ISSN 2563-7568

SHORT COMMUNICATION

Building Foundation Models in Biology

Sumanth Pareekshit Venkatesh Murthy *

Data Science Manager, Artefact, 62-64 Queen St Place, London, United Kingdom

Abstract

Over the last three years, Foundation models such as Dall-E and ChatGPT have taken the world by storm and have ushered in the "AI Boom". The next challenge is to build such models in Biology. This article examines the way text-based foundation models were built and the ways in which the approach has to be tweaked to build Foundation models in Biology. More specifically, it looks at the three components of the scaling laws - Data, Architecture and Compute and how they can be adapted to build foundation models in Biology. These Foundation models can then be used for a variety of downstream tasks such as Identification and if possible, prevention of conditions, treatment planning and nutrition planning.

Key Words: *Foundation models; Digital biology; Neural networks; Proteins; Graph-based neural networks; Multimodal architecture; Clinical datasets; Bayesian reasoning*

1. Introduction

Foundation Models (models which are trained on large corpuses of data, typically hundreds of gigabytes) in Biology which can integrate different forms of data is the next frontier in Digital Biology. This requires training multimodal architectures capable of understanding text, imaging, video and any other forms of data. These models would represent an extension of the current State of the Art (SOTA) such as EVO or AlphaFold3 which are respectively focused on DNA sequence creation and Protein-structure prediction. This article explores how multimodal Foundation Models in Biology can be built by integrating data from different sources and the challenges associated with such an endeavour.

2. What is a Foundation Model?

A foundation model is a neural network that is trained on a massive corpus of data and can derive insights using this large training dataset. The dominant players in this space are Anthropic, OpenAI, Mistral, Meta and others. Their foundation models have been trained on data available on the internet, like Reddit, Wikipedia and X. However, these models themselves do not have access to the internet, for obvious reasons of safety and security.

***Corresponding Author**: Sumanth Pareekshit Venkatesh Murthy, Data Science Manager, Artefact, 62-64 Queen St Place, London, United Kingdom, E-mail: sumanthpv.venkateshmurthy@alumni.utoronto.ca*

***Received Date**: November 21, 2024, **Accepted Date**: December 03, 2024, **Published Date**: December 11, 2024*

***Citation:** Murthy SPV. Building Foundation Models in Biology. Int J Auto AI Mach Learn. 2024,4(2):150-154.*

3. Why Biology?

Jensen Huang, the CEO of Nvidia says [1] that the "next big innovation" will come from Digital Biology. Biology is a field where research keeps happening every day. Even then, we haven't been able to completely map the networks in the human brain so far. Only very recently have researchers managed to map the connectome of a common fruit fly; doing the same for a human is a mammoth and daunting task. Progress in this domain can yield to faster diagnoses, faster identification of causes and faster development of drugs.
Let us look at two of the most important components to build Foundation models, data and architecture. The cost of compute infrastructure has gone down significantly so that it can be addressed at a later stage.

4. Data Sources for a large-scale Biology Foundation Model

4.1. Scientific literature and preprint

BioRxiv, PubMed Central, Nature/Science and Google Scholar, among others.

4.2. Molecular and structural data

Protein related data from PDB, UniProt, AlphaFold DB, PFAM and String. Genetic Data from GenBank, Encode, GTEx and dbSNp.

4.3. Clinical and medical data

UK Biobank, TCGA, ImageNet Med, ClinVar. Disease data from OMIM, DisGenET, DrugBank and PharmaGKB. Data from other countries such as the USA, Nigeria and South Africa should also be included to maintain diversity of the dataset and to account for different conditions that develop in these populations. The data from USA (and the UK) are valuable resources since both of these countries are very diverse with people from all parts of the world and unfortunately, many of the malign conditions in the populations from mainland find their way into the diaspora as well. An example here would be heart conditions and Type-2 diabetes among the South Asian diaspora in the UK and the USA.

4.4. Biological networks

KEGG, Reactome, BioGrid and MetaCyc.

4.5. 100000 genomes and UK biobank

For sequenced data to start with. For more accurate data, sequenced genome from Ilumina can be obtained (with one genome costing about USD 250). Genomes and DNA data can be included at a later stage since 99.9% of the genetic makeup is very similar. However, sequenced genomes are necessary for specific edge cases, and this can add further insights into the development of the condition.

4.6. Real-time data sources

Once a model is built it needs to continuously learn, and real-time data can be injected from multiple sources. A uniform way to manage this is to build a platform where people such as

clinicians, lab managers, hospital staff and even civilians can add data continuously. Clinicians can add their latest reports which can be stored in a vector database for fast access and inference, lab managers and hospital staff can upload the latest results and observations onto the portal. The biggest challenges here are data and feature engineering as well as system design to handle massive loads for model inference. Consumer apps which track diets can also be connected to the portal (which obviously operates with the model as a backend for inference) and upload data continuously. Here, we need to be cognizant of user and data privacy.

5. Bundling Models: Horizontal Vs. Vertical

One of the best parts about the scaling laws [2] is that models learn much better when fed huge corpuses, from sources listed in the point above. Following a similar trajectory, research teams at DeepMind and other organizations have trained models focused on tasks such as Prediction of the structure of protein [3] and Represent protein dynamics [4]. Despite large breakthroughs, these models remain highly localized, i.e. capable of analyzing one form of input data and doing one task, unlike general-purpose foundation models.

6. Vertical Approach

The approach in which individual foundation models are built, for tasks such as Protein Folding or Molecular Dynamics analysis, for example, is called the vertical approach, because combining many such models to act as one foundation model will not yield optimal results as these are disparate units.

7. Horizontal Approach

Instead of looking at individual components such as Proteins, Genome sequences and Biomarkers and building models based on these, a more effective approach is to consider specific diseases and conditions, since mapping such conditions involve protein measurements, biomarker measurements and potentially genome sequences. Imaging data from CRT, MRI and Whole Slide Imaging (WSI) can also be added to such datasets to obtain a comprehensive dataset suitable for training. Such a model would learn to develop casual relationships as well as learn how different biological pathways are present in the human body, based on the most updated information. This model can be further broken down into several models, each of which caters to a particular disease/condition, which again requires integrating several different forms of data.

A foundation model in biology, in theory, should then be capable of:

- Giving a root cause analysis for a disease
- Identifying the relevant biomarkers/biomarker levels for a condition
- Tracing the network pathway for a signal, or lack thereof
- Locating binding sites
- Drug discovery is based on the mapped network pathways.

Going further, we have a choice of using one or more of the following techniques to build such large Foundation models:

7.1. Graph-based neural networks

In graph based neural networks, entities (proteins, biomarkers, genes) are encoded into nodes while relationships are modeled as the edges between these nodes.

7.2. Bayesian reasoning

Pre-train the model with the existing data and available methods and fine-tune using the updated research. This allows the model to continuously update itself based on the priors and the new information.

7.3. Multimodal dataset and architecture

Biological data comes in many forms, hence the data preprocessing and reading steps should be capable of handling image, text, video and numerical data. A simple rule of thumb can be "Start Simple" since many problems can be solved with more interpretable methods such as Random Forests or Decision Trees. A recent paper [5] which has introduced the Proteomic Aging clock has used ensemble methods such as Cox Regression to identify features that affect age-related conditions.

8. Conclusion

Foundation Models in Biology which handle a variety of tasks can be built since the hardware is scaling quickly and is available. The primary challenge remains the collection and curation of the most relevant and cleanest datasets available from multiple sources since no single source contains all the required dataset. Proper methods need to be chosen for training and continuous learning since new data from patients can be incorporated to further improve the understanding of various conditions and diseases.

9. Further Reading

The second biggest barrier is assembling enough compute power to train the models. The cost to train ChatGPT 4 was more than [6] $100 million and it's only bound to increase. AlphaFold2's dataset has a size of around 23TB [7] and other foundation models have been trained on datasets which are as big. Naturally, it is a challenge to obtain such enormous compute infrastructure, alongside the power required to run these machines during training and inference.

References

1. AI Is Moving Biology From Science To Engineering, Advancing Medicine. https://www.forbes.com/sites/gilpress/2024/04/30/ai-is-moving-biology-from-science-to-engineering-advancing-medicine/ (accessed April 2024).

2. Askell A, Bai Y, Chen A, et al. Scaling Laws and Interpretability of Learning from Repeated Data. https://www.anthropic.com/research/scaling-laws-and-interpretability-of-learning-from-repeated-data. (accessed May 2022).

3. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. nature. 2021;596:583-9.

4. Hou C, Shen Y. SeqDance: A Protein Language Model for Representing Protein Dynamic Properties. BioRxiv. 2024.

5. Argentieri MA, Xiao S, Bennett D, et al. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. Nature medicine. 2024;30:2450-60.

6. Knight W. OpenAI's CEO says the Age of Giant AI Models is Already Over. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/. (accessed April 2023).

7. https://github.com/google-deepmind/alphafold/blob/main/afdb/README.md