Extrapolative benchmarking of model-based discrete sampling methods for RNA design

Joseph D. Valencia and David A. Hendrix

Oct 6, 2023

Abstract

Many problems in biomolecular design can be framed as sampling a biological sequence predicted by a computational model to satisfy some property of interest. This paradigm must overcome two major difficulties: (1) efficiently searching a discrete input space which is exponentially large in sequence length, and (2) preventing the sampling of out-of-distribution inputs on which the model makes inaccurate predictions. Consider the example of optimizing a candidate mRNA therapeutic given sequence $\rightarrow$ property predictors for RNA stability and translational efficiency. The number of possible nucleic acids of length L approaches the number of atoms in the universe at around $L = 130$, but any realistic model will have been trained on a mere fraction of sequence space. The computational genomics community has had success in using neural network backpropagation to efficiently probe learned sequence features [1, 2], and recent efforts [3] similarly exploit the differentiable nature of deep models to actively redesign sequences. We are working to benchmark approaches to differentiable model-based sampling on a variety of RNA design tasks, incorporating uncertainty estimation and extrapolative evaluations as partial solutions to the second challenge of generalization.

The chosen search algorithms survey recent work on probabilistic reparameterization (PR) of discrete input spaces [3, 4] and gradient-guided Markov-Chain Monte Carlo (MCMC) [5, 6]. Both classes of algorithms use automatic differentiation of neural network models $f(x)$ to efficiently compute input gradients $\nabla_x f(x)$ and do not require evaluating the models on infeasible relaxations of sequences. Beyond that, the methods differ in their treatment of input gradients – PR uses $\nabla_x f(x)$ to update a persistent set of continuous parameters θ for a categorical distribution $p(x|\theta)$, while the selected MCMC approaches use the gradients as transient biases to guide discontinuous jumps in sample space. PR has proven effective in diverse nucleic acid design tasks, and [3] presents a starting place for our work. However, low-variance gradient estimators [7, 8] which use sample gradients to augment the classical score function estimator $\nabla_\theta \mathbb{E}_{x\sim p(x|\theta)}[f(x)] = \mathbb{E}_{x\sim p(x|\theta)}[f(x)\nabla_\theta \log p(x|\theta)]$ could outperform the heuristic straight-through estimators (e.g $\nabla_\theta f(x) \approx \nabla_x f(x)$) that have been used so far. Gradient-guided MCMC has begun to be applied to protein design [9, 10], but has yet to be widely adopted in nucleic acid design, where its capacity to draw diverse samples would be similarly useful.

Using data from prior work, we have trained simple convolutional neural network (CNN) models to predict ribosomal occupancy of 5' UTRs [11], degradation properties of mRNAs [12], and

activity of RNA toehold switches [13]. To evaluate the extrapolation capabilities of each design algorithm, we can exclude the examples with the highest property values from our training process. Each model can be equipped with an uncertainty estimator using evidential deep learning [14] in order to optimize inputs for high property values and low uncertainty using each discrete search algorithm.

We are developing a general-purpose PyTorch library for wrapping arbitrary sequence models for differentiable sampling. We have so far implemented four PR methods, including three straight through estimators - the original (STE), the softmax/slope-annealed (STE-SM), the Gumbel-softmax (STE-GS), and the basic score function estimator, also known as REINFORCE. These algorithms can be run with or without instance normalization of the θ parameters, as suggested by [3]. We also have a working implementation of Gibbs-with-Gradients [5]. DMALA [6] is another promising MCMC method that approximates Langevin dynamics over discrete distributions for fast mixing, and we are currently experimenting with adaptive preconditioning schemes to make this effective in practice. Fig 1 depicts a small-scale experiment on a model predicting the mean ribosome load of a 5' UTR. We trained this model with evidential regression, excluding the highest 10% of mean ribosome load values from the trianing set. We optimized the lower confidence bound $\mu - \sigma^2$, a conservative estimate made by the network. The red line denotes the score of the initial sample, a random 50 nt sequence, and the black line indicates the highest mean ribosome load in the training data.

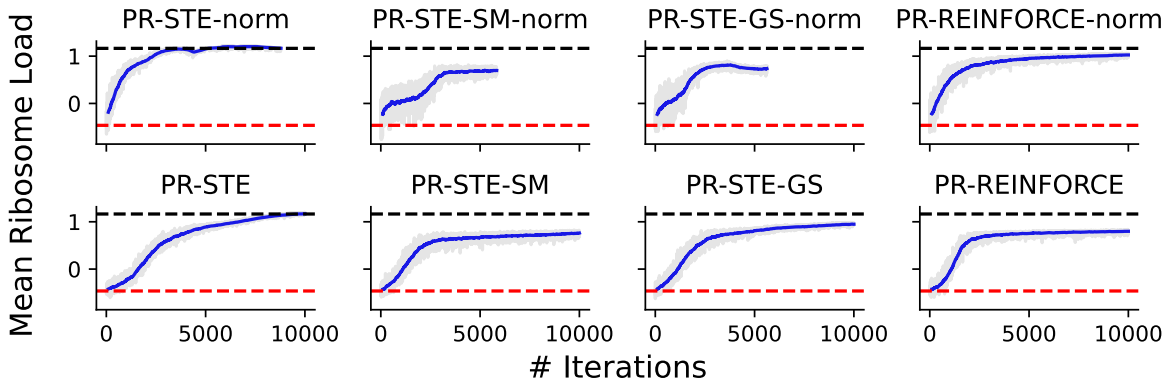

Figure 1: A typical optimization trajectory from a random 50 nt sequence.

Such preliminary results suggest that even without the advanced variance reduction techniques typically used alongside the score function estimator, the naive formula can perform nearly on par with STE. This motivates us to prioritize adding two recently developed gradient estimators [7, 8] which combine REINFORCE with Taylor approximations over a sequence neighborhood, possibly combining the benefits of the straight through and score function estimators. For a fair comparison, we will allow for multiple parallel Markov chains to equalize the computational budget with PR, which can draw and differentiate multiple samples at each iteration. We will also assess the sample diversity of each algorithm. Finally, we intend to score candidate designs using the original pretrained models, which learned using the full range of property values, as a better proxy for the ability to extrapolate in wet lab experiments.

References

1. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv:1704.02685 [cs].* arXiv: 1704.02685. http://arxiv.org/abs/1704.02685 (2020) (Oct. 2019).
2. Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W. & Mostafavi, S. Obtaining genetics insights from deep learning via explainable artificial intelligence. en. *Nature Reviews Genetics.* Publisher: Nature Publishing Group, 1–13. ISSN: 1471-0064. https://www.nature.com/articles/s41576-022-00532-2 (2022) (Oct. 2022).
3. Linder, J. & Seelig, G. Fast activation maximization for molecular sequence design. *BMC Bioinformatics* **22,** 510. ISSN: 1471-2105. https://doi.org/10.1186/s12859-021-04437-5 (2022) (Oct. 2021).
4. Daulton, S. *et al.* Bayesian Optimization over Discrete and Mixed Spaces via Probabilistic Reparameterization. en. *Advances in Neural Information Processing Systems* **35,** 12760–12774. https://proceedings.neurips.cc/paper_files/paper/2022/hash/531230cfac80c65017ad0f85d3031edc-Abstract-Conference.html (2023) (Dec. 2022).
5. Grathwohl, W., Swersky, K., Hashemi, M., Duvenaud, D. & Maddison, C. *Oops I Took A Gradient: Scalable Sampling for Discrete Distributions* en. in *Proceedings of the 38th International Conference on Machine Learning* ISSN: 2640-3498 (PMLR, July 2021), 3831–3841. https://proceedings.mlr.press/v139/grathwohl21a.html (2022).
6. Zhang, R., Liu, X. & Liu, Q. *A Langevin-like Sampler for Discrete Distributions* arXiv:2206.09914 [cs, stat]. June 2022. http://arxiv.org/abs/2206.09914 (2022).
7. Shi, J., Zhou, Y., Hwang, J., Titsias, M. K. & Mackey, L. Gradient Estimation with Discrete Stein Operators. en.
8. Lin, C.-H., Wu, S.-R., Lee, H.-Y. & Chen, Y.-N. *TaylorGAN: Neighbor-Augmented Policy Update for Sample-Efficient Natural Language Generation* arXiv:2011.13527 [cs]. Nov. 2020. http://arxiv.org/abs/2011.13527 (2023).
9. Emami, P., Perreault, A., Law, J., Biagioni, D. & John, P. S. Plug & play directed evolution of proteins with gradient-based discrete MCMC. en. *Machine Learning: Science and Technology* **4.** Publisher: IOP Publishing, 025014. ISSN: 2632-2153. https://dx.doi.org/10.1088/2632-2153/accacd (2023) (Apr. 2023).
10. Kirjner, A. *et al. Optimizing protein fitness using Gibbs sampling with Graph-based Smoothing* arXiv:2307.00494 [cs, q-bio, stat]. July 2023. http://arxiv.org/abs/2307.00494 (2023).
11. Sample, P. J. *et al.* Human 5' UTR design and variant effect prediction from a massively parallel translation assay. en. *Nature Biotechnology* **37,** 803–809. ISSN: 1087-0156, 1546-1696. http://www.nature.com/articles/s41587-019-0164-5 (2023) (July 2019).
12. He, S., Gao, B., Sabnis, R. & Sun, Q. RNAdegformer: accurate prediction of mRNA degradation at nucleotide resolution with deep learning. *Briefings in Bioinformatics* **24,** bbac581. ISSN: 1477-4054. https://doi.org/10.1093/bib/bbac581 (2023) (Jan. 2023).

13. Valeri, J. A. *et al.* Sequence-to-function deep learning frameworks for engineered riboregulators. en. *Nature Communications* **11.** Number: 1 Publisher: Nature Publishing Group, 5058. ISSN: 2041-1723. https://www.nature.com/articles/s41467-020-18676-2 (2023) (Oct. 2020).

14. Amini, A., Schwarting, W., Soleimany, A. & Rus, D. *Deep Evidential Regression* tech. rep. arXiv:1910.02600. arXiv:1910.02600 [cs, stat] type: article (arXiv, Nov. 2020). http://arxiv.org/abs/1910.02600 (2022).