Pre-training Sequence, Structure, and Surface Features for Comprehensive Protein Representation Learning

Hasun Yu
AI-based Drug Discovery Team at Kakao Brain

Proteins

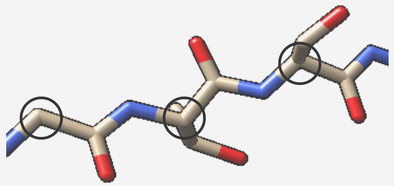
Proteins are vital components of biological systems and can be represented in various ways, including their sequences, 3D structures, and surfaces

Protein representation learning

Recent studies have successfully employed sequence- or structure-based representations to address multiple tasks in protein science

…ENNSPEHLKD…

Sequence



Structure

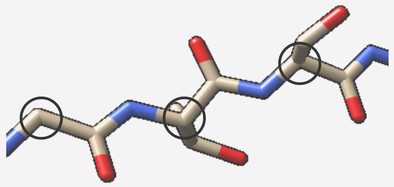
Protein representation learning

Recent studies have successfully employed sequence- or structure-based representations to address multiple tasks in protein science

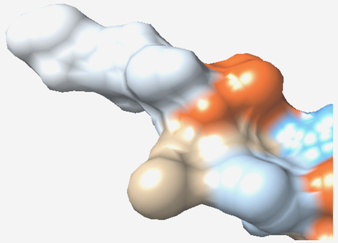
However, there has been an oversight in incorporating protein surface characteristics

…ENNSPEHLKD…

Sequence



Structure



Surface

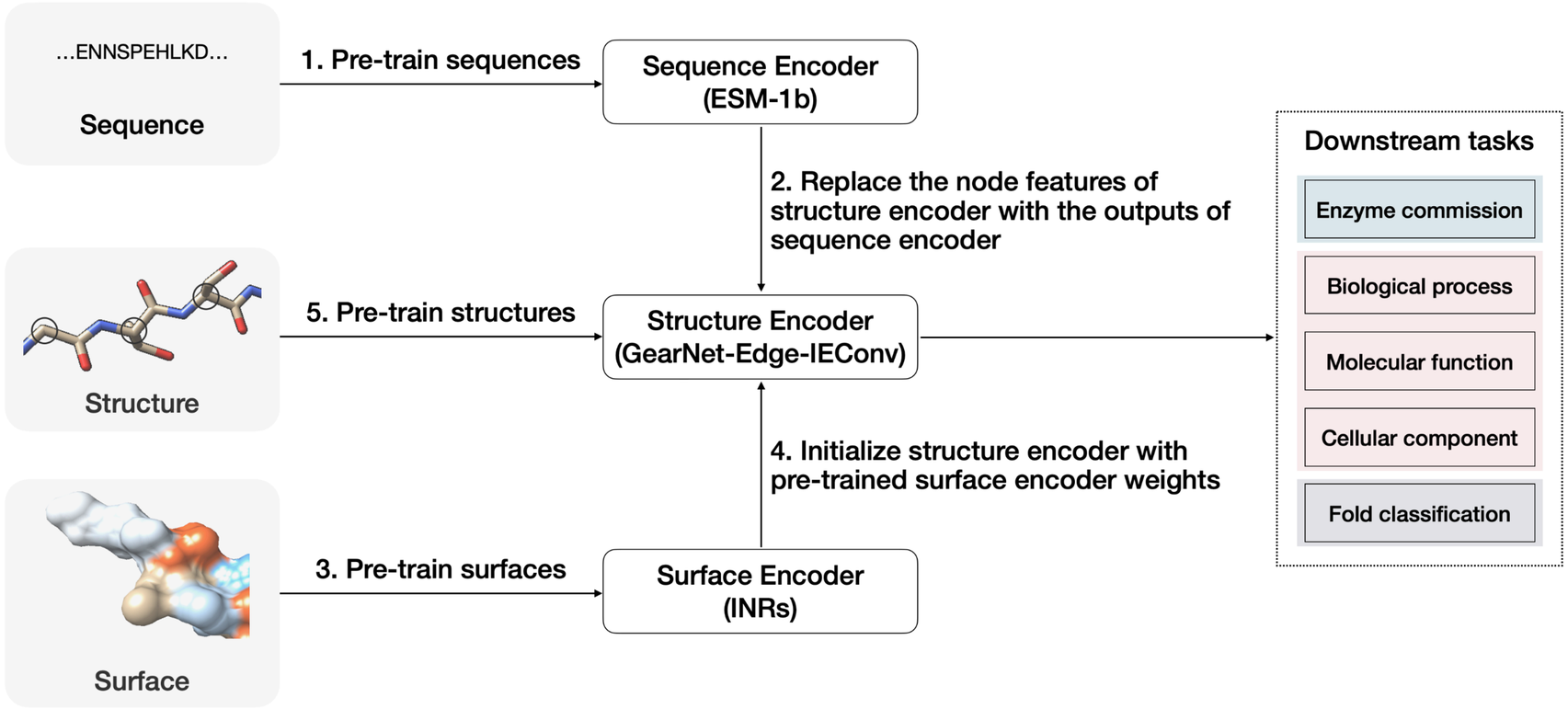
Objective

We propose a pre-training strategy that incorporates all three essential aspects of proteins: sequences, 3D structures, and surfaces

Method	Sequence Encoder	Structure Encoder	Sequence Pre-training	Structure Pre-training	Surface Encoder	Surface Pre-training
CNN	✓					
Transformer	✓					
GVP		✓				
GearNet		✓				
ESM-1b	✓		✓			
ProtBert	✓		✓			
DeepFRI	✓	✓	✓			
LM-GVP	✓	✓	✓			
ESM-GearNet	✓	✓	✓			
GearNet-MC		✓		✓		
GearNet-DP		✓		✓		
ESM-GearNet-MC	✓	✓	✓	✓		
ESM-GearNet-INR-MC (Ours)	✓	✓	✓	✓	✓	✓

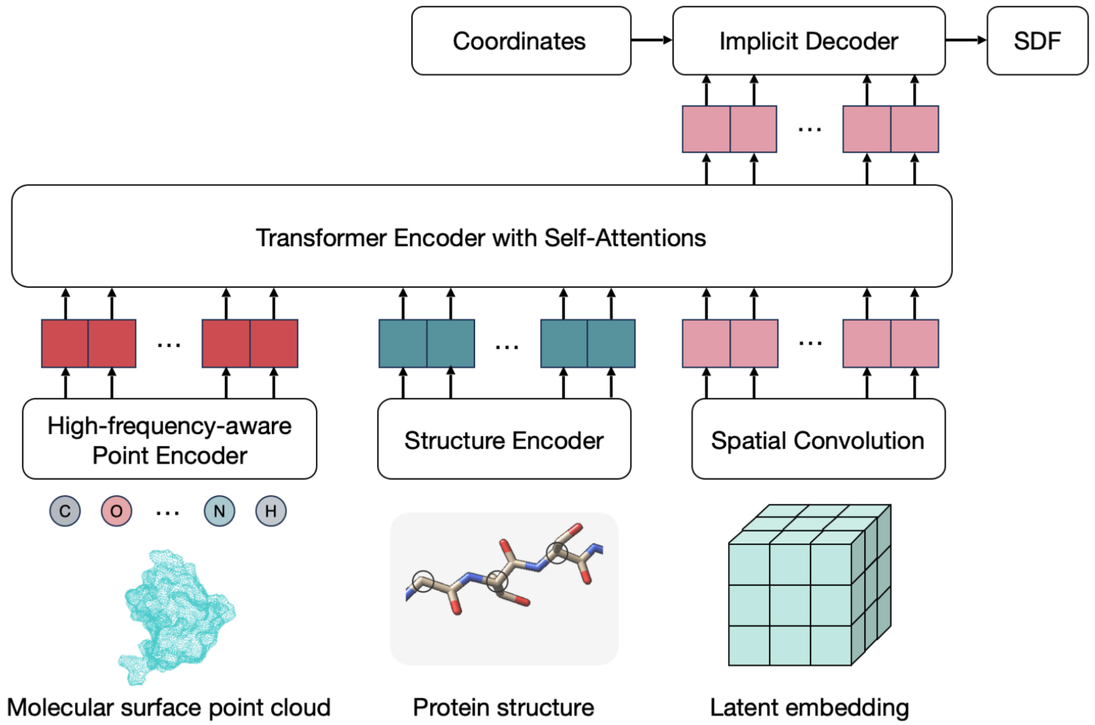
Overall strategy

Our strategy for pre-training sequences, structures, and surfaces to solve downstream tasks



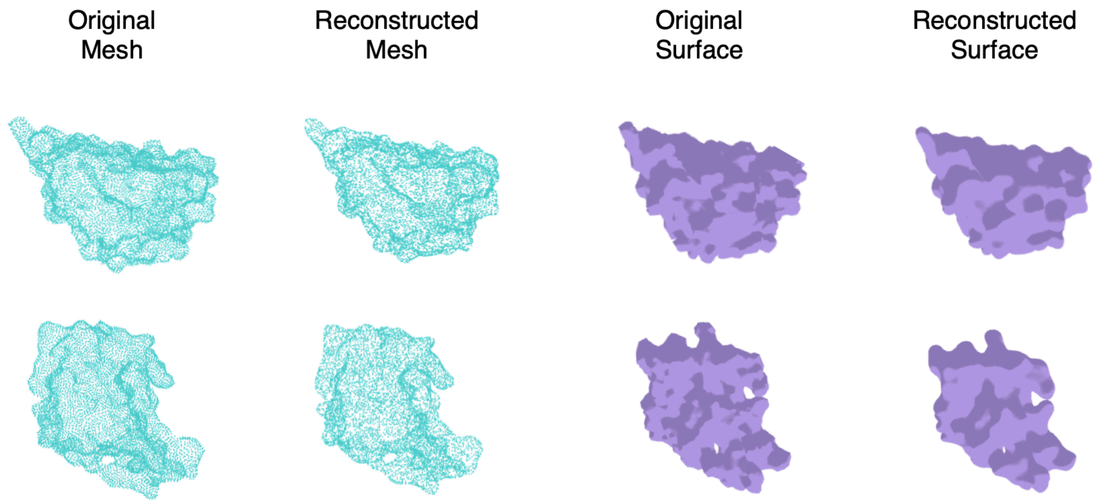
ProteinINR

An overview of our ProteinINR architecture



ProteinINR

Reconstructed meshes and surfaces from ProteinINR for given proteins



Results

Our results show improved performance on downstream tasks with ESM-GearNet-INR and ESM-GearNet-INR-MC

Method	EC		GO-BP		GO-MF		GO-CC		FC	Sum
	F_{max}	AUPR	F_{max}	AUPR	F_{max}	AUPR	F_{max}	AUPR	Acc	
ESM-1b†	86.9	88.4	45.2	33.2	65.9	63.0	47.7	32.4	-	-
ESM-2†	87.4	88.8	47.2	**34.0**	66.2	**64.3**	47.2	35.0	-	-
GearNet	81.6	83.7	44.8	25.2	60.4	52.9	43.3	26.8	46.8	465.5
GearNet-INR	81.4	83.7	44.7	26.5	59.9	52.1	43.0	27.2	47.6	466.1
GearNet-MC	87.2	88.9	49.9	26.4	64.6	55.8	46.9	27.1	51.5	498.3
GearNet-INR-MC	86.9	88.9	49.8	26.0	65.4	56.1	47.7	26.6	51.1	498.5
ESM-GearNet-MC	89.0	89.7	**53.5**	27.5	**68.7**	57.9	49.4	32.4	**53.8**	521.9
ESM-GearNet-INR	89.0	90.3	50.8	33.4	67.8	62.6	**50.6**	**36.9**	48.9	**530.3**
ESM-GearNet-INR-MC	**89.6**	**90.3**	51.8	33.2	68.3	58.0	50.4	35.7	50.8	528.1

Take home message

We propose a pre-training strategy that includes the surfaces of proteins using INR, which can lead to better protein representation

Limitation

- **Likely low performance for proteins without known structures**

Future work

- **Generating new proteins from latent representations of surfaces**
- **Extending this approach to other types of molecules**