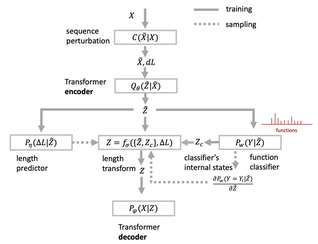
Multi-segment preserving sampling for deep manifold sampler

Daniel Berenberg, Jae Hyeon Lee, Simon Kelow, Ji Won Park, Andy Watkins, Richard Bonneau, Vladimir Gligorijević, Stephen Ra, Kyunghyun Cho

Prescient Design
A Genentech Accelerator

Motivation

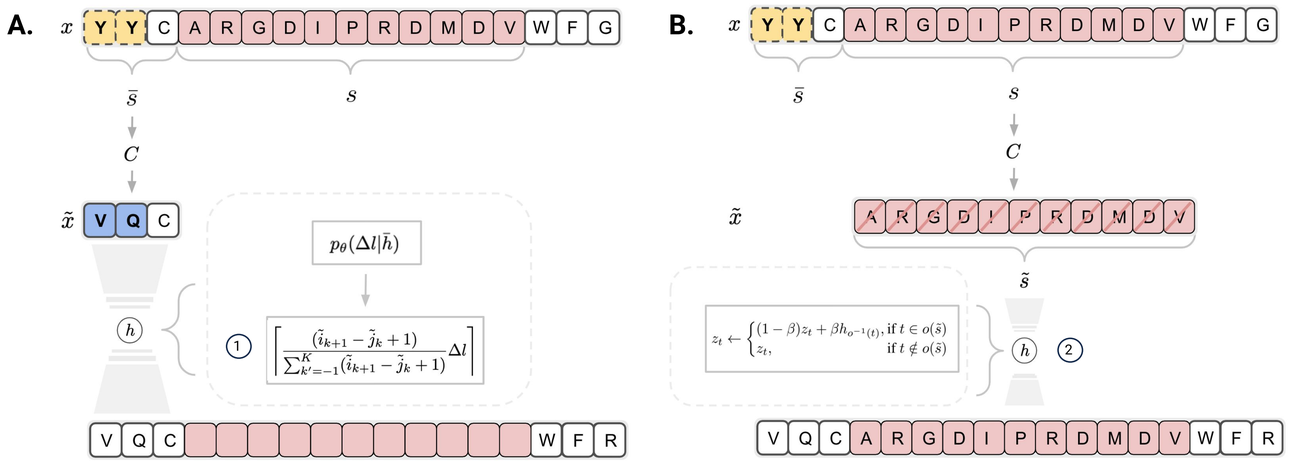
- Deep generative modeling for biological sequences presents challenges in reconciling the bias-variance trade-off between explicit biological insight and model flexibility.
- The deep manifold sampler was recently proposed as a means to iteratively sample variable-length protein sequences [1].
- For protein design, domain knowledge is often used to constrain combinatorial search space [2, 3].
- Significant challenges exist in explicitly incorporating this existing knowledge in an end-to-end learning and sampling procedure.



Summary

- The **deep manifold sampler** [1] consists of a denoising autoencoder (DAE) [2] that learns a manifold of protein sequences in a self-supervised manner and where sampling is done by iteratively denoising a sequence while exploiting the gradients from the function predictor.
- We introduce an approach called **multi-segment preserving sampling** which enables the inclusion of domain-specific knowledge by designating preserved and non-preserved segments along the input sequence, thereby restricting variation to only regions outside of the preserved segments.

Multi-segment preserving sampling

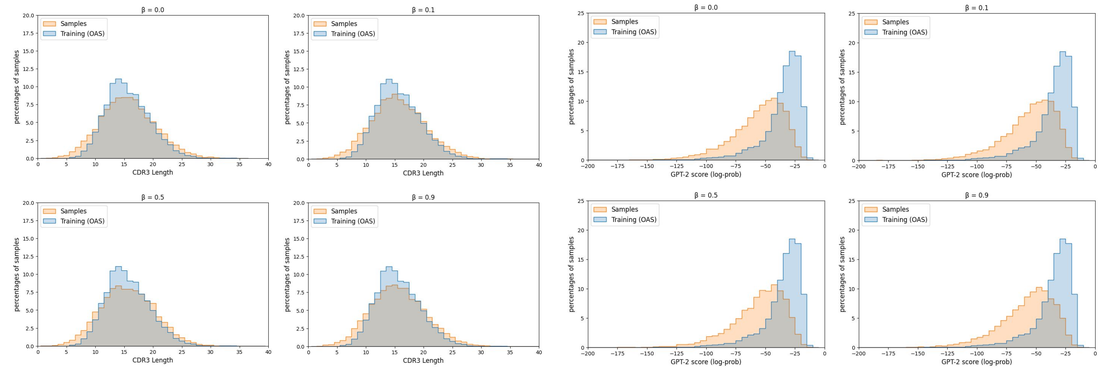


Illustrating multi-segment preserving sampling. **(A)** Non-preserved segments $\bar{s}$ are corrupted using corruption process C, for which a given token (yellow) may be randomly perturbed (blue). This is encoded as hidden vector set $h = (h_1, h_2, \ldots, h_{|\tilde{x}|})$. The length change predictor $p_\theta(\Delta l|\bar{h})$ takes in pooled, single-vector representation $\bar{h}$ and is trained to output $\Delta l^* = |\tilde{x}| - |x|$, which is distributed across $\bar{s}$ proportional to their original lengths (**Eq. 1**).

- **(B)** During the adaptive length conversion [5, 6], hyperparameter $\beta \in [0, 1]$ modulates the strength of the carry-over of hidden vectors. For example, when $\beta = 1$, we maintain the original hidden vector h_t as part of the conversion into hidden sequence z (**Eq 2**). The designated preserved-segment set s (red) remains unaltered throughout and is preserved during sampling.
- Non-autoregressive decoding [1, 5, 6, 7] yields a categorical distribution for a preserved segment's token, assigning entire probability mass to the original token's identity (**Eq 3.**) and forcing the sampled outcome to preserve the token identity:

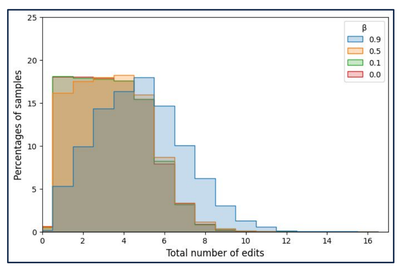
$$\tilde{y}_t^v \leftarrow \begin{cases}\infty, & \text{if } t \in o(\tilde{s}) \text{ and } v = \tilde{x}_{o^{-1}(t)} \\ -\infty, & \text{if } t \in o(\tilde{s}) \text{ and } v \neq \tilde{x}_{o^{-1}(t)} \\ \tilde{y}_t^v, & \text{if } t \notin o(\tilde{s})\end{cases}$$

Experiments



"Residue-preserving sampling" of *IGHV1-18*. All unique human antibody sequences with the *IGHV1-18* gene from the Observed Antibody Space [8] were used to sample exclusively from the CDR3 region using a deep manifold sampler while preserving all other residues, including framework regions.

- Training and sample distributions for CDR3 lengths **(top left)** were similar, with a moderate increase in sample diversity.
- Likewise, both distributions of normalized GPT-2 scores **(top right)** showed overlapping support and invariance to β.
- Mean of edit distance distributions between samples and seed sequences increases slightly with higher values of β **(bottom left)**; examples of aligned samples of CDR3 sequences under different settings of β **(bottom right)**.



β	Aligned CDR3 sequence	Edit distance
N/A (original)	ARDPEWDPF-QANY-YYYGMDV	0
0.0	ARDPEWDPF-QAN--YYYGMDV	3
0.1	ARDPEWDPFFQANYNYYYGMVD	3
0.5	KRDPEWDRF-QAPY-YTVGMDV	5
0.9	ARGPECDPH-QAV-DIYYGMDV	6

ICLR

References

[1] Gligorijević *et al.* (2021). Function-guided protein design by deep manifold sampling. bioRxiv:2021.12.22.473759

[2] Street and Mayo. (1999). Computational Protein Design. Structure, 7(5)

[3] Woolfson (2021). A brief history of *de novo* protein design: minimal, rational, and computational. Journal of Molecular Biology, 433(20)

[4] Vincent *et al.* (2008). Extracting and composing robust features with denoising autoencoders. International Conference on Machine Learning

[5] Lee *et al.* (2018). Deterministic non-autoregressive neural sequence modeling by iterative refinement. arXiv:1802.06901

[6] Shu *et al.* (2020). Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta prior Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34

[7] Gu *et al.* (2018). Non-autoregressive neural machine translation. International Conference on Learning Representations

[8] Olsen *et al.* (2022). Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Science, 31(1)