Concept Paper

On Using Large Language Models to Understand the Language of Life

Joao Pedro Magalhaes * and George M. Church

Posted Date: 14 February 2026

doi: 10.20944/preprints202602.1094.v1

Keywords: AI; aging; biology; genetics; machine learning

 doi:10.20944/preprints202602.1094.v1

Concept Paper

On Using Large Language Models to Understand the Language of Life

João Pedro de Magalhães [1,2,*] and George M. Church [3,4]

[1] Genomics of Ageing and Rejuvenation Lab, Department of Inflammation and Ageing, College of Medicine and Health, University of Birmingham, B15 2WB, United Kingdom.
[2] The Institute for Data and AI, University of Birmingham, B5 7SW, United Kingdom.
[3] Department of Genetics, Harvard Medical School, Boston, MA, USA.
[4] Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA.
* Correspondence: jp@senescence.info; Tel.: +44 121 3713643

Abstract

The application of machine learning to large datasets has ushered in a new era, exemplified by large language models like ChatGPT that represent accurate statistical representations of human languages. With advances in high throughput methods for assaying living systems, such as DNA sequencing, there is a growing number of applications of machine learning and AI in biology. Despite this progress, our understanding of biological systems remains limited, and we are still far from predicting human biology. Here, we argue that decoding the functioning of the human genome and its gene products on a large scale would enable the creation of predictive models for human biology. By modeling the interactions of gene products over time and space, leading to cell functions that collectively contribute to tissues and a functional organism, we could potentially predict human biological functions, processes and phenotypes. This approach has the potential to revolutionize biology and biomedical research, offering computational models for development, human physiology, and diseases. To understand human biology and disease, however, biological time is a key variable, and we discuss the need to decode the principles of cellular transitions. A predictive model of the language of life, with temporal and spatial resolution, is ambitious yet, in theory, technologically feasible and would have profound implications for comprehending human biology in health and disease.

Keywords: AI; aging; biology; genetics; machine learning

Introduction

Recent advances in deep learning models have showcased the capability of machine learning and artificial intelligence (AI) methods to interpret and decipher large datasets. This is particularly striking in large language models (LLMs) like ChatGPT, a statistical model of the English language, employing probabilities to associate words and construct sentences [1]. More specifically, an LLM is a deep learning system using a transformer architecture, characterized by billions of parameters trained on massive data [2]. LLMs like ChatGPT, DeepSeek and Gemini have been able to decode human languages with remarkable results in generating human-like text across multiple topics and contexts. These models are also increasingly being applied in diverse fields, including the life sciences.

Deep learning methods have been successfully employed to infer and predict protein structures [3,4], gene regulatory networks [5] and other biological phenomena [6–9]. There is a growing prospect that coupling deep learning methods with large data sets can help in decoding biology and transform biomedical research [10]. Ultimately, the goal is to employ large language models to understand the language of life. But what does it mean to crack the language of life?

Here, we argue that to understand life and human biology, we must decode the genome at a level that enables us to replicate and model its functions across cellular and physiological contexts.

 doi:10.20944/preprints202602.1094.v1

Decoding the genomic information and understanding how it gives rise to phenotypes is the closest to understanding life at a level that allows predictive biological changes in health and disease states to be inferred.

Unravelling the Language of Genes

The digital core of information from which cellular functions and human phenotypes arise is ultimately knowable [11]. Every human being originates from a single-cell zygote with two sets of chromosomes that – almost miraculously – guide processes like cell division, growth, differentiation, and development, shaping the intricate functions of an adult human being. Decoding the genome is tantamount to decoding life, as it encodes every biochemical, cellular, physiological, and anatomical aspect of human existence. Not everything in biology is deterministic, but the fundamental aspects of human biology are genetically determined and, in theory, can be predicted and modelled from the genome. Additionally, the genome encodes interactions crucial for life, such as environmental influences, interactions between gene products, and cell-to-cell interactions.

Recent advances in predicting genomic features [12], the effects of mutations on phenotypes [13], gene function [14] and protein function [3,4,7,15] and their interactions through in silico analysis of large datasets are a first step in decoding the genome. Our overarching objective in building predictive models of human biology is to unravel the language of genes. Yet, our ultimate goal must be to unravel not only the functions of individual genes but also how these genes collectively form a functioning cell and how a myriad of interacting cells give rise to tissues and, ultimately, an organism.

The goal is to understand and model not only how a cell works [16], but predict sequential events for every cell type, for every tissue and every life stage, from fertilization to old age. To put it another way, develop a computational model depicting how a single cell evolves into an adult human, encompassing transitions from a zygote to a fully developed human being. This model would elucidate how each cell type and organ function during normal physiological conditions as well as in response to external stimuli, perturbations such as drugs, and dysfunction in multiple diseases. In simpler terms, the objective is to mathematically predict how cells in each tissue differentiate, function, grow, multiply, and die, driven by specific genetic programs. By capturing all life stages, such foundational model of human biology would also encompass aging processes that underpin the major human maladies and diseases.

While developing such a highly predictive model of life may seem overambitious and well beyond our current capacity, we think that building a sophisticated model of the human genome is achievable. Despite the existing limitations in scope and accuracy of current models predicting cell transitions and reprogramming [17,18], we argue below that developing an accurate, predictive model of human biology is, in theory, within our technological reach.

Data Requirements for Decoding the Human Genome

How much data would be necessary to crack the human genome? In this context, "cracking the genome" goes beyond merely understanding the functions of individual genes; it involves modeling how their products and interactions give rise to sentences and the language of life. To put it another way, such an approach requires understanding the sequential interactions between gene products in a cell, modeling changes in gene expression and resulting proteins, predicting phenotypic outcomes, and understanding how – guided by genetic programs – cells change in time and space in the human body.

How much do we need to understand biological systems in order to model and recapitulate human biology? This is not just a philosophical question; it has important practical implications. While we may not need to understand the quantum properties of every molecule in a cell to understand the cell's functioning, grasping the rules that govern the dynamics and interactions between gene products is essential. This entails understanding the activation, repression, and interactions of gene products, as well as their relationship to cellular functions and dysfunction in

 doi:10.20944/preprints202602.1094.v1

disease. How much data on gene products would be necessary is, at present, impossible to say for sure, but some estimates are possible.

Machine and deep learning models require huge amounts of data. Indeed, models are only as good as the data underpinning them. Extrapolating from ChatGPT's data requirements, we estimate the necessary data to reach a level of genome understanding that enables the development of a comprehensive biological model, a generative model of life.

Data Requirements Extrapolation from GPT-3 to the Human Genome

ChatGPT's groundbreaking version GPT-3 was previously described and trained on $4x10^{11}$ tokens [1]. Briefly, frequent character combinations are grouped into tokens, approximately 4 characters in length, which generally align with words. As individuals typically know around 40,000 words, as a conservative estimate, we have $4x10^{11}$ / 40,000 words = $1x10^{7}$ data instances per word to train GPT-3. If we extrapolate this approach to biology, then let us assume we need a dataset with $1x10^{7}$ measurements per unit for training. But then, how many different biological units do we have? Table 1 shows data from the human genome. Just like to understand human language, we focus on words rather than letters, to understand the genome, we need to focus on genes and their products rather than on individual nucleotides. Although admittedly an oversimplification, one could argue that transcripts, encoding proteins and non-coding RNAs and representing the level of expression that determine cellular processes, serve as the equivalent to words in the genome or to tokens in ChatGPT.

Table 1. Human genome (GRCh38.p14) statistics (http://www.ensembl.org/Homo_sapiens/Info/Annotation).

Base Pairs	3,099,750,718
Gene counts	
Coding genes	19,830
Non coding genes	26,462
Pseudogenes	15,222
Gene transcripts	252,989
Gene variants	
Short Variants	1,110,229,688
Structural variants	7,861,655

In contrast to words that are typically either present or absent in a sentence, genes can undergo substantial modifications, such as through epigenetic mechanisms, to control gene expression. Gene products can also undergo important modifications, such as protein post-translational modifications. Besides, while words in a text are discrete variables, genes in biology represent continuous variables. How many different levels of expression can a gene product have, each with distinct biological impacts? While debatable, if we assume five expression states, ranging from no expression to high expression, and three intermediate states, we would need 252,989 x 5 x $1x10^{7}$ = $1.3x10^{13}$ data instances. Assuming 10 states would result in 10 times more data needed or $2.6x10^{13}$ data instances to unravel the regulation of the genome and so on.

How does this amount of data compare to existing biological datasets? If for the sake of simplicity we focus only on transcripts, a whole transcriptome array can capture the expression of virtually all expressed gene transcripts, making a single array sufficient to obtain data on all "words" in the genome. Consequently, to gather an equivalent amount of data on the genome as used to train ChatGPT, we would need around $5x10^{7}$ to $1x10^{8}$ whole transcriptome arrays. ARCHS4 (https://maayanlab.cloud/archs4/) and recount3 (https://rna.recount.bio/) have in the range of $1.3x10^{5}$ to $3.2x10^{5}$ human RNA-seq samples (plus ~$3-4x10^{5}$ mouse samples), or two to three orders of magnitude less than required.

 doi:10.20944/preprints202602.1094.v1

The Need for Dynamic and Diverse Data

While a text has logic and meaning by stringing together related words and sentences to convey a story or message, a single RNA-seq or protein array is meaningless on its own. Besides, while text is structured as a sequence of symbols and words, gene expression and the proteome are ensembles of interacting molecules across time and space. As such, we need dynamic data that reflects genome regulation and gene product changes across various time points. Crucially, we need data from multiple contexts – ages, organs, cell types, disease states, etc. A key feature both in health and disease is temporal changes in cell phenotypes, which would be crucial for data to capture. The goal is to decode which gene products change, how and in which order in response to stimuli or instructed by normal physiological processes – in turn, encoded in the genome.

Another difference between languages and the genome is that words that are rarely used are unlikely to be important, yet genes that are only rarely used may play crucial biological roles in specific contexts. As such, the type of data necessary to capture relations between gene products for a large-scale LLM-like modeling of the human genome would need to be not only much more abundant than what is presently available but also have much greater temporal resolution and diversity. In this context, by diversity we mean different life stages, cell types, stimuli and other perturbations that allow us to capture the many genetic programs operating in our cells across space – i.e., human organs – and time – i.e., lifespan.

We envision a multi-omics integration of transcriptomics, proteomics, epigenomics, and metabolomics [19], including data from multiple species and evolutionary conservation. One recent foundational model, for example, based on chromatin accessibility data and sequence information, was able to predict gene expression across human cell types [20]. Besides, DNA sequence data can also be used to predict gene expression [21,22], and non-coding RNAs would be included as another layer of epigenetic gene regulation to help infer the grammar of the genome. Moreover, proteins are arguably a more accurate equivalent to words in the genome, since the proteome is closer to phenotypes than the transcriptome; proteomics data in health and disease would thus be essential. Models would progress gradually from modeling cells to tissues, allowing approaches to be refined. Reinforcement from human feedback would ensure the accuracy of the models and prevent hallucinations. Indeed, experimental feedback would be essential in model development and fine-tuning, as in other approaches [23], based on validation experiments.

Caveats and Limitations

Creating a model reflecting the functions of the human genome and its products would revolutionize biology and medicine. It would enable us to simulate normal processes at the cell and tissue level with temporal and spatial resolution, including development and normal physiology, as well as predict sequences of events in numerous diseases and in response to treatments – including drugs. With large gaps in our understanding of complex diseases and the high failure rate of drug discovery in spite of rising costs, tackling the complexity of human biology and of our genome is imperative. However, there are several caveats and limitations.

Understanding human creations like chess or language is more straightforward for a machine than biological processes due to abundant training data and our ability to comprehend the logic behind our creations. Biology, however, is shaped by evolution and poses challenges due to limited data and the lack of conceptual frameworks that can explain its extraordinary complexity. As pointed out by others [10], the language of life is much more complex than any human language. As such, while advances in foundational models have been impressive, some models still struggle to beat simple baselines. For example, single-cell foundation models do not yet outperform linear baselines for perturbation prediction [24]. Our data requirements calculations are consequently rife with unknowns and should be considered educated guesses. We think, however, that more than the amount of data necessary, data diversity and perturbations would be essential.

Importantly, our proposed framework is not designed to model molecular structures or atomic-level representations of biochemical functions, only their inputs and outputs – which will suffice for

 doi:10.20944/preprints202602.1094.v1

our purposes. As such, we arguably have the technology to obtain the necessary measurements and data to develop such a foundational model of human biology. Moreover, while a computational model of how the genome guides development, cellular functions and human physiology could recapitulate the development of human brain functions of its components, it will not capture the emergence of thinking as readouts would not focus on neural connections. Indeed, our proposed model would be limited to the phenotypes whose readouts are included in the training data, emphasizing again the need for data diversity, including temporal and spatial resolution as well as disease states. Because the major causes of death in modern societies are age-related diseases, such as cancer, cardiovascular disease, and neurodegenerative conditions, a model encompassing the entire life course would include the aging processes that predispose to – and may even drive – the most prevalent and devastating human diseases.

Discussion

Predictive applications of AI in biology, such as learning the language of proteins to understand their structures, properties, and functions [15,25], are widespread and impressive but primarily static. It is like understanding how changing a letter changes the meaning of a word, but it does not tell us how to construct sentences or interpret our genetic blueprint. To understand human biology, we need to understand the principles of biological transitions, namely at the cell level. In this context of biological time as a key variable, dynamic networks and the genome need decoding. We know what some genes do some of the time, but we do not know what most genes do most of the time. As such, we need a multimodal foundation model that allows us to put together how genes (words) give rise to sentences (cell functions) and narratives (phenotypes). While others have discussed the application of AI to develop a virtual cell [16], our ultimate ambition is to predict spatially and temporally coordinated trajectories in cells across the life course.

While admittedly ambitious, the goal in making biology predictable is to build a computational model that replicates all human life stages, including developmental steps and aging, from the genetic sequence alone. Here we argue that developing such a model is not only technologically feasible but also the long-term vision to understand the language of life, with profound implications for comprehending human biology in health and disease.

Acknowledgements: We are grateful to current and past members of the Genomics of Ageing and Rejuvenation Lab for valuable discussions, and in particular Ludovic Senez and Priyanka Raina for help with the data estimates calculations. During the preparation of this work the authors used AI-assisted tools, Grammarly, Quillbot and ChatGPT, to improve readability and language. After using these tools, they reviewed and edited the text as needed and take full responsibility for the content of the publication. Work in JPM lab is supported by grants from the Biotechnology and Biological Sciences Research Council (BB/V010123/1), Longevity Impetus Grants and Hevolution Foundation, and LongeCity.

References

1. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A: **Language models are few-shot learners.** *Advances in neural information processing systems* 2020, **33:**1877-1901.
2. Sarumi OA, Heider D: **Large language models and their applications in bioinformatics.** *Comput Struct Biotechnol J* 2024, **23:**3498-3505.
3. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, et al: **Highly accurate protein structure prediction with AlphaFold.** *Nature* 2021, **596:**583-589.
4. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al: **Evolutionary-scale prediction of atomic-level protein structure with a language model.** *Science* 2023, **379:**1123-1130.
5. Zrimec J, Borlin CS, Buric F, Muhammad AS, Chen R, Siewers V, Verendel V, Nielsen J, Topel M, Zelezniak A: **Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure.** *Nat Commun* 2020, **11:**6141.

 doi:10.20944/preprints202602.1094.v1

6. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, Wang B: **scGPT: toward building a foundation model for single-cell multi-omics using generative AI.** *Nat Methods* 2024.
7. Hwang Y, Cornman AL, Kellogg EH, Ovchinnikov S, Girguis PR: **Genomic language model predicts protein co-regulation and function.** *Nat Commun* 2024, **15:**2880.
8. Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, Wang T, Ma J, Zhang X, Song L: **Large-scale foundation model on single-cell transcriptomics.** *Nat Methods* 2024.
9. Fabris F, Palmer D, Salama KM, de Magalhães JP, Freitas AA: **Using deep learning to associate human genes with age-related diseases.** *Bioinformatics* 2020, **36:**2202-2208.
10. Topol EJ: **Learning the language of life with AI.** *Science* 2025, **387:**eadv4414.
11. Hood L: **Systems biology: integrating technology, biology, and computation.** *Mech Ageing Dev* 2003, **124:**9-16.
12. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Lopez Carranza N, Grzywaczewski AH, Oteri F, Dallago C, Trop E, de Almeida BP, Sirelkhatim H, et al: **Nucleotide Transformer: building and evaluating robust foundation models for human genomics.** *Nat Methods* 2025, **22:**287-297.
13. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, Mort M, Cooper DN, Sebat J, Iakoucheva LM, et al: **Inferring the molecular and phenotypic impact of amino acid variants with MutPred2.** *Nat Commun* 2020, **11:**5918.
14. Raina P, Guinea R, Chatsirisupachai K, Lopes I, Farooq Z, Guinea C, Solyom CA, de Magalhaes JP: **GeneFriends: gene co-expression databases and tools for humans and model organisms.** *Nucleic Acids Res* 2023, **51:**D145-D158.
15. Bepler T, Berger B: **Learning the protein language: Evolution, structure, and function.** *Cell Syst* 2021, **12:**654-669.e653.
16. Bunne C, Roohani Y, Rosen Y, Gupta A, Zhang X, Roed M, Alexandrov T, AlQuraishi M, Brennan P, Burkhardt DB, et al: **How to build the virtual cell with artificial intelligence: Priorities and opportunities.** *Cell* 2024, **187:**7045-7063.
17. Kamimoto K, Stringa B, Hoffmann CM, Jindal K, Solnica-Krezel L, Morris SA: **Dissecting cell identity via network inference and in silico gene perturbation.** *Nature* 2023, **614:**742-751.
18. Rackham OJ, Firas J, Fang H, Oates ME, Holmes ML, Knaupp AS, Consortium F, Suzuki H, Nefzger CM, Daub CO, et al: **A predictive computational framework for direct reprogramming between human cell types.** *Nat Genet* 2016, **48:**331-335.
19. Cui H, Tejada-Lapuerta A, Brbić M, Saez-Rodriguez J, Cristea S, Goodarzi H, Lotfollahi M, Theis FJ, Wang B: **Towards multimodal foundation models in molecular cell biology.** *Nature* 2025, **640:**623-633.
20. Fu X, Mo S, Buendia A, Laurent AP, Shao A, Alvarez-Torres MDM, Yu T, Tan J, Su J, Sagatelian R, et al: **A foundation model of transcription across human cell types.** *Nature* 2025, **637:**965-973.
21. Avsec Z, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR: **Effective gene expression prediction from sequence by integrating long-range interactions.** *Nat Methods* 2021, **18:**1196-1203.
22. Linder J, Srivastava D, Yuan H, Agarwal V, Kelley DR: **Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation.** *Nat Genet* 2025, **57:**949-961.
23. Bryant DH, Bashir A, Sinai S, Jain NK, Ogden PJ, Riley PF, Church GM, Colwell LJ, Kelsic ED: **Deep diversification of an AAV capsid protein by machine learning.** *Nat Biotechnol* 2021, **39:**691-696.
24. Ahlmann-Eltze C, Huber W, Anders S: **Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines.** *Nat Methods* 2025, **22:**1657-1661.
25. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, Olmos JL, Jr., Xiong C, Sun ZZ, Socher R, et al: **Large language models generate functional protein sequences across diverse families.** *Nat Biotechnol* 2023, **41:**1099-1106.