Are Current AI Systems Unlocking Knowledge Discovery in Genomics?

Antonio Orvieto

The last few years of building large foundation models in language and vision domains have revealed how steady progress can be achieved by scaling resources while keeping the model architecture (a softmax attention-based Transformer) fixed.[1] This strategy has been readily applied in science, where large-scale open-source foundation models have been released, and scaling laws indicate that pretraining (or non-task-specific) performance improves with increased compute and data availability.[2]

Despite the impressive versatility and zero- and few-shot performance of such AI-for-science models, recent studies have questioned whether they are genuine foundational architectures, and whether they truly let scientists process data with unprecedented accuracy compared to pre-2023 machine learning. While scaling can reveal pleasant surprises for the future, we now have evidence that simple supervised machine-learning approaches from the last decade – while less versatile and requiring sufficient training data – can still offer a performance edge over large foundation models at a fraction of the computational resources needed. For instance, computational biologist Ziqi Tang and colleagues evaluated the representational power of pretrained genomic foundation models for major functional genomics prediction tasks spanning DNA and RNA regulation, and found that probing their representations can offer no substantial advantage over machine-learning approaches from the last decade.[3]

This is, in a way, to be expected: the most profound aspect of modern AI systems is their versatility. Yet can progress in biological data processing come solely from enhanced versatility? As data availability increases, it is crucial not to follow text-driven AI insights uncritically, but to dedicate our efforts to rethinking and understanding how best to design and (pre)train specialized models with unprecedented predictive power for scientific tasks. A likely path forward, exemplified by recent progress in stateful memory models (such as Mamba and the xLSTM), is to shift away from the needs of pure language modeling toward neural networks inspired by human memory and smooth structures in natural data.[4] It is helpful to remember that the transformer architecture was initially designed for the text

https://doi.org/10.1162/DAED.a.993

domain and in-context retrieval, and was later applied to other domains due to its ease of multimodal adaptation and the technical needs of our computational infrastructures. In light of recent literature showing concerning limitations of this neural network (such as state tracking, which relies on dialogue to create records and context for conversations over time), perhaps it is now time we abandon architectures defined by the needs of text and instead design truly domain-specific foundation models, beginning with the unique characteristics of scientific data.[5]

The DNA structure offers an excellent motivation for this paradigm shift: citing historian of biology Nathaniel Comfort, "if a genome is text, it is badly edited. Most DNA is gibberish, full of stutters, snippets of doggerel from other species." DNA also offers a compelling example of a nontext domain in which data availability is rapidly increasing.[6] The cost of sequencing is dropping sharply, and as genomic data volumes grow, AI gains enormous potential to reveal new knowledge. Yet we still lack understanding of the hidden logic in 98 percent of our genome that does not code for proteins – the noncoding regions whose patterns describe vital yet mostly unknown regulatory functions. Unveiling novel structures in these long sequences could unlock new treatments for cancer, autoimmune disorders, and neurological conditions.

In my theory-focused lab at the Max Planck Institute for Intelligent Systems, part of the newly established ELLIS Institute in Tübingen, Germany, we develop new architectures and training techniques for processing and pattern matching in extremely long sequences of symbols, such as those in our DNA.[7] Specifically, we are developing new efficient models with adaptive hierarchical computation that can increase the reasoning budget based on data, ensuring mathematical guarantees of enhanced expressivity toward Turing-complete reasoning. Our latest Fixed-Point RNN model can surpass both transformers and long short-term memory networks (LSTMs) on a wide range of reasoning benchmarks, including state tracking, with no need for chain-of-thought reasoning or language modeling pretraining.[8]

Yet providing a good fit for the biological domain is not as straightforward as expected: benchmark results show only minor improvements, and we lack solid intuitions about the reasoning archetypes (memorization, recall, tracking) that could unlock understanding of long nucleotide sequences. Despite our clear direction toward improving long-sequence DNA processing machines – with enhanced expressivity through recurrent depth, hierarchical reasoning, and dynamical context pruning – the path forward can only be successful through close collaboration between AI experts and biologists to design robust, hard benchmarks that assess progress in a quantifiable way and direct researchers worldwide toward challenging problems.

So are current AI systems unlocking knowledge discovery in genomics? Perhaps not yet; but they are teaching us how to ask the right questions, illuminating

the path toward systems that eventually will. As artificial intelligence advances and data costs fall, we must approach new genomic technologies with an open mind: Progress will depend not on scaling alone but on our willingness to question assumptions. Rather than following insights from other domains, we should build a scientific foundation. Across my lab and many others, AI in genomics represents both a technical and philosophical opportunity – one that can advance drug discovery, benefit society, and push forward the theory of application-aware neural network design.

ABOUT THE AUTHOR

Antonio Orvieto is a Hector Endowed Principal Investigator at the ELLIS Institute Tübingen and an Independent Group Leader at the Max Planck Institute for Intelligent Systems, where he leads the Deep Models and Optimization group. He is also a Lecturer at the University of Tübingen, faculty for PhD programs in Tübingen and Zurich, and an AI2050 Early Career Fellow. He is an Action Editor for *Transactions on Machine Learning Research* and an Area Chair for NeurIPS and ICML. His papers have been accepted and published at such major AI conferences as NeurIPS, ICLR, ICML, and AISTATS.

ENDNOTES

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, et al., "Scaling Laws for Neural Language Models," arXiv (2020), https://doi.org/10.48550/arXiv.2001.08361.

[2] Žiga Avsec, Natasha Latysheva, Jun Cheng, et al., "Advancing Regulatory Variant Effect Prediction with AlphaGenome," *Nature* 649 (8099) (2026): 1206–1218, https://doi.org/10.1038/s41586-025-10014-0; and Garyk Brixi, Matthew G. Durrant, Jerome Ku, et al., "Genome Modelling and Design Across All Domains of Life with Evo2," *Nature* 652 (8112) (2026): 1349–1361, https://doi.org/10.1038/s41586-026-10176-5.

[3] Ziqi Tang, Nirali Somia, Yiyang Yu, and Peter K. Koo, "Evaluating the Representational Power of Pre-Trained DNA Language Models for Regulatory Genomics," *Genome Biology* 26 (1) (2025).

[4] Albert Gu and Tri Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," paper presented at the First Conference on Language Modeling (COLM 2024), Philadelphia, Pennsylvania, October 7, 2024, https://openreview.net/pdf?id=tEYskw1VY2; and Maximilian Beck, Korbinian Pöppel, Markus Spanring, et al., "xLSTM: Extended Long Short-Term Memory," in *NIPS '24: Proceedings of the 38th International Conference on Neural Information Processing Systems*, ed. Amir Globerson, Lester Mackey, Danielle Belgrave, et al. (Curran Associates, Inc., 2024).

[5] William Merrill, Jackson Petty, and Ashish Sabharwal, "The Illusion of State in State-Space Models," in *ICML '24: Proceedings of the 41st International Conference on Machine Learning*, ed.

Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, et al. (Journal of Machine Learning Research, 2024).

[6] Nathaniel Comfort, "Genetics: We Are the 98%," *Nature* 520 (7549) (2015): 615–616.

[7] Antonio Orvieto, Samuel L. Smith, Albert Gu, et al., "Resurrecting Recurrent Neural Networks for Long Sequences," in *ICML '23: Proceedings of the 40th International Conference on Machine Learning*, ed. Andreas Krause, Emma Brunskill, Kyunghyun Cho, et al. (Journal of Machine Learning Research, 2023).

[8] Sajad Movahedi, Felix Sarnthein, Nicola Muca Cirone, and Antonio Orvieto, "Fixed-Point RNNs: Interpolating from Diagonal to Dense," paper presented at the Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025), San Diego, California, December 1, 2025, https://openreview.net/pdf?id=KT8y9pFgJE.