Protein B-Factor prediction based on 3D structure using graph neural networks

congruence TX

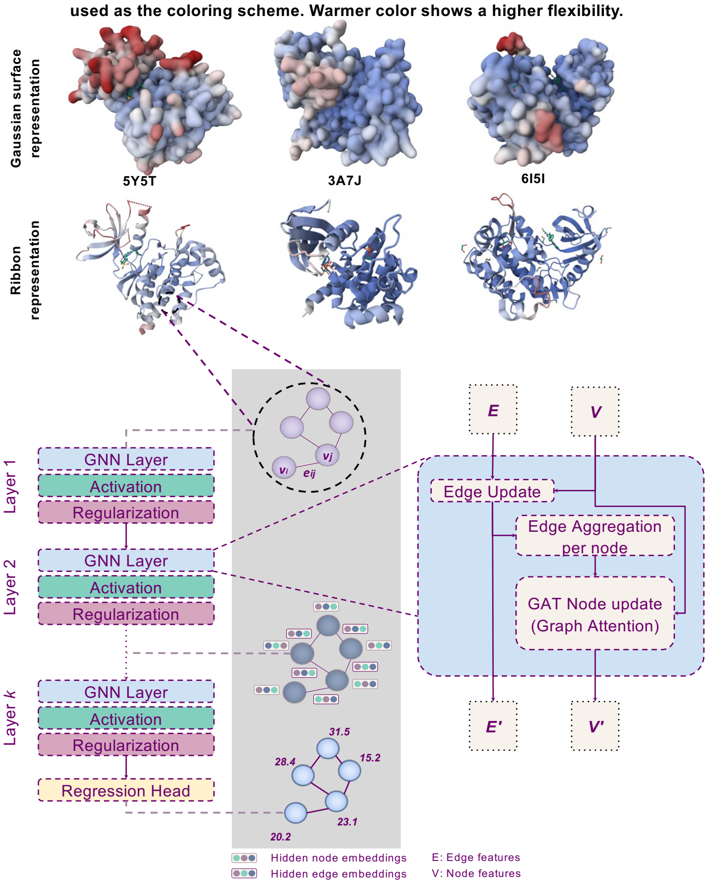
#33

Sina Sarparast*†, Aldo Zaimi*, Maximilian Ebert†, and Michael-Rock Goldsmith†
*Mila, †Computational Molecular Sciences Congruence[TX]

Abstract

Protein dynamics play a crucial role in many biological processes and drug interactions. Most existing methods for protein representation learning operate at the residue level, ignoring the finer details of atomic interactions. In this work, we propose for the first time to use graph neural networks (GNNs) to learn protein representations at the atomic level and predict B-factors from protein 3D structures. The B-factor reflects the atomic displacement of atoms in proteins and can serve as a surrogate for protein flexibility. We compared different GNN architectures to assess their performance.

Two representation of three proteins from the dataset with b-factor values used as the coloring scheme. Warmer color shows a higher flexibility.



Data and Representations

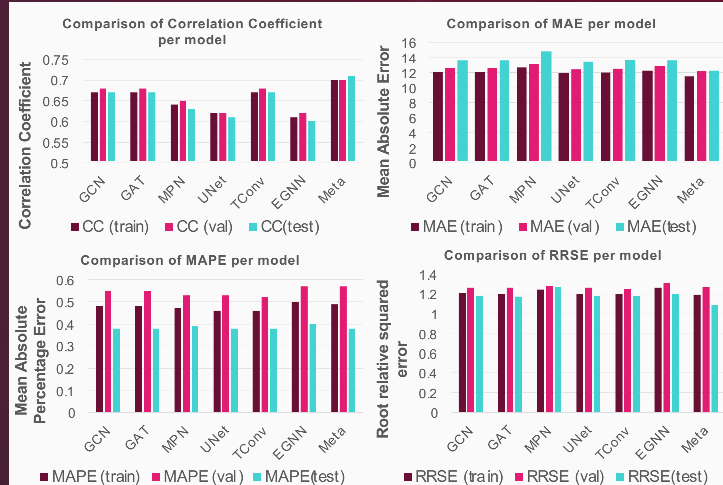
Feature Category	Feature Name	Description
Node	Atom type	The atom element is included. Most heavy atoms are carbon (C), oxygen (O), sulfur (S), nitrogen (N). Hydrogen atoms are not included.
Node	Atom relative location	Information about the atom's relative location (locant) in the residue is included. Atoms closer to the backbone typically present less mobility.
Node	Atom degree	The atom degree is computed as the number of atom neighbors present in the 5Å radius of the 3D space and normalized.
Node	Residue type	Properties like size, hydrophilicity, charge, and aromaticity are encoded for each of the 20 natural amino acids.
Edge	Covalent bond type	The type of bond (i.e., single, double, triple, delocalized) between two atoms.
Edge	Distance	The Euclidean distance between two connected nodes with 3D coordinates is computed and incorporated into the edge features.

Data and Representations

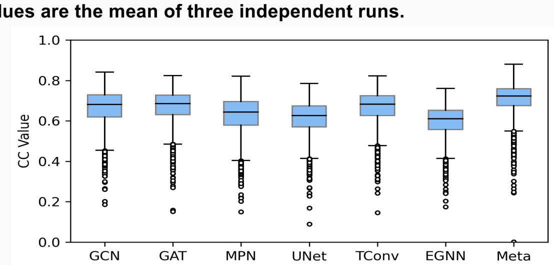
- Proteins are clustered into families based on sequence similarity, each exhibiting distinct functions and dynamics.
- Our data is sourced from the RCSB1 database, which contains over 100,000 protein structures.
- For training and validation: Diversely sampled the lowest resolution representative (Apo or Holo structure) from each protein family.
- For testing purposes: A separate set of Kinase proteins was sampled in the same manner.

Dataset	Number of proteins	Average number of residues	Average number of atoms	Total number of residues	Total number of atoms	Average resolution	Average B-Factor
Apo/Holo	2661	1004	5452	2.6M	14.5M	1.89 Å	31.3 Å²
Kinase	4313	663	3886	2.9M	17.4M	1.96 Å	36.9 Å²

Results



Training and validation performance metrics were computed on the Apo/Holo set. Values are the mean of three independent runs.



To assess quality of predictions per protein we visualize the Pearson correlation coefficient (CC) distributions of individual proteins from the Kinase test set for each model.

Key takeaways

- The Meta-GNN model achieves a correlation coefficient of 0.71 on a diverse test set of over 4,000 proteins (17 million atoms), outperforming other methods.
- Our work highlights the potential of GNN-learned representations for predicting protein flexibility and related tasks.
- Unlike most residue-based models, our models operate at the atomic level. This atomistic approach enables the prediction of ligand and co-factor B-factors.
- The models can be re-trained or fine-tuned on new PDB structures as they become available online, enhancing performance and generalization.
- These models offer a rapid means to estimate B-factors for homology-modeled protein structures, such as those generated by AlphaFold.
- Additionally, our flexible and modular framework can be easily adapted to predict other properties of interest (e.g., thermostability, surface exposure, surface electrostatics, and charge distribution) with minimal modifications.