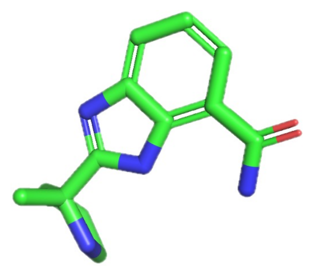
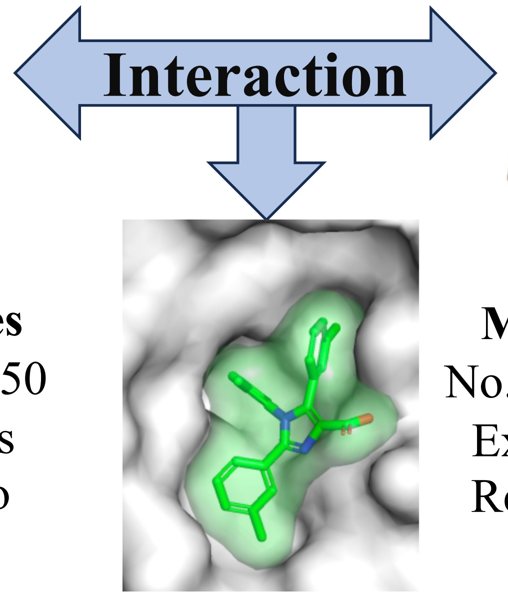
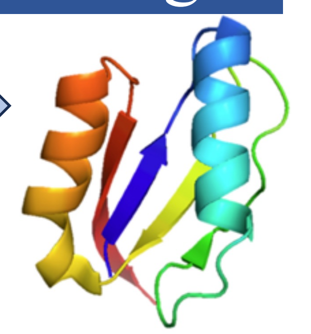
ESM All-Atom: Multi-scale Protein Language Model for Unified Molecular Modeling

Kangjie Zheng*, Siyu Long*, Tianyu Lu, Junwei Yang, Xinyu Dai, Ming Zhang, Zaiqing Nie, Wei-Ying Ma, Hao Zhou

What is Multi-scale Modeling?







Small Molecules
No. of Atoms < 150
Example: SMDs
Atom scale Info

Macromolecules
No. of Atoms > 1000
Example: Proteins
Residue scale Info

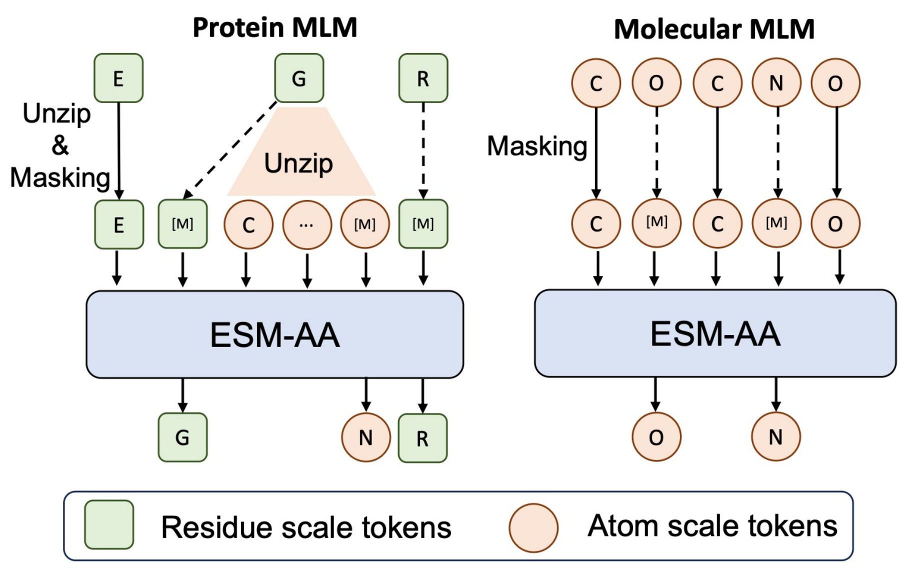
Multi-scale Modeling to Capture Multi-scale Information.

ESM-AA achieves this by pre-training on **multi-scale code-switch protein sequences** and utilizing a **multi-scale position encoding** to capture the positional relationships among residues and atoms.

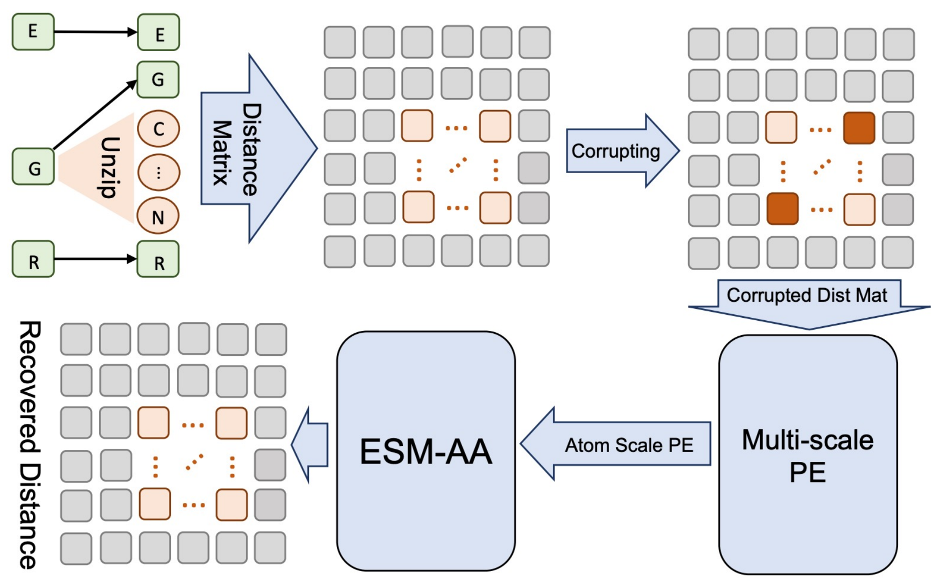
Codes: https://github.com/zhengkangjie/ESM-AA

Multi-scale Pre-training

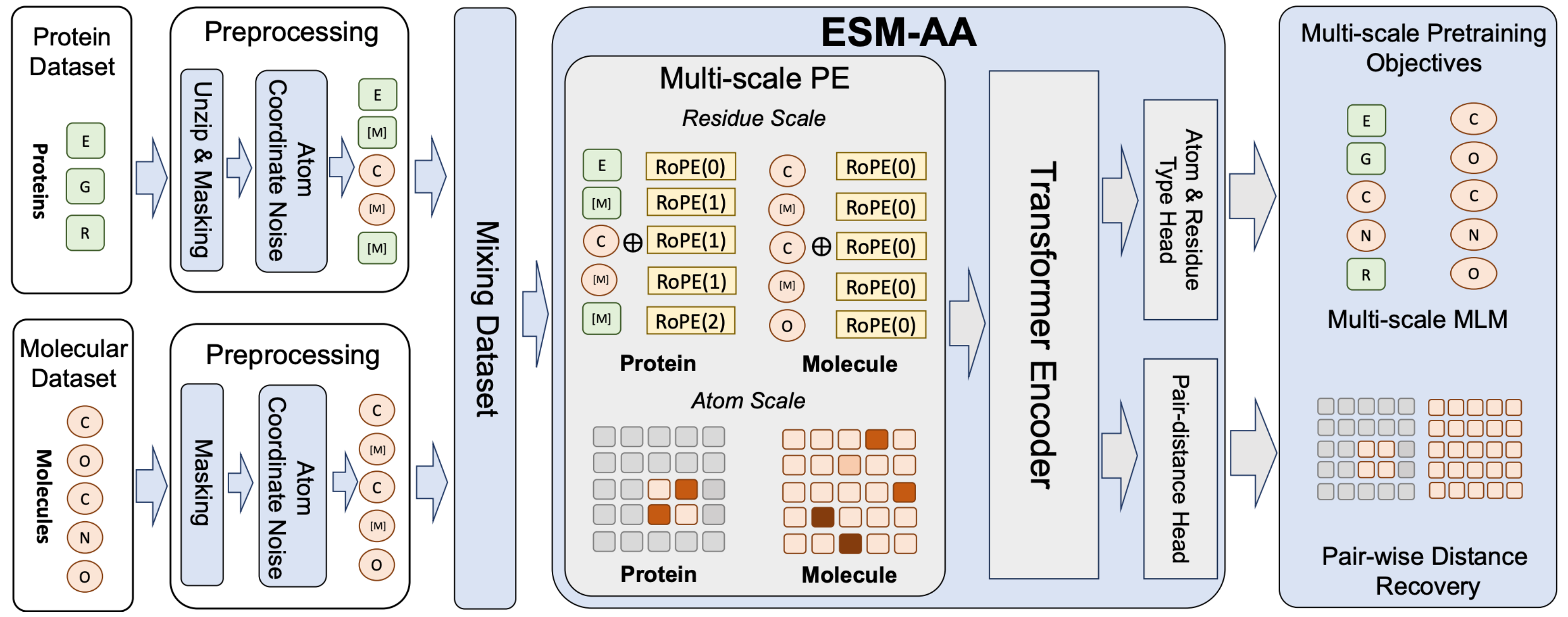
- **Inspired by the Multilingual Pre-training: Code-Switch Protein Sequence**
  - To construct a code-switch protein sequence , we randomly select a group of residues and insert their corresponding atoms into the sequence, which is the unzipping process.
- **Pre-training Objective for Multi-scale Information : Multi-scale Masked Language Modeling**
  - Randomly masking a portion of the atoms or residues in and then ask the model to predict the original atoms or residues using the surrounding context.
- **Pre-training Objective for Atom-scale Information: Pair-wise Distance Recovery**
  - We use the corrupted atoms coordinates as model input and ask model to recover the accurate Euclidean distances between these atoms. We only calculate PDR within residues.



(a) Multi-scale MLM

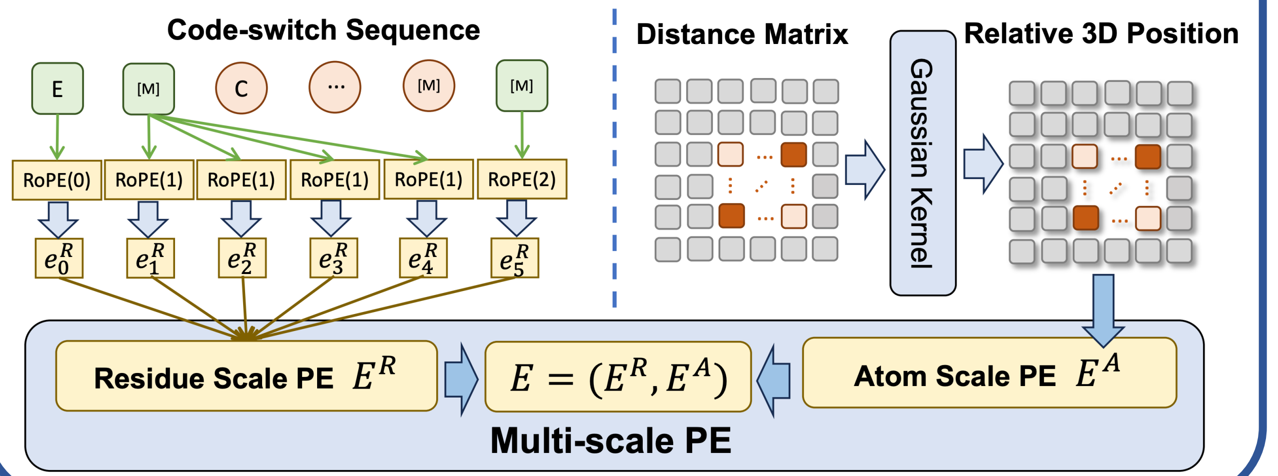


(b) Pair-wise Distance Recovery



Multi-scale Position Encoding

- **Residue Scale Position Encoding: RoPE**
  - For encoding the relationship between two residues, the PE should be consistent with the mainstream encoding method.
  - For atoms from the same unzipped residue, the PE should not introduce any ambiguous position information.
- **Atom Scale Position Encoding: 3D Spatial PE**
  - Atom-scale structural information is crucial for modeling atomic level semantics.
  - The model needs to have the ability to capture structural information at the atomic scale



Experimental Results

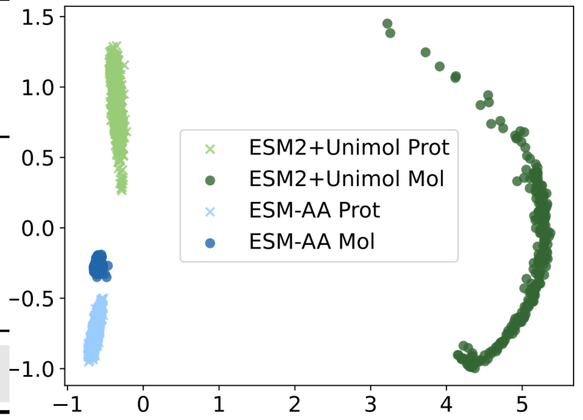
- **Pre-training Datasets: AlphaFold DB and Uni-Mol Molecular Dataset**
  - For the protein dataset, we use AlphaFold DB dataset, which contains 8M protein sequences and structures predicted by AlphaFold2 with high confidence (pLDDT > 90).
  - For the molecule dataset, we use the dataset provided by Uni-Mol, which contains 19M molecules and 209M conformations generated by ETKGD and Merck Molecular Force Field.
- **Performance on Protein-Molecule Tasks: Unified Modeling Can Harness the Full Potential of Pre-training Techniques**
  - ESM-AA outperforms other models and achieves the state-of-the-art results, which indicates that our unified modeling can harness the full potential of PLMs.
- **Performance on Protein-Only Tasks : ESM-AA Preserves the Strong Ability of Protein Understanding**
  - The table demonstrates that ESM-AA can perform similarly to ESM-2 in unsupervised contact prediction task. This indicates that ESM-AA does not sacrifice its understanding of proteins.
- **Visualization of Proteins' and Molecules' Embedding : ESM-AA Preserves the Strong Ability of Protein Understanding**
  - ESM-AA model is capable of creating a more cohesive semantic representation encompassing both proteins and molecular data, which makes ESM-AA outperform two separate models.

Method	Protein Pre-training	Molecule Pre-training	ESAR			ESPC		
			MSE ↓	R^2 ↑	Pearson ↑	ACC ↑	MCC ↑	ROC-AUC ↑
Gollub et al. (2023)	/	/	/	0.463	0.680	/	/	/
Kroll et al. (2021)	/	/	0.653	0.527	0.728	/	/	/
Baseline $_{\text{XGBoost}}$	ESM-2 35M	Uni-Mol 48M	0.652	0.528	0.727	89.9%	0.729	0.941
Baseline $_{\text{ProSmith}}$	ESM-2 35M	Uni-Mol 48M	0.642	0.536	0.733	90.8%	0.754	0.943
Ours $_{\text{XGBoost}}$	ESM-AA 35M	ESM-AA 35M	0.620	0.551	0.744	90.4%	0.743	0.949
Ours $_{\text{ProSmith}}$	ESM-AA 35M	ESM-AA 35M	**0.607**	**0.560**	**0.752**	**92.3%**	**0.797**	**0.957**

ESAR: Enzyme-Substrate Affinity Regression, ESPC: Enzyme-Substrate Pair Classification

Method	Short Range ↑			Medium Range ↑		
	P@L	P@L/2	P@L/5	P@L	P@L/2	P@L/5
TAPE 92M	0.10	0.12	0.16	0.10	0.13	0.17
ESM-1 43M	0.11	0.13	0.16	0.12	0.15	0.19
ESM-2 35M	0.20	0.29	0.46	0.22	**0.32**	**0.45**
ESM-AA 35M	**0.21**	**0.31**	**0.48**	**0.23**	**0.32**	**0.45**

Unsupervised Contact Prediction



Visualization on DTAR Dataset