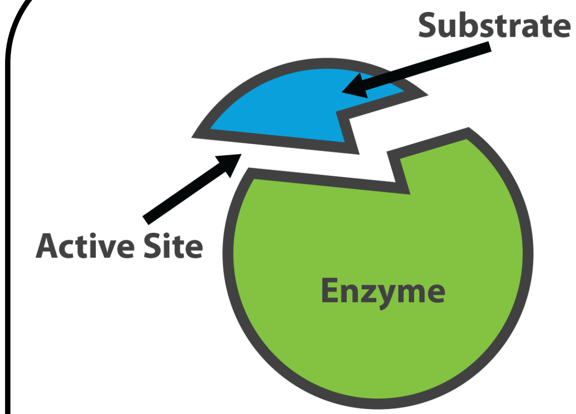
Pre-Training Protein Encoder via Siamese Sequence-Structure Diffusion Trajectory Prediction

Zuobai Zhang*, Minghao Xu*, Aurélie Lozano,
Vijil Chenthamarakshan, Payel Das, Jian Tang

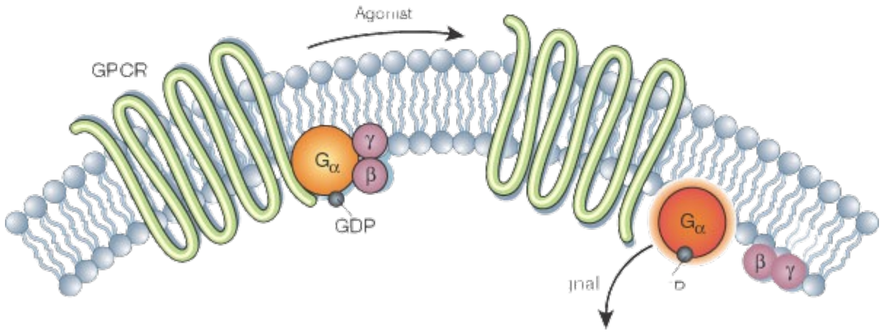
Mila IBM

Protein

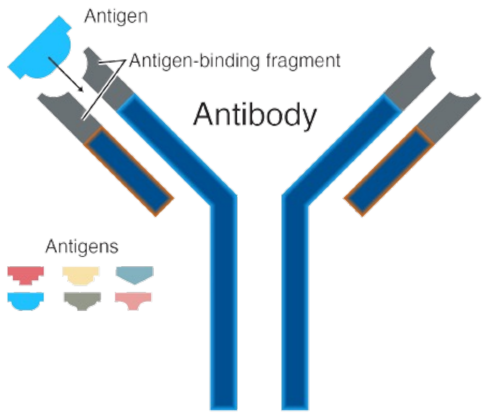
- Fundamental components in our life
  - Involved in many biological processes



Catalysis



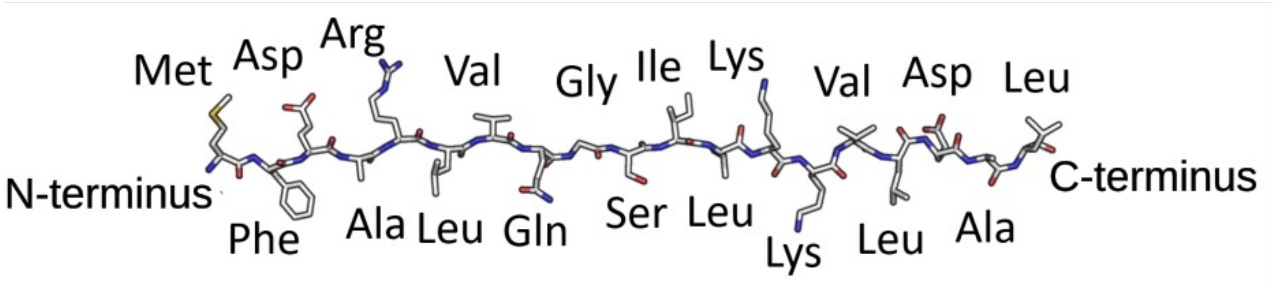
Signaling

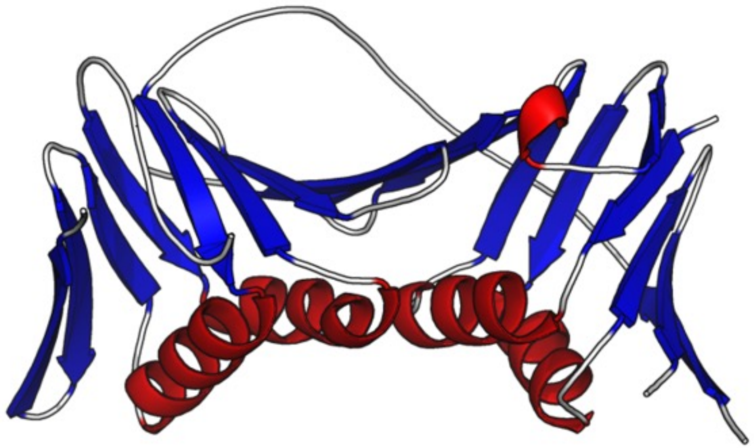


Defense

Protein Sequence and Structure

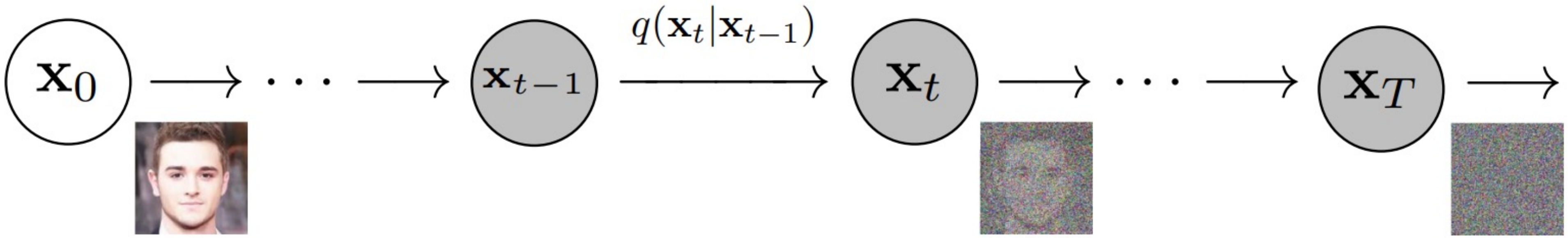
- Protein sequence consists of amino acids, *a.k.a.*, residues
- Protein sequences determines structures





Joint Pre-Training

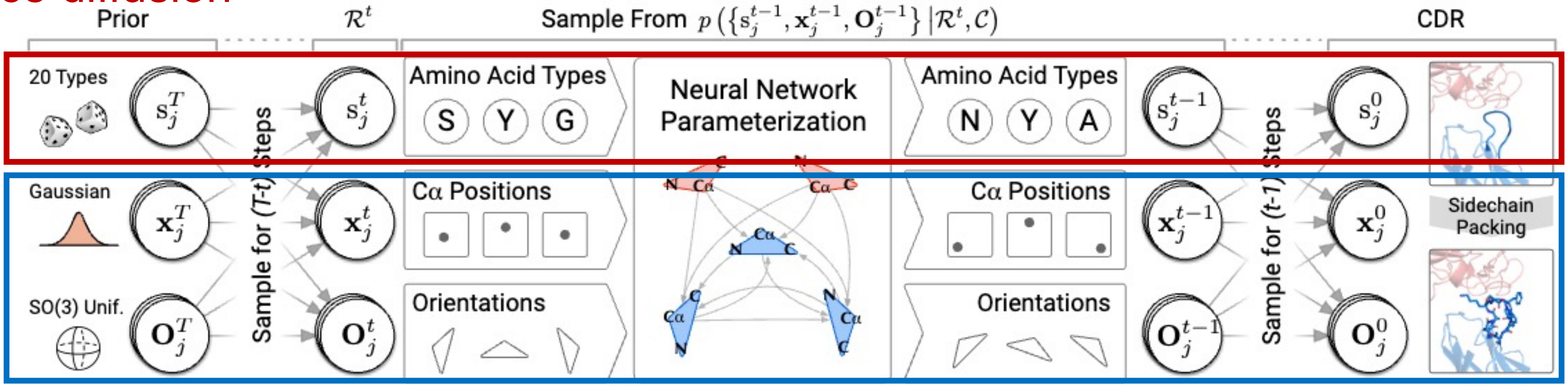
- Existing works
  - Pre-training objectives on either sequences or structures
- How to use both modalities for pre-training?
  - Diffusion models!



Diffusion Models on Proteins

- Diffusion models capture joint distribution of sequences and structures.
  - Diffusion models are equivalent to multi-level denoising.

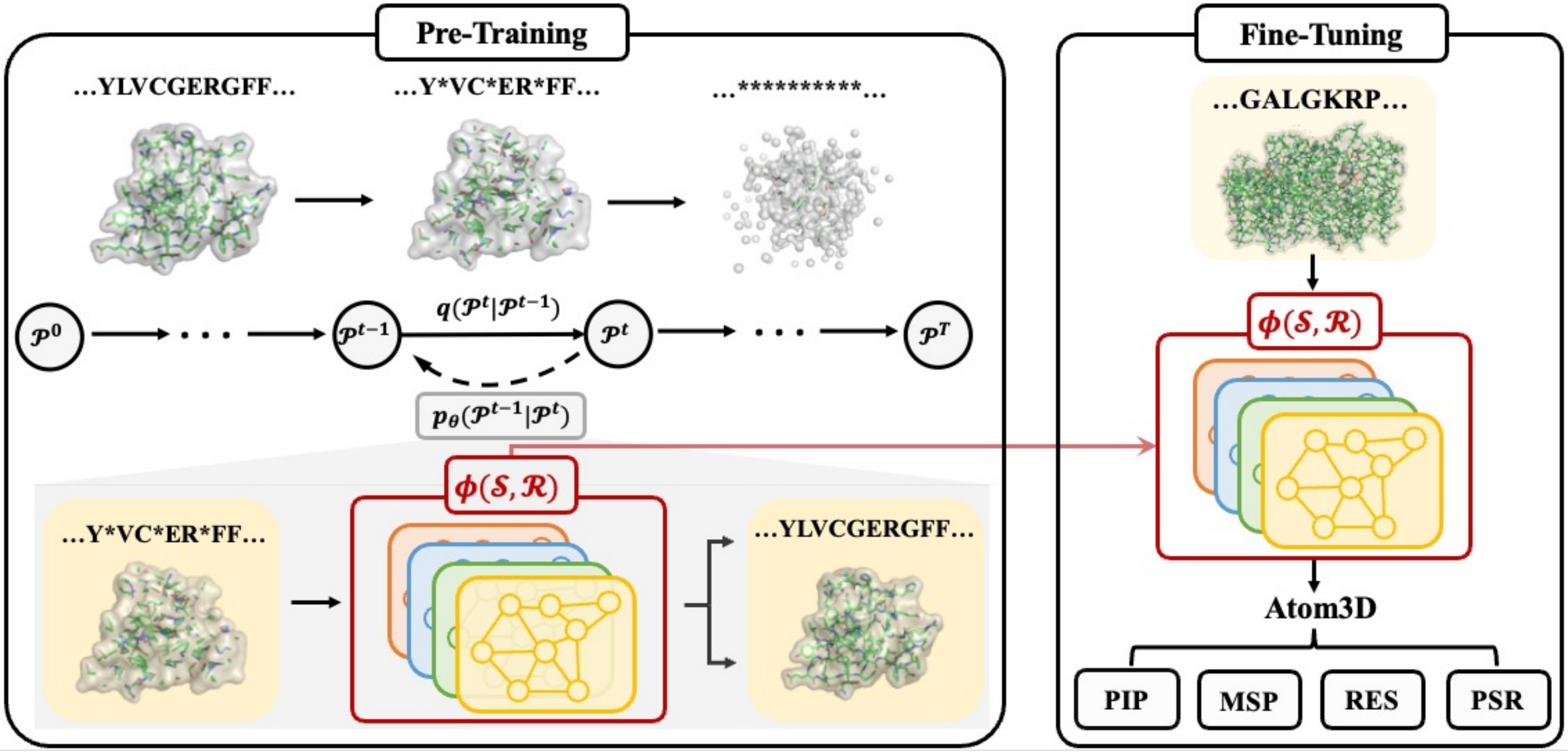
Sequence diffusion



Structure diffusion

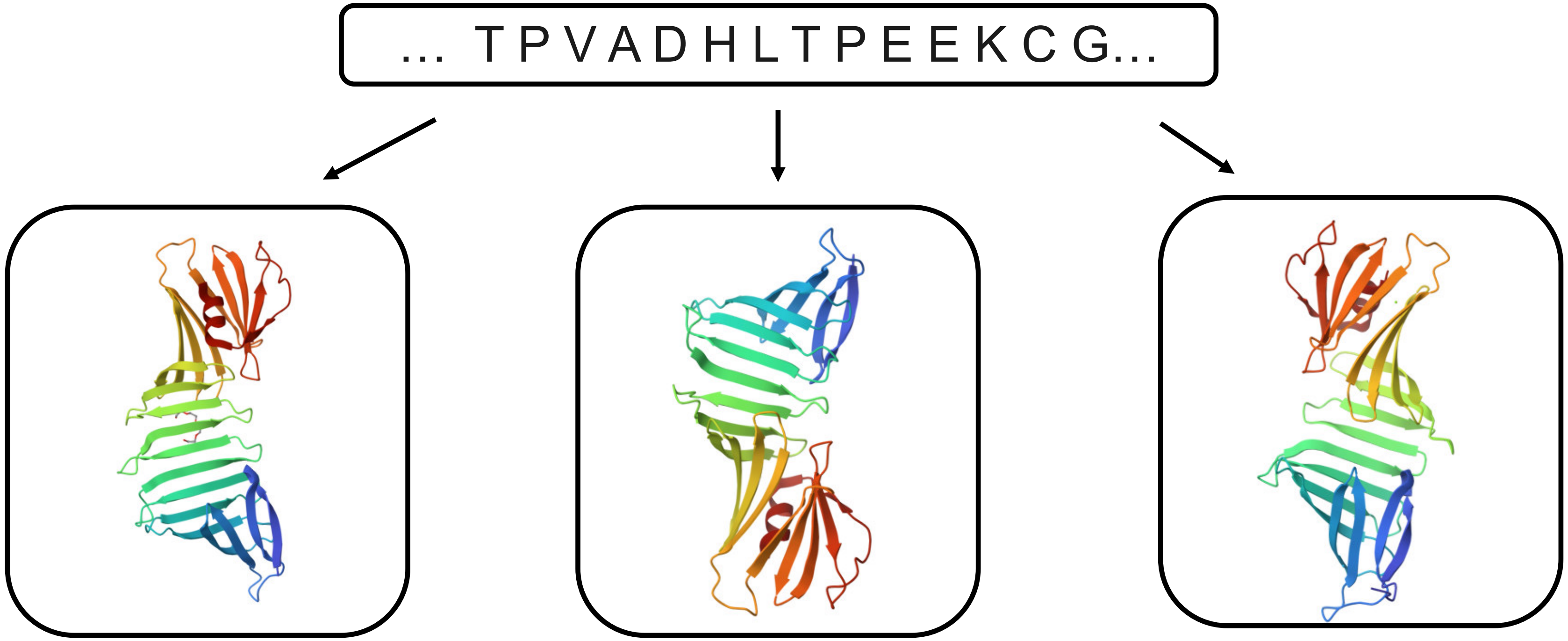
Image source: [Luo et al. 2022]()

Diffusion Models for Pre-Training (DiffPreT)



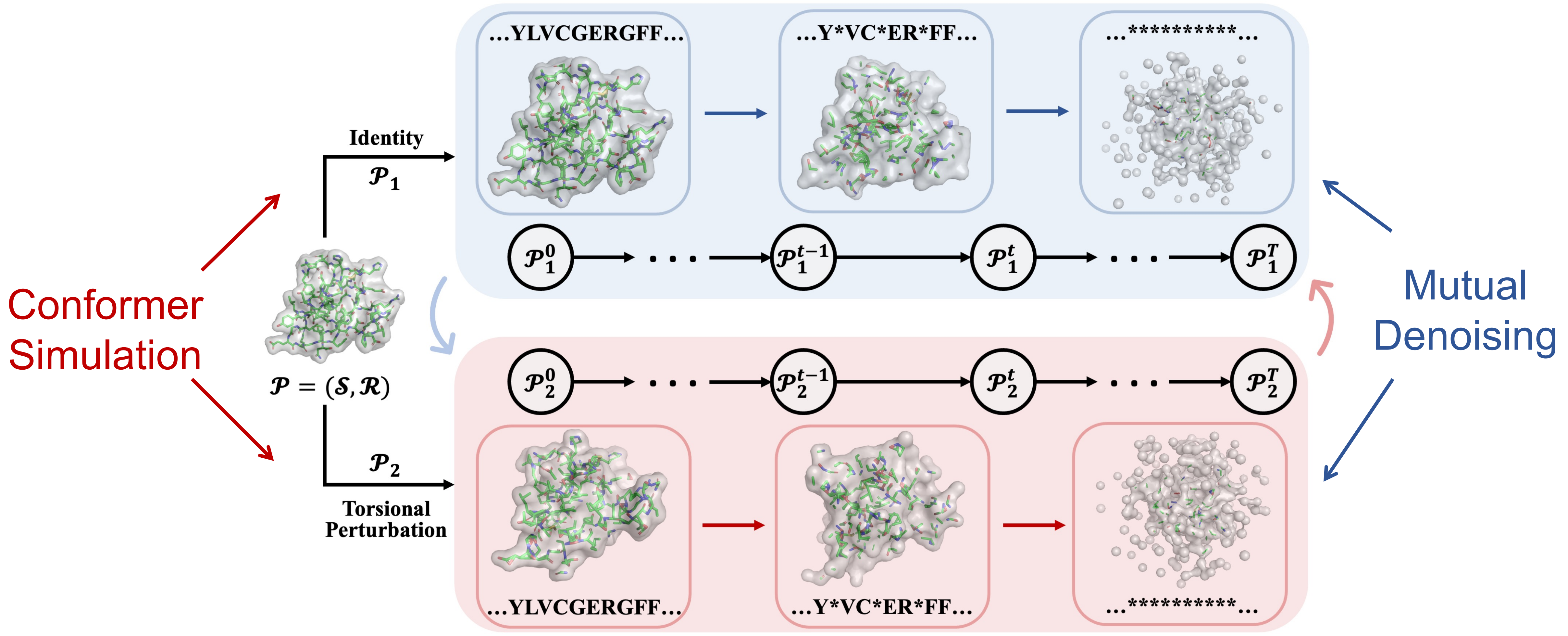
Protein Conformer

- Sequence -> Multiple structures, *i.e.*, conformers



How to capture conformer information during pre-training?

Siamese Diffusion Trajectory Prediction (SiamDiff)



Mutual information maximization between conformers

Multi-Level Denoising

- Multiple noise levels

$$\mathcal{L} := \mathbb{E}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}\left(q(\mathcal{P}^{t-1}|\mathcal{P}^{t},\mathcal{P}^{0})||p_{\theta}(\mathcal{P}^{t-1}|\mathcal{P}^{t})\right)\right]$$

  - Better than treating noise level as a hyperparameter[1]
  - Large noise – coarse-grained - easy
  - Small noise – fine-grained - difficult

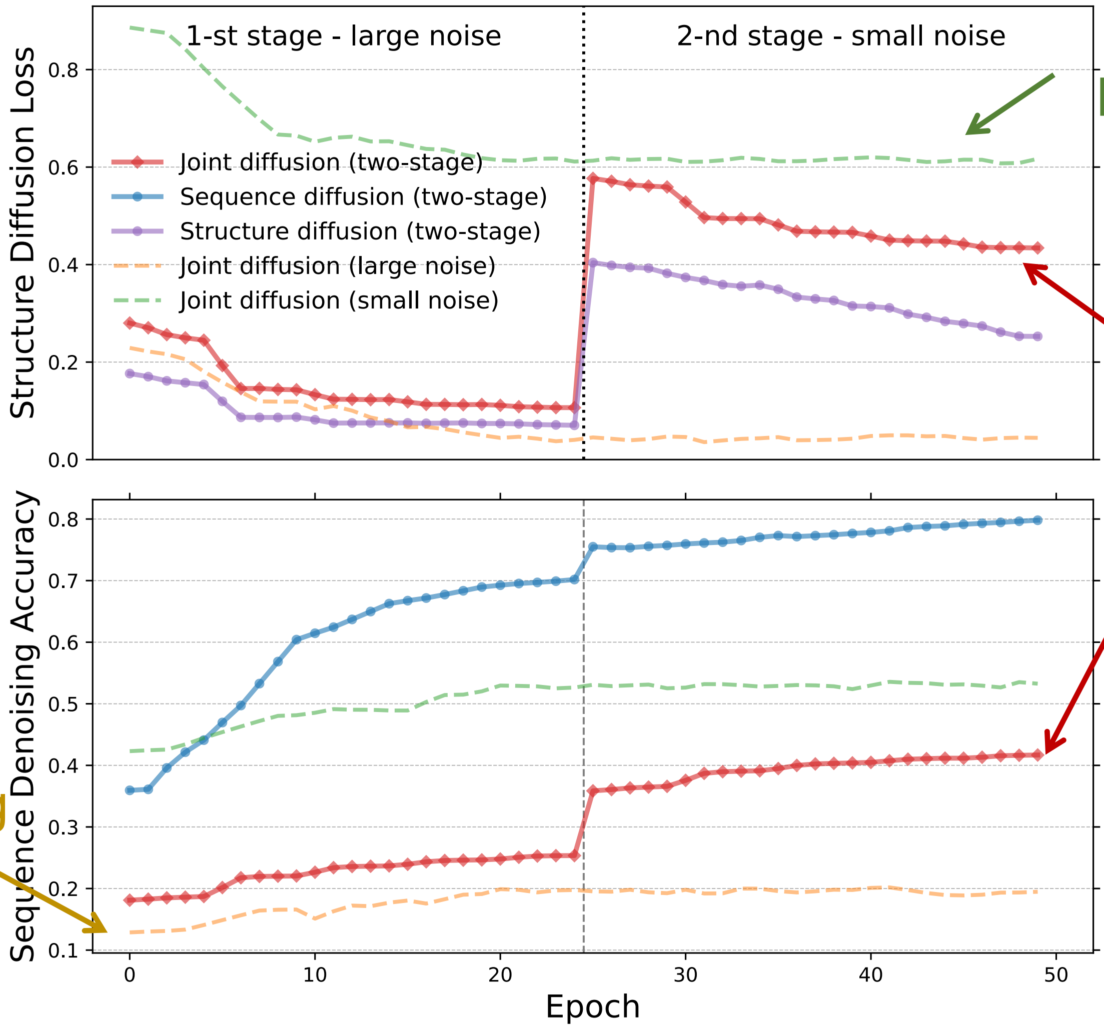
- ***However, this is very different for joint diffusion!***

[1] Zaidi, et al. "Pre-training via Denoising for Molecular Property Prediction." *ICLR 2023*.

Two-Stage Noise Scheduling

Structure perturbation makes it harder to do sequence denoising!!!

Small noise, large loss
Difficult for structure denoising



Our solution:
Two-stage noise scheduling

Large noise, small acc.
Difficult for sequence denoising

Results

Table 1: Atom-level results on Atom3D tasks.

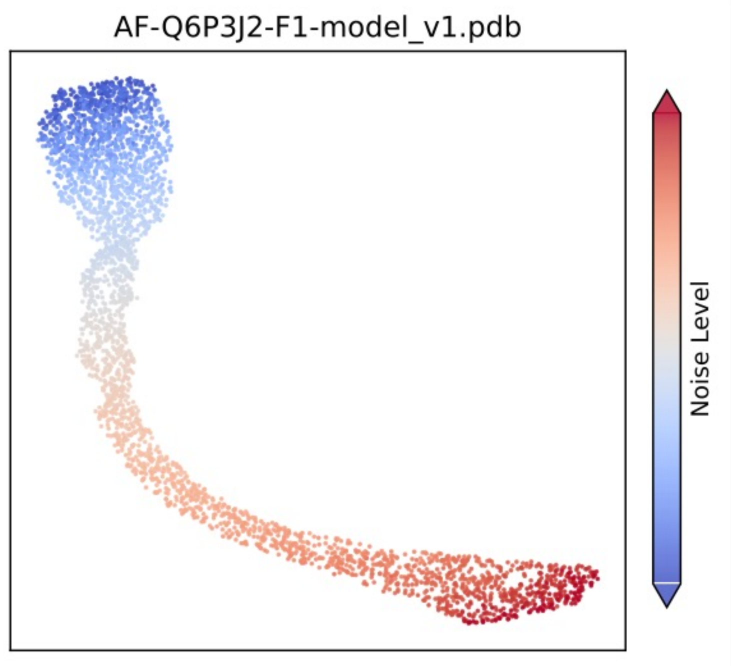
	Method	PIP	MSP	RES	PSR		Mean Rank
		AUROC	AUROC	Accuracy	Global ρ	Mean ρ	
	GearNet-Edge	0.868±0.002	0.633±0.067	0.441±0.001	0.782±0.021	0.488 ±0.012	7.6
w/ pre-training	Denoising Score Matching	0.877±0.002	0.629±0.040	0.448±0.001	0.813±0.003	0.518±0.020	5.2
	Residue Type Prediction	0.879±0.004	0.620±0.027	0.449±0.001	0.826±0.020	0.518±0.018	4.4
	Distance Prediction	0.872±0.001	0.677±0.020	0.422±0.001	**0.840±0.020**	0.522±0.004	4.0
	Angle Prediction	0.878±0.001	0.642±0.013	0.419±0.001	0.813±0.007	0.503±0.012	6.2
	Dihedral Prediction	0.878±0.004	0.591±0.008	0.414±0.001	0.821±0.002	0.497±0.004	6.8
	Multiview Contrast	0.871±0.003	0.646±0.006	0.368±0.001	0.805±0.005	0.502±0.009	7.2
	DiffPreT	0.880±0.005	0.680±0.018	0.452±0.001	0.821±0.007	0.533±0.006	2.4
	SiamDiff	**0.884±0.003**	**0.698±0.020**	**0.460±0.001**	0.829±0.012	**0.546±0.018**	**1.2**

Table 2: Residue-level results on EC and Atom3D tasks.

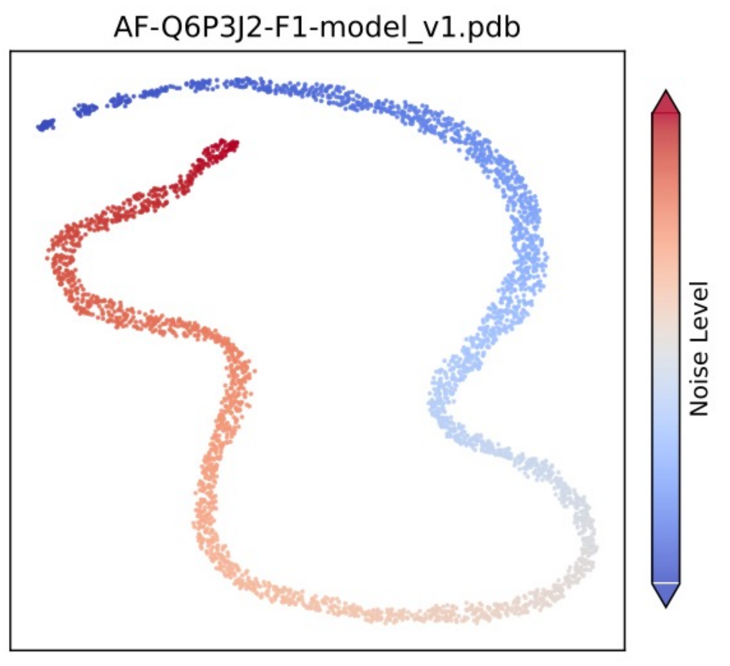
	Method	EC		MSP	PSR		Mean Rank
		AUPR	F_{max}	AUROC	Global ρ	Mean ρ	
	GearNet-Edge	0.837±0.002	0.811±0.001	0.644±0.023	0.763±0.012	0.373±0.021	7.8
w/ pre-training	Denoising Score Matching	0.859±0.003	0.840±0.001	0.645±0.028	0.795±0.027	0.429±0.017	5.0
	Residue Type Prediction	0.851±0.002	0.826±0.005	0.636±0.003	0.828±0.005	0.480±0.031	5.4
	Distance Prediction	0.858±0.003	0.836±0.001	0.623±0.007	0.796±0.017	0.416±0.021	6.4
	Angle Prediction	0.873±0.003	0.849±0.001	0.631±0.041	0.802±0.015	0.446±0.009	4.2
	Dihedral Prediction	0.858±0.001	0.840±0.001	0.568±0.022	0.732±0.021	0.398±0.022	7.2
	Multiview Contrast	0.875±0.003	**0.857±0.003**	**0.713±0.036**	0.752±0.012	0.388±0.015	4.0
	DiffPreT	0.864±0.002	0.844±0.001	0.673±0.042	0.815±0.008	0.505±0.007	3.2
	SiamDiff	**0.878±0.003**	**0.857±0.003**	0.700±0.043	**0.856±0.007**	**0.521±0.016**	**1.2**

Good results on all considered tasks

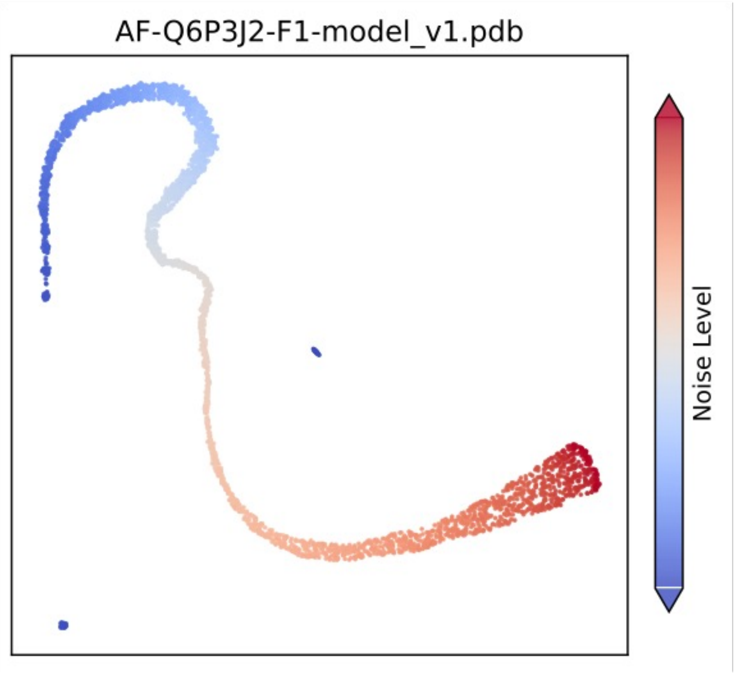
Visualization Results



Random Initialization



First-stage SiamDiff



Second-stage SiamDiff

Thanks!