HOW WELL DO GENERATIVE PROTEIN MODELS GENERATE?

Han Spinner
Department of Systems Biology
Harvard Medical School

Aaron W. Kollasch
Department of Systems Biology
Harvard Medical School

Debora S. Marks
Department of Systems Biology
Harvard Medical School

ABSTRACT

Protein design relies critically on the generation of plausible sequences. Yet, the efficacy of many common model architectures from simple interpretable models, like position-specific scoring matrix (PSSM) and direct couplings analysis (DCA), to newer and less interpretable models, like variational autoencoders (VAEs), autoregressive large language models (AR-LLMs) and flow matching (FM), for sequence sampling remains uncertain. While some models offer unique sequence generation methods, issues such as mode collapse, generation of nonsensical repeats, and protein truncations persist. Trusted methods like Gibbs sampling are often preferred for their reliability, but can be computationally expensive. This paper addresses the need to evaluate the performance and limitations of different generation methods from protein models, considering dependencies on multiple sequence alignment (MSA) depth and available sequence diversity. We propose rigorous evaluation methods and metrics to assess sequence generation, aiming to guide design decisions and inform the development of future model and sampling techniques for protein design applications.

1 INTRODUCTION

Using machine learning to design proteins is useless unless we can generate plausible sequences, regardless of training data or model type. Many different approaches to protein design to achieve different goals have been quite successful (Shin et al., 2021; Madani et al., 2023; Lian et al., 2022; Hawkins-Hooker et al., 2021), and all of these projects have hinged on generating sequences that 'make sense'. In almost all protein engineering and protein design quests, we want to create proteins that fold and function. However, conditions that encourage stability, dynamic movements, tolerance to stressors, proper expression levels, etc, are often specific protein-to-protein and project-to-project. In order to increase efficacy of these studies we must ask the simple question: How well do generative protein models generate?

Newer model architectures, such as variational autoencoders (VAE) and autoregressive large language models (AR-LLM), have shown some promise for function and structure prediction Frazer et al. (2021); Hsu et al. (2022); Notin et al. (2022). And the importance of comparing to simpler, more interpretable models has also been noted Zhang et al. (2024). Often, However, there are no theoretical guarantees that models successful for structure or fitness predictions are also guaranteed to be better for sampling new sequences. These models' architectures enable unique ways of generating: for instance, sampling sequences from a learned latent space from a VAE and ancestral sampling for AR-LLMs. But often, there are malignancies that come from these generation methods that go ignored.

When drawing sequences from the latent space of a VAE, it is common to observe issues with the diversity of sequences that are generated: (a) mode collapse, (b) posterior collapse, (c) even distribution of sequence diversity across the length of the protein, and/or (d) low quality sequences that have mutated active site or other key residues (Figure 1a). Examples of malignancies from ancestral sampling from AR-LLMs include:

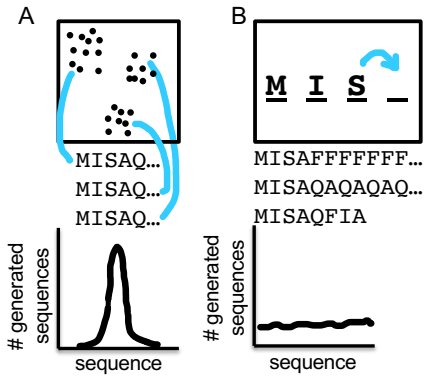

Figure 1: **Common Pathologies in VAE and AR-LLM Geneartion (a)** VAEs can suffer from mode collapse where variation captured in the latent space collapses into a single entity, and if sequences are pulled from various places they will either be fully identical or very similar. **(b)** AR-LLMs can suffer from nonsensical sequence generation where the sequences can have repeats or early termination.

(a) nonsensical and highly variable lengths, (b) repeats of single amino acids or small motifs, and/or (c) sequences completely out of distribution (Figure 1b). These issues have been noted in previous publications Hsu et al. (2022); Lucas et al. (2019), but are largely not explicitly interrogated in the field of protein design. Instead, tried and true methods like Gibbs sampling are frequently used to design proteins. Though, Gibbs sampling can become costly with calculating the probability of all mutations at each step of sampling; and if sampling one mutation at a time, it can ignore epistatic effects between multiple mutations. There are many sampling strategies that traverse all model types and several that are specific to different architectures (Figure 2). Where's the balance and what is optimal for different design tasks? Recently, an interrogation of generating GFP sequences from LLMs underscores the importance of compiling rigorous benchmarks for generating from the full suite of models at our fingertips Darmawan et al. (2023).

1.1 Benchmarked Proteins

First, we chose 9 example proteins from a benchmark of protein fitness models Notin et al. (2023) to serve as the explorative set (Table 1). These 9 proteins are a useful starting point for generation comparison because they all contain deep mutational scanning data with one and more mutations from wildtype Weng et al. (2022); Ding et al. (2023); Chen et al. (2023); Pokusaeva et al. (2019); Faure et al. (2022); Melamed et al. (2013); Gonzalez Somermeyer et al. (2022); Sarkisyan et al. (2016). We can therefore assess generation capacity with metrics like recall of top performing multi-mutants from the experimental datasets. Additionally, these proteins span diverse parts of the tree of life including humans, fungi, bacteria, and algae and they range in length with the longest protein being 724 residues and the shortest being 93 residues. In this way, we can ensure that results will not be unintentionally biased towards particular kinds of proteins. Finally, they have varying MSA depths as defined previously in ProteinGym Notin et al. (2023), spanning 610,129 sequences to 15 sequences. It is known that evolutionary model performance is dependent on the MSA depth used in training Hopf et al. (2017) and because of this, we want to evaluate model performance across a wide range of depths.

We want to interrogate the features and bugs of different generation methods from protein models in order to understand what models and sampling methods are better suited for different design tasks. Is the protein: (a) being enhanced for natural function, (b) switching specificity from substrate A to substrate B, (c) designed

Protein	Organism	Length	MSA Depth (N eff)
Phototropin (PHOT)	C. reinhardtii	118	610129
GTPase KRas (RASK)	H. sapiens	188	27851
Addiction module antidote protein (F7YBW8)	M. opportunism	93	16262
Imidazoleglycerol-phosphate dehydratase (HIS7)	S. cerevisiae	220	5191
Growth factor receptor-bound protein 2 (GRB2)	H. sapiens	217	1485
Polyadenylate-binding protein (PABP)	S. cerevisiae	577	855
Disks large homolog 4 (DLG4)	H. sapiens	724	354
Green-fluorescent protein 10 (D7PM05)	C. gregaria	235	138
Green fluorescent protein (GFP)	A. victoria	238	15

Table 1: Example proteins were chosen across the tree of life and spanning different evolutionary sequence depths. Each protein listed here has a deep mutational scan that explores ≥ 1 mutational depth from wild-type.

for a new-to-nature function (d) stabilized in certain pH or (e) salt conditions, or (f) needing a change in its isoelectric point? The list of possibilities is vast Notin et al. (2024) and the methods for assessing these qualities are lacking.

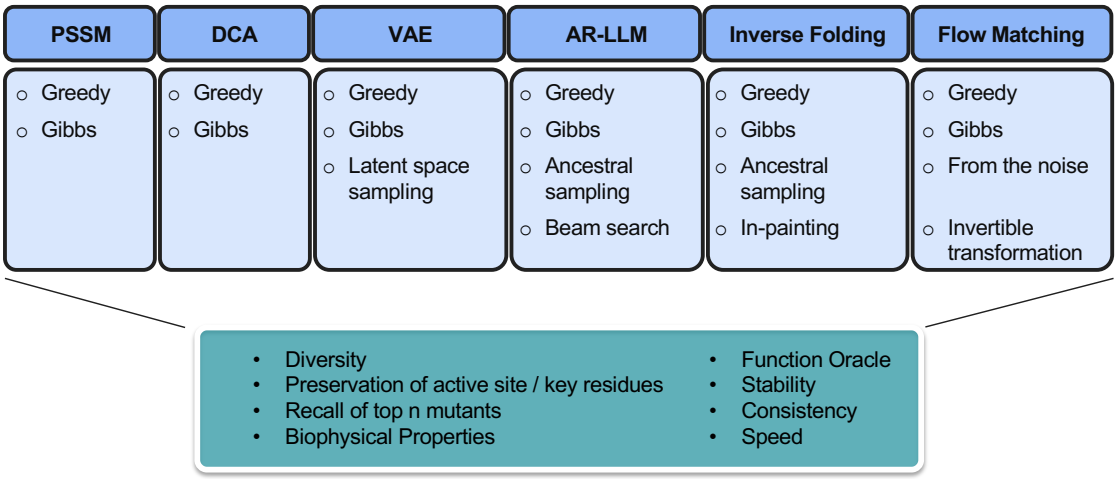

Figure 2: **Matrix of Model Types and Generation Methods.** Greedy and Gibbs sampling can be done with all model types whereas latent space and ancestral sampling are architecture-enabled.

2 *In silico* METRICS

We propose a set of *in silico* benchmark methods and metrics for evaluating sequence generation from a suite of model architectures and sampling methods (Figure 2) Marks et al. (2011); Frazer et al. (2021);

Hawkins-Hooker et al. (2021); Dauparas et al. (2022); Lipman et al. (2022). Ideally, in a protein generation framework, engineerable features are tunable. We want to have a quantified understanding of:

- Diversity metrics
  - Site-wise column entropy
  - Hamming distance from wildtype
  - Hamming distance from any known protein
  - Hamming distance from other generated sequences
- Perplexity
- Repeats
  - Longest stretch of single amino acid repeats
  - Longest stretch of double amino acid repeats
- Preservation of key residues
- Recall of top n mutants from experimental fitness measurements
- Biophysical properties
  - Molecular weight
  - Aromaticity
  - Isoelectric point
  - Molecular extinction coefficient (oxidized and reduced)
  - Percent Cysteine
- Function
  - ESM oracle Meier et al. (2021)
- Stability
  - Instability index Guruprasad et al. (1990)
  - EVcouplings delta hamiltonians Marks et al. (2011)
  - ESM-IF likelihoods Hsu et al. (2022)
  - pLDDT Pak et al. (2023)
- Consistency of model generation
  - Comparing sequences generated from different random seeds
- Scalability
- Training cost
- Inference cost
- Speed
- Resource consumption

3 RESULTS

We generated several thousand protein sequences from the proteins listed in Table 1 and calculated the diversity and biophysical metrics listed above Chaudhury et al. (2010); Cock et al. (2009).

Preliminary experiments suggest that different models are better equipped to generate sequences with different properties. When comparing simple Gibbs sampling across PSSM, DCA Marks et al. (2011), and VAE

Frazer et al. (2021), we observe two obvious differences: Generating sequences from the VAE gives consistent column entropy across the length of the protein whereas PSSM and DCA generate different levels of entropy in different parts of the protein (Figure 3a). The VAE also tends to generate sequences with a higher molecular weight, suggesting it is exchanging side chains in different ways than PSSM or DCA (Figure 3b). While specific examples are depicted in the figure, these trends were general across all proteins tested and across all temperatures of generation.

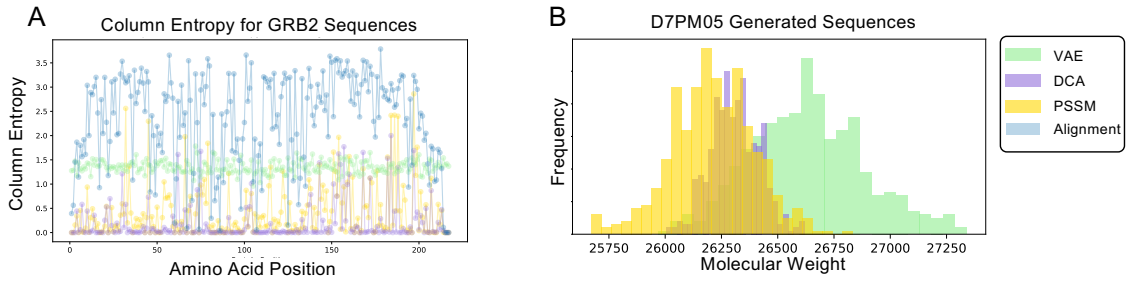

Figure 3: **Different Model Architectures Generate Sequences with Different Qualities (a)** Greedy and Gibbs sampling can be done with all model types whereas latent space and ancestral sampling are architecture-enabled.

These criteria will be crucial in choices of design decisions and will (a) help inform protein designers what needs to be checked before beginning experiments and (b) the next generation of model/sampling methods optimizing their generative capacity.

4 Future Directions

In the future, this work will cover the full extent of evaluation metrics outlined above. We will generate thousands of sequences from each of these model/sampling combinations and compare across all metrics listed. We are also going to release all of the code and create an easy-to-use interface such that people can upload their generated sequences, alignment, and any experimental data and get automatic analysis of their proposed sequences. Additionally, we are thinking about integrating this into the ProteinGym framework to enable end-to-end prediction and design for any protein of interest.

Acknowledgments

D.S.M., A.W.K. and H.S. are supported by a Chan Zuckerberg Initiative Award (Neurodegeneration Challenge Network, CZI2018-191853) and a NIH Transformational Research Award (TR01 1R01CA260415).

References

Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J Gray. PyRosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. *Bioinformatics*, 26(5):689–691, March 2010.

Yongcan Chen, Ruyun Hu, Keyi Li, Yating Zhang, Lihao Fu, Jianzhi Zhang, and Tong Si. Deep mutational scanning of an Oxygen-Independent fluorescent protein CreiLOV for comprehensive profiling of mutational and epistatic effects. *ACS Synth. Biol.*, 12(5):1461–1473, May 2023.

Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

Jeremie Theddy Darmawan, Yarin Gal, and Pascal Notin. Sampling protein language models for functional protein design. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

J Dauparas, I Anishchenko, N Bennett, H Bai, R J Ragotte, L F Milles, B I M Wicky, A Courbet, R J de Haas, N Bethel, P J Y Leung, T F Huddy, S Pellock, D Tischer, F Chan, B Koepnick, H Nguyen, A Kang, B Sankaran, A K Bera, N P King, and D Baker. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022.

David Ding, Ada Shaw, Sam Sinai, Nathan Rollins, Noam Prywes, David F Savage, Michael T Laub, and Debora S Marks. Protein design using structure-based residue preferences. June 2023.

Andre J Faure, Júlia Domingo, Jörn M Schmiedel, Cristina Hidalgo-Carcedo, Guillaume Diss, and Ben Lehner. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature*, 604(7904): 175–183, April 2022.

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, November 2021.

Louisa Gonzalez Somermeyer, Aubin Fleiss, Alexander S Mishin, Nina G Bozhanova, Anna A Igolkina, Jens Meiler, Maria-Elisenda Alaball Pujol, Ekaterina V Putintseva, Karen S Sarkisyan, and Fyodor A Kondrashov. Heterogeneity of the GFP fitness landscape and data-driven protein design. *Elife*, 11, May 2022.

K Guruprasad, B V Reddy, and M W Pandit. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.*, 4(2):155–161, December 1990.

Alex Hawkins-Hooker, Florence Depardieu, Sebastien Baur, Guillaume Couairon, Arthur Chen, and David Bikard. Generating functional protein variants with variational autoencoders. *PLoS Comput. Biol.*, 17(2): e1008736, February 2021.

Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta P I Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*, 35(2): 128–135, February 2017.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. April 2022.

Xinran Lian, Niksa Praljak, Subu K Subramanian, Sarah Wasinger, Rama Ranganathan, and Andrew L Ferguson. Deep learning-enabled design of synthetic orthologs of a signaling protein. December 2022.

Yaron Lipman, Ricky T Q Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. October 2022.

James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. Understanding posterior collapse in generative latent variable models. April 2019.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, James S Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.*, January 2023.

Debora S Marks, Lucy J Colwell, Robert Sheridan, Thomas A Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, 6 (12):e28766, December 2011.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. November 2021.

Daniel Melamed, David L Young, Caitlin E Gamble, Christina R Miller, and Stanley Fields. Deep mutational scanning of an RRM domain of the saccharomyces cerevisiae poly(a)-binding protein. *RNA*, 19(11):1537–1551, November 2013.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16990–17017. PMLR, 2022.

Pascal Notin, Aaron W Kollasch, Daniel Ritter, Lood Van Niekerk, Steffan Paul, Han Spinner, Nathan J Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Rose Orenbuch, Yarin Gal, and Debora Susan Marks. ProteinGym: Large-Scale benchmarks for protein fitness prediction and design. November 2023.

Pascal Notin, Nathan Rollins, Yarin Gal, Chris Sander, and Debora Marks. Machine learning for functional protein design. *Nat. Biotechnol.*, 42(2):216–228, February 2024.

Marina A Pak, Karina A Markhieva, Mariia S Novikova, Dmitry S Petrov, Ilya S Vorobyev, Ekaterina S Maksimova, Fyodor A Kondrashov, and Dmitry N Ivankov. Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLoS One*, 18(3):e0282689, March 2023.

Victoria O Pokusaeva, Dinara R Usmanova, Ekaterina V Putintseva, Lorena Espinar, Karen S Sarkisyan, Alexander S Mishin, Natalya S Bogatyreva, Dmitry N Ivankov, Arseniy V Akopyan, Sergey Ya Avvakumov, Inna S Povolotskaya, Guillaume J Filion, Lucas B Carey, and Fyodor A Kondrashov. An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS Genet.*, 15(4):e1008079, April 2019.

Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, Natalya S Bogatyreva, Peter K Vlasov, Evgeny S Egorov, Maria D Logacheva, Alexey S Kondrashov, Dmitry M Chudakov, Ekaterina V Putintseva, Ilgar Z Mamedov, Dan S Tawfik, Konstantin A Lukyanov, and Fyodor A Kondrashov. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603): 397–401, May 2016.

Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.*, 12(1):2403, April 2021.

Chenchun Weng, Andre J Faure, and Ben Lehner. The energetic and allosteric landscape for KRAS inhibition. December 2022.

Zhidian Zhang, Hannah K Wayment-Steele, Garyk Brixi, Haobo Wang, Matteo Dal Peraro, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. January 2024.

A Appendix

A.1 Equations

A.1.1 General Gibbs Sampling

$$P(x_i = a | x_{\neg i}, \text{model}) \propto e^{\text{log\_prob}(a | x_{\neg i}, \text{model})}$$

Where:

- x_i represents the amino acid at position i in the protein sequence.
- $x_{\neg i}$ represents all other amino acids except the one at position i.
- model is the probabilistic model used to generate log probabilities for amino acid mutations.
- $\text{log\_prob}(a | x_{\neg i}, \text{model})$ is the log probability of amino acid a at position i given the surrounding sequence context $x_{\neg i}$ according to the model.
- $P(x_i = a | x_{\neg i}, \text{model})$ represents the conditional probability of amino acid a at position i given the surrounding sequence context $x_{\neg i}$ and the model.

A.1.2 Column Entropy

$$\text{Column Entropy} = -\sum_{i=1}^{n} p_i \log(p_i)$$

Where:

- p_i is the frequency of amino acid i in the column.
- n is the number of unique amino acids in the column.

A.1.3 Molecular Weight

$$\text{Molecular Weight} = \sum_{i=1}^{n} (\text{weight}_i \times \text{count}_i)$$

Where:

- weight_i is the molecular weight of the i-th amino acid.
- count_i is the number of occurrences of the i-th amino acid in the protein sequence.
- n is the total number of different amino acids in the sequence.