Diffusion on Language Model Encodings for Protein Sequence Generation

V. Meshchaninov*[1] · P. Strashnov*[2] · A. Shevtsov*[2] · F. Nikolaev[2] · N. Ivanisenko[2] · O. Kardymon[2] · D. Vetrov[1]
[1]Constructor University, Bremen, Germany · [2]AIRI, Moscow, Russia. *Core contributor.

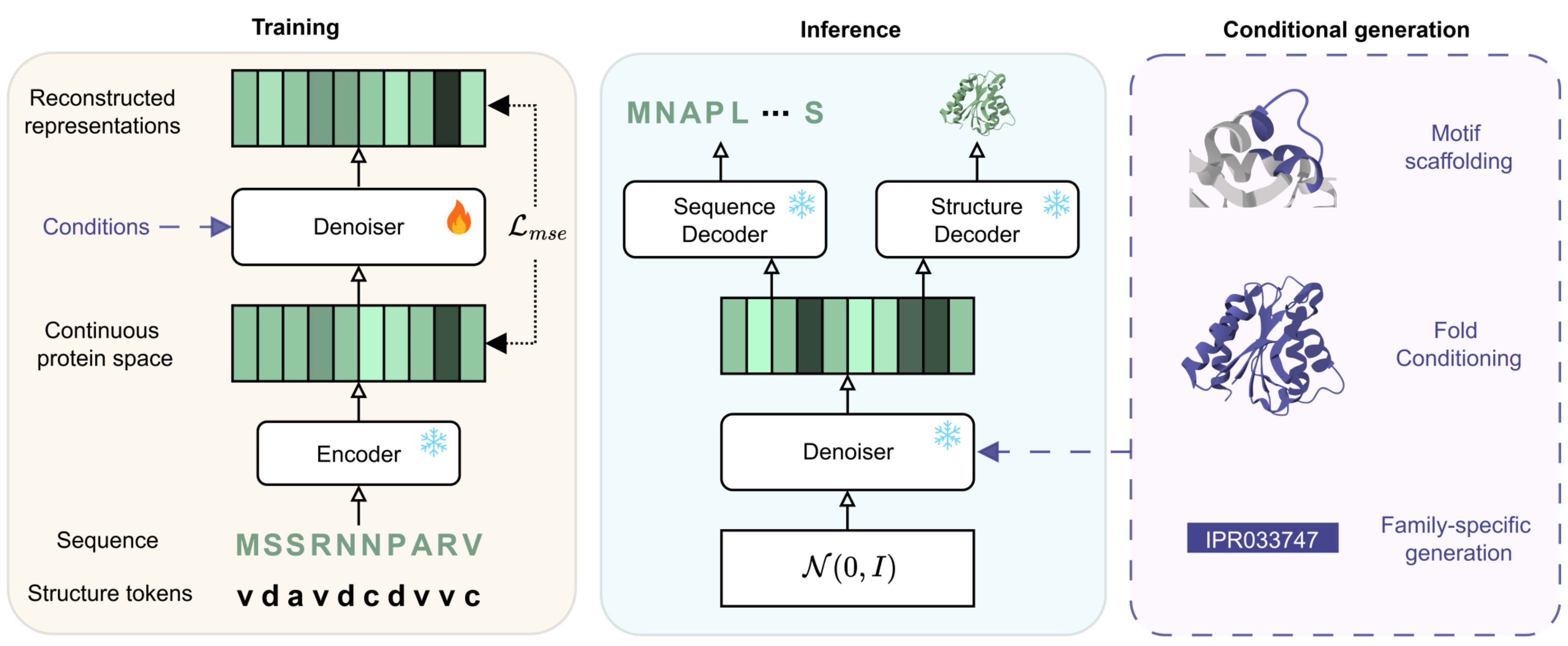
Background & Motivation

- Protein language models encode sequence and structure
- Continuous diffusion works well for images but not explored for proteins
- Discrete methods produce repetitive, low-diversity sequences.
- Goal: develop a continuous diffusion framework that works across multiple protein language models to enable efficient protein generation and design tasks.

Main contributions

- First encoder-agnostic continuous diffusion framework for proteins
- Single 35M architecture generalizes across diverse pLMs (8M-3B parameters)
- Achieves SOTA quality & diversity in unconditional generation
- Conditional fine-tuning enables family-specific design with high fidelity
- Structure-aware encoders (SaProt) solve motif-scaffolding and sequence infilling
- Pretraining with CHEAP enables fold-conditioned generation

DiMA: Diffusion Model for Aminoacids

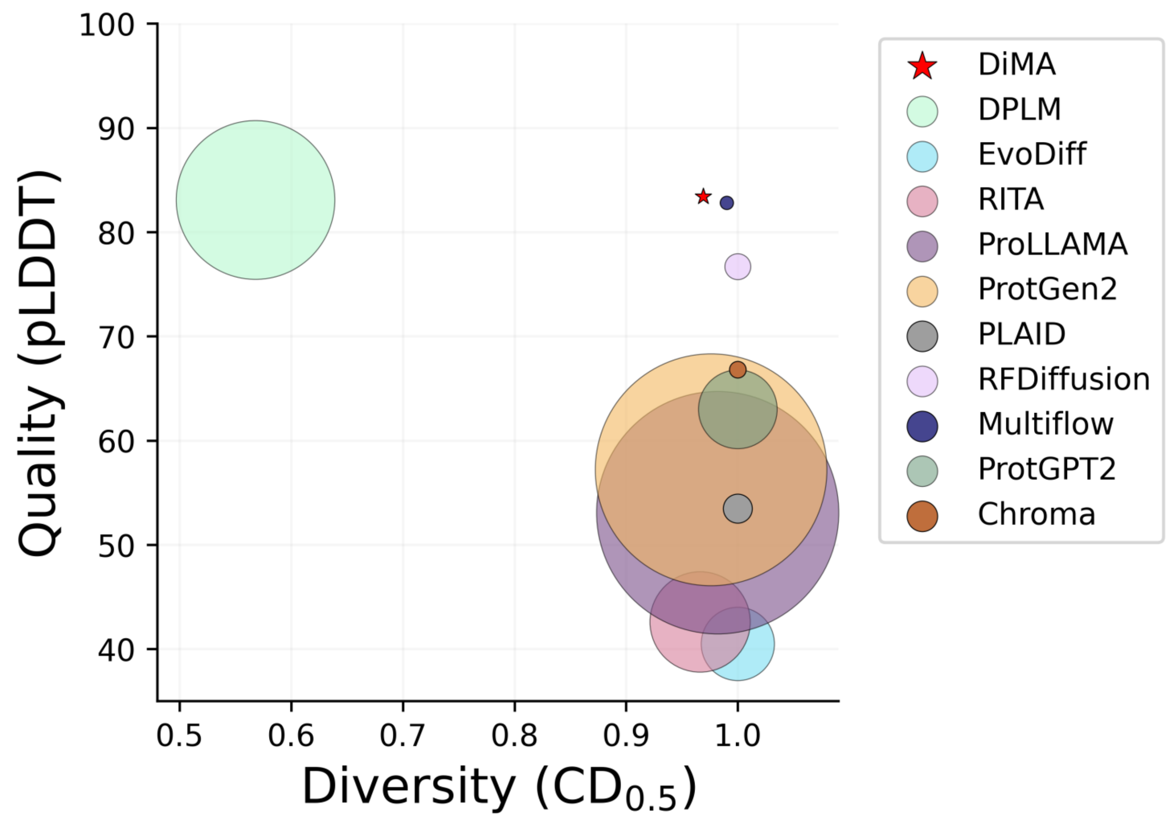


The framework consists of three main components:
1. a pre-trained protein language model encoder that maps amino acid sequences to continuous latent representations
2. a diffusion denoiser that generates latent vectors from Gaussian noise
3. decoders that reconstruct protein sequences and structures from the generated latent representations.
The approach enables joint sequence-structure generation while operating entirely in continuous latent space.

Performance Scaling with ESM-2 Encoders

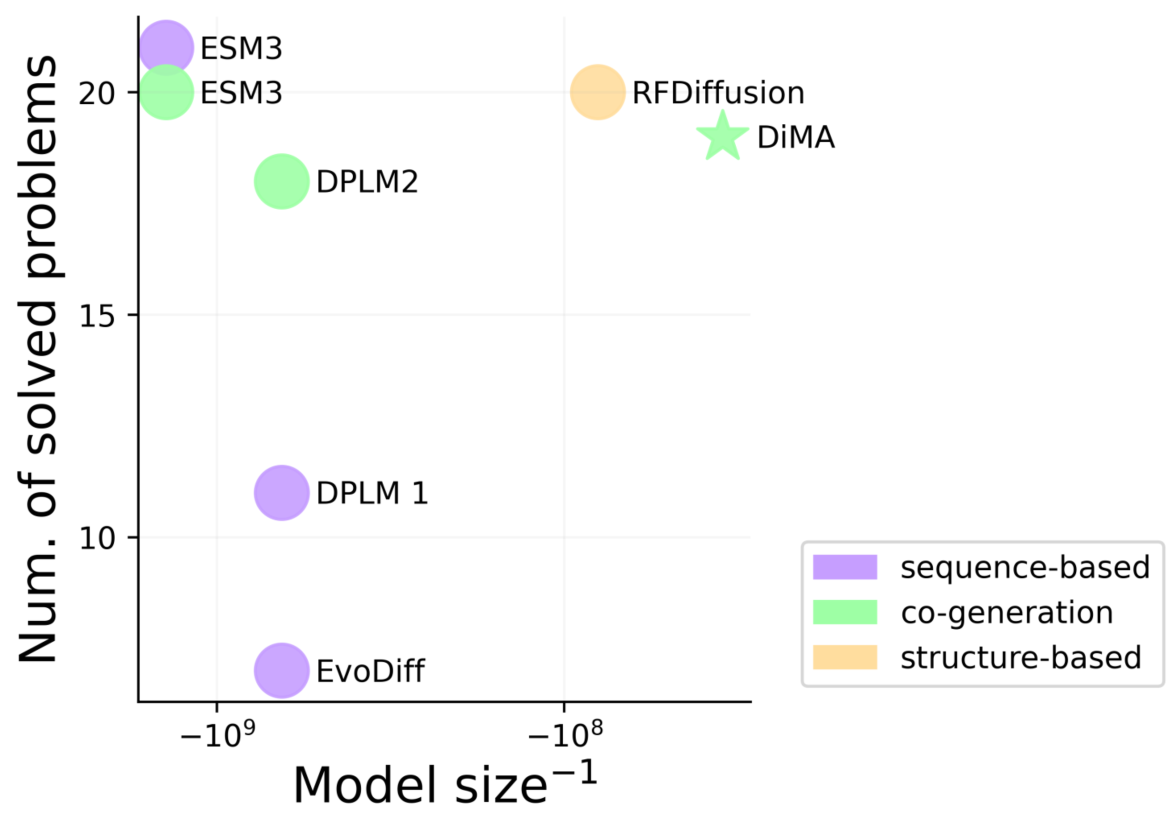
Encoder	FD-seq (↓)	pLDDT (↑)	$CD_{0.5}$ (↑)	Novelty (↑)
Dataset	0.11	83.9	0.994	57.6
Random	2.55	22.16	1.000	84.7
ESM-2 8M	0.560	74.25	0.981	68.0
ESM-2 35M	0.340	75.71	0.986	**69.1**
ESM-2 150M	0.323	80.07	**0.988**	65.6
ESM-2 650M	0.318	82.48	0.986	64.1
ESM-2 3B	**0.314**	**83.40**	0.969	63.0
ESMc 300M	0.326	82.70	0.963	64.2
CHEAP-shorten-1	0.346	81.92	0.951	64.6
CHEAP-shorten-2	0.340	78.81	0.946	66.2
SaProt 35M	0.366	82.23	0.976	65.5
SaProt 650M	0.411	83.01	0.980	65.7

Comparison with Pretrained Models



Quality-Diversity Trade-off. Circle area represents parameter count.

Functional-motif Scaffolding



Performance on Functional Motif-Scaffolding as a function of model size.

Fold-conditioned Generation