**AI-Driven Genomics**

**Risks, Ethics, and Governance of “EVO 2”**

**Section 1**

1. **Introduction**

Evo 2 is a recent AI model designed to revolutionize the way we work and learn with genomes. It was released on February 19, 2025, through collaboration between the Arc Institute, NVIDIA, and Stanford University, UC Berkeley, and UC San Francisco researchers [1]. Evo 2 is an open-source model, so researchers and developers around the world can use it and help improve it. Evo 2 employs a larger and more diverse training set that includes bacteria, archaea, eukaryotes, and bacteriophages. This enables it to identify intricate gene structures and predict correctly how gene changes influence their function.

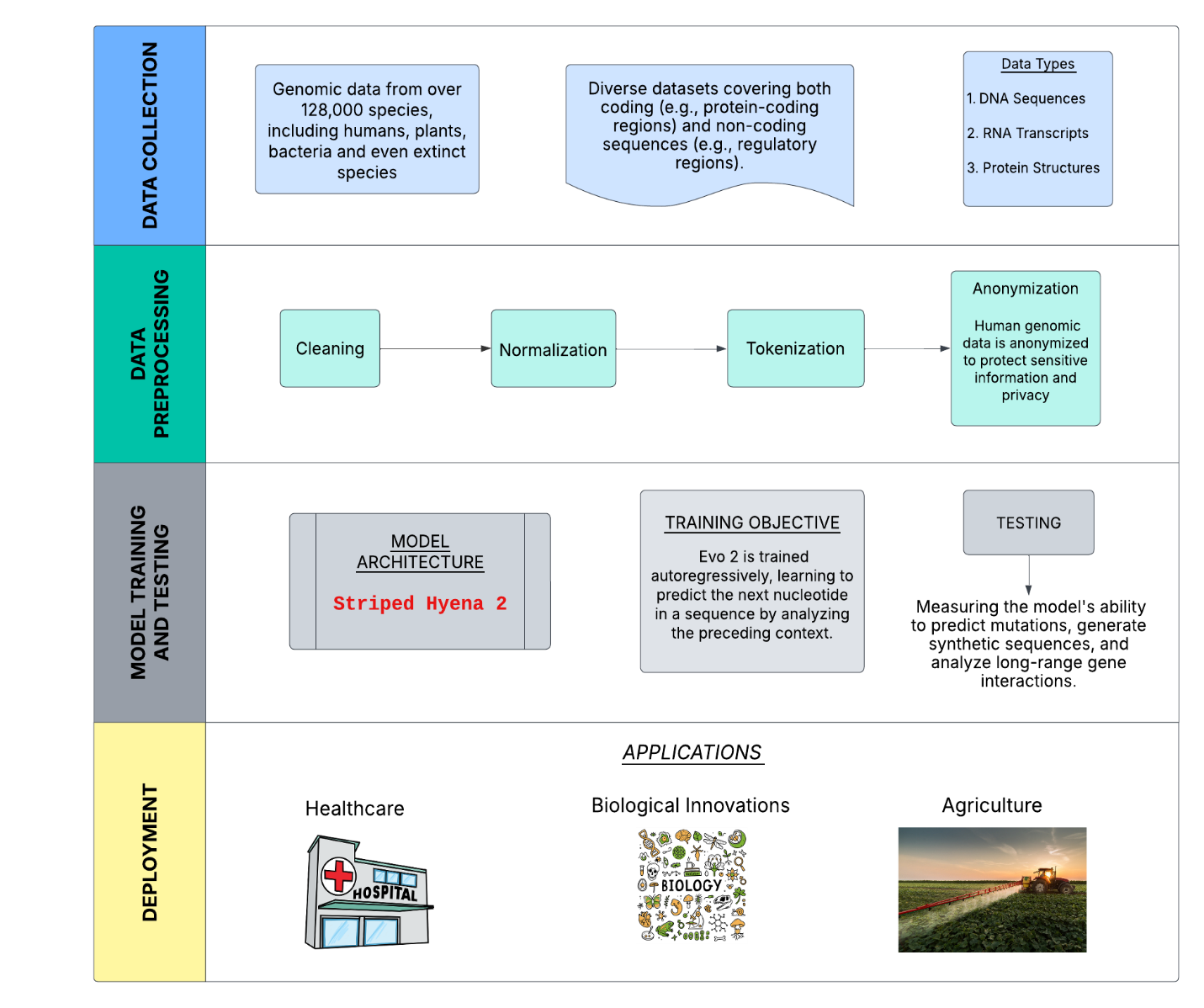
Evo 2 is designed to process extremely long sequences of genes, up to 1 million base pairs. This enables it to examine interactions in huge regions of DNA. The model is available in two forms: one with 7 billion parameters for regular work and the other with 40 billion parameters for hard and complex work [1].

1. **Data Pipeline**

The Evo 2 data pipeline is highly crucial in training its model, which has 40 billion parameters and processes sequences as long as 1 million base pairs. It starts with gathering and sorting the “OpenGenome2” dataset, which contains over 9.3 trillion nucleotides from various organisms [1]. The dataset is sorted in a precise manner to ensure that it is high quality and contains no eukaryotic viral sequences to avoid bias.

Preprocessing transforms raw nucleotide sequences into usable forms for training. Sequences are divided into smaller segments, focusing on gene regions in the initial stage of training. Data are divided into training, validation, and test sets to avoid data leakage and ensure proper evaluation. Training has two components. The first component pretrains the model on 8,192 tokens. It is concentrated on nearby things such as genes and regulatory elements. This component assists in making training faster and more efficient. The second component, referred to as midtraining, expands the context to 1 million tokens. This enables the model to comprehend larger regions in the genome.

The model employs the StripedHyena 2 architecture, which combines convolution techniques and attention mechanisms to perform well on both short and long sequences. It is significantly faster than standard Transformer models [2]. The model is tried to understand how well it performs on synthetic and biological tasks. For instance, it tests whether it can recall a 100-base-pair sequence from a larger 1-million-base-pair background, predict mutational effects and structure them without individual changes for every task.



*Figure 1. Data Pipeline of EVO 2*

1. **Risks, Ethics, and Governance Assessment:**

**3.1 Risks:**

Evo 2 is crucially risky in its application to genomics. There are a few areas of risk namely data privacy. Then, as Evo 2 must handle much genetic data, there have to be robust mechanisms to secure individual information. A very important thing is to anonymize and store personal genetic data safely from misuse and unauthorized use. The potential applications of Evo 2, such as genetic modifications, could lead to ethical dilemmas regarding the extent and nature of human intervention in natural processes. In addition, the people whose genetic data are uploaded need to be provided informed consent that they know their data are being used in this way, and they know what that usage might lead to. It is a major issue. This means that Evo 2, which was trained on different kinds of data, needs to be very well monitored so that it does not inherit old biases or introduce new ones. In other words, the model has to be periodically monitored and trained because its predictions and outcomes are fair and equitable.

The use of AI in genome editing practices presents safety risks because off-target effects alongside unintended genetic modifications can occur. Proper management strategies must be established to prevent dangerous consequences from occurring [8]. For example, altering a genetic code can modify a gene which can be dangerous to humans. Therefore, it is vital to establish strong ethical standards and legislations to mitigate risks and guarantee that Evo 2 is utilised properly for genome editing [4].

**3.2 Ethics:**

Evo 2 genome editing is of great importance because of the high ethical implication of the two concepts of autonomy and informed consent. Genetic data and gene changes are related to individual rights of making the decisions of self-informed. Evo 2 received no base data about pathogens that infect humans or complex organisms and the system blocked valuable research results when users posed queries about these pathogens[1]. This should inspire the researchers and the practitioners to give out clear and concise information about how the evo 2 gene editing works, its pros and the cons and the future effects towards it. We must respect people’s autonomy and obtain informed consent for ethical practice of the technology as well as for people’s confidence [4].

Over the past decade, governments and companies have spent a lot of time and money to apply AI to genomic science. The primary cause of this expansion is based on advances in machine learning and deep learning as exemplified by Evo 2 with progressing to greatly impact how genomic data is collected, researched, and applied [6]. The fact that there is more interest about such a subject made the number of scientific publications on AI and genomics, including Evo 2, increase each year starting from 2017 [6]. The AI market in genomics is projected to grow in next decade and the projections are diverse and positive. These developments could have a very big impact, such as new advances in drug discovery, as well as being able to predict complicated human traits with genomic data [6].

**3.3 Governance:**

Addressing the ethical and risk issues that Evo 2 will involve requires the use of governance structures. Imposing some rules on how to use Evo 2 genomics requires international cooperation and coming to a common set of rules. Public values and priorities should guide the governance models that control AI applications in genomic research. The process includes public discussions to comprehend the effects of AI-based genomic health prediction together with guiding technology developments to serve public needs [7]. Comprehensive rules on data privacy, accountability, and transparency in the use of AI-driven genomics can be developed on these basis of EU laws [5].

The current data stewardship scope moves beyond historical data processing to encompass a responsibility for real-time data streams. Consistent data policies need development together with agile frameworks combined with human oversight to handle risks that emerge in AI systems such as bias-related problems [9]. The development and application of AI-based genomics, like Evo 2, are amongst the most controversial technologies of the 21st century, and even EI ethical, political, and legal issues are involved [6].

1. **Conclusion:**

AI ethics and guidelines for genomics with Evo 2 are crucial to apply AI in genetic research and medicine in an ethical and equitable manner. Evo 2 has the potential to tackle challenges such as data privacy, unequal algorithmic biases, unanticipated outcomes, individual liberty, informed consent, social equity, and engaging in global collaboration; however we can do this with the help of ethical guidelines and with public confidence.

**5. References:**

[1] Arc Institute, "Genome modeling and design across all domains of life with Evo 2," GitHub, 2025.

[2] Together Computer, "StripedHyena: Repository for StripedHyena, a state-of-the-art beyond Transformer architecture," GitHub, 2023.

[3] R. Bouderhem, "Shaping the future of AI in healthcare through ethics and governance," *Humanit Soc Sci Commun*, vol. 11, 2024.

[4] O. Francis I., "Ethical Considerations in AI-Driven Genome Editing," *Faculty of Medicine Kampala International University Uganda*, vol. 10, 2024.

[5] Challen, R., Denny, J., Pitt, M., et al., "Artificial intelligence, bias and clinical safety," *BMJ Quality & Safety*, vol. 28, 2019.

[6] Raza, S., "Artificial Intelligence for Genomic Medicine," PHG Foundation, Mar. 2020.

[7] H. Farmer, A. Strait, F. Bennett, C. Joynson, and P. Mills, "DNA.I.: Early findings and emerging questions on the use of AI in genomics," Ada Lovelace Institute and Nuffield Council on Bioethics, Aug. 2023.

[8] M. Chustecki, "Benefits and Risks of AI in Health Care: Narrative Review," Interactive Journal of Medical Research, vol. 13, 2024.

[9] A. A. Taddese, A. C. Addis, and B. T. Tam, "Data stewardship and curation practices in AI-based genomics and automated microscopy image analysis for high-throughput screening studies: promoting robust and ethical AI applications," *Human Genomics*, vol. 19, no. 1, Feb. 2025.