



GeomMotif

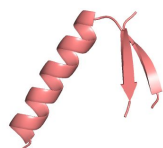
A Benchmark for Arbitrary Geometric Preservation in Protein Generation

Andrey Shevtsov

Research Associate, AIRI

Why do we need new proteins?

Scaffolding
Grafting



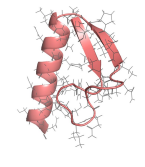
Create efficient
surrounding

- vaccines
- enzymes

Sequence
design



????????????????????

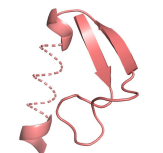


NAILAPVNLASQNSQGGVLNGFYSAN

New synthetic
protein generation

- antibiotic properties
- enzymes
- new active peptides

Protein
fragment
design



GTNAILAPVNLASQ???????FYSAN

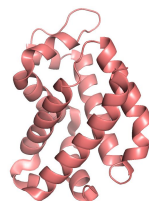
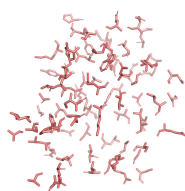


GTNAILAPVNLASQGGVLNGFYSAN

Completing
a protein region

- specific antibodies
- proteins with improved properties

De novo
generation

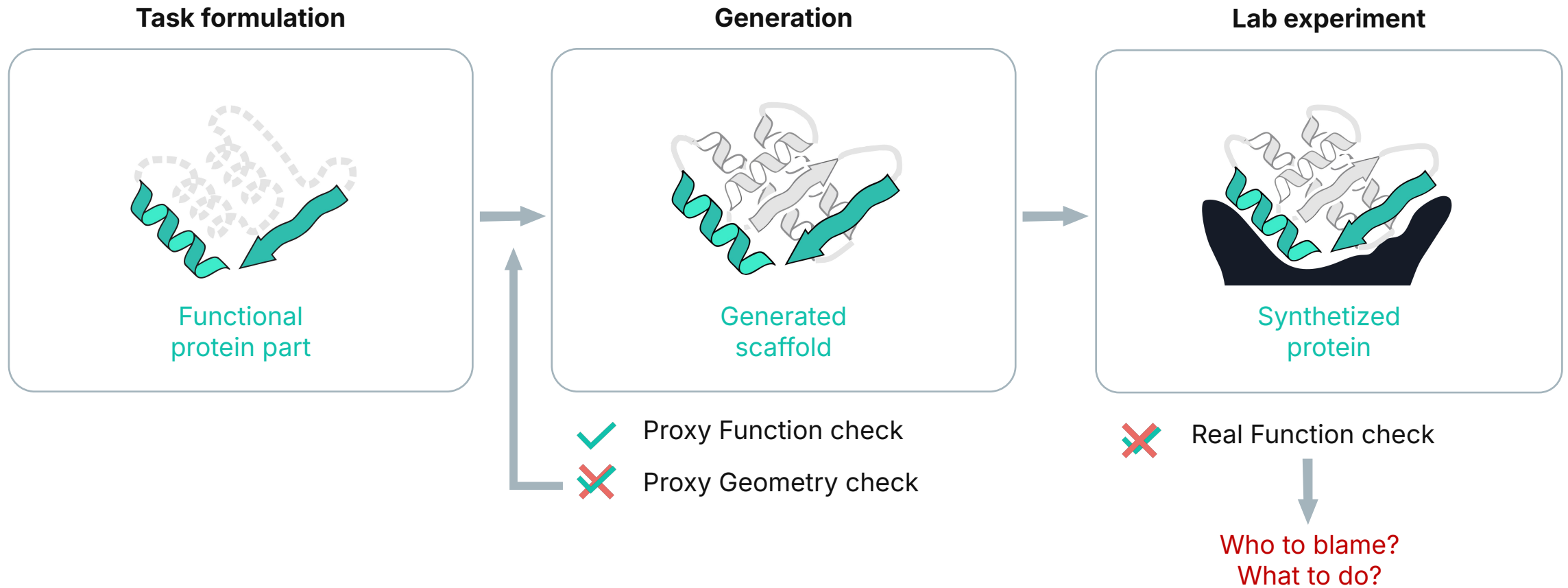


Create new
proteins

- vaccines
- new receptor binders

A Benchmark for Arbitrary Geometric Preservation in **Protein Generation**

Geometry and function



A Benchmark for Arbitrary **Geometric Preservation** in Protein Generation

Benchmark principles

GeomMotif

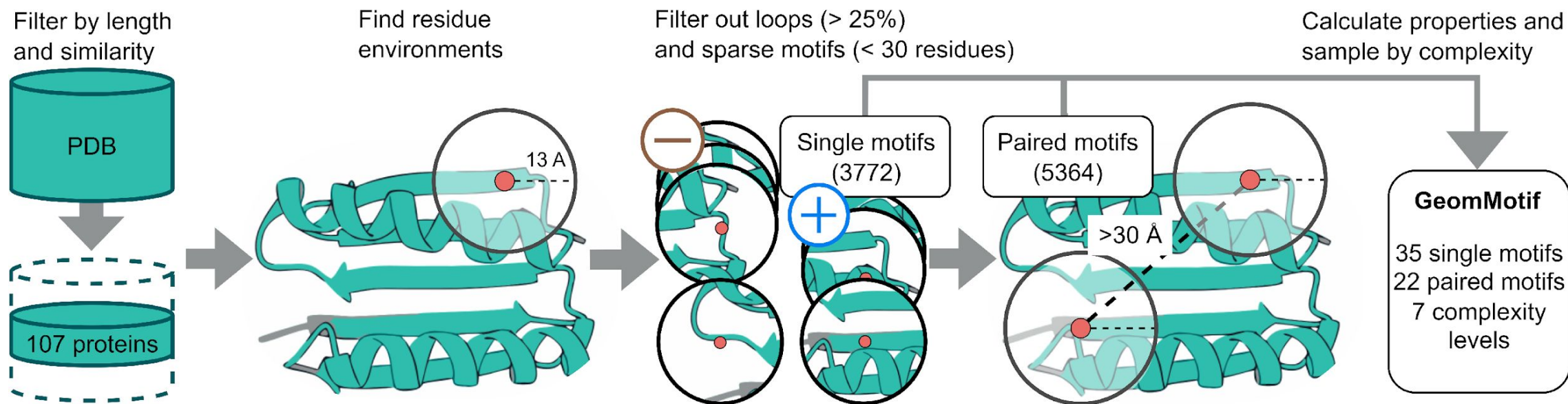
- ✓ Automatic coverage based choice
- ✓ Solvability
- ✓ Complexity levels
- ✓ Interpretability

Current Scaffolding Benchmark

- ✓ Manual functional choice
- ✗ Solvability
- ✗ Complexity levels
- ✗ Interpretability

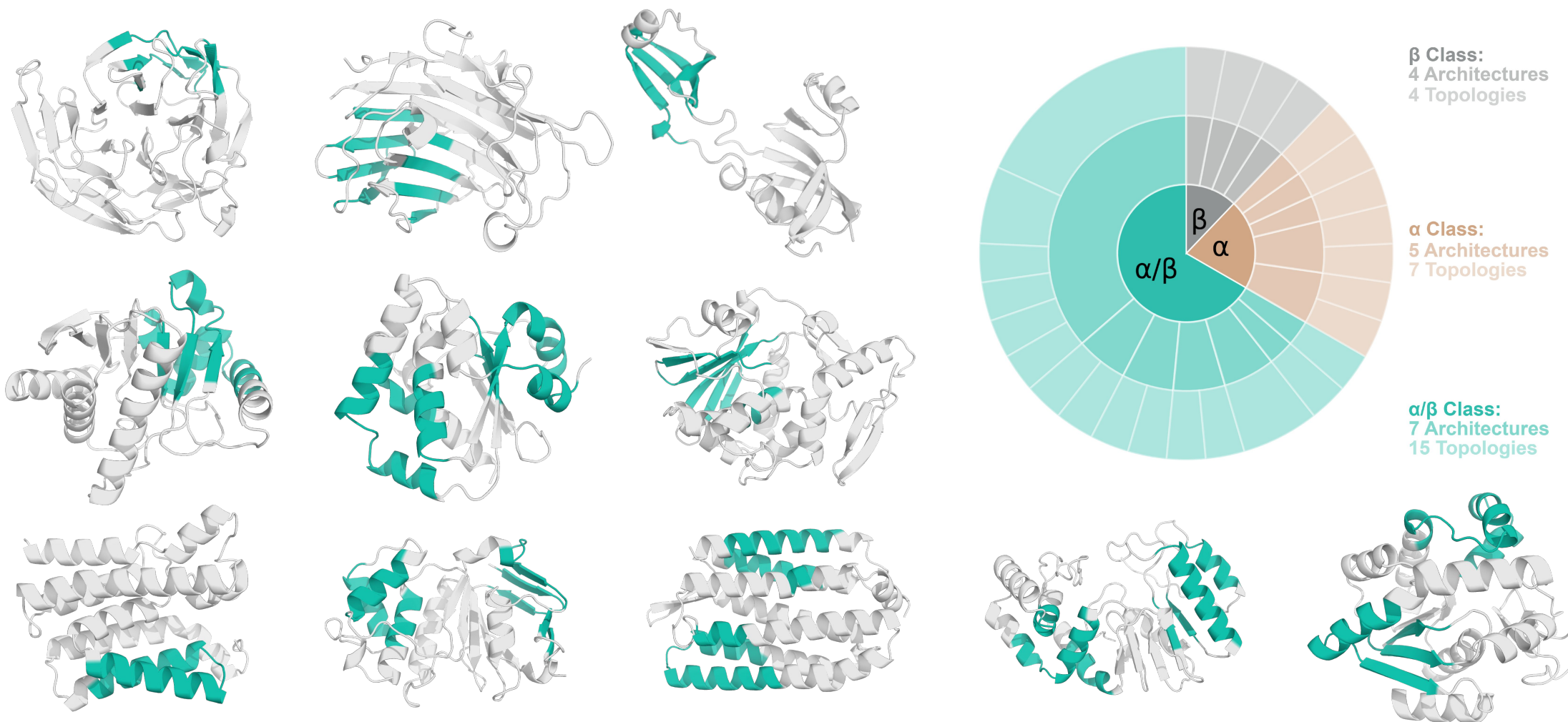
A Benchmark for **Arbitrary** Geometric Preservation in Protein Generation

Benchmark tests creation

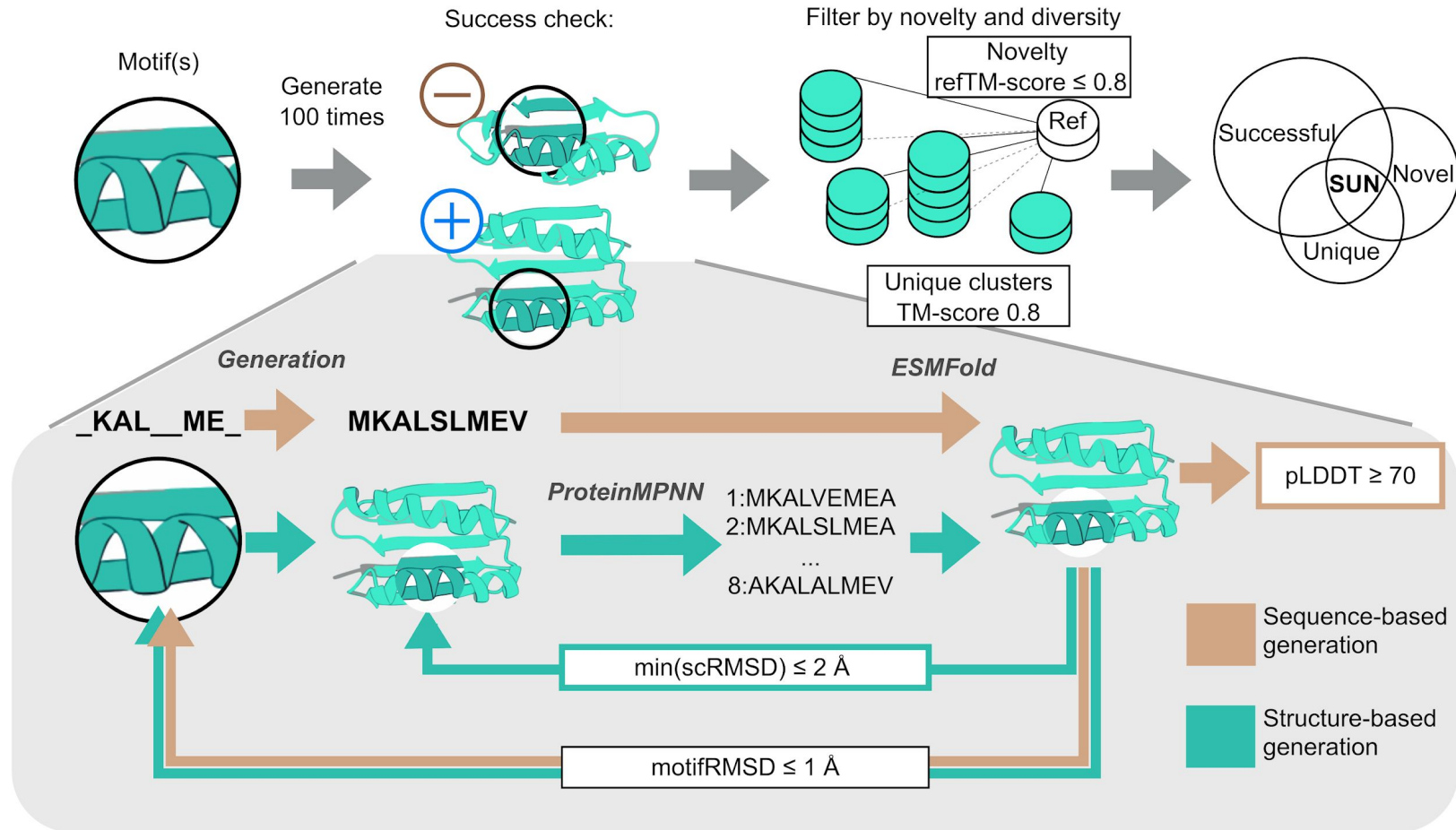


A **Benchmark** for Arbitrary Geometric Preservation in Protein Generation

Benchmark test cases



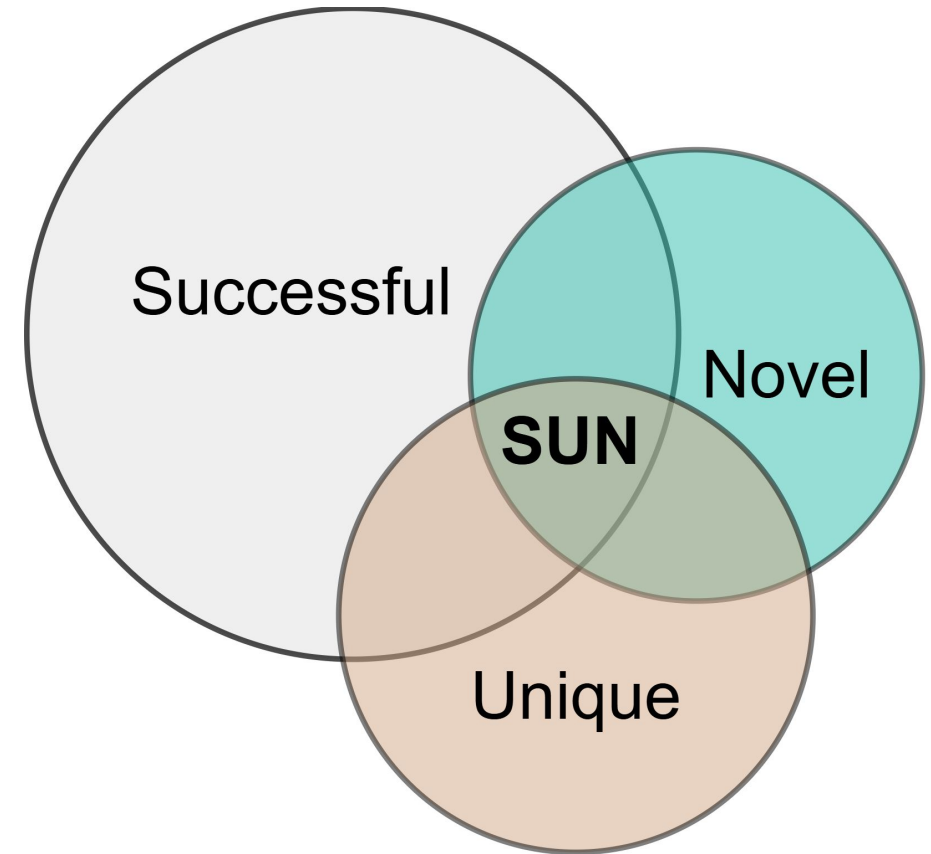
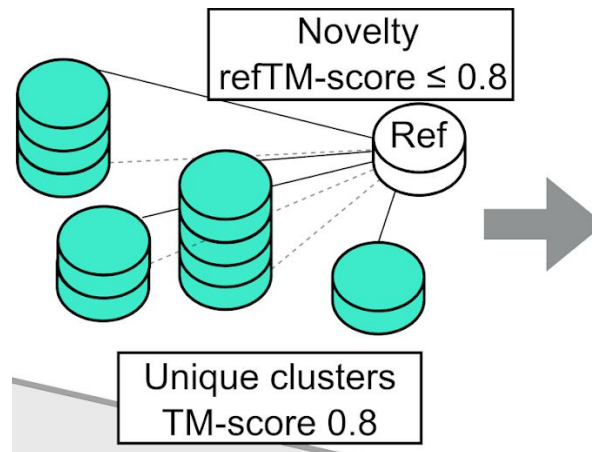
Benchmark validation



SUN metric

Practical intuition: an upper bound on the experimental success rate

Can be divided to the components for diagnostics



$$SUN = P(\text{Successful} \cap \text{Unique} \cap \text{Novel})$$

Models comparison

Structure:

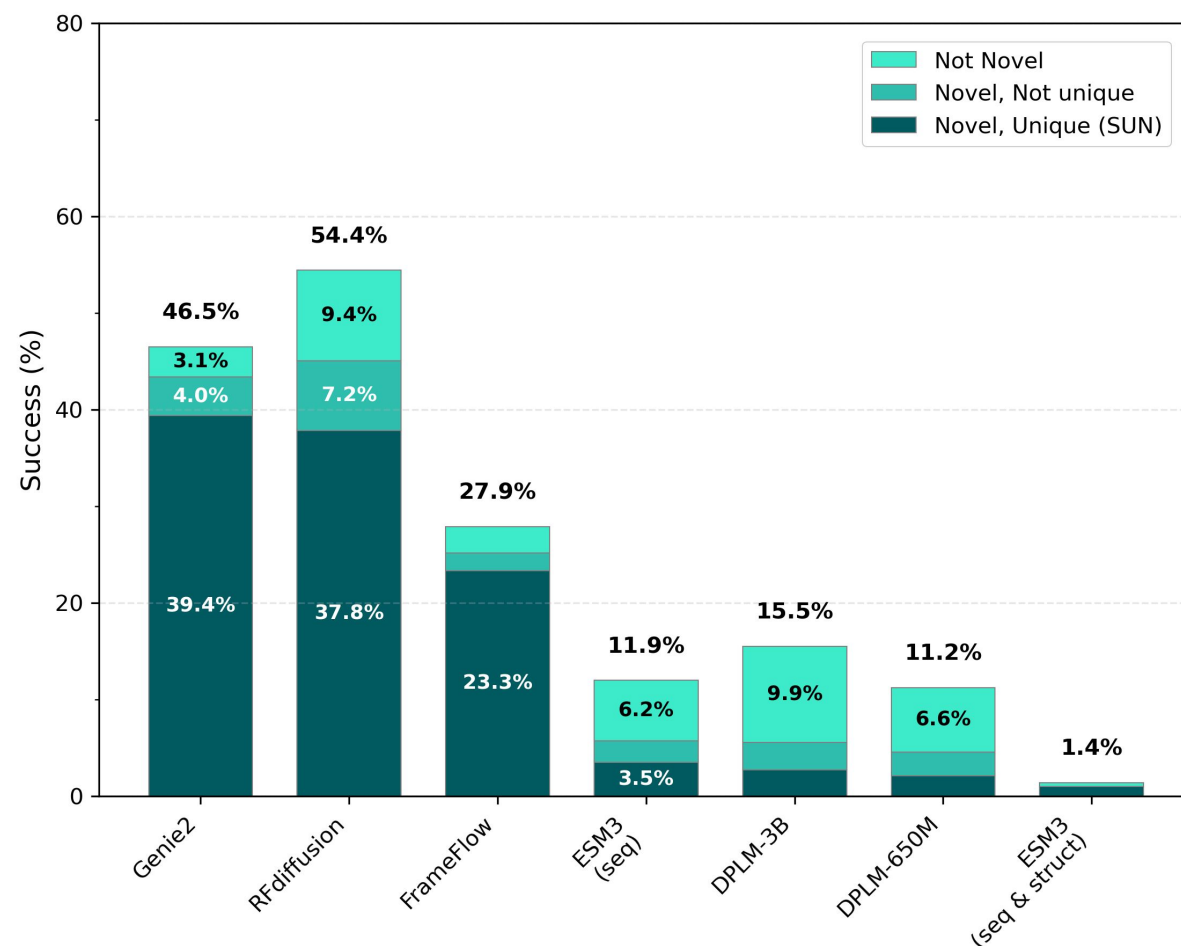
- Genie2
- Rfdiffusion
- FrameFlow

Sequence:

- ESM3 (seq)
- DPLM-650M

Sequence & Structure:

- ESM3 (seq & struct)



Models comparison

Model	Number of fragments in single motif							Number of fragments in paired motifs				
	1	2	3	4	5	6	7	3	4	5	6	7
Genie2	76.7 $\pm$ 0.5	75.4 $\pm$ 1.1	84.4 $\pm$ 1.6	38.7 $\pm$ 1.3	33.4 $\pm$ 2.8	49.9 $\pm$ 2.3	56.8 $\pm$ 0.7	39.2 $\pm$ 1.2	33.0 $\pm$ 0.7	15.2 $\pm$ 0.7	7.0 $\pm$ 0.8	12.9 $\pm$ 1.2
RFdiffusion	91.9 $\pm$ 0.7	72.8 $\pm$ 1.9	83.6 $\pm$ 0.8	57.9 $\pm$ 2.2	42.0 $\pm$ 2.8	53.8 $\pm$ 1.1	38.7 $\pm$ 0.8	14.8 $\pm$ 1.6	22.8 $\pm$ 1.3	5.9 $\pm$ 0.2	14.8 $\pm$ 0.2	8.0 $\pm$ 0.4
FrameFlow	61.9 $\pm$ 1.8	43.8 $\pm$ 1.2	49.9 $\pm$ 2.4	19.4 $\pm$ 0.8	9.8 $\pm$ 0.9	15.3 $\pm$ 1.5	12.8 $\pm$ 1.0	61.6 $\pm$ 1.8	17.1 $\pm$ 1.7	11.1 $\pm$ 0.9	9.7 $\pm$ 1.4	9.4 $\pm$ 0.9
DPLM-3B	4.6 $\pm$ 0.4	2.7 $\pm$ 0.3	2.4 $\pm$ 0.4	4.3 $\pm$ 0.6	3.4 $\pm$ 0.8	7.5 $\pm$ 1.0	10.4 $\pm$ 0.6	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	2.2 $\pm$ 0.4	0.0 $\pm$ 0.0
DPLM-650M	5.7 $\pm$ 0.7	2.0 $\pm$ 0.3	0.9 $\pm$ 0.3	1.3 $\pm$ 0.4	1.5 $\pm$ 0.3	7.7 $\pm$ 1.2	7.0 $\pm$ 0.6	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	1.0 $\pm$ 0.3	0.0 $\pm$ 0.0
ESM3 (seq)	12.1 $\pm$ 1.0	7.0 $\pm$ 0.7	1.0 $\pm$ 0.1	3.3 $\pm$ 0.7	4.2 $\pm$ 0.8	8.0 $\pm$ 0.9	12.2 $\pm$ 1.0	0.2 $\pm$ 0.2	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0
ESM3 (seq&struct)	7.9 $\pm$ 0.5	3.0 $\pm$ 0.6	0.3 $\pm$ 0.2	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	1.1 $\pm$ 0.7	1.8 $\pm$ 0.7	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0

Models comparison: main outcomes

- Geometry preservation is not solved yet
- Structure models are better than sequence models
- Multiple modalities can mess the generation



Authors

P. Strashnov, A. Shevtsov, V. Meshchaninov, O. Kardymon, D. Vetrov



@Andr_Shevtsov





airi.net



[airi_research_institute](https://t.me/airi_research_institute)



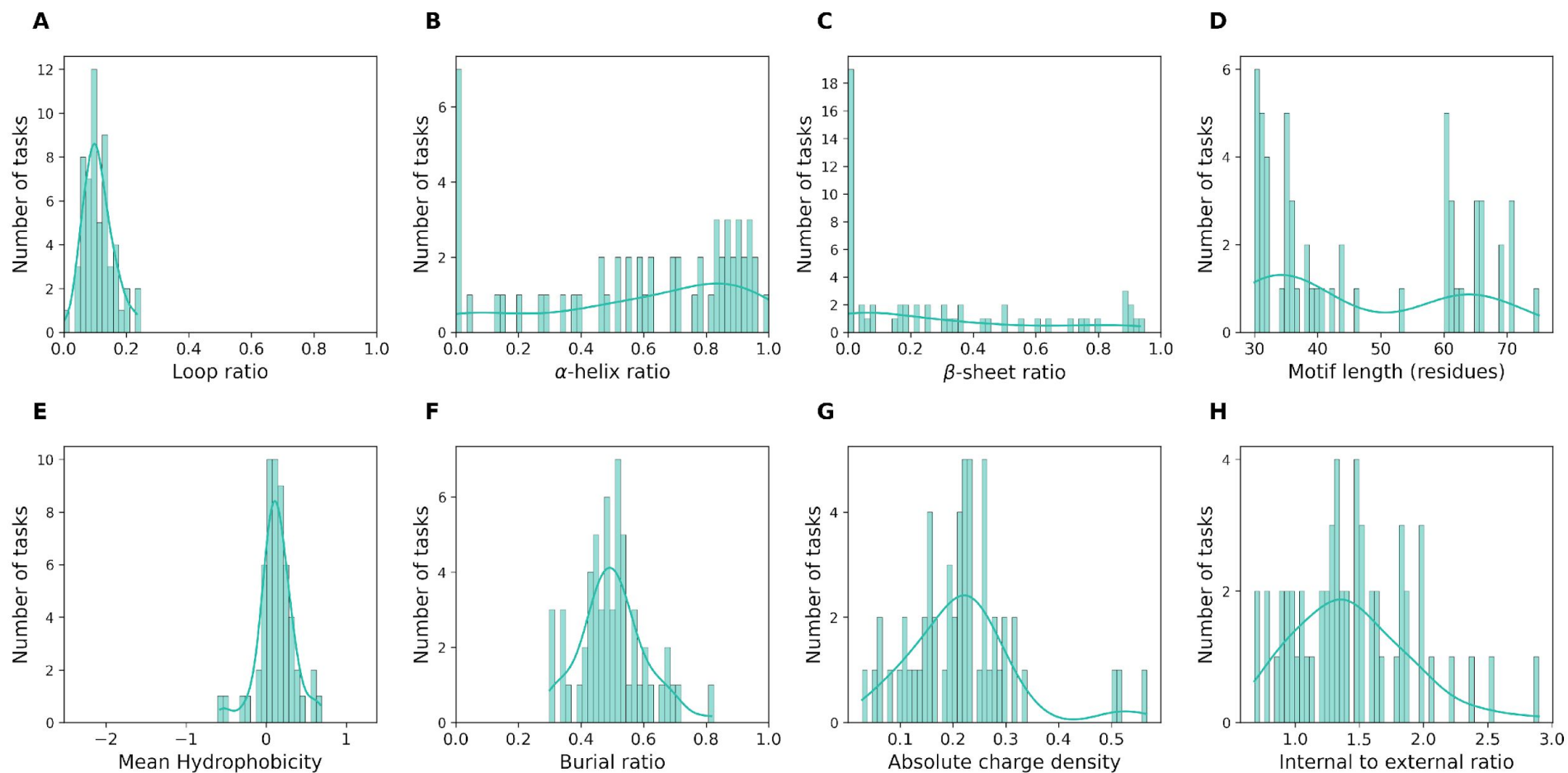
[AIRI Institute](https://vk.com/AIRI_Institute)



Telegram

AIR
I

Interpretation: Biological properties



Interpretation: Biological properties vs SUN

