Genolator: A Multimodal Large Language Model Fusing Natural Language, Genomic and Structural Tokens for Protein Function Interpretation

Martin Danner[1,2], Tanhim Islam[1], Matthias Begemann[1], Florian Kraft[1], Miriam Elbracht[1], Ingo Kurth[1], Jeremias Krause[1]

1. Center for human genetics and genomic medicine, Medical Faculty, Uniklinik RWTH Aachen, Pauwelsstrasse 30, Aachen, 52074, North-Rhine-Westphalia, Germany
2. scieneers GmbH, Kantstraße 1a, Karlsruhe, 76137, Baden-Wuerttemberg, Germany

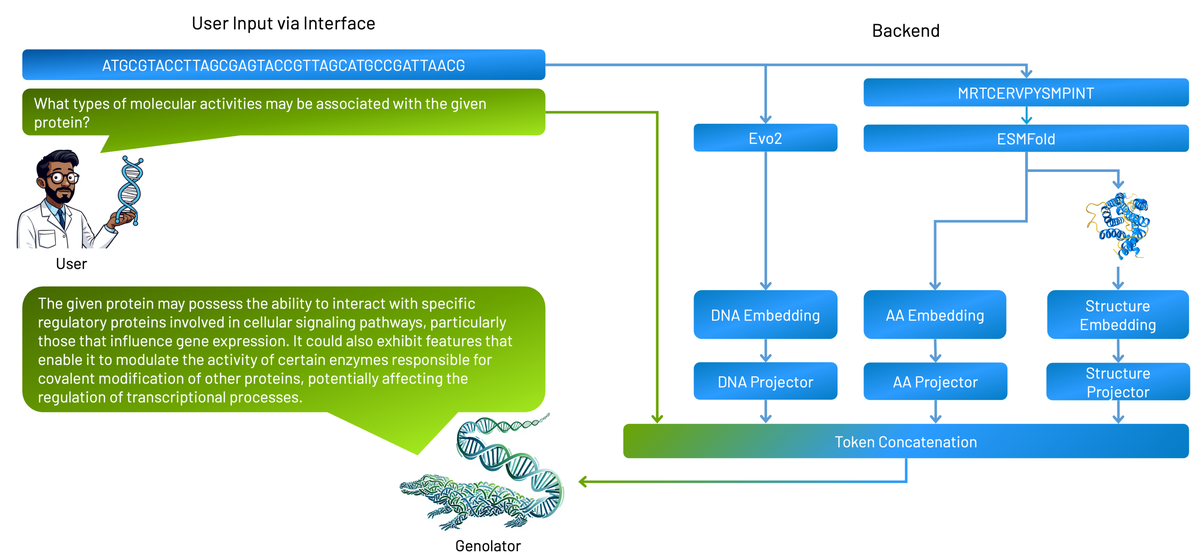
POSTER AWARD CANDIDATE

Background

Decoding the genetic code to unveil its genome functionality is a monumental task which would greatly advance the understanding of disease mechanisms and development of targeted treatment approaches. Although large language models (LLMs) have transformed natural language processing across diverse domains, translating the complex language of DNA into human-readable form remains challenging due to genomic data complexity and unexplored regions of the human genome. Current (genomic) language models either are capable of processing natural language or the genomic code. Models fusing both aspects are largely lacking[1].

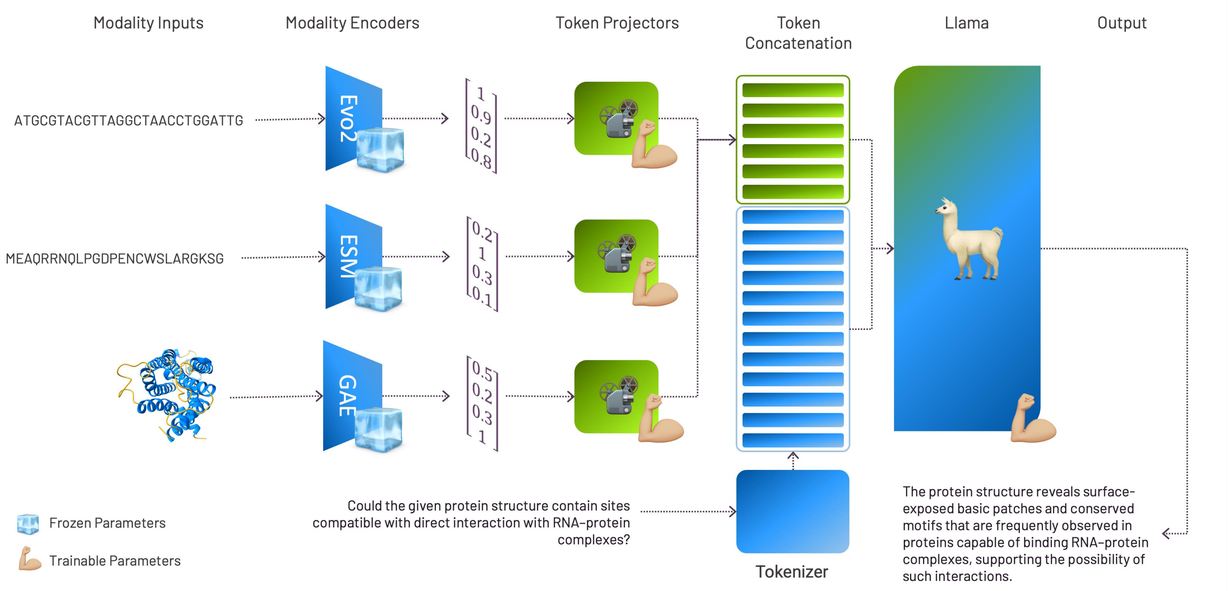
A multimodal LLM regarding gene functionality

Genolator is a finetuned Llama[2] model, designed to fuse natural language with genomic language. It receives a question, formulated in natural language (English) and information from DNA-sequences, amino acid sequences and protein structures in tandem. Each modality is presented to Genolator in the form of an embedding, a latent representation of the encoded information, from a machine learning model trained on the specific modality. It can handle confirmative/denial questions, as well as openly formulated questions regarding gene functionality.



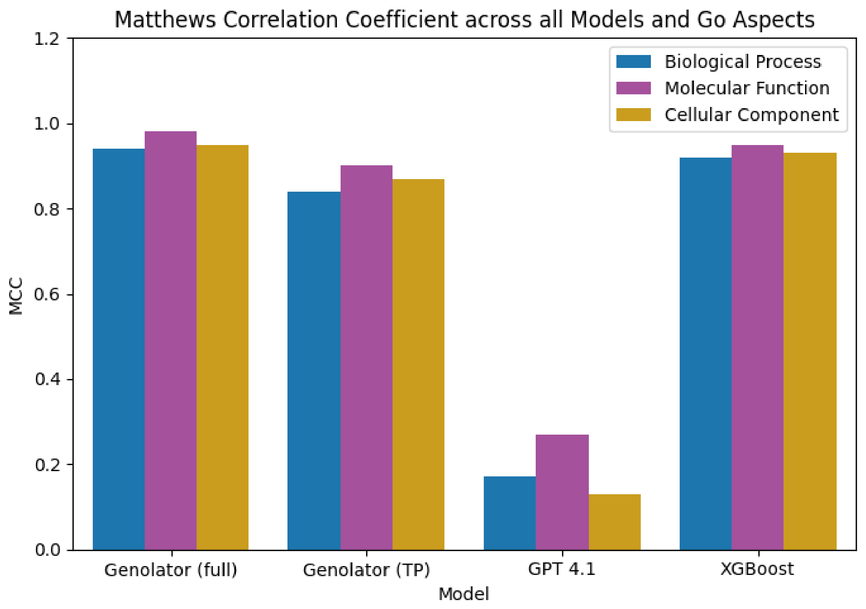
Methods

Genolator was fine-tuned on over 365,000 question-answer pairs generated using abstracted Gene-Ontology (GO)[6] terms. To obtain genomic representations, we utilized Evo2 7B[3]. Amino acid sequence embeddings were generated using the ESMFold[4] (3B) model. Structural protein embeddings were generated using a Graph Convolutional Autoencoder[5]. For the Language Model part we selected Bio-Medical-Llama-3 8B provided by ContactDoctor on Hugging Face. To enable multimodal integration within the Llama[2] architecture, we mapped the modality embeddings into Llama's[2] native token embedding space. This was achieved by designing three separate virtual token projectors. Each projector applied a learnable linear transformation to map the input embedding dimension to the hidden size expected by Llama[2] (4098). For each sample, eight virtual tokens per modality were produced, yielding a total of 24 multimodal virtual tokens, which were concatenated and prepended to the textual input tokens.



Results 1: Answering Protein Function Queries

Genolator effectively answers comfirmative and denial queries regarding protein subcellular localization, molecular function and biological processes. Evaluation demonstrates high accuracy and MCC[7] in confirming or denying protein function associations, outperforming baseline models such as openly available allrounder LLMs like GPT 4.1[8] as well as smaller domain specific models integrating knowledge from foundation models like Evo2[3] and ESM2[4].



Results 2: Genolator Hidden State Analyis

To assess whether the model effectively integrated natural language and biological information, we investigated the features learned by Genolator. We extracted the hidden states from the final layer of the Llama[2] model prior to the output layer and projected them into a two-dimensional space using t-distributed stochastic neighbour embedding (t-SNE)[9]. For most GO[6] terms, Genolator achieved a clear separation between confirmed and denied associations in the t-SNE[9] space, producing distinct clusters across all GO-term[6] aspects. Additionally, within each GO[6] aspect, Genolator clustered related terms in close proximity, while clearly separating less related terms.

Literature

1. de Almeida, B. P. et al. A multimodal conversational agent for DNA, RNA and protein tasks. Nat. Mach. Intell. 7, 928–941 (2025).
2. Grattafiori, A. et al. The Llama 3 Herd of Models. Preprint at https://doi.org/10.48550/arXiv.2407.21783 (2024).
3. Brixi, G. et al. Genome modeling and design across all domains of life with Evo 2. 2025.02.18.638918 Preprint at https://doi.org/10.1101/2025.02.18.638918 (2025).
4. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130 (2023).
5. Danner, M., Begemann, M., Elbracht, M., Kurth, I. & Krause, J. Utilizing protein structure graph embeddings to predict the pathogenicity of missense variants. 2024.11.15.623748 Preprint at https://doi.org/10.1101/2024.11.15.623748 (2024).
6. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29 (2000).
7. Chicco, D. & Jurman, G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Min. 16, 4 (2023).
8. OpenAI et al. GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2024).
9. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605 (2008).

Disclosure Statement of Financial Interest: I, Jeremias Krause , DO NOT have a financial interest or affiliation with one or more organizations that could be perceived as a real or apparent conflict of interest in the context of the subject of this presentation.
This research project was funded by the START-Program and the Clinician-Scientist-Program of the Faculty of Medicine, Uniklinik RWTH-Aachen.