Oral Presentation –
Top 1.2%

Protein Discovery with Discrete Walk-Jump Sampling

Nathan C. Frey*, Daniel Berenberg*, Karina Zadorozhny, Joseph Kleinhenz, Julien Lafrance-Vanasse, Isidro Hotzel, Yan Wu, Stephen Ra, Richard Bonneau, Kyunghyun Cho, Andreas Loukas, Vladimir Gligorijević, **Saeed Saremi**

Prescient Design
A Genentech Accelerator

Summary

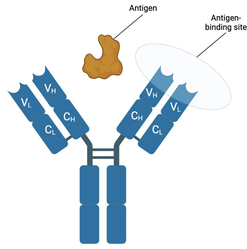
We resolve difficulties in training and sampling from a discrete generative model by learning the gradient of a ***smoothed energy function***, sampling from the smoothed data manifold with Langevin MCMC, and projecting back to the true data manifold with one-step denoising, requiring ***only a single noise level***.

Contact | Paper | Code

@nc_frey
@dabkiel1

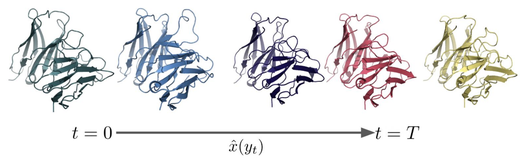
Background

Antibody Discovery



- Find promising lead candidates by searching animal immune repertoires or synthetic libraries

Sequence & Structure



Neural Empirical Bayes

- from X to $Y = X + N$, $N \sim \mathcal{N}(0, \sigma^2 I)$.
- Empirical Bayes [1, 2, 3] as the probabilistic machinery for denoising:

$$\hat{x}(y) = \frac{\int x p(y|x) p(x) dx}{\int p(y|x) p(x) dx} = \mathbb{E}[X|y] = y + \sigma^2 \underbrace{\nabla \log p(y)}_{\text{score function}}.$$

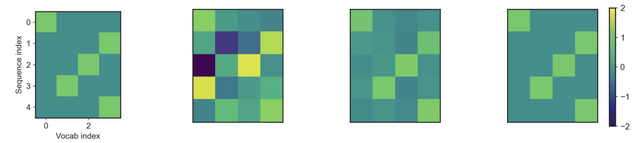
- Learn $g_\phi(y) \approx \nabla \log p(y)$ by denoising (*the noise level σ can be any value*):

$$\mathcal{L}(\phi) = \mathbb{E}_{x \sim p(x),\, \varepsilon \sim \mathcal{N}(0,I)} \|X - \hat{x}_\phi(X + \sigma\varepsilon)\|^2,$$
$$g_\phi(y) = \frac{1}{\sigma^2}\left(\hat{x}_\phi(y) - y\right).$$

Input → Output

Sequence → One hot encoding → Noisy one hot → Denoised → Argmax decoding

$x = \text{EVQLV...} \quad y = x + \varepsilon \quad \hat{x}_\phi = y + \sigma^2 g_\phi(y) \quad s = \operatorname{argmax} \hat{x}_\phi$



$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_d)$

Method

Walk

$$\mathcal{L}(\theta) = \underbrace{\mathbb{E}_{x \sim p(x),\, \varepsilon \sim \mathcal{N}(0,I)}[f_\theta(x + \sigma\varepsilon)]}_{\mathbb{E}[f_\theta(Y_+)]} - \underbrace{\mathbb{E}_{y \sim e^{-f_\theta(y)}}[f_\theta(y)]}_{\mathbb{E}[f_\theta(Y_-)]}$$

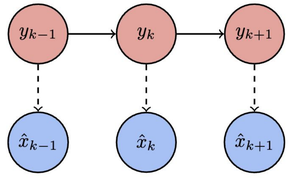
$$x \xrightarrow{+\mathcal{N}(0, \sigma^2 I_d)} y \to \boxed{\text{dEBM}} \to f_\theta(y)$$

$$y \sim e^{-f_\theta(y)} \to \boxed{\text{ByteNet}} \to \hat{x}_\phi(y)$$

$$x_{k+1} = x_k - \delta \nabla f_\theta(x_k) + \sqrt{2\delta}\, \varepsilon_k,\ \varepsilon_k \sim \mathcal{N}(0, I_d).$$

Jump

Use the learned least-squares denoiser $\hat{x}_\phi(y_k)$ for any draw $y_k \sim p(y)$ obtained by Langevin MCMC.



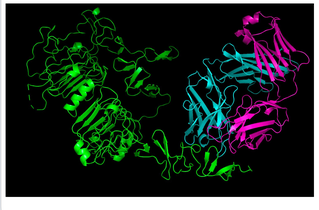
Results

In silico

Model	$W_{\text{property}} \downarrow$	Unique ↑	E_{dist} ↑	IntDiv ↑	DCS ↑
dWJS (energy-based)	**0.056**	**1.0**	58.4	55.3	0.38
dWJS (score-based)	0.065	0.97	**62.7**	**65.1**	0.49
SeqVDM	0.062	**1.0**	60.0	57.4	0.40
DEEN	0.087	0.99	50.9	42.7	0.41
GPT 3.5	0.14	0.66	55.4	46.1	0.23
IgLM	0.08	**1.0**	48.6	34.6	**0.533**
ESM2	0.15	**1.0**	70.99*	77.56*	0.061

- Benchmarked against diffusion, autoregressive and masked LMs, LLMs+ICL

In vitro



- 70% binding Trastuzumab CDR H3 designs, compared to 25% (discrete diffusion) [4] and 22% (structural diffusion) [5]

References

1. Robbins, “An empirical Bayes approach to statistics” (1956).
2. Miyasawa, “An empirical Bayes estimator of the mean of a normal population” (1961).
3. Saremi, Saeed, and Aapo Hyvärinen. "Neural empirical bayes." (2019).
4. Gruver, Stanton, Frey, et al., “Protein Design with Guided Discrete Diffusion” *NeurIPS* (2024).
5. Martinkus, et al. "AbDiffuser: full-atom generation of in-vitro functioning antibodies." *NeurIPS* (2024).

***Equal contributions**

Correspondence to:
{frey.nathan.nf1, saremi.saeed}@gene.com

ICLR