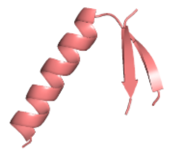
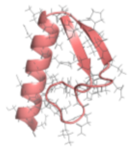
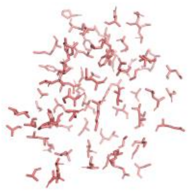
AIRI

Robust Evaluation Strategies for Protein Design

Andrey Shevtsov

Research engineer, AIRI

Why protein generation

Scaffolding			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

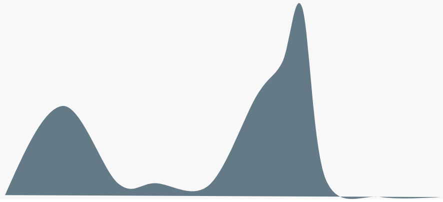
Generative models

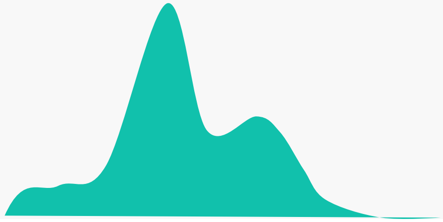
- → Autoregressive models: ProGen, RITA, ProtGPT2
- → Diffusion models: Evodiff, DPLM, Rfdiffusion
- → Flow matching models: FoldFlow, AlphaFlow, FrameFlow, MultiFlow
- → GAN: ProteinGAN

What makes a generative model great?

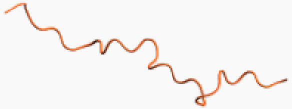
✗ ✓

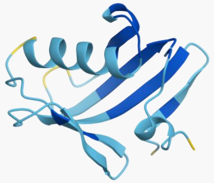
→ Generated distribution is similar to the training set



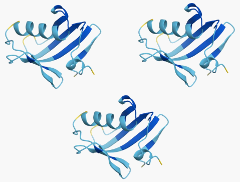


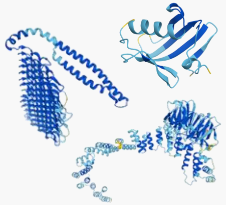
→ High quality samples





→ Diverse samples





Challenges

- Limited human feedback capability
- Lack of standardized metrics
- Insufficient metric validation

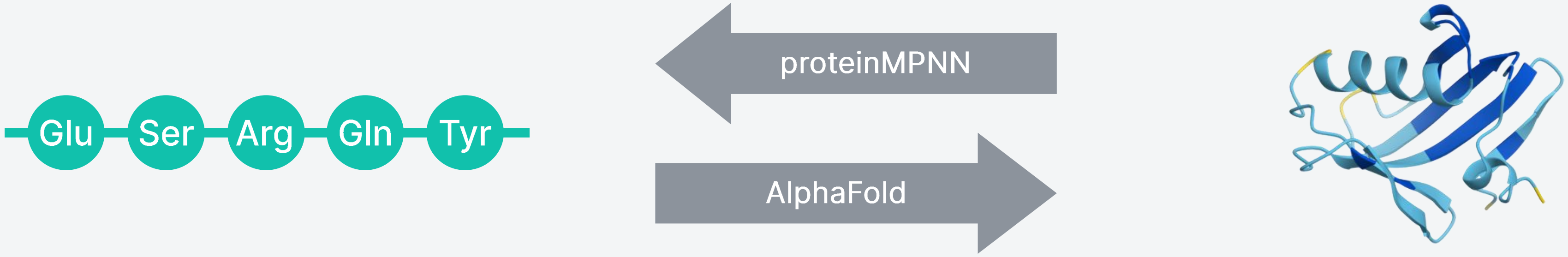
We created new SOTA model.
It generates the GREENEST proteins

We created new SOTA model.
It generates the HARDEST proteins

We created new SOTA model.
It generates the SMARTEST proteins

The bright side of protein GenAI evaluation

Protein modalities



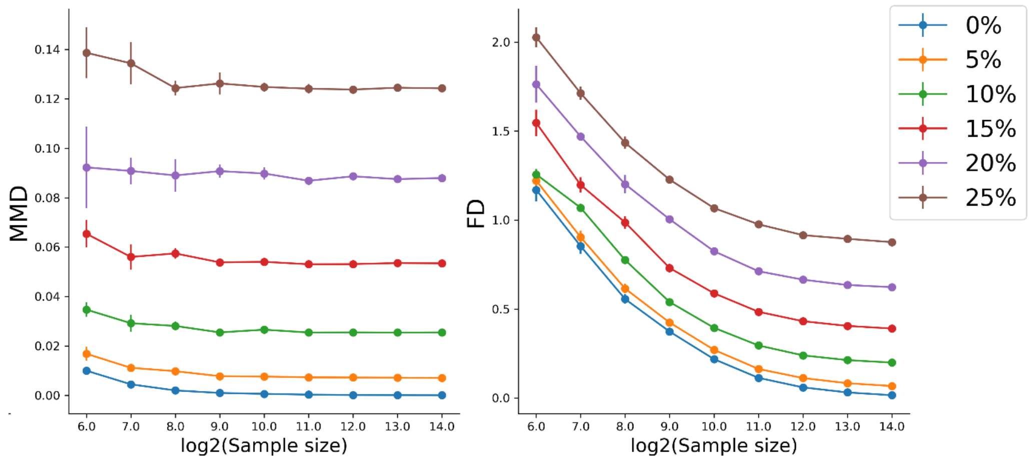
Experience from other fields

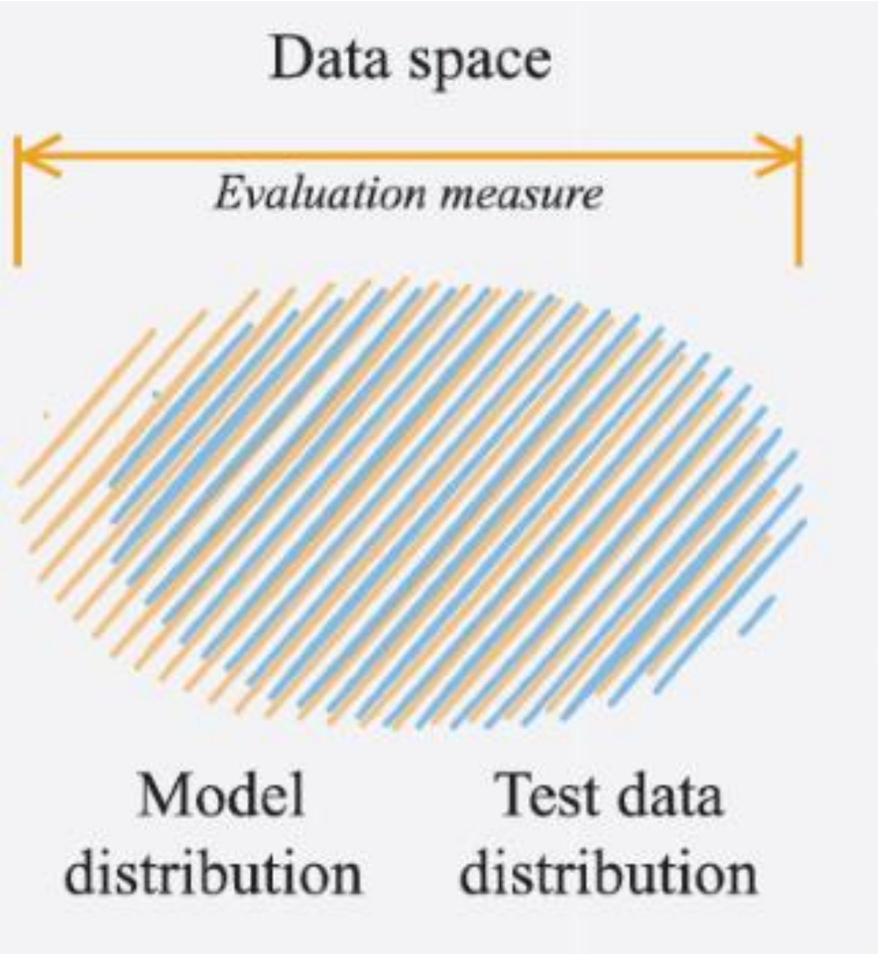
NLP

Image generation

Distribution similarity metrics

- → Fréchet distance (1000+ samples)
- → MMD (500+ samples)
- → MMD kernel: RBF (sigma=10)
- → Latent space: ProtT5 sequence embeddings



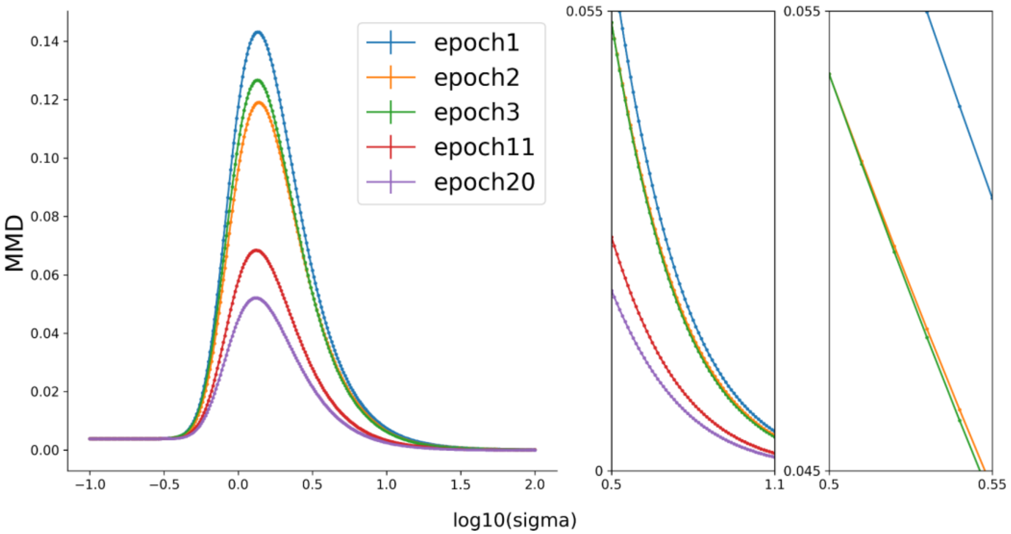


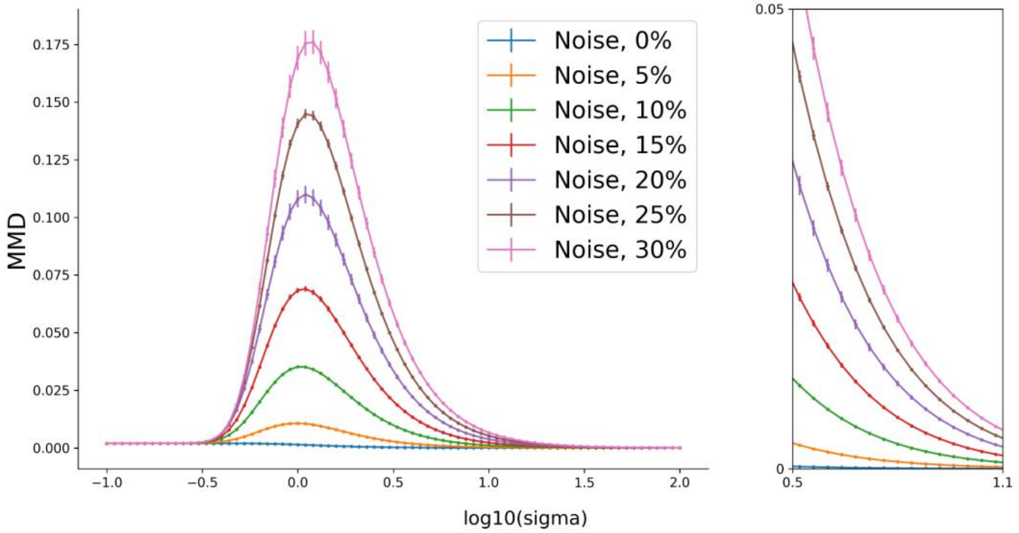
MMD Sigma choice

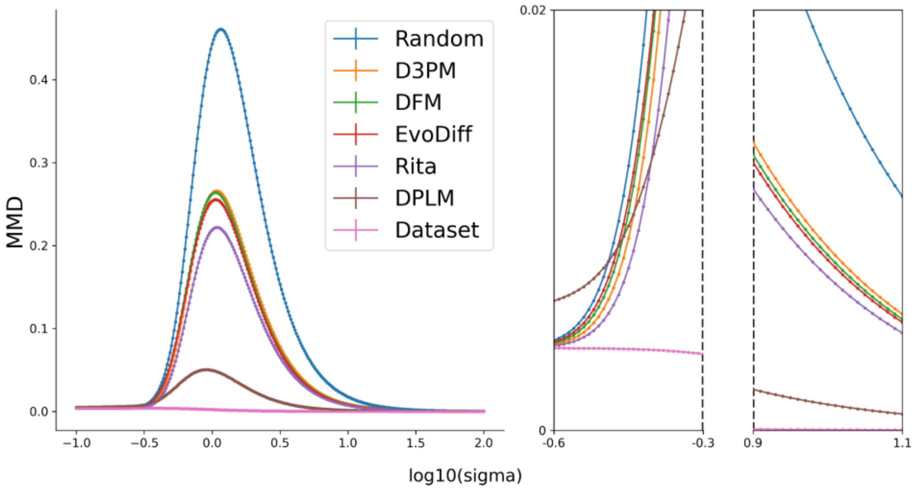
$$MMD^2(P,Q) = \mathbb{E}_{\mathcal{X}\sim P}\left[k(x,x')\right] +$$

$$+\,\mathbb{E}_{\mathcal{Y}\sim Q}\left[k(y,y')\right] - 2\mathbb{E}_{\mathcal{X},\mathcal{Y}\sim P,Q}\left[k(x,y)\right]$$

$$K(X_1,X_2) = \exp(-\frac{\|X_1 - X_2\|^2}{2\sigma^2})$$

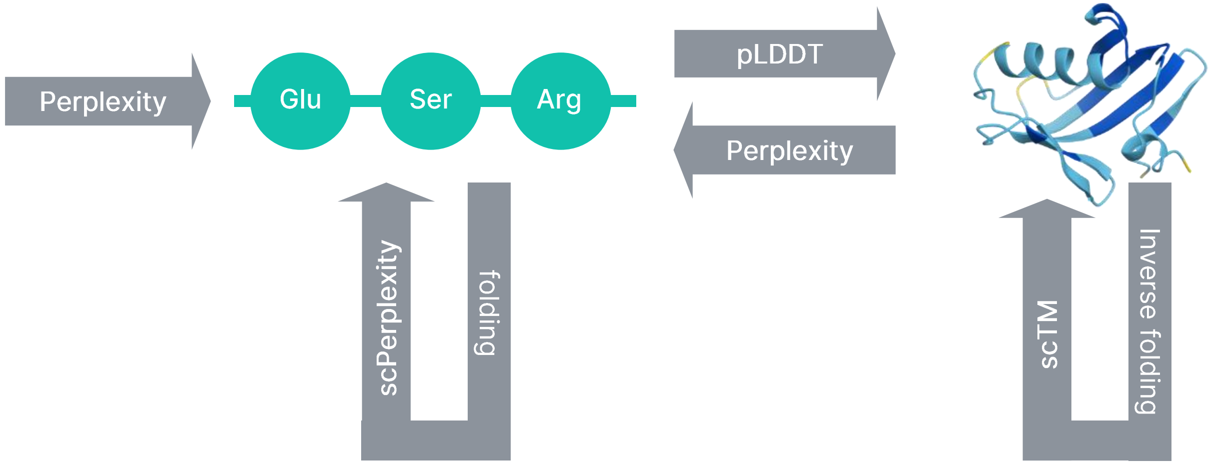






Quality metrics

- → Do not use only pLDDT or perplexity
- → Use both pLDDT and Perplexity
- → Use scPerplexity/ scTM
- → OmegaFold, ESMFold, ProteinMPNN



Perplexity

Glu

Ser

Arg

pLDDT

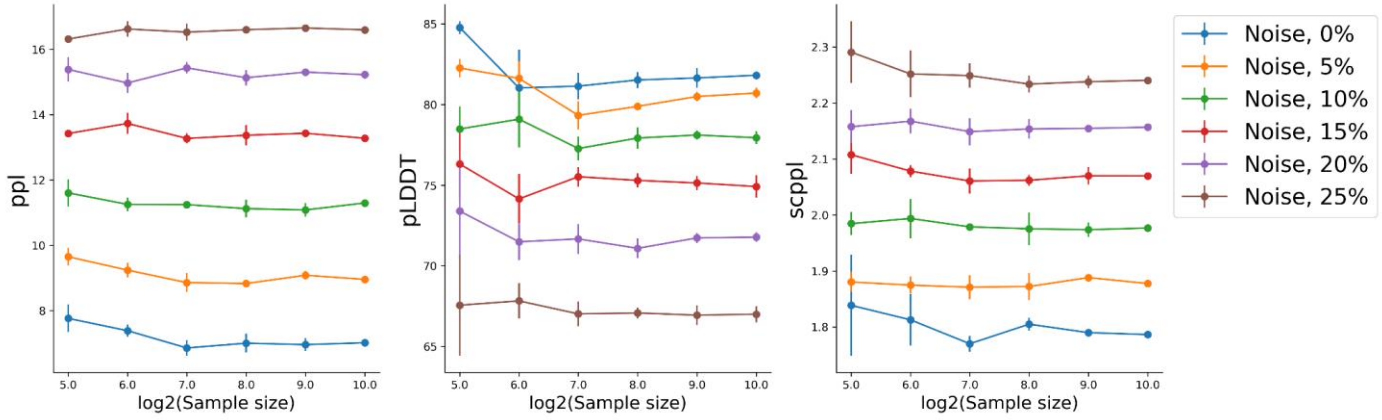
Perplexity

scPerplexity

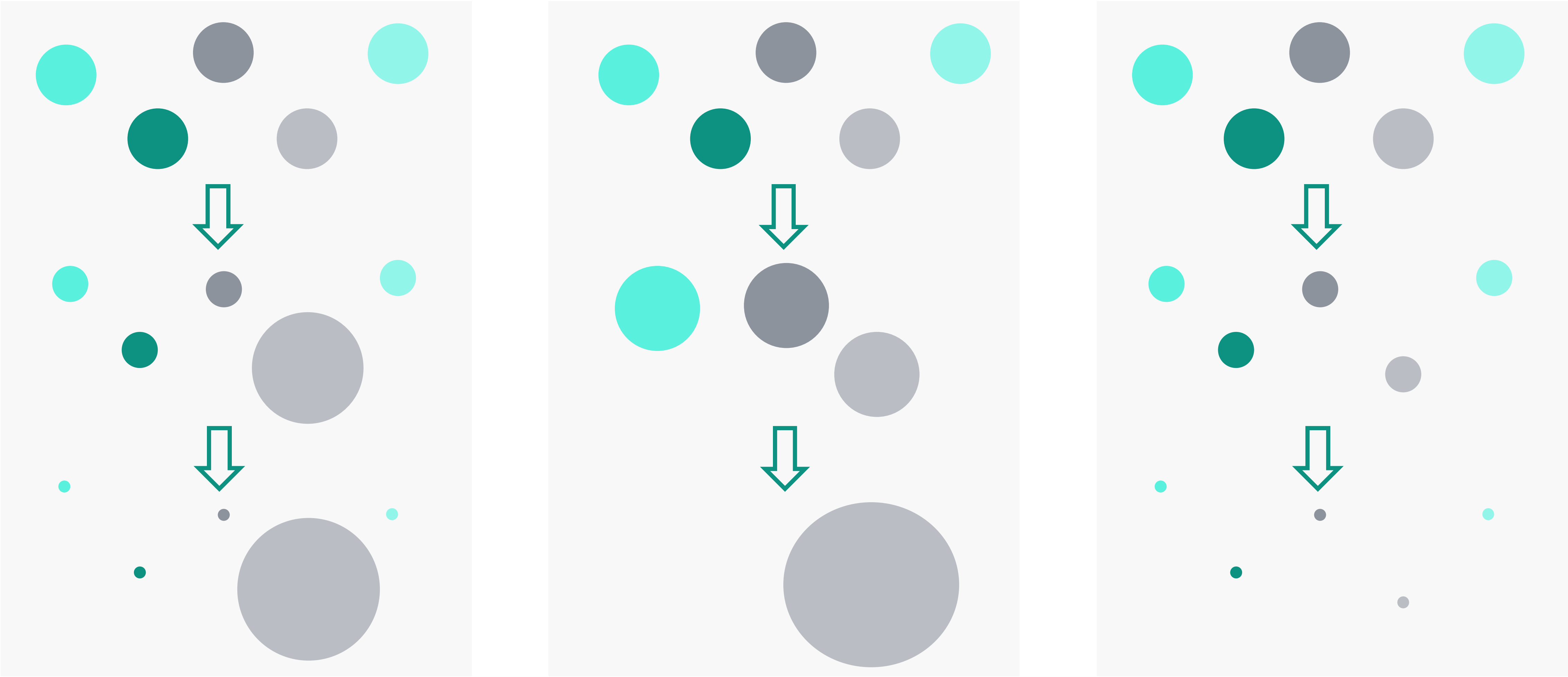
folding

scTM

Inverse folding



How to identify diverse generation?

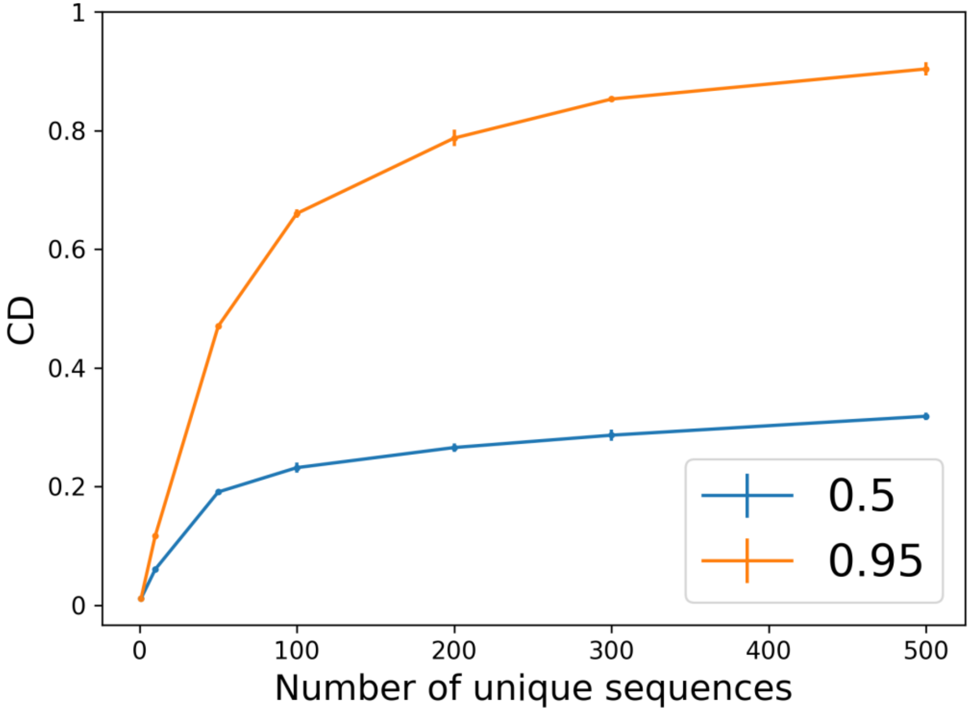


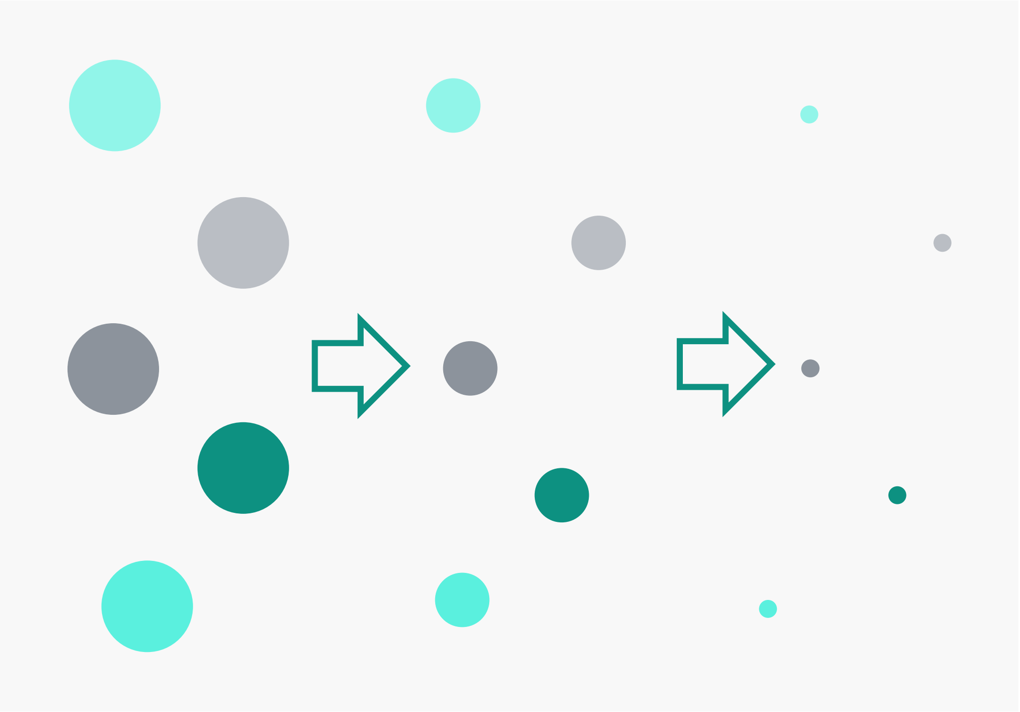
Inner diversity

→ Clustering method: MMseqs2

→ 2 thresholds: 0.5 and 0.95

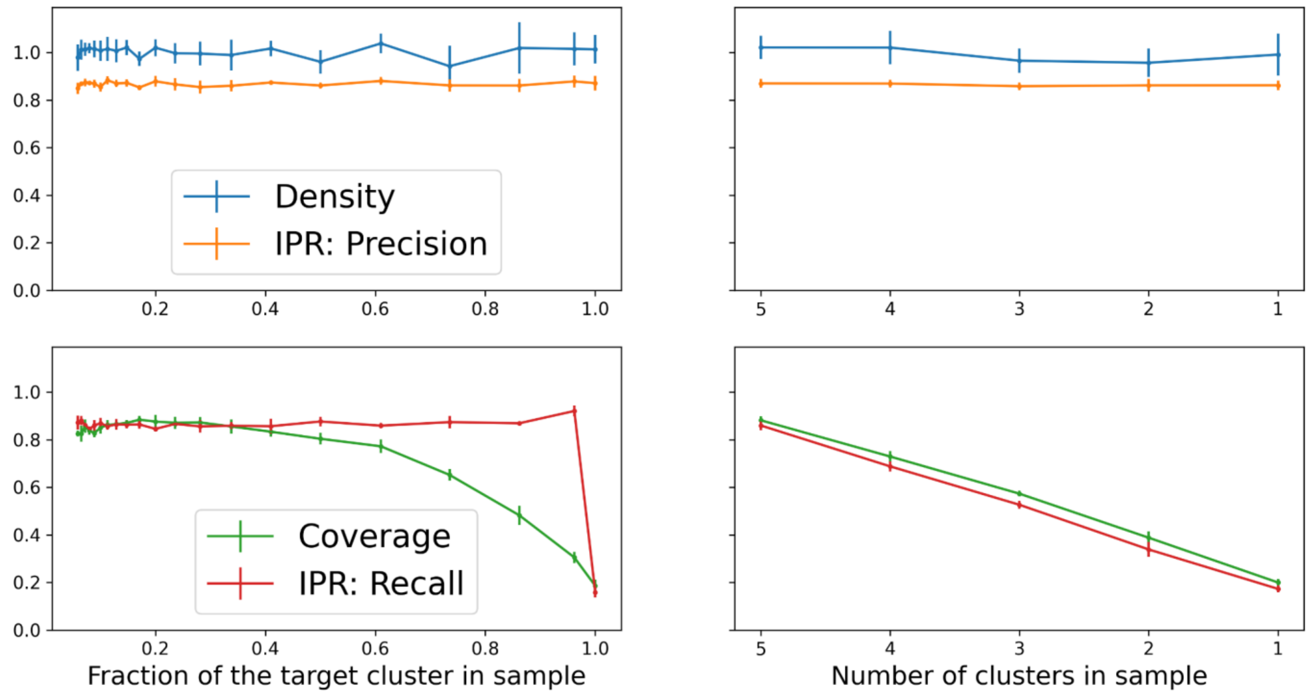
$$CD = \frac{\#Clusters}{\#Seqs}$$

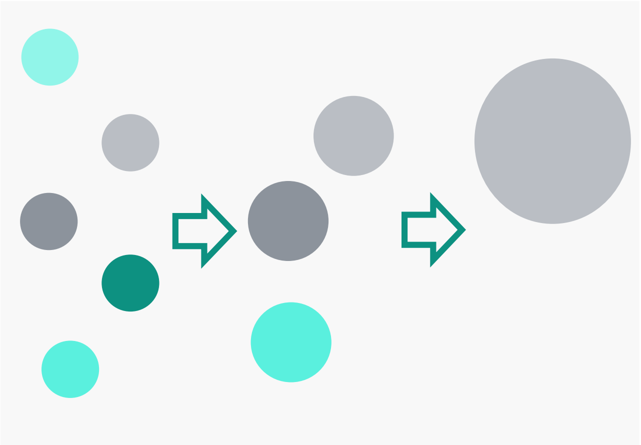


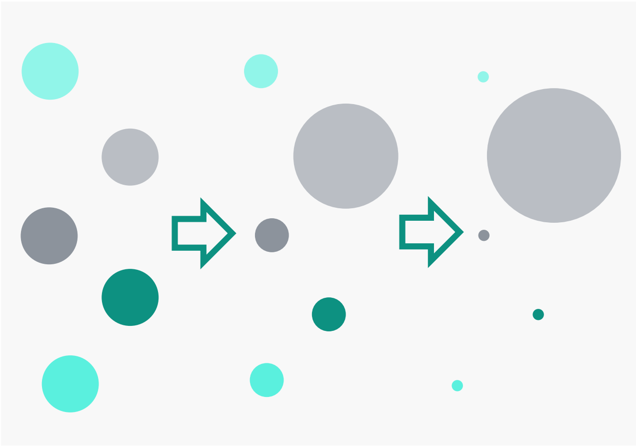


Mixed metrics: quality+diversity

→ DC is better than IPR



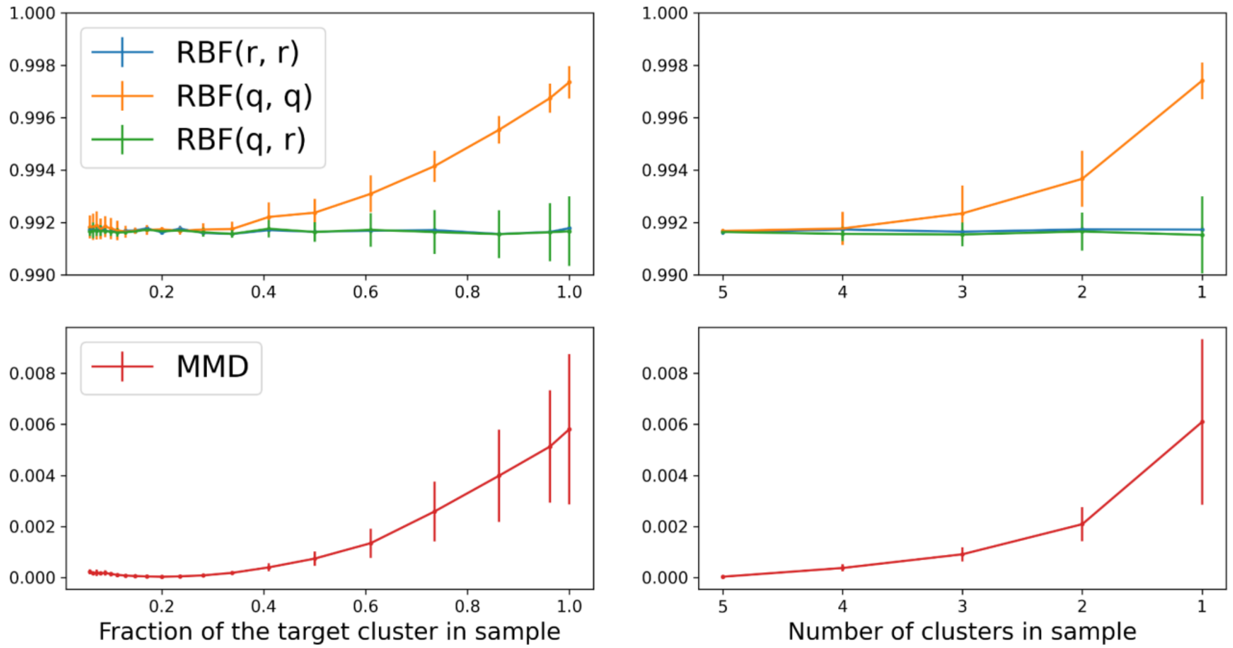


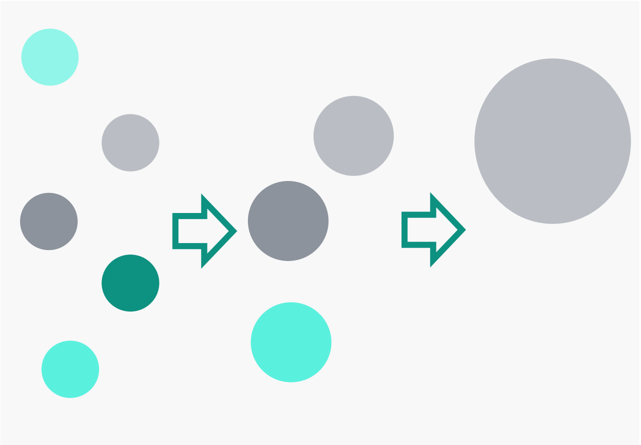


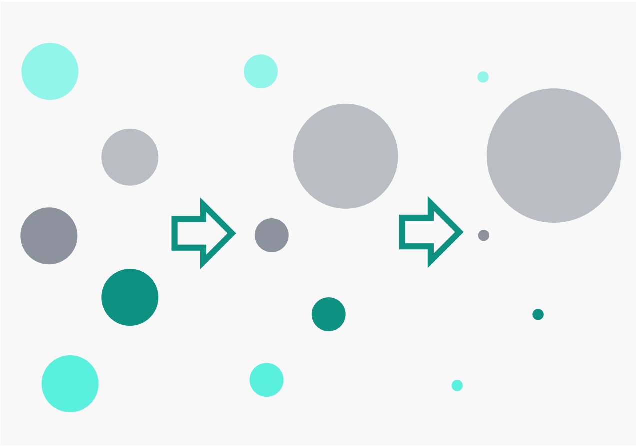
https://doi.org/10.48550/arXiv.2002.09797
https://doi.org/10.48550/arXiv.1904.06991

MMD can be used this way

$$MMD^2(P,Q) = \mathbb{E}_{\mathcal{X}\sim P}\left[k(x,x')\right] +$$
$$+\,\mathbb{E}_{\mathcal{Y}\sim Q}\left[k(y,y')\right] - 2\mathbb{E}_{\mathcal{X},\mathcal{Y}\sim P,Q}\left[k(x,y)\right]$$

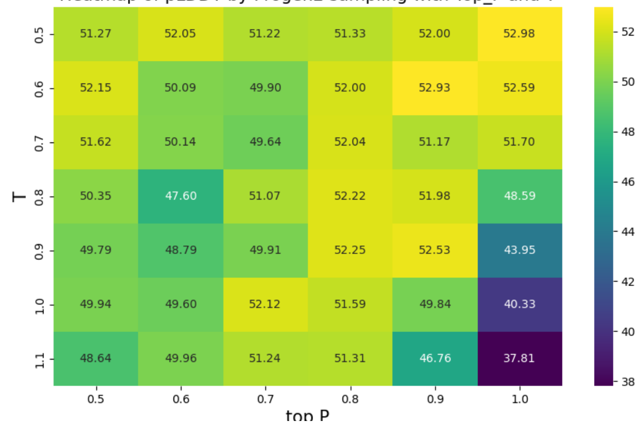


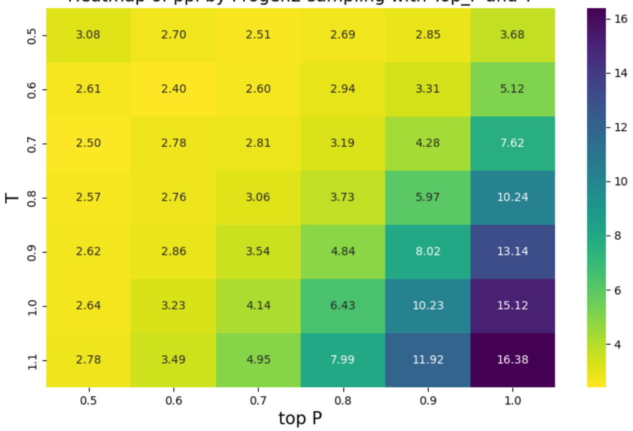




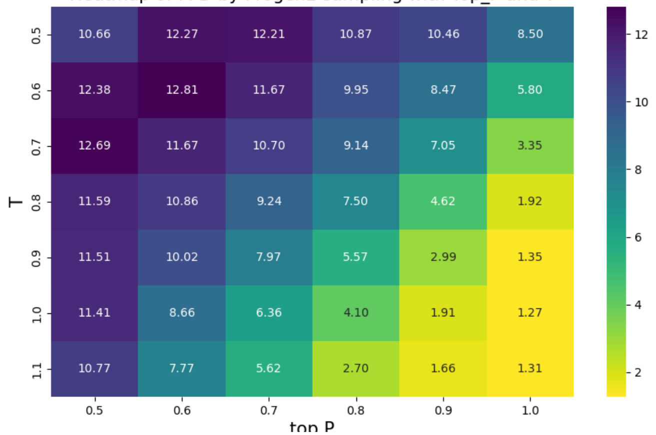
Quality vs Diversity tradeoff

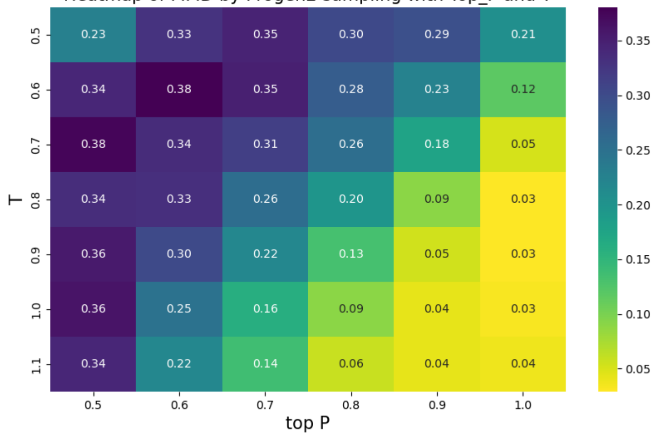
Quality


Heatmap of pLDDT by Progen2 sampling with Top_P and T

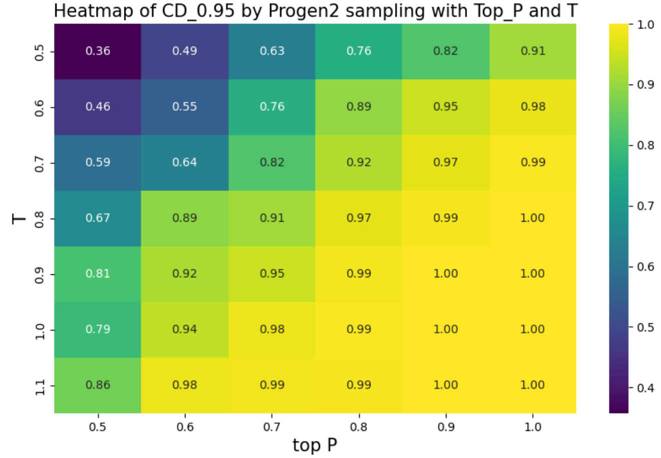

Heatmap of ppl by Progen2 sampling with Top_P and T

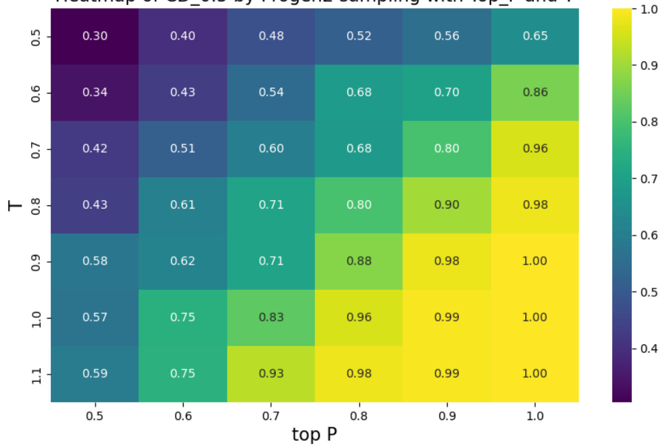
Distribution


Heatmap of FPD by Progen2 sampling with Top_P and T

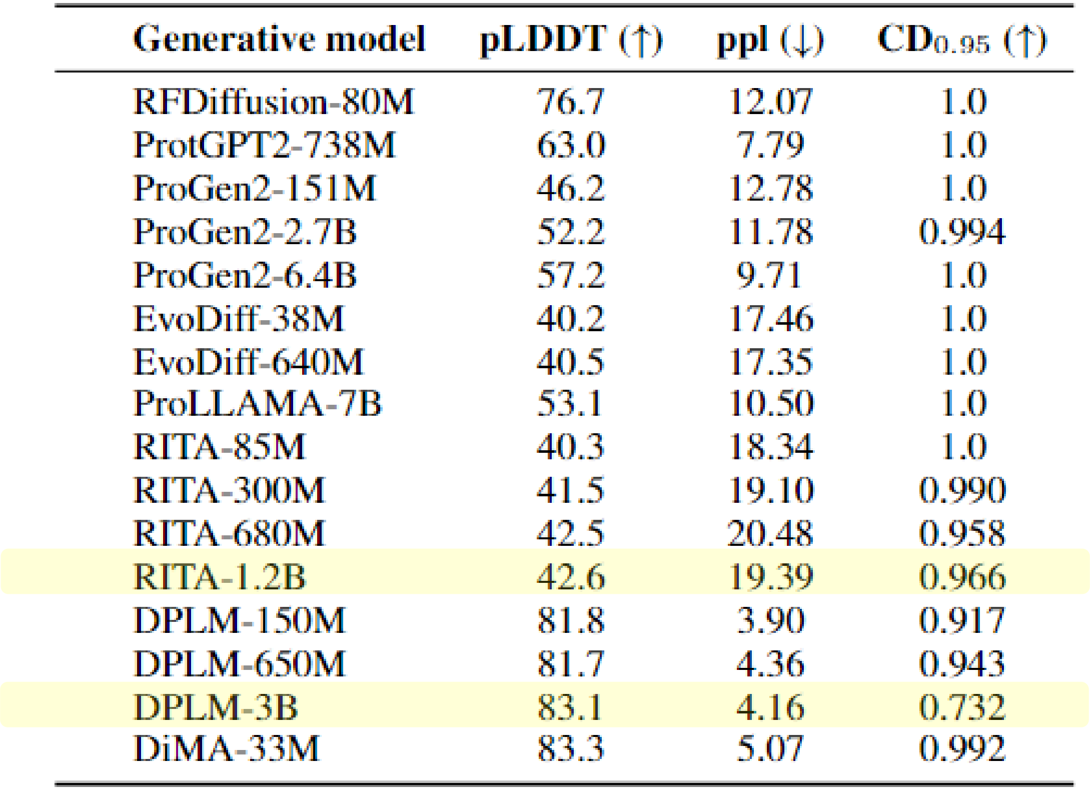

Heatmap of MMD by Progen2 sampling with Top_P and T

Diversity


Heatmap of CD_0.95 by Progen2 sampling with Top_P and T

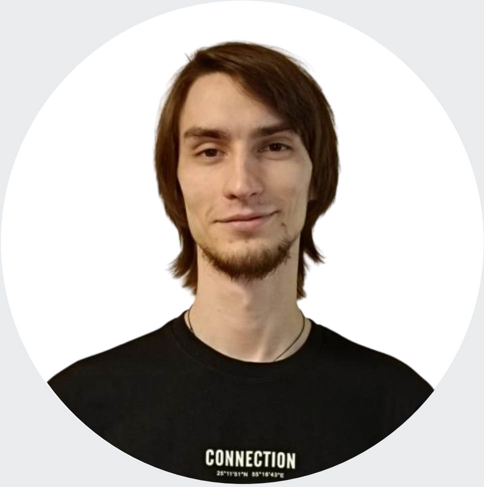

Heatmap of CD_0.5 by Progen2 sampling with Top_P and T

Models comparison



Generative model	pLDDT (↑)	ppl (↓)	$\text{CD}_{0.95}$ (↑)
RFDiffusion-80M	76.7	12.07	1.0
ProtGPT2-738M	63.0	7.79	1.0
ProGen2-151M	46.2	12.78	1.0
ProGen2-2.7B	52.2	11.78	0.994
ProGen2-6.4B	57.2	9.71	1.0
EvoDiff-38M	40.2	17.46	1.0
EvoDiff-640M	40.5	17.35	1.0
ProLLAMA-7B	53.1	10.50	1.0
RITA-85M	40.3	18.34	1.0
RITA-300M	41.5	19.10	0.990
RITA-680M	42.5	20.48	0.958
RITA-1.2B	42.6	19.39	0.966
DPLM-150M	81.8	3.90	0.917
DPLM-650M	81.7	4.36	0.943
DPLM-3B	83.1	4.16	0.732
DiMA-33M	83.3	5.07	0.992

Contacts



Andrey Shevtsov

Research engineer, AIRI

Mail: Shevtsov@airi.net @Andr_Shevtsov

AIRI

airi.net

airi_research_institute

AIRI Institute

Telegram

AIRI