SLAE: Strictly Local All-atom Environment for Protein Representation

Yilin Chen[1*], Tianyu Lu[1], Cizhang Zhao[2], Hannah K. Wayment-Steele[2], Po-Ssu Huang[1]

[1] Stanford University, [2] University of Wisconsin–Madison

ICML International Conference On Machine Learning

Code Available
Question & Collaboration:
yilinc5@stanford.edu

Stanford University

WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

Motivation

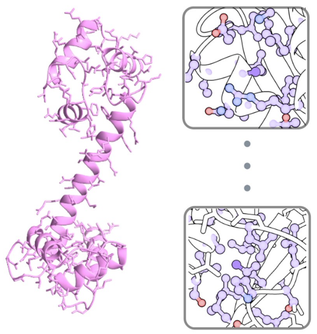
Proteins are large 3D molecular systems whose functions depend on how their atoms are arranged and interact.

Computational Protein Design

Protein Property Prediction

Current protein representations often rely on sequence-level evolutionary signals or backbone geometry, making them efficient but less sensitive to local chemistry.

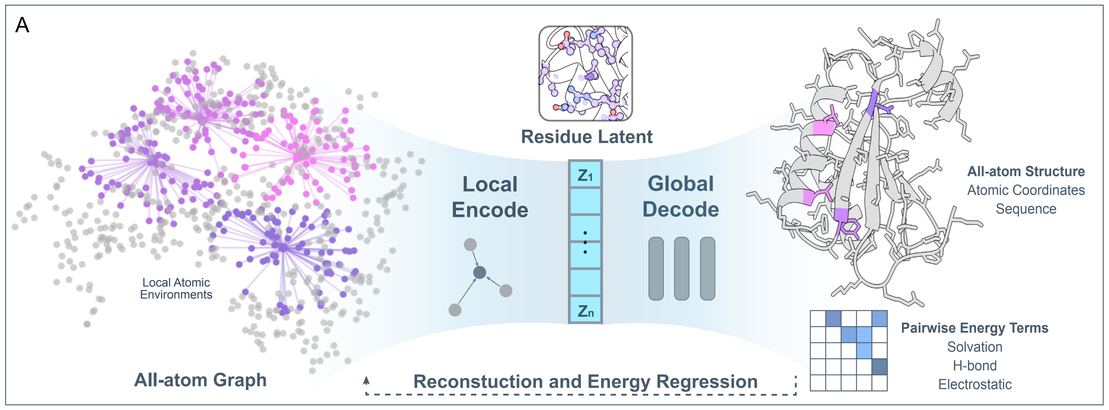
How can we preserve *atom-level physical information* while keeping protein representations compact and learnable?

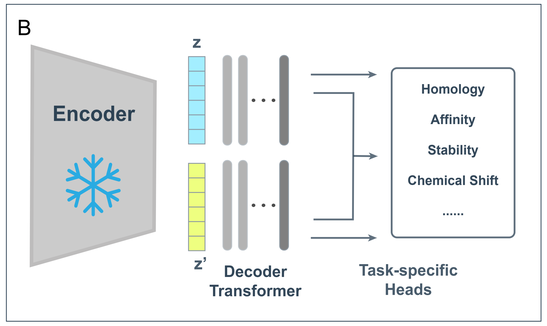


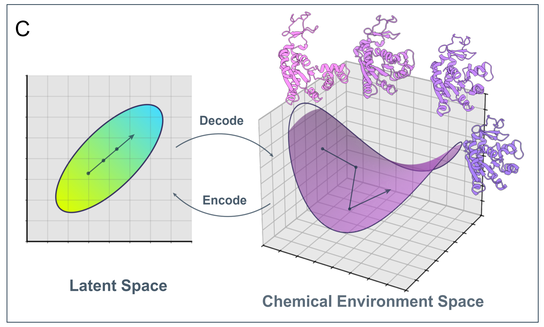
The "**semantics**" of a protein is determined by each residue's local chemical environment:

which atoms are nearby, and
how they are arranged in 3D

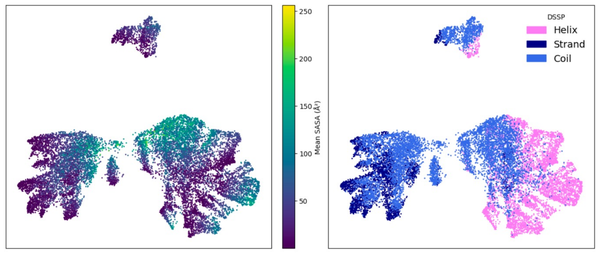
Framework Overview



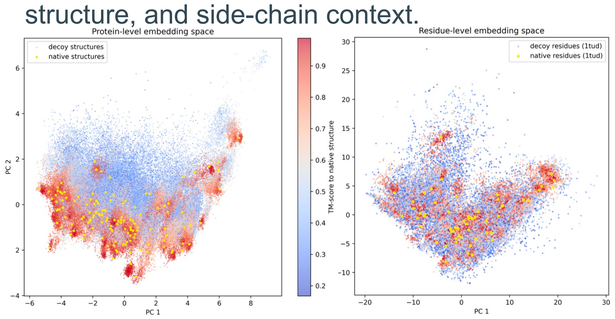




Chemical Environment Change

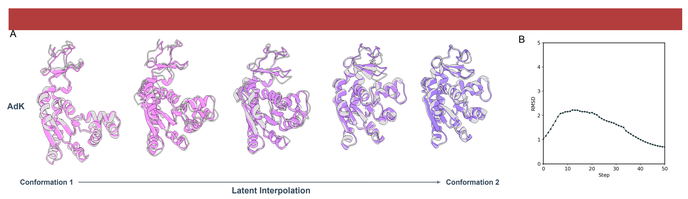


SLAE latent space captures **residue environment changes** across solvent exposure, secondary structure, and side-chain context.

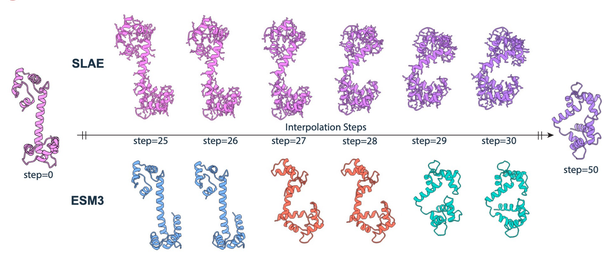


Distinguish native-like from decoy residue environments.

Conformational Transition



Linear interpolation between residues decodes to valid intermediate structures **without explicit smoothness regularization**.



SLAE produces smoother, physically consistent interpolations compare to backbone-only models, while modeling side chains.

Contribution

- **Local all-atom representation**: encode each residue by its nearby atoms and 3D geometry.
- **Physics-informed pretraining**: use energy prediction to guide structure reconstruction.
- **Local-to-global architecture**: SE(3)-equivariant encode local environments, decode all-atom protein structure.
- **Transferable embeddings**: achieve strong performance across diverse protein tasks.
- **Interpretable latent space**: captures chemistry, structure, and environment changes.

Transfer Learning across Tasks

Task Type	Autoencoder Structure Tokenizer			SLAE			Ablated SLAE			
	FoldSeek	ESM3	AminoAseed	Continuous	Quantized	Hidden Rep	Radius=4Å	Radius=10Å	BB_Seq	BB_SC
Functional Site Prediction (Average AUROC%)↑	51.90	69.24	72.43	75.20 (+3.83%)	72.45 (+0.03%)	**79.54** (+9.80%)	70.04 (–3.30%)	69.20 (–4.46%)	69.92 (–3.46%)	67.84 (–6.34%)
Structural Flexibility Prediction (Average Spearman's ρ)↑	7.80	37.35	38.08	48.19 (+26.59%)	43.81 (+15.06%)	**57.05** (+49.90%)	36.95 (–2.96%)	40.37 (+6.03%)	46.21 (+21.28%)	46.68 (+22.58%)

SLAE **transfers across tasks at multiple scales**, from individual atoms to protein interfaces and whole structures, matching or outperforming specialized methods.

Homology: Entire Structure

Affinity: Interface

Stability: Mutated Residue

Chemical Shift: Nitrogen Atom