Attention is not enough*

Hassan Hassan[a*], Kyle Puhger[b], Ali Saadat[c], Alexander Chen[d] and Maximilian Sprang[e*,*]

[a*]*DeOxy Tech, Manchester, UK*

[b]*University of California, Davis, California*

[c]*School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*

[d]*Johns Hopkins University, Baltimore, Maryland*

[e*]*Department of Biology, Institute of Qualtitative and Computational Biology, Johannes Gutenberg University Mainz, Mainz, Germany*

Abstract

Recent years have seen a flurry of generative nucleotide models, mostly of limited utility. In this short paper, we extend the theoretical unification of ecological and evolutionary change by Duthie & Luque to the the problem of synthetic DNA models. Through this extension, we provide, from first principles, methods for training improved models, grouping species as well as creating a road map to scale.

Keywords: Synthetic Biology, DNA

1. Introduction

DNA/RNA molecules are the physical substrate for which life carries information. They allow for the exploration of vast combinatorial spaces and in doing so, form the backbone of evolution as a driver of change. Learning how the subcomponents of these molecules are arranged to achieve specific goals (e.g. viral replication in mammalian cells) is a key problem of synthetic biology [1].

To that end, D&L[2] have proposed a unified equation of biological evolution and population ecology. In their paper, they state, that individuals give rise to new individuals through birth such that β_i is the number of births attributable to individual i. Individuals are removed from the population through death such that δ_i is an indicator variable that takes a value of 1 (death of i or 0 (persistence of i)). All individuals are defined by some characteristic z_i, and Δz_i defines any change in z_i from one time step t to the next $(t+1)$. The total number of individuals in the

*Corresponding author. E-mail address: masprang@uni-mainz.de

population at t is N. From this foundation, we can define Ω to be a summed characteristic across N entities:

$$\Omega = \sum_{i=1}^{N} (\beta_i - \delta_i + 1) * (z_i + \Delta z_i) \tag{1}$$

As a result of the above equation, since the rates of deaths and births are ultimately a function of the optimality of an individuals traits at a particular niche, represented by x, the sum of a genomes characteristics simplifies to:

$$\Omega(x) = f(x) + \varepsilon\sigma_\Omega \tag{2}$$

Where f is a function that selects for optimal traits and $\varepsilon\sigma_\Omega$, non-fatal variations. As Ω is a summation of characteristics, we expect the cumulative summation of the nucleotides of trait-relevant genes to be fairly stable at ecological niche points. This is because all biological traits (in the limit) are determined by nucleotides

Focusing on DNA/RNA as the backbone of biology, many projects have aimed to learn this function using different architectures and datasets, all ultimately aiming to scale to foundational level capabilities ala Claude, GPT-4 etc. Benegas et all [3] provide a good overview of the current state-of-art for Genomic Language Models.

The key goals of this paper is to hopefully answer two questions:

1. What is Ω?
2. What are the limits of the current paradigm with respect to learning f?

We hypothesize that the overall structure of Ω should be detectable and f learnable. The structure of Ω, in the context of genomic modeling, does not refer to 3D-chromatin structures such as topologically associated domains (TADs) as described in [4], but rather to the composition of the one-dimensional symbolic representation of the genome. These include repeating elements such as tandem repeats [5], regular sequence motifs and cis-regulatory regions such as TATA boxes [6] and the differences between coding and non-coding regions [7], [8],[9], [10]. We focused our study on viral genomes, as they are small and easy to handle with low computational resources.

All models, code, and datasets will be available at GitHub on https://github.com/dna-llm/learning-nucleotides. However, pre-trained models will only be available at a reasonable request as the dataset is based on pathogenic viruses.

2. What is Ω, really?

2.1. Ω is the output of evolution

There are a few essential points we should be aware of regarding evolution as a highly context specific, complex, adaptive framework. One, whether that framework can be thought of as deterministic or stochastic is dependent on the particularities of the organism. Two, its constraints are functional not sequential, backward not forward looking, developmental not constructive. There is no goal or aim to evolution. Three, as it is case for selection, mutations are non-random. Four, evolution's fitness landscape is non-static. And finally five, it operates at the systemic level. No organism exists outside of an ecosystem. See [11], [12], [13], [14] for more discussions on these points.

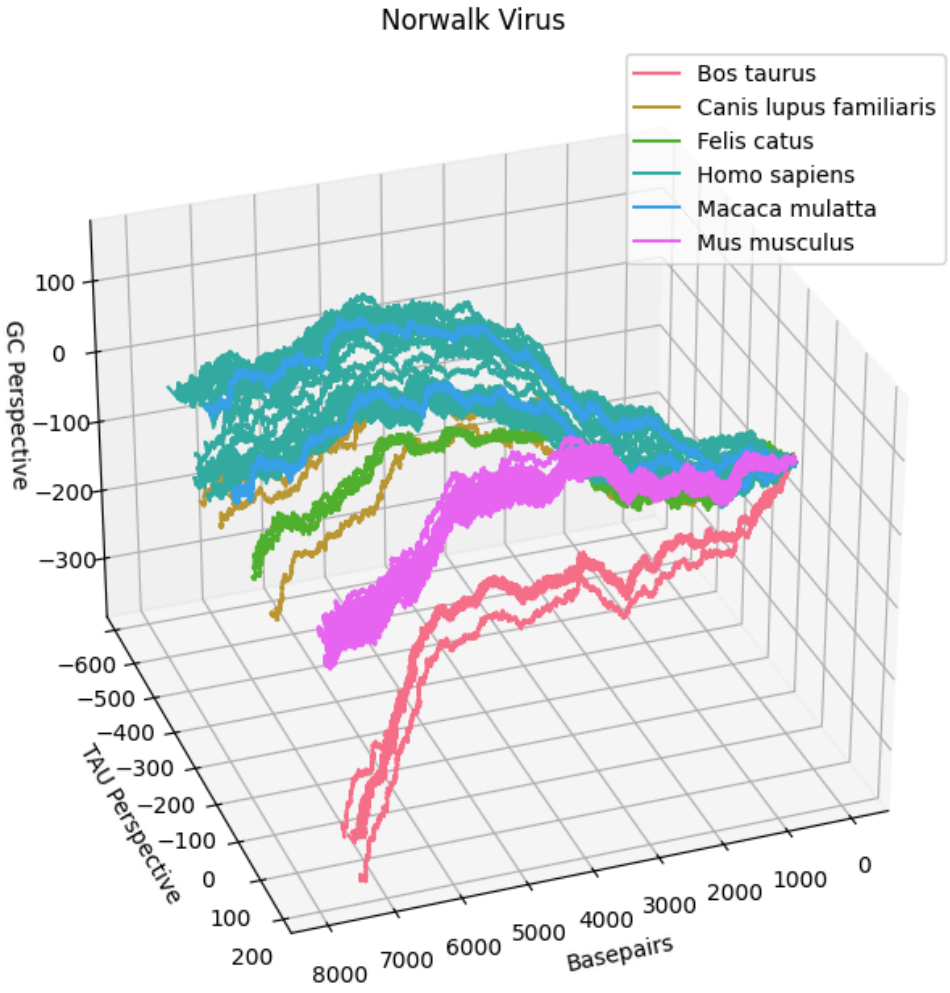


Figure 1: 3D representation of Norwalk Virus samples colored by their hosts.

Taking Norwalk Virus as an illustrative example, in Fig. 1, we can see evolutionary constraints guide sequences towards predictable patterns (particular loci based on host), with deviations from

those patterns correlated. This is due to the fact that each virus can only exist if it can successfully replicate, thus forcing emphasis on genomes with the most useful traits for a particular niche (which in this case is represented by host). Said differently, although the change at any replication point is necessarily Brownian and viral-cell interactions are chaotic, genomes approximate a Lyapunov stable point for a particular niche to minimise system-wide energy expenditure.

This means that for any model to learn biological dynamics, it must learn a complex, modular, compositional function that takes into account epigenetic regulation, regulatory networks of host cells, and the constructive, interchangeable and fractal nature of development.

3. What are the limits of the current paradigm with respect to learning the Ω generating function?

3.1. Can a transformer model learn complex, modular, compositional functions?

3.1.1. Synthetic Sequences

In a word no. Although Jesse [15] proved that decoder models under reasonable conditions can be considered Turing complete, others showed that transformer decoder models struggle to learn slow mixing hidden Markov models as well as sharp OOD functions. To test model limits ourselves, we trained a vanilla transformer decoder model (of the Pythia variant) on synthetic sequences and evaluated its performance on out-of-distribution sequences as we scaled (Fig. A.1, Fig. A.2, Fig. A.3). As can be seen on Fig. A.3, although model size was positively correlated with OOD performance, that correlation followed a power law. This is inline with the literature and what we find from other experiments.

3.1.2. Genetic Sequences

With regard to genetic sequences, we find the results of our experiments similar to those of the synthetic sequences. Transformer models need to be of sufficient size with ample examples to learn from complex datasets, as can be seen from Fig. 2. Note from Fig. 2 that OOD scores do not improve with model size. That's because OOD samples in viral genomes come from completely different populations unlike those in the synthetic experiments.

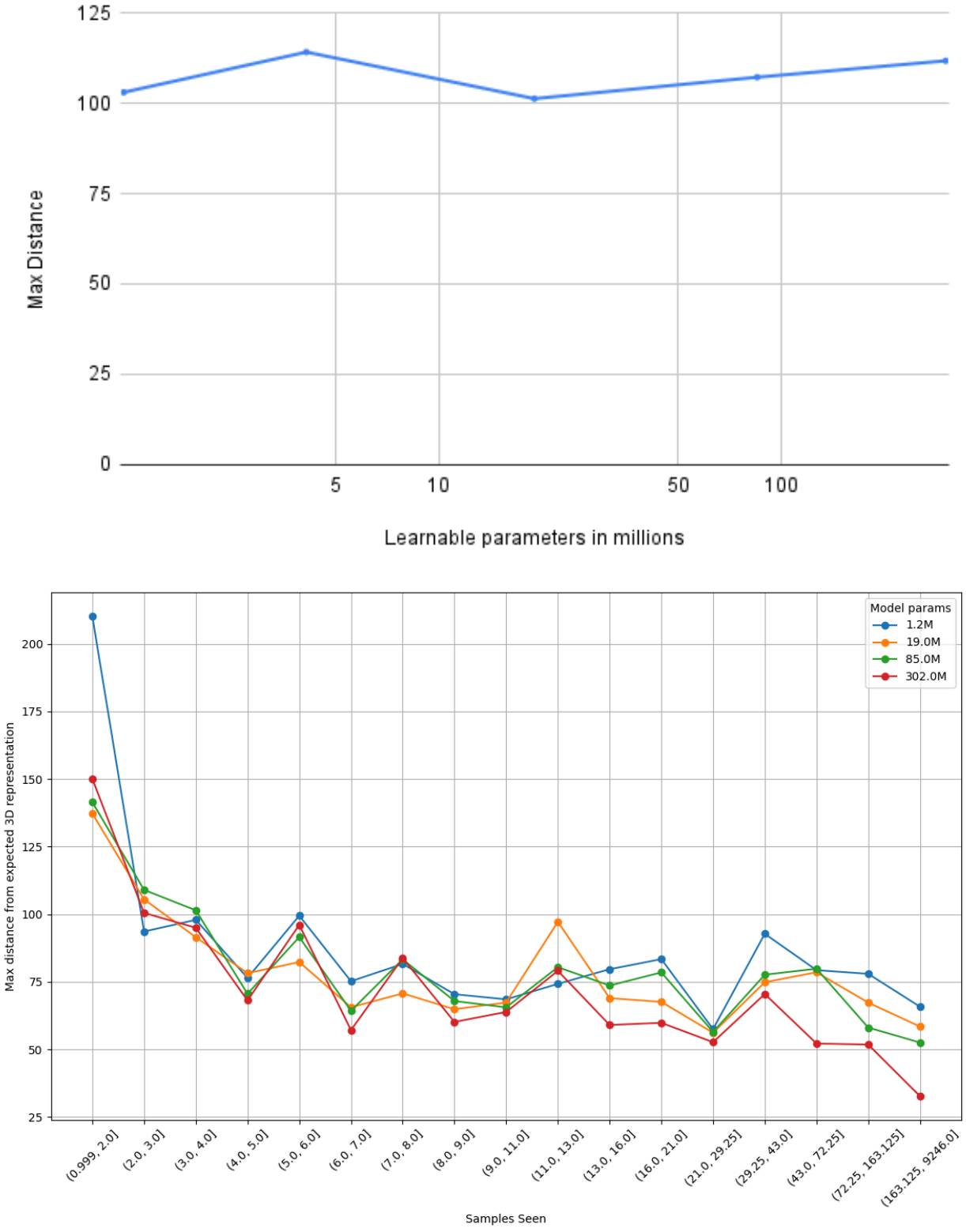


Figure 2: First chart shows max euclidean distance between expected OOD sequences and model size. Second shows max euclidean distances samples and generated seqs

4. Conclusion

A. Appendix A

A.1. Synthetic Sequences Experiment

```python
def synth_seq_gen(seq_length, wavelength, second_wavelength, second_wavelength_cos,
amplitude, amplitude_ratio, shift_exp, power, k, direction):
    # Generate the x values
    x = np.linspace(0, seq_length, seq_length)

    # Convert wavelengths to sequence lengths
    wavelength = seq_length / wavelength
    second_wavelength = seq_length / second_wavelength
    second_wavelength_cos = seq_length / (second_wavelength_cos * 10 *
np.random.rand(1))

    # Calculate the shift for the exponential component
    shift_exp = seq_length / shift_exp

    # Exponential component
    exp_component = amplitude * np.cos(2 * np.sin(2 * np.cos(2 * np.pi * (x +
shift_exp) / second_wavelength)))

    # Large amplitude component
    large_component = (amplitude * amplitude_ratio) * np.sin(2 * np.pi * x /
wavelength)

    # Second wavelength component
    second_component = amplitude * np.sin(2 * np.pi * x / second_wavelength) *
np.cos(1.1 * np.pi * x / second_wavelength_cos)

    # Polynomial component
    polynomial_component = direction * (k * x**power)

    # Noise component
    noise_component = 2 * np.random.standard_normal(seq_length) *
np.random.randint(2, size=seq_length)

    # Combine all components
    return large_component + second_component + polynomial_component +
exp_component + noise_component
```

Figure A.1: Screenshot of the function that was used to generate synthetic sequences.

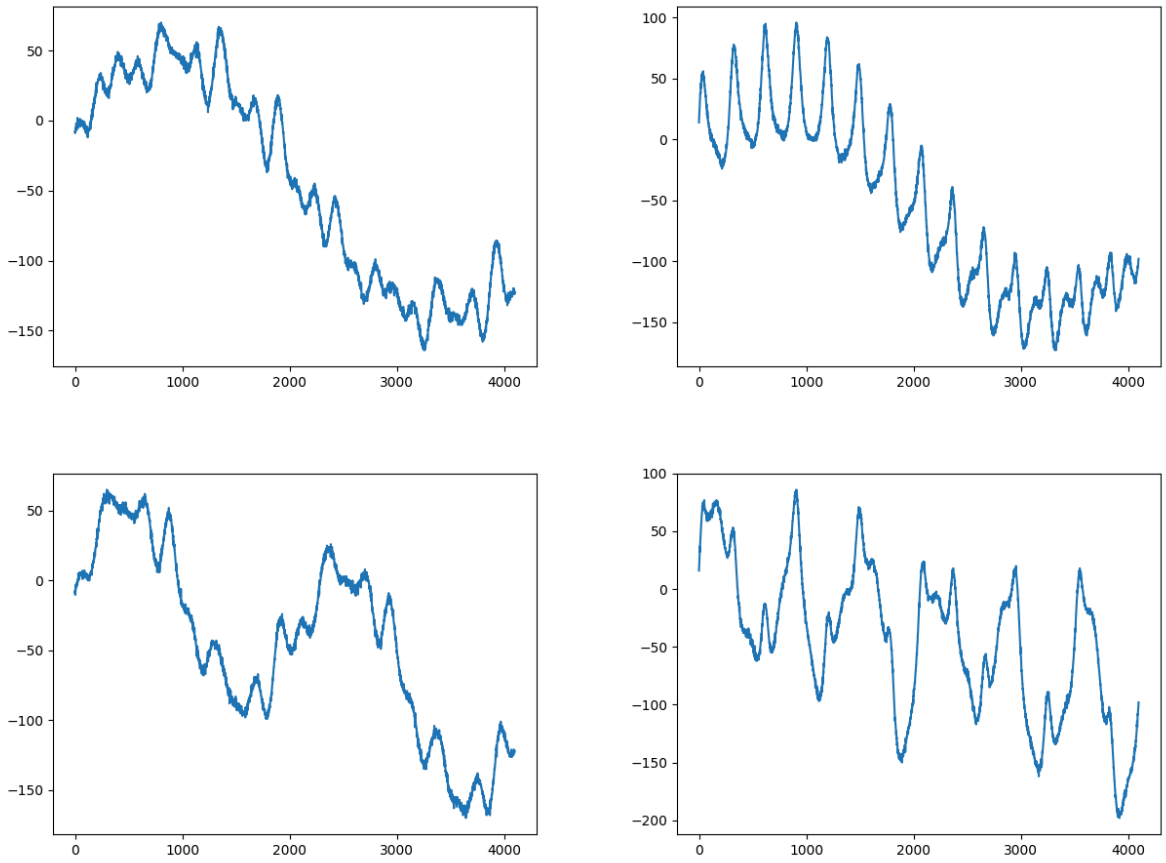
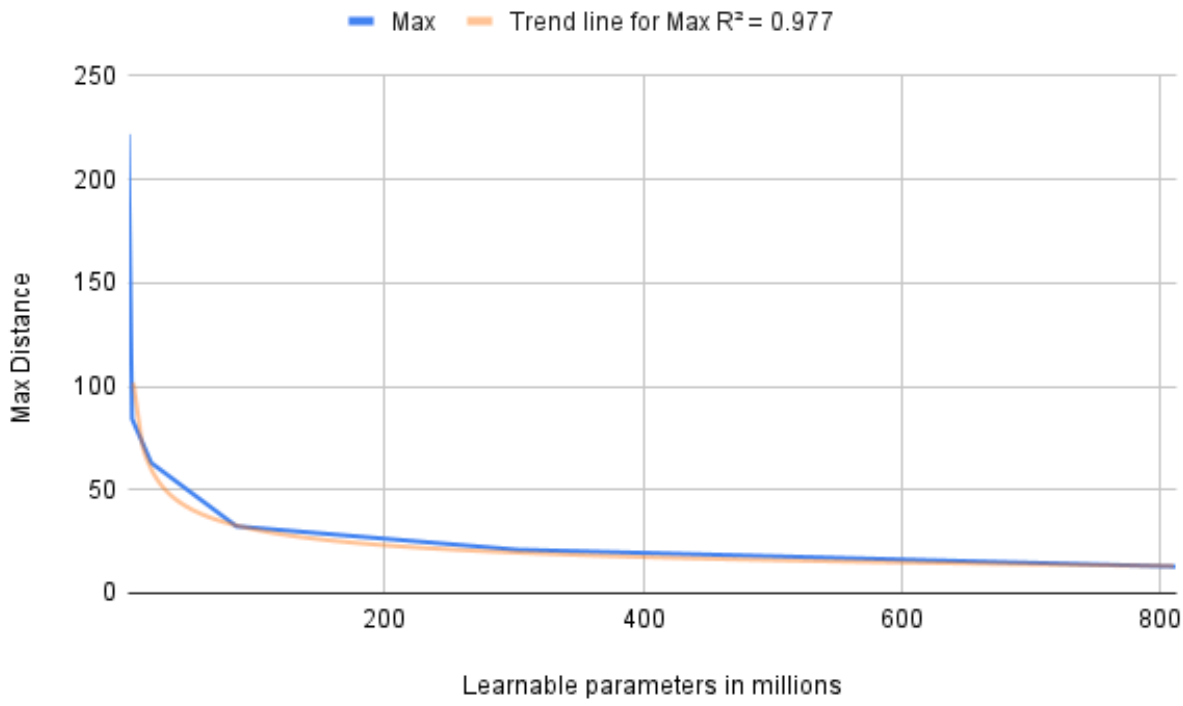


Figure A.2: Example synthetic sequences.



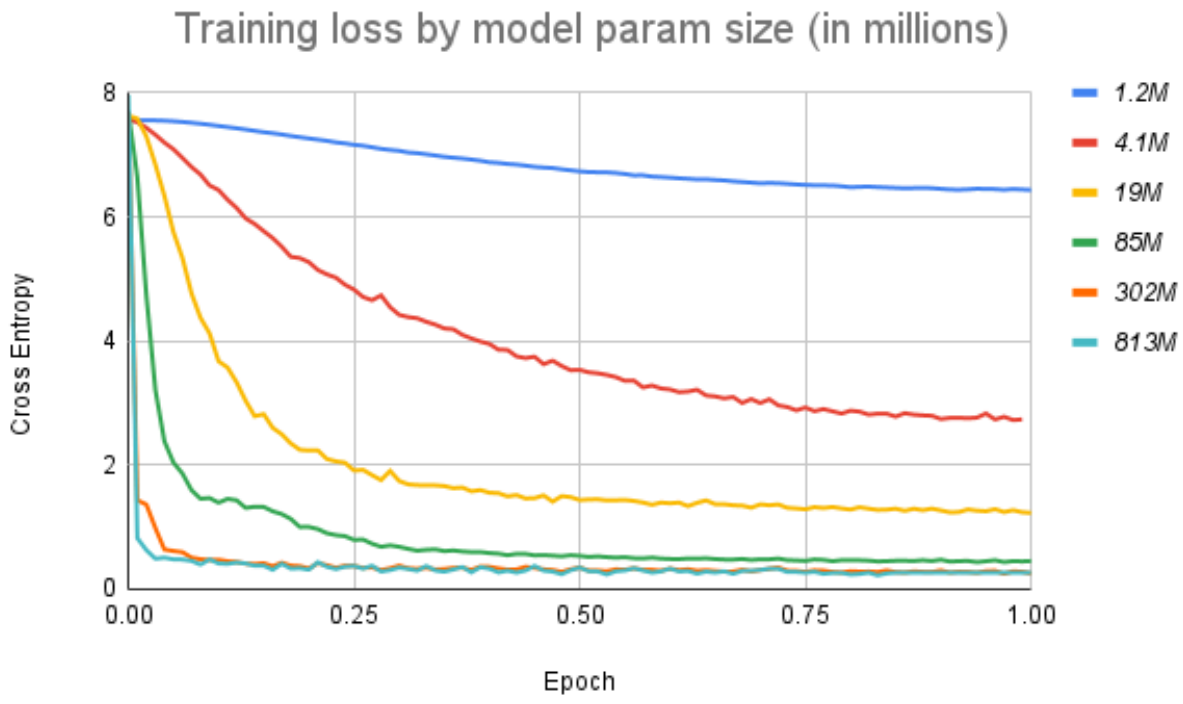


Figure A.3: Synthetic sequence results

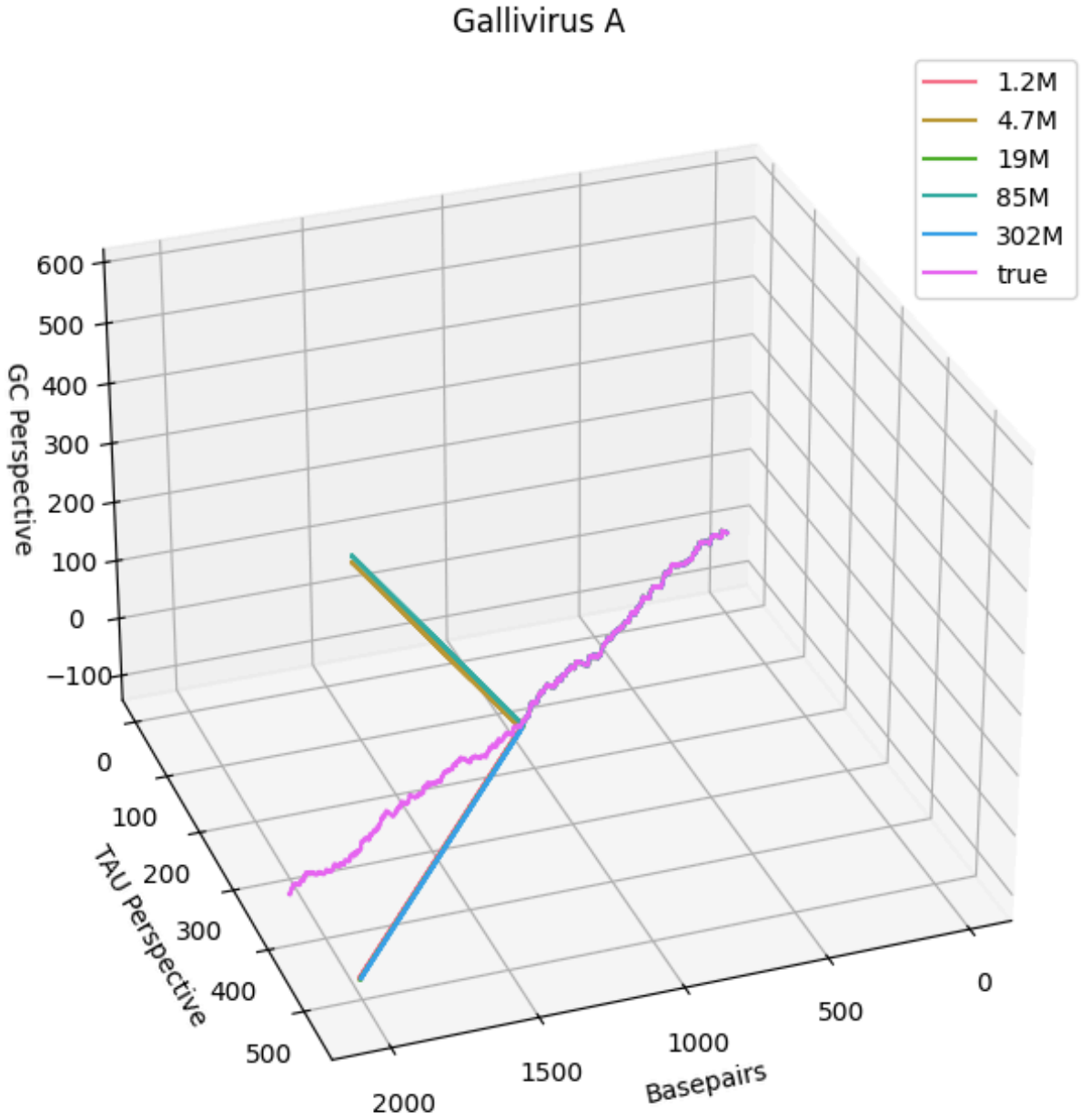


Figure A.4: Screenshot of the function that was used to generate synthetic sequences.

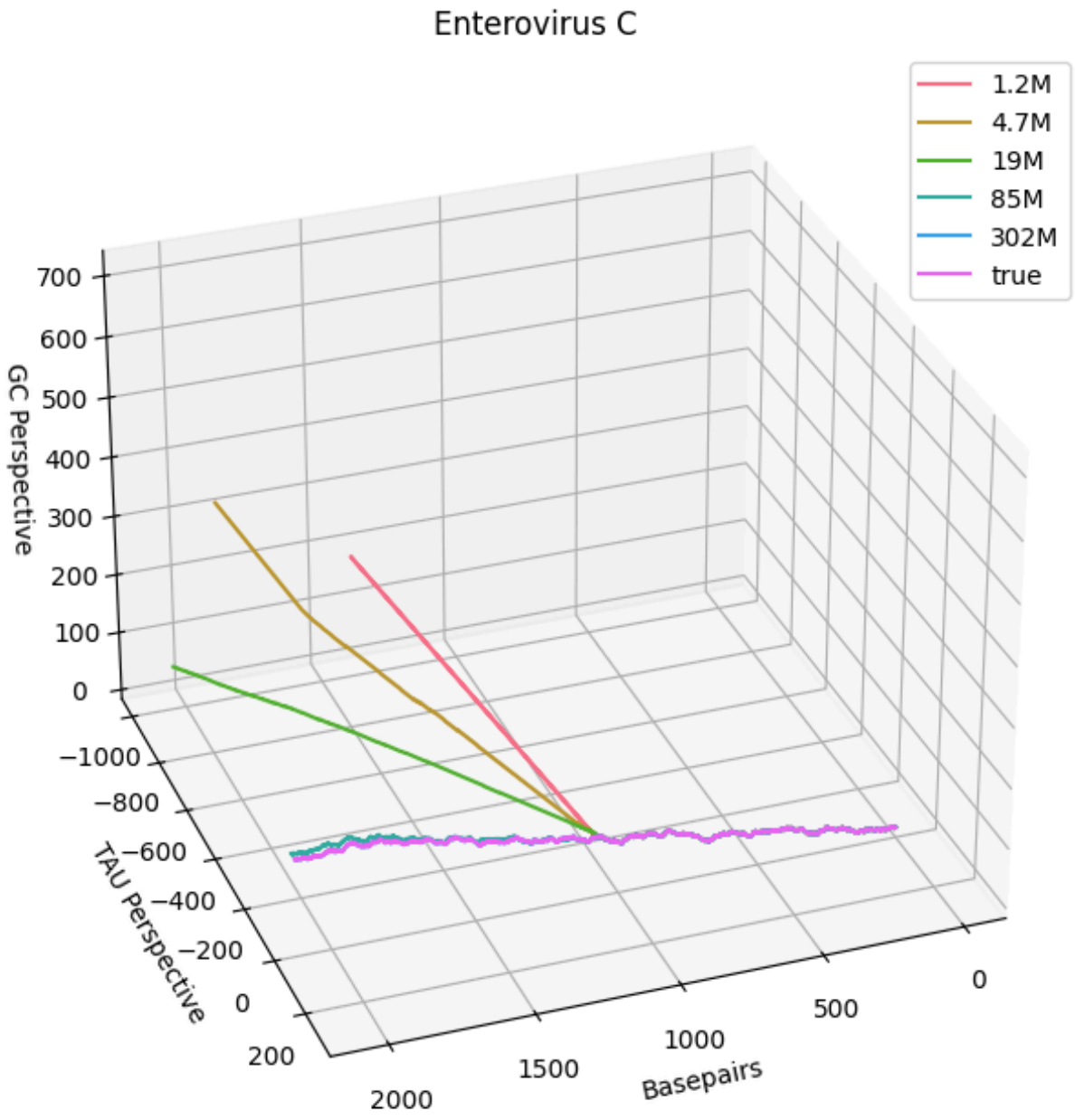


Figure A.5: Screenshot of the function that was used to generate synthetic sequences.

References

[1] E. Nguyen *et al.*, "Sequence modeling and design from molecular to genome scale with Evo," *bioRxiv preprint*, 2024, doi: 10.1101/2024.02.27.582234.

[2] A. B. Duthie and V. J. Luque, "Foundations of ecological and evolutionary change," *arXiv preprint arXiv:2409.10766*, 2024.

[3] G. Benegas, C. Ye, C. Albors, J. C. Li, and Y. S. Song, "Genomic language models: opportunities and challenges," *arXiv preprint arXiv:2407.11435*, 2024.

[4] M. J. Rowley and V. G. Corces, "Organizational principles of 3D genome architecture," *Nature Reviews Genetics*, vol. 19, no. 12, pp. 789–800, 2018, doi: 10.1038/s41576-018-0060-8.

[5] K. Verstrepen, A. Jansen, F. Lewitter, and G. R. Fink, "Intragenic tandem repeats generate functional variability," *Nature Genetics*, vol. 37, pp. 986–990, 2005, doi: 10.1038/ng1618.

[6] S. Kim and J. Wysocka, "Deciphering the multi-scale, quantitative cis-regulatory code," *Molecular Cell*, vol. 83, no. 3, pp. 373–392, 2023, doi: https://doi.org/10.1016/j.molcel.2022.12.032.

[7] E. P. Locey Kenneth J. AND White, "Simple Structural Differences between Coding and Noncoding DNA," *PLOS ONE*, vol. 6, pp. 1–8, 2011, doi: 10.1371/journal.pone.0014651.

[8] K. Zhou, A. Aertsen, and C. W. Michiels, "The role of variable DNA tandem repeats in bacterial adaptation," *FEMS Microbiology Reviews*, vol. 38, no. 1, pp. 119–141, Jan. 2014, doi: 10.1111/1574-6976.12036.

[9] S. Erdozain, E. Barrionuevo, L. Ripoll, P. Mier, and M. A. Andrade-Navarro, "Protein repeats evolve and emerge in giant viruses," *Journal of Structural Biology*, vol. 215, no. 2, p. 107962, 2023, doi: https://doi.org/10.1016/j.jsb.2023.107962.

[10] H.-U. Bernard, "Regulatory elements in the viral genome," *Virology*, vol. 445, no. 1, pp. 197–204, 2013, doi: https://doi.org/10.1016/j.virol.2013.04.035.

[11] L. Fromhage, M. D. Jennions, L. Myllymaa, and J. M. Henshaw, "Fitness as the organismal performance measure guiding adaptive evolution," *Evolution*, vol. 78, no. 6, pp. 1039–1053, 2024, doi: 10.1093/evolut/qpae043.

[12] L. Fromhage and A. I. Houston, "Biological adaptation in light of the Lewontin–Williams (a)symmetry," *Evolution*, vol. 76, no. 7, pp. 1619–1624, 2022, doi: 10.1111/evo.14502.

[13] B. Aaby and G. Ramsey, "Three Kinds of Niche Construction." [Online]. Available: https://philsci-archive.pitt.edu/16718/

[14] A. Stoltzfus, *Mutation, Randomness, and Evolution*. Oxford University Press, 2021. doi: 10.1093/oso/9780198844457.001.0001.

[15] J. Roberts, "How Powerful are Decoder-Only Transformer Neural Models?," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8. doi: 10.1109/IJCNN60899.2024.10651286.