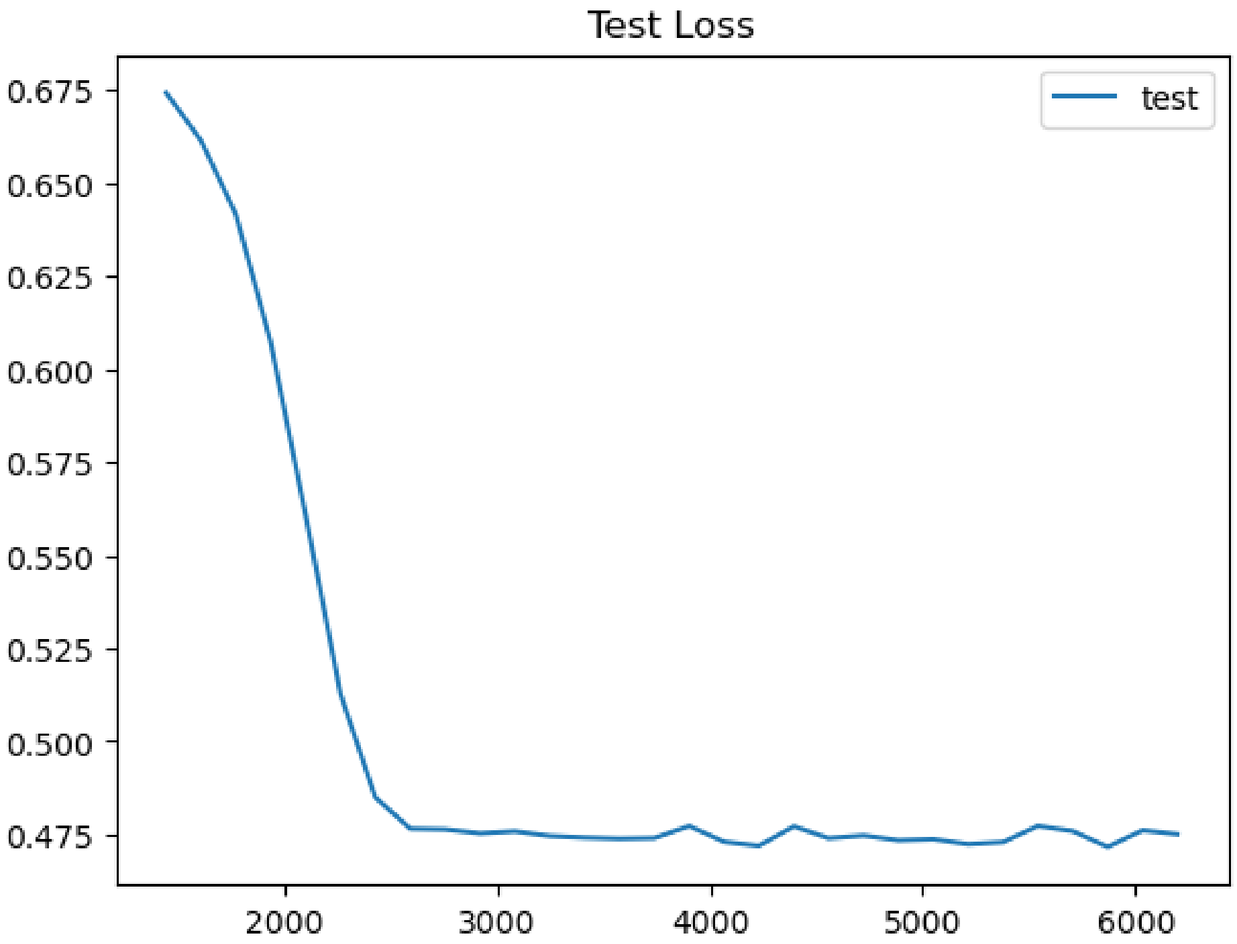
Goal: Predict chemical characteristics of the side chain of every residue starting from its backbone.

How:

- Extract the backbone atoms
- Geometric transformer
- Pooling layer

→ Classification task (5 classes)

1. hydrophobic
2. positive charge
3. neutral
4. negative charge
5. special cases



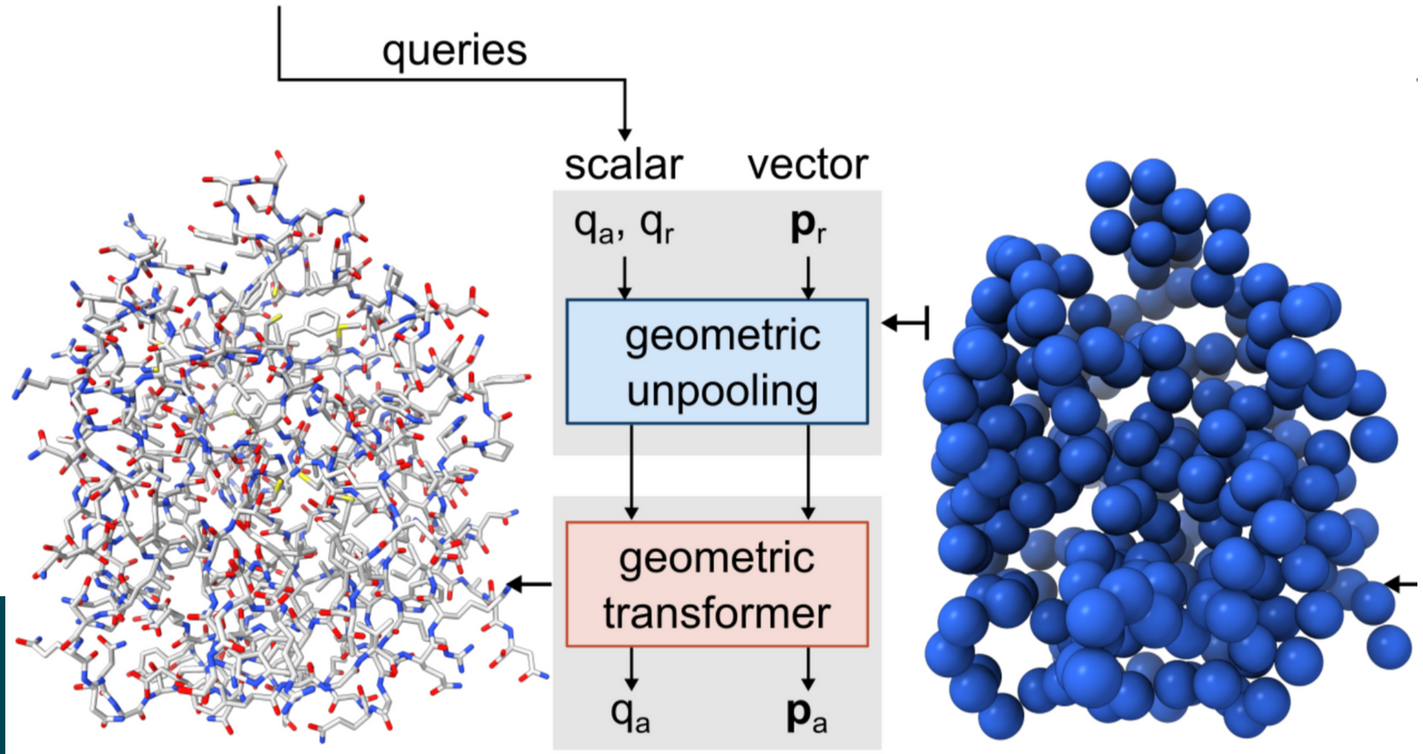
criterion = BCEWithLogitsLoss

The Unpooling Layer

Expansion: from residue-level to atomic-level representation of the protein

How: attention mechanism and membership of an atom in a residue
(reverse of the pooling layer)

Skip connections: from residue-level to atomic-level representation of the protein



Different Unpooling Structures

How many: 3 increasingly complex models

1

LINEAR EXPANSION THROUGH MATRIX MULTIPLICATION

`qa = pt.matmul(M, qr)`

2

- SELF-ATTENTION AT RESIDUE LEVEL
- LINEAR EXPANSION THROUGH MATRIX MULTIPLICATION

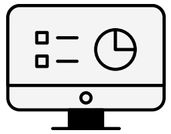
3

- SELF-ATTENTION AT RESIDUE LEVEL
- SELF-ATTENTION EXPANSION

Protocol

- Dataset
- Models
- Training
- Evaluation

Dataset



alphafold_structures_16384_64nn_v2.h5

Proteins

The dataset contains 555349 proteins, among which 22742 are human proteins

Atoms

A single protein is made of N atoms. Attributes of an atom:

- element
- name
- residue id
- residue name
- atomic coordinates

Preprocessing

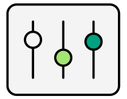
Extract relevant information

- X0: **atomic coordinates**
- qe: **atomic element names**
- qr: **residue names**
- M: **membership atom-residue matrix**

Remark: qe, qr and M are **one-hot** encodings of the corresponding quantities

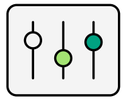
Model Architecture - Encoder

Many slightly different versions



Atomic Level Geometric Transformer:

- Multiple Layers (**8**) of **Geometric Transformers**
- **Increasing** amount of **nearest neighbours** (16, 32, 64)
- **State** dimension: **32**
- **Attention** Heads: **2**
- Batch size: N (number of atoms)

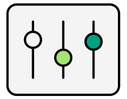


Pooling Layer

- **State** dimension: **64**
- **Attention** Heads: **2**
- Input and output size: **N --> N_res** (number of residues)

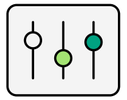
Model Architecture - Decoder

Many slightly different versions



Residue Level Geometric Transformer:

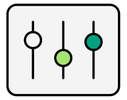
- Multiple Layers (**8**) of **Geometric Transformers**
- **Increasing** amount of **nearest neighbours** (8, 16, 32)
- **State** dimension: **64**
- **Attention** Heads: **4**
- Batch size: N_res (number of atoms)



Unpooling Layer

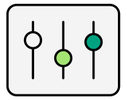
- **State** dimension: **32**
- **Attention** Heads: 2
- Input and output size: **N_res --> N**

Model Architecture - Embeddings and Projections



Features Encoding Models

- **Multi Layer Perceptron** with ELU activation function
- Embedding dimension: 32



State Vector Projection

- State dimension: **64+32**
- **Projection onto the 3D space** (coordinates updates)

Training

Train - Test Ratio	**5% - 95% (94 - 1722)** **20% - 80% (370 - 1446)** **50% - 50% (923 -893)**
Noise/Scramble Variance	**0.2** **0.5** **1** **2**
Layers of Geometric Transformer	**1 (16 atomic nn, 8 residue nn)** **2 (16+32 atomic nn, 8+16 residue nn)** **3 (16+32+64 atomic nn, 8+16+32 residue nn)**

Remark: mask the dataset over a given **minimum and maximum residue size** (from 22742 to 1816 proteins)

Fixed parameters

- Learning rate = **1e-4**
- Epochs = **100**
- Batch size = 1 protein (variable dimension)
- Optimizer = **Adam**

Evaluation

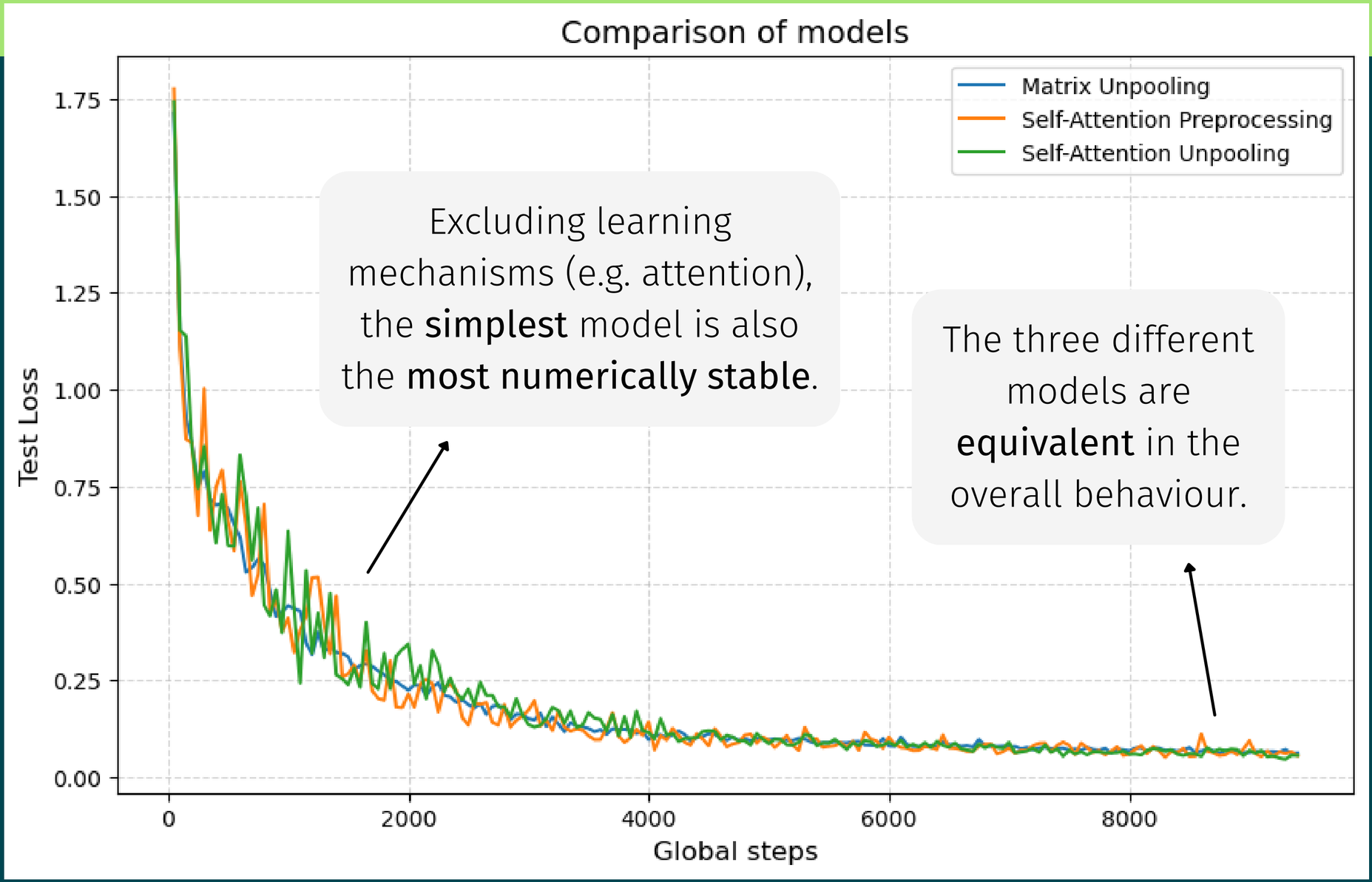
Different metrics for evaluation, computed based on the original atomic coordinates and their predictions.

1. **Mean Square Error (MSE) (Backpropagation)**
2. **Root- Mean-Square Deviation (RMSD)**
3. **Local Distance Difference Test (lDDT)**

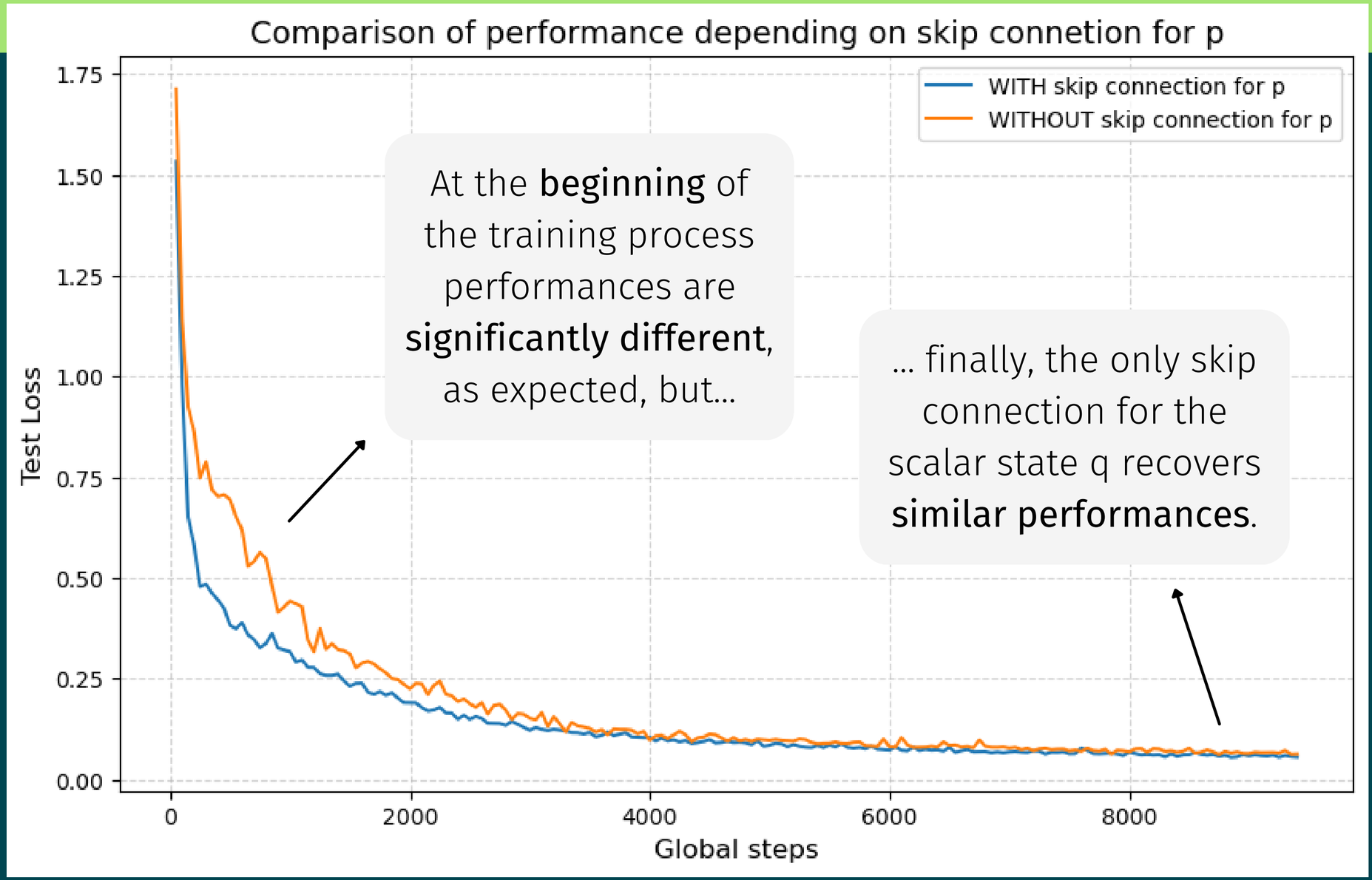
Results

- Skip connection for the vector state
- Different Unpooling models
- MSE, RMSD, lDDT
- Larger model architectures
- Increasing the training set size
- Noise threshold for reconstruction and scrambling coordinates

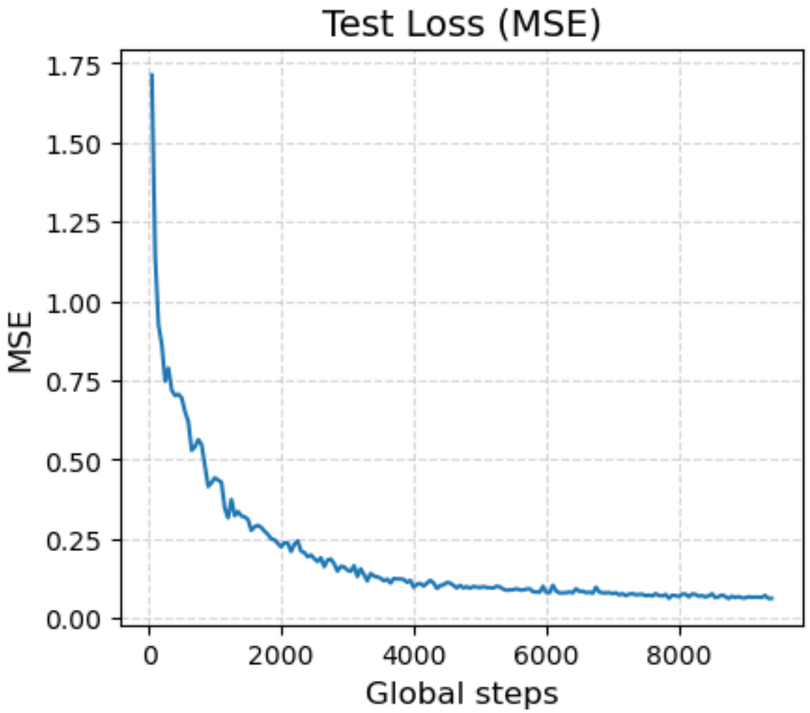
Different Unpooling Models

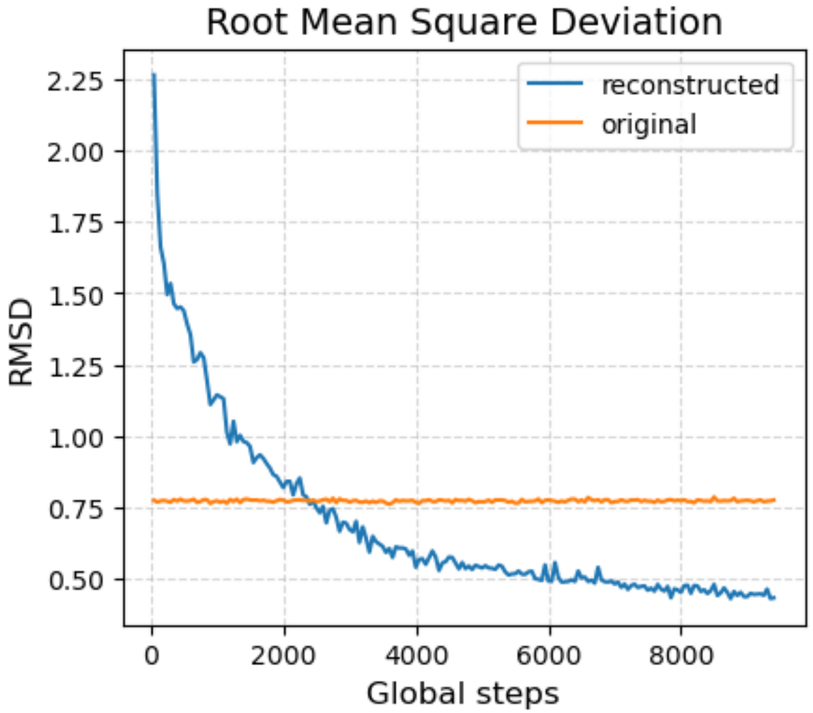


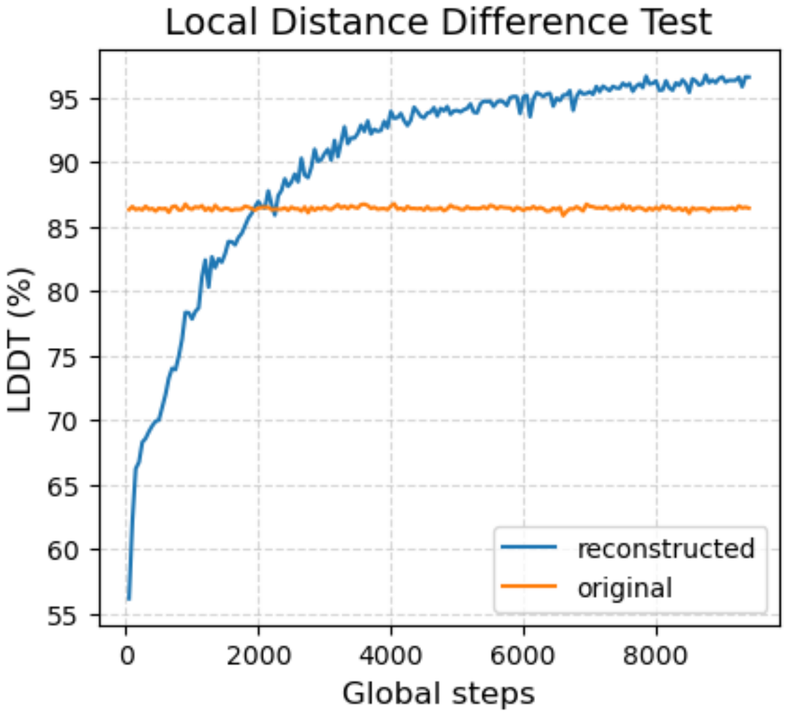
Skip Connection for the state vector p



MSE, RMSD, lDDT



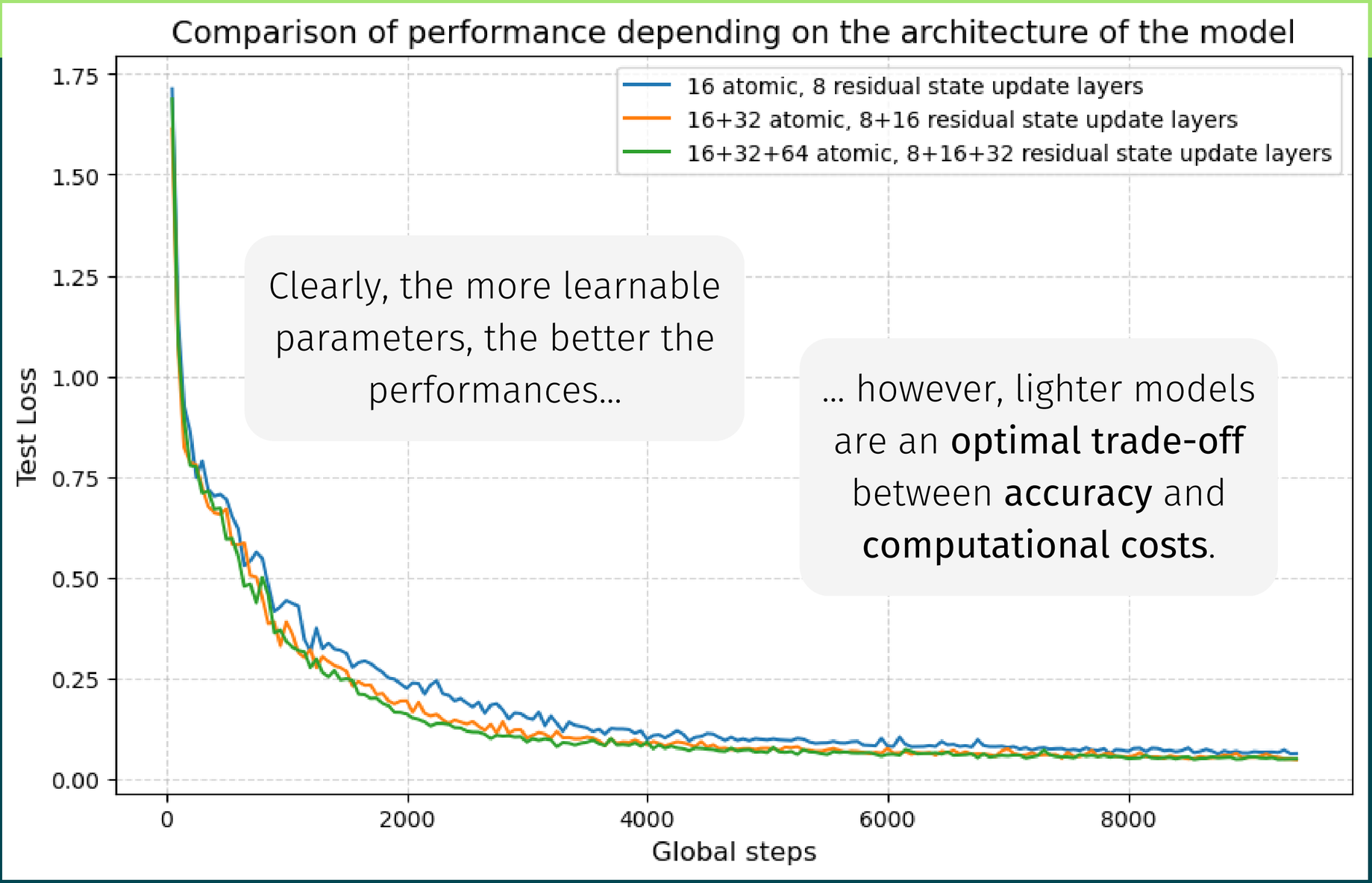




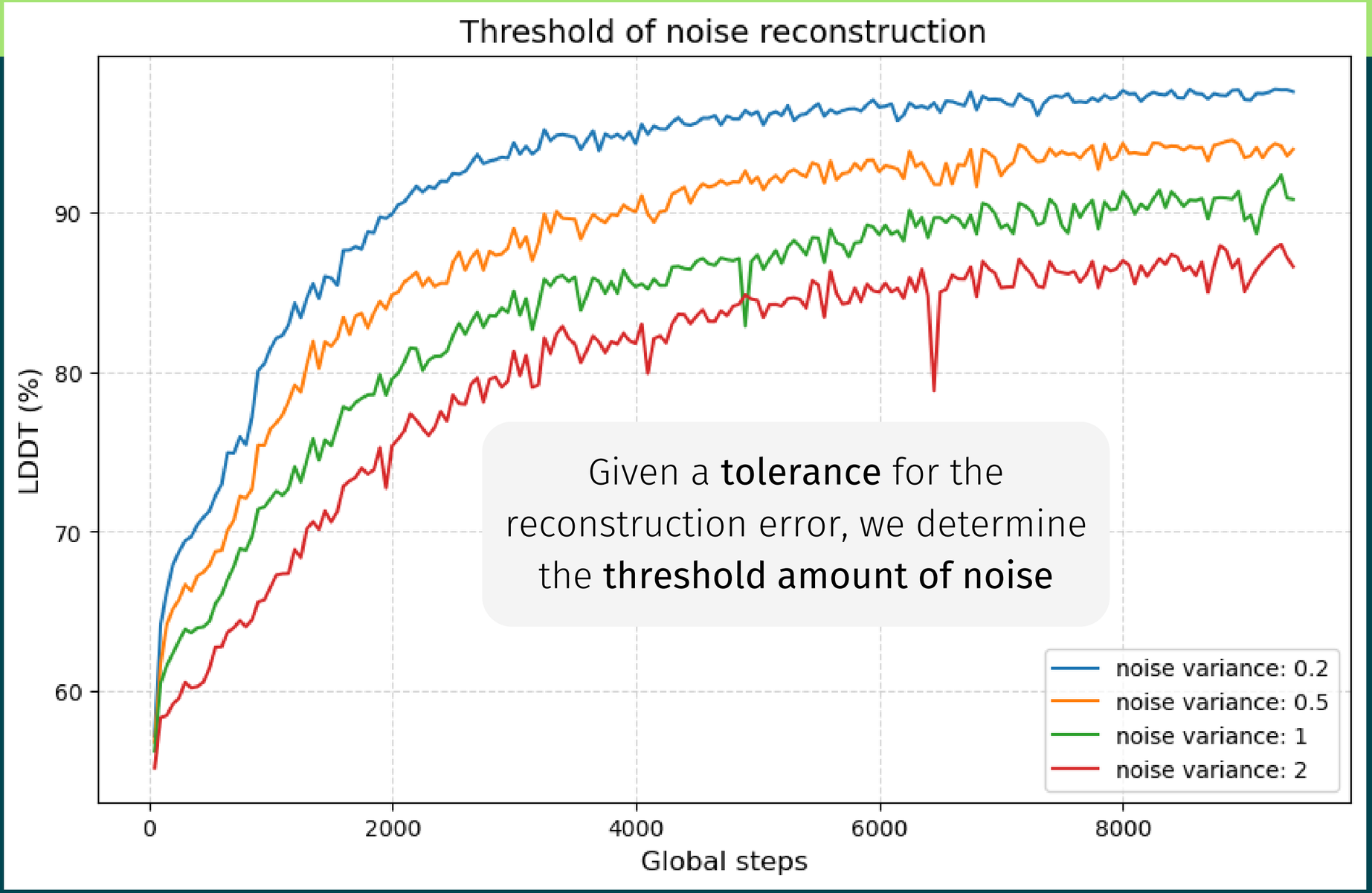
The **loss decreases**, so the model is learning.

The reconstruction error is completely recovered and, most importantly, **no overfitting** occurs.

Larger Model Architecures

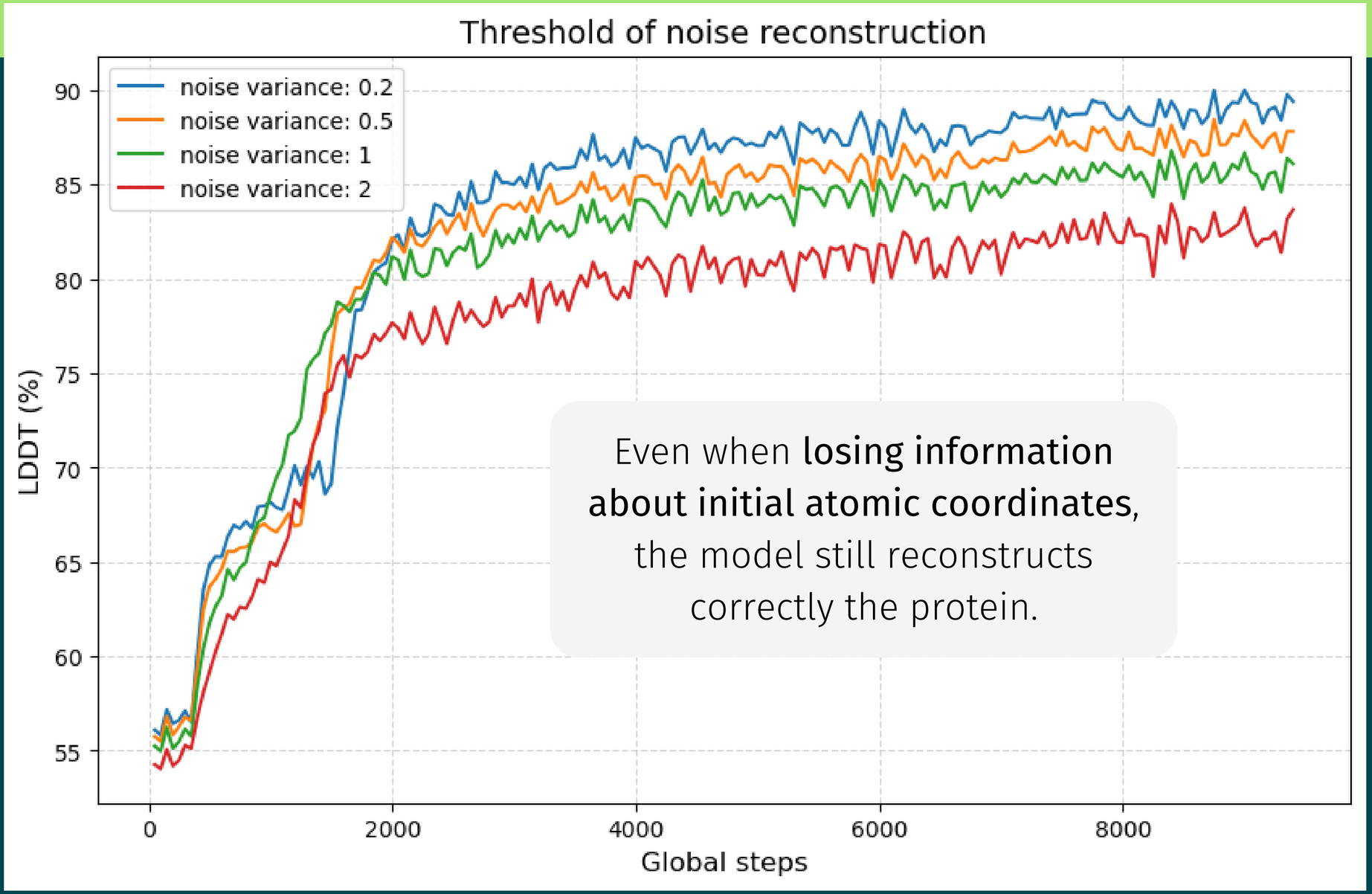


Noise threshold for reconstruction



Given a **tolerance** for the reconstruction error, we determine the **threshold amount of noise**

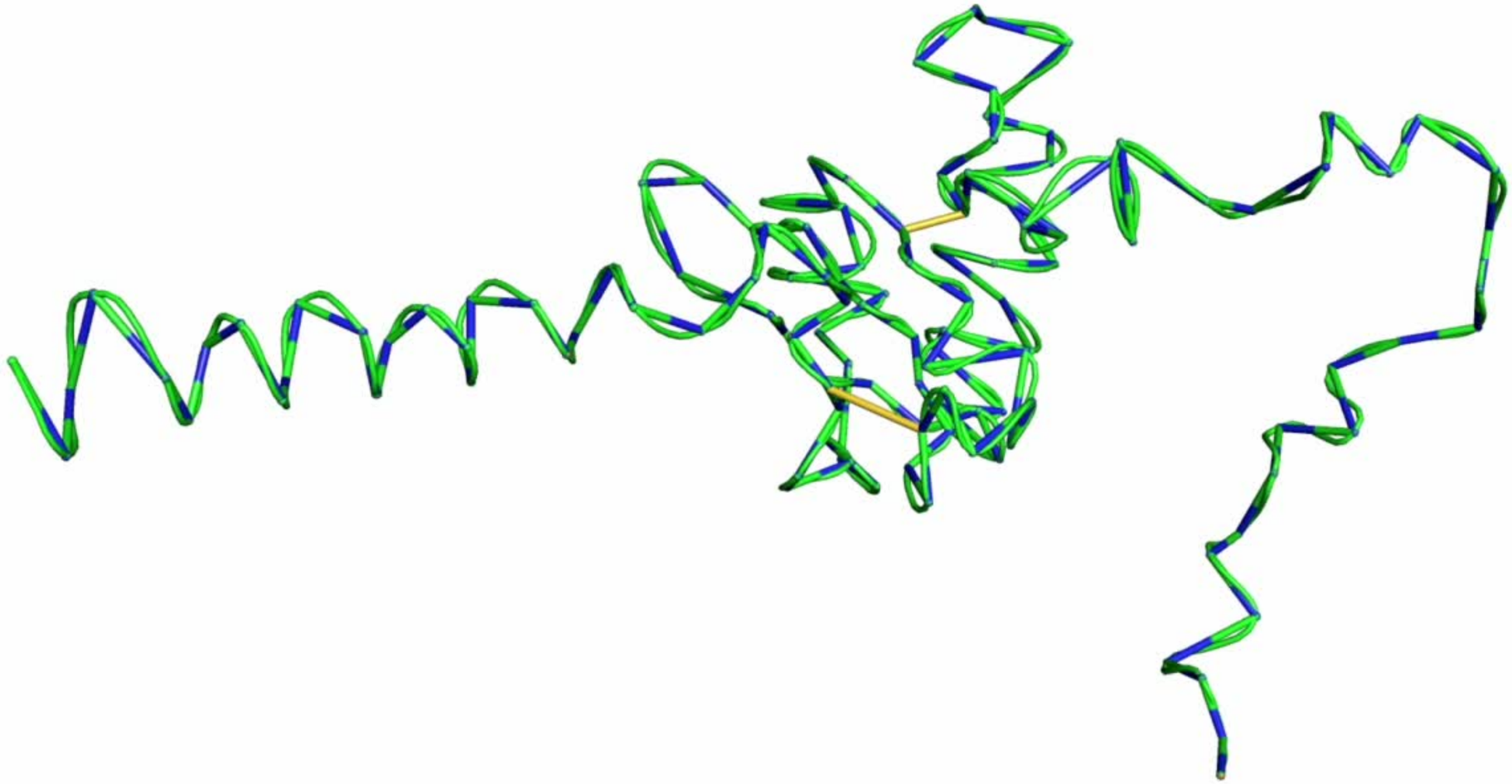
Scrambling coordinates



Even when **losing information about initial atomic coordinates**, the model still reconstructs correctly the protein.

What happens when training

NO MORE PLOTS AND NUMBERS,
JUST A WAY MORE INTUITIVE VISUALISATION OF THE LEARNING PROCESS:



An open eye to the future work

- Deeper Pooling and Unpooling:
  - Residues → Grouped Residues
  - Grouped Residues → Structure

- Learn the membership matrix M

- From Denoising to Variational AutoEncoders

- More complex Unpooling Layer's structure and architecture

Questions?

REFERENCES:

- Krapp, Lucien Fabrice, et al. "PeSTo: parameter-free geometric deep learning for accurate prediction of protein interacting interfaces." bioRxiv (2022): 2022-05.
- Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space
- Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
- Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021): 583-589.
- Varadi, Mihaly, et al. "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models." Nucleic acids research 50.D1 (2022): D439-D444.
- Tunyasuvunakool, Kathryn, et al. "Highly accurate protein structure prediction for the human proteome." Nature 596.7873 (2021): 590-596.
- van Kempen, Michel, et al. "Fast and accurate protein structure search with Foldseek." Nature Biotechnology (2023): 1-4.