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 discuss 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.

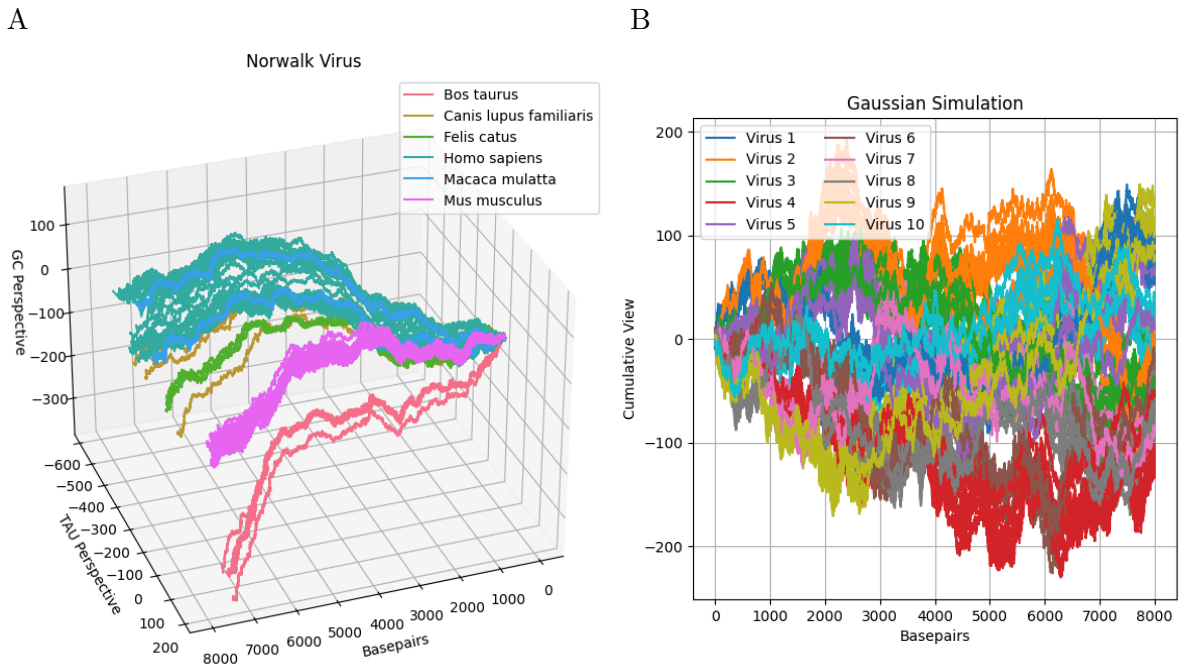


Figure 1: A. 3D representation of Norwalk Virus samples colored by their hosts. The x-Axis represents the basepairs, for every basepair we move exactly one. TAU perspective represents the cumulative summation of T (0,1,1), A (0,-1,1) or U (0, 1, 1) basepairs along the x-axis. GC perspective represents the cumulative summation of G (1,0,1) and C (-1,0,1). The viruses adapted to different host species clearly occupy different regions in the overall information representation space of Norwalk virus genomes. Hosts that are closer to each other taxonomically result in overlapping regions, e.g. humans and rhesus monkeys. B. Gaussian simulation of sub-region development in the space of related viral sequences. The simulation uses 10^-5 as a mutation rate, as found in RNA viruses. x-axis represents the basepair equivalent, y-Axis ….

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 [11]. Two, its constraints are functional not sequential, backward not forward looking, developmental not constructive [11], [12]. There is no goal or aim to evolution. Three, as it is case for selection, mutations are non-random [13]. Four, evolution's fitness landscape is non-static [11], [14]. And finally five, it operates at the systemic level. No organism exists outside of an ecosystem [12], [14].

Taking Norwalk Virus as an illustrative example, in Fig. 1 A, 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 the host). Said differently, although the change at any replication point is necessarily Brownian (through different mutations occurring at each replication) and viral-cell interactions are chaotic, viral genomes adapted to the host (the particular niche) approximate a temporary Lyapunov stable point to minimise system-wide energy expenditure (the system representing the viral cell populations along with the host's cells).

To simulate this in a Gaussian framework, we can sample 10 sequences from a random sequence matrix of shape [samples, length, basepairs]. Those 10 sequences represent successful replications within a host. Each of those sequences can then be replicated 1000 times with a level of accuracy similar to viral mutation rates (10^{-5} per bp [15]). The results of such a simulation can be seen in Fig. 1 B, showing clear speciation of the initial samples, similar to what we observe in nature. What we observe here arises just from the initial randomness combined with Brownian noise. Together, 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 [16].

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

To test our hypothesis in a controllable fashion, we conducted an experiment with synthetic sequences, represented as composite functions, based on combining sinusoidal terms with varying wavelengths, exponential modulation, polynomial growth, and additive noise. Samples from these functions can be found in Fig. A.1. Similar to the DNA representation above they exhibit micro and macro structure, although their periodicity is stronger. We trained a transformer model of the Pythia variant on these sequences on model scales from 1.2 million to 800 million parameters and evaluated its performance on out-of-distribution (OOD) sequences. See Section B.1.2.1 for a description of the sequence generation.

As shown in Fig. A.2 A, although model size was positively correlated with OOD performance, that correlation followed a power law and did not increase strongly after 300 million parameters. This is inline with the literature [CITE] and what we find from other experiments. Fig. A.2 B shows the training loss also stagnating after 300 million parameters, pointing at a bound for possible learning of these functions for transformer architectures.

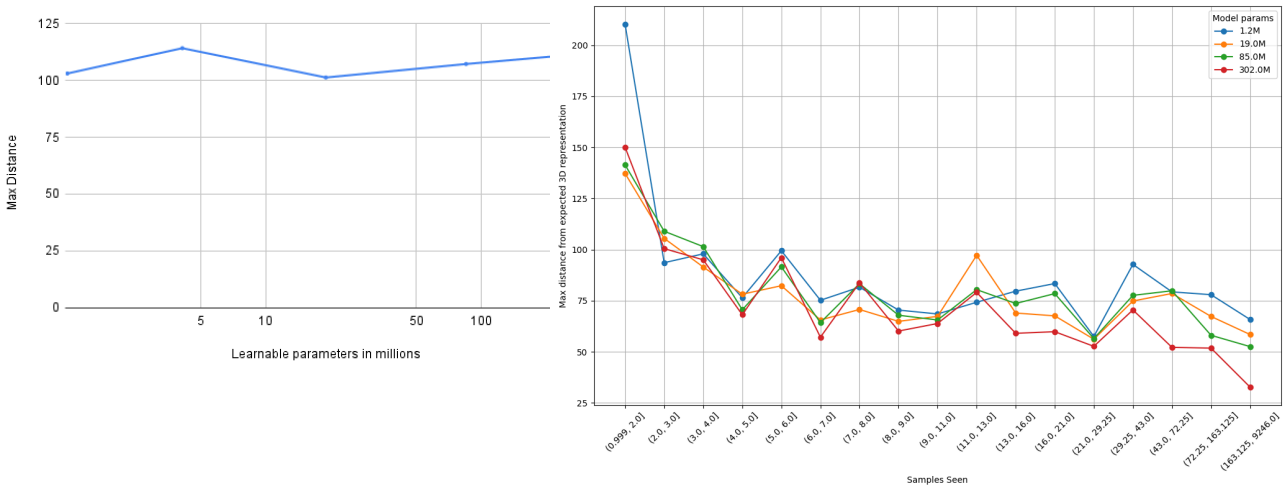


Figure 2: A. First chart shows the maximum euclidean distance between expected 3D representation of OOD sequences and generated sequences on the y-axis against the model size on the x-axis. This means that the model cannot learn OOD sequences irrespective of scale. B. shows the maximum euclidean distance between original sequences and respecticve generated sequences against the number of samples of a species that has been seen during training. Here we can observe the distance decreasing with number of examples by species seen in the training set.

3.1.2. Genetic Sequences

With the already lacing performance in synthetic and considerably less chaotic sequences, we also tested this on real genetic sequences of virus genomes from NCBI. Gennomes were fed into the models with a context window size of 2048 basepairs and an overlap of [XXX] %. A simple 8 word tokenizer was used. See section B.1.4 for a description of the dataset and tokenizer.

When training on 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 B. However, Fig. 2 A shows that OOD sequences (and even sequences that have have a low number in the training set, as indicated by the left-hand side of plot B) remains high regardless of scale of the model, clearly contrasting the synthetic experiment and showing that transformers struggle to learn the underlying functions comprising DNA sequences.

Taking a closer look at single sequences, Fig. 3 compares a virus with low numbers of samples in the dataset (Gallivirus A) with a more abundant virus type (Enterovirus C). The true sequences is given in violet and the generated sequences are colored by the parameter size of the generating model. Note that in Gallivirus (panel A) none of the models can generate a sequence close to the original after being prompted with the first half of the original sequence. Only the biggest model is able to learn the direction of the sequence representation in the function space. In contrast in panel

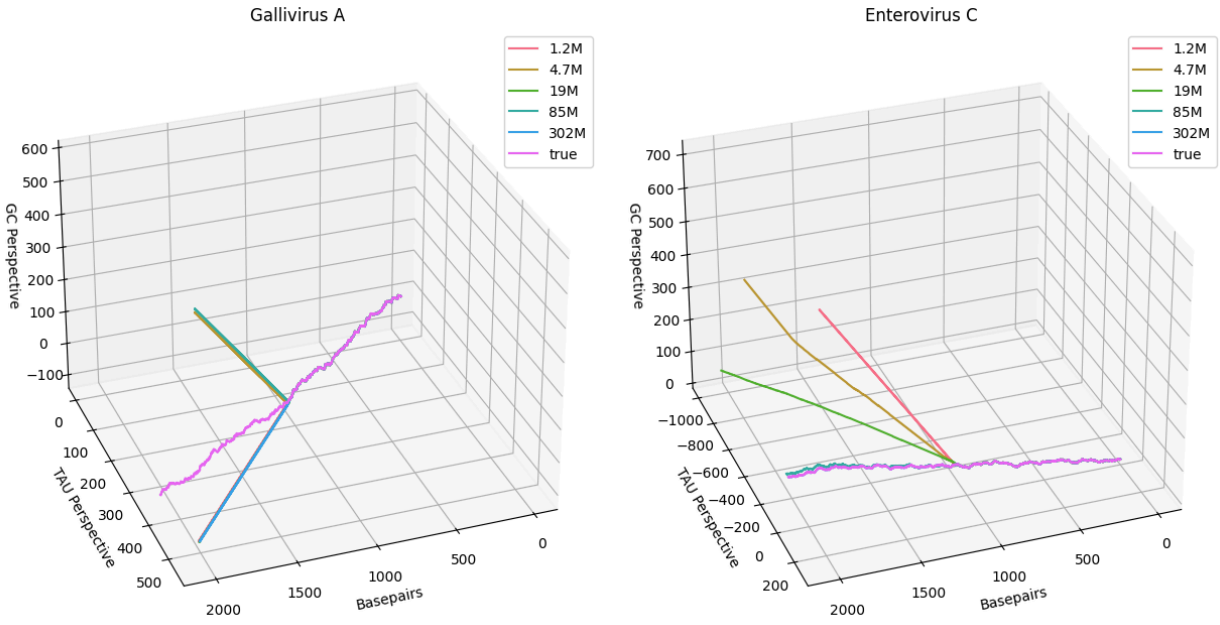


Figure 3: A. 3D representation of generated Gallivirus virus sequences colored by model size vs true sequence. Note half the sequence (1024 basepairs) is fed into each model before generation. Also note that 302M is the only model able to correctly assign the right direction, however none learn the microstructure. B. Entrovirus C virus sequences colored by model size vs true sequence. Note that as we increase model size, the outpute increase in correctness (probably due to copying). This different behaviour is likely based in the number of samples for each species: Enterovirus is much more abundant with [XXX] number of samples against [XXX] for Gallivirus. The model is able to memorize the Species overall structure, when it sees enough samples and has a large enough parameter size. However small sample sizes are not sufficient to memorize a species'profile but the model is not able to learn underlying functions that could allow knwoledgetransfer from one to the other species. Figure axes as in Figure 1.

B the models larger than 82M learned the function almost identically, pointing at memorization rather then learned knowledge, which is a known issue of LLM paradigm [CITE].

4. Discussion

With this work, we want to provoke a discussion about the feasibility of using LLM training paradigms to learn DNA sequences and apply this knowledge to downstream tasks in the life sciences and medicine. We discuss findings from mathematical evolutionary theory that we see as applicable to the general problem of what needs to be learned about biology and use them to reason about what a model needs to learn for foundational capabilities that depict biology well enough to make predictions or connections. We show that with a 3D representation of sequences the character of DNA as a function is apparent and provide an intuition on how evolution in viruses shapes the function space of the sequences.

Based on this we show with synthetic composite functions similar to what we observe in our numerical representations of DNA, that transformer models struggle to learn the underlying functions of these sequences and suffer from a power law when it comes to predict OOD samples. When moving on to the real world, we observe this shortcoming to an even larger extent: OOD Samples, and even sequences of viral species that have been seen, but with low numbers over all, will no be generated correctly and result in large differences between the original and generated sequence.

[**TODO** Hassan, add points about topology stuff, cut on different ways to see DNA (e.g. phase transition, ODE/SPDE)]

Together, these findings point at a known problem of transformers to learn complex function, although they are in theory Turing-complete under reasonable conditions as shown by Jesse et al. [17]. We suggest that these models' ability to learn the macro and micro structure of the DNA as a function could be improved with knowledge infusion. For example, this could be achieved by using topological losses to inform the model. Such a loss could be based on representations similar to what we see in this work or Yau et al. [18]. Similarly, persistent homology based losses could be used to learn the macrostructure of the sequences and enable the model to keep direction and recurring patterns.

Another possibility to gain foundational capabilities and generate viable DNA sequences on genome scale might be a switch in training paradigms and model architectures: Discrete diffusion models… [can still be see as LLM paradigm, learns ODEs?] Flow matching models …. [is likely be able to learn function space despite its complex nature, SPDEs?]

A. Appendix A

A.1. Synthetic Sequences Experiment

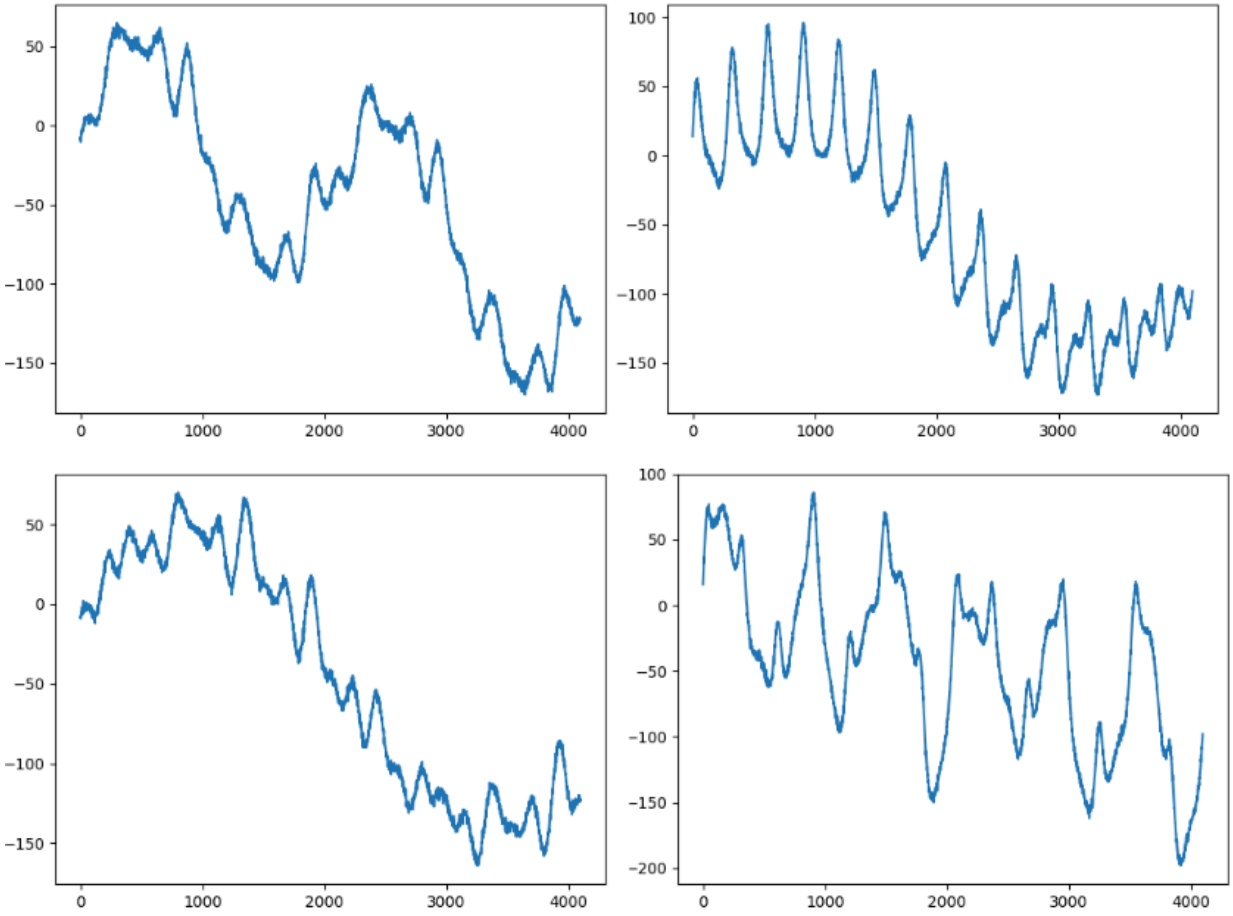
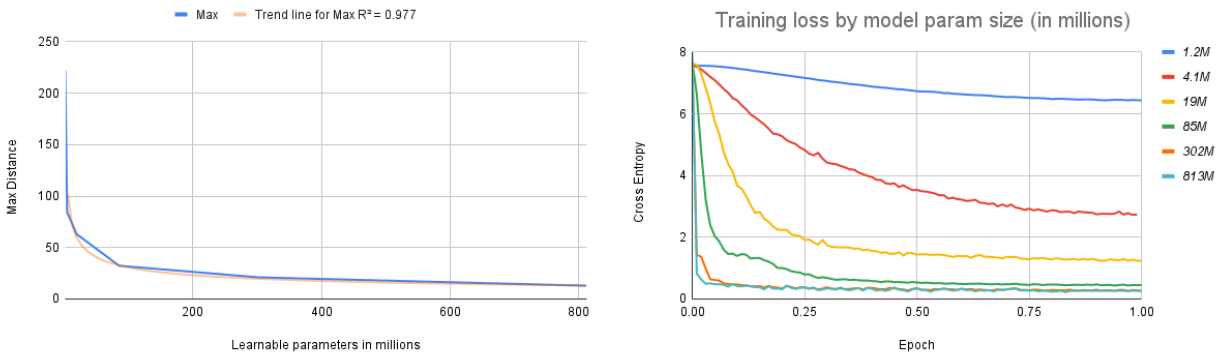


Figure A.1: Example synthetic sequences. The x-axis represents token location and the y-axis represents the token value, which is an integer conversion of our synthetic sequence generator described in Section B.1.3.



A B

Figure A.2: A. First chart shows the maximum euclidean distance between expected 3D representation of OOD sequences and generated sequences on the y-axis against the model size on the x-axis, in coiintrast to the real world sequences, the less haotic synthetic sequences can be learned better with scale. However, learning is restricted by a power law, approaching negligible error decrease at 300 million parameters. B. Second chart shows cross entropy loss across a single epoch grouped by learnable model parameter size in millions.

B. Appendix B

B.1. Methods

B.1.1. Simplified 3D representation of DNA

TODO: Upload code to github, write methods sec (chatGPT) ***TODO: Compare to Energy view and why we used simplified version***

B.1.2. Gaussian Noise based Virus sub-speciation simulation

TODO: Upload code to github, write methods sec

B.1.3. Synthetic Sequence production

To evaluate the ability of transformer models to learn complex functions, we generated synthetic sequences characterized by a combination of periodic, exponential, polynomial, and noise components. The sequence generation process was driven by a function that combined these components mathematically, incorporating sinusoidal terms with varying wavelengths, exponential modulation, polynomial growth, and additive noise. Key parameters such as amplitude, wavelength, power, and direction were systematically varied using a Cartesian product of predefined ranges. For example, wavelengths ranged from 1 to 10 units, amplitudes from 10 to 50, and amplitude ratios from 1 to 6. This systematic exploration resulted in a diverse dataset of synthetic sequences, each with a length of 4096 samples, generated using the synth_seq_gen function (find the implementation in the notebooks section of the repository). A total of N sequences were produced, and metadata for each sequence, including the parameters used, were stored for analysis. Periodically, visualizations of the generated sequences were saved to ensure quality and diversity.

The generated sequences were used to train transformer models, enabling an evaluation of their ability to approximate these synthetic functions and generalize across parameter configurations. We analyzed model performance as a function of dataset size, sequence length, and the complexity of the underlying parameters, focusing on the impact of scaling. Our results revealed that while transformers could partially learn these functions, their generalization capacity was limited, and their performance exhibited a pronounced dependence on scale. Specifically, we observed a power-law degradation in performance, highlighting intrinsic limitations in the transformer architecture for learning such structured, multi-component functions.

TODO: How have OOD seqs been picked?

B.1.4. Experiments with Virus Genome Data

TODO: Upload code to github, write methods sec (chatGPT)

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

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

[15] K. M. Peck and A. S. Lauring, "Complexities of Viral Mutation Rates," *Journal of Virology*, vol. 92, no. 14, p. 10, 2018, doi: 10.1128/jvi.01031-17.

[16] "Synthesis," *Properties of Life: Toward a Theory of Organismic Biology*. The MIT Press, 2023. doi: 10.7551/mitpress/14739.003.0007.

[17] 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.

[18] S. S.-T. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, and Y.-K. Ho, "DNA sequence representation without degeneracy," *Nucleic acids research*, vol. 31, no. 12, pp. 3078–3080, 2003.