Protein Structure Variance Prediction using Deep Learning and Molecular Dynamics Simulations

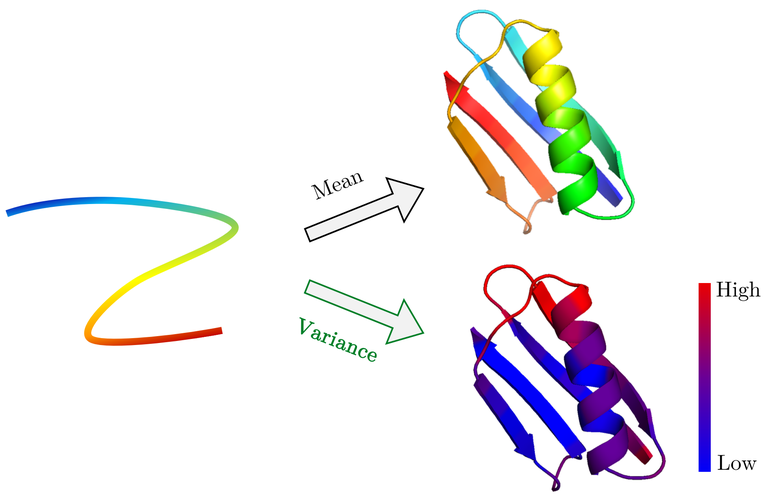
Marloes Arts ✉ ma@di.ku.dk , Wouter Boomsma ✉ wb@di.ku.dk

UNIVERSITY OF COPENHAGEN

Protein Structure Variance

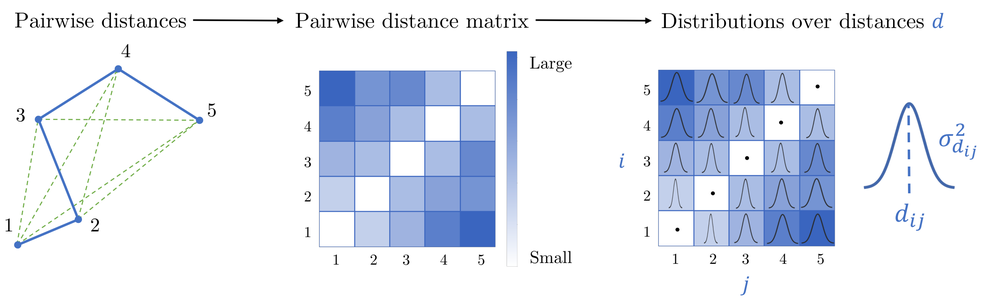
Proteins are involved in virtually all processes within cells. They consist of a string of amino acids, folded up into a **specific 3D structure** which is directly linked to the function of the protein.

Recently, there has been huge progress in predicting the static structure of proteins, but this is not the whole story. **Proteins are dynamic molecules**, and some parts have more variance than others.



Goal

Given the amino acid sequence and static structure of a protein, **predict the distribution over all pairwise distances within a protein**.



Motivation:

- Insight in protein **function**
- **Sampling** structures (e.g. data augmentation)
- *Future direction:* **weigh distances according to their variance** when predicting the mean structure from amino acid sequence (i.e. distances with less variance are probably more important for the main structure)

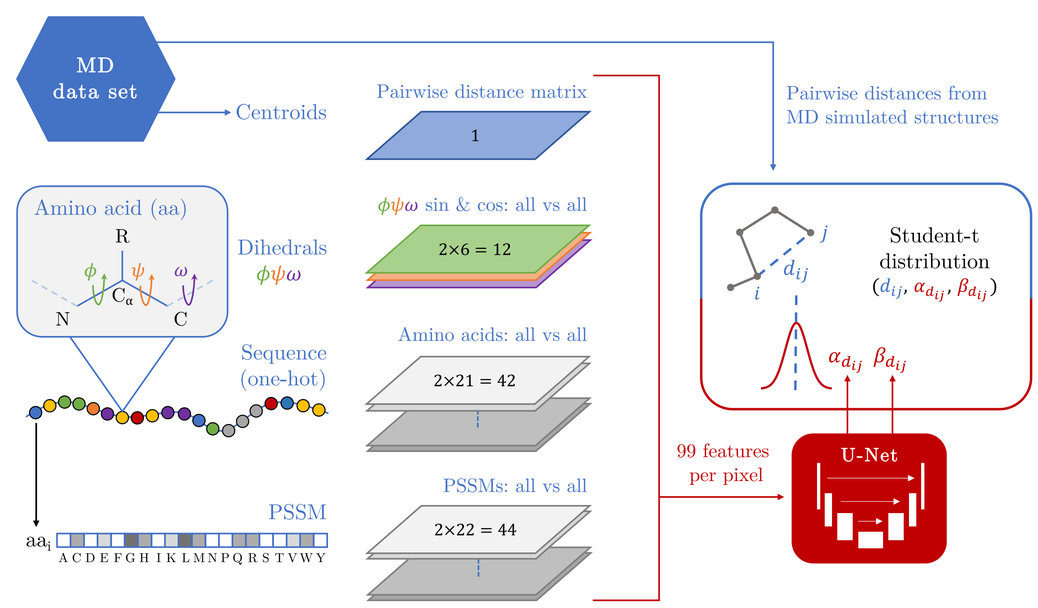
Data Set: Molecular Dynamics

We train and validate our network on a (soon to be published) data set that was constructed using **Molecular Dynamics (MD)**[1]. For each protein, we have 399 simulated structures with an interval of 50 ps. For this poster, we use all proteins with al length of ≤200 amino acids (474 train, 19 validation, 62 test).

Methods

- **Model**

Our model is based on a **U-Net**, a type of architecture that is well-known in imaging tasks such as segmentation. The input features are processed to form an "image" with 99 features per pixel.



- **Reparameterization of $\sigma_{d_{ij}}$**

For more robust training, we choose the variance to be distributed according to the **conjugate prior** of Gaussian data: the inverse-Gamma distribution.
This results in a **student-t distribution**[2].

$$\begin{aligned}\text{NLL}_{d_{ij}} &= -\log p_\theta\left(y_{d_{ij}}\right) \\ &= -\log \int \mathcal{N}\left(y_{d_{ij}} \mid \mu_{d_{ij}}, \sigma^2_{d_{ij}}\right) d\sigma^2_{d_{ij}} \\ &= -\log t_{\mu_{d_{ij}}, \alpha_{d_{ij}}, \beta_{d_{ij}}}\left(y_{d_{ij}}\right)\end{aligned}$$

Gaussian

Integrate out $\sigma^2_{d_{ij}}$

$y_{d_{ij}}$: target distance

$\mu_{d_{ij}}$: mean (centroid) distance

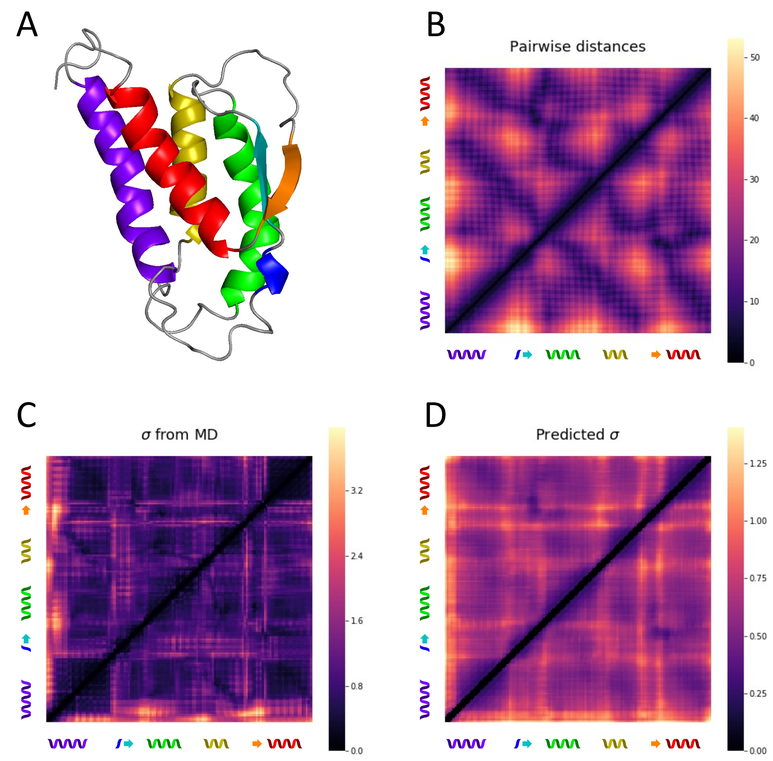
$\sigma^2_{d_{ij}} \sim$ Inv-Gamma$(\alpha_{d_{ij}}, \beta_{d_{ij}})$

(Preliminary) Results

Our network captures the pairwise distance **variance pattern** quite well.
Importantly, the predicted σ is a result of minimizing the NLL, **without the use of a ground truth σ**.

Example →

A. 3D structure ('1irl'), elements are colored and indicated in B-D
B. Pairwise distances
C. "Ground truth" σ from Molecular Dynamics
D. Predicted σ



Conclusion:

We propose a U-Net based network that can capture protein structure variance patterns given their amino acid sequence and static structure, solely by minimizing the negative log likelihood on structures simulated through Molecular Dynamics.

1. Data set built by Tone Bengtsen, previously a postdoc in our group.
2. Nicki S Detlefsen, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. arXiv preprint arXiv:1906.03260, 2019.