nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00787-2

Variational autoencoder for design of synthetic viral vector serotypes

Supplementary Notes

1. MLP-VAE Training and Sampling

MLP-VAE was trained on aligned hexon MSA calculated by ClustalOmega[1]. MSA training data was one-hot encoded. Different combination of learning rate ($5e^{-3}$ - $5e^{-5}$), weight decay ($1e^{-2}$, $1e^{-4}$) and dropout probability (0.3, 0.4) were tested (Supplementary Table 1). Model was trained with a NVIDIA V100 GPU with 32GB memory. To prevent overfitting, the training was stopped with earlystopping when validation cross-entropy loss has not improved in the last 250 epochs, and the checkpoint with the lowest validation perplexity was selected for evaluation. Model with highest test F1 score was used for generation (learning rate of $5e^{-5}$, weight decay of $1e^{-2}$ and dropout probability of 0.4). Only 40 unique sequences were found when decoding 1000 randomly sampled latent vectors.

2. ProtBert Fine-tuning and Sampling

ProtBert is a component of the ProteinVAE training pipeline. Therefore, it was fine-tuned to test if this component alone can generate hexons. 8 AMD MI50-32GB GPUs were in parallel, with a total effective batch size of 48. Learning rate from $2e^{-1}$ to $6e^{-6}$ was tested with weight decay from 0 to 10^{-5}(Supplementary Table 2). To prevent overfitting, the training was stopped with earlystopping when validation loss has not improved in the last 10 epochs, and the checkpoint with the lowest validation perplexity was selected for evaluation. The model with the lowest test perplexity was use for generation (learning rate: $2e^{-4}$, weight decay: 10^{-1}). Gibbs sampling was performed to generate sequences[2].Gibbs sampling is a Markov-chain Monte-Carlo (MCMC) sampler with a local proposal distribution, and it often becomes trapped in one mode of distribution[3]. To encourage sequence diversity, all natural sequences were used for sequence initialization, and 2 chains were sampled for each initialization. Despite 711 unique initializations, only 460 distinct sequences were generated, and they lack the sequence and structural characteristics inherent to natural hexon sequences. Additionally, Gibbs sampling was also slow due to the long sequence length, each sampling chain takes over 1.5 minutes to compute. Although higher sequence quality could potentially be obtained with Metropolis-Hastings (MH) sampling[4], the computational complexity is quadratic of the sequence length, which results in prohibitively long sampling time for long sequences like hexons, where each chain takes ~312 hours to sample.

3. Sequence Repetitiveness

In order to compare sequence repetition patterns across different positions throughout the sequence, a sliding window with a fixed length l was defined. For a sequence of length L, there are a total of $L - l + 1$possible positions for the start of window, each position can be normalized $norm\_pos = pos/L$. All windows were separated into 20 bins according to the normalized start of their start position. In each window, the number of repeated amino acids was counted. Mean repetition number and mean start of window position were plotted for all 20 bins.

4. Alphafold2 structure prediction

All structures were predicted with the Alphafold2.0.0. Due to limited computing resources, the reduced genetic database was used to save computation time. After folding a small number of natural hexons, it was observed that model_1 predictions ranked first for all sequences. To conserve computing time, only model_1 was used for all predictions reported here. For each structure, the genetic database search takes 20 min on 4 CPUs, and the modeling prediction takes 10 min on a A100 GPU with 40GB memory. Predicting all 711 Ad structures would take 948 CPU hours (39.5 days) and 118h of A100 GPU computing time (4.92 days). Prediction cannot be done on GPUs with memory of 32GB or less. In addition, no structural prediction of acceptable quality could be obtained with less resource-intensive language-model-based single-sequence structure prediction models, e.g. ESMFold[5], RGN2[6], OmegaFold[7].

5. Human AdV Hexon Classifier

All training sequences were labeled as human or non-human AdV hexon according to their fasta description. All training samples were converted to latent vectors with the previously trained encoder. A simple logistic regression classifier is trained to predict whether a training sequence is from human or non-human AdV from its latent vector. Learning rate is set to $5e^{-5}$. All samples previously generated from each cluster were encoded, and the trained classifier predicted whether they come from human AdV from their latent vectors.

6. Phylogenetic Analysis

Hexon sequences of generated sequences that are predicted to be human AdV and unique natural hexons with known serotype are aligned using ClustalOmega[1]. Maximum likelihood-based phylogenetic analysis was performed using PhyML 3.0[8] with BLOSUM62 substitution model , and support values were calculated using the Bayesian-like transformation of aLRT method (aBayes)[9]. Phylogenetic tree was visualized with iTOL[10].

7. Imputed Serotyping

Imputed serotyping was conducted as previously described[11]. All natural human AdV hexons with known serotypes were extracted and aligned with all generated sequences that are predicted to be human AdV hexon. Location of hexon loop1 and loop2 were identified according to the sequence reported[11]. Pairwise amino acid divergence was computed for all possible pairs in the MSA to distinguish if generated sequences are from any known human serotypes. It was reported that amino acid divergence higher than 4.2% in Loop1 and 2.1% in Loop2 to all previously reported serotypes supports that a new serotype is likely identified.

8. Simulation of ProT-VAE Results on Hexon Dataset

Given that ProT-VAE model has not been released, reconstruction F1 in ProT-VAE model was simulated by introducing a small Gaussian noise to the language model hidden state before decoding (Supplementary Figure 9). Since the ProtT5 model (48 M parameters) used in ProT-VAE was not publicly available, another larger ProtT5 (3 B parameters) was used. As a comparison, 5000 phenylalanine hydroxylase (PAH) sequences were found via psiBLAST with a human PAH variant (2PAH) as described in ProT-VAE. Owing to the high computational resources required by running the large T5 model, an equal size PAH dataset was randomly selected from the psiBLAST results for comparison (n = 711). When a small Gaussian noise (variance < 0.0625) was injected to the encoder output, the simulated reconstruction performance on the PAH dataset remained at a high level (F1 = 0.81). When tested on the hexon dataset (n = 711), the T5 model can faithfully decode all hexon sequences in the training dataset (F1 = 1) without noise. However, when a smaller level of Gaussian noise (variance < 0.04) was added, significantly worsened F1-score (0.19) was observed on the hexon dataset. Because of the significantly low performance with noise, it is likely that the ProT-VAE will not be able to generate high-quality hexon sequences using a ProtT5 pre-trained on sequences less than 512 amino acids.

Due to the model design in ProT-VAE, exchanging the PLM would not only require retraining the VAE part on family-specific sequences, but also retraining the generic CNN network on a large protein database which is computationally prohibitive. In comparison, accommodating another PLM in ProteinVAE only requires training the VAE component on sequences from a specific protein family. It is also unlikely that a pre-trained PLM decoder can be used directly to reconstruct hexon, because this small family is likely underrepresented in the pre-training database. Moreover, in (Sevgen et al., 2023), the diversity of sequences was only measured against one natural sequence. This could lead to overestimation of sequence novelty, because the generated sequences could be very different from the compared single natural sequences while sharing high identity with other natural sequences. Although, T5 model was previously demonstrated to be able to extrapolate up to another 600 tokens when trained on 512 tokens in natural language processing tasks (Press et al., 2021), it is unclear if ProT-VAE can be used to design longer proteins like hexon.

9. ReLSO Training and Sampling

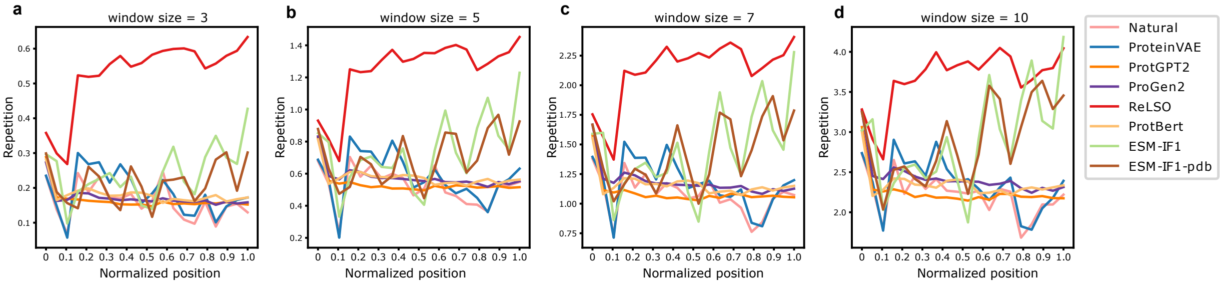
ReLSO model was developed to model sequences with a numerical label. To adapt it for training on the unlabelled hexon dataset, all loss terms related to regression were removed, leaving only the cross-entropy term and the spectral-norm penalty. Model was trained with a NVIDIA V100 GPU with 32GB memory. Different combination of learning rate ($2e^{-3}$ - $2e^{-5}$), weight decay (0 or 0.1), dropout rate (0.2 - 0.4), and number of transformer layers (5 or 10; 10M or 13M trainable parameters, respectively) were tested (Supplementary Table 3). To prevent overfitting, the training was stopped with earlystopping when validation loss has not improved in the last 250 epochs, and the checkpoint with the lowest validation perplexity was selected for evaluation. The model with the lowest test perplexity was use for generation (learning rate of $2e^{-4}$, weight decay of 0, dropout rate of 0.2, and number of transformer layers 5, 10M parameters). Since the ReLSO model is not a VAE model, and no method has been suggested for unconditional generation, 1,500 latent vectors were sampled from a Gaussian distribution with the posterior mean and standard deviation. Generated sequences were ranked according to perplexity, and the top $\frac{1}{7}$ sequences were used for comparison.

10. Ablation study

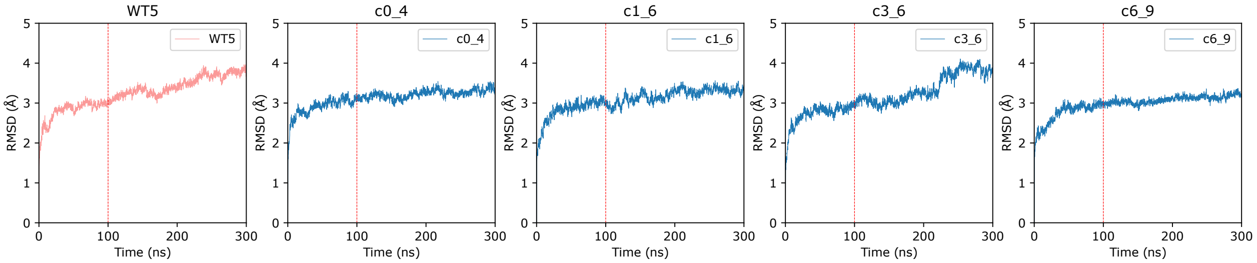
The goal of this study is to design a model that is intrinsically easy to sample where only latent variables are required during generation. This would also allow convenient sequence manipulation in latent space. Thus, instead of passing amino-acid-level embedding to the decoder, any encoder-decoder information flow other than the bottleneck vector was eliminated. Then, the generation from the bottleneck vector can be viewed as an upsampling process. Similarly, when provided with the same input vector at each time-step, traditional LSTM-based[12] sequence generation can also be viewed as an upsampling process. To test the hypothesis that non-autoregressive generation is more suitable in the case of protein design, deconvolution-based upsampling was also compared with the traditional LSTM method. After tuning the number of layers, processing direction and hidden dimension, the best performing LSTM-based model (14.0M parameters) only achieved a reconstruction accuracy of 0.18 on the validation set, while the deconvolution-based (12.4M parameters) model achieved 0.86. As an alternative to using LSTM as an upsampling method, training LSTM with additional per-time-step token information in a teacher forcing manner was experimented with, and this led to detrimental overfitting. No generation with the LSTM-based model was attempted, given the extremely low reconstruction accuracy.

To deepen our understanding of each component's contribution to the generation capacity, an ablation study was conducted by leaving out non-standard elements of the network one at a time. Reconstruction and generation performance was summarized for each version of the ablated model in Supplementary Table 4. The full model performs best overall. Using one-hot encoding instead of the ProtBert embedding resulted in worsened performance. ProtBert was also fine-tuned with the masked language model objective on the hexon dataset. However, using embeddings extracted by the fine-tuned ProtBert did not increase the performance any further. The encoder CNN module improved the reconstruction accuracy, which is likely due to the addition of local features that reflects detailed differences among sequences. The bottleneck attention module has more effect on improving generation, which is likely tied to its ability to produce a better protein level representation that complies with the sampling process[13]. Surprisingly, the amino acid attention improved generation in terms of secondary structure ratio and sequence variability profile, but not the amino acid usage. In addition, the secondary structure reweighting on the cross-entropy loss only improved amino acid usage and sequence variability profile slightly. This could be related to the error in the secondary structure prediction. The model does not benefit from larger number of parameters, and the current number of parameters cannot be compressed further without performance degradation. The ProtBert model alone cannot generate sequences with comparable quality (Supplementary Figure 1, 3-9).

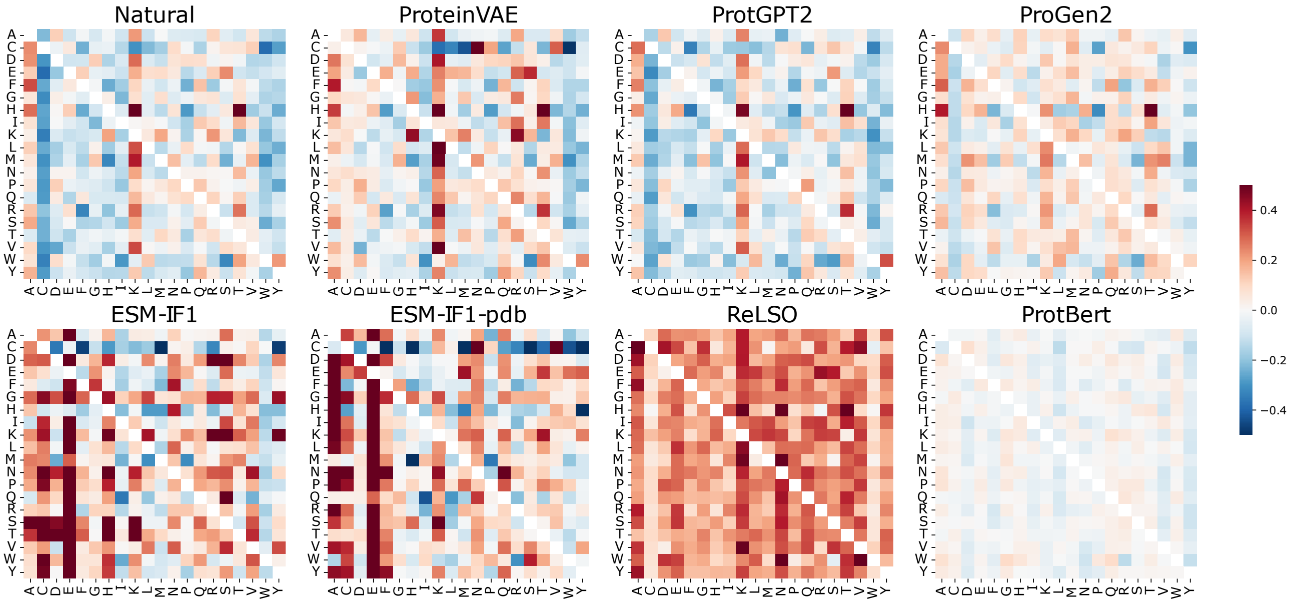
Supplementary Figures



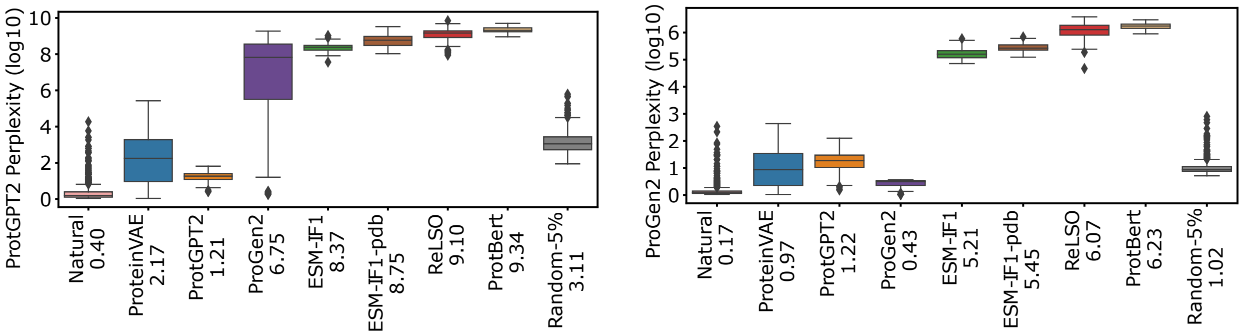
Supplementary Figure 1 (a - d): Repetition profile for natural hexons and sequences generated by all models with window size of 3, 5, 7, and 10, respectively.



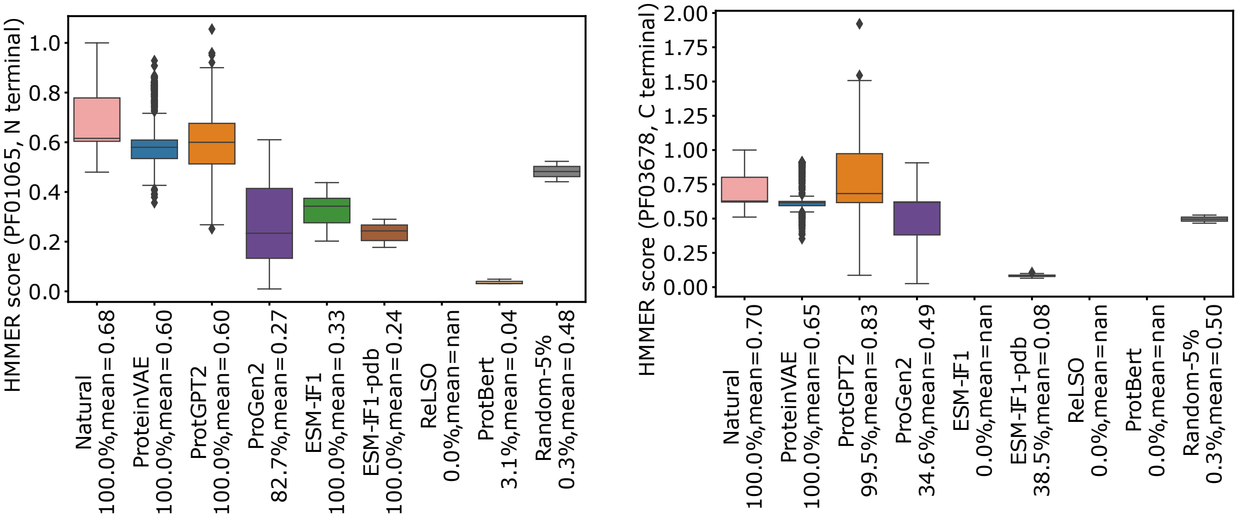
Supplementary Figure 2 RMSD for randomly selected natural (n = 1) and ProteinVAE generated sequences (n = 4) with extended simulation time of 300 ns. Dash line indicates simulation at 100 ns. As can be seen, RMSD only showed negligible changes past 100 ns. Therefore, all simulations analyzed in main results section were stopped at 100 ns.



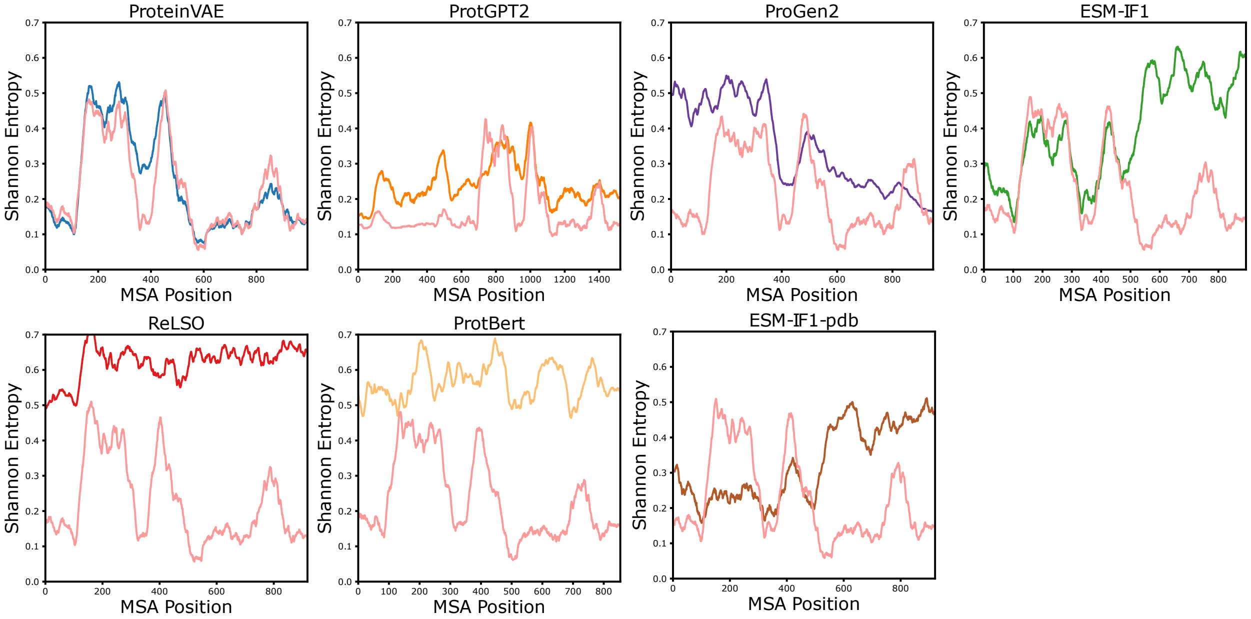
Supplementary Figure 3 Amino-acid pair association scores for sequences generated by all models. Negative values (blue) indicate shorter distances compared to random shuffled sequences.



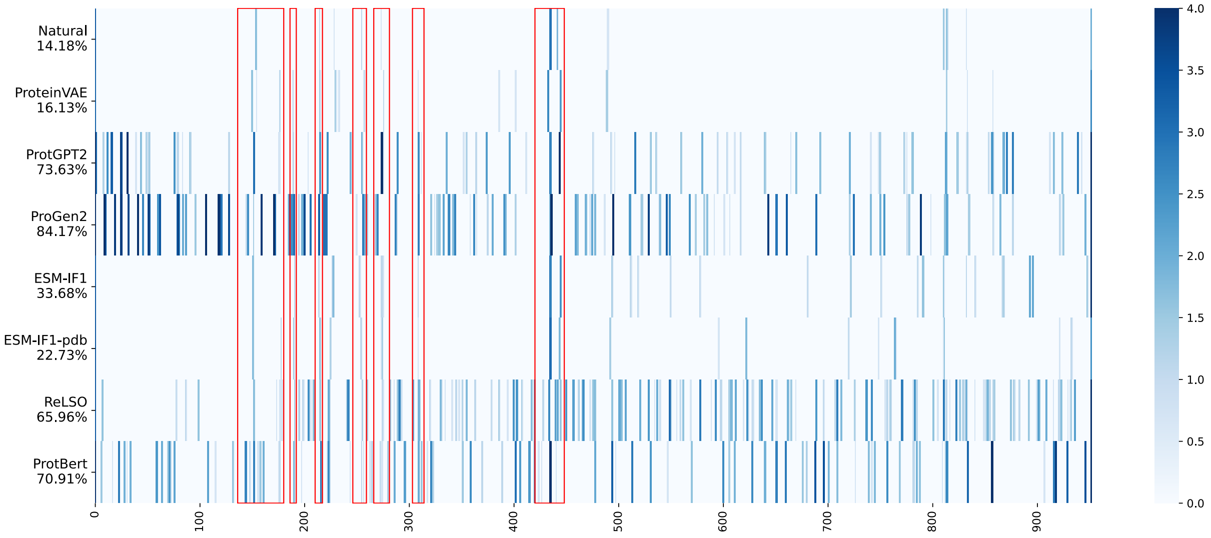
Supplementary Figure 4 Sequence perplexity (log10 transformed) from fine-tuned ProtGPT2 (left) and ProGen2(right) for sequences generated by all models. The number on the X-axis is the average perplexity for each group of sequences. Lower perplexity means better fit in the training data of natural hexons. All natural sequences were used for analysis (n = 711). For all models, the same-ratio of higher quality sequences were compared (ProteinVAE: n = 1000, all other models: n = 214). Each box-plot shows the first and third quartiles, central line is median, and whiskers show range of data with outliers displayed individually.



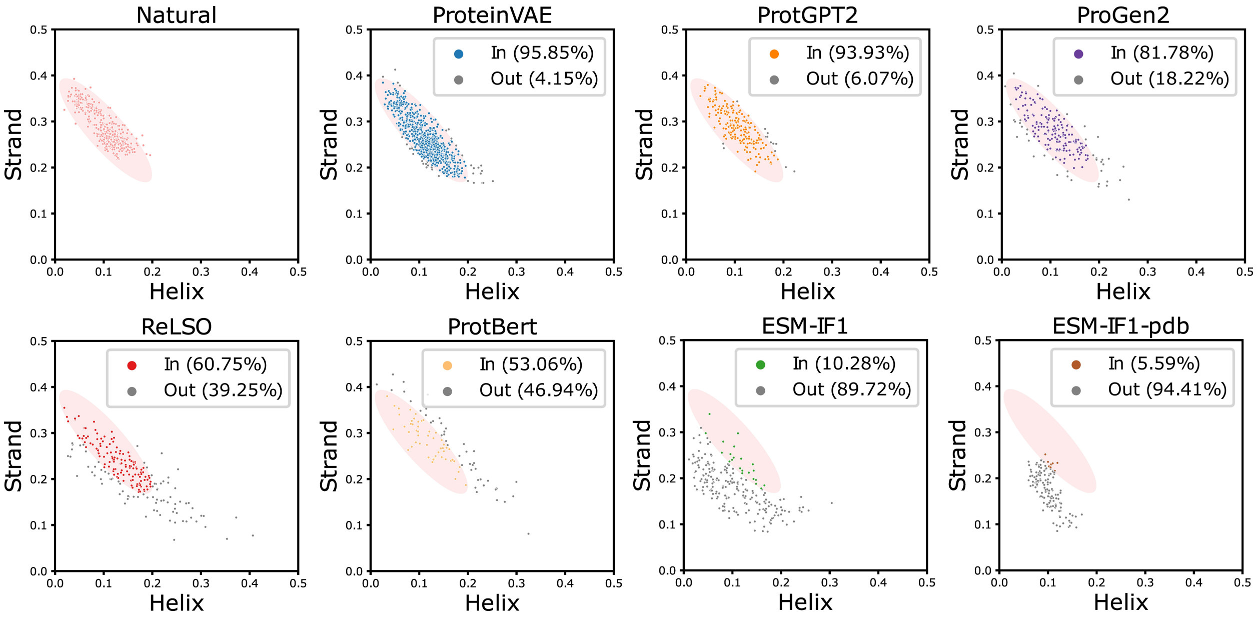
Supplementary Figure 5 HMMER score for the hexon N-terminal (left) and C-terminal (right) domain. The percentage of hits and the average score is labeled next to the model name. Scores were normalized by the highest score seen in natural sequences. Higher score means higher likelihood of a sequence containing a domain. All natural sequences were used for analysis (n = 711). For all models, the same-ratio of higher quality sequences were compared (ProteinVAE: n = 1000, all other models: n = 214). Each box-plot shows the first and third quartiles, central line is median, and whiskers show range of data with outliers displayed individually.



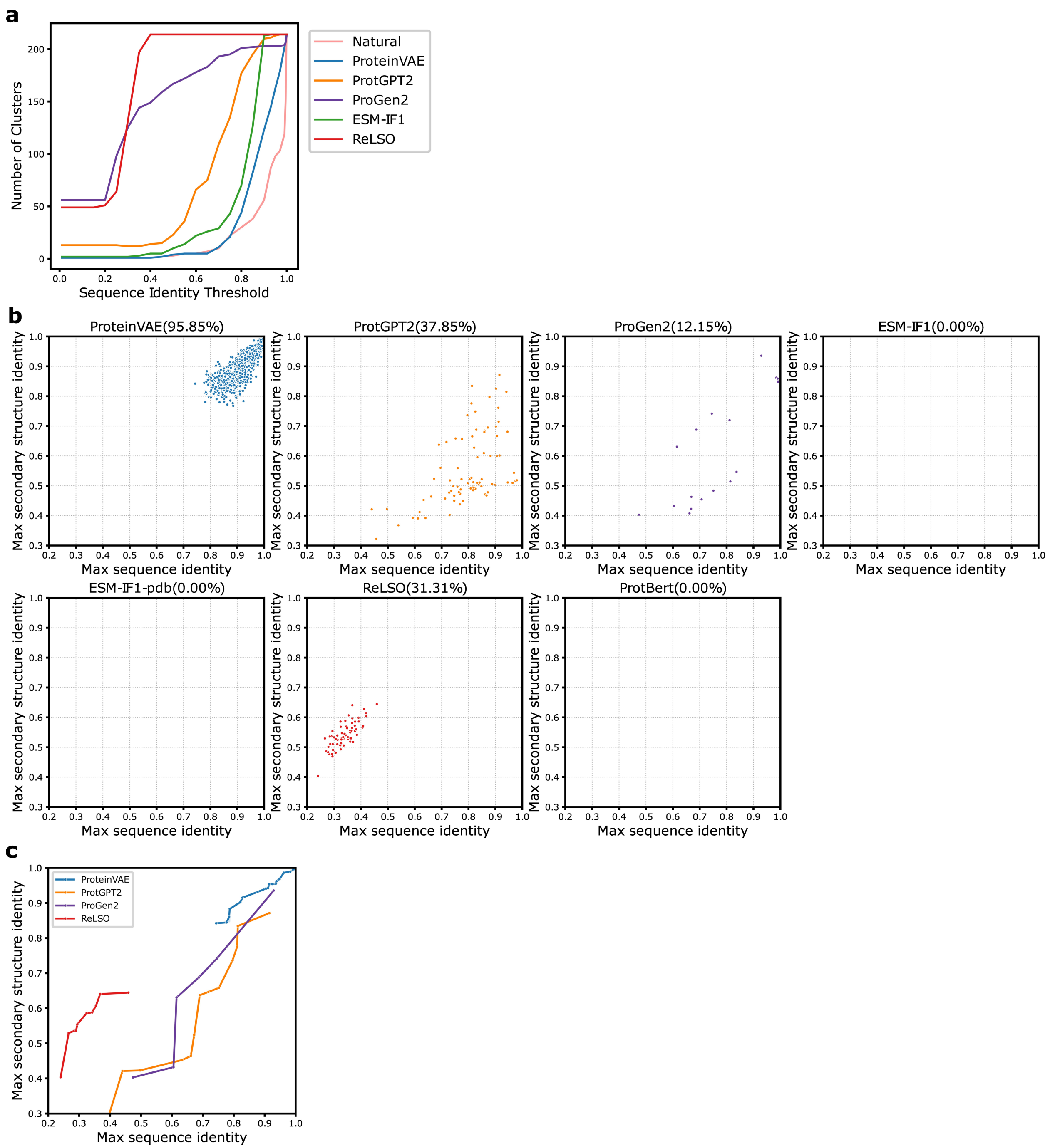
Supplementary Figure 6 Shannon-entropy for natural hexons and sequences generated by all models in MSA columns with above 20% occupancy in each dataset. Higher value reflects higher sequence variability across samples.



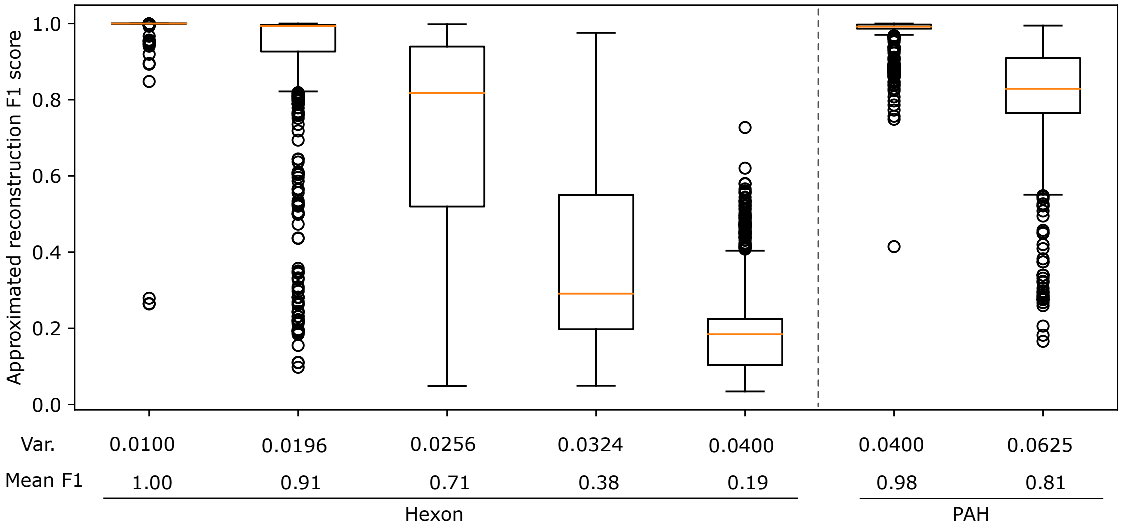
Supplementary Figure 7 Positions of invalid columns in MSA (less than 80% occupancy) in the reference sequence of human adenovirus serotype 5 hexon (P04133). Color indicates number of invalid columns (log transformed). Red squares show the location of hypervariable regions.



Supplementary Figure 8 Helix and strand ratio in natural and hexons generated by all models. Pink shade in all plots shows the area in the bi-variate normal distribution fitted on natural samples (α = 0.05). In generated sequence plots, gray points represent outliers, while colored points are sequences considered within the natural distribution.



Supplementary Figure 9 Comparing sequence diversity against sequence quality across all models **(a)** Number of clusters at different identity thresholds. **(b)** Scatter plot for sequence diversity and secondary structure similarity. X-axis is the maximum sequence identity on all aligned pairs. Y-axis is the maximum percentage identity of 3-state secondary structure on all aligned pairs of generated and natural sequences. Sequences closer to the top-left corner are ideal, as they are structurally similar to natural protein but more novel in sequence. **(c)** Pareto frontiers: The optimal sequences designed by each model are highlighted along respective frontier.



Supplementary Figure 10 Simulation of ProT-VAE result on hexon dataset. Approximated reconstruction F1 score is compared between the hexon dataset (n = 711) and a PAH dataset (n = 711) extracted the same way as described in ProT-VAE. Small Gaussian noises with different variance were introduced to the encoder output to simulate inaccuracy injected by the CNNs and VAE component in ProT-VAE. As can be seen, the simulated reconstruction F1 worsened even at very low variance for the hexon dataset, but the reconstruction performance was maintained even at higher variance in the PAH dataset. Each box-plot shows the first and third quartiles, central line is median, and whiskers show range of data with outliers displayed individually.

Supplementary Tables

Supplementary Table 1 MLP-VAE training. Test F1 is shown for each run with different learning rate, weight decay, and dropout probability. Checkpoint used for generation is highlighted in bold.

	Dropout Probability			
	0.3		**0.4**	
	Weight Decay		**Weight Decay**	
Learning Rate	**0.0001**	**0.01**	**0.0001**	**0.01**
0.0005	0.7829	0.7829	0.783	0.7829
0.005	0.7830	0.7827	0.7829	0.7830
5.00e-05	0.7827	0.7828	0.7827	**0.7831**

Supplementary Table 2 ProtBert fine-tuning. Test perplexity is shown for each run with different learning rate and weight decay. Checkpoint used for generation is highlighted in bold.

	Weight Decay					
Learning Rate	**0**	**1e-1**	**1e-2**	**1e-3**	**1e-4**	**1e-5**
2e-1	2.21e+6	2.93e+1	1.34e+3	4.35e+3	2.83e+6	3.47e+5
2e-2	18.77	18.77	18.78	18.77	18.76	18.76
2e-3	18.67	18.68	18.68	18.67	18.67	18.67
2e-4	1.14	**1.13**	1.15	1.15	1.15	1.15
2e-5	1.18	1.18	1.19	1.19	1.17	1.16
2e-6	1.18	1.18	1.18	1.18	1.18	1.18

Supplementary Table 3 ReLSO training. Test perplexity is shown for each run with different learning rate, weight decay, dropout probability, and number of transformer layer. Checkpoint used for generation is highlighted in bold. Model with 5 transformer layers has 10M parameter, and model with 10 transformer layers has 13M parameters.

	Weight Decay											
	0.1						**0.0**					
	Dropout Probability											
	0.4		**0.3**		**0.2**		**0.4**		**0.3**		**0.2**	
	Transformer Layer											
Learning Rate	**5**	**10**	**5**	**10**	**10**	**5**	**5**	**10**	**5**	**10**	**5**	**10**
2e-03	8.76	8.02	8.75	8.07	8.22	8.78	10.82	10.85	10.83	10.86	10.89	10.91
2e-04	5.18	5.36	5.18	5.35	5.35	5.19	2.54	2.45	2.27	2.33	**2.22**	2.28
2e-05	5.86	5.99	5.85	5.93	5.91	5.84	2.75	3.26	2.53	2.73	2.39	2.54

Supplementary Table 4 Ablation study illustrating model component contribution. Each setting was repeated with 3 different random seeds. **ECNN**: CNN feature vector extractor in encoder; **BA**: Autobot attention mechanism[13]; **AAA**: Attention mechanism across different amino-acid channel; **SSCE**: cross entropy reweighted to assigned 1.2x penalty to the strand positions; **Test F1**: reconstruction F1 score on test set; **JSD (Amino acid usage)**: Jesen-Shannon distance between amino acid usage frequency; **JSD (Secondary Structure)**: Jesen-Shannon distance between secondary structure ratio; **MSA Entropy Correlation**: Pearson correlation between MSA entropies in valid columns of generated and natural sequences. Bold numbers denote the best performancec in each respective column.

	ECNN	**BA**	**AAA**	**SSCE**	**Test F1**	**JSD (Amino acid usage)**	**JSD (Secondary Structure)**	**MSA Entropy Correlation**
Final (12.4M)	+	+	+	+	0.861	0.022	0.015	0.911
w/o encoder cnn	-	+	+	+	0.836	0.023	0.015	0.878
w/o bottleneck attn	+	-	+	+	**0.868**	0.039	0.038	0.500
w/o aa attn	+	+	-	+	0.847	**0.013**	0.024	0.669
w/o ss weighted ce	+	+	+	-	0.858	0.023	**0.012**	0.902
Base	-	+	-	-	0.790	0.015	0.044	0.436
Big (15M)	+	+	+	+	0.860	0.102	0.104	0.206
Medium (10.2M)	+	+	+	+	0.835	0.021	0.020	**0.913**
small (8.5M)	+	+	+	+	0.763	0.038	0.031	0.634
OneHot	+	+	+	+	0.822	0.092	0.079	0.280
Final+finetuned ProtBert	+	+	+	+	0.856	0.050	0.035	0.448
ProtBert	/	/	/	/	/	0.099	0.041	-0.014

Supplementary Table 5 ProtGPT2 fine-tuning. Test perplexity is shown for each run with different learning rate and weight decay. Checkpoint used for generation is highlighted in bold.

	Weight Decay						
Learning Rate	**0**	**1e-1**	**1e-2**	**1e-3**	**1e-4**	**1e-5**	**1e-6**
1e-1	7.93e+8	1.52e+3	1.61e+3	1.99e+10	3.37e+7	7.24e+10	1.37e+10
1e-2	114.46	104.86	93.67	113.68	120.64	95.47	114.46
1e-3	**2.24**	2.36	2.38	2.39	2.34	2.24	2.24
1e-4	2.77	2.95	2.97	2.70	2.77	2.77	2.77
1e-5	3.07	3.09	3.09	3.07	3.07	3.07	3.07
1e-6	10.41	10.40	10.41	10.41	10.41	10.41	10.41

Supplementary Table 6 ProtGen2 fine-tuning. Test perplexity is shown for each run with different learning rate and weight decay. Checkpoint used for generation is highlighted in bold.

	Weight Decay						
Learning Rate	**0**	**1e-1**	**1e-2**	**1e-3**	**1e-4**	**1e-5**	**1e-6**
6e-3	15.61	13.34	16.50	13.11	13.27	20.08	16.07
6e-4	1.34	13.31	1.36	1.47	**1.31**	1.34	1.34
6e-5	1.48	1.47	1.45	1.48	1.48	1.48	1.48
6e-6	20.81	22.59	21.05	22.92	24.87	20.44	24.10

Supplementary Table 7 ProtGen2 generation parameter tuning. Log likelihood is shown for each generation condition with different nucleaus sampling probability (top_p) and temperature. Final generation condition used in analysis is highlighted in bold.

	Temperature				
Top_p	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
0.7	-0.44	**-0.33**	-0.39	-0.49	-0.51
0.9	-0.38	-0.40	-0.46	-0.63	-0.78
1.0	-0.42	-0.45	-0.59	-0.80	-1.03

Supplementary Table 8 ESM-IF1 sampling temperature tuning with predicted structures as a template. Mean likelihood is shown to assess sequence quality, and mean number of unique 7-grams is shown to reflect changes in diversity. When temperature is reduced below 0.1, no substantial improvement in sequence quality is observed, while there is a drastic drop in sequence diversity. Therefore, sampling temperature for final generation is set to 0.1.

	Temperature						
	1e-06	**1e-05**	**1e-04**	**1e-03**	**1e-02**	**1e-01**	**1**
Mean likelihood	-0.74	-0.74	-0.74	-0.74	-0.73	-0.76	-1.56
Mean unique 7-gram	883.00	883.00	971.60	1205.20	3699.80	12099.20	44814.00

Supplementary Table 9 ESM-IF1 sampling temperature tuning with experimental structures as a template. Mean likelihood is shown to assess sequence quality, and mean number of unique 7-grams is shown to reflect changes in diversity. When temperature is reduced below 0.1, no substantial improvement in sequence quality is observed, while there is a drastic drop in sequence diversity. Therefore, sampling temperature for final generation is set to 0.1.

	Temperature						
	1e-06	**1e-05**	**1e-04**	**1e-03**	**1e-02**	**1e-01**	**1**
Mean likelihood	-0.86	-0.86	-0.86	-0.86	-0.86	-0.89	-1.90
Mean unique 7-gram	819.22	819.22	860.22	1086.78	3100.56	11259.00	44076.00

Supplementary References

1. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).

2. Wang, A. & Cho, K. BERT has a mouth, and it must speak: BERT as a Markov random field language model. *ArXiv Prepr. ArXiv190204094* (2019).

3. Tieleman, T. & Hinton, G. Using fast weights to improve persistent contrastive divergence. in *Proceedings of the 26th annual international conference on machine learning* 1033–1040 (2009).

4. Goyal, K., Dyer, C. & Berg-Kirkpatrick, T. Exposing the Implicit Energy Networks behind Masked Language Models via Metropolis–Hastings. *ArXiv Prepr. ArXiv210602736* (2021).

5. Lin, Z. *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv* **2022**, 500902 (2022).

6. Chowdhury, R. *et al.* Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.* **40**, 1617–1623 (2022).

7. Wu, R. *et al.* High-resolution de novo structure prediction from primary sequence. *BioRxiv* 2022–07 (2022).

8. Guindon, S., Delsuc, F., Dufayard, J.-F. & Gascuel, O. Estimating maximum likelihood phylogenies with PhyML. *Bioinforma. DNA Seq. Anal.* 113–137 (2009).

9. Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C. & Gascuel, O. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst. Biol.* **60**, 685–699 (2011).

10. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).

11. Madisch, I., Harste, G., Pommer, H. & Heim, A. Phylogenetic analysis of the main neutralization and hemagglutination determinants of all human adenovirus prototypes as a basis for molecular classification and taxonomy. *J. Virol.* **79**, 15265–15276 (2005).

12. Sundermeyer, M., Schlüter, R. & Ney, H. LSTM neural networks for language modeling. in *Thirteenth annual conference of the international speech communication association* (2012).

13. Montero, I., Pappas, N. & Smith, N. A. Sentence bottleneck autoencoders from transformer language models. *ArXiv Prepr. ArXiv210900055* (2021).